2019

2nd year project: Report

Paul Schonewille, Jelle Willekes

DHL LLP

27-6-2019

Jelle Willekes
Paul Schonewille

Table of Content

Introduction

DHL LLP has received data from one of their partners for which DHL LLP manages the logistics. The question asked by the partner was quite simple, they wanted savings on their logistics. We are asked to create time series for the weight, volume and amount of shipments. Afterwards we are asked to check for seasonality. Is activity higher on certain days than on others, which days and how certain are we about this. Do the all days share common features across a lane? Do the same days have about the same impact across different lanes, or is there no relation between different places only between time? Are the measures we found comparable across different transport routes? These are some of the questions we are asked to solve for the first part of the case.

The second part of the case is mainly about the clustering that DHL LLP used. What are the values of weight, volume and amount of shipments for different clusters? We need to create a code to find these values. We also need to see if the clustering done by DHL can be improved, in what way, how and why we think our way reduces costs is discussed in the main part of our report.

The conclusion of this report will be a small business summary in which we give advice on a managerial level and this advice is easily readable for all readers of this report. From these two pages one can easily see what he or she should do in certain circumstances and how to improve on those circumstances if possible, to do so.

Jelle Willekes
Paul Schonewille

Econometric methods

Data pre-processing

The first question asks us to choose between the most used choosing an origin cluster and five short ranged destinations or find the five most important origin clusters and a long-distance destination. We chose the most used origin destination and then took the five most used clusters as destination between 20 and 400 kilometres.

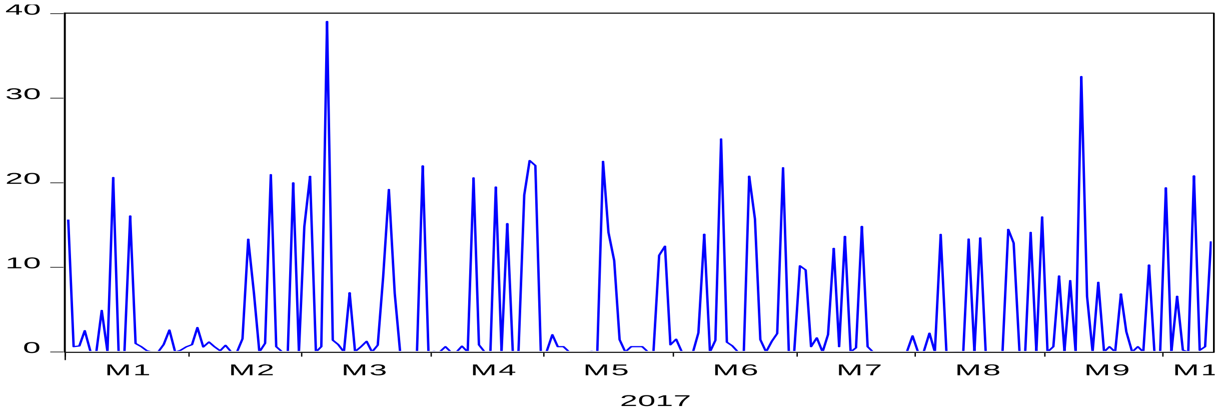
The most used cluster as origin cluster was cluster 2. We did this first part with excel, the file including the methods we used is in the zip file in the Econometric Methods tab. We started with creating a unique column of origin clusters. Afterwards we used the excel counter to count the amount of times that unique value appeared in the total column of data. The conclusion was that cluster 2 was used 13483 times, which is by far the highest amount a cluster is used. Then we had to find five clusters with a short distance to cluster 2. As mentioned, we took the five most used lanes, because we wanted to have as many datapoints as possible. We continued in excel and created now a unique column of destinations now only starting at cluster 2. Then we as before counted how many times each unique destination was in the list and our result is that the clusters 552,547,551,549 and 611 are the most used clusters as destination. The five lanes we now have are 2-547, 2-549, 2-551, 2-552 and 2-611.

552
551
2
611
547
549

Our origin cluster is in Toulouse in the south of France, which is almost in the middle of our four European countries. Our chosen destination clusters are also present in France, which makes sense since we took the five lanes with most activities at a short distance (20-400 km). These clusters are between 44 km and 212 km away. The most used lane is 2-552 which has a distance of 123 km. This surprised us a bit, since we thought that the shortest lane would be used most, also because the shorter lane was located in a more populated area, while the most used lane was located in a relatively small village.

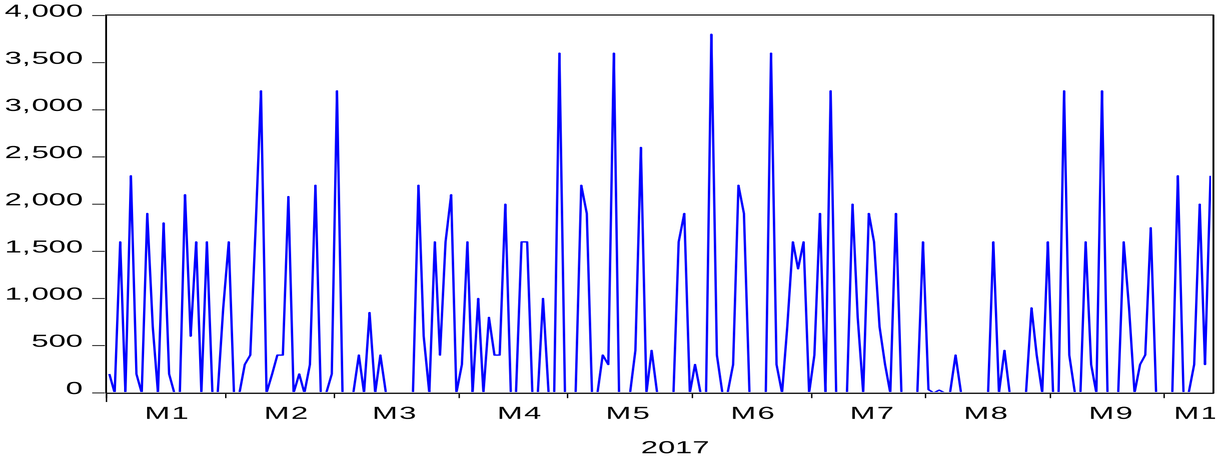
For the last part of the first question we are asked to create series per weight, volume and amount of shipments, which leads to a total of 15 series, 3 per lane. For these 15 series we need to create time series. We aggregated the weight, volume and shipments, if more transports on a day were made, to get a single value per day. We did this still in excel, we used the SUMIF function in excel to separately add the weight, volume and shipments per day, with the range of addition either the weight, volume or shipments, and then check if the day is equal to the day checked and we run this through all days of the sample. The graphs below represent some lanes with Volume, Weight and shipments, we included the outliers to get a better view for now of the actual data.

Gross Volume 2-547



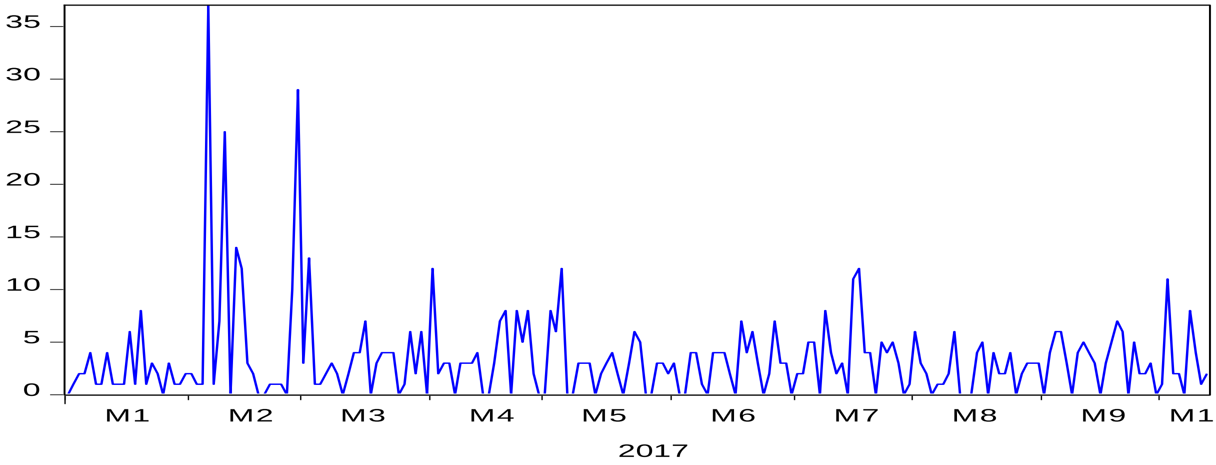
Graph of the gross volume for the lane 2-547

Gross Weight 2-551



Graph of the gross weight for the lane 2-551

Total units of shipments 2-552



Graph of the total amount of shipments per day for the lane 2-552

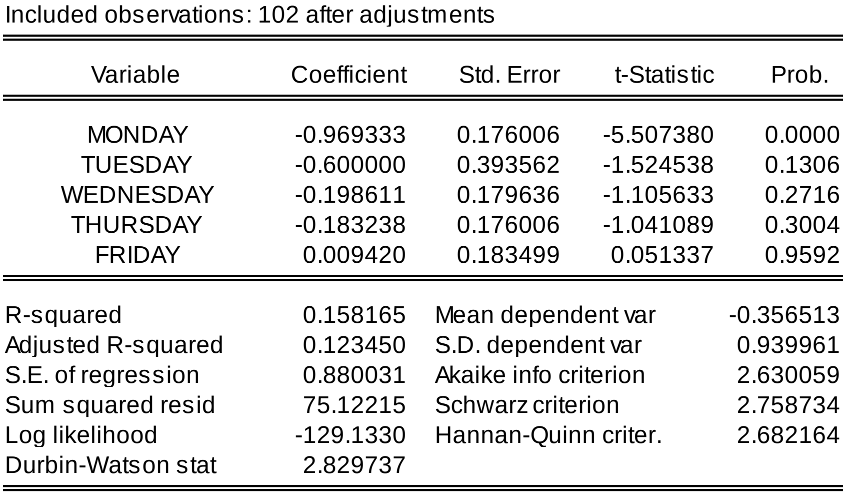
We checked for outliers and decided to control for some of these in the 2nd question, because now we do not yet need to analyse the data.

Daily seasonality

The second question then continues to start the investigation of a possible daily seasonality. We then decided to take the growth rates. We preferred growth rates over log-levels because growth rates better depict an increase or decrease in activity, while log-levels just take the log of the values and not necessarily depict an increase or decrease. This reason is also why we took growth rates over normal levels. With normal levels we would not really know whether we had increases or decreases on certain days, just what the estimated values would be. We want to know whether activity is different on certain days. If we see a negative growth rate, we immediately know that activity decreases on that particular day.

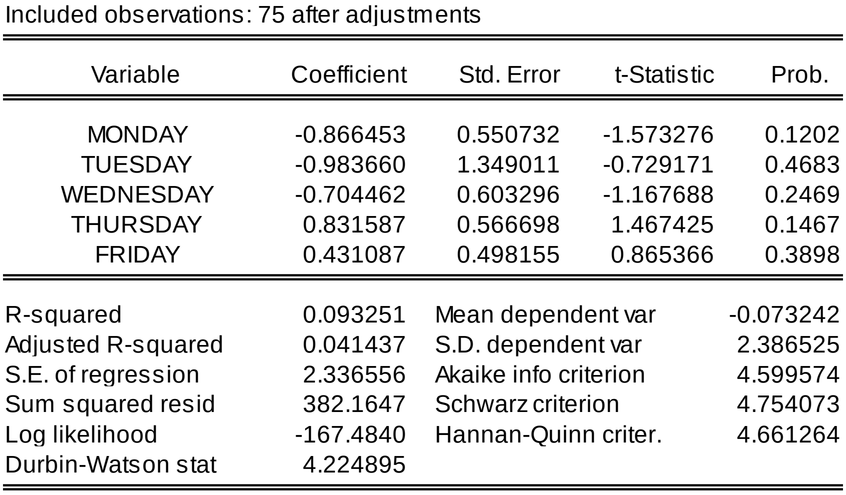
After transforming the data to growth rates, a process which reduced the sample size, because some observations did not have a direct predecessor, thus dividing by 0. These values are set to 0, which is no problem since the growth would be infinite, thus an outlier and set to 0. We decided to set all detected outliers to 0, our criteria for being an outlier can be found in our excel file under the tab 'code for Eviews'. With growth rates setting outliers to 0 is not as radical as with normal levels, due to the fact that activity then would be constant and not completely removed. We first thought about taking ¼ of the value of the outlier to reduce the extremeness of the graphs, this did not help with our adjusted R squared or the likeliness of our estimates, thus rejected this idea. We also thought about 2.5 times the standard deviation, but because we removed all outliers in one go and not per day, we couldn't justifiably take an average of standard deviation. We then looked at the different standard deviations and some days had an increase and others a decrease, we then decided to make our lives easier and set the outliers to 0.

When doing the estimation, we made a choice between relatively high adjusted R-squared or likeliness of our estimates, we preferred our estimates to be likely and thus sacrificing some fit, we changed our likeliness and R-squared by removing or not removing outliers.



```
Included observations: 102 after adjustments
=========================================================================
      Variable           Coefficient    Std. Error    t-Statistic    Prob.
=========================================================================
       MONDAY             -0.969333     0.176006      -5.507380     0.0000
       TUESDAY            -0.600000     0.393562      -1.524538     0.1306
      WEDNESDAY           -0.198611     0.179636      -1.105633     0.2716
      THURSDAY            -0.183238     0.176006      -1.041089     0.3004
       FRIDAY              0.009420     0.183499       0.051337     0.9592
=========================================================================
R-squared                  0.158165   Mean dependent var       -0.356513
Adjusted R-squared         0.123450   S.D. dependent var        0.939961
S.E. of regression         0.880031   Akaike info criterion     2.630059
Sum squared resid          75.12215   Schwarz criterion         2.758734
Log likelihood            -129.1330   Hannan-Quinn criter.      2.682164
Durbin-Watson stat         2.829737
=========================================================================
```

Table of the shipments for lane 2-551, with a 'good' adjusted R-squared and fine probabilities



```
Included observations: 75 after adjustments
==========================================================================
      Variable          Coefficient    Std. Error    t-Statistic    Prob.
==========================================================================
       MONDAY            -0.866453     0.550732      -1.573276     0.1202
      TUESDAY            -0.983660     1.349011      -0.729171     0.4683
     WEDNESDAY           -0.704462     0.603296      -1.167688     0.2469
      THURSDAY            0.831587     0.566698       1.467425     0.1467
       FRIDAY             0.431087     0.498155       0.865366     0.3898
==========================================================================
R-squared                 0.093251    Mean dependent var         -0.073242
Adjusted R-squared        0.041437    S.D. dependent var          2.386525
S.E. of regression        2.336556    Akaike info criterion       4.599574
Sum squared resid         382.1647    Schwarz criterion           4.754073
Log likelihood           -167.4840    Hannan-Quinn criter.        4.661264
Durbin-Watson stat        4.224895
==========================================================================
```

Table of the weight for lane 2-611, with significant probabilities, but low adjusted R-squared

Our estimates show that activities decrease on Monday, this is something we found in all our estimates, although the fit is in general not very good and often the likeliness of the estimates is also not that high, we still think that we can conclude that in general it is safe to assume that weight, volume and shipments decrease on Mondays for each lane we have investigated.

The other result we commonly see is that an increase in weight, volume and shipments occurs on Thursdays. Although we still need to investigate the validity of these results, we think that is quite likely that a decrease on Monday and an increase on Thursday are significant.

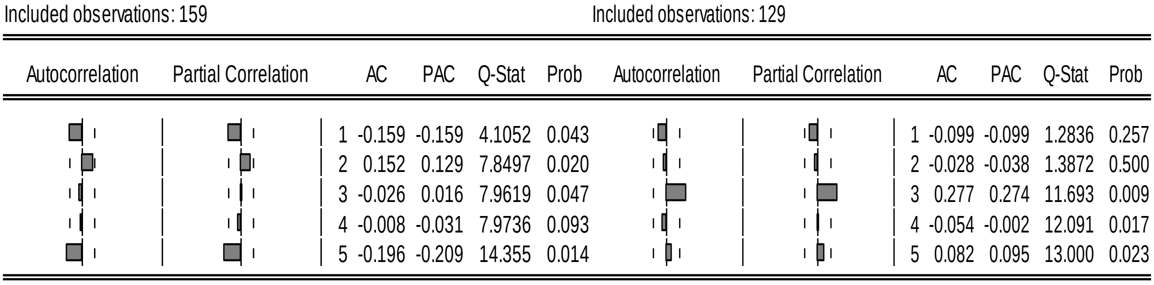
We also needed to check if seasonality is visible, so is there a significant daily effect in total. To check for this, we used a Wald test and tested for $a_1+a_2+a_3+a_4+a_5=0$, if so then a daily effect may be there, but it is cancelled by other days. If rejected, we know for sure that some days have more impact than other days.

We only had lane 2-552 with a significant daily effect of 10%, this is the only time we take 10% as critical value due to otherwise having no complete lane with significance, all other lanes had either volume, weight or shipment with a possibility of no daily effect that was not insignificant.

test Statistic	value	df	probability
t-stat	1,756412	154	0,0006
F-stat	3,084981	(1;154)	0,0006
Chi-squared	3,084981	1	0,0005
volume			
t-stat	3,397432	154	0,0009
F-stat	11,54255	(1;154)	0,0009
Chi-squared	11,54255	1	0,0007
weight			
t-stat	3,484532	154	0,081
F-stat	12,14197	(1;154)	0,081
Chi-squared	12,14197	1	0,079
shipments			

Table of lane 2-549 with Null hypotheses c(1) +c(2)+c(3)+c(4)+c(5)=0

The autocorrelation test we performed is the Q-statistic test in Eviews. The only interesting values we observed were for the shipments of the lane 2-547. According to our results we have autocorrelation starting with a lag of 3 days up to 5 days. So yesterday and the day before have no autocorrelation, but combined with 3, up to 5 days ago, then we have autocorrelation. In our further results we include 3 lags, because we have to few observations on the lane 2-547 to include more lags. The other interesting lane was 2-552 shipments. In this lane we have autocorrelation for the first three lags, then not with 4 lags, but again significant autocorrelation with 5 lags.



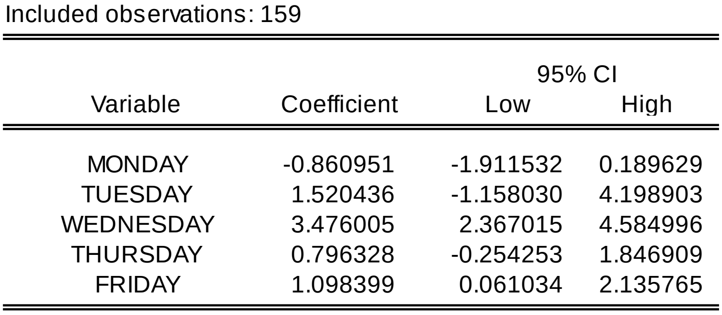
```
Included observations: 159                                    Included observations: 129

Autocorrelation  Partial Correlation     AC     PAC   Q-Stat  Prob   Autocorrelation  Partial Correlation     AC     PAC   Q-Stat  Prob

                                     1 -0.159 -0.159  4.1052  0.043                                     1 -0.099 -0.099  1.2836  0.257
                                     2  0.152  0.129  7.8497  0.020                                     2 -0.028 -0.038  1.3872  0.500
                                     3 -0.026  0.016  7.9619  0.047                                     3  0.277  0.274  11.693  0.009
                                     4 -0.008 -0.031  7.9736  0.093                                     4 -0.054 -0.002  12.091  0.017
                                     5 -0.196 -0.209  14.355  0.014                                     5  0.082  0.095  13.000  0.023
```

The Q-statistic values for lane 2-552 for the growth rates of shipments and right the values for the lane 2-547 for shipments

The next question was about to test whether the days separately are significantly different from 0, with our growth rates, this implies that on certain days a significant increase or decrease in activity could be visible. We do not have a trend in our data for significant days, we have all days significant for some lane, but no trend. This is because when doing the estimation at the beginning we did not try to minimize the standard error for our estimates. When we then check the 95% confidence interval we have that most of our estimates are not significantly different from 0, this was something we knew would happen when we carried out the regression, but we decided to add more value to a higher adjusted R-squared or high likeliness of our estimate than to lower standard errors.



Included observations: 159

		95% CI	
Variable	Coefficient	Low	High
MONDAY	-0.860951	-1.911532	0.189629
TUESDAY	1.520436	-1.158030	4.198903
WEDNESDAY	3.476005	2.367015	4.584996
THURSDAY	0.796328	-0.254253	1.846909
FRIDAY	1.098399	0.061034	2.135765

Table of the confidence interval for the lane 2-552 with regards to volume, Wednesday and Friday are significantly different from 0, although the estimate for Wednesday has a 0 probability.

Then we needed to check whether a significant difference between days exists. Thus, is Monday significantly different from Tuesday, Wednesday and the other days? For volume, weight and shipments in the lane 2-549 we do not reject the null that each day has the same impact on activity. This is our only complete lane with no significant difference between days, so no adaptations need necessarily to be made for this lane.

Test Statistic	value	df	Probability
F-Statistic	0,535332	(4;65)	0,7102
Chi-squared		4	0,7098
Volume			
F-Statistic	0,675749	(4;65)	0,6112
Chi-squared	2,702995	4	0,6087
weight			
F-Statistic	1,010312	(4;65)	0,4087
Chi-squared	4,041247	4	0,4006
Shipments			

The tables for testing significant difference between days for lane 2-549 for respectively growth rates in volume, weight and shipments. Null Hypotheses c(1) =c(2) =c(3)=c(4)=c(5)

The last part of the second question is to test for heteroskedasticity, for this we used the White-test in Eviews. The results we got display that often we have heteroskedasticity for shipments, while volume and weight display homoskedasticity. These tests do however also display the impact of choice, since a lot of p-values are between 5-20%, so if we would maintain a higher α we would have far more heteroskedasticity, we choose a 5% level of significance, due to the common use of 5%.

Volume	Weight	Shipments	ProbF(a, b)	prob. Chi-squared	null at 5%	HCSE
547			0,3649	0,3572	no	
549			0,6535	0,6354	no	
551			0,0953	0,0953	no	*
552			0	0	yes	*
611			0,7602	0,7456	no	
	547		0,1895	0,1864	no	
	549		0,8295	0,8166	no	
	551		0,4775	0,4656	no	
	552		0,0298	0,0886	no	*
	611		0,174	0,1695	no	
		547	0,0206	0,0222	yes	*
		549	0,5835	0,565	no	
		551	0,023	0,0249	yes	*
		552	0,0642	0,0649	no	*
		611	0,0363	0,0389	yes	

Pooling equations

The third question asked us something we had not yet done in Econometric Methods I. We needed to use SURE estimation to detect a pattern between the different lanes for the different days. We needed to code the data we wanted to cross-estimate on, which was not too difficult. The estimation was carried on across the different lanes with different estimators. We then tested using the Wald test, as we have done before to check for difference between estimators, to see for example if Monday has a different impact on lane 2-551 than on lane 2-611. We chose not to pool lags since this would disturb the elegance of the regression and decrease the observations thus making the results less viable.

Our results are that for amount of shipments only the Wednesdays have a different impact across lanes, we do no reject the null for all our other days. Thus Monday, Tuesday, Thursday and Friday can be handled across lanes without too much problems.

Monday	Tuesday	Wednesday	Thursday	Friday	Shipments
0,9924	0,2887	0	0,2089	0,171	total prob
-0,057389	1,008004	1,557597	-0,618543	0,666766	547-611
-0,061689	0,057571	0,652247	-1,13710	0,109903	549-611
-0,034345	0,310879	0,478618	-0,943806	-0,280771	551-611
0,098298	1,707491	1,651343	-0,561418	-0,106718	552-611

Table with values for the Wald test in the SURE test for the probability across lanes for the amount of shipments

For weight we can conclude that we can only reject the Wednesday across lanes at a 5% level. Although the Thursday can be rejected at 10%, we take the 5% as critical value. So, for both weight and shipments we cannot use different lanes to predict other lanes, again for the other days we can use different lanes to predict our lane of interest.

Monday	Tuesday	Wednesday	Thursday	Friday	Weight
0,9995	0,2308	0	0,0746	0,6105	total prob
-0,107595	4,584781	1,207601	0,630845	-0,160177	547-611
-0,138679	-0,131815	0,865575	-0,96501	-0,510195	549-611
-0,126354	0,758632	1,384931	-0,596451	0,64678	551-611
0,003919	3,257783	5,051822	1,319838	0,171862	552-611

Table with values for the Wald test in the SURE test for the probability across lanes for the weight

Lastly for volume we discovered that we do not reject the null for any day in the week, we have a Wednesday with 5,04%>5%. Thus, we have for volume a similarity across all lanes and for all weekdays.

Monday	Tuesday	Wednesday	Thursday	Friday	Volume
0.9994	0.5860	0,0504	0,8915	0.7023	total probability
0.284254	1.229.744	1,271073	0,093753	-0,449105	547-611
0.325013	0.349385	0,800673	-1,41768	-0,025588	549-611
0.343354	5.686.014	1,951661	-0,836814	1,49714	551-611
0.202599	2.270.293	3,179218	-0,455546	-0,331254	552-611

Table with values for the Wald test in the SURE test for the probability across lanes for volume

Jelle Willekes
Paul Schonewille

This concludes the first part of this report. For further information and all other tables and graph, see the other files in the map.

Jelle Willekes
Paul Schonewille

Operations Research

A challenging task within Operations Research is clustering. In the following part of the report we are going to group a set of shipments in such a way that the shipments in the same group are more similar in some sense. The similarity of the shipments will be mainly based on location. The goal of the following tasks is to find clusters that minimize the total distance from the origin of a shipment to the origin of the cluster it belongs to.

Data pre-processing

The data input for the clustering task consists of 29 data variables with a total of 8120 shipments which were carried out during the time frame from January to October 2017. Not all the 29 data variables are useful when minimizing the total distance of the shipments, therefore we only used the 13 data variables which are necessary. The other 16 data variables can be obtained by a transformation of our 13 important data variables. The excel file of the 13 data variables is converted to a .txt file such that we can read it easily in our java program. The first class dataVariable we wrote in java saves all 13 variables, which include *weight, units, originCluster, originClusterLat, originClusterLong, originLat, originLong, destinationCluster, destinationClusterLat, destinationClusterLong, destinationLat, destinationLong* and *volume.*

Later on, we will be interested in only the cluster information, therefore we wrote another java class called clusterVariables which saves the 5 most important data variables to do calculations on cluster locations. The clusterVariables class contains the variables *originCluster, originClusterLat, originClusterLong, originLat* and *originLong.*

Clustering I

The task for Clustering I is to develop an appropriate datastructure that will allow us to analyse the clustering of the shipments that is given in the data set. We will implement methods in java that calculate *the total distance of all shipments from the origin location to the origin cluster, the total weight that is being transported from the origin locations to the origin clusters, the total volume that is being transported from the origin locations to the origin clusters, the total number of shipments that is being transported from the origin locations to the origin clusters* and these measures for the destination clusters too. The output will be a summary of these calculations for the total amount of cluster, the average cluster and the 3 lowest and highest cluster values.

The summeryClustering1 class prints all data such that the user will get an overview of the summery statistics. The summary statistics *Total Weight, Total Volume, Total Number of Shipments* and *Total distance of all shipment to origin cluster* as well as *destination cluster* are necessary to compare multiple clusters in the tasks coming up. The summeryClustering1 class contains the following methods to calculate the above mentioned summary statistics: *totalWeight, totalVolume, totalNumberOfShipments,readArray, distance, degreeRadian, radianDegree, totalDistanceInfo*(for origin values)*, totalDistance2Info*(for desination values)*, lowestValues* and *highestValues.* The summary statistic can be found in the tables below.

Summary statistics of all clusters	
Total Weight	4.698326424399955E7
Total Volume	1060455.0830000355
Total Number of Shipments	236116.0
Total distance of all shipments to their origin cluster locations	97942
Total distance of all shipments to their destination cluster locations	90189

Summary statistics of average cluster	
Average Weight	72505.03741357954
Average Volume	1636.5047577161042
Average Number of Shipments	364.37654320987656
Average distance of all shipments to their origin cluster locations	151.0
Average distance of all shipments to their destination cluster locations	139.0

Summary statistics for 3 Lowest and Highest cluster values		
	Lowest	Highest
Total Weight	[0.0, 0.0, 0.0]	[242000.0, 70000.0, 53159.0]
Total Volume	[0.0, 0.0, 0.0]	[83950.026, 81468.607, 63438.236]
Total Number of Shipments	[1.0, 1.0, 1.0]	[485.0, 203.0, 200.0]

Clustering II

Now that we have the most important summery statistic we can develop and implement an algorithm that can find a different assignment of the locations to clusters. We want to find different assignment such that the total distance of all shipments from the location to the assigned cluster is minimized. We will take an algorithm that can take the number k of cluster as input. We choose to use the well-known K-means algorithm. The main reason for using the K-means algorithm is that it is in *O(n)* which implies that it is easy to use for huge data sets. Since we have 81520 shipments and 13 data variables, we have a very big data set. We will use the variables saved in clusterVariables to do our calculations within this class.

First the program creates an input field where the user can enter the desired amount of cluster for which he wants to optimize. Then the program creates k random clusters on location between a latitude of 36 and 57 and a longitude between -9 and 26. These points are the minimum and maximum points within our data set. These points will be added do an array of doubles and then it finds for each point its closest random cluster location and adds these points to the cluster. Now we have random clusters which will definitely not give us the minimum total distance, therefore we implemented to following method. The program calculates the average points (lat, long) of all elements within the random cluster and sets the average latitude and longitude as its new latitude and longitude. In case there are no shipments closest to a specific cluster we included a count option in the updating method such that if there are no shipments closest to a specific cluster, the random cluster will be assigned to a random location again until it has some elements. With this method the total distance will be smaller as there will be new shipments closest to this new random cluster location. This method keeps calculating new averages up on request. Namely we can change the number of updates we want to include in the program. The output of the program is a calculation of the total distance given for the number of updates the program takes and the optimal cluster points that will minimize the total distance.

Total distance (km) for different numbers of Clusters and updates				
Number of Clusters	Updates			
	1	3	10	1000
5	5.5198423E7	1.7571227E7	1.1508461E7	1.1496157E7
10	2.1548694E7	9642228.0	6300402.0	6300402.0
50	1.6580864E7	2470202.0	2162003.0	2062145.0
648	3844897.0	376080.0	206927.0	55901.0

Obviously, as we can see in the table above, the greater number of clusters we use for input, the smaller the total distance will be. A more interesting point to look at is the number of updates we use to recalculate new cluster locations. When we include only 1 update, we get a very large number for total distance. The more updates we include in our program, the more accurate the optimal cluster locations will be, hence the shorter the total distance. We can see that the total distance converges to a specific value, which is the optimal value.

Now it might be interesting to check if we get this same value for total distance if we let the program run it again. When we implement only 1 update, the program assigns random values to all cluster locations. This cluster location can be very inefficient; hence we get large total distances. We will check for 5 clusters what the consequence to the total distance is if the number of observations (running the program again) increases.

Total distance (km) for different observations and updates with 5 Clusters					
	1 Update	3 Updates	5 Updates	10 Updates	30 Updates
Observation 1	4.0564979E7	2.3312876E7	2.1708824E7	1.1496157E7	1.1496157E7
Observation 2	3.1659775E7	1.5722566E7	1.3322691E7	1.1496157E7	1.1496157E7
Observation 3	3.5408118E7	1.6596357E7	1.5528651E7	1.1495441E7	1.1496157E7
Observation 4	2.9565324E7	2.3921464E7	1.1604065E7	1.1496157E7	1.1496157E7

As we can see the total distances at 1 update differ a lot from each other. Interesting to see is that after 30 iterations the total distance converges to 1.1496157E7 km. We can conclude here that if we implement enough updates in our code, we can use the total distance give at 1000 updates.

It seems that we managed to find better cluster locations for our shipments. The total distance at 648 clusters converges to 55,901 Km. Compared to the total distance of 97,942 Km we found in Clustering I, we can conclude that we managed to find a decrease of nearly 43%.

Clustering III

In Clustering II we found a method to minimize the total distance of all shipments from the location to the assigned cluster. For Clustering III we will try to find alternative ways of clustering the shipments. We think that we can minimize the total distances even more when we center cluster at locations that are most frequently used. In the algorithm of Clustering II we looked at all distances which are widely spread over four countries. For some large numbers of k we found that the total distance reduced a lot. With our new method of optimizing we can increase the efficiency of the most important clusters. A downside of using this method is that we might remote an isolated area so much that the total distance will increase again. The goal of this algorithm would be to gain more efficiency from the most occurring areas hedged against the increase in distance from isolated areas. Unfortunately, we did not manage to get the code for this advanced algorithm working within the past 2 weeks. If we could extend the project for a longer time it would be very interesting to see our new results in shorter distance.

Another idea to think about regarding the logistic provider is not to cluster the shipments on total distance, but to use a different parameter. It would be more useful for our logistic partner to sort the packages on volume such that our partner can choose which method of transportation to use for different types of packages. The volume of a package can heavily affect the costs of transporting it. We can create clusters that deal with the biggest packages. The warehouses of small packages can be smaller; hence our partner can save costs in the sense of transportation costs and fixed costs for the warehouses.

Business Summary

In the business summary we will give advice on a managerial level about how to possible adapt in order to gain possible savings. We will start with advice on the seasonality of the data, thus, how to adapt to certain week days, which may at a later stage be adapted to certain months, in order to save.

Our results show a general decrease of activity on Mondays. Thus, we suggest that the company would decrease the amount of hiring for Mondays at least on the lanes we investigated, being from cluster 2 to clusters 547,549,551,552 and 611. We found that weight, volume and amount of shipments can be compared across lanes for the Monday, which means that the company could set one policy for Mondays based on one lane and extrapolate that policy to the other lanes, without too much loss in savings.

We also found a general increase in activity on Wednesdays and Thursdays, on the lanes we had no increase, those values were not with certainty a decrease, but could be an increase too. So, for Wednesdays and Thursdays we recommend to in general have more labour active and more methods of transport ready. So, on Mondays decrease and on Wednesdays and Thursdays increase compared to what is present now. Furthermore, Thursday are comparable across different transport lanes, just as the Monday. Thus, policies for Thursdays can be used for at least the different lanes we used. For Wednesdays we cannot do this, the only policy that could be justifiably extrapolated is a policy based on weight.

We also found an overall increase in activity on Fridays, which makes sense since this right before the weekend. The increase is not significant however, which means that the increase in activity could be a decrease actually. So, in general we would recommend a similar policy on Fridays as on Wednesdays and Thursdays. The advantage that Friday has, is that weight, volume and amount of shipments can be compared across different routes without loss of savings.

The only day of the week we have not yet discussed is Tuesday. However, no clear general results are visible for Tuesday. Tuesdays are different per route and even a relation between volume, weight and amount of shipments is not necessarily present. Tuesdays are comparable across different routes; thus, we recommend a policy for more routes, but differing in shipment, weight and volume. We cannot give clear instructions to increase or decrease.

Another important result is that for certain lanes a dependency on previous activities is present, higher activity yesterday means higher activity today. In our case these lanes are the shipments for lane 2-547 and weight and shipment for lane 2-552, with a dependency of respectively 5,3 and 5 days. So, if yesterday had high activity than today will probably be higher than normal and the company could hire extra labour or increased methods of transport to handle the extra activity.

Total distance (km) for different numbers of Clusters and updates				
Number of Clusters	Updates			
	1	3	10	1000
648	3844897.0	376080.0	206927.0	55901.0

The new locations of the clusters as explained and calculated in Clustering II has surprisingly good results and can save a lot of costs for our partner regarding shipments. With the current cluster locations, the total distance of each shipment location to its cluster location is 90189 Km (as calculated in Clustering I). The new method of clustering managed to allocate the clusters in such a way that the new total distance can be reduced to 55901 Km. This is a reduction of approximately 43%. Although we have a good result, we have to make some notes on these numbers. The distances we used are absolute distances what implies that we did not consider any driving distances. Also, the transportation of packages from the mainland to the UK will cause an increase in total distance. Another side note is that the new cluster locations are placed at random locations which implies that one of the 648 clusters is located in the North Sea. It is not very likely that our partner is looking for a cluster in the middle of a sea, since it will increase the costs of transportation as our partner should implement new methods of transportation.

At this stage of research, we would not advice our partner to implement the new clustering, since we did not take look at real world problems yet. The most important result from our research is that the clustering our partner is currently using is not optimal. With our report we proved that there are opportunities for our client to improve their method of clustering.