UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Visualisatie en interpretatie van het rijgedrag van heftruckchauffeurs aan de hand van een framework gebaseerd op indoor positie bepaling en clustering

Loris Gallo

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :
Prof. dr. Benoit DEPAIRE

COPROMOTOR :
Prof. dr. Niels MARTIN

2020
2021

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Visualisatie en interpretatie van het rijgedrag van heftruckchauffeurs aan de hand van een framework gebaseerd op indoor positie bepaling en clustering

Loris Gallo

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :
Prof. dr. Benoit DEPAIRE

COPROMOTOR :
Prof. dr. Niels MARTIN

Visualisatie en interpretatie van het rijgedrag van heftruckchauffeurs aan de hand van een framework gebaseerd op indoor positiebepaling en clustering

Voornaam, Naam: Loris, Gallo

Masterproef BI - 3528
Faculteit Bedrijfseconomische Wetenschappen, Universiteit Hasselt, Campus Diepenbeek, Diepenbeek, België

20 augustus 2021

Abstract: In de voorbije decennia is er al meerdere malen onderzoek verricht naar de besturing van gemotoriseerde voertuigen, evoluerend van een puur psychologisch onderzoek tot data gedreven bepalingen van het rijgedrag. Voor de logistieke sector, meer bepaald de warehouses, is het verrichte onderzoek naar rijgedrag gering. Vertrekkend vanuit deze context zal dit onderzoek trachten, om aan de hand van een nieuw opgesteld framework, het rijgedrag te visualiseren, te categoriseren en te interpreteren. De locatie data zal meerdere stappen moeten ondergaan om het mogelijk te maken deze data te visualiseren over verschillende tijdsbestekken. De clustering naar gelang de variabelen brengt met zich mee dat er een onderscheid gemaakt kan worden tussen de verschillende rijgedragingen doorheen de tijd. Om de uiteindelijke resultaten en visualisaties te kunnen interpreteren en te categoriseren zullen eerst de categorieën en de categorisatie variabelen worden gedefinieerd om de bevindingen correct te kunnen behandelen. Deze bevindingen zullen verder worden toegelicht aan het einde van deze paper.[a]

Sleutelwoorden: Rijgedrag, Real-time, Clustering

[a]Deze masterproef werd geschreven tijdens de COVID-19 crisis in 2020-2021. Deze wereldwijde gezondheidscrisis heeft mogelijk een impact gehad op het schrijf- en verwerkingsproces, de onderzoekshandelingen en de onderzoeksresultaten die aan de basis liggen van dit werkstuk.

1. Introductie

De besturing van voertuigen in de logistieke wereld, meer specifiek het besturen van heftrucks zorgt jaarlijks voor een dodental van ongeveer 85 personen. Verder zijn er jaarlijks rond de 34.900 accidenten die leiden tot ernstige blessures[1]. Vele technische aanpassingen zijn de revue gepasseerd met als doelstelling de veiligheid te bevorderen. Met de constante technologische vooruitgang ontstaan er ook andere technieken die zouden kunnen bijdragen tot deze bevordering. Deze vooruitgang laat toe specifiekere onderzoeken uit te voeren die voorheen ondenkbaar waren. Één van die onderzoeken die mogelijk werd gemaakt door middel van de vooruitgang is het bepalen van de positie van een heftruck in een indoor omgeving. In tegenstelling tot indoor positiebepaling staat outdoor positiebepaling al een heel stuk verder in de evolutie. De GPS-technologie wordt in zowat elk gemotoriseerd voertuig voor de openbare weg geïnstalleerd. Deze voorsprong aan uitrol van de technologie zorgt er voor dat onderzoeken gerelateerd aan outdoor positie bepaling al verder uitgewerkt zijn. Echter zijn er in de wereld van indoor positie bepaling verschillende technieken ontstaan doorheen de tijd: RFID-tags, GPS, WLAN ... [1] Een andere optie is het gebruiken van Bluetooth. Het is deze laatste techniek die in dit onderzoek zal worden gehanteerd.

Onderzoek naar het rijgedrag van heftruckchauffeurs is zeer beperkt. De technische aspecten van een heftruck zijn al meerdere malen bestudeerd, alsook de indeling van een warehouse. Echter worden er weinig middelen besteed om de wijze van verplaatsingen onder de loep te nemen. Aan de hand van de real-time locatie data die wordt aangeleverd voor dit onderzoek is er de mogelijkheid om deze rijgedragingen te visualiseren en te interpreteren. De bevindingen uit deze dataset kunnen vervolgens bijdragen tot een categorisering van de rijgedragingen. Zodoende zou het mogelijk zijn om in de toekomst de bestuurders te kunnen corrigeren aan de hand van de real-time data. Het onderzoek zal zich toespitsen op het visualiseren, categoriseren en interpreteren van het rijgedrag van heftruckchauffeurs aan de hand van een nieuw opgesteld framework. Dit framework zal worden getest met behulp van een real-time dataset waarop verschillende clustering technieken zullen worden toe-

[1]https://www.mccue.com/content/forklift-accident-statistics

gepast. Door middel van dit framework wordt er getracht om de rijgedragingen van heftruckchauffeurs te visualiseren, te categoriseren en te interpreteren.

Voor het categoriseren van de data kijken we naar reeds verrichte onderzoeken betreffende categorisatie. Uit deze verschillende onderzoeken blijkt dat drie parameters kunnen bepalen in welke mate een bestuurder als agressief kan worden gecategoriseerd. Snelheid, acceleratie en vertraging zijn de drie hoofdvariabelen [2]. Een constante overdreven snelheid gaat de bestuurder in kwestie bijgevolg eerder plaatsen in de 'zeer agressieve' categorie. Vijf verschillende soorten categorieën zullen worden gebruikt. Evoluerend van 'niet agressief' tot 'agressief'.

Data wordt in het onderzoek aangeleverd door BlooLoc. Blooloc weet aan de hand van Bluetooth technologie de positie van assets te tracken, in deze case heftrucks. De geleverde dataset zal worden gebruikt om het onderzoeksdoel zo compleet mogelijk te volbrengen.

Het onderzoek zal na toelichting van het framework aanvatten met het opwaarderen van de aangeleverde data alsook met de nodige data transformaties. Deze datasets zullen worden opgesplitst in verschillende verdelingen naar gelang de tijd. Zo kan het rijgedrag onderzocht worden per dag, per uur, per kwartier of zelfs per minuut. Op de data zullen clustering analyses worden toegepast. Dit laat ons toe de data op te splitsen in meerdere partities waarbij de data punten in dezelfde partitie gelijkend zijn aan elkaar en data punten behorend aan een andere cluster verschillend zijn [3]. In deze paper zullen clustering technieken K-Means en BIRCH worden toegepast. Verdere toelichting van deze stappen zijn terug te vinden in onder sectie 3.

Verder zal in dit onderzoek onder sectie 2 de relevante literatuur worden toegelicht. Sectie 3 focust zich op de bouw van het framework en de te volbrengen stappen om het framework toe te passen. In sectie 4 wordt het empirisch onderzoek volbracht en ten slotte worden in de resterende secties 5 en 6 respectievelijk de conclusie en het aanvullend werk aangekaart.

2. Gerelateerd werk

2.1. Positiebepaling heftruck

In 2017 werd al getracht een basis te leggen voor de indoor positiebepaling van vorkliften. De groei van de economie en de stijging in het gebruik van vorkliften zou een potentieel gevaar kunnen vormen voor de veiligheid van de bestuurder en de omstaanders. De data verzameling in [4] is gelijkend aan de wijze waarop dit onderzoek wordt gevoerd. De tag wordt geïnstalleerd op de vorklift en deze communiceert met 'locators'. Twee jaar later werd een gelijkaardig onderzoek uitgevoerd in Spanje. Ook hier lag de focus op het onderzoeken van de zuiverheid van de positiebepaling. Één technologische toepassing die interessant is, vindt men terug bij de positionering van de tags op de vorklift. De tags worden geplaatst op de cabine en op de vork zelf [5]. Een techniek die een extra dimensie geeft aan de dataverzameling wanneer het aankomt op positiebepaling. Het zijn deze twee papers die aantoonden dat de werkwijze bij het installeren van de 'locators' op de heftrucks als correct kunnen worden beschouwd. Een verdere aanpassing bij de installatie was niet nodig. In 2020 werd een eerste soort onderzoek opgesteld waarbij verschillende meettechnieken werden toegepast en geanalyseerd aan de hand van clustering. Een onderzoek zo uitgebreid als het geval was in [6] is een serieuze stap naar het visualiseren van rijgedrag alsook het toekennen van een rijsentiment aan de bestuurder. De paper focust zich op meerdere aspecten die kunnen bijdragen tot het categoriseren van bestuurders.

2.2. Gedragsprofilering rijgedrag voertuigen

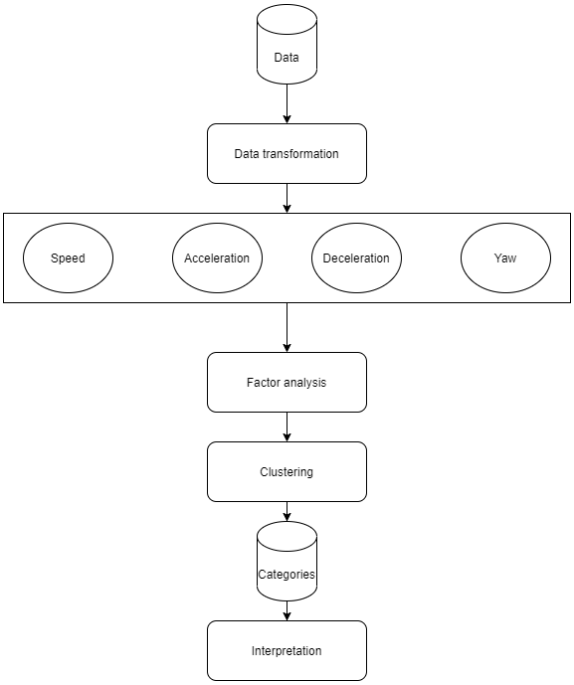
Het profileren van rijgedrag is een onderzoek dat in meerdere scenario's boven komt drijven. Zo werd in 2010 door Wang, J., Lu, M., & Li, K. een onderzoek verricht naar het categoriseren van longitudinaal rijgedrag. In [7] werd het onderzoek verricht aan de hand van verschillende parameters, variërend van de gemiddelde acceleratie tijd tot de gemiddelde tijd voor het veranderen van accelereren tot vertragen. Om te kunnen categoriseren werd de data verwerkt in het K-means clustering algoritme. Uit dit voorgaand onderzoek werd aangetoond op welke wijze het rijgedrag kan worden gevisualiseerd en geïnterpreteerd. Deze werkwijze, waarbij K-means clustering werd toegepast, zal bijgevolg ook in dit onderzoek worden gehanteerd. Verder werd in 2013 het artikel 'Design of Driving Behavior Pattern Measurements Using Smartphone Global Positioning System Data' [8] gepubliceerd. In dit artikel geraken we dichter tot het doel dat dit onderzoek tracht te bereiken.Hier heeft men aan de hand van vooropgestelde parameters en verkregen datapunten een beeld gevormd van het rijgedrag van de persoon in kwestie. In eerste instantie was het noodzakelijk om de data op te waarderen om vervolgens hiermee tot inzichten te komen. Deze paper vormt een basis tot het categoriseren van rijgedrag voor dit onderzoek.

2.3. Warehouse veiligheid

Over de veiligheid in de warehouses zijn er enkele bronnen die wijzen op aspecten die een invloed kunnen hebben op de veiligheid alsook artikels die zich focussen op het gedrag van werknemers in deze sector. Er zijn verschillende factoren die het gedrag ten opzichte van veiligheid beïnvloeden. Zo zal ook het totale gevoel van veiligheid binnen een warehouse een aanzienlijke invloed uitoefenen wat betreft het rijgedrag. De subcultuur in een setting heeft een significante invloed op het sentiment omtrent de veiligheid [9]. Dit sentiment zou wel eens kunnen worden veranderd zodra er aan de hand van real-life data kan worden aangetoond of dit sentiment wel degelijk zo veilig is. Een studie uit 2017 [10] richt de aandacht op de veiligheid wanneer het aankomt op het rijgedrag van heftruckchauffeurs in een warehouse. In deze studie werd veiligheid beschreven aan de hand van de relatie tussen het liften van de vork tijdens beweging. Een chauffeur die de vork reeds bewoog terwijl de heftruck nog niet volledig gestopt was, kreeg een lagere score wanneer het aankomt op veiligheid. Boehning, M. deed reeds onderzoek naar de mogelijke verbetering in de veiligheid en efficiëntie bij kruisingen in een warehouse. Aan de hand van PAN-Robots is een heftruck ertoe in staat om het pad van tegenliggers te herkennen in een warehouse. Een verbetering in de veiligheid aan de hand van deze techniek is zeker mogelijk zonder inefficiënt te remmen bij elke kruising indien er geen ander object in de buurt is [11]. Deze onderzoeken dienen als een impuls voor het onderzoek dat verder in deze paper zal worden uitgewerkt. De moeite gestoken in deze onderzoeken en bijgevolg de aandacht voor veiligheid in warehouses tonen aan dat het onderzoek dat gevoerd wordt een bijdrage kan leveren tot het verbeteren van de veiligheid in de warehouses op langere termijn.

3. Rijgedrag bepaling framework

Zoals besproken in sectie 2 is er nog maar weinig onderzoek gedaan in de heftrucksector die gericht is naar de veiligheid waarbij real-time locatie data wordt gebruikt. Het is omwille van deze schaarste dat in deze paper er een propositie wordt gemaakt voor een eerste framework die de rijgedragingen van heftruckchauffeurs visualiseert, categoriseert en interpreteert. Het framework zal steunen op de aangeleverde real-time locatiedata afkomstig van de heftrucks alsook op verschillende clustering technieken. Verder richt het zich tot het begrijpen van de rijgedragingen om toekomstig de gedragingen van een heftruckchauffeur op een automatische manier te verwerken. Het voorgestelde framework is zichtbaar in



Figuur 1: Framework

figuur 1. Om tot een uiteindelijke interpretatie en categorisatie te geraken van het rijgedrag van een heftruckchauffeur dienen verschillende stappen te worden genomen. Zo zal allereerst de data moeten worden aangeleverd. Vervolgens wordt deze data getransformeerd en opgewaardeerd. Eens deze stap is volbracht is het mogelijk om de multi-dimensionale dataset aan de hand van factor analyse om te vormen tot een gereduceerd formaat. Deze laatste vorm zal dan worden gebruikt om te clusteren waarbij de resultaten de verschillende categorieën zullen weergeven. Wat rest is een interpretatie van de bekomen resultaten om de categorieën te kunnen definiëren.

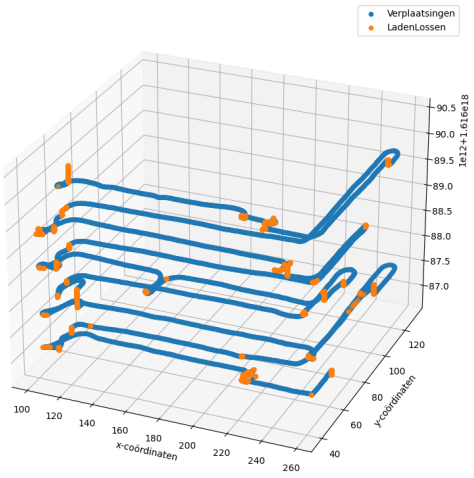
3.1. Data

De aangeleverde data zal van groot belang zijn in het succesvol gebruiken van het framework. Om de rijgedragingen van heftruckchauffeurs real-time te kunnen categoriseren dient de aangeleverde data ook naar gelang te zijn. Zo zal de data voor dit framework moeten voldoen aan verschillende eisen. Allereest dient er genoeg datapunten aanwezig zijn. Bij het gebruiken van een dataset waarbij het tijdsinterval tussen twee verschillende punten groter is dan 2 seconden is het mogelijk dat bepaalde manoeuvres niet zullen worden opgevangen. In dit geval zou een scherpe draaibeweging niet het juiste gewicht met zich meekrijgen bij de uiteindelijke clustering omdat deze simpelweg niet werd opgemerkt. Een tweede vereiste bij de data is dat er voldoende meeteenheden aanwezig zijn. Om de data te kunnen transformeren en opwaar-

deren moet de ontvangen data voldoende inhoud bevatten. De noodzakelijke factoren zijn zichtbaar in tabel 1.

Dataset	
Kolom	**Uitleg**
time	Het tijdstip van de meting
x_pos	Positionering van de heftruck ten opzichte van de x-as
y_pos	Positionering van de heftruck ten opzichte van de y-as
heading	Geeft aan in welke richting het voertuig zich verplaatst (rad)

Tabel 1: Verwachte dataset



Figuur 2: Evolutie data punten ten aanzien van de tijd

3.2. Data transformatie

Startend met deze tabel aan factoren uit de data wordt het mogelijk om de data te transformeren en op te waarderen. Verschillende stappen zullen nodig zijn om te data om te vormen tot een dataset die toelaat om verder te gaan tot de factor analyse. Vier nieuwe variabelen zullen worden gecreëerd namelijk, snelheid, acceleratie, vertraging en draaisnelheid. We starten met de bepaling van de snelheid. Deze wordt berekend door middel van de afgelegde afstand over de tijd tussen twee opeenvolgende datapunten. Acceleratie en vertraging komen voort uit de net verkregen snelheidsvariabele. Acceleratie is de verandering in snelheid over de tijd tussen twee opeenvolgende datapunten. Bij een vermindering van de snelheid over de tijd zien we een negatieve waarde verschijnen die instaat voor de vertraging. Acceleratie en vertraging behoren tot de top indicatoren wanneer het aankomt op de bepaling van agressief rijgedrag [12]. Ten slotte is er nog de draaisnelheid, ook wel de yaw velocity genoemd. Aan de hand van de yaw velocity is het mogelijk om de verandering in de 'heading' doorheen de tijd te gebruiken als indicator van rijgedrag. Yaw geeft ons in die zin een beeld van de agressiviteit van de verandering in richting. Een yaw gelijk aan 0,5 is gelijk aan een verandering van 28,64 graden per seconde. De creatie van deze vier indicatoren zorgt ervoor dat we de dataset kunnen opwaarderen alsook zuiveren.

In het bepalen van het rijgedrag van een heftruckchauffeur zijn er twee soorten van verplaatsingen in een warehouse. Enerzijds zijn er de verplaatsingen naar de laad- en lospunten en andersom, anderzijds zijn het de acties van het laden en lossen zelf. Het categoriseren van acties tijdens het laden en het lossen zijn veel sterker gericht op behandeling van de vork van de heftruck en minder op de verplaatsing van de heftruck zelf. Dit framework zal bijdrage leveren wanneer er gekeken wordt naar de verplaatsingen van heftrucks. Om het onderscheid tussen deze twee acties te maken werden de punten in de dataset waarbij voor zowel de x als de y-coördinaat met minder dan 0.5% verschilde van de voorgaande coördinaten en waarbij de snelheid lager was dan 2,5 km/h (trial-and-error), aanzien als een laad- of lospunt. Figuur 2 geeft direct aan, zoals vermeld in de legende, waar de laad- en lospunten zich bevonden en dat deze overeenstemden met al deze punten in de visualisatie. Op de visualisatie in figuur 2 is duidelijk te zien dat de punten waarop een heftruck aan het laden of het lossen is in het oranje worden gemarkeerd. Dit stemt overeen met de beweging die zich op dat moment voordoet. Flauwe bochten of simpele draaibewegingen om over te gaan naar een volgend gangpad worden niet erkend als laad- of lospunten, ook dit is een bevestiging van de correcte filtering. Dit framework zal als gevolg enkel waardevol zijn voor de *verplaatsingen*.

Na de initiële opwaardering naar de vier noodzakelijke indicatoren en de afbakening van de verplaatsingen van het laden en het lossen dient er een laatste set van aanpassingen te worden doorgevoerd. Op dit moment beschikt het framework over een dataset met duizenden datapunten die elk een snelheid, acceleratie, vertraging en draaisnelheid hebben. Om verder te gaan tot de factor analyse en de uiteindelijke clustering dient deze dataset enkele extra indicatoren te verkrijgen aan de hand van de vier nieuwe variabelen en vervolgens te worden samengevat. Indien het framework zicht toespitst op intervallen van een

uur moeten voor elk van de vier indicatoren een gemiddelde per uur worden genomen. Verder worden volgende indicatoren gecreëerd voor de finale dataset waarbij de data transformatie uiteindelijk beschikt over 12 variabelen per tijdsinterval:

- *snelheid boven 11,5 km/h*: percentage van de tijd ($V_{11.5}$) - Aangetoond door *Material Handling Equipment Distributors Association* wordt een maximale snelheid van 11,5 km/h aangeraden. Elk punt met een hogere snelheid wordt hierin opgenomen.

- *snelheid*: gemiddelde snelheid (V_{mn}) en de standaard deviatie (V_{sd}) worden als statistische waarden mee opgenomen.

- *acceleratie boven de 8,83 km/h/s*: percentage van de tijd ($A_{8.83}$) - een acceleratie met een kracht van meer dan 0,25g wordt gezien als gevaarlijk [13]. Een kleine omzetting van g-krachten naar km/h/s brengt ons dat 8.83 km/h/s het maximum is. Elk punt met een hogere acceleratie wordt hierin opgenomen.

- *acceleratie*: gemiddelde acceleratie (A_{mn}) en de standaard deviatie (A_{sd}) worden als statistische waarden mee opgenomen.

- *vertraging boven de 8,83 km/h/s*: percentage van de tijd ($D_{8.83}$) - een vertraging met een kracht van meer dan 0,25g wordt net zoals bij acceleratie gezien als gevaarlijk [13]. Ook hier wordt 8,83 km/h/s als benchmark gebruikt. Elk punt met een hogere vertraging wordt hierin opgenomen.

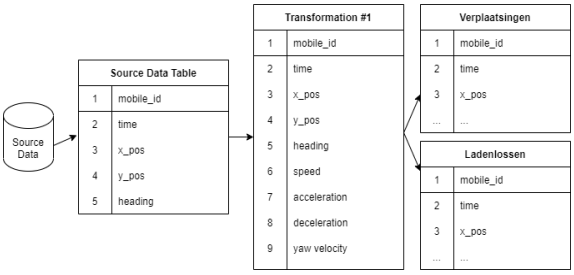
- *vertraging*: gemiddelde vertraging (D_{mn}) en de standaard deviatie (D_{sd}) worden als statistische waarden mee opgenomen.

- *yaw rate*: maximum (Y_{max}), gemiddelde (Y_{mn}) en standaarddeviatie (Y_{sd}) worden gehanteerd om de draaisnelheid te vertegenwoordigen in de berekeningen.

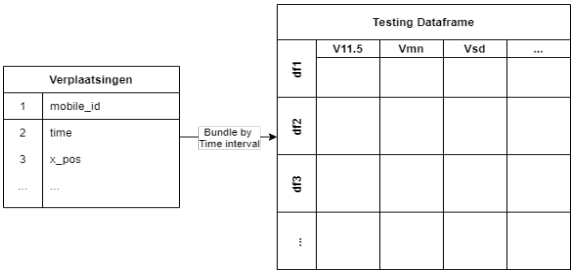
De volledige stappen voor de datatransformatie die nodig zijn om het framework te kunnen volgen zijn afgebeeld in figuren 3 en 4.Een laatste bewerking vooraleer we kunnen overgaan naar de factor analyse is het normaliseren van de data voor het tijdsinterval.

3.3. Factor analyse

De volgende stap die volbracht dient te worden is het analyseren van de factoren. Om de factoren te verminderen in dimensionaliteit en de resultaten visueler te maken zal de factor analyse worden toegepast. Het bepalen van het aantal dimensies die noodzakelijk zijn om de dataset te beschrijven kan op meerdere wijzen gebeuren. Er kan gebruik gemaakt worden van een scree plot. Deze zal visueel aangeven wat de meerwaarde is bij het gebruiken van een extra-dimensie door middel van de eigenwaarden. De eigenwaarden boven de 1 geven aan dat de bijdrage van deze factor significant is. Een tweede mogelijkheid is kijken naar het percentage van de hoeveelheid informatie die wordt meegegeven bij elke dimensie. De factoren zullen naar mate ze in aantal stijgen een lagere marginale opbrengst opleveren. Het moment dat de stijging van de variantie die wordt toegelicht door een factor een stuk lager ligt dan de toelichting van de vorige factor wordt het aangeraden om het aantal dimensies te nemen bij de laatst significante factor. Eens het aantal dimensies zijn bepaald kan de vermindering aanvatten. Het verminderen van de dimensionaliteit kan er echter voor zorgen dat door middel van een sterke onderlinge correlatie de factoren meerdere basisvariabelen zullen toelichten. De variantie van een factor zal op die wijze voor meerdere basisvariabelen dicht bij 1 liggen. Het moment dat het gaat over variabelen die bij een toelichting door middel van één factor zo hun onderscheid verliezen en bijgevolg geen verschillende patronen kunnen herkennen, dient men een andere methode te hanteren. De factoren zullen worden geanalyseerd aan de hand van een rotatie methode om de matrix met de ladingen van de factoren te roteren. Rotatie metho-



Figuur 3: Transformatie bron data tot 'verplaatsingen' en 'ladenlossen'



Figuur 4: Verplaatsingen dataset tot onderzoek set

des slagen erin om het percentage van de variantie per factor te beïnvloeden. Er zijn verschillende rotatie methodes die slagen in deze opzet: quartimax, varimax, equamax, parsimax en factor parsimony. In dit framework wordt varimax toegepast omwille van de populariteit alsook het feit dat varimax minder snel een algemene factor gaat produceren, precies wat we proberen te vermijden [14]. Varimax zelf is een statistische techniek die tracht de relatie tussen de verschillende factoren aan te geven. De rotatie bij varimax heeft de intentie om de variantie onderling zo sterk mogelijk te verhogen. Deze verhoging van de variantie zorgt ervoor dat elke rotated component de correlatie met de basisvariabelen beter weergeeft. Om er dus voor te zorgen dat enkel basisvariabele *a* wordt beschreven door rotated component 1 zal varimax de correlaties roteren. Zo zal rotated component 1 enkel basisvariabele *a* beschrijven en niet meerdere basisvariabelen. De correlatie tussen basisvariabele a en rotated component 1 dient in dit opzicht dus te worden vermeerderd en de de correlatie met de andere variabelen dient te worden afgenomen [15] [16].

3.4. Clustering

De data verkregen uit de factor analyse kan vervolgens worden geclusterd. Het doel van het framework is, zoals aangehaald in de inleiding, om unsupervised het rijgedrag van een heftruckchauffeur te categoriseren. Om te kunnen categoriseren op een wijze waarbij voorgaande kennis betreffende het rijgedrag van een chauffeur niet wordt gebruikt, werd de keuze gemaakt om verder te gaan met clustering. Clustering is een methode van unsupervised learning die wordt gebruikt om individuele objecten te categoriseren in groepen met gelijkaardige eigenschappen [2]. Aan de hand van clustering zullen we dus te zien krijgen welke groepen van objecten zich dicht genoeg bij elkaar bevinden om deze als één groep te definiëren. Veel clustering methodes zijn vandaag de dag voor handen. Elk van deze clustering methodes behoort dan ook nog eens tot een groter geheel van soorten clustering methodes die onder de algemene term 'clustering' vallen. Clustering kan worden opgedeeld in twee grote groepen namelijk, de hiërarchische methodes en de partitie methodes. Voor dit onderzoek zal van beide groepen één methode worden uitgekozen afhankelijk van de eigenschappen van deze specifieke algoritmes. Na een verkennende studie van de verschillende methodes werd gekozen voor BIRCH en K-means [17]. Deze keuze wordt hieronder verder toegelicht.

3.4.1. K-means clustering algoritme

K-means clustering is een welbekend clustering algoritme voor het uitvoeren van unsupervised learning taken [3]. K-means werd gekozen uit de verschillende partitie methodes om de verschillende voordelen. Allereerst is dit algoritme het simpelste en meest gebruikte algoritme om te clusteren. Op deze manier dient K-means in dit onderzoek als een benchmark algoritme. De meerwaarde die kan worden geleverd is al bewezen en hiermee kan het dus worden gebruikt om mogelijks tot een eerste set van resultaten te komen die voldoen aan de verwachtingen van het framework [18]. Verder wordt K-means gehanteerd vanwege de lineaire complexiteit. Dit laat ons toe om met relatieve grote datasets aan de slag te kunnen. Ook ten opzichte van de hiërarchische methodes is dit een voordeel aangezien deze te maken hebben met nonlineaire complexiteit [17]. Op een iteratieve manier gaat K-means trachten een finaal resultaat te genereren. Een bepaald aantal clusters worden meegegeven samen met de bijhorende dataset. K-means zal starten met het bepalen van de centroids om deze iteratief te verbeteren. Elk punt van de dataset wordt toegekend aan een centroid gebaseerd op de euclidische afstand tot die centroid. De euclidische afstand wordt gekozen aangezien het onderzoek wordt verricht met een dataset van rijen met numerieke waarden. Een nadeel dat verbonden is aan het gebruik van K-means is, dat de eerste selectie van centroids slecht kan zijn. Op dat moment zal het clustering algoritme meerdere iteraties nodig hebben om te komen tot het verwachte resultaat. Bij een grote dataset kan dit wel eens voor problemen zorgen en langer duren. K-means wordt in dit onderzoek omwille van de erkenning en de meerdere voordelen gebruikt als een basis clustering algoritme. K-means clustering wordt als effectief en deterministisch beschouwd [19].

Om de clustering stap in het framework unsupervised toe te passen moet er ook voor de bepaling van het aantal clusters beroep worden gedaan op een unsupervised methode. Bij de clustering volgens K-means wordt de *Elbow method* gehanteerd. De *Elbow method* kijkt naar het percentage van de variantie die wordt toegelicht door de hoeveelheid clusters. Dit wilt dus zeggen dat wanneer we het percentage berekenen van de variantie, we het aantal clusters moeten kiezen waarbij het toevoegen van een extra cluster niet opweegt tegen de minimale marginale opbrengst van deze extra cluster [20]. Het percentage van de variantie verkregen door de clusters wordt afgebeeld ten aanzien van het totaal aantal clusters. Deze visualisatie zal aantonen hoe groot de marginale bijdrage is van een extra cluster. De eerste clusters zullen de

grootste bijdragen leveren, echter zal op een gegeven moment de bijdrage drastisch dalen waardoor er een hoek zichtbaar wordt in de grafiek. Deze hoek is de "elbow". Op dit punt bevindt de optimale clusterhoeveelheid zich [21].

3.4.2. BIRCH clustering algoritme

Het tweede clustering algoritme dat werd gekozen is BIRCH, Balanced Iterative Reducing and Clustering. BIRCH is zoals voorheen aangehaald een hiërarchische methode wat in compleet contrast staat met K-means. Het hanteren van dit clustering algoritme geeft ons dus twee zeer verschillende clustering algoritmes. BIRCH wordt gekozen uit de tak van hiërarchische clustering methodes omwille van de capaciteit tot het verwerken van grote datasets. Het moment dat de dataset exponentieel groeit, bij inzet van meerdere heftrucks over meerdere periodes is het noodzakelijk dat er zich geen beperking voordoet op de kwaliteit van het proces. Dit komt omdat BIRCH een lokaal algoritme is, wat met zich meebrengt dat het algoritme niet alle data bekijkt bij het maken van clustering beslissingen. Wanneer men een grote dataset gaat clusteren, dan gaat BIRCH regionen van datapunten met een hoge dichtheid niet individueel beschouwen, maar als groep [22]. Ten tweede zorgt BIRCH er als eerste clustering algoritme voor dat er iets gedaan wordt met uitschieters [23]. Verder is BIRCH ook incrementeel. Dit houdt in dan niet altijd de volledige dataset aanwezig moet zijn. In het geval van real-time clustering waarbij elk datapunt direct wordt geleverd hoeft BIRCH niet de volledige dataset te gebruiken, één scan voldoet om de punten toe te kennen aan de passende groepen[24]. BIRCH zou in vergelijking met K-means er ook in moeten slagen om sneller uit de lokale minima partities te geraken.

Ook voor BIRCH zal er een unsupervised methode worden gehanteerd om het aantal clusters te bepalen. In tegenstelling tot K-means zal bij BIRCH *Silhouette* worden gebruikt. *Silhouette* toont aan de hand van de berekeningen aan welke hoeveelheid van clusters het meest geschikt is om de clustering mee aan te vangen. Deze methode heeft het voordeel dat het enkel afhankelijk is van de partitie van de verschillende data objecten en niet van het clustering algoritme. Het is vanwege deze eigenschap dat *Silhouette* kan worden gebruikt om te bepalen welke hoeveelheid aan clusters geschikt is voor de clustering. Verder zorgt *Silhouette* er ook voor dat "artificiële" samenvoegingen van clusters worden vermeden. Indien twee groepen van objecten met een grote dichtheid ver van elkaar liggen zal dit worden opgemerkt aangezien de ongelijkheid tussen de verschillende objecten wordt doorgerekend in de berekening van de silhouetten. [25]. Een Silhouette plot geeft ons een visualisatie van silhouetten waarbij een hoeveelheid aan clusters worden vertoond langs elkaar. Deze side-by-side view toont op een snelle en efficiënte manier aan welke hoeveelheid clusters het meest geschikt is voor de problematiek voor handen. Elke staaf in de visualisatie geeft de kwaliteit van de clustering per hoeveelheid clusters.

3.5. Interpretatie

Ten slotte volgt de interpretatie en de categorisatie. Het categoriseren van de rijgedragingen kan aanzien worden als een samenloop van data en sentiment. Aan de hand van de data kunnen we tot op een bepaalde correctheid ondervinden waar een chauffeur zich bevindt op een bepaald tijdstip. De feitelijke data zal worden gebruikt om te geraken tot een sentiment. Er zijn vele onderzoeken verricht om te bepalen of het rijgedrag van een bestuurder van een voertuig kan worden omschreven als risicovol of agressief. Zo werd al eerder bepaald dat jonge bestuurders agressiever rijden in vergelijking met oudere bestuurders die over meer ervaring beschikken [26]. De interpretatie van de resultaten zullen op deze wijze onderbouwd zijn met reeds bewezen theorieën wanneer het aankomt op de bepaling van risico en agressiviteit [27] [28] [29]. Het combineren van deze theorieën met de resultaten voor handen biedt de mogelijkheid tot het genereren van kennis en inzicht. Een framework dat onderbouwd is met zowel data als reeds verrichte onderzoeken is in staat om op een neutrale wijze de categorisering van de rijgedragingen te verwerken. Het framework is in dat opzicht bij de interpretatie en categorisatie dus niet biased [30]. Het interpreteren van de clusters is in samenhang met de 'loadingsmatrix'. De variantie waardes die verkregen zijn bij de varimax factor analyse dienen op een juiste manier volgens de waardes te worden geïnterpreteerd. In het geval dat de variantie bij een rotated component dicht bij -1 ligt wanneer we bijvoorbeeld kijken naar de basisvariabele snelheid, dan zou dit willen zeggen dat de cluster die ten opzichte van de as van deze rotated component aan de linkerzijde ligt, en zich dus in het negatieve deel begeeft, een hogere snelheid zal hebben. Op dat moment kunnen we de cluster die deze eigenschap bevat ten aanzien van de snelheidscomponent aanschouwen als agressiever. De labeling van de data naar gelang agressiviteit is afhankelijk van positie van de cluster en ook afhankelijk van de loadings-matrix. Twee test datasets worden meegegeven in de clustering om zo te beschikken over een referentie. Deze sets zijn gecreëerd op basis van de literatuur alsook door middel van ei-

gen interpretatie. Een gecreëerde agressieve bestuurder gaat vaker, procentueel gezien, boven de 11,83 km/h rijden als gemiddeld terwijl een niet-agressieve bestuurder dit haast nooit zal doen. Het is volgens deze redenering dat de twee referentie sets werden gecreëerd. Voor de uiteindelijke categorisatie kunnen we definiëren tot welke categorie een cluster behoort door de uitkomsten bij de verschillende componenten te combineren. Clusters worden aan de hand van de verkregen resultaten gebruikt om een afbakening te generen tussen de verschillende resultaten. Zo zullen er op de verschillende assen grenzen ontstaan die een agressief rijgedrag zullen onderscheiden van neutraal rijgedrag. Na het toekennen van de clusters aan niet-agressieve, licht niet-agressieve, neutrale, licht agressieve of agressieve rijgedragingen kunnen we deze uitkomsten voor de verschillende componenten samenvoegen. We gaan op dit moment voor elke erkende cluster bepalen welke karakteristieken aanwezig zijn voor de componenten. In het geval dat verminderen naar drie componenten voldoende is en dat een cluster als resultaat voor die drie verschillende componenten beschikt over niet-agressief, neutraal en agressief dan is de uiteindelijke uitkomst neutraal. We kunnen dus waardes toekennen aan de verschillende toestanden. Niet-agressief wordt aanzien als -1, licht niet-agressief als -0.5 neutraal als 0, licht-agressief als 0.5 en agressief als 1. Indien de optelling van de verschillende componenten kleiner is als -1 dan kan het rijgedrag als niet-agressief worden beschouwd. Bij een resultaat groter als 1 is het rijgedrag agressief. Uitkomsten gelijk aan -1 of 1 worden respectievelijk aanzien als licht niet-agressief of licht agressief. Een sommatie die leidt tot een uitkomst in het bereik [-0.5;0.5] is tenslotte neutraal.

4. Empirisch onderzoek

Het gedefinieerde framework onder sectie 3 zal worden toegepast in een empirisch onderzoek. Hierbij is het de bedoeling om het vooropgestelde framework te valideren. In deze sectie worden resultaten gegenereerd volgens de stappen van het framework om zo te kijken of deze bruikbaar zijn voor een categorisatie van het rijgedrag. De dataset die gebruikt wordt is een csv-bestand dat bestaat uit 5 kolommen met 10.349.916 bruikbare rijen. Elke rij staat gelijk aan een datapunt in de tijd. De dataset beschrijft de periode van 15-03-2021 tot en met 07-04-2021. De site heeft een lengte van 237m en een breedte van 120m. De datapunten zelf kennen een onderling tijdsverschil tussen de 100ms en 300ms. Een voorbeeld van de gevisualiseerde data is zichtbaar in figuur 2 voor de periode van 15-03-2021 15:30 - 16:00.

4.1. Data transformatie

De data zal zoals vermeld in sectie 3 stapsgewijs worden opgewaardeerd. We starten met de bepaling van de snelheid. Een tweede en derde berekening die bijkomend zullen bijdragen, zijn de acceleratie alsook de vertraging. Deze twee indicatoren worden berekend steunend op de voorgaand berekende snelheid. Ten slotte is er nog de draaisnelheid die zal worden berekend. Startend met de vier standaard opwaarderingen berekenen we een totaal van 12 parameters die gaan dienen bij het categoriseren van de dataset *verplaatsingen*.

Om de opgewaardeerde data te kunnen gebruiken voor visualisatie en interpretatie van het rijgedrag dient bij de creatie van de 12 variabelen de data gebundeld te worden. Om data te verkrijgen die ons iets vertelt over een bepaalde tijdsspanne zullen gemiddeldes worden gehanteerd. In het geval van 'df1' uit tabel 2 worden van alle datapunten uit het eerste uur van de dag (00:00 - 01:00) het gemiddelde genomen. Het tijdsverschil zoals reeds verduidelijkt tussen twee opeenvolgende punten bevindt zich tussen de 100ms en 300ms. In dit geval beschikt elk uur over voldoende datapunten om gebruik te maken van het gemiddelde. Verdere verfijningen wat betreft het tijdsinterval zullen ook worden toegepast, zo wordt de data gesplitst om de 5 minuten alsook om de 15 minuten en een uur.

Aan de hand van deze laatste getransformeerde dataset omtrent de *verplaatsingen* zal het empirisch onderzoek worden gevoerd. De data voor de periodes 18-03-2021, 19-03-2021, 22-03-2021, 23-03-2021, 26-03-2021, 27-03-2021, 29-03-2021, 31-03-2021, 01-04-2021 en 06-04-2021 dienen als het onderzoeksdomein omwille van de data kwaliteit en kwantiteit. De andere periodes worden vermeden omdat de data te gering is en bijgevolg de data voor handen niet representatief is voor het vertoonde rijgedrag. Bovendien worden deze niet gebruikte periodes gefilterd omwille van de kwaliteit. Inconsistente data met onmogelijke verplaatsingspatronen waarbij een heftruck zich door meerdere muren verplaatst worden vermeden. De 12 parameters zullen vervolgens voor deze gekozen periodes worden gebruikt om de verschillende patronen van elkaar te onderscheiden. Voorts worden twee finale rijen als testdata meegenomen. De gelezen literatuur in combinatie met reeds verrichte onderzoeken werden twee fictieve rijen gecreëerd met als doel representatief te zijn voor niet-agressief en agressief rijgedrag. In tabel 2 wordt de agressieve rij weergegeven als dfa en wordt de niet-agressieve rij weergegeven als dfna.

Tabel 2: Testdata voor de clustering

Dataframe	$V_{11.5}$	V_{mn}	V_{sd}	$A_{8.83}$	A_{mn}	A_{sd}	$D_{8.83}$	D_{mn}	D_{sd}	Y_{max}	Y_{mn}	Y_{sd}
df1	3.77	8.09	2.99	12.52	4.43	3.33	10.75	4.24	3.28	0.98	0.11	0.18
df2	4.13	8.11	2.94	14.41	4.71	3.39	14.48	4.61	3.46	0.99	0.12	0.19
df3	2.68	7.57	3.04	14.48	4.69	3.39	12.36	4.47	3.35	0.97	0.11	0.18
df4	3.14	7.74	2.99	13.54	4.53	3.34	11.46	4.27	3.32	0.99	0.15	0.21
df5	3.37	7.89	2.92	10.93	4.44	3.21	12.15	4.49	3.34	0.95	0.13	0.19
df6	3.75	8.18	2.79	11.76	4.30	3.30	14.51	4.52	3.42	0.99	0.13	0.19
df7	1.43	6.59	3.03	12.47	4.58	3.27	9.90	4.06	3.14	0.99	0.18	0.22
df9	5.86	6.41	4.07	16.30	4.63	3.79	13.40	3.85	3.46	0.94	0.14	0.20
df12	37.50	9.80	6.26	26.31	4.54	4.58	9.52	2.72	3.24	0.91	0.11	0.18
df13	3.78	6.14	2.89	22.22	5.85	3.57	13.33	4.25	3.62	0.92	0.24	0.24
df14	2.94	5.65	3.52	15.46	4.97	3.47	15.08	4.39	3.57	0.97	0.13	0.21
df15	3.19	6.92	3.10	23.46	5.67	3.68	19.55	5.06	3.72	0.99	0.09	0.17
df16	5.90	7.22	3.05	22.69	5.06	3.92	15.09	4.45	3.74	0.82	0.11	0.19
df17	3.68	7.96	2.91	19.30	5.27	3.60	20.24	5.32	3.71	0.98	0.11	0.19
df18	3.11	7.68	2.91	18.79	5.24	3.64	16.40	4.93	3.50	0.95	0.10	0.16
df19	3.82	8.07	2.80	21.35	5.48	3.76	18.91	5.27	3.65	0.99	0.14	0.20
df20	2.63	6.82	3.00	20.10	5.33	3.67	16.78	4.87	3.58	0.96	0.12	0.17
df21	4.18	7.31	3.04	15.40	4.96	3.51	18.26	4.98	3.74	0.92	0.18	0.19
df22	3.44	7.97	2.87	19.21	5.23	3.69	17.68	5.07	3.59	0.98	0.14	0.20
df23	3.51	8.12	2.79	20.14	5.31	3.70	17.63	5.05	3.61	0.99	0.12	0.19
dfa	10.00	10.00	2.00	15.00	6.83	3.80	15.00	6.83	3.80	0.99	0.30	0.20
dfna	0.00	6.00	2.00	0.00	4.00	3.50	0.00	4.00	3.50	0.20	0.10	0.20

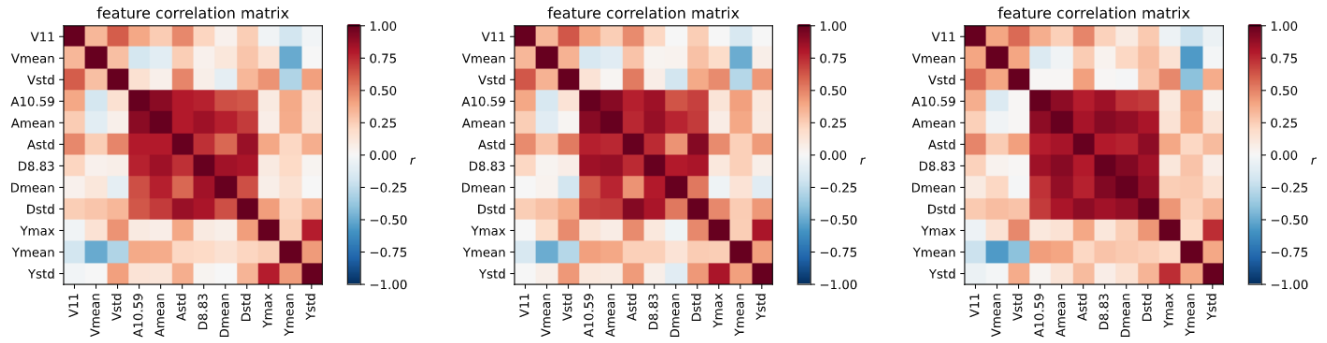
4.2. Resultaten

Op basis van de opgewaardeerde data zullen we door middel van factor analyse en de verschillende clustering technieken trachten de rijgedragingen te visualiseren en te categoriseren. Op de datasets met verschillende tijdsintervallen zal de varimax methode worden toegepast om de datapunten te clusteren. Ten slotte wordt er nader ingezoomd op de data van 18-03-2021 omwille van de variatie aan rijpatronen die zichtbaar zijn.

4.2.1. Factor analyse

De data waarop reeds verschillende bewerkingen zijn toegebracht, geeft ons een 12-dimensionale dataset. Vooraleer we beginnen we met het verminderen van het aantal dimensies, moet de data worden genormaliseerd. Dit wordt gedaan via python *StandardScaler*. Verder worden ook in figuur 5 de verschillende correlatie plots weergegeven. Deze veranderen als verwacht minimaal onderling wat met zich meebrengt dat er een vorm van consistentie teruggevonden wordt. Eens de normalisering is volbracht en we de correlatie plots hebben gevisualiseerd, continueren we met de factor analyse. Er zijn verschillende manieren om te bepalen hoeveel dimensies er moeten worden overgehouden. Een eerste mogelijkheid is het analyseren van de scree plot in figuur 6. Deze geeft ons mee dat bij de 5de dimensie de eigenwaarde nog net groter is dan 1. Vijf dimensies worden in dit geval dus aangeraden. Een tweede mogelijkheid is kijken naar het percentage van de hoeveelheid informatie die wordt meegegeven bij elke dimensie. De factor analyse geeft ons in dit geval bij 3 dimensies een percentage van 80,95% en bij 4 dimensies 89% wanneer we kijken naar de dataset met het tijdsinterval van een uur. In deze situatie wordt de afweging gemaakt tussen de meerwaarde bij het nemen van één extra dimensie en de meerwaarde bij de mogelijkheid tot het visualiseren van de uitkomst. Omwille van het feit dat een driedimensionale voorstelling 80,95% van de variantie beschrijft wordt gekozen om met deze dimensionaliteit voort te zetten. De varimax methode roteert de componenten van de correlatie matrices van de factor analyse en biedt ons zodoende nieuwe matrices [15]. Deze ladingen worden getoond in tabel 3.

Op dit moment is het mogelijk om de rotated components te linken aan de originele variabelen. Hier wordt duidelijk dat RC1 in staat voor zowel de acceleratie als de vertraging van de heftruck. RC2 vertelt meer over de gehanteerde draaisnelheid. Zowel Y_{max} en Y_{sd} zijn gecorreleerd met deze rotated component.



(a) 5 min (b) 15 min (c) 1 hour

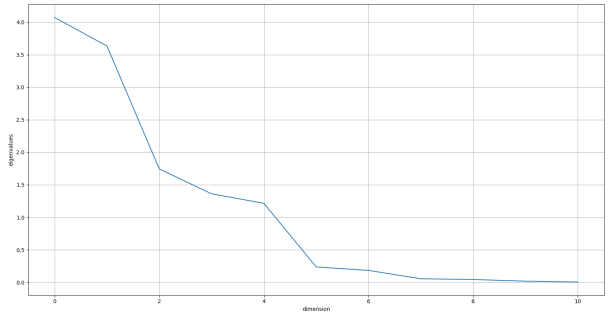
Figuur 5: Correlatie plots tijdsintervallen

5 min	RC1	RC2	RC3
$V_{11.5}$	-0.32	0.04	0.59
V_{mn}	0.01	-0.01	***0.67***
V_{sd}	-0.11	-0.42	***0.70***
$A_{8.83}$	***-0.91***	-0.06	-0.04
A_{mn}	***-0.96***	-0.02	-0.11
A_{sd}	***-0.85***	-0.33	0.36
$D_{8.83}$	***-0.88***	0.04	0.02
D_{mn}	***-0.78***	0.09	-0.10
D_{sd}	***-0.77***	-0.32	0.27
Y_{max}	-0.08	***-0.83***	0.14
Y_{mn}	-0.31	-0.41	***-0.59***
Y_{sd}	-0.08	***-0.93***	-0.03

15 min	RC1	RC2	RC3
$V_{11.5}$	-0.32	0.00	***0.54***
V_{mn}	0.00	-0.08	***0.74***
V_{sd}	-0.10	-0.48	***0.64***
$A_{8.83}$	***-0.93***	-0.08	0.08
A_{mn}	***-0.97***	-0.04	0.13
A_{sd}	***-0.82***	-0.43	-0.34
$D_{8.83}$	***-0.90***	0.01	-0.04
D_{mn}	***-0.77***	0.12	0.08
D_{sd}	***-0.77***	-0.42	-0.29
Y_{max}	-0.09	***-0.85***	-0.14
Y_{mn}	-0.32	-0.41	***0.62***
Y_{sd}	-0.08	***-0.92***	0.06

1 uur	RC1	RC2	RC3
$V_{11.5}$	-0.29	0.00	-0.49
V_{mn}	-0.10	-0.07	***-0.75***
V_{sd}	-0.06	-0.48	***-0.67***
$A_{8.83}$	***-0.88***	0.03	0.15
A_{mn}	***-0.95***	-0.03	0.13
A_{sd}	***-0.86***	-0.30	-0.20
$D_{8.83}$	***-0.96***	0.03	0.04
D_{mn}	***-0.92***	-0.06	0.00
D_{sd}	***-0.88***	-0.32	-0.15
Y_{max}	-0.15	***-0.86***	-0.09
Y_{mn}	-0.29	-0.29	***0.79***
Y_{sd}	-0.08	***-0.89***	0.11

Tabel 3: Rotated components correlatie tijdsintervallen



Figuur 6: Scree plot

Ten slotte staat RC3 in voor de correlatie betreffende de snelheidsvariabelen alsook Y_{mn}. Dit geldt voor alle tijdsintervallen. Bij het uitvoeren van de varimax methode blijft het bekomen resultaat gelijkend.

Tot slot wordt één specifieke dag meegenomen als voorbeeld. De periode van 18-03-2021 toont na visueel onderzoek veel verandering in het rijgedrag. De metingen zijn zichtbaar in tabel 2. Hier wordt gekozen om het tijdsinterval van een uur te gebruiken aangezien de correlatie plots in figuur 5 aantonen dat de correlaties bij de verschillende tijdsintervallen minimaal verschillen. Bijgevolg is dit voorbeeld representatief en duidelijk te visualiseren voor het algemeen begrip van het onderzoek. Ook voor deze periode zal de varimax methode worden toegepast. In het verdere verloop van het onderzoek zal deze periode met data over de verplaatsingen worden gebruikt om het rijgedrag te clusteren en te categoriseren. De resultaten van de varimax methode zijn in tabel 4 te zien.

De ladingen voorkomend uit de varimax methode geven in deze specifieke periode een beter onderscheid wanneer het aankomt op de variantie van acceleratie en vertraging. Werkend met de rotated components zien we dat RC2 in staat voor de informatie betreffende $V_{11.5}$ en V_{sd}, maar ook voor $A_{8.83}$ en A_{sd}. RC1 beschrijft de A_{mn} alsook alle componenten betreffende de vertraging. Ten slotte staat RC3 in voor de samenhang met de draaisnelheids componenten Y_{mn} en Y_{sd}. Ook voor de dataset *verplaatsingen* zullen volgens deze componenten zowel K-means als BIRCH worden toegepast.

4.2.2. K-means clustering

Om te bepalen hoeveel clusters het beste de dataset beschrijven, werd zoals in de sectie 3 aangehaald, de elbow method gehanteerd. In figuur 7 waar respectievelijk de elbow method voor een tijdsinterval van 5 minuten, 15 minuten en een uur worden voorgesteld

	RC1	RC2	RC3
$V_{11.5}$	-0.20	***0.95***	0.07
V_{mn}	0.22	0.43	0.01
V_{sd}	-0.34	***0.86***	-0.19
$A_{8.83}$	0.51	***0.69***	-0.17
A_{mn}	***0.87***	0.16	0.35
A_{sd}	0.10	***0.80***	-0.03
$D_{8.83}$	***0.86***	0.13	-0.29
D_{mn}	***0.84***	-0.33	0.14
D_{sd}	***0.70***	-0.08	0.06
Y_{max}	0.45	0.25	-0.07
Y_{mn}	-0.33	0.02	***0.95***
Y_{sd}	-0.14	-0.08	***0.58***

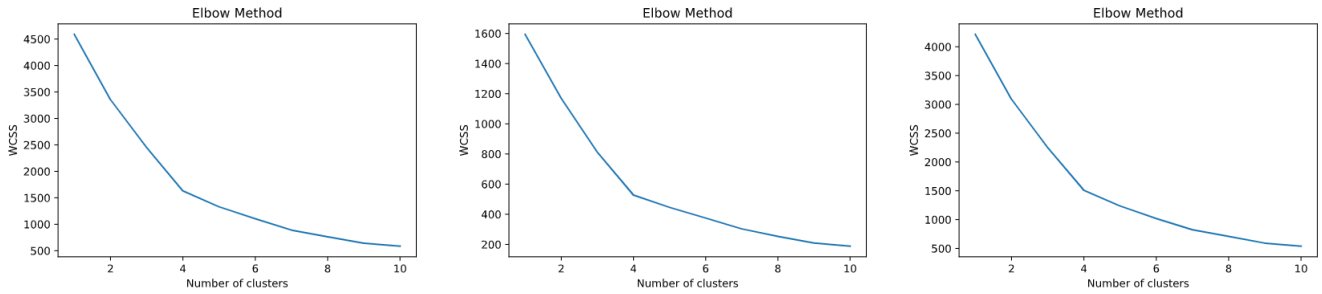
Tabel 4: Rotated components correlatie 18-03-2021

zit de *elbow* op 4 clusters. Echter is de daling na deze elbow voor alle drie de gevallen van die mate dat een vijfde cluster zeer zeker een bijdrage gaat leveren.

Er wordt dus gekozen om de clustering uit te voeren op basis van vijf clusters. Het gebruiken van de 3D-plots in figuur 9 maakt het mogelijk om visueel waar te nemen dat er zich verschillende rijgedragingen voordoen. Zo zien we, kijkend naar de RC1-as, die in dit geval instaat voor de acceleratie en vertraging, dat er twee grote groepen kunnen worden onderscheiden van elkaar. De eerste groep bevindt zich in het gebied]:0] terwijl de andere groep zich aan de andere zijde bevindt. De RC2-as, verantwoordelijk voor de draaisnelheid, geeft ons in de ruime zin 3 groepen]:-1], [-1:1] en [1:[. Ten slotte is het mogelijk om de snelheids-as (RC3) op te delen in 3 groepen]:-1], [-1:1] en [1:[. Het is deze opsplitsing die het mogelijk maken om de rijgedragingen van de heftruckchauffeurs te categoriseren. De toegekende RC-waardes geven zodoende de kans om toe te lichten waarom een patroon in een bepaalde categorie wordt geplaatst. Vooraleer we overgaan tot die stap is het waardevol om de verschillende datapunten chronologisch naast elkaar te plaatsen. Deze visualisatie brengt de mogelijkheid tot interpretatie van de resultaten met zich mee. Bij het bestuderen van deze figuren met een verschillend tijdsinterval in figuur 10 wordt direct duidelijk waarom een specifieker onderzoek naar de periode van 18-03-2021 interessant kan zijn. De clustering van de datapunten in deze periode is zeer verschillend met de rest van de totale periode. Voorts zien we ook de toegevoegde waarde verschijnen in het kiezen voor meerdere tijdsintervallen. Zo is het zichtbaar op het 5 minuten tijdsinterval dat een bepaald rijpatroon zich elke dag opnieuw voordoet voor een korte periode. Wanneer we enkel clusteren op uurbasis is dit minder duidelijk te ondervinden. Zulke bevindingen kunnen in een later onderzoek onder de loep worden genomen.

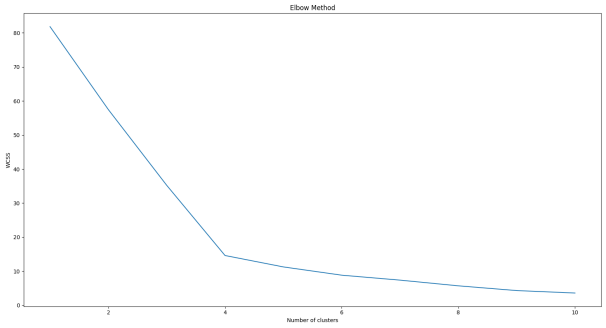
Om de periode van 18-03-2021 grondig te bestuderen werden dezelfde stappen uitgevoerd specifiek op deze periode. Uit figuur 10 (c) is het duidelijk dat het afwijkend rijpatroon in deze periode op uurbasis duidelijk zichtbaar is. Het is met deze ondervinding dat het verdere onderzoek naar de periode 18-03-2021 op uurbasis zal worden uitgevoerd. Er wordt ook voor deze specifieke periode gekozen om de clustering uit te voeren door middel van vijf clusters. Dit wordt duidelijk in figuur 8. Met de 3D-plot in figuur 11 is het mogelijk om de verschillende rijpatronen te categoriseren. Wanneer we kijken via de RC1-as, kunnen we 4 verschillende groepen onderscheiden]:-1], [-1:0], [0:1] en [1:[. De RC2-as geeft ons 3 groepen]:-1], [-1:1] en [1:[. Ten slotte is het mogelijk om de RC3-as op te delen in 3 groepen]:-1], [-1:1] en [1:[. De assen gelinkt aan de rotated components beschrijven in dit geval andere variabelen. Zo zijn zoals reeds vermeld de snelheidsvariabele en de acceleratie variabele gelinkt aan de RC2-as. Vertraging wordt afgebeeld op de RC1-as en de draaisnelheid op de RC3-as. Volgens deze opsplitsing kunnen de trajecten van de heftruckchauffeurs worden gecategoriseerd. Ook in deze specifieke case worden de RC-waardes gebruikt om toe te lichten waarom een patroon in een bepaalde categorie wordt geplaatst. RC1 gaat agressieve chauffeurs eerder aan de positieve kant van de as zetten aangezien een hoog vertragingsgemiddelde wordt gemultipliceerd met een positieve variantie. RC2 en RC3 zullen net zoals RC1 de aggressieve chauffeurs aan de rechterkant van de as positioneren.

Net zoals bij de totale periode werd hier een tweede plot gecreëerd om de evolutie van het rijgedrag doorheen de tijd te evalueren in figuur 12(a). In deze plot komt naar voren dat later op de dag, na 15u er een bepaald rijgedrag zich voordoet, waarbij de snelheden omhoog gaan en de kracht van het remmen versterkt. In de vroege uren van de ochtend is net het tegenovergestelde vast te stellen, hier zijn de snelheden wat lager en wordt er minder hard geremd. Uit de verkregen inzichten valt er af te leiden dat een agressieve chauffeur zich eerder naar het punt (2,2,2) zal begeven dan naar (-2,-2,-2). Dit wegens het feit dat de rotated components in dit geval allemaal positief zijn. Dit brengt met zich mee dat bij een hogere snelheid, een stevige acceleratie, een sterkere vertraging en een scherpe draaibeweging het punt zich rond het punt (2, 2, 2) zal situeren.



(a) 5 min (b) 15 min (c) 1 hour

Figuur 7: Elbow method K-means verschillende tijdsintervallen



Figuur 8: Elbow method K-means 18-03-2021

4.2.3. BIRCH clustering

Een tweede clustering algortime, namelijk BIRCH, wordt ook gehanteerd om de verschillende rijpatronen te categoriseren. Deze methode laat toe te kunnen kijken hoe verschillend de verkregen resultaten zijn bij beide technieken. De Silhouette methode in figuur 13 verteld ons, verschillend met de elbow method, dat 3 clusters worden aangeraden. Een eerste teken dat BIRCH minder verfijnd is in de afbakening van de verschillende rijgedragingen. Om de vergelijking te kunnen maken met de K-means clustering wordt alsnog gekozen om 5 clusters te hanteren.

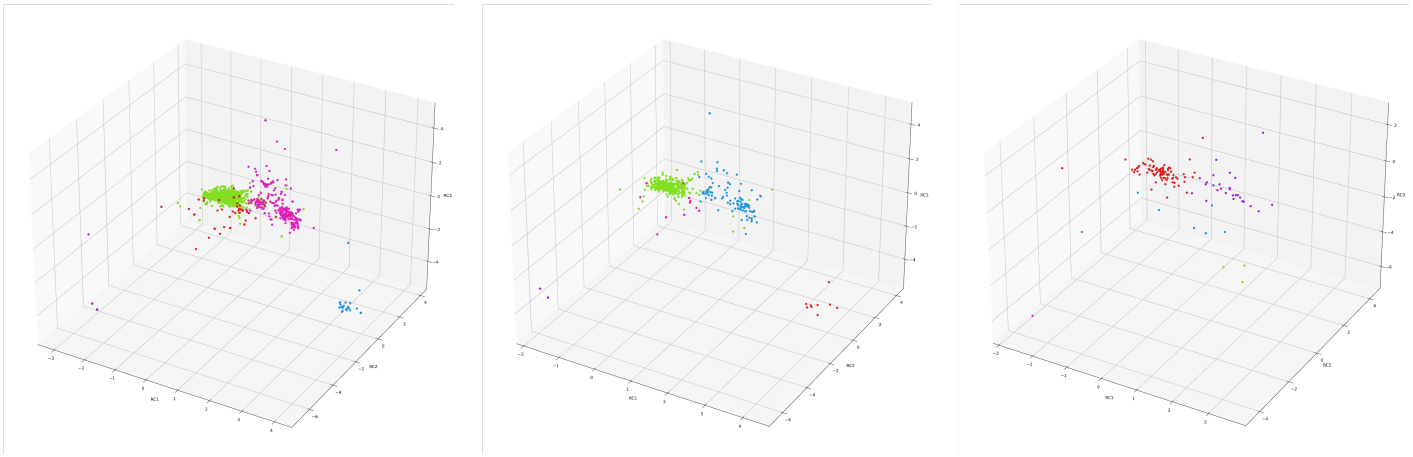
Na het uitvoeren van de BIRCH clustering zijn er enkele duidelijke constatataties. Zo is de verdeling van de datapunten onder BIRCH in figuur 14 bij het 5 minuten en 15 minuten interval veel minder duidelijk. Wanneer we ons richten op het 5 minuten tijdsinterval zien we één dominerende cluster die zowat alle datapunten bevat. In vergelijking met het 5 minuten tijdsinterval bij K-means is dit een groot verschil. Daar zien we een veel fijnere verdeling van de verschillende datapunten. Het 15 minuten tijdsinterval heeft te kampen met identiek hetzelfde fenomeen. Er lijkt in het geval van BIRCH clustering geen onderscheid gemaakt te worden langs de RC1-as. Dit maakt dat de snelheid variabele niet mee in acht wordt genomen. Een serieuze aderlating wanneer het aankomt op het categoriseren en interpreteren van de data volgens clustering. Ook bij BIRCH worden de datapunten chronologisch gevisualiseerd. De resultaten wijken hier duidelijk af ten opzichte van K-means wanneer we kijken naar de tijdsintervallen van 5 en 15 minuten zoals zichtbaar in figuur 15.

Ook kiezen we er voor om zoals bij K-means de focus te leggen op 18-03-2021. Bij het richten naar één enkele periode zien we dat BIRCH beter tot zijn recht komt. Zo is het verschil in clustering resultaten niet van diezelfde grootte als bij de totale periode. We zien één groot verschil in de clustering in figuur 17 waardoor onze criteria ook moet worden aangepast. Cluster nummer 1 uit K-means wordt hier toegekend aan het groter geheel en vormt zodoende in de BIRCH visualisatie cluster nummer 3. Op deze wijze geeft RC1 maar 3 groepen, namelijk]:-0.5], [-0.5:0.5] en [0.5:[. RC2 geeft dezelfde verdeling als bij K-means en ook RC3 wijzigt niet.

Het gebrek aan verfijning bij BIRCH geeft wel een duidelijker beeld wanneer we ons richten tot de totale periode van 18-03-2021. De clusters in figuur 12(b) tonen aan dat wanneer we kijken naar de evolutie doorheen de tijd, dat bijna alle patronen tussen 0:00 en 09:00 gelijkend zijn en dat ze zich onderscheiden van de patronen na 15:00. Zonder reeds de uiteindelijke categorisatie en interpretatie aan te vatten van de resultaten biedt het framework wel de verwachte uitkomsten. K-means clustering geeft een duidelijk onderscheid tussen verschillende rijgedragingen doorheen de tijd die af te leiden zijn uit de verkregen resultaten. BIRCH is zoals reeds vermeld minder verfijnd in de resultaten en lijkt op dit moment minder geschikt voor de uiteindelijke categorisatie van de rijgedragingen.

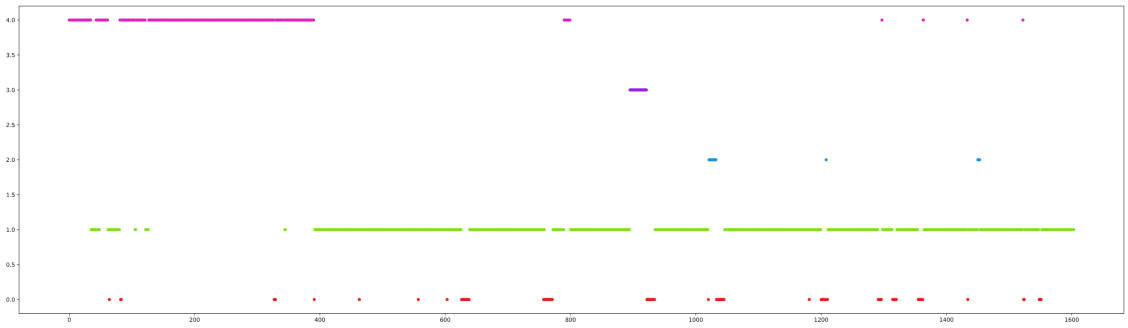
4.3. Interpretatie

De bekomen resultaten voor de periode 18-03-2021 laten ons uiteindelijk toe om de gedragingen van de heftruckchauffeurs te categoriseren. Het is in de ta-

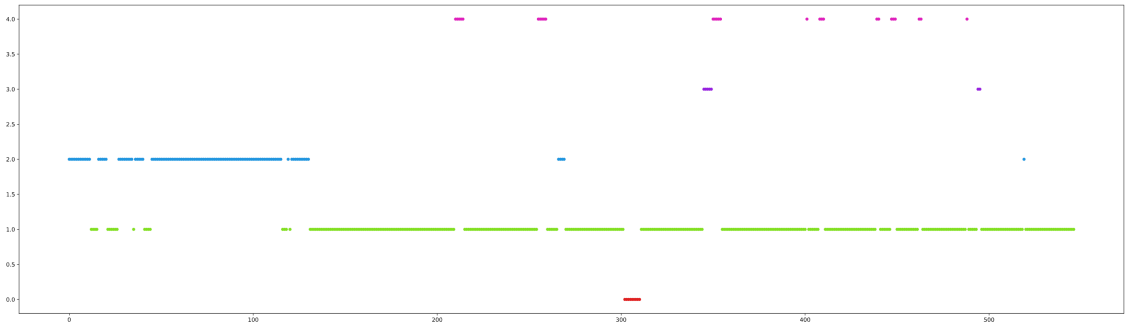


(a) 5 min (b) 15 min (c) 1 hour

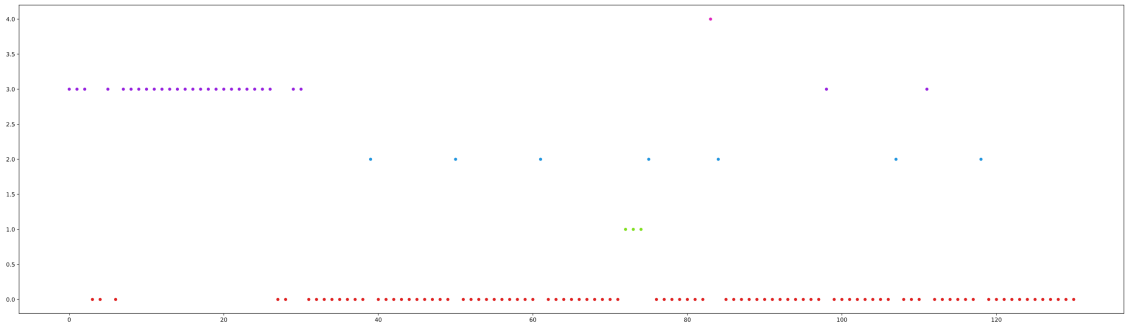
Figuur 9: K-means clustering verschillende tijdsintervallen



((a)) 5 min

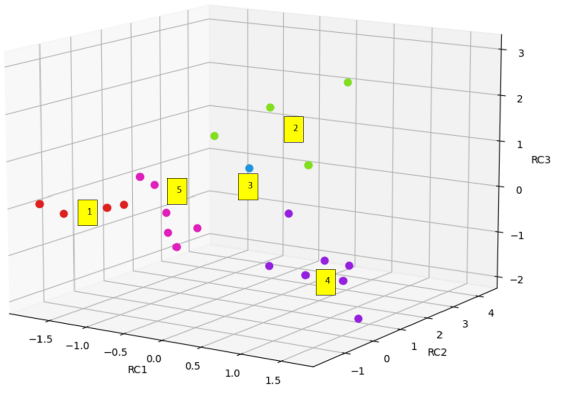


((b)) 15 min



((c)) 1 hour

Figuur 10: Chronologische scatterplots K-means



Figuur 11: K-means clustering 18-03-2021

bellen 6 en 8 dat de clusters verkregen in de figuren 11 en 17 zullen worden geïnterpreteerd. Beide tabellen beschikken over een *test*-kolom. Deze kolom geeft mee in welke cluster de controle set zich bevindt met de fictieve waardes. Lege waardes in deze kolom geven aan dat er geen controle set aanwezig is in de cluster. Deze referentie sets zullen gebruikt worden als een validatie set. In het geval dat de reële clusters sterk afwijken van de referentie sets wanneer we te maken hebben met een agressieve of niet-agressieve dataset, dan kunnen de resultaten niet als nuttig worden beschouwd.

De afbakening van de verschillende categorieën werd al kort aangehaald voor zowel de K-means als de BIRCH resultaten. In de tabellen 5 en 7 worden de componenten weergegeven gevolgd door de waardes voor de afbakening, de interpretatie en de toekenning van de clusters. Uit de bekomen resultaten van de clustering weten we dat de RC1 component in staat voor de vertragingsvariabele, RC2 de snelheid en de acceleratie beschrijft en tenslotte RC3 gelinkt kan worden aan de draaisnelheid. Alle verkregen clusters uit de resultaten worden vervolgens toegekend aan de verschillende sectoren met verschillende interpretaties. Deze toekenning is puur afhankelijk van de locatie van de cluster in de drie-dimensionale plot. Deze eerste interpretatie volgens de verschillende componenten biedt de optie om de interpretatie van de verschillende componenten samen te nemen.

Wanneer we dan kijken naar het resultaat uit de K-means clustering in tabel 6, zien we dat de 1ste cluster kan worden aanzien als *niet-agressief*. De niet-agressieve cluster nummer 1 bevat ook de controle set. Hiermee wordt dus bevestigd dat de clustering methode en de gebruikte criteria de groepen goed toekennen. De tweede cluster, met de agressieve controle set, geeft als resultaat *agressief*. Ook hier worden de criteria en de clustering bevestigd. Mocht het zijn dat de waardes bij zowel cluster 1 als cluster 2 verschillend waren van niet-agressief en agressief, dan werd het model niet geaccepteerd en waren de resultaten dus niet doeltreffend. Cluster 3 is in dit geval een speciaal geval. De vertraging kan worden beschreven als *niet-agressief* en de draaisnelheid als *neutraal*. Enkel in het geval een *agressieve* acceleratie/snelheid kan deze cluster aanzien worden als *neutraal* in totaliteit. Toch staat bij de Totaal-kolom *licht-agressief*. Dit komt omdat cluster 3 veel hogere snelheids- en acceleratiewaardes heeft in vergelijking met alle andere groepen. Er wordt dus een uitzondering gemaakt en als waarde bij acceleratie/snelheid *zeer agressief* meegegeven. Zo komen we tot het resultaat *licht-agressief*. De laatste twee clusters kennen als resultaat beiden *neutraal*. Cluster 4 spreekt voor zich aangezien *neutraal* voorkomt uit een *agressieve* vertraging, een *neutrale* acceleratie/snelheid en een *niet-agressieve* draaisnelheid. Cluster 5 kent voor de vertraging slechts een *licht niet-agressieve* waarde waarmee de twee neutrale uitkomsten minimaal overhellen. De uitkomst van een optelling zou bijgevolg -0.5 zijn wat gelijk is aan een *neutraal* rijgedrag.

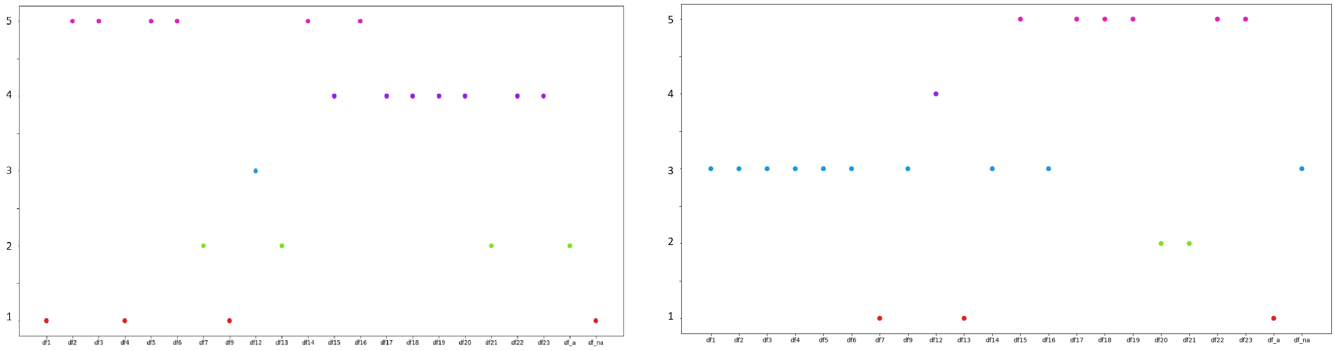
Wanneer we kijken naar de bekomen resultaten voor de BIRCH clustering in tabel 8, zien we dat de categorisatie toch wel verschilt van de K-means clustering. Cluster 1, die net zoals bij K-means de controle set bevat, heeft als totaal *licht niet-agressief*. Dit wilt dus zeggen dat er slecht één keer *niet-agressief* voorkomt in de drie rotated components. Ook cluster 5, die de controle set *agressief* bevat, geraakt niet tot *agressief*, maar belandt in *licht-agressief*. Aangezien beide extremen niet tot in de zwaarste categorie geraken, zijnde *agressief* en *niet-agressief*, hebben we te maken met een te grote veralgemening in de clustering. Dit werd ook al eerder aangehaald wanneer het duidelijk werd dat cluster 3 in de BIRCH clustering de clusters 1 en 5 van K-means omvatte. Er zijn dus minder extremen in de dataset volgens BIRCH.

BIRCH slaagt door de grotere groepering er wel in om de verschillende rijpatronen in het algemeen doorheen de dag beter van elkaar te onderscheiden, maar dit brengt met zich mee dat het verschil tussen de extremen en de normale gedragingen verwatert.

Na een complete doorloop van het vooropgestelde framework zijn de resultaten veelzeggend. Het is mogelijk om op basis van de vooropgestelde stappen verschillende rijgedragingen doorheen de tijd te onderscheiden van elkaar. K-means slaagt beter in deze opzet ten opzichte van BIRCH.

5. Conclusie

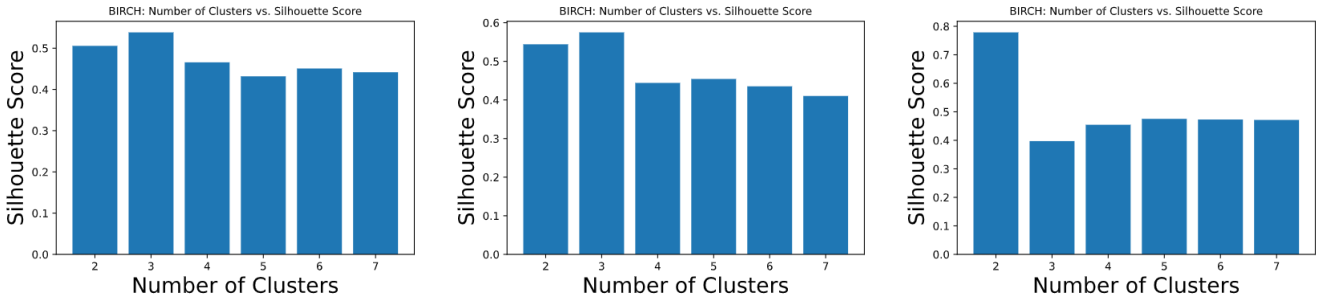
Het onderzoek heeft aangetoond dat aan de hand van het vooropgesteld framework het mogelijk is om



(a) K-means tijdsevolutie

(b) BIRCH tijdsevolutie

Figuur 12: Tijdsevolutie

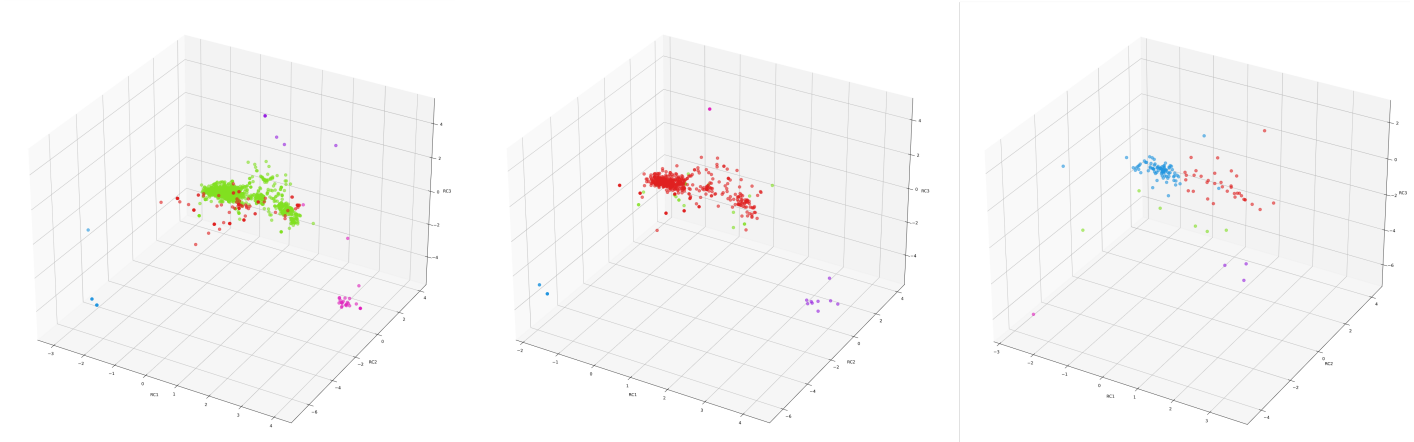


(a) 5 min

(b) 15 min

(c) 1 hour

Figuur 13: BIRCH Silhoutte verschillende tijdsintervallen



(a) 5 min

(b) 15 min

(c) 1 hour

Figuur 14: BIRCH clustering verschillende tijdsintervallen

Tabel 5: Rotated components ranking K-means

Component	Waardes	Interpretatie	Clusters
RC1	$<$ -1 -1 $<$ 0 0 $<$ 1 $>$ 1	niet-agressief licht niet-agressief licht agressief agressief	1,3 5 / 2,4
RC2	$<$ -1 -1 $<$ 1 $>$ 1	niet-agressief neutraal agressief	1 2,4,5 3
RC3	$<$ -1 -1 $<$ 1 $>$ 1	niet-agressief neutraal agressief	4 1,3,5 2

Tabel 6: Categorisatie K-means

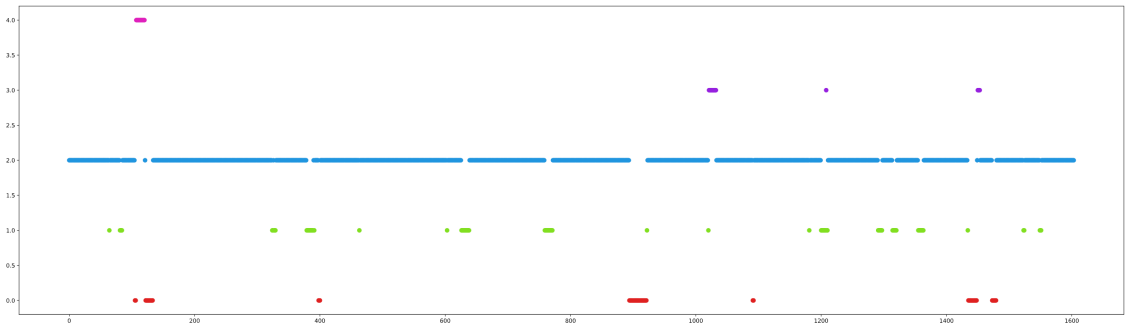
Cluster	Test	Vertraging (RC1)	Acceleratie/speed (RC2)	Draaisnelheid (RC3)	Totaal
1	df_na	niet-agressief	niet-agressief	neutraal	niet-agressief
2	df_a	agressief	neutraal	agressief	agressief
3		niet-agressief	*ZEER* agressief	neutraal	licht-agressief
4		agressief	neutraal	niet-agressief	neutraal
5		licht niet-agressief	neutraal	neutraal	neutraal

Tabel 7: Rotated components ranking BIRCH

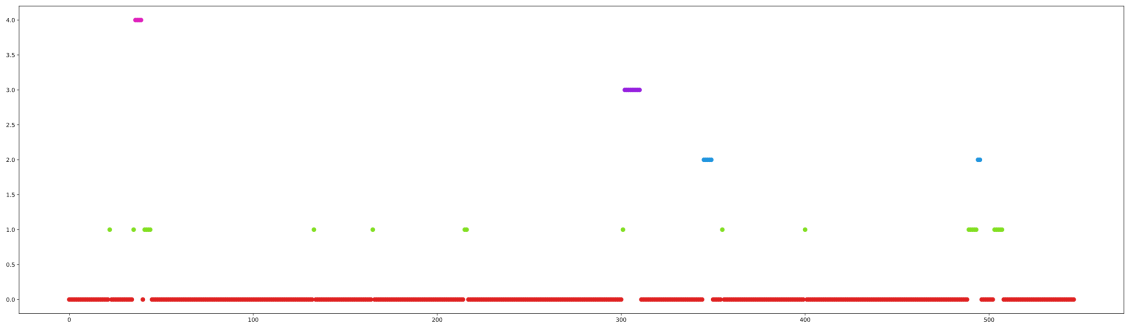
Component	Waardes	Interpretatie	Clusters
RC1	$<$ -0.5 -0.5 $<$ 0.5 $>$ 0.5	niet-agressief licht niet-agressief agressief	1,4 5 2,3
RC2	$<$ -1 -1 $<$ 1 $>$ 1	niet-agressief neutraal agressief	/ 1,2,3,5 4
RC3	$<$ -1 -1 $<$ 1 $>$ 1	niet-agressief neutraal agressief	2 1,3,4 5

Tabel 8: Categorisatie BIRCH

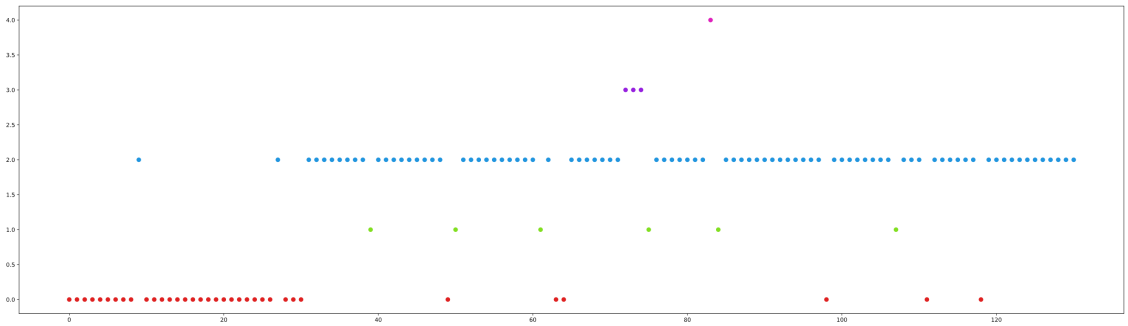
Cluster	Test	Vertraging (RC1)	Acceleratie/speed (RC2)	Draaisnelheid (RC3)	Totaal
1	df_na	niet-agressief	neutraal	neutraal	licht niet-agressief
2		agressief	neutraal	niet-agressief	neutraal
3		agressief	neutraal	neutraal	licht-agressief
4		niet-agressief	*ZEER* agressief	neutraal	licht-agressief
5	df_a	neutraal	neutraal	agressief	licht-agressief



((a)) 5 min

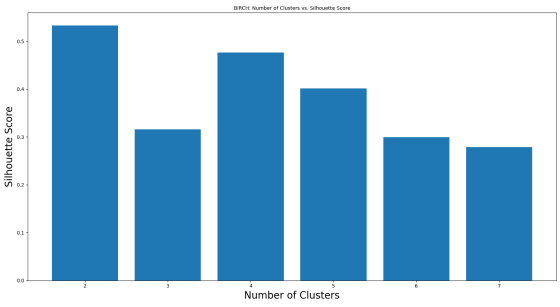


((b)) 15 min

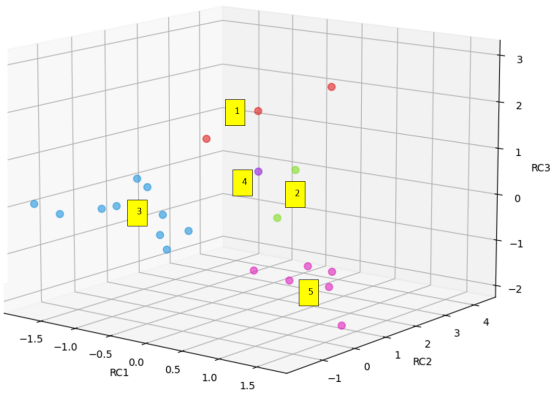


((c)) 1 hour

Figuur 15: Chronologische scatterplots BIRCH



Figuur 16: BIRCH Silhouette 18-03-2021



Figuur 17: BIRCH clustering 18-03-2021

rijgedragingen van heftruckchauffeurs in een indoor locatie te visualiseren, te categoriseren en te interpreteren. De geleverde real-life dataset, waarbij elke heftruck nauwgezet wordt gevolgd, biedt dus de mogelijkheid om op basis van de coördinaten, richtingsvariabele en de tijd verschillende categorieën te definiëren en de resultaten te interpreteren doorheen de tijd. Het is met deze verkregen resultaten dat het framework kan worden gevalideerd. De COVID-19 periode waarin het onderzoek werd gevoerd heeft er helaas voor gezorgd dat een bezoek op locatie om zodoende de verkregen resultaten met de heftruckchauffeurs te bespreken niet mogelijk was. Het direct aftoetsen van de resultaten had het onderzoek nog een laatste maal onder het vergrootglas kunnen leggen, echter is dit dus niet mogelijk gebleken.

De dataset betreffende laden en lossen is moeilijker te categoriseren wanneer er gebruik wordt gemaakt van acceleratie, vertraging en draaisnelheid. Het onderscheid tussen de gedragingen bij het laden en lossen wordt niet vertaald door de variabelen die werden opgesteld in het framework.

Zowel BIRCH als K-means leverde interessante resultaten op. K-means gaf ons een specifiekere classificatie, terwijl BIRCH erin slaagde om het verschil in gedragingen doorheen de tijd zichtbaar te maken wanneer we kijken naar de periode van 18-03-2021. Voor een clustering van de totale periode was BIRCH minder geschikt ten op zichte van K-means. De robuustheid bij het clusteren met BIRCH was te hoog waarmee verschillende patronen van rijgedrag niet konden worden opgemerkt. K-means daarentegen gaf een duidelijk beeld van de verschillende gedragingen. Verder viel er te constateren voor 18-03-2021 dat het rijgedrag in de periode tussen 15:00 en 23:00 als gelijkend kon worden beschouwd en dus verschillend was van het rijgedrag tussen 01:00 en 09:00.

Een opmerking die gemaakt dient te worden bij het framework en bijgevolg de volledige doorloop van het proces is, dat de interpretatie van belang is. Wanneer een categorisatie wordt volbracht op een periode waarbij de verschillen in rijgedraging minimaal zijn en de correlatie van de verschillende factoren onderling zeer hoog is, zal het niet direct mogelijk zijn om verschillende patronen te onderscheiden van elkaar. Dit zal direct duidelijk worden in de ladingen van de varimax methode. Vele variabelen zullen worden toegelicht door één enkele principal/rotated component. Dit is ook logisch aangezien de loadings een direct gevolg zijn van de onderlinge correlatie. Een optie die kan worden overwogen is om deze periode van rijgedragingen, die niet significant van elkaar verschillen, in totaliteit te vergelijken met een andere periode om dus toch tot een inzicht te geraken. Een voorbeeld hiervan is de data vanaf 19-03-2021. Zowel de acceleratie als de vertraging worden toegelicht door RC1, wat maakt dat we hierop niet kunnen differentiëren. Een optie is dan om de dataset van 19-03-2021 in totaliteit te nemen en te vergelijken met de patronen op 18-03-2021.

6. Aanvullend werk

De verkregen dataset, waarop het onderzoek werd uitgevoerd, bevat zeer veel datapunten, echter is het beperkt in de aangeboden informatie. Zo zijn er veel verschillende factoren die het rijgedrag van een heftruckchauffeur nog kunnen beïnvloeden waar dit framework geen belang aan heeft kunnen hechten. Enkele van de factoren die op het eerste zicht een significante invloed zouden kunnen hebben, zijn de lading waarmee de heftruck zich beweegt, de drukte, de voetgangerszones, de mentale staat van de bestuurder en de individuele karakteristieken van de bestuurder. Implementatie van dit soort factoren zouden het framework zeer zeker kunnen bevorderen en verbeteren. Zo zou een een logboek kunnen worden gebruikt om de ladingen te tracken die verplaatst zijn of het afnemen van een korte vragenlijst om de staat van de chauffeur te bepalen.

Verder verplaatst in een warehouse een heftruck zich niet enkel van de laadpunten naar de lospunten en andersom, maar er zijn ook de momenten van het laden en het lossen zelf. Enkel aan de hand van de verkregen data wordt dit een moeilijke, inzichtloze oefening. Het is door middel van extra variabelen zoals de snelheid waarmee de vork wordt bewogen, de mogelijkheid tot kantelen bij een bepaalde beweging ... dat het framework kan worden bevorderd.

Om het framework functioneel te maken dient men de informatie op constante basis te verwerken. Het categoriseren van het rijgedrag van een heftruckhauffeur zou mits de juiste omvorming van de data direct kunnen gebeuren. Voor de praktische verwezenlijking moet een verder onderzoek uitwijzen hoe dit het meest efficiënt gebeurt.

Referenties

[1] Liu, H.; Darabi, H.; Banerjee, P.; Liu, J. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **2007**, *37*, 1067–1080.

[2] Constantinescu, Z.; Marinoiu, C.; Vladoiu, M. Driving style analysis using data mining techniques. *International Journal of Computers Communications & Control* **2010**, *5*, 654–663.

[3] Ding, C.; He, X. In *Proceedings of the twenty-first international conference on Machine learning*, **2004**, 29.

[4] Sun, E.; Ma, R. In *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, **2017**, 86–90.

[5] Barral, V.; Suárez-Casal, P.; Escudero, C. J.; Garcıa-Naya, J. A. Multi-sensor accurate forklift location and tracking simulation in industrial indoor environments. *Electronics* **2019**, *8*, 1152.

[6] Halawa, F.; Dauod, H.; Lee, I. G.; Li, Y.; Yoon, S. W.; Chung, S. H. Introduction of a real time location system to enhance the warehouse safety and operational efficiency. *International Journal of Production Economics* **2020**, *224*, 107541.

[7] Wang, J.; Lu, M.; Li, K. Characterization of longitudinal driving behavior by measurable parameters. *Transportation Research Record* **2010**, *2185*, 15–23.

[8] Zhu, X.; Hu, X.; Chiu, Y.-C. Design of driving behavior pattern measurements using smartphone global positioning system data. *International Journal of Transportation Science and Technology* **2013**, *2*, 269–288.

[9] Hofstra, N.; Petkova, B.; Dullaert, W.; Reniers, G.; De Leeuw, S. Assessing and facilitating warehouse safety. *Safety Science* **2018**, *105*, 134–148.

[10] Al-Shaebi, A.; Khader, N.; Daoud, H.; Weiss, J.; Yoon, S. W. The effect of forklift driver behavior on energy consumption and productivity. *Procedia Manufacturing* **2017**, *11*, 778–786.

[11] Boehning, M. In *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, **2014**, 245–249.

[12] Tasca, L., A review of the literature on aggressive driving research; Ontario Advisory Group on Safe Driving Secretariat, Road User Safety Branch . . . : **2000**.

[13] Railsback, B. T.; Ziernicki, R. M. In *ASME International Mechanical Engineering Congress and Exposition*, **2010**; *44489*, 421–424.

[14] Browne, M. W. An overview of analytic rotation in exploratory factor analysis. *Multivariate behavioral research* **2001**, *36*, 111–150.

[15] Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.

[16] Abdi, H. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA* **2003**, 792–795.

[17] Rokach, L.; Maimon, O. In *Data mining and knowledge discovery handbook*; Springer: **2005**, 321–352.

[18] Xu, D.; Tian, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science* **2015**, *2*, 165–193.

[19] Likas, A.; Vlassis, N.; Verbeek, J. J. The global k-means clustering algorithm. *Pattern recognition* **2003**, *36*, 451–461.

[20] Marutho, D.; Handaka, S. H.; Wijaya, E., et al. In *2018 International Seminar on Application for Technology of Information and Communication*, **2018**, 533–538.

[21] Bholowalia, P.; Kumar, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications* **2014**, *105*.

[22] Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery* **1997**, *1*, 141–182.

[23] Abbas, O. A. Comparisons between data clustering algorithms. *International Arab Journal of Information Technology (IAJIT)* **2008**, *5*.

[24] Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record* **1996**, *25*, 103–114.

[25] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **1987**, *20*, 53–65.

[26] Baxter, J. S.; Manstead, A. S.; Stradling, S. G.; Campbell, K. A.; Reason, J. T.; Parker, D. Social facilitation and driver behaviour. *British Journal of Psychology* **1990**, *81*, 351–360.

[27] French, D. J.; West, R. J.; Elander, J.; WILDING, J. M. Decision-making style, driving style, and self-reported involvement in road traffic accidents. *Ergonomics* **1993**, *36*, 627–644.

[28] Doherty, S. T.; Andrey, J. C.; MacGregor, C. The situational risks of young drivers: The influence of passengers, time of day and day of week on accident rates. *Accident Analysis & Prevention* **1998**, *30*, 45–52.

[29] Huestegge, L.; Skottke, E.-M.; Anders, S.; Müsseler, J.; Debus, G. The development of

hazard perception: Dissociation of visual orientation and hazard processing. *Transportation research part F: traffic psychology and behaviour* **2010**, *13*, 1–8.

[30] Musselwhite, C. Attitudes towards vehicle driving behaviour: Categorising and contextualising risk. *Accident Analysis & Prevention* **2006**, *38*, 324–334.