Erasmus University Rotterdam

Erasmus School of Economics

Business Analytics and Quantitative Marketing

Master Thesis

A Constrained Clustering Algorithm to Examine the Effects of Consolidation on the CO_2 Emissions in the Logistics Sector

Quinty Vishnudatt

503622

Supervisor: Dr. Oguzhan Vicil

Second Assessor: Prof. Dr. Ir. R. Dekker

October 2022

Abstract

In recent years, the detrimental effects of CO_2 emissions on the global climate have become more important, and strategies have been put in place to reduce these emissions. A high leverage strategy that can be used in the transport sector is freight consolidation, which can be accomplished by clustering shipments into vehicles, resulting in lower CO_2 emissions. In this paper, we will construct a constrained clustering algorithm based on the k-means algorithm, the MCF network and PC k-means, where the constraints are based on time and vehicle capacity. The goal of the algorithm is to consolidate shipments into clusters, to examine the effect of this consolidation on the CO_2 emissions. The dataset is provided by SSC[1], which is an organization that calculates the CO_2 emission of shipment routes for different companies. The results indicate that consolidation is possible and leads to a decrease in CO_2 emissions and a better utilization of the vehicle capacities. In the original situation, 14,979 vehicles were used, with an average utilization rate of 13.1% of the vehicle capacities. The number of vehicles has been decreased to 9,894 as a result of the consolidation employed in this research, while the average vehicle capacity utilization has increased to 41.8%. The combined network's total CO_2 emissions are lower by 32.6% at 2,323,806.88 kg compared to the old network's total of 3,447,679.09 kg.

[1] https://sustainingsupplychains.com/nl/home/

1 Introduction

The negative impact of greenhouse gas (GHG) emissions on the global climate have been a present topic over recent years and concrete plans have been implemented to reduce these emissions. Any gas that can absorb infrared radiation emitted from the surface of the Earth and reradiating it to the surface, hence enhancing the greenhouse effect, can be considered a greenhouse gas. They are measured in 'carbon dioxide-equivalents' (CO_2e), which attempts to convert the global warming impact of the different GHGs together into a single measure. Today, we produce approximately 50 billion tonnes of CO2 annually, which is 40% more than in 1990 (Ritchie & Roser, 2020). Carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) are the most important greenhouse gases (Jeffry et al., 2021). During this research, we will focus on the most dominant GHG, namely CO_2. Ritchie and Roser (2020) show that 74.4% of the total GHG emissions measured in CO_2e are due to CO_2.

Several climate change problems involve mitigation (reducing emissions) and adaptation (preparing for unavoidable consequences) (Rolnick et al., 2022). We will focus on the mitigation of GHG emissions, which involves changes to transportation, buildings, industry, and more. The variety of problems within these sectors can be seen as an opportunity: there are many ways to have a significant impact (Change et al., 2014). Several of these high-impact problems that have been identified regarding mitigation, can be resolved by machine learning (ML), through either engineering or innovative research. The optimal way to use the tools to combat climate change still needs to be determined through a focused effort. Many professionals that are familiar with ML want to take action, but are unsure how (Rolnick et al., 2022). High-impact problems where ML can be applied include for example modeling demand, where ML can provide information about mobility patterns, and the use of electric vehicles (Bektaş et al., 2019), where ML methods can help with battery energy management (Ali & Söffker, 2018; Hansen & Wang, 2005) and modeling charging behavior (Wang et al., 2019). In this paper, we will focus on ML applications within the transport sector.

Transport is fundamental to our economy and society and will continue to have an increasing importance (Kelle et al., 2019). Globally, the transport sector accounts for 21% of CO_2 emissions (IPCC, 2018; Ritchie, 2020) and it has not made a significant progress to lower this (Bektaş et al., 2019; Creutzig et al., 2015). By 2050, it is predicted that the world's transportation needs would have tripled, resulting in a doubling of CO2 emissions (Forum, 2019; Greene & Lewis, 2019). Due to the relationship between GHGs and the quantity of vehicle movement and, consequently, fuel consumption, emissions have increased. Travel can be done by road, rail, water, or air and these different transportation modes have different carbon emissions. According to the International Transport Forum (ITF) (Forum, 2019), freight transport accounts for about 39% of transport CO_2 emissions and around 8% of CO_2 emissions worldwide. Road travel is responsible for 62% of CO_2 emissions, making it the most dominant mode of freight transport. Sea contributes 27%, rail 3%, and inland waterways 2% to the total transport CO_2 emissions (Schellnhuber et al., 2018). Air travel accounts for only 6% of the total transport CO_2 emissions, while it usually gets the most attention in plans to tackle climate change. Despite efforts made over time to increase the effectiveness of supply chain logistics operations, freight transportation still has a negative impact on the environment (Bektaş et al., 2019).

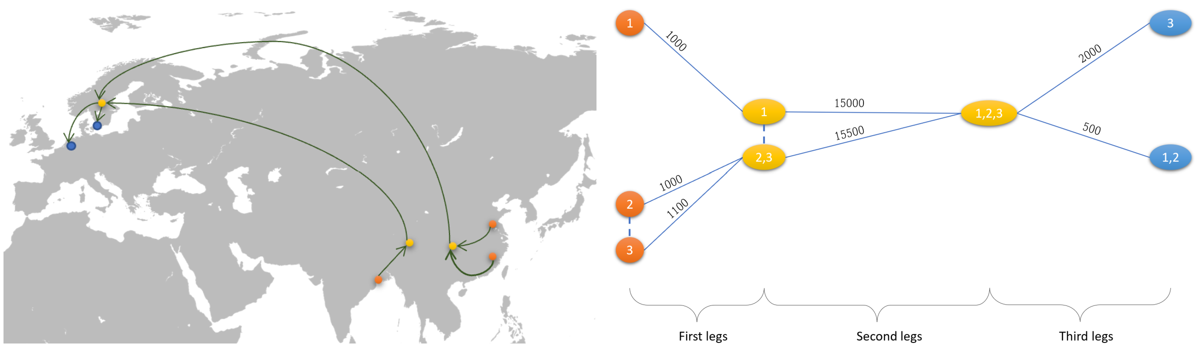
Reducing transport activity by bundling shipment routes, also referred to as freight consolidation, can help with reducing emissions (McKinnon, 2018). Consolidated shipments work by combining partial loads from one or several shippers into a full, or almost full, truckload going to common destinations. Combining numerous minor shipments allows for the dispatch of a bigger, more cost-effective load on the same vehicle. The goal is to better utilize a vehicle's capacity, which can lead to lower costs, and thus increased efficiency. Freight consolidation can reduce the number of shipments, which in turn can reduce the CO_2 emissions. This emission reduction can lower a company's carbon footprint and additionally their operating expenses (Vaillancourt, 2016) as well. It is significantly cheaper to ship consolidated loads because companies only pay for the space they need. Another cost advantage is flexibility. Companies can often allow for a larger pickup window, allowing carriers to pick up the freight as soon as possible, rather than making them wait until a specified time. Apart from the benefits that can be gained for the company, CO_2 reduction also benefits the impact of global warming overall. For smaller shippers who don't typically carry whole truckloads of freight, or businesses that send smaller batches of freight more frequently, consolidation is ideal. It increases transport efficiency and encourages economies of scale. Rolnick et al. (2022) identify freight consolidation as a bottleneck that domain experts have identified in climate change mitigation. Additionally, they state that research regarding freight consolidation is particularly well-suited to tools from ML.

The dataset is provided by Sustaining Supply Chains (SSC), which is a small organization that calculates the carbon dioxide (CO_2) emission of shipment routes for different companies. These routes, also referred to as lanes, are predetermined and known and can be divided into one or multiple legs. A leg can be seen as part of a lane that is covered by one transportation mode (rail/road/air/ocean/parcel). Each leg has its own sub-origin and sub-destination. Origins and destinations belong to the final lanes and can be seen as the definite beginning and end points, where the sub-origins and sub-destinations represent the beginning and end points of each leg. The calculation for the CO_2 emission is mostly based on the distance between the origin and destination points, the weight of the shipment, and the transportation mode, but more factors influence the CO_2. With these factors, they calculate the emission of each leg. The CO_2 emission of the total shipment is then determined by summing the emissions for each leg that is part of the lane over which a shipment is transported. Determining the emission has helped their clients give insights into their CO_2 footprint. However, the emission calculation does not tell them which actions they could take to reduce this footprint.

In the current situation, the transportation modes and corresponding capacities are not fully utilized. Our goal is to increase the vehicle load and thereby improve the utilization of the transportation modes by consolidating shipments within the fashion, textile, and sports (FTS) sector while taking capacity and time constraints into account. The network with lanes and legs has already been determined by the existing decision support system of the company. This network is a feasible schedule from which the CO_2 emission can be calculated, where shipments are picked up by third-party logistics and transported to the next (sub-)destination. For every leg in the current dataset, each shipment is transported in a separate vehicle. We will not be forming new routes, since the existing routes already provide a feasible schedule. Instead, we will use this feasible schedule to examine the possibilities of freight consolidation, going from one feasible solution to another feasible solution.

SSC has stated that its clients will benefit from freight consolidation. They pointed out that, especially in the fashion industry, shipment vehicles and containers are not optimally loaded. However, there are multiple limitations related to consolidation which justify why freight consolidation within the company has not been applied yet, even for shipments that are transported over the same leg. Several important reasons for this are that shipments may be too large to consolidate or the pickup and/or delivery times of these shipments are too far apart. Another reason for this is the complexity of creating a properly working network of customers, carriers, processing centers, and lanes. Coordinating these complex systems can be expensive for companies, since they may have to dedicate specific resources. Besides this, there are more limitations regarding freight consolidation, which will be discussed in Section 4.2.2.

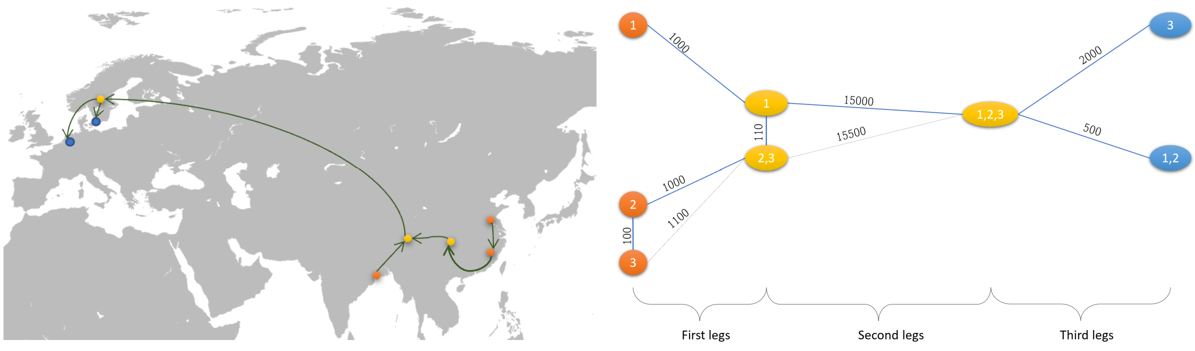
A simple example of the current approach is given in Figure 1, where Figure 1b shows the legs on the world map and Figure 1a shows a graphical representation. This network contains six unique lanes with three shipments that are transported over different legs. For simplicity, we assume that each shipment has a weight of 100 kg. The orange circles represent the origin points, the yellow ovals represent the sub-destinations and the blue ovals represent the final destinations. Next, the blue lines represent existing legs, while the dashed blue lines are non-existing legs. The corresponding values represent the distances in kilometers. The first legs in this example all are unique in the sense that every shipment is transported over a different leg. Shipments 2 and 3 share the second leg, while shipment 1 has a separate leg. However, all shipments go to the same sub-destination. Shipments 1 and 2 go to the same final destination, while shipment 3 has a different final destination. In the current situation, shipments 1, 2 and 3 will all be carried in separate vehicles for each leg, leading to 3 (vehicles) $\times$ 3 (legs) $=$ 9 vehicles in total. For simplicity, we assume that the CO_2 emissions can be calculated by adding the distance and shipment weight. The CO_2 emission of these shipments will then be $(1000 + 1000 + 1100) + 300 + (15000 + 15500\times2\ ) + 300 + (2000 + 500\times2) + 300 = 53{,}000$ kg CO_2e. This computation first sums the distances of one leg and then adds the shipment weight, which is equal to 300 since each shipment is 100 kg.



(a) (b)

Figure 1: Example of a network with three shipments

With our freight consolidation approach, we want to consolidate the shipments that are geographically close and go to the same (sub-)destination. In this specific example, shipments 2 and 3 can be clustered for the first leg, shipments 1, 2 and 3 can be clustered for the second leg, and shipments 1 and 2 can be clustered for the third leg, leading to 5 vehicles in total. However, this form of consolidation is only possible when the non-existing legs (the blue dashed lines) become existing legs. We have not created these new legs in our algorithm, but propose them as a suggestion to add to the existing network for a more efficient consolidation process. Nonetheless, consolidation is still possible if we only use the existing legs, which in that case gives us a total of 7 vehicles used with a CO_2 emission of (100 + 1000 + 1000) + 300 + (110 + 15000) + 300 + (2000 + 500) + 300 = 20,610 kg CO_2e. This is a toy example and should be treated as such since capacities and time windows are not included. Figure 2 shows the consolidated network. The dashed lines are now solid and the legs that are not used are shown in grey. After consolidation, we always choose the leg with the shortest distance, since the calculation of the CO_2 emissions is largely dependent on the distance. An additional feature of our approach which is not included in the Figure is that we will make use of flexible vehicle and container types within a transportation mode. This means that shipments do not have to be transported in the same vehicle/container that they were originally transported in according to the dataset. This allows for a more flexible consolidation process since the corresponding maximum capacities of these vehicles and containers are not assigned to specific clusters.



(a) (b)

Figure 2: Example of a network with three shipments

In this paper, we will construct a constrained clustering algorithm to answer the research question:

"Is freight consolidation possible for the current network of the company while taking time and vehicle capacity constraints into account?",

and analyze the effects of freight consolidation on the CO_2 emission reduction. Additionally, we will examine the effects of consolidation on the vehicle utilization, which we define as the percentage of the vehicle capacity that is used. When vehicles are loaded properly, the environmental impact can be significantly decreased (McKinnon & Edwards, 2010), which is why we decided to focus on this aspect as well. Freight consolidation via cluster analysis has not been intensively researched, especially in the sustainability context (Rolnick et al., 2022; Van Andel, 2018). To apply freight consolidation, we will construct a constrained clustering algorithm based on k-means. Cluster analysis is an unsupervised learning method that is used to classify objects with similar characteristics into subgroups (Tryon, 1939). Constrained clustering, a form of semi-supervised learning, was developed to extend clustering algorithms to incorporate existing domain knowledge in the form of labeled data or constraint sets (Wagstaff, 2010). The algorithm in this paper makes use of instance-level constraints in the form of time constraints and feature-level constraints in the form of cluster size constraints. The time constraints are incorporated through cannot-link (CL) constraints (Švehla, 2018) and are dependent on the maximum lead times. The constraints regarding the cluster sizes correspond with the maximum capacities of the vehicles that carry the shipments. These are incorporated in the Minimum Cost Flow (MCF) linear network optimization problem (Levy-Kramer & Klaber, 2022), which also performs the cluster assignment. The goal of the MCF network is to find the 'cheapest' possible way of sending an amount of flow through a network, where the edges of the network can have certain capacities (Sifaleras, 2016). Because of these features, the MCF network is a suitable algorithm that can be applied to examine the research question. The predefined shipments weights can be seen as the amounts of flow that is send through the network, and are distributed over the vehicles. The 'cheapest' possible way is in our case related to minimizing the CO_2 emissions, while taking the vehicle capacities into account by assigning these capacities to the edges of the network. The consolidation process allows for flexible shipment allocation with flexible vehicle types per transportation mode. An important note is that our approach can be used on top of the existing system with existing data, which means that it does not substitute but rather *complement* the existing decision support system.

To obtain the results, we divide the dataset manually into 46 unique subsets, and apply the constrained clustering algorithm to each of these subsets. After applying the constrained clustering algorithm to the subsets, our findings are as follows. The first results show that consolidation is possible with the formulated time and vehicle capacity constraints. The results show that clustering is possible for 40 of these subsets, which implies that consolidation is possible for these subsets. In the original situation, 14,979 vehicles were used to transport the shipments to their final destinations, where 13.1% of the vehicle capacities were utilized on average. After the consolidation approach used in this paper, we can conclude that the number of vehicles has been reduced to 9,894, while the average vehicle capacity utilization has increased to 41.8%. The total CO_2 emission of the original network is equal to 3,447,679.09 kg, while the total emissions of the consolidated network are 2,323,806.88 kg, which is a reduction of 32.6%. These results show that consolidation has improved both the CO_2 emissions and the utilization of the vehicle capacities.

Note there are several important limitations in this study. First, we will not take costs in terms of expenses into account and will therefore mainly focus on the sustainable outlook. The dataset provided by SSC does not contain any information about expenses, which is why we cannot include costs into our approach. This means we cannot make a trade-off between operating expenses and CO_2 emissions, and examine the practical possibility. The results we obtain regarding the consolidation are therefore compared to the original situation, which is shipment transportation without consolidation. Furthermore, the CO_2 emission calculations do not take the vehicle weight or other aspects of the vehicle into account during the emission calculations, according to the GLEC framework (Greene & Lewis, 2019). This means that even though consolidation results in a sizeable reduction in the number of vehicles needed for transport, it does not automatically lead to an emission reduction. Especially the CO_2 emission calculations of the transportation modes parcel and rail are mostly dependent on the distance and shipment weight. This results in almost no emission reduction for these transportation modes, even though consolidation takes place. The transportation modes air, ocean, and road do show a substantial reduction in CO_2

emissions, whereas road shows the most significant reduction. These modes incorporate more factors besides the distance and shipment weight (Greene & Lewis, 2019), which explains this result.

The remaining part of this paper is organized as follows. Section 2 describes the previous literature related to our research, while the data is described in Section 3. The methodology is stated in Section 4. Section 5 presents the results of the constrained clustering algorithm. Lastly, a conclusion is given in Section 6. Furthermore, we included an Appendix for additional tables and figures.

2 Literature

2.1 Sustainability

From the viewpoint of both the research community and practitioners, interest in the subject of environmental sustainability is growing (Marchet et al., 2014). Increased environmental concerns are the main reason for this growing significance, but other important elements like governmental restrictions and the creation of global certification standards have prompted businesses to carefully consider sustainability projects (Eltayeb et al., 2011). Unfortunately, the sustainability outlook is often foreshadowed by economic interest as the primary requirement (Marchet et al., 2014). Limitations related to costs, lead times, and the complexity of systems have led to a trade-off between decarbonization and economic interest. Recent work has investigated this trade-off and has come up with various suggestions to approach this. Rolnick et al. (2022) describe how ML can be a powerful tool in reducing GHG emissions and provide an overview of high-impact problems in various domains, such as electricity (Mosavi et al., 2019), industrial production (Tsoumakas, 2019) and the transport sector (McKinnon, 2018). In this section, we will specifically focus on high-impact problems within the logistics and transport sector.

To analyze these high-impact problems, the CO_2 emissions need to be modeled. There exists a variety of emission models that differ in their estimation approach or differ with regard to the parameters they take into account during the estimations (Demir et al., 2011). Calculating and reporting emissions is important, but a relatively new step in decarbonization. To obtain a simple approach for determining the global warming impact, the Global Logistics Emissions Council (GLEC) was formed to develop the GLEC Framework (Greene & Lewis, 2019). This framework offers a starting point rather than a one-size-fits-all solution, outlining emission limits, base methodology that may be used, reporting process considerations, and guidance on how to produce the best results using the information at hand. As described in Section 1, the majority of GHG emissions for logistic activities consists of CO_2, making CO_2e the most suitable unit to represent the global warming impact. The calculation and reporting of emissions should be a priority for not only businesses and organizations, but also for countries (Punte et al., 2019). It is the first step toward decarbonization.

There are multiple strategies to reduce the transport emissions, such as shifting freight from road to rail, improving vehicle utilization, and switching transport operations to renewable energy (McKinnon, 2018). Numerous measures to address the damaging effects of logistics on the environment have fallen short because they couldn't simultaneously fulfill the goals of stakeholders including transport companies, the government, and consumers (Muñoz-Villamizar et al., 2018). To account for this problem, eco-innovation is positioned as a target for organizations to be more sustainable while satisfying their stakeholders (Garcıa-Granero et al., 2018). Eco-innovation provides both environmental and economic benefits and has become a means for firms to gain a competitive advantage (Cai & Li, 2018). However, little is still understood about the difficulties adopting eco-innovation to lessen the negative environmental effects of freight transport, which is partially caused by ignorance of difficulties and restrictions (Orji et al., 2019).

2.2 Freight consolidation

Despite the fact that consolidation occurs in operations, there is little to no literature on the practice of consolidation in logistics (Vaillancourt, 2016). Schulz and Blecken (2010) describe consolidation as a side effect of horizontal cooperation but recognize it as an area of future focus because of its benefits, such as cost and time reductions. Horizontal cooperation is defined by Union (2001) as concerted practices between companies operating at the same level(s) in the market. It has received more attention recently and has been regarded as a successful method for sustainable logistics and freight movement according to Pan et al. (2019). They provide a survey of the development of horizontal collaborative transport, where they provide guidelines to logistics stakeholders who wish to embark on this. In addition to this, they recognize that accurate and comprehensive logistics metrics concerning sustainability are needed. Zhou et al. (2011) consider two collaboration modes for freight consolidation: strategic alliance and full collaboration, where a strategic alliance is described as an agreement on collaboration among firms and full collaboration implies that the firms operate as an independent unit. They provide a fundamental framework for investigating cooperative strategies that businesses may use to compete with other practitioners of freight consolidation.

In this paper, we will investigate the possibilities of freight consolidation by further focusing on improving vehicle utilization concerning the capacity. McKinnon (2018) and Punte et al. (2019) identify consolidation as one of the solutions for decarbonizing the logistics sector, especially by optimizing vehicle loading. According to McKinnon (2018) between a fifth and a third of truck-kms are run empty. However, several projects such as the Empty Miles program (MILES, 2009) have been launched to share unused transportation capacity and reduce empty-trip inefficiencies. Thanks to this initiative, 61.65 tons of CO_2 have been eliminated and annualized transportation costs have on average been reduced by 25.000 dollars.

Furthermore, Punte et al. (2019) describe how stakeholder collaboration is important to achieve vehicle utilization in freight consolidation. They state that all stakeholders should be aware of the design and implementation of solutions. Moreover, they define multi-modal optimization as an additional solution to achieve the decarbonizing goal. Proper mode selection and switching for freight transportation can increase productivity, dependability, flexibility, and sustainability. SteadieSeifi et al. (2014) give an overview of previous research regarding multi-modal freight transportation planning and state that many challenges remain. Kelle et al. (2019) overcomes one of these challenges by building a simulation model that shows the impact of transportation mode change. However, their data is limited to one region, making it difficult to extend their approach to other regions. Since the shipment data used in this paper consists of lanes, where one leg is covered by one mode, we are content with this solution.

2.3 Cluster analysis

Globally, the number of logistics clusters is growing as a result of successful models (Sheffi, 2012). These clusters are significantly important for companies that use and provide freight transportation services. There are various ways in which clustering can be done, such as clustering based on companies (Sheffi, 2013), geographical clustering (Francis et al., 2008) or demand clustering based on lanes (Mesa-Arango & Ukkusuri, 2015). Sheffi (2013) describes how clusters consisting of companies include mainly three types of companies: logistics services providers, businesses with logistics-intensive operations, and businesses with industrial firms' logistics operations, like the distribution operations of retailers and aftermarket parts suppliers. They summarize the main operational advantages achieved by carriers, which include sharing of resources and costs. Schiele (2008) conforms to this clustering view and illustrates the relevance of clusters for the strategic management of firms. They define a cluster as a group of firms and institutions of one industrial sector that are complementing each other along a value chain and state that within a cluster, forms of cooperation, as well as productivity and innovation advantages, are possible.

By breaking down the initial network into smaller groups of geographically-close nodes, geographic clustering has mostly been utilized to simplify the computations involved in solving vehicle routing difficulties (Francis et al., 2008). For the purpose of coordinating vehicle routing in massive post-disaster distribution and evacuation operations, Özdamar and Demir (2012) suggest a hierarchical cluster and route procedure. Their procedure consists of a multi-level clustering algorithm that clusters demand nodes, intending to minimize the total travel time of vehicles and promote efficient resource utilization. Dondo and Cerdá (2007) present a three-phase heuristic/algorithmic approach for the vehicle routing problem with time windows, where they use a heuristic-based clustering algorithm from which they obtain near-optimal solutions for a significant number of problems. Cortes and Suzuki (2020) introduce a model that combines the vehicle routing problem with shipment consolidation, which considers shipment exchanges between different vehicles at certain customer locations, also known as mid-route shipment consolidation. They show that up to 10% savings may be attained through their approach, but only for trucking companies and private carriers.

Geographical clustering has also been used within a sustainability framework. Kagawa et al. (2015) identify the growth in CO_2 emissions within supply-chain clusters using a geographical clustering model. Their research offers insights into where climate policies can be effectively directed by identifying the dominant CO_2 emission clusters in global supply chains. In order to decrease the total CO2 emissions across groups and maximize the total CO2 emissions within groups, the hard clustering approach is utilized to find exhaustive and mutually exclusive clusters. Zhang et al. (2016) evaluate the industrial CO_2 emission efficiency, emission reduction potential, and profits brought by emission reduction for 30 provinces in China by using geographical clustering. They adopt a modified Data Envelopment Analysis (DEA) window approach proposed by Leleu (2013) to evaluate the carbon emission performance and reduction potential.

For strategic research, decision-making, and business improvement, demand clustering in freight logistics networks is crucial, particularly for truckload (TL) companies. But discovering these lanes is not a simple task. This lane bundling problem is a hard combinatorial problem, which requires the computation of several NP-hard sub problems (Mesa-Arango & Ukkusuri, 2015). Bidding advisory models have been created to bundle lanes in TL combinatorial auctions in order to examine this issue (Kuyzu et al., 2015; Triki et al., 2014; Xu & Huang, 2014). The idea behind this is that fluctuations in shipment volume may cause shippers to seek the services of other carriers for shorter periods of time. Then, if they anticipate insufficient capacity, carriers can supplement contractual customer lanes or subcontract lanes. This way, multiple shipments can be bundled onto one lane, leading to economies of scope (Caplice, 1996).

Mesa-Arango and Ukkusuri (2015) investigate demand clustering in freight logistics networks, where they present a novel approach to detect and cluster cooperative lanes on demand, i.e. lanes that minimize empty trips when operated together from a TL perspective. They show that geographical location is not the only feature to consider while clustering, motivating why more features should be included. Van Andel (2018) examines the performance of partitioning and hierarchical clustering on freight consolidation, but not in a sustainability framework. They mostly focus on the cost advantage and evaluate how different time windows affect these costs, where they only use geographical information during their clustering. They use partitioning based and hierarchical clustering, which are commonly used and easy algorithms, but they do have some drawbacks. Partitioning based methods are relatively scalable and simple to implement but perform poorly in high dimensional cases or if the clusters are not well-separated. Hierarchical based clustering methods are flexible but are expensive for high dimensional datasets. Partitioning and hierarchical clustering are also known as hard clustering methods, where an observation can only belong to one cluster, making each cluster distinct. In this paper we will use partitioning based clustering, specifically a constrained k-means algorithm, to examine the feasibility of freight consolidation. The dataset used for clustering will only have two dimensions, which are well separated, making partitioning based clustering a suitable candidate.

3 Data

The shipment data provided by SCC is available in Excel format, see Appendix A. Each shipment route, also referred to as a lane, is divided into one or multiple legs, where each leg is covered by one transportation mode. Thus, a leg can be seen as part of a lane executed by one specific transportation mode, with its own (sub-)origin and (sub-)destination. In this paper, we make a distinction between origins and sub-origins, and destinations and sub-destinations. Origins and destinations belong to the final lanes (routes) and can be seen as the definite beginning and end points. The sub-origins and sub-destinations are the beginning and end points of each leg. If legs are at the beginning or end of a lane, the sub-origins and sub-destinations can be seen as final origins ans destinations, respectively. Lanes can consist of a different number of legs, with a maximum of four legs per lane. These legs do not have to be unique, in the sense that two lanes can have the same first leg, but different second, third or last legs. The lanes and legs have already been determined by the existing decision support system of the company. Hence, we are not creating new lanes or legs. A shipment order, or shipment, is transported over a specific lane that is given in the dataset. Shipments can have the complete same lane, but can also partially have the same lane. The latter occurs when shipments are for example transported over the same second leg within a lane, but come from different first legs and have different last legs, see Figure 1 for an example. Since each leg is associated with only one transportation mode, shipments that are transported over a leg can only be in a vehicle of that same particular transportation mode. However, the vehicle type per transportation mode may differ. These vehicle types will be discussed in Section 3.1.

SSC has datasets of several clients from different sectors, such as fashion, textiles and sports (FTS), fast-moving consumer goods (FMCG), and agri-food. Some datasets have 977 observations, and some have around 2.000.000 observations. We will focus on a dataset within FTS and use the inbound shipments. This dataset contains 12666 shipments, leading to a total of 14979 observations with a time period from October 2020 to December 2021. Each observation is equal to one shipment on a specific leg, which means one shipment can have multiple rows, but every row gives information about a different leg, see Appendix A. For every shipment, all feature values are thus available on the leg level. This means that we have all information regarding where shipments should be picked up and where they should be delivered for every sub-origin and sub-destination. Observations with the same (sub-)origin and (sub-)destination points are considered to be on the same leg. Furthermore, each shipment on a leg is transported into one vehicle, making the number of observations therefore equal to the number of vehicles.

The dataset contains many information and features, such as the estimated time of departure (ETD) and the estimated time of arrival (ETA) dates, geographical information (such as (sub-)origin/(sub-)destination point), package information (such as kg/number of items), transportation modes (such as rail/road/air/ocean/parcel) and container information. The ETD dates are only available for the definitive origins, and not the sub-origins. The ETA dates are only available for the final destination, and not the sub-destinations. Many features provide information about the geographical locations, but we will mainly use the features *SubLatFrom*, *SubLonFrom*, *SubLatTo* and *SubLonTo*, which are the latitude and longitude coordinates of the origin and destination point of one leg, respectively. The reasoning behind this is that these features are numerical, which works well with the clustering algorithm. We will use the shipment weight, denoted as the feature KGS, for the package information. Regarding transportation modes, we will use the feature $SubLaneModality$, which is the transportation mode used on a specific leg. The features *Booked as* and *Container type* will be used to obtain information about the containers. *Booked as* tells us if a container is booked as a full truckload/full container load (FTL/FCL) or a less than truckload/less than container load (LTL/LCL). The feature *Container type* tells us which type of container is used. Table 1 shows the existing types in this dataset. Furthermore, we have a variable *Shipment ID*, which gives each shipment a unique identification, and a variable *Lane ID*, which gives each lane a unique identification. Lastly, we have the *Index* feature, which gives us the leg index and indicates what part of the lane we are on.

The dataset will first be cleaned by removing or imputing several instances. After this, we will make subsets of the data based on the (sub-)destinations. We will elaborate further on this in Section 3.2. Besides the shipment information, the dataset also contains information about the CO_2 emission calculation. This includes emission categories, emission factors, the energy used during a shipment in megajoules (MJ), and the total CO_2 emissions.

3.1 Cleaning

We will clean the dataset by removing or imputing several instances. Some feature values regarding the total emissions and origin points are missing. To account for this, we will remove the complete shipment for these instances. To obtain unique shipment IDs we will create a new feature named *ID*, which is a combination of the features *Shipment ID* and *Lane ID*. Furthermore, the N/A values of the feature *Container type* have to be imputed when the transportation mode is ocean or rail. If the shipment weight is smaller than 4000 kg, we will impute these N/A values with '20GP' and otherwise with '40GP'. With the features *SubLaneModality*, *Booked as* and *Container type* we can then construct a new feature named *Capacity* which specifies the maximum vehicle capacity for that specific shipment. The maximum capacity per transport mode is shown in Table 1. The capacities for the transportation modes air, ocean, and rail will only be used if the container for the shipment is booked as an FTL or FCL. If we are dealing with LCL or LTL containers, we have assumed in consultation with SSC, that only 60% of the capacity of the container is available and can be used, since other companies are making use of the container as well. By parcel, we mean small vans that are meant for short distances. Road transport refers to larger trucks and is meant for larger distances.

Table 1: Maximum capacities in kg for the transportation modes

Transportation mode	Type	Maximum kg	Description
	20GP	21750	20ft General purpose
	20GOH	28180	20ft Garment on hangers
	40HC	26580	40ft High Cube
	40GP	26760	40ft General purpose
Ocean/Rail	40NOR	28510	Refrigerated
	40GOH	28750	40ft Garments on hangers
	45HC	25780	45ft Hi-Cube
	40RE	25960	40ft Reefer
	40REHC	26109	40ft High Cube Reefer
	45HC	25780	45ft Hi-Cube
Road	Truck	27000	
Air	Freight	120000	
Parcel	Van	150	
	Truck	762	

3.2 Subsetting

Not all features mentioned before will be used in the constrained clustering algorithm. We will shortly explain which features will be used and what their characteristics are in the final subsets of the data. We will use *SubLatFrom*, *SubLonFrom*, *SubLatTo* and *SubLonTo* as the latitude and longitude coordinates of the origin and destination point of one leg, respectively. Next, we need the shipment weight *KGS*, the *ETD*, the *Capacity*, the *ID*, the *Index* and the *SubLaneModality*. The shipment weights and vehicle capacities will be multiplied by 100 to avoid decimal values since decimal values are not handled well in the algorithm. For our clustering algorithm, we will create subsets from the dataset based on the destination points for each leg, also specified as sub-destinations points. This is manually done by grouping the dataset based on the sub-destinations. We create a new variable called *SubDestination* that concatenates the variables *SubLatTo* and *SubLonTo*. Subsets are then crated by grouping the dataset based on the sub-destinations. Each subset will correspond to one unique sub-destination, where all shipments in that specific subset must go to that specific sub-destination. These subsets are stored in a dictionary, where the key corresponds to the *SubLatTo* and *SubLonTo* values of the sub-destination, and the values correspond w The sub-destinations of the subsets where the *Index* feature is equal to 'Last' or 'Direct' can also be considered as final destinations. However, the (sub-)origin points in a subset are not unique, implying that shipments of different (sub-)origins can be clustered together as long as they are transported to the same (sub-)destination. This also implies that a subset can contain different legs, since legs consist of one (sub-)origin and one (sub-)destination, whereas subsets can consist of multiple (sub-)origins and one (sub-)destination. This approach of aggregating the data leads to 46 unique subsets, where the size of a subset varies between 1 and 10.332 observations. Furthermore, each subset will contain only one unique value of the feature *SubLaneModality* and one unique value of *Index*. Thus, a subset will always have the same leg index and the same transportation mode. This is not manually determined, but already incorporated in the structure of the current dataset of the company. Therefore, subsetting based on (sub-)destinations automatically leads to subsets with a unique *SubLaneModality* and *Index*. However, the vehicle type of this transportation mode may differ per subset, which is incorporated in the *Capacity* feature. In the subset, the number of observations corresponds with the number of shipments, which differs from the original, complete dataset. This is because observations within a subset always have the same leg index, and since a shipment is only transported over a leg once this implies that only unique shipments can be in a subset. An example of a subset is given in Table 18 in Appendix C.

Table 2 shows the counts of several features. There are many origins (production sites) and fewer destinations (distribution centers) for both the lanes and legs. The lane destinations are the final warehouses, whereas the leg destinations also include the transfer points. All shipments must go to one of the eight final lane destination points. The feature *Booked as* shows the type of container that is booked for ocean, rail, or road transport. Air, parcel and some road vehicles do not use containerized shipments. *Transportation mode* shows the distribution of the main mode used during a leg. The dataset contains 10867 shipments that are transported over direct lanes, which are mostly covered by the transportation mode parcel, which explains why parcel is the most used transportation mode. Table 3 shows the basic statistics of the shipment weight, the distance, and the total CO_2 emissions of the legs. The minimum distance and CO_2 emission can be equal to 0 in the case of direct shipments where the origin and destination point are in the same postal code and are thus close to one another. The distance cannot be obtained in these cases, which results in a total CO_2 emission of 0 kg.

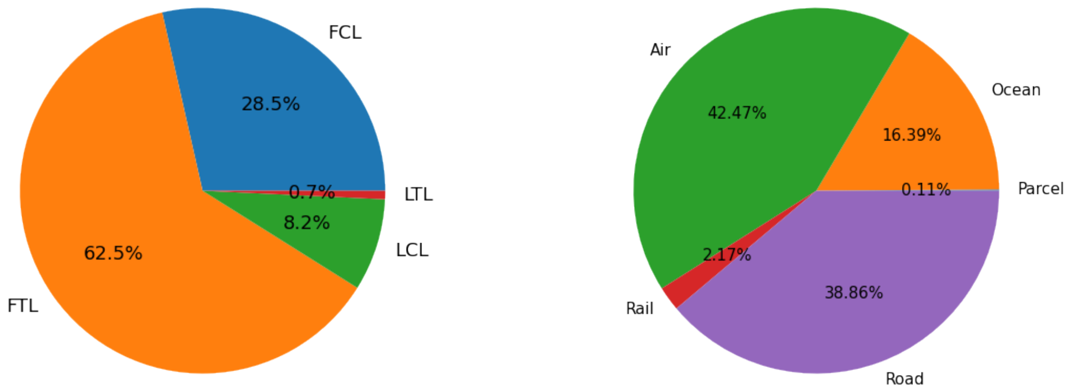
Table 2: Counts of several features

Geographical features		*Booked as*		*Transportation mode*	
Lane origins	2670	FTL	464	Parcel	10905
Lane destinations	8	LTL	43	Air	682
Leg origins	2694	FCL	686	Ocean	968
Leg destinations	46	LCL	1802	Rail	376
Number of lanes	2805	N/A	11982	Road	2049
Number of legs	2823				

Table 3: Basic statistics

	Count	Mean	St. error	Min	Max
Shipment weight (kg)	15007	607.79	1894.56	0.50	22834.09
Distance (km)	14953	2177.78	4340.53	0.00	20733.14
Total CO_2 emission (kg)	14755	231.66	1026.28	0.00	23866.04

Figure 3 shows the emission distribution of the features *Booked as* and *Transportation mode*, respectively. We see that FTL and FCL account for most of the emissions for containerized shipments, while LCL is the most common method of booking. Furthermore, parcel contributes the least to the total emission, while most lanes are covered by this mode. This is not surprising, since parcel routes are in general short. Air is the main contributor, followed by road.



(a) *Booked as*

(b) *Transportation mode*

Figure 3: Emission distributions

3.3 CO_2 emission calculation

We will use the emission related columns of the dataset to elaborate on the factors that are used to calculate the CO_2 emissions and what their impact is. As our dataset falls within the FTS sector, we will only focus on ambient shipments, which are shipments that do not need to be refrigerated. Appendix B shows graphic representations of the calculations, with a numerical example. The general equation for the CO_2 emission is

$$Emission = (tonkm \text{ or } TEUkm) \times \frac{EmissionFactor}{1000} \tag{1}$$

where tonkm and TEUkm are given by

$$tonkm = ton \times distance = \frac{Shipment\ weight}{1000} \times distance \tag{2}$$

$$TEUkm = TEU \times distance = \frac{Shipment\ weight}{CFactor} \times distance \tag{3}$$

The distance in kilometers (km) and shipment weight in kilograms (kg) determine the ton-kilometer (tonkm), which is a unit of measurement that is used in the logistics sector that gives the weight in tons of material transported, multiplied by the number of kilometers driven. The distance, shipment weight, and the conversion factor (CFactor) determine the TEUkm, which is a unit of measure of container transport which represents the transport of one twenty-foot equivalent unit (TEU) over one kilometer. The CFactor is used to convert the shipment weight into TEU, which is equal to 4500 in the FTS sector. TEUkm is used for the transportation mode ocean and tonkm is used for the remaining transportation modes. The emission factor differs per transportation mode.

Air The distance determines if the modal type is 'AirShort', 'AirMedium' of 'AirLong', which then determines the emission factor. The modal type is 'AirShort' if the distance is smaller than 1000 km, 'AirMedium' if the distance is larger than or equal to 1000 km and smaller than 3700 km, and 'AirLong' for every other distance.

Ocean/Rail The emission factor for the transportation modes ocean and rail is dependent on the origin and destination region.

Parcel The emission factor for the transportation mode parcel is dependent on the region and the distance. If the distance < 100 km, the emission factor is a constant. If the distance is larger ≥ 100 km, the emission factor can be calculated by

$$EmissionFactor_{Parcel} = \frac{distance - 100}{distance} \times \text{RoadTruck<40t} + \frac{100}{distance} \times \text{RoadVan<3.5t}$$

where RoadTruck<40t and RoadVan<3.5t are constants that represent emission factors for specific vehicle types of parcel. The intuition behind this is that for distances larger than 100 km, a hub is used. The hub is 50 km away and the route to and from the hub is covered by a van. Thus, a total distance of 100 km is covered by a van, and the remaining part of the distance is then covered by a truck.

Road The emission factor for the transportation mode road is dependent on the calculation type, where we make a distinction between a default and modeled calculation. For the default calculation, the emission factor is determined by the origin region and the category. In our case, this is always RoadTruck<40t because we only have shipments within the FTS sector. If we have a shipment that is not from the region of North America and is booked as FTL or FCL, we have a modeled calculation with a different emission factor. To determine the emission factor we first have to calculate the actual distance, which is the distance multiplied by a distance correction factor, and the empty distance, which is the distance traveled where the vehicle is empty. These distances are dependent on the origin region and the type of fuel, which is always diesel in this paper. Next, the energy in megajoules (MJ) for a loaded and empty truck is calculated with the previously calculated distances. This energy is then divided by the MJ/l factor to determine how much liter diesel is needed, which is then converted to the actual CO_2 emission through the CO_2e/l factor. As before, the components used to determine the energies are constant since we only have shipments within the FTS sector. The final emission factor can be calculated by

$$Emission_{Road} = \frac{Total\ Energy\ Used}{MJ/l} \times (CO_2e/I)/kgs/distance \times 10^6 \times \text{RegionFactor} \times \text{ConditionFactor}$$

4 Model and Solution

In this section, we will define a partitioning based constrained clustering algorithm, where the constraints will be based on vehicle capacity and maximum lead times. The idea is to apply freight consolidation on a cluster-to-destination level, where shipments of geographically close (sub-)origins are grouped and shipped to the same (sub-)destination. As stated before, origins and destinations belong to the final lanes (routes) and can be seen as the definite beginning and end points. The sub-origins and sub-destinations are the beginning and end points of each leg, where a leg is defined as a segment of a lane. For every (sub-)destination, we will make a subset of the data to ensure all shipments within this subset will arrive at the same precise (sub-)destination.

Consider, per subset, n less-than-truckload shipments, with an origin o_i, destination d_i and shipment weight w_i. Each shipment has a has a (generated) ETA eta_i based on time deadlines, which states when the shipment has to be delivered. In the original situation, each shipment is sent separately in its own vehicle from its (sub-)origin o_i to its (sub-)destination d_i, where $i = 1, ..., n$. In the consolidation approach we consider consolidated origins co_c and consolidated destinations cd_c, with $c = 1, ..., k$. The problem is to allocate consolidated shipments to vehicles in consolidation origins (pre-haulage) and ship them to a consolidation destination (end-haulage), such that the final time deadline is met and shipment weight restrictions are met. The costs in this case consist of the pre-haulage, end-haulage and consolidation shipment costs. However, since there is no information available about the expenses, the focus of this algorithm is to consolidate based on minimizing the CO_2 emissions. As a result, we will use the results of the algorithm to compare the total CO_2 emission before consolidation to the total CO_2 emission after consolidation, and analyse the difference in the utilization of the vehicle capacities.

Subsets can contain different sub-origin points, and can therefore contain different legs. This implies that if different sub-origins are clustered together, new legs must be created to make this clustering possible. In this research, however, no focus will be put on this area. As stated in Section 1, the new legs that are needed to connect these sub-origins are given as suggestions for a more optimal consolidation. The sub-origin which has the shortest distance to the sub-destination will be used to transport the consolidated shipment. We will refer to this sub-origin as the main origin. Shipments from other sub-origins are then first transported to the main origin, before travelling to the sub-destination. Furthermore, the vehicle or container type in which a shipment is transported over a leg is flexible. This means that shipments do not have to be transported in the same vehicle/container that they were originally transported in.

Proceeding with the approach, we will use the subsets in the clustering algorithms, which means that the constrained clustering algorithm will be executed for every unique subset, and examine how and where consolidation is possible. The clustering algorithm takes capacity constraints and time constraints into account, which will be explained separately, followed by several limitations of k-means and freight consolidation. Algorithm 1 gives the pseudo-code of the constrained clustering algorithm.

The costs defined in this algorithm are based on the distances between the (sub-)origins and the cluster centroids. The inputs and outputs of the algorithm are as follows.

Input a subset, with the corresponding variables *SubLatFrom*, *SubLonFrom*, *SubLatTo*, *SubLonTo*, *KGS*, *ETD*, *ID*, *Index* and *SubLaneModality*, the number of clusters k, the cannot link constraints, which are further explained in Section 4.1.3, the minimum size of the cluster, which we have specified as 0, the maximum size of the cluster, an array containing the shipment sizes

Output a dataframe containing the cluster centroids, cluster sizes and maximum capacity for each cluster labels, a dataframe containing the cluster label, the original ETD, the shipment ID, the shipment weight and the emission for each observation in the subset that was given as a an array containing the final cluster labels

4.1 Cluster analysis

Using the explanatory data analysis technique of cluster analysis, objects with similar properties can be divided into smaller groups (Tryon, 1939). A cluster is defined as a set of objects aggregated together because of certain similarities. It is a form of unsupervised learning, which refers to identifying patterns in datasets that are not classified or labeled. Contrarily, supervised learning requires that all training examples are labeled. It is nevertheless often the case that neither of these learning types is suitable. Semi-supervised learning methods use both labeled and unlabeled data to resolve this discrepancy. Constrained clustering, a form of semi-supervised learning, was developed to extend clustering algorithms to incorporate existing domain knowledge in the form of labeled data or constraint sets (Wagstaff, 2010). In this paper, the semi-supervised cluster approach will be based on constraints. Constrained clustering methods aim to find a partition of clusters of the dataset that (ideally) satisfies all constraints in the constraint set. González-Almagro et al. (2020) describe three main types of constrained clustering: cluster-level (Bradley et al., 2000), instance-level (Davidson & Ravi, 2007) and feature-level constrained clustering (Schmidt et al., 2011). Cluster-level constraints pose restrictions on the size and form of the cluster. The must-link (ML) and cannot-link (CL) requirements, on the other hand, specify whether two specific instances of a dataset must be placed in the same or different clusters. Feature-level constraints specify whether instances must or must not be grouped into a cluster, depending on specific attribute values. In this paper, we will make use of instance-level constraints in the form of time constraints, and feature-level constraints in the form of vehicle capacity constraints.

4.1.1 Partitioning based clustering

In this paper, we will construct a constrained K-means algorithm based on Levy-Kramer and Klaber (2022), and use this algorithm in the context of sustainability. K-means (MacQueen, 1967) is one of the most commonly used partitioning based clustering methods. The k-means algorithm is excellent for clustering large datasets when compared to other clustering techniques (Anderberg, 1973). It is an iterative algorithm that aims to partition a dataset into k clusters, where the number of clusters k is fixed and predetermined. The k-means algorithm of Levy-Kramer and Klaber (2022) is defined such that a

Algorithm 1 Pseudo-code clustering algorithm

```
for each subset do
    initialize centroids via k-means++ (Algorithm 2)
    iteration = 0
    best_labels, best_centers, best_costs are none
    previous centroids are none
    while new centroids are different from previous centroids and iteration is smaller than the maximum number of iterations do
        determine for every observations cluster labels and costs with the MCF network
        for every observations do
            apply the objective function to check if the assigned cluster label violates time constraints
            assign penalty if a time constraint is violated (Algorithm 4)
            reassign observation to cluster with lowest penalty to obtain new cluster label
        end for
        determine new centroids and new costs
        if best_costs is none or new costs are lower than best_costs then
            set best_costs as new costs
            set best_labels as new labels
            set best_centers as new centroids
        end if
        iteration += 1
    end while
    for every cluster do
        if the cluster size is larger than the maximum capacity then
            order the shipments by size from largest to smallest
            for every ordered shipment do
                locate the biggest shipment in the cluster
                assign biggest shipment to an empty cluster until the cluster size is smaller than or equal to the maximum capacity
            end for
        end if
    end for
end for
```

minimum and/or maximum size for each cluster can be specified in terms of the number of observations, where the cluster assignment step is modified by formulating it as a Minimum Cost Flow (MCF) linear network optimization problem. The goal of the MCF network is to find the 'cheapest' possible way of sending a certain amount of flow through a network, where the edges of the network can have certain capacities (Sifaleras, 2016). Considering we have predefined shipment weights w_i, which can be seen as the network flows, that need to be distributed over the vehicles, while minimizing the CO_2 emissions and taking vehicle capacities into account, the MCF network is a suitable algorithm. The network is then solved using a cost-scaling push-relabel algorithm using the `SimpleMinCostFlow` implementation of Google's Operations Research tools (OR-Tools) (Perron & Furnon, n.d.). The k-means algorithm of Levy-Kramer and Klaber (2022) is based on the algorithm of Bradley et al. (2000), where their algorithm has been modified such that maximum cluster sizes can be specified along with a minimum cluster size, where size is again specified in terms of the number of observations n. We have modified the algorithm of Levy-Kramer and Klaber (2022) such that a minimum and/or maximum size for each cluster can be specified in terms of vehicle capacity in kg instead of the number of observations, and where each cluster will be covered by one vehicle. In the remainder of this paper, we have set the minimum size equal to zero for simplicity. The capacities, supply nodes, and demand nodes of the MCF formulation are modified such that the network flows consist of the shipment weights in kg. The algorithm is initialized such that the default number of clusters is equal to the length of the subset, thus equal to the number of shipments. The default number of iterations in the constrained clustering algorithm is 100 and the random state is not specified.

The value of the number of clusters k can be overwritten. This input variable has a default value that is equal to the length of the dataset n, but for large datasets, this could cause problems regarding the run time. To account for this, we have made some extra assumptions about the number of clusters when the dataset becomes large. If the number of observations in the dataset n is smaller than 300, we still use the length of the dataset as the number of clusters. If the number of observations n is between 301 and 500, 501 and 800, and 801 and 2000, we use k=300, k=500 and k=700 respectively. If the number of observations is larger than 2000, we use k=2000.

4.1.2 MCF formulation with capacity constraints

Let $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be a subset of the dataset with n observations that go to the same sub-destination, where each x_i for $i = 1, ..., n$ consists of the latitude and longitude coordinates of the origin point, the shipment weight w_i, the leg index $index_i$, the ETD etd_i, the shipment ID id_i, the transportation mode $mode_i$ and the ETA eta_i. Let $C = \{C_1, ..., C_k\}$ then be set of k clusters, representing k vehicles, for this subset, where we assume k is given. We will elaborate further on the specification of k and provide limitations in Section 4.2.1. Each cluster, or vehicle, k corresponds with a maximum vehicle capacity of vc_c, where $c = 1, ..., k$. $\bar{X} = \{\bar{\mathbf{x}}_\mathbf{1}, ..., \bar{\mathbf{x}}_\mathbf{k}\}$ are the cluster centroids we want to determine. The initialization of the centroids is done by the k-means++ algorithm, where the first cluster centroid is chosen uniformly at random from the subset, see the pseudo-code in Algorithm 2. Each subsequent cluster centroid is chosen from the remaining data points with probability proportional to its squared distance from the point's closest existing cluster centroids. The inputs and outputs of the algorithm are as follows.

Input a subset containing *SubLatFrom* and *SubLonFrom*, the number of clusters

Output the cluster centroids, which are chosen from the subset and will therefore have the same coordination points as the observation to which they correspond

Algorithm 2 Pseudo-code initialization of centroids by k-means++

```
initialize first centroid
for each cluster in the range of the number of clusters − 1 do
    compute squared euclidean distance between centroid(s) and observations
    if all squared distances are zero then
        set probability equal to 1/#observations
    else
        set probability by dividing the distance by the sum of all distances
    end if
    choose next centroid(s) based on the highest probability until all centroids are initialized
end for
```

The cluster centroids $\bar{X}$, distances, shipment weight w_i, and maximum vehicle capacity $\overline{vc}$ are then used as input in the MCF network of OR-Tools. We want to use this network to assign the observations, or shipments, to the resulting cluster centroids and give each of these shipments a label that corresponds with the cluster it is in. The number of shipments in a subset is equal to n. The MCF network constructs a graph $G = (V, E)$ where V denotes the set of nodes in the graph and E denotes the set of arcs. Each arc (u, v) is associated with a capacity $c(u, v) > 0$, a flow $f(u, v) > 0$ and a unit cost $uc(u, v)$, with $u \in V$ and $v \in V$. In our case, the unit cost of an arc is defined as the distance between the nodes that are connected through this arc. There can be one or multiple source nodes $s_i \in S$, with $S \in V$, associated with supplies, and one sink node $t \in T$, with $T \in V$, associated with demand. The number of source nodes is equal to n, because each source node corresponds with one shipment.

Figure 4 shows an example of an MCF graph, where the cluster centroids $\bar{X}$, distances $uc(u, v)$, shipment weight w_i, and maximum vehicle capacity $\overline{vc}$ are used as input. The goal of this network is to assign n shipments to k clusters by using k dummy nodes, while incorporating the capacity constraints. The cluster centroids are used to determine the coordinates of the dummy and cluster nodes. Each source node $s_i \in S$ stands for one shipment i with the shipment weight w_i as the node's supply. All shipments must end in the sink node $t \in T$, where the total demand is specified as the total shipment weight of all shipments within a subset, or $\sum_{i=1}^{n} w_i$, with n the total number of shipments in the subset. Because a shipment can be put in every cluster c, each source node, which corresponds with one shipment, has arcs going to all dummy nodes with a capacity equal to the shipment weight w_i of the corresponding shipment. The costs of these arcs are determined by taking the distance between the origin point of the shipments and the cluster centroids $\bar{X}$, multiplied by 1000 to obtain a more precise value. The dummy nodes are necessary to assure that the capacity constraints are not violated. Each dummy node is connected with one cluster node, where the capacity of these arcs is equal to the maximum vehicle capacity for that specific cluster with zero costs for each arc. These zero costs follow from the fact that the dummy nodes and cluster nodes have the same coordinates, and thus a distance $uc(u, v)$ equal to zero. The maximum vehicle capacities do not have to be the same, they may differ per cluster. The arcs that go from the cluster nodes to the sink nodes have a capacity equal to vc_c for that specific cluster, to make sure the demand is met and the flow is not restricted. This is again with zero costs because the distances $uc(u, v)$ between the cluster nodes and sink node are equal to zero. This specification of the MCF allows clusters to be empty and since each cluster is covered by one vehicle, it means that not every vehicle has to be utilized. These empty vehicles can be removed or can be used as backup if necessary. The MCF network will determine which clusters will be filled and which clusters will remain empty. This is dependent on the distance between the cluster centroids $\bar{X}$ and the observations, where the cluster centroids with small $uc(u, v)$ are likely to be filled before the centroids with large $uc(u, v)$. Appendix C contains a detailed numerical example of a MCF network.

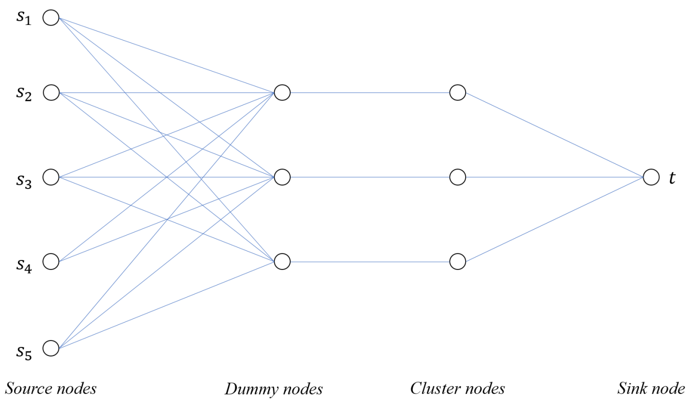


Figure 4: Example of MCF graph

The MCF network can be formulated as follows:

$$\text{minimize} \quad \sum_{(u,v)\in E} uc(u, v) \cdot f(u, v) \tag{4}$$

$$\text{s.t.} \quad f(u, v) \leq c(u, v), \qquad \forall (u, v) \in E \tag{5}$$

$$f(u, v) = -f(v, u), \qquad \forall v \in V \tag{6}$$

$$\sum_{v\in V,\ (u,v)\in E} f(u, v) - \sum_{q\in V,\ (q,v)\in E} f(q, u) = 0, \qquad \forall u, v, q \in V \setminus S \setminus T \tag{7}$$

$$\sum_{v\in V\setminus S,\ (s_i,v)\in E} f(s_i, v) = d_i, \qquad i = 1, ..., n \tag{8}$$

$$\sum_{v\in V,\ (v,t)\in E} f(v, t) = -\sum_{i=1}^{n} d_i \tag{9}$$

The goal is to minimize the total cost of the flow over all edges, as stated in the objective function 4. The capacity constraints are stated in 5, indicating that a flow cannot exceed a capacity. Constraint 6 is a symmetry constraint, and constraint 7 is for flow conservation, which means that no node, except the source and sink node, of a flow network creates or stores flow. Constraints 8 and 9 state the required flow, where d_i is the amount of flow to be sent from source s_i to sink t. We associate a value of 0, which indicates that no supply or demand should be stored in transshipment nodes.

Another input variable is the maximum vehicle capacity $\overline{vc}$, which corresponds with the $c(u, v)$ defined above in constraint 5 and the *size_max* defined in Algorithm 3. This is an array including the maximum capacities of each cluster, thus $\overline{vc} = \{\mathbf{vc}_1,..., \mathbf{vc}_k\}$. This input parameter determines the capacities of the flows between the dummy and the cluster nodes, or the maximum $f(u, v)$ that can flow between the dummy and the cluster nodes. Each observation i in the dataset is currently transported by a separate vehicle v_i, and since each vehicle has a corresponding capacity $c(u, v)$, these capacities can be linked to the observations. Each observation i has thus one maximum capacity $c(u, v)$ that corresponds with the vehicle v_i in which that observation, or shipment, is transported. Since we have n observations, this leads to n corresponding maximum capacities. If the number of clusters k is not equal to the number of observations n, but smaller than the number of observations, thus $k < n$, we will have to make some adjustments such that the maximum capacities still give a good representation of the dataset. We need a maximum capacity for every flow $f(u, v)$ that goes from a dummy to a cluster node, which means we need k values of the maximum capacity. We have several approaches to achieve this. The maximum capacity is dependent on the transportation mode. If the transportation mode is air or parcel, we are only dealing with one type of vehicle and thus only one maximum capacity value. For every cluster, we will set each capacity equal to this unique maximum value. However, if the transportation mode is road, ocean or rail, we are dealing with different capacities since the containers of these vehicles can be booked as a full truckload/full container load (FTL/FCL) or a less than truckload/less than container load (LTL/LCL). In the case of FTL/FCL, we assume that the whole container is available, whereas for LTL/LCL we have assumed in consultation with SSC, that only 60% of the container is available. Additionally, for the transportation modes ocean and rail we have different types of containers that can be used, see Table 1. To assure that all different values are incorporated in the maximum capacity of the clusters, we will make use of the fraction of the available capacities. For each subset, we will determine the fraction of every capacity. This leads to a distribution of the maximum capacities that are included in the subset. These fractions are then used to determine how many flows $f(u, v)$ will have the same capacity. To give a more concrete example, if vehicle type A has a maximum capacity c and $x\%$ of the shipments in the subset are originally transported by a vehicle with maximum capacity c, then the number of flows that will have this maximum capacity c will be $x\% \times k$, where k is the total number of clusters. A pseudo-code is given in Algorithm 3.

Input an array containing the maximum capacities of size n, the number of clusters

Output an array containing the maximum capacities of size k

A limitation of the `SimpleMinCostFlow` implementation of OR-Tools is that this method allows for partial shipments. Shipments are divided over multiple clusters, while it is preferred to have shipments as a whole. To account for this complication, we will rearrange the clusters. This will manually be done ad-hoc by removing the shipments that are divided over multiple clusters and assigning these shipments to other clusters. A label of -1 will be given to the divided shipments to remove them from the clusters. Next, the cluster sizes in terms of shipment weight will be calculated. After this, the clusters will be rearranged by assigning the divided shipments to other clusters that contain shipments with the same

Algorithm 3 Pseudo-code determining maximum capacities

```
for every maximum capacity in size_max do
    count the number of times the maximum capacity appears
    divide this count by the total number of observations in the subset to obtain the capacity fraction
end for
if all maximum capacities of the observations have the same value then
    set the maximum capacities of all clusters to this value
else
    create an empty list final_size_max to store the maximum capacities
    for every capacity fraction do
        create an array with length equal to the number of clusters times the capacity fraction
        fill this array with the corresponding maximum capacity of the capacity fraction
        add array to the list final_size_max
    end for
end if
```

origin coordinates, while still taking the capacity constraints into account. This means that the total shipment weight of a cluster remains below the maximum capacity, thus $\sum_{i=1}^{n} w_i < vc_c$ for $c \in k$. If no clusters are available, a new cluster will be formed to assign the shipment.

4.1.3 Time constraints

Besides the capacity constraints, we also want to incorporate time constraints into the constrained clustering algorithm. These time constraints will be based on the maximum lead times, which are dependent on the transport modes. The maximum lead time is defined as the maximum time a shipment can be stored at a location without being transported to the next location. Table 4 shows the maximum lead times in days for each transportation mode and the duration in days for each transportation mode, where the duration is defined as the number of days it takes to transport a vehicle over a leg. We have assumed fixed values per transportation mode in consultation with SSC.

Table 4: Maximum lead times and duration in days for the transportation modes

Transportation mode	Rail	Road	Air	Ocean	Parcel
Maximum lead time	14	14	7	21	14
Duration	16	2	2	31	2

We define the time constraints based on the maximum lead time as follows. Shipments within the specific subsets cannot be clustered together if the time between their ETA is more than or equal to the maximum lead time. To incorporate these constraints into the clustering algorithm, we will first make a list of cannot-link (CL) constraints. Each CL constraint consists of two observations that cannot be clustered together because the time between their ETA exceeds the maximum lead time. The time constraints are applied after the MCF network, which means that the observations are already clustered into vehicles based on capacity constraints. For every observation, we check if it is clustered whilst violating one or more CL constraints. If this is the case, a fixed penalty of w will be added to this observation for that specific cluster. The total penalty consists of the fixed value w and the distance $uc(u, v)$ between the observation and the specific cluster centroid. The objective function of the PC K-means algorithm of Švehla (2018) is modified and used to determine these penalties. It takes the cluster labels from the MCF output as input, together with the CL constraints, cluster centroids, and observations. Algorithm 4 shows the pseudo-code of assigning the penalties. After this, we obtain for every observation an array with penalties for each cluster. With this, we determine for every observation which cluster has the lowest penalty and assign it to that specific cluster. The resulting cluster labels that come forth from this, are used to determine the new cluster centroids. The new centroids are then used in the MCF network at the next iteration. This process is iterated until convergence, where the default number of iterations is set to 100. The input and output of Algorithm 4 are stated below.

Input a subset containing *SubLatFrom* and *SubLonFrom*, the cluster centroids, the cluster labels, the CL constraints

Output for every observation an array with penalties for each cluster

Algorithm 4 Pseudo-code assigning penalties

```
for every observation x do
    for every cluster y do
        determine the distance between observation x and cluster y
        set penalty = distance
        for every constraint in the CL constraints do
            if observation x violates constraint for cluster y then
                penalty += w for observation x for cluster y
            end if
        end for
    end for
end for
```

Incorporating the time constraints leads to breaking the initial clustering structure that follows from the MCF network. This is a result of the reassignment, where observations can be reassigned in the case of high penalties. This reassignment does not take any capacity constraints into account. As a result, some capacity constraints may be violated if observations with high penalties are reassigned to other clusters. To make sure that the capacity constraints are not violated, the clusters need to be rearranged one final time. This will again be realized via an ad-hoc approach. The clusters that violate the capacity constraints are rearranged in the following way. The biggest shipments within the clusters are allocated to other available and empty clusters until the capacity constraint is satisfied. This way, we make sure that both types of constraints are satisfied. The ad-hoc approach of the final rearrangement is incorporated in lines 21-29 in Algorithm 1.

However, it could happen that a large shipment needs to be reassigned, but the available empty clusters, thus the available empty vehicles, are not big enough. In this case, rearrangement is not possible and the shipment will be put in a smaller cluster, which leads to the capacity constraint being violated. An error message is returned that states: 'The final cluster size is too large for cluster x. Some shipments are too large for the empty clusters or the number of clusters is incorrect.'. The algorithm will still be completed, such that the other clusters may be used. When the algorithm is completed, it will return a data frame that contains the cluster centroids, cluster sizes, and maximum capacity for each cluster label. Furthermore, it returns a data frame dedicated to the time constraints, which includes the shipment ID, cluster label, and ETA for each shipment. Lastly, it returns a list containing the cluster labels for each observation.

4.1.4 Generating the ETA

As stated before, each observation has its own estimated time of departure (ETD) and estimated time of arrival (ETA). Unfortunately, the dataset only includes the ETD etd_i for the first leg and the ETA eta_i of the last leg. This means the ETD and ETA dates for all legs in between missing in the dataset and thus not known for each specific leg, which is why we have to generate this ourselves. To determine this information we will use the leg index $index_i$, which indicates what part of the lane we are on. The subsets where the leg index equals 'First' will first be used in the clustering algorithm since these subsets contain the starting points of our shipments, and are provided with an ETD. The clusters that are formed by the algorithm can contain shipments with different etd_i values. After the clusters are made for the subsets where the leg index equals 'First', we can take the latest ETD, thus the maximum etd_i of each cluster as the final ETD for the whole cluster, and add the duration of the transportation mode of the leg to generate the eta_i at the next leg. This implies that shipments within a cluster have the same eta_i for the next leg. Next, we cluster the subsets with a leg index equal to 'Second' and generate the eta_i at the next leg by using the generated eta_i at the previous clustering as the etd_i for this clustering. The maximum etd_i of each cluster is again taken as the final ETD plus the transport duration of the leg. This process is repeated until all subsets are clustered.

However, the first leg is missing for some shipments, meaning these shipments start at the second leg in the dataset. As a result, these shipments do not have a generated ETA from a previous leg since this is missing. Even though the leg index is 'Second', these shipments do have an ETD value because they represent the 'begin point' of the shipment in the dataset. For this reason, we will set the ETA equal to the ETD that is given in the dataset.

Fortunately, we can generate the eta_i and etd_i without taking into consideration the final ETA at the last leg. This is because the company is flexible with its timing schedule. It does not cause big problems if the generated eta_i slightly differs from the final ETA. However, in the case of deadlines (such as seasonal collections in the fashion industry), the ETA is important. In this case, the waiting times of the transportation modes can be shortened to assure that these shipments will arrive on time. In this research, however, no focus will be put on this area since we do not take express shipments into account.

Algorithm 5 shows the pseudo-code of the generation of the eta_i.

Input a dataframe containing the cluster label, original ETD and shipment ID of every observation in a subset

Output ETA values for the next leg of every observation in the subset

Algorithm 5 Pseudo-code generating ETA

```
make separate dataframes for every cluster through grouping by cluster label for the input dataframe
for every cluster do
    locate maximum ETD value
    for every shipment ID in the cluster do
        locate index for next leg
        change ETA value of next leg to the maximum ETD value plus the duration of the leg
    end for
end for
```

4.1.5 Alternative constraint formulation

Besides the constraints based on maximum lead times, there is another possibility to formulate constraints. We can look at vehicle departure times and use these values as etd_i. If these departure times are known, we can incorporate this in the constraints as well by stating that shipments can only be in the same cluster if their eta_i is before the vehicle departure time. In the current setting, this information is not known, so no further focus will be put on this approach.

4.1.6 Suggestion of extra legs

As explained in Section 1, some forms of consolidation can only be possible if non-existing legs are added. This is the case when shipments from different (sub-)origins are clustered together and shipped to the same (sub-)destination. Due to the absence of necessary information, we have decided to propose the extra legs as a suggestion rather than adding them to the existing network ourselves. However, to compute the total CO_2 emission of a consolidated shipment, we need to calculate the CO_2 emission of these extra legs. To determine this, we have calculated the euclidean distance of these legs. Additionally, we assumed that these extra legs will be covered by trucks for simplicity, meaning that the transportation mode is road, and calculate the corresponding CO_2 emissions based on this distance and transportation mode. Algorithm 6 shows the pseudo-code of the emission calculation of the extra legs. Line 3-5 cover subsets that have unique legs, and therefore have unique origins and distances. The main origin in line 8 is the origin that will be used for consolidation. All shipments will go to this origin and will be transported to the next (sub-)destination as one consolidated shipment. The inputs and outputs are as follows.

Input a dataframe containing the cluster label, original ETD, shipment ID, shipment weight and maximum capacity of every observation in a subset, the transportation mode of the subset

Output CO_2 emissions of extra legs

Algorithm 6 Pseudo-code calculating emission of extra legs

```
make separate dataframes for every cluster through grouping by cluster label for the input dataframe
for every cluster do
    calculate the total shipment weight and maximum weight capacity
    if the (sub-)origin and distance in the cluster are unique then
        set distance to the unique distance
        set extra emission to zero
    else
        set distance to the minimum distance
        set (sub-)origin with smallest distance as main origin
        for every origin except the main origin do
            calculate the length of the shortest path between origin and the main origin
        end for
        calculate the extra emission based on additional distance
    end if
```

4.2 Limitations

4.2.1 Limitations of k-means

K-means has multiple limitations which are also applicable in the algorithm described above (Celebi et al., 2013). The biggest drawback of the k-means clustering is that it requires prior knowledge of the data because the number of groups k, must be predefined. This issue can be resolved by running the k-means algorithm for various values of k, comparing the results, and selecting the best value of k. However, this procedure can take some time. In this case, the number of clusters can be seen as the number of available vehicles for one transport mode at a specific moment in time. Since the algorithm allows clusters to be empty, the number of available vehicles at that moment in time can be used as k, if this is known. Furthermore, the clusters obtained by k-means cannot overlap, because each observation in the dataset will be assigned to one distinct cluster. K-means can therefore not find embedded and nested clusters. In this case, that means that one complete shipment can only be assigned to one cluster with a probability of 1. This implies that the clustering output of k-means provides only one consolidation solution since shipments are put in clusters with a probability of 1. In reality, this is not always the case. Shipments can be put in multiple clusters with different probabilities, leading to different consolidation solutions. Other soft clustering algorithms such as fuzzy clustering can resolve this problem. With soft clustering methods, an observation can belong to multiple clusters with different probabilities. This makes it possible for shipments to be clustered in multiple ways, making it an insightful and promising method for freight consolidation.

4.2.2 Limitations of freight consolidation

Several factors such as shipment weight, transportation mode, container size (and more) can determine if and how freight consolidation is possible. For example, if several shipments on the same leg have a small weight with close ETD dates, these shipments can be bundled. However, if a container is completely filled or if the weight of the shipment is large, freight consolidation may not always be applicable even if the CO2 emission is high.

As with most systems, there is a trade-off for using consolidated shipping. Zhou et al. (2011) describe several limitations that may occur when applying freight consolidation within one company or throughout multiple companies. The first one is the dispatching limit of a vehicle, which tells us that at least $x\%$ of the vehicle capacity must be utilized before departing. A higher dispatching limit requires more shipments per vehicle and also incurs a higher probability of shipment deadline expiration. Next, they identify the shipment deadline. Longer shipment deadlines will lead to the use of larger vehicles, which is not always beneficial. Short deadlines may lead to not fully loaded vehicles, which is not cost-effective or sustainable. Besides the shipping deadline, time plays another important role. Due of the handling and processing requirements, consolidated shipping can take longer. It can cost money in downtime for these operations to take longer, as well as for carriers to mix shipments to make a complete load.

Vaillancourt (2016) identify additional limitations regarding consolidation throughout companies. For freight consolidation to function effectively, a network of clients, carriers, processing facilities, and lanes must be coordinated. However, coordinating these complex systems can be expensive for companies, since they may have to dedicate specific resources. Besides these aspects, access to a consolidation area is also important. Costs related to security, capacity, and resources can play a role within certain organizations. Consolidation with a larger number of organizations can be more complex, which can lead to increasing costs. In addition, a higher number of companies included in consolidation may make it difficult to decide who to collaborate with and how to best organize that collaboration while minimizing information gaps and unpredictability. In addition to this, smaller organizations might not want to consolidate with larger organizations as they might receive fewer benefits from cooperating. However, consolidation with multiple organizations does increase the amount of material handled, making it easier to reach the dispatching limit, which leads to more efficient shipment volumes and therefore more easily created gains.

5 Results

In this section, we will discuss the results of the constrained clustering algorithm. As stated in Section 3, we will have 46 unique subsets, with varying sizes between 1 and 10,332 observations. The subsets are ordered according to their leg index, which indicates which part of the lane the shipment is on. This is because the algorithm will first be executed for subsets with a leg index equal to 'First', followed by subsets with a leg index equal to 'Second', 'Third', and 'Last', subsequently. One of the subset contains 10,332 observation, which takes too much time for the algorithm to run. We will execute the algorithm separately for this subset, also referred to as the big subset. Next, we will shortly describe the results and examine the output of the constrained clustering algorithm for each specific leg index and the big subset. After this, the results for the emission reduction will be examined and analyzed, where we will again present the results for each specific leg index and the big subset. Additionally, we will investigate the consolidation results in the case where all shipments are treated as FTL or FCL. An important remark is that the run times of the algorithm should be interpreted as approximate run times and can slightly differ per execution. Furthermore, the results shown here are the results after the constrained clustering algorithm including the additional ad-hoc adjustments.

5.1 Constrained clustering algorithm

***Index*=’First’** There are 25 unique subsets where the leg index equals ’First’. These subsets have lengths that vary between 2 and 58 observations. As stated in Section 3, each observation is equal to one shipment on a specific leg, and each shipment is transported into one vehicle. The number of observations is thus equal to the number of vehicles. The number of clusters is dependent on the size of the subset, as stated in Section 4.1.1. After the algorithm is completed, we remove the clusters that contain zero shipments, also referred to as empty clusters. An example of the output of one of the subsets is shown in Table 5 and 6, where the cluster size and maximum capacity are not converted back and thus still multiplied by 100. This subset has parcel as transportation mode and key ’31.7420005798339 , 118.861999511718’, which corresponds to the sub-destination of Nanjing, China. This means that all shipments in this subset must go to the sub-destination of Nanjing, China. Table 5 shows the results on a cluster level. It contains the final clusters with cluster centroids $SubLatFrom$ and $SubLonFrom$, which are in this case equivalent to the origin point in Nanjing, China, the cluster sizes or shipment weights in kg, and the maximum capacity of the cluster in kg. Table 6 shows the results on an observation level, where the cluster label, the original ETD, ID, shipment weight (multiplied by 100), and the emission in kg for each observation are given. These results provide a detailed overview of the clusters and their content. For this specific cluster, we see that consolidation is possible. This subset has 6 observations and results in 2 clusters, which means that only 2 vehicles are required, instead of the 6 vehicles in the original situation. These tables also demonstrate why further consolidation is not possible. For this subset, this is due to the ETA dates which are too far apart.

Table 5: Results on cluster level of subset ’31.7420005798339 , 118.861999511718’

Cluster label	SubLatFrom	SubLonFrom	Cluster size	Maximum Capacity
0	32.05838	118.79647	39300.0	76200.0
1	32.05838	118.79647	16400.0	76200.0

***Index*=’Second’** There are 12 unique subsets where the leg index equals ’Second’. These subsets have lengths that vary between 1 and 453 observations. An example of the output of one of the subsets is shown in Table 7 and 8. This subset has ocean as the transportation mode and key ’57.7 , 11.95’, which corresponds to the sub-destination of the Port of Gothenburg, Sweden. This subset has 309 observations and results in 105 clusters, which would result in a large table. We will therefore only display part of the results. Table 7 shows the results on a cluster level. Here, the cluster centroids $SubLatFrom$ and $SubLonFrom$ do not always correspond with an origin point. If this is the case, shipments from multiple

Table 6: Results on observation level of subset '31.7420005798339 , 118.861999511718'

Cluster label	Original ETD	ID	Weight	Emission
0	2021-10-09	8339378736-NANJING-CN–CN–NL-5705-HELMOND-NL	1600.0	0.52
0	2021-10-09	8339378736-NANJING-CN–CN–SE-26151-Landskrona-SE	1600.0	0.52
0	2021-10-08	5390797506-NANJING-CN–CN–NL-5705-HELMOND-NL	18050.0	5.90
0	2021-10-08	5390797506-NANJING-CN–CN–SE-26151-Landskrona-SE	18050.0	5.90
1	2021-06-21	8994690935-NANJING-CN–CN–NL-5705-HELMOND-NL	8200.0	2.68
1	2021-06-21	8994690935-NANJING-CN–CN–SE-26151-Landskrona-SE	8200.0	2.68

origin points are bundled together for efficiency. However, since legs between these origin points do not yet exist, a separate smaller network must be created to make sure consolidation for these specific shipments is possible. Due to limitations regarding the time, we could not create this network ourselves and will therefore propose these new legs as suggestions. Section 4.1.6 explains how we deal with these extra legs and which assumptions are made. Noticeably, the cluster labels in Table 7 are not consecutive, which is a result of the removal of empty clusters. We have chosen not to replace the cluster numbers for clarity and coherence. This way, it can easier be compared to the results on the observation level. Table 8 shows the results on the observation level, where the IDs of each observation are not fully shown because of their length. It contains part of the clusters in Table 7. Clusters 130 and 160 have close ETA values, but are not clustered together due to their cluster centroids, even though their shipments will fit in the same cluster. On the other hand, clusters 130 and 137 have the same cluster centroid, but cannot be clustered together due to ETA values, even though their shipments will fit in the same cluster. These illustrations show how well the algorithm handles the constraints, and why certain shipments cannot be clustered together. Noticeably, some clusters contain only one shipment. These shipments can unfortunately not be consolidated with the existing shipments from the company itself.

Table 7: Results on cluster level of subset '57.7 , 11.95'

. Cluster label	SubLatFrom	SubLonFrom	Cluster size	Maximum Capacity
0	24.45000	118.033333	59450.0	2175000.0
1	36.11667	120.300000	2576467.0	2676000.0
2	18.90000	72.816667	284200.0	2175000.0
...	...	...	...	...
130	31.25000	121.500000	920500.0	2658000.0
137	31.25000	121.500000	1103833.0	2658000.0
143	22.21667	91.800000	192333.0	2175000.0
145	22.21667	91.800000	39650.0	2175000.0
160	24.80000	66.983333	50000.0	2175000.0

Table 8: Results on observation level of subset '57.7 , 11.95'

Cluster label	Original ETD	ID	Weight	Emission
...	...	...	...	...
130	2021-08-17	FLEX-1216938-Shanghai, China-CN–CN-Gothenburg...	920500.0	1990.83
137	2021-11-10	FLEX-1299227-Shanghai, China-CN–CN-Gothenburg...	1103833.0	1990.83
143	2021-10-06	FLEX-1310803-Chattogram, Bangladesh-BD–BD-Gothenburg...	192333.0	328.02
145	2021-09-06	FLEX-1288165-Chattogram, Bangladesh-BD–BD-Gothenburg...	39650.0	67.62
160	2021-08-13	FLEX-1254190-Karachi, Pakistan-PK–PK-Gothenburg...	5000.0	66.32

***Index*='Third'** There is only one subset that has a leg index equal to 'Third'. The transportation mode is ocean and the sub-destination '56.1619676005914 , 14.8199431724983' corresponds with Karlshamn, Sweden. This subset takes approximately 5 seconds to run and has 198 observations, resulting in 18 clusters, which is a significant reduction. Clustering is only done based on the capacity and time restrictions, and not on origin points that are close by. This is because there only is one origin point in this subset, which makes the clustering process more simple.

***Index*='Last' or 'Direct'** There are 7 unique subsets where the leg index of the observations equals 'Last' or 'Direct'. Direct shipments only consist of one leg and can therefore use the original ETD dates as ETA, instead of the generated dates. The subsets have lengths that vary between 1 and 10.332 observations. Unfortunately, for the dataset containing 10,332 observations, the algorithm takes more than 100 hours to run, making it impracticable to use. As a result, we have decided to divide this big subset into smaller subsets. We do this based on origin countries, such that the shipments of each subset of this big subset, are dispatched from one country. In the remainder of the paper, we will refer to this subset as the big subset, and further elaboration will be given after discussing the final datasets. The second largest dataset has 1,037 observations and takes 37 minutes and 6 seconds to run, which is already a large difference. Table 9 shows the results of several clusters for the subset with key '51.4619706999999 , 5.6788948', which corresponds to the destination of Helmond, the Netherlands. This subset has the transportation mode parcel and has only one unique origin point, which implies that this subset corresponds with one unique leg in the dataset. We see that for cluster 32, the cluster size is bigger than the maximum capacity, which is not allowed. For this dataset, the error message 'The final cluster size is too large for cluster 32. Some shipments are too large for the empty clusters or the number of clusters is incorrect.' is given. After examining the output further by analyzing Table 10, we see that cluster 32 consists of one large shipment with a shipment weight of 16250 kg. The clusters with a larger maximum capacity, which is equal to 76200 kg in this case, are already filled with other shipments and are therefore unavailable. This means that this shipment had no other choice than to be put in a cluster with a smaller capacity. If this happens and the error message appears, the company can choose to split the shipment in several smaller shipments, such that it can be divided over multiple trucks, or it can reallocate the shipment to a different vehicle or container if one is available.

Table 9: Results on cluster level of subset '51.4619706999999 , 5.6788948'

Cluster label	SubLatFrom	SubLonFrom	Cluster size	Maximum Capacity
0	51.4501	5.37453	14600.0	15000.0
1	51.4501	5.37453	73250.0	76200.0
2	51.4501	5.37453	72900.0	76200.0
...	...	...	...	...
30	51.4501	5.37453	33700.0	76200.0
32	51.4501	5.37453	16250.0	15000.0
...	...	...	...	...
62	51.4501	5.37453	9050.0	15000.0
63	51.4501	5.37453	5550.0	15000.0

Table 10: Results on observation level of subset '51.4619706999999 , 5.6788948'

Cluster label	Original ETD	ID	Weight	Emission
...	...	...	...	...
30	31/08/2021	9281418845-HEFEI-CN–CN–NL-5705-HELMOND-NL	1250.0	0.30
30	29/08/2021	8691075956314001-JIAXING-CN–CN–NL-5705-HELMOND-NL	11100.0	2.62
30	29/08/2021	9602529366314001-HANGZHOU-CN–CN–NL-5705-HELMOND-NL	6450.0	1.52
30	03/09/2021	3286243516-SHENZHEN-CN–CN–NL-5705-HELMOND-NL	9200.0	2.17
30	31/08/2021	8691273334-HEFEI-CN–CN–NL-5705-HELMOND-NL	5700.0	1.35
32	03/01/2021	6760878666-BEIJING-CN–CN–NL-5705-HELMOND-NL	16250.0	3.84
...	...	...	...	...

The big subset The big subset contains 10,332 observations. As stated before, this dataset is split based on origin countries, resulting in multiple subsets. This does not mean that each subset contains only one origin point. It means that all origin points in one subset are from one country. The shipments of each subset of the big subset, are now dispatched from one country and are transported to the same (sub-)destination. Table 11 and Table 12 show the results of a subset on cluster and observation level respectively. This subset has 8 observations and consists of shipments that have Lithuania (LT) as the origin country, resulting in 6 clusters. We notice that the clustering results are not optimal. Some shipments are not consolidated even though their consolidation would still meet the capacity and time constraints since the capacity of the clusters is by no means fully utilized, while the ETD is for most shipments the same. Moreover, the origin points are also close to one another. Clusters 0, 1, 3 and 6 can for instance be combined for a more efficient consolidation. We suspect that the clustering algorithm does not cluster these shipments since the origins are still different, even though their similarity. This may be solved by lowering the number of clusters and thereby forcing consolidation. A possible approach to accomplish this is to incorporate a penalty for each extra cluster in the MCF network. This way, the algorithm may automatically use a lower number of clusters while still satisfying the capacity and time constraints.

Table 11: Results on cluster level of 'LT'

Cluster label	SubLatFrom	SubLonFrom	Cluster size	Maximum Capacity
0	54.687156	25.279651	200.0	15000.0
1	54.967093	23.920942	350.0	15000.0
2	54.687156	25.279651	100.0	15000.0
3	54.678393	25.286509	100.0	15000.0
6	54.687156	25.279651	300.0	15000.0
7	54.687156	25.279651	100.0	15000.0

Table 12: Results on observation level of 'LT'

Cluster label	Original ETD	ID	Weight	Emission
0	2021-09-01	1Z758AA99901633196-VILNIUS-LT–LT–PL-...	100.0	0.13
0	2021-09-01	1Z758AA99901633196-VILNIUS-LT–LT–PL-...	100.0	0.13
1	2021-09-01	1Z758AA99925394945-DOMEIKAVOS-LT–LT...	350.0	0.44
2	2021-10-01	1Z758AA99901633196-VILNIUS-LT–LT–PL-...	100.0	0.13
3	2021-09-01	1Z758AA99904210664-SENAMIESCIO-LT–LT...	100.0	0.13
6	2021-09-01	1Z758AA99901633196-VILNIUS-LT–LT–PL-...	200.0	0.26
6	2021-09-01	1Z758AA99901633196-VILNIUS-LT–LT–PL-...	100.0	0.13
7	2021-12-01	1Z758AA99901633196-VILNIUS-LT–LT–PL-...	100.0	0.13

5.2 Emission reduction

Table 13 shows the results regarding the emission reduction for the subsets where the leg index equals 'First'. The table consists of multiple columns, where the sub-destination shows the longitude and latitude coordination points of the location where the shipments are transported. The number of vehicles used in the subset, which is the same as the original size of the subset, is shown as the old size, while the new size shows the number of vehicles needed after consolidation. Furthermore, the CO_2 emissions before and after consolidation in kg are presented, which are the old emission and new emission respectively. The transportation mode used in the subset is also shown, which is important for the explanation of the emission reduction. Lastly, the run time is shown, which is in the format of *hh:mm:ss.s*. These run times should be interpreted as approximate run times and can slightly differ per execution.

For 23 subsets, the number of clusters is lower than the number of observations, indicating that consolidation is possible. The total run time of this algorithm for these subsets is 37.73 seconds, which is approximately 1.51 seconds per subset. It is noticeable that the CO_2 emissions remain the same, even though shipments have been consolidated. This is mostly due to emission calculation of parcel. The emission calculation of parcel is dependent on the distance and shipment weight. The legs in the dataset are predetermined, which makes the distance also predetermined and thus fixed. Furthermore, none of these subsets have different origins clustered together, thus no new legs were created. The shipment weight does increase for the consolidated shipments, leading to a lower relative emission per shipment, but this does not change the final emission. The emission calculation does not account for the vehicle weight or other aspects related to the vehicle, which subsequently means that the number of vehicles does not directly matter for the emission calculation. Since the emission calculation of parcel is mostly dependent on the distance and weight, the total CO_2 emissions remain the same, even though consolidation has taken place. Nevertheless, the vehicle utilization does improve. In the original situation, only 37.8% of the maximum vehicle capacity was used on average. After consolidation, this percentage increased to 52.4%. We note that this is still not optimal, but it is a substantial improvement compared to the original situation.

Table 13: Emission results for *Index* = 'First'

Sub-destination	Old size	New size	Old emission	New emission	Transportation	Run time
22.308901 , 113.915001	2	1	24.05	24.05	Parcel	00:00:00.04
22.6392993927001 , 113.810997009277	47	39	115.55	115.55	Parcel	00:00:12.52
23.0499992371 , 114.599998474	4	2	27.71	27.71	Parcel	00:00:00.05
23.0832996367999 , 113.069999695	13	10	41.53	41.53	Parcel	00:00:00.16
23.3924007415771 , 113.299003601074	8	6	20.76	20.76	Parcel	00:00:00.17
24.5440006256103 , 118.12799835205	4	3	1.76	1.76	Parcel	00:00:00.13
24.7964 , 118.589996	14	12	5.54	5.54	Parcel	00:00:00.70
25.9351005554199 , 119.66300201416	2	2	13.12	13.12	Parcel	00:00:00.09
27.912201 , 120.851997	6	5	11.53	11.53	Parcel	00:00:00.16
29.3446998596 , 120.03199768100001	10	9	4.56	4.56	Parcel	00:00:00.30
30.2294998168945 , 120.43399810791	24	14	62.57	62.57	Parcel	00:00:00.84
31.1979007720947 , 121.335998535156	15	11	28.04	28.04	Parcel	00:00:00.44
31.2631 , 120.401001	9	8	45.99	45.99	Parcel	00:00:00.23
31.7420005798339 , 118.861999511718	6	2	18.19	18.19	Parcel	00:00:00.10
31.7800006866455 , 117.297996520996	10	6	5.03	5.03	Parcel	00:00:00.16
36.2661018372 , 120.374000549	58	30	68.63	68.63	Parcel	0:00:20.39
38.2924003601 , 27.156999588	6	3	1.32	1.32	Parcel	00:00:00.08
39.7827987670898 , 116.388000488281	10	6	41.06	41.06	Parcel	00:00:00.17
40.2551994324 , 29.5625991821	4	3	18.47	18.47	Parcel	00:00:00.10
40.976898 , 28.8146	18	10	16.03	16.03	Parcel	00:00:00.24
41.138198852539 , 27.9190998077392	2	1	20.51	20.51	Parcel	00:00:00.03
41.275278 , 28.751944	8	4	7.77	7.77	Parcel	00:00:00.09
44.020302 , 12.6117	2	1	0.01	0.01	Parcel	00:00:00.03
44.5354 , 11.2887	2	1	0.02	0.02	Parcel	00:00:00.03
52.1656990051 , 20.9671001433999	2	2	1.42	1.42	Parcel	00:00:00.09

Table 14 shows the results regarding the emission reduction for the subsets where the leg index equals 'Second'. The total run time for these subsets is 43 minutes and 35 seconds, which is approximately 3 minutes and 38 seconds per subset. This is significantly longer than the previous run time, which is due to the larger subsets in the dataset. The largest subset takes about 27 minutes to run and the second largest subset about 14 minutes, while the smaller subsets only take seconds. For 11 subsets consolidation is possible. However, the remaining subsets still show an emission reduction. This is due to the fact that the container types of these subsets have changed. The transportation mode ocean has various container types, which gives us flexibility in choosing an appropriate container. Again, due to the specification by the GLEC framework (Greene & Lewis, 2019), most emission calculations are not dependent on vehicle characteristics or the number of vehicles. The significant reduction in the number of vehicles needed does therefore not necessarily lead to a significant emission reduction. We see that the subsets with transportation mode rail show *no* emission reduction. However, almost all subsets with transportation mode air do show an emission reduction. In these subsets, shipments of different origins are clustered together. Since these shipments first have to be shipped to one main origin, this means that part of the original distance is now covered by road for several shipments. The main origin in this case is closer to the (sub-)destination, which means that the total distance travelled by air is shorter. The CO_2 emission is therefore substantially reduced. However, since transportation by air is the least sustainable way of transportation, the total emission remains high. In the original situation, only 7.5% of the maximum vehicle capacity was used on average. After consolidation, this percentage increased to 17.1%. This is again an improvement compared to the original situation.

Table 14: Emission results for *Index* = 'Second'

Sub-destination	Old size	New size	Old emission	New emission	Transportation	Run time
51.23333 , 4.466667	1	1	1911.15	1234.11	Ocean	00:00:00.02
51.382520008315 , 6.68178509781164	178	21	35586.81	35586.81	Rail	00:00:04.17
51.4500999451 , 5.37452983856	144	100	55662.35	55474.29	Air	00:02:06.69
51.91667 , 4.5	452	176	370106.02	346877.16	Ocean	00:27:14.37
52.308601 , 4.76389	216	123	932414.70	930863.97	Air	00:05:48.26
52.36667 , 4.883333	1	1	1774.30	1250.29	Ocean	00:00:00.02
53.55 , 9.983333	6	3	10671.04	10547.37	Ocean	00:00:00.52
54.4842056346856 , 13.5854501232011	198	20	39093.45	39093.45	Rail	00:00:04.96
55.536305364 , 13.3761978149	161	91	366523.51	366284.34	Air	00:02:44.18
56.296100616455 , 12.8471002578735	143	103	48132.38	47928.88	Air	00:02:01.56
57.662799835205 , 12.279800415039	18	16	61481.96	61481.96	Air	00:00:01.54
57.7 , 11.95	309	105	179494.47	167337.85	Ocean	00:13:58.90

As before, there is only one subset where the leg index equals 'Third'. Unfortunately, there is no emission reduction for this subset, even though consolidation takes place. However, the utilization of vehicle capacities has improved. In the original situation, only 4.6% of the maximum vehicle capacity was used on average. After consolidation, this has increased to 51.2%, which is a significant improvement.

Table 15 shows the results regarding the emission reduction for the subsets where the leg index equals 'Last' or 'Direct', where the big subset with 10.377 is removed and evaluated later on. The total run time for these subsets is 1 hour and 9 minutes, which is approximately 9 minutes and 52 seconds per subset. For 4 subsets, consolidation is possible. As stated before, we do not see any emission reduction for the subsets with transportation mode parcel. However, for the subsets with transportation mode road, we see a substantial difference. The emission calculation for road shipments is largely dependent on the energy uses, which is in turn dependent on the loaded distance, which explains this result. Consolidated shipments lead to improved use of the vehicle capacity, which sequentially leads to a higher loaded distance. This results in lower emissions and a lower relative emission per shipment. Furthermore, the utilization of vehicle capacities has again improved. Before consolidation, the vehicles were using only 14.2% of their maximum capacities on average. After consolidation, this increased to 85.8%, which is a large improvement.

Table 15: Emission results for *Index* = 'Last' or 'Direct'

Sub-destination	Old size	New size	Old emission	New emission	Transportation	Run time
50.7281655379089 , 4.239441284082	1	1	26.35	26.35	Road	00:00:00.02
51.4619706999999 , 5.6788948	142	49	241.64	241.64	Parcel	00:00:03.71
51.4658083 , 5.6777953	1	1	1.98	1.98	Parcel	00:00:00.02
51.4792547 , 5.65700959999999	1137	137	804251.08	350779.38	Road	00:38:16.14
55.8703477 , 12.8300802	933	83	535547.45	218549.38	Road	00:31:08.31
55.8789960999999 , 12.8750083	120	46	328.63	328.63	Parcel	00:00:03.09
55.8790871 , 12.8571104	1	1	2.98	2.98	Parcel	00:00:00.02

The big subset took an unusually long time to run, despite the extra subsets that are made based on the origin countries. The total run time is equal to 13 hours and 34 minutes, which is approximately 32 minutes and 35 seconds per subset. If we compare these run times with the previous run times, we see that datasets of the same lengths now take a much longer time. The subset with leg index equal to 'Direct' and key '51.4792547 , 5.65700959999999' has 1,137 observations and only took 38 minutes and 16 seconds, while the subset of the Netherlands has 1.446 observations and takes 8 hours 29 minutes and 48 seconds to run. We do not have an explanation for this yet, but it is something worth investigating in the near future. Due to these longer run times, we have decided to leave out two bigger subsets. These subsets correspond with the origin countries Germany and Denmark, which have a size of 3,159 and 3,040 observations respectively. Table 16 shows the emission results of the remaining subsets of the big subset. The results are similar to what we have seen before. The transportation mode of this dataset is parcel, which means it is unlikely to see any emission reduction, even though consolidation takes place. However, for the Netherlands, we do see a small emission reduction, which is due to the consolidation of multiple origin points. The distance between these origin points was smaller than the original distances from the origin points to the destination point. As a result, the total CO_2 emission is slightly reduced. Before consolidation 1.4% of the maximum capacity was used on average. After consolidation, this increased to 2.3%, which is a small improvement. This could be the result of the inefficient clustering process of the big subset, as explained in Section 5.1.

Summarizing all results gives us the following main findings. In the original situation, 14,979 vehicles were used to transport the shipments to their final destinations. Furthermore, 13.1% of the vehicle capacities were utilized on average. The total emission of this network is equal to 3,447,679.09 kg. After consolidation, the number of vehicles has been reduced to 9.894, while the average vehicle utilization has increased to 41.8%. The total emissions of the consolidated network are now 2,323,806.88 kg, which is a reduction of 32.6%. These results show that consolidation has a substantial impact on the CO_2 emissions, but also leads to an improved vehicle utilization.

Table 16: Emission results for the big subset

Origin country	Old size	New size	Old emission	New emission	Transportation	Run time
Austria	319	188	80.18	80.18	Parcel	00:13:48.72
Belgium	528	404	142.64	142.64	Parcel	00:53:18.18
Switzerland	1	1	0.14	0.14	Parcel	00:00:00.05
Czechia	8	6	1.70	1.70	Parcel	00:00:00.08
Estonia	8	6	3.44	3.44	Parcel	00:00:00.10
Spain	11	9	5.64	5.64	Parcel	00:00:00.10
Finland	83	51	30.09	30.09	Parcel	00:00:30.22
France	518	421	166.92	166.92	Parcel	00:53:54.95
Greece	13	10	6.03	6.03	Parcel	00:00:00.11
Croatia	7	6	1.63	1.63	Parcel	00:00:00.08
Hungary	14	7	4.86	4.86	Parcel	00:00:00.98
Ireland	11	9	2.90	2.90	Parcel	00:00:00.75
Italy	70	64	26.29	26.29	Parcel	00:00:01.25
Lithuania	8	6	1.49	1.49	Parcel	00:00:00.55
Luxembourg	13	12	4.48	4.48	Parcel	00:00:00.11
Latvia	9	8	1.96	1.96	Parcel	00:00:01.70
The Netherlands	1446	687	<u>334.48</u>	<u>334.46</u>	Parcel	08:29:47.51
Poland	101	85	15.85	15.85	Parcel	00:00:47.42
Portugal	6	5	2.91	2.91	Parcel	00:00:00.05
Romania	3	2	1.23	1.23	Parcel	00:00:00.05
Sweden	947	430	304.15	304.15	Parcel	03:01:42.90
Slovenia	3	3	0.74	0.74	Parcel	00:00:00.05
Slovakia	6	6	2.10	2.10	Parcel	00:00:00.06

5.3 FTL/FCL

If we are dealing with LTL or LCL containers other companies are making use of the container as well. However, this is practically not preferred due to various reasons. As stated before, for freight consolidation to function effectively, a network of clients, carriers, processing facilities, and lanes must be coordinated. Explicit and transparent agreements need to be made concerning risk accountability, which can generate extra expenses. If a container is booked as FTL or FCL, these extra expenses do not occur as the container is dedicated to the shipments of the company itself. Furthermore, there is more certainty about capacities because the company is not dependent on other companies. When LTL or LCL containers are booked, it may happen that the vehicle capacity is not large enough to transport the complete shipment. In terms of CO_2 emissions, FTL and FCL can reduce the relative emission per kg for a company. Having a vehicle dedicated to your own shipments gives you control about how to fill it. With LTL or LCL vehicles, you do not always know how the vehicle capacity is utilized, which could result in a higher relative emission per kg. Furthermore, the ocean and road emission calculations are also dependent on whether a container is booked as FTL/FCL or LTL/LCL, see Figure 7 in Appendix B. As a result, we have decided to run the algorithm again, but this time treating all shipments as FTL or FCL, which implies that all vehicle capacities are now fully available. We will only examine the transportation modes ocean, rail, and road, considering the type of container is only applicable to these transportation modes.

Table 17 shows the sub-destinations for which the results have changed when treating all shipments as FTL or FCL. There are not many differences regarding emission reduction, but we do see that the new size changes. This means that treating all shipments as FTL or FCL leads to even fewer vehicles needed for transportation. Furthermore, we see that the run time has decreased. For larger subsets, this can save up to 10 minutes, but for smaller subsets this run time reduction is insignificant. Treating all shipments as FTL or FCL does lead to considerably similar results as the situation where LTL or LCL shipments are included. This is not surprising, since there are not many LTL or LCL shipments. Only 8.9% of all shipments are LTL or LCL, which means that merely a few vehicle capacities have changed.

Table 17: Emission results for the FTL/FCL case

Sub-destination	Old size	New size	Old emission	New emission	Transportation	Run time
51.382520008315 , 6.68178509781164	178	16	35586.81	35586.81	Rail	00:00:00.04
51.91667 , 4.5	452	148	370106.02	346841.78	Ocean	00:22:20.79
54.4842056346856 , 13.5854501232011	198	19	39093.45	39093.45	Rail	00:00:03.58
56.161967600591 , 14.8199431724983	198	14	1335.52	1335.2	Ocean	00:00:02.90
51.4792547 , 5.65700959999999	1137	136	804251.08	350690.43	Road	00:29:50.97

6 Conclusion

In this paper, we have constructed a constrained clustering algorithm to answer the question: "Is freight consolidation possible for the current network of the company while taking time and vehicle capacity constraints into account?". The main goal is to examine the effects of freight consolidation on the CO_2 emissions within the fashion, textile, and sports (FTS) sector and investigate if this leads to an emission reduction. An additional objective is to increase the vehicle load and thereby improve the utilization of vehicles by consolidating shipments. Freight consolidation via cluster analysis has not been intensively researched, especially in the sustainability context. For this reason, costs are not taken into account during this research to remain focused on the sustainable outlook. The algorithm in this paper makes use of instance-level constraints in the form of time constraints and feature-level constraints in the form of cluster size constraints. The time constraints are incorporated through cannot-link (CL) constraints, based on the algorithm of Švehla (2018). The constraints regarding the cluster sizes correspond with the maximum capacities of the vehicles that transport the shipments. These are incorporated in the Minimum Cost Flow (MCF) linear network optimization problem, based on the algorithm by Levy-Kramer and Klaber (2022).

The constrained clustering algorithm has been applied to subsets of the dataset provided by Sustaining Supply Chains (SSC), where the shipment routes in this dataset are predetermined and known. In the original situation, 14,979 vehicles were used to transport the shipments to their final destinations, where 13.1% of the vehicle capacities were utilized on average. The total CO_2 emission of this network is equal to 3,447,679.09 kg. After consolidation, we can conclude that the number of vehicles has been reduced to

9,894, while the average vehicle utilization has increased to 41.8%. The total emissions of the consolidated network are now 2,323,806.88 kg, which is a reduction of 32.6%. These results show that consolidation has a substantial impact on the CO_2 emissions, but also on the efficiency of the transportation network since fewer vehicles need to be used to transport the same number of shipments, leading to an improvement in vehicle utilization.

The results show that consolidation is possible for 40 of the 46 subsets. This means that the number of vehicles needed for transportation can be substantially reduced, but it does not automatically mean that the CO_2 emission of every subset has been reduced. The emission calculations for the transportation modes parcel and rail are mostly dependent on the distance and shipment weight, and not on the characteristics of the vehicle itself, such as the vehicle weight. This means that although shipments are consolidated, the CO_2 emissions are *not* reduced and remain the same. The emission calculations of the transportation modes air, ocean, and road are dependent on more factors. For these transportation modes, we do see an emission reduction, especially for the transportation mode road. The emission calculation for road is dependent on the total energy used, which is in turn dependent on the payload of a vehicle. Consolidation highly influences the payload, which explains why the emission reduction for this transportation mode is remarkable. Furthermore, we see that if we treat all shipments as FTL or FCL the results do not change much. The main difference with treating all shipments as FTL or FCL is that even fewer vehicles are needed for transportation and that the run times of the algorithm have decreased. The emissions for the transportation modes ocean and road show a further reduction, but this difference is considerably small.

Unfortunately, the constrained clustering algorithm does not always satisfy the capacity constraints. This could occur when the vehicles with a larger capacity are not available anymore, and the available vehicles are not large enough to fit a shipment. The vehicles with larger capacities are in this case filled with several smaller shipments. In the current schedule provided by the company this does not happen because this shipment was originally in a larger vehicle, and the smaller shipments were in other separate vehicles. Considering that the results of the algorithm do not replace the current decision system, and rather complement it, this should not lead to difficulties. If the capacity constraints are not satisfied, the company can always decide to rearrange the shipment(s) or wait for the availability of another vehicle or container. A warning message will be returned for the cluster(s) that are not content with the capacity constraints which can be used to take further action. Furthermore, the final consolidation is not always optimal. Some shipments are not consolidated even though their consolidation would still satisfy the capacity and time constraints. We believe that the clustering algorithm does not cluster these shipments because pre-specified the number of clusters is too large. This gives the algorithm a large feasible region, which may lead to a feasible, but not an optimal solution. A possible approach to accomplish a lower number of clusters is to incorporate a penalty for each extra cluster that the algorithm uses in the MCF network during the consolidation process. This way, the algorithm may automatically use a lower number of clusters while still satisfying the capacity and time constraints.

6.1 Limitations ans suggestions for future research

The constrained clustering algorithm has several limitations. In the first step of the algorithm, the MCF network assigns observations to clusters, resulting in cluster labels for every observation. However, these cluster labels are rewritten in the next step when incorporating the time constraints. This approach is not efficient, since the clustering structure of the MCF network is broken, which can lead to a violation of the capacity constraints. More efficient would be to check if the clusters that came out of the MCF network, therefore satisfying the capacity constraints, satisfy the time constraints. If this is the case, we can keep these clusters apart and put the remaining clusters that do not satisfy these constraints can again be put in the MCF network in the next iteration. This way, the clustering structure of the MCF network is taken into account and not completely rewritten. Due to limitations regarding the time, it was not possible to incorporate this ourselves. Another suggestion is to incorporate the time constraints into the MCF network, instead of evaluating this separately. This could lead to a more efficient approach since both types of constraints are used next to one another, and not after one another. This way, the final rearrangement of the clusters is not required anymore, which again solves an inefficiency. Even though the final ad-hoc approach for rearrangement fulfilled its function, it was not optimal. It selects a feasible cluster arrangement by deleting the largest shipment and placing it in an empty cluster, which is not optimal. Nonetheless, incorporating the time constraints into the MCF network can be highly complicated and can therefore take a considerable amount of time, which is why we have chosen not to do this ourselves and propose it as a future research area.

Another point of attention is the determination of the number of clusters. The current approach is based on groundless assumptions, especially when the dataset has a large number of observations. Existing methods such as the elbow method or the gap statistic are not optimal since these clusters are mostly based on the proper separation of the observations or the within- and between-cluster distances. The number of clusters for this algorithm needs to be determined based on the optimal utilization of the cluster capacity, while minimizing emissions. In other words, the least number of clusters needed to minimize the emissions while maximizing the vehicle capacity. A suggestion for future research is to find a more valid approach to determining the number of clusters and comparing this with the existing methods.

Currently, the algorithm is constructed in such a way that it only works for this specific dataset and its corresponding values. A direction for further work could be examining the performance of the algorithm for various datasets to examine the effects of consolidation in different settings. Furthermore, the algorithm cannot handle large datasets. Datasets with more than 5.000 observations take a remarkably long time to run, it can take up to several days. This could be a problem related to the Central Processing Unit (CPU) or the Random Access Memory (RAM), which can be solved by running the algorithm on a separate server with a larger capability. Due to limited resources, it was not possible to do this ourselves.

The unavailability of ETA and ETD dates can also be seen as a limitation. We have made assumptions regarding the duration of transportation modes and stated for simplicity that these duration values are fixed. This is not realistic in practice. A suggestion for future research is to provide a more realistic approach to determining the duration of all routes since these assumptions have a large influence on the time constraints. Another limitation is the unavailability of costs, which has made it inaccessible to provide a complete overview of the effects of consolidation. Consolidation leads to a lesser number of vehicles needed for transport. This does not always lead to an emission reduction with the equations used in this paper, but it can lead to a substantial reduction in expenses. Due to the absence of this information, we cannot identify this impact. Including expenses in the analysis may help in the trade-off between decarbonization and economic interest, since this research area is still young with various opportunities for improvement.

In addition, we suggest investigating the flexibility of the transportation modes. In the current approach, the transportation modes are fixed for a given leg. However, given that each transportation mode has different emissions, it may be that using another mode may lead to lower emissions. This allows for a larger feasible region, because of the additional possibilities for consolidation. Moreover, time constraints may more easily be feasible in the case of fixed departure dates. If a shipment cannot make a departure date for a specific transportation mode, it will have to wait until the next departure before it can be transported. In the case of flexible transportation modes, this does not have to hold since it can be shipped with another transportation mode. Adding other transportation modes is possible by adding duplicate rows to the dataset that include this other mode. These rows specify the same leg covered by a different transportation mode. Adding these rows allows the clustering algorithm to make use of several transportation modes for one leg, while still minimizing the CO_2 emissions. Next, we suggest to extend the time constraints of the algorithm with must-link (ML) constraints. These ML constraints specify that certain shipments must be transported together. For the FTS sector, this can be the case when separate shipments belong to the same clothing collection and must arrive together.

Lastly, multiple clustering algorithms can be investigated regarding constrained clustering in the sustainability context. There exist soft clustering methods, where an observations can belong to multiple clusters with different probabilities. Fuzzy clustering makes it possible for shipments to be clustered in multiple ways, making it an insightful and promising method for freight consolidation.

7 References

Ali, A. M., & Söffker, D. (2018). Towards optimal power management of hybrid electric vehicles in real-time: A review on methods, challenges, and state-of-the-art solutions. *Energies*, *11*(3), 476.

Anderberg, M. R. (1973). Cluster analysis for applications. *Academic Press.*

Bektaş, T., Ehmke, J. F., Psaraftis, H. N., & Puchinger, J. (2019). The role of operational research in green freight transportation. *European Journal of Operational Research*, *274*(3), 807–823.

Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000). Constrained k-means clustering. *Microsoft Research, Redmond*, *20*(0), 0.

Cai, W., & Li, G. (2018). The drivers of eco-innovation and its impact on performance: Evidence from china. *Journal of Cleaner Production*, *176*, 110–118.

Caplice, C. G. (1996). *An optimization based bidding process: A new framework for shipper-carrier relationships* (Doctoral dissertation). Massachusetts Institute of Technology.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, *40*(1), 200–210. https://doi.org/https://doi.org/10.1016/j.eswa.2012.07.021

Change, I. C. et al. (2014). Mitigation of climate change. *Contribution of working group III to the fifth assessment report of the intergovernmental panel on climate change*, *1454*, 147.

Cortes, J. D., & Suzuki, Y. (2020). Vehicle routing with shipment consolidation. *International Journal of Production Economics*, *227*, 107622.

Creutzig, F., Jochem, P., Edelenbosch, O. Y., Mattauch, L., Vuuren, D. P. v., McCollum, D., & Minx, J. (2015). Transport: A roadblock to climate change mitigation? *Science*, *350*(6263), 911–912.

Davidson, I., & Ravi, S. (2007). The complexity of non-hierarchical clustering with instance and cluster level constraints. *Data mining and knowledge discovery*, *14*(1), 25–61.

Demir, E., Bektaş, T., & Laporte, G. (2011). A comparative analysis of several vehicle emission models for road freight transportation. *Transportation Research Part D: Transport and Environment*, *16*(5), 347–357.

Dondo, R., & Cerdá, J. (2007). A cluster-based optimization approach for the multi-depot heterogeneous fleet vehicle routing problem with time windows. *European journal of operational research*, *176*(3), 1478–1507.

Eltayeb, T. K., Zailani, S., & Ramayah, T. (2011). Green supply chain initiatives among certified companies in malaysia and environmental sustainability: Investigating the outcomes. *Resources, conservation and recycling*, *55*(5), 495–506.

Forum, I. T. (2019). *Itf transport outlook 2019*. https://doi.org/https://doi.org/https://doi.org/10.1787/transp_outlook-en-2019-en

Francis, P. M., Smilowitz, K. R., & Tzur, M. (2008). The period vehicle routing problem and its extensions. *The vehicle routing problem: Latest advances and new challenges* (pp. 73–102). Springer.

Garcıa-Granero, E. M., Piedra-Muñoz, L., & Galdeano-Gómez, E. (2018). Eco-innovation measurement: A review of firm performance indicators. *Journal of cleaner production*, *191*, 304–317.

González-Almagro, G., Luengo, J., Cano, J.-R., & Garcıa, S. (2020). Dils: Constrained clustering through dual iterative local search. *Computers & Operations Research*, *121*, 104979.

Greene, S., & Lewis, A. (2019). Global logistics emissions council framework for logistics emissions accounting and reporting. *Tech. Rep.*

Hansen, T., & Wang, C.-J. (2005). Support vector based battery state of charge estimator. *Journal of Power Sources*, *141*(2), 351–358.

IPCC, I. (2018). Summary for policymakers" in global warming of 1.5° c. an ipcc special report on the impacts of global warming of 1.5° c above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.

Jeffry, L., Ong, M. Y., Nomanbhay, S., Mofijur, M., Mubashir, M., & Show, P. L. (2021). Greenhouse gases utilization: A review. *Fuel*, *301*, 121017.

Kagawa, S., Suh, S., Hubacek, K., Wiedmann, T., Nansai, K., & Minx, J. (2015). Co2 emission clusters within global supply chain networks: Implications for climate change mitigation. *Global Environmental Change*, *35*, 486–496.

Kelle, P., Song, J., Jin, M., Schneider, H., & Claypool, C. (2019). Evaluation of operational and environmental sustainability tradeoffs in multimodal freight transportation planning [The Proceedings of the 19th International Symposium on Inventories]. *International Journal of Production Economics*, *209*, 411–420. https://doi.org/https://doi.org/10.1016/j.ijpe.2018.08.011

Kuyzu, G., Akyol, Ç. G., Ergun, Ö., & Savelsbergh, M. (2015). Bid price optimization for truckload carriers in simultaneous transportation procurement auctions. *Transportation Research Part B: Methodological*, *73*, 34–58.

Leleu, H. (2013). Shadow pricing of undesirable outputs in nonparametric analysis. *European Journal of Operational Research*, *231*(2), 474–480.

Levy-Kramer, J., & Klaber, M. (2022). K-means-constrained. https://github.com/joshlk/k-means-constrained.git

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *University of California Press.*

Marchet, G., Melacini, M., & Perotti, S. (2014). Environmental sustainability in logistics and freight transportation: A literature review and research agenda. *Journal of Manufacturing Technology Management.*

McKinnon, A. (2018). *Decarbonizing logistics: Distributing goods in a low carbon world.* Kogan Page Publishers.

McKinnon, A., & Edwards, J. (2010). Opportunities for improving vehicle utilization. *Green Logistics: Improving the environmental sustainability of logistics*, 195–213.

Mesa-Arango, R., & Ukkusuri, S. V. (2015). Demand clustering in freight logistics networks. *Transportation Research Part E: Logistics and Transportation Review*, *81*, 36–51.

MILES, E. (2009). Macy's and schneider national filling empty miles for sustainability and savings.

Mosavi, A., Salimi, M., Faizollahzadeh Ardabili, S., Rabczuk, T., Shamshirband, S., & Varkonyi-Koczy, A. R. (2019). State of the art of machine learning models in energy systems, a systematic review. *Energies*, *12*(7), 1301.

Muñoz-Villamizar, A., Santos, J., Montoya-Torres, J. R., & Jaca, C. (2018). Using oee to evaluate the effectiveness of urban freight transportation systems: A case study. *International Journal of Production Economics*, *197*, 232–242.

Orji, I. J., Kusi-Sarpong, S., Gupta, H., & Okwu, M. (2019). Evaluating challenges to implementing eco-innovation for freight logistics sustainability in nigeria. *Transportation Research Part A: Policy and Practice*, *129*, 288–305.

Özdamar, L., & Demir, O. (2012). A hierarchical clustering and routing procedure for large scale disaster relief logistics planning. *Transportation Research Part E: Logistics and Transportation Review*, *48*(3), 591–602.

Pan, S., Trentesaux, D., Ballot, E., & Huang, G. Q. (2019). Horizontal collaborative transport: Survey of solutions and practical implementation issues. *International Journal of Production Research*, *57*(15-16), 5340–5361.

Perron, L., & Furnon, V. (n.d.). *Or-tools* (Version 7.2). https://developers.google.com/optimization/

Punte, S., Tavasszy, L., Baeyens, A., & Lisa, F. (2019). A framework and process for the development of a roadmap towards zero emissions logistics 2050. *Alliance for Logistics Innovation through Collaboration in Europe (ALICE).*

Ritchie, H. (2020). Cars, planes, trains: Where do co2 emissions from transport come from? *Our World in Data*. https://ourworldindata.org/co2-emissions-from-transport

Ritchie, H., & Roser, M. (2020). Co and greenhouse gas emissions. *Our World in Data*. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al. (2022). Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, *55*(2), 1–96.

Schellnhuber, H.-J., van der Hoeven, M., Bastioli, C., Ekins, P., Jaczewska, B., Kux, B., Thimann, C., Tubiana, L., Wanngård, K., Slob, A., et al. (2018). Final report of the high-level panel of the european decarbonisation pathways initiative.

Schiele, H. (2008). Location, location: The geography of industry clusters. *Journal of business strategy.*

Schmidt, J., Brandle, E. M., & Kramer, S. (2011). Clustering with attribute-level constraints. *2011 IEEE 11th International Conference on Data Mining*, 1206–1211.

Schulz, S. F., & Blecken, A. (2010). Horizontal cooperation in disaster relief logistics: Benefits and impediments. *International Journal of Physical Distribution & Logistics Management.*

Sheffi, Y. (2012). *Logistics clusters: Delivering value and driving growth.* MIT press.

Sheffi, Y. (2013). Logistics-intensive clusters: Global competitiveness and regional growth. *Handbook of global logistics* (pp. 463–500). Springer.

Sifaleras, A. (2016). Minimum cost network flows: Problems, algorithms, and software. *Yugoslav Journal of Operations Research*, *23*(1).

SteadieSeifi, M., Dellaert, N. P., Nuijten, W., Van Woensel, T., & Raoufi, R. (2014). Multimodal freight transportation planning: A literature review. *European journal of operational research*, *233*(1), 1–15.

Švehla, J. (2018). Active-semi-supervised-clustering. https://github.com/datamole-ai/active-semi-supervised-clustering.git

Triki, C., Oprea, S., Beraldi, P., & Crainic, T. G. (2014). The stochastic bid generation problem in combinatorial transportation auctions. *European Journal of Operational Research*, *236*(3), 991–999.

Tryon, R. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Edwards brother, Incorporated, lithoprinters; publishers. https://books.google.nl/books?id=gsnrAAAAMAAJ

Tsoumakas, G. (2019). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, *52*(1), 441–447.

Union, E. (2001). Guidelines on the applicability of article 81 of the ec treaty to horizontal cooperation agreements. *European Commission Notice 2001/C 3*, *2*.

Vaillancourt, A. (2016). A theoretical framework for consolidation in humanitarian logistics. *Journal of Humanitarian Logistics and Supply Chain Management*.

Van Andel, B. (2018). A machine learning approach to shipment consolidation. *MaRBLe*, *2*.

Wagstaff, K. L. (2010). Constrained clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 220–221). Springer US. https://doi.org/10.1007/978-0-387-30164-8_163

Wang, S., Li, R., Evans, A., & Li, F. (2019). Electric vehicle load disaggregation based on limited activation matching pursuits. *Energy Procedia*, *158*, 2611–2616.

Xu, S. X., & Huang, G. Q. (2014). Efficient auctions for distributed transportation procurement. *Transportation Research Part B: Methodological*, *65*, 47–64.

Zhang, Y.-J., Hao, J.-F., & Song, J. (2016). The co2 emission efficiency, reduction potential and spatial clustering in china's industry: Evidence from the regional level. *Applied Energy*, *174*, 213–223.

Zhou, G., Hui, Y. V., & Liang, L. (2011). Strategic alliance in freight consolidation. *Transportation Research Part E: Logistics and Transportation Review*, *47*(1), 18–29. https://doi.org/https://doi.org/10.1016/j.tre.2010.07.002

Appendices

A Dataset Excel

A.1 Complete dataset

Sector	Type	Shipment ID	Transport Mode	Origin - Country Code (ISO)	Origin - City (Port / Terminal)	Origin - City Actual	Origin - Postal code Actual	Destination - Country Code (ISO)	Destination - City (Port / Terminal)	Destination - Final city
FTS	Inbound	FLEX-1431434	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1431413	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1431419	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1431425	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1427943	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1427933	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1427939	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1424217	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1427931	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1424270	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1424249	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1410934	AIR	CN		Shanghai, China			MMX - Malmo - Sweden	Landskrona
FTS	Inbound	FLEX-1410934	AIR	CN		Shanghai, China			MMX - Malmo - Sweden	Landskrona
FTS	Inbound	FLEX-1424258	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1410930	AIR	CN		Shanghai, China			AMS - Amsterdam - Netherlands	Helmond
FTS	Inbound	FLEX-1410930	AIR	CN		Shanghai, China			AMS - Amsterdam - Netherlands	Helmond
FTS	Inbound	FLEX-1424234	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1424194	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1415059	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1411406	ROAD	RO		Bucharest			Helmond	Helmond
FTS	Inbound	FLEX-1420686	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1415040	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1420664	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1411435	ROAD	RO		Bucharest			Landskrona	Landskrona
FTS	Inbound	FLEX-1415124	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1411324	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1405988	ROAD	TR		Istanbul			Helmond	Helmond
FTS	Inbound	FLEX-1415094	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1415071	ROAD	TR		Istanbul			Landskrona	Landskrona
FTS	Inbound	FLEX-1415086	ROAD	TR		Istanbul			Landskrona	Landskrona

Destination - Final Postal code	KGS	CBM	Booked as	Container	ETD	ETA	Incoterm	Shipment Method	Condition	Fuel	Carrier	Number of shipments	Number of items	Category	FromActualID
	1471,67	8.83	FTL	N/A	25/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1456,67	8.74	FTL	N/A	25/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1728,33	10.37	FTL	N/A	25/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1423,33	8.54	FTL	N/A	25/12/2021	29/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1551,67	9.31	FTL	N/A	22/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1711,67	10.27	FTL	N/A	22/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1063,33	6.38	FTL	N/A	22/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	8480	50.88	FTL	N/A	22/12/2021	30/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	993	4.8	FTL	N/A	22/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	10743,33	64.46	FTL	N/A	18/12/2021	29/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	980	5.88	FTL	N/A	18/12/2021	20/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	153,33	0.92	N/A	N/A	18/12/2021	30/12/2021	FCA	Deferred	Ambient	DieselD5	Flexport; China Cargo Airlines (CNY); Strait Air Transport				-Shanghai, China-CN
	153,33	0.92	N/A	N/A	18/12/2021	30/12/2021	FCA	Deferred	Ambient	DieselD5	Flexport; China Cargo Airlines (CNY); Strait Air Transport				-Shanghai, China-CN
	1303,33	7.82	FTL	N/A	18/12/2021	20/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	203,33	1.22	N/A	N/A	18/12/2021	29/12/2021	FCA	Standard	Ambient	DieselD5	Flexport; Bos Logistics B.V.; China Southern Airlines				-Shanghai, China-CN
	203,33	1.22	N/A	N/A	18/12/2021	29/12/2021	FCA	Standard	Ambient	DieselD5	Flexport; Bos Logistics B.V.; China Southern Airlines				-Shanghai, China-CN
	10768,33	64.61	FTL	N/A	18/12/2021	29/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1403,33	8.42	FTL	N/A	18/12/2021	22/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	6050	36.3	FTL	N/A	16/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	432	2.592	LTL	N/A	16/12/2021	21/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Bucharest-RO
	1046,3	5.66	FTL	N/A	15/12/2021	20/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	10240	61.44	FTL	N/A	15/12/2021	27/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1146	6.05	FTL	N/A	15/12/2021	21/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	352	2.112	LTL	N/A	15/12/2021	20/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Bucharest-RO
	6948,33	41.69	FTL	N/A	15/12/2021	28/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	10011,67	60.07	FTL	N/A	13/12/2021	22/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	10503,33	63.02	FTL	N/A	12/12/2021	21/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	10471,67	62.83	FTL	N/A	12/12/2021	22/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	1423,33	8.54	FTL	N/A	11/12/2021	15/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR
	816,67	4.9	FTL	N/A	11/12/2021	15/12/2021	EXW	N/A	Ambient	DieselD5	Flexport;				-Istanbul-TR

FromTerminalID	ToActualID	ToTerminalID	LaneID	Payment	Index	SubLane	SubLaneModality	Distance (km)	EmissionFactorGroup	CalculationType	SubLaneFromContinent	Cfactor
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-CN	-Landskrona-	MMX - Malmo - Sweden-	-Shanghai, China-CN--CN-MMX - Malmo - Sweden-	Full	Second	PVG-MMX	AIR	8243,907457	AirLong	Default	World	4500
-CN	-Landskrona-	MMX - Malmo - Sweden-	-Shanghai, China-CN--CN-MMX - Malmo - Sweden-	Full	Last	MMX--Landskrona-	ROAD	61,05	ROAD	Default	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-CN	-Helmond-	AMS - Amsterdam - Netherlands-	-Shanghai, China-CN--CN-AMS - Amsterdam - Neth	Full	Second	PVG-AMS	AIR	8910,452348	AirLong	Default	World	4500
-CN	-Helmond-	AMS - Amsterdam - Netherlands-	-Shanghai, China-CN--CN-AMS - Amsterdam - Neth	Full	Last	AMS--Helmond-	ROAD	130,05	ROAD	Default	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-RO	-Helmond-	Helmond-	-Bucharest-RO--RO-Helmond---Helmond-	Full	Direct	-Bucharest-RO--Helmond-	ROAD	2111,12	ROAD	Default	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-RO	-Landskrona-	Landskrona-	-Bucharest-RO--RO-Landskrona---Landskrona-	Full	Direct	-Bucharest-RO--Landskrona-	ROAD	2224,53	ROAD	Default	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Helmond-	Helmond-	-Istanbul-TR--TR-Helmond---Helmond-	Full	Direct	-Istanbul-TR--Helmond-	ROAD	2603,23	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500
-TR	-Landskrona-	Landskrona-	-Istanbul-TR--TR-Landskrona---Landskrona-	Full	Direct	-Istanbul-TR--Landskrona-	ROAD	2716,65	ROAD	Modelled	EU	4500

Figure 5: Snapshots of the Excel data

TEUkm	Tonkm	Service	SubLaneFromID	SubCityFrom	SubCountryFrom	SubLatFrom	SubLonFrom	SubLaneToID	SubCityTo	SubCountryTo	SubLatTo	SubLonTo	PaymentResponsibility	RoadEmissionCategory	EmissionFactorName	EmissionFactorRoadParcel
0	3998,012	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	3792,047	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	103,04829
0	4499,241	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	103,04829
0	3866,689	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	4215,344	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	4455,871	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	103,04829
0	2888,695	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	22075,39	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	MGV	RoadTruck<40t	103,04829
0	2585,007	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	103,04829
0	29185,87	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	MGV	RoadTruck<40t	102,0860251
0	2662,317	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	1264,038	Deferred	PVG			31,1434002	121,805	MMX			55,53630536	13,3762	TRUE	Van	AirLong	
0	9,360797	Deferred	MMX			55,5363054	13,3761978	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	680
0	3540,691	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	1811,762	Standard	PVG			31,1434002	121,805	AMS			52,308601	4,76389	TRUE	Van	AirLong	
0	26,44307	Standard	AMS			52,308601	4,76389	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	541,361015
0	28032,44	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	MGV	RoadTruck<40t	103,04829
0	3653,191	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	103,04829
0	15749,54	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	MGV	RoadTruck<40t	103,04829
0	912,0038	N/A	-Bucharest-RO			44,4497506	26,0699105	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	108,420933
0	2842,431	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	26657,08	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	MGV	RoadTruck<40t	103,04829
0	2983,302	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	Van	RoadTruck<40t	103,04829
0	783,0346	N/A	-Bucharest-RO			44,4497506	26,0699105	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	106,9719896
0	18876,18	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	MGV	RoadTruck<40t	102,0860251
0	27198,2	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	MGV	RoadTruck<40t	102,0860251
0	27342,58	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Helmond-	Helmond		51,4792547	5,65701	TRUE	MGV	RoadTruck<40t	103,04829
0	28447,86	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	MGV	RoadTruck<40t	102,0860251
0	3866,689	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251
0	2218,607	N/A	-Istanbul-TR	ISTANBUL	Turkey	41,0082376	28,9783589	-Landskrona-	Landskrona		55,8703477	12,83008	TRUE	Van	RoadTruck<40t	102,0860251

Actual distance road (km)	Empty distance road (km)	EnergyUsedFTL Empty (MJ)	EnergyUsedFTL Loaded (MJ)	EnergyUsedFTL Total (MJ)	EmissionFactorRoadModelled	Emission Factor	Unit	Emission (kg)
2852,4825	584,2434036	4790,79591	24213,79327	29004,58917	644,1894945	644,1894945	g/tonkm	2575,477526
2733,3915	559,8512711	4590,780423	23194,82614	27785,60657	650,6346849	650,6346849	g/tonkm	2467,237334
2733,3915	559,8512711	4590,780423	23340,4808	27931,26122	551,2421065	551,2421065	g/tonkm	2480,170814
2852,4825	584,2434036	4790,79591	24186,74581	28977,54172	665,4467265	665,4467265	g/tonkm	2573,075833
2852,4825	584,2434036	4790,79591	24258,5553	29049,35121	611,9196932	611,9196932	g/tonkm	2579,452194
2733,3915	559,8512711	4590,780423	23331,54828	27922,32871	556,4294427	556,4294427	g/tonkm	2479,377647
2852,4825	584,2434036	4790,79591	23985,31666	28776,11257	884,5480171	884,5480171	g/tonkm	2555,189827
2733,3915	559,8512711	4590,780423	26960,49167	31551,27209	126,9110763	126,9110763	g/tonkm	2801,611555
2733,3915	559,8512711	4590,780423	22946,2224	27537,00282	945,9015262	945,9015262	g/tonkm	2445,162435
2852,4825	584,2434036	4790,79591	29401,52266	34192,31857	104,0272197	104,0272197	g/tonkm	3036,124646
2852,4825	584,2434036	4790,79591	23938,69141	28729,48731	958,2065975	958,2065975	g/tonkm	2551,049714
						630	g/tonkm	796,3441482
64,1025	13,12942771	107,6613072	527,568464	635,2297713	6025,721927	80	g/tonkm	0,74886372
2852,4825	584,2434036	4790,79591	24119,60276	28910,39867	725,031783	725,031783	g/tonkm	2567,113831
						630	g/tonkm	1141,410234
136,5525	27,96858434	229,3423916	1125,176755	1354,519146	4548,459972	80	g/tonkm	2,11544532
2733,3915	559,8512711	4590,780423	28187,41471	32778,19513	103,8281667	103,8281667	g/tonkm	2910,556823
2733,3915	559,8512711	4590,780423	23166,22709	27757,00751	674,6699065	674,6699065	g/tonkm	2464,697866
2733,3915	559,8512711	4590,780423	25657,6101	30248,39052	170,5396637	170,5396637	g/tonkm	2685,921511
2216,676	454,0179759	3722,947402	18364,58091	22087,52832	2150,510159	80	g/tonkm	72,9603072
2852,4825	584,2434036	4790,79591	23975,78794	28766,58385	898,6476063	898,6476063	g/tonkm	2554,34372
2733,3915	559,8512711	4590,780423	27904,14252	32494,92294	108,2415644	108,2415644	g/tonkm	2885,403522
2733,3915	559,8512711	4590,780423	23028,25568	27619,03611	822,0578971	822,0578971	g/tonkm	2452,446623
2335,7565	478,4079578	3922,945254	19314,4783	23237,42356	2635,105999	80	g/tonkm	62,6427648
2852,4825	584,2434036	4790,79591	27278,12372	32068,91963	150,855535	150,855535	g/tonkm	2847,576337
2852,4825	584,2434036	4790,79591	28992,1403	33782,93621	110,2930695	110,2930695	g/tonkm	2999,773327
2733,3915	559,8512711	4590,780423	28045,33092	32636,11134	105,9863415	105,9863415	g/tonkm	2897,940419
2852,4825	584,2434036	4790,79591	29249,52199	34040,3179	106,2514868	106,2514868	g/tonkm	3022,627668
2852,4825	584,2434036	4790,79591	24186,74581	28977,54172	665,4467265	665,4467265	g/tonkm	2573,075833
2852,4825	584,2434036	4790,79591	23847,30412	28638,10003	1146,185621	1146,185621	g/tonkm	2542,934933

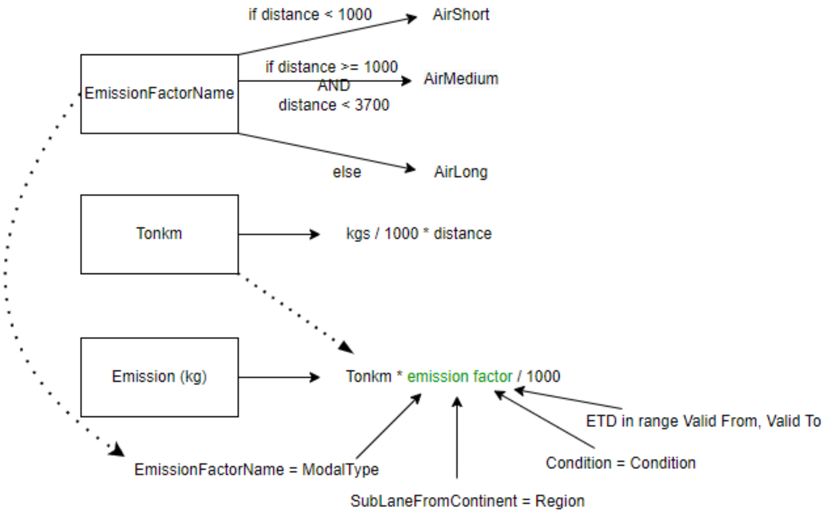
Figure 5: Snapshots of the Excel data

A.2 Example of a shipment in the dataset

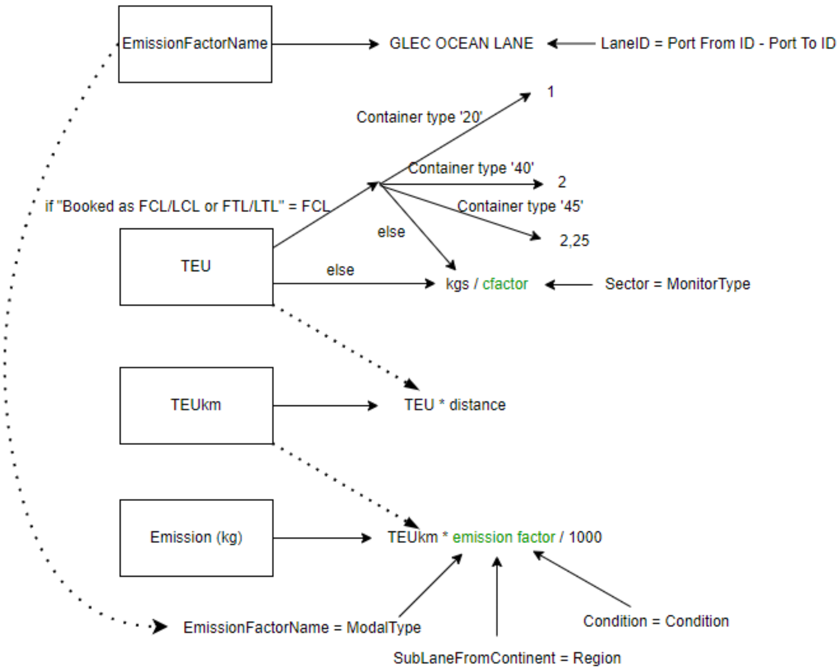
Shipment ID	Origin - City (Port / Terminal)	Destination - Final city	KGS	Booked as	Container type	LaneID	SubLane	Index	SubLaneModality	SubLatFrom	SubLonFrom	SubLatTo	SubLonTo
165178682	Xian (CNSIA)	Landskrona	895,6	LCL	40HC	--CN-Xian (CNSIA)-CN-Karlshamn (SEKAN)-SE--Landskrona-SE	PORT_Karlshamn--Landskrona-SE	Last	ROAD	56,1619676	14,81994317	55,87035	12,83008
165178682	Xian (CNSIA)	Landskrona	895,6	LCL	40HC	--CN-Xian (CNSIA)-CN-Karlshamn (SEKAN)-SE--Landskrona-SE	PORT_Stralsund-PORT_Karlshamn	Third	OCEAN	54,48420563	13,58545012	56,16197	14,81994
165178682	Xian (CNSIA)	Landskrona	895,6	LCL	40HC	--CN-Xian (CNSIA)-CN-Karlshamn (SEKAN)-SE--Landskrona-SE	RAIL_Xian-RAIL_Mukran	Second	RAIL	34,31976799	108,6340456	54,48421	13,58545

Figure 6: Shipment

B CO_2 emission calculations

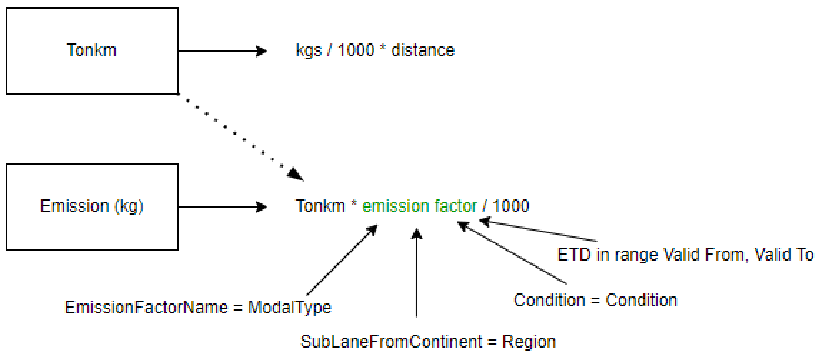


(a) Air

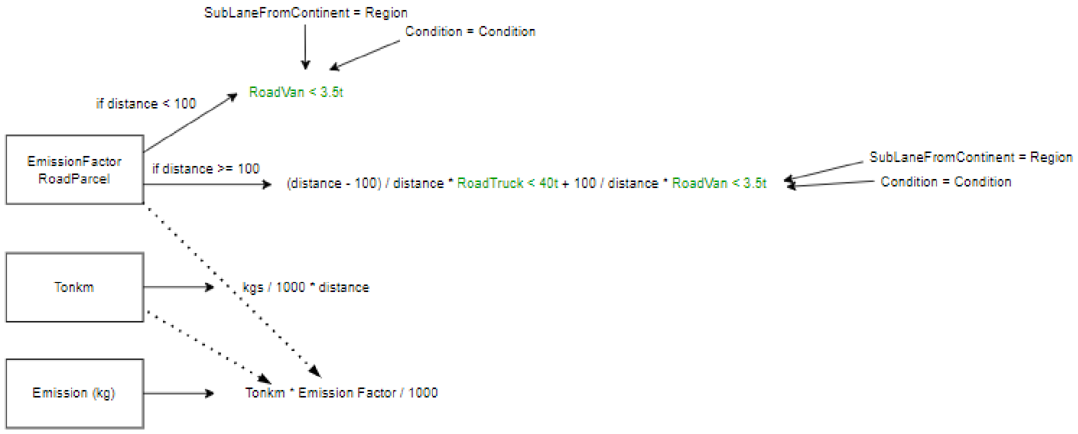


(b) Ocean

Figure 7: Graphic representations of the CO_2 emission calculations

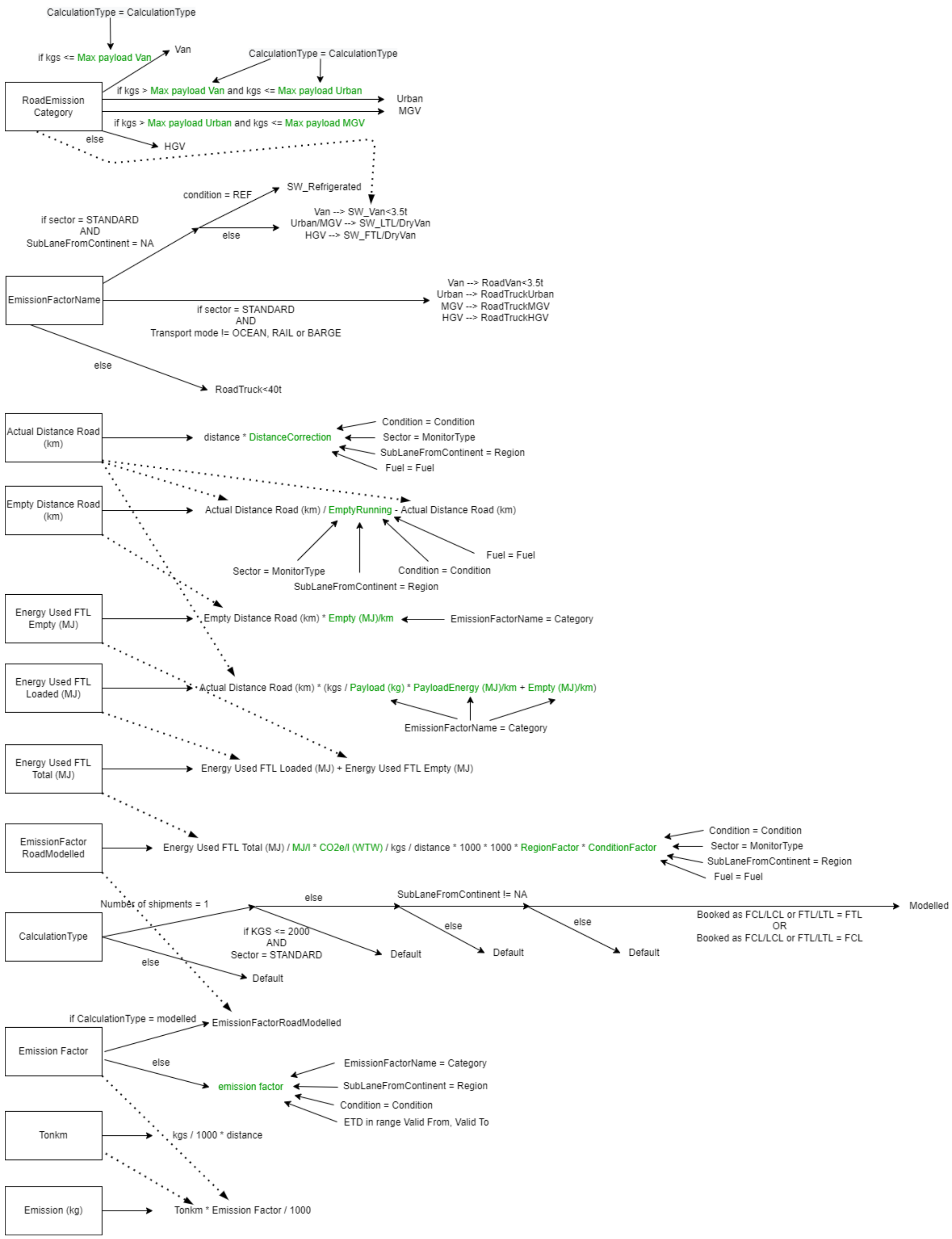


(c) Rail



(d) Parcel

Figure 7: Graphic representations of the CO_2 emission calculation



(e) Road

Figure 7: Graphic representations of the CO_2 emission calculation

C Numerical example MCF

Table 18 shows an example of a subset of the data. A subset consists of the origin points (SubLatFrom and SubLonFrom), the shipment weight (KGS), the maximum vehicle capacity for a specific shipment (Capacity), the leg index (Index), the ETD, the unique shipment ID, the transportation mode of the leg (SubLaneModality) and the ETA. The ETA is set to be zero for all leg indices that equal 'First'. This is because these legs are the starting points of the shipments. The shipments defined in this example subset all go to the same sub-destination. Figure 8 shows a simple graphical representation of the MCF network with five shipments and three clusters. Each shipment contains a latitude and longitude origin coordinate and a weight in kg. The shipments weights are 10, 5, 2, 3 and5 kg respectively, and the maximum cluster (weight) capacities are 10, 10 and 5 kg respectively. The initialization of the cluster centers starts by choosing the first centroid at random. This first centroid corresponds with one of the shipments. The second centroid is initialized by calculating every distance from the remaining shipments to the first centroid. The distances are then converted into probabilities by dividing each distance by the sum of all distances. The second centroid is then chosen randomly between the shipments that have the highest probability, and are thus farthest away from the first centroid. The third centroid is then initialized by again calculating every distance from the remaining shipments to the first and second centroid, transforming these distances to probabilities and choosing the one that is furthest away. The rest of the centroids will be chosen in the same way. The shipments will then be allocated to the dummy nodes, where each arc going from a shipment to a dummy node has a capacity equal to the shipment weight of that specific shipment. The allocation is done such that the distance is minimized and the capacity constraint is not violated. These capacity constraints are incorporated as capacities on the arcs going from the dummy nodes to the cluster nodes. The capacities of the arcs between the cluster nodes and the sink node are equal to the maximum vehicle capacity to make sure the demand is met. A possible shipment allocation is shown in Figure 8, where the orange colored arcs show which arcs are activated.

Table 18: Example of a subset

SubLatFrom	SubLonFrom	KGS	Capacity	Index	ETD	ID	SubLaneModality	ETA
1.2	2.4	10	10	First	2021-12-01	FLEX-123	ROAD	0
3.6	7.2	5	5	First	2021-12-01	FLEX-246	ROAD	0
8.7	5.9	2	10	First	2021-10-01	FLEX-369	ROAD	0
8.9	5.9	3	10	First	2021-10-01	FLEX-012	ROAD	0
8.7	5.7	5	10	First	2021-10-01	FLEX-480	ROAD	0

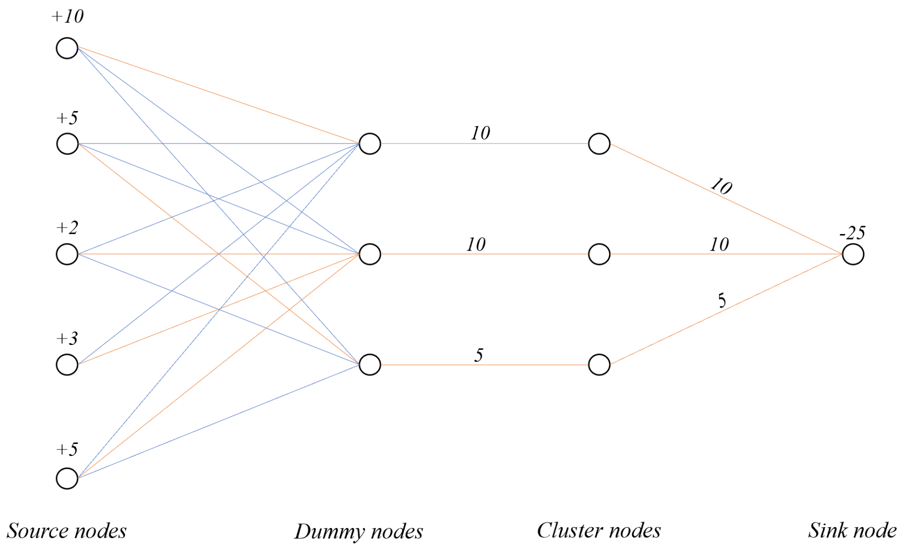


Figure 8: Numerical example