STATISTIEK EN SCHAATSEN

met Principal Component Analysis naar een eerlijker allroundtoernooi

De Internationale Schaatsunie ISU heeft in juni 2014 besloten dat op Europese allroundschaatstoernooien de 10 kilometer voor mannen en 5 kilometer voor vrouwen worden vervangen door een 1.000 meter. Sven Kramer is woedend en noemt het een 'supersprint' in plaats van een allroundtoernooi. Heeft hij gelijk? Worden sprinters in het nieuwe systeem bevoordeeld? Of hebben stayers in het huidige systeem juist voordeel? Statistiek kan ons helpen om daar een uitspraak over te doen. En om een eerlijker allroundsysteem vorm te geven.

Sven Kramer op de 10.000 meter Olympische Winterspelen 2010 in Vancouver.

Miriam Loois

Voor de ene schaatser lijkt de 500 meter soms al te lang, terwijl de andere pas op gang komt halverwege de 10 kilometer. Gevoelsmatig weten we allemaal dat er verschillende soorten rijders zijn, maar we kunnen dat ook wetenschappelijk onderbouwen. Een wiskundige methode waarmee dit kan is *principal component analysis*. Met principal component analysis kan een dataset met gecorreleerde variabelen getransformeerd worden in een set met ongecorreleerde variabelen, de principale componenten. De eerste principale component verklaart het meest van de variantie in de data, gevolgd door de tweede component, die loodrecht staat op de eerste.

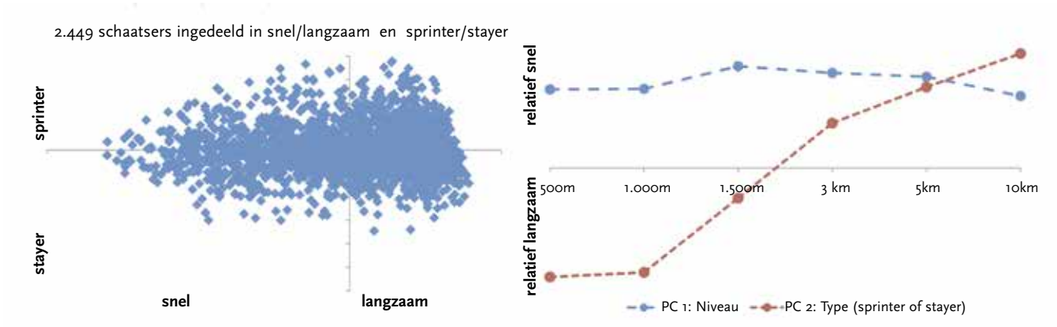
Ik heb deze techniek toegepast op de persoonlijke records van 2.449 mannelijke schaatsers op 6 afstanden (500, 1.000, 1.500, 3.000, 5.000 en 10.000 meter) uit de database van Jakub Majerski. Eerst worden de data genormaliseerd, zodat de gemiddelde tijd per afstand 0 is en de standaarddeviatie 1. Dan wordt deze matrix via *singular value decomposition* geschreven als het product van drie matrices U, Σ en V.

$$\widetilde{P}_{i,d} = \sum_j U_{ij}\Sigma_{jj}V_{dj};$$

$\widetilde{P}_{i,d}$ = Genormaliseerd persoonlijk record van schaatser i op afstand d;
U_{ij} = Interactie van schaatser i met principale component j;
Σ_{jj} = Principale component j;
V_{dj} = Interactie van persoonlijk record op afstand d met principale component j.

De matrix Σ is diagonaal en bevat de principale componenten geordend van groot naar klein. Bij elke principale component hoort een kolom van U die wat zegt over de interactie van de schaatser met de principale component, en een kolom van V die wat zegt over de interactie met het persoonlijke record op een afstand.

De eerste twee principale componenten verklaren samen 90% van de verschillen tussen de tijden van schaatsers. De eerste principale component zegt iets over het algemene niveau van de schaatser en verklaart 66% van de variantie. De ene schaatser is nu eenmaal sneller dan de andere, ongeacht de afstand. De tweede factor zegt iets over het feit of de schaatser een sprinter of een stayer is. Sprinters zijn relatief sneller op de korte afstanden, stayers zijn juist beter op de langere afstanden. Deze principale component verklaart 24% van de variantie. De andere principale componenten verklaren allemaal minder dan 5% van de variantie en zeggen iets over de relatieve snelheid op de middenafstanden (pc 3) en de relatieve snelheid op individuele afstanden (pc 4, 5 en 6). De eerste twee principale componenten zijn grafisch weergegeven in figuur 1. De linker figuur toont een spreidingsgrafiek waarbij de eerste kolom van matrix U is afgezet tegen de tweede kolom van U. De rechter figuur



Figuur 1. (links) Schaatsers indelen met behulp van de eerste twee principale componenten, niveau en type; (rechts) de eerste twee kolommen van matrix V.

	NIVEAU: SNEL (1) OF LANGZAAM (-1)	TYPE: STAYER (1) OF SPRINTER (-1)
Kramer	1,00	0,73
Verwey	1,00	0,03
Blokhuijsen	0,99	0,38
Uytdehaage	0,98	0,59
Verheijen	0,98	0,85
Wennemars	0,99	-0,60
Tuitert	0,99	-0,50

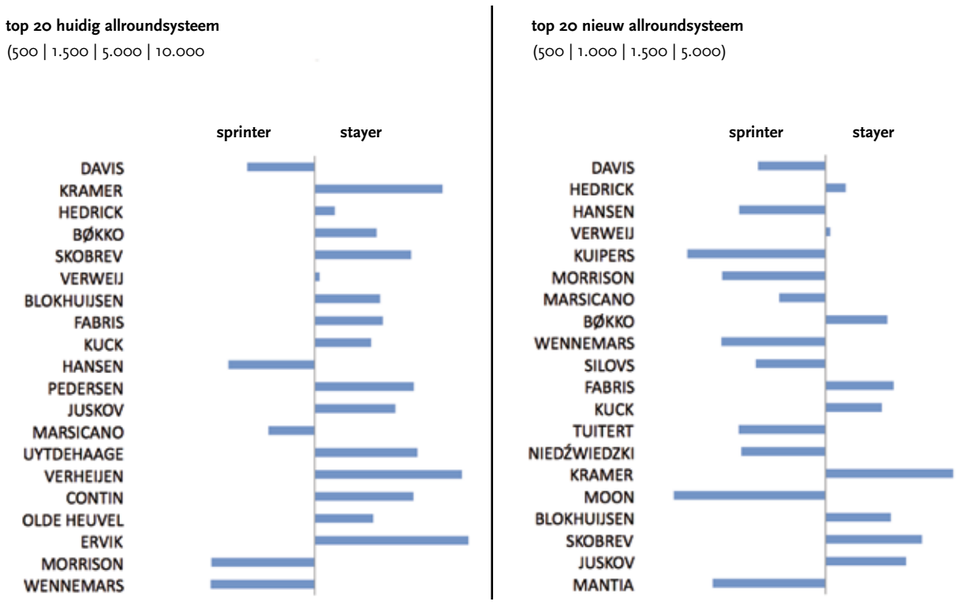
Tabel 1. Nederlandse (oud) topschaatsers en hun niveau en type. De schaatsters zijn geordend van snel naar langzaam en van stayer naar sprinter. Op basis van hun positie in die lijst krijgen ze een niveau en type toegekend tussen -1 en 1.

geeft de eerste twee kolommen van matrix *V* weer.

Deze twee principale componenten kun je op basis van percentielen in de verdeling uitdrukken in een getal tussen -1 en 1. Zie tabel 1 met Nederlandse toppers. Ze hebben allemaal een niveau dat dicht bij 1 ligt. Kramer en Verheijen blijken in de analyse, niet geheel onverwacht, echte stayers, terwijl Tuitert en Wennemars aan de sprintkant van het spectrum zitten. Verweij is een relatief neutrale schaatser.

Is het allroundtoernooi wel eerlijk?

In een allroundtoernooi worden vier afstanden geschaatst. Bij de mannen zijn dat op dit moment de 500, 1.500, 5.000 en 10.000 meter. De vrouwen rijden de 500, 1.500, 3.000 en 5.000 meter. Elke schaatser krijgt punten voor elke afstand. Per afstand wordt de tijd omgezet naar een tijd op de 500 meter, en deze tijden worden bij elkaar opgeteld. Degene met het minste aantal punten wint. Een achterstand van één seconde op de 500 meter moet dus worden gecompenseerd door 10 seconden op de 5 kilometer.



Figuur 2. Top 20 gebaseerd op persoonlijke records in het huidige systeem en in de nieuwe opzet.

Nu de hamvraag: kunnen we zeggen of een bepaald type schaatser in het voordeel is op een allroundtoernooi? Als we in figuur 2 kijken naar de top 20 van de wereld in het huidige systeem, gebaseerd op persoonlijke records (de zogenoemde Adelskalender), dan zien we relatief veel stayers. Davis is de aanvoerder van deze lijst, gevolgd door Kramer. In het nieuwe systeem, waarin de 10 kilometer is vervangen door de 1.000 meter, komen juist veel sprinters bovendrijven. Kramer moet genoegen nemen met een 15e plek. Geen wonder dat hij boos is!

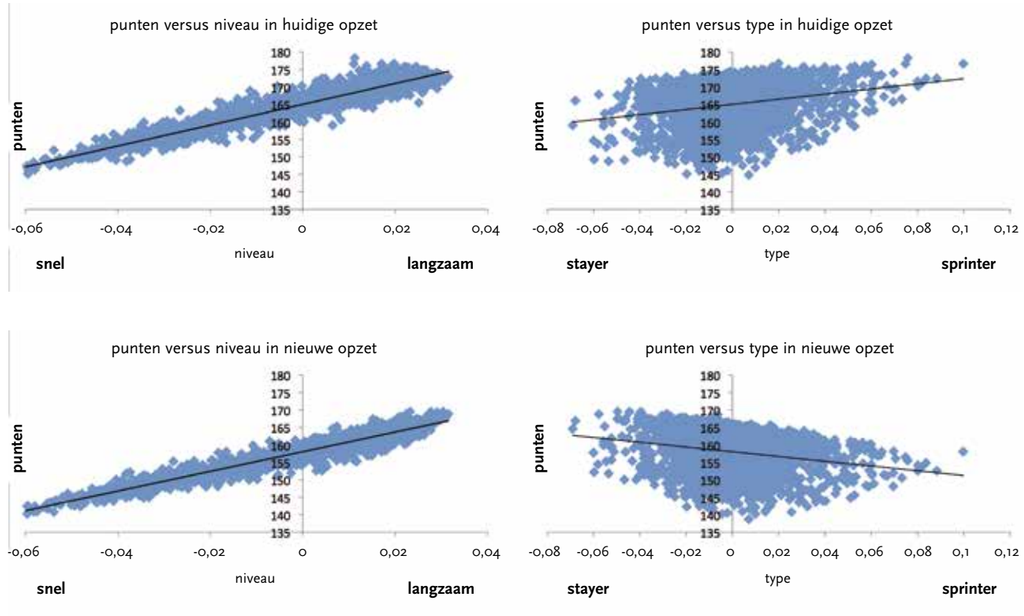
In wiskundige termen: in het huidige allroundtoernooi is de correlatie tussen puntenaantal en het niveau van de schaatser 96% en de correlatie met de principale component 'type' is 24%. Dat betekent dat stayers in het voordeel zijn. Winnen in het nieuwe allroundtoernooi hangt voor 97% samen met het algemene niveau van de schaatser en voor 23% met de vraag of hij een sprinter is. Zowel het nieuwe als het huidige systeem bevoordeelt dus een groep schaatsers. De samenhang tussen puntenaantal en niveau en type is te zien in figuur 3.

Naast het besluit om de langste afstand te vervangen door de 1.000 meter heeft de ISU besloten dat tijdens de Olympische Spelen van 2018 de 500 meter maar één keer zal worden gereden. Nu rijden de schaatsers deze afstand twee keer, een keer startend vanuit de binnenbaan en een keer startend vanuit de buitenbaan. Zowel Kamst e.a. (2010) als Hjort (1994) hebben aangetoond dat er een significant binnenbaan-buitenbaan-verschil is. Ook dit besluit leidt dus niet tot een eerlijker toernooi.

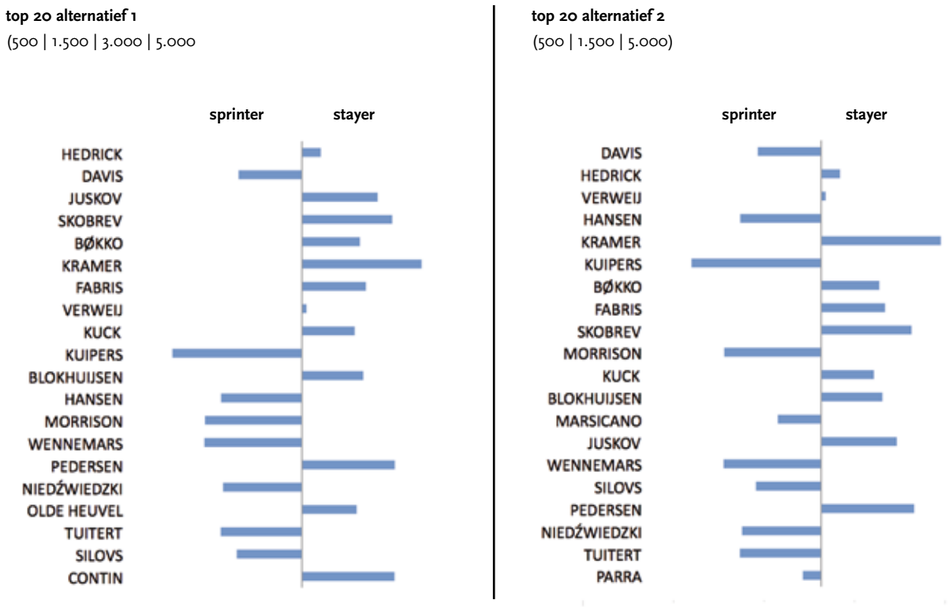
Naar een eerlijker allroundtoernooi?

Ik heb twee alternatieven bekeken. In het eerste alternatief wordt de 10 kilometer niet door de 1.000 meter, maar door de 3 kilometer vervangen. In het tweede alternatief wordt de 10 kilometer helemaal niet vervangen door een andere afstand; hij vervalt simpelweg.

In figuur 4 zien we weer de top 20 op basis van persoonlijke records. Bij het eerste alternatief zien



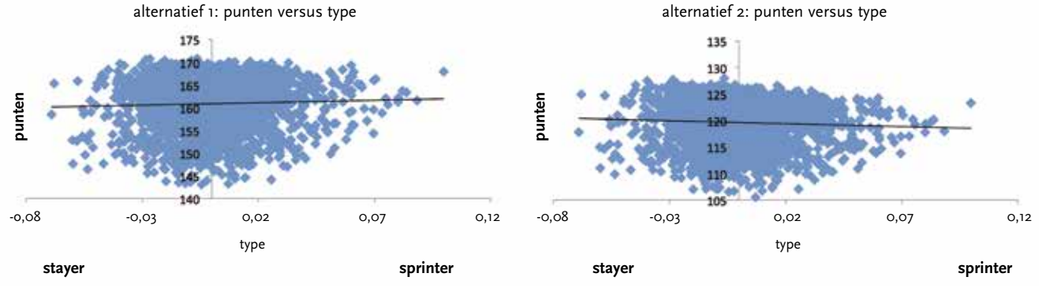
Figuur 3. Relatie tussen punten in de huidige en nieuwe opzet en het niveau en type schaatser. Een negatief 'niveau' hoort bij een relatief snelle schaatser. Een positief 'type' representeert een sprinter, een negatief 'type' een stayer. De grafieken bevatten trendlijnen.



Figuur 4. De top 20 op basis van persoonlijke records voor twee alternatieve systemen.

we al meer sprinters, hoewel de top 10 nog steeds gedomineerd wordt door stayers. Bij het tweede alternatief is de verdeling tussen sprinters en stayers aan de top nog evenwichtiger.

Dat de alternatieven evenwichtiger zijn zien we ook als we naar figuur 5 en tabel 2 met correlaties kijken. Winnen hangt nu voor maar liefst 99% samen met het algemene niveau van de schaatser en de vraag of hij een sprinter of een stayer is heeft veel minder invloed.



Figuur 5. Relatie tussen punten in de twee alternatieven en het niveau en type schaatser.

	HUIDIG SYSTEEM 500–1.500–5.000–10.000	NIEUW SYSTEEM 500–1.000–1.500–5.000	ALTERNATIEF 1 500–1.500–3.000–5.000	ALTERNATIEF 2 500–1.500–5.000
NIVEAU	0,96	0,97	0,99	0,99
TYPE SCHAATSER	0,24	- 0,23	0,04	- 0,05

Tabel 2. Correlatie tussen puntenaantal en niveau en type in de verschillende systemen. Een positieve correlatie betekent dat stayers in het voordeel zijn, bij een negatieve correlatie zijn sprinters in het voordeel.

Conclusies

Terug nu naar de woede van Kramer. Hij had gelijk: de maatregel om de 10 kilometer te vervangen door de 1.000 meter bevoordeelt de sprinters fors. Maar bij de huidige vormgeving van het allroundtoernooi zijn stayers juist erg in het voordeel.

Met behulp van principal component analysis is het mogelijk om een eerlijker allroundtoernooi vorm te geven, waarbij het algemene niveau van de schaatser daadwerkelijk de doorslag geeft. Dit kan door de 10 kilometer door de 3 kilometer te vervangen in plaats van door de 1.000 meter. Een andere mogelijkheid is om de 10 kilometer helemaal niet te vervangen en een toernooi over drie afstanden te rijden.

Moge de beste winnen!

Ook binnen de financiële sector wordt principal component analysis gebruikt, bijvoorbeeld voor het beschrijven van renterisico. De rentecurve met rentes bij verschillende looptijden wordt gebruikt voor het waarderen van verplichtingen van verzekeraars en pensioenfondsen. Bewegingen in deze curve vormen daarom (potentieel) een risico. Als principal component analysis wordt toegepast op de veranderingen in de rente komen er ook twee belangrijke factoren naar boven. De eerste is een verandering in het niveau van de rente; alle rentes stijgen of dalen. Dit is te vergelijken met het niveau van de schaatser. De tweede is een draaiing van de rente; de korte rente stijgt en de lange rente daalt of vice versa. Hier is een parallel te trekken met het type schaatser. Deze factor 'draait' als het ware de tijden op de verschillende afstanden; een snellere tijd op de korte afstanden en een langzamere tijd op de lange afstanden of vice versa.

Literatuur

Jakub Majerski's speedskating database: <http://www.sskating.com/list.php?season=2013/2014&lm=m&dist=Big6>.

Belangrijkste ISU-besluiten op een rij (ISU-congres Dublin 2014): <http://knsb.nl/nieuws/belangrijkste-isu-besluiten-op-een-rij/>.

Kamst, R., Kuper, G. H., & Sierksma, G. (2010). The Olympic 500 meter speed skating; the inner–outer lane difference. *Statistica Neerlandica, 64*(4), 448–459.

Hjort, N. J. (1994), *Should the Olympic Sprint skaters run the 500 meter twice?* Statistical Research Report. Oslo: Institute of Mathematics, University of Oslo.

Miriam Looijs heeft de Master Theoretische Natuurkunde gevolgd aan de Universiteit Utrecht en de Master Actuarial Science and Mathematical Finance aan de Universiteit van Amsterdam. Ze werkt nu als Asset Liability Manager bij pensioenuitvoerder PGGM.
E-mail: <miriamlooijs@gmail.com>