Erasmus Universiteit Rotterdam

Erasmus School of Economics

Bachelorscriptie [programma Strategy Economics]

Tennissers op zoek naar efficiëntie: Welk toernooi levert makkelijk veel punten op?

Naam student: Steef ter Horst

Studentnummer: 505574

Begeleider: heer H. Van Kippersluis

Tweede beoordelaar: heer S. Hoey

Datum definitieve versie: 16-07-2022

Abstract

In dit paper wordt een model opgesteld die het gemiddeld aantal punten van ongeplaatste tennisspelers per toernooi voorspelt, om zo het meest efficiënte toernooi om aan deel te nemen te achterhalen. De geplaatste spelers worden niet meegenomen in het model, omdat wordt verwacht dat zij meer van eigen kracht uit gaan en daarom niet de kans op punten voor de wereldranglijst als grootste drijfveer zien bij een toernooikeuze. Deze aanname is eerst met een Regressie Discontinuïteit Design getest om te onderzoeken of het terecht is dat geplaatste spelers niet zijn opgenomen in het model. Hieruit bleek dat geplaatste status een positief significant effect heeft op het aantal behaalde punten. Dit betekent dat geplaatste spelers voor aanvang van het toernooi al een voordeel hebben ten opzichte van ongeplaatste spelers en dit bevestigt dus de aanname dat het model minder interessant is voor geplaatste spelers. Vervolgens is door middel van een Meervoudige Regressie de beste combinatie qua variabelen onderzocht om het aantal punten van ongeplaatste spelers per toernooi het best te voorspellen, hierna is de daadwerkelijke voorspelling gemaakt aan de hand van de gevonden coëfficiënten. De resultaten lieten zien dat Rio de Janeiro, Hamburg en Londen de meest efficiënte ATP500 toernooien zijn en Adelaide1, Adelaide2 en Sydney de meest efficiënte ATP250 toernooien. Ook kon geconcludeerd worden dat de hogere beloningen bij een ATP500 toernooi ten opzichte van een ATP250 toernooi het concurrentie-effect irrelevant maken. Bij het maken van een toernooikeuze tussen twee toernooien met gelijke beloningen is het echter wel efficiënt om het concurrentie-effect zwaarder te laten wegen dan het kenniseffect.

Inhoudsopgave

Introductie

In het verleden is vaker onderzoek gedaan naar wat bedrijven prikkelt om tot een markt toe te treden. Zo zouden bedrijven eerder geneigd zijn om tot een markt toe te treden waar hogere omzetten worden behaald en tot een markt die een soortgelijke structuur heeft als de markt waar het bedrijf al ervaring in heeft wat betreft vraag en aanbod (Morton, 1999). In het paper van Barbosa (2007) wordt bij data van Portugese productiebedrijven gekeken of het 'kenniseffect' of het 'concurrentie-effect' zwaarder weegt bij de afweging om tot een markt toe te treden. In een markt met veel concurrentie zal het lastiger worden om succesvol te zijn als bedrijf, maar kan wel sneller geleerd worden van andere bedrijven en dus meer kennis vergaart worden. In dit specifieke onderzoek wogen de effecten van nieuwe kennis vergaren zwaarder dan de concurrentie, waardoor bedrijven dus eerder een risicozoekende keuze maakte voor een drukke markt met veel concurrentie. Deze prikkels zeggen echter niet of deze keuzes achteraf ook efficiënt bleken voor het bedrijf op basis van bijvoorbeeld winst. Bresnahan en Reiss (1991) deden in hun studie wel meer onderzoek naar de vraag of het effectief was tot een markt toe te treden met het oog op de ruimte in de markt. Wanneer te veel bedrijven tot dezelfde markt toe treden, is het volgens hen nauwelijks meer mogelijk om winst te behalen door de sterke concurrentiestrijd op het gebied van prijs. Daarom onderzochten zij welke markt eigenschappen significante invloed uitoefenden op de winst, zodat ze vervolgens konden bestuderen tot en met welk aantal bedrijven in een bepaalde markt winst behaald kon worden. Dit aantal is dan het maximale aantal bedrijven in een markt tot wanneer het efficiënt blijft om toe te treden.

In dit paper zal sportdata gebruikt worden om een soortgelijke invalshoek als Bresnahan en Reiss (1991) te testen op de tenniswereld. Niet de stimulansen om tot een markt toe te treden staan dus centraal, maar de vraag of het efficiënt is om toe te treden. Daarnaast worden ook twee soorten toernooien geobserveerd die verschillen in grootte, waardoor er ook conclusies getrokken kunnen worden over de verhouding tussen het kenniseffect en het concurrentie-effect in de tenniswereld. Sportdata biedt vaak interessante mogelijkheden om dit soort economische theorieën te testen, omdat de activiteiten vaak als natuurlijk experiment gebruikt kunnen worden en er simpelweg veel data beschikbaar is. Om de theorie over markttoetreding op de tenniswereld te testen, worden de spelers als de bedrijven gezien en de verschillende toernooien als de markten waar ze aan kunnen deelnemen. Doel van het onderzoek is vervolgens om een toernooi of toernooien te vinden waaraan een speler sowieso moet deelnemen als hij makkelijk veel punten voor de wereldranglijst wil verdienen. Dit is dan namelijk de meest efficiënte markt om zo snel mogelijk posities te stijgen op de wereldranglijst. De verwachting is dat topspelers vaak van eigen kracht uit gaan en dus minder geïnteresseerd zullen zijn in een hogere kans op iets meer punten bij een bepaald toernooi. Om die

reden bestaat de afhankelijke variabele in dit onderzoek alleen uit het gemiddeld aantal punten per toernooi van de ongeplaatste spelers, aangezien die structureel wat lager geklasseerd zijn. De onderzoeksvraag luidt dan ook:

> *Welke toernooi karakteristieken hebben invloed op het aantal behaalde punten van ongeplaatste spelers en bij welke toernooien ligt dit aantal punten het hoogst?*

De wetenschappelijke relevantie van dit paper is dat het een onderzoeksvraag probeert te beantwoorden die nog niet eerder onderzocht is. Dit maakt het onderzoek uniek en vult daarmee een leeg gat binnen de kennis van sport economie. Daarnaast laat het zien of het mogelijk is om een economische theorie ook effectief toe te kunnen passen op verschillende soorten markten zoals de tenniswereld. Het is maatschappelijk relevant aangezien ongeplaatste spelers nu een rationele keuze wat betreft punten kunnen maken voor een bepaald toernooi om op die manier hun kans te vergroten om op de wereldranglijst te stijgen. Ook biedt het onderzoekers meer inzicht in welke toernooi eigenschappen significante invloed hebben op het aantal punten behaald door ongeplaatste spelers per toernooi.

In het vervolg van dit paper zullen eerst de hypotheses besproken worden die zijn opgesteld aan de hand van de literatuur. Daarna worden de keuzes voor de variabelen bij het maken van de dataset toegelicht en de onderzoeksmethoden uitgelegd die gebruikt gaan worden. Vervolgens worden de resultaten geanalyseerd en aan de hand daarvan worden de conclusie en discussie besproken. Afsluitend worden nog aanbevelingen voor vervolgonderzoek gegeven.

Literatuur + Hypotheses

De eerste hypothese is puur om te testen in hoeverre de aanname klopt die bij het opstellen van de centrale onderzoeksvraag is gemaakt, namelijk dat een model die de kans op punten per toernooi voorspelt alleen nuttig is voor ongeplaatste spelers. Geplaatste spelers zouden namelijk meer van eigen kracht uit gaan en andere factoren als een grotere drijfveer zien bij een toernooikeuze zoals bijvoorbeeld prijzengeld.

1) *Geplaatste spelers behalen gemiddeld meer punten per toernooi dan ongeplaatste spelers*

Om deze hypothese te testen ben ik op zoek naar het causale effect van een geplaatste status op het aantal behaalde punten per toernooi. Om dit te onderzoeken zal het effect van ranking gescheiden moeten worden, aangezien geplaatste spelers automatisch een betere ranking hebben dan ongeplaatste spelers. De kwaliteit zal naar verwachting niet veel verschillen tussen bijvoorbeeld de nummer 45 en de nummer 50 op de wereldranglijst, maar omdat alleen de acht beste spelers die deelnemen aan het toernooi een geplaatste status ontvangen, kan het zijn dat de nummer 45 nog geplaatst is, maar de nummer 50 net niet. Dit geeft de nummer 45 een voordeel over de nummer 50 aangezien geplaatste spelers elkaar op basis van het speelschema ontlopen in de eerste rondes en de nummer 45 op papier dus makkelijkere tegenstanders zal treffen in de eerste rondes (Simonsson, 2022). Deze hypothese gaat dus over het achterhalen of het hebben van een makkelijker speelschema in de eerste rondes, die dus samenhangt met het hebben van een geplaatste status, een significant positief effect heeft op het aantal behaalde punten per toernooi.

De overige hypotheses zijn gebaseerd op literatuur die verwachtingen scheppen over meerdere variabelen die invloed hebben op de uitkomst van tenniswedstrijden. Het testen van deze hypotheses zorgt ervoor dat ik goed kan bepalen welke variabele een significant effect hebben op mijn afhankelijke variabele en welke variabelen dus ook belangrijk zijn om in het model te behouden wanneer ik de afhankelijke variabele wil voorspellen.

2) *Voor ongeplaatste spelers kan het efficiënter zijn aan een kleiner ATP250 toernooi deel te nemen dan aan een groter ATP500 toernooi*

Bij een ATP500 toernooi zijn meer punten voor de wereldranglijst te verdienen en wordt ook meer prijzengeld uitgereikt dan bij een ATP250 toernooi, waardoor het op voor hand aantrekkelijker lijkt om aan een ATP500 deel te nemen dan aan een ATP250. Volgens onderzoek van Morton (1999) is het namelijk zo dat een markt waar hogere beloningen te verdienen zijn tot meer intreding van bedrijven lijdt. Dit betekent echter ook dat veel topspelers er snel voor zullen kiezen om aan het ATP500 toernooi deel te nemen en dat de concurrentie hier dus sterk is. Als spelers daadwerkelijk eerder geneigd zijn

om aan het ATP500 toernooi deel te nemen vanwege de grotere beloningen, laten zij het kenniseffect dus zwaarder wegen dan het concurrentie-effect. Morton lichtte overigens wel toe dat het feit dat bedrijven sneller intreden in markten met hogere beloningen niks zegt over de effectiviteit ervan. Mijn verwachting is dan ook dat voor ongeplaatste spelers het op basis van punten effectiever is om juist het concurrentie-effect zwaarder te laten wegen dan het kenniseffect. Ongeplaatste spelers komen over het algemeen qua niveau te kort ten opzichte van de topspelers, waardoor het kleinere ATP250 meer kansen biedt om verder in het toernooi te komen. Ter vergelijking levert een 3e ronde op een ATP250 toernooi alsnog meer punten op dan de 1e ronde op een ATP500 toernooi. (SteveGTennis, 2020)

3) *Ongeplaatste spelers behalen gemiddeld meer punten per toernooi wanneer er meerdere toernooien tegelijk worden georganiseerd*

Wanneer de keuze uit toernooien groter is, wordt waarschijnlijk de spreiding van de spelers groter. Hierdoor verwacht ik dat de concurrentie per toernooi afzwakt en dat meer toernooien tegelijkertijd de kans op punten voor ongeplaatste spelers vergroot. Daarnaast zorgt het grotere aanbod van toernooien ervoor dat spelers eerder de keuze hebben uit een toernooi waar ze van tevoren meer kans achten te maken om wedstrijden te winnen. Koning (2011) toonde bijvoorbeeld aan dat mannelijke tennisspelers beduidend beter presteren voor thuispubliek en spelers kunnen ook persoonlijke voorkeuren hebben voor een bepaalde ondergrond. Anno juli 2022 heeft nummer 6 van de wereld Casper Ruud bijvoorbeeld een winstratio van 72,5% op gravel tegenover maar 33,3% op gras (UltimateTennisStatistics, 2022).

4) *Ongeplaatste spelers behalen gemiddeld meer punten per toernooi wanneer de ranking van de geplaatste spelers hoger ligt*

Uit onderzoek van Del Corral en Prieto-Rodirguez (2010) bleek dat het verschil in ranking op de wereldranglijst tussen twee spelers namelijk een significant effect had in het voorspellen van de winnaar van een tenniswedstrijd. Een groot verschil in ranking tussen beide spelers betekende vaak dat de speler met een lagere absolute ranking, wat dus in feite een betere ranking is, de wedstrijd wist te winnen. Als de gemiddelde ranking van de geplaatste spelers dus hoog ligt, zijn het relatief gezien niet de allerbeste spelers. Dit betekent dat de kwaliteit van de concurrentie relatief laag ligt en dat het concurrentie-effect dus minder zwaar weegt. De verwachting is daarom dat het aantal punten voor ongeplaatste spelers en de gemiddelde ranking van de geplaatste spelers een positief verband hebben.

Data

Om de hypotheses aan de hand van empirisch onderzoek te testen, heb ik van de periode 2012-2022 data verzameld van *ATP250* en *ATP500* tennistoernooien uit het mannenenkelspel die in 2022 op de kalender staan (ATP Tour, 2022). Bij deze toernooien kan zoals de naam suggereert maximaal 250 of 500 punten voor de wereldranglijst verdiend worden door de winnaar. Dit aantal punten maakt het de twee kleinste toernooien die gehouden worden op de ATP tour, aangezien er bij de overige ATP1000 toernooien en Grand Slams meer deelnemers per toernooi zijn en dus ook meer punten verdiend kunnen worden. De reden dat de grotere toernooien niet zijn opgenomen in de dataset is, omdat in de weken dat deze toernooien gehouden worden er niet meerdere toernooien tegelijk zijn. Dit betekent dat in deze weken de spelers niet tussen meerdere toernooien hoeven te kiezen om aan deel te nemen en dat de vraag welk toernooi efficiënter is om aan deel te nemen wat betreft punten voor de wereldranglijst overbodig is. Daarnaast worden de kleinere toernooien vaker georganiseerd, waardoor er nog steeds genoeg variatie en observaties beschikbaar blijven voor een representatieve dataset.

De afhankelijke variabele die in dit paper zo goed mogelijk voorspeld tracht te worden is *gemiddelde_punten*. Het gaat hierbij om de gemiddelde punten van de ongeplaatste spelers per toernooi. Om te achterhalen welke ronde de ongeplaatste spelers bereikte op de geobserveerde toernooien zijn alle speelschema's geanalyseerd (Flashscore, 2022). Vervolgens wordt er altijd een vast aantal punten per behaalde ronde aan de spelers toegekend voor de wereldranglijst. Deze punten verschillen per toernooi type, want zoals eerder benoemd, kunnen bij ATP500 toernooien meer punten worden verdiend dan bij ATP250 toernooien (SteveGTennis, 2020).

Eén van de variabelen die is toegevoegd aan het model is het type ondergrond waar het toernooi op gespeeld wordt (ATP Tour, 2022). In de analyse wordt onderscheid gemaakt tussen *hardcourt*, *gravel* en *gras*. Johnson en McHugh (2005) deden namelijk onderzoek naar het gemiddeld aantal slagen per servicegame. Hier kwam uit voort dat er tijdens Roland Garros op het gravel gemiddeld 6,0 keer een topspin forehand en 4,2 keer een topspin backhand gespeeld per servicegame, terwijl op het gras van Wimbledon gemiddeld 2,9 keer een topspin forehand en 1,3 keer een topspin backhand werd gespeeld. De rally's op het gras waren dus aanzienlijk korter wat aangeeft dat de service minder vaak geretourneerd werd. Dit bewijst dat de service dominanter is op gras dan op gravel, omdat in de service statistieken geen duidelijke verschillen waren tussen beide. Spelers kunnen met goed serveerwerk op gras de rally's dus korter houden en zo de zwakheden in hun spel makkelijker verbloemen. Het feit dat er zulke duidelijke verschillen zijn tussen de ondergronden maakt het belangrijk deze variabele toe te voegen aan het model.

Verder zijn er voor elk toernooi twee variabelen toegevoegd die aangeven hoeveel toernooien er gelijktijdig gehouden worden (ATP Tour, 2022). Ook hierbij wordt onderscheidt gemaakt tussen *ATP250_gelijktijdig* en *ATP500_gelijktijdig*, aangezien beide varianten een ander effect kunnen hebben op het aantal behaalde punten door ongeplaatste spelers.

Daarnaast is de variabele *continent* opgenomen in het model om te testen of de verschillende omstandigheden per continent effect hebben op de *gemiddelde_punten* (Wikipedia, 2022). Warmere temperaturen zorgen er bijvoorbeeld voor dat de hartslag omhoog gaat bij spelers, wat zorgt voor een hogere intensiteit tijdens de wedstrijd (Fernandez et al., 2006). Alle continenten zijn opgenomen in het model behalve Antarctica en Afrika. Op Antarctica wordt simpelweg geen tennistoernooi georganiseerd en in Afrika maar eentje. Aangezien maar één tennistoernooi geen betrouwbare weerspiegeling geeft van de coëfficiënt voor het continent Afrika is ervoor gekozen om het toernooi van Marrakesh niet mee te nemen in de analyse.

Om de sterkte van de concurrentie te beoordelen ben ik bij elk toernooi nagegaan wie de acht best genoteerde spelers en dus de geplaatste spelers waren (Flashscore, 2022). Vervolgens heb ik van deze geplaatste spelers hun ranking op de wereldranglijst opgezocht op het moment van het toernooi. Deze ranking is tot op de week nauwkeurig, aangezien de wereldranglijst elk begin van de week wordt vernieuwd (ATP Tour, 2022). Aan de hand van deze getallen is de variabele *gemiddelde_ranking* toegevoegd die de gemiddelde ranking van de geplaatste spelers weergeeft. Wanneer deze waarde laag ligt, betekent het dus dat theoretisch gezien betere spelers deelnemen aan het toernooi aangezien de nummer 1 van de wereldranglijst de beste positie is. Ook zijn de individuele ranking van de als eerste en de als achtste geplaatste speler toegevoegd, als respectievelijk *hoogste_ranking* en *laagste_ranking*. Dit is om te onderzoeken of er een verband te vinden is tussen de minimale en maximale ranking van de geplaatste spelers tijdens een toernooi en de *gemiddelde_punten*. De tennissport wordt bijvoorbeeld al jarenlang gedomineerd door de grote drie: Novak Djokovic, Rafael Nadal en Roger Federer. Vanwege hun dominantie zou het daarom bijvoorbeeld kunnen zijn dat de variabele *hoogste_ranking* een significant negatief effect heeft op *gemiddelde_punten*, omdat deze drie mannen enorm veel winnen.

Methodologie

De zelf gecreëerde dataset met alle zojuist behandelde variabelen zijn verzameld in een Excel document en vervolgens naar het programma Stata geïmporteerd. In dit programma zijn vervolgens alle statistische analyses uit dit onderzoek uitgevoerd.

De analyse begint door te kijken naar wat het effect van *geplaatste_status* is op *punten*. Om dit te onderzoeken is er gekozen voor een Regressie Discontinuïteit Design (RDD). Deze methode is zeer geschikt aangezien er een duidelijk omslagpunt in de data zit vanaf wanneer een speler de treatment variabele *geplaatste_status* ontvangt. Elke speler die voor de variabele *nummer* namelijk lager dan 8,5 scoort, ontvangt de treatment. Om deze reden is er ook voor een Sharp RDD gekozen over een Fuzzy RDD, omdat de verdeling tussen controle-en-treatmentgroep volledig wordt bepaald door de variabele *nummer*. De analyse is uitgevoerd op basis van de meest recente editie van elk toernooi en met spelers van *nummer* 4 tot en met 13. Ook de *ranking* van deze spelers is verzameld, zodat hiervoor gecontroleerd kan worden op zoek naar het causale effect van *geplaatste_status*.

Het uitvoeren van een RDD-analyse is een handige methode, omdat er observaties binnen een smalle bandbreedte rond de grenslijn worden vergeleken. In dit geval worden er dus spelers vergeleken die net wel een *geplaatste_status* hebben ontvangen met spelers voor wie dit net niet gold. Dit brengt als voordeel met zich mee dat er waarschijnlijk heel weinig verschil zal zitten tussen de karakteristieken van deze spelers, omdat ze bijvoorbeeld op het gebied van kwaliteit heel dichtbij elkaar liggen. Wanneer je genoeg observaties binnen die smalle bandbreedte rond de grenslijn kunt verzamelen en vergelijken, zal er uiteindelijk zo goed als perfecte randomisatie worden bereikt tussen deze observaties met dank aan de wet van de grote getallen. Door deze randomisatie kan er vervolgens vanuit gegaan worden dat de verschillen tussen observaties aan de verschillende kanten van de grenslijn, puur kunnen worden toegewezen aan het treatment effect van *geplaatste_status*. De RDD-analyse ziet er als volgt uit:

$$Punten = \beta 0 + \beta 1 * Gecentreerd\_nummer + \beta 2 * Geplaatste\_status + \beta 3 * Gecentreerd\_nummer\#Geplaatste\_status$$

De variabele *nummer* is gecentreerd door het te verminderen met 8,5, zodat de waarde hiervan op de grenslijn gelijk is aan nul. Hierdoor geeft de coëfficiënt van *geplaatste_status* in het model precies het gemiddelde treatment effect weer. Ook wordt er door het toevoegen van de interactieterm gezorgd dat het mogelijk is twee verschillende polynomen aan beide kanten van de grenslijn te hebben.

De eerste onderzoeksmethode die gebruikt wordt is een meervoudige regressie. Door middel van deze onderzoeksmethode kunnen we aan de hand van data een inschatting maken wat het effect van een

variabele is op in dit geval de afhankelijke variabele *gemiddelde_punten*. Door de richting van de coëfficiënt van de variabele en de significantie kunnen vervolgens de hypotheses bevestigd of ontkracht worden. Daarnaast geeft de Adjusted R-squared van een meervoudige regressie aan in hoeverre het gehele model de afhankelijke variabele goed voorspelt. Dit is weer belangrijk bij het onderzoeken welke combinatie van variabelen het beste model biedt om de hoofdvraag in dit onderzoek te kunnen beantwoorden. De meervoudige regressie waarbij alle variabelen uit de dataset zijn toegevoegd ziet er als volgt uit:

*Gemiddelde_punten = β0 + β1 * ATP500 + β2 * ATP500_gelijktijdig + β3 * ATP250_gelijktijdig + β4 * Ondergrond + β5 * Continent + β6 * Hoogste_ranking + β7 * Laagste_ranking + β8 * Gemiddelde_ranking*

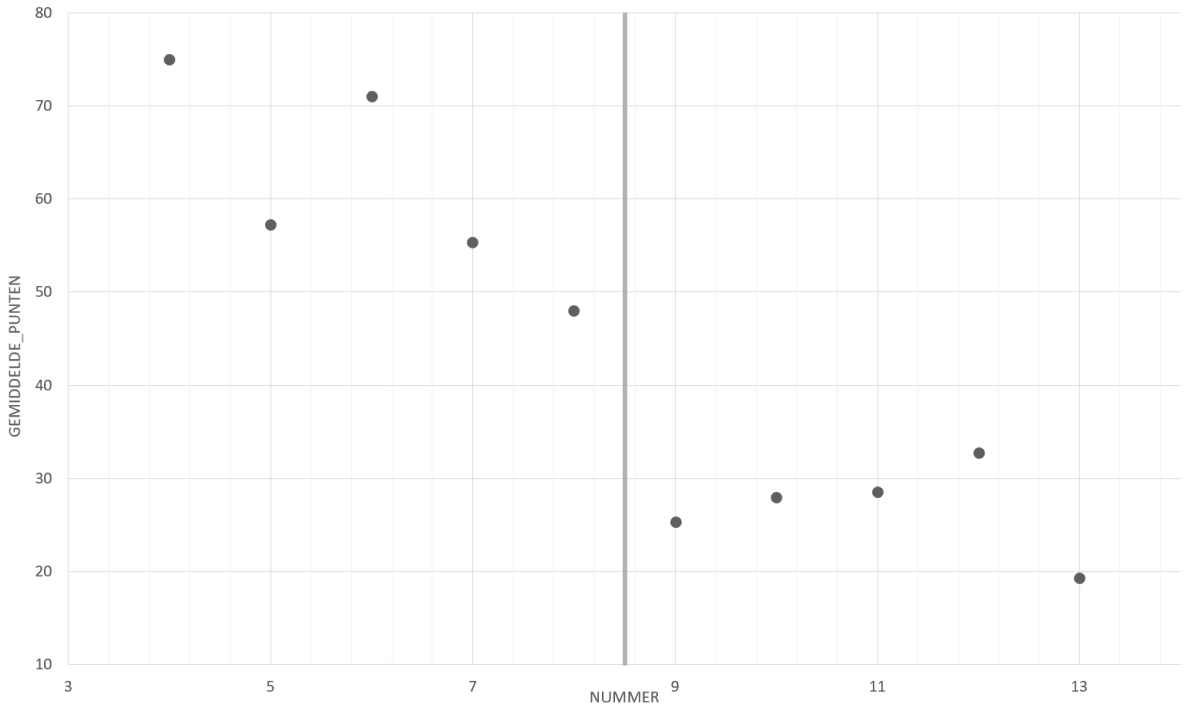
Vervolgens kan op basis van de coëfficiënten die een gemiddeld effect op *gemiddelde_punten* aangeven, een voorspelling worden gemaakt voor individuele toernooien waarvan de variabelen uit het model bekend zijn.

Aangezien alle variabelen die verzameld zijn voor deze analyses karakteristieken en statistieken van spelers en toernooien zijn, die op een ander moment in tijd niet ineens een andere waarde kunnen aannemen, is de betrouwbaarheid in dit onderzoek erg hoog. Bij herhaling van het onderzoek op een ander moment zullen namelijk ongeveer dezelfde resultaten gevonden worden. De validiteit is ook erg hoog, omdat de data van officiële organisaties is gehaald die actief zijn in of nauw betrokken bij de tenniswereld. Daarnaast is met 2012-2022 een periode gekozen die het meest actueel is om het model op te baseren en waarin ongeveer dezelfde generatie aan tennisspelers actief is. De validiteit daalt wel wat, omdat het aantal observaties wat aan de lage kant is. Na het bekijken van de resultaten zal in de discussie nogmaals terug gekeken worden op de betrouwbaarheid en validiteit van het onderzoek.

Resultaten

Het eerste gedeelte van de analyse bestudeert in hoeverre de aanname klopt dat het model alleen nuttig is voor de lager-geklasseerde ongeplaatste spelers. Bij de start van het onderzoek werd namelijk verondersteld dat geplaatste spelers eerder van hun eigen kwaliteit uit zullen gaan en daarom kans op punten niet als de belangrijkste factor zien bij het maken van een toernooikeuze. Maar is het wel daadwerkelijk zo dat het hebben van een geplaatste status een voordeel oplevert, waardoor het model niet meer interessant is voor deze spelers?

Figuur 1 - Spreidingsdiagram Punten vs Nummer



De verticale lijn representeert een grenslijn, waarbij de punten links ervan 'geplaatste spelers' zijn en rechts ervan 'ongeplaatste spelers'. Bij deze grenslijn is tevens de grootste sprong te zien.

In Figuur 1 is een spreidingsdiagram te zien met *punten* op de verticale as en *nummer* op de horizontale as. De verticale grenslijn die ligt op 8,5 scheidt de geplaatste spelers links van de lijn met de ongeplaatste spelers rechts ervan. Op basis van Figuur 1 zou dus geconcludeerd kunnen worden dat *geplaatste_status* een positief effect heeft op *punten* aangezien de grootste sprong tussen de observaties te zien is tussen de laagste geplaatste speler en de hoogste ongeplaatste speler. In dit spreidingsdiagram wordt alleen niet gecontroleerd voor andere variabelen die mogelijk effect hebben op *punten*.

Tabel 1 - Meervoudige Lineaire Regressie

PUNTEN	**COËFFICIËNT**
GEPLAATSTE_STATUS	26,476***
	(6,410)
RANKING	-0,288**
	(0,135)
CONSTANTE	44,272***
	(8,889)
OBSERVATIES	526

*Geplaatst is in dit geval de treatment-variabele. De standaardfouten staan weergegeven tussen de haakjes. Daarnaast, * betekent p ≤ 0,10, ** betekent p ≤ 0,05 en *** betekent p ≤ 0,01.*

Het feit dat in Tabel 1 wordt gecontroleerd voor de variabele *ranking* maakt de coëfficiënt van de treatment variabele *geplaatste_status* een stuk betrouwbaarder. Het toevoegen van de ranking variabele zorgt er namelijk voor dat het effect van de kwaliteit van een speler wordt meegenomen in het model. Wanneer dit niet wordt toegevoegd zal de kwaliteit van een speler verwerkt worden in de coëfficiënt van *geplaatste_status*. Dit levert een scheef beeld op van het daadwerkelijke effect aangezien deze spelers die status verdient hebben door goede prestaties in het verleden, wat indirect laat zien dat deze spelers over hoge kwaliteiten beschikken. De conclusie die getrokken kan worden uit Tabel 1 is echter wel hetzelfde als die uit Figuur 1 namelijk dat *geplaatste_status* een significant positief effect heeft op *punten*. Wanneer een speler geplaatst is, verdient hij gemiddeld 26,476 punten meer dan wanneer hij ongeplaatst is.

In Figuur 1 was een sprong in de data te zien bij de switch tussen geplaatste en ongeplaatste spelers en ook uit de regressie van Tabel 1 werd een significant positief treatment effect gevonden van *geplaatste_status*, terwijl er voor de belangrijke variabele *ranking* is gecontroleerd. Het kan echter ook zo zijn dat variabele die niet te observeren zijn een effect hebben op *punten* zoals bijvoorbeeld motivatie. Om ook voor variabelen te kunnen controleren die niet te observeren zijn, is een Sharp RDD uitgevoerd in Tabel 2. De grens is gesteld op *nummer* is 8,5 aangezien iedere speler lager dan 8,5 een geplaatste status heeft.

Tabel 2 - Sharp RDD-analyse

PUNTEN	COËFFICIËNT
GECENTREERD_NUMMER	1,196
	(3,877)
GEPLAATSTE_STATUS	24,437**
	(11,666)
GECENTREERD_NUMMER * GEPLAATSTE_STATUS	-8,823
	(5,490)
CONSTANTE	24,903***
	(9,712)
OBSERVATIES	526

*Er is een bandbreedte van 4 tot en met 13 gebruikt. De standaardfouten staan weergegeven tussen de haakjes. Daarnaast, * betekent p ≤ 0,10, ** betekent p ≤ 0,05 en *** betekent p ≤ 0,01.*

Om de betrouwbaarheid van de RDD-analyse te bewijzen zijn er van tevoren een aantal belangrijke acties uitgevoerd. Zo liet Figuur 1 al zien dat er een duidelijke sprong in de data is te zien bij de grenslijn waardoor het in ieder geval nuttig is om de RDD-analyse uit te voeren. Als hier geen sprong te zien was, was het namelijk zeer onwaarschijnlijk geweest dat er een daadwerkelijk effect aanwezig is van *geplaatste_status*. Ook wordt er door middel van de interactieterm tussen *nummer* en *geplaatste_status* gezorgd dat het mogelijk is twee verschillende polynomen aan beide kanten van de grenslijn te hebben. Als laatste is er ook gecheckt of er geen manipulatie heeft plaatsgevonden rond de grenslijn, waarbij er bijvoorbeeld opvallend veel observaties zijn die net wel een geplaatste status hebben en opvallend weinig die net niet geplaatst zijn. Dit is in dit geval ook eigenlijk niet mogelijk, omdat er altijd precies acht geplaatste spelers moeten zijn in de geobserveerde toernooien, maar voor de compleetheid wordt deze gelijke verdeling rond de grenslijn alsnog bevestigd in Figuur 2.

Op basis van de resultaten uit Tabel 2 kan gesteld worden dat het effect van *geplaatste_status* op *punten* volgens de RDD-analyse 24,437 punten is. De coëfficiënt is statistisch significant, waardoor er op basis hiervan een betrouwbare conclusie getrokken kan worden. De RDD analyse bevestigt dus eigenlijk de resultaten uit Figuur 1 en Tabel 1 waar ook al een positief effect te zien was tussen *geplaatste_status* en *punten*. Geplaatste spelers hebben dus nog voor aanvang van het toernooi een behoorlijk voordeel wat betreft kans op punten, terwijl ze op basis van resultaten uit het verleden ook nog over hogere kwaliteit beschikken. De combinatie van deze twee aspecten bevestigt voor mij dat geplaatste spelers minder geïnteresseerd zullen zijn in een model dat het toernooi probeert te voorspellen waar de kans op punten het grootst is. Hun kans op punten is namelijk al aanzienlijk hoger

dan voor ongeplaatste spelers dus zullen drijfveren als bijvoorbeeld prijzengeld en locatie belangrijker zijn voor hen dan kans op punten bij het maken van een toernooikeuze.

In het tweede gedeelte van de analyse ben ik op zoek naar het meest nauwkeurige model om het aantal punten voor ongeplaatste spelers te voorspellen. Eerst heb ik meerdere meervoudige regressies uitgevoerd met elke keer *gemiddelde_punten* als afhankelijke variabele, maar wel constant verschillende combinaties van onafhankelijke variabelen. Deze verschillende combinaties staan in Tabel 3 weergegeven als vijf verschillende modellen. Elk model heeft 381 observaties uit de jaren 2012-2022 van 52 verschillende toernooien.

De meest geprefereerde combinatie van variabelen is de samenstelling in Model 4. De Adjusted R-squared is in dit Model het hoogst wat essentieel is wanneer het doel is de afhankelijke variabele zo goed mogelijk te voorspellen. In dit geval geeft het aan dat Model 4 de variantie in *gemiddelde_punten* in 84,7% van de gevallen kan verklaren. Model 3 en 5 scoren echter ook hoog wat betreft de Adjusted R-squared en het is dus betwistbaar of dit minimale verschil de doorslag moet geven. De reden dat ik alsnog Model 4 over Model 3 prefereer, is dat het enige verschil in de variabele *hoogste_ranking* zit die in Model 4 wordt toegevoegd. Aangezien *hoogste_ranking* een significant positief effect heeft, is er geen reden om deze variabele uit het model te laten, zeker niet in combinatie met de hogere Adjusted R-squared. Bij Model 5 werkt het juist andersom aangezien hier *laagste_ranking* en *gemiddelde_ranking* worden toegevoegd ten opzichte van Model 4, terwijl ze beide geen significant effect hebben. Daarnaast is voor alle variabelen in het model getest op multicollineariteit door middel van een variantie-inflatiefactor (VIF) en daaruit bleek dat *laagste_ranking* en *gemiddelde_ranking* een sterk lineair verband hebben gezien hun VIF-scores van respectievelijk 34,72 en 42,70. Dit kan problematisch zijn voor het schatten van de coëfficiënten, omdat ze elkaar grotendeels voorspellen en er dus geen extra variantie kan worden verklaard. De VIF-scores van *laagste_ranking*, *gemiddelde_ranking* en alle overige variabelen kunnen worden teruggevonden in Tabel 4 van de Appendix.

Tabel 3 - Meervoudige Lineaire Regressie

GEMIDDELDE_PUNTEN	(1)	(2)	(3)	(4)	(5)
ATP500	21,144***	21,307***	21,672***	23,900***	22,979***
	(1,589)	(1,560)	(1,475)	(1,748)	(2,107)
ATP500_GELIJKTIJDIG	-1,667	-1,033	0,183	0,457	0,897
	(1,496)	(1,494)	(1,597)	(1,536)	(1,637)
ATP250_GELIJKTIJDIG	0,548	1,190	1,871*	2,165**	2,315**
	(1,005)	(1,031)	(1,079)	(1,043)	(1,073)
ONDERGROND					
HARDCOURT		-1,754	-3,304*	-2,996*	-3,363*
		(1,442)	(1,691)	(1,627)	(1,729)
GRAS		2,199	2,595	3,136	3,015
		(2,119)	(1,947)	(1,883)	(1,918)
CONTINENT					
NOORD-AMERIKA			-10,033***	-10,615***	-9,256***
			(2,647)	(2,552)	(3,013)
ZUID-AMERIKA			-7,977**	-7,677**	-7,053*
			(3,767)	(3,614)	(3,741)
EUROPA			-7,711***	-6,822***	-6,144**
			(2,592)	(2,519)	(2,683)
AZIË			-5,638**	-5,031*	-4,485
			(2,763)	(2,664)	(2,778)
HOOGSTE_RANKING				0,287**	0,205
				(0,133)	(0,220)
LAAGSTE_RANKING					-0,163
					(0,191)
GEMIDDELDE_RANKING					0,225
					(0,307)
CONSTANTE	26,512***	26,206***	32,914***	28,136***	29,432***
	(1,660)	(1,884)	(2,628)	(3,348)	(3,910)
OBSERVATIES	381	381	381	381	381
ADJ. R-SQUARED	0,782	0,792	0,833	0,847	0,842

*Bij de categorische variabele 'ondergrond' is 'gravel' de referentiecategorie en bij de categorische variabele 'continent' is 'Oceanië' de referentiecategorie. De standaardfouten staan weergegeven tussen de haakjes. Daarnaast, * betekent p ≤ 0,10, ** betekent p ≤ 0,05 en *** betekent p ≤ 0,01.*

In het vervolg zal dus gefocust worden op Model 4 en hierin heeft de variabele *ATP500* een significant positief effect op *gemiddelde_punten*. Vergeleken met de andere variabelen heeft het bovendien een erg sterk effect van 23,900 en om die reden en de aanhoudende significantie binnen alle modellen lijkt dit ook de variabele die het meest essentieel is in het voorspellen van *gemiddelde_punten*. De variabelen *ATP500_gelijktijdig* en *ATP250_gelijktijdig* laten zien wat het effect is als het aanbod van markten of in dit geval toernooien toeneemt. Over het effect van een stijging in het aantal ATP500 toernooien kan geen betrouwbare conclusie getrokken worden aangezien de coëfficiënt niet significant is. Een stijging in ATP250 toernooien in dezelfde week laat wel een significant positief effect zien van 2,165, waardoor geconcludeerd kan worden dat ongeplaatste spelers alleen profiteren wat betreft punten wanneer het aanbod van gelijktijdige kleinere toernooien stijgt. Over de ondergrond kan weinig zinnigs geconcludeerd worden aangezien zowel *hardcourt* als *gras* niet significant verschillen van de referentiecategorie *gravel*. Het continent daarentegen laat wel significante resultaten zien. Bij toernooien in Noord-Amerika worden namelijk gemiddeld 10,615 punten minder gescoord dan bij toernooien in Oceanië. Daarnaast worden ook in Zuid-Amerika en Europa respectievelijk gemiddeld 7,677 en 6,822 punten minder gescoord, waardoor Oceanië het meest gunstige continent is om aan een toernooideel te nemen voor ongeplaatste spelers. De mate van concurrentie wordt in het model weergegeven aan de hand van *hoogste_ranking* en dit heeft weliswaar een klein, maar wel significant positief effect. Dit wil zeggen dat wanneer de ranking van de hoogst geplaatste speler stijgt, de *gemiddelde_punten* ook stijgen. Dit is in lijn met de verwachting aangezien een hoger absoluut getal qua ranking gecorreleerd is met een lager kwaliteitsniveau van een speler.

De volgende stap in de analyse is checken hoe goed Model 4 de *gemiddelde_punten* voorspelt wanneer we het per toernooi bekijken. De 52 toernooien zijn hierbij opgesplitst in 13 ATP500 toernooien en 39 ATP250 toernooien aangezien we in Tabel 3 hebben gezien dat ze behoorlijk van elkaar verschillen wat betreft *gemiddelde_punten*. Om die reden is het niet logisch om ze met elkaar te vergelijken dus worden beide categorieën in het vervolg apart geanalyseerd.

Tabel 5 - Beschrijvende Statistieken

	ATP500	**ATP250**
GEMIDDELDE GEMIDDELDE_PUNTEN	47,05	26,77
GEMIDDELD_VERSCHIL	3,17	2,34
GROOTSTE_VERSCHIL	9,07	8,50

Het gaat hier over het verschil tussen de daadwerkelijke waarde en de voorspelde waarde. Daarnaast zijn alle getallen weergegeven in deze tabel absoluut.

In Tabel 5 wordt eigenlijk bevestigd wat de hoge waarde van de Adjusted R-squared al liet zien in Tabel 3. De getallen behorende bij *gemiddeld_verschil* en *grootste_verschil* weergeven de verschillen tussen de daadwerkelijke waarde en de voorspelde waarde. Een gemiddeld verschil van 3,17 op een variabele die gemiddeld 47,05 als waarde heeft en een gemiddeld verschil van 2,34 op een variabele die gemiddeld 26,77 als waarde heeft zijn namelijk behoorlijk accurate voorspellingen. Daarnaast is er te zien dat de grootste verschillen ook geen enorme uitschieters zijn, waardoor we kunnen stellen dat het model nagenoeg nooit een voorspelling zal maken die enorm zal afwijken van de daadwerkelijke waarde.

Nu is aangetoond dat het model in staat is *gemiddelde_punten* redelijk nauwkeurig te voorspellen, kan het gebruikt worden om toernooien individueel te analyseren en aan de hand van deze resultaten ongeplaatste spelers te adviseren. Er kan nu bijvoorbeeld geconcludeerd worden dat op basis van het model de drie ATP500 toernooien waar een ongeplaatste speler het best aan kan deelnemen zijn: Rio de Janeiro, Hamburg en Londen. Bij deze toernooien verdienen ongeplaatste spelers naar verwachting namelijk 52,11, 50,50 en 49,93 punten. Bij de ATP250 toernooien zijn Adelaide1, Adelaide2 en Sydney de meest efficiënte toernooien om aan deel te nemen voor een ongeplaatste speler die een puur rationele keuze maakt wat betreft kans op punten voor de wereldranglijst. Adelaide1 levert naar verwachting namelijk 32,91 punten op, Adelaide2 32,76 en Sydney 31,22. Alle daadwerkelijke resultaten uit het verleden, voorspellingen en het verschil per toernooi is terug te vinden in Tabel 6 en 7 van de Appendix.

Misschien nog wel interessanter is het feit dat het model ook ingezet kan worden om toernooien te voorspellen die nieuw op de kalender komen. In 2021 werd bijvoorbeeld het toernooi van Mallorca voor het eerst gespeeld en het zou zomaar kunnen dat ook in 2023 weer een nieuw toernooi georganiseerd gaat worden. Hier zijn dan geen statistieken uit het verleden over bekend, maar aan de hand van het model kan dan toch een inschatting gemaakt worden of het voor een speler efficiënt is aan het nieuwe toernooi deel te nemen.

Conclusie + Discussie

In dit paper is geprobeerd een model te creëren waarmee accurate voorspellingen gedaan kunnen worden over het gemiddeld aantal punten per toernooi dat gehaald wordt door ongeplaatste spelers. Om dit model op te stellen is getest voor meerdere variabelen, waarvan op basis van literatuur een effect verwacht werd. Op deze manier kon dan antwoord gegeven worden op de vraag: Welke toernooi karakteristieken hebben invloed op het aantal behaalde punten van ongeplaatste spelers en bij welke toernooien ligt dit aantal punten het hoogst?

Alle data die gebruikt is in dit onderzoek, is gebaseerd op officiële statistieken van spelers en toernooien die door de jaren heen niet veranderen of anders geïnterpreteerd kunnen worden. Dit betekent dat bij een herhaling van dit onderzoek nagenoeg dezelfde resultaten gevonden zullen worden en dat de betrouwbaarheid dus hoog ligt. Verder is ook de interne validiteit hoog, aangezien in het voorspellingsmodel is gecontroleerd voor multicollineariteit en met welke combinatie van variabelen de hoogste adjusted R-squared bereikt kon worden. Daarnaast is bij het onderzoeken van het causale effect van *geplaatste_status* met een RDD-analyse een zeer geschikte onderzoeksmethode gekozen, omdat de toewijzing van de treatment volledig bepaald werd door de variabele *nummer*. Hierdoor vond nagenoeg perfecte randomisatie plaats rond de grenslijn, wat wederom voor een hoge interne validiteit zorgt. De externe validiteit van het onderzoek is goed wat betreft de diversiteit in spelers en toernooien. Enkel het toernooi van Marrakesh is niet meegenomen in het model, terwijl voor de rest alle relevante toernooien en spelers uit de periode 2012-2022 zijn opgenomen. Daarentegen laat het iets wat lage aantal observaties de externe validiteit wel wat dalen.

Eerst is er getest of *geplaatste_status* een significant positief effect heeft op *punten*. Zowel de grafische weergave als de regressie waarin werd gecontroleerd voor *ranking* deden vermoeden dat er een significant effect aanwezig was. Uiteindelijk werd dit bevestigd in de RDD-analyse, waardoor geconcludeerd kan worden dat geplaatste spelers voor aanvang van het toernooi al een voordeel hebben ten opzichte van ongeplaatste spelers. De combinatie van het feit dat geplaatste spelers over het algemeen over hogere kwaliteiten beschikken en het voordeel hebben van een geplaatste status, doet vermoeden dat kans op punten voor hen minder van belang is bij het maken van een toernooikeuze. Voor ongeplaatste spelers is dit veel belangrijker, want voor hen is elk punt belangrijk om te stijgen op de wereldranglijst om zo in de toekomst wellicht zelf een geplaatste status te bemachtigen. Het feit dat het model in dit onderzoek zich dus alleen focust op ongeplaatste spelers is dus een logische aanname gebleken.

Van de uiteindelijke variabelen die zijn toegevoegd aan het model zijn er een aantal resultaten die het meest op vielen. Voorspeld werd bijvoorbeeld dat het voor ongeplaatste spelers een betere keuze zou

kunnen zijn om aan een kleiner ATP250 toernooi deel te nemen dan aan het grotere ATP500 toernooi, omdat de concurrentie daar minder sterk is. Uit de resultaten bleek dit echter niet het geval aangezien *ATP500* een sterk positief effect had op *gemiddelde_punten*. Het feit dat bij een ATP500 toernooi meer punten te verdienen valt, woog in dit geval dus zwaarder dan het minder ver komen in een toernooi door sterkere concurrentie. Hieruit kan dus ook geconcludeerd worden dat men niet altijd een afweging hoeft te maken tussen het kenniseffect en het concurrentie-effect. In dit geval zijn namelijk de grotere toernooien waar de concurrentie het sterkst is, ook de meest efficiënte toernooien om aan deel te nemen. Tegelijkertijd wordt bij deze ATP500 toernooien ook het meest geprofiteerd van het kenniseffect, omdat spelers meer kunnen leren van de sterkere concurrentie. Wanneer de beloningen in de ene markt dus zodanig beter zijn dan in de andere markt, hoeven spelers geen rekening te houden met het concurrentie-effect en kunnen ze nog altijd profiteren van het kenniseffect.

Dat meerdere toernooien tegelijkertijd het aantal punten voor ongeplaatste spelers zou verhogen is een verwachting die deels wel waar bleek te zijn. *ATP500_gelijktijdig* en *ATP250_gelijktijdig* hadden beide een positief effect, maar alleen *ATP250_gelijktijdig* was ook significant. Ongeplaatste spelers behalen gemiddeld dus wel meer punten als het aanbod van kleinere markten stijgt, maar een stijging in het aanbod van grotere markten heeft geen invloed.

Verder werd concurrentie gemeten in de vorm van de ranking van de geplaatste spelers. Verwacht werd dat sterkere concurrentie het aantal punten van ongeplaatste spelers zou doen dalen. *Laagste_ranking* en *gemiddelde_ranking* lieten geen significant effect zien, maar *hoogste_ranking* wel. Gezien het positieve effect hiervan betekent het dat wanneer een topspeler deelneemt aan een toernooi, de kans op punten voor ongeplaatste spelers minder wordt. Niet zo zeer het gemiddelde niveau van de geplaatste spelers is dus van invloed, maar wel de uitschieter naar boven. Dit bevestigt eigenlijk de dominantie in de tenniswereld van mannen als Novak Djokovic, Rafael Nadal en Roger Federer. Ook wordt hieruit duidelijk dat spelers om efficiëntie te bereiken op het gebied van punten, het concurrentie-effect wel zwaarder moeten laten wegen dan het kenniseffect als er een toernooikeuze gemaakt moet worden tussen twee toernooien van dezelfde grootte. Eerder zagen we dat de beloningen bij een ATP500 toernooi zodanig hoger zijn dan bij een ATP250 toernooi dat het concurrentie-effect er niet meer toe deed. Wanneer de keuze echter gaat tussen twee toernooien met gelijke beloningen doet het concurrentie-effect er dus wel toe.

Als laatste had het *continent* waar een toernooi op georganiseerd wordt wel invloed op *gemiddelde_punten.* Het feit dat voor elke speler de omstandigheden hetzelfde zijn, zorgde er dus niet voor dat *continent* geen invloed had. De resultaten lieten zien dat Oceanië het meest gunstig is om aan een toernooi deel te nemen voor ongeplaatste spelers. De reden hiervoor wordt in dit onderzoek niet

duidelijk, maar mogelijk heeft het te maken met de extreem warme weersomstandigheden in Oceanië of dat de meeste toernooien er gespeeld worden in het begin van het jaar. Spelers zijn in het begin van het jaar net terug van vakantie en wellicht zorgt dit ervoor dat spelers nog niet direct op hun beste niveau zijn, waardoor de verschillen tussen spelers kleiner zijn.

De uiteindelijke voorspellingen van het model lieten zien dat het over het algemeen eigenlijk altijd efficiënter is om aan een ATP500 toernooi deel te nemen dan aan een ATP250 toernooi. Er zijn echter ook veel weken waarbij alleen maar ATP250 toernooien worden georganiseerd dus is het belangrijk in beide categorieën de meest efficiënte toernooien te achterhalen. De resultaten laten zien dat op ATP500 niveau Rio de Janeiro, Hamburg en Londen de toernooien zijn waar je als ongeplaatste speler sowieso aan moet deelnemen en dat Adelaide1, Adelaide2 en Sydney de meest efficiënte ATP250 toernooien zijn.

Aanbevelingen

Bij het opstellen van het model in dit paper om het gemiddeld aantal punten voor ongeplaatste spelers per toernooi te voorspellen is data gebruikt uit de periode 2012-2022. Dit leverde 381 observaties op, een aantal dat wat aan de lage kant is om een extreem valide model mee te ontwikkelen. Om het aantal observaties en dus de externe validiteit te verhogen, zou de periode waaruit data is verzameld uitgebreid kunnen worden. Wanneer er echter ver terug in de tijd wordt gegaan wat betreft data verzamelen, zal de data die verzameld wordt van een compleet nieuwe generatie zijn. De professionele carrière van een tennisser duurt gemiddeld tussen de 10 en 15 jaar dus wanneer data verzameld wordt van meer dan 15 jaar geleden, zullen er nauwelijks tot geen spelers meer zijn die actief waren gedurende de gehele periode. Om tot uitbreiding van de dataset over te gaan, zou het daarom eerst belangrijk zijn om te onderzoeken of er geen grote verschillen te ontdekken zijn tussen verschillende tennisgeneraties. Als generaties significant verschillen van elkaar, heeft er schijnbaar een ontwikkeling plaatsgevonden binnen de tenniswereld waardoor het heden niet meer vergelijkbaar is met het verleden. Mocht dit het geval zijn, zal het uitbreiden van de dataset geen nut hebben om de validiteit te verhogen.

Bij de RDD-analyse heeft meer observaties verzamelen sowieso wel nut. De 315 observaties die gebruikt zijn in de huidige resultaten zijn alleen de observaties van de meest recente editie van elk toernooi. Wanneer net zoals bij het model data uit 2012-2022 gebruikt zou worden, ontstaat er een veel grotere dataset en stijgt de externe validiteit. Daarnaast ontstaat er niet direct het probleem die ik zojuist bij het model heb aangekaart wat betreft het moeten vergelijken van verschillende generaties.

Een andere tekortkoming van het model is het feit dat spelers van te voren nooit volledige zekerheid hebben of ze een geplaatste status of ongeplaatste status zullen krijgen bij het toernooi. Aangezien de acht hoogst genoteerde spelers een geplaatste status ontvangen, betekent dit dat spelers dus afhankelijk zijn van het wel of niet deelnemen van andere spelers. In de meeste gevallen is het wel goed in te schatten voor een speler of hij wel of niet geplaatst zal zijn, maar voor spelers met een ranking rond de 50 kan dit lastig zijn. Bij de ATP250 toernooien is het namelijk zo dat gemiddeld iedereen met een ranking van 51 of lager geplaatst is, zie Tabel 8 in de Appendix. In principe werd uit het model geconcludeerd dat een ATP500 toernooi gemiddeld gezien altijd voordeliger is om aan deel te nemen dan een ATP250 toernooi als je een ongeplaatste speler bent. Het model houdt echter geen rekening met de spelers die dus een zodanige ranking hebben waarbij ze wellicht bij een ATP500 toernooi ongeplaatst zijn, maar bij een ATP250 geplaatst. Voor deze spelers zou het dus interessant

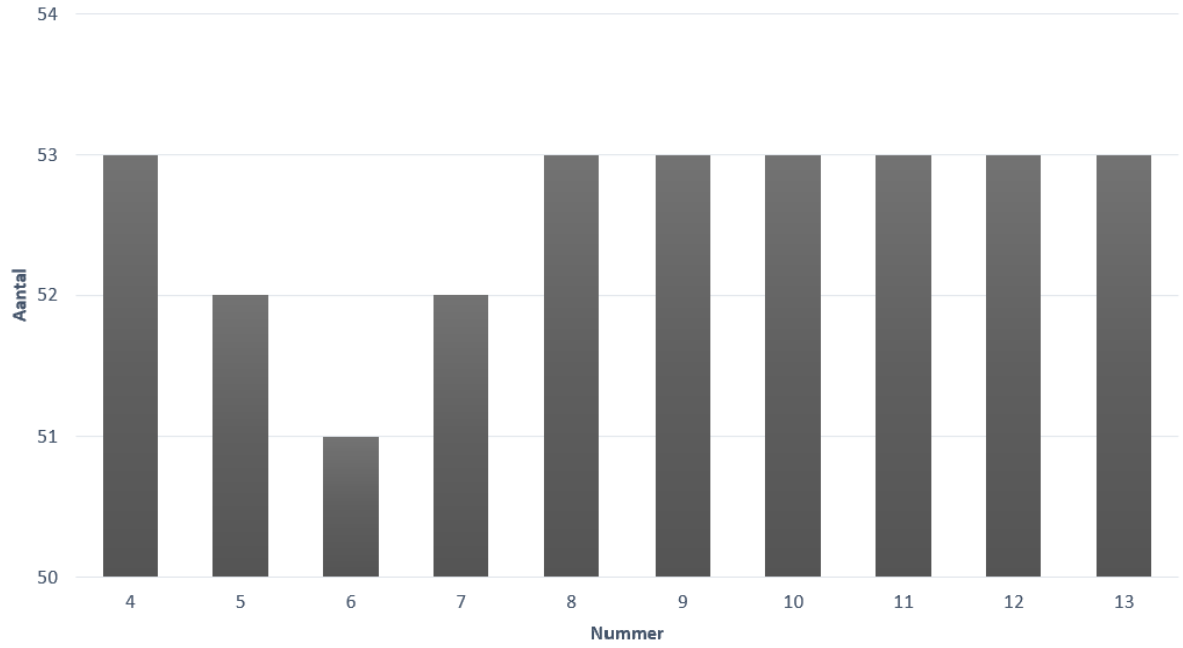
zijn om het gemiddeld aantal punten van geplaatste spelers bij een ATP250 toernooi toe te voegen aan het model, zodat ze ook in die situatie een goede afweging kunnen maken.

Verder is het door middel van het model nu mogelijk voor spelers om constant de toernooien te vergelijken die in dezelfde week gespeeld worden, zodat ze een rationele keuze kunnen maken wat betreft punten. Het model heeft deze voorspellingen echter gemaakt op basis van gemiddeldes en houdt dus geen rekening met persoonlijke voorkeuren van spelers. Eerder in dit paper is al besproken dat een speler als Casper Ruud een sterke voorkeur heeft voor gravel over gras. Dit betekent waarschijnlijk dat wanneer de afweging gemaakt moet worden tussen een gravel of een gras toernooi het voor hem persoonlijk nog altijd efficiënter is om aan het graveltoernooi deel te nemen, ook al zou het model aangeven dat het grastoernooi de efficiëntere optie is. Ter verdediging van de kwaliteit van het model, is het wel zo dat de ATP kalender is opgebouwd uit periodes verdeeld naar het type ondergrond. Zo worden bijvoorbeeld alle grastoernooien gespeeld in juni en juli. Hierdoor zijn er weinig weken waarin toernooien op verschillende ondergronden worden georganiseerd en zal in de meeste weken de persoonlijke voorkeur van een speler wat betreft ondergrond geen factor zijn waar rekening mee gehouden moet worden.

Als laatste zou het voor een vervolgstudie interessant kunnen zijn om meer onderzoek te doen naar de reden van het significante effect van *continent*. Voorafgaand aan het onderzoek had ik niet verwacht dat *continent* van invloed zou zijn, maar uit de resultaten bleek dus dat Oceanië het meest efficiënte werelddeel is om aan een toernooi deel te nemen. Over de reden hiervoor kan in dit onderzoek alleen gespeculeerd worden en daarom zou het van waarde kunnen zijn om in de toekomst aan de hand van een nieuw onderzoek de daadwerkelijke reden te achterhalen.

Appendix

Figuur 2 - Aantal Observaties per Nummer



Tabel 4 - Variantie-inflatiefactor

VARIABELE	VIF
ATP500	2,73
ATP500_GELIJKTIJDIG	2,33
ATP250_GELIJKTIJDIG	2,18
ONDERGROND	
HARDCOURT	2,42
GRAS	1,41
CONTINENT	
NOORD-AMERIKA	3,88
ZUID-AMERIKA	3,99
EUROPA	5,90
AZIË	3,30
HOOGSTE_RANKING	5,37
LAAGSTE_RANKING	34,72
GEMIDDELDE_RANKING	42,70

Tabel 6 – Gemiddelde_Punten vs Voorspelling bij ATP500

TOERNOOI	GEMIDDELDE_PUNTEN	VOORSPELLING	VERSCHIL
ROTTERDAM	50,23	47,96	2,27
RIO DE JANEIRO	53,51	52,11	1,40
DUBAI	41,48	47,23	-5,76
ACAPULCO	44,78	45,13	-0,35
BARCELONA	41,48	48,18	-6,71
LONDEN	49,20	49,93	-0,72
HALLE	58,58	49,51	9,07
HAMBURG	53,41	50,50	2,91
WASHINGTON	39,38	42,67	-3,29
BEIJING	41,72	45,08	-3,36
TOKIO	47,55	45,72	1,82
WENEN	47,29	44,17	3,12
BASEL	43,07	43,47	-0,39

De voorspelling is gemaakt op basis van Model 4 uit Tabel 1.

Tabel 7 – Gemiddelde_Punten vs Voorspelling bij ATP250

TOERNOOI	GEMIDDELDE_PUNTEN	VOORSPELLING	VERSCHIL
ADELAIDE 1	24,40	32,91	-8,50
MELBOURNE	28,64	29,03	-0,39
SYDNEY	32,13	31,22	0,91
ADELAIDE 2	40,75	32,76	7,99
CORDOBA	29,60	28,53	1,08
PUNE	28,83	27,74	1,09
MONTPELLIER	23,32	25,55	-2,23
DALLAS	17,00	22,61	-5,61
BUENOS AIRES	24,74	24,99	-0,24
DOHA	26,03	26,15	-0,12
DELRAY BEACH	29,67	23,36	6,31
MARSEILLE	22,48	25,17	-2,69
SANTIAGO	25,92	27,79	-1,87
HOUSTON	25,15	23,96	1,19
BELGRADO	21,08	25,60	-4,52
MÜNCHEN	25,09	25,69	-0,60
ESTORIL	26,80	26,78	0,02
GENEVA	29,23	25,49	3,74
LYON	28,96	26,07	2,89
STUTTGART	25,12	28,91	-3,80
DEN BOSCH	29,28	30,21	-0,93
MALLORCA	27,00	27,19	-0,19
EASTBOURNE	29,73	30,88	-1,16
NEWPORT	26,43	28,70	-2,27
BASTAD	26,14	27,15	-1,01

GSTAAD	29,16	25,51	3,65
ATLANTA	23,50	23,45	0,05
KITZBÜHEL	32,43	30,42	2,02
UMAG	26,76	30,69	-3,93
LOS CABOS	22,84	19,29	3,55
WINSTON SALEM	19,03	18,96	0,07
METZ	24,40	23,39	1,01
NUR SULTAN	30,63	29,89	0,74
CHENGDU	36,00	27,96	8,04
ZHUHAI	24,00	26,45	-2,45
SOFIA	26,53	26,38	0,15
ANTWERPEN	27,46	26,34	1,12
STOCKHOLM	22,73	24,82	-2,09
MOSKOU	25,17	26,16	-0,99

De voorspelling is gemaakt op basis van Model 4 uit Tabel 1.

Tabel 8 - Beschrijvende Statistieken ATP250

	GEMIDDELDE
HOOGSTE_RANKING	12,26
LAAGSTE_RANKING	51,18
GEMIDDELDE_RANKING	32,31

Referenties

Barbosa, N. (2007). An integrative model of firms' entry decisions. *Applied Economics Quarterly*, *53*(1), 45–67.

Bresnahan, T. F., & Reiss, P. C. (1991). Entry and Competition in Concentrated Markets. *Journal of Political Economy*, *99*(5), 977–1009. https://doi.org/10.1086/261786

Del Corral, J., & Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting*, *26*(3), 551–563. https://doi.org/10.1016/j.ijforecast.2009.12.006

Entry List. (2022). TennisTeen. https://www.tennisteen.it/entry-list.html?page=1

Fernandez, J., Mendez-Villanueva, A., & Pluim, B. M. (2006). Intensity of tennis match play. *British Journal of Sports Medicine*, *40*(5), 387–391. https://doi.org/10.1136/bjsm.2005.023168

How to Obtain Predicted Values and Residuals in Stata. (2020, 21 maart). Statology. https://www.statology.org/predicted-values-residuals-regression-stata/

Johnson, C. D., & McHugh, M. P. (2005). Performance demands of professional male tennis players * Commentary 1 * Commentary 2. *British Journal of Sports Medicine*, *40*(8), 696–699. https://doi.org/10.1136/bjsm.2005.021253

Morton, F. M. S. (1999). Entry Decisions in the Generic Pharmaceutical Industry. *The RAND Journal of Economics*, *30*(3), 421. https://doi.org/10.2307/2556056

Simonsson, F. (2022, 3 maart). *How Does Seeding Work in Tennis?* TennisPredict. https://tennispredict.com/how-does-seeding-work-in-tennis/

STEVE G TENNIS. (2020, 12 mei). *ATP Tennis Ranking Points*. https://www.stevegtennis.com/atp-tennis-ranking-points/

Tournaments | ATP Tour | Tennis. (2022). ATP Tour. https://www.atptour.com/en/tournaments

Ultimate Tennis Statistics - Casper Ruud. (2022). UltimateTennisStatistics. https://www.ultimatetennisstatistics.com/playerProfile?playerId=34553&tab=profile