Stockholms universitet

Forecasting Vehicle Quality: Developing a dynamic data-driven method to identify homogeneous sub-populations of vehicles

Sotirios Loustas

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2017:2
Matematisk statistik
Mars 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2017:2**
http://www.math.su.se

Forecasting Vehicle Quality: Developing a dynamic data-driven method to identify homogeneous sub-populations of vehicles

Sotirios Loustas*

March 2017

Abstract

This thesis aims to develop a dynamic data-driven method to identify homogeneous sub-populations/clusters of vehicles to be used in Scania's Basic Warranty Forecast but also for e.g. EPC and contract calculations. The main goals are:

1. To define "good enough" homogeneous sub-populations/clusters as well as minimum population size.
2. To identify factors governing lambda and failure rate of homogeneous sub-populations/clusters of complete vehicles
3. To identify suitable method(s) to find homogenous sub-populations/clusters of complete vehicles
4. To identify homogeneous sub-populations/clusters of complete vehicles.

To achieve those targets, data from different sources are combined in a way to maximize inclusion of recent events and a variety of clustering algorithms are applied (k-means, ward's algorithm, EM algorithm). The results are then validated using cluster validation metrics and also, empirically.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: sotloustas@gmail.com. Supervisor: Taras Bodnar.

Contents

Acknowledgements

I would like to thank my external supervisor, Annette Hultåker, for her guidance and support that made everything easier for me during the 5 months I spent at Scania. I would also like to thank my supervisor at Stockholm University, Taras Bodar, for his encouragement and interest. Also, a big thank you to Scania's YQI Department - it was really great working with you. Last, but not least, I'd like to thank Zoi.

Chapter 1

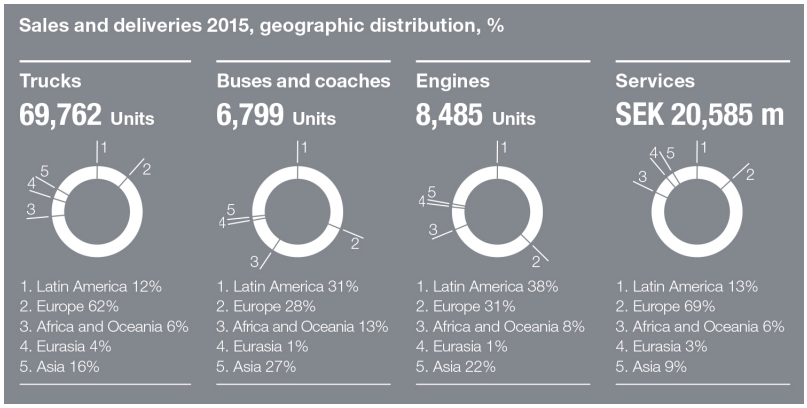
Introduction

1.1 Scania facts

Scania AB is a Swedish automotive industry manufacturer, focused in heavy trucks, busses and sole engines. It was founded in 1891 by Surahammars Bruk and Philip Wersén under the name VABIS (Vagnfabriks AktieBolaget i Södertelge). AB Scania-Vabis was established in 1911 after the merge of VABIS with Maskinfabriks-aktiebolaget Scania, a bicycle company founded in Malmö which by then had expanded and manufactured trucks and various other products. Scania merged with Saab AB in 1969 and formed Saab-Scania AB until the company split in 1995. Since then the company operates as Scania AB.

Volkswagen Group is the major shareholder since Scania's aquisition in 2007. Currently, Volkswagen AG holds 82.63 percent of the shares in Scania AB and MAN SE holds 17.37 percent of the shares in Scania AB. MAN SE is controlled by Volkswagen AG, thus Volkswage AG directly or indirectly owns 100 percent of the shares in Scania and consquently, directly or indirectly, controls all of the voting rights in Scania AB [1]. Annual revenue in 2015 was SEK 94,897 billion and net income rose to SEK 6,753 billion [2]. In 2015 Scania delivered 69,762 trucks, 6,799 busses and 8,485 engines [1].

Figure 1.1: Distribution of sales



1.2 Basic Warranty Period

Scania busses, trucks and engines are sold with a 12-month warranty. The Basic Warranty Period starts on the delivery date and ends one year later. In a rare case when a vehicle is delivered more than 12 months after it was assembled, the Basic Warranty Period expires exactly 24 months after assembly date. The Basic Warranty is solely for the benefit of the first buyer and not for the benefit of any other subsequent buyers of the product [3].

Exclusions from the warranty are costs associated with:

1. road accidents, natural causes, abnormal use, overloading, faulty servicing
2. normal wear and tear
3. additions or modifications after delivery
4. obvious failures that were not treated immediately
5. late notification of dealer and late presentation of product for repairs

1.3 Basic Warranty Forecast

Scania's YQI (Complete Vehicle Quality Information Department) is responsible for providing accurate forecasts for claims that occur during the Basic Warranty Period.

1. **Basic Warranty Forecast**: A forecast performed monthly, on various Scania products according to specifications (Haulage vehicles, Construction vehicles, City buses) or different geographical regions. The forecast is aiming to predict the number and the cost of claims for vehicles that have not yet finished their 12 month warranty period.
2. **Vehicle Off Road**: A forecast performed only for vehicles that have suffered important damage and are no longer fit for use.

In this case, the analysis is constrained in the Basic Warranty Forecast. The key components of the Basic Warranty Forecast are variables λ_t and F_l.

1. λ_t: The distribution of repairs during the warranty year.
2. F_l: Distribution of delay in claim reporting.

Parameters λ_t and F_l are used to calculate a forecast factor for each of the assembly months. The estimation of claims per vehicle (q_x) for a specific assembly month x is done from the currently known number of claims $\widetilde{n}_x$ divided by the forecast factor.

1.4 The purpose of this paper

The goal of this Thesis Project is to develop a method that will assign trucks in groups according to their specifications, examine the group claim number/rate evolution through time and finally create homogeneous clusters of vehicles. So, λ_t and F_l will be calculated for each vehicle cluster, optimizing estimations and reducing errors associated with poor choice of samples. To achieve this goal, a variety of statistical/machine learning approaches are suggested and compared.

1.5 Software used

Calculations were performed in R Studio Version 0.99.903 and the R packages used were "dplyr", "NbClust" and "clValid".

Chapter 2

Data Selection

One of the most important parts of the project is to optimize the data selection procedure. Scania YQI has different databases that need to be combined in order to produce meaningful results. In this Chapter, one can find information about the variables that are going to be used later on. Also, some empirical groupings that were created by some of Scania's departments are mentioned. Finally, two data selection approaches are suggested and compared.

2.1 Product and Claim Variables

All data used in this paper were extracted from SQL Databases located in a Scania Data Warehouse. Each database contains different variables regarding product specifications, operational data, claim registration data, etc. Out of those, three different tables of Scania's **data_mart** database were used:

1. **data_mart.product** is a table with general information about vehicles. From data_mart.product the following columns were extracted:
    - **product_id** is the key variable. Every vehicle/engine is assigned a unique product ID which is also used as a link between the different databases.
    - **segment_cd** is an index used to classify vehicles according to some common specification characteristics. It takes the values C for Construction vehicles, D for Distribution vehicles, H for Haulage vehicles and P for Public vehicles.
    - **country_current_cd** is a code giving information about the country a vehicle/engine was sold to.
    - **warranty_start_dat** is the date of delivery and also the start of the basic warranty period
2. **data_mart.claim**: a table with information about claims registered so far. From data_mart.claim the following columns were extracted:
    - **product_id** is the key variable. Every vehicle/engine is assigned a unique product ID which is also used as a link between the different databases.
    - **repair_dat** is the date when the repair of the vehicle/engine took place.
    - **month_to_repair_qty** is an integer describing the number of months between warranty start date and repair date.

- **claim_settle_dat** is the date when the claim was settled and approved by both parties.

2.2 Data selection approaches

A crucial part of this project is data selection. Since the analysis is limited on the 12 month basic warranty period, interest is focused only in claims that occurred during the first 365 days of a vehicle's use. For the sake of simplicity, only vehicles described as "trucks" are included in the analysis; "buses" and "engines" are skipped. Two approaches are suggested and compared, in order to optimize data selection procedures.

2.2.1 Claim registration delay

A major obstacle that has to be overcome is data quality. The claim registration process can be divided in 4 parts:

1. Repair Date: When the vehicle was repaired.
2. Claim Registration Date: When the claim was registered.
3. Claim Settle Date: When the claim was approved by Scania.
4. Credit Date: When the payment was completed.

In general, the Repair Date differs from the Claim Settle Data, hence there is a claim registration delay that leads to insufficient information. This issue can be solved in various ways during the data selection process.

2.2.2 First approach

The first approach is a method which is currently used by Scania's YQI. Selection of vehicles is limited on product ID's whose warranty start date and current date differ by at least 390 days. Hence, we use only vehicles that have completed their 12 month basic warranty period with an addition of 25 days. This addition is an empirical estimate that is used by Scania to balance delays between claim repair date and claim settling date (claim registration delay).

After combining information from data_mart.product (product specifications) and data_mart.claim (claim information), a pivot table is created (Table 2.1). Each product ID's claim evolution during its first 12 months in use is calculated, in addition with some product specification variables like Segment Code, Country Code and Warranty Start Date (day of delivery).

Although this method is rather simple, a major disadvantage can be observed: Vehicles younger than 390 days are not included. This information gap can lead to wrong conclusions, especially if recent developments affect vehicle quality in certain vehicle subgroups. In addition, since claim registration delay is not a constant and is also varying according to country of operation, this will lead to over/underestimation of the actual sample of vehicles that have to be included in the analysis.

Table 2.1: Segment code (SC), Country code (CC),Warranty start date and Number of claims per product ID

Product ID	Month in use												SC	CC	Warranty start
	1	2	3	4	5	6	7	8	9	10	11	12			
XXXXXX1	0	1	0	0	0	0	0	0	0	0	0	0	P	FI	2014-12-05
XXXXXX2	0	0	0	0	0	0	0	0	0	2	0	3	H	KZ	2012-02-15
XXXXXX3	0	0	0	1	0	0	0	0	0	0	0	0	C	RW	2012-04-11
XXXXXX4	0	0	0	0	0	0	0	0	0	0	1	0	D	SE	2012-04-10
XXXXXX5	0	0	1	0	1	1	0	0	0	0	0	0	P	AU	2012-02-20
XXXXXX6	0	0	2	1	1	0	0	0	0	0	2	0	H	CZ	2012-07-20

2.2.3 Second approach

To tackle the disadvantages of the first data selection approach, a new one is suggested. A new variable, named "sufficient months" (SM), is created. A month is considered sufficient if its last day's distance to the current date is larger than a delay threshold value dt(j) that is calculated using registration delay data for country j. Creating this "sufficient months" variable is crucial, since it helps filtering out periods when a significant number of claims may not have been registered due to delays in the claim registration process.

Definition 1 *Given country j and the set of claim registration delays D_j, delay threshold dt(j) is defined as the 9-th decile of D_j. Consequently, 90 $\%$ of observations in D_j are smaller or equal than dt(j).*

Scania's existing approach is to set $dt(j)$ as the average claim registration delay over all countries, which is estimated at 25 days. The new method for delay calculation has some major advantages:

1. The choice of the 90 $\%$ threshold is used to provide results that are less risky compared to using the average claim registration delay as a threshold value and was selected after discussion with Scania's YQI team.

2. The calculation is performed for each country separately, hence providing more detailed information.

3. The claim registration delays are recalculated each time.

Definition 2 *The amount of sufficient months of a vehicle with product ID i are defined as:*

$$SM(i,j) = max\Big\{n \in N | n \leq 12\frac{age(i) - dt(j)}{365}\Big\}\ where:$$

j=country, age(i)=distance between warranty start date and current date for vehicle i

After calculating the number of sufficient months for each product ID, it is possible to create subsets of vehicles based on this index. This is of particular interest since it will allow calculation of monthly claim rates, monthly claim amounts and monthly active vehicles with minimal loss of information.

Table 2.2: Sufficient Months (SM), Segment code (SC), Country code (CC),Warranty start date and Number of claims per product ID

Product ID	SM	Month of use												SC	CC	Warranty start
		1	2	3	4	5	6	7	8	9	10	11	12			
XXXXXX1	22	0	1	0	0	0	0	0	0	0	0	0	0	P	FI	2014-12-05
XXXXXX2	55	0	0	0	0	0	0	0	0	0	2	0	3	H	KZ	2012-02-15
XXXXXX3	53	0	0	0	1	0	0	0	0	0	0	0	0	C	RW	2012-04-11
XXXXXX4	54	0	0	0	0	0	0	0	0	0	0	1	0	D	SE	2012-04-10
XXXXXX5	54	0	0	1	0	1	1	0	0	0	0	0	0	P	AU	2012-02-20
XXXXXX6	50	0	0	2	1	1	0	0	0	0	0	2	0	H	CZ	2012-07-20

2.2.4 Comparison

The difference between the two approaches is significant (Table 2.3). The second approach obviously leads to the inclusion of more vehicles and claims in the analysis. As more months are added, the two approaches converge, but there is a slight difference for month 12. There, the second approach is including more vehicles and claims than the first one. This can be explained by the fact that a number of claims that have occurred during month 12 are not taken into account because they were registered during a month that was described as insufficient for those specific vehicles. The reason behind this is that the current method is using a constant (+25 days) as a claim registration delay threshold to hedge against lack of available data, which proves to be an optimistic estimator.

Apart from the fact of larger samples, another important advantage of the second approach is the inclusion of recent events in the analysis. Each month's products are different; they are sold to different countries and have different specifications. Hence, there is an increased need to incorporate claims that are associated with vehicles that have not yet completed their first 12 months in use, a requirement that is fulfilled when the second approach is used.

Consequently, the suggested approach for vehicle selection is the second. Because of this fact, all further calculations for this project will be based on data that were selected using the second approach.

Table 2.3: Number of claims per approach

	Claims, approach 1	Claims, approach 2	difference
month 1	5889	10511	4622
month 2	5150	9328	4178
month 3	5330	9444	4114
month 4	4651	7816	3165
month 5	5216	8211	2995
month 6	5270	7985	2715
month 7	6383	8759	2376
month 8	6494	8410	1916
month 9	7188	8649	1461
month 10	8018	9080	1062
month 11	8627	9254	627
month 12	14832	14590	-242
total	83048	112037	28989

Chapter 3

Data transformation

3.1 Existing vehicle groupings

Vehicle classification and categorization is a process that is being handled by various Scania departments. Each department has a unique method to categorize vehicles, according to some predetermined criteria. Also, categorization is closely correlated with the problem that has to be solved. So, any method for vehicle categorization is subjective and depends on various factors.

3.1.1 Groupings used for the Basic Warranty Forecast

YQI uses three different clusters of vehicles to forecast claims that will occur in the future. These clusters are:

1. CWP (Common Warranty Process) vehicles: Vehicles that are operating in a country which uses an online system for claim registration. Claims associated with CWP vehicles tend to have small claim registration delays.

2. non-CWP vehicles: Vehicles that are operating in a country which is not using an online system for claim registration. Claims associated with CWP vehicles tend to have large claim registration delays.

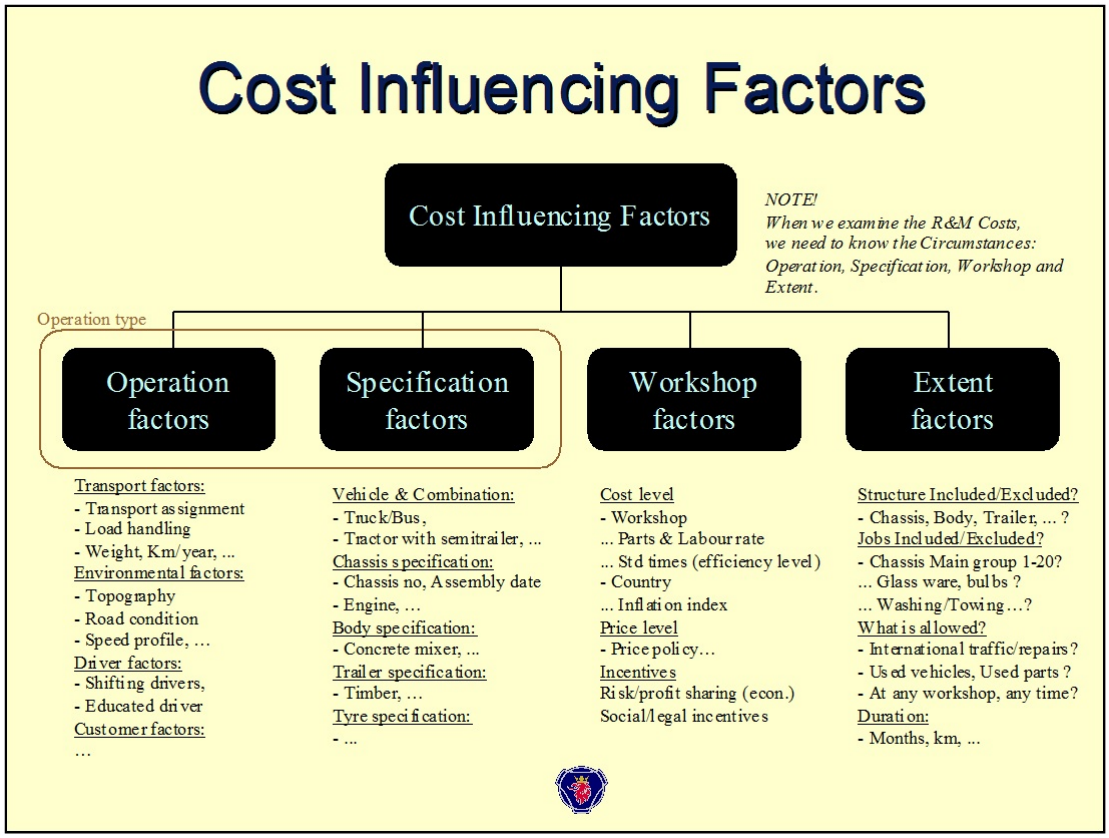
3. Brazilian vehicles: Vehicles that are operating in Brazil are placed in a separate group due to some unique behaviors observed in the past and due to the large market size.

3.1.2 Other Groupings

There are various other approaches to cluster vehicles based on some predetermined criteria. One of the most interesting is a study trying to identify the cost influencing factors for Scania's R & M Contract Specification Manual. Note that, this study is based purely on experience and not on statistical analysis of vehicle data.

The factors influencing claims are split into 4 categories: Operational Factors, Specification Factors, Workshop Factors and Extent Factors (Figure 3.1). Each category includes different sub-factors that play major or minor role. Here, only factors that are believed to be of importance are mentioned.

Figure 3.1: Cost Influencing Factors



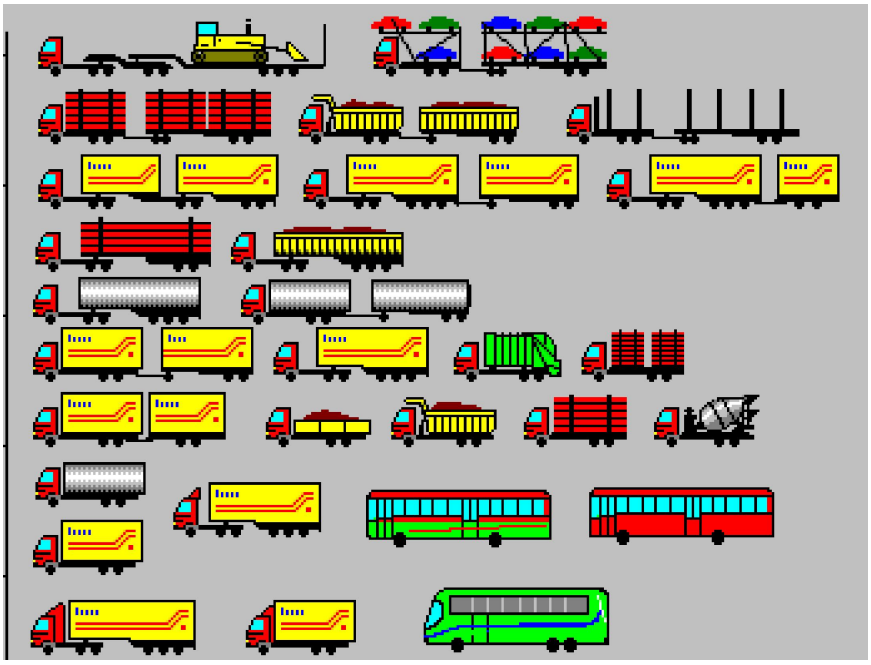
1. Operational Factors

   - Transport application: This is probably one of the most important cost influencing factors. Example of transport application: Tank, Timber, Bulk.
   - Yearly km: Estimated number of km/year. This is probably the most important Cost Influencing Factor.
   - Average used GTW (ton): Estimated average over distance of GTW.
   - Road surface: Asphalt, Asphalt/gravel, Soft gravel, Off Road
   - Speed Profile: Highway, Secondary roads, Urban Areas, City traffic, City centre
   - Topography: Flat, hilly, very hilly
   - Climate zone variation: Normal, Desert, Tropical, Arctic
   - Driver Factors: Shifting drivers, Scania Trained driver

2. Specification Factors

   - Vehicle combination type (Figure 3.2): Rigid single truck, Truck with full trailer, Truck with centre axle/bogie trailer, Tractor with semitrailer, Truck with multiple trailers, Tractor with multiple trailers

Figure 3.2: Vehicle combination types



3.2 Determine important variables

Selection of variables that are considered important was made according to the RM Contract Specification Manual and data availability. It is highly suggested to pursue a more detailed analysis, to combine data and knowledge from various Scania departments in order to detect variables that have an important effect on claim rates. Here, the vehicle subgroups were created using the variables **segment_cd** and **country_current_cd**.

Other possible variables that were available but not used are **pru** (Place of product assembly) and **wheelconfig** (Wheel configuration). The variable **pru** was not used because it is highly correlated with **country_current_cd**, since vehicles sold in certain country groups are assembled in the same place. **Wheelconfig** was not used because the **segment_cd** index is incorporating it.

3.3 Partitioning

Partitioning the data into small subgroups is an important part of the project. The methodology applied here is aiming at creating subgroups with an adequate amount of vehicles/claims. In order to properly balance between number of subgroups and subgroup size, smaller subgroups are merged.

3.3.1 Merging segment_cd and country_current_cd

The first step in partitioning the data is to create a new index by merging **segment_cd** and **country_current_cd**. The merged specification index is named **group_ID**. Consequently, each vehicle is assigned a **group_ID** index (Table 3.1). In total there are 310 different combinations of **segment_cd** and **country_current_cd**, hence there are 310 **group_ID** elements (Figure 3.3). After creating the **group_ID** index, vehicles are grouped according to it and aggregations are performed (Table 3.2).

Figure 3.3: Group ID's - Before partitioning

FI-P	CZ-C	JP-H	TN-D	SE-P	ZZ-D	BR-C	LB-H	CH-D
KZ-H	HU-H	DZ-D	PT-C	TR-C	BO-H	NZ-H	AZ-C	IL-P
RW-C	MK-H	SK-H	NO-NA	FI-D	UY-H	ES-P	QA-C	TZ-C
SE-D	BA-H	RO-H	ZZ-NA	NZ-P	UY-D	MY-D	SD-C	AB-P
AU-P	IE-C	BY-H	KE-P	NO-P	CL-D	TW-D	UG-C	LT-D
CZ-H	AB-C	ES-H	CH-NA	PL-D	AR-C	NZ-C	DU-H	LB-C
CZ-D	GR-H	BG-H	ZZ-P	NL-H	DO-D	KZ-P	SA-H	FR-P
HK-C	AB-H	TZ-H	IQ-D	ZA-C	KE-D	BE-C	RW-H	GR-P
TH-C	DU-P	PT-H	ID-D	LT-H	VE-C	MN-C	SG-D	GR-C
TH-H	AE-P	SI-H	TZ-D	PL-H	AO-C	FR-C	ET-C	DZ-C
ZZ-H	AM-H	IS-C	IQ-P	DK-H	EC-H	ZA-H	RO-D	ES-C
RU-H	IS-P	BY-D	CY-P	IR-H	EC-C	BE-D	HU-D	NZ-D
CN-H	AM-D	IS-H	SI-P	ZA-D	UY-P	ZZ-C	HU-C	DZ-H
TH-D	DU-C	LV-D	HU-P	OM-C	VE-D	MA-H	EG-C	PT-D
MY-H	RO-C	AU-C	PY-C	PL-C	VE-H	QA-P	BG-C	
HK-P	BG-D	RO-P	BO-C	DE-C	AR-P	LV-C	TN-C	
GB-C	TR-H	RU-P	HN-D	AT-H	CL-P	LV-H	TN-H	
KR-C	CN-D	PT-P	CR-C	DE-H	CO-C	SG-C	ID-H	
GB-H	IT-D	IS-D	CR-H	CH-P	UY-C	UA-H	RS-C	
CN-P	SA-C	BE-H	QA-D	NO-C	MZ-H	MY-P	HR-D	
KR-H	AE-C	AT-D	VN-D	ZA-P	PA-H	FR-D	OM-H	
GB-P	IE-D	SK-C	GH-H	CN-C	PY-H	NO-D	BY-C	
RU-C	IL-H	UG-P	CY-D	IE-P	BO-D	DE-D	HR-C	
FI-C	LT-P	UA-C	RS-D	ID-C	ZW-H	EE-P	SK-D	
RU-D	PE-C	IQ-H	ZW-D	IT-H	PE-D	NL-C	QA-H	
AU-H	DO-C	CZ-P	VG-C	PL-P	IR-D	FR-H	IN-H	
NL-D	DO-H	MA-C	HN-C	TW-H	IR-C	IL-C	MA-D	
DE-P	AR-H	IT-P	MM-H	ID-P	SI-C	EE-D	AB-D	
SE-H	CL-C	TH-P	RW-P	TW-C	SY-H	IT-C	ES-D	
NO-H	PE-H	RS-H	TW-P	IE-H	AU-D	JP-C	CY-H	
HK-H	BR-D	IQ-C	VG-H	DK-P	IL-D	DK-D	MK-C	
FI-H	KE-C	BR-P	HN-H	NL-P	SY-C	EE-C	ZW-C	
BE-P	MY-C	BR-H	MM-D	EE-H	LB-D	CH-H	MM-C	
HK-D	AR-D	UA-D	GH-C	CH-C	EG-H	AT-P	SA-P	
SE-C	KE-H	HR-H	GR-D	AT-C	SG-H	LT-C	BA-D	
SG-P	VN-C	KR-P	TR-D	KZ-C	UG-H	IN-C	AE-H	
DK-C	CL-H	SI-D	MK-D	GB-D	VN-H	AM-C	BA-C	

Table 3.1: Sufficient Months (SM), Group ID (GID),Warranty start date and Number of claims per product ID

Product ID	SM	Month of use												GID	Warranty start
		1	2	3	4	5	6	7	8	9	10	11	12		
XXXXXX1	22	0	1	0	0	0	0	0	0	0	0	0	0	P-FI	2014-12-05
XXXXXX2	55	0	0	0	0	0	0	0	0	0	2	0	3	H-KZ	2012-02-15
XXXXXX3	53	0	0	0	1	0	0	0	0	0	0	0	0	C-RW	2012-04-11
XXXXXX4	54	0	0	0	0	0	0	0	0	0	0	1	0	D-SE	2012-04-10
XXXXXX5	54	0	0	1	0	1	1	0	0	0	0	0	0	P-AU	2012-02-20
XXXXXX6	50	0	0	2	1	1	0	0	0	0	0	2	0	H-CZ	2012-07-20

Table 3.2: Number of claims per Group ID (GID), per month in use

GID	Month of use											
	1	2	3	4	5	6	7	8	9	10	11	12
C1-C	6	5	14	10	13	7	10	13	13	25	16	26
C2-C	105	112	99	69	88	75	89	74	81	79	91	127
C3-C	13	7	13	9	22	15	7	4	8	12	16	15
C3-D	57	31	36	37	38	35	35	27	31	35	29	52
C3-H	234	194	198	186	199	199	231	180	206	219	192	238
C4-C	141	139	132	108	103	113	119	103	112	129	136	252

3.3.2 Merging small sized group_ID's

In order to reduce implications that may occur due to small sample size of **group_ID**'s, every **group_ID**, containing less than t claims (threshold) in at least one month, will be merged into a new **group_ID** named **SMALL_SAMPLE**. The choice of t is based on the size of the **SMALL_SAMPLE** subgroup. The goal is to select t in a way such that not more than 5% of vehicles will be treated as outliers. So, from the 310 initial subgroups, some will not be affected and others will be merged into the **SMALL_SAMPLE** subgroup.

Table 3.3: Number of vehicles assigned to SMALL_SAMPLE subgroup, per threshold value

t (threshold)	Subgroups	Outlier group size	Percent of total vehicles
1	168	4594	1.4%
2	155	6297	1.9%
3	143	7796	2.4%
4	135	8439	2.6%
5	128	10078	3.1%
6	121	11738	3.6%
7	116	12900	3.9%
8	114	13320	4.1%
9	111	14247	4.4%
10	110	14388	4.4%
11	108	14836	4.5%
12	107	15328	4.7%
13	106	15822	4.8%
14	104	17867	5.5%

Consequently, according to Table 3.3 a threshold value of t=13 would reduce the subgroups to 106 and keep the SMALL_SAMPLE subgroup's size below the 5% limit. Out of 310 existing subgroups 205 are merged into the **SMALL_SAMPLE** subgroup which corresponds to 4.8% of vehicles.

3.4 Claim rate and claim number tables

The claim rate and claim number tables are used as inputs in the various clustering algorithms and also for validating.

Definition 3 *The claim rate for subgroup i and month t is defined as:*

$$CR(i,t) = \frac{N_c(i,t)}{N_p(i,t)}$$

$N_p(i,t)$ is the number of active vehicles of subgroup i in month t and $N_c(i,t)$ is the number of claims for active vehicles of subgroup i in month t. Note that $N_p(i,t)$ is not the same for every t. $N_p(i,t)$ is declining when t increases.

Next step is to group the vehicles according to their **Group_ID** and calculate:

1. the average number of claims for each month (Claim Rate - Appendix A.1)
2. the sum of claims for each month (Claim number - Appendix A.2)

Figure 3.4: Claim rate for C9-C vehicles

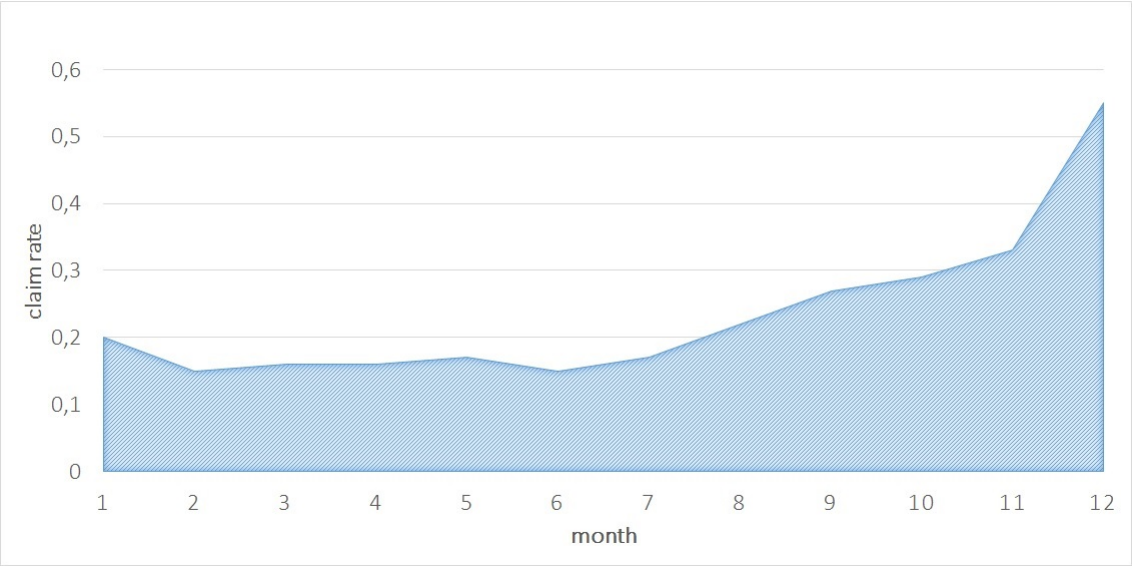


Figure 3.4 is very common claim rate evolution. In general, there are more claims during the 12th month in use because customers usually visit the workshop for one last time before the Basic Warranty expires.

Chapter 4

Clustering

Cluster analysis is aiming to divide data into clusters that are meaningful, useful, or both. The elements of each group should have either similar characteristics or should be differing from elements that belong in other groups. The balance between intra-cluster similarities and cluster dissimilarities is what provides optimal results [10].

A common approach to create homogeneous groups is to cluster elements according to some predetermined criteria. In this case, the goal is to cluster based on similar claim rate evolution during the basic warranty period, among the different partitions. Consequently, the clustering algorithms that are going to be used will be applied on the claim rate table calculated in chapter 3 (Appendix, Table A.1). Since the aim is to produce homogeneous clusters, intra-cluster similarities are considered more important than dissimilarities among the clusters and the tools that are used were selected according to this fact.

4.1 Determining the number of clusters

One of the important problems that have to be addressed is the number of clusters that will have to be created. A common approach is to cluster for a variety of cluster numbers and apply some indices on each result. Then, the selection of the optimal number of clusters (k) is relying on the criteria of each index. A useful summary of 30 indices is provided by NbClust, an R Package for determining the optimal number of clusters [11]. According to literature [12], the best performing indices for a similar dataset were CH and Duda. In this study, a newer index (D-index) is also applied, mainly because it is focused in measuring intra-cluster homogeneity.

4.1.1 Notations

- X = n×p data matrix of p variables measured on n independent observations
- x_i = vector of observations of the i-th object in cluster C_j
- $\bar{x}$ = centroid of matrix X
- c_j = centroid of cluster j
- k = number of clusters

- n = number of observations
- n_j = number of observations in cluster C_j
- p = number of variables
- $B_k = \sum_{j=1}^{k} n_j(c_j - \bar{x})(c_j - \bar{x})^T$ is the between-group dispersion matrix for data clustered into k clusters
- $W_k = \sum_{j=1}^{k} \sum_{i \in C_j}(c_j - x_i)(c_j - x_i)^T$ is the within-group dispersion matrix for data clustered into k clusters

4.1.2 CH index

The Calinski and Harabasz (CH) index [4] is defined as:

$$CH(k) = \frac{trace(B_k)/(k-1)}{trace(W_k)/(n-k)}$$

note that $trace(X) = \sum_{i=1}^{n} x_{i,i}$, where X is a n×n matrix. The value k, maximizing CH(k) is the suggested number of clusters.

4.1.3 Duda index

Duda and Hart [5] suggested a ratio criterion Je(2)/Je(1) in order to decide if a cluster should be partitioned further, hence creating more clusters.

- Je(2) is the sum of squared errors within clusters when the data are partitioned into two clusters (C_2 and C_3)
- Je(1) gives the squared errors when only one cluster is present (C_1)

Here, C_1=$C_2 \cup C_3$. Consequently,

$$Duda = \frac{Je(2)}{Je(1)}$$

The optimal number of clusters k is the smallest number that satisfies [13]:

$$CriticalDuda \leq Duda$$

where,

$$CriticalDuda = 1 - \frac{2}{p\pi} - z\sqrt{\frac{2(1 - \frac{8}{\pi^2 p})}{n_1 p}}$$

Here, z is a standard normal score and n_1 is the number of observations in cluster C_1. After several tests for different score values, z=3.2 is considered to be the optimal value. [12]

4.1.4 D-index

D-index is a graphical method, based on clustering gain on intra-cluster inertia. Intra-cluster inertia measures the degree of homogeneity between the data associated with a cluster. It calculates all distances compared to a reference point representing the profile of the cluster. Here, the reference point is the cluster centroid c_j and the distance $d(x_i, c_j)$ is the euclidean distance. Intra-cluster inertia ($w(P^k)$) is defined as:

$$w(P^k) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x_i \in C_j} d(x_i, c_j)$$

Given two partitions, P^{j-1} composed of j-1 clusters and P^j composed of j clusters, the clustering gain on intra-cluster inertia is:

$$gain = w(P^{j-1}) - w(P^j)$$

This clustering gain should be minimized. The optimal number of clusters can be observed by a sharp knee that corresponds to a signicant decrease of the first dierences of clustering gain versus the number of clusters. This can be observed better in the second differences of clustering gain versus the number of clusters [6].

4.2 Clustering Algorithms

Clustering is a fundamental data mining task, used mostly for unsupervised learning. Its goal is to group similar elements in same clusters and different elements in different clusters. Fraley and Raftery [14] divide the clustering methods into two distinct groups: hierarchical and partitioning methods. In addition, Han and Kamber [15] suggest three extra categories: density-based methods, model-based clustering and grid-based methods.

In this paper the clustering algorithms that are used are:

- k-means (partitioning)
- Agglomerative hierarchical Clustering: Ward's Method (hierarchical)
- EM-algorithm (model based)

The choice of clustering algorithms was based on the general aim of the thesis project - minimizing cluster variance and thus creating homogeneous clusters, but also using algorithms that cluster observations using different approaches.

4.2.1 k-means

The most common and simple clustering algorithm is k-means; a partitioning algorithm. The core idea is to create a clustering structure that minimizes SSE (Sum of Squared Error).

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} (c_i - x)^2$$

The value of c_j that minimizes SSE is the cluster mean $\bar{c_j}$ since:

$$\frac{\partial}{\partial c_j} SSE = \frac{\partial}{\partial c_j} \sum_{i=1}^{k} \sum_{x \in C_i} (c_i - x)^2$$

$$= \sum_{i=1}^{k} \sum_{x \in C_i} \frac{\partial}{\partial c_j} (c_i - x)^2 = \sum_{x \in C_j} 2(c_i - x) = 0$$

$$\Rightarrow n_j c_j = \sum_{x \in C_j} x_j \Rightarrow c_j = \frac{1}{n_k} \sum_{x \in C_j} x_j = \bar{c_j}$$

K-means starts with an initial random set of k cluster centroids c_1,...,c_k. In each loop, each data point is assigned to its nearest cluster centroid, thus forming k clusters. Then, the cluster centroids are recalculated and the loop starts again. The algorithm stops when the recalculated cluster centroids do not change, which means that none of the cluster assignments change in future iterations. [7]

In this paper, Euclidean distance is used in order to calculate distances between points and cluster centroids, although there are numerous distance metrics that can be applied (maximum distance, manhattan distance, canberra distance, binary distance, minkowski distance, etc.). [11]

4.2.2 Agglomerative hierarchical Clustering: Ward's Method

Hierarchical clustering is a method that is forming clusters with 2 different approaches: top-down or bottom-up. [7]

1. Agglomerative hierarchical clustering (bottom-up): Each element initially forms a cluster of its own. Then, these clusters are successively merged and a cluster structure is obtained.

2. Divisive hierarchical clustering (top-down): All elements initially form a single cluster. This cluster is successively partitioned into smaller subclusters and a cluster structure is obtained.

Partitioning or merging is performed according to a similarity criterion. In this paper, agglomerative hierarchical clustering is combined with ward's minimum variance criterion. This leads to the formation of clusters with small intra-cluster variance.

Ward's method defines the distance between two clusters C_A and C_B as $\Delta(C_A,C_B)$, where $\Delta(C_A,C_B)$ denotes the sum of square increase if C_A and C_B are merged. Hence,

$$\Delta(C_A, C_B) = \sum_{i \in A \cup B} \|\vec{x_i} - \vec{c_{A \cup B}}\|^2 - \sum_{i \in A} \|\vec{x_i} - \vec{c_A}\|^2 - \sum_{i \in B} \|\vec{x_i} - \vec{c_B}\|^2 \Leftrightarrow$$

$$\Delta(C_A, C_B) = \frac{n_A n_B}{n_A + n_B} \|\vec{c_A} - \vec{c_B}\|^2$$

where c_j is the centroid of cluster j, and n_j is the number of points in it. Δ is also called "merging cost" of combining clusters A and B. Consequently, Ward's method creates clusters with minimal intra-cluster variance. [8]

4.2.3 EM-algorithm

In model-based clustering, data are supposed to be generated from a mixture density $f(x) = \sum_{j=1}^{k} P(C_j)f_j(x)$, where f_j is the PDF of the distribution that generated observations contained in cluster j and $P(C_j)$ is the probability that an observation comes from cluster C_j. In this paper each component is modelled by the Normal Distribution with mean μ_j, covariance matrix Σ_j and hence, it has a probability density function f_j that corresponds to:

$$P(x_i|C_j) = \frac{1}{\sqrt{det(2\pi\Sigma_j)}} exp(-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)) \tag{4.1}$$

The likelihood for n observations spread into k clusters is hence:

$$L = \prod_{i=1}^{n} \sum_{j=1}^{k} P(C_j)P(x_i|C_j)$$

Here $P(C_j)$ is an empirical estimate of the "density" of the cluster (prior probability), n is the number of observations, k is the number of clusters and let d denote the dimension of x_i. [9] In the clustering framework, EM clustering algorithm is an iterative method to calculate parameters μ_k and Σ_k for each cluster k. It can be described by the scheme below:

1. Initialize cluster assignments by selecting random μ_j and Σ_j and also set initial prior probabilities $P(C_j)$ for clusters j=1,...,k.

2. For every observation i=1,...,n and cluster j=1,...,k calculate $P(x_i|C_j)$ (equation 4.1) and the posterior probability:
$$P(C_j|x_i) = \frac{P(x_i|C_j)P(c_j)}{\sum_{l=1}^{k} P(x_i|C_l)P(C_l)}$$

3. Re-calculate empirical cluster densities (prior probabilities) and model parameters:

$$P(C_j) = \frac{1}{n} \sum_{i=1}^{n} P(C_j|x_i)$$

Expected value $\mu_{C_j,m}$ of attribute m of Gaussian (cluster) C_j:

$$\mu_{C_j,m} = \sum_{i=1}^{n} \frac{P(C_j|x_i)}{nP(C_j)} x_{i,m}$$

Covariance between p,q attributes in Gaussian (cluster) C_j:

$$(\Sigma_{C_j})_{p,q} = \sum_{i=1}^{n} \frac{(P(C_j|x_i)}{n(P(C_j)} (x_{i,p} - \mu_{C_j,p})(x_{i,q} - \mu_{C_j,q})$$

4. If convergence criterion is not fulfilled, go to step 2.

After convergence is achieved, an observation x_i is assigned to the cluster that satisfies:

$$\arg\max_l P(C_l|x_i)$$

[9]

4.3 Cluster validation

Cluster Validation is performed in order to examine which of the clustering algorithms is performing better. Here, 5 different validation approaches are tested: Cohesion, Connectivity, Dunn Index, Silhouette and Chi-square test for homogeneity.

4.3.1 Cohesion

Cohesion is defined as the sum of intra-cluster variances or within cluster sum of squares (WSS) [10]:

$$Cohesion = \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - c_j)^2$$

In this paper, since the observations are 12-dimensional, the sum of their attributes is used as a cohesion index for simplifying comparisons.

4.3.2 Connectivity

Let $nn_{i(j)}$ be the j-th **nearest neighbor** of observation i. Then, $x_{i,nn_{i(j)}}$=0 if i and $nn_{i(j)}$ are in the same cluster or 1/j otherwise. Then, for a clustering partition C = $C_1, \ldots, C_K$ of the N observations into K clusters:

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

Here L is a parameter giving the number of nearest neighbors to use. Connectivity has a value in [0,∞) and should be minimized to achieve optimal clustering. [16]

4.3.3 Dunn Index

Given a clustering partition C = $C_1, \ldots, C_k$, the Dunn index is the ratio of the smallest distance of observations that aren't in the same cluster to the largest distance between observations that are in the same cluster.

$$Dunn(C) = \frac{\min_{C_p, C_q \in C, C_p \neq C_q} (\min_{x \in C_p, y \in C_q} dist(x, y))}{\max_{C_m \in C} Diam(C_m)}$$

Here Diam(C_m) is the maximum distance between all observations that are in cluster C_m. The Dunn index has a value in [0,∞) and should be maximized. [16]

4.3.4 Silhouette Width

For an observation i, the silhouette is defined as:

$$S(i) = \frac{b_i - a_i}{max(a_i, b_i)}$$

where, a_i is the average distance between i and all other observations contained in the same cluster, and b_i is the average distance between i and the observations contained in the nearest neighboring cluster C_j.

The Silhouette Width is the average Silhouette value of all N observations:

$$S.W. = \frac{1}{N}\sum_{i=1}^{N} S(i)$$

The Silhouette Width has a value in [-1,1] and should be maximized. [16]

4.3.5 Chi-square test for homogeneity

The chi square test for homogeneity is applied on contingency tables and is a hypothesis test where the test statistic is χ^2 distributed under the null hypothesis. Given a contingency table T (Table 4.1), a null hypothesis H_0, that the proportion of C_1 under Variable 1 is identical to the proportion of C_2 under Variable 1, etc., the null hypothesis is rejected if P($\chi^2(df)$ > x) < α where:

- $df = (N-1)(M-1)$ (degrees of freedom)
- $E_{r,c} = \frac{\sum_{i=1}^{N} t_{i,c} \sum_{j=1}^{M} t_{r,j}}{n}$ (expected frequency)
- $x = \sum(t_{r,c} - E_{r,c})^2 / E_{r,c}$ (test statistic)

Table 4.1: contingency table T

Category	Variable 1	...	Variable N	Row Total
C_1	$t_{1,1}$	...	$t_{1,N}$	$\sum_{j=1}^{N} t_{1,j}$
...	...	...	...	...
C_M	$t_{M,1}$	...	$t_{M,N}$	$\sum_{j=1}^{N} t_{M,j}$
Column Total	$\sum_{i=1}^{M} t_{i,1}$	...	$\sum_{i=1}^{M} t_{i,N}$	$\sum_{j=1}^{N}\sum_{j=1}^{M} t_{i,j}$

In order to validate the test, it is necessary that at least 80% of $E_{r,c}$ must be greater than 5. [17]

Chapter 5

Results

The application of the various methods is performed on 6 different datasets:

- Set 1: Vehicles delivered between 2014-12-01 and 2016-12-01
- Set 2: Vehicles delivered between 2014-06-01 and 2016-06-01
- Set 3: Vehicles delivered between 2013-12-01 and 2015-12-01
- Set 4: Vehicles delivered between 2013-06-01 and 2015-06-01
- Set 5: Vehicles delivered between 2012-12-01 and 2014-12-01
- Set 6: Vehicles delivered between 2012-06-01 and 2014-06-01

This will provide different results, hence a way to examine the stability of the algorithms when samples are changing.

5.1 Number of clusters

The number of clusters that should be created will be determined by 3 indices: CH, duda and D-index. Those indices will be calculated for each of the 6 vehicle sets. This is done in order to examine how each index is changing over time. All indices will be calculated on clusters formed using k-means and ward's algorithm because EM algorithm is not supported in the R Package NbClust.

5.1.1 CH and duda results

Table 5.1: Suggested number of clusters - CH and duda

Vehicle set	CH-kmeans	CH-ward	duda-kmeans	duda-ward
Set 1	2	2	2	4
Set 2	2	2	2	5
Set 3	2	2	3	5
Set 4	2	4	2	4
Set 5	2	2	2	4
Set 6	2	2	2	5

It is easy to notice that all sets of vehicles have similar results. CH index using k-kmeans is suggesting 2 clusters for all vehicle sets and CH index using ward's algorithm provides the same suggestions, except for Set 4, where the suggested number of clusters that need to be formed is 4. The suggested number of clusters based on the duda index using k-means is also similar to what CH index suggested, with the exception of Set 3, where the number of numbers should be 3. Duda index applied on clusters formed with ward's algorithm is providing different results. Here, the suggested number of clusters is evenly split between 4 and 5.

5.1.2 D-index results

D-index is a graphical method, hence we need to estimate the number of clusters empirically, by finding the most "extreme" peak in the second differences graph.

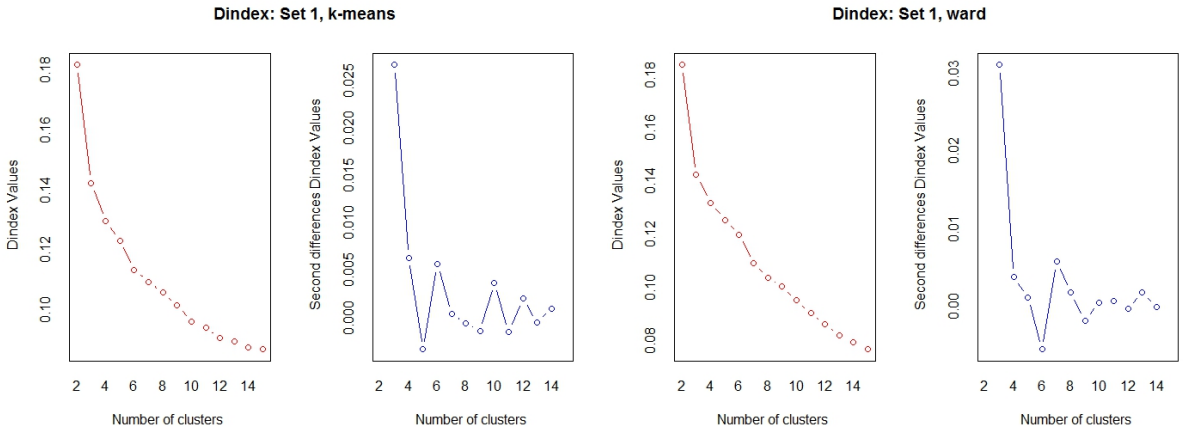


Figure 5.1: D-index: Vehicle set 1

For vehicle set 1, the optimal number of clusters (k) according to D-index (Figure 5.1) is:

- k=6 if k-means is used
- k=7 if ward's algorithm is used

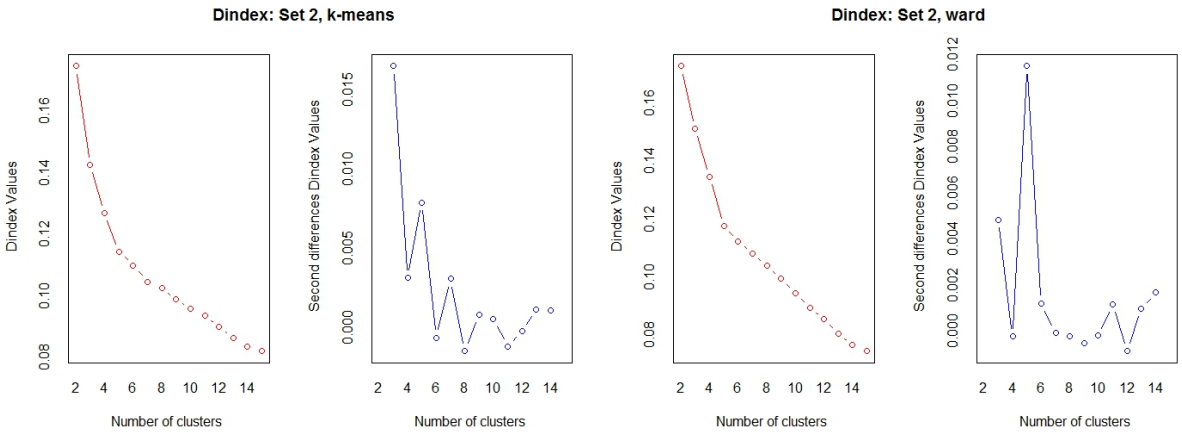


Figure 5.2: D-index: Vehicle set 2

For vehicle set 2, the optimal number of clusters (k) according to D-index (Figure 5.2) is:

- k=5 if k-means is used
- k=5 if ward's algorithm is used

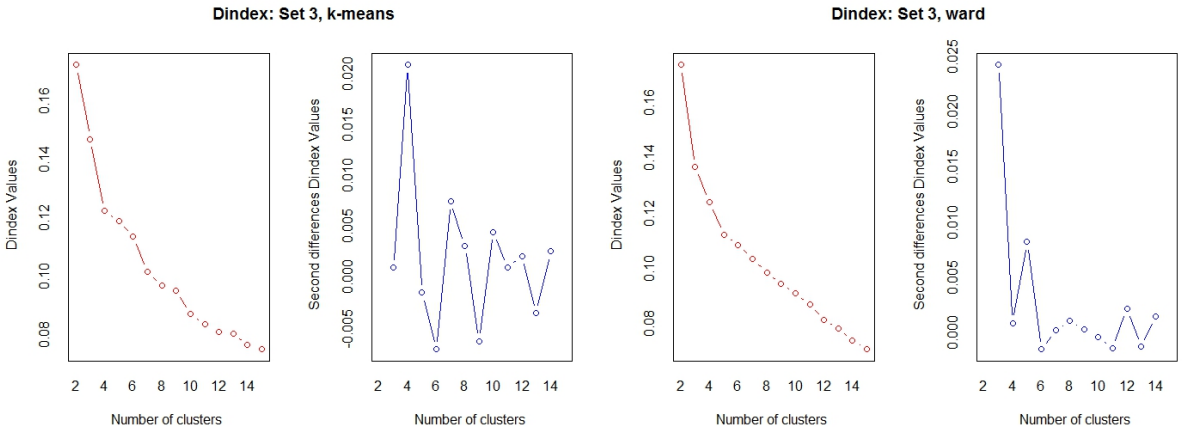


Figure 5.3: D-index: Vehicle set 3

For vehicle set 3, the optimal number of clusters (k) according to D-index (Figure 5.3) is:

- k=4 if k-means is used
- k=5 if ward's algorithm is used

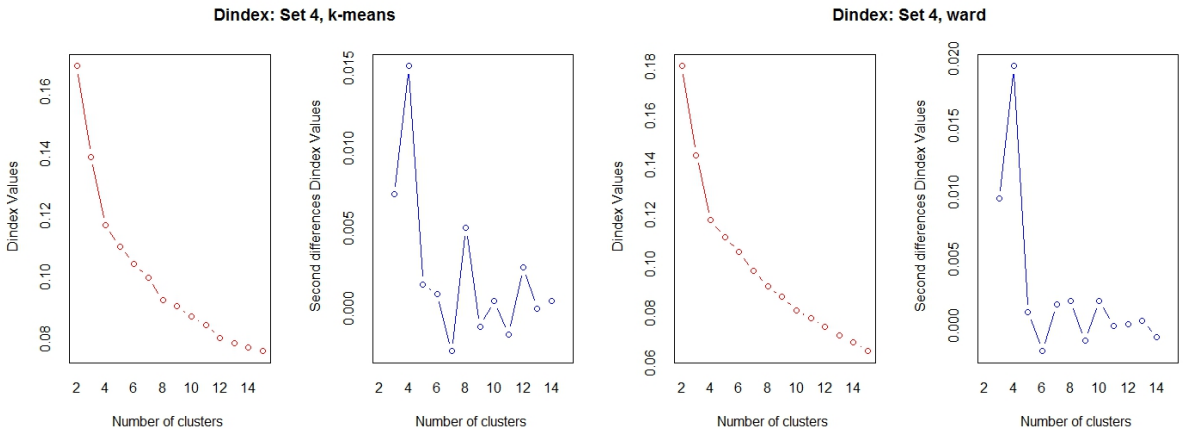


Figure 5.4: D-index: Vehicle set 4

For vehicle set 4, the optimal number of clusters (k) according to D-index (Figure 5.4) is:

- k=4 if k-means is used
- k=4 if ward's algorithm is used

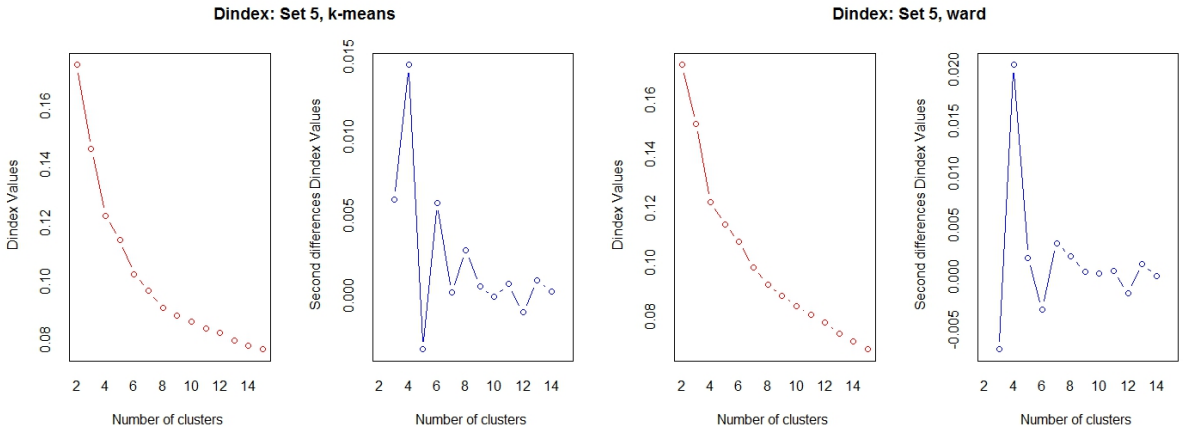


Figure 5.5: D-index: Vehicle set 5

For vehicle set 5, the optimal number of clusters (k) according to D-index (Figure 5.5) is:

- k=4 if k-means is used
- k=4 if ward's algorithm is used

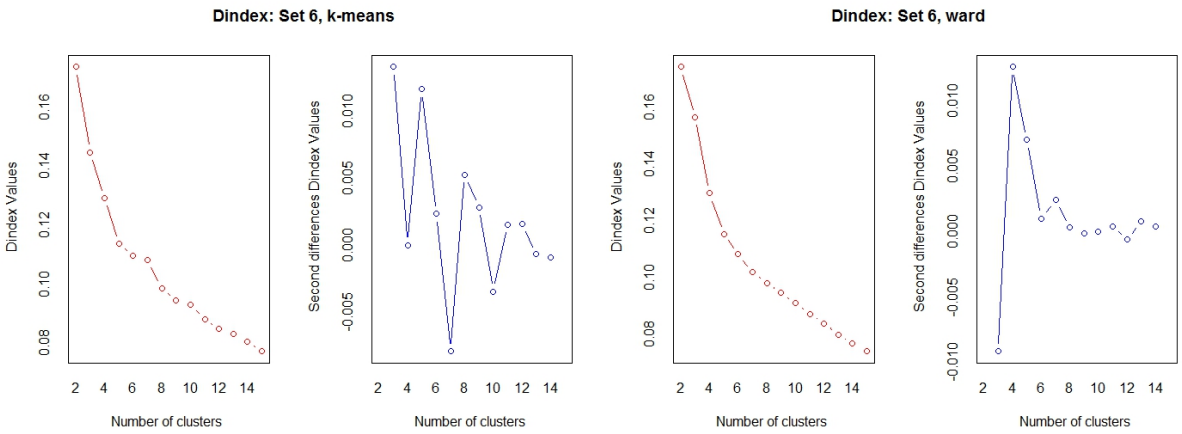


Figure 5.6: D-index: Vehicle set 6

For vehicle set 6, the optimal number of clusters (k) according to D-index (Figure 5.6) is:

- k=4 if k-means is used
- k=5 if ward's algorithm is used

After examining the second differences in the D-index graphs, the results are completely different compared to CH and duda. The number of clusters suggested for each set are concentrated in Table 5.2.

Table 5.2: Suggested number of clusters - Dindex

Vehicle set	Dindex-kmeans	Dindex-ward
Set 1	6	7
Set 2	5	5
Set 3	4	5
Set 4	4	4
Set 5	4	4
Set 6	5	4

Since the analysis will be constrained on Set 1, the possible number of clusters is 2(suggested by CH/k-means,CH/ward,Duda/k-means), 4(suggested by Duda/ward), 6(suggested by D-index/k-means) or 7(suggested by D-index/ward), with k=2 being the most frequent suggestion. By combining results for Set 1 and results based on the other sets (Table 5.3), k=4 seems a probable solution as well, whereas k=6 and k=7 appear only once. Consequently, clustering is going to be performed with k=2 and k=4.

Table 5.3: Suggested number of clusters - Comparison of results

k	count
2	16
3	0
4	10
5	8
6	1
7	1

5.2 Clustering results and validation

Clustering with k-means, ward's algorithm and EM algorithm where k=2 and k=4 will provide 6 different ways to partition data into the desired number of clusters. Each clustering result will be validated using several indices.

5.2.1 Cohesion

Table 5.4: Cluster Cohesion, k=2

Cluster	Cluster SSE, k-means	Vehicles	Cluster SSE, ward	Vehicles	Cluster SSE, EM	Vehicles
1	167246.44	112915	167937.61	113108	115385.14	29516
2	69067.67	13612	68464.78	13419	118646.99	97011
Cohesion	236314.11	-	236402.39	-	234032.13	-

For k=2, k-means and ward's algorithm provide similar results. The 2 clusters are almost identical with approximately 10% of vehicles assigned into a small cluster and the rest is assign into a larger one. There are also similar cluster SSE values. To the contrary, EM algorithm is suggesting that approximately

20% of vehicles are assigned into the small cluster and the cluster SSE is almost identical in both clusters, despite their size difference. Cohesion is minimized when EM algorithm is applied.

Table 5.5: Cluster Cohesion, k=4

Cluster	Cluster SSE, k-means	Vehicles	Cluster SSE, ward	Vehicles	Cluster SSE, EM	Vehicles
1	48958.28	10876	69830.87	27674	59957.44	23497
2	69830.87	27674	94878.85	85434	42071.17	49238
3	94878.85	85434	53394.93	11583	57789.44	39160
4	18710.64	2543	14113.52	1836	73269.48	14632
cohesion	232378.64	-	232218.17	-	233087.53	-

For k=4, k-means and ward's algorithm again provide similar results. In fact 2 out of 4 clusters are identical (k-means cluster 2 = ward cluster 1, k-means cluster 3 = ward cluster 2). EM is again splitting vehicles into completely different clusters compared to k-means and Ward's algorithm. Cohesion is minimized when Ward's algorithm is applied.

5.2.2 Connectivity , Dunn Index Silhouette Width

Connectivity, Dunn Index and Silhouette Width provide similar results for k-means and Ward, whereas EM is the worst performing method.

Table 5.6: Connectivity , Dunn Index Silhouette Width

Method	Validation Index	k=2	k=4
ward	Connectivity	6.1647	21.9651
	Dunn	0.2014	0.1053
	Silhouette	0.5886	0.4068
k-means	Connectivity	8.1425	22.1266
	Dunn	0.2014	0.1144
	Silhouette	0.5861	0.4098
EM	Connectivity	43.4794	33.6671
	Dunn	0.0785	0.0512
	Silhouette	0.3732	0.2312

According to Table 5.6, Connectivity is **minimized** for k=2 using Ward's algorithm, Dunn index is **maximized** for k=2 using both k-means and Ward's algorithm and Silhouette width is **maximized** when Ward's algorithm is applied with k=2. Consequently, EM algorithm is the method that produces worse clustering results according to all three indices. This could be different if normal distributions in the mixture would be replaced. Another explanation may be that EM algorithm is in fact unsuitable for this type of clustering problems.

5.2.3 Chi Square Test for homogeneity

Table 5.7: Chi Square Test for homogeneity, k=2

	k-means		Ward		EM	
cluster	p-value	X^2	p-value	X^2	p-value	X^2
1	0	3700.02	0	3722.51	0	4061.01
2	0	2259.76	0	2236.76	0	1494.35

Table 5.8: Chi Square Test for homogeneity, k=4

	k-means		Ward		EM	
cluster	p-value	X^2	p-value	X^2	p-value	X^2
1	0	1412.42	0	2075.60	0	1771.69
2	0	2075.60	0	1211.06	0	681.97
3	0	1211.06	0	1493.43	0	595.92
4	0	635.86	0	352.09	0	2515.73

The application of Chi Square Test for homogeneity on the claim number table of each cluster has provided results that lead to the conclusion that the frequency counts are not distributed identically across different group ID's in each cluster. In addition, since p-values are always zero, the Chi-Square Test results do not provide us with information useful for selecting the best clustering method. This could be caused by the large dimensions of the contingency tables and is an interesting result that could be studied further.

Chapter 6

Conclusions

It is clear that among the various methods tested in this paper, clustering according to k-means or ward's algorithm is providing the best results. This makes sense, since both algorithms are focused in minimizing errors associated with cluster variance. All validation metrics confirm that k-means and ward's algorithm perform better, with the exception of Chi Square Test for homogeneity, which is in general failing to provide meaningful results. Lastly, clustering results are confirming claim behaviors that have been observed empirically by Scania. For example, Indian and Russian Construction Trucks are known to produce more claims. As one can observe in the clustering results (Appendix B, Internal Version) these vehicles are the ones forming a small cluster when k=4.

The choice of k on the other hand isn't that straightforward. Although k=2 seems to be the optimal solution, it may be better to use k=4 since it will be preferable to deal with more clusters of vehicles in the forecast. In addition, since interest is constrained in cluster variance, forming more clusters may be a better choice because variances are correlated with cluster sizes.

A feature that requires further research is the **Group_ID** variable. The current choice of country and segment code was made based on empirical estimations, so it would be beneficial to find which variables affect claim rates and how important they are. This will improve the data partitioning and clustering processes.

Appendix A

Claim rate, claim number tables

*All tables are based on Vehicle Set 1

A.1 Claim rate table

CONFIDENTIAL DATA

A.2 Claim number table

CONFIDENTIAL DATA

Appendix B

Clustering Results

CONFIDENTIAL DATA

Bibliography

[1] Annual and Sustainability Report 2015,
`https://www.scania.com/group/en/scania-at-a-glance/`

[2] Scania Key Figures,
`https://www.scania.com/group/en/key-figures/`

[3] Terms and Condition of Basic Warranty Period

[4] Calinski T, Harabasz J. *"Dendrite Method for Cluster Analysis."*. Communications in Statistics Theory and Methods, 3(1), 1-27, 1974

[5] Duda RO, Hart PE *"Pattern Classification and Scene Analysis."*. John Wiley & Sons, New York. 1973

[6] Lebart L, Morineau A, Piron M *"Statistique Exploratoire Multidimensionnelle."*. Dunod, Paris. 2000

[7] Oded Maimon, Lior Rokach *"The Data Mining and Knowledge Discovery Handbook"*. Springer, 2005

[8] Ward JH *"Hierarchical Grouping to Optimize an Objective Function."*. Journal of the American Statistical Association, Vol. 58, No. 301, p. 236-244. 1963

[9] C. Fraley and A. E. Raftery *"Model-based clustering, discriminant analysis, and density estimation."*. Journal of the American Statistical Association, Vol. 97, No. 458, p. 611-631. 2002

[10] Pang-Ning Tan, Michael Steinbach, Vipin Kumar *"Introduction to Data Mining"*. Ch.8, Pearson Addison Wesley, 2006

[11] Malika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs *"NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set"*. Journal of Statistical Software, Volume 61, Issue 6, 2014

[12] Glenn W. Milligan, Martha C. Cooper *"An examination of procedures for determining the number of clusters in a data set"*. Psychometrika, VOL. 50, NO. 2, p. 159-179, 1985

[13] A.D. Gordon *"Classification"*. Chapman & Hall, 1985

[14] Fraley C. and Raftery A.E. *"How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis"*, Technical Report No. 329. Department of Statistics University of Washington, 1998.

[15] Han, J. and Kamber M. *"Data Mining: Concepts and Techniques"* Morgan Kaufmann Publishers, 2001.

[16] Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta *"clValid, an R package for cluster validation"* Journal of Statistical Software, Volume 25, Issue 4, 2008.

[17] Voinov V., Nikulin M., Balakrishnan N. *"Chi-Squared Goodness of Fit Tests with Applications"* Academic Press, 2013.