Eindhoven University of Technology

MASTER

Comparing GHG emissions of outsourced transport operations

Cornelissen, Laura M.

Award date:
2024

Link to publication

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 22. Mar. 2025

Eindhoven University of Technology

Master Thesis

Comparing GHG emissions of outsourced transport operations

Author:

Laura Cornelissen 1260596

Supervisors:

Layla Martin	TU/e
Yeqing Zhou	TU/e
Sofie E.	Company
Sofie W.	Company

February 9, 2024

Abstract

This study introduces a novel methodology designed to compare the emission intensities in gram CO2e per tonne-kilometer of Logistics Service Providers (LSPs) across varying data quality and availability levels. By incorporating confidence and prediction intervals into a regression model, we offer a robust method of quantifying the uncertainty and variability associated with emission intensity estimations. Through careful feature selection and prediction model refinement, we ensure small confidence and prediction interval widths, thereby enabling accurate and insightful comparisons between LSPs. This study reveals significant differences between the emission intensities of two LSPs and industry averages. This discrepancy can be partly explained by dedicated contracted shipments, which typically entail high empty distances and result in high emission intensities. Additionally, this research emphasizes the importance of data quality and availability on the accuracy of the prediction.

Preface

Completing this thesis has been a challenging yet rewarding journey. It has allowed me to deeply explore the nuances of sustainable transportation and contribute to the existing body of knowledge in this field.

Throughout the process of researching, analyzing data, and writing, I have benefited from the support and guidance of many individuals. First, I want to thank the anonymous industry partner for the opportunity to perform my thesis and especially Sofie E. and Sofie W. for the guidance and meetings. Besides, I want to thank my supervisor from the University, Layla, for her guidance and feedback throughout the process. Your handwritten feedback and advice helped me write, structure, and take a more critical look at my thesis. Additionally, I would like to thank my second supervisor Yeqing Zhou for her helpful feedback, encouragement, and valuable insights.

Lastly, I am grateful for the mentorship of my academic advisors, the collaboration of my peers, and the encouragement of my friends and family.

Laura Cornelissen, Eindhoven, February 9, 2024

Contents

List of Tables

List of Figures

Executive Summary

Objective

As transportation accounts for 23-30% of GHG emissions, accurately measuring emissions in transportation and distribution is crucial (Herold & Lee, 2017; Schmied, Knörr, & Hepburn, 2012; Wild, 2021). However, companies often face data quality and availability challenges, making it difficult to estimate emissions accurately. These challenges result in a lack of comparability between Logistics Service Providers (LSPs) due to variations in calculation methods, data quality and availability levels. This issue is particularly problematic when transportation is outsourced to multiple LSPs. Therefore, comparing LSPs based on their emissions can provide valuable insights into their environmental performances and help companies make better decisions. As of 2024, the European Union mandates reporting of Scope 3 (indirect) emissions, a step toward implementing a carbon tax for transportation. In this system, carriers are responsible for paying the transportation carbon taxes, which they will then pass on to the end users. From an economic standpoint, it is beneficial for companies to collaborate with LSPs to have low emissions to ensure compliance with current and future environmental regulations. In light of these factors, this research aims to determine the impact of data quality and availability on the accuracy of estimating indirect GHG emissions for companies that outsource transport operations and make a comparison between Logistics Service Providers (LSPs).

Methodology

This study provides a case study conducted at a chemical production company that outsources all transport operations. It aims to compare the emissions of Logistics Service Providers while differing in data quality and availability. For this purpose, we develop an algorithm that considers all combinations of features and excludes combinations that violate assumptions to include confidence and prediction intervals of a multiple regression model that predicts emission intensities. This algorithm enables us to compare LSPs while quantifying the uncertainty and variability of the estimations. We use Ordinary Least Squares to determine the regression model. Model validation involves in-sample and out-of-sample techniques, including Leave-One-Out Cross-Validation and comparison with industry averages.

Numerical experiments

The numerical experiments consist of two phases to compare LSPs with different data quality and availability. First, we determine and validate the best regression model with high predictive accuracy for each LSP per modality. For this purpose, we perform a sensitivity analysis with different thresholds for the exhaustive feature selection algorithm. We find that careful selection of the significance level threshold in the exhaustive feature selection algorithm can yield comparable results to stepwise regression while offering automation and ease of use. Furthermore, the exhaustive feature selection algorithm provides valid models even with limited data. In road transportation, including empty and loaded distance data significantly improves predictive accuracy, showing a 2 to 3-fold enhancement over using only great circle distance. This highlights the importance of detailed data for precise emission predictions. For intermodal transportation, utilizing great circle distance results in a narrower confidence interval than actual loaded distance, potentially due to a misclassification error in one lane's transportation modality, emphasizing the critical need for data accuracy in predictive model performance. The prediction model for LSP 2 captures less variability but exhibits higher relative variability in individual observations compared to LSP 1, resulting in a wider prediction interval width for LSP 2. Additionally, prediction and confidence interval widths increase with higher proportions of empty distance in road transportation, indicating

greater variability and uncertainty likely due to unaccounted influential factors or data errors.

Our analysis reveals markedly lower emission intensities for LSP 2 than for LSP 1. This discrepancy can be partly attributed to LSP 1's involvement in dedicated contracted shipments, which typically entail high empty distances and result in elevated emission intensities. Furthermore, we observed that the emission intensities for full truckload shipments at LSP 2 are 1.3 to 1.6 times lower than industry averages. This could stem from underestimated emissions, superior carbon performance, or overestimated default factors. Conversely, spot-contracted emission intensities for LSP 1 were found to be 1.2 to 1.6 times larger than industry averages. This disparity may arise from overestimated empty distances, errors in distance allocation, misclassification errors, measurement inaccuracies, or elevated proportions of empty distances compared to industry norms. Moreover, LSP 2 demonstrated superior performance in emission intensity and total well-to-wheel emissions compared to LSP 1 for similar lanes.

Conclusions

In conclusion, this study underscores the pressing need for accurate estimation and comparison of greenhouse gas emissions in the transportation sector. By developing a novel approach to compare emissions intensity among LSPs, this research advances our understanding of environmental impact assessment and regulatory compliance. The findings highlight the importance of addressing data quality and availability challenges to ensure accurate emission estimates and facilitate informed decision-making.

1 Introduction

Greenhouse gas (GHG) emissions have a significant environmental impact, requiring reduction (WRI & WBCSD, 2004). Consistent with this, the European Commission has mandated that large companies report their carbon footprint in the supply chain from the 1st of January, 2024 (EFRAG, 2023). This highlights the importance for companies to measure not only their own GHG emissions but also those throughout their value chain. The logistics sector finds value chain (Scope 3) emissions particularly relevant because purchased transportation represents the largest source of value chain (Scope 3) emissions (Herold & Lee, 2017). Additionally, European countries are obligated to achieve carbon neutrality by 2050, as stated by the European Commission. As transportation accounts for 23-30% of GHG emissions, accurately measuring emissions in transportation and distribution is crucial (Herold & Lee, 2017; Schmied et al., 2012; Wild, 2021).

Accurately estimating greenhouse gas emissions and the corresponding uncertainties is essential for reducing emissions and creating a transparent inventory (IPCC, 2006). However, accurately measuring indirect GHG emissions poses challenges due to complex and diverse supply chains and a lack of standardised methodologies (Royo, 2020). Insufficient data quality and availability further produce inaccurate results and hinder intercompany emissions comparisons. According to Auvinen et al. (2014), any comparison must meet a certain standard of accuracy to be meaningful. Therefore, high-quality data need to be gathered to achieve increasing levels of accuracy, increasing the complexity of the demands placed on all parties within the supply chain. On the contrary, there is a strong need for a uniform method to measure carbon footprints. Comparing Logistics Service Providers (LSPs) based on their emissions can help companies that outsource transportation to improve decision-making. From 2024 onwards, it will be mandatory to report indirect emissions. This means the European Union is now closer to implementing a carbon tax for transportation. Carriers, or emitters, will be responsible for paying the transportation carbon taxes, which will then be passed on to the comparing outsourcing transportation. Therefore, it is also, from an economic perspective, interesting to choose LSPs with low emissions. Choosing LSPs with lower emissions can ensure compliance with current and future environmental regulations. To address these issues, we determine the impact of data quality and availability on the accuracy of estimating GHG emissions for companies that outsource transport operations and make a comparison between LSPs.

1.1 Problem context

This study extends the limited research on comparing companies regarding their GHG emissions in the transportation sector. The ability to compare GHG emissions is crucial to assess carbon performance and to ensure compliance and data accuracy (Herold & Lee, 2017). Besides, comparing a company's sustainability performance to its competitors is beneficial to determine its competitive (dis)advantage. Moreover, it is helpful for companies that outsource their transportation to decide which partner to cooperate with to minimise their emissions. Despite the importance, literature often fails to compare company emissions. We found little literature on comparing GHG emissions and no literature on comparing GHG emissions in the transportation sector. To our knowledge, no study successfully compares GHG emissions of outsourced transport operations. However, studies recognise the lack of comparability (Davydenko et al., 2014; Royo, 2020; Radonjič & Tompa, 2018; Wegener et al., 2019). Therefore, this study fills this gap by comparing GHG emissions of transportation companies. For this purpose, we first address the challenges of comparability. We recognise the two main problems that hinder such comparability: the lack of a globally harmonised measurement method and the different data quality

and availability levels. We discuss these problems in further detail in the following subsections.

1.1.1 Towards a globally harmonised method

The first challenge is the lack of a globally harmonised measurement method. Multiple standard agencies, governments, and industry bodies provide advice on calculating and reporting GHG emissions, leading to a proliferation of different methods, tools, and databases (Royo, 2020). This lack of communication and standardisation creates confusion among stakeholders, making it challenging to select an appropriate method for emissions estimation (Royo, 2020). Radonjič and Tompa (2018) underscore the irrelevance of comparing carbon footprints among organisations unless the boundary conditions for including GHG emissions and the used estimation methods are similar enough to permit meaningful comparisons. Additionally, the costs associated with using accurate GHG estimation methods can often outweigh the benefits, leading some companies to opt for less precise methodologies, resulting in inaccurate measurements (Wegener et al., 2019). The lack of a harmonised measurement method for GHG emissions poses a significant challenge in accurately assessing and comparing emissions across companies, industries, and sectors (Davydenko et al., 2014; Royo, 2020; Radonjič & Tompa, 2018; Wegener et al., 2019). A method must be harmonised to enable comparable calculations for the wide range of modes, operational characteristics, and vehicle types that may be used along the value chain (A. Lewis, 2016).

Currently, many methods are available to calculate emissions. However, none applies to various data quality and availability levels. The GLEC Framework represents a first step towards a worldwide harmonised standard (Davydenko et al., 2014). Recently, the new standard ISO 14083:2023 was published based on the GLEC Framework. The GLEC framework presents the only worldwide harmonised calculation approach that focuses on transport, covers all transport modes, is fully regionally applicable, and incorporates the entire transport chain (Davydenko et al., 2014). Although many recommend the GLEC framework for its ability to standardise logistics emissions measurement and reporting, some critics have raised concerns about the framework's complexity and the resources required to implement it. Royo (2020) states that the framework requires reliable and representative data. However, obtaining reliable and representative data is difficult since measuring value chain (Scope 3) emissions involves many actors. Besides, Hörandner et al. (2023) state that comparison of emissions is still difficult using the GLEC Framework due to huge differences between fuel-based and distance-based calculations using emission intensity factors. Furthermore, assessing the estimate's uncertainty is crucial to compare emissions with multiple data quality and availability levels (Waldman et al., 2020). The GLEC Framework does not include this uncertainty. Another problem regarding the GLEC Framework is the value chain (Scope 3) emission calculation with limited data quality. The GLEC Framework provides emission intensity factors for some data quality and availability levels. However, an emission intensity factor is available only for a few specific combinations of variables. For example, for a road transportation leg in an articulated truck with up to 40 tonnes Gross Vehicle Weight, with Average/mixed load characteristics and LNG as fuel type, the emission intensity factor is 88 g CO_2e per tonne-kilometer. However, a company that outsources transportation often does not know the vehicle type of a shipment, making it impossible to determine the emission intensity factor. Thus, there is no method to determine the emission intensity for different data quality and availability levels; thus, we cannot estimate the emissions. This is an enormous problem since many companies cope with limited data, and it will become obligatory to report Scope 3 emissions.

To further explain the problem regarding the comparability using the GLEC Framework, we consider an example of two LSPs that perform logistics services for a company that outsources transport: LSPs 1 and 2. We have data on one specific road shipment, both LSPs. Both shipments have the load characteristics of a container shipment. For LSP 1, we know that the vehicle type is an articulated truck with up to 40 tonnes gross vehicle weight, having fuel type Diesel. Thus, we obtain a default emission intensity of 75 grams of CO2e per tonne-kilometer. For LSP 2, we know that the fuel type is Diesel, but we do not know if the articulated truck has a maximum capacity of 34 tonnes, 40 tonnes or 44 tonnes, resulting in emission intensity of 100, 75 and 67 grammes per tonne-kilometer, respectively. We cannot determine the emission intensity of LSP 2 using the GLEC framework because we do not know the truck's capacity. The inability to determine emission intensity is a huge problem, as companies often do not know vehicle characteristics if transportation is outsourced. Even if we knew the vehicle, fuel, and load characteristics, we would obtain a point estimate of the emission intensity, regardless of other parameters such as cleanings, heatings, traffic conditions, journey, or cargo type. Point estimates based on different input data result in incomparable results. For example, LSP 1 and 2 both have a shipment conducted with an articulated truck with a capacity of 40 tonnes with fuel-type Diesel carrying a container. Both would obtain an emission intensity of 75 grams per tonne-kilometer. However, the shipment performed by LSP 1 needs cleanings and heatings, while the shipment performed by LSP 2 does not. In reality, the shipment performed by LSP 1 would have a higher emission intensity than that of LSP 2, while the emission intensities obtained by GLEC are similar. This results in inaccurate emission intensities and could be improved by incorporating uncertainty using an interval rather than a point estimate. Currently, no method or tool estimates emissions while quantifying uncertainty to compare emissions of LSPs. Appendix 1 discusses other tools and databases and their inapplicability as a globally harmonised method for data quality and availability levels.

1.1.2 Data quality and availability

Data quality and availability present another significant challenge in estimating and comparing GHG emissions. This challenge is particularly true in transportation and distribution, as outsourcing logistics services is common. When services are outsourced, the value chain becomes more complex, making it more difficult to obtain data. Besides, as the supply chain becomes more intricate, data visibility and traceability requirements become stricter, posing challenges in making informed decisions for emission reduction strategies (Royo, 2020). Data quality and availability significantly influence the quality of the estimates. Velázquez-Martínez et al. (2014) quantify the influence of using more detailed estimation methods on the uncertainty of the estimates. Results reveal that highly aggregate methods, using standard emission factors and assumptions of full truckload utilisation, exhibited magnitude errors of up to 25%. Conversely, models assuming average load factors show a magnitude error of 16%. Certain estimation models systematically overestimate carbon emissions, while others systematically underestimate emissions in specific situations, indicating the presence of substantial and systematic aggregation errors (Velázquez-Martínez et al., 2014).

Moreover, the GLEC Framework recommends the use of data on fuel use. However, if this is unavailable, the distance-based method should be used. This method depends on information on parameters such as vehicle type, vehicle mode, weight, distance or load characteristics to determine the emission intensity factor. This results in the inability to accurately compare emissions of different companies with different data quality and availability levels. For companies, it is interesting to know what parameters are

relevant to determine the emission intensity factor. Companies could save time by only collecting the relevant parameters to improve the precision. Besides, governments could determine which parameters are obligatory to collect for companies to collect for emissions estimates.

According to Waldman et al. (2020), these significant data-quality issues, such as lack of common data sources, overuse of generic data sets, poor reliability of results, are caused by the use of point estimates instead of interval estimates and due to limited data availability. Using confidence intervals is an excellent method to quantify uncertainty and to cope with the lack of data quality and availability (Tong et al., 2012; Waldman et al., 2020).

1.2 Contributions and scope

The present study addresses multiple gaps. First and foremost, we introduce a novel methodology designed to compare the emission intensities of LSPs across varying levels of data quality and availability. By incorporating confidence and prediction intervals into our analysis, we offer a robust method of quantifying the uncertainty and variability associated with emission intensity estimations. Through careful feature selection and prediction model refinement, we aim to minimise confidence and prediction interval widths, thereby enabling accurate and insightful comparisons between LSPs.

We are the first to develop a freight transport emission intensity prediction model applicable to varying data quality and availability levels while quantifying uncertainty. For this purpose, we develop an algorithm that returns the regression model with the most appropriate set of features, predicting the emission intensity for various sets of features as input. The first phase involves utilising an algorithm for exhaustive feature selection and model filtering based on assumptions. This step yields the most accurate model for predicting emission intensity while considering confidence and prediction intervals to quantify uncertainty of the estimate. In a case study, we show that this algorithm works for various data quality and availability levels. This approach is novel since it combines exhaustive feature selection, stringent model filtering based on statistical assumptions, and confidence and prediction interval estimation within the context of predicting emission intensity. This comprehensive method, specifically tailored for emission intensity estimation, is unique in the domain. This phase sets the foundation by establishing reliable predictive models for various data quality and availability levels.

While numerous emission calculation methods exist, none sufficiently address the diverse data quality and availability encountered across companies while quantifying uncertainty. The GLEC Framework stands as the closest approximation to a harmonised approach. However, our assessment reveals that the GLEC Framework lacks universal applicability across varying data quality and availability levels and does not incorporate mechanisms for assessing uncertainty. Our study recommends augmenting the GLEC Framework by integrating our proposed algorithm. This integration would enable companies to determine emission intensity factors, including confidence and prediction intervals, accommodating different data quality and availability scenarios. Such enhancements hold promise for significantly refining the accuracy of GHG emissions estimation, ultimately fostering a more effective reduction of emissions.

While many studies focus on reducing carbon emissions in transport and logistics, they often fail to define carbon performance. This study uses the metric emission intensity to assess carbon performance in terms of emitted grammes CO_2e per tonne-kilometer. According to the European norm EN 16258,

using the CO_2 equivalent is necessary instead of CO_2 only. Therefore, the emission intensity includes carbon dioxide (CO_2), nitrous oxide (N2O), methane (CH4), hydrofluorocarbons (HFC), perfluorocarbons (PFC) and sulfur hexafluoride (SF6).

We are the first to compare the emissions of outsourced transport operations. The research comprehensively compares emission intensities between different LSPs, considering factors such as distance, empty distance, weight, transportation mode, and industry averages. By comparing emission intensities between different LSPs, the research provides valuable insights into the factors influencing environmental performance in the transportation sector. It identifies significant discrepancies in emission intensities between LSPs and highlights potential reasons behind these differences, such as variations in shipment types and operational practices. The findings have practical implications for decision-makers in the logistics industry, as they offer insights into strategies for improving environmental performance and reducing greenhouse gas emissions. For example, identifying factors contributing to lower emission intensities in certain LSPs can inform decision-making processes to enhance sustainability practices and meet regulatory requirements.

The findings highlight the importance of detailed data, such as empty and loaded distance data, in enhancing predictive accuracy for emission intensity predictions in road transportation. This emphasises the significance of incorporating comprehensive datasets to improve the precision of emission estimates and inform decision-making processes. The analysis identifies potential data accuracy issues, such as misclassification errors in transportation modality, which can affect the performance of predictive models, particularly in intermodal transportation. This underscores the critical need for ensuring data accuracy and reliability to minimise uncertainties in model predictions.

From a practical perspective, this research is particularly relevant in light of the Corporate Sustainability Reporting Directive (CSRD) announced by the European Commission in 2022. The CSRD requires large EU and EU-listed companies to report more detailed sustainability reports, including direct and indirect emissions and other environmental factors (EFRAG, 2023). The first companies must follow the new reporting rules for 2024, with reports published at the beginning of 2025. This research, therefore, holds significant relevance for companies that outsource their transportation to logistics service providers, providing them with valuable insights and guidance in meeting the upcoming reporting requirements.

1.3 Thesis outline

The document is organised as follows. Chapter 2 describes a background on GHG estimation methods and predicting emissions with uncertainty, followed by the problem statement and methodology in Chapter 3. Then, Chapter 4 introduces the case study of our anonymous industry partner, including the data description and preparation. We compare and evaluate multiple prediction models and compare LSPs in Chapter 5, providing the numerical experiments. Lastly, Chapter 6 concludes and discusses the methodological and practical results.

2 Background

This section consists of two subsections. The first section is about estimating greenhouse gases in the transportation sector. We discuss some important concepts, guidelines and standards. Besides, we examine how the GLEC Framework works. In the second subsection, we discuss the uncertainty of emission intensity and the sources of uncertainty. We also discuss emission intensity prediction methods and uncertainty quantification methods.

2.1 Greenhouse gas emission estimation

The three most widely used standards and norms for calculating GHG emissions in transportation and distribution are GHG Protocol, ISO 14083:2023 and EN 16258. According to Schmied et al. (2012), the choice of standard is based on the purpose of the calculation. We could use EN 16258 and ISO 14083 to determine the Transport Services Footprint and the GHG Protocol to calculate the Value Chain carbon footprint. Appendix A provides an overview of the most widely used tools and databases.

2.1.1 GHG Protocol

The GHG protocol establishes frameworks and guidelines to measure and calculate GHG emissions. In 2001, the initial edition of "The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard" was published. The Corporate Accounting and Reporting standard evolved into the most widely used standard to measure carbon emissions. Business leaders use this standard to understand, quantify and manage their carbon footprint (WRI & WBCSD, 2004). According to WRI and WBCSD (2004), we can classify GHG emissions into three categories. Scope 1 emissions are the direct GHG emissions that the company controls or owns. For example, emissions from combustion in boilers, vehicles or furnaces that are owned or controlled by the reporting company. Besides, emissions from chemical production in owned or controlled process equipment belong to Scope 1 emissions. Moreover, Scope 2 emissions are the GHG emissions from power generation purchased by the reporting company. Scope 2 emissions occur physically at the electricity generation facility. Moreover, Scope 3 emissions are the most significant source of emissions for companies and present the greatest opportunity to impact GHG reductions and achieve several of GHG-related business objectives (WRI & WBCSD, 2011). Scope 3 emissions refer to the indirect greenhouse gas emissions that occur through a company's activities but are not directly owned or controlled by the company.

2.1.2 EN 16258

EN 16258 was, until recently, the only standard focusing on transport. The standard EN 16258, "Methodology for calculating and declaring energy consumption and greenhouse gas emissions of transport services," offers guidelines to systematically calculate GHG emissions for passenger and freight transport. This standard outlines a methodology, defines system boundaries, addresses allocation, and identifies data sources. In this context, a transport service refers to the conveyance of goods from the sender to any destination (Schmied et al., 2012). The calculation necessitates dividing this transport service into segments where the item travels on a specific vehicle without changing vehicles. We also refer to these segments as "legs". EN 16258 differentiates between three definitions of energy consumption and emissions. Well-to-tank (WTT) emissions encompass all indirect emissions generated by processes ranging from the energy source (the well) through extraction, processing, storage, and delivery phases up to the point of utilisation (the tank). Well-to-tank emissions are classified as Scope 3 emissions. Tank-to-wheel

emissions represent the emissions resulting from fuel combustion used to power Scope 1 activities (the wheel). Tank-to-wheel emissions are designated as Scope 1 emissions. For electricity, hydrogen fuel cells, and biofuels, Tank-to-Wheel is zero since all emissions occur in the Well-to-Tank stages at the point of use. Furthermore, Well-to-Wheel (WTW) emissions combine Well-to-Tank and Tank-to-Wheel, representing emissions across the fuel life cycle. Well-to-wheel encompasses both direct and indirect emissions.

2.1.3 The GLEC Framework

Since the development of the GLEC framework in 2016, the GLEC framework has become increasingly popular. This framework presents the only worldwide harmonised calculation approach that focuses especially on transport, covers all modes of transportation, is fully regionally applicable, and incorporates the entire transport chain (Davydenko et al., 2014). GLEC Framework is in line with the GHG Protocol. The GLEC framework is a standardised methodology that enables companies to consistently and transparently measure and report their logistics emissions, including Scope 1, 2, and 3 emissions. The GLEC Framework harmonised existing methodologies and practices into one framework. Table 1 shows base methodologies for each transportation mode. The base methodologies could provide valuable tools to calculate emissions. Some tools, such as EcoTransIT, are commercial tools that calculate emissions after submitting a list with information. Other tools, such as IATA RP 1678 and CCWG, are guidelines for estimating emissions. These guidelines are already included in the GLEC framework, so there is no need to study these. However, it might be useful to take a look at these guidelines. SmartWay is also a base methodology containing valuable program data for air, inland waterways, road and rail transport modes. However, SmartWay only provides program data on carriers in the US and Canada. This method does not apply to this study since the scope of this research is Europe Middle-East Africa (EMEA). If a company has already implemented a methodology by GLEC, the company can continue using that method. However, small adjustments must be made to align with GLEC. For each transportation mode, the GLEC framework provides an overview of the global impact, the scope of activities included, information on base methods, and tips for emissions accounting.

Transportation mode	**Base methodology**
Air	IATA RP 1678 SmartWay Air Cargo Tool
Inland Waterways	SmartWay Barge Carrier Tool GHG Emission Factors for Inland Waterways Transport IMO Ship Energy Efficiency Operation Index
Sea	CCWG (only for container shipping) IMO Ship Energy Efficiency Operation Index
Road	EN 16258 SmartWay Road Carrier Tool
Rail	EcoTransIT SmartWay Rail Carrier Tool
Logistics Sites	Guidance for Greenhouse Gas Emissions Accounting at Logistics Sites Guidance for Greenhouse Gas Emission Footprinting for Container Terminals

Table 1: Base methodologies GLEC Framework

The GLEC framework identifies three steps. The first step is to set boundaries and goals. The second step is to calculate scope 1 and 2 emissions. For companies that outsource all transport and distribution, scope 1 and 2 emissions are zero. Therefore, we only focus on the third step, estimating Scope 3 emissions.

Input data

To estimate Scope 3 GHG emissions, different levels of data quality could be used as input data, depending on the availability of the data. The type of data influences the accuracy of the results. Therefore, it is essential to collect high-quality data. The GLEC Framework recommends using primary data. However, this is not always possible. Therefore, it is possible to estimate emissions using GLEC with multiple data levels of input data. The GLEC Framework distinguished four data categories (A. Lewis & Greene, 2019). Assumptions made on primary data have higher validity than assumptions made on default data (A. Lewis & Greene, 2019).

There are four possible input data categories. The first category includes primary data. Transport buyers should aim to collect qualitatively good primary data from carriers to calculate their Scope 3 emissions. Primary data can range from exact information, from, for example, fuel receipts or annual spending, to aggregated values that reflect fuel or emission intensity for a year's worth of vehicle movements. Another input data category is program data. Program data is data from the green freight programs to guide carrier selection and identify potential energy, cost and emission-saving strategies, for example, SmartWay carrier performance data. Green freight programs are essential in connecting shippers and carriers around the world. The third category includes modelled data. Models combine shipment data, such as goods types, journey origin and destination, and intermediate handling locations, with information about vehicles and fleets to model fuel use and emissions. Modelled data is used for some of the methodologies in Table 1, such as EcoTransIT. Modelled data is secondary data. Lastly, default data can be used if no other data are available. The data contain industry averages using standard assumptions of vehicle efficiency, load factor and empty running. An example could be the GLEC default emissions factors. Default data is secondary data.

Determine distances

Furthermore, we can determine the distance using the Great Circle Distance (GCD), Shortest Feasible Distance (SFD), or planned distance. Different methods to determine distance are used for different transport modes depending on data availability and quality. We can determine actual based on knowledge of the actual route or odometer readings. Mostly, the actual distance is only known by carriers. Great Circle Distance is the shortest distance between two points measured along the sphere's surface. Great Circle Distance is mostly used for air transportation. The shortest Feasible Distance is the shortest route between two locations, mostly found by route planning software. This method neglects real operating conditions such as vehicle restrictions, weight or height, road types or constructions. The planned distance is the shortest distance found by planning software while considering real operating conditions.

Calculate tonne-kilometers

To calculate Scope 3 emissions, we must calculate tonne-kilometres. A tonne-kilometer represents one tonne of cargo moving for one kilometer. The tonne-kilometer can be used to express the efficiency of the vehicle mode. The GLEC Framework recommends not using attributes other than weight, such as volume or density, for consistency. We could calculate tonne-kilometers by multiplying the weight and the distance. We choose the sum of tonne-kilometres for each transport mode to find the total tonne-kilometres. If accurate tonne-kilometres data are unavailable, we can calculate the tonne-kilometres by multiplying the total weight by the average shipment leg distance or the average total weight by the total shipment leg distance.

Fuel efficiency or emissions intensity factors
There are two methods to determine the total emissions. If there is data on fuel use, the fuel-based method must be used. This method is more accurate than the distance-based method. The following formula represents the fuel-based method:

$$\text{kgCO}_2\text{e emissions} = \sum_1^n \left(\text{ total tkm} \cdot \text{fuel efficiency factor } \left(\frac{\text{kg fuel}}{\text{tonne-km}}\right) \cdot \text{fuel emission factor } \left(\frac{\text{kgCO}_2\text{e}}{\text{kg fuel}}\right)\right)$$

The following formula represents the distance-based method:

$$\text{kgCO}_2\text{e emissions} = \sum_1^n \left(\text{ total tkm} \cdot \text{CO}_2\text{e intensity factor } \left(\frac{\text{kgCO}_2\text{e}}{\text{tonne-km}}\right)\right)$$

The GLEC framework provides a list of fuel efficiency and emission intensity factors. The emission intensity factors are based on industry averages and are also referred to as default factors. These default factors are often higher than the actual emissions, encouraging companies to seek primary data. Default factors are mode-specific and depend on vehicle characteristics such as Gross Vehicle Weight, which is the vehicle weight including the maximum load, the fuel type, long-haul/short-haul and vehicle type. Furthermore, the GLEC framework provides tips for each calculation, such as how distance should be measured for each mode. Additionally, GLEC framework contains many assumptions, so calculations could always be performed despite unavailable data. For example, average load factors given in caseload factors are unknown.

The Smart Freight Center extended the GLEC framework for the chemical industry and established guidelines to estimate GHG emissions Smart Freight Centre and Cefic (2021). These guidelines include sector-specific emission intensity factors and guidance.

2.2 Emission intensity prediction and uncertainty

This subsection provides an overview of methods to predict emission intensity. We discuss regression analysis and different feature selection methods. Additionally, we discuss uncertainty since this is an essential element in estimating GHG emissions. We consider the types and primary sources of uncertainty. Furthermore, we discuss methods to quantify uncertainty.

2.2.1 Prediction methods

Predicting freight transportation emissions could provide insights into the main features influencing emissions, future emissions trends and insight for developing appropriate environmental policies and strategies to mitigate carbon emissions. There are many methods to predict energy-related emissions. According to Ding et al. (2017), the forecasting methodologies of emissions can be generally divided into three categories: non-linear intelligent models, statistical analysis models, and grey prediction models. Non-linear intelligent models such as artificial neural networks (ANN) (Acheampong & Boateng, 2019), support vector machines and fuzzy regression have been developed to predict GHG emissions effectively. However, one fundamental limitation needs to be overcome: accurate prediction results are highly dependent on the amount of training data, which are used to discover potentially predictive relationships in intelligent systems, machine learning, genetic programming and statistics (Ding et al., 2017).

Statistical analysis models, including univariate models, trend analysis, and multivariate regression analysis models (Piecyk & McKinnon, 2010) have been widely utilised for predicting energy-related CO2 emissions. Multiple regression gives a clear understanding of how variables influence the emissions. However, a limitation of causal models as regression is that it depends on the availability and reliability of independent variables over the forecasting period, which requires further efforts in data collection and estimation (Zhou et al., 2006). Besides, variables for modelling carbon emissions can be chaotic, non-stationary, or non-linear. In these cases, the classical statistical and econometric approaches are unsuitable. The grey prediction model is a prediction model that can achieve accurate projections even in sparse samples (Ding et al., 2017). Grey prediction models require historical time series data on emissions.

2.2.2 Multiple linear regression

Regression analysis falls under the category of supervised learning because the algorithm is trained using input and output labels (Koza et al., 1998). Regression is a mathematical technique in machine learning that helps data scientists predict a continuous outcome y_i based on the values of one or more predictor variables x by estimating how one variable influences the other. Linear regression is the most commonly used type of regression analysis due to its easy application in forecasting and predicting. Multiple regression analysis is a statistical technique for analysing the relationship between a dependent variable and multiple predictor variables (Hair et al., 2010). To understand the concept of multiple regression analysis, we first explain the concept of simple linear regression. The equation gives simple linear regression

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i \tag{1}$$

where x_i is the predictor variable and y_i is the response variable of the ith data pair where $(x_i, y_{i0}), i = 1, ..., n$. β_0 and β_1 represent the intercept and the slope of the regression, respectively. ε_i is the error term. The error term ε_i is the difference between the actual value of y_i and the unobservable expected value of y_i. In cases where the dependent variable is linearly dependent on multiple independent variables, the equation that defines the relationship between these variables x and the dependent variable y_i takes on a general form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \epsilon_i \tag{2}$$

or in vector form:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i,$$

The standard error of the estimate measures the variance in the predicted values. We can use the standard error of the estimate to develop confidence intervals around the mean response. It is similar to the standard deviation of a variable around its mean, but instead reflects the expected distribution of predicted values that would occur if we would take multiple samples of the data (Hair et al., 2010).

Assumptions

Multiple assumptions are necessary for multiple regression: linearity, homoscedasticity, linearity, and normality.

Linearity

The assumption of linearity includes that the bivariate relationship between the independent and dependent variables is linear. A residual plot of each independent variable x_i and the dependent variable y is useful to visualise this relationship. Besides, scatterplots could be used to visualise the relationship.

In case there is a non-linear relationship, corrective action is required. Corrective actions can take three forms (Hair et al., 2010): (1) transforming data values (e.g. logarithm, square root) of one or more independent variables to achieve linearity, (2) including the non-linear relationships in the regression model, through for example the creation of polynomial terms or (3) using specialised methods such as non-linear regression specifically designed to accommodate the curvilinear effects of independent variables or more complex non-linear relationships.

Independence of the residuals
Another assumption deals with the carryover effect that occurs when one observation is dependent on the previous one, making the residual not independent (Hair et al., 2010). In such cases, like in time series data, we should identify the potential sequencing variables, for instance, time in a time series problem, and plot the residuals using this variable. For instance, suppose the identification number denotes the order in which we collect our responses. In that case, a residual plot can identify emerging patterns. A common test to check this is the Durbin-Watson test (Durbin & Watson, 1950).

Normally distributed error term
Normality is a crucial assumption in multivariate analysis. It refers to the shape of the data distribution for a single metric variable and its similarity to the normal distribution, which is the benchmark for statistical methods. In case there is a significant deviation from the normal distribution, all statistical tests become invalid because normality is necessary to use F and t statistics (Hair et al., 2010).

The Shapiro-Wilk test is a statistical test of the hypothesis that the distribution of the data as a whole deviates from a comparable normal distribution. We can conduct this test for the individual independent variables and the error term of the variate (Hair et al., 2010). It is crucial to understand that assumptions in multiple regression analysis apply to both dependent and independent variables and the relationship as a whole (Hair et al., 2010).

Homoskedasticity
An assumption of Ordinary Least Squares is homoskedasticity, which means that there is a constant variance of the error term. Residual plots or simple statistical tests are useful to diagnose heteroskedasticity. A popular test for heteroscedasticity is the Breusch-Pagan test. The Breusch-Pagan test has a null hypothesis that homoskedasticity is present. If the p-value is less than some significance level (i.e. $\alpha = 0.05$), we reject the null hypothesis and conclude that heteroscedasticity is present. If heteroskedasticity exists, there are two remedies: (1) change model specification or (2) compute heteroskedasticity-robust standard errors.

Exogeneity
An assumption of OLS holds that there is no endogeneity, which is present if (1) there are omitted variables from the model or (2) when the outcome variable is a predictor of x and not simply a response to x.

Least squares method

Least squares estimation is a parameter estimation method in regression analysis based on minimising the sum of the squares of the residuals. Ordinary least squares is the most common estimator. The most fundamental reason for the widespread use of ordinary least squares regression is that it is easy to calculate and it is part of statistical software packages (Keles, 2018). The Ordinary Least Squares method determines the unknown parameters β by minimising the sum of the squares of the differences between the observed dependent variables $\sum (y_i - \widehat{y}_i)^2$. The minimisation problem of multiple regression is described as follows:

$$\hat{\beta} = arg \min_{(\beta)} \left| |y - \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta}| \right|^2$$

To minimise the squared error terms, we take the derivative of the sum of squared errors equal to 0, resulting in the following formula to determine the parameter vector β:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y \quad \text{Where} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_m \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Where m is the number of independent variables included in the model and n is the number of observations.

Ordinary Least Squares assumes no errors in the independent variables (Keles, 2018). Ordinary least squares are inappropriate when substantial errors exist in describing variables. In these cases, errors-in-variable models should be considered. Total Least Squares is a method that accounts for errors in both dependent and independent variables (Markovsky & Van Huffel, 2007). Total Least Squares is also known as "orthogonal regression" or "errors-in-variables" (Markovsky & Van Huffel, 2007). The ordinary least-squares and the total least-squares methods evaluate the fitting accuracy in distinct ways. The ordinary least-squares method minimises the sum of the squared vertical distances between the data points and the fitting line. In contrast, the total least squares method minimises the sum of the squared orthogonal distances between the data points and the fitting line (Markovsky & Van Huffel, 2007). According to Keles (2018), total least squares perform better than ordinary least squares if there are measurement errors in the independent variable. However, the computation of orthogonal regression is complex and challenging.

Generalised Least Squares (GLS) is an extension of the Ordinary Least squares method that allows efficient estimation of β when heteroscedasticity or correlations are present among the error terms of the model, as long as the form of heteroscedasticity and correlation are known independently of the data. Weighted Least Squares (WLS) is a form of Generalized Least Squares.

2.2.3 Feature selection

Selecting a proper set of variables is essential to build a statistical model. Feature selection methods are developed to select this set of variables. There are three main categories of feature selection methods: filter, wrapper, and embedded methods (Desboulets, 2018). Filter methods are techniques for selecting features based on statistical methods using selection thresholds. They typically rely on statistical measures or heuristics to rank or score features based on their relationship with the dependent variable. Filter methods can be divided into Rank-Based and Subset evaluation-based methods (Khaire &

Dhanalakshmi, 2022). Rank-based methods use univariate statistical techniques to evaluate the rank of each feature. Rank-based methods do not consider the interrelationships between the features and do not work well in identifying redundant features (Khaire & Dhanalakshmi, 2022). Subset Evaluation methods use multivariate statistics to rank feature subsets, considering feature dependency, without a classifier and with higher computational efficiency than the wrapper technique. However, it is slower and less stable than univariate ranking (Khaire & Dhanalakshmi, 2022). Wrapper methods are an iterative feature selection procedure incorporating supervised learning algorithms and rank features based on a subset evaluation technique, considering correlation and dependencies. The technique optimises performance by considering the bias of the prediction algorithm but has high computational expense and a high risk of overfitting. Examples include Recursive Feature Elimination (RFE) and Greedy Forward Selection (GFS). Although the embedded technique shares the same advantages as the wrapper technique, it is considered advantageous in terms of computational complexity compared to the wrapper technique (Khaire & Dhanalakshmi, 2022). Examples of embedded methods are LASSO or ridge regression.

In regression analysis, Desboulets (2018) divides feature selection algorithms into three types of algorithms: test-based, penalty-based and screening-based. Test-based algorithms are based on statistical tests, including normality tests, t-tests, and other similar tests. This method eliminates variables that do not contribute to the model. Examples of this method include stepwise regression and autometrics. The second approach is penalty-based procedures, which impose constraints on parameters to promote sparsity among them. Ridge and LASSO are examples of this category. Penalty-based methods are beneficial when (1) there are high degrees of multi-collinearity or (2) the number of variables exceeds the number of observations in the sample Hair et al. (2010). Finally, there are screening procedures that rank variables by importance. Although not designed for selection, they are helpful for significant dimensional problems where the number of features is higher than the number of observations. This approach considers additive models, meaning variables can be treated independently.

2.2.4 Uncertainty and variability

Uncertainty and variability are concepts that play a crucial role in estimating emissions. Uncertainty represents a lack of knowledge regarding the true value of a variable. It is commonly expressed by a probability density function (PDF) that depicts the possible values and their likelihood (Frey, 2007). It depends on the analyst's level of knowledge, which is influenced by the quality and quantity of relevant data, as well as their understanding of underlying processes and inference methods (Smith et al., 2015). Uncertainty arises from the unknown nature of an undetermined quantity or the true distribution of variability within a population. Conversely, variability refers to the inherent heterogeneity exhibited by a variable across different temporal or spatial dimensions (Morgan & Henrion, 1990; Cullen & Frey, 1999). It arises from differences in design or operational conditions among various emitters and manifests as inter-plant (spatial) or intra-plant (temporal) variability. Variability is an intrinsic attribute of the system or nature itself and is not dependent on the analyst's perception or knowledge.

The fundamental difference between uncertainty and variability lies in their origins and dependencies. Variability is a property of the system itself, arising from natural heterogeneity. In contrast, uncertainty arises from a lack of complete knowledge or information about a variable's true value or distribution (IPCC, 2006). Additionally, both uncertainty and variability are affected by the averaging time. Short-term emissions exhibit higher variability than long-term emissions, reflecting greater fluctuations over

shorter time intervals. Similarly, uncertainty associated with short-term estimates is larger than that associated with long-term estimates, owing to the higher unpredictability inherent in short-term observations (Wang et al., 2021). The interplay between time intervals and their associated variations highlights the differential impacts of uncertainty and variability on environmental analysis.

2.2.5 Types of uncertainty

In emissions estimation, it is crucial to determine the sources of uncertainty and if these are applicable. These sources of uncertainty influence two types: model uncertainty and parameter uncertainty. Model uncertainty concerns the uncertainty that arises from flaws in how the chosen conceptualisations are modelled. These imperfections may occur because of limitations of available data or other sources of structural errors in the model, such as failure to properly consider emissions' sensitivity to ambient conditions or other factors. Modelling can provide a basis for estimating emissions or removals for specific categories and managing data in the entire inventory (IPCC, 2006). Model uncertainty refers to the uncertainty associated with the mathematical equations (i.e. models) used to characterise the relationships between various parameters and emission processes. Multiple factors could influence the model uncertainty. For example, model uncertainty may arise due to an incorrect mathematical model or inappropriate parameters (i.e. inputs) in the model (GHG Protocol, 2023). Sometimes, data or measurements may not be available due to various reasons. One such reason is the lack of completeness, which occurs when the process is not yet recognised or a measurement method is not yet developed. This can result in incomplete conceptualisation, leading to bias and random error depending on the situation.

Parameter uncertainty is the uncertainty associated with the parameters used in the model. There are multiple sources of parameter uncertainty. Firstly, measurement errors cause parameter uncertainty. Measurement error can occur due to random and systematic errors during measuring, recording and transmitting data. It can also result from finite instrument resolution, inexact values of measurement standards and reference materials, and inexact values of constants and other parameters obtained from external sources that are used in the data-reduction algorithm. Other factors contributing to measurement error include approximations and assumptions made during the measurement method and estimation procedure, and variations in repeated observations of the emission or removal or associated variable even under identical conditions (IPCC, 2006). Misclassification errors occur when categorical variables are incorrectly categorised. Additionally, uncertainties arise when measurements attempt to obtain values below the detection limit. This can lead to both bias and random error. Besides, there are situations in which data needed to characterise emissions is unavailable. Due to this lack of data, proxy data is needed to make estimates, leading to uncertainty. Moreover, the lack of representativeness of data causes uncertainty. This type of uncertainty arises when the available data does not completely match the conditions associated with real-world emissions or activities. For instance, emissions data may be available only when a plant operates at full load, while data for start-up or load changes may be missing. As a result, the available data may only partially relate to the desired emission estimate. This lack of representativeness often leads to bias.

Parameter uncertainty can be classified into statistical and systematic uncertainty (GHG Protocol, 2023). Statistical uncertainty is the uncertainty due to random variability of sample data. Increasing the number of independent samples can often reduce variability. It is important to distinguish between variability and uncertainty, as the previous section defines. Systematic uncertainty is associated with systematic biases

in the estimation process. It occurs when data are systematically biased, meaning that the average of the measured or estimated value is always less or greater than the true value. These biases can arise due to various reasons such as non-representative samples used in constructing emissions factors, incomplete identification of relevant source activities or categories, or usage of incorrect or incomplete estimation methods or faulty measurement equipment. As the true value is unknown, such systematic biases cannot be detected through repeated experiments and hence cannot be quantified through statistical analysis. However, it is possible to identify biases and sometimes quantify them through data quality investigations and expert judgments. Parameter uncertainty influences the model uncertainty.

2.2.6 Confidence intervals

Many authors state that using confidence intervals is a good method to quantify uncertainty and to cope with the lack of data quality and availability (Tong et al., 2012; Waldman et al., 2020; IPCC, 2006). The confidence interval is the probability range that contains the true value based on an estimated value. That is, there is a probability of 80% that the true value is between 45 and 55 if the estimated value is 50 and the 80% confidence interval is [45,55]. A smaller, narrow-range model performs better when comparing models using the confidence interval for the same confidence interval. Besides, Harnett & Murphy (1980) define a confidence interval as an estimation of the mean value of the outputs. The 95% confidence interval of true emission can be used to quantify the uncertainty regarding emission estimates due to sampling error in terms of their mean or other statistics (IPCC, 2006). We distinguish parametric and non-parametric methods to determine confidence intervals. Parametric methods rely on a certain distribution. Parametric methods are often preferred over non-parametric methods since the confidence intervals are smaller when we determine confidence intervals based on parametric assumptions (Vickers, 2005). For example, the sum of normal distributions is normally distributed, and the product of lognormal distributions is lognormally distributed (Frey, 2007). While non-parametric tests can be applied to a broader range of data types and require fewer assumptions, parametric tests are generally preferred due to their higher sensitivity when it comes to detecting differences between samples or the effect of independent variables on dependent variables (Abdulazeez, 2014). Non-parametric methods such as bootstrapping should be used if the data distribution is unknown or not symmetric (IPCC, 2006; Vickers, 2005). Non-parametric methods are known for their flexibility in handling different types of data. The essential advantage of the non-parametric bootstrap method is that it does not rely on any specific distributional assumptions. This makes it a useful tool for estimating the sampling distribution of statistics (such as the sample mean), assuming that the sample is representative of the population it is drawn from and that the observations in the sample are independent and identically distributed (Tong et al., 2012). Tong et al. (2012) quantify the uncertainty of emission estimates for national GHG emissions using bootstrap confidence intervals. They conclude that they obtain similar results using either a 95% confidence interval or a bootstrap confidence interval when the sample size exceeds 30. Besides, they show that a larger sample size results in a smaller interval mean and smaller interval standard deviation of the confidence interval.

Suppose the data sample size is large enough. In that case, standard statistical goodness-of-fit tests, combined with expert judgement, can help decide which Probability Density Function to use to describe variability in the data (IPCC, 2006). If data is normally distributed, we can apply the classical confidence interval developed by Neyman (1935):

$$CI = \bar{x} \pm Z^{(\alpha)} \cdot \frac{\sigma}{\sqrt{n}}$$

Where CI is the confidence interval, α is the significance level, and Z is the critical value of the Z-distribution. Besides, $\bar{x}$ is the sample mean, σ is the sample standard deviation, and n is the sample size.

Prediction interval

Another interval to quantify uncertainty is through prediction intervals. The prediction interval provides a range within which we expect a future observation to fall with a certain probability, whereas the confidence interval focuses on past or current events. The prediction interval is wider than a confidence interval and considers data variability and uncertainty associated with individual predictions. Confidence intervals measure the uncertainty in model structure, while prediction intervals concern total variance. With a known probability, the prediction interval uses sample data to predict a new observation, given a set of values for the independent variables (Olive, 2007).

3 Problem statement and methodology

3.1 Problem statement

Companies often cope with a lack of data quality and availability, resulting in difficulties in estimating emissions. Due to different calculation methods and data quality and availability levels, there is a lack of comparability of emissions between LSPs. This lack of data is especially a problem when transportation is outsourced. For companies that outsource their transportation to multiple LSPs, comparing LSPs based on their emissions is interesting to gain insights into their environmental performances. Comparing LSPs could improve decision-making. Since reporting scope 3 emissions is obligatory from 2024 and onwards, the European Union is one step closer to implementing a carbon tax for transportation. The emitters (carriers) must pay the transportation carbon taxes, and the emitters will forward these taxes. Therefore, it is also, from an economic perspective, interesting to choose LSPs with low emissions. Choosing LSPs with lower emissions can ensure compliance with current and future environmental regulations. Therefore, this research aims to develop an approach to compare the emission intensity of LSPs while differing in data quality and availability.

To enable a fair and standardised comparison between different LSPs in terms of emission intensity, we must first predict the emission intensity for each LSP. Many features influence the emission intensity, such as distance, vehicle type, journey type or traffic conditions. However, we often do not have access to all features. There is currently no method to estimate emissions for different feature combinations. For example, the GLEC Framework requires knowledge of a road shipment's vehicle characteristics and size to determine the emission intensity. Companies that outsource transportation do not always know what vehicle the LSP uses. However, we still want to be able to estimate the emissions. Therefore, we develop a model to predict emission intensity for each LSP.

Besides, we must quantify uncertainty for each prediction model to compare different models with different input data quality and availability. In the background section, we discussed that there is model uncertainty and parameter uncertainty. An essential cause of model uncertainty is the lack of data. For example, we do not have data on traffic conditions, but traffic conditions do influence emission intensity. This results in uncertainty. Besides, a model can have parameter uncertainty if we have a model predicting the emission intensity of an LSP with weight as the describing variable and if there are observational errors in the data on weight. For example, the LSP does not include the packaging weight in the shipment weight.

We could use a parametric or non-parametric method to predict emission intensity while quantifying uncertainty while having different data quality and availability. Non-parametric methods, while flexible, often require extensive training data and may struggle to provide precise estimations with small sample sizes (Frey, 2007; Acheampong & Boateng, 2019). Additionally, while non-linear intelligent models offer potential advantages in capturing complex relationships within the data, they may be impractical due to their large amount of training data and lack of interpretability (Acheampong & Boateng, 2019). A parametric method is preferred if the case study data is normally distributed after transformations. In contrast, a non-parametric is preferred if the data are not normally distributed and cannot be transformed to normal distribution. In the case study, parameters are normally distributed after data transformations. Besides, the sample size is relatively small. In this context, choosing multiple linear regression as the modelling approach presents a pragmatic solution. By developing separate regression models for

each subset corresponding to different LSPs and transport modes, we can tailor the modelling process to accommodate variations in data characteristics while providing interpretable and actionable insights. Furthermore, the parametric nature of multiple linear regression allows us to leverage the distributional properties of the data, facilitating more precise estimations even with small sample sizes. Additionally, the transparency and simplicity of the regression framework make it accessible to stakeholders with varying levels of technical expertise, ensuring that the results are easily understandable and actionable in real-world decision-making scenarios.

For this purpose, we develop a model for each subset S_j for $j \in \{1, 2, \ldots, J\}$, where each subset corresponds to one LSP and one transport mode. Consider multiple linear regression models M_j for $j \in \{1, 2, \ldots, J\}$, where each model explains the emission intensity, y, of one modality for one specific LSP as follows:

$$M_j : y = \beta_{0j} + \sum_{i \in S_j} \beta_{ij} \cdot x_i + \varepsilon \tag{3}$$

This model is a classical linear regression model (Poole & O'Farrell, 1971). ε is the random error term of the model M_j. In each M_j, x_i represents a describing variable, and the sets of describing variables S_j differ between the models. Each describing variable represents one parameter that influences the dependent variable y. We want to compare the predicted dependent variable emission intensity y of the different LSPs under similar values of x_i to assess which LSP performs better. For example, we determine which LSP performs better for short-distance shipments below 200 kilometres with Full Truckload (FTL).

However, we must distinguish between predictions of the mean response and predictions of future observations. Therefore, we must decide if we are predicting the emission intensity for a particular new shipment with characteristics x_i or the mean response for a shipment for a given x_i. This depends on the billing of the carbon emissions taxes of LSPs to the company that outsources emissions. If the LSP bills emission taxes based on the average of each lane, we are interested in the confidence interval of the mean response. We are interested in the prediction interval of whether the LSP bills emission taxes based on every single shipment. Therefore, we calculate the confidence interval or prediction interval to quantify the uncertainty of emission intensity, depending on carbon tax billing characteristics per LSP.

To compare emission intensities of LSPs under similar circumstances, we determine the prediction and confidence interval. To adequately compare LSPs, we want to obtain narrow intervals. If there are differences between logistics service providers, we can only find them if confidence and prediction intervals are small. Therefore, reducing the confidence and prediction interval width as much as possible is essential. Consequently, we must determine the final prediction model M_j by finding the corresponding set of features S_j so that S_j minimises the standard error of the estimate. Then, we calculate the mean response variable y for different input variables of each model M_j. We can compare the prediction intervals and confidence intervals of the mean response y between models M_j under similar values of input variables.

Figure 1 shows a typical intermodal and road shipment in the chemical industry. The total emissions CO_2e include the emissions caused by empty runs and loaded tonne-kilometers. To calculate the emission intensity of a shipment, divide the total CO2e (black arrow) by the loaded tonne-kilometres (blue arrow). Thus, the emission intensity represents the emitted grams of CO2e per loaded tonne-kilometer.

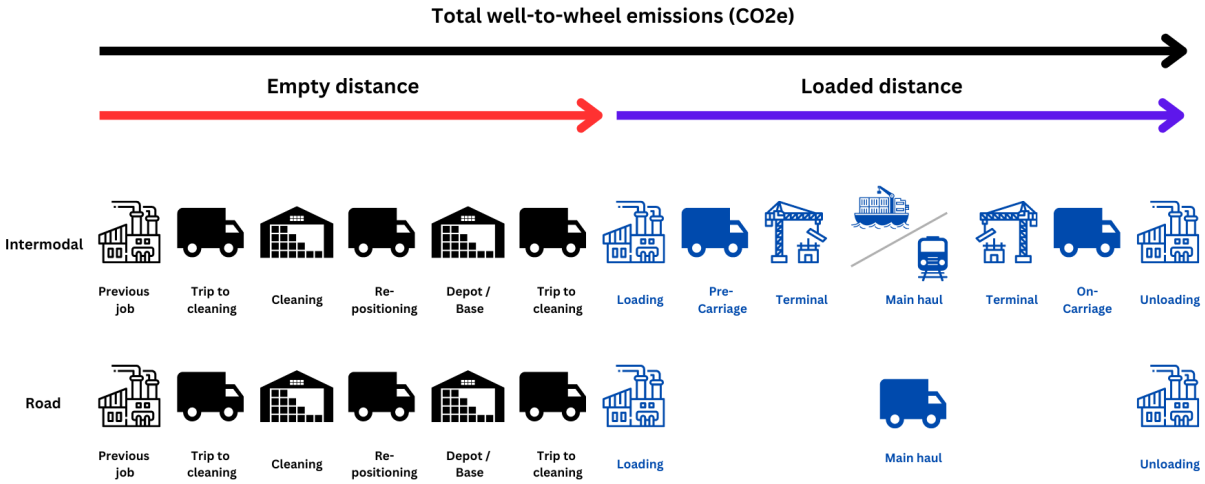

Figure 1: Structure of the shipment (copied from Smart Freight Centre & Cefic (2021))

3.2 Methodology

This research aims to determine prediction and confidence intervals of the mean response to compare emission intensities of LSPs, differing in data quality and availability. This methodology consists of two phases: (1) determining the regression model for each LSP per modality having a set of features that accurately predicts emission intensity and (2) comparing emission intensities of LSPs and the corresponding confidence and prediction intervals to assess which LSP performs better.

3.2.1 Least Squares method

Several least squares estimation methods exist to determine a regression model: total least squares, generalised least squares, weighted least squares and ordinary least squares. The required method highly depends on the available data.

Emission intensity is the outcome of a model from logistics service providers that is externally validated, meaning that the dependent variable is estimated instead of measured. In regression analysis, the residual can be divided into two components when the dependent variable is estimated. The first component is the sampling error, which represents the difference between the true value of the dependent variable and its estimated value. The second component is the random shock that would have occurred even if the dependent variable had been directly observed instead of estimated. The first component is heteroscedastic if the sampling variance differs across observations. However, the second component can still be homoscedastic (J. Lewis & Linzer, 2005). If the sampling error variance is small or the variance of second component residuals is large relative to the sampling error, then ordinary least squares perform quite well. We do not know which proportion of the emission intensity is due to the first or second component. Therefore, we cannot use Feasible Generalized Least Squares (FGLS), recommended for Estimated Dependent Variable (EDV) models (J. Lewis & Linzer, 2005). Thus, the only methods to estimate parameters are Ordinary Least Squares or Weighted Least Squares while having errors in the dependent variable is possible (J. Lewis & Linzer, 2005). Thus, we can use OLS with robust standard error estimates, which may result in inefficient parameter estimates if the sampling error variance is large but provides accurate estimates of the parameter uncertainty (J. Lewis & Linzer, 2005). Alternatively, we can use Weighted Least Squares to obtain more efficient parameter estimates, but there is a risk of

getting inaccurate standard errors. Inaccurate standard errors influence the validity of the confidence and prediction intervals, which would be a huge problem. To maintain accurate parameter estimates and standard error estimates, we must obtain data on emission intensities with a sampling error that is relatively small or with equal variances to ensure accurate estimations and valid confidence and prediction intervals. For this purpose, we use ordinary least squares.

Although Total Least Squares (TLS) regression might be considered a potential alternative due to its ability to handle errors in independent and dependent variables, it is not feasible with our case study data (Markovsky & Van Huffel, 2007). Our dataset lacks the information to accurately estimate the proportion of error variance attributed to both variables, making it challenging to implement TLS regression effectively. Additionally, orthogonal regression assumes that the errors in the independent and dependent variables have a ratio of 1 due to using a similar measurement method. This is not necessarily the case as different techniques are used to, for example, measure the distance of a shipment or the emission intensity of a shipment. Therefore, we maintain consistency by utilising OLS regression throughout our methodology, ensuring a coherent approach to modelling emission intensity across different LSPs and modalities. Additionally, ordinary least squares estimates are more straightforward to interpret and more accessible to perform as many statistical software packages include ordinary least squares compared to total least squares.

3.2.2 Data requirements

Some data requirements must hold to calculate the confidence interval and prediction interval based on parametric assumptions. To use ordinary least squares, the variance of the sampling error residuals of the emission intensity must be small relative to the random error variance.

The sample must represent the entire population. In this case, the population for which the regression model must predict is one LSP and one modality. For example, we obtain three regression models to compare three LSPs for road transportation. The regression model predicts the emission intensity of one LSP of one modality. Additionally, each observation must be independent of other observations. The independent and dependent variables must be normally distributed after transformations.

3.2.3 Feature selection algorithm

To derive regression models for each Logistics Service Provider (LSP) and modality, we employ a feature selection algorithm to identify the optimal set of predictors for predicting emission intensity. We adopt exhaustive feature selection, considering all possible combinations of predictors. While exhaustive search offers thorough exploration, it may become computationally intensive for datasets with numerous features (Hair et al., 2010). In the case of scope 3 emissions, if we have data on all features influencing emission intensity according to A. Lewis and Greene (2019), we would have 15 features, resulting in $2^{15} = 32,768$ combinations. Thus, computational feasibility is achievable using programming languages like R or Python. Additionally, as our case study does not exhibit high multi-collinearity or exceed the variables-to-observations ratio, test-based methods such as stepwise regression or best subset selection are more appropriate than penalty-based approaches. Thus, we develop an algorithm that performs a linear regression analysis for M_j for all possible combinations of parameters in subset S_j, while it excludes models if assumptions are violated.

We determine a vector with all describing variables x_i in S_j. Besides, we determine an identity matrix representing all possible combinations of describing variables x_i in S_j. We determine the regression model for each combination and exclude models that violate assumptions. We return the model with the smallest standard error of the estimate since the standard error directly influences the confidence and prediction interval. The assumptions of Ordinary Least Squares must hold. Therefore, the error terms of the variate must be normally distributed. Therefore, the algorithm excludes models without normally distributed error terms through the Shapiro-Wilk test. As the Shapiro-Wilk test is preferred for sample sizes ($<$50), compared to the Kolmogorov-Smirnov test, we exclude all models with a p-value of the Shapiro-Wilk test smaller than 0.05. If the Shapiro-Wilk test is non-significant ($p>0.05$), we conclude that the distribution of the sample is not significantly different from a normal distribution.

The set of describing variables in the model must have a linear relationship with the dependent variable. Besides, we need homoskedasticity to obtain unbiased confidence intervals. Therefore, we exclude models when heteroskedasticity is present. When the model contains heteroskedasticity, the estimated standard error is wrong, and since the confidence interval relies on the standard error, the confidence intervals are not calculated correctly. Therefore, we conduct the Breusch-Pagan test and exclude all models if $p<0.05$. Furthermore, we exclude models where the p-value of the F-statistics is $p>0.05$. Insignificant F statistics indicate that there is no relation between any of the independent variables and the dependent variable.

Multi-collinearity could lead to inflated standard errors and thus affect the confidence interval. Insignificant coefficients do not necessarily invalidate the entire model. However, insignificant coefficients increase the width of the confidence interval of the mean response. Models with high multi-collinearity result in very complex models, which are difficult to interpret. Models are excluded if there is multi-collinearity. (Hair et al., 2010) state that the model should be reconsidered if the Variance Inflation Factor (VIF) is larger than four, and if the VIF is larger than 10, the variable should be removed. Therefore, we determine which threshold of the VIF is most appropriate. We exclude models if the VIF exceeds this threshold. If the model only contains one describing variable, we cannot calculate the Variance Inflation Factor and the model is included since multi-collinearity cannot exist. Additionally, we analyse the influence of the significance level of coefficients to determine the optimal threshold. Therefore, we perform a sensitivity analysis for a p-value threshold of the independent variables of 1, 0.05, 0.01 and 0.001.

If there is no valid model, there might be no feature that explains the dependent variable, or we have insufficient observations. All models may be excluded due to heteroscedasticity. If this is the case, weighted least squares might be more appropriate due to heteroscedasticity in the sampling errors. Since one of the requirements is that the data of the dependent variable follows a normal distribution, we can calculate the confidence interval using the classical confidence interval and back-transform these confidence intervals. If no describing variables affect the response variable, the confidence interval of the mean response reflects the variability of the dependent variable itself.

3.2.4 Parameter uncertainty

After determining the best model for each subset, we need to verify the data of the input parameter. Therefore, we determine if there are measurement errors or misclassification errors. If there is also uncertainty in one of the model's input parameters, we should include this in the model. This uncertainty

Input : List of all predictors, response variable
Output: Best model adhering to assumptions minimising Root Mean Squared Error, list of valid models adhering to assumptions

```
Initialise an empty list for valid models;
Initialize variables for best model metrics;
Initialize thresholds for assumption tests;
Generate a matrix of all possible combinations of predictors with binary indicators
for each combination in all predictor combinations do
    selected predictors ← extract predictors from the combination;
    if selected predictors not empty then
        filtered data ← filter data based on selected predictors;
        model ← fit linear regression on filtered data;
        Calculate VIF ;
        Perform Breusch-Pagan test;
        Perform Shapiro-Wilk test
        if Shapiro-Wilk test and Breusch Pagan test > 0.05 and all coefficients p-values <
          threshold and VIF < threshold then
            Append model to the list of valid models;
            Calculate RMSE of the model;
            if RMSE of model < RMSE best model then
                Best model = model;
                RMSE best model = RMSE;
            end
        end
    end
end
```

Algorithm 1: Linear Regression Model Feature Selection

consists of both statistical and systematic uncertainty. We test for errors in each describing variable in the final model. We can validate the parameter input data when both datasets contain a similar parameter. Therefore, we compare the data from two sources and conduct a t-test to determine whether the differences between the measurements are significant. If differences are significant, we use the linear errors-in-variable model:

$$\begin{cases} y_t = \alpha + \beta x_t^* + \varepsilon_t \\ x_t = x_t^* + \eta_t \end{cases}$$

If there are measurement errors in the independent variables, total least squares could obtain more accurate estimations than ordinary least squares. However, the intercept and parameter slopes are more complex to interpret for total least squares. Additionally, the ratio of variances of the dependent and independent variables is assumed to be known. Thus, total least squares can only be applied if validation data is available. If measurement errors are present in a categorical describing variable, we call these misclassification errors. Misclassification errors are used for dummy regressors. We can only include misclassification errors if $\alpha + \beta < 1$, where α and β represent the probability of type I and II errors.

When we have data on an independent variable of a model in only one dataset, we cannot determine if there are measurement errors. We can verify if the values are in line with the industry averages from the GLEC framework.

3.2.5 Confidence and prediction intervals

In cases where the carbon tax will be forwarded per lane, we are interested in the confidence interval of the mean response. In cases where the carbon tax will be forwarded per individual shipment, we are interested in the prediction interval. As there are no carbon taxes yet for freight transportation, we do not know how LSPs will forward the taxes. Therefore, we determine both confidence and prediction intervals for the emission intensity.

IPCC (2006) guidelines suggest using the 95% confidence interval to quantify uncertainty in estimating GHG emissions. Therefore, we take $\alpha = 0.05$. We want to obtain confidence intervals for specific values of describing variables. We must determine confidence intervals if future carbon taxes must be paid per shipment lane. Therefore, we determine the confidence interval of the mean response as follows:

$$CI = \hat{y}_h \pm t_{(\alpha/2)} \cdot s_e \cdot \sqrt{\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)} \tag{4}$$

Where n is the number of observations, and x_h is the value for determining the confidence interval of the mean response. This formula only holds when error terms are normally distributed. Besides, the describing and response variables must be normally distributed. If errors in the regression are normally distributed, β_1 is normally distributed with mean β_1 and variance $\sigma^2 / \sum (x_i - \bar{x})$, where the σ^2 is the variance of the error terms. Besides, the intercept is normally distributed with mean β_0 and variance σ^2/n. Additionally, the sum of squared residuals Q is distributed proportionally to χ^2 with $n-2$ degrees of freedom and independently from $\hat{\beta}$.

The confidence intervals for multiple regression are calculated as follows:

$$\hat{y}\left(X_h\right) \pm t_{(1-\alpha/2,n-p)} \cdot s_e \sqrt{X_h^\top \left(X^\top X\right)^{-1} X_h}$$

This is the confidence interval of the mean response μ_Y when the independent variable values are $\mathbf{X}_h = \left(1, X_{h,1}, X_{h,2}, \ldots, X_{h,k}\right)^{\mathrm{T}}$. $t_{(1-\alpha/2,n-k-1)}$ is the t-multiplier with $n-k-1$ degrees of freedom, where k is the number of describing variables. s_e is the standard error of the regression, the standard error of the fit, or the root mean square error (RMSE). s_e is the square root of the Mean Square Error (MSE). MSE is the sum of squared errors divided by the degrees of freedom of the error:

$$MSE = s_e^2 = \frac{SSE}{df_e} = \frac{\sum \left(y_i - \widehat{y}_i\right)^2}{n-k-1}$$

Multiple factors affect the width of the confidence interval of the mean response. Firstly, as the mean square error (MSE) decreases, the width of the interval decreases. Besides, when we decrease the confidence level, the t-multiplier decreases, and hence, the width of the interval decreases. Moreover, as we increase the sample size n, the width of the interval decreases. Moreover, the more spread out the predictor values, the larger $\sum (x_i - \bar{x})^2$, hence the narrower the interval. Lastly, the closer x_h is to the average of the sample's predictor values $\hat{x}$, the smaller the quantity $(x_h - \bar{x})^2$, and hence the narrower the confidence interval.

Additionally, in cases where the carbon taxes must be paid for each shipment, we must determine the

95% prediction interval as follows:

$$PI = \hat{y}_h \pm t_{(\alpha/2)} \cdot s_e \cdot \sqrt{\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)} \tag{5}$$

The prediction interval is calculated similarly to the confidence interval of the mean response but has an additional MSE term. This means a confidence interval will always be narrower than the corresponding prediction interval at x_h. By calculating the interval around the sample's mean of the predictor values and increasing the sample size, the standard error of the confidence interval can approach zero. However, the prediction interval includes an additional mean squared error (MSE) term, which means that its standard error cannot approach zero, no matter how large the sample size.

3.2.6 Validation

We validate the algorithm and final model using different techniques. There are two types of model validation: in-sample and out-of-sample model validation. In-sample validation is the goodness of fit, which is how well the model fits the data it has been trained on. Out-of-sample validation refers to how well the model works on new data. To assess in-sample accuracy, we assess the goodness-of-fit

Algorithm validation
First, we validate the algorithm. The algorithm must be generalisable for input data quality and availability levels. Therefore, we test the algorithm using different subsets. We try the algorithm using data from our industry partner to identify if the algorithm is applicable if data is limited. Therefore, we use multiple sets of input variables S_i and compare the obtained model.

Additionally, we perform a sensitivity analysis for different thresholds for excluding models to enhance the algorithm. We use Variance Inflation Factor thresholds of 4 and 10 and thresholds for coefficient significance levels of 0.001, 0.01, 0.05 and 1. We determine which thresholds obtain the best models regarding in-sample and out-of-sample accuracy.

Furthermore, we compare the model obtained using the algorithm with a model obtained through forward stepwise regression. In the first step, we select the highest bivariate correlation, using a correlation matrix (Hair et al., 2010). The next step is (1) to check and delete any variables in the equation falling below the significance threshold and then (2) add the variable with the highest statistically significant partial correlation. We repeat this step until there are no more variables that significantly contribute. Lastly, we check if the assumptions are met.

Model fit
For the in-sample validation, we use similar data to train and validate the model. We use metrics such as Mean Squared Error, R-squared and adjusted R-squared to assess the goodness of fit. Additionally, we determine the in-sample confidence interval. We use the R-squared to compare models with the same dependent variable. It has a value between 0 and 1. The R-squared depends on the variance or error term of the model. The R-squared always increases if extra variables are added to the model. A R-squared close to 1 means the regression model explains the explanatory variables well. The adjusted R-squared does not increase if we add independent variables. To compare models with different numbers

of independent variables, we use the adjusted R-squared.

$$AdjustedR^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1} \tag{6}$$

Where p is the number of independent variables, and n is the number of observations in the sample.

K-fold cross validation

We also test for out-of-sample accuracy to assess how the results of the final regression model will generalise to independent data from a similar LSP. We test for overfitting since overfitting may affect the reliability of the confidence and prediction intervals. Therefore, we use k-fold cross-validation. In K-fold cross-validation, a dataset is repeatedly split into several training and test data. We use Leave-One-Out Cross-Validation (LOOCV) since we have very small datasets for the case study. This is the most extreme case of k-fold cross-validation.

In this method, each observation is individually assigned to the test set, i.e. $k = n$ and $p = 1/n$ (Molinaro, Simon, & Pfeiffer, 2005) where p is the proportion, n is the sample size. The distribution of Sn places mass $1/n$ on the n binary vectors, which assign each of the n observations to the learning and test sets. LOOCV and the corresponding $p = 1/n$ represent the best example of a bias-variance trade-off (Molinaro et al., 2005). We evaluate the test observation using the MSE. We can validate the model if the validation error is slightly higher than the training error. Moreover, we determine the out-of-sample confidence and prediction intervals. Additionally, we determine the out-of-sample 95% confidence and prediction intervals. We determine the percentage of left-out observations within the prediction interval. Additionally, we compare different models based on the out-of-sample confidence and prediction interval with.

Comparison with industry averages

According to Snee (1977), using theory can enable us to understand whether the regression model makes sense. Therefore, we compare the emission intensities with confidence intervals with the emission intensity factors developed by the GLEC Framework. We determine if the industry averages are within the confidence and prediction intervals of the emission intensities of LSPs.

3.2.7 Comparison of LSPs

We want to compare emission intensities of multiple LSPs. We use Equation (3) to compare different logistics service providers. Consider two companies with models M_1 and M_2 and sets of describing variables S_1 and S_2. We calculate confidence intervals for the mean response and the prediction intervals for similar values of X_h. We determine the mean response for all observations. To determine the individual effect of an independent variable, we set other describing variables in the model equal to the average value of the describing variable. For continuous describing variables, we determine the confidence and prediction interval of the emission intensity for each observation X_h. We visualize these effects.

We use the confidence interval to determine whether there is a significant difference between the companies. When the confidence intervals do not overlap, the difference between the groups is statistically significant. When there is overlap, there might be a significant difference. We conduct t-tests to determine if the difference is significant.

4 Case study

We conduct this case study for the logistics department of our anonymous industry partners in Europe, the Middle East, and Africa (EMEA) business. Our industry partner is an American multinational manufacturing company that outsources all transportation and distribution activities. They must report their scope 3 emissions from 2024 onwards and thus estimate them. The company manufactures chemicals, fibres, and plastics. Our industry partner wants to know if it is also possible to determine the emission intensity with the current data quality and availability. Therefore, we use their data as an input subset of the exhaustive feature selection algorithm. Furthermore, we compare two LSPs that perform our industry partner's transportation and operate mainly in road transportation. Our industry partner has the opportunity to change to a biofuel per lane. Therefore, comparing the two LSPs and the lanes is very useful. Besides, as emitters must pay carbon taxes in the future, comparing the emission intensities of LSPs is also economically interesting.

The case study includes shipments transported within or from Europe and Middle-East Africa. We have data from shipments in 2022 from two logistics service providers that we want to compare. We have data on the emission intensity in grams per loaded tonne-kilometer from both logistics service providers. We must assume that the sampling error variance of emission intensities is small and, thus, that the emission intensities have minimal measurement errors to ensure the data requirements to use ordinary least squares. The LSPs determine the emissions using a model. An external party validates their model, and thus, the model properly predicts the emission intensity. This suggests a level of confidence in the accuracy of their estimations. However, there will always be an error as emission intensities are estimated and not exactly measured, but we must make this assumption by using ordinary least squares. Assuming accurate emission intensity values allows a better interpretation of the regression model's coefficients and predictions. If there were measurement errors in the dependent variable, it could complicate the interpretation of the model's results and undermine the reliability of any conclusions drawn.

Multiple parameters affect the emission intensity factor. For the entire model to accurately predict the emission intensity of road transportation, we need data per shipment leg: the shipment weight, fuel type, empty running, load factor, vehicle year, vehicle weight class, vehicle volume, cargo type (mail and parcel, bulk, containers, pallets, mass-limited cargo and volume-limited cargo), condition (ambient or temperature controlled), journey type (point-to-point (long haul) or multiple collection and delivery), contract type (shared or dedicated), topography, road type, long-distance vs short haul and traffic conditions (A. Lewis & Greene, 2019). If we have high-quality data on all these describing variables, the confidence interval width would be minimal and only contain statistical uncertainty. Not knowing parameters influences the model uncertainty. When certain factors are not included in the model, while they influence the emission factor omitted, the uncertainty increases, and the confidence and prediction interval widens.

According to Smart Freight Centre and Cefic (2021), we can distinguish packed goods shipments from bulk goods shipments for road transportation. Bulk goods shipments can be categorised into dedicated and spot transportation services. The equipment is dedicated to a specific product and company for a dedicated transport service. This is more common in the chemical industry because of the specialist nature of the equipment, cargo and cleaning requirements. Dedicated transportation often leads to a higher level of empty running. Empty running for dedicated transportation is about 50% (Smart Freight Centre & Cefic, 2021). Packed goods shipments can be categorised into full truckload (FTL) and less-

than-truckload (LTL). When a company has enough product to fill a vehicle, either by volume or weight, it is called a full truckload. On the other hand, in less than truckload (LTL), the company has one or two consignments that are individually not large enough to fill the vehicle. There are two types of LTL: partial load and groupage. Transportation of a partial load involves a single LTL consignment that is not large enough to fill a vehicle by weight or other dimensions and is transported alone from a single point of origin to a single destination. Groupage refers to consolidating multiple less-than-truckload (LTL) shipments from different chemical companies and various origins by a logistics service provider. This consolidation allows for a higher load factor during main haul transportation than individual shipments. The consolidated shipments may be delivered to one or multiple end destinations. FTL and partial load shipments have an average empty running of 22%, while groupage has an average empty running of 17% of the distance.

LSP 1 is a transportation company specialising in handling liquid goods in four industries: chemicals, gases, petroleum products, and food. LSP 1 ships by road, rail, and by sea. They use EcoTransIT to calculate their GHG emissions. Extra 81.5 kg and 50.0 kg CO_2e are added to the Well-to-Wheel emissions for shipments where cleaning and heating are needed. They use diesel fuel for the shipments of our industry partner. The data from LSP 1 contains all 943 shipments conducted in 2022, where 44% intermodal and 56% road. We only have data per shipment and not per shipment leg. For each shipment, we have data on the total Well-to-Wheel emissions, the starting and ending date, shipment weight (excluding vehicle weight), the actual distance including empty miles, contract type (shared or dedicated), condition (ambient or temperature controlled) and if cleaning is required. Shipment numbers are provided; thus, we can match the shipment numbers with the data from our industry partner. Therefore, we also have data on the planned and great circle distance. The shipments are bulk goods. Appendix B shows a part of the gathered data from LSP 1. Some values are changed for reasons of confidentiality. We also obtain data per lane per contract type on the average loaded distance and the average emission intensity in grams per tonne-kilometer. As we only have data on emission intensity per lane, we must aggregate the data per lane. Besides, the 943 shipments are not independent observations since many shipments are conducted from a similar origin-destination lane, having similar characteristics. Besides, we only have data for LSP 2 per origin-destination lane per shipment profile. Therefore, we aggregate the data by taking the mean value per origin-destination lane per contract type (Dedicated or spot). We can calculate the average empty distance since we have the average loaded and total distance. We do not have data on the traffic conditions, road type, vehicle year, vehicle weight class, engine class, journey type, and cargo type. LSP 1 uses about 41% tank containers, 3% road tankers, 54% Intermediate Bulk Containers and 2% trucks for all customers. However, we do not know the container and vehicle type per shipment. Besides, the data on the emission intensity follow a bimodal distribution; therefore, we split the dataset into two datasets: one for intermodal and one for road transportation. We have 23 lanes for intermodal and 26 lanes for road transportation. We validate the data on modality from LSP 1 with the data from our industry partner. However, the data from our industry partner states that 9/23 intermodal lanes are road lanes. The data from our industry partner states that 1/23 of road lanes are intermodal. Since the LSP transports the shipments, we assume that the data from LSP 1 is valid and that the data from our industry partner contains misclassification errors. Our industry partner misses 31/943 shipments (3.28%) that are present in the data from LSP 1.

We do not normalise the data because this does not improve or only slightly improve the model's performance. We perform min-max scaling, but this does not improve the performance of the model in-sample

and out-of-sample validity. Besides, it increases the complexity of interpretation. Additionally, since we have small datasets, normalisation can amplify the impact of outliers or small variations.

LSP 2 is a transportation company mainly using trucks but also operates using rail and inland waterway transport or combined. LSP 2 transports chemicals, healthcare and infrastructure goods. This company primarily operates using their vehicles instead of hired vehicles. They calculate emissions using the GLEC Framework with the tool EcoTransIT. In total, they performed 779 shipments in 2022. They emitted 396.79 $CO2e$ in 2022 for our industry partner. Furthermore, LSP 2 only uses Euro 6 trucks with fuel type Diesel 7 for our industry partner's shipments, which means that the vehicle limits are according to the European Emission Standard. LSP 2 uses the fuel-based method for shipments that are not outsourced. Approximately 40% of the shipments are outsourced to another carrier. Otherwise, the distance-based method is used. LSP 2 allocates their emissions per Business Unit (for example, chemical or healthcare). The average fuel use per tonne-km per business unit allocates the GHG emission. They aim to allocate the emissions per vehicle type at the beginning of 2024. If an order is Less Than Truckload, the emissions are allocated based on the pallet amount. We have data per lane per shipment leg. For each leg, we have data on the emission intensity (Well-to-Wheel), modality, shipment weight (excluding vehicle weight), tonne-kilometers, and shipment profile (Full Truckload, Less-then-Truckload or groupage). Appendix B gives an example of a piece of data. Besides, we know that the capacity of most trucks is 40 tonnes or 33 pallets. Each vehicle is replaced after four years. These trucks are articulated trucks. LSP 2 uses the information on the weight of the shipments of our industry partner. We do not have data on the traffic conditions, road type, vehicle year, vehicle weight class, engine class, empty running (%), journey type, cargo type, contract type (shared or dedicated), and condition (ambient or temperature controlled). Data on LSP 2 contain 779 shipments in 2022. We have 120 road lanes, of which 51 are Less-than-Truckload, 17 are Groupage, and 56 are Full Truckload. For rail, we have 20 lanes. When we look at the data for rail transport, we see that the distance is the same for 16 out of 20 lanes. Each lane is displayed as the origin-destination lane for the whole shipment, while the emission intensity and distance are only for the rail component. In total, 44 out of 48 shipments are transported using this lane. The data is thus not random. This rail lane is between the Netherlands/Belgium and Romania/Hungary. Rail and Inland waterways contain three lanes each. We cannot validate the data from LSP 2 with the data from our industry partner since no shipment number is given. We do not know the starting point and end point of one leg. We only know the starting point and end point of the whole shipment. Furthermore, the data from our industry partner is not complete. Only 549/749 shipments are allocated to LSP 2.

For the regression model, we need data with a normal distribution. Besides, we need a linear relationship between the independent variable and emission intensity. According to IPCC (2006), a lognormal probability density function may be appropriate when uncertainties are large for a non-negative parameter and known to be positively skewed. If many uncertain variables are multiplied, the product asymptotically approaches lognormality. Because concentrations result from mixing processes, which are, in turn, multiplicative, concentration data tend to be distributed similarly to a lognormal. Therefore, we expect the emission intensity to have a lognormal distribution since many factors cause the uncertainty of emission intensity. The emission intensity is the output of a model, having many antecedents of uncertainty.

We expect a positive relation between empty kilometres and emission intensity and a negative relation between loaded kilometres and emission intensity. We split the total distance into these two, as we

want to determine the separate effect. We expect the relationship between loaded distance and emission intensity to be positive with diminishing returns. For this relationship, a negative reciprocal, logarithmic or square root transformation of distance is appropriate for both independent and dependent variables (Hair et al., 2010). Typically, as the distance increases, vehicles have the opportunity to operate at more constant speeds, which can lead to better fuel efficiency. Long-haul highway transportation may be more fuel-efficient than shorter, stop-and-go urban trips. Moreover, traffic congestion, frequent stops, and acceleration/deceleration patterns in urban areas can increase fuel consumption and emission intensity. Longer distances with less congestion may allow for more consistent and fuel-efficient driving, contributing to lower emission intensity. We expect that the emission intensity increases if the empty distance increases since the emission intensity is calculated over the total CO2e, including the empty distance divided by the loaded tonne-kilometers. When the empty distance increases, the total CO2e increases, while the tonne-kilometers remain the same. We expect that weight negatively influences the emission intensity. As the weight increases, the vehicle will be fuller, resulting in a lower emission intensity in grams per loaded tonne-kilometer.

We calculate each parameter's mean, standard deviation, skewness and kurtosis to determine the appropriate transformations. For each continuous parameter, we conduct the Shapiro-Wilk test to determine normality. Based on this, we try different transformations to obtain normality while maintaining linearity. We visualise the distributions in a probability density plot. We try logarithmic, square root, squared, cubic and reciprocal transformations to obtain normal parameters. If the distribution is between square root and logarithmic, we also try power transformations with the power of 0.1 and 0.4. For categorical variables, we include dummy variables. The appendix C shows the descriptive statistics of each variable and the corresponding p-value of the Shapiro-Wilk test. Table 2 shows the final transformations for each parameter. The distribution of great circle distance and planned distance of LSP 1 are still different from a normal distribution after transforming the data for road transportation ($p<0.05$). However, we do not exclude this parameter since ($p>0.01$).

We did not find an appropriate transformation to transform weight to a normal distribution for road transportation for both LSPs. For LSP 1, the parameter weight for road transport has a multi-modal distribution with three peaks: one around 21, one around 24 and one around 29 tonnes. The 4x2 tractor with a 3-axle semi-trailer is the most commonly used long-distance vehicle. Today, its kerb weight is around 14,900kg, which, at a 40-tonne gross weight allowance, leaves 25,100kg for potential payload. The maximum allowed weight in Europe differs per country. In most countries, this weight allowance is 40 tonnes. However, in some countries, this is higher. The small peak of around 29 tonnes contains two lanes. These are within the Netherlands, where the maximum weight allowance is 50 tonnes. When looking at the emission factors in the CEFIC, the tank trucks are typically loaded with 21 tonnes, a silo hopper with 26 and a tank container with 24 tonnes. The peak of around 24 tonnes could be tank containers, and the peak of around 21 could be tank trucks. Therefore, we transform the weight for road transportation into a categorical parameter. If the weight of the shipment is 22.5 tonnes or smaller, the shipment is categorised as a tank truck, and if the shipment is 22.5 or higher, the shipment is categorised as a tank container.

For LSP 2, there is an extremely high correlation between weight and Full Truckload, which aligns with our expectations since the shipment weight of Full Truckload shipments is much higher than Groupage and Less than truckload. Therefore, if we include dummies for the two peaks in this probability density

plot, we would obtain dummies similar to those for FTL. Therefore, we exclude weight.

LSP	Modality	Parameter	Transformation
LSP 1	Road	Great Circle Distance	Power of 0.2
LSP 1	Road	Planned Distance	Power of 0.2
LSP 1	Road	Empty Distance	Power of 0.3
LSP 1	Road	Loaded Distance	Logarithmic
LSP 1	Road	Total Distance	Power of 0.2
LSP 1	Road	Weight	dummy
LSP 1	Road	Emission intensity	logarithmic
LSP 1	Intermodal	Great Circle Distance	Power of 0.2
LSP 1	Intermodal	Planned Distance	Power of 0.2
LSP 1	Intermodal	Empty Distance	Power of 0.3
LSP 1	Intermodal	Loaded Distance	Logarithmic
LSP 1	Intermodal	Total Distance	Power of 0.2
LSP 1	Intermodal	Weight	None
LSP 1	Intermodal	Emission intensity	logarithmic
LSP 2	Road	Distance	square root
LSP 2	Road	Weight	Exclude
LSP 2	Road	Tonne-kilometer	Power of 0.2
LSP 2	Road	Emission intensity	logarithmic

Table 2: Final transformations

5 Numerical experiments

This study aims to compare emissions of outsourced transport operations and determine whether significant differences exist. Therefore, the experiments consist of two phases: (1) determining the regression model for each LSP per modality having a set of features that accurately predicts emission intensity and (2) comparing emission intensities of LSPs and the corresponding confidence and prediction intervals to assess which LSP performs better. In section 5.1, we determine the final regression model using the exhaustive feature selection algorithm that excludes models with heteroscedasticity and non-normally distributed error terms. We perform a sensitivity analysis to determine the optimal algorithmic thresholds regarding the confidence and prediction interval width. We validate the algorithm through a comparison with stepwise regression. We also show that the algorithm performs well with limited data. To compare emission intensities of transport operations, we develop a regression model for each LSP per modality to predict emission intensities. We also include confidence and prediction intervals to compare LSPs with varying input data.

We have three subsets. Subset 1 contains data from LSP 1 for road transportation, subset 2 from LSP 1 for intermodal transportation, and subset 3 includes data from LSP 2 for road transportation. Subset 1 contains the following parameters: Great Circle Distance (GCD), Planned distance, empty distance, loaded distance, weight (dummy), Dedicated, and cleaning. Subset 2 contains similar parameters, except that weight is a continuous variable instead of a dummy variable. Subset 3 contains Distance, Weight, Tonne-kilometers, full truckload, groupage, and LTL.

5.1 Algorithmic performance and model comparisons

We want to obtain one prediction model for each subset having narrow confidence intervals of the mean response and prediction intervals. Therefore, the model must have a high predictive accuracy. This section aims to find the best model for each LSP per modality.

We compare different scenarios to determine the best prediction model for each subset. First, we compare how the results are influenced if we include insignificant independent variables. We determine the in-sample and the out-of-sample accuracy of both models. We obtain the best regression model for each subset S_j in terms of RMSE. We implement the exhaustive feature selection algorithm, determining the model with the smallest RMSE using the programming language R. We compare the features resulting from the algorithm with models determined using stepwise regression. Then, we select the best model for each subset.

5.1.1 Threshold Sensitivity Analysis

In this section, we explore the impact of different thresholds and variable inclusions on the performance of our predictive models. The analysis involves considering two thresholds for the Variance Inflation Factor (VIF) and evaluating the effects of including or excluding insignificant variables. We also investigate the trade-offs between model complexity and performance metrics. Table 3 presents an in-depth overview of the in-sample performance, highlighting variations across different models and thresholds. Additionally, we assess out-of-sample performance through Leave-One-Out cross-validation, providing insights into the models' generalization capabilities.

To determine the models with the smallest Root Mean Square Error (RMSE) of the model for different algorithm thresholds. We consider two Variance Inflation Factor (VIF) thresholds: 4 and 10. According to Hair et al. (2010), variables with a VIF between 4 and 10 should be reconsidered, and variables in models with a VIF above ten should be removed. Besides, we determine the effect of the significance level of variables on the uncertainty and variability of the emission intensity prediction. Therefore, we use five different thresholds for the t-test of each parameter in the model: 0.001, 0.01, 0.05, 0.1 and 1. Table 3 displays the in-sample accuracy of the different models. All models are determined using an exhaustive feature selection algorithm, selecting the model with the smallest RMSE while fulfilling all assumptions of Ordinary Least Squares. Table 4 shows an overview of the obtained independent variables per model.

We obtain the same model when excluding models with variables having a VIF of 4 or higher as 10 or higher. The exhaustive feature selection algorithm obtains the valid model with the smallest RMSE. Multi-collinearity could cause inflated error terms, resulting in a high RMSE. Models with multi-collinearity thus do not have a small RMSE. Additionally, we find different best models when setting the p-value threshold of the t-test to 0.001, 0.01, 0.05, 0.1 and 1. Models 1a, 2a, and 3a, which have a threshold of 1 and thus allow for insignificant variables, have the best goodness-of-fit. However, models 1c, 2f and 3b have the narrowest in-sample confidence interval. These models all exclude variables with significance levels of 0.01 or higher. Including insignificant variables includes a source of uncertainty, increasing the confidence interval width. For LSP 2, the confidence interval width is the smallest for a p-value threshold of the t-test of 0.01.

Model	Modality	LSP	VIF threshold	T-test p-value threshold	R-squared	Adjusted R-squared	RMSE	MSE	CI width in-sample
1a	Road	1	4;10	1	0.867	0.843	0.131	0.017	21.21
1b	Road	1	4;10	0.1	0.861	0.843	0.131	0.017	18.75
1c	Road	1	4;10	0.05;0.01;0.001	0.837	0.823	0.140	0.019	**16.74**
2a	Intermodal	1	4;10	1	0.801	0.721	0.332	0.110	24.93
2b	Intermodal	1	4;10	0.1	0.758	0.701	0.344	0.118	22.36
2c	Intermodal	1	4;10	0.05	0.714	0.667	0.363	0.132	20.65
2f	Intermodal	1	4;10	0.01;0.001	0.607	0.587	0.404	0.163	**15.51**
3a	Road	2	4;10	1; 0.05	0.423	0.408	0.353	0.125	15.75
3b	Road	2	4;10	0.01	0.414	0.404	0.354	0.126	**13.27**
3c	Road	2	4;10	0.001	0.401	0.391	0.358	0.128	14.11

Table 3: In-sample performance models with different thresholds

To determine how the model predicts new observations, we perform Leave-one-out cross-validation. We calculate each model's average confidence and prediction interval width of the left-out observation. Table 5 displays the out-of-sample performance. The prediction and confidence intervals are back-transformed. Model 2a has a high Mean Squared Error (MSE) and Standard Deviation (SD) for test data and a wide out-of-sample prediction and confidence interval. Model 2a includes variable heating compared to models 2b and 2c. We find that only one observation includes heating. This observation's out-of-sample standard error is exceptionally high, resulting in a high average MSE and SD. Therefore, we can conclude that model 2b is overfitting. Therefore, we exclude this model.

Moreover, we find that for LSP 1 for road transport, the prediction interval is the narrowest for a p-value

Model	Independent variables
1a	Loaded distance, empty distance, dedicated, and cleaning
1b	Loaded distance, empty distance, and dedicated
1c	Loaded distance, empty distance
2a	Loaded distance, empty distance, dedicated, cleaning, heating and weight
2b	Loaded distance, dedicated, cleaning and weight
2c	Planned distance, dedicated, cleaning
2f	Loaded distance
3a	Full truckload, groupage and tonne-kilometers
3b	Full truckload and tonne-kilometers
3c	Less than truckload and groupage

Table 4: Independent variables per model

			Training		Testing				
Model	VIF threshold	T-test p-value threshold	MSE	SD	MSE	SD	% in PI	PI width	CI width
1a	4;10	1	0.017	0.001	0.022	0.028	96.30	56.67	23.42
1b	4;10	0.1	0.017	0.001	0.020	0.026	**96.30**	**55.44**	20.39
1c	4;10	0.05; 0.01; 0.001	0.019	0.001	0.022	0.034	92.59	57.55	**17.93**
2a	4;10	1	0.111	0.008	0.624	2.257	100.00	$1.46{\cdot}10^6$	$1.42{\cdot}10^6$
2b	4;10	0.1	0.118	0.008	0.144	0.153	95.45	62.41	30.88
2c	4;10	0.05	0.132	0.009	0.152	0.166	100.00	60.91	26.23
2f	4;10	0.01;0.001	0.163	0.009	0.173	0.180	**100.00**	**59.40**	**17.51**
3a	4;10	1; 0.05	0.124	0.002	0.129	0.192	93.33	94.88	16.10
3b	4;10	0.01	0.126	0.002	0.129	0.191	92.50	**94.41**	**13.46**
3c	4;10	0.001	0.128	0.002	0.132	0.204	**95.76**	95.76	14.35

Table 5: Out of Sample performance - models with different thresholds

threshold of 0.1 (model 1b), while the confidence interval is the narrowest for a p-value threshold of 0.001 to 0.05 (model 1c). However, we would expect that model 1c also has the smallest prediction interval. An explanation could be that these prediction intervals do not accurately reflect the uncertainty associated with predictions on new data. However, the percentage of observations within the interval increases, indicating that model 1b is not overfitting. The difference between model 1b and model 1c is that model 1b contains a dedicated describing variable, while 1c does not. The p-value of dedicated in model 1b is 0.051, suggesting that there might be some evidence for the variable's impact. However, the evidence is not strong enough to reach conventional statistical significance, leading to differences in the prediction interval but not necessarily in the confidence interval.

The prediction interval considers both the uncertainty with the estimated mean response and the variability of individual observations. Including more variables, especially those not strongly significant, can lead to a wider confidence interval. Additional variables introduce more complexity into the model, capturing noise or random fluctuations in the training data. However, paradoxically, including some slightly insignificant variables reduces the variability of individual observations. In regression modelling, variability refers to how much individual observations deviate from the model's predicted values. Higher variability means individual observations can differ more from the predicted mean. Including dedicated as an independent variable reduces variability more than it increases the uncertainty of the mean and contributes to a smaller prediction interval.

Insight 1: The prediction interval width is smaller for a model with a slightly insignificant variable ($0.05<p<0.1$) than for a model with only significant variables ($p<0.05$) for road transportation.

For LSP 2, model 3b performs best regarding prediction and confidence interval width. This model includes full truckload and tonne-kilometers as describing variables. However, model 3c contains more predictions that are within the prediction interval, indicating that the prediction interval of model 3b is smaller but less accurate.

Our analysis highlights the trade-offs between model complexity and performance metrics such as RMSE, R-squared, and adjusted R-squared. Models with higher complexity, characterized by including more variables or lenient significance thresholds, often achieve higher R-squared values on the training data. However, this improvement in explanatory power may not translate to better predictive performance on out-of-sample data. Conversely, simpler models, obtained by imposing stricter thresholds or excluding insignificant variables, may exhibit lower R-squared values but demonstrate better generalization capabilities and narrower prediction intervals. It is essential to balance model complexity and predictive accuracy, considering the intended application and interpretability of the models in real-world scenarios. We find a p-value threshold of 0.01 overall, which results in the smallest average out-of-sample prediction and confidence interval width.

5.1.2 Stepwise regression

To validate the feature selection algorithm, we compare the models with one using forward stepwise feature selection to evaluate the exhaustive feature selection algorithm performance. Therefore, we include the transformed parameter having the highest correlation with the log-transformed emission intensity and determine if the model is significant. Then, we check stepwise if the parameter with the highest correlation has a significant t-test, thus significantly improving the model. Then, we repeat this until no more variables are improving the model. Besides, we check assumptions for the model and re-specify the model if assumptions are violated.

We obtain the same models 1c, 2f and 3b using forward stepwise feature selection in Table 3 and 5. These models are similar to those obtained using the exhaustive feature selection algorithm, implying that the exhaustive feature algorithm can perform similarly to stepwise regression, depending on the significance level of the independent variables we include. The exhaustive feature selection algorithm has advantages over stepwise feature selection because we can use any independent variables as input to obtain the model without manually testing the assumptions.

The models with only partial significance levels below 0.01 result in smaller confidence intervals. When we obtain a model using an exhaustive feature selection algorithm with a t-test p-value threshold of 0.01, we obtain the same models as the stepwise regression. Thus, the exhaustive feature selection algorithm works well as we choose a significance level threshold of 0.01 to reduce uncertainty. We obtain similar models by using stepwise feature selection. However, the exhaustive feature selection algorithm is more accessible for multiple data inputs, as the algorithm performs all assumption tests. We only have to give a set of parameters as input, and the algorithm obtains the model with the smallest RMSE.

Insight 2: We obtain the same models for stepwise regression as for the exhaustive feature selection algorithm with a p-value threshold of 0.01, implying that careful selection of the significance level threshold in the exhaustive feature selection algorithm can yield comparable results to stepwise regression while offering automation and ease of use.

5.1.3 Different data input

The purpose of comparing different data inputs is to evaluate the impact of varying datasets on our models' predictive accuracy and robustness. By assessing how models perform with data from the Logistics Service Provider (LSP) and our industry partner, we aim to determine how well the exhaustive feature selection algorithm performs on limited data regarding accurate predictions of emission intensity in road and intermodal transportation scenarios. The subset from our industry partner includes great circle distance, planned distance, and weight (dummy). In contrast, the subset from LSP 1 includes great circle distance, planned distance, empty distance, loaded distance, weight (dummy), dedicated, and cleaning. We only have data from our industry partner on LSP 1, and thus cannot perform this analysis for LSP 2. We only include models having a Variance Inflation Factor (VIF) smaller than 4.

We first determine the in-sample performance. Table 6 includes the models having the smallest out-of-sample prediction and confidence interval width, obtained in subsection 5.1.1 (Model 1b, 1c, 2c). We compare these models with the models using data from the LSP with models determined with limited data from our industry partner. Model 1d includes two variables: great circle distance and weight (dummy). We find that for road transportation, the predictive accuracy (adjusted R-squared) is three times higher when we include data on empty and loaded distances (model 1c vs model 1d). Additionally, the in-sample confidence interval width is two times smaller using data from the LSP for road transportation. For intermodal transport, this discrepancy is smaller. Data on actual loaded distance improves the adjusted R-squared and R-squared with 5%, compared to great circle distance to predict emission intensity. The confidence interval width even decreases with 2% if we include the actual loaded distance.

Insight 3: For road transportation, including data on empty and loaded distance significantly enhances predictive accuracy, resulting in a 2 to 3-fold improvement compared to using only great circle distance, which underscores the importance of incorporating more detailed data, such as empty and loaded distances, for more precise emission intensity predictions.

Model	Input data	Modality	T-test threshold p-value	R-squared	Adjusted R-squared	RMSE	MSE	CI width in-sample
1b	LSP	Road	0.1	**0.861**	**0.843**	**0.131**	**0.017**	18.75
1c	LSP	Road	0.05	0.837	0.823	0.140	0.019	**16.74**
1d	Industry partner	Road	1&0.05	0.333	0.278	0.282	0.079	34.23
2c	LSP	Intermodal	0.05	**0.714**	**0.667**	**0.363**	**0.132**	20.65
2f	LSP	Intermodal	0.01;0.001	0.607	0.587	0.404	0.163	15.51
2e	Industry partner	Intermodal	0.05	0.581	0.560	0.417	0.174	**15.25**

Table 6: In-sample performance models with different input data

To further analyze the decreasing confidence interval width, when including the actual loaded distance compared to the great circle distance, we perform Leave One Out Cross-Validation (LOOCV). We calculate each model's average back-transformed confidence and prediction interval width of the left-out observation. Table 7 shows the results. For intermodal transport, the out-of-sample confidence interval is also narrower for the model with great circle distance as the independent variable (model 2e) compared to the model with actual loaded distance (model 2f). We would expect data from the LSP to be more accurate, reducing uncertainty and thus reducing the interval widths. Model 2e also has a higher rate of new observations within the prediction interval, which indicates that the model is not overfitting.

One possible reason great circle distance better predicts emission intensity regarding confidence interval width is that there are errors in the data for intermodal transport. For example, according to A. Lewis and Greene (2019), distance should be calculated using actual rail network distance. However, rail distance can be difficult to find. The actual loaded distance data obtained from LSP 1 may contain errors. We cannot validate the loaded distance since we only have data on the total distance from our industry partner. Emission intensity is calculated as emissions per unit of loaded distance (e.g., metric tons of CO2 per tonne-kilometer). If the loaded distance is mismeasured, it can significantly impact the accuracy of emission intensity calculations.

We dive deeper into the individual out-of-sample observations to identify a potential cause of this discrepancy. For the model with loaded distance as describing variable (model 2f), we find one observation for which the out-of-sample confidence interval is extremely wide, namely 136.94. This lane is from Belgium to the Netherlands, with a great circle distance of 63.96 kilometres. This lane may be a misclassification error since such a short-distance lane is often conducted with a truck. For model 2e, which contains great circle distance as an independent variable, the out-of-sample confidence interval is less extreme, namely 96.74 for this lane. If we exclude this lane from the testing dataset of both models, the model with loaded distance and the model with great circle distance have a confidence interval width of 10.71 and 12.62, respectively. Assuming that this is indeed a misclassification error, we can conclude that the model with loaded distance performs better regarding out-of-sample and in-sample accuracy.

				Training		Testing				
Model	Input data	Modality	T-test threshold p-value	MSE	SD	MSE	SD	% in PI	PI width	CI width
1b	LSP	Road	0.1	0.017	0.001	0.020	0.026	**96.3**	**55.44**	20.39
1c	LSP	Road	0.05	0.019	0.001	0.022	0.034	92.6	57.55	**17.93**
1d	Industry partner	Road	1;0.05	0.079	0.006	0.092	0.155	92.6	122.02	38.41
2c	LSP	Intermodal	0.05	0.132	0.009	0.152	0.166	100.0	60.91	26.23
2f	LSP	Intermodal	0.01;0.001	0.163	0.009	0.173	0.180	**100.0**	**59.40**	17.51
2e	Industry partner	Intermodal	0.05;0.1	0.174	0.009	0.182	0.181	100.0	59.67	**16.44**

Table 7: Out of Sample performance - models with different input data

Insight 4: In intermodal transportation, utilizing great circle distance as an independent variable results in a narrower confidence interval than actual loaded distance. This discrepancy is attributed to a misclassification error in the transportation modality of one lane. This finding underscores the importance of data accuracy and highlights the potential impact of errors on predictive model performance in intermodal

transport analysis.

Data availability consistently reduces the prediction interval width of new observations. This observation suggests that our models become more precise in predicting individual data points with more parameters. Despite the constraints of limited data, the exhaustive feature selection algorithm generates valid models, highlighting the algorithm's ability to predict emission intensities even from limited data availability. Moreover, this analysis underscores the importance of data accuracy and the influence of errors on the confidence and prediction interval widths.

5.1.4 Model selection

We aim to select the most suitable model based on the prediction and confidence interval widths, as these predict emission intensity with limited uncertainty and variability. To compare emissions of LSPs, we want to accurately predict emission intensity to identify potential differences in carbon performance. For LSP 2, model 3b demonstrates the smallest prediction and confidence intervals, making it the optimal choice as a prediction model for road transportation. Regarding road transportation for LSP 1, model 1b shows the best performance regarding the prediction interval, while model 1c outperforms the confidence interval. Although both models exhibit strong performance, model 1c is selected as more straightforward and easier to interpret while maintaining comparable performance. Examining Table 8, we observe that models 1b and 1c differ primarily in including the "Dedicated" variable, which has an insignificant p-value of 0.051. We observe that the coefficients for 'Loaded Distance' increase and 'Empty Distance' decrease in magnitude when the 'Dedicated' variable is included in the model. This reduction is attributed to dedicated shipments, which account for a portion of the empty distance. Given the negligible difference in performance and the desire for simplicity, model 1c is chosen. Model 2f performs best after excluding the misclassified lane from the testing set.

Model	Intercept	$GCD^{0.2}$	Planned $Distance^{0.2}$	Loaded Distance (log)	Empty $Distance^{0.3}$	Dedicated	Tank Container	Cleaning	Heatings	Weight
1a	5.135***	-	-	-0.282 ***	0.162***	0.125*	-	0.064	-	-
1b	5.112***	-	-	-0.274***	0.163***	0.107	-	-	-	-
1c	**5.102*** **	**-**	-	**-0.267*** **	**0.170*** **	-	-	-	-	-
1d	5.698***	-0.321**	-	-	-	-	-0.367*	-	-	-
2a	8.232***	-	-	-0.697***	0.082	0.378	-	0.468	0.500	-0.053
2b	7.686***	-	-	-0.585***	-	0.552*	-	0.593*	-	-0.045
2c	5.390***	-	-0.7547***	-	-	0.626*	-	0.729*	-	-
2d	7.126***	-0.828***	-	-	-	-	-	-	-	-0.038
2e	6.351***	-0.857***	-	-	-	-	-	-	-	-
2f	**7.690*** **	-	-	**-0.649*** **	-	-	-	-	-	-

Table 8: Potential models with features, coefficients and significance levels of LSP 1

In the remainder of this section, we discuss the final models of each LSP per modality. We need these models to compare emission intensities while incorporating confidence and prediction intervals in section 5.2.

Road transportation

We obtain two models for road transportation, one predicting the emission intensity of LSP 1 and one of LSP 2. For LSP 1, we obtain a final model as described in Table 9. According to the default factors in the chemical industry, dedicated shipments have approximately 30 to 50% higher emission intensity compared to spot shipments (Smart Freight Centre & Cefic, 2021). Dedicated shipments emit more due

to high empty kilometres in dedicated shipments. Empty distance is already a describing variable in the model, which could explain that the variable dedicated is insignificant. Additionally, we did not find a significant relationship between cleaning and emission intensity. This lack of significance may stem from the fact that only 30% of the lanes require cleaning, thus minimizing its overall impact. Similarly, the weight dummy variable was found to be insignificant and, therefore, not included in the model. This insignificance could be attributed to the unequal distribution of tank containers (85%) and tank trucks (15%) within the dataset.

	Estimate	2.50%	97.50%	S.E.	t value	p-value
(Intercept)	5.10166	4.805291	5.398025	0.1436	35.528	<0.001
Loaded distance (log)	-0.26732	-0.32714	-0.20749	0.02899	-9.222	<0.001
Empty distance$^{0.3}$	0.17045	0.133625	0.207265	0.01784	9.554	<0.001

Table 9: Final regression model of LSP 1 for road transportation

For LSP 2, we use S_2 as input containing the transformed tonne-kilometer, weight, distance, less than truckload, full truckload and groupage. Table 10 displays the final model. The transformed tonne-kilometer (b=−0.08, p<0.001) negatively correlates with the logarithmic emission intensity, meaning that if the tonne-kilometer increases, the emission intensity decreases. However, the impact of including tonne-kilometers as the effect is quite low, with an adjusted R-squared of 0.40 compared to 0.42.

We expect a positive relationship between distance and emission intensity since long-distance legs often include more highways and thus better traffic conditions, reducing emission intensity. However, the variable 'Distance' is excluded from the model due to its minimal correlation with the log emission intensity. Despite efforts to find a more suitable transformation, no significant relationship emerged. We expect a negative relationship between distance and emission intensity with diminishing returns. We do not observe any noticeable pattern when visually representing this relationship by plotting the data on a scatterplot. This scatterplot is shown in Appendix E.

Additionally, we are 95% confident that the emission intensity of full truckload is 22.88% to 43.4% lower than the less-than-truckload or groupage shipments (b=−0.415, p<0.001). This aligns with our expectations, as full truckload shipments are often directly driven from origin to destination.

	Estimate	2.50%	97.50%	S.E.	t value	p-value
(Intercept)	4.69096	4.447105	4.93482	0.12313	38.097	<0.001
Tonne-kilometer$^{0.2}$	-0.0788	-0.12828	-0.02932	0.02498	-3.154	<0.01
FTL	-0.41471	-0.5696	-0.25981	0.07821	-5.302	<0.001

Table 10: Final regression model of LSP 2 for road transportation

Remarkably, loaded distance does not influence the emission intensity of LSP 2, but it does for LSP 1. There are some explanations for why distance does not influence the emission intensity since emission intensity is in grams of CO_2e per tonne-kilometer. We expect that loaded distance has a larger impact on emission intensity for short-distance shipments. For short-distance shipments, vehicles may not have sufficient time to reach optimal operating conditions. Cold starts and frequent stops, characteristic of

short trips, can lead to lower fuel efficiency and higher emission intensity. The average loaded distance is 2.1 times smaller for LSP 1 than for LSP 2, which could explain this difference in the model. Figure 2 shows that the effect of loaded distance on emission intensity diminishes for loaded distances higher than 250 kilometres.

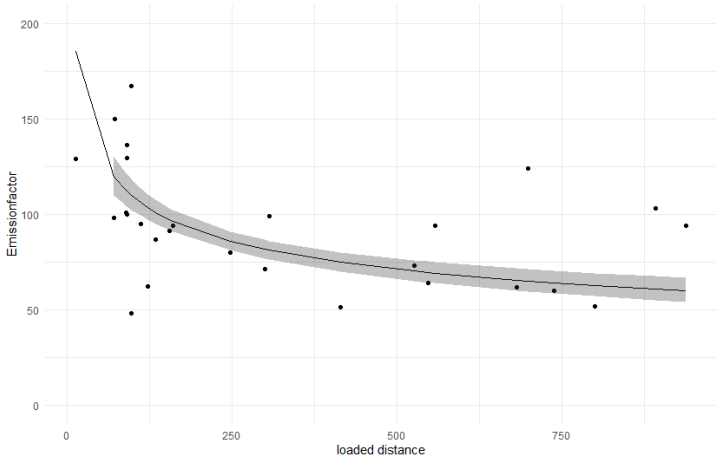

Figure 2: LSP 1 Road - Confidence interval of the mean response (constant empty distance)

Insight 5: While loaded distance significantly influences emission intensity for LSP 1, with an average loaded distance significantly 2.1 times shorter than LSP 2, it does not exhibit a significant effect for LSP 2. This suggests a relationship between loaded distance and emission intensity that diminishes for longer distances.

Intermodal transport

Table 11 shows the final model for intermodal transport. This model contains one describing variable: loaded distance. There is no clear relationship between empty miles and emission intensity. A potential reason could be that empty miles do not influence emissions in rail transport since trains have very low emissions overall. Besides, cleaning is already in the model and could also represent the empty runs for a part. One crucial problem with intermodal transport is that we do not know what part of the trip is road and what part is rail or sea transport.

	Estimate	2.50%	97.50%	S.E.	t value	p-value
(Intercept)	7.6902	6.006738	9.37358	0.807	9.529	$<$0.001
Loaded distance (log)	-0.6494	-0.89306	-0.4057	0.1168	-5.559	$<$0.001

Table 11: Final regression model of LSP 1 for intermodal transportation

5.2 Comparison of LSPs

In this section, we compare LSPs regarding the confidence intervals of the mean response or the prediction interval. As discussed in the problem statement, we are interested in the confidence interval of the mean emission intensity if the LSP would bill future emission taxes based on the average of each lane. We are interested in the prediction interval of the emission intensity if the LSP would bill future emission taxes for every shipment. Therefore, the confidence interval or prediction interval is calculated to quantify

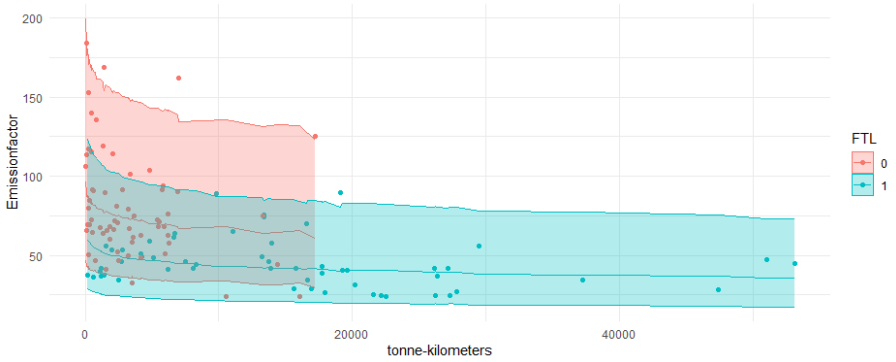

Figure 3: Out-of-sample prediction interval of emission intensity for road transport of LSP 2

the uncertainty of emission intensity, depending on carbon tax billing characteristics per LSP. However, emission taxes are not obligatory yet; thus, we do not know how LSPs will forward the taxes. Therefore, we make a comparison in terms of the prediction interval and the confidence interval of the mean response.

5.2.1 Prediction interval widths

Section 5.1 describes the average out-of-sample prediction interval width of 55.44 and 94.41 for LSP 1 and 2, respectively. The interval width of LSP 2 is 1.7 times wider than that of LSP 1. A potential reason could be that the R-squared for LSP 1 is two times higher than LSP 2. A higher R-squared suggests that the model captures a more significant proportion of the variability in the response variable. In other words, a high R-squared suggests less unexplained variability or randomness in the data. The prediction interval considers the uncertainty associated with estimating the mean response and the variability of individual observations around that mean. If there is more randomness in the data, it implies that individual observations may deviate more from the predicted mean, leading to a wider prediction interval. We have less data available on LSP 2 than on LSP 1. LSP 2 also performs groupage and LTL shipments, while LSP 1 only performs FTL shipments.

Figure 3 shows the prediction interval of LSP 2. There are no values for groupage and LTL above 20000 tonne-kilometers; thus, we cannot predict above this value. We find that the prediction interval is wider for Groupage and Less than truckload shipments than full truckload shipments. This difference could have multiple potential reasons. Groupage and less-than-truckload shipments often involve smaller and lighter loads than full-truckload shipments. The variability in the size and weight of shipments can contribute to increased uncertainty in predicting emissions, leading to a wider prediction interval. Besides, Groupage and Less than Truckload shipments typically involve multiple origins and destinations within a single shipment. The diverse nature of these shipments makes it challenging to accurately predict emission intensities due to the varying conditions at different stops. Moreover, groupage and Less-than-truckload shipments may have more complex routing and scheduling requirements, with multiple stops and changes in route. The increased complexity in transportation can introduce additional sources of variability, widening the prediction interval.

LSP 2 has a wider prediction interval, while LSP 1 has a wider confidence interval. Since the prediction interval consists of both the variability of individual observations and uncertainty in the mean response, this wide confidence interval is primarily due to the wide variability of individual observations since the confidence interval of the mean response is relatively small. For this purpose, we determine the

Coefficient of Variance (CV) of both LSPs to determine if LSP 2 has a wider variability of observations. Table 12 shows the mean, standard deviation (SD) and Coefficient of Variance (CV) within both LSPs. We find that the relative variability of individual observations of LSP 2 is 1.5 times larger than that of LSP 1. Thus, the model of LSP 2 captures two times less variability in the data and contains 1.5 times more variability than LSP 1, which could result in a wider prediction interval.

	Mean	SD	CV
LSP 1	93.16	30.94	33.21
LSP 2	64.20	32.17	50.10

Table 12: variability of emission intensity

Insight 6: The prediction model of LSP 2 captures two times less variability in the emission intensity and contains 1.5 times higher relative variability of individual observations than that of LSP 1, which could result in a 1.6 times larger prediction interval width of LSP 2 compared to LSP 1, that only conducts full truckload shipments. LSP 2 also conducts less than truckload and groupage shipments, which have a wider prediction interval than full truckload shipments.

5.2.2 Varying interval widths

We find that the confidence and prediction interval widths are not constant for each proportion of empty distance compared to the total distance of LSP 1. This subsection analyses the potential causes and consequences of the varying prediction and confidence interval widths.

The final model predicting emission intensity for road transportation of LSP 1 contains two describing variables: loaded distance and empty distance. Appendix F contains two figures that show the relationship between loaded distance and emission intensity and empty distance and the emission intensity individually. Figure 4 shows the proportion of the empty distance and the corresponding emission intensity with prediction intervals. The red dots represent the left-out value of the Leave-One-Out Cross-Validation (LOOCV). A linear positive trend exists between the empty distance proportion and the emission intensity. Figure 4 shows that the prediction interval widens if the proportion of empty kilometres increases. The widening prediction interval could be due to higher variability of observations for higher tonne-kilometres, as the red dots are further apart for higher empty distances, meaning that the mean squared error is higher. Moreover, there are two actual values outside the prediction interval due to multiple potential reasons. Shipments with a small proportion of empty distances might have more reliable and accurate data. Shipments having extremely high empty distances seem off, as the industry average is 22%. Inaccurate or incomplete data can negatively impact the model's performance. Higher data quality for shipments with low empty distances contributes to better predictions. We cannot validate the data on empty and loaded distances as our industry partner does not have this data. Additionally, shipments with low empty distances may have lower variability in factors influencing transportation, such as delivery windows, loading/unloading times, and operational constraints. Reduced variability makes it easier for the model to capture and predict patterns.

Figure 5 shows the confidence intervals of the mean response of LSP 1 for road transportation for each proportion of empty distance. The observations do not fall for 95 % within the confidence interval because the interval represents the 95 % probability that the mean emission intensity falls within the interval

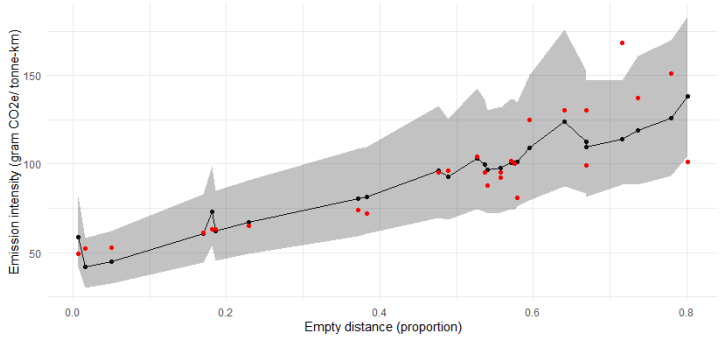

Figure 4: LSP 1: prediction interval out-of-sample proportion empty distance

and not the individual observation. Around 65 % empty distance, the confidence interval gets very wide and the emission intensity zigzags. The lane around 65 % is also markable in Figure 4 for the prediction interval, as the predicted emission intensity increases and decreases slightly. This zigzag pattern could be due to a variable not being incorporated into the model while influencing the emission intensity. For example, it could be that traffic conditions or road characteristics also influence the emission intensity, while this is not incorporated in the model.

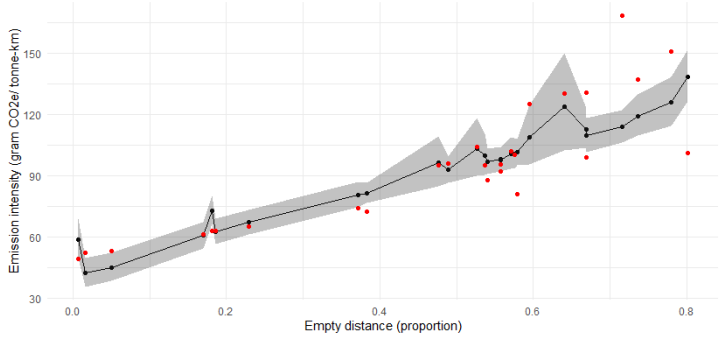

Figure 5: LSP 1: Confidence interval out-of-sample proportion empty distance

We further investigate the individual predictions that result in a high interval width. For example, the observation having 64 % empty distance results in a confidence interval width of 47.86, which is extremely high compared to the average of 17.9. This lane is spot-contracted. The great circle distance is only 8.8 kilometres, which is extremely low, and the emission intensity is 129.09, which is high. The model may not work well on such short distances due to limited data on extremely short distances because there is not enough training data on very short distances.

To reduce the prediction interval width and to improve the predictions for high proportions of empty kilometres, it is essential to identify the causes of these high proportions of empty kilometres and to validate the data. Therefore, we would need data on the routes driven, including the empty distance.

Insight 7: Prediction interval widths increase with higher proportions of empty distance in road transportation and confidence interval widths increase for specific lanes with high proportion of empty distance both due to increased variability and uncertainty, potentially due to unaccounted additional influential factors, lack of training data or due to errors in the data.

Figure 6 shows the confidence intervals of the mean response of LSP 2 road transportation. The interval gets wider as the tonne-kilometers increase for groupage or less than truckload shipments. We have fewer data points for values with high tonne-kilometers, leading to increased uncertainty in estimating the parameters. Fewer data points can result in wider confidence intervals when less information is available to make accurate predictions.

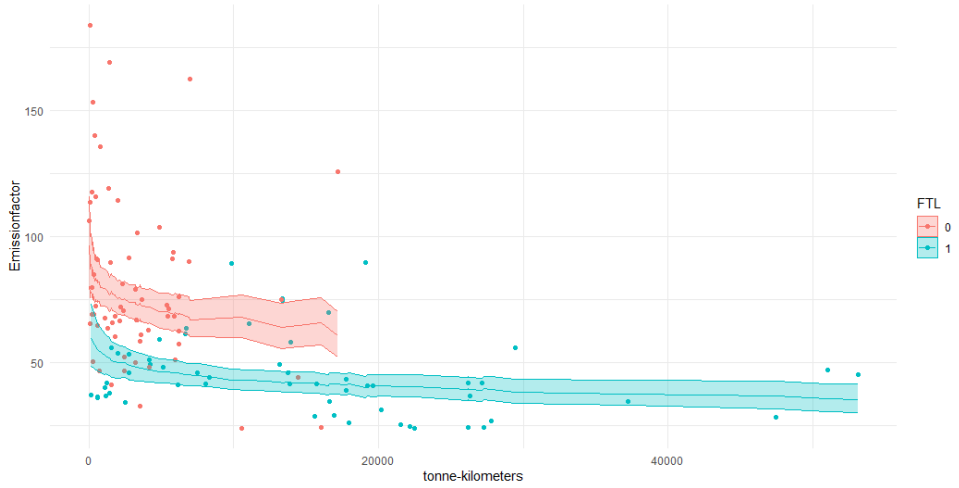

Figure 6: LSP 2 Road - Confidence interval of the mean response

5.2.3 Comparing emissions of LSPs

Figure 7 depicts emission intensity and confidence intervals against loaded distance for LSP 1 and LSP 2, enabling comparison despite loaded distance not being a model variable for LSP 2. Regarding LSP 2, the upper blue confidence interval range represents full truckload shipments, while the lower range denotes less than truckload and groupage shipments. LSP 1 only conducts shipments with distances under 1000 kilometres; thus, we cannot compare LSP 1 and 2 for higher distances. The emission intensity of LSP 1 also highly depends on the empty distance, causing a zigzag pattern. For example, LSP 1 conducts shipments using seven lanes between 75 and 100 kilometres, where the empty distance fluctuates significantly, resulting in high discrepancies in emission intensity. Unfortunately, we do not know the empty distance of LSP 2 and thus cannot compare. To compare LSP 1 and 2, we must compare LSP 1 with the lower blue range, as these represent the full truckload shipments and LSP 1 only conducts bulk transportation shipments, which are full truckload. LSP 2 shows superior emission intensity performance compared to LSP 1. In certain instances, the emission intensity for LSP 1 exceeds that of LSP 2 by more than double.

This discrepancy between LSP 1 and 2 could have multiple reasons. Therefore, we compare the emission intensities with the industry averages. Full truckload ambient and temperature-controlled packed goods have an emission intensity of 63 and 71, respectively. Industry averages exceed the full truckload shipments of LSP 2 of 44.9 grams of CO2e per tonne-km. Thus, LSP 2's emission intensities are 1.3 to 1.6 times lower than the chemical transportation industry averages. Potentially, LSP 2 performs better than the industry averages, or LSP 2 underestimates the emissions in their calculations. LSP 2 uses the fuel-based method if shipments are not outsourced, which is the case for about 60 % of the shipments. The fuel-based method often results in smaller emission intensities than the distance-based method using

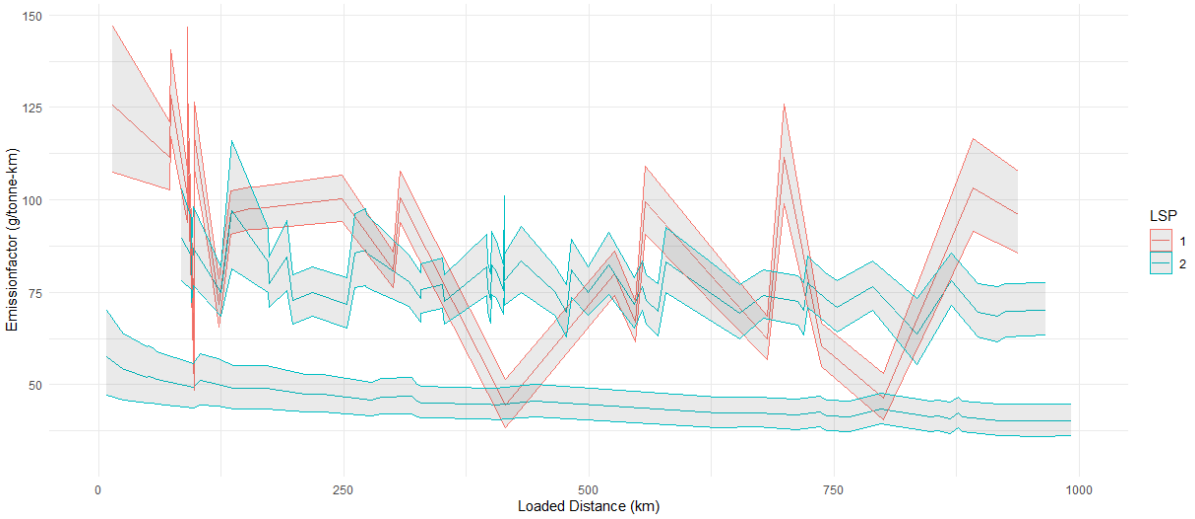

Figure 7: Comparison of emission intensities including confidence intervals for road transportation

default factors to encourage LSPs to use the fuel-based method (A. Lewis & Greene, 2019). LSP 2 uses a modern fleet and offers eco-driving courses as part of its sustainability efforts. However, LSP 2 does not (yet) use biodiesel as a fuel type. LSP 2 allocates emissions based on the average fuel use per business unit per tonne-kilometer, specifically the chemical business unit. This approach may not accurately capture the variability in fuel consumption across different shipments or business units. As a result, emissions related to certain types of shipments or activities within the logistics chain may be underestimated. We do not know if LSP 2 also includes emissions from cleaning and heating the vehicle. If these emission sources are not accounted for in their calculations, it could lead to underestimating emissions. Omitting such factors may result in an incomplete picture of the environmental impact of LSP 2's operations. Considering these factors, it is more plausible that LSP 2's reported emission intensities do not fully represent the actual environmental footprint of their operations. Further investigation into their emission estimation methods and including additional emission sources could provide a clearer understanding of the extent of underestimation and help improve the accuracy of their emission assessments. However, it is difficult to determine if LSP 2 underestimates emissions because the industry averages from the GLEC Framework are higher than the actual emission intensity.

Insight 8: The full truckload emission intensities are 1.3 to 1.6 times lower for LSP 2 than the industry averages due to underestimated emissions, high carbon performance or overestimated default factors.

LSP 1 has an average emission intensity of 93.16 g CO2e per tonne-km. LSP 1 only transports bulk goods. According to A. Lewis and Greene (2019) bulk goods have an average load of 22 tonnes and empty running of 22%. Ambient bulk goods have an emission intensity of 61 g CO2e per tonne-km, while temperature-controlled bulk goods have an emission intensity of 68 g CO2e per tonne-km. For LSP 1, dedicated shipments are, on average, 98.87 g CO2e per tonne-km, which is within the industry average of 76 to 101 g CO2e per tonne-km (depending on the vehicle type and heating/cooling requirements). The spot-contracted shipments emit 86.03 grams per tonne-km for LSP 1, 1.2 to 1.6 times larger than the industry averages between 55 and 70 g per tonne-km. This difference from industry norms suggests potential discrepancies in data accuracy or kilometre allocation mistakes to our industry partner.

Further analysis reveals that dedicated lanes in LSP 1 experience an average of 49.2% empty distances, while spot lanes show 43.6%, compared to industry averages of 50% and 22 %, respectively. There is a large discrepancy between the industry averages and the spot-contracted shipments. A possible reason could be errors in the data on empty and loaded distances, as we can only validate the total distances. Errors in the loaded distance would have enormous consequences, resulting in wrongly calculated emission intensities. LSP 1 allocates too many empty kilometres to our industry partner. Another possibility is misclassification errors in the parameter contract type, meaning that some spot-contacted lanes are dedicated contracted. It is also possible that LSP 1 drives many empty kilometres for spot-contracted lanes, resulting in very high emission intensities. The reasons could be that LSP 1 might face operational constraints, such as tight delivery schedules or limited access to alternative routes, resulting in less efficient transportation practices and higher emissions. Another reason could be that the routes taken by LSP 1 for spot-contracted shipments may involve more challenging terrain or longer distances, leading to higher fuel consumption and emissions. Additionally, external factors such as traffic congestion, weather conditions, or road closures could impact the efficiency of transportation operations and contribute to higher emissions for spot-contracted shipments.

It is more likely that the discrepancies in emission intensities between spot-contracted shipments and industry averages result from errors in data accuracy or kilometre allocation rather than LSP 1 performing significantly more empty kilometres. To determine if LSP 1 performs significantly more empty kilometres than LSP 2, we must gather data on the loaded distance and empty distance to validate the data of LSP 1.

Insight 9: The spot-contracted emission intensities for LSP 1 are 1.2 to 1.6 times larger than industry averages due to overestimated empty distances, distance allocation errors, misclassification errors, measurement errors or high proportions of empty distance compared to industry averages.

To further analyze the differences between LSP 1 and 2, we compare two similar lanes where LSP 1 and 2 both perform shipments. Table 13 shows the predicted emission intensity and the corresponding confidence interval for lane 1 and 2 characteristics. Table 14 shows average data on the lanes. Neither the model for LSP 1 nor LSP 2 accurately predicts the mean emission intensity of lane 1. The actual emission intensity is lower than the predicted emission intensity for both models. This error could be due to this lane's short distance, and the model does not work well for very low distances. For lane 2, the predictions are very close to the actual emission intensity. However, the confidence and prediction interval widths for the model of LSP 1 are wider for lane 2 than lane 1. LSP 1 does not perform many long-distance shipments, resulting in higher uncertainty. We also find that LSP 1 has broader confidence interval widths for both lanes, while LSP 2 has broader prediction interval widths for lane 1.

Furthermore, LSP 2 performs better for both lanes regarding total well-to-wheel emissions and emission intensity. LSP 1 does not conduct cleaning or heating procedures for shipments on either lane. LSP 1's loaded distance is 1.8 times larger than that of LSP 2 for lane 1. A potential reason for this discrepancy is that LSP 1 and 2 may have different routing strategies, leading to variations in the distance covered. Routing strategies could involve choosing different highways or roads with varying distances between the same origin and destination. However, this is not very likely as Google Maps only gives us routes between 56 and 61 kilometres. Some intermediate stops may not be visible in our industry partner's data. However, it is more likely that there are errors in the loaded distance data. For lane 2, the emission intensity of LSP 1 is 2.5 times higher than that of LSP 2. While the loaded distance of LSP 1 exceeds

that of LSP 2 by only 2%, the significant disparity in emission intensities suggests that empty distances substantially contribute to LSP 1's higher emissions. For lane 2, the empty distance of 992.52 kilometres is similar to the industry average of 50% empty distance for dedicated shipments.

Insight 10: LSP 2 performs better than LSP 1 in emission intensity and total well-to-wheel emissions for similar lanes. As LSP 1 contains dedicated contracted lanes, the empty distance is larger than industry averages, which could explain the higher emission intensities of LSP 1 compared to LSP 2. The loaded distance of LSP 2 is smaller than LSP 1 due to more efficient driving or errors in the data, resulting in lower emission intensities of LSP 2 compared to LSP 1.

Lane	LSP	WTW emissions (tonnes)	Actual emission intensity	Fit	CI lower	CI upper	CI width	PI lower	PI upper	PI width
1	1	0.11	48.26	56.08	48.59	64.72	16.13	40.66	77.35	36.69
1	2	0.05	42.07	51.95	44.98	60.01	15.03	25.38	106.35	80.97
2	1	1.94	103.00	103.20	91.38	116.54	25.15	75.50	141.05	65.55
2	2	0.76	40.75	40.84	36.86	45.25	8.39	20.10	83.01	62.91

Table 13: Comparison of LSPs for similar lanes and the prediction and confidence intervals

Lane	LSP	Shipment weight	Tonne-km	Shipment Profile	GCD	Planned distance	Total distance	Loaded distance	Empty distance	Contract type
1	1	24.28	2354	-	44.39	60.00	97.64	96.97	0.66	Spot
1	2	22.04	1211	FTL	-	-	-	55.02	-	-
2	1	21.16	18844	-	619.73	803.00	1883	890.48	992.52	Dedicated
2	2	22.13	19213	FTL	-	-	-	869.07	-	-

Table 14: Data of the lane comparison

6 Conclusion and discussion

Greenhouse Gas emissions have a significant environmental impact, requiring reduction. Therefore, the European Commission mandates companies to report emissions. This is the first step towards implementing carbon taxes for transportation. Choosing LSPs with lower emissions ensures compliance with current and future regulations. However, companies often face data quality and availability challenges, making it difficult to estimate emissions accurately. These challenges result in a lack of comparability between different logistics service providers (LSPs) due to variations in calculation methods, data quality and availability levels. Therefore, this research developed an approach to compare the interval estimates of emission intensity of LSPs while differing in data quality and availability. We first determine one prediction model for each LSP per modality that predicts interval estimates. Thus, we obtain models with narrow confidence and prediction interval estimates to identify differences between LSPs. The interval in which we are interested depends on the future billing characteristics of the LSP. If the LSP bills the carbon taxes per average emissions per transportation lane, we are interested in the confidence interval of the emission intensity. If the LSP forwards carbon taxes per individual shipment, we are interested in the prediction interval of the emission intensity. As we do not know how LSPs will bill carbon taxes in the future, we determine both confidence and prediction intervals. After determining a prediction model for each LSP per modality, we compare the emission intensity intervals of LSPs. For this purpose, we use an exhaustive feature selection algorithm that considers all combinations of features and excludes combinations that violate assumptions. We obtain the best model minimizing the Root Mean Squared Error (RMSE).

Our analysis reveals that the exhaustive feature selection algorithm with a p-value threshold of 0.01 obtains the same models as for forward stepwise regression, implying that careful selection of the significance level threshold in the exhaustive feature selection algorithm can yield comparable results to stepwise regression while offering automation and ease of use.

Furthermore, we find that the model with the narrowest confidence interval is not always similar to the model with the narrowest prediction interval. The confidence interval width highly depends on the significance levels of the independent variables. Models with only independent variables with p-values below 0.001 result in narrower confidence intervals. However, prediction interval widths are also influenced by how well the variability is explained by the model.

We find that distance does not influence emission intensity for LSP 2, while it does influence emission intensity for LSP 1 for road transportation. We find that the loaded distance of LSP 2 is 2.1 times larger than that of LSP 1. Due to frequent stops and cold starts, short distances impact emission intensity in grams of CO2e per tonne-kilometers. As emission intensity is in grams per tonne-kilometers, emission intensity reaches a constant level as the distance increases. Additionally, we find that the prediction intervals for LSP 2 are 1.6 times wider than LSP 1. As the prediction interval captures the variability of individual observations, while the confidence interval does not, LSP 2 contains more variability in the data. The coefficient of variation is 1.5 times larger for LSP 2, and the model of LSP 2 captures two times less variability in the emission intensity than LSP 1. This results in a wide prediction interval.

Our analysis reveals markedly lower emission intensities for LSP 2 than for LSP 1. This discrepancy can be partly attributed to LSP 1's involvement in dedicated contracted shipments, which typically entail

high empty distances and result in elevated emission intensities. Furthermore, we observed that the emission intensities for full truckload shipments at LSP 2 are 1.3 to 1.6 times lower than industry averages. This could stem from underestimated emissions, superior carbon performance, or overestimated default factors. Conversely, spot-contracted emission intensities for LSP 1 were found to be 1.2 to 1.6 times larger than industry averages. This disparity may arise from overestimated empty distances, errors in distance allocation, misclassification errors, measurement inaccuracies, or elevated proportions of empty distances compared to industry averages. Moreover, LSP 2 demonstrated better performance in emission intensity and total well-to-wheel emissions than LSP 1 for similar lanes.

We employed multiple linear regression to determine the confidence and prediction intervals in predicting and comparing Logistics Service Providers' (LSPs) emissions. While this approach provides valuable insights into the relationships between emissions and various variables, it is crucial to acknowledge the limitations of parametric methods, especially when dealing with non-normally distributed variables. While our study primarily focused on parametric methods for emission estimation, future research could explore the application of non-parametric methods as alternatives. Non-parametric methods could offer more robust comparisons across different LSPs, particularly for variables with non-normal distributions, such as shipment weight. These methods are not sensitive to the shape of the distribution and can provide accurate results even in the presence of skewed or non-normally distributed data. However, non-parametric methods do not work well with limited data points and could determine less precise intervals. Future research should compare non-parametric methods with parametric methods.

Data availability and quality are often issues as we collect data for outsourced transport operations. While we obtained data from two LSPs, the data still lacks quality and availability. For example, we did not have data on empty distances for one LSP. To further compare the emission intensities of LSPs, it is interesting to compare the empty distances of LSPs and the antecedents of empty distances. Given the high amount of empty distances of LSP 1, conducting a detailed analysis of empty distance patterns across various LSPs and transportation modes would provide valuable insights. Understanding the factors influencing empty distances, such as route planning strategies and operational efficiency, could inform strategies for emissions reduction. Future research should prioritize efforts to enhance the quality and availability of data related to outsourced transport operations. This includes the validation of data on empty distances and the development of comprehensive datasets to facilitate more detailed analyses.

We developed a method to compare LSPs. However, we must rely on emission intensities determined by LSPs. While these emission intensities are determined using validated methods, these may still contain errors. Additionally, we cannot validate all independent variables in the prediction model for emission intensity. We used ordinary least squares to obtain parameter estimates due to a lack of validation data. However, using validation data, we can determine the magnitude of the random and measurement errors in both the independent variables and the dependent variable of the model. Therefore, future research should focus on gathering validation data. With validation data, it is possible to perform Total Least Squares (TLS) if measurement errors are present in the independent variable or Feasible Generalized Least Squares (FGLS) if the dependent variable is estimated instead of measured.

References

Abdulazeez, A. (2014). *Differences between parametric and non-parametric tests plus their advantages and limitations.*

Acheampong, A. O., & Boateng, E. B. (2019). Modelling carbon emission intensity: Application of artificial neural network. *Journal of Cleaner Production*, *225*, 833–856.

Auvinen, H., Clausen, U., Davydenko, I., Diekmann, D., Ehrler, V., & Lewis, A. (2014). Calculating emissions along supply chains—towards the global methodological harmonisation. *Research in Transportation Business & Management*, *12*, 41–46. doi: 10.1016/j.rtbm.2014.06.008

Cullen, A. C., & Frey, H. C. (1999). *Probabilistic techniques in exposure assessment: A handbook for dealing with variability and uncertainty in models and inputs.* Springer.

Davydenko, I., Ehrler, V., de Ree, D., Lewis, A., & Tavasszy, L. (2014). Towards a global co2 calculation standard for supply chains: Suggestions for methodological improvements. *Transportation Research Part D: Transport and Environment*, *32*, 362–372.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*, *6*(4), 45.

Ding, S., Dang, Y., Li, X., Wang, J., & Zhao, K. (2017). Forecasting chinese co2 emissions from fuel combustion using a novel grey multivariable model. *Journal of Cleaner Production*, *162*, 1527–1538.

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, *37*(3/4), 409–428. Retrieved 2024-01-04, from `http://www.jstor.org/stable/2332391`

EFRAG. (2023). *European Sustainability Reporting Standards E1 CLIMATE CHANGE* (Tech. Rep.).

Frey, H. C. (2007). Quantification of uncertainty in emission factors and inventories. *Unpublished manuscript, Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh*.

GHG Protocol. (2023). *GHG Protocol Guidance on Uncertainty Assessment in GHG Inventories and Calculating Statistical Parameter Uncertainty* (Tech. Rep.). GHG Protocol. Retrieved from `https://ghgprotocol.org/calculation-tools-and-guidance`

Hair, J., Barry, J. B., Rolph, E. A., & Rolph, E. A. (2010). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall.

Herold, D. M., & Lee, K. H. (2017). Carbon management in the logistics and transportation sector: An overview and new research directions. *Carbon Management*, *8*(1), 79–97.

Hörandner, L., Egger, L. M. P., & Beil, D. (2023). Calculating emissions along multimodal transport chains-standards, difficulties and problems. In *Proceedings of pianc smart rivers 2022: Green waterways and sustainable navigations* (pp. 1338–1341). Springer.

IPCC. (2006). Ipcc guidelines for national greenhouse gas inventories.

Keles, T. (2018). Comparison of classical least squares and orthogonal regression in measurement error models. *International Online Journal of Educational Sciences*, *10*(3), 200–214.

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, *34*(4), 1060–1073.

Koza, J., Bennett III, F., Andre, D., & Keane, M. (1998, 09). Automated design of both the topology and sizing of analog electrical circuits using genetic programming. , 151–170. doi: 10.1007/978-94-009-0279-4_9

Lewis, A. (2016). Towards a harmonized framework for calculating logistics carbon footprint. *Sustainable Logistics and Supply Chains: Innovations and Integral Approaches*, 163-181.

Lewis, A., & Greene, S. (2019, 7). *GLEC Framework for Logistics Emissions Accounting and Reporting* (Tech. Rep.). Retrieved from `https://doi.org/10.46461/glecframework` doi: 10.46461/glecframework

Lewis, J., & Linzer, D. (2005). Estimating regression models in which the dependent variable is based on estimates. *Political Analysis*, *13*(4), 345–364.

Markovsky, I., & Van Huffel, S. (2007). Overview of total least-squares methods. *Signal processing*, *87*(10), 2283–2302.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press.

Neyman, J. (1935). On the problem of confidence intervals. *The Annals of Mathematical Statistics*, *6*, 111-116.

Olive, D. J. (2007). Prediction intervals for regression models. *Computational Statistics & Data Analysis*, *51*(6), 3115–3122.

Piecyk, M. I., & McKinnon, A. C. (2010). Forecasting the carbon footprint of road freight transport in 2020. *International Journal of Production Economics*, *128*(1), 31–42.

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, 145–158.

Radonjič, G., & Tompa, S. (2018). Carbon footprint calculation in telecommunications companies–the importance and relevance of scope 3 greenhouse gases emissions. *Renewable and Sustainable Energy Reviews*, *98*, 361-375.

Royo, B. (2020). Measuring and allocating scope 3 ghg emissions. *Towards User-Centric Transport in Europe 2: Enablers of Inclusive, Seamless and Sustainable Mobility*, 200–211.

Schmied, D. M., Knörr, D. W., & Hepburn, L. (2012). *Calculating ghg emissions for freight forwarding and logistics services in accordance with en 16258.*

Smart Freight Centre, & Cefic. (2021). *Calculating GHG transport and logistics emissions for the European Chemical Industry* (Tech. Rep.).

Smith, J., Brown, E., Johnson, A., Garcia, M., & Lee, D. (2015). Understanding and characterizing uncertainties in real-time chemical measurements of atmospheric trace gases and aerosols: Part i—general concepts. *Atmospheric Measurement Techniques*, *8*(6), 2363–2387.

Snee, R. D. (1977). Validation of regression models: methods and examples. *Technometrics*, *19*(4), 415–428.

Tong, L.-I., Chang, C.-W., Jin, S.-E., & Saminathan, R. (2012). Quantifying uncertainty of emission estimates in national greenhouse gas inventories using bootstrap confidence intervals. *Atmospheric Environment*, *56*, 80-87.

Velázquez-Martínez, J. C., Fransoo, J. C., Blanco, E. E., & Mora-Vargas, J. (2014). The impact of carbon footprinting aggregation on realizing emission reduction targets. *Flexible Services and Manufacturing Journal*, *26*, 196-220.

Vickers, A. J. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology*, *5*(1), 1–12.

Waldman, B., Huang, M., & Simonen, K. (2020). Embodied carbon in construction materials: a framework for quantifying data quality in epds. *Buildings and Cities*, *1*(1), 625–636.

Wang, Z., Chen, L., Park, J., Kim, M., & Patel, S. (2021). Assessing the impact of uncertainty and variability in environmental data using probabilistic models. *Environmental Modeling & Assessment*, *26*(1), 71–85.

Wegener, M., Labelle, R., & Jerman, L. (2019). Unpacking carbon accounting numbers: A study of the commensurability and comparability of corporate greenhouse gas emission disclosures. *Journal of Cleaner Production*, *211*, 652-664.

Wild, P. (2021). Recommendations for a future global co2-calculation standard for transport and logistics. *Transportation Research Part D: Transport and Environment*, *100*, 103024.

WRI, & WBCSD. (2004). The greenhouse gas protocol: a corporate accounting and reporting standard, revised edition.

WRI, & WBCSD. (2011). *Corporate Value Chain (Scope 3) Accounting and Reporting Standard Supplement to the GHG Protocol Corporate Accounting and Reporting Standard GHG Protocol Team* (Tech. Rep.).

Zhou, P., Ang, B., & Poh, K. L. (2006). A trigonometric grey prediction approach to forecasting electricity demand. *Energy*, *31*(14), 2839–2847.

Appendices

A Comparison tools and databases

Table 15 shows that only a limited number of tools are globally applicable. Besides, most tools and databases do not apply to all transportation modes. All transport modes are air, rail, road, sea, and inland waterways. The only tools that incorporate all transport modes are BigMile and EcoTransIT, which are commercial tools that are very expensive. However, these two tools are expensive commercial initiatives and do not incorporate emissions from logistics sites. Therefore, companies using multiple transport modes or operating in different world regions often need to combine tools and databases to estimate GHG emissions. Besides, some tools such as International Maritime Organization (IMO), Smartway, International Air Transport Association (IATA) and Clean Cargo Working Group (CCWG) denote the emissions solely in Tank-to-Wheel. However, this does not include the indirect emissions. It is more convenient to note emissions in both Tank-to-Wheel (TTW) and Well-to-Wheel (WTW). Besides, comparing results in Tank-to-Wheel with results in Well-to-Wheel is impossible. Therefore, it is necessary to convert results to Well-to-Wheel to enable comparison between different methods.

Tools & databases	**Legal base**	**Geographical scope**	**Transport modes**	**Logistics sites**	**WTW**
ARTEMIS	Initiative	Europe	All except air	No	No
EcoTransIT	Commercial	Global	All	Yes	Yes
NTM	Initiative	Europe	Air, rail, road, sea	No	No
DEFRA	Initiative	Europe	Road, rail	No	No
SmartWay	Program	North America	All	No	No
BigMile	Commercial	Global	All	Yes	Yes
IMO	Official	Global	Sea & inland waterways	No	No
IATA	Association	Global	Air	No	No
CCWG	Initiative	Global	Sea	No	No
Guide for GHG Assessment for Logistics Sites	Initiative	Europe	Logistics sites	Yes	Yes
Guidance for GHG Emission Footprinting for Container Terminals	Research	Europe	Logistics sites	Yes	Yes

Table 15: Overview of most important tools and databases

Furthermore, none of these calculation methods include confidence intervals, which makes it very difficult to compare LSPs using different input data. Therefore, the calculations often over- or underestimate the emissions. Moreover, many tools are commercial and not openly accessible.

B Case study data

Shipment No	From City	From Country	From Date	To City	To Country	To Date	Spot / Dedicated	Modality	Transported Weight (t)	CO2 WTW (t)	Distance (km)	Cleanings	Heatings
1234567	GENT	BE	18/08/2022	Arteixo	ES	30/08/2022	Spot	Intermodal	24.94	1.007	2,630	1	0
1234568	GENT	BE	19/01/2022	Kallo	BE	14/01/2022	Dedicated	Road	24.48	0.147	140	0	1

Table 16: LSP 1 - data example per shipment from LSP

Load City	Unload City	Average Loaded Distance	Modality	Spot / Dedicated	WTW emission intensity (g per tonne-km)
GENT	ARTEIXO	2104.16	Intermodal	Spot	18.58
GENT	KALLO	138.60	Road	Dedicated	48.26

Table 17: LSP 1 - data example per lane per contract type (spot/dedicated) from LSP

Order_Lane	**Pick_Up Country**	**Pick_Up Postal_Code**	**Delivery Country**	**Delivery Postal_Code**	**Trip Year**	**Trip Month**	**Main mode**	**Shipment Profile**	**Count of Ordernumber**	**Ton**	**Sum of TonKm**	**CO2e TTW (Ton)**	**CO2e WTW (Ton)**	**CO2e Intensity TTW (g/tonkm)**	**CO2e Intensity WTW (g/tonkm)**
BE20-BE23	BE	2040	BE	2300	2022	1	Truck	FTL	5	105.666	5,136	0.17	0.22	33.52	42.41092
BE20-BE23	BE	2040	BE	2300	2022	5	Truck	LTL	1	11.282	2,458	0.10	0.13	41.22	52.1523

Table 18: LSP 2 - data example from LSP

C Transformations

								Test of normality		
LSP	Modality	Parameter	Data points	Mean	Standard deviation	Skewness	Kurtosis	Statistic	Significance	Transformation	Significance after remedy
1	Road	Great Circle Distance	27	246.12	228.33	0.78	2.02	0.8	1.46E-04	Power of 0.2	0.012
1	Road	Planned Distance	27	306.67	282.39	0.83	2.17	0.81	2.32E-04	Power of 0.2	0.015
1	Road	Loaded Distance	27	334.34	293.56	0.74	2.08	0.84	8.18E-04	Logarithmic	0.033
1	Road	Empty Distance	27	279.93	281.94	1.59	4.55	0.78	6.46E-05	Power of 0.3	0.337
1	Road	Total Distance	27	614.27	510.83	1.32	3.82	0.83	4.93E-04	Power of 0.2	0.734
1	Road	Weight	27	24.07	1.91	1.3	5.9			Dummy	
1	Road	Emission intensity	27	93.16	30.94	0.55	2.73	0.95	1.86E-01	Logarithmic	0.471
1	Road	Dedicated	27	0.56	0.51	-0.22	1.05			Dummy	
1	Road	Cleanings	27	0.29	0.44	0.95	1.97			Dummy	
1	Road	Heatings	27	0	0	NA	NA			Exclude	
1	Intermodal	Great Circle Distance	22	790.73	661.23	2.66	11.04	0.7	2.14E-05	Power of 0.2	0.133
1	Intermodal	Planned Distance	22	1036.58	842.91	2.3	9.24	0.76	1.17E-04	Power of 0.2	0.270
1	Intermodal	Loaded Distance	22	1239.40	999.83	2.17	8.04	0.77	1.53E-04	Logarithmic	0.481
1	Intermodal	Empty Distance	22	556.51	429.56	1	2.93	0.87	8.20E-03	Power of 0.3	0.401
1	Intermodal	Total Distance	22	1795.9	1275.88	1.76	5.79	0.81	6.07E-04	Power of 0.2	0.566
1	Intermodal	Weight	22	23.14	3.03	0.15	1.92	0.94	0.165	None	0.165
1	Intermodal	Emission intensity	22	29.84	22.28	1.68	5.16	0.8	4.38E-04	Logarithmic	0.831
1	Intermodal	Dedicated	22	0.23	0.43	1.3	2.69			Dummy	
1	Intermodal	Cleanings	22	0.81	0.39	-1.48	3.44			Dummy	
1	Intermodal	Heatings	22	0.05	0.2	4.33	19.82			Dummy	
2	Road	Distance	120	702.81	551.96	1.22	4.53	0.9	1.26E-07	Square root	0.201
2	Road	Weight	120	12.29	8.3	0.01	1.46	0.9	2.31E-07	Exclude	-
2	Road	Tonne-kilometer	120	8869.1	10671.58	1.93	7.21	0.77	1.72E-12	Power of 0.2	0.300
2	Road	FTL	120	0.47	0.5	0.13	1.02			Dummy	
2	Road	LTL	120	0.4	0.49	0.41	1.17			Dummy	
2	Road	Groupage	120	0.13	0.34	2.16	2.65			Dummy	
2	Road	Emission intensity	120	64.2	32.17	1.38	5.06	0.89	4.74E-08	Logarithmic	0.213

Table 19: Descriptive statistics, final transformations and Shapiro-Wilk test

D Stepwise regression

	Loaded distance (log)	Empty distance $^{0.3}$	GCD $^{0.2}$	Planned distance $^{0.2}$	Dedicated	Cleaning	Tank container	Emission intensity (log)
Loaded distance (log)	1.000	0.437	0.962	0.959	0.224	0.279	-0.349	-0.463
Empty distance $^{0.3}$	0.437	1.000	0.400	0.403	0.281	0.117	-0.511	0.507
GCD $^{0.2}$	0.962	0.400	1.000	0.996	0.146	0.263	-0.349	-0.439
Planned distance $^{0.2}$	0.959	0.403	0.996	1.000	0.158	0.244	-0.374	-0.426
Dedicated	0.224	0.281	0.146	0.158	1.000	-0.223	-0.373	0.206
Cleaning	0.279	0.117	0.263	0.244	-0.223	1.000	0.279	-0.109
Tank container	-0.349	-0.511	-0.349	-0.374	-0.373	0.279	1.000	-0.199
Emission intensity (log)	-0.463	0.507	-0.439	-0.426	0.206	-0.109	-0.199	1.000

Table 20: Covariance matrix LSP 1 road transportation

	Heatings	Loaded distance (log)	Empty distance (0.3)	GCD (^0.2)	Planned distance (0.2)	Dedicated	Cleaning	Weight	Emission intensity (log)
Heatings	1.000	0.075	-0.065	0.078	0.021	-0.127	0.176	0.343	0.017
Loaded distance (log)	0.075	1.000	0.439	0.968	0.962	-0.652	0.507	0.083	-0.779
Empty distance (0.3)	-0.065	0.439	1.000	0.392	0.403	0.025	0.201	-0.046	-0.103
GCD (^0.2)	0.078	0.968	0.392	1.000	0.992	-0.630	0.491	0.143	-0.762
Planned distance (0.2)	0.021	0.962	0.403	0.992	1.000	-0.616	0.536	0.124	-0.755
Dedicated	-0.127	-0.652	0.025	-0.630	-0.616	1.000	-0.627	0.035	0.596
Cleaning	0.176	0.507	0.201	0.491	0.536	-0.627	1.000	-0.072	-0.209
Weight	0.343	0.083	-0.046	0.143	0.124	0.035	-0.072	1.000	-0.289
Emission intensity (log)	0.017	-0.779	-0.103	-0.762	-0.755	0.596	-0.209	-0.289	1.000

Table 21: Covariance matrix LSP 1 intermodal transportation

	Tonne-km $^{0.2}$	Distance $^{0.5}$	FTL	LTL	Groupage	Emission intensity (log)
Tonne-km $^{0.2}$	1.000	0.705	0.559	-0.165	-0.583	-0.522
Distance $^{0.5}$	0.705	1.000	0.084	0.011	-0.140	-0.051
FTL	0.559	0.084	1.000	-0.764	-0.367	-0.603
LTL	-0.165	0.011	-0.764	1.000	-0.320	0.336
Groupage	-0.583	-0.140	-0.367	-0.320	1.000	0.401
Emission intensity (log)	-0.522	-0.051	-0.603	0.336	0.401	1.000

Table 22: Covariance matrix LSP 2 road transportation

E LSP 2 road scatterplot distance

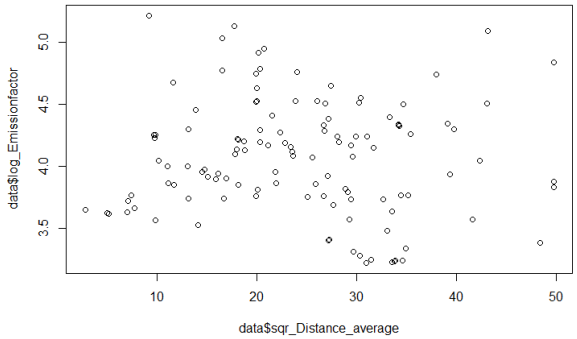

Figure 8: Scatterplot distance of LSP 2

F Prediction intervals - LSP 1

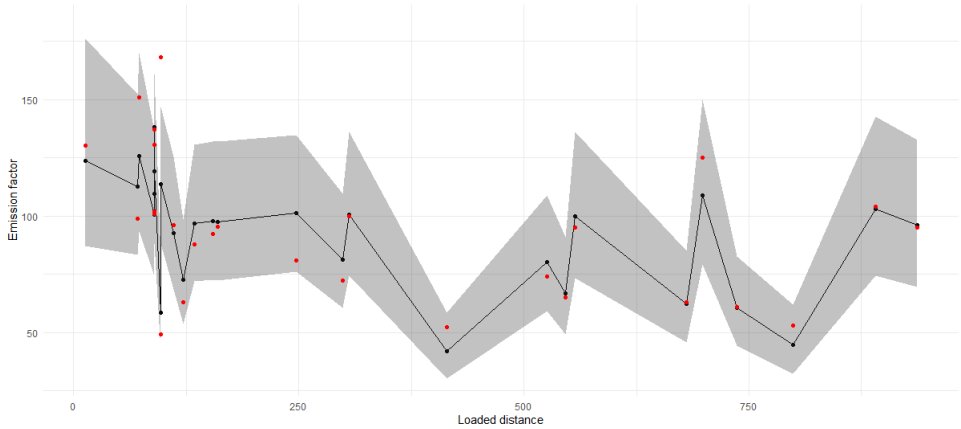

Figure 9: Out-of-sample prediction interval of LSP 1 with loaded distance for road transportation

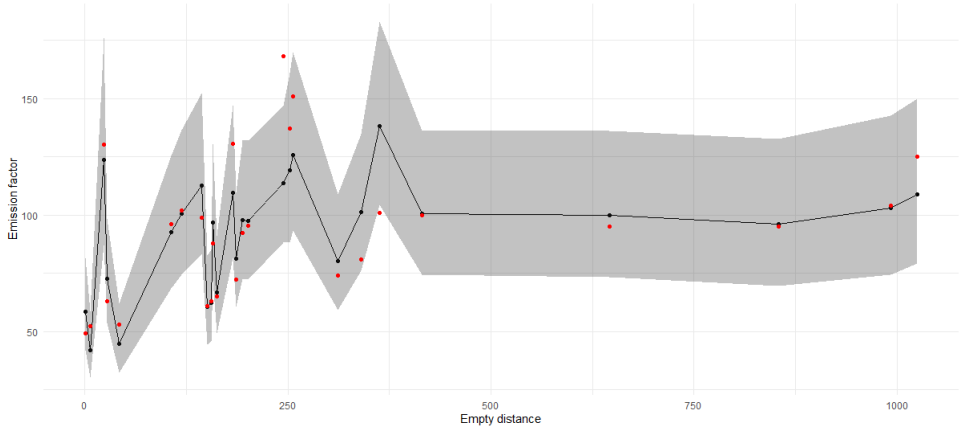

Figure 10: Out-of-sample prediction interval of LSP 1 with empty distance for road transportation

Figure 9 shows the prediction intervals for LSP 1 and the corresponding loaded distance. Figure 10 shows the empty distance and the prediction intervals. The red dots represent the left-out value of the Leave-One-Out Cross-Validation (LOOCV). Both figures have a zigzag pattern. Since loaded distance negatively and empty distance positively influence emission intensity, there is a trade-off. For example, if the loaded distance is very high and the empty distance is very low, the emission intensity will be very low. Thus, the loaded distance on its own or the empty distance on its own does not say much.

G Confidence intervals - LSP 1

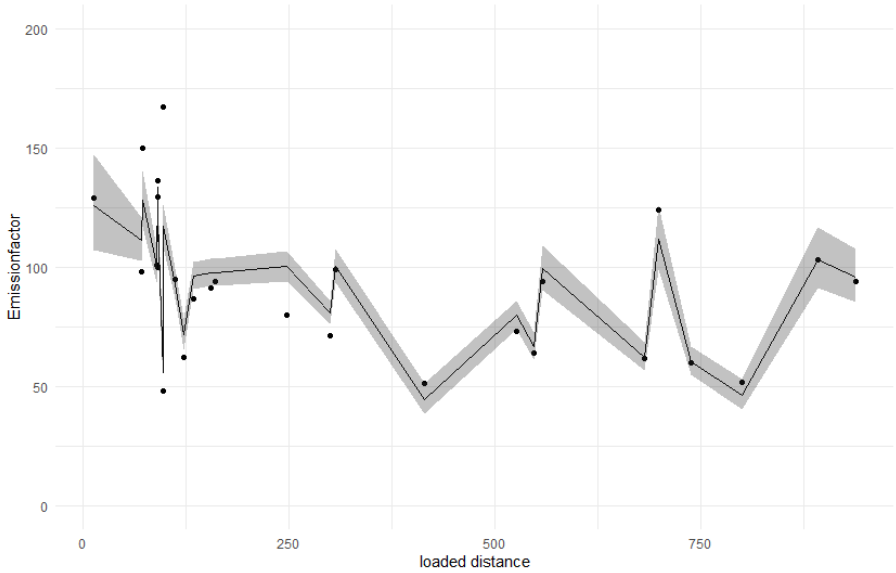

Figure 11: LSP 1 Road - Confidence interval of the mean response (varying empty distance)

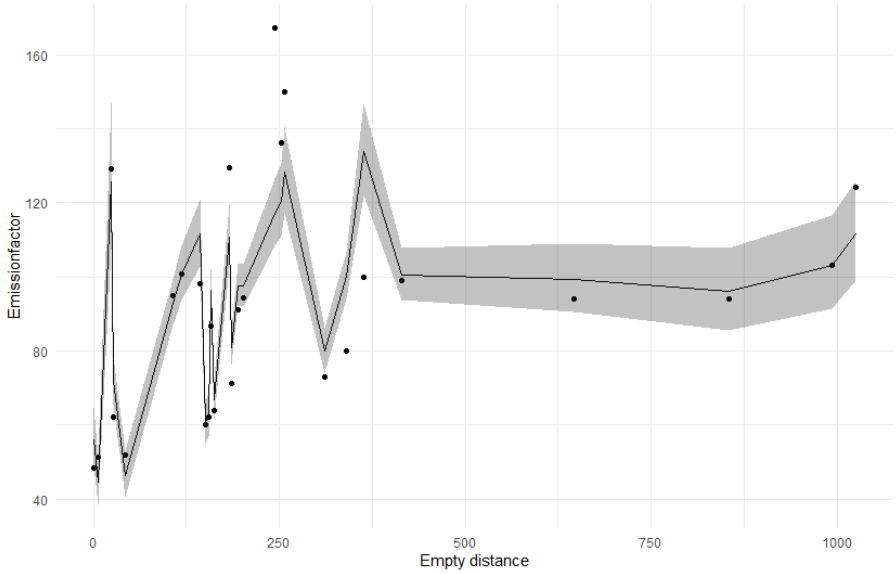

Figure 12: LSP 1 Road - Confidence interval of the mean response (varying loaded distance)