Miljøministeriet
Miljøstyrelsen

Dynamisk værktøj til vurdering af grundvandets PFAS-sårbarhed 'proof of concept'

Miljøprojekt, nr. 2319

Oktober 2026

Udgiver: Miljøstyrelsen

Redaktion:
Christian Nørby Friis Sørensen & Bastian Germundsson, COWI A/S

Kvalitetssikring:
Bo Tegner Bay, COWI A/S

ISBN: 978-87-7564-163-5

Miljøstyrelsen offentliggør rapporter og indlæg vedrørende forsknings- og udviklingsprojekter indenfor miljøsektoren, som er finansieret af Miljøstyrelsen. Det skal bemærkes, at en sådan offentliggørelse ikke nødvendigvis betyder, at det pågældende indlæg giver udtryk for Miljøstyrelsens synspunkter. Offentliggørelsen betyder imidlertid, at Miljøstyrelsen finder, at indholdet udgør et væsentligt indlæg i debatten omkring den danske miljøpolitik.

Indhold

1.	**Resume**	**4**
1.1	Dansk	4
1.2	English	4
2.	**Indledning**	**6**
2.1	Formål	6
2.2	Strategi og metodebeskrivelse	6
3.	**Metode**	**8**
3.1	Software og programmeringssprog	8
3.2	Målværdi	8
3.3	Anvendt data	11
3.3.1	Kilder	11
3.3.2	Parametre	12
3.4	Datavask	15
3.5	Modellen	16
3.6	Dynamisk opdatering	17
4.	**Resultater**	**19**
4.1	Nøgletallene for modellen	19
4.2	Parameterafhængigheder	19
4.3	Udstilling og forankring	21
4.4	Estimater	22
5.	**Konklusion**	**24**
6.	**Referencer**	**25**

1. Resume

1.1 Dansk

Rapporten præsenterer et projekt, der har udviklet en machine learning-model til at vurdere sandsynligheden for tilstedeværelsen af PFAS i grundvandet. Ved at bruge offentlige data fra blandt andet GEUS' Jupiter-database og anvende kunstig intelligens, kan modellen estimere sandsynligheden for tilstedeværelse af PFAS i forskellige grundvandsmagasiner. Formålet er at hjælpe kommuner, regioner og andre interessenter ved at stille et indledende screenings-værktøj til rådighed ved undersøgelses-, oprydnings- og anlægsindsatser.

Projektet bygger på et eksisterende værktøj udviklet af COWI til vurdering af grundvandets sårbarhed over for pesticider. Modellen bruger de samme 11 parametre, såsom magasintype, vandstand, nettoinfiltration og arealanvendelse, men er tilpasset til at estimere PFAS-sårbarheden. Der er anvendt algoritmen 'XG Boost Classifier' for at opnå de bedste prædiktive resultater. Modellen har en høj nøjagtighed på 83% og præcision på 79%, men også en tendens til at underestimere sandsynligheden for fund. På trods af usikkerheder og bias, anses modellens estimater for at være anvendelige i praksis, men bør suppleres med faglige vurderinger og lokale data.

Efterhånden som nye data indsamles, kan modellen let gentrænes, hvilket sikrer, at estimaterne altid er baseret på de nyeste data og modellen forbedres således løbende i takt med, at mere data bliver tilgængeligt.

Værktøjet er tilgængeligt gennem Danmarks Miljøportal, hvor implementeringen sikrer bred adgang og nem anvendelse af værktøjet. Modellen udgør et værdifuldt værktøj til vurdering af PFAS-sårbarhed i grundvandet og forventes at kunne udgøre en vigtig ressource for fremtidige beslutninger om f.eks. miljøbeskyttelse og offentlige sundhedsindsatser i Danmark.

1.2 English

This report presents a project that has developed a machine learning model to assess the probability of PFAS presence in groundwater. By using publicly available data, including the GEUS Jupiter database, and applying artificial intelligence, the model estimates the likelihood of PFAS being present in different groundwater aquifers. The purpose is to support municipalities, regions, and other stakeholders by providing an initial screening tool for investigation, remediation, and construction activities.

The project builds on an existing tool developed by COWI for assessing groundwater vulnerability to pesticides. The model uses the same 11 parameters, such as aquifer type, water level, net infiltration, and land use, but has been adapted to estimate PFAS vulnerability. The 'XGBoost Classifier' algorithm has been applied to achieve strong predictive performance. The model demonstrates an accuracy of 83% and a precision of 79%, but also shows a tendency to underestimate the probability of PFAS detections. Despite uncertainties and inherent biases, the model's estimates are considered applicable in practice, provided they are supplemented with professional judgement and local data.

As new data becomes available, the model can be retrained with limited effort, ensuring that estimates are continuously updated and improved over time.

The tool is available through the Danish Environmental Portal (Danmarks Miljøportal), ensuring broad access and ease of use. The model represents a valuable tool for assessing PFAS

vulnerability in groundwater and is expected to contribute to future decision-making related to environmental protection and public health in Denmark.

2. Indledning

PFAS (per- og polyfluoralkylstoffer) består af en stor gruppe syntetisk fremstillede fluorstoffer, der har været brugt siden begyndelsen af 1950'erne. Stofferne er stabile, svært nedbrydelige og derudover er de meget vandopløselige, hvorfor de kan ende i vores grundvand. P.t. viser estimater fra regionerne, at op imod 15.000 lokaliteter har haft en aktivitet, der kan have resulteret i forurening med PFAS og som kan have påvirket miljøet. Dette stiller regionerne i et dilemma, for hvordan fordeles ressourcerne til undersøgelsesaktiviteter bedst, hvilke ejendomme udgør den største risiko og er der nogle lokaliteter, der kan frasorteres alene på baggrund af deres geografiske placering?

Ved machine learning udnyttes brugen af et stort antal definerende parametre, mange flere end den menneskelige hjerne kan processere på en gang. Det giver mulighed for at finde mønstre, afhængigheder og korrelationer mellem mange forskellige parametre. Derved kan der opstilles modeller, som kan anvendes til at estimere eller udpege, hvordan et givent parameter vil forhold sig i områder, hvor kun de andre parametre er kendt. Med denne metode i hånden har projektet udviklet et værktøj, som kan hjælpe med at finde nålen i høstakken eller sagt på en anden måde; belyse hvilke potentielt forurenede lokaliteter der ligger i områder, hvor PFAS-forbindelserne sandsynligt vil føre til fund i grundvandet. Ved at benytte offentligt tilgængeligt data herunder GEUS' Jupiter database til at træne modellen med, opnås en model med de nyeste og mest opdaterede data vedrørende PFAS og ved at lade modellen have en dynamisk tilgang vil modellen kunne holde sig selv opdateret, som der løbende kommer nye data til.

Med en landsdækkende geografisk udbredelse, foruden Bornholm og andre ikke brofaste øer, vil værktøjet kunne finde anvendelse hos ikke kun regionerne, men også forsyninger, kommuner og private og derfor har en forankring hos Danmarks Miljøportal efterstræbt. Værktøjet kan således i dag tilgås gennem blandt andet 'Arealinformation'.

2.1 Formål

Grundlæggende er formålet med projektet at udvikle en machine learning model på baggrund af eksisterende data og målinger, som kan udpege områder i de 13 danske grundvandsmagasiner, hvor der er størst sandsynlighed for at påvise PFAS. Der er konkret fokuseret på de 22 stoffer, der i dag indgår i standardanalyseprogrammerne. Det er ikke projektets formål at vurdere, hvorvidt PFAS udgør en risiko overfor grundvandet.

Projektet står på skuldrene af et eksisterende og offentligt tilgængeligt værktøj, som COWI har udviklet til vurdering af pesticidsårbarhed, hvorfra blandt andet tilgangen og parametre genbruges. Det nye er, at PFAS-modellen er dynamisk, så den altså kan holde sig opdateret sideløbende med, at nyt data kommer til.

2.2 Strategi og metodebeskrivelse

Metodisk er kunstig intelligens, konkret machine learning, blevet brugt som et avanceret regressions- eller prædiktionsværktøj. Dette gøres ved at opsætte og træne en model på eksisterende data, som deraf kan estimere sandsynligheden for at finde PFAS i et givent grundvandsmagasin i et konkret geografisk punkt. Det vurderes at pesticiderne og PFAS stoffernes sårbarhed deler flere definerede parametre og derfor er disse genbrugt.

Projektet har indeholdt et 'proof of concept' element for at undersøge, hvorvidt det er muligt at etablere en model, som kan gentræne sig selv, når nye data kommer til og derved konstant være opdateret. Det er nemlig essentielt på et område som PFAS, hvor ny viden og nye data kommer ind med stor hast.

Machine learning er en familie af algoritmer og modeller indenfor kunstig intelligens, der udmærker sig ved deres evne til at finde komplekse mønstre i store datamængder. For at kunne træne en model skal man have et datasæt, hvor man i hvert punkt kender de beskrivende parametre, såvel som den parameter man ønsker at træne modellen i at identificere – i dette tilfælde, om der er målt indhold af PFAS eller ej. Den trænede model kan herefter estimere sandsynligheden for at påvise stofferne i punkter, hvor der ikke er foretaget målinger. Med andre ord kan man ud fra et repræsentativt datasæt med målinger samt beskrivende parametre træne en model, der kan generaliseres til alle punkter, hvor de beskrivende parametre kendes og estimere sandsynligheden for fund i disse. Antallet af PFAS-data anvendt til at træne modellen er afgørende for præcisionen af værktøjet.

Da området indenfor kunstig intelligens er i en rivende udvikling, udvikles nye modeller og algoritmer næsten dagligt. Projektet har taget udgangspunkt i det eksisterende pesticidværktøj, udviklet af COWI, og anvendt de nyeste og mest optimale metoder til at udføre opgaven.

3. Metode

3.1 Software og programmeringssprog

Selve modellen er programmeret udelukkende i Python 3.11. Python er et såkaldt 'high-level' programmeringssprog, der blandt andet udmærker sig ved at være det et af de førende sprog indenfor databehandling, -analyse og machine learning.

En af årsagerne til at Python er blevet så populært, er at det er et såkaldt open source-sprog, hvilket vil sige at sproget udvikles og vedligeholdes af et stort netværk af frivillige. Når der kommer nye modeller og algoritmer, vil de derfor hurtigt være at finde i Python, og evt. fejl opdages og udbedres hurtigt. Dette sker ved, at der udgives såkaldte pakker, man kan importere, hvis man ønsker at benytte en given funktionalitet. Det kunne være machine learning algoritmer såsom XG Boost, men er ikke begrænset hertil, og kunne f.eks. også være pakker til at hente data fra databaser, eller til at visualisere data.

I Python-koden er Pandas-modulet benyttet til at håndtere data, og XG Boost modulet benyttes til at modellere PFAS data og estimere sårbarheden. Data opbevares i to forskellige databaser på en MSSQL server.

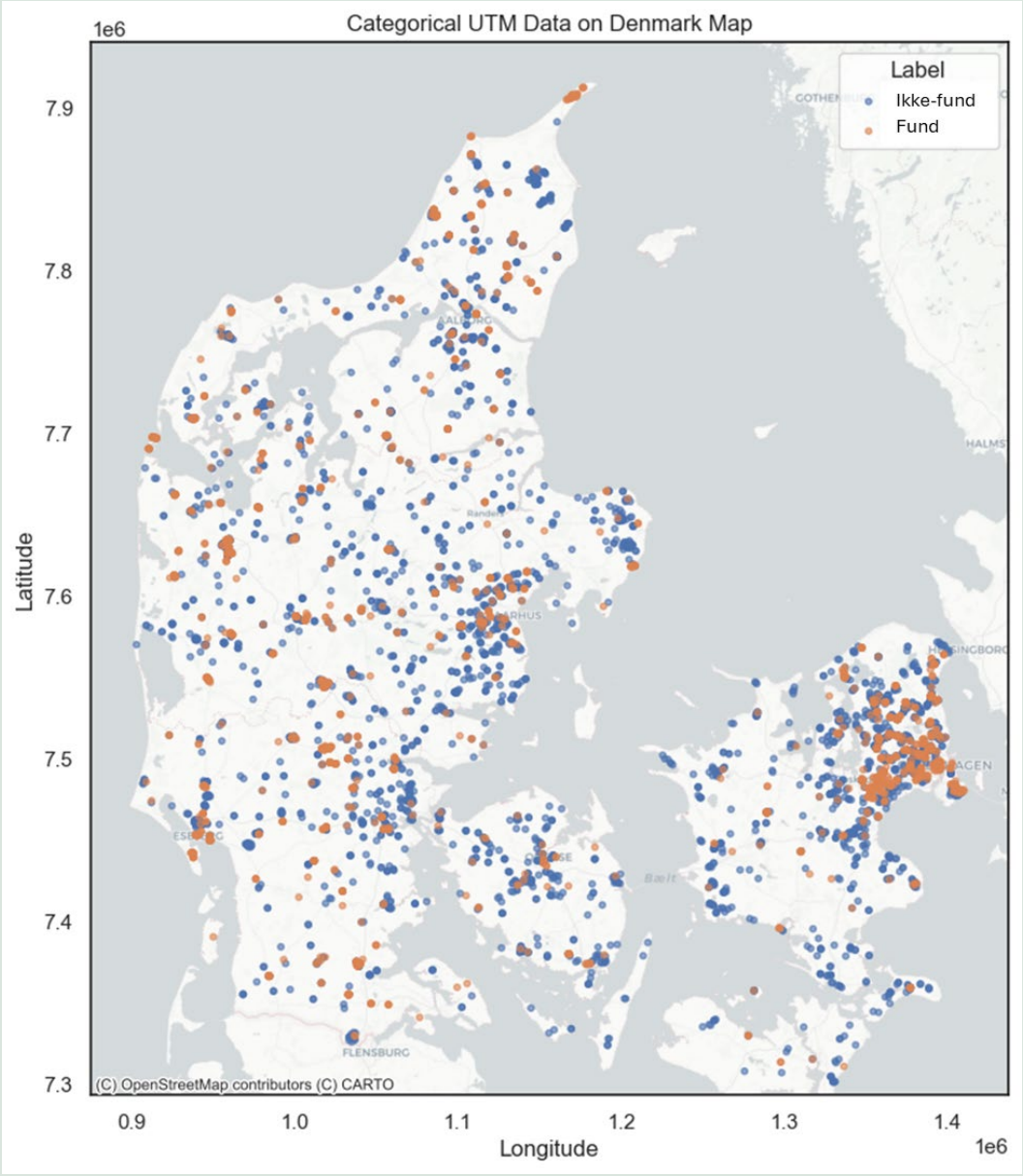
3.2 Målværdi

Målværdien er den værdi, som modellen skal trænes til at kunne estimere. I dette projekt er målværdien derfor, om der er i et givent geografisk punkt i Danmark, findes PFAS eller ej. Målværdierne er derfor defineret som klasserne 'fund' eller 'ikke-fund' og helt konkret som enten 1 eller 0. Tilgangen er baseret på reglen:

Et indtag vurderes at være påvirket af PFAS, hvis der er detekteret PFAS i seneste analyse, eller hvis der er detekteret PFAS mere end en gang historisk.

Årsagen til reglen er, at når man vurderer, hvorvidt et boringsindtag er påvirket af forurening, kan man ikke udelukkende se på de seneste analyseresultater. Dels har den seneste prøvetagning ikke altid været analyseret for alle stoffer, dels er der usikkerhed tilknyttet prøvetagning og analyse.

Datagrundlaget for målværdierne er modtaget som såkaldte bacpac filer fra Danmarks Miljøportal og indeholder analysedata for alle PFAS-prøver i GUES' Jupiter database. Grundet den dynamiske tilgang vil datamængden vokse med etablering af nye boringer og udtagning af nye vandprøver i nye og eksisterende filtre/indtag. For den trænede model var der PFAS-analyser i 5018 forskellige boringsindtag og dette udgør den grundlæggende datamasse. Af de 5018 analyser er der påvist fund, på baggrund af ovenstående regel, i 37 % af prøverne. Den geografiske spredning af indtag, hvor der er analyseret for PFAS, er set af FIGUR 1. Det er her tydeligt at de fleste data er koncentreret omkring de større danske byer, især København, Århus, Odense og trekantområdet, men dog ses det, at der er analyser fra hele Danmark uden betydelige geografiske huller. Det er fra figuren ikke entydigt klart, at fundene af PFAS skulle være koncentreret på baggrund af geografi, men disse findes også i hele landet.

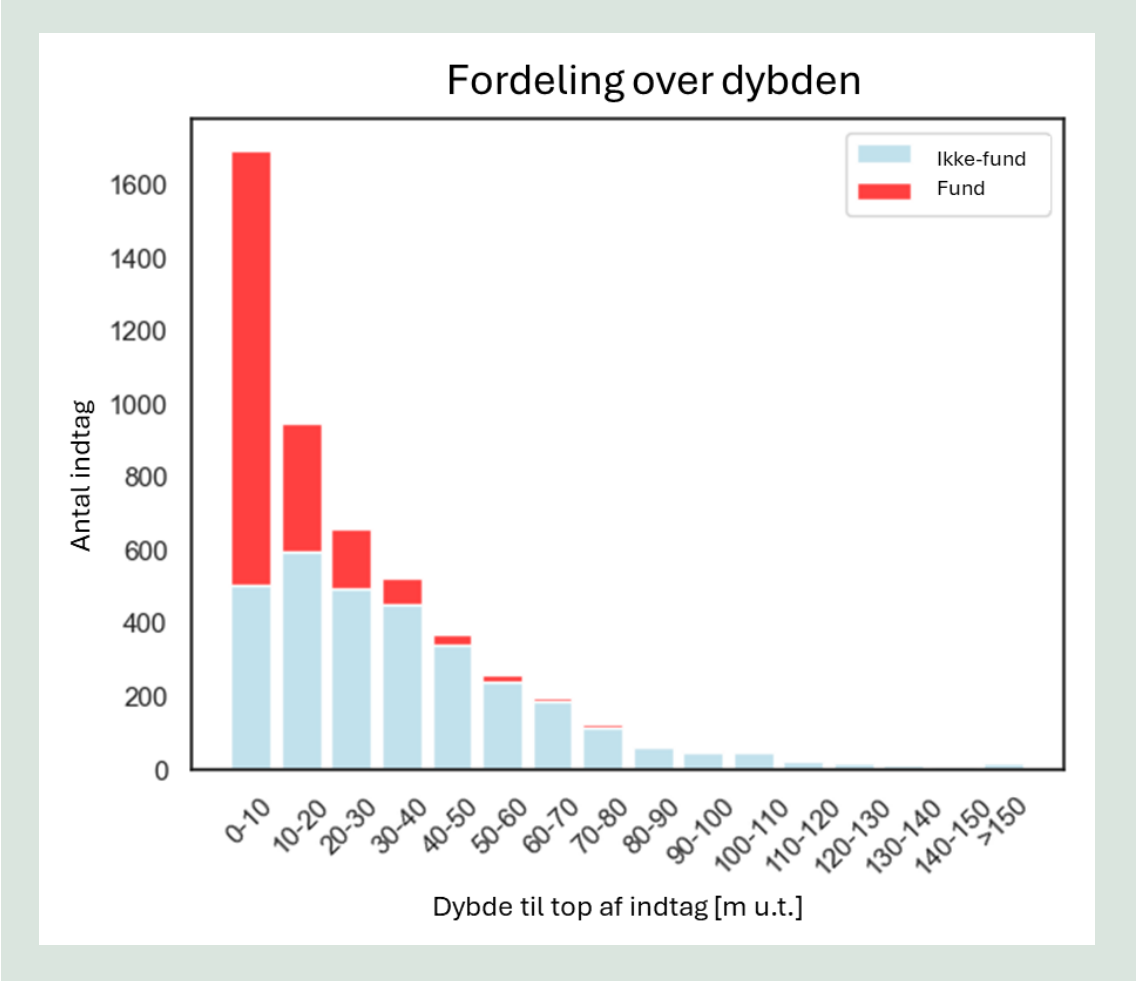


FIGUR 1. Geografisk spredning af analyser for PFAS. De blå punkter repræsenterer ikke-fund mens de orange repræsenterer fund. Grundlæggende er der en udmærket geografisk spredning af både analyser og fund. Der er selvfølgeligt set en større koncentration omkring de større danske byer.

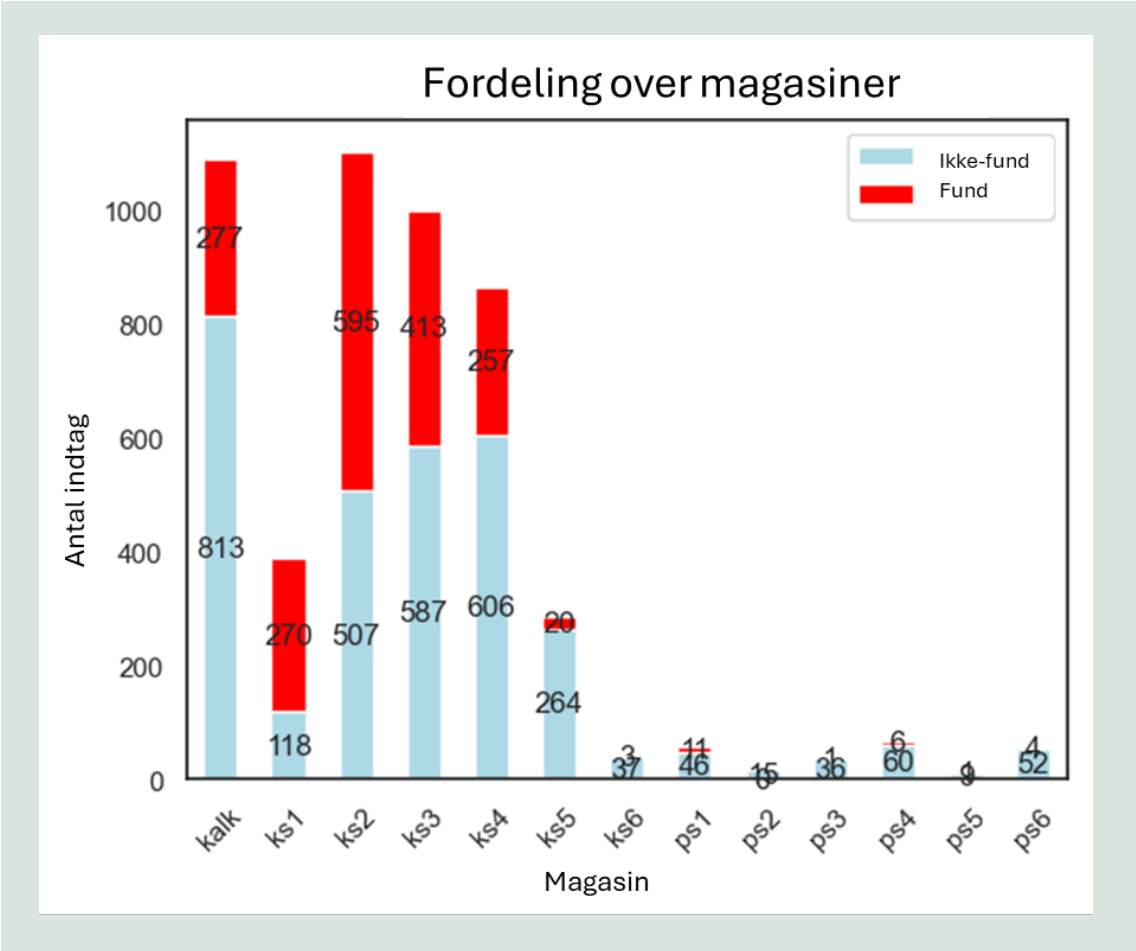
Fordeling af analyser er også undersøgt for både dybder og magasiner, se hhv. FIGUR 2 og FIGUR 3. For dybder gælder det, at langt de fleste analyser stammer fra indtag, som er er filtersat terrænnært; 0-10 m u.t. Der findes også analyser fra større dybder, men antallet falder med dybden, således der er relativt få indtag dybere end 50 m u.t. I forhold til fund af PFAS er det tydeligt at langt de fleste fund, er set ved lave dybder. For dybden 0-10 m u.t. er der således flere analyser, hvor der er fundet PFAS, end indtag hvor der ikke er fundet PFAS. Dette ændrer sig med dybden, således der ved stigende dybde er påvist PFAS i en mindre og mindre andel af indtagene. Det vurderes at være et klart billede på grundvandets alder, som stiger med dybden og reelt vil vandet i disse filtre være ældre end anvendelsen af PFAS-stofferne.

For magasintilknytning gælder, at langt de fleste analyser er udtaget fra indtag i de grundvandsmagasiner, som kaldes KS samt kalk. KS dækker de kvartære sandenheder og disse er, sammen med kalk, nationalt udbredte og findes både i Jylland, på Fyn og Sjælland. PS dækker de prækvartære sandenheder og disse er i modsætning til KS ikke nationalt udbredte og findes kun i dele af Jylland. Overvægten af indtag i KS og kalk skyldes derfor grundlæggende, at disse magasiner findes i langt større dele af Danmark. For en detaljeret gennemgang af grundvandsmagasiner i Danmark (og DK-modellen) samt udbredelse heraf henvises til publikationen af GEUS 'Afgrænsning af de danske grundvandsforekomster' /2/. Der findes overordnet 6 grundvandsmagasiner i kvartære sandenheder og 6 grundvandsmagasiner i prækvartære sandheder.

Fordelingen mellem fund og ikke-fund i de 6 KS magasiner viser, at der er en tendens til, at der bliver færre og færre fund af PFAS jo dybere man kommer i formationen (stigende KS betegnelse). Dette er igen vurderet at afhænge af grundvandets alder, som grundlæggende vil være højere i KS6 end i KS1. De kvartære og prækvartære sandenheder er nemlig alle efterfulgt af et tal fra 1-6. Disse tal skal ses som en kronologi i aflejringerne, således 1 er mest terrænnær og 6 mindst. PS6 vil derfor oftest have en større dybde end PS1 og derfor en højere grundvandsalder.



FIGUR 2. Fordeling af analyser i forhold til dybde af indtaget. Det er tydeligt at antallet af indtag falder med dybden og at der er langt flest data 0-10 m u.t. Der er også ligeledes her langt de fleste indhold af PFAS er påvist.



FIGUR 3. Fordeling af analyser i forhold til magasintilknytning. Da Danmark er opbygget geologisk som det er, vil langt de fleste analyser være udtaget i kalk eller en af de kvartære sandenheder (KS1-KS6). Der er grundlæggende ikke meget data i de prækvartære sandenheder (PS1-PS6), da disse kun findes i dele af Jylland.

3.3 Anvendt data

Udgangspunktet for projektet har været at stå på skuldrene af pesticidværktøjet og helt grundlæggende at anvende det store arbejde, der allerede er lavet i forhold til udvælgelse af parametre og fund af diverse kilder. Det er nemlig vurderet, at det der er afgørende for grundvandets sårbarhed overfor pesticider, også er afgørende for PFAS-forbindelserne. F.eks. forventes det, at dybden til grundvandet eller områdets arealanvendelse, for bare at nævne to eksempler, også er vigtige parametre for PFAS. Derfor har udgangspunktet været at undersøge, hvordan modellen præsterer ved anvendelse af samme parametre på PFAS-data og så herefter afgøre, om der burde laves en reevaluering af de valgte parametre.

3.3.1 Kilder

Danmark er et foregangsland i forhold til at gøre data offentligt tilgængelige og det er den primære årsag til, at projekter som dette kan lade sig gøre. De vigtigste kilder til data for dette projekt er GEUS' boringsdatabase Jupiter sammen med den nationale hydrologiske model (DK-modellen). Desuden ligger COWI inde med flere egne databaser med lignende information om koordinater, geologi, vandstand og analyser. DK-modellen indeholder processerede data, som anvendes til hydrologisk modellering. Modellen dækker hele landet og kan derfor give de samme data for alle punkter. Modellen indeholder lagflader med nettoinfiltration, grundvandsdannelse, top- og bundkote af hver lagflade, beregnede vandtryk i hvert magasin

mv. Modellen er senest opdateret i 2019, men indeholder data, som ikke varierer nævneværdigt over tid. Projektet har genbrugt modellen fra pesticidværktøjet, som har en diskretisering på 100 x 100 meter, da diskretiseringen i den offentligt tilgængelige model er højere.

Derudover er der anvendt følgende litteratur fra forskningsverdenen:

Arealanvendelse: Levin, G., Iosub, C.-I. & Jepsen, M.R. 2017. Basemap02. Technical documentation of a model for elaboration of a land-use and land-cover map for Denmark. Aarhus University, DCE – Danish Centre for Environment and Energy, 64 pp. Technical Report from DCE – Danish Centre for Environment and Energy No. 95

Makroporestrømning: Iversen, Bo & Børgesen, Christen & Lægdsmand, Mette & Greve, Mogens & Heckrath, G. & Kjaergaard, Charlotte. (2011). Risk Predicting of Macropore Flow using Pedotransfer Functions, Textural Maps, and Modeling. Vadose Zone Journal. 10. 1185. 10.2136/vzj2010.0140.

Humusindhold: Adhikari, Kabindra. (2013). Soil mapping in Denmark using digital soil mapping Techniques.

Dybde til redoxgrænsen: Koch, Julian & Stisen, Simon & Refsgaard, Jens & Ernstsen, Vibeke & Jakobsen, Peter & Højberg, Anker. (2019). Modeling Depth of the Redox Interface at High Resolution at National Scale Using Random Forest and Residual Gaussian Simulation. Water Resources Research. 55. 10.1029/2018WR023939.

3.3.2 Parametre

Grundlæggende gælder, at en model skal bruge lige så mange beskrivende parametre som et menneske ville skulle bruge, for at kunne lave en vurdering og udgangspunktet har været de samme 13 parametre, som anvendt i pesticidværktøjet. Beskrivelsen af parametrene er enslydende med den tekniske rapport /1/, som ligger til grund for pesticidværktøjet:

Parameter 1: Topografi

For datasættet er terrænoverfladen i muligt omfang trukket fra oplysninger i kildedatabaserne. Alternativt er terrænoverfladen trukket ud fra den terrænmodel, som er anvendt i DK-modellen. For beregningsgriddet er topografien udelukkende trukket fra DK-modellen.

Terrænmodellen er opskaleret til DK-modellen, så den har en diskretisering på 100 x 100 m.

Parameter 2: Magasintype

DK-modellen er opbygget med lagflader, som er sammenhængende, men som dog nogle steder har en tykkelse på 0 m.

Bornholm har imidlertid en helt anderledes geologi med grundfjeld, hvilket gør udtræk fra DK-modellen ubrugeligt i denne sammenhæng. DK-modellen dækker ligeledes heller ikke de mindre øer i Danmark. Derfor dækker projektet desværre ikke Bornholm og andre små øer, såsom Samsø og Læsø.

For at koble boringsindtag til geologiske lag fra DK-modellen er hvert indtag tildelt navnet på det magasin, som toppen af indtaget befinder sig i.

I nogle tilfælde har boringerne en indtagsdybde, som ligger over terrænet fra DK-modellen. Dette kan skyldes fejl i data eller opskaleringen af terrænets diskretisering. Sådanne indtag er tilknyttet KL1, som er det øverste lerlag i DK-modellen.

DK-modellen har dannnet datagrundlag for denne parameter for både datasættet og beregningsgriddet.

Parameter 3: Vandstand
Data om vandstand/vandtryk findes kun i kildedatabaserne og har derfor været en begrænsende faktor for størrelsen af datasættet med målværdier.

For hvert indtag, hvortil der ligger kemiske analyser, er datoen for seneste vandprøve fundet. Derefter er den vandstandsmåling, som er taget tættest på datoen for seneste analyse, knyttet til indtaget. Dette gøres for at anvende data, som i størst muligt omfang beskriver den samme tilstand.

Der er lavet to tjek af vandstandsdata for at frasortere fejl:

Vandstandsmålingen skal være taget efter boringsdatoen.
Pejlingen (m u.t.) skal være over bund af indtag.
Vandstande angivet som både 'ro' og 'drift' er medtaget.

Parameteren er trukket fra det beregnede trykniveau fra DK-modellen til beregningsgriddet.

Parameter 4: Nettoinfiltration
Nettoinfiltrationen til både datasættet og beregningsgriddet stammer fra DK-modellen. Laget er baseret på forskellige parametre, såsom nedbør, befæstelsesgrad og fordampning for perioden 2000-2010.

For at frasortere fejldata er følgende tjek udført:

Punkter med negativ nedbør er sat til 0.
Punkter med nedbør på over 1000 mm/år er sat til 1000 mm/år.

Parameter 5: Tykkelse af umættet zone
Tykkelsen af den umættede zone angiver, hvor tykt det tørre jordlag er, dvs. hvor langt der er ned til øverste grundvandsspejl.

Tykkelsen af den umættede zone til både datasættet og beregningsgriddet er trukket ud fra DK-modellen i 100 x 100 m grid.

Parameter 6: Grundvandsdannelse
Grundvandsdannelsen til hvert magasin er i DK-modellen givet som infiltrationen fra et geologisk lag til et andet. Grundvandsdannelsen er derfor givet for det lag, hvori toppen af indtaget sidder.

Grundvandsdannelsen til magasinet til både datasættet og beregningsgriddet er trukket ud fra DK-modellen i 100 x 100 m grid.

Parameter 7: Dybde til magasin
For hvert indtag er dybden fra terræn i DK-modellen givet, til det lag, som toppen af indtaget befinder sig i. Dette gælder både for datasættet og beregningsgriddet.

Parameter 8: Akkumuleret lertykkelse over indtag
Til denne parameter findes der to kilder, dels de indrapporterede jordprøver og lagfølger i kildedatabaserne, dels de geologiske lagfølger, som trækkes ud fra DK-modellen.

Den samlede tykkelse af overliggende lerede jordarter er beregnet for hvert indtag for hver datakilde. I de indtag, som har jordprøver registreret, er den akkumulerede lertykkelse baseret på disse data, mens indtag uden registrerede jordprøver er knyttet til beregningen på data fra DK-modellen. Data til beregningsgriddet er ligeledes taget fra DK-modellen.

Hvis toppen af et indtag sidder i et lerlag (enten jordprøve data eller DK-model lagflader) er tykkelsen af dette lag kun beregnet fra top af indtag.

I nogle tilfælde er toppen af det øverste lag fra DK-modellen ikke kalibreret, hvilket vil sige at toppen af laget overstiger terrænet. I disse tilfælde er lertykkelsen sat til 0 m.

Parameter 9: Akkumuleret lertykkelse over indtag under redoxgrænsen
Denne parameter svarer til beregningen fra den akkumulerede lertykkelse over indtag (8), men regner kun lerlag under redoxgrænsen med (se parameter 10). Hvis redoxgrænsen er midt i et lerlag, er tykkelsen af dette lag kun beregnet op til denne grænse.

Parameter 10: Dybde til redoxgrænsen
Dybden til redoxgrænsen er taget fra GEUSs kort over redoxgrænsen (Koch et al.).

Kortet er bla. dannet ud fra jordprøveinformationer i Jupiter, kortlægning af lerindhold i topjorden, forskellige topografiske parametre, landskabstyper og lavninger. Diskretiseringen er 100 x 100 m.

Parameter 11: Makroporestrømning (K_jump)
Makroporestrømingen er beskrevet af Iversen et al. Parameteren K_jump angiver potentialet for strømning i overjordens makroporer. Jævnfør Iversen et al. beregnes parameteren som differencen imellem den mættede (Ks) og den nær-mættede (Kh) hydrauliske ledningsevne i overjorden.

Kildedata for Ks og Kh er udarbejdet for tre jordhorisonter A (0-30 cm), B (30-70 cm) og C (70-100 cm). Datasættet for horisont A har en diskretisering på 250 x 250 m, mens horisont B og C har en diskretisering på 500 x 500 m.

Der er anvendt et gennemsnit af de tre horisonter, hvorefter Kh er trukket fra Ks for at danne en parameter til beskrivelse af potentialet for makroporestrømning.

Parameter 12: Humusindhold
Hummusindholdet er beregnet af Adhikari et al., ligeledes for de tre horisonter (A, B og C) i den øverste meter af overjorden. Datasættet har en diskretisering på 30 x 30 m.

Der er i dette projekt anvendt et gennemsnit af de tre horisonter.

Parameter 13: Arealanvendelse
Arealanvendelsen er baseret på en kortlægning fra 2016 af Levin et al. Kortlægningen har dannet flere forskellige kort, hvorfra dette projekt har brugt data fra laget LU_01_2016.Kolonnen AGG_LU_Name_DK er anvendt som kategori. Dog er nogle af kategorierne grupperet efter deres betydning for grundvandets sårbarhed. De kategorier for arealanvendelse, so er anvendt er vist nedenfor.

De kategorier for arealanvendelse, som er anvendt, er følgende:

Bebyggelse
Ikke kortlagt
Industri

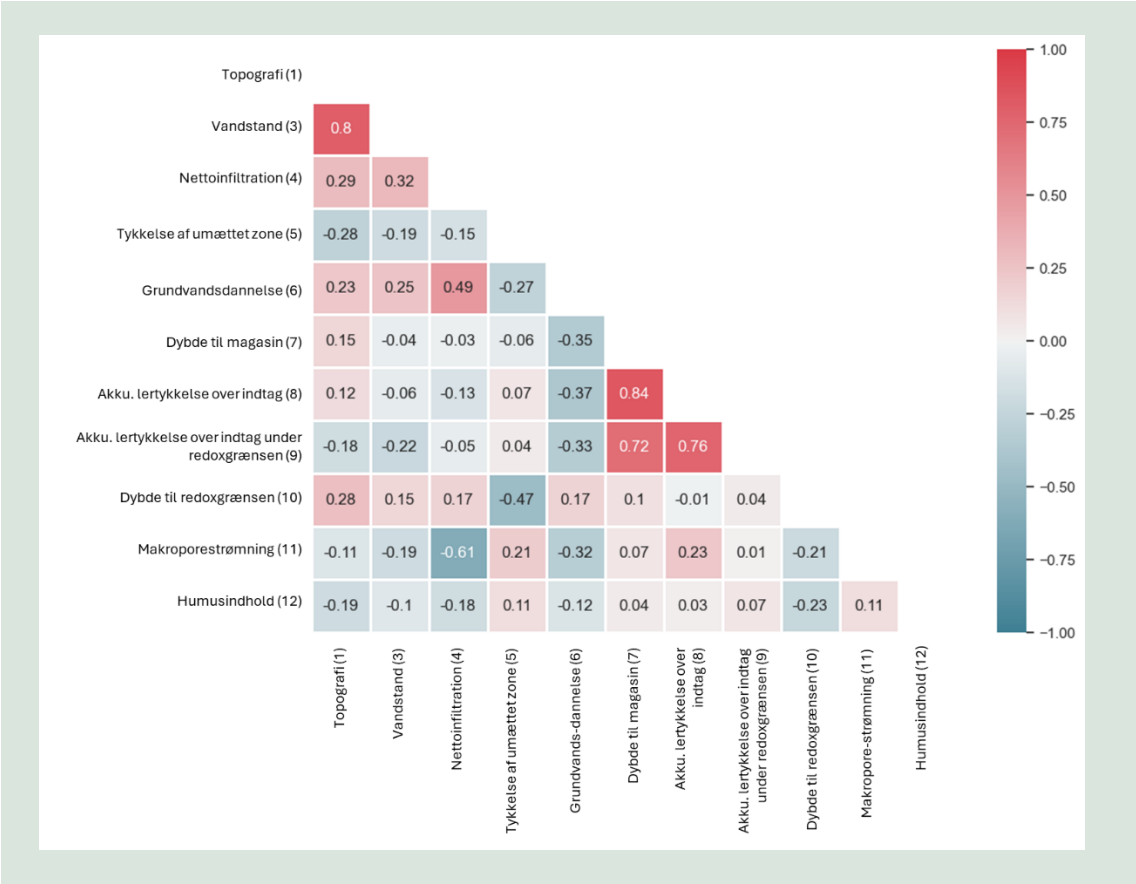
Jernbane
Landbrug ikke klassificeret
Landbrug intensivt
Landbrug ekstensivt
Lufthavn
Natur
Råstofområde
Sø
Skov
Vandløb
Vej

3.4 Datavask

Forud for modelleringen blev der foretaget en datavask, ved at tjekke de valgte parametre samt målværdierne for fejl og spredning for at sikre, at der udelukkende blev anvendt pålidelige data og at der ikke blev indført unødvendige fejlkilder.

Der er først og fremmest lavet histogrammer for alle numeriske parametre for at tjekke spredning på data. I samme omfang er PFAS-observationerne gennemgået for afvigelser og her er fjernet 10 observationer, da deres koordinater ikke stemte overens med dem fra DK-modellen.

FIGUR 4 nedenfor, viser korrelationsmatricen for de numeriske parametre. Magasintilknytning og arealinformation fremgår ikke af korrelationsmatricen, da begge er kategoriske parametre. Fra matricen kan parametrenes interne korrelation, enten positiv eller negativ undersøges og på baggrund heraf er det tydeligt, at flere parametre korrelerer tæt. Som det også var set fra analysen af parametrene for pesticidværktøjet, korrelerer typografi og vandstand tæt og derfor besluttes det at lade typografi udgå. Desuden er det set, at dybde til magasin, akkumuleret lertykkelse over indtag og akkumuleret lertykkelse over indtag under redoxgrænsen alle har en positiv korrelation. Da begge parametre omkring den akkumulerede lertykkelse har enslydende indhold, er det valgt at lade parameter 8; akkumuleret lertykkelse over indtag udgå.



FIGUR 4. Korrelationsmatrix for de numeriske parametre. Magasintilknytning og arealinformation indgår ikke, da de er kategoriske parametre. Matricen viser korrelationerne mellem parametrene, hvor en kraftig farve (enten rød eller blå) indikerer en enten positiv eller negativ korrelation og en svag farve indikerer let eller ingen korrelation. På baggrund heraf udgik to parametre.

3.5 Modellen

Af datasættet på nu 5.008 PFAS-observationer er ¼ af observationer blevet udtaget til validering og modellen er trænet på de resterende ¾ af datasættet. Dette er en gængs metode til at validere modellers nøjagtighed. Alternativet ville være at udtage nye prøver forskellige steder i Danmark, men det ville ikke give samme robuste datasæt til valideringen.

Som for pesticidværktøjet blev det besluttet, på baggrund af den indledende datavask og dataanalyse, at anvende en klassifikationsalgoritme til estimaterne. Hvor der tidligere blev anvendt 'Random Forest' algoritmen blev det besluttet at anvende klassifikationsmodellen 'XG Boost Classifier' til denne model. For at evaluere modellens anvendelighed anvendes et sæt metrikker herunder nøjagtighed, præcision og recall, hvilke er almindelige enheder, der anvendes til at evaluere ydeevnen af machine learning modeller. Disse metrikker måler, hvor godt modellen klarer sig i forhold til at estimere korrekte resultater. Grundlæggende gælder det at:

- Nøjagtighed er en metrik, der måler, hvor godt modellen klarer sig generelt. Det besvarer spørgsmålet: "Hvor mange af alle estimaterne var korrekte?" nøjagtighed er forholdet mellem antallet af korrekte estimater og det samlede antal estimater. Det er vigtigt at nævne, at nøjagtighed ikke tager højde for ubalancer mellem de forskellige klasser (fund af PFAS eller ikke-fund af PFAS) i dataene.

- Præcision er et mål for, hvor præcist modellen identificerer de positive tilfælde. Det besvarer spørgsmålet: "Hvor mange af de positive estimater var faktisk korrekte?" Med andre ord, er

det forholdet mellem antallet af korrekte positive estimater og det totale antal positive estimater. Høj præcision betyder, at modellen har en lav falsk positiv rate og er god til at klassificere de positive tilfælde korrekt.

- Recall er et mål for, hvor godt modellen identificerer de korrekte positive tilfælde. Det besvarer spørgsmålet: "Hvor mange af de faktiske positive tilfælde blev korrekt identificeret?" Med andre ord, det er forholdet mellem antallet af korrekte positive estimater og det totale antal faktiske positive tilfælde. Høj recall betyder, at modellen har en lav falsk negativ rate og fanger de fleste af de positive tilfælde.

Det er vigtigt at bemærke, at metrikkerne er indbyrdes forbundne, og at en høj værdi for den ene kan føre til en lavere værdi for en anden, især når der er ubalancer imellem de to klasser, som modellen prøver at estimere, her PFAS-sårbarhed eller ej.

Det sidste parameter til at klarlægge modellens ydeevne er F1-scoren, som er en metrik, der kombinerer både præcision og recall for at give en enkelt værdi til at vurdere ydeevnen af en machine learning model. Den tager højde for begge metrikker for at give et mere balanceret billede af modellens præstation.

Det gælder generelt set, at en score på 0,7 eller højere anses for at være god, mens en score på 0,8 eller højere betragtes som fremragende. Dog afhænger det af specifikationerne for den givne opgave og de krav, der er stillet til præcision og recall. Da modellen er baseret på 'XG Boost' og da pesticidværktøjet fungerer som en blåstempling af de valgte parametre er det forventet, at modellen vil have en grundlæggende acceptabel gengivelse af virkeligheden. Algoritmen forsøger nemlig at maksimere nøjagtigheden uden at vægte hverken 'fund' eller 'ikke-fund'.

3.6 Dynamisk opdatering

Når en machine learning model skal køres, vil der typisk være en række trin, der på et eller andet tidspunkt skal udføres:

1. **Hentning af rådata fra interne og eksterne kilder.**
   Data befinder sig ofte et andet sted, end den database, som modellen bruger til at læse og skrive fra og til. Derfor skal data, enten manuelt eller via en kode, overføres fra kilden.

2. **Præprocessering og samkøring af rådata**
   Ofte vil data være i et format, som ikke er det optimale ift. hvad machine learning modellen har brug for. Det kan være alt fra datatyper der skal ændres og filtrering af dårlige data, til data der skal samkøres, geokodes, osv.

3. **Træning af modellen**
   For at modellen kan finde mønstre i data, den ikke har set før, skal den se mange eksempler på data, hvor forholdet mellem det vi ønsker at estimere, og de forhold der altid er kendte, kan identificeres og indlæres.

4. **Kørsel af resultater**
   Når modellen er trænet, kan den viden den har tillært sig, benyttes til at komme med estimater på data, den ikke har set før.

5. **Publicering af resultater**

Når resultaterne er klar, skrives de til en database. Ofte vil det herudover også være nødvendigt at publicere dem igennem en løsning, f.eks. via WFS (Web Feature Service).

Ovenstående trin giver et godt overblik over de trin, der indgår i at komme fra rådata til resultater. Det er dog også en forenklet gennemgang, og virkeligheden er typisk mere mudret. Det betyder også at mange modeller typisk kun køres en eller to gange, og derfra bruges resultaterne så fremadrettet. I mange tilfælde kan det være tilstrækkeligt at køre en model et par gange, når den har opnået den ønskede kvalitet. Andre gange er der dog situationer, hvor man kan have et ønske om let og fleksibelt at kun gentræne modellen og lave nye resultater, f.eks. i en situation hvor den viden modellen er trænet på, løbende ændrer sig.

For modellen til at screene PFAS sårbarheden, har vi derfor opsat en pipeline, der med et enkelt klik kan genkøre trin 1-5 fra flowet ovenfor. På den måde vil modellen hele tiden blive bedre og bedre i takt med, at der indsamles ny viden om PFAS-forureninger i Danmark.

4. Resultater

4.1 Nøgletallene for modellen

Ved at anvende klassifikationsalgoritmen er følgende metrikker opnået for den endelige model:

TABEL 1. Nøgletal for modellen.

Metrik		
Nøjagtighed (accuracy)	0,83	83 %
Præcision (precision)	0,79	79 %
Recall	0,72	72 %
F1-score	0,75	75 %

Modellen har, som set fra TABEL 1, opnået en nøjagtighed på 0,83, hvilket betyder, at 83 % af PFAS-estimaterne på valideringsdatasættet på landsplan var korrekte. Datasættet til validering består af 1.252 boringsindtag, hvoraf 464 indtag repræsenterer PFAS fund og de resterende 788 indtag repræsenterer PFAS ikke-fund. Af de 1.252 boringsindtag estimerer modellen 1.033 indtag korrekt, hvilket altså svarer til en modelnøjagtighed på 83 %. For pesticidværktøjet blev der opnået en nøjagtighed på 72 %.

Ligeledes er præcisionen også i den højere ende og det indikerer at model har en acceptabel evne til at estimere PFAS-fund. Modellen estimerer, at der i datasættet er 423 PFAS-påvirkede indtag og ud af disse 423 indtag har modellen ret i 334 gange. Det giver altså en præcision på 79 %.

For recall præsterer modellen ikke lige så godt som for nøjagtighed og præcision. Recall siger noget om, i hvor høj grad modellen evner at genfinde de reelle PFAS-fund og en lav recall giver derfor et højere antal falsk negative estimater. Ud af de 464 reelle fund kan modellen genfinde de 334 og der er altså 130 reelle PFAS-fund, som modellen ikke kan genfinde – en falsk negativ. Det giver et recall på 0,72, som betyder at modellen i 72 % af sine estimater er i stand til at genfinde de reelle påviste indtag. Modellen er faktisk væsentligt bedre til at genfinde ikke-fundene. Fra de 788 ikke-fund i det reelle data kan modellen genfinde 699. Der er derfor ikke ret mange falske positive og derfor kan man være ret sikker på, at der hvor modellen peger på, at der er PFAS-påvirket faktisk er korrekt.

Samlet vurderes modellen at give et acceptabelt grundlag for at kunne estimere PFAS sårbarhed i forskellige geografier, hvor der ikke i forvejen findes PFAS målinger. Det bør dog noteres, at modellen for nuværende vil have tendens til at underestimere PFAS sårbarheden, hvilket er set på baggrund af modellens svagere recall af de reelle fund. Resultaterne bør derfor ikke anvendes til fravalg af undersøgelse i områder med lav sandsynlighed.

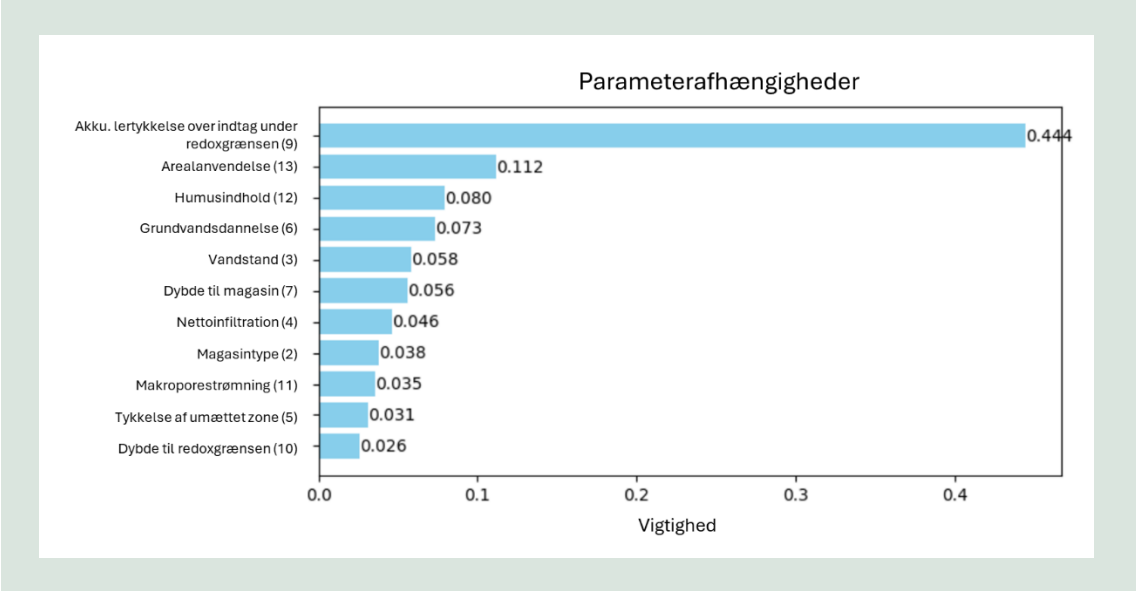
Som med alle andre modeller er resultaterne ikke bedre end de data der bliver anvendt i modellen. Derfor er det forventet, at både præcision og recall vil stige med en stigende datamængde. Det er netop derfor, at der er lagt stor fokus på at udarbejde en dynamisk model, som træner sig selv og bliver bedre med tiden.

4.2 Parameterafhængigheder

Da projektet står på skuldrene af pesticidværktøjet, har der været en stor sikkerhed for, at de udvalgte parametre ville give en acceptabel model og det er også det metrikkerne viser. Men

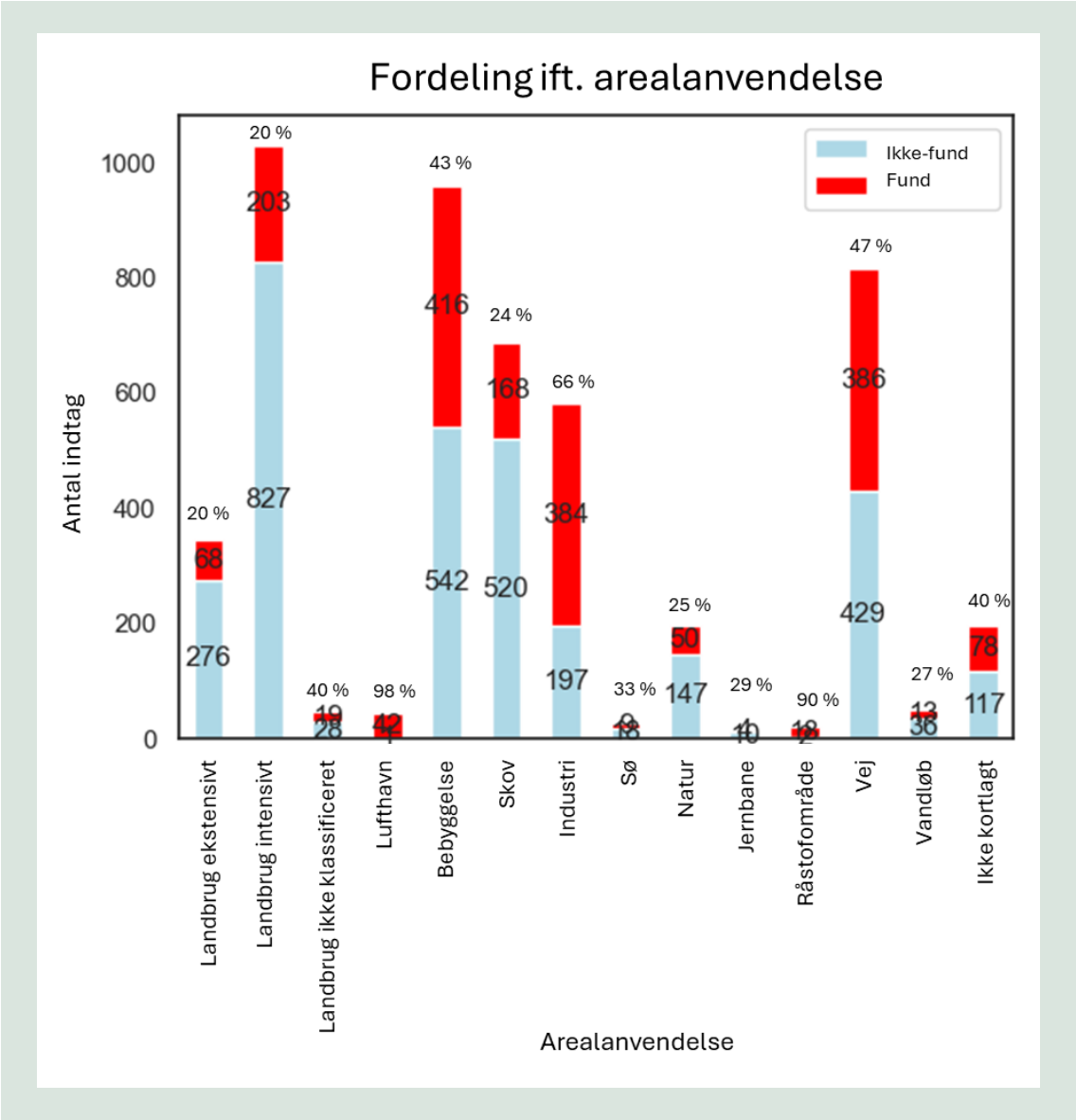
hvilke parametre der ville vise sig at være afgørende for PFAS har der ikke været sikkerhed for.

Figuren med parameterafhængighederne nedenfor, FIGUR 5, viser at lertykkelsen over indtaget er det vigtigste parameter i forhold til fund eller ikke fund af PFAS. Dette hænger fint sammen, rent geologisk, da lertykkelsen over en pågældende indtag vil være positivt korreleret med grundvandets alder og da man først begyndte at anvende PFAS-stofferne i 1950'erne vil grundvand, der er ældre end dette, ikke være påvirket. Den næst vigtigste parameter er arealanvendelse og det betyder at PFAS-forurening generelt er tæt korreleret, med den anvendelse arealerne har.



FIGUR 5. Grafen viser modellens og estimaternes parameterafhængigheder. Altså hvilke parametre er vigtigt for fund eller ikke fund.

Hvis man kigger dybere ned i data, FIGUR 6, viser det sig, at det især er arealanvendelser som lufthavne, industri og råstofområdet, som har en tæt korrelation med fund af PFAS. For lufthavne er der i 98 % af de analyserede indtag påvist PFAS, mens det samme gælder for 1/3 af det, der er karakteriseret som industri. I forhold til ikke-fund af PFAS er denne klasse især knyttet til forskellige typer af landbrug, samt skov og natur. Det bør bemærkes, at der i datasættet er indlejret forskellige usikkerheder og bias, se afsnit 4.3 nedenfor.



FIGUR 6. Fordeling af fund / ikke-fund af PFAS i forhold til arealanvendelsen. Over hver søjle er angivet en procentsats, som viser andelen af fund i forhold antallet af indtag. Det viser sig, at der er flere anvendelser, som især korrelerer stærkt med fund af PFAS, her f.eks. lufthavne (98%), industri (66 %) og på råstofområdet (90 %). I forhold til arealanvendelser, der korrelerer stærkere med ikke-fund, er f.eks. forskellige landbrugs aktiviteter (20 %), skov (24 %) og natur (25 %).

4.3 Usikkerheder og bias

Da data, i f.eks. GEUS' Jupiter database, ikke er indsamlet med henblik på statistisk repræsentativitet, vil data afspejle historiske prøvetagnings- og analysepraksisser. Det betyder, at datasættet indeholder en række systematiske skævheder (bias), som vil påvirke modellens resultater. Mulige bias kunne være:

Prøvetagningsbias: Prøver til analyse for PFAS er historisk udtaget på lokaliteter, hvor der har været en forventning om PFAS-forurening (industri og lufthavne). Dette medfører en højere fundrate i disse områder end i en tilfældig geografisk fordeling.

Analysebias: Datasættet omfatter kun de PFAS-forbindelser, der indgår i gældende analyseprogrammer (aktuelt 22 stoffer). Andre relevante stoffer (f.eks. TFA) indgår derfor ikke i

datagrundlaget. Det betyder, at aktiviteter som f.eks. langbrug giver færre fund, end hvad forventeligt er korrekt.

Dybdebias: En stor del af analyserne er udført i terrænnære boringer, hvilket betyder, at modellen primært beskriver forhold i det øvre grundvand.

Disse usikkerheder er et resultat af det tilgængelige datagrundlag og ikke et valg i modeludviklingen. Den dynamiske tilgang forventes over tid at reducere nogle af skævhederne i takt med, at datagrundlaget udvides til flere geografier og anvendelsestyper, men det forventes ikke, at bias vil kunne elimineres.

Modellens resultater skal derfor bruges som et estimat af sandsynlighed for fund af PFAS, og ikke som en fuldstændig eller sandfærdigt billede af faktiske PFAS-forureninger eller risiko herfra.

4.4 Udstilling og forankring

Da formålet med projektet var at lave et offentligt tilgængeligt værktøj, som kunne benyttes af forskellige offentlige og private institutioner samt andre interessenter, har der været et stort ønske om at kunne forankre og udstille resultaterne et sted, hvor de kommer ud til det størst mulige publikum. Arbejdet er baseret på gode offentlige data og i dette tilfælde har det været ønsket, at projektet tilbagebetaler med resultater baseret på disse data. Udgangspunktet har derfor været, at resultaterne skulle udstilles gennem Danmarks Miljøportal, hvor det blandt andet skulle kunne tilgås fra webservicen 'Arealinformation'.

Gennem en udstilling internt i COWI og ved deling af såkaldte WMS (Web Map Service) og WFS (Web Feature Service) filer herfra, er det lykkedes at få værktøjet implementeret hos Danmarks Miljøportal og det er derfor muligt at tilgå resultaterne via deres 'Lagvælger', som kan tilgås i blandt andet 'Arealinformation' og 'Kemidata' men også hentes til lokale GIS systemer, som f.eks. QGIS.

For at tilgå data i Arealinformation gøres følgende:

Arealinformation åbnes: https://danmarksarealinformation.miljoeportal.dk/
I lagvælgeren til højre trykkes *'Tilføj lag'*
I toppen af det nye vindue søges efter: *'Grundvandets PFAS-sårbarhed (samlet)'*
Tryk på *'+'* under datasættet og luk igen vinduet

Det er i vinduet muligt at vælge data for konkret det af de 13 forskellige grundvandsmagasiner, man er interesseret i eller vælge et grupperet lag, som indeholder alle 13 grundvandsmagasiner samlet. Hvis det samlede lag er valgt, er det nu muligt at skifte mellem de forskellige grundvandsmagasiner i lagvælgeren, hvorved resultaterne opdateres på danmarkskortet.

4.5 Estimater

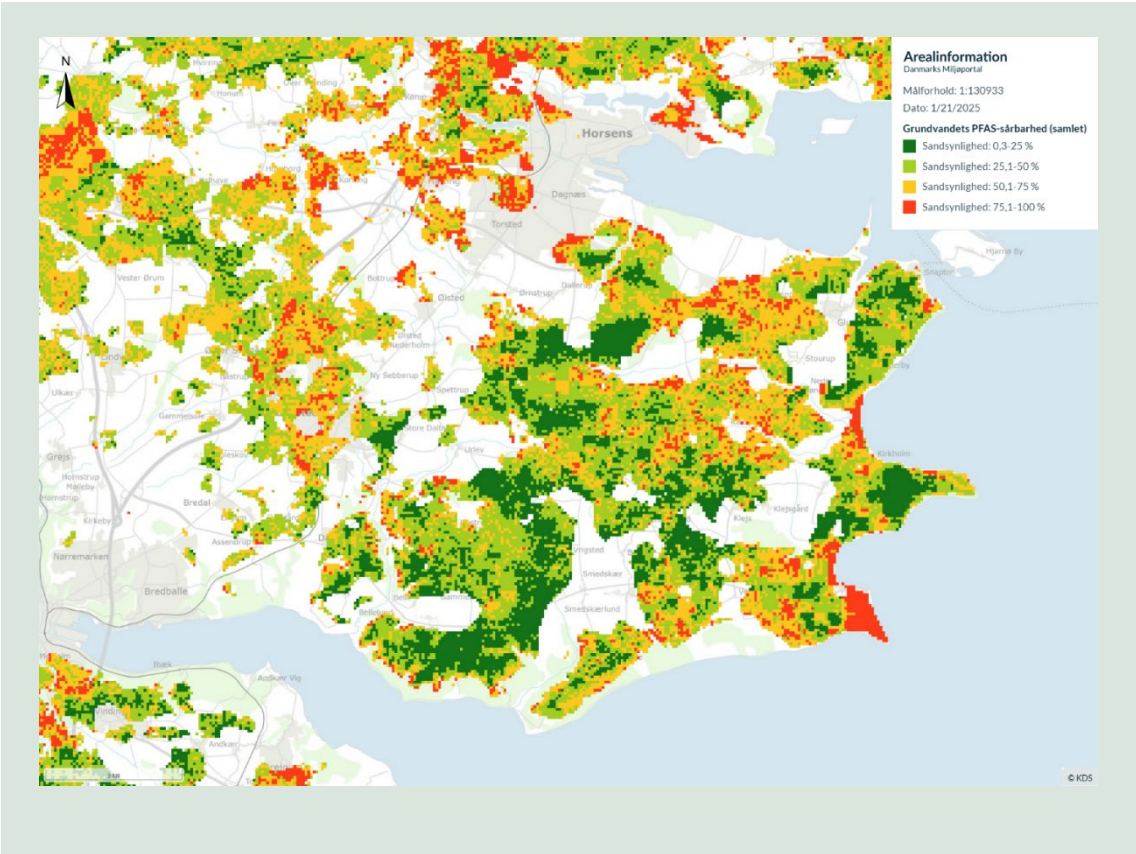
For at kunne estimere PFAS-sårbarheden benyttes DK-modellens 100x100m grids. For hvert grid er der lavet et punkt i hver af magasinerne og det er i disse punkter estimaterne er knyttet. Der er flere steder i Danmark, hvor ikke alle de geologiske lag optræder og i disse områder er der selvfølgeligt ikke knyttet et estimat.

I hvert punkt og magasin ses der en værdi for, hvor sikker modellen er på hhv. fund eller ikke fund af PFAS. Hvert punkt i kortene viser en beregnet sandsynlighed for, om der findes PFAS i et bestemt grundvandsmagasin. Farverne angiver sandsynligheden:

- Rød = > 75 % sandsynlighed for fund
- Gul = 50-75 % sandsynlighed for fund
- Lysegrøn = 25-50 % sandsynlighed for fund

- Mørkegrøn = < 25 % sandsynlighed for fund

Der hvor magasinet ikke optræder, ses blot baggrundskortet. Se FIGUR 7 for forståelse.



FIGUR 7. Udsnit fra Arealinformation. Kortet viser sandsynligheden for fund af PFAS i grundvandsmagasinet KS2. De røde områder indikerer stor sandsynlighed, de gule områder mindre stor sandsynlighed og de grønne og mørkegrønne områder hhv. mindre og mindst sandsynlighed for at påvise PFAS i KS2 i det pågældende område.

5. Konklusion

Projektet har sigtet efter at udarbejde en machine learning model, som kunne hjælpe offentlige institutioner og andre interessenter med at estimere PFAS-sårbarheden i forskellige grundvandsmagasiner. Vi står som land over for en udfordring i forhold til disse kemikalier, som er meget svært nedbrydelige i naturen. Der er derfor af stor interesse at have et værktøj, som kan hjælpe med at prioritere forureningsundersøgelser, vandindvinding, anlægsprojekter eller andet med relation til PFAS.

Gennem en klassifikationsalgoritme, 'XG Boost' er det lykkedes at udarbejde en model, som har en nøjagtighed på 82 % af sine estimater af PFAS på tværs af Danmark. Det betyder at modellen kan anvendes som et screeningsværktøj for de forskellige grundvandsmagasiners sårbarhed for PFAS. Modellen har for nuværende tendens til at estimere flere falske negative end falske positive og det betyder i praksis, at den kan have tendens til at underestimeresandsynligheden for fund. Omvendt vil de områder den udpeger som belastet, med stor sandsynlighed være korrekt. Det er håbet og forventningen, at modellen med tiden vil blive endnu bedre til at estimere korrekte målværdier og derfor vil usikkerhederne falde. Men det er vigtigt at brugeren har øje for de bias, der naturligt ligger i modellen og at det ikke er forsøgt at minimere eller undgå disse i arbejdet. Brug af modellen bør derfor altid suppleres med faglige vurderinger og lokale data.

Da undersøgelser for PFAS er et område med stor fremgang, vil der nationalt blive udtaget en stor mængde prøver de kommende år. For at modellen ikke allerede indenfor kort tid vil være forældet, er den blevet lavet dynamisk, så den kan opdateres med nyeste data ved meget lille indsats. Det kræver blot at modellen med få klik køres på det nyeste data og herefter vil en ny version blive udstillet hos Danmarks Miljøportal.

6. Referencer

/1/ Værktøj til vurdering af grundvandets pesticidsårbarhed - Teknisk baggrundsrapport, COWI 2020
/2/ Afgrænsning af de danske grundvandsforekomster – Ny afgrænsning og delkarakterisering samt fagligt grundlag for udpegning af drikkevandsforekomster, GEUS v. Lars Troldborg, 2020

Resumé
Rapporten præsenterer et projekt, der har udviklet en machine learning-model til at vurdere sandsynligheden for tilstedeværelsen af PFAS i grundvandet. Ved at bruge offentlige data fra blandt andet GEUS' Jupiter-database og anvende kunstig intelligens, kan modellen estimere sandsynligheden for tilstedeværelse af PFAS i forskellige grundvandsmagasiner. Formålet er at hjælpe kommuner, regioner og andre interessenter ved at stille et indledende screenings-værktøj til rådighed ved undersøgelses-, oprydnings- og anlægsindsatser.

Projektet bygger på et eksisterende værktøj udviklet af COWI til vurdering af grundvandets sårbarhed over for pesticider. Modellen bruger de samme 11 parametre, såsom magasintype, vandstand, nettoinfiltration og arealanvendelse, men er tilpasset til at estimere PFAS-sårbarheden. Der er anvendt algoritmen 'XG Boost Classifier' for at opnå de bedste prædiktive resultater. Modellen har en høj nøjagtighed på 83% og præcision på 79%, men også en tendens til at underestimere sandsynligheden for fund. På trods af usikkerheder og bias, anses modellens estimater for at være anvendelige i praksis, men bør suppleres med faglige vurderinger og lokale data.

Efterhånden som nye data indsamles, kan modellen let gentrænes, hvilket sikrer, at estimaterne altid er baseret på de nyeste data og modellen forbedres således løbende i takt med, at mere data bliver tilgængeligt.

Værktøjet er tilgængeligt gennem Danmarks Miljøportal, hvor implementeringen sikrer bred adgang og nem anvendelse af værktøjet. Modellen udgør et værdifuldt værktøj til vurdering af PFAS-sårbarhed i grundvandet og forventes at kunne udgøre en vigtig ressource for fremtidige beslutninger om f.eks. miljøbeskyttelse og offentlige sundhedsindsatser i Danmark.

Miljøstyrelsen
Lerchesgade 35
5000 Odense C

www.mst.dk