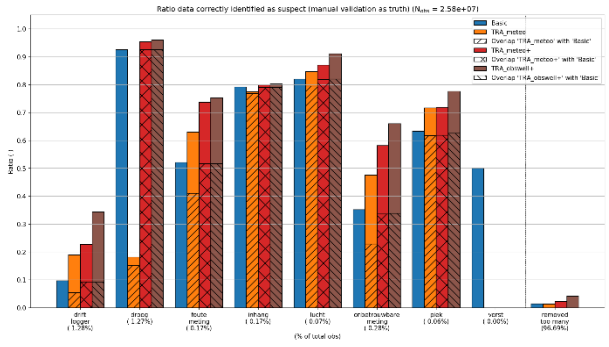
ARTESIA

WATER RESEARCH UNLIMITED

Automatische foutherkenning met tijdreeksanalyse

TRAVAL



Schoonhoven 2020

ARTESIA
WATER RESEARCH UNLIMITED

Automatische foutherkenning met tijdreeksanalyse

TRAVAL

© 2020 Artesia B.V.

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.

Projectnummer: 19.014.000

Datum: 11-3-2020

Auteur(s): Davíd Brakenhoff

Borging: Frans Schaars

Inhoud

1 Inleiding

1.1 Achtergrond

Het valideren (en eventueel corrigeren) van gemeten grondwaterstanden gebeurt in de praktijk nog vaak met de hand (al dan niet ondersteund door software). In sommige gevallen wordt er automatische foutherkenning uitgevoerd, bijvoorbeeld bij het invoeren van de data in databases, maar dan is de foutherkenning beperkt tot simpele eenvoudig herkenbare gevallen. In veel gevallen zijn de correcties zelfs helemaal niet uitgevoerd, onvolledig, of wordt er data onterecht verwijderd (bijvoorbeeld gedurende uitzonderlijke weersomstandigheden). Automatische foutherkenning van data met tijdreeksanalyse zou de kwaliteit van de gegevens aanzienlijk kunnen verbeteren. Het voordeel van automatische foutherkenning is dat het snel is (meteen beschikbaar), tijd (en dus kosten) bespaart, het reproduceerbaar is en kwalitatief betere reeksen oplevert (minder gaten door tijdige signalering foute metingen). Een bijkomend voordeel van tijdreeksanalyse is dat het ook mogelijk is om met een bandbreedte de periodes waarin metingen ontbreken op te vullen. Er zijn ook nadelen aan automatische validatie, bijvoorbeeld het onterecht corrigeren van (juist) interessante data of dat er minder aandachtig met de data gewerkt wordt doordat het proces geautomatiseerd is (minder “in gesprek” met de data). Deze nadelen wegen naar verwachting echter niet op tegen de voordelen van een robuust en betrouwbare methode voor het automatisch herkennen van fouten in data. Ook is het mogelijk om in de implementatie van de automatische foutherkenning ervoor te zorgen dat de verdachte metingen zichtbaar blijven, en niet voor de gebruiker verdwijnen.

In de praktijk is er vraag naar dergelijke automatische methodes om meetnetbeheerders te helpen om sneller te reageren als er meetfouten worden geconstateerd in een meetpunt, of om de kwaliteit van het meetnet te verbeteren.

Hoewel het nu al mogelijk is om een tijdreeksmodel op te stellen voor een bepaald meetpunt en daarna het model te vergelijken met nieuwe metingen, ontbreekt het momenteel aan kennis over algoritmes om te bepalen wanneer een meting verdacht is, of kennis over welke methode goed werkt om bepaalde type fouten te ontdekken. Ook is nog niet onderzocht wat de invloed is van de onzekerheid van het tijdreeksmodel op de bruikbaarheid ervan voor het signaleren van verdachte metingen.

In dit project willen we deze onderwerpen onderzoeken, en vervolgens de uitkomsten daarvan toepassen op data uit de praktijk. Het project is dus opgedeeld in een deel theoretisch onderzoek en een deel dat de verworven kennis toepast op datasets uit de praktijk. Uit het onderzoek hopen we statistisch onderbouwde methodes af te leiden die in de praktijk toegepast kunnen worden voor het valideren van metingen.

1.2 Doel

Het doel van dit project is om te onderzoeken in hoeverre stijghoogte metingen op een verantwoorde manier automatisch gevalideerd kunnen worden met behulp van tijdreeksanalyse. Om dit te onderzoeken zijn de volgende onderzoeksvragen geformuleerd:

- Hoe bereken je een betrouwbaarheidsinterval voor een simulatie met een tijdreeksmodel?
- Hoe beïnvloedt onzekerheid van een tijdreeksmodel (de model fit en de parameter onzekerheid) de bruikbaarheid ervan voor automatische validatie?

- Welke algoritmes of methodes kunnen worden afgeleid om meetfouten in data te identificeren?
- Hoe kan ik tijdreeksanalyse gebruiken om automatisch verdachte metingen te signaleren?
- Hoe verhoudt een methode op basis van tijdreeksanalyse zich tot een methode op basis van eenvoudige statistische regels? Hoe verhouden beide methodes zich tot de handmatige validatie?

1.3 Aanpak

In dit onderzoek zijn de bovenstaande onderzoeksvragen beantwoordt door automatische foutherkenningsmethodes te ontwikkelen en die vervolgens toe te passen op vijf verschillende datasets. Daarbij is als uitgangspunt gehanteerd dat de handmatige validatie goed is. Deze aanname is nodig om een objectieve vergelijking te kunnen maken tussen verschillende methodes. Met de handmatige validatie als “de waarheid” is gekeken hoe complexere foutherkenningsmethodes op basis van tijdreeksanalyse zich verhouden tot foutherkenningsalgoritmes gebaseerd op simpele statistische regels. Er is gekozen om de vergelijking te maken voor de gehele datasets om vast te kunnen stellen wat de methode oplevert in de praktijk.

Aanvullend onderzoek naar het herkennen van drift in tijdreeksen is opgenomen in Bijlage 5.

1.4 Leeswijzer

In hoofdstuk 2 wordt de data per organisatie kort toegelicht en wordt uitgelegd hoe een uniforme dataset is gecreëerd voor de verdere analyses. In hoofdstuk 3 worden de tools beschreven die zijn gebruikt voor het uitvoeren van de analyses. De automatische foutherkenningsmethodes zijn in hoofdstuk 4 beschreven. De resultaten zijn opgenomen in hoofdstuk 5. Hoofdstuk 6 bevat de discussie van de resultaten. Hoofdstuk 7 bevat de conclusies van dit onderzoek en in hoofdstuk 8 zijn aanbevelingen opgenomen voor vervolgonderzoek.

2 Data

2.1 Inleiding

In dit hoofdstuk is de beschikbare data beschreven. Er is niet gekeken naar individuele datasets, maar naar kenmerken van de datasets als geheel. Voor gedetailleerde informatie over de verschillende datasets wordt verwezen naar Bijlage 1.

De data van de verschillende organisaties wordt op verschillende manieren behandeld en beheerd. Dit betekent dat de datasets eerst omgezet moeten worden naar een uniform formaat zodat de automatische foutherkenningsmethodes toegepast kunnen worden. Alle datasets zijn ingelezen en omgezet naar een standaard formaat in een database (met behulp van de Python packages Pandas, Pystore en Arctic[1]). Daarna is per dataset gekeken naar de verschillen tussen de ruwe en gevalideerde tijdreeksen en de opmerkingen bij de validatie. De resultaten van deze vergelijking zijn in onderstaande paragrafen opgenomen.

2.2 Verschil tussen ruw en gevalideerd

In Figuur 1 is de verdeling weergegeven van de wijzigingen ten opzichte van de ruwe reeksen na de handmatige validatie. In het validatieproces kunnen er drie dingen gebeuren met een ruwe meting (de kleur tussen haakjes komt overeen met de kleur voor die categorie in Figuur 1):

- **Identical (groen):** De meting werd als goed beoordeeld en blijft hetzelfde;
- **Changed (rood):** De meting werd als fout beoordeeld maar kan worden gecorrigeerd en wordt dus aangepast;
- **Removed in validation (paars):** De meting werd als fout beoordeeld en kan niet worden gecorrigeerd en wordt dus 'verwijderd'.

In Figuur 1 geeft de binnenste cirkel aan welk aandeel van elke groep is voorzien van een opmerking (indien deze data beschikbaar was). In een ideale wereld zouden alle gewijzigde of verwijderde metingen voorzien zijn van een opmerking, maar dat is niet het geval. Niet alle locaties konden worden meegenomen in deze analyse om diverse redenen:

- Geen ruwe en gevalideerde tijdreeks beschikbaar;
- Tijdreeks bevatte geen data, of kon om een andere reden niet ingelezen worden.

Opvallend (maar niet verrassend) zijn de verschillen in de verdelingen tussen de verschillende datasets. Bij Rijn en IJssel wordt metingen vrijwel nooit aangepast, terwijl bij Vitens ca 83% van de metingen wordt aangepast. Dit heeft alles te maken met de verschillende handmatige validatieprocessen die bij de verschillende organisaties zijn geïmplementeerd.

[1] Voor meer technische informatie over hoe het databeheer is ingericht voor dit project, neem contact op met Artesia.

Figuur 1 Verdeling van veranderingen ten opzichte van de ruwe tijdreeks na de handmatige validatie. In de binnenste cirkel is aangegeven hoeveel van de metingen binnen een groep zijn voorzien van een opmerking. Bovenaan de diagrammen staat het aantal locaties dat is meegenomen in deze vergelijking.

2.3 Opmerkingen bij validatie

In dit project is gekozen om negen verschillende categorieën te hanteren voor foute metingen. Dit zijn grotendeels de standaard categorieën voor het labelen van foute metingen in het softwareprogramma ArtDiver:

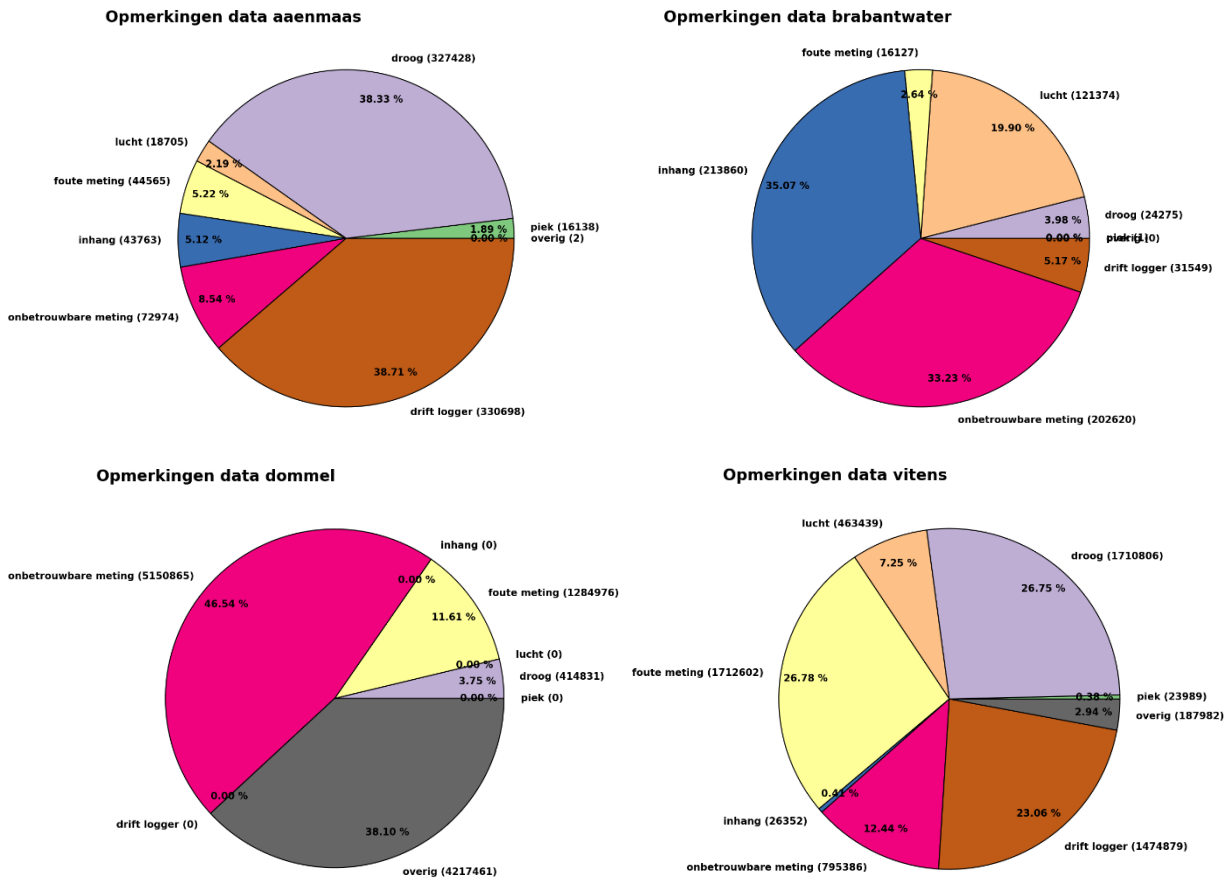
- **Piek:** een plotse opvallende toename van de meetwaarde gevolgd door plotse opvallende afname van vergelijkbare grootte.
- **Droog:** de logger hangt droog en meet dus de luchtdruk, vaak zichtbaar door een vlakke periode onderaan een tijdreeks.
- **Lucht:** de logger meet de luchtdruk omdat de logger niet in de peilbuis hangt, vaak zichtbaar als neerwaartse pieken in de tijdreeks.
- **Inhang:** een fout gerelateerd aan de inhangdiepte van de logger, vaak gekenmerkt door periodes in de tijdreeks die op een ander niveau lijken te liggen.
- **Foute meting:** de meting ligt buiten het meetbereik van de logger
- **Onbetrouwbare meting:** de meting ligt binnen het meetbereik van de logger maar is toch verdacht.
- **Drift logger:** de logger metingen gaan in de tijd steeds meer afwijken van de daadwerkelijke grondwaterstand door het "driften" van de logger. De mate van drift kan variëren tussen enkele cm per jaar tot tientallen cm in een korte periode.
- **Vorst**[2]**:** Het water in de peilbuis is bevroren waardoor de metingen niet meer te relateren zijn aan een grondwaterstand.
- **Overig:** alle fouten die niet in een van de andere categorieën ondergebracht kan worden.

Alle opmerkingen in de datasets van de verschillende organisaties zijn vertaald naar een van deze categorieën. Voor de koppeltabellen die daarvoor zijn gebruikt wordt verwezen naar Bijlage 1. In sommige gevallen waren meerdere opmerkingen beschikbaar per meting. In die gevallen is de eerste opmerking gebruikt om de vertaling te maken.

In Figuur *2* is de verdeling van de opmerkingen voor vier datasets weergegeven in taartdiagrammen. Uit de figuur is duidelijk te zien dat de frequentie waarmee fouten voorkomen per dataset sterk verschilt. Dat is niet verrassend omdat verschillende validatieprocessen toegepast worden en verschillende mensen fouten anders zullen categoriseren. Toch is het interessant om de verdeling van het type fout per organisatie te zien, omdat dit toch iets zegt over welke fouten veel voorkomen.

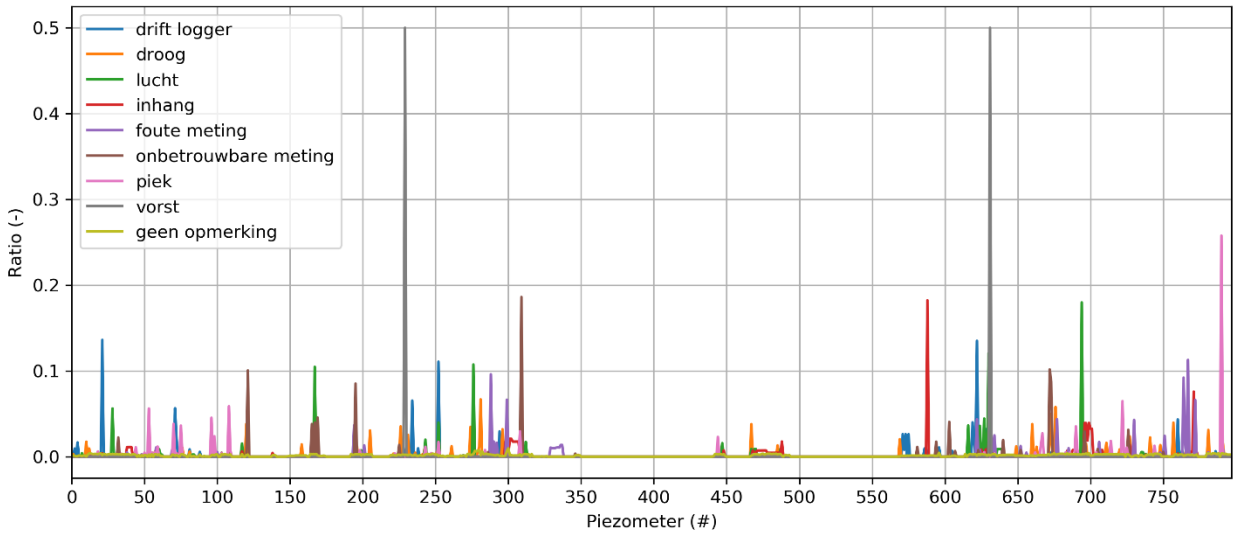
De data van Rijn en IJssel is niet beschouwd omdat er geen opmerkingen beschikbaar zijn. Voor de dataset Aa en Maas (telemetrie) is er ook geen figuur opgenomen omdat de figuur weinig zegt. Van de opmerkingen vallen 99.9% (7858 metingen, 0.04% van het totale aantal metingen) in de categorie "inhang". Verder zijn er nog 4 metingen voorzien van het label "foute meting" en 3 van het label "onbetrouwbare meting".

[2] Deze categorie had eigenlijk onder "overig" moeten vallen maar is in de analyse per ongeluk apart gebleven. Er valt maar een relatief klein aandeel van alle foute metingen in deze categorie.

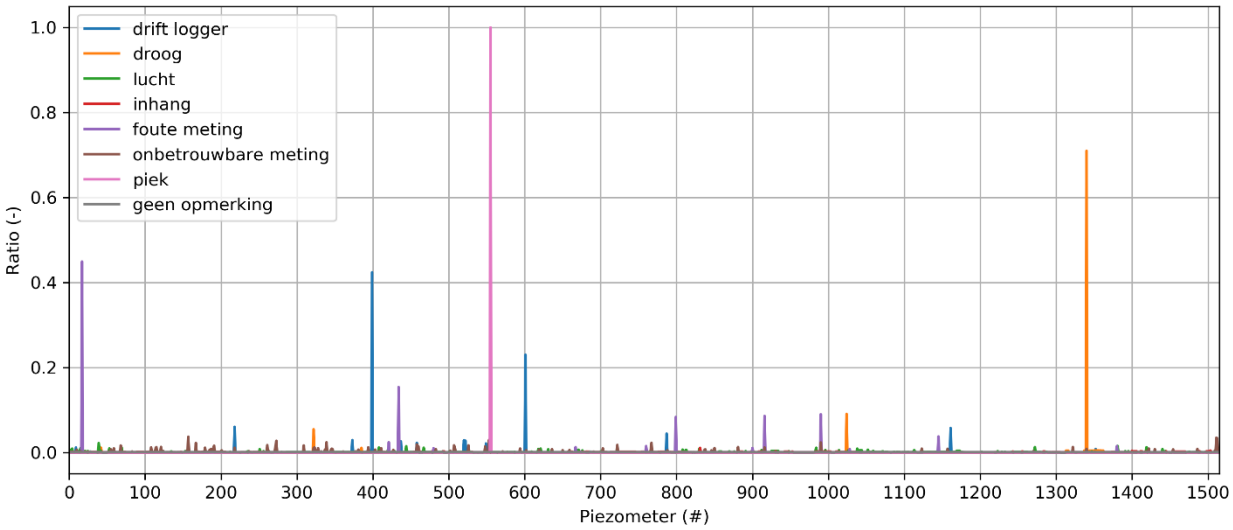


Figuur 2 De verdeling van de opmerkingen per dataset (na vertaling naar de 8 standaard categorieën gebruikt in dit onderzoek). In deze figuur is de categorie "vorst" ook onder "overig" geplaatst. De datasets Aa en Maas (telemetrie) en Rijn en IJssel zijn niet weergegeven.

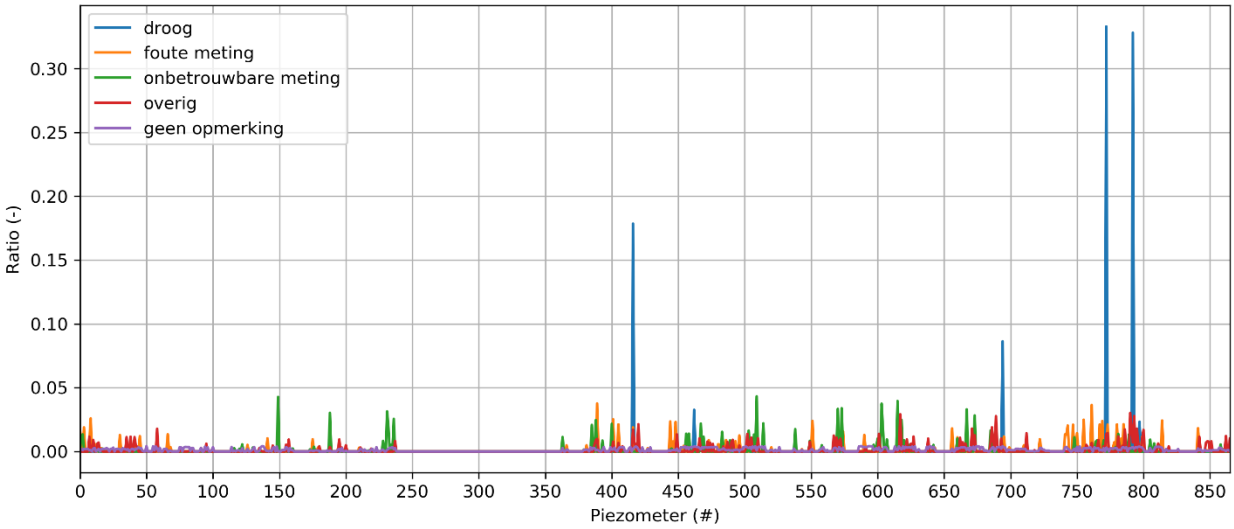
De verdeling van de opmerkingen in de handmatige validatie over de verschillende tijdreeksen is hieronder per dataset weergegeven. Hieruit is zichtbaar dat sommige fouten voornamelijk in een klein aantal tijdreeksen voorkomen, bijvoorbeeld dat 70% van de droogval in de dataset van Brabant Water voorkomt in één peilbuis (zie Figuur 4). De uitschieters zijn interessant om te bekijken omdat de resultaten van de automatische foutherkenning erg afhankelijk zijn van de score voor die ene peilbuis.



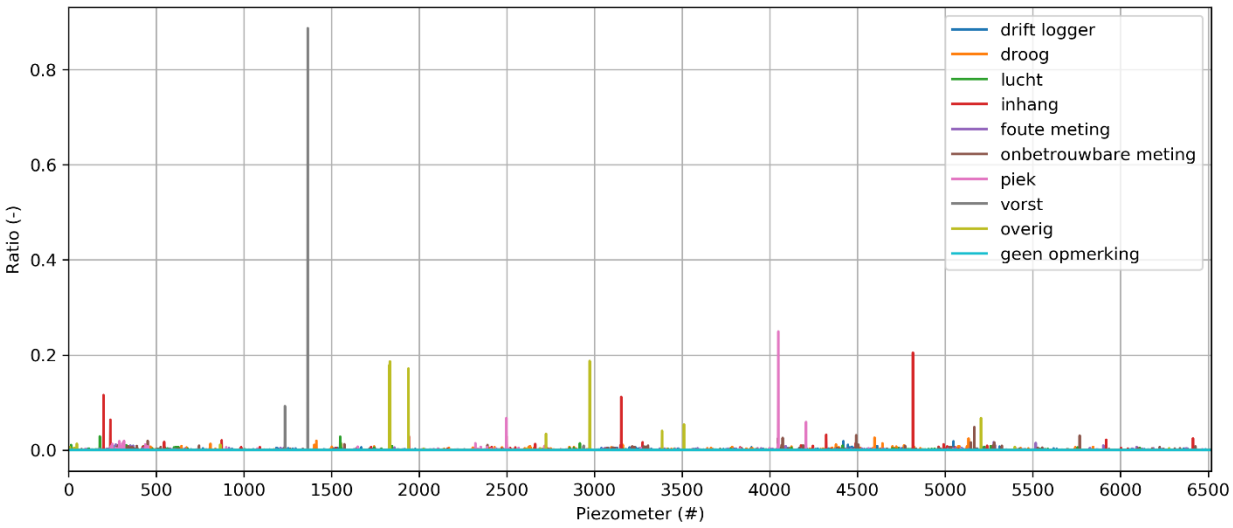
Figuur 3 Verdeling van opmerking in de handmatige validatie over alle tijdreeksen in de dataset van Aa en Maas.



Figuur 4 Verdeling van opmerking in de handmatige validatie over alle tijdreeksen in de dataset van Brabant Water.



Figuur 5 Verdeling van opmerking in de handmatige validatie over alle tijdreeksen in de dataset van De Dommel.



Figuur 6 Verdeling van opmerking in de handmatige validatie over alle tijdreeksen in de dataset van Vitens.

2.4 Voorbewerking voor automatische foutherkenning

In de handmatige validatie worden foute metingen aangepast of verwijderd. Het verwijderen van een meting duidt meestal op een meetfout terwijl een aanpassing vaker te maken heeft met het feit dat er een verschil geconstateerd werd tussen de logger meting en een handpeiling. De loggermetingen worden in dat geval in hoogte bijgesteld zodat deze op de handpeilingen vallen. Deze aanpassing is meestal in de ordegrootte van enkele centimeters. Deze kleine aanpassingen zijn lastig vindbaar voor automatische algoritmes (die niet de handpeilingen beschouwen). De mensen die de handmatige validatie uitvoeren kunnen dit type “fout” tenslotte ook alleen vinden met behulp van de handpeilingen.

Gezien het grote aandeel aanpassingen van de metingen in een aantal van de datasets (Vitens 83%, Aa en Maas 62%, Brabant Water 44%) is het niet logisch om dit type “fouten” mee te nemen in de beoordeling van automatische foutherkenningsalgoritmes. Deze fouten zijn niet vindbaar met de beoogde methodes en dus is het niet praktisch om die in de beoordeling mee te nemen. Om deze reden is ervoor gekozen om de methodes alleen te scoren op het identificeren van fouten waarbij de meting in de handmatige validatie is verwijderd. Dit heeft als gevolg dat de ruwe tijdreeks aangepast moet worden, zodat alleen de grove fouten nog in de reeks zitten, en niet de aanpassingen van de metingen. Voor deze studie is dus een zogenaamde “synthetisch ruwe” tijdreeks gecreëerd. In deze synthetisch ruwe reeks zitten alleen de grove fouten, en alle andere metingen zijn gelijk aan de handmatige gevalideerde reeks.

3 Framework automatische foutherkenning

3.1 Inleiding

Er is een generiek framework ontwikkeld in voor het toepassen van verschillende automatische foutherkenningsalgoritmes op datasets. Dit framework bestaat uit een aantal basiscomponenten:

- Vergelijking van tijdreeksen
- Een generieke structuur voor vastleggen foutherkenningsalgoritmes
- Een bibliotheek van generieke foutherkenningsregels
- Tools voor het toepassen van foutherkenningsalgoritmes en het opslaan van de resultaten

Verder zijn nog enkele geavanceerdere tools voor het optimaliseren van parameters en het beheren van grote hoeveelheden data. Deze tools worden in de laatste paragraaf omschreven.

3.2 Vergelijking van tijdreeksen

De vergelijking van tijdreeksen wordt gedaan door de SeriesComparison tool. Met deze tool zijn twee type vergelijkingen mogelijk:

- Twee tijdreeksen worden onderling vergeleken
- Twee tijdreeksen worden met elkaar vergeleken ten opzichte van een derde tijdreeks

In de onderlinge vergelijking wordt gekeken welke waardes identiek zijn, welke verschillen, en welke wel in de een maar niet in de ander aanwezig zijn. In de vergelijking kan een grenswaarde ingesteld worden waarboven twee metingen significant van elkaar verschillen. Standaard staat deze waarde op 1 mm. Deze vergelijking levert de volgende statistieken op:

- Welke metingen zijn in beide reeksen aanwezig en zijn identiek (verschil kleiner dan grenswaarde)
- Welke metingen zijn in beide reeksen aanwezig en zijn verschillend (verschil groter dan grenswaarde)
- Welke metingen zitten alleen in tijdreeks 1
- Welke metingen zitten alleen in tijdreeks 2

	In beide, identiek of verschillend (kept in both, identical/different)	Alleen in reeks 1 (only in series1)	Alleen in reeks 2 (only in series2)
Tijdreeks 1	X	X	-
Tijdreeks 2	X	-	X

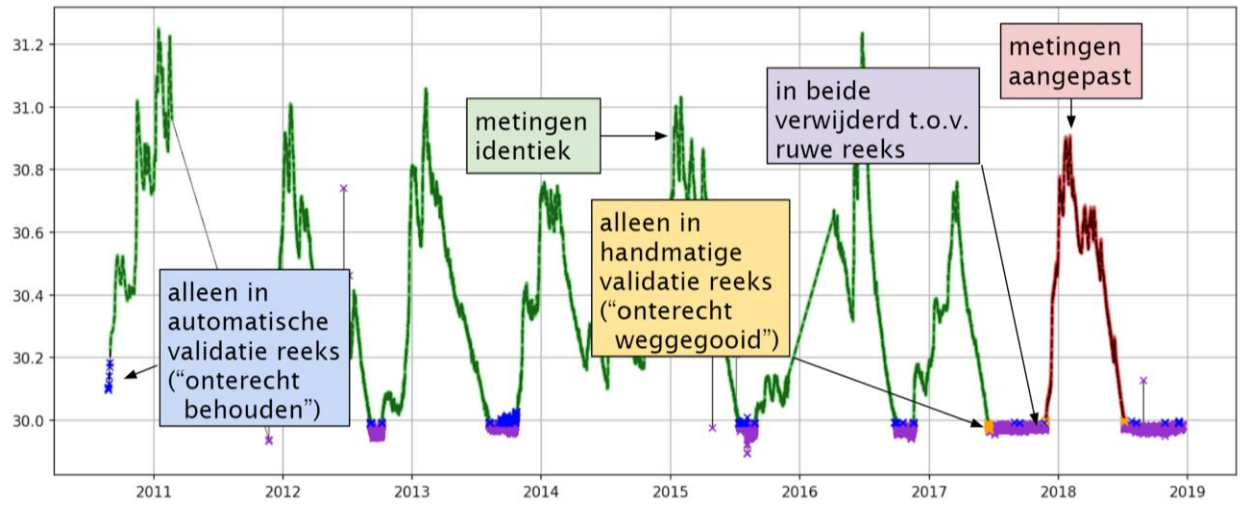
Het tweede type vergelijking voegt hier nog een extra categorie aan toe door de twee tijdreeksen nog met een derde (ruwe) tijdreeks te vergelijken. Hiermee worden twee verschillende gevalideerde tijdreeksen met een ruwe tijdreeks vergeleken. Hierdoor kunnen de volgende categorieën worden onderscheiden:

- Welke metingen zijn in beide tijdreeksen behouden t.o.v. de ruwe tijdreeks
- Welke metingen zitten alleen in tijdreeks 1 en de ruwe tijdreeks
- Welke metingen zitten alleen in tijdreeks 2 en de ruwe tijdreeks
- Welke metingen zijn in beide tijdreeksen verwijderd vergeleken met de ruwe tijdreeks.

Uitgangspunt in de vergelijkingen is dat metingen met een invalide waarde (b.v. NaN) worden beschouwd als ontbrekend (dus niet in de tijdreeks zitten). Deze metingen worden op dezelfde manier beschouwd als metingen die wel aanwezig zijn in een reeks maar niet in de ander (dus helemaal niet zijn geregistreerd).

	In beide behouden (kept in both)	Alleen in reeks 1 (only in series1)	Alleen in reeks 2 (only in series2)	In beide verwijderd (Removed in both)
Tijdreeks 0	X	X	X	X
Tijdreeks 1	X	X	–	–
Tijdreeks 2	X	–	X	–

In Figuur 7 is een vergelijking van twee reeksen (in dit geval twee verschillende gevalideerde reeksen) met een derde reeks (hier de ruwe data) visueel weergegeven voor een willekeurige peilbuis uit de dataset van Aa en Maas.



Figuur 7 Voorbeeld van een vergelijking van 2 reeksen (b.v. twee op verschillende manieren gevalideerde tijdreeksen) met een derde (ruwe) reeks.

3.3 Structuur foutherkenningsalgoritmes

De structuur voor het vastleggen van foutherkenningsalgoritmes is een boomstructuur die is vastgelegd in een tabelvorm. Elke foutherkenning regel wordt genummerd, en er wordt per regel aangegeven op welk tussenresultaat die regel moet worden toegepast. Met deze structuur kunnen bijvoorbeeld ook algoritmes worden vastgelegd waarbij alle regels op de originele (ruwe) tijdreeks worden toegepast en vervolgens worden de resultaten samengevoegd tot een automatisch gevalideerde reeks.

Hieronder staat een voorbeeld tabel, waarin twee regels worden toegepast om verdachte metingen te signaleren: “droogval” en “spikes”. Deze twee regels worden beide toegepast op de originele tijdreeks en leveren beide een resultaat op waarin verdachte metingen zijn gesignaleerd. De laatste stap is het combineren van de resultaten van deze twee regels.

Stap	Regel	Toepassen op
1	Droogval detecteren	0 (originele reeks)
2	Spikes vinden	0
3	Combineren resultaten van stap 1 en 2	(1, 2)

3.4 Bibliotheek met generieke foutherkenningsregels

Er is een bibliotheek met generieke foutherkenningsregels geschreven waarmee verdachte metingen gesignaleerd kunnen worden. Deze regels zijn geschreven als functies die een tijdreeks opleveren met daarin zogenaamde “correcties” op de oorspronkelijke tijdreeks. Verdachte metingen worden omgezet naar een invalide waarde (b.v. NaN). In deze functies worden overige parameters meegegeven (b.v. de maximale NAP waarde waarboven metingen als verdacht moeten worden gemarkeerd) waarmee de regel toegepast kan worden op een willekeurige tijdreeks.

Voor de parameters van deze regels kunnen ook functies worden opgegeven (in plaats van een vaste waarde). In die situatie wordt de functie aangeroepen met de naam van de tijdreeks en wordt het resultaat daarvan als argument gebruikt in de foutherkenningsregel. Hiermee kan bijvoorbeeld data worden opgehaald uit een database op het moment dat die data nodig is om fouten op te sporen voor een bepaalde tijdreeks. Dit geeft de flexibiliteit om met dezelfde foutherkenningsregel verschillende parameters te gebruiken voor verschillende peilbuizen.

In onderstaande tabel is een overzicht opgenomen van de regels die zijn toegepast in deze studie. Daarin staat per regel een korte en langere omschrijving en de functienaam zoals die in Python is gedefinieerd. Voor een tabel met alle beschikbare regels wordt verwezen naar Bijlage 2, voor meer informatie wordt verwezen naar de documentatie van de betreffende Python functies.

Tabel 1 Korte omschrijving, Python functienaam en uitgebreidere omschrijving van de verschillende generieke foutherkenningsregels die zijn toegepast in dit onderzoek.

Regel	Python functienaam	Omschrijving
Groter dan tijdsafhankelijke grenswaarde	val_gt_threshold_series	Als een meting groter is dan een bepaalde grenswaarde die verandert in de tijd is de meting verdacht
Kleiner dan tijdsafhankelijke grenswaarde	val_lt_threshold_series	Als een meting kleiner is dan een bepaalde grenswaarde die verandert in de tijd is de meting verdacht
Verschil tussen twee andere tijdreeksen voldoet aan een willekeurige conditie	val_diff_others_condition	Als het verschil tussen twee andere tijdreeksen voldoet aan een bepaalde conditie is de meting verdacht. De conditie is een python functie en kan alles zijn (b.v. verschil is kleiner dan een bepaalde waarde)
Detectie van spikes	val_spikes	Als er sprong boven een bepaalde grootte plaatsvindt die direct wordt gevolgd door een sprong van vergelijkbare grootte in de tegenovergestelde richting is de meting daartussenin verdacht
Buiten bandbreedte	val_tra_outside_pi	Als een meting buiten een bandbreedte ligt is de meting verdacht.
Dood signaal	val_flat_signal	Uitgebreide regel om een dood signaal te signaleren. De volgende checks kunnen worden uitgevoerd: - Maximale afwijking gedurende een periode kleiner dan een bepaalde waarde

		- Maximale spreiding t.o.v. de mediaan in die periode - De helling van een lineaire fit door de metingen in die periode is kleiner dan een bepaalde waarde - De helling voor en na de beschouwde periode hebben een tegenovergesteld teken (dal of plateau) - De meetpunten liggen onder of boven een bepaald niveau (uitgedrukt in een percentage, b.v. onder 5%-waarde van de metingen)
Combineren resultaten	combine_nan	Combineer het resultaat van N andere regels. Waar een van de reeksen een NaN bevat, bevat het resultaat ook een NaN.

3.5 Tools voor toepassen foutherkenning

Voor het toepassen en bijhouden van de resultaten van het foutherkenning algoritme is een tool ontwikkeld waarin een ruwe tijdreeks en een foutherkenning algoritme worden opgegeven. Deze tool heet ValidateTimeseries. De tool past de stappen uit het algoritme toe op de tijdreeks en slaat de tussenresultaten op. Zowel de correcties als het eindresultaat worden opgeslagen. De relatie tussen de correctie reeks en het resultaat is hieronder omschreven aan de hand van onderstaand voorbeeld. De correctie reeks bevat de aanpassingen die, als ze bij een andere reeks worden opgeteld, het resultaat van die stap opleveren.

- Resultaat stap 1 = Originele reeks + Correctie reeks stap 1
- Resultaat stap 2 = Resultaat stap 1 + Correctie reeks stap 2

Met deze resultaten is het mogelijk om na elke stap het resultaat te bekijken en is het mogelijk om te onderzoeken welke regel welke metingen als verdacht heeft gemarkeerd. Deze tool vormt samen met de het foutherkenning algoritme de basis voor het toepassen van automatische foutherkenning op een willekeurige dataset.

3.6 Overige tools

Binnen het framework zijn nog andere tools beschikbaar die of te maken hebben met het optimaliseren van parameters van de foutherkenning regels, of het managen van grote hoeveelheden data.

3.6.1 Optimalisatie tools

Er is een tool ontwikkeld (met de naam OptimizeValidation) om de parameters in een foutherkenning algoritme te optimaliseren zodanig dat het resultaat “zo min mogelijk afwijkt” van een opgegeven “waarheid”. Deze waarheid kan bijvoorbeeld de handmatig gevalideerde tijdreeks zijn. De afwijking tussen de “waarheid” en de automatische foutherkenning wordt gedefinieerd als de som van het aantal valse positieven (goede metingen die onterecht als verdacht zijn gesignaleerd) en het aantal valse negatieven

(foute metingen die onterecht zijn behouden). Door dit aantal te minimaliseren worden parameter waardes afgeleid waarbij de automatische foutherkenning de opgegeven “waarheid” zo dicht mogelijk benaderd.

Het optimalisatie algoritme maakt gebruik van Scipy minimize met als standaard instelling de “Nelder-Mead” methode. Er is ook een optimalisatie tool die werkt met een verzameling van tijdreeksen: OptimizeValCollection (zie ook volgende alinea). Daarmee worden de parameters bepaald zodanig dat de gehele dataset zo min mogelijk afwijkt van de opgegeven ‘waarheid”.

N.B. Het optimalisatie algoritme is niet uitvoerig getest en hoewel het in de voorbeelden goed lijkt te werken is het mogelijk niet geheel geschikt voor een probleem met een discrete doelfunctie. Voorzichtig gebruiken!

3.6.2 Databeheertools

De automatische foutherkenning wordt toegepast op grote datasets met veel tijdreeksen. Om dit te structureren zijn twee verschillende tools ontwikkeld, een die alles in het geheugen (RAM) van de PC berekend, en een ander die gebruik maakt van een database om resultaten tussentijds weg te schrijven.

In memory (RAM)

De tool om de foutherkenning toe te passen op een hele dataset in het geheugen heet ValidateTsCollection. Deze tool creëert eigenlijk een verzameling van ValidateTimeseries objecten en bevat functies om foutherkenning toe te passen op de hele verzameling tijdreeksen. Ook zijn er functies beschikbaar om de vergelijking te maken met bijvoorbeeld de handmatig gevalideerde tijdreeksen waarmee statistieken over de foutherkenning berekend kunnen worden voor de gehele dataset.

Database (MongoDB)

Omdat een aantal van de datasets dusdanig groot is dat het niet meer in het geheugen (32GB in de PC die hiervoor gebruikt is) past, is er ook een tool gemaakt die tussentijds alles ophaalt en wegschrijft naar een database. De tool heet ValidationProject en maakt verbinding met een MongoDB database en gebruikt de Python module Arctic om data weg te schrijven en in te lezen. Bij het toepassen van het foutherkenning algoritme wordt de ruwe reeks opgehaald uit de database. Het resultaat van de foutherkenning wordt weggeschreven in de database, en indien gewenst wordt de vergelijking gemaakt met een opgegeven “waarheid”, waarvan het resultaat ook wordt weggeschreven naar de database. Dit is langzamer, maar het geheugengebruik blijft beperkt waardoor statistieken kunnen worden berekend voor hele grote datasets.

4 Methode

In dit hoofdstuk zijn de verschillende automatische foutherkenningsalgoritmes beschreven die op de verschillende datasets zijn toegepast. Ook wordt beschreven hoe de resultaten zijn beoordeeld.

4.1 BASIS methode

Het eerste algoritme dat is toegepast is gebaseerd op simpele regels en is bedoeld om een referentiemaat te geven van hoeveel verdachte metingen terug gevonden kunnen worden met behulp van automatische regels. Dit algoritme is de BASIS methode genoemd.

Er zijn twee verschillende implementaties van deze methode toegepast omdat voor bepaalde datasets geen extra informatie beschikbaar was om droogval mee te detecteren. In de eerste methode gebeurt het detecteren van droogval op basis van de inhangdiepte van de logger of op basis van het verschil tussen de ruwe luchtdruk en de ruwe waterdruk. In de tweede methode wordt alleen de tijdreeks van de grondwaterstand beschouwd, en wordt een algoritme toegepast om “levendigheid” te toetsen. De methodes zijn hieronder opgesomd met tussen haakjes de datasets waarop die methode is toegepast).

- Basis met droogval detectie
    - Inhangdiepte (Aa en Maas (divers), Brabant Water, Vitens)
    - Ruwe druk reeksen (De Dommel)
- Basis met levendigheid (Aa en Maas (telemetrie), Rijn en IJssel)

4.1.1 BASIS met droogval detectie

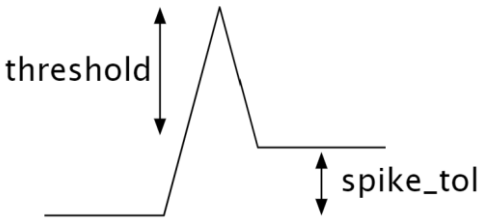
Het BASIS algoritme met droogval detectie is weergegeven in onderstaande tabel. Het algoritme bestaat uit drie regels die ieder worden toegepast op de oorspronkelijke reeks. Het algoritme zoekt spikes, droge periodes, en metingen die boven de bovenkant van de peilbuis liggen en markeert deze als verdacht. In de laatste stap worden die tussenresultaten samengevoegd.

Stap	Regel	Toepassen op
1	Spike detectie	0 (originele reeks)
2	Droogval detectie	0
3	Hard maximum	0
4	Combineren resultaten van stappen 1, 2 en 3	(1, 2, 3)

Hieronder worden de regels en de geselecteerde waardes van de parameters toegelicht. In de uitleg zijn de parameters en de opgegeven waardes voor die parameters steeds dikgedrukt.

Spike detectie

Een meting is verdacht als er een opwaartse sprong plaatsvindt die meteen (na 1 **tijdstap**, indien kleiner dan **7 dagen**) gevolgd wordt door een neerwaartse sprong, en beide sprongen groter zijn dan een grenswaarde (**threshold**, hier gekozen als **0.15 m**), of andersom. Het absolute verschil tussen de twee sprongen moet minder zijn dan een bepaalde waarde (**spike_tol**, hier gekozen als **0.15 m**).



Droogval detectie

Een meting is verdacht als de inhangdiepte t.o.v. NAP van de logger plus een **tolerantie (0.05 m)** hoger ligt dan de NAP waarde van de meting. Hiervoor moet de inhangdiepte bekend zijn. De inhangdiepte mag veranderen in de tijd.

Deze methode is toegepast voor de datasets van Aa en Maas (divers), Brabant Water en Vitens.

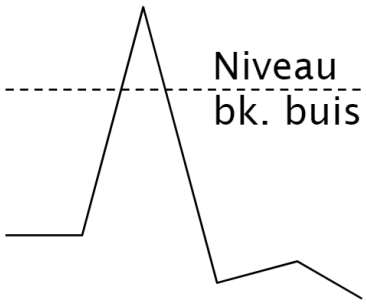
OF

Een meting is verdacht als de ruwe waterdruk reeks minus de ruwe luchtdruk reeks kleiner is dan een bepaalde **tolerantie**.

Deze methode is toegepast op de dataset van De Dommel.

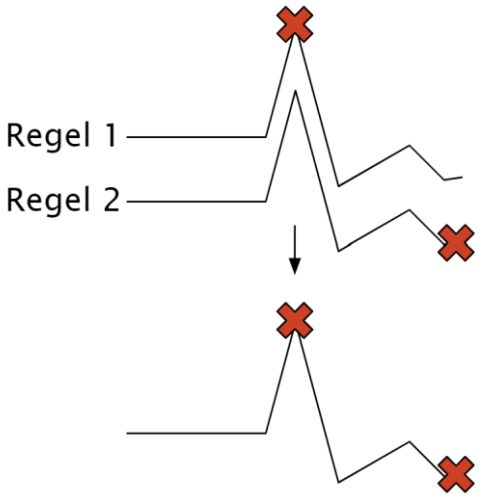
Hard maximum

Een meting is verdacht als deze hoger ligt dan de **NAP waarde van de bovenkant van de peilbuis**. Deze regel is uiteraard niet geschikt voor afgesloten peilbuizen waarbij de stijghoogte hoger ligt dan de bovenzijde van de peilbuis. In een vervolgstadium kan bij afgesloten peilbuizen de parameter voor niveau bovenkant buis aangepast worden om hier rekening mee te houden. Deze regel is toch toegepast omdat het aantal afgesloten peilbuizen klein is.



Combineren

Als een van de detectie methodes een meting als verdacht signaleert, dan is de meting verdacht. Een voorbeeld is hieronder visueel weergegeven met 2 regels:



4.1.2 BASIS met levendigheid

Voor de datasets waarvoor geen inhangdieptes van de loggers bekend zijn en er geen ruwe drukmetingen beschikbaar zijn is het niet mogelijk om de droogval regel uit het vorige algoritme toe te passen. Hiervoor is een alternatieve BASIS methode opgesteld die de levendigheid van het signaal beoordeeld. De methode zoekt naar een vlak ("dood") signaal. In dit algoritme wordt eerst spike detectie toegepast en vervolgens wordt op het resultaat het vlakke signaal algoritme toegepast.

Stap	Regel	Toepassen op
1	Spike detectie	0 (originele reeks)
2	Levendigheid	1

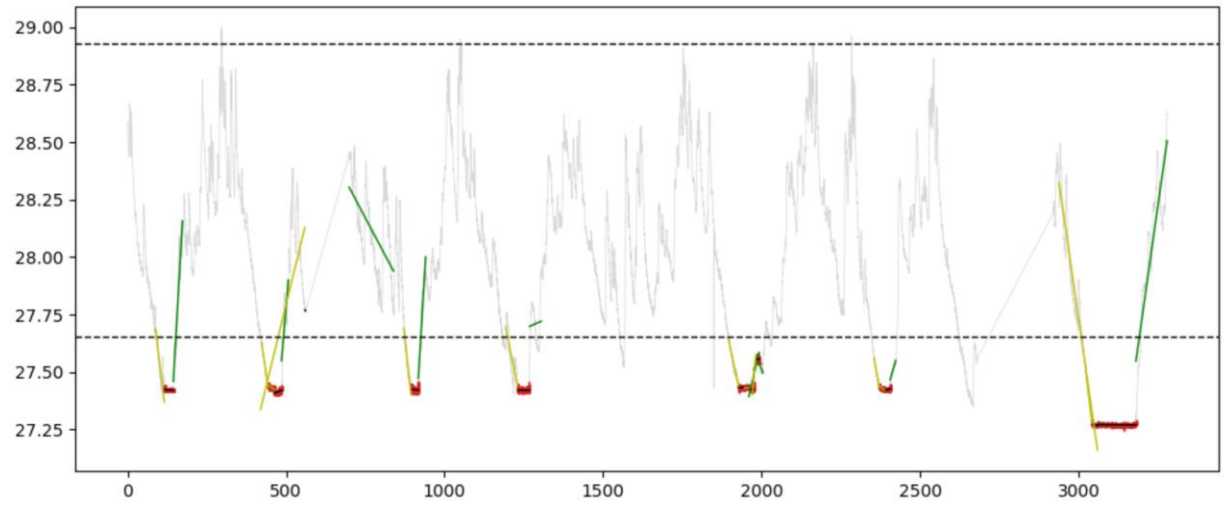
Deze methode is toegepast voor de datasets Aa en Maas (telemetrie) en Rijn en IJssel.

Levendigheid

Het zoeken naar droge periodes op basis van de eigenschappen van de metingen zelf blijkt lastig met uurlijkse metingen. De variatie van de reeks in droge periodes is vergelijkbaar met de variatie andere periodes waarin de logger niet droog hangt. Hiervoor is een algoritme ontwikkeld dat op basis van verschillende criteria iets probeert te zeggen over de levendigheid van het gemeten signaal.

Metingen worden als verdacht bestempeld als:

- De variatie tussen metingen binnen **een periode** van minimaal **10 dagen** binnen ± een grenswaarde (**0.025 m**) van elkaar liggen.
- Het **aantal metingen** in deze periode bedraagt minimaal **2 stuks.**
- De absolute waarde van **de helling** van een lineaire fit door deze punten bedraagt maximaal **0.001.**
- De **spreiding van de punten** in deze periode t.o.v. de mediaan bedraagt minder dan een grenswaarde (**0.035**).
- De metingen in deze periode liggen onder de **5%-waarde** of boven de **99%-waarde** van de reeks. Metingen die tot voor deze stap als verdacht zijn gemarkeerd zijn niet meegenomen in het berekenen van de percentielen. Hiermee worden alleen vlakke periodes die aan de onder- of bovenzijde van een reeks liggen als verdacht gemarkeerd.



Figuur 8 Voorbeeld resultaat van het algoritme voor levendigheid. In het rood zijn punten gemarkeerd waar het algoritme het signaal als "dood" inschat. De gele en groene lijnen geven de helling van de grondwaterstand voor en na de "dode stukken weer. De zwarte gestreepte horizontale lijnen geven de grenzen aan; een signaal kan alleen "dood" zijn als de grondwaterstand buiten deze grenzen ligt.

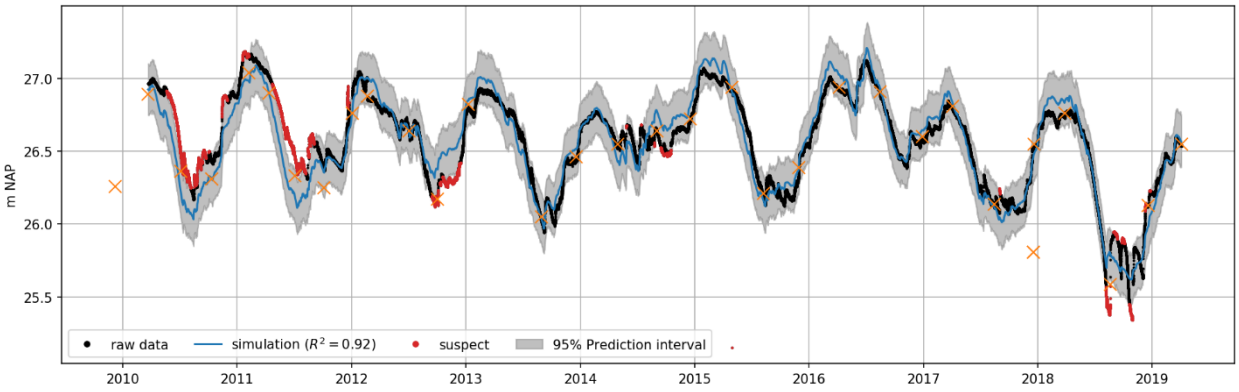
4.2 TRA meteo methode

Deze methode is gebaseerd op de uitkomsten van een tijdreeksmodel voor het betreffende meetpunt waarbij het weer (neerslag en verdamping) zijn opgenomen als de verklarende variabelen. Deze methode heeft de naam TRA_meteo gekregen.

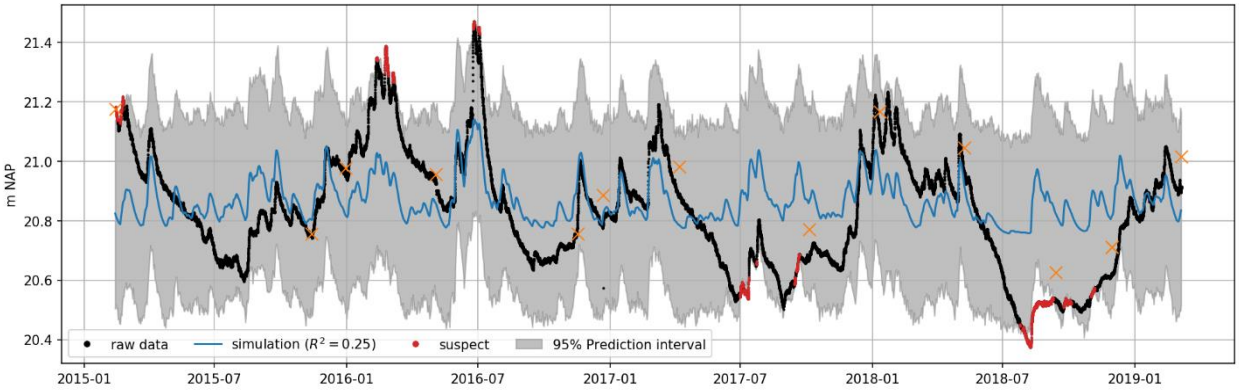
Stap	Regel	Toepassen op
1	Buiten voorspellingsbandbreedte	0 (originele reeks)

Voor alle meetpunten is een tijdreeksmodel gemaakt op basis van de dichtstbijzijnde KNMI stations waar neerslag en verdamping worden gemeten. Daarbij is de handmatig gevalideerde tijdreeks gebruikt om het model te optimaliseren. Het model is geoptimaliseerd op dag basis. Dit model is vervolgens gebruikt om een voorspellingsbandbreedte te berekenen. De voorspellingsbandbreedte is gebaseerd op de onzekerheid van de geoptimaliseerde parameters en de residuen van het model. Voor de bandbreedte moet een **waarde alfa** opgegeven worden. In deze studie is de waarde van **0.01** voor alfa gekozen. Dit komt overeen met de 99%-voorspellingsbandbreedte, dat wil zeggen dat ca. 99% van de handmatig gevalideerde metingen binnen deze bandbreedte moeten liggen. Dat betekent ook dat per definitie 1% van de handmatig gevalideerde metingen al buiten de bandbreedte ligt, en dus dat er waarschijnlijk 1% van de ruwe metingen per definitie als "onterecht verwijderd" zal worden bestempeld. De bandbreedte zou geoptimaliseerd kunnen worden om het aantal valse positieven te verkleinen.

De ruwe reeks wordt vergeleken met de bandbreedte, en alle metingen die buiten de bandbreedte liggen worden als verdacht bestempeld. Dit concept is weergegeven in Figuur 9 met een goed tijdreeksmodel en in Figuur 10 met een slecht tijdreeksmodel. In blauw is de simulatie van de grondwaterstand volgens het tijdreeksmodel weergegeven, en in het grijs de 99%-voorspellingsbandbreedte. In zwart zijn de ruwe metingen weergegeven die binnen de bandbreedte liggen, en in het rood de metingen die daarbuiten vallen. De oranje kruizen vertegenwoordigen de handpeilingen. In Figuur 9 is aan het begin van de tijdreeks zichtbaar dat de ruwe metingen afwijken van de handmetingen en dat de TRA_meteo methode die metingen terecht signaleren als verdacht. Dat lukt niet met het model uit Figuur 10.



Figuur 9 Voorbeeld van een bandbreedte op basis van een goed tijdreeksmodel waarmee ruwe metingen als verdacht worden gemarkeerd. Oranje kruisjes geven de handmetingen aan.



Figuur 10 Voorbeeld van een bandbreedte op basis van een slecht tijdreeksmodel waarmee ruwe metingen als verdacht worden gemarkeerd. Oranje kruisjes geven de handmetingen aan.

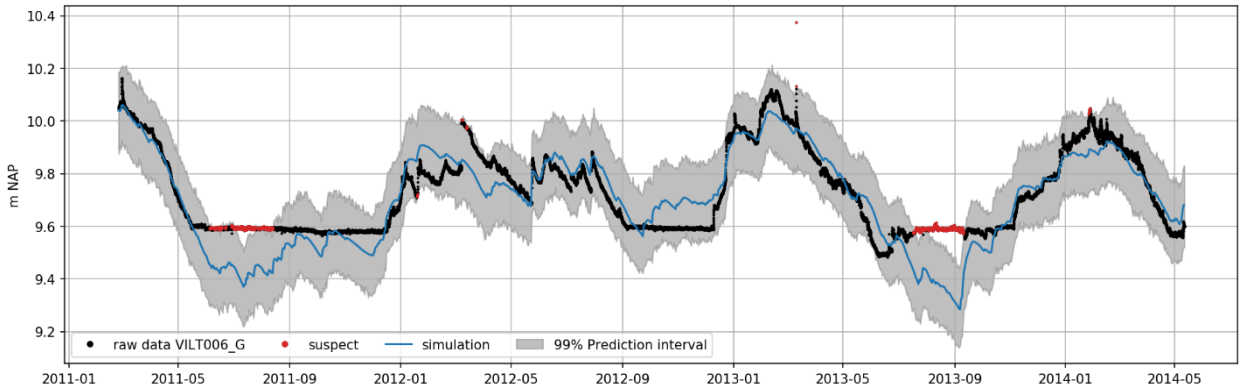
4.3 TRA meteo+ methode

Deze methode is een combinatie tussen de droogval detectie regel uit de BASIS methode en de TRA meteo methode omschreven in de vorige sub paragraaf. Deze methode heeft de naam TRA_meteo+.

Stap	Regel	Toepassen op
1	Droogval detectie	0 (originele reeks)
2	Buiten voorspellingsbandbreedte	1

In deze methode wordt eerst de droogval detectie zoals omschreven in paragraaf 4.1.1 toegepast, en vervolgens wordt het resultaat vergeleken met de bandbreedte op basis van het meteorologisch tijdreeksmodel zoals omschreven in paragraaf 4.2.

Deze methode is toegevoegd omdat uit analyses bleek dat de methode TRA_meteo relatief slecht scoorde bij het signaleren van metingen die in de handmatige validatie als droog waren bestempeld. De reden dat de tijdreeksanalyse modellen niet goed in staat zijn droogval te signaleren is waarschijnlijk omdat de loggers maar net droogvallen waardoor de stijghoogte minder ver onder de logger ligt dan de grootte van de onzekerheidsbandbreedte van het tijdreeksmodel. In Figuur 11 is een voorbeeld van de toepassing van TRA_meteo op een peilbuis met droogval opgenomen. De droge periodes worden maar gedeeltelijk als verdacht gemarkeerd omdat de droge metingen vaak nog binnen het voorspellingsinterval liggen. Als de droogval regel wordt toegevoegd worden deze metingen wel gesignaleerd.



Figuur 11 Toepassing TRA_meteo op peilbuis VILT006_G uit de dataset van Aa en Maas (divers). In het blauw is de modelsimulatie weergegeven, in het grijs het berekende voorspellingsinterval en in het zwart de ruwe tijdreeks. De rode puntjes liggen buiten de bandbreedte en zijn dus verdacht volgens de methode.

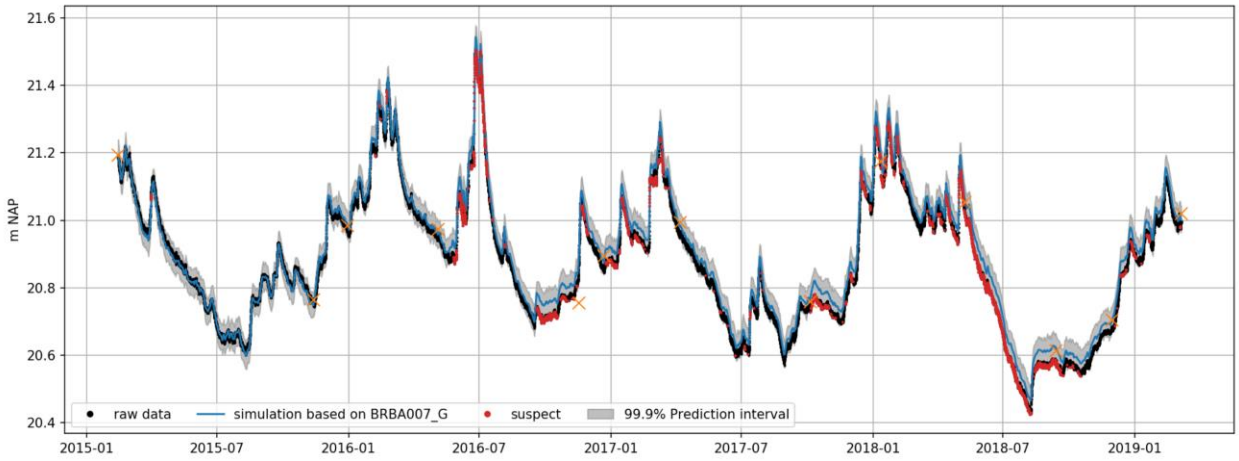
4.4 TRA obswell+ methode

Deze methode is een combinatie van de droogval detectie regel uit de BASIS methode en een tijdreeksmodel op basis van een nabijgelegen meetpunt. Deze methode heeft de naam TRA_obswell+.

Stap	Regel	Toepassen op
1	Droogval detectie	0 (originele reeks)
2	Buiten voorspellingsbandbreedte	1

Voor uitleg over de droogval detectie wordt verwezen naar paragraaf 4.1.1. De volgende stap is de vergelijking van de tijdreeks met een bandbreedte (net zoals bij TRA_meteo), maar deze bandbreedte is afkomstig van een ander type tijdreeksmodel.

Deze tijdreeksmodellen zijn gebaseerd op andere peilbuizen in de buurt. Voor elke peilbuis zijn de 5 dichtstbijzijnde meetpunten opgezocht en is voor elk punt een tijdreeksmodel afgeleid met de handmatig gevalideerde tijdreeks. Het beste model (model met de beste fit) is gebruikt om een bandbreedte te bepalen. Ook voor deze bandbreedte moet een **waarde alfa** geselecteerd worden. in dit onderzoek is **0.001** aangehouden. Dit komt overeen met het 99.9% voorspellingsinterval, met andere woorden, 99.9% van alle handmatig gevalideerde metingen ligt binnen deze bandbreedte. Een voorbeeld van de toepassing van deze regel is weergegeven in Figuur 12. Net als in de vorige paragraaf, geeft de blauwe lijn de simulatie van de grondwaterstand weer volgens het tijdreeksmodel. In grijs is de 99.9% voorspellingsbandbreedte weergegeven. In het rood zijn de punten gemarkeerd die buiten die bandbreedte vallen. Om in dit voorbeeld vast te stellen of de met rood gemarkeerde metingen daadwerkelijk verdacht zijn, moeten de handpeilingen van beide reeksen nader beschouwd worden.



Figuur 12 Voorbeeld van een bandbreedte berekend met een tijdreeksmodel op basis van een nabijgelegen meetpunt dat goed correleert met de beschouwde grondwaterstand. Oranje kruisjes geven de handpeilingen aan.

Een voordeel van deze methode is dat meetpunten in de buurt vaak goed correleren met de beschouwde peilbuis, en dus dat er hele goede modellen beschikbaar zijn. Als die twee peilbuizen een afwijking ten opzichte van elkaar vertonen is het dus zeer aannemelijk dat er iets aan de hand is, en dus dat de meting interessant of verdacht is. Het nadeel van deze methode is dat het niet eenvoudig is om automatisch vast stellen welke van de twee peilbuizen de fout bevat. Het is dus mogelijk dat goede metingen als verdacht worden bestempeld omdat in de verklarende reeks (de andere peilbuis) metingen zitten die verdacht zijn.

4.5 Samenvatting toegepaste methodes per dataset

In onderstaande tabel is samengevat welke methodes op welke dataset zijn toegepast. De toepassing van de verschillende BASIS methodes is bepaald door de aanwezigheid van informatie over inhangdieptes of de beschikbaarheid van ruwe drukmetingen. De toepassing van de TRA_meteo+ is ook afhankelijk van de beschikbaarheid van deze gegevens. De methode TRA_obswell+ is alleen op de diver dataset van Aa en Maas toegepast vanwege de eindigheid van de beschikbare tijd.

Dataset	BASIS droogval	BASIS levendigheid	TRA_meteo	TRA_meteo+	TRA_obswell+
Aa en Maas (divers)	X	–	X	X	X
Aa en Maas (telemetrie)	–	X	X	–	–
Brabant Water	X	–	X	X	–
De Dommel	X	–	X	X	–
Rijn en IJssel	–	X	X	–	–
Vitens	X	–	X	X	–

5 Resultaten

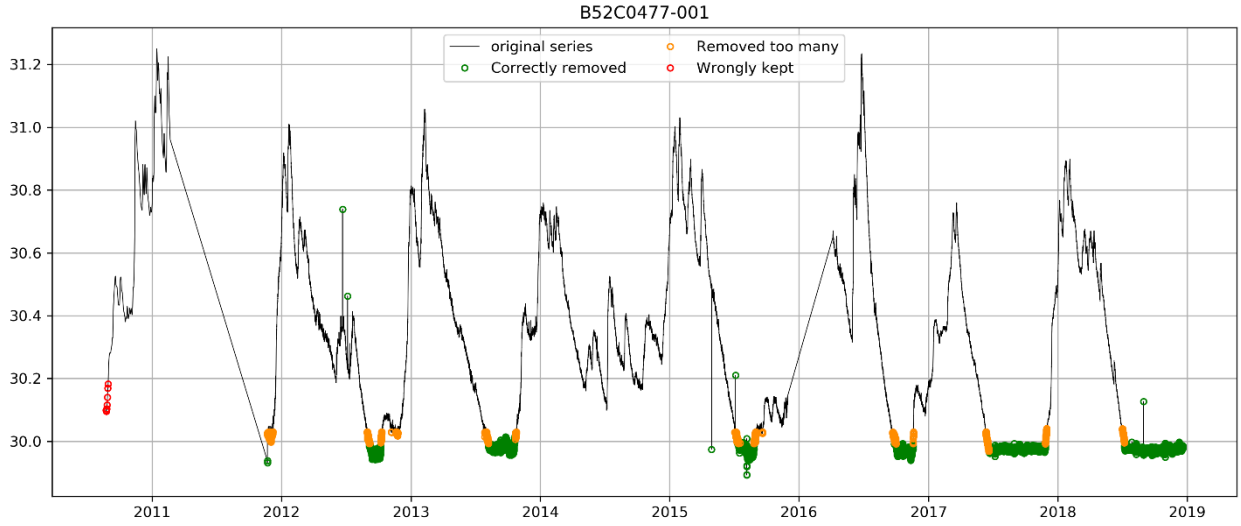
In dit hoofdstuk zijn de resultaten van de toepassing van de verschillende foutherkenningsmethodes op de datasets van de verschillende organisaties gepresenteerd. Eerst is een voorbeeld resultaat voor een enkele peilbuis gepresenteerd. Vervolgens is een overzicht gepresenteerd van hoe de verschillende methodes scoren ten opzichte van de handmatige validatie. Vervolgens zijn de resultaten gepresenteerd per dataset. Verder zijn enkele voorbeelden uitgelicht om de bruikbaarheid van automatische foutherkenning op basis van tijdreeksanalyse nader te analyseren.

5.1 Voorbeeld resultaat per peilbuis

Het resultaat van de automatische foutherkenning en de vergelijking met de handmatige validatie voor een enkele peilbuis is weergegeven in Figuur 13. De automatische foutherkenning levert een oordeel op of een ruwe meting verdacht is of niet. Als deze uitkomst wordt vergeleken met de handmatige validatie kunnen vier categorieën worden onderscheiden:

- Terecht gesignaleerd (correctly removed)
- Onterecht gesignaleerd (removed too many)
- Onterecht behouden (wrongly kept)
- Terecht behouden (correctly kept)

In Figuur 13 zijn deze categorieën ook weergegeven, hoewel de categorie “terecht behouden” niet expliciet gelabeld is. De figuur laat zien welke metingen die in de handmatige validatie zijn verwijderd wel (groen) en niet (rood) door het automatische algoritme worden gesignaleerd. Ook heeft het algoritme een bijvangst ten opzichte van de handmatige validatie. Dat zijn metingen die in de handmatige validatie niet verdacht zijn bevonden, maar door het automatische algoritme wel (oranje).



Figuur 13 Resultaat vergelijking automatische foutherkenning (BASIS methode) met handmatige validatie voor een peilbuis.

5.2 Overzicht methodes

In onderstaande tabellen is een overzicht gepresenteerd van hoe goed de verschillende automatische methodes de handmatige validatie kunnen nabootsen. Als de handmatige validatie als de waarheid wordt beschouwd (hierover later meer in de discussie in hoofdstuk 6), hoeveel van die verdachte metingen

worden dan teruggevonden door de automatische methodes, en hoeveel metingen worden er onterecht als verdacht gemarkeerd?

In Tabel 2 en Tabel 3 is zichtbaar dat de mate waarin de handmatige validatie kan worden nagebootst met automatische methodes afhankelijk is van de beschouwde dataset. Een algemene beschouwing is lastig te maken gezien de verschillen tussen de datasets, de variatie in de kwaliteit van de reeksen en van de handmatige validatie. Wel is zichtbaar dat de TRA_meteo+ methode over het algemeen beter scoort dan de BASIS methode qua percentage dat goed wordt gesignaleerd (Tabel 2) maar dat dat gepaard gaat met een toename van het aantal onterecht gesignaleerde metingen (Tabel 3). De toename van het aantal onterecht gesignaleerde metingen ligt ongeveer tussen de 0 en 3% van het totale aantal metingen. Dit zijn geen grote percentages, maar het gaat wel om een groot aantal metingen dat in deze categorie valt. In de dataset van Aa en Maas (divers) zijn het er bijvoorbeeld (25.8 – 0.85 ≈) 25 miljoen.

Met de TRA_meteo+ methode lukt het in de datasets van Aa en Maas (divers), Brabant Water en Vitens, om tussen de 68% en 83% van alle verdachte metingen te signaleren. Daar staat tegenover dat ca. 2% van de metingen onterecht worden gesignaleerd als verdacht.

In de datasets van De Dommel, Rijn en IJssel en Aa en Maas (telemetrie) zijn de scores van de automatische methodes aanzienlijk lager (tussen de 28% en 38%), maar scoren de methodes gebaseerd op tijdreeksanalyse wel hoger dan de BASIS methode. Wederom gaat deze hogere score bij de Dommel en Aa en Maas (telemetrie) gepaard met een hoger aandeel valse positieven. Bij Rijn en IJssel scoort de TRA_meteo methode beter (het aandeel valse positieven is lager).

Tabel 2 Aantal goed signaleerde metingen gegeven dat de handmatige validatie als waarheid wordt beschouwd. Het percentage goed gesignaleerd is het aantal goed gesignaleerde metingen uit het totale aantal signaleert. Groen is beter, rood is slechter.

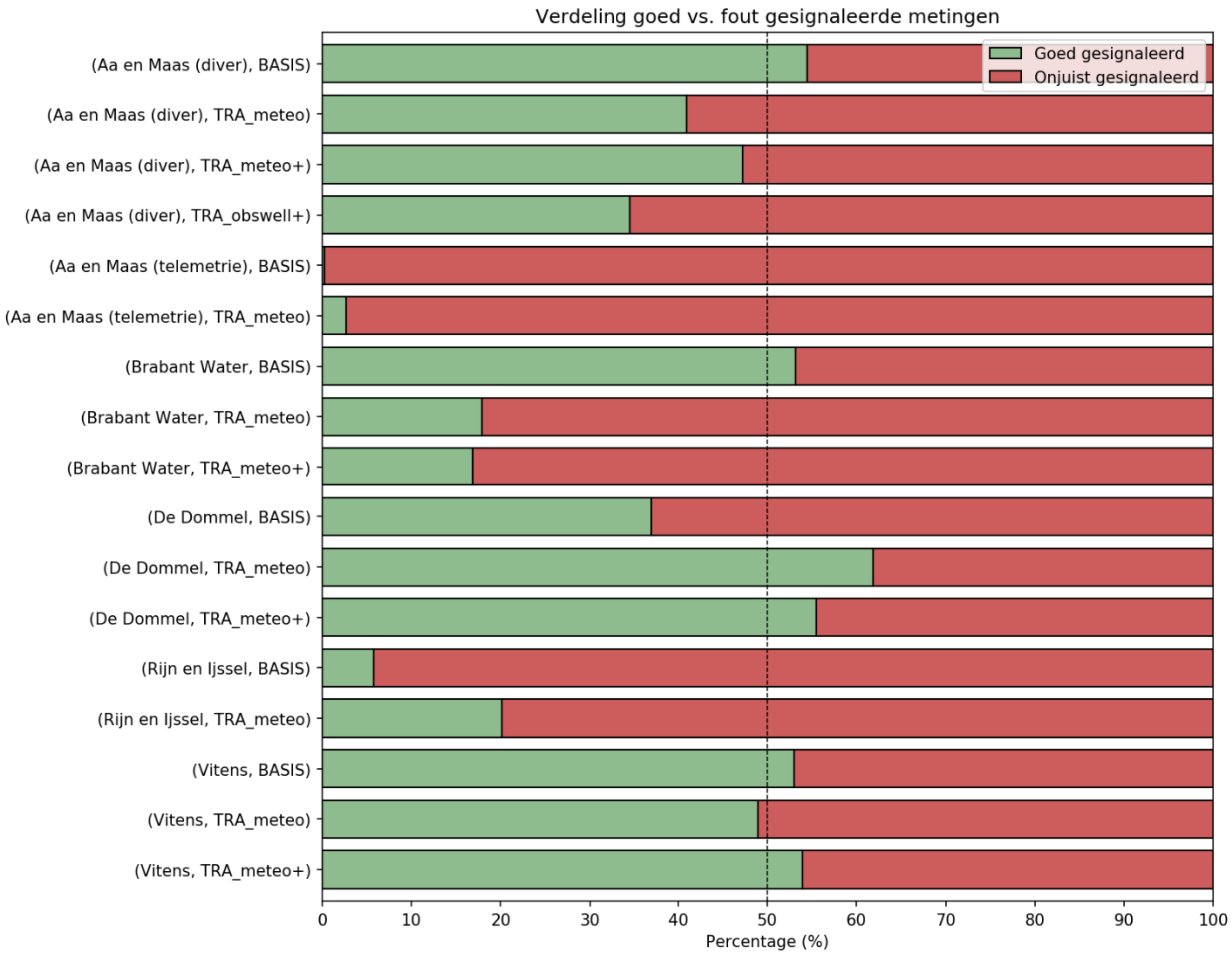
				Basic	TRA_meteo	TRA_meteo+	TRA_obswell+
Dataset	Aantal metingen (miljoenen)	Aantal gesignaleerd (miljoenen)	Aandeel gesignaleerd (%)	Goed gesignaleerd (%)	Goed gesignaleerd (%)	Goed gesignaleerd (%)	Goed gesignaleerd (%)
Aa en Maas (divers)	25.8	0.85	3.28	52.5	29.0	62.2	67.9
Aa en Maas (telemetrie)	7.6	0.02	0.23	1.9	28.2	-	-
Brabant Water	98.2	0.61	0.62	69.6	80.6	83.4	-
De Dommel	6.4	0.98	15.27	5.2	33.2	36.0	-
Rijn en Ijssel	7.2	0.07	0.98	20.7	37.4	-	-
Vitens	149.7	6.40	4.28	56.2	47.8	67.7	-

Tabel 3 Aantal teveel verwijderde metingen gegeven dat de handmatige validatie als waarheid wordt beschouwd. Het percentage teveel verwijderd is het aandeel van het aantal niet gesignaleerde metingen (aantal metingen – aantal gesignaleerd) dat onterecht als verdacht is gesignaleerd. Groen is beter, rood is slechter.

				Basic	TRA_meteo	TRA_meteo+	TRA_obswell+
Dataset	Aantal metingen (miljoenen)	Aantal gesignaleerd (miljoenen)	Aandeel gesignaleerd (%)	Teveel verwijderd (%)	Teveel verwijderd (%)	Teveel verwijderd (%)	Teveel verwijderd (%)
Aa en Maas (divers)	25.8	0.85	3.28	1.44	1.37	2.28	4.21
Aa en Maas (telemetrie)	7.6	0.02	0.23	1.83	2.34	-	-
Brabant Water	98.2	0.61	0.62	0.38	2.3	2.55	-
De Dommel	6.4	0.98	15.27	1.35	3.13	4.41	-
Rijn en Ijssel	7.2	0.07	0.98	3.34	1.46	-	-
Vitens	149.7	6.40	4.28	2.13	2.13	2.47	-

In Figuur 14 is de verdeling tussen het aantal goed en onjuist gesignaleerde metingen procentueel weergegeven. Van het aantal metingen dat als verdacht wordt bestempeld door de automatische foutherkenningsalgoritmes, geven de balken aan hoeveel er goed worden gemarkeerd (groen), en hoeveel valse positieven er zijn (rood). Uit deze figuur kan ingeschat worden wat de kans is dat een meting daadwerkelijk verdacht is als deze door een automatische methode wordt gesignaleerd. Bijvoorbeeld, voor de dataset van Vitens, als de methode TRA_meteo+ wordt toegepast, is er een ca. 53% kans dat een meting die als verdacht wordt bestempeld ook daadwerkelijk fout is.

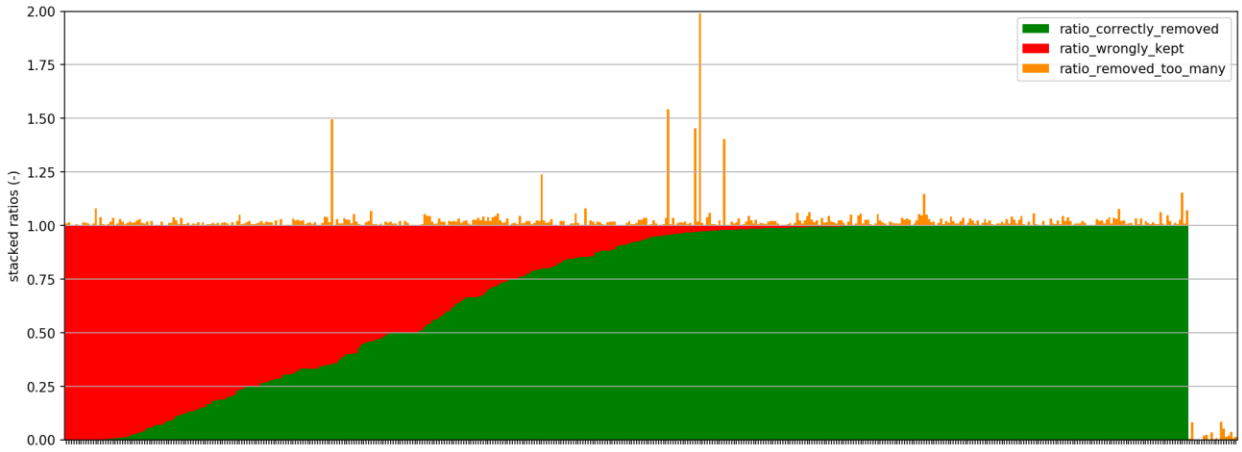
Deze uitkomsten zijn sterk afhankelijk van de handmatige validatie. Aangezien die niet perfect is moet er voorzichtig met deze resultaten worden omgegaan. Hoe beter de handmatige validatie, dus hoe dichter de handmatige validatie bij de waarheid ligt, hoe objectiever deze getallen beoordeeld kunnen worden. Als de handmatige validatie de waarheid benadert, geeft deze figuur aan welk aandeel van de gesignaleerde metingen ook daadwerkelijk als fout bestempeld zouden moeten worden bij toepassing van een automatische foutherkenningsmethode. Bij het ontwikkelen van een automatisch algoritme is het doel om het aandeel goed gesignaleerd zo hoog mogelijk te krijgen zonder dat er overdreven veel metingen onterecht worden gesignaleerd. De parameters die de foutherkenningsregels sturen hebben daar een belangrijke invloed op en hebben dus ook een invloed op deze grafiek.



Figuur 14 Verdeling aantal goede versus onjuist gesignaleerde metingen per dataset en automatische foutherkenning methode.

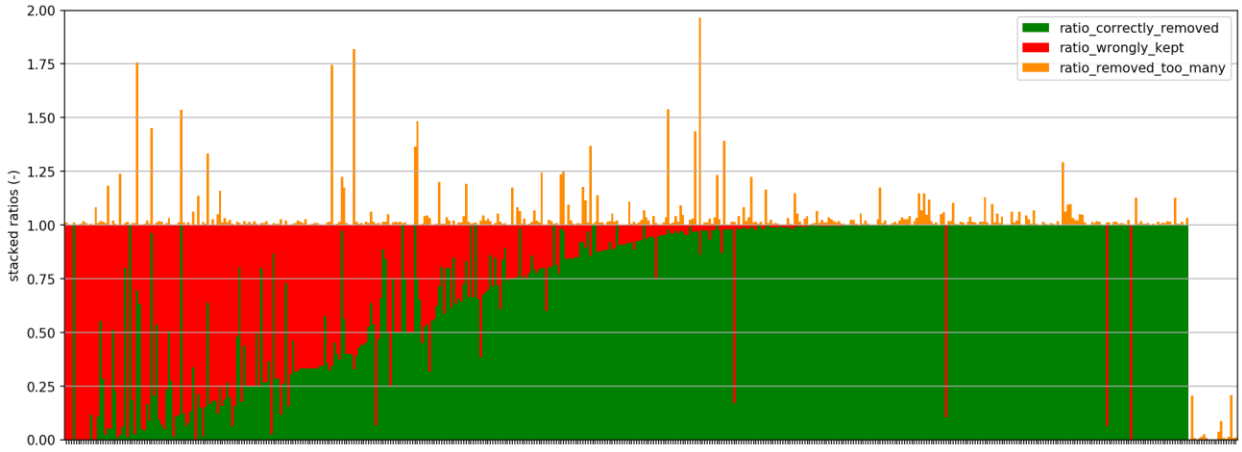
Het is belangrijk te onthouden dat er achter deze bulk statistieken allerlei verschillende tijdreeksen zitten en dat als 60% van de fouten gevonden worden met een automatisch algoritme dat dat niet betekent dat 60% van de fouten in elke individuele reeks worden gevonden. Voor sommige peilbuizen worden 100% van de fouten gevonden, en voor andere bijvoorbeeld maar 10%. Figuur 15 laat deze percentages per peilbuis zien voor de TRA_meteo+ methode toegepast op de dataset van Aa en Maas (divers). In het groen is het aandeel (tussen 0 en 1) weergegeven van fouten die door het automatische algoritme worden gesignaleerd. In het rood is het aandeel gepresenteerd van de niet-gevonden fouten. In het oranje erboven is het aandeel gepresenteerd van hoeveel van de niet foute metingen nog gesignaleerd zijn door de automatische methode (de bijvangst, valse positieven). Let op dat deze grafiek het aandeel goed en fout beschouwd, en bevat dus geen informatie over absolute getallen. Het aantal totale metingen en het aantal foute metingen verschilt per peilbuis.

Voor een uitgebreidere uitleg van de figuur en de figuren met de resultaten van de overige methodes voor alle datasets, zie Bijlage 4.



Figuur 15 Aandeel "terecht gesignaleerd" (correctly removed), "onterecht behouden" (wrongly kept) en "onterecht gesignaleerd" (removed too many) per peilbuis voor de dataset van Aa en Maas (divers). Elk balkje is het resultaat voor een peilbuis van de vergelijking van de automatische foutherkenning met TRA_meteo+ met de handmatige validatie. Aan de rechterzijde van de grafiek zijn er voor enkele peilbuizen geen groene/rode balken omdat de ruwe reeksen volgens de handmatige validatie geen fouten bevatten.

Deze grafiek is in Figuur 16 gepresenteerd voor de methode TRA_obswell+ maar dan met dezelfde volgorde van meetpunten als in de vorige figuur. Hierdoor is het mogelijk om iets te zeggen (in kwalitatieve zin) over de overeenkomsten tussen de TRA_meteo+ en TRA_obswel+ methodes. Worden in beide methodes in dezelfde reeksen de fouten goed herkend? Of zijn er ook duidelijke verschillen waarbij de ene methode de fouten wel herkent maar de andere niet. Uit de figuur kan opgemaakt worden dat een groot deel van de reeksen waarvoor de methode TRA_meteo+ goed scoort, de methode TRA_obswell+ ook goed scoort (het oplopende patroon blijft zichtbaar). Aan de rechterzijde zijn enkele rode balken zichtbaar: dit zijn reeksen waar de methode TRA_obswell+ slechter presteert dan TRA_meteo+. De groene balken aan de linkerzijde zijn juist punten waarin de methode TRA_obswell+ veel beter presteert dan TRA_meteo+. In veel gevallen presteren beide methodes vergelijkbaar, maar zijn er ook duidelijk reeksen waarbij de ene methode een goede aanvulling is op de andere.



Figuur 16 Aandeel "terecht gesignaleerd" (correctly removed), "onterecht behouden" (wrongly kept) en "onterecht gesignaleerd" (removed too many) per peilbuis voor de dataset van Aa en Maas (divers) met de methode TRA_obswell+. De volgorde van de meetpunten is hetzelfde gehouden als in Figuur 15.

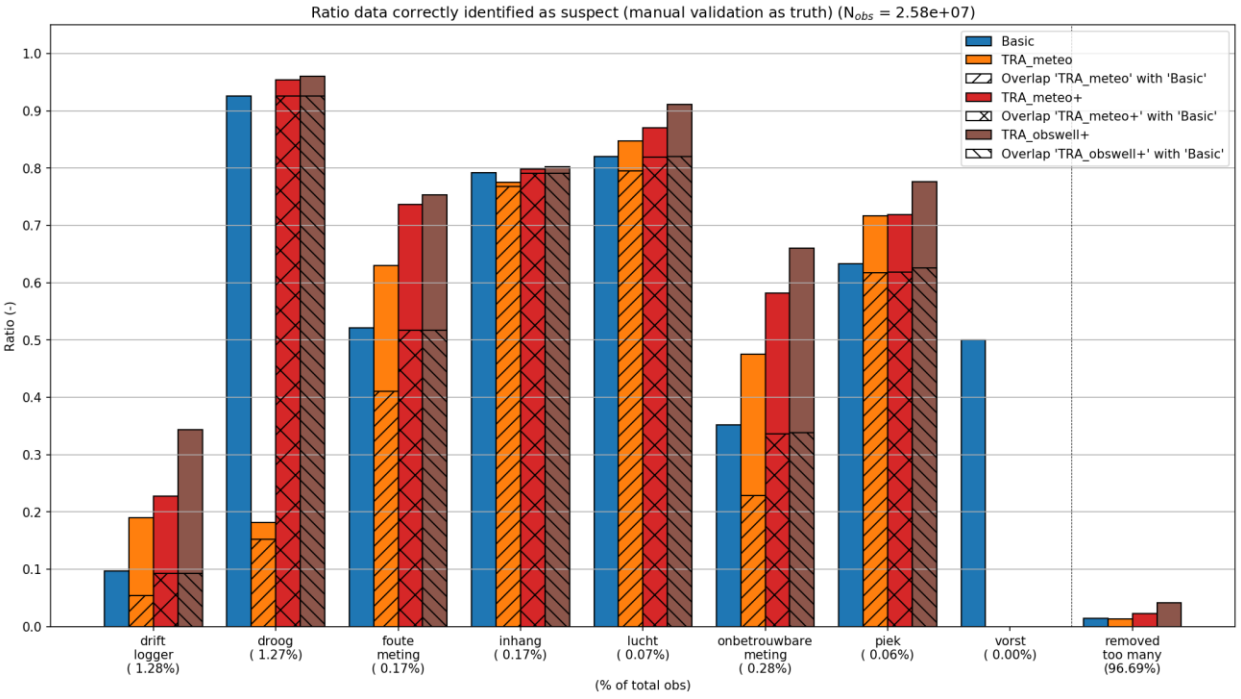
5.3 Waterschap Aa en Maas

De resultaten van de foutherkenningsalgoritmes toegepast op de twee datasets van Aa en Maas zijn hieronder opgenomen. De eerste paragraaf beschouwd de data afkomstig van divers, de tweede de telemetrie data.

5.3.1 Divers

In Figuur 17 is het resultaat van het toepassen van de automatische foutherkenningsalgoritmes op de divers dataset van Aa en Maas weergegeven. Algemeen kunnen de volgende conclusies worden getrokken op basis van de figuur:

- De TRA_meteo methode werkt niet goed voor het detecteren van droogval ten opzichte van de BASIS methode.
- De TRA_meteo+ en TRA_obswell+ methodes scoren voor alle categorieën beter dan de BASIS methode. De verbetering is echter beperkt voor de categorieën “droogval”, “lucht” en “inhang”. Dit zijn alle drie vergelijkbare typen fouten die relatief makkelijk te detecteren zijn mits de inhangdiepte van de loggers bekend is. Voor de categorieën “drift”, “foute meting”, “onbetrouwbare meting”, en “piek” zijn de verbeteringen groter. Dit zijn fouten die vaak lastiger zijn om op te sporen.
- De methodes op basis van tijdreeksanalyse modellen scoren slechter (hoger) in de categorie “removed too many” (teveel verwijderd). Dit zijn de valse positieven (volgens de foutherkenningsmethode verdacht, maar volgens de handmatige validatie niet). De verbetering in het vinden van verdachte metingen gaat dus gepaard met een toename in het aantal valse positieven.



Figuur 17 Aandeel fouten uitgesplitst per categorie dat gesignaleerd wordt door verschillende automatische foutherkenningsalgoritmes voor de dataset Aa en Maas (divers). Het gearceerde deel geeft de overlap aan met de BASIS methode.

Hieronder is per categorie fout aangegeven welke methodes aanbevolen worden voor het signaleren van dat type fout:

- **Drift logger:** De methodes TRA_meteo+ en TRA_obswell+ (hierna de TRA+ methodes) zijn in staat om respectievelijk 22% en 35% van dit type foute metingen te signaleren. Dit zijn geen hoge scores, maar wel aanzienlijk beter dan de 10% die de BASIS methode scoort. Drift is een van de lastigste type fouten om te detecteren omdat de meetfout (bij beperkte drift) in de eerste periode klein is. Voor deze categorie fout zijn de foutherkenningsmethodes op basis van tijdreeksanalyse van toegevoegde waarde. Met name de modellen op basis van andere peilbuizen lijken geschikt om drift mee op te sporen.
- **Droog:** De BASIS methode is in staat om ca. 93% van de droge metingen te signaleren. De verbetering in de TRA+ methodes is beperkt. Voor dit type fout is de droogval regel uit de BASIS methode voldoende goed en biedt tijdreeksanalyse geen meerwaarde.
- **Foute meting:** De TRA+ methodes scoren hier ca. 75% t.o.v. de 53% in de BASIS methode. Dit is een aanzienlijke verbetering en dus wordt geconcludeerd dat de TRA+ methodes geschikt zijn voor het detecteren van dit type fout.
- **Inhang:** De verschillen tussen de methodes zijn minimaal, allemaal signaleren ze ca. 80% van de metingen in deze categorie. Gezien de eenvoud van de BASIS methode zou voor dit type fout de BASIS methode volstaan.
- **Lucht:** De TRA_meteo+ scoort ca. 5% beter dan de BASIS methode die ca. 83% van deze categorie correct signaleert. De TRA_obswell+ methode scoort ca. 9% beter dan de BASIS methode. Dit is een beperkte verbetering. De TRA+ methodes scoren dus iets hoger, maar het verschil t.o.v. de BASIS methode is klein.
- **Onbetrouwbare meting:** De BASIS methode scoort hier ca. 35%, de TRA_meteo+ methode 58% en de TRA_obswell+ methode 66%. De TRA+ methodes bieden een meerwaarde voor het signaleren van dit type fout.
- **Piek:** De BASIS methode scoort ca. 63% voor dit type metingen, de TRA_meteo methodes ca. 72% en de TRA_obswell+ methode ca. 77%. Ook hier zijn de methodes op basis van tijdreeksanalyse van meerwaarde voor het opsporen van dit type fout.
- **Vorst:** Er zijn in de dataset 2 metingen (uit 25.8 miljoen) die dit label hebben gekregen. Gezien het beperkte aantal metingen is het niet mogelijk een uitspraak te doen over de geschiktheid van de methodes om fouten van deze categorie op te sporen.

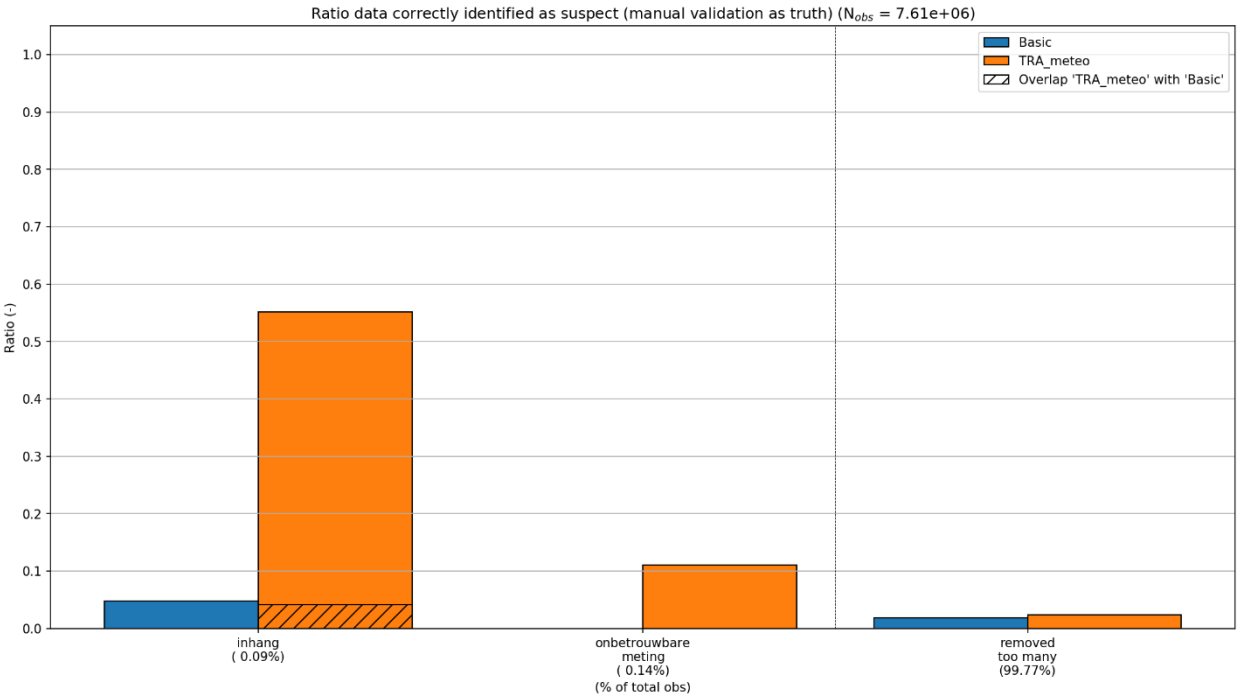
De TRA_meteo+ methode is in staat 63% van de fouten te signaleren waarbij ca. 2% van de goede metingen ook als verdacht wordt gesignaleerd. Dat betekent dat iets minder dan de helft van de gesignaleerde metingen goed is gesignaleerd, de rest zijn dus valse positieven. De TRA_obswell+ methode signaleert ca. 68% van de fouten, maar geeft ook ca. 4% valse positieven. Daarmee is het aandeel goed gesignaleerd ca. 1/3 van het totale aantal metingen dat wordt gesignaleerd door deze methode.

5.3.2 Telemetrie

Figuur 18 toont de resultaten uitgesplitst per type fout voor de twee automatische foutherkenningsalgoritmes toegepast op de telemetrie dataset van Aa en Maas. In de telemetrie data van Aa en Maas is het aantal geïdentificeerde fouten in de handmatige validatie beperkt, waardoor alleen de categorieën “inhang” en “onbetrouwbare meting” bestaan. Vanwege het gebrek aan informatie over inhangdieptes zijn alleen de BASIS en TRA_meteo methodes toegepast.

De kwaliteit van de handmatige validatie is zeer waarschijnlijk ontoereikend om conclusies te trekken over de prestaties van de automatische foutherkenningsmethodes. De TRA_meteo methode scoort beter dan de BASIS methode, maar van het totale aantal gesignaleerde metingen is maar 3% daadwerkelijk fout. Met andere woorden, de fouten in deze dataset zijn lastig te signaleren met behulp van automatische

methodes. Maar gezien het beperkte aantal peilbuizen dat handmatig gevalideerd is, is de verwachting dat een deel van de valse positieven eigenlijk wel foute metingen betreft. De automatische methodes (met name de BASIS methode) moeten mogelijk nog verder geoptimaliseerd worden, maar om ze goed te scoren moet eerst de handmatige validatie verbeterd worden.

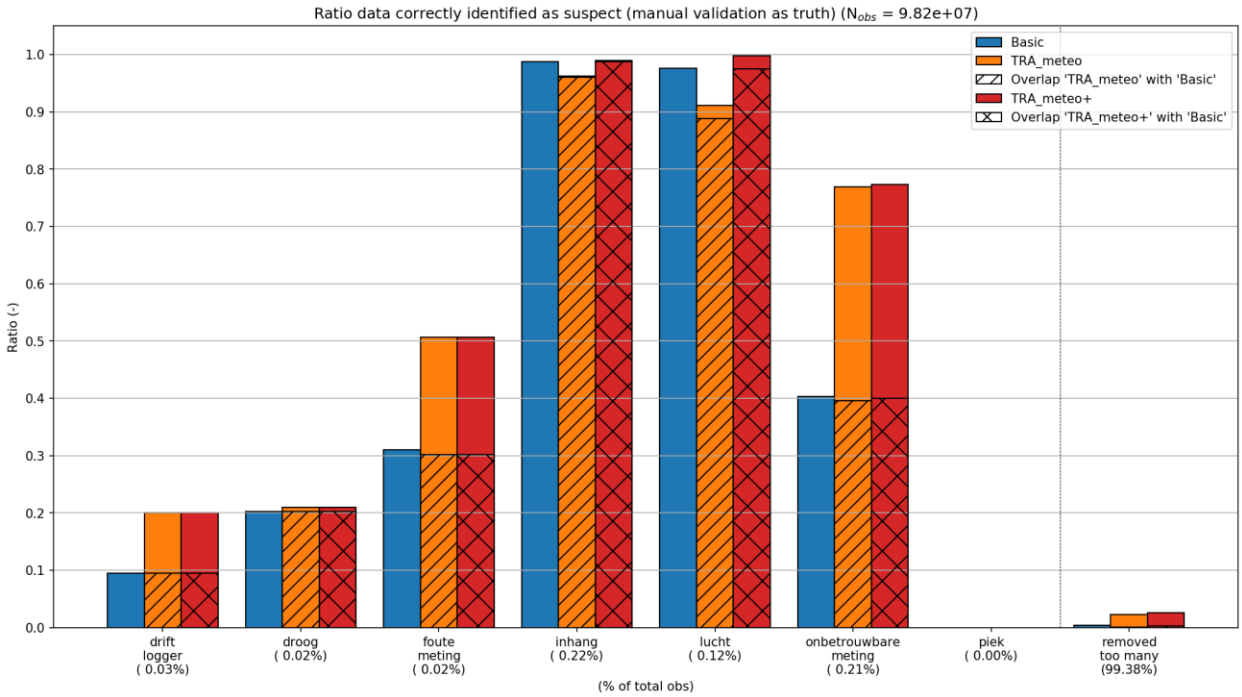


Figuur 18 Aandeel fouten uitgesplitst per categorie dat gesignaleerd wordt door verschillende automatische foutherkenningsalgoritmes voor de dataset van Aa en Maas (telemetrie). Het gearceerde deel geeft de overlap aan met de BASIS methode.

5.4 Brabant Water

Figuur 19 toont de resultaten uitgesplitst per categorie voor de automatische foutherkenningsalgoritmes toegepast op de dataset van Brabant Water. Over het algemeen kunnen de volgende conclusies worden getrokken:

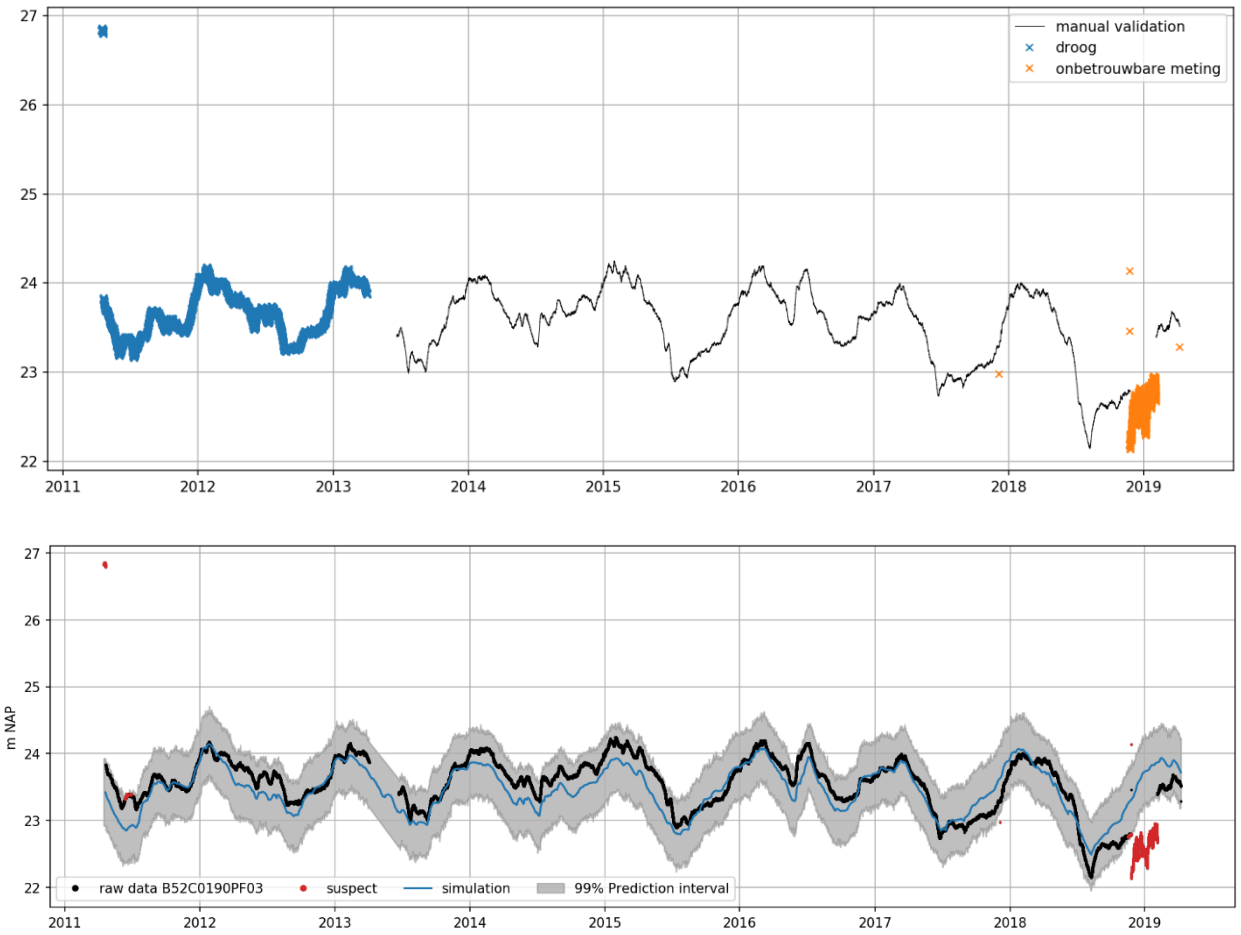
- Voor de categorieën “droog”, “inhang” en “lucht” is het verschil tussen de BASIS methode en de methodes op basis van tijdreeksanalyse klein. Opvallend is dat in tegenstelling tot resultaten bij de andere datasets, de automatische foutherkenningsmethodes niet goed in staat zijn metingen uit de categorie “droog” te signaleren.
- Voor de categorieën “drift”, “foute meting” en “onbetrouwbare meting” scoren de methodes op basis van tijdreeksanalyse beter dan de BASIS methode.
- Het verschil tussen de TRA_meteo en TRA_meteo+ methode is klein ten opzichte van de verschillen in andere datasets. Dat betekent dat de droogval regel uit de BASIS methode maar beperkt van waarde is voor het signaleren van verdachte metingen.
- Het aantal valse positieven is relatief laag bij de BASIS methode (ten opzichte van de overige datasets): ca. 0.4%. Ten opzichte van de BASIS methode is het aantal valse positieven gesignaleerd door de TRA methodes aanzienlijk hoger (ca. 2%). Het aandeel goed gesignaleerde metingen neemt dus toe met de TRA methodes, maar het aandeel valse positieven neemt ook toe.



Figuur 19 Aandeel fouten uitgesplitst per categorie dat gesignaleerd wordt door verschillende automatische foutherkenningsalgoritmes voor de dataset van Brabant Water. Het gearceerde deel geeft de overlap aan met de BASIS methode.

Hieronder is de prestatie van de automatische methodes per categorie nader toegelicht:

- **Drift logger:** De BASIS methode scoort ca. 10% en de TRA_meteo methodes ca. 20%. De TRA_meteo methodes scoren niet hoog, maar zijn wel beter in staat drift fouten te signaleren dan de BASIS methode. Van de fouten die gelabeld zijn als drift zit 65% in 2 tijdreeksen.
- **Droog:** Alle drie de automatische methodes scoren ca. 20%. Het is opvallend dat de BASIS methode in dit geval zo laag scoort bij het detecteren van droogval, gezien de hoge scores in andere datasets waarbij dezelfde informatie over inhangdieptes e.d. bekend is. Dit is nader onderzocht en het blijkt dat ca. 70% van de droogval metingen in één reeks aanwezig zijn (zie ook Figuur 4). In Figuur 20 is deze tijdreeks te zien waarbij de opmerking droog onterecht is gebruikt om metingen af te keuren. Het resultaat van TRA_meteo is ook weergegeven en geeft geen aanleiding om deze metingen als verdacht te bestempelen.



Figuur 20 Opmerkingen in handmatig gevalideerde tijdreeks B52C0190PF03 (boven). De opmerking droog is hier onterecht gebruikt voor het afkeuren van de metingen. Resultaat van de TRA_meteo methode (onder).

- **Foute meting:** De TRA_meteo methodes scoren ca. 20% beter dan de BASIS methode en signaleren daarmee ca. 50% van de metingen in deze categorie. De TRA_meteo methode is van de automatische methodes dus het meest geschikt om dit type fout te signaleren.
- **Inhang:** Alle drie de methodes scoren ongeveer gelijk en signaleren 95%+ van dit type fout. De BASIS methode is voldoende om de meeste va dit type fout te identificeren.
- **Lucht:** De BASIS methode (97%) scoort beter dan de TRA_meteo methode (91%). De TRA_meteo+ methode signaleert 100% van de metingen in deze categorie. De BASIS en TRA_meteo+ methodes zijn dus zeer geschikt voor het signaleren van dit type fout.
- **Onbetrouwbare meting:** De TRA_meteo methodes geven bijna een verdubbeling van het aantal gesignaleerde fouten van dit type (ca. 78% t.o.v. 40%). De TRA_meteo+ methode biedt weinig meerwaarde t.o.v. de TRA_meteo methode.
- **Piek:** Deze categorie komt maar 1x voor in de hele dataset, dus kan er geen uitspraak worden gedaan over welke methode het meest geschikt is voor het signaleren van dit type fout.

De BASIS methode vindt 70% van de fouten die in de handmatige validatie zijn gesignaleerd. Van het totale aantal gesignaleerde metingen met deze methode is iets meer dan de helft goed gesignaleerd. De BASIS methode is dus al zeer geschikt voor het automatisch herkennen van fouten in de data.

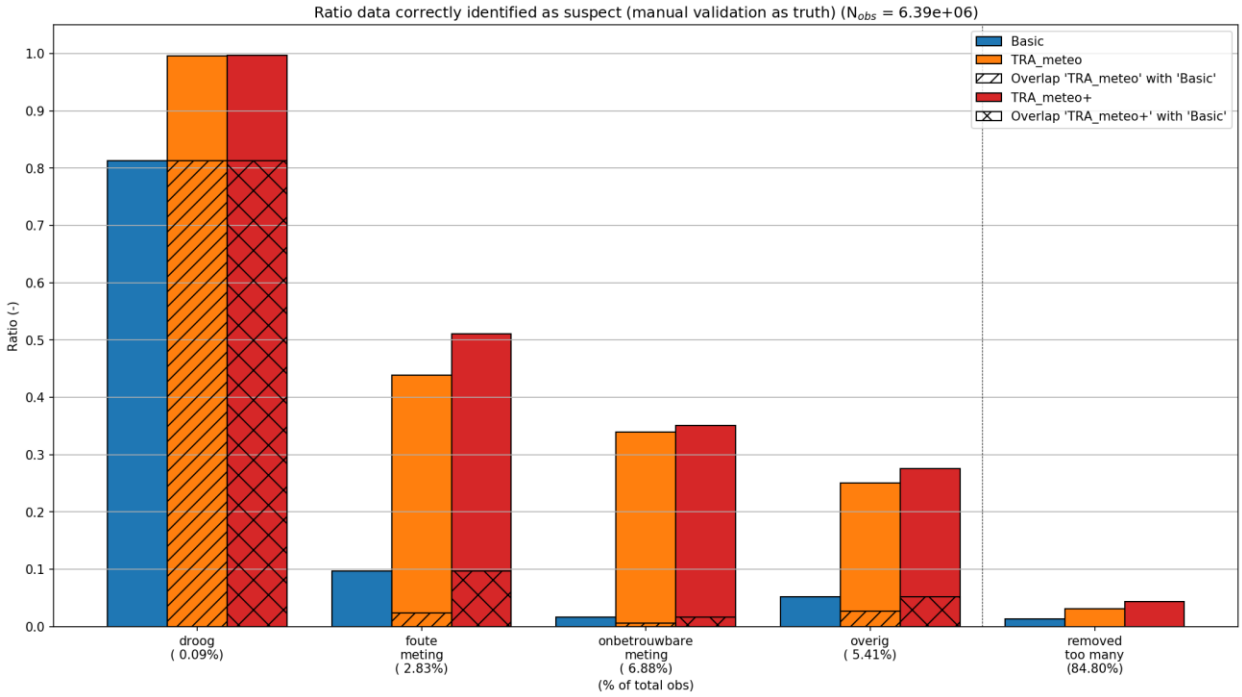
De TRA_meteo (ca. 81%) en TRA_meteo+ (ca. 83%) methodes scoren vergelijkbaar op het signaleren van meetfouten. De meerwaarde van de TRA methodes is dat er ca. 11-13% meer foute metingen worden

gesignaleerd ten opzichte van de BASIS methode (ca. 70%). Het aandeel valse positieven neemt wel met ca. 2% toe bij de TRA methodes ten opzichte van 0.4% valse positieven in de BASIS methode.

5.5 Waterschap de Dommel

De resultaten van de automatische foutherkenningsmethodes toegepast op de dataset van De Dommel zijn weergegeven in Figuur 21. In de dataset van De Dommel zijn relatief veel metingen als verdacht gesignaleerd in de handmatige validatie. Dit wordt grotendeels veroorzaakt vanwege de regel die eist dat de logger meting en de handmeting niet te ver van elkaar mogen liggen. Als dat wel het geval is wordt alle data tot de volgende handmeting die wel aan de eisen voldoet als verdacht gemarkeerd. De opmerkingen bij de gevalideerde metingen zijn vertaald naar de ArtDiver categorieën, maar niet alle categorieën worden onderscheiden. De categorieën 'drift' en 'pieken' worden bijvoorbeeld niet onderscheiden in de opmerkingen.

In de figuur is zichtbaar dat in alle gevallen de foutherkenningsmethodes op basis van tijdreeksanalyse beter scoren dan de BASIS methode. Voor de categorie "droog" werken de met name de TRA_meteo methodes zeer goed (bijna 100% van de fouten worden gesignaleerd). In de overige categorieën wordt maximaal ca. 50% van de fouten correct gesignaleerd. Dit is mogelijk omdat er in de handmatige validatie veel metingen als verdacht zijn gemarkeerd die nog best plausibel zijn al liggen ze verder dan 5 cm van de handmeting. Het aantal metingen dat door de automatische methodes wordt onterecht wordt gesignaleerd (gegeven de handmatige validatie als waarheid) ligt hoger voor de methodes op basis van de tijdreeksanalyse. De Dommel heeft aangegeven dat er de kwaliteit van de handmatige validatie te wensen over laat, en dus moet er voorzichtig worden omgesprongen met harde conclusies op basis van deze figuur.



Figuur 21 Aandeel fouten uitgesplitst per categorie dat gesignaleerd wordt door verschillende automatische foutherkenningsalgoritmes voor de dataset van De Dommel. Het gearceerde deel geeft aan welk aandeel ook door de BASIS methode wordt geïdentificeerd.

Hieronder is per categorie beschreven hoe goed de verschillende automatische methodes in staat zijn om bepaalde typen fouten te signaleren.

- **Droog:** De BASIS methode is in staat om ca. 80% van de droge metingen te signaleren. De TRA_meteo methodes signaleren ca. 100%. Het is mogelijk dat de BASIS regel hier lager scoort omdat de ruwe druk tijdreeksen (de P1 en P2 reeksen) vaak korter waren dan de gevalideerde tijdreeks. Er zijn maar 4 tijdreeksen die samen vrijwel alle fouten gelabeld als droogval bevatten. De TRA_meteo methodes zijn goed in staat om in deze 4 tijdreeksen deze fouten te signaleren.
- **Foute meting:** De BASIS methode signaleert ca. 10% van de metingen in deze categorie. De TRA_meteo+ methode signaleert ca. 50% van deze metingen. De methodes op basis van tijdreeksmodellen hebben dus een meerwaarde t.o.v. de BASIS methode.
- **Onbetrouwbare meting:** De BASIS methode signaleert bijna geen metingen in deze categorie, terwijl de TRA_meteo methodes ca. 35% scoren.
- **Overig:** De Basis methode signaleert ca. 5% van de metingen, terwijl de TRA_meteo methodes ca. 25% weten te signaleren. Dat is een verbetering maar er worden veel metingen in deze categorie niet gesignaleerd.

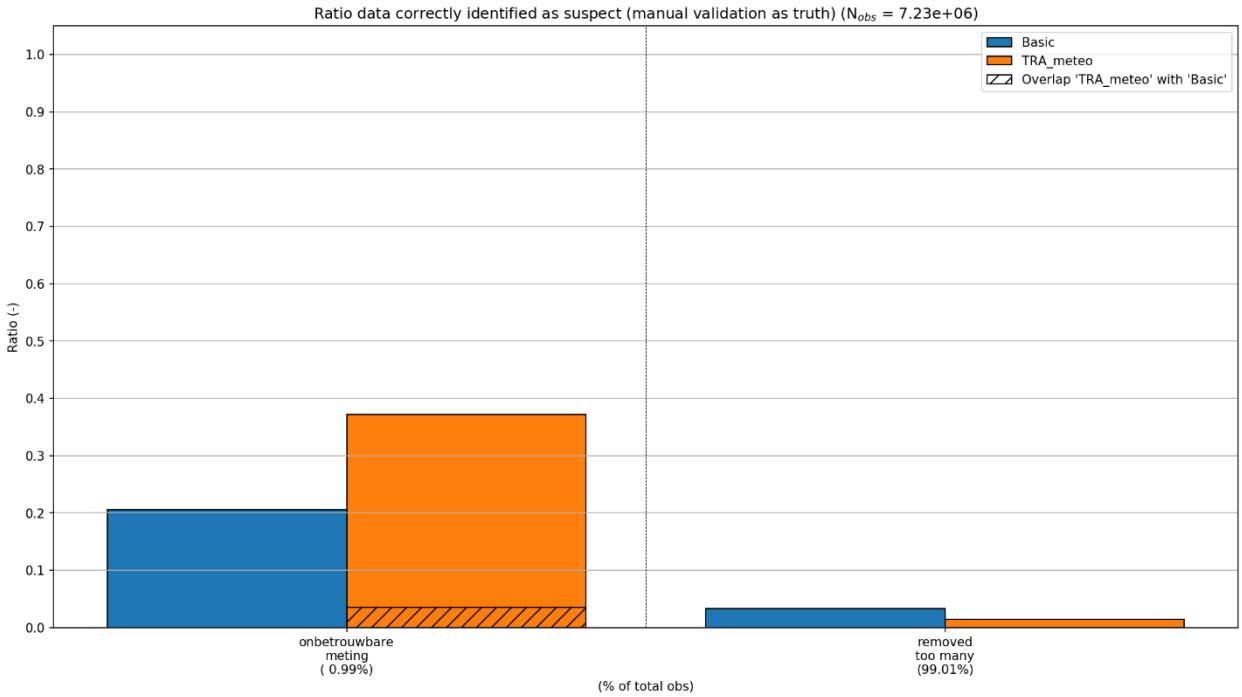
De BASIS methode vind slechts 5% van de metingen die in de handmatige validatie als verdacht zijn gemarkeerd. De methodes op basis van tijdreeksanalyse scoren aanzienlijk beter en vinden ca. 33% van de verdachte metingen terug. Wel wordt maar een beperkt deel van de metingen die in de handmatige validatie als verdacht zijn gemarkeerd teruggevonden. Zoals eerder aangegeven kan dit te maken hebben met de kwaliteit van de handmatige validatie. Bij de TRA_meteo methodes is de kans dat een meting daadwerkelijk verdacht is gegeven dat deze door de automatische methode wordt gesignaleerd hoog, met een kans van ca. 0.6. Met andere woorden, 60% kans dat een meting daadwerkelijk fout is, gegeven dat de automatische methode de meting als verdacht heeft gemarkeerd.

5.6 Waterschap Rijn en IJssel

In Figuur *22* zijn de resultaten van de BASIS en TRA_meteo methodes op de dataset van Rijn en IJssel weergegeven. De dataset van Rijn en IJssel bevatte helaas geen opmerkingen, waardoor er geen onderverdeling gemaakt kon worden in verschillende categorieën. Deze opmerkingen zijn wel beschikbaar maar kunnen helaas niet in bulk geëxporteerd worden.

Uit de figuur kan opgemaakt worden dat de methode op basis van tijdreeksanalyse beter presteert en minder valse positieven oplevert, dus eigenlijk in alle opzichten beter werkt dan de BASIS methode. Daarbij moet wel aangegeven worden dat de BASIS methode maar beperkt geoptimaliseerd is, en dat na optimalisatie mogelijk een betere score haalbaar is. Zie ook de aanbevelingen ten aanzien van het optimaliseren van methodes in hoofdstuk 8.

In totaal vindt de TRA_meteo (37%) methode ca. twee keer zoveel verdachte metingen als de BASIS methode (20%) terwijl het aandeel valse positieven ongeveer half zo groot is. Van het totale aantal gesignaleerde metingen met TRA_meteo is maar 20% daadwerkelijk verdacht volgens de handmatige validatie. De TRA_meteo methode presteert dus beter, maar het aantal gevonden fouten is wel beperkt in vergelijking met de resultaten van een aantal van de andere datasets.

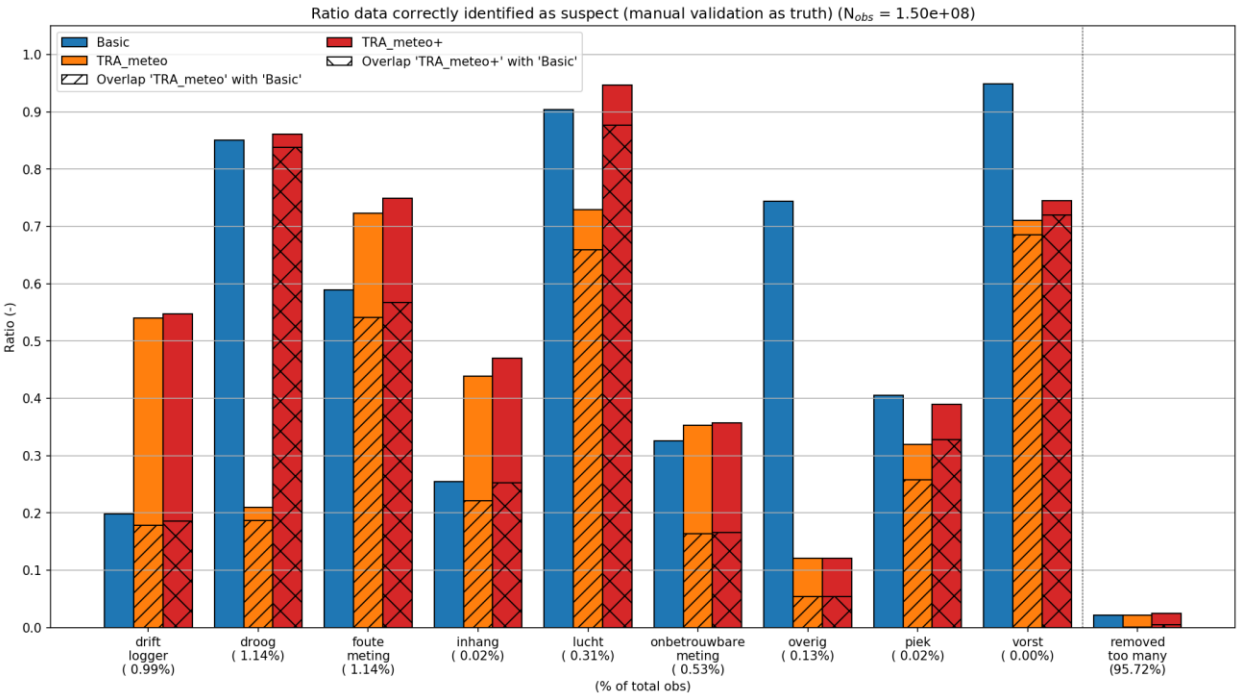


Figuur 22 Aandeel fouten dat gesignaleerd wordt door verschillende automatische foutherkenningsalgoritmes voor de dataset van Rijn en IJssel. Het gearceerde deel geeft aan welk aandeel ook door de BASIS methode wordt geïdentificeerd.

5.7 Vitens

In Figuur 23 zijn de resultaten voor de automatische methodes toegepast op de dataset van Vitens weergegeven. Uit de figuur kunnen de volgende algemene conclusies worden getrokken:

- De BASIS methode scoort goed op de volgende categorieën: “droog”, “lucht”, “overig” en “vorst”. Met name in de categorieën “overig” en “vorst” zorgt een van de regels ervoor dat de score significant hoger ligt dan de van de TRA_meteo methodes.
- De TRA_meteo methodes scoren aanzienlijk beter in de categorieën “drift”, “foute meting” en “inhang”.
- In tegenstelling tot andere resultaten (Aa en Maas (divers) en Brabant Water), is de BASIS methode in sommige gevallen beter in staat fouten te herkennen dan de TRA_meteo methodes. Dit betekent dat een van de regels in de BASIS methode goed werkt voor het herkennen van bepaalde typen fouten. Het opnemen van deze regel in een gecombineerde methode zou het resultaat verder kunnen verbeteren.
- Het aantal valse positieven van alle methodes is nagenoeg gelijk.

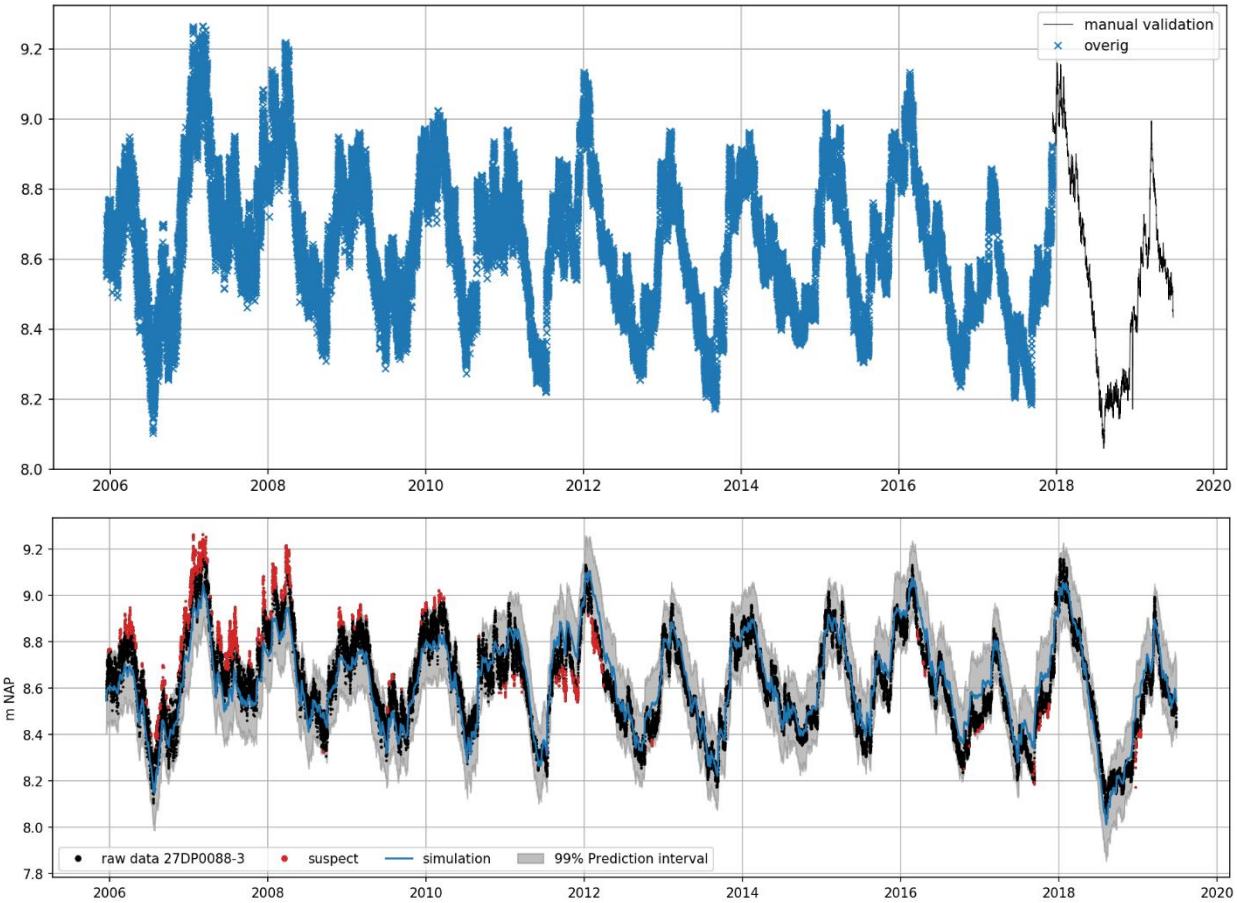


Figuur 23 Aandeel fouten uitgesplitst per categorie dat gesignaleerd wordt door verschillende automatische foutherkenningsalgoritmes voor de dataset van Vitens. Het gearceerde deel geeft aan welk aandeel ook door de BASIS methode wordt geïdentificeerd.

Hieronder zijn de resultaten per categorie nader omschreven:

- **Drift logger:** De TRA_meteo methodes scoren hier ca. 55% terwijl de BASIS methode op 20% blijft hangen. Voor drift herkenning zijn de tijdreeksanalyse methodes van toegevoegde waarde.
- **Droog:** De BASIS methode voor het detecteren van droogval werkt zeer goed en is in staat 85% van dit type foute meting te signaleren. De TRA_meteo methodes voegen hier weinig aan toe.
- **Foute meting:** De TRA_meteo methodes scoren minimaal 13% beter dan de BASIS methode. De TRA_meteo+ methode vindt ca. 75% van de foute metingen terug.
- **Inhang:** De tijdreeksmethodes werken hier aanzienlijk beter dan de BASIS methode. De BASIS methode signaleert ca. 25% van de metingen in deze categorie terwijl de TRA_meteo+ methode bijna 50% signaleert.
- **Lucht:** De BASIS vindt ca. 90% van dit type foute meting terug. Het toevoegen van foutherkenning op basis van een tijdreeksmodel levert 5% meer gesignaleerde fouten op.
- **Onbetrouwbare meting:** alle drie de methodes scoren vergelijkbaar en kunnen ca. 35% van dit type fout signaleren. Wel is de overlap tussen de methodes maar ca 15%. Het combineren van de BASIS en TRA_meteo methodes zou dus theoretisch een methode opleveren die 50% van dit type fout terug vindt.
- **Overig:** De BASIS methode weet ca. 75% van alle metingen in deze categorie te vinden. Dit komt omdat de 'hardmax'-regel een groot deel van dit type fout weet te identificeren. Er zijn een paar reeksen waarin een groot aandeel van het totale aantal fouten met het label "overig" in voorkomt. In deze meetpunten ligt de bovenkant buis (grotendeels) onder de gemeten niveaus, waardoor de metingen als verdacht worden gezien. In Figuur 24 is een voorbeeld hiervan opgenomen voor peilbuis 27DP0088-3. Hierin is zichtbaar dat een groot deel van de tijdreeks verdacht is bevonden door de handmatige validatie maar dat met een tijdreeksmodel de metingen best goed

gesimuleerd kunnen worden. Dit roept de vraag op of deze metingen wel fout zijn zoals de handmatige validatie beweert.



Figuur 24 Opmerkingen in tijdreeks 27DP0088-3 (boven). Toepassing foutherkenning TRA_meteo op peilbuis (onder).

- **Piek:** De BASIS en TRA_meteo+ methodes scoren vergelijkbaar en vinden ca. 40% van de metingen in deze categorie. De BASIS methode scoort net iets hoger en heeft dus de voorkeur boven de TRA_meteo methodes.
- **Vorst:** In deze categorie vallen een beperkt aantal metingen (ca. 6000). De BASIS methode signaleert 95% van deze fouten, terwijl de TRA_meteo methodes rond de 70% scoren. Ca. 90% van alle fouten gelabeld als vorst zitten in één tijdreeks. De 'hardmax'-regel is blijkbaar in staat deze fout te identificeren.

Met de BASIS methode worden ca. 56% van de verdachte metingen volgens de handmatige validatie gevonden. Het aantal valse positieven is ca. 2% waardoor iets meer dan de helft van de gesignaleerde metingen met de BASIS methode daadwerkelijk ook foute metingen zijn. De TRA_meteo+ methode vind ca 68% van alle foute metingen terug, en het aantal valse positieven neemt iets toe naar ca. 2.5%. Toch is de kans dat een meting als deze gesignaleerd wordt ook daadwerkelijk fout is vergelijkbaar met de BASIS methode. De kans dat een meting daadwerkelijk verdacht is als deze gesignaleerd wordt door TRA_meteo+ is ca. 54%.

6 Discussie

In dit hoofdstuk worden de resultaten nader geanalyseerd. De impact van bepaalde aannames op de resultaten en onzekerheden die invloed hebben op de conclusies worden besproken.

6.1 De handmatige validatie als "waarheid"?

Om een vergelijking te kunnen tussen de automatische foutherkenningsmethodes is het noodzakelijk om een waarheid te definiëren. Er moet een objectieve maat zijn om te kunnen zeggen dat een methode beter in staat is om foute metingen te herkennen dan een andere methode. In deze studie is dat de handmatige validatie. Een van de bronnen van onzekerheid in de uitkomsten van deze studie is de aanname dat de handmatige validatie de waarheid vertegenwoordigt. We weten dat de handmatige validatie onvolledig is en fouten bevat. Wat is de impact daarvan op de resultaten van dit onderzoek?

Hieronder staan de drie typen fouten in de handmatige validatie en de impact daarvan op het scoren van de automatische foutherkenning:

- **"Niet gevalideerd" –** Foute metingen die niet geïdentificeerd zijn en dus in de gevalideerde dataset zijn blijven zitten. Indien deze fouten kunnen door de automatische methodes terecht gesignaleerd worden, worden ze gerekend als "removed too many" of "onterecht gesignaleerd".
- **"Teveel gevalideerd" –** Goede metingen die onterecht als fout geïdentificeerd zijn en dus als verdacht worden gezien maar dat eigenlijk niet zijn. Als deze metingen door de automatische methodes niet gesignaleerd worden, krijgen ze daardoor het label "wrongly kept" of "onterecht behouden".
- **"Verkeerd geïdentificeerd" –** Foute metingen die verkeerd zijn gelabeld zorgen ervoor dat fouten verkeerd gecategoriseerd worden. Dit type fout heeft minder impact op de resultaten uit deze studie, maar als er gekeken wordt naar specifieke regels om bepaalde fouten te herkennen, maken dit soort fouten het moeilijker om een regel te beoordelen.

De scores berekend in het vorige hoofdstuk geven aan in hoeverre de handmatige validatie benaderd kan worden middels automatische foutherkenningsmethodes. Hoe dichter de handmatige validatie bij de perfecte validatie ligt, hoe meer de score iets zegt over hoe goed de methode scoort in absolute zin. Het is lastig te beoordelen hoe ver een handmatige validatie van de perfecte validatie af ligt. De scores moeten dus vooral gezien worden als een middel om foutherkenningsmethodes mee te vergelijken. Ook zeggen ze iets over de mate waarin de huidige handmatige validatie geautomatiseerd kan worden.

De impact van deze onzekerheid is dus vooral dat er geen absolute conclusies getrokken kunnen worden over hoeveel foute metingen er goed gesignaleerd zijn door de automatische methodes. Een methode die 80% goed gesignaleerde fouten scoort in een relatief goed gevalideerde dataset kan net zo veelbelovend zijn als een methode die 30% scoort in een relatief slecht gevalideerde dataset. In het eerste geval kan met meer zekerheid gesteld worden dat de methode van waarde is, maar die zekerheid is lastig te kwantificeren. In het tweede geval is eigenlijk een her-validatie nodig om de methode echt goed te beoordelen.

Ook bij het beoordelen van de geschiktheid van een bepaalde regel of methode voor het identificeren van bepaalde type fouten moet er rekening gehouden worden met de imperfectie in de handmatige validatie. Een perfecte score (bijvoorbeeld 100% droogval vinden) is vaak niet haalbaar, dus moeten we de scores vooral gebruiken om methodes en regels te vergelijken. Bij optimaliseren van regels of methodes moet

hier ook rekening mee gehouden worden. Het nastreven van 100% van de fouten identificeren kan ervoor zorgen dat veel te veel goede metingen worden als verdacht worden gesignaleerd. De categorieën “teveel verwijderd” en “onterecht behouden” zijn interessant om nader te beschouwen, bijvoorbeeld in een her-validatie.

6.2 Wat zit er allemaal in de groep valse positieven?

Valse positieven zijn metingen die door de automatische methodes als verdacht worden gemarkeerd maar in de handmatige validatie niet als verdacht zijn geïdentificeerd. De metingen in deze groep vallen eigenlijk in twee categorieën:

- De automatische methode doet het niet goed (de handmatige validatie heeft gelijk).
- De handmatige validatie is niet goed (de automatische methode heeft dus gelijk).

De eerste categorie is interessant om de automatische methode beter af te stellen. Dit type fout moet namelijk geminimaliseerd worden. Ook kunnen hier metingen tussen zitten die hydrologisch interessante periodes (bijvoorbeeld extreme situaties) vertegenwoordigen, die niet met het huidige tijdreeksmodel gesimuleerd kunnen worden. De tweede categorie is interessant omdat dit metingen zijn die eigenlijk wel verdacht zijn maar in de handmatige validatie niet zijn gevonden. De valse positieven zijn dus een interessante subset metingen om nader te beschouwen.

6.3 Fouten laten zitten of goede metingen weggooien?

Een automatische methode om foute metingen te detecteren wordt beoordeeld door de resultaten van de automatische methode naast die van de handmatige validatie te leggen. Gegeven dat de automatische methode niet perfect is, moet er een keuze worden gemaakt tussen het aantal foute metingen dat juist wordt gesignaleerd en het aantal metingen dat onterecht als verdacht wordt gezien. In de meeste gevallen zal meer fouten detecteren ook betekenen dat er meer valse positieven tussen zitten. Andersom geldt als er minder foute metingen worden gesignaleerd, zullen er minder valse positieven tussen zitten. Deze afweging zal door de gebruikers van deze methodes gemaakt moeten worden: “Is het acceptabel om 1 op de 100 metingen weg te gooien als ik daarmee 9 van de 10 aanwezige fouten automatisch kan signaleren?”.

Een belangrijk aspect daarbij is de meetfrequentie. Bij het meten van stijghoogtes is het waarschijnlijk sneller acceptabel om een meting weg te gooien in een periode waar er per uur gemeten wordt dan in een periode waar er per dag of 2-wekelijks wordt gemeten. Dit kan betekenen dat de eisen t.a.v. een automatische methode dus afhankelijk zijn van de meetfrequentie. In dagelijkse data accepteren we meer foute metingen, omdat we liever niet teveel weggooien. In uur-data zijn we bereid meer weg te gooien om meer fouten eruit te filteren.

6.4 Tijdreeksanalyse voor foutherkenning

In deze paragraaf zijn verschillende aspecten met betrekking tot de toepassing van tijdreeksanalyse geanalyseerd.

Goede versus slechte tijdreeksmodellen

In de resultaten uit het vorige hoofdstuk is geen onderscheid gemaakt tussen goede en slechte tijdreeksmodellen. Dat is omdat het voorspellingsinterval per definitie rekening houdt met de kwaliteit van de fit. Het 99% voorspellingsinterval betekent dat ca. 99% van de metingen binnen die grenzen liggen. Bij een goed model kunnen die grenzen veel strakker gekozen worden omdat het model de metingen goed

volgt. Bij een minder goed model (bijvoorbeeld het gemiddelde van de tijdreeks) moeten die grenzen veel groter gekozen worden om 99% van alle metingen te bevatten. Dus ook met minder goede modellen kunnen nog wel fouten gedetecteerd worden, maar zijn dat alleen de extremere fouten. Echter, is het verbeteren van de tijdreeksmodellen wel aan te bevelen, omdat met een beter model ook subtielere fouten gedetecteerd kunnen worden.

Wanneer werkt tijdreeksanalyse niet?

Voor het toepassen van een foutherkenningsalgoritme op basis van tijdreeksanalyse is voldoende data nodig om een goed tijdreeksmodel af te kunnen leiden. Dat betekent dat bij een nieuw meetpunt tijdreeksanalyse niet toegepast kan worden tot er voldoende data is om een model mee af te leiden.

Voor zeer slechte modellen is het waarschijnlijk net zo nuttig om een bandbreedte te berekenen op basis van de gevalideerde data. Dat kan een bandbreedte ten opzichte van het gemiddelde of een voortschrijdend gemiddelde zijn. De methode op basis van een slecht tijdreeksmodel werkt vergelijkbaar met zo'n bandbreedte, maar is eigenlijk onnodige moeite voor een regel die ook simpeler kan.

Wat is het beste tijdreeksmodel?

Wat is het beste tijdreeksmodel voor het berekenen van een bandbreedte om verdachte metingen te signaleren in de toekomst? Enerzijds is het fijn om een lange tijdreeks te hebben zodat allerlei klimatologische omstandigheden zijn gemeten (droge en natte zomers, etc.). Het tijdreeksmodel dat daarop gefit wordt probeert dan rekening te houden met al die situaties. Anderzijds is het mogelijk dat in de loop der jaren omstandigheden veranderen die het gedrag van de grondwaterstand beïnvloeden. Dat kan niet gevangen worden in het model, omdat het model ervan uitgaat dat de relatie tussen de verklarende variabelen en de grondwaterstand constant blijft. Om deze reden kan het dus beter zijn om een minder lange periode te beschouwen bij het maken van het model. Dat model is dan mogelijk beter in staat om de huidige situatie te simuleren.

Dit is uiteraard een effect van het vereenvoudigen van de werkelijkheid met een imperfect model. Een perfect model zou ook de veranderingen die het beschreven gedrag veroorzaken meenemen. Idealiter gebruiken we een model dat alle situaties goed omschrijft, maar aangezien dat model zeer waarschijnlijk niet bestaat moeten we een model kiezen dat het beste in staat is de huidige situatie te simuleren. Een pragmatische aanpak met een tijdreeksmodel dat niet de hele tijdreeks maar een (voldoende lange) recente periode beschouwd is dus mogelijk een goede oplossing voor foutherkenning. Daarbij zijn er nog wel vragen die onderzocht moeten worden:

- Welke periode moet beschouwd worden door het tijdreeksmodel om de beste simulatie van de huidige situatie te krijgen?
- Hoe lang moet de periode/tijdreeks minimaal zijn om een voldoende goed model af te leiden?

7 Conclusie

In dit onderzoek zijn verschillende automatische foutherkenningsalgoritmes toegepast op de datasets van vijf verschillende organisaties. Het doel van dit onderzoek is om te onderzoeken in hoeverre meetfouten automatisch gesignaleerd kunnen worden met behulp van tijdreeksanalyse. Om het onderzoek mogelijk te maken is een framework ontwikkeld om foutherkenningsalgoritmes op een generieke manier op te stellen. In Python zijn tools ontwikkeld om deze algoritmes op datasets toe te passen en om tijdreeksen met elkaar te vergelijken. Met behulp van dit framework en de ontwikkelde tools zijn een aantal automatische foutherkenningsalgoritmes opgesteld. De automatische foutherkenningsmethodes zijn toegepast op de datasets van de vijf deelnemende organisaties en de uitkomsten zijn vergeleken met resultaten van de handmatige validatie.

Hoe bereken je een betrouwbaarheidsinterval voor een simulatie met een tijdreeksmodel?

Bij het optimaliseren van een tijdreeksmodel (met een ruismodel) wordt ook de zekerheid waarmee een parameter kan worden vastgesteld berekend. De kansverdeling van de parameterwaarde wordt als normaal verdeeld aangenomen en de standaardafwijking zegt dus iets over de spreiding van de gevonden waardes. Door heel vaak uit deze parameterverdelingen parametersets te trekken en het model opnieuw te simuleren kan een bandbreedte worden berekend als gevolg van de parameteronzekerheid. Als daarbij de onzekerheid van het model wordt opgeteld resulteert dat in het voorspellingsinterval. De onzekerheid van het model wordt bepaald door de verdeling van de residuen (de vergelijking van de metingen met de simulatie). Deze voorspellingsbandbreedte heeft een parameter die bepaald hoeveel metingen binnen dat interval moeten liggen. De 99%-voorspellingsbandbreedte betekent dat ca. 99% van de metingen tussen die bovengrens en ondergrens liggen.

Hoe beïnvloedt onzekerheid van een tijdreeksmodel (de model fit en de parameter onzekerheid) de bruikbaarheid ervan voor automatische validatie?

De bandbreedte houdt per definitie rekening met de fit van het tijdreeksmodel. Een slecht model heeft een veel grotere bandbreedte dan een goed model. In beide gevallen wordt geëist dat 99% van de metingen binnen die bandbreedte moeten liggen. Een slecht model (b.v. het gemiddelde van de tijdreeks) moet dan een hele brede bandbreedte hebben om ervoor te zorgen dat aan deze eis wordt voldaan. Een goed model volgt de gemeten grondwaterstand veel beter waardoor de bandbreedte veel smaller kan zijn om alsnog aan die 99%-eis te voldoen.

Er is in deze studie geen onderscheid gemaakt in de kwaliteit van de tijdreeksmodellen die gebruikt zijn voor de foutherkenning. Er zitten dus hele goede modellen tussen, maar ook hele slechte modellen. De bandbreedte houdt rekening met de fit van het model mee bij het bepalen of een meting verdacht is of niet. Een bandbreedte op basis van een slecht model is dus minder kritisch met het signaleren van metingen. Subtiele fouten worden daarmee niet ontdekt, maar grove fouten nog wel.

Welke algoritmes of methodes kunnen worden afgeleid om meetfouten in data te identificeren?

De eerste van de automatische foutherkenningsmethodes is de BASIS methode, die dient als een maatstaf om de andere automatische methodes mee te vergelijken. Deze methode is gebaseerd op relatief eenvoudige statistische regels en geeft dus een indicatie in hoeverre de handmatige validatie nagebootst kan worden met een relatief simpel algoritme. De BASIS methode verschilt per dataset omdat niet altijd dezelfde metadata beschikbaar was. Over het algemeen bestaat de methode in ieder geval uit een regel om droogval te en een regel om spikes te detecteren.

De overige methodes zijn gebaseerd op tijdreeksmodellen waarbij het weer (neerslag en verdamping) of een nabijgelegen peilbuis als verklarende variabele is gebruikt. Met de modellen wordt een bandbreedte berekend, die gebruikt wordt om van ruwe metingen in te schatten of deze plausibel zijn of verdacht. In de meeste gevallen is een tijdreeksmodel gebruikt op basis van neerslag en verdamping (de meteorologie) om de stijghoogte te simuleren. Deze methode heet in dit onderzoek TRA_meteo. In de methode TRA_meteo+ is de tijdreekanalyse samengevoegd met een eenvoudige regel om droogval te detecteren uit de BASIS methode. Voor één dataset zijn ook tijdreeksmodellen gemaakt op basis van het nabijgelegen meetpunt dat het sterkst correleert met de stijghoogte in het beschouwde meetpunt. Deze methode heet in deze studie TRA_obswell+.

Hoe kan ik tijdreeksanalyse gebruiken om automatisch verdachte metingen te signaleren?

De foutherkenningsmethodes TRA_meteo, TRA_meteo+ en TRA_obswell+ zijn gebaseerd op tijdreeksmodellen. Op basis van de onzekerheid van het tijdreeksmodel en de onzekerheid van de parameters wordt een bandbreedte berekend om in te schatten of een meting verdacht is. Deze bandbreedte geeft de grenzen aan waar bijvoorbeeld 99% van de gevalideerde metingen (de metingen waar het model op is gefit) tussen liggen. Een ruwe meting die buiten die grenzen ligt wordt dan gesignaleerd als verdacht.

Hoe verhoudt een methode op basis van tijdreeksanalyse zich tot een methode op basis van eenvoudige statistische regels? Hoe verhouden beide methodes zich tot de handmatige validatie?

De automatische foutherkenningsmethodes zijn toegepast op de datasets van de vijf organisaties en de uitkomsten zijn vergeleken met resultaten van de handmatige validatie. De resultaten geven aan in hoeverre de automatische methodes de handmatige validatie kunnen nabootsen. Hoeveel fouten zijn er gevonden, hoeveel valse positieven zijn er? De berekende scores voor de automatische methodes geven aan in hoeverre de automatische methodes in staat zijn om de handmatige validatie na te bootsen. De score zegt dus niet hoe goed een automatische methode het in absolute zin doet. Dat laatste is alleen mogelijk als de handmatige validatie perfect is (en dat is het niet). Wel kan gesteld worden dat als de handmatige validatie goed is, de score een benadering geeft van de absolute prestatie van de automatische foutherkenningsmethode.

De resultaten verschillen aanzienlijk per organisatie en per automatische methode. Gezien de verschillen in de datasets en de kwaliteit van de handmatige validatie is het vergelijken van uitkomsten tussen datasets niet zinvol. Het vergelijken van de uitkomsten binnen een dataset kan wel, en daaruit blijkt over het algemeen dat de methodes op basis van tijdreeksanalyse (de TRA methodes) meer fouten goed weten te signaleren dan de BASIS methode. Daar staat wel tegenover dat er ook een toename is van de valse positieven. Met andere woorden, het meenemen van een model dat rekening houdt met het weer of nabijgelegen meetpunten levert een meerwaarde op t.o.v. relatief simpele statistische regels, maar betekent ook dat er meer metingen als verdacht worden gezien. De absolute verhouding tussen het aantal juist gesignaleerd en valse positieven blijft bij sommige datasets nagenoeg gelijk tussen de verschillende methodes. In de datasets van Aa en Maas (divers) en Vitens is, ondanks het grotere aantal valse positieven met de TRA_meteo methodes, de toename van het aantal valse positieven ongeveer gelijk aan de toename van het aantal correct gesignaleerde metingen.

De foutherkenning op basis van tijdreeksanalyse op nabijgelegen meetpunten (TRA_obswell+) is maar op één dataset toegepast (Aa en Maas (divers)) maar scoorde daar ook meteen het hoogst van alle methodes op het juist signaleren van foute metingen. Wel was bij die methode ook het aantal valse positieven veruit het hoogst.

De resultaten zijn ook uitgesplitst naar type foute meting voor zover dat mogelijk was binnen de verschillende datasets. Daaruit is op te maken welk aandeel van een bepaald type fout met een bepaalde automatische foutherkenningsmethode goed gesignaleerd kan worden.

De droogval regel uit de BASIS methode lijkt zeer goed te werken voor het detecteren van lucht metingen (droogval, logger uit peilbuis, etc.). De TRA_meteo methode scoort in vergelijking daarmee relatief slecht op het detecteren van droogval. Dit was ook de reden om de gecombineerde methodes TRA_meteo+ en TRA_obswell+ ook te testen. De reden hiervoor is dat de logger in veel gevallen maar net droog is gevallen, waardoor de lucht metingen nog gewoon binnen de toelaatbare bandbreedte van het tijdreeksmodel liggen. In de categorieën “onbetrouwbare meting”, “foute meting” en “drift logger” scoren de methodes op basis van tijdreeksanalyse significant beter dan de BASIS methode, hoewel voor drift de score in absolute zin relatief laag blijft (zie volgende alinea).

Drift van een logger blijkt met de toegepaste methodes een van de lastigste fouten om te herkennen. De methodes op basis van tijdreekanalyse scoren hoger dan de BASIS methode maar signaleren maar een beperkt deel van dit type fout in datasets. Aanvullend onderzoek naar het herkennen van drift in tijdreeksen is opgenomen in Bijlage 5.

8 Aanbevelingen

In dit hoofdstuk worden verschillende aanbevelingen die voortkomen uit het uitgevoerde onderzoek beschreven.

Bulk versus individuele resultaten

In deze studie is voornamelijk gefocust op het toetsen van automatische foutherkenningsalgoritmes op grote datasets en de resultaten zijn ook in bulk beschouwd. Daarmee is het mogelijk om bredere conclusies te trekken over de effectiviteit van de methodes, maar het nadeel is dat er minder aandacht is voor individuele gevallen. Vragen als: “waarom vindt TRA_meteo geen droogval in dataset X, maar wel in dataset Y?” zijn niet volledig beantwoord in deze studie, maar zijn wel essentieel voor het verbeteren van de algoritmes. Voor het implementeren van automatische foutherkenning is het belangrijk om de parameters van de foutenherkenningsregels goed te kiezen en ook goed te onderzoeken in welke gevallen ze niet goed werken en waarom dat zo is.

Verbeteren tijdreeksmodellen

Een logische aanbeveling is om de tijdreeksmodellen te verbeteren. Betere modellen zorgen ervoor dat subtielere fouten gesignaleerd kunnen worden omdat de berekende bandbreedte veel smaller is. Bij het verbeteren van de modellen kan bijvoorbeeld gedacht worden aan het meenemen van belangrijke hydrologische invloeden, of het aanpassen van de optimalisatieperiode.

Methode berekenen voorspellingsbandbreedte

De berekende voorspellingsbandbreedte van de tijdreeksmodellen wordt berekend met een Monte-Carlo trekking uit de berekende parameterverdelingen. Bij deze trekking is het uitgangspunt dat de parameters normaal verdeeld zijn met een bepaalde standaardafwijking. Maar in de optimalisatie zijn de parameters in veel gevallen begrensd, dus de aanname van een standaard normale verdeling klopt niet helemaal. Er wordt aanbevolen om in het berekenen van het voorspellingsinterval rekening te houden met deze grenzen. Hiermee wordt de berekende bandbreedte nauwkeuriger.

Her-validatie

Door de kwaliteit van de handmatige validatie te verbeteren, kan ook beter ingeschat worden hoe goed de automatische foutherkenningsmethodes scoren in absolute zin. De resultaten van de automatische methodes kunnen gebruikt worden om de her-validatie te sturen, de metingen in de categorieën “teveel verwijderd” en “onterecht behouden” zijn waarschijnlijk interessant om te bekijken. Door feedback te geven op het resultaat van de automatische foutherkenning kan het algoritme ook verder bijgestuurd worden zodat het in de toekomst beter wordt. Daarmee wordt een soort “lerend” systeem gecreëerd, dat ook wel eens wordt omschreven met de term “machine learning”. Het leren van het systeem wordt vastgelegd in de parameters die de foutherkenningsregels sturen. Het optimaliseren van deze parameters wordt in de volgende sub-paragraaf omschreven.

Optimaliseren van automatische foutherkenningsmethodes

De regels voor het detecteren van fouten bevatten parameters die de regel sturen. Een voorbeeld: “wat is het maximale het verschil tussen luchtdruk en de ruwe waterdruk waarbinnen we een meting nog als droog beschouwen?”. In dit onderzoek is weinig aandacht besteedt aan het optimaliseren van deze parameters. De parameters zijn in veel gevallen ingesteld door te testen op enkele reeksen, en vervolgens zijn die parameters toegepast voor de gehele dataset. Er valt dus winst te behalen door deze optimalisatie wel uit te voeren. Zo kan er bijvoorbeeld een enkele set parameters afgeleid worden die het beste werkt

voor een hele dataset maar ook kunnen er per individuele reeks een set parameters afgeleid worden, die alleen gelden voor die ene reeks. De optimale parameters worden bepaald door het minimaliseren van het verschil met de handmatige validatie (dus minimaliseren van het aantal valse positieven en onterecht gesignaleerde metingen).

Operationele implementatie automatische foutherkenning

Op basis van de resultaten uit dit onderzoek zou een organisatie kunnen beslissen dat zij een (of een combinatie van) automatische foutherkenningsmethode(s) willen toepassen in hun validatieproces. De automatische foutherkenningsmethode wordt dan als tool ingezet om nieuwe binnenkomende metingen te analyseren.

Eerst moet de tool opgezet en "getraind" worden. Dat betekent o.a. de regels programmeren en de parameters voor die regels optimaliseren. Als tijdreeksanalyse wordt toegepast moeten ook alle tijdreeksmodellen worden opgesteld in deze stap. Idealiter is hiervoor is een dataset beschikbaar met ruwe en gevalideerde data. Daarna is de tool gereed voor gebruik.

Tijdens gebruik komt nieuwe ruwe data binnen die door de tool wordt geanalyseerd. Het resultaat is een reeks waarin metingen gelabeld worden als verdacht of niet verdacht. Indien gewenst kan ook worden aangegeven welke regel een meting als verdacht classificeert (b.v. de droogval regel, of door het tijdreeksmodel). De volgende stap is dan het accepteren of negeren van de suggestie door de automatische foutherkenningstool. De uitkomst van die stap zou zelfs weer terug naar de tool gestuurd kunnen worden om de optimalisatie van de parameters nogmaals uit te voeren.

Toepassing methodes zonder gevalideerde tijdreeks

In deze studie is alleen gekeken naar locaties waar zowel een ruwe als een gevalideerde tijdreeks beschikbaar zijn. In de praktijk is dit niet altijd het geval. Het zou interessant zijn om te kijken of er een methode ontwikkeld kan worden om alsnog een tijdreeksmodel af te leiden op basis van een ruwe reeks (waar de fouten dus nog in zitten). Een idee is om automatische foutherkenning iteratief toe te passen op de ruwe reeks waarbij telkens de verdachte metingen worden weggegooid en opnieuw een tijdreeksmodel wordt afgeleid. Als het dan mogelijk is om na enkele stappen een voldoende goed tijdreeksmodel te maken, kan dat model dienst doen om fouten te detecteren.

Bijlage 1 Overzicht van de data

Inleiding

Deze bijlage bevat een overzicht van de beschikbare gegevens per deelnemende organisatie aan het project TRAVAL.

Databeschikbaarheid

In deze paragraaf wordt per deelnemende instantie beschouwd welke data beschikbaar is. Er is niet naar individuele meetpunten gekeken, maar naar de dataset in algemene zin. De deelnemende organisaties zijn:

- Waterschap Aa en Maas
- Waterschap De Dommel
- Waterschap Rijn en IJssel
- Brabant Water
- Vitens

Waterschap Aa en Maas

Bron: ArtDiver (offline drukopnemers) + FEWS (telemetrische drukopnemers)

De data van Aa en Maas bestaat uit een dataset afkomstig uit FEWS en uit ArtDiver.

De ArtDiver dataset bestaat uit alle actieve waarnemingsfilters die momenteel voorzien zijn van een Diver. Deze meetpunten hebben een offline drukopnemer die drie keer per jaar wordt uitgelezen. Daarnaast wordt er gebruik gemaakt van een ArtDiver dataset met alle inactieve Diver meetpunten. Dit zijn meetpunten die in de afgelopen drie jaar zijn vervallen of voorzien zijn van een telemetrisch meetsysteem van Ellitrack. De volgende ArtDiver datasets zijn gebruikt:

- 201906026.adi
- adi_historie_meetreeksen.adi

In totaal bevatten de adi-bestanden 486 waarnemingsfilters met meetreeksen. De ArtDiver dataset is in het meest uitgebreide XML formaat weggeschreven en bevat alle ruwe en gevalideerde metingen, (tijdreeksen van) metadata en de gebruikte controle-handpeilingen. Deze dataset is op 25-6-2019 aangeleverd.

De volgende parameters zijn beschikbaar in de dataset van Aa en Maas (alleen van de offline Diver meetpunten):

PARAMETER	BESCHRIJVING
MEETPUNT.HOOGTE	Bovenkant peilbuis in mNAP
MAAIVELD.HOOGTE	Maaiveldhoogte in mNAP
INHANG.DIEPTE	Opgemeten kabellengte in m
GW.METING.TOTAALCORRECTIE	Gevalideerde meetreeks in mNAP, inclusief alle correcties met comment en flag
LD.METING.BAROCORRECTIE	Luchtdruk in m (kan 0 zijn bij geventileerde loggers of telemetrie systemen)
LD.METING	Ruwe waterdruk logger in m

GW.METING.RUW	Ongevalideerde meetreeks in mNAP. (= Meetpunt.hoogte – Inhang.diepte + LD.meting – LD.meting.barocorrectie)
G.METING.INHANGDIEPTE	Berekende (of gevalideerde) inhangdiepte in m. Dit is de gemeten kabellengte + correctiewaarde bepaald ahv afwijking met handpeiling
T.METING	Temperatuur logger in gr C
C. METING	Conductivity logger in mS/cm (inden CTD logger)
GW.HANDMETING	Handpeiling in mNAP, inclusief flag en evt. comment (opmerking uit het veld)

Daarnaast zijn vanuit FEWS de telemetrie meetpunten geëxporteerd uit ruw en uit productie. Hier geldt als selectie dat het meetpunt apparatuur bevat van Ellitrack of Realsense telemetrische drukopnemers. De ruwe meetreeks bestaat alleen uit NAP metingen. De productie meetreeks bestaat uit de gevalideerde NAP meetreeks inclusief eventuele validatie comments en flags (kwaliteitslabels).

De telemetrische meetpunten bevatten ook de historische NAP metingen die afkomstig zijn van een Diver of eventuele handpeilingen. Bij de dataset afkomstig uit FEWS zal tijdens de analyse gericht worden op het telemetrische deel van de meetreeks. Deze dataset is op 25-6-2019 uit FEWS geëxporteerd.

De koppeltabellen om opmerkingen in de gevalideerde data te vertalen naar de gebruikte categoriën in deze studie zijn hieronder weergegeven.

Tabel 4 Koppeltabel opmerkingen in de data Aa en Maas (divers) en de categorie die in deze studie is gebruikt.

Opmerking data	Categorie ArtDiver
geen opmerking	geen opmerking
vorst	vorst
foute meting	foute meting
piek	piek
onbetrouwbare meting	onbetrouwbare meting
lucht	lucht
droog	droog
inhang	inhang
drift logger	drift logger

Tabel 5 Koppeltabel opmerkingen in de data Aa en Maas (telemetrie) en de categorie die in deze studie is gebruikt.

Opmerking data	Categorie ArtDiver
geen opmerking	geen opmerking
Batterij leeg	geen opmerking
Buis was defect daarom hiervoor geen data	geen opmerking
Diver defect	foute meting
Ellitrack geinstalleerd	geen opmerking
Ellitrack is vervangen	geen opmerking
Ellitrack vervangen ivm intern vocht	geen opmerking
Ellitrack vervangen vanwege dag-nacht ritme	geen opmerking
Geen data, door fout in data verzenden	geen opmerking
Lege batterij	geen opmerking
Ombouwing ivm fout in verzending	geen opmerking
Omgemaaid	geen opmerking

Sensor stoort daarom geen data	geen opmerking
foute inhang	inhang
foute inhang logger	inhang
foute kabellengte	inhang
foute meting	foute meting
inhang LevelTrack	geen opmerking
laatste Diver meting	geen opmerking
laatste Divermeting	geen opmerking
meterfout	onbetrouwbare meting
peilbuis vernield	geen opmerking
start Ellitrack	geen opmerking
uitschieters verwijderd en lineair geinterpoleerd	onbetrouwbare meting
verkeerde meting tijdens uitvoeren handmeting. meting geinterpoleerd	onbetrouwbare meting

Waterschap De Dommel

Bron: FEWS

In onderstaande tabel staan de verschillende reeksen die per meetpunt beschikbaar zijn. De inhangdiepte van de loggers en de bovenkant van de meetpunten is niet bij ons bekend waardoor het niet mogelijk is om vanuit de ruwe gegevens (de KELLER.P1 en KELLER.P2 reeksen) de stijghoogte af te leiden.

PARAMETER	BESCHRIJVING
KELLER.P1	Luchtdruk meting bij de peilbuis
KELLER.P2	Ruwe waterdruk in logger in
STIJGHOOGTE	Ongevalideerde meetreeks in mNAP
STIJGHOOGTE_BEWERKT	Gevalideerde meetreeks in mNAP

De regels die zijn toegepast om van stijghoogte naar stijghoogte_bewerkt zijn opgenomen in een PDF met een stroomschema. Hiervoor zijn onder andere de HardMin en HardMax parameters per locatie gebruikt. Deze waardes zijn beschikbaar gesteld via een Excel sheet met een export vanuit Ultimo.

In zowel de 'stijghoogte' als de 'stijghoogte_bewerkt' reeks zijn opmerkingen te vinden over de betrouwbaarheid van bepaalde metingen.

Opvallend is dat in het ene voorbeeld dat ik heb beschouwd, de 'stijghoogte_bewerkt' reeks langer is dan de ruwe 'stijghoogte' reeks.

Waterschap Rijn en IJssel

Bron: WISKI

De data van Waterschap Rijn en IJssel is afkomstig van WISKI. De volgende parameters zijn aangeleverd:

PARAMETER	BESCHRIJVING
MOMENTAAN.O	Ruwe tijdreeks (eenheid nog onbekend)
MOMENTAAN.V	Gevalideerde tijdreeks (eenheid nog onbekend)
HANDMATIG.O	Ruwe handmetingen (eenheid nog onbekend)
HANDMATIG.V	Gevalideerde handmetingen (eenheid nog onbekend)
CALCULATED.SENSOR	Sensorhoogte
CALCULATED.DIVER	Tijdreeks in mNAP
CALCULATED.HAND	Handpeiling in mNAP

CALCULATED.MBOVPB	Bovenkant peilbuis in mNAP
CALCULATED.MMV	Maaiveld in mNAP

De metadata van de peilbuizen is aangeleverd in een CSV file. De bovenkant buis, inhangdiepte, is bekend, maar er is wel maar een waarde beschikbaar per locatie, dus wijzigingen in de tijd van dergelijke parameters zijn niet beschikbaar in de dataset.

De metingen zijn voorzien van Interpolation Codes en Quality Codes. De interpolation codes doen in principe niks. Wel is aangegeven (telefonisch Peter Kloosterman d.d. 16/09/2019) dat bij een aantal meetpunten interpolatie is uitgevoerd bij de export van de data. In de gevalideerde reeksen is dus geïnterpoleerde data toegevoegd. De Quality Codes geven aan wat er met de ruwe data is gebeurd. De meeste data wordt automatisch gevalideerd op basis van een min/max bereik (deze zijn ruim gekozen op basis van de minimale en maximale NAP waardes in het gebied). Ook wordt gecontroleerd of de laatste meting minder dan een dag geleden heeft plaatsgevonden. Als aan deze checks wordt voldaan krijgt de meting een Quality Code 80. Als de meting buiten het meetbereik valt of niet door de automatische validatie komt krijgt het de Quality Code 254.

Er zijn bij de reeksen handmatig ingevoerde opmerkingen beschikbaar maar deze kunnen (nog) niet in bulk geëxporteerd worden en zijn dus niet beschikbaar.

INTERPOLATION TYPE	
101	No interpolation
102	Linear interpolation
103	Const until next

QUALITY CODE	SFSQ_DESC_S	SFSQ_CODE_S	OPM
-1		Missing	
40	Handmatig gevalideerd (Goed)	Goed	Alleen chemie?
60	Handmatig gevalideerd (Handvalid)	Handvalid	
80	Automatisch gevalideerd (Autovalid)	Autovalid	
120	Handmatig opgevuld (Opgevuld)	Opgevuld	
200	Ruwe Data	Unchecked	
220	-	Verdacht	Alleen chemie
254	Uit meetbereik (Poor)	Invalid	

De koppeltabel om opmerkingen in de gevalideerde data te vertalen naar de categorieën die gebruikt zijn in dit onderzoek is hieronder weergegeven. De opmerkingen in de data hebben vaak veel verschillende manieren waarop ze gespeld zijn, de koppeltabel bevat steeds maar een variant van vergelijkbare opmerkingen.

Tabel 6 Koppeltabel opmerkingen in de data van De Dommel en de categorie die in deze studie is gebruikt.

Opmerking data	Categorie ArtDiver
handmeting wijkt teveel van loggerwaarde	onbetrouwbare meting
Peilbuis verplaatst en hernoemd als kempen05a	overig
meetwaarde boven max meetbereik sensor (r)	foute meting

meetwaarde onder min bereik sensor	foute meting
foutieve sensorregistratie	foute meting
handmatige interpolatie	overig
afkeur op springerigheid (r), nieuwe inhangdiepte	onbetrouwbare meting
sensor defect	foute meting
baro defect	foute meting
historische gegevens ontbreken	overig
handmeting ontbreekt	overig
Droog	droog
Lucht	droog
werkzaamheden locatie	overig
handmatig goedgekeurd	overig
handmatig afgekeurd	onbetrouwbare meting
inhang	inhang
kabellengte aangepast	overig
loggerwaarde ontbreekt	overig
afkeur op levendigheid	onbetrouwbare meting
foute meting	foute meting
logger verplaatst	overig
logger verwijdert	overig
nieuwe logger geplaatst	overig
onbetrouwbare meting	onbetrouwbare meting
ontbrekende waarde, logger wordt geijkt	overig
p2 ontbreekt	foute meting
too many comments	overig
geen opmerking	geen opmerking

Brabant Water

Bron: ArtDiver (offline drukopnemers) + FEWS (telemetrische drukopnemers)

De data van Brabant Water is voor het grootste gedeelte afkomstig uit ArtDiver en bestaat uit 36 pompstations/looprondes (adi-bestanden). Iedere adi-file bevat circa 50 actieve waarnemingsfilters. In totaal bevatten de adi-bestanden 1516 waarnemingsfilters met meetreeksen.

De ArtDiver dataset is in de meest uitgebreide XML variant weggeschreven en bevat alle ruwe en gevalideerde metingen, meta-data en de gebruikte controle-handpeilingen. Deze dataset is op 28-6-2019 aangeleverd. De volgende parameters zijn beschikbaar in de dataset van Brabant Water (alleen van de offline Diver meetpunten):

PARAMETER	BESCHRIJVING
MEETPUNT.HOOGTE	Bovenkant peilbuis in mNAP
MAAIVELD.HOOGTE	Maaiveldhoogte in mNAP
INHANG.DIEPTE	Opgemeten kabellengte in m
GW.METING.TOTAALCORRECTIE	Gevalideerde meetreeks in mNAP, inclusief alle correcties met comment en flag

LD.METING.BAROCORRECTIE	Luchtdruk in m (kan 0 zijn bij geventileerde loggers of telemetrie systemen)
LD.METING	Ruwe waterdruk logger in m
GW.METING.RUW	Ongevalideerde meetreeks in mNAP. (= Meetpunt.hoogte – Inhang.diepte + LD.meting – LD.meting.barocorrectie)
G.METING.INHANGDIEPTE	Berekende (of gevalideerde) inhangdiepte in m. Dit is de gemeten kabellengte + correctiewaarde bepaald ahv afwijking met handpeiling
T.METING	Temperatuur logger in gr C
C. METING	Conductivity logger in mS/cm (inden CTD logger)
GW.HANDMETING	Handpeiling in mNAP, inclusief flag en evt. comment (opmerking uit het veld)

Brabant Water heeft naast haar eigen meetnet (in ArtDiver) ook het beheer van het provinciale meetnet van Noord-Brabant en van een aantal gemeentes. Het provinciale meetnet bestaat uit telemetrische meetpunten (type Koenders) waarvan de data via Lizard in het FEWS van waterschap de Dommel terecht komt. Voor de gemeentelijke meetnetten komt de data tevens in het FEWS van waterschap de Dommel terecht.

De koppeltabel om van de opmerkingen in de gevalideerde datasets naar de categoriën die in deze studie zijn gebruikt te komen is hieronder weergegeven.

Tabel 7 Koppeltabel opmerkingen in de data van Brabant Water en de categorie die in deze studie is gebruikt.

Opmerking	Categorie ArtDiver
geen opmerking	geen opmerking
inhang	inhang
lucht	lucht
droog	droog
foute meting	foute meting
onbetrouwbare meting	onbetrouwbare meting
verstoorde meting	foute meting
onbruikbare meting	onbetrouwbare meting
monsternemer	lucht
monstername	lucht
drift	drift logger
onbruikbaar (drift)	drift logger
drift logger	drift logger
uithaal	inhang
piek	piek

Vitens

Bron: ArtDiver

De data van Vitens is alleen afkomstig uit ArtDiver en bestaat uit 120 looprondes (adi-bestanden) verdeeld over vijf provincies:

- Gelderland: loopronde 1001 t/m 1080
- Overijssel: loopronde 2001 t/m 2046
- Friesland: loopronde 3001 t/m 3032

- Utrecht: loopronde 4001 t/m 4020
- Flevoland: loopronde 5004 en 5005

Iedere adi-file bevat circa 50 actieve waarnemingsfilters. In totaal bevatten de adi-bestanden 6519 waarnemingsfilters met meetreeksen. Het gehele meetnet van Vitens bestaat uit offline Diver drukopnemers van Van Essen (periode 2002 t/m nu).

De ArtDiver dataset is in de meest uitgebreide XML variant weggeschreven en bevat alle ruwe en gevalideerde metingen, meta-data en de gebruikte controle-handpeilingen. Deze dataset is op 23-7-2019 aangeleverd. De volgende parameters zijn beschikbaar in de dataset van Vitens:

PARAMETER	BESCHRIJVING
MEETPUNT.HOOGTE	Bovenkant peilbuis in mNAP
MAAIVELD.HOOGTE	Maaiveldhoogte in mNAP
INHANG.DIEPTE	Opgemeten kabellengte in m
GW.METING.TOTAALCORRECTIE	Gevalideerde meetreeks in mNAP, inclusief alle correcties met comment en flag
LD.METING.BAROCORRECTIE	Luchtdruk in m (kan 0 zijn bij geventileerde loggers of telemetrie systemen)
LD.METING	Ruwe waterdruk logger in m
GW.METING.RUW	Ongevalideerde meetreeks in mNAP. (= Meetpunt.hoogte – Inhang.diepte + LD.meting – LD.meting.barocorrectie)
G.METING.INHANGDIEPTE	Berekende (of gevalideerde) inhangdiepte in m. Dit is de gemeten kabellengte + correctiewaarde bepaald ahv afwijking met handpeiling
T.METING	Temperatuur logger in gr C
C. METING	Conductivity logger in mS/cm (inden CTD logger)
GW.HANDMETING	Handpeiling in mNAP, inclusief flag en evt. comment (opmerking uit het veld)

De koppetabel om de opmerkingen in de gevalideerde dataset te vertalen naar de categorieën die in deze studie zijn gebruikt is hieronder weergegeven.

Tabel 8 Koppeltabel opmerkingen in de data van Vitens en de categorie die in deze studie is gebruikt.

Opmerking	Categorie ArtDiver
geen opmerking	geen opmerking
inhang	inhang
lucht	lucht
droog	droog
droog[verklaring: filter verstopt/lek]	droog
foute meting	foute meting
onbetrouwbare meting	onbetrouwbare meting
twijfelachtig	onbetrouwbare meting
drift logger	drift logger
vorst	vorst
piek	piek
spanningswater	overig
filterverwisseling *	foute meting
behoort niet tot de meetreeks	onbetrouwbare meting

hoort niet tot deze meetreeks	onbetrouwbare meting
hoort niet bij deze meetreeks	onbetrouwbare meting
hoort niet bij dit meetpunt	onbetrouwbare meting
[verklaring: infiltratie via peilbuis]	onbetrouwbare meting
onbetrouwbare meting[verklaring: infiltratie via peilbuis]	onbetrouwbare meting
foute meting[verklaring: infiltratie via peilbuis]	onbetrouwbare meting
piek[verklaring: infiltratie via peilbuis]	onbetrouwbare meting
[verklaring: filter verstopt/lek]	overig
foute meting[verklaring: filter verstopt/lek]	overig
piek[verklaring: filter verstopt/lek]	overig
peilput verstopt	overig
verstopt	overig
lucht[verklaring: filter verstopt/lek]	overig
verstopt[verklaring: filter verstopt/lek]	overig
[verklaring: onttrekking]	overig
[verklaring: variabele winning]	overig

Conclusie

Waterschap Aa en Maas

Voor de data afkomstig uit ArtDiver (offline meetpunten, divers) zijn alle analyses mogelijk.

Voor de telemetrische meetpunten zijn alleen de ruwe en gevalideerde NAP metingen beschikbaar en geen originele drukmetingen e.d. Dit beperkt mogelijk een aantal vergelijkingen.

Waterschap De Dommel

De ruwe en gevalideerde NAP reeksen zijn beschikbaar, dus vergelijking tussen automatische en “handmatig” gevalideerde reeksen is mogelijk. De metadata om ruwe drukmetingen te vertalen naar NAP-waardes dus mogelijk zijn bepaalde vergelijkingen niet mogelijk.

Waterschap Rijn en IJssel

De ruwe en “handmatig” gevalideerde reeksen zijn beschikbaar. Er zijn geen opmerkingen beschikbaar voor de gevalideerde metingen, dus het is niet mogelijk om per categorie fout te onderzoeken hoe goed de automatische methodes het doen. De opmerkingen zijn er wel, echter is exporteren alleen mogelijk per meetpunt en niet in bulk.

Brabant Water

Voor de data afkomstig uit ArtDiver (offline meetpunten) zijn alle analyses mogelijk.

Voor de telemetrische meetpunten zijn alleen de ruwe en gevalideerde NAP metingen beschikbaar en geen originele drukmetingen e.d. Dit beperkt mogelijk een aantal vergelijkingen.

Vitens

Alle analyses mogelijk.

Bijlage 2 Berekenen voorspellingsinterval

Als onderdeel van het project zijn nieuwe methoden toegevoegd aan Pastas waarmee de onzekerheid van TRA-modellen in kaart kan worden gebracht. Het is belangrijk onderscheid te maken tussen twee typen onzekerheidsintervallen: het betrouwbaarheidsinterval en het voorspellingsinterval. Het betrouwbaarheidsinterval geeft de onzekerheid van de gesimuleerde grondwaterstand weer als gevolg van de onzekerheid in de parameters. Het voorspellingsinterval geeft het interval weer waar nieuwe metingen binnen zullen liggen, bv. met een kans van 95%. Het betrouwbaarheidsinterval is altijd smaller dan het voorspellingsinterval.

Voor beide intervallen zijn methoden in geprogrammeerd (zie solver.py op de Pastas Github1 site). Nadat het model is gekalibreerd en de optimale parameters zijn geschat, komen deze methoden in een Pastas model beschikbaar. Tijdens de optimalisatie wordt een benadering van de covariantiematrix van de parameters berekend, op basis waarvan de parameter onzekerheid en correlaties kunnen worden geschat. Strikt genomen is de voorwaarde voor het gebruik van de onzekerheden dat de ruis (noise) die Pastas berekend voldoet aan de voorwaarden van witte ruis: de ruis bevat geen significante autocorrelatie, is homoscedastisch en alle ruis volgt dezelfde verdeling (bv. de Normaal verdeling). Het is aan de gebruiker deze eigenschappen te controleren. Wanneer niet aan deze voorwaarden wordt voldaan dan is het mogelijk dat de geschatte betrouw- en voorspellingsintervallen te smal of te breed zijn.

In Pastas worden de betrouwbaarheidsintervallen geschat met behulp van een Monte Carlo aanpak. Met behulp van de covariantiematrix en de optimale parameter set wordt een groot aantal (N) random samples getrokken uit een multivariate normaalverdeling. Vervolgens wordt de modeluitkomst N-maal berekend. Het betrouwbaarheidsinterval wordt berekend uit de N modeluitkomsten. Het 95% betrouwbaarheidsinterval ligt bijvoorbeeld tussen de 2.5 laagste en 97.5% hoogste modeluitkomst. Dezelfde aanpak kan gebruikt worden voor het schatten van de betrouwbaarheidsintervallen van de respons functies en de individuele contributies van verschillende hydrologische variabelen.

Voor het schatten van het voorspellingsinterval is nog een extra stap nodig. Bij ieder van de N Monte Carlo simulaties van de gesimuleerde stijghoogten wordt een random residu opgeteld die getrokken wordt uit een normaalverdeling met gemiddelde nul en standaardafwijking gelijk aan de standaardafwijking van de modelresiduen. Het voorspellingsinterval, bijvoorbeeld het 95% interval, wordt wederom bepaald uit de N modeluitkomsten waar nu een random residu bij opgeteld is. Om de rekentijd te beperken is de waarde voor N standaard ingesteld op 1000 simulaties voor het voorspellingsinterval en 10^p voor het betrouwbaarheidsinterval (p is het aantal parameters van het model onderdeel).

Zoals beschreven, is de geïmplementeerde methode voor het berekenen van de onzekerheidsintervallen een benadering. De benadering geeft goede resultaten als aan de genoemde voorwaarden voor de ruis (noise) voldaan wordt. Er zijn nog tal van verbeteringsmogelijkheden denkbaar die ontwikkeld kunnen worden voor situaties waarin dit niet het geval is. Daarnaast gaat de methode er van uit dat de residuen heteroscedasties zijn. Dit is niet altijd het geval, bijvoorbeeld als de residuen groter zijn in dalen en/of pieken. Daarnaast gaat de methode ervan uit dat de covariantiematrix goed geschat kan worden, wat niet altijd het geval is. Ook voor deze gevallen kunnen meer nauwkeurige methoden ontwikkeld worden in de toekomst.

1https://github.com/pastas/pastas/blob/master/pastas/solver.py

Bijlage 3 Omschrijving foutherkenningsregels

Regel	Python functienaam	Omschrijving
N*standaardafwijking	val_n_sigma	Metingen buiten ± N keer de standaardafwijking zijn verdacht.
Maximale gradient	val_max_gradient	Als de verandering tussen twee metingen groter is dan een bepaalde grenswaarde binnen een bepaalde periode is de meting verdacht.
Groter dan grenswaarde	val_gt_threshold	Als een meting groter is dan een bepaalde grenswaarde is de meting verdacht
Kleiner dan grenswaarde	val_lt_threshold	Als een meting kleiner is dan een bepaalde grenswaarde is de meting verdacht
Groter dan tijdsafhankelijke grenswaarde	val_gt_threshold_series	Als een meting groter is dan een bepaalde grenswaarde die verandert in de tijd is de meting verdacht
Kleiner dan tijdsafhankelijke grenswaarde	val_lt_threshold_series	Als een meting kleiner is dan een bepaalde grenswaarde die verandert in de tijd is de meting verdacht
Verschil met andere tijdreeks is groter dan grenswaarde	val_diff_other_series_gt_threshold	Als het verschil tussen de tijdreeks en een andere tijdreeks is groter dan een bepaalde grenswaarde is de meting verdacht
Verschil met andere tijdreeks is kleiner dan grenswaarde	val_diff_other_series_lt_threshold	Als het verschil tussen de tijdreeks en een andere tijdreeks is kleiner dan een bepaalde grenswaarde is de meting verdacht
N*standaardafwijking rondom bewogen gemiddelde	val_bandwidth_moving_avg_n_sigma	Als een meting buiten de bandbreedte valt van ±N keer standaardafwijking vanaf het voortschrijdend gemiddelde is de meting verdacht
Stap tussen twee metingen ligt buiten N*standaardafwijking	val_diff_outside_of_n_sigma	Als de sprong tussen twee opeenvolgende metingen groter is dan ±N keer de standaardafwijking (van de sprongen) is de meting verdacht
Verschil met andere tijdreeks is groter dan grenswaarde	val_diff_gt_threshold	Als het verschil tussen twee opeenvolgende metingen groter is dan een grenswaarde is de meting verdacht
Verschil met andere tijdreeks is kleiner dan grenswaarde	val_diff_lt_threshold	Als het verschil tussen twee opeenvolgende metingen kleiner is dan een grenswaarde is de meting verdacht
Verschil tussen twee andere tijdreeksen voldoet aan een willekeurige conditie	val_diff_others_condition	Als het verschil tussen twee andere tijdreeksen voldoet aan een bepaalde conditie is de meting verdacht. De conditie is een python functie en kan alles zijn (b.v. verschil is kleiner dan een bepaalde waarde)

Detectie van spikes	val_spikes	Als er sprong boven een bepaalde grootte plaatsvindt die direct wordt gevolgd door een sprong van vergelijkbare grootte in de tegenovergestelde richting is de meting daartussenin verdacht
Buiten bandbreedte	val_tra_outside_pi	Als een meting buiten een bandbreedte ligt is de meting verdacht.
Langdurige sprongen (offsets) in tijdreeks	val_offset_detection	Zoekt alle opwaartse en neerwaartse sprongen die groter zijn dan een bepaalde waarde en probeert een goede match te vinden. Alle metingen tussen twee tegenovergestelde sprongen van vergelijkbare grootte worden als verdacht gemarkeerd.
Dood signaal	val_flat_signal	Uitgebreide regel om een dood signaal te signaleren. De volgende checks kunnen worden uitgevoerd: - Maximale afwijking gedurende een periode kleiner dan een bepaalde waarde - Maximale spreiding t.o.v. de mediaan in die periode - De helling van een lineaire fit door de metingen in die periode is kleiner dan een bepaalde waarde - De helling voor en na de beschouwde periode hebben een tegenovergesteld teken (dal of plateau) - De meetpunten liggen onder of boven een bepaald niveau (uitgedrukt in een percentage, b.v. onder 5%-waarde van de metingen)
Pas metingen aan op basis van verschil met handpeilingen	val_shift_to_manual_obs	Verschuif de metingen zodanig dat het resultaat op de handpeilingen komt te liggen. Verschillende interpolatie methodes zijn mogelijk.
Behouden van bepaalde opmerkingen	val_keep_comments	Alle metingen die niet zijn voorzien van een bepaalde opmerking worden als verdacht gesignaleerd. Gebruiken om een reeks te creëren waar alleen bepaalde fouten in zitten.
Combineren resultaten	combine_nan	Combineer het resultaat van N andere regels. Waar een van de reeksen een NaN bevat, bevat het resultaat ook een NaN.

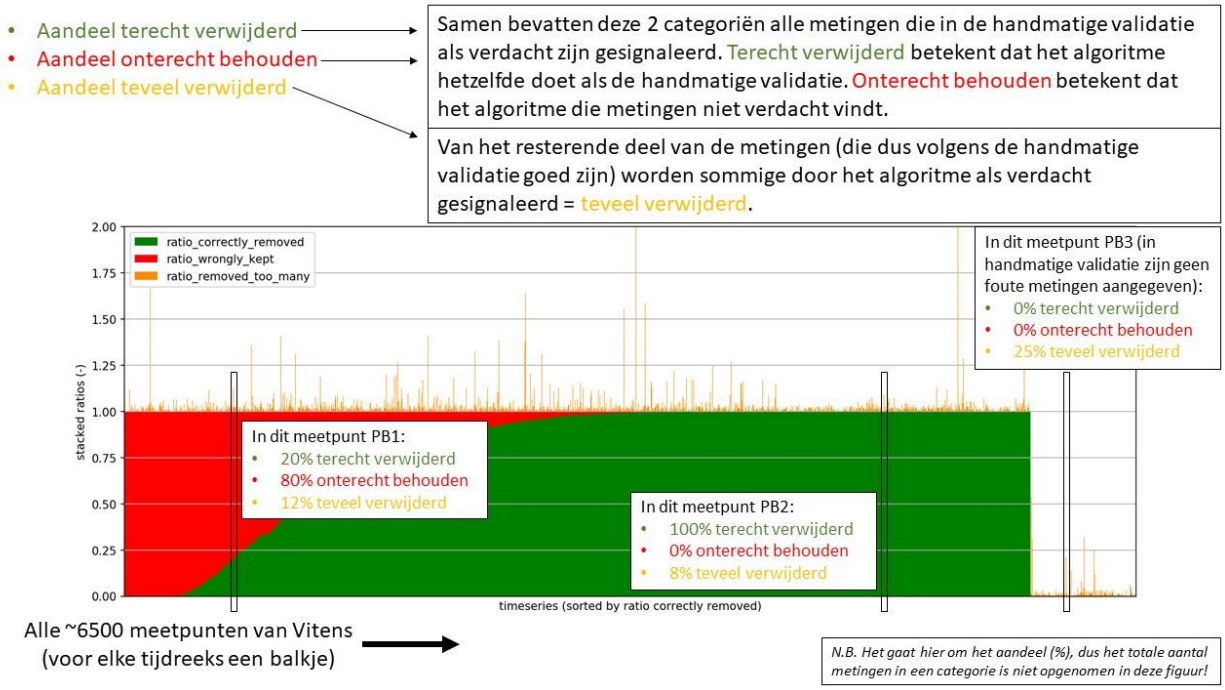
Bijlage 4 Resultaten foutherkenning

Inleiding

Deze bijlage bevat figuren van de uitkomsten van de verschillende automatische foutherkenningsalgoritmes op alle datasets. De figuren geven een overzicht van de uitkomsten van de vergelijking tussen de automatische foutherkenning en handmatige validatie per peilbuis. Elke peilbuis is opgenomen op de x-as als een gestapeld staafdiagram waarbij de verschillende categorieën resultaten zijn weergegeven. De uitkomsten geven per peilbuis het volgende aan:

- Terecht gesignaleerd of “correctly removed” (groen)
- Onterecht behouden of “wrongly kept” (rood)
- Onterecht gesignaleerd of “removed too many” (oranje)

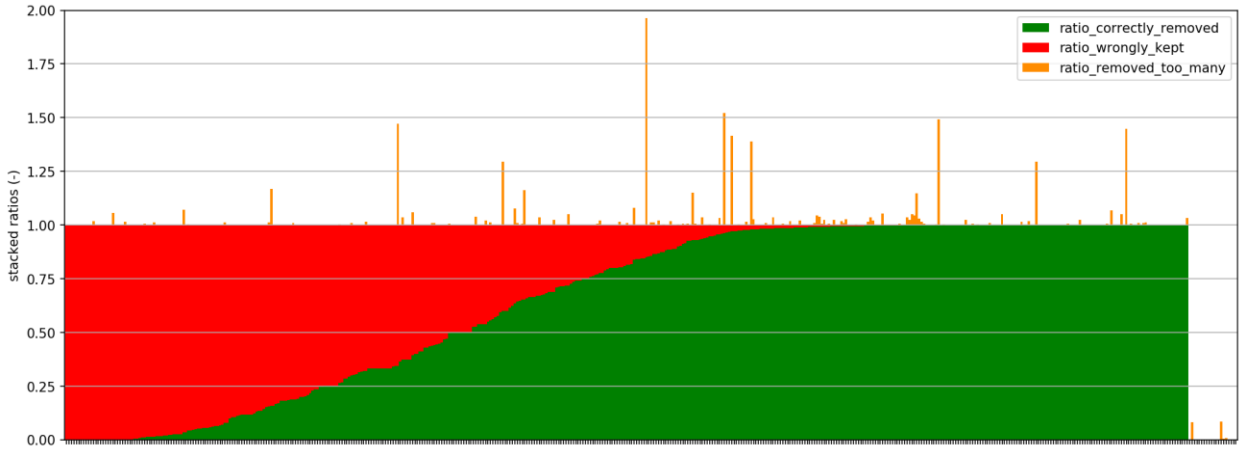
De peilbuizen zijn gesorteerd op het toenemende aantal terecht gesignaleerde fouten. De balken geven het aandeel dat in een bepaalde categorie valt weer en geven dus niet aan hoeveel fouten in absolute zin zijn gesignaleerd of behouden. Het kan voorkomen dat er reeksen zijn die volgens de handmatige validatie geen fouten bevatten. Deze peilbuizen staan aan de rechterzijde van de grafiek en hebben geen groene of rode balkjes. In onderstaande figuur is deze uitleg visueel gepresenteerd.



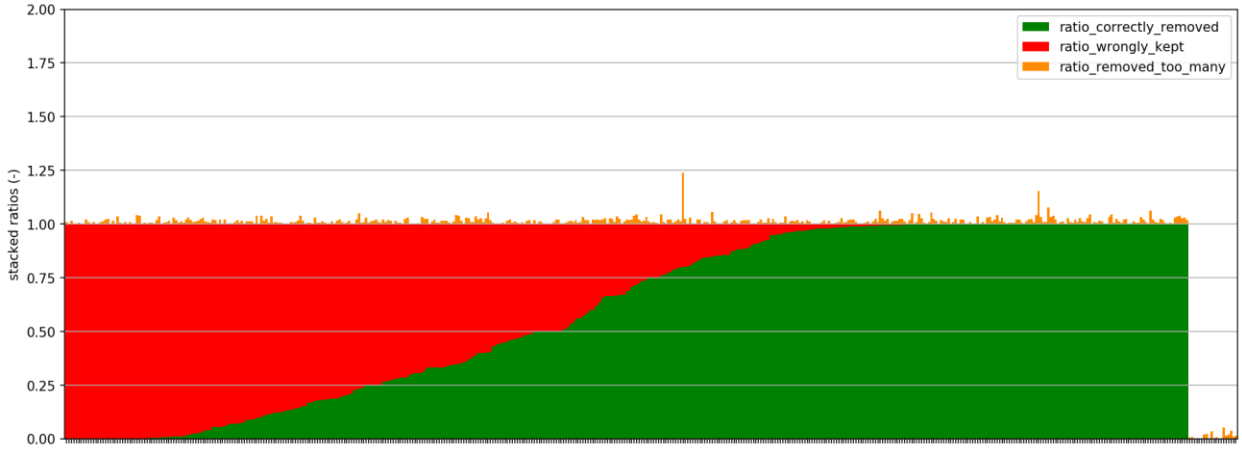
Figuur 25 Uitleg van het gestapelde staafdiagram dat de uitkomst van een automatische foutenherkenningsmethodes op een bepaalde dataset weergeeft.

Het doel van deze grafieken is om te laten zien wat de verdeling is van de resultaten van een automatische foutenherkenning toegepast op een hele dataset. Uit zo’n grafiek kan bijvoorbeeld worden opgemaakt dat in een groot aantal van de peilbuizen de automatische foutherkenning 100% van de fouten weet op te sporen.

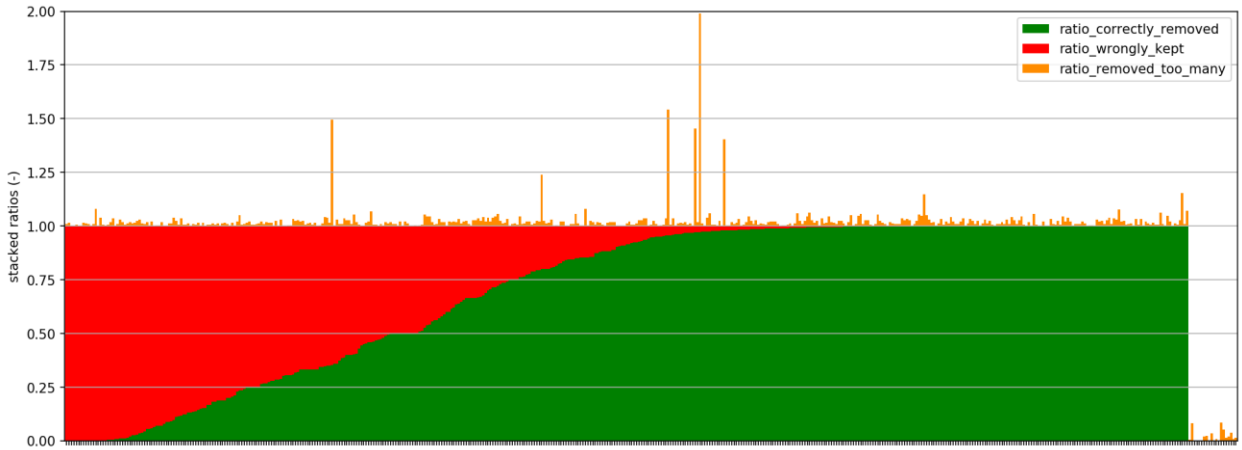
Aa en Maas (divers)



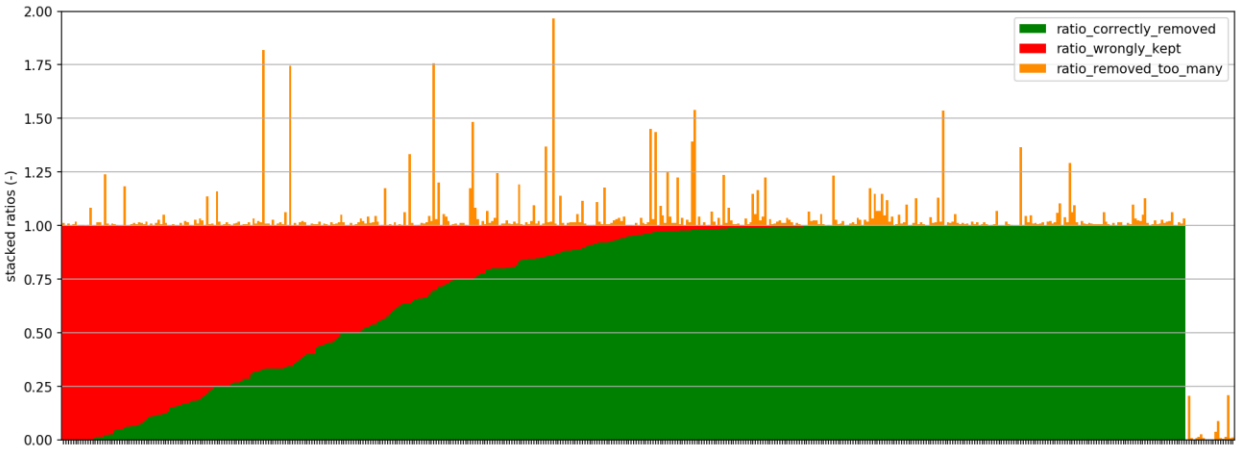
Figuur 26 Methode BASIS (Aa en Maas (divers))



Figuur 27 Methode TRA_Meteo (Aa en Maas (divers))

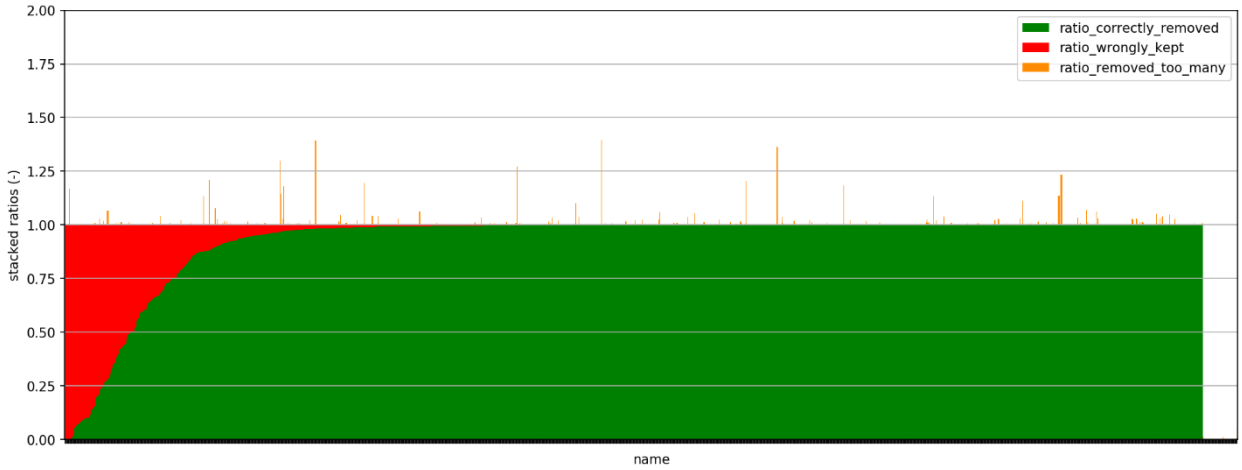


Figuur 28 Methode TRA_Meteo+ (Aa en Maas (divers))

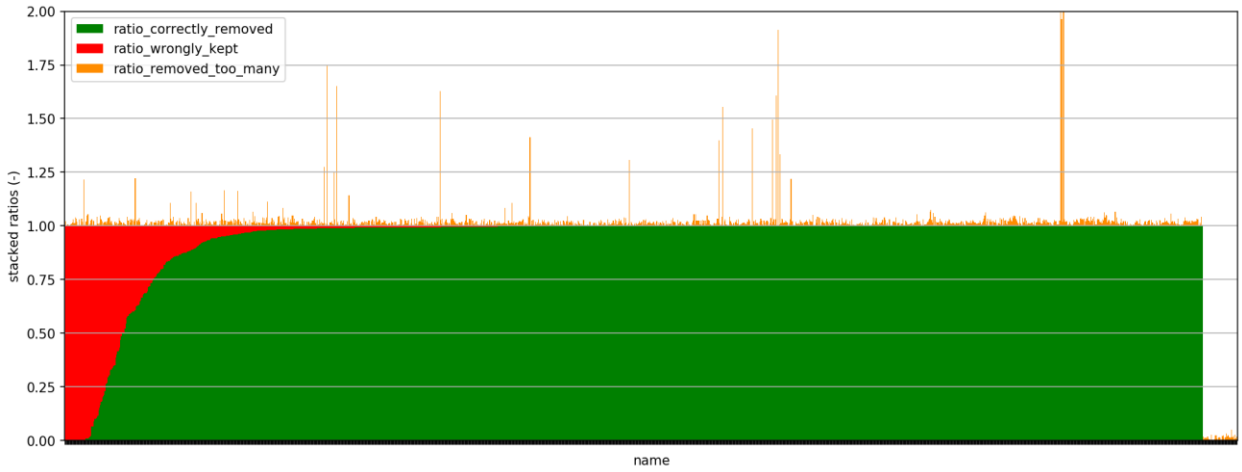


Figuur 29 Methode TRA_Obswell+ (Aa en Maas (divers))

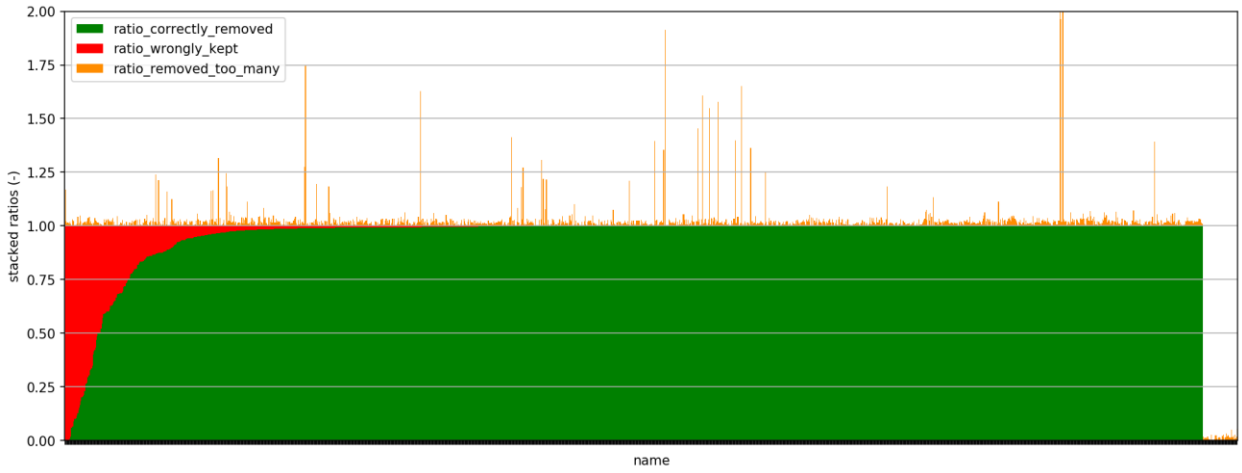
Brabant Water



Figuur 30 Methode BASIS (Brabant Water)

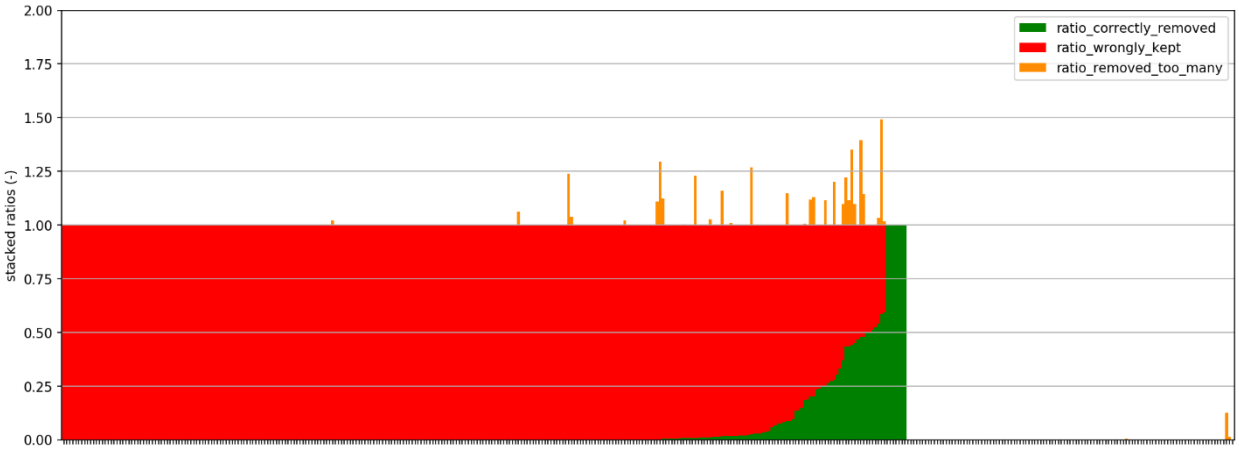


Figuur 31 Methode TRA_Meteo (Brabant Water)

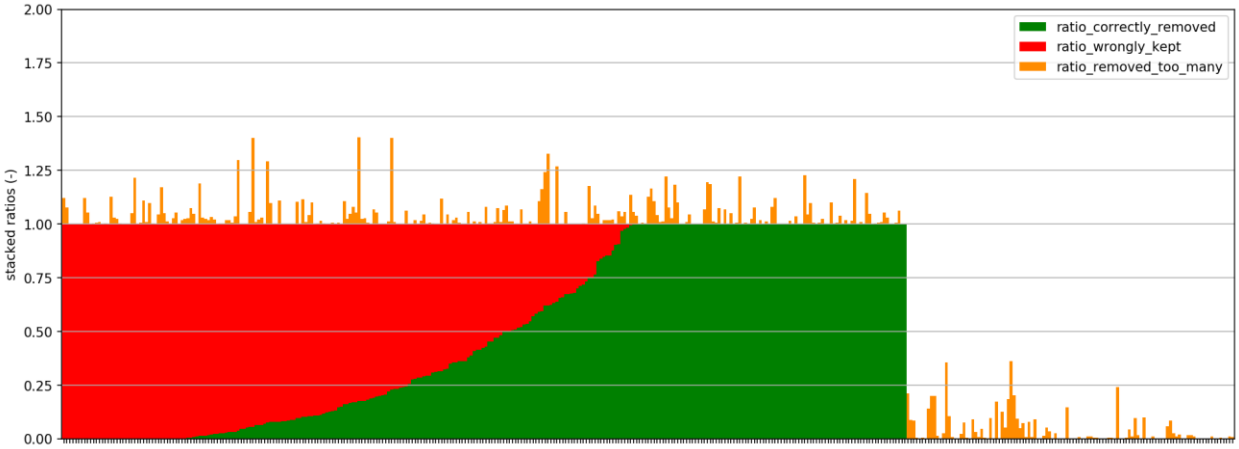


Figuur 32 Methode TRA_Meteo+ (Brabant Water)

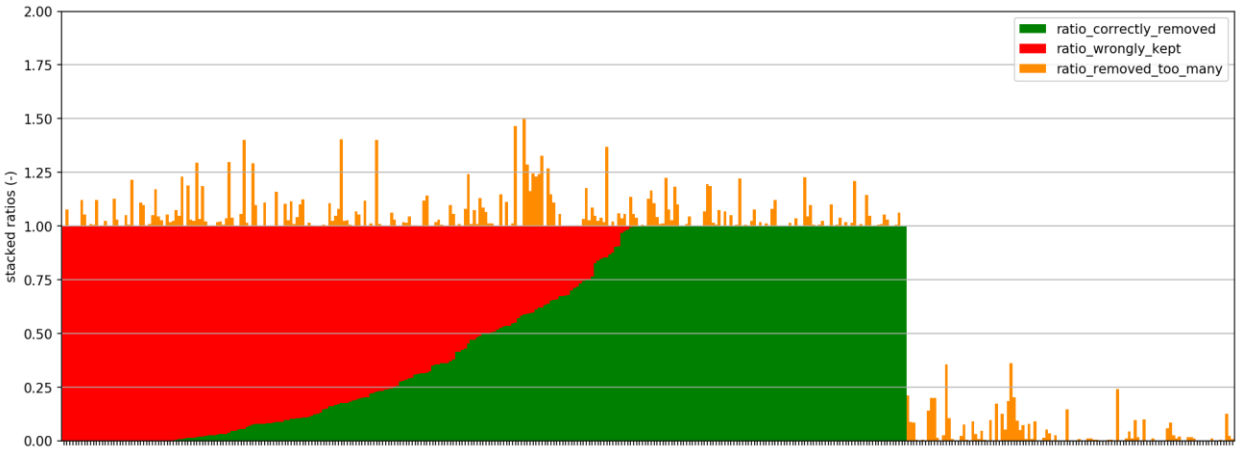
De Dommel



Figuur 33 Methode BASIS (De Dommel)

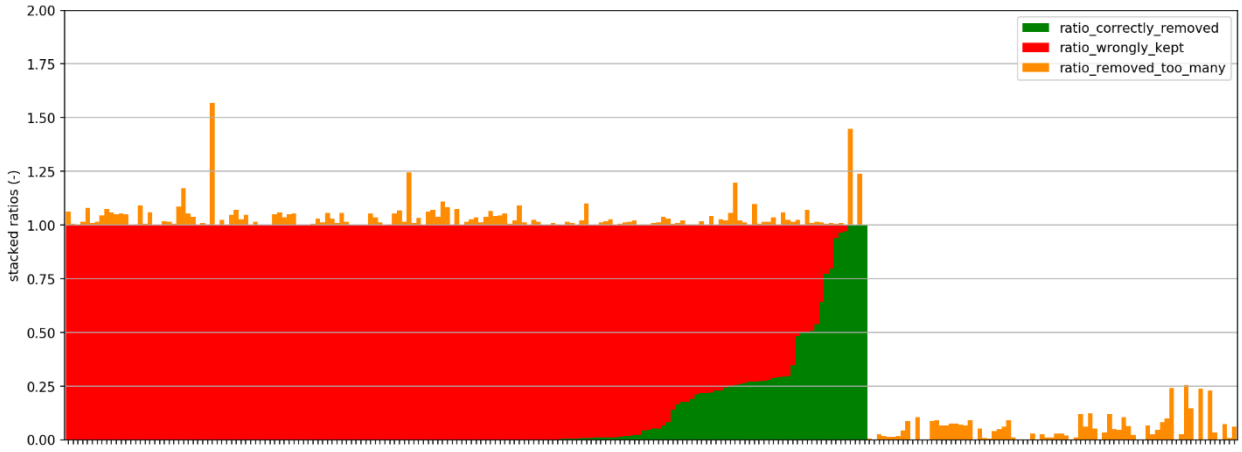


Figuur 34 Methode TRA_Meteo (De Dommel)

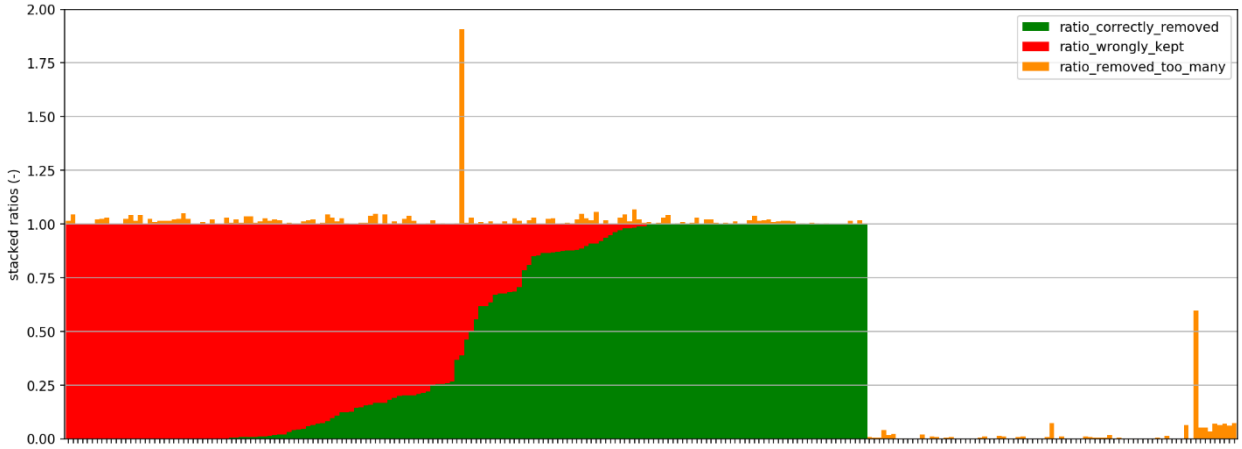


Figuur 35 Methode TRA_Meteo+ (De Dommel)

Rijn en IJssel

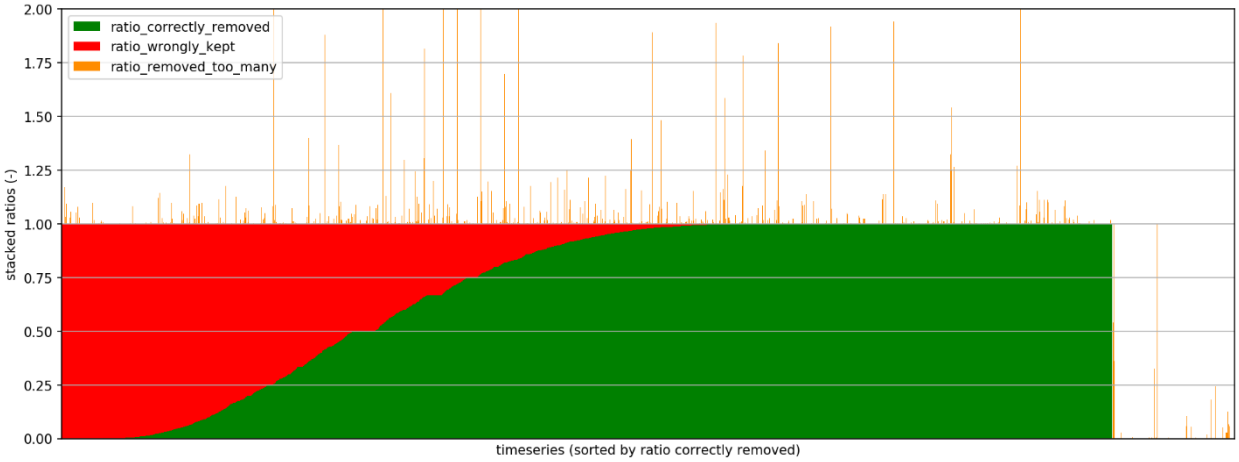


Figuur 36 Methode BASIS (Rijn en IJssel)

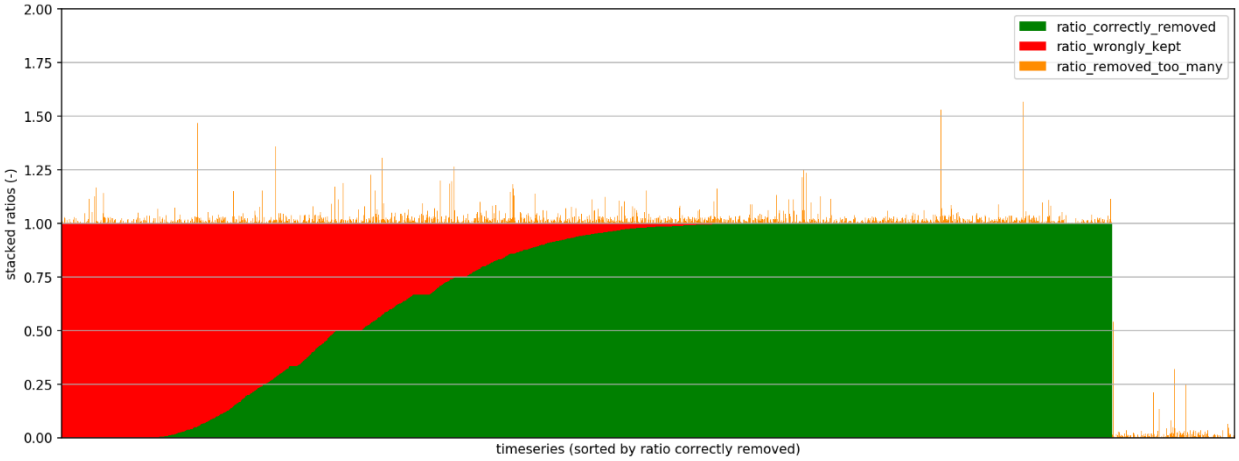


Figuur 37 Methode TRA_Meteo (Rijn en IJssel)

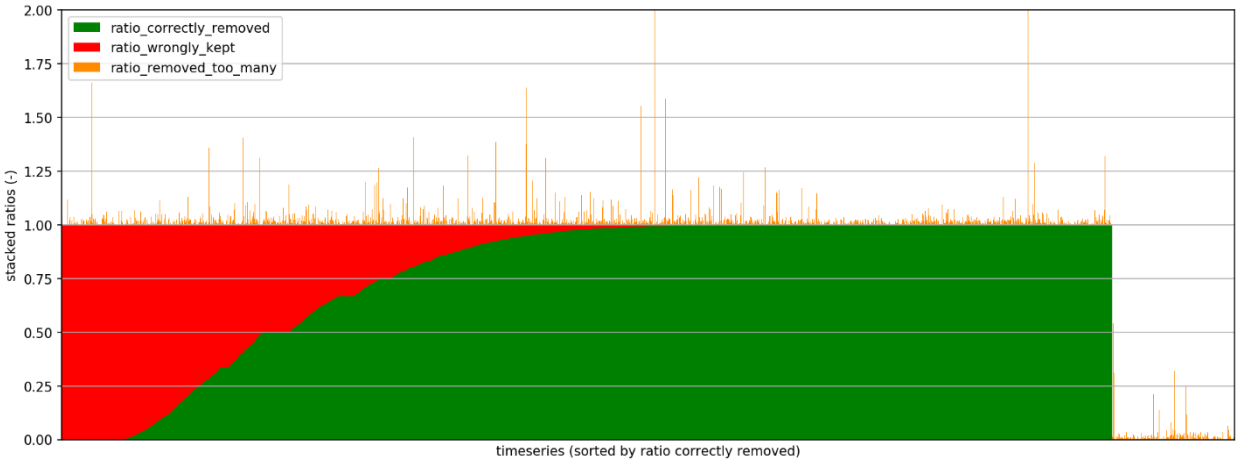
Vitens



Figuur 38 Methode BASIS (Vitens)



Figuur 39 Methode TRA_Meteo (Vitens)



Figuur 40 Methode TRA_Meteo+ (Vitens)

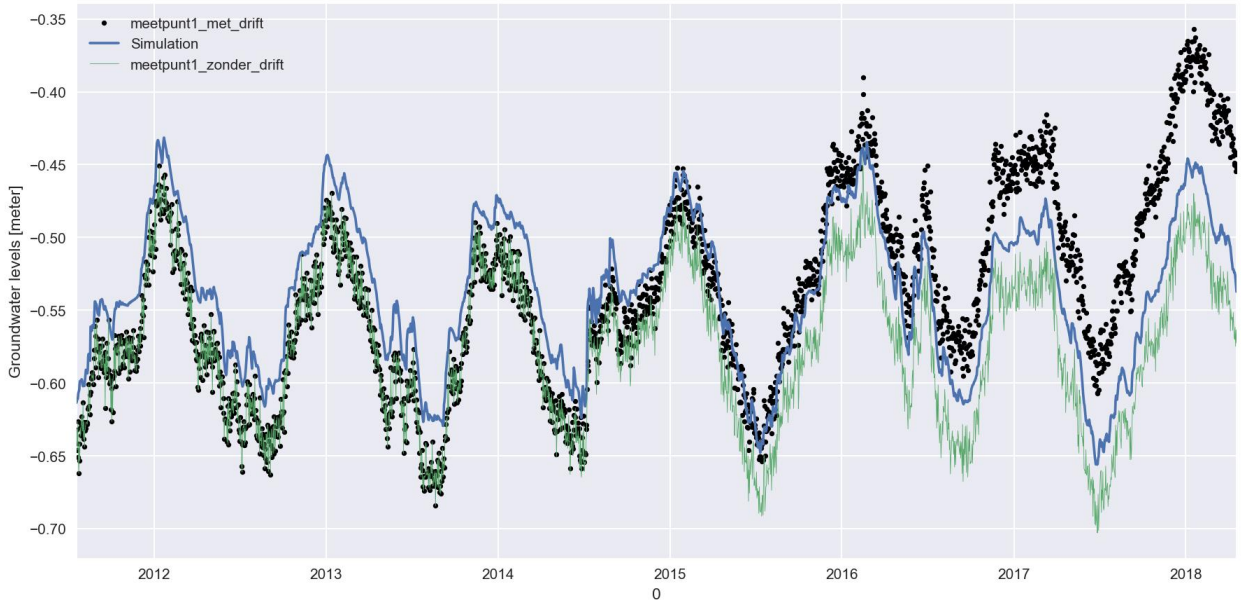
Bijlage 5 Drift herkenning met TRA

Inleiding

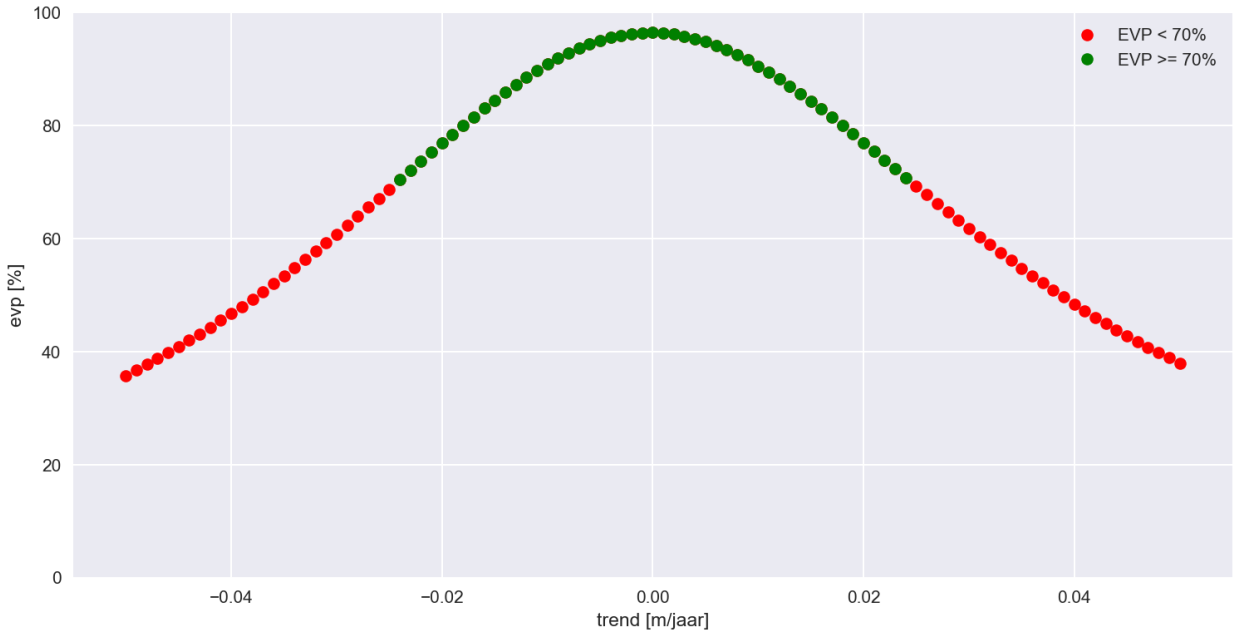
Er is aanvullend onderzoek gedaan naar het herkennen van drift in een tijdreeks met behulp van tijdreeksanalyse. Door een telkens een tijdreeksmodel te fitten op een deel van de tijdreeks kan bestudeerd worden hoe de fit veranderd in de tijd. Als er een trend in de data aanwezig is binnen de beschouwde periode (m.a.w. drift) neemt de fit statistiek al snel af. Dit gedrag is alleen geobserveerd bij een tijdreeks waarvoor een goed model kon worden afgeleid en nog niet breder toegepast. Ook is deze methode nog niet zo ver ontwikkeld dat van individuele metingen gezegd kan worden dat er drift aanwezig is, maar het lijkt wel een veelbelovende indicator te zijn voor de aanwezigheid van drift in een tijdreeks.

Invloed trend op de fit van een tijdreeksmodel

Om te onderzoeken wat de invloed is van een trend in de data op de fit van een tijdreeksmodel is voor een tijdreeks van grondwaterstandsmetingen een trend opgeteld bij de helft van de reeks. In Figuur 41 is de oorspronkelijke tijdreeks (groen) en de tijdreeks met de trend (zwarte puntjes) weergegeven. De blauwe lijn is de simulatie met het tijdreeksmodel dat is geoptimaliseerd op de tijdreeks met drift. De helling van de trend is vervolgens gevarieerd tussen −0.05 m/jaar en +0.05 m/jaar en telkens is het tijdreeksmodel opnieuw geoptimaliseerd en is de fit van het model berekend. De resultaten zijn weergegeven in Figuur 42. De trend in de data heeft een duidelijk effect op de fit van het model. Bij een trend van ±2.5 cm/jaar schiet de EVP onder de 70%.



Figuur 41 Fit van en model (blauwe lijn) op een tijdreeks met een handmatig toegevoegde trend over de helft van de data (zwarte punten). De oorspronkelijke reeks is weergegeven met de groene lijn.



Figuur 42 Helling van de trend versus de fit van het tijdreeksmodel.

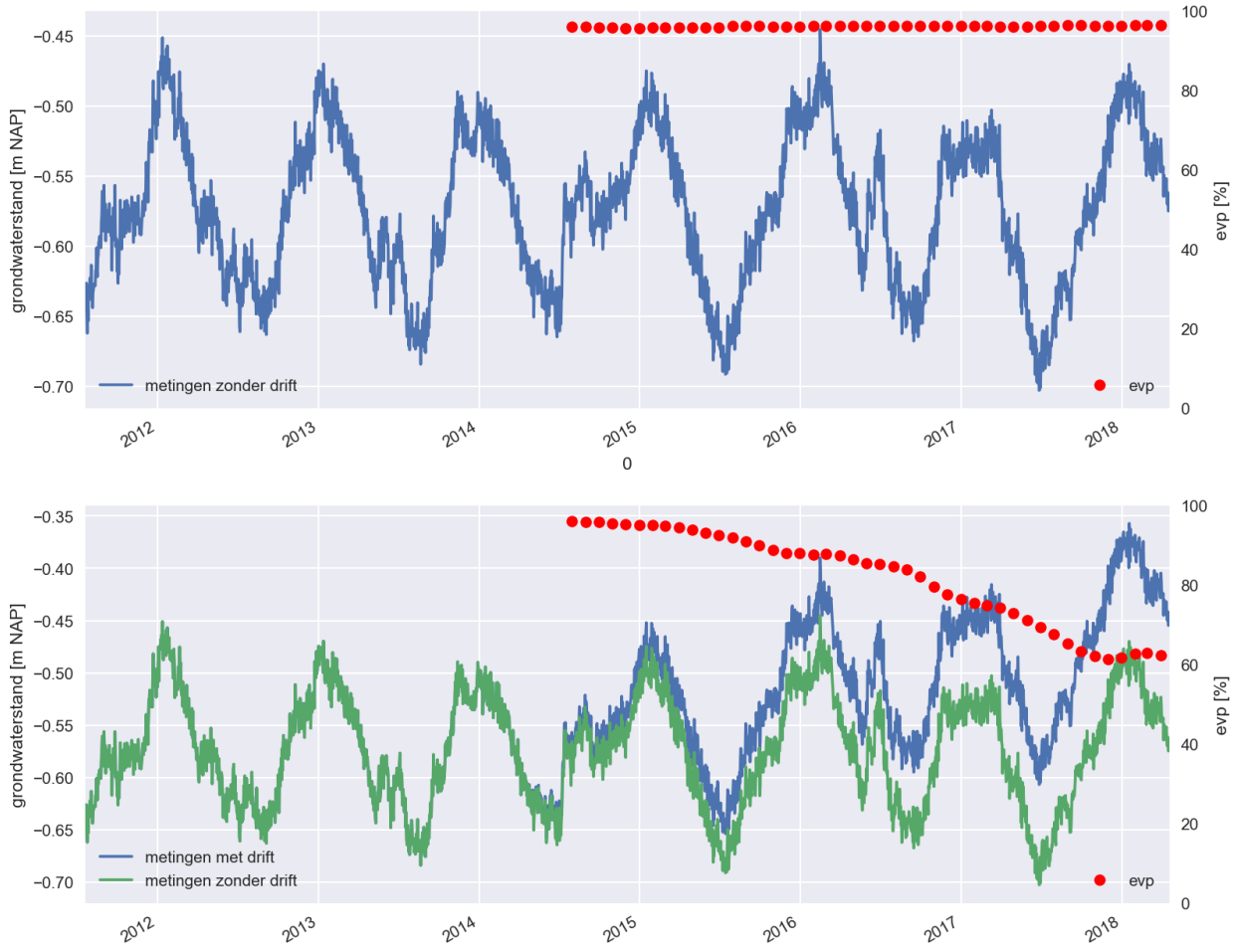
Herkenning van trend met een voortschrijdende fit tijdreeksmodel

Om drift te herkennen in een tijdreeks is gekeken hoe de fit van het tijdreeksmodel veranderd als er een deel van de reeks met een trend aanwezig is in de optimalisatieperiode. Hiervoor is een tijdreeksmodel steeds opnieuw geoptimaliseerd over een langere periode. Naarmate de periode langer wordt valt er een groter deel van de metingen met trend in de optimalisatieperiode. De resultaten zijn gepresenteerd in Figuur 43. De berekende EVP, is telkens weergegeven aan het einde van de optimalisatieperiode. Als controle is de voortschrijdende fit berekend op de tijdreeks zonder trend. Daarin is zichtbaar dat de EVP vrijwel constant blijft. In de situatie met trend met ca. 10% af als er een jaar aan metingen met een trend zijn opgenomen in de optimalisatieperiode.

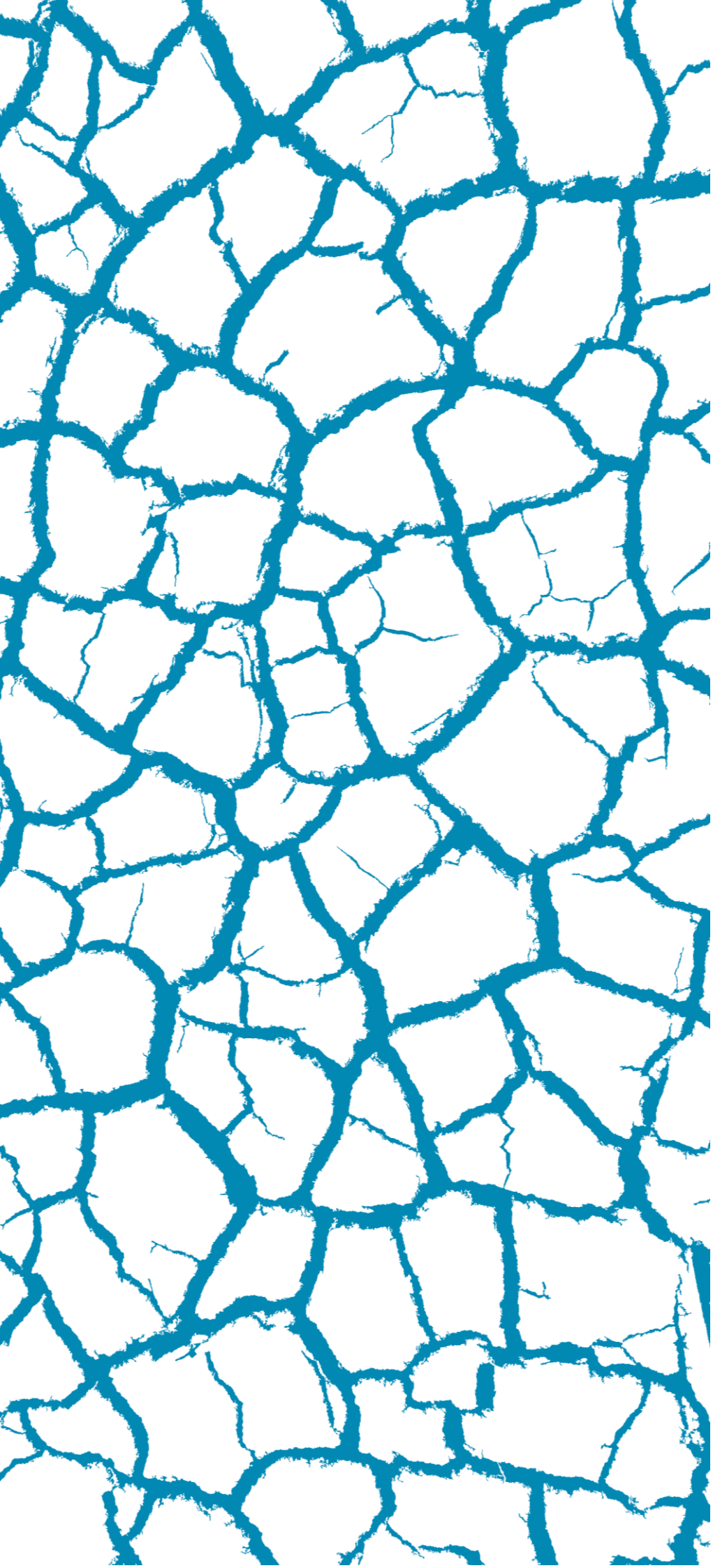
Conclusie

Uit dit voorbeeld met een goed tijdreeksmodel blijkt dat de aanwezigheid van een trend een significante invloed heeft op de fit van het model. Het herkennen van drift in een tijdreeks waarbij steeds nieuwe metingen binnenkomen lijkt ook mogelijk op basis van een voortschrijdende optimalisatieperiode. De methode is wel intensief qua rekenkracht omdat er veel tijdreeksmodellen moeten worden geoptimaliseerd. Verder onderzoek is nodig om aan te tonen hoe deze methode werkt met minder goede modellen, en of deze methode ook werkt op data uit de praktijk. De volgende vragen zouden dan als leidraad kunnen dienen:

- Hoe beïnvloedt een trend de fit van een slecht model (model met een slechte fit om mee te beginnen)?
- Hoe gedraagt de model fit zich bij een voortschrijdende optimalisatieperiode op een uitgebreide dataset uit de praktijk? Komt het vaak voor dat de fit afneemt door andere redenen dan een trend?
- Welke verandering in de fit statistiek beschouw je als een mogelijke indicatie van een trend?
- Wat is de beste lengte van de voortschrijdende optimalisatieperiode voor het herkennen van een trend?



Figuur 43 Voortschrijdende fit van een tijdreeksmodel zonder drift in de data (boven) en met drift in de data (onder).



Korte Weistraat 12

2871 BP Schoonhoven

Tel: (0182) 387138

www.artesia-water.nl