KU LEUVEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

Valpartijen in eendagskoersen in het professionele wielrennen

Onderzoek naar de factoren die bepalend zijn voor het aantal valpartijen tijdens een eendagswedstrijd in het professionele wielrennen

Astrid Buttiens

R0258162

Masterproef aangeboden tot het behalen van de graad

MASTER IN DE TOEGEPASTE ECONOMISCHE WETENSCHAPPEN: HANDELSINGENIEUR
Major Kwantitatieve Methoden

Promotor: Prof. Dr. Gerda Claeskens
Assistent: Roel Verbelen

Academiejaar 2014-2015

KU LEUVEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN

Valpartijen in eendagskoersen in het professionele wielrennen

Onderzoek naar de factoren die bepalend zijn voor het aantal valpartijen tijdens een eendagswedstrijd in het professionele wielrennen

Deze masterproef gaat op zoek naar de factoren die het gemiddeld aantal valpartijen in een eendagswedstrijd in het professionele wielrennen bepalen. De literatuurstudie bij dit onderzoek toont aan dat dit onderwerp slechts beperkt aan bod kwam in eerder wetenschappelijk onderzoek. De resultaten van de verkennende analyses en de Poisson-regressiemodellen die in dit onderzoek ontwikkeld worden, leiden tot twee belangrijke conclusies. Het gebruik van oortjes blijkt geen meerwaarde te bieden bij het voorspellen van het gemiddeld aantal valpartijen per wedstrijd. Het jaar waarin een wedstrijd gereden wordt, de categorie van een wedstrijd, de lengte van het wedstrijdparcours, de gemiddelde snelheid van de winnaar, de hoeveelheid finishers, het aan- of afwezig zijn van kasseien, de hoeveelheid hellingen in het parcours en de kijkcijfers van de overeenkomstige tv-uitzending doen dat wel. Dit onderzoek is een aanzet tot verder onderzoek naar het ontstaan en de oorzaken van valpartijen tijdens een wielerwedstrijd. De ontwikkelde Poisson-regressiemodellen kunnen immers enkel een verband verduidelijken, maar hier geen verklaring voor bieden.

Astrid Buttiens

R0258162

Masterproef aangeboden tot het behalen van de graad

MASTER IN DE TOEGEPASTE ECONOMISCHE WETENSCHAPPEN: HANDELSINGENIEUR
Major Kwantitatieve Methoden

Promotor: Prof. Dr. Gerda Claeskens
Assistent: Roel Verbelen

Academiejaar 2014-2015

Dankwoord

Een student die zijn of haar masterproef afrondt, is als een renner die een mooie overwinning boekt. Zo'n overwinning is meer dan het resultaat van één enkele persoon. Het vloeit voor uit de prestaties van een team. Daarom wil ik in dit dankwoord graag iedereen bedanken een bijdrage leverde bij het verwezenlijken van deze masterproef.

In de eerste plaats wil ik graag mijn promotor, Prof. Dr. Gerda Claeskens, bedanken. Zij gaf me de kans om dit eigen onderwerp uit te werken. Een heel jaar lang mocht ik me toeleggen op een studie over de mooiste sport ter wereld, wat van mijn masterproef een zeer fijne ervaring maakte.

Ik wil ook graag mijn werkleider, Roel Verbelen, bedanken voor het begeleiden van deze masterproef. Zijn opmerkingen en aanmoedigingen waren onmisbaar bij het schrijven van deze tekst. Ik wil hem dan ook in het bijzonder bedanken voor alle tijd en moeite die hij investeerde in het brainstormen over dit onderzoek en het nalezen van mijn teksten.

Een speciale dank gaat uit naar mijn ouders. Zij vertolken de rol van de toegewijde ploegleiders die achter hun *poulain* aan rijden om hem met raad en daad bij te staan en hem door de moeilijke zones van de wedstrijd te coachen.
Vijf jaar lang zorgden zij er voor dat ik kon genieten van het zorgeloze bestaan van een student in Leuven. Tot op de dag van vandaag is er niet één moment geweest waarop ik niet op hun onvoorwaardelijke steun kon rekenen.

Een laatste woord van dank is voor mijn zussen, broer en vrienden. Deze schare trouwe supporters stond steeds voor me klaar, gewapend met een oneindige dosis geduld, enthousiasme, steun en motivatie.

Astrid Buttiens
Mei 2015

Inhoudstafel

Algemene Inleiding

Het openingsweekend, met de Omloop het Nieuwsblad en Kuurne-Brussel-Kuurne, is de traditionele aftrap van het Belgische wielerjaar. Vanaf Milaan-Sanremo volgen de wedstrijden elkaar in sneltempo op. Haast elke keer wordt de wedstrijd ontsierd door de vele valpartijen, met vaak ernstige blessures tot gevolg. Al snel volgt dan de discussie: waarom vallen de renners toch zo vaak? De vele ploegleiders, renners en journalisten hebben elk hun eigen mening. Het peloton is te groot, er wordt te snel en/of te nerveus gereden, de oorzaak ligt bij het gebruik van de oortjes, het ligt aan de weersomstandigheden, het parcours, het materiaal, enz.

Ondanks de vele opiniestukken die het probleem van de valpartijen al aankaartten, werd nog maar weinig wetenschappelijk onderzoek gevoerd naar dit aspect van het (professionele) wielrennen. Deze masterproef wil een aanzet zijn tot het opvullen van deze leemte aan de hand van volgende onderzoeksvraag:

> **WELKE FACTOREN ZIJN BEPALEND VOOR HET AANTAL VALPARTIJEN TIJDENS EEN EENDAGSWEDSTRIJD IN HET PROFESSIONELE WIELRENNEN?**

Deze onderzoeksvraag wordt verder uitgediept aan de hand van een reeks deelvragen die elk een ander aspect van de probleemstelling belichten. Deze vragen komen aan bod in het verdere verloop van deze tekst.

Het eerste hoofdstuk van deze masterproef biedt een overzicht van de wetenschappelijke literatuur omtrent valpartijen in het wielrennen. Zoals eerder aangegeven is dit type onderzoek zeer beperkt. Daarom wordt ook gebruik gemaakt van studies die zich toeleggen op één of meerdere factoren die een verklaring kunnen bieden voor het aantal valpartijen per eendagswedstrijd. Daarnaast bespreekt deze literatuurstudie een reeks conclusies die voortvloeien uit de analyse van fietsers in het dagelijks verkeer, niet-professionele renners en sporttakken die gemeenschappelijke kenmerken vertonen met het wielrennen. In het laatste deel van dit hoofdstuk komt de doelgroep van deze masterproef aan bod. Niet alleen de renners hebben baat bij een veilige wielersport, maar ook de internationale wielerunie en haar verzekeringspartners hebben mogelijk interesse in de antwoorden op deze onderzoeksvraag.

Het karakter van de wielersport bemoeilijkt het verzamelen van data. Het samenbrengen van de gegevens voor deze analyse is dan ook een eigen werk. De manier waarop dit gebeurde, alsook de bronnen die hiervoor geraadpleegd werden, worden besproken in het tweede hoofdstuk.
Het eerste deel van dit hoofdstuk beschrijft de zestien Europese eendagswedstrijden die in de analyse opgenomen werden. Daarna volgt een overzicht van de competitiemodellen die gebruikt werden tussen 1997 en 2014. Deze hebben immers een invloed op de manier waarop de verschillende wedstrijden gecatalogeerd werden. Het derde deel behandelt de overige wedstrijdkenmerken die in de analyse aan bod komen. Hier wordt ook de betekenis van de verschillende variabelen uitgelegd die in het verdere verloop van deze masterproef aan bod komen.

In het derde hoofdstuk worden de univariate analyses van de verschillende variabelen besproken. De numerieke variabelen worden bestudeerd aan de hand van hun histogram en belangrijkste kengetallen. Daarnaast wordt ook de relatie van elke variabele tot de responsvariabele bestudeerd. Dit gebeurt aan de hand van meerdere matrix-scatterplots, die steeds aangevuld worden met twee trendlijnen. Een eerste trendlijn schat het lineaire verband tussen de twee variabelen. Deze groene rechte heeft vaak een slechte fit met de data, daarom wordt steeds een tweede curve toegevoegd. Deze rode trendlijn is het resultaat van een niet-parametrische regressie, die minder strenge assumpties maakt. Op die manier wordt een meer verfijnd beeld van de relatie van elke variabele tot het gemiddeld aantal valpartijen per wedstrijd verkregen.
De categorische variabelen worden besproken aan de hand van verschillende boxplots. Deze grafieken bieden een overzicht van de belangrijkste kengetallen van de te verklaren variabele voor elk niveau van de bestudeerde factorvariabele. Deze figuren worden aangevuld met de nodige significantietesten.

Het vierde hoofdstuk van deze tekst stelt vijf Poisson-regressiemodellen voor die het gemiddeld aantal valpartijen per wedstrijd trachten te voorspellen. Deze modellen vormen de basis voor het beantwoorden van de tien deelvragen van de onderzoeksvraag.
Eerst gaat dit hoofdstuk dieper in op wat een Poisson-regressiemodel precies is, hoe de parameters van dergelijk model geschat worden en op welke manier de kwaliteit van de voorgestelde modellen beoordeeld wordt.
Vervolgens wordt een additief regressiemodel ontwikkeld. Dit model bevat alle variabelen die in de analyse aan bod komen en geldt als referentie voor de andere modellen die besproken worden. Dit model wordt immers stapsgewijs verbeterd door de AIC-waarde van het model te minimaliseren. In de vierde sectie wordt de voorspellingsfout van het regressiemodel naar beneden gebracht, zodat een model gecreëerd wordt dat het gemiddeld aantal valpartijen per wedstrijd zo goed mogelijk kan voorspellen.
Bij elk van de modellen hoort een bespreking van de beoordelingscriteria van het model, alsook van de waarden van de geschatte parameters.

Het vijfde en laatste deel biedt een overzicht van de conclusies van deze masterproef. Deze besluiten zijn gebaseerd op de resultaten van de vijf Poisson-regressiemodellen en de verkennende analyses en trachten een antwoord te formuleren op de verschillende deelvragen van dit onderzoek. Omdat de Poisson-regressiemodellen enkel het verband tussen variabelen kunnen weergeven, kan echter geen uitspraak gedaan worden over de causaliteit van de valpartijenproblematiek.
Tot slot biedt dit afsluitend hoofdstuk een overzicht van de mogelijke onderzoeksvragen die behandeld kunnen worden in toekomstig onderzoek.

1 Literatuurstudie

1.1 Overzicht van de literatuur

Dit eerste hoofdstuk biedt een overzicht van de beschikbare literatuur over valpartijen in het wielrennen en het verband met één of meer factoren die eerder aangehaald werden. Het aanbod wetenschappelijk onderzoek omtrent deze probleemstelling is echter beperkt. Daarom maakt deze tekst ook gebruik van wetenschappelijke conclusies over fietsers in het dagelijks verkeer, over fietsers in het niet-professionele wielrennen en over andere sporttakken waarin gelijkaardig onderzoek verricht werd.

1.1.1 Evolutie van het aantal valpartijen over de tijd

Valpartijen zijn inherent aan het wielrennen. Dit hoeft echter niet te betekenen dat het aantal valpartijen per wedstrijd niet kan toe- of afnemen. Het loont daarom de moeite om de evolutie van het aantal valpartijen per wedstrijd te bestuderen. Wanneer blijkt dat er de laatste jaren meer of minder gevallen werd, dan ligt de oorzaak mogelijk bij recente veranderingen. Op deze manier wordt het mogelijk de wijzigingen van reglementen, de vernieuwing en verbetering van het materiaal en de verbeterde technieken voor het begeleiden van atleten in rekening te nemen.
Dit leidt tot een eerste en belangrijke deelvraag in deze masterproef:

NEEMT HET AANTAL VALPARTIJEN TOE?

De Organisation for Economic Co-operation and Development (OECD) onderscheidt twee soorten onderzoeken over het dragen van een helm tijdens het fietsen. Een eerste type onderzoek legt het verband tussen het dragen van een helm en de ernst van een blessure of verwonding bij een valpartij. In de meeste gevallen gaat het om letsels aan het hoofd, de hersenen of het aangezicht (Thompson en Patterson, 1998). Uit de analyse van de OECD blijkt dat een helm de impact van een valpartij kan reduceren. Dit is niet onbelangrijk, want in het competitieve baan- en wegwielrennen resulteert twee tot drie procent van de valpartijen in blessures of verwondingen (McLennan et al., 1988).
De onderzoeken van de tweede soort bestuderen het veiligheids- en gezondheidsaspect van het promoten of verplichten van een fietshelm. Dergelijk onderzoek komt minder frequent voor. De OECD beperkt zich tot het bespreken van de factoren die de veiligheid en gezondheid beïnvloeden. Vooreerst reduceert het verplichten van de fietshelm het risico op blessures of verwondingen. De OECD verwijst hierbij naar het onderzoek van Elvik et al. (2009) waaruit blijkt dat het aantal hoofdletsels met 25% is afgenomen. Het dragen van een helm creëert helaas een vals gevoel van veiligheid. Fietsers passen hun gedrag aan en durven meer risico's nemen dan wanneer ze geen helm dragen. Daarnaast beïnvloeden deze campagnes de grootte en samenstelling van de populatie fietsers. Volgens de OECD durven de meest voorzichtige fietsers niet langer de straat op, omdat de campagnes wijzen op de risico's van de sport. Daarnaast wordt er minder gefietst wanneer het dragen van een fietshelm verplicht is, waardoor het aantal ongevallen afneemt (Organisation for Economic Co-operation and Development, 2013).

Ook het aanpassen van de reglementen kan een invloed hebben op het aantal valpartijen in een wedstrijd. Lybbert et al. (2012) onderzoeken de impact van de wijziging van de Red Flag Rule op het aantal valpartijen in een etappe die eindigt op een massasprint. Deze regel werd ingevoerd om klassementsrenners te beschermen tegen tijdsverliezen die veroorzaakt worden door valpartijen aan het eind van een etappe. Daarom krijgen de renners de tijd toegewezen van de groep waarin ze zich bevonden bij het ingaan van de laatste drie kilometer wanneer er hier een valpartij plaatsvindt. Aanvankelijk had deze tijdsmeting plaats bij het ingaan van de laatste kilometer, die aangeduid wordt met een rode vlag. Sinds het wijzigen van het reglement tellen Lybbert et al. (2012) opvallend meer valpartijen bij het ingaan van de laatste drie kilometer dan voorheen. Bovendien zijn er ook meer valpartijen in de laatste fase van de wedstrijd dan voorheen.

1.1.2 Kenmerken van een wedstrijd

Mogelijk zijn er wedstrijden waarin meer (of minder) gevallen wordt dan anderen. Het is nuttig de kenmerken van deze wedstrijden meer in detail te bestuderen. Deze koersen kunnen immers een voorbeeld vormen voor de maatregelen die nodig zijn ter preventie van valpartijen. Dit aspect komt aan bod in een tweede deelvraag:

> ZIJN ER WEDSTRIJDEN WAARIN MEER (OF MINDER) GEVALLEN WORDT DAN ANDEREN?

1.1.3 Categorie van een wedstrijd

Bij het opstellen van de kalender worden wedstrijden ingedeeld in categorieën. Er wordt rekening gehouden met het type, de moeilijkheidsgraad en internationale uitstraling van elke wedstrijd. Deze categorieën vormen het fundament van het puntensysteem van de Union Cycliste Internationale (UCI). Hoe hoger de categorie van een wedstrijd, hoe meer punten een overwinning oplevert. Ook neemt de prijzenpot die de wedstrijdorganisatie kan verdelen toe met de categorie die een wedstrijd toebedeeld krijgt.

Het puntensysteem wordt gezien als een maatstaf voor de kwaliteit van een renner. Op de transfermarkt zijn het vooral de renners met een hoog puntensaldo die snel een contract weten te versieren. Ze impliceren immers een grote(re) waarde voor een team. Daarnaast is dit systeem de basis voor het ploegenklassement. Dit wordt opgesteld door de punten van de twaalf beste renners van elk team te sommeren. Dit klassement is vooral van strategisch belang; het beslist namelijk over de volgorde van de volgwagens in de wedstrijden. Een renner heeft zijn volgwagen liefst vooraan, omdat dit de wachttijden op assistentie sterk reduceert. In de aanloop van een belangrijke wedstrijd is het dus noodzakelijk voldoende punten te verzamelen, want enkel dan kan de ploegassistentie goed gepositioneerd worden. De invloed van het puntensysteem op de risicobereidheid van het peloton kan onderzocht worden aan de hand van volgende vraag:

> WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE CATEGORIE VAN EEN WEDSTRIJD?

Het reglement van de UCI beïnvloedt het gedrag in het peloton. Het eerder besproken onderzoek over de Red Flag Rule toonde al aan dat de renners meer risico nemen in de laatste drie kilometer van de wedstrijd wanneer fouten hier minder afgestraft worden (Lybbert et al., 2012). Het puntensysteem heeft mogelijk een gelijkaardig effect.

1.1.4 Lengte van de wedstrijd

De categorie van een wedstrijd bepaalt de maximale lengte van een wedstrijdparcours. Het UCI-reglement legt immers beperkingen op voor koersen uit de lagere categorieën. Het parcours van deze wedstrijden mag niet meer dan 200 kilometer lang zijn. De races uit de hoogste categorieën kennen geen beperkingen.

De lengte van het parcours is een belangrijk gegeven bij de preventie van valpartijen, in het bijzonder wanneer een parcours ingekort of verlengd wordt. Dit aspect zal verder onderzocht worden aan de hand van volgende deelvraag:

> WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN PER WEDSTRIJD EN DE LENGTE VAN HET WEDSTRIJDPARCOURS?

Het onderzoek van Townes et al. (2005) vindt dat de klachten over ziekte, blessures en verwondingen toenemen met de lengte van een wedstrijd. De conclusies van dit onderzoek zijn echter niet volledig toepasbaar op de setting van een eendagswedstrijd. Enerzijds zijn de data afkomstig van een niet-competitief evenement dat gespreid werd over meerdere dagen. Hierin speelt de vermoeidheid van de deelnemers een veel grotere rol dan bij een eendagswedstrijd. Anderzijds zijn niet alle klachten het gevolg van een valpartij.

1.1.5 Gemiddelde snelheid van de winnaar

De gemiddelde snelheid waarmee een wedstrijd wordt afgewerkt is geen onbelangrijk aspect. Thompson en Patterson (1998) toonden al aan dat de gevolgen van een valpartij ernstiger zijn bij een valpartij aan hoge snelheid. Eluru et al. (2008) komen tot een zelfde conclusie na het onderzoeken van de ernst van verwondingen bij ongevallen met fietsers in het dagelijks verkeer.

Het onderzoek van de OECD (2013) concludeert dat het aantal valpartijen kan gereduceerd worden door het verkeer te vertragen. Idealiter ligt de snelheid van het gemotoriseerd verkeer rond de fietsers niet hoger dan 30 km/u. In haar onderzoek citeert de OECD ook enkele andere onderzoeken die stellen dat de vertragende elementen die vandaag gebruikt worden niet de beste optie zijn. De vele verkeersdrempels en rotondes die geïnstalleerd werden, doen het aantal valpartijen zelfs toenemen.

Ook Roi en Tinti (2014) komen tot de conclusie dat er beter trager gereden kan worden. Zij stellen dat 63% van de aanvragen tot medische assistentie bij niet-professionele wedstrijden te wijten zijn aan valpartijen. De snelheid van de renners is hierbij een bijkomende risicofactor, waarmee ze bedoelen dat er meer ongevallen zijn wanneer aan hoge snelheid gereden wordt. Bovendien ligt de snelheid van de professionele renners haast dubbel zo hoog dan die van de recreatieve fietsers. Dit geeft mogelijk aanleiding tot nog meer valpartijen.

De gemiddelde snelheid van de renners is de laatste decennia alleen maar toegenomen. Het onderzoek van El Helou et al. (2009) analyseert de gemiddelde snelheden tussen 1892 en 2008 over de drie grote rondes (Giro, Tour en Vuelta) en acht Europese wedstrijden[1]. Dit onderzoek onderscheidt vier grote periodes, die elk gekenmerkt worden door een toename van de gemiddelde snelheden.

[1] Het gaat hier om Milaan-Sanremo, Parijs-Roubaix, Waalse Pijl, Luik-Bastenaken-Luik, Parijs-Nice, Dauphiné Libéré, Vier dagen van Duinkerke en Midi Libre.

Een eerste periode eindigt bij het eind van Wereldoorlog I. In deze periode halen de renners gemiddelde snelheden tussen 25 en 30 km/u. Tijdens het interbellum kent de gemiddelde snelheid de grootste stijging en groeit naar waarden om en bij 35 km/u. Na Wereldoorlog II klimt de gemiddelde snelheid van de renners naar waarden rond 37 km/u. Een vierde en laatste periode start in 1993. De gemiddelde snelheid van de renners neemt opnieuw toe naar een globaal gemiddelde van ongeveer 41 km/u. Dit is een toename van 6,38% ten opzichte van de vorige periode.
Aan de basis van deze evolutie liggen een reeks externe en interne factoren. De externe factoren zijn niet te controleren. In deze categorie vinden we de aerodynamische, technologische en medische verbeteringen die de afgelopen jaren geboekt werden terug. Zo nam het gewicht van de fietsen af van 40 kg (in 1864) naar 7 tot 8 kg (in 1986).
De klasse van de interne factoren bevat onder andere de verbeterde technieken voor het trainen en begeleiden van atleten. Ook deze factoren verklaren de toename van de gemiddelde snelheid tussen 1892 en 2008. Volgens de auteurs ligt ook het gebruik van prestatie-bevorderende middelen, zoals EPO, aan de basis voor de overgang tussen periode drie en vier (El Helou et al., 2009).

Perneger (2010) beperkt zich tot een analyse van de gemiddelde snelheden in Giro, Tour en Vuelta tussen 1947 en 2009. Hij is het niet volledig eens met de conclusies uit het vorige onderzoek. Perneger (2010) stelt eveneens vast dat de gemiddelde snelheden zijn toegenomen, maar hij laat de vierde periode starten bij het begin van de jaren ’80. Dit is tien jaar eerder dan in het onderzoek van El Helou et al. (2009). Uit zijn berekeningen blijkt dat de gemiddelde snelheden stagneren vanaf de jaren ’90. In de periode tussen 1990 en 2004 neemt de gemiddelde snelheid toe met amper 0,16 km/u. Na 2004 gaat de gemiddelde snelheid over de drie grote rondes zelfs dalen. Het globaal gemiddelde neemt dan af met 0,22 km/u. Perneger (2010) vergelijkt deze evolutie van de gemiddelde snelheden met de timing van beslissingen inzake het antidopingbeleid. De daling van de gemiddelde snelheid bij het begin van de jaren 2000 valt samen met het vermeerderen en het verstrengen van de dopingcontroles. Hij besluit dat het afnemen van de prestaties van de renners gerelateerd is aan het succes van deze dopingcontroles, ook omdat er geen andere verklaringen zijn voor deze evolutie. Het onderzoek benadrukt echter ook dat het onmogelijk is deze hypothese te verifiëren, omwille van het gebrek aan data.

Ook het onderzoek van Lodewijkx en Brouwer (2012) beperkt zich tot een analyse van de gemiddelde snelheden van de drie grote rondes. Zij komen tot dezelfde conclusies als Perneger (2010). In de jaren ’80 stellen ze immers een nieuwe stijging van de gemiddelde snelheden vast, al kent de Giro een tragere evolutie dan andere wedstrijden. Deze stijging vlakt af in de jaren ’90, zoals ook Perneger (2010) vaststelde.
Lodewijkx en Brouwer (2012) zoeken de verklaring voor deze toename in de jaren ’80 bij de internationalisering en commercialisering van het wielrennen. Verder wijzen ze op de wijzigingen in de structuur van een team. Voor de jaren ’80 is er één kopman die bijgestaan wordt door een team van helpers. In de jaren ’80, ’90 en 2000 kent een team meerdere kopmannen en schaduwkopmannen die elk hun eigen kans mogen wagen. Vanaf de jaren ’80 vindt ook de sportvoeding zijn ingang in het peloton. Deze drankjes en voedingssupplementen kunnen een bijdrage geleverd hebben bij het toenemen van de gemiddelde snelheid van de renners (Lodewijkx en Brouwer, 2012).
Voor het afvlakken van de gemiddelde snelheden in de jaren ’90 verwijzen Lodewijkx en Brouwer (2012) naar de eerder besproken verklaringen van Perneger (2010). Volgens de auteurs is het succes van de dopingcontroles echter niet de enige reden voor deze evolutie. Zij verwijzen naar de wet van de afnemende meeropbrengsten; de fysieke verschillen tussen de renners worden alsmaar kleiner, daarom groeien de prestaties op het topniveau steeds meer naar elkaar toe. Omdat er geen wetenschappelijk onderzoek is dat deze verklaring ondersteunt, raden de auteurs aan deze verklaring van naderbij te bekijken in toekomstig onderzoek.

Deze onderzoeken wijzen in de richting van een relatie tussen het aantal valpartijen en de gemiddelde snelheid tijdens een wedstrijd. Daarom zal dit verband meer in detail onderzocht worden met behulp van volgende deelvraag:

> WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE GEMIDDELDE SNELHEID VAN DE WINNAAR?

1.1.6 Grootte van het peloton

De opiniestukken die de oorzaken van de vele valpartijen bespreken, stellen vaak voor om het peloton te verkleinen (Lagae in De Standaard, 2011 en 2013 en Maeckelberg in De Morgen, 2014).
Dit brengt volgend aspect aan het licht:

> WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN HET AANTAL STARTERS IN EEN WEDSTRIJD?

De praktijk leert dat een peloton in het dagelijks verkeer beter niet meer dan dertig personen bevat, dit om de veiligheid te garanderen. Er is echter geen wetenschappelijk onderzoek dat deze stelling ondersteunt. Het onderzoek van Johnson et al. (2009) onderkent het belang van de grootte van een peloton. Zij raden aan dit aspect in de toekomst verder te onderzoeken.

Bij een wedstrijd wordt een parcours verkeersvrij gemaakt. Dit betekent dat het overige verkeer geen rol meer speelt, wat mogelijk een groter peloton toelaat. Er is echter geen onderzoek naar het maximale aantal renners binnen deze setting. Ter preventie raden sommige onderzoeken aan om de grootte van het peloton af te stemmen op de aard van het parcours (Roi en Tinti, 2014).

1.1.7 Gebruik van oortjes

Sinds de jaren ’90 wordt gebruik gemaakt van oortjes tijdens de wedstrijden. Dit is de draadloze communicatie tussen renners en hun ploegleiders in de volgwagen. Het gebruik van deze technologie staat al langer ter discussie in de wielerwereld.

De tegenstanders stellen immers dat deze vorm van communicatie de aard van de sport verstoort en de renners afleidt (Lippi et al., 2011). Goldenbeld et al. (2011) bestuderen een gelijkaardige situatie bij fietsers in het dagelijks verkeer. De veronderstelling dat naar muziek luisteren, telefoneren of sms’jes versturen tijdens het fietsen het risico op betrokkenheid in een ongeval verhoogt, werd echter niet significant bevonden.
Het gebruik van handsfree toestellen lost het probleem echter niet op. Uit het onderzoek van de Waard et al. (2011) blijkt dat fietsers aangeven zich veilig te voelen bij het gebruik van deze toestellen, maar dat dit gevoel mogelijk niet correct is. Uit de analyse blijkt dat de fietsers significant sneller afremmen bij het waarnemen van een stopsignaal, maar niet in staat zijn om het geluid uit de omgeving beter waar te nemen.
Deze negatieve effecten verdwijnen wanneer fietsers slechts één van de twee oortjes gebruiken. De auditieve waarneming van deze fietsers is significant beter dan deze van fietsers die twee oortjes gebruiken om muziek te luisteren (de Waard et al., 2011). In het peloton gebruiken de renners slechts één oortje. De bevindingen van de Waard et al. (2011) spreken dus in het voordeel van de voorstanders van het oortjesgebruik.

De voorstanders van het gebruik van oortjes wijzen op de mogelijkheid om gevaarlijke passages te signaleren, zodat renners zich niet moeten laten uitzakken om te kunnen overleggen met hun ploegleider en deze ook kunnen bereiken wanneer ze technische of medische assistentie nodig hebben. Gueguen (2010) sluit zich hierbij aan; hij stelt vast dat het gebruik van oortjes niet zorgt voor een toename van het aantal etappes dat eindigt in een massasprint. Noch beïnvloeden ze het massaal toekomen van de renners. Integendeel, de tijd tussen de aankomsten van de eerste dertig renners is zelfs toegenomen.

In 2011 besloot de UCI om oortjes te bannen in wedstrijden met een categorie van 1.HC[2] of lager (Lippi et al., 2011). Dit om te vermijden dat de ploegleiders de wedstrijd controleren en de renners vanuit de auto de nodige bevelen geven. Deze regel werd in de eerste plaats ingevoerd om de koersen spannender te maken. Het loont echter de moeite om te onderzoeken of deze maatregel ook de veiligheid verhoogt en het aantal valpartijen weet te verminderen.

De relatie tussen het aantal valpartijen en het gebruik van oortjes zal geanalyseerd worden aan de hand van volgende deelvraag:

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN HET GEBRUIK VAN OORTJES?

1.1.8 Weersomstandigheden

Weersomstandigheden zijn een ander gegeven bij een wielerwedstrijd. De analyse van dit aspect is eveneens van belang bij de preventie van valpartijen en wordt weergegeven in volgende vraag.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE WEERSOMSTANDIGHEDEN?

Roi en Tinti (2014) stellen vast dat de weersomstandigheden een belangrijke rol spelen bij de hoeveelheid aanvragen tot medische assistentie. Deze bepalen niet alleen de aard, maar ook de hoeveelheid aanvragen in een amateurwedstrijd. In geval van regenweer zijn er maar liefst 84% meer aanvragen tot medische assistentie.

In het huidige veiligheidsbeleid van de UCI zijn organisatoren verplicht te zorgen voor een alternatieve route bij slechte weersomstandigheden. In extreme gevallen, zoals sneeuw, kan zelfs beslist worden om een wedstrijd niet te laten doorgaan (Organiser's Guide to Road Events, 2013).

Lippi et al. (2011) delen de mening van de UCI. Ze vragen onder andere aandacht voor de verf die gebruikt wordt voor het aanbrengen van wegmarkeringen. Deze kan mogelijk glad worden bij regenweer, wat het risico op valpartijen vergroot. De verfkeuze is moeilijk na te gaan over het ganse parcours, maar zou expliciet in rekening moeten genomen worden in de start- en finishzone van de wedstrijden. Eerder onderzoek wees al uit dat de eerste en laatste kilometers het gevaarlijkst zijn (McLennan et al., 1988).

[2] Het huidige systeem van de UCI maakt een onderscheid tussen World Tour (WT) wedstrijden en het continentale circuit. Binnen dit continentale circuit zijn er drie categorieën, namelijk 1.HC, 1.1 en 1.2.

1.1.9 Parcours van de wedstrijd

Eendagswedstrijden worden gekenmerkt door hun parcours. In de eerste helft van het voorjaar worden de kasseiklassiekers afgewerkt. Deze koersen bevatten korte hellingen, waarvan de ondergrond vaak uit kasseien bestaat. Daarnaast worden verschillende kasseistroken in het parcours opgenomen. Daarna volgt een reeks wedstrijden met langere en steilere beklimmingen. Aan het eind van het seizoen, in het najaar, volgen dan de najaarsklassiekers. Het parcours van deze laatste groep koersen is zeer uiteenlopend. Een volgende vraag die gesteld kan worden is dan:

> WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN HET WEDSTRIJDPARCOURS?

Leonard et al. (2005) bestuderen het effect van een gewijzigd parcours in de motor- en autosport. Het parcours van Castle Combe werd aangepast door het toevoegen van twee chicanes, i.e. S-vormige bochten die de race moeten vertragen. Uit onderzoek blijkt dat deze wijziging het aantal blessures doet afnemen. De auteurs merken op dat er mogelijk ook invloed is van andere veranderingen, zoals de hogere veiligheidseisen die aan de voertuigen gesteld worden.
Het vertragen van een wedstrijd is echter niet de enige mogelijkheid om de veiligheid te verbeteren. Bovendien is er een trade-off met het spektakel en de historische waarde van een wedstrijd (Lippi en Guidi, 2005).
Lippi et al. (2011) onderstrepen het belang van veilige en goed onderhouden wegen. Het wijzigen van het parcours om deze wegen op te zoeken is echter niet altijd even realistisch. Sommige elementen in het parcours hebben zo'n historische waarde dat ze het karakter van de wedstrijd bepalen.

1.1.10 Media-aandacht

De middelen van een wielerploeg worden voorzien door een sponsor, die een zekere *return on investment* verlangt. Deze kan gerealiseerd worden door renners die zich in de kijker rijden tijdens een lange vlucht of de overwinning behalen (Morrow en Idle, 2008). Uiteraard ziet een sponsor zijn renners liever niet betrokken in een valpartij, of erger nog, uitvallen tijdens een wedstrijd.
Om voordeel te halen uit het sponsoren van een ploeg is het belangrijk goed te presteren in de juiste wedstrijden. Dit zijn de wedstrijden in de omgeving van de sponsor en/of die wedstrijden die veel aandacht krijgen. Met volgende vraag kan gemeten worden of renners bereid zijn meer risico's te nemen wanneer een wedstrijd veel media-aandacht krijgt:

> WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE MEDIA-AANDACHT DIE EEN WEDSTRIJD KRIJGT?

Dit aspect kwam al aan bod in het onderzoek van Van Reeth (2013). In dit onderzoek toont hij aan dat de kijkcijfers voor de Tour de France hoger zijn wanneer een etappe belangrijk is voor het eindklassement, of wanneer het parcours een spannende wedstrijd voorspelt. Verder beargumenteert hij ook dat het gemiddelde kijkcijfer een maat is voor de interesse van de 'echte' wielerfan, want deze personen kijken een volledige wedstrijd. Dit in tegenstelling tot het piekkijkcijfer, dat een maat is voor de werkelijke interesse voor het live-verslag van een wedstrijd.

1.2 Doelgroep

Het antwoord op deze vragen is in de eerste plaats relevant voor de renners die aan deze wedstrijden deelnemen. Zij hebben er alle belang bij hun job te kunnen uitoefenen in zo veilig mogelijke omstandigheden.

Daarnaast is dit probleem ook een actueel vraagstuk voor de UCI. In het verleden werd een werkgroep samengesteld die het veiligheidsaspect van de sport zou bestuderen. Deze raakte echter in de vergetelheid na het aanstellen van Brian Cookson tot nieuwe UCI-voorzitter. Toch, in een recente persmededeling kondigde de UCI de geboorte van een nieuwe commissie aan die zich zal toeleggen op veilige omstandigheden voor fietsers in het algemeen. De UCI hoopt zo meer mensen aan te zetten tot het gebruik van de fiets. Een veilige wielersport als uithangbord kan deze doelstelling alleen maar ten goede komen.

De resultaten van dit onderzoek hebben mogelijk ook impact op de organisatoren van de wedstrijden. Op welke manier kunnen zij het aantal valpartijen en de bijhorende blessures vermijden? Wanneer blijkt dat het aantal valpartijen beïnvloed wordt door het parcours of de weersomstandigheden, dan kan dit in rekening genomen worden om een beter evenwicht te vinden tussen spektakel en veiligheid. Met welk type parcours vergroot een organisatie het risico op valpartijen? Onder welke omstandigheden wordt een wedstrijd beter afgelast? Met een antwoord op deze vragen kunnen organisatoren hun steentje bijdragen in de preventie van valpartijen.

Tot slot zullen de verzekeraars van de renners een beter beeld krijgen van de risicofactoren van een bepaalde wedstrijd. Valpartijen zijn immers vaak de oorzaak van blessures, waarvan de kosten voor het behandelen en opvolgen vergoed moeten worden. Daarnaast zijn er ook kosten verbonden aan het missen van volgende wedstrijden. De afwezigheid van één of meerdere renners ten gevolge van een valpartij impliceert een verlies van inkomsten voor het hele team.

2 Data

Het karakter van de wielersport bemoeilijkt het verzamelen van gegevens. Daarom gaat dit tweede hoofdstuk dieper in op de data die verzameld werden en welke bronnen hiervoor gebruikt werden.

De dataset van dit onderzoek telt 272 observaties. Elke observatie stemt overeen met een wedstrijd die werd afgewerkt in de elite-categorie bij de mannen tussen februari 1997 en oktober 2013. In deze categorie moeten alle renners in het bezit zijn van een UCI-licentie. Een renner met dergelijke licentie gaat akkoord met de statuten en reglementen die de internationale wielerunie oplegt. In ruil krijgt hij het recht om deel te nemen aan de wielerwedstrijden die georganiseerd worden. De meerderheid van deze renners is ouder dan 23 jaar. Enkel de grote talenten maken de overstap naar het profpeloton op jongere leeftijd.

De bespreking van de dataset telt drie delen. Eerst wordt een korte beschrijving gegeven van de zestien Europese eendagswedstrijden die in de analyse aan bod komen. De wedstrijden worden beschreven in de volgorde waarin ze voorkomen op de huidige kalender. Het volgende deel behandelt de competitiemodellen van de UCI. Deze modellen zijn belangrijk, omdat ze de wielerkalender van de nodige structuur voorzien. Het derde en laatste deel biedt een overzicht van de overige wedstrijdgegevens die bijeen gebracht werden. Voor deze kenmerken zal de betekenis van de gegevens besproken worden, alsook de manier waarop deze verzameld werden.

2.1 De wedstrijden

2.1.1 Het Belgische openingsweekend

Het Belgische wielerseizoen start met een openingsweekend. In twee dagen worden twee koersen afgewerkt. Op zaterdag gaan de renners van start in de Omloop het Nieuwsblad en op zondag werkt het peloton Kuurne-Brussel-Kuurne af.

De start en finish van de Omloop het Nieuwsblad, voorheen gekend als de Omloop het Volk[3], liggen in de Gentse binnenstad. Tijdens deze wedstrijd trekken de renners een eerste keer over de hellingen en kasseien van de Vlaamse Ardennen. Deze koers is een organisatie van Flanders Classics en werd ondergebracht in het Europese continentale circuit, dat de naam UCI Europe Tour kreeg.

Ook Kuurne-Brussel-Kuurne is deel van de UCI Europe Tour. De renners fietsen van Kuurne richting Brussel, tot het keerpunt in de buurt van Ninove. Daarna keert het peloton terug naar de finish in Kuurne. De organisatie is in handen van de Koninklijke Sportingclub Kuurne. Deze wedstrijd is één van de twee Vlaamse wedstrijden die niet georganiseerd worden door Flanders Classics.

[3] In 2008 gingen het Volk en de Gentenaar op in het Nieuwsblad. De wedstrijd wijzigde haar naam naar Omloop het Nieuwsblad, omdat deze krant de nieuwe hoofdsponsor werd.

2.1.2 Voorjaarsklassiekers

2.1.2.1 Milaan-Sanremo

Wanneer de lente aanbreekt, trekken de renners naar Italië voor Milaan-Sanremo. Deze wedstrijd is niet alleen de eerste, maar ook de langste klassieker van het seizoen. Aan de finish hebben de renners ongeveer 300 km in de benen.

De wedstrijd behoort tot de UCI World Tour. In deze competitie worden de topwedstrijden van het wielerseizoen samengebracht. De koers kreeg ook een plaats tussen de monumenten van de wielersport; dit is een verzameling van de vijf oudste en zwaarste wedstrijden van het seizoen. De organisatie van dit evenement is in handen van RCS Sport. Dit bedrijf is een deel van de Italiaanse RCS Media Group, die verantwoordelijk is voor de uitgave van de roze sportkrant *Gazzetta dello Sport*.

2.1.2.2 Kasseiklassiekers en Brabantse Pijl

Na Milaan-Sanremo keert het peloton terug naar Vlaanderen. De Vlaamse wielerweek start met de wedstrijd Dwars door Vlaanderen, die georganiseerd wordt door Flanders Classics. Het parcours van deze race lijkt op dat van de Ronde van Vlaanderen, maar is korter. De wedstrijd werd ingedeeld in de UCI Europe Tour competitie.

Eind maart volgt de E3 Harelbeke, die haar naam dankt aan de oorspronkelijke naam van de snelweg tussen Harelbeke en Antwerpen. De race wordt georganiseerd door de vzw Hand in Hand. De E3 Harelbeke is samen met Kuurne-Brussel-Kuurne de enige Vlaamse voorjaarswedstrijd die niet in handen is van Flanders Classics.
Het parcours van de E3 Harelbeke is gelijkaardig aan dat van de 'oude' Ronde van Vlaanderen. In 2012 besliste de UCI om deze koers te promoveren van de UCI Europe Tour naar de UCI World Tour.

Vervolgens is het de beurt aan Gent-Wevelgem. In deze koers starten de renners in Deinze, bij Gent en rijden dan naar Wevelgem. De wedstrijd is deel van het UCI World Tour circuit en de organisatie is in handen van het Flanders Classics van Wouter Vandenhautte. In de oude kalender vond Gent-Wevelgem plaats op de woensdag tussen de Ronde van Vlaanderen en Parijs-Roubaix. Enkele jaren geleden verplaatste de UCI dit evenement naar de zondag voor de Ronde van Vlaanderen.

De Ronde van Vlaanderen is het hoogtepunt van het Vlaamse voorjaar. 'Vlaanderens Mooiste' is een World Tour wedstrijd die georganiseerd wordt door Flanders Classics. Net als Milaan-Sanremo maakt deze koers deel uit van de vijf monumenten van de wielersport. In 2012 beslisten de organisatoren het parcours te hertekenen. De start in Brugge bleef behouden, maar de finish werd verplaatst naar Oudenaarde. Deze nieuwe aankomstplaats is het gevolg van het invoeren van een lus rond de Paterberg en Oude Kwaremont. Deze hellingen vormen het nieuwe zwaartepunt van de wedstrijd en moeten meermaals beklommen worden.

Drie dagen later volgt de Scheldeprijs. Deze race wordt typisch gewonnen door een sprinter. De start van de Scheldeprijs vindt plaats in Antwerpen, daarna volgt een eerder vlak parcours. De finish ligt op de Churchilllaan in Schoten. Ook deze wedstrijd is een organisatie van Flanders Classics en behoort tot de UCI Europe Tour.

Een week na de Ronde van Vlaanderen trekken de renners naar Frankrijk voor 'de Hel van het Noorden'. De organisatie van Parijs-Roubaix is in handen van Amaury Sport Organisation (ASO), die ook de Tour de France voor hun rekening nemen. ASO is deel van de Franse mediagroep Editions Philippe Amaury (EPA). Parijs-Roubaix is het derde wielermoment van het seizoen en behoort tot het World Tour circuit.

De Brabantse Pijl maakt de overgang tussen het eerste en het tweede deel van het voorjaar. Het parcours bevat geen kasseistroken meer en de hellingen worden ingeruild voor steilere en langere heuvels. Daarom verschijnt er ook een ander type renner aan de start. De Brabantse Pijl behoort tot de UCI Europe Tour en is de laatste van de zes wedstrijden die georganiseerd worden door Flanders Classics.

2.1.2.3 Amstel Gold Race en Waalse klassiekers

Het tweede deel van de voorjaarsklassiekers vangt aan met de Amstel Gold Race. In deze race krijgen de renners de heuvels van Zuid-Limburg voor de wielen geschoven. De Amstel Gold Race is de enige Nederlandse klassieker in het World Tour circuit. De wedstrijd wordt georganiseerd door de Stichting Amstel Gold Race.

Na de Amstel Gold Race trekt het peloton terug naar België voor de Waalse klassiekers. Eerst is het de beurt aan de Waalse Pijl. Dit evenement is een organisatie van het Franse ASO en is deel van de UCI World Tour competitie. Sinds 1983 finisht de wedstrijd op de steile Muur van Huy.

Luik-Bastenaken-Luik sluit het rijtje van de voorjaarsklassiekers. Deze wedstrijd werd voor het eerst georganiseerd in 1892 en is daarmee de oudste van alle klassiekers. Het leverde de wedstrijd de bijnaam *La Doyenne* (‘de oude dame’) op, evenals een plaats tussen de vijf wielermonumenten. De organisatie van deze World Tour wedstrijd wordt verzorgd door het Franse ASO.

2.1.3 Najaarsklassiekers

In de zomer gaat de aandacht van de wielerfan vooral naar de drie grote rondes. Er staan echter ook enkele eendagswedstrijden op het programma. Eind juli of begin augustus, afhankelijk van het einde van de Tour, wordt de Clasica San Sebastian gereden. Het parcours situeert zich in en rond de stad San Sebastian in het Spaanse Baskenland. De krant *El Diario Vasco* is verantwoordelijk voor de organisatie van dit World Tour evenement.

In augustus vindt de Vattenfall Cyclassics plaats. De eerste editie van deze race vond plaats in 1996, toen nog onder de naam HEW Cyclassics[4]. De wedstrijd is daarmee de jongste klassieker op de kalender. Het parcours situeert zich in de stad Hamburg en is eerder vlak; vaak gaat een sprinter met de bloemen naar huis. Deze wedstrijd is de enige Duitse race in het World Tour circuit.

De Ronde van Lombardije is het vijfde en laatste monument van de wielersport. Traditioneel staat deze herfstklassieker ingepland aan het eind van de maand oktober. Sinds 2004 kende het parcours nog weinig wijzigingen. De renners starten in het Zwitserse Mendrisio en finishen in het Italiaanse Como. Onderweg krijgt het peloton een heuvelachtig parcours voorgeschoteld. De wedstrijd is een organisatie van RCS Sport en behoort tot het UCI World Tour circuit.

[4] HEW was een energiemaatschappij uit Hamburg. In 2006 werd het bedrijf overgenomen door het Zweedse Vattenfall. Sindsdien draagt de wedstrijd de naam van de nieuwe hoofdsponsor.

2.2 De categorieën

Het tweede deel van dit hoofdstuk neemt de competitiemodellen van de UCI onder de loep. Een competitiemodel brengt structuur aan in de kalender door aan elke wedstrijd een categorie toe te kennen. Bij het toewijzen van deze categorieën wordt rekening gehouden met het type, de moeilijkheidsgraad en internationale uitstraling van een wedstrijd.

De UCI introduceerde reeds verschillende systemen om de eendagswedstrijden te combineren in één competitie. Het verleden leert dat een model een aantal jaren gebruikt wordt en dan weer ingeruild wordt voor een nieuwe methode.

Tabel 1 biedt een overzicht van de competitiemodellen sinds 1989. Ook voor 1989 maakte de UCI al gebruik van systemen voor het organiseren van het wielerjaar, maar deze modellen worden hier niet besproken. Alle wedstrijden in de dataset werden immers afgewerkt tussen 1997 en 2013 en kregen zo een plaatsje in één van onderstaande competitiemodellen. In het verdere verloop van de tekst krijgen de verschillende modellen een korte toelichting.
Daarnaast biedt Tabel 1 de mogelijkheid om de verschillende competitiemodellen te vertalen naar één gemeenschappelijke schaal. De hoogste categorie van elk systeem wordt omgedoopt tot categorie 1. De volgende categorieën worden achtereenvolgens genummerd van categorie 2 tot categorie 7. De eerste kolom van de tabel biedt een overzicht van de manier waarop de verschillende categorieën vertaald worden naar deze nieuwe, ordinale schaal.

Nieuwe schaal		**1989-2004**	**2005-2007**	**2008-2010**	**2011-Heden**
	Competitie-model	*UCI World Cup*	*UCI ProTour*	*UCI ProTour en Historische Kalender*	*UCI World Tour*
Categorie 1	Internationale competitie	CDM	PT	PT HIS	WT
Categorie 2	Continentale competitie	.BC (of .HC)	.BC (of .HC)	.BC (of .HC)	.BC (of .HC)
Categorie 3		.1	.1	.1	.1
Categorie 4		.2	.2	.2	.2
Categorie 5		.3			
Categorie 6		.4			
Categorie 7		.5			

Tabel 1 Overzicht van de competitiemodellen vanaf 1989

2.2.1 UCI World Cup

In 1989 werd de UCI World Cup in het leven geroepen. Deze competitie verzamelde de tien belangrijkste eendagswedstrijden van het seizoen. Op de kalender werden deze wedstrijden aangeduid met de categorie *Coupe du Monde* (CDM).

De wedstrijden buiten de Wereldbekercompetitie werden ingedeeld in de categorieën .BC, .1, .2, .3, enz. De hoogste categorie bevatte de wedstrijden buiten categorie, aangeduid met .BC (of .HC, voor het Franse *hors catégorie*). Daarna volgden de wedstrijden van eerste categorie, genoteerd met .1. De volgende categorie bevatte de wedstrijden van tweede categorie; deze kregen de gelijkaardige notatie .2. De wedstrijden van derde categorie werden weergegeven als .3, enz. Om het verschil tussen een eendagswedstrijd en meerdaagse rittenkoers weer te geven werd het cijfer 1, respectievelijk 2 voor het punt van de categorie geschreven.

2.2.2 UCI ProTour

Sinds 2005 maakt de UCI een onderscheid tussen twee circuits. De drie grote rondes, enkele kleinere rittenkoersen en de tien wereldbekerwedstrijden werden samengevoegd tot een eerste circuit. Deze verzameling topwedstrijden kreeg de naam UCI Pro Tour. De wedstrijden in het ProTour circuit werden aangeduid met de letters PT.

Verder kregen alle teams de kans zich kandidaat te stellen voor een ProTour licentie. Na een sportieve en financiële doorlichting reikte de UCI deze licenties uit aan de twintig beste teams. Deze ProTour teams had het recht en de plicht om deel te nemen aan elke wedstrijd uit de ProTour. Ze werden echter niet weerhouden van deelname in andere, niet-ProTour wedstrijden. Daarnaast kregen de organisatoren de toestemming om bij elke wedstrijd een aantal *wild cards* uit te delen aan niet-ProTour ploegen.

Deze regeling was echter niet naar de zin van enkele organisatoren. ASO, RCS en Unipublic[5] wilden zelf bepalen wie toestemming kreeg om deel te nemen aan hun evenementen. Zij wilden af van de verplichte uitnodiging aan alle ProTour teams. De UCI gaf echter niet toe en de discussie bereikte haar hoogtepunt tijdens de Tour de France van 2008. ASO, RCS en Unipublic haalden hun wedstrijden uit de ProTour en creëerden een eigen competitie. De opsplitsing tussen de historische kalender (HIS) en de ProTour bleef bestaan tot het invoeren van de World Tour in 2010.

Het tweede circuit kent vijf competities; één voor elk werelddeel. In Europa spreken we van de UCI Europe Tour. De vijf competities van het continentale circuit zijn volledig onafhankelijk; dit betekent dat elke competitie haar eigen wedstrijden, deelnemers en klassementen heeft. De races behoren tot de categorieën .BC, .1 of .2 die eerder beschreven werden. Het deelnemersveld bestaat uit ProTour ploegen, pro-continentale en continentale[6] ploegen.

In het continentale circuit gelden echter een aantal beperkingen. Het aantal deelnemende ProTour teams in één wedstrijd is beperkt tot 70%. De renners uit deze ploegen komen ook niet in aanmerking voor het klassement. Deze strijd is voorbehouden aan de renners uit de (pro-)continentale teams.

---

[5] Unipublic is de organisator van de Vuelta d' España (of de Ronde van Spanje)

[6] Het onderscheid tussen pro-continentale en continentale ploegen wordt bepaald door een verzameling specificaties van de UCI.

2.2.3 UCI World Tour

In 2010 deed de UCI enkele toegevingen, waardoor de discussie met ASO, RCS en Unipublic werd bijgelegd. Dit resulteerde in de UCI World Tour (WT) en bracht de wedstrijden uit de ProTour en historische kalender opnieuw bij elkaar. Het gebruik van de teamlicenties bleef bestaan, maar de organisatoren kregen inspraak bij het uitreiken van deze licenties. Opdat de organisatoren meer *wild cards* zouden kunnen uitdelen, zijn er nog achttien licenties in omloop. Het onderscheid met de continentale circuits bleef ook behouden. Bovendien kende dit circuit geen wijzigingen; de organisatie en reglementen van deze competitie blijven gelden zoals ze eerder besproken werden.

2.3 Kenmerken van de wedstrijd

Het laatste deel van dit hoofdstuk bespreekt de verschillende variabelen die in de analyse in rekening zullen genomen worden. Voor elk van deze variabelen wordt ook aangegeven op welke manier de gegevens verzameld werden.

2.3.1 Lengte van het parcours

Het parcours van een wedstrijd varieert van jaar tot jaar. Ook de lengte van het parcours kan hierdoor van jaar tot jaar verschillen. Meestal brengt de wedstrijdorganisatie enkel kleine wijzigingen aan in de route. Ieder parcours heeft immers een aantal kenmerkende elementen die telkens terugkeren. Deze kleine wijzigingen zijn bijvoorbeeld het gevolg van wegwerkzaamheden, waardoor de organisatie verplicht is de renners om te leiden. Het is ook mogelijk dat de organisatie een nieuwe helling of kasseistrook wenst op te nemen in het parcours of simpelweg voor alternatieve wegen kiest. Deze veranderingen zorgen er voor dat het aantal af te leggen kilometers van jaar tot jaar kan variëren.

De organisatie moet hierbij wel rekening houden met de maximale afstand die ze opgelegd kreeg. Deze is afhankelijk van de categorie waartoe een wedstrijd behoort. De Raad van de UCI World Tour bepaalt de maximale lengte van elke World Tour wedstrijd. De koersen uit de vroegere Historische Kalender vormen hierop een uitzondering. Bij deze races wordt de maximale lengte van het parcours bepaald door het directiecomité. In het continentale circuit geldt de regel dat in geen enkele wedstrijd meer dan 200 km mag worden afgelegd, tenzij de UCI hierop een uitzondering toestaat.

De gegevens over de lengte van het parcours zijn gemakkelijk te verzamelen. In haast elk verslag is informatie terug te vinden over de totale afstand die renners moesten afleggen. Deze data zijn ook online beschikbaar. Voor dit onderzoek werd beroep gedaan op de websites van *Bike Race Info, WV Cycling* en *ProCyclingStats.*

2.3.2 Gemiddelde snelheid van de winnaar

De gegevens over de gemiddelde snelheid van de winnaar zijn ook online beschikbaar op websites als *Bike Race Info, WV Cycling* en *ProCyclingStats.* Bovendien kan de gemiddelde snelheid gemakkelijk afgeleid worden wanneer deze informatie ontbreekt. Deze websites bevatten immers gegevens over de totale lengte van de race en de tijd die de winnaar nodig had om de wedstrijd af te werken.

2.3.3 Aantal starters

De huidige regels van de UCI bepalen dat een peloton niet meer dan 200 renners mag bevatten. Deze deelnemers komen uit maximaal 25 ploegen. Voor een eendagswedstrijd moet een team immers minimaal vier en maximaal acht renners afvaardigen. De namen van deze deelnemers moeten vooraf doorgegeven worden aan de organisatie.

Voor de start van een wedstrijd moeten de renners hun deelname bevestigen. Ze doen dit door het wedstrijdblad te komen tekenen. Het aantal handtekeningen bepaalt de definitieve grootte van het peloton. Renners die niet komen tekenen mogen immers niet starten. Wanneer een coureur zich niet meldt voor de start, krijgt hij *did not start* (DNS) achter zijn naam. Deze renners worden niet onder de starters gerekend.

De grootste deel van de gegevens over het aantal starters werd verzameld aan de hand van de website *Bike Race Info*. Wanneer de informatie hier ontbrak, werd gebruik gemaakt van de officiële wedstrijduitslagen. Deze zijn online beschikbaar via de website van *Cycling News.* Deze pagina geeft haast altijd een volledige uitslag van de wedstrijd weer. Wanneer deze informatie ook hier niet beschikbaar bleek, werd gebruik gemaakt van de publicaties van officiële uitslagen in kranten en tijdschriften. Deze zijn eveneens online beschikbaar via de website *GoPress Academic.*

2.3.4 Aantal finishers

De officiële uitslagen worden opgemaakt aan de finish. Wanneer een renner de finish niet haalt, dan krijgt hij de letters DNF achter zijn naam. Dit is een afkorting voor *did not finish*. Er zijn verschillende oorzaken voor een DNF mogelijk. Een renner kan opgegeven hebben of uit koers genomen zijn. Renners worden uit koers genomen bij het overtreden van het wedstrijdreglement of uit veiligheidsoverwegingen. Een renner kan ook buiten tijd eindigen. Het huidige UCI-reglement legt de tijdslimiet op 8%; dit betekent dat een renner buiten tijd eindigt wanneer hij 8% meer tijd nodig heeft dan de winnaar om het parcours af te werken. Ook in dit scenario krijgt een renner een DNF achter zijn naam.

Voor het verzamelen van gegevens over het aantal finishers werden dezelfde bronnen gebruikt als bij het aantal starters. Het merendeel van de gegevens werd dus verzameld aan de hand van de website *Bike Race Info*. In tweede instantie werd gebruik gemaakt van de uitslagen op *Cycling News*. Voor de minder recente wedstrijdedities bleek deze informatie vaak niet beschikbaar. Er werd dan beroep gedaan op de officiële uitslagen die gepubliceerd worden in kranten en tijdschriften. Deze archieven werden geraadpleegd met behulp van de *GoPress Academic* website.

2.3.5 Gebruik van radiocommunicatie

Aan het eind van de jaren ’90 raakten de oortjes ingeburgerd in het peloton. Deze draadloze vorm van communicatie biedt de renners en ploegleiders de mogelijkheid om met elkaar te overleggen tijdens de wedstrijd. De literatuurstudie van deze masterproef verwees al naar de discussie omtrent het gebruik van deze technologie. De voorstanders wijzen op de vele voordelen; zo kunnen de ploegleiders gevaren signaleren en hebben de renners de mogelijkheid om assistentie te vragen tijdens de wedstrijd. De tegenstanders stellen dat de renners sneller afgeleid zijn en dat de oortjes het koersverloop verstoren. De wedstrijd zou ook minder spannend zijn, omdat de renners steeds op de hoogte zijn van de koerssituatie en de ploegtactiek.

In de tiende etappe van de Tour de France van 2009 werd bij wijze van experiment zonder oortjes gekoerst. De etappe eindigde in de verwachte massasprint. De discussie bleef duren en in 2011 besliste de UCI om de oortjes te verbieden in niet-World Tour wedstrijden. Deze regel werd opgenomen in de dataset onder de dummy-variabele *radio.* In een wedstrijd met oortjes krijgt deze variabele de waarde *yes*, in het andere geval krijgt deze variabele de waarde *no*.

Wanneer de UCI een categorie toekent aan een wedstrijd, dan is deze beslissing geldig voor meerdere jaren. Na deze periode volgt een evaluatie. Het is dan mogelijk dat een wedstrijd een nieuwe categorie toebedeeld krijgt, waardoor het gebruik van de oortjes opnieuw verboden of toegelaten wordt. In de dataset is dat het geval voor de E3 Harelbeke. In 2012 werd de wedstrijd opgewaardeerd naar het World Tour niveau, waardoor het gebruik van oortjes toegelaten werd. In 2011, toen de wedstrijd nog tot het continentale wielercircuit behoorde, was het gebruik van oortjes verboden. Voor het invoeren van deze nieuwe regel waren de oortjes toegestaan in alle wedstrijden, ongeacht de categorie waartoe ze behoorden. Alle niet-World Tour wedstrijden schakelen daarom over van een dummy-variabele met waarde *yes* naar waarde *no*.

2.3.6 Weersomstandigheden

Het karakter van de wielersport bemoeilijkt het verzamelen van gegevens over temperatuur. Een parcours beslaat 170 tot 300 km; het is bijna onmogelijk hier frequent en op een correcte manier de temperatuur te meten. De recente live-verslagen op *Sporza* en *Cycling News* vermelden soms de temperatuur aan de start en/of aankomst. Er is echter geen informatie over het moment waarop deze meting plaatsvond en welke methode hiervoor gebruikt werd. Omdat de meerderheid van de gegevens ontbreekt en de meest recente gegevens weinig accuraat zijn, zal de temperatuur niet opgenomen worden in de analyse.

Om de weersomstandigheden alsnog te capteren, werd de variabele *rain* gecreëerd. Deze dummy-variabele krijgt de waarde *no* wanneer het droogt blijft tijdens de wedstrijd. In het geval dat de renners geconfronteerd worden met regen, sneeuw of hagel, krijgt deze variabele de waarde *yes*. Er wordt geen onderscheid gemaakt naar de aard en hoeveelheid neerslag die viel. Voor het creëren van deze dummy-variabele werden verschillende artikels op *Cycling News* en *Sporza* geraadpleegd. Deze websites voorzien live-feeds bij de belangrijkste wedstrijden. In dergelijk verslag wordt de race minuut per minuut opgevolgd. Wanneer de renners te maken krijgen met een regen-, sneeuw- of hagelbui, dan wordt dit vermeld in het verslag.

2.3.7 Hindernissen op het parcours

Bij de voorbeschouwing van een wedstrijd, publiceren websites als *Sporza* en *Cycling News* een korte beschrijving van de wedstrijdroute. Hier worden de belangrijkste hindernissen van het parcours besproken. Voor de eendagswedstrijden volgt een analyse van de hellingen en/of kasseistroken. Vaak worden deze analyses ondersteund met de tijdschema's en routeplannen die de organisatie beschikbaar stelt.

De hindernissen in het parcours werden samengevat in drie variabelen. Een eerste variabele geeft het aantal hellingen in het parcours weer. Deze informatie werd afgeleid uit de plannen, tijdschema's en besprekingen die bij de voorbeschouwing aan bod kwamen. De tweede variabele telt het aantal kasseistroken in het parcours. De laatste variabele meet het aantal kilometer kasseien in een parcours; een kasseistrook is immers niet altijd even lang. De gegevens voor beide variabelen werd eveneens afgeleid uit de documenten die terug te vinden zijn op websites als *Sporza* en *CyclingNews.*

2.3.8 Kijkcijfers

Een wielerploeg leeft van sponsoring. Het is daarom belangrijk goed te presteren in de juiste wedstrijden, zodat de sponsor voldoende *return on investment* krijgt. De juiste wedstrijden zijn de races die plaatsvinden in de buurt van de hoofdkantoren van de sponsor en/of koersen die veel media-aandacht krijgen. Een sponsor creëert immers *return on investment* wanneer zijn renners reclame voor hem maken door zich in de kijker te rijden.

Renners kunnen op verschillende manieren in beeld komen. Ze kunnen meegaan in een lange vlucht; de renners in de kopgroep krijgen immers veel aandacht aan het begin van de tv-uitzending. Het logo van de sponsor komt dan ook duidelijk in beeld. De grootste aandacht gaat uiteraard naar de winnaar. Een sponsor ziet zijn renners dus graag meestrijden om de overwinning in de belangrijke wedstrijden.

De hoeveelheid media-aandacht die een wedstrijd krijgt, kan gekwantificeerd worden met behulp van kijkcijfers. Het verzamelen van deze statistieken is een taak van het Centrum voor Informatie over de Media (CIM). Zij hebben hiervoor bij 1500 gezinnen in Vlaanderen en Wallonië een audi- en kijkmeter geïnstalleerd.
Het CIM mag echter niet alle kijk- en luistercijfers vrijgeven. Zij krijgen enkel toestemming voor het publiceren van de kijkcijfers van de best bekeken tv-programma's. De overige gegevens worden niet gepubliceerd, maar bijgehouden door de verschillende zenders die de kijkcijferstudies financieren. Daarom werd er voor dit onderzoek contact opgenomen met de VRT. Zij zenden al jaren alle wielerwedstrijden uit en beschikken over de kijkcijfers van deze programma's.

Het kijkgedrag in Vlaanderen wordt gekwantificeerd aan de hand van verschillende cijfers. Een eerste variabele bevat informatie over het marktaandeel van een uitzending. Deze variabele geeft het percentage kijkers voor een bepaald tv-programma weer. Enkel de gezinnen waarvan het tv-toestel aanstaat worden in rekening gebracht.

Een volgende variabele beschrijft de kijkdichtheid. Dit is het percentage Vlamingen dat naar een bepaald tv-programma kijkt. Voor het berekenen van deze percentages worden alle potentiële Vlaamse kijkers in rekening gebracht, dus ook de gezinnen waarvan het tv-toestel niet aanstaat op het moment van de bestudeerde uitzending.

De derde variabele bevat cijfers over het gemiddeld aantal kijkers van een tv-programma. Dit gemiddelde wordt berekend over de volledige duur van de bestudeerde uitzending. Het resulterende getal wordt beschouwd als het werkelijke kijkcijfer van een programma. Volgens Van Reeth (2013) vormt deze variabele een maat voor de aandacht van de echte wielerfans. Hij verwijst hiermee naar de kijkers die een wedstrijd van de eerste tot de laatste minuut volgen. Om de algemene belangstelling voor een wedstrijd te meten, stelt hij voor om piekkijkcijfers te gebruiken. Deze cijfers geven aan hoeveel kijkers er voor de buis zitten op het drukst bekeken moment van elke uitzending. Dit cijfer neemt ook kijkers in rekening die inschakelen om de spannendste of laatste kilometers van een wedstrijd te bekijken. De tijd die zij voor tv doorbrengen is korter dan bij de echte fans. Daarom wordt hun aandeel uitgevlakt bij het berekenen van het gemiddeld aantal kijkers per uitzending.

De dataset van deze masterproef bevat gegevens over het marktaandeel, de kijkdichtheid en het kijkcijfer van elke uitzending. De eerste twee variabelen geven een percentage weer. Bijgevolg hebben deze cijfers een waarde tussen 0% en 100%. De laatste variabele geeft het gemiddeld aantal kijkers van elke wedstrijd weer. Deze informatie wordt voorgesteld als een geheel getal. Het piekkijkcijfer werd niet opgenomen in dit onderzoek, omdat de VRT deze cijfers niet beschikbaar stelde.

2.3.9 Valpartijen

Een laatste variabele registreert het aantal valpartijen per wedstrijd. Bij deze telling werd geen rekening gehouden met het aantal betrokken renners in een valpartij. Dit betekent dat valpartijen met één renner, dan wel veertig renners beiden gerapporteerd worden als één valpartij.

Deze gegevens werden verzameld aan de hand van de live-feeds op de websites van *Sporza* en *Cycling News.* Deze verslagen geven informatie over de samenstelling en voorsprong van de (eventuele) kopgroep, over de hindernissen in het parcours, de weersomstandigheden tijdens de koers en over de valpartijen die plaatsvinden.

Het is mogelijk dat deze live-feeds niet alle valpartijen rapporteren, maar ook de camera's en reporters in de wedstrijd slagen er niet in om elke valpartij te registreren. De belangrijke valpartijen worden steevast gemeld. We veronderstellen dat een valpartij belangrijk is wanneer veel renners betrokken zijn of wanneer de betrokken coureur(s) ernstige verwondingen oplopen en/of hun weg niet kunnen verderzetten. Een valpartij is eveneens belangrijk wanneer ze het verdere koersverloop beïnvloedt.

We kunnen aannemen dat de valpartijen die gemist worden plaatsvinden in de achtergrond van de wedstrijd, waar de betrokken renners in een verloren positie zitten en dat deze valpartijen geen blijvende schade met zich meebrengen.

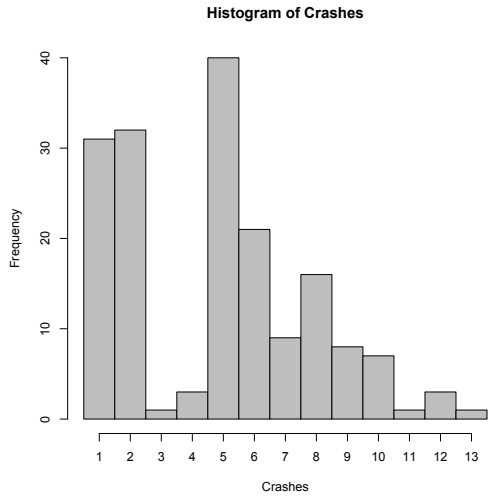
3 Univariate analyse van de variabelen

3.1 Aantal valpartijen per wedstrijd

De variabele *crashes* registreert het aantal valpartijen per wedstrijd. Het tweede hoofdstuk vermelde al dat er bij het tellen van het aantal valpartijen geen rekening gehouden werd met het aantal betrokken renners. Dit betekent dat een massale valpartij in het peloton en een val van één enkele coureur beiden als één valpartij gerekend worden.

In de bestudeerde eendagskoersen werden er één tot dertien valpartijen per wedstrijd geteld. Voor 99 van de 272 bestudeerde koersen ontbreekt de waarde van het aantal valpartijen, omdat het verslag van de wedstrijd niet teruggevonden werd of onvoldoende gedetailleerd bleek om het aantal valpartijen per wedstrijd te kunnen vaststellen. Deze observaties bleven behouden, omdat er wel informatie kon verzameld worden over de andere kenmerken van een koers. Aan de hand van deze gegevens kan het begrip eendagswedstrijd preciezer gekwantificeerd worden.

Het histogram van de variabele *crashes* is weergegeven in Figuur 1. Deze verdeling heeft een gemiddelde van 4,792 valpartijen per wedstrijd. De standaarddeviatie is gelijk aan 2,99. Deze waarde is eerder klein, want de variatiecoëfficiënt[7] heeft een waarde van 0,62 en blijft dus kleiner dan de drempelwaarde van 1.



Figuur 1 Histogram voor de variabele *crashes*

[7] De variatiecoëfficiënt meet de verhouding tussen de standaarddeviatie en het gemiddelde van een verdeling. Wanneer deze waarde kleiner is dan 1, spreken we van een verdeling met kleine variantie.

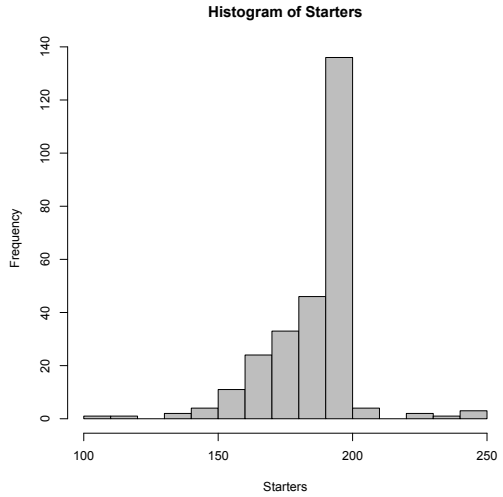
3.2 Kenmerken van het aantal deelnemers

3.2.1 Het aantal starters in een wedstrijd

3.2.1.1 Een variabele voor het aantal starters

De variabele *starters* geeft het aantal deelnemers in elke wedstrijd weer. Deze variabele weerspiegelt het aantal handtekeningen op het wedstrijdblad. Een renner die niet komt tekenen mag immers niet starten en krijgt een DNS achter zijn naam.

Het histogram van Figuur 2 toont een links-scheve verdeling. De vorm van deze verdeling wordt beïnvloed door de regels van de UCI. Het huidige reglement bepaalt immers dat een peloton niet meer dan 200 renners mag bevatten. Deze regel vormt een verklaring voor de piek in het histogram bij wedstrijden met 190 tot 200 deelnemers. De ploegen streven naar een maximale bezetting, wat resulteert in wedstrijden waarin een volledig peloton aan de start verschijnt.



Figuur 2 Histogram voor de variabele *starters*

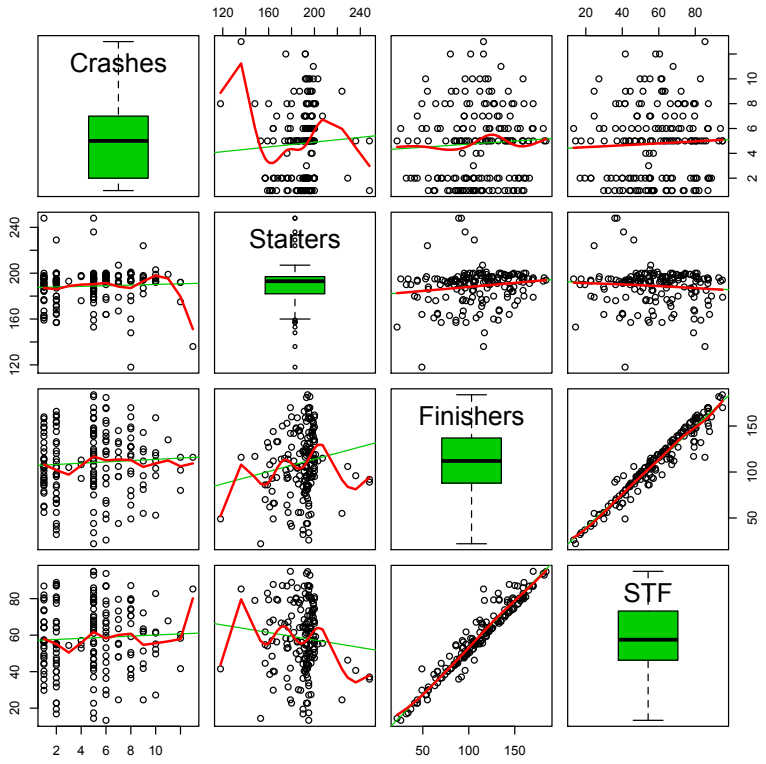
De verdeling van *starters* wordt gekenmerkt door een gemiddelde van 186,3 renners en een standaarddeviatie met een waarde van 17,16. Deze standaarddeviatie is eerder klein, want de variatiecoëfficiënt is gelijk aan 0,09.

De tweede boxplot in Figuur 3 biedt een overzicht van de overige kengetallen van *starters*. Hieruit blijkt dat de bestudeerde wedstrijden 103 tot 248 renners aan de start kregen. Enkele observaties gelden als een uitschieter[8] en worden weergegeven als een lege cirkel.
De mediaan wordt aangeduid met behulp van een zwarte horizontale streep. Dit kengetal heeft een waarde van 191,5 renners. De mediaan is dus groter dan het gemiddelde, wat een typisch kenmerk is voor een links-scheve verdeling.

[8] Een uitschieter is een boxplot is een observatie die buiten het interval $[Q_1 - \frac{3}{2}IQR, Q_3 + \frac{3}{2}IQR]$ valt, met Q_1 en Q_3 de waarden van het eerste en derde kwartiel en IQR de interkwartielafstand.

Het eerste spreidingsdiagram van Figuur 3 brengt de variabelen *crashes* en *starters* samen. Het verband tussen beide variabelen wordt weergegeven met behulp van twee trendlijnen. De groene rechte is het resultaat van een regressieanalyse. Deze techniek veronderstelt een lineair verband tussen beide variabelen. De rechte vormt echter geen goede fit voor de data, want vele observaties liggen ver van de trendlijn verwijderd. Het verband tussen het aantal starters en het aantal valpartijen is dus niet lineair.
Door de assumpties van deze regressie te versoepelen, kan een betere fit verkregen worden. Daarom werd de rode curve toegevoegd aan de grafiek. Deze is het resultaat van een niet-parametrische regressie, waarbij het softwarepakket *R* gebruik maakt van de *gamLine*-functie. De afkorting *gam* staat voor *generalized additive model* en laat de gebruiker toe om te specifiëren dat aangenomen wordt dat de responsvariabele een Poisson-verdeling volgt. Deze techniek maakt geen verdere assumpties over de vorm van het model, noch over het aantal te schatten parameters. Deze gegevens worden afgeleid uit de data. Ter compensatie vraagt deze analyse om een grotere dataset.

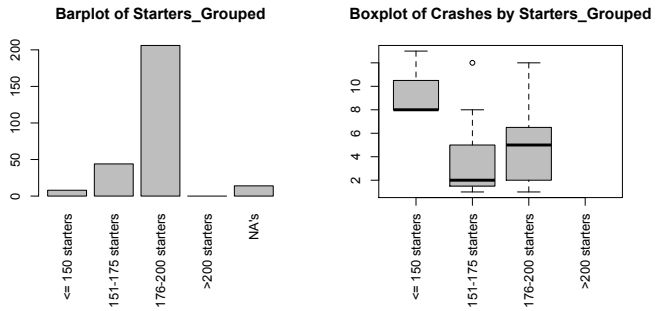
Bij een peloton met 140 tot 160 renners kent de curve een dalend verloop. Het aantal valpartijen neemt dan lineair af met de grootte van het peloton. Wanneer het aantal deelnemers toeneemt tot 170 à 200 renners gaat de trendlijn stijgen. Over dit interval neemt het aantal valpartijen per wedstrijd dus toe.



Figuur 3 Matrix-scatterplot van de variabelen *crashes, starters, finishers* en *STF*

3.2.1.2 Een categorische variabele voor het aantal starters

Het verband met de grootte van het peloton wordt verder bestudeerd aan de hand van de variabele *starters_grouped*. Het staafdiagram in Figuur 4 biedt een overzicht van de vier niveaus van deze variabele.



Figuur 4 Staafdiagram en boxplots voor de variabele *starters_grouped*

In een eerste groep wedstrijden kwamen maximaal 150 deelnemers aan de start. Dit stemt overeen met 25 teams die elk maximaal zes renners mogen afvaardigen. In deze races werd het meest gevallen. Een gemiddelde wedstrijd telt 9,667 valpartijen.
Wanneer de 25 wielerteams elk zeven renners mogen afvaardigen, bestaat een peloton uit maximaal 175 deelnemers. Deze optie wordt bestudeerd aan de hand van een tweede groep koersen waarin 151 tot 175 renners aan de start verschenen. In deze wedstrijd werd het minst gevallen; gemiddeld telt een wedstrijd amper 3,522 valpartijen.
In een volgende verzameling races kwamen 176 tot 200 renners aan de start. Deze koersen weerspiegelen het huidige UCI-reglement. Een wedstrijdorganisatie mag immers maximaal 25 ploegen uitnodigen, die elk acht renners mogen opstellen. Het peloton telt dan maximaal 200 deelnemers. In deze wedstrijden werden gemiddeld 4,835 keer per wedstrijd gevallen.
Een vierde en laatste groep koersen omvat alle wedstrijden die meer dan 200 renners aan de start verzamelden. In deze races werd er gemiddeld 5,875 keer per wedstrijd gevallen.

Om na te gaan of de gemiddeldes van deze vier wedstrijdgroepen significant verschillen, wordt een ANOVA-tabel opgesteld. De resultaten van deze analyse worden samengevat in Tabel 2.

ANOVA						
	Df.	***SS***	***MS***	***F-waarde***	***p-waarde***	
Starters_grouped	3	118,03	39,345	4,6942	0,003559	**
Residuals	169	1416,48	8,382			
TOTAAL	**172**	**1534,51**				

Tabel 2 ANOVA-tabel voor de variabele *starters_grouped*

De vijfde kolom van Tabel 2 toont de waarde van de teststatiek. Deze is gelijk aan 4,6942 en is groter dan de kritieke F-waarde van 2,658079. Deze laatste waarde stemt overeen met het 95e percentiel van een F-verdeling met respectievelijk 3 en 169 vrijheidsgraden in teller en noemer. Er wordt immers gewerkt met een significantieniveau van 5%. Deze teststatistiek wordt daarom geassocieerd met een kleine p-waarde, waardoor we de nulhypothese verwerpen.

Om de significant verschillende gemiddeldes te identificeren, wordt gebruik gemaakt van een Tukey-test. Deze test zorgt voor een paarsgewijze vergelijking van alle gemiddeldes. Deze analyse wordt samengevat in Tabel 3.

	Difference	Lower	Upper	p-value	
176-200 starters en 151-175 starters	1,312793	-0,3782209	3,003807	0,186749	
≤ 150 starters en 151-175 starters	6,144928	1,5336297	10,756225	0,0038143	***
> 200 starters en 151-175 starters	2,353261	-0,7301612	5,436683	0,1995118	
≤ 150 starters en 176-200 starters	4,832134	0,448471	9,215798	0,0243974	**
> 200 starters en 176-200 starters	1,040468	-1,6908192	3,771754	0,7562115	
> 200 starters en ≤ 150 starters	-3,791667	-8,8773787	1,294045	0,2175916	

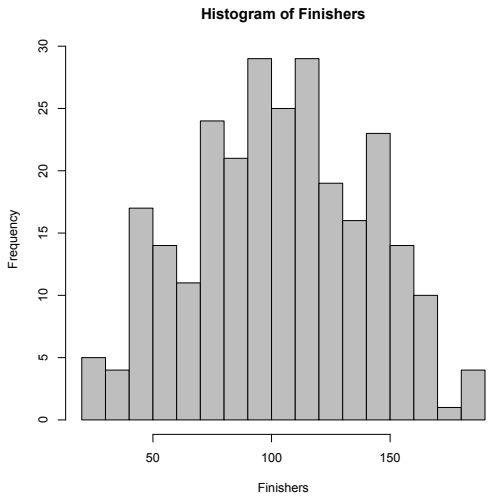
Tabel 3 Tukey-test voor de variabele *starters_grouped*

In een Tukey-test wordt het verschil tussen elk duo gemiddeldes berekend. Wanneer dit verschil significant verschillend is van nul, wordt deze combinatie geassocieerd met een p-waarde die kleiner is dan het significantieniveau van 5%.

Bovenstaande tabel bevat twee duo's met een p-waarde die kleiner is dan 5%, waaruit blijkt dat het gemiddeld aantal valpartijen voor wedstrijden met maximaal 150 deelnemers significant verschillend is van de gemiddelde hoeveelheid valpartijen in wedstrijden met respectievelijk 151 tot 175 en 176 tot 200 deelnemers.

3.2.2 Het aantal finishers

De variabele *finishers* registreert het aantal renners dat over de eindmeet rijdt binnen de vooropgestelde tijdslimiet. Dit betekent dat een renner niet finisht wanneer hij onderweg beslist om op te geven, of wanneer hij buiten tijd eindigt.



Figuur 5 Histogram voor de variabele *finishers*

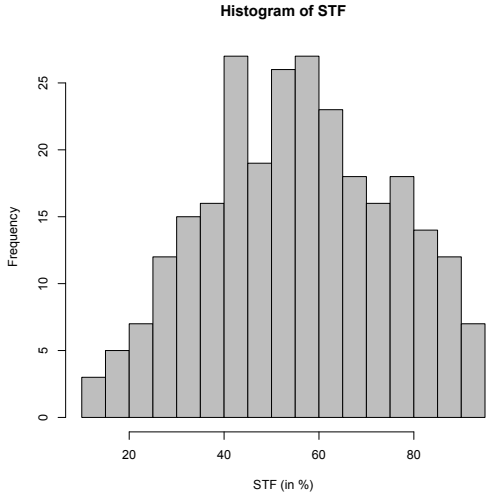
De overeenkomstige boxplot uit Figuur 3 bundelt de belangrijkste kengetallen van *finishers.* Uit deze grafiek kunnen we afleiden dat de bestudeerde koersen 22 tot 184 finishers tellen. Omdat de boxplot symmetrisch is, zijn het gemiddelde en de mediaan weinig verschillend. Deze kengetallen hebben een respectievelijke waarde van 103,6 en 104 renners. Dit kleine verschil wijst op een verdeling die niet scheef is. Deze veronderstelling wordt bevestigd door het histogram in Figuur 5, waarin een haast symmetrische verdeling wordt afgebeeld. Deze verdeling heeft een standaarddeviatie van 36,76. Wanneer we deze waarde afzetten tegen het gemiddelde, vinden we een variatiecoëfficiënt van 0,35. Deze waarde is ruim kleiner dan 1 en wijst dus op een kleine standaardfout.

Het tweede spreidingsdiagram van Figuur 3 brengt de variabelen *crashes* en *finishers* samen. Het verband tussen beide variabelen wordt verduidelijkt aan de hand van de groene rechte en rode niet-parametrische curve die eerder besproken werden. Deze trendlijnen hebben beiden een kleine positieve trend. De hoeveelheid valpartijen lijkt dus toe te nemen met het aantal finishers. Voor beide trendlijnen is er echter weinig visueel verband merkbaar. Het positieve verband tussen deze twee variabelen moet dan ook voorzichtig geïnterpreteerd worden.

3.2.3 Verhouding tussen het aantal starters en het aantal finishers

De variabele *STF* meet de verhouding tussen het aantal finishers en starters in een bepaalde wedstrijd. Dit cijfer wordt uitgedrukt als een percentage met een waarde van 0% tot 100%. Deze variabele heeft een sterk positieve correlatie met *finishers*. Dit uit zich in een lineair verband tussen *finishers* en *STF* in Figuur 3.

Het histogram van de variabele *STF* wordt weergegeven in Figuur 6. Het gemiddelde en de standaarddeviatie hebben een respectievelijke waarde van 55,69 % en 19,26%. Deze standaarddeviatie is klein, want de corresponderende variatiecoëfficiënt (COV[9] = 0,34) heeft een waarde die kleiner is dan 1.



Figuur 6 Histogram voor de variabele *STF*

[9] COV staat voor *Coefficient of Variance*, een Engelse afkorting voor variatiecoëfficiënt

De laatste boxplot in Figuur 3 geeft de overige kencijfers van *STF* weer. Deze grafiek toont aan dat de waardes variëren van 12,64% tot 94,97%. Er is dus geen enkele wedstrijd waarin alle gestarte renners de eindmeet halen. Verder toont de figuur aan dat de mediaan, die aangeduid wordt met een horizontale zwarte streep, gelijk is aan 55,44%. De mediaan is dus iets kleiner dan het gemiddelde, maar het verschil tussen beide kengetallen is zo klein dat ze nagenoeg samenvallen. Dit is een typisch kenmerk voor een symmetrische verdeling.

In het derde spreidingsdiagram bovenaan in Figuur 3 wordt de variabele *STF* in verband gebracht met de responsvariabele *crashes*. Ook dit diagram wordt aangevuld met de twee eerder besproken trendlijnen. Beide trendlijnen vertonen een licht positieve helling, net als bij de variabele *finishers*. Dit gelijkaardige resultaat is te wijten aan de hoge correlatie tussen de variabelen. Helaas bieden beide trendlijnen opnieuw weinig verband met de data.

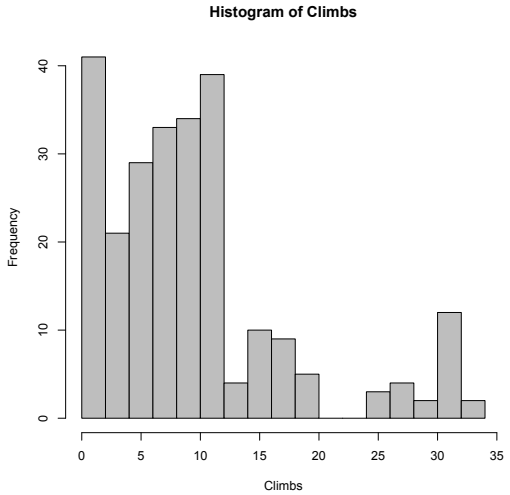
3.3 Kenmerken van het parcours

3.3.1 De hellingen in het parcours

3.3.1.1 Een variabele voor het aantal hellingen

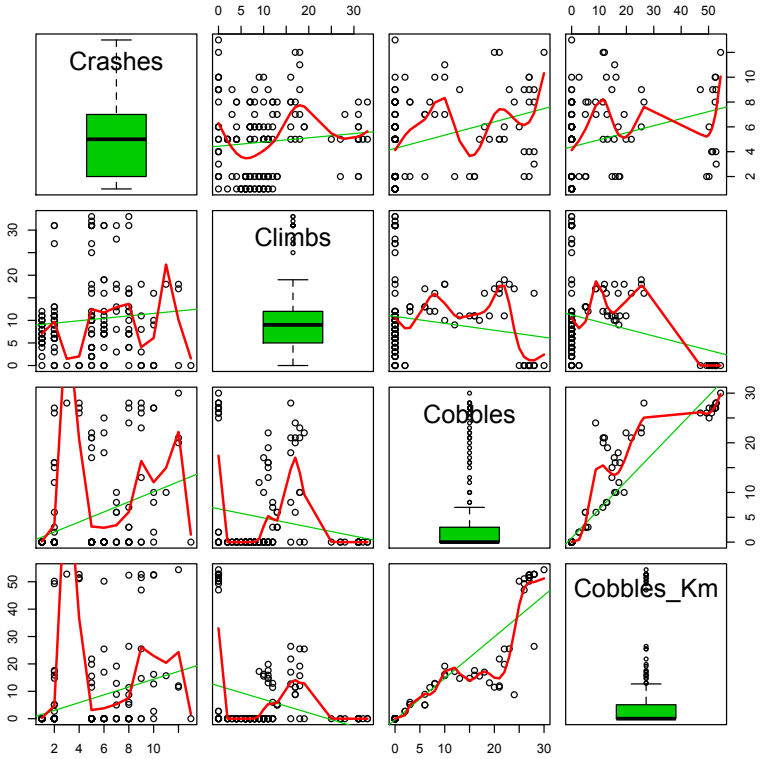
Om de kenmerken van het parcours te capteren, wordt gebruik gemaakt van drie variabelen. Eén van deze variabelen noteert het aantal hellingen in de wedstrijdroute. Deze variabele kreeg de naam *climbs*.

Het aantal hellingen in een parcours varieert van wedstrijd tot wedstrijd. In koersen als Parijs-Roubaix en de Scheldeprijs moet er niet of nauwelijks geklommen worden. In de Amstel Gold Race daarentegen, krijgen de renners gemiddeld 30,47 hellingen voor de wielen geschoven.
Gemiddeld bevatten de bestudeerde races 9,79 hellingen. De standaardfout van deze verdeling is gelijk aan 8,00. Deze fout is klein; de variatiecoëfficiënt blijft met een waarde van 0,82 immers beneden de vooropgestelde drempelwaarde van 1.



Figuur 7 Histogram voor de variabele *climbs*

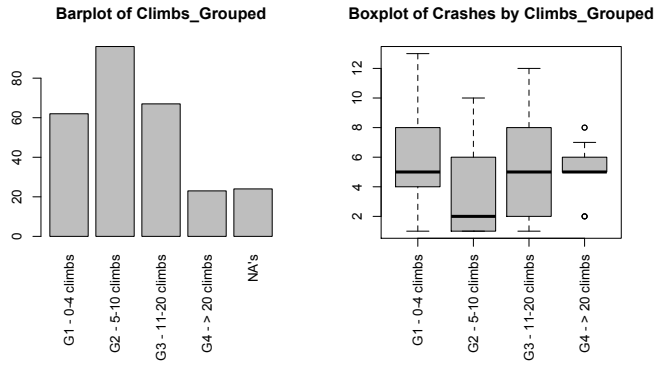
Het eerste spreidingsdiagram in Figuur 8 toont het verband tussen de variabelen *crashes* en *climbs*. Dit diagram wordt opnieuw aangevuld met twee trendlijnen. De groene rechte is het resultaat van een lineaire regressie. Vele observaties liggen echter ver boven of beneden de rechte; het verband tussen *crashes* en *climbs* verloopt dus niet lineair.
De rode curve geeft de resultaten van de eerder besproken niet-parametrische regressie weer. Hierin wordt het verband tussen de variabelen verder verfijnd. Voor wedstrijden die minder dan vijf hellingen bevatten, kent deze trendlijn een dalend verloop. Wanneer de organisatie meer hellingen toevoegt, dan neemt het aantal valpartijen sterk toe. Voor koersen met een twintigtal hellingen bereikt de curve een lokaal maximum. Wanneer het aantal heuvels verder opgedreven wordt, neemt het aantal valpartijen opnieuw af.



Figuur 8 Matrix-scatterplot van de variabelen *crashes, climbs, cobbles* en *cobbles_km*

3.3.1.2 *Een categorische variabele voor het aantal hellingen*

De variabele *climbs_grouped* wijst elke koers toe aan één van de vier wedstrijdgroepen die onderscheiden werden op basis van het histogram in Figuur 7. Het staafdiagram aan de rechterkant van Figuur 9 biedt een overzicht van de vier groepen en de hoeveelheid races die ze bevatten.



Figuur 9 Staafdiagram en boxplots voor de variabele *climbs_grouped*

De eerste categorie wordt gecodeerd als G1 en bevat de wedstrijden met minder dan vijf hellingen. De koersen met vijf tot tien hellingen krijgen het label G2. In de derde groep (G3) vinden we wedstrijden met elf tot en met twintig hellingen. De vierde en laatste verzameling koersen (G4) fungeert als een restcategorie. Hier worden de wedstrijden met meer dan twintig hellingen samengebracht.

De boxplots in Figuur 9 visualiseren de belangrijkste kengetallen van de variabele *crashes* per wedstrijdgroep. Uit deze grafiek kan afgeleid worden dat de medianen sterk verschillen. Voor de gemiddeldes komen we tot een zelfde vaststelling. De koersen in groep G1 tellen gemiddeld 5,744 valpartijen. In groep G2 gaan de renners minder vaak tegen het asfalt. Een gemiddelde wedstrijd telt immers 3,691 valpartijen. In groep G3 is het gemiddelde gelijk aan 5,644 valpartijen per wedstrijd. Voor groep G4 vinden er gemiddeld 5,176 valpartijen per wedstrijd plaats.

Om na te gaan of de verschillen tussen de gemiddeldes het gevolg zijn van het spelen van het toeval, wordt gebruik gemaakt van een ANOVA-tabel. De resultaten van deze analyse worden samengevat in Tabel 4.

ANOVA						
	Df.	***SS***	***MS***	***F-waarde***	***p-waarde***	
Climbs_Grouped	3	152,63	50,876	6,2056	0,000509	***
Residuals	165	1352,73	8,198			
TOTAAL	**168**	**1505,36**				

Tabel 4 ANOVA-tabel voor de variabele *climbs_grouped*

In deze tabel vinden we een p-waarde van 0,000509. Deze waarde is veel kleiner dan het significantieniveau van 5%, wat betekent dat de teststatistiek ruim groter is dan de kritieke F-waarde. Minstens twee gemiddeldes vertonen dus een significant verschil.

Om na te gaan welk van deze gemiddeldes significant afwijkt, wordt opnieuw gebruik gemaakt van een Tukey-test. De resultaten van deze analyse worden getoond in Tabel 5.

	Difference	Lower	Upper	p-value	
G2 en G1	-2,0524133	-3,5451192	-0,5597074	0,0026271	***
G3 en G1	-0,0991453	-1,7249568	1,5266662	0,9985831	
G4 en G1	-0,5671192	-2,7268835	1,5926452	0,9040084	
G3 en G2	1,9532680	0,5252041	3,3813318	0,0028017	***
G4 en G2	1,4852941	-0,5298201	3,5004083	0,2265624	
G4 en G3	-0,4679739	-2,5835772	1,6476295	0,9396912	

Tabel 5 Tukey-test voor de variabele *climbs_grouped*

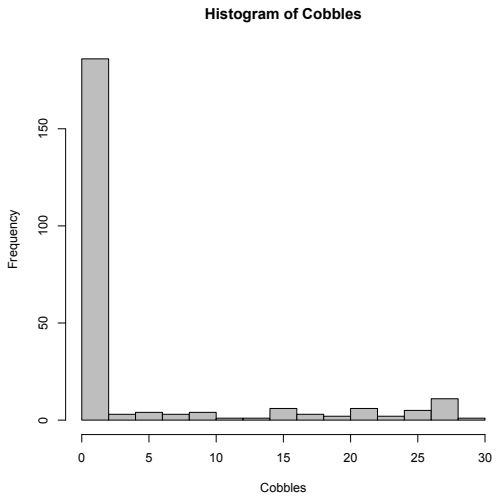
Deze tabel bevat twee p-waardes die kleiner zijn dan het significantieniveau van 5%. Dit betekent dat de gemiddelde hoeveelheid valpartijen in koersen uit groep G2 significant verschilt van het gemiddeld aantal valpartijen in wedstrijden uit de groepen G1 en G3.

3.3.2 De kasseien in het parcours

3.3.2.1 Een variabele voor het aantal kasseistroken

Naast heuvels en hellingen krijgt het peloton nog andere obstakels voorgeschoteld. Wedstrijden in Vlaanderen en Noord-Frankrijk onderscheiden zich door de aanwezigheid van kasseien in de wedstrijdroute. Het aantal kasseistroken in het parcours wordt bijgehouden door de variabele *cobbles*.

De variabelen *cobbles* en *climbs* hebben een negatieve correlatiecoëfficiënt. Dit betekent dat er een trade-off bestaat tussen het aantal hellingen en kasseistroken in een wedstrijd. Een koers die meer kasseien bevat, telt minder hellingen, en omgekeerd.



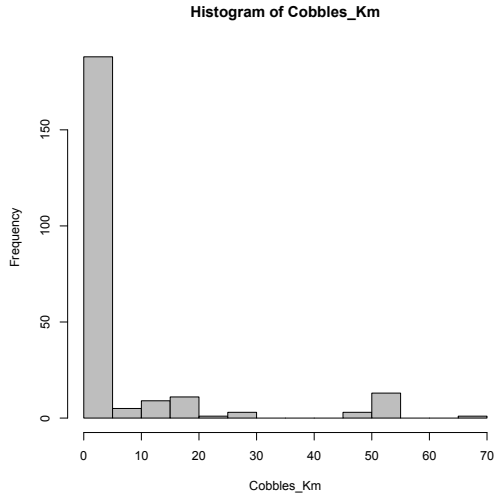
Figuur 10 Histogram voor de variabele *cobbles*

Een analyse van de percentielen van de variabele *cobbles* brengt aan het licht dat driekwart van de bestudeerde koersen geen enkele kasseistrook bevat. Deze vaststelling vormt een verklaring voor het eerder lage gemiddelde van 3,979 kasseistroken per wedstrijd. Door dit grote aantal wedstrijden zonder kasseien vertoont het histogram van Figuur 10 één grote piek, gevolgd door een lange staart van andere observaties. De vorm van deze grafiek doet vermoeden dat de verdeling een grote variantie heeft. De standaardfout van *cobbles* heeft een waarde van 8,42. Deze waarde blijkt inderdaad groot te zijn, want de variatiecoëfficiënt heeft een waarde van 2,12 en is dus ruim groter dan 1.

Het tweede spreidingsdiagram in Figuur 8 brengt de variabelen *cobbles* en *crashes* samen in één plot. Het verband tussen deze variabelen wordt verduidelijkt met een regressierechte. Deze trendlijn veronderstelt een lineair verband tussen *cobbles* en *crashes*. De rechte wordt gekenmerkt door een positieve richtingscoëfficiënt. Dit betekent dat de hoeveelheid kasseistroken het aantal valpartijen positief beïnvloedt. De visuele fit stelt echter teleur. Vele observaties bevinden zich ver van de groene regressierechte. Het verband tussen *cobbles* en *crashes* is naar alle waarschijnlijkheid niet lineair.

3.3.2.2 Een variabele voor het aantal kilometer kasseien

Een kasseistrook heeft geen uniforme lengte. Daarom wordt de hoeveelheid kasseien in een wedstrijdparcours ook op een tweede manier in rekening gebracht. De variabele *cobbles_km* meet de hoeveelheid kasseien in een parcours, uitgedrukt in kilometer.
Deze variabele is sterk positief gecorreleerd met de variabele *cobbles*. Dit hoeft niet te verwonderen, want beide variabelen meten een zelfde aspect van het wedstrijdparcours. Door de hoge correlatie lijkt de vorm van de verdeling van *cobbles_km* sterk op die van de variabele *cobbles*. Het histogram in Figuur 11 lijkt wel een kopie van het histogram van de variabele *cobbles* in Figuur 10. Het grootste verschil zijn de waardes op de horizontale as van de grafiek.



Figuur 11 Histogram voor de variabele *cobbles_km*

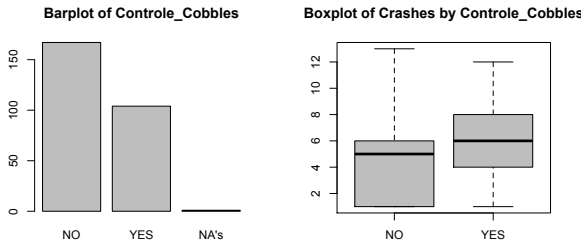
Ook de matrix-scatterplot van Figuur 8 weet de sterk positieve correlatie tussen *cobbles* en *cobbles_km* te visualiseren. Het spreidingsdiagram van de twee variabelen wordt aangevuld met een groene regressierechte en rode niet-parametrische curve. De observaties liggen op een rechte met positieve richtingscoëfficiënt, wat wijst op een sterk positieve correlatie tussen de twee variabelen.

De matrix-scatterplot benadrukt de sterke correlatie tussen de variabelen nog een tweede keer. Het spreidingsdiagram van de variabelen *cobbles_km* en *crashes* wordt immers gekenmerkt door twee trendlijnen met een patroon dat gelijkaardig is aan dat van de trendlijnen in de grafiek van *crashes en cobbles*.

3.3.2.3 Een categorische variabele voor de aanwezigheid van kasseien in het parcours

De variabele *controle_cobbles* is een factorvariabele met twee niveaus. Een wedstrijd waarin één of meerdere kasseistroken voorkomen, krijgt de waarde *yes.* Voor de overige koersen neemt deze variabele de waarde *no* aan.

De rechterkant van Figuur 12 beeldt twee boxplots af. De eerste boxplot visualiseert de kengetallen van de variabele *crashes* voor wedstrijden waarin geen kasseien voorkomen. De tweede boxplot biedt een zelfde overzicht voor de kasseikoersen. De eerste groep heeft een gemiddelde van 3,991 valpartijen per wedstrijd. Voor de kasseikoersen ligt het gemiddelde iets hoger. In deze groep telt een wedstrijd gemiddeld 6,06 valpartijen.



Figuur 12 Staafdiagram en boxplots voor de variabele *controle_cobbles*

Om na te gaan of deze gemiddeldes significant verschillend zijn, wordt gebruik gemaakt van een tweezijdige Welch-test. Er wordt aangenomen dat de data normaal verdeeld zijn en dat de varianties van beide groepen niet gelijk zijn. De waarde van de teststatistiek wordt vervolgens vergeleken met de kritieke t-waarde. Deze waarde komt uit een Student t-verdeling waarvan het aantal vrijheidsgraden (*df*) bepaald wordt met uitdrukking (1).

$$df = \frac{\left(s_1^2/n_1\right)^2 + \left(s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1) + \left(s_2^2/n_2\right)^2/(n_2-1)} \quad (1)$$

De waardes s_1 en s_2 verwijzen naar de standaarddeviaties van de bestudeerde groepen. Met n_1 en n_2 wordt het aantal observaties in groep 1, respectievelijk groep 2 aangeduid. Wanneer we deze Welch-test uitvoeren voor de variabele *controle_cobbles*, bekomen we de resultaten die weergegeven worden in Tabel 6.

	t-value	Df.	Conf. Int.	Conf. Int.	p-value	
			Upper	***Lower***		
Tweezijdige test	-4,5528	125,624	-2,968560	6,059701	1,234e-05	***

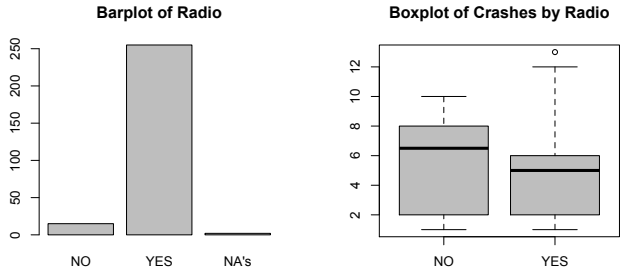
Tabel 6 Significantietest voor de variabele *controle_cobbles*

De nulhypothese van deze tweezijdige toets veronderstelt dat beide gemiddeldes gelijk zijn. Wanneer beide gemiddeldes significant verschillen, moet de absolute waarde van de teststatistiek groter zijn dan het 97,5e percentiel van een Student t-verdeling met 125,624 vrijheidsgraden. Deze kritieke t-waarde heeft een waarde van 1,979028. De teststatistiek in de eerste kolom van Tabel 6 heeft een waarde van -4,5528. De absolute waarde van deze teststatistiek is dus ruim groter dan de kritieke t-waarde. Daarom wordt deze geassocieerd met een zeer kleine p-waarde, waardoor de nulhypothese verworpen kan worden. We besluiten dat beide gemiddeldes significant verschillend zijn.

3.4 Gebruik van radiocommunicatie

In 2011 besliste de UCI het gebruik van oortjes te beperken. Sindsdien zijn de oortjes enkel toegelaten in wedstrijden van het World Tour niveau. Deze regel was een heuse vernieuwing in het wielrennen. Voorheen waren de oortjes immers in alle wedstrijden toegelaten en dit sinds de intrede van de radiocommunicatie aan het eind van de jaren '90.

Met behulp van het staafdiagram in Figuur 13 kunnen de wedstrijden over twee groepen verdeeld worden. De eerste balk bevat vijftien wedstrijden die na 2011 afgewerkt werden en waarin het gebruik van oortjes niet langer toegestaan werd omdat de koers niet tot het World Tour niveau behoorde. De tweede balk bevat de duidelijke meerderheid van de wedstrijden. In deze wedstrijden werd er met oortjes gekoerst. De meerderheid van de bestudeerde koersen bevindt zich hier, omdat de beslissing tot het verbieden van oortjes pas doorgevoerd werd in 2011. Tot slot zijn er nog twee wedstrijden die afgelast werden omwille van sneeuwval[10] . Deze wedstrijden worden aangeduid met de waarde NA.



Figuur 13 Staafdiagram en boxplots voor de variabele *radio*

De rechterkant van Figuur 13 toont twee boxplots. De linker-boxplot biedt een overzicht van de kengetallen voor de koersen waarin oortjes niet toegestaan werden. In deze wedstrijden werden er gemiddeld 5,929 valpartijen per wedstrijd geregistreerd. De tweede boxplot biedt een overzicht van het aantal valpartijen in wedstrijden waarin met oortjes gekoerst werd. Het maximaal aantal valpartijen ligt hier hoger, maar het gemiddelde is lager. In deze wedstrijden werd gemiddeld 4,692 keer gevallen.
Om na te gaan of deze gemiddeldes een significant verschil vertonen, worden opnieuw een tweezijdige t-test uitgevoerd. De resultaten van deze analyse worden weergegeven in Tabel 7.

	t-value	Df.	Conf. Int.	Conf. Int.	p-value
			Upper	***Lower***	
Tweezijdige test	1,436	15,172	-0,5971469	3,0706419	0,1713

Tabel 7 Significantietest voor de variabele *radio*

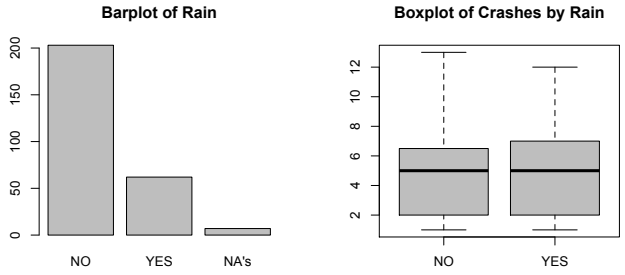
De teststatistiek is gelijk aan 1,436. Deze waarde is kleiner dan de kritieke t-waarde, die overeenstemt met het 97,5e percentiel van een Student t-verdeling met 15,172 vrijheidsgraden. De p-waarde is groter dan het significantieniveau van 5%, zodat de nulhypothese niet verworpen kan worden. Dit betekent dat de oortjes geen significante invloed hebben op de hoeveelheid valpartijen per wedstrijd.

[10] Het gaat hier om de Omloop het Nieuwsblad uit 2004 en Kuurne-Brussel-Kuurne uit 2013

3.5 Weersomstandigheden

Het peloton is een rijdend circus; op enkele uren tijd leggen de renners 170 tot 300 km af. Deze eigenschap maakt het verzamelen van informatie over de weersomstandigheden tijdens de wedstrijd een lastige taak. Eerder in deze masterproef werd daarom uitgelegd waarom enkel geregistreerd werd of er neerslag viel tijdens de wedstrijd. Dit gebeurt met behulp van de variabele *rain*.

Het staafdiagram links in Figuur 14 telt opnieuw twee groepen. De eerste balk komt overeen met 203 wedstrijden waarin het droog bleef. De volgende staaf bestaat uit 62 races waarin het peloton een regenjasje nodig had om zich te beschermen tegen een of andere vorm van neerslag. Er werd immers geen onderscheid gemaakt voor regen, hagel of sneeuw. De derde balk bestaat uit zeven wedstrijden met de waarde NA. Twee races werden afgelast omwille van de sneeuw. Voor de overige vijf wedstrijden kon niet achterhaald worden of er neerslag viel tijdens het verloop van de wedstrijd.



Figuur 14 Staafdiagram en boxplots voor de variabele *rain*

In Figuur 14 staan twee boxplots afgebeeld. De eerste boxplot biedt een overzicht van de kengetallen van het aantal valpartijen in een wedstrijd waarin het niet regende, sneeuwde of hagelde. Voor het gemiddelde vinden we een waarde van 4,734 valpartijen per wedstrijd. De tweede boxplot toont de kengetallen voor de wedstrijden waarin er wel neerslag viel. In deze groep wedstrijden werd iets meer gevallen. We tellen gemiddeld 4,939 valpartijen per wedstrijd.

Tabel 8 toont een overzicht van de resultaten van de tweezijdige t-test die uitgevoerd werd om de gemiddeldes van de twee groepen te vergelijken. Er wordt hierbij gebruik gemaakt van een significantieniveau van 5%.

	t-value	Df.	Conf. Int.	Conf. Int.	p-value
			Upper	***Lower***	
Tweezijdige test	-0,4037	87,242	-1,213652	0,803843	0,6874

Tabel 8 Significantietest voor de variabele *rain*

De nulhypothese van deze test stelt dat beide gemiddeldes gelijk zijn. Om van een significant verschil te spreken, moet de teststatistiek in het kritieke gebied liggen. Dit zijn alle waarden die niet tot het interval [-1,987531; 1,987531] behoren.

De tabel leert ons dat de teststatistiek gelijk is aan -0,4037. Deze waarde behoort niet tot het kritieke gebied. De overeenkomstige p-waarde is groter dan het significantieniveau van 5%, waardoor de nulhypothese niet kan verworpen worden. Ook de regen heeft geen significante invloed op het aantal valpartijen per wedstrijd.

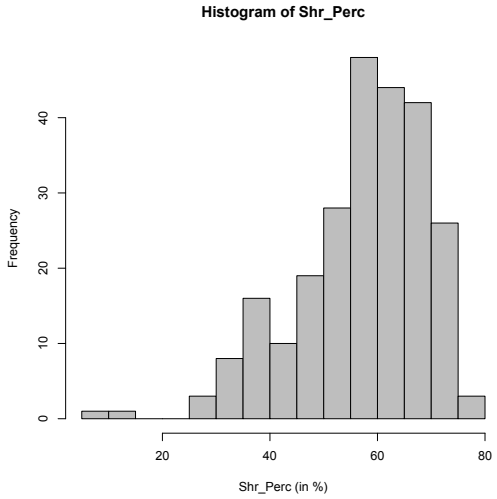
3.6 Kijkcijfers

3.6.1 Het marktaandeel

Het marktaandeel van een uitzending wordt gedefinieerd als het percentage kijkers dat naar dit programma kijkt. De populatie bestaat uit de kijkers van dat moment; dit betekent dat gezinnen waarvan het tv-toestel uitgeschakeld is niet in rekening genomen worden.

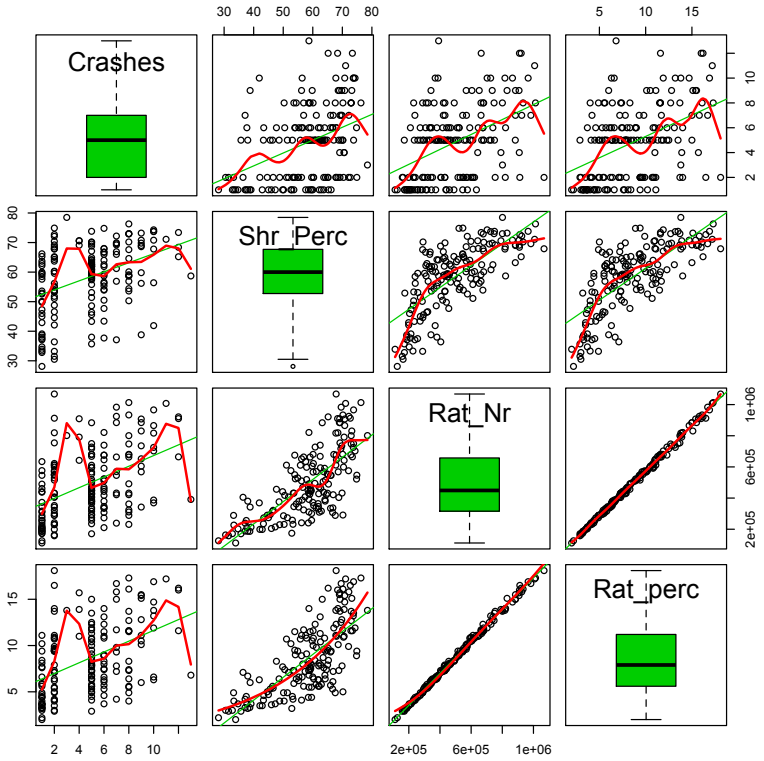
Het marktaandeel van de bestudeerde wieleruitzendingen varieert van 5,40% tot 78,50%. Het eerste percentage stemt overeen met de Scheldeprijs van 1998. Deze wedstrijd werd op een weekdag afgewerkt, waardoor ze waarschijnlijk minder kijkers lokte. Het tweede percentage werd gemeten op Paaszondag 12 april 2009. Op die dag behaalde Tom Boonen zijn derde overwinning in Parijs-Roubaix, wat heel wat kijkers overtuigde om op *Sporza* af te stemmen.

Het histogram van Figuur 15 toont de links-scheve verdeling van de variabele *shr_perc*. De standaarddeviatie van deze verdeling is klein, want de variatiecoëfficiënt is gelijk aan 0,21 en blijft dus kleiner dan 1. Deze links-scheve verdeling zorgt er ook voor dat de mediaan groter is dan het gemiddelde. Voor de mediaan vinden we een waarde van 59,10%, terwijl het gemiddelde gelijk is aan 57,15%.



Figuur 15 Histogram van de variabele *shr_perc*

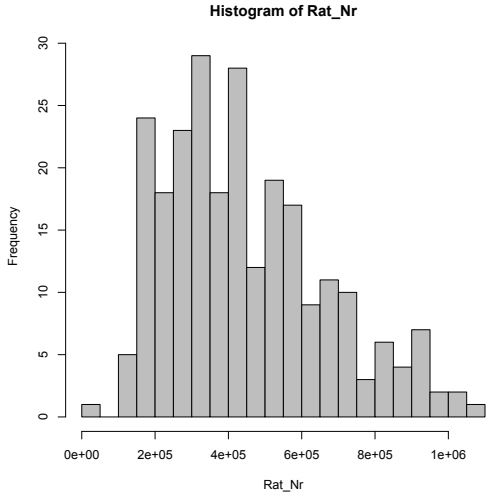
Het eerste spreidingsdiagram van Figuur 16 vindt een positief verband tussen de variabelen *shr_perc* en *crashes.* De groene regressierechte en rode niet-parametrische regressiecurve worden beiden gekenmerkt door dezelfde positieve trend. Dit resultaat kan betekenen dat het aantal valpartijen toeneemt in koersen die meer media-aandacht krijgen. Of, omgekeerd, dat er meer mensen op een wieleruitzending afstemmen wanneer ze horen dat de renners regelmatig vallen.



Figuur 16 Matrix-scatterplot van de variabelen *crashes, shr_perc, rat_nr* en *rat_perc*

3.6.2 Het kijkcijfer

Het werkelijke kijkcijfer van een wielerwedstrijd wordt genoteerd met de variabele *rat_nr*. Deze variabele geeft het gemiddeld aantal kijkers voor elke wieleruitzending weer. Eerder in deze masterproef werd al beschreven dat dit gemiddelde berekend wordt over de volledige duur van het tv-programma.



Figuur 17 Histogram van de variabele *rat_nr*

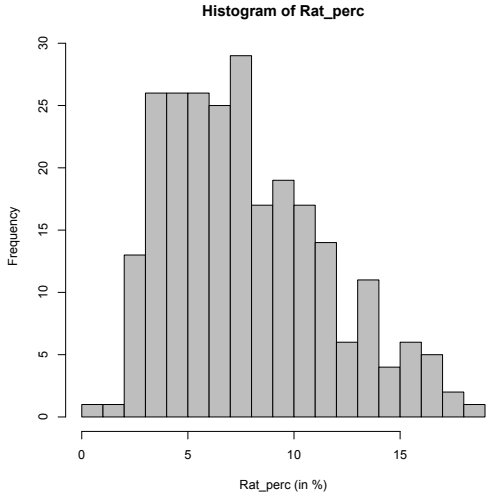
Het histogram van de variabele *rat_nr* wordt weergegeven in Figuur 17. In tegenstelling tot het marktaandeel blijkt het gemiddeld aantal kijkers een rechts-scheve verdeling te volgen. Het gemiddelde is gelijk aan 448.000 kijkers. De standaarddeviatie heeft een waarde van 216.086,7 kijkers.

Het spreidingsdiagram van de variabelen *crashes* en *rat_nr* uit Figuur 16 toont aan dat het aantal valpartijen toeneemt met het gemiddeld aantal kijkers. De groene regressierechte en de rode niet-parametrische curve worden immers beiden gekenmerkt door een stijgend verloop.

3.6.3 De kijkdichtheid

De kijkdichtheid van een tv-programma wordt gedefinieerd als het percentage Vlamingen dat naar dit programma kijkt. Voor het berekenen van dit percentage worden alle Vlamingen in rekening gebracht, ook de gezinnen waar er niet naar tv gekeken wordt op het moment van de meting. Deze informatie wordt verzameld in de variabele *rat_perc*.

Het percentage Vlamingen dat de bestudeerde wedstrijden via televisie volgde, varieert van 0,5% tot 18,10%. Deze cijfers zijn hoger dan in het buitenland. De koers is immers nergens zo populair als in Vlaanderen (Van Reeth, 2013). Een wieleruitzending trekt de aandacht van gemiddeld 7,856% van de Vlamingen. Het histogram van Figuur 18 toont aan dat de verdeling van deze variabele rechts-scheef is, net als de verdeling van het gemiddeld aantal kijkers van een wedstrijd. De standaarddeviatie van de kijkdichtheid is klein; deze heeft een waarde van 3,74. De variatiecoëfficiënt is gelijk aan 0,48.



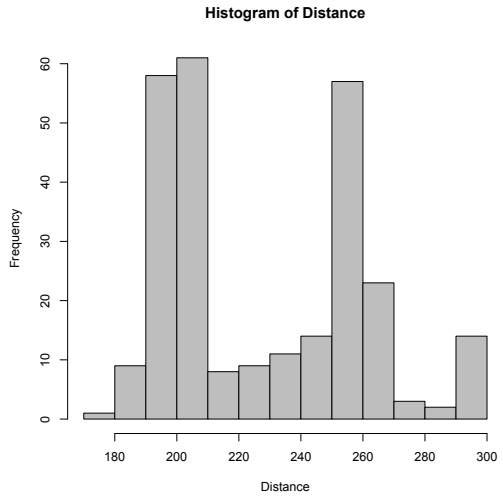
Figuur 18 Histogram voor de variabele *rat_perc*

De rechterbovenhoek van Figuur 16 bevat het spreidingsdiagram van de variabelen *rat_perc* en *crashes*. De groene rechte en rode curve hebben een positief verloop. Bovendien vertonen de curves sterke gelijkenissen met trendlijnen die gevonden werden voor de variabele *rat_nr*. Dit is geen verrassend resultaat, daar de variabelen een sterk positieve correlatie vertonen, zoals te zien op de off-diagonal van de matrix-scatterplot van Figuur 16.

3.7 Lengte van een wedstrijd

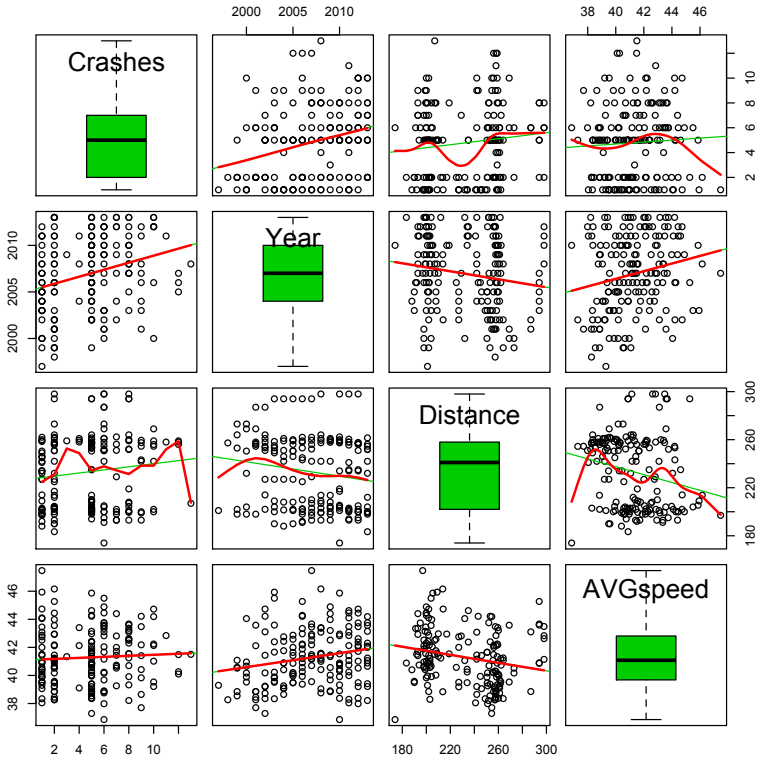
Uit het overzicht van de kengetallen van *distance* blijkt dat de lengte van de bestudeerde races varieert van 174 km (Kuurne-Brussel-Kuurne in 2010) tot 298 km (Milaan-Sanremo in de periode 2008-2012). Gemiddeld moeten de renners in één wedstrijd 228,6 km afleggen.

De verdeling van de lengte van de wedstrijden wordt weergegeven in Figuur 19. Deze grafiek vertoont twee pieken. De eerste piek bevat de wedstrijden met een lengte van 180 km tot 210 km. Dit zijn de wedstrijden uit het continentale circuit. In het huidige reglement mogen deze koersen immers niet meer dan 200 km lang zijn. De tweede piek situeert zich bij de races met een lengte van ongeveer 260 km. In deze groep vinden we de wedstrijden van het hoogste niveau terug. Voor deze races zijn er geen beperkingen op het aantal kilometers dat moet worden afgelegd.



Figuur 19 Histogram van de variabele *distance*

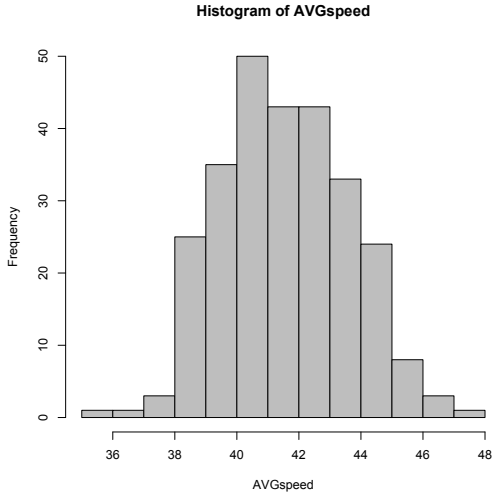
Uit de matrix-scatterplot van Figuur 20 kan afgeleid worden dat het aantal valpartijen in een wedstrijd mee-evolueert met de af te leggen afstand. De regressierechte vertoont echter geen goede fit met de data. De niet-parametrische regressie schuift de assumpties van lineariteit opzij. Het resultaat van deze analyse wordt weergegeven door de rode curve. Op basis van deze trendlijn kunnen drie wedstrijdgroepen onderscheiden worden. Een eerste groep bevat wedstrijden met een lengte tot 200 km. In deze races wordt er mogelijk nerveuzer gekoerst, waardoor het aantal valpartijen toeneemt. Het is ook mogelijk dat de renners de wedstrijd van begin af aan hard maken, wat mogelijk tot meer valpartijen leidt. Voor koersen met een lengte van 200 tot 240 km kent de trendlijn een dalend verloop. Dit betekent dat het aantal valpartijen per wedstrijd afneemt. In langere races neemt het aantal valpartijen opnieuw toe. Deze afstand stemt overeen met de lengte van de wielermonumenten. Deze wedstrijden hebben het grootste aanzien in het peloton. Er wordt dan ook op het scherpst van de snee gekoerst, wat mogelijk tot meer valpartijen leidt. Ook is het mogelijk dat de renners in een langere race langer blootgesteld worden aan het risico op een valpartij en daardoor vaker in aanraking komen met het asfalt. Het effect zwakt echter af van zodra een wedstrijd meer dan 260 km lang is.



Figuur 20 Matrix-scatterplot van de variabelen *crashes, date, distance* en *AVGspeed*

3.8 Gemiddelde snelheid van de winnaar

De variabele *AVGspeed* registreert de gemiddelde snelheid waarmee de winnaar zijn koers afwerkte. Deze informatie wordt uitgedrukt in kilometer per uur.
Het histogram van *AVGspeed* wordt weergegeven in Figuur 21. Het globale gemiddelde is gelijk aan 41,57 km/u. De standaarddeviatie heeft een waarde van 2,06 km/u. Dit is een kleine standaardafwijking, want de variatiecoëfficiënt heeft een waarde van amper 0,05.



Figuur 21 Histogram van de variabele *AVGspeed*

De matrix-scatterplot in Figuur 20 brengt *AVGspeed* in verband met drie andere variabelen. Op de laatste rij wordt *AVGspeed* weergegeven in functie van de tijd. De twee trendlijnen in dit diagram bevestigen de bevindingen van Perneger (2010), Lodewijkx en Brouwer (2012) en El Helou et al. (2012). De gemiddelde snelheid van de winnaar is de laatste jaren alleen maar toegenomen, ook in de eendagswedstrijden.

Het derde spreidingsdiagram in deze figuur toont het verband tussen *AVGspeed* en *crashes*. De groene rechte heeft een licht positieve hellingsgraad. Deze trendlijn biedt echter geen goede visuele fit voor de data, want verschillende observaties liggen ver van deze rechte verwijderd. De rode curve toont een niet-lineair verband waaruit we kunnen afleiden dat er weinig valpartijen per wedstrijd plaatsvinden wanneer de winnaar een gemiddelde snelheid van ongeveer 38 km/u tot 41 km/u haalt. Wanneer hij de koers met een hogere gemiddelde snelheid weet af te werken, dan neemt ook het aantal valpartijen toe. Er wordt het meest gevallen bij een snelheid van ongeveer 42 km/u. Daarna vlakt het effect af.

3.9 Evolutie van het aantal valpartijen over de tijd

De variabele *date* registreert de datum waarop de verschillende wedstrijdedities gereden werden. Aan de hand van deze informatie kan de tijdsevolutie van het aantal valpartijen per wedstrijd onderzocht worden.

3.9.1 Evolutie van het aantal valpartijen over de jaren heen

Een eerste analyse gaat het verband tussen het jaar waarin een koers gereden werd en het aantal valpartijen per wedstrijd na. We maken hiervoor gebruik van de informatie van de numerieke variabele *year.*

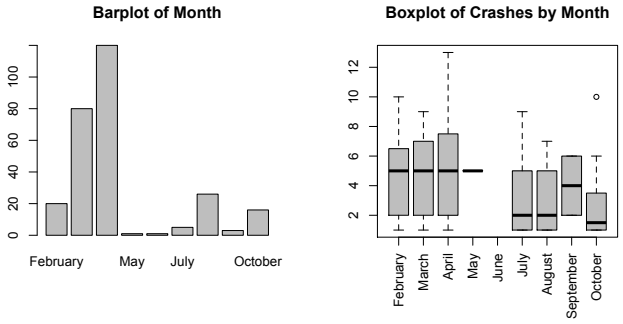
Het eerste spreidingsdiagram van Figuur 20 toont de evolutie van het aantal valpartijen per wedstrijd over de jaren heen. De groene regressierechte en rode niet-parametrische curve vallen samen. Beide trendlijnen kennen dus dezelfde positieve trend. Deze lineaire trend toont aan dat het aantal valpartijen per wedstrijd jaarlijks toegenomen is. Deze positieve trend doet vermoeden dat er in de jaren ’90 minder gevallen werd dan in 2013.

3.9.2 Evolutie van het aantal valpartijen over de maanden heen

De variabele *date* geeft ook aan in welke maand de wedstrijden afgewerkt werden. Deze informatie werd overgenomen door de variabele *month*. Het wielerseizoen wordt afgewerkt tussen februari en oktober, waardoor deze variabele slechts tien niveaus kent.

Figuur 22 verduidelijkt de opbouw van het wielerseizoen. Tussen februari en april wordt het eerste deel van de eendagswedstrijden afgewerkt. Het gaat hier om de openingswedstrijden en de voorjaarsklassiekers. Daarna volgt een onderbreking van ongeveer drie maanden. In mei, juni en juli worden de Giro, Tour en enkele kleinere rittenkoersen afgewerkt, waardoor er minder eendagswedstrijden georganiseerd worden. In de tweede helft van de zomervakantie volgen nog twee kleinere eendagswedstrijden. Vervolgens worden de najaarsklassiekers gereden. Eind oktober wordt het wielerseizoen afgesloten met de Ronde van Lombardije.

De boxplots doen vermoeden aan dat er aan het begin van het seizoen meer gevallen wordt dan in de laatste wedstrijden. Ook zien we dat de spreiding van het aantal valpartijen per wedstrijd groter is aan het begin van het seizoen, dan aan het eind.



Figuur 22 Boxplots voor de variabele *month*

Aan de hand van de ANOVA-analyse van Tabel 9 wordt nagegaan of deze gemiddeldes ook significante verschillen vertonen. Voor deze analyse wordt een significantieniveau van 5% gebruikt.

ANOVA					
	Df.	***SS***	***MS***	***F-waarde***	***p-waarde***
Month	16	163,81	10,2380	1,1652	0,3018
Residuals	156	1370,70	8,7865		
TOTAAL	**172**	**1534,51**			

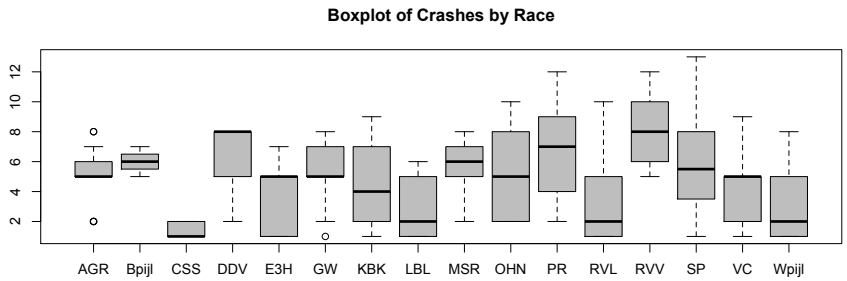
Tabel 9 ANOVA-tabel voor de variabele *month*

De F-waarde wordt geassocieerd met een p-waarde van 0,3018. Deze waarde is ruim groter dan 5%, wat betekent dat de gemiddeldes niet significant verschillen. De maand waarin een wedstrijd georganiseerd wordt, lijkt dus weinig invloed uit te oefenen op de hoeveelheid valpartijen die geteld worden.

3.10 De wedstrijden

3.10.1 Een variabele voor de wedstrijd

Het vorige hoofdstuk beschreef de zestien verschillende Europese eendagswedstrijden die in de dataset van deze masterproef voorkomen. Het is de taak van de variabele *race* om elke wedstrijdeditie te labelen met de naam van de wedstrijd.



Figuur 23 Boxplots van de variabele *race*

De zestien boxplots in Figuur 23 bieden een overzicht van de variabele *crashes* per wedstrijd. Wanneer we deze grafieken vergelijken, dan zien we dat deze verdelingen sterk verschillen. Zo ligt de medianen van Dwars door Vlaanderen (DDV) en de Ronde van Vlaanderen (RVV) een stuk hoger dan die van de Clasica San Sebastian (CSS). De vormen van de boxplots geven ook aan dat de gemiddeldes naar alle waarschijnlijkheid sterk verschillen.

Om de gemiddeldes van de verschillende wedstrijden te vergelijken maken we gebruik van een ANOVA-analyse. De resultaten van deze test zijn weergegeven in Tabel 10.

ANOVA						
	Df.	***SS***	***MS***	***F-value***	***p-value***	
Race	15	512,92	34,195	5,2552	1,933e-08	***
Residuals	157	1021,59	6,507			
Totaal	**162**	**1534,51**				

Tabel 10 ANOVA-tabel voor de variabele *race*

Opdat het gemiddelde van één van de wedstrijden significant afwijkt van dat van de anderen wedstrijden moet de F-waarde van de variabele *race* groter zijn dan 1,73052. Deze waarde stemt immers overeen met het 95e percentiel van een F-verdeling met 15 vrijheidsgraden in de teller en 157 vrijheidsgraden in de noemer.

In de vijfde kolom van Tabel 10 wordt de F-waarde voor de variabele *race* weergegeven. Deze is gelijk aan 5,2552 en is dus ruim groter dan de vooropgestelde kritieke F-waarde van 1,73052. Dit resulteert in een p-waarde die bijna gelijk is aan 0. We kunnen dus besluiten dat het gemiddelde van minstens één van de wedstrijden significant verschilt van de anderen.

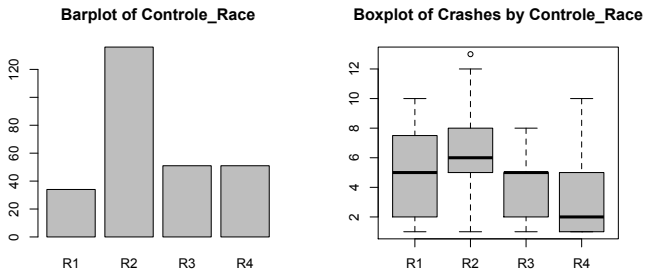
Aan de hand van een Tukey-test kan achterhaald worden welke gemiddeldes afwijken. Deze test voert een paarsgewijze vergelijking van alle gemiddeldes uit. Tabel 11 bevat de resultaten van alle gemiddeldes die verschillen op het 5%-significantieniveau.

	Difference	Lower	Upper	p-value	
CSS en AGR	-3,7667	-7,3933	-0,1400	0,033114	*
GW en CSS	3,9500	0,1463	7,7537	0,033163	*
MSR en CSS	4,3667	0,5630	8,1704	0,009229	**
OHN en CSS	4,3154	0,5788	8,0520	0,008473	**
PR en CSS	5,2714	1,5933	8,9495	0,000176	***
PR en LBL	3,4286	0,0709	6,7862	0,040012	*
RVL en PR	-3,5714	-6,9930	-0,1498	0,031280	*
RVV en CSS	6,9857	3,3076	10,6638	0,000000	***
RVV en E3H	4,7302	0,9347	8,5256	0,002590	***
RVV en KBK	3,8857	0,2076	7,5638	0,027177	*
RVV en LBL	5,1429	1,7852	8,5005	0,000038	***
RVV en RVL	5,2857	1,8641	8,7073	0,000031	***
SP en CSS	4,7000	0,4862	8,9138	0,013728	**
Wpijl en PR	-3,4286	-6,7862	-0,0709	0,040012	*
Wpijl en RVV	-5,1429	-8,5005	-1,7852	0,000038	***

Tabel 11 Tukey-test voor de variabele *race*

3.10.2 Een controlevariabele voor het type wedstrijd

Bij het beschrijven van de verschillende wedstrijden werden deze gegroepeerd naar hun volgorde op de huidige kalender. Deze groepen stemmen overeen met het type renners dat in elke wedstrijd aan de start komt.



Figuur 24 Staafdiagram en boxplots voor de variabele *controle_race*

Deze indeling wordt overgenomen door de variabele *controle_race*. In de eerste groep krijgen de koersen het label R1. Hier vinden we de wedstrijden uit het openingsweekend terug. Milaan-Sanremo, de Vlaamse klassiekers en de Brabantse Pijl worden aangeduid als groep R2. In deze wedstrijden komen renners als Tom Boonen en Fabian Cancellara aan de start. De tweede helft van de voorjaarsklassiekers wordt ondergebracht in een volgende groep R3. Omdat er in deze koersen meer geklommen moet worden, hebben de topfavorieten een kleinere en lichtere lichaamsbouw dan hun collega's uit de eerste helft van het voorjaar. De wedstrijden uit het najaar worden ondergebracht in groep R4.

Rechts in Figuur 24 wordt de boxplot met kengetallen van de variabele *crashes* getoond. De vormen van de boxplots vertonen verschillen en de gemiddeldes blijken eveneens te verschillen. Voor de wedstrijden uit groep R1 vinden we een gemiddelde van 5,087 valpartijen per wedstrijd. In groep R2 wordt het meest gevallen; deze wedstrijden tellen gemiddeld 6,067 valpartijen. De koersen uit groep R3 hebben een gemiddelde van 3,814 valpartijen per wedstrijd. In de najaarsklassiekers wordt het minst gevallen. Een koers telt dan gemiddeld 2,097 valpartijen.

Om na te gaan of de vier wedstrijdgroepen significant verschillende gemiddeldes hebben, wordt een ANOVA-test uitgevoerd. De resultaten van deze analyse worden weergegeven in Tabel 12.

ANOVA						
	Df.	***SS***	***MS***	***F-value***	***p-value***	
Controle_race	3	278,79	92,929	12,507	1,998e-07	***
Residuals	169	1255,72	7,430			
Totaal	**172**	**1534,51**				

Tabel 12 ANOVA-tabel voor de variabele *controle_race*

De F-waarde is groter dan het 95e percentiel van een F-verdeling met respectievelijk 3 en 169 vrijheidsgraden. De gemiddeldes vertonen dus significante verschillen.

Om te achterhalen welke gemiddeldes precies verschillen, voeren we een Tukey-test uit. In deze test worden de gemiddeldes van de vier groepen paarsgewijs vergeleken.

	Diff.	**Upper**	**Lower**	**p-value**	
R2 en R1	0,9797101	-0,7061518	2,6655721	0,4350312	
R3 en R1	-1,2730030	-3,1001663	0,5541602	0,2732483	
R4 en R1	-2,1807065	-4,1142145	-0,2471986	0,0201801	*
R3 en R2	-2,2527132	-3,6056567	-0,8997697	0,0001546	***
R4 en R2	-3,1604167	-4,6538636	-1,6669697	0,0000009	***
R4 en R3	-0,9077035	-2,5589995	0,7435926	0,4847309	

Tabel 13 Tukey-test voor de variabele *controle_race*

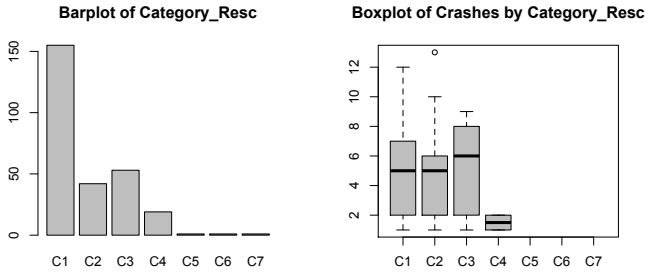
De resultaten van deze analyse worden weergegeven in Tabel 13. Hieruit blijkt dat er bij een significantieniveau van 5% drie duo's zijn die significante verschillen vertonen.

Het gemiddeld aantal valpartijen van de najaarsklassiekers (R4) is significant verschillend van de hoeveelheid valpartijen in de openingswedstrijden (R1) en de eerste helft van het voorjaar (R2). Er is echter ook een significant verschil tussen de gemiddeldes van de openingswedstrijden (R1) en de kasseiklassiekers die in het voorjaar afgewerkt worden (R2).

3.11 De wedstrijdcategorieën

Naast een plaatsje op de wielerkalender kregen de verschillende wedstrijden ook een categorie toegewezen. De UCI maakt hiervoor gebruik van een competitiemodel. Het vorige hoofdstuk behandelde de verschillende systemen die de laatste jaren gebruikt werden. Er werd immers verschillende keren een nieuw model geïntroduceerd.

De categorie van elke wedstrijd wordt bijgehouden met behulp van de variabele *category*. De variabele *category_resc* vertaalt de categorieën van de vier competitiemodellen naar één gemeenschappelijke ordinale schaal. De wedstrijden van het hoogste niveau worden aangeduid met het nummer C1. Het volgende wedstrijdniveau wordt aangeduid met het nummer C2. De overige categorieën worden achtereenvolgens genummerd met de waardes C3 tot C7, zoals aangegeven in Tabel 1.
De samenstelling van de variabele *category_resc* en de boxplots met het aantal valpartijen per wedstrijd in een bepaalde categorie worden weergegeven in Figuur 25.



Figuur 25 Staafdiagram en boxplots voor de variabele *category_resc*

Voor de niveaus C5 tot C7 kan geen boxplot getekend worden, want deze categorieën bevatten elk slechts één observatie. Tussen 1997 en 1999 werd de Brabantse Pijl (BPijl) immers ieder jaar gepromoveerd tot een hogere categorie en de andere wedstrijden uit deze categorie maken geen deel uit van de dataset. Omwille van het beperkte aantal observaties worden deze drie categorieën niet in rekening genomen in de verdere analyse.

Om te testen of het gemiddeld aantal valpartijen per wedstrijdcategorie significant afwijkt van de gemiddelde hoeveelheid valpartijen in de andere categorieën, wordt opnieuw gebruik gemaakt van een ANOVA-tabel. De resultaten van deze analyse worden samengevat in Tabel 14.

ANOVA					
	Df.	***SS***	***MS***	***F-value***	***p-value***
Category_resc	3	22,01	7,3362	0,8197	0,4846
Residuals	169	1512,50	8,9497		
Totaal	**172**	**1534,51**			

Tabel 14 ANOVA-tabel voor de variabele *category_resc*

De teststatistiek is kleiner dan de vooropgestelde kritieke F-waarde van 2,6581. Bijgevolg wordt deze waarde geassocieerd met een p-waarde van 0,4846. Deze is ruim groter dan 5%, zodat de nulhypothese niet kan verworpen worden. De categorie van een wedstrijd blijkt dus weinig invloed te hebben op het aantal valpartijen per wedstrijd.

4 Modellering van het aantal valpartijen per wedstrijd

4.1 Methodologie

In het vierde hoofdstuk van deze masterproef worden vijf Poisson-regressiemodellen ontwikkeld. Deze modellen trachten het aantal valpartijen per wedstrijd te voorspellen op basis van (een deel van) de eerder besproken kenmerken van een eendagswedstrijd. Voordat deze Poisson-regressiemodellen ontwikkeld en besproken worden, beschrijft de eerste sectie van dit vierde hoofdstuk wat een Poisson-regressiemodel precies is, hoe de parameters geschat worden en hoe de kwaliteit van de modellen beoordeeld wordt.

4.1.1 Een Poisson-regressiemodel

In het regressiemodel trachten we het aantal valpartijen per wedstrijd te verklaren. Dit betekent dat de variabele *crashes* de rol van responsvariabele aanneemt. We nemen aan dat de hoeveelheid valpartijen per wedstrijd een Poisson-verdeling volgt. Deze discrete kansverdeling telt immers hoe vaak een bepaalde gebeurtenis, in dit geval een valpartij, waargenomen wordt binnen een bepaald tijdsinterval.

Een lineair regressiemodel kan het gemiddelde van de responsvariabele beschrijven als een lineaire functie van de verklarende variabelen. Het model wordt dan beschreven met behulp van uitdrukking (2), waarbij de foutentermen ϵ een normale verdeling volgen met een gemiddelde nul en een zekere standaarddeviatie σ^2

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{2}$$

Omdat aangenomen werd dat de te verklaren variabele een Poisson-verdeling volgt, is het niet langer mogelijk het gemiddelde weer te geven als een lineaire functie van de onafhankelijke variabelen. Bijvoorbeeld omdat het gemiddelde groter is dan nul en omdat de foutentermen niet langer normaal verdeeld zijn. Daarom wordt gebruik gemaakt van een *generalized linear model* (GLM). Deze klasse parametrische regressiemodellen neemt aan dat de responsvariabele een verdeling volgt die deel uitmaakt van de exponentiële familie[11]. Met behulp van een linkfunctie $g(\,)$ wordt de gemiddelde responswaarde van het model zo getransformeerd dat het geschatte model kan geschreven worden als een lineaire functie van de verklarende variabelen. De algemene vorm van een GLM wordt weergegeven in formule (3), waarbij *Y* een verdeling uit de exponentiële familie met een gemiddelde $E(Y|x_1, \ldots, x_p)$ volgt.

$$g\left(E(Y|x_1, \ldots, x_p)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3}$$

In een Poisson-regressiemodel vervult de logfunctie de rol van de linkfunctie $g(\,)$, zodat het model kan beschreven worden zoals in formule (4), waarbij *Y* een Poisson-verdeling met een gemiddelde $E(Y|x_1, \ldots, x_p)$ volgt.

$$\log\left(E(Y|x_1, \ldots, x_p)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{4}$$

[11] Naast de Poisson-verdeling zijn onder andere de normale, de binomiale en de Gamma-verdeling deel van de exponentiële familie.

4.1.2 Werking van de iteratively reweighted least squares methode

Een bijkomende moeilijkheid is het schatten van de parameters in een GLM. Er is immers geen expliciete uitdrukking voor het bepalen van de parameterschatters, omdat het gemiddelde van de afhankelijke variabele niet langer een lineaire functie is van de parameters in de parametervector $\boldsymbol{\beta}$.

Om de waarden van de parameters te bepalen, maakt het softwarepakket *R* gebruik van de *iteratively reweighted least squares* methode. In deze methode wordt de waarde van de verschillende parameters berekend op basis van een Newton-Raphson algoritme of Fisher scoring algoritme.

Onderstaande tekst licht de werking van het Newton-Raphson algoritme kort toe. Deze beschrijving is gebaseerd op de cursustekst *Statistical Modelling* van Prof. G. Claeskens, waarin een meer uitgebreide beschrijving van dit algoritme, alsook van de werking van het Fisher scoring algoritme gegeven wordt.

In een eerste stap wordt de eerste afgeleide van de log likelihood functie ℓ_n gelijkgesteld aan nul. Deze bewerking wordt weergegeven in uitdrukking (5).

$$\frac{\partial \ell_n(\widehat{\boldsymbol{\beta}},\phi)}{\partial \beta_j} = 0 \tag{5}$$

Vervolgens wordt de Taylorontwikkeling van $\frac{\partial \ell_n(\widehat{\boldsymbol{\beta}},\phi)}{\partial \beta_j}$ rond een bepaalde waarde $\boldsymbol{\beta}$ bepaald. Wanneer enkel de eerste en tweede orde termen behouden blijven, bekomen we de uitdrukking die weergeven wordt in (6).

$$\frac{\partial \ell_n(\widehat{\boldsymbol{\beta}},\phi)}{\partial \beta_j} = 0 \doteq \frac{\partial \ell_n(\widehat{\boldsymbol{\beta}},\phi)}{\partial \boldsymbol{\beta}} + \frac{\partial^2 \ell_n(\boldsymbol{\beta},\phi)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta^t}} \cdot \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \tag{6}$$

Wanneer we deze uitdrukking herschrijven in functie van de parameterschatter $\widehat{\boldsymbol{\beta}}$, bekomen we uitdrukking (7).

$$\widehat{\boldsymbol{\beta}} \doteq \boldsymbol{\beta} + J_n(\boldsymbol{\beta},\phi)^{-1} \frac{\partial \theta(\boldsymbol{\beta})^t}{\partial \boldsymbol{\beta}} u(\boldsymbol{\beta}) \tag{7}$$

met $\frac{\partial \theta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \theta_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \dots, \frac{\partial \theta_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^t$ en $u(\boldsymbol{\beta}) = \left(\frac{\partial \ell_n(\boldsymbol{\beta},\phi)}{\partial \theta_1(\boldsymbol{\beta})}, \dots, \frac{\partial \ell_n(\boldsymbol{\beta},\phi)}{\partial \theta_n(\boldsymbol{\beta})}\right)^t$

en waarin $-J_n(\boldsymbol{\beta},\phi) = \frac{\partial \theta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \cdot \frac{\partial \log f(y,\theta(\boldsymbol{\beta}),\phi)}{\partial \theta(\boldsymbol{\beta})^2} \cdot \frac{\partial \theta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^t} + \frac{\partial \theta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \cdot u(\boldsymbol{\beta})$ de Hessiaan-matrix voorstelt, waarvoor aangenomen wordt dat deze inverteerbaar is.

Het resultaat van deze uitdrukking is een startwaarde voor het proces waarbij de waarde van $\widehat{\boldsymbol{\beta}}$ telkens opnieuw ingevuld wordt in formule (7). Dit iteratieve proces wordt beëindigd wanneer $\widehat{\boldsymbol{\beta}}$ convergeert. Op dat moment geeft de vector $\widehat{\boldsymbol{\beta}}$ immers de maximum likelihood schatters van de parameters $\boldsymbol{\beta}$ weer.

4.1.3 Beoordelingscriteria

Om de kwaliteit van de modellen te beoordelen, wordt gebruik gemaakt van twee criteria. Eerst wordt de waarde van het Akaike's Information Criterion (AIC) berekend. Dit criterium wordt gedefinieerd in formule (8).

$$AIC(\boldsymbol{\beta}) = -2\ log\text{-}likelihood_{max}\ (\boldsymbol{\beta}) + 2\ dim(\boldsymbol{\beta}) \quad (8)$$

Deze formule maakt een afweging tussen de fit en complexiteit van het bestudeerde model. Het eerste deel van de formule beoordeelt de manier waarop het model de data weergeeft. Bij een goede fit vertoont een model een grote log-likelihood, waardoor de AIC-waarde daalt. In het tweede deel van de formule wordt deze waarde gepenaliseerd voor het aantal parameters die nodig zijn om deze fit te bereiken. Omdat we een eenvoudig model met goede fit wensen, verkiezen we een model met zo klein mogelijke AIC-waarde.

Een tweede criterium beoordeelt de predictiekwaliteit van het model. De parameters van het regressiemodel worden geschat op basis van de trainingsdata, vervolgens wordt het resulterende model toegepast op de testobservaties. Deze laatste groep bevat informatie over 29 wedstrijden[12] die afgewerkt worden tussen februari 2013 en oktober 2014. Dit zijn ongeziene data, want het model wordt geschat op basis van 143 races die gereden werden tussen februari 1997 en oktober 2012. De voorspellingsfout van het model wordt dan berekend aan de hand van uitdrukking (9), waarin *n* het aantal bestudeerde observaties voorstelt.

$$Voorspellingsfout = \sum_{i=1}^{n}\left(y_{Crashes(i)} - \hat{y}_{Crashes(i)}\right)^2 \quad (9)$$

Deze formule vergelijkt het werkelijke aantal valpartijen per wedstrijd ($y_{Crashes(i)}$) met de voorspelde hoeveelheid valpartijen per wedstrijd ($\hat{y}_{Crashes(i)}$). De verschillen tussen beide waardes worden gekwadrateerd en daarna opgeteld.
Het vervolg van deze tekst zal naar dit resultaat verwijzen als de voorspellingsfout van het bestudeerde model. Een goed model kan een kleine voorspellingsfout voorleggen; dit impliceert immers dat er weinig fouten gemaakt worden bij het voorspellen van het aantal valpartijen per wedstrijd.

Voor elk model worden twee voorspellingsfouten berekend. Nadat het model geschat werd op basis van de 143 trainingsobservaties wordt het een eerste keer toegepast op diezelfde verzameling gegevens. Dit levert de voorspellingsfout van de trainingsdata. Vervolgens wordt de voorspellingsfout van de 29 test-wedstrijden berekend. Deze fout wordt aangeduid als de voorspellingsfout van de test-data.

4.1.4 Vijf modellen

Om de prestaties van de regressiemodellen te vergelijken, wordt gebruik gemaakt van een benchmarkmodel. Dit eerste model is een additief model dat alle variabelen bevat die in de univariate analyses reeds aan bod kwamen. Enkel de variabele *rat_nr* werd niet opgenomen in het model. De correlatiematrix in bijlage 2 toont immers aan dat deze variabele sterk positief gecorreleerd is met de variabele *rat_perc*. Dit is geen verassing, want *rat_perc* en *rat_nr* bevatten dezelfde informatie, maar stellen deze op een andere manier voor.

[12] Er ontbreken drie wedstrijden uit 2013; Kuurne-Brussel-Kuurne werd afgelast wegens sneeuw en voor Milaan-Sanremo en de Vattenfall Cyclassics werden geen volledige gegevens teruggevonden.

Uit de correlatiematrix in bijlage 2 blijkt dat enkele andere variabelen eveneens sterk (positief) gecorreleerd zijn. Verschillende achtergronden zijn immers grijs gekleurd, wat betekent dat de variabelen over een correlatiecoëfficiënt van meer dan 60% beschikken. Bovendien bevatten enkele niet-numerieke variabelen gelijkaardige informatie die op een andere manier gecodeerd wordt. Een model dat dergelijk duo bevat, zal moeite hebben met het onderscheiden van de verklarende waarde van elk van beide variabelen. In een tweede en derde model zal daarom geprobeerd worden om de AIC-waarde van het model te minimaliseren. Het tweede model is zuiver additief, in het derde model worden ook interactietermen in rekening genomen.

In het vierde en vijfde model wordt de voorspellingsfout van de trainingsdata naar beneden gebracht. Zo wordt een model gecreëerd dat geschikt is voor het voorspellen van het aantal valpartijen per wedstrijd op basis van de kenmerken die doorheen deze masterproef aan bod kwamen.

4.2 Een additief model

De prestaties van de drie andere regressiemodellen die zullen voorgesteld worden, worden afgetoetst aan een benchmarkmodel. Dit additieve model bevat alle variabelen die in deze masterproef aan bod kwamen, op de variabele *rat_nr* na. Zoals eerder uitgelegd is deze variabele immers sterk gecorreleerd met de variabele *rat_perc*, waardoor geen onderscheid kan gemaakt worden tussen de verklarende waarde van elk van beide variabelen.

Dit benchmarkmodel wordt weergegeven als Model 1, waarbij we aannemen dat $y_{Crashes}$ een Poisson-verdeling volgt.

$$
\begin{aligned}
log\big(E(y_{Crashes})\big) = \quad & \beta_0 && +\beta_1 \times Year && +\beta_2 \times Month \\
& +\beta_3 \times Race && +\beta_4 \times ControleRace && +\beta_5 \times Category \\
& +\beta_6 \times CategoryResc && +\beta_7 \times Distance && +\beta_8 \times AVGspeed \\
& +\beta_9 \times Starters && +\beta_{10} \times ControleStarters && +\beta_{11} \times Finishers \\
& +\beta_{12} \times STF && +\beta_{13} \times Radio && +\beta_{14} \times Rain \\
& +\beta_{15} \times Climbs && +\beta_{16} \times ControleClimbs && +\beta_{17} \times Cobbles \\
& +\beta_{18} \times ControleCobbles && +\beta_{19} \times CobblesKm && +\beta_{20} \times ShrPerc \\
& +\beta_{21} \times RatPerc
\end{aligned}
$$

Model 1 Het additieve model

Tabel 15 toont de resultaten van dit model voor de twee eerder besproken criteria. Deze scores gelden als referentiewaarde voor de modellen die in het verdere verloop van deze tekst aan bod komen.

Resultaten voor Model 1	
AIC-waarde	554,88
Voorspellingsfout (trainingsdata)	2246,113
Voorspellingsfout (test-data)	440,0689

Tabel 15 Resultaten voor Model 1

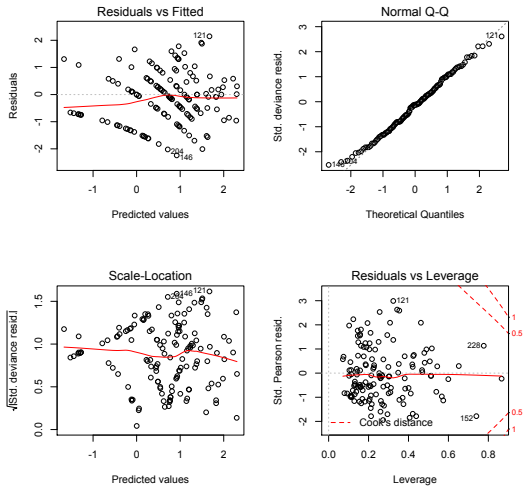
Om de kwaliteit van dit eerste model te beoordelen, wordt gebruik gemaakt van de vier plots die weergegeven worden in Figuur 26.

Het eerste diagram toont de voorspelde waardes in functie van de residuen van het model. De residuen meten de afstand tussen de werkelijke en geschatte waarde van de observaties. Uit deze grafiek blijkt dat de observaties 121, 146 en 204 uitschieters zijn. De rode trendlijn ligt net beneden de horizontale nullijn. Dit betekent dat het gemiddelde van de errortermen net iets kleiner is dan nul. Het patroon van de observaties is een gevolg van het discreet zijn van de variabele *crashes*. Bij elke waarde *k* die de variabele aanneemt, hoort precies één dalende curve.

De volgende grafiek toont de QQ-plot van het model, waarin de kwantielen van de gestandaardiseerde deviance residuen vergeleken worden met de theoretische kwantielen van een normaalverdeling. De uiteindes van de plot tonen enkele afwijkingen, maar de meeste observaties vallen samen met de gestippelde rechte. Over het algemeen kunnen we stellen dat de gestandaardiseerde deviance residuen bij benadering normaal verdeeld zijn.

De scale-location plot brengt de voorspelde waarden van de observaties en de wortel uit de gestandaardiseerde deviance residuen samen. De observaties liggen willekeurig rond de rode trendlijn. Bij zeer kleine voorspellingswaarden wordt de spreiding kleiner. Deze verschillen zijn echter niet problematisch, zodat we kunnen aannemen dat voldaan is aan de assumptie van homoskedasticiteit van de gestandaardiseerde deviance residuen.

In de laatste grafiek wordt de *Cook's distance* van de observaties bestudeerd. Omdat alle datapunten binnen de rode stippellijnen liggen, mogen we aannemen dat geen van de observaties tegelijk afwijkend en invloedrijk is.



Figuur 26 Plots voor Model 1

4.3 Twee modellen op basis van AIC-waarde

Het derde deel van dit hoofdstuk gaat op zoek naar het model met de kleinste AIC-score. Dit model is eenvoudiger dan Model 1 en/of geeft de trainingsdata beter weer, want enkel zo kan de AIC-waarde naar beneden gebracht worden.

Om deze combinatie van variabelen te selecteren die de AIC-waarde zo klein mogelijk maakt, wordt gebruik gemaakt van de *StepAIC*-functie uit de *R*-library *MASS*. Deze functie doorloopt alle deelverzamelingen van de variabelen in de dataset. Op basis van hun bijdrage aan de AIC-waarde van het model worden de variabelen één na één toegevoegd of verwijderd uit het tijdelijke model. Aan het eind van dit iteratieve proces wordt het model met zo klein mogelijke AIC-score voorgesteld.

4.3.1 Een zuiver additief model op basis van AIC-waarde

We passen deze techniek toe op het additieve Model 1, zonder rekening te houden met interactietermen. We vinden dan het model dat weergegeven wordt als Model 2.

$$log\left(E(Y_{Crashes})\right) = \beta_0 + \beta_1 \times Year + \beta_3 \times Race + \beta_7 \times Distance + \beta_{11} \times Finishers + \beta_{18} \times ControleCobbles + \beta_{20} \times ShrPerc$$

Model 2 Het additieve model met de kleinste AIC-waarde

Dit model bevat een intercept en zes variabelen. Alle variabelen die de data niet goed weergegeven en/of het model nodeloos complex maken, worden immers verwijderd. De scores van dit model worden weergegeven in Tabel 16.

Resultaten voor Model 2	
AIC-waarde	526,67
Voorspellingsfout (trainingsdata)	2163,426
Voorspellingsfout (test-data)	301,3815

Tabel 16 Resultaten voor Model 2

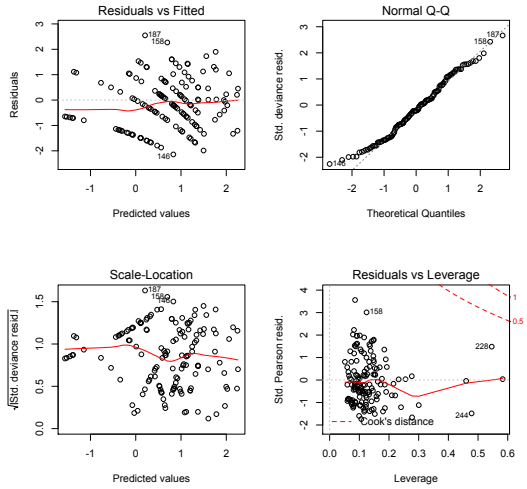
De AIC-score van het model daalt tot een waarde 526,67. Dit betekent dat Model 2 een beter evenwicht vindt tussen de fit van de trainingsdata en de complexiteit van het model. Bovendien heeft dit model meer voorspellingskwaliteit. Het model geeft ook betere voorspellingen voor de trainingsdata, want de voorspellingsfout daalt naar 2163,426. Wanneer we het model toepassen op de 29 ongeziene testobservaties, dan vinden we een voorspellingsfout van 301,3815. Ook deze waarde is lager dan de voorspellingsfout van Model 1.

We beoordelen de kwaliteit van het model een tweede keer met de plots in Figuur 27. Het eerste diagram van deze figuur gelijkt sterk op de eerste plot van Figuur 26. Observaties 146, 158 en 187 worden gemarkeerd als afwijkend, maar het vierde plot van deze figuur toont aan dat deze afwijkingen niet problematisch zijn. De rode trendlijn is horizontaal en ligt net beneden de rechte met vergelijking *y = 0*. De gestandaardiseerde deviance residuen volgen dus een verdeling met een gemiddelde net kleiner dan nul.

De uiteindes van de QQ-plot wijken af van de gestippelde 45°-lijn. Dit betekent dat de gestandaardiseerde deviance residuen mogelijk niet normaal verdeeld zijn.

In het scale-location plot kan geen patroon gevonden worden. Net als in Model 1 neemt de spreiding af bij kleine voorspellingswaarden. De meerderheid van de observaties ligt echter willekeurig in een horizontale band rond de trendlijn, zodat we kunnen stellen dat de variabiliteit in het model onafhankelijk is van de waarde van de te verklaren variabele.

De laatste grafiek toont aan dat er geen observaties zijn die tegelijk afwijkend en invloedrijk zijn. Alle datapunten liggen immers binnen het gebied dat omlijnd wordt door de rode stippellijnen van de *Cook's distance.*



Figuur 27 Plots voor Model 2

Tabel 17 toont de parameterschattingen voor Model 2. Bij een significantieniveau van 5% zijn meerdere parameters significant verschillend van nul. Deze schatters worden aangeduid met één of meerdere sterren, afhankelijk van de grootte van de p-waarde. Hoe kleiner de p-waarde, hoe meer sterren de variabele achter zijn naam krijgt.

De parameter bij de variabele *year* is significant verschillend van nul. De waarde van deze schatter is positief en geeft aan dat het aantal valpartijen jaarlijks licht toeneemt. Wanneer alle kenmerken van een wedstrijd ongewijzigd blijven, schat dit model dat dezelfde wedstrijd één jaar later $e^{0,0564}$ = 1,0580 meer valpartijen telt.

De Amstel Gold Race (AGR) fungeert als *baseline*-koers voor het schatten van de parameters bij de verschillende niveaus van *race*. Milaan-Sanremo (MSR) is de enige wedstrijd met een positieve parameterschatter. In deze koers worden dus meer valpartijen geteld dan in de Amstel Gold Race. Het verschil tussen het gemiddeld aantal valpartijen in beide wedstrijden blijkt echter niet significant, zoals ook al bleek uit het resultaat van de univariate analyses.
De schatters bij de Omloop het Nieuwsblad (OHN), Dwars door Vlaanderen (DDV), de E3 Harelbeke (E3H), Gent-Wevelgem (GW), de Scheldeprijs (SP), de Waalse Pijl (Wpijl), Luik-Bastenaken-Luik (LBL) en de Clasica San Sebastian (CSS) zijn negatief en bovendien significant verschillend van nul. Deze koersen vertonen dus significant minder valpartijen dan de Amstel Gold Race. Deze resultaten zijn een bevestiging van de univariate analyses, want ook daar werd een significant verschil tussen de gemiddeldes van de verschillende koersen vastgesteld.

De parameter bij de variabele *distance* is negatief. Een toename van de lengte van het wedstrijdparcours wordt geassocieerd met een daling van het aantal valpartijen. Deze parameter blijkt echter niet significant verschillend van nul.

De schatter bij de variabele *finishers* is positief. Een toename van het aantal finishers is dus gerelateerd aan een stijging van het aantal valpartijen. Deze parameterschatter is echter ook niet significant verschillend van nul.

De univariate analyses uit het vorige hoofdstuk toonden al aan dat in een kasseikoers meer gevallen wordt dan in een 'gewone' wedstrijd. In dit model is de parameterschatter bij de *yes*-waarde van de variabele *controle_cobbles* positief en significant verschillend van nul. Een kasseikoers telt $e^{1,0930}$ = 2,9832 meer valpartijen dan een gewone wedstrijd.

Een laatste parameter schat het verband tussen de hoeveelheid valpartijen en het marktaandeel van het tv-verslag van deze wedstrijd. De schatter bij deze variabele is positief, wat betekent dat een toename van het marktaandeel gekoppeld wordt aan een toename van het aantal valpartijen. Deze parameter is echter niet significant verschillend van nul.

	Estimate	Std. Error	z-value	Pr(>|z|)	
Intercept	-109,6000	33,6400	-3,2590	0,001119	**
Year	0,0564	0,0168	3,3630	0,000771	***
Race Bpijl	-0,5100	0,6464	-0,7890	0,430154	
Race CSS	-2,5000	0,7625	-3,2780	0,001044	**
Race DDV	-2,0220	0,8043	-2,5140	0,011933	*
Race E3H	-1,9310	0,8971	-2,1520	0,031385	*
Race GW	-1,6200	0,6521	-2,4840	0,012989	*
Race KBK	-0,9155	0,5902	-1,5510	0,120879	
Race LBL	-0,6391	0,3114	-2,0520	0,040136	*
Race MSR	0,5124	0,4151	1,2340	0,217050	
Race OHN	-1,8850	0,7253	-2,5990	0,009337	**
Race PR	-0,0717	0,5722	-0,1250	0,900328	
Race RVL	-0,6951	0,4072	-1,7070	0,087829	.
Race RVV	-0,4353	0,5636	-0,7720	0,439857	
Race SP	-1,2670	0,5639	-2,2470	0,024639	*
Race VC	-0,0284	0,3842	-0,0740	0,941096	
Race Wpijl	-1,4950	0,5498	-2,7190	0,006547	**
Distance	-0,0153	0,0082	-1,8700	0,061477	.
Finishers	0,0036	0,0021	1,6590	0,097055	.
Controle_Cobbles YES	1,0930	0,4899	2,2310	0,025688	*
Shr_Perc	0,0148	0,0103	1,4390	0,150223	

Tabel 17 Parameterschatters voor Model 2

4.3.2 Een model met interactietermen op basis van AIC-waarde

De *StepAIC*-functie wordt een tweede keer toegepast op het oorspronkelijke Model 1. Dit keer laten we toe dat het voorgestelde model interactietermen bevat. Aan het eind van de procedure stelt de functie Model 3 voor.

$$
\begin{aligned}
log\left(E(y_{Crashes})\right) = \quad & \beta_0 && +\beta_1 \times Year \\
& +\beta_2 \times Month && +\beta_3 \times Race \\
& +\beta_7 \times Distance && +\beta_9 \times Starters \\
& +\beta_{10} \times StartersGrouped && +\beta_{11} \times Finishers \\
& +\beta_{12} \times STF && +\beta_{14} \times Rain \\
& +\beta_{18} \times ControleCobbles && +\beta_{20} \times ShrPerc \\
& +\beta_{21} \times RatPerc && +\beta_{22} \times StartersGrouped \times ShrPerc \\
& +\beta_{23} \times STF \times RatPerc && +\beta_{24} \times Month \times RatPerc \\
& +\beta_{25} \times Month \times Starters && +\beta_{26} \times MonthRain \\
& +\beta_{27} \times Rain \times ControleCobbles && +\beta_{28} \times Distance \times RatPerc \\
& +\beta_{29} \times Starters \times Rain && +\beta_{30} \times StartersGrouped \times Rain
\end{aligned}
$$

Model 3 Het interactie-model met de kleinste AIC-waarde

Tabel 18 geeft de scores van dit derde model voor de beoordelingscriteria weer.

Resultaten voor Model 3	
AIC-waarde	512,85
Voorspellingsfout (trainingsdata)	2347,656
Voorspellingsfout (test-data)	550,7379

Tabel 18 Resultaten voor Model 3

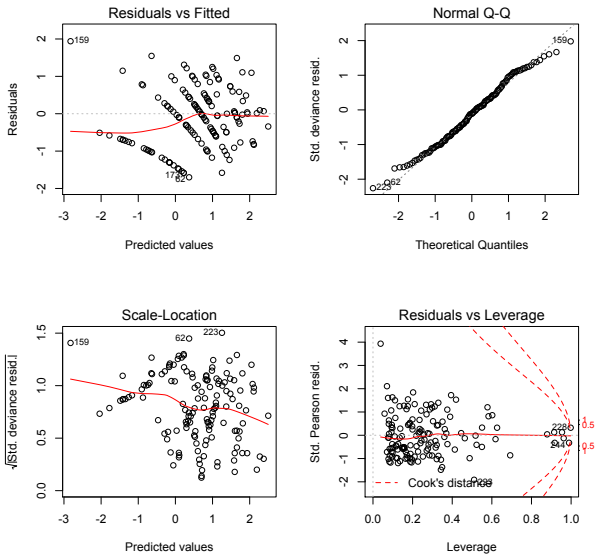
Door het toelaten van interactietermen daalt de AIC-waarde van het model naar 512,85. Het model moet echter inboeten aan voorspellingskwaliteit. Bij het voorspellen van de hoeveelheid valpartijen voor de wedstrijden in de training-set stijgt de predictiefout tot 2347,656. Deze waarde is veel hoger dan de voorspellingsfout van Model 1 en Model 2. Dit ondanks de kleinere AIC-waarde, die het model penaliseert voor het aantal parameters dat het bevat. Dit wijst in de richting van *overfitting*; dit is het fenomeen waarbij een model steeds verbeterd wordt, waardoor het nodeloos complex wordt en zelfs toevalligheden in de trainingsdata kan weergeven (Blockeel, 2010).

Net als bij de vorige modellen wordt de kwaliteit van het model bestudeerd aan de hand van vier plots. Deze worden weergegeven in Figuur 28.
De eerste grafiek toont de voorspelde waarde van de observaties in functie van de residuen van het model. Deze grafiek telt opnieuw één dalende curve voor elke waarde *k* die de responsvariabele aanneemt. De trendlijn wordt gekenmerkt door een knikpunt. De eerste helft van de curve is kleiner dan nul, maar in de tweede helft valt de curve samen met de horizontale rechte met vergelijking *y = 0*. Hierdoor wordt het gemiddelde van de gestandaardiseerde deviance residuen mogelijk kleiner dan nul.

In de QQ-plot worden de kwantielen van de wortel uit de gestandaardiseerde deviance residuen vergeleken met de theoretische kwantielen van een normaalverdeling. Omdat de observaties goed aansluiten bij de gestippelde 45°-lijn kunnen we stellen dat de gestandaardiseerde deviance residuen bij benadering normaal verdeeld zijn.

De rode trendlijn in de derde grafiek heeft een dalend verloop, omdat de wortel uit de gestandaardiseerde deviance residuen kleiner wordt naarmate het model een grotere waarde voorspelt.

De laatste grafiek leert ons dat geen van de gemarkeerde observaties tegelijkertijd afwijkend en invloedrijk is. Alle datapunten liggen immers binnen het gebied dat omlijnd wordt door de rode stippellijn van de *Cook's distance*.



Figuur 28 Plots voor Model 3

De parameterschatters van Model 3 worden weergeven in Tabel 19.

De parameterschatter bij de variabele *year* blijft positief en significant verschillend van nul. Wanneer alle andere wedstrijdkenmerken behouden blijven, schat dit model dat een wedstrijd één jaar later $e^{0,0719}$ = 1,0746 extra valpartijen telt. Het positieve verband tussen het jaar waarin een wedstrijd georganiseerd wordt en de hoeveelheid valpartijen dat geteld wordt, wordt zo opnieuw bevestigd.

De variabele *month* werd niet opgenomen in Model 2. In dit derde model wordt geschat dat het aantal valpartijen toeneemt naarmate het seizoen vordert. De parameterschatter blijkt echter niet significant verschillend van nul. De interactietermen van *month* en *starters* en *month* en *rain* blijken dat wel te zijn.

De parameterschatters bij de verschillende niveaus van de variabele *race* leren ons dat er in de Amstel Gold Race (AGR) minder gevallen wordt dan in Milaan-Sanremo (MSR), de Ronde van Vlaanderen (RVV), Parijs-Roubaix (PR) of de Vattenfall Cyclassics (VC). De schatters van deze laatste wedstrijden zijn immers groter dan nul. Enkel voor Milaan-Sanremo is er echter sprake van een significant verschil. De andere schatters zijn immers niet significant verschillend van nul.
De andere wedstrijden hebben een parameterschatter die kleiner is dan nul, wat betekent dat er minder valpartijen geteld worden dan in de Amstel Gold Race. De koersen waarin een significant verschil waargenomen wordt, krijgen één of meerdere sterren achter hun naam. Hoe kleiner de overeenkomstige p-waarde, hoe meer sterren de overeenkomstige koers krijgt.

Dit model schat een negatieve relatie tussen de variabelen *crashes* en *distance*, maar deze parameter blijkt niet significant verschillend van nul.

De variabelen *starters* en *starters_grouped* bevatten dezelfde informatie en maken beiden deel uit van het regressiemodel. Bovendien komen beide variabelen ook voor in één of meerdere (significante) interactietermen. De schatters bij *starters_grouped*, *starters* en de bestudeerde interactietermen hebben een verschillend teken, waardoor er geen eenduidig verband is met het aantal valpartijen in een wedstrijd.

Hetzelfde geldt voor de variabelen *finishers* en *STF*. Beide variabelen zijn gerelateerd aan het aantal renners dat de finish van een wedstrijd bereikt en bovendien komt *STF* nog een tweede keer voor in één van de interactietermen van het model. De parameters bij *STF, finishers* en de interactieterm met *STF* zijn allemaal significant verschillend van nul, maar hebben een ander teken waardoor het verband met het aantal valpartijen per wedstrijd afhankelijk is van de grootte van beide variabelen.

De parameterschatter bij *rain* is negatief en significant verschillend van nul. Deze variabele maakt ook deel uit van vier significante interactietermen, waardoor het verband tussen het aantal valpartijen en de neerslag niet onafhankelijk kan bestudeerd worden.
Wanneer het regent in één van de latere koersen van het seizoen, dan wordt deze wedstrijd met minder valpartijen geassocieerd dan wanneer het regent aan het begin van het seizoen. De schatter van de interactieterm van *controle_cobbles* en *rain* is eveneens positief en significant verschillend van nul. In een kasseikoers wordt het negatieve verband tussen de regen en het aantal valpartijen dus afgezwakt.
De twee laatste interactietermen brengen de variabele *rain* in verband met respectievelijk *starters* en *starters_grouped*. Het verband tussen het aantal valpartijen en de aan- of afwezigheid van regen wordt dus ook beïnvloed door het aantal starters.

De schatter bij *controle_cobbles* is positief, maar niet significant verschillend van nul. De aanwezigheid van kasseien in het parcours lijkt dus weinig invloed te hebben op het aantal valpartijen per wedstrijd. De waarde van deze variabele blijkt echter wel van belang in enkele (significante) interactietermen.

De schatters van het marktaandeel en de kijkdichtheid van de uitzending van een wedstrijd geven aan dat er een positief verband bestaat tussen het aantal kijkers en het aantal valpartijen. Enkel het marktaandeel blijkt een significante bijdrage te leveren bij het voorspellen van het aantal valpartijen per wedstrijd.

	Estimate	Std. Error	z-value	Pr(>|z|)	
Intercept	-141,0000	42,2200	-3,3390	0,000842	***
Year	0,0719	0,0208	3,4610	0,000537	***
Month	1,8780	0,9677	1,9410	0,052300	.
Race Bpijl	-1,3440	0,8163	-1,6460	0,099680	.
Race CSS	-1,9990	1,0870	-1,8400	0,065783	.
Race DDV	-3,4830	1,4230	-2,4480	0,014373	*
Race E3H	-2,5980	1,1020	-2,3570	0,018447	*
Race GW	-2,0570	0,8083	-2,5440	0,010946	*
Race KBK	-1,7030	0,8195	-2,0790	0,037637	*
Race LBL	-0,3454	0,3373	-1,0240	0,305906	
Race MSR	1,4730	0,5544	2,6570	0,007874	**
Race OHN	-2,2090	0,8979	-2,4600	0,013900	*
Race PR	0,6440	0,6551	0,9830	0,325637	

Race RVL	-0,7515	1,4180	-0,5300	0,596156	
Race RVV	0,0787	0,6426	0,1220	0,902591	
Race SP	-1,8830	0,7932	-2,3740	0,017597	*
Race VC	0,2637	0,8508	0,3100	0,756600	
Race Wpijl	-2,3560	0,8133	-2,8970	0,003765	**
Distance	-0,0148	0,0165	-0,9010	0,367719	
Starters	-0,0114	0,0260	-0,4380	0,661450	
Starters_Grouped 176-200 starters	4,6830	1,5240	3,0740	0,002115	**
Starters_Grouped <= 150 starters	-13,5600	9,9940	-1,3570	0,174905	
Starters_Grouped > 200 starters	10,7900	3,0670	3,5180	0,000435	***
Finishers	0,0645	0,0273	2,3630	0,018111	*
STF	-13,6200	4,9250	-2,7650	0,005687	**
Rain YES	-9,8560	3,9140	-2,5180	0,011794	*
Controle_Cobbles YES	0,7459	0,5541	1,3460	0,178311	
Shr_Perc	0,0724	0,0242	2,9910	0,002778	**
Rat_perc	0,2306	0,3253	0,7090	0,478373	
Starters_Grouped 176-200 starters * Shr_Perc	-0,0564	0,0226	-2,4900	0,012763	*
Starters_Grouped <= 150 starters * Shr_Perc	0,2396	0,1607	1,4910	0,135934	
Starters_Grouped > 200 starters * Shr_Perc	-0,1353	0,0420	-3,2200	0,001282	**
STF * Rat_perc	0,2201	0,1093	2,0150	0,043954	*
Month * Rat_perc	0,0409	0,0288	1,4200	0,155597	
Month * Starters	-0,0113	0,0053	-2,1380	0,032499	*
Month * Rain YES	0,2325	0,0855	2,7180	0,006566	**
Rain YES * Controle_Cobbles YES	0,5699	0,2792	2,0410	0,041280	*
Distance * Rat_perc	-0,0022	0,0012	-1,7890	0,073624	.
Starters * Rain YES	0,0552	0,0222	2,4900	0,012761	*
Starters_Grouped 176-200 starters * Rain YES	-1,9750	0,6786	-2,9100	0,003620	**
Starters_Grouped <= 150 starters * Rain YES	NA	NA	NA	NA	
Starters_Grouped > 200 starters * Rain YES	-2,3970	1,2940	-1,8520	0,064005	.

Tabel 19 Parameterschatters voor Model 3

4.4 Een model op basis van voorspellingsfout

In het vierde deel van dit hoofdstuk wordt een laatste model samengesteld. Hier worden de variabelen geselecteerd op basis van hun bijdrage aan de voorspellingsfout bij het voorspellen van het aantal valpartijen in de wedstrijden uit de verzameling test-data.

Om deze variabelen te selecteren die de predictiefout van het model minimaliseren, wordt de invloed van elke variabele op de predictiefout bepaald. We starten vanaf het oorspronkelijke Model 1, waaruit de variabelen één na één weggelaten worden.
Het schatten van de parameters gebeurt op basis van de trainingsdata. Vervolgens wordt voor elk van deze 'nieuwe' modellen de waarde van de voorspellingfout van de test-data bepaald. De resultaten van deze analyses worden weergegeven in Tabel 20, alsook de procentuele verandering van de voorspellingsfout ten opzichte van de referentiewaarde van Model 1.

	Voorspellingsfout (test-data)	Verandering van voorspellingsfout (in procent)
Referentiewaarde	*440,0689*	
Year	401,2997	-8,8098
Month	341,6648	-22,3611
Race	290,7440	-33,9322
Controle_Race	440,0689	0,0000
Category	386,4377	-12,1870
Category_Resc	440,0689	0,0000
Distance	432,2751	-1,7710
AVGspeed	439,5645	-0,1146
Starters	363,5860	-17,3798
Starters_Grouped	314,6353	-28,5032
Finishers	357,7876	-18,6974
STF	367,8626	-16,4080
Radio	446,2787	1,4111
Rain	451,1390	2,5155
Climbs	458,3018	4,1432
Climbs_Grouped	414,4566	-5,8201
Cobbles	452,8901	2,9134
Controle_Cobbles	408,1386	-7,2558
Cobbles_Km	399,9821	-9,1092
Shr_Perc	545,1768	23,8844
Rat_perc	443,1172	0,6927

Tabel 20 Overzicht van de invloed van de variabelen op de voorspellingsfout van Model 1

De variabelen *year, month, distance, AVGspeed, radio* en *rain* bevatten specifieke informatie die niet overlapt met de informatie die de andere variabelen aanbrengen. In deze eerste stap besluiten we deze variabelen te behouden, zodat deze informatie niet verloren gaat.

De impact van de overige variabelen wordt vergeleken met de verbetering die het weglaten van andere, gelijkaardige variabelen met zich meebrengt. Verschillende variabelen behoren immers tot één familie van gecorreleerde variabelen die gelijkaardige informatie bevatten.

Zo bevatten *race* en *controle_race* sterk gelijkaardige informatie, want beide variabelen beschrijven het type wedstrijd. Het weglaten van de variabele *race* levert een grotere verbetering op dan het weglaten van *controle_race*. In het volgende model zal daarom enkel *controle_race* gebruikt worden voor het beschrijven van het type wedstrijd.

Category en *category_resc* zijn twee andere variabelen die dezelfde informatie bevatten. Beide variabelen beschrijven immers de wedstrijdcategorie waartoe een race behoort. Uit de tabel blijkt dat het weglaten van *category* de grootste verbetering realiseert, daarom zal enkel de variabele *category_resc* behouden blijven.

Ook *starters* en *starters_grouped* brengen dezelfde informatie aan. De resultaten van Tabel 20 tonen aan dat de voorspellingsfout sterker daalt wanneer *starters_grouped* verwijderd wordt dan wanneer *starters* uit het model verdwijnt. In het volgende model blijft daarom enkel de variabele *starters* over.

Uit de correlatiematrix in bijlage 2 blijkt dat *finishers* en *STF* gekenmerkt worden door een correlatiecoëfficiënt van 96%. Mogelijk heeft het model moeite met het onderscheiden van de voorspellingskracht van elk van beide variabelen, daarom beslissen we enkel *STF* op te nemen in het volgende model. Het weglaten van de variabele *finishers* levert immers een grotere verbetering van de voorspellingsfout van het model op.

Zowel *climbs* als *climbs_grouped* beschrijft het aantal hellingen in het parcours. Bij het weglaten van *climbs_grouped* daalt de voorspellingsfout van het model. Bij het weglaten van *climbs* is dat niet het geval; de voorspellingsfout neemt dan toe. In het volgende model blijft de variabele *climbs* dus behouden en wordt *climbs_grouped* verwijderd.

Voor het beschrijven van de kasseistroken in het parcours wordt gebruik gemaakt van de numerieke variabelen *cobbles* en *cobbles_km* en de dummy-variabele *controle_cobbles*. Bij het weglaten van *cobbles_km* of *controle_cobbles* neemt de voorspellingsfout van het model af, maar wanneer *cobbles* weggelaten wordt, neemt de voorspellingsfout toe. Daarom besluiten we de aan- of afwezigheid van kasseien in het parcours uit te drukken met de variabele *cobbles*.

Een laatste familie variabelen bestudeert de impact van de media-aandacht voor een bepaalde wedstrijd. De variabele *rat_nr* werd al weggelaten, omwille van de sterk positieve correlatie met *rat_perc*. *Rat_perc* en *shr_perc* bevatten eveneens sterk positief gecorreleerde informatie. Uit de correlatiematrix van bijlage 2 blijkt immers dat deze variabelen een correlatiecoëfficiënt van 72% hebben. Het weglaten van beide variabelen doet de voorspellingsfout toenemen. Het verwijderen van *shr_perc* zorgt voor de grootste toename, waardoor we deze variabele verkiezen boven *rat_perc.*

Wanneer we de besproken variabelen weglaten, vinden we Model 4.

$$
\begin{aligned}
log\left(E(Y_{Crashes})\right) = \quad & \beta_0 && +\beta_1 \times Year && +\beta_2 \times Month \\
& +\beta_4 \times ControleRace && +\beta_6 \times CategoryResc && +\beta_7 \times Distance \\
& +\beta_8 \times AVGspeed && +\beta_9 \times Starters && +\beta_{12} \times STF \\
& +\beta_{13} \times Radio && +\beta_{14} \times Rain && +\beta_{15} \times Climbs \\
& +\beta_{18} \times Cobbles && +\beta_{21} \times ShrPerc &&
\end{aligned}
$$

Model 4 Voorspellingsmodel na eerste analyse van de variabelen

De waarden van de beoordelingscriteria worden weergegeven in Tabel 21.

Resultaten voor Model 4	
AIC-waarde	555,78
Voorspellingsfout (trainingsdata)	2130,402
Voorspellingsfout (test-data)	240,1126

Tabel 21 Resultaten voor Model 4

Model 4 heeft een hogere AIC-score dan Model 1, maar levert betere voorspellingen af. De voorspellingsfout van de trainingsdata daalt naar een waarde van 2130,402; deze waarde is lager dan de predictiefout van Model 1, Model 2 en Model 3. Zoals verwacht neemt ook de fout bij het voorspellen van het aantal valpartijen in de testwedstrijden af. Deze voorspellingsfout daalt naar een waarde van 240,1126.

In een volgende iteratie onderzoeken we of de voorspellingsfout van de test-data verder kan teruggedrongen worden. We laten alle variabelen in Model 4 één na één weg en berekenen telkens de waarde van de voorspellingsfout van dit submodel. Deze fout wordt vergeleken met de voorspellingsfout van Model 4, die weergegeven wordt in de laatste lijn van Tabel 21. De resultaten van deze analyse worden weergegeven in Tabel 22.

	Voorspellingsfout (test-data)	Verandering van voorspellingsfout (in procent)
Reference	*240,1126*	
Year	259,1269	7,9189
Month	243,9421	1,5949
Controle_Race	244,2481	1,7223
Category_Resc	252,7219	5,2514
Distance	238,8383	-0,5307
AVGspeed	240,9871	0,3642
Starters	239,6857	-0,1778
STF	240,5376	0,1770
Radio	221,9979	-7,5442
Rain	242,4728	0,9830
Climbs	250,8835	4,4858
Cobbles	225,5514	-6,0643
Shr_Perc	250,2006	4,2014

Tabel 22 Overzicht van de invloed van de variabelen op de voorspellingsfout van Model 4

Wanneer de variabele *year, month, controle_race, category_resc, AVGspeed, STF, rain, climbs* of *shr_perc* verwijderd wordt, neemt de waarde van de predictiefout van het model toe. Deze variabelen lijken dus een belangrijke bijdrage te leveren bij het voorspellen van het aantal valpartijen per wedstrijd en blijven daarom behouden.
Het weglaten van de overige variabelen doet de predictiefout dalen, daarom worden deze variabelen weggelaten.

Het finale regressiemodel wordt weergegeven als Model 5.

$$log\left(E(Y_{Crashes})\right) = \beta_0 + \beta_1 \times Year + \beta_2 \times Month + \beta_4 \times ControleRace + \beta_6 \times CategoryResc + \beta_8 \times AVGspeed + \beta_{12} \times STF + \beta_{14} \times Rain + \beta_{15} \times Climbs + \beta_{20} \times ShrPerc$$

Model 5 Voorspellingsmodel na tweede analyse van de variabelen

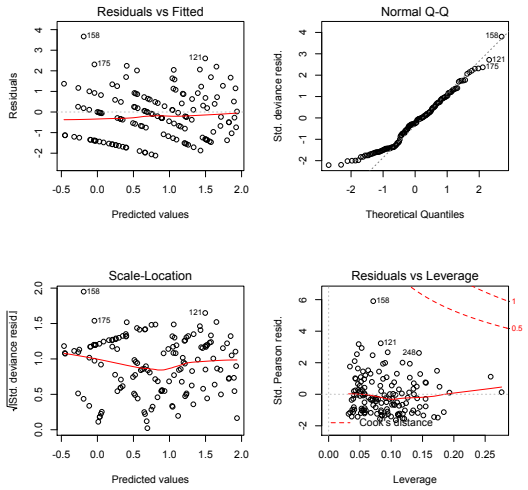
De waarden van de beoordelingscriteria van dit model worden samengevat in Tabel 23.

Resultaten voor Model 5	
AIC-waarde	566,92
Voorspellingsfout (trainingsdata)	2035,955
Voorspellingsfout (test-data)	196,4753

Tabel 23 Resultaten voor Model 5

Wanneer we de resultaten van deze tabel vergelijken met de waarden in Tabel 15 zien we dat de AIC-waarde van het model is toegenomen tot 566,92. In ruil biedt het model betere voorspellingen, want de waarde van de predictiefout van de trainingsdata daalt van 2246,113 naar 2035,955. Zoals verwacht neemt ook de voorspellingsfout van de test-data af; deze predictiefout daalt naar een waarde van 196,4753.

We beoordelen de kwaliteit van dit model aan de hand van de vier plots in Figuur 29.



Figuur 29 Plots voor Model 5

In de eerste grafiek worden de residuen en voorspelde waarden met elkaar vergeleken. Het gemiddelde van de residuen is kleiner dan nul. De rode trendlijn ligt immers beneden de stippellijn met vergelijking *y = 0.*

Uit de QQ-plot blijkt dat de gestandaardiseerde deviance residuen niet normaal verdeeld zijn. De observaties in de linker- en rechterbovenhoek wijken immers sterk af van de gestippelde 45°-lijn, die de kwantielen van een normale verdeling schetst.

Uit de derde grafiek kunnen we afleiden dat de observaties willekeurig verspreid liggen in een horizontale band rond de rode trendlijn. De trendlijn kent immers een vrij horizontaal verloop. We kunnen besluiten dat de variantie van de wortel uit de gestandaardiseerde deviance residuen onafhankelijk is van de waarde die door het model voorspeld wordt.

De laatste plot meet welke observaties gelijktijdig invloedrijk en afwijkend zijn. Geen van observaties is echter problematisch, daar alle datapunten binnen het gebied van de *Cook's distance* liggen.

De parameterschatters van Model 5 worden weergegeven in Tabel 24.

	Estimate	Std. Error	z-value	Pr(>|z|)	
(Intercept)	-102,8000	30,4300	-3,3790	0,0007	***
Year	0,0512	0,0154	3,3240	0,0009	***
Month	0,1409	0,1032	1,3650	0,1722	
Controle_RaceR2	-0,2444	0,2797	-0,8740	0,3823	
Controle_RaceR3	-1,1120	0,3316	-3,3540	0,0008	***
Controle_RaceR4	-1,7740	0,7661	-2,3150	0,0206	*
Category_RescC2	-0,5230	0,1982	-2,6390	0,0083	**
Category_RescC3	-0,4813	0,2815	-1,7100	0,0873	.
Category_RescC4	-1,2170	1,0410	-1,1690	0,2424	
AVGspeed	-0,0019	0,0359	-0,0530	0,9576	
STF	-0,0038	0,3618	-0,0100	0,9917	
RainYES	-0,0087	0,1171	-0,0750	0,9406	
Climbs	0,0020	0,0072	0,2830	0,7774	
Shr_Perc	0,0213	0,0081	2,6290	0,0086	**

Tabel 24 Parameterschatters voor Model 5

De schatter bij de variabele *year* is positief en significant verschillend van nul. Wanneer een koers één jaar later onder precies dezelfde omstandigheden georganiseerd wordt, telt deze wedstrijd gemiddeld $e^{0,0512}$ = 1,0525 meer valpartijen dan het jaar daarvoor.

De parameter bij de variabele *month* is positief, maar niet significant verschillend van nul. Hieruit kunnen we afleiden dat de wedstrijden aan het eind van het seizoen met meer valpartijen geassocieerd worden dan de koersen die plaatsvinden in het begin van het seizoen.

De schatters bij de vier niveaus van *controle_race* worden afgetoetst aan de koersen uit het openingsweekend (R1). Deze vergelijking toont aan dat de Waalse klassiekers (R3) en de najaarsklassiekers (R4) significant minder valpartijen tellen dan de wedstrijden uit het openingsweekend. De parameters bij deze niveaus van *controle_race* zijn immers negatief en significant verschillend van nul.

De waarde van de schatters bij de niveaus van *category_resc* worden vergeleken met de resultaten voor de wedstrijden uit de hoogste categorie (C1). Hieruit blijkt dat in de races van tweede categorie (C2), het huidige .BC (of .HC), significant minder gevallen wordt dan in de wedstrijden van het hoogste niveau.

De parameter bij de variabele *AVGspeed* is negatief, maar niet significant verschillend van nul. Een wedstrijd waarin de winnaar met een hogere gemiddelde snelheid over de finish rijdt, wordt dus met minder valpartijen geassocieerd dan een wedstrijd waarin de winnaar een lagere gemiddelde snelheid heeft.

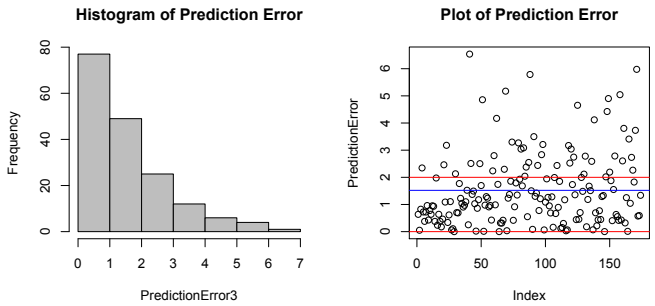
De variabele *STF* heeft een negatieve parameterschatter. Dit betekent dat een wedstrijd met een grotere verhouding tussen het aantal finishers en starters geassocieerd wordt met minder valpartijen.

Uit de tabel blijkt dat de schatter bij *rain* negatief is. Dit betekent dat een wedstrijd met regenweer met minder valpartijen geassocieerd wordt dan een wedstrijd waarin het droog blijft. Net als in de univariate analyses kan echter geen significant verschil tussen beide gemiddeldes aangetoond worden.

De parameterschatter bij *climbs* is positief. Een wedstrijd met veel hellingen wordt dus gerelateerd aan meer valpartijen. De parameter is echter niet significant verschillend van nul.

Tot slot toont de tabel aan dat de parameter bij het marktaandeel van de tv-uitzending van een wedstrijd positief en significant verschillend is van nul. Een wedstrijd met een groter marktaandeel wordt dus gerelateerd aan meer valpartijen.

Uit de onderste lijnen van Tabel 23 blijkt dat dit laatste model zowel de trainingsdata als de test-data goed kan voorspellen. In volgende analyse nemen we alle races in rekening.



Figuur 30 Analyse van de absolute waarden van de foutentermen bij het voorspellen van het aantal valpartijen per wedstrijd

Wanneer we de absolute waarde van de voorspellingsfout bij elke bestudeerde observatie *i* berekenen zoals aangegeven in uitdrukking (10), vinden we dat het model het werkelijke aantal valpartijen met gemiddeld 1,52 valpartijen over- of onderschat.

$$Absolute\ waarde\ van\ voorspellingsfout = \left|y_{Crashes(i)} - \hat{y}_{Crashes(i)}\right| \quad (10)$$

Uit Figuur 30 blijkt dat de voorspelling van het model nooit meer dan zeven valpartijen afwijkt van de werkelijke hoeveelheid valpartijen. Ook toont deze figuur aan dat het model in de meeste gevallen nul tot twee valpartijen te veel of te weinig voorspelt. De meeste observaties bevinden zich immers tussen de rode curves van de rechterplot. De blauwe trendlijn geeft het gemiddelde van de resultaten van formule (10) weer.

4.5 Besluit

De doelstelling van de onderzoeksvraag van deze masterproef bestaat er in deze factoren te identificeren die het aantal valpartijen per wedstrijd bepalen. Omdat de predictiefout van het laatste model lager is dan de voorspellingsfout van Model 1 en de modellen met de kleinste AIC-waarden, is dit regressiemodel mogelijk de beste manier om het aantal valpartijen per wedstrijd te voorspellen. Daarom zullen voornamelijk de resultaten van het laatste model gebruikt worden voor het beantwoorden van de verschillende deelvragen van de onderzoeksvraag, die besproken worden in het laatste, besluitende hoofdstuk van deze masterproef.

5 Algemeen besluit

De onderzoeksvraag van deze masterproef gaat op zoek naar de factoren die bepalend zijn voor het aantal valpartijen in een eendagswedstrijd in het professionele wielrennen. Hoewel deze vraag regelmatig aan bod komt in kranten, tijdschriften en op radio en TV is er weinig wetenschappelijk onderzoek naar valpartijen in het (professionele) wielrennen. Bij het beschrijven van de manier waarop de data voor dit onderzoek verzameld werden en welke bronnen hierbij geraadpleegd werden, wordt duidelijk waarom de wetenschappelijke literatuur omtrent dit onderwerp eerder beperkt is. Het is het dynamische karakter van de wielersport dat het verzamelen van gegevens bemoeilijkt.

Het verzamelen van data over de hoeveelheid valpartijen en de overige kenmerken van een wedstrijd is een persoonlijk werk, want de wielersport kent geen officiële databases die statistieken verzamelen over de verschillende wedstrijden die ieder jaar afgewerkt worden. Vervolgens worden de verzamelde data ingedeeld in training- en testgegevens. De eerste groep data bevat de kenmerken van de zestien edities van zestien Europese eendagskoersen die afgewerkt werden in de periode tussen februari 1997 en oktober 2012. De testgegevens beschrijven diezelfde kenmerken voor de wedstrijdedities van 2013 en 2014.

Eerst worden de gegevens bestudeerd in een reeks univariate analyses. Deze resultaten bieden een eerste kijk op het verband tussen de verschillende wedstrijdkenmerken en het aantal valpartijen per koers. Vervolgens worden alle variabelen samengebracht in één Poisson-regressiemodel. Hieruit worden deze variabelen geselecteerd die de AIC-waarde van het model laten dalen. Op die manier wordt een model gecreëerd dat niet alleen eenvoudig(er) is, maar er ook in slaagt om de trainingsdata goed weer te geven.
Deze techniek wordt een tweede keer toegepast. Dit keer wordt echter toegelaten dat het voorgestelde model ook interactietermen bevat. Op die manier wordt het verband tussen de verklarende variabelen onderling in rekening genomen.
In een laatste model wordt de voorspellingsfout van het model naar beneden gebracht. Zo wordt een model gecreëerd dat geschikt is voor het voorspellen van het aantal valpartijen per wedstrijd voor een verzameling nieuwe gegevens.

Op basis van de resultaten van de univariate analyses en de Poisson-regressiemodellen die ontwikkeld werden, kan een antwoord geformuleerd worden op de tien deelvragen van de onderzoeksvraag. De Poisson-regressiemodellen verduidelijken echter enkel het verband tussen het aantal valpartijen en de verschillende wedstrijdkenmerken en kunnen bijgevolg geen verklaring bieden voor de oorzaak van het valpartijenprobleem.

NEEMT HET AANTAL VALPARTIJEN TOE?

Het aantal valpartijen neemt ieder jaar toe. In de univariate analyse van de variabele *year* wordt duidelijk dat er een positief lineair verband bestaat tussen het jaar waarin een race gereden wordt en het aantal valpartijen in deze race.
In de Poisson-regressiemodellen is de parameterschatter bij de variabele *year* positief en significant verschillend van nul. Dit resultaat bevestigt niet alleen het toenemen van het aantal valpartijen, maar benadrukt ook dat het jaar waarin een wedstrijd afgewerkt wordt belangrijke informatie is bij het voorspellen van het aantal valpartijen per wedstrijd.

Het jaar waarin een wedstrijd afgewerkt wordt, capteert informatie over tal van factoren. Toekomstig onderzoek kan dit verband verder uitdiepen; wat is bijvoorbeeld het verband tussen het verbeterde materiaal en de wetenschappelijke begeleiding van de renners? En welke impact hebben (recente) wijzigingen in het UCI-reglement?

ZIJN ER WEDSTRIJDEN WAARIN MEER GEVALLEN WORDT DAN ANDEREN?

De ANOVA-analyse van de variabele *race* toont aan dat in verschillende koersen significant meer valpartijen geteld worden dan in andere races. De wedstrijden met een significant verschil in valpartijen werden geïdentificeerd met behulp van een Tukey-test. De twee eerste Poisson-regressiemodellen bevestigen deze vaststellingen. In beide modellen zijn de schatters bij de Omloop het Nieuwsblad, Dwars door Vlaanderen, de E3 Harelbeke, Gent-Wevelgem, de Scheldeprijs en de Waalse Pijl negatief en significant verschillend van nul. Deze wedstrijden tellen dus significant minder valpartijen dan de Amstel Gold Race. De parameterschatters bij Kuurne-Brussel-Kuurne, de Brabantse Pijl, Luik-Bastenaken-Luik, de Clasica San Sebastian en de Ronde van Lombardije zijn ook negatief, maar niet significant verschillend van nul.
Beide modellen schatten ook dat in Milaan-Sanremo meer gevallen wordt dan in de Amstel Gold Race, maar enkel in het Poisson-regressiemodel met interactietermen is deze schatter significant verschillend van nul.
Het gemiddeld aantal valpartijen per wedstrijd is dus gerelateerd aan de koers zelf. Met een Poisson-regressiemodel kan dit verschil vastgesteld, maar niet verklaard worden.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE CATEGORIE VAN EEN WEDSTRIJD?

Uit het finale Poisson-regressiemodel blijkt dat wedstrijden uit de hoogste categorie significant meer valpartijen tellen dan wedstrijden uit de tweede categorie. In de andere Poisson-regressiemodellen wordt deze informatie echter niet opgenomen.

Een overwinning of ereplaats in een wedstrijd van de hoogste categorie levert meer punten en publiciteit op voor een renner en zijn ploeg. Ook hebben deze wedstrijden een grotere prijzenpot te verdelen. Mogelijk hebben deze aspecten een invloed op het aantal valpartijen per wedstrijd. Deze analyse schiet echter tekort om een verklaring te formuleren voor deze vaststellingen. Het verband met de te verdelen punten en prijzen kan onderzocht worden in verder onderzoek.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE LENGTE VAN HET WEDSTRIJDPARCOURS?

De parameterschatters bij *distance* relateren een toename van de wedstrijdlengte met een daling van het aantal valpartijen. In de twee eerste modellen is de schatter van deze variabele echter niet significant verschillend van nul.

Deze vaststelling plaatst vraagtekens bij het inkorten van een wedstrijd in gevaarlijke omstandigheden. Een Poisson-regressiemodel kan geen causaal verband vaststellen, maar het is mogelijk dat het inkorten van een wedstrijdparcours leidt tot een toename van de hoeveelheid valpartijen.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE GEMIDDELDE SNELHEID VAN DE WINNAAR?

Eerder onderzoek toonde reeds aan dat de gemiddelde snelheid van de renners, en dus ook die van de winnaar, de laatste jaren alleen maar is toegenomen. De verkennende analyse van de gemiddelde snelheid van de winnaar leidt tot dezelfde conclusie. Het verband tussen het jaar waarin een wedstrijd plaatsvindt en de gemiddelde snelheid waarmee de winnaar de finish bereikt, is lineair en heeft een positieve richtingscoëfficiënt.

De variabele *AVGspeed* wordt enkel opgenomen in het laatste regressiemodel. Daar is de parameter negatief, maar niet significant verschillend van nul. Een wedstrijd wordt dus met meer valpartijen geassocieerd wanneer de winnaar een lagere gemiddelde snelheid heeft bij het bereiken van de finishlijn.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN HET AANTAL STARTERS IN EEN WEDSTRIJD?

De auteurs van de vele opiniestukken over de wielersport raden vaak aan om het peloton te verkleinen. In het bijzonder in de Vlaamse koersen, waar het peloton zich door de vele smalle wegen moet verplaatsen.

Uit de verkennende analyses blijkt dat een peloton van maximaal 150 deelnemers significant meer valpartijen telt dan een wedstrijd waarin respectievelijk 151 tot 175 en 176 tot 200 renners van start gaan.

Enkel het regressiemodel met interactietermen bevat specifieke informatie over het aantal starters in een wedstrijd. Het verband met het aantal starters is echter complex en wordt beïnvloed door verschillende interactietermen. Het verband tussen het aantal starters en de hoeveelheid valpartijen in een wedstrijd is dus niet eenduidig.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN HET GEBRUIK VAN OORTJES?

Een steeds terugkerende discussie is het gebruik van oortjes tijdens de wedstrijd. Deze vorm van draadloze communicatie wordt vaak afgeschilderd als de grote boosdoener van de vele valpartijen in het peloton.

Uit deze analyse blijkt echter dat het gebruik van oortjes weinig invloed heeft op het aantal valpartijen. In de univariate analyses werd geen significant verschil gevonden tussen het gemiddeld aantal valpartijen in een wedstrijd met, dan wel zonder oortjes.
De variabele maakt ook geen deel uit van de twee Poisson-regressiemodellen met de kleinste AIC-waarde en ook het laatste model neemt de variabele *radio* niet op. Dit geeft aan dat deze variabele weinig meerwaarde biedt bij het voorspellen van de hoeveelheid valpartijen per wedstrijd.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE WEERSOMSTANDIGHEDEN?

De weersomstandigheden werden slechts in beperkte mate in rekening genomen. Voor deze analyse werden immers enkel gegevens verzameld over de aan- of afwezigheid van neerslag tijdens de koers. Regen, sneeuw en hagel blijken echter geen significante invloed te hebben op de mate waarin gevallen wordt.

De resultaten van deze analyse bieden heel wat ruimte voor verder onderzoek. Zo kan het verband tussen het aantal valpartijen en de temperatuur en windsnelheden tijdens de wedstrijd nagegaan worden. Ook kan het verband met de neerslag verder verfijnd worden. Het is immers mogelijk dat het type neerslag, alsook de intensiteit van de regen-, hagel- of sneeuwbui een rol speelt bij de hoeveelheid valpartijen. Een ander aspect is het moment waarop het peloton getroffen wordt door neerslag. Het kan interessant zijn om na te gaan wat de invloed is van een regenbui aan het begin, dan wel eind van een wedstrijd.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN HET WEDSTRIJDPARCOURS?

Het parcours van een wedstrijd wordt getypeerd door de hellingen en kasseistroken die het bevat. Deze elementen hebben vaak een historische waarde en blijken verband te houden met de hoeveelheid valpartijen in een wedstrijd.

Uit de analyses blijkt dat een kasseikoers significant meer valpartijen telt dan een ‘gewone’ wedstrijd. In de univariate analyse van de dummy-variabele voor het aan- of afwezig zijn van kasseien en het eerste regressiemodel blijkt het verschil tussen beide soorten wedstrijden zelfs significant. Door in verder onderzoek rekening te houden met de plaats waar een valpartij plaatsvindt, kan onderzocht worden of de valpartijen ook plaatsvinden op de kasseistroken van het parcours.

De hellingen in het parcours blijken minder belangrijk bij het voorspellen van het aantal valpartijen per wedstrijd. In de univariate analyses werd een significant verschil vastgesteld tussen de verschillende wedstrijdgroepen, maar in het laatste regressiemodel is de parameterschatter bij het aantal hellingen per wedstrijd niet significant verschillend van nul.

WAT IS HET VERBAND TUSSEN HET AANTAL VALPARTIJEN EN DE MEDIA-AANDACHT DIE EEN WEDSTRIJD KRIJGT?

Uit de univariate analyses van het marktaandeel, de kijkdichtheid en het kijkcijfer van een wieleruitzending blijkt dat een toename van het aantal kijkers geassocieerd wordt met een stijging van het aantal valpartijen.

De regressiemodellen maken allemaal gebruik van deze informatie voor het verklaren van de hoeveelheid valpartijen per wedstrijd. Telkens wordt een positief verband geschat tussen het aantal kijkers en het aantal valpartijen. In het tweede en het laatste model is de parameter significant verschillend van nul.

De resultaten van de univariate analyses, de samenstelling van de regressiemodellen en de antwoorden op deze tien deelvragen leiden tot het besluit dat verschillende factoren een (belangrijke) rol spelen bij het verklaren van het aantal valpartijen per wedstrijd.

Deze analyse leert ons dat het gebruik van oortjes weinig invloed heeft op het aantal valpartijen in een wedstrijd. Deze variabele wordt immers niet opgenomen in één van de ontwikkelde modellen.

Het jaar en de maand waarin een wedstrijd gereden wordt, de naam van de wedstrijd, de categorie van de wedstrijd, de lengte van het wedstrijdparcours, de gemiddelde snelheid waarmee de winnaar de eindmeet bereikt, het aantal starters en finishers, de aan- of afwezigheid van kasseien en het marktaandeel van de tv-uitzending van een koers blijken wel belangrijke informatie te bevatten bij het voorspellen van het aantal valpartijen per wedstrijd, maar de parameters zijn niet altijd significant verschillend van nul.

De factoren die in deze analyse aan bod kwamen zijn echter niet de enige aspecten die een invloed kunnen hebben op het aantal valpartijen per wedstrijd. In verder onderzoek kan het verband met één of meerdere (significante) factoren verder uitgediept worden door meer gedetailleerde statistieken te verzamelen. Bij het beantwoorden van de tien deelvragen van de onderzoeksvraag werden reeds enkele voorbeelden gegeven van elementen die dieper onderzocht kunnen worden. Verder is het belangrijk na te gaan of de besproken factoren een gelijkaardige invloed hebben op het aantal valpartijen in de etappes van een rittenkoers. Ook is het mogelijk om de ernst van een valpartij in rekening te nemen. In deze analyse werd immers geen rekening gehouden met het aantal betrokken renners of de gevolgen van een valpartij. Een andere invalshoek bij het bestuderen van het aantal valpartijen in het (professionele) wielrennen is het bestuderen van de dynamiek in een peloton. De manier waarop een peloton beweegt en waar de renners die betrokken zijn in een valpartij zich bevinden, bieden waarschijnlijk meer inzicht in de manier waarop een valpartij ontstaat.

Bijlagen

Bijlage 1 – Overzicht van de variabelen

Variabele	
Crashes	Aantal valpartijen per wedstrijd.
AVGspeed	Gemiddelde snelheid van de winnaar van de wedstrijd.
Category	Originele categorie van de wedstrijd. Gebruik van verschillende competitiemodellen over de jaren heen.
Category_Resc	Aangepaste categorie van de wedstrijd. Vertaling van vier competitiemodellen naar één ordinale schaal. • C1 voor wedstrijden van categorieën CDM, PT en WT • C2 voor wedstrijden van categorie .BC • C3 voor wedstrijden van categorie .1 • C4 voor wedstrijden van categorie .2 • C5 voor wedstrijden van categorie .3 • C6 voor wedstrijden van categorie .4 • C7 voor wedstrijden van categorie .5
Climbs	Hoeveelheid hellingen in een parcours
Climbs_grouped	Categorische variabele die wedstrijden groepeert naar het aantal hellingen dat ze bevatten. • G1 voor wedstrijden die nul tot vier hellingen tellen • G2 voor wedstrijden die vijf tot tien hellingen tellen • G3 voor wedstrijden die elf tot twintig hellingen tellen • G4 voor wedstrijden die meer dan twintig hellingen tellen
Cobbles	Hoeveelheid kasseistroken in een parcours
Cobbles_Km	Hoeveelheid kasseistroken in een parcours, uitgedrukt in kilometer.
Controle_cobbles	Dummy-variabele die aangeeft of een wedstrijd kasseien bevat. • *Yes* voor wedstrijden met kasseien • *No* voor wedstrijden zonder kasseien
Controle_Race	Categorische variabele die wedstrijden groepeert naar hun volgorde op de huidige kalender. • R1 voor de openingswedstrijden • R2 voor het eerste deel van de voorjaarsklassiekers (Milaan-Sanremo, kasseiklassiekers en Brabantse Pijl) • R3 voor het tweede deel van de voorjaarsklassiekers (Amstel Gold Race en Waalse klassiekers) • R4 voor de najaarsklassiekers
Date	Datum waarop de wedstrijd doorging.
Distance	Lengte van de wedstrijd.

Finishers	Aantal finishers in een wedstrijd
Month	Maand waarin de wedstrijd gereden werd
Race	Naam van de wedstrijd. • AGR voor Amstel Gold Race • Bpijl voor Brabantse Pijl • CSS voor Clasica San Sebastian • DDV voor Dwars door Vlaanderen • E3H voor E3 Harelbeke (of E3 Prijs Harelbeke) • GW voor Gent-Wevelgem • KBK voor Kuurne-Brussel-Kuurne • LBL voor Luik-Bastenaken-Luik • MSR voor Milaan-Sanremo • OHN voor Omloop het Nieuwsblad (of Omloop het Volk) • PR voor Parijs-Roubaix • RVL voor Ronde van Lombardije • RVV voor Ronde van Vlaanderen • SP voor Scheldeprijs • VC voor Vattenfall Cyclassics (of HEW Cyclassics) • Wpijl voor Waalse Pijl
Radio	Dummy-variabele die aangeeft of de renners tijdens de wedstrijd gebruik maakten van oortjes. • *Yes* voor wedstrijden met oortjes • *No* voor wedstrijden zonder oortjes
Rain	Dummy-variabele die aangeeft of er neerslag (regen, hagel of sneeuw) viel tijdens de wedstrijd • *Yes* voor wedstrijden met neerslag • *No* voor wedstrijden zonder neerslag
Rat_Nr	Gemiddeld aantal kijkers van de tv-uitzending van een wedstrijd. Het gemiddelde wordt berekend over de volledige duur van de wedstrijd.
Rat_perc	Kijkdichtheid van de tv-uitzending van een wedstrijd. Percentage Vlamingen dat naar dit tv-programma kijkt, waarbij alle tv-toestellen in rekening gebracht worden.
Shr_Perc	Marktaandeel van de tv-uitzending van een wedstrijd. Percentage kijkers dat naar dit tv-programma kijkt, waarbij enkel de tv-toestellen die aanstaan in rekening gebracht worden.
Starters	Aantal starters in een wedstrijd.

Starters_grouped	Categorische variabele die wedstrijden groepeert op basis van het aantal deelnemers • ≤ 150 starters voor koersen met maximaal 150 deelnemers (25 ploegen met maximaal zes renners per team) • 151-175 starters voor koersen met 151 tot 175 deelnemers (25 ploegen met maximaal zeven renners per team) • 176-200 starters voor koersen met 176 tot 200 deelnemers (25 ploegen met maximaal acht renners per team) • > 200 starters voor koersen met meer dan 200 deelnemers (25 ploegen met meer dan acht renners per team)
STF	Verhouding tussen het aantal finishers en starters in een wedstrijd.
Year	Jaar waarin de wedstrijd gereden werd

Bijlage 2: Correlatiematrix van de numerieke variabelen

CORR	Year	Distance	AVGspeed	Starters	Finishers	STF	Climbs	Cobbles	Cobbles_Km	Shr_Perc	Rat_Nr	Rat_perc
Year	x	-0,097137	0,13024	-0,087405	0,223837	0,256931	0,128868	-0,021692	-0,02241	0,131346	0,272682	0,221552
Distance	-0,097137	1	-0,275733	0,201651	0,192185	0,134065	0,079155	0,209912	0,236947	0,078756	0,215856	0,224922
AVGspeed	0,13024	-0,275733	1	-0,16319	0,226975	0,266345	-0,316794	-0,114204	-0,09309	-0,076732	-0,227272	-0,241531
Starters	-0,087405	0,201651	-0,16319	1	0,22619	-0,028898	0,166678	0,088301	0,078977	0,23567	0,254029	0,260602
Finishers	0,223837	0,192185	0,226975	0,22619	1	0,96345	-0,03283	-0,19342	-0,198973	0,060037	-0,04875	-0,062725
STF	0,256931	0,134065	0,266345	-0,028898	0,96345	1	-0,08043	-0,213872	-0,219531	-0,00709	-0,118851	-0,134865
Climbs	0,128868	0,079155	-0,316794	0,166678	-0,03283	-0,08043	1	-0,16691	-0,262768	0,152014	0,144061	0,138526
Cobbles	-0,021692	0,209912	-0,114204	0,088301	-0,19342	-0,213872	-0,16691	1	0,933312	0,481178	0,637964	0,651296
Cobbles_Km	-0,02241	0,236947	-0,09309	0,078977	-0,198973	-0,219531	-0,262768	0,933312	1	0,438843	0,58691	0,599308
Shr_Perc	0,131346	0,078756	-0,076732	0,23567	0,060037	-0,00709	0,152014	0,481178	0,438843	1	0,713701	0,720224
Rat_Nr	0,272682	0,215856	-0,227272	0,254029	-0,04875	-0,118851	0,144061	0,637964	0,58691	0,713701	1	0,998211
Rat_perc	0,221552	0,224922	-0,241531	0,260602	-0,062725	-0,134865	0,138526	0,651296	0,599308	0,720224	0,998211	1

Lijst met figuren

Lijst met tabellen

6 Bibliografie

Blockeel, H. (2010). *Machine Learning and Inductive Inference.* Leuven: ACCO Uitgeverij.

Claeskens, G. (2013). *Statistical Modelling.* Leuven: ACCO Uitgeverij.

Croux, C. (2011). *Bedrijfsstatistiek.* Leuven, Belgium: Ekonomika VZW.

de Waard, D., Edlinger, K., & Brookhuis, K. (2011). Effects of listening to music, and of using a handheld and handsfree telephone on cycling behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour , 14* (6), 626-637.

El Helou, N., Berthelot, G., Thibault, V., Tafflet, M., Nassif, H., Campion, F., et al. (2012). Tour de France, Giro, Vuelta, and classic European races show a unique progression of road cycling speed in the last 20 years. *Journal of Sport Sciences , 28* (7), 789-796.

Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident and Analysis Prevention , 40* (3), 1033-1054.

Elvik, R., Høye, A., & Vaa, T. (2009). *Handbook of Road Safety Measures.* Bingley Emerald Group Publishing.

Goldenbeld, C., Houtenbos, M., Ehlers, E., & De Waard, D. (2011). The use and risk of portable electronic devices while cycling among different age groups. *Journal of Safety Research , 43* (1), 1-8.

Gueguen, G. (2011). Lorsque l'innovation modifie les règles du jeu - les oreillettes des cyclistes en question. *Le Mag des Sciences de Gestion* (3), 3 - 6.

Jansen, M., & Claeskens, G. (2010). *Kansrekenen en beschrijvende statistiek.* Leuven, Belgium: ACCO Uitgeverij.

Johnson, M., Oxley, J., & Cameron, M. (2009). *Cyclist bunch riding: a review of literature.* Victoria: Monash University Accident Research Centre.

Leonard, L., Lim, A., Chesser, T. J., Norton, S. A., & Nolan, J. P. (2005). Does changing the configuration of a motor racing circuit make it safer? *British Journal of Sports Medicine , 39*, 159-161.

Lippi, G., & Guidi, G. C. (2005). Effective measures to improve driver safety. *British Journal of Sports Medicine , 39* (9), 686-688.

Lippi, G., Sanchis-Gomar, F., & Favoloro, E. (2011). Cycling: To race or to live? - Reflections on Skewed Priorities. *International Journal of Sports Medicine , 32* (8), 648-649.

Lodewijkx, H. F., & Brouwer, B. (2012). Tour, Giro, Vuelta: Rapid Progress in Cycling Performance Starts in the 1980s. *International Journal of Sports Science , 2* (3), 24-31.

Lybbert, T. J., Lybbert, T. C., Smith, A., & Warren, S. (2012). Does the Red Flag Rule Induce Risk Taking in Sprint Finishes? Moral Hazard Crashes in Cycling's Grand Tours. *Journal of Sports Economics , 13* (6), 603-618.

McClave, J. T., Benson, G. P., Sincich, T., & Knypstra, S. (2011). *Statistiek - Een inleiding.* Amsterdam: Pearson Education Benelux.

McLennan, J. G., McLennan, J. C., & Ungersma, J. (1988). Accident Prevention in Competitive Cycling. *The American Journal of Sports Medicine , 16* (3), 266-268.

Morrow, S., & Idle, C. (2008). Understanding Change in Professional Road Cycling. *European Sport Management Quarterly , 8* (4), 315-335.

Nour El Helou, G. B.-F. (2010). Tour de France, Giro, Vuelta, and classic European races show a unique progression of road cycling speed in the last 20 years. *Journal of Sport Sciences , 28* (7), 789-796.

Organisation for Economic Co-operation and Development. (2013). Analysis of international trends in bicycle use and cyclist safety. *Cycling, Health and Safety* .

Organisation for Economic Co-operation and Development. (2013). Overview of bicycle crash characteristics in selected countries. *Cycling, Health and Safety* .

Organisation for Economic Co-operation and Development. (2013). Review of bicycle safety measures. *Cycling, Health and Safety* .

Perneger, T. V. (2010). Speed Trends of Major Cycling Races: Does Slower Mean Cleaner? *International Journal of Sports Medicine , 31* (4), 261-264.

Roi, S. G., & Tinti, R. (2014). Requests for medical assistance during an amateur road cycling race. *Accident Analysis and Prevention , 73*, 170-173.

Thompson, D. C., & Patterson, M. Q. (1998). Cycle Helmets and the Prevention of Injuries - Recommendations for Competitive Sport. *Journal of Sports Medicine , 25* (4), 213-219.

Townes, D. A., Barsotti, C., & Cromeans, M. (2005). Injury and Illness During a Multiday Recreational Bicycling Tour. *Wilderness and Environmental Medicine , 16* (3), 125-128.

Van Reeth, D. (2013). TV Demand for the Tour de France: The Importance of Stage Characteristics versus Outcome Uncertainty, Patriotism, and Doping. *International Journal of Sports Finance , 8* (1), 39-60.

Websites

Bike Race Info. (2015). *Bike Race Info*. Opgeroepen op 2015, van Bike Race Info: http://www.bikeraceinfo.com

Cycling News. (2015). *Cycling News*. Opgeroepen op 2015, van Cycling News: www.cyclingnews.com

ProCyclingStats. (2015). *ProCyclingStats*. Opgeroepen op 2015, van ProCyclingStats: www.procyclingstats.com

Sporza Wielrennen. (2015). *Sporza Wielrennen*. Opgeroepen op 2015, van Sporza: www.sporza.be/wielrennen

WV Cycling. (2002). *WV Cycling*. Opgeroepen op 2015, van WV Cycling: http://www.wvcycling.com/

Opiniestukken

Lagae, W. (2013, July). De Tour, waar spektakel primeert op veiligheid. De Standaard.

Lagae, W. (2013, March). De wielerwereld draait door. De Standaard.

Lagae, W. (2011, September). Hertekening van het Ronde parcours is een zegen. De Standaard.

Lagae, W. (2011, May). Wielersport moet veel veiliger. De Standaard.

Lagae, W. (2011, July). Wielersport zit in een Gordiaanse knoop. De Standaard.

Maeckelbergh, B. (2014, April). Zes oplossingen voor valpartijen in het peloton - Wat veroorzaakt de valpartijen in het peloton, en hoe lossen we dat op? De Morgen.

Online documenten

Centrum voor Informatie over de Media. (2011, February). *Centrum voor Informatie over de Media.* Opgeroepen op 2015, van CIM: www.cim.be

Union Cycliste International. (2013, March). *Organiser's Guide to Road Events.* Opgeroepen op 2015, van Union Cycliste International: www.uci.ch

FACULTEIT ECONOMIE EN BEDRIJFSWETENSCHAPPEN
Naamsetraat 69 bus 3500
3000 LEUVEN, België
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be