CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Data-Driven Decisions for eTruck Infrastructure: Optimal Placement of Charging Stations in a Route Network

A Case Study in Collaboration with DFDS A/S

Master Thesis

MSc Business Administration and Data Science

Contract n. 30215

Authors:

Jan Gaydoul, Student n. 149045

Emanuela Zucchetto, Student n. 149888

Supervisor:

Daniel Hain

Date: May 2023

Page count: 107

Character count: 222,720

Abstract

The goal of this thesis is to assess how to leverage Data Science methods to find optimal positions for electrical vehicle charging stations within a logistics provider's route network. More specifically, this research is carried out in collaboration with DFDS A/S, an European leader in providing both transportation and logistics services, who is aiming at becoming a carbon neutral company by 2050. In this regard, part of the company's decarbonization plan foresees to replace 25% of their truck fleet with electrical trucks (eTrucks). DFDS now is in need of a data-driven approach to answer the question where to deploy the eTrucks and where to install charging stations.

In this research, the Data Science methods we have focused on are Visual Analytics and Graph Theory. In a first step, Visual Analytics has been used used to better understand the complex transportation network at hand, both in volume and geographical terms. Here, the created dashboard has been contributed to identifying six focus countries for charging stations (and consequently eTrucks) deployment: United Kingdom, Belgium, the Netherlands, Germany, Denmark and Sweden.

Next, Graph Theory techniques have been deployed to further examine these focus areas. For each country, a graph representing the route network within the relevant distance range for eTrucks deployment has been constructed and visualized. The distance range from 5 to 300 km served as the baseline for this analysis, as this range constitutes the current common range of eTruck operations. By first deploying a community detection algorithm in order to identify important substructures within the route networks and then calculating the most important nodes based on the identified communities, we were able to identify suitable spots for EV charging stations within each network.

The results have then been compared to scenarios with increased (e.g. via technological advancements) and decreased (e.g. because of tough weather conditions) eTruck ranges in order to assess the scalability of the analyses. Lastly, based on the various analyses and calculations, tangible recommendations to DFDS were made as to which parts of the route networks to electrify and where to install charging stations, based on the identified electrification potentials for the focus countries.

Source code available at: https://github.com/em1899/master-thesis-project

Keywords: Visual Analytics, Graph Theory, Community Detection, Electric Vehicles, Charging Station Positioning, Coverage Analysis

Contents

List of Figures

List of Tables

1 About DFDS

This thesis is the result of a cooperation with DFDS A/S.

DFDS (Det Forenede Dampskibs-Selskab) A/S is a Danish shipping and logistics company headquartered in Copenhagen that has been operating for more than 150 years. The company has grown into a leading provider of transportation and logistics services in Europe, with over 8,000 employees and an annual revenue of approximately 27 billion DKK (3.6 billion USD) in 2022. (DFDS, 2023a) DFDS offers a wide range of logistics services by land and water, including shipping, logistics, and transport solutions, with the shipping segment accounting for 57% of the revenue and the logistics and haulage operations accounting for 43% of revenue. (DFDS, 2023a) Their shipping services cover both freight and passenger transport, including roll-on/roll-off (RoRo) and container shipping, while the logistics services on land involve warehousing, distribution, and other value-added services such as pick-and-pack, labeling, and assembly. Their transport solutions include road, rail, and intermodal transport, providing customers with end-to-end transportation solutions. Looking at road transport - as this will be the scope of this thesis -, DFDS offers full truckload (FTL) and less-than-truckload (LTL) services, which allow customers to transport goods efficiently and cost-effectively. Specialized solutions such as temperature-controlled transport for perishable goods, oversized and heavy transport for large and heavy cargo, and express transport for urgent shipment are also offered. In order to offer this broad product portfolio, DFDS has established a vast network of terminals, warehouses, and distribution centers throughout Europe, enabling them to provide efficient and reliable logistics services to customers across various industries. (DFDS, 2023d) In line with the industry's recent shifts towards more sustainable operations, DFDS has set an ambitious goal to become a carbon neutral company by 2050 and is heavily investing in new technologies that facilitate a more sustainable way of transportation, such as hybrid or electrical vessels and trucks or using biofuels and hydrogen fuels. (DFDS, 2023e)

2 Introduction

2.1 Introduction to the Business Problem

Without any doubt, climate change is one of the most pressing challenges of our time. Addressing this global issue has been widely recognized as urgent by governments, businesses, and individuals alike. If actions to reduce greenhouse gas emissions and mitigate the impacts of climate change are not (sufficiently) taken, the consequences will be severe and irreversible. In this context, it requires collective action and innovative solutions to transform the way energy is produced and consumed and natural resources are used.

Within Europe, the European Union (EU) is at the forefront of the fight against climate change and has set itself an ambitious goal to become climate neutral by 2050 (European Commission, 2021). This means that the EU intends to achieve net-zero greenhouse gas emissions by that date, where any remaining emissions will be offset by activities that remove CO2 from the atmosphere, such as reforestation or carbon capture and storage. To achieve this goal, the EU has set a series of interim targets, including a 55% reduction in greenhouse gas emissions by 2030, compared to 1990 levels. This target has been enshrined in legislation such as the European Climate Law and requires significant effort from all sectors of the economy, including companies. To comply with the EU's regulatory requirements and align with its long-term goals, companies need to implement measures to reduce their carbon footprint. This may include measures such as adopting renewable energy sources, improving energy efficiency or using more sustainable material. (European Commission, 2022)

In this light, DFDS set itself ambitious carbon emission reduction targets as well. The stated goal is to reduce emissions by 45% by 2030 and, in line with the EU target, to become completely CO2 neutral by 2050. Given the fact that roughly half of DFDS' revenue comes from logistics operations on land (DFDS, 2023a), a key role for this is being played by the development of more and more sustainable solutions for their truck operations. DFDS currently operates its land routes with diesel-powered trucks – which, as diesel trucks show a large carbon footprint and account for 23% of the overall carbon footprint of the traffic sector (European Commission, 2021), majorly impedes the company's efforts towards a carbon neutral future.

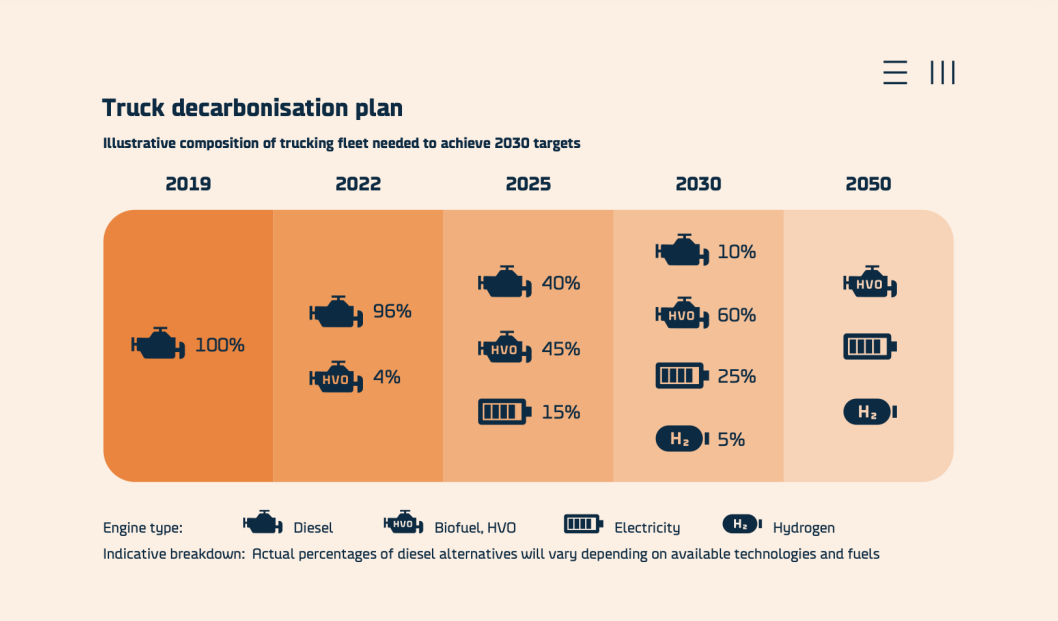

Figure 1: DFDS truck decarbonization plan: targets for tracking fleet engine type distribution to reach 2050 goals (DFDS, 2023b)

That's why DFDS made the decision to replace their fleet of diesel trucks with electrical trucks (eTrucks), trucks that run with HVO (hydro-treated vegetable oil), as well as Hydrogen trucks step by step, as those types of fuels have a significantly improved carbon footprint compared to the diesel trucks currently in use (Irles, 2023). The current truck decarbonization plan Figure 1 foresees to have 60% of the truck fleet run on HVO and biofuel by 2050, 25% on electricity, and 15% on hydrogen (these current estimates might change due to technological advancements over time). Due to the limited range of eTrucks (currently around 350 km when fully charged), DFDS – based on Hunter et al. (2021) – decided that eTrucks should take over the shorter ranges below 300 km, while trucks running on the other fuels will cover the long-distance operations, which will be a significant aspect for the analyses following later on.

However, this thesis will only focus on the deployment and distribution of eTrucks.

At the end of 2022, the first 125 electrical trucks were leased from Volvo DFDS (2023b) and are currently being delivered in batches (as can be seen in Figure 2), with the delivery of the last trucks being expected in Q4 2023. These trucks are supposed to be deployed within relevant parts of the DFDS operation and will serve customers who are willing to invest in

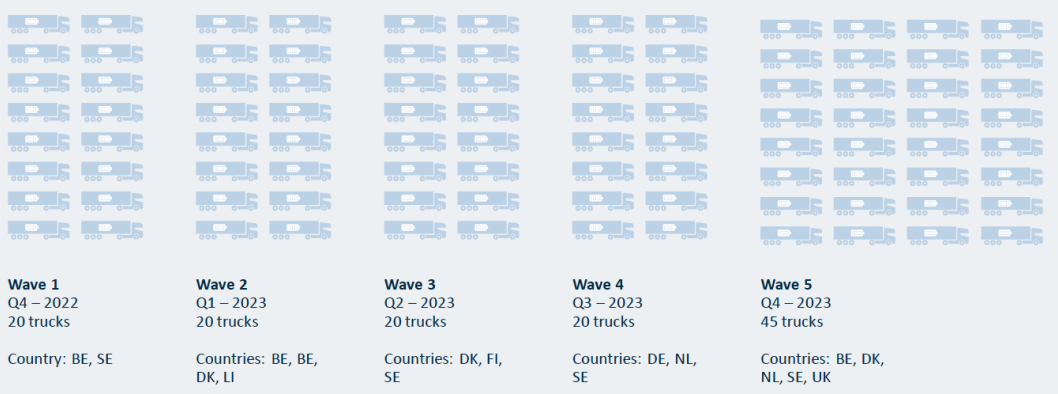

Figure 2: Distribution plan for VOLVO eTrucks leased by DFDS (directly provided by DFDS)

more sustainable transportation – as electrical trucks come with significantly higher costs than standard diesel trucks, which makes DFDS's service and products more expensive (Sharpe & Basma, 2022).

For a variety of reasons, the eTrucks project comes with high risks and complexities for DFDS. For example, the technologies behind eTrucks are relatively new, which might lead to some operational teething problems along the way as DFDS doesn't have any experience yet with the new product offering. The costs for eTrucks are significantly higher than for a standard diesel truck (Hunter et al., 2021), which makes the end product more expensive to customer. And, last but not least, charging has to be considered; while a diesel truck can easily rely on the available public charging infrastructure, no such infrastructure yet exists for eTrucks (more details on that in subsection 2.4), which is why DFDS also needs to set up a charging infrastructure for their fleet of eTrucks at strategically well positioned points with high traffic. These spots where DFDS invests into charging stations have to be picked very carefully, as the goal should be to electrify as large of a part of the network as possible with a number of charging stations that is as little as possible.

Keeping all this in mind, DFDS needs to address and answer two key questions for the eTrucks project to succeed:

1. On which routes within the DFDS network should the eTrucks be deployed?
2. Where should charging stations for the eTrucks be placed in order to cover large parts of the network?

Given the immense size of DFDS' route network across Europe, DFDS wants (and needs) to follow a data-driven approach when it comes to finding answers to these questions in order to make sure that the eTrucks are deployed in places where they can take over as many operations as possible. The problem: As of now, hardly any basis to build a data-driven approach on exists. While plenty of data on the logistical operations of DFDS is available, some key problems currently impede decision-making based on this data. For example, different datasets that have to be included in the analysis and that contain slightly different but crucial information are stored in different data warehouses with unclear data governance in place, making it difficult to drive insightful analyses. Furthermore, one result of this is that no comprehensive dashboarding solution showing information about the most important routes or the like exists as of now across the organization, making it difficult for the project managers of the eTruck project to decide on the routes where the first eTrucks should be deployed as critical information is not easily accessible.

As a result, the first eTrucks have been allocated to Belgium and Sweden (see Figure 2) not based on any data analyses but much more on "word of mouth" by talking to customers in these areas and checking for demand. For the further course of the project, the project managers now want to take a more data-driven approach - and that is where this master's thesis steps in.

2.2 Research Question and Scope

After outlining the challenges that the eTrucks project poses for DFDS and the need for a data-driven approach to optimize the deployment of eTrucks and charging stations, this Master's thesis aims to contribute to the project's success by addressing the following research questions:

- **RQ**: How can Data Science methods be used to find optimal positions for charging stations within the route network of a logistic provider?
- **Sub-RQ 1**: How can Visual Analytics be used to identify areas of high traffic and demand for charging stations within the route network of the logistic provider?
- **Sub-RQ 2**: How can Graph Theory techniques be used to optimize the placement of

charging stations within the route network of the logistic provider, taking into account traffic flows and routes distances?

By exploring these research questions, this thesis aims to provide DFDS with a data-driven approach for determining the optimal locations for charging stations in their route network.

2.3 Structure of the Thesis

As presented in the previous section, the objective of this thesis is to support logistics providers in finding optimal positions for charging stations through the use of Data Science tools. For this purpose, two different strategies are used: Visual Analytics (VA) and Graph Theory. The main rationale for this choice is that the combination of these two approaches can support in investigating the business problem from a high-level to a more granular one. For that purpose, the use of a VA tool such as Tableau provides an interactive framework for a general understanding of the logistical flows at hand, and the interactive interface helps project managers in navigating the visualizations without the need of programming skills.

While being able to identify high traffic and demand for charging stations would already contribute in providing value, the use of Graph Theory techniques can additionally help in optimizing the placement of charging stations. Here, Graph Theory would allow to obtain a more detailed perspective on the actual network in desired areas as well as clearly understand how the electrification of specific routes would help the logistic provider in reaching its targets.

For this reason, this thesis paper is organised in a way that retraces this path from general to specific: from a general understanding of a business problem, the focus shifts to data exploration in order to answer to the main research question as well as the two sub-questions related to Visual Analytics and Graph Theory in subsection 8.1.

More specifically, after a general introduction to DFDS, the logistic provider we have collaborated with for this thesis project (section 1), in subsection 2.1 an introduction to the business problem is given and the research questions are presented (subsection 2.2) before providing a background overview on eTrucks and charging stations (subsection 2.4). In section 3 the most relevant theoretical concepts are briefly presented, them being: coordinate reference systems (subsection 3.1), Visual Analytics (subsection 3.2), Graph Theory (subsec-

tion 3.3) and Community Detection Algorithms (subsection 3.4). In the following section 4 the available literature review is outlined, with a focus on AI application in the logistics industry (subsection 4.1), coverage analysis (subsection 4.2) and graph theory and community detection algorithm applications to the transportation network (subsection 4.3 and subsection 4.3.1).

Sections 5, 6, 7 heavily deal with the data provided by DFDS, starting with the Methodology section. Here, after an introduction to CRISP-DM as research methodology (subsection 5.1) and the general framework adopted (subsection 5.2), data sources (subsection 5.3), description (subsection 5.4) and pre-processing and EDA (subsection 5.5) are presented. Subsequently, in the Visual Analytics approach section (section 6), a series of visualizations and a dashboard are provided to identify areas of high traffic and demand for charging stations. In addition, subsection 7.1 introduces the graph building process and the associated assumption, followed by a country analysis for each of the selected countries (subsection 7.2).

After that, the results deriving from the two used approaches (subsection 8.1) are summarized with the intent of answering the research questions. Moreover, suggestions on how to implement the solution in DFDS' current infrastructure (subsection 8.2) are provided. Finally, the main limitations are outlined in section 9 before stepping into conclusions and proposal for future work (section 10).

2.4 Background: Rise of eTrucks, Lack of Charging Stations

When thinking about the research questions introduced above, one could be tempted to underestimate the problem and think "*Why does DFDS not just rely on publicly available charging stations for electrical vehicles?*". The simple answer is: Charging stations for eTrucks significantly differ from charging stations for electrical light commercial vehicles (e-LCVs, such as Teslas), and for eTrucks such network just does not exist yet (Bernard et al., 2022). That is why, before moving on to the methodological part of the thesis, we briefly want to address the rapidly growing demand for eTrucks as well as the corresponding charging infrastructure, which simultaneously is not keeping pace.

Despite the first development of electric vehicles can be traced back to the 1800s, their relevance has largely increased starting from the first decade of the 2000s (U.S. Department of

energy, 2023b). Based on the most recently available data (2021), now sales of EV worldwide amount to more than 6.5 million vehicles (Statista, 2022). In order to better understand the industry of electric vehicles it is important to identify the main groups in which they can be divided. One of the most relevant differences lies in the battery set-up. Based on this EV can be divided into three main categories, as summarized by the U.S. Department of energy (2023a):

- BEV (Battery Electric Vehicle): this type of vehicle can be considered as "fully" electric, in the sense that they are solely powered by electricity and rely on external sources to be recharged. The focus of this research will be on BEV, as they are the ones adopted by DFDS

- PHEV (Plug-in Hybrid Electric Vehicles): these vehicles have a combustion engine in addition to the electric battery. In this way they can cover larger distances compared to BEV, but have a more limited battery capacity.

- HEV (Hybrid Electric Vehicles): similarly to PHEV, they are powered from both batteries and combustion engine, but in this case the battery pack cannot be charged from an external source.

In addition to these main categories, others exist as well, however they have more limited applications.

Another basis for differentiation among electric vehicles can be their size and/or purpose. Indeed, the majority of the EV market refers to smaller vehicles for private or shared used, such as e-cars, e-bikes and e-scooters. However, larger electric vehicles have seen an increase in relevance as well, especially when it comes to public transport and light electric vehicles (LEVs). In particular, based on a BNEF study reported by Statista (2022) 67% of buses in use worldwide are expected to be electric by 2040. This is mainly due to the governmental commitment to lower pollution levels and to the lower long-term operational expenses – still taking into account higher upfront costs for acquisition. Nevertheless, almost the full share (95%) of e-buses expected to be in use as of 2030 will be in Mainland China, which is also the country with the highest number of electric busses currently in use, as a consequence of the combination of different factors, including the large urban population and the subsidies

system put in place by the government. For what concerns Europe, Germany has the largest electric bus fleet, closely followed by the UK and the Netherlands. (Statista, 2022)

Another type of EV that is predicted to grow is that of e-LCVs (electric light commercial vehicles). This is possible as they have a "high and predictable" (Statista, 2022, p. 113) use rate, given that they are expected to be used on fixed and predictable routes. Given the constrained geographical area of use they do not require multiple charges during the day, making their use particularly convenient. On the other hand, the predictability of heavy-duty electric trucks can be much more complex to analyse and indeed, their roll-out is still marginal. In 2020, sales of eTrucks are mostly concentrated in China and Europe, which accounted for 17% and 13% of truck sales, respectively. However, sales in China in 2021 accounted for 44% of the total, more than double compared to Europe. In Europe, in 2021 the heavy truck volume amounted to 346 units, with Switzerland having the highest with 77 units Volvo, 2022. Even expecting an overall increase in usage, with 120 deployed eTrucks, DFDS will likely have a considerable impact on eTrucks volume in the current year, being one of the first movers in adopting this solution. Nevertheless, this explains why it may be challenging for the company to rely on already available charging infrastructure, as publicly available charging options are currently low in number (Bernard et al., 2022).

A series of elements can explain the marginal relevance of this segment of the industry: firstly, long-haul vehicles need higher driving autonomy, requiring in this sense a specific battery and charging infrastructure. Secondly, commercial vehicles and heavy-duty trucks account for not even 5% of vehicles in use in Europe and therefore, automobile companies can be less incentivized to invest in this area of production (Statista, 2022). However, with the expected rise in deployement of eTrucks, this is subject to change in the coming years (McKinsey, 2023).

2.4.1 Type of EV charging stations

In order to fully understand the electric vehicle industry, it is also fundamental consider the status of EV charging opportunities. Indeed, EV charging stations have complex technical characteristics that considerably affect strategies bringing higher levels of electrification in transport systems. In order to help create standardization, several organizations and societies - such as the International Organization for Standardization (ISO), the Society of Automotive

Engineers (SAE) and the International Electromechanical Commission (IEC) - have defined standards to be followed for various aspects of EV charging. (Mastoi et al., 2022; Pareek et al., 2020)

Firstly, based on Pareek et al. (2020) it is possible to distinguish between EV charging stations based on their type: "residential charging station" are located at the EV owner residence and tend to be used for overnight charging, as they have limited charging capability. At the same time, because of this, they make it possible to diminish grid load, which is one of the main issues related to the increase of demand for charging stations (McKinsey, 2022). In addition, EV owners can capitalize on the time their vehicle would anyway be parked by using "parking charging station". Finally, "public charging station" are an efficient solution for vehicles requiring higher charging times.

Another relevant aspect to consider when differentiating charging stations is the energy level of the charging, which in a way mirrors the already-presented types. Level 1 stations only require a standard 120-V outlet and for this reason are most typically used as residential and parking charging stations. Given their maximum power of 2.4 kW, they can take up to 16 hours to charge a vehicle. Level 2 stations can be used both in residential and public locations, when the power requirements are met. Similarly to L1 stations, they rely on AC supply, but the higher "current flow capacity" (Mastoi et al., 2022, p. 11509) allows for faster charging. Finally, Level 3 chargers guarantee even higher charging speed, as 80% can be charged in 20 minutes, thanks to the DC technology in use. (Mastoi et al., 2022)

In addition to the more commonly available options, production of these ultra-fast charging stations is increasing. This solution could still guarantee charging time under an hour also for vehicles with higher capacity batteries, thanks to their higher power output. However, in order to ensure economy of scale, a global or at least regional standard is required. (Bernard et al., 2022) This becomes even more relevant in the case of the heavy electric vehicle industry, which is still considerably less developed than the light vehicle one. In addition, it is important to consider that buses and trucks are indeed designed to carry high loads and to travel longer distances- especially in the case of the latter- and therefore require more powerful batteries. Consequently, faster charging solutions and higher electrical capacity are required to be developed. For what concerns Europe and North America, in 2018 CharIN, a global association dedicated to promoting interoperability of charging infrastructure, has set up a

task force aiming at introducing standardization in charging stations with maximum power of 3.75 MW. Sales are expected to start in 2024, as in 2023 a series of pilot projects will be carried out, including the first European test event for the company in April 2023. The company, with already more that 50 thousand CSS charging points in the continent, is indeed planning to develop a new high power charging solution to meet the raising requests in the electric heavy-duty industry. Within this, they aim at introducing "requirements for the EVSE, the vehicle, communication, and related hardware" and recommendations for what concerns charging connectors. (CharIN, 2023). Also, in 2022, Milence - a joint venture of Daimler, Volvo, and Traton - was established, with the goal of "building Europe's leading network of public charging solutions for heavy-duty vehicles" (Milence, 2023). However, the baseline is that this is a fast developing market with more and more players pushing into it, but for the ongoing eTrucks project, DFDS yet has to install their own charging infrastructure as public solutions are not ready yet.

2.4.2 Other alternatives for EV charging

Despite being the most popular option, wired charging stations are not the only solution for EV charging. Bernard et al. (2022) have identified three main alternatives and the reason that is limiting their usage in the industry. The first is battery swapping, which would allow considerably reduced charging times as well as lower upfront costs for the vehicle, as batteries represent between 30 and 40% of the overall cost (IER, 2022). Nevertheless, the lack of standardization in battery production and the high cost of battery acquisition for station owners have hindered their diffusion. For these reasons, only a couple of battery swapping project have been developed outside China, which accounts for the highest density of electrical vehicles in use. However, these attempts by Better Place and Tesla did not achieve the expected success. Overhead catenary charging would use pantographs to make electricity flow from the electrical line to the vehicle. Through this technique it would be possible to leverage on the economies of scale for public transport. (Bernard et al., 2022) Despite this, cost would still remain extremely high, as much as 1 million euros per km with production at scale based on a Siemens study of 2022. Finally, another alternative would be in-road wireless charging, where the charging would happen through "magnetic coils embedded in the road to receiving coils fitted to electric vehicles" (Bernard et al., 2022, p. 7). This solution would

again be extraordinarily costly, as much as USD 1.2 million per km (Houser, 2018), with the lack of operational and technical standards making the development even more complex. Additionally, in the same way as the previous case developing such projects would require considerable public investments.

To summarize, these alternative solutions still require large investments for research as well as deployment, which are not feasible for governments on their own let alone single private corporations willing to invest. The development of these solutions would in this sense require a wide coordinated effort, which poses these options outside the scope of interest of DFDS. For this reason, the deployment of charging stations represents the most viable option for the logistics provider.

Having now established the business problem, its background and its significance in the previous sections, we now turn our attention to developing a conceptual framework to address this challenge. In the following sections, we will outline thee most important concepts we will be working with throughout the thesis, and introduce relevant theoretical and empirical literature to justify our choices, before then moving on to the analytical, data-heavy part of the thesis.

3 Relevant Theoretical Concepts

This chapter aims at providing an high level overview of the theoretical frameworks adopted in the process of answering to the research questions. Considering that this task relies on the use of a multitude of disciplines, for brevity purposes, only the most relevant concepts will be presented. More specifically, a coordinate reference system is used to understand the geographical dimension of logistics data, while Visual Analytics serves as aid to facilitate the understanding of such data in a more straightforward manner. Finally, Graph Theory is used for a more in-depth analysis of the network structure of the company's distribution flow.

3.1 Coordinate reference System

Given the high relevance that geospatial analysis has in this context, a coordinate reference system is used to identify starting and ending point for each registered booking and consequently for each route covered (subsection 5.4). In particular, the adopted system is an

angular one, where each geographical location is represented by a couple of points indicating latitude and longitude, respectively. While the latitude represents position of a point north or south in respect to the Equator, longitude measures the position eastern or western to the Greenwich meridian (Encyclopaedia Britannica, 2023). Additionally, given the constraints posed by the driving capacity of electric trucks, the distance covered in a journey results to be a critical aspect. For this reason, the distance between start and end location has been calculated in two different ways, with increasing precision. The first approach consisted in measuring the Haversine distance, also known as "great circle distance". This represents the "angular distance between two points on a sphere" (scikit-learn, 2023), where the pints are represented by their geographical coordinates. Haversine distance would represent a valuable and considerably accurate measure in the case of air transport, however it poses some limitation in the context of land freight as roads are unlikely to be straight from point to point. For this reason, performing geographical routing through the aid of the Bing Maps Routes API allows to determine more accurately the distance between points assuming that a vehicle would drive between starting and ending location (Microsoft, 2022).

3.2 Visualization - Tools and Approaches

Visual Analytics is a field combining a plethora of disciplines aiming at generating interactive visual interfaces and deriving relevant insights useful in the context of analytical reasoning and in decision making processes (Kulkarni et al., 2016). To put it into Few's words, visualization techniques and products "are rapidly becoming recognized for the rich analytical insights, they make available to our eyes" (Few, 2007, p. 5). This is indeed true especially in the current context, as BI tools are used to facilitate the understanding of an otherwise rather complicated dataset. While raw data are initially collected for the specific purpose of monitoring and organising deliveries and therefore require to be understood mainly by expert in such departments, the nature of data limits the understanding that can be derive for other scopes. Therefore, the goal of the use of Visual Analytics in this case is to encompass the complex nature of data by presenting them in a way that bridges between the more technical aspect and the business insights that can be derived from them. Through the use of a tool like Tableau, it is not only possible to generate independent visualizations, but also to combine them in dashboards, where it is possible to derive insights by centralizing and monitoring one

or more datasets in an interactive manner (Calzon, 2022).

In this thesis, it serves the purpose of monitoring the overall distribution flow of the company from a geographical, temporal and volume perspective, as presented in section 6. In particular, a floating Tableau dashboard is used (subsection 6.4), as it allows more flexibility in positioning and sizing visualizations and filters, resulting in a more pleasant product. As mentioned, one of the key aspects is interactivity and for this a series of filters and pages have been used in the construction of visualizations and dashboards. While filters more simply limit the range of data to be visualized on the criteria based on which the filter is set, pages allow for a more elaborate breakdown. In fact, by inserting a feature in the corresponding shelf a series of pages is generated, where each differentiates from the others based on a value of the field used to generate them. This would be typically used with time measures as it makes it possible to understand how data change through time (Tableau, 2023b).

3.3 Graph Theory

Graph Theory is especially used to translate complex networks into mathematical structures. For example, it is typically adopted in social network analysis, communication representation, but it can be applicable in a wide variety of other situations as well. In the context of this thesis, graphs are used to represent DFDS route network of specific countries (section 7), trying to leverage traffic flows and route distances to optimize charging station placement.

In order for a graph to be identified, the presence of a series of elements is required: a non-empty set of vertices (V) or nodes, a set of edges (E) disjoint from that of vertices and a function associating each element in E with unordered pair of vertices in V (Bondy & Murty, 1976). Therefore, a graph can be mathematically represented as:

$$G = (V, E) \tag{1}$$

There can be many criteria based on which it is possible to classify graphs, however, the most important one is between directed and un-directed graphs (Figure 3). Indeed, since they represent different structural situations, based on the category it is possible (or not) to apply certain approaches and algorithms. The characteristic that distinguishes between in-directed and directed graphs is the nature of the pair of vertices to which edges are associated. In the

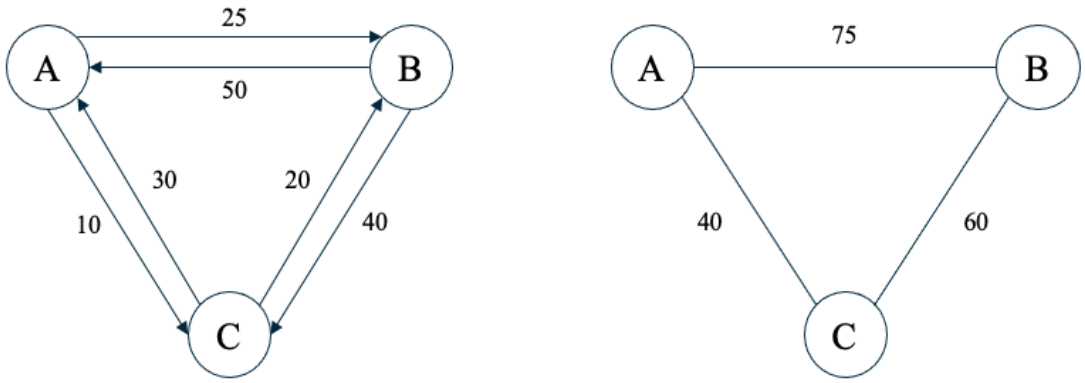

Figure 3: Example of a directed graph (left) vs. an undirected graph (right)

first case, the the order of the pair is irrelevant, while in the latter each arc in the digraph has to be associated with an ordered pair of vertices. (Bondy & Murty, 1976)

For this reason, directed graphs are generally used in the case of traffic flow problem, as there may be one-way roads or because we want to represent a real flow of vehicles which are therefore moving in specific directions.

Very frequently, a number (or weight) is associated with edges in graphs. The weight of a connection gives an indication of the relevance of that edge compared to the others. This could represent the number of times a route has been covered, or the distance between the starting and ending node of a journey. More precisely, (Bondy & Murty, 1976, p. 16) define weighted graphs as "subgraph of a weighted graph," where "the weight *w(H)* of H is the sum of the weights $\sum w(e)$ on its edges".

3.3.1 Centrality Measures

Once direction and weights are set it is possible to use the graph structure to get a deeper understanding of the network. A typical approach consists in using algorithms to find which path has the maximum or minimum weight, like with the problem of the shortest path, which aims at identifying the shortest path from one node to the other with the lowest cost. Similarly, it is relevant to obtain an understanding of the centrality of nodes in the Graph. As the name suggests, centrality measures help in measuring how central a node is, in the sense of relevance. More precisely, three main centrality measures can be identified (Hansen

et al., 2020):

- Degree centrality: indicates the number of direct connections of a node. In the case of a digraph, both In- and Out- degree are specified, with the first relating to the number of incoming connections and the second to those directed to other nodes.

- Closeness centrality: it is the summation of the geodesic distance between a given node and each of the other nodes in the network, representing in this way the "closeness" of a node to the other. Here, the geodesic distance is the amount of edges between two nodes considering the shortest path, with it being $d(a, b) = \infty$ if no path between the two exists.

- Betweenness centrality: in this case the relevance of a node is based on the number of its appearances in the shortest path between all node pairs in the graph.

Despite these measures lay the foundation for understanding node centrality, their main limitation is that they cannot be immediately applied to weighted graphs, the category the route networks graphs created here belong to (Tore Opsahl, 2011).

Initially, weight strength was used as substitute measure for Degree centrality, given that it takes into account the weight level. However, it lacks in counting the number of connection existing. Therefore, Opsahl et al. (2010) have designed a solution combining the two through the introduction of a parameter quantifying the relevance of the number of connections relative to the weight value. In this way degree centrality is measured as the product of the number of nodes connection of a given node by the mean weight adjusted by the tuning parameter.

The transposition of closeness to weighted graphs has been firstly carried out by Newman (2001), through the application of Dijkstra's shortest path algorithm. The algorithm was used to determine the least-costly path, where the cost is calculated as the inverse of the weight.

Finally, an application of betweenness centrality to weighted graph was the result of a generalization of the centrality developed by Brandes (2001). The rationale behind this relates to the fact that when more intermediate notes with strong connections exists, the connection is faster compared to having less nodes but with weak connections.

Again, these generalizations take into account the weight of each connection, but lacks in counting the number of connection existing. In order to solve this, Opsahl et al. (2010) generalisation of shortest paths can be used.

3.4 Community Detection Algorithms

Despite being defined as graph in the mathematical literature, these structures are very commonly referred to as networks, in particular in social science field. This is due to the fact that they are use to represents relationship and interactions among people or thing and especially when it comes to social network or web interactions.(Newman, 2003) From this perspective it is also simple to understand the role of the previously presented centrality, as a measure of relevance of nodes in a network. As an example, such measures could be used to calculate the degree of connection of a social media user compared to others.

Nevertheless, the increase in computational power has shifted the focus from smaller networks to others with up to billions of nodes, where the relevance of a single vertex is only minimal. For this reason, instead of identifying single nodes it can be more meaningful to understand the impact that a "group" of nodes has on the network. These groups are called *Communities* in the literature, as collection of nodes that have a "high density of edges within them, with a lower density of edges between groups"(Newman, 2003). Another reason that makes communities relevant - still connected to the magnitude of networks analysed nowadays - is related to the easiness of visualization, as with millions or billions of nodes it would be complex to clearly visualize each one individually, so the identifications of communities can help with that.

Communities have straightforward application in the case of social network as a person/node can have connection with other schoolmates and with coworkers. However, it is likely that the level of connection is higher within people attending the same school and people working at the same place, but the intra-connections between the two groups are lower in number. Similarly, the same can be applied to logistics networks, where journeys depart with a high frequency from a distribution point to serve the surrounding area, but there are still vehicles driving from the production factory, warehouse or another distribution point to the current one.

A series of community detection algorithm is available, however based on the specific application some can be more appropriate than others. Nevertheless, before presenting the main algorithms in use it is important to differentiate among two categories based on their approach (Gujral et al., 2019):

- agglomerative methods: this methods works by adding edges iteratively to a graph formed initially only of nodes. The addition is done progressively adding edges with decreasing strength, where the definition of strength depends on the algorithm used.
- divisive methods: these algorithm use the opposite approach of agglomerate methods, meaning that they iteratively remove edges from the network, starting with stronger ones.

Here, two different algorithms will be presented: the *Girvan-Newman* and the *Louvain* algorithm, the most popular one and the one used in this research, respectively. Nevertheless, many other algorithms are available as well. Both of these are based on optimizing the network modularity benefit function, in contrast to clustering techniques aiming at splitting the graph in sub-graphs. The main limitation of the latter is that the number of communities and their size has to be known beforehand, which is an unlikely event in a real-life application. On the other hand, network modularity (De Meo et al., 2011, p. 89) can be defined as

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{|E|} - \left(\frac{d_s}{2|E|} \right)^2 \right] \tag{2}$$

where l_s is the count of edges part of the s^{th} community and d_s the sum of the node degrees in the same. This function can be maximized, which is the goal of the presented strategies.(De Meo et al., 2011) The *Girvan-Newman algorithm* first ranks edges strength based on the betweenness centrality and then proceeds in deleting edges while increasing M, therefore using a divisive approach. This is based on the rationale that nodes with high centrality "connect nodes belonging to different communities" (De Meo et al., 2011, p. 89). Despite being a widely used algorithm, its main drawback is the computational complexity of calculating betweenness centrality, making it unfitting when the size of the graph increases considerably.

This issue is solved by the *Louvain algorithm*, since it is based on local information. In this case, vertex are added to communities in a way that maximizes the modularity. Consequently, a new network is built, where nodes are the communities identified in the first step. (De Meo et al., 2011). This is indeed the algorithm that will be used to identify communities within

the route network graphs of the countries in analysis in section 7.

3.5 Evaluation Metrics for Graphs

In order to better evaluate the Community Detection algorithms used and to understand the results on the application to DFDS network in subsection 7.2, a series of metrics are used and enumerated below:

- Average Degree: average number of edges a node has, calculated as the average between the sum of node degree and total nodes. This helps in giving a general understanding on how the graph may be structured (Woodall, 2008)
- Clustering Coefficient: this metric gives an indication of the level to which nodes in a graph tend to form clusters and it is represented by the ratio between the number of connections between the nodes in the proximity area of a node n over the total number of connection there can be. The coefficient gets closer to 1 (0 being the minimum) the more nodes connected to n are also connected with other nodes in the proximity (Saramaki et al., 2007)
- Modularity: this measure indicates how strongly communities are separated the one from the other. A high modularity (which can assume values between 0,1) indicates a higher number of edges in a community than you would expect by random guess (Newman & Girvan, 2004)
- Conductance: this measure indicated the ratio between connections pointing outside the community and the total number of connections it has. On a range between 0 and 1, the smaller the conductance of a community the more the links within the community compared to the links with nodes located outside the community (Bollobás, 1998)

In the previous section, we introduced the key concepts that underpin our analysis, including the use of the coordinate reference systems, visual analytics, and graph theory. In the following section, we will examine relevant literature to explore how these methods have been applied to similar problems in the past, and draw on these insights to carry out our own analysis.

4 Literature Review

This section will provide an overview on the available literature concerning the main aspects of the topics covered in this research. Firstly, a general introduction to the main AI applications in logistics is given, before narrowing down the focus to more constrained research areas. Despite being a subject of recent applications, various approaches have been used to determine the optimal location for charging stations for electric vehicles and more generally in the context of coverage analysis applied to alternative fuels. Additionally, network theory has been widely used in transport applications as well, especially in solving the location-allocation problem. Finally, the sub field of community detection algorithm has proven to be successful in clustering and segmenting transport networks, as a way to improve current infrastructures. A summary of all mentioned papers, including objectives and model(s) used is available in Appendix I.

4.1 Artificial Intelligence in Logistics and Supply Chain management

In recent years, the technological developments alongside uncertainty related to climate change and more recently pandemics has considerably impacted the logistic and supply chain industry, boosted by new advancements that can be represented by the umbrella term of fourth industrial revolution (Industry 4.0) (Woschank et al., 2020). While providing a satisfactory overview would be excessively extensive, the reviews elaborated by Singh et al. (2021), Woschank et al. (2020) and Giuffrida et al. (2022) helped us in generating a brief summary of the most relevant applications in the industry.

In their review, Singh et al. (2021) have identified the trends regarding ML applications in logistics. In particular, they have focused on four main challenges that affect logistics: intermodal transportation, uncertainty of demand, user behaviour and reverse logistics. The first issue is concerned with ensuring a smooth transition between mode of transport, when multiple of them are used to transport goods or people. Here, supervised learning models like Neural Networks and SVM are used (Abdirassilov & Sładkowski, 2018), as well as clustering (Göçmen & Erol, 2019) as example of unsupervised model. For what concerns demand uncertainty, there does not seem to be a most frequently used approach as multiple different ones have been tested, such as genetic algorithm (Zarbakhshnia et al., 2020) or combination

of long short term memory and convolutional neural network (Ren et al., 2020). For what concerns user behaviour, comparative studies on supervised models have been proposed (Xu et al., 2021). Finally, reverse logistics aims are "recapturing value" (Singh et al., 2021, p. 69) by moving goods the in the opposite direction: from point of use to origin point. This is a rather new area, with application in CNN (supervised Learning) (Schlüter et al., 2021) and Markov Decision Problems (reinforcement learning) (Tuncel et al., 2014). To summarize, it results that supervised approaches still have higher relevance in this literature, with a variety of different approaches especially in the first three groups. While unsupervised learning is used in some k-Means and clustering applications, reinforcement learning is very rarely used.

A similar approach, aiming at identifying AI applications is used by Woschank et al. (2020). In this case the focus is specifically on research associated with Smart Logistics applications, where seven main clusters are identified by the authors. Among these, "Cyber-Physical Systems in Logistics" and "Predictive Maintenance" are the most popular. CPS is used to better understand how to leverage in the best possible way the large amount of data owned by companies, by analysing them, identifying patterns with the goal of activating workflows. On the other hand, the field of predictive maintenance has flourished following the increase in use of sensors collecting data that can later be used for data analytics. Another relevant focus areas are Intelligent Transport Logistics- mainly driven by object recognition tasks- and "production planning and control systems", a more established field which is now being revamped thanks to AI leading to planning strategies in real time. Again "Strategic and Tactical Process Optimization" are heavily used for forecasting tasks both from a user and customer perspective (e.g. To predict customer demand), while improvement of Operational Processes in Logistics, aims at introducing in logistics processes that can be derived from nature. Finally, Hybrid Decision Support Systems are focused on human-centered engineering, where AI can provide support in decision making processes by leveraging big data. (Woschank et al., 2020)

The third review (Giuffrida et al., 2022) presented provides a narrowed-down analysis of the specific sub-field of last-mile logistics, from both the perspective of ML and vehicle routing applications. These two approaches, and the combination of both, are indeed the most popular ones in the optimization of this field, which is becoming important for the public sector and especially delivery companies. Among the most relevant applications in the first

category, Bricher and Müller (2020) have used DNN for logistic cargo automation and SVM and combinations of K-nearest neighbours and random forest are used for demand forecasting (Albadrani et al., 2021). Another popular topic is that of anomaly detection, with ML models to GPS tracks anomalies (Feng & Timmermans, 2015). On the other hand, Vehicle Route Optimization (VRO) models – as the name suggests- aim at finding the optimal route taking into account a series of factor, based on the sub-group of VROs considered. For example, rich vehicle routing problems use real-life constraint (Lahyani et al., 2015), while other versions introduce uncertainty of demand, introducing the possibility of not being able to satisfy it (Sumalee et al., 2011). In heterogeneous VRP models, fleets composed of different types of vehicles are considered, as it would be a typical condition of last-mile logistic. Finally, this logistics sector could benefit or dynamic VRP as well, as they introduce "dynamic routing" which could be beneficial in case of "accidents and re-scheduling" (Giuffrida et al., 2022, p. 7).

4.2 Coverage Analysis

When it comes to vehicles powered by alternative fuels, the issue of ensuring coverage from a refuelling station perspective becomes more complex as the vehicle autonomy is considerably reduced compared to that of oil-powered cars and trucks. Therefore, the assumption that a single refuelling station would be sufficient to cover the demand on that whole path cannot be used. Also, in many studies the assumption that by increasing the number of alternative fuel vehicles the number of traditional vehicle will decrease it reasonably considered valid. For this reason, analysis for deployment of new charging stations starts from the observation of current gas station locations. Despite this could solve the issue of finding a location for such purpose, it does not eliminate the problem of the large investments required for the deployment of a EV charging station area. In addition, another issue specifically affecting electric vehicles is that the electric power can be stored in limited amounts and "and must be kept in real-time balance between power generation and consumption" (Gong et al., 2016, p. 65). In the same study, Gong et al. (2016) also underline the problem of heterogeneity from different perspective: firstly, the use of EV varies significantly in different geographical regions and at the same time the flow and the traffic condition have a random component. Moreover, the technical configurations of power grids can vary considerably and are often in a sub-optimal state of efficiency. All of these considerations, in combination with the increase

in relevance and popularity of electric vehicles, makes coverage analysis a prominent issue and a popular topic in research. For this reason, an overview of the status of research on the topic is provided, focusing of the aspects that the research is trying to improve.

From the investment perspective, Davidov and Pantoš (2017) aim at minimizing the infrastructure costs while ensuring to the user two main factors: charging reliability and service quality. The first aspect is guaranteed by the placement of a station in the users driving range, while the second by the creation of a quality-of-service index, "reflecting the disposable charging time of the EV driver to complete planned trips" (Davidov & Pantoš, 2017, p. 1165). Based on their application, a higher index determines a lower placement infrastructure cost, as longer charging times are associated with lower costs for the charging station. Similarly, there are other researches interested in minimizing development and installation costs, although with a considerable focus on the energy and grid aspect. As an example, Yan et al. (2014) aim at minimizing investment costs taking into account energy losses and constraint conditions. Here, an intense focus is set on power supply structure, associated switches and the potential points for charging station deployment. In order to analyse the combinations of feeders which would minimize the energy loss, an intelligent optimization algorithm is used: the hierarchic genetic algorithm.

Another optimization strategy for the planning of a public charging infrastructure for battery electric vehicles (BEV) with the goal of minimizing the infrastructure costs reducing at the same time the impact on the power network, is developed by Gong et al. (2016). Alongside with the increase in popularity of electric vehicles, the technology aspect has seen improvements as well, including those related to the charging power. For this reason, the focus of the research at hand is specifically on fast charging solutions. The suggested solution is "an abstract-map-based multilayer optimization strategy" (Gong et al., 2016, p. 64), where in each layer as set of condition is tested and the optimal results are used as inputs for the following one. In particular, keeping the goal of minimizing infrastructure cost constant, the objective of the first layer is to reduce negative effects on the transportation system, while the second focuses on the power system. Finally, the latter tries to combine the two perspectives, leading to a result that overperforms the other two taken separately.

On the contrary, Xi et al. (2013) try to find the optimal number of first and second level

charging stations to deploy presenting different scenarios simulations based on the available budget. While most studies do not take into account the time necessary for charging - and this can be considered a valid assumption in the case of fast-charging as it generally takes only up to 30 minutes - this cannot be applied when slower charging solutions are in place. Based on the resulting observations, first level stations are recommended in the case of workplaces parking lots as the incremental speed of charging of second level stations would provide little benefit considering the longer stop in the area. On the other hand, when considering shopping centres the latter type is recommended, as parking times are shorter.

In later years, Zeb et al. (2020) have focused in optimizing the combination of current level of charging stations. The rationale behind this research objective is related to the existing complexity in compatibility of different level of EV charging stations, due to high energy consumption of the highest level of charging stations – with up to 100 Kw charging power. Again, the objective is minimizing energy losses and distribution transformer loads as well as installation costs. The resulting constraint non-linear stochastic objective function is solved through a Particle Swarm Optimization model and the applications to a real case scenario show a improvement for all three considered aspects.

While minimization of costs and impact on the grid are definitely relevant topic in coverage analysis framework, the role of demand should not be overlooked. For this reason, Xiang et al. (2016) have integrated in the research not only aspects concerning station capacity, but also traffic flow data, still with the purpose of minimizing the investment. Indeed, in the process of planning a self-standing infrastructure, it is also necessary to make estimates of the charging load that would have to be satisfied given the demand. In order to integrate this aspect, the traffic flow distribution at each time stamp is calculated and different typology options are designed, before narrowing down to the solutions that respect the constraints.

Micari et al. (2017) have created a two-level model, firstly focused in finding the optimal location for charging stations and later in finding the appropriate number of to be placed in each area. In this case, the underlying algorithm is based on 3 different criteria: vehicle technology, charging station technology and flow of vehicles. More specifically, the first two refer to the way the battery is charged from the vehicle and station perspective, while the flow indicates the volume of cars driving. For each of the two steps, they have defined two

set of functions. In order to identify the number of charging stations, the function depends on the "range anxiety" of drivers and the safety margin, which is related to the battery capacity. On the other hand, the number of charging stations is decided based on the number of charging sockets of the station, number of vehicles that can be recharged in a day, on a flow amplification factor - depending on flow at rush hours and average level- and flow at the node. Therefore, the demand component is present at both levels. They then apply the defined algorithms to the network of Italian highways with a series of different scenarios based on the values of the parameters and then define the amount of charging stations that should be set by 2030 and 2050 in each Italian region. (Micari et al., 2017)

Similarly, Shahraki et al. (2015) have analysed the paths covered by almost 12 thousand taxis in Beijing over a three-week time frame to identify the optimal charging station placement so to "maximize the amount of vehicle-miles-traveled (VMT) being electrified" (Shahraki et al., 2015, p. 166) . Since there were currently 40 available charging stations, they aimed at finding the 40 optimal locations for them. Based on the results, the optimized locations would lead to a 59 and 88% increase in the observed measure (VMT) - for slow and fast charging, respectively.

On the other hand, Andrenacci et al. (2016) have focused more on private vehicle flow analysis to provide a first approach in understanding appropriate location for public or private charging stations. In order to do so, data collected in the urban area of Rome by GPS tracking devices for insurance purposes have been used. Despite accounting for only 6% of the whole private flow, this number could reasonably indicate the fraction represented by electric vehicles and in any case the solution can easily be scaled to larger amounts. Here, the identification of potential location is done through the use of clustering with a K-mean model, with the centroid being the candidate location for the given cluster. (Andrenacci et al., 2016) Despite being a popular approach and allowing to measure the volume of flow happening within an area, it does not consider the directionality of routes. For this reason, models based on communities' identification are preferred in this thesis.

A visual approach to identify potential location is used by Qiao et al. (2018). In particular, they use a Voronoi diagram, as it "guarantees reasonable separation of the existing charging stations into service areas" (Qiao et al., 2018, p. 3). Indeed, such graphs divide the surface so that each point in the plan has a distance to the central point of the area where it is located

lower or equal to any other centre. In addition, this paper focuses on demand analysis as well, but in this case it is calculated as a function of waiting time to charge the consumers' vehicle. Therefore, after having identified services area in a region, an optimization algorithm to minimize the waiting time is developed. From a real-life application to the city of Shanghai, they are able not only to identify the busiest charging station, but also to derive an overview on users' behaviour. In fact, it results that the average probability of charging at home is above 60% compared to using a public station, and only when the battery level is below 20% the latter probability is higher than the first.

A considerably different perspective is suggested by Luo et al. (2017), who focuses on trying to maximise the service providers' profits. Nevertheless, reducing the negative impact on the power infrastructure and satisfying the quality-of-service index are criteria still taken into consideration as they have an impact on the potential profit. In order to solve this optimization problem in the context of an oligopolistic market structure, a nested logit model is used to calculate the optimal combination of the three type of charging to reach the desired result.

Finally, another relevant aspect related to charging stations deployment and more generally to electric vehicles is the environmental impact. Indeed, it can be seen as one of the main diver for transition towards EV (Hosseini & Sarder, 2019). In particular, Donateo et al. (2015) analyse the emission level of carbon dioxide and other chemicals polluting the environment such as CO and particulate matter and compare the impact from electric and fuel vehicles. In addition, these numbers are weighted up to the emission limits sets by the European legislation. The approach involved analysing the behaviour of user of EV charging stations in Rome in 2013. In general, results show that emission level of the observed particles are within the limits set by legislation and inferior to those of traditional vehicles. Additionally, researchers have observed that contrary from expectations, the periods with the "highest number of recharges (10–12 am, 1–3 pm) are also the best to recharge from the environmental point of view" (Donateo et al., 2015, p. 684) as it is when renewable energy sources make the highest amount of energy available.

This same aspect of environmental impacts is studied also by Hosseini and Sarder (2019), as element of the broader concept of "sustainability". Indeed, they explore 11 sub-criteria

that fall under the umbrella of environmental, social and economic aspects covering the overall concept of sustainability. More specifically, the deployment and maintenance cost are the main drivers of the economic dimension, while the social one covers security and impact on users' life. Finally, environmental influence is measured as a metric of water usage and waste generation as well as emissions. The main innovative aspect of this research is that it does not only take into account quantitative aspects but qualitative aspects that have been overlooked in the past. In order to include these, a Bayesian network model is adopted, as the use of various types of variables facilitates quantifying risk, making the model frequently used in the risk assessment context. The resulting model helped in assessing the optimal location to deploy charging stations as the one with the highest probability, as well as identifying the technical and the social aspect as the most and least important for the decision process. (Hosseini & Sarder, 2019)

4.3 Graph Theory Applications in Transportation Networks

As it is possible to note, coverage analysis studies are generally associated with the resolution of optimization problems, having various possible criteria as constraints. One aspect that is touched upon only marginally by Micari et al. (2017) is the origin-destination path analysis. This approach is closely related to network analysis theory and therefore use different models from the ones presented so far. In particular, the most relevant are the Flow Refueling Location Model (FRLM) and the p-median model, both presented in this section.

A relevant application of graph theory in the field of logistics is represented by the flow-allocation model initially developed by John Hodgson et al. (1996), which aims at satisfying flow demand considering the shortest path between starting and destination location. Given the point in time during which it was developed, it was assumed that a single refuelling facility along a path would have been sufficient. However, when considering vehicles using alternative fuels, this assumptions becomes obsolete, as the battery capacity can limit the route range considerably. (Kuby & Lim, 2005)

Kuby and Lim (2005, p. 127) have developed a "mixed-integer programming formulation" applicable to the Flow Refueling Location Model with the constrain of having refuelling stations exclusively at network nodes. This paper presents an example of "location-allocation model", aiming in this sense at locating facilities and allocating them demand. The result of

their test shows, as expected, that the lower the route range, the lower the number of paths that can be covered with a certain number of refuelling stations, when positioned at nodes. Moreover, for each additional facility the net addition of flows covered increases compared to the previous one. Nevertheless, given the constraint of having stations only at network nodes, it is possible that the coverage cannot be fully satisfied. This means that the volume of flow refuelled may not reach 100% if it is not possible to cover all existing paths with the available nodes. For this reason, Kuby and Lim (2007) has tested 3 different approaches to expand the initial idea of the Flow Refueling Location Model by allowing stations to be place along edges as well. Among these, the Added-Node Dispersion Problem (ANDP) results in having the overall best performance and all of the three tested models perform equally or better than the constrained model. Since nodes are better location than edges, given that they can serve all the paths crossing it in any direction, on average 86% of the refuelling stations are placed at vertex, while only 14% are on the edges. Nevertheless, this portion is fundamental to reaching the full coverage.

Another limitation of the initial model is the exclusive focus on the number and location of the refuelling stations, overlooking the nature of such station and more specifically their capacity, meant as the number of vehicles that can be refuelled simultaneously. In the case of hydrogen this applies especially to those stations creating their own hydrogen (Upchurch et al., 2009), however this issue can easily be translated to the context of electric vehicles as there is only a constrained amount of plug in each station and even more relevant is the potential impact of stations on the electrical grid. Through the Capacitated Flow Refueling Location Model (CFRLM) as defined by Upchurch et al. (2009), they aim at solving this issue by introducing a capacity constraint on stations. This model aims at maximizing the used capacity and at the same time it prioritizes covering path of flows requiring a lower number of refuelling spots compared to others. The approach is then tested on the area of Arizona and allows to find the most appropriate balance between number of stations and marginal contribution given by the additional station to the overall coverage. Nevertheless, since the CFRLM uses a greedy approach, the result could be sub-optimal as it identifies the best available solution, which main not correspond to the globally optimal one.

Another approach to solve the coverage problem with the use of graph theory is through

p-median models. These aim at minimizing the "total weighted distance travelled" (Upchurch & Kuby, 2010, p. 751) as opposed to the flow-refueling location model where the goal is maximising the quantity of routes that can be refuelled considering the shortest path. Similarly to the FRLM, also the p-median model was initially developed without taking into account route length constraints as at the time alternative fuels were not used in scaled production (Hakimi, 1964). The first approaches on alternative fuels have been published in the first half of the 2000s and are mainly focused of hydrogen applications, but can be extended to electric vehicle charging stations as well. Nicholas et al. (2004) tried to find an efficient balance between number of refuelling stations and customer coverage, given the high cost of hydrogen refuelling stations. More specifically they adopted a geographic information system model (GIS) as p-median technique, using as a base the network of fuel stations. Result showed that considering only 30% of the number of existing fuel stations would only increase the driving time to the potential (hydrogen) refuelling station by 16 seconds, in the metropolitan area observed. As a further analysis, the research was extended (Nicholas & Ogden, 2006, p. 1) to the main metropolitan areas in California. In addition, instead of deriving the required umber of stations from the population density level in the area, the "average driving time to the nearest station (convenience metric)" is used. In this way, the density would not affect the final amount.

Another model based on the extension of the p-median is the "fuel travel-back" approach developed by Z. Lin et al. (2008). Here both nodes and edges as considered as candidate locations, with the likelihood represented by the distribution of the vehicle miles travelled. In this case as well, the research is concerned with the number of hydrogen refuelling station to deploy. Based on the study results, a number of stations corresponding to 18% of the traditional fuel ones would be sufficient.

Finally, Upchurch and Kuby (2010) have compared applications of both the p-median approach and the flow-based FRLM. Here, the main objective is to evaluate which of the two performs best in reaching the other model's goal. Indeed, the p-median is more focused on refuelling close to the starting point, while FRLM focuses on refuelling on the way. The study shows that the FRLM performs better in terms of p-median objectives than vice-versa. From this, it can be derived that the first model is more capable of satisfying both theories

of placing stations at the users' home or along their way. Additionally, it is noted that the latter model is more sensitive to flow demand and that the results are also affected by the scale, with the p-median model showing an increasingly worse performance with a reduction in size of the model.

4.3.1 Community Detection Algorithms in Transportation Networks

Community detection algorithms are especially popular in social sciences, where they find a florid environment for applications in the context of social network analysis. As an example, identifying communities through social networks can be useful to tailor marketing recommendations as well as sensing sentiment towards a topic or again to identify community leaders. Nevertheless, network structure can be identified in various context, including that of transportation network. In particular, community structure is a relevant feature that can be helpful in better understanding the network structure and in supporting network design decisions (Oubaalla & Benhlima, 2018).

Indeed, multiple studies have the goal of getting a better understanding of the transportation network or the urban structure. As an example, this is the focus of Li and Zhang (2016), as a way to provide city planners the tools to improve and develop new management policies. Firstly, travel data from three different transportation modes - bus, taxi and trains - are used to define the overall network structure and afterwards community analysis is used to identify sub-networks and discrepancies among them. Based on the results of a real-life application to the city of Guangzhou, taxi and bus layers show similarities, while the rail network is much more independent. On the other hand, from a community perspective, the main criterion for separation is the belonging to different sub-regions or municipalities.

Taking this concept a step further, Majima et al. (2014) use community detection to generate bus line routes in the context of a Public Transport Network. The rationale behind this consists in assuming that a community can be seen as a potential route for a bus line. In order to solve the problem of computational complexity caused by the many criteria affecting transportation efficiency, this study adopts a Multi Agent System, where bus lines compete to increase the number of passengers picked up along the route. Based on this approach with a higher number of initial agents there is an higher likelihood for an optimal solutions, but a negative impact on the number of steps required to reach the final solution. Therefore,

even if the final solution is not optimal, the resulting quality and the increased speed are satisfactory aspects. Like in the previously mentioned case, oftentimes the dimension of networks makes time and computational complexity of methods difficult to manage. For this reason, Wandelt et al. (2021, p. 1) developed two "network attack strategies" specifically tailored to transportation networks using community algorithm approaches to remove interlink edges between communities. This helps in evaluating and improving the robustness of a network though community analysis.

In the studies presented in this section up until this point, the networks' structure was derived from GPS data of the vehicles analysed (Li & Zhang, 2016), or it was the goal of the research to build the optimal one like in the latter case. However, data for analysis can also be inferred from other sources such as data directly collected from the users rather than from the vehicle itself, which would be especially beneficial in the case of trains or busses where multiple people with different journeys are travelling at once. This is the case of a study carried out by Yu et al. (2020, p. 1), who are analysing travellers phone data to build a network and use community algorithms to define the cross-regional commuting demand with the goal of helping improve the network infrastructure in terms of "liveability and sustainability". This is deemed to be a relevant aspect in the Chinese socio-cultural context as the job-housing separation has increased commuting habits and consequently the pollution and traffic levels. Consequently, these data help in reconstructing the main patterns between residential and work areas as well.

Transport infrastructure improvement has demonstrated to be one on the most prominent applications of community algorithm applications in the infrastructure network. Nevertheless, there are example of studies defined with other goals as well, including traffic accident analysis, in which the research by L. Lin et al. (2014) represents the first example. Here, the modularity-optimizing community detection algorithm is used as preliminary step to cluster accidents to decrease heterogeneity before applying an algorithm to identify patterns within each cluster. From its application, accidents are divided into 8 different clusters, based on criteria such as: number of lanes, travel direction, weekday (or weekend), weather conditions... This strategy improved the result of the later applied algorithm, as it is was able to detect more patterns compared to its application on the non-clustered dataset.

While applications to the transportation networks have generally increased in number in the last decade, specific subdomains still have limited literature available, such as in the case of logistics industry. Despite this, the advantage that community detection algorithms could provide is considerably high, especially in trying to capture economies of scopes deriving from the identification of "symmetric flows between clusters" which could help in reducing empty loads returns (Mesa-Arango & Ukkusuri, 2015, p. 1). In this paper they develop a framework able to satisfy a threefold objective while still being efficient from a computational perspective: capturing network inter-dependencies, introducing a pricing component in clustering and considering price and volume uncertainties. This would help in grouping together paths based on synergies rather than geographical proximity.

Studies on community algorithms in logistics are not only limited to the analysis of transportation patterns, as Beckers et al. (2018) shows. Here, buyer-supplier geographical relations are observed alongside with available data on employment distribution, with the purpose of identifying co-location links between firms in the Belgian logistics industry through a novel quantitative approach. More specifically, both datasets are provided by the National Bank of Belgium, where one refers to employment data (employment in full time equivalents) in 2010 in Belgian municipalities and the second to microeconomic data mapping the inter-firm relations through company invoices where at least one of the parties is in the logistic industry. In this way, they were able to identify three different logistics concentrations: clusters serving larger regions, "spill-over cluster" in the neighbouring hinterlands and polycentric cluster connecting firms to the rest of the network. (Beckers et al., 2018)

At the same time, concerns about the possibility to adopt network analysis to the logistic sector have arisen. In particular, network nodes are assumed not to provide a large amount of information on the local context, in this sense removing the heterogenous component, as all nodes in a network are considered equal, disregarding the real-life scenario in which they are placed (Beckers et al., 2019). Therefore, the goal of Beckers et al. (2019, p. 316) is to combine a more local perspective with a network related one, intended as the "identification of its local, regional and national role and related connectivity." For this reason, they perform multiple iterations of community algorithms, in order to better understand the hierarchical layers of the Belgian logistics network based on the framework provided by Beckers et al. (2019)

Similarly,Badiee et al. (2020) tries to reconstruct the transportation network structure and existing communities through a collaboration network of drivers, based on human relationship interaction among them. The goal of the study is to "improve the transportation administrations performance" (Badiee et al., 2020, p. 2). As an example, this entails improvements in the time needed to allocate a freight to a driver based to their belonging to the community and their centrality and optimize the planning based on vehicles and drivers shared among multiple communities. On the other hand, this approach can generate advantages for driver as well: they can receive information about orders or traffic behaviour based on the community they belong to.

5 Methodology

In the previous section, we examined the relevant literature and gained insights into how researchers have tackled similar problems using the methods we will employ in our analysis as well, such as Graph Theory and community detection. Building on this knowledge, we now turn to our own approach for addressing the business problem at hand. In the following section, we will outline our methodology and approach and provide a detailed description of our analysis, drawing on the insights gained from the literature review, and introduce the data we were provided with by DFDS to carry out our analyses.

5.1 Theoretical Research Methodology

As a theoretical methodology framework, this thesis is based on the "CRoss Industry Standard Process for Data Mining" (Chapman et al., 2000), also known as CRISP-DM, a typical data mining project management framework. Given the central role data have in this thesis, this framework provides a useful and structured approach.

While multiple models have stemmed from CRISP-DM since its initial theorizing, however, based on surveys it is still considered the standard procedure for such projects (Martínez-Plumed et al., 2021). Therefore, this framework is going to be used as a base approach, knowing that it allows for flexibility and for variations to be derived from it. More specifically, the overall process can be segmented in six different sections, presented below. While the structure initially suggests a sequentiality in executing the tasks, the circularity and the

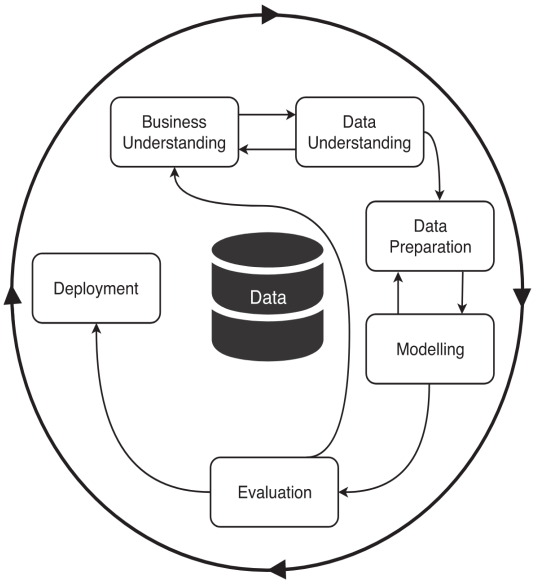

Figure 4: A visual representation of the CRISP-DM framework (Chapman et al., 2000)

interactions also underline that each phase is not rigid and that they can be updated and revised throughout the process (Figure 4). A procedure that should be maintained throughout the whole research involves keeping records of each sub-step in order to have a structure documentation of the overall project execution.

- Business understanding: firstly, it is important to frame the problem in order to understand it from a business perspective. In this context it concerns getting a better understanding of DFDS' operations and, more specifically, their eTruck project objectives. This is important as it helps assessing the current situation in terms of resources available and existing constraints. Based on that, it is then necessary to set clear objectives and define a plan for the overall project and clear success criteria to evaluate the project outcome.

- Data understanding: the second step entails collecting and/or obtaining the data before providing an initial description, regarding format, size and feature names. The next subsection would then focus on data exploration with the goal of gaining a deeper understanding of the data features, their value distribution and providing simple aggregations and data manipulation. This also helps in ensuring data quality, by verifying if and what data points are missing and if others are incorrect.

- Data preparation: the third step begins with selecting the data to be analysed, based on

quality, volume and how they can contribute to reaching the overall goal. In addition, it may be required to perform some data cleaning based on the quality status determined in the previous section and possibly find a strategy to input missing data. Moreover, it may be useful to construct new features based on the those available as well as generate new variables that could provide additional value in the modelling part. Finally, especially when working with multiple data sources, it could be meaningful to join the data together in order to collect all relevant information in the same table.

- Modelling: following on the model decision taken in the first step, this section involves applying and fine tuning the chosen data science model or approach. While this refers more generally to machine learning, statistics or database systems, it still has been deemed appropriate for this context - which the generation of visualizations and the use of Graph Theory techniques - as the pipelines created allows for the identification and analysis of patterns, which is the main purpose of data mining.

- Evaluation: this section regards model evaluation and model review, in case it is deemed necessary. This can typically mean evaluating whether performance metrics show satisfactory results and if the model generalises well. In this specific case, model evaluation concerns with making sure that the measured metrics display values in the valid ranges and if results demonstrate to be consistent with the expectations and the assumptions previously made.

- Deployment: in this last phase, results of the models are analysed and a potential strategy for the model deployment is proposed. For this project, an integration with the company current infrastructure is developed. Additionally, a final comprehensive report is produced, corresponding to the thesis at hand.

Finally, in order to provide a clear overview on the use of the framework within this project, a summary in Table Table **??** is here presented.

CRISP-DM overview		
Stage	Section title	Chapter
1. Business understanding	About DFDS	1
	Introduction to the Business Problem	2.1
2. Data understanding	Research Framework and Approach	5.2
	Data Source	5.3
	Data Description	5.4
3. Data preparation	Data Pre-Processing and Exploratory Data Analysis (EDA)	5.5
4. Modelling	Bookings Volume, Geographical Distribution of DFDS Flow, Flow Density Analysis of Focus Countries	6.1-6.3
	Graph Building and Underlying Assumptions	7.1
5. Evaluation	Dashboard	6.4
	Analyses	7.2
6. Deployment	Research Findings and Insights	8.1
	Implementing the solution in the current infrastructure	8.2

Table 1: Overview of the theoretical framework used in this project and related chapters

5.2 Research Framework and Approach

Building on the insights of the previous more theoretical sections, this paper now turns back to the case at hand. In coordination with DFDS, a two-fold approach has been identified in order to analyze the route network and identify possible sites for charging stations. After an in-depth analysis of the underlying data, (1) a dashboarding solution for Visual aAalytics of the data will be provided, followed by (2) a more in-depth analysis of the transportation network leveraging techniques from Graph Theory.

There are multiple reasons this approach has been chosen. First of all, building a dashboard is a useful approach for getting a first overview of the data at hand, especially when working with large and complex datasets. In the discussions with DFDS, it turned out that no holistic dashboard containing fast insights into their routes and operations exists across the organization, so building such dashboard would not only provide us with fast insights into the data we're working with before starting deeper analyses, but also create immediate value to DFDS.

However, while providing fast and interactive insights, dashboards do have some limitations when working with large amounts of data about a transportation network, just as in our case. It is especially the complex relationships between different elements in such networks – e.g. the presence of communities, very well connected sites, or the efficiency of different routes

– that can be difficult to capture in a simple dashboard that is designed to provide a holistic overview of the data. At this point, Graph Theory comes into place. As mentioned in the Literature Review section before (section 4), Graph Theory techniques such as community detection algorithms or centrality measures are commonly used when analyzing large transportation networks as it allows to deep dive into the exact relationships within the network we just mentioned.

Piecing all this together now, the structure for this following analytical section is as follows: to begin with, the data that this thesis builds on is introduced and findings of explanatory data analysis (EDA) are discussed. The data is then being prepared in two different ways: (1) to build a dashboard and (2) to represent the data in a graph structure (please find a more elaborate explanation why two different data formats are needed in the respective section below). The dashboard is then introduced and, based on the respective findings, focus areas are identified. Lastly, deep dives into the identified focus areas are performed leveraging Graph Theory techniques, which will eventually lead to tangible recommendations as to where DFDS should place charging stations for their eTrucks.

5.3 Data Source

Given the large dimension of DFDS and the numerous acquisitions that have characterized its development, for security reasons the company's main data warehouse (DWH) is residing on premises. Storing data on cloud would entail a series of advantages for a company present on a large geographical region such as DFDS, as cloud allows for unlimited, scalable and flexible storage as well as a high degree of mobility which would allow access from various locations. Additionally, pay as you go systems could allow higher cost efficiency as deliveries are considerably affected by yearly seasonality. More generally, it removes the need of owning hardware and ensures backup and recovery measures. Nevertheless, data migration from a on-premises to a cloud storage represent a highly complex process, requiring a tight collaboration with Subject Matter Experts (SMEs).(Capgemini, 2022) Additionally, it is possible that acquired companies rely on different systems that have to be integrated as well. For this reasons the ongoing process of digital transformation is considerably complex and will require some time to be completed.

The company's most important data storage solution is a software developed in house, *Velocity*. This tool groups data originating from a series of other platforms in tabular format and makes them retrievable though SQL queries. Among these, one of the most relevant is the software *Direct*, where customers can request deliveries and include additional requests about them. Based on such information, logistic planners would then decide what specific trucks to assign to each booking based on a series of characteristics such as location or type of delivery i.e., cold chain products. These additional details would then be inputted in *Velocity* as well. Additionally, also equipment data and DFDS logistics platform data would be integrated on the main platform.

For security, timing and technicality reasons, we were not assigned a DFDS account that would have allowed us to directly access data from *Velocity*. Instead, the relevant datasets for our analyses were extracted from *Velocity* by DFDS and made available to us via Microsoft Sharepoint. More specifically, three different files were shared:

- *Results.csv*: collection of bookings for the period 02/01/2020 to 23/01/2023, the day in which the dataset was retrieved and shared.
- *booking_with_legs_2021−2022.csv*: collection of bookings where each entry corresponds to a subbooking, having in this sense a subdivision in journey legs. The dataset covers the period 01/01/2020 to 30/04/2022.
- *collections_bookings_with_legs_2022.csv*: structured in the same way as the file *booking_with_legs_ 2021−2022.csv*, however it covers the time frame 01/05/2022 to 10/02/2023, the day in which the data was retrieved and shared with us.

The following section will now elaborate on how the data we received looked like and which data processing steps were necessary before continuing with the analyses.

5.4 Data Description

As logistics data – just like in this case study – is often very complex and non-intuitive at times, this section is essential in helping readers understand the context and reliability of the study's approach, findings and conclusions. It also serves the purpose of replication of the research and facilitates the evaluation of the study's validity and generalizability.

Essentially, the three datasets used in this study contain all DFDS logistic service operations between January 1st, 2020, and February 10th, 2023, with different levels of complexity. The data hence represents DFDS' route network, containing both the pick-up and drop-off points of bookings (these can either be DFDS entities such as warehouses, distribution centers, or the like or client sites such as distribution centers of a supermarket chain). It is very important to mention at this point that the data is delivery-based, meaning that a row in the data represents a delivery by a DFDS truck (= a booking); for example, a DFDS truck picking up a load for a customer at a DFDS distribution point and then delivering it to the client site. Along with the route, information about load (such as weight, temperature, etc.), customer, and time are being given. Please find a detailed overview of the variables in Table 22 in Appendix II.

However, there are differences as to how the data is represented between the three datasets. Coming from different sources within DFDS' own *Velocity* system, they convey similar information in a slightly different way, which requires some merges to handle and align these differences. Essentially, bookings are represented in a different way in the *Results.csv* than they are in *_with_legs_1.csv* and *df_with_legs_2.csv*. As the name suggests, the latter two split up a booking into multiple sub-bookings, where applicable, which essentially means that they also contain information whether a truck, on his way from the pick-up point to the drop-off point, had one or more stopovers to drop parts of the load. To illustrate this, please see this example of BookingId '9031814':

	BookingId	FirstCollectionCity	FirstCollectionCountry	FirstCollectionLongitude	FirstCollectionLatitude	LastDeliveryCity	LastDeliveryCountry	LastDeliveryLongitude	LastDeliveryLatitude
2703963	9031814	Liverpool	United Kingdom	-2.85863	53.47114	LEEDS	United Kingdom	-1.47493	53.72657

Figure 5: Example of BookingId entry in the dataframe originated from Results.csv

	BookingId	SubBookingName	SubBookingLegId	FromLatitude	FromLongtitude	FromCity	FromCountry	ToLatitude	ToLongtitude	ToCity	ToCountry
0	9031814	A	11006604	53.47114	-2.85863	Liverpool	United Kingdom	53.37030	-1.36706	Sheffield	United Kingdom
1	9031814	B	11006603	53.47114	-2.85863	Liverpool	United Kingdom	53.72657	-1.47493	LEEDS	United Kingdom

Figure 6: Example of BookingId entry in the dataframe originated from df_with_legs_2.csv

We can clearly see that *Results.csv* (Figure 5) only depicts this booking as a delivery from Liverpool to Leeds in the United Kingdom, while we can extract the information from *df_with_legs_2* (Figure 6) that the truck had a stop in Sheffield along the way. This information

is fairly crucial for us, which is why we decided to merge the three datasets together - with *df_with_legs_1* and *df_with_legs_2* as basis - in order not to loose these important route information. The merge was straightforward as most of the columns aligned, however, additional information about whether a truck was fully loaded (FTL) or not (LTL) as well as about the customer associated with a booking that have only been present in *Results.csv* were added in the course of the merge.

At this point, a fairly clear picture of the data we are working with in the upcoming sections should be present. That's why we are now moving on to the section about data pre-processing and exploratory data analysis (EDA), where we briefly elaborate on the most important steps we took to get the data ready for Visual Analytics and graph building.

5.5 Data Pre-Processing and Exploratory Data Analysis (EDA)

Instead of providing a detailed overview of the nitty-gritty EDA steps, this section wants to give an overview of the four main goals the data pre-processing needed to achieve: changing the representation of a route, calculating the travel distance of each route, preparing the data in three different ways – delivery-based, route-based and location-based – for Visual Analytics and graph building, and, lastly, gaining some first insights.

1) Changing representation of a route

When we briefly introduced how sub-bookings are represented in *df_with_legs_1* and *df_with_legs_2*, looking at the screenshots in the section above the attentive reader might already have realized that a major issue with the way the sub-bookings are being set up exists, which requires some handling. Sticking to the example with BookingId '9031814', we can see that a delivery from Liverpool to Leeds via Sheffield contains two sub-bookings: Liverpool to Leeds, and Liverpool to Sheffield. Obviously, this makes sense from the booking perspective, as DFDS is serving multiple customers with one truck: for one of them, a load is being delivered from Liverpool to Leeds, and for the other one, the delivery is being made from Liverpool to Sheffield using the same truck. The problem with this representation, however, is that when analyzing the routes, the data suggests that the two routes that the truck drove are the routes Liverpool – Leeds and Liverpool – Sheffield, while, in the real life scenario, it is more likely

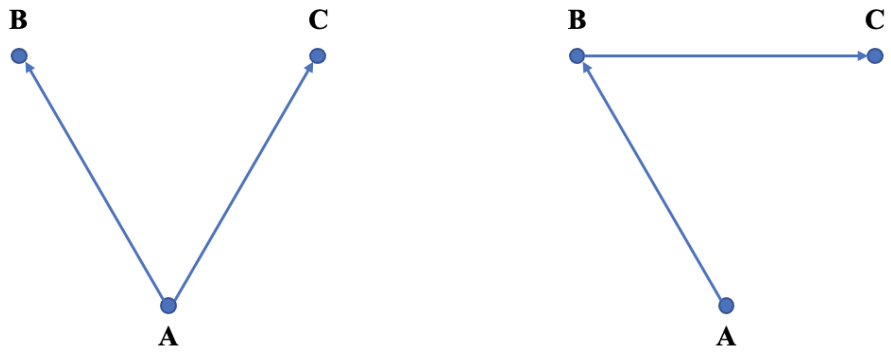

Figure 7: Example of how the function fix_legs acts: the starting coordinates of each leg are updated to match the ending coordinates of the previous leg

that the truck took the route Liverpool – Sheffield – Leeds. Please see the visualization of this problem in a generalized way below (with Liverpool = A, Sheffield = B, Leeds = C) in Figure 7.

Because of this, we majorly needed to restructure the data. We handled this in a way that we built a function, fix_legs, that first checks for re-appearing BookingIds, which means that the respective booking consists of multiple sub-bookings (legs). Essentially, this function then fixes any inconsistencies in the coordinate data of the delivery legs, by updating the starting coordinates of each leg to match the ending coordinates of the previous leg. In other words, the function replaces the information about the pick-up location of the delivery – 'FromLatitude', 'FromLongitude', 'FromCity', 'FromCountry' – with the drop-off information of the row before. To refer back to the example from above, this changes the routes in a way that it's not Liverpool – Sheffield and Liverpool – Leeds, anymore, but Liverpool – Sheffield and Sheffield – Leeds (which then fits the representation on the right side of Figure 7 above).

This step is absolutely crucial for further analyses and helps to ensure that the delivery routes are correctly identified and analyzed.

2) Calculating route distances using Bing Maps API

Now that we fixed the delivery routes, the next very crucial step is to find out about the distances of these routes – an information that was not given in the original datasets by DFDS. However, knowing the exact distance of a route is essential in our case as eTrucks

have a very limited reach ($\approx$ 300 km) and DFDS needs to be able to assess which routes they would be able to cover with eTrucks based on this.

We want to stress the "exact" part as there indeed are relatively easy measures to quickly calculate distances between coordinate points, such as the Euclidean distance or the Haversine formula. However, these measures alone may not be sufficient as they do not take into account the complexities of real-world transportation networks. The Haversine formula assumes a spherical earth and does not consider factors such as the road network and topography, while the Euclidean distance only measures the straight-line distance between two points and does not consider obstacles such as buildings or rivers that may require detours. Simply spoken, both measures just don't take into account that streets aren't straight lines, and the Euclidean distance as well as the Haversine formula differ significantly from the distances that are actually being traveled when a truck drives from a Point A to a Point B. Thus, to obtain more accurate distance measurements in transportation network analysis, other distance measures should be used.

In order to do so, we decided to calculate the route distances leveraging the Bing Maps API (Microsoft, 2022). Being a location-based service API that provides developers with a range of geospatial features and functionalities, one of the key features of the Bing Maps API is its ability to calculate distances between coordinate points. This feature is particularly useful four our case and transportation network analysis in general, as it allows for accurate measurement of travel distances between delivery locations.

To calculate distances between coordinate points using the Bing Maps API, we made use of the API's routing service. The routing service can be accessed via a RESTful web service, which accepts input in the form of start and end coordinates, as well as other parameters such as mode of transportation (e.g., driving, walking, or cycling) and routing preferences (e.g., shortest distance, fastest time). Since the dataset provides us with the start and end points of a delivery in coordinate form, this was easily possible. We extracted the unique routes - corresponding to the unique combinations of start and end points - within the original dataset, which resulted in around 220,000 unique routes we needed to calculate the distance for. Next, we defined the mode of transportation ('travelMode') as "driving" for obvious reasons but did not specify the routing preference, as in network analysis both distance and time are important measures and we hence did not want to give constrains here.

When a request is sent to the routing service with the start and end coordinates, the API calculated the distance between the two points based on the selected mode of transportation. The distance returned by the API will take into account real-world factors such as road networks, traffic congestion, and detours, hence providing a more accurate measure of the actual travel distance between the two points compared to measures like the Euclidean distance. We did this in multiple batches for the 220,000 unique routes, as the educational key we were provided with by Bing only allows for 50,000 API calls per day.

3) Preparing different datasets based on deliveries, routes, and locations

After an initial exploration of the data and taking into consideration the different approaches through which we analyse them, it seemed appropriate to generate three different datasets from the pre-processed version of the initial data.

Firstly, dropping columns that do not provide any value or are not used in the analysis makes data more manageable and easier to work with. For this reason a "slimmer" version of the original dataset is created, keeping the same structure of the original data, in a delivery-based format. As the datasets we were provided with were delivery based, this did not change the essential structure of the data.

However, in order to build the graphs afterwards, the dataset needed to be transformed from a delivery-based representation to a route-based representation, where each row doesn't represent a single delivery anymore but a unique route in the network with additional information about how frequently this route has been taken. This is also reducing the datasets' size significantly from about 5 million rows down to about 200,000 rows. Going from there, graphs representing the transportation network can easily be constructed using the aggregated data, providing a more accurate representation of the network's structure and connectivity. A graph is a collection of nodes (representing points A and B in this case) and edges (representing the routes between them). By representing the data in terms of unique routes and their frequencies, we can create edges between the nodes that represent the points A and B, weighted by the frequency of the route. This can give us a better understanding of the most frequently traveled routes and the busiest parts of the network. Moreover, representing the data in terms of unique routes and their frequencies can also make it easier to identify anomalies or outliers. For example, if a particular route is taken significantly more or less

often than expected, this could indicate a problem with the transportation network, such as a road closure or increased demand.

Finally, a third dataset is generated, in this case keeping track of the frequency each point is either a beginning or ending point for a delivery or part of it. This shifts the focus from the route per-se - like in the case of the route-based approach - to the the start and delivery location, in order to more generally assess which could be interesting location to analyse with more attention.

Following this process, there are 3 different dataset available, each of them designed with a specific purpose:

- *df_deliveries*: this dataset has 20 features and 4973604 observations, each corresponding to a delivery, which again corresponds to a booking or one of its legs. Therefore, in this case, the same structure of the original dataset is maintained.
- *df_routes*: this dataset has 12 features and 197129 observations. Here, each entry represents a route in DFDS distribution flow and since it's derived from the original datasets by grouping it based on routes, information about the frequency each route is travelled is provided.
- *df_locations*: this dataset has 9 features and 21168 observations. In this case the grouping is based on the locations, intended as departure and arrival point. In this context, the frequency each point appears in the original pre-processed dataset is provided, as a measure of the relevance of each point in the network

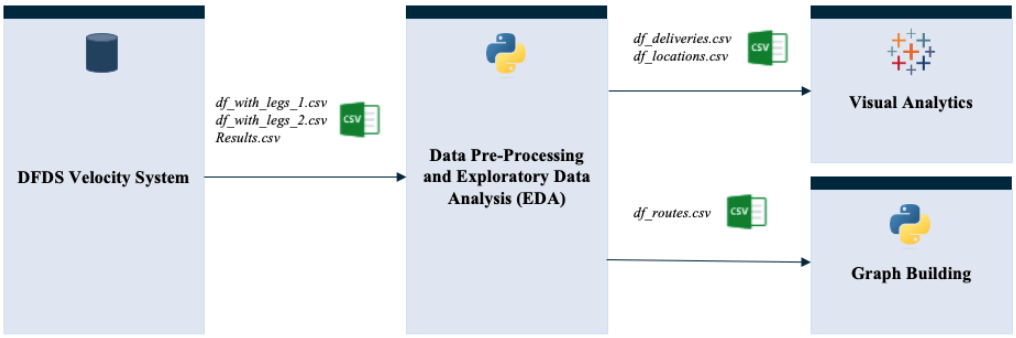

Figure 8: Data Flow Visualization

For each of them, a table summarizing name, type and description of the variables is available in Appendix II. Figure 8 illustrates the point touched upon above.

4) Some first insights

By having defined this main perspectives through which it is possible to analyse data, it is feasible to already obtain first insights. In particular, from the delivery-based dataset, an initial overview of the areas of interest for DFDS can be derived, by observing the distribution of the origin and destination countries.

UK, Sweden and Denmark result to be the top three county for what concerns both collection and delivery, with UK and Denmark having around the same number of departures and arrivals (Figure 9). This raised the question on whether the journeys are indeed within the same country and for this reason a variable indicating if the delivery is domestic - where the origin country and the destination country are the same - was created. It results that 65.7% of the deliveries are local, but still a good portion of them involves crossing borders.

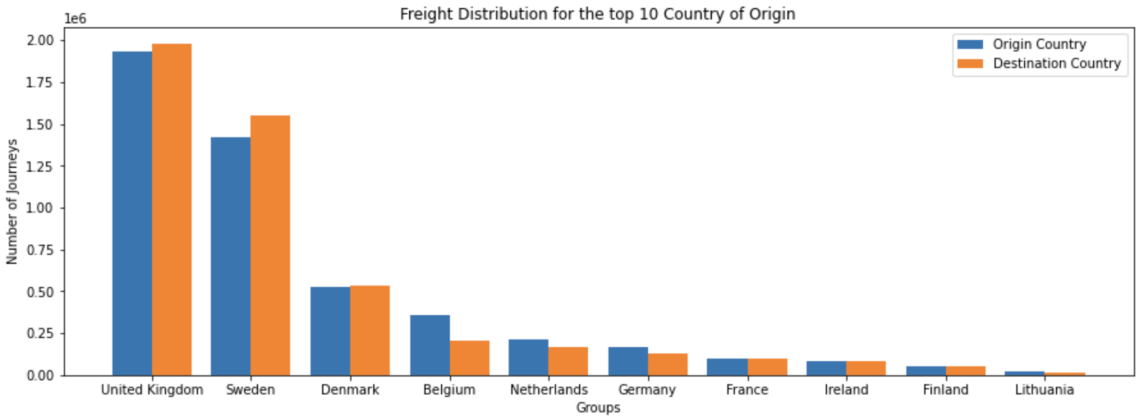

Figure 9: Number of journeys starting and ending in each of the top 10 countries of origin

In addition, it is possible to identify some features characterising the truckload of DFDS deliveries. In particular, the most interesting insights regard the temperature and the type of load. Nevertheless, as it will be discussed more in detail in section 9, these variables present a high number of null values, meaning that these data can only partially support us in understanding the type of load. As an example around 81% of delivery temperature are null values, however 13.7% of the available data indicated that the temperature is below 0°C. This indicates that DFDS deals with frozen goods, which is a relevant aspect to consider given

that frozen cargos can considerably impact the route range that can be travelled by eTrucks.

Based on the same rationale, having an overall understanding of the load type can be beneficial as well. Again, even if in smaller ratio compared to the previous case, 11% of the entries for the FullLoadIndicator variable are nulls. 57.3% of transports are FTL types of deliveries, indicating that in the majority of cases trucks will travel at their maximum capacity. Like in the case of frozen and refrigerated goods, it is an aspect to keep into consideration when it comes to evaluating eTrucks ranges of motion.

For what concerns the route-based dataset, the most relevant aspect is that the routes with the highest frequency are in reality 0km distance routes. The reason is that for organizational and planning purposes, moving of goods within warehouses are logged in the system as well, in order to ensure always having control on goods locations. This is for example the case of the two most popular "routes", happening in the main port in Gothenburg (Sweden) and in the distribution facilities in Vejen (Denmark), respectively. Consequently, since these movements of goods does not involve trucks and would not be relevant for this analysis, the values are dropped.

Finally, the location-based dataset poses the focus on location rather than routes importance, helping in identifying the main starting and ending locations of the travelled routes. The first 10 points by percentage of total starting and ending locations, are all either ports - the first, second and fourth most relevant point are located in Karlshamm (Sweden) and Immingham (UK)- or warehouses - like in Larkall and Grimsby (UK), with the third and fifth point. Moreover, no point represents more than 3.1% of the flow, with the seconds highest percentage being 1.58%. However, while this certainly indicates that the overall distribution network is largely scattered, it has to be noted that each point is represented by a combination of coordinates, as the level of detail would consider point geographically close as different entries in this dataset. As an example, the second and fourth most important locations are both located in the area of Immingham's port, but given the two slightly different coordinates they are entered separately.

After having gained an initial overview about the complex data we are working with and presenting a few preliminary insights into DFDS' transportation network, we can now move the data to Tableau for Visual Analytics, with the goal to identify focus areas within DFDS' network, which we will then examine further leveraging Graph Theory techniques later on.

6 Visual Analytics Approach

After the exploratory data analysis and data cleaning process, we still deemed relevant to generate visualizations with the purpose of deriving a complete understanding of the current DFDS infrastructure as well as being able to identify the main focus areas for further analysis. As presented in subsection 2.1, the company currently lacks a data-driven approach for identifying optimal locations for charging station placement. However, providing such approach is a necessary step to better understand the large and complex datasets DFDS can leverage on.

For this reason, the following visualizations have been created:

- a line graph, to give an overview of the volume of bookings in DFDS infrastructure over time
- a static spike graph with an overview of frequency for both starting and delivery point distribution, to get a first idea about important sites within the network
- a map indicating the major starting and delivery points over time, to add dynamism to the previous visualization and to carry out an analysis from a density perspective
- a dashboard, providing a clear overview of the current operations

The *df_deliveries* dataset has been used for the creation of all the visualization but the spike graph. In that case, the *df_locations* dataset was needed, since the graph is built based on the frequency percentage of each location in our data, which cannot be calculated in a straightforward manner in Tableau.

Based on these visualizations and a continuous dialogue with DFDS project manager for the current project it was then decided on which countries and area to focus further analysis.

While the functionalities to perform these analysis are offered by a variety of tools, with PowerBI and Tabealu being the most popular ones, the choice fell on the latter. This is mostly due to practical reasons: indeed, the department involved in the development on the EV deployment project is already familiar with the tool, as there are in-house experts working and generating dashboards in Tableau. Therefore, all the visualizations presented in this chapter are created with this tool even when the name is not made explicit.

6.1 Bookings Volume

Firstly, before understanding the geographical distribution of flows, it was deemed important to get a complete understanding of the current infrastructure. For this reason the analysis started with visualizing the volume of shipments overtime.

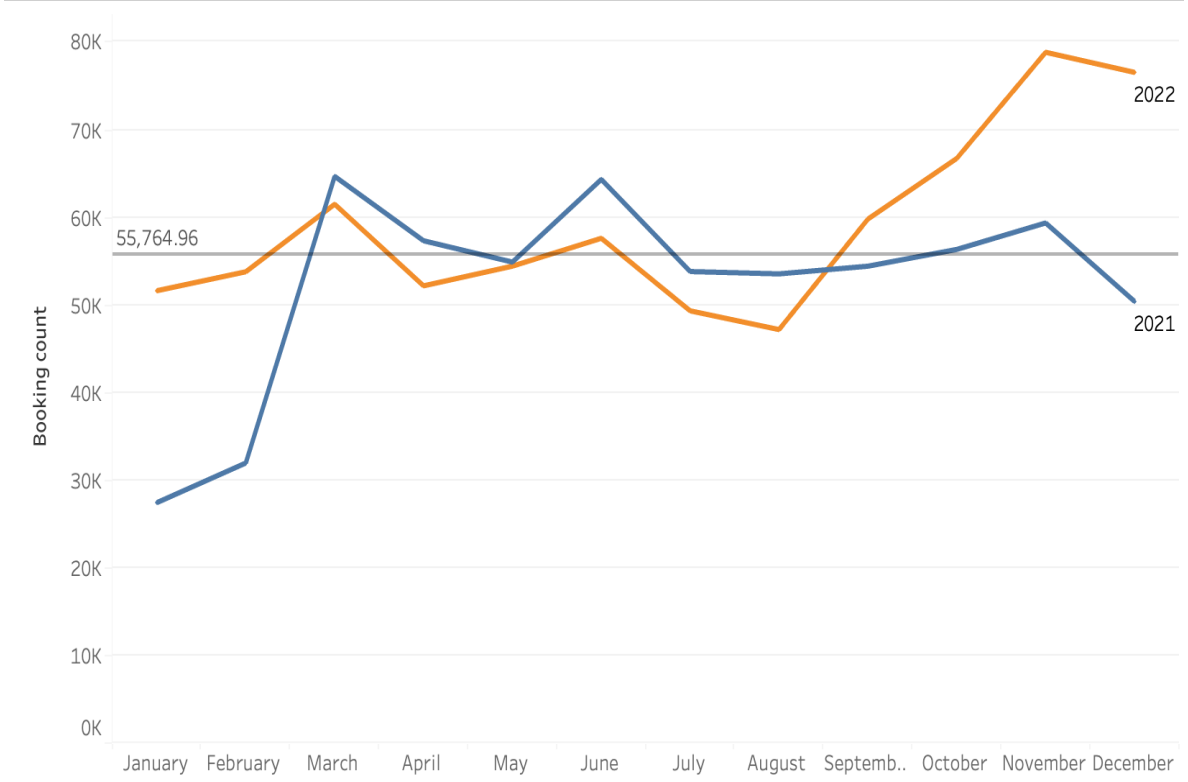

Figure 10: Month by month comparison of bookings volume for years 2021 and 2022

At first glance, it is possible to notice that the highest number of booking has been registered in November 2022 and the lowest in January 2021, for an average of around 56,000 deliveries per month. From the same graph it is possible to assume the presence of seasonality throughout the year. Indeed, both 2021 and 2022 show spikes in March, June - even if more pronounced in 2021 - and November. Similarly, the lowest points are in January, July and December, which could be associated with Christmas and summer holidays. Despite the overall similar seasonal behaviour, year 2022 is generally under performing compared to 2021 with a major increase in booking volumes starting from October 2022 and reaching a 20,000 booking increase compared to the previous year in December. With the purpose of investigating this, a breakdown by country was performed and it was discovered that

the growth is associated with higher volumes in Denmark. In order to validate the data, a discussion with DFDS project manager confirmed that this behaviour is the consequence of the acquisition of Danish companies and taking over their operations.

6.2 Geographical Distribution of DFDS Flow

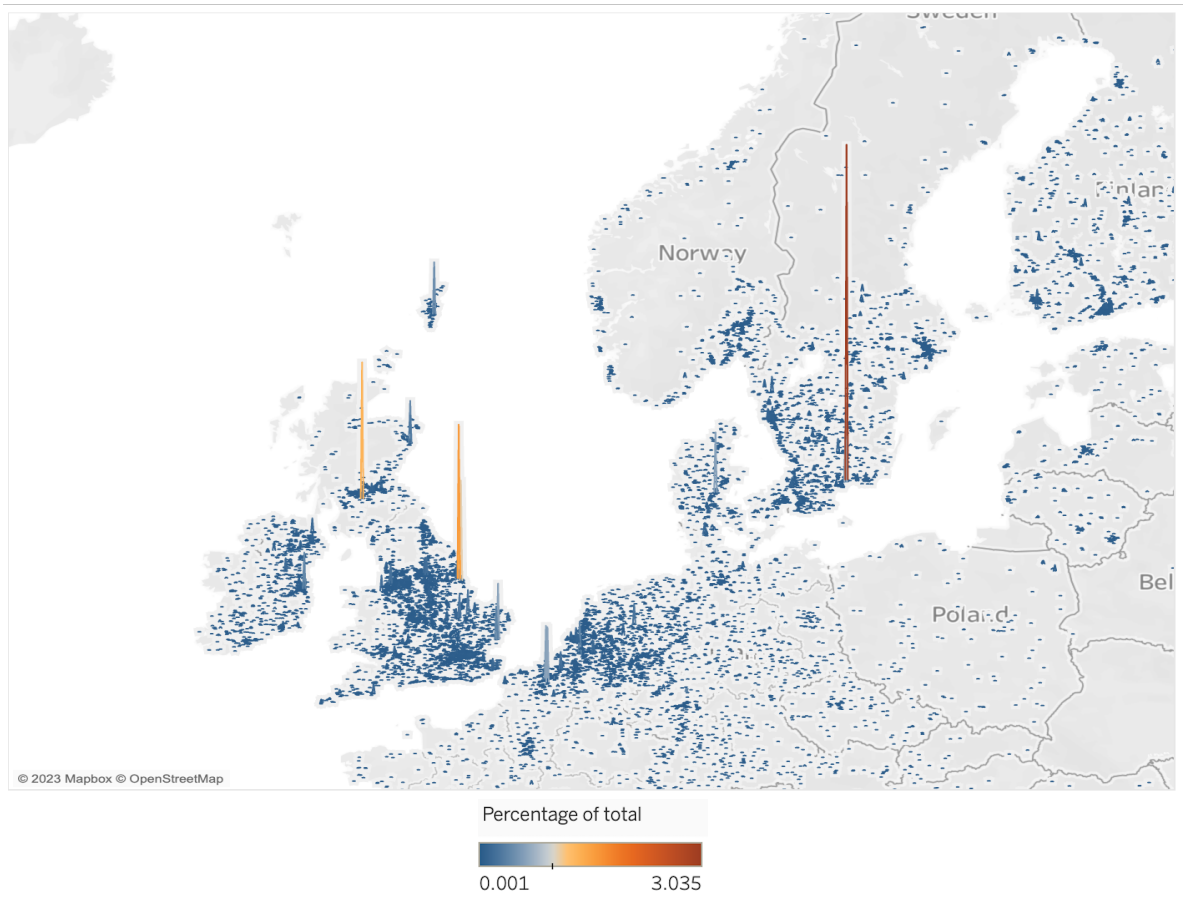

Figure 11: Spike graph indicating the percentage of total flow interesting each location

After a visual validation of bookings data, the second step involved obtaining an overview of the geographical distribution of flows. For this task a spike graph (Figure 11) has been chosen as it helps not only in identifying the areas in which there is a departure or arrival point, but the magnitude of the flow to or from those locations. Indeed, the height of each spike is proportional to the percentage of journeys departing and ending from the given location. Firstly, it is possible to notice a much higher density around Great Britain, Belgium, The Netherlands, Denmark, and the south of Sweden, with a generally progressive decrease moving

further away from those areas. The southern coast of Norway, Ireland, Germany and Northern Italy have a higher density compared to the rest of Europe, but each individual point retains no more than 0.015% of the traffic. In addition, particularly interesting are the highest spikes, located in the areas of Gothenburg (Sweden) and Hull (UK), which represents major ports for DFDS businesses, and in Glasgow (UK/Scotland) which is another major distribution point. In order to avoid result biased by the short journeys happening inside the port, which would likely not be carried out by long haul trucks, all trips with a route distance of 0 were filtered out.

Among the other high-frequency locations, the central-east part of the UK has points with high frequency in Manchester, Leeds, Peterborough and Norwich. Similarly, in Scotland two high spikes in Aberdeen cumulatively accounting for around 1% of beginning and end of journeys. Still in the UK, the Shetland Islands presents an high peak. However, this should not be considered relevant for the scope of this research as goods reach the island through sea freight and given the dimensions of the island, the distribution on site is not done by long haul trucks. On the other side of the English Channel, In the high density areas in continental Europe the most important locations are at border between France and Belgium on the Roubaix area and in Ghent.

Given the constraint of placing charging stations along DFDS routes, this visualization serves an important purpose in narrowing down the geographical areas in which the company should focus its investments. Clearly, the identified countries with the highest densities and frequencies represent ideal candidates from this numerical analysis. However, those results should be matched with the more qualitative data the company has collected on the presence of policies providing financial advantages for the deployment of electric vehicles in the given country. By discussing with the DFDS project managers, it was decided to narrow down the analysis to six countries: UK, Belgium, The Netherlands, Germany, Denmark and Sweden.

6.3 Flow Density Analysis of Focus Countries

Having defined a narrower selection of countries on which to focus the analysis allows to adopt a more thorough approach to identify key locations not only in absolute terms, but also throughout time. For this reason, it was decided to use a Page filter in this map visualization to better understand whether there is and increase or decrease of traffic month by month,

making it also possible to introduce a representation of the existing seasonality in a map. The period covered goes from January 2021 to January 2023, the months for which we have fully available data, with the exclusion of 2020 as it introduce misleading values due to the effect of the pandemic on specific product consumption.

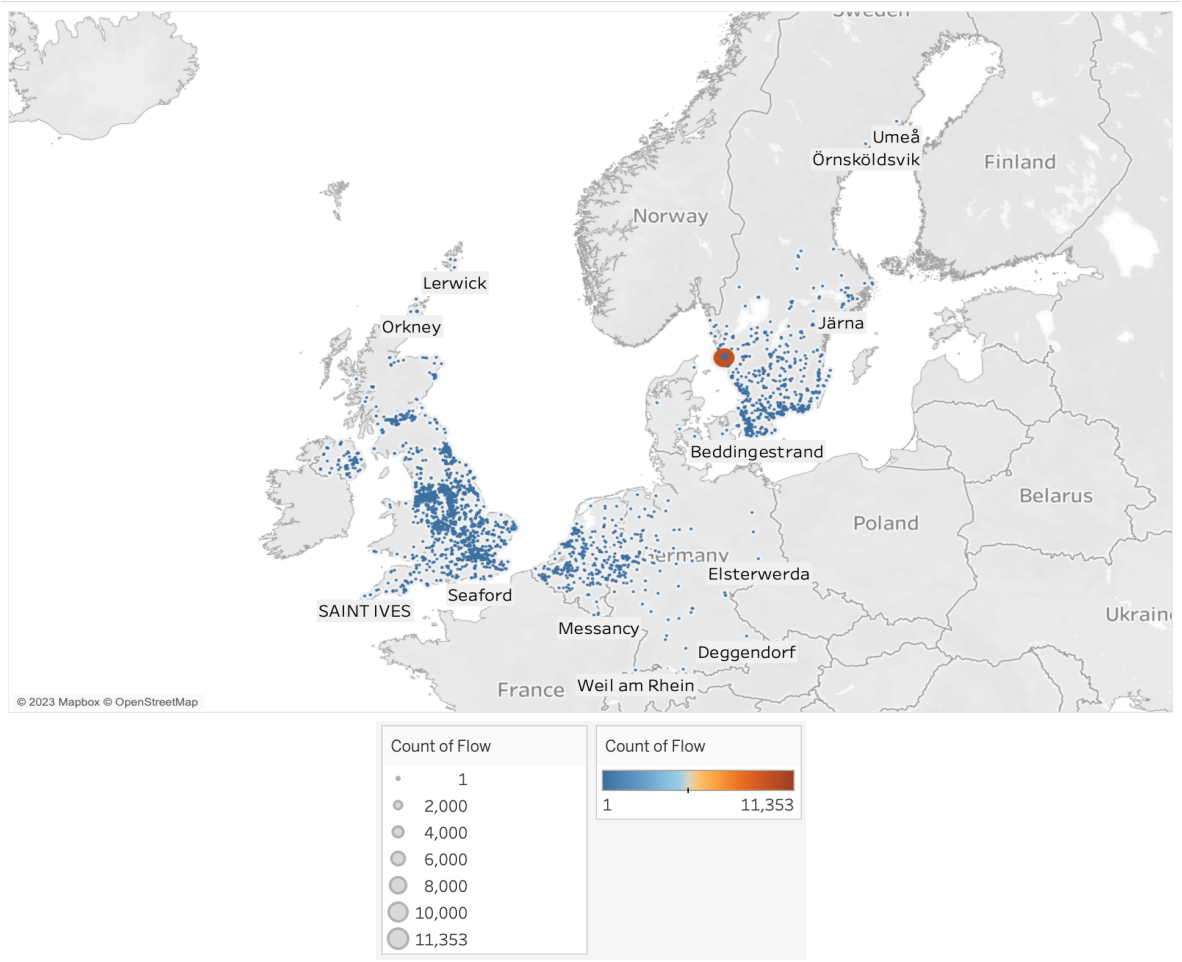

Figure 12: Flow Density Analysis of Focus Countries

In order to visualize the magnitude of each point in the distribution infrastructure, a map with Size Marks is used. More specifically, each circle in the map indicates the beginning or end of a leg, based on the latitude and longitude provided in the booking dataset. Since each location can have a different relevance, this aspect is conveyed in two ways: firstly, the higher the frequency count of a point, the larger the radius of the circle used to represent it. Secondly, the higher the same frequency, the darker the tone of the colour filling the circle, with blue being associated with lower values and red with higher ones.

Additionally, some filters have been added to make the visualization more dynamic. In particular, a route distance filter allows to select the range of route length to visualize, with the purpose of showing specifically routes that are in the scope of this analysis, meaning those longer than 5Km and shorter than 300 km, the distance used as capacity limit for the EV employed by DFDS. In this sense, the country is another relevant criteria in defining the geographical scope. Therefore, a filter based on the specific 6 countries considered of interest is set. Finally, the "Domestic delivery" filter enables the user to choose to show only journeys having the same country as starting and ending point. This is important to avoid distance measures to be biased by the presence of sea freight (for example between France and UK or within the UK between the main island and Shetland).

6.4 Dashboard

Previously, the realized visualizations have helped in having a complete overview of DFDS deliveries from both a volume and a geographical perspective. Indeed, as previously stated these are among the main quantitative criteria that could help in narrowing down the areas of potential candidate points for EV charging stations. Nevertheless, these tools are useful in defining the current status, but would not take into account potential upcoming changes that could affect the company decision for what concerns additional future deployments. As an example, the acquisition of Danish companies in the 3rd Quarter of 2022 has substantially increased booking volumes not only for the single country but the overall level in general. Similarly, frequency flow changes in addition to the introduction of new favourable policies for electric vehicles could open up to deployment opportunities in new areas not previously considered.

For this reason, a dashboard could be an useful monitoring tool that can be updated whenever data can be a relevant support in decision-making processes. Indeed, further explanation on how this could be implemented are present in subsection 8.2.

6.4.1 Design Choices

In order to serve its functions of maintaining an overview on flow, only a small number of relevant element has been selected to be presented in the dashboard in a clean and effective manner. In addition to visualizations, a series of key numbers has been selected as well with

the purpose of providing a high-level summary of the topic at hand.

Given that the current interest for deployment of EV charging station has been narrowed down to a limited number of countries, it seemed reasonable to constrain the countries that visualizations can be filtered with only to the relevant ones in this analysis. However, since the scope of analysis could expand with time it is possible to extend the current list showed in the dashboard filter.

For what concerns features related to the overall aspect, it is important to mention that since the dashboard has been designed specifically for DFDS and with the goal of having it introduced in their own tool-stack, it has been branded with the company logo, present in the top right corner of the dashboard. With the same idea in mind, colours used in visualizations are part of the company's colour palette, when the number of elements in the graph allow for that. Indeed, in some visualizations the required number of color tones did non make it possible to ensure full consistency.

6.4.2 Dashboard description

The dashboard consists of four different graphs and three filters (one of which is a Page filter) as well as a pane with key numbers in the top left corner. While two graphs are related to the volumes of booking over time, a map visually shows how delivery locations are scattered. In the top left section of the dashboard three key numbers are displayed. All of these numbers refer to routes within the ranges considered in scope of this project, which is between 5 and 300 km. However, the filter - applied to all sheets using the same data source - can be tweaked from the sheet of any visualization using it. This easily allows the user to set a new route range, in case improvements to the battery capacity make it possible to slightly increase the route length taken into consideration. The first one shows the average route length, while the central one displays the average number of daily routes in the countries of interest. Lastly, the right number represents the percentage of routes that is within the desired range compared to the total journeys in the dataset.

Moving to the right upper area of the board, it is possible to see the used filters. The first filter on the right is a page filter which is applied to the *Starting and Ending Location Density* map. This allows to visualize the desired month, that can be selected from the drop down menu. Alternatively, it is possible to use the start button and pages will start to flip through

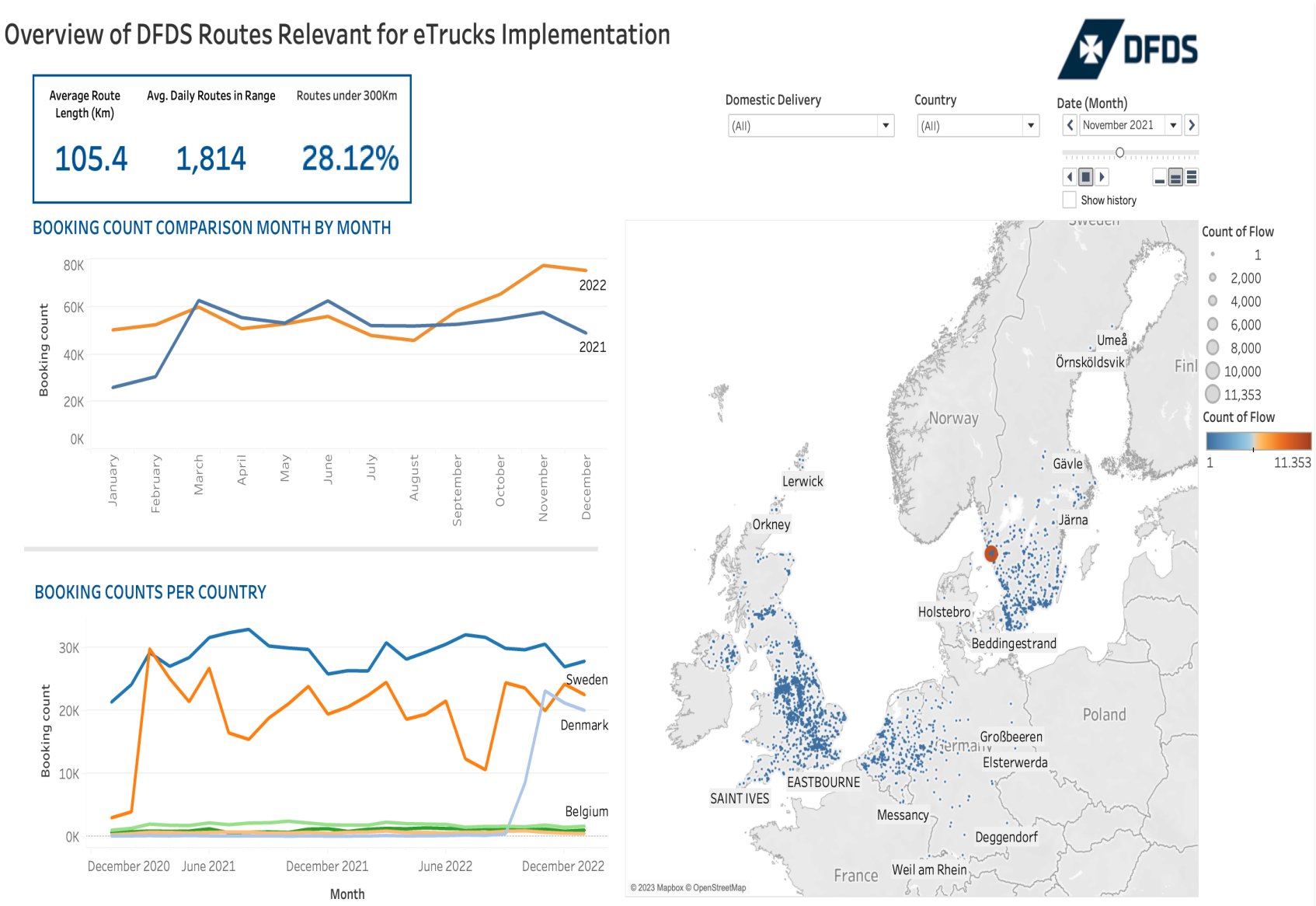

Figure 13: Tableau dashboard providing a general overview of DFDS flow in the eTruck route capacity range [5,300km]

consecutively. On the bottom right corner of the page, three boxes allow to set the speed for visualizing the pages. The central filter, related to the countries is applied to all graphs in the dashboard and enables the user to select which countries they want to visualize data of. This allows for both a higher level or more granular analysis in case it is required. Finally, the filter on the left makes it possible to select whether to visualize only domestic journey or all of them. This can be useful especially when electrical trucks get deployed in contiguous countries.

For what concerns the graphs, on the right side the same map presented in subsection 6.3 is placed. This has proven to be extremely relevant in helping DFDS understanding which areas should be interest by the positioning of charging stations for their eTrucks. On the right side of the map, a legend facilitates obtaining a better understanding. Since a thorough explanation of this visualization has already been presented, no additional detail will be provided in this section. On the left hand side two line graphs are shown, helping in understanding quantitative values related to bookings made. The top one compares the count of booking in each year on a month by month basis, while the bottom chart presents the count of bookings for each country of interest through time. This can be used by project managers to have a first overview of the flows in the countries considered, both in terms of overall volumes and with a breakdown by country, respectively.

This dashboard represents the final product for what concerns the use of Visual Analytics in supporting the goal to find ideal locations for charging stations within DFDS' route network. Indeed, before starting to identify specific locations where to place charging stations, it is necessary to get a deeper understanding of the overall distribution flow, which has been the goal of this chapter.

Here, the line graph for booking volume has helped us understand how the demand changes with time, both due to potential seasonality patterns and external acquisitions. On the other hand, the spike graph and the map have provided a geographical overview of where pick up and drop off points are located as well as their relevance within the network. Finally, the dashboard aims at putting the pieces together to provide a full overview allowing the company to monitor its operations throughout time to facilitate its decision-making process with regards to the development of eTrucks.

In this sense, this chapter also presents the steps performed in order to answer to the first sub-research question, as it will be discussed more in detail in subsection 8.1. From here we'll move to the next section which will be concerned with presenting the next approach adopted which is Graph Theory. These techniques will help in optimizing the placement for charging station within DFDS network by progressing on the insights gathered thanks to this Visual Analytics section.

7 Graph Theory Approach

In the previous sections, we used Visual Analytics techniques and built a dashboard about DFDS' route operations across Europe in order to identify focus areas to concentrate deeper analyses on. In that regard - and also in close collaboration with DFDS' eTrucks project management team - six focus countries have been identified: the United Kingdom, Sweden, Denmark, Germany, Belgium and the Netherlands. The next step now entails looking at these countries' route networks individually in order to find out important routes to deploy eTrucks on and ideal spots to place charging stations.

In order to do these deep dives, we decided to represent each of the country's route network as a graph, as graphs offer numerous advantages for analyzing complex networks like the route network of a logistics company like DFDS. In these graphs, the start and end points of routes (e.g. distribution centers, warehouses, client sites, etc.) represent the nodes in the graphs, while the routes between these points constitute the edges of the graph. Going from there, we can quantify and assess structural properties of the network, identify key nodes and edges that play critical roles in the overall connectivity of the network, and maybe uncover hidden patterns with significant implications for DFDS' operations.

Hence, the next section is concerned with how the graphs for the focus countries were built, not only from a technical perspective, but also which assumptions we made along the process. In subsection 7.2 we then deep dive into the route network of the focus countries and will make tangible recommendations to DFDS as to where to put charging stations and which routes to electrify.

7.1 Graph Building and Underlying Assumptions

Figure 14 depicts the process of how the graphs were constructed for all of the six focus countries and the steps that were taken to be able to make recommendations to DFDS. This section is meant to walk through these steps, explain our reasoning behind them, and elaborate on underlying assumptions we had to make.

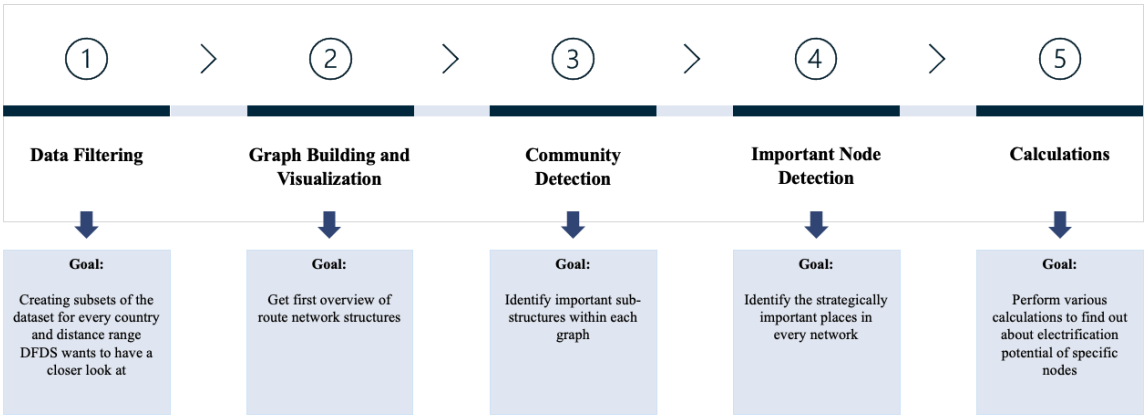

Figure 14: Steps of the Analysis

In a first step, the *df_routes* dataframe that was created earlier, where each row represents a unique route in the network, is filtered into multiple dataframes to create subsets of the data for the respective country and distance range that needs to be examined. For this purpose, the code requires the user to give three parameters to an input function (country, minimum distance and maximum distance), after which the respective filtered dataset is exported automatically. However, one baseline distance range had to be defined to drive the major part of the analysis. In that regard, we agreed with DFDS to work with a distance range [5; 300] for the baseline analysis. While 5 km was set as the lower bound as shorter operations would be covered by smaller vehicles, a couple of reasons played into the decision for the 300 km upper bound. While eTrucks may be able to drive longer distances, it is important to consider that many factors, like the weather condition or the geographical nature of the route can negatively impact the range - for example, the use of A/C in the eTruck during the summer period would reduce the driving capacity. Similarly, the nature of the load itself has to be considered, not only in term of weight, but also in terms of type of products. As an example, DFDS has many supermarket chains among its customers for which it provides

cold-chain services. Here, the use of refrigerated containers also reduces the amount of km that can be covered before the next charge. And lastly, "range anxiety" of the driver - the fear of running out of charge before reaching the destination - has to be considered as well.

However, in order to be able to drive insightful analysis, three other data subsets with the following distance ranges have been exported as well: [5; max], [5; 250] and [5; 500]. We did this mainly for two reasons, 1) for calculation purposes and 2) to be able to compare results in case the eTruck range will decrease (e.g. because of certain weather conditions) or increase (e.g. because of technological advancements).

In a next step, directed graphs were created for all the exported dataframes using the NetworkX library. The coordinates of the pick up and drop off locations of a delivery were used as nodes in the graph, and the respective routes between these nodes constitute the edges of these graphs. The term 'directed graph' refers to the fact that between every node, there's two respective edges: one going from node A to node B, the other one from node B to node A (subsection 3.3). The number how often each unique route has been taken ('RouteCount' variable in *df_routes*) serves as the weight of the edges. Also, for every node the sum of the weights of the attached edges was calculated, in order to get information about the traffic happening at each node.

In a next step, the graphs have been visualized using the folium library - a versatile Python library that allows to visualize data structures on top of base maps (Rob Story, 2013). This comes in especially useful in our case, as it allows us to visualize the nodes at their exact graphic location defined by their coordinates. By doing so, we were able to visualize the complex graph structures with thousands of nodes and edges in a comprehensible way, which we further facilitated by color coding the edges using a cool-warm color scheme based on their weights - more frequently used routes are marked orange-red ("warm"), while less frequently used routes are in blue shades ("cool"). It is worth mentioning at this point that we made the decision to depict the routes as a straight line between nodes, and not by visualizing the actual route the truck would take on highways, etc. There are two reasons for that: first, visualizing the actual route would have come with a loss of information, as many trucks would e.g. take the same highways, which would mask the information as to which node-to-node connections are especially important. And second, we are working under the assumption that charging can only happen at nodes anyways (see *assumption 1* below), so for the business case

at hand the information about which highways are especially being used in a country just is not as important. But to also give the reader an idea how these finalized graphs visualized via the folium library looked like, here's an example:

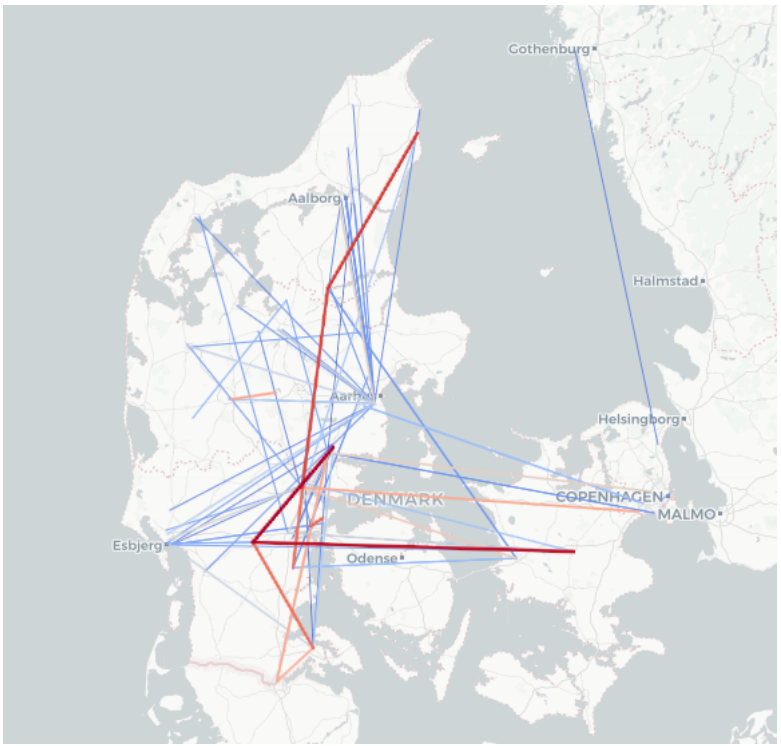

Figure 15: Example for a Graph Visualization using folium library

Having set up these graphs and corresponding maps allowed us to then perform various calculations and examine the route network's properties in order to find answers to our research questions and visualize our results. In order to do so, the next important step was to partition each graph into multiple communities, and then find the most important nodes in the graph. It is essential to mention at this point that the following analyses were also conducted under the assumption that charging can only happen at nodes of the graph, not on edges. This decision is relevant because choosing the locations has an impact on the further methodology. In this case, the choice was made based on the dialogue with the DFDS project managers for various reasons. In the context of this project, nodes represent the starting and ending point of a delivery (or at least of one of its legs) and therefore they are locations owned by either DFDS or one of its customers. In the first case, there would be no impediments for DFDS to install charging station in their own sites. At the same time, since transportation

with eTrucks is the result of an agreement between the parties (as customers agree to having their good delivered for a higher price with eTrucks), it would be possible to insert the deployment of charging station within the agreement. On the contrary, edges (representing the routes travelled between the two points) could be either publicly or privately owned with large differences within countries, limiting in this sense the possibility for DFDS to install its charging stations alongside routes.

- **Assumption 1**: Charging can only happen at nodes, not edges.

The next step was concerned with community detection. For this task, a couple of popular algorithms are available, which usually can be differentiated by the type of graph they are used for - either directed graphs, where two edges exist between two nodes (one for each direction with respective weights) or undirected graphs, which only has one connection between two nodes that sums together the weights of both directions.

Given its popularity and the general high performance level, we used the Louvain algorithm for community detection (subsection 3.4). However, this algorithm only works with undirected graphs - which required some handling, as in our case we are working with directed graphs, given that the direction of journeys is represented in the data. However, the fact that charging can only happen at nodes takes away the importance of knowing about the directions of the edges - simply put, a charging station doesn't care about where the trucks using it are coming from and going to. Therefore, we converted the directed graphs in their undirected version, where the new edge attributes (and weights) "are a combination of the attributes of the directed edges" (Networkx, 2023). This assumption is widely used in the context of transportation networks (as exemplified by Guerra et al. (2022)), and especially feasible in our case. An exemplary community visualization can be found in Figure 16.

the next step was concerned with how to pick the most important nodes in every graph. Essentially two options were available: deciding on the most important nodes based on 1) the node traffic, or 2) on different centrality measures. Given the fact that the traffic at nodes is heavily skewed (few nodes with lots of traffic, many with very little traffic, as presented in subsection 7.2), we decided to go with option 1) and decide on the most important nodes based on node traffic, as making the decision purely based on centrality measures would have yielded the risk to ignore nodes that are highly frequented. Based on the extensive reasoning

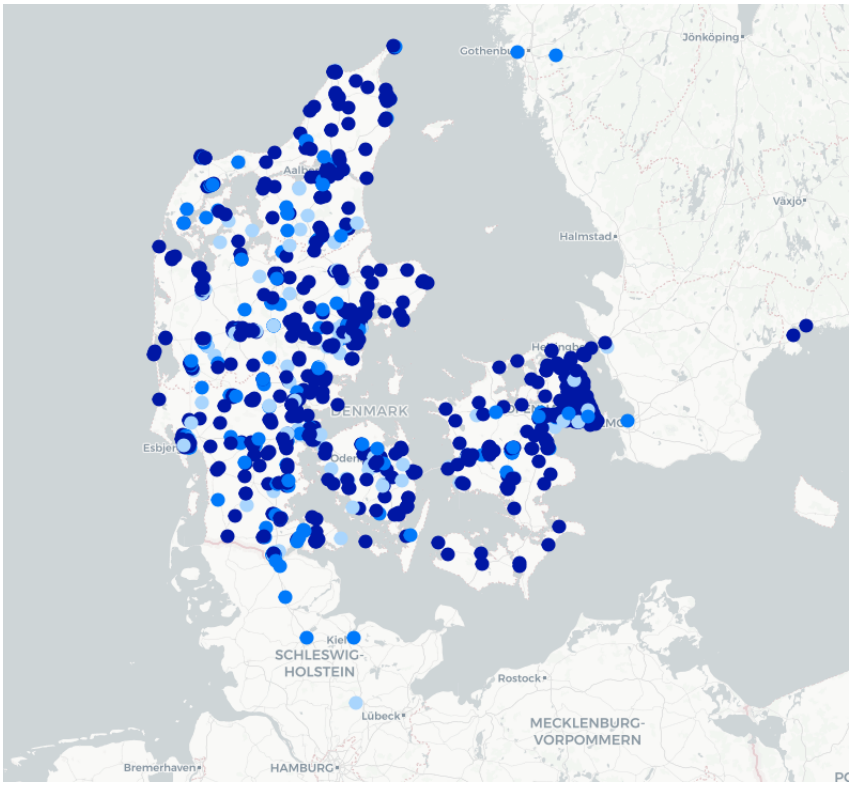

Figure 16: Visualization of Communities

stated above, the next two assumptions for our analyses read as follows:

- **Assumption 2**: The importance of a node is measured by its traffic.
- **Assumption 3**: To identify key nodes based on total traffic, we can transform directed graphs into undirected ones, as they offer a suitable approximation for detecting overall interaction patterns, which hold more significance than edge directionality.

From there, the next question was whether to select the most important nodes from the graph as a whole, or the most important nodes from the largest communities. Given the fact that all graphs we created showed very strong and clearly defined community structures (subsection 7.2), we decided to pick the five most important nodes of the three largest communities for each graph. We are aware that we might ignore important nodes with this methodology (e.g., the 6th most important node of the largest community might be more important than the most important node of the third largest community), but on the other side it was important to us to have the different communities represented in the result as their strong structure can have various implications in a transportation network: e.g. cold chain vs. warm chain

transportation, or the representation of an important client/group of clients, such as in the case that DFDS delivers to multiple warehouses of the same supermarket chain.

- **Assumption 4**: The most important nodes within a network can be identified at the community level.

Based on these four assumptions, we were then able to proceed to calculate the most important nodes in each graph and make recommendations as to where DFDS should put charging stations based on the electrification potential of these nodes. In order to calculate this electrification potential, one last assumption was made:

- **Assumption 5**: There are no charging capacity constraints at a node.

We essentially assume that if DFDS installs charging stations at a node, all attached routes - incoming and outgoing - could potentially be electrified as the eTrucks covering these routes could charge there before departure. In the real world scenario, it would depend on a number of factors - like the number of installed charging stations, the number of eTrucks deployed at this node, the charging duration, just to name a few - whether this in fact could be possible, and coordination with DFDS' route planning department would be needed, as it will be presented in section 9.

As a last step, we then compared how "bulletproof" our results for the [5; 300] distance range were by doing the same calculations and analyses for the [5; 250] range and the [5; 500] range, respectively, which gives DFDS tangible insights into the scalability of the eTruck project.

In conclusion, we deemed this fairly extensive theoretical overview, including the introduction of our five assumptions, necessary in order for the reader to be able to follow the analyses in the following section.

7.2 Analyses

7.2.1 United Kingdom

The first country we are taking a deeper look at is the United Kingdom (UK), as it is the country with the most overall DFDS operations. In the dataset we were working with - which essentially contains all DFDS logistic operations in Europe - 71.8% of all routes are associated with the UK, meaning they either have their start or end point (or both, in the case of domestic deliveries) in the UK. Even more, 78.9% of all the deliveries are associated with the UK (where routes correspond to edges of the graph and deliveries to weights of the edges of the graph). Hence, the UK constitutes a key market for DFDS and choosing the right routes to deploy eTrucks and the right spots to deploy charging stations will be absolutely crucial.

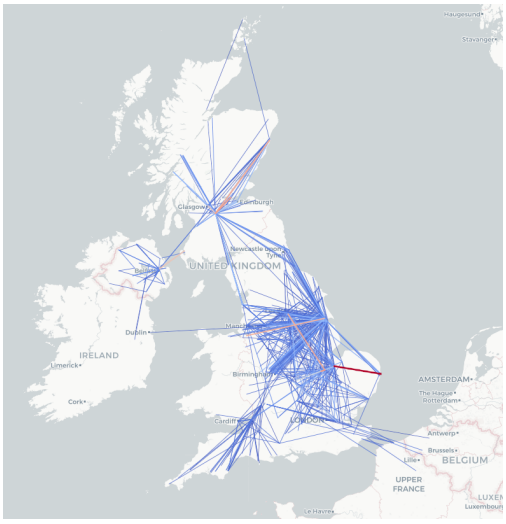

Figure 17: Top 2.5% routes in Uk

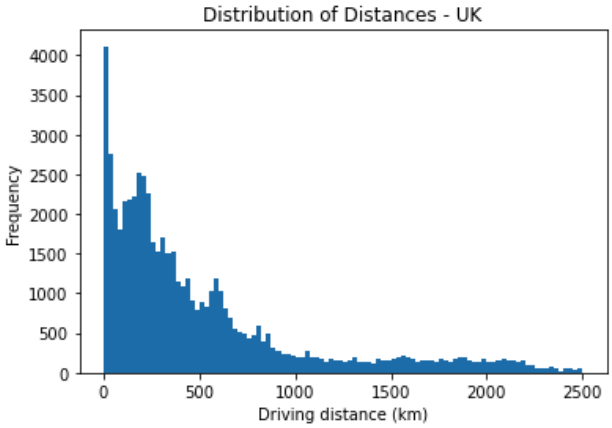

Figure 18: Route frequency distribution in Uk

First of all, we filtered for the 2.5% most important routes within the [5; 300] range to get a first overview (Figure 17), which in the UK still leaves us with multiple routes that are frequently used by DFDS. Nonetheless, we can identify a few important hubs: By far the most routes are connected to the harbour in Immingham, where DFDS has a large hub including terminals, warehouses, and more (DFDS, 2023c). Going from there, especially three main route directions can be identified: towards the North (to Newcastle and onwards to Glasgow), West (Manchester and Liverpool region) and Southwest (Birmingham). Surprisingly, however,

the route which is used by far the most often (which is indicated by the dark red line) runs from the Harbour in Lowestoft in the East to a DFDS warehouse in Wisbech. Further examination of this node revealed though that this warehouse (and hence the associated routes) is especially used for frozen goods, which makes the route not very suitable for eTrucks as cold chain goods would reduce the range the eTruck can travel. Further important hubs can be found in Cardiff, Belfast and Glasgow, from which many routes are connected to places in Wales, Northern Ireland, and Scotland, respectively.

Regarding the route distances within the UK network, we can see quite the typical, steadily declining distribution with many short range routes (<500km). The largest part of the routes is shorter than the critical range of 300km, which suggests that, hypothetically, large portions of the network could be electrified as they fall within the range that can be covered by eTrucks.

Measure	**Value**
Number of Nodes	12,459
Number of Edges	26,900
Avg. Degree	4.318
Avg. Clustering Coefficient	0.067
Number of Communitieses	446
Modularity of Community Structure	0.741

Table 2: Graph Characteristics for UK (Distance Range: 5-300km)

After this first high-level overview, the next step now was to build the graph for the UK network, following the approach laid out in the previous section, and analyse the network based on the metrics defined in subsection 3.5. As Table 2 shows, the UK graph has 12,459 nodes and 26,900 edges, making it a large and highly complex graph. This is also shown by the average degree of 4.318 (the average number of edges a node has). Given the nature of transportation networks and also analyzing the UKs network visually, this suggests that there are few nodes with many edges (e.g. distribution centers, warehouses) and many nodes with very few or even just one edge (e.g. a client site that is being delivered to by only one DFDS center). The average clustering coefficient is very low (0.067), which implies that the graph is not tightly clustered, and the nodes within the graph don't tend to form tightly connected groups. In other words: nodes are found all over the country and are not necessarily clustered together in geographical groups, and nodes that are geographically close to each other aren't

necessarily also connected in the graph. This seems counter-intuitive with the high modularity score of 0.741 at first, which indicates that the communities of the graph - 446 of which have been detected by the Louvain algorithm - are very well defined, as modularity scores can range from -1 to 1, with values closer to 1 indicating strongly connected community structures. However, low average clustering coefficients and a high modularity score can indeed co-exist: this scenario is often the case in graphs with many (smaller) communities, whose nodes, however, are strongly connected despite not being geographically close to each other (Newman, 2006). This is a typical scenario in logistics networks, as the connection between nodes often times isn't defined by pure proximity, but factors independent of that - such as the type of goods that are being transported or the customers that are being served. (subsection 3.5)

Community	**Size (Nodes)**	**Size (Conductance Score)**
Community #1	1,897	0.329
Community #2	1,521	0.490
Community #3	1,474	0.604

Table 3: Largest Communities in the UK Graph (Distance Range: 5-300km)

In line with the approach we laid out earlier, we then took a look at the three largest communities (out of 446). They combine 39% of the nodes of the network (Table 3) - which might be a reason for the mediocre at best conductance scores (0.329, 0.490, 0.604 respectively), which suggest that these communities aren't perfectly separated from the rest of the graph - however, given their size of more than 1,400 nodes each, that was to be expected. Overall though, the high modularity score validates our conceptual choice to then move on to select the five most important nodes - based on node traffic (representing the weights of all the attached edges summed together) in every community and play with scenarios about deploying charging stations at these nodes.

The geographical locations of these identified nodes can be seen in Figure 19. Seven of them are located in harbour areas (four in the Immingham area, three in Felixstowe/Harwich), while the rest is especially scattered around Liverpool, Birmingham and Leeds. What are the implications of putting up charging stations at these nodes? To recap, the assumption was

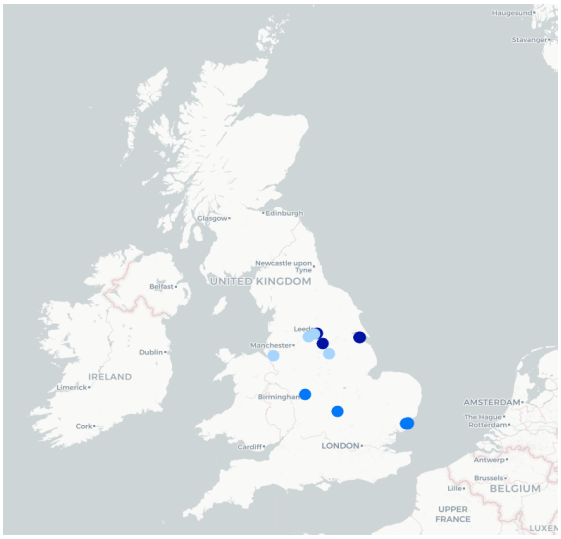

Figure 19: Top nodes in UK for the range [5,300]

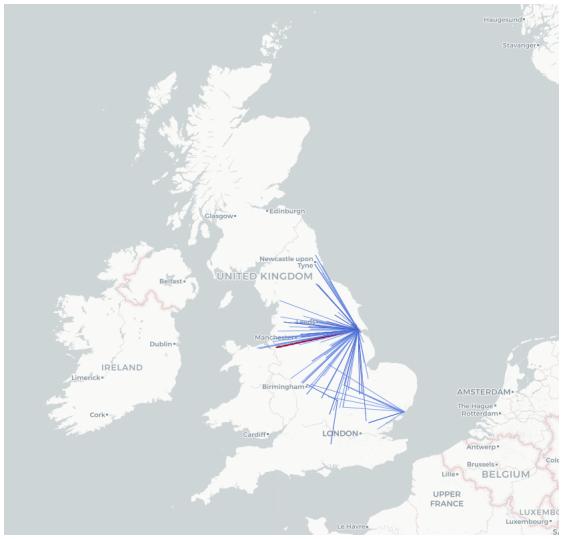

Figure 20: Route overlap in UK

made that if there are charging stations at a node, all routes connected to this node (that are within the 300 km distance range) can be covered by eTrucks and hence be electrified.

Table 4 shows the locations of the 15 most important nodes in the UK graph and their respective influence on a possible electrification for the network, while Figure 20 shows the overlap between the top 2.5% most important routes and the routes that could be electrified if charging stations were placed at all of the 15 most important nodes. Only 21.42% of the most important routes could be electrified by charging stations at these nodes, and especially routes departing from and going to the harbours in Immingham and Harwich would be covered by this. Looking more closely at the most important node at the port in Immingham, setting up charging stations there (but nowhere else) could lead to an electrification of 2,151 routes (as this node has 2,151 edges) and 43,900 deliveries (equalling the sum of these edges), which translates to 2.34% of the routes and 1.91% of the deliveries in the complete route network in the UK without any distance threshold. Within the subset of the 300 km distance threshold, installing charging stations at this node could electrify 8.00% of the routes and 6.19% of the deliveries, as is shown in Table 4. The way this table can be read further is that every row gives information about the share of the network that could be electrified if charging stations would be set up at the respective node *additionally* to the node(s) before. For example, looking at the 9th most important node in the graph in Wakefield tells us that if all of the nine most important nodes would be equipped with charging stations, a total of 5,910 routes

Node	City	Electrified Routes	% Electrification (Routes) Total	% Electrification (Routes) Subset	Electrified Deliveries	% Electrification (Deliveries) Total	% Electrification (Deliveries) Subset
(53.62811, -0.18739)	Immingham (port)	2151	2.34%	8.00%	43900	1.91%	6.19%
(53.62560417, -0.20199861)	Immingham (port)*	3870	4.20%	14.39%	63874	2.78%	9.01%
(53.2761, -2.87574)	Ellesmere Port	3870	4.20%	14.39%	63874	2.78%	9.01%
(53.63414, -0.19931)	Immingham (port)	4128	4.48%	14.39%	71762	2.78%	10.12%
(51.95229, 1.32532)	Felixstowe (port)	5198	5.65%	19.32%	79176	3.12%	11.16%
(53.62719, -0.18007)	Immingham (port)	5374	5.84%	19.98%	85644	3.44%	12.07%
(51.9474, 1.25301)	Harwich (port)	5894	6.40%	21.91%	91193	3.96%	12.86%
(53.50873, -1.33107)	Wath upon Dearne	5899	6.41%	21.93%	91340	3.97%	12.88%
(53.71286, -1.51991)	Wakefield	5910	6.42%	21.97%	91411	4.09%	12.88%
(52.18454, -0.8867)	Northampton	5971	6.49%	22.20%	94122	4.17%	13.27%
(51.94752, 1.32417)	Felixstowe (port)	6425	6.98%	23.88%	96050	4.18%	13.54%
(52.52019, -1.88804)	Birmingham	6427	6.98%	23.89%	96065	4.18%	13.54%
(53.64801, -1.77422)	Huddersfield	6427	6.98%	23.89%	96065	4.18%	13.54%
(53.69978, -1.60723)	Dewsbury	6427	6.98%	23.89%	96065	4.18%	13.54%
(53.31605, -1.13454)	Gateford	6427	6.98%	23.89%	96065	4.18%	13.54%

Table 4: Impact of most important nodes in the UK (Distance Range: 5-300km)

within the network can be electrified, which translates to 6.42% of the routes in the whole network and 21.97% of routes in the subset.

Essentially, this table shows that after electrifying the seven most important nodes, there would be little to no additional benefit for DFDS if charging stations were to be placed at the other eight identified nodes as well, as they would only add very few additional routes and deliveries that could be electrified, as the last three wouldn't add any new electrification potential at all. The reason for this is that these are nodes with a very small amount of edges, and most of their edges (all of them in the case of the last three nodes) are already connected to one ore more of the more important nodes, which is why electrifying these nodes wouldn't yield to any additional electrification benefit as these routes could already be covered by charging stations at other nodes. For example, these nodes could be client sites which are receiving deliveries frequently by the same DFDS distribution center, which is why they only have one edge.

Overall, the electrification potential in the route network of the UK remains quite low.

Even if putting up charging stations at eight strategically important nodes, only 6.42% of routes and 4.09% of deliveries can be electrified. The reason for this is the size and complexity of the UK's route network, which has a large number of less frequently used nodes and edges instead of very few key nodes and edges, which limits the electrification potential even when setting up charging stations at important nodes. We will see later that indeed this is different in smaller countries with less complex networks.

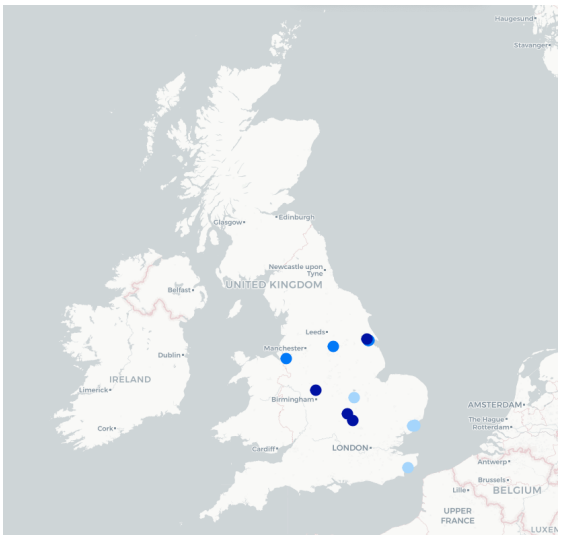

Figure 21: Top nodes in UK for the range [5,250]

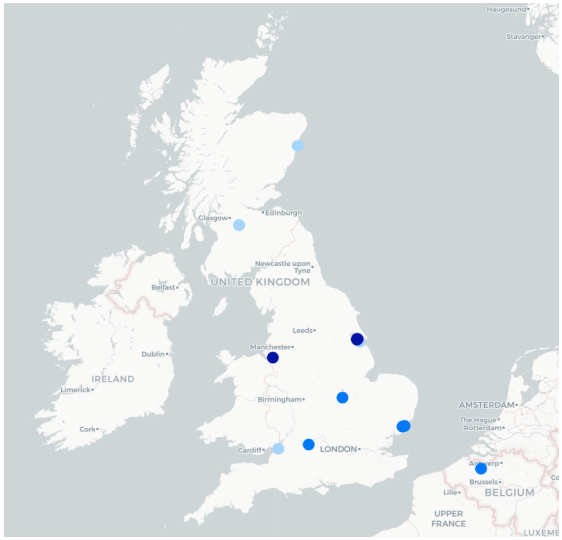

Figure 22: Top nodes in UK for the range [5,500]

Lastly, the scalability of these results need to be discussed. What happens if for some reasons, such as very cold temperatures in winter, the range of eTrucks decreases? Or, on the contrary, what if technological advancements allow eTrucks to drive more than 300 km without charging five years from now? These questions are extremely relevant to DFDS when thinking about scalability of our results and the eTruck project in general, which is why the last step of the country analysis deals with the question how "bulletproof" the results are if the eTruck distance range will change. In order to do so, we extracted two more subsets from the original dataset - one with a decreased [5; 250] range, and one with an increased [5; 500] range - and then calculated the most important nodes in the newly generated graphs again. In in ideal case, the most important nodes would not differ between the three graphs with the different ranges.

It turns out that ten out of the 15 detected most important nodes belong to the most

important nodes of the [5; 250] graph as well; for the [5; 500] graph, there is still an overlap of six nodes Figure 22. The seven most important nodes are the same in all three graphs, with the exception of one node at the port in Immingham, which isn't among the most important nodes for the [5; 500] graph. This indeed shows a solid scalability for the eTrucks project in the UK, as DFDS could set up charging stations at the seven most important detected nodes and be sure that these nodes would keep their importance even if the eTruck range would decrease or increase for various reasons.

Recommendation: Based on the previous discussion of results, we would advise DFDS to especially look into the seven most important identified nodes to install charging stations. Given that four of those are located in the port of Immingham, looking into these spots should be a priority. While the overall electrification potential remains quite low for the UK, this should not change this recommendation as the overall importance of the UK within the DFDS network suggests that still a tremendous amount of routes and deliveries could be electrified.

7.2.2 Sweden

The next country we are looking at is Sweden, which ranks 2nd when it comes to the share of the overall DFDS truck operations. 43.43% of all the routes and 61.8% of all deliveries are associated with Sweden. This gap between routes and deliveries is an interesting thing to note, as it suggests that in Sweden a relatively high number of deliveries is carried out on a relatively low number of routes, which should yield in a potentially high impact of charging stations at the right nodes as there are many highly frequented routes.

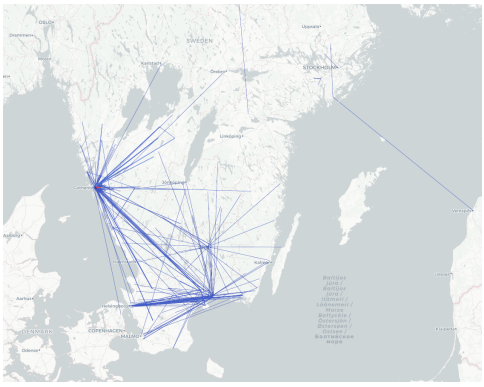

Figure 23: Top 2.5% routes in Sweden

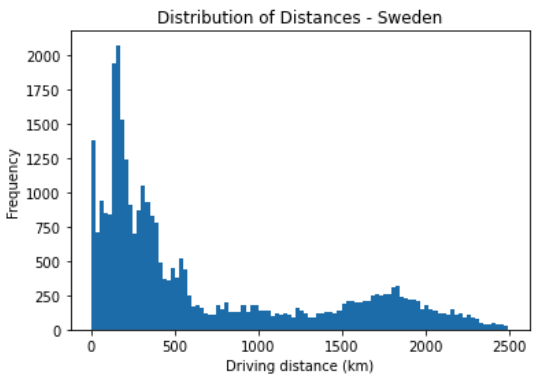

Figure 24: Route frequency distribution in Sweden

Filtering for the most important 2.5% of routes based on the number of deliveries (Figure 23 reveals that two especially important hubs exist in Sweden: not very surprisingly, one in the Gothenburg area, where DFDS has multiple large sites around the harbour, and then a second one in the southern part of the country in Karlshamm. From there, the most important routes in Sweden's network cover the majority of Southwestern Sweden, including e.g. Jönköping, Växjö, Malmö and Helsingborg. We can also see that the most frequently taken routes are very short routes (< 50km) within the Gothenburg area – this would have been even more visible if we would have included route distances < 5km in the analysis, as our initial data revealed that many extremely short distance operations are executed in this area around the harbour. Regarding the route distances in Sweden's network, we can see that the majority of routes indeed has a distance in the eTruck relevant range of maximum 300 km, which was not necessarily to be expected due to the large size of the country that facilitates long routes. Also, it is interesting to see that many routes with distances > 1,200km exist as well – these usually are the routes departing from/arriving at Gothenburg or other ports and that were executed by sea freight. However, the API we used calculated the distances using "driving" as the travel mode, which results in these long distances although it in reality it maybe was only a shipment over a few hundred kilometers, e.g. between Gothenburg and Frederikshavn in northern Denmark. In any case, these routes are not relevant for the eTruck project, which is why we do not need to take them into consideration.

Measure	**Value**
Number of Nodes	10,703
Number of Edges	13,636
Avg. Degree	2.548
Avg. Clustering Coefficient	0.045
Number of Communitieses	424
Modularity of Community Structure	0.745

Table 5: Graph Characteristics for Sweden (Distance Range: 5-300km)

Upon constructing the graph for Sweden's route network, as depicted in Table 5, we observe a total of 10,703 nodes and 13,636 edges. Just as for the UK, this results in a sizable and intricate graph. The average degree of 2.548 further highlights the complexity, indicating the average number of edges each node possesses. The number is significantly lower

compared to the average degree of the UK (4.318); one reason for that could be that Sweden is significantly larger and less densely cluttered with nodes, which leads to more transportation routes existing in the UK (26,99 edges in the UK 13,600 in Sweden). However, the values for the average clustering coefficient (0.045), the number of communities detected (424) and the modularity of the community structure (0.745) are very similar to the UK values, which indicates that in general we are dealing with a very similar graph in terms of its properties (a.o. nodes not showing any sign of clustering and the community structure being fairly well defined). Yet, there's a striking difference between the conductance scores for the largest communities in the graphs for Sweden and the UK. While in the UK the values have been between 0.3 and 0.6, the conductance scores for the three largest communities in Sweden are 0.061, 0.097 and 0.131, respectively, which suggests that the communities in Sweden are very tightly connected internally and have fewer connections with nodes outside the community – again, this is most likely reasoned in the geographical constitution of the country. Overall though, the graph shows strong community structures, which again justifies the next step of selecting the five most important nodes of the three biggest communities for further analysis.

Community	**Size (Nodes)**	**Size (Conductance Score)**
Community #1	4,233	0.061
Community #2	685	0.097
Community #3	609	0.131

Table 6: Largest Communities in the Sweden Graph (Distance Range: 5-300km)

The geographical locations of these identified nodes can be seen in Figure 25. Other than two nodes in Jönköping and Värnamo, all of the important nodes are located close to ports – e.g. six of them in Karlshamm, two in Kalmar, and two in Gothenburg/Kungsbacka.

Accordingly, Figure 26 - illustrating the intersection between the top 2.5% of the most important routes and the routes that could be electrified if charging stations were installed at each of these 15 key nodes – shows that DFDS should especially look into routes departing from/going to Karlshamm. This is somewhat surprising, as the port in Gothenburg is one of DFDS' biggest hub in general – when it comes to route electrification, however, a focus on other hubs seems to be suggested by the data, which is reasoned by the very short routes in the Gothenburg area.

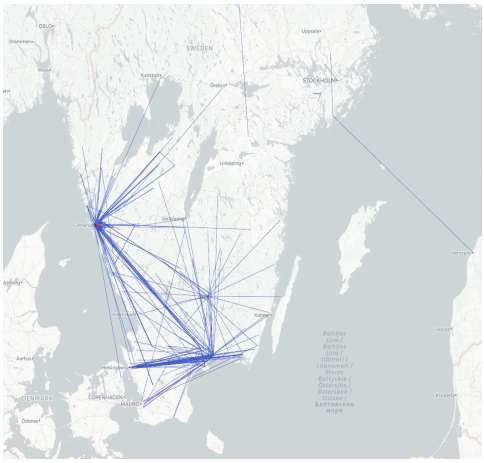

Figure 25: Top nodes in Sweden for the range [5,300]

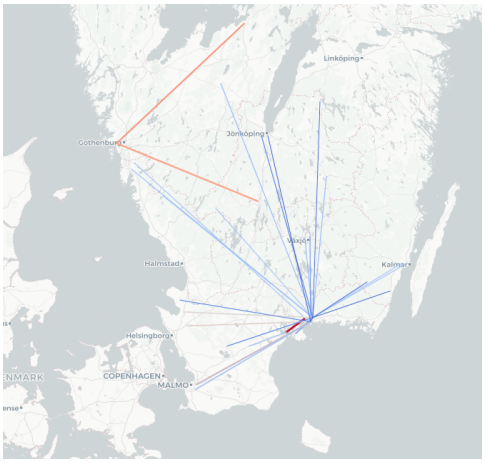

Figure 26: Route overlap in Sweden

Analogous to the UK methodology, Table 7 displays the positions of the 15 most crucial nodes in Sweden and their respective impact on potential network electrification. It is visible on a first glance that just by electrifying the two most important nodes, both located in Karlshamm, already the largest part of the electrification potential Sweden has to offer could be achieved. This is reasonable when looking back at the map showing the routes that could be electrified by placing charging stations at the top 15 nodes (Figure 25) – almost all these routes depart from or go to Karlshamm, with the by far most important route being between Karlshamm and Bromölla (3rd most important node). Electrifying the two most important nodes already impacts 13.14% of all the routes in Sweden and even 41.26% of the 300 km subset – quite impressive, considering the size of the graph. However, only a very small share of the deliveries is affected by this – 1.54% for the total graph, 2.57% of the subset. This seems way off in the beginning, but has a very good reason: the initial dataset (which is used for these calculations) contained thousands of "deliveries" that happened within the port of Gothenburg. These operations are important to have in the system for DFDS and were included in these calculations, but since the graph and corresponding maps are built on a [5;300] subset, they are not included here. However, these extremely short operations, especially within ports, are not covered by eTrucks anyways, which is why we do not need to be concerned with them. Overall, given that DFDS would already be able to electrify 12.14%/41.39% by setting up charging stations at only two nodes, Sweden offers high electrification potential in this regard.

Node	City	Electrified Routes	% Electrification (Routes) Total	% Electrification (Routes) Subset	Electrified Deliveries	% Electrification (Deliveries) Total	% Electrification (Deliveries) Subset
(56.20318, 14.8734)	Karlshamn*	4787	11.18%	35.11%	9655	1.16%	1.94%
(56.16279, 14.84092)	Karlshamn (port)*	5626	13.14%	41.26%	12834	1.54%	2.57%
(56.07341, 14.46605)	Bromölla (port)	5644	13.18%	41.39%	14053	1.69%	2.82%
(56.19039, 14.74404)	Mörrum	5662	13.22%	41.52%	14159	1.70%	2.84%
(57.69723, 11.85502)	Göteborg (port)	5667	13.23%	41.56%	14770	1.78%	2.96%
(56.1628, 14.81778)	Karlshamn (port)	5766	13.47%	42.29%	15353	1.85%	3.08%
(57.2099, 14.03045)	Värnamo	5778	13.49%	42.37%	15609	1.88%	3.13%
(55.61147, 13.08601)	Arlöv	5778	13.49%	42.37%	15609	1.88%	3.13%
(56.06076, 14.6093)	Sölvesborg	5779	13.50%	42.38%	15614	1.88%	3.13%
(56.35238, 12.83254)	Förslöv	5779	13.50%	42.38%	15614	1.88%	3.13%
(57.76524, 14.08892)	Jönköping	5779	13.50%	42.38%	15614	1.88%	3.13%
(56.6759, 16.25046)	Kalmar	5779	13.50%	42.38%	15614	1.88%	3.13%
(56.67573, 16.32088)	Kalmar	5779	13.50%	42.38%	15614	1.88%	3.13%
(56.19914, 15.63073)	Karlskrona	5779	13.50%	42.38%	15614	1.88%	3.13%
(57.4793, 12.08572)	Kungsbacka	5779	13.50%	42.38%	15614	1.88%	3.13%

Table 7: Impact of most important nodes in Sweden (Distance Range: 5-300km)

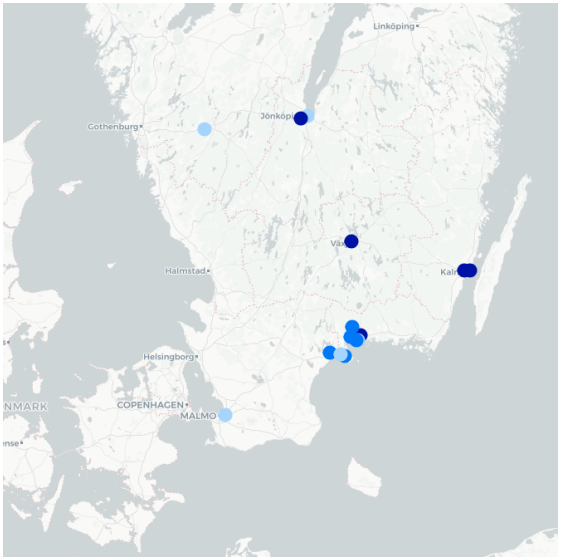

Figure 27: Top nodes in Sweden for the range [5,250]

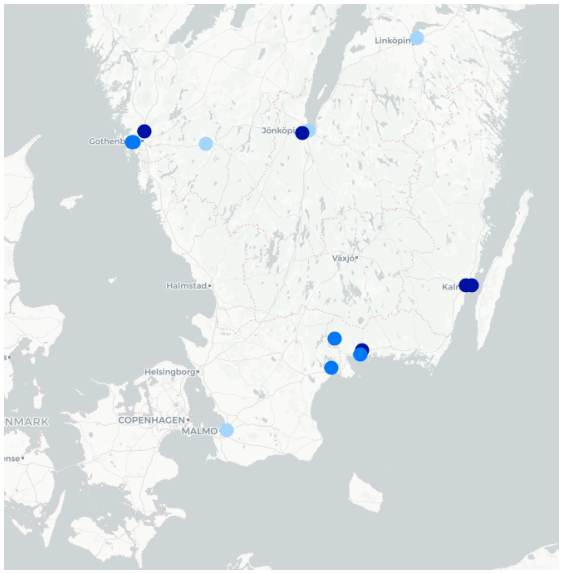

Figure 28: Top nodes in Sweden for the range [5,500]

Lastly, looking at the important nodes in the graphs in case the ranges change to [5; 250] and [5; 500] respectively, we can see that only six out of the 15 detected nodes are also relevant in the [5; 250] graph and eight in the [5; 500] graph, with only little overlap existing within the top seven or eight nodes, which is most likely caused by the highly complex network structure

in Sweden (Table 5). However, this also suggests that the results are not too scalable for Sweden, suggesting the conclusion that DFDS should be very certain about the [5; 300] range if they plan to set up charging stations at the identified important nodes.

Recommendation: For Sweden, the recommendation would be to install EV charging stations at the two most important nodes in Karlshamm, as they already cover large parts of the graph and installing charging stations at additional nodes did not yield to considerable additional electrification potential. However, it should not be ignored that Gothenburg is one of the most important overall hubs of the route network - so, DFDS should assess the feasibility of eTruck deployment on very short routes, as many operations happen within the port in Gothenburg, and eventually install charging stations in that area as well.

7.2.3 Denmark

The next country to analyze is Denmark, DFDS' "home turf". Denmark accounts for 6.55% of the routes in DFDS' network across Europe, but with 21.39% for more a significantly higher portion of all the deliveries – which makes it a very interesting country to look at, as also the the fact that Denmark is relatively small promises high electrification potential if the right locations for charging stations are being chosen.

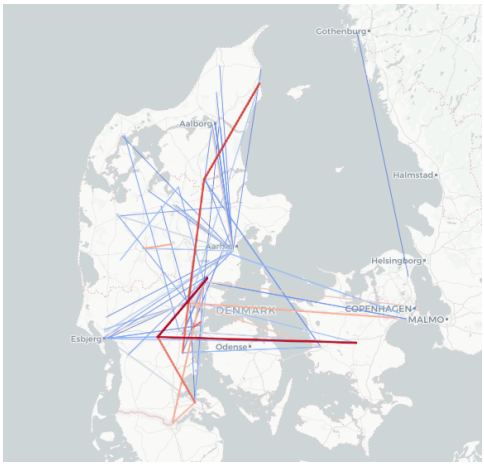

Figure 29: Top 2.5% routes in Denmark

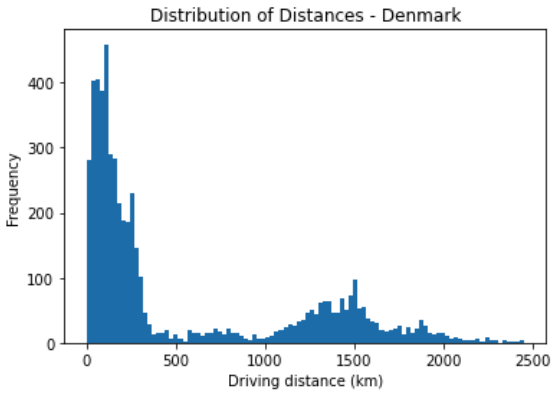

Figure 30: Route frequency distribution in Denmark

Looking at the most important routes in Denmark (Figure 29), two main directions of travel can be identified: One between the North of Jutland (Frederikshaavn/Saeby) all the

way down to the border of Germany, and the second between central Jutland (e.g. Vejle, Horsens, Vejen) and Sealand in the Eastern part of the country (especially to Ringsted and Copenhagen). Compared to the networks of Sweden or the UK, for Denmark there seem to be more clearly defined important routes – the maps displaying the most frequently used routes is less cluttered and there are multiple dark red routes, which might be caused by the network being significantly smaller overall. Another indicator for this can be seen in Figure 30, displaying the route distances within Denmark's route network: They are either below 300 km or longer than 1,000 km, which can clearly be separated by routes being taken by truck on land and routes being taken by ships to other countries.

Measure	**Value**
Number of Nodes	1,592
Number of Edges	3,508
Avg. Degree	4.407
Avg. Clustering Coefficient	0.181
Number of Communities	82
Modularity of Community Structure	0.571

Table 8: Graph Characteristics for Denmark (Distance Range: 5-300km)

After setting up the graph for Denmark's route network, we identify 1,592 nodes, 3,508 edges, and 4.407 edges per node. We hence deal with a much smaller graph than the ones before, which however comes with a similar level of complexity as the UK graph, given the relatively high number of edges per node. The average clustering coefficient is fairly low as well with 0.181, however, some moderate clustering of nodes is detected. The Louvain algorithm has identified 82 distinct communities within the graph, and the modularity score of 0.571 once again demonstrates a fairly well separated community structure.

Community	**Size (Nodes)**	**Size (Conductance Score)**
Community #1	683	0.480
Community #2	200	0.880
Community #3	196	0.913

Table 9: Largest Communities in the Denmark Graph (Distance Range: 5-300km)

The three biggest communities (Table 9) include 67.78% of all the nodes, most of them belonging to the most important community. This structure most likely is the reason for the very high conductance scores of communities #2 and #3, as high conductance scores for smaller communities can often times be observed when there's one large, dominating community.

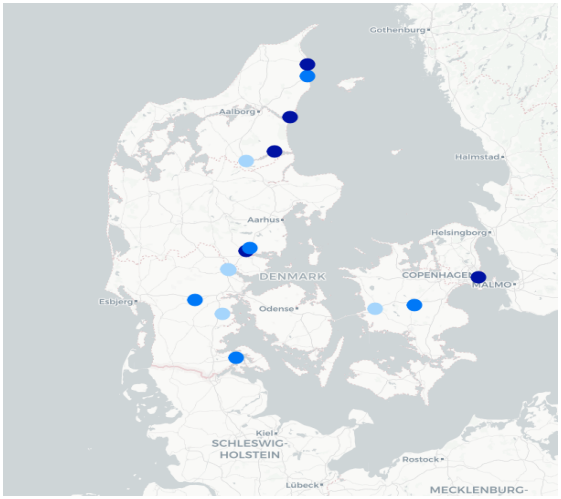

Figure 31: Top nodes in Denmark for the range [5,300]

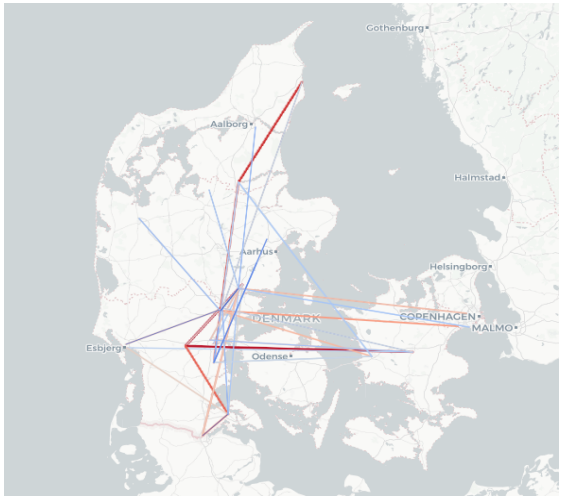

Figure 32: Route overlap in Denmark

Looking at the 15 most important nodes in Figure 31, we can see that they clearly are located along the two main routes identified earlier, northern Jutland – southern Jutland and central Jutland – Sealand. More specifically, three hubs can be identified: three of those locations are on Sealand, seven in central and southern Jutland, and the remaining five in northern Jutland. Just as expected earlier, by electrifying all of these nodes, large parts of the most important routes – namely, 67,81% - could be electrified (Figure 32). This large difference to Sweden or the UK, where electrifying the most important nodes yielded significantly less potential, can clearly be attributed to the smaller size of the Denmark graph and the existence of more distinct top routes.

Accordingly, looking at Table 10, we can see that the electrification potential in Denmark is relatively high. By only electrifying the most important node, a DFDS site located in Horsens, already 1,323 routes and 20,386 deliveries – corresponding to 17.49% of the routes and 37.71% of the deliveries of the total graph and 17.72% / 24.6% for the subset, respectively – could be electrified. Electrifying the eleven most important nodes (as there is little to now further

Node	City	Electrified Routes	% Electrification (Routes) Total	% Electrification (Routes) Subset	Electrified Deliveries	% Electrification (Deliveries) Total	% Electrification (Deliveries) Subset
(55.88724, 9.7797)	Horsens	1323	17.49%	37.71%	20386	17.72%	24.60%
(55.35537, 9.49505)	Christiansfeld	1454	19.23%	41.45%	27574	23.97%	33.28%
(55.47409, 9.15797)	Vejen	1501	19.85%	42.79%	33936	29.50%	40.95%
(55.91521, 9.82681)	Gedved	1583	20.93%	45.13%	37318	32.44%	45.04%
(54.98001, 9.65443)	Sønderborg	1638	21.66%	46.69%	40376	35.09%	48.73%
(55.42908, 11.79541)	Ringsted	1674	22.13%	47.72%	42582	37.01%	51.39%
(55.72787, 9.57226)	Vejle	1709	13.49%	22.60%	45900	39.90%	55.39%
(56.63791, 9.77947)	Hobro	1753	13.49%	23.18%	47123	40.96%	56.87%
(57.33164, 10.51244)	Saæby	1786	23.61%	50.91%	49620	43.13%	59.88%
(55.39743, 11.3298)	Slagelse	1814	23.99%	51.71%	50237	43.66%	60.63%
(55.73234, 9.56031)	Vejle	1865	24.66%	53.16%	52215	45.38%	63.01%
(55.66764, 12.56366)	Copenhagen	1865	24.66%	53.16%	52215	45.38%	63.01%
(56.71611, 10.11689)	Hadsund	1866	24.66%	53.19%	52216	45.38%	63.01%
(57.4273, 10.51456)	Frederikshavn	1866	24.66%	53.19%	52216	45.38%	63.01%
(56.99785, 10.30737)	Hals	1866	24.66%	53.19%	52216	45.38%	63.01%

Table 10: Impact of most important nodes in Denmark (Distance Range: 5-300km)

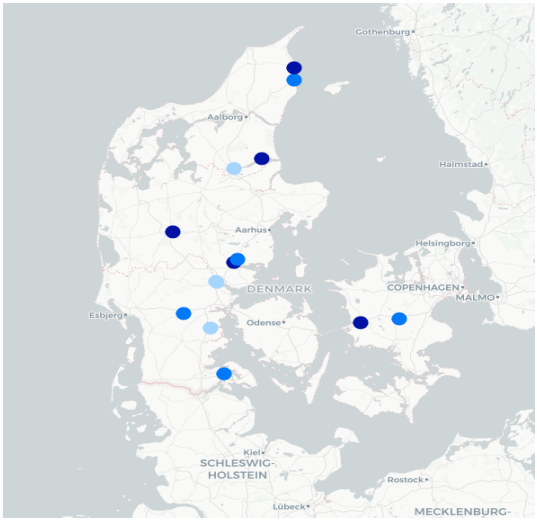

Figure 33: Top nodes in Denmark for the range [5,250]

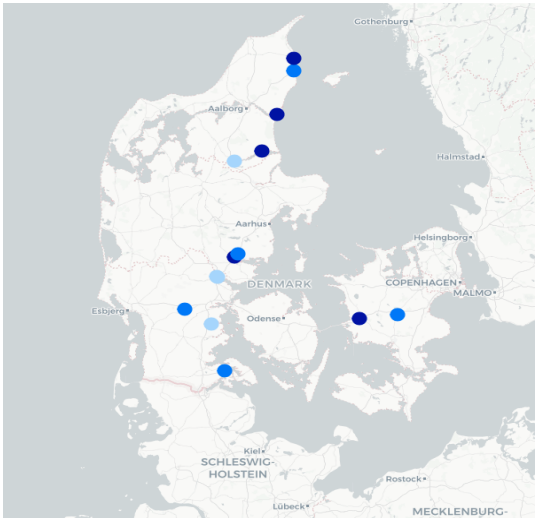

Figure 34: Top nodes in Denmark for the range [5,500]

potential by electrifying nodes 12-15) could potentially electrify close to half of the deliveries for the whole Denmark graph – which is huge, considering that Denmark accounts for 21.39% of the deliveries in the complete DFDS network. Again, these way higher electrification potentials compared to the UK or Sweden can be explained by the much smaller size of the country itself and also its graph, which makes it possible to reach large portions of the routes by electrifying only a few nodes.

These results are also resembled when comparing them to the [5; 250] and [5; 500] graphs. For these graphs, 13 ([5; 250]) and 14 ([5; 500]) of the most important nodes are also part of the identified important nodes of the [5; 300] graph, which can also be seen when looking at the geographical locations of these nodes (**??**). This indicates high scalability of the eTruck project in Denmark, meaning that the charging stations would still be at the right places in case eTruck range will decrease or increase.

Recommendation: In Denmark, an electrification of large parts of the route networks can be achieved by installing charging stations and deploying eTrucks on only a few nodes. In that regard, DFDS should especially look into the site in Horsens, as it is located along both main directions of travel - from Northern Jutland to Southern Jutland and from Sealand to Jutland. From there, the other identified nodes can be looked into, as the results are also very robust to possible future changes of the eTruck range.

7.2.4 Germany

While for the three countries we already analyzed there have always been significantly more deliveries than routes themselves, this is the opposite for Germany. Germany is associated with 13.1% of all the routes but only with 4.31% of all the deliveries, which suggests that in Germany we indeed deal with many routes that aren't highly frequented. Remembering the results of the Visual Analytics section of this paper and the geographical location of Germany, this also points to the assumption that the German routes are mostly used as "transit" routes in order to transport goods from the north (Denmark, Sweden) on to the (south-)west towards France or the UK.

This also is clearly visible when looking at the route distance distribution for Germany (Figure 36), which looks entirely different to the distributions in other countries; In Germany, we see only a low amount of routes within the current eTruck range of 300 km, but a very large amount of long haul operations with distances between 300 and 2,000 km. On a first glance, this makes Germany not the most attractive country for DFDS to focus on – on the other side, Germany heavily subsidizes eMobility (BMWK-Federal Ministry for Economics Affairs and Climate Action, 2017), which is why DFDS still wants to look into the implications of electrifying the shorter routes within their network in Germany. Looking at the most

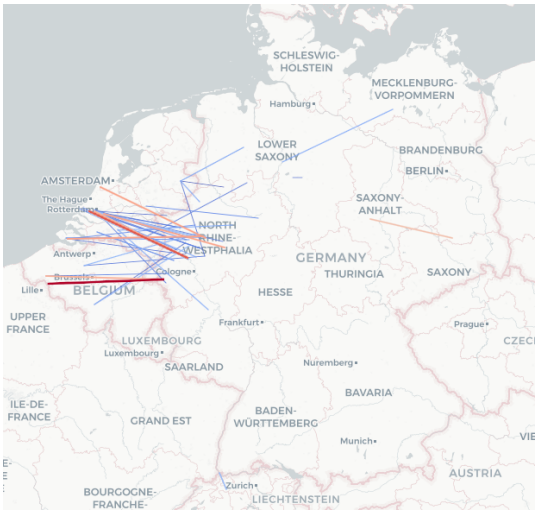

Figure 35: Top 2.5% routes in Germany

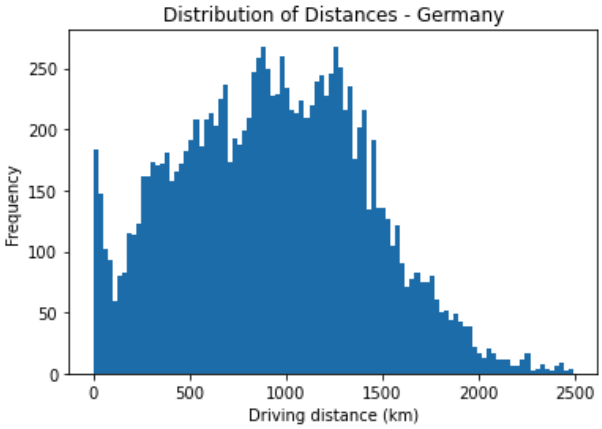

Figure 36: Route frequency distribution in Germany

important routes in Germany (Figure 35), the assumption of Germany especially being a transit country is confirmed. Significant routes can almost exclusively be found in the Western part of the country, connecting places in the state of North-Rine Westphalia to places abroad, such as in the Netherlands and Belgium. The by far most important route connects Würselen, a small city close to the Belgian border, with Mouscron, a city in the western part of Belgium bordering France. In general, it is interesting to see how the route network for Germany heavily includes places in Belgium and the Netherlands.

Measure	Value
Number of Nodes	3,114
Number of Edges	2,741
Avg. Degree	1.760
Avg. Clustering Coefficient	0.013
Number of Communities	801
Modularity of Community Structure	0.959

Table 11: Graph Characteristics for Germany (Distance Range: 5-300km)

Constructing the graph for this network reveals 3,114 nodes and 2,741 edges, the latter one being significantly less than in the way smaller country Denmark, again proving the assumption about Germany as a "transit" country. However, in all the other relevant graph measures: average clustering coefficient (0.013), number of communities and modularity of the community structure (801/0.959) as well as the conductance scores for the largest three com-

munities (0.004, 0.037, 0.011) (Table 12), the graph and its communities look exceptionally well defined. Especially the very high modularity score in combination with the low conductance scores has to be pointed out here, which suggests very clearly distinct and defined communities within the graph.

Community	Size (Nodes)	Size (Conductance Score)
Community #1	225	0.004
Community #2	214	0.011
Community #3	89	0.037

Table 12: Largest Communities in the Germany Graph (Distance Range: 5-300km)

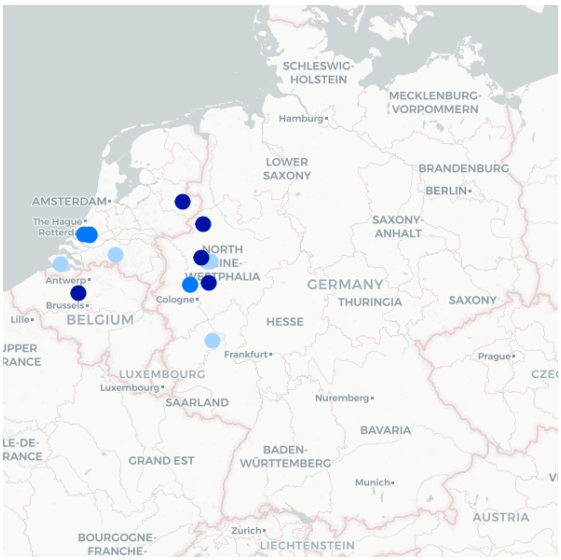

Figure 37: Top nodes in Germany for the range [5,300]

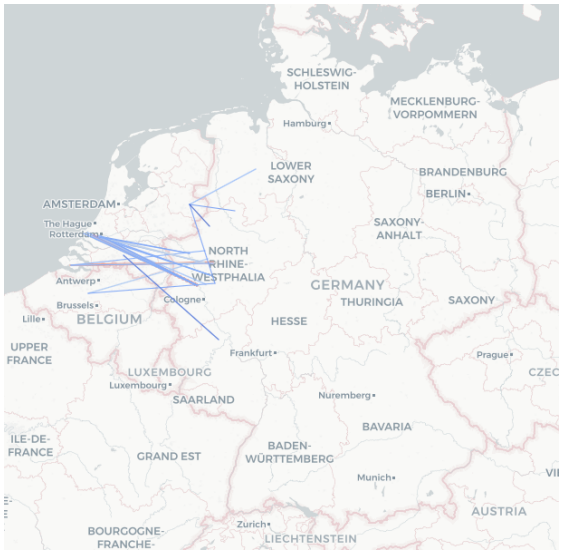

Figure 38: Route overlap in Germany

In line with the previous findings, the top 15 nodes (Figure 37) are scattered around the Western part of Germany (eight), the Netherlands (six), and Belgium (one) – so, contrary to the previous countries, the route network in this case shows important nodes abroad, which might suggest looking at Germany, the Netherlands and Belgium as one entity.

Accordingly, the share of the routes that could be electrified if charging stations were being put up at these nodes are especially routes crossing the borders as well and cover 58.82% of the most important routes. However, the most interesting place to look into in this network can definitely be found in Neuss, next to Dusseldorf in western Germany, as the two most

Node	City	Electrified Routes	% Electrification (Routes) Total	% Electrification (Routes) Subset	Electrified Deliveries	% Electrification (Deliveries) Total	% Electrification (Deliveries) Subset
(51.1524, 6.7797)	Neuss	28	0.13%	1.02%	2222	1.15%	14.00%
(51.15239, 6.77973)	Neuss	48	0.22%	1.75%	3664	1.89%	23.08%
(51.89208, 4.28846)	Botlek Rotterdam (port) (Netherlands)	59	0.27%	2.15%	3992	2.06%	25.15%
(51.88743, 4.42584)	Rotterdam (port) (Netherlands)	70	0.32%	2.55%	4255	2.20%	26.80%
(52.36467, 6.61535)	Almelo (Netherlands)	268	1.22%	9.78%	5081	2.63%	32.01%
(51.45185, 3.72609)	Nieuwdorp (port) (Netherlands)	274	1.25%	10.00%	5855	3.03%	36.88%
(51.88259, 4.41928)	Pernis (port) (Netherlands)	289	1.32%	10.54%	6160	3.18%	38.81%
(51.4921, 7.17789)	Bochum	289	1.32%	10.54%	6160	3.18%	38.81%
(51.02602, 4.1479)	Dendermonde (Belgium)	310	1.41%	11.31%	6386	3.30%	40.23%
(51.49675, 7.26455)	Bochum	312	1.42%	11.38%	6393	3.30%	40.27%
(51.59159, 5.0236)	Tilburg (Netherlands)	346	1.58%	12.62%	6559	3.39%	41.32%
(51.54411, 7.06145)	Gelsenkirchen	346	1.58%	12.62%	6559	3.39%	41.32%
(50.31059, 7.30857)	Polch	352	1.61%	12.84%	6676	3.45%	42.06%
(52.03955, 7.09178)	Ledgen	353	1.61%	12.88%	6677	3.45%	42.06%
(51.18532, 7.22614)	Remscheid	354	1.62%	12.91%	6679	3.45%	42.08%

Table 13: Impact of most important nodes in Germany (Distance Range: 5-300km)

important nodes are located there. Interestingly, these nodes only account for 48 routes in total – however, these 48 routes handle 23.08% of the deliveries of the subset network with distance <=300 km. The percentages for the whole network are expectedly very low, which goes in line with the earlier discovery that the network especially features long haul operations. Overall, the largest parts of electrification potential can be achieved by focusing on the nine most important nodes (which contain five nodes in the Netherlands, three in Germany and one in Belgium), which would yield to being able to electrify about 40% of the deliveries within the 300k m distance range.

The comparison with the [5; 250] graph and the [5; 500] graph reveals mixed results. While the first five of the most important nodes are identified in all graphs, the rest of the important nodes only shows little overlap (Figure 38). Hence, DFDS could confidently electrify the five most important nodes in Germany, but would need to be very careful in selecting further nodes to electrify if need be.

Recommendation: It showed that the DFDS route network in Germany is very inter-

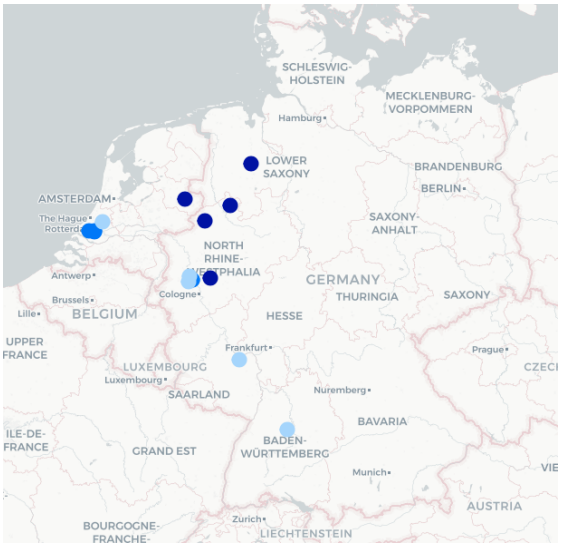

Figure 39: Top nodes in Germany for the range [5,250]

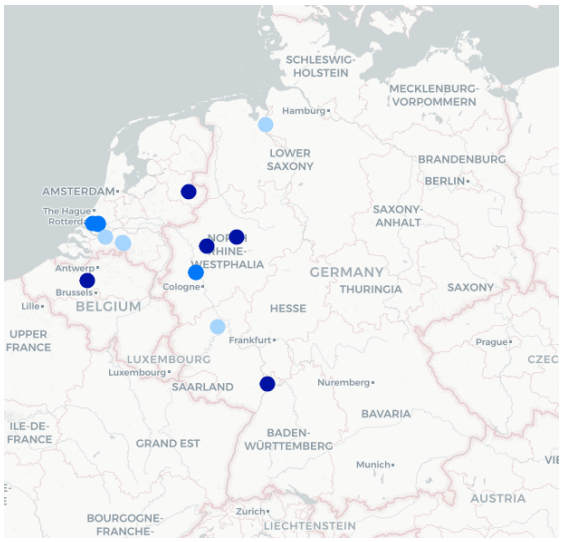

Figure 40: Top nodes in Germany for the range [5,500]

twined with Belgium and the Netherlands, and many of the detected important nodes in this network are located in one of these two countries as well. That makes the recommendation for Germany fairly clear, as the two most important nodes were found in Neuss, North-Rhine Westphalia. Electrifying these nodes could result in an important step to also electrify the "transit" routes from France,Belgium and the Netherlands in the west to Denmark and onwards to Sweden in the north. If further capacities should be available, the next sites to look into should be the ones in Bochum.

7.2.5 Belgium

The second to last country we are looking at is Belgium, which accounts for 7.95% of the routes and 8.38% of the deliveries and hence has a fairly balanced routes-to-deliveries ratio.

By looking at the top 2.5% routes (Figure 41, it can be observed that the majority of DFDS flows in Belgium are concentrated in the northern-central area of Belgium, between Kortrijk (and more specifically Wevelgem and Moeskroen), Ghent and Brussels, where we find the routes with the highest frequencies. These connect the industrial centers close to Kortrijk to those in Ghent and the the top route linking the port in Ghent to another industrial hub in its proximity, where DFDS distribution facilities are placed. However, a lot of them extend to the very south end of Belgium and over the country borders, to Paris, London, Amsterdam and Gouda in the Netherlands and Würselen in Germany. Multiple routes link the central

area of Belgium with London, therefore, understanding the mode of transport is in this case relevant. However, given that the extremities of the routes are not located in port towns, but in the hinterland it is likely to assume that these trucks cross the Channel passing through France, making them still - at least partially - relevant for the analysis.

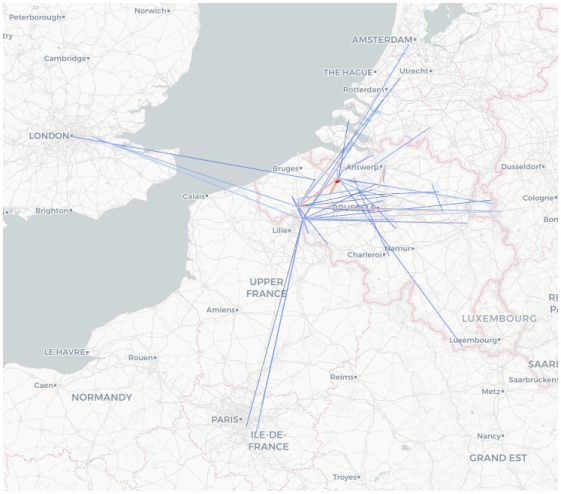

Figure 41: Top 2.5% routes in Belgium

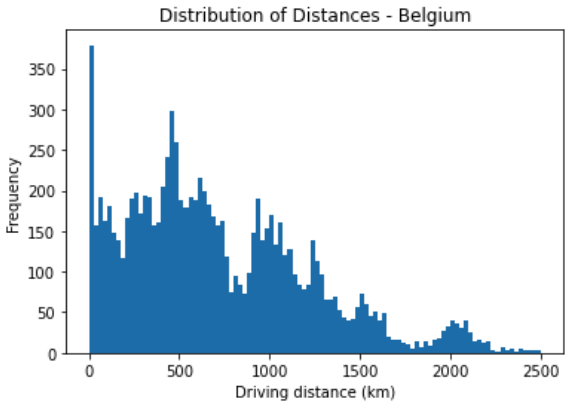

Figure 42: Route frequency distribution in Belgium

In terms of distance distribution (Figure 36), the highest frequencies are recorded for the shortest trips, which can be easily related to the route between the port of Ghent and the industrial area or similar scenarios. Differently from countries like UK, Sweden or Denmark, which show very high distribution for shorter ranges with a fast decline over 500km, here the decline is less abrupt. Indeed, routes between 500 and around 1250 Km still have a high frequency, compared to the maximum reached. Also, a considerable amount of journeys can be placed between 300 and 500 Km, which means their electrification potential is currently limited, but in case of technological developments increasing the battery capacity, they could potentially become more relevant for DFDS' goals.

Moving our focus to the network size presented in Table 14, the Belgian network is formed by 2437 nodes and 2732 edges, for an average degree of 2.24 edges per node. While being much more reduced in size compared to a country like Sweden, the average degree is very close. This means that despite the smaller size, the Belgian network shows an equal complexity. Considering the top routes in Figure 41 this can be visually seen as well.

The coefficient is again close to 0, recalling this typical behaviour of logistic networks by

Measure	Value
Number of Nodes	2,437
Number of Edges	2,731
Avg. Degree	2.241
Avg. Clustering Coefficient	0.045
Number of Communities	280
Modularity of Community Structure	0.846

Table 14: Graph Characteristics for Belgium (Distance Range: 5-300km)

creating strong connections - and eventually communities - between nodes not necessarily geographically close with each other. More specifically, 280 communities are identified, with a modularity score of 0.846, indicating strong connections within its members. This is validated when looking at the rather low conductance scores of the top three communities - 0.128, 0.016 and 0.222, respectively -, meaning that they are clearly identifiable, with reduced connection with nodes that are not part of them (Table 15.

Community	Size (Nodes)	Size (Conductance Score)
Community #1	347	0.128
Community #2	324	0.016
Community #3	89	0.222

Table 15: Largest Communities in the Belgium Graph (Distance Range: 5-300km)

When plotting the nodes of the top communities (Figure 43, we can see that there are two groups of nodes clustered around Ghent and Kortrijk and two lines of edges plotted consecutively, in addition to the single one located in the Paris area. This seems to represents a distribution structure, with goods departing from warehouses and then being distributed in different directions. The fact that the nodes in the same direction do not belong to the same community should not be concerning, as it is possible that the each community deals with a specific type of product or a specific customer and therefore this create strong connection between locations far apart from each other. This network structure also explains the fact that the overlap with the important routes that could potentially be electrified if charging stations are put up at these nodes is quite high, like shown in Figure 44.

Keeping these aspect in mind, the next analysed data are concerned with the number of

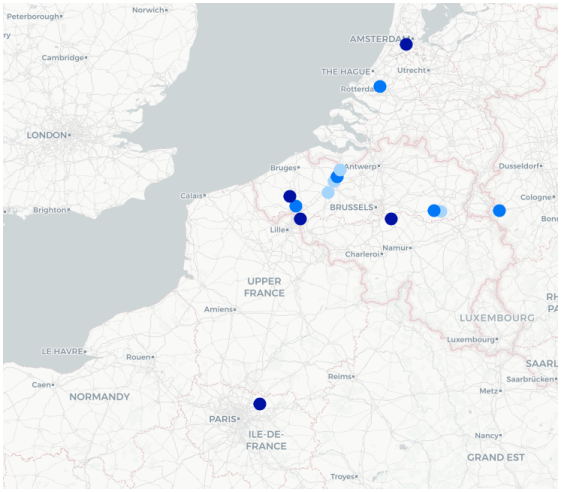

Figure 43: Top nodes in Belgium for the range [5,300]

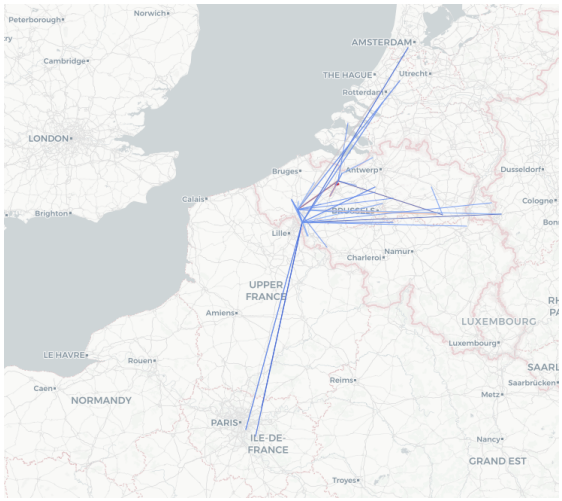

Figure 44: Route overlap in Belgium

routes and deliveries electrifiable by placing charging stations at top nodes (Table 16. Again, among the top 5 nodes, the first two (Wevelgem and Moeskroen) are located just close to Kortrijk and the following three in the Ghent area. Going more in detail, adding EV charging stations in Wevelgem would only allow to cover 2.76% of total routes and 4.93% of deliveries, rising up to 6.30% and 7.9% respectively if charging station on the second most important node were added. These numbers would increase again when considering the third node, with 7.32% of routes and 10.02% of deliveries. However, after that the additional coverage would only be marginal, reaching a maximum total of 7.9% routes and 12.16% deliveries covered in total. This means, that it is not possible to electrify certain routes as longer than 500 km. However, the shorter routes generally have higher relevance as with the electrification of slightly less than 40% of routes allow to have more than 60% of deliveries relying on eTrucks.

Finally, when comparing the other observed scenarios in Figure 45 and Figure 46, there does not seem to be major difference, with almost all the node in Belgium remaining the same. In the shorter range, the only relevant modification is that the node in Amsterdam would be substituted by another one in the Paris area. On the other hand, in the [5,500] case, nodes in France would not be as relevant, while one in Peterborough would result among the top nodes (Figure 46). Overall, this results in high scalability for the eTruck project in Belgium as the identified important nodes remain almost the same in different scenarios.

Recommendation: The route network in Belgium and the geographical distribution of

Node	City	Electrified Routes	% Electrification (Routes) Total	% Electrification (Routes) Subset	Electrified Deliveries	% Electrification (Deliveries) Total	% Electrification (Deliveries) Subset
(50.85454, 3.18812)	Wevelgem	376	2.76%	13.77%	12419	4.93%	24.90%
(50.73771, 3.24732)	Moeskroen	857	6.30%	31.38%	19888	7.90%	39.87%
(51.11924, 3.77834)	Ghent	996	7.32%	36.47%	25226	10.02%	50.57%
(51.08508, 3.74928)	Ghent (port)	1014	7.45%	37.13%	29330	11.65%	58.80%
(50.97667, 3.65531)	Nazareth	1016	7.47%	37.20%	29332	11.65%	58.80%
(51.12622, 3.78695)	Ghent	1016	7.47%	37.20%	29332	11.65%	58.80%
(50.80449, 5.30291)	Borgloon	1021	7.50%	37.39%	29676	11.79%	59.49%
(50.81017, 6.16851)	Würselen (Germany)	1022	7.51%	37.42%	29677	11.79%	59.49%
(50.81449, 5.2071)	Sint-Truiden	1022	7.51%	37.42%	29677	11.79%	59.49%
(51.9441, 4.41363)	Rotterdam (Netherlands)	1022	7.51%	37.42%	29677	11.79%	59.49%
(51.18492, 3.83621)	Wachtebeke	1055	7.75%	38.63%	30540	12.13%	61.22%
(50.94423, 3.09295)	Roeselare	1063	7.81%	38.92%	30577	12.15%	61.30%
(52.32232, 4.8003)	Schiphol (Netherlands)	1063	7.81%	38.92%	30577	12.15%	61.30%
(50.73689, 4.57978)	Wavre	1075	7.90%	39.36%	30621	12.16%	61.39%
(48.99551, 2.65332)	Compans (France)	1075	7.90%	39.36%	30621	12.16%	61.39%

Table 16: Impact of most important nodes in Belgium (Distance Range: 5-300km)

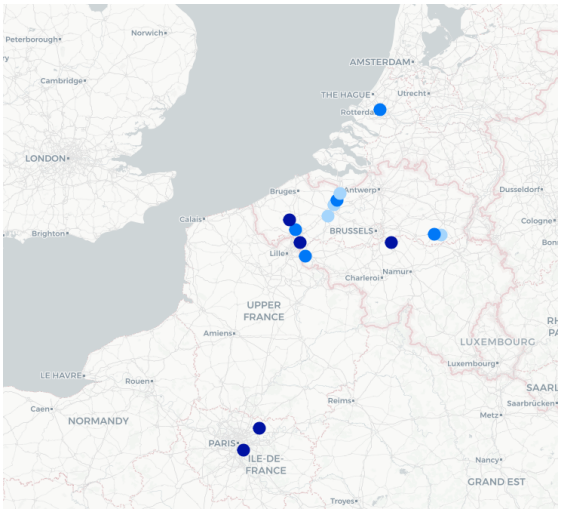

Figure 45: Top nodes in Belgium for the range [5,250]

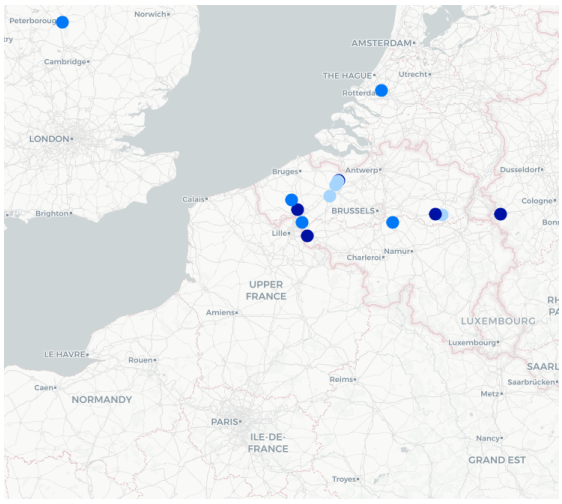

Figure 46: Top nodes in Belgium for the range [5,500]

the important nodes suggest to especially look into two places to install charging stations and deploy eTrucks: One would be around Ghent, as three of the six most important nodes are located in the area in and around the city. The second spot to look into would be somewhere near to the French border (e.g. Wevelgem or Moeskroen) in order to be able to electrify more routes going in a westerly direction (e.g. towards France and the UK) as well.

7.2.6 Netherlands

The last country identified to analyze in the previous sections is the Netherlands, which is associated with 15.68% of the routes and 7.17% of the deliveries. Similar to Germany, this is most likely reasoned by the Netherlands being a transit country for many goods, which results in many unique routes, many of them with start or end points abroad, with relatively little deliveries in comparison.

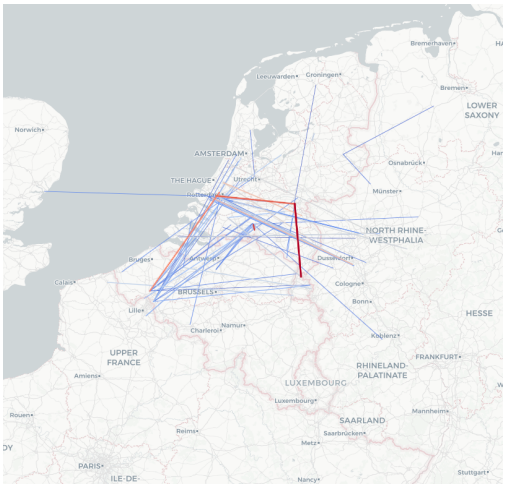

Figure 47: Top 2.5% routes in the Netherlands

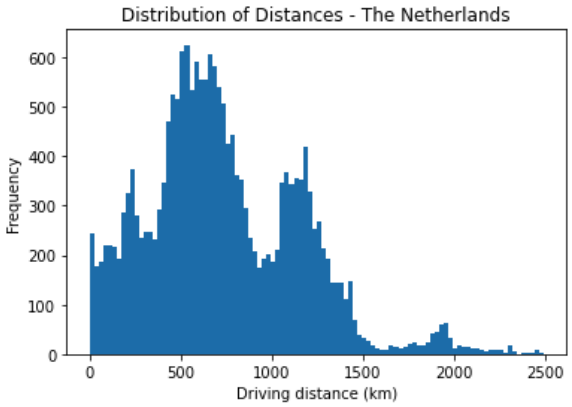

Figure 48: Route frequency distribution in the Netherlands

This is evident when looking at Figure 47 with the top 2.5% routes, as most of the routes cross the borders with either Belgium or Germany. Nevertheless, the most frequent routes are all domestic, with the exception of the one connecting Rotterdam with Kortrijk.

In terms of distance distribution (Figure 48), the highest frequencies are registered in the driving distance range between 500 and 1000 m, with two additional other relevant groups between 5 and 500 km and 1000 and 1500 km. However, since the majority of journeys is longer that 500 km, the country does not seem the most suitable candidate for eTruck deployment, even considering the most optimistic scenario of having a maximum driving range of 500 km.

When considering the overall DFDS network involving the Netherlands (Table 17), we can observe a total of 3,508 nodes, making it around the same size as the German network, but with 3,700 edges - representing around 1,000 additional routes compared to Germany - much more connected. Indeed, the average degree of edges per node is 2,111, a number comparable

Measure	Value
Number of Nodes	3,508
Number of Edges	3,702
Avg. Degree	2.111
Avg. Clustering Coefficient	0.013
Number of Communities	556
Modularity of Community Structure	0.916

Table 17: Graph Characteristics for the Netherlands (Distance Range: 5-300km)

to the one of Sweden, which on the other hand has more than 10,000 and 13,000 thousand nodes and edges, respectively. The average clustering coefficient is very low (0.013), indicating the absence of strong geographical clustering of nodes. The small clustering coefficient paired with a modularity value close to 1, replicates a scenario similar to the one seen before, with a large number of well-defined communities involving nodes not necessarily close to each other.

Community	Size (Nodes)	Size (Conductance Score)
Community #1	238	0.013
Community #2	235	0.029
Community #3	177	0.136

Table 18: Largest Communities in the Netherlands Graph (Distance Range: 5-300km)

As shown in Table 9, among all 556 communities, the three major ones represent 17% of the total number of nodes. Still, the conductance scores are close to 0 for the first couple and still low for the third. As mentioned before, this indicated that the communities are clearly separated, contrary to what could be visually expected considering the high number of routes passing through the southern part of the country.

When looking at the location of the most important nodes through Figure 49, it is interesting to see that 2/3 of them are actually located in Belgium. However, this in not surprising if considering the geographical proximity of these countries and the vast extension of the company flows. In the Netherlands, the most relevant locations are in the two major ports in Rotterdam and Amsterdam.

This same pattern can be seen also by observing Table 19 with the most important nodes, with the first two being in Belgian towns. By placing charging stations in Mouscron (Belgium)

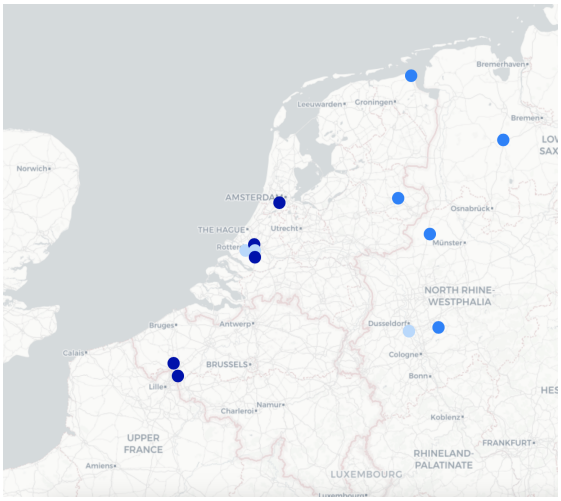

Figure 49: Top nodes in the Netherlands for the range [5,300]

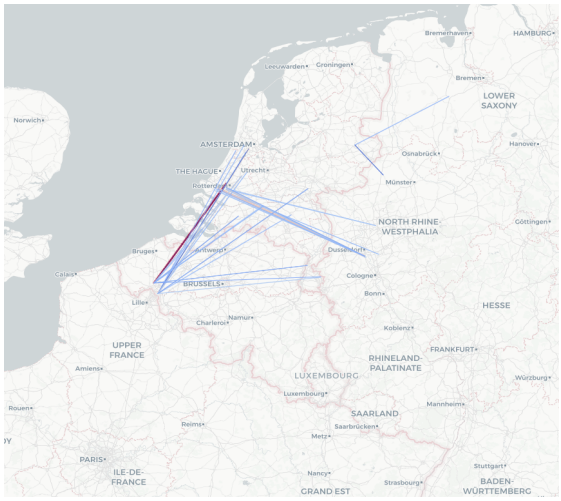

Figure 50: Route overlap in the Netherlands

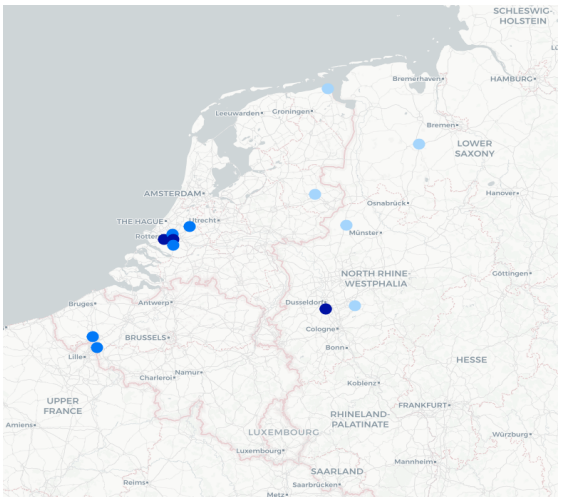

Figure 51: Top nodes in the Netherlands for the range [5,250]

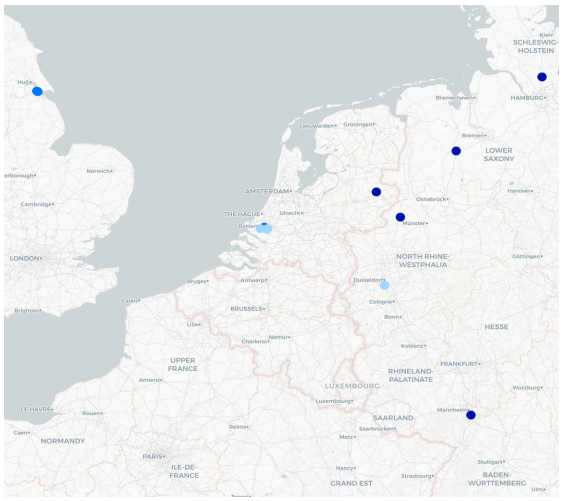

Figure 52: Top nodes in the Netherlands for the range [5,500]

it would be possible to electrify 172 of all the routes in the country, reaching almost 1% of the total if other EV charging stations were to be placed in Wevelgem (Belgium) as well. In general, by adding charging stations in a new location the total amount of routes electrified would increase, but given the structure of the Dutch network, this would be valid until the 10th node, as any additional location would not increase the number of electrified routes above 612. Based on these observations, the opportunity for electrification for DFDS remains considerably low, as by electrifying the 10 most important nodes, not even 10% of all routes would be covered and an even smaller percentage of deliveries (only 3.76%). However, if considering the electrifiable range below 300 km, the percentage of routes raises to 16.53%

Node	City	Electrified Routes	% Electrification (Routes) Total	% Electrification (Routes) Subset	Electrified Deliveries	% Electrification (Deliveries) Total	% Electrification (Deliveries) Subset
(50.73771, 3.24732)	Mouscron (Belgium)	172	0.61%	4.65%	2322	0.86%	5.99%
(50.85454, 3.18812)	Wevelgem (Belgium)	268	0.96%	7.24%	4543	1.68%	11.71%
(51.1524, 6.7797)	Neuss (Germany)	294	1.05%	7.94%	6751	2.49%	17.40%
(51.15239, 6.77973)	Neuss (Germany)	313	1.12%	8.45%	8191	3.02%	21.11%
(51.89208, 4.28846)	Botlek Rotterdam (port)	332	1.19%	8.97%	8610	3.18%	22.19%
(51.9441, 4.41363)	Rotterdam	337	1.20%	9.10%	8642	3.19%	22.28%
(51.88743, 4.42584)	Rotterdam (port)	356	1.27%	9.62%	8971	3.31%	23.12%
(52.36467, 6.61535)	Almelo	590	2.11%	15.94%	9941	3.67%	25.62%
(51.88267, 4.4193)	Rotterdam (port)	611	2.18%	16.50%	10199	3.76%	26.29%
(52.32232, 4.8003)	Schiphol	612	7.51%	16.53%	10201	3.76%	26.29%
(51.82961, 4.43223)	Oud-Beijerland	612	7.75%	16.53%	10201	3.76%	26.29%
(52.03955, 7.09178)	Ledgen (Germany)	612	7.81%	16.53%	10201	3.76%	26.29%
(52.88172, 8.21503)	Großenkneten (Germany)	612	7.81%	16.53%	10201	3.76%	26.29%
(51.18532, 7.22614)	Remscheid (Germany)	612	7.90%	16.53%	10201	3.76%	26.29%
(53.45059, 6.8101)	Eemshaven (port)	612	7.90%	16.53%	10201	3.76%	26.29%

Table 19: Impact of most important nodes in the Netherlands (Distance Range: 5-300km)

and that of deliveries to 26.29%. This can easily be explained by the fact that the vast majority of routes involving the country is longer than 500k, meaning that they have no potential to be covered by eTrucks as of now.

Finally, the other possible scenarios are observed. When considering shorter routes (Figure 51), 14 of the top nodes would still be among the most relevant. When increasing the span considered (Figure 52, the picture changes significantly as only eight nodes would remain relevant. Especially the fact that many nodes abroad (close to Hamburg, in Mannheim and in Hull) appear to be relevant in this selected range, would prove it difficult to scale the project if eTruck ranges increases to 500km.

Recommendation: The route network in the Netherlands has its four most important nodes in Mouscron, Wevelgem (both Belgium) and Neuss (Germany), all of which have been pointed out as locations to set up EV charging stations already in the recommendations of Belgium and Germany, respectively. As for the Netherlands themselves, the most important spot to look into would be Rotterdam and its port, as three of the remaining important nodes are located there.

8 Discussion of Results

8.1 Research Findings and Insights

In the previous two chapters, the two main approaches used in this research - Visual Analytics and Graph Theory - were presented and the results deriving from their application were illustrated. In this section, we aim at exploring and discussing the results in light of the initially defined research questions, in order to define what are the business implications of this and to provide DFDS with useful recommendations for the further development of the eTruck project.

Recalling subsection 2.2, the main goal of this study is to assess how Data Science methods can optimize positions for charging stations within the route network of a logistics provider, in this case DFDS. More specifically, for reasons discussed earlier, the choice fell on Visual Analytics and Graph Theory. Therefore, for each of these approaches, we will now try to give an answer to the respective subquestions and finally to the main research question, in light of the analyses laid out in the previous sections.

How can Visual Analytics be used to identify areas of high traffic and demand for charging stations within the route network of the logistic provider?

In the case of Visual Analytics, the goal was to identify areas of high traffic and demand for placing EV charging stations in order to decide on focus areas. Here, the creation of maps was useful in identifying the areas with higher densities in the company's network. These are mainly located in northern Europe (United Kingdom, Belgium, The Netherlands, the north and west of Germany, Denmark, and the south of Sweden) and the intensity decreases moving further away from those areas. Based on this distribution it was possible to assume that areas like Belgium, the Netherlands and northern Germany serve more a transit function, covering legs in a larger distribution flow from Sweden to the UK (and vice versa) and the rest of Europe. This also resonates with the name of some of the main customers in the booking logs (not available to external parties due to legal and privacy reason) and the network analysis of section 7. This does not mean that these location should be overlooked, but on the other hand, helped in gaining a deeper understanding of the type of infrastructure DFDS has.

However, the use of Visual Analytics was not only important to understand the geographical distribution of demand and it functionalities, but also to understand its behaviour through time. Reason for this is that introducing eTrucks implicates a large investment for the company and therefore it is valuable only if they can ensure that these new eTrucks are used in the optimal locations, where they can electrify as large parts of the network as possible. For this reason, a series of visualizations have been generated with a two-fold goal: on one hand, understanding the delivery patterns throughout the year (and between different years) and on the other the focusing on the delivery volumes of each country in the analysis.

To combine these perspectives, a dashboard was created: in this way it would be possible to have a overview in one screen of the demand over time as well as on a country basis and an interactive historical control of distribution from a geographical perspective.

This final product already provides an answer to the first research sub-question. However, as a standalone solution this would represent more a snapshot at present, while it would be more meaningful to create a solution able to reflect the current status whenever needed, given that the project is still in its early stages. As an example, the acquisition of a number of companies has highly impacted the distribution volumes in Denmark towards the end of 2022 and a similar event in the future could also modify the main candidates for further electrification. In this sense, the implementation suggested in the subsection 8.2 can provide support on that, suggesting a bi-yearly dashboard update with the introduction of the newly collected data.

How can Graph Theory techniques be used to optimize the placement of charging stations within the route network of the logistic provider, taking into account traffic flows and routes distances?

As outlined before, Visual Analytics comes with constraints, which made it necessary to deploy other techniques as well in order to dive deeper into DFDS' route network. Here, Graph Theory techniques were used, as by doing so we could transform DFDS' route networks in the identified focus areas into mathematical structures, graphs, which gave us a complete tool box of calculations at hand in order to answer the research question. Additionally, visualization techniques allowed to always follow the steps taken. The graphs were created at country level,

considering those countries identified based on the results of the Visual Analytics before and discussions with DFDS. For each country, multiple scenarios have been analyzed. First, the range that might be best suited for eTrucks to cover given the current technological status quo, with up to 300 km; Then, an extended scenario up to 500 kilometers that takes possible developments into consideration that may increase battery capacity, and, lastly, a scenario with a decreased range down to 250 km, knowing that various possibilities exist that might in fact limit the range of en eTruck (such as cold chain products, range anxiety, etc.).

In general, it's possible to say that Graph Theory and its visualizations through the folium library has at first helped in providing a comprehensive overview of the networks filling in the gaps of the Visual Analytics tools (e.g. by visualizing the most important routes) as they do not have the computational power to generate those visualizations. Additionally, we have been able to provide a more technical evaluation of the nature of these networks as well as the substructures and communities that have been identified within them through the use of Community Detection algorithms. This has made it possible to identify the most important nodes in the most important communities, starting from which, we could measure the contribution that each node could give to the electrification of routes and deliveries.

On a more tangible level, these insights could help the logistic provider in a multitude of ways. Firstly, the identified top nodes would already give suggestions on where to install charging stations and deploy eTrucks. While it is true that this analysis does not take into account the number of charging station to deploy, it is reasonable to start electrifying the routes with the highest identified potential. Certainly, the decision should be made by in-depth -evaluating each country's situation. However, in some cases, a cross-country comparison could also provide an additional perspective. Indeed, like in the case of transit countries such as Germany, Belgium, and the Netherlands, multiple of the top nodes actually overlap. For this reason, considering the three countries as a single region when deciding on charging station deployment could be reasonable, as it would allow to electrify the routes that provide the most shared contribution.

Also, the company should balance this quantitative data with additional information that can have an impact on the eTruck deployment potential, first of which being the grid capacity. In this sense, countries with high electrification potential may turn out to be less appealing if the grid capacity is unable to satisfy the company needs. Another aspect the logistic

provider should consider when deciding about positioning charging stations, is the type of goods transported. While here we have presented more in detail the 5 km to 300 km range, it could be more relevant to analyse the scenario with shorter ranges when knowing that a route is involved in cold-chain logistics.

In general, while for some countries the potential for electrification seems to be quite limited, this should not be a concern at this point in time. Indeed, the replacement of diesel trucks with eTrucks is still in its early stages, therefore for now it would be more relevant to target the most appealing routes in the different country networks.

Additionally, in some countries like Germany and the Netherlands the limits in electrification are due to the presence of a large number of routes outside the range of distance suitable for eTrucks. While this research cannot provide a solution for those cases, it still lays out interesting information that could be useful in the future, when DFDS will be starting to deploy hydrogen trucks which are more suited in those scenarios.

These explanations, will then contribute to answering the main research question:

How can Data Science methods be used to find optimal positions for charging stations within the route network of a logistic provider?

The first way Data Science methods can contribute to identify the optimal placement for charging station is by providing a data-driven approach to the project. Indeed, the suggested approaches are helpful in leveraging the large amount of data logistic providers have available, and more specifically plannings and registrations of deliveries. Through this, it is then possible to get a full understanding of the complete network, which is a necessary step before making decisions regarding specific locations. Finally, Data Science tools can be used to identify all potential candidate locations for charging station deployment, as well as measuring the contribution that each would give in the electrification process. In addition, given that different factors impact the battery capacity, this allows to create multiple scenarios, in order to adopt the one most suited for the specific situation.

To summarize, this section has explained how Data Science models can contribute to find optimal positions for charging stations. On one hand, Visual Analytics was more concerned with a broader understanding of the network in general, posing the necessary foundation for

further analysis. On the other, Graph Theory focused more on each specific country rather than the network as a whole, providing a higher level of detail for each of them. While both approaches have contributed to the the overall goal of the thesis, they also show shortcomings that in a way limit their potential. This and other limitation identified throughout the process, will be discussed in section 9.

8.2 Handing over the Solution to DFDS

As one of the last sections before concluding, we briefly want to elaborate on how our findings and solutions will be handed over to DFDS and possibly integrated into their IT infrastructure.

As seen on subsection 5.3, DFDS' data infrastructure setup is fairly complex, also considering that it is composed of a series of software applications interacting with each other and that part of the data is still residing on premises. Nevertheless, given the constrained scope of this project and the procedure of data anonymization taking place, it would be suggested to transfer data necessary for the decision making process in charging station deployment to the cloud. In particular, the used datasets were retrieved from DFDS' in-house *Velocity* software. From this data storage it is possible to pull data locally, but it could also be possible to store them in a cloud environment, by synchronising it with the data source. More specifically, the current DFDS cloud provider is Amazon Web Services (AWS).

When it comes to implementing the visualizations and dashboard created in Tableau, we believe having real-time data would not be necessary. On the contrary, it would be more meaningful to perform periodic updates on the initial database by adding newly available data. This would enable the project managers to monitor the development of DFDS flows and to derive new deployment plans or to make modifications to previous forecasts. Therefore, synchronising the transfer of data from *Velocity* to a AWS database (such as *DynamoDB*) would provide little benefit while increasing costs. A feasible solution, would be to use an event trigger like the one provided by AWS *EventBridge*. This service allows to trigger events base on specific criteria and/or time periods. This can be particularly useful when setting up applications combining different AWS services. In this case it could set a trigger for a Lambda function after a certain amount of minutes, seconds or days. AWS *Lambda* (AWS, 2023) is a service that allows to run code without the need of managing servers, making in this sense very versatile and flexible, especially in cases where the activity is performed seldom.

For what concerns the time window after which triggering the database update, a series of elements should be taken into account. Firstly, given the continuous developments in the industry, the increasing relevance of the subject and the DFDS goal of having electric trucks representing 25% of their fleet, the timeline for updates should undergo reconsideration based on the speed of developments. However, for the initial period an update twice a year would be sufficient. Here, it is relevant to consider that decision to buy new equipment - based on the understanding of new needs and replacement that have to be made - are made at the end of Q2 with purchases executed in Q3. For this reason, it would be suggested to enrich the observed database, so to have a clear understanding of the current situation in time to present request for new equipment.

In order to visualize the data and derive insights from them, it would then be possible to connect the data from AWS querying service like Athena to a Tableau platform. Such service allows to analyse data using standard SQL language and can allow to import data to Tableau through the specific *Amazon Athena* connector (Tableau, 2023a). Alternatively, it would be possible to create a specific front-end tool, residing as well on the cloud environment.

This more dynamic structure would also allow to connect additional data source that are valuable for the company's strategy. As an example, it would be possible to add data related to the current charging stations owned by DFDS as well as data on public chargers.

Similarly, an implementation for the graph applications can be developed as well. The current set up already offers a great range of flexibility on the user side: for example, the input function used in the code allows the user to select the desired countries and the desired route length range to look at, which can indeed be very helpful when considering different scenarios. However, the execution of this Jupyter notebook would still require a bit of familiarity with the programming language and with an IDE or code editor. Additionally, currently the generated maps and summary .html files are very conveniently stored in designed repositories, facilitating their retrieval. Nevertheless, considering that this solution is finally designed to support project managers in decision making processes, the creation of a front-end solution would help increasing the usability level also for the project managers. This front-end solution could be designed by the data team we have collaborated with during this thesis project by creating an interface from which requests can be sent in order to run the code based on the

desired parameters and output the resulting map and summary.

9 Limitations

In order to support DFDS in assessing where to deploy charging station for their eTrucks, we have mainly focused on logistic data related to their current flows. Nevertheless, the research shows some limitations, deriving from the characteristic of the data itself, the source of data taken into consideration and the technical choices made as well. Therefore, this section is devoted to analyse such limitations.

Some of the main limitations are concerned with the choice of data used: while logistics booking data are ideal in providing high detail for each booking, they are designed for planning purposes rather than a backwards-looking analysis. Here, as already mentioned in the subsection 5.5, in case of bookings with multiple legs the starting location is always said to be the starting point of the whole booking, without providing details on the exact order of the legs.

For this reason, this is not a "stand-alone" solution as it requires to collaborate with route planners. In particular, this role is mainly focused on the preparation of route schedules ensuring to carry out pick-ups and deliveries in the most optimal way. With the introduction of eTrucks, the recharging time and place would also be another aspect to factor in the equation. Therefore, dialogue with route planners have not been inside the scope of this thesis but may be required to understand if routes can be completely electrified or maybe just partially.

Other limitations concerning data are more related to data quality. As an example, the "Temperature" feature has more than 81% missing values, making it complex to derive any insights from that perspective. On the contrary, information about the temperature of the loads would be very valuable to understand precisely how many deliveries transport frozen goods. This aspect would then again be important to make more reasonable estimates on the driving range capacity, as it would be more limited given that part of the energy would be needed to ensure the required temperature is maintained. A similar observation can be applied to other features related to the volume, length, width and height of the load, even if in this context having these additional information would only provide a marginal additional value.

Moreover, in some cases, certain features are only contained in specific datasets. As an example, the two data sources provided for this research were joined together as a series of attributes were only present in one of them, some of which being the full load or empty load indicator. Again, this is valuable information as the load weight has an impact on the route capacity when it comes to eTrucks. Knowing whether it is an empty load, a FTL or LTL could be taken into account when defining the network weights.

Therefore, for what concerns this aspect, we suggest DFDS to promote more coordination and consistency within its distribution network while collecting these data as a way to provide additional information and move research further.

While the data at hand provides a valuable amount of information - despite their limitations - other sources of data that would have an impact on DFDS' decision making strategy have not been taken into account, posing in this sense additional limitations. One of this is related to the electrical grid capacity and conditions, which undoubtedly have an effect on the possibility of electrifying transportation networks. As an example, based on the discussion with the project managers it was possible to understand - even if just in qualitative terms - that the grid structure in the western parts of the UK (Manchester, Liverpool) is not yet strong enough to easily handle the placement of multiple charging stations for eTrucks. Therefore, even if the solution provided in the previous analysis subsection 7.2 would considerably facilitate the route electrification process it does not take into account the constraint the current grid condition in the country poses.

Furthermore, given the high costs the network electrification determines, DFDS desires to leverage as much as possible on governmental subsidies supporting the deployment of eTrucks. However, since they are tied to the country issuing them, decisions on where to introduce eTrucks could be influenced by this aspect as well. While for the countries on which this research has been focusing there are subsidies available, there may be other favourable policies the company is unaware of. Indeed, based on their experience, retrieving information about these subsidies is not always straightforward, as it depends on the easiness of use of governmental websites and whether they are available in English as well. Moreover, it is even harder to obtain information when governments are still discussing about the possibility of introducing subsidies. Therefore, given the limited data available and the difficulty in

validating information transferred through word of mouth, these aspects were not included in the research at hand.

As for the technical choices made, it is important to recall the decision on the specific algorithm to use for community detection. While a more detailed explanation for these assumption is presented in subsection 7.1, we hereby recall that the Louvain algorithm, an algorithm applicable for undirected graphs was here used on directed graphs, which made it necessary to convert the graphs first, resulting in an approximation of it. More generally, the type of community algorithm chosen can lead to different results and only a comparative analysis could help in understanding which one leads to the best result while also taking into account the algorithm complexity.

Moreover, as presented in subsection 7.2 three different scenarios representing different battery capacity constraints have been used: 5 to 300 km, 5 to 250 km, and 5 to 500 km. While the choice of the lower bound is related to the fact that smaller operations within hubs are not carried out by eTrucks, the upper bounds represent, respectively: an average measure of capacity (300), the capacity when considering factors such as cold-chain transport, tough weather conditions or range anxiety (250), and a potentially larger capacity resulting from technological developments (500). In the execution of the analysis we have supposed that once a charging station is placed at a node, all attached routes can be electrified. Therefore, from each charging station a maximum distance corresponding to the upper bound of the ranges can be covered. However, in this case it is not ensured that the eTruck will be able to return to its origin before needing to be charged, which might pose problems as the eTruck could run out of battery and is nowhere near a charging station. Nevertheless, it was decided not to halve the range distance as that would imply that all journeys go back to the departure point, which is definitely not the case, as we showed in subsection 5.4.

Finally, it is important to mention that while this research focuses on identifying the optimal locations for eTruck charging stations in DFDS' flows, this does not cover the decision on how many charging stations to deploy in each location. Different models of charging stations may come with differing costs, also based on the number of charging sockets they provide. As we did not have information about DFDS' budget for the deployment of charging stations, this aspect had to be left aside in our analysis but is important in the real world case at hand. In addition, since an international framework for fast-charging is being established

just now (CharIN, 2023), it is very likely that new developments would come in the short future. Therefore, providing information on the number of charging station to deploy would require a deep understanding of the project budget as well as defining different scenarios based on the expected developments, both of which exceed the scope of this thesis.

10 Conclusion and Future Work

The main objective of this thesis has been assessing how Data Science approaches can contribute to find optimal positions for charging stations within the route network of a logistic provider. In particular, the attention has been devoted to Visual Analytics and Graph Theory to identify how they can help in identifying areas of high traffic and demand and optimize the placement of charging stations in the company's route network.

This research was carried out in collaboration with DFDS, a European leader in providing both transportation and logistics services. Given their goal of reaching carbon neutrality by 2050, the deployment of eTrucks is one of their main priorities, and consequently also deciding on where to allocate them and where to install the needed charging stations. However, what has been lacking was a data-oriented approach able to support in their decision making processes - which is exactly what this thesis aimed to provide.

After obtaining data related to the deliveries carried out by DFDS in Europe between 01/01/2020 and 10/02/2023, the first step has involved performing an Explorative Data Analysis, to understand the main characteristics of DFDS' route network before stepping into additional processing and data engineering activities. Once the final datasets were prepared, Visual Analytics was used to better understand the flow distributions, both in volume and geographical terms. This has allowed to provide a dashboard to have a first data-oriented overview on the previously-mentioned aspects. The resulting dashboard has indeed contributed to identifying the six potential country candidates for charging station (and consequently eTrucks) deployment: United Kingdom, Belgium, the Netherlands, Germany, Denmark and Sweden.

Thereafter, the Graph Theory applications have specifically focused on these areas. Firstly, graphs have been created for each specific network and plotted on maps, using color gradients as measure of the importance of routes. While the range 5 to 300 km was identified as most representative given the context and the battery capacity, other scenarios with decreased and

increased range have been evaluated as well in order to assess scalability of the results.

Given the size of each transportation network, it was then deemed appropriate to use community detection algorithms to identify substructures within the graphs and their characteristics. Based on this, the five most important nodes of the three most important communities have been visualized for each range scenario and the common nodes have been identified. Finally, additional relevant information about the cumulative contribution to the network's electrification given by each "top node" have been provided.

To summarize, the final deliverable provided as support to find optimal positions for charging stations in DFDS' route network are a dashboard for the overall network as well as interactive network maps and summaries in html format for each of the six countries in analysis. Furthermore, all the code written to create these outputs is shared with DFDS, in order to make the results reproducable and also to allow the company to run their own analyses with scenarios of their choosing.

While this research rather provides recommendations for the placement of charging stations instead of a holistic and all encompassing solution, it definitely yields the data-driven support needed by DFDS to identify the optimal sites. Nevertheless, improvements to this process can still be made, and this should be the focus of future research.

On one hand, refinements in data quality are still possible: as mentioned in subsection 5.4, relevant fields such as load weight and temperature are generally unavailable. Therefore, similar analysis should be carried out defining more precise scenarios, provided enough data about these aspects are collected.

Similarly, a more detailed result could be obtained if additional data sources were to be integrated, such as quantitative information about the availability and the magnitude of subsidies for the deployment of eTrucks that some European governments have drafted. Or again, the research on charging stations pricing and DFDS budget evaluation could allow not only to define the location, but to evaluate the number of charging station to be deployed to satisfy the demand.

On the other hand, future work could still focus on the use of more complex and articulated models for identifying optimal locations. As an example, in order to apply the community detection algorithm selected, the graphs had to be converted in their undirected version, mak-

ing in this sense an approximation of their original directed form. Here, a tailored algorithm could be designed in order to perform community detection while preserving the nature of the graph as well.

Additionally, an optimization algorithm (e.g., linear programming, genetic algorithm, or a custom heuristic) could be designed to determine the optimal number and location of charging stations within the prioritized nodes. This optimization should consider factors such as the charging capacity of stations, eTruck range, and the desired percentage of fleet electrification. The objective function can be set to minimize the total cost of charging infrastructure while maximizing coverage and convenience for eTrucks.

While the issue of optimally placing charging station is not a new topic in the literature, research up to date has mainly focused on the user perspective and small electric vehicles rather than long haul trucks. For this reasons, this thesis is providing a different perspective on the subject by focusing on solving the problem of optimal charging station location for a private logistics company in the process of deploying some of their first eTrucks. In addition, this thesis has used a novel approach, as rather than focusing on purely mathematical optimization approaches, it aims at creating the tools that will support the formulation of business decisions through an actionable data-driven approach. This has enabled us to identify the six countries DFDS should focus on as well as to provide suggestions on how to prioritize the electrification process. Therefore, despite not providing an exact answer in determining in which locations charging stations should be deployed, it definitely represents a fundamental tool for DFDS - and potentially other logistics providers - for reaching their electrification objectives.

Bibliography

Abdirassilov, Z., & Sładkowski, A. (2018). Application of artificial neural networks for short-term prediction of container train flows in direction of China – Europe via Kazakhstan. *Transport Problems*, *T. 13, z. 4*. https://doi.org/10.20858/tp.2018.13.4.10

Albadrani, A., Alghayadh, F., Zohdy, M. A., Aloufi, E., & Olawoyin, R. (2021). Performance and Predicting of Inbound Logistics Processes Using Machine Learning. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 0790–0795. https://doi.org/10.1109/CCWC51732.2021.9376171

Andrenacci, N., Ragona, R., & Valenti, G. (2016). A demand-side approach to the optimal deployment of electric vehicle charging stations in metropolitan areas. *Applied Energy*, *182*, 39–46. https://doi.org/10.1016/j.apenergy.2016.07.137

AWS. (2023). Serverless Computing - AWS Lambda - Amazon Web Services. Retrieved April 30, 2023, from https://aws.amazon.com/lambda/

Badiee, A., Kalantari, H., Ghazanfari, M., Fathian, M., & Shahanaghi, K. (2020). Introducing drivers' collaboration network: A two-layers social network perspective in road transportation system analysis. *Research in Transportation Business & Management*, *37*, 100532. https://doi.org/10.1016/j.rtbm.2020.100532

Beckers, J., Thomas, I., Vanoutrive, T., & Verhetsel, A. (2018). Logistics clusters, including inter-firm relations through community detection [Number: 2]. *European Journal of Transport and Infrastructure Research*, *18*(2). https://doi.org/10.18757/ejtir.2018.18.2.3229

Beckers, J., Vanhoof, M., & Verhetsel, A. (2019). Returning the particular: Understanding hierarchies in the Belgian logistics system. *Journal of Transport Geography*, *76*, 315–324. https://doi.org/10.1016/j.jtrangeo.2017.09.015

Bernard, M. R., Tankou, A., Cui, H., & Ragon, P.-L. (2022). Charging solutions for battery electric trucks.

BMWK-Federal Ministry for Economics Affairs and Climate Action. (2017). Regulatory environment and incentives for using electric vehicles and developing a charging infrastructure. Retrieved May 13, 2023, from https://www.bmwk.de/Redaktion/EN/Artikel/Industry/regulatory-environment-and-incentives-for-using-electric-vehicles.html

Bollobás, B. (1998). *Modern Graph Theory* (Vol. 184). Springer. https://doi.org/10.1007/978-1-4612-0619-4

Bondy, A., & Murty, U. S. R. (1976). GRAPH THEORY WITH APPLICATIONS.

Brandes, U. (2001). A Faster Algorithm for Betweenness Centrality [Publisher: Routledge]. *Journal of Mathematical Sociology*, *25*(2), 163. https://doi.org/10.1080/0022250X.2001.9990249

Bricher, D., & Müller, A. (2020). A Supervised Machine Learning Approach for Intelligent Process Automation in Container Logistics. *Journal of Computing and Information Science in Engineering*, *20*(3). https://doi.org/10.1115/1.4046332

Calzon, B. (2022). What Is A Data Dashboard? Definition, Meaning & Examples. Retrieved April 29, 2023, from https://www.datapine.com/blog/data-dashboards-definition-examples-templates/

Capgemini. (2022). Data engineering & Data migration - Lecture at CBS for the course "Applied Machine Learning and Data Engineering in Business Context".

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9(13), 1-73. *SPSS inc*(9(13)), 1–73.

CharIN. (2023). CharIN – Empowering the next level of e-mobility. Retrieved April 29, 2023, from https://www.charin.global/

Davidov, S., & Pantoš, M. (2017). Planning of electric vehicle infrastructure based on charging reliability and quality of service. *Energy*, *118*, 1156–1167. https://doi.org/10.1016/j.energy.2016.10.142

De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011). Generalized Louvain method for community detection in large networks [ISSN: 2164-7151]. *2011 11th International Conference on Intelligent Systems Design and Applications*, 88–93. https://doi.org/10.1109/ISDA.2011.6121636

DFDS. (2023a). Annual Report 2022. Retrieved April 18, 2023, from https://downloads.ctfassets.net/mivicpf5zews/30de54Ianj9yGhjVvOxz11/8699bff2907d57956e6538a342c06514/DFDS_NO_10_24_02_2023_ANNUAL_REPORT_2022.pdf

DFDS. (2023b). Annual Review 2022. Retrieved April 21, 2023, from https://assets.ctfassets.net/mivicpf5zews/1nUGz9g8adQEUdLylBIMc3/40f479929a7b35f7b38c9110f5d52916/DFDS_24_02_2023_Annual_Review_2022.pdf

DFDS. (2023c). Immingham Terminal — Serviceleistungen am Terminal. Retrieved May 12, 2023, from https://www.dfds.com/de-de/frachtschifffahrt/serviceleistungen-am-terminal/immingham-terminal

DFDS. (2023d). Logistics Solutions — Logistics Services. Retrieved May 13, 2023, from https://www.dfds.com/en-gb/logistics-solutions

DFDS. (2023e). Long term climate action plan — Sustainability. Retrieved April 29, 2023, from https://www.dfds.com/en/about/sustainability/climate-plan/long-term

Donateo, T., Licci, F., D'Elia, A., Colangelo, G., Laforgia, D., & Ciancarelli, F. (2015). Evaluation of emissions of CO2 and air pollutants from electric vehicles in Italian cities. *Applied Energy*, *157*, 675–687. https://doi.org/10.1016/j.apenergy.2014.12.089

Encyclopaedia Britannica. (2023). Latitude and longitude — Definition, Examples, Diagrams, & Facts — Britannica. Retrieved April 6, 2023, from https://www.britannica.com/science/latitude

European Commission. (2021). Volume of carbon dioxide emissions from commercial vehicles in the EU between 2015 and 2019 (in million metric tons) [Graph]. In Statista. Retrieved April 29, 2023, from https://www-statista-com.esc-web.lib.cbs.dk/ statistics/1230285/eu-commercial-vehicles-carbon-dioxide-emissions-volume/. Retrieved February 4, 2023, from https://www.statista.com/study/89318/zero-emission-commercial-vehicles-in-europe/

European Commission. (2022). Commission proposes certification of carbon removals. Retrieved April 21, 2023, from https://ec.europa.eu/commission/presscorner/detail/en/ip_22_7156

Feng, T., & Timmermans, H. J. P. (2015). Detecting activity type from GPS traces using spatial and temporal information [Number: 4]. *European Journal of Transport and Infrastructure Research*, *15*(4). https://doi.org/10.18757/ejtir.2015.15.4.3103

Few, S. (2007). Data Visualization - Past, Present, and Future.

Giuffrida, N., Fajardo-Calderin, J., Masegosa, A. D., Werner, F., Steudter, M., & Pilla, F. (2022). Optimization and Machine Learning Applied to Last-Mile Logistics: A Review [Number: 9 Publisher: Multidisciplinary Digital Publishing Institute]. *Sustainability*, *14*(9), 5329. https://doi.org/10.3390/su14095329

Göçmen, E., & Erol, R. (2019). Transportation problems for intermodal networks: Mathematical models, exact and heuristic algorithms, and machine learning. *Expert Systems with Applications*, *135*, 374–387. https://doi.org/10.1016/j.eswa.2019.06.023

Gong, L., Fu, Y., & Li, Z. (2016). Integrated planning of BEV public fast-charging stations. *The Electricity Journal*, *29*(10), 62–77. https://doi.org/10.1016/j.tej.2016.11.010

Guerra, S. P., George, V. K., Morar, V., Joshua, R., & Silva, G. (2022). *On using undirected graph techniques for directed graphs through Category Theory* (preprint). In Review. https://doi.org/10.21203/rs.3.rs-1995489/v2

Gujral, E., Papalexakis, E. E., Theocharous, G., & Rao, A. (2019). Hacd: Hierarchical Agglomerative Community Detection In Social Networks [ISSN: 1551-2541]. *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. https://doi.org/10.1109/MLSP.2019.8918734

Hakimi, S. L. (1964). Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph [Publisher: INFORMS]. *Operations Research*, *12*(3), 450–459. Retrieved April 15, 2023, from https://www.jstor.org/stable/168125

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2020). Chapter 3 - Social network analysis: Measuring, mapping, and modeling collections of connections. In D. L. Hansen, B. Shneiderman, M. A. Smith, & I. Himelboim (Eds.), *Analyzing Social*

Media Networks with NodeXL (Second Edition) (pp. 31–51). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-817756-3.00003-0

Hosseini, S., & Sarder, M. (2019). Development of a Bayesian network model for optimal site selection of electric vehicle charging station. *International Journal of Electrical Power & Energy Systems*, *105*, 110–122. https://doi.org/10.1016/j.ijepes.2018.08.011

Houser, K. (2018). EV-charging roads have arrived. Here's why we do (and don't) need them. Retrieved February 21, 2023, from https://futurism.com/ev-charging-roads-sweden

Hunter, C., Penev, M., Reznicek, E., Lustbader, J., Birky, A., & Zhang, C. (2021). *Spatial and Temporal Analysis of the Total Cost of Ownership for Class 8 Tractors and Class 4 Parcel Delivery Trucks* (tech. rep. NREL/TP-5400-71796, 1821615, MainId:6232). https://doi.org/10.2172/1821615

IER. (2022). Electric Vehicle Battery Costs Soar. Retrieved April 29, 2023, from https://www.instituteforenergyresearch.org/renewable/electric-vehicle-battery-costs-soar/

Irles, S. (2023). Battery electric trucks emit 63% less GHG emissions than diesel. Retrieved May 13, 2023, from https://theicct.org/battery-electric-trucks-emit-63-less-ghg-emissions-than-diesel/

John Hodgson, M., Rosing, K., Leontien, A., & Storrier, G. (1996). Applying the flow-capturing location-allocation model to an authentic network: Edmonton, Canada. *European Journal of Operational Research*, *90*(3), 427–443. https://doi.org/10.1016/0377-2217(95)00034-8

Kuby, M., & Lim, S. (2005). The flow-refueling location problem for alternative-fuel vehicles. *Socio-Economic Planning Sciences*, *39*(2), 125–145. https://doi.org/10.1016/j.seps.2004.03.001

Kuby, M., & Lim, S. (2007). Location of Alternative-Fuel Stations Using the Flow-Refueling Location Model and Dispersion of Candidate Sites on Arcs. *Networks and Spatial Economics*, *7*(2), 129–152. https://doi.org/10.1007/s11067-006-9003-6

Kulkarni, P., Joshi, S., & Brown, M. S. (2016). *BIG DATA ANALYTICS* [Google-Books-ID: j3KhDAAAQBAJ]. PHI Learning Pvt. Ltd.

Lahyani, R., Khemakhem, M., & Semet, F. (2015). Rich vehicle routing problems: From a taxonomy to a definition. *European Journal of Operational Research*, *241*(1), 1–14. https://doi.org/10.1016/j.ejor.2014.07.048

Li, J., & Zhang, W. (2016). Identifying spatial structure of travel modes through community detection method. *2016 IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, 227–231. https://doi.org/10.1109/ICITE.2016.7581337

Lin, L., Wang, Q., & Sadek, A. W. (2014). Data Mining and Complex Network Algorithms for Traffic Accident Analysis [Publisher: SAGE Publications Inc]. *Transportation Research Record*, *2460*(1), 128–136. https://doi.org/10.3141/2460-14

Lin, Z., Ogden, J., Fan, Y., & Chen, C.-W. (2008). The fuel-travel-back approach to hydrogen station siting. *International Journal of Hydrogen Energy*, *33*(12), 3096–3101. https://doi.org/10.1016/j.ijhydene.2008.01.040

Luo, C., Huang, Y.-F., & Gupta, V. (2017). Placement of EV Charging Stations—Balancing Benefits Among Multiple Entities [Conference Name: IEEE Transactions on Smart Grid]. *IEEE Transactions on Smart Grid*, *8*(2), 759–768. https://doi.org/10.1109/TSG.2015.2508740

Majima, T., Takadama, K., Watanabe, D., & Katuhara, M. (2014). Application of community detection method to generating public transport network. *Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies*, 243–250. https://doi.org/10.4108/icst.bict.2014.257865

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories [Conference Name: IEEE Transactions on Knowledge and Data Engineering]. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Mastoi, M. S., Zhuang, S., Munir, H. M., Haris, M., Hassan, M., Usman, M., Bukhari, S. S. H., & Ro, J.-S. (2022). An in-depth analysis of electric vehicle charging station infrastructure, policy implications, and future trends. *Energy Reports*, *8*, 11504–11529. https://doi.org/10.1016/j.egyr.2022.09.011

McKinsey. (2022). Electric vehicle charging stations in Europe. Retrieved May 13, 2023, from https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/europes-ev-opportunity-and-the-charging-infrastructure-needed-to-meet-it

McKinsey. (2023). Why most electric trucks will choose overnight charging. Retrieved May 13, 2023, from https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/why-most-etrucks-will-choose-overnight-charging

Mesa-Arango, R., & Ukkusuri, S. V. (2015). Demand clustering in freight logistics networks. *Transportation Research Part E: Logistics and Transportation Review*, *81*, 36–51. https://doi.org/10.1016/j.tre.2015.06.002

Micari, S., Polimeni, A., Napoli, G., Andaloro, L., & Antonucci, V. (2017). Electric vehicle charging infrastructure planning in a road network. *Renewable and Sustainable Energy Reviews*, *80*, 98–108. https://doi.org/10.1016/j.rser.2017.05.022

Microsoft. (2022). Routes API - Bing Maps. Retrieved May 13, 2023, from https://learn.microsoft.com/en-us/bingmaps/rest-services/routes/

Milence. (2023). Charging the future of road transport. Retrieved May 13, 2023, from https://milence.com/

Networkx. (2023). NetworkX 3.1 documentation (DiGraph.to_undirected). Retrieved April 23, 2023, from https://networkx.org/documentation/stable/reference/classes/generated/networkx.DiGraph.to_undirected.html

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [Publisher: American Physical Society]. *Physical Review E*, *64*(1), 016132. https://doi.org/10.1103/PhysRevE.64.016132

Newman, M. E. J. (2003). The Structure and Function of Complex Networks [Publisher: Society for Industrial and Applied Mathematics]. *SIAM Review*, *45*(2), 167–256. Retrieved April 13, 2023, from https://www.jstor.org/stable/25054401

Newman, M. E. J. (2006). Modularity and community structure in networks [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks [arXiv:cond-mat/0308217]. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nicholas, M. A., Handy, S. L., & Sperling, D. (2004). Using Geographic Information Systems to Evaluate Siting and Networks of Hydrogen Stations [Publisher: SAGE Publications Inc]. *Transportation Research Record*, *1880*(1), 126–134. https://doi.org/10.3141/1880-15

Nicholas, M. A., & Ogden, J. M. (2006). Detailed Analysis of Urban Station Siting for California Hydrogen Highway Network.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. https://doi.org/10.1016/j.socnet.2010.03.006

Oubaalla, W., & Benhlima, L. (2018). An Overview of Community Detection Methods in Transportation Networks. *Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications*, 1–6. https://doi.org/10.1145/3289402.3289546

Pareek, S., Sujil, A., Ratra, S., & Kumar, R. (2020). Electric Vehicle Charging Station Challenges and Opportunities: A Future Perspective. *2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*, 1–6. https://doi.org/10.1109/ICONC345789.2020.9117473

Qiao, Y., Huang, K., Jeub, J., Qian, J., & Song, Y. (2018). Deploying Electric Vehicle Charging Stations Considering Time Cost and Existing Infrastructure [Number: 9 Publisher: Multidisciplinary Digital Publishing Institute]. *Energies*, *11*(9), 2436. https://doi.org/10.3390/en11092436

Ren, S., Choi, T.-M., Lee, K.-M., & Lin, L. (2020). Intelligent service capacity allocation for cross-border-E-commerce related third-party-forwarding logistics operations: A deep

learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *134*, 101834. https://doi.org/10.1016/j.tre.2019.101834

Rob Story. (2013). Folium — Folium 0.14.0 documentation. Retrieved May 8, 2023, from https://python-visualization.github.io/folium/

Saramaki, J., Kivela, M., Onnela, J.-P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks [arXiv:cond-mat/0608670]. *Physical Review E*, *75*(2), 027105. https://doi.org/10.1103/PhysRevE.75.027105

Schlüter, M., Lickert, H., Schweitzer, K., Bilge, P., Briese, C., Dietrich, F., & Krüger, J. (2021). AI-enhanced Identification, Inspection and Sorting for Reverse Logistics in Remanufacturing. *Procedia CIRP*, *98*, 300–305. https://doi.org/10.1016/j.procir.2021.01.107

scikit-learn. (2023). Haversine distances. Retrieved April 6, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.haversine_distances.html

Shahraki, N., Cai, H., Turkay, M., & Xu, M. (2015). Optimal locations of electric public charging stations using real world vehicle travel patterns. *Transportation Research Part D: Transport and Environment*, *41*, 165–176. https://doi.org/10.1016/j.trd.2015.09.011

Sharpe, B., & Basma, H. (2022). A meta-study of purchase costs for zero-emission trucks.

Singh, A., Wiktorsson, M., & Hauge, J. B. (2021). Trends In Machine Learning To Solve Problems In Logistics. *Procedia CIRP*, *103*, 67–72. https://doi.org/10.1016/j.procir.2021.10.010

Statista. (2022). eMobility - In-depth Market Insights & Data Analysis. Retrieved February 4, 2023, from https://www.statista.com/study/49240/emobility---market-insights-and-data-analysis/

Sumalee, A., Uchida, K., & Lam, W. H. K. (2011). Stochastic multi-modal transport network under demand uncertainties and adverse weather condition. *Transportation Research Part C: Emerging Technologies*, *19*(2), 338–350. https://doi.org/10.1016/j.trc.2010.05.018

Tableau. (2023a). Amazon Athena. Retrieved April 30, 2023, from https://help.tableau.com/current/pro/desktop/en-us/examples_amazonathena.htm

Tableau. (2023b). Shelves and Cards Reference. Retrieved April 8, 2023, from https://help.tableau.com/current/pro/desktop/en-us/buildmanual_shelves.htm

Tore Opsahl. (2011). Node Centrality in Weighted Networks. Retrieved May 13, 2023, from https://toreopsahl.com/tnet/weighted-networks/node-centrality/

Tuncel, E., Zeid, A., & Kamarthi, S. (2014). Solving large scale disassembly line balancing problem with uncertainty using reinforcement learning. *Journal of Intelligent Manufacturing*, *25*(4), 647–659. https://doi.org/10.1007/s10845-012-0711-0

Upchurch, C., & Kuby, M. (2010). Comparing the p-median and flow-refueling models for locating alternative-fuel stations. *Journal of Transport Geography*, *18*(6), 750–758. https://doi.org/10.1016/j.jtrangeo.2010.06.015

Upchurch, C., Kuby, M., & Lim, S. (2009). A Model for Location of Capacitated Alternative-Fuel Stations [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.2009.00744.x]. *Geographical Analysis*, *41*(1), 85–106. https://doi.org/10.1111/j.1538-4632.2009.00744.x

U.S. Department of energy. (2023a). Alternative Fuels Data Center: Electric Vehicles. Retrieved April 29, 2023, from https://afdc.energy.gov/vehicles/electric.html

U.S. Department of energy. (2023b). Timeline: History of the Electric Car. Retrieved May 13, 2023, from https://www.energy.gov/timeline-history-electric-car

Volvo. (2022). Heavy-duty electric truck market volume in Europe in 2021, by selected countries [Graph]. In Statista. Retrieved April 29, 2023, from https://www-statista-com.esc-web.lib.cbs.dk/ statistics/1313051/europe-heavy-electric-truck-market-volume-by-country/. Retrieved February 4, 2023, from https://www.statista.com/statistics/1313051/europe-heavy-electric-truck-market-volume-by-country/

Wandelt, S., Shi, X., & Sun, X. (2021). Estimation and improvement of transportation network robustness by exploiting communities. *Reliability Engineering & System Safety*, *206*, 107307. https://doi.org/10.1016/j.ress.2020.107307

Woodall, D. R. (2008). The average degree of an edge-chromatic critical graph. *Discrete Mathematics*, *308*(5-6), 803–819. https://doi.org/10.1016/j.disc.2007.07.048

Woschank, M., Rauch, E., & Zsifkovits, H. (2020). A Review of Further Directions for Artificial Intelligence, Machine Learning, and Deep Learning in Smart Logistics. *Sustainability*, *12*(9), 3760. https://doi.org/10.3390/su12093760

Xi, X., Sioshansi, R., & Marano, V. (2013). Simulation–optimization model for location of a public electric vehicle charging infrastructure. *Transportation Research Part D: Transport and Environment*, *22*, 60–69. https://doi.org/10.1016/j.trd.2013.02.014

Xiang, Y., Liu, J., Li, R., Li, F., Gu, C., & Tang, S. (2016). Economic planning of electric vehicle charging stations considering traffic constraints and load profile templates. *Applied Energy*, *178*, 647–659. https://doi.org/10.1016/j.apenergy.2016.06.021

Xu, X., Shen, Y., (Amanda) Chen, W., Gong, Y., & Wang, H. (2021). Data-driven decision and analytics of collection and delivery point location problems for online retailers. *Omega*, *100*, 102280. https://doi.org/10.1016/j.omega.2020.102280

Yan, X., Duan, C., Chen, X., & Duan, Z. (2014). Planning of Electric Vehicle charging station based on hierarchic genetic algorithm. *2014 IEEE Conference and Expo Transportation Electrification Asia-Pacific (ITEC Asia-Pacific)*, 1–5. https://doi.org/10.1109/ITEC-AP.2014.6941087

Yu, Q., Li, W., Yang, D., & Zhang, H. (2020). Mobile Phone Data in Urban Commuting: A Network Community Detection-Based Framework to Unveil the Spatial Structure of Commuting Demand [Publisher: Hindawi]. *Journal of Advanced Transportation*, *2020*, e8835981. https://doi.org/10.1155/2020/8835981

Zarbakhshnia, N., Kannan, D., Kiani Mavi, R., & Soleimani, H. (2020). A novel sustainable multi-objective optimization model for forward and reverse logistics system under demand uncertainty. *Annals of Operations Research*, *295*(2), 843–880. https://doi.org/10.1007/s10479-020-03744-z

Zeb, M. Z., Imran, K., Khattak, A., Janjua, A. K., Pal, A., Nadeem, M., Zhang, J., & Khan, S. (2020). Optimal Placement of Electric Vehicle Charging Stations in the Active Distribution Network. *IEEE Access*, *8*, 68124–68134. https://doi.org/10.1109/ACCESS.2020.2984127

I Literature Review Summary

Literature Review Overview		
Artificial Intelligence in Logistics and Supply Chain management		
Author(s) and Year	Objective	Used Model(s)
Sumalee et al. (2011)	handling adverse weather conditions uncertainties in vehicle routing problem	multi-modal transport network assignment model
Tuncel et al. (2014)	reducing uncertainty in disassembly line balancing problem (DLBP)	Monte-Carlo based reinforcement learning
Feng and Timmermans (2015)	identifying activity types based on GPS traces	Bayesian belief network, decision tree, random forest
Abdirassilov and Sładkowski (2018)	container flow short-term prediction	Artificial Neural Network (ANN)
Göçmen and Erol (2019)	packing first, routing second problem optimization	k-means, genetic algorithm
Woschank et al. (2020)	review of AI, ML and DL applications in smart logistics	/
Zarbakhshnia et al. (2020)	solution proposition for the forward and reverse logistics network problem	genetic algorithm
Ren et al. (2020)	allocation of capacity in cross-border E-commerce 3PFL operations	Seq2Seq based CNN-LSTM
Bricher and Müller (2020)	process automation in container logistics	DNN
Singh et al. (2021)	review of ML trend associated to logistics issue	/
Albadrani et al. (2021)	demand forecasting in inbound logistics	K-nearest neighbors (KNN), Random Forests, Support Vector Machine (SVM)
Giuffrida et al. (2022)	review of optimization and ML application in last-mile logistics	/
Coverage Analysis		
Author(s) and Year	Objective	Used Model(s)
Xi et al. (2013)	EV charging station positioning	3-step modeling approach
Yan et al. (2014)	EV charging station planning	Hierarchic Genetic Algorithm
Shahraki et al. (2015)	public EV charging station positioning	GAMS optimization model
Donateo et al. (2015)	analysis of CO2 and air pollutants emission from EV	statistical analysis
Gong et al. (2016)	EV fast-charging station planning	abstract-map-based multi-layer optimization model
Xiang et al. (2016)	EV charging station economic planning	multi-objective optimization model
Andrenacci et al. (2016)	EV charging station deployment from a demand perspective	optimization model
Davidov and Pantoš (2017)	EV charging station planning from a charging reliability and QoS perspective	optimization model
Micari et al. (2017)	EV charging infrastructure planning	tailored algorithm
Luo et al. (2017)	EV charging station positioning	nested logit model
Qiao et al. (2018)	EV charging station deployment from a cost perspective	optimization algorithm
Hosseini and Sarder (2019)	EV charging station positioning	Bayesian Network model
Zeb et al. (2020)	EV charging station positioning	particle swarm optimization (PSO)

Graph Theory Applications in Transportation Networks		
Author(s) and Year	Objective	Used Model(s)
Hakimi (1964)	presentation of theory related to centre and medians of a graph	/
John Hodgson et al. (1996)	application of flow-capturing location-allocation model to Edmonton, Canada	FCLM
Nicholas et al. (2004)	hydrogen station positioning	p-median model
Kuby and Lim (2005)	application of FCLM for alternative fuel vehicles	FCLM with mixed-integer programming formulation
Nicholas and Ogden (2006)	hydrogen station positioning	tailored algorithm
Kuby and Lim (2007)	application of FCLM for alternative fuel vehicles with sites on arcs	FCLM, Added-Node Dispersion Problem (ANDP)
Z. Lin et al. (2008)	application of fuel-travel-back approach for hydrogen vehicles	p-median fuel-travel-back model
Upchurch et al. (2009)	application of capacitated FCLM	capacitated FCLM model
Upchurch and Kuby (2010)	comparison of the p-median and FCLM	p-median and FCLM
Community Detection Algorithms in Transportation Networks		
Author(s) and Year	Objective	Used Model(s)
Majima et al. (2014)	generating public transport networks routes	overlapping community detection algorithm /
L. Lin et al. (2014)	traffic accident analysis through clustering	modularity optimization method
Mesa-Arango and Ukkusuri (2015)	demand clustering	modularity maximization method
Li and Zhang (2016)	understanding of urban structure of transportation systems	COMBO method
Oubaalla and Benhlima (2018)	overview of Community Detection methods in transportation networks	/
Beckers et al. (2018)	logistic clustering identification	Louvain method
Beckers et al. (2019)	understanding the hierarchical structure of the logistics Belgian network	Louvain method
Yu et al. (2020)	understanding spatial structure of commuting demand through mobile phone data	modularity optimization method
Badiee et al. (2020)	identifying travellers collaboration networks	tailored community detection algorithm
Wandelt et al. (2021)	increasing transportation network robustness	Community Dismantling, Community Dismantling Edges models

Table 20: Literature Review Overview

II Dataframes Variables Overview

df_final		
Variable Name	Type	Description
BookingId	int64	Unique ID for every delivery; it is possible that one BookingId consists out of multiple legs
SubBookingName	object	Character identifying leg of a delivery (e.g. A, B, C, ...)
SubBookingLegId	int64	Unique ID for every leg of the delivery
StartLegLocationId	int64	Unique ID for pick-up location of a delivery
EndLegLocationId	int64	Unique ID for drop-off location of a delivery
FromLocation	object	Entity where delivery gets picked up
ToLocation	object	Entity where delivery is being dropped off
TransportId	float64	Unique ID for every transport
StartRequestedDate	object	in YYYY-MM-DD format; date when delivery was supposed to be delivered
EndRequestedDate	object	in YYYY-MM-DD format; date when delivery was actually being delivered
FromLatitude	float64	Coordinates of pick-up location
FromLongitude	float64	Coordinates of pick-up location
FromCity	object	Origin city of delivery
FromCountry	object	Origin country of delivery
ToLatitude	float64	Coordinates of drop-off location
ToLongitude	float64	Coordinates of drop-off location
ToCity	object	Destination city of the delivery
ToCountry	object	Destination country of the delivery
GrossWeight	float64	Weight of the load (in kilogram)
LoadMetres	float64	Size of the load (in meter)
CubicMetres	float64	Volume of the goods (in m3)
Length	float64	Length of the goods (in meter)
Width	float64	Width of the goods (in meter)
Height	float64	Height of the goods (in meter)
Temperature	float64	Temperature of the goods (in Celsius)
FullLoadIndicator	object	Indicates whether truck was fully loaded
EmptyBookingIndicator	object	Indicates whether the truck was empty, e.g. when relocating
CustomerName	object	Name of the customer the delivery is made for

Table 21: Summary table for df_final variables

df_deliveries		
Variable Name	Type	Description
BookingId	int64	Unique ID for every delivery; it is possible that one BookingId consists out of multiple legs
SubBookingName	object	Character identifying leg of a delivery (e.g. A, B, C, ...)
SubBookingLegId	int64	Unique ID for every leg of the delivery
CustomerName	float64	Name of the customer the delivery is made for
FromLocation	int64	Entity where delivery gets picked up
ToLocation	int64	Entity where delivery is being dropped off
StartRequestedDate	object	in YYYY-MM-DD format; date when delivery was supposed to be delivered
EndRequestedDate	object	in YYYY-MM-DD format; date when delivery was actually being delivered
FromLatitude	float64	Coordinates of pick-up location
FromLongitude	float64	Coordinates of pick-up location
FromCity	object	Origin city of delivery
FromCountry	object	Origin country of delivery
ToLatitude	float64	Coordinates of drop-off location
ToLongitude	float64	Coordinates of drop-off location
ToCity	object	Destination city of the delivery
ToCountry	object	Destination country of the delivery
DomesticDelivery	int64	Indicates whether pick-up and drop-off location are in the same country or not
RouteDistance	float64	Distance (in Km) between pick-up and drop-off location, calculated in "driving" mode through Bing API
Temperature	float64	Temperature of the goods (in Celsius)
FrozenLoad	float64	Indicates whether goods have a temperature below 0 or not

Table 22: Summary table for df_deliveries variables

df_routes		
Variable Name	Type	Description
FromLatitude	float64	Coordinates (latitude) of pick-up location
FromLongitude	float64	Coordinates (longitude) of pick-up location
ToLatitude	float64	Coordinates (latitude) of drop-off location
ToLongitude	float64	Coordinates (longitude) of drop-off location
RouteDistance	float64	Distance (in Km) between pick-up and drop-off location, calculated in "driving" mode through Bing API
FromLocation	int64	Entity where delivery gets picked up
ToLocation	int64	Entity where delivery is being dropped off
FromCity	object	Origin city of delivery
ToCity	object	Destination city of the delivery
FromCountry	object	Origin country of delivery
ToCountry	object	Destination country of the delivery
RouteCount	int64	Frequency the given route has been travelled

Table 23: Summary table for df_routes variables

df_locations		
Variable Name	Type	Description
Location	object	Tuple with latitude and longitude coordinates of each pick-up or drop-off location
Latitude	float64	Latitude coordinates of the "Location"
Longitude	float64	Longitude coordinates of the "Location"
StartLocationFrequency	int64	Frequency the "Location" is a drop-off point
EndLocationFrequency	int64	Frequency the "Location" is a pick-up point
LocationFrequency	int64	Frequency the "Location" is either a pick-up or drop-off point
StartPct	float64	StartLocationFrequency expressed as percentage of total starting point frequencies
EndPct	float64	EndLocationFrequency expressed as percentage of total ending point frequencies
TotPct	float64	LocationFrequency expressed as percentage of total frequencies

Table 24: Summary table for df_locations variables