BTO 2018.019 | Maart 2018

BTO rapport

Sensornetwerken in het distributienet dragen bij aan het vergroten van de systeemkennis

KWR Watercycle Research Institute

BTO 2018.019

Sensornetwerken in het distributienet dragen bij aan het vergroten van de systeemkennis

BTO 2018.019 | Maart 2018

Opdrachtnummer
40554-215

Projectmanager
Stefan Kools

Opdrachtgever
BTO - Thematisch onderzoek - Nieuwe meetmethoden en sensoring

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

Auteur(s)
dr.ir. D. (Dirk) Vries, dr. J.R.G. (Joost) van Summeren

Verzonden aan
Dit rapport is verspreid onder BTO-participanten en is na 12 maanden openbaar

Jaar van publicatie
2018

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

Keywords

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

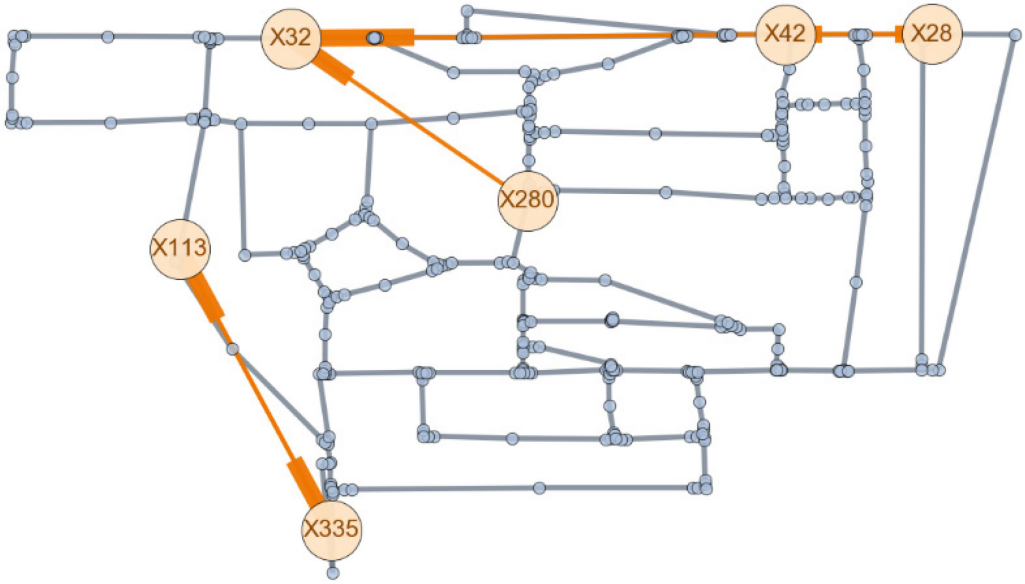
KWR Watercycle Research Institute

BTO *Managementsamenvatting*

Sensoren plus een innovatieve methodiek vergroten de systeemkennis van een bemeten leidingnetwerk

Auteur(s) dr.ir. D. (Dirk) Vries, dr. J.R.G. (Joost) van Summeren

In potentie zijn sensoren in het leidingnet bruikbaar om behalve het detecteren van incidenten ook andere informatie over het onderliggende systeem te verkrijgen. Dat blijkt uit een studie waarmee met slimme algoritmes en computersimulaties is getest in hoeverre analysetechnieken in staat zijn continu de toestand van het modelsysteem in realtime te ijken. Gekeken is naar het waarnemen van afwijkingen zoals dichte afsluiters die in werkelijkheid openstaan, of een kalibratiefout van een sensor. In dit onderzoek zijn de prestaties van een algoritme op niet-geïdealiseerde sensorgegeven met experimenten in een proefinstallatie van Vitens getest.



Leidingnet (blauwe lijnen) met sensorlocaties (cirkels), waarin relaties tussen sensorsignalen met het algoritme zijn berekend (oranje verbindingslijnen: verdikkingen geven oorzaak-gevolg aan, bij twee verdikkingen in één lijn is dit verband onbepaald)

Belang: beter inzicht in onderliggend systeem dankzij sensornetwerken

Waterbedrijven baseren beslissingen met betrekking tot kennis van het distributiesysteem op gegevens die in de praktijk niet altijd correct, volledig en actueel zijn. Dat komt bijvoorbeeld door modelfouten in afsluiterstanden, stroompatronen, of door sensoren die storen. Ontwikkelingen in sensortechnieken en algoritmes bieden hiervoor mogelijk een oplossing, omdat zij steeds meer zicht bieden op realtime situaties. Naast verstoringen in waterkwaliteit en waterlevering, kan sensorinformatie ook informatie geven over het onderliggende systeem. Realtime analysetechnieken maken het mogelijk afwijkingen ten opzichte van veronderstelde kennis te identificeren. In een voorgaand BTO-project (KMTS, Van Summeren et

al., 2016) is de basis van dit idee uiteengezet met een informatietheoretisch raamwerk. Dit onderzoek bouwt daarop voort.

Aanpak: toetsing prestatie realtime sensoren met numerieke simulaties en een proefinstallatie

Op basis van virtuele sensorgegevens zijn twee benaderingen uit de literatuur uitgewerkt en getoetst. Berekeningen zijn gedaan met een numeriek model van een leidingnet. Drie leveringsscenario's zijn gesimuleerd: één met een reguliere levering, en twee waarin respectievelijk een dichte afsluiter of een onjuiste sensormeting voor afwijkingen zorgden. Na een eerste toetsing werd de meest veelbelovende benadering geselecteerd. Vervolgens zijn de prestaties van een algoritme getest op niet-geïdealiseerde sensorgegeven met experimenten in een proefinstallatie: een schaalmodel van Leeuwarden en omgeving, waarin verschillende watertypes zijn nagebootst aan de hand van gedoseerde markeerstoffen. Vitens stelde de proefinstallatie beschikbaar en hielp bij uitvoering van een experiment. Meting van de waterkwaliteit gebeurde met realtime sensoren voor het elektrisch geleidend vermogen.

Resultaten: algoritme herkent met sensorgegevens afwijkingen van normale operatie

De eerste resultaten met de meest veelbelovende methodiek zijn positief: uit realtime sensoren in het leidingnet kan de informatiestroom en -richting worden herleid. Uit de numerieke toetsing blijkt dat het geselecteerde algoritme bij voldoende data, afwijkingen van normaal functioneren (foutieve afsluiterstand of sensormeting) herkent. De experimenten boden vooralsnog te weinig data voor betrouwbare uitspraken. Voor een grondige evaluatie van toepassing van het algoritme in de praktijk zijn meer experimenten in een testomgeving nodig.

Implementatie: ondanks veelbelovende resultaten realtime sensor zijn meer testen nodig

Slechts enkele stappen in de ontwikkeling van de methodiek zijn nodig om verbeterde kennis over het leidingnet én de sensoren te bewerkstelligen. Toepassing van het huidige algoritme werkt als voldoende gegevens met een bepaald patroon beschikbaar zijn over verblijftijden in het leidingnet. Alleen dan kunnen uit realtime sensorgegevens automatisch de informatiestromen en -richtingen afgeleid worden. Met deze kennis kan de juistheid van een bestaand model gecontroleerd worden, een inschatting van de werkelijkheid gemaakt worden of sensorfouten opgespoord worden. Het is mogelijk om het algoritme aan te passen zodat het geschikt is voor een scala aan meetreeksen met 'dynamische' patronen. Door verdere ontwikkeling van de methodiek kan de aanpak geschikt worden gemaakt voor het opschalen naar grote sensornetwerken. Daarnaast bestaat de behoefte om de prestaties van de techniek te vergelijken met alternatieven.

Tenslotte wordt het gebruik van een proefinstallatie aanbevolen om reguliere en extreme omstandigheden snel en veilig te kunnen testen.

Rapport

Dit onderzoek is beschreven in rapport *Sensornetwerken in het distributienet dragen bij aan het vergroten van de systeemkennis* (BTO 2018.019).

Meer informatie
dr. Ir. Dirk Vries
T +31 30 606 9671
E dirk.vries@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

KWR Watercycle Research Institute

Inhoud

1 Inleiding

1.1 Aanleiding

Hoewel het ondergrondse netwerk lastig is te monitoren bieden sensornetwerken van waterkwaliteit, druk en volumestroom in combinatie met innovatieve data-analysetechnieken steeds meer mogelijkheden voor (realtime) inzicht in de kwaliteit van het water en kennis van het systeem. Het laatste aspect is van belang omdat in de praktijk de kennis van het distributienet niet altijd correct, volledig en actueel blijkt te zijn. Denk daarbij aan fouten in afsluiterstanden, stroompatronen, lekken, etc. Daarnaast is bekend dat sensoren ook niet perfect de werkelijkheid meten (fout-positieven, fout-negatieven, sensordrift, etc.). Veel veronderstellingen over de werking van het systeem zijn wél gebaseerd op deze informatie: dit zijn bijvoorbeeld aannames over de topologie en het functioneren van het systeem enerzijds, of over de kwaliteit van metingen anderzijds. Dit gebeurt zowel direct als via modellen waarin deze gegevens zijn verwerkt. Voor een betrouwbaardere bedrijfsvoering is het daarom belangrijk om te zoeken naar een methodiek om de kwaliteit van deze informatie (en de modellen die hierop worden gebaseerd) systematisch te verbeteren.

Hoewel er reeds enige tijd waterkwaliteitssensoren voor drinkwaterdistributienetwerken op de markt zijn, vindt grootschalige toepassing hiervan nog slechts op beperkte schaal plaats. Kennis van de betreffende parameters wordt op zichzelf als waardevol beschouwd, maar het blijft onduidelijk hoe de opbrengsten van een netwerk van dergelijke sensoren in de bredere zin in de bedrijfsvoering zich verhouden tot de kosten van een dergelijk netwerk (Van Thienen et al., 2015).

Daarnaast stelt het toepassen van grotere hoeveelheden data in operationele beslissingen eisen aan zowel de kwaliteit van de data als aan die van de gegevens die worden gebruikt bij de interpretatie van de data (al dat niet met behulp van modellen). In het eerste geval kan men bijvoorbeeld denken aan het optreden van fout-positieven en fout-negatieven; bij het tweede punt bijvoorbeeld aan de topologie van het distributienetwerk, standen van afsluiters, etc. en de representatie hiervan in modellen.

De toepassing van netwerken van waterkwaliteitssensoren is veel gemakkelijker kostenefficiënt te maken, wanneer een sensornetwerk tegelijkertijd meerdere doelen kan dienen (Kroll en King, 2010). Een voorbeeld hiervan is de monitoring van wettelijk voorgeschreven of voor klanten relevante waterkwaliteitsparameters gecombineerd met bewaking op incidenten (besmettingen) in het distributienet. In het voorgaande project *Kostenefficiënte Meervoudige Toepassing van Sensornetwerken* (KMTS, Van Summeren et al., 2015) is voorgesteld in welke mate een dergelijke combinatie met de huidige beschikbare technologie te realiseren valt. Bovendien is er een basis gelegd voor benaderingen om, naast 1) monitoring van de waterkwaliteit, ook 2) de kennis van het distributiesysteem te vergroten en 3) mogelijk het vertrouwen van de klant te behouden dan wel te vergroten.

Het eerste punt is voldoende ontwikkeld en uitgetest, implementatie vond via het TKI-project INTEREST plaats (Van Summeren, 2016). Het derde punt, behoud dan wel vergroting van het klantvertrouwen, is binnen KMTS een punt van discussie gebleken. Op basis van de bevindingen is nader onderzoek met een sociaalwetenschappelijke basis aanbevolen. Het tweede punt, vergroten van de systeemkennis, wordt als nuttig gezien omdat het kan leiden tot:

- een betere identificatie van afwijkingen in het functioneren van onderdelen van de infrastructuur en/of;
- een betere identificatie van afwijkingen ten opzichte van veronderstelde kennis van het systeem (foute afsluiterstanden, lekken, foute sensormetingen, verkeerd geregistreerde leidingeigenschappen, etc.)

en daardoor de inzet (en baten) van sensoren breder toepasbaar kan zijn. In het KMTS-project is de basis van een theoretisch raamwerk voor het vergroten van de systeemkennis gepresenteerd. De aanbeveling in dit project was om dit verder te ontwikkelen en te gaan toepassen, zodat de genoemde opbrengsten concreet worden gemaakt en worden gekwantificeerd.

1.2 Doelstelling

Het doel van het huidige project is deze verdere ontwikkeling voor het vergroten van systeemkennis te realiseren en de methodiek te toetsen. Een centrale onderzoeksvraag is hoe aanvullende informatie over fouten, inconsistenties en nauwkeurigheden in een meet- en distributiesysteem kan worden verkregen uit sensormetingen middels een systeemtheoretische benadering. Specifiek gaat het om afwijkingen ten opzichte van een reguliere situatie als gevolg van een inconsistente afsluiterstand en een foutieve sensormeting. Ten behoeve van de beheersing en verifieerbaarheid van experimenten is besloten om de methode toe te passen op het schaalmodelnetwerk van de VIP in het Vitens Innovation Center (Meeting of Waters). Hiermee beoogt het project een opstap te vormen naar een pilot in een echt distributienetwerk.

Op de mogelijke route naar slimme netwerken –waarin met een hoge dichtheid wordt gemeten en gestuurd wordt op diverse parameters– is de implementatie van dergelijke methoden om de systeemkennis te vergroten een logische stap. In elke situatie waarin overwogen wordt een sensornetwerk voor een specifiek doel aan te leggen (bijvoorbeeld event detectie), kan een dergelijke methodiek extra opbrengsten genereren, en daarmee de waarde van het netwerk vergroten en het doel verbreden.

1.3 Aanpak en leeswijzer

Voor de uitwerking van een methodiek om de systeemkennis te vergroten, is uitgegaan van benaderingen uit de systeem- en informatietheorie. In het bedrijfstakonderzoek is dit nieuw terrein. De bestaande mogelijkheden zijn geëvalueerd op basis van numerieke testen. Voor deze uitwerking is gebruik gemaakt van virtuele waterkwaliteit- en -kwantiteitgegevens, bepaald met een hydraulisch rekenpakket. De meest geschikte van de twee onderzochte benaderingen is getoetst met praktijkdata, gegenereerd met het schaalmodel van Leeuwarden in het Vitens Innovation Center. De bovengenoemde stappen (uitwerking en werking van de methodiek) zijn beschreven in Hoofdstuk 2. De resultaten van de twee benaderingen, d.w.z. hoe deze presteren op de virtuele en praktijkdata zijn beschreven in Hoofdstuk 3. Hoofdstuk 4 bevat de conclusies en aanbevelingen.

2 Aanpak

2.1 Het gebruik van grafen

Voor de methode maken we gebruik van grafentheorie gericht op het inzichtelijk maken van de informatiestromen. Een graaf is een verbinding van knooppunten met lijnstukken, een gerichte graaf een set van knooppunten en verbindingen waarbij de verbindingen een richting hebben. De eigenschappen van een graafmodel maken het bij uitstek geschikt voor het modelleren van sensornetwerken. Een voorwaarde van beide methodieken is dat er geen terugkoppellussen in het netwerk zijn (Spirtes en Clark, 1996). Bij terugkoppellussen is de gemeten informatie afhankelijk van een andere informatiebron (sensor) en de eigen informatiestroom. Dit is voor waterleidingnetwerken een plausibele aanname omdat een waterpakketje doorgaans niet meer dan een keer bij dezelfde plek komt. De methode is ontwikkeld en vervolgens met een leidingnetmodel en een proefmodelnetwerk die het leidingnet in en rondom Leeuwarden representeert, gesimuleerd. Het leidingnetmodel wordt in Paragraaf 2.4 verder toegelicht.

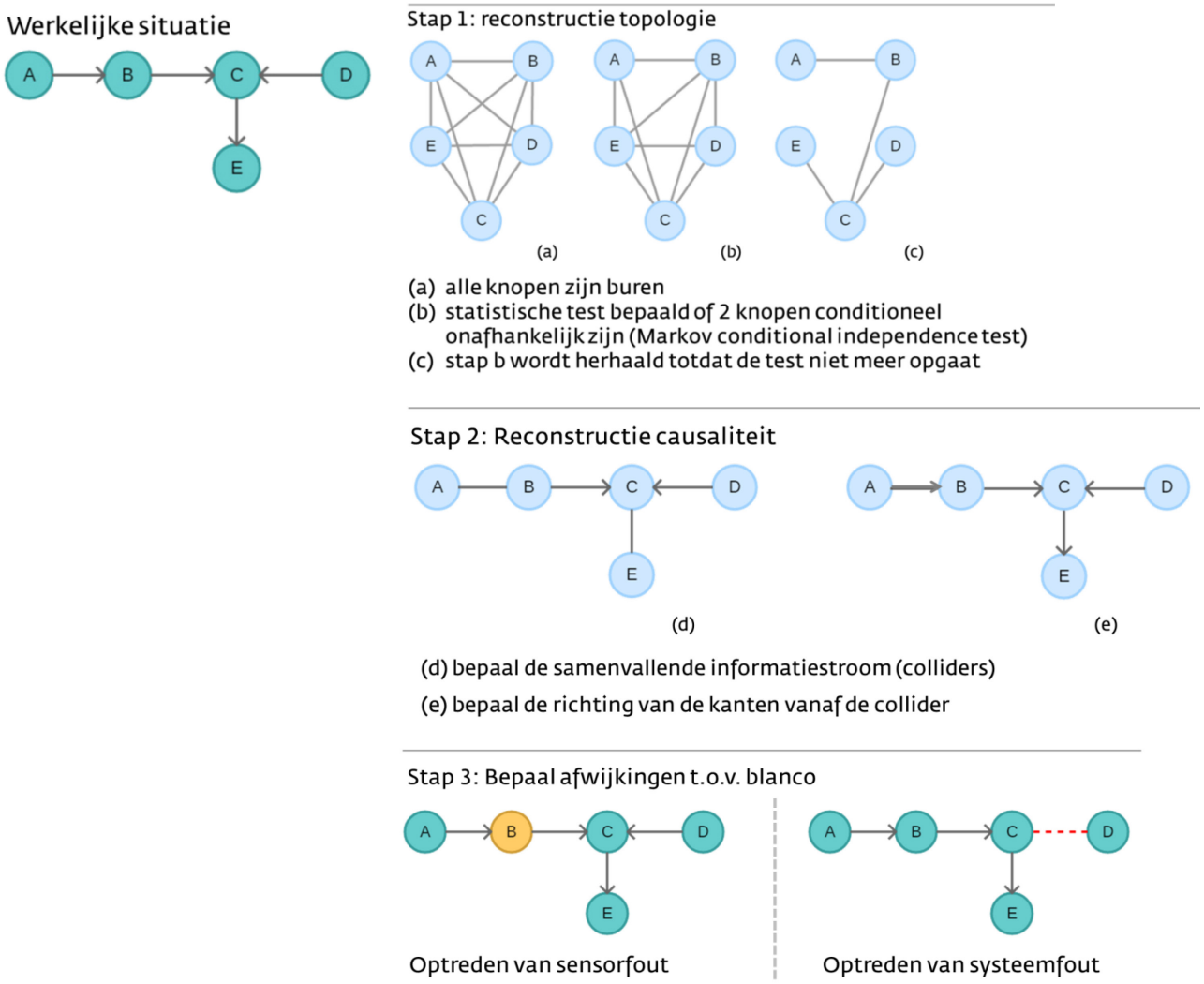
2.2 Heuristische aanpak

We richten ons op een heuristische benadering om (veranderingen in) de informatiestromen van sensoren in een (drink)waternetwerk in kaart te brengen en de methodiek te toetsen. Met een heuristische methode kan een oplossing numeriek worden benaderd voor problemen waarvoor geen exacte oplossing bestaat of niet eenvoudig deterministisch (denk aan bijvoorbeeld een tracing-methode) gevonden kan worden. De gekozen aanpak kenmerkt zich door twee aannames:

1. Alleen meetgegevens van waterkwaliteitssensoren zijn beschikbaar, m.a.w. we maken geen gebruik van bestaande (hydraulische) netwerkmodellen om de informatiestromen in kaart te brengen;
2. De informatie afkomstig uit sensoren (waterkwaliteits- en/of waterleveringsgegevens) volgen een ‘oorzaak-gevolg’ pad. Een signaal van een sensor kan daarom relateren met signalen van een of meerdere sensoren elders in het netwerk. Dit hoeft niet per sé een oorzakelijk, fysisch verband te zijn, maar ook een transportverschijnsel. Zo kost het in een waterleidingnet enige tijd voordat een waterpakket de route van (sensor)punt A naar (sensor)punt B heeft afgelegd, waarbij informatiestroom B zal correleren met A.

De methodiek voor het vergroten van de systeemkennis is opgebouwd in een drietal stappen (Figuur 2-1):

1. Graafreconstructie van het bestaan van verbanden tussen meetlocaties;
2. Toetsen van de causaliteit in alle mogelijke verbindingen van de graaf, op basis van topologie en correlatie van elk sensortriplet. De (statistische) toetsing is uitgevoerd met een 5%-onzekerheidsdrempel;
3. Herleiden van fouten of onzekerheden als gevolg van:
    a. veranderingen in de waterinfrastructuur (bijv. andere afsluiterstanden of lekkende leidingen);
    b. veranderingen in sensoren (vervuilde sensor/sensordrift, foutieve sensorkalibratie).



FIGUUR 2-1. CONCEPTUELE VOORSTELLING VAN DE METHODIEK. IN DE "WERKELIJKE SITUATIE" (LINKS) ZIJN 5 SENSOREN WEERGEGEVEN ALS BLAUWE CIRKELS (A-E) MET ONDERLINGE CAUSALE RELATIES (PIJLEN). DE METHODIEK BEVAT 3 STAPPEN (ZIE DE TEKST).

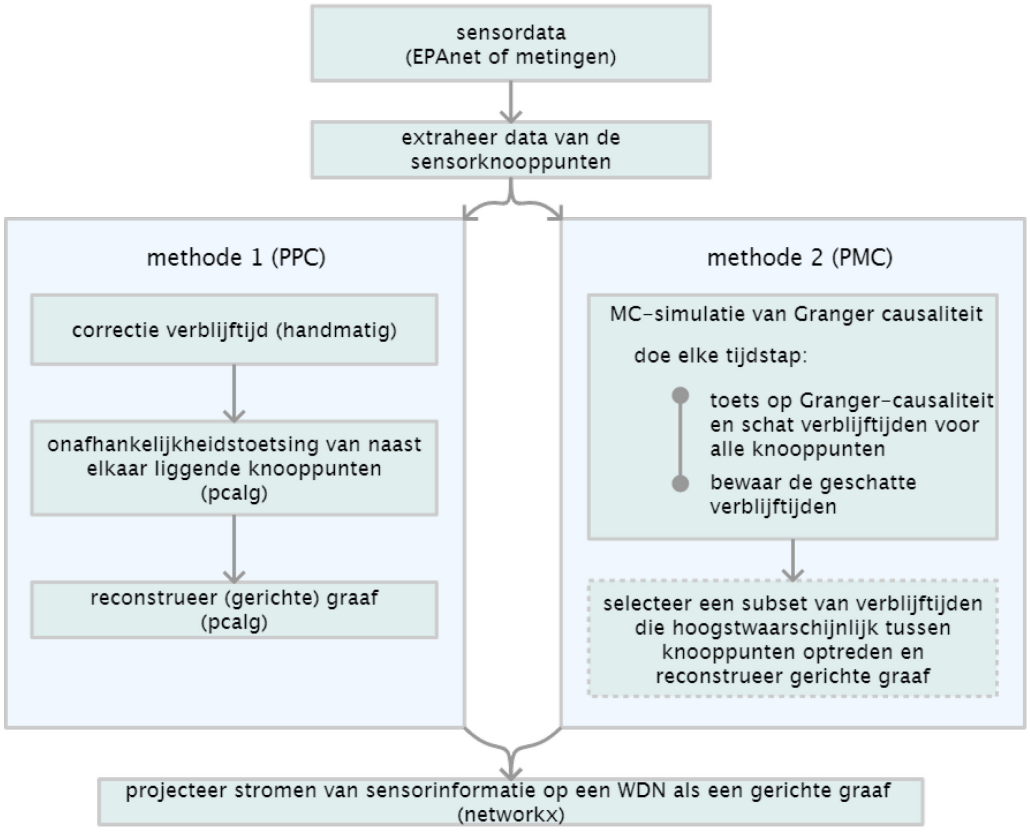
In stap 1 en 2 zijn twee methodes onderzocht:

- *Protocol voor de PC-routine (PPC)*: een Bayesiaanse[1] aanpak gebaseerd op het PC-algoritme (Spirtes en Clark, 1996; Kalisch en Bühlmann, 2007). Bij dit algoritme is het resultaat een zogenaamd 'Bayesian Belief Network'. Dit wel zeggen dat elke connectie in de graaf gebaseerd is op een (Bayesiaanse) kans van connectiviteit en causaliteit.
- *Protocol met MC-simulatie:* een Monte Carlo-simulatie waarin elke combinatie van knooppunten wordt getest op een statistisch causaliteitsprincipe. Bij PMC is elk lijnstuk

[1] De stelling van Bayes is een regel uit de kansrekening die ervan uitgaat van de kans van een bepaalde mogelijkheid die ten grondslag ligt aan een gebeurtenis (zeg P(A|B), m.a.w. de kans P voor een gebeurtenis A gegeven B) en wordt uitgedrukt in de voorwaardelijke kansen op de gebeurtenis bij elk van de mogelijkheden P(A), P(B) en P(B|A).

van een graaf gebaseerd op een causaliteitsprincipe uit de statistiek (Granger's causaliteit[2]), waarbij gewerkt wordt met een onzekerheidsdrempel.

De routines zijn schematisch weergegeven in Figuur 2-2 en gedetailleerd omschreven in Bijlage I. De routines zijn geprogrammeerd in een Python-raamwerk met functie-aanroepen van het statistisch softwarepakket R en het hydraulische simulatiemodel EPAnet, zie Bijlage II.



FIGUUR 2-2: GEVOLGDE AANPAK OM DE INFORMATIESTROMEN TUSSEN SENSOREN IN KAART TE BRENGEN.

2.3 Scenario's

De methodiek is met simulaties getoetst op twee praktische toepassingen, d.w.z. (C1) detectie van gewijzigde afsluiterstanden en (C2) detectie van foutieve sensormetingen en vergeleken met een situatie waarin er een 'foutloze' werkelijkheid is. Deze situatie wordt het basisscenario genoemd. Het basisscenario (B) is met het model en een experiment gesimuleerd. De scenario's B, C1 en C2 zijn voor beide algoritmen (PPC en PMC) numeriek getest. De resultaten staan beschreven in Vries en Van Summeren (2017), zie ook Bijlage I. Uit deze resultaten blijkt dat de heuristische PPC-procedure (nog) niet robuust genoeg is om causaliteit aan te tonen. Op basis van deze testen is besloten om alleen de prestaties van de PPC-procedure met werkelijke metingen te testen. Alleen het basisscenario is met een experiment nagebootst.

Een overzicht van scenario's is weergegeven in Tabel 2-1.

[2] Granger's causaliteitsprincipe zegt dat X een Granger-oorzaak is van Y, als Y beter voorspeld kan worden door de geschiedenis van X én Y te gebruiken dan door de geschiedenis van alleen Y te gebruiken (zie bijvoorbeeld Dahlhaus en Eichler, 2003).

TABEL 2-1: OVERZICHT VAN SCENARIO'S

Afkorting	Situatie	Aantal sensoren	Sensordata	Metingen in proefnetwerk	Numerieke simulatie	Hoofdstuk
B-CCWI	Basis-scenario	6	x	-	x	H. 3*
B-L	Basis-scenario	3	x	x (EGV)	x	H. 3
C1	Detectie van gewijzigde klepstand	6	x	-	x	H. 3*
C2	Detectie van sensorfout	6	x	-	x	H. 3*

** De volumestromen zijn wel berekend, maar niet gebruikt door het algoritme.*

*** Deze resultaten staan ook in Bijlage I.*

2.4 Proefinstallatie

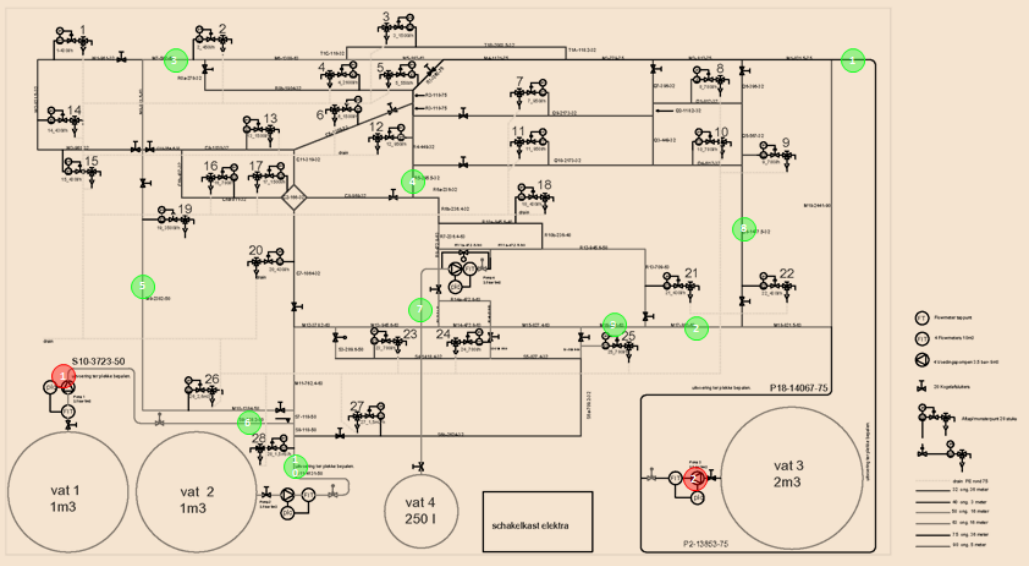
Voor het genereren van praktijkgegevens om de methodiek te toetsen, zijn metingen verricht op een door Vitens beschikbaar gestelde proefinstallatie. Deze installatie is ontwikkeld en gebouwd binnen het speerpuntonderzoek van Vitens. Voor een gedetailleerde beschrijving van de werking wordt de lezer verwezen naar de bijbehorende literatuur: Van Summeren et al., 2014; Van Summeren & Meijering, 2015; Van Summeren et al., 2017. In deze paragraaf wordt volstaan met de relevante hoofdzaken.

Het betreft een schaalmodel van ca. 4 m x 7 m van het distributienet in en rond Leeuwarden (Van Summeren et al., 2017). De leidingdiameters en -lengte en de tijdschaal zijn geschaald zodat snel experimenten zijn uit te voeren. De schaling is zodanig uitgevoerd dat de hydraulische condities (Reynolds-getal, schuifspanningen langs leidingwanden en turbulente diffusie) zo goed mogelijk de werkelijk situatie representeren.

FIGUUR 2-3. PROEFINSTALLATIE IN DE PROEFHAL VAN VITENS TE LEEUWARDEN.

Het leidingwerk in de proefinstallatie bestaat uit transparante PVC-buizen voor de grotere(300+ mm) leidingen in het gebied. Zoals te zien in de bouwtekening (Figuur 2-4)

bevat de installatie o.a. 3 voedingslocaties, 28 verbruikslocaties, 2 doseersystemen voor het toevoegen van markeerstoffen (zoutoplossing en kleurstof). Er zijn 3 realtime sensoren geplaatst voor het meten van het elektrisch geleidend vermogen (EGV). Deze “langsstroom”-sensoren zijn in de waterstroom geplaatst en meten de EGV zonder tijdsvertraging. De bewegende onderdelen (pompen, regelbare afsluiters bij de tappunten) worden aangestuurd via een centrale bediening.



FIGUUR 2-4. BOUWTEKENING VAN DE PROEFINSTALLATIE. DE GROENE CIRKELS GEVEN POTENTIËLE LOCATIES VAN SENSORAANSLUITINGEN OP DE LEIDINGEN WEER, DE EGV-SENSOREN ZIJN GEÏNSTALLEERD OP LOCATIES 1, 4 EN 6. DE RODE CIRKELS GEVEN DE 2 DOSEERSYSTEMEN WEER. VATEN 1 EN 2 REPRESENTEREN DE VOEDINGSLOCATIES WIRDUM, VAT 4 DE REINWATERKELDER GREUNS (INACTIEF IN DE SIMULATIES) EN VAT 3 DE VOEDINGSLOCATIE NOARDBURGUM.

Deze maakt het mogelijk om tijdsafhankelijke patronen voor de verbruikspunten en doseersystemen in te stellen.

KWR heeft een memo opgesteld voor de uitvoering van experimenten B, C1 en C2 met de proefinstallatie. Daarbij werd beoogd om minimaal 5 EGV-sensoren te installeren. In verband met de verhuizing van de proefinstallatie en project-deadlines zijn uiteindelijk 3 realtime EGV-sensoren geïnstalleerd. In één van deze sensoren stelt Vitens geen vertrouwen in de kwaliteit van de gelogde data. Het beperkte aantal experimenten en betrouwbare sensoren reduceert de zeggingskracht van de bijbehorende resultaten (Hoofdstuk 3) aanzienlijk, met 2 sensoren is bovendien de identificatie van de scenario’s C1 en C2 niet mogelijk. De metingen zijn derhalve slechts indicatief.

3 Resultaten en discussie

3.1 Graafreconstructie voor data uit 6 sensoren

De resultaten in deze paragraaf staan tevens beschreven in de CCWI-congresbijdrage (Bijlage I).

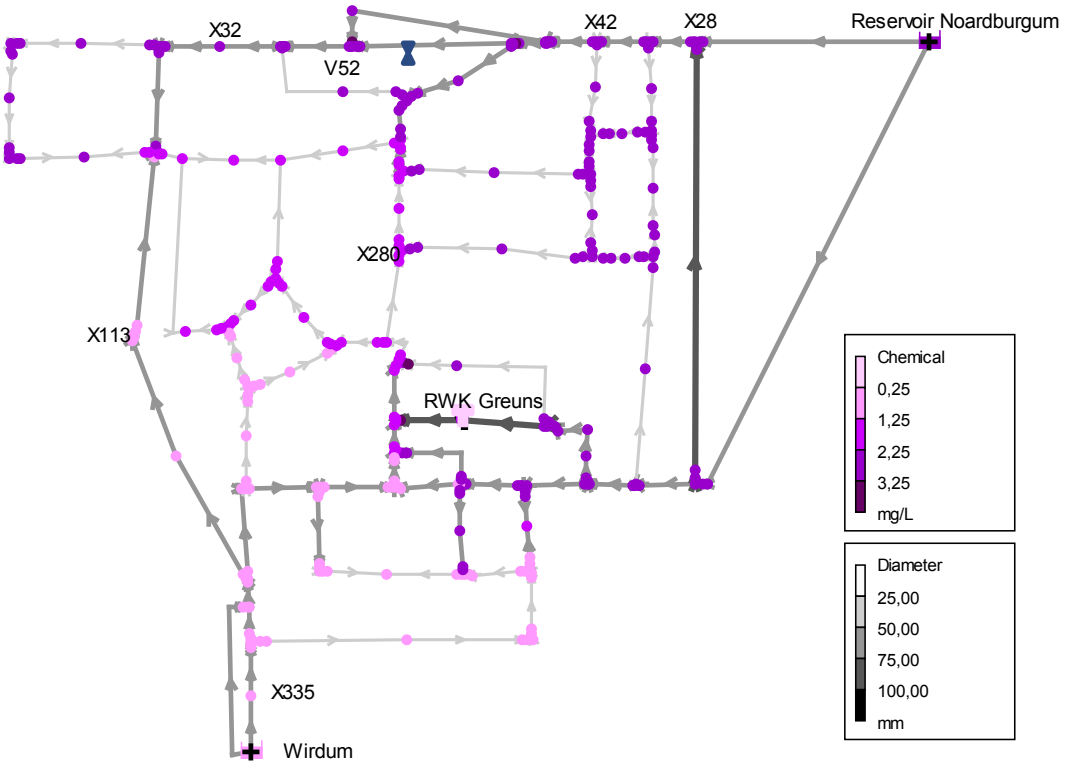
3.1.1 PMC: Monte Carlo-methodiek

De resultaten met de PMC-methodiek lijken veelbelovend indien gebruik wordt gemaakt van een klein (academisch) voorbeeld met 4 knooppunten en Gaussiaanse ruis (ruis met een amplitude die als functie van de tijd een normale verdeling heeft). De methode blijkt echter geen betrouwbare resultaten te geven indien gegevens zijn gesimuleerd met EPANET. Verblijftijden worden in dat geval alleen geschat tussen X42 wijzend naar X28 (3 uur verblijftijd) en X32 wijzend naar X28 (1,45 uur verblijftijd). Deze schattingen wijken ver af van de vertragingen die afgeleid zijn uit de gemeten responses. Bovendien blijkt het causale effect precies in de tegenovergestelde richting bepaald te zijn. De methodiek is op basis van deze resultaten niet verder uitgewerkt, en de focus is gelegd op de toepassing van het PC-algoritme.

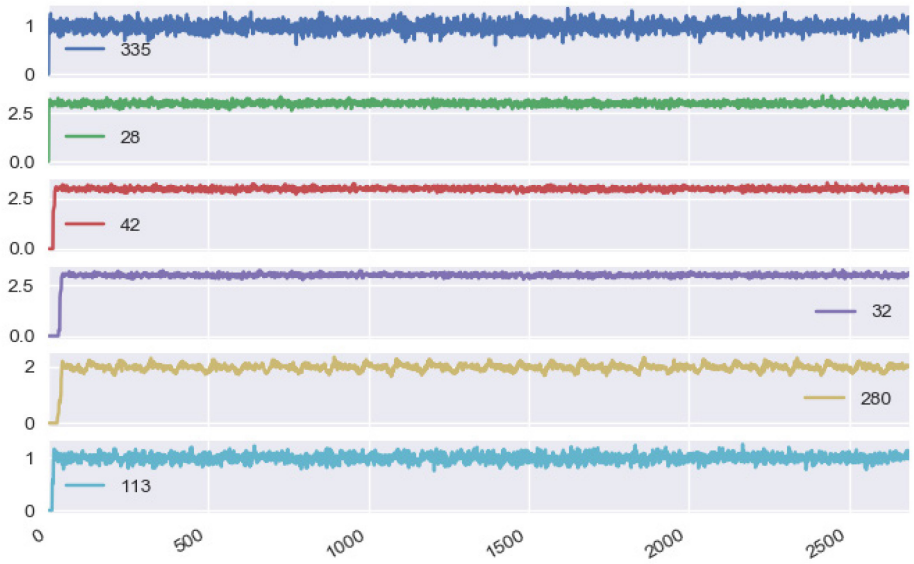
3.1.2 PPC: scenario B-CCWI

Situatie B-C, het basisscenario, is gesimuleerd met EPAnet. In dit scenario zijn heeft de waterkwaliteit afkomstig van Noordbergum een andere samenstelling dan die vanuit Wirdum. Dit is gesimuleerd met verschillende tracerconcentraties van respectievelijk 3.0 en 1.0 (eenheid niet relevant), resultaten zijn getoond in Figuur 3-1. Op basis van de EPANET-simulatie, verwachten we dat informatie vanuit Wirdum (Wd) via X335 naar X113 en X280 vloeit, terwijl informatie van reservoir Noordbergum (Nb) via X28 naar X42 zal stromen, dan via V52 naar X32 stroomt en mogelijk van X42 of X32 naar X280. De meetfrequentie is 4 keer per uur, de gesimuleerde sensorsignalen zijn weergegeven in Figuur 3-2.

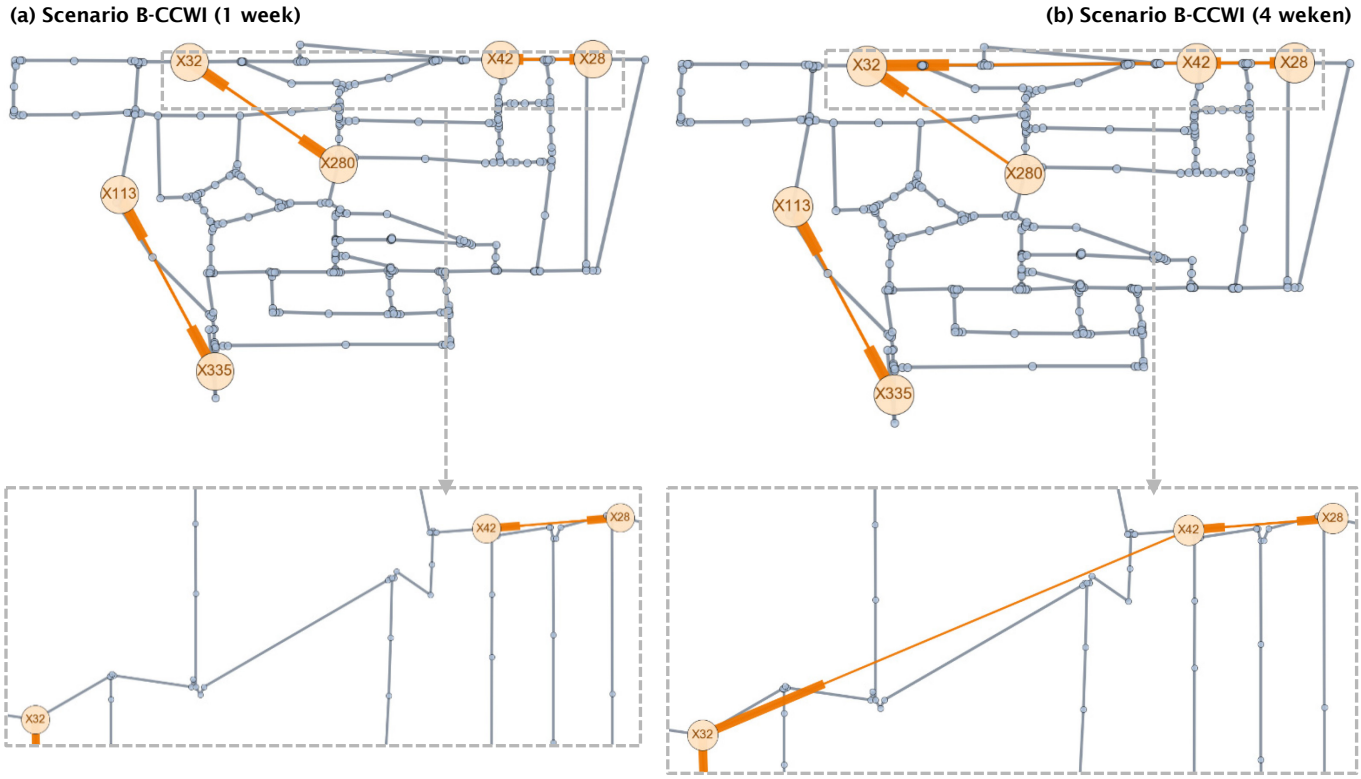
De resultaten van PPC worden weergegeven voor de situatie dat er een meetset van 1 week aan gegevens beschikbaar is (Figuur 3-3a) en een periode van 4 weken, ofwel 672 uur (Figuur 3-3b). Het PC-algoritme vindt bij een meetperiode van 1 week een correlatie maar geen causaliteit tussen de sensorparen X335- X113 en X28-X42, zoals blijkt uit de oranje lijnen met dubbele pijlpunten. Een gegevensperiode van 4 weken resulteert in meer gevonden (cor)relaties. De verbinding X42 - X32 is nu toegevoegd en de informatiestroom is correct bepaald in de richting van X32. Verder wordt X32 blijkbaar beïnvloed door X280. De richting in de causaliteit is opmerkelijk, omdat de EPANET-simulaties aantonen dat X280 8,75 uur ten opzichte van het Nb-reservoir vertraagd is, terwijl X32 een vertraging van 7,75 uur heeft ten opzicht van Nb. Merk op dat er geen causaliteit kan worden afgeleid uit gegevens van X335 en X113.



FIGUUR 3-1: HET EPANET-MODEL VAN DE PROEFINSTALLATIE, MET LEIDINGDIAMETER WEERGEGEVEN IN GRIJSWAARDEN (ZIE LEGENDA ONDERAAN), DE STROOMRICHTING GETOOND MET PIJLEN EN DE CONCENTRATIEWAARDE VAN TRACERCHEMICALIËN PAARS GEKLEURD (ZIE LEGENDA). SENSOREN ZIJN GEPLAATST BIJ NODES X335, X113, X32, X280, X42 EN X28. BIJ WIRDUM IS DE GEMIDDELDE TRACERCONCENTRATIE 1.0, BIJ NOARDBURGUM IS DE GEMIDDELDE CONCENTRATIEWAARDE INGESTELD OP 3.0. DE KLEP (V54) DIE IN SCENARIO C1 IS GEBRUIKT, WORDT GETOOND MET EEN BLAUW SYMBOOL.



FIGUUR 3-2: SENSORSIGNALEN IN HET BASISSCENARIO (B-CCWI) BIJ X335, X28, X42, X32, X280 EN X113.



FIGUUR 3-3: GRAAF VAN DE INFORMATIESTROOM IN HET SENSORNETWERK. ORANJE GEKLEURDE LIJSTUKKEN VERBINDEN DE KNOOPPUNTEN (HIER SENSOREN), EEN VERDIKKING VAN DE LIJN REPRESENTEERT EEN PIJLPUNT. DE GRAAF IS GEPROJECTEERD OP HET SCHAALMODEL (BLAUW) VOOR HET B-CCWI-SCENARIO VOOR (A) 1 WEEK EN (B) 4 WEKEN VAN GEGEVENS. DE ONDERSTE PANELEN ZOOMEN IN OP DE REGIO VAN HET NOORDELIJKE STROOMPAD (X28 → X42 → X32).

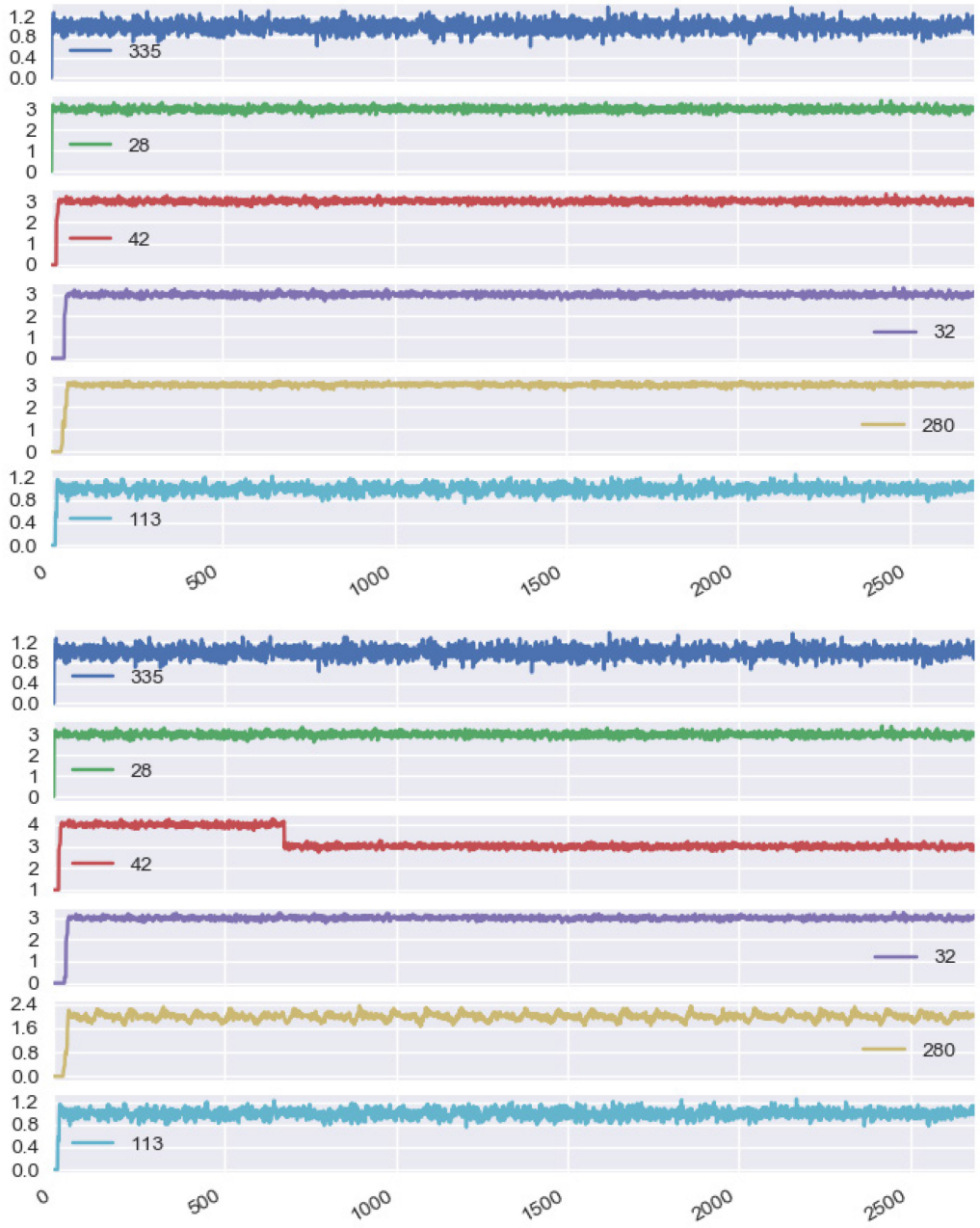
3.1.3 PPC: scenario C1 en C2

Het PC-algoritme is opnieuw toegepast, dit keer op de scenario's C1 (een afsluiterstand is anders dan in situatie B) en C2 (sensorstoring). De tijdreeksen afkomstig van de sensoren zijn weergegeven in Figuur 3-4. Op de y-as staan de (EGV)-signalen, op de x-as het aantal meetmomenten (hier 4 weken aan waterlevering gesimuleerd, met een meetmoment per kwartier).

Voor C1 (Figuur 3-5a) is informatie geblokkeerd tussen de sensoren in de directe nabijheid van de afsluiterstand (X42, X32, X280). In vergelijking met de resultaten in Figuur 3-3b, detecteert het algoritme een onderbroken informatiestroom (X42 - X32 en X32 - X280) in de buurt van de gesloten afsluiter, terwijl de andere correlaties en enkele van de richtingen ongewijzigd blijven.

(C1)

(C2)

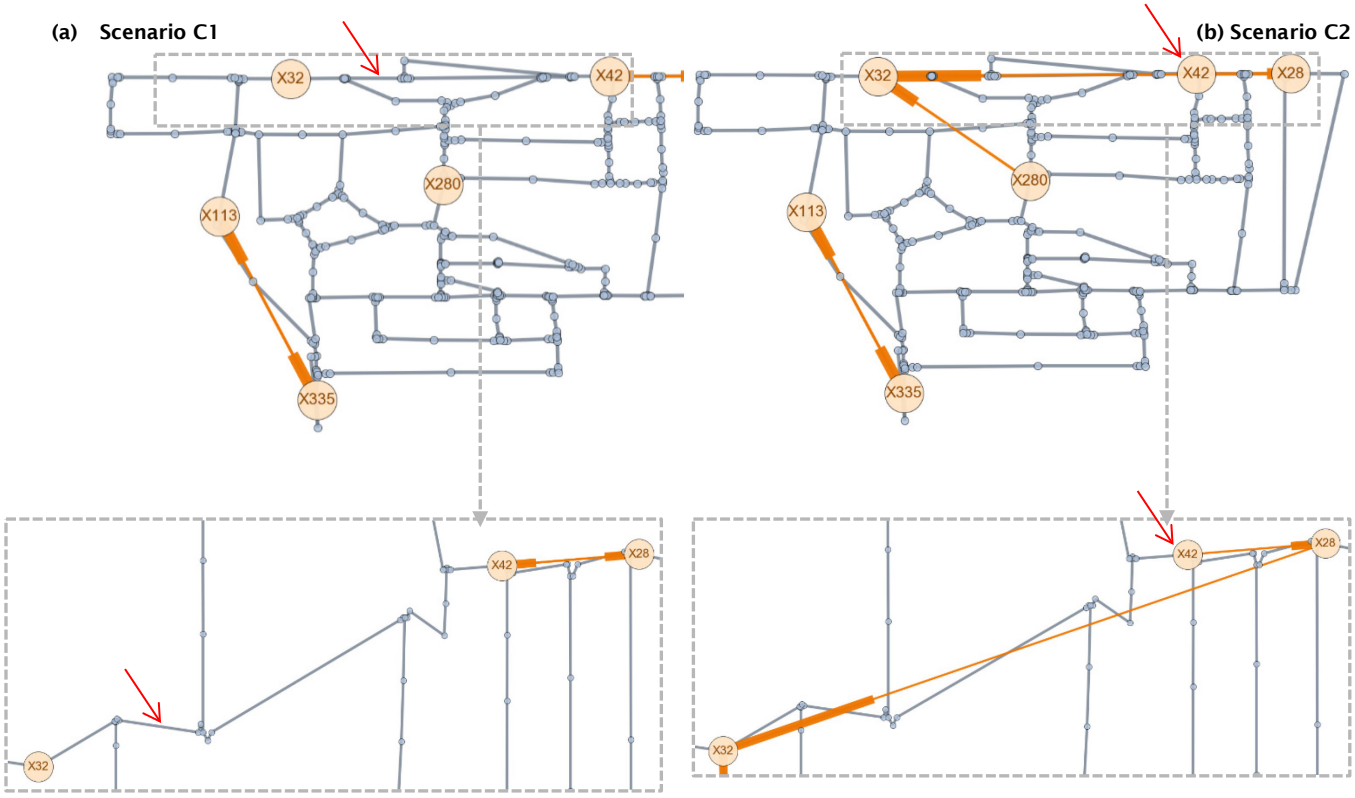


FIGUUR 3-4: SENSORSIGNALEN IN SCENARIO C1 (AFSLUITER V54 DICHT) EN C2 (BIAS IN SENSOR X42) BIJ X335, X28, X42, X32, X280 EN X113, OP DE X-AS STAAN HET AANTAL MEETMOMENTEN (MET TIJDSCHALING: ELK KWARTIER 1 MEETMOMENT).

De afgeleide graaf van scenario C2 verschilt ook van het basisscenario, maar nu in de knooppunten X28, X42, X32 en X280. Het afwijkende sensorsignaal X42 resulteert in een herverdeling van verbindingen: X42 heeft nu een causaal effect op X28 en X28 heeft een effect op X32. Merk op dat het verstoorde signaal van X42 niet meer Gaussisch is, wat een belangrijke voorwaarde schendt voor de onafhankelijkheidstests van het PC-algoritme.

De uit paragraaf 3.1 verkregen inzichten suggereren dat de gevolgde aanpak kan worden gebruikt om afwijkingen in het distributienet en sensornetwerk te detecteren. Voor de realisatie van een robuust algoritme dat toepasbaar is in werkelijke leidingnetwerken is het

vooraf schatten van de informatiereistijd tussen de ene en de andere sensor onwenselijk. Hiervoor zijn verdere doorontwikkeling en testen van de methodiek nodig.

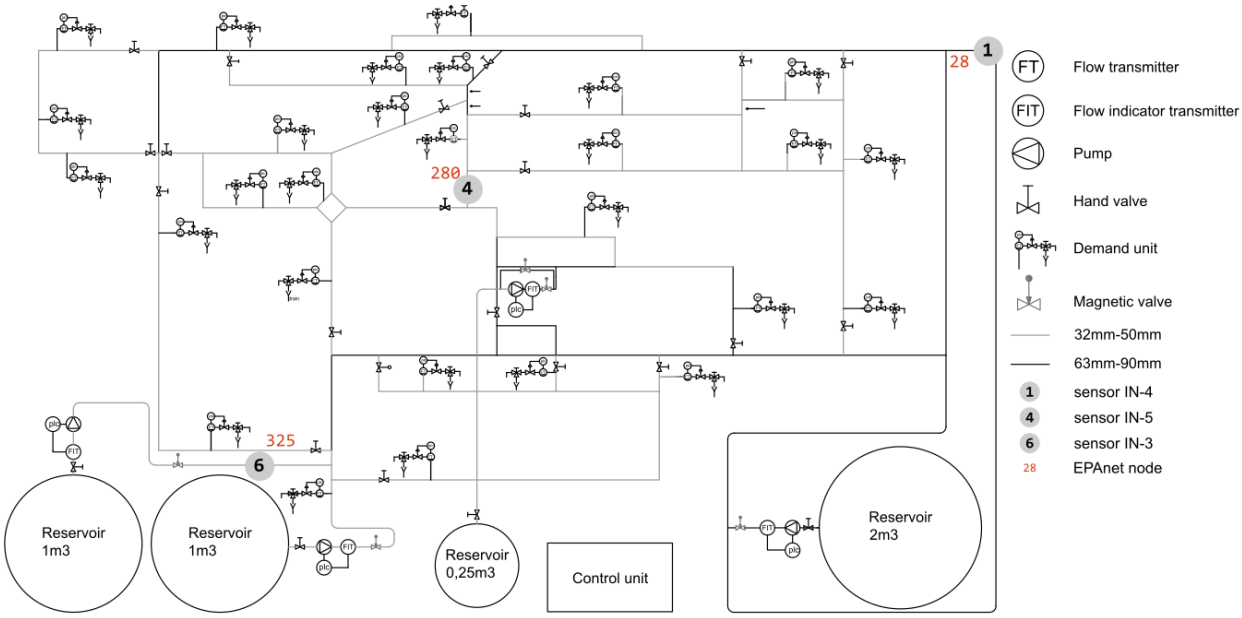


FIGUUR 3-5: GRAAF VAN DE INFORMATIESTROOM IN HET SENSORNETWERK. ORANJE GEKLEURDE LIJSTUKKEN VERBINDEN DE KNOOPPUNTEN (HIER SENSOREN), EEN VERDIKKING VAN DE LIJN REPRESENTEERT EEN PIJLPUNT. DE GRAAF IS GEPROJECTEERD OP HET SCHAALMODEL (BLAUW) VOOR HET C1-SCENARIO (A) EN (B) HET C2-SCENARIO. DE RODE PIJLEN GEVEN LOCATIES AAN VAN EEN FOUTIEVE AFSLUITERSTAND (C1-SCENARIO OF EEN SCENARIO MET SENSORSTORING (SCENARIO C2). DE ONDERSTE PANELEN ZOOMEN IN OP DE REGIO VAN HET NOORDELIJKE STROOMPAD (X28 → X42 → X32).

3.2 Graafreconstructie voor data uit 3 sensoren

3.2.1 Geïnstalleerde sensoren in het geschaalde modelnetwerk

Ter referentie is het proces- en instrumentatiediagram van het modelnetwerk dat gebruikt is in experiment B-L weergegeven in Figuur 3-6. De grijze bolletjes met de nummers 1, 4 en 6 in het figuur geven de EGV-sensoren aan. De rode tekst naast de sensorlocaties geven de EPAnet-nummering weer (sensor 1: X28, sensor 4: X280, sensor 6: X325). De signalen van sensor 4 zijn als onbetrouwbaar aangemerkt.



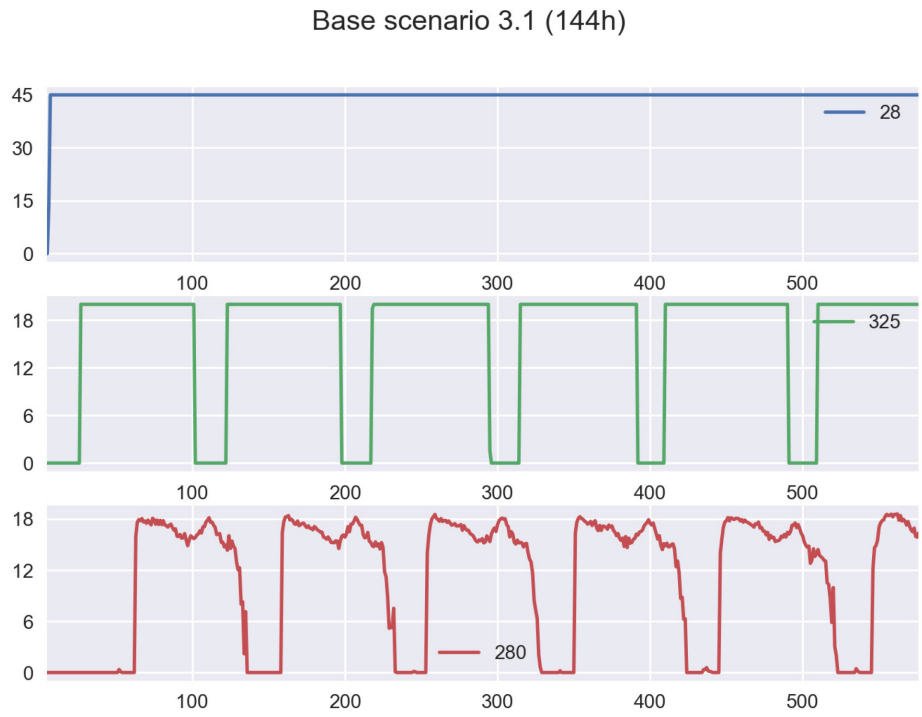
FIGUUR 3-6: PROCES- EN INSTRUMENTATIEDIAGRAM VAN HET MODELNETWERK. IN ROOD DE NUMMERING DIE DOOR HET EPANETMODEL IS GEBRUIKT VOOR DE EGV-SENSOREN IN-3, IN-4, EN IN-5.

3.2.2 Nabootsen sensorsignalen met model

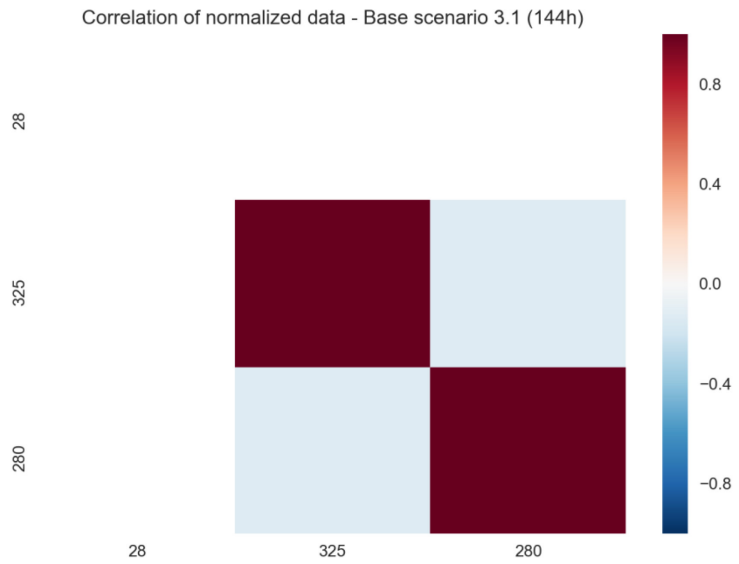
Als eerste stap is situatie B-L nagebootst in EPAnet met afnamepatronen zonder toegevoegde Gaussische ruis. Een meetfrequentie van 4 metingen per uur is aangehouden, een 24-uurspatroon heeft dus 96 meetmomenten. Volumestromen en tracerconcentraties zijn met dit model doorgerekend. De gemeten concentraties bij locaties X28, X325 en X280 zijn getoond in Figuur 3-7. Merk op dat de blokvormige signalen duidelijk verschillen met de Gaussische ruissignalen van het modelexperiment in paragraaf 3.1. De blokpatronen zijn het gevolg van tijdelijke bijmenging van water afkomstig van het reservoir Wirdum waarin geen markeerstoffen zijn gedoseerd. Deze afwijkende stroompatronen ten opzichte van scenario B-CCWI hangen samen met het inactief maken van 18 van de 28 verbruikspatronen. Dit was nodig om tegemoet te komen aan de beperkte toevoer in de proefhal. Uit Figuur 3-7 wordt duidelijk dat het blokpatroon vanuit sensorlocatie X325 (nabij Wirdum) is terug te zien bij de sensorlocatie in het midden van het net (X280). Daarnaast zijn de verblijftijden goed te schatten uit de bloksignalen: blijkbaar kost het 26 tijdstappen (1u38m in werkelijkheid) voordat het patroon vanuit Wirdum locatie X325 bereikt, het constante signaal in Noordbergum bereikt vrijwel instantaan locatie X28. Locatie X280, de sensor in het ‘midden’, heeft een verblijftijd van 51-26=25 tijdstappen ten opzichte van sensorlocatie X325. Op basis van deze resultaten is gekozen om het PPC-algoritme uit te voeren op de meetset vanaf de 26e tijdstap en een verblijftijd van 25 tijdstappen tussen sensor X280 en X325, omdat het algoritme niet in staat is rekening te houden met verblijftijden. Na weglaten van de eerste periode wordt de data genormaliseerd[3] naar waarde 1. Vervolgens is een correlatiematrix bepaald, zie Figuur 3-8. Deze correlatiematrix staat aan de basis van de PPC-routine. Merk op dat X325 inderdaad correleert met X280 met een waarde -0.13 (normalisatie zorgt voor veel negatieve waarden, waardoor de correlatie in dit geval ook negatief uitvalt), en dat de correlatie met sensor X28 een nulwaarde heeft (witte vlakken) omdat het een constant signaal is.

[3] Bij normalisatie naar waarde 1 wordt het gemiddelde van een dataset bepaald en die van een (meet)waarde afgetrokken. Het resultaat wordt vervolgens gedeeld door het verschil tussen de maximum- en minimumwaarde.

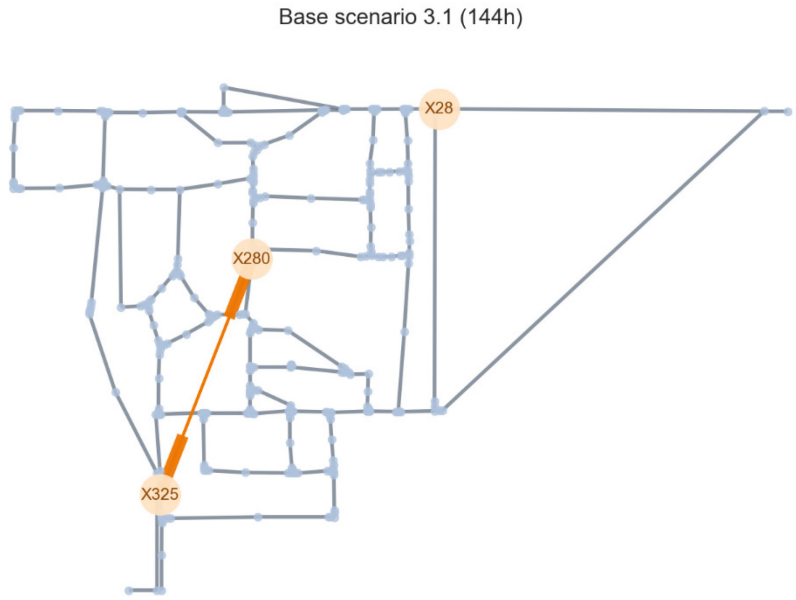
De graafreconstructie op basis van de PPC-methodiek is getoond in Figuur 3-9. Merk op dat het blokpatroon en de correlatie tussen sensor X280 en X325 inderdaad wordt opgemerkt door het algoritme, maar dat de richting onbepaald is gebleven. De onbepaalde richting is waarschijnlijk het gevolg van de complexe EGV-patronen die ontstaan door bijmenging van water zonder doseerstoffen vanuit Wirdum.



FIGUUR 3-7: GESIMULEERDE PATRONEN BIJ SENSORLOCATIE X28, X325 EN X280.



FIGUUR 3-8: CORRELATIEMATRIX VAN DE GESIMULEERDE SIGNALEN NA NORMALISATIE. ER IS GEEN CORRELATIE MET SENSOR 28.



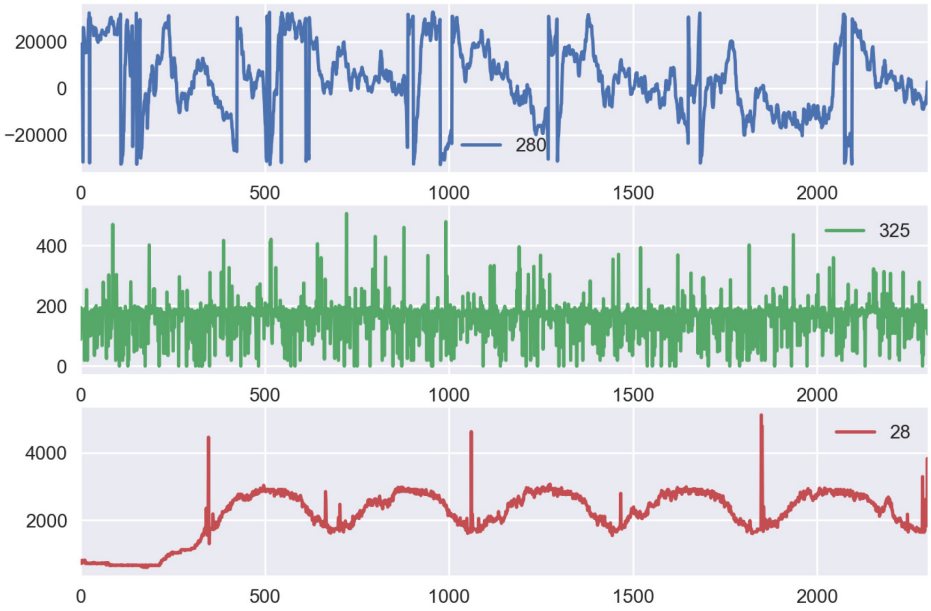
FIGUUR 3-9: GRAAFRECONSTRUCTIE (ORANJE LIJN TUSSEN SENSOR X325 EN X280) VAN DE SENSORDATA DIE MET EPANET OP SENSORLOCATIES X325, X28 EN X280 IS GESIMULEERD.

3.2.3 Experiment met de proefinstallatie

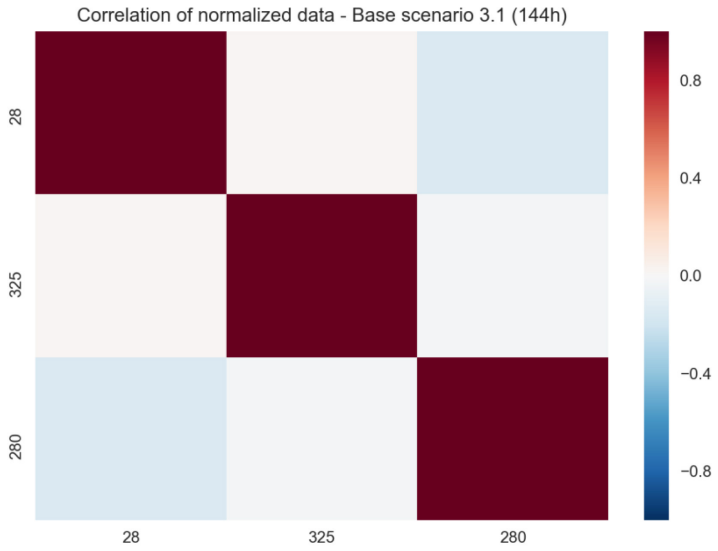
In totaal zijn er 2303 meetmomenten per tijdreeks beschikbaar met een tijdschaalfactor van 1:60, overeenkomend met 6 dag-vraagpatronen (144 uur). Sensoren hebben elke 3,75 seconden gelogd, dat overeen komt met een opgeschaalde tijd van 3,75 minuten. Figuur 3-10 toont de gemeten EGV-signalen bij de EPAnet-knopen X280, X325 en X28. Merk op dat er veel variatie zit op sensorsignaal X325, de sensor dichtbij Wirdum. Deze lijkt invloed te hebben op de waterkwaliteit gemeten op locatie X280. Ook lijkt dit signaal niet het blokpatroon te hebben dat eerder numeriek bepaald is met EPAnet (Figuur 3-7). Daarentegen lijkt het signaal in locatie X28, nabij Noordbergum, wel een terugkerend patroon te hebben. Sensorsignaal X28 lijkt ook dit golfpatroon in locatie X280 te hebben, hoewel interpretatie

lastig lijkt door de hoeveelheid ruis. Na 214 meetmomenten is het afnamepatroon nabij Noordbergum zichtbaar, daarvóór lijkt er een 'nulsignaal' te zijn. De periode ervoor wordt daarom als verblijftijd beschouwd. Verder lijken de ruispatronen tussen de 600 en 800 meetmomenten, 1000 en 1200, 1400 en 1500 en 1700 en 1900 monsters bij X28 te zorgen voor extra verstoringen in X280 met een vertraging van ca. 200 monsters (750 sec., overeenkomend met ca. 12 uur). Andere verblijftijden ten opzichte van de voeding, of de verblijftijd van signaal X280 ten opzichte van signaal X28 en X325, konden op basis van deze meetset niet worden bepaald. De eerste 214 meetmomenten worden buiten beschouwing gelaten voor de analyse met het PC-algoritme.

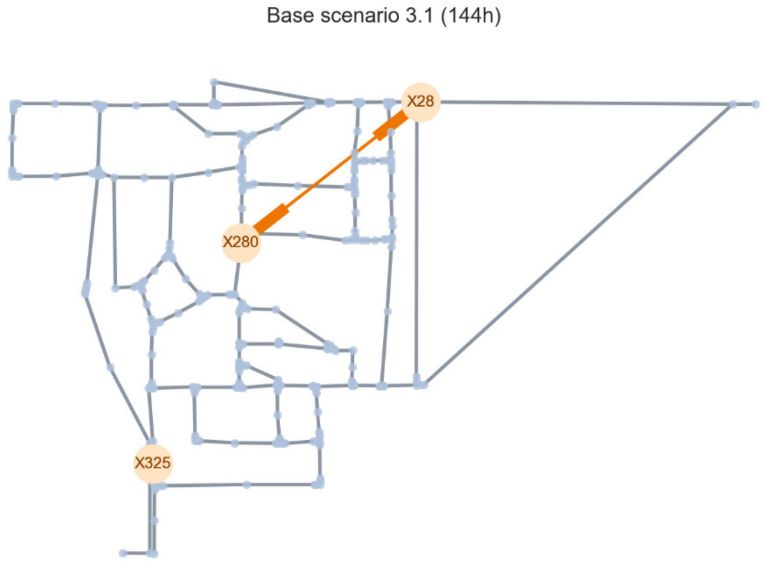
Uit de meetset kon met het PPC-algoritme de graaf als getoond in Figuur 3-12 worden herleid. Blijkbaar lijkt het patroon uit Noordbergum (X28) te correleren met het signaal uit locatie X280. Dit wordt bevestigd door de correlatiewaarde van -0.15, getoond in de matrix in Figuur 3-11. De correlatie tussen X325 en X280 is vrijwel verwaarloosbaar (0.02). De richting van de informatiestroom tussen X28 en X280 kon echter niet worden bepaald.



FIGUUR 3-10: EGV-METINGEN (IN mS) OP SENSOREN X280, X325 EN X28.



FIGUUR 3-11: CORRELATIEMATRIX VAN DE GENORMALISEERDE MEETDATA.



FIGUUR 3-12: POSITIE VAN SENSOREN IN HET EPANET-MODEL (ORANJE BOLLETJES) EN DE GERECONSTRUEERDE GRAAF (ORANJE LIJN) GEPROJECTEERD OP HET GESCHAALDE PROEFNETWERK (LICHTBLAUW). EEN VERDIKKING VAN DE LIJN REPRESENTEERT EEN PIJLPUNT, TWEE PIJLPUNTEN OP EEN LIJNSTUK BETEKENT DAT DE RICHTING ONBEPAALD IS.

3.3 Discussie

De resultaten van de gesimuleerde en gemeten situatie B-L verschillen aanzienlijk:

- de signalen afkomstig van de locaties dichtbij de voedingspunten Wirdum (X325) en Noordbergum (X28), maar ook het signaal gemeten in X280, hebben in het daadwerkelijke experiment veel ruis die in de numerieke situatie niet gesimuleerd zijn.
- een golf- of blokpatroon is in sensorsignaal X28 (nabij NB) waarneembaar, terwijl deze in de EPAnet-simulatie niet is aangenomen. De EPAnet-simulatie gaat juist uit van een

blokpatroon vanuit X325 (Wirdum) en een constant signaal uit X28 (Noordbergum). Dit is waarschijnlijk het gevolg van hoe de inkomende volumestromen zijn verdeeld over de drie voedingslocaties. Deze verdeling is ingesteld zoals in de simulatie, maar het is gebleken dat deze in de proef lastig met veel precisie zijn te sturen. Dit kan leiden tot afwijkende volumestroompatronen en de daarmee samenhangende sensorsignalen.

Als gevolg van deze verschillen, zijn er verschillen in de correlaties. Het PC-algoritme gaat uit van de correlatiematrix voor het reconstrueren van een ongerichte graaf. De graaf die de informatiestroom in beeld brengt, is daardoor verschillend voor de numerieke resultaten ten opzichte van de graaf bepaald uit de meetdata. Uit de signalen van de numerieke en de experimentele situatie is geen 'Bayesiaanse' causaliteit ontdekt.

3.4 Synthese

Deze studie heeft inzicht opgeleverd in de toepassing en bruikbaarheid van informatietheoretische concepten voor het vergroten van systeemkennis in het leidingnetwerk. Met synthetische experimenten zijn twee verschillende methodieken getest op bruikbaarheid. Gedurende het ontwikkel- en testtraject is gebleken dat bepaalde karakteristieken van een methodiek meer of minder de potentie van doorontwikkeling en toepassing in de praktijk bepalen.

De resultaten beschreven in paragraaf 3.1 tonen dat de Bayesiaanse aanpak van het PC-algoritme in de PPC-methode vrij robuust de (causale) connectiviteit binnen een netwerk kan aantonen en herleiden, een belangrijke eigenschap die ontbrak in de Monte Carlo benadering van de PMC-methodiek. Een nadeel van het PC-algoritme is de conditie dat voor het herleiden van de connectiviteitsstructuur (de graaf), de signalen stationair (Gaussisch) moeten zijn: m.a.w. geen verblijftijdseffecten en dynamiek in het signaal, bijvoorbeeld de afhankelijkheden met andere variabelen (co-lineariteit). Verder zijn een groot aantal meetpunten nodig om betrouwbaar een graafreconstructie te kunnen bepalen. De Granger's causaliteitstoets die toegepast is in het PMC-methode is juist wel ingericht op het compenseren van verblijftijdseffecten. Ondanks deze relevante tekortkomingen van het PC-algoritme, is het 'standaard' PC-algoritme beter toepasbaar voor de analyse van sensorsignalen in leidingnetwerken. Het PC-algoritme dient wel aangepast te worden zodat het kan omgaan met (a) andere signalen en historie van een signaal (auto)regressieve karakteristieken en (b) verblijftijden. Een dergelijke doorontwikkeling is toegepast in bijvoorbeeld het werk van Demiralp en Hoover (2003) voor een klein netwerk. Nadeel van de methodiek is dat er heel veel verschillende combinaties van regressiemodellen en structuren mogelijk zijn, waardoor elke mogelijkheid gefit en getoetst moet worden met het PC-algoritme. Dat maakt de methode ook potentieel gevoelig voor vals-positieve verbindingsstukken en dus een foutieve interpretatie van de informatiestroom. Weer een ander algoritme is om uit te gaan van netwerk*opbouw*[4] waarbij een Bayesiaans waarschijnlijkheidscriterium wordt gebruikt, zoals het *Greedy Equivalence Search* (Meek 1997, concept toegepast in bijv. neurologie: Ramsey *et al*., 2009). De voor- en nadelen zijn vergelijkbaar met het PC-algoritme. Een overzicht van voor- en nadelen is gegeven in Tabel 3-1.

[4] het PC-algoritme gaat uit van een volledig verbonden vermaasd netwerk en haalt verbindingsstukken weg op basis van Bayes conditionaliteitstoetsing.

TABEL 3-1: OVERZICHT VAN INFORMATIETHEORETISCHE CONCEPTEN OM DE KENNIS VAN HET SENSORNETWERK IN HET LEIDINGNET TE VERGROTEN

Informatietheoretische benaderingen	Toegepast op leidingnetwerken	Beschrijving	Voordelen	Nadelen
MC en Granger's causaliteit (PMC)	ja*	Monte Carlo toetsing van Granger's causaliteit tussen 2 knooppunten	Houdt rekening met dynamiek (verblijftijden)	Huidige methodiek houdt geen rekening met connectiviteit (mazen in netwerk bijv.)
PC-algoritme	ja*	PC-algoritme detecteert causaliteit en connectiviteit in signalen van sensornetwerken	Houdt rekening met (inter)connectiviteit	- Statische methode, houdt geen rekening met dynamiek en verblijftijden; - Alleen toepasbaar op Gaussische (ruis)signalen
PC-algoritme en SVAR-identificatie[1]	nee	Zoals hierboven, maar vooraf gegaan door Monte Carlo simulatie van mogelijke SVAR-modellen	Houdt rekening met connectiviteit en (regressieve) tijdreeksen	- vergt meer rekentijd dan standaard PC-algoritme; - gevoelig voor fouten in SVAR-identificatie
(G)ES [2, 3]	nee	Netwerkopbouw op basis van een Bayesiaans criterium	Houdt rekening met (inter)connectiviteit	- modificaties nodig om rekening te houden met verblijftijden [3] en andere dynamica

**Dit werk, [1] DemirAlp en Hoover (2003), [2] Meek (1997), Ramsay et al., 2009.*

Tenslotte verdient het aanbeveling om alle sensoren te ijken en te toetsen op betrouwbaarheid, om vervolgens het experiment én de simulaties te herhalen met als doel het kalibreren van het EPAnet-model en het toetsen van een algoritme. Pas dan kan geconstateerd worden in hoeverre het algoritme gevoelig is voor (toevallige) correlaties enerzijds en overige factoren, zoals de gevoeligheid voor verblijftijd, hoeveelheid datapunten, fitmethodieken of de 95%-betrouwbaarheidsgrens anderzijds.

4 Conclusies en aanbevelingen

4.1 Conclusies

De algemene conclusies van het onderzoek zijn:

- De toetsing van de methodiek met virtuele sensorgegevens uit numerieke berekeningen toont dat het PC-algoritme verschillende uitkomsten geeft bij de verschillende scenario's (foutieve sensormeting of afsluiterstand). Gelet op de theorie en toepasbaarheid van dit algoritme is het belangrijk dat voldoende (Gaussische) data beschikbaar is en de verblijftijden tussen sensorlocaties bekend zijn om betrouwbare resultaten te verkrijgen.
- De methodiek is in principe toepasbaar voor uiteenlopende vraagstukken en schaalbaar naar toepassingen in het werkelijke net waarbij tientallen of honderden sensoren worden gebruikt. Een voorwaarde voor drinkwaterdistributiesystemen is dat een oplossing wordt gevonden om de verblijftijd te schatten (kalibratie met tracer-stapfuncties) en sensorsignalen hiervoor te corrigeren.
- Het toepassen van de methodiek op sensorgegevens uit experimenten met de proefinstallatie van Vitens heeft geen aanvullende antwoorden opgeleverd over het presteren van de methodiek onder niet-geïdealiseerde omstandigheden.
- Het is op dit moment nog te vroeg voor implementatie van de uitgewerkte methodiek in de praktijk.

Specifieke conclusies wat betreft de mogelijkheden van de uitgewerkte methodiek zijn:

- Het PC-algoritme is veelbelovend om (veranderingen in) tijdscorrelatie (Granger's causaliteit) te bepalen uit sensornetwerkdata van een DWDN. Dit is gebleken in een gesimuleerde omgeving waar bronpatronen van de waterkwaliteit zijn gedefinieerd als signalen met Gaussische ruis en de geobserveerde uitvoersignalen op de sensorknoop-punten zijn gecorrigeerd voor vertraging door transport van water.
- Een Monte Carlo aanpak voor het toetsen van Granger causaliteit tussen sensorparen gevolgd door een selectie van de meest frequent voorkomende vertragingen heeft geen bevredigende resultaten opgeleverd. Voorbewerking, vertragings-consistentie checks en verdere toetsing zijn nodig om de methodiek robuuster in te richten en vervolgens te valideren en evalueren of de voorziene aanpak voldoende betrouwbaar is.
- Het PC-algoritme is gevoelig voor de signaalinhoud. Een deel van de causaliteit in het sensornetwerk werd niet opgelost wanneer de signaalduur te kort was of wanneer de standaarddeviatie van de ruis te klein was.

4.2 Aanbevelingen voor vervolgonderzoek

Voor toepassing van de methodiek in de praktijk is het nog te vroeg. Verder onderzoek is nodig om een aantal overgebleven of vervolgvragen te beantwoorden.

Algemene kennisvragen zijn:

- Kan de combinatie van het algoritme met andere data (bijvoorbeeld een hydraulisch model, of een leidinginformatiesysteemmodel) niet (nog) meer opleveren? Voor een dergelijke benadering dient ook onderzocht te worden hoe op een systematische manier modelfouten van afwijkingen in de werkelijkheid onderscheiden kunnen worden.
- De vraag is open gebleven hoe er uit sensoren zoveel mogelijk informatie afgeleid kan worden voor meervoudige toepassingen (incidentdetectie, systeemkennis,

(klant)vertrouwen in het systeem). Op welke wijze kan deze informatieverwerking zo slim mogelijk? Sensoren kunnen voor dit doel apart worden geïnstalleerd, of er kan gebruik worden gemaakt van sensoren die al voor een ander doel zijn geïnstalleerd. Toetsing van het PC-algoritme kan dan als bijvangst dienen.

Kennisvragen ten aanzien van het PC-algoritme zijn:

- Met welke precisie en zekerheid (vals-positieven, vals-negatieven) zijn afwijkingen te herkennen? Is het type afwijking te bepalen? Hoeveel, welke en waar zijn sensoren hiervoor nodig? Uit wetenschappelijke literatuur blijkt dat de methodiek ook zou moeten werken met alleen sensorinformatie door gebruik te maken van (vector) autoregressieve modellen en recursieve toepassing van een graafreconstructiemethode (bijvoorbeeld het PC-algoritme). Hiermee wordt het toepassingsbereik groter: ook tijdreeksen met niet-Gaussische signatuur en verblijftijdseffecten worden dan meegenomen. De vraag rijst hoe betrouwbaar een dergelijke methodiek is. Om deze vraag te beantwoorden is verdere ontwikkeling en validatie nodig.
- Hoe verhoudt de meerwaarde (of opbrengst) van het PC-algoritme zich tot andere methodieken? Voor het herkennen van afsluiterstanden moet de aanpak met het PC-algoritme zich bijvoorbeeld meten met de bepaling hiervan op basis van een bij de afsluiter geplaatste geluidsensor. Ook met betrekking tot het bepalen van sensorfouten is te denken aan alternatieven zoals het aanbrengen van een intern alarmsysteem bij de sensoren. Ook voor toepassingen van het PC-algoritme die niet in het huidige project zijn getoetst is een vergelijk met alternatieve methodieken en/of doorontwikkeling aan te bevelen.

Ten aanzien van toetsing in de praktijk, wordt het gebruik van een proefinstallatie aanbevolen als validatiestap ter aanvulling op numerieke simulaties en voorafgaand aan grootschalige toepassing, vanwege drie redenen. Reden (1): het gebruik van niet-geïdealiseerde hydraulische omstandigheden en sensormetingen levert een realistischere toetsing op, (2) een proefinstallatie biedt een veilige testomgeving en (3) incidenten of (extreme) afwijkende situaties komen in de praktijk niet vaak voor, waardoor toetsing langdurig of complex wordt. Dit laatste aspect is een probleem gebleken in eerdere projecten waarbij gegevens uit het werkelijke netwerk werden gebruikt (Vries et al., 2016; Maessen et al. 2017).

Tenslotte, de ontwikkelingen in dit project lopen vooruit op een toekomst waarin realtime monitoring (en op termijn realtime aansturing) van het leidingnet met sensorgegevens een steeds grotere rol spelen in het operationele beheer. Indien bovenstaande vervolgvragen zijn beantwoord, wordt het interessant om de methodiek in het werkelijke leidingnet te toetsen.

5 Literatuur

Dahlhaus, R. & Eichler, M. (2003). *Causality and graphical models in time series analysis*. Oxford Statistical Science Series, 115-137.

Kalisch, M. & Bühlmann, P. (2007). *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*. Journal of Machine Learning Research 8: 613-636.

Kroll, D. & King, K. (2010). *Methods for evaluating water distribution network early warning systems*. American Water Works Association, 102(1), 1-11.

Demiralp, S. & Hoover, K.D. *Searching for the causal structure of a vector autoregression (2003).* Oxford Bulletin of Economics and statistics 65: 745-767.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A. en Glymour, C. (2010). *Six problems for causal inference from fMRI*. NeuroImage 49 (2): 1545–58.

Spirtes, P. and G. Clark (1991). An algorithm for fast recovery of sparse causal graphs. Social science computer review 9.1: 62-72.

P. van Thienen, B. de Graaf, M. van de Roer, P. Schaap, V. Sperber (2014*). Sensoring van waterkwaliteit in het distributienet: een rationele benadering*. H2O online.

Van Summeren, J. (2016). *Investeringen en rendementen van sensornetwerken ten behoeve van waterkwaliteitsbewaking – TKI INTEREST*. KWR Watercycle Research Institute, KWR 2016.052.

Van Summeren, J. & Meijering, S. (2015). *Meeting of Waters – Handleiding proefinstallatie.* KWR Watercycle Research Institute, BTO 2015.221(s).

Van Summeren, J., Meijering, S., Hijnen, W., Beverloo, H. & Van Thienen, P. (2014). *Meeting of Waters: Ontwerp van een proefinstallatie voor drinkwatertransport in de Vitens Innovation Playground*. KWR Watercycle Research Institute, BTO 2014.041.

Van Summeren, J., Meijering, S., Beverloo, H. & Van Thienen, P. (2017). *Design of a distribution network scale model for monitoring drinking water quality*. Journal of Water Resources Planning and Management, 143(9). doi:10.1061/(ASCE)WR.1943-5452.0000799.

Van Summeren, J., Van Thienen, P., Vries, D., Vertommen, I., Korevaar, M., Brouwer, S. (2016). *Kostenefficiënte toepassing van sensoren voor meerdere doelen in het drinkwaterdistributiesysteem.* KWR Watercycle Research Institute, BTO 2016.048.

Vries, D., Van den Akker, B., Vonk, E., De Jong, W., Van Summeren, J. (2016) *Application of machine learning techniques to predict anomalies in water supply networks,* Water Science & Technology: Water Supply, 16(6), 1528-1535.

Vries, D., Van Summeren, J. (2017). *Valve status verification and sensor error detection via causal inference drom sensor data* (F94), CCWI 2017 –Computing and Control for the Water Industry, Sheffield, U.K.,

Bijlage I Conferentiebijdrage CCWI

Valve Status Verification and Sensor Error Detection via Causal Inference from Sensor Data

Dirk Vries[1,*], Joost van Summeren[1]
[1] KWR Watercycle Research Institute, Groningenhaven 7, 3433PE Nieuwegein, The Netherlands
[*]*dirk.vries@kwrwater.nl*

ABSTRACT

Recent developments in (near) real-time sensor applications have the potential to provide operators and managers with useful information on drinking water distribution supply and need of its maintenance. A systematic methodology based on causal inference from observational data is proposed to increase knowledge of water supply distribution systems equipped with sensor networks. This methodology can be used to help identify deviations from expected operation of water supply and sensor infrastructure, using only observational data. We outline the first steps of two distinct procedures that use data from a sensor network, to infer a map of a causal dependence structure. These procedures are applied to scenario studies where an unexpected change in operation occurs, i.e. a valve status is different and a sensor bias is introduced. A draft outline of future steps is given that could improve and validate the methodology.

Keywords: graph modelling, sensor networks, distribution model

I.1 BACKGROUND

Recent developments in (near) real-time sensor applications have the potential to provide operators and managers with useful information on drinking water distribution supply and need of maintenance. Although implementation of sensor networks is not yet common practice, numerous numerical studies have demonstrated potential benefits of a sensor network, such as real time event detection of water quality contaminations (e.g. [1]) or leakage and pipe burst detection and localization (e.g. [2]). We also foresee that sensor networks will provide operational benefits such as improving distribution network models and the effectiveness of sensor networks. Automated monitoring and control of water supply services using sensor data and models imply a strong reliability on sensor data and network models. This reliance poses a challenge because knowledge of distribution networks is not always correct, comprehensive, and up-to-date and sensors are known to be imperfect (false positives and negatives, drift and failure, etc.). A systematic methodology to increase actual knowledge of the systems can help identify deviations from expected operation of the drinking water distribution and sensor infrastructure.

A novel method investigated in this work is aimed at quantifying operational benefits using a heuristic approach and testing the methodology with a laboratory scale model of a real-life distribution network. In this paper, we focus on the development of a graph theoretical

procedure aimed at improving the quality of system information and models that rely on such information. We outline the first steps of two distinct procedures to use only observational data, i.e. data from a sensor network, to infer a map of a causal dependence structure. We test these steps and give a draft outline of the remaining steps that could improve the methodology in the identification of deviations from expected operation in water supply networks. In order to provide tangible and quantified benefits of a sensor network we narrow the work down to two practical applications, i.e. (C1) detection of changed valve statuses and (C2) detection of erroneous sensor measurements.

I.2 METHODS

The information flow of a sensor network provides the basis of our procedure to infer a graph model for a baseline situation, i.e. where the operation of the drinking water distribution network (DWDN) is assumed normal. It is our hypothesis that any deviations with respect to this baseline case (BC) occurring from sensor faults, leakage or changed valve status values, etc. is revealed as a change in the newly estimated graph. Hence the estimated information flow is presented as a graph, i.e. each node represents a (sensor) variable such as flow, pressure, or electrical conductivity, and each edge represents a correlation (undirected graph) or a direct cause between nodes (directed acyclic graph, DAG). It is assumed that there are no feedback loops, hence acyclic, or hidden variables. We implement and evaluate two notions of causality inference from synthetic DWDN data by the following steps:

1. a framework for graph theoretical analysis is set up and written in the Python programming language to determine if there are causal relationships in the sensor network (Figure 1). The framework consists of calls to the R statistical software package (via python module rpy2), calls to EPANET (epanettools), calls to statistical tests (stattools) and methods to construct and visualise the sensor and DWDN via the python module networkx.

   The framework enables testing and evaluating two procedures (Figure 2):

   - a graph theoretic methodology that uses the Peter and Clark (PC)-algorithm directly onto sensor data [3,4]. Conceptually, the algorithm starts with a complete, undirected graph between each node (here: each sensor signal) and recursively deletes edges based on Markov conditional independence tests until a minimal set of connected nodes is reached. The R package pcalg provides pre-defined functions to perform these independence tests on Gaussian, discrete or binary data and discover directed (acyclic) graphs. The PC algorithm is evaluated on data which is pre-processed to remove time lags. These time lags are (manually) estimated on the basis of delays in step response signals. This procedure will be referred to as PPC;
   - a Monte Carlo (MC) analysis of Granger causality tests and detected time lags for every sensor couple in the set of sensors (nodes) using the sensor data. Granger causality means that 'X Granger-causes Y, if Y can be better predicted using the histories of both X and Y than it can by using the history of Y alone'. The same data as in the PPC procedure is used, but now the lag estimation is part of the causality tests. Based on a set of estimated lags and causal relations between nodes, a subset of most likely occurring lags (and thus Granger-causal relations) are selected to draw a directed graph. This will be further called the PMC procedure.

2. a DWDN model (EPANET) was constructed with a layout and sensor network depicted in Figure 2. The DWDN represents an experimental scale model of a real-life supply zone [5]. It includes two main water supply sources with different water quality in the West (Noard-Burgum, NB) and South (Wirdum) and the transport and large distribution mains (of real-life diameters of >300 mm). Demands are represented by 31 demand nodes. A tank is included, but is not actively taken into account in the calculations. Three sensors (X28, X42, X32) are placed along a North-side transport

main of eastward water flow from the water supply source in the West, one sensor (X280) is placed near the centre of the DWDN, one (X335) near the water supply source in the South and at the Eastside (X113). All sensors measure water quality, i.e. the concentration of a chemical tracer.

3. We define and evaluate test conditions (variation in signals, number of sensors) and define the baseline scenario BC and case scenarios C1 and C2:
    - BC: 1 week and 4 weeks of tracer data is simulated by EPANET with a 15 minutes sampling frequency to check the sensitivity of the method to the amount of available data. Repeating weekly demand patterns are set. A chemical tracer is supplied at NB with an average concentration value of 3.0 and perturbed with Gaussian noise with standard deviation 0.1. Similarly, the water supply at Wirdum contains a tracer with an average concentration of 1.0 and a Gaussian noise perturbation with a standard deviation of 0.1.
    - C1: similar to BC with 4 weeks of data, except that one valve, i.e. valve V52 (Figure 2) is closed.
    - C2: similar to BC with 4 weeks of data, except that one sensor, i.e. sensor X42 (Figure 2) has a bias of +1.0 during a period of 168 hours from the start of the simulation.
4. Simulation tests of cases C1 and C2 for an EPANET model of a water supply distribution network (DWDN) in order to determine the performance and applicability of the procedures.

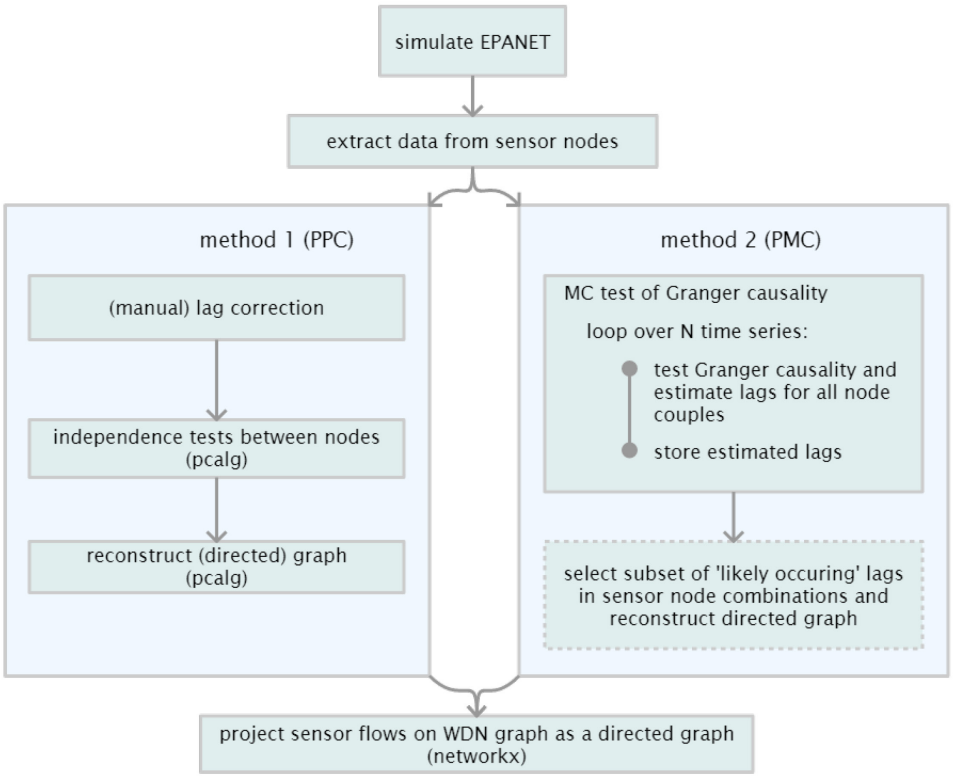


FIGURE 1. SCHEME OF FOLLOWED PROCEDURES TO ESTIMATE A CAUSAL STRUCTURE BETWEEN SENSOR NODES IN A DWDS.

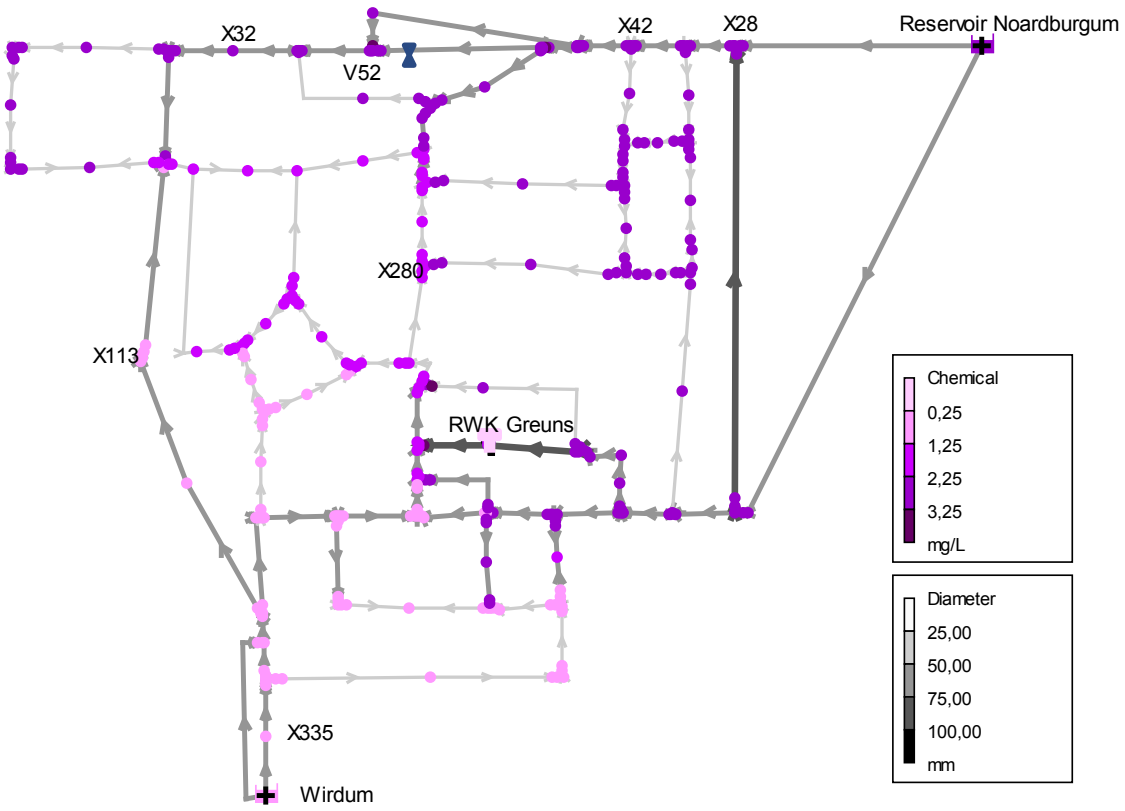


FIGURE 2. LAYOUT OF THE EPANET SCALE MODEL. PIPE WIDTHS ARE PLOTTED ON A GRAY SCALE, FLOW DIRECTIONS ARE SHOWN BY ARROWS, AND THE CONCENTRATION OF TRACER CHEMICALS ARE SHOWN BY THE PURPLE COLORS. SENSORS ARE PLACED AT NODES X<NUMERIC VALUE>. AT WIRDUM, THE AVERAGE TRACER CONCENTRATION IS 1.0, AT NOARDBURGUM THE AVERAGE CONCENTRATION VALUE IS 3.0. THE VALVE USED IN SCENARIO C1 IS SHOWN BY THE BLUE SYMBOL LABELED V52.

I.3 RESULTS

PMC Method

The results with the PMC method vary to a large extent. While the results were promising for a small, academic example with 4 nodes and Gaussian noise; the method does not work well with (sensor) data simulated with EPANET. Lags are only estimated between X42 pointing towards X28 (lag: 3 hours) and X32 pointing towards X28 (lag: 1.45h) and are off from the estimated lags deduced from step responses and the causal effect is exactly in the opposite direction (X28 to X42: 3.75 hours).

PPC Method

The PPC procedure relies on the R library ‘pcalg’ to infer causal maps (directed graphs). Results are shown in Figure 3. When no causal relation between two nodes is found, no edge is drawn. The conditional independence tests are run with an uncertainty threshold of 5% (orange lines). Results of the simulated baseline scenario (BC) are shown for the case where water quality data during a period of 1 week is available (Figure 3a) and a period of 4 weeks (Figure 3b). Based on the simulation in EPANET (Figure 2), we expect that information flows from Wirdum (no sensor present) via X335 to X113 and most of the time to X280, while information from reservoir Noardburgum (NB) will flow via X28 towards X42, passing V52 towards X32 and possibly from X42 or X32 to X280.

The BC lag corrected case when using *1 week* of chemical tracer data is shown in Figure 3a. In this graph, the algorithm finds that X335 is correlated to X113, but no causality is found. (Partial) correlation is also resolved between nodes X28 – X42. No (causal) relation between X42 – X32 is found. When a *4 week* period of data is available, the graph looks different (Figure 3b). The edge X42 – X32 is now added, and the information flow is correctly resolved towards X32. Furthermore, X32 is apparently effected by X280. The direction in causality is remarkable, because from the EPANET simulations we know that X280 is lagging 8.75 hours w.r.t. the NB reservoir, while X32 has a delay of 7.75 hours. Note that no correlation or causality could be inferred from data of X335 and X113.

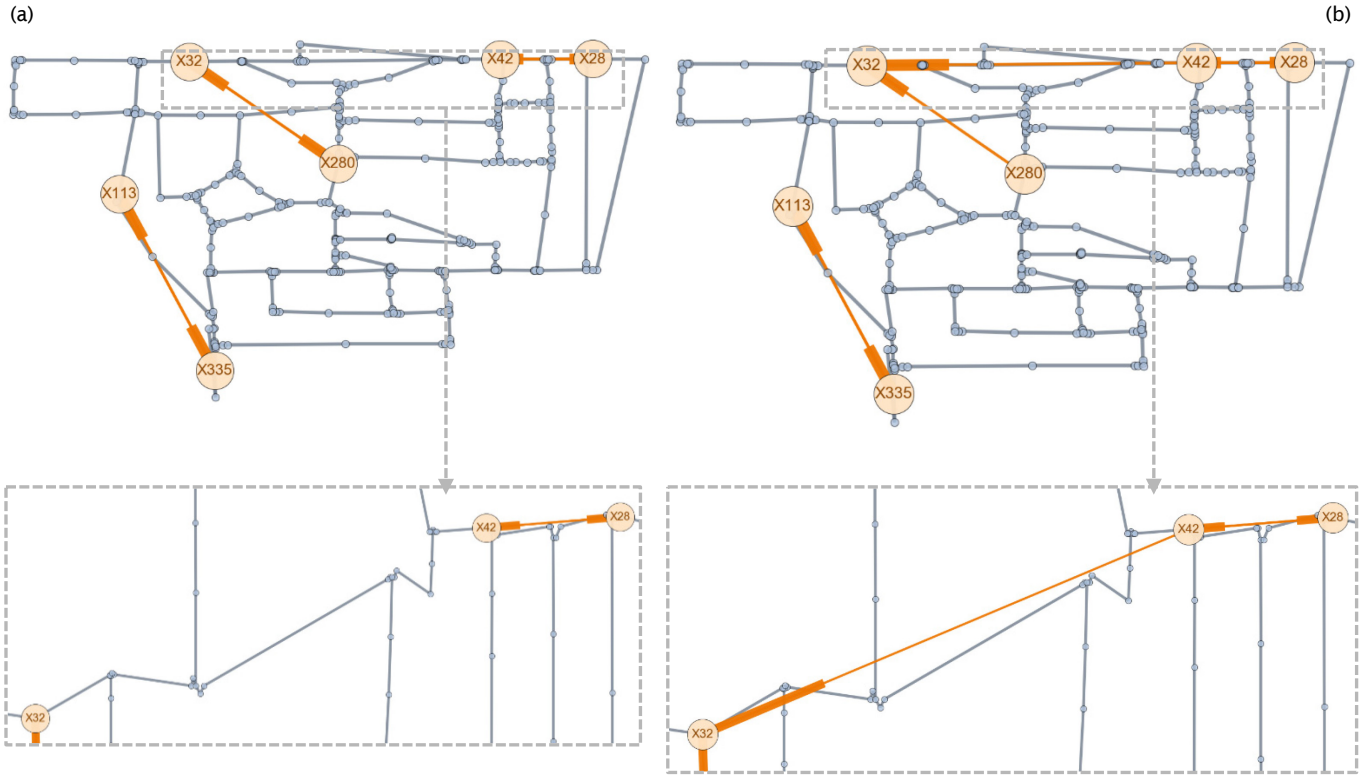


FIGURE 3: DAG REPRESENTATION OF THE SENSOR NETWORK OBSERVATIONS (ORANGE COLOURED EDGES, A STUB REPRESENTS AN ARROWHEAD) PROJECTED ONTO THE SCALE MODEL (BLUE) BY PPC FOR THE BC SCENARIO. DAG RECONSTRUCTION BY (A) 1 WEEK AND (B) 4 WEEKS OF DATA , RESPECTIVELY. THE LOWER PANELS ZOOM IN ON THE REGION OF THE NORTHERN TRAJECTORY (X28 > X42 > X32).

Figure 4 shows the results of the PPC procedure applied to the practical cases of an unexpected valve status (C1) and sensor failure (C2). For the case C1, the graph shows that information is obstructed between the sensors in the direct vicinity of the valve (X42, X32, X280). Compared to Figure 2b, the procedure indicates an interrupted flow of information in the vicinity of the valve that was closed in this simulation (X42 – X32 and X32 – X280), while the other connections and some of the directions remain unchanged. The inferred graph of scenario C2 in Figure 4 also differs from the BC graph (Figure 3b), but now in the node set X28, X42, X32 and X280. Apparently, the introduced bias in sensor X42 leads to a re-shuffling of links: X42 now has a causal effect on X28 and X28 has an effect on X32. It should be noted that the perturbed signal of X42 is not Gaussian anymore, which violates an important assumption for the independence tests of the PC algorithm. These preliminary results must be substantiated with further tests, but suggest that the followed approach can be used to detect deviations from expected operation, even with a limited number of sensors.

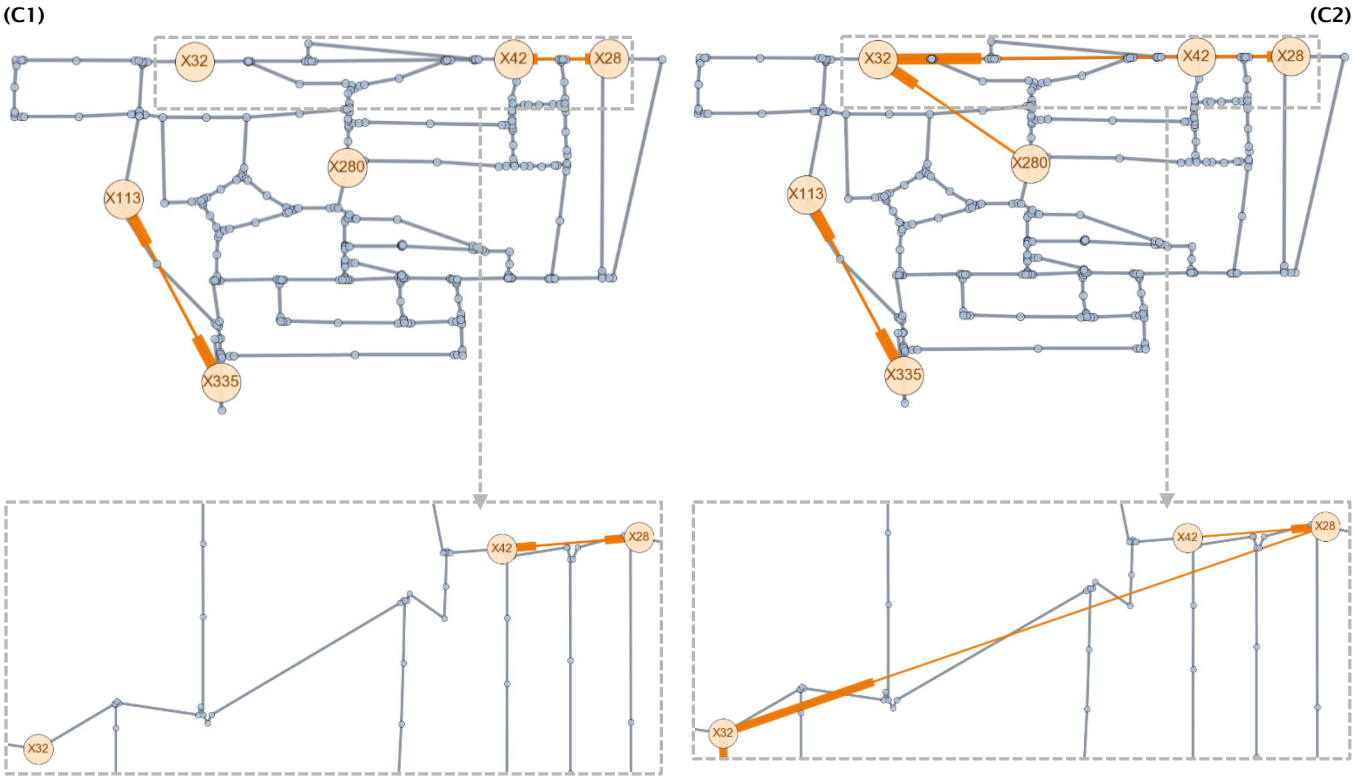


FIGURE 4: GRAPH NOTATION AS FIGURE 3, FOR PRACTICAL APPLICATIONS OF AN UNEXPECTED VALVE STATUS (C1) AND ERRONEOUS SENSOR MEASUREMENTS (C2). THE LOWER PANELS SHOW THE ZOOMED IN REGION OF THE NORTHERN TRAJECTORY (X28 > X42 > X32).

I.4 DISCUSSION AND FUTURE STEPS

While the PC algorithm as introduced by Peter and Clark is - by definition - not equipped for lagged signals, a drawback in applying the PMC method to the presented application is that it relies on a heuristic approach that searches for a high probability in connectivity based on time lag information between node couples, but not for the whole network. As a consequence, there is a significant risk of introducing inconsistencies in the resulting graph with respect to the sum of estimated lags if interconnected nodes are present. Furthermore, it is known that Granger causality tests should be applied to data that is stationary, and, possible co-integration should be corrected first. However, in this work, stationarity is not guaranteed due to several reasons, e.g. possible mixing of water, changing flow directions during the course of a day or when bias occurs (as is the case for C2).

Based on experiences with both techniques, we propose to pursue the following steps to improve the method:

- Check whether the use of other data (flow, pressure or other water quality data) reduces the amount of data needed to capture the information flow.
- Extend the PMC procedure with (i) a check whether the data is stationary, semi-stationary over a time window, or has an order of integration with the augmented Dickey–Fuller test and proceed with (ii) the estimation of a VAR (vector autoregressive regression) using the procedure as outlined in [6] and (iii) check for Granger causality.
- For PPC, there are different options to cope with lagged sensor signals:
    - With the assumption that the unlagged (source) signal is the only cause for the lagged signal and there is no co-integration or autoregression present, time

delays can be estimated by minimisation of the squared difference of the unlagged (source) signal and lagged signal [7]. This rather straightforward approach resembles the work here, except that the time delays are now estimated from data.

- As a first step, it is assumed that all sensor signals can be modelled by a structured VAR (SVAR) process, i.e. $A(L)Y_t = E_t$, where Y_t is a n x 1 column vector of n sensor variables at time step t; A is a n x n conformable matrix whose terms are polynomials in a fixed lag value L, and E is a column vector of errors at time t. The idea is to generate SVAR models by a Monte Carlo approach with N realisations of A, estimate the SVAR and use the residuals (filtered Y_t minus Y_t) as an input for the PC algorithm in each realisation. Note that N is typically in the order of 10^4 to 10^5. Then, the selected graph is compared to a reference graph ('PC true graph'), i.e. the graph that indexes the equivalence class to which the true graph belongs. Finally, every possible link is evaluated. See [8] for more details.

In addition, we plan to apply the technique to a real-life scale model of a distribution network from the Dutch water company Vitens and evaluate the results to relate this new information to tangible benefits. The use of a scale model (as opposed to numerical simulations) allows for testing the methodology under (close-to) real-life circumstances, including realistic imperfections of sensors, drinking water and pipes in a controlled environment. Another research subject is to address the optimal sensor placement problem based on maximising causal inference by the PMC or PPC method.

I.5 CONCLUDING REMARKS

- A Monte Carlo approach for testing Granger causality between any sensor pair followed by selection of the most frequently occurring lags did not yield satisfactory results. Pre-processing, lag consistency checks and further tests are needed to validate whether this approach holds promise.
- The PC algorithm seems promising to infer (changes in) causalities from sensor network data of a DWDN, at least in a simulated environment where source patterns of water quality are defined as signals with Gaussian noise and the observation output signals at the sensor node positions are corrected for time lags.
- The PC algorithm is sensitive to the 'information content' of the signal, or more generally speaking, whether the system excitation was sufficient. Part of the causality in the sensor network was not resolved when either the signal duration was too short or when the standard deviation in the noise sequence was relatively small.

Acknowledgements

This research was conducted within a project of the Joint Research Program (BTO) of the Dutch Water Companies. We acknowledge the valuable comments by Mirjam Blokker (KWR) on the manuscript, the support of Vitens to use the scale model of Leeuwarden for experiments and fruitful discussions with Peter van Thienen (KWR).

References

[1] Zhao, Y., Schwartz, R., Salomons, E., Ostfeld, A., Vincent Poor, H. New formulation and optimization methods for water sensor placement (2016). Environmental Modelling & Software, 76 Issue C, 128-136.

[2] Mounce, S. R., Mounce, R. B., Jackson, J., Boxall, J.B. Pattern matching and associative artificial neural networks for water distribution system time series data analysis (2014). Journal of Hydroinformatics, 16.3, 617-632.

[3] Spirtes, P. and G. Clark. An algorithm for fast recovery of sparse causal graphs. Social science computer review 9.1 (1991): 62-72.

[4] Kalisch, M. and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research 8.Mar (2007): 613-636.

[5] Van Summeren, J., Meijering, S., Beverloo, H. and Van Thienen, P. Design of a distribution network scale model for monitoring drinking water quality (2017). Journal of Water Resources Plannning and Management, 143(9), 1-10.

[6] Toda, H.Y., and Yamamoto, T. Statistical inference in vector autoregressions with possibly integrated processes (1995). Journal of econometrics 66.1: 225-250.

[7] Giovanni, J and Scarano, G. Discrete time techniques for time delay estimation (1993). *IEEE Transactions on signal processing* 41.2: 525-533.

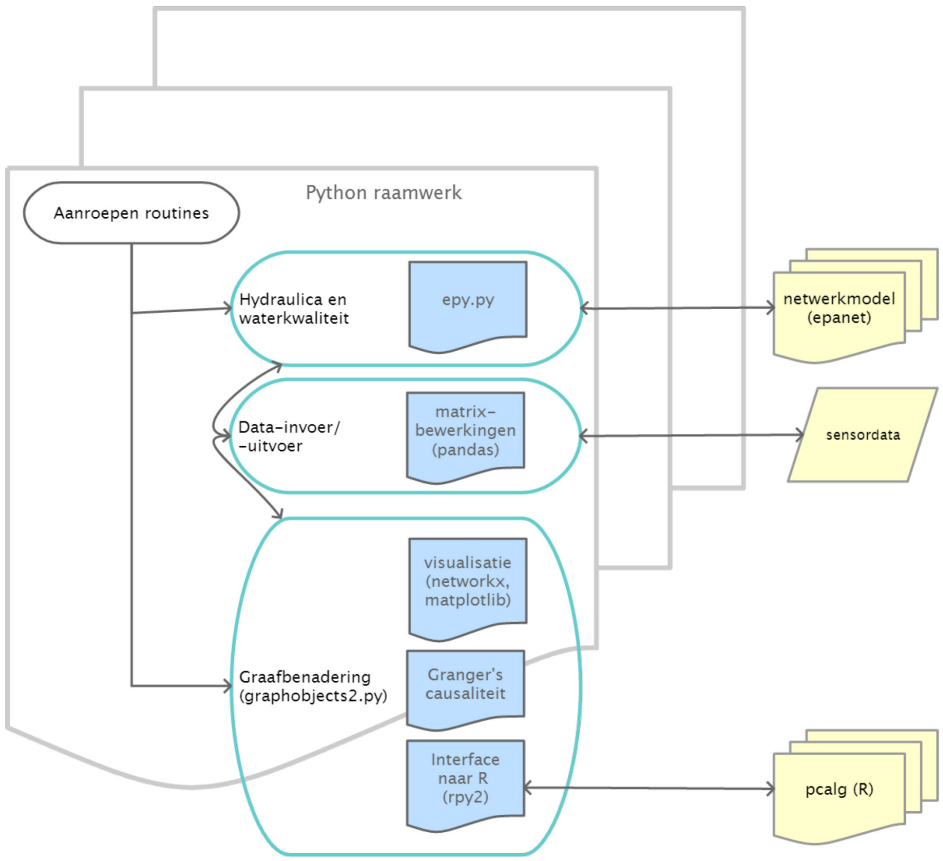
[8] Demiralp, S., and Hoover, K.D.. Searching for the causal structure of a vector autoregression (2003). *Oxford Bulletin of Economics and statistics* 65: 745-767.

Bijlage II Programmatuur

Structuur van softwarecode

Er is gebruik gemaakt van voornamelijk Python en het statistische pakket R om graaftheoretische concepten numeriek te testen op (a) sensordata en (b) gesimuleerde waterkwaliteit. Causaliteitstesten worden bepaald met het PC-algoritme (PPC-routine, via de pcalg-bibliotheek van R) en Granger's causaliteitstoetsen inclusief verblijftijdschattingen (PMC). Deze bepalingen en visualisatie zijn in graphobjects2.py door KWR object-georiënteerd ontwikkeld, waarbij o.a. gebruik is gemaakt van bestaande visualisatieroutines in Python (networkx, matplotlib). Sensordata worden in een matrix opgeslagen, waarop de graafbenaderingen kunnen worden getest. Data worden opgeslagen in matrices met behulp van de pandas-module. In plaats van sensordata kunnen ook gesimuleerde data (van waterkwaliteit) met behulp van het softwarepakket EPAnet worden aangeroepen, opgeslagen en getest worden op causaliteit tussen de tijdreeksen. Aanroep en het verwerken van data is in een epy.py object geprogrammeerd.

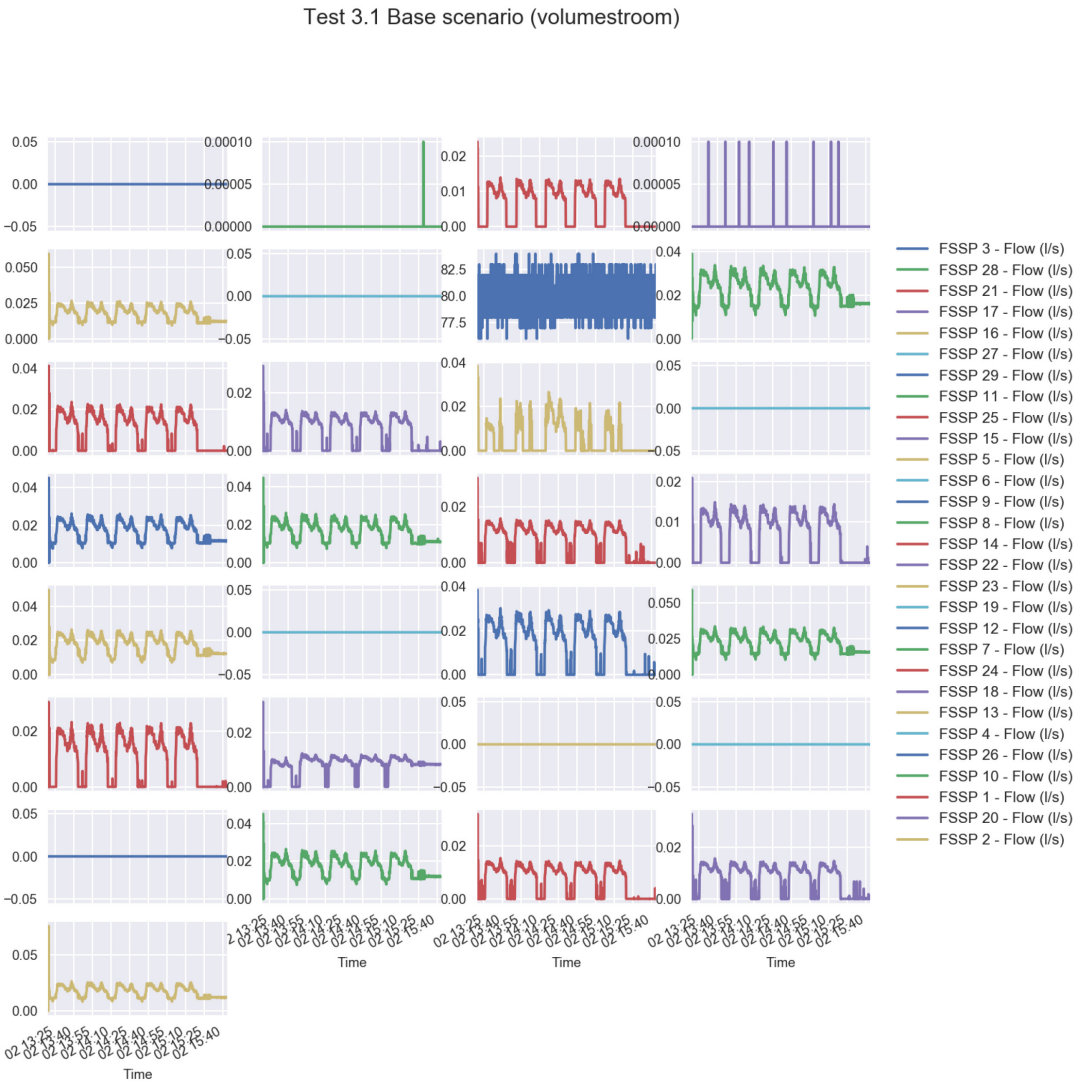
Het Python-raamwerk is schematisch weergegeven in Figuur 5-1.



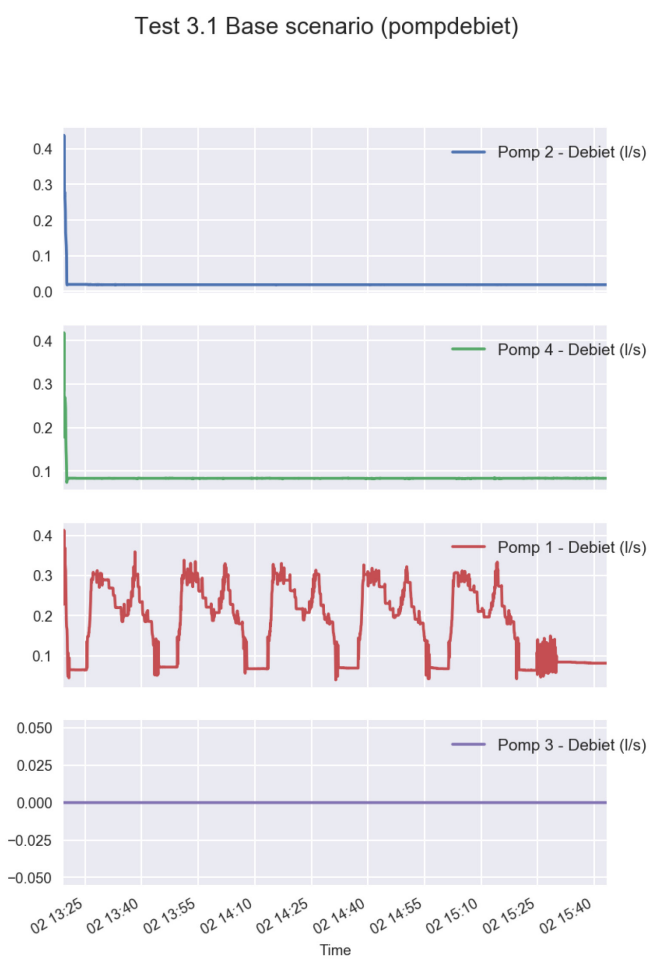
FIGUUR 5-1: SCHEMATISCH OVERZICHT VAN GEBRUIKTE PROGRAMMATUUR.

Bijlage III Data

Experiment B-L



FIGUUR 5-2: VOLUMESTROMEN (L/S) GEMETEN OP VERSCHILLENDE PUNTEN IN HET MODELNETWERK.



FIGUUR 5-3: GEMETEN POMPDEBIETEN. POMP 1 IS VOOR VOEDINGSLOCATIE WIRDUM (MET DOSEERSYSTEEM), POMP 2 VOOR WIRDUM (ZONDER DOSEERSYSTEEM), POMP 3 VOOR REINWATERKELDER GREUNS (INACTIEF IN DEZE PROEF) EN POMP 4 VOOR NOARDBURGUM.