A Machine Learning Approach to Shipment Consolidation

Bas van Andel*

Abstract

This research analyzes the current approach used by a client of DHL LLP for the transportation of shipments from suppliers to production sites. As a result of this analysis, several improvements are proposed that can be used to reduce the costs of transportation in the logistics network. Focus is put on shipment consolidation, rather than on the rerouting of shipments. In this paper, two consolidation methods are introduced. Cluster analysis groups together suppliers that are geographically close and generally ship to the same production sites, and a time-based policy that introduces a maximum waiting time for shipments before they are released. While only marginal improvements are obtained when applying these techniques independently, a combination of the two provides a powerful synergy. A trade-off between the savings and proportion of on-time shipments arises, when a maximum waiting time is introduced. The potential savings, therefore, depends on the company's tolerance for late shipments. While the proposed techniques work in theory, practical and organizational challenges emerge when applying them in the real world.

1 Introduction

1.1 Introduction to DHL Global Forwarding

DHL's business is organized into four different divisions: Post - eCommerce - Parcel, Express, Supply Chain and Global Forwarding (2016 Annual Report, 2017). In 2016, DHL earned a total revenue of more than EUR 57 billion. The Global Forwarding division accounted for almost 24% of this revenue, or about EUR 14 billion. In contrast to DHL's other divisions, it has a very asset-light business model, which is based upon the brokerage of transport services between clients and freight carriers. As a result, the division accounts for less than 10% of DHL's 459,000 employees. Moreover, DHL Global Forwarding is market leader in the air freight business and the second-largest provider of

*Bas van Andel received a bachelor degree in Econometrics & Operations Research at Maastricht University in 2017.
Contact: basvanandel@gmail.com

ocean and overland freight services.

The Lead Logistics Partner (LLP) of DHL Global Forwarding is responsible for instigating and managing change across the entire supply chain of a client. Clients include major businesses and industry leaders. The outsourcing sector is highly competitive, but simultaneously offers a lot of growth opportunities. By introducing lean logistics processes, optimizing logistics networks and bringing continuous improvement and cost reduction, the Lead Logistics Partner of DHL remains to be a front runner in this category (Lead Logistics Partner (LLP), n.d.).

1.2 Problem description

The problem at hand involves the optimization of a logistics network of a client of DHL. The client develops and markets systems, equipment and services for the transport sector. In order to accomplish this, there are numerous suppliers all over the world that ship items to their production sites, via air, ocean and road transportation. DHL's Lead Logistics Partner is currently in charge of managing the supply chain network and is aiming at optimizing this process in a generic manner, in order to reduce the costs of transportation.

In total, there are 825 different suppliers, or origins, from all over the world. Over the course of the past two years, they made around 15,000 shipments to a total of 85 production sites, or destinations. Every shipment consists of one or more transportation requests (TR), that correspond to one or more purchase orders (PO). A purchase order is made up of one or several items, and can be shipped at once or in parts. If an order is shipped at once, it will be assigned a single transportation request. If it is shipped in parts, multiple transportation requests will be assigned to the PO. Several purchase orders can also obtain the same TR in case they have been consolidated at the supplier and shipped together. Moreover, every order has a predefined lead time. That is, the time frame after the order has been made, in which the shipment has to be completed, i.e. delivered at the client's production site. In order to satisfy these lead times, a choice can be made between air, ocean and road transportation. Depending on the lane, there are different schedules for flights and ships that have to be taken into account as well. For the purposes of this research, focus will be put on the optimization of air and ocean transportation, as road transportation only accounts for a small fraction of the total shipments (13%). Furthermore, it only accounts for a tiny proportion of the total costs and, thus, less savings can be achieved for this type of transportation, as will become evident in the cost analysis of the current approach.

1.2.1 Research objective

The aim of this research is to find a generic approach that can be used to reduce the costs of transportation, while satisfying all the lead time constraints of the shipments.

1.2.2 Motivation for improvements

Currently, DHL's client has implemented a straightforward approach towards managing the logistics network, as discussed in Section 3. Optimizing this network is a natural move forward, because it can produce major cost savings for the client and, thus, potential revenue for DHL's Lead Logistics Partner. Moreover, a generic optimization can yield results that can be applied to networks of other clients too. Finally, a more efficient approach minimizes the unnecessary use of vehicles and therefore also contributes to environmental sustainability.

1.3 Related literature

Since the problem at hand concerns a very broad topic and, thus, comes with a variety of solutions, it is difficult to find papers that collectively exhaust all these possibilities. Instead, shipment consolidation has been given the main focus, for reasons that are discussed in Section 3.2. Several academia have researched the field of shipment consolidation. Tan et al. (2013) discuss various clustering techniques that can potentially be used to reinforce shipment consolidation. A distinction is made between partitional and hierarchical clustering. Both techniques will be implemented, and eventually compared. Moreover, Mutlu, Cetinkaya, and Bookbinder (2010) describe three different shipment consolidation policies and subsequently find analytical expressions for their implementation. The consolidation policies include a time-based, quantity-based and time-and-quantity-based approach. A time-based policy releases shipments after waiting for a certain maximum amount of time, whereas a quantity-based policy only releases a shipment after a predefined target load has been accumulated through consolidation. The time-and-quantity based policy combines the two and, hence, releases shipments whenever one of these two targets has been met. Another approach to shipment consolidation is bin-packing (Deng, 2013). As described in Section 2.1.2, ocean shipments can be transported using Full Container Load, which implies the entire container is rented for transportation. Leaving empty space in such containers results into larger expenses per kilogram or cubic meter. Therefore, it is important to minimize this empty space, which can be done by bin-packing. This paper, however, does not analyze the implications of this technique any further. Finally, the problem also shares similarities with a vehicle routing problem with time windows. Desrochers, Desrosiers, and Solomon (1992) describe an improved algorithm to solve such a problem. In this research, however, no focus will be put on this area.

1.4 Outline

This research focuses on analyzing the current approach of DHL's client and proposes improvement methods that can help reducing the costs of transportation. First, the shipment data will be cleaned and the necessary interpolation will be performed. Second, the current approach is examined by analyzing the costs of the different transportation methods. Third, several cost reducing improvements upon this approach will be introduced and investigated thoroughly. Finally, the results, as well as the practical challenges of these techniques, will be discussed.

2 Data pre-processing

By examining data in the pre-processing phase, possible flaws and errors can be found and corrected, before moving on to the cost calculation and improvements. It is therefore a crucial step in the data analysis process. First, the structure of the costs will be explained as well as the corresponding assumptions. Second, some shipments have been assigned an incorrect volume. The volume of these shipments will be corrected by investigating shipments with an unrealistic relation between the weight and volume. Finally, the lead times of shipments will be analyzed and improved accordingly.

2.1 Structure of the costs

The costs of different methods of transportation consist of three parts: pre-carriage, mainfreight and destination-carriage costs. Calculating the costs of the three parts for air and ocean transportation, happens in a different manner and hence a distinction will be made below.

2.1.1 Air transportation costs

Pre-carriage costs include pickup costs (per kg), a fuel surcharge for the pickup, handling and security, and costs for export customs clearance. These costs depend on the airport of departure. Mainfreight costs include air freight (per kg using bulk discount rates), a fuel surcharge and an insurance. In case that volume in cubic meters multiplied by 166.7 is larger than the number of kilograms, this number (i.e. the taxable weight) is used instead of the real weight when calculating the freight costs. The destination-carriage costs consist of import customs clearance, handling costs and costs for delivery to the production site, with a corresponding fuel surcharge. The cost structure is summarized in Table 1.

Pre-carriage	**Mainfreight**	**Destination-carriage**
Pickup (per kg)	Air freight (per taxable kg, bulk rates)	Delivery (per kg)
Handling and security (per kg)	Fuel surcharge (per taxable kg)	Handling and security (per kg)
Export customs (fixed)	Insurance (per kg)	Import customs (fixed)

Table 1: Cost structure of air transportation

2.1.2 Ocean transportation costs

For ocean transportation, the cost structure is very similar, apart from the fact that a distinction has to be made between Full Container Load (FCL) and Less than Container Load (LCL). In case a shipment is sent using FCL, the items will be sent using one or several containers. The containers exist in two sizes: the smaller 20' container that has a capacity of 33.0 cubic meter (cbm) or 22,100 kg and the larger 40' container that has a capacity of 67.3 cbm or 27,397 kg. The number of small and large containers have to be computed first, before the costs can be calculated. Shipments that are sent using LCL only pay for the weight and volume it uses. Pre-carriage costs include pickup costs and a corresponding fuel surcharge, handling and security costs, and export customs clearance. For FCL, the costs are calculated per container and for LCL, the pickup costs are per kg (with bulk rate discounts) and the handling and security costs per cbm. Mainfreight costs consist of freight costs and costs for BAF (Bunker Adjustment Factor). Again, a distinction has been made between FCL and LCL. That is, ocean shipments transported using FCL will be charged per container, while LCL shipments will be charged based on the actual volume being shipped. Furthermore, for the calculation of the mainfreight of LCL, 1000 kg is equal to 1 cbm and, thus, the largest of the two should be used when calculating the costs. The destination-carriage costs include import customs clearance, handling costs and delivery costs, with a corresponding fuel surcharge. The costs for FCL are based on the number and sizes of the containers used, and the costs for LCL are per cbm for handling and per kg for the delivery. The cost structure, including the distinction between FCL and LCL, is summarized in Table 2.

Pre-carriage	**Mainfreight**	**Destination-carriage**
Pickup (LCL: per kg, FCL: per cont.)	Ocean freight (LCL: per cbm, FCL: per cont.)	Delivery (LCL: per kg, FCL: per cont.)
Handling and security (LCL: per cbm, FCL: per cont.)	BAF (LCL: per cbm, FCL: per cont.)	Handling and security (LCL: per cbm, FCL: per cont.)
Export customs (fixed)		Import customs (fixed)

Table 2: Cost structure of ocean transportation

2.1.3 Road transportation costs

Although, no improvements will be discussed regarding road transportation, it is important that one backs this choice by numbers. Since there are no predefined costs from different carriers for road transportation, the costs are calculated in a different manner, using several heuristics. First, the distance between the origin and destination is calculated. Given the latitude and longitude of the two locations, the distance can be calculated as follows (Dutch, 2016). Convert the latitude and longitude from degrees to radius and solve:

$$distance = \arccos\left(\sin lat_1 * \sin lat_2 + \cos lat_1 * \cos lat_2 * \cos\left(lon_1 - lon_2\right)\right) * 180 * \pi * 60 * 1.1515 * 1.609344$$

In this equation, lat_1 and lon_1 correspond to the latitude and longitude of the origin and lat_2 and lon_2 correspond to the latitude and longitude of the destination. Second, the number of trucks is determined that is necessary to transport the shipment. If the shipment fits into one truck, it is considered Less than Truck Load (LTL) and a percentage of the capacity will be used in the calculation. Only trucks with a capacity of 67.3 cbm and 24,000 kg are considered. Finally, a price per kilometer per truck is chosen and the costs are calculated:

$$costs = distance * price\_per\_kilometer * percentage\_of\_truck\_used * (1 + factor)$$

In the calculation of the costs, a price of \$1.25 per kilometer will be used. The *factor* is used to adjust for the fact that in the real world, roads are no straight lines from an origin to a destination. The calculated distance does not take this into account, and, thus, a factor is introduced to make the costs more realistic. A factor of 20% will be used in the final calculation. Finally, as some shipments only use a tiny fraction of the truck, a minimum cost of \$50 per shipment is used in the cost calculation.

2.1.4 Carriers

There are several carriers with different cost structures that can be used for the transportation. Each origin-destination pair, on a country-to-country level, has been assigned a preferred carrier, which will be used for all transportation on this lane. Both air and ocean transportation have different preferred carriers, and no deviations from these shippers will be made.

2.2 Assumptions about the costs

The first assumption that will be made when calculating the costs addresses the pickup costs for both air and ocean transportation. Depending on where the supplier is located, pickup costs to the same airport or port can vary slightly. As the difference between these costs is negligible, the following assumption will be made: the pickup costs for every supplier to their closest airport or port are the same. For the same reason, a similar assumption will be made for the delivery costs of the destination-carriage: no distinction will be made between delivery costs to production sites or destinations that have the same closest airport or port. In case of FCL ocean transportation, additional costs can occur if the recipient of a shipment fails to collect the items from the container within a certain time frame. For the purpose of this research, these costs have been omitted.

Moreover, as the costs of transportation are all calculated in different currencies, a conversion to United States Dollars will be made at the end of the computation. This is done using a fixed exchange rate and, thus, ignores foreign exchange rate effects. Finally, the responsibility of the total costs of a shipment is contingent on international commercial terms (incoterms). Combining shipments with different incoterms can lead to complicated cost structures and therefore has to be taken into account. For the problem at hand, however, no differences in incoterms will be assumed.

2.3 Incorrect volumes of shipments

In the data analysis process, it became clear that for some shipments the ratio between the weight and volume was very unrealistic. For example, one shipment had a weight of 44 kilograms and a volume of almost 64,000 cubic meters. Clearly, an error occurred when inserting the data. According to the client, these mistakes, due to human error, are only prevalent in the 2015 shipment data, not in the 2016 data. Therefore, the 2016 data has been used as training data for a linear regression: $volume = \beta_0 + \beta_1 * weight$. This regression, and subsequent estimation, will be done under the assumption that the weight of the shipments has been inserted correctly.

For all the shipments in 2015, the ratio between the weight (kg) and volume (liters), i.e. the density, will be used to determine whether the volume is inaccurate. This is done using the following heuristic: every shipment should have a density (kg/liters) between 0.001 and 100. These values correspond to the density range of the correct 2016 shipments and, thus, deal with most of the errors. All the shipments that do not satisfy this, around 500 shipments, will be placed in the test set. Before the right volumes of these shipments can be estimated, a separation has to be made between the different methods of transportation. Subsequently, the volumes of the shipments in the test set will be estimated using the regression results. In the remainder of this paper, the volumes that result from the regressions will be considered the true volumes of these shipments and, thus, used for further research.

Air transportation regression The training set of air shipments consists of 3,128 observations. A simple linear regression of the volume on the weight yields the results in Figure 1 ($R^2 = 0.5932$).

	Coefficient	Standard Error	P-value
Constant	0.2399	0.03716	$1.25 * 10^{-10}$
Weight (kg)	0.003016	$4.469 * 10^{-5}$	$< 2 * 10^{-16}$

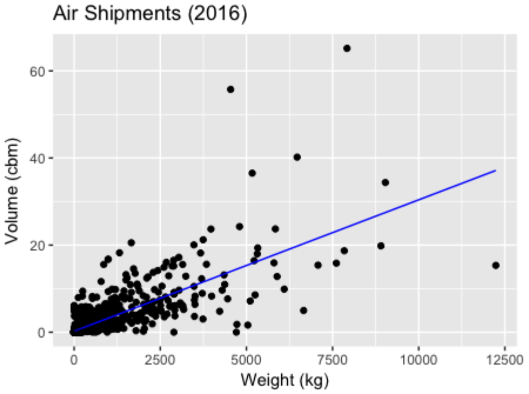


Figure 1: Air transportation regression

Ocean transportation regression The training set of ocean shipments consists of 3,344 observations. A similar regression yields a $R^2 = 0.6294$ and the results in Figure 2.

	Coefficient	Standard Error	P-value
Constant	3.676	0.2473	$< 2 * 10^{-16}$
Weight (kg)	0.002083	$2.768 * 10^{-5}$	$< 2 * 10^{-16}$

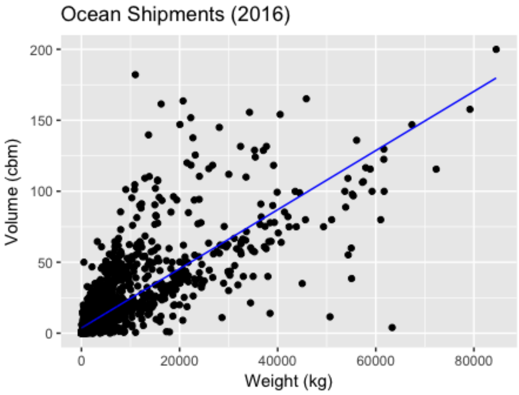


Figure 2: Ocean transportation regression

Road transportation regression Finally, the same regression will be applied to the 1,070 road shipments. With $R^2 = 0.5654$, the results in Figure 3 are obtained:

	Coefficient	Standard Error	P-value
Constant	0.9197	0.1836	$6.39 * 10^{-7}$
Weight (kg)	0.002935	$7.946 * 10^{-5}$	$< 2 * 10^{-16}$

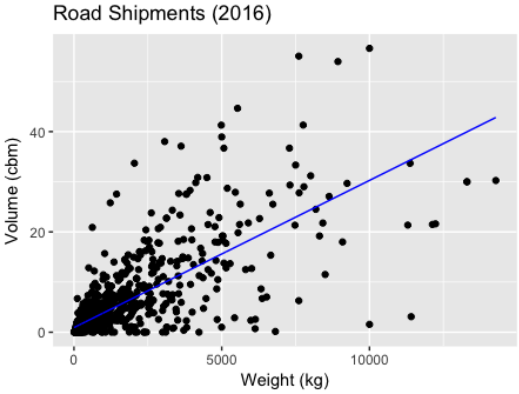


Figure 3: Road transportation regression

2.4 Lead times

Every purchase order has a predefined lead time. That is, the order has to be delivered at the production site within a certain time frame. These lead times form the main constraint of the optimization problem, as it restrains suppliers from combining all the orders that have to be sent to the same destination. In order to find the lead time of a shipment, the lead times of the underlying purchase orders need to be found first. Subsequently, the minimum of these times is taken and assigned to the shipment as its lead time.

2.4.1 Lead times analysis

When calculating the lead times of the shipments, several difficulties occurred. First, due to missing data for the lead times of purchase orders or missing orders for the shipment, it was impossible to assign every shipment a lead time. As a result, some shipments have been assigned a lead time of zero, which corresponds to no lead time at all. The next section will elaborate on these shipments a little more. Second, some shipments were assigned an unrealistic lead time of a few hours, while others had a very large lead time of more than a year. Figure 4 shows the distribution of the lead times per month for all methods of transportation. Shipments without a lead time (i.e. zero lead time) have been omitted.

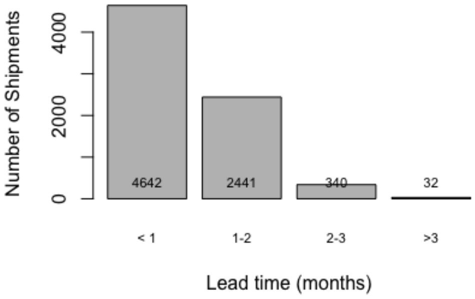


Figure 4: Lead times per month

Most of the shipments have a lead time of less than one month. In order to examine these lead times even more, the number of shipments for lead times with daily intervals have been plot in Figure 5.

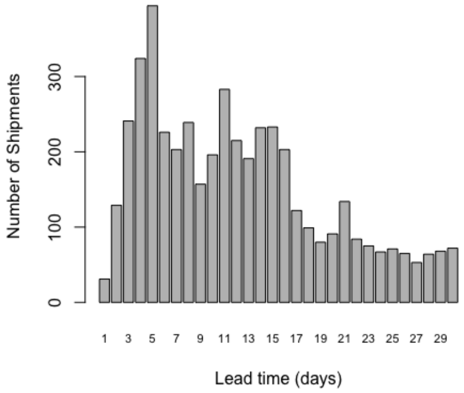


Figure 5: Lead times per day

When examining the shipments for both air and sea transportation independently, the legitimacy of some lead times is questionable. Figure 6 shows that for air and ocean transportation, respectively, extraordinary low lead times are found.

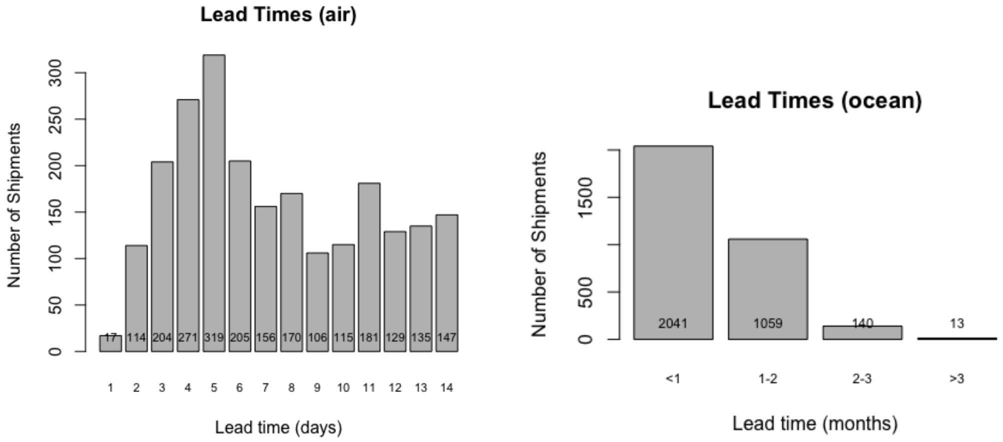


Figure 6: Lead times for air and ocean transportation separately

As the different stages in the shipment process (pre-carriage, mainfreight and destination-carriage) each take a few days to progress, most of the lead times above are very unrealistic. Hence, it is necessary to take these shipments into account too, when interpolating the missing lead times.

2.4.2 Interpolation

Before interpolating the missing lead times, a choice has to be made regarding the extraordinary low lead times. For both air and ocean transportation, a minimum lead time has to be chosen. All shipments with a lead time below this minimum are considered wrong and will be assigned a new time. For the purpose of this research, a minimum lead time for air transportation of 7 days, and 30 days for ocean transportation will be chosen. After a lower bound has been established, all shipments having a lead time below this bound will be assigned a new time using a simple interpolation technique. The shipment data is split into origin-destination pairs, on a country-to-country level. Subsequently, the average lead time of each pair will be found, excluding lead times below the minimum and above some maximum value. In this study, lead times of three months or higher are left out of the calculation, in order to adjust for outliers. Next, every shipment without a lead time or with a lead time that is less than the lower bound will be assigned a lead time that is equal to the average for that particular lane. After repeating this process for every lane, all the shipments will have a realistic lead time, i.e. a time above the lower bound.

3 Current approach

Currently, a straightforward approach is implemented and used for the transportation of shipments. First, one or more purchase orders are collected at a certain supplier. Second, a choice has to be made between air, ocean or road transportation, depending on the lead times and destinations of the orders. Third, purchase orders with similar lead times that have to be sent to the same destination will be combined and, depending on the method of transportation, shipped from the closest airport or port. The shipments will be sent to the airport or port that is the closest to the production site. Finally, the shipments will be delivered at the destination.

3.1 Cost calculation

Due to the absence of details on the costs of certain lanes, it was only feasible to find the costs of 13,527 out of 14,899 shipments. Of those 13,527 shipments, 5,834 (43%) were shipped by air transportation, 5,922 (44%) by ocean transportation and the remaining 1,771 (13%) by road transportation. The results have been summarized in Table 3 and Table 4.

Method	Pre-carriage	Mainfreight	Destination-carriage	Total costs
Air	$1,049,077 (12%)	$7,059,468 (82%)	$514,805 (6%)	$8,623,351 (39%)
Ocean	$4,278,224 (34%)	$4,710,199 (37%)	$3,597,689 (29%)	$12,586,114 (56%)
Road	-	$1,100,411 (100%)	-	$1,100,411 (5%)
Total	$5,327,302 (24%)	$12,870,078 (58%)	$4,112,494 (18%)	$22,309,875 (100%)

Table 3: Current transportation costs

Method	Shipments	Sum of weight (kg)	Sum of volume (cbm)
Air	5834 (43%)	1,961,216 (6%)	27,503 (16%)
Ocean	5922 (44%)	26,896,490 (85%)	128,571 (75%)
Road	1171 (13%)	2,771,791 (9%)	16,036 (9%)
Total	13527 (100%)	31,629,497 (100%)	172,110 (100%)

Table 4: Shipment data

In Table 3, the percentages for the pre-carriage, mainfreight and destination-carriage correspond to the total costs of the different methods, whereas the percentages for total costs correspond to the last row with the aggregated results.

3.1.1 Air transportation

The majority of the costs come from the mainfreight. The reason for this is that in the mainfreight cost calculation, the volumes are most dominant. That is, the volume in cubic meters multiplied by 166.7 is usually larger than the kilogram amount and, thus, this number is used when calculating the mainfreight costs. The mean of the total costs for air transportation is at $1,478, while the median is only $375. The minimum cost is $66 and the maximum $1,307,800. This implies that most of the air shipments are relative cheap, but there are a few large outliers. It also highlights the fact that there are probably still some inaccurate volumes in the data, even after an attempt to correct these. Although $1,307,800 seems an unlikely amount for a shipment, a choice regarding the density has been made in Section 2.3 and will hence be respected.

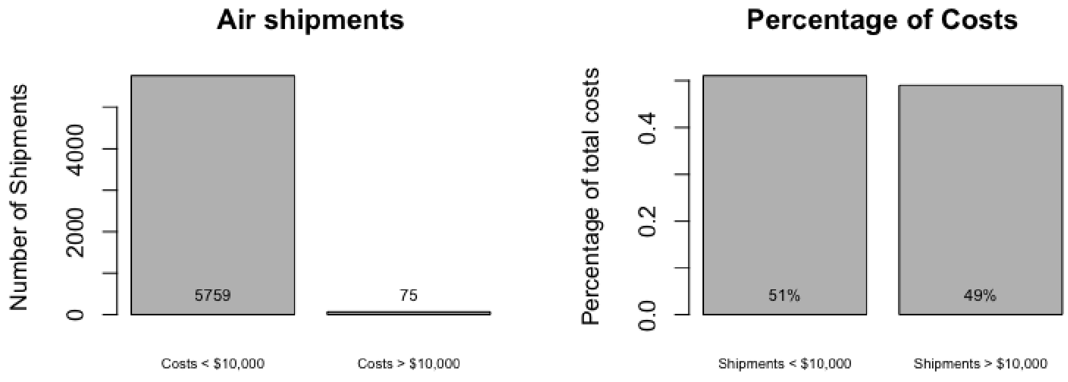


Figure 7: Distribution of air transportation costs

Figure 7 shows that there are 5,759 shipments with a cost of less than $10,000, or 98.7% of the shipments, that account for only 51% of the costs. The 75 shipments with a cost that is larger than $10,000 account for 49% of the total costs of air transportation. Figure 8 highlights that by breaking down the shipments under $10,000 even further, the majority of the shipments has a cost of less than $1,000.

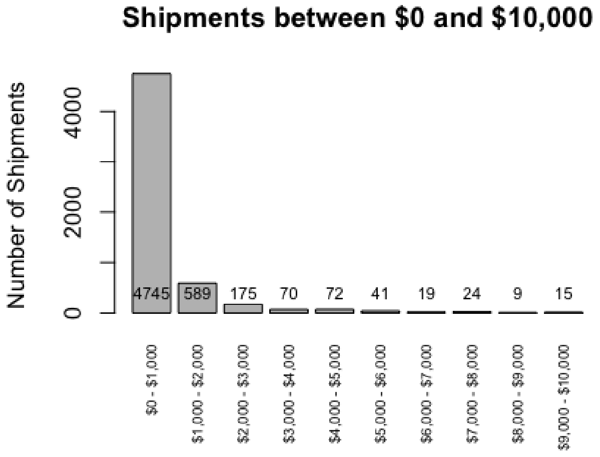


Figure 8: Distribution of small air shipments

An important takeaway from this analysis is that a lot of the savings potential for air transportation can be achieved by consolidation the smaller shipments (i.e. shipments having costs less than $1,000).

3.1.2 Ocean transportation

In contrast to air transportation, the costs of ocean transportation are not mainly dependent on the mainfreight costs. The reason for this is that the volume in cubic meters is used in the calculation of the mainfreight costs, unless the number of kilograms divided by 1,000 is larger than this value. The mean of the total costs for ocean transportation is \$2,125 and the median is \$1,096. The cost per shipment ranges from \$179 to \$617,052. Figure 9 shows the distribution of the costs.

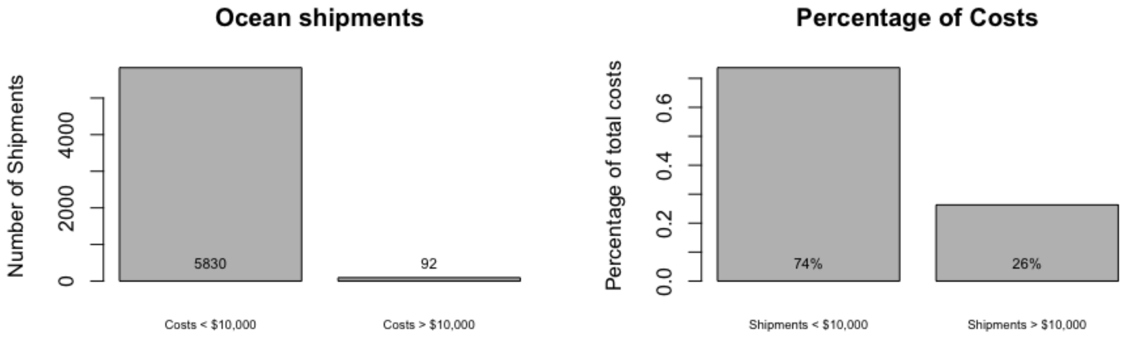


Figure 9: Distribution of ocean transportation costs

The costs of the different ocean shipments are more scattered, in comparison to the costs of air shipments. Therefore, the shipments under \$10,000 still account for 74% of the total costs. As illustrated in Figure 10, the majority of the shipments has a cost of less than \$1,000, but this is only 46% of the shipments, compared to 82% for air transportation.

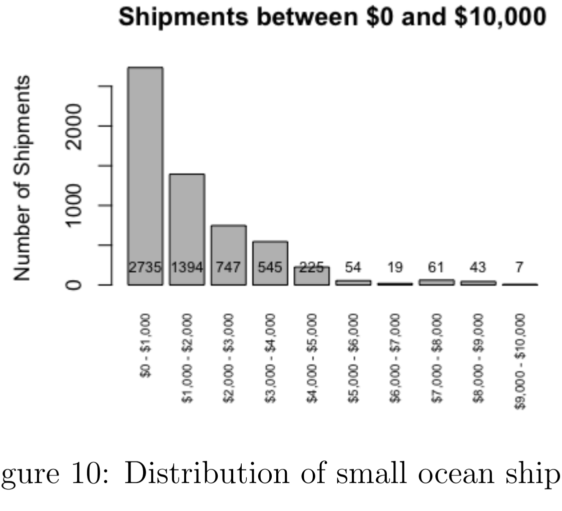


Figure 10: Distribution of small ocean shipments

3.1.3 Road transportation

The distribution of the costs for road transportation looks very similar to the air transportation costs. Most of the shipments are relatively small (i.e. under $1,000) and only a few are large. The 20 shipments that cost more than $10,000 account for around 47% of the total costs for road transportation. The average cost is $621.35 and the median cost is $81.68, ranging from a minimum of $50 to a maximum of $80,292. Figure 11 and Figure 12 show the distribution of the costs.

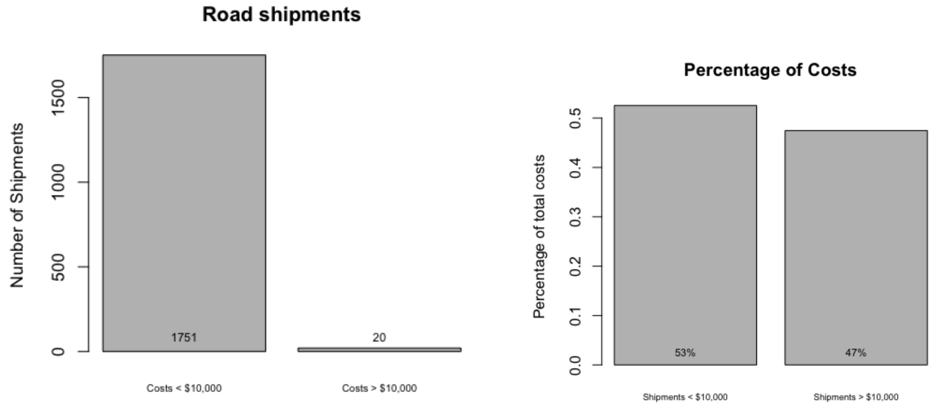


Figure 11: Distribution of road transportation costs

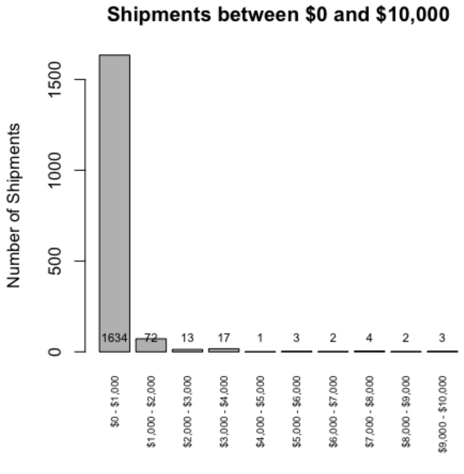


Figure 12: Distribution of small road shipments

Even though potential cost savings might be achieved here, no focus will be put on the optimization of these road transportation costs. This research will focus on the optimization of air and ocean transportation, as these methods account for a total of $21,209,465 or 95% of the costs. They also account for over 91% of the total weight and the total volume.

3.2 Improvements

Using the cost analysis above, several ideas for improvements can be identified. As for both air and ocean transportation, the majority of the shipments have a cost of less than \$1,000, a lot of savings can be achieved by consolidating these shipments. The reason for this is that there are minimum pickup and delivery costs, and fixed fees that have to be paid regardless of the size of the shipment. Moreover, freight costs work with bulk rates, which means the average price per kilogram or cubic meter decreases when the size of the shipment increases. By combining different shipments, these fixed fees are only paid once and pickup, delivery and freight costs might be lower.

Furthermore, there are some very large shipments that account for an enormous amount of the total costs. This is evident for air transportation, as there are 75 shipments that account for about 49% of the total costs. Rerouting these shipments via cheaper airports or ports can provide major cost savings. This paper will, however, solely focus on the consolidation rather than the rerouting of shipments. Section 4 and 5 will elaborate on several techniques that can be used to combine shipments from different origins and on the same lane, respectively.

4 Cluster analysis

The first method to improve upon the current approach involves the consolidation of shipments by combining together origins that are geographically close and that ship to the same destinations. These groups of origins will subsequently be regarded as one origin that combines all the orders that are collected at one of the suppliers in the group. Since the number of origins is much larger than the number of destinations, there is a converging flow of shipments and, thus, the choice has been made to group together the origins instead of the destinations. The method used to do this is clustering and will be examined in this part of the paper. As there are multiple clustering techniques known in the literature, they will be discussed independently and eventually compared.

4.1 Introduction

Cluster analysis is a technique in unsupervised learning that groups together data that share similar characteristics. In contrast to supervised learning, in unsupervised learning, the data have no target attribute and, therefore, intrinsic structures or classes have to be explored (Unsupervised Learning, n.d.). The goal of clustering is that these classes have a high intra-class similarity and a low inter-class similarity. Similarity is often defined by a distance measure such as Euclidean distance. Two different types of clustering techniques will be considered in this paper: partitional clustering and hierarchical clustering. Partitional clustering only deals with unnested clusters, while hierarchical clustering can involve nested clusters that are organized as a tree. Another way in which the two types differ is that hierarchical clustering initially assigns each point in a cluster by itself, whereas partitional clustering maintains a certain set of clusters (Johnson, 2014).

4.2 Data preparation

In order to successfully perform any cluster algorithm, the data has to be prepared appropriately. For the problem at hand, the different data points that have to be clustered are the origins or suppliers. Every origin has latitude and longitude coordinates of its location. However, it is also important that the suppliers generally ship to the same destinations, otherwise no possible synergies can be obtained. With this in mind, the data matrix is constructed as follows. The rows represent the origins and the columns include the latitude, longitude and all the available destinations. The values at latitude and longitude are the coordinates of the different origins and the values for the different destinations represent the number of shipments that the origin sent to that particular destination, over the course of two years. In this way suppliers are clustered together that are both geographically close to each other and that send shipments to the same destinations. The clustering is done under the assumption that suppliers generally ship to a number of destinations, and do not deviate from these production sites too often.

Table 5 shows an example of how the data will be structured for m origins and n destinations. In this specific instance, origin 1 will likely not be grouped together with origin m, even though they have very similar coordinates. The reason for this is that they both ship to very different destinations.

	Latitude	Longitude	Shipments to dest. 1	⋯	Shipments to dest. n
Origin 1	31.3	120.6	0	⋯	300
⋮	⋮	⋮	⋮	⋮	⋮
Origin m	31.2	120.4	200	⋯	0

Table 5: Example of the data matrix for the cluster analysis

Since shipments are sent via air and ocean transportation, a distinction has to be made in the clustering as well. Moreover, the cluster analysis will be done for each continent (Asia Pacific, Europe and Americas) separately, as these suppliers are not located close to each other and might cause disruptions to the results. For the remainder of this section, air transportation of Asian suppliers will be used for illustrative purposes. The latitude and longitude of the origins in Asia are plotted in Figure 13 with China in blue, India in red and the other countries in black.

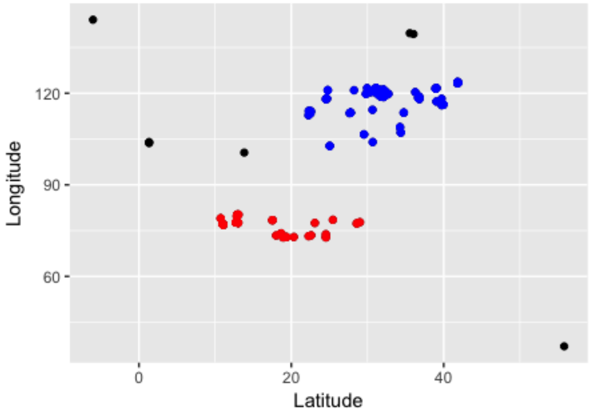


Figure 13: Coordinates of origins in Asia

4.2.1 Feature scaling

Since the different features, i.e. columns of the Table 5, are in different units, it is important to scale accordingly. In order to do this, the data in each column has to be standardized. That is, the mean and standard deviation of each column will be calculated, and subsequently every new value in row i and column j, x'_{ij}, is derived as follows:

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

Where x_{ij} is the current value for row $i = 1, ..., m$ and column $j = 1, ..., n+2$, and μ_j and σ_j are the mean standard deviation of column j, respectively. After applying this feature scaling procedure, all the variables have an equal weight in the cluster analysis. For obvious reasons, however, the origins in each cluster have to be geographically close to each other. Therefore, more weight has to be put on the latitude and longitude. This will be done by multiplying the corresponding features by some number α.

$$x''_{ij} = x'_{ij} * \alpha \tag{2}$$

In this equation, x''_{ij} corresponds to the updated values for $i = 1, ..., m$ and $j = 1, 2$ (the coordinates). The cluster analysis can be done for multiple values of α and eventually compared. For now, $\alpha = 1$ will be assumed.

4.3 Partitional clustering: k-means

K-means is arguably one of the most popular techniques in the partitional clustering literature (Tan et al., 2013). It is a centroid-based algorithm that decomposes the data into a set of K disjoint clusters, for some predefined K. K-means aims to minimize the following objective function:

$$\text{Total Within Sum of Squares} = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} dist(\boldsymbol{c}_i, \boldsymbol{x})^2 \tag{3}$$

Where K is the number of clusters, C_i is the set of data points in cluster i and c_i is the centroid of cluster i, which is defined by the mean of its objects. Every cluster C_i should have a size of at least one and $C_i \cap C_j = \emptyset$ for all $i \neq j$. Let m_i be the number of items in cluster C_i, then:

$$\boldsymbol{c_i} = \frac{1}{m_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x} \tag{4}$$

The distance between $\boldsymbol{c_i}$ and $\boldsymbol{x}$, $dist(\boldsymbol{c_i}, \boldsymbol{x})$, is often calculated as the Euclidean distance. That is, $dist(\boldsymbol{c_i}, \boldsymbol{x}) = \sqrt{\sum_{j=1}^{n} (c_{ij} - x_j)^2}$, in case there are n features. Although there are other distance measures that can be used for k-means, such as cosine and Manhattan distance, this paper will solely focus on minimizing the Euclidean distance (Tan et al., 2013).

Finding the optimal solution to the objective function has been proved to be NP-hard for $K \geq 2$ (Aloise, Deshpande, Hansen, Popat, 2009). The two algorithms discussed to solve this problem will always converge to a solution, but they will not necessarily converge to a global minimum. Convergence is, among other things, contingent on the choice of initial centroids. Section 4.3.2 will elaborate more on this.

4.3.1 K-Means algorithms

Lloyd's algorithm This algorithm is probably the most straightforward version of k-means. The procedure works as follows. After K initial centroids have been chosen, each data point will be assigned to the closest one in terms of Euclidean distance. Subsequently, the centers will be recalculated by taking the mean of all the objects that have been assigned to them. This process of assigning the data points and updating the centers will be repeated until the centroids do not change anymore, i.e. the algorithm has converged. Lloyd's algorithm has a complexity of $O(I * K * m * n)$, where I is the number of iterations required for convergence, K is the number of clusters, m is the number of data points and n the number of features (Tan et al., 2013). Below is a formal definition of the algorithm.

Algorithm 1 K-means (Lloyd)

```
Select $K$ initial centroids.
repeat
    Form $K$ clusters by assigning each point to its closest centroid, using Euclidean distances.
    Recompute the centroid of each cluster by taking the mean of all the data points.
until Centroids do not change.
```

Hartigan and Wong The main difference between Lloyd's algorithm and the algorithm proposed by Harginan and Wong is that for the latter the centroids will be updated every time a single data point is assigned to a different cluster (Hartigan Wong, 1979). This is in contrast to Lloyd's algorithm, which only updates the centroids after all the objects have been assigned to a cluster. Multiple studies have shown that this results in a more efficient algorithm and, hence, this is also the algorithm that will be used to solve the problem at hand. Moreover, the complexity is similar to Lloyd's algorithm and, thus, equal to $O(I * K * m * n)$ (Slonim, Aharoni, Crammer, 2013). For the remainder of this paper, k-means and Hartigan and Wong's algorithm will be used interchangeably.

Algorithm 2 K-means (Hartigan and Wong)

```
Assign all the data points to $K$ random clusters and calculate the respective centroids.
repeat
    Starting from the first object, find the closest centroid and assign the point to that cluster. If the data point has now been assigned to a different cluster, update the centroid of this cluster as well as the centroid of the cluster that this object left.
    Loop through all the points in order to find the new centroids.
until Centroids do not change.
```

4.3.2 Initialization of the centroids

A wrong initialization of the centroids in the first step of k-means, might cause the algorithm to converge to a poor local minimum that is far removed from the global optimum (Algorithms for k-means clustering, 2013). In order to prevent this from happening, certain techniques can be used to ensure a good convergence. A very common technique from the literature is k-means++. The intuition behind the algorithm is that it initializes the centers in such a way that they are as far apart from each other as possible. It adjusts for outliers by choosing the center at random, using probabilities that are proportional to the distance from the data point to the closest center. The k-means++ algorithm is $O(\log K)$-Competitive and is formally defined below, with X being the set of all data points and $D(x)$ the (Euclidean) distance from object x to the closest center (Arthur Vassilvitskii, 2007). Once this initialization has been completed, the k-means algorithm can be executed in a normal fashion.

Algorithm 3 K-means++ initialization

```
Choose one center $c_1$ uniformly at random from $X$.
Set a new center $c_i$ equal to $x \in X$ with probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$
Repeat step 2 until $K$ centers have been found.
```

4.3.3 Determining the number of clusters

Up until now, no assumption has been made regarding the number of clusters in the k-means algorithm. As there is no one right way to determine the number of clusters in a data set, this is a major drawback of the k-means algorithm. There are, however, several techniques that can enhance the decision of how many clusters are optimal for the data set. The outcome of a good cluster analysis yields classes with a high intra-class similarity and a low inter-class similarity, while keeping the number of clusters as low as possible. These similarities can be measured inversely by the total within sum of squares (TWSS) and the between sum of squares (BSS), respectively. The TWSS declines with an increasing number of clusters, while the BSS increases in that case. Hence, there is a trade-off between optimizing the similarities and minimizing the number of clusters. A formal definition of the TWSS is given by Equation (3), and the BSS equals the Total

Sum of Squares (TSS) minus the TWSS:

$$BSS = TSS - TWSS = \sum_{i=1}^{m} dist(\boldsymbol{c}, \boldsymbol{x_i})^2 - \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} dist(\boldsymbol{c_i}, \boldsymbol{x})^2 \tag{5}$$

Where $\boldsymbol{c}$ is the global mean of the data and $\boldsymbol{x_i}$ represents each of the m data points. The process of analyzing the number of clusters is best shown by means of an example. The Fisher's Iris flower data set is one of the best known sets, mainly used for classification. Within the set, there are 150 records of flowers of three different types with four features (Fisher, 1936). After applying Algorithm 2 (k-means) to the data set for $K = 2, ...8$, the results in Figure 14 have been obtained.

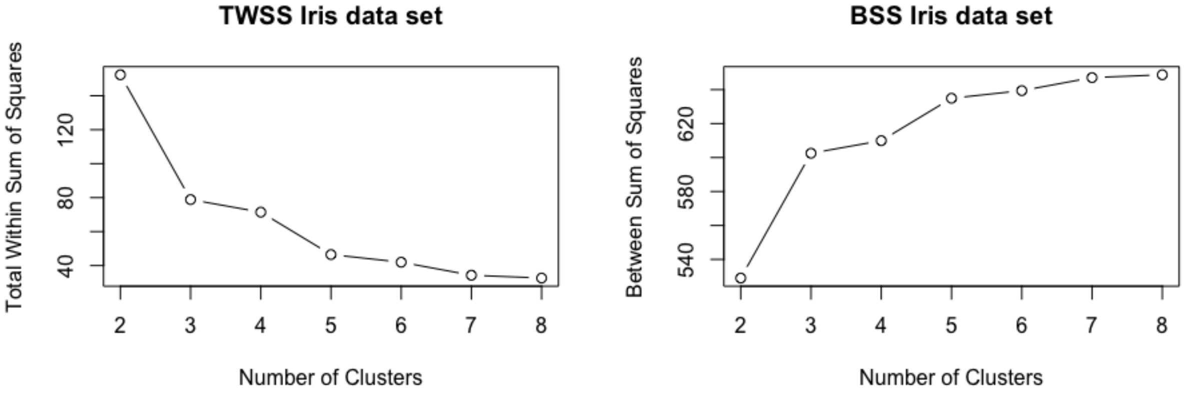


Figure 14: TWSS and BSS for an increasing number of clusters

The graphs above clearly show that the TWSS decreases significantly when moving from two to three clusters, and hence the BSS experiences a steep increase. The decrease in the TWSS from three to four clusters is very small and, therefore, it can be concluded that the 'optimal' K is equal to 3. This procedure is known in the literature as the elbow method (Kodinariya Makwana, 2013). There are also other approaches that can be used to find the number of clusters such as the Knee Point Detection in Bayesian Information Criterion (Zhao, Hautamaki, Franti, 2008). As these methods are outside the scope of this research, only the elbow method will be applied to the problem at hand.

4.4 Hierarchical clustering

Hierarchical clustering is another renowned clustering method. It shares some characteristics with partitional clustering, but also has major differences. There are two approaches for generating a hierarchical cluster analysis: agglomerative and divisive clustering (Tan et al., 2013). Agglomerative clustering can be considered a bottom-up approach in which initially every data point is an individual cluster and, at each step, the two closest clusters are merged. This requires a notion of clustering proximity, which will be discussed in Section 4.4.1. Conversely, divisive clustering is a top-down approach in which there is one cluster initially and, at each step, a cluster is split into two, until only singleton clusters remain. For the remainder of this section, focus will be put on agglomerative clustering. Tan et al. (2013) described the algorithm for this type of clustering as follows:

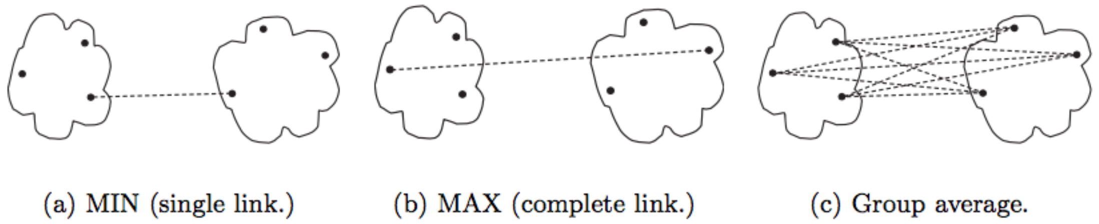
Algorithm 4 Agglomerative hierarchical clustering

```
Compute the distance (proximity) matrix
repeat
    Merge the closest two clusters
    Update the distance matrix to reflect the proximity between the new cluster and the original clusters
until Only singleton clusters remain
```

In case of m data points, the initial distance matrix is an m by m matrix that represents the Euclidean distance between each cluster, or data point. In order to perform the merge and update steps of the algorithm, a proximity measure has to be chosen. Section 4.4.1 will elaborate on different approaches. The time compexity of Algorithm 4 is $O(m^2 \log m)$ (Tan et al., 2013)..

4.4.1 Proximity measures

The main parameter of the agglomerative clustering algorithm is the proximity measure used for merging two clusters. Three different techniques will be discussed in this section: MIN (single linkage), MAX (complete linkage) and the Group Average.



(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Figure 15: Graph-based definitions of cluster proximity (Tan et al., 2013).

Figure 15 visualizes the three different proximity measures. The MIN technique defines the distance between two clusters by the Euclidean distance between the two closest data points. Conversely, the MAX method uses the farthest two points to measure the distance between two clusters. The final approach involves taking the average of all pairwise distances (i.e. the average length of edges). The three proximity measures will be compared in Section 4.5.

4.4.2 Determining the number of clusters with Dendrograms

Even though Algorithm 4 does not have some input parameter K for the number of clusters, it is still necessary to determine the number of clusters that is appropriate for the data, based on the results. A technique that can be used for this is a dendrogram.

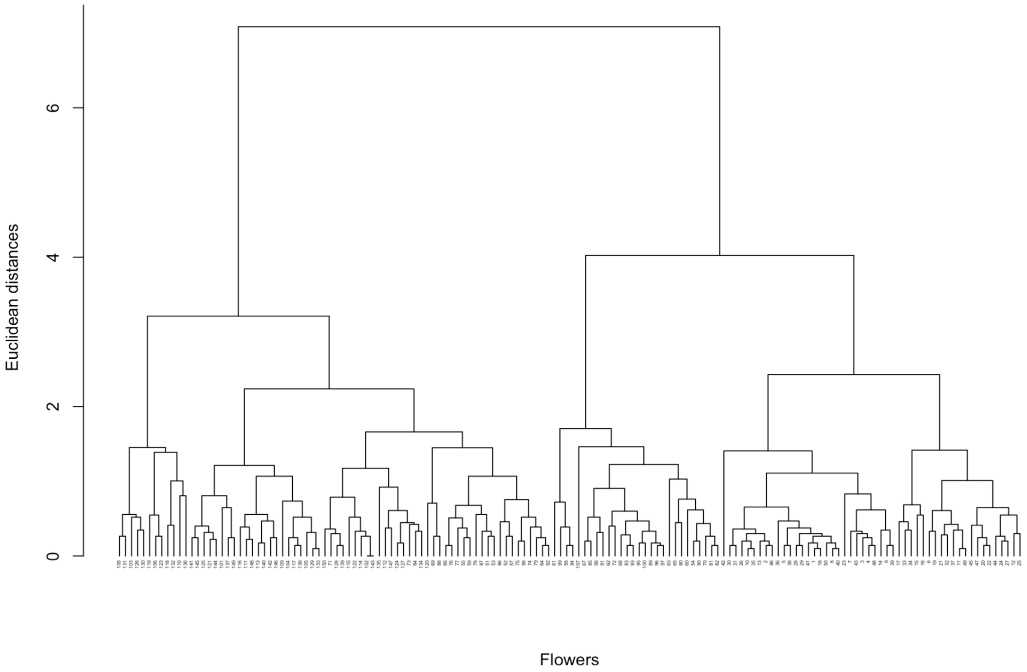


Figure 16: Dendrogram of Fisher's Iris flower data set (complete linkage)

A dendrogram visualizes the hierarchical clustering process. The x-axis represents all the objects, or flowers. The y-axis shows the Euclidean distance at which the clusters are merged. This value is equal to the total within sum of squares (TWSS) from Equation 3. Choosing the number of clusters happens in a similar fashion as for k-means: minimize the TWSS, while keeping the number of clusters as low as possible. Using the bottom-up approach from agglomerative clustering, every time two clusters are merged, it should not add too much distance to the TWSS. Again, there is no one right answer here and, thus, the choice depends purely on heuristics. In the case of Figure 16, either 3 or 4 clusters seems to be a natural choice.

4.4.3 Comparison to k-means

There are two major differences between k-means and hierarchical clustering. The first contrast is that hierarchical clustering is only suitable for small data sets (Tan et al., 2013). This is the case because of the distance matrix that needs a lot of storage (m by m) and the time complexity of $O(m^2 \log m)$. K-means has a complexity of $(O(I * K * m * n))$, but it will only perform the cluster analysis for one specific K. Hence, if one wants to use the elbow method of Section 4.3.3, the algorithm has to be executed m times and, thus, the complexity will deteriorate.

Another major difference is the randomness of the results. Hierarchical clustering will always obtain the same results, contingent on the proximity measure. K-means' results, however, depend on the initialization of the centroids. For every initialization, k-means can converge to a different local minimum. Even after applying the k-means++ (Algorithm 3), there is still randomness involved in the determination of the initial centroids and, hence, in the final results.

4.5 Clustering results

In this section, the techniques discussed above will be applied to the problem at hand. The data will be prepared as discussed in Section 4.2 and the cluster analysis will be performed for each continent and transportation method independently. Due to the small number of origins and far-spread locations in the Americas, no clustering will be applied there. Hence, there are four different groups remaining: air transportation in Europe, air transportation in Asia, ocean transportation in Europe and ocean transportation in Asia. For illustrative purposes, the origins for air transportation in Asia are used as an example throughout this section.

4.5.1 K-means results

For air transportation in Asia, there are 167 origins that ship to a total of 45 destinations. After applying the k-means algorithm to this data, using $\alpha = 3$ for feature scaling, Figure 17 is obtained for further analysis:

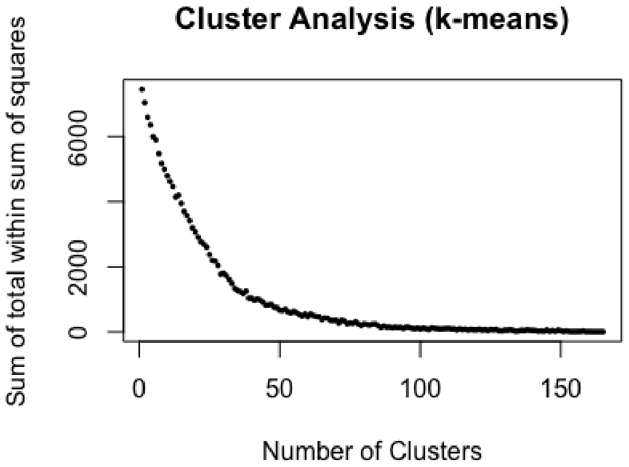


Figure 17: Elbow analysis for air transportation origins in Asia

Based on the graph above and the use of the elbow principle from Section 4.3.3, a reasonable number of clusters would be around $K = 50$. Now, the number of unique origins has basically decreased from 167 to 50. The same approach has also been applied to the three other sets of origins.

Savings If the shipments from origins that are in the same cluster that were sent from the same (air)port on the same day would have been combined, the total costs for air and ocean transportation would have been \$20,602,314. This is a 2.9% reduction, compared to the current costs of \$21,209,464. The number of shipments would decrease from 10,757 to 10,624. Note, however, that the real savings will be higher, as shipments that are not shipped from the same (air)port can possibly be combined as well. Section 7.1 will elaborate on this.

4.5.2 Hierarchical clustering results

The dendrogram, with complete linkage as its proximity measure, for air transportation origins in Asia can be found in Figure 18. Unfortunately, it is not easy to determine the number of clusters visually, and, thus, the number of clusters will be set equal to 50, as for k-means.

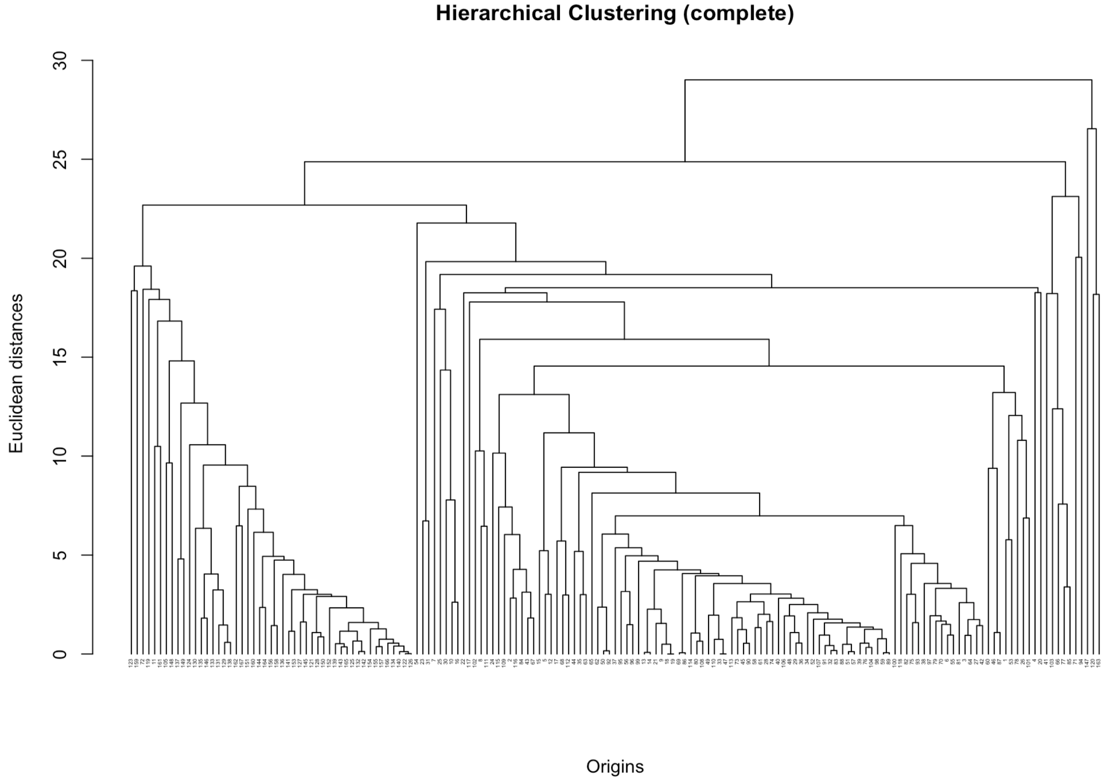


Figure 18: Dendrogram of air transportation origins in Asia

Savings After applying the hierarchical clustering procedure for air and ocean transportation separately for both Asia and Europe, the following results have been found. This again, only concerns the consolidation of same-day shipments that are in the same cluster. The savings for each proximity measure differ substantially, as shown in Table 6. Clearly, the single linkage method yields the best results, while the savings of average linkage are similar to those obtained using k-means. Again, the same applies to these results: the savings would have been higher if shipments departing from different (air)ports could be combined too.

Proximity measure	Total costs	Savings	Number of shipments
Single	$20,532,972	3.2%	10574
Complete	$20,680,683	2.5%	10727
Average	$20,608,400	2.8%	10664

Table 6: Hierarchical clustering results

4.6 Improvements

Even though the current results yield reasonable savings, there are possible improvements that can be implemented. This section will implement these improvements and subsequently compare the results with the ones from Section 4.5.1 and 4.5.2.

4.6.1 Reconstruction of the clustering matrix

The first possible improvement involves a change in the columns of the clustering matrix, as defined in Section 4.2. Keeping the latitude and longitude columns the same, it would be an option to have the different airports and ports as the others columns instead of the actual destinations. After all, it is important that the suppliers ship to the same (air)ports, and not necessarily to the same production site. The second improvement deals with the fact that there are a few origins that only have a few shipments in total, over the course of the past two years. Figure 19 shows the distribution of origins, based on the number of shipments. Due to this low number of shipments, the savings potential when clustering these origins is marginal. Therefore, a possible solution would be to only cluster the origins that have at least x shipments.

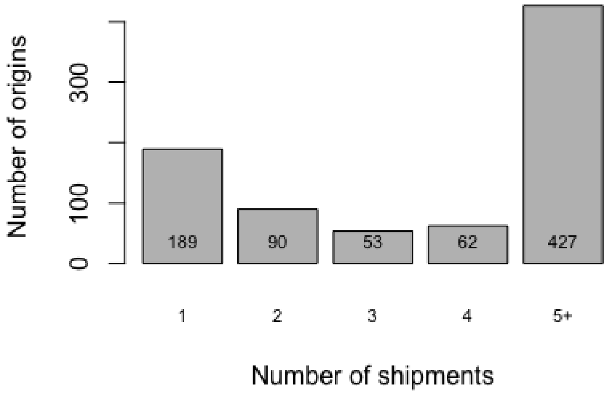


Figure 19: Distribution of the number of shipments per origin

For the remainder of this section, x is chosen to be equal to 5. Applying these improvements to the clustering matrix, the example given in Section 4.2 would now look like Table 7, with $m' \leq m$ and $n' \leq n$.

	Latitude	Longitude	Shipments to (air)port 1	⋯	Shipments to (air)port n'
Origin 1	31.3	120.6	0	⋯	300
⋮	⋮	⋮	⋮	⋮	⋮
Origin m'	31.2	120.4	200	⋯	0

Table 7: Revamped example of the data matrix for the cluster analysis

4.6.2 Improved results

From the 821 different origins, there are only 427 origins that have sent at least 5 shipments over the course of the past two years. Because of the marginal savings potential of the other origins, they are left out of the clustering analysis. Also, the 85 possible destinations are narrowed down to 24 (air)ports. For air transportation in Asia, there are now 127 origins that ship to 11 different airports. In order to determine the number of clusters, the elbow method will be applied again, using Figure 20.

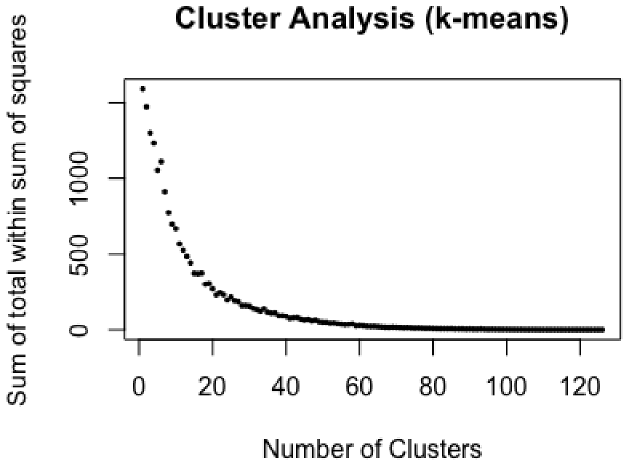


Figure 20: Elbow analysis for air transportation origin in Asia

Around $K = 35$ seems a reasonable choice for the number of clusters. K-means has been applied to the different continents and to both transportation methods. The same number of clusters will be used for hierarchical clustering, using the three different proximity measures. The costs savings can be found in Table 8.

Method	Total costs	Savings	Number of shipments
K-means	$20,583,811	2.9%	10567
Single linkage	$20,416,639	3.7%	10212
Complete linkage	$20,573,753	3.0%	10503
Average linkage	$20,550,742	3.1%	10379

Table 8: Improved results of cluster analysis

4.6.3 Stability of clusters

Up until now, the cluster analysis has only been performed on the full data set of 2015 and 2016. Over the course of these two years, the clusters might have to change, due to a change in suppliers or components necessary at the production sites. In this section, it will be tested whether better results can be obtained by changing the frequency of the cluster analysis. The savings of the basic k-means solution will be used as a benchmark. Table 9 shows the results for different durations of the clusters.

Duration	Total costs	Savings	Increase in savings
2 years	$20,602,314	2.9%	-
1 year	$20,540,926	3.2%	10.1%
6 months	$20,515,531	3.3%	14.3%
3 months	$20,507,152	3.3%	15.7%
1 month	$20,463,141	3.5%	22.9%

Table 9: Savings for different cluster durations

By decreasing the time frame in which a cluster is used, before it is updated (from two years to one month), savings are increased by almost 23%. Not every supplier will send out shipments every month and, thus, optimal clusters will differ over time. Clearly, the figures in Table 9 concern potential savings and not necessarily true savings, as in the real world, clusters have to be chosen before each time period, based on past data. Section 7.1 will elaborate on these implications.

5 Renewal Theory

The second improvement that will be discussed is based on renewal theory. The cluster analysis, discussed in Section 4, focuses on the consolidation of shipments from different origins that are geographically close to each other. Another way to consolidate shipments is to combine shipments on the same lane. That is, many orders do not have to be sent immediately and could possibly be combined with other orders, if the shipping process was to be delayed for a few days. This section will start with an introduction on renewal theory, followed by its application in the consolidation of shipments. Subsequently, the improvement will be implemented and potential savings will be computed.

5.1 Introduction to Poisson processes

A Poisson process is usually involved with the arrival of customers at a bank (Durrett, 2011). The interarrival times of the customers $\tau_1, \tau_2, ...$ are independently exponentially distributed with some rate λ. Let $T_n = \tau_1 + \tau_2 + \cdots + \tau_n$ be the arrival time of the nth customer, and $N(s) = \max\{n : T_n \leq s\}$, i.e. the number of arrivals by time s. That is, $T_n \sim Gamma(n, \lambda)$ and $N(s) \sim Poisson(\lambda s)$. From $\tau_i \sim Exponential(\lambda)$ it follows that $E(\tau_i) = \frac{1}{\lambda}$ and $\mathrm{var}(\tau_i) = \frac{1}{\lambda^2}$. The definitions given in this section are fundamental for the application of Poisson processes in shipment consolidation.

5.2 Shipment consolidation using renewal theory

According to Mutlu et al. (2010), there are three different types of shipment consolidation policies popular in the literature: time-based, quantity-based and time-and-quantity-based consolidation policies. Under a time-based policy, consolidated shipments are released at fixed intervals, regardless of the size of the shipment. Alternatively, under a quantity-based policy, orders are combined until a predefined target load is accumulated. As the latter does not take into account the lead times of the orders and can therefore cause enormous delays to shipments, focus is put on the time-based policy. A time-and-quantity-based consolidation policy could also be applied, but is not very appropriate for the problem at hand. The lead times are the main constraint and if there is a possibility of combining more orders while satisfying this, it does not make sense to release the shipment when a certain target load has been accumulated. Hence, the remainder of this section will focus on a time-based consolidation policy.

Using the definitions from Section 5.1, and replacing the customers arriving at a bank for orders at a supplier, the Poisson process framework can be applied. Mutlu et al. (2010) use this approach in order to find an analytical model for the time-and-quantity-based consolidation policy. They do so by minimizing the expected long-run average cost, denoted by $G(q, T)$. The parameter q defines the target quantity and T the waiting time limit, such that the shipment is released when one of these values is exceeded. As only the time-based consolidation policy is considered in this research, the function can be simplified to $G(\infty, T)$, i.e. setting $q = \infty$. In the paper by Mutlu et al. (2010), every shipment has a fixed cost of K, in addition to a per unit cost c for each order in the load. Although this is a tremendous simplification compared to the complicated cost structure of the considered network, it still shares similar characteristics. Let w be a waiting cost that is incurred for delaying each order for a unit time. Mutlu et al. (2010) show that, in the case of $q = \infty$:

$$G(\infty, T) = \frac{K}{1+T} + wT\left(1 + \frac{1}{1+T}\right) \tag{6}$$

This expression is minimized for:

$$T^* = \sqrt{\frac{K-w}{w}} - 1 \tag{7}$$

The complication is, however, that there is no waiting cost available for the problem at hand. Rather, there are lead times for each order that need to be satisfied. A possible way to bypass this problem is by using w as some sort of proxy of the lead times; choosing a large w when the difference between the transportation time of a shipment and its lead time is small and a small w when the reverse is true. Another option would be to set w equal to the difference between the costs of air and ocean transportation, divided by difference in transportation time. The idea behind this is that if an ocean transportation shipment waits at the supplier for some time, it has to be sent via air transportation in order to satisfy its lead time, which is generally more expensive. These approaches are, however, outside the scope of this research and will, thus, not be further discussed. Instead, two different approaches are proposed in the next sections.

5.2.1 Choosing a waiting time for all shipments

The first method used to shipment consolidation on the same lane involves the introduction of a $T \in \mathbb{N}$, as defined above. This T acts as a upper bound on the time that an order will wait before it is released. An 'optimal' T can be chosen by analyzing the relation between its value and the savings achieved. Clearly the savings will increase with T, but this comes at the expense of delaying shipments. While delaying shipments is not necessary a problem, they have to arrive at the production site within their predefined lead times. Therefore, when choosing T, the proportion of on-time shipments has to be taken into account as well, which is inversely related to the value of T.

For $T = 0$, i.e. the current approach, 94.8% of the shipments arrive on time. That is, the transportation time of the shipment is less than its lead time. As illustrated in Figure 21, when $T = 1$ is chosen, savings of 3.1% are achieved, at the expense of a decrease in on-time shipments from 94.8% to 93.3%.

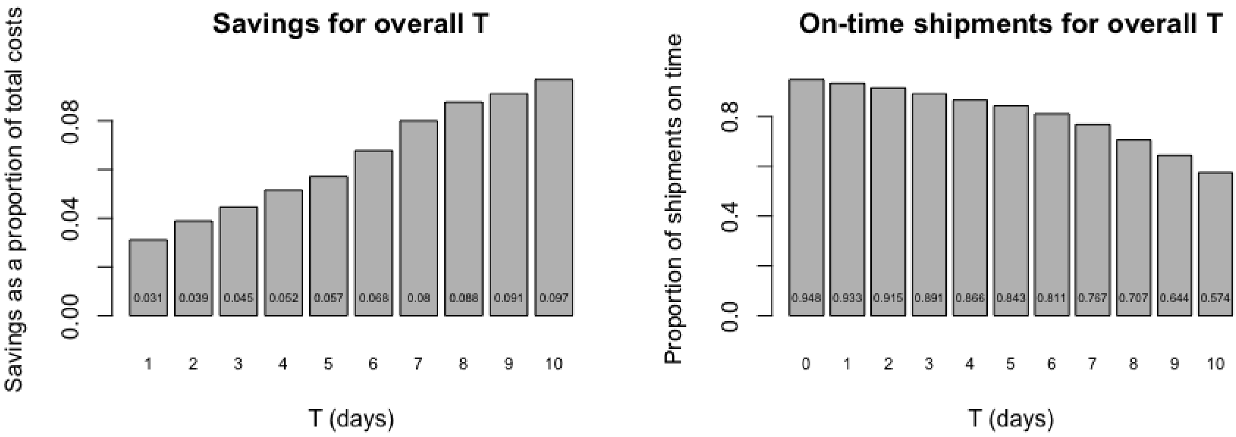


Figure 21: Savings versus on-time shipments per policy T (without clustering)

Figure 21 shows that a lot of the savings are achieved when choosing a small T. When increasing T, a lot of shipments will fail to meet their lead time requirement, as is obvious from the steep decreases in the percentage of on-time shipments. The choice of T depends on the client's priorities, but it will probably not exceed 7, since for larger values the increase in savings is marginal compared to the surge in late shipments.

5.2.2 Choosing a waiting time per lane

The second approach will choose a different waiting time T for each origin-destination pair or lane. This is a logical choice as some lanes only have a few shipments a year, while others receive new orders on a daily basis. Moreover, the interarrival times of the shipments at some lanes might be very large, which makes consolidation of these shipments more difficult. In order to choose a different T for each lane, a matrix is created, both for air and ocean transportation, with the rows being the origins and the columns being the destinations. Subsequently, the number of shipments of each origin-destination pair is found as well as the average interarrival time. That is, the shipments, on each lane, are sorted based on their initialization date in ascending order, and the time between each of the shipments is found, i.e. the interarrival times. The average of these interarrivals times is saved in the matrix. Finally, only the origin-destination pairs with an average interarrival time below some value x (and larger than zero, in the case of one or only same-day shipments) will qualify for a waiting time policy T. For the purpose of this research, $x = 10$ as larger waiting time may substantially worsen the number of on-time shipments. Hence, only lanes with an average interarrival time of less than 10 will be considered.

As each lane can viewed as a Poisson process on its own, the interarrival times, τ_i, are exponentially distributed with rate λ (Cetinkaya Bookbinder, 2003). In order to estimate the parameter of an exponential distribution, a maximum likelihood estimation approach can be applied. In the case of an exponential distribution, Foster (n.d.) shows that the maximum likelihood estimator is:

$$\lambda_{MLE} = \frac{1}{\bar{\tau}} \tag{8}$$

where $\bar{\tau}$ is the sample mean of the interarrival times, i.e. $\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n} \tau_i$. Equivalently, the mean of an exponential distribution is $\frac{1}{\lambda}$. Since all the λ_{MLE}'s have now been established, a choice has to be made regarding the waiting time T_j for each lane j. Assuming a rate of 0.2, an example that illustrates the consequences of the choice of T can be found in Table 10. The probability row represents the chance of collecting at least one more order for the lane that can be consolidated for a given T. These values are found using the Cumulative Density Function of the exponential distribution for $\lambda = 0.2$.

T	1	2	3	4	5	6	7	8	9	10
Probability	0.18	0.33	0.45	0.55	0.63	0.70	0.75	0.80	0.83	0.86

Table 10: Probability of at least one more order given policy T

Due to the right-skewed shape of the distribution, it is most likely for another shipment to arrive within the next few days. In fact, at its mean 5, the probability is equal to 0.63.

This can be taken into account when choosing a value for T. Figure 22 supports this result.

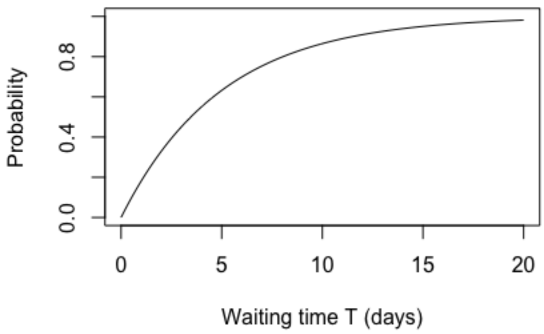


Figure 22: Cumulative Density Function (CDF) of exponential distribution ($\lambda = 0.2$)

In this research, the waiting time T_j for each for each lane j will be set equal to the sample mean of the interarrivals times for that lane, rounded up. That is, $T_j = *\bar{\tau}_j$. This choice is made for the sake of simplicity and because, theoretically, it provides at least a 63% chance of consolidation with at least one more shipment, while keeping the number of late shipment to a minimum. When applying this to the problem at hand, savings of 3.64% are achieved, with 93.2% of the shipments arriving on time.

5.3 Summary

In Section 5.2.1 and 5.2.2, two different approaches have been introduced. The first policy, choosing a waiting time T for all the shipments, had to deal with the trade-off between savings and the percentage of on-time shipments. The choice for T depends on the client's allowance for late shipments, but will probably lie somewhere between 1 and 6 days. The second approach made use of the renewal theory from Section 5.1. Instead of choosing a waiting time T for all the lanes, a different T_j was chosen for each lane j. While the same trade-off still exists, the choice of T can also be based on the Cumulative Density Function of the exponential distribution. When the waiting time for each lane j is set equal to the sample mean of the interarrival times, i.e. $T_j = *\bar{\tau}_j$, the savings are equal to 3.64%, with 93.2% of the shipments arriving on time. This can be compared to setting the overall waiting time $T = 1$, which has a similar percentage of on-time shipments, but only yields savings of about 3.1%. While the second approach provides better results, it also comes with additional complexity. Especially in the real world, the first approach might be more suitable, due to its simplicity, yet decent savings.

6 Synthesis

Most of the results have been discussed in the sections on cluster analysis and renewal theory, independently. This section, however, first summarizes these findings and subsequently combines the two techniques in order to demonstrate a powerful synergy. Finally, the results are improved by decreasing the duration of the clusters from two years to one month.

6.1 Cluster analysis

Partitional and hierarchical clustering yielded different results, ranging from savings of 2.5% to 3.2%. When restructuring the data matrix, savings up to 3.7% could be achieved using hierarchical clustering with a single linkage proximity method. Finally, using different durations of the clusters, i.e. changing the clusters after a fixed period of time, the savings will increase when this duration decreases. Especially annual and monthly clusters seem interesting, since semi-annual and quarterly clusters did not significantly improve the annual results. In the case of k-means clustering, savings increased from 2.9% to 3.5%, or about 23%, when the duration decreased from 2 years to 1 month.

6.2 Time-based policy

Two different time-based policies have been discussed: choosing a waiting time for all shipments, and choosing a waiting time per lane. Due to the complexity, yet marginal improvement in savings of the latter, only the former will be analyzed further. When increasing the waiting policy T for all the lanes, the savings expand, while the percentage of on-time shipment deteriorates. This trade-off is shown in Figure 21.

6.3 Combination

Combining the two methods above provides very powerful results. As shipment consolidation is then done on a cluster-to-destination level, there will be more shipments on this lane, and, thus, there will be more consolidation opportunities using a time-based policy. Hence, the combination of the two approaches can lead to an effective synergy. Table 11 shows a comparison between the results from Section 5.2.1 and the potential improvement with hierarchical clustering (single linkage), both for the savings and the on-time shipments.

T	Savings	Savings (cluster)	On-time shipments	On-time shipments (cluster)
0	0%	3.7%	94.8%	94.4%
1	3.1%	6.0%	93.3%	92.8%
2	3.9%	7.3%	91.5%	90.8%
3	4.5%	8.5%	89.1%	88.0%
4	5.2%	9.6%	86.6%	85.1%
5	5.7%	10.4%	84.3%	82.4%
6	6.8%	11.8%	81.1%	79.2%
7	8%	13.1%	76.7%	74.5%
8	8.8%	13.9%	70.7%	60.8%
9	9.1%	14.5%	64.4%	61.1%
10	9.7%	15.2%	57.4%	53.2%

Table 11: Savings versus on-time shipments under different time-policies T, with and without clustering (hierarchical, single linkage)

Depending on the client's allowance for late shipments, extraordinary results can be obtained by combing the two shipment consolidation methods. Considering that under the current approach around 5.2% of the shipments are late, the client might decide to increase this number to around 10% on average. This means it will choose for the waiting time policy $T = 2$ and, therefore, reduce the total costs for air and ocean transportation by about 7.3%. In addition, it is also possible to increase the number of on-time shipments by changing the method of transportation. As most of the late shipments are sent via ocean transportation, it can be considered to use air transportation instead and, therefore, reduce the transportation time significantly. Air transportation is, however, generally more expensive and hence the trade-off between savings and on-time shipments will remain. It is also important to note that the lateness of shipments is not binary: when incrementing T, the difference between the lead time and transportation time of late shipments will also increase, and might cause problematic delays.

6.4 Shorter cluster duration with a time-based policy

Section 4.6.3 discussed the duration of clusters and the potential influence on the achieved savings. Hierarchical cluster (single linkage) yielded 3.7% in savings, for a two year duration of the clusters. The total savings of clustering in combination with a time-based policy can be reinforced when a shorter cluster duration has been chosen. Table 12 shows the increase in savings for a one month duration, for different values of waiting time policy T.

T	Savings clustering (2 years)	Savings clustering (1 month)
0	3.7%	4.6%
1	6.0%	7.1%
2	7.3%	8.4%
3	8.5%	9.6%
4	9.6%	10.7%
5	10.4%	13.4%

Table 12: Savings for hierarchical clustering (single linkage) per two years versus hierarchical clustering (single linkage) per month for different values of policy T

Clearly, the savings found in Table 11 can be expanded by decreasing the duration of the cluster from two years to one month. The increase in savings ranges from 0.9 to 3.0 percentage points.

7 Conclusion and discussion

7.1 Practical challenges

Currently, cluster analysis has been done using past data, and subsequently these clusters have been applied to the same data. In the real world, however, clustering has to be done using this same past data, but will be used on future shipments. For this reason, only potential clustering savings have been calculated up until now. Since only two years of data are available, it is not possible to test the cluster analysis on new data. When decreasing the duration of the cluster, for instance, to one month, clustering on past data can now be used to cluster origins of new shipments, within the boundaries of the two years of data. When considering this, the following question arises: which past data should be used in the cluster analysis for the consolidation of future shipments? Is it best to use e.g. January in the cluster analysis in order to apply this to the next month's shipments, or is there a certain seasonality in the data that makes January 2015 a good candidate for January 2016? Conversely, it would also be possible to use e.g. the cluster analysis of January and February for March, February and March for April, and so on. The answers to these questions are, however, outside the scope of this research.

Another challenge that arises when consolidating shipments in clusters concerns the fact that every cluster needs a central location where all the orders are combined. This also means that it takes more time and effort to send out the shipments, which may involve additional costs. In this research, the costs that are associated with clustering, are not taken into account.

Finally, consolidating shipments from origins that are close to each other, but in different countries, can lead to organizational and practical issues. Every lane, on a country-to-country level, has its own preferred carrier, as described in Section 2.1.4. When consolidation takes place, a choice has to be made regarding the carrier. Combining shipments between countries also leads to complications for any of the carriers. Hence, in this study only shipments from the same country are consolidated, even though potential savings could be achieved by relaxing this constraint.

7.2 Conclusion

This paper examined the process of analyzing DHL LLP's current approach and proposed several improvement techniques in order to reduce the costs of transportation. It started by giving a formal problem description, the goal of the research and the client's motivation for optimizing the network. Subsequently, some related literature was discussed on shipment consolidation in general, and clustering in specific. In order to find possible improvement methods, the costs of the current approach had to be calculated and analyzed. This process included changing incorrect data from 2015 and interpolating missing data from both years. During the analysis of the current costs, two potential improvements were identified: shipment consolidation and shipment rerouting. The paper subsequently focuses on examining machine learning methods that can help with the former improvement. First, a clustering approach is introduced that groups together origins that are geographically close to each other and generally ship to the same destinations. Second, a time-based policy is analyzed that assigns a maximum waiting time for an order at each supplier. Finally, these two methods are combined and become a very powerful synergy. Based on the client's tolerance for late shipments, extraordinary savings can be achieved. These results can be reinforced by decreasing the duration of the clusters from two years to one month. With a maximum allowance of 10% for late shipments, savings up to 8.4% can be gained. Although these findings only apply to this particular data set, the techniques proposed in this research are generic, and can therefore be implemented in any similar network.

7.3 Further research

Even though multiple methods have been discussed in this paper, there is still ample room for improvement. First, as mentioned in Section 7.1, the data can be analyzed in order to determine the right implementation for cluster analysis, using past data and future shipments. Second, different clustering techniques can be analyzed and compared. In this study, only partitional and hierarchical clustering have been considered. Using a technique such as fuzzy clustering can provide new insights, as in that case an origin can belong to multiple clusters (Gath Geva, 1989). Third, the possibilities of changing ocean shipments to air shipments, can be examined further. For each lane, it can be analyzed what the optimal waiting time policy would be, by delaying ocean shipments and changing them to air shipments in case they will otherwise arrive too late. Finally, several rerouting techniques can be explored. For each cluster, a number of close (air)ports can be compared. Subsequently, the (air)port that is on average the cheapest can be assigned to the cluster. In this case, transportation costs to the different (air)ports have to be taken into account as well. This is only one possible approach for the rerouting of shipments.

References

2016 annual report. (2017). Deutsche Post DHL Group. Retrieved from http://www.dpdhl.com/content/dam/dpdhl/Investors/Events/Reporting/2017/FY2016/DPDHL_2016_Annual_Report.pdf

Algorithms for k-means clustering. (2013, Jan). University of California San Diego. Retrieved from https://cseweb.ucsd.edu/ dasgupta/291 -geom/kmeans.pdf

Aloise, D., Deshpande, A., Hansen, P., Popat, P. (2009). *Np-hardness of euclidean sum-of-squares clustering.* Machine learning, 75(2), 245–248.

Arthur, D., Vassilvitskii, S. (2007). *k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms* (pp. 1027–1035).

Cetinkaya, S., Bookbinder, J. H. (2003). *Stochastic models for the dispatch of consolidated shipments. Transportation Research Part B: Methodological*, 37(8), 747–768.

Deng, N. (2013). *Shipment consolidation and distribution models in the international supply chain (Unpublished doctoral dissertation).* University of Missouri–Columbia.

Desrochers, M., Desrosiers, J., Solomon, M. (1992). *A new optimization algorithm for the vehicle routing problem with time windows.* Operations research, 40(2), 342–354.

Durrett, R. (2011). *Essentials of stochastic processes.* Springer.

Dutch, S. (2016, Jan). *Converting utm to latitude and longitude.* University of Wisconsin - Green Bay. Retrieved from https://www.uwgb.edu/dutchs/UsefulData/UTMFormulas.HTM

Fisher, R. A. (1936). *The use of multiple measurements in taxonomic problems.* Annals of eugenics, 7(2), 179–188.

Foster, P. (n.d.). *Maximum likelihood estimation.* The University of Manchester. Retrieved from http://www.maths.manchester.ac.uk/ peterf/CSI_ch4_part1.pdf

Gath, I., Geva, A. B. (1989). *Unsupervised optimal fuzzy clustering.* IEEE Transactions on pattern analysis and machine intelligence, 11(7), 773–780.

Hartigan, J. A., Wong, M. A. (1979). *Algorithm as 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108.

Johnson, R. (2014). *Clustering and association.* University of Notre Dame. Retrieved from http://www3.nd.edu/ rjohns15/cse40647.sp14/www/content/lectures/13%20-%20Hierarchical%20Clustering.pdf

Kodinariya, T. M., Makwana, P. R. (2013). *Review on determining number of cluster in k-means clustering.* International Journal, 1(6), 90–95.

Lead logistics partner (LLP). (n.d.). Deutsche Post DHL Group. Retrieved from http://www .dhl.com/en/logistics/lead_logistics_provider_llp.html

Mutlu, F., Cetinkaya, S. i. l., Bookbinder, J. H. (2010). *An analytical model for computing the optimal time-and-quantity-based policy for consolidated shipments.* Iie Transactions, 42(5), 367–377.

Slonim, N., Aharoni, E., Crammer, K. (2013). *Hartigan's k-means versus lloyd's k-means-is it time for a change?* In Ijcai.

Tan, P.-N., Steinbach, M., Kumar, V. (2013). *Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining.*

Unsupervised learning. (n.d.). The University of Iowa. Retrieved from http://homepage.divms.uiowa.edu/ hzhang/c145/notes/chap18b.pdf

Zhao, Q., Hautamaki, V., Franti, P. (2008). *Knee point detection in BIC for detecting the number of clusters.* In International conference on advanced concepts for intelligent vision systems (pp. 664–673).