Universiteit Twente

Bacheloropdracht

Het vermogen om te winnen

Auteurs:
Imke Gerritsma
Tom Sniekers
Elieke van Sark

Begeleider:
Dr. Ir. Maurits de Graaf

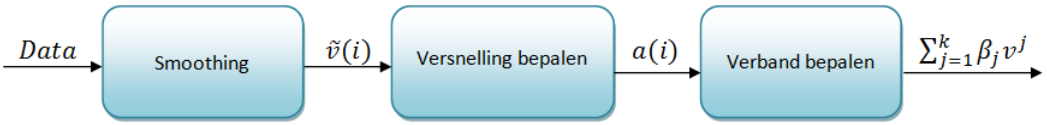
17 juni 2013

Samenvatting

Dit verslag is geschreven naar aanleiding van een vraagstuk van het Solar Team. Het Solar Team Twente heeft het doel om de World Solar Challenge te winnen. Ze ontwerpen en bouwen een zonne-auto om het parcours door Australië af te leggen. Deze zonne-auto ontvangt energie met behulp van zonnecellen. Echter, de capaciteit van de accu is beperkt. Daarom moet er efficiënt met de energie omgegaan worden. Daarentegen moet ook een race gewonnen worden, dus de snelheid moet hoog liggen. Het is voor het opstellen van een goede strategie tijdens de race van belang om te weten hoeveel vermogen (en dus energie) geleverd moet worden om met een bepaalde snelheid te rijden. Ons doel van het onderzoek is het ontwikkelen van een methode die, uit een ruwe dataset van de snelheden ten op zichte van de tijd, een zo nauwkeurig mogelijk verband vindt voor het vermogen bij een bepaalde snelheid.

Om dit vraagstuk op te kunnen lossen kijken we eerst naar de theorie. Er bestaan formules voor de krachten die op een (personen)auto werken, waaruit een verband voor het vermogen bij een snelheid kan worden afgeleid. Uit eerder onderzoek is gebleken dat, door de specifieke karaktereigenschappen van de zonne-auto, de formules niet direct toepasbaar zijn. We hebben daarom de formules niet direct toegepast, maar als leidraad gebruikt. We hebben alle constantes in de theoretische formule samengenomen. Hierdoor kunnen we ons probleem specifieker formuleren: we zoeken een polynoom die het verband tussen versnelling en snelheid representeert.

De twee datasets die we ontvangen om het verband te bepalen zijn afkomstig van uitroltests van de testauto, mock-up genoemd. De data zet de snelheid tegen de tijd uit. Echter door meetonnauwkeurigheden is deze data niet meteen bruikbaar, maar zullen we deze gladder moeten maken. Hierna willen we uit de gesmoothde snelheden de versnellingen bepalen. Als we de versnellingen hebben bepaald, kunnen we een polynoom vinden voor de versnelling afhankelijk van de snelheid. Ons proces ziet er als volgt uit.



Voor het eerste blok, smoothing, hebben we twee methodes getest: het Nearest-neighbor smoothen en de Savitzky-Golay methode. Voor het tweede blok hebben we vier methodes ontwikkeld. Dit zijn twee numerieke methodes, A en B, een analytische methode en een toepassing van de Savitzky-Golay methode. Deze methodes hebben we met elkaar vergeleken en getest. Voor het bepalen van het polynoom gebruiken we lineaire regressie met als schatter de kleinste kwadraten schatter.

We hebben in eerste instantie de methodes vergeleken en de aannames gecontroleerd. Hiervoor hebben we de metingen van de eerste dataset van een uitroltest van de mock-up gebruikt. Uit het vergelijken van de vier methodes in combinatie met het smoothen volgt dat de verschillende methodes verschillende polynomen vinden. Om te bepalen welke methode het beste is hebben we een programma geschreven aan de hand van een bekend verband. Dit programma genereert meetwaarden aan de hand van het verband waar een normaalverdeelde fout bij op wordt geteld. Vervolgens hebben we gekeken welke methode het beste in staat is om uit deze datapunten het bekende verband terug te vinden. Het blijkt dat de Savitzky-Golay methode het gezochte verband exact terugvindt en daarom concluderen we dat dit de methode is die we moeten gebruiken om datapunten te smoothen en de versnellingen te bepalen. We zullen met behulp van de bepaalde versnellingen een verband voor de snelheid tegen de versnelling moeten vinden. Uit de literatuur blijkt dat lineaire regressie met als schatter de kleinste kwadratenschatter hiervoor de beste methode is. Onze uiteindelijke methode wordt als volgt weergeven.

Uiteindelijk wordt de ontwikkelde methode uitgevoerd op de tweede dataset van de mock-up. Deze dataset bestaat wederom uit snelheidsmetingen tijdens een uitroltest. Het gevonden verband ziet er als volgt uit.

$$P(v) = 11,96v + 3,22v^2 + 0,23v^3.$$

Dit verband hebben we besproken met het Solar Team Twente en zij veronderstelden dat de orde van grootte van de coëfficiënten realistisch was.

Om te bepalen hoeveel data er nodig is om een betrouwbaar verband te krijgen hebben we op de tweede dataset een gevoeligheidsanalyse toegepast. Het blijkt dat bij gegevens van 6 uitroltest, met 80% van de datapunten eenzelfde verband gevonden wordt als bij alle datapunten. Ook blijkt dat dit verband een willekeurige 20% van de datapunten nog steeds goed beschrijft. Ook merken we op dat metingen bij de hoogste en laagste waarde van de snelheid grote invloed hebben op het verband dat gevonden wordt. Een laag aantal ritten is dus voldoende, maar hierbij is het van belang een zo groot mogelijk interval aan snelheden mee te nemen.

De conclusie is dat we aan de hand van uitroltests van de mock-up een methode hebben ontwikkeld die uit een ruwe dataset een goed verband vindt voor het vermogen bij een bepaalde snelheid. Dit gebeurt door de Savitzky-Golay methode te gebruiken om uit ruwe datapunten voor de snelheid meteen bruikbare datapunten voor de versnelling te verkrijgen. Met deze datapunten vinden we door gebruik te maken van lineaire regressie met de kleinste kwadraten schatter het verband. Onze methode kan echter meer dan dat. Op datasets die het Solar Team zal verkrijgen met de zonne-auto, zal onze methode toepasbaar zijn om een goed verband te krijgen. Ook als de snelheden veel hoger liggen dan de snelheden waar we nu mee hebben gewerkt en zelfs als het gevonden verband, zoals het Solar Team verwacht, afwijkt van de theorie zal het verband als goed moeten worden beschouwd. Ook zal het mogelijk zijn de wind als variabele mee te nemen, ondanks dat wij de wind als constant hebben beschouwd.

Tenslotte hebben we in het statistisch computerprogramma SPSS een syntax gemaakt, waarmee het Solar Team met enkele drukken op de knop het gewenste verband zal vinden. We hebben in een korte handleiding geschreven hoe wij SPSS en Matlab in ons onderzoek hebben gebruikt.

Inhoudsopgave

1 Introductie probleem

Het Solar Team Twente is een team dat bestaat uit studenten met als doel de World Solar Challenge te winnen. Dit is een wedstrijd waarbij er 3000 kilometer door Australië afgelegd moet worden met een zonneauto. De belangrijkste specificaties van deze auto zijn als volgt: de afmetingen van de auto mogen maximaal 4500 mm in de lengte, 1800 mm in de breedte en 1800 mm in de hoogte zijn. Daarbij mogen de zonnecellen in totaal 9 m^2 aan oppervlakte beslaan [Challenge, 2013]. De capaciteit van de accu in de zonne-auto is gelimiteerd op 5 kWh. Alle overige energie komt van de zon, of wordt teruggewonnen uit de kinetische energie van het voertuig. Om in de race optimaal te presteren is het belangrijk zo efficiënt mogelijk met de energie om te gaan. Er zal dus een goede verdeling gevonden moeten worden voor het energieverbruik.

Het is voor het Solar Team van belang om te weten hoe het werkelijk verbruikte vermogen afhangt van de snelheid, zodat er een goede rijstrategie bepaald kan worden. Zij willen zo precies mogelijk weten welk vermogen er geleverd moet worden om een bepaalde snelheid te rijden. De opdracht die het Solar Team voor ons heeft is als volgt geformuleerd.

Bepaal het verband tussen het werkelijk verbruikte vermogen en de snelheid.

De data die we gebruiken om het verband te bepalen is afkomstig uit tests met een testauto, ook wel de mock-up genoemd. Het verband voor het vermogen is afhankelijk van de soort auto. Gezien de mock-up niet dezelfde specificaties heeft als de zonne-auto en de weerstanden die op de mock-up werken niet overeenkomen met de weerstanden die op de zonne-auto werken, zullen we met de verkregen data geen verband voor de zonne-auto kunnen vinden. We zullen dus een methode moeten vinden, waarmee het verband voor het vermogen gevonden kan worden. De formulering van het probleem wordt dan als volgt.

Vind een methode om het verband tussen het vermogen en de snelheid te bepalen.

De tests waaruit data verkregen wordt, zijn uitroltests. Elke uitrolproef bestaat uit het versnellen tot een vastgestelde snelheid, waarna de motoren uitgezet worden. Vanaf dat moment, tijdstip 0, worden de metingen uitgevoerd. De snelheid wordt op drie manieren gemeten om eventuele meetfouten in kaart te kunnen brengen. Het Solar Team heeft geen apparatuur beschikbaar om de windsnelheid te meten. Deze waarden zijn dus onbekend. We zullen aannemen dat de windkracht (en windrichting) die gedurende de uitroltests op de auto werkt constant is.

Er zouden ook tests gedaan kunnen worden, waarbij de mock-up met constante snelheid rijdt. Echter is het verkrijgen van betrouwbare data uit deze tests erg moeilijk. Dit komt doordat de cruisecontrol er voor zorgt dat bij een snelheid onder de vastgestelde snelheid er versneld gaat worden, waarbij vermogen geleverd wordt. Op het moment dat de gereden snelheid weer boven de vastgestelde snelheid komt, zal er tijdelijk geen vermogen geleverd worden. Dit zorgt voor relatief grote schommelingen in het verbruikte vermogen. Ook zijn er extreme schommelingen in de stroom, aangezien de mock-up op snelheid wordt gehouden door steeds ófwel maximale stroom door de motor te sturen ófwel geen stroom door de motor te sturen. Hierdoor moet er gedurende een lange tijd op constante snelheid gereden worden voordat de gemiddelde stroom nauwkeurig bepaald kan worden. Dit is in werkelijkheid moeilijk te realiseren. Er zullen dus alleen uitroltests gedaan worden en geen tests met constante snelheid.

De enige gegevens die bekend zijn, zijn het gewicht van de mock-up (230 kg), de snelheid ten opzichte van de tijd in de uitroltest en de afgelegde afstand. Met deze gegevens zullen we een methode moeten vinden om het verband tussen het vermogen en de snelheid te bepalen.

Het verslag is als volgt opgedeeld. In hoofdstuk 1 introduceren we het probleem en proberen we inzicht te geven in de bezigheden van het Solar Team Twente. In het tweede hoofdstuk zullen

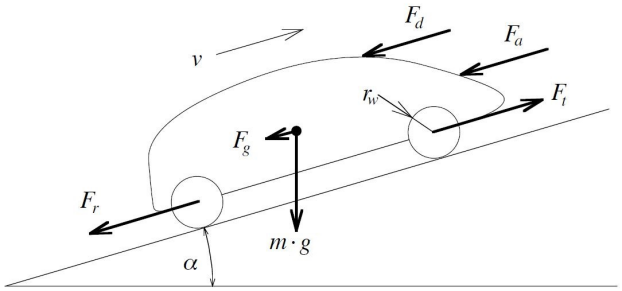
we de literatuur naslaan en onderzoeken wat er bekend is over het vermogen, de krachten en de weerstanden van een (zonne-)auto. In hoofdstuk 3 en 4 zullen we het probleem wiskundig formuleren met behulp van de gevonden voorkennis en hier een gepaste aanpak voor zoeken. Deze aanpak zal opgedeeld worden in drie delen en per deel zullen er een aantal methodes besproken worden in hoofstuk 5, 6 en 7. Vervolgens zullen deze methodes in hoofdstuk 8, 9 en 10 vergeleken en getest worden. Door middel van deze tests bepalen we welke methode het beste is. Met behulp van deze methode zullen we in hoofdstuk 11 een verband bepalen voor de tweede dataset van het Solar Team Twente. Er kan dan gekeken worden of dit verband overeenkomt met de verwachtingen en dus realistisch is. In hoofdstuk 12 zullen we een gevoeligheidsanalyse doen. We zullen kijken hoeveel data er nodig is, zodat de gekozen methode nauwkeurig wordt. Tot slot zullen er in hoofdstuk 13 conclusies worden getrokken en aanbevelingen worden gedaan.

2 Theoretisch verband

Er bestaan theoretische verbanden om met behulp van een aantal gegevens de krachten die op een auto werken te bepalen. Guzzella en Sciarretta beschrijven welke krachten dit zijn en welke formules hierbij horen [Guzzella and Sciarretta, 2005]. Het vermogen is te bepalen als deze krachten en de snelheid bekend zijn. Hiervoor geldt het volgende verband.

$$\vec{P} = \vec{F} \cdot \vec{v}. \tag{1}$$

Gegeven de snelheden, zullen we dus de kracht moeten bepalen. Om met constante snelheid te kunnen rijden, moeten de voorwaartse krachten van de auto gelijk zijn aan de terugwaartse krachten. Er moet dus net zoveel kracht geleverd worden als de weerstanden veroorzaken. De krachten die we hiervoor in acht nemen zijn de totale voorwaartse kracht F_t, de luchtweerstand F_a, de rolweerstand F_r, de kracht F_g veroorzaakt door rijden op niet-horizontale wegen en een storingskracht F_d, veroorzaakt door alle niet-gespecificeerde weerstanden.



Figuur 1: Krachten op auto

Als de voorwaartste en tegenwaartste kracht elkaar niet volledig opheffen ontstaat er een resulterende kracht F_{res}. Deze kracht zorgt voor een versnelling of een vertraging, afhankelijk van de richting van deze resulterende kracht. Voor de versnelling van een voertuig in de lengterichting geldt de volgende vergelijking.

$$F_{res} = m \cdot a. \tag{2}$$

Daar de residuele kracht F_{res} de som van krachten is, geldt:

$$m \cdot a = F_t(t) - (F_a(t) + F_r(t) + F_g(t) + F_d(t)), \tag{3}$$

waarbij $a = \frac{d}{dt}v$. In deze vergelijking is m de massa van het voertuig, a de versnelling van het voertuig en zijn de krachten F_t, F_a, F_r, F_g, F_d zoals hierboven beschreven. We bekijken de situatie waarin de auto al een snelheid heeft en vervolgens gaat vertragen. We kijken dus naar de dynamische rolweerstand en niet naar de statische rolweerstand, welke alleen geldt wanneer de auto vanuit stilstand gaat rijden. Daarnaast beschouwen we het wegoppervlak zonder hellingen ($F_g = 0$) en nemen we de storingsterm niet in acht ($F_d = 0$). Gezien het feit dat de data verkregen wordt uit uitroltests, nemen we ook de voorwaartse kracht niet mee ($F_t = 0$). Vergelijking (3) wordt gereduceerd tot:

$$m \cdot a = -(F_a(t) + F_r(t)). \tag{4}$$

De vergelijkingen voor de luchtweerstand en rolweerstand zien er als volgt uit.

$$F_a = \frac{1}{2} \cdot \rho \cdot A \cdot c_w \cdot v_{netto}^2. \tag{5}$$

$$F_r = (c_{r0} + c_{r\eta} \cdot v_{netto}^{\eta}) \cdot m \cdot g. \tag{6}$$

De snelheidsafhankelijke term in de rolweerstand wordt in de theorie meestal aangenomen als lineair ($\eta = 1$) of kwadratisch ($\eta = 2$) [Kulakowski, 1994]. In de praktijk is de significantie van deze term vaak zo laag dat formule (6) gereduceerd wordt tot:

$$F_r = c_r \cdot m \cdot g. \tag{7}$$

Als we de vergelijkingen (5) en (7) invullen in vergelijking (4) krijgen we de volgende vergelijking.

$$m \cdot a = -\frac{1}{2} \cdot \rho \cdot A \cdot c_w \cdot v_{netto}(t)^2 - m \cdot g \cdot c_r. \tag{8}$$

Hiervan zijn de variabelen weergeven in de onderstaande tabel.

Tabel 1: Grootheden formules

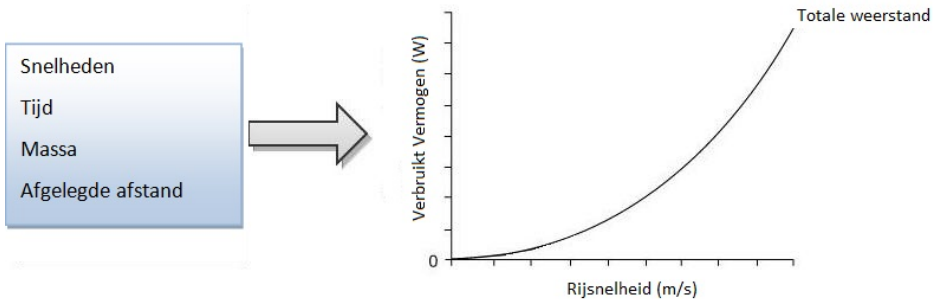
Symbool	Grootheid	Dimensie
P	Vermogen	W
F	Kracht	N
v_{netto}	Snelheid van het voertuig ten opzichte van het medium	m/s
F_a	Luchtweerstand	N
ρ	Dichtheid van de stof waarin de auto zich voortbeweegt, de luchtdichtheid	kg/m^3
A	Oppervlakte van het vooraanzicht van de auto	m^2
c_w	Weerstandscoëfficiënt	-
F_r	Rolweerstand	N
c_r	Wrijvingscoëfficiënt	-
p	Druk in de banden	bar
m	Massa van het voertuig	kg
g	Gravitatieconstante	m/s^2

In formules (5) en (7), voor de luchtweerstand en de rolweerstand, vinden we respectievelijk de weerstandscoëfficiënt c_w en de wrijvingscoëfficiënt c_r. De coëfficiënten c_w en c_r worden meestal als constante beschouwd. Door het Solar Team wordt echter de verwachting uitgesproken dat deze "coëfficiënten" in het geval van de zonne-auto niet constant zijn. Zij denken dat dit functies zijn, die afhankelijk zijn van verscheidene variabelen. Wij zullen gedurende ons onderzoek bekijken of we deze aanname kunnen bevestigen of verwerpen. Gezien het feit dat de snelheid de grootste variabele is, verwachten wij dat deze de grootste invloed heeft op de weerstands- en wrijvingsfunctie. Ook kunnen de windsnelheid, de druk in de banden en andere factoren van invloed zijn.

Om gebruik te kunnen maken van vergelijking (8) zullen we de weerstands- en wrijvingsfunctie moeten bepalen. Ook moet de rest van de gegevens door middel van metingen bekend zijn. Echter zijn niet de juiste middelen beschikbaar om alle gegevens te bepalen. Vandaar dat we de theorie niet direct toepassen. In plaats hiervan zullen we al de gegevens, op de snelheid na, zien als één grote coëfficiënt.

De metingen die we aangeleverd krijgen door het Solar Team zijn de snelheid, tijd, massa en afgelegde afstand van de mock-up. Met deze gegevens zal het verband tussen het vermogen en de snelheid bepaald moeten worden. We verwachten dat dit verband er uit gaat zien als in de onderstaande grafiek [van Dalen, 2010].

Figuur 2: Verwacht verband



3 Probleemstelling

We zullen niet rechtstreeks de natuurkundige formules gebruiken om het gezochte verband te bepalen, zoals in het vorige hoofdstuk beschreven. Hiervoor is al een reden genoemd. Een tweede reden is dat in deze theoretische modellen een aantal gegevens in beschouwing genomen moeten worden die in werkelijkheid lastig te isoleren zijn uit de aangeleverde data. Zo kunnen we de luchtdichtheid en de oppervlakte van het vooraanzicht van de auto niet exact meten. We kunnen hier hooguit een geschatte waarde voor invoeren. Als de werkelijke waarde echter afwijkt van onze schatting, zullen we niet weten waar de afwijking vandaan komt. De afwijking kan voortkomen uit een verkeerde meting van de luchtdichtheid, of een verkeerde bepaling van de oppervlakte van het vooraanzicht van de auto. Hierdoor wordt het onmogelijk om het gevonden verband te verbeteren. Ook de windsnelheid kan niet gemeten worden. Hierdoor moeten we de windsnelheid als constant aannemen en kunnen we deze verwerken in onze constantes. Hierdoor worden de formules eenvoudiger.

Bovendien zullen we alleen metingen krijgen van de mock-up. Voor deze mock-up gelden andere waarden dan voor de zonne-auto. Als we voor de mock-up de theoretische formules gebruiken, zal dit niet automatisch ook gelden voor de zonne-auto. Het gaat dus niet om het verband dat we vinden, maar om de methode die we gebruiken om het verband te vinden. Omdat de mogelijkheid bestaat dat de theorie niet gevolgd wordt, zullen we met behulp van een andere methode een functie moeten vinden die de data zo goed mogelijk beschrijft. In deze functie zullen de verschillende elementen als luchtdichtheid en oppervlakte van het vooraanzicht niet geëxpliciteerd worden. Kortom, de theorie zal maar deels gebruikt worden.

Vergelijking (8) zal ons op de goede weg helpen. Als eerste stap delen we aan de linkerkant van de vergelijking de massa weg. Dit leidt tot de volgende vergelijking.

$$a = -\frac{1}{m} \cdot \frac{1}{2} \cdot \rho \cdot A \cdot c_w \cdot v_{netto}^2 - g \cdot c_r. \tag{9}$$

Vervolgens reduceren we alle constantes uit vergelijking (9) tot twee constantes, β_0 en β_2. Ook schrijven we de versnelling (a) vanaf nu afhankelijk van de snelheid ($a(v)$). Dit geeft ons het volgende model.

$$a(v) = \beta_2 v_{netto}^2 + \beta_0, \tag{10}$$

waarin

$$\beta_2 = -\frac{\rho \cdot A \cdot c_w}{2m}, \quad \beta_0 = -g \cdot c_r. \tag{11}$$

Dit is een algemeen verband tussen de versnelling en de snelheid. In dit verband wordt de netto snelheid (v_{netto}) genoemd. Deze netto snelheid kunnen we in verschillende situaties bekijken. We lichten twee situaties, die we in het verslag zullen gebruiken, uitvoeriger toe. Dit zijn de situaties waarbij er geen wind staat en waarbij er tegenwind is. Als er geen wind staat wordt de netto snelheid (v_{netto}) gelijk aan de snelheid die gemeten wordt in de auto (v). Het model wordt dan als volgt.

$$a(v) = \beta_2 v^2 + \beta_0, \tag{12}$$

waarin

$$\beta_2 = -\frac{\rho \cdot A \cdot c_w}{2m}, \quad \beta_0 = -g \cdot c_r. \tag{13}$$

In het geval dat er tegenwind (v_{wind}) staat, wordt de netto snelheid gezien als $v_{netto} = v + v_{wind}$. Aangezien de windsnelheid niet accuraat gemeten kan worden, weten we alleen een gemiddelde windsnelheid. Daarom kunnen we de windsnelheid niet als variabele meenemen en beschouwen we deze als constant. We kunnen deze waarde verwerken in onze β's. Als we de definitie van de netto snelheid invullen geeft dit de volgende afleiding naar een model met tegenwind.

$$\begin{aligned} a(v) &= \beta_2 (v + v_w)^2 + \beta_0 \\ &= \beta_2 (v^2 + 2v_w v + v_w^2) + \beta_0 \\ &= \beta_2 v^2 + 2\beta_2 v_w v(t) + \beta_2 v_w^2 + \beta_0. \end{aligned}$$

Duidelijker opgeschreven geeft dit ons:

$$a(v) = \tilde{\beta}_2 v^2 + \tilde{\beta}_1 v + \tilde{\beta}_0, \tag{14}$$

waarin

$$\tilde{\beta}_2 = \beta_2, \quad \tilde{\beta}_1 = \beta_2 \cdot v_w, \quad \tilde{\beta}_0 = \beta_2 v_w^2 + \beta_0. \tag{15}$$

We hebben nu drie verschillende modellen gezien. Een algemeen model, een model zonder wind en een model met tegenwind. Welke machten van de snelheid meegenomen worden verschilt per model. Voor ons onderzoek is dit niet handig en willen we het model veralgemeniseren. Bovendien heeft het Solar Team het vermoeden dat de c_w en de c_r ook afhankelijk zijn van de snelheid. In dat geval kunnen we niet zeggen dat we deze "coëfficiënten" meenemen in de constante. Ook weten we dan de maximale graad van ons polynoom niet. Om dit probleem op te lossen generaliseren we het probleem nog verder. We gaan op zoek naar een methode om een polynoom te vinden die het verband tussen de versnelling en de snelheid zo goed mogelijk omschrijft. Het polynoom komt er als volgt uit te zien:

$$a(v) = \sum_{j=1}^{k} \beta_j v^j. \tag{16}$$

4 Probleemaanpak

Het probleem geformuleerd in hoofdstuk 1 kunnen we opdelen in meerdere delen. Allereerst moeten we overwegen of het nodig is de data te behandelen om deze bruikbaar te maken. De data die we gekregen hebben is namelijk niet glad. Dit is te zien in de volgende afbeelding. Als we kijken

Figuur 3: Voorbeeld data

	A	B	
1	Time	MC_avg_Vel	
2	38,6	1,3235	-
3	38,8	1,2995	+
4	39	1,322	-
5	39,2	1,286	=
6	39,4	1,286	+
7	39,6	1,3145	-
8	39,8	1,304	-
9	40	1,2055	

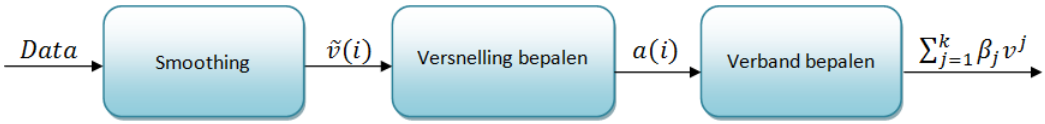
naar de kolom MC_avg_Vel, zien we de gemeten snelheid van de mock-up. Zoals te zien is daalt deze snelheid soms, stijgt deze soms en blijft deze soms gelijk. De versnelling die hier uit volgt zal hierdoor soms negatief, soms positief en soms nul zijn. Dit heeft tot gevolg dat als we hier direct een verband door zoeken, dit verband erg veel "hobbels" bevat. Dit zal een onnauwkeurig model opleveren. We kunnen de data behandelen om het verloop van de snelheid gladder maken. Hier moeten we een methode voor kiezen. Vervolgens moet de versnelling bepaald worden vanuit de snelheden en de tijd, die we als data verkregen hebben. Ook hier zijn verschillende manieren voor. Tot slot moet met de gevonden versnellingen een verband bepaald worden waarin de versnelling tegen de snelheid uitgezet wordt.

In hoofdstuk 3 is geconcludeerd dat dit verband een polynoom is. Voor dit polynoom moeten de coëfficiënten en de maximale graad bepaald worden. Dit leidt tot de volgende opsomming van problemen.

1. Gegeven $v(t_1)$,$v(t_2)$,...,$v(t_n)$, hoe bepalen we $\tilde{v}(t_1)$,$\tilde{v}(t_2)$,...,$\tilde{v}(t_n)$, waarbij $\tilde{v}(t_i)$ een glad verloop heeft.
2. Gegeven $\tilde{v}(t_1)$,$\tilde{v}(t_2)$,...,$\tilde{v}(t_n)$, hoe bepalen we $a(t_1)$,$a(t_2)$,...,$a(t_n)$.
3. Gegeven $v(t_1)$,$v(t_2)$,...,$v(t_n)$ en $a(t_1)$,$a(t_2)$,...,$a(t_n)$, hoe bepalen we

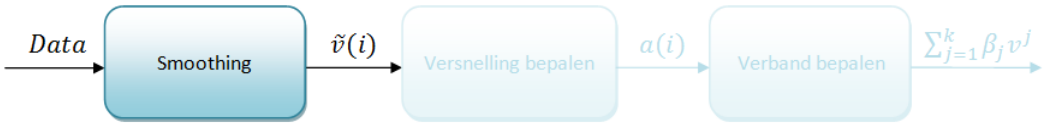
$$a(v) = \sum_{j=1}^{k} \beta_j v^j.$$

In het vervolg zullen we voor het gemak $v(t_i)$, $\tilde{v}(t_i)$ en $a(t_i)$ reduceren tot respectievelijk $v(i)$, $\tilde{v}(i)$ en $a(i)$. De probleemaanpak staat ook in de onderstaande figuur weergeven.



Vanaf nu zullen we bovenstaande figuur gebruiken om aan te geven met welk deelprobleem we bezig zijn. Voor elk deelprobleem zullen we verschillende methodes bekijken. Deze methodes zullen we testen, waarna er voor elke stap één methode gekozen wordt. Deze drie methodes vormen samen onze uiteindelijke methode om het verband te bepalen tussen de snelheid en het benodigd vermogen om deze snelheid te rijden.

5 Smoothing



In dit hoofdstuk bespreken we methodes om het eerste deelprobleem op te lossen. In hoofdstuk 4 is verteld dat we gaan kijken of we de data moeten behandelen om het verloop gladder te maken. Het gladder maken van data wordt smoothen genoemd. Het smoothen kan belangrijk zijn om het uiteindelijk gewenste polynoom (16) zo nauwkeurig en bruikbaar mogelijk te maken. Voor het smoothen van data zijn vele mogelijkheden. Enkele bekende en veelgebruikte methodes staan in de onderstaande tabel toegelicht.

Tabel 2: Smoothingmethodes

Methode	Specificatie	Toepasbaarheid
Nearest-neighbor smoothing	Voor elk datapunt y_i kijk naar q dichtstbijzijnde punten en middel de waardes voor deze punten.	Deze methode is erg omslachtig en de nieuwe datapunten worden met het toenemen van q al snel onnauwkeurig. Toch zullen we deze methode als eerste gaan gebruiken (met kleine q) om de grootste ruis in ieder geval al weg te filteren.
Savitzky-Golay smoothing	Voor elk datapunt y_i, kijk naar q dichtstbijzijnde punten en neem van deze punten een gewogen gemiddelde. Het Savitzky-Golay algoritme geeft de wegingsfactoren (coëfficiënten) voor de verschillende q_i. Ook is het mogelijk om naar de afgeleide van de datapunten te kijken.	Door de vele mogelijkheden die de Savitzky-Golay methode geeft, wordt deze in veel gevallen beschouwd als de beste methode voor het smoothen van een ruwe dataset. Wij zullen deze methode dan ook intensief behandelen.
Kernel-average smoother	Voor elk datapunt y_i, middel binnen een straal r rondom het punt y_i de waardes van de liggende datapunten.	Het resultaat van de nieuwe data is al beter dan bij de Nearest-neighbor smoothing, omdat een datapunt ver weg van het punt y_i nooit meer wordt gebruikt voor het smoothen van y_i. Dit is in theorie wel mogelijk bij Nearest-neightbor smoothen. De nieuwe data heeft echter nog steeds te veel afwijkingen.
Locale lineaire regressie	Voor elk datapunt y_i worden de datapunten op de randen van een cirkel met straal r middels een rechte lijn met elkaar verbonden met als middelpunt het punt y_i.	De uiteindelijke functie lijkt steeds meer op hetgeen waar we naar op zoek zijn, maar doordat je lokaal lineaire regressie toepast, verlies je min of meer data. Dit is zonde, omdat met de juiste computerprogramma's het mogelijk is om lineaire regressie toe te passen op alle datapunten.

Methode	Specificatie	Toepasbaarheid
Ramer-Douglas-Peucker algorithm/End-point fit algoritme	Voor elke set datapunten elimineert dit algoritme punten om uit een grote set datapunten een kleinere set bruikbare datapunten te verkrijgen.	Wat het algoritme als bruikbaar beschouwd, zijn de datapunten die relatief de grootste afwijkingen hebben van de andere datapunten in een interval. Omdat wij te maken hebben met meetfouten en ruis, is het niet handig dit algoritme te gebruiken, aangezien de punten die het algoritme uiteindelijk overhoudt niet betrouwbaar zijn.

Er bestaan zoals te zien veel manieren om te smoothen. Enkele manieren lijken op het eerste gezicht veel op elkaar, anderen wijken hier weer totaal van af. De keuze van de smoothingmethode is dus belangrijk. Hieronder staan de twee smoothingmethodes die wij gekozen hebben duidelijk toegelicht.

5.1 Nearest-neighbor smoothing

De eerste smoothingmethode die we gaan gebruiken is het Nearest-neighbor smoothen. Zoals in de tabel staat beschreven kies je bij deze methode voor elk datapunt y_i de q dichtsbijzijnde punten en gebruik je het gemiddelde van deze datapunten inclusief y_i zelf, om het punt y_i te smoothen. Wat de methode niet vereist, maar wat bij ons wel het geval is, is dat we aan weerszijde van het punt y_i naar hetzelfde aantal punten m ($=\frac{q-1}{2}$) punten kijken. Dit is omdat onze datapunten met een constant tijdsinterval verkregen zijn. Het smoothen van een punt is wiskundig te schrijven als

$$\tilde{y}_i = \frac{y_{i-m} + y_{i-(m-1)} + .. + y_i + y_{i+1} + ... + y_{i+m}}{2m+1}. \tag{17}$$

Met het toenemen van m (en dus ook q) worden de gesmoothde punten al snel onnauwkeurig. Wel gaan we kijken of het gebruiken van kleine m nuttig is voor het smoothen van onze data.

5.2 Savitzky-Golay methode

De tweede smoothingmethode die we gebruiken is de Savitzky-Golay methode. De Savitzky-Golay methode voor het smoothen van een dataset is een goede en veelgebruikte manier om uit een ruwe dataset een nieuwe dataset te maken, waarbij de meetfouten aanzienlijk verminderen.
De Savitzky-Golay methode lijkt in zekere zin op het Nearest-neighbor smoothen. Bij het Nearest-neighbor smoothen nemen we het gemiddelde van $q = 2m + 1$ punten om een ruw datapunt te smoothen, met m een te kiezen aantal datapunten links en rechts van het te smoothen punt y_i. In principe benader je hier de q datapunten middels een rechte lijn. Bij de Savitzky-Golay methode kun je een datapunt smoothen door q ruwe datapunten te benaderen met een polynoom van willekeurige orde. Voor het punt y_i wordt een gewogen gemiddelde genomen, waarbij deze wegingsfactoren (coëfficiënten) berekend worden door het Savitzky-Golay algoritme. Bij een gekozen orde polynoom om de q datapunten rondom het punt y_i te benaderen, vindt het Savitzky-Golay algoritme de coëfficiënten voor deze q punten om het polynoom te beschrijven. Dit gebeurt door gebruik te maken van de kleinste kwadraten methode, die later in het verslag aan bod zal komen. De coëfficiënten noemen we in het vervolg ${p_s}^{(0)}$, waarbij $p_0^{(0)}$ de coëfficiënt is voor het te smoothen punt y_i, $p_{-1}^{(0)}$ de coëfficiënt is voor het punt y_{i-1}, $p_1^{(0)}$ de coëfficiënt is voor het punt y_{i+1} etc. Het gesmoothde punt $\tilde{y}_i$ is hierdoor te schrijven als

$$\tilde{y}_i = \sum_{s=-m}^{m} (p_s^{(0)} \cdot y_{i+s}).$$

Savitzky en Golay hebben ontdekt dat naast de mogelijkheid om een punt y_i te smoothen op de hierboven besproken manier, het mogelijk is om te kijken naar de d^e orde afgeleide van het punt y_i. Het smoothen van de d^e orde afgeleide van het punt y_i gebeurt op dezelfde wijze als bij de 0^e orde:

$$\frac{d^d \tilde{y_i}}{dx^d} = \sum_{s=-m}^{m} (p_s^{(d)} \cdot y_{i+s}). \tag{18}$$

Hierin is $p_s^{(d)}$ de coëfficiënt voor het punt y_{i+s} om de d^e orde afgeleide van het punt y_i te smoothen. Doordat het mogelijk is om naar de willekeurige orde afgeleide van een ruw datapunt y_i te kijken, kun je van het gesmoothde punt $\tilde{y_i}$ een Taylorreeks opstellen. Deze taylorreeks voor een datapunt y_i komt er uit te zien als

$$Y(x) = c_0 + \frac{c_1 \cdot x}{1!} + \frac{c_2 \cdot x^2}{2!} + ... + \frac{c_j \cdot x^j}{j!} = \sum_{d=0}^{j} \frac{c_d \cdot x^d}{d!}. \tag{19}$$

Hierin is c_0 gelijk aan $\tilde{y_i}$, c_1 gelijk aan $\frac{d^1 \tilde{y_i}}{dx^1}$ etc. Je kunt dus zelf bepalen hoeveel datapunten (keuze van m) je gebruikt en hoe groot de orde van het polynoom is (beïnvloed je coëfficiënten $p_s^{(d)}$) dat je gebruikt om één datapunt y_i te smoothen.

Uit wiskundige berekeningen zijn de algemene formules afgeleid en bijbehorende tabellen gemaakt voor de coëfficiënten $p_s^{(0)}$ en $p_s^{(d)}$ bij veelgebruikte waardes van m en d. We zullen verderop in het verslag onze keuze voor m maken en de orde van het polynoom waarmee we de datapunten gaan smoothen bespreken, motiveren en de bijbehorende formules en tabellen geven.

5.2.1 Voor- en nadelen

Een belangrijk voordeel van de Savitzky-Golay methode is het resultaat dat de methode geeft. De data die verkregen wordt na het toepassen van deze methode is bruikbaarder om een polynoom te bepalen, omdat er zoveel ruis wordt weggenomen. Hierbij is de orde van het polynoom waarmee je gaat smoothen van groot belang voor je resultaat. Een lage orde pakt de meetfouten het hardst aan, maar het kan er ook voor zorgen dat "uitschieters" die wel degelijk van invloed zijn op je latere model worden weggefilterd. Zoals genoemd kun je ook eenvoudig naar de afgeleide van de datapunten kijken. Dit is voor ons een groot voordeel. Wij kunnen immers uit de data, die de snelheid tegen de tijd uitzet, naar de afgeleide van de snelheid, de versnelling (a) kijken.

Uiteraard kleven aan deze methode ook enkele kleine nadelen. Zo zijn de buitenste m datapunten niet op dezelfde manier te smoothen als de rest, omdat deze datapunten te weinig datapunten links, dan wel rechts van zich hebben. Dit is op te lossen door simpelweg de buitenste m datapunten buiten beschouwing te laten of de m naar 0 te laten convergeren, naarmate het datapunt zich dichterbij de buitenkant bevindt. De laatste optie geeft echter niet voor elk datapunt een even nauwkeurige smoothing, waardoor er, bij voldoende dataputen, vaak voor wordt gekozen om de buitenste punten m buiten beschouwing te laten.

5.2.2 Toepassing

In deze sectie bespreken we hoe de Savitzky-Golay methode toegepast gaat worden met onze data. Ook motiveren we onze keuzes voor m en de orde van het polynoom dat we door de $2m+1$ punten willen fitten.

Ten eerste hebben we het over de orde van het polynoom door de q datapunten. Om te beginnen benaderen we de q datapunten met een rechte lijn, omdat we verwachten dat hierdoor de meetfouten het best weggefilterd worden en de gesmoothde data het meest realistisch is. Bovendien willen we kijken of de nieuwe data overeenkomt met de nieuwe data bij Nearest-neighbor smoothen.

Ook gaan we kijken naar een smoothing met een polynoom van orde 3. Verder zal er gekeken worden naar de 0^e en 1^e orde afgeleide. Hogere orde afgeleides zijn voor ons niet van belang. We hoeven immers alleen tot de 1^e afgeleide, in ons geval de versnelling, te kijken. In de tabel hieronder staan de formules om de coëfficiënten $p_s^{(d)}$ te vinden voor 0^e en 1^e afgeleide.

Tabel 3: Formules voor coëfficiënten Savitzky-Golay methode.

Orde afgeleide	Orde polynoom	Formule
0	1	$p_s^{(0)} = \frac{1}{2m+1}$
1	1	$p_s^{(1)} = \frac{3s}{(2m+1)(m+1)(m)}$
0	3	$p_s^{(0)} = \frac{3(3m^2+3m-1-5s^2)}{(2m+3)(2m+1)(2m-1)}$
1	3	$p_s^{(1)} = \frac{5[5(3m^4+6m^3-3m+1)s-7(3m^2+3m-1)s^3]}{(2m+3)(2m+1)(2m-1)(m+2)(m+1)(m)(m-1)}$

Als we deze formules in gaan vullen voor waardes van m die we gaan gebruiken krijgen we de volgende tabellen.

Tabel 4: Tabellen voor orde 1.

(a) Tabel voor $p_s^{(0)}$ bij orde 1.

2m+1	h	$p_0^{(0)}$	$p_1^{(0)}$	$p_2^{(0)}$	$p_3^{(0)}$	$p_4^{(0)}$
5	5	1	1	1	0	0
7	7	1	1	1	1	0
9	9	1	1	1	1	1

(b) Tabel voor $p_s^{(1)}$ bij orde 1.

2m+1	h	$p_0^{(1)}$	$p_1^{(1)}$	$p_2^{(1)}$	$p_3^{(1)}$	$p_4^{(1)}$
5	10	0	1	2	0	0
7	28	0	1	2	3	0
9	60	0	1	2	3	4

Tabel 5: Tabellen voor orde 3.

(a) Tabel voor $p_s^{(0)}$ bij orde 3.

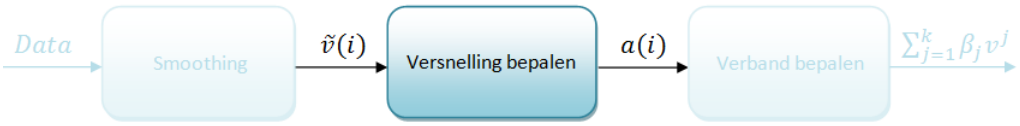
2m+1	h	$p_0^{(0)}$	$p_1^{(0)}$	$p_2^{(0)}$	$p_3^{(0)}$	$p_4^{(0)}$
5	35	17	12	−3	0	0
7	21	7	6	3	−2	0
9	231	59	54	39	14	−21

(b) Tabel voor $p_s^{(1)}$ bij orde 3.

2m+1	h	$p_0^{(1)}$	$p_1^{(1)}$	$p_2^{(1)}$	$p_3^{(1)}$	$p_4^{(1)}$
5	504	0	326	414	0	0
7	2520	0	587	726	−33	0
9	7128	0	756	1158	852	516

Hierbij is $p_s^{(0)}$ gelijk aan $p_{-s}^{(0)}$, $p_s^{(1)}$ gelijk aan $-p_{-s}^{(1)}$ en is h het product van de noemer zoals in tabel 3 staat beschreven.

6 Versnelling bepalen



Uit de proeven van het Solar Team krijgen we alleen gegevens over de snelheid ten opzichte van de tijd. Het doel is echter om de versnelling tegen de snelheid uit te zetten in een te bepalen verband. Omdat hier gegevens over de versnelling voor nodig zijn, moeten we een methode bedenken om deze versnelling te bepalen. Hier zijn veel verschillende manieren voor. We hebben er vier ontwikkeld en die staan hieronder besproken.

6.1 Numerieke methode A

Uit de natuurkunde is bekend dat de versnelling gezien kan worden als de afgeleide van de snelheid. Aangezien er geen continu verband bekend is voor de snelheid, kan deze niet afgeleid worden naar de tijd om de versnelling te verkrijgen. Wel kan de versnelling numeriek bepaald worden op ongeveer dezelfde manier:

$$a(i) = \frac{\Delta v}{\Delta t} = \frac{v(i) - v(i-1)}{\Delta t}, \tag{20}$$

waarbij Δt in onze data gelijk is aan 0.2. Zoals in hoofdstuk 4 staat beschreven, kan het nemen van een te klein interval in sommige gevallen tot negatieve versnellingen leiden. Dit is een gevolg van onzuivere metingen. Om met een klein interval wel goede versnellingen te krijgen, hebben we de datapunten $v(i)$ gesmoothd met een simpele Nearest-neighbor smoothing met 5 punten. De nieuwe punten zijn:

$$\tilde{v}(i) = \frac{v(i+2) + v(i+1) + v(i) + v(i-1) + v(i-2)}{5}. \tag{21}$$

Vervolgens wordt de versnelling berekend uit de nieuwe punten $\tilde{v}(i)$ met behulp van vergelijking (20) waardoor er veel van de ruis verdwenen zal zijn.

6.2 Numerieke methode B

Zoals bij de voorgaande methode beschreven wordt, is de versnelling onderhevig aan de ruis op de snelheid. Dit komt doordat er een klein verschil tussen opeenvolgende snelheden bekeken wordt. In de eerste methode wordt dit opgelost door de snelheid te smoothen en daarna pas de versnelling te berekenen. Als tweede methode wordt gekeken of deze stap overgeslagen kan worden door gelijk een groter verschil tussen twee snelheden te bekijken. Hiervoor kiezen we dan ook weer een interval van vijf punten. De versnelling wordt dan als volgt berekend.

$$a(i) = \frac{v(i+2) - v(i-2)}{5 \cdot 0,2} = v(i+2) - v(i-2). \tag{22}$$

6.3 Analytische methode

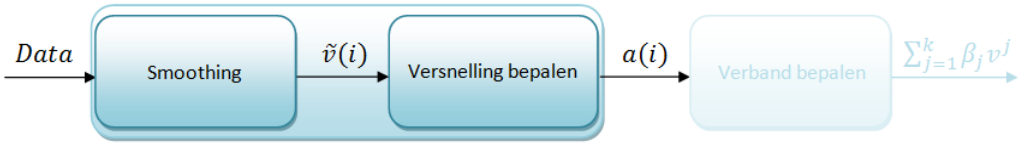
In numerieke methode A wordt verondersteld dat de versnelling gezien kan worden als de afgeleide van de snelheid naar de tijd. Het probleem hierbij is dat er geen functie bekend is voor de snelheid die afgeleid kan worden. Om deze afgeleide toch te kunnen gebruiken, kan ook eerst een verband tussen v en t bepaald worden, welke vervolgens afgeleid kan worden om een verband te vinden voor de versnelling. Voor de snelheid zal met behulp van lineaire regressie een verband gezocht worden, waarna de versnelling ten opzichte van de tijd als volgt berekend wordt.

$$a(t) = \frac{d}{dt} v(t). \tag{23}$$

Deze formule geeft een verband voor a naar t, terwijl we a tegen v uit willen zetten. Echter, door $a(t)$ en $v(t)$ in te vullen voor verschillende t krijgen we datapunten voor a en v die we tegen elkaar uit kunnen zetten.

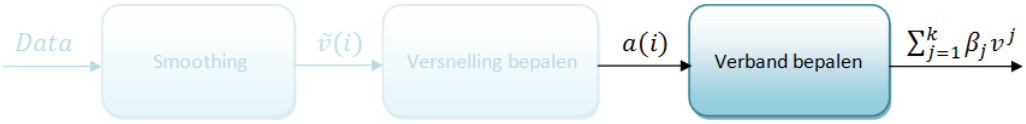
6.4 Savitzky-Golay methode

Als vierde methode zal gekeken worden naar het gebruik van een Savitzky-Golay filter. Dit filter, beschreven in paragraaf 5.2, kan direct de afgeleide bepalen rond een punt. De Savitzky-Golay methode kan als een methode om te smoothen gezien worden, maar ook als een methode om de versnelling te bepalen. Dit is hieronder schematisch weergeven.



Het Savitzky-Golay filter kan voor elk meetpunt $v(i)$ een punt $a(i)$ bepalen. Om deze punten te bepalen wordt een Matlab-programma gebruikt. Dit programma wordt besproken in Appendix, deel 2.

7 Verband bepalen



Zodra er een goede methode gekozen is om datapunten voor de versnelling te bepalen, moeten we ook een methode kiezen om voor deze datapunten een verband te vinden. Een optie hiervoor is interpolatie. Als de data betrouwbaar is, kunnen opeenvolgende punten met elkaar verbonden worden. Hierdoor vind je een exact verband door de bestaande datapunten, maar kun je niets voorspellen voor punten buiten het model. Een andere optie is lineaire regressie. Dit is een proces waarbij geprobeerd wordt een bepaald verband zo goed mogelijk te passen op de datapunten. Hierdoor wordt uiteindelijk een verband verkregen met een zo klein mogelijke afwijking tot de datapunten. Met dit verband is het ook mogelijk om punten buiten het interval te voorspellen. Deze methode zal hier verder uitgediept worden.

7.1 Soort verband

Volgens de theorie uit hoofdstuk 3 is het verband tussen de versnelling en de snelheid polynomiaal. Van welke graad het polynoom moet worden is aan ons om te bepalen. Een polynoom van graad 1 ziet er uit als $y = ax + b$, , een polynoom van graad 2 als $y = ax^2 + bx + c$, polynoom van graad 3 als $y = ax^3 + bx^2 + cx + d$ etc. Een polynoom van graad 1 is een lijn die precies past op 2 (data)punten, een polynoom van graad 2 is een curve die precies past op 3 (data)punten. In theorie kan er dus een polynoom van graad $n - 1$ opgesteld worden die precies door alle n datapunten heen loopt. Dit is echter geen realistisch verband. De vraag is dus van welke graad het polynoom moet zijn om de fout zo klein mogelijk te maken, maar tegelijkertijd het verband realistisch te houden.

In de literatuur worden hier uitspraken over gedaan. Bij technische en bedrijfskundige toepassingen worden er vaak polynomen gevonden met hoogstens een tweede orde term [Administration, 2012]. Als we dit toepassen op de coëfficiënten c_w en c_r en dit doorrekenen in het verband voor de versnelling ten opzichte van de snelheid, zal er hooguit een vierde orde polynoom voor gelden. De maximale graad van polynoom (16) zal dus twee, drie of vier zijn. In [Faraway, 2002] lezen we dat het onwaarschijnlijk is dat wanneer de onafhankelijke variabele x geen deel uit maakt van het polynoom, x^2 wel deel uit maakt van het polynoom. Stel dat we het polynoom $y = \alpha_0 + \alpha_2 x^2$ vinden. Dan betekent dit, dat bij een verschuiving $x \to x + a$ het model $y = \alpha_0 + \alpha_2 a^2 + 2\alpha_2 ax + \alpha_2 x^2$ wordt. Dit houdt in dat in het verschoven model een term x is verschenen die niet in het oorspronkelijke model zit, terwijl verschuivingen geen significante verschillen op zouden moeten leveren. In ons geval houdt dit in dat het onwaarschijnlijk is bijvoorbeeld een model afhankelijk van v^4 te vinden, maar niet van v^3.

7.2 Lineaire regressie

Het analyseren van data waarin mogelijk een specifieke samenhang tussen verschillende variabelen bestaat, wordt regressieanalyse genoemd. In ons geval houdt dit in dat wordt onderzocht of èn hoe de versnelling afhangt van verschillende machten van de snelheid. In het algemeen wordt de afhankelijke variabele Y genoemd en de onafhankelijke variabelen x. Voor de metingen is dan een model op te stellen:

$$Y = f(x) + \epsilon, \tag{24}$$

waar Y de gemeten waarde is, $f(x)$ de gezochte functie die het verband omschrijft tussen de afhankelijke en onafhanke variabelen en ϵ de storing die is opgetreden bij het meten van Y. In ons geval wordt $f(x)$ een polynoom. Met behulp van een lineaire regressieanalyse worden de relevante

machten van de snelheid in dit polynoom gevonden en de bijbehorende coëfficiënten. Met de hand is een lineaire regressieanalyse lastig uit te voeren. Er bestaan echter computerprogramma's die het werk erg makkelijk maken. Om de lineaire regressieanalyse uit te voeren gebruiken wij het programma SPSS. Hoe dit in zijn werk gaat is beschreven in Appendix 1.

7.3 Aannames

Er is besloten regressieanalyse te gebruiken voor het bepalen van relevante variabelen. Om dit te gebruiken moeten er een aantal aannames gedaan worden met betrekking tot de data [Siero et al., 2009]. Als we er vanuit gaan dat de data aan deze aannames voldoet, kan regressieanalyse toegepast worden. De aannames zijn als volgt.

1. De data bestaat uit onafhankelijke waarnemingen.
2. Er is een lineair verband tussen de afhankelijke en onafhankelijke variabelen. In ons geval de versnelling a en de machten van de snelheid v,...,v^k.
3. De variantie van de residuen is gelijk voor alle datapunten.
4. De residuen zijn normaal verdeeld.

We nemen aan dat de data bestaat uit onafhankelijke waarnemingen. Dit wil zeggen dat er geen samenhang bestaat tussen de verschillende datasets. Deze aanname betekent dat de correlatie tussen de residuen uit de regressieanalyse gelijk moet zijn aan nul. Mocht de data niet onafhankelijk zijn, dan zijn geschatte standaardfouten onzuiver en dus kleiner dan in werkelijkheid. Hierdoor kun je concluderen dat er een relatie tussen de waarnemingen bestaat, terwijl dit in werkelijkheid niet het geval is. Verder zal er aangenomen worden dat het verband beschreven kan worden door een polynoom. Met behulp van de residuen, het verschil tussen de gemeten waardes en de geschatte waardes, kunnen we controleren of deze aanname klopt. Als derde wordt aangenomen dat de variantie van de residuen gelijk is voor alle datapunten, ook wel homoscedasiticiteit genoemd. Deze aanname kan gecontroleerd worden door naar de residuenplots te kijken. Is er sprake van homoscedasiticiteit dan zal in deze plot de spreiding van de punten namelijk overal even groot zijn. Als laatste zal aangenomen worden dat de residuen normaal verdeeld zijn, zodat toetsen uitgevoerd kunnen worden. Mocht deze aanname niet opgaan, kan het namelijk zijn dat de betrouwbaarheidsintervallen niet goed zijn en dat er verkeerde conclusies worden getrokken. Als blijkt dat de residuen niet normaal verdeeld zijn, zullen we gebruik moeten gaan maken van de Centrale-Limietstelling. Deze stelling zegt dat wanneer een dataset voldoende groot is, er bij benadering nog steeds een normaal verdeelde verzameling ontstaat.

7.4 Het bepalen van de relevante variabelen

In de voorgaande paragraaf 7.3 zijn een aantal aannames gedaan. Ervan uitgaande dat deze aannames gelden voor de data waar mee gewerkt wordt, kan een regressiemodel opgesteld worden voor ons probleem. Er kan gekozen worden uit enkelvoudige lineaire regressie en meervoudige lineaire regressie. De eenvoudigste van beiden is de enkelvoudige lineaire regressie. Hierbij hoort het volgende model.

$$Y = \beta_0 + \beta_1 x + \epsilon. \tag{25}$$

Hierin is Y de afhankelijke variabele, x de onafhankelijke variabele, $\beta = (\beta_0,\beta_1)$ de coëfficiënten die je wil bepalen en ϵ de storingsterm. In paragraaf 7.3 is gezien dat deze storingen als normaal verdeeld worden aangemomen met verwachting 0 en standaardafwijking σ. In dit model wordt de afhankelijke variabele voorspeld door slechts één onafhankelijke variabele. Als dit model toegepast zou worden op het theoretische verband tussen de versnelling en de snelheid, zou Y de versnelling zijn en moet er voor x een keuze gemaakt worden. We kunnen v of v^2 invoeren, maar niet beide. Als er een kwadraat wordt genomen, valt het model, zoals besproken is in paragraaf 7.3, nog steeds

onder lineaire regressie. Zo lang het model namelijk uitgedrukt kan worden als lineaire combinatie van de variabelen valt het model onder lineaire regressie [Seber and Lee, 2003]. In de theorie komen echter al twee variabelen voor in het model. Bovendien willen we onderzoeken of er ook nog hogere machten van de snelheid in het model voorkomen. Enkelvoudige lineaire regressie kan daarvoor niet gebruikt worden. Hiervoor kan het meervoudige lineaire regressiemodel wél gebruikt worden. Dit model ziet er als volgt uit.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon. \tag{26}$$

In dit model is Y de afhankelijke variabele, zijn $x = (x_1, ..., x_k)$ de onafhankelijke variabelen, $\beta = (\beta_0,...,\beta_k)$ de coëfficiënten die je wil bepalen en ϵ de storingsterm. Voor Y kan weer de versnelling genomen worden en voor elke x_i kan een andere macht van de snelheid ingevuld worden. Met behulp van het programma SPSS kan onderzocht worden welke machten van de snelheid relevant zijn voor het model en welke coëfficienten hierbij horen. De literatuur die we hier over gevonden hebben in paragraaf 7.1 moeten we natuurlijk wel meenemen bij het evalueren van een gevonden verband.

7.5 Schatters voor variabelen

Om een verband tussen de afhankelijke en onafhankelijke variabelen te vinden wordt een schatter gekozen. Een schatter T is in dit geval een steekproeffunctie die een schatting geeft van de parameters, θ genoemd. In ons geval zijn dit de coëfficiënten in het polynoom, $\beta = (\beta_0,...,\beta_k)$ genoemd. De gekozen schatter noemen we B die de schattingen $\hat{\beta} = (\hat{\beta_0}, ..., \hat{\beta_k})$ geeft. Er zijn verschillende schatters waaruit gekozen kan worden. De schatters verschillen qua moeilijkheid, toepasbaarheid, zuiverheid en consistentie van elkaar. Goede schatters zijn zuiver en consistent. Het consistent zijn van een schatter is een drempelvoorwaarde om slecht gekozen schatters te elimineren. Per definitie betekent het consistent zijn van een schatter dat voor een schatting T_n van θ, $\lim_{n\to\infty} P(|T_n - \theta| > \epsilon) = 0 \; \forall \epsilon > 0$ en $\forall \theta \in \Theta$. In ons geval wordt dit $\lim_{n\to\infty} P(\left|\hat{\beta}_n - \beta\right| > \epsilon) = 0$ waarbij $\hat{\beta}_n$ een schatting is voor β op basis van n datapunten. Kortom, consistentie betekent dat hoe groter de steekproef is, hoe kleiner de fout in de schatting wordt. Intuïtief is dit ook logisch. Het kan namelijk niet zo zijn dat wanneer er meer data beschikbaar is om informatie uit te halen, de schatting voor de parameters verder af komt te liggen van de daadwerkelijke waarde.

Daarnaast is er de voorwaarde dat de schatter zuiver moet zijn. Het zuiver zijn van een schatter betekent dat de schatter gemiddeld genomen (over alle steekproeven) gelijk is aan de parameters. Net gedefinieerd betekent dit dat een schatter T zuiver heet als $E[T] = \theta$ voor alle $\theta \in \Theta$. Dit ziet er in ons geval uit als $E[B] = \beta$ voor alle β. Door deze definitie wordt bijvoorbeeld uitgesloten dat de schatter constant is, gezien dan de verwachting van de schatter alleen gelijk is aan een enkele parameter en niet geldt voor alle parameters [Albers, 2010]. Hieronder zullen de meest gebruikte schatters kort besproken worden, waarna een uitendelijke keuze gemotiveerd wordt.

7.5.1 De kleinste-kwadratenmethode

In [Hansen et al., 2012] wordt een uitgebreide omschrijving van de kleinste-kwadratenmethode gegeven. Deze methode in het simpelste geval noemt men lineaire kleinste-kwadratenmethode. Deze methode kijkt naar de som van de kwadratische afwijking op de y-as van de datapunten met een te bepalen lijn $y = a + bx$. Gezocht wordt nu naar de schatter voor a en b die zorgt dat de som van de residuen in het kwadraat geminimaliseerd wordt. Deze schatter is zuiver en consistent. Verder is deze methode altijd nauwkeurig te gebruiken wanneer de residuen eindige variantie hebben en homoscedastisch zijn.

7.5.2 Gegeneraliseerde kleinste-kwadratenmethode

In [Kariya and Kurata, 2004] staat dat deze uitbreiding van de lineaire kleinste-kwadratenmethode als voordeel heeft dat deze ook nauwkeurig te gebruiken is wanneer de variantie van de geschatte parameters niet gelijk is (heteroscedasticiteit) of wanneer er enige afhankelijkheid bestaat tussen de residuen.

7.5.3 Percentuele kleinste kwadraten

In deze methode wordt weer gekeken naar het minimaliseren van de som van de gekwadrateerde residuen. Echter kan men nu door een verdeling toe te kennen aan het percentage waarin de gekwadrateerde residuen meetellen, de rol van grote afwijkingen in de afhankelijke variabele reduceren.

7.5.4 Totale kleinste kwadraten

In tegenstelling tot lineaire kleinste kwadraten waarbij alleen de sommatie van de gekwadrateerde residuen langs de y-as gesommeerd worden, wordt bij de totale kleinste kwadraten ook naar de residuen langs de x-as gekeken. In het lineaire geval wordt dus een lijn $y = a + bx$ bepaald aan de hand van de gekwadrateerde euclydische afstanden. Dit wordt gebruikt als er naast een fout in de verklarende variabelen ook een fout in de afhankelijke variabele wordt verwacht.

7.5.5 Maximum-likelihood schatting en gerelateerde technieken

Deze methode gaat uit van een bekende stochastische verdeling van de datapunten. In het meestvoorkomende geval een normale verdeling met verwachting 0 en variantie σ^2. Zoals de naam van de schatter al doet vermoeden, maximaliseert deze schatter de likelihoodfunctie. De likelihoodfunctie van een parametervoorstelling (θ) moet per definitie gelijk zijn aan de kans op deze datapunten (x_i), gegeven de parametervoorstelling. Kortom, $\Lambda(\theta|x) = P(x|\theta)$, met Λ de likelihoodfunctie [Margenau and Murphy, 1943]. Omdat we er niet vanuit gaan dat onze datapunten komen uit een stochastische verdeling, zullen we deze schattingsmethodes niet gebruiken.

7.5.6 De Theil-Sen schatter

In [Wilcox, 2005] wordt nog een andere schatter geopperd. De Theil-Sen schatter is een schatter die de mediaan neemt van de helling van de lijn tussen alle datapunten. Deze methode is zeer robuust, simpel en weinig gevoelig voor afwijkende datapunten. Deze schatter geeft vaak een betere schatting dan de kleinste-kwadratenmethode in het geval dat er wordt gezocht naar een lijn van graad één. We zijn echter op zoek naar een polynoom van graad groter dan één om de datapunten te benaderen en dus zal deze schatter geen goede keuze zijn.

7.5.7 Motivatie gekozen schatter

Voor enkelvoudige en meervoudige lineaire regressie is het bewezen dat de kleinste kwadraten schatter de beste lineaire zuivere schatter is [Seber, 1997]. Wij zullen gebruik maken van deze schatter. Met behulp van deze methode worden er schattingen gedaan voor de parameters, door de kwadratische onderlinge afwijkingen tussen de data en de verwachte waardes te minimaliseren. We schatten β af met de waardes $\hat{\beta}$ die zorgen dat de kromme het best passend is. De kleinste kwadraten schatter B geeft de waardes $\hat{\beta}$ die $\sum_{i=1}^{n}(y_i - f_{\hat{\beta}}(x_i))^2$, met n het aantal datapunten, minimaliseert [van de Geer, 2005]. Deze formule is als volgt opgesteld: y_i is de gemeten waarde in het i^e datapunt en $f_{\hat{\beta}}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ...$ is hier de regressiefunctie, waarin we één datapunt x_i invullen. We kijken dan voor welke $\hat{\beta}$ in dit geval het verschil tussen de gevonden functie en het ingevulde datapunt zo klein mogelijk is. Deze optimale waarden $\beta = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ vormen de parameters $\beta = (\beta_0, \beta_1, ..., \beta_k)$ in het lineaire regressiemodel.

8 Een voorbeeld van lineaire regressie

Om te laten zien hoe we lineaire regressie gebruiken, beschrijven we hieronder een voorbeeld. De data hiervoor genereren we met behulp van een Matlabprogramma. In dit programma wordt uitgegaan van een bekend verband. Voor dit verband worden datapunten gegenereerd, met een fout op de snelheid. Vervolgens bekijken we hoe SPSS het originele verband destilleert uit de gegenereerde data. Het programma om de data te genereren ziet er als volgt uit.

```
i=1;
for v=[15:1:40]
    l(i)=v;
    Y(i)=F_total(1.2,v,0.07,225,0,1.225)+normrnd(0,1);
    i=i+1;
    l(i)=v;
    Y(i)=F_total(1.2,v,0.07,225,0,1.225)+normrnd(0,1);
    i=i+1;
end
B=[l' Y'];
xlswrite('data.xlsx', B, 'data1', 'B1');
```

In dit programma wordt een formule F_total aangeroepen. Deze formule wordt ingevuld voor snelheden van 15 m/s tot 40 m/s. De formule omschrijft het theoretische verband tussen de snelheid en de totale weerstand bij deze snelheid voor de Nuna 6 zonne-auto, volgens vergelijking (8). De gegevens die daarvoor nodig zijn, staan beschreven in tabel 6 [Wikipedia, 2012]. Dit zullen niet de meest nauwkeurige of betrouwbare waardes zijn, maar omdat het om een voorbeeld gaat is dat ook niet nodig.

Tabel 6: Gegevens Nuna 6

A	C_w	C_r	m	ρ
$1,2\ m^2$	$0,07$	$0,0025$	$225\ kg$	$1,225\ kg/m^3$

Het programma geeft voor verschillende waardes van de snelheid een output. Deze output bestaat uit de totale weerstand plus een standaard-normaal verdeelde fout. Ten slotte schrijft het progamma deze waardes naar een Excelbestand. Het is de bedoeling om met dit programma de gegevens na te bootsen die ook uit een proef zouden volgen. Bij bekende waardes van de snelheid worden "gemeten" waardes van de weerstand verkregen, waar een meetfout in kan meespelen. Het theoretische verband dat we terug moeten vinden is als volgt.

$$Y = 5,5181 + 0,052v^2. \tag{27}$$

Uit de theorie is bekend dat het verband tussen de weerstand en de snelheid polynomiaal is. Als afhankelijke variabele zal de weerstand genomen worden en als verklarende variabelen worden de machten van de snelheid genomen. De waardes voor de weerstand en snelheid laden we in SPSS en de hogere machten van de snelheid worden daar berekend. Vervolgens voeren we een regressieanalyse uit op de manier die beschreven staat in Appendix 1. Bij 52 variabelen geeft SPSS het volgende model.

$$Y = 6,266 + 0,051v^2. \tag{28}$$

Het 95%-betrouwbaarheidsinterval voor de constante term is van 5,511 tot 7,021 en voor de coëfficiënt van v^2 is het betrouwbaarheidsinterval van 0,050 tot 0,052. De coëfficiënt van v^2 is dus behoorlijk nauwkeurig afgeschat, deze zal in 95% van de gevallen een afwijking hebben van maximaal 0,001. De constante term is minder nauwkeurig, deze kan een afwijking van 0,755 naar boven of naar beneden hebben.
Het model is in 95% van de gevallen goed, want de coëfficiënten van de originele formule vallen binnen de betrouwbaarheidsintervallen. De adjusted R square is 0,997, wat aangeeft dat het verschil van het model en de datapunten al erg klein is.

Om te kijken hoeveel invloed de hoeveelheid data op de nauwkeurigheid van het model heeft, veranderen we de stapgrootte in het Matlabprogramma van 1 naar 0,5. Hierdoor wordt nu twee keer zoveel data gegenereerd, namelijk 104 datapunten. Als dezelfde procedure gevolgd wordt als met de vorige data, vindt SPSS het volgende model.

$$Y = 5,851 + 0,051v^2. \tag{29}$$

De 95%-betrouwbaarheidsintervallen zijn respectievelijk van 5,476 tot 6,226 en van 0,051 tot 0,052. De grootte van de betrouwbaarheidsintevallen is gehalveerd, wat aangeeft dat het model een stuk nauwkeuriger wordt bij tweemaal zoveel data. Om de verschillende modellen verder met elkaar te vergelijken, kunnen de verschillende waardes van de adjusted R square bekeken worden. Bij 104 variabelen is deze adjusted R square al 0,999, wat aangeeft dat het model nog maar een kleine afwijking heeft met de datapunten.

Dit voorbeeld laat zien dat een lineaire regressie analyse met SPSS een goed resultaat kan geven om het verband te bepalen. Niet alleen wordt er een verband gegeven, ook volgen er betrouwbaarheidsintervallen en andere hulpmiddelen om de kwaliteit van het model te evalueren. Doordat er een fout op de metingen zit is niet te garanderen dat het daadwerkelijke verband precies gevonden wordt. Daarom is het belangrijk om naar de betrouwbaarheidsintervallen te kijken. Deze betrouwbaarheidsintervallen bevatten bij het voorbeeld de daadwerkelijke waarde wel. Ook worden deze betrouwbaarheidsintervallen steeds kleiner als er meer data beschikbaar is om het model op te baseren. Aangezien in het daadwerkelijke probleem meer dan 104 datapunten beschikbaar zijn, zal deze methode zeker een goede manier zijn om ons verband tussen de snelheid en de versnelling af te schatten.

9 Methodes vergelijken: dataset 1

In hoodstuk 6 zijn verschillende methodes besproken die gebruikt kunnen worden om de versnelling te bepalen. Uiteindelijk kiezen we één methode die gebruikt gaat worden. De eerste dataset zullen we gebruiken om de methodes met elkaar te vergelijken.

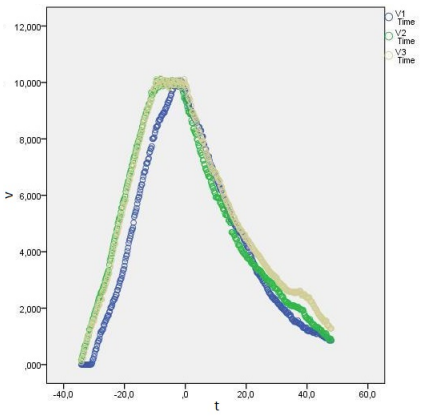
9.1 Dataset 1

Dataset 1 bevat de meetresultaten van de eerste uitroltest van het Solar Team. Er is hiervoor met een mock-up wagen gereden. Deze mock-up is eerst opgetrokken tot 10 m/s waarna de motor uitgezet is. Dit moment wordt tijdstip nul genoemd en daarna wordt per 0.2 seconden de snelheid van de mockup gemeten in de motorcontrolunits (MCU's) en op de display. Dit is drie keer herhaald. Per ritje is er een Excelbestand gemaakt. Het Excelbestand met de data ziet er als volgt uit: in de eerste kolom vinden we de tijd (in seconden), in de tweede en derde kolom vinden we de snelheden (in meters per seconden) gemeten door de twee MCU's, in de vierde kolom staat het cruise-control-setpoint, in de vijfde kolom de snelheid (in meters per seconden) die weergeven wordt op de display van de mock-up en in de laatste kolom vinden we het gemiddelde van de twee snelheden gemeten door de MCU's. Door het Solar Team is de verwachting uitgesproken dat deze laatste gemiddelde snelheid het meest representatief is voor de werkelijkheid. Deze snelheid wordt dan ook gebruikt in de komende methodes. In deze dataset wordt er vanuit gegaan dat de wind constant is voor de drie verschillende ritjes. De wind wordt daarom niet meegenomen als variabele. Een deel van het Excelbestand van een specifieke uitroltest ziet er als volgt uit.

	A	B	C	D	E	F
1	Time	MC0 vehicle_velocity	MC1 vehicle_velocity	Steer setpoint	Steer velocity	MC_avg_Vel
2	0	9,593	9,806	0	9,629	9,6995
3	0,2	9,479	9,476	0	9,629	9,4775
4	0,4	9,479	9,324	0	9,429	9,4015
5	0,6	9,479	9,298	0	9,429	9,3885
6	0,8	9,479	9,298	0	9,429	9,3885
7	1	9,304	9,171	0	9,298	9,2375
8	1,2	9,304	9,124	0	9,298	9,214
9	1,4	9,139	9,029	0	9,298	9,084
10	1,6	9,065	9,016	0	9,298	9,0405
11	1,8	9,065	9,016	0	9,298	9,0405
12	2	8,992	8,928	0	9,298	8,96

Figuur 4: Excelbestand uitroltest 1, rit 1.

Hoewel dezelfde proef drie keer herhaald is, zijn er toch verschillen in de data. Dit komt onder andere doordat er een afwijking zit in het moment waarop de motor precies uitgezet wordt. In het eerste ritje is de snelheid namelijk $9,6995$ m/s als de motor uit wordt gezet, terwijl dit bij de tweede rit $9,397$ m/s is en bij de derde $9,924$ m/s. Hierdoor zijn de grafieken ten opzichte van elkaar verschoven. Dit is te zien in de afbeelding hiernaast, waarin ook nog andere verschillen zichtbaar zijn. De blauwe grafiek heeft weinig overeenkomsten met de groene en de gele grafiek. De groene en de gele grafiek hebben meer overeenkomsten. In deze grafieken zitten er op dezelfde momenten "hobbels" in de grafiek. De grootte van deze hobbels is wel verschillend. Een verklaring van deze hobbels zou gegeven kunnen worden door variaties in de wind. Het verschil in grootte zou kunnen worden verklaard door het verschil in windsnelheid. Zo



Figuur 5: Datapunten uitroltest 1.

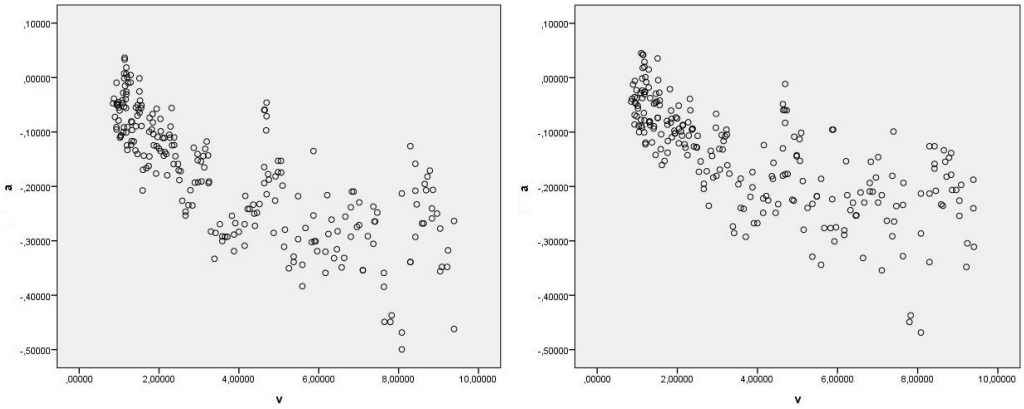
nemen wij aan dat rit 2 (groen) en rit 3 (geel) tussen de 30 en 50 seconden een meewerkende windkracht ondervonden. Hierdoor heeft de mock-up relatief minder vertraging dan buiten dit interval. De verschillen tussen de grafieken zouden ook het gevolg kunnen zijn van onnauwkeurigheden tijdens de uitroltests. Uiteindelijk zal het onmogelijk zijn om een verband te vinden dat met alle drie de ritten een kleine afwijking heeft.

We willen de methodes voor het bepalen van de versnelling vergelijken. Dit betekent dat we de gegeven snelheden per tijdstip zullen omrekenen naar versnellingen. Uit deze versnellingen zullen we dan een verband bepalen en zullen we met behulp van SPSS bekijken hoe groot de afwijking van dit verband is. Het bepalen van de versnellingen kan alleen per rit. Gezien we op dit moment alleen willen weten welke methode het beste de versnelling bepaald en het dus niet van belang is welk verband we vinden, zullen we daarom naar één specifieke rit kijken. Hiervoor kiezen we de rit die er het meest representatief uit ziet, in ons geval dus rit 1 (blauw).

Met behulp van SPSS zullen we kijken of er eenzelfde verband wordt gevonden met de verschillende methodes. Is dit het geval dan zou het niet uitmaken welke methode we uiteindelijk gebruiken. We hebben in Appendix 1 vermeld dat we naar de adjusted R square kunnen kijken om te zien hoeveel procent van de variantie van de gemeten variabele door het model verklaard wordt. Echter kunnen we deze gevonden waardes bij de verschillende methodes niet met elkaar vergelijken. De verschillende methodes veranderen namelijk de data. Dit betekent dat de versnellingen die als input gebruikt worden per methode verschillend zijn. De adjusted R square zegt dus in het algemeen niets over welke methode het beste model oplevert.

9.2 Numerieke methode A en B

Als eerste worden de numerieke methodes A en B getest. Aangezien deze methodes redelijk op elkaar lijken worden deze methodes hier naast elkaar getest. De waardes voor $\tilde{v}(i)$ en $a(i)$ worden in het Excelbestand berekend op de manier beschreven in hoofdstuk 6. De verkregen versnellingen zetten we uit tegen de gemeten snelheden. Dit leidt tot de volgende afbeeldingen.



(a) Versnelling tegen snelheid, methode A. (b) Versnelling tegen snelheid, methode B.

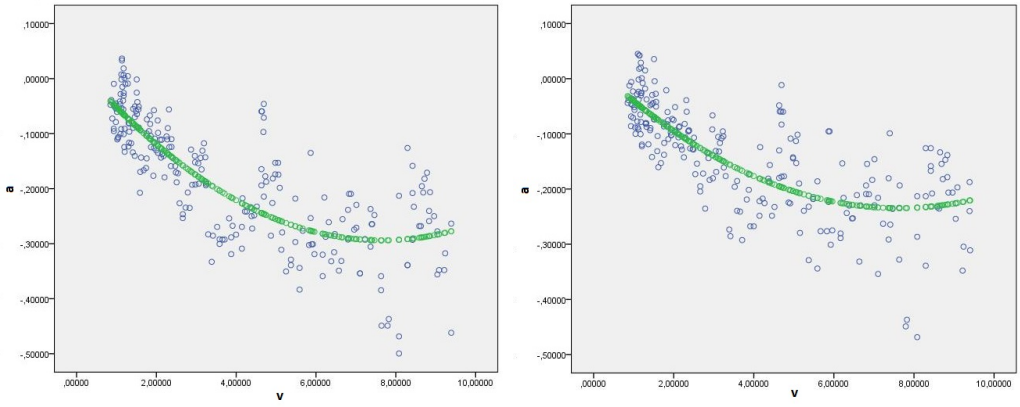
Figuur 6: Methode A en B, versnelling tegen snelheid.

Vervolgens wordt met deze data een lineaire regressieanalyse in SPSS gedaan zoals beschreven staat in Appendix 1. De modellen die hier uit volgen zijn als volgt:

$$a_A(v) = 0,026 - 0,084v + 0,005v^2, \tag{30}$$
$$a_B(v) = 0,022 - 0,067v + 0,004v^2. \tag{31}$$

Deze modellen kunnen ook grafisch weergeven worden. Zo kunnen we bekijken hoe het model

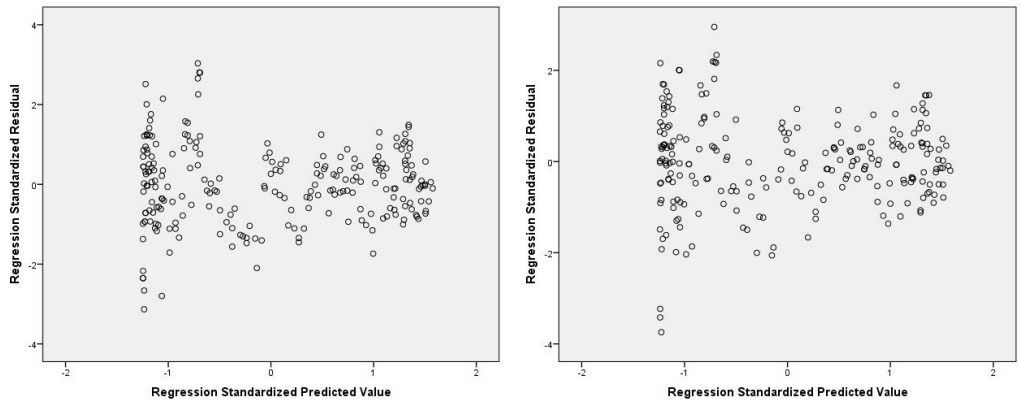
eruit ziet ten opzichte van de data. Dat is in de volgende afbeelding te zien.



(a) Gevonden verband met methode A. (b) Gevonden verband met methode B.

Figuur 7: Methode A en B, gevonden verbanden.

De juistheid van de aannames voor het maken van het model kan gecontroleerd worden aan de hand van de scatterplots. Als er in deze afbeeldingen chaos te zien is, houdt dat in dat er geen onverklaard verband meer tussen de meetdata zit. Deze plots zien er als volgt uit.

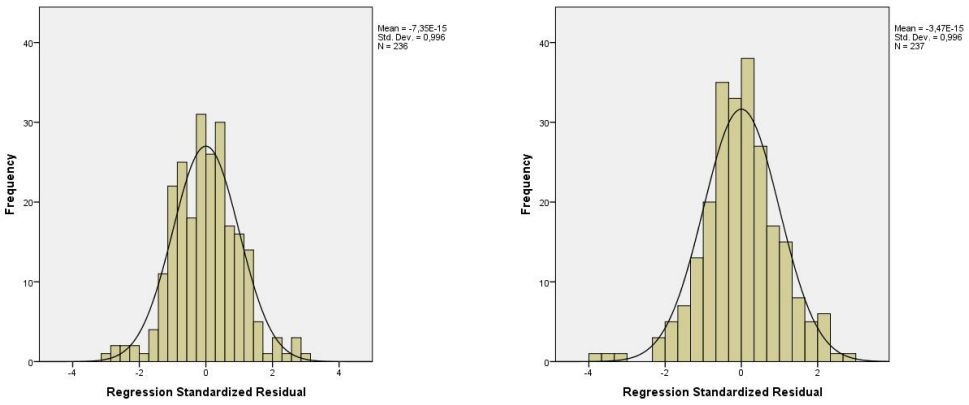


(a) Scatterplot methode A. (b) Scatterplot methode B.

Figuur 8: Methode A en B, scatterplots.

In deze plots is nog niet totale chaos te zien. Dit zou kunnen aangeven dat er nog een onverklaard element in de data zit. Dit kan komen doordat de wind niet meegenomen wordt als variabele, of door effecten van de omgeving. De metingen zijn gedaan op een weg met zijstraten. Er zou een variatie in de wind kunnen zijn in een vast patroon, afhankelijk van de plaats van de zijstraten. Een andere reden voor het ontbreken van totale chaos zijn de metingen. Doordat het een uitrolproef betreft, zijn er veel meer datapunten in de buurt van snelheid 0 m/s dan bij een snelheid van bijvoorbeeld 8 m/s.

Om de aanname te controleren dat de afwijkingen normaal verdeeld zijn wordt de histogram van de residuen bekeken. Deze zien er als volgt uit.



(a) Histogram methode A.

(b) Histogram methode B.

Figuur 9: Methode A en B, histogrammen.

Zoals te zien is, zijn de afwijkingen inderdaad normaal verdeeld.

Bij de numerieke methodes A en B wordt er dus aan de aannames voldaan. Deze methodes kunnen toegepast worden om een model te vinden. Echter zit er wel verschil in de gevonden modellen. Dit betekent dat één van beide methodes een model vindt dat dichter bij de werkelijkheid ligt. Welke van de twee dit is zullen we nog moeten testen.

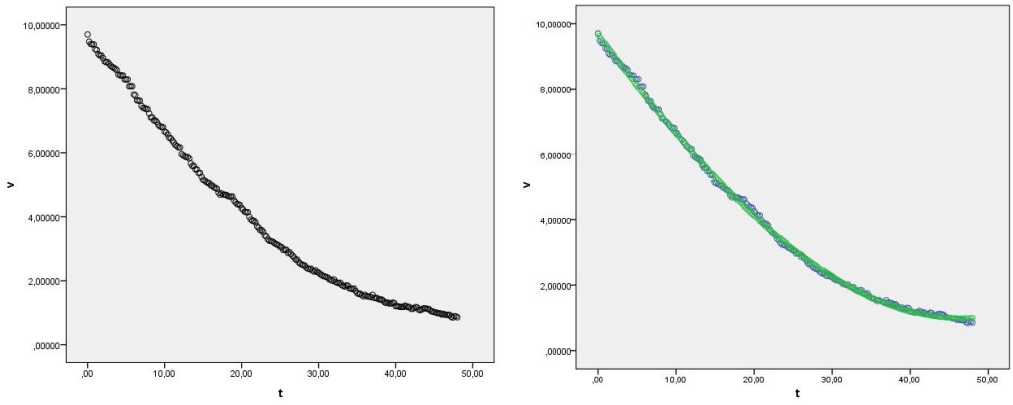
9.3 Analytische methode

Als de versnelling met behulp van de analytische methode bepaald wordt, gaat dit op een heel andere manier dan bij de numerieke methodes. Deze methode wordt daarom apart behandeld. Bij de analytische methode wordt om te beginnen een verband gevonden voor de snelheid naar de tijd. Vervolgens zal dit verband afgeleid worden. Om het verband $v(t)$ te bepalen, voeren we een lineaire regressieanalyse uit met als afhankelijke variabele de snelheid en als verklarende variabelen de machten van de tijd.

Het is moeilijk te bepalen welk model het meest passend is voor de datapunten. Het model past zich namelijk steeds aan, als er hogere ordes van de onafhankelijke variabelen beschikbaar gesteld worden. In SPSS betekent dit dat de Stepwise methode, uitgelegd in paragraaf 1 steeds een model oplevert die afhangt van een constante, t, t^2, t^3 en een erg hoge macht van t. Gezien de theorie uit paragraaf 7.1 zegt dat dit niet wenselijk is wordt er opnieuw een lineaire regressie analyse uitgevoerd, waarbij de methode Enter wordt gekozen. Hierbij worden alleen t, t^2 en t^3 opgegeven als verklarende variabelen. De data waar we een verband voor willen vinden is te zien in afbeelding 10a. Na een lineaire regressieanalyse wordt het volgende verband gevonden.

$$v = 9,681 - 0,325t + 0,002t^2 + 2,238 \cdot 10^{-5}t^3. \tag{32}$$

Om te controleren of het verband realistisch is, bekijken we in afbeelding 10b het gevonden verband ten opzichte van de datapunten.

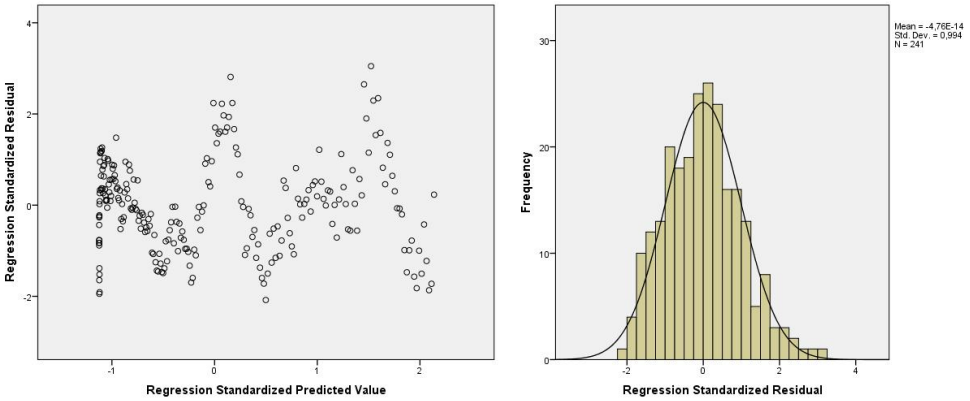


(a) Datapunten, snelheid tegen tijd. (b) Gevonden verband, snelheid tegen tijd.

Figuur 10: Analytische methode, snelheid tegen tijd.

Om dit model te bepalen hebben we de snelheden rechtstreeks uit de data gehaald. De adjusted R square van het model dat we met behulp van deze "ruwe" snelheden hebben bepaald is $0,999$. Dit betekent dat het model een erg kleine afwijking heeft met de datapunten. Gezien deze waarde al zo dicht bij 1 ligt, is het niet nodig om de snelheden te smoothen voordat een lineaire regressieanalyse wordt uitgevoerd.

Aan de scatterplot is te zien dat er een onnauwkeurigheid in onze data zit. Ook wordt gecontroleerd of de afwijkingen normaal verdeeld zijn. In de afbeelding is zowel de scatterplot als de histogram te zien.



(a) Scatterplot Analytische methode, snelheid tegen tijd. (b) Histogram Analytische methode, snelheid tegen tijd.

Figuur 11: Analytische methode, scatterplot en histogram.

De residuen zijn wederom normaal verdeeld.
Het verkregen model geeft het verband van de snelheid ten opzichte van de tijd. Er wordt echter gezocht naar een verband van de versnelling ten opzichte van de tijd. Om het verband van de versnelling te krijgen wordt het verband van de snelheid afgeleid naar de tijd.

$$a(t) = \frac{d}{dt}v(t) = \frac{d}{dt}(9,681 - 0,325t + 0,002t^2 + 2,238 \cdot 10^{-5}t^3). \quad (33)$$

$$a(t) = -0,325 + 0,004t + 6,714 \cdot 10^{-5}t^2. \quad (34)$$

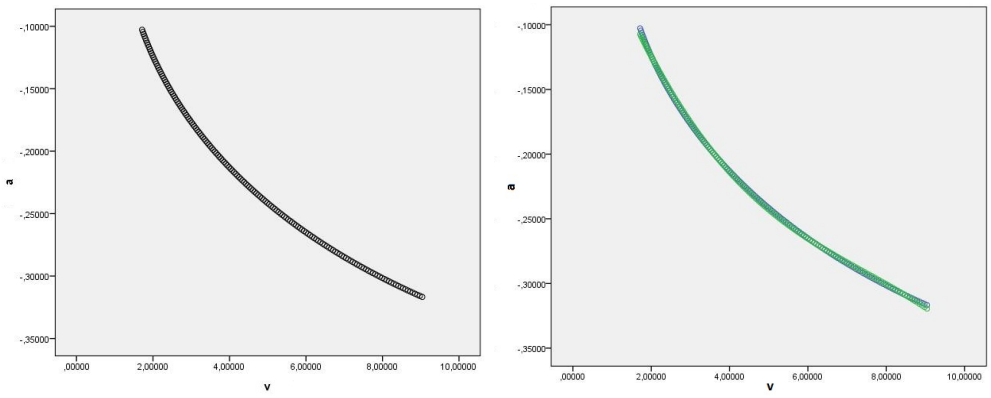
Om nu net als bij de numerieke methodes een verband te vinden voor de versnelling ten opzichte van de snelheid, worden met behulp van een Matlabprogramma datapunten gegenereerd voor v en a. Het programma dat hiervoor gebruikt wordt ziet er als volgt uit.

```
i=1;
for t=2:0.2:35
    x(i)=V(t);
    y(i)=a(t);
    i=i+1;
end
B=[x' y'];
xlswrite('mockup.xlsx', B , 'Blad1', 'A1');
```

In dit programma worden er voor tijdstippen van 2 tot 35 datapunten van de snelheid en de versnelling uitgerekend. Deze waardes worden naar een Excelbestand geschreven, waar de snelheid in de eerste kolom en de versnelling in de tweede kolom staat. Deze data kan met behulp van SPSS tegen elkaar uitgezet worden, zodat er een verband bepaald wordt. De data wordt weergeven in afbeelding 12a. Met SPSS wordt een verband bepaald tussen de versnelling en de snelheid. Dit geeft het volgende model.

$$a(v) = 0,024 - 0,093v + 0,010v^2 + 0,000448v^3. \tag{35}$$

Dit model bekijken we ten opzichte van de data. Dit is te zien in afbeelding 12b.

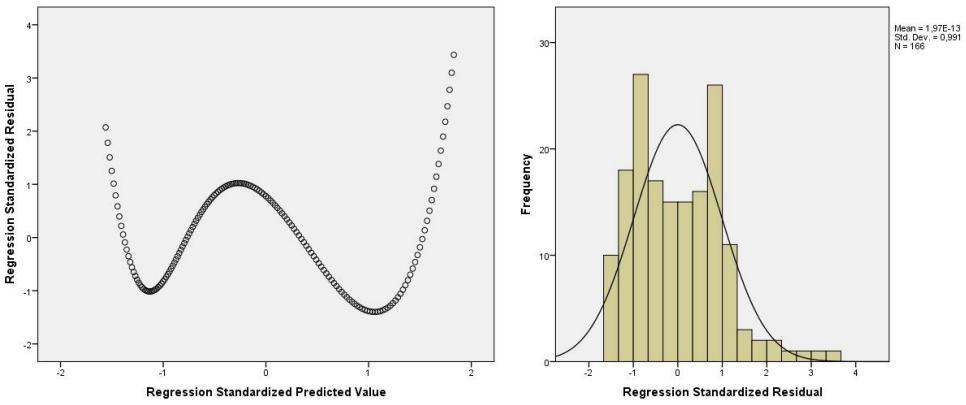


(a) Berekende datapunten, versnelling tegen snelheid.

(b) Gevonden verband, versnelling tegen snelheid.

Figuur 12: Analytische methode, versnelling tegen snelheid.

Dit model wijkt weer enigszins af van de modellen gevonden met de numerieke methodes. Dit betekent dat ook deze methode getest moet worden ten opzichte van de andere methodes. De scatterplot en histogram werpen nog wat twijfels op over dit model.



(a) Scatterplot Analytische methode, versnelling tegen snelheid.

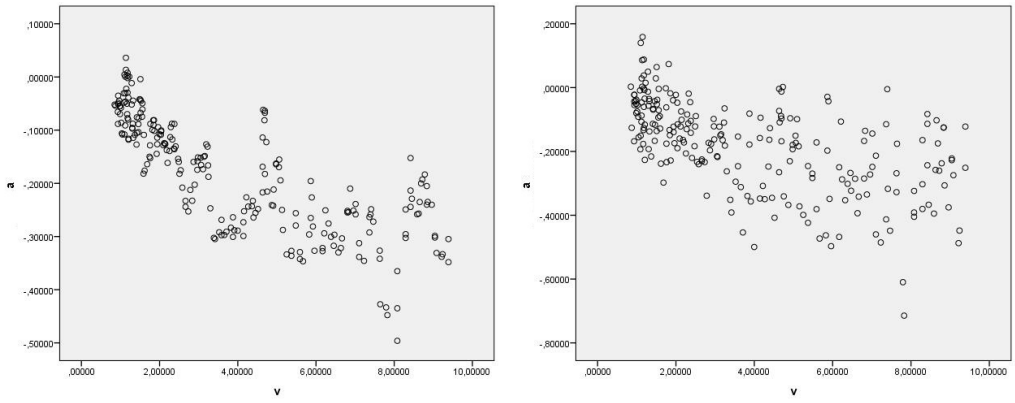
(b) Histogram Analytische methode, versnelling tegen snelheid.

Figuur 13: Analytische methode, scatterplot en histogram.

Zoals in afbeelding 13a en 13b te zien is, vertoont de scatterplot geen chaos en hebben de residuen geen uitgesproken normale verdeling. Er wordt dus niet voldaan aan de aannames die nodig zijn om lineaire regressie uit te voeren. In vergelijking met de numerieke methodes lijkt de analytische methode dus niet wenselijk om te gebruiken.

9.4 Savitzky-Golay

De laatste methode die we gaan testen is de Savitzky-Golay methode. Zoals in paragraaf 5.2 beschreven is, kunnen we met deze methode direct naar een willekeurige orde afgeleide van datapunten kijken. Waar de uitroltesten datapunten geven voor de snelheid (v) tegen de tijd (t), kunnen we met deze methode dus direct kijken naar de gesmoothde versnelling (a) op elk tijdstip. Met behulp van Matlab, waar met het commando 'Savitzkygolayfilt('data','orde polynoom','orde afgeleide','framesize')' de methode eenvoudig toegepast kan worden, hebben we meteen de nieuwe gesmoothde datapunten voor de versnelling verkregen. Dit hebben we gedaan door een polynoom van orde 1, oftwel een rechte lijn en een polynoom van graad 3, op q punten te passen. De datapunten van de gesmoothde versnelling tegen de snelheid staan hieronder.



(a) Versnelling tegen snelheid, Savitzky-Golay orde 1.

(b) Versnelling tegen snelheid, Savitzky-Golay orde 3.

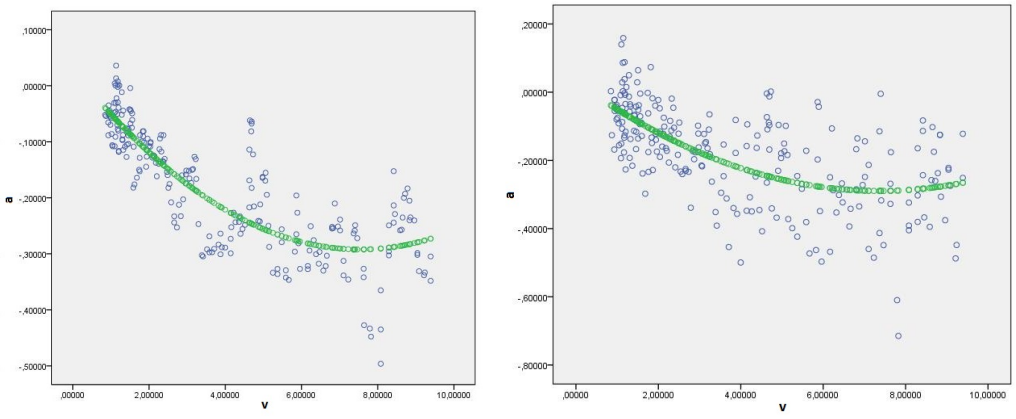
Figuur 14: Savitzky-Golay orde 1 en 3, versnelling tegen snelheid.

Met deze datapunten hebben we weer een lineaire regressie analyse uitgevoerd in SPSS, met

de volgende modellen als resultaat:

$$a_{orde1}(v) = 0,029 - 0,085v + 0,006v^2, \tag{36}$$
$$a_{orde3}(v) = 0,031 - 0,087v + 0,006v^2. \tag{37}$$

Als we deze verbanden in een grafiek weergeven, ziet het er als volgt uit.

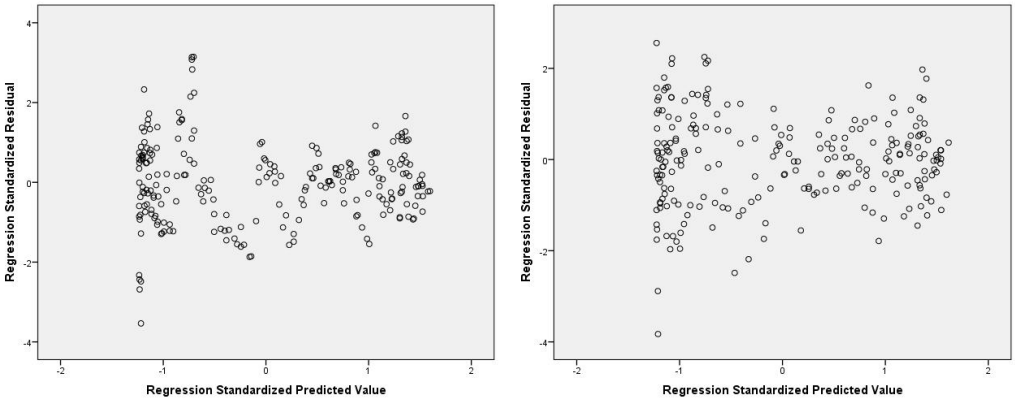


(a) Gevonden verband met Savitzky-Golay methode orde 1.

(b) Gevonden verband met Savitzky-Golay methode orde 3.

Figuur 15: Savitzky-Golay orde 1 en 3, gevonden verband.

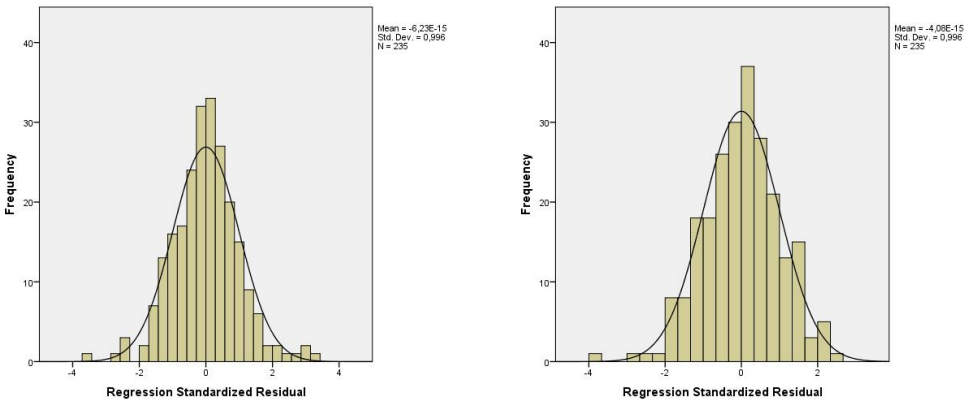
De scatterplots en histogrammen die bij het gevonden verband horen zijn als volgt.



(a) Scatterplot Savitzky-Golay orde 1.

(b) Scatterplot Savitzky-Golay orde 3.

Figuur 16: Savitzky-Golay orde 1 en 3, scatterplots.



(a) Histogram Savitzky-Golay orde 1. (b) Histogram Savitzky-Golay orde 3.

Figuur 17: Savitzky-Golay orde 1 en 3, histogrammen.

Aan de scatterplot is te zien dat bij het model van orde 3 meer sprake te zijn van chaos dan bij orde 1. In de histogram kijken we naar de verdeling van de residuen. Zoals te zien, zijn de residuen normaal verdeeld.

9.5 Vergelijking van de resultaten

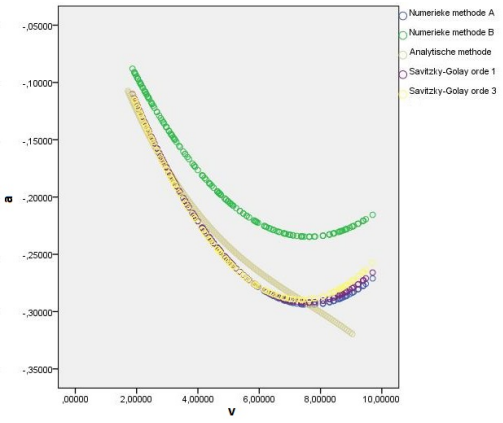
We hebben alle methodes toegepast op de eerste rit van dataset 1. Hier zijn modellen voor de versnelling afhankelijk van de snelheid uitgekomen. De coëfficiënten in deze modellen staan in de onderstaande tabel.

Tabel 7: Modellen voor $a(v)$.

$a(v)$	v^0	v^1	v^2	v^3
Numerieke methode A	0,026	-0,084	0,005	0
Numerieke methode B	0,022	-0,067	0,004	0
Analytische methode	0,024	-0,093	0,010	0,000448
Savitzky-Golay orde 1	0,029	-0,085	0,006	0
Savitzky-Golay orde 3	0,031	-0,087	0,006	0

De verschillende verbanden kunnen we ook grafisch vergelijken. De verbanden staan weergeven in een figuur 18.

Kortom, er is te zien dat de meeste verbanden op elkaar lijken. Het verband van de analytische methode wijkt qua vorm wel significant af van de andere verbanden. Ook het verband van numerieke methode B is anders. We kunnen op basis van de resultaten echter niets zeggen over de juistheid van de gevonden verbanden en ook niet concluderen dat één methode boven een andere gekozen zou moeten worden. Om wél te kunnen concluderen welke methode het beste is, zullen we verder moeten testen.



Figuur 18: Gevonden verbanden.

10 Methodes testen

In hoofdstuk 9 hebben we verschillende methodes toegepast om een verband te vinden die de versnelling tegen de snelheid uitzet en deze te vergelijken. We hebben geconcludeerd dat we op het eerste gezicht niet kunnen zien welke van deze methodes het best passende verband geeft. We zullen dus een manier moeten bedenken om te bepalen welke methode we gaan gebruiken. De enige kennis die we hebben over een model tussen de snelheid en de versnelling zijn de theoretische vergelijkingen. Deze vergelijkingen zullen we gaan gebruiken om te testen of onze methodes een realistisch verband vinden.

10.1 Data genereren

Om de methodes te testen is een Excelbestand gemaakt. In dit bestand wordt het theoretische model gebruikt om data te genereren. Het genereren van deze data gaat als volgt.

Allereerst worden een aantal vaste waardes gedeclareerd. Zo wordt de voorwaartse kracht op 0 N gezet. De rolweerstand is onafhankelijk van de snelheid en wordt op 5 N gezet. Voor de luchtweerstand wordt een formule gebruikt: $c_l(v(t) - v_{wind}(t))^2$. De constante c_1 wordt vastgesteld op 2. De windsnelheid wordt in deze formule ook genoemd, maar die wordt op 0 gezet. Ten slotte wordt de massa vastgesteld op 100 kg, de gravitatieconstante op $9,81$ m/s^2, de beginsnelheid op 10 m/s en de grootte van de tijdstappen op $0,1$ seconden.

Het genereren van data gaat nu als volgt. Op tijdstip nul wordt gestart met een snelheid van 10 m/s. Voor deze snelheid zijn de theoretische weerstanden te bepalen. De rolweerstand is 5 N en de luchtweerstand is $2 \cdot 10 = 200$ N. Vervolgens kan hiermee een versnelling berekend worden met behulp van de in hoofdstuk 2 genoemde formule $F_{res} = m \cdot a = -F_a - F_r$. Voor de versnelling wordt dus de volgende formule verkregen: $a = \frac{-F_a - F_r}{m}$. Deze versnelling kan bepaald worden met bepaalde weerstanden. Vervolgens kan met deze versnelling de snelheid voor het volgende tijdstip bepaald worden, met deze snelheid de nieuwe weerstanden, hiermee de nieuwe versnelling enzovoorts. Op deze manier wordt iteratief de data gegenereerd. Het verband dat volgt uit de theorie is als volgt.

$$a(v) = -0,02v^2 - 0,05. \tag{38}$$

Om de methodes te testen wordt er een meetfout bij de snelheid opgeteld. In het Excelbestand wordt in een kolom voor elke snelheid een onafhankelijke meetfout bepaald door een standaardnormaalverdeeld getal te vermenigvuldigen met $0,1$. Vervolgens wordt deze meetfout opgeteld bij de snelheid zodat er snelheden met meetfouten verkregen zijn. Met deze "gemeten"snelheden kunnen de methodes getest worden. Als er een methode is die hetzelfde verband vindt voor de versnelling ten opzichte van de snelheid, kan dit gezien worden als een goede methode. Als er meerdere methodes zijn die het verband vinden, zal er een keuze gemaakt moeten worden op basis van andere criteria. Als er geen enkele methode is die het verband terug vindt, moet een keuze gemaakt worden tussen het bedenken van een nieuwe methode of het kiezen van de methode die het dichtst in de buurt komt van het gezochte verband.

10.2 Gevonden verbanden

Met behulp van het Excelbestand zijn datapunten voor de tijd en de snelheid gegenereerd. Op deze data kunnen de methodes om de versnelling te bepalen getest worden. De methodes die getest worden staan beschreven in hoofdstuk 6. De datapunten voor de versnelling worden bepaald zoals beschreven staat per methode. Vervolgens wordt met deze data een lineaire regressieanalyse toegepast zoals beschreven staat in Appendix 1.

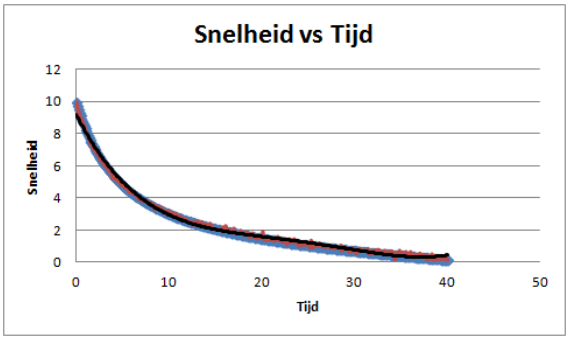
Bij Savitzky-Golay zijn nog bepaalde keuzes te maken. Voor het bepalen van versnelling, kan de grootte van het interval gekozen worden en de graad van het polynoom dat gebruikt wordt. Als intervalgrootte wordt een interval van 7 punten gekozen. Dit is in het algemeen groot genoeg om

een datapunt te smoothen. Voor de graad van het polynoom bekijken we zowel graad 1 als graad 3. Graad 1 omdat meetfouten hierbij het snelste worden gecorrigeerd. Graad 3 kiezen we omdat we denken dat dit een realistischer beeld geeft van het verloop van de snelheid bij een uitroltest dan andere orde polynomen. De resultaten van alle methodes worden hieronder weergeven.

Tabel 8: Gevonden model per methode.

Methode	Verband
Numerieke methode A	$a(v) = -0,02v^2 - 0,05$
Numerieke methode B	$a(v) = -0,016v^2 - 0,038$
Analytische methode	$a(v) = 9,869 \cdot 10^{-5}v^3 - 0,002v^2 + 0,009v - 0,276$
Savitzky-Golay orde 1	$a(v) = -0,02v^2 - 0,05$
Savitzky-Golay orde 3	$a(v) = -0,02v^2 - 0,06$

Het is te zien dat het resultaat van numerieke methode A erg lijkt op het gezochte verband (38). Het verband van numerieke methode B wijkt een stuk meer af van (38) dan het verband gevonden met numerieke methode A. Deze methode werkt dus minder goed. De resultaten van de analytische methode wijken erg af van het gezochte verband (38). Na een korte analyse is hier ook een vermoedelijke oorzaak voor gevonden. In deze methode wordt allereerst het verband $v(t)$ benaderd met een polynoom. Dit verband is echter niet polynomiaal, waardoor er al een grote afwijking ontstaat. Dit is ook in de volgende afbeelding te zien.



De zwarte lijn geeft de benadering met een polynoom weer, welke door het werkelijke verband heen golft. Dit zorgt voor een afwijking, die meegenomen wordt bij het berekenen van de versnelling. Aan de hand van deze test kan dus de conclusie getrokken worden dat deze methode niet goed werkt.
Het resultaat voor Savitzky-Golay met graad 1 is erg vergelijkbaar met het resultaat van de numerieke methode A. Dit is logisch, aangezien de methodes praktisch hetzelfde doen. Savitzky-Golay smoothen met graad 1 komt namelijk op hetzelfde neer als Nearest-neighbor smoothing. Er zit wel een verschil in het berekenen van de versnelling.

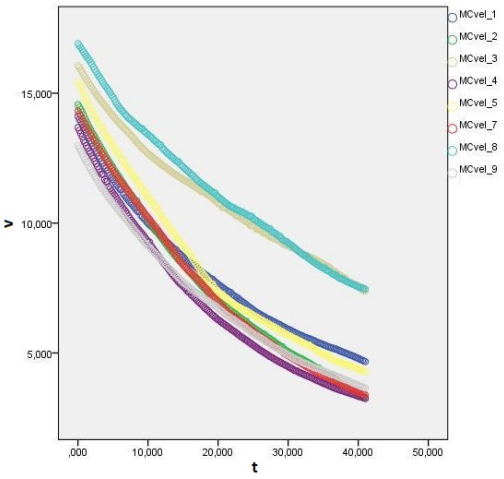
10.3 Keuze van methode

Uit de analyse die hierboven gedaan is, blijkt dat de numerieke methode A en de Savitzky-Golay methode het beste resultaat geven. Voor Savitzky-Golay zijn er Matlabfuncties beschikbaar, welke het berekenen van de versnelling erg gemakkelijk maken. Ook kan in deze functie de intervalgrootte makkelijk aangepast worden. Er wordt dus gekozen om met de Savitzky-Golay methode orde 1 verder te werken. Deze zullen we vanaf nu de Savitzky-Golay methode noemen.

11 Terugkoppeling naar de praktijk

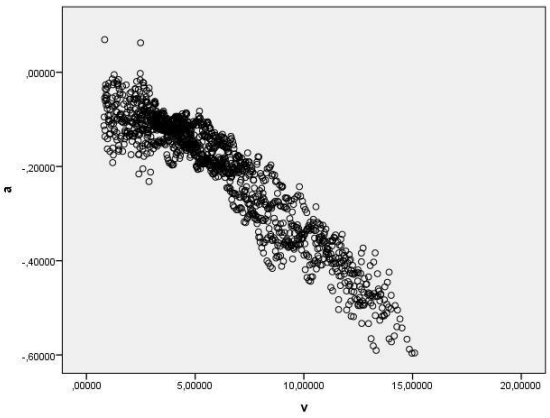
We hebben bepaald dat de Savitzky-Golay methode de beste methode is om een verband tussen de versnelling en de snelheid te vinden. We zullen nu met behulp van deze methode het verband bepalen uit de tweede dataset. Hiervoor worden eerst de metingen bestudeerd.

De tweede dataset is op eenzelfde manier verkregen als de eerste dataset, maar dit keer slechts met één MCU. Dit houdt in dat er versneld wordt tot ongeveer 15 m/s en dat vervolgens de motor uitgezet wordt. Er wordt elke 0,2 seconden een meting van de snelheid gegeven. De windsnelheid tijdens de ritten was niet te meten. Wel is bekend dat de windsnelheid die dag ongeveer 4 m/s was en dat enkele ritten met wind mee en enkele met wind tegen zijn gereden. In de grafiek is te zien dat rit 3 en rit 8 met meewind gereden zijn. Gezien we maar twee metingen voor de wind hebben en deze niet nauwkeurig zijn, hebben we besloten de windsnelheid niet als variabele mee te nemen, maar als constante te beschouwen. We kijken dus enkel naar rit 1, 2, 4, 5, 7 en 9 (de metingen van rit 6 waren onbruikbaar). In de grafiek is ook te zien dat de snelheid waarop de motor uit wordt gezet erg varieert. De ritten zien er verschillend uit, maar zullen wel eenzelfde verband representeren.



Figuur 19: Datapunten uitroltest 2.

We kunnen dus alle data bij elkaar nemen en daar een gemiddeld verband op bepalen. De datapunten van de versnelling uitgezet tegen de snelheid zijn zichtbaar in de volgende afbeelding.



Figuur 20: Datapunten versnelling tegen snelheid.

Nu zullen we door middel van de Savitzky-Golay methode de versnellingen bepalen. De versnellingen uitgezet tegen de snelheden zijn hiernaast weergeven. Vervolgens wordt met deze data een lineaire regressieanalyse in SPSS gedaan zoals beschreven staat in Appendix 1.

Het model dat hieruit volgt is:

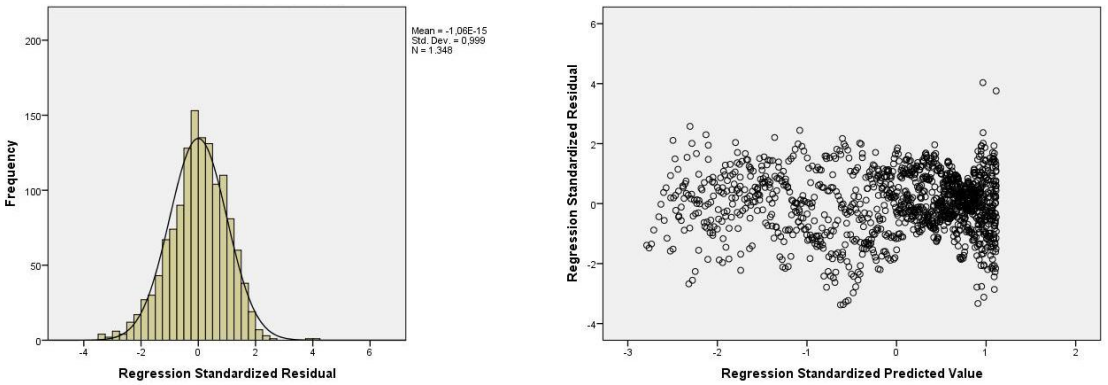
$$a(v) = -0,075 - 0,003v^2 + 5,271 \cdot 10^{-6}v^4.$$

Het model heeft een adjusted R square van 0,903. Dit betekent dat 90,3% van de variantie van de versnelling verklaard wordt door het verschil in snelheid. Er is dus een sterke samenhang. Vervolgens wordt er gekeken naar het betrouwbaarheidsinterval van de geschatte coëfficiënten. De 95%-betrouwbaarheidsintervallen zijn als volgt.

Tabel 9: 95% betrouwbaarheidsintervallen.

Coëfficiënt van	Ondergrens	Bovengrens
v^0	-0,079	-0,071
v^2	-0,003	-0,003
v^4	$4{,}0{\cdot}10^{-6}$	$6{,}0{\cdot}10^{-6}$

Er is te zien dat de betrouwbaarheidsintervallen erg klein zijn. Dit betekent dat het model nauwkeurig is. De juistheid van de aannames voor het maken van het model kan gecontroleerd worden aan de hand van de histogram en de scatterplot.



(a) Verdeling van de afwijking.

(b) Scatterplot versnelling afhankelijk van snelheid.

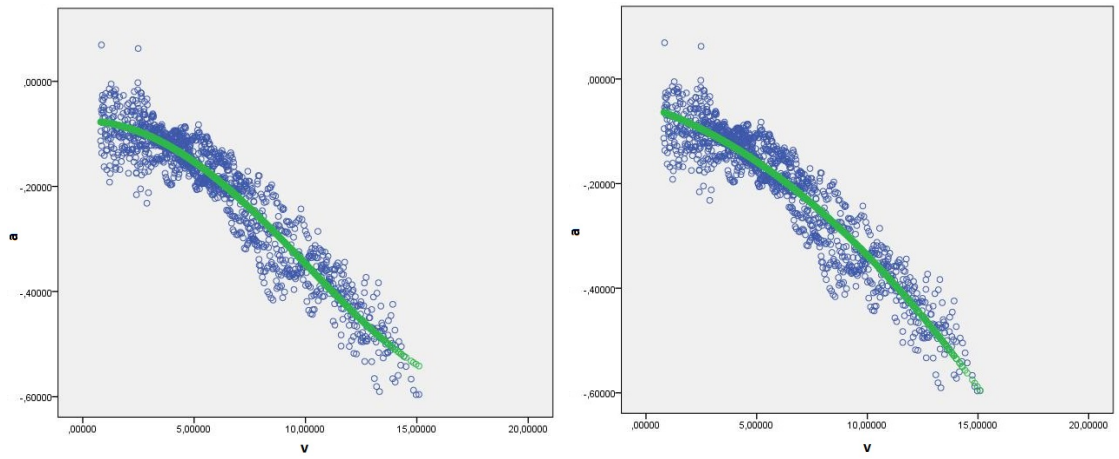
Figuur 21: Controle aannames.

Aan de histogram is te zien dat de residuen normaal verdeeld zijn. De scatterplot vertoont niet totale chaos. Dit kunnen we verklaren door de manier waarop de metingen gedaan zijn. Er worden namelijk per tijdstip metingen gedaan. Hierdoor zijn er meer datapunten bij lagere snelheden. Dit verklaart de opstapeling van datapunten aan de rechterzijde van de scatterplot.

Hoewel de aannames bevestigd worden en het model nauwkeurig is, zijn we er toch niet tevreden mee. Het model voldoet namelijk niet aan alle eisen die besproken zijn in paragraaf 7.1. In het model staan een tweede en een vierde orde term. Volgens de literatuur moeten er dan ook een eerste en derde orde term in het model staan. Bovendien zou de mock-up moeten voldoen aan de theoretische vergelijkingen. We verwachten dus een verband van dezelfde vorm als vergelijking (14). Als we het gevonden verband op de data leggen, zoals in afbeelding 22a te zien is, zien we ook dat de vierde orde term zorgt voor raar gedrag bij hogere snelheden. Rond de 15 m/s begint de vertraging namelijk af te nemen, terwijl we verwachten dat hoe harder er gereden worden hoe sneller de wagen vertraagd. We zullen dus ook het model beschouwen waarin we de vierde orde term verwaarlozen. We doen dit door opnieuw een lineaire regressieanalyse uit te voeren, met als enige beschikbare variabelen v en v^2.
Het model wordt dan als volgt:

$$a(v) = -0,052 - 0,014v - 0,001v^2. \tag{39}$$

Dit model geldt alleen voor een situatie met een tegenwind van ongeveer 4 m/s.



(a) 4^e-graads model.

(b) 2^e-graads model.

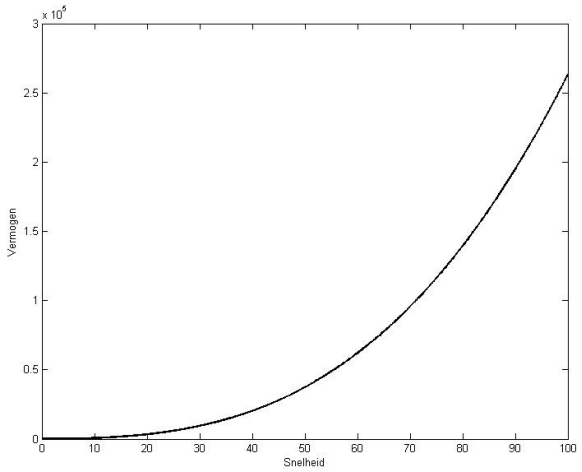
Figuur 22: Gevonden modellen.

Als we de grafieken van het 2^e- en 4^e-graads model met elkaar vergelijken zien we dat vooral bij de hogere snelheden de verbanden afwijkend zijn. Het 4^e-graads model vertoont in het uiteinde namelijk onrealistisch gedrag. Het 2^e-graads model gedraagt zich zoals we verwachten. Als we de twee modellen vergelijken zien we dat de betrouwbaarheidsintervallen ongeveer even groot zijn. De adjusted R square ligt bij het tweede model lager. Deze is namelijk 0,898. Toch zullen we op basis van de theorie en de bovenstaande figuren kiezen voor het 2^e-graads model.

Met behulp van de vergelijkingen voor de kracht (2) en het vermogen (1), de versnelling (39) en de bekende waarde m, vinden we:

$$P = -F \cdot v = -m \cdot a \cdot v = -230 \cdot (-0,052 - 0,014v - 0,001v^2) \cdot v = 11,96v + 3,22v^2 + 0,23v^3. \quad (40)$$

De grafiek bij deze vergelijking ziet er als volgt uit.



Figuur 23: Vermogen tegen snelheid.

Als we dit vergelijken met figuur (2) is te zien dat ons model overeenkomt met de theorie.

12 Gevoeligheidsanalyse

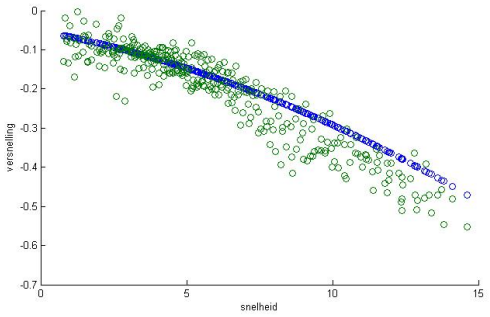
In dit hoofdstuk testen we de gevoeligheid van de gekozen methode voor de hoeveelheid data. Hiervoor bekijken we twee verschillende situaties. In het eerste onderdeel wordt een deel van de data weggefilterd, waarbij de data verspreid ligt over de totale rit. Vervolgens bekijken we hoe goed het model op deze data past. In het tweede onderdeel bekijken we een klein onderdeel van een rit. Hiermee testen we hoe goed een kleiner interval het totale verband kan voorspellen.

12.1 20% van data wegfilteren

In dit onderdeel wordt bekeken hoe goed het model, bepaald op 80% van de data, past op een deel van de data die niet is meegenomen om het model te bepalen. Dit zegt iets over de overdraagbaarheid van het model. Deze 20% van de data kan namelijk een extra rit representeren. Deze rit is dan niet gebruikt om het model te bepalen, maar het model moet natuurlijk ook voor deze rit goed kunnen voorspellen.
Om deze test uit te voeren, is allereerst 20% data weggefilterd. Dit is gedaan door steeds de vijfde meting apart te nemen. Zo krijgen we 20% van de data, verspreid over de hele dataset. Vervolgens wordt ons model bepaald op de resterende data. Dit geeft het model dat bepaald is in het hoofdstuk 11, namelijk:

$$a(v) = -0,052 - 0,014v - 0,001v^2.$$

Vervolgens kijken we hoe goed dit model op de gefilterde data past. Dit bekijken we eerst in een figuur. We zien dat de vorm van de data gerepresenteerd wordt door het model. Wel kunnen



Figuur 24: Gevonden verband op gefilterde data

we zien dat naarmate de snelheid hoger wordt, de datapunten veelal onder het gevonden model liggen. Een volgende stap in de evaluatie is het berekenen van de adjusted R square. Dit doen we met behulp van Matlab. De functie die we hiervoor gebruiken ziet er als volgt uit.

```
function ARsq = ARsq(v,a,d)
versnellingfit=versnelling(v);
yresid=a-versnellingfit;
SSresid=sum(yresid.^2);
SStotal=(length(a)-1)*var(a);
rsq=1-SSresid/SStotal;
ARsq=(length(a)-1)/(length(a)-d-1)*rsq;
```

Deze functie berekend allereerst de residuen, het verschil tussen de gemeten waarde en de voorspelde waarde. Dit gebeurt met behulp van de functie versnelling(v), waar het gevonden verband door SPSS ingevoerd is. Vervolgens worden de residual sum of squares en de total sum of squares berekend. Met behulp van deze waardes kan de adjusted R square berekend worden. Als we onze gegeven invoeren in het programma krijgen we een adjusted R square van $0,839$.

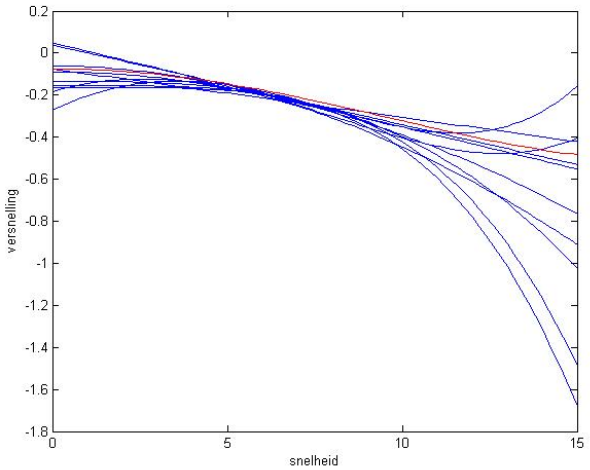
12.2 Klein interval

Als tweede test kijken we hoeveel invloed het begin en het eind van de uitroltest heeft. Dit doen we door steeds een deel data van bepaalde grootte uit het midden te nemen, en hier een model op te baseren. Vervolgens vergelijken we dit model met het gevonden model (39).
De data bestaat uit alle gegevens van 6 uitroltests. Deze uitroltests hebben echter niet allen even lang geduurd. Hierdoor is het "midden" lastig aan te duiden. Om het probleem te versimpelen nemen we als totale data de data vanaf tijdstip 0 tot het tijdstip waarbij het eerste ritje ophoudt. Dit is na 40 seconden. Aangezien er per $0,2$ seconden een meting gedaan wordt, levert dit 200 datapunten per ritje op. Het midden zit nu bij 20 seconden. Om dit midden nemen we vervolgens intervallen van $n \cdot 10\%$, waarbij we $n \in [1, .., 10]$ nemen. Op dit interval doen we een lineaire regressieanalyse zoals beschreven staat in Appendix 1. Als we in SPSS als methode om variabelen in te voeren de methode Stepwise kiezen, krijgen we de volgende resultaten.

Tabel 10: Model gevonden met Stepwise methode

Interval	Model	adjusted R square
10%	$-0,076 - 0,023v$	0,046
20%	$-0,155 - 3,016 \cdot 10^{-5}v^4$	0,298
30%	$-0,162 - 2,616 \cdot 10^{-5}v^4$	0,431
40%	$-0,134 - 2,640 \cdot 10^{-4}v^3$	0,590
50%	$-0,090 - 0,003v^2$	0,662
60%	$0,048 - 0,040v$	0,727
70%	$-0,271 + 0,080v - 0,012v^2 + 3,189 \cdot 10^{-5}v^4$	0,795
80%	$-0,183 + 0,046v - 0,009v^2 + 2,202 \cdot 10^{-5}v^4$	0,836
90%	$0,039 - 0,038v$	0,864
100%	$-0,061 - 0,004v^2 + 1,500 \cdot 10^{-5}v^3$	0,895

Deze modellen zijn allemaal van enigszins andere vorm, waardoor ze lastig te vergelijken zijn. Dat is ook te zien aan de grafieken. Op het middenstuk komen de grafieken overeen, maar in de uiteindes zijn er grote verschillen.



Figuur 25: Model gevonden met stepwise methode

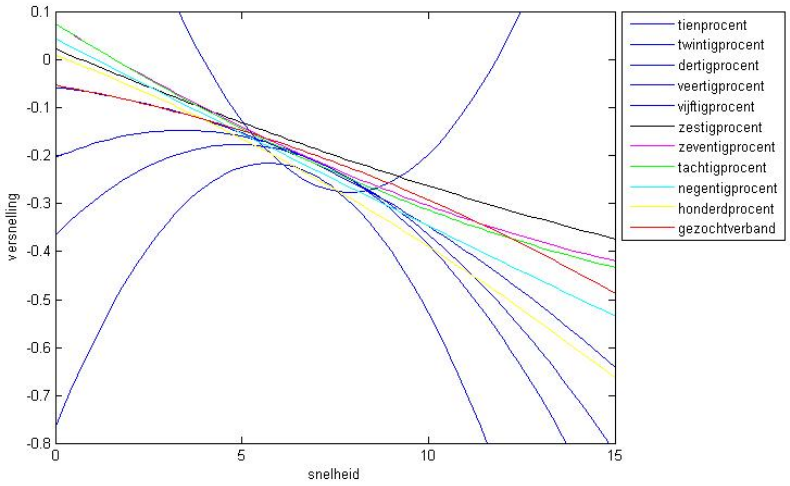
Aangezien we in hoofdstuk 11 gekozen hebben voor een model die aan de theoretische vergelijking (10) voldoet, passen we dat hier ook toe. We voeren op elk interval opnieuw een lineaire regressieanalyse uit, maar nu met de methode Enter. Hierdoor dwingen we SPSS de variabele v

en v^2 in het model te plaatsen en vergelijken we de resultaten met verband (10) Dit leidt tot de volgende resultaten.

Tabel 11: Model gevonden met enter methode

Interval	Model	adjusted R square
10%	$0,844 - 0,284v + 0,018v^2$	0,046
20%	$-0,768 + 0,194v - 0,017v^2$	0,299
30%	$-0,366 + 0,078v - 0,008v^2$	0,431
40%	$-0,204 + 0,034v - 0,005v^2$	0,591
50%	$-0,057 - 0,009v - 0,002v^2$	0,661
60%	$0,024 - 0,033v + 4,340 \cdot 10^{-4}v^2$	0,727
70%	$0,076 - 0,048v + 0,001v^2$	0,785
80%	$0,077 - 0,049v + 0,001v^2$	0,827
90%	$0,045 - 0,040v + 9,023 \cdot 10^{-5}v^2$	0,864
100%	$0,011 - 0,030v - 0,001v^2$	0,893

Deze modellen worden ook weergeven in de volgende grafiek.



Figuur 26: Model gevonden met enter methode

In deze grafiek is te zien dat er nog steeds geen duidelijke conclusies te trekken zijn. Het model bij 100% van de data (de eerste 40 seconden van de 6 ritten) komt weinig overeen met model (39) gevonden bij alle beschikbare data. Dit is ook in de grafiek te zien: de gele kromme begint boven de rode kromme, maar eindigt er onder. Het is dus erg van belang om data in het begin en aan het eind mee te nemen, voor een realistisch verband. Dit kan bevestigd worden door naar de lagere percentages te kijken. De functies bij 10% tot 50% van de data, komen enkel bij de top in de buurt van het gevonden model. Deze functies hebben een grotere kromming dan de functies bij hogere percentages.

13 Conclusies en aanbevelingen

In dit hoofdstuk bespreken we de conclusies die we kunnen trekken op basis van het onderzoek dat besproken wordt in dit verslag. Ook bespreken we een aantal onderwerpen waar in een vervolgstudie verder op ingegaan kan worden. Deze aanbevelingen zullen de resultaten verbeteren.

13.1 Conclusies

De hoofdvraag van ons onderzoek luidt als volgt.

Bepaal het verband tussen het werkelijk verbruikte vermogen en de snelheid.

We krijgen echter geen meetdata van de zonne-auto, de wagen waar we eigenlijk het verband voor willen weten. Het gaat dus niet zozeer om het verband dat we vinden, maar om de methode die we gebruiken om het verband te bepalen.

Deze hoofdvraag hebben we opgedeeld in drie deelproblemen.

1. Gegeven de snelheden, hoe bepalen we de aangepaste snelheden, zodat deze snelheden uitgezet tegen de tijd een glad verloop hebben en daardoor goed bruikbaar zijn voor de volgende stap.
2. Gegeven de (gladde) snelheden, hoe vinden we de versnellingen.
3. Gegeven de snelheden en versnellingen, hoe vinden we het verband dat de relatie tussen de versnelling en de snelheid omschrijft.

De eerste twee deelvragen zijn opgelost met de Savitzky-Golay methode. Met behulp van deze methode verkrijgen we in één stap versnellingen die goed bruikbaar zijn. De derde deelvraag wordt beantwoord met de literatuur en een lineaire regressieanalyse. Door de literatuur te bestuderen hebben we bepaald dat het verband een polynoom moet worden. Vervolgens wordt met een lineaire regressieanalyse de maximale graad en de coëfficiënten van het polynoom bepaald. Deze twee methodes maken samen onze methode die een verband kan bepalen tussen het werkelijk verbruikte vermogen en de snelheid.

Onze methode toegepast op dataset 2 van de mockup geeft het volgende resultaat:

$$P(v) = 11,96v + 3,22v^2 + 0,23v^3$$

De vorm van het verband komt overeen met vergelijking (10), het verband dat volgens de theorie moet gelden. De adjusted R square bij dit model is $0,843$. Dit betekent dat $84,3\%$ van het gedrag van de versnelling verklaard wordt door de verschillen in de snelheid. Dit geldt echter alleen voor snelheden die binnen het gemeten interval liggen. Het is lastiger te zeggen hoe goed het model is voor snelheden die buiten de gemeten waardes vallen. Het model is gebaseerd op uitroltests, waarbij de hoogste snelheid rond de 16 m/s ligt. Het Solar Team wil tijdens de race hogere snelheden rijden, maar we kunnen niet controleren of het model voor hogere snelheden accuraat is. Hier zijn namelijk geen metingen voor gedaan.

Als de methode toegepast wordt op dataset 2 krijgen we een model dat voldoet aan de theoretische vergelijking (10). Onze methode kan echter ook verbanden opleveren die niet aan de theorie voldoen. Dit hebben we gezien in hoofdstuk 9, waar we een verband vinden op dataset 1. In hoofdstuk 2 bespreken we dat het Solar Team verwacht dat een verband voor de zonne-auto de theorie niet zal volgen. Deze verwachting kan onderzocht worden door onze methode toe te passen op data van de zonne-auto.

De gevoeligheid van de methode voor de hoeveelheid data is getest. Hier zijn twee belangrijke resultaten uitgekomen. Er is getest hoe goed het model voorspellingen doet op waarden van de snelheid buiten de waarden waar het model op bepaald is. We hebben gezien dat het van belang is het interval zo groot mogelijk te kiezen om een betrouwbaar verband te krijgen. Ook is getest hoeveel invloed het heeft als 20% van de data weggehaald wordt, en hoe overdraagbaar het model is. Hieruit blijkt dat op 80% van de data nog steeds een goed model bepaald wordt. Dit model past ook op de 20% data die niet gebruikt is om het model te bepalen. Als er een nieuwe dataset gemaakt wordt kunnen we aan de hand hiervan het volgende adviseren. Het is van belang dat een zo groot mogelijk interval aan snelheden meegenomen wordt. Het aantal ritjes is minder van belang, zes ritten is meer dan genoeg om een geldig verband te bepalen.

Tot slot bespreken we het grote verschil tussen dataset 1 en dataset 2. In hoofdstuk 9 wordt onze methode toegepast op dataset 1. Dit geeft een resultaat dat sterk verschilt van het model gevonden met dataset 2. Dit laat het belang van de testomgeving zien. Bij het maken van de eerste dataset waren de omstandigheden erg ongunstig. De weg was kort en er waren veel omgevingsfactoren die invloed hebben op het uiteindelijke verband. Bij het maken van de tweede dataset waren een aantal factoren verminderd. Doordat de weg langer was en er minder omgevingsfactoren van invloed waren, geeft dataset twee een verband wat meer in het algemeen zal gelden.

13.2 Aanbevelingen

Om een gevonden verband zo goed mogelijk te maken, kunnen we aanbevelingen doen tot verbetering van de methode. Hieronder bespreken we deze aanbevelingen en leggen we kort uit waarom dit tot verbetering zal leiden.

Ten eerste raden we aan een manier te zoeken om de windsnelheid te meten. Voor ons is het niet mogelijk de windsnelheid mee te nemen in het model omdat er geen meetdata voor beschikbaar is. De windsnelheid is echter naast de snelheid van de wagen de belangrijkste variabele die de versnelling kan voorspellen. Als de windsnelheid wel als variabele meegenomen wordt in het model, zal het verband verbeteren.

Een tweede advies is metingen doen bij hogere snelheden. In onze gevoeligheidsanalyse is het belang hiervan gebleken. Het blijkt dat een verband voor een bepaald domein van de snelheid niet goed overeenkomt op een verband voor een veel groter domein van de snelheid. Hoe groter de variatie in snelheid is, hoe beter het model wordt.

Een derde advies is het doen van metingen in de omgeving waar de race wordt verreden. De omstandigheden zullen dan het beste overeenkomen met de omstandigheden die het Solar Team in de race zal aantreffen. De metingen zullen het meest betrouwbaar zijn.

Bibliografie

[Administration, 2012] Administration, U. C. D. T. (2012). Nist/sematech e-handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/.

[Albers, 2010] Albers, W. (2009-2010). *Dictaat van het vak Wiskundige Statistiek*, chapter 3.2 Schatters en criteria, pages III.3–III.9. Universiteit Twente.

[Challenge, 2013] Challenge, W. S. (2013). World solar challenge. http://www.worldsolarchallenge.org/.

[Faraway, 2002] Faraway, J. (2002). Practical regression and anova using r. http://csyue.nccu.edu.tw/Practical%20Regression%20and%20Anova%20using%20R.pdf.

[Guzzella and Sciarretta, 2005] Guzzella, L. and Sciarretta, A. (2005). *Vehicle Propulsion Systems, Introduction to Modeling and Optimization*, chapter 2 Vehicle Energy and Fuel Consumption Basic Concepts, pages 13–16. Springer, Berlin.

[Hansen et al., 2012] Hansen, P., Pereyra, V., and Scherer, G. (2012). *Least Squares Data Fitting with Applications*, chapter 2 The Linear Least Squares Estimation. Johns Hopkins University Press.

[Kariya and Kurata, 2004] Kariya, T. and Kurata, H. (2004). *Generalized Least Squares*, chapter 2 Generalized Least Squares Estimators. John Wiley & Sons, Ltd.

[Kulakowski, 1994] Kulakowski, B. T. (1994). *Vehicle-road interaction*, chapter Rolling Resistance Characteristics of New Zealand Road Surfaces, pages 249–252. ASTM, Philadelphia.

[Margenau and Murphy, 1943] Margenau, H. and Murphy, G. (1943). *The Mathematics of Physics and Chemistry*. Princeton, D. Van Nostrand.

[Seber, 1997] Seber, G. A. F. (1997). *Linear Regression Analysis*, chapter 3 Properties of Least Squares Estimates and 8 Polynomial Regression, pages 49 and 214–217. John Wiley & Sons, Inc., New York.

[Seber and Lee, 2003] Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, chapter 1.3 Linear Regression Models, page 4. John Wiley & Sons, Inc., Hoboken, New Jersey.

[Siero et al., 2009] Siero, F., Huisman, M., and Kiers, H. (2009). *Voortgezette regressie- en variantieanalyse*, chapter 3 Assumpties en generalisatie, pages 47–75. Bohn Stafleu van Loghum.

[van Dalen, 2010] van Dalen, A. (2010). Energy requirements of cycling. http://www.avdweb.nl/solar-bike/energy-requirements-of-cycling.html.

[van de Geer, 2005] van de Geer, S. (2005). *Encyclopedia of Statistics in Behavioral Science, Volume 2*, chapter Least Squares Estimation, pages 1041–1045. John Wiley & Sons, Ltd, Chichester.

[Wikipedia, 2012] Wikipedia (2012). Nuna 6. http://nl.wikipedia.org/wiki/Nuna_6.

[Wilcox, 2005] Wilcox, R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, chapter 10.2 Theil-Sen Estimator, pages 424–427. Elsevier Science.

Appendices

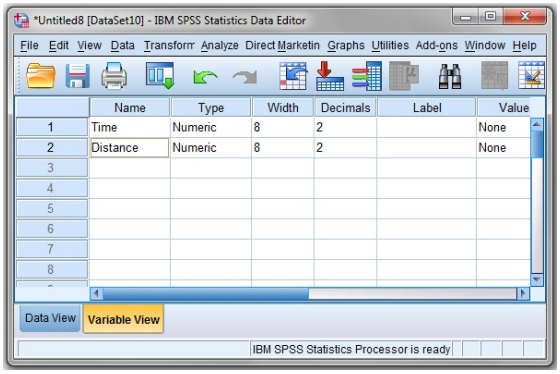
In het verslag is naar voren gekomen dat we gebruik maken van het statistische computerprogramma SPSS om met lineaire regressie het verband te vinden wat we willen. Uit gesprekken met de opdrachtgevers uit het Solar Team is gebleken dat zij niet bekend zijn met het programma SPSS. Zij hebben ons gevraagd een "handleiding"te schrijven zodat het Solar Team inzicht krijgt in wat het programma kan, hoe wij het programma gebruiken en hoe output geanalyseerd moet worden. Hieronder zal stap voor stap worden uitgelegd hoe we het programma hebben gebruikt en hoe we tot ons model zijn gekomen.

1 Lineaire regressie met SPSS

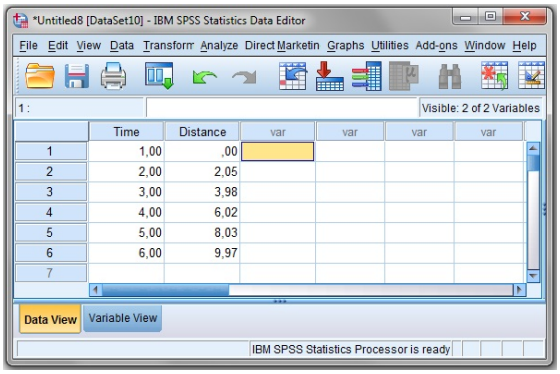
Het programma SPSS is een statistisch computerprogramma. SPSS kan data lezen, bewerken en visualiseren. Met SPSS kan een lineaire regressieanalyse gedaan worden en kunnen de geschatte waardes voor de parameters bepaald worden. Als SPSS geopend wordt is er een een leeg Datafile te zien. Links onderin zijn twee knoppen, Data View en Variable View. Dit zijn twee manieren om de data te bekijken. SPSS werkt met twee schermen, het datascherm dat te zien is bij openen, en het outputscherm.

1.1 Data invoeren

De data waar mee gewerkt moet worden, wordt door het Solar Team aangeleverd in een Excel bestand. Dit Excel bestand is als volgt opgebouwd. In de eerste rij staat de naam van de variabele genoemd. Onder elke naam zijn in een kolom alle metingen genoemd. Metingen van verschillende variabelen in dezelfde rij horen bij elkaar. In SPSS kan dit Excel bestand geïmporteerd worden. Hiervoor wordt SPSS geopend en klikken we het openingsscherm weg. We hebben nu een leeg Datafile voor ons. Vervolgens gaan we naar File → Open → Data. In het nieuwe schermpje selecteren we als bestandstype Excel (*.xls, *.xlsx, *.xlsm) en zoeken we het gewenste bestand. Er verschijnt een nieuw scherm, aan deze instellingen veranderen we niks, we klikken op OK. Nu is het Excel bestand geïmporteerd in SPSS. Het is verstandig om te controleren of alles goed is overgezet, zowel in de Data View als in de Variable View. In de Data View zien we de namen van de variabelen en alle getallen staan. In de Variable View is het belangrijk om te kijken naar Type, Width en Decimals. Als de variabelen van het type Numeric zijn en de grootte en het aantal decimalen klopt met wat je wil, dan is de data goed ingevoerd.



(a) Invoeren data

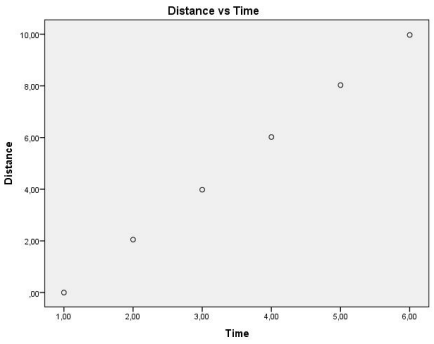


(b) Invoeren data

1.2 Te bepalen verband bekijken

Nu alle benodigde variabelen in SPSS staan, kan het te bepalen verband bekeken worden. Door de gewenste afhankelijke variabele uit te zetten tegen de gewenste verklarende variabele in een scatterplot, wordt het verband visueel gemaakt. Dit kan door naar **Graphs** → **Legacy Dialogs** → **Scatter\Dot** te gaan en vervolgens voor **Simple Scatter** te kiezen.

In het scherm dat verschijnt moeten de afhankelijke en verklarende variabele ingevoerd worden. Door aan de linkerkant de gewenste afhankelijke variabele te kiezen en vervolgens op het pijltje naast **Y Axis** te klikken, wordt deze variabele ingevoerd als afhankelijke variabele. De verklarende variabele kan op dezelfde wijze bij **X Axis** ingevoerd worden. Rechtsboven kan via de knop **Titles** een titel, ondertitel en voetnoot voor de afbeelding ingevoerd worden. Verder hoeft er niks ingevoerd te worden, druk om de plot te maken op **OK**. Als er een scatterplot gemaakt wordt van het voorbeeld in de vorige afbeeldingen ziet deze er als volgt uit. In deze plot is Distance de afhankelijke variabele, Time de verklarende variabele en is als titel 'Distance vs Time' meegegeven. Aan deze plot is te zien dat het gewenste verband waarschijnlijk lineair is.



Figuur 28: Plot

1.3 Variabelen berekenen

Als uit de scatterplot volgt dat het te bepalen verband niet lineair is, moeten er variabelen berekend worden. Als het verband bijvoorbeeld kwadratisch lijkt te zijn, kan SPSS niet zelf zorgen dat het kwadraat van de tijd wordt meegenomen bij het bepalen van het verband. Dit kwadraat van de tijd zal als nieuwe variabele ingevoerd moeten worden. Dit is te doen via **Transform** → **Compute Variable**. Linksboven bij **Target Variable** moet de gewenste naam ingevoerd worden, bijvoorbeeld t_kwadraat. Daarnaast moet de functie ingevoerd worden waarmee de nieuwe variabele berekend wordt. Dit kan met behulp van de knoppen in het midden van het scherm en de bekende variabelen die links staan. In het geval wordt dit Time * Time, dit kan zowel getypt worden als geselecteerd uit de linker kolom en de middenknoppen. Als tot slot op **OK** gedrukt wordt, verschijnt de nieuwe variabele in de datafile.

1.4 Lineaire regressieanalyse uitvoeren

Als alle variabelen die mogelijk in het model moeten in de datafile staan, kan een lineaire regressieanalyse uitgevoerd worden. Omdat er uitgegaan wordt van lineaire regressie, gaat dit als volgt: **Analyze** → **Regression** → **Linear**. In ons geval wordt de afhankelijke variabele ingevoerd door de gewenste variabele uit de linkerkolom te selecteren en op het pijltje naast **Dependent** te klikken. De verklarende variabelen worden op eenzelfde manier ingevoerd bij **Independent(s)**. Als het gewenst is verschillende modellen te proberen, kan een tweede set van verklarende variabelen ingevoerd worden door bij **Independent(s)** op **Next** te klikken. Onder het vak voor **Independent(s)** kan een regressiemethode gekozen worden. Er kan gekozen worden uit de opties *Enter, Stepwise, Remove, Backward* of *Forward*. De methode *Enter* zorgt ervoor dat alle variabelen gedwongen in het model komen. De methode *Forward* begint met een model dat alleen een constante bevat. Vervolgens gaat het programma op zoek naar de verklarende variabele die de grootste correlatie heeft met de afhankelijke variabele. Deze variabele wordt toegevoegd en het effect van het toevoegen wordt bepaald. Als blijkt dat de variabele geen significante verbetering van het model veroorzaakt zal de variabele niet toegevoegd worden. Dit proces wordt herhaald tot er geen variabelen meer zijn die voldoen aan de criteria om ingevoerd te worden. De methode *Backward* begint met een model waar alle variabelen in gebruikt worden. De zwakste verklarende variabele wordt vervolgens verwijderd en de regressie wordt herberekend. Als het model hier significant slechter van wordt, zal de verklarende variable heringevoerd worden. Anders blijft de variabele verwijderd. Deze procedure wordt herhaald tot er alleen nuttige verklarende variabelen in het model overblijven. De methode *Stepwise* is de meest elegante methode. Het proces werkt ongeveer hetzelfde als bij de *Forward* methode. Echter wordt de significantie van de variabele bepaald aan de hand van de F-waarde. Ook wordt er elke keer als een variabele toegevoegd is, de significantie van alle andere variabelen opnieuw bekeken. Als er een variabele tussen zit die,

na toevoeging van de nieuwe variabele, niet langer een significant belang speelt in het model, dan zal deze variabele weer worden verwijderd. Deze methode zorgt ervoor dat je zo weinig mogelijk verklarende variabelen opneemt in je model.

Als het zeker is welke variabelen in het model moeten komen, is de methode *Enter* een goede keus. Als het niet zeker is welke variabelen in het model gebruikt moeten worden, is de *Stepwise* methode een goede methode om te gebruiken. Rechtsboven in het scherm zijn nog een aantal knoppen waarmee bepaalde instellingen gekozen kunnen worden. Bij **Statistics** worden de volgende opties aangevinkt: Estimates, Confidence intervals (95%), Durbin-Watson, Casewise diagnostics (Outliers outside 3 standard deviations), Model fit, R square change. Bij **Plots** is een plot van de gestandardiseerde residuen tegen de gestandardiseerde voorspelde waarden wenselijk. Hiervoor wordt *ZRESID geselecteerd uit de linkerkolom en met het pijltje naast **Y** ingevoerd als afhankelijke variabele. *ZPRED wordt op dezelfde manier ingevoerd bij **X** als verklarende variabele. Ook worden in dit scherm de opties Histogram en Normal probability plot geselecteerd. Bij **Save** wordt alleen de Unstandardized Predicted Values aangevinkt. Hiermee worden de waardes van het bepaalde model in de datafile opgeslagen, waarmee later een grafiek gemaakt kan worden om het model de evalueren. Bij **Options** kan gekozen worden of er wel of geen constante in het model meegenomen wordt. Verder hoeft er niks aangepast te worden, door op **OK** te drukken wordt de regressieanalyse uitgevoerd.

1.5 Output analyseren

Nu alle stappen voor een lineaire regressieanalyse gevolgd zijn, verschijnt in de Outputfile een reeks gegevens. Deze gegevens zullen in hier van boven naar beneden beschreven worden. Bij elke beschrijving zal een voorbeeld van een output gegeven worden. De eerste tabel waar naar gekeken wordt is **Variables Entered/Removed**. In deze tabel wordt per stap beschreven welke variabele is ingevoerd of verwijderd. Ook wordt hierin beschreven welke methode gebruikt is en welke criteria deze methode gebruikt heeft.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Time	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: Distance

In het voorbeeld is te zien dat er maar 1 stap nodig is om tot het model te komen dat SPSS het beste vindt. In de eerste stap is de variabele Time ingevoerd.
De volgende tabel is **Model Summary**. Deze tabel geeft per stap de R, R Square, Adjusted R square, de standaardafwijking van de schatter en de veranderingen in R square en F. Het is in deze tabel belangrijk om naar R Square en de Adjusted R Square te kijken. Hier staat R voor de pearson correlatiecoëfficiënt. De R Square is een maat voor de juistheid van het model. Hoe meer deze waarde in de buurt van één ligt, hoe beter. Dit geldt ook voor de Adjusted R Square. Deze waarde is eigenlijk nog beter om naar te kijken, omdat deze waarde meeneemt hoeveel variabelen in het model gebruikt worden. Hierbij zorgt een groot aantal variabelen voor een verlaging van de Adjusted R Square.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	1,000[a]	1,000	1,000	,03267	1,000	65248,380	1	4	,000	2,933

a. Predictors: (Constant), Time

b. Dependent Variable: Distance

In het voorbeeld wordt de afhankelijke variabele Distance perfect voorspeld door de verklarende variabele Time. De R, R Square en Adjusted R Square zijn allemaal 1.
De volgende tabel wordt ANOVA genoemd. Deze kan gebruikt worden om de relevantie van variabelen te toetsen. Dit is vooral relevant als er gekozen is voor de methode Enter. Echter is hier gekozen voor een Stepwise methode, waarbij SPSS zelf de relevantie van de variabelen al onderzocht heeft. Op deze tabel wordt verder niet ingegaan.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	69,660	1	69,660	65248,380	,000[b]
	Residual	,004	4	,001		
	Total	69,665	5			

a. Dependent Variable: Distance

b. Predictors: (Constant), Time

Vervolgens komt een tabel genaamd Coefficients. In de meest linkerkolom is te zien welke variabelen gebruikt zijn in het model. Daarna komen de coëfficiënten die bij deze variabelen horen en de standaardafwijking van deze geschatte coëfficiënt. Verder is het belangrijk om naar de betrouwbaarheidsintervallen te kijken. Aangezien de coëfficiënten geschat zijn, valt niet met honderd procent zekerheid te zeggen dat bij een nieuwe meting het model dezelfde coëfficiënten zal krijgen. Wel kunnen we zeggen dat 95 van de 100 keer deze coëfficiënten uit het 95%-betrouwbaarheidsinterval komen.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1,975	,030		-64,917	,000	-2,059	-1,890
	Time	1,995	,008	1,000	255,438	,000	1,973	2,017

a. Dependent Variable: Distance

Hierna volgt nog een tabel Excluded Variables. Deze tabel omschrijft welke variabelen buiten het model gelaten zijn en welke toelatingsgrenzen gehanteerd moeten worden voordat deze variabele wel in het model komt.

In het voorbeeld is te zien dat de variabele t_kwadraat buiten het model gelaten is. Deze variabele had blijkbaar geen toegevoegde waarde meer.
De laatste tabel heet Residuals Statistics. Deze tabel geeft informatie over de voorspelde waardes

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	t_kwadraat	-,015[b]	-,720	,523	-,384	,042

a. Dependent Variable: Distance

b. Predictors in the Model: (Constant), Time

Figuur 29: Verdeling residuen

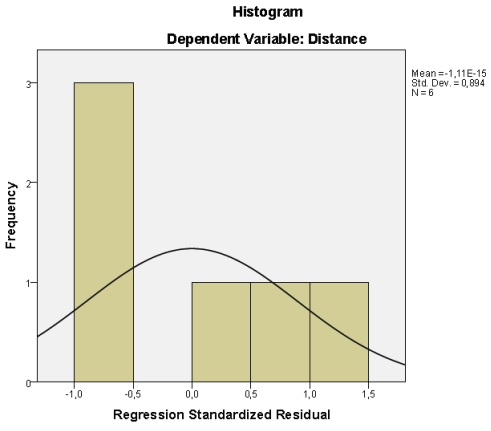
en de afwijking tussen de metingen en de voorspelde waardes. Er wordt voor nu niet verder op deze tabel ingegaan.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	,0205	9,9962	5,0083	3,73257	6
Residual	-,03076	,03438	,00000	,02922	6
Std. Predicted Value	-1,336	1,336	,000	1,000	6
Std. Residual	-,941	1,052	,000	,894	6

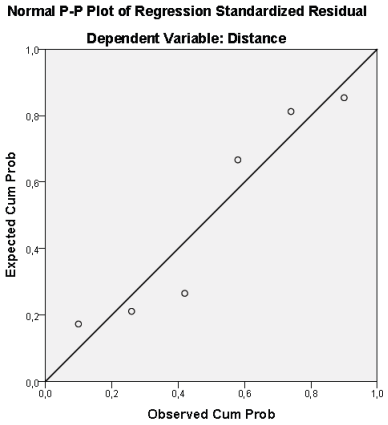
a. Dependent Variable: Distance

Na deze tabellen volgen nog een aantal grafieken. De eerste grafiek is een histogram van de Standardized Residual. Dit is de gestandardiseerde waarde van het verschil tussen de voorspelde en de gemeten waarde. In deze histogram staat ook een lijn, die de standaardnormaalverdeling voorsteld. Hoe meer de histogram deze lijn lijkt te volgen, hoe meer het aannemelijk is dat de residuen normaal verdeeld zijn. Dit is een aanname die gedaan wordt om lineaire regressie toe te passen, dus het is wenselijk dat de histogram standaard normaal verdeeld lijkt.

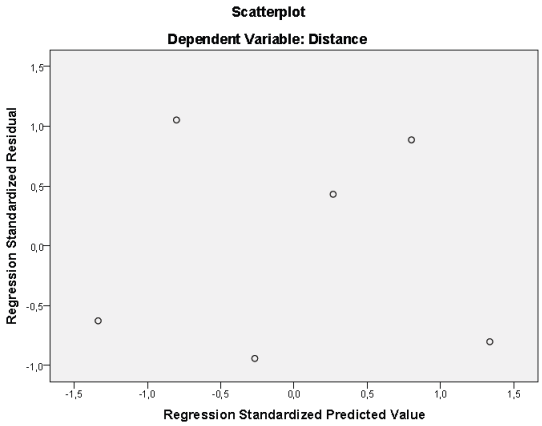


In het voorbeeld is te zien dat de histogram niet echt lijkt op de lijn die een normaalverdeling weergeeft. In dit geval komt dit vooral omdat er erg weinig metingen zijn waar het model op gebaseerd is. Er zijn immers maar zes punten gemeten. Als er echter bij een groot aantal metingen nog steeds een significante afwijking is van de lijn, dan moet de aanname van normaal verdeelde residuen heroverwogen worden.

De volgende afbeelding laat de afwijking van de metingen zien ten opzichte van het bepaalde model. Het bepaalde model is afgebeeld als een rechte lijn. De meetpunten worden daaromheen geplaatst. Hoe meer de meetpunten de rechte lijn volgen, hoe beter het model de meetpunten voorspeld. Deze plot kan gebruikt worden om te zien of de afwijkingen goed verdeeld zijn, of dat er juist op een bepaald punt een grote afwijking is.



In het voorbeeld zijn erg weinig punten om het model echt te evalueren. Wel is te zien dat de punten allemaal ongeveer even ver van de lijn liggen. Er is dus niet een bepaald punt die een erg grote afwijking heeft en misschien wel verkeerd gemeten is.
De laatste grafiek is een scatterplot van de gestandardiseerde residuen tegen de gestandardiseerde voorspelde waardes. In deze afbeeldingen moeten de punten chaos vertonen. Als er in de verdeling van de punten een bepaald patroon te zien is, geeft dit aan dat er nog een onverklaard element in de data zit. Als de punten een chaotische verdeling hebben, betekend dit dat de grootte van de afwijking niet afhangt van waar deze afwijking gevonden is.



In deze plot zijn erg weinig punten om echt te beoordelen of het chaotisch verdeeld is. We zien echter niet gelijk een patroon in de verdeling van de punten. Dit lijkt dus een goed model te zijn.

1.6 Syntax

Om het gebruik van SPSS eenvoudiger te maken kan er een syntax gemaakt worden. Deze onthoudt de stappen die er uitgevoerd moeten worden om met lineaire regressie een passend verband te vinden. Het bepalen van een verband kan dan in een stappenplan samengevat worden:

1. De data invoeren zoals hierboven beschreven 1.

2. De Syntax openen. Met behulp van de knoppen: **Run** → **All** zal Spss een lineaire regressieanalyse uitvoeren. Er verschijnt een Outputfile met een reeks gegevens.

3. Vervolgens kan zoals in de paragraaf Output analyseren 1 de resultaten bekeken worden. Belangrijk is dat er bepaald wordt of er aan de aannames voldaan wordt en of er dus conclusies getrokken kunnen worden.

2 Savitzky Golay in Matlab

In Methode 4 wordt gebruik gemaakt van een Savitzky-Golay filter om datapunten voor de versnelling te verkrijgen. Dit filter wordt toegepast met behulp van Matlab. In de standaard Matlabfuncties zit een functie voor de Savitzky-Golay smoothing filter. Deze functie kan echter alleen een signaal glad maken, geen afgeleides berekenen. Om dit toch te kunnen doen met Matlab is er een programma gedownload van de File Exchange pagina van de Matlab website. Dit programma is ingediend door `http://www.mathworks.com/matlabcentral/fileexchange/authors/62607`. Het programma heeft goede commentaren en ratings en is door ons getest. Na deze evaluatie is aangenomen dat het programma goed werkt.

2.1 Het programma

Het programma bestaat uit twee files. De eerste file heet savitzkyGolay.m. Deze functie doet nagenoeg hetzelfde als de Matlabfunctie sgolay. Echter zijn bij deze functie enkele restricties voor de data opgelost en werkt de functie sneller. Het programma bepaald de coëfficiënten die de Savitzky-Golay filter gebruikt.
De tweede file heet savitzkyGolayFilt.m. Deze functie filtert de input met behulp van de coëfficiënten die berekend worden door de andere file. De functie kan zowel gesmoothe punten geven als punten voor de afgeleides.

2.2 Gebruik van het programma

Om gebruik te kunnen maken van het programma moet eerst de data goed ingelezen worden in Matlab. Voor elk ritje kan een vector V_k gemaakt worden waar de snelheden in staan. Het is hiervan bij belang dat de metingen op volgorde staan, aangezien deze volgorde gebruikt wordt om te smoothen danwel de versnelling te berekenen.

$$V_k = \begin{pmatrix} v(1) \\ \vdots \\ v(n) \end{pmatrix} \tag{41}$$

Vervolgens kan op deze vector de functie savitzkyGolayFilt(X,N,DN,F) gebruikt worden. In deze functie staat X voor de vector die behandeld wordt, in ons geval geldt dus X=V_k. De variabele N staat voor de orde van het polynoom die gebruikt wordt. Zoals beschreven is in hoofdstuk 5.2 wordt N=1 gebruikt. De variabele DN staat voor de orde van differentiatie. Aangezien de eerste afgeleide bekeken moet worden, geldt DN=1. Tot slot staat de variabele F voor de grootte van het interval. Deze moet oneven zijn. Wij kiezen voor F=7. Als al deze waarden in de functie ingevuld worden wordt er een output gegeven. Dit is echter nog niet onze versnelling. Als er gewerkt wordt met afgeleides moet deze functie namelijk geschaald worden met T^{-DN} waarbij T staat voor de grootte van het meetinterval. In ons geval worden metingen verkregen per 0.2 seconden. Om de vector van versnellingen van rit k (A_k) te krijgen, wordt dus het volgende ingetypt: $A_k = 0.2^{-1}\cdot$ savitzkyGolayFilt(V_k,1,1,7). Aangezien er wordt gewerkt met intervallen van zeven punten, kunnen de eerste en laatste drie punten niet goed bepaald worden. Vervolgens

moet dus een nieuwe vector met de versnellingen $\tilde{A_k}$ gemaakt worden, waar deze punten uitgelaten worden.

$$A_k = \begin{pmatrix} a(1) \\ \vdots \\ a(n) \end{pmatrix}, \tilde{A_k} = \begin{pmatrix} a(4) \\ \vdots \\ a(n-3) \end{pmatrix} \tag{42}$$

Dit moet ook gedaan worden voor de vector V_k, aangezien deze twee vectoren vervolgens samen in Excel gezet moeten worden en dus even lang moeten zijn.
Deze procedure kan herhaald worden voor elk ritje. Vervolgens kan alle data naar Excel gelezen worden, zodat er vervolgd kan worden met lineaire regressie.

De hierboven beschreven manier om met Matlab de Savitzky-Golay methode toe te passen, zodat datapunten bruikbaar worden gemaakt om een goed verband voor te vinden, is in een stappenplan weer te geven. We nemen hierbij als voorbeeld dataset 2.

3 Handleiding

In dit hoofdstuk zal kort onder elkaar gezet worden welke stappen er ondernomen moeten worden om een verband voor het vermogen afhankelijk van de snelheid te vinden. Hiervoor moet allereerst de aangeleverde data in beschouwing worden genomen. Mochten er metingen gedaan zijn met twee MCU's, dan adviseren we om naar de gemiddelde snelheid te kijken. Houdt vervolgens het onderstaande stappenplan aan. Dit stappenplan is specifiek voor één uitroltest van bedoeld.

1. Open Matlab.
2. Klik op File → Import Data... en zoek de het betreffende Excelbestand.
3. Selecteer de kolom waarin de gemiddelde snelheid staat vanaf tijdstip 0 en noem deze V. Klik vervolgens op Import.
4. Voer vervolgens in het Command Window het commando: $A = 0.2\text{^}(-1)*savitzkyGolayFilt(V, 1, 1, 7);$ in.
5. Verwaarloos de eerste en de laatste 3 punten van A en V en maak een matrix van A en van V.
6. Exporteer deze matrix naar een Excelbestand.
7. Open Spss en het hierboven genoemde Excelbestand.
8. Klik op het tabblad Variable View en voeg bij Name op rij 1 V en op rij 2 A toe en verhoog het aantal decimalen naar 3.
9. Klik op het tabblad Data View en plak hier de kolommen uit het Excelbestand.
10. Open de Syntax met behulp van de knoppen: File → Open → Syntax...
11. Klik in de Syntax op de knoppen: Run → All.
12. Analyseer de output aan de hand van hoofdstuk 1.

Het is ook mogelijk om meerdere uitroltests tegelijkertijd te analyseren. Hiervoor volgt bijna eenzelfde stappenplan. In Matlab moet er enkel voor gezorgd worden dat er van alle snelheden één vector wordt gevormd en de bijhorende versnellingen er in een vector naast geplaatst worden, waarbij weer de eerste en de laatste drie punten verwaarloosd zijn. De vervolgstappen met SPSS worden op dezelfde manier gedaan.

Bedenk dat er volgens het bovenstaande stappenplan een verband voor versnelling en de snelheid gevonden wordt. Om een verband voor het vermogen en de snelheid te vinden, moeten de formules uit het hoofdstuk 2 gebruikt worden.