KWR 2016.006 | April 2016

Kennis uit waterdata in en rondom het leidingnet

Toepassing van machine learning en statistiek op data van Brabant Water

Kennis uit waterdata in en rondom het leidingnet

Toepassing van machine learning en statistiek op data van Brabant Water

KWR 2016.006 | April 2016

Opdrachtnummer
400552

Projectmanager
ing. J.A. (Ton) van Leerdam

Opdrachtgever
TKI

Kwaliteitsborger(s)
Dr. Ir. E.J.M. (Mirjam) Blokker

Auteur(s)
ir. E. (Erwin) Vonk, dr. ir. D. (Dirk) Vries, dr. J.R.G. (Joost) van Summeren, ir. J.M. (Jan-Maarten) Verbree (Nelen en Schuurmans), dr. ir. B.A. (Bas) Wols, drs. B.W. (Bernard) Raterman, ing. R. (Roel) Diemel (Brabant Water) en ing. J. (Johan) van Erp (Brabant Water)

Verzonden aan
Brabant Water, Nelen en Schuurmans, Witteveen+Bos

Dit onderzoek is uitgevoerd in samenwerking met Brabant Water, Nelen en Schuurmans en Witteveen+Bos en mede gefinancierd uit de Toeslag voor Topconsortia voor Kennis en Innovatie (TKI's) van het ministerie van Economische Zaken.

Jaar van publicatie
2016

Meer informatie
dr.ir. D. (Dirk) Vries
T 671
E dirk.vries@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
F +31 (0)30 60 61 165
E info@kwrwater.nl
I www.kwrwater.nl

KWR | April 2016 © KWR

Alle rechten voorbehouden.
Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of enig andere manier, zonder voorafgaande schriftelijke toestemming van de uitgever.

Samenvatting

Bij drinkwatervoorziening wordt door toenemend gebruik van online sensoren, automatisering en informatisering van processen een steeds groter wordende hoeveelheid 'water'-data gegenereerd. Daarnaast is er ook een toenemend aanbod van zogenaamde 'open data': datasets van diverse instanties die publiek toegankelijk zijn. Hierbij kan gedacht worden aan bijvoorbeeld datasets met demografische kenmerken van openbare ruimte, omgevingstemperatuur en grondwaterstanden. Knowledge Discovery in Databases (KDD) is een analysemethodiek waarbij getracht wordt waardevolle kennis uit een combinatie van dergelijke databronnen te destilleren, die vervolgens ingezet kan worden om op operationeel niveau informatievergaring en besluitvorming te verbeteren.

Het doel van dit project is om inzicht te krijgen in de mogelijkheden van datamining en KDD teneinde een betere bedrijfsvoering van drinkwaterlevering te realiseren. Als bronhouder van veel drinkwater-gerelateerde data heeft Brabant Water de ambitie uitgesproken om te komen tot effectievere assetmanagement-tools op basis van alle data die zij reeds genereren vanuit de dagelijkse bedrijfsvoering. Gelet op de specifieke databronnen die voor dit project beschikbaar waren, zijn er twee specifieke onderzoeksvragen geformuleerd:

- Hoe kan KDD worden toegepast op klantmeldingen, gegevens over de distributie-infrastructuur en (sociaal-)demografische gegevens **(casus K)**?
- Hoe kan KDD worden toegepast op storingsregistratie, real-time data uit het procesinformatiesysteem (druk- en volumestroomgegevens) en gegevens van het leidingnet **(casus P)**?

Brabant Water heeft daartoe in consortium met KWR, Nelen en Schuurmans en Witteveen+Bos samengewerkt om data van diverse bronnen slim te combineren en te analyseren met KDD technieken. KWR heeft voor de twee casussen een proof-of-principle van inzet en nut van deze technieken neergezet. Nelen en Schuurmans heeft de data en modelresultaten gevisualiseerd met inzet van hun platform Lizard™. Witteveen+Bos heeft in dit project een adviserende rol gehad.

Per casus is bekeken welke analyse en methodiek het meest geschikt was. Bij casus K stonden klantmeldingen en gegevens over het leidingnet, panden, omgevingstemperatuur en demografie van buurten en wijken centraal. Het beschikbare aantal gegevens en de frequentie (in tijd) van de observaties hebben geleid tot een aanpak met correlatie-analyse en toetsing op significantie. De methodiek heeft een correlatie tussen temperatuur en het aantal meldingen over bruin water aan het licht gebracht. Een tweede ontdekking was de correlatie tussen geluidsmeldingen over watermeters in woningen en het bouwjaar van die woning: een indicatie dat de locatie van een watermeter binnen een woning van doorslaggevend belang kan zijn. Visualisatie van de meldingen in Lizard ondersteunde het onderzoek naar toetsing op gevonden correlaties.

Bij casus P lag het zwaartepunt op real-time data, waarbij datamining is ingezet om nieuwe inzichten in de frequentie van storingen te ontdekken en bestaande hypothesen te verifiëren. De analyse bevestigde het vermoeden dat er een verband is tussen het drukverschil over een dag en de storingsfrequentie van cementhoudende leidingen: hoe groter dit drukverschil, hoe groter het aantal storingen voor dit type leidingen. Voor leidingen van andere materialen was

deze correlatie minder sterk: daar bleek juist bij zeer gangbare drukverschillen (tussen 109 en 129 kPa) de storingsfrequenties hoger liggen.

Algemeen kan geconcludeerd worden dat datagedreven analyses het onderzoek naar fysische processen niet kunnen vervangen, maar dat deze wel een nuttige aanvulling kunnen zijn op dergelijk onderzoek. Zo illustreert casus K dat statistische analyses een eerste screening vormen voor een veelvoud aan mogelijke verklaringen. Nader (fysisch) onderzoek kan vervolgens inzicht geven in de oorzakelijke verbanden tussen temperatuur en bruin water, of watermeters die geluidsoverlast geven en het bouwjaar van het pand waar de watermeter is geplaatst. Casus P laat zien dat met behulp van KDD, bruikbare parameters uit hoogfrequente metingen zijn te extraheren, die vervolgens weer in een multivariate analyse gebruikt kunnen worden om relaties te analyseren. Daarmee is het nut van data-visualisatie en data-analyse met statistische methoden en KDD aangetoond. Het is ook duidelijk geworden dat kennis over het leidingnet zijn waarde opnieuw bewijst, evenals kennis over de wijze waarop informatie wordt verzameld, geaggregeerd en gecombineerd. De stappen hierin zijn niet eenvoudig in een kant-en-klaar recept voor informatieverwerking te vatten. Wel kunnen degelijk datamanagement, het delen van data (over storingen en meldingen tot operationele data) bedrijfstak-breed, als ook data-visualisatie de inzet van KDD en statistiek ondersteunen en de analyse versnellen.

Inhoud

Kennis uit waterdata in en rondom het leidingnet 1

Samenvatting 3

Inhoud 5

1 Datamanagement van de toekomst 6
1.1 Introductie 6
1.2 Projectdoelstellingen, -consortium en aanpak 6
1.3 Leeswijzer 7

2 Casusbeschrijving en databronnen 8
2.1 Casus K 8
2.2 Casus P 11

3 Analysemethodiek 13
3.1 Casus K 13
3.2 Casus P 16

4 Resultaten 20
4.1 Casus K 20
4.2 Casus P 29

5 Visualisatie in Lizard 37
5.1 Lizard 37
5.2 Visualisatie voor drinkwater 37
5.3 Ervaringen opslaan en visualiseren van drinkwatergegevens 40

6 Conclusies en aanbevelingen 41
6.1 Conclusies en aanbevelingen casus K 41
6.2 Conclusies en aanbevelingen casus P 42
6.3 Samenvattende aanbevelingen 43

7 Literatuur 44

Bijlage I Methodiek berekening vermazingsgraad en dimensioneringsmaat 45
Indicatoren leidingnetontwerp 45
Databronnen 46
Uitgangspunten 46
Methodiek 46

Bijlage II Methodiek definiëren drukzones 49

1 Datamanagement van de toekomst

1.1 Introductie

Door verdergaand gebruik van online sensoren, automatisering en informatisering van processen beginnend bij de bron (bijv. puttenbeheer) tot aan de tap (bijv. smart metering) zal de hoeveelheid 'water'-data in de toekomst groter worden. Daarnaast is er een steeds grotere hoeveelheid openbaar beschikbare data (open data) die betrekking heeft tot metingen in de openbare ruimte (temperatuur, (grond)waterstanden) of demografische gegevens.

De urgentie om uit deze datastromen effectiever informatie te destilleren is in een workshop met als onderwerp 'Big waterdata' te KWR begin oktober 2013 uitgesproken door de deelnemende waterbedrijven. Een ontwikkeling die in deze context bij uitstek geschikt kan zijn, is het proces van combineren van input data en vinden van bruikbare informatie of kennis (*Knowledge Discovery in Databases* (KDD)), als ook het trainen en gebruiken van datagedreven modellen door zogenaamde *machine learning*. De succesvolle inzet van deze datagedreven methoden in andere sectoren, zoals bijvoorbeeld de financiële sector, marketing en sociale media, is ook veelbelovend voor de drinkwatersector (Manyika et al. 2015). In het recente BTO rapport over datamining in de drinkwatersector zijn diverse voorbeelden genoemd voor mogelijke toepassingsgebieden (Vonk and Vries 2015).

Met KDD kan op operationeel niveau informatievergaring en besluitvorming rondom waterlevering en integriteit van de waterinfrastructuur ondersteund worden, of met andere woorden een ondersteuning op het gebied van assetmanagement gerealiseerd worden. De inzet van deze technieken kan tevens nieuwe hypothesen en onderzoeksvragen voor nieuw gevonden verbanden opwerpen. Naar aanleiding van de uitkomsten van de workshop is het TKI project DiAMANT gestart.

1.2 Projectdoelstellingen, -consortium en aanpak

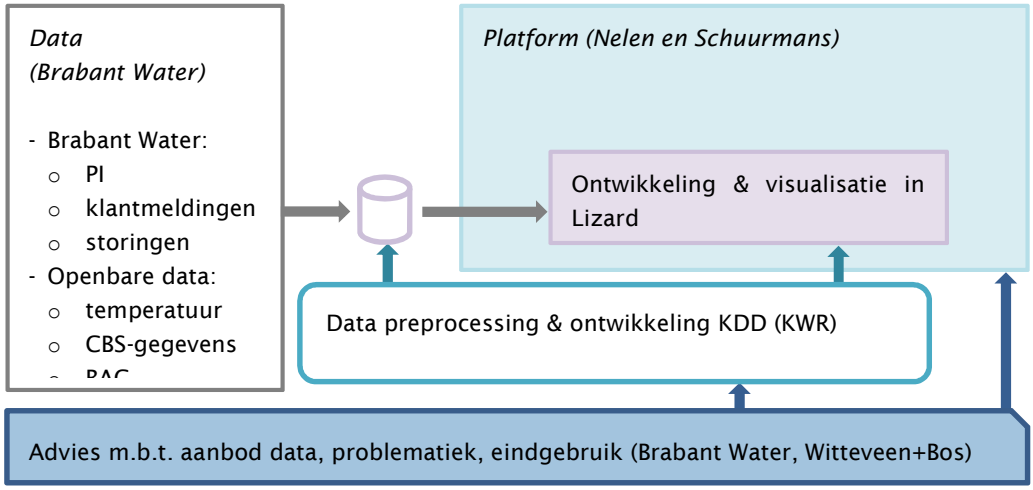
Het doel in dit project is om inzicht te krijgen in de mogelijkheden van datamining en KDD teneinde een betere bedrijfsvoering van drinkwaterlevering te realiseren. Als eindgebruiker heeft Brabant Water bij de uitdaging van datamanagement, de ambitie uitgesproken om te komen tot effectievere assetmanagement-tools met visuele ondersteuning van beheerstaken en daardoor verbeterde dienstverlening en intelligentere bedrijfsvoering in de toekomst. Daartoe zijn voor dit onderzoek twee specifieke onderzoeksvragen geformuleerd:

- Hoe kan KDD worden toegepast op klantmeldingen, gegevens over de distributie-infrastructuur en (sociaal-)demografische gegevens **(casus K)**?
- Hoe kan KDD worden toegepast op storingsregistratie, real-time data uit het procesinformatiesysteem (druk- en volumestroomgegevens) en gegevens van het leidingnet **(casus P)**?

Voor de analyse van de data bij beide casussen is de volgende getrapte aanpak gehanteerd:

(i) onderzoek verbanden en ontwikkel gereedschap, gebruikmakend van KDD en statistiek en met als invoer data in en rondom drinkwaterdistributienetten (o.a. productiemetingen en klantmeldingen);
(ii) breng de informatie samen en visualiseer zodanig dat afwijkingen of correlaties met reeds bekende patronen zichtbaar worden, en
(iii) integreer deze visualisatie in een analysetool die inzicht in de integriteit en prestaties van het leidingnet verschaft.

De rol van de verschillende partners en hun (in-kind) bijdrage in het project is schematisch weergegeven in Figuur 1.



FIGUUR 1. SCHEMATISCHE WEERGAVE VAN INBRENG TECHNOLOGIE EN DATA PER PARTNER. GEBRUIKTE AFKORTINGEN: CBS, CENTRAAL BUREAU VOOR STATISTIEK; BAG, BASISADMINISTRATIE GEMEENTES.

1.3 Leeswijzer

Dit rapport heeft twee sporen, casus K en P, die uiteindelijk samenkomen door een beschouwing in hoofdstuk 6: Conclusies en aanbevelingen. In hoofdstuk 2 worden beide casussen toegelicht en wordt ingegaan op de beschikbare databronnen. Vervolgens wordt in hoofdstuk 3 verder ingegaan op de analysemethodiek, waarna in hoofdstuk 4 de resultaten gepresenteerd worden. Daarna komt in hoofdstuk 5 de visualisatie van de resultaten in het Lizard-platform van Nelen en Schuurmans aan bod. In hoofdstuk 6 worden de conclusies en aanbevelingen per casus, en algemene conclusies gepresenteerd.

2 Casusbeschrijving en databronnen

2.1 Casus K[1]

2.1.1 Doel en vraagstelling

In deze casus worden allerlei data onderzocht om (a) het vóórkomen van bruinwatermeldingen en (b) meldingen van klanten over tikkende watermeters nader te onderzoeken. Hiervoor combineren we een grote verscheidenheid aan data, waaronder 5 jaar aan klantcontactgegevens, demografische gegevens en temperatuurgegevens. De analyse is gericht op een beter begrip van mogelijke invloeden van leidingnetwerkkarakteristieken, temperatuur en demografische factoren, zonder sterke aannames te maken over de processen of afhankelijkheden.

Vanwege de beperkte omvang van beschikbare gegevensbronnen, is hier gekozen voor de inzet van statistische technieken en niet voor de inzet van KDD-technieken. De vraagstelling bij deze casus is als volgt geformuleerd:

> *Wat is de invloed van temperatuur, netwerkkarakteristieken en demografische factoren op klantmeldingen gerelateerd aan bruin water en geluiden afkomstig van watermeters?*

2.1.2 Achtergrond bruinwatermeldingen

Het optreden van bruin water in gedistribueerd drinkwater is een wijd verspreid fenomeen dat wereldwijd optreedt in veel drinkwaterdistributiesystemen. Hoewel bruin water voornamelijk slechts een esthetisch probleem is en ongevaarlijk voor de gezondheid van consumenten, is het vaak reden voor klanten om een melding te maken bij drinkwaterbedrijven. Merk op dat bruinwatermeldingen bij Brabant Water geen groter probleem zijn dan elders in Nederland.

Onderzoek heeft aangetoond dat bruin water wordt veroorzaakt door opwerveling van in het net geaccumuleerde deeltjes die voort kunnen komen uit een combinatie van bronnen, waaronder de waterzuivering, roestvorming in leidingen en biofilmmateriaal op de leidingwand (Vreeburg 2007). Het vóórkomen van een bruinwatermelding hangt af van ten minste 4 condities: 1) deeltjesmateriaal accumuleert in de distributieleidingen, 2) opwerveling: geaccumuleerd deeltjesmateriaal wordt in suspensie gebracht door hydraulische krachten, 3) een klant neemt het resulterende bruinwaterincident waar en 4) de klant besluit het incident te melden bij het waterbedrijf.

De invloed van hydraulische omstandigheden en leidingnetwerkontwerp op het bruinwaterrisico is aangetoond in eerder onderzoek. Traditioneel werd bij het netwerkontwerp rekening gehouden met een overcapaciteit (bijvoorbeeld om tegemoet te komen aan de bluswatervraag van de brandweer), wat resulteert in lage stroomsnelheden en een hoge potentie van deeltjesaccumulatie in leidingen (Vreeburg 2007). Dit inzicht heeft geleid tot het ontwerp en succesvolle implementatie van vertakte, zelfreinigende netwerken, waarin hoge

[1] Het materiaal over bruinwatermeldingen dat wordt gepresenteerd in casus K is gepubliceerd in Van Summeren et al. (2015)

snelheden in één richting dagelijks deeltjes uit de leidingen spoelen. Dit voorkomt hoge accumulatieniveaus en verlaagt het bruinwaterrisico (Blokker et al. 2010).

Recent onderzoek toont aan dat bruinwatermeldingen vaker voorkomen bij hogere temperaturen en dat dit niet gerelateerd is aan het optreden van leidingbreuken of watervraagpatronen (Cook et al. 2001). Een systematisch onderzoek in het drinkwater-distributienet van PWN toont geen bewijs voor een toename van deeltjesmateriaal dat het distributienet binnenkomt vanaf de zuivering bij hogere temperaturen en suggereert in plaats daarvan een invloed van een proces dat versterkt wordt door hogere temperaturen in het distributienet (Blokker and Schaap 2015).

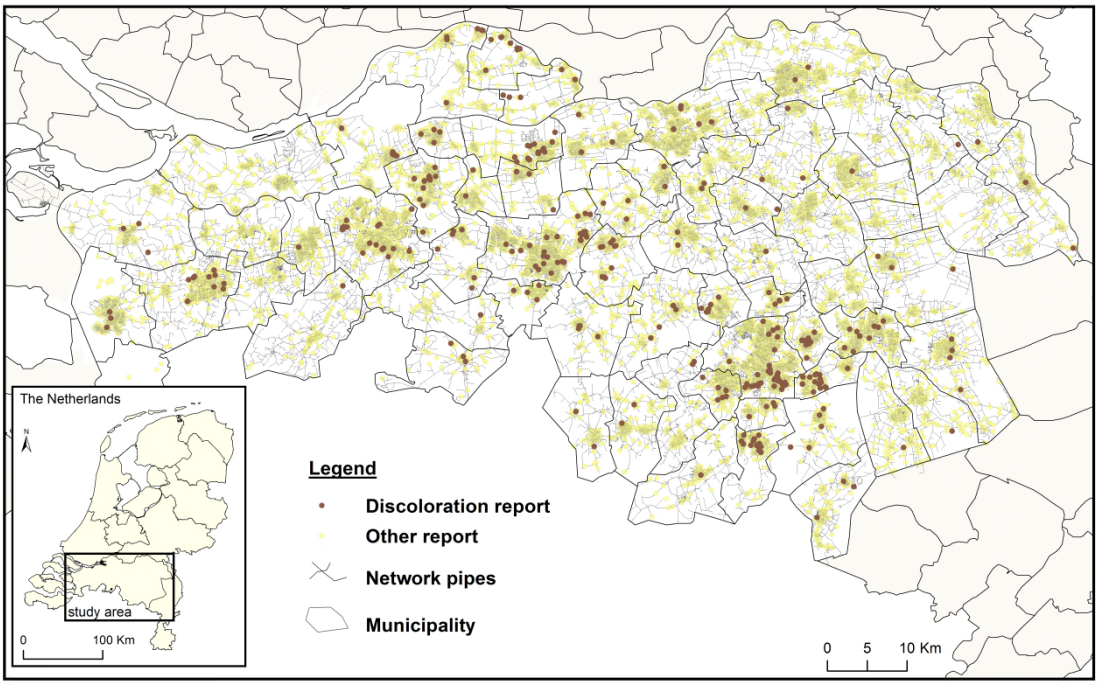
Omdat continue troebelheidmetingen niet algemeen beschikbaar zijn, is onderzoek naar bruin water vaak afhankelijk van klantcontactgegevens. Helaas kunnen klantmeldingen een verkeerde voorstelling geven van netwerkprestaties, vanwege inherent demografische invloeden die het beeld mogelijk vervormen, zoals de levensstijl van klanten en aanwezigheid in huis. Brabant Water en KWR hechten waarde aan het begrijpen van de relevante achterliggende processen bij de bruinwatermeldingen. Met meer inzichten in deze processen tracht Brabant Water het aantal bruinwatermeldingen verder te kunnen reduceren. In dit onderzoek wordt getracht om informatie met betrekking tot bruinwater te destilleren uit data en klantmeldingen.

2.1.3 Achtergrond klantmeldingen van geluiden afkomstig van watermeters

Brabant Water heeft de afgelopen jaren slimme meters (V200 volumemeters) geïnstalleerd bij huisaansluitingen. Een aantal van deze meters blijken een tikkend geluid te produceren waarover klanten meldingen maken. De oorzaak van de tikkende meters is niet bekend. In dit deelonderzoek wordt gekeken of met analyse van de beschikbare data een verband kan worden gevonden met omgevingsdata, waaronder informatie over panden afkomstig van de gemeentelijke basisadministratie gemeentes (BAG).

2.1.4 Beschikbare databronnen

De uitgevoerde analyse is toegepast op het hele voorzieningsgebied van Brabant Water, waarin 2,4 miljoen klanten worden bediend (Figuur 2).



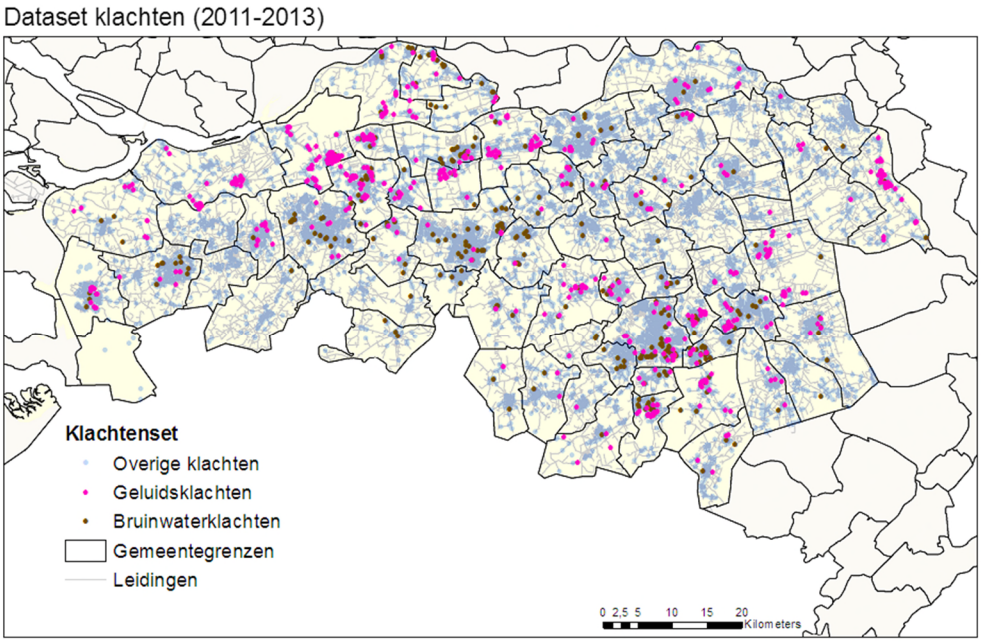
FIGUUR 2. VOORZIENINGSGEBIED VAN BRABANT WATER MET NETWERKLEIDINGEN (GRIJS), BRUINWATER-GERELATEERDE KLANTMELDINGEN (BRUINE PUNTEN), ALGEMENE MELDINGEN ("REFERENTIEMELDINGEN", LICHTGELE PUNTEN) EN GEMEENTEGRENZEN (ZWART).

Het distributienetwerk bevat zowel vermaasde als vertakte, zelfreinigende netwerksecties. Het totale netwerk bestaat uit verschillende distributiesystemen en de 15 bijbehorende InfoWorks netwerkmodellen zijn samengevoegd en geëxporteerd als één enkel model voor geostatistische databewerkingen in ArcGIS. In de statistische analyse hebben we de gegevens onderverdeeld in wijken (zwarte lijnen in Figuur 2 tonen de grovere onderverdeling met gemeentegrenzen). De buurtgrenzen zijn bepaald door het Centraal Bureau voor Statistiek (CBS, "CBSBuurten2011"). Dit is het meest gedetailleerde niveau waarop de demografische data is onder te verdelen. Buurtgrenzen geven een natuurlijke onderverdeling van zowel het distributienetwerk als de demografische data.

TABEL 1. OVERZICHT BESCHIKBARE DATASETS VOOR CASUS K

Data	Bron	Ruimtelijk?	Meetfrequentie	Databeheer/ software
Leidingnet	Brabant Water	ja	n.v.t.	InfoWorks
Klantmeldingen	Brabant Water	ja	onregelmatig	Software klantcontacten
Buurtinformatie (demografie, bouwjaar woningen, etc.)	CBS	ja	jaarlijks	BAG (basisadministratie gemeentes)
Meetreeksen omgevingstemperatuur (lucht)	KNMI	ja	dagelijks	KNMI

Voor de analyse van meldingen over geluiden van volumemeters is gebruik gemaakt van een dataset van 1731 vervangen meters (Figuur 3). Daarnaast is een dataset van 651 geluid gerelateerde meldingen geëxtraheerd uit een totale set van klantmeldingen in de periode 2010-2014. Gegevens van bouwjaar van woningen/panden zijn afkomstig uit de Basis Administratie Gemeentes (BAG).



FIGUUR 3. OVERZICHT VAN GELUIDSKLACHTEN (PAARS) IN HET VOORZIENINGSGEBIED VAN BRABANT WATER.

2.2 Casus P

2.2.1 Doel en vraagstelling

Bij deze casus is de relatie tussen data uit het procesinformatiesysteem bij verschillende meetlocaties in het distributiegebied met gegevens van het leidingnet onderzocht. Procesdata omvat druk- en volumestroomgegevens en leidingnetdata omvat storingen en leidingnetkarakteristieken. Specifiek is de volgende vraagstelling geformuleerd:

> *Is er een relatie tussen het aantal storingen per eenheid leidinglengte en het gemeten, dynamische, drukregime van een toeleverend pompstation?*

Met de onderzoeksresultaten wordt voorzien dat het vervangingsvraagstuk bij Brabant Water van leidingen beter kan worden beantwoord. De verwachting is dat met een beter inzicht over de integriteit van leidingen, investeringen met grotere zekerheid kunnen worden geprioriteerd.

2.2.2 Achtergrond

Brabant Water is actief op zoek naar factoren die invloed hebben op de conditiedegradatie alsmede de storingsfrequentie van leidingen in het distributienet. Daarbij is het einddoel om op basis van dergelijke factoren te komen tot een effectiever saneringsbeleid, dat prioriteit geeft aan leidingen die de grootste kans hebben om te storen.

Uit recent onderzoek van KWR blijkt dat de storingsfrequentie van leidingen verband houdt met het drukregime van pompstations (Wols et al. 2016). Dit verband is aangetoond voor de drukverdeling die hoort bij de 'criteriumdag'. De criteriumdag is een 'benchmark' verbruiksdag bij Brabant Water, gebaseerd op het verbruikspatroon op het 80% percentiel van alle verbruiksdagen in een historisch jaar. Een logische vervolgstap is daarmee om het verband tussen storingsfrequentie en drukregime ook te onderzoeken voor daadwerkelijk gemeten drukken bij diverse pompstations (zodat ook wisselingen in de tijd en ruimtelijke effecten meegenomen kunnen worden).

2.2.3 Beschikbare databronnen

Voor deze analyse is gebruik gemaakt van de volgende datasets:

- **Meetreeksen druk en volumestroom (PI)**
  Dit zijn de meetgegevens per uitgaande reinwatertak in de pompstations. De data is voorhanden over de periode 19 juni 2014 tot en met 31 maart 2015.
- **Lijst met storingen (USTORE)**
  Deze dataset bevat de volgende kolommen: melddatum, materiaal, diameter, jaar van aanleg, locatie van storing (RD-coördinaten), etcetera. De lijst van storingen komt uit USTORE. Daarbij zijn de oorzaak derden en foutieve aanleg verwijderd. Alleen storingen in dezelfde periode als druk en volumestroom metingen zijn meegenomen in de verder analyse.
- **Lijst met alle leidingen (Infoworks)**
  Deze dataset bevat de volgende kolommen: materiaal, diameter, jaar van aanleg, locatie van leiding (RD-coördinaten).

Een overzicht van de databronnen weergegeven in Tabel 2.

TABEL 2. OVERZICHT DATABRONNEN EN BIJBEHORENDE KARAKTERISTIEKEN VOOR CASUS P

Data	Bron	Ruimtelijk?	Meetfrequentie	Databeheer/ software
Leidingnet	Brabant Water	ja	eenmalig	InfoWorks
Meetreeksen druk	Brabant Water	Ja	1/minuut	OSISoft PI
Meetreeksen volumestroom	Brabant Water	Ja	2/minuut	OSISoft PI
Storingen	Brabant Water	Ja	onregelmatig	USTORE
Begrenzing voorzieningsgebieden	Brabant Water	Ja	eenmalig	ArcGIS
Locaties WPB's, aanjagers en opjagers	Brabant Water	Ja	eenmalig	ArcGIS
Vertaaltabel sensorcoderingen-leidingtakken	Brabant Water	Ja	eenmalig	Excel

3 Analysemethodiek

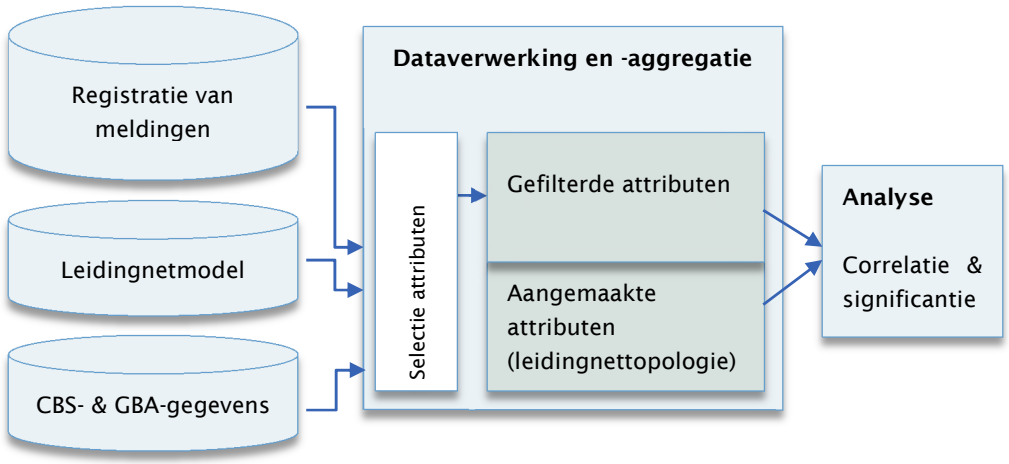
3.1 Casus K

3.1.1 Methodiek op hoofdlijnen

De databronnen, benoemd in het vorige hoofdstuk, zijn gekoppeld, gecombineerd en geanalyseerd. De hoofdstappen in deze methodiek zijn de voorbewerking en selectie van data, het kwantificeren van de leidingnettopologie en de uitvoering van statistische analyses. Dit is schematisch weergegeven in Figuur 4. Er zijn twee analysemethodieken toegepast:

1. Het uitvoeren van een correlatieanalyse van demografische gegevens, leidingnetgegevens en klantmeldingen met betrekking tot bruinwaterincidenten en algemene klantmeldingen;
2. Uitvoering van significantietesten tussen (enkelvoudige) relaties van temperatuur, leidingdiameter en leidingmateriaal met meldingen over vuilwaterlast.

De stappen bij dataverwerking en analyse zijn in de navolgende paragrafen verder toegelicht.



FIGUUR 4. VAN DATABRON NAAR ANALYSE VAN RELATIES TUSSEN ATTRIBUTEN IN CASUS K

3.1.2 Voorbewerking en selecties uit beschikbare datasets

De (totale) set van klantmeldingsgegevens beslaat een periode van 5 jaar (januari 2010-december 2014) en bevat 13.848 meldingen. Hiervan zijn 306 meldingen geïdentificeerd als bruinwater-gerelateerd. In totaal 651 meldingen zijn geïdentificeerd als gerelateerd aan geluiden van watermeters. De selectie is uitgevoerd door te zoeken op een combinatie van steekwoorden in de beschrijving van meldingen, waarna ter controle nog een handmatige selectie is uitgevoerd. De overige meldingen vormen de basis voor een referentiedatabase. Deze referentieset omvat meldingen die betrekking hebben op administratieve zaken en onderbrekingen in waterlevering en druk. Meldingen die specifiek refereren aan spui-acties zijn verwijderd uit de referentieset (ook met een zoekactie op kernwoorden), omdat ze het beeld van spontaan ontstane bruinwaterincidenten mogelijk vervormen. De referentieset bevat 12.823 meldingen.

Door het KNMI gemeten daggemiddelde buitentemperaturen hebben we gekoppeld aan individuele klantmeldingen, gebaseerd op de dag van de melding en kleinste afstand tot één van de drie weerstations Eindhoven, Gilze-Rijen en Volkel. Er is gebruikt gemaakt van dagtemperaturen omdat continu gemeten temperatuurmetingen van het drinkwater niet beschikbaar zijn. Het is redelijk om aan te nemen dat (1) de bodemtemperatuur een dominante invloed heeft op de daggemiddelde drinkwatertemperatuur (Blokker and Pieterse-Quirijns 2013) en (2) de bodemtemperatuur sterk gerelateerd is aan de daggemiddelde buitentemperatuur.

De gebruikte demografische data is afkomstig van het CBS. Data zijn beschikbaar voor 1488 van de 1594 buurten in Noord-Brabant. De overige 106 buurten zijn uitgesloten (door het CBS) omdat het aantal inwoners lager is dan 50. We hebben de volgende 14 parameters gebruikt in de analyse: percentage personen in 5 leeftijdscategorieën (0-14, 15-24, 25-44, 45-64, >65 jaar), percentage ongehuwden, percentage gehuwden, percentage eenpersoonshuishoudens, percentage huishoudens zonder kinderen, percentage huishouden met kinderen, grootte van huishoudens, percentage gescheiden personen, percentage verweduwde personen en bevolkingsdichtheid.

3.1.3 Kwantificeren leidingnettopologie

Om het leidingnetwerkontwerp te karakteriseren hebben we twee maten gedefinieerd en bepaald op basis van netwerk(model)karakteristieken op buurtniveau:

- De vermazingsgraad (G_{maas}): Een indicator voor de vertakte of vermaasde netwerken, berekend als de verhouding van het aantal mazen (*M*) ten opzichte van het aantal aansluitingen (*A*):

$$G_{maas} = \frac{M}{A} = \frac{L - K + S}{A}$$

  met *L* het aantal leidingen, *K* het aantal knooppunten, *S* het aantal niet-verbonden sub netwerken in een gebied.

- De netwerkdimensioneringsgraad (G_{dim}): Een indicator voor de totale verbruikslast op het distributienet, gedefinieerd als de verhouding van het totale volume van distributieleidingen (*V*) tot het gemiddelde uurlijks klantverbruik (*Q*) in een gebied. Een lage waarde van G_{dim} is indicatief voor een relatief klein distributievolume voor een gegeven watervraag, wat relatief hoge gemiddelde stroomsnelheden en korte verblijftijden bevordert. We gebruiken de volgende formule:

$$G_{dim} = \frac{V}{Q} = \frac{\pi \sum_{i=1}^{n} D_i^{\,2}}{4 \sum_{j}^{m} Q_j}$$

  met D_i en L_i de diameter en lengte in meters van een distributieleiding *i*. De som in de teller is over *n* netwerkleidingen. De totale watervraag in een buurt (in m^3 per jaar) is gebaseerd op de som van de vraag van *m* klanten. Het totaal van de geadministreerde watervraag kan licht afwijken van de hoeveelheid geproduceerd water in een distributiegebied. Dergelijke verschillen worden veroorzaakt door bijvoorbeeld administratieve fouten, leidinglekken, en meetfouten en worden in rekening gebracht met een kleine correctie in de klantverbruiken. De verbruiken zijn van 2011 en worden representatief geacht voor de analyseperiode (2010-2014). We verwaarlozen daarmee de kleine mogelijke variaties in buurtverbruiken en netwerkvolumes tijdens de analyseperiode.

3.1.4 Statistische analyse op niveau van klantadressen en buurten

Sommige parameters kunnen rechtstreeks worden gekoppeld aan klantadressen (daggemiddelde temperatuur en leidinglengte, -materiaal en –diameter). We hebben leidingeigenschappen toegekend van de leiding die geografisch het dichtst bij het klantadres ligt, c.q. de daggemiddelde temperatuur van het dichtstbijzijnde weerstation. Vervolgens hebben we distributies van deze parameters berekend voor de bruinwater-gerelateerde en referentie-klantmeldingen en onderzocht of verschillen in de gemiddeldes van de distributies statistisch significant zijn, op basis van p-waarden van een statistische t-toets.

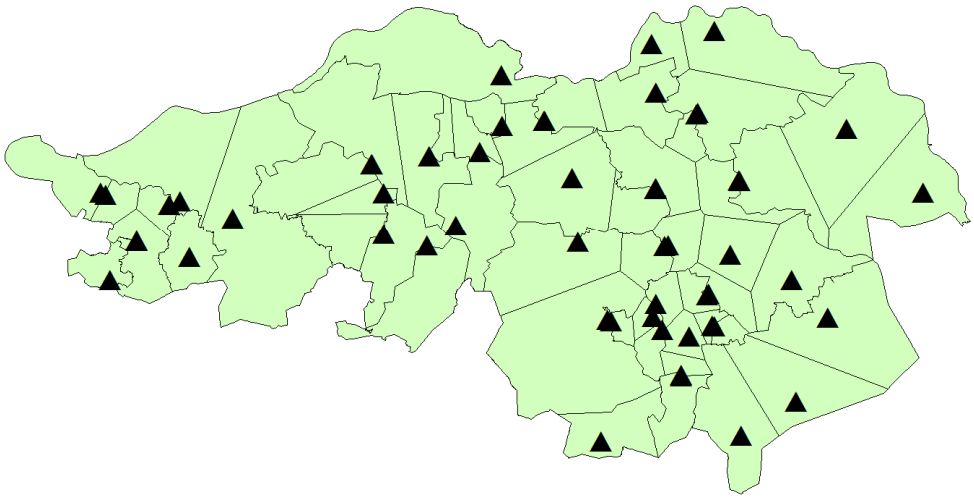
Andere parameters zijn enkel beschikbaar op buurtniveau (de 14 demografische factoren) of hebben weinig zeggingskracht wanneer ze niet worden geaggregeerd (netwerkkarakteristieken G_{maas} en G_{dim}). Voor deze parameters hebben we mogelijke afhankelijkheden onderzocht door correlaties te berekenen tussen de parameters en de klantmeldingsfrequenties per buurt. De meldingsfrequenties zijn berekend door de som van meldingen over de periode 2010-2014 in een buurt te delen op het aantal aansluitingen in diezelfde buurt.

3.2 Casus P

3.2.1 Methodiek

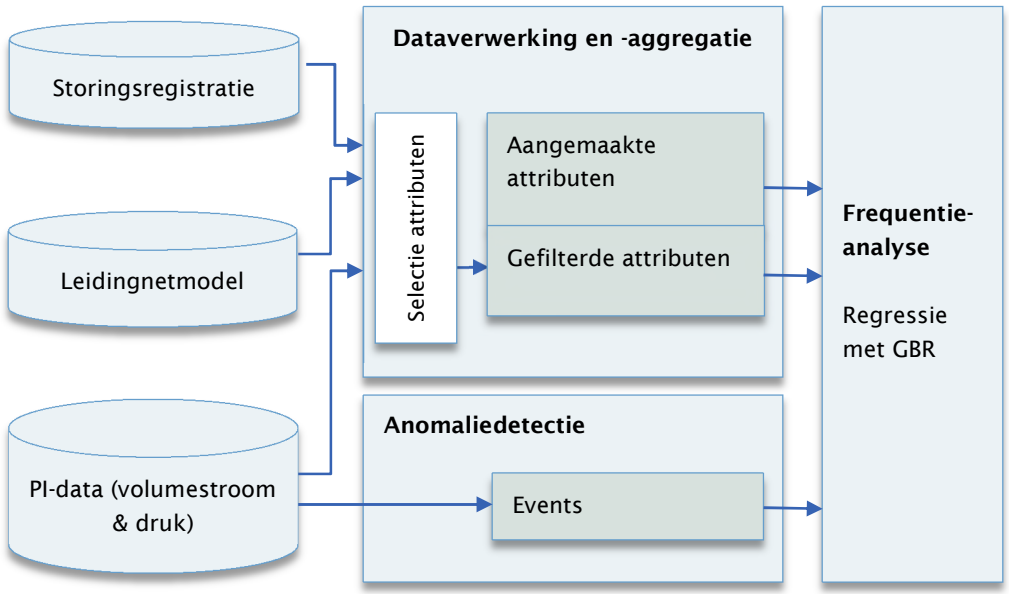
Er worden door Brabant Water automatische drukmetingen gedaan op verschillende locaties, zoals uitgaande reinwatertakken van Water Productie Bedrijven (WPB's), transportleidingen, aanjagers, opjagers en op een aantal plekken in het leidingnet. Voor dit onderzoek waren alleen drukmetingen bij WPB's en op-/aanjagers beschikbaar.

Om een relatie te leggen tussen drukschommelingen zoals gemeten op dit beperkte aantal puntlocaties en het optreden van storingen in een geografisch groot gebied, moet een belangrijke aanname over de verdeling van druk worden gemaakt. We hebben aangenomen dat elke meetlocatie met een pompstation een 'drukzone' bewerkstelligt: dat wil zeggen een gebied met eenzelfde, heersende druk. Middels deze drukzones kunnen dan de drukkarakteristieken zoals gemeten op een zekere meetlocatie gekoppeld worden aan storingen in het omringende gebied. De gehanteerde methodiek voor het definiëren van drukzones op basis van meetlocaties en voorzieningsgebieden is omschreven in Bijlage II. Het resultaat is een indeling van Brabant Water in 55 drukzones (Figuur 5).



FIGUUR 5. INDELING DRUKZONES IN DE PROVINCIE NOORD-BRABANT. MEETLOCATIES DRUK ZIJN WEERGEGEVEN MET ZWARTE DRIEHOEK.

Per drukzone zijn (a) meetreeksen voor volumestroom en druk opgevraagd uit de PI-database, (b) storingen opgehaald uit USTORE en (c) relevante leidingnetkarakteristieken opgehaald uit Infoworks exports. Drukreeksen zijn daarna geaggregeerd tot de 5-minuteninterval, om vervolgens direct gebruikt te worden als invoer voor de uiteindelijke analyse naar relaties tussen drukregime en storingen. In een parallel spoor zijn de drukreeksen en volumestroomreeksen gebruikt om middels een anomaliedetectiealgoritme geautomatiseerd zogenaamde 'events' te genereren. Een event is daarbij gedefinieerd als een moment waarop in de meetreeksen een afwijkend patroon te zien is. Een dergelijke afwijking kan veroorzaakt worden door bijvoorbeeld een leidingbreuk, verandering in de pompaansturing, openstaande brandkraan, bijzondere (feest-)dag, etcetera. De zogenaamde 'events' per drukzone zijn, samen met de geaggregeerde drukreeksen, gebruikt als invoer voor de analyse naar correlaties tussen drukregime en storingsfrequentie. Deze methodiek op hoofdlijnen is schematisch weergegeven in Figuur 6.



FIGUUR 6. VAN DATABRON NAAR ANALYSE VAN RELATIES TUSSEN ATTRIBUTEN IN CASUS P

3.2.2 Karakteriseren drukregime met directe parameters

De volumestroom- en drukmeetreeksen voor elke drukzone zijn geaggregeerd op dagniveau, waarbij het gemiddelde, minimum en maximum van druk en volumestroom op de dag is bepaald. Daarnaast is voor de druk ook het maximale drukverschil op de dag bepaald (verschil tussen hoogste en laagste druk op een dag).

Alleen de storingen die binnen het tijdsinterval vallen waar druk- en volumestroommetingen voorhanden waren zijn gebruikt. Ook zijn met behulp van de druk- en volumestroomgegevens de drukken op de dag van storen en de twee dagen voorafgaand aan de storing bepaald. Voor de gemiddelde druk is hier het gemiddelde genomen, voor de maximale druk op een dag de maximale waarden over deze 3 dagen, voor de minimale druk de minimum waarde over de 3 dagen en voor het drukverschil de maximum waarde over de drie dagen. Daarnaast zijn het aantal anomalieën voor zowel volumestroom als druk opgeteld die op de dag van storen of de twee dagen hiervoor zijn opgetreden.

3.2.3 Events bepalen met anomaliedetectie

Er is een automatische anomaliedetectie uitgevoerd op alle druk- en volumestroommetingen. Hiervoor is gebruik gemaakt van een tijdreeks-regressiemodel, dat voor elk tijdstip een voorspelling doet voor de te verwachten meetwaarde op basis van type dag (weekdag, zaterdag, zondag) en op basis van het tijdstip van de dag. Vervolgens wordt de daadwerkelijke meetwaarde vergeleken met de modelvoorspelling, waarbij de mate van verschil gebruikt wordt om te bepalen of de meetreeks onverklaarbaar gedrag vertoont of juist in lijn ligt met wat verwacht mag worden. Voor deze methodiek is gebruik gemaakt van een zogenaamd Support Vector Regressiemodel (SVR), volgens de Scikit-learn LibSVM implementatie in Python (Pedregosa et al. 2011). Anomaliedetectie wordt gedaan binnen een ‘moving window’ van 20 tijdstappen, waarbij een patroon van meer dan 10 opeenvolgende afwijkingen van meer dan drie standaardafwijkingen ten opzichte van de voorspelling wordt aangemerkt als een anomalie. Deze methodiek is grotendeels gebaseerd op de methodiek als beschreven in Mounce et al. (2011). Omwille van rekentijd en voor een robuustere afschatting van het verbruik maken we in dit onderzoek echter gebruik van één regressiemodel, waar voorgenoemde onderzoekers werken met aparte modellen voor elk tijdstip van de dag.

De ‘events’ die door het algoritme als uitvoer worden gegeven worden getypeerd door een zekere duur (aantal tijdstappen dat het afwijkende gedrag gedetecteerd is), een ‘surprise score’ (de ernst van de afwijking, uitgedrukt in het aantal standaardafwijkingen ten opzichte van de modelvoorspelling voor iedere tijdstap, gesommeerd over het aantal afwijkende tijdstappen). Deze uitkomsten van deze anomaliedetectie zijn gebruikt als input voor de uiteindelijke regressie-analyse voor het verklaren van de storingsfrequentie (Figuur 6). Daartoe is voor iedere dag bepaald hoeveel druk- en volumestroomanomalieën opgetreden zijn en wat de som is van de duur en de surprise-score. Als er geen anomalieën zijn opgetreden is deze waarde nul, als er geen data beschikbaar was voor de gegeven locatie is de waarde NaN (Not-a-Number). Een uitgebreidere beschrijving van de methodiek is beschreven in het TKI-rapport KWR 2016.024 (Vries et al., 2016).

3.2.4 Frequentieanalyse

Diverse statistische experimenten zijn uitgevoerd om het verband tussen storingen en druk/volumestroom te bepalen. Voor al deze analyses is gebruik gemaakt van Gradient Boosting Regression (Prettenhofer and Louppe 2014) om de relatie tussen de genoemde attributen en de storingsfrequentie te bepalen. Meer informatie over deze techniek is beschikbaar in het BTO-rapport over datamining voor assetmanagement (Vonk and Vries 2015). De techniek heeft als voordelen dat deze inzetbaar is voor heterogene data (met informatie op verschillende schalen gemeten), en het op robuuste wijze niet-lineariteiten kan beschrijven. Nadelen zijn er ook: de techniek vereist enige optimalisatie van meta-parameters en resultaten kunnen niet geëxtrapoleerd worden.

1. **Analyse op basis van drukzone**
   Voor iedere drukzone is een storingsfrequentie berekend door het aantal storingen te delen door de totale leidinglengte in deze drukzone en de tijdsperiode.
   Vervolgens zijn voor iedere drukzone nog de ‘directe parameters’ en aanvullende attributen toegevoegd om te onderzoeken of deze verband houden met de storingsfrequentie. Dit zijn:

   - Aanwezigheid van materiaalsoorten, diameterklassen en aanlegjaarklassen. Voor iedere klasse is de fractie bepaald ten opzichte van het aantal storingen die geldt voor deze klasse. Bijvoorbeeld mat_CH laat de fractie zien van het aantal storingen van cementhoudende leidingen t.o.v. het totaal aantal in deze drukzone.
   - Karakteristieken van druk- en volumestroom: de gemiddelde druk, drukverschil en volumestroom per locatie.
   - Aantal anomalieën, duur anomalieën en ‘surprise score’ van druk en volumestroom per drukzone.

2. **Analyse op basis van variaties over de tijd**
   Hier zijn het aantal storingen over alle drukzones gezamenlijk beschouwd en is de verandering in de tijd onderzocht. Op weekbasis zijn de volgende attributen bepaald:

   - Totaal aantal storingen per week.
   - Totaal aantal anomalieën voor volumestroom en druk, evenals de duur en surprise scores, per week.
   - Gemiddelde druk en volumestroom per week

3. **Analyse op basis van een druk- of volumestroomklasse**
   In deze analyse wordt gebruik gemaakt van de druk- en volumestroomdata op de dag van storen. Hierbij worden een aantal klassen van druk en volumestroom gedefinieerd,

en wordt gekeken hoeveel storingen, hoeveel leidinglengte en gedurende welke tijdsduur een druk of volumestroomklasse optreedt. De klassen zijn zo gekozen dat in iedere klasse evenveel storingen vallen. Voor iedere klasse wordt vervolgens een storingsfrequentie berekend, gebruik makend van de gegevens van alle drukzones. De uitvoering is als volgt:

a. Voor iedere drukzone en drukklasse wordt het aantal storingen opgeteld.
b. Voor iedere drukzone en drukklasse wordt het aantal dagen waarop een bepaalde drukklasse optrad opgeteld.
c. Voor iedere drukzone wordt de totale leidinglengte berekend.
d. Vervolgens wordt voor iedere drukzone en drukklasse de lengte vermenigvuldigd met de tijdsduur (de lengte per drukklasse zal niet verschillen, maar de tijdsduur wel).
e. De storingsfrequentie per drukklasse wordt bepaald door het aantal storingen op te tellen over ieder drukzone en deze te delen door de som van de lengte maal tijdsduur per locatie.

De regressie geeft als output (o.a.) de afhankelijkheden tussen data-attibuten. Het verband tussen een bepaalde drukklasse of volumestroomklasse en de storingsfrequentie wordt daarmee inzichtelijk. Dezelfde analyse is uitgevoerd voor het aantal anomalieën voor volumestroom en druk die gevonden zijn op de dag van de storing.

Voor elk van deze drie experimenten kan van tevoren een specifieke analyse gemaakt worden, van bijvoorbeeld leidingen van een bepaalde materiaalsoort. Vanwege de beperkte omvang van het aantal storingen per type of subklasse is ervoor gekozen om de analyse te verrichten op alle leidingen, alleen CH-leidingen (cementhoudende leidingen) en ‘alles behalve CH’-leidingen.

4 Resultaten

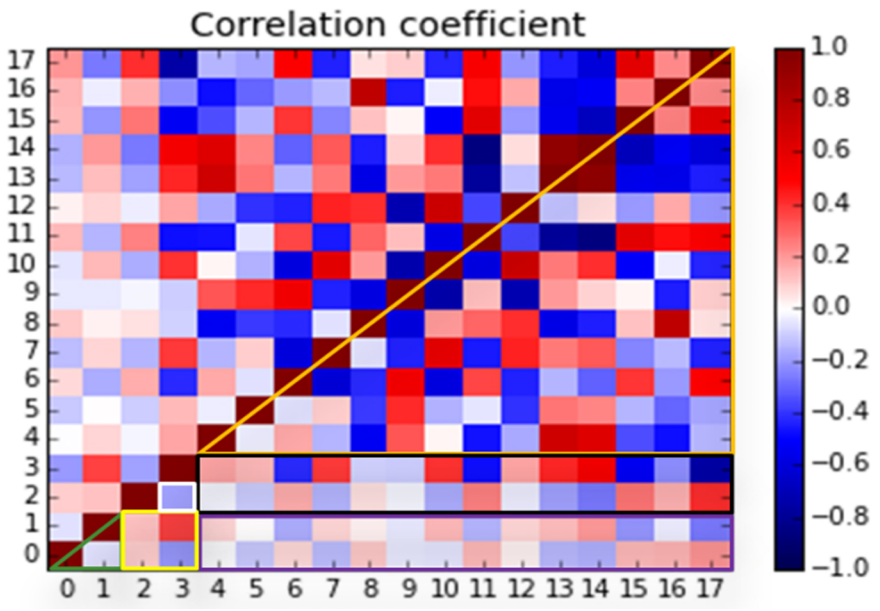
4.1 Casus K

4.1.1 Correlatieanalyse klantmeldingen en omgevingsfactoren

Om mogelijke relaties te onderzoeken tussen geaggregeerde parameters op buurtniveau, hebben we correlaties berekend tussen klantmeldingen (bruinwater-gerelateerd en referentie), netwerkkarakteristieken (G_{maas} en G_{dim}) en de 14 demografische factoren benoemd in paragraaf 3.1.2. We gebruiken een maat voor de afhankelijkheid van twee parameters, namelijk Spearman's rank correlatie ρ, die afhankelijkheden beschrijft waarbij een monotone opgaande of neergaande trend (correlatie) kan worden aangetoond.

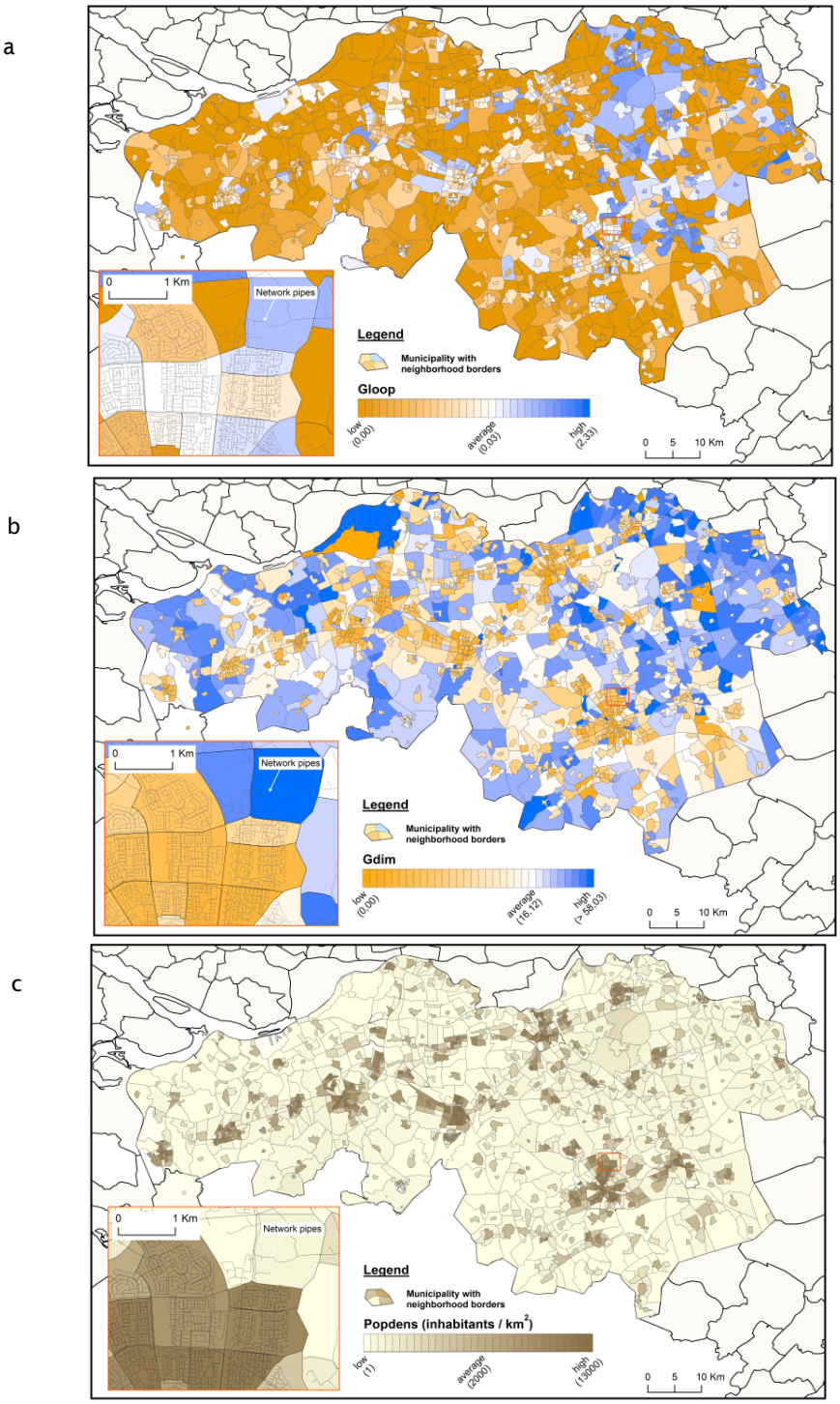
Figuur 7 toont de resultaten van de correlatieanalyse. De resultaten laten een erg zwakke correlatie zien tussen demografische factoren enerzijds en klantmeldingen anderzijds (paarse rechthoek in Figuur 7). Dit geldt zowel voor de bruinwater-gerelateerde klachten ($|\rho|<0,21$) als voor de algemene klachten (referentieset, $|\rho|<0,26$). Dit demonstreert dat de gebruikte klantgegevens slechts een zwakke demografische vertekening kennen. De twee set klantgegevens zijn bovendien grotendeels onafhankelijk, wat wordt weerspiegelt door hun onderling erg zwakke correlatie ($\rho=0,06$, groene driehoek in Figuur 7). Een groot aantal demografische factoren zijn onderling sterk gecorreleerd (oranje driehoek in Figuur 7) en dit weerspiegelt plausibele demografische verbanden, zoals de sterke correlatie tussen getrouwde personen en de leeftijdscategorie 44-65 jaar ($\rho=0,61$).

Netwerkkarakteristiek G_{maas} is slechts zwak gecorreleerd met zowel de bruinwater- als de referentiemeldingen (correlatiecoëfficiënten van respectievelijk $\rho_{02}=0,09$ en $\rho_{12}=0,12$, gele rechthoek in Figuur 7). Dit suggereert dat er geen significante invloed is van vermazing/vertakking van het netwerk op bruinwatermeldingen. Ook is er geen sterke correlatie tussen bruinwater-gerelateerde meldingen en G_{dim}: de correlatiecoëfficiënt is licht negatief ($\rho_{03}= -0,19$).



FIGUUR 7. OVERZICHT VAN CORRELATIECOËFFICIËNTEN. ELK ELEMENT IN DEZE SYMMETRISCHE MATRIX GEEFT EEN COËFFICIËNT WEER VOOR DE CORRELATIE TUSSEN TWEE PARAMETERS ZOALS AANGEGEVEN DOOR DE HORIZONTALE EN VERTICALE ASSEN. EEN HOGE WAARDE (ROOD, DICHT BIJ 1) BETEKENT EEN STERK POSITIEVE CORRELATIE, EEN LAGE WAARDE (BLAUW, DICHT BIJ -1) EEN STERK NEGATIEF. PARAMETERS MET EEN WAARDE 0 ZIJN NIET GECORRELEERD. DE INDICES OP DE ASSEN CORRESPONDEREN MET DE VOLGENDE PARAMETERS: (0) FREQUENTIE VAN BRUINWATER-GERELATEERDE KLANTMELDINGEN, (1) FREQUENTIE VAN ALGEMENE KLANTMELDINGEN (REFERENTIESET), (2) GMAAS, (3) GDIM, (4) PERCENTAGE VAN PERSONEN IN DE LEEFTIJDSCATEGORIE 0-14, (5) 15-24, (6) 25-44, (7) 45-64, EN (8) >65 JAAR, (9) PERCENTAGE ONGETROUWDEN, (10) PERCENTAGE GETROUWDEN, (11) PERCENTAGE 1-PERSOONSHUISHOUDENS, (12) PERCENTAGE HUISHOUDENS ZONDER KINDEREN, (13) PERCENTAGE HUISHOUDENS MET KINDEREN, (14) GROOTTE VAN HUISHOUDENS, (15) PERCENTAGE GESCHEIDEN PERSONEN, (16) PERCENTAGE VERWEDUWDE PERSONEN, (17) BEVOLKINGSDICHTHEID. DE COËFFICIËNTEN VAN DE DIAGONAALELEMENTEN ZIJN PER DEFINITIE GELIJK AAN 1. DE GEKLEURDE RECHTHOEKEN EN DRIEHOEKEN ZIJN UITGELEGD IN DE TEKST.

Figuur 8 toont hoe G_{maas} en G_{dim} en de bevolkingsdichtheid geografisch zijn verdeeld over het voedingsgebied. Te zien is dat G_{maas} en G_{dim} grotendeels onafhankelijk zijn verdeeld, wat tevens blijkt uit de lage bijbehorende correlatiecoëfficiënt (ρ_{23}=-0,18, witte rechthoek). G_{maas} is slechts zwak gecorreleerd met demografische factoren, terwijl G_{dim} sterker afhangt van demografische factoren (respectievelijk rij 2 en 3, zwarte rechthoek in Figuur 7). Opvallend hierbij is dat G_{dim} vrij sterk afneemt met toenemende bevolkingsdichtheid ($\rho_{3,17}$=-0,74) , wat ook blijkt uit de geografische verdelingen van deze parameters (vergelijk Figuur 8b en Figuur 8c). Lagere verbruik-netwerkvolume-verhoudingen (hogere capaciteit) in landelijke gebieden ten opzichte van stedelijke gebieden worden veroorzaakt doordat in landelijk gebied veel leidingkilometers nodig zijn om klanten te verbinden met het leidingnet, terwijl in steden de onderlinge afstanden kleiner zijn. Vanwege deze relatie is G_{dim} ook gecorrelateerd met andere demografische factoren die op hun beurt samenhangen met de bevolkingsdichtheid.



FIGUUR 8. NETWERK KARAKTERISTIEKEN. (A) VERMAZINGSGRAAD G_{MAAS}. WARME EN KOUDE KLEUREN GEVEN RESPECTIEVELIJK EEN LAGE EN HOGE VERMAZINGSGRAAD AAN. (B) DIMENSIONERINGSGRAAD G_{DIM}, WARME KLEUREN TONEN BUURTEN MET EEN RELATIEF HOGE VERBRUIKSLAST (LAGE VOLUME-VERBRUIKSVERHOUDING). (C) BEVOLKINGSDICHTHEID WAARBIJ STEDELIJKE EN LANDELIJKE GEBIEDEN HERKENBAAR ZIJN. DE INZETTEN IN ELK PANEEL TONEN LEIDINGEN EN BUURTGRENZEN.

4.1.2 Significantie relaties klantmeldingen en omgevingsfactoren

Op klantadresniveau hebben we verdelingen van temperatuur en leidingdiameter, -lengte en –materiaalsoort uitgerekend die samenhangen met de bruinwater- en de referentiemeldingen. We hebben daarbij bepaald of verschillen in de distributies statistisch significant zijn en gebruiken daarvoor een t-toets voor datasets van ongelijke monstergrootte en ongelijke variantie (de zogenaamde Welch's t-toets voor "double-tailed events"). Hieruit volgt voor elke vergelijking van 2 verdelingen een p-waarde tussen 0 en 1, waarbij een hogere waarde betekent dat het minder waarschijnlijk is dat de twee verdelingen hetzelfde gemiddelde hebben (en dus meer aannemelijk dat ze op een statistisch significante manier van elkaar afwijken). Voor de temperatuurverdelingen hebben we naast de referentiemeldingen ook de drie weerstations gebruikt als een referentie voor bruinwatermeldingen.

Resultaten van de berekende p-waarden staan in Tabel 3. Meest opvallend daarbij is dat de temperatuursverdeling voor bruinwatermeldingen significant afwijkt van die voor de referentiemeldingen, wat wordt weerspiegelt door de lage p-waarde van 1,5x10-6. Het temperatuurhistogram in Figuur 9a illustreert hoe bruinwatermeldingen worden gedaan bij gemiddeld hogere temperaturen dan de referentiemeldingen. Bovendien wijkt de bruinwatermelding-gerelateerde temperatuursverdeling af van de verdelingen voor de 3 weerstations, wat het verder aannemelijk maakt dat bruinwatermeldingen gemiddeld worden gedaan bij relatief hoge temperaturen. De temperatuursverdelingen van de 3 weerstations onderling zijn vrijwel inwisselbaar, zoals ook blijkt uit de bijbehorende hoge p-waarden [2]van 0,17 – 0,57 (Tabel 3) en uit de grotendeels overlappende verdelingen (Figuur 9b). De enigszins bimodale verdeling van de temperaturen komt overigens goed overeen met de gegevens voor de lange-termijn (1951-2015, niet getoond) en weerspiegelt de seizoensafhankelijke variatie in lichtinval.

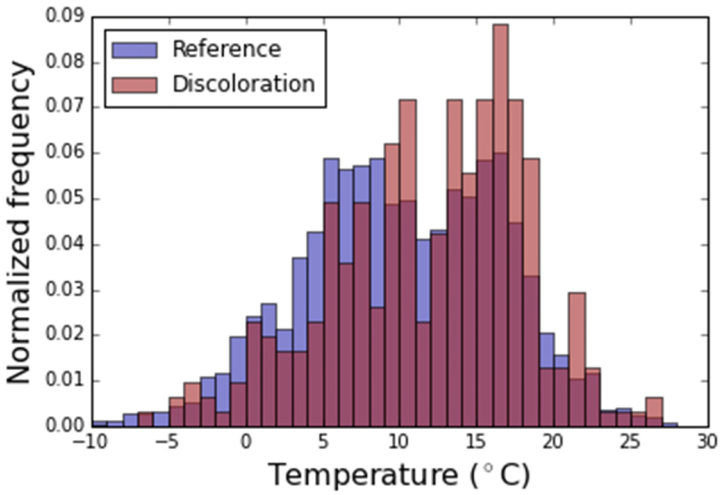
Temperaturen van de referentiemeldingen wijken minder (dan de bruinwatermeldingen) af van de weerstationsmetingen, met een p-waarde >1,5x10^{-3} en een verschil in gemiddelde waarde <0,5°C. De afwijking is bovendien in omgekeerde richting: "referentiemeldingen" vinden plaats bij een ietwat lagere temperatuur dan op basis van het jaargemiddelde mag worden verwacht, en dit contrasteert met de relatief hoge temperaturen voor bruinwatermeldingen. Het zoeken naar mogelijke verklaringen valt buiten de strekking van dit onderzoek, maar het vaker voorkomen van meldingen bij koude temperaturen als gevolg van klant-, administratie,- of drinkwatergerelateerde verklaringen.

TABEL 3. STATISTISCHE SIGNIFICANTIE VAN AFWIJKENDE GEMIDDELDE WAARDEN VAN VERDELINGEN VAN PARAMETERS (A EN B). HET GEMIDDELDE VAN ELKE VERDELING STAAT TUSSEN HAAKJES.

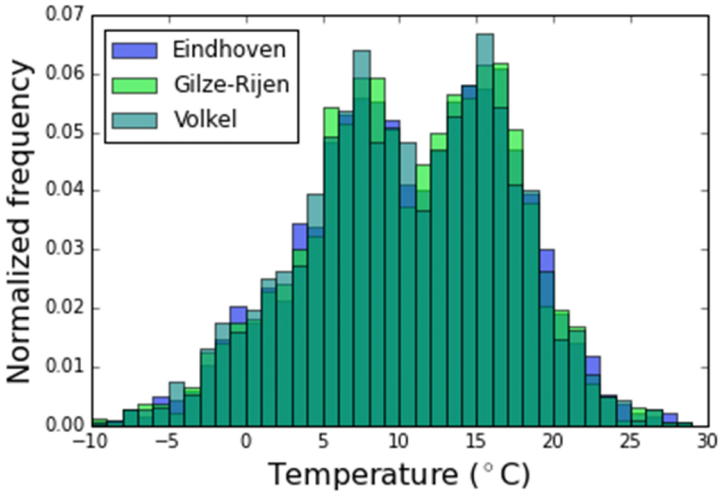
Grootheid (eenheid)	Verdeling A (gem.)	Verdeling B (gem.)	p-waarde (-)
Temperatuur (°C)	Bruinwater-gerelateerd (12.9)	Referentie (10.1)	1.5x10^{-6}
	Bruinwater-gerelateerd (12.9)	Weerstation 1 (10.6)	1.4x10^{-3}
	Bruinwater-gerelateerd (12.9)	Weerstation 2 (10.5)	4.3x10^{-4}
	Bruinwater-gerelateerd (12.9)	Weerstation 3 (10.3)	8.0x10^{-5}
	Referentie (10.1)	Weerstation 1 (10.6)	1.5x10^{-3}
	Weerstation 1 (10.6)	Weerstation 2 (10.5)	0.57
	Weerstation 2 (10.5)	Weerstation 3 (10.3)	0.42
	Weerstation 3 (10.3)	Weerstation 1 (10.6)	0.17
Leidingdiameter (mm)	Bruinwater-gerelateerd (99.8)	Referentie (108,2)	3.0x10^{-4}
Leidinglengte (m)	Bruinwater-gerelateerd (77.0)	Referentie (79.5)	0.43
Leidingmateriaalsoort (-)	-		n.v.t.

[2] De p-waarde is de maat van waarschijnlijkheid waarbij de hypothese van gelijke gemiddeldes van twee datasets ontkracht wordt.

a



b



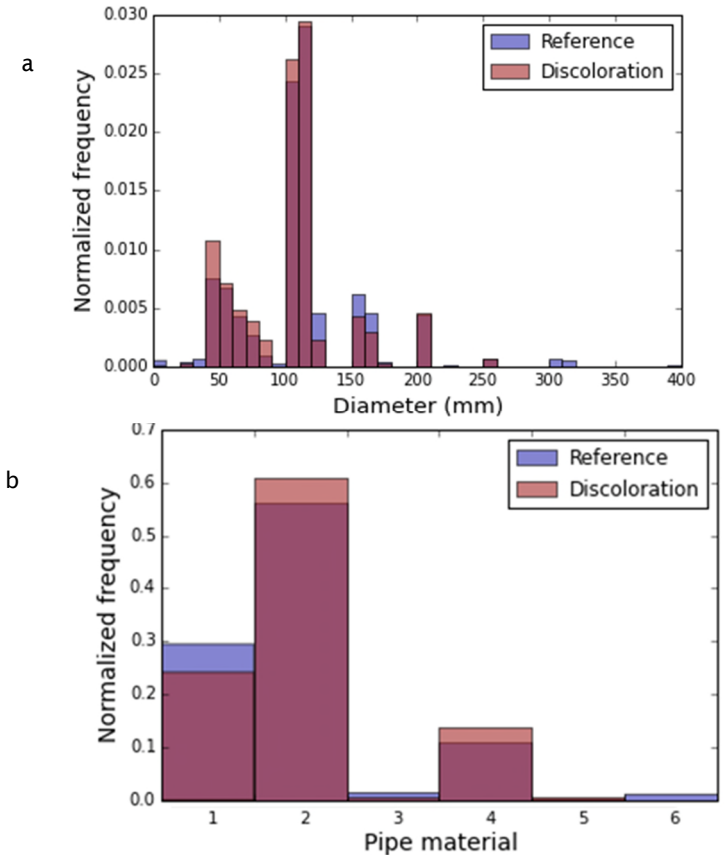
FIGUUR 9. TEMPERATUURHISTOGRAM VOOR (A) BRUINWATERMELDINGEN (BRUIN) EN REFERENTIEMETINGEN (BLAUW) EN (B) DRIE WEERSTATIONS IN HET DISTRIBUTIEGEBIED IN DEZELFDE PERIODE (2010-2014).

De invloed van leidingdiameter op bruinwatermeldingen is veel zwakker dan de invloed van temperatuur, maar nog steeds statistisch significant, zoals blijkt uit de p-waarde in Tabel 3. Bruinwaterincidenten worden gerapporteerd bij (gemiddeld) kleinere diameters dan meldingen van de referentieset, zoals is te zien in Figuur 10a. De intervalfrequenties zijn ietwat gelimiteerd voor de bruinwater-gerelateerde set (de maximum frequentie is 90 in het interval voor diameters 100-110 mm), wat betekent dat ieder interval is blootgesteld aan een zekere mate van statistische ruis. Niettemin zijn de intervalfrequenties structureel hoger voor bruinwater-gerelateerde meldingen dan voor de referentieset voor alle intervallen in het diameterdomein 40-120 mm, wat de statistische significantie verhoogt.

Een (directe) invloed van diameter via hydraulica is mogelijk omdat (ongeacht het netwerkontwerp) stroomomkeringen en sterke snelheidsgradiënten waarschijnlijker zijn voor kleine diameters, en dit versterkt de opwervelingspotentie. Echter, een (indirecte) invloed van diameter via netwerkontwerp is ook mogelijk, aangezien zelfreinigende netwerken over het algemeen worden aangelegd met kleinere leidingdiameters dan niet-zelfreinigende netwerken.

Zoals voor de meeste voedingsgebieden heeft het distributiesysteem in deze studie minder zelfreinigende dan vermaasde netwerksecties omdat eerstgenoemde nog relatief nieuw zijn (sinds 2000). Om een beter onderscheid te kunnen maken tussen een directe invloed van de diameter en een mogelijke indirecte invloed van netwerkontwerp is het aan te bevelen om de diameterrelaties voor beide netwerktypen afzonderlijk te onderzoeken. Hiervoor is echter wel voldoende data nodig om beide sub-netwerken voldoende goed te kunnen bemonsteren.

Onze resultaten tonen geen significante afhankelijkheid van meldingen met leidinglengte of leidingmateriaal. Omdat leidingmaterialen zijn gecategoriseerd (in plaats van gekwantificeerd), hebben we hierop geen t-test kunnen uitvoeren. In plaats daarvan hebben we de data in histogrammen gevisualiseerd (Figuur 10b). Deze tonen dat de verschillen in materialen tussen de verdelingen van de bruinwatermeldingen en referentiemeldingen klein zijn. Het berekende verschil in leidinglengte is niet significant, zoals blijkt uit de hoge p-waarde van 0,43 (Tabel 3).



FIGUUR 10. HISTOGRAMMEN VAN (A) LEIDINGDIAMETER EN (B) LEIDINGMATERIAAL, GEASSOCIEERD MET BRUINWATER-GERELATEERDE KLANTMELDINGEN (BRUIN) EN ALGEMENE MELDINGEN (REFERENTIESET, BLAUW). LEIDINGMATERIAAL CATEGORIËN ZIJN (1) CH, (2) PVC, (3) PE, (4) GIETIJZER, (5) NODULAIR GIETIJZER EN (6) OVERIGE MATERIALEN.

4.1.3 Discussie analyse bruinwatermeldingen

De door ons berekende zwakke correlatie tussen meldingen en demografische factoren maakt het aannemelijk dat gebruikte klantmeldingen slechts blootgesteld zijn aan een lage demografische vervorming. Dit suggereert dat de meldingen een nuttige indicator zijn voor bruinwaterincidenten, zelfs zonder een correctie voor de gedragspatronen van klanten zoals

klanten die het incident waarnemen of de bereidwilligheid van klanten om melding te maken bij het waterbedrijf. Het koppelen van klantmeldingsfrequenties met buurtgemiddelden introduceert een statistische fout omdat we gevolgtrekkingen van het gedrag van individuen afleiden uit de gevolgtrekking voor de groep waartoe die individuen behoren (ecologische bedrieglijkheid). De invloed van deze ecologische bedrieglijkheid vermindert met een toenemend aantal klantmeldingen in een buurt. Dit benadrukt het belang van het gebruik van een grote hoeveelheid klantcontactgegevens.

Onze bevinding van een sterke positieve correlatie tussen bruin water gerelateerde klantmeldingen en temperatuur kan verschillende oorzaken hebben. Het is mogelijk dat hogere temperaturen de microbiologische activiteit en formatie van biofilms bevordert, met als gevolg een verhoogde accumulatiepotentie van deeltjesmateriaal op de leidingwand en een verhoogd bruinwaterrisico. Deze hypothese komt overeen met eerder werk waarin wordt gedemonstreerd dat hogere temperaturen de biofilmpotentie bevorderen (Kooij 2001) en dat een waargenomen hogere accumulatiesnelheid in distributiesystemen bij hogere temperatuur niet kan worden verklaard door variaties in de kwaliteit van inkomend water van de zuivering alleen, maar waarschijnlijk worden versterkt door hogere temperaturen in het leidingnet (Blokker and Schaap 2015). Deze mogelijke verklaringen zijn gerelateerd aan conditie 1 (paragraaf 2.1.2). Het is mogelijk dat temporele variaties de frequentie van bruinwater-gerelateerde klantmeldingen beïnvloeden. Zo is het mogelijk dat de watervraag bij hogere temperaturen toeneemt omdat mogelijk mensen bijvoorbeeld vaker douchen of hun tuin sproeien. Dit zou tot een toename kunnen leiden van het aantal hydraulische verstoringen wat de zelfreinigende werking van (delen van) het leidingnetwerk kan beïnvloeden (conditie 2, paragraaf 2.1.2). De invloed van vraagpatronen op resuspensie kan verder worden onderzocht door continue monitoring van de waterkwaliteits- (turbiditeitsmetingen) en –kwantiteitspatronen (vraag, volumestromen, snelheden) op meerder locaties en gedurende verschillende seizoenen. Een ander mogelijk seizoensgebonden effect is een afwijkende demografie tijdens de zomervakantieperiode vanwege afwijkende (thuis-)aanwezigheid van klanten wat de waarschijnlijkheid om incidenten waar te nemen zou kunnen verhogen (conditie 3, paragraaf 2.1.2). Ook de bereidheid van klanten om contact op te nemen met het waterbedrijf (conditie 4, paragraaf 2.1.2) zou tijdens de zomervakantie hoger kunnen zijn. Een demografische oorzaak (conditie 3 en 4) verklaart echter niet het significante verschil tussen bruinwater-gerelateerde en “referentie”-meldingen. Hoewel wij niet expliciet de seizoensafhankelijkheid hebben onderzocht, toont onze studie slechts een zwakke correlatie tussen bruinwater-gerelateerde meldingen en demografische factoren. Deze zwakke correlatie suggereert dat de meldingen een bruikbare indicator zijn voor werkelijk optredende, waarneembare drinkwaterfenomenen. Personeelscapaciteit bij drinkwaterbedrijven kan ook met de seizoenen samenhangen, wat de frequentie van reparaties en/of spuiacties kan beïnvloeden en daarmee ook de bruinwaterrisico’s. Bij Brabant Water zijn spuiacties uniform over het jaar verdeeld (afgezien van pauzes samenhangend met vorst of langdurige droogte). Hoewel de seizoensinvloed van spuiacties daarom klein of afwezig is, kan toevoeging van data van spuiactie-programma’s en reparaties de analyse verder verrijken.

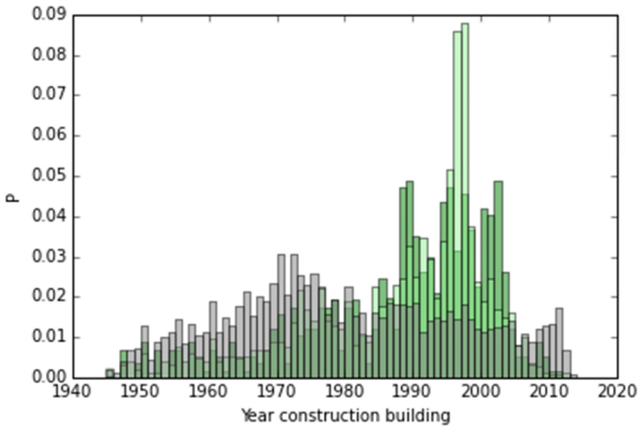
Bruinwaterklachten hangen mogelijk samen met seizoensinvloeden van de kwaliteit van het inkomende water. Deze mogelijkheid lijkt echter in tegenspraak met een recente studie uitgevoerd bij PWN, die toont dat de seizoensafhankelijkheid van bruin water een bron heeft in het distributienet en niet in het inkomende water van de zuivering (Blokker and Schaap 2015). Daarnaast zijn bij Brabant Water slechts kleine kwaliteitsvariaties in het gezuiverde water te verwachten omdat alleen grondwater wordt gebruikt voor productie. Niettemin is het raadzaam om mogelijke variaties in inkomende waterkwaliteit te verifiëren voor verschillende netwerklocaties door de meest waarschijnlijke waterbron te bepalen op basis van

expertkennis, berekening van de kortste hydraulisch afstand tot een waterbron, of geavanceerde hydraulische modelberekeningen.

De resultaten van dit onderzoek tonen dat bruinwater-gerelateerde meldingen vaker voorkomen bij leidingen met een kleine diameter. Mogelijk weerspiegelt dit een hydraulische oorzaak, bijvoorbeeld vanwege een toegenomen depositie-opwervelingspotentie voor kleine diameter leidingen. Een andere mogelijkheid voor een bruinwater-diameter afhankelijkheid is een microbiologisch proces, aangezien een toename in de oppervlakte-volume-verhouding voor dunne leidingen een relatief groot substratum biedt voor microbiologische activiteit, wat kan leiden tot een verhoogde deeltjesaccumulatiepotentie. Het is tevens mogelijk dat water in dunne leidingen relatief vaak aan hoge temperaturen wordt blootgesteld omdat het water sneller de temperatuur van de omgeving aanneemt. Als dit zo is, manifesteert een bruinwater-temperatuur-relatie (zie de discussie hierboven over mogelijke oorzakelijke verbanden) zich in een negatieve bruinwater-diameter relatie. Om het leidende proces achter de gevonden bruinwater-diameter relatie te achterhalen is verder onderzoek nodig.

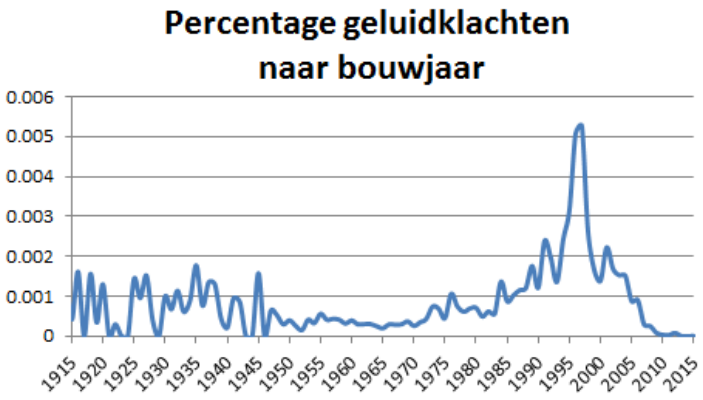
4.1.4 Analyse geluidsmeldingen van watermeters

De analyse toont dat geluidsmeldingen van watermeters vaker voorkomen bij huizen gebouwd rond 1995-2005: er is een significant verschil tussen de verdeling van bouwjaar van panden voor geluidsklachten, ten opzichte van die voor overige klachten (Figuur 11). Geluidsklachten van watermeters zijn slechts zwak-gecorreleerd aan socio-economische parameters en leidingnetkarateristieken (vermazings- en dimensioneringsgraad). Er is nagegaan dat het groter aantal meldingen van huizen uit deze periode niet het gevolg is van de grotere aanwezigheid van woningen uit deze periode (Figuur 12). De correlatie tussen bouwjaar en meldingen is ook herkenbaar[3] in een geografische weergave (Figuur 13). Een voorbeeld is de gemeente Oss waarin de meeste vervangen meters voorkomen in nieuwbouwwijken gebouwd na 1971.

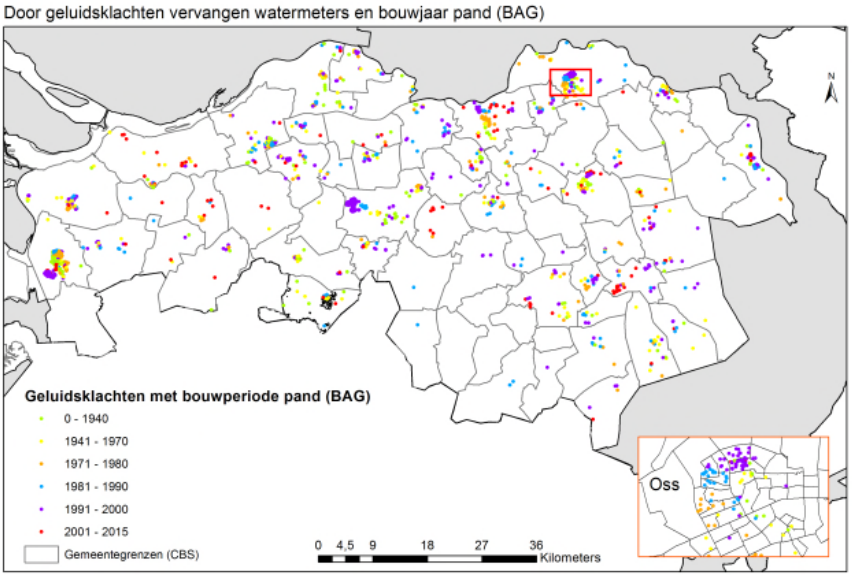


FIGUUR 11. VERDELING VAN GELUIDSKLACHTEN NAAR BOUWJAAR VAN HET PAND. IN GROEN DE VERDELING VOOR 1731 VERVANGEN METERS. IN GRIJS DEVERDELING VOOR 14624 ALGEMENE MELDINGEN.

[3] De correlatie tussen bouwjaar en het optreden van geluidsklachten is met deze geografische weergave niet herleidbaar (omdat het bouwjaar van panden zonder storing niet is getoond), maar illustreert wel het eerder genoemde statistisch relevante verband.



FIGUUR 12. PERCENTAGE VAN WONINGEN EN PANDEN WAARVAN WATERMETERS ZIJN VERVANGEN,. VOOR ELK JAAR IS GENORMALISEERD MET HET TOTAAL AANTAL WONINGEN.



FIGUUR 13. GEOGRAFISCHE VERDELING VAN GELUIDSKLACHTEN MET KLEURLABELS NAAR BOUWJAAR PAND (RECHTS).

4.1.5 Discussie van analyse meldingen geluiden watermeters

Oorzaken voor het verband tussen geluidsmeldingen van watermeters en bouwjaar zijn vooralsnog niet duidelijk aan te geven. In overleg met Johan van Erp (Brabant Water) zijn de volgende punten benoemd:

1. Volumetrische watermeters zorgen over het algemeen voor meer geluidsklachten. Dit komt door (a) een fysiek contactmoment in en (b) door het drukverschil optredend als gevolg van het meetprincipe.
2. Bij een watermeter met een hogere nauwkeurigheid wordt (1) kritischer, vanwege de nauwere dimensionering.

Het tikkende geluid dat ontstaat kan versterkt worden door de binneninstallatie. Deze varieert sterk door het voorzieningsgebied van Brabant Water. Een beschouwing hierover is als volgt:

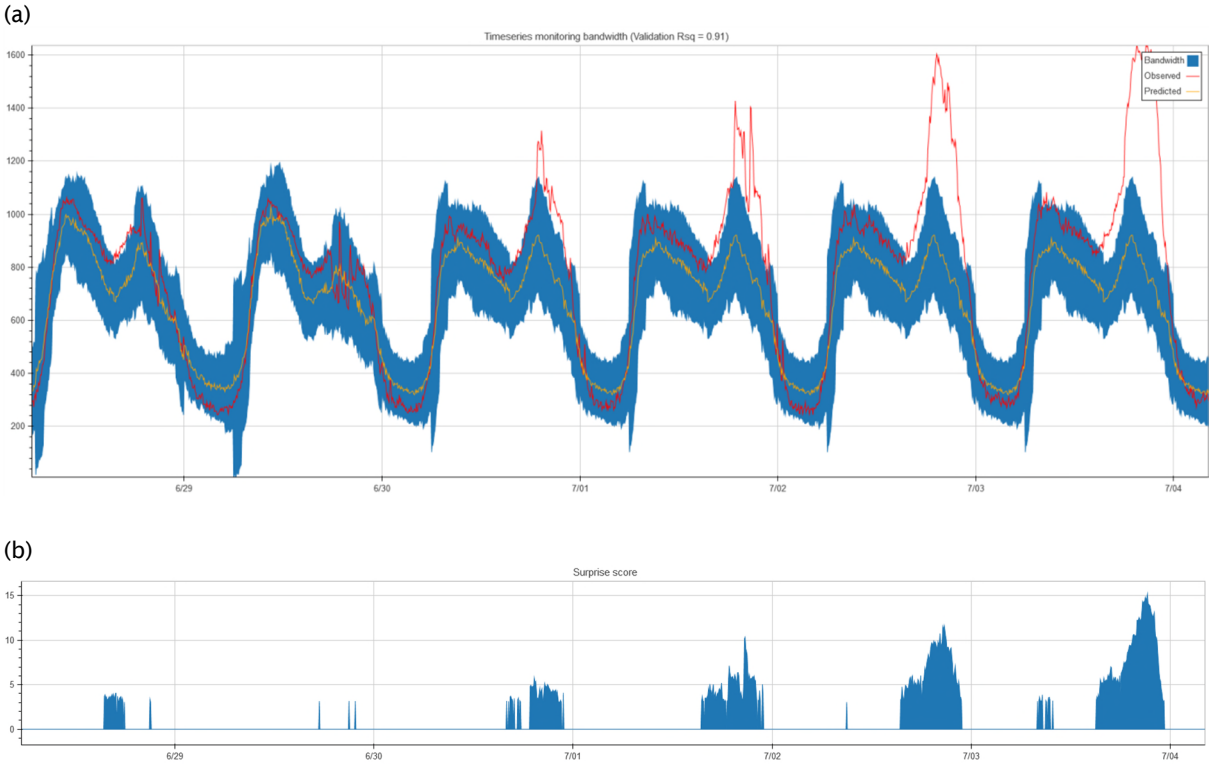
1. Appartementencomplexen hebben vaak lange stijgleidingen, waarin het geluid getransporteerd wordt.

2. Oudere panden zijn vaak gehoriger dan nieuwere panden die beter geïsoleerd zijn.
3. Per aannemer kan er een verschil zijn wat betreft de aanleg van binnen-installaties. Een leiding die niet goed vast zit verplaatst het geluid eenvoudiger.
4. In grofweg de helft van de panden met een watermeter is deze aangebracht in de meterput. Bij overige panden is de meter doorgaans aangebracht in de meterkast. De tweede groep levert de meeste klachten (meterkast wordt klankkast). Bij een zeer klein deel van de panden is de watermeter ergens anders in de woning aangelegd. In sommige gevallen levert dit klachten (toilet, slaapkamer etc.).

4.2 Casus P

Voor casus P is alle data voorbewerkt. In totaal zijn er 426 storingen over de periode van juni 2014 tot en met maart 2015 (de periode die overlapt met beschikbare meetreeksen voor volumestroom en druk). Voor 317 respectievelijk 321 van deze storingen is druk- of volumestroomdata beschikbaar (niet op alle meetlocaties en alle momenten zijn deze data beschikbaar). Locaties waarvoor geen (of onvolledige) data beschikbaar was zijn opgesomd in Bijlage II.

De anomaliedetectie is uitgevoerd op alle volumestroom- en drukreeksen, de methodiek is uitgelegd in het TKI-rapport KWR 2016.024 (Vries et al., 2016) en het BTO-rapport (Vonk and Vries 2015). Een voorbeeld hiervan is weergegeven in Figuur 14. Op een aantal reeksen geeft het regressiealgoritme echter een slechte fit ($R^2_{test} < 0,5$). Voor 85 storingen zijn er ook drukanomalieën op de dag van storen of tijdens de twee dagen daarvoor. Voor de andere storingen zijn er geen anomalieën gevonden. Het aantal anomalieën in de volumestroommetingen is hoger; hier zijn er in totaal 168 gevonden.



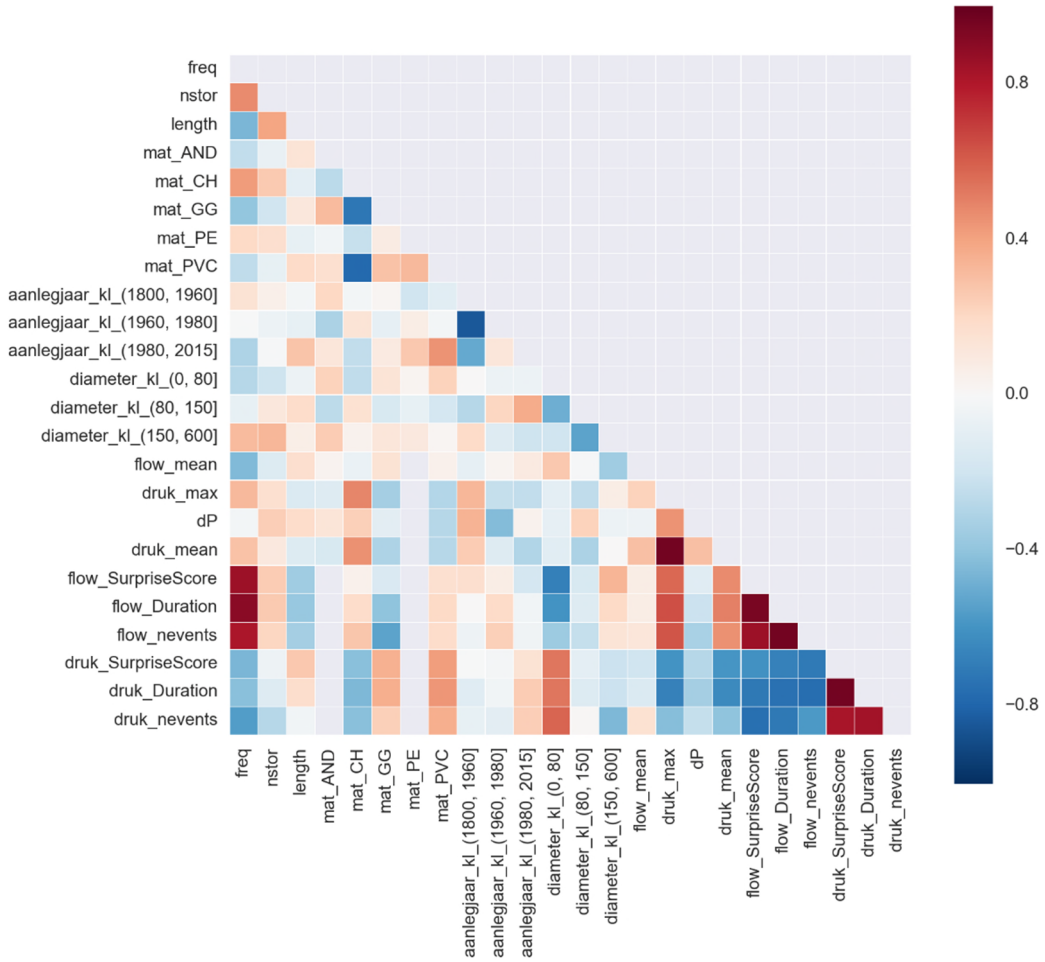
FIGUUR 14. (A) BANDBREEDTEMONITORING MET HET ANOMALIEDETECTIEALGORITME EN (B) DE MATE VAN AFWIJKING VAN GEMETEN VOLUMESTROOM TEN OPZICHTE VAN MODELVOORSPELLING IN DE VORM VAN DE SUPRISE SCORE. DIT IS EEN UITSNEDE UIT DE VOLUMESTROOM-METINGEN BIJ DRUKZONE SEPPE1.

4.2.1 Resultaten analyse 1 (drukzone-gewijs)

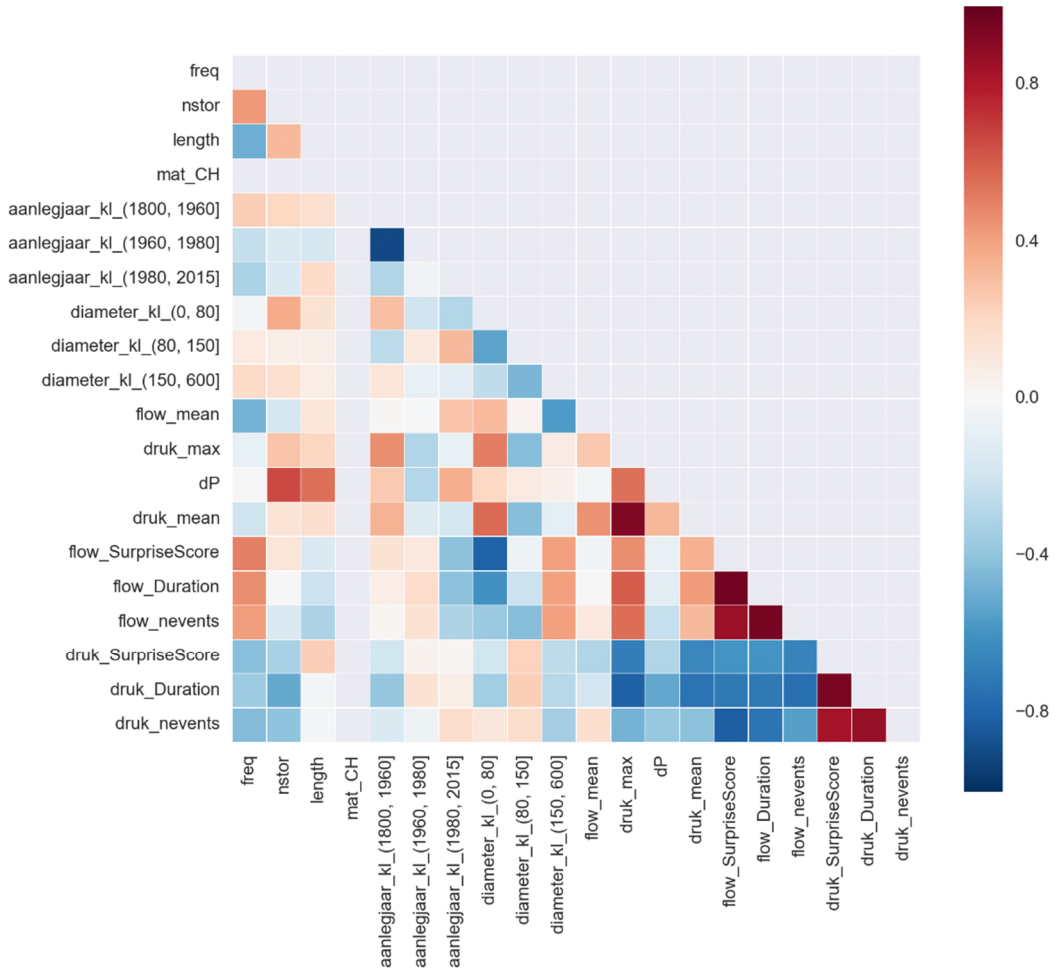
In totaal zijn er 55 drukzones. Na verwijdering van alle drukzones met minder dan 5 storingen blijven er 31 over. Vervolgens vallen er 6 drukzones af, omdat hiervan geen meetgegevens beschikbaar zijn, zodat er 25 drukzones overblijven. Dan zijn er nog 2 drukzones waar geen anomalie-informatie beschikbaar is, zodat er 23 drukzones overblijven. Voor ieder attribuut zijn de uitbijters verwijderd (gedefinieerd als een afwijking groter dan 3 keer de standaarddeviatie).

Voor de analyse op basis van locatie is een correlatie-matrix gemaakt die de correlatie tussen parameters aangeeft voor alle leidingen (Figuur 15) en voor alleen CH leidingen (Figuur 16). Opvallend is dat er geen duidelijk aanwijsbare één op één relatie bestaat tussen storingsfrequentie en de onderzochte variabelen. De gemiddelde volumestroom lijkt echter negatief gecorreleerd te zijn met de storingsfrequentie. Daarnaast vertoont de materiaalsoort CH (cementhoudend) een licht positieve correlatie met de storingsfrequentie. Dit komt overeen met observaties in de praktijk dat CH relatief vaak stoort, vergeleken met andere leidingmaterialen.

Datamining is toegepast op de verkregen data uit deze analyse, echter de gevonden fits (R^2 op basis van kruisvalidatie) waren erg laag. Dit komt ofwel omdat het aantal attributen vrij groot is ten opzichte van het aantal metingen (dit werkt overfitting in de hand, waardoor bij validatie er een lage R^2-score is), of omdat er geen aantoonbare verbanden zijn.



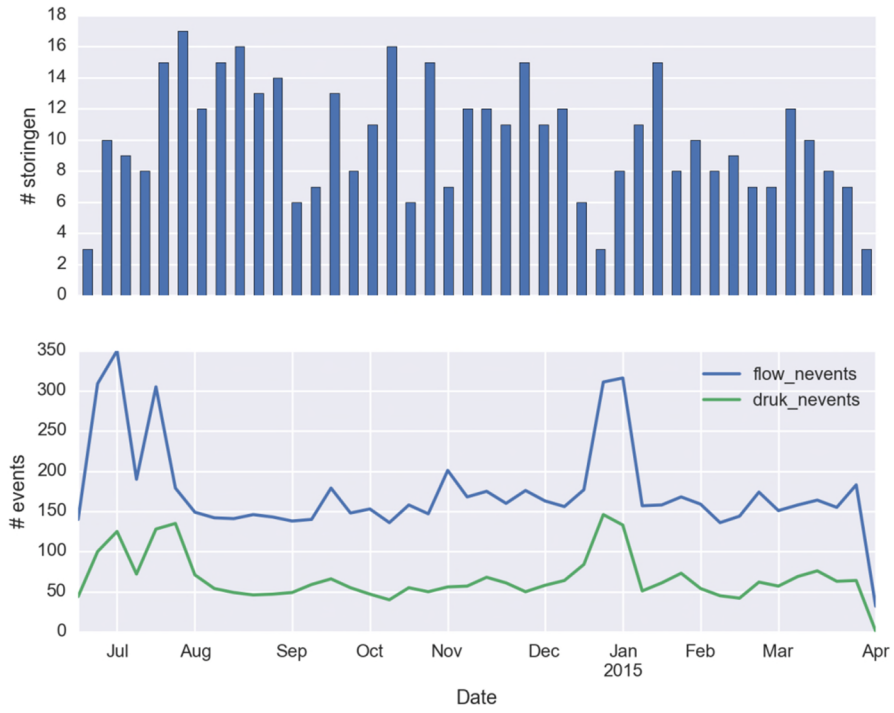
FIGUUR 15. CORRELATIE-MATRIX (OP BASIS VAN SPEARMAN-COEFFICIENT) TUSSEN DE VERSCHILLENDE ATTRIBUTEN BEPAALD PER DRUKZONEGEBIED (ALLE LEIDINGEN). IN DEZE PLOT IS 'FREQ' DE STORINGSFREQUENTIE, 'NSTOR' HET AANTAL STORINGEN (ABSOLUUT) EN 'DP' HET MAXIMALE DRUKVERSCHIL OP DE TWEE DAGEN VOORAFGAAND AAN DE STORING.



FIGUUR 16. CORRELATIE-MATRIX (OP BASIS VAN SPEARMAN-COEFFICIENT) TUSSEN DE VERSCHILLENDE ATTRIBUTEN BEPAALD PER DRUKZONEGEBIED (ALLEEN CH LEIDINGEN). IN DEZE PLOT IS 'FREQ' DE STORINGSFREQUENTIE, 'NSTOR' HET AANTAL STORINGEN (ABSOLUUT) EN 'DP' HET MAXIMALE DRUKVERSCHIL OP DE TWEE DAGEN VOORAFGAAND AAN DE STORING.

4.2.2 Resultaten analyse 2 (variaties over de tijd)

Voor deze analyse is alle data van Brabant Water samengenomen en zijn op weekbasis het aantal storingsmeldingen vergeleken met het aantal 'events' dat door de anomaliedetectie is gekoppeld. Dit is gedaan door alle storingen en anomalieën op te tellen over alle gebieden en vervolgens op weekbasis een tijdreeks te genereren (Figuur 17). Te zien is dat het aantal anomalieën voor druk en volumestroom sterk correleert. Opvallend is verder dat op de momenten dat er veel anomalieën zijn, zoals in de zomervakantie en rond de kerst, er juist minder storingen zijn. Naast het aantal anomalieën zijn er op weekbasis ook andere druk en volumestroom parameters bepaald, zoals gemiddelde druk, maximale druk en maximum drukverschil. Met behulp van regressie (Gradient Boosting Regression) is hier een verband gezocht tussen deze attributen en storingen. Ook hier is geen goede fit gevonden tussen de attributen en het aantal storingen. Wat betreft het aantal anomalieën lijkt een omgekeerde trend zichtbaar: minder storingen bij meer anomalieën. Dit heeft mogelijk te maken met de vakantieperiodes, waarin veel anomalieën werden geregistreerd, maar juist minder storingen (zie ook Figuur 17). Een mogelijke verklaring is, dat veel mensen op vakantie zijn en daarom geen storingen melden, terwijl het drukregime door diezelfde vakantie-afwezigheid ook afwijkt ten opzichte van normale patronen.



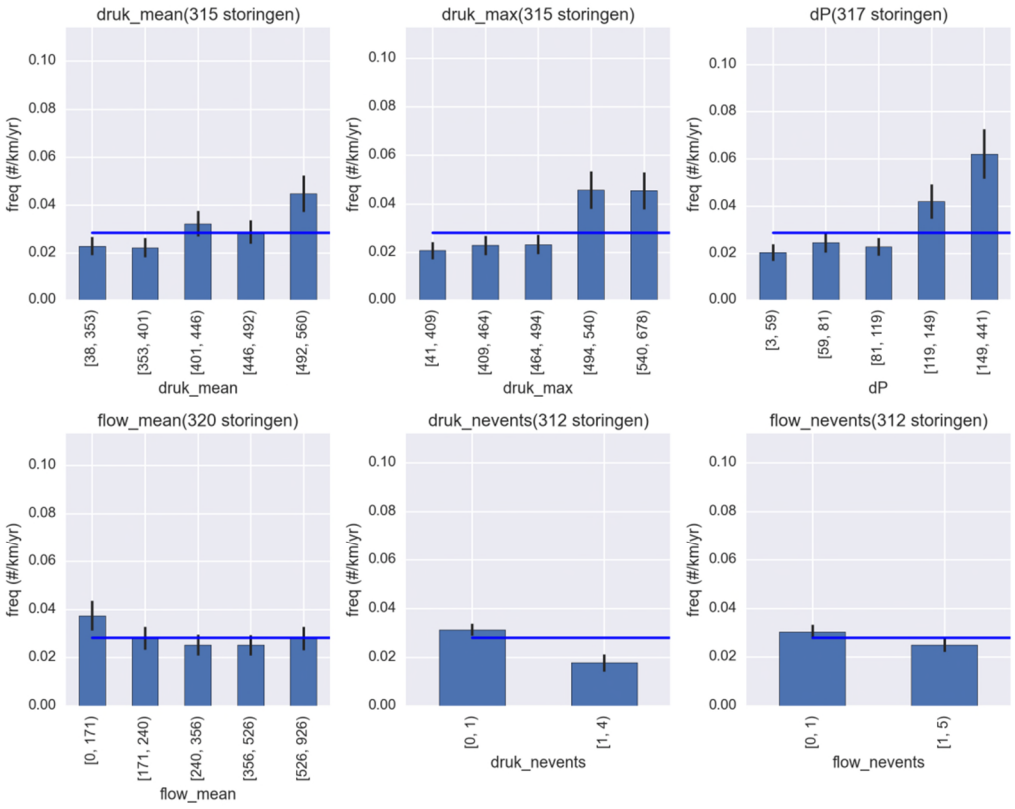
FIGUUR 17. TIJDREEKSEN VAN AANTAL STORINGEN PER WEEK (BOVEN) EN AANTAL ANOMALIEN VOOR DRUK (DRUK_NEVENTS) EN VOLUMESTROOM (FLOW_NEVENTS) PER WEEK.

4.2.3 Resultaten analyse 3 (indeling in druk- of volumestroomklasse)

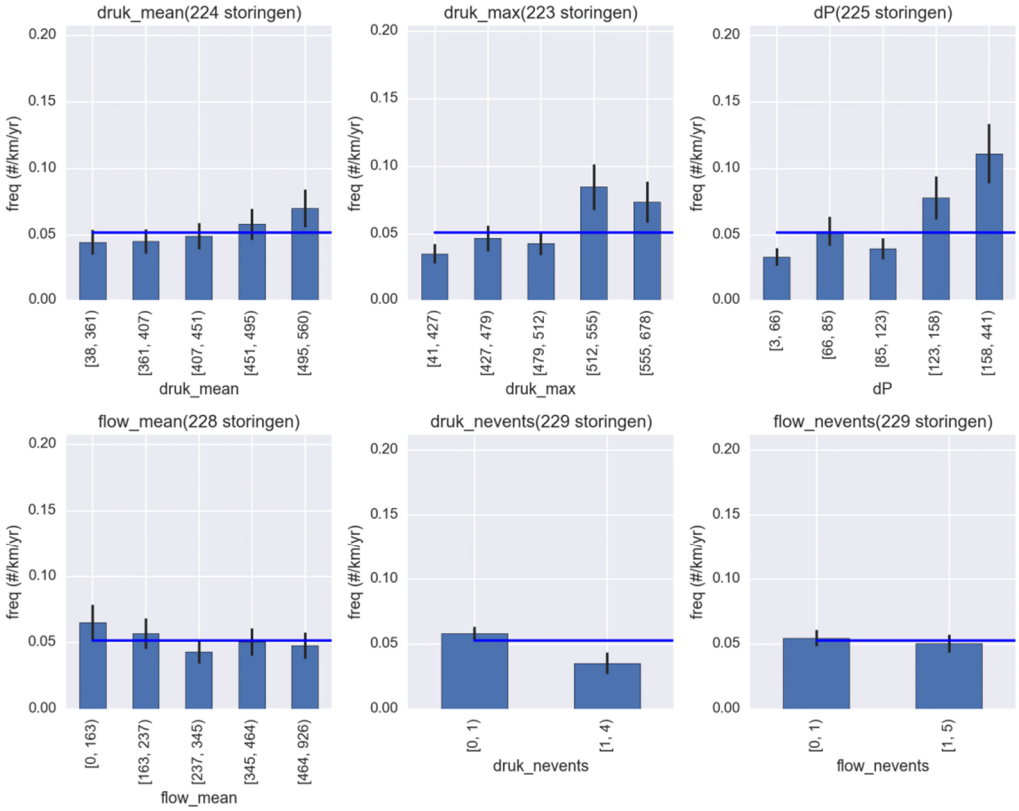
Voor deze analyse zijn er grafieken gemaakt die de storingsfrequentie uitzet tegen de druk of volumestroomklassen voor alle leidingen. De analyse is herhaald voor alleen CH leidingen en overige leidingen (geen CH materiaal). Uit deze drie analyses (Figuur 18, Figuur 19 en Figuur 20) valt op te maken dat:

- Er voor alle materialen geen verband lijkt te zijn met volumestroom en de storingsfrequentie;
- CH stoort meer bij hogere gemiddelde druk, terwijl andere materialen geen toenemende storingsgevoeligheid hebben voor gemiddelde druk;
- CH vanaf ongeveer 500 kPa maximale druk, een stuk hogere storingsfrequentie heeft dan bij een lager maximaal drukregime, terwijl bij overige materialen vaker een storing optreedt bij het drukregime tussen 435 en 463 kPa. Bovendien is de storingsgevoeligheid voor maximale druk bij CH een factor 3 hoger dan bij de andere materialen;
- CH ten opzichte van andere materialen een ander patroon volgt ten aanzien van de storingsfrequenties gerelateerd aan het maximale drukverschil op een dag: storingen treden in toenemende mate op vanaf een max-dag-drukverschil van 119 kPa, terwijl bij de andere materialen alleen hogere storingsfrequenties optreden bij een regime van 109 tot 129 kPa en deze circa 5x zo laag zijn als bij CH.

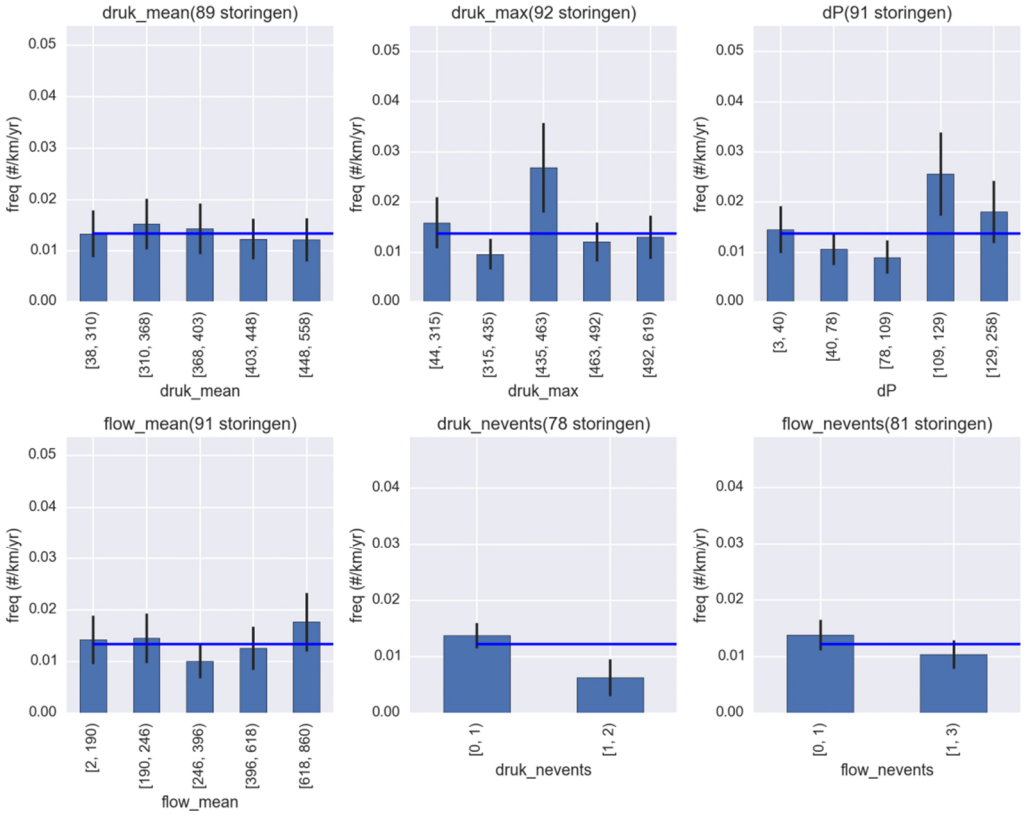
Hieruit valt op te maken dat met name de CH leidingen gevoelig zijn voor druk en storingsfrequentie.



FIGUUR 18. STORINGSFREQUENTIE (**ALLE** MATERIALEN) PER VERKLARDENDE VARIABELEN ZIJN OPGEDEELD IN KLASSEN VAN GELIJKE GROOTTE: (DRUK_MEAN) GEMIDDELDE DRUK BIJ MEETPUNT, (DRUK_MAX) MAXIMALE DRUKOP EEN DAG, (DP) MAXIMALE VERSCHILDRUK OP EEN DAG (FLOW_MEAN) GEMIDDELDE VOLUMESTROOM (M3/H), (DRUK_NEVENTS) AANTAL ANOMALIEN, (FLOW_NEVENTS)) AANTAL ANOMALIEN VOLUMESTROOMREEKS. DE BLAUWE LIJN IS DE GEMIDDELDE STORINGSFREQUENTIE VOOR ALLE COHORTEN.



FIGUUR 19. STORINGSFREQUENTIE (ALLEEN **CH**) PER VERKLARDENDE VARIABELE, OPGEDEELD IN KLASSEN VAN GELIJKE GROOTTE: (DRUK_MEAN) GEMIDDELDE DRUK BIJ MEETPUNT, (DRUK_MAX) MAXIMALE DRUKOP EEN DAG, (DP) MAXIMALE VERSCHILDRUK OP EEN DAG (FLOW_MEAN) GEMIDDELDE VOLUMESTROOM (M3/H), (DRUK_NEVENTS) AANTAL ANOMALIEËN, (FLOW_NEVENTS)) AANTAL ANOMALIEËN VOLUMESTROOMREEKS.



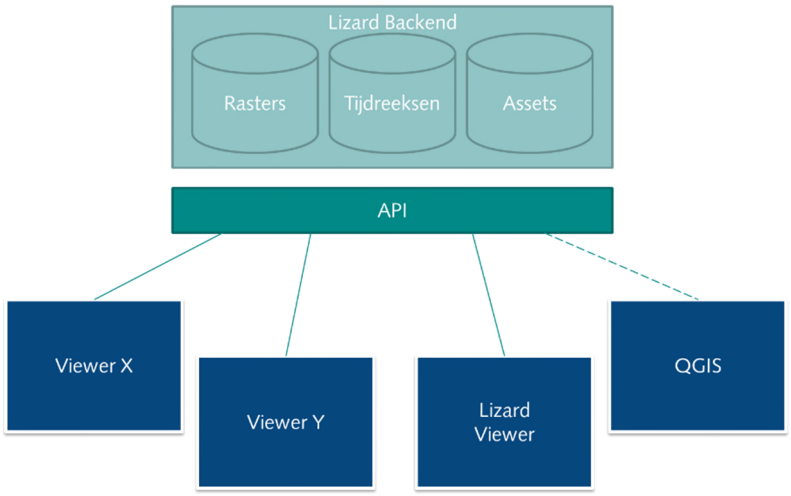
FIGUUR 20. STORINGSFREQUENTIE (ALLES BEHALVE **CH**) PER VERKLARENDE VARIABELEN, OPGEDEELD IN KLASSEN MET EEN GELIJK AANTAL STORINGEN: (DRUK_MEAN) GEMIDDELDE DRUK BIJ MEETPUNT, (DRUK_MAX) MAXIMALE DRUKOP EEN DAG, (DP) MAXIMALE VERSCHILDRUK OP EEN DAG (FLOW_MEAN) GEMIDDELDE VOLUMESTROOM (M3/H), (DRUK_NEVENTS) AANTAL ANOMALIEN, (FLOW_NEVENTS)) AANTAL ANOMALIEN VOLUMESTROOMREEKS.

5 Visualisatie in Lizard

5.1 Lizard

Binnen het project is Lizard gebruikt als middel voor online data visualisatie. Hierbij is gekeken hoe de data goed gepresenteerd kan worden om te helpen bij het zoeken naar verbanden tussen verschillende data. Lizard is een online platform voor dataopslag en visualisatie. Het platform bestaat uit drie onderdelen: een backend, een API en een viewer. De backend is de dataopslag, waar data gestructureerd wordt opgeslagen zodanig dat het gecombineerd kan worden. De dataopslag is ontwikkeld voor grote hoeveelheden data die real time binnen kunnen komen en is daarmee geschikt voor operationeel gebruik. De API is de Application Programming Interface die zorgt voor de ontsluiting van de database op een eenduidige manier, zodat ook andere programma's er mee kunnen communiceren. De viewer van Lizard is een webviewer waarin tijd en ruimte centraal staan om de data te bekijken. Hierdoor wordt het mogelijk om ruimtelijke beelden te combineren, zoals neerslag-informatie, waterstandsmetingen, drinkwaterleidingen, waterverbruik, enz.

Naast de viewer van Lizard is het mogelijk om andere viewers te koppelen aan de Lizard backend via de API. Ook kunnen andere pakketten zoals QGIS en ArcGIS gekoppeld worden om de data op te halen, te bewerken en terug te sturen.



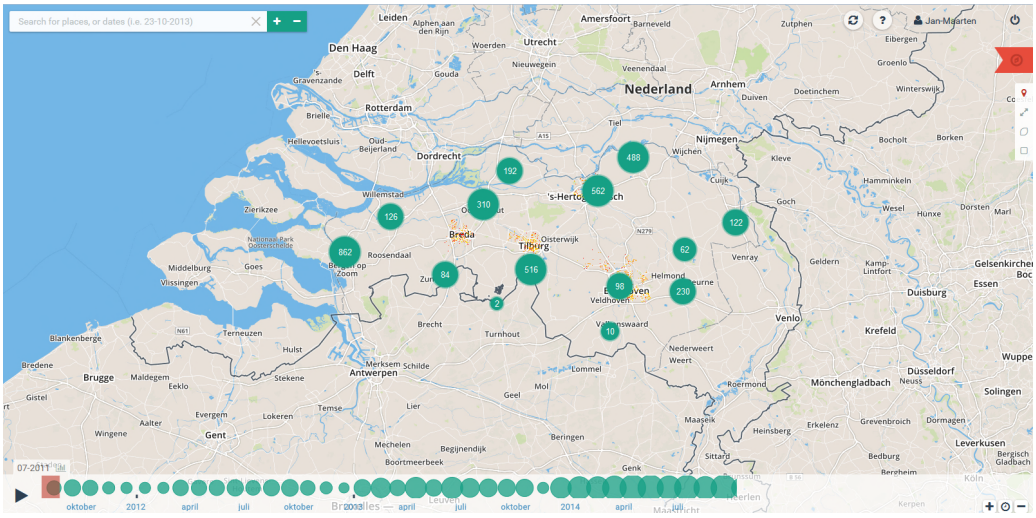
FIGUUR 21. ARCHITECTUUR LIZARD MET DE BACKEND IN LICHTGROEN, DE API IN DONKERGROEN EN DE VIEWERS IN BLAUW.

5.2 Visualisatie voor drinkwater

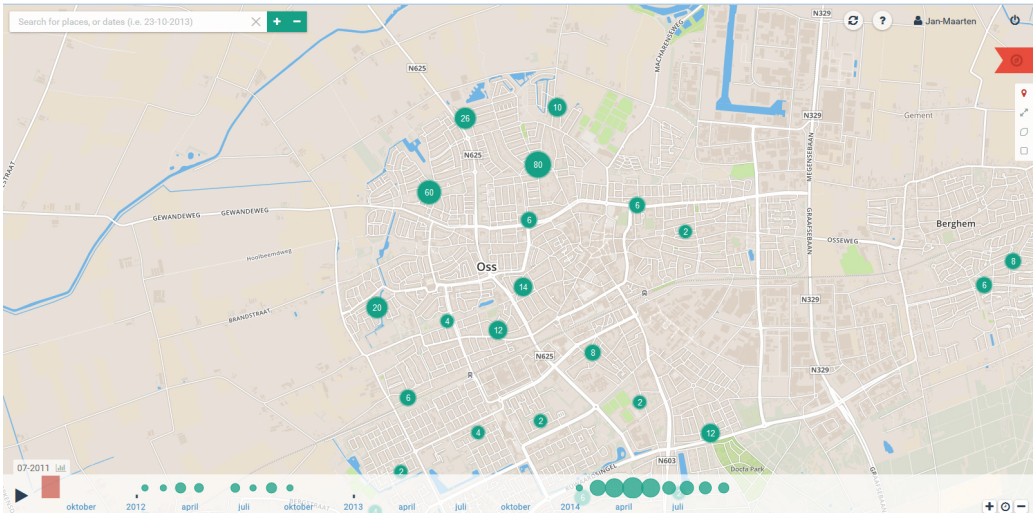
Voor dit project is gebruik gemaakt van de tijd-ruimte visualisatie van Lizard om geluidsmeldingen en waterverbruik geografisch te tonen. De geluidsmeldingen hebben allemaal een datum en een locatie en kunnen zo op de juiste plek weergegeven worden. Om het overzicht te krijgen, is een aggregatie-methodiek ontwikkeld zodat de meldingen bij een hoger zoomniveau geclusterd worden. Door op een 'cluster' te klikken zoom je in naar een lager niveau, waar weer clusters (maar dan op een lager niveau) aanwezig zijn. Hiermee is ruimtelijk een goed beeld te vormen waar de geluidsmeldingen plaatsvinden en kunnen 'hot spots' ontdekt worden. Deze visualisatievorm kwam bijvoorbeeld goed tot recht in het zoeken

naar geluidsmeldingen in de plaats Oss. De meldingen bleken grotendeels in één wijk te zijn gedaan.

Om ook zicht te krijgen op de spreiding van de meldingen door de tijd, wordt onderaan in de tijdbalk (Figuur 22) het aantal meldingen weergeven. De grootte van het bolletje in de tijdbalk wordt bepaald door het aantal meldingen in die tijdsperiode (maand, dag, uur; afhankelijk van zoomniveau) voor het gebied dat op het scherm zichtbaar is. Met deze visualisatie kunnen temporele effecten snel gezien worden.

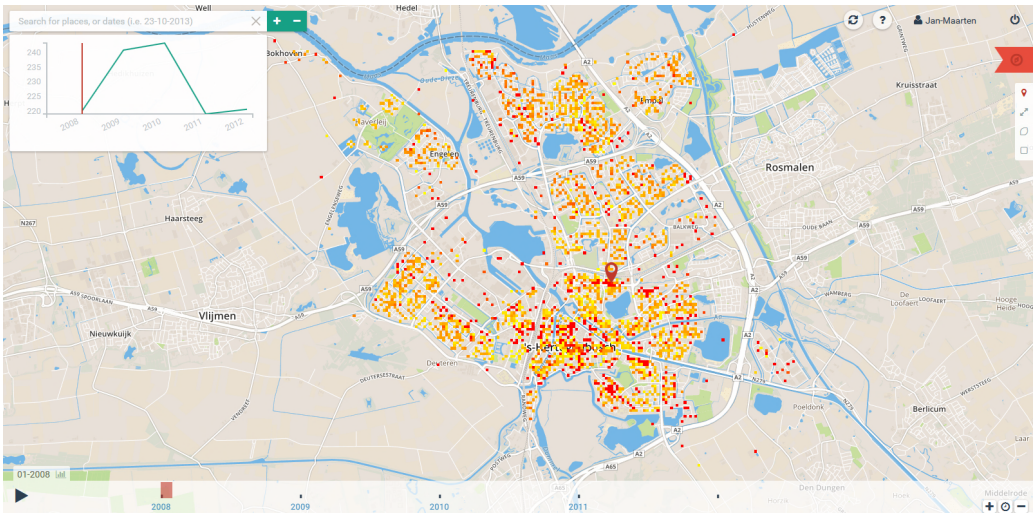


FIGUUR 22. WEERGAVE VAN GELUIDSMELDINGEN IN LIZARD OP EEN HOOG ZOOMNIVEAU, WAARBIJ MEERDERE MELDINGEN GECLUSTERD WORDEN GEVISUALISEERD. DE TIJDBALK ONDERIN GEEFT DE HOEVEELHEID MELDINGEN IN DE TIJD AAN.

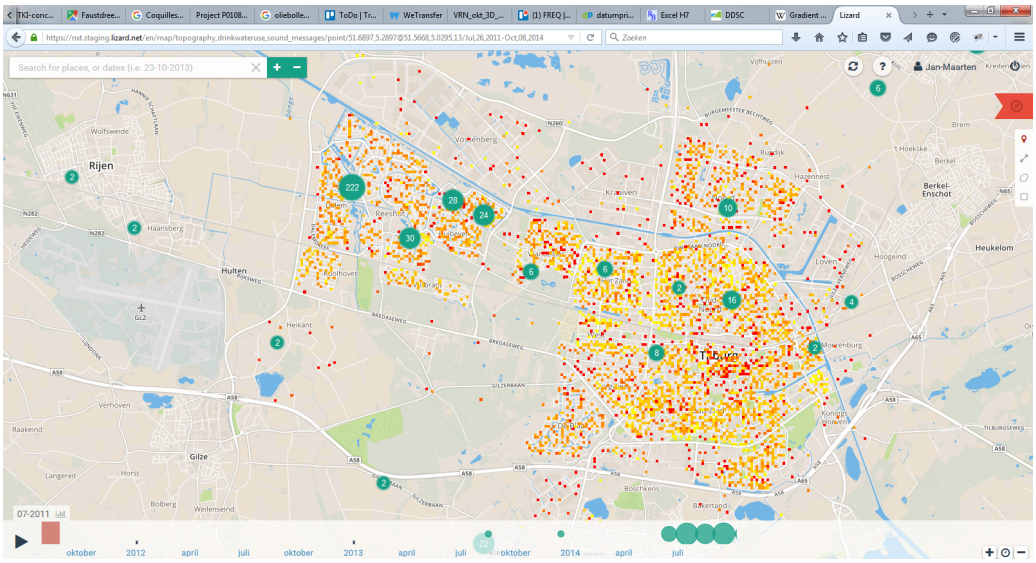


FIGUUR 23. WEERGAVE VAN GELUIDSMELDINGEN IN LIZARD OP EEN LAGER ZOOMNIVEAU ROND OSS. OPVALLEND IS DAT IN BEPAALDE WIJKEN EEN HOGERE MELDINGSDICHTHEID IS. DAARNAAST IS GOED ZICHTBAAR DAT IN 2013 GEEN MELDINGEN IN DE DATASET AANWEZIG ZIJN.

Voor het drinkwaterverbruik zijn jaarlijkse gegevens van 2008 tot en met 2012 beschikbaar gesteld op postcodeniveau. Deze data is bewerkt om tot geïnterpoleerde vlakdekkende rasters te komen (Figuur 24). Deze rasters kunnen worden afgespeeld om zo een animatie te vormen, waarbij het watergebruik over de tijd zichtbaar wordt. Door te klikken op een rasterpunt wordt het waterverbruik voor dat punt door de tijd inzichtelijk gemaakt. Zowel de geluidsmeldingen als het drinkwaterverbruik zijn losse kaartlagen die gecombineerd kunnen worden (Figuur 25). Ook kaartlagen als een digitaal hoogtebestand en landgebruik zijn aanwezig.



FIGUUR 24. WEERGAVE VAN WATERVERBRUIK IN LIZARD. DE GRAFIEK GEEFT HET WATERVERBRUIK VOOR DE JAREN 2008 TOT EN MET 2012 WEER.



FIGUUR 25. DE COMBINATIE VAN INTERACTIEVE KAARTLAGEN ZOALS WATERVERBRUIK EN GELUIDSMELDINGEN BIEDT KANSEN OM TE ZOEKEN NAAR MOGELIJKE RELATIES.

5.3 Ervaringen opslaan en visualiseren van drinkwatergegevens

In het project is een aantal ervaringen opgedaan rondom de inzet van data voor visualisatie in Lizard, namelijk:

- Het opvragen, duiden en verwerken van de data kost veel tijd. Bij de start van het project was de hoop dat makkelijk een grote hoeveelheid aan (real-time) data gevisualiseerd zou kunnen worden. In de praktijk blijkt het lastig om aan deze data (zoals klantmeldingen, druk- en volumestroommetingen, etcetera) te komen. Toestemming, de juiste personen die het kunnen aanleveren en een goede uitleg van de data blijken cruciaal te zijn om de data juist te kunnen verwerken.
- Het gekozen dataformaat/aggregatie niveau kan verbeterd worden. Bijvoorbeeld het waterverbruik werd aangeleverd op postcode niveau, wat een aggregatie is van de beschikbare data. Postcodegebieden zijn voor visualisatie in Lizard geen logische gebieden, omdat ze in grootte erg kunnen verschillen, soms niet aaneengesloten zijn en door de tijd heen kunnen variëren. In Lizard wordt gewerkt met rasters en daarom is de data verrasterd om rasterfunctionaliteiten te kunnen gebruiken, zoals animaties en dwarsprofielen. Als de data in de ruwe vorm beschikbaar wordt gesteld (of direct geaggregeerd naar gelijkmatige rasters), zullen de resultaten beter zijn.
- De visualisaties in Lizard hebben geholpen om inzicht te krijgen in de data en zijn aansprekend. Bij de demonstraties en overleggen waar Lizard is ingezet bleek dat het eenvoudig was om interactief te visualiseren waar bijvoorbeeld geluidsmeldingen waren. Dit leidde onder andere tot het vermoeden dat geluidsmeldingen gerelateerd waren aan bepaalde bouwjaren, waar vervolgens verder naar gekeken kon worden. Het visualiseren van gegevens in tijd en ruimte, biedt mogelijkheden richting te geven aan onderzoeksvragen.

6 Conclusies en aanbevelingen

6.1 Conclusies en aanbevelingen casus K

Om (mogelijke) invloedsfactoren op bruinwaterincidenten in het drinkwater-distributiesysteem te onderzoeken, is een correlatieanalyse uitgevoerd met 5 jaar aan klantcontactgegevens van het Brabant Water-voorzieningsgebied met 2,4 miljoen klanten. Dit onderzoek complementeert daarmee studies gericht op onderzoek naar specifieke oorzakelijke verbanden, gerelateerd aan fysische, biologische en/of chemische processen. Deze studie bevestigt dat meerdere factoren invloed hebben op het optreden van bruinwaterincidenten en meldingen daarvan. Op basis van de bevindingen worden de volgende conclusies getrokken:

- Een zwakke afhankelijkheid van demografische factoren demonstreert dat de patronen van klantmeldingen niet sterk onderhevig zijn aan demografische vertekeningen.
- De vermazingsgraad is slechts zwak gecorreleerd met zowel de bruinwater- als de referentiemeldingen. Een aantal aspecten kunnen hieraan ten grondslag liggen: (a) er is geen significante invloed van vertakking van het netwerk op bruinwatermeldingen, (b) het aantal meldingen per gebied is te laag voor statistische significantie, of (c) andere factoren hebben meer invloed. Ook is er geen sterke correlatie tussen bruinwater-gerelateerde meldingen en de dimensioneringsmaat.
- De resultaten tonen een statistisch significante, positieve afhankelijkheid van bruinwatermeldingen met de buitentemperatuur. Dit resultaat ondersteunt eerder onderzoek dat suggereert dat hoge temperaturen het optreden van bruin water bevorderen.
- Watermeters die zijn vervangen in verband met geluidsmeldingen komen relatief vaak voor in woningen en panden met een bouwjaar tussen 1995-2005. Het verdient nader onderzoek naar de oorzaak. Geluidsmeldingen zijn niet sterk gecorreleerd met socio-economisch factoren of leidingnetkarakteristieken (vermazingsgraad en dimensioneringsmaat).

Op basis van de conclusies en gevolgde methodiek kunnen wij de volgende aanbevelingen doen voor nader onderzoek. Hierbij wordt terugverwezen naar de 4 condities die in paragraaf 2.1.2 zijn benoemd:

- Om mogelijke hydraulische en/of microbiologische oorzaken voor de gevonden relaties met temperatuur en leidingdiameter te onderzoeken, wordt aanbevolen om de analyse uit te breiden met: opwervelingspotentiemetingen gecorrigeerd voor nabijgelegen spuiacties uit het verleden, dagelijkse maximumsnelheden als maat voor de lokale zelfreinigende capaciteit, schattingen van het meest nabije waterproductiebedrijf voor elke netwerklocaties, en data om (mogelijke) seizoensafhankelijkheden in bedrijfsoperatie, klantdemografie en watervraag te onderzoeken.
- Om meer inzicht te krijgen in de oorzaken van bruinwatermeldingen is verder gericht onderzoek nodig om de relatieve invloeden te onderscheiden van accumulatie van deeltjesmateriaal (stap 1) of opwerveling van materiaal als gevolg van hydraulische krachten (stap 2). Om meer inzicht te verkrijgen in accumulatie (stap 1), bevelen we aan om troebelheidsmetingen aan de analyse toe te voegen, omdat daarmee de mate van bruin water is te kwantificeren. De troebelheidsmetingen zouden idealiter moeten worden gecorrigeerd voor tijd sinds de laatste spuiactie, omdat het lokale bruinwaterrisico na een spuiactie in het algemeen toeneemt over de tijd. Om meer inzicht te krijgen in

opwerveling (stap 2), bevelen we aan om de zelfreinigende werking van een netwerk met gebruik van de nauw verwante dagelijkse lokale maximumsnelheid te karakteriseren en als invloedsfactor te onderzoeken. Deze snelheid kan worden berekend met de lokale maximum-stroming geschat met een vraagmodel op kleine tijdschaal.

6.2 Conclusies en aanbevelingen casus P

Er is een relatie gevonden tussen de storingsfrequentie en het drukregime van pompstations. Zowel de gemiddelde druk, maximale druk als de maximale verschildruk op een dag laten een toename in storingsfrequentie zien bij een toenemend(e) druk(-verschil). Dit verband met druk lijkt met name aanwezig bij CH leidingen. Er is geen direct verband gevonden tussen het optreden van anomalieën en storingen.

Op basis van onze bevindingen doen wij voor deze casus de volgende aanbevelingen:

- Het gebruikte algoritme voor anomaliedetectie biedt potentie, indien aan de volgende verbeteringen wordt gewerkt:
- expliciet vakanties en feestdagen meenemen in de regressie-component, teneinde beter het waterverbruikssignaal te kunnen voorspellen, in lijn met Bakker et al. (2013);
- rekening houden met seizoensinvloeden. Als extra variabelen kunnen de buitentemperatuur en de maand van het jaar meegenomen worden in het anomaliedetectiealgoritme. Het meenemen van deze extra parameters vereist dat langere meetreeksen gebruikt worden, omdat bij de huidige reeksen (~1 jaar) overfitting van het model zal optreden bij het meenemen van deze variabelen.
- nu is onduidelijk of een door het anomaliedetectie-algoritme aangemerkt ‘event’ ook daadwerkelijk iets belangrijks was, of dat het gaat om een vals alarm. Door de gedetecteerde events te labelen kunnen anomalieën in de toekomst accurater, dat wil zeggen: met minder vals-positieve hits, worden gedetecteerd.
- Er is aangenomen dat het leidingnetmodel volledig de werkelijkheid presenteert gedurende de periode waarin de PI-data is geregistreerd, inclusief bijvoorbeeld afsluiterstanden of vervangen leidingen. Wijzigingen aan het leidingnet in de werkelijkheid vanaf 2011 leiden ertoe dat het model niet geheel meer overeenkomt met de werkelijkheid. Beter zou zijn om elke mutatie van het leidingnet geregistreerd te hebben en per tijdslot het bijbehorende leidingnetmodel te gebruiken in de analyse.
- Door de gedefinieerde drukzones met modelberekeningen te bepalen en/of te valideren, wordt de analyse betrouwbaarder. Er kunnen bijvoorbeeld meer (en dus ook kleinere) drukzones gedefinieerd worden door ook drukmeetpunten in het leidingnet mee te nemen. In de huidige opzet waren deze metingen en meetlocaties niet beschikbaar. Het blijft van belang om per drukzone voldoende storingen geregistreerd te hebben.
- Tot slot is het aan te raden om bij vervolgonderzoek meer (van dezelfde) data te betrekken in de analyse. Nu was de beschouwde periode vanaf zomer 2014 tot zomer 2015. Door een langere periode te nemen is het mogelijk om enerzijds de anomaliedetectie te verbeteren door rekening te houden met seizoensinvloeden (zie eerste aanbeveling) en anderzijds meer storingen mee te nemen in de analyse (zie ook aanbeveling bij vorige bullet).

6.3 Samenvattende aanbevelingen

Met dit TKI-project is een verkenning uitgevoerd naar de mogelijkheden van datamining voor de bedrijfsvoering van Brabant Water. De verkenning heeft enkele inzichten opgeleverd, die resulteren in de volgende aanbevelingen.

Het hebben van een grote hoeveelheid data is zonder uitzondering essentieel bij de inzet van KDD. Verreweg de grootste belemmering in de huidige analyses is dat er, in verhouding tot het aantal verschillende variabelen (attributen), nog steeds weinig data per attribuut [4] beschikbaar is. Voor toekomstige data-analyses is het advies daarom:

- huidige registraties zoals bijvoorbeeld in USTORE, te continueren en uit te breiden naar observaties met betrekking tot operationele opvallende zaken (labeling anomalieën) als ook klantmeldingen;
- daarnaast aandacht te besteden in de opslag en beschikbaarheid van data (in bijvoorbeeld PI). Het adagium 'eerst meten, dan weten' is hier van toepassing: indien een fenomeen zich voordoet waarin men geïnteresseerd is, dan is het noodzakelijk om hier eerst meerdere observaties over te verzamelen. Met statistische methodieken, inclusief KDD, geldt bovendien dat het aantal metingen een veelvoud aan observaties[5] moet bevatten alvorens de data-analyse betrouwbaar genoeg wordt.

Projectmatig zit de grootste inspanning in het verzamelen, structureren en correct interpreteren van gegevens die nodig zijn voor de data-analyse. Vakspecialistische kennis is hierbij een noodzaak om geen fundamentele fouten te maken bij het koppelen van databronnen en om effectief te kunnen overleggen over de data. Omdat de inspanning 'aan de voorkant' van het data-analyseproces zo groot is maar toch inzichten kan opleveren die het belang van één waterbedrijf kunnen overstijgen, wordt aanbevolen om:

- datasets zoveel mogelijk uniform en bedrijfstakbreed te organiseren en beschikbaar te stellen. Hierdoor kunnen obstakels met betrekking tot weinig data (zie ook de vorige aanbeveling) worden weggenomen, en zijn ook onderlinge vergelijkingen tussen waterbedrijven, of validaties mogelijk;
- delen van data en efficiënte data-analyses vergen, naast een bedrijfstakbrede organisatie, consistentie in data-labeling en -coderingen. Dit laatste aspect dient eenduidige antwoorden te geven op de vragen: wat is er gemeten, waar is het gemeten en hoe is het gemeten? Dit is een behoorlijke opgave die het consistent gebruik van eenheden, eenduidige manier van meten en het duidelijk vastleggen van meta-informatie vergen. Data uit pilot-metingen (proeftuinen) en bedrijfstakbrede storingsregistraties zijn een eerste aanzet.

[4] Verhoudingsgewijs is de huidige data nog verre van 'big data' te noemen. Hiervan is pas sprake indien bijvoorbeeld grootschalige sensornetwerken hoogfrequente metingen zouden doorsturen naar een centraal punt, alwaar het opgeslagen en/of verwerkt zou worden. Het gaat dan om ordegrootte gigabytes aan data per seconde. Ter vergelijking: de totale PI-database die we voor deze analyses hebben gebruikt, met data van ongeveer 1 jaar, is 'slechts' 15 GB. De nuttige meetreeksen daarin (die voor deze analyse gebruikt zijn) hebben een omvang van ordegrootte 2 GB.

[5] bij statistische toetsen betreft het aantal observaties vaak in de ordegroote van honderd tot duizend, afhankelijk van kennis over de verdeling van data. Bij KDD wordt, door het gebruik van classificatie- en /of regressiemethodieken, het aanbevolen om een ordegrootte van tienduizend observaties of meer aan te houden.

7 Literatuur

Bakker, M., J. H. G. Vreeburg, K. M. Van Schagen, en L. C. Rietveld. 2013. A fully adaptive forecasting model for short-term drinking water demand. Environmental Modelling en Software 48:141–151.

Blokker, E. J. M., en E. J. Pieterse-Quirijns. 2013. Modeling temperature in the drinking water distribution system. Journal - American Water Works Association 105:E19–E29.

Blokker, E. J. M., en P. G. Schaap. 2015. Temperature influences discolouration risk. Procedia Engineering 119:280–289.

Blokker, E. J. M., J. H. G. Vreeburg, P. G. Schaap, en J. C. Van Dijk. 2010. The Self-Cleaning Velocity in Practice. Pages 187–199 WDSA2010. . ASCE.

Cook, D. M., P. S. Husband, en J. B. Boxall. 2001. Operational management of trunk main discolouration risk. Urban Water Journal:1–14.

Kooij, D. Van der. 2001. Heterotrophic Plate Counts en Drinking-water Safety. . World Health Organization, London.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, en A. H. Byers. 2015. Big data: The next frontier for innovation, competition, en productivity. . McKinsey Global Institute.

Mounce, S. R., R. B. Mounce, en J. B. Boxall. 2011. Novelty detection for time series data analysis in water distribution systems using support vector machines. Journal of Hydroinformatics 13.4:672–686.

Pedregosa, F., G. Varoquaux, A. Gramfort, A. Michel, B. V. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, en V. Dubourg. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830.

Prettenhofer, P., en G. Louppe. 2014. Gradient Boosting Regression Trees in Scikit-Learn. . PyData 2014.

Vonk, E. en D. Vries. 2015. Datamining voor assetmanagement - inventarisatie en voorbeelden uit de watersector. KWR - BTO 2016.007.

Vreeburg, J. H. G. 2007. Discolouration in drinking water systems: a particular approach. . Delft University of Technology, Enschede.

Vries, D., E. Vonk, W. de Jong, H. van Duist, P. van der Marel, J. van der Wielen. 2016. Herkennen van anomalieën in waterdata: demo in de Vitens proeftuin. KWR - BTO 2016.024.

Wols, B., J. Van Summeren, G. Mesman, en B. Raterman. 2016. Fysieke kwetsbaarheid leidingen voor klimaatverandering. KWR - BTO 2016.016.

Bijlage I Methodiek berekening vermazingsgraad en dimensioneringsmaat

Indicatoren leidingnetontwerp

Als onderdeel van dit project is gezocht naar eenvoudige indicatoren die een ruimtelijk beeld geven van het leidingnetontwerp en -topologie op lokaal niveau. Vanuit functioneel oogpunt is enerzijds het geometrisch ontwerp van een leidingnet kenmerkend (de zogenaamde vertakte of vermaasde ontwerpen). Anderzijds is ook de dimensionering een eigenschap die van belang is (het totale 'leidingvolume' aanwezig in bepaald een gebied), ten opzichte van het waterverbruik. Teneinde deze kenmerken ruimtelijk zichtbaar te maken zijn de volgende indicatoren ontwikkeld:

1. **Vermazingsgraad**
   De vermazingsgraad wordt bepaald door de verhouding tussen het aantal mazen dat aanwezig is in een bepaald gebied, ten opzichte van het aantal afnemers (aansluitingen) in datzelfde gebied. In formulevorm:
   $$VG = \frac{L - K + N}{A}$$
   In deze vergelijking is VG de vermazingsgraad (dimensieloos), L het aantal leidingsegmenten, K het aantal knopen, N het aantal 'deelnetwerken' (stukken leidingnet binnen een buurt die onderling niet met elkaar verbonden zijn) en A het aantal aansluitingen. De teller van deze vergelijking levert het aantal mazen binnen een bepaald gebied op, dat genormaliseerd wordt middels de aansluitingen.

2. **Dimensioneringsmaat**
   De dimensioneringsmaat (dimensieloos) is de verhouding tussen het totale volume aan distributieleidingen dat aanwezig is in een bepaald gebied, gedeeld door het gemiddelde uurverbruik van de afnemers in dat gebied. Een lage waarde voor de dimensioneringsmaat geeft aan dat er een relatief klein leidingvolume aanwezig is ten opzichte van het verbruik. Daardoor zijn relatief hoge stroomsnelheden van het water aannemelijk en zal de zelfreinigende functie goed zijn. Een hoge dimensioneringsmaat betekend dat er veel leidingvolume in een gebied aanwezig is om een relatief kleine hoeveelheid water bij de klanten af te leveren. Gemiddeld over een dag zullen de stroomsnelheden laag zijn. In formulevorm:
   $$DM = \frac{\sum_{i=1}^{N} {}^{1}/_{4}\pi {D_i}^2 L_i}{Q}$$
   In deze vergelijking is DM de dimensioneringsmaat voor een gebied, bepaald op basis van het aantal leidingsegmenten N in dat gebied. Daarbij is D_i de diameter van leidingsegment i (in m) en L_i de lengte van leidingsegment i (in m). Het leidingvolume wordt genormaliseerd met het totaalverbruik Q in het betreffende gebied (in m^3/jaar).

Beide indicatoren zijn interessant om eventuele correlaties aan te tonen met vuilwaterklachten. Daarnaast kan een dergelijke visualisatie patronen onthullen in 'gebruiksdruk' op het leidingnet. Tot slot is het hiermee mogelijk om structurele ingrepen op het leidingnet (implementatie streefstructuren, verschillen in ontwerpstrategieën) inzichtelijk te maken.

Databronnen

Voor het bepalen van deze leidingnetindicatoren is gebruik gemaakt van de Infoworks leidingnetmodellen van Brabant Water. Er zijn in totaal 15 voorzieningsgebieden, elk met een eigen model. Uit de leidingnetmodellen zijn de verbruiksgegevens en de leidingnetgeometrie geëxporteerd. De basisverbruiken in de modellen hebben als bron de verbruiksadministratie "Accent". Per model kan het 2010, 2011 en mogelijk 2012 zijn (dat is niet in alle rapportages even duidelijk), maar het grootste deel is afkomstig uit 2011.

Het basisverbruik geeft de ruimtelijke verdeling van de verbruiken over het gebied op basis van de gemeten verbruiken per afnemer. Vervolgens is de balans bepaald van het gebied en zijn de verbruiken vermenigvuldigd naar dat balanscijfer. Bijvoorbeeld: als in een gebied 900.000 m3/jaar verbruik aanwezig is volgens Accent (2011) en de balans sluit op 1.000.000 m3/jaar (2012), dan wordt het verbruik vermenigvuldigd met 10/9 waarmee de balansverschillen vereffend zijn. In de vereffening zitten alle administratiefouten, niet verrekend verbruiken (lekverliezen), meetfouten, verschil in afname tussen 2011 en 2012 etcetera.

Naast de Infoworks modellen is voor deze analyse gebruik gemaakt van de CBS 'Wijk- en buurtkaart 2011' en de CBS 'Provinciegrenzenkaart 2012'

Uitgangspunten

De volgende uitgangspunten liggen ten grondslag aan de bepaling van de leidingnetindicatoren (vermazingsgraad en dimensioneringsmaat):

- Gezien het beschikbare basismateriaal zijn de leidingnetindicatoren bepaald voor het jaar 2011.
- Leidingnetindicatoren zijn bepaald per CBS buurt. De 'buurt' is de meest fijne administratieve indeling (op hogere schaalniveaus bestaan er ook 'wijken' en 'gemeenten'). Het gebruik van een administratieve indeling is verkozen boven een indeling in geometrische rastercellen, aangezien wijken en buurten over het algemeen grenzen hebben die corresponderen met het onderliggende leidingnet. Tevens is een buurt en wijk tastbaarder dan een abstracte cel die willekeurig over het leidingnet heen gelegd wordt.
- Indien bij de bepaling van de dimensioneringsmaat alle leidingen uit het leidingnet worden meegenomen, dan kan dit zorgen voor een 'bias' in de uitkomsten. Er kan bijvoorbeeld een buurt zijn waar toevallig een grote transportleiding doorheen loopt, terwijl deze transportleiding zelf geen aantakkingen heeft op de afnemers in deze buurt. Om juist een beeld te krijgen van de dimensionering van het lokale distributienet (tertiair leidingnet) zijn alleen leidingen van kleiner dan rond 200 mm meegenomen.

Methodiek

Voorbewerking

1. *Exporteren uit Infoworks*
   De verbruikspunten en leidingsegmenten zijn uit Infoworks geëxporteerd naar zogenaamde shapefiles voor verdere analyse met ArcGIS.
2. *Samenvoegen data uit voorzieningsgebieden*
   De verbruikspunten en leidingsegmenten uit de 15 individuele leidingnetmodellen zijn samengevoegd tot één totaalmodel voor Brabant.
3. *Waterverbruik standaardiseren naar gemiddeld uurverbruik (m3/uur)*
   Afhankelijk van het verbruikspunt zijn de verbruiksgegevens op verschillende manieren vastgelegd. Alles is voor deze berekeningen gestandaardiseerd naar kubieke meter per uur.
4. *Opschonen leidingnetmodel*

Het Infowork leidingnetmodel bevat fictieve secties, met diameters 0, 888 of 1888 millimeter. Deze 'virtuele' secties worden gebruikt om (de hydraulica van) afsluiters, aanjagers en pompstations te kunnen modelleren. Voor deze analyse hebben deze secties echter geen nut. Ze zijn daarom verwijderd uit de dataset.

5. *Opknippen leidingnet op buurtniveau*
   In ArcGIS is het gehele leidingnet 'doorgeknipt' op de buurtgrenzen om zo een analyse op buurtniveau mogelijk te maken. Na het opknippen dienen de lengte en begin- en eindpunten van elk leidingsegment opnieuw berekend te worden.
6. *Multiparts converteren*
   Door Infoworks worden leidingen standaard als zogenaamde 'multipart features' geëxporteerd. Dit betekend dat elke leiding uit meerdere losstaande delen kan bestaan. Zo komt het voor dat leidingen uit delen bestaan die over elkaar heen liggen, waardoor het beginpunt identiek is aan het eindpunt. Ook zijn er gevallen waarbij een enkele leiding bestaat uit twee delen die onderling niet met elkaar verbonden zijn. In dergelijke situaties zijn de leidingen geautomatiseerd 'platgeslagen' (geconverteerd naar een enkele leiding) of gesplitst in twee individuele leidingen. Voor deze bewerkingsstap worden alle leidingeigenschappen gebruikt, met uitzondering van de codering voor de begin- en eindknoop waaraan de leiding in het oorspronkelijke leidingnetmodel gekoppeld was.
7. *Herberekenen geometrische eigenschappen leidingsegmenten*
   Na het converteren van de multiparts naar singleparts dienen de lengte en begin- en eindpunten van elk leidingsegment opnieuw berekend te worden.
8. *Genereren van netwerkknopen*
   Na het opknippen en opschonen van het leidingnet zijn voor alle eindpunten van leidingsegmenten knopen gegenereerd. Overlappende knopen (bij twee onderling verbonden leidingsegmenten) zijn samengevoegd.

Bepaling vermazingsgraad

Stappen:

1. *Sommeren knopen per buurt*
   Alle knopen, aangemaakt bij de voorbewerking, zijn opgeteld op buurtniveau.
2. *Sommeren aantal 'deelnetwerken' per buurt*
   Essentieel in de berekening van de vermazingsgraad is het corrigeren voor het aantal afzonderlijke netwerken in een buurt. Daartoe is per buurt het aantal 'losse' netwerken bepaald.
3. *Sommeren aantal leidingsegmenten per buurt*
   Optelling van het aantal afzonderlijke leidingsegmenten (na splitsing).

Bepaling dimensioneringsmaat

Stappen:

1. *Sommeren leidingvolume per buurt*
   Op basis van diameters en lengte van elk leidingsegment is het totale leidingvolume per buurt berekend. Alleen leidingen met een diameter kleiner dan 200 mm zijn meegenomen in deze berekening.
2. *Sommeren gemiddeld uurverbruik per buurt*
   Gemiddelde uurverbruiken van elk verbruikspunt zijn gesommeerd tot buurtniveau.

Er is gebruik gemaakt van Python scripting (ArcPy library) om alle hiervoor genoemde handelingen in ArcGIS te automatiseren. De resulterende indicatoren zijn gevisualiseerd door in ArcMap te kiezen voor het classificeren van elke buurt in 5 klassen, met daarbij een indeling gebaseerd op standaarddeviaties.

Bijlage II Methodiek definiëren drukzones

Uitgangspunt bij het definiëren van de drukzones is de standaardindeling van Brabant Water in voorzieningsgebieden. Deze voorzieningsgebieden geven reeds een globaal beeld van de druk die in een gebied heerst. Toch zijn voorzieningsgebieden nog te grof om direct als drukzone gebruikt te worden:

1. Diverse pompstations liggen zelf binnen een bepaald voorzieningsgebied, maar hebben een of meerdere takken die leveren aan aangrenzende voorzieningsgebieden.
2. Bepaalde voorzieningsgebieden hebben naast een primair waterproductiebedrijf (WBP) nog een aantal aanjagers of opjagers, die een duidelijke scheiding in drukregime veroorzaken binnen een voorzieningsgebied. Bij aanjagers is er daarnaast ook een duidelijke scheiding in druk tussen de zuig- en de perszijde.
3. Een aantal pompstations leveren in het geheel niet direct aan het leidingnet, maar alleen aan industriële gebruikers of alleen aan reinwaterkelders. Feitelijk zijn drukmetingen van deze pompstations dus niet relevant voor de druk in het leidingnet.
4. Voortvloeiend uit (1) en (2) kan een enkel voorzieningsgebied meerdere drukmeetpunten bevatten.

De drukzones zijn als volgt tot stand gekomen:

5. Pompstations met één of meer takken die direct leveren aan een aangrenzend voorzieningsgebied hebben feitelijk twee drukmeetpunten: een op de eigenlijke locatie van het pompstation en een op de plek waar de betreffende transportleiding aantakt op het aangrenzende voorzieningsgebied. In de meetpunten-dataset is voor dergelijke gevallen daarom een extra meetpunt aangemaakt op de plek van aantakking in aangrenzend voorzieningsgebied.
6. Indien een aanjager zich fysiek in een bepaald voorzieningsgebied bevindt, maar de uitstromende leiding uitkomt in een ander voorzieningsgebied, dan is deze aanjager in het model verplaatst naar het toeleverende gebied. Dit is gedaan voor aanjager Oudgastel (verplaatst van voorzieningsgebied Seppe naar Wouw).
7. Aanjagers zijn gesplitst in twee verschillende meetpunten: één punt voor de zuigzijde en één meetpunt voor de perszijde.
8. Pompstations die alleen leveren aan industriële gebruikers of reinwaterkelders van een ander pompstation zijn verwijderd uit de meetpunten-dataset (Schijf en Zevenbergen).
9. Bij pompstations met meerdere reinwatertakken, die allemaal leveren aan het omliggende leidingnet, is de drukreeks van de eerste tak gebruikt als representatief voor het omringende gebied.
10. Daar waar op basis van kennis binnen KWR bekend is hoe de begrenzing van drukzones in het leidingnet lopen zijn deze handmatig ingetekend. Het gaat hierbij bijvoorbeeld om de begrenzing tussen pers- en zuigzijde van aanjagers of specifieke drukzones zoals de regio Waalwijk, waar vanwege frequent voorkomen van storingen een afwijkend drukregime heerst.

11. In alle overige gevallen zijn drukzones gegenereerd door zogenaamde Thiessen-polygonisatie toe te passen. Dit is een wiskundige standaardmethodiek om de geometrische invloedssfeer van punten op een oppervlak te bepalen.

Van enkele locaties zijn geen metingen bekend. Meetlocaties zonder druk/volumestroom-data zijn:

- 146-Aanj Dongen-DrinkwaterDrukZuigzijde
- 135-Macharen-MacharenOss
- 104-Vessem-VessemEindhoven
- 152-Aanj Waalwijk-DrinkwaterDrukPerszijde
- 109-Lieshout-LieshoutSenH1
- 146-Aanj Dongen-DrinkwaterDrukPerszijde
- 148-Aanj Oud Gastel-DrinkwaterDrukZuigzijde
- 140-Tilburg-TilburgTilburg1
- 142-Aanj Veldhoven 1-DrinkwaterDrukPerszijde
- 110-Someren-SomerenPeelkant1
- 148-Aanj Oud Gastel-DrinkwaterDrukPerszijde