Chapter 1

Computational Results

This chapter will present the results from the following experiments:

- Wild operator evaluation
- Operators evaluations
- ALNS evaluation

Before going into the experimental results we will describe the experimental setup. This icludes the techical side aswell as the setup of the analytical experiments themselves. After that we will explain the generation ofthe instances used in the final result section.

1.1 Experimental Setup

In this section we describe the technical setup of the experiments aswell as the Analytical setup of our eperiments.

1.1.1 Technical Setup

The computational experiments in this paper is run on two different computers. The more demanding experiments are run on a 64-bit Windows 8 computer with a 2.7 Ghz i7-8500 processor and 16GB RAM. We will shorten the name of this computer to "Windows computer". The less demanding experiments are run on a 64-bit Ubuntu 18.04 computer with a 1.8 Ghz quad core i7-8550u processor and 16GB RAM. We call this computer for short "Ubuntu computer".

The instance generator described in the next section was implemented in Java (version number). The mathematical model from section 2 is setup in AMPL (version number), using the Gurobi solver. All AMPL experiments were run on the Windows computer. The ALNS heuristics are implemented in Java (verison number). These experiments were run partly on the Ubuntu and partly on the Windows computer. The statistical experiments are performed in Matlab (version number) on the Ubuntu computer.

1.1.2 Analytics Setup

In section 3 we described seven operators, aswell as a wild operator, some invented in this paper and others based on known ALNS heuristics. For our testing we generated five instance sets of each five instances, totally 25 instance. Each instance set has 5 instances with sizes ranging from $4 - 150$ orders. The sizes used in this paper are 4, 12, 35, 80 and 150 orders. While testing the wild operator, we used 5 of these instances of each of the sizes. The test was run 10 times using all of the seven described operators from section 3. The results from these To analyse the performance of each of the seven operators, 5 reasonably sized instances was solved 10 times using each of the $2^7 = 128$ combinations of operators. To determine which of the operators influence the result we have performed several statistical tests, including ANOVA and regressional analysis, and t-tests. After the analysis of the operators, 5 heuristics are chosen for further testing. For these tests we compare the results toward the best known solution from AMPL and ..... TODO. We use a 95% confidence level for all statistical experiments in this paper.

1.2 Instances

The instance generator was created based on real data from an anonumous costomer of 4flow. Following, we will describe how we designed the generator and how we generated the instances used in the analytical part of the paper.

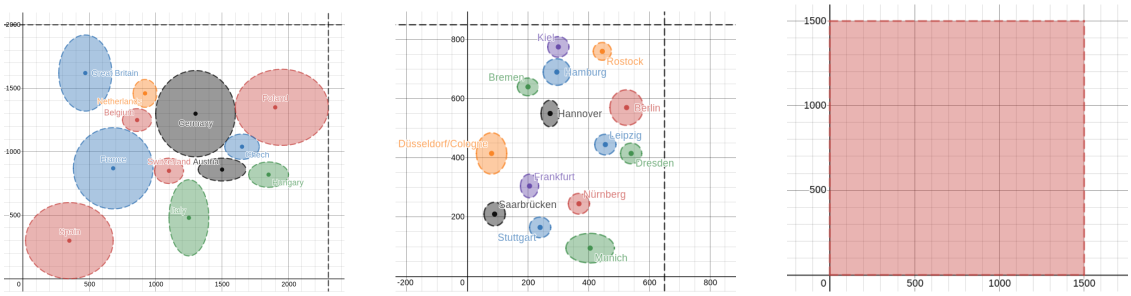
1.2.1 Generate Instances based on real data

The 4flow data gives information about the number of orders $|N|$, locations $|L|$, factories $|F|$. The amount of vehicles $|V|$ are kept large compared to the amount of orders to make sure there are enough but not too many vehicles. A solution that uses all but one vehicle is the ideal here and we found $|N|/2 \leq |V| \leq 2/3|N|$ to be fitting. $|N|, |V|, |L|, |F|$ are given as input to the generator. In addition the data from 4flow gave us information about the size of a vehicle and its compatabilities. Some vehicles might have cooling possibilities amd some speical equipment required for transport of special goods or equipment required at the pickup or delivery location etc.). Other information aquired by the data was travel distances, cost structure etc.

To keep the instances feasible but still as realistic as possbile it makes sense to limit the data to different possibilities. Our data was generated with the following properties:

- Orders are assigned to pickup and delivery locations randomly. Orders assigned to the same location are given the same stop L_s
- Each delivery location is assigned to a factory at random N_f.
- Each location aswell as each order is assigned a special property with 5% probability. This will decide which vehicle can pickup which order, N_v^P and N_v^D
- We let the vehicles types be split up in 3 different vehicle types, small, medium, large, each with extected capacitied and capabilities.

Figure 1.1: Area of random point generation, colored parts are area of possible point generation



(a) Europe **(b)** Germany **(c)** Uniform

- Large vehicle: slower but compatible with all locations and orders, with $Q_v^{kg} = 24'$ and $Q_v^{vol} = 102$
- Medium vehicle: medium fast and compatible with all locations but not orders with special compatability, with $Q_v^{kg} = 18'$ and $Q_v^{vol} = 71$
- Small vehicle: fastest but not compatible with special locations and special orders, with $Q_v^{kg} = 12'$ and $Q_v^{vol} = 55$

- The distances were calculated using pytagoras on the randomly generated points described in the next section
- The time T_{ijv} were scaled with 60% of the travel distance, added with a random variation of $+/-10\%$ of the travel distance, multiplied by the speed of the vehicle (slow: $* = 105\%$, medium: $* = 102.5\%$).
- The cost matrices $C_{v\alpha\beta}^{km}$, $C_{v\alpha\beta}^{kg}$, $C_{v\alpha\beta}^{fix}$ were based on a real 4flow cost matrix, scaled to the size of the instance and to the size of the vehicle.
- The cost of no transport was set to a minimum lower bound scaled based on the weight/volume/distance of the order.
- Time windows $[\underline{T_{ip}}, \overline{T_{ip}}]$ were generated randomly based on typical factory opening hours. 1-2 timewindows per day, and 3-7 days per week based on the instance size.

Random locations based on real georaphical data

Most of 4flow's customers are based either in Germany or in Europe. To make the instance generator as realistic as possible we have decided to split the instances into 3 geographical types; European, German and uniform geographically distributed locations. We made 2 maps based on real scale approximations of geographical data from National Geographics, in km. To simplify we have sticked to geographical points with an eliptic uniformly distributed area surrounding the point to represent a country or a city. figure 1.1 illustrates the areas of possible locations used in the generator. Larger elipses are more likely to be selected by the generator than smaller elipses.

For the selected elipse a point was selected within the elipse at random with a uniform distribution. For figure 1.1c points were generated at random within the limits shown. From our 5 instance sets, two were generated using figure 1.1a on the preceding page, two with figure 1.1b on the previous page and one with the uniform distribution from figure 1.1c on the preceding page. If a point belong in the same factory as a previously generated point, that point was generated within a reasonable radius of three kilometer.

1.2.2 Generated Instances

For testing our algorithm in this paper, 5 instance sets of each 5 instances of varying sizes were generated. Among the 5 instance sets, 2 were generated based on figure 1.1a on the previous page, two based on figure 1.1b on the preceding page and one based on figure 1.1c on the previous page. Each instance set contains the following sizes: TODO: Describe the instances I have generated and how I frefer to them in the paper. Finish the presentation of the results first to, know how I do this.

1.3 Initial Results

In the first two parts of this section we will present the data from our experiments, that leads to the final composition of our model. The results for the final algorithm are presented in the third part of this section.

1.3.1 Evaluation of the wild algorithm

To help our algorithm out of a stuck position we designed a wild algorithm described in section section **??** on page ??. To evaluate the implementaion of this operator we have decided to evaluate the result of running the complete ALNS with all heuristics included with three different modifications. First we run the algorithm without any modification. Secondly we ran the algorithm including the wild algorithm. Lastly we ran the algorithm with a modification that resets the algorithm each time we get stuck and start from a new random solution. Comparing these three options towards eachother will help us evaluate if our wild algorithm is helping us in general and see if our wild algorithm is different than doing a reset and just starting from a new solution somewhere.

Results of running ALNS

We performed runs on one instance set based on the european locations of 5 representative sizes of our problem. To represent each instance we have chosen the following parameters:

- $\wp_n$ represent the amount of orders where n is the instance order size
- $\Re_v$ represent the amount of vehicles where v is the instance vehicle amount
- $\Im_l$ represent the location amount where l is the instance location amount

As an example $\wp_{10}\Re_4\Im_7$ would be an instance with ten orders, four vehicles and seven locations.

The result from these three runs are summarized in table 1.8 on page 14.

Table 1.1: ALNS with all heuristics using 3 different reset algorithms

Instance	Init.	AObj no	AObj res	AObj wld	BObj no	BObj res	BObj wld	rt no	rt res	rt wld
$\wp_4\Re_3\Im_7$	609 680.26	3 444.67	3 444.67	3 444.67	3 444.67	3 444.67	3 444.67	0.25	0.20	0.32
$\wp_{12}\Re_7\Im_9$	1 023 745.51	149 692.64	154 832.88	149 692.39	149 692.39	149 692.39	149 692.39	0.25	0.51	0.72
$\wp_{35}\Re_{20}\Im_{22}$	2 682 067.86	10 639.09	10 849.54	10 294.25	10 404.91	10 358.57	9 880.20	0.25	1.61	3.86
$\wp_{80}\Re_{45}\Im_{45}$	6 422 128.60	22 262.23	25 802.91	21 339.57	20 761.22	21 777.43	20 972.44	0.25	8.05	25.26
$\wp_{150}\Re_{80}\Im_{85}$	12 059 380.27	40 667.18	38 313.04	34 035.86	34 316.00	34 345.00	33 008.34	0.25	48.92	105.26

The columns contains in order: the instance size, the initial objective value (no transport), the average objective found over 10 runs first for no stuck algorithm, then for random restart, then our wild algorithm, then follows the best objective found for no stuck algorithm, then random restart, then our wild operator, finally the average run time of for again no algorithm, restart and wild algorithm.

Observations of results from wild algorithm evaluation

The table 1.8 on page 14 shows that on average using our wild algorithm clearly outperforms both the restart algorithm and not including a reset algorithm. We see that the average result for all instance sizes are lowest using our wild algorithm. In addition our wild algorithm outperforms the others for all instances in finding the best solution, except for $\wp_{80}\Re_{45}\Im_{45}$, where no algorithm finds a better best solution. This is however probably due to the algorithm ending in a lucky neighbourhood and since the wild operator is vastly improving the solution on average we still think its safe to conclude that it is the better option.

With regards to the running time the random restart is here the fastest, followed by no algorithm, and lastly by our wild algorithm. This is expected as the wild operator adds additional iterations to the search every time it gets stuck. Also the random restart is provided with the solutions and does not generate them itsself. This would add significant time to this algorithm and the increased performance is making it worth it.

1.3.2 Initial evaluation of heuristics

Different heuristics have different strengths and weaknesses. Some heuristics are not performing well on their own, but work very well in combination with other heuristics. Other heuristics are stron on their own, and their performance is prohibited by other heuristics, or not getting enough running time when too many heuristics are included.

To find the right combination of heuristics we have done an evaluation of their performance to search for the best possible combination of operators. To do this selected the instance set with a representative size of ... and ran our algorithm, without the wild algorithm adaptation, with each combination of operators. This results in $2^7 = 128$ each running 10 times. To be able to compare the results from differnt instances we calculated the improvements from the initial solution which resulted in the improvement from the best solution found during the 10 runs, and an average improvement from the initial solution from the 10 runs of each combination. Each observation was also multiplied by 1000 to make the analysis more readable.

We evaluated the data resulting from the different combination in three steps:

- In the first part we ranked the combinations based on the best improvement and the average improvement

- In the second part we ran ANOVA and regression analysis to see which heuristics have a significant impact on the result
- In the third part we run t-tests to see if certain heuristics in combination with others have a positive or a negative impact on the final result

Ranking the combinations

We shorten the names of each heuristic and will refer to them for here on as components. The components are assigned the following names:

- C1: this is the swap heuristic described in section **??** on page ??
- C2: exchange heuristic from section **??** on page ??
- C3: is the 2-opt heuristic in section **??** on page ??
- C4: is the random removal first fit insertion heuristic from section **??** on page ??
- C5: clustering heuristic described in section **??** on page ??
- C6: this is the worst removal and greedy insert heuristic in section **??** on page ??
- C7: is the shaw removal and regret-3 insertion heuristic described in section **??** on page ??

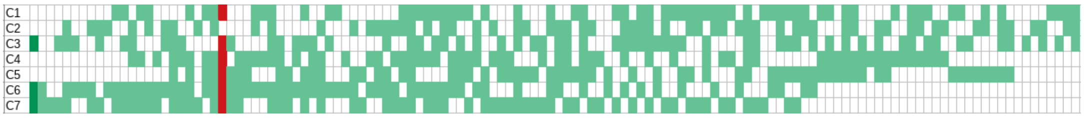
Figure 1.2 on the facing page shows the ranking of each combination of components from left to right. A colored tile indicates that the current component is in use. The further a combination is to the left, the higher improvement was from the initial solution compared to the other combinations.

The figure 1.2a shows the combination of components ranked based on the average improvment over the 10 runs for that combination. Then figure 1.2b shows the combination of components ranked based on the best improvement overall during the 10 runs. Finally figure 1.2c shows the ranking of the average improvement + the best improvement to see which combinations performs best overall.

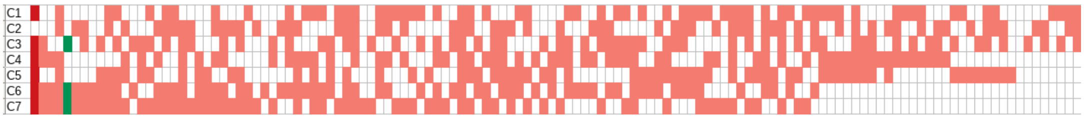
ANOVA and regression analysis

The ranking gives us an overview over which components are working and is always part of a good combination. It also gives us information on which components are not performing well overall and which combination of components are not working well. However, ANOVA and regression analysis can help us evaluate which components have a significant positive impact on the result. Table 1.2 on the next page shows the results of ANOVA (III) analysis performed using each component as a source of variance. We extended the model with the instances as random effects to give us a better explanation of the result and less noise in the model. The result from the ANOVA can give us information on which component has a significant impact on the result. We did one ANOVA analysis for the average improvement shown in figure 1.2a, and one for the best improvement in figure 1.2b.

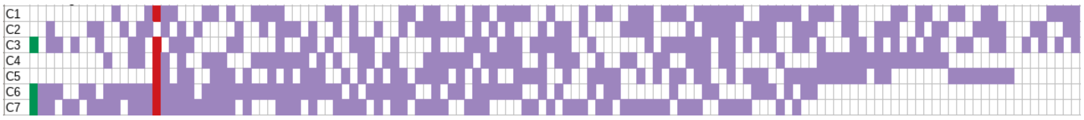
Figure 1.2: Ranking of improvements from initial solution, the highlighted tiles indicate use of a component. Strongly highlighted combinations represent the best combination in one of the figures.



(a) Average improvement



(b) Best improvement



(c) Best+Average improvement

Table 1.2: Analysis of variance

Source	Sum sq.	df	Mean sq.	F	P>F
C1	0.052	1	0.052	0.09	0.7682
C2	0.748	1	0.748	1.26	0.2628
C3	0.038	1	0.038	0.06	0.8017
C4	22.432	1	22.432	37.66	0
C5	12.734	1	12.734	21.38	0
C6	117.945	1	117.945	198	0
C7	112.406	1	112.406	188.7	0
Inst	371.704	4	92.926	156	0
Error	350.257	588	0.596		
Total	975.207	599			

(a) Average improvement statistics

Source	Sum sq.	df	Mean sq.	F	P>F
C1	0.075	1	0.0745	0.23	0.6306
C2	0.195	1	0.1949	0.61	0.4369
C3	0	1	0.0001	0	0.988
C4	8.372	1	8.3718	26	0
C5	3.679	1	3.6789	11.43	0.0008
C6	66.236	1	66.2357	205.71	0
C7	67.081	1	67.0813	208.33	0
Inst	344.971	4	86.2428	267.84	0
Error	189.331	588	0.322		
Total	671.378	599			

(b) Best improvement statistics

The columns contains in order: the source of the variability and for each source the sum of the sqares, the degrees of freedom, the mean squares, the F-statistic and the p-value.

Table 1.3: Results of multiple linear regression model

Term	Estimate	SE	tStat	pValue
Intercept	994.58	0.11713	8491.5	0
C1	-0.018582	0.063017	-0.29487	0.7682
C2	-0.07063	0.063017	1.1208	0.26283
C3	-0.01583	0.063017	-0.2512	0.80175
C4	0.39108	0.063729	6.1366	1.5502e-09
C5	-0.29466	0.063729	-4.6236	4.6407e-06
C6	0.89676	0.063729	14.071	5.7098e-39
C7	0.87544	0.063729	13.737	1.923e-37
Inst1	0.50928	0.099639	5.1113	4.3337e-07
Inst2	-0.051052	0.099639	-0.51237	0.60859
Inst3	1.6601	0.099639	16.662	2.4375e-51
Inst4	1.755	0.099639	17.614	4.4128e-56

(a) Average improvement statistics, $R^2 = 0.641$

Term	Estimate	SE	tStat	pValue
Intercept	994.97	0.086114	11554	0
C1	0.022291	0.046332	0.48112	0.63061
C2	-0.036047	0.046332	-0.77801	0.43687
C3	-0.00069693	0.046332	-0.015042	0.988
C4	0.23891	0.046855	5.099	4.6112e-07
C5	-0.15838	0.046855	-3.3801	0.00077249
C6	0.67202	0.046855	14.342	3.2e-40
C7	0.67629	0.046855	14.434	1.2062e-40
Inst1	0.64176	0.073257	8.7605	2.0732e-17
Inst2	0.16783	0.073257	2.291	0.022316
Inst3	1.837	0.073257	25.077	6.2887e-95
Inst4	1.6686	0.073257	22.778	8.2582e-83

(b) Best improvement statistics, $R^2 = 0.718$

The columns contains in order: the term and for each term the coefficient estimate, the standard error of the coefficients, t-statistics to test if the term is significant, and the p-value

In addition to this we performed a multiple linear regression in table 1.3 on page 8 on the same data as the ANOVA. The regressional analysis gives us insight into wether a component is positively or negatively influencing the result aswell as how well the components explain the result through the R^2. Also here we used the Instances as a random effect, and the final instance is when all Instance terms are 0. We did two different

Observations from initial evaluation of heuristics

The results from table 1.2 and table 1.3 split our components into two groups. The first obvious group are the significant heuristics. The ANOVA results from table 1.2a and table 1.2b on the preceding page tells us that four components, C4-C7, have a significant impact on the result. However from 1.3a and

Regarding the remaining heursitics, C1, C2, and C3, we can safely conclude that they are not contributing singificantly on the result on their own. However we cannot conclude that they do not have a positive effect in combination with the significant heuristics. We therefore conclude that we need further testing to know if these heuristics have a positive or negative influence on the result. We refer to the heuristics C1,C2,C3 and C5 as the undecided group, or G in our further evaluation of the heuristics.

1.3.3 Further evaluation of heuristics

To further analyse the performance of the heuristics we want to analyse how the heuristics are performing as a group (G) to see if they have an effect on the results. The result from section 1.3.2 on page 5 tells us that it is out of the question to use any combination of components where only undecided components are used. These components will alone have a negative impact on the result but it is still possible that using them combined with the significant components could have a positive impact on the result We have therefore removed the observations where the undecided group appear alone in the following testing. We want to test if using one or several of the undecided heuristics in combination with other heuristics have a positive impact on the result.

Table 1.4: Results of T-tests on the undecided group mean vs no undecided heuristics

Populations	Impr type	Tail	H-stat	p-value	tStat	conf-int
$P_H - P_N$	Avrg	both	0	0.5187	-0.6458	-0.3785 – 0.1912
$P_H - P_N$	Avrg	right	0	0.7407	-0.6458	-0.3326 – inf
$P_H - P_N$	Avrg	left	0	0.2593	-0.6458	-inf – 0.1453
$P_H - P_N$	Best	both	0	0.6903	0.3986	-0.2174 – 0.3281
$P_H - P_N$	Best	right	0	0.3452	0.3986	-0.1734 – inf
$P_H - P_N$	Best	left	0	0.6548	0.3986	-inf – 0.2841

The columns contain in order: the populations tested agains eachother, type of data tested(average or best improvement), tail which determines the alternative hypothesis, h-value (1 rejects the null hypothesis, 0 failure to reject), p-value of the test, t-Statistic of the test, confidence interval for the true population mean

We have done this in two parts. We first wanted to see if the components as a group has a positive or negative influence on the result. To test this we have done T-tests to compare the mean of the population where we combine the undecided heuristics with some significant heuristic, to the mean of the pooulation when we are not using an undecided heuristic. Then we wanted to see if using the undecided group with specific combinations of the significant components have significant impact on the model. We did this using further ANOVA (III) statistical analysis and multi linear regression model.

Evaluation of undecided components as a group

The first thing we did was to run a T-test on that compares the mean of the population which include some combination of the undecided group heuristics and the significant heuristics, towards the population that does not contain any heuristic from the undecided group. To represent the population without any of the undecided heuristics we have the used the parameter P_N and to represent the population with some combination of the undecided heuristics and the significant heuristics we have used the parameter P_H. The results from the T-test are summarized in

We continued the testing of the undecided group by performing ANOVA (III) analysis on different combinations of the undecided group and the significant heuristics. We did this to see if a combination of the undecided group and the significant heuristics could help explain the variations in the result and to see if a combination of some or all of the significant heuristics work better than others.

The results are summarized in table 1.5 and table 1.6 on the following page. As an example a source of G+C4, contains all observations where at least one of the undecided components are combined exclusively with C4.

Observations from results of the heuristics further evaluation

The results from table 1.4 on the previous page tells us that we cannot reject the H_0 null hypothesis that the mean of the two populations are different using a 95% confidence interval, for neither average or best improvement. This tells us that the components from the undecided group either have no significant impact on the result, or that the effect from some are nulling out the others. Further testing is needed to make any further conclusions.

Table 1.5: Analysis of variance with the undecided group in combination with the significant heuristics

Source	Sum sq.	df	Mean sq.	F	P>F
G+C4	19.024	1	19.0236	538.34	0
G+C6	0	1	0	0	0.9859
G+C7	0.005	1	0.0045	0.13	0.7204
G+C4+C6	0.059	1	0.0589	1.67	0.1973
G+C4+C7	0.001	1	0.0009	0.03	0.8722
G+C6+C7	0.22	1	0.2202	6.23	0.0128
G+C4+C6+C7	0.166	1	0.1657	4.69	0.0308
Inst	309.022	4	77.2555	2186.21	0
Error	19.365	548	0.0353		
Total	385.274	559			

(a) Average improvement statistics

Source	Sum sq.	df	Mean sq.	F	P>F
G+C4	12.879	1	12.8793	609.42	0
G+C6	0.131	1	0.1307	6.19	0.0132
G+C7	0.239	1	0.2389	11.31	0.0008
G+C4+C6	0.223	1	0.2227	10.54	0.0012
G+C4+C7	0.153	1	0.1525	7.22	0.0074
G+C6+C7	0.42	1	0.4195	19.85	0
G+C4+C6+C7	0.395	1	0.395	18.69	0
Inst	295.738	4	73.9346	3498.45	0
Error	11.581	548	0.0211		
Total	352.547	559			

(b) Best improvement statistics

The columns contains in order: the source of the variability and for each source the sum of the sqares, the degrees of freedom, the mean squares, the F-statistic and the p-value.

Table 1.6: Results of multiple linear regression model with the undecided group in combination with the significant heuristics

Term	Estimate	SE	tStat	pValue
Intercept	995.85	0.035525	28032	0
G+C4	-0.89285	0.038481	-23.202	1.8003e-83
G+C6	0.00068095	0.038481	0.017696	0.98589
G+C7	0.013782	0.038481	0.35815	0.72037
G+C4+C6	0.049672	0.038481	1.2908	0.19731
G+C4+C7	-0.0061945	0.038481	-0.16097	0.87217
G+C6+C7	0.096062	0.038481	2.4963	0.012841
G+C4+C6+C7	0.083317	0.038481	2.1651	0.030808
Inst1	0.53458	0.02512	21.281	1.0604e-73
Inst2	0.031671	0.02512	1.2608	0.20793
Inst3	1.71	0.02512	68.073	1.6221e-269
Inst4	1.5959	0.02512	63.531	6.3233e-255

(a) Average improvement statistics, $R^2 = 0.95$

Term	Estimate	SE	tStat	pValue
Intercept	995.88	0.027473	36249	0
G+C4	-0.73464	0.029759	-24.686	5.012e-91
G+C6	0.07402	0.029759	2.4873	0.013167
G+C7	0.10006	0.029759	3.3625	0.00082635
G+C4+C6	0.096599	0.029759	3.246	0.0012417
G+C4+C7	0.079948	0.029759	2.6865	0.0074396
G+C6+C7	0.13259	0.029759	4.4554	1.0157e-05
G+C4+C6+C7	0.12865	0.029759	4.3232	1.8269e-05
Inst1	0.70773	0.019426	36.432	1.6371e-148
Inst2	0.27728	0.019426	14.273	1.5823e-39
Inst3	1.873	0.019426	96.415	0
Inst4	1.5781	0.019426	81.237	8.6749e-308

(b) Best improvement statistics, $R^2 = 0.967$

The columns contains in order: the term and for each term the coefficient estimate, the standard error of the coefficients, t-statistics to test if the term is significant, and the p-value

Table 1.5 and table 1.6 gives us further insight into our model. First of all the result from table 1.5a on the previous page tells us that only two combinations of the undecided group and the significant variables have a significant impact on the result using a 95% confidence interval. Using some heuristics from the undecided group combined with component C6 and C7, aswell as C4 has a significant impact on the average improvement result. From table 1.6a we also see that the combinations are also getting a positive coefficient. We also observe that the $R^2 = 0.95$ is very high so this model explains the results very well and this supports the use of these heuristic combination in regards to the average improvement.

The same combinations are significant and positive in reagards to the best improvement tabels table 1.5b on the previous page and table 1.6b. And even though more combinations are significant for best improvement the combinations with highest positive estimated coefficients are still the combinations with C6 and C7, and C6 C7 and C4. We also observe here the increased $R^2 = 0.967$ which supports our conclusion that this should consistently lead to good results on best and average improvement.

The results from the further testing tells us that there are combinations of the undecided group and the significant components that have a positive significant impact on the result of both average and best improvement. It supports our results from section 1.3.2 on page 5 of significant components C6, C7 and C4 and leads us to conclude that there might be a positive influence from the undecided group components, however further testing is nessecary to determine which components should be included.

1.3.4 Evaluation of individual components

Until now, the heuristic components C4, C6 and C7 have been proven significant. The components C1, C2, C3 and C5 have been proven significant in combination with the significant components but not alone. We continue refering to them as the undecided group components or G.

We want to figure out which of the components in the undecided group, if any, have a positive, or negative, influence on the result. To do this we performed t-tests to check if an undecided component is significantly improving or decreasing the best and average improvement. Like in section 1.3.3 on page 8 it is out of the question to use any combination of heuristics where only undecided heuristics are used, so we remove these observations from the data also in these tests. In our tests we test if the mean of the populations where we combine the undecided heuristics with some significant heuristic is significantly different than the pooulation when we are not using an undecided heuristic.

T-tests of individual components

Similar to the previous section we use the parameter $P_{NA}^{C_i}$ to refer to the population without a specific component C_i for the average improvement, indicated by A, and the parameter $P_{HB}^{C_i}$ as the population including the component C_i for the best improvement, indicated by B. Here C_i represents one of the component heuristics described in the previous section. We did the test for all 4 of the undecided heuristics from the previous section C_1, C_2, C_3 and C_5. The results from the T-tests are summarized in 1.7.

Table 1.7: Results of T-tests on individual components

Populations	Tail	H-stat	p-value	tStat	conf-int
$P_{HA}^{C_1} - P_{NA}^{C_1}$	both	0	0.5427	-0.6090	-0.1580 – 0.0727
$P_{HA}^{C_1} - P_{NA}^{C_1}$	right	0	0.7286	-0.6090	-0.1325 – inf
$P_{HA}^{C_1} - P_{NA}^{C_1}$	left	0	0.2714	-0.6090	-inf – 0.0472
$P_{HB}^{C_1} - P_{NB}^{C_1}$	both	0	0.9240	-0.0955	-0.1428 – 0.1296
$P_{HB}^{C_1} - P_{NB}^{C_1}$	right	0	0.5380	-0.0955	-0.1208 – inf
$P_{HB}^{C_1} - P_{NB}^{C_1}$	left	0	0.1076	-0.0955	-inf – 0.1076
$P_{HA}^{C_2} - P_{NA}^{C_2}$	both	1	3.4853e-10	-6.5098	-0.0692 – 0.0412
$P_{HA}^{C_2} - P_{NA}^{C_2}$	right	0	1.0000	-6.5098	-0.0661 – inf
$P_{HA}^{C_2} - P_{NA}^{C_2}$	left	1	1.7426e-10	-6.5098	-inf – 0.0443
$P_{HB}^{C_2} - P_{NB}^{C_2}$	both	1	4.5425e-10	-3.5486	-0.0370 – 0.0106
$P_{HB}^{C_2} - P_{NB}^{C_2}$	right	0	0.9998	-3.5486	-0.0349 – inf
$P_{HB}^{C_2} - P_{NB}^{C_2}$	left	1	2.2712e-04	-3.5486	-inf – 0.0127
$P_{HA}^{C_3} - P_{NA}^{C_3}$	both	1	0.0244	-2.2635	-0.0276 – -0.0043
$P_{HA}^{C_3} - P_{NA}^{C_3}$	right	0	0.9878	-2.2635	-0.0250 – inf
$P_{HA}^{C_3} - P_{NA}^{C_3}$	left	1	0.0122	-2.2635	-inf – -0.0069
$P_{HB}^{C_3} - P_{NB}^{C_3}$	both	0	0.6271	-0.4864	-0.0134 – 0.0081
$P_{HB}^{C_3} - P_{NB}^{C_3}$	right	0	0.6865	-0.4864	-0.0116 – inf
$P_{HB}^{C_3} - P_{NB}^{C_3}$	left	0	0.3135	-0.4864	-inf – 0.0063
$P_{HA}^{C_5} - P_{NA}^{C_5}$	both	0	0.4702	-0.7231	-0.0301 – 0.0118
$P_{HA}^{C_5} - P_{NA}^{C_5}$	right	0	0.7649	-0.7231	-0.0255 – inf
$P_{HA}^{C_5} - P_{NA}^{C_5}$	left	0	0.2351	-0.7231	-inf – 0.0071
$P_{HB}^{C_5} - P_{NB}^{C_5}$	both	1	1.7956e-04	-3.7972	-0.0167 – 0.0528
$P_{HB}^{C_5} - P_{NB}^{C_5}$	right	1	8.9779e-05	-3.7972	-0.0197 – inf
$P_{HB}^{C_5} - P_{NB}^{C_5}$	left	0	0.9999	-3.7972	-inf – 0.0499

The columns contain in order: the populations tested agains eachother, type of data tested(average or best improvement), tail which determines the alternative hypothesis, h-value (1 rejects the null hypothesis, 0 failure to reject), p-value of the test, t-Statistic of the test, confidence interval for the true population mean

Observations from the results of the individual component evaluations

The first thing we see from the T-tests is that we were right in assuming that the effect of the components were cancelling eachother out. We go though each of the components results from table 1.7 on the next page here:

- C1: The population mean of both best and average improvement is not significantly different using a 95% confidence interval. It follows then that the tails are also not significantly different.

- C2: We reject the null hypothesis that the means are equal using a 95% significanse interval for the average and best improvement regarding this component. The left tail alternative hypothesis' that the means of $P_{HA}^{C_2}$ and P_{HB}^{C2} is lower than $P_{NA}^{C_2}$ and $P_{NB}^{C_2}$ are accepted, while the right tail alternative hypothesis is rejected for both best and average improvement.

- C3: The null hypothesis that the population mean of $P_{HA}^{C_3}$ is equal to the population mean of $P_{NA}^{C_3}$ is rejected and accepted for $P_{HB}^{C_3}$ and $P_{NB}^{C_3}$. The left tail alternative hypothesis that the means of $P_{HA}^{C_3}$ is with 95% confidence lower than $P_{NA}^{C_3}$, is accepted.

- C5: The null hypothesis is rejected for $P_{HB}^{C_5}$ and $P_{NB}^{C_5}$ and accepted for $P_{HA}^{C_5}$ and $P_{NA}^{C_5}$. The alternative hypothesis of the right tail of the best improvement, that the mean is significantly higher in $P_{HB}^{C_5}$ than $P_{NB}^{C_5}$ is accepted.

The results summarized above tells us that using C1 will have no effect on the outcome of the result. Also C2 and C3 are significantly decreasing the average for both components and also for best for C2. This indicates that it is not beneficial to include these operators in a model. Finally C5 is not affecting the average improvement however it is positively effecting the best improvement, indicating that it could be beneficial to include this component.

1.3.5 Deciding on the final model composition

The results from section ?? on page ?? tells us that our final model should include our wild algorithm. This will influence our running time a little but give us much more reliable results.

As for selecting heuristics or components to include, the results from section 1.3.2 and section 1.3.3 on page 8, indicates that we need to include C4, C6 and C7 in our final model. We observed from our testing of the wild algorithm that the running time of C1 is significantly lower than our other operators figure **??** on page ??. Even though section 1.3.4 on page 10 showed us that C1 had no effect on the model, the low running time is leading us to rather include this operator than not to get us a little bit better running times. It will as section 1.3.4 on page 10 also showed not negatively effect our results. C5 is also not effecting the average improvement of our model, but it also has a good running time, and it significantly effects the result of the best improvement found. This tells us that we should include this operator in our final composition.

The results from section 1.3.3 on page 8 indicated that a combination of C4, C6 and C7 + some of combination of the undecided group components is significantly

Table 1.8: Model performance in European Instance sets

			Set 1				Set 2			
Ord	Veh	Loc	Avrg Imp	Best Imp	Best knwn	Dual gap	Avrg Imp	Best Imp	Best knwn	Dual gap
4	3	7	99.44%	99.44%	99.44%	0.00%				

The columns contains in order: the instance size, the initial objective value (no transport), the average objective found over 10 runs first for no stuck algorithm, then for random restart, then our wild algorithm, then follows the best objective found for no stuck algorithm, then random restart, then our wild operator, finally the average run time of for again no algorithm, restart and wild algorithm.

effecting both the average and best improvement. Therefore our final model composition will be the ALNS algorithm with the wild algorithm and heuristic components C1, C4, C5, C6 and C7.

1.4 Final results

After deciding on a model best suited to our problem we did a final run of our algorithm using composition described in the previous section. The algorithm was run using 5 instance sets of each 5 representative sizes. Two instance sets were generated using the map from figure 1.1a, two instance sets were generated based on the map from figure 1.1b and one set was generated based on figure 1.1c on page 3.

1.4.1 Evaluation of final model

As in section ?? on page ?? we choose to represent each instance using the following parameters:

- $\wp_n$ represent the amount of orders where n is the instance order size
- $\Re_v$ represent the amount of vehicles where v is the instance vehicle amount
- $\Im_l$ represent the location amount where l is the instance location amount

In addition we choose to add either E_i for europe instances, G_i for german instances and U_i for the unifor distribution instance. i is differenciating each instance.

As an example $\wp_{10}\Re_4\Im_{7E}$ would be an instance with ten orders, four vehicles and seven locations based on the Europe location instance.

For each instance we also used the Windows computer to let AMPL try for 10 000 seconds to find an optimal solution for each instance. The results from the runs are summarized in table **??** on page ??.

TODO: Finish tables here for each instance...

We also haphazardly selected some runs from different instances to analyse how the operators are performing. The results are shown in figure **??** on page ??. TODO: finish figures with operator performances

1.4.2 Observations of results the final composition

Table **??** on page ?? shows that our model is TODO: finish observations part.

However, from the previous sections result we also cannot conclude that they are not significant in combination with other heuristics or if they are leading to a worse solution than not including them, see **??**.