Een instrument voor het voorspellen van de bezettingsgraad van parkeergarages

BSc Project Plan

Liza Snellen
STUDENT TECHNISCHE BEDRIJSKUNDE
Universiteit Twente

UNIVERSITEIT TWENTE

Dr.Ir. M.R.K. Mes
Industrial Engineering and Business Information Systems

Dr. M.C. van der Heijden
Industrial Engineering and Business Information Systems

GOUDAPPEL GROEP

H. Palm
DAT.Mobility, Consulting

Ing. A.H. Oldenburger
Goudappel Coffeng, Mobiliteitsdiensten

Managementsamenvatting

Goudappel Groep bestaat uit een aantal samenwerkende bedrijven die mobiliteitsvraagstukken gezamenlijk op de beste manier proberen op te lossen. De bekendste labels zijn Goudappel Coffeng en DAT.Mobility. In opdracht van DAT.Mobility is er onderzoek gedaan naar de mogelijkheden om de bezettingsgraad van parkeergarages te kunnen voorspellen aan de hand van historische gegevens.

Probleemanalyse

Het meest gebruikte en bekendste systeem in Nederland is het parkeerverwijssysteem. Het doel van dit systeem is om de bereikbaarheid van steden te bevorderen en de bezetting van de parkeerfaciliteiten te vergroten. Echter blijkt uit onderzoek dat maar 56% van de automobilisten het systeem opmerkt en het door 11,4% daadwerkelijk wordt gebruikt. Er is behoefte aan een systeem dat vroegtijdig kan voorspellen hoe druk het is in parkeergarages. Hierdoor kunnen automobilisten de keuze van de parkeerlocatie van tevoren nemen en zal dit leiden tot minder zoekverkeer en minder wachtrijen in steden rondom parkeergarages. Daarnaast is het ook goed voor gemeenten om deze informatie tot hun beschikking te hebben. Zij kunnen analyseren of er bijvoorbeeld meer parkeerplaatsen in een stad nodig zijn.

Theoretisch kader

Aan de hand van een literatuuronderzoek is gekeken naar de variabelen die invloed hebben op de bezettingsgraad en naar voorspellingsmethoden. Er zijn vijf onafhankelijke variabelen die als belangrijkste in vergelijkbare onderzoeken naar voren komen: *tijdstip van de dag, dag van de week, vakantie, weer* en *speciale gebeurtenissen.* Daarnaast spelen ook de variabelen *capaciteit* en *locatie* een rol, omdat er verschillende parkeergarages worden meegenomen in het onderzoek.

De methoden die in de literatuur naar voren zijn gekomen zijn: beslissingsboom/regressieboom, bewegend gemiddelde, ARIMA, kunstmatig neuraal netwerk, lineaire regressie/veelvoudige regressie. De regressieboom komt in meerdere onderzoeken als beste naar voren en is daarom het meest geschikt om voor dit onderzoek te gebruiken. Neurale netwerken hebben hun eigen kanttekening en worden in meerdere onderzoeken afgeschreven. ARIMA is een model dat ingewikkeld is en niet goed kan omgaan met ruis. Daarnaast is lineaire regressie een eenvoudig model, maar komt het niet goed naar voren uit de vergelijkbare onderzoeken.

Data- analyse

De vijf variabelen die al aanwezig zijn in de historische gegevens voor 128 parkeergarages in Nederland zijn: *de datum, het tijdstip, de locatie, het aantal vrije plaatsen* en *de capaciteit.* Een duidelijk onderscheid dat wordt gemaakt tussen de parkeergarages is de locatie; in het centrum of juist bij een bedrijventerrein. Voor het onderzoek worden 12 pakeergarages gebruikt, die allemaal in het centrum van verschillende steden gevestigd zijn. Er zijn duidelijke patronen te zien bij verschillende variabelen zoals: *type dag, tijd* en *vakantie*. Hierdoor is vast te stellen dat de bezettingsgraad van parkeergarages zeker van deze onafhankelijke variabelen afhangt. Ook zijn er duidelijke cycli in de gegevens te zien, die de weken weergeven, waarin de weekenden drukker zijn dan werkdagen. Sommige uitschieters zijn te verklaren door verkeerd opgeslagen historische gegevens en andere zijn te verklaren door feestdagen, evenement, koopzondagen en vakantie.

Er zijn enkele beperkingen voor het onderzoek, zoals de relatief kleine dataset. De voorspellingen zouden nauwkeuriger zijn als er meer historische gegevens beschikbaar waren. Verder is het niet te achterhalen hoe lang één auto geparkeerd staat in de parkeergarage, waardoor er weinig te zeggen valt over de drukte naar en van een parkeergarage. Tot slot heeft het systeem zelf ook zijn beperkingen, omdat gegevens soms niet correct worden opgeslagen.

Implementatie

Er is een onderscheid tussen een train dataset, waarmee het model gemaakt wordt en een test dataset om het model te testen. Er zijn drie verschillende combinaties van train en test datasets gemaakt, om uit te sluiten dat de uitkomsten berusten op toeval. Om de regressieboom te kunnen vergelijken, worden er ook voorspellingen gegenereerd aan de hand van een eenvoudige methode. Deze methode gebruikt alleen de variabele *tijdstip* en *dag van de week* en berekend hiervan de gemiddelden. De regressieboom wordt berekend met de regressieformule die focust op een zo klein mogelijke error. De paden in de boom zijn de uitkomst en leiden naar de voorspellingen. De boom begint met alle waarnemingen in één klasse met daarvan de gemiddelde bezettingsgraad. Per splitsing worden de waarnemingen verdeeld over de volgende twee klasse. De boom eindigt met veel verschillende klassen met ieder een aantal waarnemingen en de gemiddelde bezettingsgraad hiervan.

Resultaten

De validatie van de modellen wordt gecontroleerd door het analyseren van de regressie tussen de voorspellingen en de werkelijke waarden. De samenhang tussen deze twee waarden kan worden weergegeven met de correlatie coëfficiënt R. Een waarde van $R > 0.7$ geeft aan dat er een sterk verband is tussen de variabelen. Daarnaast is voor elke dataset ook de determinatiecoëfficiënt R^2 berekend, die groter moet zijn dan 0.6. Ook de residuen zijn bekeken, die het verschil tussen de werkelijke en de voorspelde bezettingsgraad weergeven. Het eenvoudige model voldoet aan geen enkel criteria en is daarom niet geschikt om voorspellingen mee te genereren.

De regressieboom daarentegen heeft een gemiddelde R^2-waarden van ruim 0.80 en een gemiddelde R-waarden van 0.91. Ook de regressieformule is bekeken per train dataset en voldoet aan het criterium. De gegevens van de verschillende train datasets lijken veel op elkaar. De residuen geven allemaal een random patroon en lijken normaal verdeeld te zijn. Er zijn wat extreme afwijkingen, maar toch kan er met enige zekerheid gezegd worden dat de residuen normaal verdeeld zijn en dat de regressieboom dus goed presteert op de train dataset.

Het eenvoudige model scoort duidelijk slechter dan de regressieboom, daarom zijn de verdere analyses met de test datasets alleen uitgevoerd voor de regressieboom. De waarden van R^2 blijven gemiddeld 0.80 en met een R-waarde van gemiddeld 0.90 is ook de samenhang van deze datasets zeer sterk. Ook de regressievergelijkingen voldoen aan het criterium. Het analyseren van de residuen geeft iets meer verschillen weer tussen de test datasets. Toch kan ook nu worden gezegd dat de residuen normaal verdeeld zijn en dat de regressieboom accurate voorspellingen genereert.

Conclusie en aanbevelingen

Uiteindelijk kan uit de resultaten geconcludeerd worden dat de prestatie van de regressieboom zeer goed is. Er kan nog met de complexiteit gespeeld worden om de betrouwbaarheid van het model te verbeteren. Hierdoor zal de nauwkeurigheid van de voorspelling flink omhoog gaan, maar de kans op overfitten is groter. Voor nu was het streven een betrouwbaarheid van 80%. Met dit instrument kan de bezettingsgraad van parkeergarages in Nederland met 81% nauwkeurigheid worden voorspeld.

Verder liggen er mogelijkheden om parkeergarages in één stad met elkaar te vergelijken. Hier komen andere variabelen zoals de prijs en de loopafstand naar een centrum om de hoek kijken. Daarnaast is het ook mogelijk om real-time gegevens aan het model toe te voegen, waarmee het model nauwkeuriger de bezettingsgraad over een kleine tijdshorizon kan voorspellen. Tot slot is het belangrijkste vervolgonderzoek wat plaats zal vinden het onderzoeken van de in- en uitrijtijden van parkeergarages. Als de gemiddelde verblijftijd van een auto voorspeld kan worden en dit gecombineerd wordt met de voorspelling van de hoeveelheid auto's die aanwezig zijn, kunnen er analyses gedaan worden voor de intensiteit van het wegennetwerk rondom parkeergarages.

Voorwoord

Voor u ligt mijn Bachelor opdracht “Een instrument voor het voorspellen van de bezettingsgraad van parkeergarages”. Dit verslag is geschreven in het kader van het afronden van mijn Bachelor Technische Bedrijfskunde aan de Universiteit Twente. Het onderzoek richt zich op het ontwikkelen van een instrument waarmee de bezettingsgraad van parkeergarages in Nederland voorspeld kan worden. In de periode van september 2018 tot en met januari 2019 heb ik bij DAT.Mobility in Deventer aan dit onderzoek gewerkt.

Graag wil ik in het voorwoord van de gelegenheid gebruik maken om mijn begeleiders en collega’s van DAT.Mobility te bedanken. Met veel plezier heb ik aan mijn opdracht gewerkt, mede door de prettige en open sfeer binnen het bedrijf. Als stagiair werd ik overal voor uitgenodigd en dit gaf mij een hartverwarmend gevoel en leuke herinneringen. In het bijzonder wil ik Henri Palm bedanken voor de tijd en energie die hij in mij en mijn onderzoek heeft gestoken. Ik kon elk moment bij hem terecht voor advies, kennis of een gezellig onderonsje. Daarnaast wil ik ook André Oldenburger bedanken voor de begeleiding en zijn interesse in mijn onderzoek. Als laatste wil ik alle andere werknemers van DAT.Mobility en Goudappel Coffeng bedanken voor alle hulp en leuke momenten.

Verder wil ik ook mijn begeleider van de Universiteit Twente, Martijn Mes, bedanken voor de begeleiding. Hij heeft mij van nuttige feedback voorzien en meegedacht over de juiste richting van mijn onderzoek. Het was fijn dat hij zoveel tijd had om samen met mij mijn werk door te nemen.

Tot slot gaat mijn dank uit naar mijn familie en vrienden voor het bieden van steun en het geven van advies. De interesse van velen deed mij goed.

Ik wens u veel leesplezier!

Liza Snellen

Enschede, februari 2019

Inhoud

Hoofdstuk 1 – Introductie

Dit onderzoek begint met een introductie van het bedrijf met daarbij hun strategie en focuspunten (§1.1). Vervolgens wordt de aanleiding van het onderzoek besproken (§1.2) en wordt aan de hand van een probleemkluwen het handelingsprobleem vastgesteld (§1.3). Daarnaast wordt ook het doel van het onderzoek uitgestippeld (§1.4). Uiteindelijk volgt er een beschrijving over de aanpak van het onderzoek en zal de verdere opbouw van het verslag worden toegelicht (§1.5).

1.1 Bedrijfsinformatie Goudappel Groep

Goudappel Groep bestaat onder andere uit Goudappel Coffeng en DAT.Mobility. Goudappel Coffeng verbindt expertises, belangen en partijen vanuit kennisleiderschap. Samen met DAT.Mobility zijn zij gespecialiseerd in (geografische) optimalisatie-vraagstukken, Big Data, het ontwikkelen en toepassen van verkeersmodellen, ontwikkelen van (reistijd)-voorspellers, algoritmiek en het realiseren van grootschalige software platformen. Goudappel Groep maakt zich sterk voor een zo efficiënt en schoon mogelijk mobiliteitssysteem, waarin mobiliteit staat voor alle bewegingen van mensen en goederen. De focus ligt op mobiliteit, maar wel in de volle breedte en diversiteit die recht doet aan dit specialisme. Dit is een uniek kenmerk van Goudappel Groep en dit heeft geleid tot een bedrijf met focus, netwerk, diversiteit, cultuur en omvang om klanten één loket te bieden voor hun complexe mobiliteitsvraagstukken.

De strategie binnen Goudappel Groep bestaat uit drie verschillende gebieden: advies, IT-oplossingen en mobiliteitsdiensten. Advies wordt gegeven van strategie tot en met uitvoering. Goudappel Groep bestrijkt de hele beleidskolom van doel- en strategiebepaling tot en met ontwerp, uitvoering, implementatie en evaluatie. Hierin wordt vakkennis gecombineerd met kennis van de regio en een duidelijke procesaanpak. Daarnaast worden er operationele IT-toepassingen ontwikkelt op straat of web, zoals intelligente verkeersmanagement-oplossingen, multimodale reisinformatie of geografische informatiesystemen. Ook zijn er gereedschappen voor het modelleren en analyseren van mobiliteitsvraagstukken aanwezig. Tot slot is het derde speerpunt het wegnemen van de zorgen van de klanten tijdens beheer en exploitatie. De mobiliteitsdiensten gaan over het operationeel houden van het verkeersmodel tot de volledige uitvoering hiervan.

1.2 Aanleiding onderzoek

Al lange tijd is Goudappel Groep bezig met de ontwikkeling op het gebied van voorspellingen op korte en (middel)lange termijn. Deels door gebruik van aangeboden technologieën en deels door inhoudelijke eigen ontwikkelingen. De propositie van Goudappel Groep hangt samen met de core competences waar producten of diensten op ontwikkeld worden: verkeersdata, (real-time) fuseren, voorspellen korte en lange termijn, multimodale reisinformatie, data analytics en visualisatie. Hierdoor positioneren zij zich op het domein van het verrijken van ruwe data tot beslisinformatie voor stakeholders.

Goudappel Groep wil via algoritmes voor datafusie, verificatie/validatie, verrijking en correlaties in staat zijn enerzijds de actuele en dynamische status weer te geven, maar ook voorspellingen te genereren over kortere operationele en/of langere tactische termijn. Het gaat hierbij om een zo compleet mogelijk beeld van de verkeerskundige situatie voor zowel auto/vracht/OV/fiets op de diverse netwerken, zoals afwikkeling, parkeerbezettingsgraden, brugopeningen en maatregelinzet. Deze kennis wordt aangevuld met korte en lange termijn voorspellingen over hoe die situatie zicht gaat ontwikkelen en biedt de mogelijkheid om Key Performance Indicators (KPI's) af te leiden, zodat de stakeholders individueel de performance op hun eigen KPI's kunnen monitoren, de beste beslissingen kunnen nemen of gecoördineerd maatregelen in kunnen zetten.

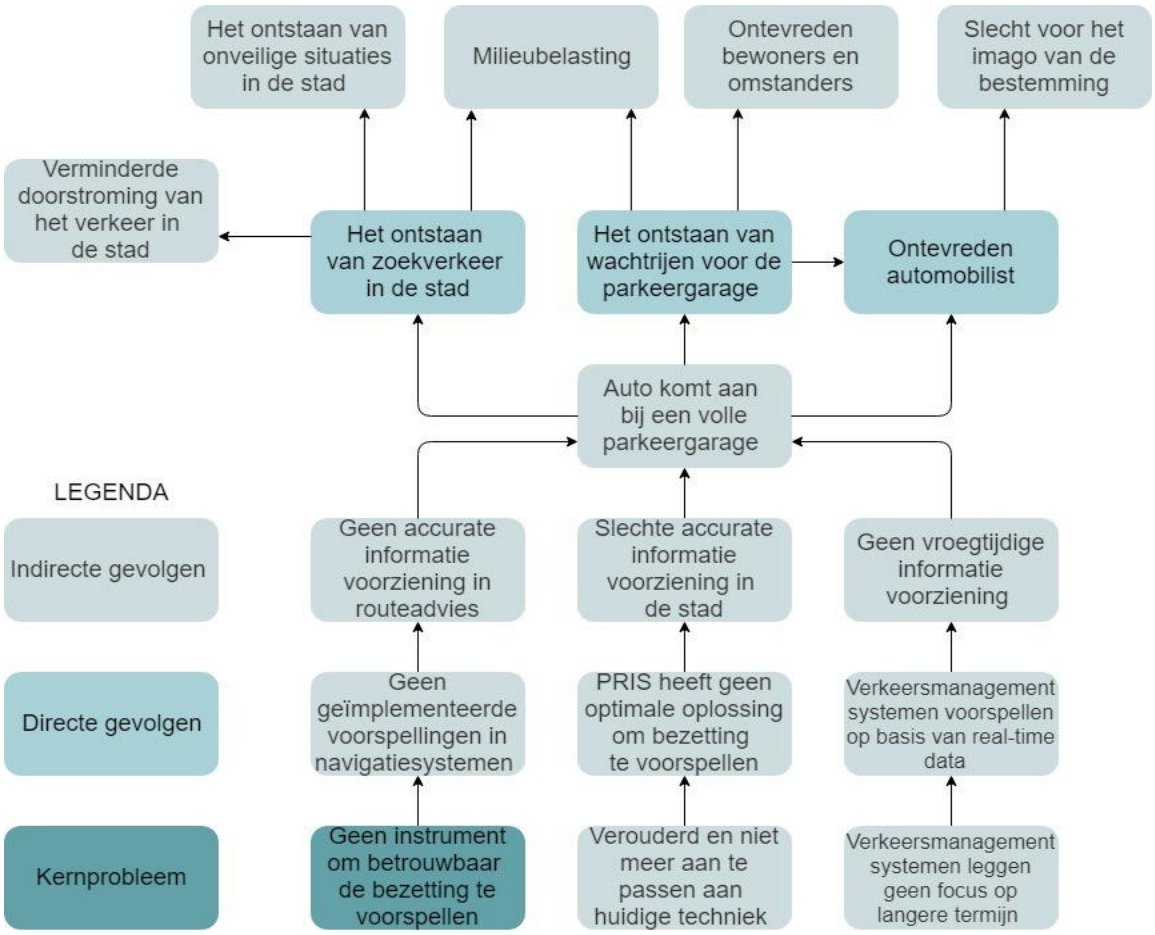
Goudappel Groep werkt aan een softwareapplicatie genaamd OmniTRANS Next. Een compleet data en model gedreven platform waarin demografische- en verkeersgegevens samenkomen en worden verwerkt tot bruikbare prognoses en analyses voor relevante tijdshorizonnen, waarbij de prognoses gebruik maken van state-of-the-art model- en datatechnieken. Dit onderzoek richt zich op het onderdeel parkeren door het leggen van een basis voor een model wat door Goudappel Groep geïntegreerd kan worden. Het voorspellen van de bezettingsgraad van parkeergarages staat hierin centraal.

1.3 Probleemidentificatie

Er zijn verschillende problemen die opgelost kunnen worden met elk andere gevolgen. Voordat er een handelingsprobleem vast gesteld kan worden, moeten de verschillende problemen met hun onderlinge relaties in kaart worden gebracht. Deze relaties kunnen worden weergegeven in een probleemkluwen om structuur aan te brengen in de probleemcontext en het handelingsprobleem te identificeren (Heerkens & Winden, 2012).

1.3.1 Probleemkluwen

De vraag vanuit DAT.Mobility is om meer inzicht te krijgen in de bezetting van parkeergarages en de mogelijkheden om deze te kunnen voorspellen op korte en lange termijn. Deze vraag komt voort uit verschillende perspectieven en stakeholders die hier baat bij hebben. Met deze informatie worden er indirecte en directe gevolgen aangepakt. Om de relaties tussen problemen in kaart te brengen is er een probleemkluwen gemaakt, zie Figuur 1. Er zijn er meerdere oorzaken die leiden naar de uiteindelijke directe gevolgen: het ontstaan van zoekverkeer in de stad, het ontstaan van wachtrijen voor de parkeergarage en ontevreden automobilisten.



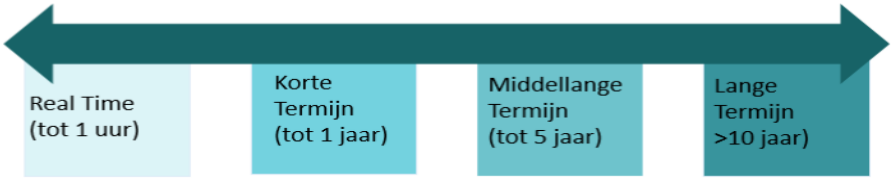
FIGUUR 1 - PROBLEEMKLUWEN

1.3.2 Handelingsprobleem

In de probleemkluwen staan veel problemen genoemd, maar deze kunnen niet allemaal tegelijk aangepakt worden. Door de onderlinge relaties zal het oplossen van elk kernprobleem indirect gevolgen hebben voor meerdere problemen. Momenteel is er geen instrument binnen Goudappel Groep beschikbaar om betrouwbaar de bezettingsgraad van parkeergarages te voorspellen. Dit gebrek is cruciaal voor verwijssystemen, reisplanners, navigatiesystemen, omdat er geen vroegtijdige informatie verschaft kan worden met betrekking tot het parkeren op de bestemming. Ook kan er op dit gebied geen advies worden gegeven aan gemeenten, omdat er weinig inzicht is in deze gegevens. Het model dat hiervoor nodig is moet aan een bepaalde betrouwbaarheid voldoen. Echter, omdat er nog geen model aanwezig is, kan er niet makkelijk worden gezegd wat de mate moet zijn van deze betrouwbaarheid. Dit is iets wat gedurende het onderzoek duidelijk moet worden.

Door een gebrek aan informatievoorziening voor parkeerroutes komen auto's aan bij volle parkeergarages. Hierdoor zijn de bestuurders niet alleen ontevreden, maar dit resulteert ook in zoekverkeer en wachtrijen. Deze drie hoofdproblemen zijn een direct gevolg van het kernprobleem. Echter hebben deze drie problemen nog een aantal kleinere indirecte gevolgen: verminderde doorstroming van het verkeer in de stad, het ontstaan onveilige situaties in de stad, milieubelasting, ontevreden bewoners & omstanders, slecht voor het imago van de bestemming. Het kernprobleem is het donkerste gemarkeerd, zoals te zien is in Figuur 1.

Voor het maken van voorspellingen is er een belangrijk onderscheid te maken tussen real time, korte, middellange en lange termijn voorspellingen, zie Figuur 2. Er ligt veel markt in het tonen van real time data op het gebied van laadpalen, werkelijke filezwaarte, bezetting parkeergarages et cetera. Daarnaast zijn de voorspellingen van de korte termijn een hard groeiende markt. Middellange en lange termijn voorspellingen blijven relevant om investeringen van de overheid te verantwoorden.



FIGUUR 2 - TIJDSHORIZON

De grens tussen verschillende tijdshorizonnen is een vaag begrip. Dit komt door het feit dat voor elke voorspellingen meerdere variabelen nodig zijn dan alleen de tijdshorizon. Daar komt bij dat veel van deze variabelen overeenkomen, maar de methode voor de voorspellingen kan een groot verschil maken. Echter kan er wel altijd zonder real time data voorspeld worden, maar voor een korte tijdshorizon zullen de voorspellingen minder nauwkeurig zijn. Als er wordt gevraagd voor een voorspelling over vijf minuten is het aantrekkelijk om dit te doen aan de hand van actuele gegevens. Bij korte termijn voorspellingen in dit onderzoek wordt er gekeken naar de rest van de dag, morgen, maar ook naar volgende maand. Het handelingsprobleem voor dit onderzoek is als volgt gedefinieerd:

Handelingsprobleem: Er is geen instrument om de bezettingsgraad van parkeergarages op de korte termijn te voorspellen.

Op dit moment is er geen instrument, dus kan er niet worden vastgesteld hoe accuraat het model moet zijn. De toepassing van het model staat op dit moment nog niet vast, de eerste stap is het onderzoeken of er een instrument de bezettingsgraad van parkeergarages kan voorspellen. Daarom ligt de focus van dit onderzoek op het ontwikkelen van een model met een zo hoog mogelijke

betrouwbaarheid. Als deze voorspellingen een hoge betrouwbaarheid hebben en als ze worden meegenomen in de informatievoorziening voor parkeerroutes, zal dit een grote impact hebben op de situaties rondom parkeergarages en de tevredenheid van verschillende stakeholders. Ook gemeenten kunnen aan de hand van het model verschillende analyses en ontwerpen testen. Er zijn KPI's nodig om het handelingsprobleem meetbaar te maken en om hiermee aanwijzingen te geven over de mate van kwaliteit:

- ✓ De samenhang van de werkelijke waarneming en de voorspelling;
- ✓ Betrouwbaarheid van het model;
- ✓ De normaalverdeling van de residuen.

1.4 Onderzoeksdoel

Het doel van dit onderzoek is tweeledig. Ten eerste is er behoefte aan uitgebreid onderzoek naar verschillende voorspel methodieken. In meerdere opdrachten binnen Goudappel Groep zijn voorspellingen nodig, elk over een andere tijdshorizon. Ieder vraagstuk heeft andere kenmerken en andere doelen, waardoor voor ieder vraagstuk een andere methode de beste methode kan zijn. Hierdoor is het belangrijk om een goed overzicht te hebben van deze methoden met hun specificaties om zo de juiste te kunnen kiezen. Hierbij komt dat er ook onderzoek gedaan moet worden naar welke variabele relevant zijn en ook daadwerkelijk bijdragen aan de voorspelling. De combinatie van de juiste methode en de juiste variabelen maakt het mogelijk om een betrouwbare voorspelling te maken.

Daarnaast is het doel om een instrument te ontwikkelen om de bezettingsgraad van parkeergarages te voorspellen. Dit onderzoek richt zich dus op de drukte in de parkeergarage en niet eromheen. Een vervolgonderzoek, wat al gepland staat, zal zich richten op een instrument dat de in- en uitrijtijden kan voorspellen. Gecombineerd met de uitkomst van dit onderzoek, kan de drukte op het wegennetwerk rondom de parkeergarages voorspeld en geanalyseerd worden. Later in het proces kan het worden geïmplementeerd in het programma OmniTRANS Next. Aan de hand van deze software kunnen gemeenten de verkeersstromen analyseren en eventueel ingrijpen aan de hand van de voorspellingen. Ook de automobilisten zelf zullen baat hebben bij dit onderzoek, omdat zij vroegtijdig van informatie voorzien kunnen worden.

1.5 Probleemaanpak en verslagopbouw

Voor het oplossen van het handelingsprobleem uit paragraaf 1.3 is het belangrijk dat er genoeg informatie is over het kernprobleem. Het oplossen van een kennisprobleem helpt om het handelingsprobleem verder aan te pakken. Er staat één onderzoeksvraag vraag centraal die antwoord geeft op het beschreven handelingsprobleem:

Kennisprobleem: Welk instrument is geschikt voor het voorspellen van de bezettingsgraad van parkeergarages op de korte termijn?

Het onderzoek is verdeeld in vijf fases: de probleemanalyse, de formulering van alternatieve oplossingen, de ontwikkeling, resultaten en tot slot de evaluatie. Voor het beantwoorden van het kennisprobleem zijn meerder onderzoeksvragen nodig. Per onderzoeksvraag wordt aangegeven in welke fase deze behandeld wordt en het onderzoeksontwerp wordt beschreven. In Appendix A is een overzicht te zien van de indeling van de hoofdstukken die aangehouden zal worden. Ook is bij elk hoofdstuk aangeven welk hoofd- en deelvragen beantwoord zullen worden.

1.5.1 Probleemanalyse

Tot deze fase behoort de introductie die in de paragrafen hiervoor te lezen is. Als aanvulling voor de probleemidentificatie uit Hoofdstuk 1 richt Hoofdstuk 2 zich op de analyse van de huidige informatievoorzieningen voor het gebruik van parkeergarages. Het is van belang om te kijken naar de huidige situatie voordat er gekeken kan worden naar methoden om de huidige informatievoorzieningen te verbeteren of nieuwe modellen te ontwikkelen. Dit wordt gedaan aan de hand van de volgende vragen:

- Hoe worden weggebruikers van informatie voorzien voor het parkeren in parkeergarages?
    - Welke verwijssystemen en apps zijn er in Nederland in gebruik?
    - Welke technieken worden gebruikt voor het bepalen van de bezettingsgraad?
    - Wat zijn de prestaties van de huidige systemen?

1.5.2 Formulering van alternatieve oplossingen

Als de huidige situatie in kaart is gebracht, wordt er vervolgens onderzoek gedaan in de literatuur naar voorspel methodieken die een mogelijke oplossingen kunnen zijn voor het handelingsprobleem. Naast dat alleen de theorie van verschillende methoden wordt beschreven in Hoofdstuk 3, wordt ook gekeken naar hoe deze methoden van toepassing kunnen zijn. In de literatuur zal gezocht worden naar mogelijke oplossingen aan de hand van de volgende vragen:

- Wat zijn potentiele methoden om voorspellingen te kunnen doen op basis van historische data?
    - Welke variabelen komen in de literatuur voor die relevant zijn en bijdragen aan de voorspelling?
    - Welke methoden zijn er in de literatuur bekend om patronen en trends in data te kunnen analyseren?
    - Welke methoden zijn er in de literatuur bekend om een vergelijkbaar voorspel probleem op te lossen?

1.5.3 Ontwikkeling

De fase ontwikkeling bestaat uit twee hoofdstukken. Eerst komt in Hoofdstuk 4 de analyse van de historische gegevens aan bod. Het analyseren van deze gegevens wordt gedaan aan de hand van het toevoegen van verschillende onafhankelijke variabelen. Deze verbanden van deze variabelen met de bezettingsgraad van parkeergarages worden onderzoek. Met alle aanwezige kennis wordt er antwoord gegeven op de volgende vragen:

- Welke verklaringen zijn er voor de patronen in de dataset?
    - Welke verschillen zijn er tussen de parkeergarages?
    - Wat zijn de beperkingen van de dataset?
    - Hoe passen alle variabelen bij elkaar in één dataset?
    - Welke patronen zijn er in de dataset aanwezig?

Vervolgens richt Hoofdstuk 5 zich op de implementatie van de gekozen oplossing uit Hoofdstuk 3. De methode zal worden gebruikt en toegespitst op het onderwerp van dit onderzoek. Het model wordt ontwikkeld en het proces hiervan wordt beschreven. Ook wordt het uiteindelijke model toegelicht. De vragen die hier centraal staan zijn:

- Hoe kan het proces worden weergegeven in een model?
    - Hoe wordt de data opgesplitst in een train en testdataset?
    - Wat zijn de in- en output variabelen van het model?
    - Hoe komt het model en de visualisatie hiervan tot stand?

1.5.4 Experimenten & analyses

Tot deze fase behoort Hoofdstuk 6 waarin de resultaten worden beschreven. Zodra het model is ontwikkeld en de methodiek om te voorspellen is geïmplementeerd, kunnen de uitkomsten met elkaar worden vergeleken. De verwachte impact van het ontworpen model wordt geanalyseerd. De vraag die in deze fase wordt beantwoord is:

- Wat is de prestatie van het instrument op de train en test datasets?

1.5.5 Evaluatie

Tot slot wordt het model geëvalueerd en worden er conclusies getrokken en aanbevelingen gedaan. In Hoofdstuk 7 wordt eerst naar het model gekeken en hoe dit toegepast zou kunnen worden in binnen Goudappel Groep. Ook wordt er gekeken naar welke uitbreidingen nog mogelijk zijn en hoe deze in een toekomstig onderzoek uitgewerkt kunnen worden. De vragen dit centraal staan in dit hoofdstuk zijn als volgt:

- Wat is het antwoord op de hoofdvraag?
    - In welke andere situaties en applicaties kan dit nieuwe instrument van belang zijn?
    - Welke verdere uitbreidingen zijn er mogelijk met het model voor Goudappel Groep?

Hoofdstuk 2 – Huidige informatiesystemen

Dit hoofdstuk zal ingaan op de verschillende systemen die automobilisten van informatie voorzien in de buurt van parkeergarages (§2.1). Er wordt daarmee antwoord gegeven op de onderzoeksvraag: *"Hoe worden weggebruikers van informatie voorzien voor het parkeren in parkeergarages?"*. De deelvragen zullen in dit hoofdstuk steun bieden om deze vraag te beantwoorden:

- Welke verwijssystemen en apps zijn er in Nederland in gebruik?
- Welke technieken worden gebruikt voor het bepalen van de bezettingsgraad?
- Wat zijn de prestaties van de huidige systemen?

De applicatie van DAT.Mobility wordt toegelicht en ook het verband met dit onderzoek (§2.2). Verder volgt er een beschrijving over hoe de systemen informatie opslaan en of deze gegevens bruikbaar zijn voor analysedoeleinden (§2.3). Tot slot wordt de prestatie van het parkeersysteem voor parkeergarages onderzocht (§2.4) en wordt de onderzoeksvraag beantwoord (§2.5).

2.1 Informatievoorziening omtrent parkeergarages

Nederland telt in 2018 ruim 11,2 miljoen wegvoertuigen (CBS, 2018). De auto is dan ook niet meer weg te denken uit het straatbeeld van nu. Echter staat een auto volgens Krol (2017) van milieudefensie 95% van de tijd stil. Er zijn dus veel parkeerplaatsen nodig en in stedelijk gebied zorgt dit voor grote problemen. Gemiddeld bestaat ongeveer dertig procent van al het stadsverkeer uit automobilisten die op zoek zijn naar een vrije plek (Wentink, 2014). Dit is een bedreiging voor de kwaliteit van leven en de bereikbaarheid van de stad. Uit het onderzoek van Marsden (2006) komt naar voren dat de gemiddelde tijd dat automobilisten zoeken naar een vrije parkeeplaats ruim 8 minuten bedraagt. Gemeenten nemen maatregelen om het probleem te verminderen. Er worden maatregelen getroffen zoals het aantal parkeerplaatsen vergroten waar dat mogelijk is, aanpassingen in de parkeertarieven of (nieuwe) vormen van informatievoorzieningen. Het is echter niet altijd duidelijk of de nieuwe methoden van informatievoorzieningen ook daadwerkelijk het probleem verhelpen.

2.1.1 Parkeerverwijssysteem

In veel gemeenten worden verwijssystemen gebruikt om op een heldere en directe wijze automobilisten naar vrije parkeervoorzieningen te leiden. Het Parkeer Route Informatie Systeem (PRIS) is een veelgebruikt systeem en informeert het verkeer over de parkeermogelijkheden in of rond een stad, dorp of object met borden naast de weg. De eerste parkeerverwijssystemen in Nederland bestonden uit statistische verwijzingen die enkel de route aangaven naar de dichtstbijzijnde parkeerlocatie. Echter wist de automobilist die het systeem volgde niet of er bij aankomst op de parkeerlocatie parkeerruimte vrij was (CROW, 2007). Daarom volgde al snel de eerste dynamische parkeerverwijssystemen. De eerste dynamische parkeerverwijssystemen informeerden de automobilist over de route naar de parkeerlocatie en de bezetting van de parkeerfaciliteit. Deze bezetting werd binair weergegeven, VOL of VRIJ. Een groot voordeel van de dynamische parkeerverwijssystemen is de aanzienlijke vermindering van het zoekverkeer in de stad. Doordat automobilisten worden voorzien van informatie over route en bezetting van de parkeerlocatie, wordt sneller een vrije parkeerplaats gevonden en verbetert hiermee de bezetting van de parkeerfaciliteiten (CROW, 2007). Het doel van de eerste dynamische parkeerverwijssystemen was vooral het informeren van automobilisten en nog niet zozeer het sturen.

De tweede generatie dynamische parkeerverwijssystemen deed in de jaren negentig haar intrede. In plaats van VOL/VRIJ-aanduidingen kan met verbeterde voorspellingsalgoritmes het aantal vrije parkeerplaatsen worden weergegeven. Naast de bezetting kunnen de systemen ook andere informatie verstrekken, zoals openingstijden, loopafstand en parkeerkosten. Deze tweede generatie

parkeerverwijssystemen heeft net als de eerste generatie het doel om de bereikbaarheid te bevorderen en de bezetting van de parkeerfaciliteiten te vergroten (CROW, 2007). Naarmate de afstand van een dynamische informatiepaneel tot de parkeerlocatie groter wordt dient de weergave afgestemd te worden op de toekomstige situatie. Het systeem dient dus een voorspelling te doen van het toekomstige aantal beschikbare plaatsen en hier de informatie op aan te passen voor het moment dat een automobilist arriveert (Wegenwiki, 2012).

Tegenwoordig wordt vaak een combinatie van statische en dynamische parkeerverwijssystemen gebruikt. Op basis van actuele situaties worden automobilisten via de snelste, kortste of gewenste route naar de meest geschikte parkeerfaciliteit nabij hun bestemming geleid (CROW, 2007). Naast verwijssystemen zijn er ook verkeersmanagementsystemen om de doorstroming van het verkeer te regelen. Verkeersregelinstallaties (VRI) zijn op dit gebied essentieel, zoals verkeerslichten. Met het geven van optische signalen aan weggebruikers kunnen een of meerdere verkeersstromen geregeld worden.

Momenteel zijn er ontwikkelingen op het gebied van de integratie van parkeerverwijssystemen met andere verkeersmanagementsystemen. Verkeersregelinstallaties kunnen bijvoorbeeld rijrichtingen naar beschikbare parkeergarages bevoordelen en rekening houden met de geregistreerde uitstroom bij een parkeerfaciliteit. Omgekeerd kunnen parkeerverwijssystemen verwijzingen richten op de parkeerlocatie die op dat moment goed bereikbaar zijn. Het is ook mogelijk om op grond van de actuele verkeersituatie een alternatieve route naar een parkeerlocatie aan te bieden. De invloed van parkeergelegenheden kan cruciaal zijn voor de doorstroom van het verkeer. Voorspellingen op de in- en uitstroom hiervan kunnen een groot verschil maken.

2.1.2 Parkeerapp

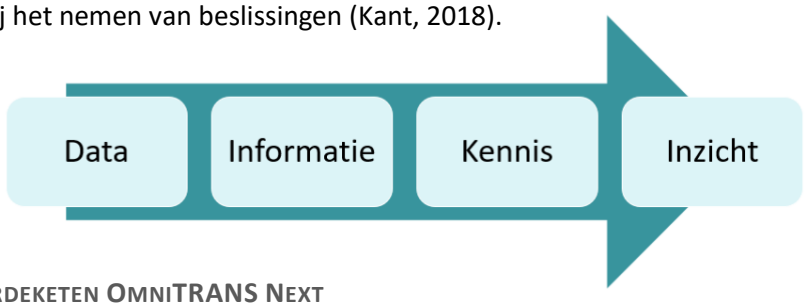
Ook parkeerapps hebben hun intrede gedaan. Tegenwoordig is het mogelijk om in ruim honderd gemeenten te betalen met een app. Er hoeft niet van tevoren worden nagedacht over de duur van het parkeren. Er zijn wel verschillen tussen de apps, maar ze bieden globaal dezelfde mogelijkheden. Ze tonen een overzicht van parkeerlocaties in de buurt. Bij het arriveren start je de parkeeractie en wanneer je de parkeerlocatie wil verlaten, stop je de parkeeractie. Echter is een nadeel van de parkeerapps, dat registratie vrijwel altijd noodzakelijk is en het bedrijf brengt vrijwel altijd extra servicekosten in rekening per parkeeractie. Daarnaast bestaat het risico dat een gebruiker niet goed afmeldt waardoor de rekening torenhoog oploopt. Het is ook een cruciaal punt dat je nog niet kunt zien waar er nog vrije parkeerplaatsen zijn. Volgens ikvergelijkhet.nl (2018) is ParkMobile de beste app om je parkeerplek te betalen met de telefoon. Er zijn geen registratiekosten of maandelijkse kosten gekoppeld aan deze app, wat een groot voordeel is tegenover de andere bestaande apps, zoals YellowBrick en Park-Line. Daarnaast is ook het tarief per uur relatief laag. Het parkeren kan betaald worden via bellen, sms of een app. Daarnaast zijn er ook nog eens veel steden gekoppeld aan deze app, waardoor het in bijna 200 steden in Europa mogelijk is om te betalen met ParkMobile.

Naast deze app is TimesUpp een andere grote speler op de markt die zich meer richt op navigatie. (Kuijten, 2016). Deze app is gekoppeld met je agenda en geeft een melding als je moet vertrekken om op tijd bij je afspraak te zijn. In de route zijn files, vertragingen en ook parkeertijd meegerekend. Het doel van de app is daarom ook dat je nooit meer te laat hoeft te komen op je eerstvolgende afspraak. Toch heeft deze app ook zijn nadelen. Het is namelijk niet mogelijk om vervoersmiddelen te combineren, zoals eerst de auto pakken naar een treinstation en vanaf daar het laatste stuk met de trein te reizen. Daarnaast geeft de app wel de huidige bezettingsgraad van parkeergarages, maar neemt hier geen voorspellingen in mee. Als de verwachte aankomsttijd over anderhalf uur is, dan geeft de app de huidige parkeerbezetting weer.

2.2 OmniTRANS Next

DAT.Mobility is volop bezig met het ontwikkelen van voorspellingen op korte en lange termijn. Deze kunnen worden toegevoegd in de applicatie OmniTRANS Next, die nu in de ontwikkelfase is. De verwachting is dat deze applicatie binnen 3 jaar in heel Nederland te gebruiken is op het gebied van strategie, beleid en operationeel verkeermanagementniveau. OmniTRANS is een data-gedreven platform waarin demografische- en verkeersgegevens samenkomen en worden verwerkt tot bruikbare prognoses en analyses voor relevante tijdshorizonnen. Het is uitermate geschikt voor het modelleren van de interacties tussen de verschillende transportmiddelen binnen een stedelijke context. OmniTRANS maakt een duidelijk onderscheid tussen statische en dynamische toewijzingsmethoden en biedt tegelijk een soepele overgang tussen de twee methoden. Het dynamische toedeling algoritme (DTA) van StreamLine maakt gebruik van een bewezen methodologie, modelleert de effecten van de spreiding van het verkeer in de tijd op de verkeersafwikkeling en de op- en afbouw van files en wachtrijen (Brederode, z.d.).

OmniTRANS is voorzien van een gedetailleerd en ingrijpend algoritme voor het modelleren van alle types van kruispunten en het biedt rijke interfaces om de geometrische lay-out, de signaalplannen en de verschillende configuratieperiodes van kruispunten te definiëren. Het concept van 'dimensies' zorgt voor een juiste opslag van de gegevens en vereenvoudigt de samenwerking tussen meerdere gebruikers. Nieuwe innovaties en algoritmes kunnen eenvoudig geïntegreerd worden via speciaal ontworpen modelklassen. OmniTRANS biedt een breed scala van formaten voor de uitwisseling van gegevens en mechanismen voor uitwisseling met andere systemen. In Figuur 3 is de bijbehorende waardeketen te zien, die begint bij het verzamelen van data. Gevolgd door informatie genereren uit deze databronnen om vervolgens hier kennis uit te halen en stakeholders inzicht te geven en te ondersteunen bij het nemen van beslissingen (Kant, 2018).



FIGUUR 3 - WAARDEKETEN OMNITRANS NEXT

De informatie van het model wat in dit onderzoek centraal staat, maakt veel nieuwe analyses mogelijk. Vooral als ook de in- en uitrijtijden worden geanalyseerd kan de intensiteit van het wegennetwerk rondom parkeergarages beter verklaard worden. Het zal een klein onderdeel zijn van de applicatie en wellicht niet voor elke gemeenten interessant, maar het biedt wel kansen om te laten zien dat ze het kunnen.

2.3 Technieken voor het bepalen van de bezettingsgraad

Om onderzoek te kunnen doen naar de bezettingsgraad van parkeergarages en om deze uiteindelijk te kunnen voorspellen, zijn historische gegevens nodig over de bezetting van parkeergarages. Verschillende systemen worden gebruikt om deze parkeergegevens te werven en op te slaan. Het is van groot belang dat parkeergegevens worden opgeslagen om zo analyses te kunnen doen voor de problemen omtrent parkeren. Het geeft gemeenten inzicht waar de knelpunten liggen en waar veranderingen noodzakelijk zijn. Elk systeem heeft zijn eigen manier om historische data op te slaan en heeft zo zijn eigen voor- en nadelen. Er kunnen er gegevens verkregen worden uit parkeermeters, parkeergarages en parkeersensoren. In deze paragraaf worden alle drie de systemen besproken, maar uiteindelijk zal de focus vooral liggen op de datagegevens uit parkeergarages.

2.3.1 Parkeerrechtendatabase

Vroeger moesten automobilisten betalen bij een parkeermeter, die terugliep zodra er munten in waren gevoerd totdat de parkeertijd voorbij was. Parkeerwachters konden op die manier controleren of een auto betaald had voor de parkeerplaats. Later volgden parkeerautomaten die bonnetjes printten, die in de auto gelegd dienden te worden. Ook dit kon visueel gecheckt worden.

Tegenwoordig zijn er ook andere mogelijkheden om te betalen en om te controleren of een automobilist heeft betaald. Automobilisten kunnen nu inbellen of hun kenteken invoeren waardoor in het parkeerautomaat een digitaal parkeerrecht ontstaat dat uiteindelijk ook in de parkeerrechtendatabase wordt opgeslagen. Een parkeerwachter kan vervolgens met behulp van kentekenscanapparatuur automatisch controleren of de auto op het bewuste tijdstip voor die specifieke buurt een geldig parkeerrecht heeft. Dit zorgt voor een groot efficiëntievoordeel vergeleken met de methode uit het verleden. Daarnaast is het ook een groot voordeel dat deze gegevens in een gestructureerd format worden opgeslagen. Door de parkeerrechtendatabase is het voor een onderzoeker mogelijk om hier analyses mee uit te voeren, die zonder een dergelijk systeem niet mogelijk zijn (Grooten, Gelder, & Spruijt, 2012).

2.3.2 Parkeersensoren

Er zijn ook ontwikkelingen die nog in de kinderschoenen staan. Parkeersensoren is er hier een van. Dit zijn draadloze parkeersensoren die de real-time bezetting per parkeerplek monitoren en de tijdsduur dat een voertuig geparkeerd staat registreert. Dat doen ze zowel aan de hand van veranderingen in het lokale aardmagnetische veld als met infrarood. Slimme algoritmes bepalen steeds wat het beste bruikbaar is: een van beide technologieën of de combinatie. Zo kunnen automobilisten verwezen worden naar vrije parkeerplaatsen in de buurt van de bestemming. De beschikbare parkeercapaciteit wordt hierdoor beter benut en parkeerplekken kunnen efficiënter worden gehandhaafd (Wentink, 2014). De real-time informatie kan ook op dynamische verwijsborden worden weergegeven.

Het systeem minimaliseert gegevensverlies door op verschillende niveaus in het netwerk lokaal te bufferen. Wanneer er een stroomstoring is, blijven de gegevens bewaard. De datacollector verzendt de parkeerstatusgegevens naar een softwaretoepassing, die via een webbrowser direct te gebruiken zijn. Hiermee is de bezettingsgraad van parkeerplaatsen te configureren, te monitoren en te analyseren om het verkeer effectief te begeleiden (Wentink, 2014).

2.3.3 Parkeerverwijssysteem

Naast de bovengenoemde parkeergelegenheid op straat ligt de focus in dit onderzoek op het parkeren in parkeergarages. Het parkeerverwijssysteem geeft dynamisch weer hoeveel plekken er nog vrij zijn of dat de garage vol is. Hiervoor zijn de parkeerverwijssystemen uitgerust met een database met daarin per garage en per parkeeracties de in- en uitrijtijdstippen en het type parkeerder (abonnementhouder of kortparkeerder). Per tijdsperiode is ook de capaciteit en de bezetting beschikbaar (Grooten et al, 2012). Naast de mogelijkheden van deze database voor het informeren van de automobilist kunnen deze gegevens ook gebruikt worden voor analysedoeleinden. De informatie uit het verkeerverwijssysteem is per garage per 15 minuten beschikbaar.

Een nadeel van parkeerverwijssystemen is dat deze pas op de bestemming te zien zijn. Er is geen mogelijkheid om aan het begin van je reis dit op te zoeken wat de bezetting zal zijn op het moment van arriveren. Het is dus de bedoeling dat je naar je bestemming rijdt en daar verwezen wordt naar beschikbare parkeerplaatsen. Dit kan als lastig ervaren worden omdat de naam van de parkeerplaats op de borden staat en niet de locatie. Ook kan dit onhandig zijn als de automobilist de stad niet kent en niet precies weet waar de bestemming is ten op zich van de parkeerlocatie.

2.4 Beoordeling parkeerverwijssysteem

In de afgelopen jaren zijn er een aantal onderzoeken gedaan naar de effectiviteit van de parkeerverwijssystemen. In het onderzoek van Van de Zande (2013) is het gebruik van de parkeerverwijssystemen geanalyseerd. In totaal heeft 56% van de ondervraagde automobilisten de parkeerverwijssystemen opgemerkt. Echter heeft hiervan maar 11.4% zich ook daadwerkelijk laten sturen door het systeem. De reden die de automobilisten geven voor het niet opmerken en niet gebruiken van het systeem is de bekendheid met de verkeers- en parkeersituatie. Van de automobilisten die het systeem wel hebben gebruikt, geeft 87% aan de naam van de faciliteit te hebben gezien, 40% de route en tot slot heeft 95% de beschikbaarheid gezien. De automobilisten die de parkeerverwijsborden wel hebben waargenomen, maar niet hebben gebruikt, geven bijna allemaal aan dat dit komt omdat zij bekend zijn met de parkeervoorzieningen.

Van de Zande (2013) concludeert in haar onderzoek dat parkeerverwijssystemen voornamelijk waardevol zijn voor bepaalde type reizen, zoals werken of winkelen. Het gebruik van parkeerverwijssystemen zal toenemen als de systemen worden afgestemd op specifieke doelgroepen. Het kennisniveau van de automobilist over de verkeers- en parkeersituatie is de belangrijkste factor die het effect beïnvloedt. Echter is de installatie en het beheer van parkeerverwijssystemen kostbaar en daardoor moeten gemeenten zich afvragen of het gebruik van dit systeem rendabel is. Daar komt bij dat er steeds meer technologische ontwikkelingen zijn die dynamische systemen op straat wellicht overbodig maken. Bijna elke automobilist heeft tegenwoordig een navigatiesysteem in de auto die hen navigeert naar de eindbestemming en daarnaast zijn smartphones met bijbehorende apps ook in opkomst. Deze laatste twee technieken kunnen de automobilist gepersonaliseerde informatie geven over de route en parkeren wat parkeerinformatie op straat overbodig kan maken.

Daarnaast is het een discutabel punt wanneer een parkeergarage vol is. In het eerste opzicht zou men zeggen als er géén parkeerplaats meer vrij is. Echter is uit onderzoek van prof. ir. P. Hakkesteegt (ca 1990) van de TU Delft een algemeen aanvaard criterium in de parkeerbranche opgesteld: *De gemiddelde acceptabele bezettingsgraad wordt vastgesteld dat op het drukste uur van de dag de bezettingsgraad op de maatgevende dag nog ongeveer 10 % vrije parkeerplaatsen beschikbaar zijn.* Dit komt voort uit het feit dat bij een hogere bezettingsgraad dan 90% aankomend verkeer niet altijd makkelijk een vrije parkeerplaats meer kan vinden. De parkeerlocatie wordt dan als (te) vol ervaren.

Volgens wethouder Kammeijer-Luycks (2017) van de gemeente Bergen op Zoom is het huidige parkeerverwijssysteem verouderd en niet meer aan te passen aan de huidige stand der techniek. Volgens hem is investeren in deze systemen niet rendabel. Zoals ook uit het onderzoek van Van de Zande (2013) naar voren komt, trekken automobilisten zich vaak weinig tot niets aan van de informatie op de borden en gaan naar een parkeerlocatie die ze bij vertrek als doel hadden. Daar aangekomen gaan zij voor de betreffende parkeerlocaties in de rij staan. Daarnaast zegt wethouder Kammeijer-Luycks dat er steeds meer technologische ontwikkelingen zijn die dynamische systemen naast de weg op termijn overbodig maken. Kortom, investeren in een parkeerverwijssysteem met digitale informatie is een oplossing waarvan de effectiviteit te wensen overlaat en in de nabije toekomst door technologie in de auto en smartphones wordt vervangen.

2.5 Conclusie

In dit hoofdstuk is er gekeken naar de verschillende technieken die worden gebruikt om automobilisten van informatie te voorzien. Daarnaast zijn er ook slimme technieken en ontwikkelingen die het parkeren met alles erop en eraan makkelijker maken. Kortom, er is antwoord gegeven op de eerste onderzoeksvraag:

- Hoe worden weggebruikers van informatie voorzien voor het parkeren in parkeergarages?
    - Welke verwijssystemen en apps zijn er in Nederland in gebruik?
    - Welke technieken worden gebruikt voor het bepalen van de bezettingsgraad?
    - Wat zijn de prestaties van de huidige systemen?

Het meest gebruikte en bekende systeem in Nederland is het parkeerverwijssysteem. Het doel van dit systeem is om de bereikbaarheid te bevorderen en de bezetting van de parkeerfaciliteiten te vergroten. Doormiddel van een combinatie van statische en dynamische parkeerverwijssystemen worden automobilisten van informatie voorzien. De bezettingsgraad van de garage wordt weergegeven met aantallen of een VOL/VRIJ-aanduiding. Ook de locatie, openingstijden en parkeerkosten kunnen worden weergegeven.

Door het telefonisch inbellen of het invoeren van je kenteken, kan tegenwoordig veel gemakkelijker gecontroleerd worden of er is betaald voor een parkeerplek. Omdat dit met kentekenscanapparatuur gebeurd, worden deze gegevens direct opgeslagen en kunnen deze ook voor analyses worden gebruikt. Daarnaast zijn er ook parkeersensoren, die kunnen 'zien' of er een auto geparkeerd staat op de desbetreffende parkeerplaats of niet. Ook deze gegevens worden automatisch opgeslagen en verzonden naar een softwaretoepassing waardoor ze direct te gebruiken zijn in een webbrowser. Het parkeerverwijssysteem is uitgerust met een database met daarin de in- en uitrijtijdstippen van elke automobilist en de totale bezetting per kwartier. Er zijn dus meerdere technieken die elk op een andere manier gegevens opslaan.

Uit onderzoek blijkt dat het parkeerverwijssysteem niet het doel bereikt. Zo merkt maar 56% van de automobilisten het systeem op en wordt het door 11,4% daadwerkelijk gebruikt. Dit is een dermate laag percentage, dat het de vraag oproept of het systeem wel rendabel is. Daar komt bij dat er veel technologische ontwikkelingen zijn, die dit systeem overbodig maken. Zo worden de bezettingsgraden al meegenomen in navigatie apps. Dit kan zeker nog beter en zal nog verder ontwikkeld moeten worden, maar het is wel een begin van nieuwe mogelijkheden. Een ander belangrijk punt wat naar voren is gekomen is het gevoel van automobilisten wanneer een parkeergarage vol is. Voor de acceptabele bezettingsgraad is vastgesteld dat er nog ongeveer 10% van de parkeerplaatsen vrij moeten zijn, wil een automobilist niet het gevoel hebben dat een parkeergarage vol is.

Er is dus behoefte aan een systeem dat vroegtijdig kan voorspellen hoe druk het zal zijn in een parkeergarage op het moment van arriveren. Dit kunnen automobilisten dan al van tevoren meenemen in hun keuze waar ze zullen parkeren en eventueel een andere parkeergarage kiezen. Dit zal leiden tot minder zoekverkeer en minder wachtrijen in steden rondom parkeergarages. Daarnaast is het ook goed voor gemeenten om deze informatie tot hun beschikking te hebben. Zo kunnen gemeenten analyses doen om te onderzoeken of er meer parkeerplaatsen in een stad nodig zijn. Tot slot zijn gemeenten ook altijd geïnteresseerd in de *"Common Operational Picture"*. Dit wil zeggen dat er een overzicht is van alle posities en de status van belangrijke infrastructuur, waar parkeergarages ook een belangrijk onderdeel van uitmaken.

Hoofdstuk 3 – Theoretisch kader

Dit hoofdstuk bevat een literatuuronderzoek om te onderzoeken welke methoden er zijn om voorspellingen te kunnen genereren op basis van historische data. Daarom is er een onderzoeksvraag geformuleerd die luidt: "*Wat zijn potentiele methoden om voorspellingen te kunnen doen op basis van historische data?".* Deze onderzoeksvraag wordt aan de hand van enkele deelvragen beantwoord:

- Welke variabelen komen in de literatuur voor die relevant zijn en bijdragen aan de voorspelling?
- Welke methoden zijn er in de literatuur bekend om patronen en trends in data te kunnen analyseren?
- Welke methoden zijn er in de literatuur bekend om een vergelijkbaar voorspelprobleem op te lossen?

Volgens Waller & Fawcett (2013) is *data science* de applicatie van kwantitatieve en kwalitatieve methoden om relevante problemen op te lossen en uitkomsten te voorspellen. Voordat er voorspellingen gemaakt kunnen worden is er eerst een dataset nodig met daarin alle relevante onafhankelijke variabelen en de afhankelijke variabele. Er wordt eerst gekeken naar welke variabelen in andere onderzoeken op het gebeid van parkeren zijn gebruikt en welke relevant zijn (§3.1).

Als het duidelijk is welke variabelen relevant zijn en alle gegevens zijn verzameld is de volgende stap het "schoonmaken" van deze data. Denk hierbij aan ontbrekende waarden of het berekenen van gemiddelde waarden over de tijdshorizon. De patronen van deze data worden bekeken en afwijkend gedrag kan wellicht al verklaard worden. In paragraaf 3.2 zal daarom onderzoek gedaan worden naar methoden die patronen en trends in data te kunnen herkennen en daar ook betekenis aan kunnen geven.

Om de stap te maken naar het voorspellen, is het belangrijk dat de correlaties tussen de verschillende onafhankelijke variabelen en de afhankelijke variabele worden gecheckt. Verschillende methoden worden uiteindelijk beschreven die voorspellingen kunnen maken op basis van historische data (§3.3). Door resultaten uit andere onderzoeken zullen de voor- en nadelen naar voren komen, zodat er een duidelijk beeld ontstaat van de methoden (§3.4). Tot slot worden de onderzoeksvragen van dit hoofdstuk samengevat (§3.5).

3.1 Variabelen definiëren

De drukte in parkeergarages hangt van verschillende factoren af en daarom kan de lijst van de onafhankelijke variabelen ook heel lang zijn. Om dit te beperken is er gekeken naar 11 onderzoeken met vergelijkbare onderwerpen om te bepalen welke variabelen volgens andere onderzoekers relevant zijn. De afhankelijke variabele in dit onderzoek is de bezettingsgraad van parkeergarages. Dus specifiek gezegd het percentage geparkeerde auto's op een bepaald punt op de tijdshorizon. Hieruit kan natuurlijk gemakkelijk het aantal vrije plaatsen geconcludeerd worden. Deze variabele is afhankelijke van verschillende onafhankelijke variabelen. Een samenvatting van al deze variabelen met het bijbehorende onderzoek waarin de variabelen voorkomen, is te zien in Tabel 1.

De variabele die in elk onderzoek naar voren komt is de variabele *tijdstip van de dag.* Het is logisch te beredeneren dat een parkeergarage tijdens kantooruren drukker bezet is dan midden in de nacht. Daarnaast kan het ook zijn dat de parkeergarage in het centrum ligt en daardoor zal de bezetting dus ook afhankelijk zijn van de openingstijden van de winkels eromheen.

Een tweede variabele die in de literatuur veel voorkomt is *de specifieke dag van de week.* Het grootste verschil is tussen werkdagen en weekenddagen. In het weekend kan het drukker zijn in de

parkeergarage omdat er veel mensen vrij zijn en in het centrum kunnen winkelen of gebruik kunnen maken van andere vrijetijdsdoeleinden. Toch kan er ook verschil zitten in de werkdagen en dit verschil is daarom ook belangrijk om mee te nemen.

Het *weer* is een volgende variabele die ook in het Nederlandse onderzoek van Weijermans & Berkum (2004) wordt meegenomen. Mensen worden vaak beïnvloed door het weer, wellicht gaan ze normaal op de fiets, maar als het regent pakken ze misschien toch liever de auto. Deze variabele kan opgesplitst worden in verschillende factoren, zoals de temperatuur, luchtvochtigheid, totaal aantal zonuren, regen, wind et cetera.

Verder wordt door Chen, Pinelli, Sinn, Botea, & Calabrese (2013) en Reinstadler, Braunhofer, Elahi, & Ricci (2013) de variabele *vakantie* genoemd. Hierin wordt gekeken naar de schoolvakantie die invloed kunnen hebben op de drukte in parkeergarages. Hierdoor kunnen er op de werkdagen ook hele andere patronenen ontstaan.

In het onderzoek van Fabusuyi, Hampshire, Hill, & Sasanuma (2014) komt de variabele *speciale gebeurtenis* naar voren. In de buurt van een parkeergarage kunnen grote evenementen plaatsvinden of andere vrijetijdsbestedingen. Denk hierbij aan een theater, bioscoop of voetbalstadion. De schemas van deze evenemten kunnen invloed hebben op de drukte van parkeergarages in de buurt.

Onderzoek	Tijdstip van de dag	Dag van de week	Vakantie	Weer	Speciale gebeurtenis
Chen (2014)	X	X			X
Chen et al. (2013)	X		X	X	
Fabusuyi et al. (2014)	X	X		X	X
Kunjithapatham et al. (z.d.)	X	X			
Nieuwborg (2013)	X	X	X	X	X
Richter et al. (2014)	X	X			
Reinstadler et al. (2013)	X		X	X	X
Soler (2015)	X	X			
Vlahogianni et al. (2014)	X	X			
Weijermans et al. (2004)	X	X	X	X	X
Zheng et al. (2015)	X	X		X	

TABEL 1 - ONAFHANKELIJKE VARIABELEN

Naast deze bovengenoemde variabelen zijn er nog andere onafhankelijke variabelen die relevant zijn als er meerdere parkeergarages met elkaar worden vergeleken. Dan wordt de *locatie* van de parkeergarage namelijk belangrijker, omdat de locatie invloed zal hebben op de drukte van een parkeergarage. Zo kan een parkeergarage midden in een stadcentrum liggen omringd door winkels of juist meer in de buurt van bedrijven en scholen waar het in het weekend rustiger zal zijn. Ook de *capaciteit* van de parkeergarage speelt een rol als meerder garages met elkaar worden vergeleken. In deze variabele kan onderscheid gemaakt worden tussen tussen klein, middelgroot en groot.

Daarnaast zijn er ook variabelen zoals *prijs, afstand* en *beschikbaarheid* die naar voren komen in onder anderen het onderzoek van Chen (2014) en van Vlahogianni et al. (2014). Echter zijn voor dit onderzoek deze variable niet direct van belang. Deze variabelen hebben meer te maken met de keuze en het gedrag van mensen terwijl dit onderzoek verkeerkundig is ingestoken. Deze variabelen zijn wel van belang als er meerdere parkeergarages in één stad met elkaar worden vergeleken.

3.2 Data-analyse

Nu in de vorige paragraaf alle belangrijke variabelen zijn beschreven, is het van belang dat alle gegevens voor deze variabelen worden verzameld. Deze verzamelde informatie is uiteindelijk de dataset waarmee het model gevoed zal worden. Daarom is het van belang dat er geen gaten in de data zitten en dat de gegevens overeenkomen met elkaar. Denk hierbij aan het feit dat alle data dezelfde tijdseenheid moet hebben en dat er geen lege cellen in de data kunnen zitten. Als er namelijk lege cellen in een dataset zitten kan dit in de voorspelling leiden tot een lagere betrouwbaarheid van het model. Het klaarmaken en analyseren van deze gegevens heet in de literatuur *datamining*.

Datamining verwijst naar het proces van semiautomatische, betrouwbare en intelligente analyses van grote sets onbewerkt gegevens en de ontdekking van nuttige kennis, informatie, instructies en antwoorden (Zarafani, Abbasi, & Liu, 2014). Dataminingtechnieken worden steeds efficiënter en goedkoper en daarom zeer interessant voor bedrijven om te gebruiken. Deze methoden kunnen niet alleen expliciete informatie ontdekken, maar ook impliciete kennis doormiddel van data-analyses. Volgens Arumugam (2017) is de eerste stap in het proces het genereren of verzamelen van een dataset met verschillende variabelen. De informatie in deze dataset moet in verband staan met de kennis die nodig is bij het voorspellen van de afhankelijke variabele. Gegevens opschonen is een methode om onvolledige, inconsistente en nietszeggende gegevens te vervangen. Om een betrouwbare en zekere voorspelling te krijgen, is het belangrijk dat nietszeggende waarden worden geëlimineerd. Ook moet de data getransformeerd worden in de passende lay-out die aansluit bij de andere gegevens. Alle gegevens worden dus verzameld, opgeschoond en getransformeerd.

3.2.1 Clusteren

Na het verzamelen van alle gegevens komt de stap clusteren. Deze stap is het proces van het groeperen van data-objecten, zodanig dat objecten in hetzelfde cluster sterk op elkaar lijken en objecten die tot verschillende clusters behoren, sterk van elkaar verschillen. Het clusteren van data presenteert hoogwaardige en bruikbare patronen. In dit onderzoek is het meest voor de hand liggend om per cluster één parkeergarage in te delen. Elke parkeergarage heeft zo zijn eigen kenmerkende patroon en zal dus altijd op zichzelf blijven lijken.

De meest gebruikte techniek om data te clusteren is de **K-means clusteringstechniek**. Bij deze techniek worden K-zwaartepunten gekozen, waarbij K het aantal gewenste clusters is. Elk punt wordt toegewezen aan het dichtstbijzijnde zwaartepunt en elke verzameling punten toegewezen aan een zwaartepunt is een cluster. Het zwaartepunt van een cluster wordt vervolgens bijgewerkt op basis van de punten die aan het cluster zijn toegewezen. Deze stappen kunnen worden herhaald en bijgewerkt totdat er geen enkel punt of zwaartepunt meer veranderd. Dit algoritme is wiskundig weergegeven in Figuur 4.

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

FIGUUR 4 - K-MEANS CLUSTERINGSTECHNIEK

3.2.2 Classificeren

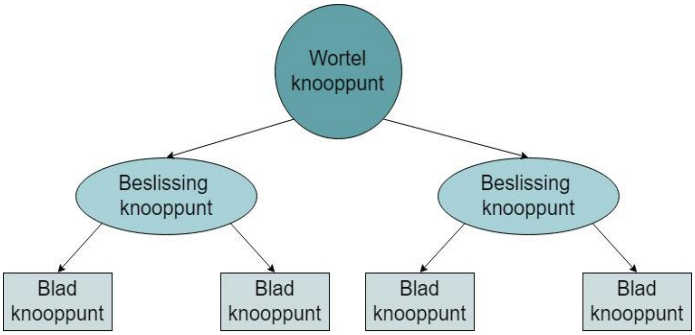
Naast clusteren is er nog een andere techniek die classificatie wordt genoemd en werkt net op een andere manier. Classificatie technieken zijn in staat om een grote hoeveelheid gegevens te verwerken. Het doel van classificatie is het groeperen van data in klassen van bekende labels. Het proces bestaat

uit twee fasen: de leerfase en de classificatie fase. In de leerfase worden training datasets geanalyseerd en wordt een classificatiemodel gebouwd. In de classificatie fase worden de testgegevens gebruikt om de nauwkeurigheid van de classificatie te schatten. De prestatie van een classificatiemodel is gebaseerd op het aantal punten dat correct is voorspeld door het model. De nauwkeurigheid is als volgt gedefinieerd in het onderzoek van Tomovic, Stanisic, & Kadic (2018):

$$Nauwkeurigheid = \frac{Aantal\ correcte\ voorspellingen}{Totaal\ aantal\ voorspellingen}$$

De meest bekende techniek om data te classificeren is de **beslissingsbom**, ook wel desicion tree genoemd. Deze techniek is vooral populair door het vermogen om complexe relaties te modeleren doormiddel van logische regels (Richardson & Lidbury, 2013). De studie van Tomovic et al. (2018) stelt dat de beslissingsboom drie soorten knoopunten heeft:

- Een wortelknooppunt die geen binnenkomende takken heeft, maar wel veel uitgaanden takken.
- Het beslissingsknooppunt, elk met precies één inkomend tak en ten minste twee uitgaande takken.
- Bladknooppunt, elk met precies één inkomende tak en geen uitgaande takken.



FIGUUR 5 - BESLISSINGSBOOM

De beslissingsboom combineert de variabelen op een hiërarchische manier zodat de belangrijkste beslissing zich in het wortelknooppunt bevindt. Elk knooppunt in de boom verwijst naar een van de variabelen. Beslissingsbomen worden als vanzelfsprekend beschouwd en het is daarom niet nodig om een expert te zijn om kennis uit een beslissingsboom te halen (Svoray, 2011). Een besluit in een beslissingsboom wordt volgens Tan et al. (2005) gemaakt op een recursieve wijze door de train dataset op te splitsen in subsets. Daarin is D_t de set van trainingspunten die zijn gekoppeld aan knooppunt t met $y = \{y_1, y_2, \ldots, y_c\}$ als de klasse labels. De volgende recursieve definitie van het algoritme voor de beslissingsboom is als volgt.

STEP 1: Als alle punten in D_t tot dezelfde klasse y_t t behoren, dan is t een bladknooppunt met label y_t.

STEP 2: Als de punten in D_t tot meer dan één klasse behoren, worden de punten verder onderverdeeld in kleinere subsets. Het algoritme wordt vervolgens recursief toegepast op elke set D_t.

Een variant van de beslissingsboom is een **regressieboom**. Het grootste verschil is dat een beslissingsboom de dataset verdeeld in klassen die horen bij een bepaalde variabelen. Vaak is deze uitkomst “Ja” of “Nee”. Het kan ook zijn dat er meer dan twee klassen worden onderverdeeld. Bij een regressieboom gaat het om numerieke en continue waarden, bijvoorbeeld de prijs van een product. Een ander verschil is de manier waarop de regressieboom tot stand komt. Aangezien de doelvariabele geen klassen heeft, wordt het regressiemodel toegepast op de doelvariabele met behulp van elk van de onafhankelijke variabelen. Elke onafhankelijke variabelen wordt op verschillende punten gesplitst en op elk punt wordt de error tussen de voorspelde waarde en de werkelijke waarden vastgesteld. Elke keer wordt de kleinste error gekozen als basisknooppunt en dit proces wordt recursief voortgezet. Aan de hand van dit model is het ook mogelijk voorspellingen te genereren.

3.3 Voorspellingsmethoden

Als alle benodigde gegevens zijn verzameld, opgeschoond, getransformeerd en geanalyseerd, dan is het tijd om te kijken naar de methoden die er zijn om voorspellingen mee te doen. In de afgelopen decennia zijn er veel voorspel methoden ontwikkeld om de toenemende variëteit en complexiteit aan te kunnen. Elke methode heeft zijn eigen specifieke kenmerken en voor elke toepassing moet de juiste methode gekozen worden. De keuze van een methode hangt van veel factoren af, zoals de relevantie en beschikbaarheid van historische gegevens, de gewenste mate van nauwkeurigheid, de te voorspellen tijdshorizon en de context van de voorspelling.

Voor bedrijven kunnen voorspellingen baanbrekend zijn. Voorspellingen helpen bedrijven om te zien wat er nog komt en zij kunnen hun acties daarop af stemmen (Debnath, 2018). Wanneer er wordt gesproken over voorspellen wordt de techniek bedoelt die doormiddel van historische gegevens de toekomst kan voorspellen. Het klinkt vrij eenvoudig, toch hebben bedrijven hier moeite mee. Grotendeels komt dit door de vele verschillende methoden die er zijn en ze niet goed weten welke methode het meest geschikt is voor hen. In dit onderzoek wordt er gekeken naar deze methoden en hoe ze beoordeeld kunnen worden voor verschillende toepassingen. Grofweg kunnen deze methoden in twee groepen gesplitst worden: *kwalitatieve* en *kwantitatieve* modellen (Reid & Sanders, 2011)

3.3.1 Kwalitatieve voorspelmodellen

Kwalitatieve modellen passen kennis van het bedrijf, markt, product en klant toe om een oordeel te vormen over een voorspelling. Deze modellen maken bijna geen gebruik van historische data, maar juist van enquêtes en observaties waaruit data wordt gehaald. De meest recente informatie kan zo verkregen worden (Reid et al., 2011). Dit is te vergelijken met het voorspellen van een voetbalwedstrijd. Eigen kennis wordt gebruikt om te voorspellen welke club er zal winnen. Echter worden deze voorspellingen gemaakt door mensen zelf en zijn ze vaak bevooroordeeld.

Een aantal veel gebruikte methoden worden in deze paragraaf besproken, te beginnen bij de **Delphi-methode**. Deze methode heeft als doel het bereiken van een consensus tussen een groep deskundigen waarbij hun anonimiteit behouden wordt. Er wordt een panel samengesteld op een bepaald vakgebied. Het proces omvat het verzenden van een vragenlijst aan de panelleden, het samenvatten van de bevindingen en het verzenden van een bijgewerkte vragenlijst waarin de bevindingen zijn verwerkt. Dit proces gaat zo door totdat er een consensus is bereikt, waardoor het een tijdrovende methode is (Reid et al., 2011).

Een andere bekende methode is **Markt Onderzoek** waarbij enquêtes en interviews worden gebruikt om wensen, voorkeuren, meningen en ideeën voor nieuwe producten te identificeren. Deze methode werkt goed om te weten te komen wat de klanten willen. Echter, zijn er ook een aantal tekortkomingen. Een van de meest voorkomende heeft te maken met hoe de enquêtevragen zijn opgesteld. Er zijn meerdere eisen waaraan interview vragen moeten voldoen en het is lastig om een goede enquête te ontwikkelen. Zo moet het mogelijk zijn om tussen verschillende antwoorden te kiezen. Als er maar een beperkt aantal antwoorden zijn en het gewenste antwoord staat er niet tussen, kan dit tot verkeerde interpretaties leiden (Reid et al., 2011).

Tot slot is er de veel gebruikte methode **Executieve Opinion** (bestuur oordeel). Voor deze methode wordt een groep managers gevormd die gezamenlijk een voorspelling ontwikkelen. Vaak wordt deze methode gebruikt voor strategische voorspellingen of het voorspellen van het succes van een nieuw product of een nieuwe dienst. Hoewel managers goede inzichten in de voorspellingen kunnen brengen, heeft deze methode een groot nadeel. Vaak kan de mening van één persoon de voorspelling domineren als die persoon meer macht heeft dan de andere leden van de groep.

3.3.2 Kwantitatieve voorspelmodellen

In tegenstelling tot kwalitatieve modellen, zijn kwantitatieve modellen gebaseerd op wiskundig onderzoek en daarom zijn deze voorspellingen consistent. Een model zal altijd exact dezelfde voorspelling geven uit dezelfde set gegevens. Deze modellen zijn dus objectief en kunnen daarnaast ook in één keer veel informatie verwerken. Echter, moet deze grote hoeveelheid data wel tot je beschikking staan om dergelijke voorspellingen te kunnen doen. De voorspellingen zijn even goed als de beschikbare historische gegevens. Dat wil zeggen als er een dataset ter beschikking is van één maand en de voorspelling wordt gedaan voor een heel jaar, dan kan van tevoren al worden bedacht dat de voorspelling flink zal afwijken (Reid et al., 2011). Binnen de groep van kwantitatieve voorspelmodellen is er nog een onderscheid te maken tussen *Tijdreeksmodellen* en *causale modellen.*

Tijdreeksmodellen

Tijdreeksmodellen gaan ervanuit dat tijdsreeksen van gegevens alle informatie bevatten die nodig zijn om een voorspelling te genereren. Een tijdreeks is een reeks waarnemingen die met regelmatige tussenpozen gedurende een bepaalde periode worden gedaan. Tijdreeksmodellen gaan ervan uit dat er een voorspelling mee gemaakt kan worden op basis van patronen in gegegevens (Reid et al., 2011).

De eenvoudigste manier om patronen te identificeren is om de gegevens te plotten en de resulterende grafieken te bekijken. Er zijn vier basispatronen die aanwezig kunnen zijn in een tijdreeks van gegevens en die informatie kan gebruikt worden om voorspellingen te genereren:

- Horizontaal niveau: Patroon waarin waarden schommelen rond een constant gemiddelde.
- Trends: Patroon waarin gegevens in de loop van de tijd toenemende of afnemenden waarden vertonen.
- Seizoensgebondenheid: Elk patroon dat zich regelmatig herhaalt en een constante lengte heeft.
- Cycli: Patronen gecreëerd door economische schommelingen.

Naast deze patronen die herkend kunnen worden, kan er ook ruis aanwezig zijn, die niet voorspeld kan worden. Tijsreeksen bestaan dus uit de vier genoemde patronen plus ruis. Sommige gegevens hebben veel ruis waardoor het moeilijker is om nauwkeurig te voorspellen. Veel voorspelmodellen proberen daarom zoveel mogelijk de ruis te elimineren.

$$Data = niveau + trends + \ seizoensgebondenheid + cycli + ruis$$

Verschillende tijdreeksmodellen worden toegelicht. Waarvoor kunnen ze gebruikt worden, hoe kunnen ze gebruikt worden en waarom moet er wel of niet voor een bepaald model gekozen worden?

De Bewegende Gemiddelde Techniek (Moving Average Method) is een simpele methode die snel uitgevoerd kan worden. Bij deze methode wordt het gemiddelde genomen over n, het aantal waarnemingen, van de meeste recente periodes. Naarmate er nieuwe gegevens beschikbaar komen, worden de oudste gegevens verwijderd. Het aantal waarnemingen dat is gebruikt om het gemiddelde te berekenen, wordt constant gehouden. Op deze manier "beweegt" het gemiddelde door de tijd. De formule is als volgt:

$$F_{t+1} = \frac{\sum A_t}{\mathrm{n}} = \frac{A_t + \ A_{t-1} + \cdots + A_{t-n}}{n}$$

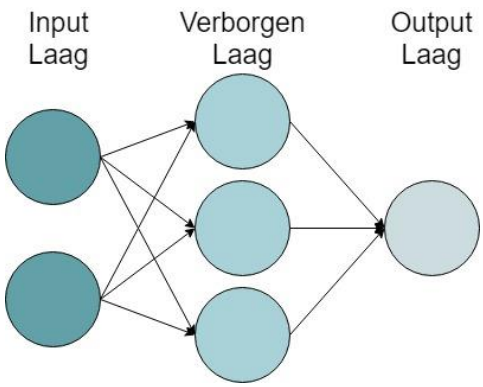
Waarbij F_{t+1} = voorspelling van de vraag voor de volgende periode $t + 1$
A_t = werkelijke waarde voor de huidige periode t
n = aantal periodes of gegevenspunten gebruikt voor het gemiddelde

Deze methode is alleen goed te gebruiken voor gegevens die schommelen rond een constant gemiddelde, een horizontaal niveau. Als de methode wordt toepast op gegevens met een trend, dan ontstaan er achterblijvende voorspellingen. Verder geldt hoe kleiner het aantal waarnemingen, des te beter de prognose reageert op de verandering in de vraag. De voorspelling is echter ook meer onderhevig aan de random veranderingen in de gegevens. Als de gegevens veel ruis bevatten, zou een hoge responsiviteit tot grotere fouten kunnen leiden. Aan de andere kant, hoe groter het aantal waarnemingen, des te minder de voorspelling reageert op de verandering in de vraag, maar juist wel minder ruis.

Een andere methode die hier verder op voort borduurd it de methode **ARIMA**. Dit is een statistische methode die gebruik maakt van tijdreeksgegevens om de toekomst te voorspellen. Een ARIMA-model combineert drie processen: autoregressief, geïntegreeerd delen van de reeks en het voortschrijnend gemiddelde van de gegevensset. ARIMA correleert in wezen automatisch zijn eigen eerdere afwijkingen van het gemiddelde, waardoor het belang wordt gehecht aan het tijdreeksgedeelte van de gegevens. Het zorgt voor trends, seizoensgebondenheid, cycli, errors en niet-stationaire aspecten van een gegevensverzemeling bij het maken van voorspellingen. Een belangrijke overweging voor ARIMA is dat de dataset ten minste 36-40 historische gegevenspunten moet hebben met minimale uitschieters. Daarnaast is het opstellen van een ARIMA model vaak een dure en tijdrovende zaak voordat er een nauwkeurig model tot uitwerking komt.

Op het eerste zicht lijkt deze methode goed aan te sluiten bij dit onderzoek, toch zijn er wat minpunten. Zo is er al een onderzoek gedaan over de bezetting van parkeergarages waarin de methode ARIMA wordt vergeleken ten opzichte van regressie modellen. Uit dit onderzoek van Chen (2014) blijkt dat het ARIMA model het slechtste presteert met de grootste training én test error. Dit komt doordat het model zeer goed presteert wanneer er weinig uitschieters zijn. Wanneer er dus één parkgeergarge zou worden gebruikt in dit onderzoek, zou het een goed model zijn. Door de gegevens van meerdere parkeergarages in een dataset te gebruiken, zal dit model dus niet goed kunnen omgaan met de variatie en een hoge error geven.

Een andere veel voorkomende voorspellingsmethode is het **Kunstmatig Neuraal Netwerk**, oftewel Artificial Neural Network (ANN). Een neuraal netwerk is gebaseerd op biologische neurale netwerken en is een niet-linear model. Het is altijd opgebouwd uit een input laag, een of meerdere verborgen lagen en de output laag, zie Figuur 6. In elke laag zijn 'neuronen' aanwezig die worden aangestuurd, die elk een andere input krijgen. Deze neuronen geven zo weer hun eigen signaal af aan de volgende laag met bijbehorende neuronen. Elke neuron heeft zijn eigen kenmerken en daarom ook zijn eigen gewicht. Dit gewicht bepaalt hoe sterk het signaal via die connectie aankomt bij de volgende laag en dit signaal kan positief en negatief zijn.



FIGUUR 6 - NEURAAL NETWERK

ANN werkt efficient met gegevens die ruis bevatten of lage correlaties en is daarnaast ook flexibel bij het verwerken van multidimensionale gegevens. Het vermogen om te generaliseren en te leren maakt dit model effectief in de ontwikkling van voorspellingsmodellen en daarnaast ook zeer geschikt voor het voorspellen van gebeurtenissen waar weinig bekend is over de relaties tussen de verschillende variabelen (Karlaftis & Vlahogianni, 2011). De verborgen lagen stellen de voorspellingsmodellen in staat om niet-lineaire relaties en interacties tussen variabelen te beschouwen (Pflüger, Köhn, Schreieck, Wiesche, & Kremar, 2016).

Een probleem van ANN modellen is echter hun "Black Box" concept. In andere methoden is het mogelijk om het effect en de invloed van elke variabele te vinden, terwijl het in ANN modellen moeilijk of helemaal niet mogelijk is om dit soort relaties te vinden (Kumar, Parida, & Katiyar, 2013). Ook is het niet haalbaar met ANN om een hypothesetest uit te voren tussen de verklarende varibalen en het beoogde resultaat. Daarnaast is het ook nog eens zo dat deze modellen computationeel intensief zijn. Dus voor deze modellen geldt dat een grotere dataset zorgt voor een hogere betrouwbaarheid, maar dat daardoor de rekentijd aanzienlijk lang wordt. Zo is het voor real-time voorspellingen vaak niet acceptabel om een ANN model te gebruiken vanwege de rekentijd.

Causale modellen

Naast tijdreeskmodellen zijn er ook causale modellen die een andere logica gebruiken om een voorspelling te genereren. Ze veronderstellen dat de variable die voorspeld moet worden op de een of andere manier gerelateerd is aan andere variabelen in de omgeving. De taak van de forecaster is om te ondekken hoe deze variabelen gerelateerd zijn in wiskundige termen en deze informatie te gebruiken om de toekomst te voorspellen. Uit historische gegevens kan een model gebouwd worden dat de relatie van de variablen verklaart en deze relaties gebruikt om te voorspellen (Reid et al., 2011). Causale modellen kunnen erg complex zijn, vooral als er rekening wordt gehouden met de relaties tussen de variabelen. Ze zijn makkelijk te gebruiken en kunnen sneller een voorspelling genereren dan tijdreeksmodellen, waarvoor modelbouw vereist is (Reid et al., 2011).

Naast het niet-lineaire model ANN, zijn er ook lineaire modellen te gebruiken, zoals **lineaire regressie**. Hierbij is de voorspelling van de afhankelijke variabele gerelateerd aan de onafhankelijke variabele op een lineaire manier. De relatie tussen twee variabelen is de vergelijking van een rechte lijn. Er kunnen vaak meerdere rechte lijnen door gegevens worden getrokken. Lineaire regressie selecteert de parameters a en b, die een rechte lijn definiëren, door de som van de gekwadrateerde fouten of afwijkingen van de lijn te minimaliseren. Het berekenen van de waarden voor a en b kan ingewikkeld zijn. De stappen voor het berekenen van de lineaire regressievergelijkingen zijn als volgt:

Stap 1: Bereken parameter b

$$b = \frac{\sum XY - \mathrm{n}\overline{XY}}{\sum X^2 - \mathrm{n}\overline{X^2}}$$

Waarbij $\bar{\bar{Y}}$ = gemiddelde van de Y waarden
$\bar{\bar{X}}$ = gemiddelde van de X waarden
n = aantal data punten

Stap 2: Bereken parameter a

$$a = \bar{Y} - b\bar{X}$$

Stap 3: Substitueer de waarden a en b om de lineaire regressievergelijking te vormen

$$Y = a + bX$$

Stap 4: Voor het maken van een voorspelling voor de afhankelijke variabele (Y), vervang de juiste waarde voor de onafhankelijke variabele (X).

Bij het uitvoeren van lineaire regressie is het nuttig om de **correlatie coëfficiënt** te berekenen, die de richting en sterkte van de lineaire relatie tussen de onafhankelijke en afhankelijke variabelen meet.

$$R = \frac{\mathrm{n}\sum(XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2]} * \sqrt{[n(\sum Y^2) - (\sum Y)^2]}}$$

De waarde van R kan varieren tussen -1 en +1 en heeft de volgende betekenis:

- $R = +1$: Er is een perfece positive lineaire relatie tussen de twee variablen. Voor elke toename van 1 eenheid van de onafhankelijke variable, is er een toename van 1 eenheid van de afhankelijke variabele.
- $R = -1$: Er is een perfecte negatieve lineaire relatie tussen de twee variabelen. Ondanks dat de relatie negatief is, bewegen de twee variabele in tegengestelde richting. Een toename van 1 eenheid van de onafhankelijke variabele gaat gepaard met een afname van 1 eenheid van de afhankelijke veriabele.
- $R = 0$: Er is geen lineair verband tussen de variabelen.

Hoe dichter de waarde bij -1 of +1 ligt hoe sterker de lineaire relatie tussen de variabelen is. Daarnaast geeft R^2 aan hoe goed de regressielijn past bij de gegevens. Hoe hoger R^2, des te beter.

Het is ook mogelijk dat er meerdere onafhankelijke variabele een relatie hebben met de afhankelijke variabelen. **Veelvoudige Regressie** is een krachtig hulpmiddel voor voorspellingen en moet worden gebuikt wanneer meerdere factoren de variabele beïnvloeden die wordt voorspeld. Veelvoudige Regressie verhoogt echter significant de benodigde gegevens en software vereisten voor voorspellingen. De formule voor veelvoudige Regressie is als volgt:

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_kX_k$$

Waarbij
Y = afhankelijke variable
B_0 = het Y-snijpunt
$B_1 \ldots B_k$ = coëfficiënten die de invloed van de onafhankelijke variabelen op de afhankelijke variabelen weergeven
$X_1 \ldots X_k$ = onafhankelijke variabelen

3.4 Vergelijking tussen de voorspellingsmethoden

Verschillende methoden om voorspellingen te generen zijn in de vorige paragraaf aan bod gekomen. Het blijft lastig om te besluiten welke methode het beste aansluit bij de probleemstelling van dit huidige onderzoek. Daarom is er in de literatuur gezocht naar onderzoeken met vergelijkbare cases waar deze methoden worden gebruikt. Er zijn een aantal onderzoeken geweest die meerdere methoden onderling hebben vergeleken. Natuurlijk blijft elke case altijd anders, maar het is een goede indicatie om te zien welke methode goed scoort.

Zo is er in twee vergelijkbare onderzoeken van Zeng et al. (2015) en van Shinde, Bhagwat, Pharate & Paymode (2016) gekeken naar welke methode het beste presteert voor het bepalen van de beschikbaarheid van parkeerplaatsen. De methoden regressieboom, support vector regressie en neuraal netwerk zijn met elkaar vergeleken. Hierin scoorde in beide onderzoeken het neurale netwerk het slechtste. Dit wordt verklaardt door het feit dat een neuraal netwerk de interne correlaties van de onafhankelijke variabelen meeneemt. Tussen de tijd van de dag en de dag van de week zal echter geen sterke correlatie aanwezig zijn. De prestaties van de regressieboom zijn duidelijk beter dan de andere twee vergeleken methoden op basis van alle criteria.

Kunjithapatham et al. (z.d.) heeft ook onderzoek verricht naar de mogelijkheid om de parkeerbeschikbaarheid in San Francisco Area te voorspellen. Het doel van zijn onderzoek is om het verkeer te verminderen door bestuurders te helpen bij het zoeken naar de best mogelijke lege parkeerplekken. De methoden beslissingsboom en lineaire regressie zijn in dit onderzoek met elkaar vergeleken. Het resultaat van het model op basis van de beslissingsboom had een foutmarge van 5% in tegenstelling tot de lineare regressie methode, die een foutmarge had van 23%. Kunjithapatham schrijft wel dat er meer onderzoek nodig is om vast te stellen of dit resultaat zomaar kan worden overgenomen voor andere steden.

Verder heeft ook Chen (2014) in zijn onderzoek naar parkeerbezettingen methoden met elkaar vergeleken. De methoden die in zijn onderzoek naar voren kwamen zijn: ARIMA, lineare regressie, support vector regressie en ANN. In zijn onderzoek komt het model van het neurale netwerk als het beste naarvoren. Een belangrijk verschil met dit huidige onderzoek is dat er onderzoek is gedaan naar parkeerplaatsen langs de straat, die allemaal gelijkwaardig aan elkaar zijn in tegenstelling tot verschillende parkeergarages in verschillende steden. Daarnaast komt het ARIMA model het slechtste naar voren, in dit onderzoek komt het vooral doordat er niet al te veel datapunten zijn gebruikt. Voor het ARIMA model geldt namelijk des te meer datapunten, des te beter de voorspelling zal zijn.

In een onderzoek van Soler (2015) worden enkele voorspellingsmethoden naast elkaar geleged en wordt er gekeken of deze methoden wel of niet aansluiten bij het voorspellen van parkeerbezettingen. Zo wordt de methode van neurale netwerken door Soler (1015) afgewezen, omdat er meerdere nadelen aan deze methoden zitten. Ten eerste is het een kennisnetwerk in de vorm van gewichten, waardoor het model lastiger is om te begrijpen. Het model kent intern gewichten toe aan de klassen op basis van kennis uit de train dataset, maar deze kennis wordt niet naar buiten gebracht door het model. Er is dus geen duidelijk beeld van de beschrijvende variabelen en de invloed daarvan op de voorspellende waarden. Ten tweede is het moeilijk om je eigen kennis in het netwerk toe te voegen, omdat het model vooral uitgaat van de train dataset. Dus als er al kennis aanwezig is over enkele onafhankelijke variabelen is het bijna onmogelijk om deze kennis in het model te stoppen.

3.5 Conclusie

In de literatuur is gekeken naar de verschillende variabelen die relevant zijn voor het voorspellen van de bezettingsgraad van parkeergarages. Verder is er gekeken hoe deze variabelen een dataset vormen waar analyses mee gedaan kunnen worden. Daarnaast zijn er ook methoden die worden ontwikkeld om deze bezettingsgraden te kunnen voorspellen. Kortom, er is antwoord gegeven op de volgende vragen die horen bij dit hoofdstuk:

- Wat is een potentiele methode om voorspellingen te kunnen doen op basis van historische data?
    - Welke variabelen komen in de literatuur voor die relevant zijn en bijdragen aan de voorspelling?
    - Welke methoden zijn er in de literatuur bekend om patronen en trends in data te kunnen analyseren?
    - Welke methoden zijn er in de literatuur bekend om een vergelijkbaar voorspelprobleem op te lossen?

Er zijn vijf onafhankelijke variabelen die als belangrijkste naar voren zijn gekomen voor het voorspellen van de bezettingsgraad van parkeergarages. De variabelen waar het om gaat zijn: t*ijdstip van de dag, dag van de week, vakantie, weer* en *speciale gebeurtenis.* Daarnaast zijn er nog twee andere variabelen die in dit onderzoek ook van belang zijn, omdat er meerdere parkeergarages worden

meegenomen. Daarom spelen *capaciteit* en *locatie* ook een belangrijke rol. De capaciteit is belangrijk omdat er natuurlijk niet meer auto's geparkeerd kunnen worden dan de capaciteit. Daarnaast is de locatie belangrijk, omdat er verwacht kan worden dat de parkeergarage die het dichts bij het centrum ligt, sneller vol zal lopen dan een parkeergarage aan de rand van een stad.

Alle informatie van deze variabelen zorgen samen voor de dataset die uiteindelijk gebruikt wordt voor het voorspellen van de bezettingsgraad. Eerst zal deze informatie één geheel moeten vormen en daarvoor opgeschoond en getransformeerd worden. Voordat er direct aan de hand van deze dataset voorspellingen worden gedaan, is het goed om eerst zelf analyses te doen met de data. Komen er patronen uit die logisch te verklaren zijn? Zijn er verschillende parkeergarages die dezelfde patronen weergeven? Geven alle afhankelijke variabelen logische reacties? Dit zijn een paar vragen die gesteld kunnen worden over de dataset, zonder direct voorspellingen te maken.

Na deze analyses moeten er voorspellingen worden gegenereerd aan de hand van de methode die het beste hierbij past. De methoden worden verdeeld in kwalitatieve en kwantitatieve voorspelmodellen. In dit onderzoek gaat het om wiskundige modellen en niet over enquêtes of interviews en daarom ligt de focus op kwantitatieve modellen. De methoden die in deze categorie in de literatuur naar voren zijn komen zijn als volgt:

- Beslissingsboom / Regressieboom
- Bewegend Gemiddelde
- ARIMA
- Kunstmatig Neuraal Netwerk
- Lineaire Regressie / Veelvoudige Regressie

Uit verschillende onderzoeken is er één methode die duidelijk als beste naar voren komt om de bezetting van parkeergarages te voorspellen. De beslissingsboom/regressieboom komt in meerdere onderzoeken als beste naar voren en is daarom het meest geschikt voor dit onderzoek om mee door te gaan. Neurale netwerken hebben zo hun eigen kanttekening en worden in meerdere onderzoeken afgeschreven. ARIMA is een model wat ingewikkeld is en daarom ook een wisselend resultaat levert. Daarnaast is lineaire regressie een makkelijk model dat eventueel nog kan worden onderzocht, maar het komt niet goed naar voren uit vergelijkbare onderzoeken.

Hoofdstuk 4 – Data analyse

Nu het duidelijk is welke methode gebruikt zal worden tijdens dit onderzoek, is de volgende stap het compleet maken van de dataset. Deze dataset zal gebruikt worden voor het genereren van de voorspelling en zal daarom alle variabelen moeten bevatten die eerder benoemd zijn. Als deze dataset helemaal compleet is, kunnen er analyses gedaan worden. De vraag voor dit hoofdstuk is daarom: "*Welke verklaringen zijn er voor de patronen in de dataset?*". Om deze vragen te beantwoorden zijn ook in dit hoofdstuk een aantal sub vragen opgesteld:

- Welke verschillen zijn er tussen de parkeergarages?
- Wat zijn de beperkingen van de dataset?
- Hoe passen alle variabelen bij elkaar in één dataset?
- Welke patronen zijn er in de dataset aanwezig?

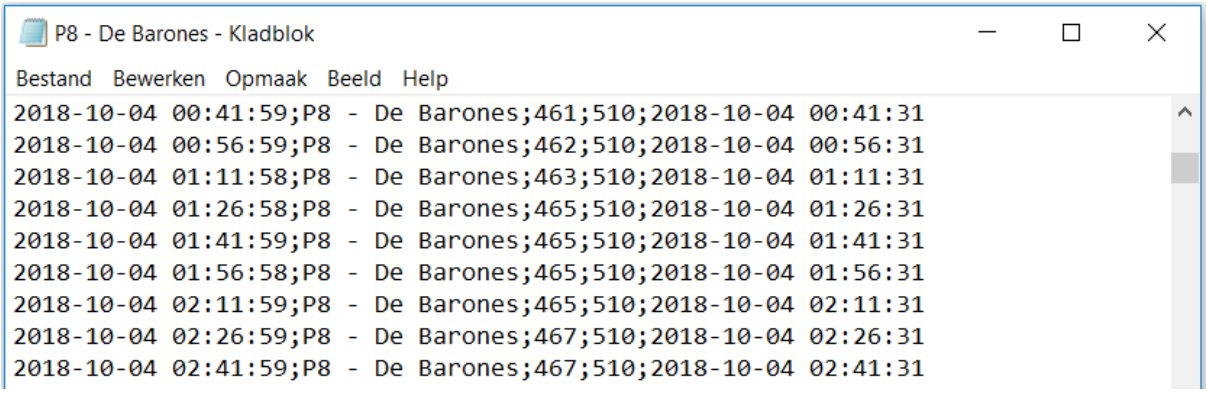
De eerste paragraaf begint met een toelichting over hoe de gegevens binnenkomen en opgeslagen worden (§4.1). Vervolgens wordt er globaal gekeken naar alle parkeergarages en hoe deze verschillen van elkaar (§4.2). Ook worden er in paragraaf 4.2 keuzes gemaakt over de parkeergarages die in het verder onderzoek worden meegenomen. De beperkingen van de dataset die aanwezig zijn worden besproken in paragraaf 4.3. Daarna wordt er toegelicht hoe alle variabelen worden toegevoegd en wat voor impact deze variabelen hebben op de bezetting van de parkeergarages (§4.4). De patronen die in de dataset aanwezig zijn worden toegelicht en verklaard (§4.5). Aan het eind van het hoofdstuk zal een conclusie volgen die alle vragen beantwoord (§4.6).

4.1 De historische gegevens

Zoals eerder besproken in het literatuuronderzoek is de eerste stap het analyseren van de gegevens, voordat er voorspellingen gegenereerd kunnen worden. Het gaat hier om de historische gegevens van 128 parkeergarages in Nederland. Binnen het bedrijf is er de mogelijkheid om deze gegevens te verzamelen op elk gewenst moment. Wanneer dit systeem wordt geactiveerd, worden elk kwartier nieuwe gegevens opgeslagen met daarin het huidige aantal vrije parkeerplaatsen in de desbetreffende parkeergarages. Er zijn vier momenten geweest tijdens mijn onderzoek dat het systeem geactiveerd was en daardoor zijn er vier afzonderlijke dataset ontwikkeld:

- Dataset 1: 4 oktober t/m 28 oktober
- Dataset 2: 21 november t/m 5 december
- Dataset 3: 20 december t/m 27 december
- Dataset 4: 5 januari t/m 8 januari

De onbewerkte gegevens die binnen komen uit het systeem worden opgeslagen als tekstdocumenten. Elke parkeergarage heeft zijn eigen document, hiervan is een voorbeeld te zien in Figuur 7.



FIGUUR 7 - HISTORISCHE GEGEVENS

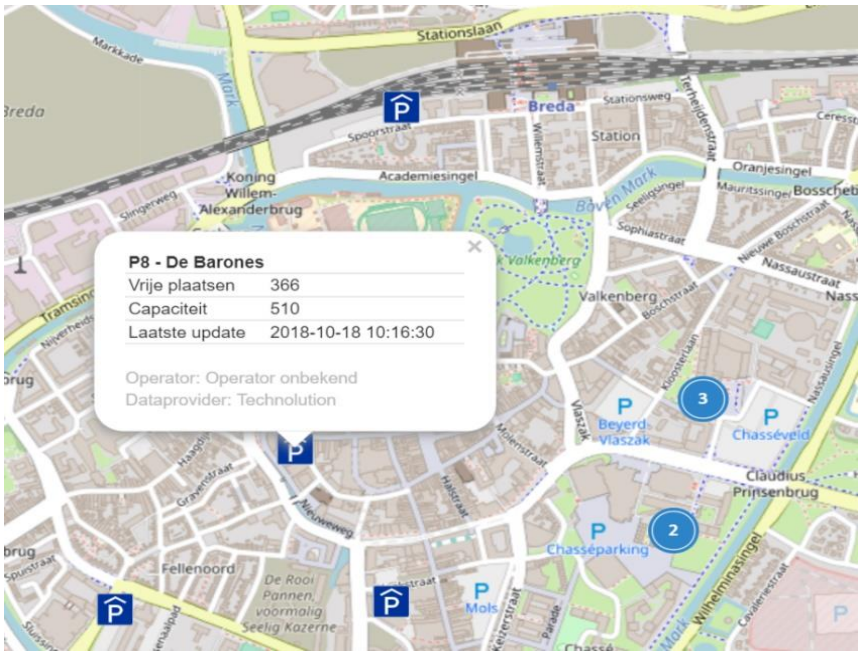
Het bestand heeft geen header, dus het is niet meteen te zien hoeveel variabelen er beschreven worden. Elke regel in het bestand geeft de informatie van een bepaald moment in de tijd met de bijbehorende variabelen, die worden gescheiden door een puntkomma. Eerst wordt de datum en tijd weergegeven, gevolgd door de naam van de parkeergarage. Daarna het aantal vrije parkeerplaatsen, de capaciteit en als laatste nog een keer de datum en tijd. De tijdnotatie die wordt gehanteerd is uu:mm:ss, oftewel uur, minuten en seconden. In de eerste regel uit dit voorbeeld zijn er dus in de parkeergarage De Barones 461 parkeerplaatsen vrij van de 510 om bijna 42 over 12 in de nacht van 4 oktober 2018. In Tabel 2 zijn deze vijf variabelen uit de originele dataset weergegeven. De reden voor het feit dat de datum en tijd twee keer worden genoteerd, is dat de eerste notatie het moment van de meeting weergeeft en de laatste fungeert als check. Dit is iets wat zo in het systeem verwerkt zit en verder niet als belangrijk wordt geacht in de verdere analyse van de dataset.

Variabele	Beschrijving
Datum	De datum van de meeting weergegeven met de notatie: dd-mm-jj.
Tijd	Het uur van de dag weergegeven met de notatie: uu:mm:ss.
Locatie	De naam van de parkeergarages met soms de plaatsnaam erbij.
Vrije plaatsen	Dit getal geeft aan hoeveel vrije plekken er nog zijn in de parkeergarage.
Capaciteit	Dit getal geeft aan hoeveel plekken er in totaal zijn in de parkeergarage.

TABEL 2 - VARIABELEN UIT DE HISTORISCHE DATA

4.2 Het uitzoeken van de gegevens

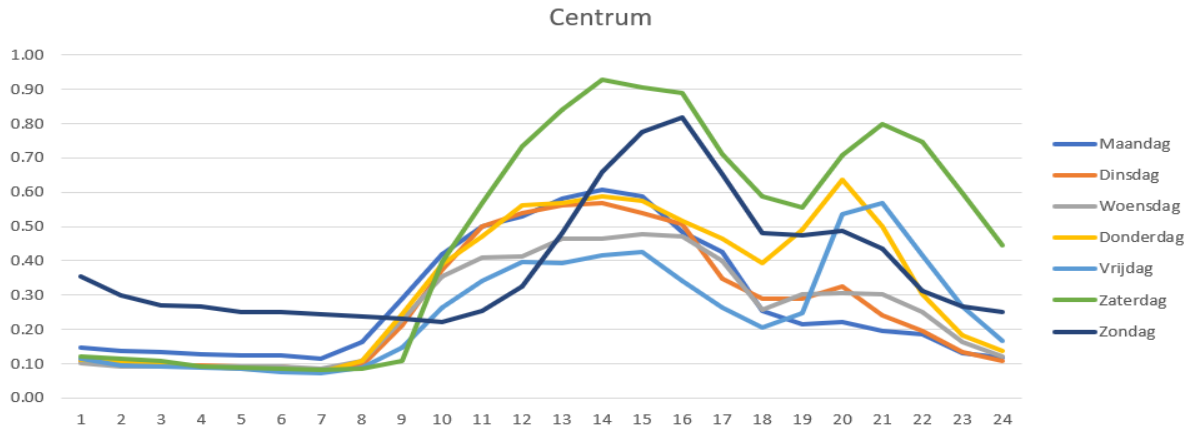
De gegevens worden binnen Goudappel Groep verzameld van 128 verschillende parkeergarages. De lijst van parkeergarages met hun bijbehorende capaciteit en locatie staan in Appendix B. In het systeem worden niet alle gegevens helemaal nauwkeurig verwerkt, hierdoor komt het voor dat er van een aantal garages geen gegevens worden genoteerd. Deze garages worden verder niet meegenomen in het onderzoek, omdat er geen historische gegevens beschikbaar zijn. Er is er ook een webpagina met de kaart van Nederland, waarop alle locaties van de parkeergarages zijn weergegeven. Dit geeft een mooi overzicht en maakt het makkelijk om de verschillende locaties van de garages te onderscheiden. In Figuur 8 is een gedeelte van de kaart te zien met enkele parkeergarages in Breda.



FIGUUR 8 - KAART VAN BREDA

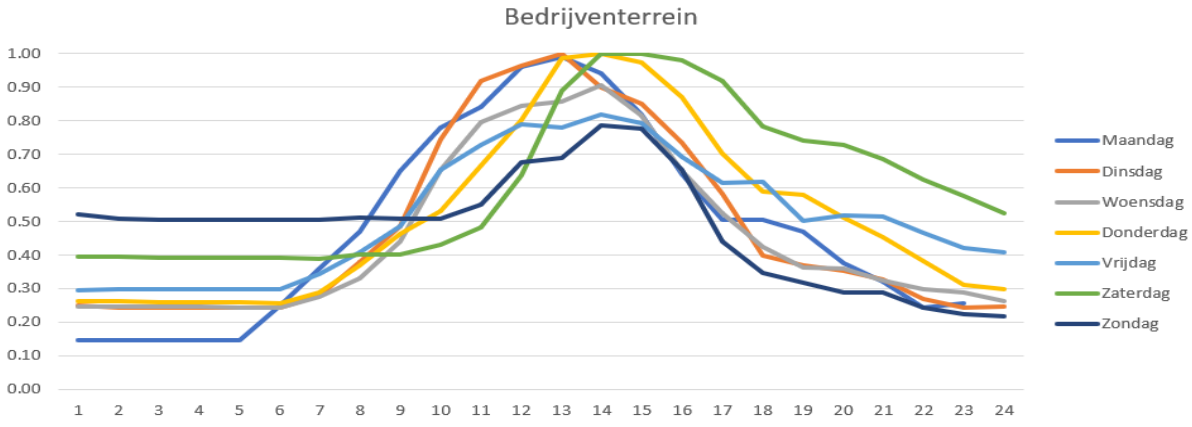
In dit onderzoek is er een duidelijk verschil tussen het voorspellen van de bezettingsgraad van parkeergarages zonder gedragsfactoren en mét gedragsfactoren. Van de Zande (2013) heeft haar onderzoek gericht op deze gedragsfactoren en daarmee onderzocht door welke factoren automobilisten gestuurd worden. Hier gaat het om factoren zoals prijs en afstand tot het centrum. In dit onderzoek ligt daarom de focus op het vergelijken van parkeergarges in verschillende steden, zodat de keuze van de automobilist hier geen grote rol in speelt.

Voor het onderzoek kunnen niet alle parkeergarages worden meegenomen. In Appendix B is een overzicht van alle parkeergarages in Nederland waarvan de gegevens worden opgeslagen. Er zijn 24 parkeergarages waarvan de gegevens niet of niet correct zijn opgeslagen, dus deze garages kunnen sowieso niet worden gebruikt. Daarnaast zijn er twee grote verschillen te onderscheiden in de locatie van de parkeergarages. Natuurlijk zijn ze in verschillende steden gevestigd, maar het belangrijkste is de precieze locatie in de stad. Is het juist in het centrum omringd door winkels en andere recreatie doeleinden of is het meer een locatie in de buurt van scholen en bedrijven? Twee parkeergarages zijn uit de dataset gehaald en zijn te zien in Figuur 9 en Figuur 10 om de verschillen duidelijk weer te geven.



FIGUUR 9 - BEZETTINGSGRAAD CENTRUM

Er zijn verschillende redenen die de pieken in Figuur 9 verklaren. De bezetting is in het centrum op een zaterdag en zondag duidelijk hoger dan de doordeweekse dagen. Dit is natuurlijk het weekend waarop mensen vaak vrij zijn en gaan winkelen. Het valt ook op dat deze twee lijnen snel stijgen naar hun piek. Op zaterdag is dit eerder dan op zondag, omdat de winkels op zondag later open gaan. Als tweede is er ook te zien dat er nog een tweede piek is op donderdag, vrijdag en zaterdag. Dit zijn dagen die populair zijn om bijvoorbeeld uit eten te gaan, maar ook vaak koopavond op een donderdag. Tot slot is het hoge begin van de zondag te verklaren door het uitgaansleven op zaterdagavond.

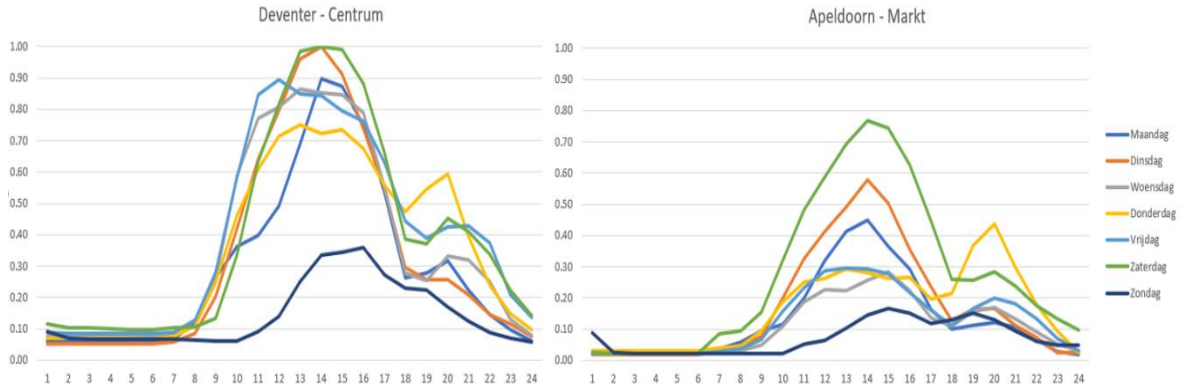


FIGUUR 10 - BEZETTINGSGRAAD BEDRIJVENTERREIN

Het bedrijventerrein in Figuur 10 laat een ander patroon zien. Het eerste dat opvalt is dat er maar één duidelijke piek is rond de middag. Ook hier zijn er nog een aantal andere zaken die opvallen en te verklaren zijn. Zo is de piek van woensdag en vrijdag lager dan de andere werkdagen. Dit zijn dagen waarop er gemiddeld minder mensen werken, omdat dit pupulaire dagen zijn om thuis te werken of voor part-time mensen om hun vrije dag in te plannen. Verder is ook te zien dat op zondag er wel iets gebruik wordt gemaakt van de parkeergarage maar duidelijk minder dan de andere dagen. De piek op zaterdag is te verklaren door omliggende recreatiemogelijkheden, zoals sportverenigingen.

In de lijst met alle parkeergarages uit Appendix B is er onderscheid gemaakt tussen locaties in het centrum, station, buitenwijk of parkeergarages naast snelwegen. De focus van het onderzoek ligt op de ontwikkeling van een voorspellingsmodel en niet op bepaalde parkeergarages. Er is gekozen om parkeergarages in het centrum van steden te gebruiken voor het onderzoek naar de mogelijkheden voor voorspellingen. Na het selecteren van alleen centrum garages blijft er een selectie over van 51 parkeergarages met wisselende capaciteit. Van deze overgebleven parkeergarages blijven er 47 over waarvan alle gegevens compleet zijn en die mogelijk gebruikt kunnen worden voor het voorspellen van de parkeergarages.

Door het plotten van de afzonderlijke datasets van de parkeergarages, die zijn overgebleven uit de vorige analyse, zijn nog steeds duidelijke verschillen te zien. In totaal zijn er 12 parkeergarages uit het overzicht genomen, waarvan de locaties allemaal in het centrum zijn en dus worden gebruikt voor het ontwikkelen van een voorspellingsmodel. In Figuur 11 zijn twee grafieken te zien van de parkeergarages in het centrum. Op het eerste gezicht lijken de grafieken niet direct op elkaar, maar als de kleuren afzonderlijk vergeleken worden zijn er ook duidelijke overeenkomsten. De groene piek die verwijst naar zaterdag is bij beide parkeergarages het hoogste, gevolgd door de piek van dinsdag. De gele lijn in de grafiek heeft duidelijk een tweede piek die ook in beide grafieken het hoogste is. Tot slot is het goed te zien dat de donker blauwe lijn van zondag bij beide grafieken een minimale piek heeft als gevolg dat de winkels die week gesloten waren. De overige grafieken van de andere 10 parkeergarages zijn te vinden in Appendix C.



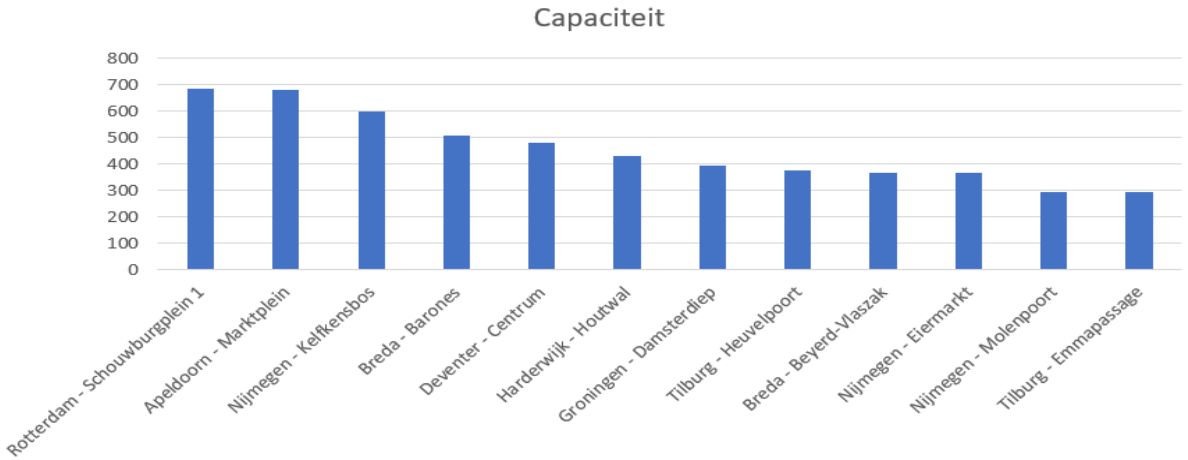
FIGUUR 11 - DE BEZETTINGSGRAAD ÉÉN WEEK

De 12 parkeergarages die verder worden meegenomen in het onderzoek kunnen op meerdere manieren vergeleken worden. Zo is in Tabel 3 de bezettingsgraad per maand te zien. De bezettingsgraad is een percentage ten opzichte van de capaciteit van de desbetreffende parkeergarage. Wat opvalt is dat de gemiddelde bezettingsgraad significant laag is, wat te verklaren valt doordat de garages in de nacht niet tot nauwelijks bezet zijn. Daarnaast is het ook te zien dat de bezettingsgraad per parkeergarage in de verschillende maanden relatief dicht bij elkaar liggen. Dit is voor het voorspellen van de bezettingsgraad een belangrijk punt, omdat de maanden dus ook met elkaar te vergelijken vallen.

Onderzoek	Capaciteit	Bezettings graad oktober	Bezettings graad november	Bezettings graad december	Bezettings graad Januari	Gemiddelde bezetting
Deventer Centrum	481	0.41	0.43	0.37	0.30	172
Apeldoorn Marktplein	680	0.21	0.22	0.25	0.12	134
Harderwijk Houtwal	432	0.28	0.30	0.33	0.29	129
Nijmegen Eiermarkt	365	0.28	0.39	0.31	0.30	125
Nijmegen Kelfkensbos	600	0.34	0.36	0.37	0.28	208
Nijmegen Molenpoort	295	0.28	0.21	0.28	0.16	78
Tilburg Emmapassage	295	0.25	0.32	0.32	0.21	80
Tilburg Heuvelpoort	375	0.33	0.35	0.36	0.23	124
Breda Barones	510	0.34	0.45	0.50	0.47	208
Breda Beyerd-Vlaszak	365	0.34	0.34	0.35	0.25	124
Rotterdam Schouwburg Plein 1	685	0.34	0.41	0.38	0.30	246
Groningen Damsterdiep	392	0.20	0.18	0.20	0.13	75

TABEL 3 - SPECIFICATIES VAN DE PARKEERGARAGES

Wel zijn er grote verschillen in de capaciteit van de parkeergarage. Dit is dan ook de reden dat er niet wordt gekeken naar de bezetting van een parkeergarage maar naar het percentage. Dit zegt in het algemeen meer dan een willekeurig getal. Een bezettingsgraad van 0.50 geeft dus aan dat de parkeergarage half vol is, maar met een bezetting van 50 kun je hier niks over zeggen. De capaciteit van de parkeergarages is weergegeven in Figuur 12.



FIGUUR 12 - DE CAPACITEIT VAN DE 12 PARKEERGARAGES

4.3 De beperkingen van de dataset

De informatie uit de dataset en de dataset zelf hebben ook enkele beperkingen. In het optimale geval zou de dataset uit meer historische gegevens bestaan. Het zou namelijk waardevol zijn om dezelfde dag uit verschillende jaren te kunnen vergelijken. Nu kan het zo zijn dat er speciale uitschieters zijn, die niet direct te verklaren vallen door bijvoorbeeld vakantie of een speciale gebeurtenis. Ook is het lastig te achterhalen welke evenementen er hebben plaatsgevonden en of er een directe link te leggen is tussen de bezetting en het evenement.

Daarnaast is het niet mogelijk om te achterhalen hoelang één auto in de parkeergarage aanwezig is. Daarom valt er ook niks te zeggen over de in- en uitstroom van een parkeergarage. De patronen van alle auto's afzonderlijk kunnen veel zeggen over de reden van het bezoek aan de parkeergarage. Ook kan er met de in- en uitrijtijden iets over de drukte op de omliggende wegen worden gezegd. Dit is informatie die nuttig is voor gemeenten voor verkeersanalyses, omdat ze dan precies weten waar deze auto's vandaan komen en wat voor invloed parkeergarages hebben op het wegennetwerk. Dit is helaas iets wat niet in dit onderzoek meegenomen kan worden, maar wel wat verder onderzocht zal worden binnen het bedrijf.

Verder komt het ook voor dat het systeem een tijdelijke error heeft, waardoor er sommige gegevens verloren gaan. Voor één parkeergarage is dit geen ramp, maar dit maakt het lastig voor het vergelijken van meerdere parkeergarages. Zo komt het voor dat de bezetting van een parkeergarage drie dagen lang hetzelfde is. Het is niet te achterhalen of de garage ook daadwerkelijk dicht was of dat het echt incorrecte data is en het systeem het fout heeft genoteerd. Dit zorgt natuurlijk voor gekke grafieken en vergelijkingen.

4.4 Het vormen van een dataset

Voor dit onderzoek moeten de historische gegevens getransformeerd worden in een dataset, waarin alle variabelen opgenomen zijn, die in het literatuuronderzoek uit hoofdstuk 3 naar voren zijn gekomen. De historische gegevens zijn genoteerd per kwartier, maar voor de voorspellingen worden de gegevens per uur meegenomen. Het doel van het onderzoek is namelijk het genereren van een voorspelling over een uur, een dag, een week of een maand. De kleinste tijdseenheid is daarin het belangrijkste en daarom wordt die meegenomen in de nieuwe dataset. Naast de variabelen die al aanwezig waren in de originele gegevens, zijn ook nieuwe variabelen toegevoegd. De uiteindelijke lijst van variabelen is weergegeven in Tabel 4.

Variabele	Beschrijving
Datum	De datum van de meeting weergegeven met de notatie: dd-mm-jj.
Tijd	Het uur van de dag weergegeven met de getallen 1 tot en met 24.
Type dag	Alle dagen van de week.
Locatie	De naam van de parkeergarages met soms de plaatsnaam erbij.
Vrije plaatsen	Dit getal geeft aan hoeveel vrije plekken er nog zijn in de parkeergarage.
Capaciteit	Dit getal geeft aan hoeveel plekken er in totaal zijn in de parkeergarage.
Bezetting	Dit getal geeft aan hoeveel plekken bezet zijn in de parkeergarage.
Bezettingsgraad	Dit percentage is de bezetting ten opzichte van de capaciteit.
Moment	De dag is in vier delen verdeeld: nacht, ochtend, middag en avond.
Speciale gebeurtenis	Een binaire waarde die aangeeft of er een event plaatsvond.
Temperatuur	De verwachte temperatuur van die dag.
Regen	De verwachte mate regen van die dag, verdeeld in zes schalen.
Vakantie	Een binaire waarde die aangeeft of het vakantie is of niet.

TABEL 4 - VARIABELEN IN DE DATASET

Het tekstdocument uit Figuur 7 wordt in Excel geladen, waaruit een lange tabel met vijf verschillende kolommen volgt. Een gedeelte van de bovenkant van de tabel is weergegeven in Figuur 13. De datasets van de vier verschillende periodes worden in Excel geladen en in één tabel samengevoegd. Zodat in één werkmap alle gegevens staan van een specifieke parkeergarage. Dit is het begin van de totale dataset, die het uiteindelijk moet worden met alle genoemde variabelen. Stap voor stap worden deze variabelen toegevoegd en worden de gegevens getransformeerd tot de complete dataset.

Column1	Column2	Column3	Column4	Column5
4-10-2018 00:11	P8 - De Barones	454	510	4-10-2018 00:11
4-10-2018 00:26	P8 - De Barones	460	510	4-10-2018 00:26
4-10-2018 00:41	P8 - De Barones	461	510	4-10-2018 00:41
4-10-2018 00:56	P8 - De Barones	462	510	4-10-2018 00:56
4-10-2018 01:11	P8 - De Barones	463	510	4-10-2018 01:11
4-10-2018 01:26	P8 - De Barones	465	510	4-10-2018 01:26
4-10-2018 01:41	P8 - De Barones	465	510	4-10-2018 01:41
4-10-2018 01:56	P8 - De Barones	465	510	4-10-2018 01:56
4-10-2018 02:11	P8 - De Barones	465	510	4-10-2018 02:11
4-10-2018 02:26	P8 - De Barones	467	510	4-10-2018 02:26
4-10-2018 02:41	P8 - De Barones	467	510	4-10-2018 02:41
4-10-2018 02:56	P8 - De Barones	467	510	4-10-2018 02:56

FIGUUR 13 - TEKSTDOCUMENT INGEVOERD IN EXCEL

4.4.1 Datum

Om de dataset compleet te maken, moeten alle gegevens getransformeerd worden, zodat het allemaal bij elkaar aansluit. Uiteindelijk moeten er geen lege gaten meer in de dataset aanwezig zijn. De eerste stap is het splitsen van de datum en de tijd. In de originele historische data staan deze samen in één cel genoteerd en kan het niet worden gezien als twee onafhankelijke variabelen. Daarom wordt de cel eigenschap van deze kolom aangepast naar alleen de datum met de dag-maand-jaar notatie. Verder worden ook de eerste en laatste dag verwijderd uit de dataset, omdat deze beide dagen maar voor de helft zijn waargenomen. Zo is de eerste kolom gereed en staat de datum zoals gewenst genoteerd.

4.4.2 Tijd

Voordat de tijd verwijderd wordt uit de beide kolommen, wordt er eerst naar de verschillende tijdstippen gekeken. Zoals al eerder genoemd, moet de bezetting er uur genoteerd staan. Voor het verwijderen van alle tussenliggende kwartieren is een macro geschreven die telkens drie regels verwijdert en dan een regel verder springt, zie de code in Figuur 14. De regels worden verwijderd en niet gebruikt voor een gemiddelde, omdat er anders kans is op het wegnemen van pieken. In de tabel in Figuur 13 zal dus eerst drie keer regel twee verwijderd worden, waardoor de eerste drie regels uit de tabel verdwijnen. De regel met de tijd 00:56 zal blijven staan en vervolgens zullen de volgende drie regels worden verwijderd, waarna de regel met de tijd 01:56 zal blijven staan. Dit wordt voor de dataset doorgevoerd en zo zijn in één keer alle kwartieren uit de dataset verwijderd. Tot slot wordt er een extra kolom toegevoegd, waarin de uren vermeld worden, waarin 00:56 gelijk staat aan 1, 01:56 aan 2 enzovoorts. Dit gaat respectievelijk door tot 24 en begint daarna voor de volgende dag weer opnieuw bij 1.

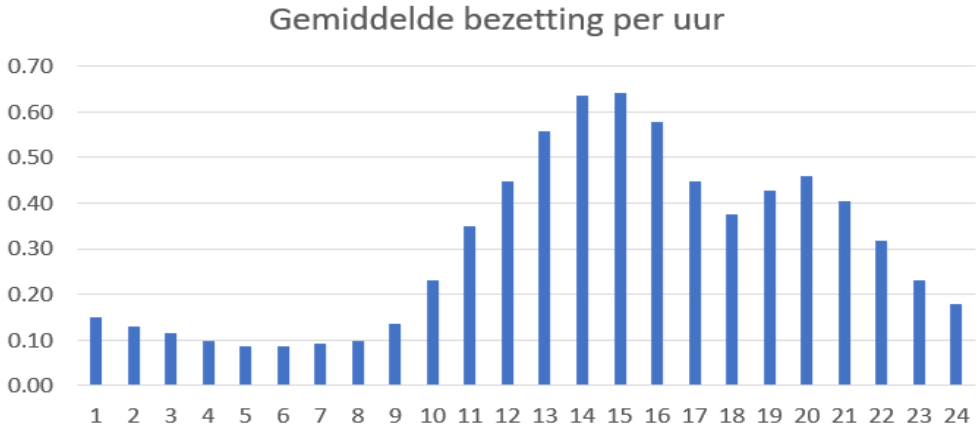
```
For x = 2 To 275

    Rows(x).EntireRow.Delete
    Rows(x).EntireRow.Delete
    Rows(x).EntireRow.Delete

Next x
```

FIGUUR 14 - MACRO VERWIJDEREN VAN KWARTIEREN

In Figuur 15 is de gemiddelde bezetting per uur te zien over alle dagen en alle parkeergarages in de dataset. Hierin is een duidelijk trendlijn te zien met beginnend een klein dal na middernacht. Waarna een stijging volgt vanaf het moment dat winkels openen om negen uur. Na deze stijging komt er een piek rond de middag tussen twee en drie uur. De waardes dalen weer tot zes uur in de avond, het tijdstip dat de winkels sluiten. Toch volgt er weer een piek na zes uur door de recreatievoorzieningen en restaurants die vooral druk zijn in de avond. Deze tweede lagere piek is rond acht uur in de avond. Vervolgens daalt deze tweede piek tot het dal van 6 uur in de ochtend.

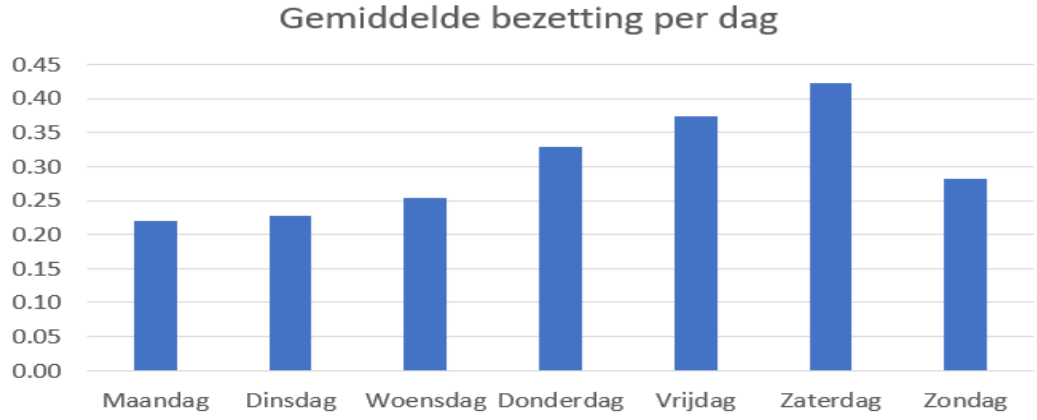


FIGUUR 15 - GEMIDDELDE BEZETTING PER UUR

4.4.3 Type dag

De volgende variabele die moet worden toegevoegd is de dag van de week. Voor het verdere proces is dit een belangrijke variabele, omdat er dan snel en makkelijk gekeken kan worden naar bijvoorbeeld alle vrijdagen. Dit staat dan los van de datum en is goed door te voeren. Ook is het een variabelen die mooie verschillen laat zien tussen verschillende dagen. Voor het toevoegen van de dag wordt de eerste datum opgezocht. In dit geval is dat 04-10-2018, wat een donderdag was. Voor elke set van 24 regels, elk uur op een dag, is het type dag hetzelfde. Hierdoor verandert dus na elke 24 regels het type dag en na alle dagen van de week start dit natuurlijk weer van voor af aan.

De gemiddelde bezetting is per dag weergegeven in Figuur 16 voor de gehele dataset. Het geeft een duidelijk patroon van welke dagen over het algemeen drukker zijn dan andere. Zo is de zaterdag de drukste dag in parkeergarages, wat logische wijs ook vaak verwacht wordt. De relatief lage gemiddelde bezetting van zondag kan verklaard worden doordat er koopzondagen en geen koopzondagen zijn. De koopzondagen zijn over het algemeen heel druk, terwijl zondagen waarop de winkels gesloten zijn juist het tegenovergestelde laten zien. De gemiddelde bezetting hangt daar dus tussenin. Het type dag is dus een variabele met grote invloed op de bezettingsgraad in parkeergarage.



FIGUUR 16 - GEMIDDELDE BEZETTING PER DAG

4.4.4 Locatie

De variabele locatie is voor een dataset van één garage voor elke regel gelijk. Echter is het wel een variabele die belangrijk is om mee te nemen. In een bestand waar alle parkeergarages worden samengevoegd, kan het van belang zijn om naar enkele parkeergarages afzonderlijk te kijken. Of juist twee specifieke parkeergarages te vergelijken, door de variabele locatie kunnen deze garages er makkelijk uitgefilterd worden. De locatie wordt weergegeven met eerst de plaatsnaam gevolgd door de naam van de garages. In dit geval wordt de locatie dus weergegeven door: Breda – De Barones. Zo kan er dus ook nog gekeken worden naar verschillende parkeergarages in dezelfde stad.

4.4.5 Capaciteit

Elke parkeergarage heeft een eigen capaciteit, wat het maximale aantal parkeerplaatsen aangeeft die er in de parkeergarage zijn. Voor het genereren van voorspellingen is het goed om deze variabelen in de gaten te houden. Het kan niet zo zijn dat de voorspelling namelijk hoger is dan de capaciteit. Net zoals de variabele locatie is deze variabelen voor één garage in de hele dataset gelijk.

4.4.6 Vrije plaatsen

De laatste variabele die door de originele historische data wordt weergegeven is het aantal vrije plaatsen. Er hoeft aan deze variabelen niks veranderd te worden. Doordat er al regels zijn verwijderd uit de dataset, staat deze variabele gelijk in de goede tijdseenheid genoteerd.

4.4.7 Bezetting

Uit het aantal vrije plaatsen en de capaciteit volgt de bezetting van een parkeergarage. Deze variabele is te verkrijgen door in Excel het aantal vrije plaatsen van de capaciteit af te trekken. Door dit in elke regel opnieuw te doen, wordt het meteen duidelijk hoeveel auto’s er geparkeerd staan omdat deze plekken niet meer vrij zijn.

4.4.8 Bezettingsgraad

Naast de bezetting kan ook de bezettingsgraad berekend worden. Voor het analyseren van één parkeergarage is de bezetting en de bezettingsgraad lineair met elkaar verbonden. Echter voor het vergelijken van meerdere parkeergarages is het heel nuttig. De bezettingsgraad is een percentage wat wordt berekend door de bezetting te delen door de capaciteit. Als de bezetting van een parkeergarage 171 is en van een andere 320, dan wil dit nog steeds niks zeggen over hoe druk het is in beide parkeergarages, omdat de capaciteit hier niet bij wordt vermeld. Stel de bezettingsgraad van deze benoemde bezettingen is respectievelijk 0.58 en 0.47, dan is dus de eerste parkeergarage met een bezetting van 171 drukker. Dit is precies waarom het belangrijk is bij het vergelijken van parkeergarages met een andere capaciteit om naar de bezettingsgraad te kijken.

4.4.9 Moment

Het moment van de dag is een variabele die het mogelijk maakt om een grovere schattingen te maken. De dag kan worden opgesplitst in vier delen: nacht, ochtend, middag en avond. Ieder deel van de dag bestaat uit 6 uur. Doormiddel van deze variabelen kunnen verschillende delen van de dag met elkaar vergeleken worden. Zo zijn de zondagmiddagen een interessant voorbeeld om te vergelijken. Daarnaast zouden er ook grove schattingen gemaakt kunnen worden met de aanname dat de bezetting in grotendeels in elk deel van de dag hetzelfde is. Zo kunnen er eventuele voorspellingen gemaakt worden voor de middagen.

4.4.10 Speciale gebeurtenis

De variabel speciale gebeurtenis is een belangrijke maar ingewikkelde variabelen. Het is logisch dat parkeergarages drukker zijn op een zondag als het koopzondag is dan wanneer de winkels op een zondag gesloten zijn. Deze variabelen wordt binair weergegeven, dat wil zeggen 0 als er geen event is

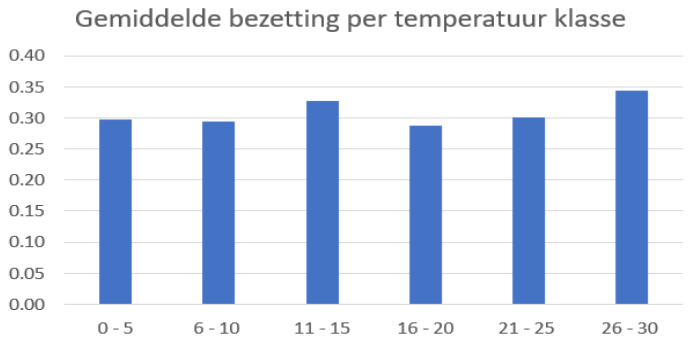
en 1 als er wel een event is. Koopzondagen worden meegenomen als speciale gebeurtenis. Verder is er gezocht naar andere evenementen die drukte in een stad veroorzaken. Hierbij kan worden gedacht aan marathons door een stad of kerstmarkten.

Het is lastig om vast te stellen welke evenementen drukte in een parkeergarage veroorzaken. De grootste uitschieters in de datasets zijn in ieder geval verklaard. In Appendix D is een overzicht te vinden waarin per stad de evenementen zijn weergegeven die zijn achterhaald. Vervolgens is er per stad gekeken of deze evenementen ook daadwerkelijk waren terug te vinden in de bezettingsgraad van de parkeergarages, dit is weergegeven in Appendix E. Het is te zien dat veel losse evenementen geen tot weinig invloed hebben op de bezetting van parkeergarages. Dit komt vooral door de locatie van evenementen die niet midden in het centrum is. Wat verder opvalt in de grafieken van de bezettingsgraad is dat er een gekke lijn te zien is op 25 en 26 december, de kerstdagen. Deze dagen zijn zo uniek en eigenlijk niet te vergelijken met de andere dagen in de dataset. Om deze reden zijn deze twee dagen uit de dataset gehaald, omdat ze anders teveel voor ruis zouden zorgen. Als er meer van dit soort extreme gevallen waren geweest of meer jaren aan data, had er een extra variabelen 'feestdag' kunnen zijn. Op dit moment is het logischer om deze twee dagen uit de dataset te halen.

4.4.11 Temperatuur

Uit de literatuur is de variabelen temperatuur in veel onderzoeken naar voren gekomen. Via het KNMI (2018) zijn alle weergegevens terug te vinden. In Nederland zijn 50 weerstations die gegevens opslaan en waarvan alle historische gegeven te verkrijgen zijn. Er is gekeken welk weerstation het dichtst bij welke parkeergarage ligt. In Appendix F is te vinden welke weersstations zijn gebruikt. De gegevens van het KNMI worden weergegeven per uur. Dat is de perfecte tijdseenheid voor de dataset, toch is er voor gekozen om de variabele temperatuur anders te definiëren. Met het oog op de voorspellingen is het handig om de voorspelling van het weer mee te nemen voor een bepaalde dag. Dus als de bezetting voor volgende week vrijdag voorspeld moet worden, is het handig als de verwachte temperatuur voor die vrijdag ingevoerd kan worden. Het zou onhandig zijn als er dan per uur de verwachte temperatuur ingevoerd moet worden.

De verwachte temperatuur per dag is niet meer te achterhalen. Daarom is doormiddel van de historische data gekeken naar de maximale temperatuur van die dag en die wordt aangenomen als de verwachte temperatuur. Doormiddel van een macro in Excel wordt er voor elke dag gekeken naar de maximale temperatuur. Deze wordt overgenomen en afgerond naar een geheel getal. Dit getal is voor die dag de verwachte temperatuur. In Figuur 17 is de gemiddelde bezetting van alle dagen en alle parkeergarages te zien verdeeld over 6 verschillende klassen. Door de relatief kleine dataset is het lastig om een goed beeld te krijgen van de invloed van deze variabelen. De maanden in de dataset hebben in temperatuur maar enkele uitschieters. Zo bevat de laatste klasse '26 – 30' maar drie dagen, waarvan twee weekenddagen. Dit heeft een grote invloed op de drukte in parkeergarages. Door deze toevalligheden is het niet duidelijk genoeg wat de invloed is van de temperatuur.



FIGUUR 17 - GEMIDDELDE BEZETTING PER TEMPERATUUR KLASSE

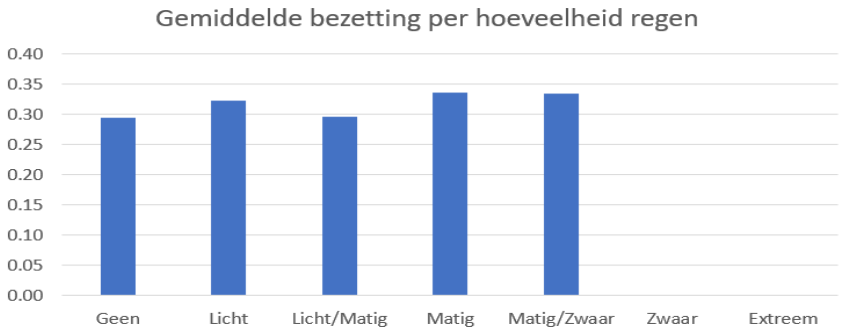
4.4.12 Regen

In de categorie weer wordt naast de temperatuur ook gekeken naar de hoeveelheid neerslag die wordt verwacht voor een bepaalde dag. Deze variabele werkt volgens hetzelfde principe als de variabele temperatuur. Er is vanuit de historische data per uur van het KNMI gekeken naar de maximale neerslag op een dag. Het weerbericht geeft namelijk aan of er op een dag neerslag verwacht wordt. Volgens weer.nl is de standaard verdeling voor de neerslagintensiteit als volgt:

- 0.1 tot 0.3 mm per uur
- 0.3 tot 1.0 mm per uur
- 1.0 tot 3.0 mm per uur
- 3.0 tot 10.0 mm per uur
- 10.0 tot 30.0 mm per uur
- Meer dan 30.0 mm per uur

Deze zes schalen worden in de dataset weergegeven met de volgende termen: lichte, lichte/matige, matige, matige/zware, zware en extreme neerslag. De neerslag op een dag wordt aangenomen als de van tevoren verwachte neerslag voor die dag, omdat net zoals de temperatuur de verwachtingen niet meer te achterhalen zijn. Per dag is er dus één term die wordt weergegeven, zodat bij de voorspellingen aan het eind van het model er voor een dag ook maar één term hoeft worden ingevoerd.

In Figuur 18 is per schaal de gemiddelde bezetting te zien over alle dagen en alle parkeergarages. Het eerste wat opvalt is dat er geen bezetting is weergegeven bij de schaal 'Zwaar' en 'Extreem'. Dit komt doordat in de dataset deze twee schalen niet voorkomen. Op de dagen in de dataset heeft het dus niet harder geregend dan 10.0 mm per uur. Verder zijn er geen grote verschillen te zien tussen de andere schalen. Dit kan met twee redenen verklaard worden. Dit kan komen door de relatief kleine dataset, want er zijn niet veel dagen te vergelijken met dezelfde schaal. Daarnaast kan het ook zo zijn dat door toeval het op drukkere dagen zoals voor kerst of in het weekend het telkens licht heeft geregend. Dit heeft direct een grote invloed op de weergave van deze variabele. De invloed van regen op de bezetting is dus op het eerste zich niet duidelijk te verklaren.



FIGUUR 18 - GEMIDDELDE BEZETTING PER HOEVEELHEID REGEN

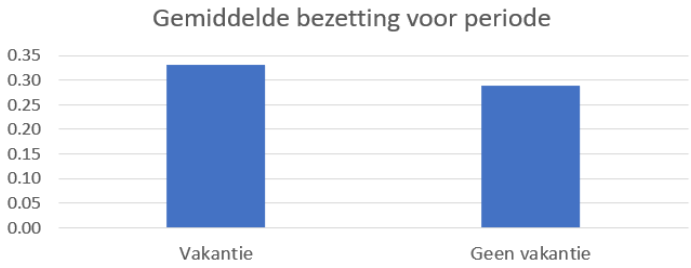
4.4.13 Vakantie

De laatste variabele die wordt meegenomen is vakantie. In veel onderzoeken is naar voren gekomen dat vakantie een significatie invloed uitoefend op de drukte in parkeergarages. Ook dit is een binaire variabele die 0 is als er geen vakantie is en 1 als het in die periode wel vakantie is. Er zijn twee vakantieperiodes in de dataset die ook per regio in Nederland verschillen. De vakanties zijn verkregen van de website van de Rijksoverheid (2018) en zijn in Tabel 5 weergegeven.

	Regio Noord	Regio Midden	Regio Zuid
Herfstvakantie	20 oktober t/m 28 oktober 2018	20 oktober t/m 28 oktober 2018	13 oktober t/m 21 oktober 2018
Kerstvakantie	22 december 2018 t/m 6 januari 2019	22 december 2018 t/m 6 januari 2019	22 december 2018 t/m 6 januari 2019

TABEL 5 - VAKANTIEPERIODES

Deze variabele is dus niet voor elke parkeergarage hetzelfde, maar er zijn wel veel overeenkomsten. In Figuur 19 is de gemiddelde bezetting weergeven voor de twee verschillende periode's; wel of geen vakantieperiode. Het is direct te zien dat de gemiddelde bezetting tijdens vakantieperiodes hoger is dan wanneer het geen vakantie is. Het is een verschil van 4%, wat relatief laag is. Echter door de kleine dataset valt de conclusie te trekken dat de variabele vakantie wel invloed heeft op de bezetting.



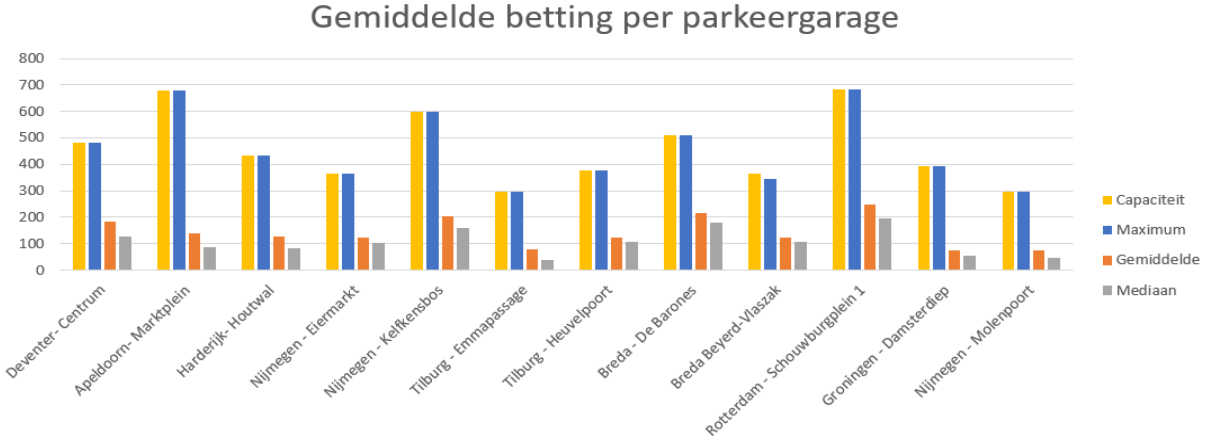
FIGUUR 19 - GEMIDDELDE BEZETTING PER PERIODE

4.5 Patronen in de bezetting

De gevormde dataset met de genoemde variabelen uit de vorige paragraaf kan geanalyseerd worden om inzicht te krijgen in de patronen over tijd, de piek uren en de gemiddelde bezetting. In deze analyse zijn de twaalf parkeergarages meegenomen, die eerder zijn genoemd.

4.5.1 Bezetting per parkeergarage

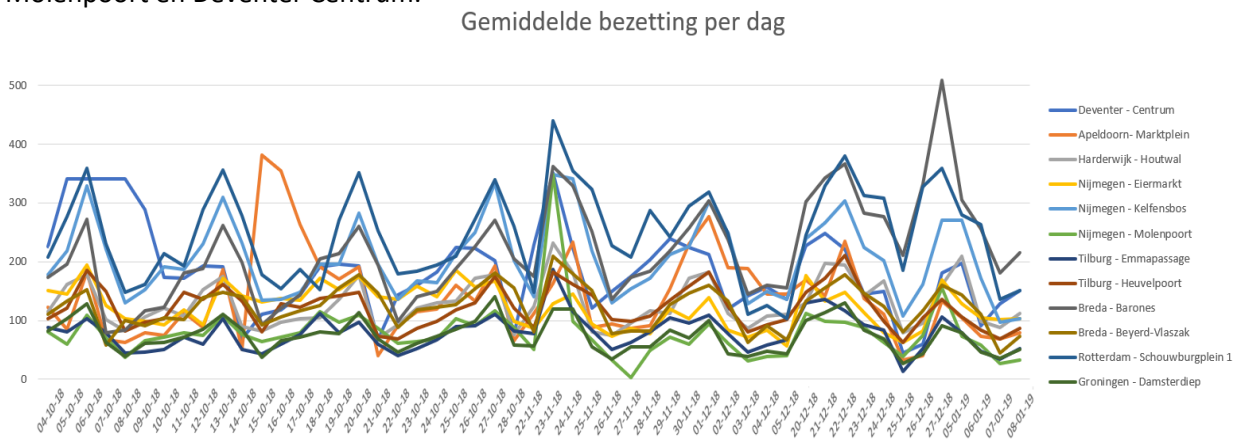
Het analyseren van de bezetting per parkeergarage geeft een globaal overzicht van de bezetting en de piekwaardes. In Figuur 20 zijn voor alle parkeergarage verschillende waarden weergegeven. Eerst is de capaciteit te zien met daarnaast de maximale bezetting die voorkomt in de dataset. De grafiek laat zien dat elke parkeergarage wel eens helemaal vol zit, behalve de garage Beyerd-Vlaszak in Breda. Verder is de gemiddelde bezetting en de mediaan weergegeven, deze zijn berekend over alle dagen per parkeergarage. Het gemiddelde omvat overal ongeveer 1/3 van de capaciteit. Hieruit kan geconcludeerd worden dat de parkeergarages wel degelijk met elkaar te vergelijken zijn en hetzelfde gedrag vertonen. Daarnaast is de mediaan bij elke parkeergarage lager dan het gemiddelde. Wat deze grafiek ook impliciet laat zien is dat de bezettingen rechtsscheef verdeeld zijn.



FIGUUR 20 - GEMIDDELDE BEZETTING PER PARKEERGARAGE

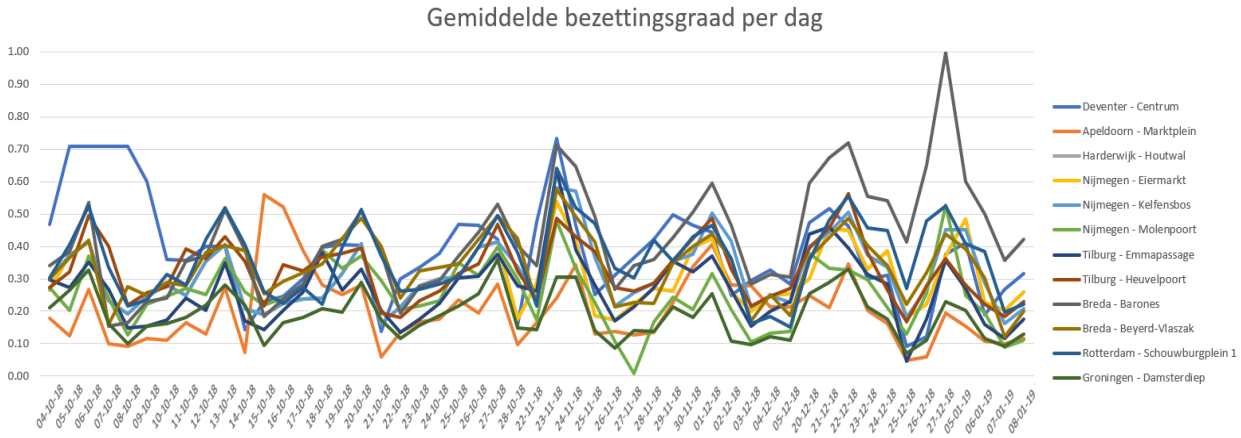
Nu het duidelijk is dat de parkeergarages met hun verdeling op elkaar lijken, is de volgende stap om dieper in te gaan op de patronen die de garages geven. Figuur 21 laat de gemiddelde bezetting per dag per parkeergarage zien. De precieze informatie is weergegeven in Appendix G. Er zijn duidelijke cycli te zien die de weken weergeven. In het weekend is het duidelijk drukker dan door de week en dit verklaart dan ook om de 5/6 dagen een hoge piek. De hoogte van de pieken verschilt echter wel veel, maar dit komt vooral door de verschillende capaciteiten van de parkeergarages.

Verder zijn er een paar uitschieters te zien. De duidelijkste is de oranje piek op 15 oktober, van de parkeergarage Marktplein in Apeldoorn. Deze piek is niet te verklaren en wordt in de dataset dus meegenomen als event, om de ruis zoveel mogelijk in te perken. Verder valt nog op dat in het midden één piek relatief hoger is dan de rest en dat dit ook voor meerdere parkeergarages geldt. Dit is het weekend vóór de kerstdagen, gewoonlijk wordt er dan al drukte verwacht in en rondom winkelcentrums. Voor de parkeergarages die deze duidelijk grotere piek weergeven, wordt deze zondag als speciale gebeurtenis gezien om in de dataset duidelijker te maken dat het om een drukkere dag gaat dan normaal. Dit geldt voor de parkeergarages: Schouwburgplein 1, De Barones, Kelfensbos, Molenpoort en Deventer Centrum.



FIGUUR 21 - GEMIDDELDE BEZETTING PER DAG

Naast de bezetting, is in Figuur 22 de bezettingsgraad weergegeven per dag per parkeergarage. Ook hier zijn duidelijk uitschieters te zien, waarvan sommige te verklaren zijn. Zo is de donkerbruine piek op 27 december van de parkeergarage De Barones te verklaren door een systeem fout. Op deze dag is voor elk uur de bezetting de volledige capaciteit. Hiervan kan met gezond verstand gezegd worden dat dit een fout is, daarom wordt deze dag uit de dataset verwijderd. Hetzelfde geldt voor de eerste dagen van de paarkeergarage in Deventer. Ook hier wordt drie dagen achter elkaar dezelfde bezetting gemeten. Het zou kunnen zijn dat door omstandigheden een parkeergarage wordt gesloten, maar voor het onderzoek wordt er aangenomen dat het systeem de metingen niet goed heeft opgeslagen.



FIGUUR 22 - GEMIDDELDE BEZETTINGSGRAAD PER DAG

Als deze twee onlogische pieken uit de grafiek worden gedacht, is te constateren dat de patronen van de bezetingen van de parkeergarages op elkaar lijken. Verder worden er ook nog andere evenementen meegenomen zoals de intocht van Sinterklaas, Kerstmarkten, marathons en een aantal andere evenementen. Voor elk evenement is gekeken of er ook een significant verschil te zien is tegenover dezelfde dag in andere weken. Door deze vergelijkingen is er een lijst tot stand gekomen, die zo goed mogelijk de ruis zal wegnemen. De variabele speciale gebeurtenis zorgt dus voor een zo zuiver mogelijke voorspelling.

4.6 Conclusie

In dit hoofdstuk is er een dataset ontwikkeld en geanalyseerd. Deze is gevormd door historische gegevens samen te voegen met verklarende variabelen. Ook is er gekeken naar de beperkingen van de dataset en hoe deze invloed hebben op het verdere onderzoek. Verder is er gekeken naar patronen in deze dataset en of deze patronenen ook te verklaren zijn. Kortom, er is antwoord gegeven op de volgende vragen:

- Welke verklaringen zijn er voor de patronen in de dataset?
    - Welke verschillen zijn er tussen de parkeergarages?
    - Wat zijn de beperkingen van de dataset?
    - Hoe passen alle variabelen bij elkaar in één dataset?
    - Welke patronen zijn er in de dataset aanwezig?

De dataset begint met ruwe data die wordt verzameld door het systeem, waarin per kwartier het aantal vrije plaatsen van een parkeergarage worden genoteerd. De variabelen die in deze historische gegevens worden waargenomen zijn: *de datum, het tijdstip, de locatie, het aantal vrije plaatsen* en *de capaciteit.* Deze gegevens worden bijgehouden voor 128 parkeergarages in Nederland. Deze parkeergarages hebben overeenkomsten, maar ook duidelijke verschillen. Zo is een belangrijk verschil de locatie van de parkeergarge. Een duidelijk onderscheid wat er gemaakt kan worden in de locatie is of ze in het *centrum* liggen of juist bij een *bedrijventerrein.* Voor het onderzoek worden 12 pakeergarages gebruikt, die allemaal in het centrum van verschillende steden gevestigd zijn.

De totale dataset van verzamelde gegevens is voldoende om voorspellingen mee te generen, echter zijn er ook een aantal beperkingen. Zo is de dataset relatief klein en zouden de voorspellingen naukeuriger zijn als er meer historische gegevens beschikbaar zouden zijn. Verder is het niet te achterhalen hoelang één auto geparkeerd staat in de parkeergarage, waardoor er niets te zeggen valt over de drukte naar en van een parkeergarage. Tot slot heeft het systeem zelf ook zijn beperkingen, omdat soms gegevens niet correct worden opgeslagen. Zo zitten er wat gaten in de dataset en kunnen de dagen per parkeergarage ook met elkaar verschillen.

Na het analyseren van de historische gegevens is de volgende stap het compleet maken van de dataset met alle variabelen die uit het literatuuronderzoek naar voren zijn gekomen. Er zijn direct duidelijke patronen te zien bij verschillende variabelen zoals: *type dag, tijd* en *vakantie*. Ook de gemiddelde bezetting, maximale bezetting en de mediaan zijn weergegeven. Hierin is te zien dat de bezetting rechtsscheef verdeeld is. Tot slot zijn de gemiddelde bezetting(sgraad) per dag in een grafiek te zien en valt te concluderen dat er duidelijk patronen en cycli terug te vinden zijn bij de 12 parkeergarages. De cycli geven duidelijk de weken aan, waarin het in een weekend drukker is dan op werkdagen. Sommige uitschieters zijn te verklaren door verkeerd opgeslagen historische gegevens en andere zijn te verklaren door feestdagen, koopzondagen en vakantie.

Hoofdstuk 5 – De implementatie

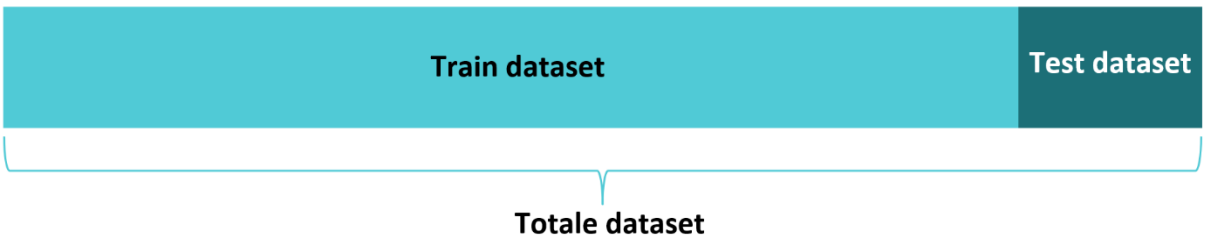
De gegevens zijn compleet en met deze gegevens kan de volgende stap worden gemaakt, de ontwikkeling van de regressieboom. Het model moet de dataset goed begrijpen om verbanden tussen de verschillende variabelen te kunnen leggen. Er kan niet meteen naar de resultaten gekeken worden, voordat het model goed in elkaar zit. De onderzoeksvraag voor dit hoofdstuk is daarom: “Hoe kan de het proces worden weergegeven in een model?”. Voor het beantwoorden van deze vraag, zijn de volgende deelvragen opgesteld:

- Hoe wordt de data opgesplitst in een train en test dataset?
- Wat zijn de in- en output variabelen van het model?
- Hoe komt het model en de visualisatie hiervan tot stand?

Eerst zal de complete dataset opgedeeld worden in een train dataset en een test dataset (§5.1). Verder in het hoofdstuk wordt ook uitgelegd waarom dit verschil zo belangrijk is. Vervolgens wordt er aan de hand van deze gegevens gekeken hoe een simpele voorspellingstechniek gebruikt kan worden om voorspellingen te genereren (§5.2). Daarna volgt de implementatie van de regressieboom. Hierin worden verschillende aspecten bekeken, zoals de complexiteit van het model en de invloed van de verschillende variabelen, om zo tot het meest geschikte model te komen (§5.3). Dit uiteindelijke model wordt gevisualiseerd en besproken om duidelijk te maken hoe de regressieboom gelezen kan worden (§5.4). Tot slot volgt er een conclusie van het hoofdstuk waarin de bovenstaande vragen worden beantwoord (§5.5).

5.1 Train en test dataset

Nu de dataset helemaal compleet is, moet er nog één stap gemaakt worden voordat het generen van voorspellingen kan beginnen. Als het model af is en er voorspellingen gegenereerd kunnen worden, is het belangrijk om te weten hoe betrouwbaar deze voorspellingen zijn. Om hier achter te komen moeten de voorspellingen vergeleken kunnen worden met werkelijke historische gegevens. De gegevens waarop het model getest wordt, kunnen niet dezelfde gegevens zijn als de input van het model. Daarom wordt er een onderscheid gemaakt tussen een train dataset en een test dataset. Met de train dataset wordt het model zogezegd getraind, om tot een goede voorspelling te komen. Aan de hand van de test dataset worden deze voorspelingen getest. In Figuur 23 is de verdeling te zien van de dataset.

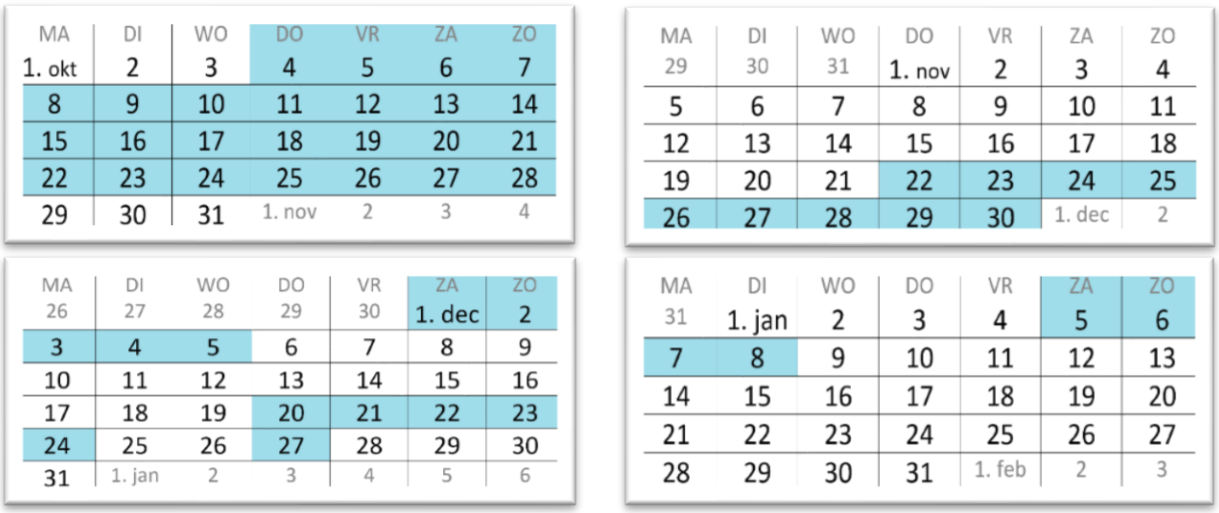


FIGUUR 23 - VERDELING VAN DE DATASET

Er zijn verschillende manieren om de dataset op te splitsen. De techniek die in veel onderzoeken wordt gebruikt is *kruisvalidatie*, ook wel bekend in het Engels als cross-validation. Deze techniek gebruikt één parameter, genaamd k, die verwijst naar het aantal groepen waarin een dataset gesplitst moet worden. Eerst worden de gegevens door elkaar gehusseld, waarna de dataset in k-groepen verdeeld wordt. Vervolgens worden het gewenste aantal k-groepen in de train dataset gestopt en het overblijfsel is uiteraard de test dataset. Zo kunnen er verschillende combinaties gemaakt worden, waardoor toeval wordt uitgesloten. De methode is populair, omdat hij eenvoudig te begrijpen is en over het algemeen een goede verdeling maakt voor train- en test datasets.

Voordat de dataset door elkaar wordt gehusseld is het van belang om te weten op basis waarvan de groepen moeten worden ingedeeld. De dataset, zoals die nu is, kan per regel worden bekeken maar ook per 24 regels, of te wel per dag. Dit is een belangrijk verschil wat moet worden meegenomen voordat er een splitsing wordt gemaakt. Het doel van het onderzoek is om een uur, een dag, een maand of zelfs langer vooruit te kunnen voorspellen. Als de dataset per regel wordt bekeken en er enkel losse uren uit filteren voor de test dataset, kunnen er ook alleen uren met elkaar worden vergeleken. Om deze reden worden de uren van een dag bij elkaar gehouden en horen de 24 regels van één dag dus bij elkaar. Zo kan uiteindelijk de betrouwbaarheid van een voorspelling voor een hele dag én per uur berekend worden.

In Figuur 24 zijn de dagen blauw gemarkeerd als ze in de dataset aanwezig zijn. In totaal zijn het dus 49 verschillende dagen waarvan de gegevens zijn opgeslagen. Er zijn 12 parkeergarages in de dataset meegenomen dus twaalf maal 49 dagen betekent in totaal 588 dagen met waarnemingen. Echter waren er wat momenten waarop het systeem geen of verkeerde gegevens heeft opgeslagen. In totaal zijn er 12 dagen verwijderd uit de dataset, omdat deze dagen onlogische gegevens lieten zien. De dagen die uit de dataset zijn gehaald kunnen worden gevonden in Appendix H. Na het verwijderen van deze dagen blijven er in totaal 576 dagen over in de dataset, die allemaal gebruikt kunnen worden voor verder onderzoek. Deze uiteindelijke dataset wordt verdeeld in een train- en test dataset.



FIGUUR 24 - DE DAGEN VAN DE WAARNEMINGEN

De dataset is relatief klein voor het onderzoek wat verricht wordt. In veel gevallen wordt 10% van de gegevens als test dataset gezien, maar in dit geval wordt de test dataset kleiner. Per parkeergarage worden er drie willekeurige dagen uit de dataset gehaald. Dit betekent dus, 3 maal 12, dat er 36 dagen in de test dataset komen en er 540 dagen in de train dataset eindigen. De verdeling wordt dan 93,75% van de data in de train dataset en 6,25% van de gegevens in de test dataset.

Het model kan één keer worden gerund met deze specifieke train- en test dataset, maar de kans op toevallige resultaten wordt dan groot. Daarom is het van belang kruisvalidatie wordt toegepast en het model getest wordt op meerdere train datasets met bijbehorende test datasets. Zo kan er worden gevalideerd of de resultaten van het model puur toeval zijn voor elke set, of dat er een duidelijk patroon in gegevens zitten. Voor verder toelichting van het model in dit hoofdstuk wordt maar naar één combinatie train en test dataset gekeken. De andere datasets met verschillende verdelingen van train en test dataset worden niet besproken, maar worden op dezelfde manier uitgevoerd. De resultaten zijn uiteraard terug te lezen in het volgende hoofdstuk. Daar wordt gekeken naar de nauwkeurigheid van het model en of deze bij de verschillende dataset vergelijkbaar is.

5.1.1 Het verdelen van de train- en test set

De 49 verschillende dagen die in de dataset aanwezig zijn krijgen allemaal hun eigen nummer. Doormiddel van een macro in Excel worden willekeurige getallen gekozen tussen 1 en 49, die refereren naar de datum horend bij het nummer. Zo worden er per parkeergarage drie nummers toegewezen, die corresponderen met de 49 verschillende dagen in de dataset. Er wordt direct gecontroleerd of er per parkeergarage wel drie verschillende type dagen gebruikt worden. De lijst met data die hieruit voort komt is terug te vinden in een overzicht in Appendix I. Aan de hand van de belangrijke variabele type dag is gekeken of het percentage van de verschillende dagen in de train en test dataset op elkaar lijken om te voorkomen dat er bijvoorbeeld geen maandagen in de train dataset zitten of juist in de train dataset.

Type dag	Aantal in Train dataset	Percentage in train dataset	Aantal in Test dataset	Percentage in Test dataset
Maandag	75	0.14	5	0.14
Dinsdag	63	0.12	6	0.17
Woensdag	55	0.10	4	0.11
Donderdag	89	0.16	5	0.14
Vrijdag	77	0.14	6	0.17
Zaterdag	90	0.17	5	0.14
Zondag	90	0.17	5	0.14
Totaal	**540**	**1.00**	**36**	**1.00**

TABEL 6 - HET PERCENTAGE TYPE DAG PER DATASET

In Tabel 6 is te zien dat het percentage maandagen in beide datasets gelijk is aan elkaar. Verder zijn er geen overeenkomsten, maar ook geen significante verschillen. Het type dag is een variabele die veel invloed heeft op de bezettingsgraad. Dat is ook de reden waarom naar dit percentage gekeken wordt en daaruit te concluderen valt dat de test dataset een goede verdeling van verschillende dagen bevat.

5.2 Eenvoudige voorspelling

Voor het voorspellen van bepaalde situaties of patronen wordt er al snel gedacht aan complexe en moeilijke algoritmes. Toch kan het soms zijn dat een simpel model of enkel het gemiddelde al voldoende is om accurate voorspellingen te genereren. Eenvoudige methoden kunnen daarnaast ook het begrip vergroten en de kans op fouten verminderen (Green & Armstrong, 2015). Om er zeker van te zijn dat de regressieboom ook daadwerkelijk iets toevoegt en beter presteert dan een simpel model, wordt er eerst een eenvoudige techniek toegepast om de voorspellingen te generen.

De oorspronkelijke historische gegevens zijn opgeslagen met daarin de tijd en datum verwerkt. Het is ook te verwachten dat deze twee variabelen de meeste invloed hebben op de bezettingsgraad, misschien wel als enige variabelen invloed hebben. Om dit na te gaan, worden er voorspellingen gegenereerd op basis van het gemiddelde van de variabele tijd en de variabele datum. Aan de hand van deze resultaten kan het duidelijk worden of het regressiemodel wellicht overbodig is en dat de andere variabelen overbodig zijn. De procedure begint bij het berekenen van het gemiddelde van elk uur afzonderlijk en het gemiddelde van elke dag afzonderlijk. Daarna wordt het gemiddelde genomen voor elke combinatie uur en dag, waaruit de verwachte bezettingsgraad volgt per uur per dag. Voor het voorspellen van een bepaalde bezettingsgraad wordt dus alleen gekeken naar het tijdstip en het type dag en op basis daarvan volgt de verwachte bezettingsgraad. In het volgende hoofdstuk zullen de resultaten van dit model worden weergegeven en geanalyseerd. Ook zal het model naast de regressieboom worden gelegd om te zien of ze even goed presteren of dat er een methode duidelijk als beste naar voren komt.

5.3 De implementatie

De test dataset is nu helemaal klaar om te gebruiken voor het generen van voorspellingen. In het literatuuronderzoek in hoofdstuk 3 is naar voren gekomen dat de regressieboom het meest geschikte algoritme is voor dit onderzoek. In deze paragraaf wordt het model geformuleerd dat de uiteindelijke regressieboom vormt waaruit de bezettingsgraad te voorspellen is voor parkeergarages in het centrum van steden. Om de regressieboom te visualiseren en ermee te rekenen is gekozen voor het programma R. Dit is een softwarepakket en programmeertaal dat veel wordt gebruikt bij dataverwerking en statistische toepassingen. Het programma biedt uitgebreide technieken zoals, lineair en niet lineaire modellen, statistische toetsen, tijdreeksanalyses, classificatie en clustering algoritmes. De techniek die in dit onderzoek gebruikt wordt heet 'rpart', wat staat voor *Recursive Partitioning And Regression Trees.* Dit houdt in wezen in dat er een beslissingsboom wordt gecreëerd, maar de uitkomsten zijn niet binair. Zo komt er geen algemene formule uit die een bepaalde kans genereerd, maar het creëert een patroon dat zegt: als *a, b* en *c* waar zijn, dan is de kans *x*.

5.3.1 Het inlezen van de gegevens

Het programma R eist dat gegevens worden geformatteerd in een .csv bestand om het te verwerken, in het Engels staat dit voor *comma-seperated values*. Het voordeel van dit type bestand is dat het nog steeds met Excel geopend kan worden, maar ook als tekst bestand. In het tekst bestand worden alle waardes, zoals de naam van het bestand al zegt, gescheiden met een puntkomma en dat is ook hoe R het bestand leest. Zodra de data is ingelezen, wordt alles in precies dezelfde rijen als in Excel weergegeven, zie Figuur 25.

```
> head(traindata)
             Locatie Capaciteit     Datum1   TypeDag VrijePlaatsen Bezetting Bezettingsgraad
1 Deventer - Centrum        481 10-10-2018 Woensdag           445        36            0.07
2 Deventer - Centrum        481 10-10-2018 Woensdag           445        36            0.07
3 Deventer - Centrum        481 10-10-2018 Woensdag           445        36            0.07
4 Deventer - Centrum        481 10-10-2018 Woensdag           445        36            0.07
5 Deventer - Centrum        481 10-10-2018 Woensdag           445        36            0.07
6 Deventer - Centrum        481 10-10-2018 Woensdag           446        35            0.07
    Moment Tijd Event Temperatuur Temp.1 Regen.1 Regen Vakantie
1    Nacht    1     0          24    6.2       0  Geen        0
2    Nacht    2     0          24    5.0       0  Geen        0
3    Nacht    3     0          24    6.0       0  Geen        0
4    Nacht    4     0          24    6.2       0  Geen        0
5    Nacht    5     0          24    7.8       0  Geen        0
6 Ochtend     6     0          24    7.0       0  Geen        0
```

FIGUUR 25 - TRAIN DATASET INGELEZEN IN R

Vervolgens is ook de structuur uit te lezen van de train dataset. De structruur is weergegeven in Figuur 26 waarin boven aan als eerst te lezen is dat er 12786 regels aan waarnemingen zijn met elk 15 verschillende variabelen. Vervolgens worden alle 15 variabelen weergegeven met daar achter het datatype en het aantal waarnemingen, zoals ze voorkomen in de dataset. Er zijn vier verschillende datatypes die voorkomen.

```
> str(traindata)
'data.frame':   12786 obs. of  15 variables:
 $ Locatie        : Factor w/ 12 levels "Apeldoorn - Marktplein",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Capaciteit     : int  481 481 481 481 481 481 481 481 481 481 ...
 $ Datum          : Factor w/ 49 levels "1-12-2018","10-10-2018",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ TypeDag        : Factor w/ 7 levels "Dinsdag","Donderdag",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ VrijePlaatsen  : int  445 445 445 445 445 446 439 420 357 199 ...
 $ Bezetting      : int  36 36 36 36 36 35 42 61 124 282 ...
 $ Bezettingsgraad: num  0.07 0.07 0.07 0.07 0.07 0.07 0.09 0.13 0.26 0.59 ...
 $ Moment         : Factor w/ 4 levels "Avond","Middag",..: 3 3 3 3 3 4 4 4 4 4 ...
 $ Tijd           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Event          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Temperatuur    : int  24 24 24 24 24 24 24 24 24 24 ...
 $ Temp.1         : num  6.2 5 6 6.2 7.8 7 9.9 13.2 17 20 ...
 $ Regen.1        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Regen          : Factor w/ 5 levels "Geen","Licht",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Vakantie       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

FIGUUR 26 - DE STRUCTUUR VAN DE TRAIN DATASET

Het type ‘factor’ komt vijf keer voor in de dataset bij de variabelen: locatie, datum, type dag, moment en regen. Deze vijf variabelen worden als factor gezien, omdat ze uit letters bestaan in plaats van uit cijfers. Elke variabelen heeft zo zijn eigen aantal levels dat gelijk staat aan het aantal verschillende waarnemingen voor de desbetreffende variabelen. Zo heeft de variabele Type Dag zeven levels, omdat deze gelijk staan aan het aantal dagen in een week.

Het type ‘integer’, wat staat voor een geheel getal komt ook vijf keer voor. De variabelen capaciteit, vrije plaatsen, bezetting, tijd en temperatuur worden gelezen als integer. Deze variabelen kunnen alleen maar gehele getallen aannemen. Ook de tijd wordt gezien als getal, omdat het niet noodzakelijk is voor het programma dat het wordt gezien als een tijdseenheid. De getallen 1 tot en met 24 zijn makkelijk te begrijpen als de uren van een dag. Verder is het ook logisch dat deze variabelen geen kommagetallen kunnen zijn, want de capaciteit kan bijvoorbeeld niet bestaan uit halve auto’s.

Ook het type ‘numeric’, komma getallen, komt voor in de dataset bij de variabelen: bezettingsgraad, temperatuur.1 en regen.1. De bezettingsgraad is het percentage wat bezet is en wat wordt weergegeven met een kommagetal. Daarnaast zijn de andere twee variabelen de gegevens per uur, die uiteindelijk zijn omgezet en niet direct worden meegenomen in het model. Deze gegevens staan er puur ter controle voor de nieuwe variabelen temperatuur en regen.

Tot slot is het laatste type dat in de dataset voorkomt ‘logical’. De variabelen speciale gebeurtenissen, event genoemd in de dataset, en vakantie worden binair weergegeven. In Excel hadden deze variabelen alleen de waarden 0 en 1, nu worden deze respectievelijk weergegeven met de woorden FALSE en TRUE.

5.3.2 Het algoritme van de regressieboom

Al kort is benoemd dat het algoritme genaamd ‘rpart’ verwijst naar de regressieboom, die wordt gebruikt om de bezettingsgraad van de parkeergarages te voorspellen. Deze inzichtelijke techniek geeft de relaties weer tussen een variabele y en verklarende variabelen $x_1,\ x_2,\ x_3, \ldots.\ , x_i$. Hierbij mogen deze relaties tussen de variabelen niet-lineair zijn. Een belangrijk onderscheid tussen een beslissingsboom en een regressieboom is dat de voorspellende variabele y bij een beslissingsboom binair is, dus bijvoorbeeld 0 of 1. Bij een regressieboom is y een continue variabele, die alle waarden tussen een bepaald interval kan aannemen. De variabel y is in dit onderzoek een percentage, dus kan daarom alle waarden tussen 0 en 1 aannemen.

Het algoritme van een regressieboom kijkt of de variabiliteit van de y-variabele verminderd kan worden door de waarde van één van de x-variabelen in verschillende stukken te verdelen. Zo kan bijvoorbeeld de set van y-waarden verdeeld worden over twee sets waarbij geldt: set 1 bevat alle y-waarden waarvoor $x_2 \leq 0.4$ en set 2 bevat alle y-waarden waarvoor $x_2 > 0.4$. Een dergelijke verdeling is alleen voordelig als de variabiliteit van de y-variabele in beide sets minder is dan de variabiliteit zonder deze splitsing. Het algoritme zoekt naar deze verbanden en welke relatie het meeste bijdraagt aan het reduceren van de variabiliteit in y. Dus eerst kijkt het algoritme naar welke x-variabele veel invloed heeft op de variabiliteit en vervolgens wordt er gekeken waar de optimale splitsing zit.

Deze optimale splitsing wordt berekend in het algoritme door het testen van elke waarde van elke variabele, waarbij de splitsing wordt gekozen waarbij de waarde van de som van de kwadratische fout het kleinste is. Dit wordt gedaan aan de hand van de volgende formule:

$$Som\ van\ de\ kwadratische\ fout = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2$$

In deze vergelijking is y_i de werkelijke waarden van de voorspellende variabele, oftewel de werkelijk gemeten bezettingsgraad. Daarnaast is $\bar{y}_1$ het gemiddelde van de voorspelde y-waarden over de linkerkant van de mogelijke splitsing. Hetzelfde geldt voor $\bar{y}_2$, alleen dit is het gemiddelde voer de rechterkant van de mogelijke splitsing. In feite wordt de werkelijke bezettingsgraad minus de voorspelde bezettingsgraad berekend. Dit is de error die gesommeerd wordt over dus de linker- en rechterkant van de boom. De splitsing waarbij deze error het kleinste is, wordt gezien als de optimale splitsing.

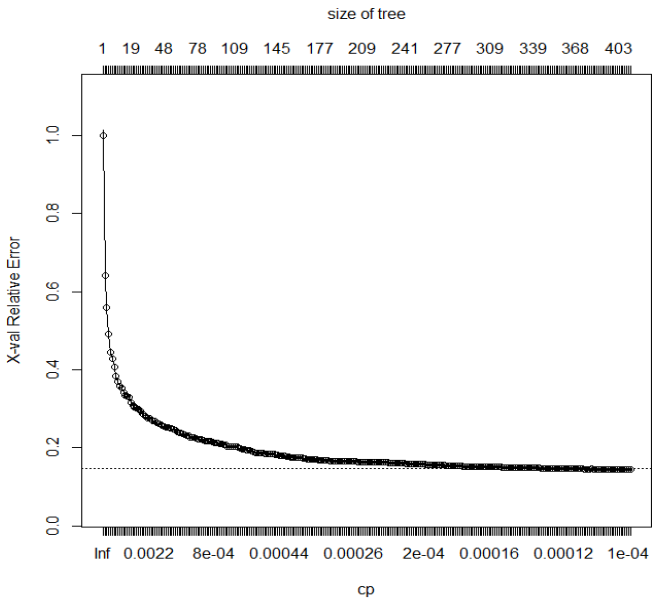
Na één splitsing wordt dit proces herhaald om de volgende mogelijke splitsing te berekenen. Het algoritme stopt afhankelijk van de vooraf opgegeven restrictie. Dit kan zijn wanneer er minder dan een bepaald aantal y-waarden overblijven na een splitsing. Het kan ook afhankelijk zijn van de complexiteit die wordt gehanteerd. Het splitsproces kan worden weergegeven in een boomstructuur. Uiteindelijke eindigt de boom in een eindknooppunt met daarin de gemiddelde waarde van een groep y-data die voldoet aan het vooraf afgelegde pad. Het betekent dus dat er een patroon is tussen de waarden die samen in één eindknooppunt vallen en daarom dezelfde gemiddelde waarde kunnen hanteren. Bij een regressieboom is een groot voordeel dat er op basis van een goede interpretatie gekeken kan worden of er variabelen gedeeltelijk of helemaal uit de analyse weggelaten kunnen worden, omdat er duidelijk naar voren is gekomen dat ze geen voorspellend verband hebben met de voorspellende variabelen.

5.3.3 Het programmeren van het model

De gegevens zijn in het programma geladen dus de volgende stap is om de codes te schrijven voor het berekenen van de regressieboom. In het software programma R zijn slimme algoritmes ingebouwd, deze kunnen worden aangeroepen en worden ingevuld met verschillende variabelen en restricties. Uiteindelijk volgt, na enkele eerdere stappen, de volgende formule:

$$
\begin{aligned}
&> Regressieboom \\
&\quad < -rpart\ (Bezettingsgraad \sim Capaciteit + TypeDag + Tijd + \ Event \\
&\quad + Temperatuur + Regen + Vakantie, \\
&\quad data = traindata, \quad method = \text{"}anova\text{"}, \quad cp = 0.0001)
\end{aligned}
$$

Eerst wordt de naam weergegeven van het model dat wordt gecreëerd en dat is in dit geval *Regressieboom*. Na de naam van het model wordt de formule geschreven, beginnend met welk algoritme er gebruikt moet worden. In dit geval verwijst *rpart* naar het pakket dat er een classificatie of regressieboom moet worden gebouwd. Vervolgens wordt er ingevuld met welke variabelen dit moet gebeuren. Als eerste wordt de variabele bezettingsgraad genoemd, omdat dit de variabelen is die voorspeld moet worden en dus onderaan als einde van de boom moet worden weergegeven. Na het ~ teken, worden alle variabelen weergegeven die de variabele bezettingsgraad moeten verklaren. Als alle variabelen zijn genoemd, moet ook worden benoemd waar deze informatie te vinden is. Daarom wordt er geschreven dat alle data in de train dataset is weergegeven, genaamd traindata. Verder zijn er nog twee dingen die belangrijk zijn om te vermelden. De methode die moet worden gebruikt binnen het pakket heet rpart. Zoals gezegd, kan er een classificatie boom en een regressieboom gecreëerd worden. Respectievelijk worden deze aangeduid met de begrippen *class* en *anova*, daarom wordt in dit geval voor de methode anova ingevuld. Als laatste moet de complexiteit van het model worden ingevuld, afgekort met *cp*. Deze kan een waarden hebben tussen 0 en 1. Als de complexiteit gelijk is aan 1, betekent dit dat er geen complexiteit is en zal de boom maar 1 blad hebben, oftewel alle waarnemingen zitten in één klasse. Hoe dichter het getal bij nul komt, hoe complexer de boom wordt.



	CP	nsplit	rel error	xerror	xstd
1	3.5878e-01	0	1.000000	1.00009	0.0129122
2	8.5520e-02	1	0.641222	0.64135	0.0088905
3	6.7264e-02	2	0.555702	0.56279	0.0081274
4	4.1124e-02	3	0.488439	0.49340	0.0073251
5	1.9818e-02	4	0.447315	0.44749	0.0064919
6	1.4193e-02	5	0.427497	0.43047	0.0061060
7	1.2227e-02	7	0.399111	0.40820	0.0059321
818	1.1491e-07	1034	0.089025	0.13352	0.0028365
819	9.8195e-08	1035	0.089025	0.13352	0.0028365
820	9.0013e-08	1036	0.089025	0.13352	0.0028365
821	7.5636e-08	1037	0.089025	0.13352	0.0028365
822	7.0322e-08	1038	0.089025	0.13352	0.0028365
823	6.4295e-08	1039	0.089024	0.13352	0.0028365
824	5.0632e-08	1040	0.089024	0.13352	0.0028365
825	4.5944e-08	1041	0.089024	0.13352	0.0028365
826	1.5002e-08	1042	0.089024	0.13352	0.0028365
827	0.0000e+00	1043	0.089024	0.13352	0.0028365

FIGUUR 27 - DE COMPLEXITEIT VAN HET MODEL

Met de waarden van de complexiteit kan gespeeld worden. Door nul in te vullen, kan de complexiteit geanalyseerd worden. In Figuur 27 is een grafiek te zien met op de onderste x-as de waarden voor de complexiteit, op de bovenste x-as het totaal aantal splitsingen in de boom en op de y-as de *xerror*. De waarden van de xerror is de kruisvalidatiefout, die in het programma zit ingebouwd. Dit zijn schattingen van de gevalideerde voorspellingsfout voor het aantal splitsingen in de boom.

Verder is er naast de grafiek ook een tabel met getallen van de complexiteit te zien, waar ook de relatieve error is weergegeven. De correlatiecoëfficiënt is in het literatuuronderzoek al naar voren gekomen en wordt aangeven met de letter R. De relatieve error is 1 - R^2. De correlatiecoëfficiënt geeft informatie over de mate waarin een model de werkelijke data benadert. Dus een relatieve error van 0.2 geeft een R^2 0.8, wat betekent dat het model met 80% zekerheid de ingevoerde data kan voorspellen. De xerror is een objectievere maat voor de nauwkeurigheid van de voorspelling. Dat is ook de reden dat deze over het algemeen hoger is dan de relatieve error. De formule voor het berekenen van de nauwkeurigheid van het model is dus als volgt te berekenen:

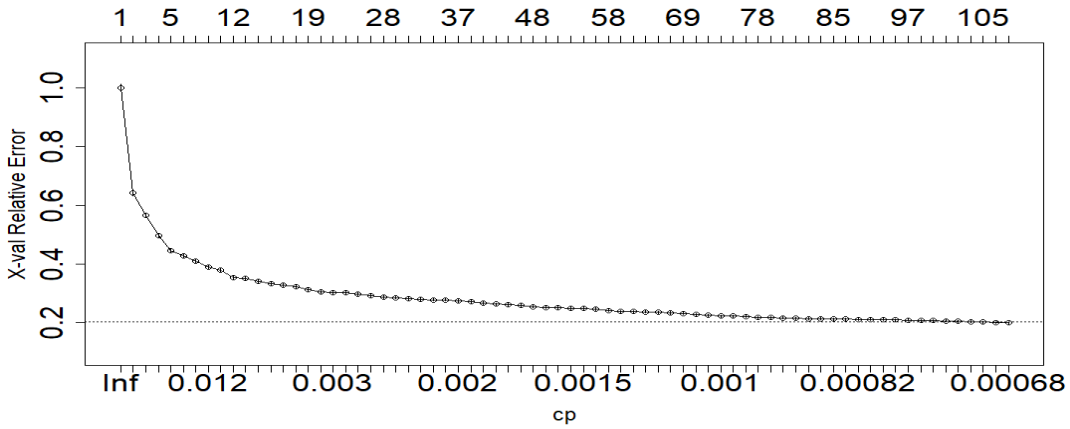
$$Voorspellingsnauwkeurigheid = (1 - \text{xerror}) * 100\%$$

In de tabel van Figuur 27 is te zien dat de kleinste waarden voor de complexiteit geen significant verschil meer maken. Uiteindelijk convergeert de grafiek naar een xerror van 0.13352, wat betekent dat het model in dit geval maximale een voorspellingsnauwkeurigheid heeft van:

$$\begin{aligned} Voorspellingsnauwkeurigheid &= (1 - xerror) * 100\% \\ &= (1 - 0.13352) * 100\% \\ &= 86.7\%\% \end{aligned}$$

De maximale nauwkeurigheid van het model is op dit moment 86.7%. Omdat de directe gebruiker van het model nog niet bekend is, wordt het model in eerste instantie gemaakt voor Goudappel Groep zelf. Er zijn daarom geen eisen aan welk percentage het model moet voldoen, daarom wordt er voor nu naar een betrouwbaarheid gestreefd van 80%. Later kan dit makkelijk aangepast worden in het model, mocht dit nodig zijn als de toepassing bekend is.

De complexiteit heeft een grote invloed op de mate van nauwkeurigheid van het model. Zo was het in Figuur 27 te zien dat een kleine waarden voor de complexiteit zorgt voor een nauwkeuriger model. In deze lijst is gekeken naar de waarden waarvoor de nauwkeurigheid 80% is, oftewel bij een xerror van 0.20. Ook moet er rekening gehouden worden met het overfitten van het model. Dit houdt in dat er zoveel specifieke verbanden worden gevonden in de train dataset dat dit juist onnauwkeurigheden geeft in de test dataset. Deze relaties zijn toevallig aanwezig in de train dataset, maar dit zijn verbanden weer waarop voorspellingen gebouwd kunnen worden. Dit heeft veel invloed op de nauwkeurigheid van de voorspelling. Om dit te voorkomen wordt de optimale complexiteit waarde gehanteerd. In Figuur 28 is dit duidelijk te zien door de lijn die erbij is getekend. Zodra de punten net onder deze lijn duiken is de optimale complexiteit waarde benaderd. Uit de lijst en deze grafiek komt een complexiteit waarde van 0.00068 naar voren met een xerror van 0.19936.



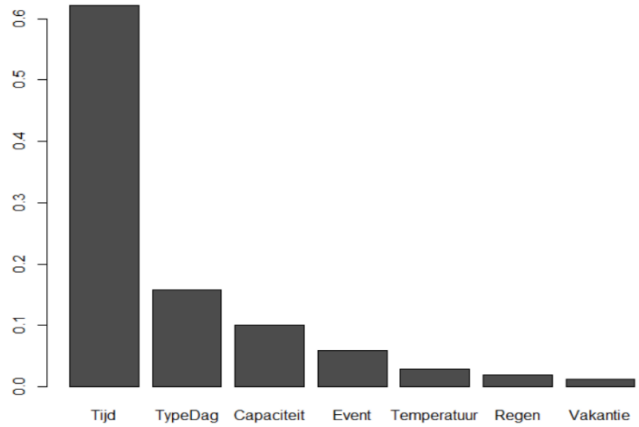
FIGUUR 28 - DE COMPLEXITEIT TEGENOVER DE XERROR

De grafiek laat zien dat de eerste splitsingen in de boom en dus de waarden van de complexiteit een groot verschil maken in de nauwkeurigheid van het model. Hoe kleiner de complexiteit hoe dichter de punten bij elkaar liggen en dus hoe minder invloed de veranderingen hebben op de nauwkeurigheid. De lijn bij een xerror van 0.20 ligt net boven het midden van het laatste bolletje, wat bewijst dat voor een xerror van 0.2 een complexiteit van 0.00068 nodig is. Daarnaast is ook te zien dat dit resulteert in 108 splitsingen in de regressieboom. De nauwkeurigheid van het model waar verder mee wordt gegaan is dus als volgt:

$$\begin{aligned} Voorspellingsnauwkeurigheid &= (1 - xerror) * 100\% \\ &= (1 - 0.19936) * 100\% \\ &= 80.1\%\% \end{aligned}$$

5.3.4 De invloed van variabelen

In de formule zijn alle variabelen beschreven die mee worden genomen bij het voorspellen van de bezettingsgraad van de parkeergarages. Voor elke variabele geldt dat deze een andere invloed heeft op de bezettingsgraad. In Figuur 29 is de kracht van elke variabelen weergegeven als percentage van het totaal. De variabele tijd is een grote uitschieter met een invloed van 62%. Dit is ook te verwachten, omdat er grote verschillen zitten in de bezettingsgraad per tijdstip. Daarnaast is te zien dat de temperatuur, regen en vakantie weinig invloed hebben. De temperatuur en regen hebben weinig invloed, omdat het een relatief kleine dataset is waarin geen grote en duidelijke verschillen zitten. Als er meer seizoenen met elkaar vergeleken zouden worden, dan zou de invloed van deze variabelen ook zeker hoger zijn.

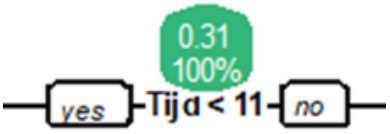


FIGUUR 29 - DE INVLOED VAN VARIABELEN ALS PERCENTAGE

5.4 De regressieboom

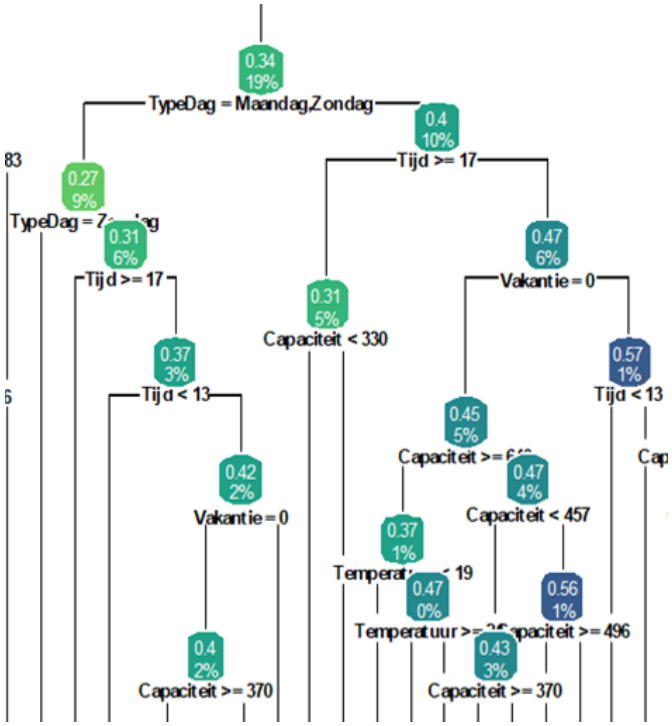
De formule is geschreven, de variabelen zijn gedefinieerd en de complexiteit van het model is vastgesteld, dus de regressieboom is gemaakt alleen nog niet gevisualiseerd. Zoals in Figuur 28 is te zien, zijn er voor deze boom 108 splitsingen nodig daarom is de boom moeilijk in zijn geheel weer te geven, zie Appendix J voor de regressieboom. Om de boom te kunnen toelichten zijn er delen van de boom in deze paragraaf toegevoegd. De eerste geeft het begin van de boom weer, gevolgd door een gedeelte van het midden van de boom en tot slot een gedeelte van het einde van de boom. De kleuren die in de boom zijn weergegeven staan voor de waarden van de bezettingsgraad. Hoe lager de bezettingsgraad, des te lichter het blokje. Hoe hoger de bezettingsgraad, des te donkerder het blokje.

Het begin van de boom is in Figuur 30 weergegeven. In het groene blokje zijn twee getallen geschreven. Het bovenste getal geeft aan dat de gemiddelde bezettingsgraad van de waarnemingen in die klasse 0.31 is. Het percentage eronder geeft aan hoeveel waarnemingen er van de totale dataset in deze klasse zitten. Dit is het begin van de boom, dus alle waarnemingen van de dataset zitten in deze klasse (100%). Vervolgens wordt de eerste splitsing gemaakt op basis van de variabele tijd. Dit is logisch, omdat in de vorige paragraaf te zien was dat de tijd overduidelijk de meeste invloed heeft op de bezettingsgraad. Onder het groene blokje staat de voorwaarden van de splitsing. Bij elke splitsing zijn er twee opties om te antwoorden op de voorwaarden: *Ja* en *Nee.* In het vervolg van de regressieboom zijn deze twee opties niet meer weergegeven, maar bij elke splitsing geldt: naar links is ja en naar rechts is nee.



FIGUUR 30 - HET BEGIN VAN DE REGRESSIEBOOM

In een middenstuk van de regressieboom in Figuur 31 is goed te zien dat er na elke splitsing minder waarnemingen overblijven in de volgende klassen. Door de vele splitsingen in de boom, zijn niet alle splitsingen even goed te lezen. Alle paden die door de boom worden beschreven zijn ook uit te lezen in tekst. De voorwaarden in de bovenste splitsing is gebaseerd op de variabelen type dag. Als het een maandag of zondag is, dan gaat het pad verder naar links, is het een andere dag dan maandag of zondag, gaat het pad naar rechts. Stel het is een donderdag, dan is voor deze waarneming de volgende voorwaarden of de tijd groter of gelijk is aan 17. Zo worden alle splitsingen beoordeeld en wordt het pad steeds langer. Het is ook goed te zien dat de splitsingen ook veel invloed hebben op de gemiddelde bezettingsgraad van de waarnemingen in die klassen. Door alleen de eerste beslissing zit er al een verschil in bezettingsgraad van 0.27 tot 0.40.



FIGUUR 31 – EEN DEEL VAN DE REGRESSIEBOOM

Het einde van de boom is een lange rij met klassen waarin de hele train dataset is verdeeld. Alle paden beginnen dus in dezelfde klassen en eindigen onderaan in de boom in verschillende klassen, zie Figuur 32. In sommige klassen is beschreven dat 0% van de waarnemingen in die klasse terecht zijn gekomen. Dit betekent niet dat er geen waarneming in die klasse zit, maar het percentage wordt afgerond naar een heel getal. Zo kan het bijvoorbeeld zijn dat er in één klasse 0.21% van alle waarnemingen zitten, wat inhoudt dat er 26 waarnemingen van het totaal in de klasse zijn beland. Dit percentage wordt echter wel afgerond naar beneden, waardoor er 0% komt te staan. Ook hier geldt hoe lichter het blokje hoe lager de gemiddelde bezettingsgraad en andersom hoe donkerder het blokje hoe hoger deze gemiddelde bezettingsgraad. Deze uiteindelijke klassen zijn ook de uitkomst van de regressieboom, omdat aan de hand van de paden de voorspellingen gegenereerd worden.

FIGUUR 32 - HET EINDE VAN DE REGRESSIEBOOM

5.5 Conclusie

In dit hoofdstuk is de ontwikkeling van de regressieboom stap voor stap doorgenomen. De totale dataset is opgesplitst in verschillende groepen van een train en een test dataset. De train dataset is ingelezen en de verschillende datatypen zijn behandeld. Ook is er gekeken naar verschillende aspecten die invloed hebben op de prestatie van het model. De formule voor de regressieboom is opgesteld en de uiteindelijke boom is toegelicht. Kortom, de volgende vragen zijn beantwoord:

- Hoe kan het proces worden weergegeven in een model?
  - Hoe wordt de data opgesplitst in een train en test dataset?
  - Wat zijn de in- en output variabelen van het model?
  - Hoe komt het model en de visualisatie hiervan tot stand?

Een model wordt gebaseerd op een bepaalde dataset. Om vervolgens te controleren of er ook echt voorspellingen gedaan kunnen worden met het model, moet het getest worden op een andere dataset. Daarom is er een onderscheid tussen een train dataset, waarmee het model gemaakt wordt, en een test dataset om het model te testen. De splitsing in de dataset is gemaakt op basis van hele dagen. In totaal zit er 93,25% van de totale gegevens in de train dataset en de overige 6,25% zit in de test dataset. Er is gecontroleerd of de verschillende variabelen in beide datasets vergelijkbaar zijn, zodat dus alle informatie in het model zit, maar ook getest kan worden. Er zijn drie verschillende combinaties van train en test datasets gemaakt, om uit te sluiten dat de uitkomsten berusten op toeval.

Om de regressieboom te kunnen beoordelen is goed om deze resultaten te vergelijken met een andere methode. Daarom worden er ook voorspellingen gegenereerd doormiddel van een eenvoudige methode, die gebruik maakt van het gemiddelde van de twee belangrijkste variabelen. De variabele tijd en het type dag worden gebruikt om de voorspellingen te genereren. Door het vergelijken van deze resultaten kan er worden vastgesteld of een complexere methode in dit onderzoek beter presteert of dat een eenvoudige methode al voldoet aan de eisen.

De train datasets is vervolgens in het software programma R ingelezen. In totaal zijn er vier verschillende datatypes in de gegevens aanwezig. Als eerste komt het type *factor* voor, die aangeeft dat de input van de variabelen tekst bevat, zoals het type dag van de week. Daarnaast is er het type *integer*. Dit type geeft aan dat het alleen maar om hele getallen gaat, zoals bij de capaciteit van de parkeergarages. Ook zijn er variabelen die weergegeven worden met een kommagetal, zoals de bezettingsgraad. Deze variabelen worden weergeven met het type *numeric.* Tot slot is er het type *logical*, die aangeeft dat het om een binaire variabelen gaat. Dit zijn de twee variabelen vakantie en speciale gebeurtenis, die alleen 0 of 1 kunnen aannemen.

De volgende stap naar de regressieboom is het schrijven van de formule die deze boom berekend en ontwikkelt. In deze formule wordt de variabelen bezettingsgraad als eerste genoemd, gevolgd door alle variabelen die de bezettingsgraad beschrijven. Ook wordt er aangegeven dat het algoritme van de regressieboom moet worden gebruikt. Tot slot kan ook een waarde worden ingevuld voor de mate van de complexiteit die gehanteerd moet worden. De complexiteit heeft invloed op de grote van de boom en ook op de nauwkeurigheid van de voorspellingen. Echter met een te hoge complexiteit kan het ook zijn dat er verbanden worden gelegd in de train dataset, die eigenlijk niet met elkaar in verband gelegd kunnen worden. Naast de complexiteit kan er ook gekeken worden naar de invloed van de verschillende variabelen. De variabele tijd is ver uit het belangrijkste met 62%, gevolgd door het type dag met 16%. De weer variabelen hebben een lage invloed, wat verklaard kan worden door het feit dat er eigenlijk maar 1 seizoen is meegenomen in de gegevens.

Doormiddel van de formule zijn alle takken in de boom berekend met kleinste som van de kwadratische fout. De boom kan gevisualiseerd worden, maar is lastig om goed in kaart te brengen. In totaal zijn er 108 splitsingen in de boom, dat wil zeggen er is 108 keer een keuze is gemaakt om naar link of naar rechts te gaan. Deze paden in de boom zijn uiteindelijk de uitkomst en leiden naar de voorspellingen. De boom begint met alle waarnemingen in één klasse met daarvan de gemiddelde bezettingsgraad. Per splitsing worden de waarnemingen verdeeld over de twee volgende klasse en opnieuw worden van die klassen het gemiddelde berekend. Zo eindigt de boom met veel verschillende klassen met ieder een aantal waarnemingen en hun eigen gemiddelde bezettingsgraad. Voor een nieuwe waarneming wordt er vanaf het begin van de boom keuzes gemaakt totdat de opgegeven complexiteit is bereikt om er zeker van te zijn dat overfitten niet plaatsvindt. Dit einde is een klasse waarin de voorspelde bezettingsgraad staat beschreven voor die waarneming.

Hoofdstuk 6 – Resultaten

De eenvoudige techniek en de regressieboom zijn ontwikkeld, maar hoe goed representeren deze methoden de werkelijke bezettingsgraad van parkeergarages? In dit hoofdstuk zullen er verschillende analyses gedaan worden op de train en test dataset. Uit deze analyses moet naar voren komen hoe goed de modellen presteren. De onderzoeksvraag voor dit hoofdstuk is daarom:

- Wat is de prestatie van het instrument op de train en test datasets?

Eerst wordt de prestatie bekeken van het eenvoudige model met als input de train dataset (§6.1). Hierin wordt er gekeken naar de regressielijn en de residuen. Vervolgens zullen dezelfde stappen worden toegepast op de regressieboom om te zien welke methode het beste presteert (§6.2). Het model dat als beste naar voren komt, wordt gebruikt om de test dataset te voorspellen. Ook hierop worden de analyses toepast om te kijken hoe het model presteert op deze nieuwe gegevens (§6.3). Deze voorspellingen zullen gevisualiseerd worden om in te zoomen op de afwijkingen (§6.4). Tot slot wordt er aan het eind van dit hoofdstuk een conclusie gevormd waarin de onderzoeksvraag wordt beantwoord (§6.5).

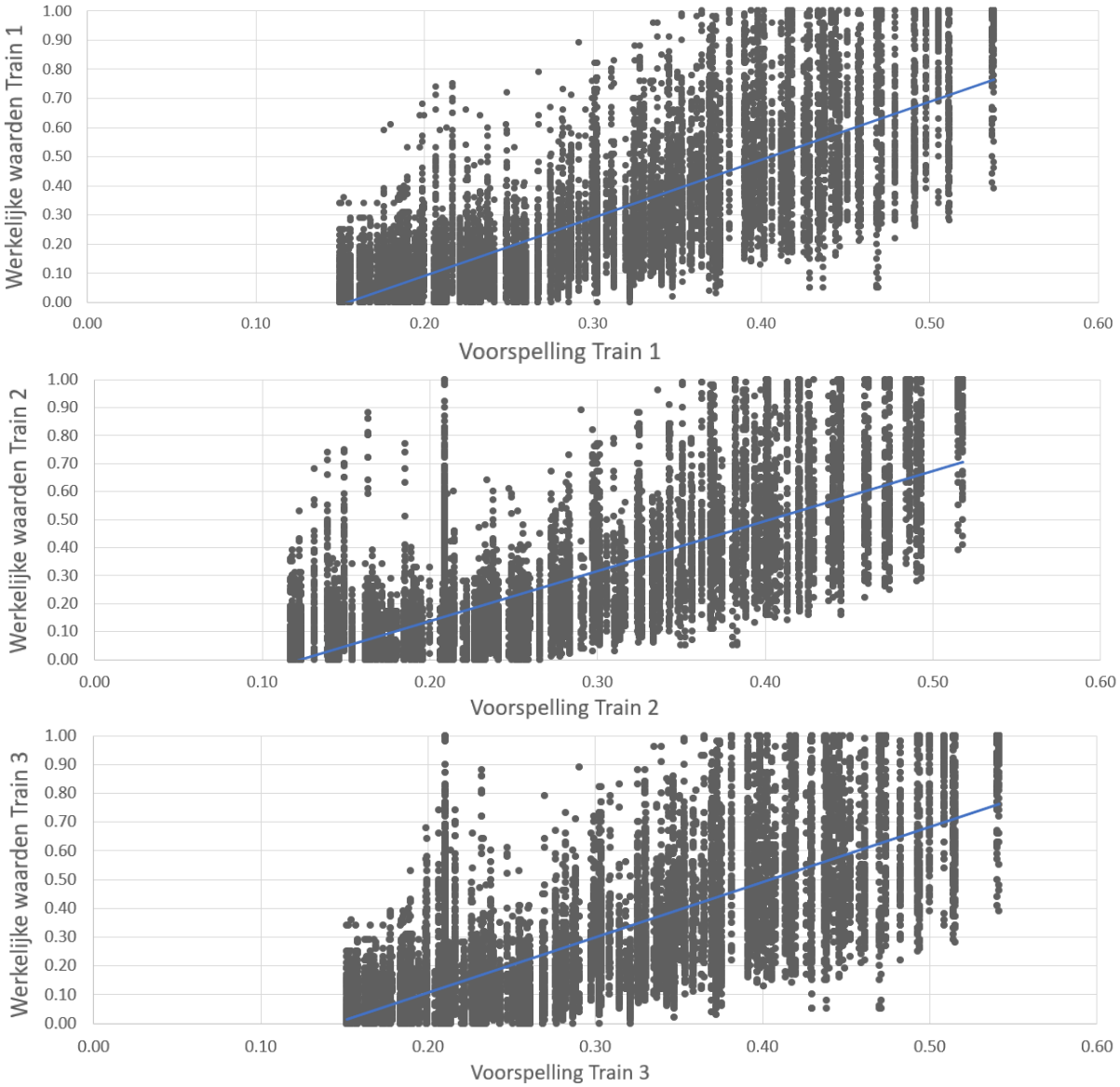
6.1 Eenvoudige techniek

Zoals in hoofdstuk 5 aan bod is gekomen, is er een methode gebruikt om eenvoudig voorspellingen te genereren aan de hand van het gemiddelde van twee belangrijke onafhankelijke variabelen. Deze methode kan aantonen of het zinvol is om met complexere modellen te werken of dat een eenvoudig model al een accurate voorspelling kan genereren. De gemiddelde waarden zijn berekend met behulp van de train dataset. Daarom wordt er eerst gekeken naar de betrouwbaarheid en de nauwkeurigheid van het model als er voorspellingen worden gedaan op de train dataset. Het beoordelen van het model gebeurt aan de hand van twee verschillenden methoden. Eerst wordt er gekeken naar de samenhang tussen de werkelijke en de voorspelde waarden. Vervolgens wordt ook gekeken naar de afwijkingen hiervan en hoe deze afwijkingen beoordeeld kunnen worden.

6.1.1 De regressielijn

Het doel van een voorspelling is dat deze waarde zo dicht mogelijk bij de waarde van de werkelijkheid ligt. Daarom is het belangrijk om in een grafiek te kijken naar de samenhang tussen de werkelijkheid en de voorspelling. De lijn die door deze puntenwolk getekend kan worden is de regressielijn. De best passende lijn wordt bepaald door het totaal van de gekwadrateerde afwijkingen, verticaal gemeten. Hierdoor ligt de lijn precies op de punten waar het totaal van de afwijkingen tussen de voorspelling en de werkelijkheid minimaal is. Uiteindelijk geldt er hoe dichter de punten uit de hele puntenwolk bij de lijn liggen, des te beter de voorspelling.

Er zijn drie verschillende verdelingen gemaakt in de totale dataset. In totaal zijn er dus drie groepen met elk een train en test dataset waarmee het model berekend en getest kan worden. Dit is gedaan om er zeker van te zijn dat de uitkomsten van een model niet op toeval berusten. Als elke groep ongeveer dezelfde betrouwbaarheid en dezelfde resultaten levert, dan is het model consistent te noemen. Voor elke train dataset is de simpele techniek toegepast, daarom wordt er in de vergelijking telkens gekeken naar drie verschillende resultaten. De opbouw van het model is hetzelfde alleen de input gegevens zijn dus verschillend. In Figuur 33 zijn drie grafieken te zien waar de werkelijke waarnemingen uitgezet zijn tegenover de voorspelde waarde. De drie verschillende input gegevens worden respectievelijk weergegeven met de naam: Train 1, Train 2 en Train 3. In elke train dataset zijn ongeveer 12000 waarnemingen. Elke stip in de grafiek geeft een van deze waarnemingen weer. Omdat het over de bezettingsgraad gaat en dit een percentage is kunnen de punten niet hoger uitkomen dan 1, zowel op de x-as als de y-as. De bijbehorende regressielijn is in het blauw te zien.



FIGUUR 33 - VOORSPELLING VERSUS WERKELIJKHEID

Om de grafiek te kunnen beoordelen, moet het duidelijk zijn waar op gelet moet worden en wanneer een model als 'goed' beschouwd kan worden. Volgens het onderzoek van Golbraikh & Tropsha (2002) zijn er drie belangrijke criteria voor de validiteit van een model:

- $R^2 > 0.6$
- Correlatie coëfficiënt R dicht bij 1
- De regressielijn gaat (bijna) door de oorsprong en de richtingscoëfficiënt ligt dicht bij 1

De drie criteria samen vormen het ultieme model, wat bijna onmogelijk is. Daarom wordt er naar gestreefd dat de waarden zo dicht mogelijk in de buurt komen van de waarden in de criteria. De determinatie coëfficiënt, R^2, is een statistische maat die aangeeft welk gedeelte van de totale variantie in de variabele verklaard wordt door het model. Daarom geeft een R^2 van 0% aan dat het model geen enkele variantie van de werkelijk waarnemingen kan verklaren. Daarentegen geeft een R^2 van 100% aan dat het model alle variantie van de werkelijke waarnemingen kan verklaren. Er geld dat hoe hoger deze waarde is, hoe beter de voorspellingen zijn van het model. Daarnaast is de correlatie coëfficiënt in hoofdstuk 3 al naar voren gekomen. De correlatie coëfficiënt meet hoe sterk het verband is tussen twee variabelen. Vanaf R > 0.7 wordt over het algemeen gesproken over een sterk verband.

Voor de analyse van de modellen worden de drie criteria gecontroleerd. In Tabel 7 zijn alle gegevens samengevat per train dataset en per criterium. Al eerder is benoemd dat er in dit onderzoek wordt gestreefd naar een betrouwbaarheid van 80%. De waarde van R^2 ligt gemiddeld rond de 0.57, wat aangeeft dat 57% van de variantie in de bezettingsgraad over de gemeten periode verklaard wordt. Voor het eerste criterium geldt dus dat er niet aan wordt voldaan, omdat de waarden lager zijn dan 0.6. Ook wordt er dus duidelijk niet aan de 80% voldaan, die wordt verwacht voor dit onderzoek. De correlatie coëfficiënt R, is bij elke train dataset net boven de 0.7, de grens die aangeeft dat er gesproken kan worden van een sterk verband. Echter ligt de waarden niet dicht bij 1 wat voor criterium 2 belangrijk is.

Train dataset	Train 1	Train 2	Train 3
$R^2 > 0.6$	$R^2 = 0.5984$	$R^2 = 0.5440$	$R^2 = 0.5583$
R = 1	R = 0.774	R = 0.738	R = 0.747
Regressieformule	Y = 1.9956x - 0.3076	Y = 1.7835x – 0.2188	Y = 1.9204x -0.2763

TABEL 7 - REGRESSIE ANALYSE

Voor het derde criterium wordt er gekeken naar de regressielijnen die door de puntenwolken zijn getekend. De regressieformule voor train dataset 1 ziet er in formule vorm als volgt uit:

$$y = 1.9956x - 0.3076$$

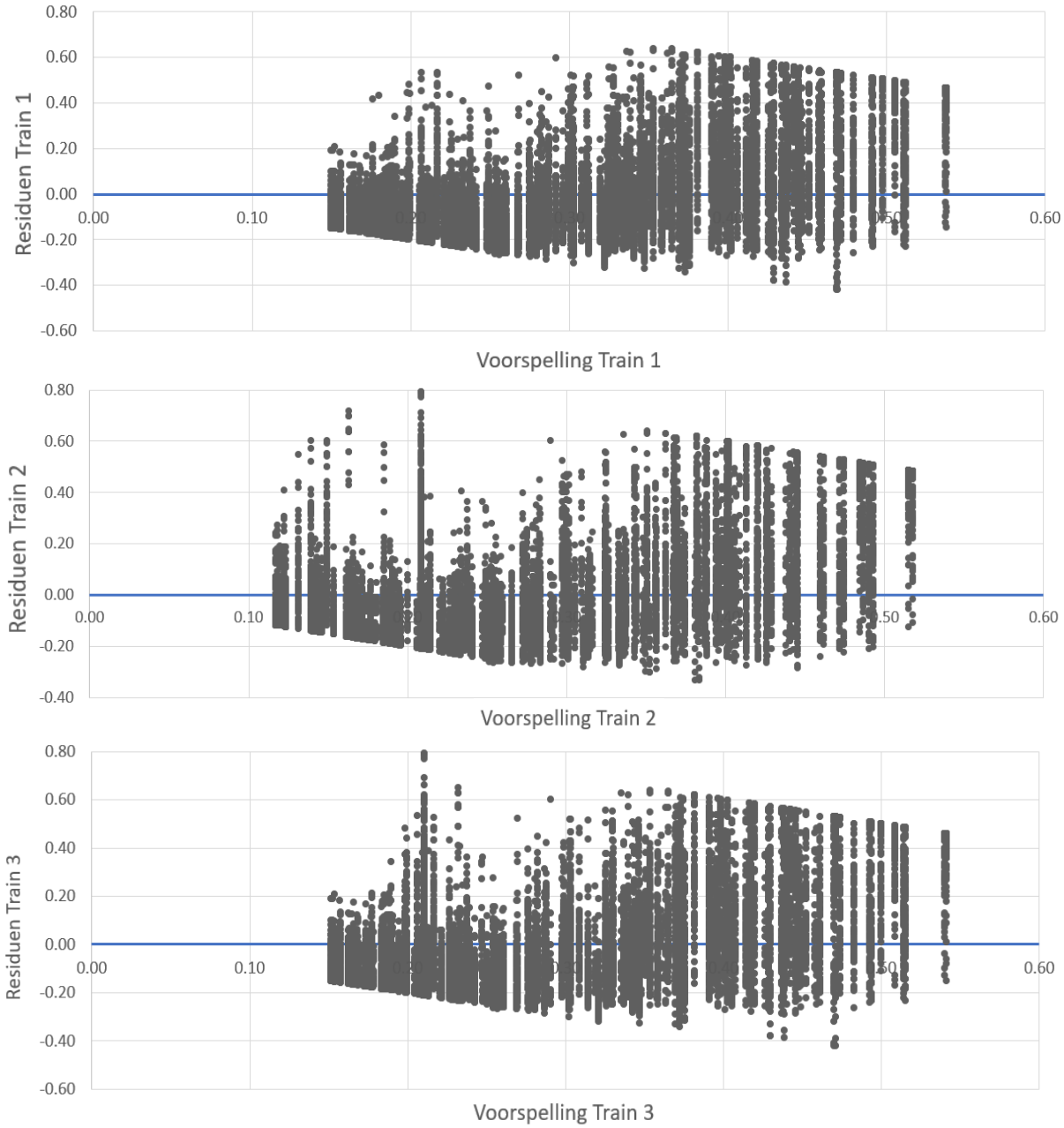
In het optimale geval gaat de regressielijn door de oorsprong en is de richtingscoëfficiënt 1. Dat houdt in dat als de voorspelling 0 is de werkelijke waarneming ook 0 was en als de voorspelling 0.1 is de werkelijke waarde ook 0.1 was en zo verder. Als dit precies waar is, geeft het een voorspelling met een nauwkeurigheid van 100%. In de grafieken in Figuur 33 is al duidelijk te zien dat de lijn niet door de oorsprong gaat. Om dit te controleren wordt x vervangen door 0, waardoor de y-waarden -0.3076 is. Daarnaast is de richtingscoëfficiënt 1.9956 van train dataset 1, wat ook niet dicht bij 1 ligt.

De eenvoudige methode is uitgevoerd op drie verschillende train dataset, waarvan geen enkele voldoet aan alle drie de criteria. Op basis hiervan kan er gezegd worden dat deze methode geen goede voorspellingen kan genereren. Toch is het van belang om ook nog andere metingen uit te voeren die ook de kwaliteit van het model beoordelen.

6.1.2 Residuen

Naast het kijken naar de regressielijn is het ook mogelijk om te kijken naar de residuen. Dit is een maat om de afwijking te berekenen tussen de werkelijke bezetting en de voorspelde bezetting. De residuen zijn een gevolg van onwillekeurige en onbeheersbare invloeden, ook wel ruis genoemd. Om de bezetting goed te schatten, moeten de residuen normaal verdeeld zijn met een gemiddelde van nul. Dit is een belangrijke controle voor de validiteit van het model.

In Figuur 34 zijn per train dataset de residuen uitgezet tegenover de voorspelde waarden. Elk punt geeft een waarde weer die berekend is door de voorspelde bezettingsgraad van de werkelijke bezettingsgraad af te trekken. Alle punten in één verticale lijn hebben dezelfde voorspelde bezettingsgraad, maar de afwijking is anders door de verschillen in de werkelijke waarden. In een optimaal scenario zouden alle waarden op de blauwe nul lijn liggen, omdat alle voorspelde waarden dan precies gelijk zijn aan de werkelijke waarnemingen. De grafieken lijken veel op elkaar. Het valt op dat er rechtsboven en linksonder een duidelijke lineaire trend aanwezig is. Dit is te verklaren door het feit dat bijvoorbeeld een voorspelde bezettingsgraad van 0.8 geen grotere afwijking kan hebben dan 0.2. Het zijn immers percentages en die kunnen niet boven de 100% uitkomen. Dus de voorspelde waarde plus het residu kan nooit groter zijn dan 1.



FIGUUR 34 – VOORSPELLING VERSUS RESIDUEN

Het belangrijkste wat er te zien is in de grafieken is dat er geen duidelijke patroon te vinden is. Residuen moeten op het eerste gezicht random lijken, zodat ze niet te verklaren zijn of met een andere variabelen te voorspellen zijn. Het gedrag van de residuen in alle drie de modellen is vrijwel gelijk.

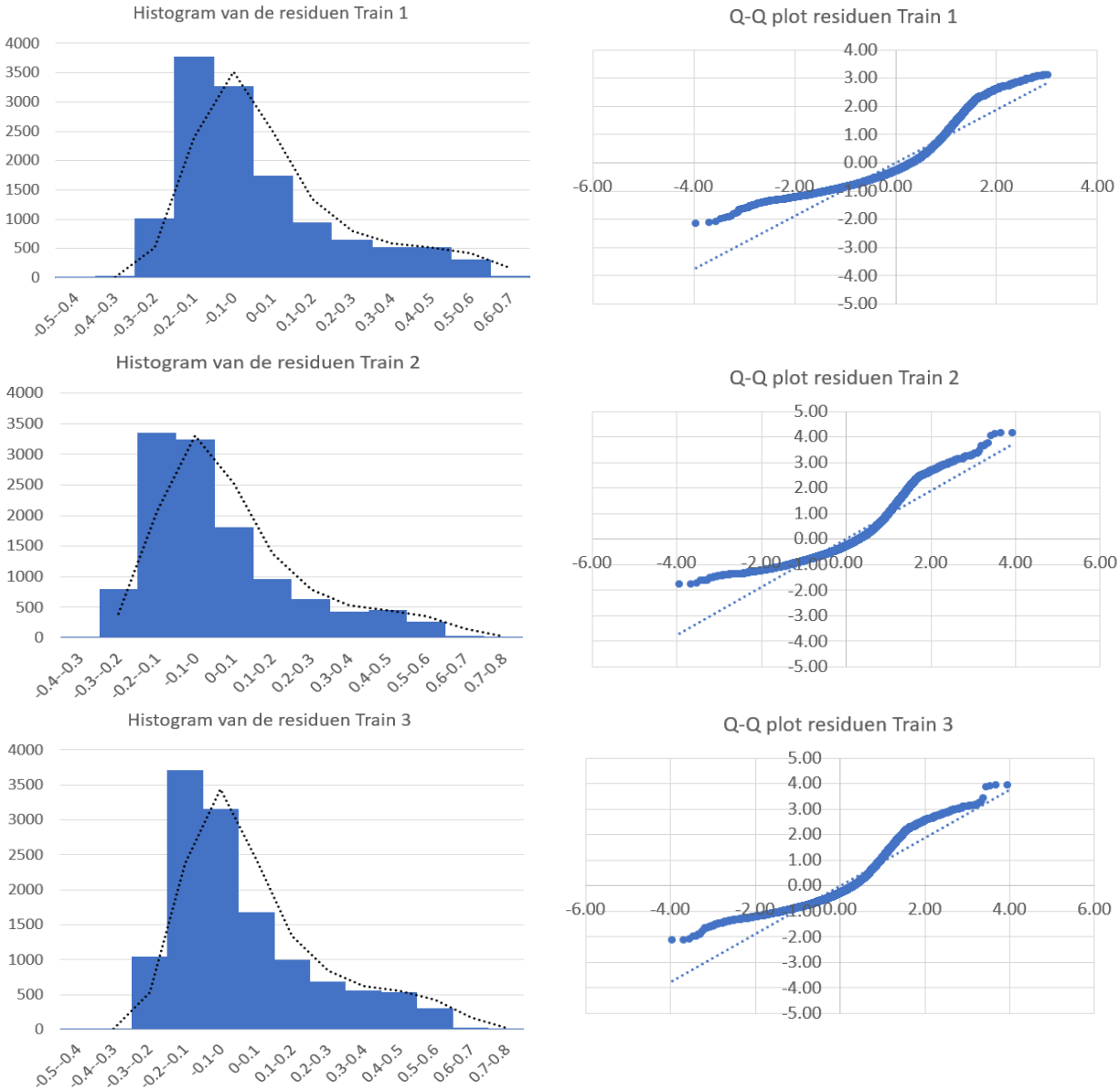
6.1.3 Normaal verdeling residuen

Zoals al kort genoemd, moeten de residuen naast een random patroon ook normaal verdeeld zijn met een gemiddelde van nul. De belangrijkste reden hiervoor is dat het model dan enige consistentie kent. Als de afwijkingen niet normaal verdeeld zijn, betekent dit dat niet alle trends in de dataset verklaard zijn. Bij voorspellingen zal er altijd enige ruis aanwezig blijven, maar onbekende trends horen niet in de dataset achter te blijven. In Tabel 8 is de statistische samenvatting weergegeven van de residuen per train dataset. Het gemiddelde is alleen nul bij train 1, echter liggen de andere gemiddelden wel dicht bij nul. Verder zijn er een aantal verschillen te zien in de waarden.

Residuen	Minimum	Q1	Mediaan	Gemiddelde	Q3	Maximum
Train 1	-0.41929	-0.14075	-0.05795	0.00000	0.08368	0.63708
Train 2	-0.33343	-0.12888	-0.04889	0.00485	0.08132	0.79132
Train 3	-0.42103	-0.14047	-0.05648	0.00456	0.09518	0.78953

TABEL 8 - STATISTISCHE SAMENVATTING VAN DE RESIDUEN

De maximale waarde die een afwijking kan aannemen in dit geval is 1 en de minimale waarde is -1. Dit zou voorkomen als de voorspelde bezettingsgraad 0.00 is en de werkelijke bezettingsgraad 1.00 of andersom. Het spreidingsinterval van train 2 is het grootste en loopt van -0.33 tot 0.79. Op een schaal van -1 tot 1 zijn dit grote afwijkingen en lijkt het model daarom niet goed. Echter, kan het voorkomen dat dit interval wordt gevormd door enkele uitschieters. In Figuur 34 is te zien dat er weinig punten zijn boven de 0.6. Om dit duidelijker in beeld te krijgen zijn in Figuur 35 de drie bijbehorende histogrammen weergegeven met daarnaast een Q-Q plot. Alle genoteerden residuen tussen 0.00 en 0.10 bevinden zich in één klassen. Vervolgens is geteld hoeveel residuen er in één klasse zitten en deze zijn weergegeven in de histogrammen.



FIGUUR 35 – HISTOGRAMMEN EN Q-Q PLOTS VAN DE RESIDUEN

Uit de statistische samenvatting in Tabel 7 was al te zien dat de verdelingen erg op elkaar lijken. Ook de drie histogrammen en Q-Q plots lijken veel op elkaar. Het is te zien dat er meer waarnemingen rechts van nul worden geteld dan links en dat de waarden boven nul ook groter kunnen zijn. Dit verschijnsel wordt ook wel rechtsscheef genoemd. Als in een Q-Q plot alle punten op een rechte liggen kan er een normaal verdeling worden vastgesteld. In de Q-Q plots, die naast de histogrammen zijn weergegevens is het ook te zien dat de verdeling rechtsscheef is doordat de punten rechtsboven flink boven de lijn van de normaal verdeling liggen. Toch liggen de punten links onder ook flink boven de lijn, wat er naar verwijst dat het model meer extreme waarden geeft dan verwacht wordt als het model een volledig normaal verdeeld zou zijn.

Daarnaast wordt de bezettingsgraad in het model met grotere waarden onderschat, dan overschat. Een residu is namelijk de werkelijke waarden minus de voorspelling. Te zien is dat er een aantal grote residuen zijn boven nul, wat betekent dat de voorspelling lager uitkwam dan de werkelijke waarden. Op basis van deze grafieken kan er worden gezegd dat het niet zeer aannemelijk is dat de residuen volledig normaal verdeeld zijn.

Twee verschillende manieren zijn gebruikt om te kijken naar de validiteit van het model. Naast de regressielijn is er gekeken naar de nauwkeurigheid van de voorspellingen en ook de afwijking van de voorspellingen. Het model voldoet niet aan alle eerder genoemde criteria en ook de verdeling van de residuen lijkt niet op die van een normaal verdeling. Voor deze eenvoudige techniek kan er dus geconcludeerd worden dat er geen accurate voorspellingen mee gegeneerd kunnen worden.

6.2 Regressieboom

In de vorige paragraaf is gebleken dat een eenvoudige methode er niet toe in staat is om accurate voorspellingen te generen. Daarom zal in deze paragraaf een complexere methode gebruikt worden, namelijk de regressieboom. In de regressieboom worden de bezettingsgraden, die zijn waargenomen, aan de hand van verschillende variabelen in verschillende klassen ingedeeld. Er worden relaties gevonden tussen de verschillende variabelen, waarop voorspelling gebaseerd kunnen worden. Door deze keuzes in het model kan er worden gekeken naar de nauwkeurigheid van de voorspellingen. Voor elke uiteindelijke klasse in de boom is er vanaf het begin een specifiek pad tot het eind waarde waarin de bezettingsgraad van die klasse wordt beschreven. Alle paden samen bevatten alle mogelijke waarnemingen. De voorspellingen worden op basis van deze paden naar verschillende klasse geleid. Zo kan een pad er als volgt eruit zien:

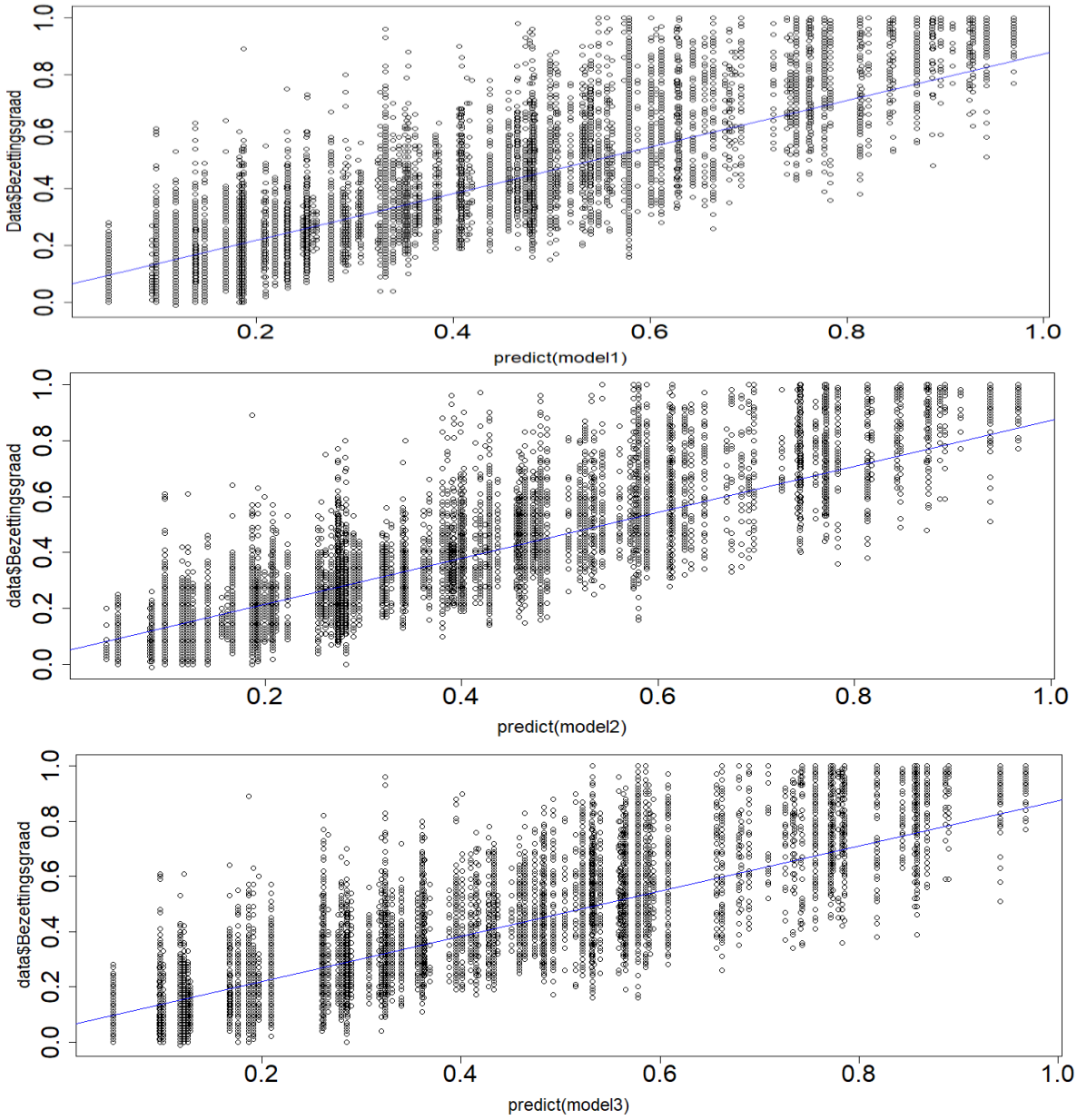
- **Als** de *Tijd* is kleiner dan **4**
- **En** het *Type dag* is **zaterdag**
- **En** de *Capaciteit* is groter dan **330,**
    - **Dan** is de *Bezettingsgraad* gelijk aan **0.183**

Elke waarneming die hieraan voldoet, krijgt dus de voorspelde bezettingsgraad van 0.183. Deze waarde komt voort uit het algoritme, omdat dit het gemiddelde is van alle waarnemingen die aan dit pad voldoen. Het algoritme heeft een relatie gevonden in de waarnemingen die hieraan voldoen, waarmee de som van de kwadratische fout zo klein mogelijk is. Elk pad verschilt in de variabelen en de voorwaarden waaraan deze variabelen moet voldoen. Zo komen uiteindelijk alle waarnemingen in andere klassen terecht. Om de methode te beoordelen worden dezelfde analyses gedaan als in de vorige paragraaf op het eenvoudige model. De samenhang tussen de werkelijke waarnemingen en de voorspellingen zal geanalyseerd worden. Ook de eerder genoemde criteria opgesteld door Golbraikh & Tropsha (2002) zullen gecontroleerd worden. Tot slot zal er gekeken worden naar de verdeling van de afwijkingen tussen de werkelijke en voorspelde waarden.

6.2.1 De regressielijn

De eerste stap is om de relatie tussen de werkelijke waarnemingen en de voorspellingen te analyseren, dit wordt ook wel de regressieanalyse genoemd. De regressieboom wordt geprogrammeerd voor de drie verschillende train datasets, die respectievelijk worden aangegeven met model 1, model 2 en model 3. Vervolgens worden voor dezelfde waarnemingen in de train dataset de voorspellingen gegenereerd. Deze werkelijke waarden en voorspelde waarden zijn in Figuur 36 tegen elkaar uitgezet. Ook hier is in het blauw de regressielijn door de puntenwolken getekend. De lijn ligt precies op de punten waar het totaal van de afwijkingen tussen de voorspelling en de werkelijkheid minimaal is.

Alle verticale verzamelingen van punten geven één klasse weer waarin de waarneming valt. Een groot verschil wat al direct te zien is in de grafiek ten opzichte van het eenvoudige model, is dat de voorspelde waarden op de x-as van 0 tot en met 1 lopen. Dit geeft aan dat er meer variabiliteit in de voorspellingen zitten dan bij het eenvoudige model. Dit wil echter niet zeggen dat de afwijkingen ook minder groot zijn. Voor het analyseren van de regressie worden de drie criteria opnieuw gecontroleerd.



FIGUUR 36 - VOORSPELLINGEN VERSUS WERKELIJKHEID

Het eerste criteria controleert of de determinatie coëfficiënt, R^2, groter is dan 0.6. Deze maat geeft aan of het model de variantie in de werkelijke waarnemingen kan verklaren. Hoe hoger dit getal, des te beter de voorspellingen de werkelijkheid benaderen. In Tabel 9 zijn alle waarden samengevat van de grafieken hierboven. Het valt op dat de waarden voor R^2 hoog zijn. Dat geeft aan dat dit model de variantie in de werkelijke waarden beter kan voorspellen dan het eenvoudige model. Er wordt in dit onderzoek gestreefd naar een nauwkeurigheid van 80%. Model 1 heeft een R^2 van 0.8176, wat aangeeft dat 81.8% van de variantie in de bezettingsgraad over de gemeten periode verklaard wordt. Ook de andere twee modellen hebben een hogere nauwkeurigheid dan 80%, dus alle drie de modellen voldoen aan het eerste criterium.

Train dataset	Model 1	Model 2	Model 3
$R^2 > 0.6$	$R^2 = 0.8176$	$R^2 = 0.8201$	$R^2 = 0.8177$
R = 1	R = 0.904	R = 0.906	R = 0.904
Regressieformule	Y = 0.8176x + 0.0564	Y = 0.8201x + 0.0522	Y = 0.8177x + 0.0565

TABEL 9 – REGRESSIE ANALYSE

De waarde van de correlatie coëfficiënt moet volgens het tweede criterium dicht bij 1 liggen. In Tabel 9 is te zien dat de waarden bijna gelijk zijn voor elk model. Ze liggen allemaal boven de grens van 0.7, wat aangeeft dat er in elk model een zeer sterk verband is tussen de voorspelde en werkelijke bezettingsgraad. Daarom voldoet elk model ook aan het tweede criterium. Voor het derde criterium wordt de formule van de regressielijn geanalyseerd. De regressieformule van model 1 ziet er als volgt uit:

$$y = 0.8176x + 0.0564$$

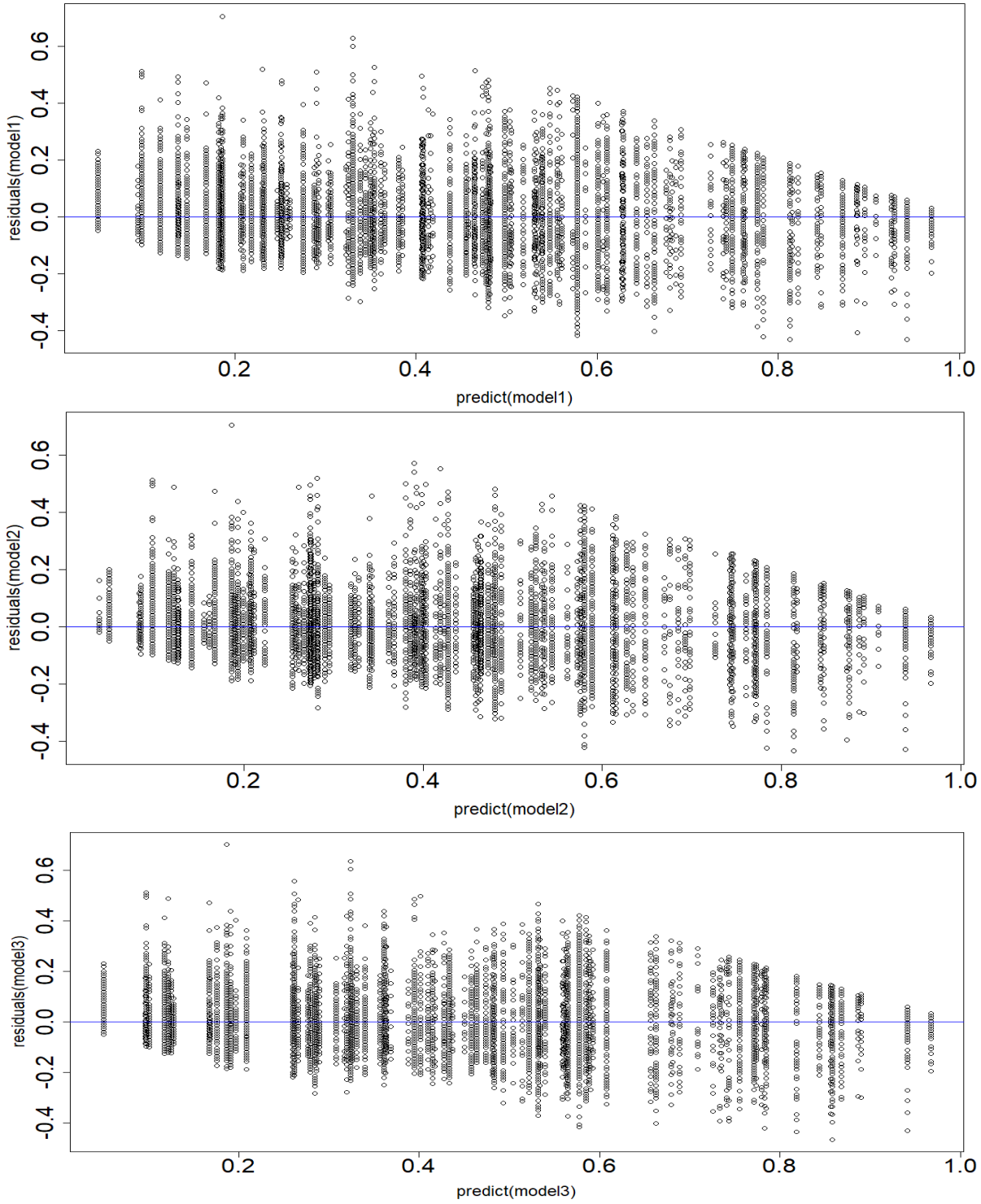
Er moet gelden dat de lijn ongeveer door de oorsprong gaat. Dit is makkelijk te testen door 0 in te vullen voor x wat voor y een waarde van 0.0564 oplevert. Hieruit is te concluderen dat de lijn dicht door de oorsprong gaat, dus aan het eerste gedeelte van het criterium wordt voldaan. Tot slot wordt de richtingscoëfficiënt van de regressielijn bekeken, die ook in de formule wordt weergegeven. De richtingscoëfficiënt van de regressielijn van model 1 is gelijk aan 0.8176, wat relatief dicht bij 1 ligt zoals het criterium vraagt.

Alle drie de modellen voldoen met enige significantie aan alle drie de criteria. Op basis hiervan zou er geconcludeerd kunnen worden dat de modellen goed zijn. Echter moet het model ook getest worden op de verdeling van de residuen.

6.3.2 Residuen

De tweede stap is het analyseren van de afwijkingen tussen de werkelijke waarde en de voorspelling. De afwijking wordt berekend door de werkelijke waarde minus de voorspelling, ook wel residu genoemd. In de verdeling van de residuen komen uitschieters duidelijk naar voren. De residuen moeten een onvoorspelbaar patroon weergeven en daarnaast ook normaal verdeeld zijn met een gemiddelde rond nul. Een voorspelling blijft een benadering, maar in een optimaal geval zouden alle residuen nul zijn. In dat geval is de werkelijke waarde namelijk gelijk aan de voorspelde waarden, precies wat gewenst is. Om de residuen te analyseren zijn in Figuur 37 de waarden van de voorspelling uitgezet tegenvoer de residuen. Ook hier zijn de verschillende test dataset aangegeven met de namen model 1, model 2 en model 3. De blauwe lijn geeft het optimale scenario weer, waarin alle afwijkingen nul zouden zijn. Het is te zien dat de grafieken veel op elkaar lijken. Daarnaast is ook hier de lineaire trend rechtsboven en linksonder duidelijk aanwezig.

Ook hier geldt dat alle punten samen op één verticale lijn samen tot één klasse behoren. Dus elke verticale lijn heeft dezelfde voorspelde bezettingsgraad. De residuen zijn echter anders door de verschillende werkelijke waarnemingen. Ook de blauwe optimale lijn is weergegeven, waarop alle afwijkingen nul zouden zijn. Het is een duidelijk verschil dat de voorspellingen gevarieerder zijn tegenover de eenvoudige methode. Ook de y-as heeft een kleiner interval, wat aangeeft dat de residuen minder grote uitschieters hebben. Waar het interval bij de eenvoudige methode van -0.6 tot met 0.8 liep, loopt de y-as nu van -0.4 tot en met 0.6. Dit is positief, omdat dus de voorspellingen beter overeenkomen met de werkelijke waarnemingen.



FIGUUR 37 – VOORSPELLING VERSUS RESIDUEN

Het belangrijkste wat er te zien is in de grafieken is dat er geen duidelijke patroon te vinden is. Als er wel een patroon in de residuen te vinden is, betekent dit dat er nog informatie in de gegevens zit, die niet in het model zijn vastgelegd. Residuen moeten op het eerste gezicht random lijken, zodat ze niet te verklaren zijn of met een andere variabelen te voorspellen zijn. Het gedrag van de residuen in alle drie de modellen is vrijwel gelijk.

6.3.3 Normaal verdeling residuen

De laatste stap is het controleren of de residuen normaal verdeeld zijn met een gemiddelde rond nul. Als de residuen normaal verdeeld zijn, betekend dit dat het model consistent is en dat alle trends in de dataset verklaard zijn. In Tabel 10 is de statistische samenvatting weergegeven van de residuen per model. Het gemiddelde van elke traindataset is gelijk aan nul. Dit is goed voor het model, omdat dit betekend dat het gemiddelde van alle afwijkingen samen nul is.

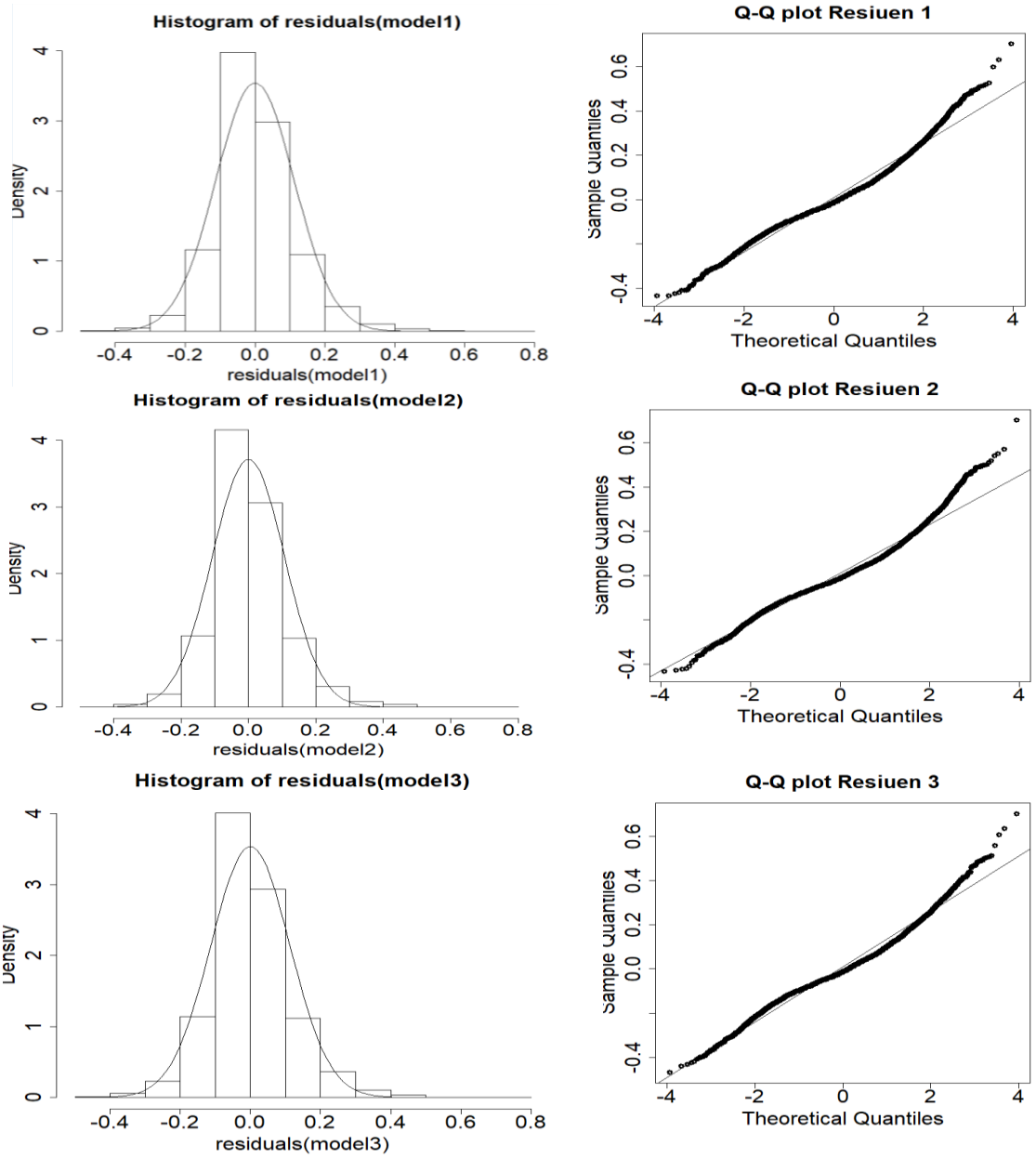
Residuen	Minimum	Q1	Mediaan	Gemiddelde	Q3	Maximum
Model 1	-0.43322	-0.06818	-0.01073	0.00000	0.05968	0.70369
Model 2	-0.43322	-0.06465	-0.01156	0.00000	0.05408	0.70327
Model 3	-0.46774	-0.06792	-0.01196	0.00000	0.05981	0.70339

TABEL 10 - STATISTISCHE SAMENVATTING VAN DE RESIDUEN

Verder is te zien dat de spreiding relatief groot is. Het interval van de spreiding kan lopen van een minimale waarde van -1 tot een maximale waarde van 1. De maximale waarde voor een residu uit de modellen is 0.7. Op een schaal van -1 tot 1 is dit een grote afwijking en daardoor lijkt het model in eerste instantie niet zo goed. Hoewel dit niet waar hoeft te zijn, omdat deze uiterste waarden ook naar voren kunnen komen door eventuele uitschieters. In de grafieken is al te zien dat er weinig punten zijn buiten het interval [-0.4; 0.4]. Om een goed beeld te krijgen en te controleren of de residuen ook normaal verdeeld zijn, zijn in Figuur 38 de histogrammen en bijbehorende Q-Q plots weergegeven.

In Tabel 9 is duidelijk te zien dat de gegevens van de verschillende train dataset veel op elkaar lijken. Hieruit volgen daarom ook drie vergelijkbare histogrammen. Deze drie histogrammen geven in eerste instantie een mooi beeld van een normaal verdeling. De lijn die door het histogram is getekend, geeft een duidelijke piek precies boven 0.0. Daaruit kan geconcludeerd worden dan de meeste residuen een waarde hebben rond 0.0. Verder valt het op de x-as bij alle drie de histogrammen geen symmetrisch interval heeft. Er zijn grotere uitschieters te vinden boven nul dan onder nul, waardoor de verdeling links scheef genoemd kan worden. Echter gaat het om zo weinig waarnemingen dat de residuen nog steeds normaal verdeeld kunnen zijn. Om dit te controleren zijn ok de Q-Q plots weergegeven.

Als in een Q-Q plot alle punten op dezelfde rechte lijn liggen, kan er met zekerheid gezegd worden dat de waarden normaal verdeeld zijn. Er zijn geen grote afwijkingen te zien in de Q-Q plots. Wat wel opvalt is dat er rechts boven enkele stippen afwijken van de rechte lijn. Dit komt voort uit de lichte links scheve verdeling, die ook in de histogrammen te zien was. Voor het model geeft dit aan dat er meer extreme waarden voorspeld zijn dan verwacht wordt als het model een volledig normale verdeling weergeeft.



FIGUUR 38 – HISTOGRAMMEN EN Q-Q PLOTS VAN DE RESIDUEN

Twee verschillende manieren zijn gebruikt om te kijken naar de validiteit van regressieboom. Naast de regressielijn is er gekeken naar de nauwkeurigheid van de voorspellingen en ook de afwijking van de voorspellingen. Aan alle drie de criteria werd voldaan en ook de verdeling van de residuen lijkt sterk op die van een normaal verdeling. Dit is niet direct het bewijs dat het model juist is, echter voldoet het wel aan alle bovenvermelde eisen. Daarom kan er met een redelijke zekerheid worden aangenomen dat het model goed is.

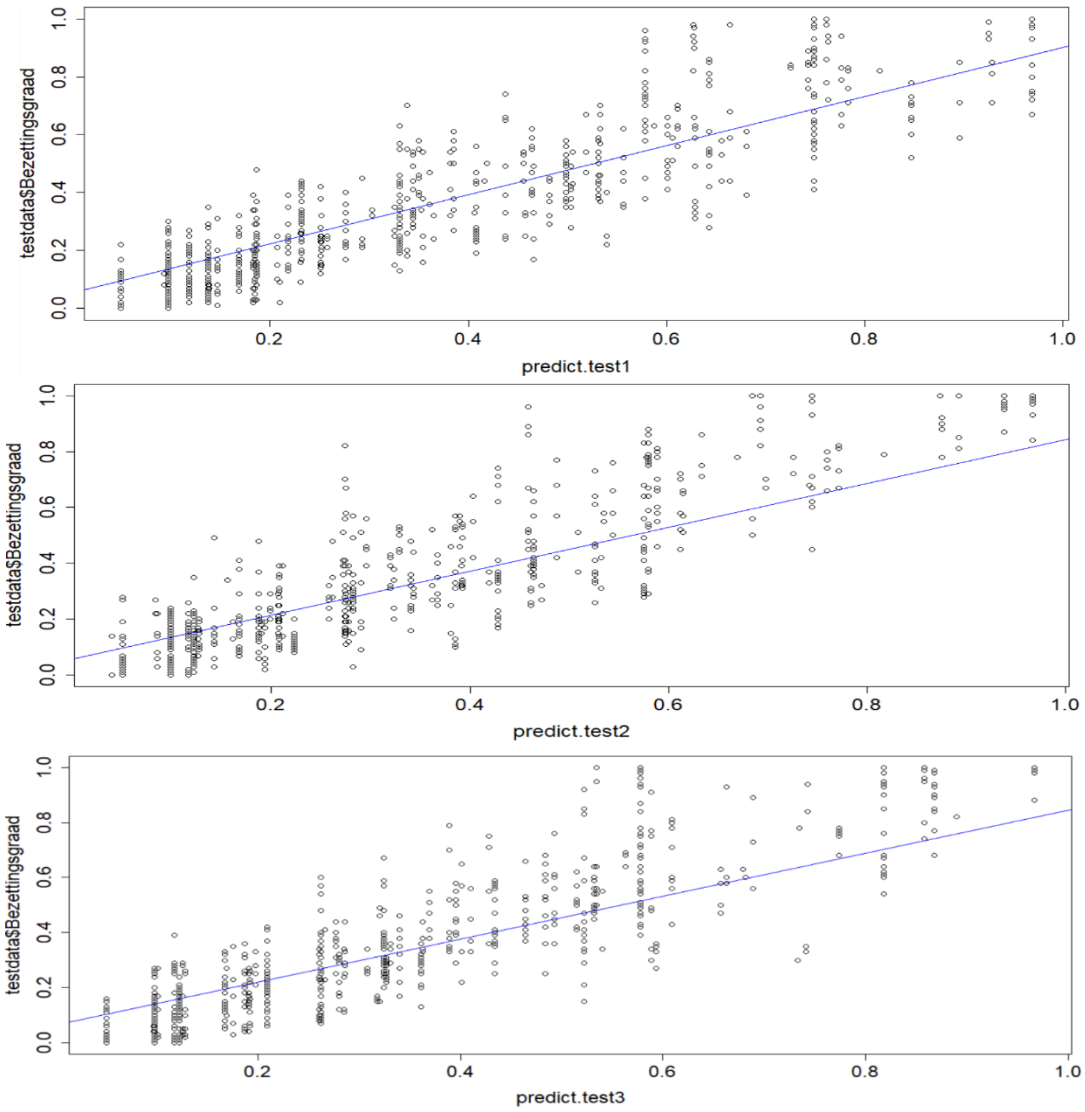
In de vergelijking tussen het eenvoudige model en de regressieboom komt de regressieboom duidelijk als beste naar voren. Dit model heeft een nauwkeurigheid van ruim 80% en alle afwijkingen lijken sterk normaal verdeeld te zijn. Omdat het zo duidelijk naar voren komt dat de regressieboom beter presteert, wordt dit model gebruikt in de verdere resultaten. Of het model deze hoge betrouwbaarheid behoudt als er nieuwe gegevens worden gebruikt, wordt in de volgende paragraaf geanalyseerd aan de hand van dezelfde technieken. Zo kunnen ook de gegevens tussen de train dataset en de test dataset met elkaar vergeleken worden.

6.3 Validatie Regressieboom

Verschillende metingen en analyses zijn gedaan op de train dataset. Het model is gemaakt vanuit de traindataset, dus daarom kan er worden verwacht het model goed presteert op bovenstaande eisen. Het is daarom van belang dat het model ook getoetst wordt op de test dataset. Deze gegevens zaten niet in het model tijdens de ontwikkeling en daarom zijn het nieuwe waarnemingen waarop het model getest kan worden. Voor het vergelijken van het model worden precies dezelfde stappen herhaalt als hiervoor. Te beginnen bij de grafieken waarin de werkelijkheid is uitgezet tegenover de voorspellingen gevolgd door het analyseren van de afwijkingen.

6.3.1 De regressielijn

Voordat de regressielijn getekend kan worden, moet het model eerst gedraaid worden op de test dataset. Het model doorloopt precies dezelfde paden als bij de train dataset. Zo wordt er voor elke regel van de waarnemingen gekeken in welke klasse deze regel terecht komt. Alle mogelijke scenario's zitten in de regressieboom verwerkt en dus worden al deze nieuwe combinaties zonder enige problemen aan de klassen toegekend. Ook hier geld dat elke verticale lijn van punten één klasse weergeeft. Op de horizontale as staan de voorspelde waarden en op de verticale as staan de werkelijke waarnemingen. In het blauw is de regressielijn er door heen getekend. In Figuur 39 zijn de grafieken te zien. De verschillende test datasets worden weergegeven met test1, test 2 en test 3.



FIGUUR 39 - VOORSPELLINGEN VERSUS WERKELIJKHEID

Grotendeels lijken de drie grafieken op elkaar. In deze grafieken zijn minder punten weergegeven omdat de test dataset kleiner is dan de train dataset. Daardoor kan het ook dat eventuele uitschieters meer impact hebben. Ook nu geldt dat de waarden niet boven de 1.0 uit zouden moeten komen omdat dit de maximale bezettingsraad is, namelijk 100%. Het is te zien dat alle test gegevens hier aan voldoen. De volgende stap is om te kijken naar de eerder genoemde drie criteria. Er kan worden gekeken naar de samenhang tussen de voorspelde bezettingsgraad van het model en de werkelijke waarden die is waargenomen.

Het eerste criterium voor de validiteit van het model is de nauwkeurigheid van de voorspelling. Deze kan benaderd worden door de samenhang tussen de voorspelling en de werkelijkheid oftewel de waarden van R^2. In Tabel 11 zijn de waarden voor elk model op de test dataset weergegeven. Alle waarden voor R^2 zijn hoger dan de 0.6 uit het eerste criterium, dus hieraan is voldaan. Wat ook te zien is, dat alle waarden lager zijn dan de nauwkeurigheid in uit de train dataset. Dit is logisch te verklaren doordat deze waarnemingen nieuw zijn voor het model, die kunnen afwijken van de gegevens waarop het model gebouwd is. Echter zijn de verschillen niet groter dan 0.02, oftewel 2%.

Test dataset	Test 1	Test 2	Test 3
$R^2 > 0.6$	$R^2 = 0.8097$	$R^2 = 0.8038$	$R^2 = 0.7947$
R = 1	R = 0.900	R = 0.897	R = 0.891
Regressieformule	Y = 0.8481x + 0.0561	Y = 0.7850x + 0.0584	Y = 0.7804x + 0.0651

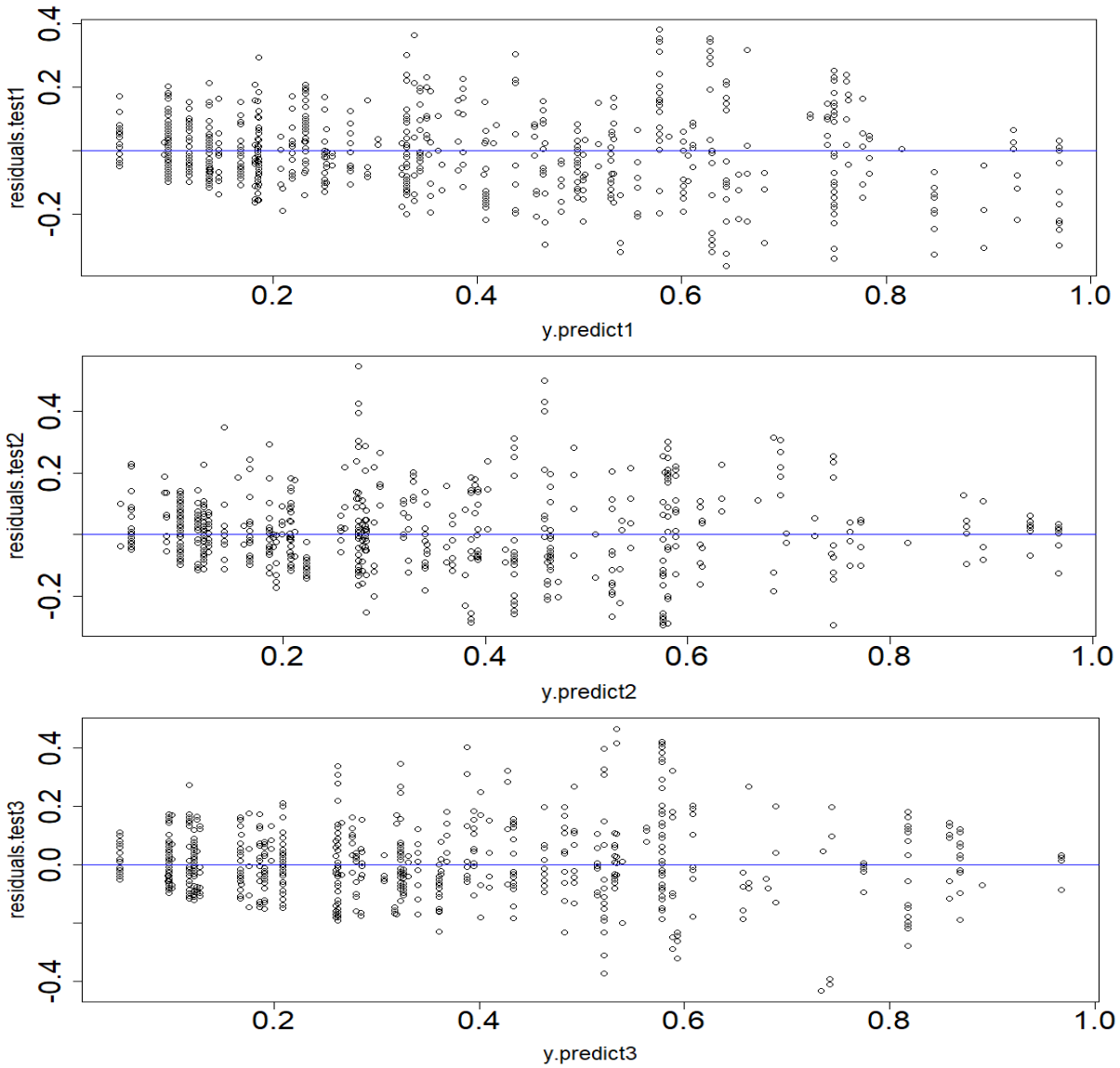
TABEL 11 - REGRESSIE ANALYSE

Het tweede criterium gaat over de correlatie coëfficiënt R, die zo dicht mogelijk bij 1 moet liggen. Ook deze waarden zijn iets gezakt, maar niet veel. Hierdoor kan ook zeker vastgesteld worden dat de afwijking van de correlatie niet significant groot is en dus ook aan het tweede criterium wordt voldaan.

Voor het laatste criterium wordt er naar de blauwe regressielijnen in de grafieken gekeken. Hierbij hoort voor elke lijn een eigen regressieformule. De richtingscoëfficiënt van deze lijn moet zo dicht mogelijk bij 1 liggen en de lijn moet ongeveer door de oorsprong gaan. De richtingscoëfficiënten zijn respectievelijk 0.8481, 0.7850 en 0.7804 en liggen dus relatief gezien dicht bij de optimale waarden 1. Daarnaast gaan de lijnen allemaal dicht door de oorsprong als de getallen 0.0561; 0.0584 en 0.0651, omdat deze alle drie heel dicht bij 0 liggen. Dus ook de regressieformules voldoen goed aan het criterium. De waarden lijken misschien niet al te dicht bij 1 en 0 te liggen, maar deze waarden zijn het optimale scenario, wat realistisch gezien niet mogelijk is. Een voorspelling blijft altijd een benadering. De drie modellen met de test dataset voldoen dus alle drie met enige significantie aan de criteria.

6.3.2 Residuen

De voorspelde waarden en de werkelijke waarnemingen lijken een goede samenhang te hebben. Daarom is de volgende stap om te kijken of ook de residuen een goede verdeling hebben. De residuen geven de afwijkingen tussen de voorspelling en de werkelijkheid. Deze afwijkingen moeten geen voorspelbaar patroon hebben en daarnaast normaal verdeeld zijn met een gemiddelde rond nul. Een normaal verdeling voor de residuen geeft aan dat de meeste afwijkingen rond nul liggen en hoe verder weg van nul hoe minder afwijkingen er worden waargenomen. Dit is dus beter voor het model, omdat er dan minder grote afwijkingen zijn en meer kleine afwijkingen. In Figuur 40 zijn de residuen uitgezet tegenover de voorspelling, waarbij de voorspellingen van drie test dataset respectievelijk zijn aan gegeven met predict1, predict2 en predict3. De residuen van deze voorspellingen tegenover de werkelijke waarden uit de test datasets zijn weergegeven met test1, test2 en test3.



FIGUUR 40 – VOORSPELLING VERSUS RESIDUEN

Uit de grafieken van de residuen, die ook nu weer veel op elkaar lijken, vallen een aantal dingen op. Er zijn meer punten links in het begin van de grafiek te vinden dan aan het einde. Dit kan verklaard worden door het feit dat de gemiddelde bezettingsgraad van alle dagen van alle parkeergarages op 0.31 ligt. Het is dus logisch dat er meer waarnemingen laag uitvallen in de test dataset. Daarnaast is ook nu de lineaire trendlijn aanwezig. Dit betekent dat het model dus hier wel goed rekening mee houdt, omdat er geen bezettingsgraad hoger dan 1 voorspeld wordt. Tot slot is het ook duidelijk te zien dat de puntenwolk zich vooral om de blauwe lijn y = 0 bevindt. Dit geeft aan dat het gemiddelde van de afwijkingen van de puntenwolk rond de nul zal liggen, wat ook gewenst is voor een normale verdeling.

Verder is de y-as met daarop de residuen kleiner dan bij de train dataset, waar de as van -0.4 tot -0.6 liep. Dit is positief, want dat betekent dat de afwijkingen op de test dataset dus minder groot zijn. Dit betekent niet direct dat het model goed is of dat de test dataset goed presteert, want het gemiddelde van de residuen kan nog steeds anders zijn. Om dit verder te bepalen wordt er gekeken naar de verdeling van de residuen.

6.3.3 Normaal verdeling residuen

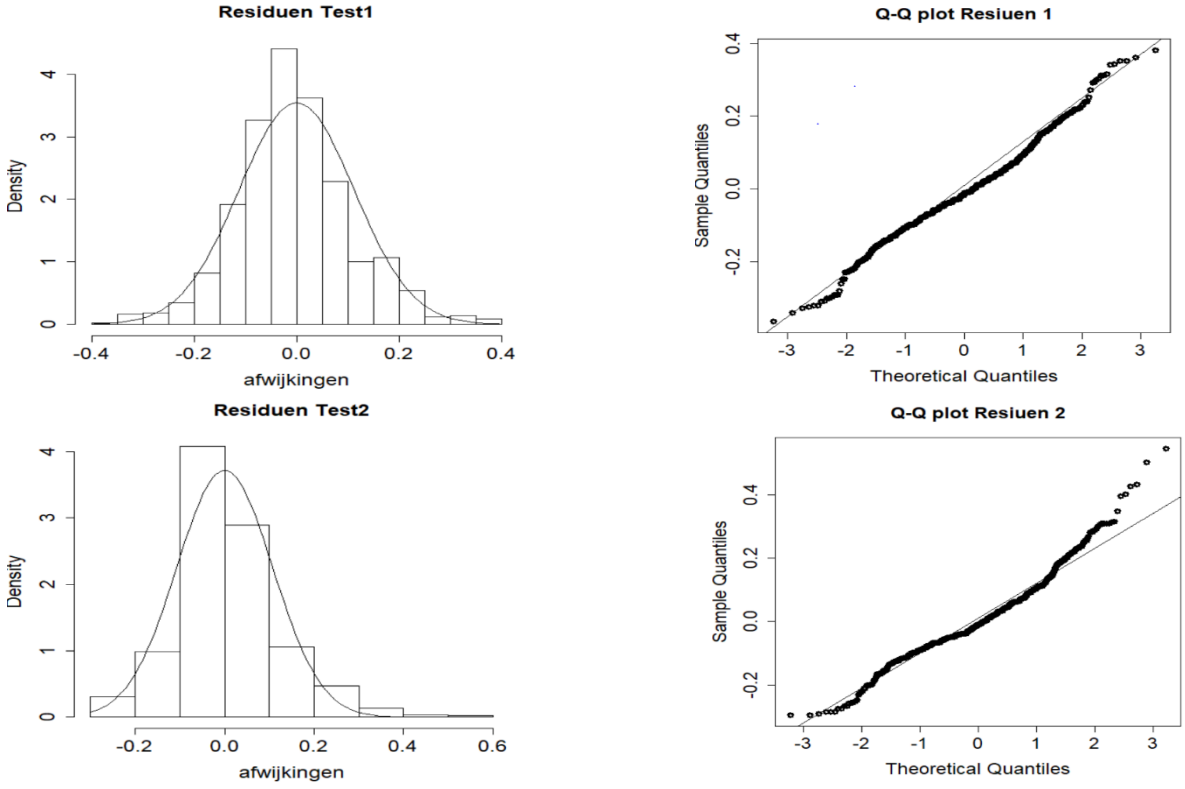
Om te kunnen bepalen of de residuen ook daadwerkelijk normaal verdeeld zijn, worden de statistische waarden van de residuen van elke dataset bekeken. Hierin is het belangrijkste dat het gemiddelde van de residuen nul is, om normaal verdeeld te zijn. In Tabel 12 zijn al deze waarden weergegeven en is te zien dat geen enkel gemiddelde precies nul is. Test 3 heeft een gemiddelde het dichts bij nul, maar ook de andere afwijkingen zijn relatief klein. Het is hieruit niet meteen vast te stellen of de residuen wel of niet normaal verdeeld zijn door deze kleine verschillen.

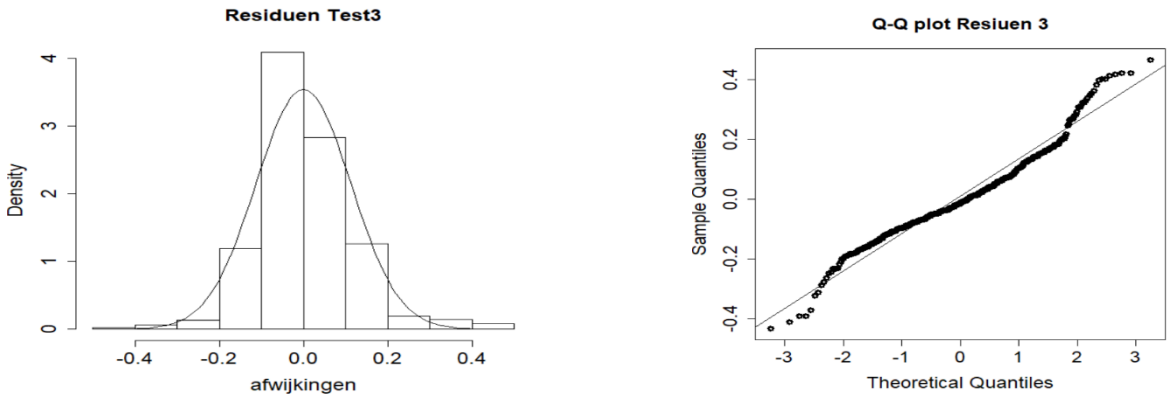
Residuen	Minimum	Q1	Mediaan	Gemiddelde	Q3	Maximum
Test 1	-0.36328	-0.07524	-0.01393	-0.00751	0.05249	0.38165
Test 2	-0.29540	-0.06718	-0.00992	-0.00535	0.06304	0.54574
Test 3	-0.43333	-0.06966	-0.01369	-0.00057	0.05995	0.46594

TABEL 12 - STATISTISCHE SAMENVATTING VAN DE RESIDUEN VAN DE TEST DATASET

In de gegevens zijn wat meer verschillen te zien dan tussen dezelfde gegevens van de train dataset. Waar bij de train dataset alle waarden ongeveer gelijk waren, zijn nu de minimum en maximum waarden wel verschillend van elkaar. Toch kan er niet meteen geconcludeerd worden of dit juist goed of slecht is, omdat het te maken heeft met eventuele uitschieters. Test 1 en test 3 lijken het symmetrisch goed te doen, met wat kleine afwijkingen. Test 2 daarentegen lijkt minder symmetrisch, door een minimum van -0.29540 en een maximum van 0.54574. Daarentegen zijn de waarden van kwartiel 1 en kwartiel 3 van test 2 juist weer het meest symmetrisch.

Het blijft belangrijk dat de verdeling symmetrisch is gecentreerd rond nul en dat de afwijkingen vanaf nul steeds onwaarschijnlijker worden als de afwijking groter wordt. Echter uit alleen deze getallen kan niet de conclusie getrokken worden of de residuen echt symmetrisch verdeeld zijn. Uitschieters kunnen hier een grote rol in spelen, die voor de verdeling wel genegeerd kunnen worden. Om de symmetrie te controleren zijn er histogrammen gemaakt van de residuen, Figuur 41. De verschillende test datasets worden respectievelijk weergegeven met test 1, 2 en 3. Ook de Q-Q plots zijn erbij te zien die weergeven of residuen daadwerkelijk op één lijn liggen.





FIGUUR 41 - HISTOGRAMMEN EN Q-Q PLOTS VAN DE RESIDUEN

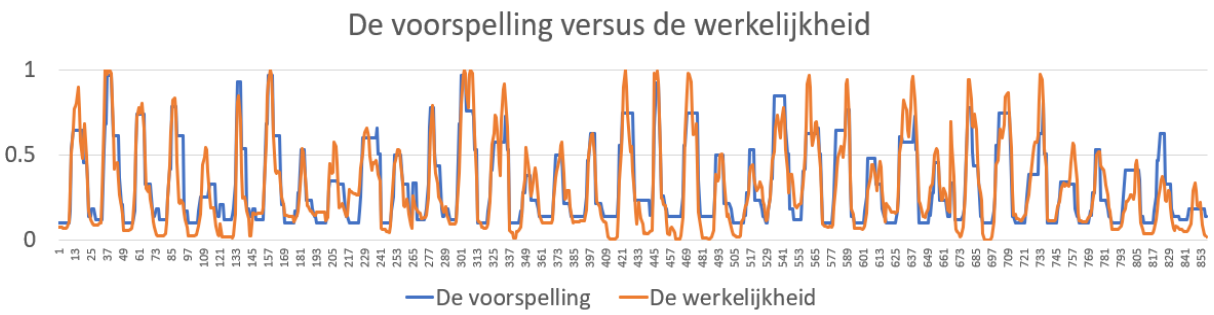
Het histogram van test 1 lijkt op een perfecte normale verdeling. De piek rond de nul waarden met beide staarten even dik en even lang. Ook in de Q-Q plot is te zien dat alle waarden op of net naast de lijn liggen. Daarom is het zeer aannemelijk om te zeggen dat de residuen voor test dataset 1 normaal verdeeld zijn. Test dataset 2 geeft een ander beeld weer. De spreiding op de x-as van de afwijkingen loopt van -0.3 tot en met 0.6. Toch lijkt de verdeling nog redelijk symmetrisch, maar de staart aan de rechter kant is langer dan aan de linker kant, dit wordt ook wel rechtsscheef genoemd. Dit is ook te zien in de Q-Q plot waar de punten rechtsboven flink afwijken van de lijn. De lijn wordt hier ook dunner en eindigt met een aantal punten. Dit geeft dus aan dat er maar weinig waarnemingen zo flink afwijken en dat er dus een aantal extreme verschillen tussen de voorspelling en de waarneming zaten. De rest van de punten liggen grotendeels allemaal op de lijn. Tot slot ziet het histogram van test dataset 3 er redelijk symmetrisch uit. Wat wel opvalt is dat er relatief veel waarnemingen in de staarten van het histogram aanwezig zijn in vergelijking met de normale verdeling. In de Q-Q plot is ook te zien dat de punten een soort uitgerekte s vormen, waarbij de uiteinden flink afwijken van de lijn. Dit gedrag betekent dat de gegevens extremere waarden bevatten dan zou worden verwacht als ze echt afkomstig waren van een normale verdeling.

In de test dataset 2 en 3 zijn ten opzichte van train dataset 2 en 3 een aantal extreme waarnemingen geconstateerd, terwijl test dataset 1 juist normaal verdeeld lijkt. Dit kan komen doordat in de train dataset meerdere uiteenlopende waarnemingen zitten, waardoor de waarnemingen in de test dataset makkelijker te voorspellen zijn. Voor de set 2 en 3 kan het zijn dat de waarnemingen in de test dataset zo afwijken van de waarnemingen die zijn gebruikt om het model te maken, dat de voorspellingen ook afwijken. Op alle relatief kleine afwijkingen na, voorspelt het model ook de test dataset met een nauwkeurigheid van 80%.

6.4 Visualisatie van de voorspellingen

In de vorige paragraaf is duidelijk te zien dat de verdeling van train en test dataset 1 het beste presteert. Daarom wordt het model uit train dataset 1 gebruikt om te kijken naar de voorspellingen die aan de hand van de regressieboom berekend worden. Alle paden die aan de verschillende klassen zijn verbonden in de boom zijn verwerkt in Excel.

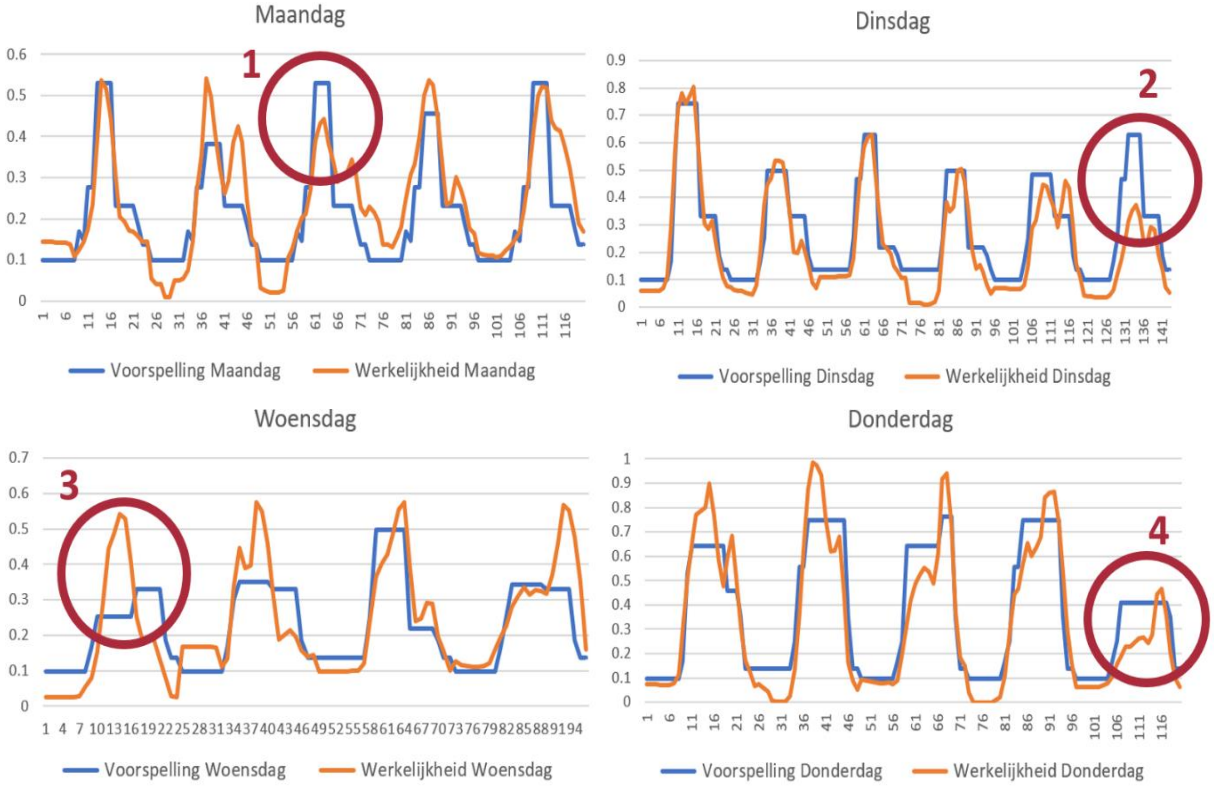
Voor elke regel in de test dataset wordt gekeken aan welke voorwaarden wordt voldaan en welk pad hieraan precies voldoet. Zo krijgt elke regel, die een waarneming representeert een eigen voorspelling van de bezettingsgraad. In test dataset 1 zijn 857 waarnemingen. In Figuur 42 zijn al deze waarnemingen met hun bijbehorende voorspelling weergegeven. Op de y as is de bezettingsgraad weergegeven en op de x-as het nummer van de regel waarin de waarneming zich bevindt. De blauwe lijn is de voorspelling en de oranje lijn is de werkelijkheid. Elke piek representeert één dag van een bepaalde parkeergarage. In totaal zijn er 36 dagen weergegeven.

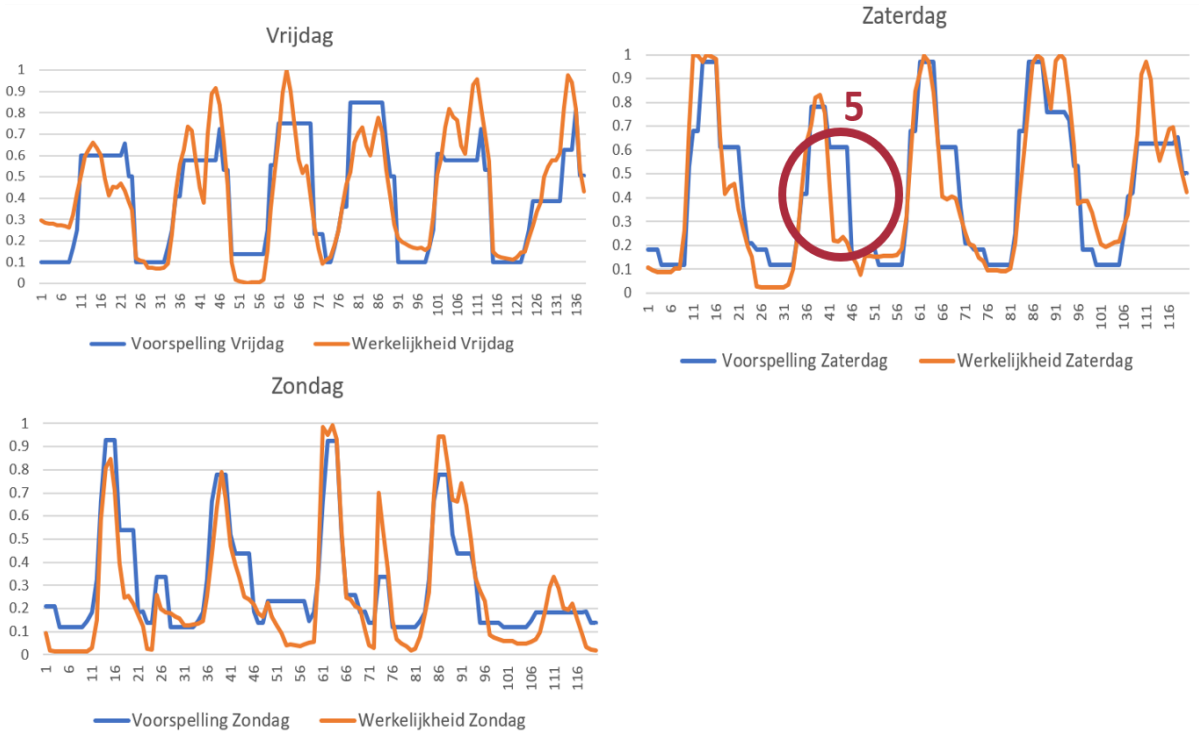


FIGUUR 42 - DE VOORSPELDE BEZETTINGSGRAAD VERSUS DE WERKELIJKHEID

Het is goed te zien dat de blauwe lijn van de voorspelling grotendeels gelijk loopt met de werkelijkheid. Het model voorspelt met een zekerheid van 81%, oftewel 19% van de voorspellingen wijken af. Daarnaast is te zien dat de blauwe lijn wat meer rechte stukken laat zien. Dit komt door de verschillende klassen waarmee gewerkt wordt. In model 1 zijn 109 klassen waarover de bezettingsgraad in klassen van 0.000 tot 1.000 verdeeld zijn. Niet elke mogelijkheid zit hierin, dus vandaar dat een aantal waarnemingen achter elkaar dezelfde voorspelling kunnen krijgen, waardoor er dus rechte lijnen en haakse hoeken in de voorspellingslijn zitten.

Naast de voorspelling over alle waarnemingen is het interessant om te zien hoe het model het doet op alle verschillende dagen in de week afzonderlijk. In Figuur 43 zijn de voorspellingen tegenover de werkelijkheid weergeven per dag. De 36 dagen uit de test dataset zijn verdeeld over de grafieken per dag typen. In één grafiek staan dus verschillende parkeergarages en verschillende data door elkaar, maar de waarnemingen van één dag van de week staan wel bij elkaar. Dus één piek geeft één dag van een parkeergarage. Grotendeels lijken ook hier de blauwe lijn van de voorspellingen en de oranje lijn van de werkelijkheid samen te vallen. Toch zijn de afwijkingen nu wat duidelijker te zien. Een aantal verschillen zijn met een rode cirkel aangegeven. Deze grotere afwijkingen worden bekeken om te zien of er een trend aanwezig is in de afwijkingen.





FIGUUR 43 - DE VOORSPELLINGEN PER TYPE DAG

De gegevens per cirkel zijn weergegeven in Tabel 13 om te kijken of de afwijkingen te verklaren zijn of dat er een bepaald patroon te herkennen is. Voor elke afwijking zijn de gegevens van de variabelen ingevuld. Er zitten wat overeenkomsten in de gegevens. Zo komt de parkeergarage in Groningen en in Apeldoorn twee keer voor. Voor de garage in Apeldoorn geldt echter dat in cirkel 3 de bezettingsgraad eerst flink wordt onderschat en vervolgens wordt overschat, terwijl juist in cirkel 5 de bezettingsgraad alleen wordt overschat. De twee pieken uit cirkel 1 en cirkel 2 lijken op elkaar en in de gegevens is te zien dat ze gemeenschappelijk hebben dat het op die gemeten dag vakantie was. Het model verwacht een hogere bezetting in verband met de vakantie, maar in werkelijkheid was dit niet zo. Het model heeft moeite met het voorspellen van de bezettingsgraad in de vakantie op werkdagen. Dit kan komen doordat er weinig gegevens in de train dataset zitten die duidelijk het verband met de vakantie kunnen weergeven. Daarnaast zijn dit ook dagen die van zichzelf veel variatie zullen leveren, omdat elke vakantie anders beleefd wordt. De variabele regen kan ook enige invloed hebben op de afwijkingen. Echter zitten hier én onderschattingen én overschattingen in, waardoor er dus geen duidelijk patroon te herkennen is. De andere variabelen lijken ook geen duidelijke patroon weer te geven.

	Cirkel 1	Cirkel 2	Cirkel 3	Cirkel 4	Cirkel 5
Naam	Tilburg – Heuvelpoort	Groningen – Damsterdiep	Apeldoorn – Marktplein	Groningen – Damsterdiep	Apeldoorn – Marktplein
Capaciteit	375	392	481	392	481
Tijdstip	12:00 – 17: 00	10:00 – 17:00	10:00 – 22:00	10:00 – 18:00	17:00 – 22:00
Event	Nee	Nee	Nee	Nee	Nee
Temperatuur	24	14	15	23	21
Regen	Geen	Licht/Matig	Licht/Matig	Geen	Licht/Matig
Vakantie	Ja	Ja	Nee	Nee	Nee

TABEL 13 - GEGEVENS VAN DE AFWIJKINGEN

6.5 Conclusie

Aan de hand van een regressie analyse en het bekijken van de residuen is in dit hoofdstuk de prestatie van beide modellen beoordeeld. Er is een onderscheid gemaakt tussen het valideren van het beste model op de train dataset en op de test dataset. Er zijn drie verschillende combinaties van train en test dataset gebruikt om toeval uit te sluiten. Met de analyses die zijn gedaan, kan er antwoord gegeven worden op de volgende onderzoeksvraag:

- Wat is de prestatie van het instrument op de train en test datasets?

De validatie begint met het checken van de regressie tussen de voorspelde waarden en de werkelijke waarnemingen van het eenvoudige model. Al snel blijkt dat het model aan geen enkel criteria voldoet en slecht scoort op de eisen. Zo is de gemiddelde waarde voor de determinatie coëfficiënt R^2 gelijk aan 0.57. In dit onderzoek wordt naar een nauwkeurigheid van 80% gestreefd, wat dit model dus niet haalt. Na het analyseren van de regressie is de volgende stap het bekijken van de residuen, het verschil tussen de werkelijke bezettingsgraad en de voorspelde bezettingsgraad. Hoe kleiner deze afwijkingen zijn, des te beter het model presteert. Daarnaast moeten deze residuen geen patroon weergeven en normaal verdeeld zijn. De gegevens van de verschillende train datasets lijken veel op elkaar. Echter zijn deze residuen duidelijk niet normaal verdeeld en daarom kan er worden geconcludeerd dat dit eenvoudige model de werkelijke bezettingsgraad niet goed kan voorspellen.

Vervolgens worden dezelfde analyses uitgevoerd voor de regressieboom. De samenhang tussen de werkelijke waarden en de voorspelling wordt weergegeven met de correlatie coëfficiënt R, die gemiddeld voor elke train dataset gelijk is aan 0.91. Een waarde van $R > 0.7$ geeft aan dat er een sterk verband is tussen de variabelen. In dit geval betekent dat dus dat de voorspellingen dicht bij de werkelijkheid liggen. Daarnaast moet voor elke dataset ook de determinatie coëfficiënt R^2 groter dan 0.6 zijn. De waarden van R^2 geeft de nauwkeurigheid van het model. Voor elke train dataset ligt deze waarden boven de 0.8, waardoor de nauwkeurigheid van het model minimaal 80% is op de train dataset. Ook de regressieformules zijn bekeken per dataset en voldoen aan het criterium. Verder geven de residuen allemaal een random patroon en lijken ze normaal verdeeld te zijn. Er zijn enkele extremen afwijkingen, die niet in een normaal verdeling verwacht zouden worden. Toch kan er met enige zekerheid gezegd worden dat de residuen normaal verdeeld zijn en het model goed presteert op de train dataset.

Door de duidelijke verschillen tussen het eenvoudige model en de regressieboom, wordt in het onderzoek verder gegaan met de regressieboom. Alle analyses laten duidelijk zien dat deze methode het beste presteert, daarom wordt deze methode verder beoordeeld aan de hand van de test datasets. Alle stappen zijn nog een keer herhaalt en ook nu komt de regressieboom goed naar voren. De nauwkeurigheid van het model is gezakt, maar relatief weinig. De waarden van R^2 blijft gemiddeld 0.80 en daardoor de blijft de gemiddelde R waarde 0.90. Ook hier zijn de regressievergelijkingen bekeken en voldoen ook deze aan alle criteria. Het analyseren van de residuen geeft iets meer verschillen weer tussen de verschillende test datasets. De test dataset 1 geeft het best een normale verdeling weer als er wordt gekeken naar de residuen. In de andere datasets zijn meer extreme waarden aanwezig, waarvan het model de verbanden niet helemaal goed kan leggen. Daardoor worden er grotere afwijkingen gevonden. Toch is dit niet in de meerderheid van de voorspellingen, waardoor nog steeds gezegd kan worden dat de regressieboom accurate voorspelling kan genereren.

De voorspellingen van de test dataset zijn samen met de werkelijke waarden in een grafiek gezet. Aan de hand hiervan zijn de afwijkingen tussen deze twee lijnen duidelijk zichtbaar. De grotere afwijkingen zijn uitgelicht en de informatie van deze metingen zijn samengevat. Toch lijken de afwijkingen geen

patroon te vormen. De afwijkingen kunnen met allerlei factoren te maken te hebben, denk aan een wegafsluiting, waardoor de parkeergarage slecht te bereiken is. Het is hierdoor wel aannemelijk dat alle verbanden die aanwezig zijn, door het model eruit gehaald zijn.

Uiteindelijk kan geconcludeerd worden door de analyses dat de prestatie van de regressieboom goed is. Er kan nog met de complexiteit gespeeld worden om de betrouwbaarheid van het model hoger te maken. Hierdoor zal de nauwkeurigheid van de voorspelling omhoog gaan. Daarnaast wordt dan ook de kans op overfitten groter. Dit houdt in dat er relaties worden gelegd die er niet zijn. De prestatie van het model wordt daardoor ook lager. Voor nu was het streven een betrouwbaarheid van 80% en de regressieboom geeft een gemiddelde betrouwbaarheid van 80%. Dus met dit instrument kan de bezettingsgraad van parkeergarages in Nederland met 80% zekerheid worden voorspeld.

Hoofdstuk 7 – Conclusie en aanbevelingen

In dit laatste hoofdstuk van het onderzoek is de conclusie van het hele onderzoek te lezen. Op basis daarvan worden er aanbevelingen gedaan voor DAT.Mobility. Ook is er gekeken naar eventueel toekomstig onderzoek en wat er nog verder ontwikkeld zou kunnen worden. De conclusie in dit hoofdstuk wordt gevormd door het beantwoorden van de volgende vragen:

- Wat is het antwoord op de hoofdvraag?
    - In welke andere situaties en applicaties kan dit nieuwe instrument van belang zijn?
    - Welke verdere uitbreidingen zijn er mogelijk met het model voor Goudappel Groep?

Aan de hand van deze vragen wordt eerst de conclusie van het onderzoek besproken (§7.1). Vervolgens worden de aanbevelingen voor Goudappel Groep toegelicht in paragraaf 2. Verder worden mogelijke toekomstige onderzoeken besproken, die aansluiten bij het ontworpen model (§7.3).

7.1 Conclusie

Het onderzoek is begonnen met het analyseren van de informatievoorzieningen in de buurt van parkeergarages in verschillende steden. De eerste bevinding was dat er veel verschillende technieken gebruikt worden om automobilisten te voorzien van informatie over parkeerplekken. Echter wordt dit pas gedaan op het moment van arriveren. Op het moment van vertrek is er nog geen informatie beschikbaar over waar de automobilist zijn auto zeker weten kan parkeren.

Binnen Goudappel Groep zijn er historische gegevens beschikbaar over de bezetting van parkeergarages in Nederland. Echter wordt er niks met deze gegevens gedaan en kunnen zij ook (nog) geen analyses doen op dit gebied. De focus van het onderzoek lag daarom op het onderzoeken van verschillende voorspelmethoden en de ontwikkeling van een voorspellingsmodel. Dit model kan door Goudappel Groep worden meegenomen in hun eigen applicatie OmniTRANS Next, maar er kunnen ook analyses voor gemeenten mee gedaan worden.

Uit het literatuuronderzoek is gebleken dat de methode regressieboom het beste kan worden gebruikt voor het voorspellen van de bezettingsgraad van parkeergarages. Ook is er gekeken naar welke variabelen in eerdere onderzoeken een duidelijk verband hadden met de bezettingsgraad. De onafhankelijke variabelen die uiteindelijk zijn meegenomen in het model zijn: tijd, type dag, capaciteit, event, temperatuur, regen en vakantie. Uit analyses van deze gegevens kwamen duidelijke verschillen naar voren. De bezetting is op werkdagen gemiddeld lager dan in het weekend. Ook de variabele tijd heeft duidelijk een overheersende invloed op de bezetting. Een ander groot verschil tussen de parkeergarages is de locatie. De bezetting van parkeergarages in het centrum van een stad laten duidelijk een ander patroon zien dan parkeergarages in de buurt van bedrijven en scholen. In een centrum waar winkels op zondag open zijn zal het heel druk zijn, terwijl rond een bedrijven terrein op zondag weinig automobilisten te vinden zullen zijn. Door deze scheiding in de parkeergarages is in het onderzoek de focus gelegd op parkeergarages in het centrum van steden om deze met elkaar te vergelijken. In totaal zijn er 12 parkeergarages meegenomen voor het analyseren en modelleren van de bezettingsgraad.

Aan de hand van alle historische gegevens en verschillende onafhankelijke variabelen van deze parkeergarages is het model uitgewerkt. De totale dataset is verdeeld in drie groepen met elk een train en een test dataset. Dit is gedaan om toeval uit te sluiten en om de uitkomsten met elkaar te kunnen vergelijken. In het onderzoek wordt gestreefd naar een nauwkeurigheid van 80%. Om de regressieboom te kunnen vergelijken, is er ook een eenvoudig model toegevoegd met maar twee onafhankelijke variabelen. Met beide modellen zijn voorspellingen gegeneerd met als input gegevens de drie verschillende train datasets.

De voorspellingen zijn geanalyseerd aan de hand van de regressielijn en de residuen. De regressielijn geeft de samenhang weer tussen de voorspelling en de werkelijkheid, die zo dicht mogelijk bij elkaar moeten liggen. De residuen zijn de afwijkingen tussen de werkelijke waarneming en de voorspelling. Deze afwijkingen moeten normaal verdeeld zijn, alleen dan kan er worden vastgesteld dat de modellen de werkelijkheid zo goed mogelijk benaderen. Uit de analyses bleek al snel dat het eenvoudige model niet voldoet aan verschillende criteria en ook niet aan de nauwkeurigheid van 80%. De regressieboom daar tegenover voldoet wel aan deze criteria en kan ook met ruim 80% zekerheid de bezettingsgraden van parkeergarages voorspellen.

Het is logisch dat een model de input gegevens met een relatief hoge zekerheid kan voorspellen. Om te controleren of de regressieboom ook goed presteert op nieuwe gegevens werd de test dataset gebruikt. Dezelfde analyses zijn uitgevoerd en het model scoort lager, maar dit verschil is niet significant. Ook de residuen lijken nog steeds sterk normaal verdeel te zijn. Er zijn enkele kleine afwijkingen terug te vinden, maar deze worden vooral verklaard door de extreme uitschieters. Deze uitschieters kunnen komen door de kleine dataset die is gebruikt in het onderzoek. In een optimaal geval zouden de dagen van verschillende jaren met elkaar vergeleken worden, om zo extreme en uitzonderlijke gevallen er tussen uit te filteren.

Met deze kleine dataset in gedachte kan het worden vastgesteld dat het model zeer goed presteert. Uit deze analyses komt naar voren dat de prestaties van de modellen van de regressieboom weinig tot niks van elkaar verschillen. De residuen van de test datasets lijken normaal verdeeld te zijn. Daarnaast is de nauwkeurigheid van de modellen gemiddeld 80%, wat inhoudt dat er in 80% van de gevallen een goede voorspelling wordt gedaan.

In het begin van het onderzoek was ook een hoofdvraag opgesteld, die als luidt volgt:

- Welk instrument is geschikt voor het voorspellen van de bezettingsgraad van parkeergarage op de korte termijn?

Het antwoord op deze vraag is de regressieboom. Er was behoefte aan een model dat de bezettingsgraad van parkeergarages kon voorspellen. Dat model ligt nu op de planken bij Goudappel Groep. Hiermee kunnen nog veel nieuwe ontwikkelingen worden doorgevoerd. Verschillende variabelen kunnen worden toegevoegd of juist worden verwijderd. Verder kan er gespeeld worden met de waarde van de complexiteit. Hierdoor wordt de grootte van de boom bepaald en daarmee ook direct de nauwkeurigheid van het model.

Ook de doelstellingen van het onderzoek zijn gehaald. Er is uitgebreid onderzoek gedaan naar de verschillenden methoden die er zijn om voorspellingen met te generen. Deze zijn in eerdere onderzoeken al met elkaar vergeleken en ook deze resultaten zijn bekend. Uit deze analyse komt de regressieboom als het betrouwbaarste model naar voren. Daarnaast was het doel om een instrument te ontwikkelen en ook dit doel is behaald. Met het model kunnen analyses gedaan worden en het kan worden toegevoegd in de applicatie OmniTRANS Next.

7.2 Aanbevelingen

Het is nu de taak aan Goudappel Groep om het model te gebruiken, verder te ontwikkelen en uit te breiden. Een belangrijke taak is om via onderzoek bij adviseurs te achterhalen welk percentage voor de betrouwbaarheid van het model voldoende is. Dit zal per toepassing grote verschillen weergeven. Het kan zijn dat analyses van verschillende parkeergarages globaal bekeken moeten worden om informatie aan gemeenten te verstrekken. Het kan ook zijn dat het model direct van toepassing is bij de automobilisten. Dan moet het model een zeer hoge betrouwbaarheid hebben, omdat er anders

nog steeds zoekverkeer, wachtrijen en ontevreden automobilisten ontstaan. Dit onderzoek lag niet binnen mijn scope, maar de informatie is binnen Goudappel Groep zeker aanwezig.

Daarnaast kan het model informatie aanleveren voor de applicatie OmniTRANS Next. Vooral als er ook nog in- en uitrijtijden aan worden gekoppeld, kan dit model veel zeggen over de intensiteit op het wegennetwerk rond een parkeergarage. Het is goed voor Goudappel Groep om een toepasbaar instrument klaar te hebben liggen. Met een aantal verdere ontwikkelingen in de applicatie, kunnen ze de voorspellingen overal en altijd generen. Ook voor verdere projecten is het handig als ze het model begrijpen en kunnen aanpassen naar hun eigen wensen.

7.3 Toekomstig onderzoek

Er zijn veel uitbreidingen mogelijk die de toepassing van het model beter en breder te maken. Als eerste zouden er meer details in het model kunnen worden toegevoegd, zodat er (nog) betere keuzes in de regressieboom gemaakt kunnen worden. Verschillende verkeerssituatie in de steden kunnen bekeken worden die invloed hebben op de parkeergarages. De ene parkeergarage kan beter bereikbaar zijn dan een andere. Ook werkzaamheden zouden invloed kunnen hebben. Zo zijn er nog ontelbare invloeden van buitenaf die kunnen verklaren waarom een piek/dal in de bezetting van de parkeergarage waar te nemen is.

Daarnaast ligt er ook een grote mogelijkheid om parkeergarages in één stad met elkaar te vergelijken. Hier komen andere variabelen om de hoek kijken, die veel invloed zullen hebben op de bezettingsgraad. Als er meerdere parkeergarages in één stad zijn, zullen de loop afstanden bijvoorbeeld verschillen. Het is niet logisch om deze variabele mee te nemen als parkeergarages in verschillende steden worden vergeleken, omdat de automobilist dan de keus niet zelf kan maken. Deze variabele moet dus wel meegenomen worden als de parkeergarages allemaal in één stad liggen. De automobilist kan dan zelf de keuze maken in welke parkeergarage hij/zij de auto parkeert. Ook de prijs en de openingstijden van een parkeergarage kunnen dan interessant zijn. In deze analyses kan er ook worden gekeken naar hoe snel de verschillende parkeergarages in één stad vollopen. Is er echt een favoriete parkeergarage en waarom is dat dan zo? Dit zijn vragen die dan interessant zijn om te onderzoeken.

Het is ook mogelijk om real-time data te koppelen aan het model. Dit houdt in dat als de voorspelling over een uur is, dat de huidige bezettingsgraad bekend is en wordt meegenomen in de voorspelling. Het is logisch voor te stellen dat er dan minder grote afwijkingen zullen zijn in de voorspellingen. Het is geen lastige opgave om de real-time gegevens te koppelen aan de dataset. Echter is het wel lastig om deze real-time gegevens in het model mee te nemen. Het vergt nieuw onderzoek om te kijken hoe dit in het huidige onderzoek toepast kan worden of dat er wellicht een andere methode beter is. Daarnaast is het ook onzeker wat de invloed is van deze real-time gegevens.

Het belangrijkste vervolg onderzoek is al kort genoemd. Vanaf februari gaat een andere student verder met dit onderzoek, waarbij zijn focus ligt op het voorspellen van de intensiteit op het wegennetwerk rond parkeergarages. Dit is natuurlijk een mooie aansluiting op dit onderzoek. Als deze student kan voorspellen hoelang auto's gemiddeld in een parkeergarage verblijven en dit kan combineren met de voorspelling van hoeveel auto's er aanwezig zijn, dan kunnen hieruit mooie analyses volgen die de intensiteit op het wegennetwerk rond de parkeergarages kunnen voorspellen. Dit is ook waardevol voor de applicatie OmniTRANS Next, die aan de hand van verschillende algoritmes, toedelingsmodellen en historische gegevens analyses maakt over deze intensiteiten. Het model kan hieraan worden toegevoegd, waardoor aan de hand van deze voorspellingen gevisualiseerd kan worden wat de invloed is op het wegennetwerk van de in- en uitstroom van parkeergarages.

Bibliografie

Arumugam, A. (2017). A predictive modeling approach for improving paddy crop productivity using data mining techniques . *Turkisch Journal of Electrical Engineering & Computer Sciences*.

Brederode, L. (n.d.). *OmniTrans*. Retrieved from archief.data.nl: http://archief.dat.nl/nl/producten/omnitrans/

CBS. (2018). *StatLine: Motorvoertuigenpark*. Retrieved from statline.cbs.nl: https://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=82044NED

Chen, B., Pinelli, F., Sinn, M., Botea, A., & Calabrese, F. (2013). Uncertainty in urban mobility: Predicting waiting times for shared bicycles and parking lots. *16th Interational IEEE Annual Conference on Intelligent Transportaation Systems*, 53-58.

Chen, X. (2014). Parking Occupancy Prediction and Pattern Analysis. *Stanford*.

CROW. (2007). *Handboek Parkeerverwijssystemen. Mobiliteit en Transport.* CROW.

Debnath, P. (2018, januari 10). *Choosing The Right Forecasting Technique.* Retrieved from Brillio: https://info.brillio.com/insights/blog-posts/choosing-the-right-forecasting-technique

Fabusuyi, T., Hampshire, R. C., Hill, V. A., & Sasanuma, K. (2014). Decision Analytics for Parking Availability in Downtown Pittsburgh. *Interfaces 44(3)*, 286-299.

Golbraikh, A., & Tropsha, A. (2002). Beware of q2! *Journal Mol. Graph. Model*, Volume 20;269-276.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Busniess Research 68*, 1678-1685.

Grooten, J., Gelder, M. v., & Spruijt, W. (2012). Innovatie in Amsterdamse bereikbaarheidsanalyses: meer inzicht tegen minder kosten. Amsterdam.

Heerkens, H., & Winden, A. v. (2012). *Geen probleem.* Buren: Business School Nederland.

ikvergelijkhet.nl. (2018). *Mobiel parkeren*. Retrieved from ikvergelijkhet.nl.

Kammeijer-Luycks, Y. (2017, februari 14). Aan de leden en duoburgerleden van de gemeenteraad van Bergen op Zoom. *Aanpak parkeerroute nformatie systeem (PRIS)*. Bergen op Zoom.

Kant, P. (2018, september 18). OmniTRANS Next . Deventer.

Karlaftis, M., & Vlahogianni, E. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Elsevier*.

KNMI. (2018). *uurgegevens van het weer in Nederland*. Retrieved from KNMI.nl: https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens

Krol, A. (2017, november 28). *Niet de zelfrijdende auto, maar delen zorgt voor een fijnere stad*. Retrieved from Milieudefensie : https://milieudefensie.nl/actueel/niet-de-zelfrijdende-auto-maar-delen-zorgt-voor-een-fijnere-stad

Kuijten, B. (2016, maart 7). *Review: TimesUpp 2.0 is de vernieuwde reisassistent die je waarschuwt voor vertrek*. Retrieved from iCulture: https://www.iculture.nl/apps/review-timesupp-2-vernieuwde-reisassistent/

Kumar, K., Parida, M., & Katiyar, V. (2013). Short term traffic flow prediction for a non urban highway using Artificial Neural Network. *Elsevier*, 755–764.

Kunjithapatham, A., Tripathy, A., Eng, C.-t., Chiriyankandath, C., & Rao, T. (z.d.). Predicting Parking Lot Availability in San Francisco: Modeling the Occupancy Rate of Individual Spaces and Parking Lots. *University of California - Santa Cruz*.

Marsden, G. (2006). *The evidence base for parking policies - a review.* Transport Policy 13 (6) .

Nieuwborg, R. (2013). Masterproef: parkeerplaatskeuze in parkeerfaciliteiten. *Universiteit Hasselt*.

Pflüger, C., Köhn, T., Schreieck, M., Wiesche, M., & Kremar, H. (2016). Predicting the Availability of Parking Spaces with Publicy Available Data. *Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn*, 361–374.

prof. ir. Hakkesteegt, P. (ca 1990). *Bezettingsgraad, toelaatbare bezettingsgraad.* TU Delft: collegedictaat hb 15, par. 13.4.3.

Reid, R. D., & Sanders, N. R. (2011). *Operations Management: An Integrated Approach* . Hoboken : John Wiley & Sons.

Reinstadler, M., Braunhofer, M., Elahi, M., & Ricci, F. (2013). Predicting parking lots occupancy in Bolzano. *Institute of computer science, Free University of Bolzano, Italy* .

Richardson, A. M., & Lidbury, B. A. (2013). Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. *US National Library of Medicine*.

Richter, F., Di Martino, S., & Mattfeld, C. D. (2014). Temporal and spatial clustering for a parking prediction service. *IEEE 26th International Conference on Tools with Artificial Intelligence*.

Rijksoverheid. (2018). *Overzicht schoolvakanties 2018-2019*. Retrieved from Rijksoverheid.nl.

Shinde, S., Bhagwat, S., Pharate, S., & Paymode, V. (2016). Prediction of Parking Availability in Car Parks Using Sensors and IOT: SPS. *International Journal of Engineering Science and Computing*, Volume 6, issue No. 4.

Soler, S. (2015). Creation of a web application for smart park system and parking prediction study for system integration. *Institute of new imaging technologies*.

Svoray, T. (2011). Predicting gully initiation: comparing data mining techniques, analytical hierarchy processes and the topographic threshold. *Earth Surface Processes and Landforms*.

Tomovic, S., & Stanisic, P. &. (2018). Data Mining Approach in Climate Classification and Climate Network Construction - Case Study Montenegro. *University of Montenegro*.

van de Zande, T. T. (2013). *Parkeerverwijssystemen.* Radboud Universiteit Nijmegen.

Vlahogianni, E. I., Kepaptsoglou, K., Tsetsos, V., & Karlaftis, M. G. (2014). Exploiting new sensor tecnologies for real time parking prediction in urban areas. *TRB Annual Meeting*.

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management . *Journal of Business Logistics* , 34(2): 77-84.

Wegenwiki. (2012, februari 8). *Parkeerroute-informatiesysteem*. Retrieved from Wegenwiki: https://www.wegenwiki.nl/Parkeerroute-informatiesysteem

Weijermars, W., & Berkum, E. v. (2004). Stedelijke verkeerspatronen . *Vakgroep Verkeer, Vervoer en Ruimte Universiteit Twente*.

Wentink, I. (2014, juni). Life at nedap. *Bits&Chips*. Retrieved from Case: parkeersensoren op straat verminderen zoekverkeer.

Zarafani, R., Abbasi, M. A., & Liu, H. (2014). Social Media Mining. *Cambridge University Press, New York*.

Zheng, Y., Rajasegarar, S., & Leckie, C. (2015). Parking Availability Prediction for Sensor-Enabled Car Parks in Smart Cities. *The University of Melbourne, Australia*.

Appendices

Appendix A

Globale stappen	Hoofdstuk	(Deel)vragen
Probleemanalyse	1. Introductie	
	2. De bezettingsgraad van parkeergarages	**Hoe worden weggebruikers van informatie voorzien voor het parkeren in parkeergarages?** Welke verwijssystemen en apps zijn er in Nederland in gebruik? Welke technieken worden gebruikt voor het bepalen van de bezettingsgraad? Wat zijn de prestaties van de huidige systemen?
Formulering van alternatieve oplossingen	3. Theoretisch kader	**Wat zijn potentiele methoden om voorspellingen te kunnen doen op basis van historische data?** Welke variabelen komen in de literatuur voor die relevant zijn en bijdragen aan de voorspelling? Welke methoden zijn er in de literatuur bekend om patronen en trends in data te kunnen analyseren? Welke methoden zijn er in de literatuur bekend om een vergelijkbaar voorspelprobleem op te lossen?
Ontwikkeling	4. Data-analyse	**Welke verklaringen zijn er voor de patronen in de dataset?** Welke verschillen zijn er tussen de parkeergarages? Wat zijn de beperkingen van de dataset? Hoe passen alle variabelen bij elkaar in één dataset? Welke patronen zijn er in de dataset aanwezig?
	5. De implementatie	**Hoe kan het proces worden weergegeven in een model?** Hoe wordt de data opgesplitst in een train en test dataset? Wat zijn de in- en output variabelen van het model? Hoe komt het model en de visualisatie hiervan tot stand?
Resultaten	6. Resultaten	**Wat is de prestatie van het instrument op de train en test datasets?**
Evaluatie	7. Conclusie en aanbevelingen	**Wat is het antwoord op de hoofdvraag?** In welke andere situaties en applicaties kan dit nieuwe instrument van belang zijn? Welke verdere uitbreidingen zijn er mogelijk met het model voor Goudappel Groep?

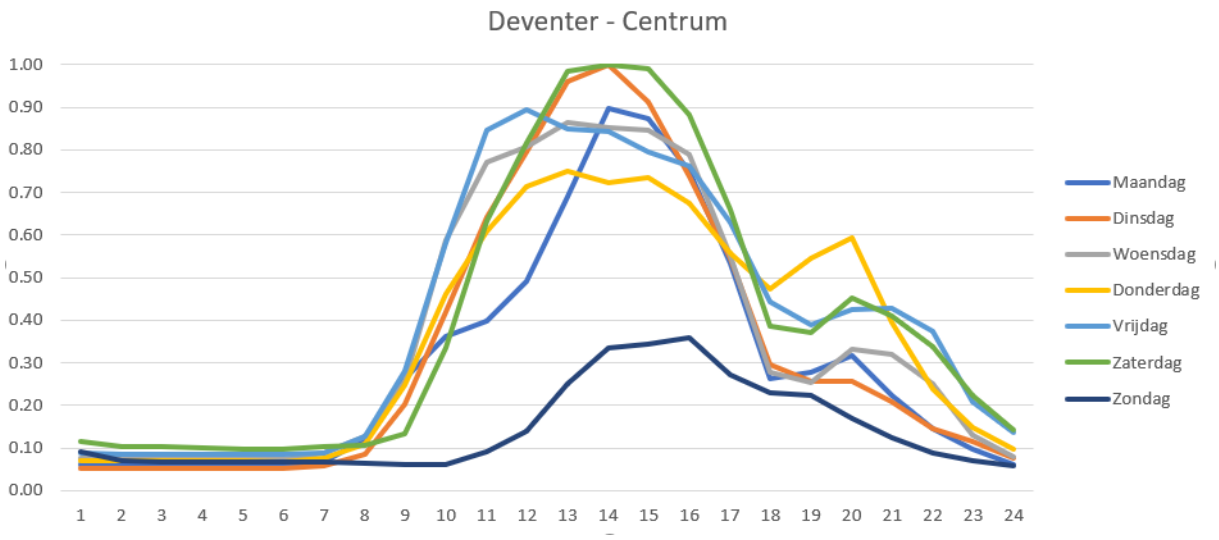
Appendix B

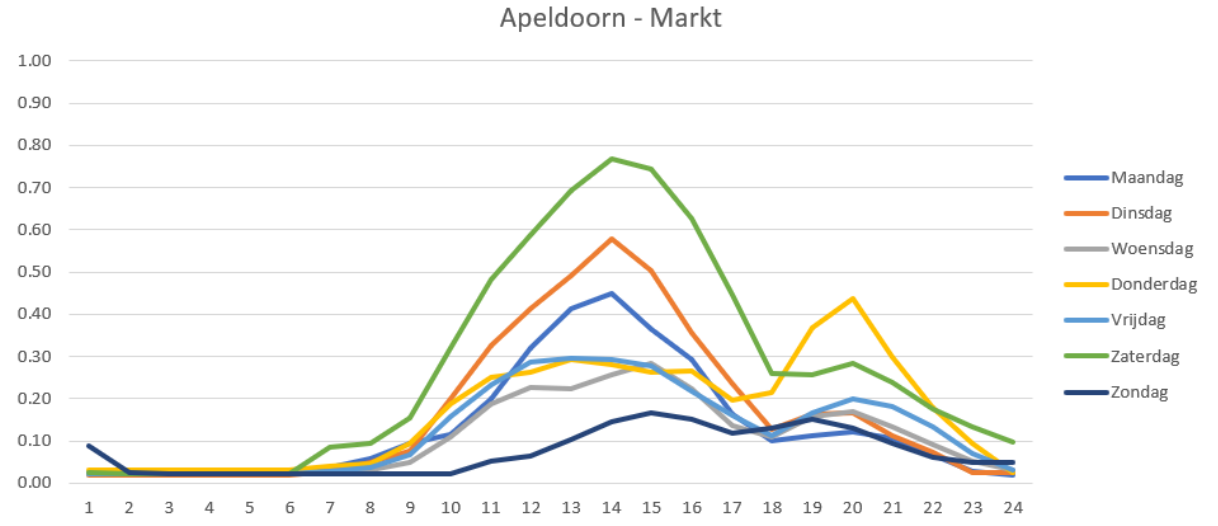
Naam	Capaciteit	Stad	Locatie	Correcte gegevens
Boterdiep	400	Groningen	Centrum	
Boulevard P5-P6	Geen data	Harderwijk	Dolfinarium	
Circus	175	Groningen	Buiten	Nee
Citadel	550	Almere	Centrum	
Damsterdiep	392	Groningen	Centrum	
De Nieuwe Kolk	500	Assen	Centrum	
Drent Museum	238	Assen	Centrum	
Flevoweg P3	Geen data	Utrecht	Centrum	
Garage Bijenkorf	417	Eindhoven	Centrum	Nee
Garage Noorderbrug	Geen data	Maastricht	Centrum	
Garage Oranjekwartier	Geen data	Amsterdam	Buiten	
Grifthoek	320	Utrecht	Buiten	
Havendijk P1	Geen data	Zaltbommel	Haven	
Hoeksterend	390	Leeuwarden	Centrum	
Hoog Catharijne P1	355	Utrecht	Station	
Hoog Catharijne P2	316	Utrecht	Station	
Hoog Catharijne P3	441	Utrecht	Station	
Hoog Catharijne P4	385	Utrecht	Station	
Hoog Catharijne P5	1200	Utrecht	Station	
Hoog CatharijneP6	279	Utrecht	Station	
Houtwal	432	Harderwijk	Centrum	
Jaarbeurs P1	560	Utrecht	Station	
Jaarbeurs P2	900	Utrecht	Station	
Jaarbeurs P3	995	Utrecht	Station	
Jaarbeurs P4	1900	Utrecht	Station	
Klanderij	625	Leeuwarden	Centrum	
Klooster	68	Harderwijk	Centrum	
Kop van Lombok	200	Utrecht	Buiten	
Kruisstraat	100	Utrecht	Buiten	
La Vie	344	Utrecht	Centrum	
Lelykade P2	Geen data	Harderwijk	Buiten	
Mercurius	288	Assen	Centrum	
Neptunus	190	Assen	Centrum	
Oldehove	525	Leeuwarden	Centrum	
Oosterstraat	250	Leeuwarden	Centrum	
Oud Amelisweerd P3	140	Bunnik	Buiten	
P+R station Breukelen NO P2	Geen data	Utrecht	Station	
P+R station Breukelen W P3	Geen data	Utrecht	Station	
P+R station Breukelen ZO P1	Geen data	Utrecht	Station	
P+R Aan de drie Heren	Geen data	Beek (L)	Station/snelweg	
P+R Haren 1	320	Haren (Gr)	Snelweg	
P+R Haren 2	535	Haren (Gr)	Snelweg	
P+R Hoogkerk	567	Groningen	Station/snelweg	

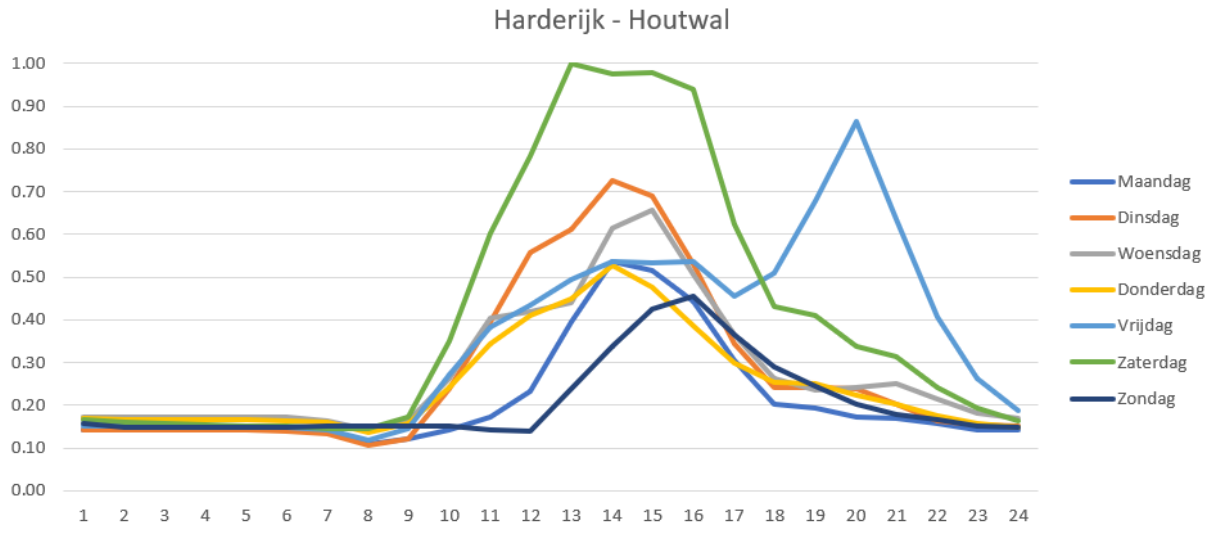
P+R Jeustraat	Geen data	Voerendaal	Station	
P+R Kardinge	880	Groningen	Station/snelweg	
P+R Mareheiweg	Geen data	Klimmen	Station	
P+R Meerstad	438	Groningen	Station/snelweg	
P+R Reitdiep	300	Groningen	Station/snelweg	
P+R Slochterenlaan	155	Bussum	Station	
P+R Spoorstraat	Geen data	Echt-Susteren	Station	
P+R Spoorstraat	Geen data	Meerssen	Station	
P+R Station Maastricht-Noord	Geen data	Maastricht	Station	
P+R Stationsstraat	Geen data	Echt-Susteren	Station	
P+R Stationsstraat	Geen data	Meerssen	Station	
P+R Stationsweg	Geen data	Echt-Susteren	Station	
P01 Stationsplein	173	Deventer	Station	
P1 ChassAparking	600	Breda	Centrum	
P1 Emmapassage	295	Tilburg	Centrum	
P1 Keizer Karel	683	Nijmegen	Station	
P1-P Erasmusbrug	200	Rotterdam	Buiten	
P02 Brink	330	Deventer	Buiten	
P2 Behouden Haven	170	Zaandam	Centrum	
P2 Holland Casino	600	Breda	Centrum	
P2 Koningsplein	169	Apeldoorn	Centrum	
P2 Oude Stad	320	Nijmegen	Centrum	
P2-P Kiphof	112/212	Rotterdam	Centrum	Nee
P2a Dirk van den Broek	Geen data	Zaandam	Centrum	
P03 De Stadspoort	30	Goes	Centrum	
P3 ChassAveld	675	Breda	Centrum	
P3 Heuvelpoort	375	Tilburg	Centrum	
P3 Kelfkensbos	600	Nijmegen	Buiten/snelweg	
P04 Kelder Stadspoort	150	Ede	Centrum	
P4 Het Turfschip	600	Breda	Centrum	
P4 Knegtel	188	Tilburg	Centrum	
P4 Molenpoort	295	Nijmegen	Centrum	
P4-P Schouwburgplein 1	692	Rotterdam	Centrum	
P4 Zaantheather	90	Zaandam	Centrum	
P05 Centrumgarage	481	Deventer	Centrum	
P5 Beyerd-Vlaszak	365	Breda	Centrum	
P5 De Burcht	236	Zaandam	Centrum	
P5 Marienburg	320	Nijmegen	Centrum	
P5-P Oude Haven	97/147	Rotterdam	Centrum	Nee
P5 Schouwburg	285	Tilburg	Centrum	
P06 Sluisstraat	70	Deventer	Centrum	
P6-013 Tivoli	310	Tilburg	Centrum/station	
P6 Eiermarkt	365	Nijmegen	Centrum	
P7 De Prins	470	Breda	Centrum	
P7 Pieter Vredeplein	0	Tilburg	Centrum/station	Nee
P8 De Barones	510	Breda	Centrum	
P8 Stappegoor	783	Tilburg	Bedrijventerrein	

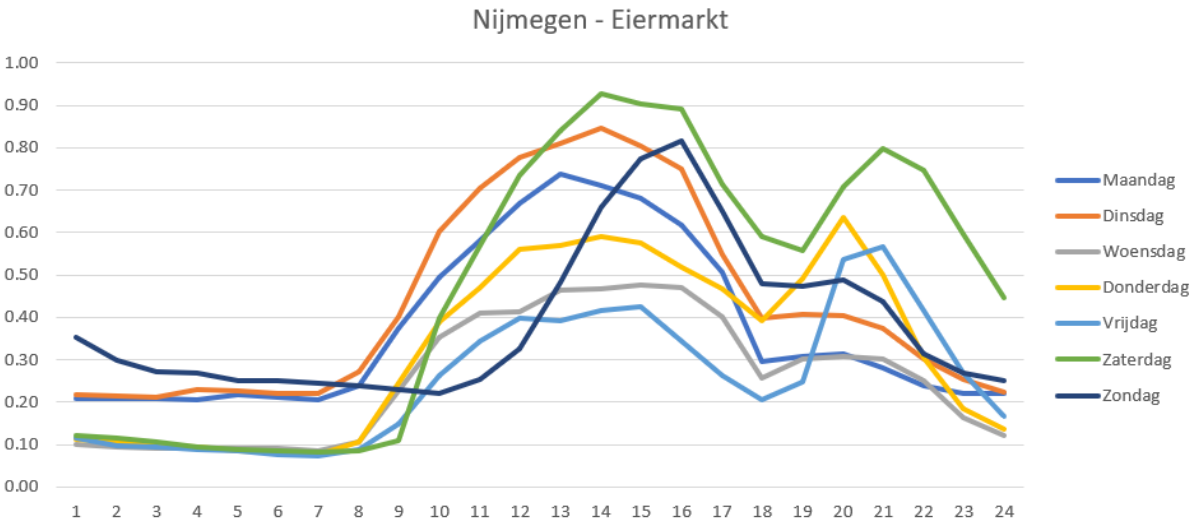
P10 Concordiastraat	280/455	Breda	Centrum	Nee
P11 Gasthuisvelden Arsenaalpad	335	Breda	Centrum	
P18-P Boompjes	200	Rotterdam	Buiten/haven	
P19-P Museumpark	352	Rotterdam	Centrum	
P22-P Kralingse Zoom	1690	Rotterdam	Buitenwijk	
P23-P Schouwburgplein 2	731	Rotterdam	Centrum/station	
P27-P Colosseumweg	180	Rotterdam	Buitenwijk	
P30-P+R Slinge Benedendek	415	Rotterdam	Buitenwijk	
P34-P+R Meijersplein	497	Rotterdam	Buitenweg/snelweg	
P35-P+R Schiedam Centrum	168	Schiedam	Centrum	
P36-P+R Alexander	497	Rotterdam	Buitenwijk/snelweg	
P37-P+R Beverwaard	505	Rotterdam	Buitenwijk/snelweg	
P40-P+R Vijfsluizen	88	Schiedam	Station/snelweg	
P41-P+R Capelsebrug	370	Rotterdam	Transferium/snelweg	
P43-P+R Noorderhelling	541	Rotterdam	Buitenwijk/snelweg	
P47-P Hoogvliet Centrum	177	Rotterdam	Buitenwijk	
P301-P+R Slinge Bovendek	420	Rotterdam	Buitenwijk	
Paardenveld	384	Utrecht	Buitenwijk	
PAP01 Marktplein	680	Apeldoorn	Centrum	
PAP02 Koningshaven	500	Apeldoorn	Centrum	
PAP03 Brinkpark Kortparkeerders	109	Apeldoorn	Centrum	
Park & Ride Spoorstraat	Geen data	Breda	Station	
Rhijnauwen Lus (P2)	65	Utrecht	Buitenwijk	
Rhijnauwen Theehuis (P1)	95	Utrecht	Buitenwijk	
Scheepssingel	96	Harderwijk	Haven	
Springweg	410	Utrecht	Centrum	
Stadhuis	190	Harderwijk	Centrum	
Station Europapark	100	Groningen	Buitenwijk	
Stille Wei	Geen data	Harderwijk	Centrum	
Terrein Stadspark 1	Geen data	Maastricht	Centrum	
Terrein Stadspark 2	Geen data	Maastricht	Centrum	
Terrein Stadspark 3	Geen data	Maastricht	Centrum	
Triade	540	Assen	Centrum	
Vaartsche Rijn	209	Utrecht	Buitenwijk	
Vuldersbrink	84	Harderwijk	Centrum	
Westeinde	146	Harderwijk	Station	
Zaailand	720	Leeuwarden	centrum	

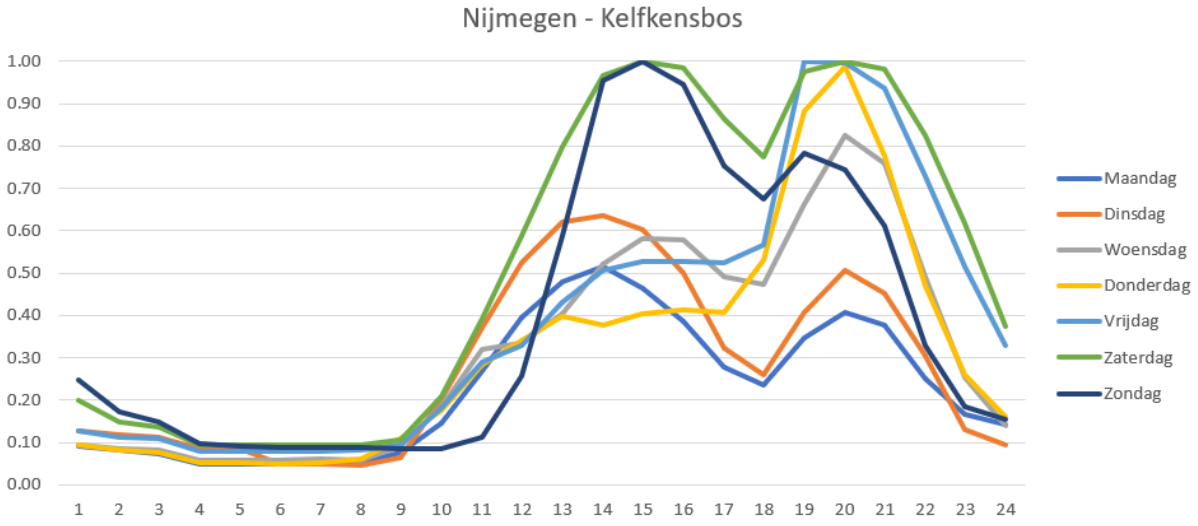
Appendix C

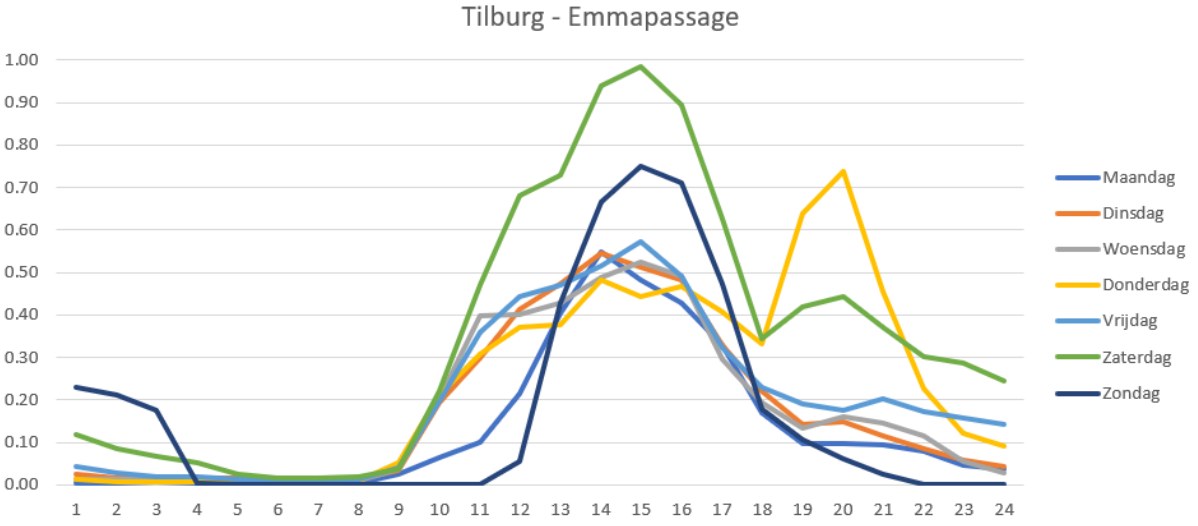

Deventer - Centrum
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag

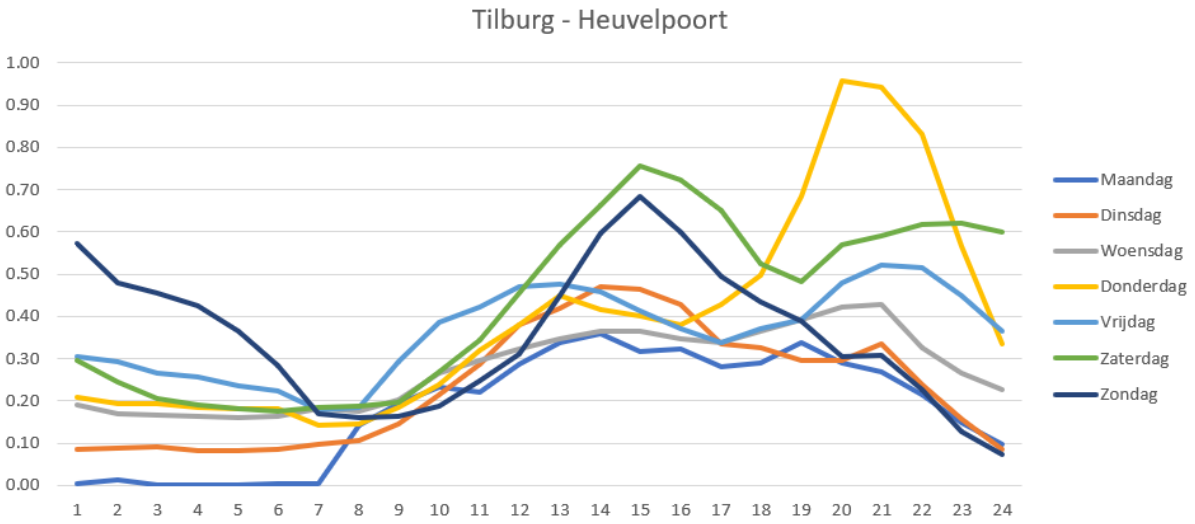

Apeldoorn - Markt
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag

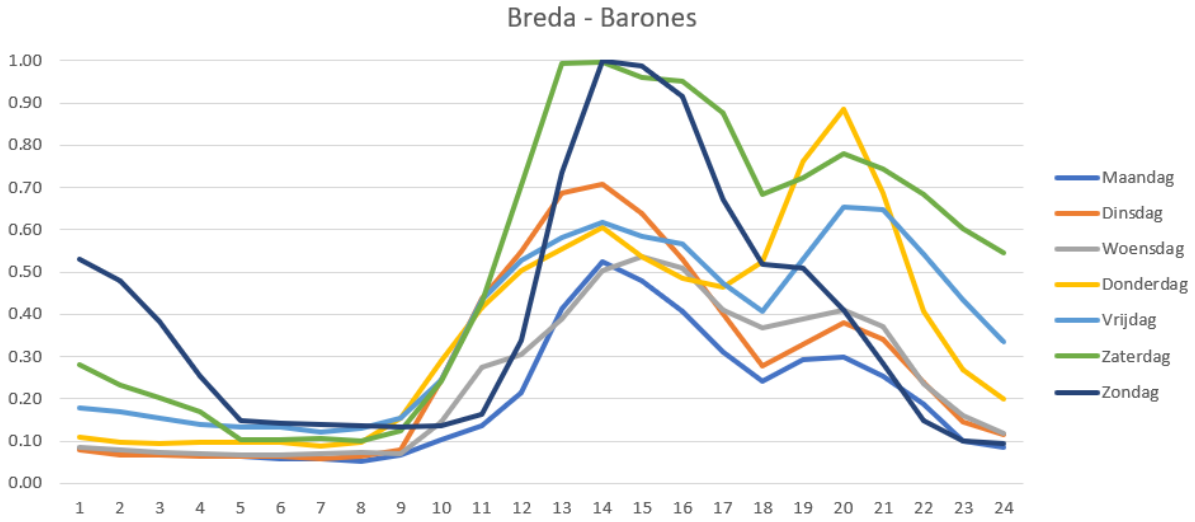

Harderijk - Houtwal
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag

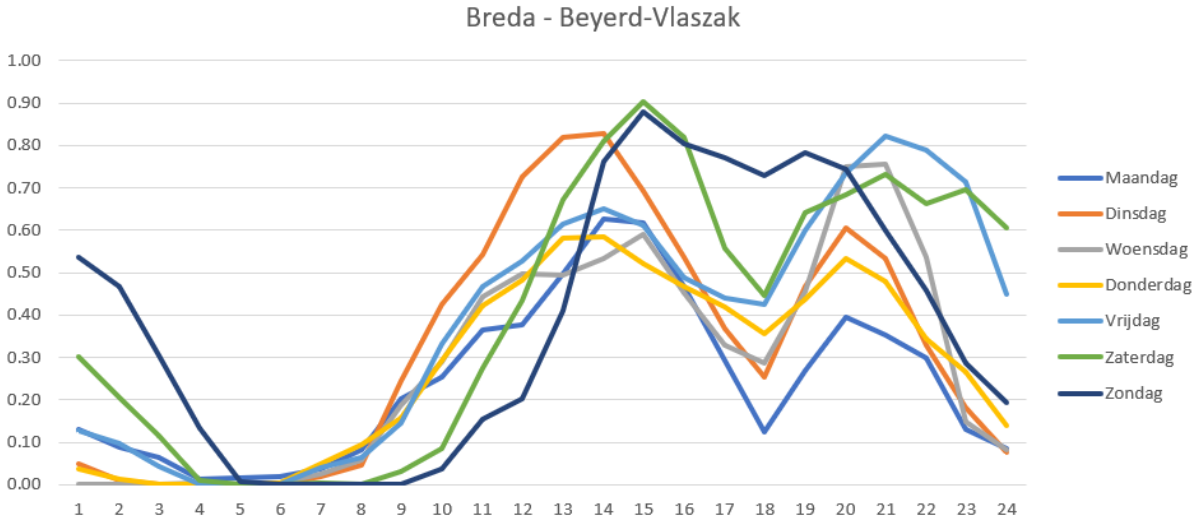

Nijmegen - Eiermarkt
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag

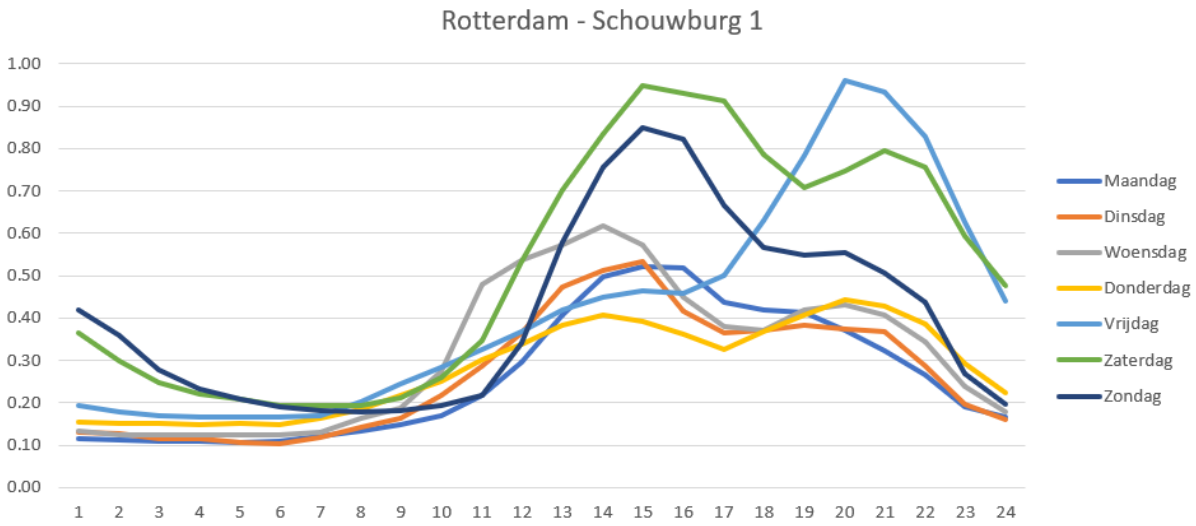

Nijmegen - Kelfkensbos
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag

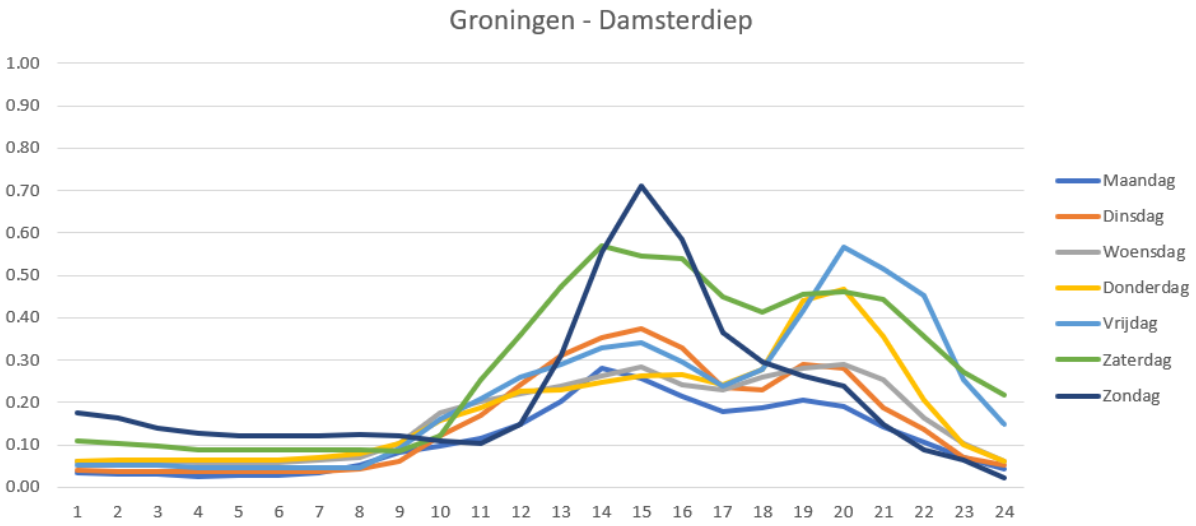

Tilburg - Emmapassage
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag

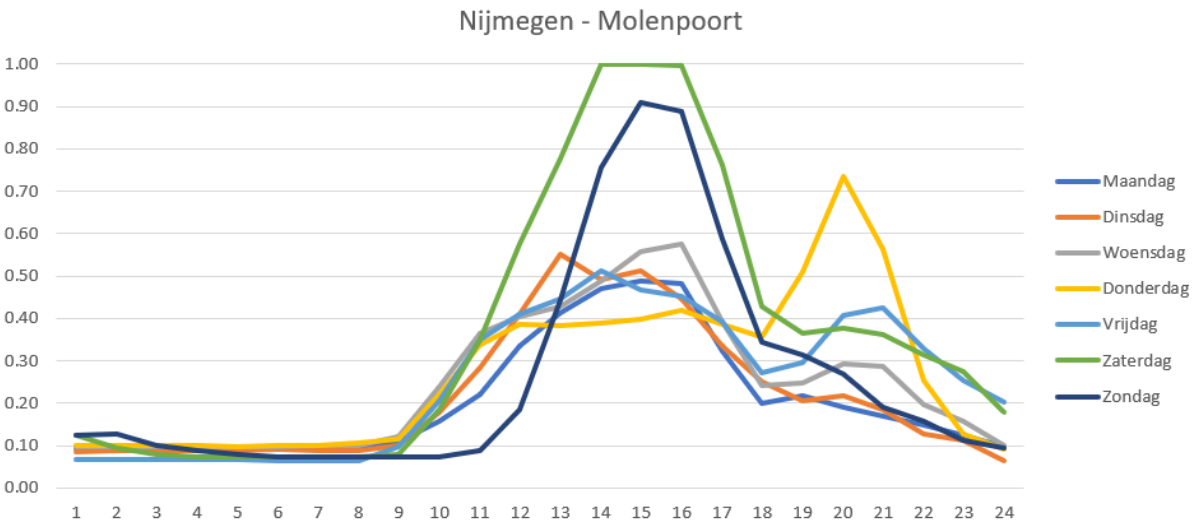

Tilburg - Heuvelpoort
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag


Breda - Barones
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag


Breda - Beyerd-Vlaszak
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag


Rotterdam - Schouwburg 1
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag


Groningen - Damsterdiep
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag


Nijmegen - Molenpoort
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maandag
Dinsdag
Woensdag
Donderdag
Vrijdag
Zaterdag
Zondag

Appendix D

DEVENTER – OVERIJSEL	
Vakantie (regio Noord)	Herfstvakantie: 20 oktober t/m 28 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Koopzondag: 7 oktober
	Koopzondag: 4 november
	Koopzondag: 25 november
	Koopzondag: 2 december
	Koopzondag: 16 december
	Koopzondag: 23 december
	Stoffenbeurs: 7 oktober (checken)
	Intocht sinterklaas: 1 december
	Dickens Festijn: 15 & 16 december
	Brink tot Brinkloop: 30 december

APELDOORN - GELDERLAND	
Vakantie (regio Midden)	Herfstvakantie: 20 oktober t/m 28 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Koopzondag: 7 oktober
	Koopzondag: 28 oktober
	Koopzondag: 4 november
	Koopzondag: 25 november
	Koopzondag: 2 december
	Koopzondag: 9 december
	Koopzondag: 16 december
	Koopzondag: 23 december
	Koopzondag: 30 december
	Fashionshow & najaar markt: 6 oktober
	Brocante markt: 20 & 21 oktober (checken)
	Avond Halloween shopping: 26 oktober
	Winter markt: 25 november
	Sinterklaas intocht: 1 december
	Wijnwandeling: 8 december (checken)
	Winterfair: 15 & 16 december (checken)

HARDERWIJK - GELDERLAND	
Vakantie (regio Midden)	Herfstvakantie: 20 oktober t/m 28 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Koopzondag: 7 oktober
	Koopzondag: 4 november
	Koopzondag: 25 november
	Koopzondag: 2 december
	Koopzondag: 9 december
	Koopzondag: 16 december
	Koopzondag: 23 december
	Koopzondag: 30 december
	Popronde: 14 oktober
	Intocht sinterklaas: 17 november
	Kofferbak verkoop in de garage: 25 november
	Kerstmarkt: 13 december

NIJMEGEN - GELDERLAND	
Vakantie (regio Zuid)	Herfstvakantie: 13 oktober t/m 21 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Iedere zondag koopzondag, wel op de eerste zondag van de maand de meeste winkels open. Niet echt een evenement
	Kermis: 6 t/m 14 oktober
	Intocht sinterklaas: 17 november
	Zondagsmarkt: 2 december
	Extra koopavond: 4 december
	Extra koopavond: 21 december
	Koopzondag + extra's: 23 december

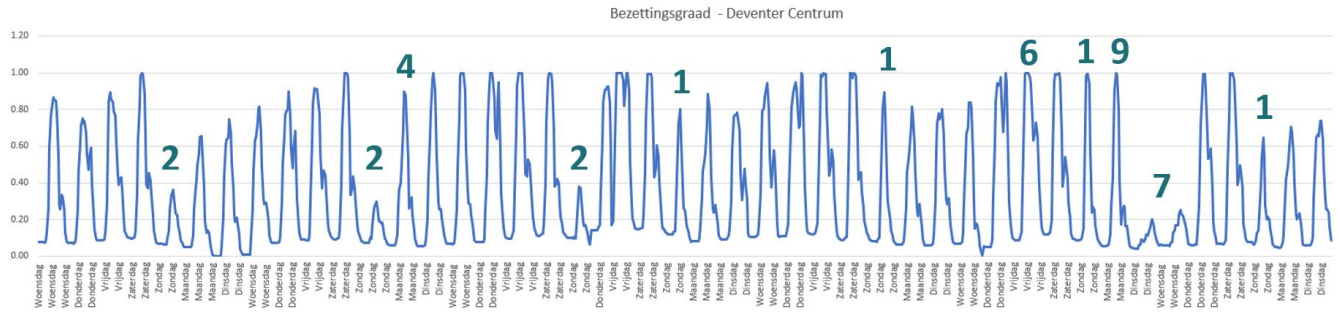
TILBURG – NOORD BRABANT	
Vakantie (regio Zuid)	Herfstvakantie: 13 oktober t/m 21 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Iedere zondag koopzondag, niet echt een event
	Popronde: 12 oktober
	Elfde van de elfde: 11 november
	Intocht sinterklaas: 18 november
	Tilburg Zingt Wintereditie: 15 december
	Kerstmarkt: 16 december

BREDA – NOORD BRABANT	
Vakantie (regio Zuid)	Herfstvakantie: 13 oktober t/m 21 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Iedere zondag koopzondag, niet echt een event
	Singelloop: 7 oktober
	Popronde: 15 november
	Intocht sinterklaas: 17 november
	Rotary Santa Run: 16 december

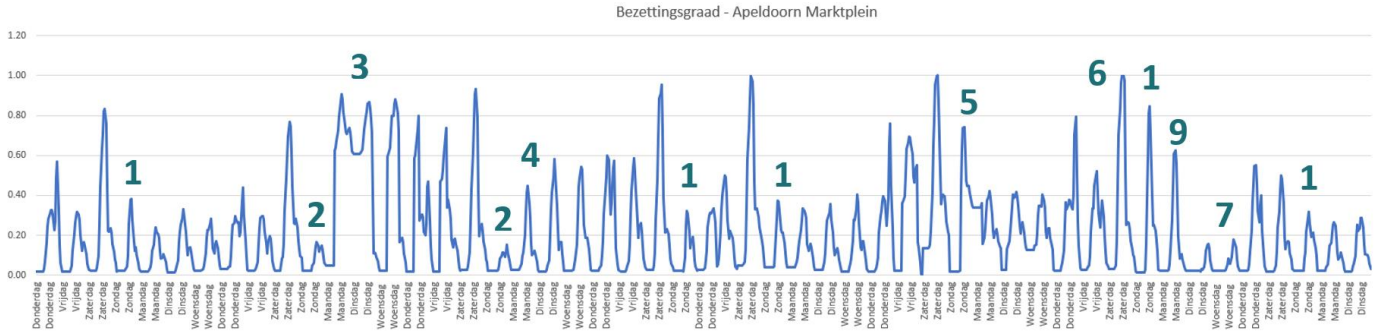
ROTTERDAM – ZUID HOLLAND	
Vakantie (regio Midden)	Herfstvakantie: 20 oktober t/m 28 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Iedere zondag koopzondag, niet echt een event
	PoPronde: 3 november
	Intocht Sinterklaas: 17 november
	Black Friday: 23 november

GRONINGEN – GRONINGEN	
Vakantie (regio Noord)	Herfstvakantie: 20 oktober t/m 28 oktober
	Kerstvakantie: 22 december t/m 6 januari
Event	Iedere zondag koopzondag, niet echt een event
	Popronde: 4 oktober
	4 mijl Groningen: 14 oktober
	Kadepop: 22 oktober
	Sinterklaasintocht: 17 november
	Kerstplaza/winterkermis: 25 december t/m 6 jan

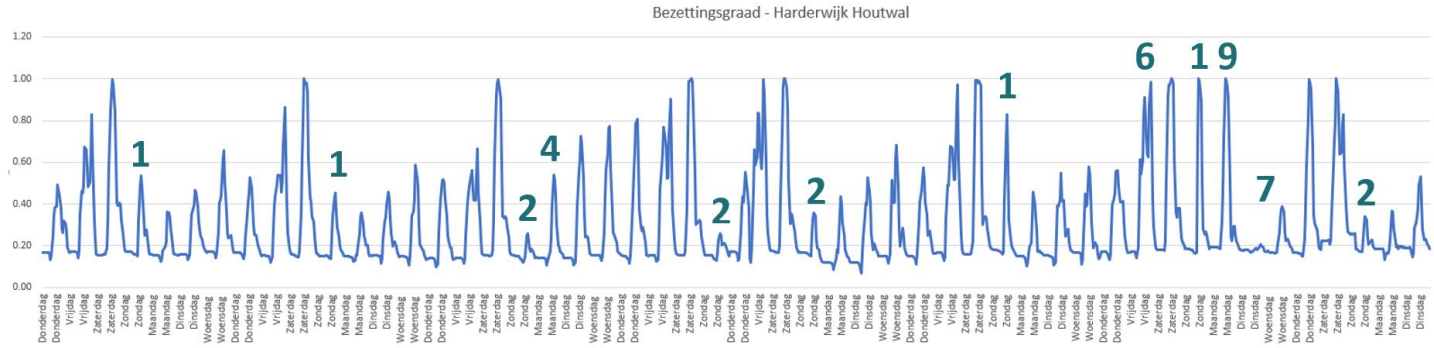
Appendix E



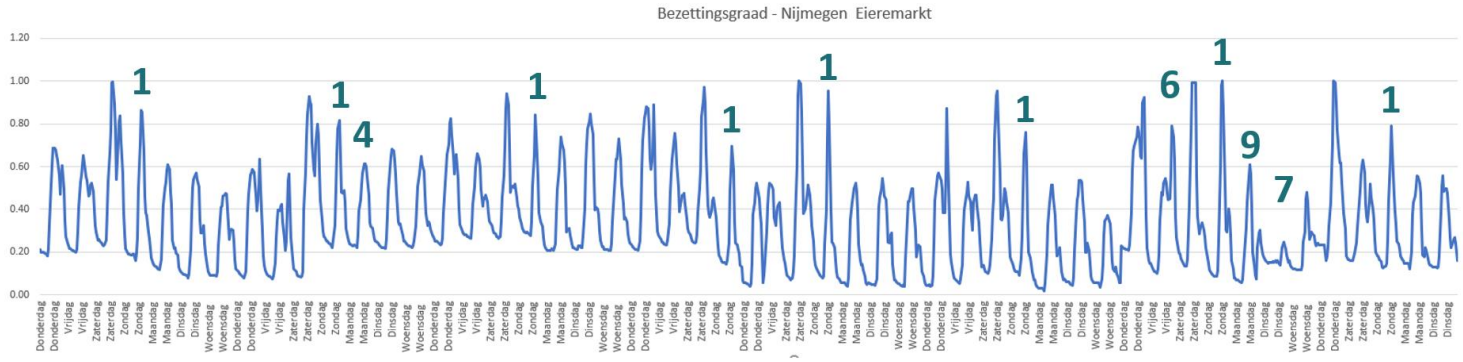
Deventer – centrum	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



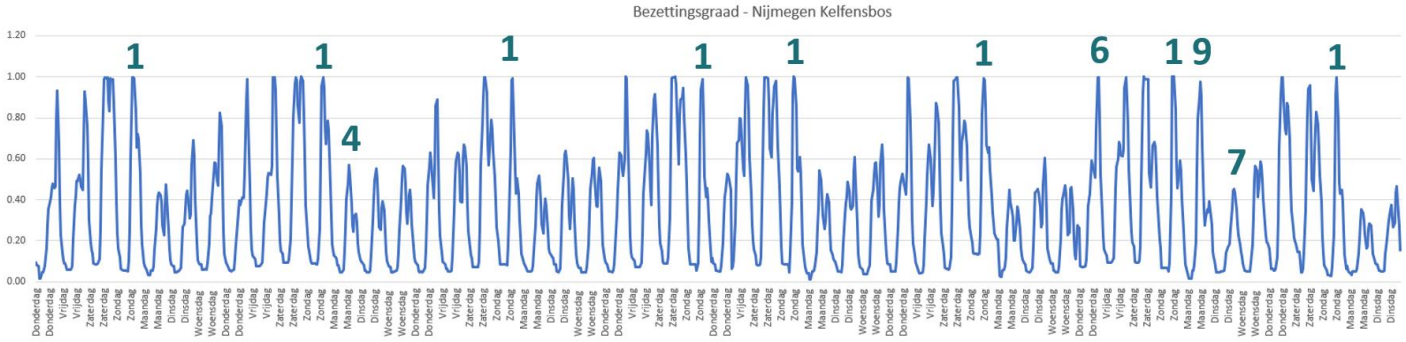
Apeldoorn - Marktplein	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



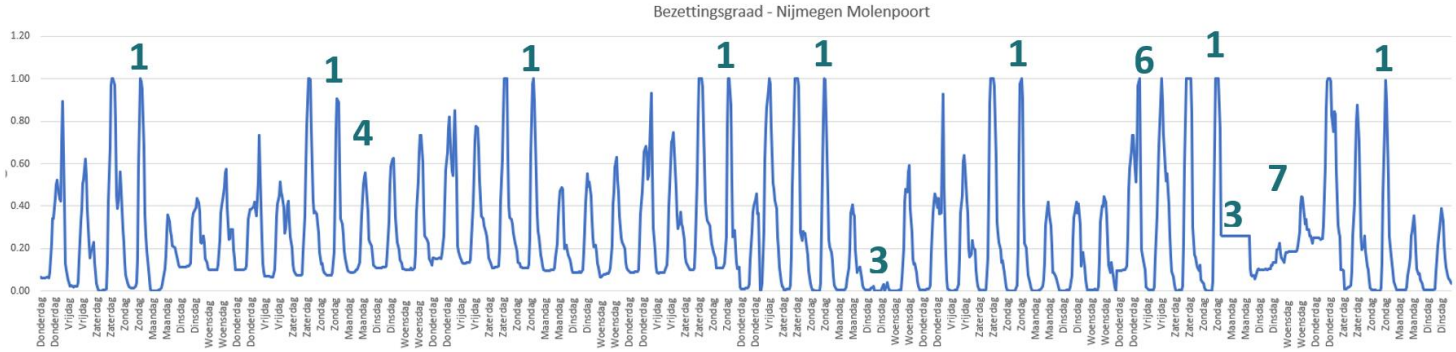
Harderwijk – Houtwal	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



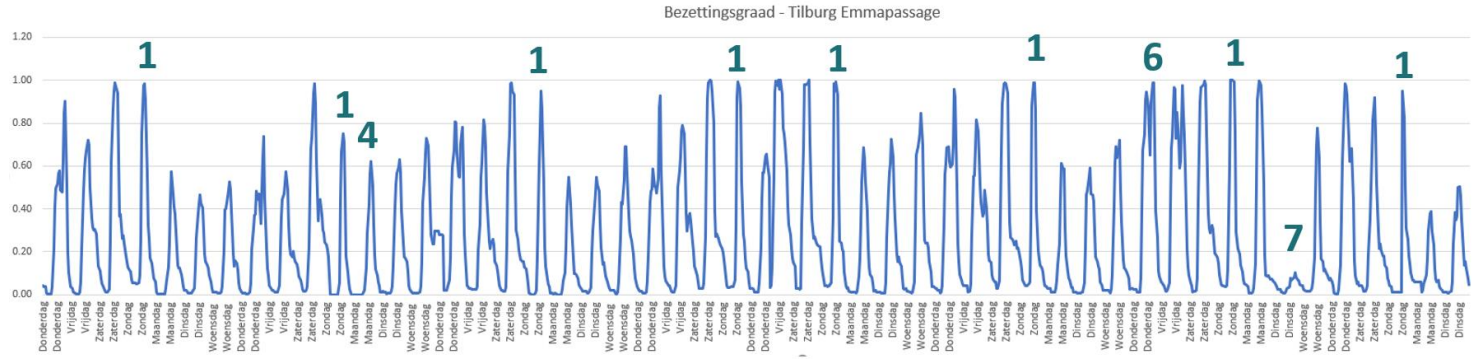
Nijmegen – Eiermarkt	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



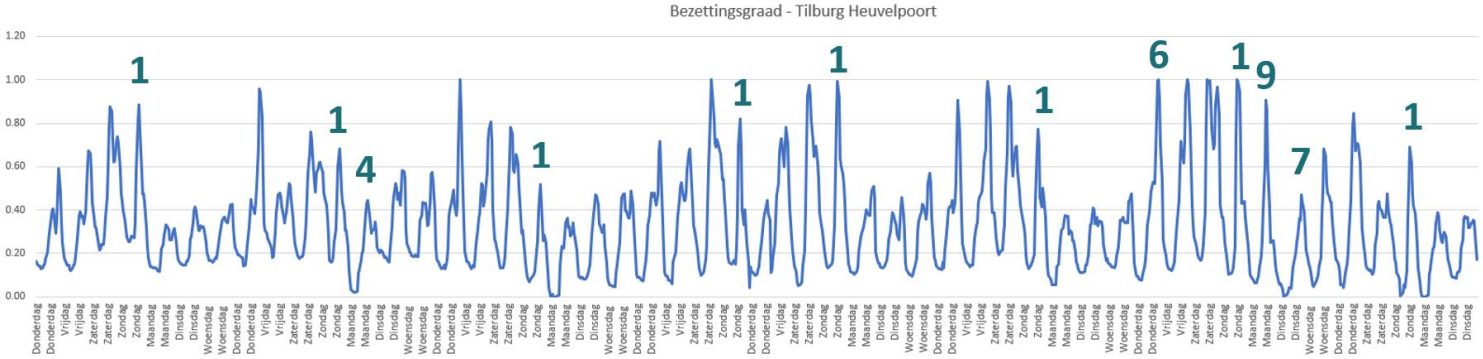
Nijmegen – Kelfensbos	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



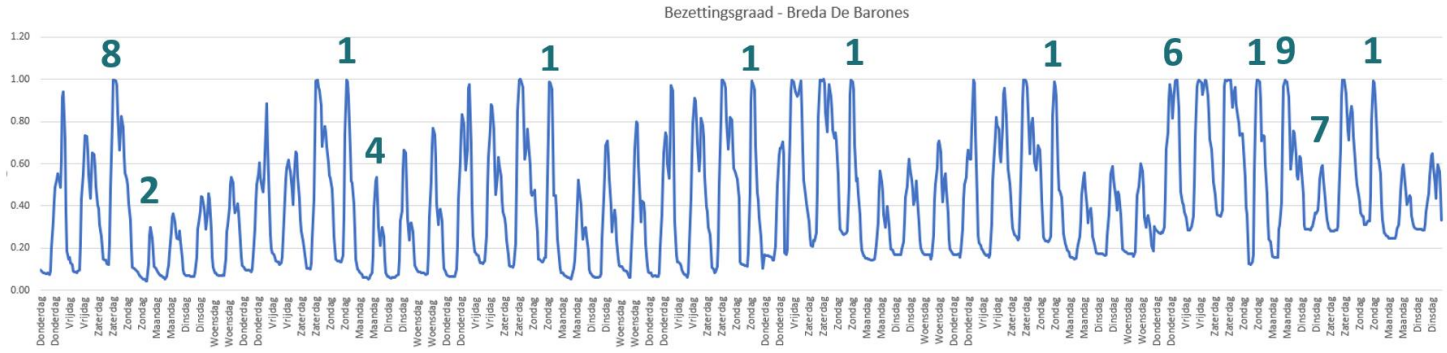
Nijmegen – Molenpoort	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



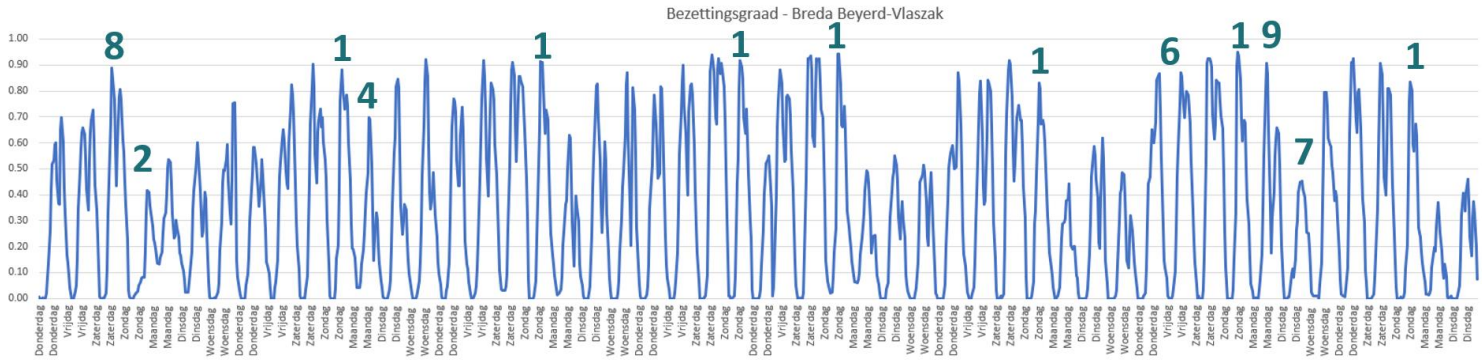
Tilburg - Emmapassage	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



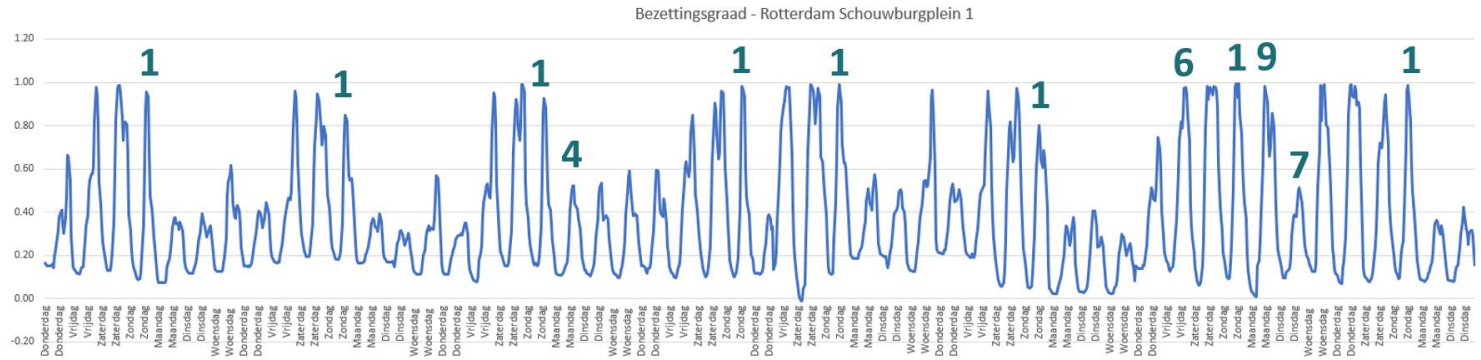
Tilburg - Heuvelpoort	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
9	Dag voor kerst



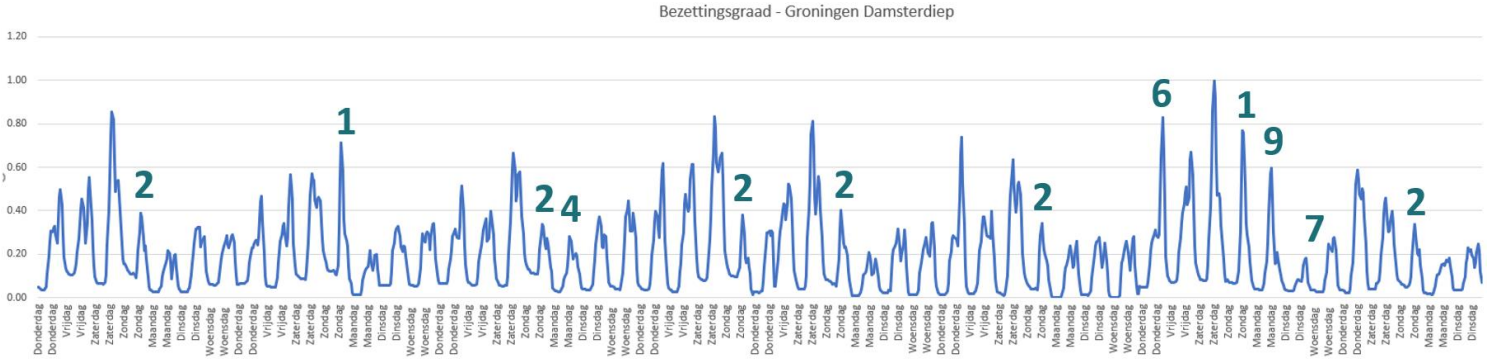
Breda – De Barones	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
8	Singelloop/marathon
9	Dag voor kerst



Breda – Beyerd-Vlaszak	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
8	Singelloop/marathon
9	Dag voor kerst



Rotterdam – Schouwburgplein 1	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
8	Singelloop/marathon
9	Dag voor kerst



Groningen – Damsterdiep	Verklaring
1	Koopzondag
2	Geen koopzondag
3	Error in de data
4	Herfstvakantie
5	Intocht Sinterklaas
6	Begin kerstvakantie + Inkopen kerst
7	Eerste & tweede kerstdag
8	Singelloop/marathon
9	Dag voor kerst

Appendix F

Stad	Locatie weerstation	Afstand (km)
Deventer	Hupsel	52
Apeldoorn	Deelen	19
Harderwijk	De Bilt	52
Nijmegen	Volkel	52
Tilburg	Gilze Rijen	12
Breda	Woensdrecht	48
Rotterdam	Rotterdam	0
Groningen	Eelde	12

Appendix G

Parkeergarage	Capaciteit	Maximum	Gemiddelde	Mediaan
Deventer - Centrum	481	481	184	126
Apeldoorn – Marktplein	680	680	140	86
Harderwijk – Houtwal	432	432	127	83
Nijmegen – Eiermarkt	365	365	123	102
Nijmegen – Kelfkensbos	600	600	205	161
Nijmegen – Molenpoort	295	295	78	40
Tilburg - Emmapassage	375	375	122	108
Tilburg – Heuvelpoort	510	510	216	180
Breda – De Barones	365	346	123	107
Breda – Beyerd-Vlaszak	685	685	247	195
Rotterdam – Schouwburgplein 1	392	391	74	54
Groningen – Molenpoort	295	295	74	46

Appendix H

Parkeergarage	Periode	Reden
Deventer - Centrum	4-10 t/m 9-10	Error
Breda - De Barones	27 december	Error
Apeldoorn - Marktplein	15 & 16 oktober	Error
Nijmegen - Molenpoort	27 & 28 november	Error
Nijmegen - Molenpoort	24 december	Error

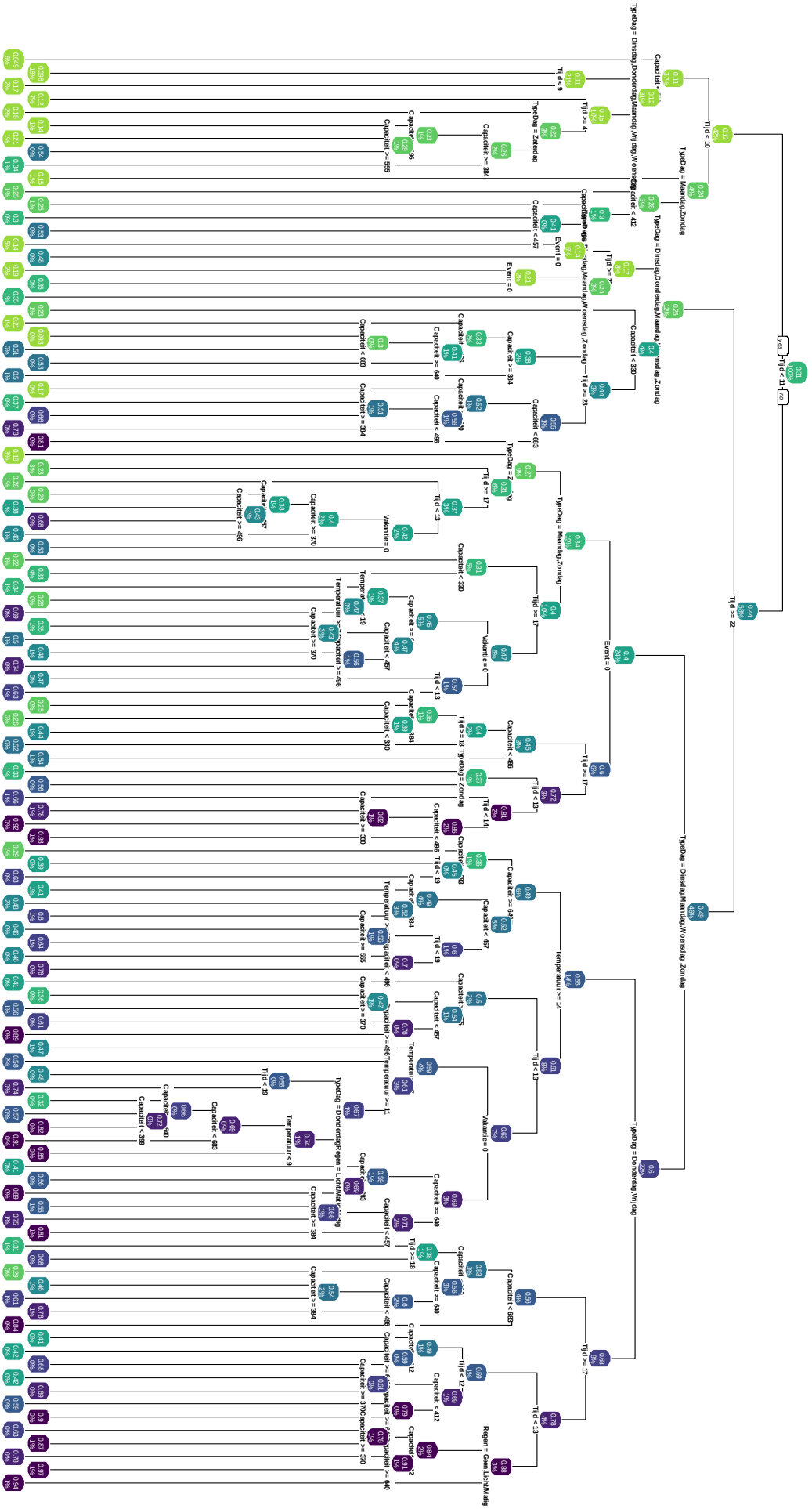
Appendix I

Test set 1	Datum	Type dag
Deventer – Centrum	18-10-2018	Donderdag
	1-12-2018	Dinsdag
	4-12-2018	Zaterdag
Apeldoorn – Marktplein	6-10-2018	Zaterdag
	24-10-2018	Woensdag
	23-12-2018	Zondag
Harderwijk – Houtwal	6-10-2018	Zaterdag
	22-10-2018	Maandag
	5-12-2018	Woensdag
Nijmegen – Eiermarkt	19-10-2018	Vrijdag
	4-12-2018	Dinsdag
	6-1-20194	Zondag
Nijmegen – Kelkensbos	13-10-2018	Zaterdag
	26-10-2018	Vrijdag
	26-11-2018	Maandag
Nijmegen – Molenpoort	10-10-2018	Woensdag
	16-10-2018	Dinsdag
	21-12-2018	Vrijdag
Tilburg – Emmapassage	25-11-2018	Zondag
	27-12-2018	Donderdag
	8-1-2019	Dinsdag
Tilburg – Heuvelpoort	15-10-2018	Maandag
	23-11-2018	Vrijdag
	1-12-2018	Zaterdag
Breda – De Barones	4-10-2018	Donderdag
	9-10-2018	Dinsdag
	30-11-2018	Vrijdag
Breda – Beyerd-Vlaszak	8-10-2018	Maandag
	25-11-2018	Zondag
	20-12-2018	Donderdag
Rotterdam – Schouwburgplein	5-10-2018	Vrijdag
	17-10-2018	Woensdag
	22-10-2018	Maandag
Groningen – Damsterdiep	11-10-2018	Donderdag
	23-10-2018	Dinsdag
	6-1-2019	Zondag

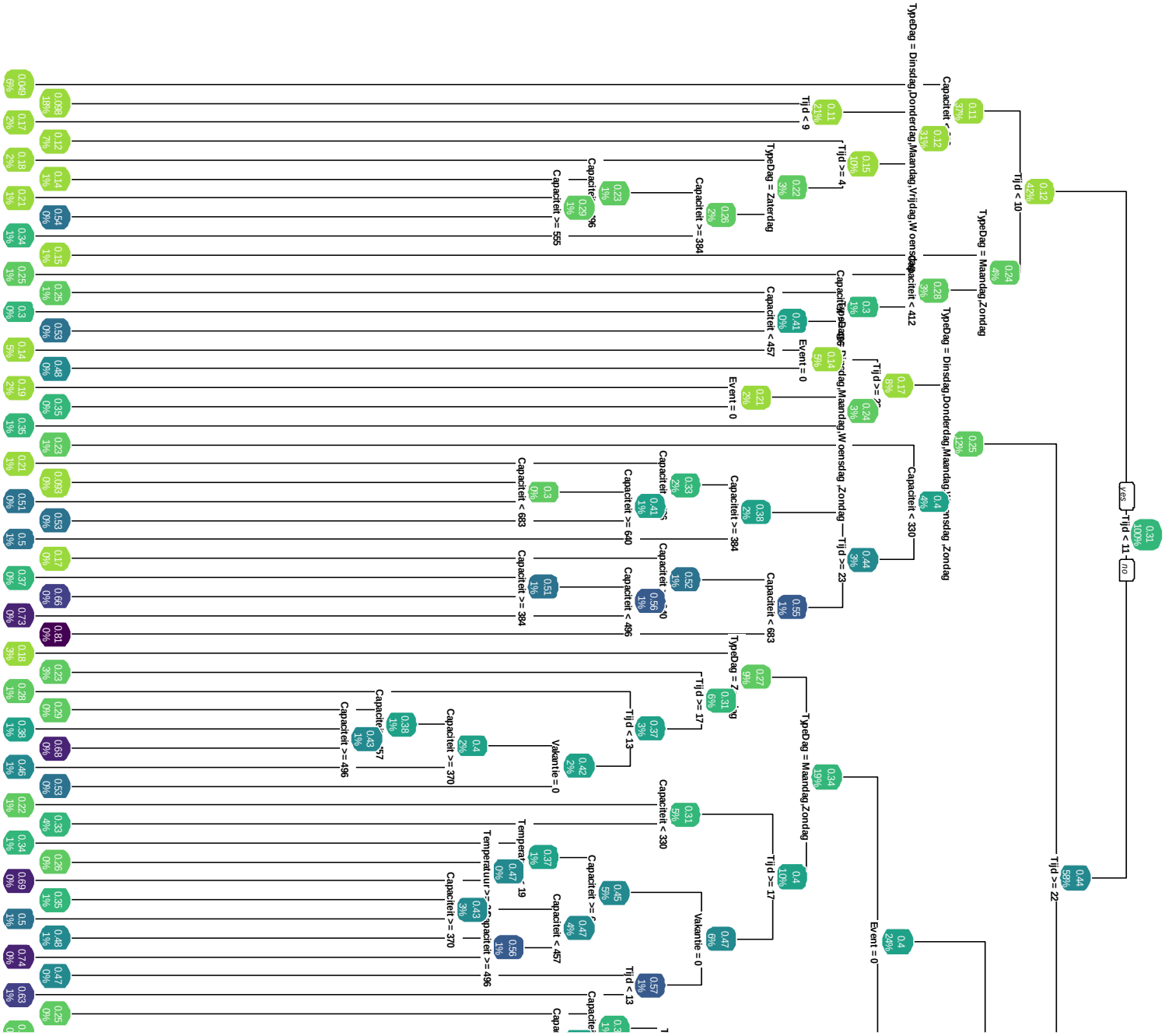
Test set 2	Datum	Type dag
Deventer – Centrum	23-11-2018	Vrijdag
	1-12-2018	Zaterdag
	4-12-2018	Dinsdag
Apeldoorn – Marktplein	4-10-2018	Donderdag
	24-10-2018	Woensdag
	1-12-2018	Zaterdag
Harderwijk – Houtwal	6-10-2018	Zaterdag
	20-12-2018	Donderdag
	23-12-2018	Zondag
Nijmegen – Eiermarkt	22-10-2018	Maandag
	28-10-2018	Zondag
	5-12-2018	Woensdag
Nijmegen – Kelkensbos	22-11-2018	Donderdag
	6-1-2019	Zondag
	7-1-2019	Maandag
Nijmegen – Molenpoort	6-10-2018	Zaterdag
	21-10-2018	Zondag
	26-11-2018	Maandag
Tilburg – Emmapassage	18-10-2018	Donderdag
	26-10-2018	Vrijdag
	27-11-2018	Dinsdag
Tilburg – Heuvelpoort	25-11-2018	Zondag
	28-11-2018	Woensdag
	21-12-2018	Vrijdag
Breda – De Barones	10-10-2018	Woensdag
	16-10-2018	Dinsdag
	27-10-2018	Zaterdag
Breda – Beyerd-Vlaszak	23-11-2018	Vrijdag
	1-12-2018	Zaterdag
	8-1-2019	Dinsdag
Rotterdam – Schouwburgplein	15-10-2018	Maandag
	17-10-2018	Woensdag
	30-11-2018	Vrijdag
Groningen – Damsterdiep	8-10-2018	Maandag
	9-10-2018	Dinsdag
	20-12-2018	Donderdag

Test set 3	Datum	Type dag
Deventer – Centrum	13-10-2018	Zaterdag
	27-11-2018	Dinsdag
	20-12-2018	Donderdag
Apeldoorn – Marktplein	5-10-2018	Vrijdag
	21-10-2018	Zondag
	5-12-2018	Woensdag
Harderwijk – Houtwal	22-11-2018	Donderdag
	26-11-2018	Maandag
	6-1-2019	Zondag
Nijmegen – Eiermarkt	11-10-2018	Donderdag
	13-10-2018	Zaterdag
	26-10-2018	Vrijdag
Nijmegen – Kelkensbos	16-10-2018	Dinsdag
	28-11-2018	Woensdag
	21-12-2018	Vrijdag
Nijmegen – Molenpoort	7-10-2018	Zondag
	9-10-2018	Dinsdag
	8-1-2019	Dinsdag
Tilburg – Emmapassage	15-10-2018	Maandag
	23-11-2018	Vrijdag
	1-12-2018	Zaterdag
Tilburg – Heuvelpoort	4-10-2018	Donderdag
	30-11-2018	Vrijdag
	3-12-2018	Maandag
Breda – De Barones	8-10-2018	Maandag
	25-11-2018	Zondag
	5-12-2018	Woensdag
Breda – Beyerd-Vlaszak	6-10-2018	Zaterdag
	17-10-2018	Woensdag
	29-11-2018	Donderdag
Rotterdam – Schouwburgplein	18-10-2018	Donderdag
	22-10-2018	Maandag
	2-12-2018	Zondag
Groningen – Damsterdiep	17-10-2018	Woensdag
	27-11-2018	Dinsdag
	1-12-2018	Zaterdag

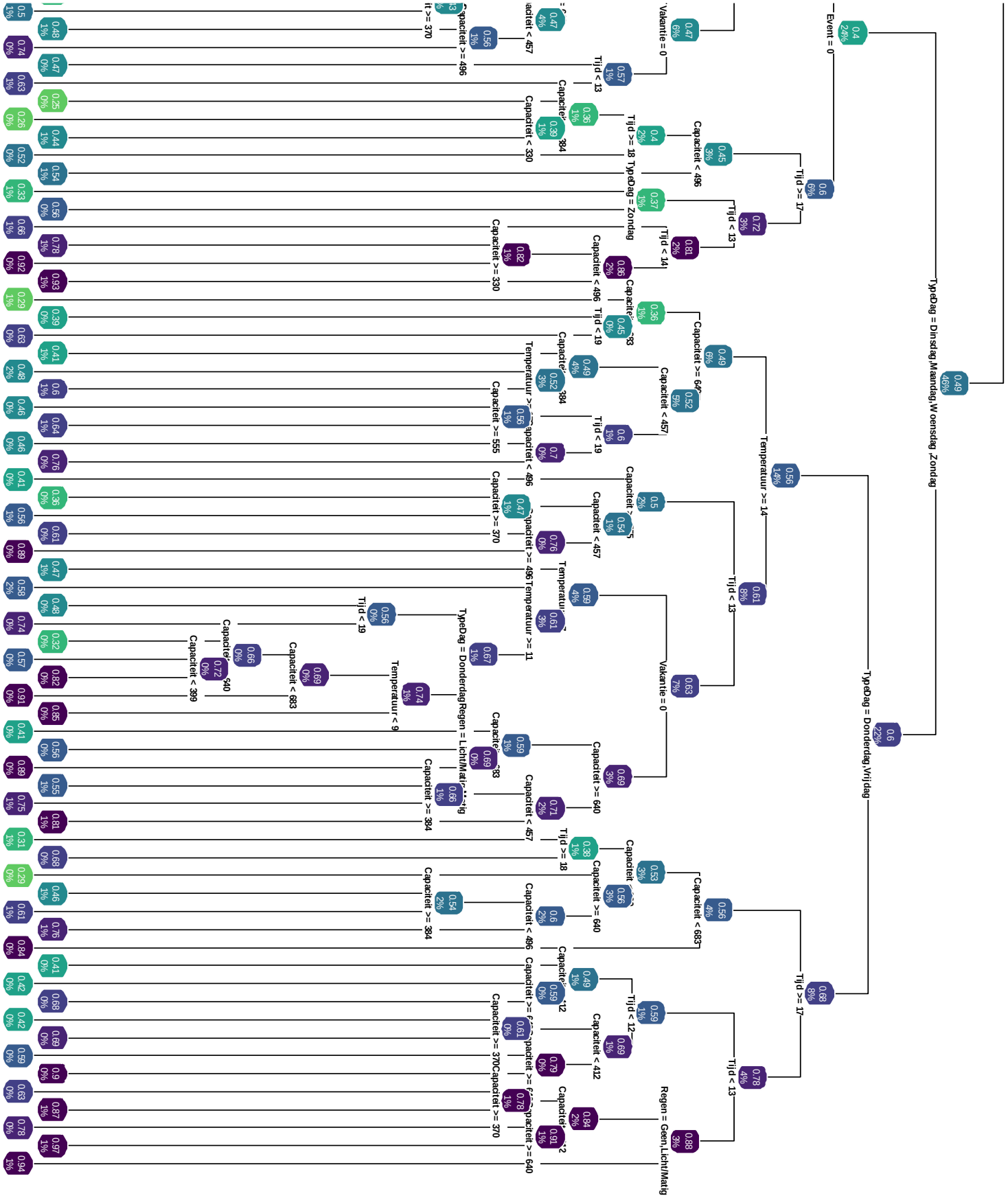
Appendix J



DEEL 1


TypeDag = Dinsdag,Donderdag,Maandag,Vrijdag,Woensdag,Zondag
Capaciteit < 412
TypeDag = Maandag,Zondag
Tijd < 10
Tijd < 11
Tijd >= 22
Capaciteit < 330
Event = 0
Tijd >= 23
Capaciteit < 683
TypeDag = Zaterdag
Tijd >= 17
Tijd < 13
Capaciteit >= 384
Capaciteit >= 640
Capaciteit >= 555
Capaciteit >= 496
Capaciteit >= 370
Capaciteit < 457
Vakantie = 0
Tijd >= 4
Tijd < 9

DEEL 2


0.4
24%
Event = 0
TypeDag = Dinsdag,Maandag,Woensdag,Zondag
0.49
46%
0.6
22%
TypeDag = Donderdag,Vrijdag
0.6
6%
Tijd >= 17
0.56
14%
Temperatuur >= 14
0.68
8%
Tijd >= 17
0.72
3%
Tijd < 13
0.61
8%
Tijd < 13
0.78
4%
Tijd < 13
0.47
6%
Vakantie = 0
0.63
7%
Vakantie = 0
0.88
3%
Regen = Geen,Licht/Matig
0.57
1%
Tijd < 13
0.4
2%
Tijd >= 18
0.81
2%
Tijd < 14
0.49
6%
Capaciteit >= 640
0.5
2%
0.56
4%
Capaciteit < 683
0.59
1%
Tijd < 12
0.86
2%
Capaciteit < 496
0.52
5%
Capaciteit < 457
0.54
1%
Capaciteit < 457
0.69
3%
Capaciteit >= 640
0.53
3%
0.56
3%
Capaciteit >= 640
0.49
1%
0.69
1%
Capaciteit < 412
0.84
2%
0.91
1%
0.47
4%
Capaciteit < 457
0.56
1%
Capaciteit >= 496
0.36
1%
0.39
1%
Capaciteit < 384
0.45
3%
Capaciteit < 496
0.37
1%
TypeDag = Zondag
0.82
1%
Capaciteit >= 330
0.36
1%
0.45
0%
0.49
4%
0.52
3%
Capaciteit >= 384
0.56
1%
Capaciteit >= 555
0.7
0%
Capaciteit < 496
0.47
1%
Capaciteit >= 370
0.76
0%
Capaciteit >= 496
0.59
4%
0.61
3%
Temperatuur >= 11
0.67
1%
TypeDag = Donderdag
0.59
1%
0.69
0%
0.71
2%
Capaciteit < 457
0.38
1%
Tijd >= 18
0.54
2%
Capaciteit >= 384
0.6
2%
Capaciteit < 496
0.59
0%
0.61
0%
0.79
0%
0.78
1%
Capaciteit >= 370
0.91
1%
Capaciteit >= 640
0.56
0%
Tijd < 19
0.66
0%
0.72
0%
Capaciteit < 640
0.69
0%
Capaciteit < 683
0.74
1%
Temperatuur < 9
0.66
1%
Capaciteit >= 384
0.6
1%
Tijd < 19
0.76
0%
Capaciteit >= 496
0.61
3%
Temperatuur >= 11
0.67
1%
TypeDag = Donderdag
0.59
1%
0.69
0%
0.71
2%
Capaciteit < 457
0.38
1%
Tijd >= 18
0.54
2%
Capaciteit >= 384
0.6
2%
Capaciteit < 496
0.59
0%
0.61
0%
0.79
0%
0.78
1%
Capaciteit >= 370
0.91
1%
Capaciteit >= 640
0.56
0%
Tijd < 19
0.66
0%
0.72
0%
Capaciteit < 640
0.69
0%
Capaciteit < 683
0.74
1%
Temperatuur < 9
0.66
1%
Capaciteit >= 384