

KWR 2017.059 | December 2017

Analyse van 'slimme meter'-data voor het in kaart brengen van hotspots in het distributienet

Analyse van 'slimme meter'-data voor het in kaart brengen van hotspots in het distributienet

KWR 2017.059 | December 2017

Opdrachtnummer

401366

Projectmanager

drs. Nellie (P.G.G.) Slaats

Opdrachtgever

TKI onderzoek (Koel water: warmwaterzintuigen voor het zorg dragen voor verfrissend drinkwater)

Kwaliteitsborger(s)

dr. ir. Mirjam (E.J.M.) Blokker

Auteur(s)

dr. Joost (J.R.G.) van Summeren, dr. ir. Dirk (D.) Vries, ir. Monique (M.F.M.A.) Albert, ir. Jan-Maarten (J.) Verbree

Verzonden aan

Brabant Water, Nelen & Schuurmans

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

Jaar van publicatie
2017

Meer informatie

dr. Joost (J.R.G.) van Summeren

T +31 30 606 9667

E Joost.van.Summeren@kwrwater.nl

Samenvatting

Belang

Het belang van fris drinkwater is het comfort en de gezondheid van de klant. Hoewel een precieze grens voor drinkwater zonder microbiologische risico's lastig is aan te geven, is in de Drinkwaterwet een maximumwaarde voor drinkwater aan de tap vastgesteld van 25°C. Deze grens wordt in sommige gevallen overschreden. Het is aannemelijk dat warm weer en een stedelijke omgeving de kans op een warme ondergrond en hoge drinkwatertemperatuur vergroten. Eerder onderzoek heeft aannemelijk gemaakt dat vanwege klimaatverandering en verstedelijking dergelijke overschrijdingen in de toekomst vaker zullen voorkomen als geen actie wordt ondernomen.

Brabant Water heeft een nieuw type watermeter geïnstalleerd waarmee dagelijks de minimumtemperatuur van drinkwater in binneninstallaties van klanten wordt gemeten. Hoewel de hoeveelheid data en het aantal meters op dit moment nog beperkt is, biedt toepassing op grote schaal kansen om de drinkwatertemperatuur in het leidingnet inzichtelijk te maken, bijvoorbeeld door een actuele "waterwarmtekaart" te maken. Dit geeft inzicht in wanneer en waar in het leidingnet er temperatuuroverschrijdingen voorkomen en wat de bepalende invloedfactoren zijn.

Aanpak

In het voorliggende rapport is onderzocht hoe een machinelearning-methodiek is in te zetten om uit grote hoeveelheden meetgegevens en omgevingsdata verbanden te vinden en de temperatuur van het water bij de klant te voorspellen. Hiervoor zijn de gegevens verzameld van de watermeters (dagelijkse minimumtemperatuur en volumestroom), het leidingnet (leidingdiameter, materiaaltype, aanlegjaar), het weer (zonneshijn, zonnestraling, bewolingsgraad, luchttemperatuur) en de omgeving (stedelijk hitte-eiland effect, gebouwhoogte, landgebruik, grondsoort). Bij de data-voorbewerking is rekening gehouden met de directe omgeving bovengronds (door variabelen te middelen over een cirkelvormig gebied rond de meters) en ondergronds (door variabelen te middelen langs het stroompad, teruggerekend vanaf de watermeter). Om systematisch de samenhang tussen de parameters te onderzoeken is gebruik gemaakt van Gradient Boosting Regressie. Dit is een machinelearning-techniek waarbij bij elke vertakking van een 'regressieboom' de fout tussen predictie en meting (residu) wordt verkleind door een attribuut in de regressieboom mee te nemen. De gradiëntmethode bij GBRT en een van te voren geselecteerd criterium (bijvoorbeeld kleinste kwadraten) zorgen ervoor dat GBRT de predictiefout minimaliseert. Voor het inzichtelijk ontsluiten van data is onderzocht wat de mogelijkheden zijn van toepassing van het Lizard visualisatieplatform.

Resultaten

Voor 9 onderzochte huishoudens is, na training van het algoritme, met omgevingsfactoren de locatie-specifieke minimum drinkwatertemperatuur op een dag te voorspellen met een nauwkeurigheid van 0,2°C. Van de onderzochte omgevingsfactoren zijn de zonnestraling, buitentemperatuur, zonneshijnuur en bewolingsgraad de belangrijkste bepalende factoren. Onderzochte statische factoren (bodemsot, gebouwhoogte, landgebruik) dragen

nauwelijks bij aan de voorspelling van de locatie-specifieke dynamiek van de drinkwatertemperatuur. De voorspellingen worden nauwkeuriger als (i) de luchttemperatuur wordt gecorrigeerd voor temperatuurveranderingen in de tijd als gevolg van het warmte-eiland effect en (ii) als perioden van laag verbruik worden weggelaten uit de analyse, waarschijnlijk omdat de watermeter dan de temperatuur binnenshuis bemonstert, in plaats van de temperatuur in het leidingnet. Bij middeling over een stroompad is er geen substantiële afhankelijkheid van de onderzochte statische factoren of lengte van het stroompad.

Het Gradient Boosting algoritme dat in dit onderzoek is toegepast, is geschikt voor het beoogde doel. Het is mogelijk om drinkwatertemperaturen te voorspellen en de onzekerheid in de voorspellingen te bepalen. Ook kan kwantitatief worden gemaakt in welke mate factoren bijdragen aan de voorspellingen. Gradient Boosting is geschikt voor opschaling naar grote hoeveelheden data, waarmee de kwaliteit van de voorspelling zal toenemen. Bij opschaling naar grote datasets is het Lizard-platform geschikt om metingen, modeluitkomsten, gegevens van het leidingnet en praktijkervaring bij elkaar te brengen en snel en in hoge kwaliteit te visualiseren. Op deze manier zijn de gegevens beter te duiden.

Vervolgonderzoek en implementatie

Voor het monitoren van de temperatuur aan de tap is het nodig het sensornetwerk uit te breiden. De onderzochte machine-learningtechnieken en visualisatieplatform zijn geschikt om op te schalen naar grote hoeveelheden data, indien er in de toekomst meer sensorgegevens beschikbaar komen.

In dit project is gebleken dat het verzamelen, overdragen en duiden van data veel tijd en inspanning vergt. Een aanbeveling die aansluit bij het TKI-project DIAMANT is om het databeheer te stroomlijnen door de dataset zoveel mogelijk uniform te maken, consistente data-coderingen en eenheden te gebruiken en de metingen eenduidig, transparant en met meta-data vast te leggen.

Er wordt aanbevolen een implementatieplan op te stellen om de meerwaarde voor werkprocessen te identificeren en te bepalen welke eisen dat stelt aan de data en welke visualisatievormen daarvoor gewenst zijn. Implementatie van een machinelearning-methodiek en visualisatietool in het werkproces heeft als voordeel dat grote hoeveelheden data beschikbaar komen en blijven en de informatie te gebruiken binnen de organisatie of te delen met andere partijen.

Inhoud

1	Inleiding	5
1.1	Aanleiding	5
1.2	Het belang van koel drinkwater	6
1.3	Temperatuurontwikkeling in het distributienet	7
1.4	De datamining-aanpak	8
1.5	Vraagstelling en aanpak op hoofdlijnen	8
1.6	Leeswijzer	9
2	Methode	10
2.1	Aanpak data-analyse	10
2.2	Beschrijving van data	11
2.3	Voorbewerking data	18
2.4	Correlatie en regressie	20
3	Data-analyse	23
3.1	Correlatieanalyse	23
3.2	Regressieanalyse	24
3.3	Synthese	32
4	Perspectief voor een alarmsysteem voor hoge temperaturen	34
4.1	Overwegingen bij toepassing	34
4.2	Mogelijkheden voor opschaling naar grote hoeveelheden data	35
4.3	Visualisatie met Lizard	36
4.4	Resultaten en perspectief	39
5	Conclusies en aanbevelingen	40
5.1	Conclusies	40
5.2	Aanbevelingen voor vervolgonderzoek en implementatie	40
6	Literatuur	42

1 Inleiding

1.1 Aanleiding

Met een nieuwe type watermeter wordt het mogelijk om de temperatuur van drinkwater bij de binneninstallaties van klanten te meten. Met toepassing op grote schaal wordt het –in combinatie met andere metingen– mogelijk om de temperatuur van het leidingwater in een gebied te monitoren en met een warmtekaart hot- en coolspots te bepalen. Het doel van het project dat wordt beschreven in het voorliggende rapport is om te onderzoeken of met watermetergegevens en datagedreven technieken een dynamische warmtekaart is te maken, of er relaties met andere data zijn en of hiermee gezondheidsrisico's kunnen worden berekend.

Onderzoek naar de mogelijkheden op dit gebied sluit aan bij de belangen en vragen van Brabant Water: hoe kan het toepassen van sensoren in combinatie met datagedreven methoden (intelligentie in het drinkwatersysteem) bijdragen aan een verbeterde klantgerichte drinkwatervoorziening, processen en kostenefficiënte bedrijfsvoering. In het TKI-project DIAMANT (Vonk, Vries et al. 2016) is in hetzelfde samenwerkingsverband (Brabant Water, Nelen & Schuurmans en KWR) verkent hoe datagedreven methoden kunnen worden ingezet om bedrijfsprocessen te verbeteren. De positieve resultaten waarbij hypothesen over storingsgedrag en de relatie tussen temperatuur en klantmeldingen konden worden geverifieerd, vormde de aanzet om de mogelijkheden verder te onderzoeken in het huidige TKI-project.

Bij het inzetten van kennis van de drinkwatertemperatuur voor operationele doeleinden is het inzichtelijk maken en operationaliseren van de data een belangrijke stap. In dit verband biedt het project een kans voor Nelen en Schuurmans om platforms en tools op het gebied van waterbeheer aan te bieden en te ontwikkelen. Voor KWR is dit project interessant om kennis op te doen en te ontsluiten over de temperatuur in het leidingnet. Daarbij zet het de kennis van waterketen, modellen en data-analysetechnieken in.

Gaandeweg het project is gebleken dat het lastiger dan gedacht om met de beschikbare gegevens een dynamische hotspotkaart te genereren. Dit rapport richt zich daarom, meer dan oorspronkelijk was voorzien, op de potentie van sensorgegevens en datagedreven technieken ten behoeve van het verbeteren van bedrijfsprocessen en welke stappen nodig zijn voor implementatie. Om dezelfde reden is besloten om een tweede werkpakket dat oorspronkelijk was voorzien niet uit te voeren binnen dit TKI-onderzoek. Van dit tweede werkpakket was het doel om de relatie tussen bruinwatermeldingen en temperatuur te onderzoeken.

1.2 Het belang van koel drinkwater

In de Drinkwaterwet (2011) is een maximumwaarde vastgelegd voor drinkwater aan de tap van 25°C. In Nederland is het drinkwater over het algemeen van uitstekende kwaliteit. Omdat drinkwater wordt gedistribueerd zonder toevoeging van desinfectiemiddelen is de microbiologische stabiliteit gevoeliger voor temperatuur dan elders in de wereld. Het leveren van water van voldoende lage temperatuur is in het belang van de volksgezondheid, hoewel precieze grenzen van een "veilige" temperatuur lastig zijn aan te geven. Het is waarschijnlijk dat een hoge temperatuur de groei van pathogenen versnelt (als die in het leidingnet aanwezig zijn). De precieze invloed is echter onduidelijk: de veranderingen in de microbiologische waterkwaliteit zijn het resultaat van een complexe interactie tussen verschillende organismen (Prest, Hammes et al. 2016; Van Bel 2017). Onderzoek naar biofilms heeft uitgewezen dat een temperatuur hoger dan ongeveer 30°C langer dan 7 dagen een groei van *Legionella pneumophila* veroorzaakt waarbij gezondheidsrisico's groot zijn (Van der Wielen 2014). Omdat microbiologisch onderzoek nog geen eenduidige grenswaarde voor een "veilige" drinkwatertemperatuur heeft opgeleverd, is dit vraagstuk in hoeverre het herkennen van hotspots in het leidingnet gezondheidsrisico's kan beperken niet nader onderzocht in dit project. Kennis van de temperaturen in het leidingnet is wel een nuttige stap om uiteindelijk gezondheidsgerelateerde indicatorparameters te bepalen.

Bij pompstations wordt de watertemperatuur gemeten maar in het distributienet is de temperatuur meestal onbekend. Onderzoek naar de kwaliteit van drinkwater in Nederland in 2006 toont dat drinkwater bij pompstations voldoet aan de temperatuurnorm, maar dat de norm soms wordt overschreden in het distributienet: in 2006 werd een overschrijding van 25°C vastgesteld bij 0,1% van de routinematige waterkwaliteitsmetingen, met een maximum gemeten temperatuur van 29,2°C (Versteegh and Dik 2007). De oorzaak zou de bovengemiddeld warme zomer van 2006 kunnen zijn. Vanwege verstedelijking en klimaatverandering is te verwachten dat warmte en, zonder verdere maatregelen, overschrijdingen van de drinkwatertemperatuur vaker zullen voorkomen.

Temperatuur is van invloed op drinkwaterkwaliteitsparameters, zoals het kalkafzettend vermogen, hardheid en de zuurgraad. (Real-time) inzicht in de watertemperatuur kan bijdragen aan een betere bepaling van dergelijke (afgeleide) parameters. Het softsensor-principe is een geschikte methode om afgeleide parameters te bepalen in het gehele distributienet op basis van een beperkt aantal metingen, zoals in een proof-of-principle uitgewerkt door Vries, Van den Akker et al. (2015). Een andere motivatie voor het kennen van de drinkwatertemperatuur is dat fris drinkwater bijdraagt aan het comfort van klanten.

Er zijn nog andere doelen te noemen die echter nog wat verder in de toekomst liggen. Een voorbeeld is het perspectief om klanten proactief te informeren in geval de leidingtemperatuur de maximumwaarde dreigt te overstijgen. Ook biedt inzicht in de temperatuur mogelijk nieuwe handelingsperspectieven om warmteterisico's te beperken, bijvoorbeeld via aanpassingen in het ontwerp van streefstructuren of stedelijke indeling.

Brabant Water heeft een nieuw type watermeter geïnstalleerd waarmee de temperatuur van drinkwater in binneninstallaties van klanten wordt gemeten. Om dataverkeer te beperken (en de levensduur van de batterijen te verlengen) wordt elke dag de minimumtemperatuur doorgegeven. Dit lijkt in eerste instantie beperkt, maar toepassing op grote schaal moet het mogelijk maken om de temperatuur in het distributienet inzichtelijk maken, bijvoorbeeld door dagelijks een actuele "waterwarmtekaart" te maken. In het huidige project is een data-

analyse methodiek uitgewerkt om systematisch relaties tussen de gemeten temperatuur aan de tap en omgevingsfactoren te onderzoeken.

1.3 Temperatuurontwikkeling in het distributienet

Tijdens de distributie van drinkwater door het leidingnet wisselt het water warmte uit met de omgeving. Eerder onderzoek heeft aangetoond dat voor veelvoorkomende bodem- en leidingeigenschappen in Nederlandse drinkwatersystemen het drinkwater in enkele uren de bodemtemperatuur aanneemt (Blokker and Pieterse-Quirijns 2013). Dit betekent dat in de meeste gevallen de bodem de watertemperatuur van het pompstation "overschrijft" voordat het de klant bereikt. Verschillende factoren bepalen de bodemtemperatuur op de diepte van waterleidingen, meestal ongeveer 1 meter onder de oppervlakte:

- *Weersomstandigheden.* De bodemtemperatuur is sterk gerelateerd aan de luchttemperatuur en wordt beïnvloed door de hoeveelheid zonne-instraling. De weerseffecten worden vertraagd en uitgedempt vanwege het bufferend vermogen van de bodem.
- *Oppervlaktegebruik.* Bebouwing, groenvoorziening, wateroppervlak, antropogene warmtebronnen zijn van invloed op de warmtehuishouding. Als gevolg van het stedelijk hitte-eilandeffect (urban heat island effect, hierna ook aangeduid met UHI(E)) kan de temperatuur in stedelijke gebieden enkele graden toenemen ten opzichte van het platteland (Stewart and Oke 2012).
- *Bodemeigenschappen.* Warmtegeleiding door de bodem wordt bepaald door de bodemsoort en grondwaterstanden.

In een voorgaand bedrijfstakonderzoek is het bodemtemperatuurmodel van Blokker and Pieterse-Quirijns (2013) uitgebreid met termen voor urbane evapo-transpiratie, antropogene warmtebronnen en de warmteopslag van gebouwen (Agudelo-Vera, Blokker et al. 2015). De gemodelleerde temperaturen kwamen goed overeen met temperatuurmetingen in urbane (Rotterdam) en peri-urbane (De Bilt) gebieden. De resultaten laten zien dat in de zomer het drinkwater warmer is in urbane gebieden dan in peri-urbane gebieden. De meest extreme consequenties van opwarming zijn te verwachten bij zogenaamde "hotspots: locaties in het leidingnet met gemeten temperaturen boven de gemodelleerde maximum waarden. Bij toepassing van het model op het meest extreme KNMI klimaatscenario (W+) stijgt het aantal dagen in een jaar met een overschrijding van de 25°C-grens van 9 dagen voor de huidige situatie naar 55 dagen in 2030 en naar 83 dagen in 2050. Als gevolg van opwarming van de bodem zullen ook andere bodemfuncties –elektriciteitskabels, leidingen en riolering, voorraad grondwater en archeologische waarde– negatieve effecten ondervinden, zoals aannemelijk gemaakt met een hittestressmodellering voor de Spaarnedammerbuurt in Amsterdam (Mantel and Gevers 2017).

In een vervolgproject van Agudelo-Vera, Blokker et al. (2015) zijn mogelijke bronnen voor ondergrondse hittestress bepaald en de relatie met bodemtemperatuurmetingen onderzocht voor het stedelijke gebied Rotterdam-Noord (Agudelo-Vera 2017). Vermoedelijke boven- en ondergrondse antropogene warmtebronnen (o.a. elektrische verdeelstations, waternetwerken en hoogspanningskabels) en stedelijke kenmerken (o.a. grondsoort, grondwaterstand, verharding/bestrating) zijn in kaart gebracht. De resultaten tonen dat bodemtemperaturen boven de 25°C voorkomen gedurende een aantal weken, met verschillen tussen locaties tot wel 10°C. Lokale omstandigheden zijn hierbij van groot belang. Bepalende factoren voor het optreden van hotspots zijn: (i) locatie altijd in de zon, (ii) locatie onder tegeloppervlak, (iii) de nabijheid van minstens één antropogene warmtebron. Bij de aanleg van leidingen kunnen waterbedrijven dan ook het best deze drie bepalende factoren voor hotspots proberen te vermijden.

1.4 De datamining-aanpak

Brabant Water overweegt het aantal watermeters met een thermometer in de toekomst uit te breiden. Naarmate de hoeveelheid gegevens groeit, wordt het lastiger en bewerkelijker om handmatig nuttige informatie uit de gegevens te destilleren. Het ligt dan voor de hand het analyse-proces te analyseren. Datamining is hiervoor een geschikte methode. Datamining wordt gedefinieerd als het zoeken naar patronen of correlaties in grote datasets met als doel de hieruit verkregen informatie op toegankelijke wijze aan de gebruiker te presenteren (Vonk, Vries et al. 2016). Het uitgangspunt is dat er nuttige informatie is te destilleren uit grote hoeveelheden data van verschillende bronnen. De ruwe data wordt daarbij stapsgewijs omgevormd tot inzichtelijke informatie die is te gebruiken voor operationele doeleinden. Er zijn correlaties aan te tonen, maar geen oorzakelijke verbanden. Datamining wordt inmiddels op commerciële basis ingezet met toepassingen op het gebied van sociale netwerken, militair toezicht, financiële dienstverlening, mobiele telefonie en wetenschappelijk onderzoek (biologie, astronomie, life sciences, etc.) om er slechts een paar te noemen (Sagiroglu and Sinanc 2013). Ook voor de drinkwatersector lijkt bigdata-management kansrijk om bedrijfsprocessen te verbeteren. Datamining is in een voorgaand TKI-project datamining ingezet om storingsfrequenties te ontdekken en bestaande hypothesen over storingsgedrag te verifiëren (Vonk, Vries et al. 2016). In hetzelfde project zijn gegevens over het leidingnet, omgevingstemperatuur en demografische gegevens gecombineerd en heeft een correlatie-analyse een positief verband aan het licht gebracht tussen temperatuur en het aantal klantmeldingen over bruin water (Van Summeren, Raterman et al. 2015). Visualisatie van de meldingen met het Lizard-platform, ontwikkeld door Nelen en Schuurmans, ondersteunde het onderzoek.

1.5 Vraagstelling en aanpak op hoofdlijnen

Het voorliggende onderzoek is gericht op de analyse van grote hoeveelheden data voor het genereren van een actueel inzicht van de temperatuur in het leidingnet en het bepalen van invloedfactoren voor het optreden van hotspots. Dit kan echter nog niet worden gerealiseerd met de warmtemeters die op dit moment zijn geïnstalleerd in Veldhoven. Dit wordt onderkend en de onderzoeksvragen zijn daarop aangepast:

1. Welke verbanden zijn te leggen op basis van een beperkt aantal sensorlocaties? Wat zijn de verwachtingen bij gebruik van meer watermeters?
2. Welke aspecten zijn belangrijk bij het verzamelen, duiden en voorbewerken van data?
3. Is machinelearning geschikt voor het leggen van verbanden tussen temperatuurmetingen en omgevingsdata? Is de data-analysetechniek geschikt voor opschaling naar grote hoeveelheden data?
4. Welke meerwaarde kan het Lizard-platform bieden bij het inzichtelijk ontsluiten van data voor eindgebruikers?

De gebruikte databronnen zijn temperatuur- en volumestroommetingen, weergegevens, leidingneteigenschappen en kenmerken van de stedelijke omgeving. Deze bronnen zijn voorbewerkt om met statistische analysetechnieken de samenhang (correlaties en afhankelijkheden) tussen de parameters te onderzoeken. Daarbij is gekozen voor Gradient Boosting Regressie (Friedman 1999; Friedman 2001). In de toepassing is gebruik gemaakt van de scikit-learn machinelearning-toolbox in Python (Pedregosa, Varoquaux et al. 2011). Deze techniek is voor verschillende applicaties succesvol gebleken (Moisen, Freeman et al. 2006; Vonk, Vries et al. 2016). Voor de ontsluiting van informatie aan de gebruikers is een inzichtelijke datavisualisatie essentieel. Visualisatie-mogelijkheden en schaalbaarheid zijn onderzocht door Nelen & Schuurmans. Daartoe zijn de gegevens geïntegreerd in het Lizard-platform.

1.6 Leeswijzer

Het voorliggende rapport is als volgt opgebouwd:

- Hoofdstuk 2 beschrijft welke gegevens van de watermeter en omgeving zijn gebruikt, welke voorbewerkingen zijn toegepast en welke stappen zijn genomen in de correlatie- en regressie-analyse.
- In Hoofdstuk 3 worden de resultaten van de correlatie- en regressieanalyse gepresenteerd.
- In Hoofdstuk 4 wordt het perspectief en aanbevelingen voor toepassing van de gevolgde methode gepresenteerd. Hier wordt ingegaan op de uitbreiding van een sensornetwerk, databeheer en datavisualisatie met het Lizard platform.
- In Hoofdstuk 5 staan de belangrijkste conclusies van het onderzoek en aanbevelingen voor vervolgonderzoek.

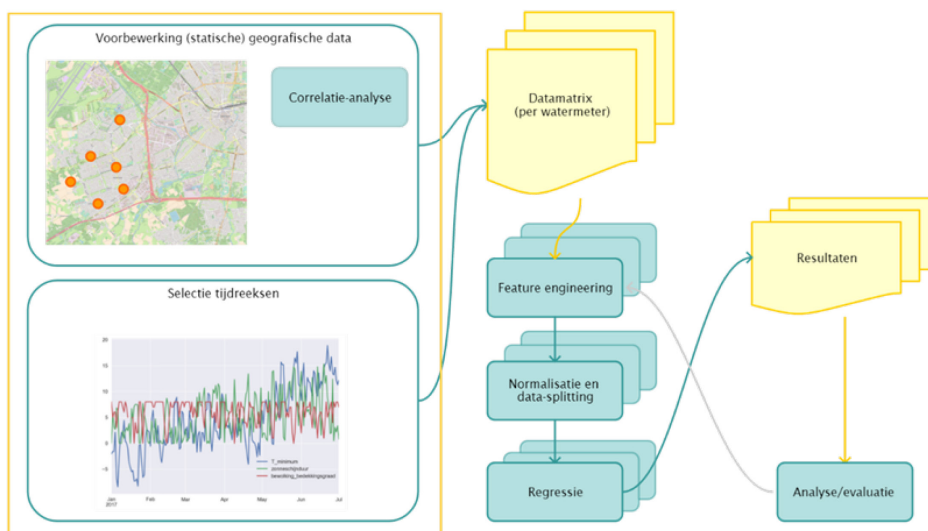
2 Methode

2.1 Aanpak data-analyse

De totstandkoming van een data-analyse is een iteratief proces dat typisch bestaat uit een aantal stappen (de stappen zijn recursieerd benoemd):

- de selectie van data en de bewerkingen erop (*voorbewerking en selectie*);
- het verzamelen van de data en structureren tot een of meerdere matrices (*datamatrix*);
- het selecteren en combineren ('*feature engineering*') van relevante attributen, eventueel aangevuld met *correlatie-analyse*;
- de bewerkingsstap voor de regressie of classificatie door de data te normaliseren en de set te splitsen in een training- en een validatieset (*normalisatie en data-splitting*);
- de *regressie* (of bij andere type problemen: classificatie);
- en tenslotte een *analyse en evaluatie* van resultaten.

Uit de analyse van de eerste regressieanalyses volgen vaak aanwijzingen voor de tweede iteratie, het selecteren en combineren van relevante attributen. Het gehele proces is schematisch weergegeven in Figuur 2-1. In Paragraaf 2.3 worden de data en databronnen van dit project gespecificeerd, paragraaf 2.4 behandelt de stappen die uitgevoerd zijn op de *datamatrix* voor de regressie.



FIGUUR 2-1: STAPPEN DIE GEVOLGD ZIJN OM DE REGRESSIE UIT TE VOEREN. DATA ZIJN WEERGEGEVEN MET GEELOMRANDE KADERS, STAPPEN ZIJN ZEEBLAUW GEKADERD.

Er wordt gekozen voor twee routes om antwoord te geven op de vraagstellingen 1 en 3 in Paragraaf 1.4. De volgende specifieke vragen sluiten hierop aan:

1. Is met alleen lokale aspecten de temperatuur van het leidingwater bepalen?
2. Zijn lokale aspecten belangrijker dan de weg die het water van pompstation tot tap aflegt?

De twee routes zijn beide met regressieanalyses onderzocht, de aanpak staat beschreven in Paragraaf 2.4. Eerst volgt een beschrijving van de data (Paragraaf 2.2) en de datavoorbewerking die nodig was voor de twee routes (Paragraaf 2.3).

2.2 Beschrijving van data

Tabel 1 geeft een overzicht van de databronnen die opgenomen zijn in de datamatrix die als input is gebruikt voor de regressieanalyse. In de tabel is aangegeven of de data is gemiddeld over de directe omgeving van de watermeter (lokaal) of langs de weg die het water aflegt van het waterproductiebedrijf tot aan de watermeter bij de klant (stroompad)

TABEL 2-1 OVERZICHT GEBRUIKTE DATABRONNEN. "LOKALE REGRESSIE" EMN "REGRESSIE LANGS STROOMPAD" REFEREERT AAN HET TYPE VOORBEREIKING, ZOALS UITGELEGD IN DE TEKST. DE RESOLUTIE IN TIJD EN RUIMTE VARIEERT PER BRON EN IS BESCHREVEN IN DE TEKST.

Dataset	Eenheid	Lokale regressie	Regressie langs stroompad	Bron
Meterdata				Brabant Water
- Volmestroom	m ³ /s of l/s	X	X	
- Omgevingstemperatuur	°C	X	X	
- Minimum watertemperatuur	°C	X		
Bodemsoort	-	X	X	PDOK (Alterra)
Gebouwhoogtedata	m	X		PDOK (Kadaster)
Meteorologische data				KNMI
- Minimale luchttemperatuur	°C	X		
- Zonneschijnduur	0,1 uur	X		
- Zonnestraling	J cm ⁻²	X		
- Bewolking bedekkingsgraad	achtsten (octa's)	X		
Landgebruik:				Klimaateffectatlas
- Percentage bebouwing	} % (oppervlak als fractie van het totale oppervlak)	X	X	
- Percentage groen		X	X	
- Percentage verhard		X	X	
- Percentage water		X	X	
- Percentage overig		X	X	
Hittestress				Klimaateffectatlas
- Stedelijk hitte-eilandeffect (dag)	°C		X	
- Stedelijk hitte-eilandeffect (nacht)	°C	X	X	
Leidingkarakteristieken				Brabant Water
- Leidingmateriaal	-		X	
- Leidinglengte	m		X	
- Leidingdiameter	mm		X	

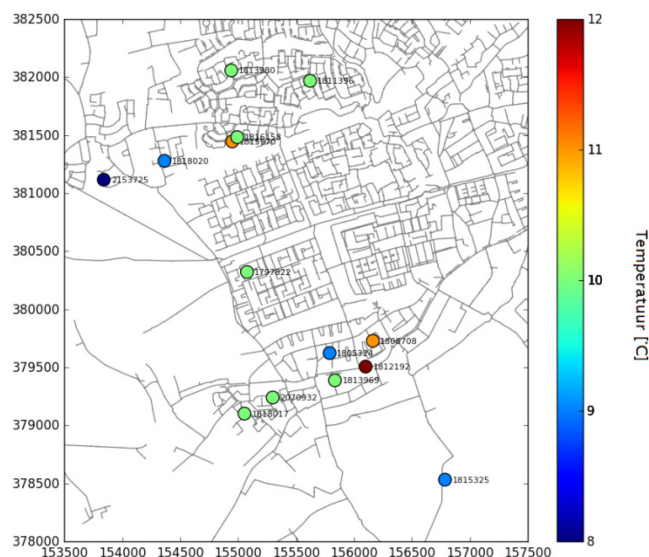
2.2.1 'Slimme meter'-data

Brabant Water heeft de in dit onderzoek gebruikte data van de slimme watermeters aangeleverd. Het gaat om data van 39 meters vanaf de periode van 1 mei 2016 tot en met oktober 2017. De meters zijn geplaatst in 20 gemeenten in Noord-Brabant. Veldhoven heeft de hoogste concentratie van geïnstalleerde watermeters (14 stuks) en is daarom als focusgebied gekozen in dit onderzoek.

Data van de watermeters zou dagelijks gelogd moeten worden, maar voor verschillende meters is dit soms niet gebeurd. Ontbrekende data is meestal het gevolg van een (tijdelijk) onderbroken dataverbinding tussen watermeters en het netwerkpunt (gateway). Vanwege ontbrekende data zijn 6 van de 14 watermeters in Veldhoven onvoldoende bruikbaar bevonden voor verdere analyse.

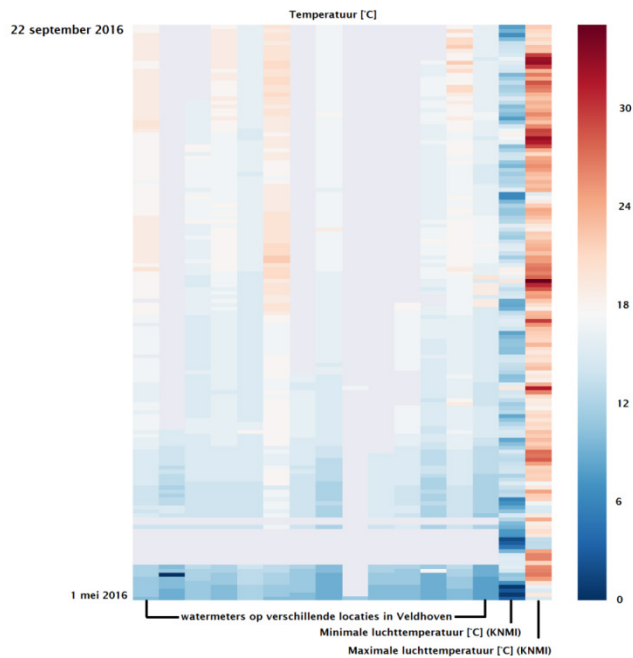
De grootheden die dagelijks door de watermeter worden geregistreerd zijn de volumestroom, de minimum temperatuur van het water dat de watermeter passeert, de omgevingstemperatuur bij de watermeter, en de temperatuur van de gateway. Omdat de gateway niet altijd bij de watermeter is geplaatst en daarom lastig is te duiden is deze temperatuur niet in dit onderzoek meegenomen. Wanneer er water wordt verbruikt, is te verwachten dat de temperatuur aan de tap die van het distributienet zal aannemen. Indien er (voor langere periode) geen water wordt verbruikt, zal de omgeving van de watermeter de minimumtemperatuur die de watermeter registreert bepalen.

Figuur 2-2 laat de locaties van de watermeters met de bijbehorende gemeten temperatuur voor 1 dag zien. Daarnaast geeft figuur 2-3 een voorbeeld van een dynamische temperatuurkaart (de gemeten temperatuurdatta in een periode in de zomer van 2016), en figuur 2-4 de dynamische volumestroomkaart voor dezelfde periode, die uit deze dataset verkregen zijn.

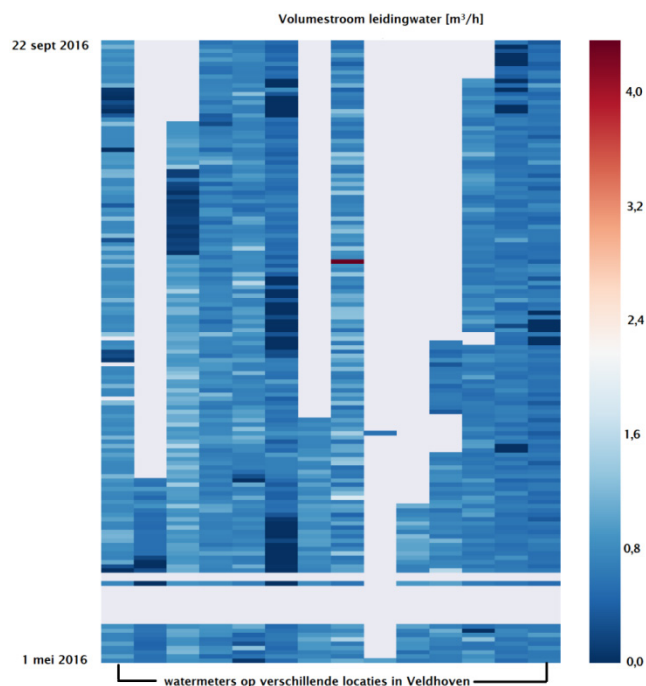


FIGUUR 2-2. VISUALISATIE VAN GEMETEN WATERTEMPERATUUR DOOR WATERMETERS IN VELDHOVEN VOOR 1 DAG (HIER: 1 MEI 2016). HET LEIDINGNET IS WEERGEGEVEN MET DE GRIJZE LIJNEN, DE LOCATIES

VAN DE WATERMETERS. DOOR DE GEKLEURDE BOLLETJES. DE KLEUR VAN DE BOLLETJES GEEFT DE TEMPERATUUR WEER.



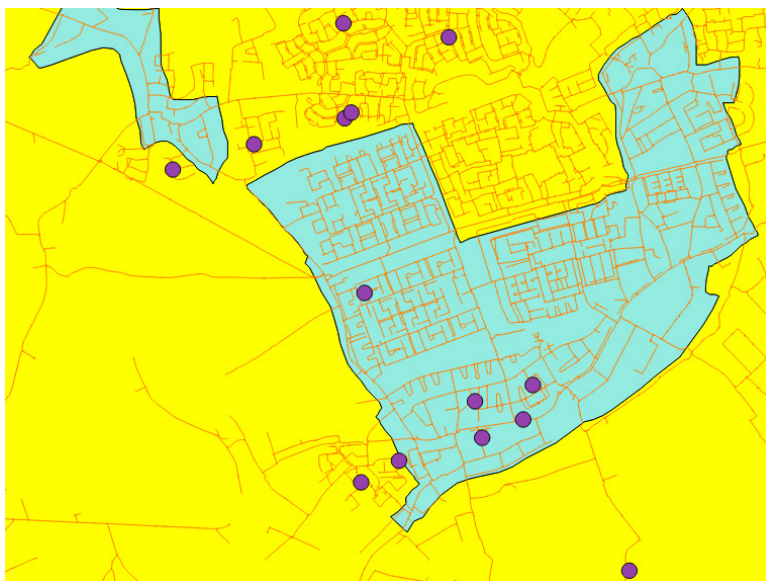
FIGUUR 2-3 DYNAMISCHE TEMERATUURKAART WAARBIJ IEDERE RIJ EEN DAG VOORSTELT (1 MEI -22 SEPT 2016) EN IEDERE KOLOM DE GEMETEN TEMPERATUUR (DE MINIMALE TEMPERATUUR DOOR DE 14 WATER METERS, EN DE MIN. EN MAX. LUCHTTEMPERATUUR GEMETEN DOOR METEOROLOGISCH STATION EINDHOVEN). GRIJS BETEKENT GEEN DATA.



FIGUUR 2-4 DYNAMISCHE VOLUMESTROOMKAART WAARBIJ IEDERE RIJ EEN DAG IS (1 MEI –22 SEPT 2016) EN IEDERE KOLOM DE GEMETEN VOLUMESTROOM (IN m³/UUR) DOOR DE 14 WATER METERS. GRIJS BETEKENT GEEN DATA.

2.2.2 Bodemkaart en kaart van type landgebruik

Om de verschillende typen landgebruik (water, groen, verhard, onverhard, etc.) op te kunnen nemen in de datamatrix zijn deze vertaald naar een numerieke waarde. Er is voor gekozen om de warmtegeleidingscoëfficiënt van de grondsoort mee te nemen in de regressie. Figuur 2-3 laat de grondsoortenkaart voor het focusgebied zien.



FIGUUR 2-3. TYPE LANDGEBRUIK VOOR VELDHOVEN. BLAUW: BEBOUWING, GEEL: ZAND. PAARSE PUNTEN GEVEN DE LOCATIES AAN VAN WATERMETERS. HET LEIDINGNET IS WEERGEGEVEN MET DE ORANJE LIJNEN

2.2.3 Gebouwhoogtedata

Het effect van gebouwhoogte op de te meten watertemperatuur is niet triviaal. De gebouwen zorgen voor schaduw, kan warmte gedurende lange tijd vasthouden en afgeven indien convectie beperkt is door hoogbouw. Figuur 2.2 laat de gebouwhoogtekaart voor Veldhoven zien.

Het gegevensbestand achter deze kaart, '3D gebouwhoogte NL', bevat de voor deze studie gebruikte gemiddelde hoogte van 99,7% van alle gebouwen uit TOP10NL.



FIGUUR 2-4. GEBOUWKAART VAN VELDHOVEN. GROENE BLOKJES STELLEN GEBOUWEN VOOR, RODE STERREN MARKEREN DE LOCATIES VAN DE WATERMETERS.

2.2.4 Meteorologische data

De meteorologische data is vrij beschikbaar van het KNMI. In dit project gebruiken we het dagelijkse minimum van de luchttemperatuur, zonneschijnduur, zonnestraling en bewolgingsbedekkingsgraad. Gegevens zijn afkomstig van het KNMI-weerstation Eindhoven. De minimale luchttemperatuur wordt gemeten op een meethoogte van 1,5 m, en is de laagst gemeten waarde van 24 uurlijkse waarnemingen in een etmaal. De zonneschijnduur (in 0,1 uur) wordt berekend uit 10-minuten waarden van de zonnestraling. Voor zonneschijnduur kleiner dan 0.05 uur wordt een waarde -1 voor zonneschijnduur gegeven. De bewolking-bedekkingsgraad is het gemiddelde van de 24 uurlijkse waarnemingen in een etmaal en wordt weergegeven in achtsten (of octa's) (zie Tabel 2-2, Bron: KNMI).

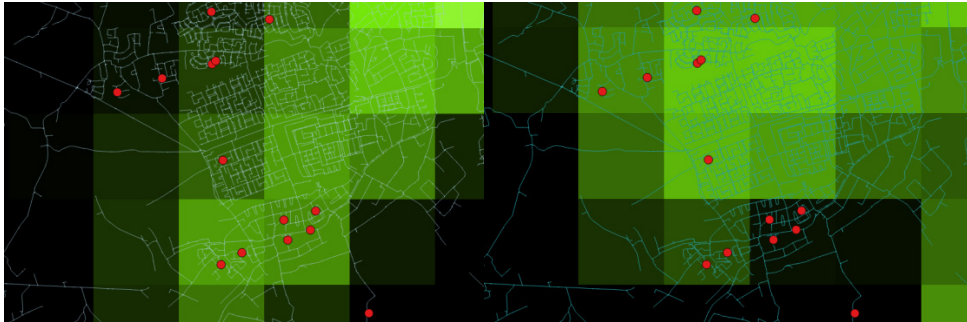
TABEL 2-2 VERTALING VAN OCTA NAAR BEWOLKINGSGRAAD

Octa	Klasse	Octa	Klasse
0	Onbewolkt	5	Half tot zwaar bewolkt
1	Vrijwel onbewolkt	6	Zwaar bewolkt
2	Licht bewolkt	7	Vrijwel geheel bewolkt
3	Half bewolkt	8	Geheel bewolkt
4	Half bewolkt	9	Bovenlucht onzichtbaar

2.2.5 Landgebruik en hitte-eiland effect

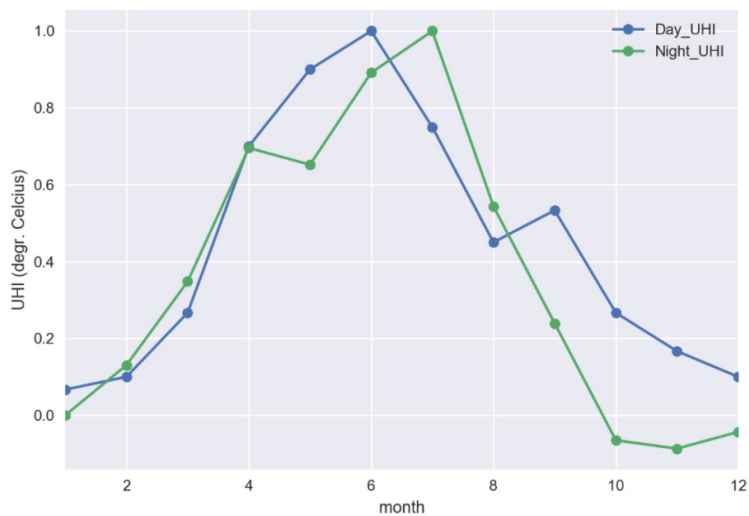
Landgebruikdata en hitte-eiland effect data zijn afkomstig uit de klimaateffectatlas. De data is beschikbaar op wijkniveau. Het landgebruik is onderverdeeld in 5 categorieën: “bebouwd”, “verhard”, “groen”, “water” en “overig” die samen optellen tot 100%. Onder “overig” landgebruik wordt veelal stedelijk gebied (o.a. particuliere tuinen of erven), duinen, zand of heide verstaan. Akkerbouw wordt tot “groen” gerekend. De data is opgedeeld in sectoren op wijkniveau. De dataset achter de kaart bevat voor iedere sector de fractie van het totale oppervlak wat door de vijf typen landgebruik wordt ingenomen.

Het oppervlakte hitte-eiland effect is het verschil in oppervlaktetemperatuur tussen stad en buitengebied (in °C). Figuur 2-5 laat het oppervlakte hitte-eiland effect voor Veldhoven voor zowel overdag als 's nachts zien. Het oppervlakte hitte-eiland effect is 's nachts lager dan overdag wat samenhangt met de instraling van de zon. Bebouwing en verharding dragen in het algemeen bij aan een groter hitte-eiland effect.



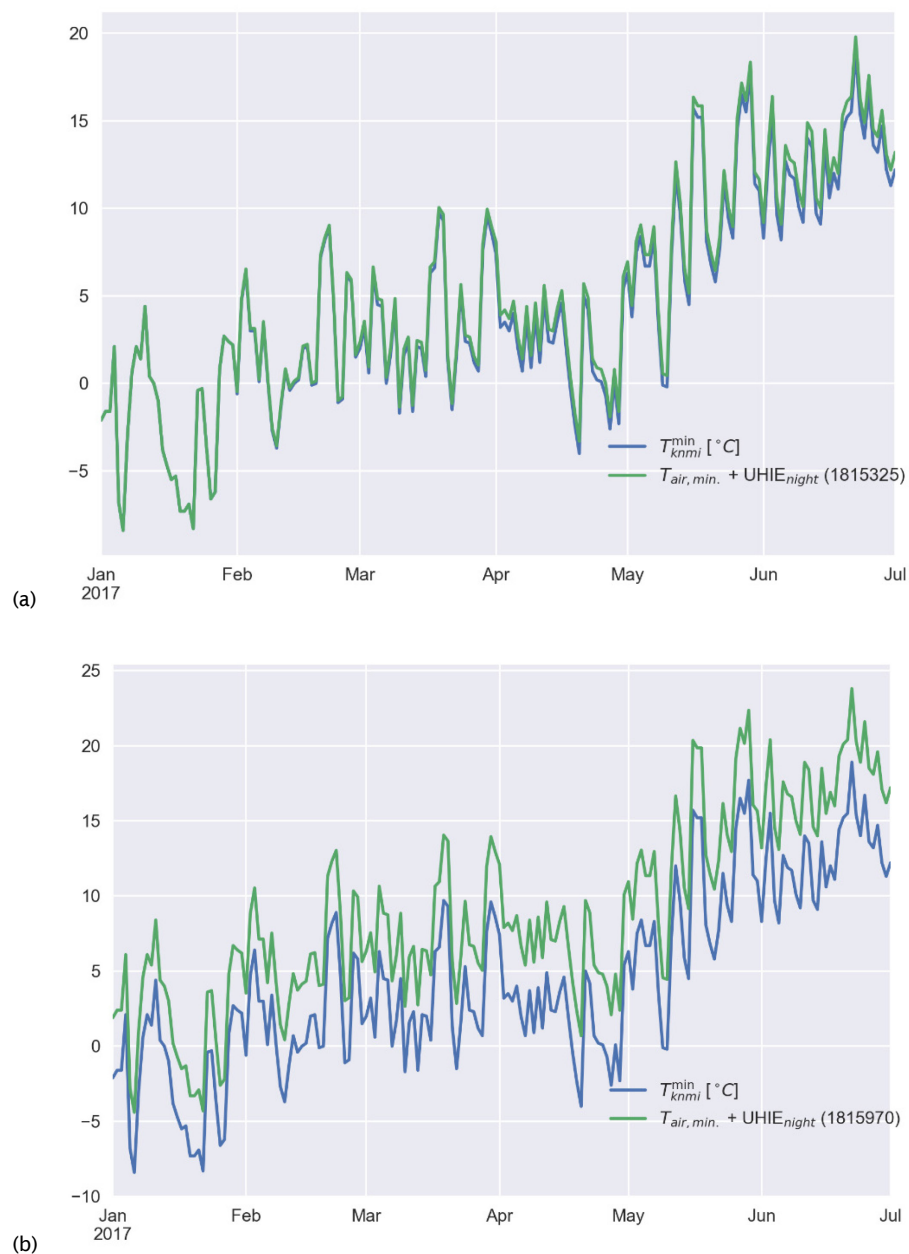
FIGUUR 2-5. OPPERVLAKTE HITTE-EILAND EFFECT VOOR VELDHOVEN (LINKS: DAG; RECHTS:NACHT). DE KLEURSCHAAL VAN DE LINKER KAART (VOOR OVERDAG) LOOPT VAN 0°C (ZWART) TOT 9 °C (LICHTGROEN). DE SCHAAL VAN DE RECHTER KAART (VOOR 'S NACHTS) LOOPT VAN 0°C (ZWART) TOT 4°C. HET LEIDINGNET IS WEERGEGEVEN MET DE BLAUWE LIJNEN.

Ten behoeve van de regressie-analyses (Paragraaf 2.4), zijn de UHI's opgeteld bij de buitentemperatuur van het KNMI-weerstation Eindhoven. Daarbij is rekening gehouden met seizoensvariatie: in de zomer is de UHIE (in Nederland) hoger dan in de winter. Deze correctie is op maandelijkse basis uitgevoerd met gemeten effecten die zijn gepubliceerd door Wolters and Brandsma (2012), zie ook Figuur 2-6.



FIGUUR 2-6: UHIE PER MAAND (NAAR: WOLTERS AND BRANDSMA (2012))

Afhankelijk van omgevingsfactoren, kunnen de gesommeerde buitentemperatuur en UHIE tussen de locaties verschillen. Als voorbeeld worden de UHIE-gecorrigeerde temperatuur bij watermeter 1815325 en 1815970 getoond in respectievelijk Figuur 2-7 (a) en (b).



FIGUUR 2-7: ONGECORRIGEERDE (BLAUWE LIJN) EN UHI-GECORRIGEERDE (GROENE LIJN) BUITENTEMPERATUUR VOOR SENSORLOCATIE 1815325 EN 1815970.

2.3 Voorbewerking data

2.3.1 Lokale factoren: middeling over bovengronds oppervlak

De geografische invloedssfeer van omgevingsparameters is niet op voorhand bekend. Daarom wordt, voor de parameters waarbij dit van toepassing is, voorafgaand aan de lokale regressieanalyse op verschillende afstanden van de watermeters een middeling van de verschillende voorkomende waarden voor deze betreffende parameters gedaan. Er is gekozen voor 4 stralen binnen de afstand van een gemiddelde straatlengte: 10m, 20m, 50m en 100m. De straal rondom de watermeters is visueel gemaakt in Figuur 2-8. Voor de gebouwhoogtedata wijkt de middeling iets af omdat alleen waardes zijn toegekend aan gebouwen (niet aan onbebouwde gebieden). De middeling vindt daarom plaats over het totale oppervlak van de gebouwen in plaats van naar het totale oppervlak.

Of de middeling daadwerkelijk effect heeft is afhankelijk van de locatie van de watermeter en de variabiliteit van de omgevingsparameter. Dit blijkt het geval te zijn voor 'grondsoort', 'gebouwhoogte' en het 'stedelijk oppervlakte hitte-eiland-effect (dag)'.



FIGUUR 2-8. GEBOUWHOOGTEDATA VOOR VELDHOVEN. IN PAARS GEBIEDEN RONDOM WATERMETERS MET EEN STRAAL VAN 100M.

2.3.2 Middeling langs stroompaden

Het is aannemelijk dat sommige omgevingsparameters de drinkwatertemperatuur beïnvloeden tijdens de reis van het water naar de taplocatie. Een voorbeeld is de materiaalsoort van leidingen. Om de grootte van dergelijke invloeden te bepalen zijn een aantal omgevingsfactoren gemiddelen langs het stroompad (in plaats van over een bovengronds oppervlak, Paragraaf 2.3.1). Om dit te realiseren zijn de volgende stappen gezet:

- *Stroompaden.* Er is gebruik gemaakt van stroompad-bepaling met het Dijkstra-algoritme (Dijkstra 1959) voor alle locaties in Veldhoven, zoals bepaald door (Van Rooij 2016). Aan ieder leidingdeel is een weging toegekend op basis van de leidinglengte en -diameter. Door de som van de weegfactoren te minimaliseren zoekt het algoritme naar de "weg van de minste weerstand". Dit wordt beschouwd als het meest waarschijnlijke stroompad.
- *Tijdafhankelijke volumestromen per leidingdeel.* Het totale geleverde volume (vanaf het ingangspunt vanaf Eindhoven naar Veldhoven) is opgesplitst naar ratio van het aantal huishoudens per aansluiting. Grootverbruikers zijn hierin verdisconteerd. Door deze informatie te combineren met het volumestroompatroon op het ingangspunt zijn de volumestroompatronen berekend voor alle leidingen in Veldhoven.

- *Stroomhistorie*. Met de informatie uit de eerste twee stappen is voor elke sensormeting vanaf het verbruikspunt, stroomopwaarts, richting het aanvoerpunt te bepalen waar het water zich bevond op een bepaald tijdstip. De verblijftijd per leidingdeel wordt bijgehouden tijdens deze berekening.
- *Middeling*. De omgevingsfactoren worden gekoppeld aan de leidingdelen. De uitvoerwaarde is het gemiddelde over een instelbare tijdsperiode, gewogen naar verblijftijd per leidingdeel.

Om de invloedssfeer van omgevingsparameters is niet op voorhand bekend. Daarom is de berekening uitgevoerd voor een aantal verschillend aantal instelbare tijdsperiodes: ½, 1, 2, 4 en 8 uur. Deze data is de input voor de regressieanalyse. De middeling over stroompaden is toegepast op de parameters zoals aangegeven in Tabel 1.

2.4 Correlatie en regressie

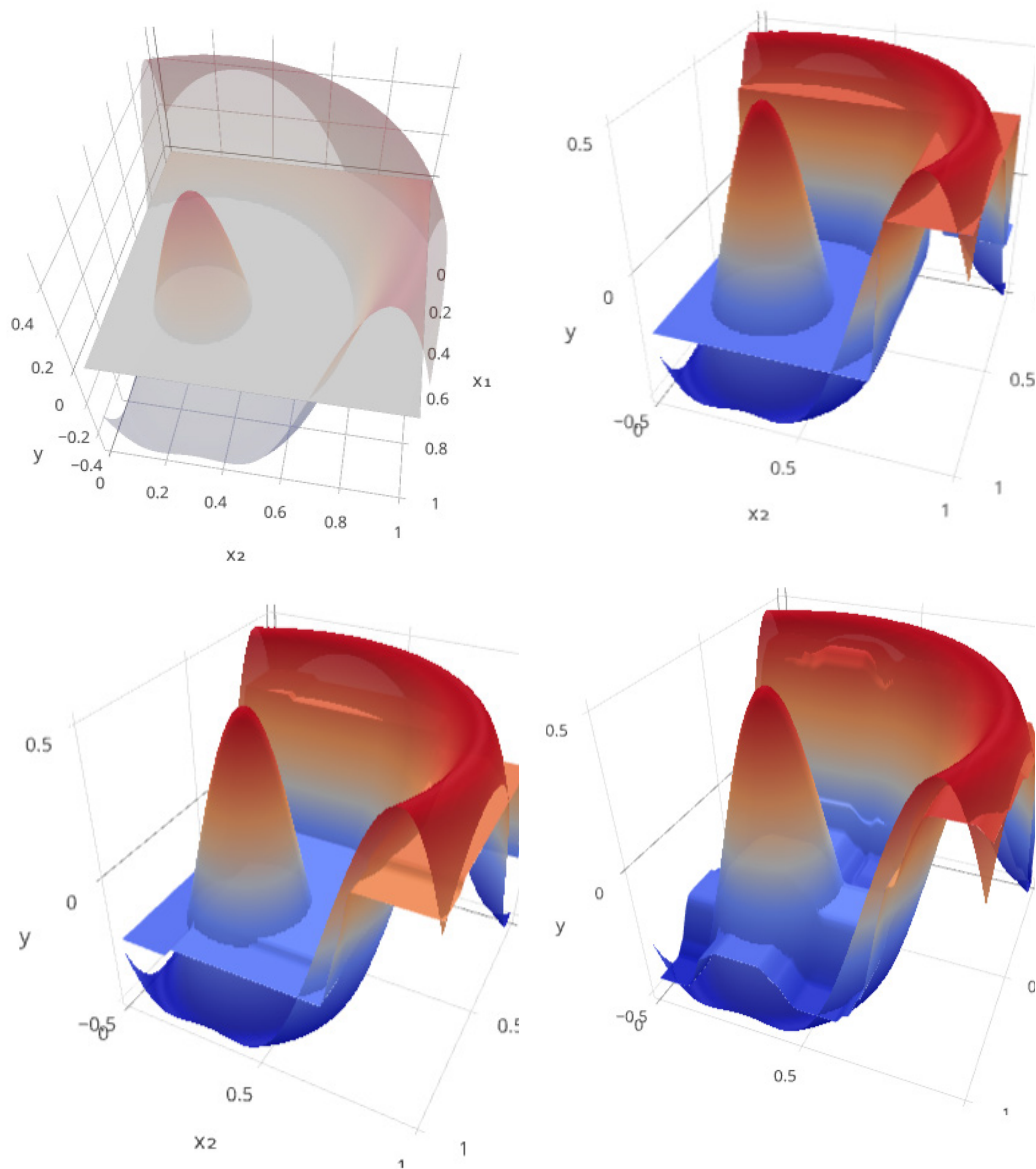
2.4.1 Stappen in de machinelearning

Na de voorbereiding van de data (zie Paragraaf 2.3) zijn de volgende stappen genomen in de data-analyse:

1. *Z-normalisatie* om de data te normaliseren. Deze stap is eigenlijk ook een voorbereiding, maar wordt hier genoemd omdat de machinelearning-programmatuur deze functionaliteit ook bevat. Een z-normalisatie schaalde de data door van elk meetpunt het gemiddelde van de totale meetset van de meetvariabele af te trekken en het resultaat te delen met de standaarddeviatie. De verkregen getallen worden z-scores genoemd. Een z-score heeft daardoor altijd een verwachtingswaarde van 0 en een standaarddeviatie 1 en kan daardoor met z-scores van andere variabelen vergeleken worden;
2. *Clusteranalyse* van correlatie-matrices om de hoeveelheid gecorreleerde attributen zoveel mogelijk te beperken. Correlaties van enkele subsets zijn onderzocht (Tabel 1);
3. *Selectie* van de subset aan attributen die voor de regressie wordt gebruikt. De selectie gebeurt op basis van de clusteranalyse in stap 2;
4. *Gerandomiseerd opsplitsen* van de dataset in een *test* (of ook wel validatie)-set waarbij 25% van de datapunten apart wordt gehouden en een set aan trainingdata (de overige 75%);
5. Geautomatiseerde, *gerandomiseerde selectie* van de *belangrijkste* attributen voor regressie. De mate van belangrijkheid van elk attribuut wordt bijgehouden en gesorteerd;
6. Gradiënt-gebaseerde 'boosting' regressieboom (GBRT), (Friedman 1999; Friedman 2001). Dit is een getrapte regressietechniek ('*tree regression*'), waarbij bij elke vertakking van een 'regressieboom' de fout tussen de voorspelling en meting (residu) wordt verkleind door een attribuut in de regressieboom mee te nemen. Dit betekent dat hoe meer vertakkingen de voorspelling heeft (dus hoe meer attributen van de dataset worden meegenomen), hoe kleiner de residuen. Dit heeft als gevolg dat de fout in de voorspelling (*bias*) kleiner wordt, maar de variantie op voorspellingen groter. Het risico op het fitten van meetruis bij het meenemen van meer attributen wordt groter. De gradiëntmethode bij GBRT, de selectie van de beste regressiebomen en een van te voren geselecteerd criterium (bijvoorbeeld kleinste kwadraten) zorgen ervoor dat GBRT de predictiefout minimaliseert en beter presteert dan 'conventionele' regressiebomen of gradient-boosting methodieken. De GBRT-methodiek en de regressieboom- en gradient-boosting zijn geïllustreerd in Figuur 2-9.
7. *Berekenen prestaties van regressie*: de determinatiecoëfficiënt (R^2) en de kleinste kwadraten tussen meetpunt en predictie (MSE: mean squared error) wordt bepaald.

TABEL 1: OVERZICHT VAN ATTRIBUTEN WAARVOOR CORRELATIES ZIJN ONDERZOCHT

	Attributen
gebouwhoogte	Straal van 10, 20, 50 en 100 m
landgebruik	Onverhard, verhard, groen, water en overig
geleidingscoëfficiënt van de bodem	Straal van 10, 20, 50 en 100 m



FIGUUR 2-9: LINKSBOVEN: DE GOLFFUNCTIE DIE BENADERD MOET WORDEN DOOR EEN REGRESSIEMODEL, RECHTSBOVEN: BENADERING MET REGRESSIEBOOM MET VERTAKKINGSNIVEAU 2, LINKSONDER: BENADERING MET GBRT BESTAANDE UIT 5 BOMEN, VERTAKKINGSNIVEAU 2, RECHTSONDER: BENADERING MET GBRT BESTAANDE UIT EEN VERZAMELING VAN 100 BOMEN, VERTAKKINGSNIVEAU 2.

2.4.2 Studie naar invloed van UHIE, effecten van stroompaden en volumestroom

Door te variëren in de attributen van de dataset en de GBRT-techniek toe te passen op de verschillende datasets, kunnen we kijken welke attributen effect hebben op de prestaties van de regressie.

We onderzoeken de volgende vragen en hypothesen, een schema staat in Tabel 2:

- Is met alleen lokale aspecten de temperatuur van het leidingwater bepalen? Dit onderzoeken we met regressies met codering RL:
 - Verondersteld wordt dat water binnenshuis opwarmt indien er geen waterconsumptie is. De hypothese is dat de prestatie van de regressie (RL-f) verbetert (MSE wordt lager, R^2 hoger) als gefilterd wordt op een lage, maximale volumestroom ($F < 0.1 \text{ m}^3$) ten opzichte van de regressie met ongefilterde data (RL). Het is te verwachten dat het regressiemodel waarin de lage volumestromen zijn weggefilterd (en dus de opwarming binnenshuis minder een rol zal spelen), beter zal correleren met omgevingsdata.
 - We onderzoeken of het meenemen van het Urban Heat Island Effect (UHIE) de prestaties van een regressie (RL-u) en (RL-f-u) verbetert ten opzichte van de regressies waarbij alleen de KNMI buitentemperatuur is meegenomen.
- Kunnen we met data langs de stroompaden het resulterende regressieboom-model (RS) verbeteren ten opzichte van de regressies die zonder stroompadinformatie (RL) wordt bepaald?
- Zien we de invloeden van attributen op de voorspelde watertemperatuur verschillen als we verschillende watermeters bestuderen?

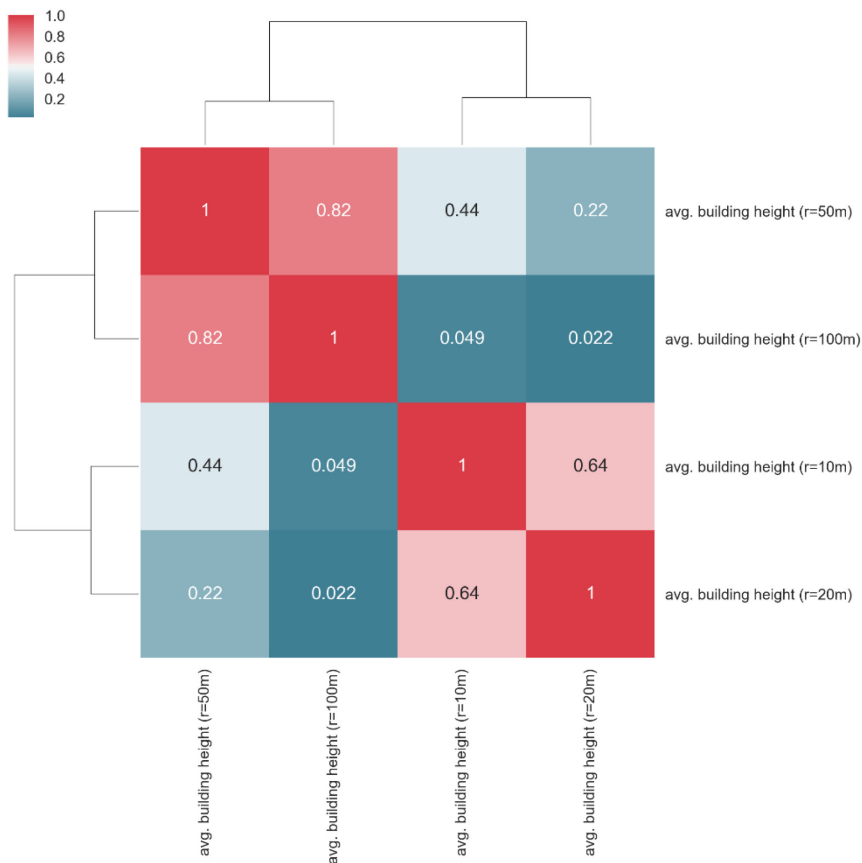
TABEL 2: OVERZICHT VAN REGRESSIES (A.A.: ANDERE ATTRIBUTEN, F: VOLUMESTROOM (FLOW))

Codering regressie	Luchttemperatuur Eindhoven + a.a.	Luchttemperatuur Eindhoven + UHIE + a.a.	Geen filtering op F	Filter indien $F < 0.1$	Stroompaden
RL	x		x		
RL-u		x	x		
RL-f	x			x	
RL-f-u		x		x	
RS				x	x

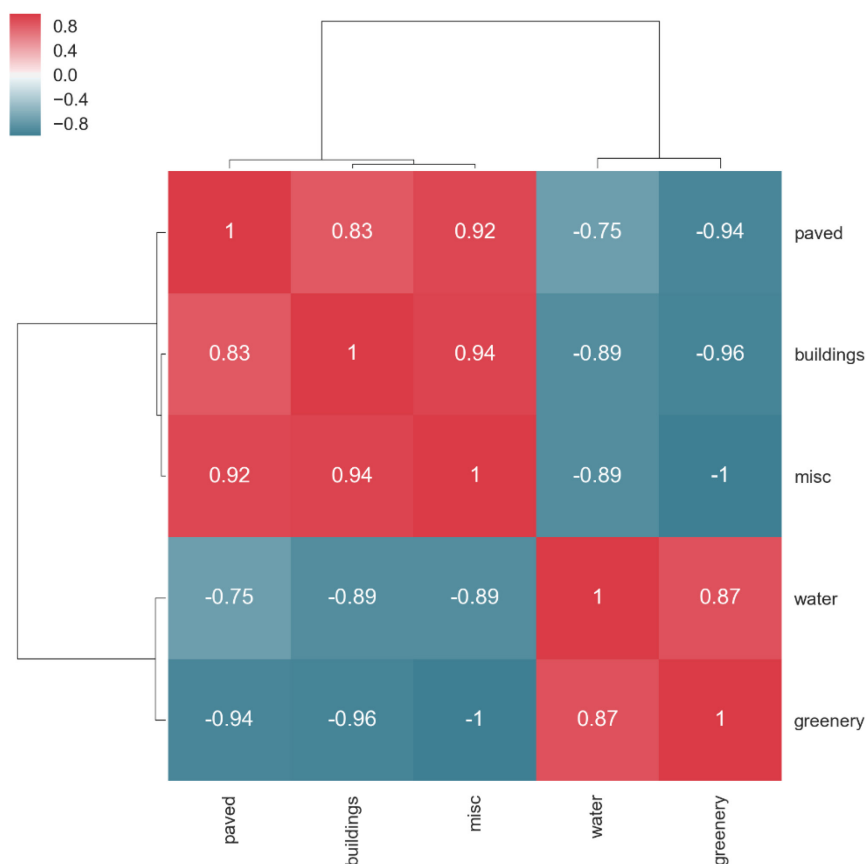
3 Data-analyse

3.1 Correlatieanalyse

Met clustering zijn correlatie-matrices opgeknipt in groepen van attributen die sterk gecorreleerd zijn: de gebouwhoogte, het land(oppervlak)gebruik, bodemgeleidbaarheid en UHI-effecten. De datasets zijn opgenomen in Bijlage I.1. Van de gebouwhoogte (gemiddeld over verschillende bovengrondse oppervlaktes) en het gebruik van het landoppervlak worden de clusters getoond in Figuur 3-1 en Figuur 3-2, respectievelijk. Er is een hoge onderlinge correlatie tussen gebouwhoogtes gemiddeld over oppervlakten met straal 10 en 20 m, en oppervlakten met straal 50 en 100 m. Een mogelijke verklaring voor deze clustering is dat gebouwen binnen een straal van 20 m vaak aaneengeschaald en van een zelfde type en hoogte zijn, terwijl er op een straal van 50 meter of groter de gemiddelde gebouwhoogte meer kan variëren door straten en andere type bebouwing.



FIGUUR 3-1: GEMIDDELDE GEBOUWHOOGTE BEPAALD OVER EEN VERSCHILLENDE STRAAL T.O.V. DE WATERMETER



FIGUUR 3-2: CORRELATIEMATRIX EN CLUSTERS VAN VERHARDE EN ONVERHARDE EIGENSCHAPPEN.

De clusteranalyses van de correlatiematrix voor de geleidingscoëfficiënt van de bodem en de UHIE worden niet getoond, de geleidingscoëfficiënten en de dag-UHIE's blijken per locatie niet, of slechts enkele promille van elkaar te verschillen (met als resultaat dat alles sterk gecorreleerd is). Zie de data in Tabel 9 en Tabel 10 in Bijlage I. Dit hangt samen met de zeer uniforme grondsoortverdeling in Veldhoven (bijna overal bebouwing, Figuur 2-3) en de grove rasterindeling voor het UHIE (Figuur 2-5) ten opzichte van de oppervlaktes waarover gemiddeld wordt.

3.2 Regressieanalyse

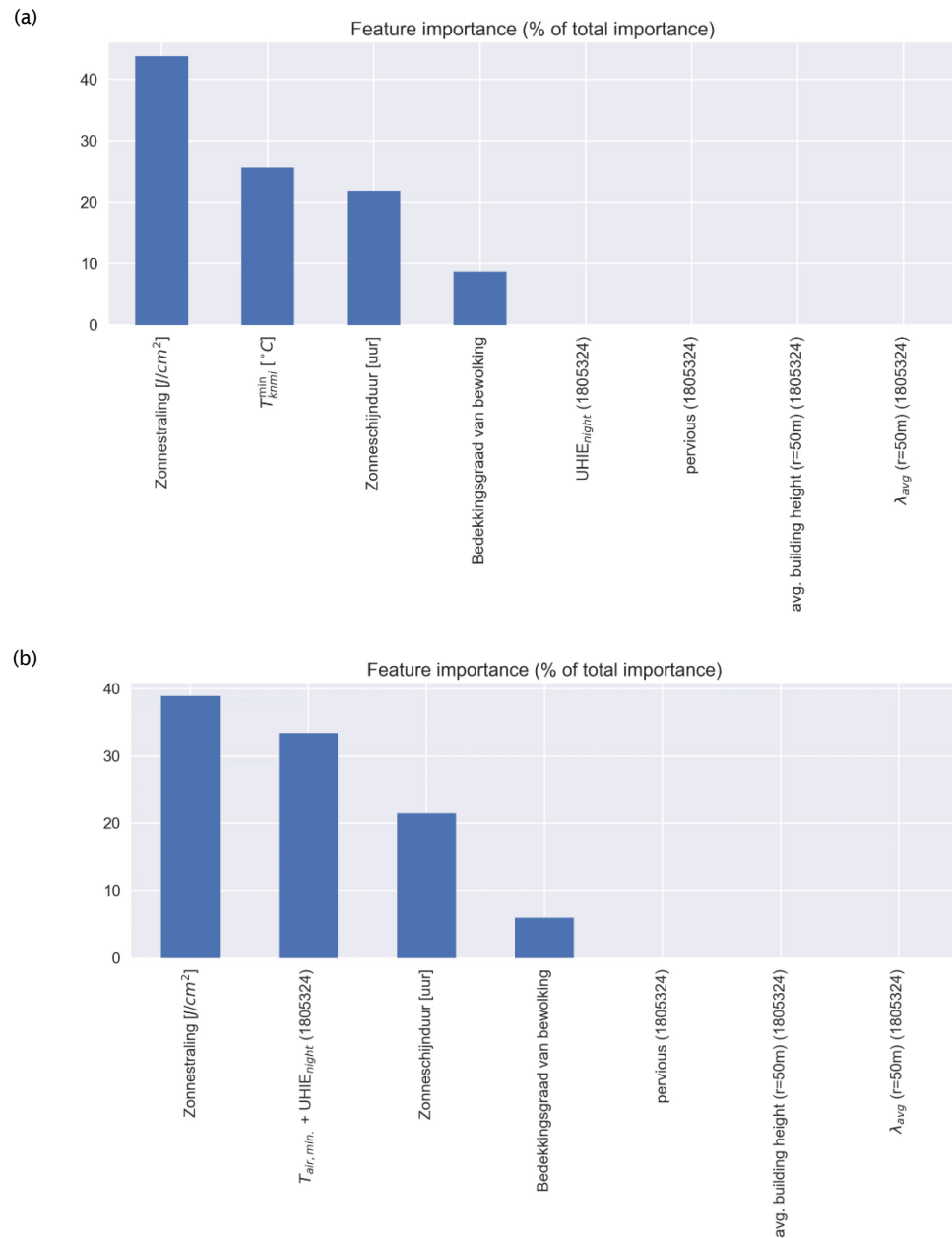
De resultaten van de regressieanalyses RL en RS worden getoond door:

- de mate van belangrijkheid van een attribuut die de machinelearning-procedure procentueel (ten opzichte van de totale bijdrage van alle attributen) heeft bepaald, uit te zetten voor de dataset met en zonder correctie op de UHIE, en data gefilterd op lage volumestroom (Figuur 3-3);
- de prestaties van de regressies te rapporteren aan de hand van:
 - de residuen die bepaald zijn bij één watermeter (1805324) voor ongecorrigeerde en UHI-gecorrigeerde buitentemperatuur. Resultaten voor overige sensorlocaties blijken vergelijkbaar. Deze resultaten zijn weergegeven in Figuur 3-4;
 - een tabel met determinatiecoëfficiënt (R^2) en MSE-waarden per regressie (Tabel 3).
 - een tabel waarin de prestaties van de verschillende regressies op significantie worden getoetst (Tabel 4).

3.2.1 Lokale effecten (RL)

Invloed van attributen

Figuur 3-4 toont de belangrijkheid van attributen aan voor de RL-analyses RL-f en RL-u-f. voor sensorlocatie 1805324. De figuur is op grote lijnen representatief voor alle sensorlocaties, Merk op dat aan alleen de tijdreeks-attributen (zonnestraling, temperatuur, zonneschijnduur en bedekkingsgraad van bewolking) gewicht wordt toegekend in de regressie. De statische attributen (UHI gedurende de nacht: UHI_{night} , onverhard: pervious, gebouwhoogte: building height en bodemgeleidbaarheid: λ_{avg}) hebben zoals verwacht geen substantiële invloed op het (tijdsvariërende) patroon van de watertemperatuur. Verder is het opvallend dat het belang van buitenluchttemperatuur groter wordt als rekening wordt gehouden met het lokale UHI-effect.



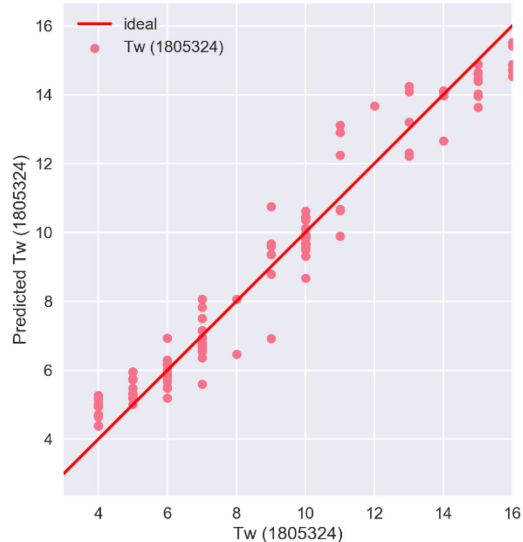
FIGUUR 3-3: BELANGRIJKHEID VAN DATA-ATTRIBUUT UITGEDRUKT IN PROCENTEN VOOR (a) DE DATASET MET ONGECORRIGEERDE BUITENTEMPERatuur EN (b) DE UHIE-GEcorrIGEERDE BUITENTEMPERatuur.

Prestaties regressie

Figuur 3-4 laat zien dat de residuen dichter op de ideale predictielijn (daar waar de predicties exact overeenkomen met de waarnemingen) liggen bij een UHIE-gecorrigeerde buitentemperatuur dan wanneer de buitentemperatuur niet gecorrigeerd is voor lokale omstandigheden. Het verschil in de optelsom van residuen wordt duidelijk uit de MSE-waarden, zie daarvoor Tabel 3. De resultaten suggereren een invloed van het UHIE op de drinkwatertemperatuur, aangezien de temperatuurvariatie per sensorlocatie beter wordt

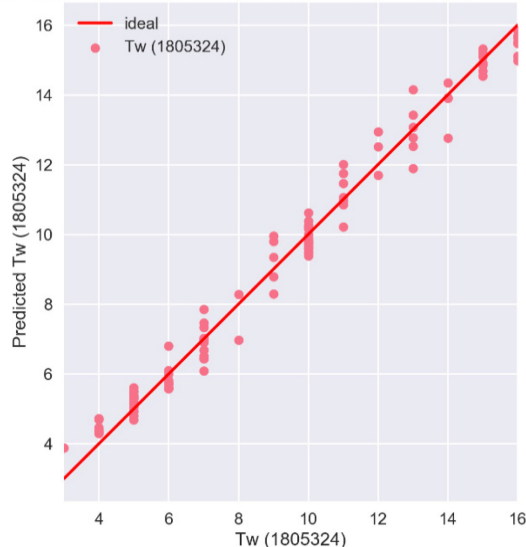
voorspeld als het UHIE wordt meegenomen in de analyse. De richting (positieve of negatieve correlatie) en magnitude van het UHIE op de drinkwatertemperatuur vergt nader onderzoek. De databasebaseerde predictie met Gradient Boosting Regression Trees biedt functionaliteit om deze aspecten kwantitatief te bepalen (hier niet getoond).

Predicted versus measured, data fit quality: $R^2 = 0.983$, $MSE = 0.227$



(a)

Predicted versus measured, data fit quality: $R^2 = 0.987$, $MSE = 0.169$



(b)

FIGUUR 3-4: SPREIDING VAN REGRESSIERESIDUEN VOOR DE REGRESSIE MET ONGECORRIGEERDE BUITENTEMPERatuur (a) EN DE UHIE-GEcorrIGEERDE BUITENTEMPERatuur (b).

TABEL 3: PRESTATIE VAN GBRT-REGRESSIES IN DETERMINATIECOëFFICIëNT EN MSE.

locatie	RL		RL-u		RL-f		RL-f-u		N datapunten
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	
1805324	0.983	0.227	0.987	0.169	0.986	0.192	0.990	0.140	114
1797822	0.982	0.349	0.984	0.308	0.984	0.313	0.987	0.261	129
1811396	0.984	0.282	0.987	0.227	0.985	0.265	0.987	0.224	126
1813969	0.976	0.686	0.979	0.603	0.988	0.277	0.990	0.233	102
1815325	0.983	0.247	0.986	0.203	0.983	0.247	0.986	0.203	135
1815970	0.986	0.261	0.987	0.239	0.986	0.261	0.987	0.239	110
1816158	0.984	0.279	0.987	0.229	0.984	0.279	0.987	0.229	127
2070932	0.986	0.203	0.986	0.198	0.986	0.195	0.991	0.130	85
<i>gemiddeld</i>	<i>0.983</i>	<i>0.317</i>	<i>0.985</i>	<i>0.272</i>	<i>0.985</i>	<i>0.254</i>	<i>0.988</i>	<i>0.207</i>	

De vraag of de regressies zonder UHI en zonder de drempelwaarde op de volumestroom (RL) verschillen met de regressies waarbij wél rekening is gehouden met de UHI en de volumestroom-drempelwaarde is onderzocht met een statistische toets. Hiervoor is de Mann-Whitney-U-test gebruikt. Met deze test wordt het significantieniveau (p-waarde) en een U-statistiek bepaald voor de test dat een groep1 verschilt van groep2 op basis van een rangordesckikking. Voor het huidige aantal waarnemingen ($n=8$ sensorlocaties) blijkt dat de

drempelwaarde voor de U-statistiek $U_{\text{kritisch}} = 15.0$ is: een waarde lager dan U_{kritisch} betekent dat de H_0 -hypothese (groep2 komt uit dezelfde populatie als groep1) kan worden verworpen. Resultaten van de toets voor de MSE en R^2 zijn opgenomen in Tabel 3-2. Uit deze tabel zien we dat op basis van:

- de MSE, er geen significant verschillen tussen de RL en RL-u of RL-f regressies zijn. Dit is te verklaren uit de variaties van MSE tussen de verschillende watermeters: de variaties onderling zijn dusdanig groot en verschillend tussen de regressies, dat de rangordeschikking verschillend uitpakt en er geen significant onderscheid kan worden gemaakt tussen de regressies.
- de determinatiecoëfficiënt, er een significante verbetering te zien is voor de RL-u en de RL-f regressies ten opzichte van RL.

De regressie-analyses zijn uitgevoerd voor afzonderlijke watermeters op basis van omgevingsfactoren en trainingsdata. Op locaties waar metingen ontbreken zijn geen regressies uitgevoerd omdat daar de ondergrens van hoeveelheid data was bereikt. Voldoende data zijn wel nodig bij het toewerken naar een dynamische warmtekaart. Hiervoor moet de invloed van geografische variatie in omgevingsfactoren worden bepaald: bijvoorbeeld, hoe correleert de drinkwatertemperatuur met een hitte-eiland effect van 5°C in een omgeving met veel water en hoge gebouwen. Om dergelijke vragen te beantwoorden is de plaatsing van meer sensoren nodig in een gebied met veel variatie in de omgevingsfactoren. Naast de huidige aanpak, is het dan ook interessant om machinelearningtechnieken toe te passen om de invloeden van geografische variatie in omgevingsfactoren te onderzoeken.

TABEL 4: STATISTISCHE TOETSING VAN VERSCHILLEN IN REGRESSIERESULTATEN.

Hypothese	W	p	$U < U_{\text{kritisch}}$, $p < 0.05$ (5%)	Conclusie
MSE (RL) > MSE (RL-u)	19.0	0.086	Nee	Geen significant verschil in MSE met data met/zonder UHIE en met/zonder volumestroom
MSE (RL) > MSE (RL-f)	39.5	0.215	Nee	
R^2 (RL-u) > R^2 (RL)	10.0	0.010	Ja	Significant betere prestatie(s) wanneer R^2 wordt bepaald indien UHIE wordt meegenomen of op volumestroom wordt gefilterd
R^2 (RL-f) > R^2 (RL)	14.5	0.032	Ja	

3.2.2 Stroompadeffecten

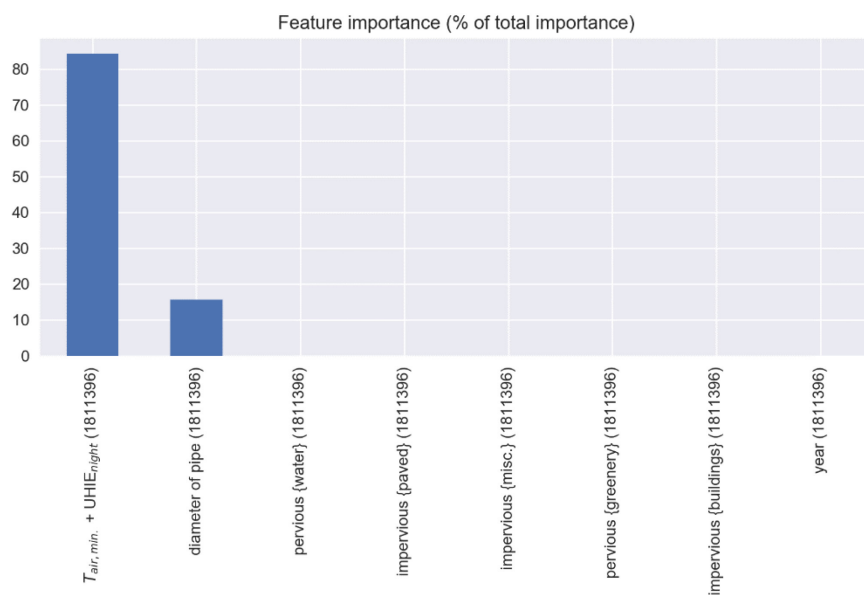
Invloed van attributen

De invloed van elk attribuut is per sensorlocatie anders. We zien 4 verschillende patronen:

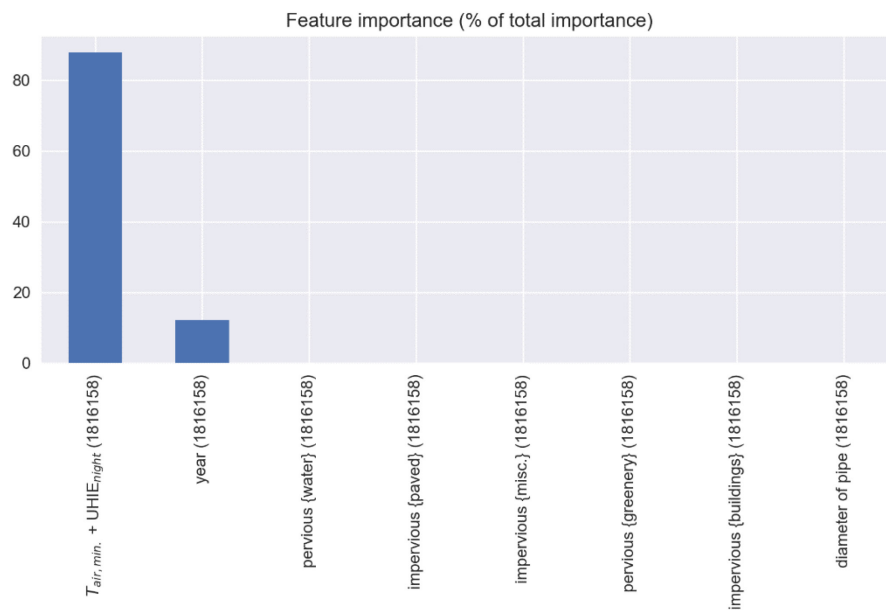
1. Alleen buitentemperatuur-gedomineerd, waarbij alle overige attributen minder dan 3% invloed hebben. Dit is het geval voor sensor 1805324 (alleen deze is getoond, zie Figuur 3-7), 2070932, 1797822 en 1815970.
2. Naast de buitentemperatuur, ook de invloed van de diameter van de leiding. Dit is het geval bij sensor 1811396 (Figuur 3-5).
3. Naast de buitentemperatuur, ook de invloed van het aanlegjaar van de leidingen. Dit is het geval voor sensor 1816158 (Figuur 3-6).
4. Naast de buitentemperatuur, ook een kleine invloed van het aanlegjaar én de diameter (allebei rond de 5%). Dit is het geval voor sensor 1815325 (Figuur 3-8).

Voor alle gevallen geldt dat de (UHI-gecorrigeerde) buitentemperatuur een heel groot aandeel heeft in het resulterende regressiemodel, en dat de attributen die relateren aan gebruik van landoppervlak een verwaarloosbare invloed. Bij de (stroompad)berekening van 1813969 ontbraken er attributen, waardoor alleen de buitentemperatuur werd meegenomen in de regressie.

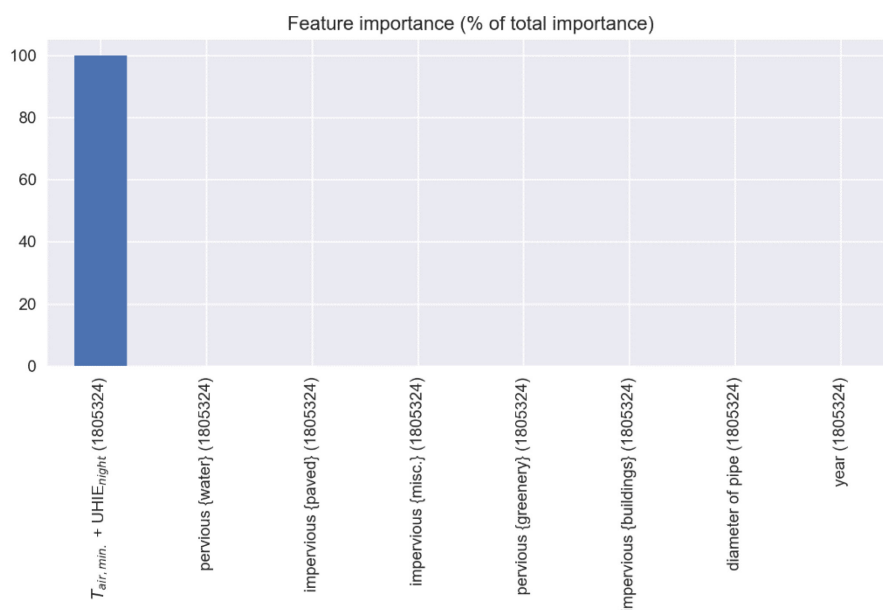
Dat de buitentemperatuur het grootste aandeel heeft in het regressiemodel is verklaarbaar: dit is de enige parameter met een duidelijke variatie in de tijd en een aannemelijke correlatie met de watertemperatuur (via de bodemtemperatuur). Het kleine aandeel van leidingdiameter en aanlegjaar ligt minder voor de hand. Mogelijk weerspiegelt dit een indirect effect: tijdens verbruikspieken zijn – voor een vaste verblijftijd – de stroompaden waarover wordt gemiddeld langer, wat vaak gepaard zal gaan met grotere gemiddelde diameters (vanwege afnemende diameters van bron tot tap) en variatie in het aanlegjaar. Aangezien de verbruikspieken plaatsvinden op specifieke momenten tijdens de dag is ook een correlatie met de temperatuur te verwachten. Of het inderdaad om indirecte invloeden gaat en wat de richting en amplitude is van deze invloeden zou in een vervolgonderzoek kunnen worden onderzocht. De gebruikte machinelearning-technieken bevat functionaliteit om deze aspecten te bepalen (hier niet getoond).



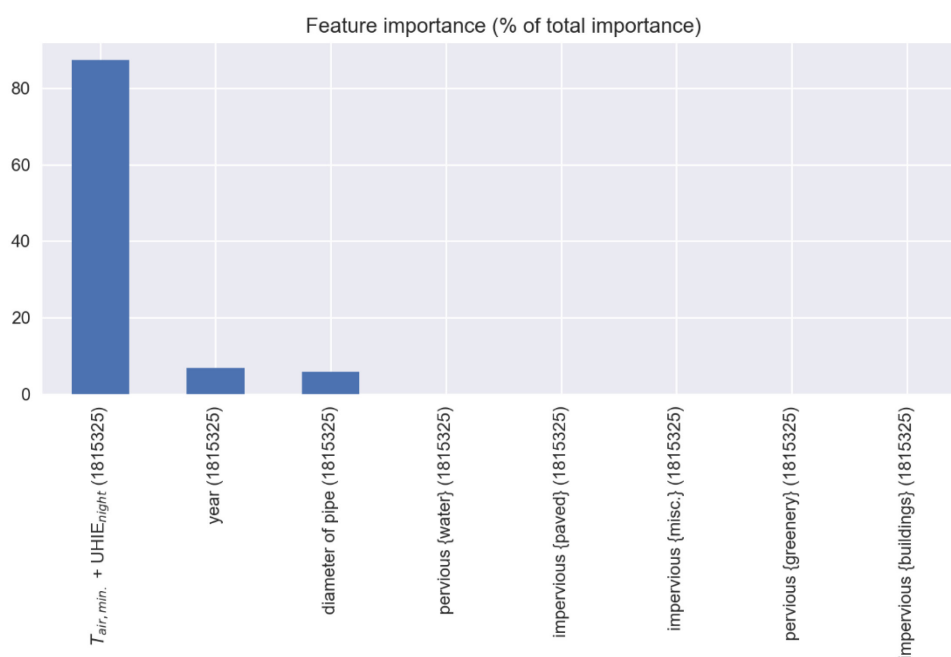
FIGUUR 3-5: INVLOED VAN ATTRIBUTEN VOOR DE RS-ANALYSE VAN WATERMETER 1811396.



FIGUUR 3-6: INVLOED VAN ATTRIBUTEN VOOR DE RS-ANALYSE VAN WATERMETER 1816158.



FIGUUR 3-7: INVLOED VAN ATTRIBUTEN VOOR DE RS-ANALYSE VAN WATERMETER 1805324.



FIGUUR 3-8: INVLOED VAN ATTRIBUTEN VOOR DE RS-ANALYSE VAN WATERMETER 1815325.

Prestatie regressie

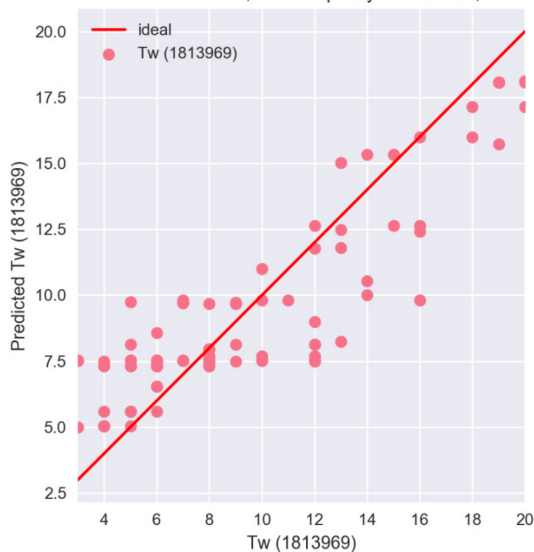
De middeling met verschillende tijdduur blijkt niet of nauwelijks effect te hebben op de resulterende datamatrix en regressie, zie Tabel 5. De prestatie-indicatoren van de regressie pakken slechter uit dan die van de RL-analyse. Dat de regressie met stroompaddata minder goed presteert is goed zichtbaar in een grafiek met voorspelde versus gemeten waarden (Figuur 3-9, willekeurige sensor). De resultaten zijn echter niet één op één te vergelijken met RL omdat het aantal en het type attribuut verschillen. Met meerdere watermeters en lokale

metingen wordt het perspectief om het UHI-effect, en de effecten daarvan op de watertemperatuur, te bepalen en te voorspellen ook echt haalbaar.

TABEL 5: PRESTATIE-INDICATOREN VAN STROOMPADREGRESSIE (RS) EN AANTAL ATTRIBUTEN.

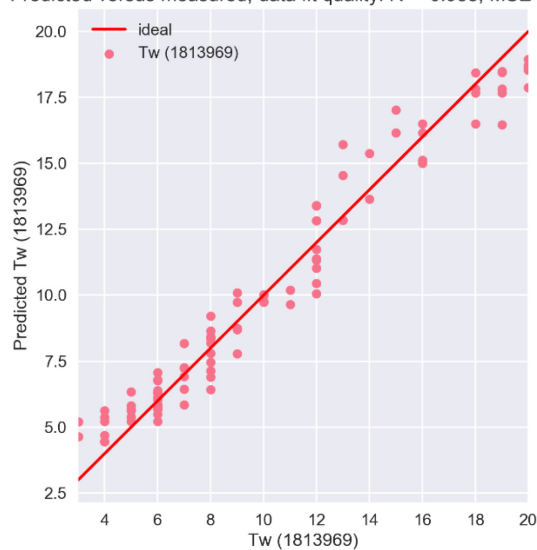
watermeter	0,5 uur		1 uur		2 uur		4 uur		8 uur		N data-punten	N attributen
	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE		
1805324	0.90	1.30	0.90	1.30	0.90	1.30	0.90	1.30	0.90	1.31	115	8
1797822	0.91	1.72	0.91	1.72	0.91	1.72	0.91	1.72	0.91	1.80	131	8
1811396	0.96	0.71	0.96	0.71	0.96	0.71	0.96	0.71	0.96	0.71	128	8
1813969	0.92	2.00	0.92	2.00	0.92	2.00	0.92	2.00	0.92	2.00	104	1
1815325	0.90	1.45	0.95	0.66	0.96	0.63	0.96	0.63	0.96	0.63	136	8
1815970	0.90	1.75	0.90	1.75	0.90	1.75	0.90	1.75	0.89	1.92	110	8
1816158	0.96	0.60	0.96	0.60	0.96	0.60	0.96	0.60	0.96	0.61	129	8
2070932	0.92	1.07	0.92	1.07	0.92	1.20	0.92	1.20	0.97	0.36	85	8

Predicted versus measured, data fit quality: $R^2 = 0.915$, $MSE = 1.999$



(a)

Predicted versus measured, data fit quality: $R^2 = 0.988$, $MSE = 0.277$



(b)

FIGUUR 3-9 VOORSPELDE WATERTEMPERATUUR VERSUS GEMETEN WATERTEMPERATUUR VOOR WATERMETER 1813969 BIJ DE STROOMPAD- (a) EN LOKAAL-EFFECT-ANALYSE (b).

3.3 Synthese

De resultaten van de regressie-analyses laten met een beperkt aantal watermeters en een periode van slechts 6 maanden (van 1 januari 2017 tot en met 1 juli 2017), zien dat de watertemperatuur is te voorspellen met metingen uit de omgeving. Omgevingsdata-bronnen die het weerpatroon (temperatuur en zonnestraling) volgen, blijken een significante invloed te hebben op de prestaties van de regressie. Bij de lokale-effectanalyse blijken alle weerfactoren van groot belang voor het model. De gedachte achter de stroompad-analyse

nodigt uit om de effecten van de afgelegde route van het water nader te onderzoeken voor een grotere set aan watermeters en omgevingsdata: is bijvoorbeeld de temperatuur van het water gedurende het stroompad ook meer afhankelijk van zonnestraling dan van leidingmaterialen? En bij een grotere set aan watermeters, zien we dan factoren die lokaal zijn bepaald, zoals antropogene bronnen, terug in de regressieresultaten? Een uitgebreidere set aan watermeters over een divers gebied (ruraal, stedelijk, verschillende bebouwing) eventueel aangevuld met metingen van de temperatuur op verschillende plekken bieden een goed startpunt om deze vragen verder te onderzoeken met machine-learningtechnieken en deze te complementeren en analyseren met bodem- en watertemperatuurberekeningen.

4 Perspectief voor een alarmsysteem voor hoge temperaturen

4.1 Overwegingen bij toepassing

In het toewerken naar het monitoren van hotspots monitoren zijn verschillende benaderingen mogelijk. Bij de keuze over toekomstige installatie van sensoren moet goed worden nagedacht over het doel of doelen van de temperatuurmetingen, de verwachte opbrengsten en investeringen en de mogelijkheden voor handelingsperspectieven. De te volgen aanpak zou hierop moeten worden afgestemd. Verschillende benaderingen en bijbehorende kenmerken zijn weergegeven in Tabel 6 en worden hieronder nader besproken:

- *Ontwikkeling van een temperatuurkaart op basis van metingen met een volledige dekking.* Verwezenlijk een volledige dekking door bij elke klant een temperatuurmeter te plaatsen. Een voordeel is dat zo de meest nauwkeurige informatie beschikbaar komt en geen informatie van de omgeving meer nodig is. Een hoge resolutie van informatie gewenst, aangezien lokale oorzaken de temperatuur kunnen beïnvloeden. Een nadeel is de grote investering die samengaat met de plaatsing van sensoren. Ook ontstaat op deze manier geen inzicht in de bepalende factoren.
- *Datagedreven aanpak.* Bepaal de temperatuur met metingen zonder volledige dekking in combinatie met omgevingsdata en machinelearning. De methodiek beschreven in dit rapport is geschikt voor uitwerking van deze aanpak.
- *Hybride aanpak.* Koppel de datagedreven aanpak aan een op fysische wetten gebaseerd model. Dit heeft als voordeel dat de relaties tussen omgeving en temperatuur fysisch worden onderbouwd, wat de temperatuurbepalingen zal verbeteren ten opzichte van de datagedreven aanpak.

TABEL 6. OVERZICHT VAN VOOR- EN NADELEN VOOR VERSCHILLENDE BENADERINGEN VOOR HET BEPALEN VAN HOTSPOTS IN HET LEIDINGNET. DE BENADERINGEN ZIJN NADER BESCHREVEN IN DE TEKST. GROENE VELDEN GEVEN AAN WAAR HET EEN VOORDEEL TEN OPZICHTE VAN OVERIGE BENADERINGEN BETREFT.

		Datagedreven (Big Data)		Modelstudie
	Volledige dekking sensoren	Datagedreven aanpak	Hybride aanpak	Modelstudie (o.b.v. fysische wetten) gevalideerd met metingen
Inzicht hotspots (nauwkeurigheid en resolutie)	Optimaal	Afhankelijk van sensordichtheid en interpolatie	Gebaseerd op combinatie modellen & sensoren	Afhankelijk van modelaannames + validatie
Inzicht in bepalende factoren	Nee (tenzij aangevuld met omgevingsdata)	Ja	Ja	Ja, voor zover in het model aanwezig
Inzicht in causale verbanden	Nee	Nee	Ja	Ja
Aantal benodigde sensoren	Bij elke aansluiting	Afhankelijk van gewenst inzicht	Afhankelijk van gewenst inzicht	Ter validatie
Omgevingsdata nodig?	Nee	Ja	Ja	Ja
Scenario's modelleren mogelijk?	Nee	Ja	Ja	Ja

In de hybride aanpak kunnen sensorgegevens en modelberekeningen elkaar aanvullen. Bodemmodellen geven in veel gevallen een goede indicatie van de drinkwatertemperatuur (Blokker and Pieterse-Quirijns 2013; Agudelo-Vera, Blokker et al. 2015; Agudelo-Vera 2017). Maar afwijkingen zijn mogelijk, vanwege gemaakte modelaannames (het niet expliciet beschrijven van de warmte-uitwisseling tussen leiding en ondergrond, of het beschouwen van alleen rechte leidingen in een oneindige halfruimte) en de onvolledige kennis van omgevingsparameters. Zo is bijvoorbeeld meer kennis over antropogene bronnen gewenst (Agudelo-Vera 2017) en sensorgegevens zouden dit hiaat kunnen opvullen zodat antropogene bronnen uit de gegevens zijn af te leiden.

Een keuze in de te volgen aanpak hangt samen met de bedrijfsspecifieke vragen: welke nauwkeurigheid en detailniveau van een temperatuurkaart zijn gewenst.

Een keuze in de te volgen aanpak hangt samen met de bedrijfsspecifieke vragen: welke nauwkeurigheid en detailniveau van een temperatuurkaart zijn gewenst en hoeveel budget is beschikbaar voor investeringen? Dergelijke afwegingen zijn te onderbouwen met een kosten-baten analyse. In een eerder TKI-onderzoek (Van Summeren 2016) zijn de kosten van waterkwaliteitssensoren afgezet tegen de verwachte opbrengsten (gedefinieerd als de detectiekans van waterkwaliteitsafwijkingen in het distributienet. Een soortgelijke systematische afweging kan worden toegepast op het temperatuurvraagstuk.

Doelstellingen die al ter sprake zijn gekomen in de inleiding van dit rapport zijn: voldoen aan vastgestelde maximumwaarden voor de drinkwatertemperatuur, beperken van gezondheidsrisico's waarbij de temperatuur een rol speelt en vergoten van het klantcomfort. Een mogelijk ander doel is om temperatuurmetingen te gebruiken in de berekening van afgeleide waterkwaliteitsparameters volgens het softsensor-principe. Veel van deze doelen zijn lastig te kwantificeren of te vertalen naar financiële baten. Het bepalen van de verwachte opbrengsten is echter een nuttige eerste stap om met een kosten-opbrengsten-relatie keuzes voor sensorplaatsing te onderbouwen.

De rest van dit hoofdstuk gaat in op de vraag of de in dit project gebruikte tools geschikt zijn voor het verwerken van grote hoeveelheden sensorgegevens naar nuttige informatie voor de eindgebruiker. De opgedane ervaring in dit onderzoek wordt daarbij vertaald naar een aantal inzichten rond het gebruik van sensormetingen, machinelearning en visualisatietools.

4.2 Mogelijkheden voor opschaling naar grote hoeveelheden data

Om temperaturen voor een heel leidingnet in kaart te brengen is het nodig om de beschikbare hoeveelheid data uit te breiden. Het uitbreiden van het aantal sensoren zal hieraan bijdragen. Het loggen van alleen de minimum temperatuur op een dag is begrijpelijk vanuit praktisch oogpunt (het verlengen van de levensduur van batterijen), maar het is te verwachten dat er meer informatie is te destilleren indien er meerdere keren er dag wordt gelogd.

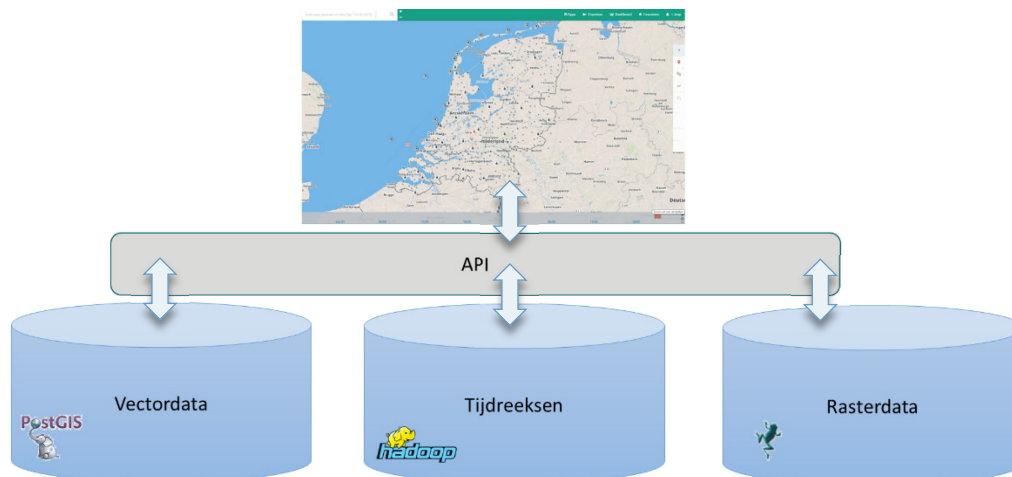
De onderzochte machinelearningtechniek Gradient Boosting regressie is geschikt voor de systematische analyse van grote datasets. Gradient Boosting is geschikt om temperaturen van het water aan de tap te voorspellen en te bepalen hoeveel elke omgevingsfactor bijdraagt aan de voorspellingen (Hoofdstuk 3). Ook is de methodiek geschikt om onzekerheden van de voorspellingen te kwantificeren en de sterkte en richting (positief of negatief) van de afhankelijkheden van de factoren te bepalen. Dit kan in mogelijk vervolgonderzoek worden gebruikt om een datagedreven model op te stellen waarbij ook de temperaturen zijn te bepalen voor locaties waarvoor metingen ontbreken.

Naast het verzamelen, voorbereiden en analyseren van data is een belangrijke stap het ontsluiten van nuttige informatie voor het maken van operationele beslissingen door de eindgebruiker. De mogelijkheden hiervoor voor visualisatie met Lizard zijn onderzocht en worden beschreven in de rest van dit hoofdstuk.

4.3 Visualisatie met Lizard

Binnen het onderzoek is veel gebruik gemaakt van datasets die werden geanalyseerd met specialistische software of eigen (python)-scripts. Dit is voor onderzoekers een geschikte methode om meer inzicht te krijgen. Voor de terugkoppeling naar de andere belanghebbenden, zoals bijvoorbeeld Brabant Water, is meer visualisatie en toegankelijkheid gewenst.

Voor het project is gekozen om Lizard te gebruiken als data visualisatie-tool. Lizard is een online platform gericht op het snel en stabiel ontsluiten van gegevens. Daarbij ligt een sterke focus op informatie dat betrekking heeft op water. Het platform bestaat uit drie onderdelen: een backend, een API en een viewer (zie Figuur 4-1). De backend is de dataopslag, waar data gestructureerd wordt opgeslagen zodanig dat het gecombineerd kan worden. De dataopslag is ontwikkeld voor grote hoeveelheden data die real time binnen kunnen komen en is daarmee geschikt voor operationeel gebruik. De API is de Application Programming Interface die zorgt voor de ontsluiting van de database op een eenduidige manier, zodat ook andere programma's er mee kunnen communiceren. De API wordt ook gebruikt om data in de backend op te slaan. De warmtemetingen voor dit project zijn ook op deze manier in de database gezet voor verdere verwerking. De viewer van Lizard is een webviewer waarin tijd en ruimte centraal staan om de data te bekijken. Hierdoor wordt het mogelijk om ruimtelijke beelden te combineren, zoals neerslag-informatie, waterstandsmetingen, drinkwaterleidingen, waterverbruik, enz. Naast de viewer van Lizard is het mogelijk om andere viewers te koppelen aan de Lizard backend via de API. Ook kunnen andere pakketten zoals QGIS en ArcGIS gekoppeld worden om de data op te halen, te bewerken en terug te sturen.



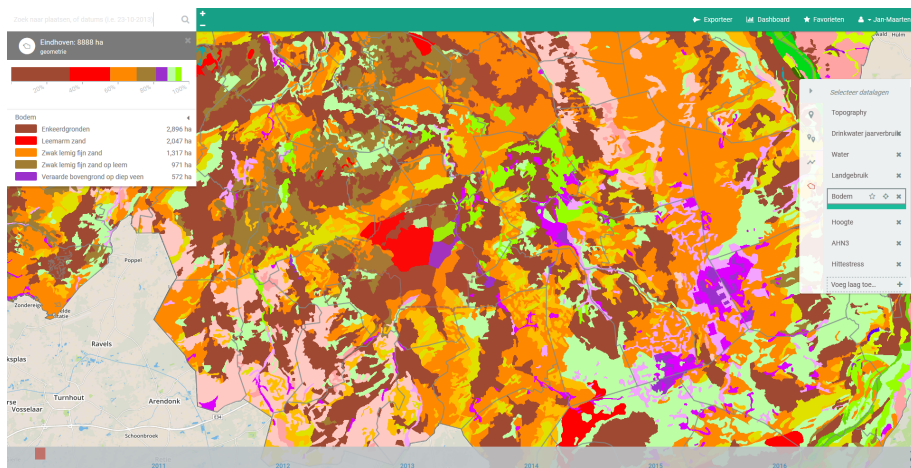
FIGUUR 4-1: GLOBALE LIZARD ARCHITECTUUR: DE OPDELING IN VERSCHILLENDE COMPONENTEN

Lizard heeft een architectuur die gericht is op het snel visualiseren van de gegevens. Om dit te bereiken is per datatype een eigen database ingericht of ontwikkeld. Via één API is deze data te bevragen door de Lizard-viewer of andere viewers. De gegevens in de database zijn opgedeeld in drie categorieën: vectordata (data over een punt, lijn of vlak), tijdreeksen en rasterdata (data voor een grid, waarbij elke cel een waarde heeft). Alle waterleidingen,

rioleringsleidingen en meetstations (locatie en attributen) zijn opgeslagen in de vector database, temperatuurmetingen en overige metingen in de tijdreeks database en de verrasterde drinkwater verbruikgegevens, hoogteinformatie en neerslaginformatie in de rasterdatabase.

Met Lizard is de gebruiker vooral in staat om informatie bij elkaar te brengen en daardoor de gegevens te duiden. Het is daarbij mogelijk om zowel ruimtelijk als in de tijd in en uit te zoomen. Zo wordt de data snel toegankelijk op het juiste gewenste niveau. De kwaliteit van de informatievoorziening gaat daarbij omhoog. In Lizard worden metingen, modeluitkomsten, statische data (leidingnet) en praktijkervaring bij elkaar gebracht en inzichtelijk gemaakt voor iedereen. De ambitie van Nelen & Schuurmans is om de komende 10 jaar Lizard onderdeel te laten zijn van oplossingen die informatie leveren en ondersteuning bieden aan de belangrijke problemen van de mondiale waterwereld. Hierbij ligt de focus op de problemen op het vlak van klimaatverandering, voedselzekerheid en drinkwaterzekerheid.

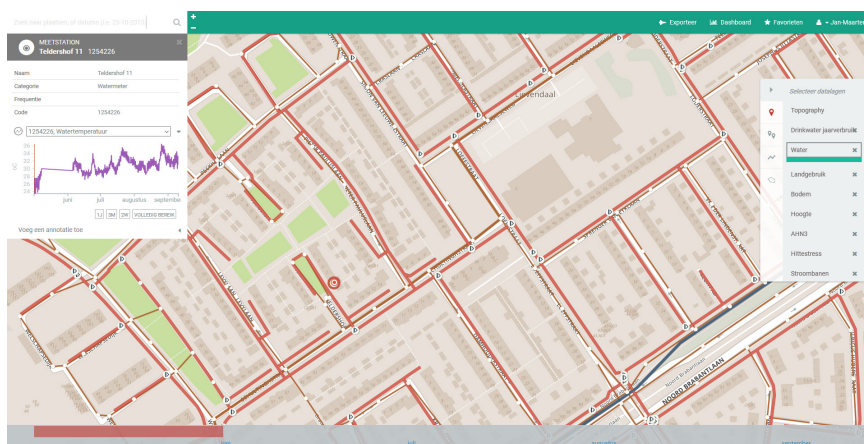
Hieronder staan enkele voorbeelden van Lizard gegeven met data die relevant is en gebruikt is voor het onderzoek.



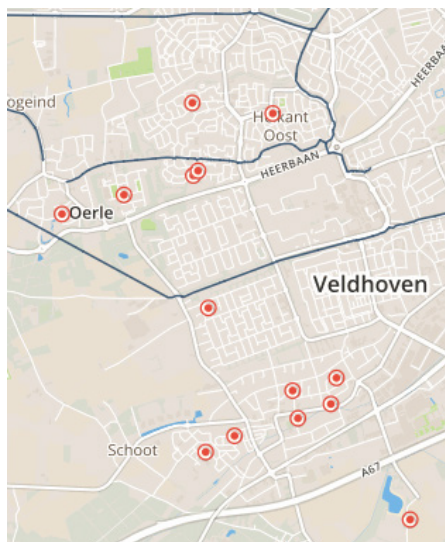
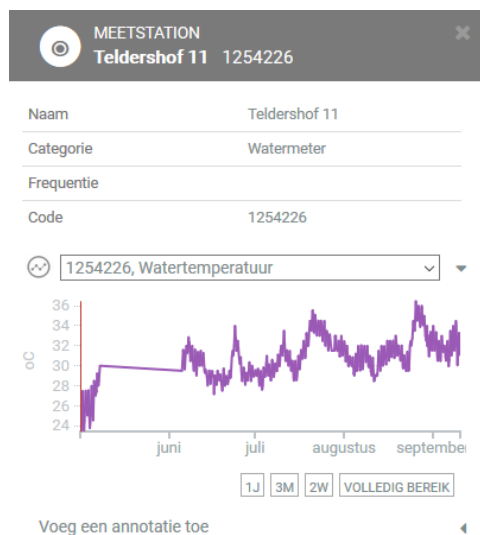
FIGUUR 4-1: BODEMKAART VISUALISATIE IN LIZARD VOOR DE GEMEENTE EINDHOVEN



FIGUUR 4-3: VISUALISATIE HITTESTRESS IN LIZARD



FIGUUR 4-4: VISUALISATIE MEETSTATIONS VAN WATERTEMPERATUUR IN LIZARD



FIGUUR 4-5: LINKS: WATERTEMPERATUUR VAN EEN MEETSTATION IN LIZARD. RECHTS: DE MEETSTATIONS RONDOM VELDHOVEN.

4.4 Resultaten en perspectief

Door het project hebben we meer inzicht gekregen in de beschikbaarheid en uitwisselbaarheid van gegevens om zogenaamde 'hotspots' te ontdekken. Niet alleen voor goed onderzoek, maar zeker voor operationele monitoring is een groter netwerk van sensoren nodig. Daarnaast zal goed gekeken moeten worden hoe deze data het beste opgeslagen en uitgewisseld kan worden. De leveranciers van de temperatuurmetingen heeft er voor gekozen om per tijdvakje de maximale waarde op te slaan, waardoor er minder frequent een geaggregeerde waarde beschikbaar is. Het zou goed zijn als het mogelijk wordt om hiernaast ook de geregistreerde waarden op te kunnen vragen.

De grootste uitdaging is te zorgen dat de data gedeeld wordt en de applicatie geïmplementeerd wordt in de organisatie. Daarmee wordt bedoeld dat onderzoek vaak stopt bij het eindresultaat van de analyse, maar dat de meerwaarde pas echt gecreëerd wordt als het in een werkproces meegenomen wordt. Zo kan voorkomen worden dat het een eenmalige exercitie is en wordt op een natuurlijke wijze afgedwongen de data en kennis regelmatig te gebruiken. Nelen & Schuurmans ziet in Lizard een product dat als centrale datahub geschikt is om data uit bronsystemen in te lezen, via de API beschikbaar te stellen aan onderzoekers en andere partijen en te visualiseren via de Lizard-viewer. Dat heeft voordelen voor (TKI-) onderzoeken, maar ook het dagelijks werkproces. Als vervolgstap is specifiek voor dit project nodig dat de datalevering verbeterd wordt: er zijn meer databronnen nodig en de databronnen moeten 'ruwere' data leveren. Hiermee wordt bedoeld dat met de aggregatie naar een maximum temperatuur teveel informatie verloren gaat. Als breder gekeken wordt, is het nodig dat er een implementatieplan komt, waarin beschreven staat voor welke werkprocessen inzicht in de watertemperatuur waarde heeft, welke eisen dat aan de data stelt en welke visualisatievormen daarvoor gewenst zijn. In het nog lopende TKI-onderzoek "Big Data Epidemiology" wordt de uitkomst bevestigd dat er behoefte is aan een implementatieplan.

Er ligt een kans om grote stappen te nemen in de datauitwisseling. Vanuit de waterschappen zijn daar al veel initiatieven voor opgestart, waarbij een samenbindend initiatief de Digital Delta heet. Binnen dit programma worden standaarden opgesteld voor de manier vanuit uitwisseling, niet de semantische standaarden. Digital Delta is een samenwerking tussen waterschappen, rijkswaterstaat en adviesbureaus. Het is op zoek naar een bredere blik op de waterwereld. De focus ligt momenteel op waterkwantiteit, waterkwaliteit en biologische data. Voor de drinkwaterwereld zouden hier kansen liggen om aansluiting te vinden. Het wordt dan ook aanbevolen om deze mogelijkheden verder te verkennen.

5 Conclusies en aanbevelingen

5.1 Conclusies

In dit onderzoek zijn machinelearning-technieken ingezet om verbanden te herkennen tussen de minimum dagtemperatuur gemeten bij binneninstallaties van 9 huishoudens in Veldhoven en gegevens van het weer, het leidingnet en de stedelijke omgeving. De belangrijkste conclusies van dit onderzoek zijn:

- Voor 9 onderzochte huishoudens is, na training van het algoritme, met omgevingsfactoren de locatie-specifieke minimum drinkwatertemperatuur op een dag te voorspellen met een nauwkeurigheid van 0,2°C. Van de onderzochte omgevingsfactoren zijn de zonnestraling, buitentemperatuur, zonneschijnduur en bewolgingsgraad de belangrijkste bepalende factoren. Deze vaststelling sluit aan bij eerder onderzoek waarin aannemelijk is gemaakt dat het weer een grote invloed heeft op de temperatuur van het drinkwater, via de bodem (Blokker and Pieterse-Quirijns 2013). Onderzochte statische factoren (bodemsoort, gebouwhoogte, landgebruik) dragen nauwelijks bij aan de voorspelling van de (locatie-specifieke) dynamiek van de minimale drinkwatertemperatuur. De voorspellingen worden nauwkeuriger als (i) de luchttemperatuur maandelijks wordt gecorrigeerd voor het warmte-eiland effect en (ii) als perioden van laag verbruik worden weggelaten uit de analyse, omdat de watermeter dan de omgevingstemperatuur aanneemt, in plaats van de temperatuur in het leidingnet.
- Bij middeling over een stroompad is er geen sterke afhankelijkheid van de onderzochte factoren: leidingmateriaal, -diameter, -aanlegjaar, bodemgebruik en hitte-eilandeffect. Ook de invloed van de padlengte waarover de parameters worden gemiddeld is niet substantieel voor de meeste locaties. De resultaten zijn niet één-op-één te vergelijken met de resultaten met lokale middeling.
- Het Gradient Boosting algoritme dat in dit onderzoek is toegepast, is geschikt voor het beoogde doel. Het is mogelijk om drinkwatertemperaturen te voorspellen en de onzekerheid in de voorspellingen te bepalen. Ook kan kwantitatief worden gemaakt in welke mate factoren bijdragen aan de voorspelling. De sterkte en richting van de afhankelijkheden van de afzonderlijke factoren zijn te bepalen (hoewel deze zaken in het huidige onderzoek niet zijn onderzocht).
- Gradient Boosting is geschikt voor opschaling naar grote hoeveelheden data, waarmee is te verwachten dat de kwaliteit van de voorspelling zal toenemen.
- Bij het opschalen van grote hoeveelheden data is een platform als Lizard geschikt om in een hoge kwaliteit van informatievoorziening te voorzien door metingen, modeluitkomsten, gegevens van het leidingnet en praktijkervaring bij elkaar te brengen en snel te visualiseren zijn de gegevens beter te duiden.

5.2 Aanbevelingen voor vervolgonderzoek en implementatie

Op basis van de ervaringen in het voorliggende onderzoek worden de volgende aanbevelingen gedaan voor vervolgstappen:

- De ontwikkeling van een temperatuurkaart van het leidingnet moet worden afgestemd op het gewenste detailniveau en onnauwkeurigheid. Daarbij zijn verschillende ontwikkelrichtingen tegen elkaar af te zetten: (i) ontwikkeling op basis van sensormetingen die het leidingnet volledig dekken, (ii) de datagedreven aanpak waarbij

sensorgegevens en omgevingsdata met machinelearning worden geanalyseerd, (iii) de hybride aanpak waarbij de datagedreven aanpak en een bodemmodel op basis van fysische wetten elkaar versterken. Een voordeel van de hybride aanpak is dat deze niet volledig leunt op meetgegevens (en de bijbehorende kosten van sensoren) of op de aannames van fysische wetten.

- Omdat het plaatsen van sensoren kosten met zich meebrengt, moet goed worden nagedacht over het doel, de verwachte opbrengsten en kosten van sensorplaatsing. Een eerder TKI-onderzoek waarbij kosten van sensoren zijn uitgezet tegen de verwachte opbrengsten (Van Summeren 2016) kan daarbij als voorbeeld dienen. De kosten-opbrengsten-balans zal verbeteren door sensoren in te zetten voor verschillende doeleinden, zoals klantcomfort, reductie van mogelijke gezondheidsrisico's en voldoen aan de vastgelegde maximum waarde voor de temperatuur.
- Om de temperatuur aan de tap voor een gebied in detail te monitoren, is het nodig het sensornetwerk uit te breiden. Een uitgebreidere set aan watermeters over een divers gebied (ruraal, stedelijk, verschillende bebouwing) en meerdere jaren, eventueel aangevuld met metingen van de temperatuur op verschillende plekken bieden een goed startpunt om een aantal vervolgvragen te onderzoeken met machinelearning-technieken en deze te complementeren en analyseren met bodem- en watertemperatuurberekeningen. Specifieke vervolgvragen zijn: Met welke nauwkeurigheid is een dynamische kaart van de drinkwatertemperatuur te maken? Zijn voorspellingen te doen op locaties waar geen temperatuurmetingen voorhanden zijn? Welke sensordichtheid en meetperiode zijn hiervoor nodig?
- In dit project is gebleken dat het verzamelen, overdragen en duiden van data veel tijd en inspanning vergt ten opzichte van de daadwerkelijke regressie. Ook is vakspecialistische kennis noodzakelijk om data op waarde te schatten. Gestroomlijnd data-management kan de inspanningen van de voorbewerking en duiding beperken. Om tot een dergelijke stroomlijning te komen is het belangrijk om de dataset zoveel mogelijk uniform te maken, consistente data-coderingen en eenheden te gebruiken en de metingen eenduidig, transparant en met meta-data vast te leggen. De stroomlijning is tijdens dit onderzoek sterk verbeterd omdat Brabant Water de sensormetingen uniform vastlegt en makkelijk kan delen.
- Het verdient aanbeveling om een applicatie op te nemen in een werkproces. Zo wordt "afgedwongen" dat grote hoeveelheden data over een lange periode worden opgenomen in een werkproces en dit maakt het makkelijker om data beschikbaar te maken binnen de organisatie of te delen met anderen. Een platform als Lizard kan hierbij als datahub fungeren. Een nuttige vervolgstap is het opstellen van een implementatieplan, waarin beschreven staat voor welke werkprocessen toegevoegde waarde is te verwachten en welke eisen dat aan de data stelt en welke visualisatievormen gewenst zijn.

6 Literatuur

Agudelo-Vera, C., et al. (2015). *Drinking water temperature in future urban areas*. BTO 2015.012, KWR Watercycle Research Institute.

Agudelo-Vera, C. M. (2017). *Hotspots in het leidingnet*. BTO 2017.023 (in concept), KWR Watercycle Research institute.

Blokker, E. J. M. and E. J. Pieterse-Quirijns (2013). *Modeling temperature in the drinking water distribution system*. Journal - American Water Works Association 105(1): E19-E29.

Dijkstra, E. (1959). *A note on two problems in connexion with graphs*. In E. Dijkstra, Numerische Mathematik: 269-271.

Friedman, J. H. (1999). *Greedy function approximation: a gradient boosting machine*. Stanford University.

Friedman, J. H. (2001). *Greedy function approximation: a gradient boosting machine*. Annals of statistics: 1189-1232.

Mantel, P. and R. Gevers (2017). *Hittestress in de bodem -Het effect van klimaatverandering en hittestress op de bodemfuncties*. Gron Weg en Waterbouw. Sweco Nederland B.V., Hogeschool Van Hall Larenstein. B.Sc.

Moisen, G. G., et al. (2006). *Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods*. Ecological modelling 199(2): 176-187.

Pedregosa, F., et al. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research 12(Oct): 2825-2830.

Prest, E., et al. (2016). *Biological Stability of Drinking Water: Controlling Factors, Methods, and Challenges*. Frontiers in Microbiology 7(45).

Sagiroglu, S. and D. Sinanc (2013). *Big data: A review*. Collaboration Technologies and Systems (CTS): 42-47.

Stewart, I. D. and T. R. Oke. (2012). *Local climate zones for urban temperature studies*

Van Bel, N. (2017). xxx. BTO 2017.204(s) (in concept). KWR Watercycle Research Institute.

Van der Wielen, P. W. J. J. (2014). *Effect van waterkwaliteit, seizoen, drinkwaterinstallatie en verblijftijd/afstand op opportunistische pathogenen in drinkwater*. BTO 2014.015, KWR Watercycle Research Institute.

Van Rooij, D. J. M. (2016). Clusteranalyse bruinwatermeldingen Eindhoven. Dept. Applied Mathematics, Fontys Hogescholen. BSc Thesis (in Dutch).

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

Van Summeren, J., et al. (2015). *Influence of temperature, network diagnostics, and demographic factors on discoloration-related customer reports*. Procedia Engineering 13th Computing and Control for the Water Industry Conference, CCWI 2015. Leicester, U.K.

Versteegh, J. F. M. and H. J. J. Dik. (2007). *The quality of drinking water in the Netherlands in 2006..* RIVM Report 703719022.

Vonk, E., et al. (2016). *Kennis uit waterdata in en rondom het leidingnet. Toepassing van machine learning en statistiek op data van Brabant Water*. KWR 2016.006, KWR Watercycle Research Institute.

Vries, D., et al. (2015). *Prototype softsensor voor het kalkafzettend vermogen van leidingwater*. BTO 2015.030, KWR Watercycle Research Institute

Wolters, D. and T. Brandsma (2012). "Estimating the Urban Heat Island in residential areas in the Netherlands using observations by weather amateurs. (Type." *Journal of Applied Meteorology and Climatology* 51(4): 711-721.

Bijlage I Data

Datasets

I.1 Geografische, statische data

TABEL 7: GEMIDDELDE GEBOUWHOOGTE RONDOM DE LOCATIE VAN DE WATERMETER

Watermeter	gemiddelde gebouwhoogte (r=10m)	gemiddelde gebouwhoogte (r=20m)	gemiddelde gebouwhoogte (r=50m)	gemiddelde gebouwhoogte (r=100m)
1797822	19.0	9.2	8.7	6.3
1805324	10.9	3.5	5.0	4.4
1808708	16.0	7.4	8.6	6.1
1811396	14.1	6.2	6.1	5.3
1812192	15.5	8.9	4.1	4.1
1813969	4.7	6.8	4.1	4.6
1813980	8.3	4.4	6.2	5.2
1815325	13.8	6.6	2.1	0.7
1815970	12.8	7.0	6.1	4.5
1816158	8.0	4.2	5.2	4.8
1818017	9.7	5.5	5.2	6.3
1818020	11.7	8.2	5.3	4.4
2070932	15.1	7.8	5.7	4.4

TABEL 8: SCORE VAN TYPE OPPERVLAK PER WATERMETER

Watermeter	gebouwen	groen	overig	verhard	water
1797822	0.17	0.27	0.44	0.12	0.00
1805324	0.15	0.32	0.44	0.09	0.00
1808708	0.15	0.32	0.44	0.09	0.00
1811396	0.15	0.15	0.54	0.15	0.00
1812192	0.15	0.32	0.44	0.09	0.00
1813969	0.15	0.32	0.44	0.09	0.00
1813980	0.16	0.19	0.53	0.12	0.00
1815325	0.02	0.79	0.13	0.02	0.03
1815970	0.16	0.19	0.53	0.12	0.00
1816158	0.16	0.19	0.53	0.12	0.00
1818017	0.15	0.32	0.44	0.09	0.00
1818020	0.07	0.57	0.29	0.05	0.01
2070932	0.15	0.32	0.44	0.09	0.00
2153725	0.07	0.57	0.29	0.05	0.01

TABEL 9: GEMIDDELDE GELEIDINGSCOËFFICIËNT VAN DE BODEM RONDOM DE WATERMETER

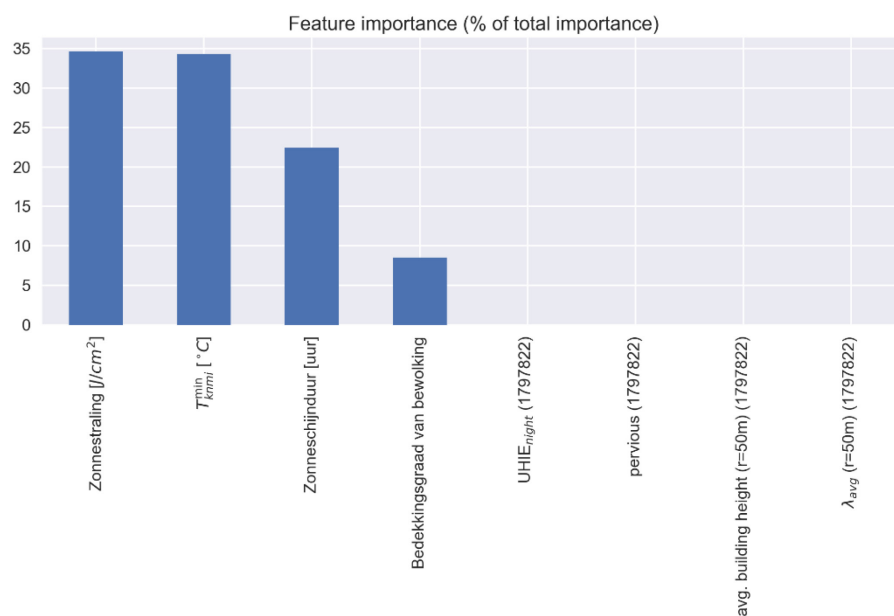
Watermeter	geleidingscoëfficiënt			
	(r=10m)	(r=20m)	(r=50m)	(r=100m)
1797822	2.9	2.9	2.9	2.9
1805324	2.9	2.9	2.9	2.9
1808708	2.9	2.9	2.9	2.9
1811396	1.7	1.7	1.7	1.7
1812192	2.9	2.9	2.9	2.9
1813969	2.9	2.9	2.9	2.9
1813980	1.7	1.7	1.7	1.7
1815325	1.7	1.7	1.7	1.7
1815970	1.7	1.7	1.7	1.7
1816158	1.7	1.7	1.7	1.7
1818017	1.7	1.7	1.7	1.7
1818020	1.7	1.7	1.7	1.7
2070932	1.7	1.8	2.0	2.1
2153725	1.7	1.7	1.7	1.7

TABEL 10: GEMIDDELDE UHIE RONDOM DE WATERMETER

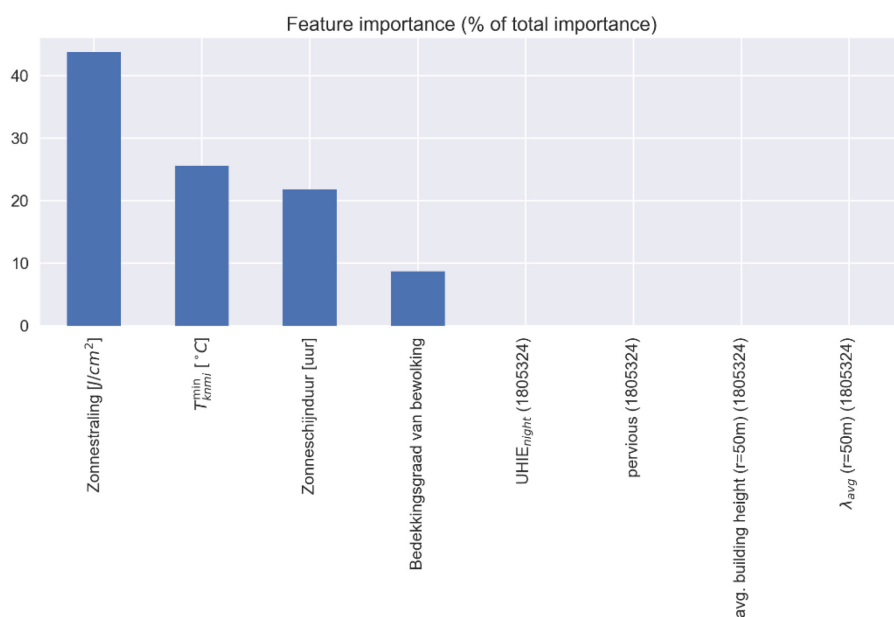
Code	UHIE over de dag				UHIE over de nacht
	(r=10m)	(r=20m)	(r=50m)	(r=100m)	
1797822	3.0	3.0	3.0	3.0	4.0
1805324	5.0	5.0	5.0	5.0	0.0
1808708	5.0	5.0	5.0	5.0	0.0
1811396	3.0	3.0	3.0	2.7	3.0
1812192	5.0	5.0	5.0	5.0	0.0
1813969	5.0	5.0	5.0	5.0	0.0
1813980	1.0	1.0	1.0	1.0	3.0
1815325	0.0	0.0	0.0	0.0	0.0
1815970	2.0	2.0	2.0	2.0	4.0
1816158	2.0	2.0	2.0	2.0	4.0
1818017	5.0	5.0	5.0	5.0	2.0
1818020	1.0	1.0	1.0	1.0	3.0
2070932	5.0	5.0	5.0	5.0	2.0
2153725	1.0	1.0	1.0	1.0	3.0

Bijlage II Figuren lokale regressie

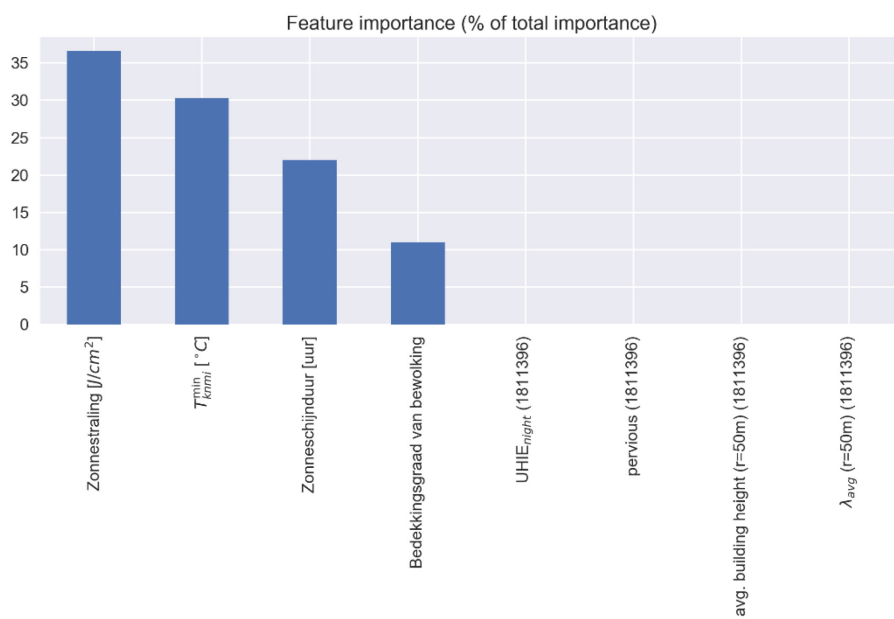
Relevantie van attributen per sensorlocatie



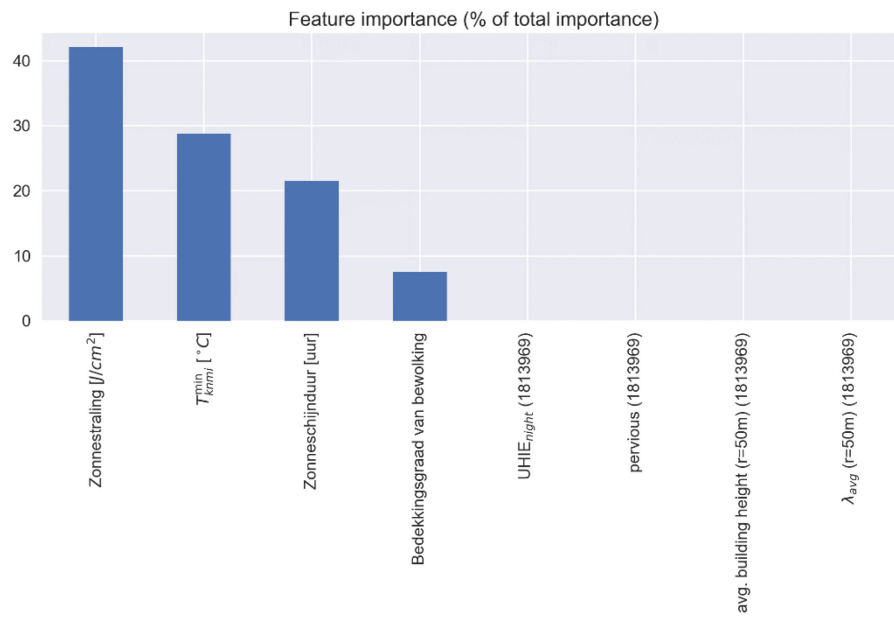
FIGUUR 6-1: LOKALE REGRESSIE, SENSORLOCATIE 1797822



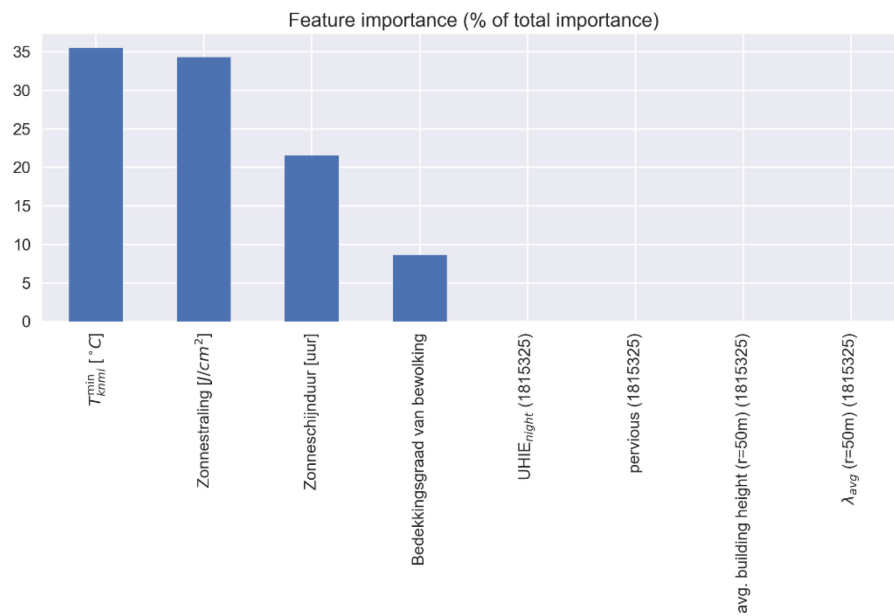
FIGUUR 6-2: LOKALE REGRESSIE, SENSORLOCATIE 1805324



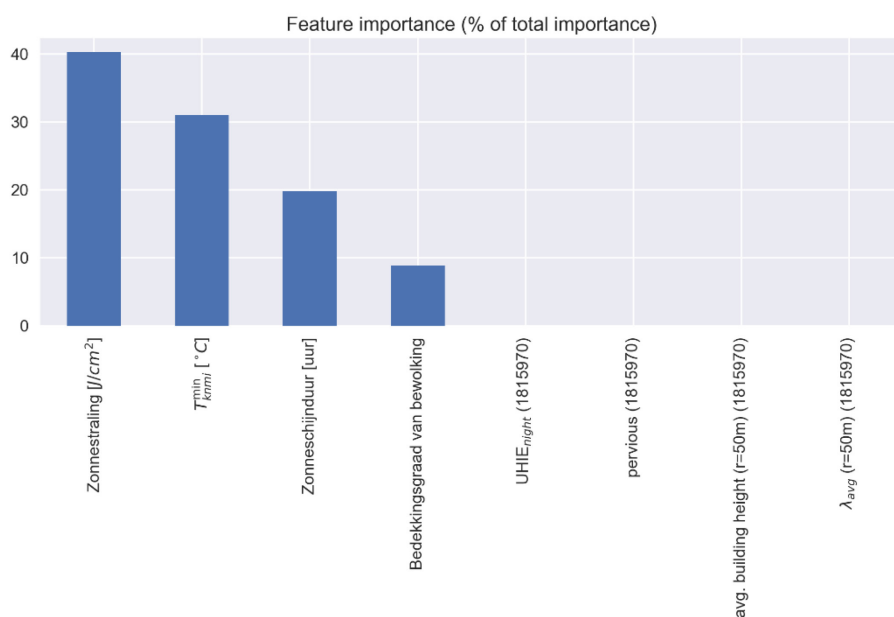
FIGUUR 6-3: LOKALE REGRESSIE, SENSORLOCATIE 1811396



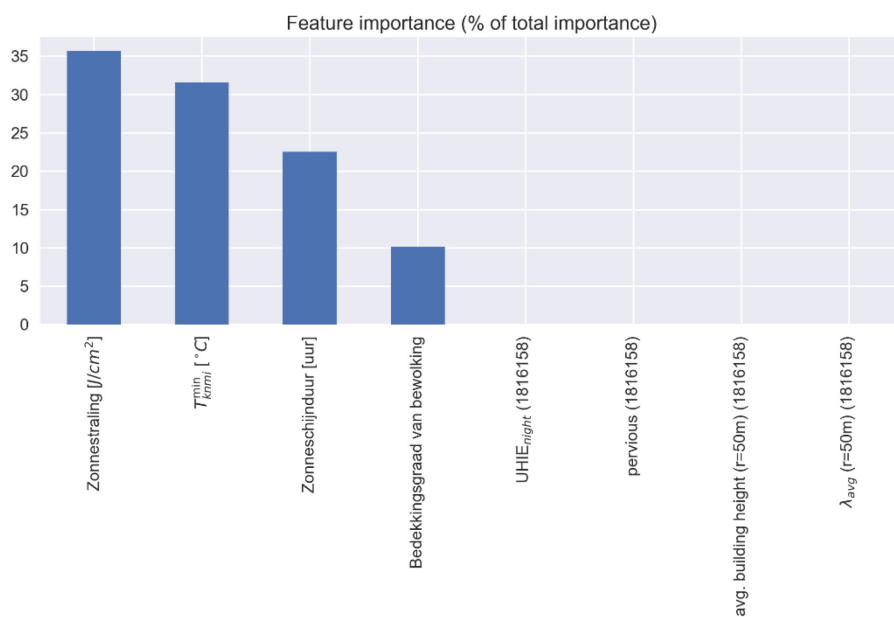
FIGUUR 6-4: LOKALE REGRESSIE, SENSORLOCATIE 1813969



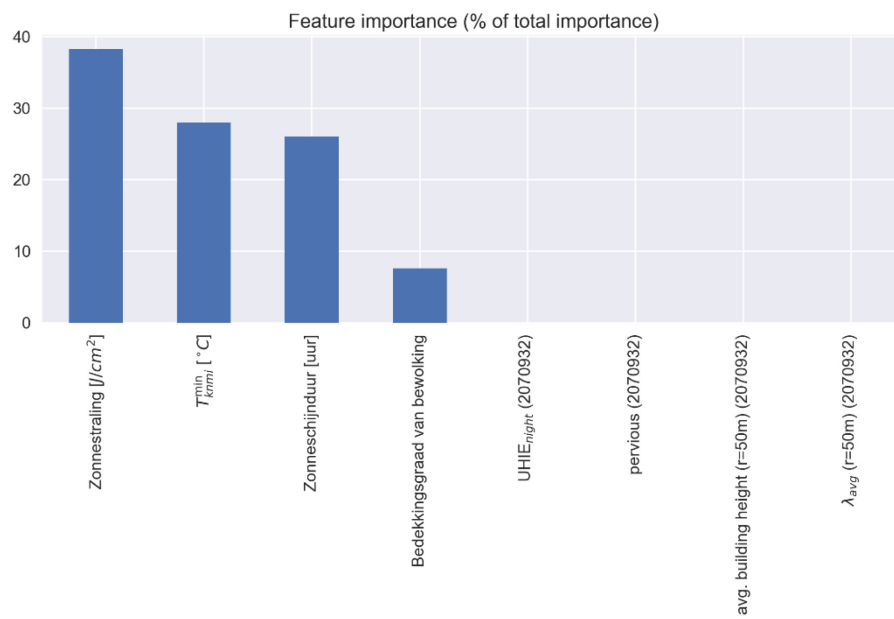
FIGUUR 6-5: LOKALE REGRESSIE, SENSORLOCATIE 1815325



FIGUUR 6-6: LOKALE REGRESSIE, SENSORLOCATIE 1815970



FIGUUR 6-7: LOKALE REGRESSIE, SENSORLOCATIE 1816158.



FIGUUR 6-8: LOKALE REGRESSIE, SENSORLOCATIE 2070932.