Sleutelfactor Toxiciteit

Vuistregels voor zuiveringsefficiëntie van stoffen op basis van stofeigenschappen

Input voor de waterkwaliteitsindex zuiveringsinspanning

Auteurs:
Tessa Pronk, Roberta Hofman (KWR Water Research Institute)

Data levering:
DPWE bedrijven (meetgegevens verwijdering)
Versie: september 2021

1. Inleiding

Voor een goede kwaliteit drinkwater is een bepaald niveau van zuiveringsinspanning nodig. De belangrijkste doelstelling van de KRW is de bescherming en verbetering van de waterkwaliteit, dat is inclusief water bestemd voor de productie van drinkwater. KRW-artikel 7.3 zegt: “De lidstaten dragen zorg voor de nodige bescherming van de aangewezen waterlichamen met de bedoeling de achteruitgang van de kwaliteit daarvan te voorkomen, teneinde het niveau van zuivering dat voor de productie van drinkwater is vereist, te verlagen.” RIWA-Rijn heeft om dit te kunnen kwantificeren een waterkwaliteitsindex laten ontwikkelen (RIWA-Rijn jaarrapport, 2018; Pronk et al., 2019). Hierover is ook gepubliceerd in een internationaal tijdschrift (Pronk et al., 2020). In feite gaat het om een raamwerk van indices: een voor de zuiveringsopgave (alles wat boven een beoogde grenswaarde aanwezig is) en een voor de zuiveringsinspanning (hoe makkelijk dit gezuiverd kan worden, op basis van stofeigenschappen). Gecombineerd geven deze het benodigde niveau van zuivering.

In deze notitie gaan we de berekening van de zuiveringsinspanning voor microverontreinigingen op basis van stofeigenschappen verbeteren. We gaan deze data-gedreven afleiden, voor verschillende zuiveringsstappen. Eerder was deze stap gebaseerd op een fictieve 'simpele' zuivering op basis van biodegradeerbaarheid en adsorptie (Pronk et al., 2020). In de bestaande literatuur worden wel veelvuldig bepaalde eigenschappen gekoppeld aan een mate van zuivering maar dit is altijd kwalitatief. De data-gedreven vuistregels zijn dus nieuw.

Tabel 1 geeft een overzicht van de zuiveringstechnieken die te onderscheiden zijn. We definiëren 'inspanning' niet heel precies. De technieken in Tabel 1 worden ingedeeld van conventioneel naar geavanceerd vanwege benodigde (verwachte) investeringen in de implementatie of onderhoud.

Tabel 1. Overzicht van technieken voor zuivering, van boven naar onder met toenemende inspanning.

Klasse inspanning	Techniek	Bijzonderheid
Conventioneel	Coagulatie Flocculatie (sedementatie) met snelfiltratie	Bij oppervlaktewater kunnen hier diverse vlokmiddelen worden toegevoegd. Bij grondwater is dit niet nodig.
Extra (een van deze technieken)	Oxidatie (Ozon)	Dit volgt op conventioneel. Na oxidatie wordt actieve kool toegepast.
	MembraanFiltratie (ultra- of nano-)	Dit volgt op conventioneel.
	Actieve Kool (poeder of granulair)	Dit volgt op conventioneel. Stoffen kunnen hier na verloop van tijd verdrongen worden door andere, beter adsorberende stoffen. Granulaire kool kan geactiveerd worden. De standtijd en contacttijd is bij granulaire kool belangrijk voor de zuiveringsefficiëntie. De kwantiteit is belangrijk bij poederkool.
Geavanceerd (een van deze technieken)	Omgekeerde osmose (RO)	Dit volgt op conventioneel. In plaats van 'Extra'.
	Geavanceerde oxidatie (AOP) (H_2O_2 en Ozon of UV)	Dit volgt op conventioneel. In plaats van 'Extra'. Na oxidatie wordt actieve kool of een ander filtratieproces toegepast.

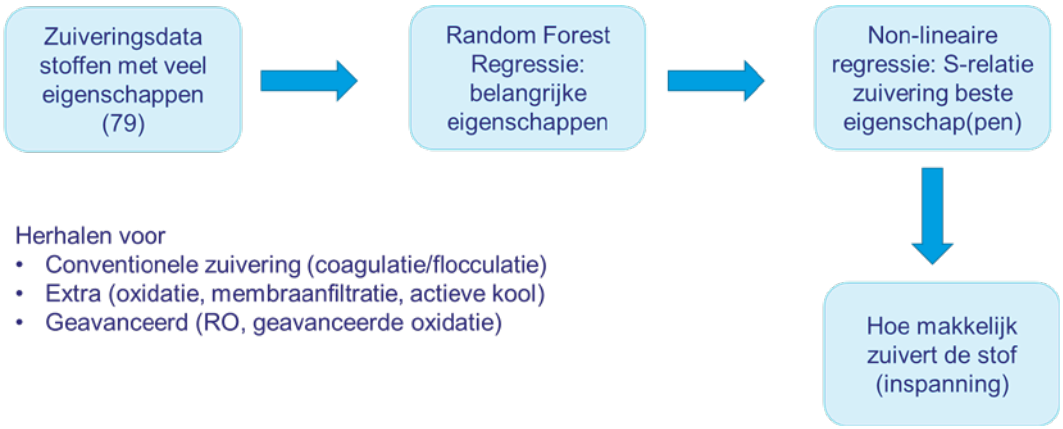
Om de zuiveringsinspanning voor water met microverontreinigingen te berekenen zullen we moeten bepalen hoe goed stoffen worden verwijderd in elke zuiveringstechniek. De verwijdering van stoffen wordt bepaald door stofeigenschappen, naast specifieke implementaties van zuiveringsinstallaties bij individuele waterbedrijven en de omstandigheden

zoals de aanwezigheid van natuurlijk organisch materiaal (NOM). Voor de zuivering op basis van stofeigenschappen zullen we data-gedreven vuistregels afleiden. Deze data-gedreven aanpak stelt de relatie vast tussen eigenschappen en zuivering in de praktijk, in plaats van uit de theorie.

De vuistregels die de verwijdering voor elke zuiveringsinspanning gaan voorspellen, zullen grove indicatoren aan de hand van stofeigenschappen zijn, voor een gemiddelde zuivering van een bepaalde zuiveringstechniek. De vuistregels geven een indicatie van de waterkwaliteit en geven niet de kwaliteit aan van specifieke zuiveringsinstallaties.

2. Methode

Voor elk van de zuiveringsstappen in Tabel 1 bepalen we een vuistregel voor de verwijdering van de stoffen in de zuiveringsstap op basis van stofeigenschappen, die op grond van data het beste past. In Figuur 1 staan de stappen bij elkaar. Voor de (organische) stoffen waarvoor zuiveringsdata bekend zijn in de datasets die we voor deze notitie gebruiken, zoeken we stofeigenschappen op (stap 1, Figuur 1).



Figuur 1: ***De stappen om tot een vuistregel voor zuiveringsefficiëntie te komen.***

Deze eigenschappen komen uit diverse bronnen. Het zijn met de OPERA modellen voorspelde eigenschappen die te downloaden zijn via het online ‘Chemistry Dashboard’ of de OPERA gebruikers interface (Mansouri et al., 2018). Ook voegen we ook de eigenschappen zoals opgeslagen in Open Babel (O’Boyle et al., 2011), en berekend door PubChem toe. Deze eigenschappen kunnen worden gedownload via een script in de open source statistische software R met het ChemmineR pakket (Cao et al., 2008). Ook voegen we eigenschappen toe die voorspeld zijn met modellen in EPIsuite (US EPA, 2012). In totaal worden er 79 (voorspelde) stofeigenschappen toegevoegd. Van deze eigenschappen controleren we of de logaritme van de eigenschap beter een “normale” verdeling laat zien dan de eigenschap zelf. Als dat zo is, nemen we de log van de eigenschap.

Van al deze eigenschappen is het de vraag welke geassocieerd zijn met de zuiveringsefficiëntie die behaald wordt in de zuiveringsstap. Voor het selecteren van eigenschappen gebruiken we een machine learning techniek; 'Random Forest' regressie (Breiman, 2001). Machine learning technieken in het algemeen maken gebruik van geavanceerde statistische modellen om een goed voorspellingmodel te 'leren'. Random Forest werkt met het maken van willekeurige beslisbomen. Het algoritme komt tot de best verklarende eigenschappen door de beslisbomen met de eigenschappen die het beste presteren om de zuiveringsefficiëntie te verklaren steeds te prioriteren. Met een Random Forest regressie in R bepalen we de belangrijkheid van de stofeigenschappen om de zuivering te voorspellen per zuiveringstechniek (zie stap 2, Figuur 1). Sommige stofeigenschappen zijn hetzelfde, er zijn bijvoorbeeld meerdere bronnen voor logKow (experimenteel of modelmatig). Alle min of meer dezelfde eigenschappen zullen logischerwijs met dezelfde belangrijkheid eindigen. De meest gecorreleerde eigenschappen (correlatie >0.97) zijn daarom vantevoren al verwijderd. Eigenschappen die een standaard deviatie hebben van nul (deze variëren niet), worden ook verwijderd vantevoren. Als er meerdere eigenschappen zijn die minder hoge correlatie hebben maar allemaal hoog eindigen na het Random Forest model, dan wijst dat erop dat mogelijk meerdere eigenschappen belangrijk zijn.

Na het bepalen van de meest belangrijke eigenschappen, willen we een simpele vuistregel vaststellen voor de relatie van de eigenschap met de zuiveringsefficiëntie (Stap 3 in Figuur 1).

Voor een vuistregel willen we graag dat de verwijdering tussen de 0 en 100% ligt. Een S-vormige relatie van een eigenschap met verwijdering is daarvoor een goede oplossing. Met een non-lineaire regressie fitten we in R een zo goed mogelijk passende S-curve door de data, die het verband tussen zuivering en stofeigenschap(pen) aangeeft. In Basisvergelijking 1 en 2 staan de twee vergelijkingen die mogelijk zijn.

$$V = \frac{C}{1+e^{A\cdot(Eigenschap-B)}}$$ Basisvergelijking 1

Hierbij is C het maximale bereik, B is het midden van de S-curve, en A is een maat voor steilheid. 'Eigenschap' is de waarde van de stofeigenschap. Als de relatie omgekeerd is (een afnemende waarde correleert met een toename in zuiveringsefficiëntie), gebruiken we de volgende vergelijking:

$$V = C - \frac{C}{1+e^{A\cdot(Eigenschap-B)}}$$ Basisvergelijking 2

We bekijken of de modellen voor de vuistregels die we afleiden significant zijn. Hoe goed de vuistregel de zuivering voorspelt bepalen we door de model fit (met de determinatiecoëfficiënt R^2 die de verklaarde variantie weergeeft) en de significantie van het verband tussen vuistregel en de meetwaarden te bepalen.

3. Vuistregels per zuiveringstechniek

3.1 Conventionele zuivering

Voor de conventionele zuivering is een dataset van één Nederlands DPWE drinkwaterbedrijf beschikbaar. Voor een selectie van stoffen is diverse keren gemeten wat de zuiveringsefficiëntie is. Incidenteel wordt een negatieve verwijdering gemeten. Dit komt waarschijnlijk doordat de kleine hoeveelheden stof moeilijk te meten zijn. Ook is het mogelijk dat de moederstof weer wordt terug gevormd uit de eerder gevormde metabolieten bij afbraak van de stof. We zetten alle meetpunten die negatieve verwijdering hebben, op nul. We houden de stoffen die een negatieve verwijdering hebben wel in de dataset. Deze set heeft voor een aantal stoffen veel meetpunten per stof. Voor een robuust resultaat selecteren we stoffen met 9 of meer meetpunten. Hierbij blijven er 29 stoffen over.

We kijken welke van de stofeigenschappen het meest geassocieerd zijn met de verwijderingspercentages in de dataset voor conventionele zuivering. Daarvoor gebruiken we een Random Forest regressie (Stap 2 in Figuur 1) die de belangrijkheid van variabelen voor het verklaren van de verwijderingsefficiëntie kan vaststellen. Uit de Random Forest komen de eigenschappen in Tabel 2 naar voren als meest belangrijke.

Tabel 2. De meest verklarende eigenschappen voor de zuiveringsefficiëntie van stoffen in de conventionele zuivering dataset uit de Random Forest regressie. De top 5 voor beide maten is weergegeven.

Naam	Bron	Uitleg	%IncMSE*	IncNodePurity*
OCTANOL_WATER_PARTITION_LOGP	Opera	Log octanol water coëfficiënt	562	229550
PredlogKow	EpiSuite	Log octanol water coefficient	662	227310
KM_DAYS	Opera	Biotransformatie snelheid	385	150269
TPSA	Pubmed	Totale polaire oppervlakte	396	131055
HBondAcceptorCount	Pubmed	Aantal H-binding acceptoren	301	112330

** %IncMSE staat voor de toename in mean square error bij het weglaten van deze eigenschap. Hoe hoger, hoe belangrijker de eigenschap. IncNodePurity staat voor de mate van de verhoging van hoe goed een uitkomst voorspeld kan worden door een beslispunt in de beslisboom met deze eigenschap toe te voegen. Hoe hoger, hoe beter.*

Uit Tabel 2 blijkt dat de stofeigenschap Log octanol water coëfficiënt (logKow) bovenaan staat in beide maten van belangrijkheid voor het behalen van en goed resultaat in de RF regressie (%IncMSE en IncNodePurity). Dit is daarom de meest geassocieerde eigenschap voor het bepalen van de zuiveringsefficiëntie. De RF regressie geeft een behoorlijk goed resultaat (78%

van de variatie verklaard). Een nadeel van een machine learning model is dat het zo ingewikkeld is, dat het alleen als computermodel kan worden toegepast. Het model bestaat voor RF regressie uit zeer veel (duizenden) beslisbomen, die samen het resultaat geven. Het is daardoor niet praktisch in gebruik of intuïtief gemakkelijk interpreteerbaar. Daarom offeren we nauwkeurigheid op voor een simpele formule. We gebruiken logKow om de S-curve (Basisvergelijking 1) te fitten. De coëfficiënten A,B,C (zie Basisvergelijking 1) zijn significant vast te stellen in dit model, zoals hieronder te zien is in de rapportage van de fit zoals gegenereerd in R.

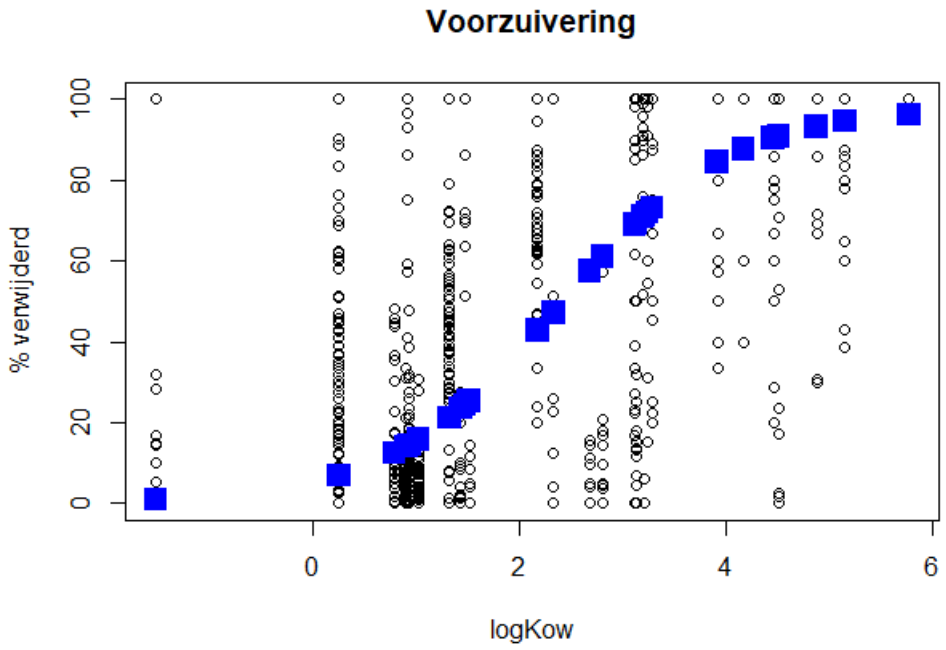
Non-lineaire regressie fit voor de vuistregel conventionele zuivering met logKow:

```
Formula: percentage_adj ~ C/(1 + exp(A*(OCTANOL_WATER_PARTITION_LOGP_OPERA_PRED - B)))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
A -1.20279    0.08550  -14.07   <2e-16 ***
B  2.39456    0.09721   24.63   <2e-16 ***
C 97.89476    3.12950   31.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vullen we de waarden van deze coefficienten in bij Basisvergelijking 1, komen we op de volgende vergelijking voor de vuistregel:

$$V_{conv} = \frac{98}{1+e^{-1.2\cdot(logP-2.4)}} \qquad \text{Vergelijking 1}$$

In Figuur 2 staat de S-curve die is gefit voor de conventionele zuivering (Vergelijking 1). Het verband tussen voorspelde waarden met deze vuistregel op basis van logKow en gemeten zuivering heeft een R^2 van 0.56. Een deel van de variatie van de metingen in deze figuur komt door variatie *binnen* een stof. Zouden we de mediaan uitzetten, dan is de correlatie beter te zien (niet getoond). De weergave in Figuur 2 geeft echter beter de werkelijke variatie in meetwaarden aan. Het lijkt misschien of de relatie niet goed past, maar het feit dat de coëfficiënten in Vergelijking 1 significant kunnen worden vastgesteld en dat de relatie tussen gemeten en berekende zuivering ook significant is, zit deze relatie wel degelijk in de data, ondanks de variabiliteit.



Figuur 2: ***Verwijdering van stoffen in relatie tot logKow. Zie Vergelijking 1 voor de formule voor de blauwe punten. Dit zijn de waarden uit de vuistregel.***

Als de tweede eigenschap, KM_DAYS, ook wordt meegefit geeft dit een R^2 van 0.58. Deze verbetering weegt niet op tegen de voordelen van een simpel model.

De vuistregel is afgeleid met data voor oppervlaktewater. Bij grondwater is het proces in principe hetzelfde. Tijdens beluchting word Fe(II) tot Fe(III) geoxideerd, en vervolgens werkt het weer hetzelfde als toevoegen van Fe(III) en daardoor coagulatie. Mogelijk speelt bij grondwater wel de stofeigenschap 'vluchtigheid' een grotere rol, omdat bij de beluchting relatief vluchtige stoffen eerder zullen vervliegen. Ook de samenstelling van het water kan anders zijn.

3.2 Actieve kool

Voor deze vuistregel kijken we naar data met actieve kool (granulaire actieve kool, GAC en poeder actieve kool, PAC). Er zijn drie bedrijven in de dataset waarvoor data beschikbaar is. In heel Nederland zijn er meer bedrijven die dit toepassen. Bij zowel PAC als GAC zien we een aantal stoffen met een negatieve verwijdering. Bij actieve kool worden op een gegeven moment stoffen die minder goed hechten, verdrongen door microverontreinigingen die beter aan actieve kool hechten. Aanvankelijk zullen deze verdreven stoffen wel goed geadsorbeerd zijn geweest. Dit geeft een extra uitdaging voor het bepalen van een vuistregel voor een eigenschap die de verwijdering met actieve kool weergeeft. Omdat PAC en GAC veel overeenkomsten hebben, verwachten we dat een eigenschap die een vuistregel kan zijn voor actieve kool voor beide toepassingen geldt. We kijken welke eigenschappen naar voren komen bij de RF regressie voor de set in totaal. We zien in Tabel 3 de eigenschappen die hoog eindigen.

Tabel 3. De meest verklarende eigenschappen voor AC dataset uit de Random Forest regressie.

Naam	Bron	Uitleg	% Inc MSE	Inc Node Purity
OCTANOL_AIR_PARTITION_COEFF_LOGKOA	Opera	LogKoa; partitionering gedrag tussen lucht en omgevingsmatrices (bodem, vegetatie, aerosolen)	172	31964
VAPOR_PRESSURE_MMHG	Opera	Dampdruk	142	28366
BOILING_POINT_DEGC	Opera	Kookpunt (graden Celsius)	137	27623
HENRYS_LAW_ATM.M3.MOLE	Opera	Evenwichtssituatie van een oplosmiddel dat in contact is met een gas	113	27278
ATMOSPHERIC_HYDROXYLATION_RATE_.AOH._CM3.MOLECULE.SEC	Opera	Introduceert een hydroxylgroep in een organische verbinding, vaak de eerste stap in het degradatieproces van organische stoffen in lucht	81	21031

De Random Forest regressie verklaart 12.3%. Dat is laag voor dit geavanceerde statistische model. Nemen we als aanvulling de factor ‘conditie’ mee wat de verschillende omstandigheden weergeeft in de verschillende bedrijven (PAC/GAC, contacttijd, versheid, seizoen), stijgt de verklaarde variantie naar 56%. Dit betekend dat de omstandigheden inderdaad voor een groot deel de verwijdering bepalen. Maar, omdat we de vuistregel aan de hand van en stofeigenschap willen maken, nemen we toch de hoogst verklarende stofeigenschap logKoa hiervoor.

De eigenschap logKoa wordt in literatuur niet in verband gebracht met zuiveringsefficientie van microverontreinigingen met actieve kool. Op GAC worden eigenschappen zoals logKow en biodegradatie verwacht (Korotta-Gamage et al. 2017; Garcia et al. 2021). Op PAC alleen logKow, omdat er geen microorganismen groeien op deze vorm van actieve kool (eenmalig gebruik) en biodegradatie daardoor geen rol speelt. Ook HBA1 (aantal H bond acceptoren), zou verwacht zijn, dit is het aantal H-bruggen dat gevormd kan worden. Hoe meer H-bruggen er gevormd kunnen worden, des te hydrofieler een stof is, en hoe minder adsorptie op actieve kool je verwacht. LogKoa is een maat voor verdeling van semi-vluchtige stoffen tot aerosolen (Finizio et al., 1997), ofwel een parameter die de partitioning definieert tussen organische stoffen en atmosfeer. Het is ongeveer gelijk aan de ratio Kow met de constante van Henry (Meylan et al., 2005) maar heeft ook een relatie met molair volume (Busca, 2020). Hoe lager de logKoa, hoe vluchtiger de stof. Een hoge logKoa is in de literatuur onder andere geassocieerd met een hoge mate van bioaccumulatie in organismen.

Dampdruk is de tweede belangrijke eigenschap. Dit heeft naar verwachting niet te maken met het vervluchtigen van de stof zelf tijdens de behandeling met actieve kool. De stoffen worden voor het bepalen van de zuiveringsefficiëntie voor de DPWE data wel aan het water toegevoegd (spiken). Maar, in de DPWE set zijn alleen stoffen meegenomen die daadwerkelijk in de

bronnen zijn aangetroffen. Dit zijn dus geen stoffen die direct zullen vervluchtigen na toevoeging en het is de vraag of ze in de dampfase komen. Inderdaad zijn er maar drie stoffen in de dataset die een Henry constante hebben kleiner is dan 10^{-5} atm m^3/mol en mogelijk vervluchtigen (niet getoond). Ook kookpunt is een eigenschap die in Tabel 3 hoog eindigd. Zowel het kookpunt van een stof als de interacties met het actieve kooloppervlak worden grotendeels bepaald door molmassa en polariteit/lading. Dat kan verklaren waarom het kookpunt een maat kan zijn voor de effectiviteit van adsorptie.

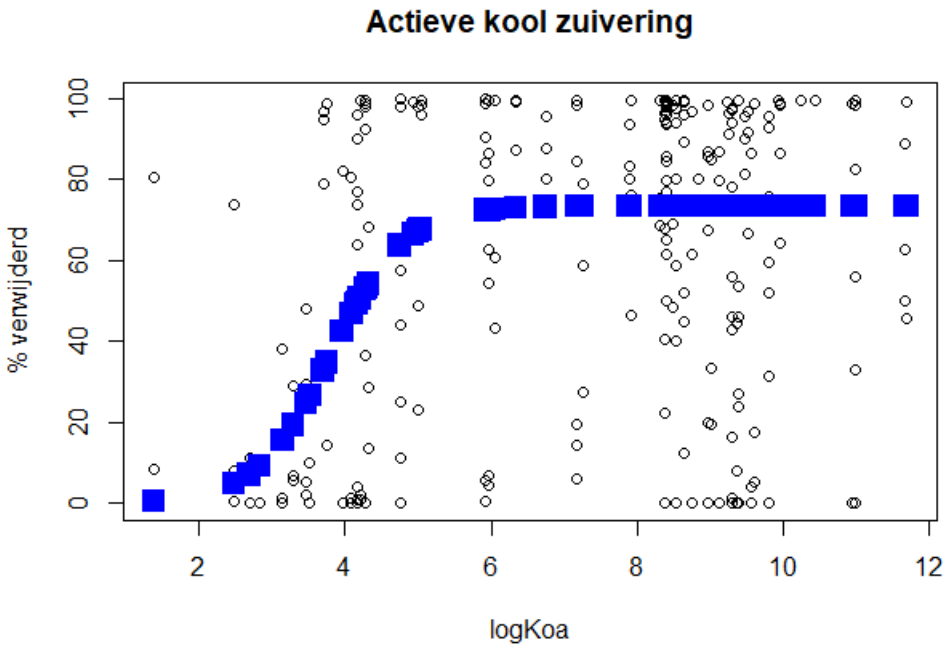
Als we de fit met logKoa proberen, komt daar in de S curve (zie Figuur 3) een significante modelfit voor. De coëfficiënten A,B,C (zie Basisvergelijking 1) zijn significant vast te stellen in dit model, zoals hieronder te zien is in de rapportage van de fit zoals gegenereerd in R.

Non-lineaire regressie fit voor de vuistregel conventionele zuivering met logKoa:

```
Formula: measurement_adj~(C/(1+exp(A*(OCTANOL_AIR_PARTITION_COEFF_LOGKOA_OPERA_PRED-
B))))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
A  -1.9942     0.6404  -3.114  0.00202 **
B   3.7981     0.1521  24.979  < 2e-16 ***
C  73.4368     2.2916  32.047  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.74 on 314 degrees of freedom
```

Een lineair model op verwijdering en voorpelde verwijdering op basis van de vuistregel geeft een R^2 van 0.17. Fitten we in aanvulling op logKoa ook het kookpunt mee dan wordt de R^2 verhoogd tot 0.18. Omdat we simpele vuistregels willen, kiezen we voor een model van alleen logKoa.

De lage R^2 geeft aan dat er veel ruis en variabiliteit in de data zitten die niet verklaard worden door de stofeigenschap. De trend is wel significant (zie Tabel 8). Dit betekent dat de voorspellende variabele nog steeds informatie geeft over de respons, ook al vallen de datapunten verder van de regressielijn. De voorspellingen zullen alleen niet heel *precies* zijn (hebben een hoog predictie-interval). Voor meer informatie over de combinatie lage R^2 met hoge significantie, zie https://blog.minitab.com/en/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values.



Figuur 3: ***Verwijdering van stoffen in relatie tot logKoa. Zie Vergelijking 2 voor de formule voor de blauwe punten. Dit zijn de waarden uit de vuistregel.***

$$V_{ak} = \frac{73.4}{1+e^{-1.99\cdot(\log Koa-3.8)}}) \qquad \text{Vergelijking 2}$$

De vuistregel verklaart niet veel van de variatie, maar wat het verklaart is wel significant. Bij de interpretatie moet rekening gehouden worden met het feit dat dit niet de zuivering weergeeft bij een specifiek bedrijf of op een bepaald moment (bij elk bedrijf zijn de omstandigheden weer anders) maar dat dit een algemene vuistregel is voor een ‘gemiddelde’ situatie en dat de spreiding in de praktijk groot is. Voor de waterkwaliteit op basis van verwijderbaarheid van stoffen gebaseerd op logKoa hebben we de omstandigheden bij bedrijven niet nodig. De vuistregel geeft immers niet aan wat de verwijdering bij een bedrijf wordt, maar een algemene indicatie van de waterkwaliteit. Wat we wel moeten controleren is of deze vuistregel niet bijvoorbeeld voor maar één bedrijf geldt, of alléén voor PAC danwel GAC.

Als we de RF regressie analyse doen voor de drie bedrijven apart, komt logKoa niet bij alle bedrijven hoog in de RF analyse voor geassocieerde eigenschappen (Tabel 4). Wel spelen bij alle bedrijven eigenschappen die te maken hebben met vluchtigheid (Vapor pressure, Henry’s law). Het is opvallend dat de twee bedrijven met GAC de eigenschappen logKow (‘logP’) en biodegradeerbaarheid toch in de top 5 komen. De eigenschap ‘Solubility’ is ook geassocieerd met logKow, die twee eigenschappen zijn negatief gecorreleerd. LogKoc is juist positief gecorreleerd met logKow.

Tabel 4. Eigenschappen die volgens de Random Forest regressie geassocieerd zijn met de verwijdering met PAC (links) en GAC (midden, rechts). De eigenschappen staan op volgorde van belangrijkheid, de meest belangrijke boven. GAC2 wordt toegepast na oxidatie.

PAC	GAC1	GAC2 (na oxidatie)
BOILING_POINT_DEGC	BioDegr	MELTING_POINT_DEGC
OCTANOL_AIR_PARTITION_COEFF_LOGKOA	logKOC	WATER_SOLUBILITY_MOL.L
ATMOSPHERIC_HYDROXYLATION_RATE_.AOH._CM3.MOLECULE.SEC	HENRYS_LAW_ATM.M3.MOLE	VP
VP	PredlogKow	logP
MR	WATER_SOLUBIL-ITY_MOL.L	MR
Complexity	FeatureRingCount3D	BioDegr

Hoe minder data, hoe meer de resultaten beinvloedbaar zijn door specifieke stoffen of meetfouten en specifieke omstandigheden. Dit kan de reden zijn voor de variatie in de meest belangrijke eigenschappen in Tabel 4. Voor de individuele bedrijven waren we niet op de huidige vuistregel uitgekomen volgens Tabel 4, maar mogelijk geldt de vuistregel wel voldoende. Als we de vuistregel (Vergelijking 2) *toepassen* op de data van de drie individuele bedrijven die actieve kool gebruiken, dan is de vuistregel voor PAC significant geassocieerd met de gemeten zuivering ($p<0.001$, R^2 0.28), voor GAC1 ook ($p<0.001$, R^2 0.14), maar voor GAC2 is de vuistregel niet significant ten opzichte van de gemeten zuiveringsefficienties ($p<0.95$). Mogelijk is het verschil veroorzaakt doordat bij GAC2 de watersamenstelling anders is met betrekking tot natuurlijk organisch materiaal. Dit komt doordat de GAC2 wordt toegepast na oxidatie. Oxidatie heeft er zeker toe geleid dat stoffen, waaronder NOM, zijn veranderd, qua structuur, en er dus sprake is van een andere matrix.

De variatie in verwijderingsefficientie die met de vuistregel verklaard wordt (R^2 =0.17) is wel erg laag. Omdat we data-gedreven werken houden we dit voorlopig toch aan als ‘beste optie’. Vuistregel 2 zal idealiter nog moeten worden bevestigd door een nieuwe, onafhankelijke dataset voor zuiveringsefficientie van stoffen met behandeling van actieve kool.

3.3 Oxidatie

Een behandeling met O_3 (ozon) is oxidatie. Er zijn maar data beschikbaar van één bedrijf in de dataset dat dit toepast. In Tabel 5 staan de stofeigenschappen die volgens de RF regressie het meest geassocieerd zijn met de verwijdering door oxidatie.

Tabel 5. De meest 6 verklarende eigenschappen voor oxidatie dataset uit de Random Forest regressie.

Naam	Bron	Uitleg	% Inc MSE	Inc Node Purity
abonds	Open Babel	Aantal aromatische bindingen	316	19174
OCTANOL_AIR_PARTITION_COEFF_LOGKOA	Opera	verdeling tussen lucht en omgevingsmatrices (bodem, vegetatie, aerosolen)	296	18349
ATMOSPHERIC_HYDROXYLATION_RATE_.AOH._CM3.MOLECULE.SEC	Opera	Introduceert een hydroxylgroep in een organische verbinding, vaak de eerste stap in het degradatieproces van organische stoffen in lucht (en zeker ozon)	266	17143
FeatureRingCount3D	Pubmed	aantal ringstructuren per stof features per compound	243	14920
BOILING_POINT_DEGC	Opera	Kookpunt (graden Celsius)	182	14046
BioDegr	EPIsuite	Biodegradatie snelheid (uren tot jaren)	230	13417

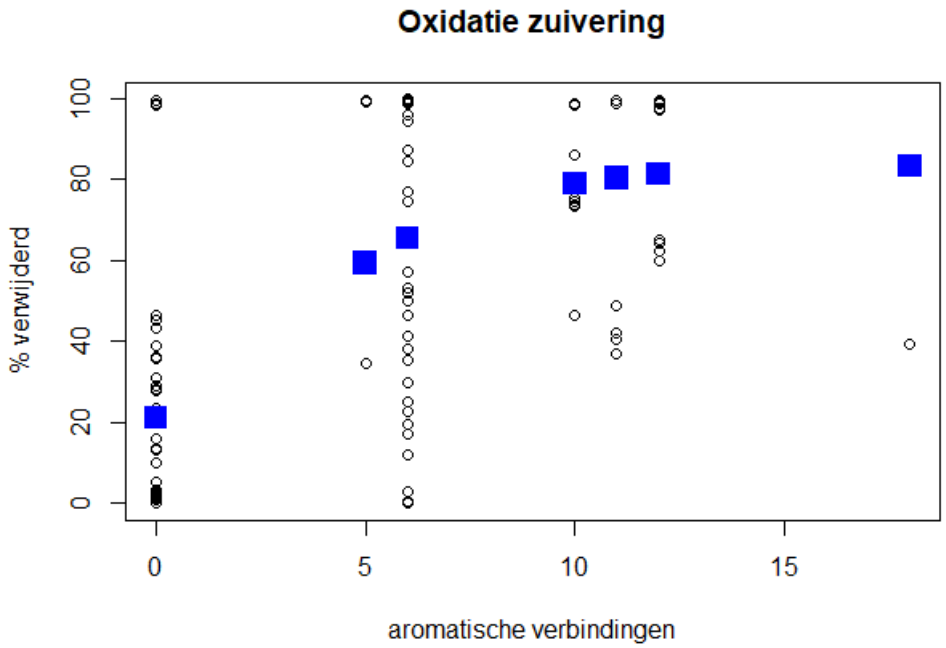
Voor oxidatie is het aantal aromatische verbindingen (abonds) belangrijk, blijkt uit de analyse. Aromatische verbindingen zijn een bekende voorspeller van oxidatie met ozon (Sonntag et al. 2012). Random Forest regressie komt tot een acceptabele 71.9% verklaarde variantie. Maar dat model is te ingewikkeld om als vuistregel te gebruiken.

De coëfficiënten A,B,C (zie Basisvergelijking 1) zijn significant vast te stellen in dit model, zoals hieronder te zien is in de rapportage van de fit zoals gegenereerd in R.

Non-lineaire regressie fit voor de vuistregel oxidatie met abonds:

```
Formula: measurement_adj ~ (C/(1 + exp(A * ((OpenBabelabonds) - B))))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
A -0.39478    0.09488  -4.161 5.96e-05 ***
B  2.71482    0.84309   3.220  0.00165 **
C 83.50323    7.13869  11.697  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

De voorspelde verwijdering met de S-curve (Vergelijking 3, Figuur 4) op basis van abonds heeft een R^2 van 0.41 met de gemeten verwijdering.



Figuur 4: ***Verwijdering van stoffen in relatie tot abonds. Zie Vergelijking 3 voor de formule voor de blauwe punten. Dit zijn de waarden uit de vuistregel.***

$$V_{ox} = \frac{83.5}{1+e^{-0.39\cdot(abonds-2.71)}}$$ Vergelijking 3

Voor deze vuistregel geldt wederom dat deze is vastgesteld op basis van een beperkte set data. Omdat de eigenschap bekend is, is het wel redelijk om te veronderstellen dat de vuistregel in de goede richting zit. Een validatie met een onafhankelijke dataset zal dit kunnen bevestigen. Naar aromatische dubbele bindingen kunnen ook gewone dubbele bindingen geoxideerd worden. Voor een aantal stoffen kan dit in dit verband ook belangrijk kunnen zijn voor de verwijdering maar dat komt niet uit de analyse van deze dataset.

3.4 Geavanceerde oxidatie

Geavanceerde oxidatieprocessen (AOP) zijn processen waarbij hydroxylradicalen worden gevormd, die erg effectief kunnen reageren met een breed scala aan verbindingen. Voorbeelden van AOP zijn UV in combinatie met O_3 of H_2O_2, of de combinatie O_3/H_2O_2. Voor AOP zijn voor twee bedrijven data beschikbaar in de dataset. AOP worden op dit moment ook bij twee drinkwaterbedrijven toegepast.

Tabel 6. De meest verklarende eigenschappen voor oxidatie dataset uit de Random Forest regressie.

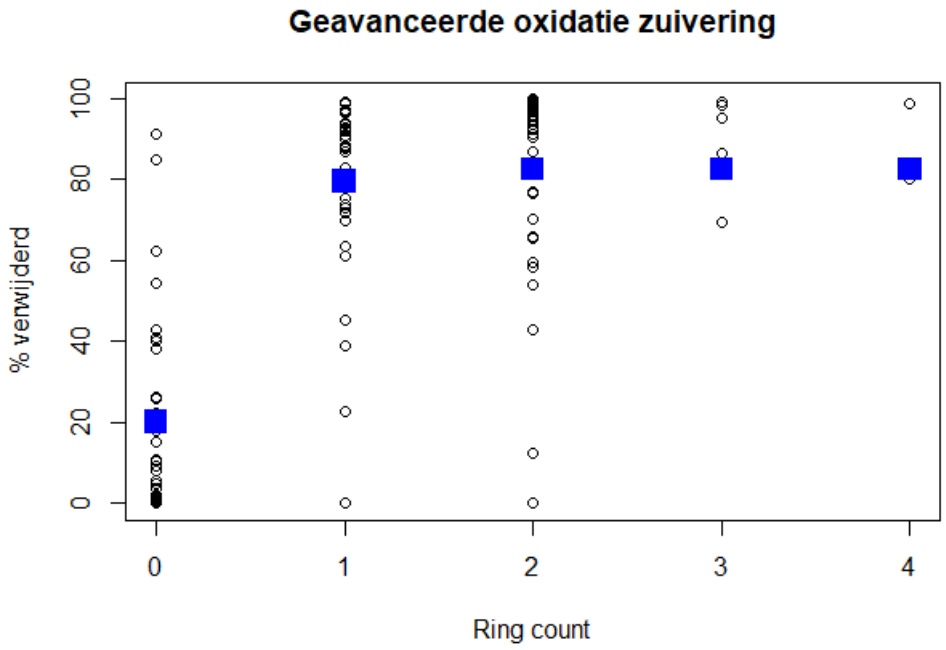
Naam	Bron	Uitleg	%IncMSE	IncNodePurity
FeatureRingCount3D	Pubmed	Count of ring features per compound	321	19846
HBA1	Open Babel	H bond acceptor aantal	278	16376
ATMOSPHERIC_HYDROXYLATION_RATE_.AOH._CM3.MOLECULE.SEC_OPERA_PRED	Opera	Introduceert een hydroxylgroep in een organische verbinding, vaak de eerste stap in het degradatieproces van organische stoffen in lucht (en ozon)	236	15684
BioDegr	EPIsuite	Biodegradatie snelheid (uren tot jaren)	173	13228
nF	Open Babel	Number of Fluorine Atoms	190	11535

Voor geavanceerde oxidatie zijn ring eigenschappen en HBA1 (Number of Hydrogen Bond Acceptors) belangrijke eigenschappen. Het aantal ringen is een logische stofeigenschap om de verwijdering te duiden (Jenkin et al. 2020; Jenkin et al. 2018; Jin et al. 2020). De eigenschap die bij oxidatie het meest belangrijk is, abonds, staat bij geavanceerde oxidatie op plek 7 en staat daardoor net niet in Tabel 6. RF regressie komt tot een acceptabele 73.4% verklaarde variantie.

De coëfficiënten A,B,C (zie Basisvergelijking 1) zijn significant vast te stellen in dit model, zoals hieronder te zien is in de rapportage van de fit zoals gegenereerd in R.

Non-lineaire regressie fit voor de vuistregel AOP met FeatureRingCount3D:

```
Formula: measurement_adj ~ (C/(1 + exp(A * ((PMFeatureRingCount3D) - B))))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
A  -4.3928     1.5845  -2.772  0.00647 **
B   0.2560     0.1095   2.338  0.02108 *
C  82.5572     3.4432  23.977  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.1 on 118 degrees of freedom
```



Figuur 5: ***Verwijdering van stoffen in relatie tot aantal ringen. Zie Vergelijking 4 voor de formule voor de blauwe punten. Dit zijn de waarden uit de vuistregel.***

$$V_{advox} = \frac{82.6}{1+e^{-4.4\cdot(\text{FeatureRingCount3D}-0.26)}} \qquad \text{Vergelijking 4}$$

De beste vuistregel is op basis van FeatureRingCound3D (Figuur 5), en heeft een R^2 van 0.59 met gemeten zuiveringsefficienties.Als we in de vuistregel het aantal ringstructuren combineren met HBA1 neemt de R^2 toe tot 0.68. Dit is een behoorlijke verbetering. Omdat het testen van een ingewikkelder model meer tijd kost gebruiken we toch alleen de eerste eigenschap, het aantal ringstructuren. De vuistregel voor AOP is niet heel erg onderscheidend. Van ring count 1-4 is de voorspelling ongeveer gelijk. Alleen ring count 0 geeft een lage verwijdering. In de praktijk wordt AOP niet toegepast zonder een aanvullende filtering met actieve kool of bijvoorbeeld duinfiltratie, om bijvoorbeeld de overmaat H_2O_2, transformatieproducten en gevormd AOC (assimileerbaar organisch koolstof) te verwijderen. Dat laatste is om de stabiliteit van het water te behouden/verbeteren.

3.5 Omgekeerde Osmose

Voor ‘geavanceerde’ zuivering Omgekeerde Osmose (RO) zijn data beschikbaar van één bedrijf. We kijken eerst wat de meest belangrijke eigenschappen zijn, daarna kijken we of we een versimpelde vuistregel kunnen maken. In Tabel 7 staan de eigenschappen die het hoogst eindigden na het toepassen van de RF regressie.

Tabel 7. De meest verklarende eigenschappen voor geavanceerde zuivering omgekeerde osmose (RO) uit de RF regressie.

Naam	Bron	Uitleg	% Inc MSE	Inc Node Purity
logAVERAGE_MASS	Chemistry Dashboard	Gemiddelde Massa	252	12328
logEPISFUG	EpiSuite	Fugacity, potentieel van een stof om van het ene naar het andere compartiment te gaan	121	7577
logOpenBabelsbonds	OpenBabel	'Sigma' bindingen, sterkste type covalente binding	71	7271
logPMMonoisotopic Mass	Pubmed	Monoisotopische massa	104	6844
EPISBioDegr	EPIsuite	Biodegradatie snelheid (uren tot jaren)	80	6291

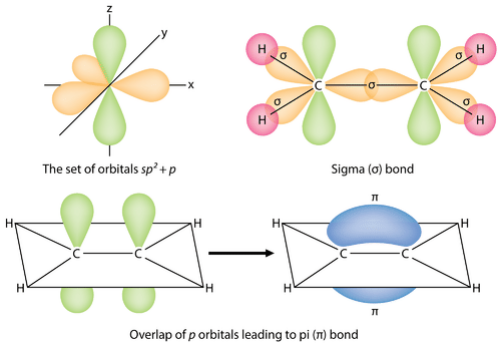
Met het Random Forest model wordt 56.7% van de variantie verklaard. Maar, dit model is te ingewikkeld om als vuistregel te dienen. Daarom fitten we weer een S-curve (zie Figuur 7) met een non-lineaire regressie in de statistische software 'R' met de meest verklarende eigenschap (Vergelijking 5).

De coëfficiënten A,B,C (zie Basisvergelijking 1) zijn significant vast te stellen in dit model, zoals hieronder te zien is in de rapportage van de fit zoals gegenereerd in R.

Non-lineaire regressie fit voor de vuistregel RO met log average mass:

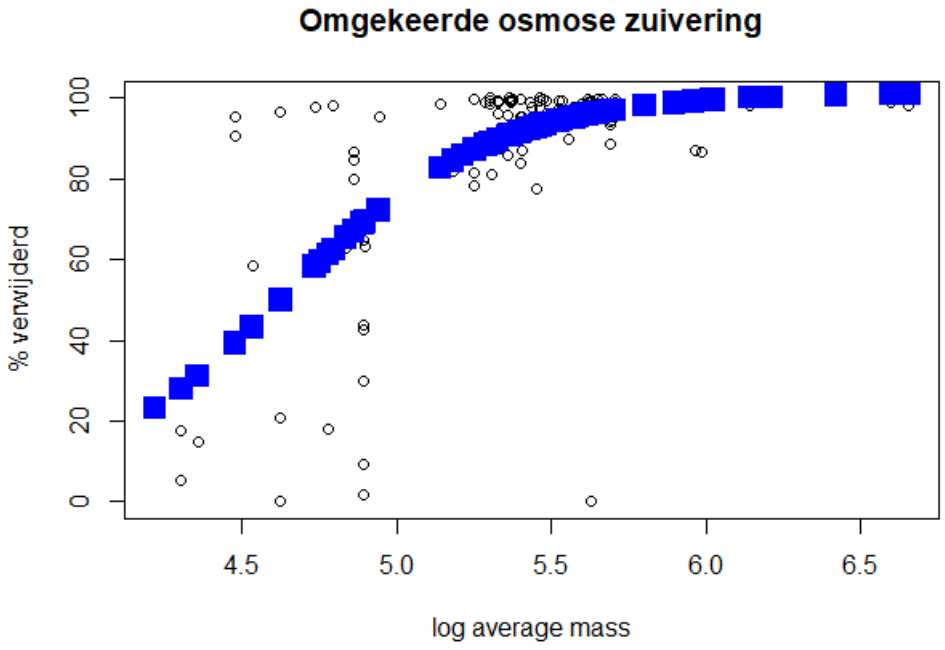
```
Formula: measurement_adj ~ (C/(1 + exp(A * (log(AVERAGE_MASS) - B))))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
A  -2.94922    0.61177  -4.821 4.01e-06 ***
B   4.63332    0.06045  76.650  < 2e-16 ***
C 101.11235    3.97122  25.461  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De R^2 van dit model met de gemeten verwijdering is 0.48. De eigenschap log (gemiddelde massa) is veruit het meest verklarend en er is daarom geen reden om een extra eigenschap mee te fitten. Dit maakt het model ook simpeler. Gemiddelde massa is een bekende en logische stofeigenschap om voor RO de verwijdering te voorspellen (Yangali-Quintanilla et al. 2009; Verliefde et al. 2008). Vluchtigheid is een gerelateerde eigenschap waarschijnlijk vanwege de grootte van het molecuul: hoe kleiner hoe vluchtiger, en dus ook hoe moeilijker te verwijderen met behulp van een membraan. Waarom sigma bonds op plek 3 staat, valt niet goed te verklaren. C-H bindingen zijn bijvoorbeeld sigmabindingen. C=C bevat zes sigma- en één pi-binding (zie Figuur 6).



Figuur 6: sigma- en pi-bindingen in C2H4 (https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3A_Introductory_Chemistry_(CK-12)/09%3A_Covalent_Bonding/9.18%3A_Sigma_and_Pi_Bonds)

Misschien kun je stellen dat met meer sigma-bindingen en dubbele bindingen het molecuul vaak compacter kan zijn. Heel erg voor de hand liggend is deze associatie niet.



Figuur 7: ***Verwijdering van stoffen in relatie tot log average mass. Zie Vergelijking 5 voor de formule voor de blauwe punten. Dit zijn de waarden uit de vuistregel.***

$$V_{ro} = \frac{100}{1+e^{-2.95\cdot(\log(Mass)-4.63)}}$$ Vergelijking 5

De eigenschap is gebaseerd op data van een enkel bedrijf. De eigenschap molmassa staat wel bekend als bepalende eigenschap in RO zuivering. De precieze vorm van de vuistregel zal geverifieerd moeten worden met een onafhankelijke dataset.

4. Discussie en overzicht van de afgeleide vuistregels

In Tabel 8 staat een overzicht van de afgeleide vuistregels.

Tabel 8. Overzicht van Technieken waar een vuistregel voor is afgeleid.

Zuiveringniveau	Techniek	Eigenschap	R^2 vuistregel-meetwaarden	P-waarde vuistregel-meetwaarden
Conventioneel	Coagulatie flocculatie	logKow	0.56	2.2e-16
Extra	Actieve kool	logKoa	0.17	5.67e-15
Extra	Oxidatie	Aromatische bindingen	0.41	5.34e-16
Geavanceerd	Geavanceerde oxidatie	Ring aantal 3D	0.59	2.2e-16
Geavanceerd	Omgekeerde osmose	logavMassa	0.48	2.2e-16

Stoffen op één eigenschap verbinden met een zuiveringsefficiëntie zal niet voor elke stof even goed gelden. Benzeen bijvoorbeeld is vluchtig (lage logKoa) en zal volgens de vuistregel slecht adsorberen op actieve kool, maar kan in de praktijk heel goed op actieve kool worden geadsorbeerd omdat de ringen in de stof een goede interactie geven met het kooloppervlak. Dit zijn verbeteringen aan de vuistregels die nog kunnen worden aangebracht, bijvoorbeeld door een tweede eigenschap mee te nemen in de vuistregel of voor bepaalde groepen een andere data-gedreven vuistregel af te leiden, speciaal voor de stofgroep.

Voor beide typen ‘extra’ zuivering waar in de DPWE set data voor beschikbaar zijn, actieve kool en oxidatie, blijkt dat de verwijdering erg variabel is. Dit kan aan de verschillende stoffen maar ook aan de diverse geteste omstandigheden liggen. Een enkele stof kan variëren van weinig tot veel verwijdering. De stoffen zijn getest in verschillende waterbedrijven onder verschillende condities. Daarnaast kunnen er verschillen optreden doordat verschillende datasets zijn gebruikt die bijvoorbeeld betrekking hadden op een ander type actieve kool (geproduceerd uit een andere grondstof), of kool met een andere beladingsgraad. Dat maakt dat een vuistregel per definitie veel ruis zal kennen; er is grote spreiding in de praktijk in verwijderingsefficiënties binnen een stof, terwijl de vuistregel op basis van en eigenschap altijd slechts een enkele waarde geeft.

De berekende zuiveringsefficiënties zijn gebaseerd op een algemene en grove vuistregel gebaseerd op de stofeigenschappen. Elke afzonderlijke drinkwaterzuiveringsinstallatie zal de zuiveringstechniek voor de omstandigheden hebben geoptimaliseerd, waardoor de zuivering afhankelijk van omstandigheden slechter of beter kan gaan dan berekend via de vuistregel (die een gemiddelde waterkwaliteit aangeeft op basis van stofeigenschappen). Daarnaast zijn de vuistregels zijn afgeleid voor data van een beperkt aantal drinkwaterbedrijven. Het is mogelijk dat de curves in andere datasets anders lopen. Door een ander gehalte of andere samenstelling natuurlijk organisch materiaal kan de relatie tussen adsorptie en logKow bij bijvoorbeeld de conventionele zuivering net iets anders uitvallen.

De regels moeten daarom gezien worden als een eerste aanzet om stofeigenschappen te linken aan zuiveringsefficiëntie. Met het doel om de waterkwaliteit in termen van zuiveringsinspanning te bepalen. Ze zullen bestendigd moeten worden door validatie met een of meerdere onafhankelijke datasets. Dit geldt met name voor filtratie over actieve kool waar een eigenschap naar voren kwam (logKoa) die normaal niet geassocieerd is met de zuivering. Daarnaast zal gevalideerd moeten worden of bepaalde groepen van stoffen misschien andere curves volgen dan de algemene curves uit de vuistregels die hier vastgesteld worden. PFAS zijn hierbij bijvoorbeeld een relevante groep stoffen. De vuistregels zijn vastgesteld op een gelimiteerde set van data. Dat maakt dat data-gedreven vuistregels ook minder robuust kunnen zijn. Als de relatie tussen een eigenschap van een stof en het gedrag van die stof in een zuiveringsproces bekend is uit de literatuur, is er minder reden tot twijfel. Voor minder logische stofeigenschappen zal een onafhankelijke dataset kunnen bevestigen of deze eigenschap inderdaad ook in de nieuwe dataset significant geassocieerd is met de zuiveringsefficiëntie van de stof. Vooral logKoa als vuistregel voor actieve kool is eerder niet gerapporteerd.

Het voordeel van de vuistregels is dat deze een data gedreven, simpele indicatie voor verwijderbaarheid op basis van stofeigenschappen vormen. Deze kunnen gebruikt worden om stoffen waar geen zuiveringsdata voor bestaan te prioriteren op hun verwijderbaarheid. Tot nadere verificatie kunnen de vuistregels gezien worden als een ‘beste optie beschikbaar’ voor een eerste inschatting van de verwijderbaarheid van veel stoffen tegelijk. Voor een precieze voorspelling van de verwijderbaarheid voor specifieke zuiveringsinstallaties, kan daarna bijvoorbeeld de tool ‘AquaPriori’ (KWR, 2017) ingezet worden. AquaPriori wordt op dit moment verder ontwikkeld voor meerdere technieken en zeer veel stoffen. Dit model houdt ook rekening met onderlinge interacties van stoffen en bijvoorbeeld natuurlijk organisch materiaal, wat in de vuistregels niet gebeurt.

Referenties

Breiman, L. (2001) Random Forests, Machine Learning 45(1), 5-32.

Cao, Y., A. Charisi, L.-C. Cheng, T. Jiang, and T. Girke (2008) ChemmineR: A Compound Mining Framework for R. Bioinformatics 24 (15): 1733–4. https://doi.org/10.1093/bioinformatics/btn307.

Finizio, A., Mackay, D., Bidleman, T., Harner, T. (1997) Octanol-air partition coefficient as a predictor of partitioning of semi-volatile organic chemicals to aerosols, Atmospheric Environment,Volume 31, Issue 15, Pages 2289-2296

García, L., J. C. Leyva-Díaz, E. Díaz and S. Ordóñez (2021). "A review of the adsorption-biological hybrid processes for the abatement of emerging pollutants: Removal efficiencies, physicochemical analysis, and economic evaluation." Science of the Total Environment 780.

Jenkin, M. E., R. Valorso, B. Aumont, M. J. Newland and A. R. Rickard (2020). "Estimation of rate coefficients for the reactions of 3with unsaturated organic compounds for use in automated mechanism construction." Atmospheric Chemistry and Physics 20(21): 12921-12937.

Jenkin, M. E., R. Valorso, B. Aumont, A. R. Rickard and T. J. Wallington (2018). "Estimation of rate coefficients and branching ratios for gas-phase reactions of OH with aromatic organic compounds for use in automated mechanism construction." Atmospheric Chemistry and Physics 18(13): 9329-9349.

Jin, H., D. Liu, J. Zou, J. Hao, C. Shao, M. Sarathy and A. Farooq (2020). "Chemical kinetics of hydroxyl reactions with cyclopentadiene and indene." Combustion and Flame 217: 48-56.

KWR water (2017) AquaPriori: a priori het verwijderingsrendement bepalen. Rapport KWR 2017.027 D. Vries, B. Wols, M. W. Korevaar, E. Vonk https://www.kwrwater.nl/projecten/aquapriori/

Korotta-Gamage, S. M. and A. Sathasivan (2017). "A review: Potential and challenges of biologically activated carbon to remove natural organic matter in drinking water purification process." Chemosphere 167: 120-138

Mansouri K, Grulke CM, Judson RS, Williams AJ (2018) OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform. 10(1):10. doi: 10.1186/s13321-018-0263-1

Meylan, William & Howard, Philip. (2005). Estimating octanol–air partition coefficients with octanol–water partition coefficients and Henry's law constants. Chemosphere. 61. 640-4. 10.1016/j.chemosphere.2005.03.029.

O'Boyle, N.M., Banck, M., James, C.A. et al. (2011) Open Babel: An open chemical toolbox. J Cheminform 3, 33 https://doi.org/10.1186/1758-2946-3-33

Pronk, T. E., Hofman-Caris, R. C. H. M., Vries, D., Kools, S. A. E., ter Laak, T. L., Stroomberg, G. J. (2021) A water quality index for the removal requirement and purification treatment effort of micropollutants. Water Supply 21 (1): 128–145. doi: https://doi.org/10.2166/ws.2020.289

Sonntag, c. v. and U. V. Gunten (2012). Chemistry of ozone in water and wastewater treatment; from basic principles to applications. London, IWA publishing.

US EPA (2012) Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11. United States Environmental Protection Agency, Washington, DC, USA.

Verliefde, A. R. D., S. G. J. Heijman, E. R. Cornelissen, G. L. Amy, B. Van Der Bruggen and J. C. Van Dijk (2008). "Rejection of trace organic pollutants with high pressure membranes (NF/RO)." Environmental Progress 27(2): 180-188.

Yangali-Quintanilla, V., A. Verliefde, T. U. Kim, A. Sadmani, M. Kennedy and G. Amy (2009). "Artificial neural network models based on QSAR for predicting rejection of neutral organic compounds by polyamide nanofiltration and reverse osmosis membranes." Journal of Membrane Science 342(1-2): 251-262.