Ruimtelijk beeld nitraatconcentraties in grondwater

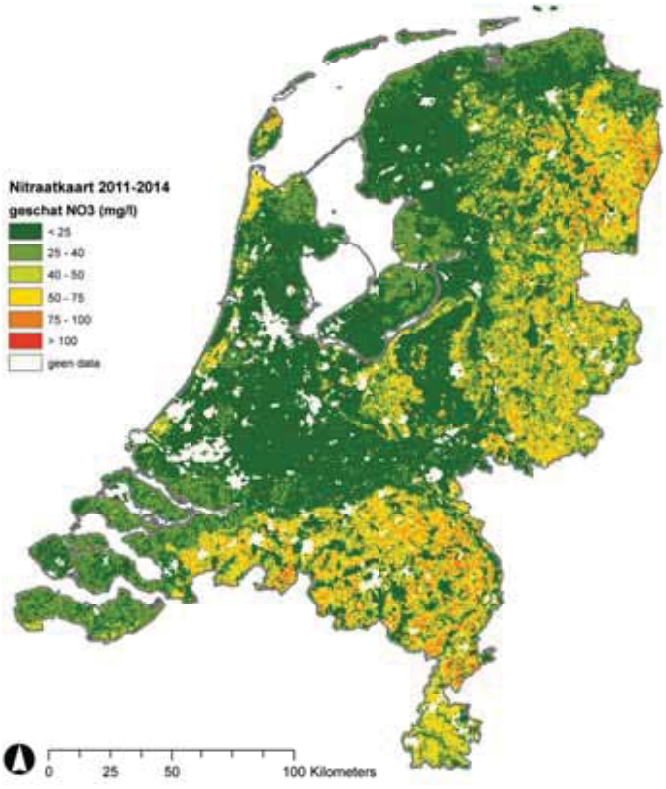
Data Science is de wetenschap achter Big data

De opkomst van Big Data wordt ook wel de vierde industriële revolutie genoemd. Daarin is data de grondstof van nieuwe industriële processen. Met Big Data kwam ook het vakgebied van de 'data science' op. Terwijl Big Data verwijst naar de hele grote bak met getallen, gaat het bij data science om wat je met die getallen doet. Data omzetten in bruikbare producten, dat is de 'science' achter Big Data.

Door: Job Spijker en Astrid Vrijhoef

Over de auteurs:
Job Spijker en Astrid Vrijhoef zijn wetenschappelijk medewerkers bij het RIVM

Ook de bodemprofessional gaat de ontwikkelingen in de data science merken. De technologie uit de data science wordt steeds vaker toegepast bij ruimtelijke vraagstukken. Bij het RIVM passen we zogenaamde 'machine learning' technieken toe om grote datasets van ruimtelijke informatie te verwerken, en te combineren met monitoringnetwerken. Deze machine learning technieken zijn gebaseerd op zelf lerende computer algoritmen. Door de combinatie van machine learning, ruimtelijke gegevens (GIS-data) en resultaten van de monitoringnetwerken kunnen landsdekkende kaarten gemaakt worden. Een voorbeeld van zo'n kaart is de nitraatkaart. Dit is een kaart met een ruimtelijk beeld van de nitraatconcentraties in water dat in een bepaalde periode uitspoelt uit de wortelzone van landbouwpercelen en natuurgebieden naar het grond- en oppervlaktewater in Nederland.



FIGUUR 1: DE NITRAATKAART VAN NEDERLAND VOOR DE PERIODE 2011-2014. DE KAART GEEFT DE NITRAATCONCENTRATIES IN WATER DAT UITSPOELT UIT DE WORTELZONE VAN LANDBOUWPERCELEN EN NATUURGEBIEDEN NAAR HET GROND- EN OPPERVLAKTEWATER IN NEDERLAND.

BIG DATA EN DATA SCIENCE

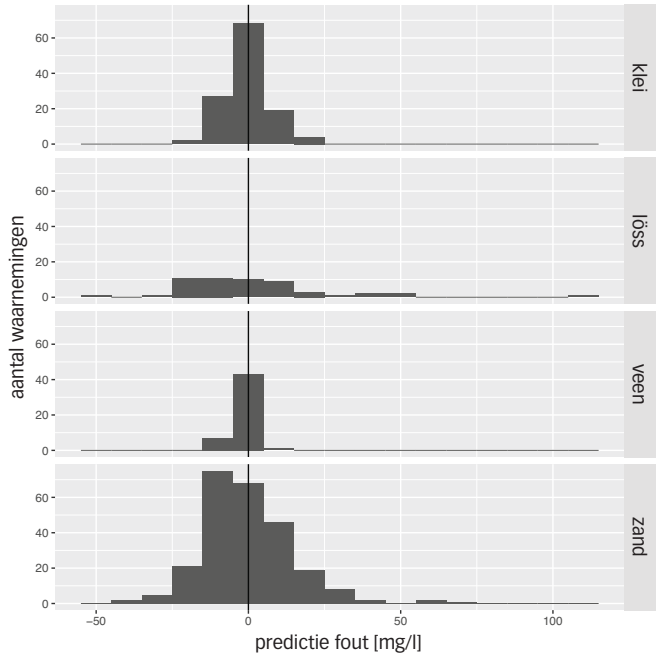
Big data gaat niet alleen maar om veel data. Big Data gaat vooral om het combineren van meer verschillende databestanden. Ieder met een eigen bron, eigen eigenschappen van de data en eigen data structuur.

Bij data science draait het om het creëren van producten op basis van die data. Producten waarbij de data is omgezet in informatie. Deze informatie kan gebruikt worden om direct te handelen of om je nieuwe inzichten te geven. Daarnaast kan de informatie uitgegeven worden als een zelfstandig product. Bij de data analyse

> Data science is het creëren van producten op basis van Big Data

draait het niet alleen om de klassieke statistiek of de wiskunde alleen. Een data scientist heeft verschillende eigenschappen, zoals expert kennis van het domein waarin de data wordt gebruikt, het kunnen gebruiken van programmeertalen in plaats van kant en klare software en het kunnen toepassen van machine learning.

Maar is data science nu oude wijn in nieuwe zakken? Nee, er zijn belangrijke verschillen. Door de omvang van de data zijn de gebruikelijke softwarepakketten zoals Microsoft Excel niet meer toereikend. Voor data science worden voornamelijk programmeeromgevingen zoals Python en R gebruikt, en de gebruikte



FIGUUR 2: HISTOGRAM VAN DE VERSCHILLEN TUSSEN GEMETEN NITRAATCONCENTRATIES OP LMM-BEDRIJVEN EN DE GEÏNTERPOLEERDE NITRAATCONCENTRATIES VAN HET MODEL OP DEZE BEDRIJVEN.

code voor analyses is veelal gebaseerd op open source computercode. Daarnaast is er nog een belangrijk verschil. Waar we vroeger zelf het model moesten bedenken en opstellen laten we dit nu door de machine learning algoritmes van de computer doen.

Een klassieke modelleur gebruikt zijn expert kennis en (statistische) formules om gedrag van variabelen te verklaren of te voorspellen op basis van relaties met andere variabelen. Het opstellen van een wiskundig model is een secure en tijdrovende klus en vereist veel gedetailleerde mechanistische kennis van het onderliggende systeem. De data scientist benadert het op een andere wijze, hij of zij gaat uit van de data en veel minder, zoals bij de modelleur, uit van a priori kennis over het systeem. De data scientist gaat uit van algoritmes zoals Random Forest, Support Vector Machines of neurale netwerken om patronen in de data zichtbaar te maken. Deze patronen zijn een respons op basis van de verklarende variabelen in de dataset.

De kracht van data science zit erin dat het relatief eenvoudig is om tot voorspellingen, zoals een kaart, te komen op basis van data. Als de data zijn verzameld en in de juiste vorm is gegoten is het verder aan de computer om patronen en relaties in de data te duiden. Dit kan heel snel tot nieuwe en onverwachte inzichten leiden. Maar waar de modelleur veel tijd besteedt aan het opstellen van het model en a priori uitgaat van causaliteit, besteedt de data scientist vooral tijd aan het interpreteren en beoordelen van het resultaat. De gevonden relaties en patronen hebben niet per se een oorzakelijk verband. Daarom is analyse en duiding van de onzekerheden in de voorspelling van het model ook erg belangrijk.

Een ander groot voordeel van het werken met machine learning algoritmes is de reproduceerbaarheid. Een klassiek model bevat veel aannames en keuzes die per modelleur kunnen verschillen. Een algoritme is meer onbevooroordeeld en maakt zijn 'eigen' keuzes. De uitkomst van een machine learning algoritme is robuuster voor de invloed van de 'bias' van de gebruiker.

Machine learning is geen kwestie van een druk op de knop. Er zal vooral heel zorgvuldig naar de data gekeken moeten worden. Op basis van de onderzoeksvraag en statistische eigenschappen van de data moet er een algoritme gekozen worden. De data moet vervolgens in de juiste vorm worden gemasseerd. Vaak worden hierbij aparte procedures geprogrammeerd voor het omgaan met ontbrekende waarden en uitbijters. Via weer andere procedures wordt het gekozen algoritme geoptimaliseerd. Statische eigenschappen, zoals verklaarde variantie en predictie fout, zijn hierbij vaak leidende criteria. Door slimme keuzes te maken in datavoorbewerking, selectie van verklarende variabelen, algoritmes, en optimalisatie-procedures probeert de data scientist een zo optimaal mogelijk model te creëren.

DE NITRAATKAART

Het RIVM meet met het Landelijk Meetnet effecten Mestbeleid (LMM) onder andere nitraatconcentraties in het water dat uitspoelt uit percelen op landbouwbedrijven en in slootwater op deze bedrijven (RIVM, 2017, Buis e.a., 2013). Hiermee volgt het LMM de kwaliteit van grond- en oppervlaktewater op landbouwbedrijven, gerelateerd aan de bedrijfsvoering op deze bedrijven. Dit om de effecten van de veranderingen in de bedrijfsvoering, als gevolg van het mestbeleid, op de waterkwaliteit snel in beeld te krijgen. Een van de onderdelen waar het LMM specifiek op monitort, is de uitspoeling van nitraat naar het grondwater. Dit nitraat is een omzettingsproduct van stikstof in de bodem, bemesting is een belangrijke bron van dit stikstof. Het LMM meet op circa 450 bedrijven verspreid over Nederland. Dit geeft wel ruimtelijk verdeelde puntinformatie, maar geen ruimtelijk dekkend beeld.

> Zelf lerende algoritmen leiden tot nieuwe inzichten

Om de nitraatkaart te maken worden meetgegevens uit het LMM gecombineerd met veel verschillende databronnen, zoals landgebruiksgegevens, bodemkaarten, grondwaterstanden en statistieken over bemesting (stikstofbelasting). Om te komen tot een landsdekkende kaart worden de gegevens ook gecombineerd met de RIVM data van het Trendmeetnet Verzuring (TMV).

In het LMM wordt op basis van de bodemkaart onderscheid gemaakt naar grondsoortregio's waarbij vier hoofdgrondsoortregio's worden gehanteerd: Klei, Veen, Löss, en Zand. Voor de

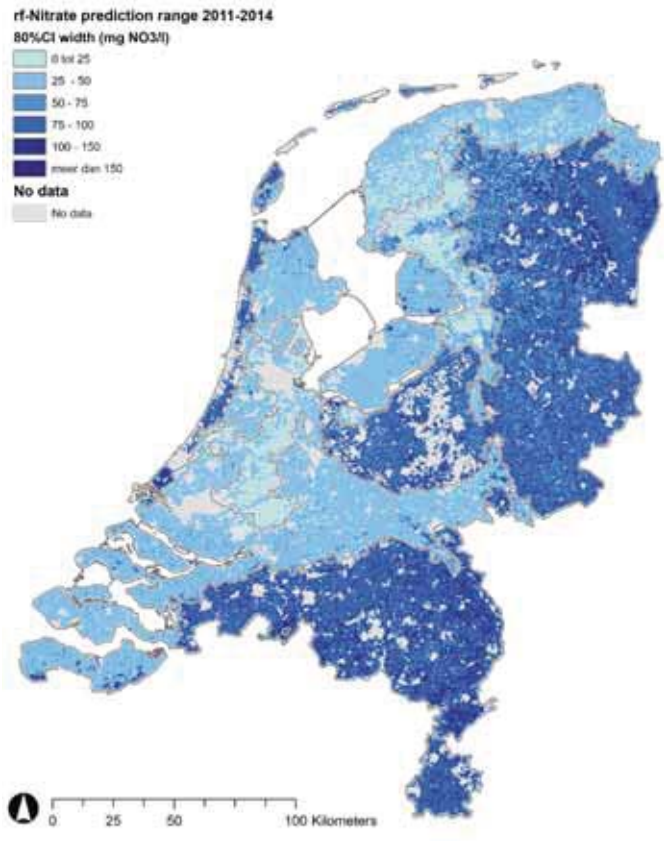
Op basis van welke gegevens is de nitraatkaart samengesteld?

Landbouwgebieden:

- Nitraatconcentraties LMM-locaties
- Bedrijfsoppervlakken LMM-locaties
- Grondsoortregiokaart LMM
- Landgebruikskaart Nederland
- Grondwatertrappenkaart (Gt)
- Bodemkaart
- Grondsoortenkaart meststoffenwet
- Mestgebruik per gemeente/grondgebruik (MAMBO-data)

Natuurgebieden:

- Nitraatconcentraties TMV-locaties
- Landgebruikskaart
- Stikstofdepositiekaart
- Bodemkaart



FIGUUR 3: KAART MET HET 10 - 90% BETROUWBAARHEIDSINTERVAL VOOR DE INDIVIDUELE 500*500 METER BLOKKEN OVER 2011-2014.

kaart zijn voor de Klei-, Zand- en Lössregio de gegevens voor uitspoelingswater en voor de Veenregio de gegevens voor slootwater gebruikt. Nitraatconcentraties in het uitspoelingswater van natuurgebieden werden door het RIVM gemeten in het Trendmeetnet Verzuring tot 2015 Beide meetnetten geven data die betrekking hebben op een beperkt oppervlak, namelijk bedrijf of natuurlocatie. Door deze gegevens te combineren met kaartinformatie en bemestingsdata op nationaal niveau en gebruik te maken van een statistisch model, is de nitraatconcentratie geschat voor heel het landelijk gebied in Nederland, ook daar waar niet is gemeten. Dit is in de nitraatkaart weergegeven. Ieder model kent zijn onzekerheden, ook deze hebben we in een kaart weergegeven.

METHODE

De nitraatkaart (Figuur 1) is gebaseerd op de in het LMM en TMV beschikbare data.

Voor landbouwgebieden zijn de beschikbare kaarten, zoals de bodemkaart of bemestingskaart, eerst omgezet naar gridkaarten en in hetzelfde coördinatenstelsel gezet (zie Tekstkader). Vervolgens zijn de percelen van de LMM-bedrijven digitaal over deze kaarten gelegd. Van elke locatie zijn zo de verschillende karakteristieken bekend (Boumans e.a. 2008). Deze karakteristieken hebben we gecombineerd met de bekende nitraatconcentratie op deze bedrijven. Daarna zijn, met behulp van het Random Forest algoritme (Breiman, 2001), de nitraatconcentraties geschat voor heel Nederland (verdeeld in blokken van 500 bij 500 meter). Random Forest is een algoritme gebaseerd op zogenaamde regression trees, een veel gebruikte statistische techniek om voorspellingen te doen.

Voor natuurgebieden is op elke locatie uit het TMV een stikstofbelasting berekend vanuit de omringende blokken en de stikstofdepositie. Deze en de andere karakteristieken zijn gebruikt om met behulp van dezelfde statistische aanpak als bij landbouwgebieden voor heel Nederland nitraatconcentraties te berekenen.

De nitraatkaart is samengesteld uit een combinatie van deze twee kaarten. Daar waar de 500 m blokken van natuur en landbouw overlappen is uitgegaan van het gemiddelde.

MODEL ONZEKERHEDEN

De nitraatkaart is gebaseerd op een voorspelling van een statistisch model. Elk model heeft onzekerheden (Figuur 2). Op basis van de gebruikte gegevens kan het model ongeveer 50% van de waargenomen regionale verschillen in de nitraatconcentraties verklaren, dit is de zogenaamde statistisch verklaarde variantie in het model. De overige 50% van de verschillen wordt veroorzaakt door factoren die niet in het model zijn opgenomen. Het Random Forest algoritme doet een groot aantal voorspellingen voor de nitraatconcentraties in Nederland. De nitraatkaart is de gemiddelde voorspelling. Om een indruk te geven van de bandbreedte van de voorspellingen is er ook een kaart gemaakt met het 10-90%-interval van de voorspelde waarden per 500 bij 500 m blok. Deze kaart staat in Figuur 3. De onzekerheden vertonen een duidelijk ruimtelijk patroon, daar waar de nitraatconcentratie het hoogst is, is ook de onzekerheid hoog. Dit is niet ongebruikelijk in statistische modellen. In absolute waarden is de onzekerheid ook groot. De onzekerheid van de kaart betekent dat er geen uitspraken gedaan kunnen worden over de waterkwaliteit op lokaal niveau, bijvoorbeeld op het niveau van enkele pixels. Wel kan de kaart gebruikt worden voor uitspraken op regionaal niveau. Zo kunnen regio's of provincies wel met elkaar vergeleken worden. De verwachting is dat door het verder ontwikkelen van het model en gebruik te maken van meer (open) data, zoals de gegevens over de gewasrotatie, de onzekerheden in komende versies de kaart zullen afnemen.

CONCLUSIE

Data science, het vakgebied rond Big Data, is een andere manier om data om te zetten in bruikbare producten. Bij data science gaat men meer uit van de patronen in de data zelf, en niet per se van a priori causaliteit. Dit kan leiden tot snellere en robuustere inzichten in de overvloed aan data die nu beschikbaar is. Maar er zijn ook valkuilen. Het is de computer die met zijn 'machine learning' algoritmen het leeuwendeel van de data-analyse uit handen neemt. Een analyse en duiding van de onzekerheden van het resultaat zijn onlosmakelijk onderdeel van een product op basis van machine learning.

REFERENTIES

1. Boumans, L.J.M., Fraters, B. and Van Drecht, G.,2008, Mapping nitrate leaching to upper groundwater in the sandy regions of The Netherlands, using conceptual knowledge. Environ Monit Assess (2008) 137:243–249.
2. Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.
3. Buis, E., Hoogeveen, M.W., Fraters, D., Van Leeuwen, T.C., 2013, De gevolgen van 20 jaar mestbeleid op het milieu. Het Landelijk Meetnet effecten Mestbeleid: het meten van landbouwpraktijk en waterkwaliteit. Bodem, 23(1): 19-21.
4. RIVM, 2017, Website Landelijk Meetnet Mestbeleid, http://www.rivm.nl/lmm, geraadpleegd 1 mei 2017.