Waterdatamining – de eerste ervaringen uit de drinkwaterpraktijk

Dirk Vries, Erwin Vonk (KWR Watercycle Research Institute), Johan van Erp (Brabant Water), Roel Diemel (Brabant Water), Wybren de Jong (Vitens)

KWR heeft onlangs in samenwerking met Brabant Water en Vitens onderzoek uitgevoerd naar de mogelijkheden van datamining voor de drinkwatersector. Een inventarisatie van assetmanagement-kennisvragen, een literatuurstudie naar datamining en de eerste praktijkervaringen uit twee pilotprojecten vormen voor drinkwaterbedrijven een eerste, voorzichtige stap richting een datagedreven bedrijfsvoering. Drinkwaterbedrijven kunnen de verzamelde kennis inzetten om hun operationele processen (bijvoorbeeld assetmanagement) te verbeteren. Het onderzoek heeft geleid tot waardevolle inzichten voor toekomstige projecten en nieuwe kansen. Daarbij blijft het wel zaak om de kwaliteit en kwantiteit van beschikbare data met datamining-ambities af te stemmen en experts van verschillende vakgebieden samen te laten werken.

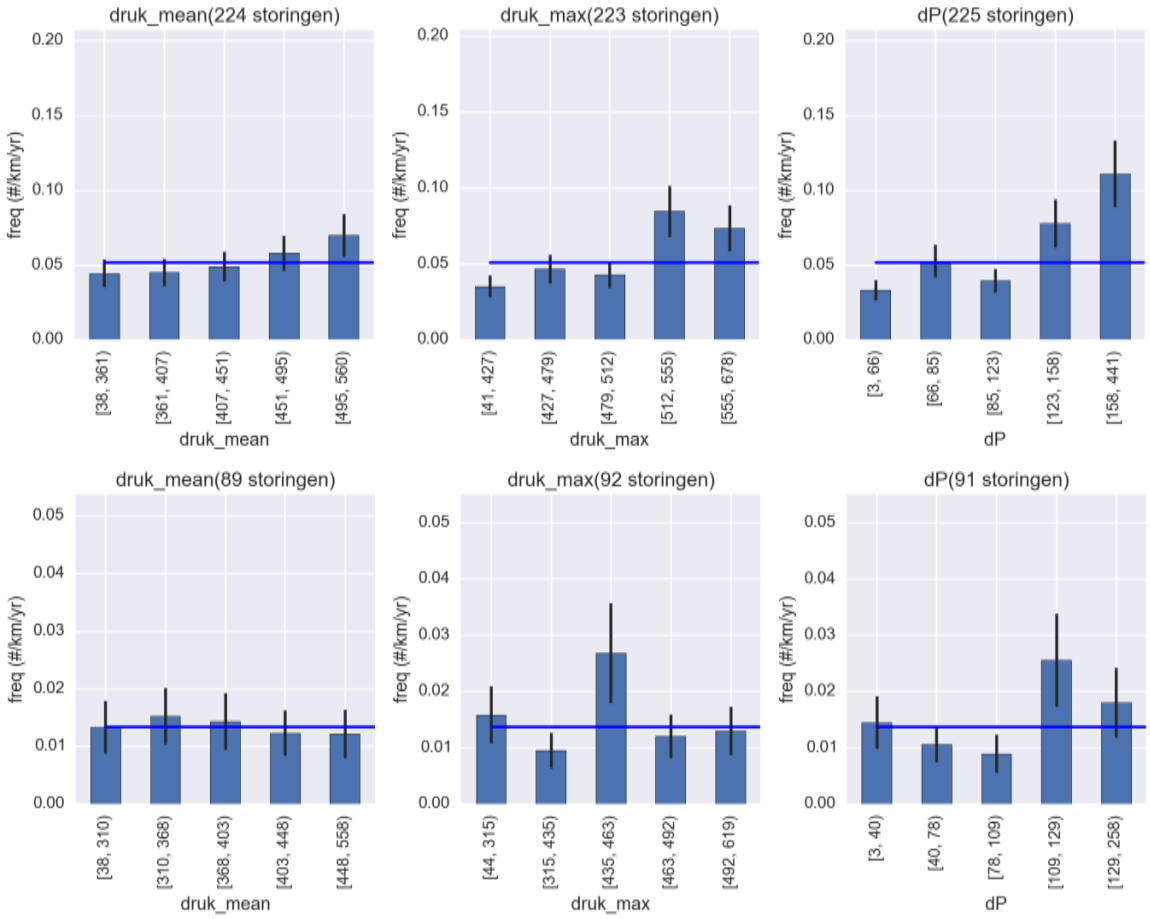
Ook voor de drinkwatersector kan datamining worden gebruikt om bedrijfsprocessen te verbeteren. Uit een inventarisatie door KWR is gebleken dat assetmanagers bij drinkwaterbedrijven behoefte hebben aan meer inzicht in de conditie en faalkansen van diverse typen *assets* [1]. Een aantal faalmechanismen kan reeds herkend worden met gerichte metingen en monitoringsystemen. Voorbeelden zijn detectie van grote lekken door het meten van volumestroom en druk en het analyseren van deze gegevens, zoals met de VLPV-methode [2] of anomaliedetectie (bijv. [3]). Datamining kan vooral uitkomst bieden bij vraagstukken waarbij nog geen duidelijk inzicht bestaat in de fysische processen die een rol spelen en waar data continu gemeten worden of ruimschoots beschikbaar zijn. Mogelijke toepassingen van datamining voor de drinkwaterinfrastructuur zijn bijvoorbeeld het herkennen en detecteren van faalcondities of andere afwijkingen, het koppelen aan kosten voor de levenscyclus of het inschatten van de slijtage van pompen. Tot slot biedt datamining mogelijkheden om (bijna) *real-time* storingen in het leidingnet te signaleren.

Verkenning naar toepassingen in de drinkwatersector

In het kader van het Bedrijfstakonderzoek (BTO) van de Nederlandse drinkwaterbedrijven heeft KWR geïnventariseerd welke mogelijkheden datamining biedt om bestaande datasets van drinkwaterbedrijven in te zetten om operationeel assetmanagement te kunnen verbeteren. Daartoe zijn BTO-breed zowel de vraag (kennisbehoefte) als het aanbod (datasets en datamining-technieken) geïnventariseerd door middel van interviews bij de drinkwaterbedrijven, een gezamenlijke workshop en literatuuronderzoek. Ook zijn bij zowel Vitens als Brabant Water zogenaamde TKI (Topconsortia voor Kennis en Innovatie) -pilotprojecten uitgevoerd rondom datamining. Dit heeft interessante informatie opgeleverd voor toekomstige projecten. Tot slot heeft KWR een *toolbox* ontwikkeld, waarmee het analyseren van grote datasets wordt versneld. Deze 'Machine Learning Toolbox' biedt kant-en-klare functies die onderzoekers kunnen gebruiken om datagedreven modellen te ontwikkelen en interactief resultaten zichtbaar te maken.

In het pilotproject bij Brabant Water is samengewerkt met Nelen & Schuurmans en Witteveen+Bos. Centraal in deze casus stonden de datasets met klantmeldingen van Brabant Water, de

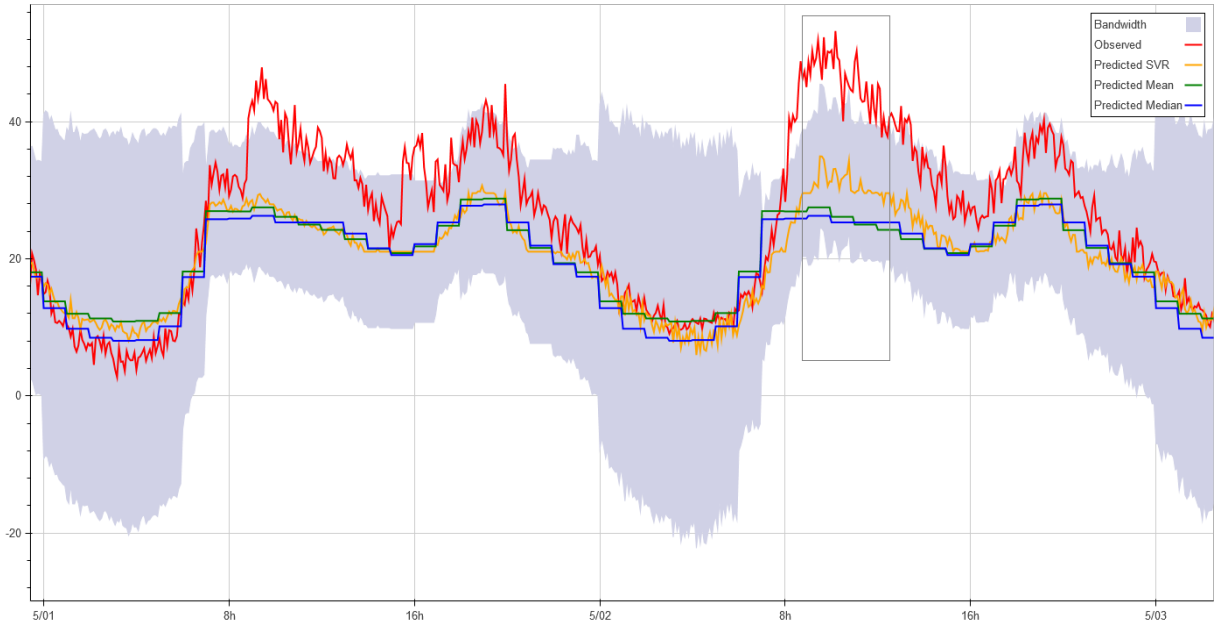
storingsregistratie van de Nederlandse drinkwaterbedrijven (USTORE [4]), procesdata (volumestroom- en drukmetingen bij drinkwater pompstations) en leidingnetgegevens. Door deze bronnen onderling met elkaar te combineren en aan te vullen met openbare databronnen (CBS-wijkgegevens en KNMI-temperatuurmetingen) is een aantal interessante verbanden aan het licht gekomen, die bestaand fysisch-experimenteel onderzoek grotendeels bevestigen. Zo blijkt er een verband te bestaan tussen buitentemperatuur en bruinwatermeldingen [5]. Ook kon een statistisch verband worden aangetoond tussen de storingsfrequentie van cementhoudende leidingen en de waterdruk die aanleverende pompstations geven [6]. Deze informatie helpt Brabant Water bij het opstellen van leidingnetsaneringsplannen en onderhoud (afbeelding 1).



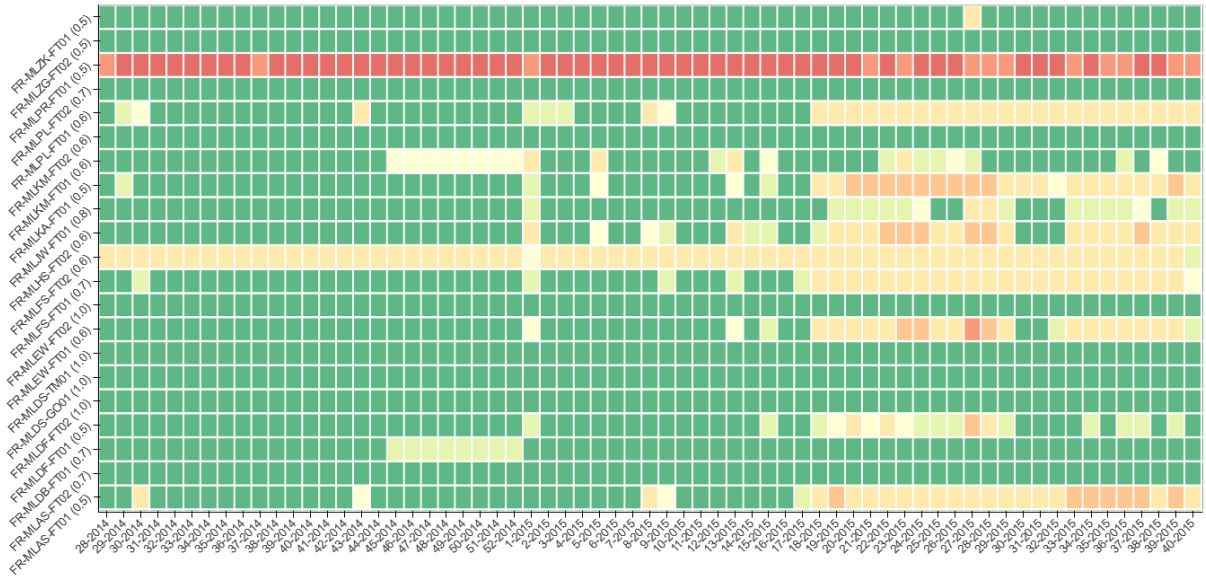
Afbeelding 1. Storingsfrequentie-histogrammen voor cementhoudende leidingen (bovenste rij plots) en niet-cementhoudende leidingen (onderste rij). Zichtbaar is dat bij cementhoudende leidingen de storingsfrequentie (aantal storingen per kilometer per jaar) samenhangt met de maximale druk (druk_max) op een dag en de verschildruk (dP). De blauwe lijn is de gemiddelde storingsfrequentie voor alle leidingen. Gemiddelde druk laat ook een trend zien, maar die is minder sterk. Voor overige leidingmaterialen is geen duidelijke correlatie gevonden.

Een tweede pilot bij Vitens was gericht op *real-time* analyse van datastromen die gegenereerd worden in de Vitens Innovation Playground (VIP), een proeftuin voor sensoren in het leidingnet in en rondom Leeuwarden. De verwachting is dat het in de toekomst mogelijk is om met behulp van de sensorsignalen (o.a. volumestroom, druk en elektrische geleidbaarheid) direct vast te stellen op welke locaties een lek of andere afwijking optreedt. Daarvoor zijn tijdens dit TKI-project diverse machine learning-algoritmes uitgetest op de gecombineerde meetgegevens, waarbij werd geprobeerd afwijkingen te herkennen (afbeeldingen 2 en 3). Uit de pilot bleek dat veel verschillende

technieken afwijkingen kunnen aantonen, maar dat validatie van deze technieken op vals-negatieve en vals-positieve alarmeringen pas kan uitsluiten welke methodiek voldoende betrouwbaar is, alvorens deze op grote schaal toe te kunnen passen [7, 8]. Het toekennen van meta-informatie aan deze alarmeringen is een stap die niet alleen belangrijk is voor validatiedoeleinden, maar ook voor het herkennen en groeperen van vergelijkbare alarmeringen.



Afbeelding 2. Een afwijking in gemeten volumestroom op één specifieke sensorlocatie in de Vitens-proeftuin. Oranje, groen en blauw zijn modelvoorspellingen op basis van verschillende modeltypen. Rood is het gemeten signaal. Zodra dit gedurende een aantal tijdstappen buiten de monitoringsbandbreedte uitkomt (grijs) wordt een alarmsignaal uitgegeven.



Afbeelding 3. Resultaten anomaliedetectie in de Vitens-proeftuin. Verticaal de sensorlocaties in de proeftuin en horizontaal de weeknummers (lopend vanaf week 28 in 2014 tot en met week 40 in 2015). Afwijkingen waarbij de sensorwaarde veel afwijkt van hetgeen voorspeld is, zijn rood gemarkeerd, kleinere afwijkingen worden oranje gekleurd. Indien er geen sprake is van een afwijking, is het betreffende vakje groen gekleurd.

De vier ingrediënten voor succes

Naast nieuwe inzichten door de toepassing van datamining, heeft het recente onderzoek waardevolle ervaringen opgeleverd:

1. Gebruik domeinkennis en registreer meta-informatie

Het is niet zo dat wanneer data van allerlei verschillende soorten bronnen, zonder dat er een mens aan te pas komt, in een slim algoritme worden gestopt, dit meteen nieuwe inzichten of resultaten oplevert. Het overgrote deel van een datamining-project gaat op aan het (al dan niet geautomatiseerd) op orde brengen van datasets en *feature engineering* (zie punt 2). Er is veel kennis nodig om ervoor te zorgen dat getallen op een juiste manier geïnterpreteerd worden. Uniform geregistreerde meta-informatie helpt bij het zoeken naar bepaalde patronen of situaties. Beheer en interpretatie van data vereisen een nauwe samenwerking tussen de dataspecialisten en experts op het gebied van hydraulica, materiaalkennis en bedrijfsvoering in de waterinfrastructuur.

2. Pas feature engineering toe

Feature engineering is een procedure waarbij een bestaande dataset verrijkt wordt met afgeleide parameters, die verkregen worden door simulaties met (andere) modellen of berekeningen op bestaande data. Het gaat om het opwerken van brondata naar invoerparameters die het *machine learning*-model zo accuraat en betrouwbaar mogelijk maken. Het is belangrijk goede invoerparameters te kiezen. Zo worden *machine learning*-algoritmes 'geholpen' om verbanden te leren (en vervolgens te herkennen) of te ontdekken. Storingen zijn bijvoorbeeld vaak gerelateerd aan de levensduur, maar databronnen geven meestal alleen informatie over begin- en einddatum. Door uit deze databronnen de levensduur als eenvoudige afgeleide parameter te bepalen en die vervolgens te correleren met het optreden van storingen, kan een *machine learning*-model betrouwbaardere resultaten opleveren.

3. Stel de data centraal

Een model is hooguit zo goed als de data waarmee het gevoed is. Het adagium 'eerst meten, dan weten' is hier van toepassing: eerst moet voldoende informatie worden verzameld voordat er gemodelleerd kan worden. Veelvoorkomende problemen zijn: (1) een te korte meetperiode of een te lage meetfrequentie, waardoor te weinig gegevens met voldoende variatie beschikbaar zijn, (2) het ontbreken van 'labels' of meta-informatie, die modelkalibratie en validatie mogelijk maken, (3) te weinig (uniform) geregistreerde gebeurtenissen en (4) onvoldoende kwaliteit van data, variërend van falende sensoren tot menselijke typefouten. Bij de laatste categorie kunnen datagedreven technieken juist ook worden ingezet om de datakwaliteit te verbeteren. Hiervoor is het nodig het gedrag van de sensoren vast te leggen en ontdekte fouten te registreren. De eerste categorie komt waarschijnlijk het vaakst voor: er is over een periode wel gemeten wat de waarde was van variabelen X en Y, maar niet die van variabele Z. Zodoende is het onmogelijk om een model op te stellen dat de waarde van Z voorspelt op basis van X en Y. Het is belangrijk om te beseffen dat, als men wil weten onder welke omstandigheden vaker storingen voorkomen, er een groot aantal storingen bij verschillende condities moet zijn opgetreden en geregistreerd.

4. Wees bewust van de valkuilen

Ondanks de vele kansen die (big) datamining biedt, is de methodiek gevoelig voor onjuist gebruik of foutieve interpretatie. Resultaten dienen door betrokkenen altijd kritisch bekeken te worden met het oog op veelvoorkomende valkuilen. Een voorbeeld is het gebruik van datasets die niet representatief zijn voor de werkelijkheid (bijvoorbeeld door fouten in de meetopstelling, variaties in meetmethodieken door de tijd of het selectief kiezen van steekproeven uit een grote populatie). Een klassieke valkuil bij datamining is 'over-fitting'. Het datagedreven model is in dat geval volledig geoptimaliseerd om de data die bij het opstellen zijn gebruikt (inclusief de daarin aanwezige ruis en irrelevante gegevens) te reproduceren. Een overgefit model zal daardoor bij nieuwe data onbetrouwbare resultaten leveren. Tot slot kunnen foutieve interpretaties ontstaan door 'data dredging' - correlaties zoeken zonder deze te valideren en uit correlaties oorzakelijke conclusies trekken (correlaties hoeven geen causaal verband te hebben).

Datamining

Het gericht zoeken naar verbanden of patronen in databases.

Feature engineering

Procedure waarbij een bestaande dataset verrijkt wordt met afgeleide parameters die verkregen worden door modelsimulaties of berekeningen op bestaande parameters. Een eenvoudig voorbeeld is het gebruik maken van de leeftijd van een waterleiding, in plaats van de bestaande parameters: aanlegjaar en huidig jaar.

Meta-informatie

Toelichting op gegevens in een database of een (historische) meetreeks, bijvoorbeeld gebruikte eenheden, beschrijving van een meetopstelling, gebeurtenis of (kalibratie)procedure en begeleidende tekst over metingen.

Machine learning

Onderzoeksveld grenzend aan kunstmatige intelligentie, statistiek en optimalisatie, dat zich bezig houdt met de ontwikkeling van algoritmes en technieken die computers in staat stellen patronen of verbanden te leren en te ontdekken.

Dankwoord

Dit onderzoek is uitgevoerd binnen het kader van het bedrijfstakonderzoek (BTO) van de Nederlandse waterbedrijven. Genoemde casussen zijn mede gefinancierd uit de Toeslag voor Topconsortia voor Kennis en Innovatie (TKI's) van het ministerie van Economische Zaken. De auteurs danken Jan-Maarten Verbree (Nelen & Schuurmans), Bas Wols (KWR), Joost van Summeren (KWR) en betrokken waterbedrijven voor hun bijdrage aan dit onderzoek.

Referenties

1. Vonk, E. & Vries, D. (2015). Datamining voor assetmanagement – inventarisatie en voorbeelden uit de watersector, KWR, BTO 2015.077, Nieuwegein.
2. Thienen, P. van, Pieterse-Quirijns, I., Kater, H. de & Duifhuizen, J. (2012). Nieuwe lekverliesbepalingsmethoden voor het drinkwaterdistributienet. H2O, vol. 8, p. 41-44.
3. Armon, A., Gutner, S., Rosenberg, A., & Scolnicov, H. (2011). Algorithmic network monitoring for a modern water utility: a case study in Jerusalem. *Water Science & Technology*, *63*(2).

4. Beuken, R., Slaats N. & Bont, R. de (2011). "Naar een duurzame balans tussen prestaties, kosten risico's voor waterdistributie." H2O 44.8: 37.
5. Summeren, J. van, Raterman, B., Vonk, E., Blokker, M., Erp, J. van & Vries, D. (2015). Influence of Temperature, Network Diagnostics, and Demographic Factors on Discoloration-Related Customer Reports. Procedia Engineering, 119, 416-425.
6. Vonk, E., Vries, D., Summeren van, J.R.G., Verbree, J.M., Wols, B.A., Raterman, B.W. (2016). Kennis uit waterdata in en rondom het leidingnet, KWR, 2016.006, Nieuwegein
7. Vries, D., Vonk, E., Jong, W. de, Duist, H. van, Marel, H. van der, Wielen, J. van der (2015). Herkennen van anomalieën in waterdata: demo in de Vitens proeftuin, KWR, 2015.108, Nieuwegein.
8. Vries, D., Akker, B. van den, Velickov, S., Jong, W. de & Summeren, J. van, Application of machine learning techniques to predict anomalies in water supply networks (2015). New developments in Water and IT conference, 8-10 februari, Rotterdam.