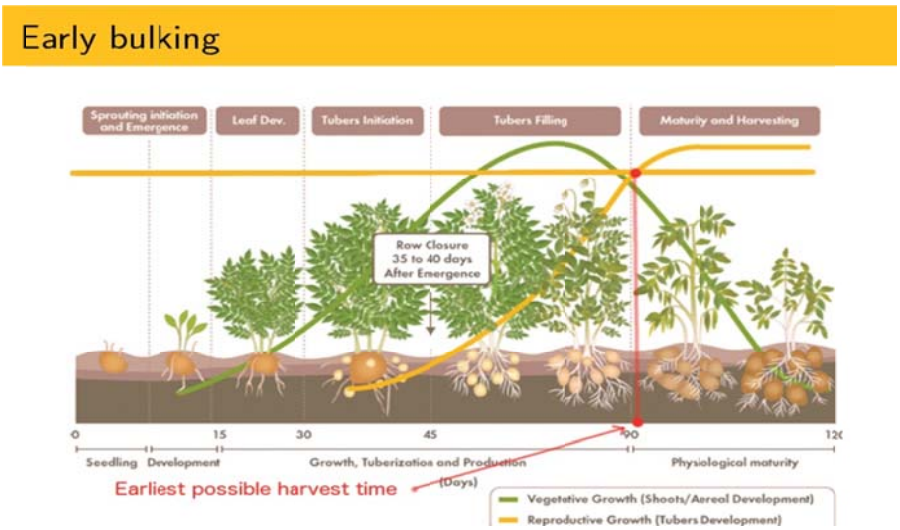
Zoek de snelst uit de kluiten gewassen aardappels

HZPC is een bedrijf dat door kruisen telkens nieuwe aardappelrassen produceert, bijvoorbeeld voor ziekteresistentie. Maar het wil ook de genetisch markers *in kaart brengen van aardappelrassen die snel groeien. De wiskundigen brachten daarvoor statistisch zwaar geschut in stelling. Pieter-Jelte Lindenbergh van HZPC toonde zich onder de indruk: 'Wiskundigen kunnen echt iets toevoegen aan deze business.'*

HZPC produceert door kruisen elk jaar nieuwe variëteiten aardappelen, die over de hele wereld en in allerlei klimaten worden geplant. Een van de belangrijkste kwaliteiten van een variëteit is *early bulking:* het moet zo snel mogelijk na het planten grote aardappelen vormen, zodat er snel geoogst kan worden. Vorig jaar heeft HZPC een experiment gedaan met honderd variëteiten, die elk zijn ingezaaid op twee stukken grond en op vier verschillende tijdstippen geoogst. Bij elke oogst zijn alle aardappelen gemeten en gewogen.
Verder is ooit van een groot aantal variëteiten het hele genoom onderzocht op de aanwezigheid van SNP's. Een SNP (spreek uit 'snip') is een goed identificeerbaar stukje DNA, een *marker* (merkteken) met een vaste locatie in het genoom, waarvan het aardappel-genoom er duizenden bevat. In dit onderzoek zijn er ruim elfduizend geïdentificeerd. SNP's zijn geen genen: ze hebben zelf geen functie in de plant. Maar als een SNP in het DNA vlak naast een gen ligt dat, bijvoorbeeld, zorgt voor snelle groei, dan erft deze snip vrijwel altijd samen over met de eigenschap 'snelle groei'. Ligt een SNP een heel stuk verwijderd van dit gen, dan erven ze vrijwel onafhankelijk van elkaar over. Het is dus wel degelijk nuttig om SNP's, indien mogelijk, te identificeren met bepaalde erfelijke eigenschappen, zelfs als over het gen dat daarvoor zorgt niets bekend is.

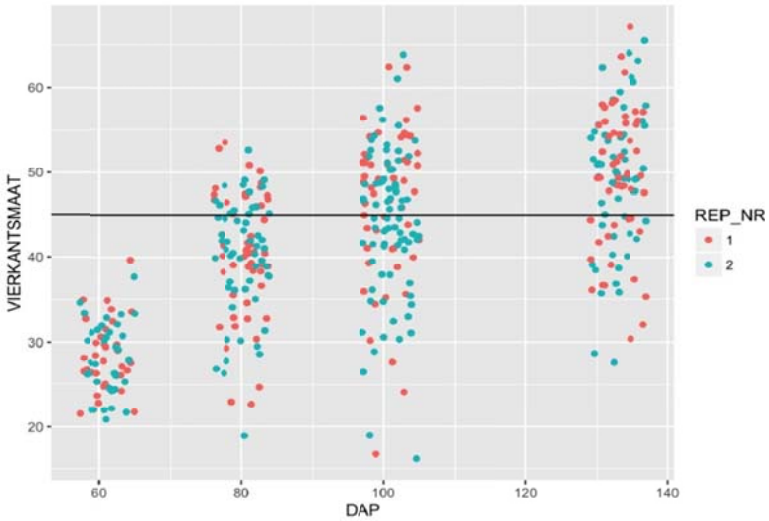
Nu HZPC al deze data beschikbaar heeft, leidt dat tot twee concrete onderzoeksvragen:
– maak een model voor de groei van de aardappelen, waarmee je kunt voorspellen welke variëteiten het snelst rijp zijn om te oogsten.
– Welke SNP's zijn goede indicatoren voor een variëteit die snel oogstrijp is?

In eerder onderzoek, vertelt Lindenbergh, is al diverse malen een link gevonden tussen één SNP en resistentie tegen een bepaalde ziekte. Maar hij verwacht dat de opbrengst aan aardappelen van een variëteit een veel complexere eigenschap is, zodat er links zullen zijn met vele SNP's. Het is wiskundig bepaald niet triviaal, om uit de experimenten die HZPC vorig jaar gedaan heeft, die SNP's voor snelle groei te halen - als het al mogelijk is.



[bijschrift] Algemeen schema voor de groei van een aardappelplant. *Tuber* is de vakterm voor één aardappel.

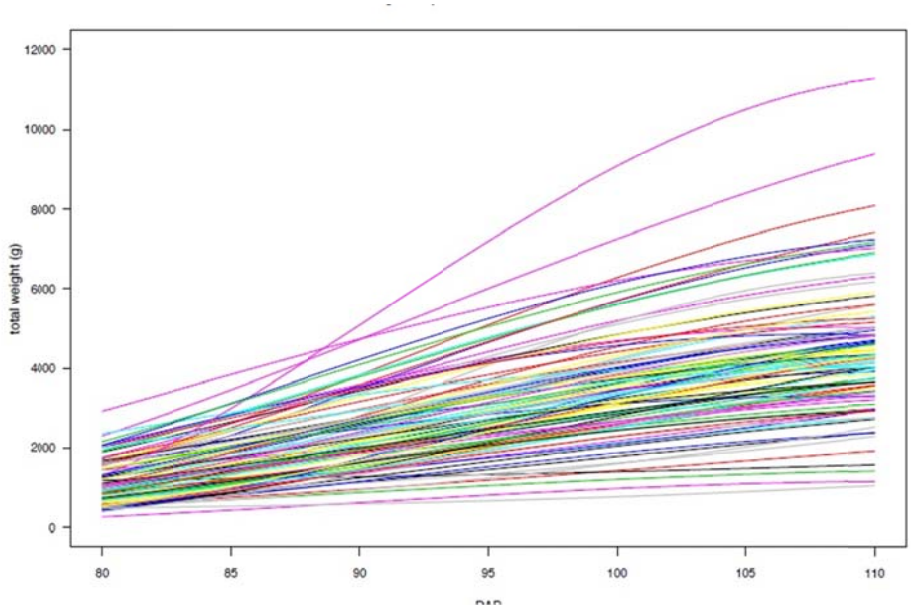
In de groep die deze twee onderzoeksvragen aan gaat pakken, zitten wiskundigen die uitgebreide ervaring hebben met statistische technieken. De eerste opgave - maak een model voor de groei van de diverse variëteiten aardappel - levert daarom weinig problemen op. Uitgangspunt is een experiment dat door HZPC in 2015 is gedaan, waarbij honderd variëteiten aardappels elk in twee stukken grond zijn geplant. Van alle aardappels is op vier tijdstippen een deel geoogst: 60, 80, 100 en 130 dagen na het planten. Van elke geoogste plant is gemeten en gewogen hoe groot de opbrengst was.
Als je die ruwe cijfers in een grafiek zet, ontstaat zelfs met één variëteit al een heel rommelig beeld.



[bijschrift] De grootte van individuele aardappels wisselt sterk. Hier zijn van één variëteit de groottes van alle aardappels aangegeven (verticale as), die op vier verschillende tijden (horizontale as) geoogst werden. Om ze te kunnen onderscheiden, zijn de stippen in de grafiek niet precies op de oogsttijd geplaatst, maar er omheen. Na de derde oogsttijd zit er weinig groei meer in de aardappels. De kleur geeft aan in welk van twee proefveldjes de planten opgroeiden, waaruit blijkt dat de opbrengst in beide proefvelden maar weinig verschilt.

Determinisme en chaos

Om dit chaotische beeld om te zetten in een groeicurve die optimaal de gemiddelde groei van een variëteit weergeeft, neem je aan dat de groei van elke aardappel bestaat uit de som van een deterministisch gedeelte en een chaotisch, door toeval bepaald gedeelte.
Het deterministische deel kun je weergeven door een formule die alleen afhangt van de tijd (het aantal dagen na het planten). Voor het toevallige deel neem je aan dat die variatie voldoet aan een zekere standaard statistische verdeling. Dit model bevat nog drie parameters, als het ware de knoppen waar je aan kunt draaien. De optimale groeicurve vind je, door aan de knoppen te draaien totdat het toevallige deel zo klein mogelijk is.
Dit doe je voor elke variëteit apart, wat dus 100 verschillende groeicurves oplevert.



[Bijschrift] Resultaten van een model dat voor alle 100 variëteiten zo goed mogelijk de groei voorspelt. Verticaal het totaal gewicht van de aardappelen, horizontaal het aantal dagen na het planten.

Voor het bedrijf is van belang, dat ze nu per variëteit voor elke mogelijke oogsttijd een optimale schatting van de groei hebben. Zo kan men betere beslissingen nemen over welke variëteit, gegeven de omstandigheden en locatie, het best geplant kan worden, en wanneer geoogst. Als je vooral een hoge opbrengst wilt, kan tragere groei acceptabel zijn. Lindenbergh: 'Maar in sommige landen, bijvoorbeeld Ethiopië, wordt het op zeker moment gewoon te heet en moet je wel oogsten.' Dan is een variëteit die goed is in *early bulking* het meest gewenst.

Het drama met de SNP's

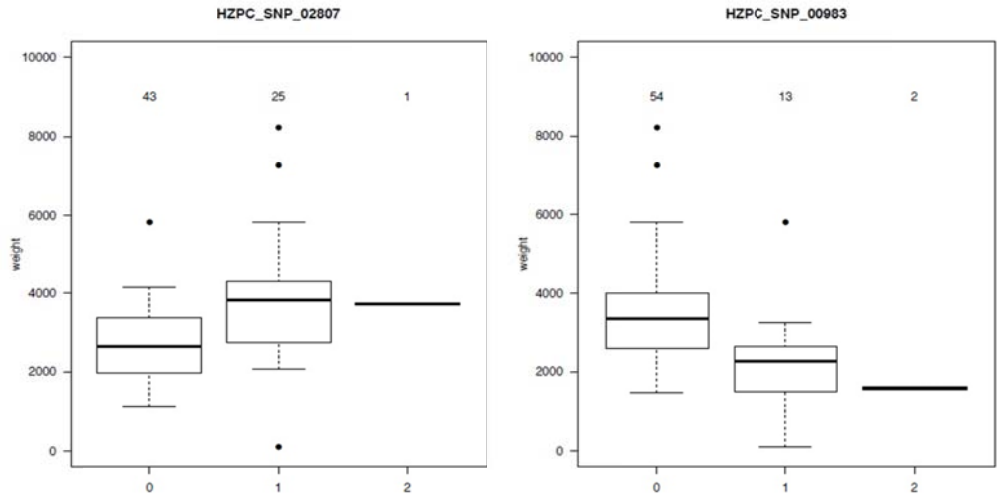
Tot zover waren er geen echte obstakels voor de wiskundigen. Maar bij de tweede onderzoeksvraag – zoek uit welke SNP's relevant zijn voor snelle groei – gingen de zaken

minder van een leien dakje.
Op zich is het een bekend probleem uit de genetica: zoek uit welke genen een functie hebben voor kwaal of eigenschap X, op basis van een aantal genetische profielen. Het lastige is, dat er duizenden genen in aanmerking komen, terwijl men slechts over een veel kleiner aantal DNA-profielen beschikt (want het is duur en tijdrovend om die te maken).
Dat geldt bijvoorbeeld voor DNA-profielen van patïenten met een specifiek type kanker, maar ook hier: er zijn 11.673 SNP's, terwijl men maar van 113 variëteiten een profiel heeft. Erger nog, van maar 69 variëteiten met een bekend profiel zijn ook voldoende nauwkeurige gegevens bekend over de opbrengst.
De techniek om uit zulke 'ijle' datasets relevante verbanden te destilleren is nog lang niet af. Zo kreeg Aad van der Vaart, wiskundige aan de Leidse Universiteit, in 2015 nog een Spinoza-premie van 2,5 miljoen euro om deze tak van statistiek verder te ontwikkelen.

Er bestaan wel diverse methoden om dit probleem aan te pakken, maar daarbij moet de wiskundige soms nog op intuïtie keuzes maken. Op initiatief van Alessandro Di Bucchianico van de Technische Universiteit Eindhoven en Fetsje Bijma van de Vrije Universiteit kiest de groep voor 'elasticnet', een variant van een zogeheten 'lineaire regressie model'. De wiskunde achter zulke modellen is gecompliceerd, maar op zijn simpelst gezegd komt het er op neer, dat je begint met aan te nemen dat alle voorspellers – in dit geval alle 11.673 SNP's – enige invloed op de groei van de aardappel hebben, en dat je de invloed van alle afzonderlijke SNP's simpelweg kunt optellen voor het totale effect. Elasticnet is een procedure om alleen de SNP's met de meeste invloed op de groei uit te selecteren.
De input voor die procedure zijn de 69 variëteiten aardappel waarvan bekend is welke SNP's ze hebben en hoe groot hun opbrengst precies was.
Aardappelen zijn genetisch een beetje vreemd, omdat ze tetraploïde zijn: ze hebben van elk gen vier kopieën (de mens heeft van elk gen twee kopieën, en is diploïde). Van elke SNP kan een variëteit dus 0,1,2,3 of 4 exemplaren hebben. In principe is dat juist gunstig voor de analyse, want als in een variëteit één exemplaar van een SNP gunstig samenhangt met snelle groei, zou een variëteit met drie of vier exemplaren nog sneller moeten groeien. De tetraploïdie geeft in principe dus extra informatie.

In de loop van de week zet de groep het elasticnet-model in de steigers en begint met de eerste tests. Maar ook na het maken van vele lange uren aan de diverse laptops komt er nog geen witte rook uit het lokaal van de groep. Zo wordt het donderdagmiddag vijf uur, en de groep overleeft op koffie en M&M's. 'Het is een drama,' verzucht Bijma desgevraagd, 'geen enkele SNP doet mee aan snelle groei als je streng bent.'
Maar terwijl ze dit zegt, loopt er nog een run met het elasticnet-model, en al een paar minuten later klinkt het: 'Net toen je binnenkwam vonden we er een.'
'Maar één?'
'Nee, dat zou niet goed zijn. We weten al dat er méér zijn.'

Positief of negatief
Tijdens de presentatie, de volgende dag, blijkt dat donderdagmiddag inderdaad het moment van een bescheiden doorbraak was. Volgens het model hangt ongeveer 1 procent van alle SNP's samen met snelle groei, maar dit kan zowel positief als negatief zijn.



[bijschrift snips] Twee voorbeelden van SNP's die samenhangen met snelle groei. Links is de samenhang positief. Dat blijkt uit het feit dat het totale gewicht van de aardappelen per plant gemiddeld hoger is als de SNP wel in het DNA zit (1; rechter balkje), dan wanneer dat niet zo is (0; linker balkje). De SNP kan ook twee keer (of zelfs maximaal vier keer) voorkomen, maar dat is hier maar bij één variëteit het geval (de streep rechts).
De dikke streep geeft het mediane gewicht over de variëteiten aan (een soort gemiddelde) , de box en de dunne streepjes geven een maat voor de variatie rond de mediaan.
In het rechter plaatje is de associatie van de SNP met groei juist nadelig, want het balkje aan de linkerkant ligt hoger dan het balkje rechts in dat plaatje.

Dit resultaat is idealiter nog maar de aftrap voor verder onderzoek. Om echt te weten hoe de SNP's samenhangen met de groei, zou je het model moeten uitbreiden. Ten eerste door van meer dan 69 variëteiten de SNP's te bepalen. Ook moet je niet slechts onderzoeken hoe afzonderlijke SNP's samenhangen met groei, maar ook hoe groepjes SNP's dat eventueel doen. Als dat allemaal betrouwbaar is vastgesteld, kan HZPC heel gericht variëteiten met de juiste SNP's met elkaar gaan kruisen om een variëteit te maken met een gewenst groeiprofiel: *early bulking*, of juist een hogere opbrengst na een langere groeifase.
Lindenbergh was na afloop van hun presentatie vol lof over wat de groep deze week had gepresteerd. Hij deed zelfs een oproep aan het publiek in de zaal: 'Ik hoop dat studenten van jullie contact met ons opnemen voor een stage of een baan. Want wiskundigen kunnen echt iets toevoegen aan deze business.'