Schelpen op druksensoren digitaal opsporen met datascience

Laura Snip, Inke Leunk (Hoogheemraadschap De Stichtse Rijnlanden), Jeroen Gorter (waterschap De Dommel)

In het beheergebied van Hoogheemraadschap De Stichtse Rijnlanden staan 661 druksensoren die gebruikt worden voor peilbeheer. Met behulp van een datascience-model op basis van kwantielenregressie is het mogelijk om tijdig vervuiling door schelpdieren op de sensoren op te sporen. Door deze automatiseringsstap kan meer data vaker gecontroleerd worden en kan vervuiling eerder opgespoord worden. Hierdoor zal de data beter op orde zijn voor gebruik van hydrologische modellen.

Hoogheemraadschap De Stichtse Rijnlanden (HDSR) meet op 661 locaties met druksensoren geautomatiseerd de waterstand. In het traject van meting tot eindgebruiker kunnen op allerlei punten dingen misgaan die tot foute data leiden. Dit is beschreven in een eerder H2O-vakartikel: ‘Datavalidatie: voorbeelden uit de praktijk van waterkwantiteitsmetingen’ [1].
Bij HDSR wordt het controleren en valideren van de metingen deels geautomatiseerd en deels handmatig gedaan. Het handmatig controleren en valideren gebeurt door de waterstand in een grafiek weer te geven en afwijkingen visueel op te sporen. Deze handmatige controle vindt circa drie keer per jaar plaats, waardoor het soms enkele maanden kan duren voordat foute data worden opgespoord. Afhankelijk van de locatie, het kunstwerk en het normale gedrag van de waterstand worden fouten eerder of later opgespoord. Vanwege de grote hoeveelheid sensoren is het wenselijk de controle en validatie verder te automatiseren.
Het opsporen van vervuiling van de sensor door schelpdieren is daarvan een begin. Soms groeien er schelpen op een sensor van de waterstand en beïnvloeden dan de druk die er gemeten wordt. Dit geeft een heel specifieke, ‘piekachtige’ afwijking. Deze afwijkingen beginnen vaak klein en groeien langzaam, waardoor ze pas na langere tijd zichtbaar worden. Op een locatie met sterk wisselende waterstand zullen de pieken minder snel opvallen; vaak blijft de stand lange tijd binnen min of meer normale grenzen (zie afbeelding 1). Dit zorgt ervoor dat de huidige automatische validatiemethodes de afwijking vaak niet registreren en het lang kan duren voordat de fout in de meting wordt opgespoord. Doordat de afwijking klein begint en steeds groter wordt, is het ook moeilijk te bepalen vanaf welk moment de data moeten worden afgekeurd.

Afbeelding 1. Periode met foute metingen in gele cirkel. De metingen zijn fout, maar blijven absoluut gezien redelijk binnen normale bandbreedte

In dit artikel wordt beschreven hoe een proof of concept is uitgevoerd om deze vervuiling door schelpdieren op te sporen. Er zijn verschillende datascience-modellen getest, maar hier wordt alleen de toegepaste methode getest.

Uitgangspunten

Voorafgaand aan het onderzoek zijn de uitgangspunten besproken met de datavalidator van HDSR.
Zij heeft aangegeven dat ze de vervuiling door schelpdieren herkent op basis van de metingen van de waterstand zelf, zonder gebruik van andere meetpunten of weerdata. Daarom is gekozen voor een datascience-methode die alleen gebruik maakt van de data van de sensor zelf en om verder geen externe databronnen te gebruiken.
Daarnaast zijn alle modellen getraind op data die goedgekeurd zijn. De bedoeling is namelijk dat het model de afwijkingen niet kan voorspellen, dus moet het model geen kennis hebben van die afwijkingen.
Een afwijking als gevolg van vervuiling door schelpdieren kenmerkt zich door een ‘springerigheid’ in de metingen (zie afbeelding 1). Dit houdt in dat de frequentie van data die gebruikt wordt, deze springerigheid ook moet hebben. De waterstand wordt iedere vijftien minuten gelogd, of als de waarde verandert. Er is voor gekozen om alle data om te zetten naar een tijdsinterval van drie minuten. Hierdoor blijft de springerigheid behouden, maar blijft de grootte van de databestanden nog handelbaar.
In overleg zijn de modellen zo getraind dat ze eerder een afwijking te veel aangeven dan een afwijking missen. De modellen worden ingezet als waarschuwing voor de datavalidator, niet voor automatische afkeuring. De datavalidator kijkt daarom liever een keer te veel dan dat ze iets mist.

Methodes

Dataverzameling uit FEWS-WIS

Uit het in Delft ontwikkelde waterinformatiesysteem FEWS-WIS zijn data verzameld van de waterstanden tot en met 2016. Deze data zijn voor het grootste gedeelte gevalideerd en bevatten periodes waar de datavalidator de metingen heeft afgekeurd wegens schelpengroei. Om verdere verwerking van de data te versimpelen, zijn de ruwe data omgezet naar gelijkmatige tijdsintervallen. Zoals hierboven uitgelegd is gekozen voor een interval van drie minuten en zijn de data lineair geïnterpoleerd. Hierbij is aangenomen dat met dit kleine tijdsinterval de sprongen in data door schelpen nog goed te detecteren zijn en als er geen waarde beschikbaar is, de verandering van de waterstand lineair verloopt. Voor de proof of concept hebben is gemaakt van een export uit FEWS-WIS, waarbij de ruwe data per sensor in een Excelbestand is gezet met de opmerkingen van de datavalidator erbij. Door de opmerkingen kunnen afgekeurde data worden herkend en uit de trainingsset gehaald, maar kan ook het model worden getest op het herkennen van afwijkingen.

Datascience-model Kwantielenregressie

Het detecteren van schelpen op de sensor is in principe een anomalie(afwijkings)detectie. Met een datascience-model wordt voorspeld wat de waarde zou moeten zijn, gebaseerd op voorgaande metingen. Vervolgens wordt gekeken hoever de daadwerkelijke waarde afwijkt van de voorspelling. Hoe groter de afwijking, hoe groter de kans dat het een foute meting is, bijvoorbeeld als gevolg van schelpen op de sensor. Om de waterstand op een volgend moment te voorspellen zijn verschillende datascience-modellen toe te passen. Voor deze toepassing werkte kwantielenregressie het beste. Bij kwantielenregressie worden er drie lineaire regressiemodellen ontwikkeld, één voor de mediaan en één voor elk gevraagd kwantiel [2]. Het resultaat is dan een onder- en bovengrens waarbinnen de waterstand verwacht wordt.

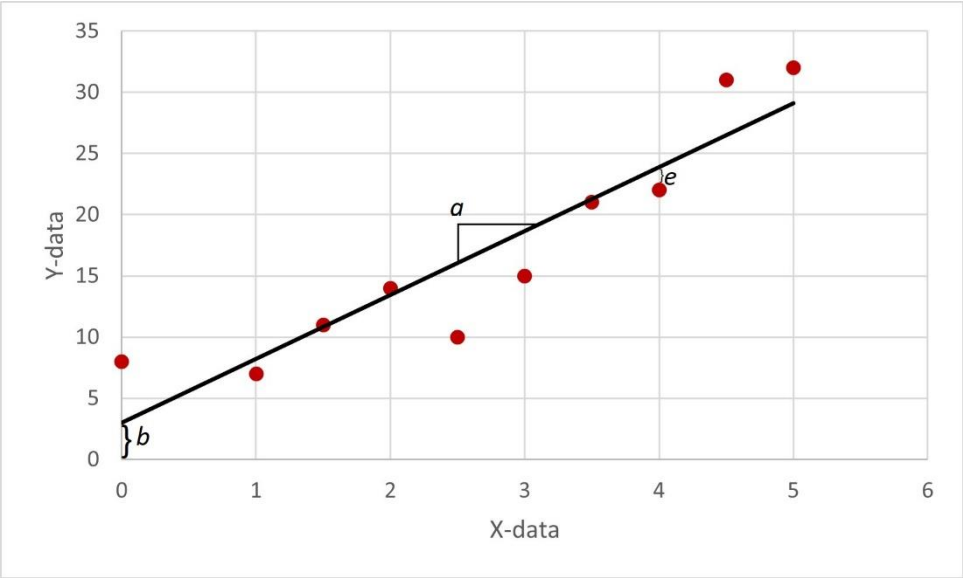
Lineaire regressie

Bij lineaire regressie probeert het model een lineair verband tussen verschillende punten te leggen waarbij de volgende vergelijking wordt gebruikt:

$y=a*X+b+e$

y is de output (in dit geval de gemeten/voorspelde waterstand)
a is de helling van de lijn
X is de inputdata
b is de kruising van de grafiek bij de y-as
e is de ruis die bij de metingen aanwezig kan zijn

Deze parameters zijn ook weergegeven in afbeelding 2, waar een voorbeeld van lineaire regressie bij een dataset te zien is.



Afbeelding 2. Voorbeeld van een schatting van een lineair regressiemodel waarin aangegeven is hoe de parameters a, b en e geschat kunnen worden

Door gebruik te maken van trainingsdata leert het model de parameters a, b en e zo in te stellen dat de voorspelde waarde (y) zo dicht mogelijk bij de werkelijk gemeten waardes in de buurt komt. De inputdata (X) hoeft bij een datascience-model niet één parameter te zijn, het kunnen er veel meer zijn. Omdat de datavalidator heeft aangegeven dat ze geen externe factoren wil gebruiken, bestaat de inputdata alleen uit de daadwerkelijk gemeten waterstanden. Het model kreeg niet alleen de waterstand van een tijdstap eerder mee, maar ook van de negen stappen daarvoor. Er worden in totaal dus tien waardes meegegeven, wat overeenkomt met een half uur aan waterstanden.

Kwantielenregressie

Bij kwantielenregressie worden er drie lineaire regressiemodellen getraind en getest. Eén model berekent het 0.5-kwantiel (de mediaan) door het absoluut gemiddelde verschil tussen de voorspelling en de meting zo klein mogelijk te maken. Dit model probeert dus de gemeten waarde zo goed mogelijk te voorspellen. De twee andere modellen worden gebruikt om de boven- en ondergrens te bepalen op basis van vooraf gekozen kwantielen. Deze modellen kunnen worden beschouwd als de (on)zekerheidsmarge waarbinnen de gemeten waardes naar verwachting zullen vallen. In deze toepassing kan de waterstand vrij nauwkeurig bepaald worden met een lineair regressiemodel, waardoor strakke kunnen worden ingesteld. De kwantielen zijn daarom op 0.001 en 0.999 gezet.
Per sensor is één model getraind met 80% van de data en getest met 20% van de data. Met de testdata wordt bepaald hoe goed het getrainde model daadwerkelijk de gemeten waardes kan voorspellen. Het is belangrijk om te testen met data die het model nog niet heeft ‘gezien’, om de werkelijke nauwkeurigheid van het model te bepalen. Zodra het model als voldoende nauwkeurig wordt beschouwd, kan het worden getest op het detecteren van de afwijkende periodes als gevolg van schelpengroei.

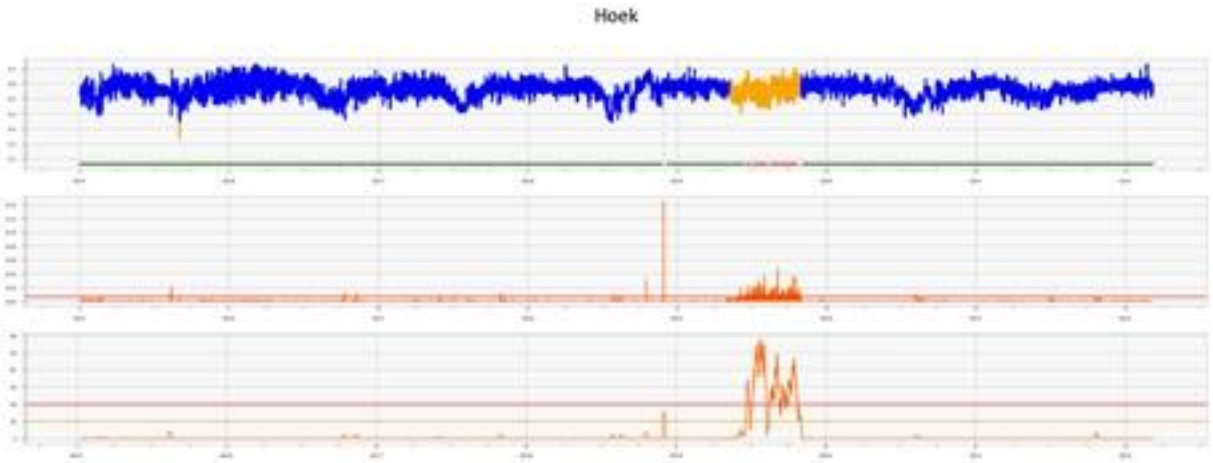
Signalering afwijkingen

Met het getrainde kwantielenregressiemodel kunnen, in een nieuwe dataset, afwijkingen opgespoord worden. Maar niet elke gedetecteerde afwijking hoeft meteen een schelp te zijn. Daarom worden de

afwijkingen per dag opgeteld en als deze boven een grens komen, is het aannemelijk dat er schelpen op de sensor zitten. De grens is nu vastgesteld op 20 afwijkingen. Als er dus meer dan 20 afwijkingen per dag zijn, is het aannemelijk dat er schelpen op de sensor zitten.

Resultaten

Doordat de datavalidator al in de data had aangegeven wanneer er vervuiling door schelpen optrad (afbeelding 3), kon de daadwerkelijke aanwezigheid van schelpen worden vergeleken met de momenten die door het datascience-model werden aangewezen. Hierdoor kon worden getest of het model een afwijking kon detecteren. In afbeelding 3 zijn drie grafieken weergegeven. De bovenste is de gemeten waterstand, waarbij de geel gekleurde data zijn afgekeurd door de datavalidator. De middelste grafiek geeft het verschil tussen de voorspelling van het datascience-model en de gemeten waarde weer per tijdsstap. Hierbij is een acceptabel verschil vastgesteld op 85% en weergegeven als de rode lijn. De onderste grafiek telt het aantal verschillen buiten de vastgestelde kwantielen bij elkaar op. Zoals te zien is in de onderste twee grafieken van afbeelding 3 komen de momenten van een verschil tussen de voorspelling en de meting overeen met de gele afgekeurde data. In de middelste grafiek is te zien er ook op andere moment uitschieters in de waterstand plaatsvinden (vooral naar beneden). Deze afwijkingen worden dus ook gedetecteerd.



Afbeelding 3. Resultaten van het datascience-model van locatie Hoek, met in de bovenste figuur de gemeten waterstanden en de afgekeurde data in geel. De middelste grafiek geeft het verschil tussen de voorspelling van de waterstand en de gemeten waterstand, met de rode lijn als grens van een acceptabel verschil. In de onderste figuur is het aantal momenten per dag waarop een verschil tussen voorspelling en afwijking hoger is dan de acceptabele grens opgeteld. De rode lijn geeft 20 verschillen per dag aan

Conclusie

Het ontwikkelde datscience-model is in staat om vervuiling door schelpen op druksensoren te detecteren. Er zijn meer dan 45 sensoren getest, op verschillende locaties met verschillende patronen in de waterstanden. Zo is er een zaagtandpatroon getest bij een locatie met een gemaal, een vrij stabiel zomer/winterpatroon op een locatie zonder kunstwerk en een sterk wisselend patroon op een locatie bij een stuw. Tot dusver lijkt de methode robuust genoeg om bij verschillende patronen afwijkende data te detecteren.

De datavalidator is tevreden met de behaalde resultaten. Haar vermoedens kunnen nu bevestigd worden door het datascience-model, waarna de sensor schoongemaakt kan worden en de data weer betrouwbaar zijn. Geconcludeerd wordt daarom dat de proof of concept geslaagd is.

Aanbevelingen

De methode is nog niet ingebouwd in het softwareprogramma FEWS-WIS, waarin de data van het peilbeheer verzameld wordt. Het is de wens van de datavalidator dat deze extra datavalidatiestap in FEWS-WIS wordt gebouwd, zodat alle informatie in één systeem beschikbaar komt. Als de methode in FEWS-WIS beschikbaar is, kunnen de modellen geautomatiseerd periodiek gedraaid worden voor alle locaties. Het is niet mogelijk alle locaties regelmatig te controleren zolang de data handmatig in het model gevoerd moet worden en het model daarna handmatig gedraaid moet worden.

Daarnaast moet uitgezocht worden hoe ieder getraind model van een sensor opgeslagen kan worden. Er zijn 661 sensoren en dat betekent ook dat er 661 getrainde modellen nodig zijn. Voorlopig zal het model gebruikt worden voor locaties waar vaker schelpen zijn gedetecteerd, voordat het uitgebreid wordt naar andere locaties.

Het patroon van de waterstanden zal zeer waarschijnlijk niet erg veel veranderen in de tijd en daarom kan het getrainde model op data uit 2016 lang gebruikt worden. Mocht het getrainde model toch niet meer goed werken, dan is het nodig om het model te hertrainen met nieuwe data. Hiervoor moet ook een robuuste methode ontwikkeld worden.

De methode is erg effectief gebleken in het opsporen van schelpen omdat dit plotselinge veranderingen zijn. Langzame veranderingen, zoals een drift van de sensor, zullen waarschijnlijk niet gesignaleerd worden. Hiervoor moet een andere methode gebruikt worden.

Referenties

1. Hagedooren, H., Jungermann, N., Benjamin, E., Leunk, I. (2020). ‘Datavalidatie: voorbeelden uit de praktijk van waterkwantiteitsmetingen’. *H2O-vakartikelen*, 2 juli 2020. https://www.h2owaternetwerk.nl/vakartikelen/datavalidatie-voorbeelden-uit-de-praktijk-van-waterkwantiteits-metingen

2. Scikit (2022.) *Quantile regression.* https://scikit-learn.org/stable/auto_examples/linear_model/plot_quantile_regression.html