AALBORG UNIVERSITET
STUDENTERRAPPORT

Sensorbaseret varsling af badevandskvalitet i Aarhus Havn

Mathias S. Kristensen & Ørjan Heggdal

Kandidatspeciale | Vand og Miljø | 7. juni 2019

AALBORG UNIVERSITET
STUDENTERRAPPORT

Det Ingeniør- og
Naturvidenskabelige Fakultet
Institut for Byggeri og Anlæg
Thomas Manns Vej 23
9220 Aalborg Øst
www.civil.aau.dk

Titel

Sensorbaseret varsling af badevandskvalitet i Aarhus Havn

Projekttype

Kandidatspeciale

Projektperiode

September 2018 - Juni 2019

Forfattere

Mathias Schandorff Kristensen
Ørjan Heggdal

Vejledere

Jesper Ellerbæk Nielsen
Michael Robdrup Rasmussen

Sideantal: 122
Afsluttet: 7. juni 2019

Synopsis:

Nærværende rapport undersøger hvordan der kan opstilles et system til varsling af forringet badevandskvalitet i Aarhus Havn. Formålet med projektet er at opstille en model, der på baggrund af sensorbaserede parametre, kan indikere hvorvidt koncentrationen af fækale bakterier overskrider grænseværdierne for tilfredsstillende badevandskvalitet. Gennem en systemanalyse er det undersøgt hvilke parametre, der egner sig som inddirekte kvalitetsparametre. Her er det konkluderet, at vandets elektriske ledningsevne egner sig bedst og kan bruges til at modellere badevandskvaliteten med en præcision på 82%. Ved brug af machine learning er det undersøgt, hvorvidt der kan opnås en bedre varslingsmodel ved at benytte sig af flere inputparametre. Her konkluderes det, at der ved at benytte vandets ledningsevne i kombination med maksimale regnintesiteter for de foregående 48 timer, kan opnås en præcision på 89 % samt en væsentlig reduktion i antallet af dage, hvor der skal varsles om dårlig badevandskvalitet. Til implementeringen af varslingssystemet er der udviklet en målestation, der kan begrænse algevækst på sensoren samt opstillet en simpel model, der er baseret på machine learning modellerne.

ABSTRACT

During this project a complete warning system has been built to warn about bad bathing water quality in Aarhus Harbour using real-time sensor data. The current methods used to measure the bacteria concentration in water is insufficient to warn about bad bathing water quality, due to the long incubation period of the samples of minimum 24 hours. This, combined with the generally bad water quality in the harbour, leads to the need of a warning system.

With a statistical approach the suitable indicators for bathing water quality have been investigated. The conclusion is that the best indicator is the electrical conductivity of the water. This is due to the fact that the main source of pollution to the harbour is the river, Aarhus Å, which has its delta in the harbour. Aarhus Å acts as a recipient of a large amount of CSO's and outlets from wastewater treatment plants. The electrical conductivity therefore acts as a tracer of how much polluted fresh water is present at the place of measurement. The correlation between the bacteria concentration and the electrical conductivity has been modelled with an exponential regression with good results. The performance has been improved by including rain depth and intensity to a new model based on machine learning. Besides the improved performance of the model this results in a significant reduction in the number of closing days.

To complete the warning system a measuring station to monitor the electrical conductivity has been developed. The main task of the new measuring station is to avoid the unwanted growth of algae on the sensor. This problem is solved by placing the sensor in a chamber which can be emptied from water with compressed air.

RESUME

I dette projekt er der opbygget et komplet varslingssystem der, ved brug af sensordata, kan varsle om forringet badevandskvalitet i Aarhus Havn. De nuværende metoder til analyse af bakteriekoncentrationer i badevand er ikke egnede til at varsle om forringet badevandskvalitet grundet lang opformeringstid for prøverne på mindst 24 timer. Den lange opformeringstid, kombineret med generelt dårlig vandkvalitet i havnen, medfører et behov for et varslingssytem.

Med en statistisk tilgang er der undersøgt hvilke parametre der er egnede indikatorer for badevandskvalitet. Det konkluderes herudfra, at den bedste indikator er vandets elektriske ledningsevne. Dette skyldes, at den primære kilde til fækal forurening af havnen er Aarhus Å, der har sit udløb i havnen. Aarhus Å er belastet af en lang række regnbetingede udløb fra fælleskloak og udledninger fra renseanlæg. Den elektriske ledningsevne kan derfor benyttes som tracer for hvor stor en andel af forurenet ferskvand, der er tilstede der, hvor der måles. Der er opbygget en eksponentiel model for sammenhængen mellem bakteriekoncentration og ledningsevne, der viser en høj præcision. Præcisionen er forbedret ved at inddrage nedbørsmængder og -intesiteter i en ny model baseret på machine learning. Foruden de forbedrede prædiktionsresultater resulterer den nye model i en markant reduktion i antallet af dage, hvor badeområde skal lukkes.

Til at fuldende varslingssystemet er der udviklet en målestation til at monitere ledningsevnen. Formålet med den nye målestation er at løse et problem med uønsket algevækst på sensoren. Dette problem er løst ved at placere sensoren i et kammer, der kan tømmes for vand ved brug af trykluft.

FORORD

Denne rapport er udarbejdet i forbindelse med vores afgangsprojekt på kandidatuddannelsen i Vand og Miljø på Aalborg Universitet. Afgangsprojektet er udført i perioden fra september 2018 til juni 2019 og er kulminationen på fem år som studerende ved Institut for Byggeri og Anlæg. Projektet er blevet til i samarbejde med Aarhus Kommune, Amphi-Bac ApS og Aalborg Universitet.

Baggrunden for projektet er behovet for varsling af badevandskvaliteten i Bassin 7 i Aarhus Havn, da området ikke opfylder badevandsbekendtgørelsens kvalitetskrav. Projektets formål er derfor at undersøge mulighederne for at opbygge et varslingssystem, der kan indikere, hvorvidt badevandskvaliteten er god eller dårlig.

Vores motivation for at vælge netop dette projekt, har været mulighederne for at arbejde med IoT og Machine Learning til indsamling og behandling af data samt at udvikle enkle, sensorbaserede løsninger på problemer, der typisk løses med store deterministiske modeller.

Vi vil gerne sige tak til Nikolaj Kruse Christensen fra Teknik og Miljø, Aarhus Kommune, for et godt samarbejde. Der skal også lyde et tak til Søren Bastholm fra Amphi-Bac ApS for vejledning omkring badevandskvalitet og bakterieanalyser. Det største tak skal dog gå til vores vejledere Jesper Ellerbæk Nielsen og Michael Robdrup Rasmussen for jeres engagement i vores projekt, utallige idéer og altid kyndige vejledning.

Læsevejledning

I rapporten angives passive kilder med [Efternavn, År] og aktive kilder med Efternavn [År]. Hvis kilden er angivet før et punktum, refereres der til sætningen, og hvis kilden er angivet efter et punktum, refereres der til hele det foregående afsnit. Alle kilder fremgår af litteraturlisten på side 83.

Figurer, formler og tabeller er nummereret efter det kapitel, de befinder sig i, således den første figur i kapitel 2 hedder figur 2.1 og så fremdeles. Kortmateriale til udarbejdelse af figurer stammer fra Miljøstyrelsen [2019] og Styrelsen for Dataforsyning og Effektivisering [2019]. Et overblik over de vedlagte digitale bilag fremgår af appendiks A. En beskrivelse af det anvendte data fremgår af appendiks B.

INDHOLDSFORTEGNELSE

1 Indledning **1**
- 1.1 Aarhus Havn: Fra industrihavn til aktivt bymiljø . . . 1
- 1.2 Forvaltning af badevandskvalitet . . . 2
- 1.3 Måling af badevandskvalitet . . . 3
- 1.4 Badevandskvaliteten i Aarhus Havn . . . 4
- 1.5 Forbedring af badevandskvalitet i Aarhus havn . . . 5
- 1.6 Projektets mål og problemformulering . . . 7

2 Systemanalyse **9**
- 2.1 Opland til Aarhus Å . . . 9
- 2.2 Havnesystemet i Aarhus . . . 11
- 2.3 Indikatorer på dårlig badevandskvalitet . . . 15
- 2.4 Opsummering . . . 23

3 Undersøgelse af elektrisk ledningsevne **25**
- 3.1 Dataindsamling . . . 25
- 3.2 Statistisk overblik over ledningsevnedata . . . 26
- 3.3 Undersøgelse af tidslig variation af ledningsevne . . . 28
- 3.4 Fourier-analyse af svingninger i ledningsevnen . . . 29
- 3.5 Undersøgelse af stedslig variation af ledningsevne . . . 30
- 3.6 Påvirkning af ledningsevnen ved enkelthændelser . . . 31
- 3.7 Sammenhæng mellem vind og ledningsevne . . . 35
- 3.8 Opsummering . . . 37

4 Modellering af badevandskvalitet på baggrund af ledningsevne **39**
- 4.1 Eksponentiel model til vurdering af badevandskvaliteten . . . 39
- 4.2 Optimering af kritisk ledningevne . . . 42
- 4.3 Sandsynlighedsmodel for badevandskvalitet . . . 43
- 4.4 Kritisk ledningsevnes betydning for lukketid . . . 44
- 4.5 Opsummering . . . 46

5 Modellering af badevandskvalitet med machine learning **47**
- 5.1 Hvad er machine learning? . . . 47
- 5.2 Oversigt over udvalgte modelalgoritmer . . . 47
- 5.3 Fremgangsmåde . . . 49
- 5.4 Udvælgelse af inputparametre . . . 50
- 5.5 Opbygning af regressionsmodel . . . 54
- 5.6 Opbygning af klassifikationsmodel . . . 57
- 5.7 Sammenligning af modeller . . . 60
- 5.8 Opsummering . . . 62

6 Implementering af varslingsmetode **63**
- 6.1 Forbedring af metode til måling af ledningsevne . . . 64
- 6.2 Valg af model til implementering . . . 71
- 6.3 Hjemmeside til varsling af badevandskvalitet . . . 72
- 6.4 Opsummering . . . 73

7 Konklusion **75**

8 Perspektivering: Er metoden generel? 77
8.1 Undersøgelsesområde 77
8.2 Dataindsamling i Østre Havn og ved Aalborg Havnefront 79
8.3 Potentiale for udbredelse af sensorbaseret varsling 80

Litteratur 83

Appendiks A Oversigt over digitale bilag 87

Appendiks B Dataoverblik 89

Appendiks C Interkalibrering af elektrisk ledningsevnemålere 93

Appendiks D Temperaturkorrigering af elektrisk ledningsevne 95

Appendiks E Opbygning af logistisk regressionsmodel 99

Appendiks F Beskrivelse af machine learning algoritmer 101

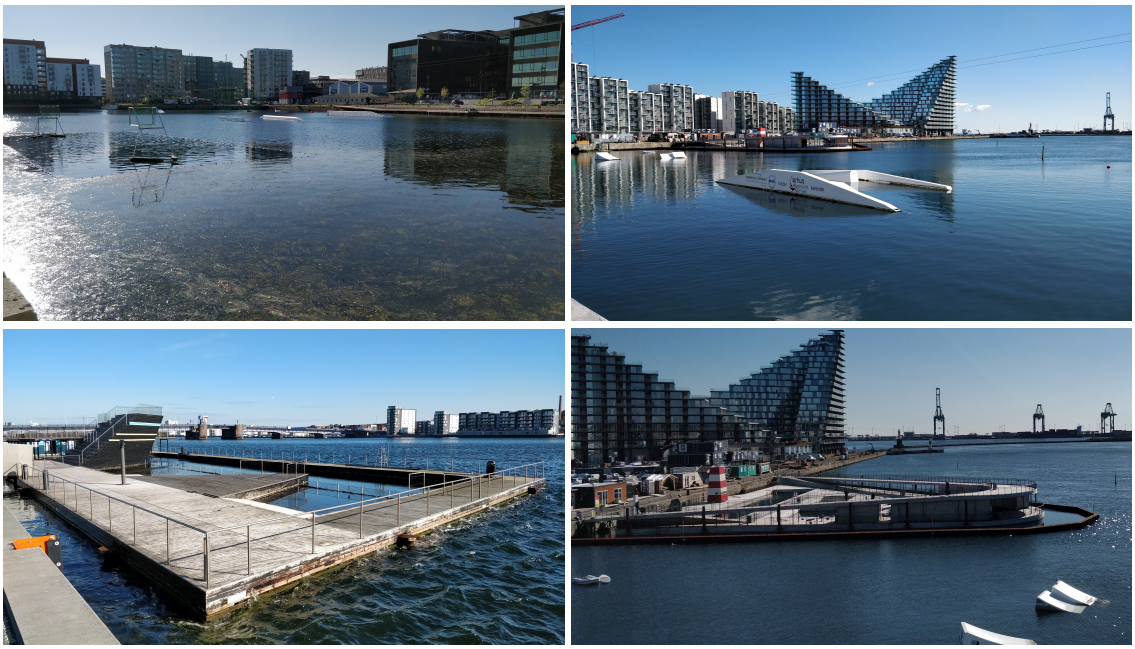
Appendiks G Beskrivelse af performance-parametre 107

Appendiks H Analyse af E. Coli i vandprøver 109

Kapitel 1

INDLEDNING

De danske havne er i disse tider i gang med en omdannelse fra grå industrihavne til kulturelle områder, der indbyder til rekreative aktiviteter. Havnebade etableres i store og små havne over hele Danmark, og vandsportsklubberne oplever en stigning i medlemmer [Nordea Fonden, 2018]. Omdannelse fra industri til kultur er en stor del af den politiske dagsorden, hvor omdannelse af industrihavne er blevet til prestigeprojekter over hele landet [Kulturarvsstyrelsen, 2009].



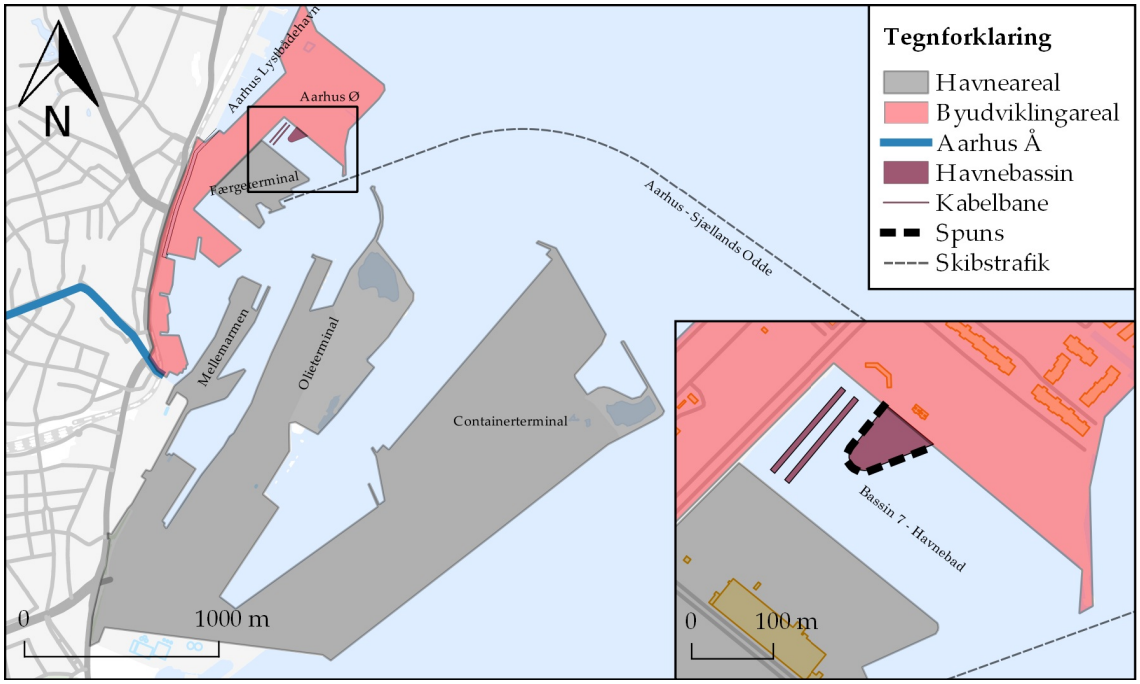
Figur 1.1. Øverst fra venstre: Aalborg Kabelpark og Kabelparken i Bassin 7. Nederst fra venstre: Aalborg Havnebad og havnebadet i Bassin 7.

Udbyggelse af havneområder til rekreative formål giver en ny dimension til byudviklingen, da det giver mulighed for at skabe nye områder, der lægger op til aktiviteter på vandet i de tætbebyggede urbane område. Det er dog samtidig kendt, at vandkvaliteten i havne generelt er dårlig, da havne ofte er forbundet med større byer, hvor regnbetingede udledninger er hyppige [Luna m.fl., 2018]. Dette medfører en sundhedsrisiko for brugerne og er derfor et forhold, man skal tage hensyn til, når havne inddrages til rekreative formål. Et af de områder, hvor bynære rekreative vandområder bliver etableret, er ved Aarhus Havn.

1.1 Aarhus Havn: Fra industrihavn til aktivt bymiljø

Aarhus Havn er en stor industrihavn, der blandt andet huser Danmarks største containerhavn, en færgeterminal og en krydstogtterminal. En oversigt over havnearealet fremgår af figur 1.2. Området omkring Nordhavnen, der er markeret med rødt på figur 1.2, er siden 2007 blevet byudviklet, og den nye bydel er døbt *Aarhus Ø*. I den nye bydel opføres

mange store etagebyggerier og skal, når bydelen står færdig, huse 10.000-12.000 beboere og 10.000 arbejdspladser. [Aarhus Kommune, 2018b]



Figur 1.2. Oversigt over Aarhus Havn.

Ét af de mange prestigebyggerier i Aarhus Ø er det nye havnebad. Havnebadet er placeret i Bassin 7 centralt i Aarhus Ø. Det arkitekttegnede havnebad er blevet etableret med en spuns rundt om selve havnebadet. Dette har været nødvendigt for at skåne konstruktionen imod bølger og samtidig sikre vandkvaliteten i havnebadet.

Udover havnebadet i Bassin 7 benyttes havnebassinerne flere steder rekreativt. Ved Bassin 7 ligger Aarhus Watersports Complex, der driver en kabelbane i det åbne havnebassin. Ved Lystbådehavnen ligger Surf Agency, der afholder kurser i vandsportsaktiviteter. Modsat havnebadet er der kun lavet få foranstaltninger til at skåne brugerne af den øvrige del af havnen mod forurenet vand. Dette til trods, for at der er et stort ønske om at benytte havnen rekreativt.

På trods af industriens store rolle i Aarhus Havn bliver havnen benyttet til badning og kan dermed betragtes som badevand [Miljø- og Fødevareministeriet, 2016]. Dette medfører, at badeområdet er underlagt lovgivning omkring badevandskvalitet.

1.2 Forvaltning af badevandskvalitet

I EU klassificeres badevandskvaliteten i overfladevande efter EU's direktiv om forvaltning af badevandskvalitet [Europa-parlamentet og rådet, 2006], der i Danmark er implementeret gennem bekendtgørelse om badevand og badeområder. Bekendtgørelsen stiller krav til udarbejdelse af badevandsprofiler og prøvetagning til klassifikation af badevandskvalitet ved danske badevandsområder. Ved badevandsområder forstås ferskvand og havvand, som kommunalbestyrelsen forventer at blive anvendt til badning

og som ikke er omfattet af et badevandsforbud. Klassifikationen af badevandsområdet udarbejdes på baggrund af de gældende kvalitetskrav, der fremgår af tabel 1.1. Overholder badevandsområdet ingen af kvalitetskravene, klassificeres kvaliteten som værende ringe, og der skal derfor udstedes et permanent badeforbud. Et badevandsområde kan desuden klassificeres som ringe, hvis forekomsten af cyanobakterier, affald, tang eller fytoplankton udgør en sundhedsfare. [Miljø- og Fødevareministeriet, 2016]

	Intestinale enterokokker [cfu/100 ml]	**Escherichia coli [cfu/100 ml]**
Udemærket kvalitet	100*	250*
God Kvalitet	200*	500*
Tilfredsstillende kvalitet	185**	500**

Tabel 1.1. Kvalitetskrav for kystvande og overgangsvande. *Ud fra en vurdering af 95-% fraktilen. **Ud fra en vurdering af 90-% fraktilen. [Miljø- og Fødevareministeriet, 2016]

Badevandskvaliteten klassificeres på baggrund af koncentrationen af to indikatorer på fækal forurening; intestinale enterokokker og Escherichia coli (herefter benævnt enterokokker og E. Coli). Disse to bakterier benyttes, da de opfylder kravene for fækale indikatorer. Dette betyder, at de optræder, når der er sket en fækal forurening, og at mængden af indikatorerbakterier er proportionale til graden af fækal forurening. Grænseværdien er fastsat ud fra sandsynligheden for, at badende bliver syge. Sandsynligheden for at blive syg stiger med koncentrationen af indikatorerbakterier, og der er dermed mindre sandsynlighed for at blive syg af badevand af udemærket kvalitet i forhold til god kvalitet [Kay m.fl., 1994]. Grænseværdien er forskellig for badning i havvand og ferskvand. Dette skyldes, at der er større sandsynlighed for at sluge vandet i havvand, hvor der ofte er bølger i forhold til ferskvand, hvor der ofte ikke er bølger.

Jævnfør Miljø- og Fødevareministeriet [2016] kræves der minimum 4 stikprøver pr. badevandssæson. Klassifikationen vurderes ud fra prøver, som er taget igennem de sidste 4 sæsoner. Opfylder klassifikationen ingen af kravene i tabel 1.1, klassificeres badevandskvaliteten som værende ringe. Badevandet kan dog stadig klassificeres som værende udemærket, god eller tilfredsstillende på trods af kortvarig forurening, forudsat at badende varsles tidligt. Tidlig varsling af badende indebærer oplysning til badende om dårlig badevandskvalitet, når den er tilstedeværende, hvilket kræver en online monitering af badevandskvaliteten.

1.3 Måling af badevandskvalitet

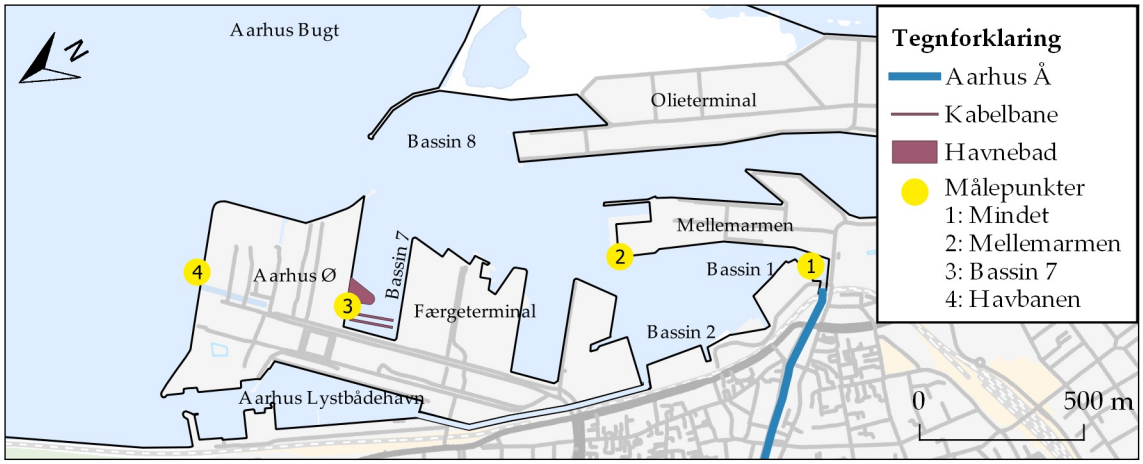
Vurdering af badevandskvalitet baseres på analyse af vandets indhold af enterokokker og E. Coli. I Danmark bestemmes mængden af E. Coli og enterokokker i overfladevand ved mikrotiterplates, hvor antallet af bakterier kan beregnes efter opdyrkning [ISO, 1999]. Denne metode kræver op til 72 timers opdyrkning. Den lange opdyrkningstid hindrer muligheden for varsling af badevandskvalitet ved direkte måling af E. Coli og enterokokker. Det kræver derfor en alternativ varslingsmetode for at kunne opretholde tilstrækkelig klassifikation af badevandskvaliteten. [Miljø- og Fødevareministeriet, 2018]

Varsling af badevandskvalitet foretages flere steder i Danmark på baggrund af deterministiske strømnings-modeller. Input til modellerne er typisk overløbsvoluminer, der måles i realtid af forsyningsselskaberne, der kombineres med input som vindpåvirkning, tidevand samt badeområdets fysiske udformning. Modellerne simulerer vandets strømninger samt nedbrydningen af bakterier og er dermed i stand til at give et estimat af bakterieindholdet. [DHI, 2019]

Andre metoder til varsling af badevandskvalitet er baseret på statistiske sammenhænge mellem bakteriekoncentrationen og en eller flere indirekte parametre. Disse parametre kan være meteorologiske parametre som eksempelvis nedbør, vindretning og vindstyrke. Parametrene kan også relatere sig mere direkte til vandet som eksempelvis turbiditet, temperatur og salinitet. Sammenhængen kan findes ved hjælp forskellige regressionsmetoder, og varslingen bliver dermed foretaget på baggrund af måling af indirekte parametre. [De Brauwere, Ouattara og Servais, 2014]

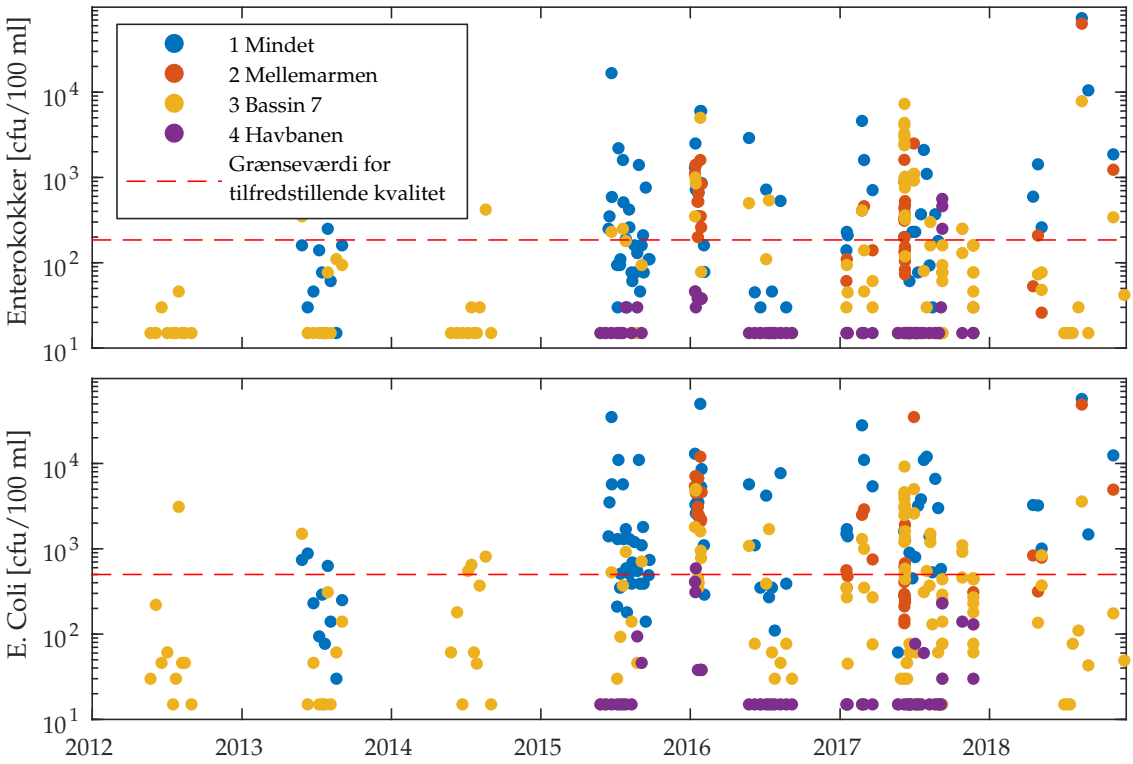
1.4 Badevandskvaliteten i Aarhus Havn

Siden 2012 er der fortaget en række målinger af fækale indikatorbakterier i Aarhus Havn. Målingerne er foretaget fire forskellige steder i havnen. Disse målepunkter fremgår af figur 1.3.



Figur 1.3. Målepunkter for bakteriemålinger i Aarhus Havn.

Det fremgår af figur 1.4, at bakterieindholdet i havnen ofte er over grænseværdien for tilfredsstillende badevandskvalitet.



Figur 1.4. Indhold af fækale indikatorbakterier i Aarhus Havn i perioden år 2012-2018. Målepunkterne fremgår af figur 1.2. Detektionsgrænsen er 15 cfu pr. 100 ml. Data fra Aarhus Kommune [2018a]

Det gennemsnitlige bakterieindhold er størst ved målepunkt 1 ved udløbet af Aarhus Å. Bakterieindholdet er generelt lavere ved Mellemarmen og Bassin 7. Ved Havbanen er bakterieindholdet under grænseværdien størstedelen af tiden. Det fremgår, at bakterieindholdet i havnen er så højt, at badevandskvaliteten ikke kan klassificeres som tilfredsstillende, med mindre der varsles, når bakteriekoncentrationen overskrider grænseværdien. Der er derfor behov for en nøjagtig varslingsmodel, således at brugerne kan adviseres, når vandkvaliteten er ringe.

1.5 Forbedring af badevandskvalitet i Aarhus havn

I år 2005 blev det af byrådet i Aarhus Kommune besluttet at igangsætte "Aarhus-Å-projektet". Det primære formål med projektet er at forbedre vandkvaliteten i Brabrand Sø, Aarhus Å og Aarhus Havn. Vandkvaliteten skal forbedres for at overholde de lovgivningsmæssige krav i forhold til at undgå permanent badeforbud og for at imødekomme de stigende krav til vandkvaliteten fra borgerne. Projektet understøtter en række andre projekter som genåbning af Aarhus Å gennem midtbyen, etablering af Aarhus Ø samt rekreativt brug af byens vandområder. [Aarhus Vand, 2018]

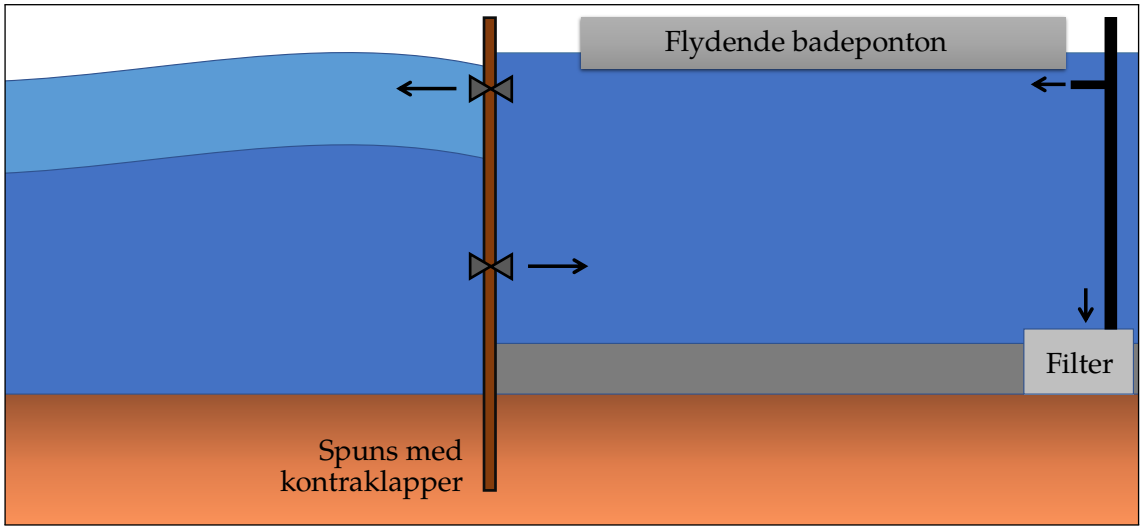
Projektet indeholder fire overordnede tiltag til at forbedre vandkvaliteten. Det mest omfattende er forbedring af afløbssystemet, herunder nedlægning af overløbsbygværker, etablering af forsinkelsesbassiner samt separatkloakering. Det andet tiltag er etablering af hygiejniseringsanlæg samt forbedret kapacitet på Åby og Viby renseanlæg for at forbedre rensningen. De sidste to tiltag indeholder styring af afløbssystemet på baggrund af

radardata, der optimerer måden, hvorpå afløbssystemt benyttes samt varsling af dårlig badevandskvalitet. [Aarhus Vand, 2018]

Efter "Aarhus-Å-projektet" har Aarhus Vand arbejdet videre med at forbedre hygiejniseringen på Åby og Viby renseanlæg. I 2016 og 2017 har Aarhus Vand undersøgt effektiviteten af hygiejniseringen på renseanlæggene. Undersøgelsen viste, at den gennemsnitlige reduktion af E. Coli og enterokokker var henholdsvis ca. 97,5 % og 93,7 % ved sandfiltrering og oxidering med peroxidet pereddikesyre. Dette medførte en koncentration umiddelbart efter udløbet fra renseanlægget svarende til udmærket badevandskvalitet i tørvejr. I regnvejrsforhold er hygiejniseringen imidlertid ikke tilstrækkelig til, at vandkvaliteten overholder badevandskvalitetskravene. [Allegue, 2017]

1.5.1 Forbedring af badevandskvaliteten i havnebadet i Bassin 7

I havnebadet i Bassin 7 er badevandskvaliteten forbedret ved at etablere en spuns rundt om havnebadet. I spunsen er der monteret et vandudskiftningssystem der, ved hjælp af to kontraventiler, udnytter tidevandet til at lukke dybtliggende havvand ind ved højvande og lukke højtliggende vand ud af bassinet ved lavvande. Samtidigt renses vandet internt i bassinet i et biologisk sandfilter. Af figur 1.5 fremgår en principskitse af rensesystemet i havnebadet. [Dalhoff, 2017]



Figur 1.5. Princip af rensesystem i havnebadet. Tegnet på baggrund af skitse fra Aarhus Kommune [2018a].

Konceptet bygger på teorien om, at det forurenede vand befinder sig øverst i vandsøjlen, mens det salte havvand nederst i vandsøjlen er renere. Det fremgår af badevandskvalitetsmålinger fra indersiden af spunsen, at systemet har den ønskede effekt, da der ikke er målt overskridelser af kvalitetskriteriet siden havnebadets åbning [Aarhus Kommune, 2019].

1.5.2 Nuværende varsling af badevandskvalitet i Aarhus Havn

I Aarhus Kommune er der på nuværende tidspunkt en varsling af badevandskvaliteten på baggrund af en deterministisk model. Modellen simulerer blandt andet vandets strømninger i tre dimensioner samt spredningen og henfaldet af bakterier. Inde i havnen bruges modellen til at varsle om badevandskvaliten i to punkter. Et punkt i Bassin 1, ved udløbet af Aarhus Å, hvor varslet er et permanent badeforbud. Det andet punkt er i selve havnebadet, inden for spunsen, hvor der altid varsles om god badevandskvalitet. [DHI, 2019]

Jævnfør DHI [2019] er der altså ingen varsling af badevandskvalitet i resten af havnen. Dette kan skyldes, at det er en udfording at tilpasse en deterministisk model til kunne simulere vandets strømninger og bakteriernes spredning i et komplekst havnesystem som Aarhus Havn.

1.6 Projektets mål og problemformulering

Målet med dette projekt er at kunne varsle om dårlig badevandskvalitet i Aarhus Havn i realtid. Behovet for realtidsvarsling er opstået, da flere områder i Aarhus Havn er blevet indraget til rekreative formål samtidig med, at vandkvaliteten i havnen i perioder er forringet i forhold til de gældende krav for badevandskvalitet. Det er omkostningstungt og tidskrævende at forhindre udledningen af urenset og renset spildevand. Derfor skal der etableres et system, der kan varsle brugerne af havnebassinerne om ringe badevandskvalitet for at kunne opretholde en tilfredsstillende klassifikation af badeområdet. Projektets formål leder frem til følgende problemformulering:

Hvordan kan badevandskvaliteten i Aarhus Havn varsles på baggrund af sensordata, der kan måles i realtid?

1.6.1 Fremgangsmåde

Varslingen skal basere sig på indirekte målinger som eksempelvis salinitet og temperatur, der kan foretages i realtid. På baggrund af de indirekte målinger opbygges en statistisk model til varsling af badevandskvaliteten.

Projektet består af fire dele; analyse af systemet, indsamling og analyse af data, udvikling af varslingsmodel samt implementering af varslingssystem. Systemanalysen vil indeholde en redegørelse for havnens opbygning, oplandets kilder til dårlig badevandskvalitet samt analyse af potentielle indikatorer på dårlig badevandskvalitet. Herefter indsamles og analyseres data for de relevante indikatorer for at udbygge den fysiske forståelse af systemet og hvorfor indikatorerne er egnede til varsling. Udvikling af varslingsmodellen gøres først ud fra simple regressioner på én inputparameter. Derefter inddrages machine learning som metode til udvikling af modeller med flere inputparametre. På baggrund af varslingsmodellen opbygges et komplet varslingssytem, der gør det muligt at implementere metoden til varsling af badende. Ydermere undersøges det hvorvidt varslingsmetoden egner sig til varsling af badevandskvalitet andre steder end i Aarhus Havn, hvilket samles i en perspektivering.

Kapitel 2

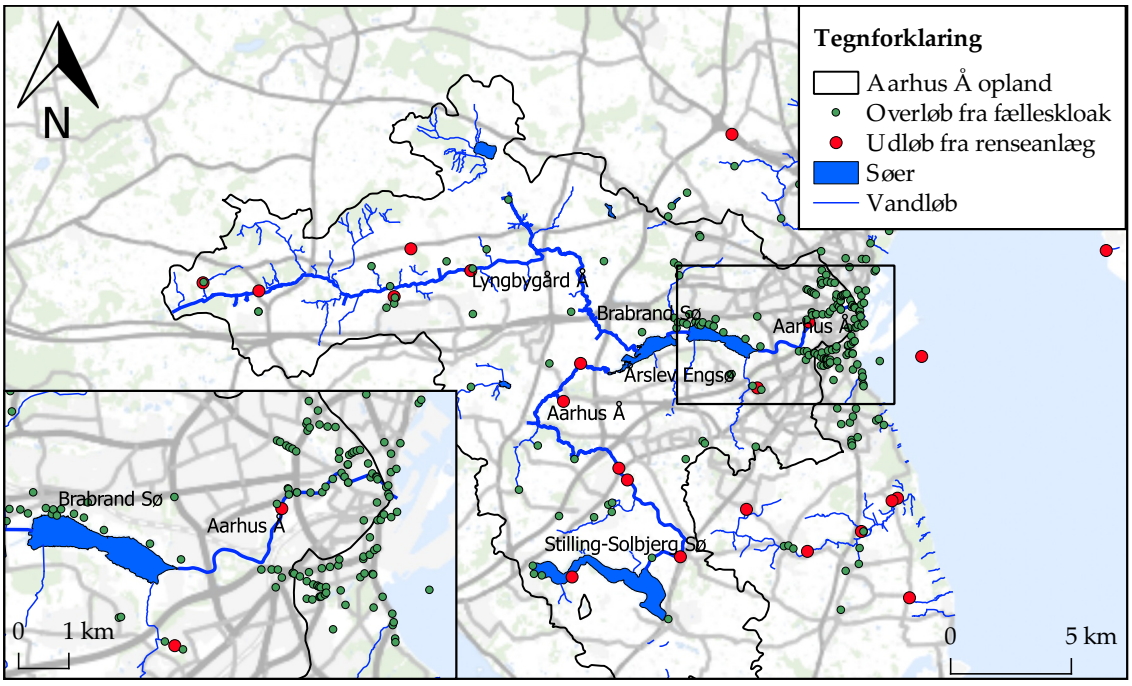
SYSTEMANALYSE

I dette kapitel dannes et overblik over systemet således at der kan identificeres mulige indikatorer til bestemmelse af badevandskvaliteten i Aarhus Havn.

Forurening af badevand er påvirket af det fysiske system, der omgiver lokaliteten. Det er derfor nødvendigt at undersøge, hvordan disse forhold er i Aarhus Havn for at kunne vurdere badevandskvaliteten. I analysen indgår lokalisering af potentielle kilder, redegørelse for det hydrologiske opland samt identifikation og undersøgelse af potentielle indikatorer på dårlig badevandskvalitet.

2.1 Opland til Aarhus Å

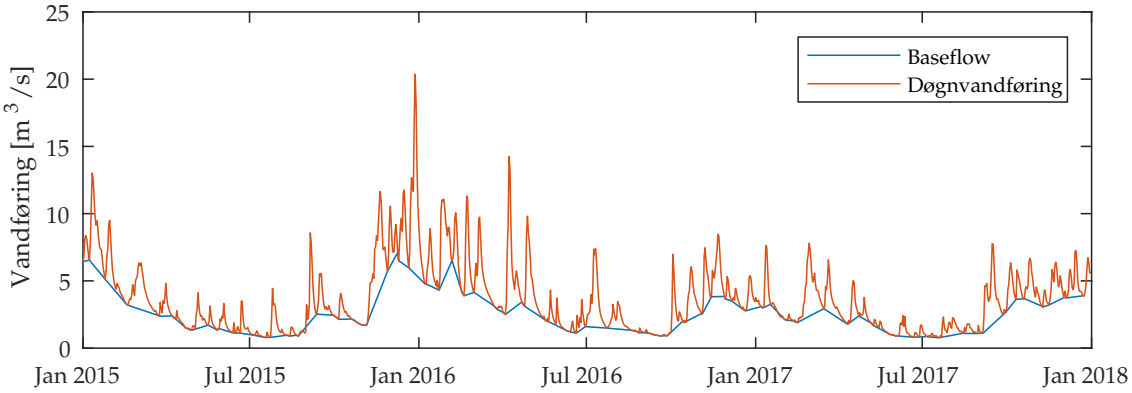
Aarhus Å har sit udløb i Aarhus Havn og bidrager med størstedelen af ferskvandstilførslen til havnen. Aarhus Å er recipient til en lang række regnbetingede udløb fra fælleskloak og dermed ofte påvirket af spildevand. Derudover har to af renseanlæggene i Aarhus Midtby sit udløb til Aarhus Å og bidrager dermed også med en belastning til Aarhus Havn.



Figur 2.1. Overblik over oplandet til Aarhus Å. Naturdata fra Miljøministeriet [2010]. Udløb og overløb fra Aarhus Kommune [2016].

Aarhus Å udspringer fra Stilling-Solbjerg Sø og strømmer videre til Årslev Engsø og Brabrand Sø, hvorefter den har sit udløb i Aarhus Bugt. Lyngbygård Å, der er den anden hovedstrækning i oplandet, strømmer fra Låsby i vest og videre til Årslev Engsø, hvor den løber sammen med Aarhus Å.

Aarhus Å er et regnvandspåvirket vandløb, da baseflowet er lavt i forhold til peak-flowet, der opstår i regnfulde perioder. Dette fremgår af figur 2.2, der viser døgnvandføringen og baseflowet for Aarhus Å. De hurtige ændringer i vandføringen skyldes, at en stor del af oplandet er befæstet, og at der dermed er stor afstrømning ved store regnhændelser. De bredere peaks repræsenterer de længere regnfulde perioder og den langsomme afstrømning. Langsom afstrømning stammer blandt andet fra marker, grøfter og bassiner.



Figur 2.2. Døgnvandføring og baseflow i Aarhus Å ved udløbet til Aarhus Havn. Data fra Danmarks Miljøportal [2018].

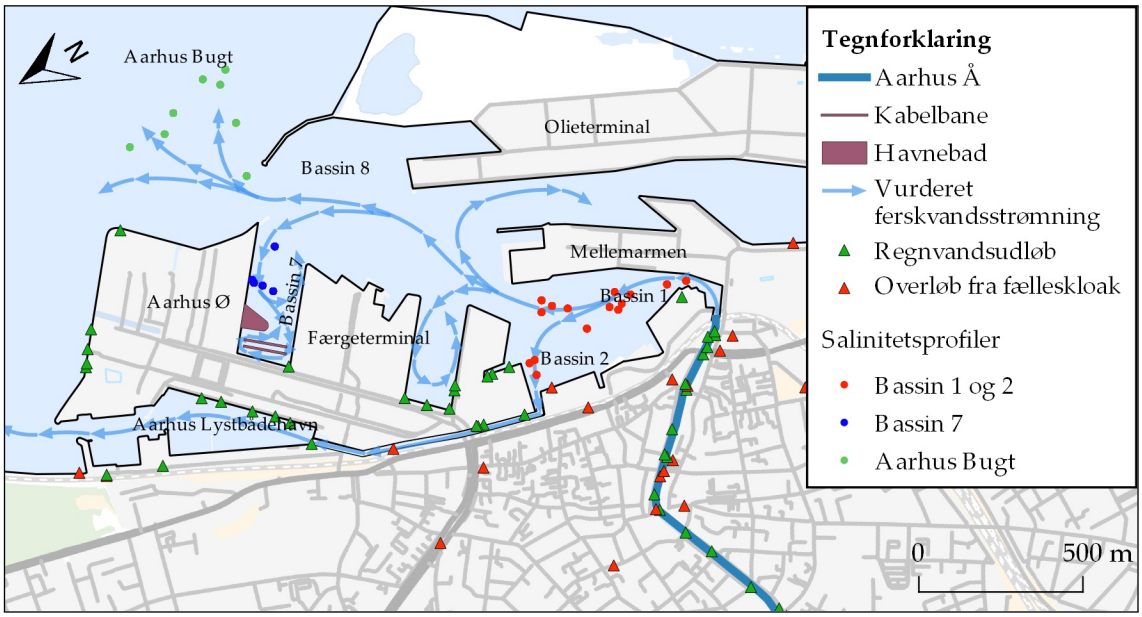
Årslev Engsø er en reetableret sø med et areal på ca. 100 hektar. Området har tidligere været drænet, men er blevet genetableret som vådområde for at reducere næringsstofbelastningen til Brabrand Sø og Aarhus Bugt. Opholdstiden i Årslev Engsø er meget varierende og forholdsvis kort; gennemsnitligt 4 dage, men op til 20 dage i sommerhalvåret. Søen er omgivet af afgræssede enge, hvilket kan være en kilde til fækal forurening. [Miljøministeriet, 2010]

Aarhus Å strømmer fra Årslev Engsø videre til Brabrand Sø. Brabrand Sø har et opland på ca. 150 ha og en gennemsnitlig opholdstid på ca. 7 dage. Brabrand Sø er tidligere kendt for at være stærkt forurenet af spildevand. Aarhus Vand har de seneste år arbejdet på at forbedre vandkvaliteten for at forbedre de rekreative forhold i søen [Aarhus Vand, 2018]. Søen er dog stadig recipient for en række regnbetingede udløb, hvilket fremgår af figur 2.1. Det fremgår, at den største tæthed af overløbsbygværkerne findes ved og efter Brabrand Sø.

Udover de regnbetingede overløb fra fælleskloak og renseanlæg udledes der renset spildevand til Aarhus Å. I Aarhus Midtby udledes der renset spildevand fra Åby Renseanlæg og Viby Renseanlæg til Aarhus Å. Viby Renseanlæg udleder op til 10 mio. m^3 renset spildevand pr. år til Døde Å, der er et lille tilløb til Brabrand Sø. Åby Renseanlæg udleder op til 6 mio. m^3 renset spildevand pr. år direkte til Aarhus Å. Dette svarer til en samlet udledning på 0,5 m^3 pr. sekund i gennemsnit. Udover udledningerne til Aarhus Å bliver der udledt renset spildevand til Aarhus Bugt fra blandt andre Marselisborg Renseanlæg og Egå Renseanlæg. [Aarhus Kommune, 2016]

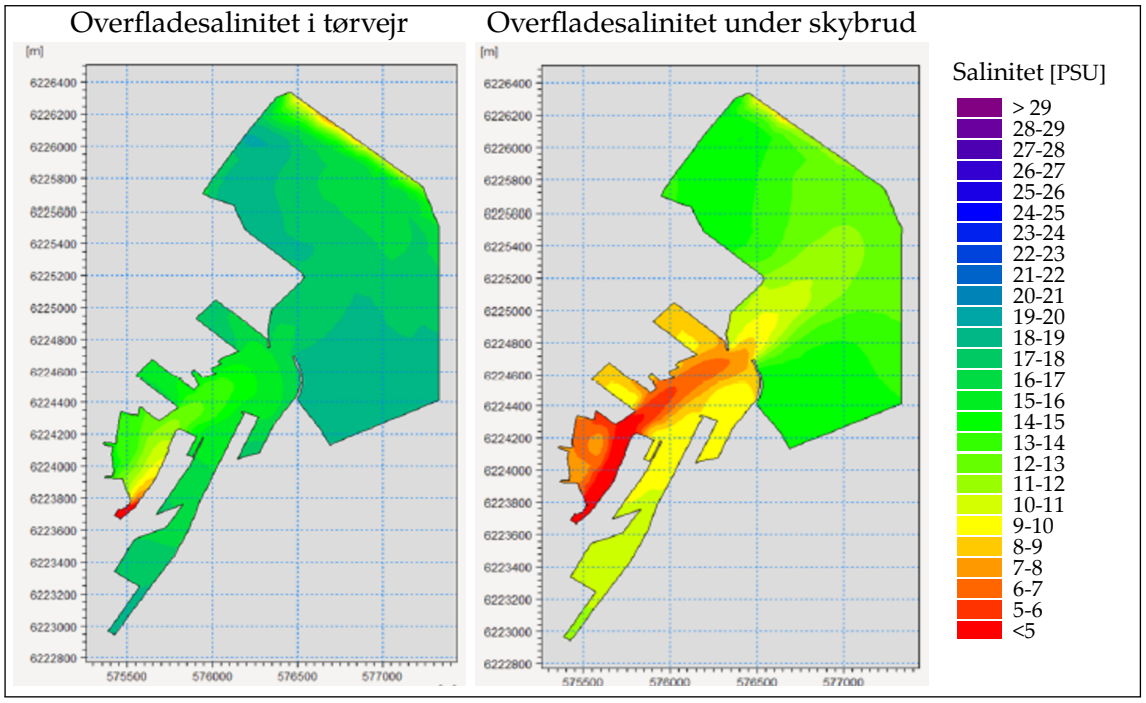
2.2 Havnesystemet i Aarhus

Aarhus Å har sit udløb til Aarhus Havn vest for Mellemarmen, hvorefter vandet strømmer videre gennem Bassin 1 og 8 og ud mellem pieren på Olieterminalen og Aarhus Ø. Ved udløbet til Aarhus Havn er Aarhus Å opstemt ved et stemmeværk, der skal opretholde vandstanden i åen. Fra Bassin 1 strømmer vandet desuden videre igennem Bassin 2 og videre gennem en ny kanal, der er etableret i forbindelse med udbygningen af Aarhus Ø, til Aarhus Lystbådehavn. Af figur 2.3 fremgår et overblik over Aarhus Havn fra udløbet af Aarhus Å til Aarhus Bugt på ydersiden af havnen.



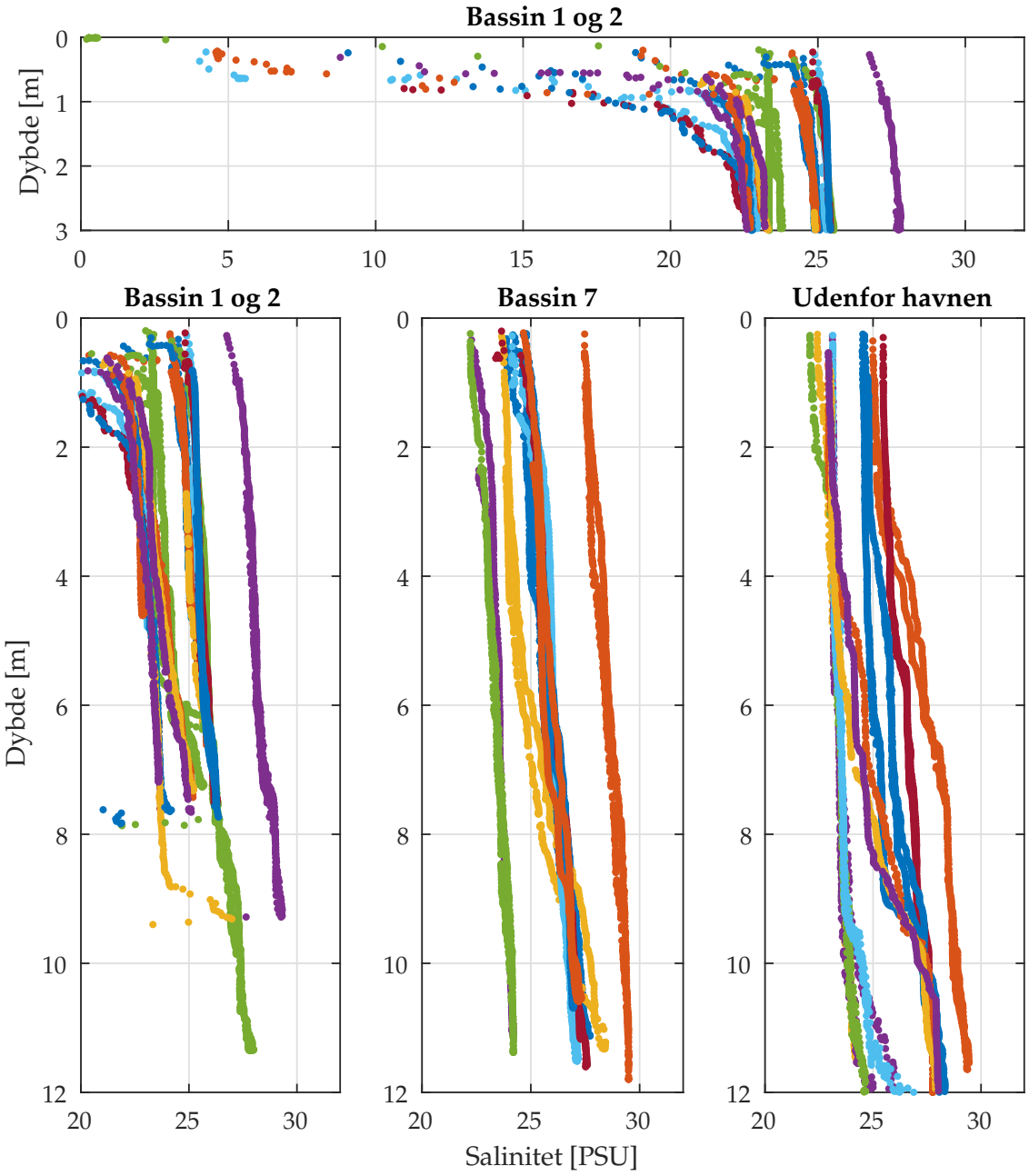
Figur 2.3. Overblik over havnen fra udløbet af Aarhus Å til Aarhus Bugt. Data fra Danmarks Miljøportal [2018] og Aarhus Kommune [2018a].

Tilførslen af ferskvand til Aarhus Havn stammer primært fra Aarhus Å. Udover Aarhus Å udledes der vand fra en række regnvandsudløb fra separatkloakerede områder. Kilderne til forurenet vand er udover Aarhus Å de overløbsbygværker fra fælleskloakerede områder, der udleder til havnen. Ferskvandsstrømningen på 2.3 er vurderet på baggrund af modelresultater fra en MIKE-3 model udarbejdet af Cowi [COWI, 2017]. På figur 2.4 fremgår disse modelresultater, der viser at hele havnen er ferskvandspåvirket i tørvejr. Ved kraftige regnhændelser er saltindholdet i størstedelen af havnen halveret og er dermed meget ferskvandspåvirket.



Figur 2.4. Deterministisk model af overfladesaliniteten. Modellen er en tredimensionel DHI-model udarbejdet af COWI [2017].

Når det ferske vand fra Aarhus Å strømmer ud i Aarhus Havn, bliver det til dels opblandet med havvandet. Da åvandet er ferskt, og dermed har en lavere densitet end det salte havvand, vil der kunne opstå lagdelte strømninger. På figur 2.5 fremgår 64 salinitetsprofiler fra Aarhus Havn. Profilerne er inddelt i tre kategorier ud fra hvor i havnen, de er foretaget. Profilernes placering fremgår af figur 2.3. Det fremgår, at salinitetsprofilerne afviger en del fra det modellerede på figur 2.4. Dette kan skyldes, at modellen er kalibreret op imod ét salinitetsprofil, hvor overfladesaliniteten er væsentligt lavere end på figur 2.5 [COWI, 2017].

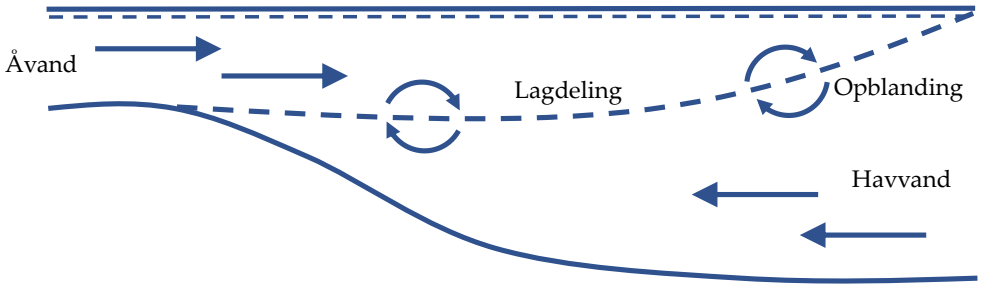


Figur 2.5. Salinitetsprofiler ned gennem vandsøjlen. Farverne adskiller de forskellige profiler. Profilerne er udført ved de på figur 2.3 angivne lokationer i perioden juni-september 2017. Bemærk de forskellige akser. Data fra Aarhus Kommune [2018a].

Af figur 2.5 fremgår det, at der i Bassin 1 og 2 er en tendens til, at der opstår lagdeling. Dette er forventet, da det er her Aarhus Å strømmer ud. Samtidig fremgår det, at der er en lille tendens til lagdeling udenfor havnen i 9 meters dybde, hvilket der ikke fremgår ved Bassin 7, selvom Bassin 7 umiddelbart er tættere på den store ferksvandskilde Aarhus Å. Dette kan skyldes, at vandets vej fra Aarhus Å til Aarhus Bugt er en mere lige strømning uden lige så mange forstyrrelser, der skaber hvirvelstrømninger, som er med til at skabe opblanding mellem havvand og ferskvand. Profilerne, der fremgår af figuren, er målt på 7 forskellige dage henover sommeren 2017. Det fremgår, at der er en forskel på profilerne fra måling til måling. Dette skyldes de forskellige forhold, der er med til at frembringe en

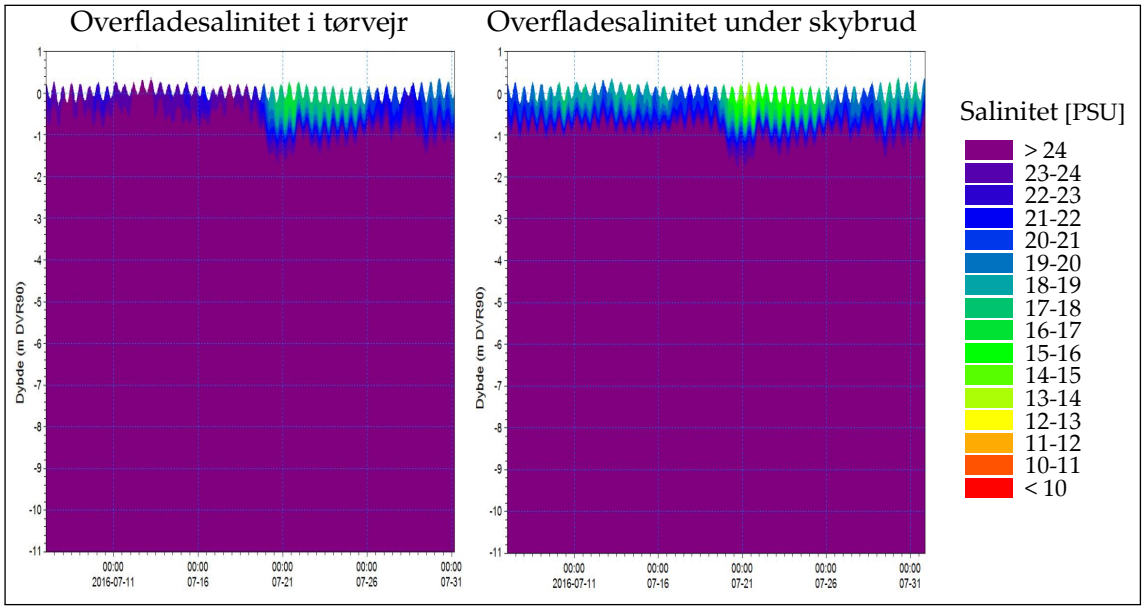
lagdeling som eksempelvis mængden af ferksvandstilstrømning og vindpåvirkningen.

Salinitetsprofilerne viser, at der er en lagdeling, som bliver svagere jo længere ud af havnen, man kommer. På figur 2.6 er de vertikale strømninger illustreret. Figuren viser, at lagdelingen brydes jo længere ud i havnen, man kommer.



Figur 2.6. Konceptuel model af strømningerne gennem havnesystemet i længdesnit.

Konceptet, med at åvandet transporteres lagdelt ud gennem havnen, underbygges af den deterministiske MIKE-3 model udarbejdet af COWI. To lodrette profiler af saliniteten i Bassin 7 fra denne model fremgår af figur 2.7. Figuren viser salinitetsprofilet i tørvejr og under en regnhændelse.



Figur 2.7. Lodret profil af saliniteten i Bassin 7. Profilerne er lavet på baggrund af en deterministisk tredimensionel DHI-model udarbejdet af COWI [2017].

Af modelresultaterne fremgår det, at åvandet med en lav salinitet breder sig ud lagdelt ud gennem havnen til Bassin 7. Modellen viser desuden, at der ved større regnhændelser er en endnu større mængde ferskvand, og at udbredelsen i dybden er større.

2.3 Indikatorer på dårlig badevandskvalitet

Ud fra de to foregående afsnit vurderes det at være forurenet vand fra Aarhus Å, der udgør en risiko for badevandskvaliteten i flere dele af Aarhus Havn. Forureningen af Aarhus Å kan skyldes regnbetingede udledninger, udledning af renset spildevand og udledning af fækalier fra græssende dyr langs Årslev Engsø. Grundet den lange analysetid for bakteriekoncentration vil det ikke være muligt at lave real-tids varsling på baggrund af direkte målinger af bakterie-indhold. Dermed er det nødvendigt at vurdere badevandskvaliteten ud fra parametre, der kan måles i realtid og som giver en indirekte indikation af badevandskvaliteten.

Da det vurderes, at det er åvand, der giver anledning til forringet badevandskvalitet, kan det være muligt at vurdere risikoen for forurening ved at undersøge parametre, der er forskellige i åvand og havvand som eksempelvis salinitet og temperatur. Varslingen kan muligvis også baseres på andre parametre som nedbør, vandføring i Aarhus Å eller vindretning. Disse parametre siger noget om, hvor meget åvand der ledes til havnen, og hvilken retning vinden påvirker åvandet. I de følgende afsnit undersøges det, om der umiddelbart kan findes en sammenhæng mellem nedbørsmængder, vandføring, vindretning eller salinitet og forringet badevandskvalitet. Formålet med dette er at identificere de parametre, der kan benyttes til varsling af dårlig badevandskvalitet i Aarhus Havn. På figur 2.8 fremgår placeringen af de målestationer, hvor der er indsamlet data til undersøgelsen.

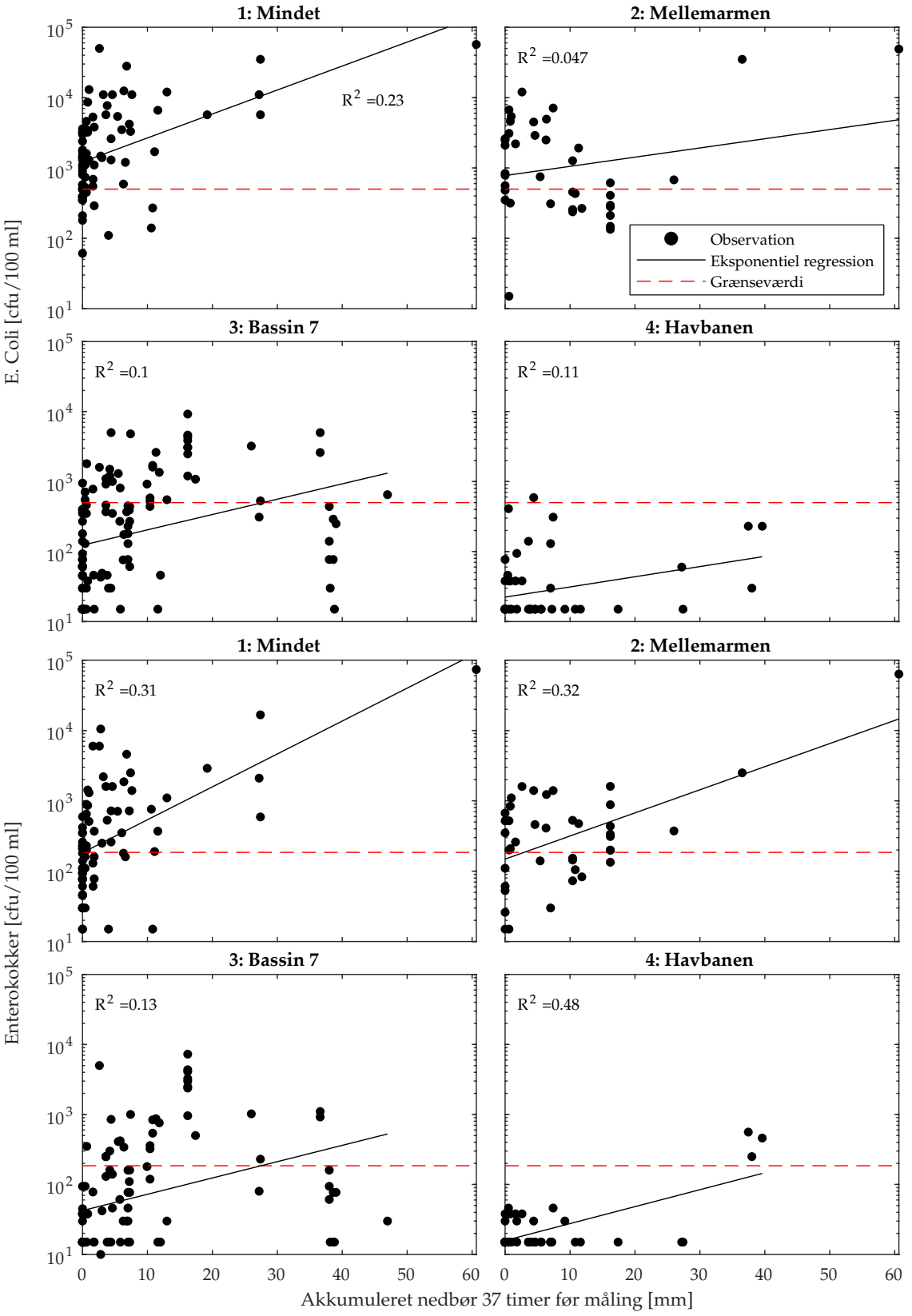
Figur 2.8. Placering af klimastationer til måling af nedbør, vandføring og vindforhold samt målepunkter til indsamling af vandprøver.

Yderligere information omkring oprindelsen af data fremgår af appendiks B.

2.3.1 Nedbørsmængde som indikator for dårlig badevandskvalitet

En af årsagerne til dårlig badevandskvalitet kan være udledning af fækale bakterier ved regnbetingede udledninger fra fællessystemer. De regnbetingede udledninger forårsages af nedbør, der ikke er kapacitet til i afløbssystemet. Det er vanskeligt at måle hvor meget spildevand, der udledes til Aarhus Å og dermed videre ud i havnen via. regnbetingede udløb. Det kan derfor være muligt at anvende nedbørsmængden til at vurdere i hvor stor grad, der har været overløb fra fællessystemet til vandmiljøet og dermed en mulig forringelse af badevandskvaliteten. På baggrund af denne hypotese undersøges det i det følgende, om en stor nedbørsmængde kan korreleres med en høj bakteriekoncentration i Aarhus Havn.

For at sammenholde baktieriemålingerne med nedbør akkumuleres nedbørsmængden fra prøvetagningstidspunktet og 37 timer tidligere. Denne akkumuleringstid benyttes, da det er her det bedste fit til en eksponentiel funktion opnås. Af figur 2.9 fremgår bakteriekoncentrationerne sammenholdt med nedbørsmængden.

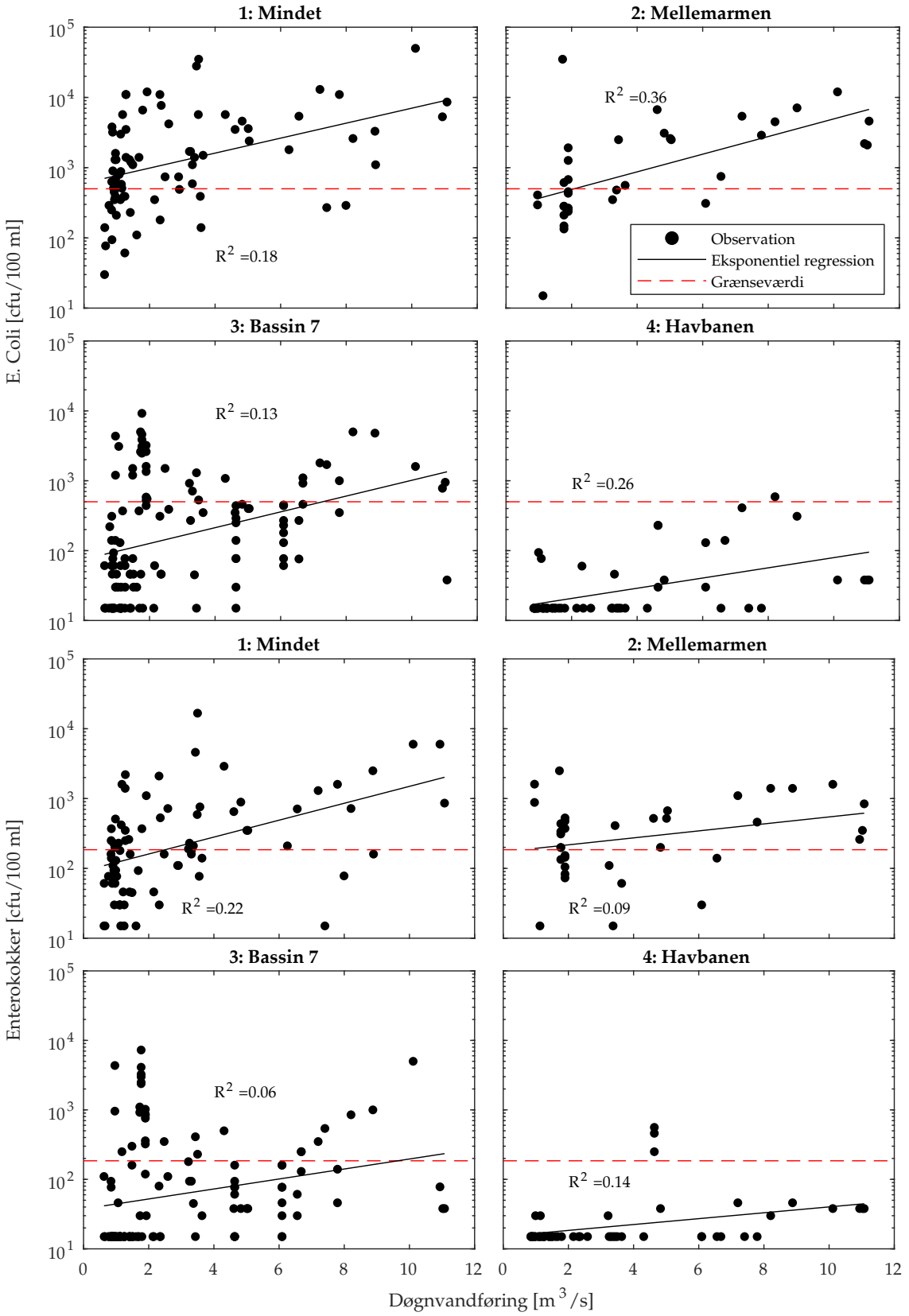


Figur 2.9. Bakteriekoncentrationen ved målepunkt 1, 2, 3 og 4 i Aarhus Havn sammenholdt mod akkumuleret nedbør i de 37 forregående timer. Data fra Spildevandskomiteens regnmålersystem [2019] og Aarhus Kommune [2018a].

Som det fremgår af figur 2.9, er der ikke et entydigt billede af sammenhængen mellem akkumuleret nedbør og bakteriekoncentration. Det fremgår dog, at meget store regnhændelser ofte giver overskridelse af badevandskriteriet. For de bakteriemålinger, der er lavet i perioder med lidt eller ingen nedbør, fremgår det, at der er en stor variation, hvilket indikerer, at bakteriekoncentrationen også afhænger af andet end nedbør.

2.3.2 Vandføring som indikator for dårlig badevandskvalitet

Et alternativ til nedbør som indikator på regnbetinget udledning af spildevand er vandføringen i Aarhus Å. Af figur 2.2 fremgår døgnvandføringen og baseflowet i Aarhus Å for år 2015-2017 ved udløbet. Det fremgår af figuren, at baseflowet i Aarhus Å er relativt lavt i forhold til de regnbetingede peaks. Dette betyder, at vandløbet er regnvandspåvirket, og vandføringen kan derfor muligvis bruges som et udtryk for overløbsmængden i oplandet. Derfor undersøges det, om der er en korrelation mellem bakteriekoncentrationen og vandføringen i Aarhus Å. Af figur 2.10 fremgår koncentrationerne af henholdsvis E. Coli og enterokokker for de 4 målepunkter i Aarhus Havn sammenholdt med døgnvandføringen ved udløbet af Aarhus Å.



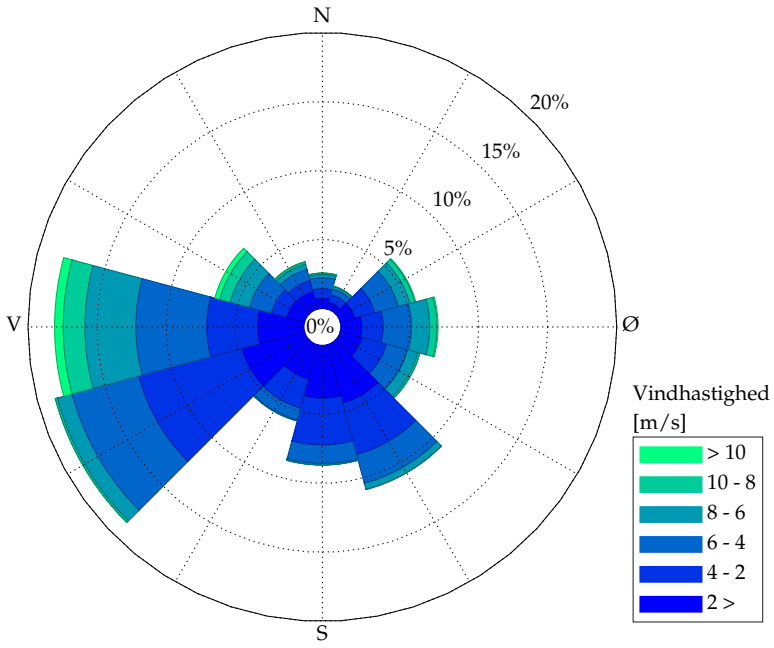
Figur 2.10. Bakteriekoncentrationen ved målepunkt 1, 2, 3 og 4 i Aarhus Havn sammenholdt mod døgnvandføringen i Aarhus Å. Data fra Danmarks Miljøportal [2018] og Aarhus Kommune [2018a].

Af figur 2.10 fremgår det, at der ikke er en god korrelation mellem døgnvandføringen og bakteriekoncetrationen. Det fremgår, at der selv ved lave vandføringer kan være et højt indhold af bakterier. Ved meget høje vandføringer er der dog typisk en høj bakteriekoncentration ved målestationerne inde i havnen.

Generelt viser undersøgelsen af nedbør og døgnvandføring, at der ikke er en overbevisende korrelation til bakteriekorrelationen. Dette kan skyldes, at bakterieindholdet i havnen ikke kun afhænger af mængden af nedbør, da koncentrationen i Aarhus Å ofte er høj i tørvejr og ikke bliver væsentligt højere efter regnhændelser.

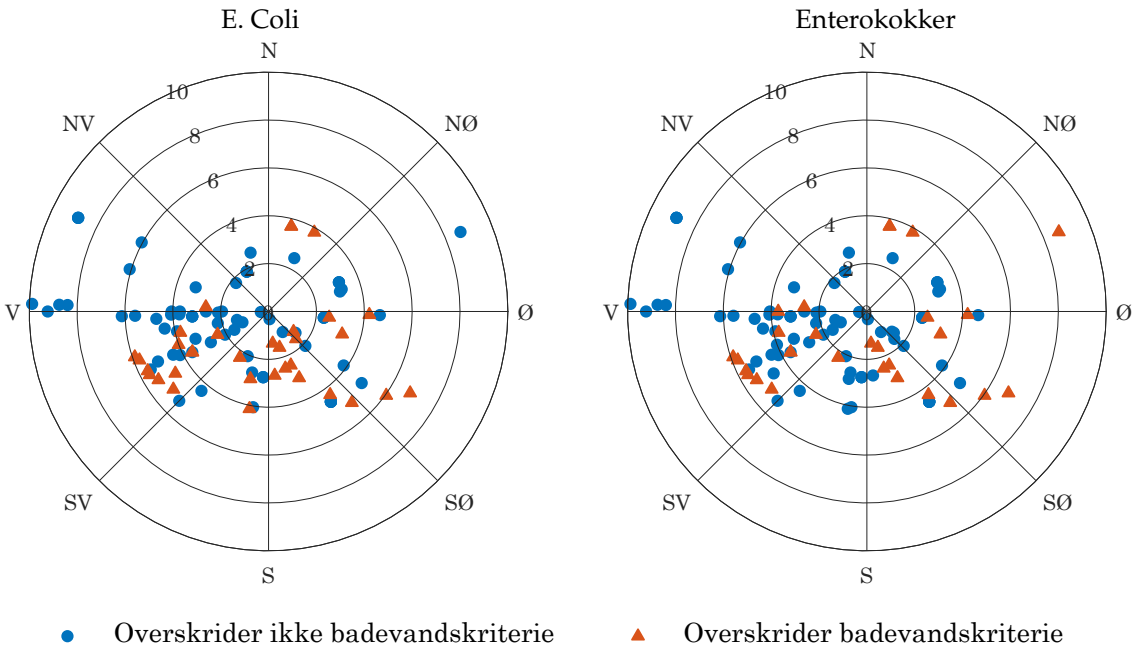
2.3.3 Vindretning og -hastighed som indikator for dårlig badevandskvalitet

Vandets strømninger i havnen forventes at være påvirket af vindretning og -hastighed. Det undersøges derfor, om der er en sammenhæng mellem vindpåvirkning og badevandskvalitet. Af figur 2.11 fremgår fordelingen af vindforholdene i Aarhus. Det fremgår, at det oftest blæser fra vest og sydvest, og at kraftige vindhastigheder på over 10 m/s kun opstår ved vestenvind.



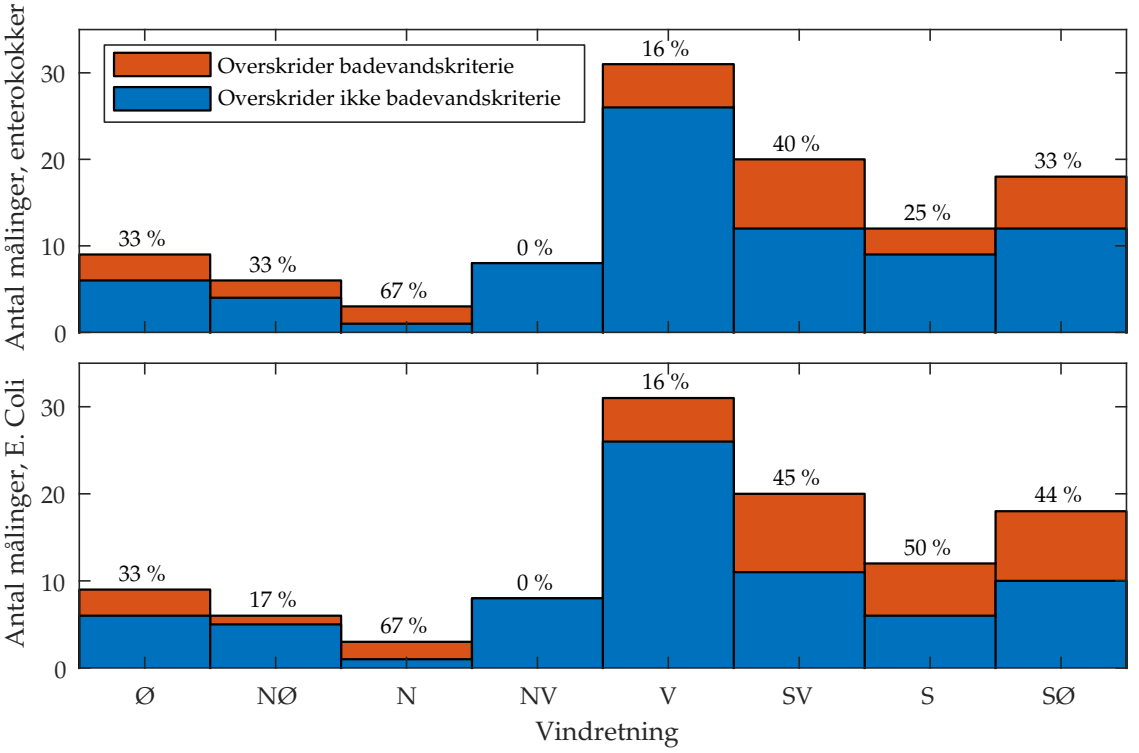
Figur 2.11. Fordeling af vindretning og -hastigheder i Aarhus. Målestationens placering fremgår af figur 2.8. Data fra The Weather Company [2018].

På figur 2.12 fremgår det, hvorvidt målingerne for badevandskvalitet ved målepunkt 3 i Bassin 7 overskrider badevandskriteriet sammenholdt med vindretning- og hastighed. Det fremgår af figuren at størstedelen af målingerne med dårlig badevandskvalitet kan korreleres med vind fra de sydlige retninger uden nogen signifikant korrelation med vindhastigheden. Det fremgår desuden, at densiteten af målinger foretaget ved vind fra de sydlige retninger er større end fra de nordlige retninger.



Figur 2.12. Sammenhæng mellem vindpåvirkning og badevandskvalitet ved målepunkt 3 i Bassin 7. Vindpåvirkningen er midlet over de foregående 6 timer. Afstanden fra cirklens centrum angiver vindhastighed i m/s. Data fra The Weather Company [2018] og Aarhus Kommune [2018a].

Fordelingen af målinger ud fra vindretning fremgår af figur 2.13.

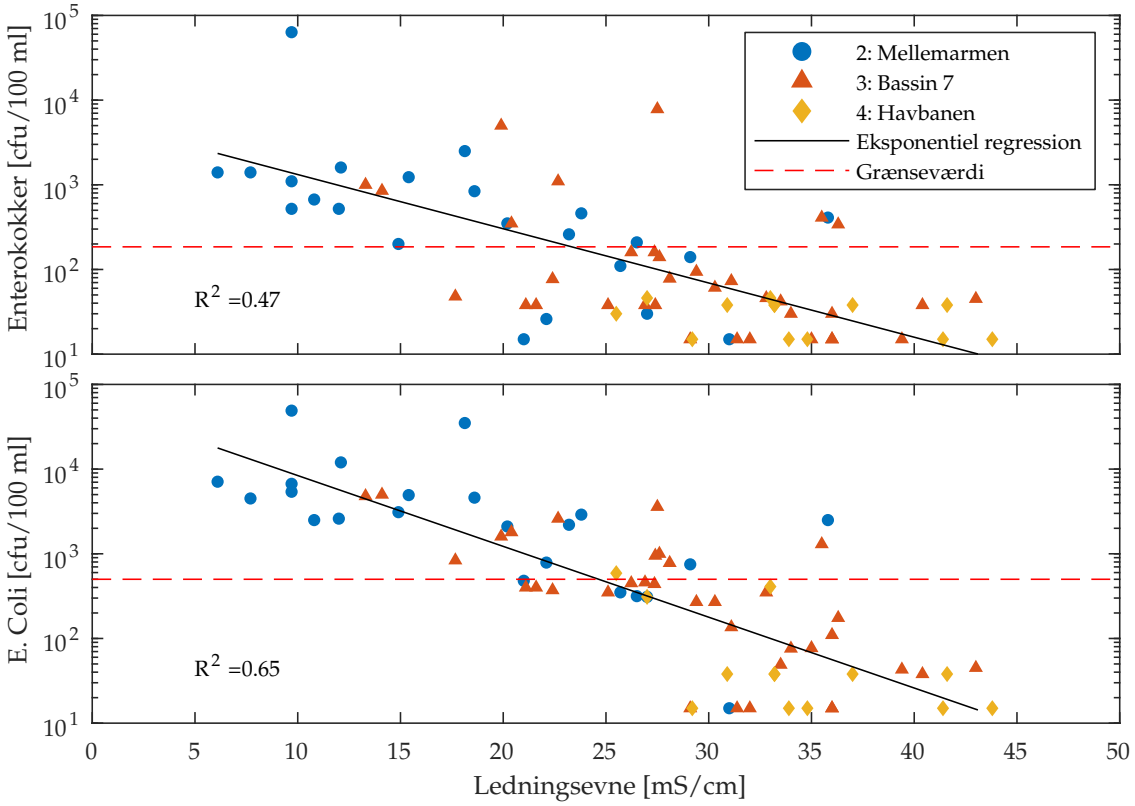


Figur 2.13. Antal målinger foretaget ved målepunkt 3 i Bassin 7 fordelt på den gennemsnitlige vindretning i de foregående 6 timer. Procentangivelsen angiver, hvor stor en andel af bakteriemålingerne, der overskrider badevandskriteriet. Data fra The Weather Company [2018] og Aarhus Kommune [2018a].

Af figur 2.13 fremgår det, at der er foretaget flest målinger ved vind fra vest. Samtidig fremgår det, at den største procentdel af målinger med dårlig badevandskvalitet er ved vind fra de sydlige retninger. Dette er vurderet på baggrund af, at de tre målinger ved vind fra nord ikke kan give en repræsentativ tendens. Overordnet set er der en korrelation mellem vinden og badevandskvaliteten i Bassin 7, hvilket formentlig har en sammenhæng med, hvordan vinden påvirker transporten og opblanding af åvandet i havnen.

2.3.4 Salinitet som indikator for dårlig badevandskvalitet

Vandføring og vindretning er to parametre, der ikke direkte relaterer sig til opblandingsgraden mellem åvandet og havvandet. Det er derfor nødvendigt at undersøge sammenhængen mellem opblandingsgraden og bakteriekoncentrationen på baggrund af andre parametre. Da havvand er salt og åvand er ferskt, er saliniteten en indikator på, hvor stor en andel af vandet i bassinet der består af vand fra Aarhus Å. Saliniteten er direkte sammenhængende med den elektriske ledningsevne af vandet, hvorfor denne benyttes som mål for saliniteten. På denne baggrund undersøges en sammenhæng mellem koncentrationen af henholdsvis E. Coli og enterokokker og den elektriske ledningsevne.



Figur 2.14. Sammenhæng mellem elektrisk ledningsevne og bakteriekoncentrationer. R^2-værdi angiver korrelationen mellem de målte værdier og den eksponentielle regression. Målingerne er foretaget ved målepunkt 2, 3 og 4. Data fra Aarhus Kommune [2018a].

Figur 2.14 viser sammenhængen mellem ledningsevnen og bakteriekoncentrationen i målepunkterne 2, 3 og 4. Årsagen til at målepunkt 1 ikke er inddraget i grafen er, at dette

punkt ligger lige ved udløbet af Aarhus Å og dermed har en konstant lav ledningsevne.

Den eksponentielle regression viser, at der er en signifikant sammenhæng mellem bakteriekoncentrationen og ledningsevnen med en R^2 på henholdsvis 0,47 og 0,65 for enterokokker og E. Coli. Denne korrelation bekræfter hypotesen om, at bakteriekoncetrationen i Aarhus Havn afhænger af opblandingsgraden mellem ferskvand og havvand. Sammenhængen kan med fordel benyttes til at vurdere badevandskvaliteten indirekte.

2.4 Opsummering

I Aarhus Havn bidrager Aarhus Å med en stor ferskvandstilførsel, der oftest har en stor bakteriekoncentration. Densitetsforskellen på ferskvandet og havvandet gør, at der opstår en lagdeling af vandet i havnen. Lagdelingen er størst ved udløbet til havnen og aftager gradvist ud af havnen. I afsnit 2.3 er det undersøgt, om nedbør, vandføring, vind eller ledningsevnen kan bruges som indikator for dårlig badevandskvalitet. Den parameter, der giver den bedste korrelation til bakteriekoncentrationen, er den elektriske ledningsevne. Korrelationen viser, at lavere ledningsevne relaterer sig til høje bakteriekoncentrationer. Derudover er der en tendens til, at vindforholdene i havnen kan være betydende for bakteriekoncentrationen, samt at kraftig nedbør ofte resulterer i dårlig badevandskvalitet.

Kapitel 3

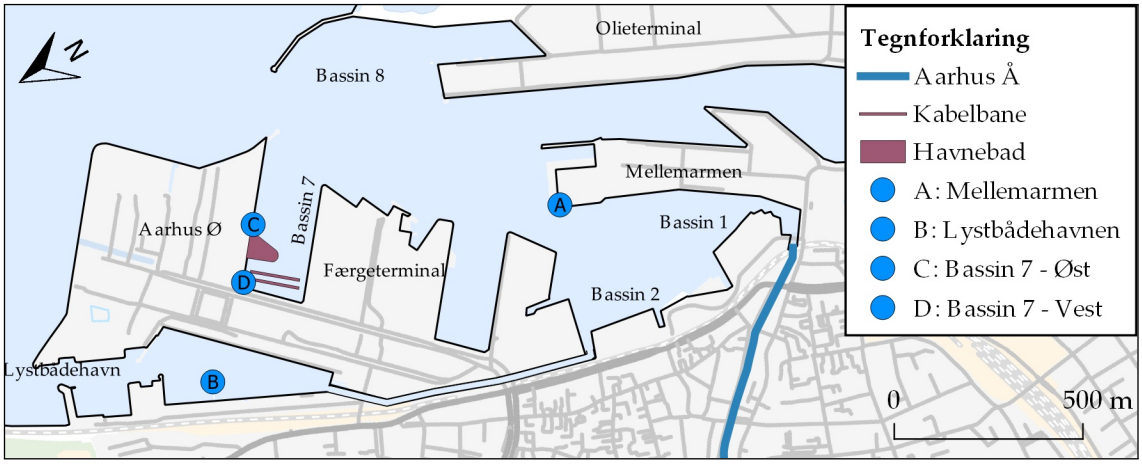
UNDERSØGELSE AF ELEKTRISK LEDNINGSEVNE

I kapitel 2 er korrelationen mellem koncentrationen af fækale bakterier i Aarhus Havn og en række parametre undersøgt. Disse undersøgelser viser, at den bedste korrelation er mellem bakteriekoncentrationen og den elektriske ledningsevne af vandet. På baggrund af dette vil der i dette kapitel foretages yderligere undersøgelser af ledningsevnen for at kortlægge, hvordan den varierer, og hvad den påvirkes af.

Præmissen for brug af ledningsevnen, som indikator for badevandskvalitet, er, at forureningen med fækale bakterier skyldes udledning af ferskvand til havnen fra Aarhus Å, der typisk har et stort bakterieindhold. Da ledningsevnen er lineært afhængig af saliniteten af vandet, kan den bruges til at indikere, hvor stor en andel af vandet i havnen der udgøres af det salte havvand [Wenner, Smith og Soule, 1930].

3.1 Dataindsamling

Til at undersøge de tidslige og stedslige variationer af ledningsevnen er der opstillet fire målestationer, der måler ledningsevnen 30 cm under vandoverfladen. De tre af målestationerne er offline HOBO ledningsevne-loggere fra Onset Computer Corporation [2019]. Disse er indstillet til at måle ledningsevnen og temperaturen med et interval på 5 minutter. Den sidste måler er en online OTT ledningsevne-logger fra OTT HydroMet [2019], der er instillet til at måle ledningsevnen og temperaturen 3 gange i timen. Målerne er interkalibreret lineært som beskrevet i appendiks C, for at kunne sammenligne observationerne. Målernes placering fremgår af figur 3.1.



Figur 3.1. Placering af ledningsevnemålere i Aarhus Havn.

Placeringen af målestationerne er foretaget ud fra en bagtragtning af, hvor der er rekreative aktiviteter, der kan blive påvirket af dårlig badevandskvalitet. Derudover er målestationerne placeret således, at der måles i forskellige afstande fra Aarhus Å. Dette

gøres for at undersøge opblandingen af ferskvand i havnebassinet. I Bassin 7 er der to målestationer. Det er valgt at placere to målstationer inde i Bassin 7 for at undersøge, hvor stor variation der er indenfor et mindre område. På figur 3.2 fremgår målestation A og B. Her fremgår det, at HOBO ledningsevne-loggeren er placeret under en bøje, der er fortøjret til kajkanten. Bøjen følger tidevandet, og der måles dermed altid i samme dybde. Der er placeret en tennisbold på loggeren for at beskytte den mod slag.

Figur 3.2. Målestation B til venstre og målestation A til højre.

3.2 Statistisk overblik over ledningsevnedata

For at vurdere den tidslige og stedslige variation i ledningsevnen skabes der først et overblik over de indsamlede data. Da ledningsevnen er afhængig af temperaturen, er rådata temperaturkorrigeret. Temperaturkorrektionen fortages efter den lineære sammenhæng, der fremgår af formel 3.1 [Sorensen og Glass, 1987].

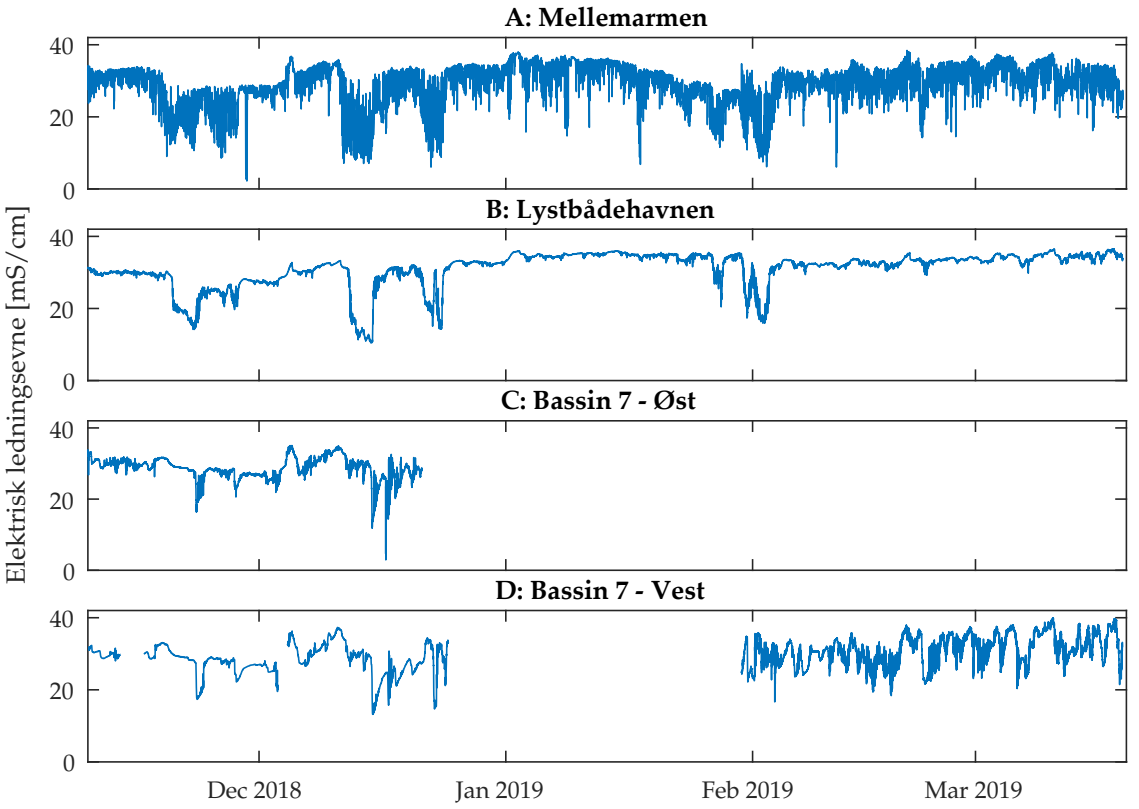
$$EC_{25} = \frac{EC_t}{1 + a \cdot (t - 25)} \tag{3.1}$$

Hvor:

EC_{25}	Elektrisk ledningsevne ved 25 °C [$\frac{\text{mS}}{\text{cm}}$]
EC_t	Elektrisk ledningsevne ved temperaturen t [$\frac{\text{mS}}{\text{cm}}$]
a	Temperaturkoefficient [$°\text{C}^{-1}$]
t	Temperatur [°C]

Temperaturkoefficienten er eksperimentielt bestemet til 0,019 pr. grad. Metoden og eksperimenterne er nærmere beskrevet i appendiks D. Alle ledningsevner i dette projekt er korrigeret til ledningsevnen ved 25 °C, således at de kan sammenlignes på tværs af datasæt.

På figur 3.3 fremgår tidsserierne over loggede data fra de fire ledningsevnemålere i Aarhus Havn.



Figur 3.3. Ledningsevnen ved målestation A, B, C og D i Aarhus Havn. Ledningsevnen er målt 30 cm under vandets overflade. Manglende målinger for målestation D i januar 2019 skyldes udfald i målestationen.

Der er indsamlet data fra november 2018 til marts 2019. Dette er ikke helt idéelt, da det ikke er i badesæsonen, men det giver dog stadig et indblik i den stedslige og tidslige variation i ledningsevnen i havnen. For Bassin 7 - øst er der indsamlet data i en kortere periode for at undersøge korrelationen mellem Bassin 7 - øst og vest.

Af figur 3.3 fremgår det, at der er en sammenhæng mellem ledningsevnen ved målestationerne. Der er især stor lighed mellem målestationerne i Bassin 7, dog med lidt større fluktuationer i ledningsevnen ved den østlige målestation, C. Dette skyldes, at målestation C er tættere på åen og dermed er mere ferskvandspåvirket. Fluktuationerne fremgår ligeledes ved Mellemarmen, hvor der er en stor variation i ledningsevnen.

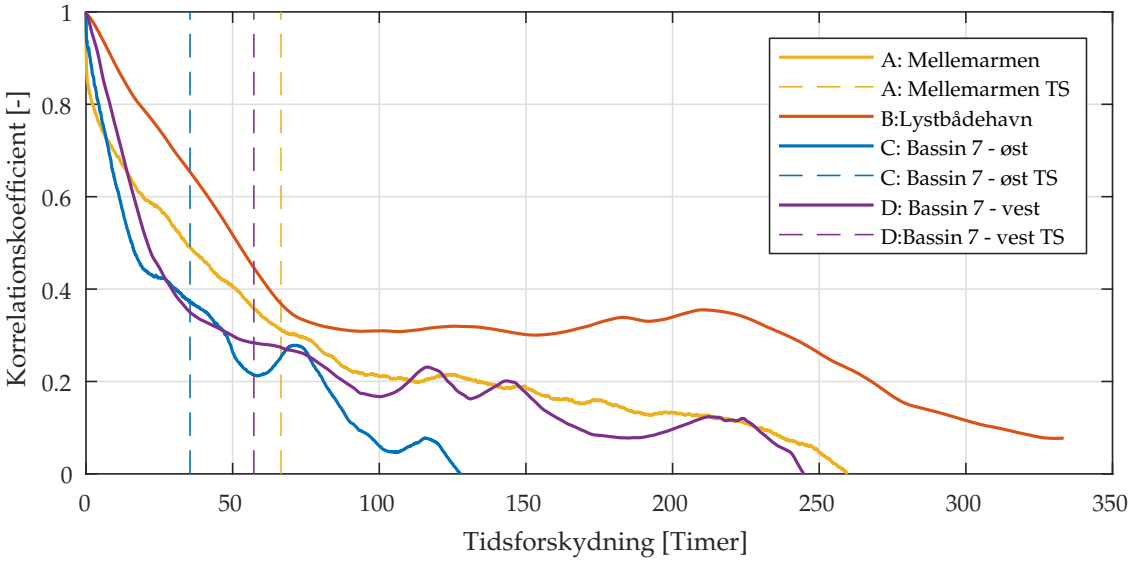
	Samplingfrekvens [h^{-1}]	**Middelværdi** [mS/cm]	**Standardafvigelse** [mS/cm]	**Median** [mS/cm]
A: Mellemarmen	12	28,55	6,07	29,89
B: Lystbådehavnen	12	30,24	5,38	31,54
C: Bassin 7 - Øst	12	28,85	3,45	28,99
D: Bassin 7 - Vest	3	28,73	4,76	28,99

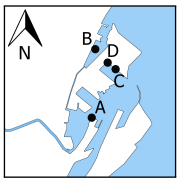
Tabel 3.1. Statistiske parametre for ledningsevnemålinger i Aarhus Havn.

Af tabel 3.1 fremgår det, at middelværdien for ledningsevnen er højest ved Lystbådehavnen. Dette indikerer, at ferskvandstilstrømningen til Lystbådehavnen er mindre end ved de resterende målestationer. Dette er i overenstemmelse med, at der i januar 2019 har været en længere periode uden fald i ledningsevnen. Der er lavet en Lillefors test på de fire datasæt, der viser, at ledningsevnen følger en normalfordeling ved et signifikansniveau på 5 %.

3.3 Undersøgelse af tidslig variation af ledningsevne

For at undersøge den tidslige variation af ledningsevnen opstilles et autokorrelogram for hver af de fire målestationer. Et autokorrelogram bruges til at undersøge vedvarenheden i en tidsserie. Er vedvarenheden i tidsserien stor, er værdierne korreleret med de forudgående målinger. Det forventes, at vedvarenheden for ledningsevnen er stor, da der flere gange er målt et fald i ledningsevnen, der varede flere dage. Er vedvarenheden derimod lille, må der forventes en større variation i ledningsevnen, og den kan derfor være vanskelig at benytte som parameter til varsling. Autokorrelogrammet beregnes ved at korrelere tidsserien med tidsserien tidsforskudt. Autokorrelogrammet for de fire målestationer fremgår af figur 3.4.





Figur 3.4. Autokorrelation for ledningsevnen ved målestation A, B, C og D. TS=tidsskala.

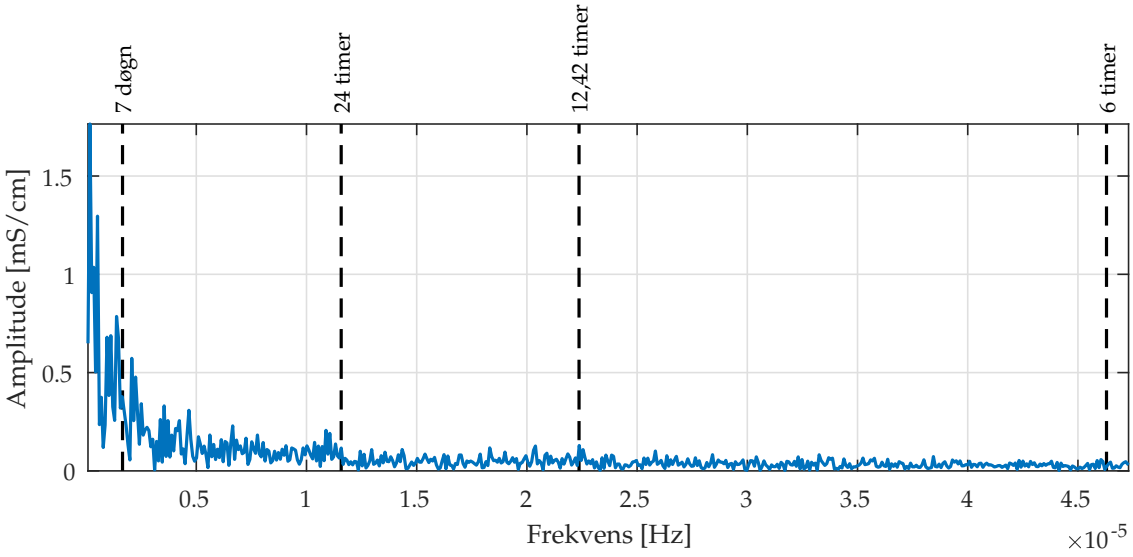
Autokorrelogrammet på figur 3.4 viser, at vedvarenheden og tidsskalaen for ledningsevnen er stor ved alle målestationer. Dette betyder, at målinger foretaget indenfor et interval, der er kortere end tidsskalaen, er korreleret med hinanden. Det er en fordel med en stor vedvarenhed, når ledningsevnen skal bruges til varsling, da samplingfrekvensen kan holdes lav, og at varslingstilstanden ikke hele tiden ændrer sig.

3.4 Fourier-analyse af svingninger i ledningsevnen

Af figur 3.3 fremgår det, at der specielt ved Mellemarmen er store svingninger i målingerne. Der kan være flere grunde til de store udsving i ledningsevnen ved Mellemarmen. En af årsagerne kan blandt andet være, at måleren er placeret i relativt åbent havområde, hvor der kan opstå kraftigere bølgegang. Andre årsager kan være, at der er en lagdeling i samme dybde, som sensoren er placeret, og der derfor skiftevis bliver målt på en ferskvandsstrøm i overfladen og derefter i lavereliggende havvand, eller at ferskvandsstrømmen fra Aarhus Å svinger horisontalt i udmundingen. Årsagen til svingningerne undersøges ved hjælp af en fourier-analyse, Der kan vise hvorvidt der indgår harmoniske svingninger i tidsserien for Mellemarmen.

Fourier-analysen udføres ved at lave en fourier-transformation af tidsserien. Resultatet af fourier-transformationen er en række fourier-koefficienter, der kan benyttes til at beskrive tidsserien ved trigonometriske funktioner. Da analysen kun benyttes til at bestemme harmoniske svingninger, er det kun interessant at bestemme tidsseriens frekvens-spektrum. Frekvens-spektrummet består af svingningernes amplitude som funktion af frekvens.

Frekvens-spektrummet for ledningsevnen ved Mellemarmen fremgår af figur 3.5.

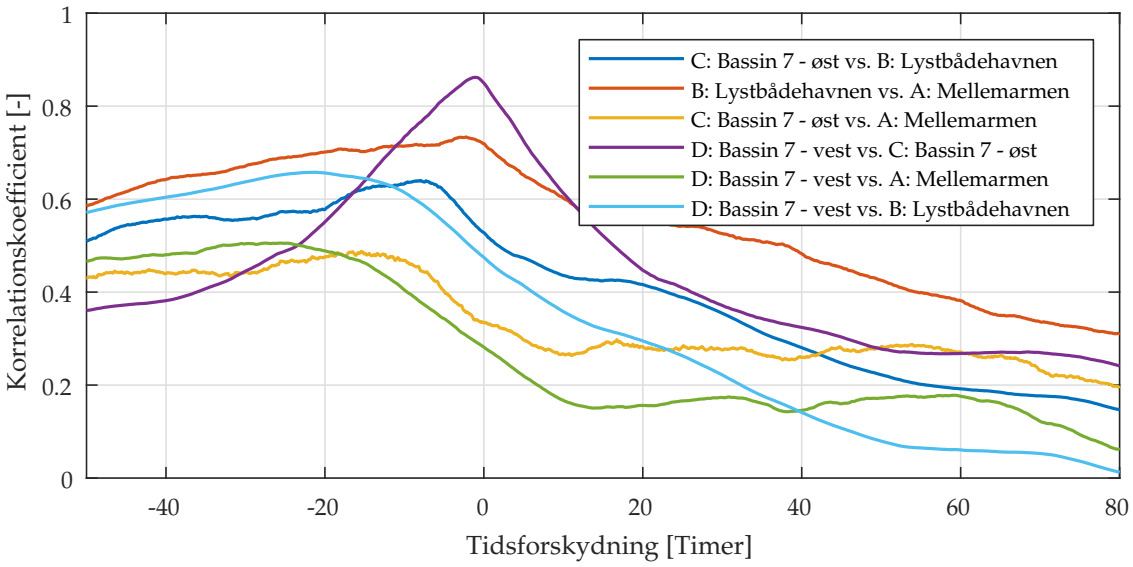


Figur 3.5. Frekvensspektrum for ledningsevnen ved Mellemarmen, bestemt ved foruier-analyse.

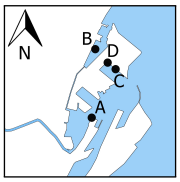
Af figur 3.5 fremgår det, at det eneste peak, der er i amplitude, er ved en meget lav frekvens svarende til en periode på tidsseriens længde. Det vurderes derfor, at der ikke er nogle harmoniske svingninger i tidsserien, hvorfor det kan konkluderes, at blandt andet tidevand ikke har en direkte indvirkning på ledningsevnen. Samme undersøgelse er lavet for de resterende datasæt med samme udfald.

3.5 Undersøgelse af stedslig variation af ledningsevne

Den stedslige variation af ledningsevnen indenfor havneområdet er interessent, da den kan give et overblik over de horisontale strømninger i havnen, og dermed hvordan en forurening vil sprede sig. Det forventes, at ledningsevnen for de to stationer i Bassin 7 er meget ens, hvorimod det er mere usikkert, om de andre målestationer er ens. For at undersøge hvordan ledningsevnen i de fire målepunkter varierer i forhold til hinanden, analyseres tidsserierne for kryds-korrelation. Kryds-korrelogrammerne for de 6 kombinationer af tidsserier fremgår af figur 3.6.



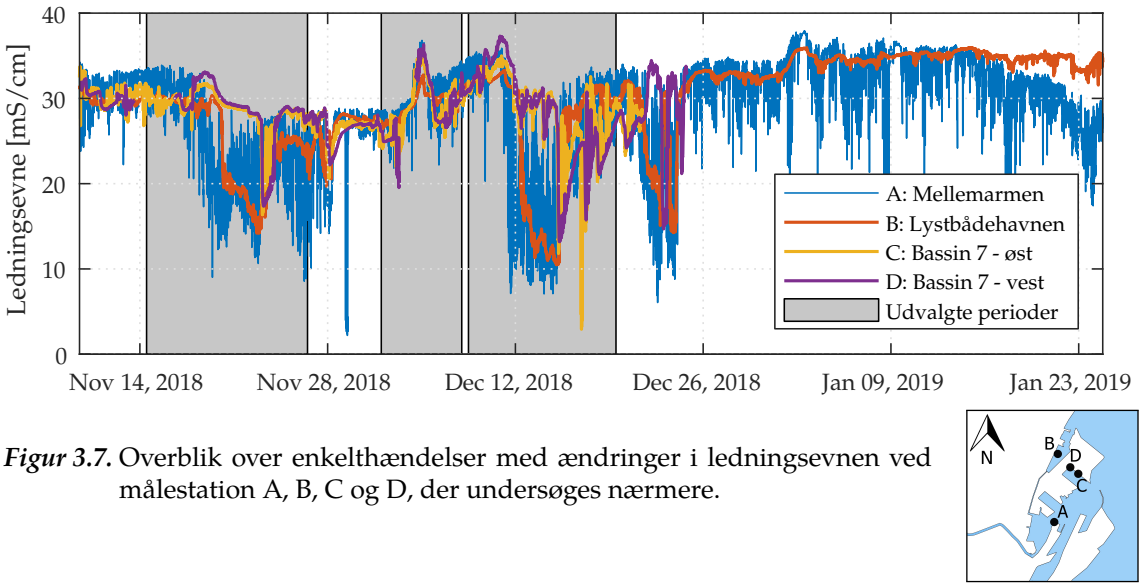
Figur 3.6. Krydskorrelation mellem ledningsevne ved målestation A, B, C og D.



Det fremgår af figur 3.6, at der er en høj korrelation mellem de to målere i Bassin 7 samt mellem Lystbådehavnen og Mellemarmen. Korrelationen mellem de to målere i Bassin 7 er som forventet høj, da afstanden mellem de to målere er kort. Korrelationen mellem Lystbådehavnen og Mellemarmen er ikke lige så forventet som korrelationen i Bassin 7, da de to målere er placeret i to forskellige dele af havnen. Korrelationen giver en indikation om, at ferskvandsstrømningerne fra Aarhus Å splittes op og strømmer henholdsvis ud forbi Mellemarmen og gennem kanalen til Lystbådehavnen. Af figur 3.3 fremgår det, at tendenserne i ledningsevnen ved de to målestationer er meget ens, både i forhold til hvornår der er lav ledningsevne og det generelle niveau. Dette kan blandt andet ses ved, at der sker nogle kraftige fald i ledningsevnen samtidig og med samme varighed. Det fremgår også, at den store forskel på de to målestationer er de kraftige svingninger, der er i ledningsevnen ved Mellemarmen, hvorimod ledningsevnen i Lystbådehavnen er mere stabil.

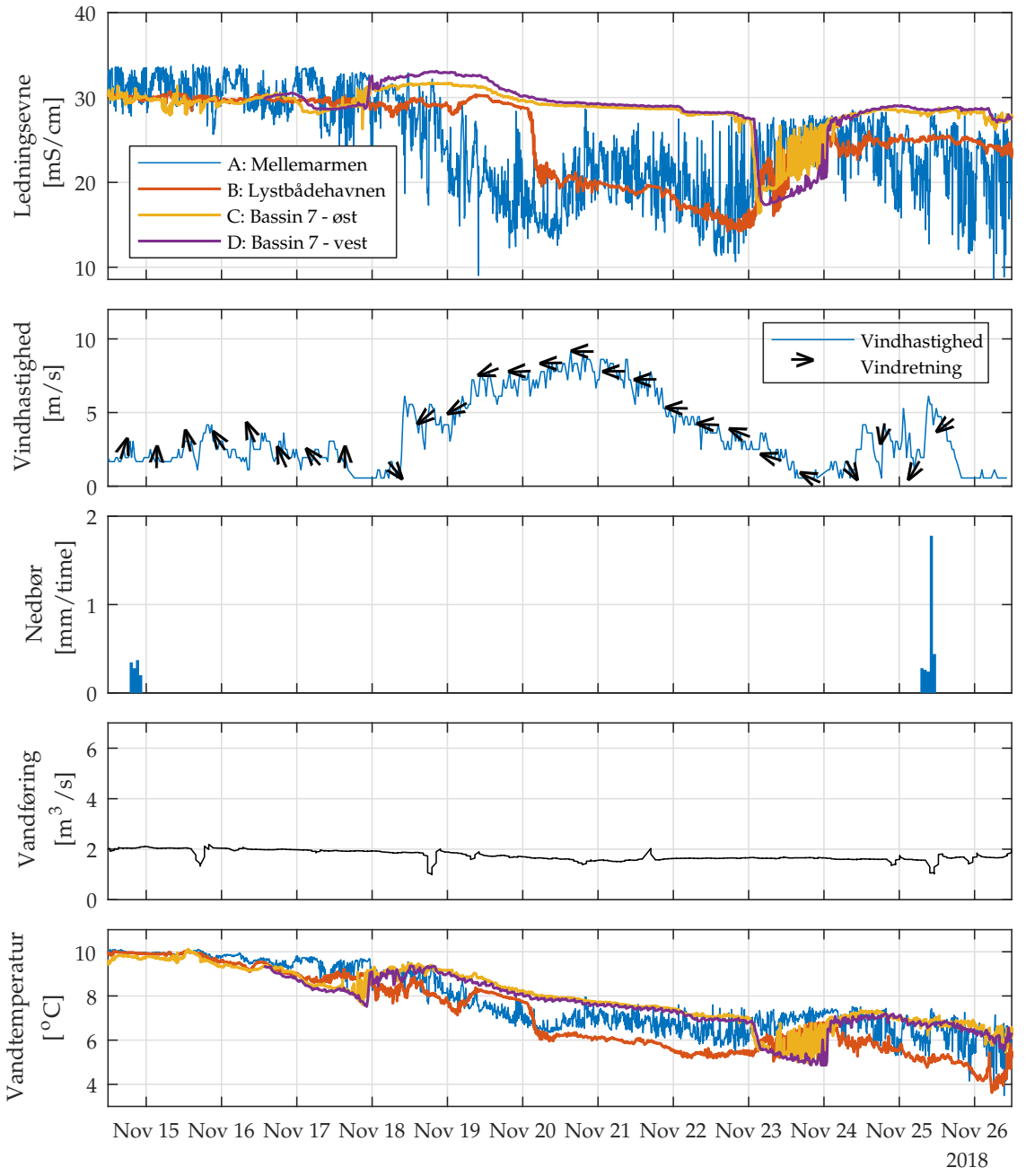
3.6 Påvirkning af ledningsevnen ved enkelthændelser

For at undersøge hvordan blandt andet vinden påvirker ledningsevnen undersøges en række enkelthændelser, hvor ledningsevnen enten har været høj eller lav. Der er udvalgt tre perioder i november og december 2018, der undersøges nærmere. Disse perioder er markeret med grå på figur 3.7.



Figur 3.7. Overblik over enkelthændelser med ændringer i ledningsevnen ved målestation A, B, C og D, der undersøges nærmere.

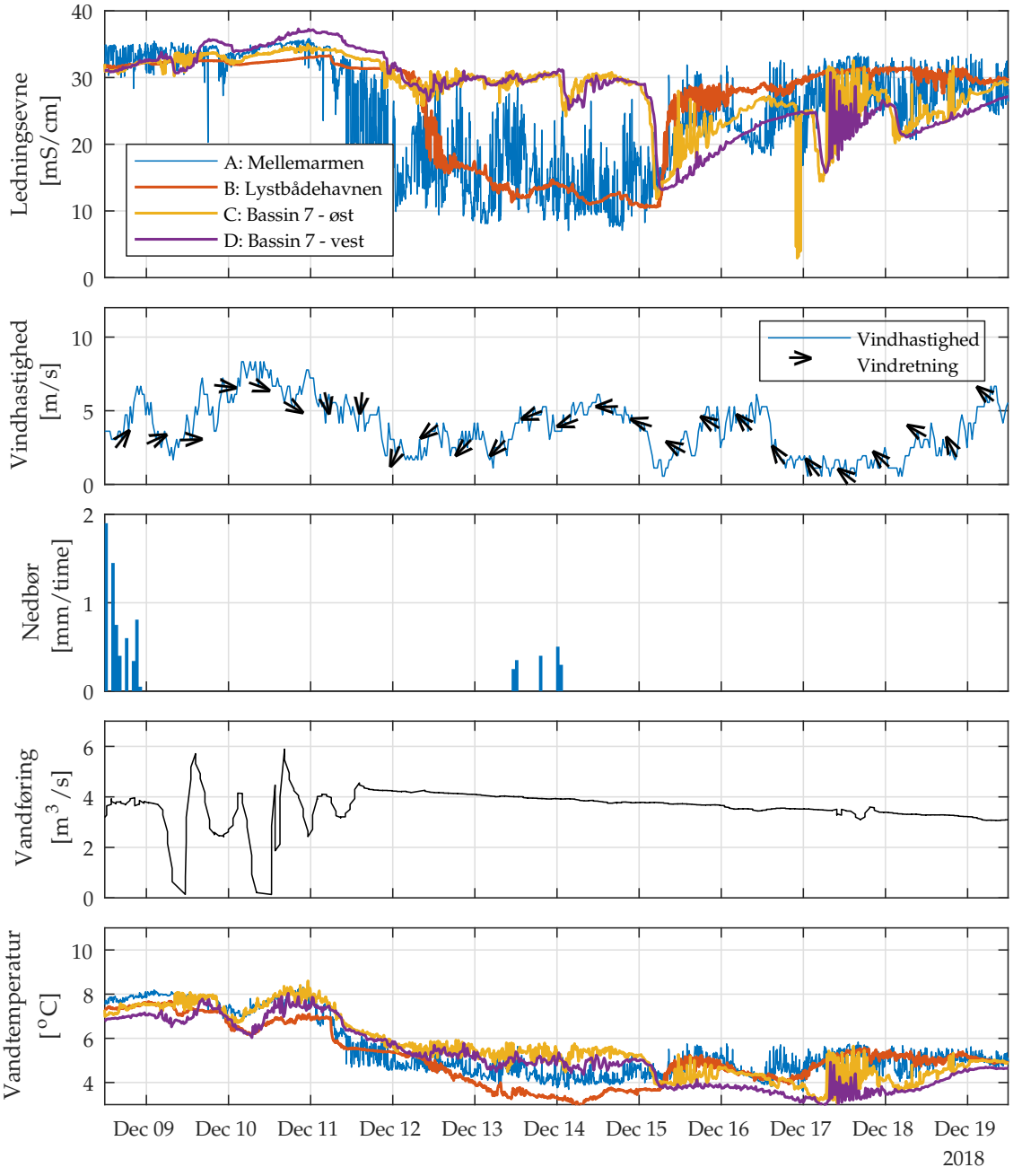
I det følgende undersøges, hvad årsagen til ændringerne i ledningsevnen i de tre perioder er. Dette gøres ved at plotte ledningsevnedata med vindretning og -hastighed, nedbør, vandføring samt vandtemperaturen.



Figur 3.8. Ledningsevnen målt ved målestation A, B, C og D sammenholdt med vinddata, nedbør, vandføring i Aarhus Å samt vandtemperatur for perioden 14. november 2018 til 26. november 2018. Vinddata, nedbørsdata og vandføringsdata er fra henholdsvis The Weather Company [2018], Spildevandskomiteens regnmålersystem [2019] og Hydrometri [2018].

Af figur 3.8 fremgår det, at hverken vandføring fra Aarhus Å eller nedbør kan være årsag til det fald, der er i ledningsevnen ved de fire målestationer. Det fremgår samtidig, at der sker et skift i vindretning fra syd til øst, når ledningsevnen falder. Dette indikerer, at vinden påvirker ledningsevnen. Derudover fremgår det, at temperaturen falder samtidigt med ledningsevnen. Dette kan skyldes, at ferskvandet der også påvirker ledningsevnen har en lavere temperatur end havvandet. Det undersøges, om tendensen også er gældende

for det fald i ledningsevne, der har været i perioden fra 8. til 19. december 2018.

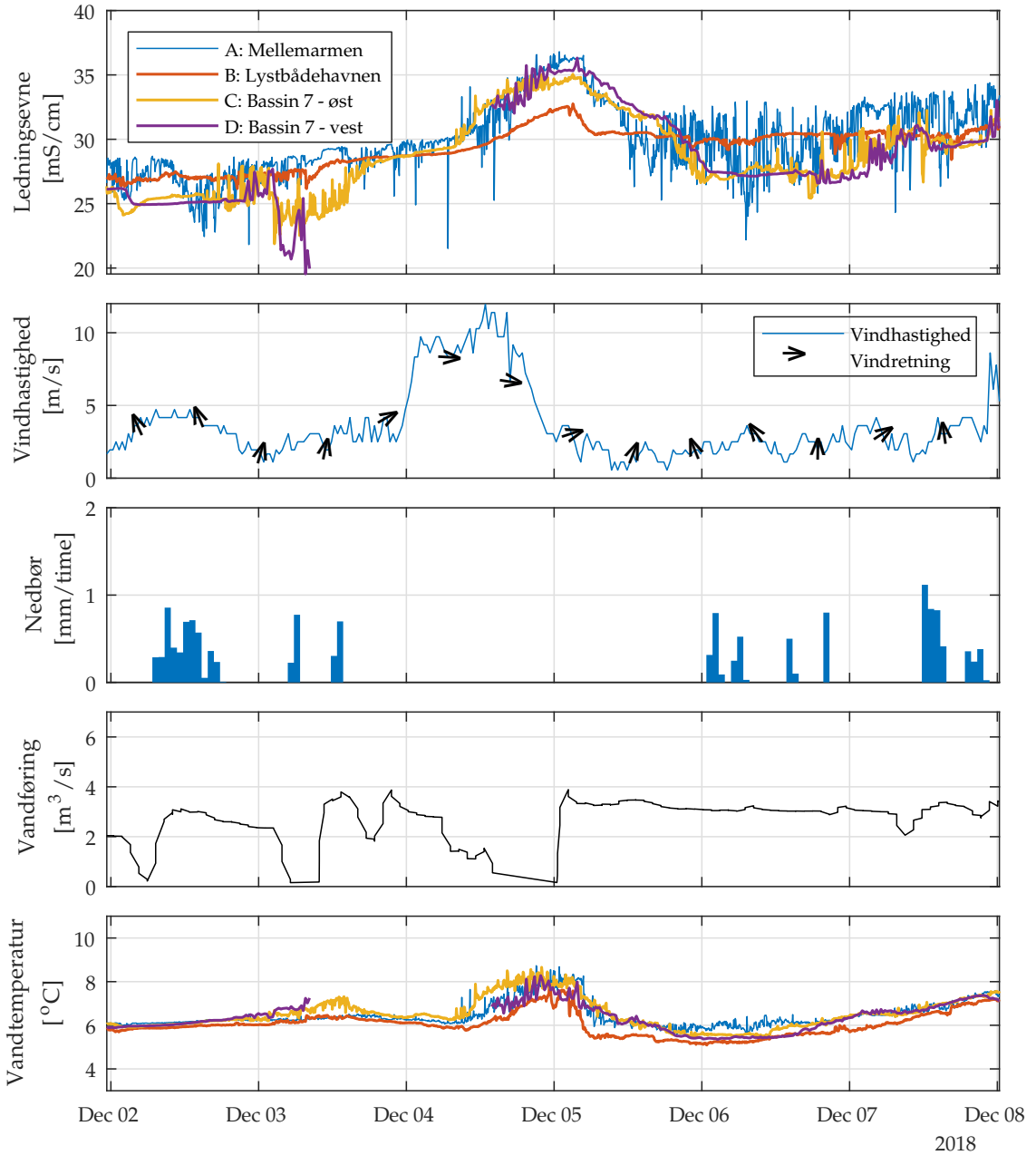


Figur 3.9. Ledningsevnen målt ved målestation A, B, C og D sammenholdt med vinddata, nedbør, vandføring i Aarhus Å samt vandtemperatur for perioden 8. december 2018 til 19. december 2018. Vinddata, nedbørsdata og vandføringsdata er fra henholdsvis The Weather Company [2018], Spildevandskomiteens regnmålersystem [2019] og Hydrometri [2018].

Af figur 3.9 fremgår det, at ledningsevnen ved Mellemarmen og Lystbådehavnen falder, idet vinden skifter om til nordøstenvind. Det fremgår ligeledes, at ledningsevnen ved Bassin 7 falder, idet vinden skifter til at komme fra sydøstlig retning. Det fremgår desuden, at der er nogle udsving i vandføring i dagene op til faldet i ledningsevne. Disse udsving

kan forklares af en periode med høj vandstand, hvor tidevandet har givet anledning til tilbagestuvning i Aarhus Å ved højvande og efterfølgende øget vandføring ved lavvande.

Hændelsen den 5. december 2018, hvor ledningsevnen er steget, undersøges for at vurdere, om vinden også har den modsatte effekt af, hvad der blev vist på figur 3.8 og 3.9. Denne hændelse fremgår af 3.10.



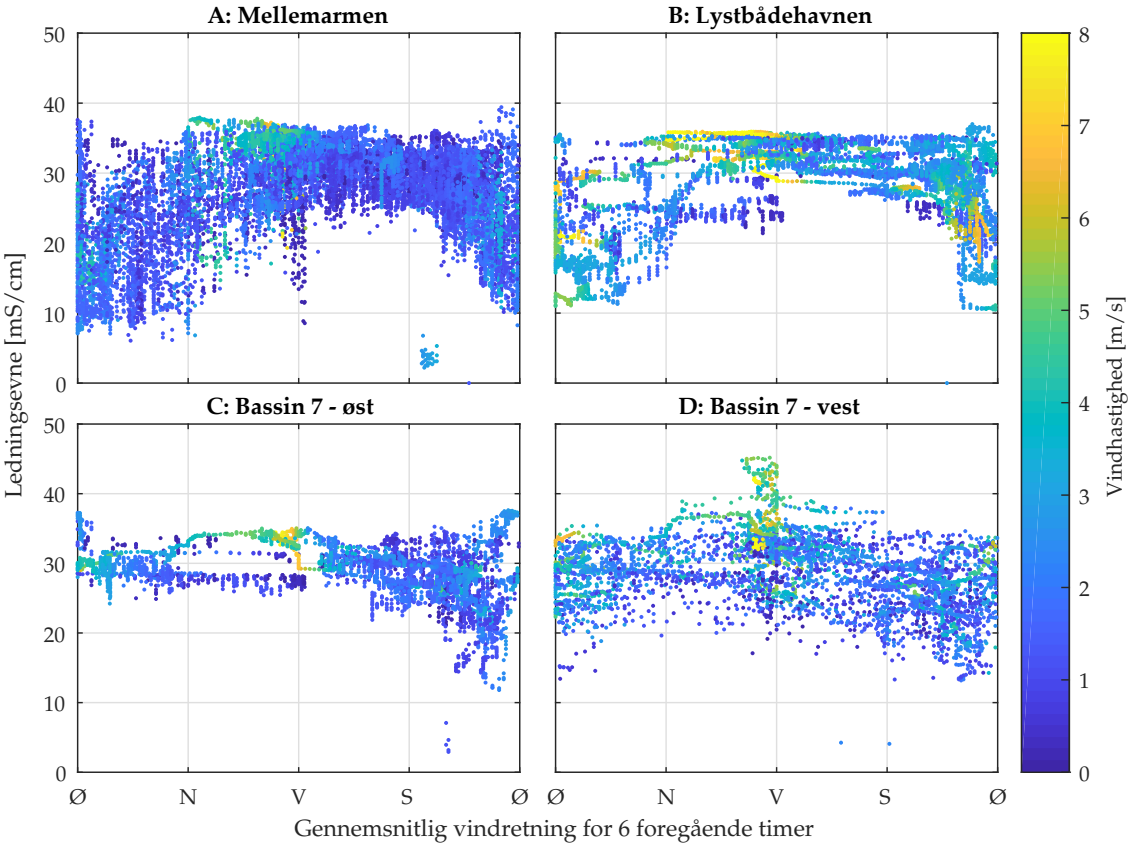
Figur 3.10. Ledningsevnen målt ved de målestation A, B, C og D sammenholdt med vinddata, nedbør, vandføring i Aarhus Å samt vandtemperatur for perioden 2. december 2018 til 8. december 2018. Vinddata, nedbørsdata og vandføringsdata er fra henholdsvis The Weather Company [2018], Spildevandskomiteens regnmålersystem [2019] og Hydrometri [2018].

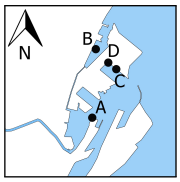
Af figur 3.10 fremgår det, at kraftig vind fra vest giver en stigning i ledningsevne ved alle 4 målestationer. Dette kan skyldes, at den kraftige vind giver anledning til en stor opblanding mellem ferskvand og havvand. Det kan også skyldes, at vinden flytter ferskvandet i overfladen ud af havnen.

På baggrund af de viste tendenser foretages der i det følgende en mere generel analyse af ledningsevnen og vindpåvirkningen.

3.7 Sammenhæng mellem vind og ledningsevne

For at undersøge hvorvidt der er en generel tendens til, at ledningsevnen afhænger af vindretningen, sammenlignes ledningsevnen, vindretning og -hastighed for hele tidsserien. Dette fremgår af figur 3.11.





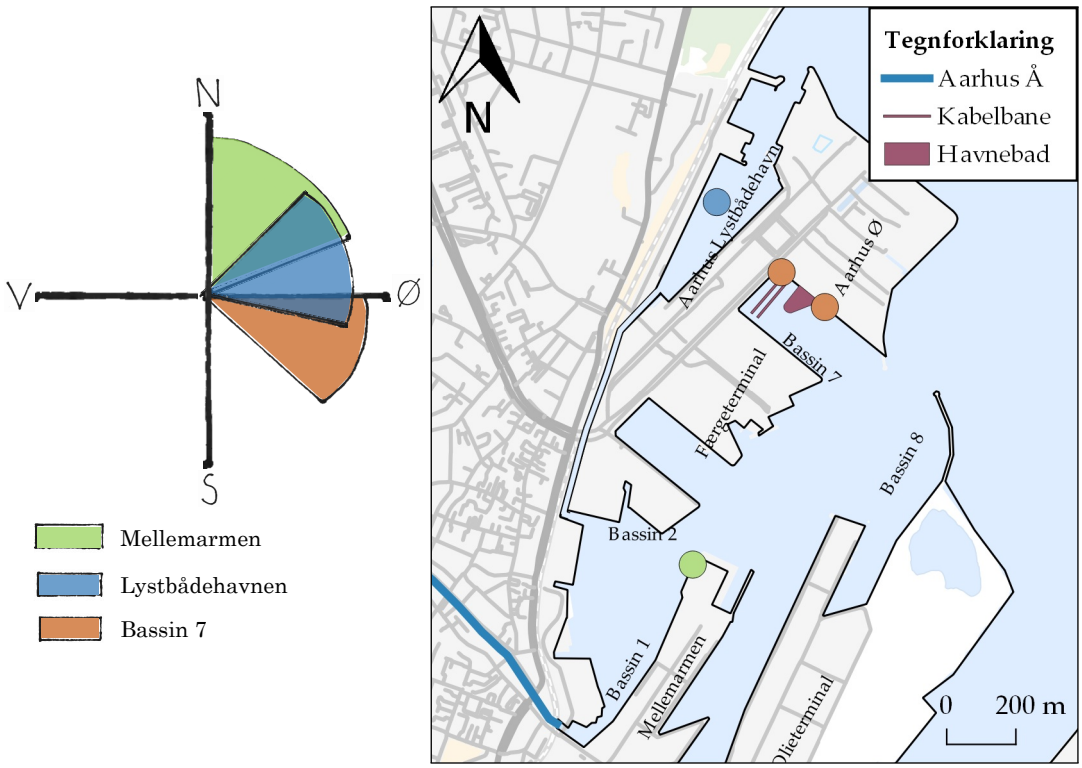
Figur 3.11. Sammenhæng mellem ledningsevne, vindretning og -hastighed ved målestation A, B, C og D. Vinddata er midlet over 6 timer og stammer fra The Weather Company [2018].

Af figur 3.11 fremgår det, at der er en sammenhæng mellem vindretning og ledningsevne. Ved Bassin 7 - vest er det tydeligst at ledningsevnen stiger ved vestenvind og mindre tydeligt, at den falder ved østenvind. For de resterende målestationer fremgår det, at vind fra sydøst, nordøst og øst kan give lave ledningsevner i henholdsvis Bassin 7 - øst, Mellemarmen og Lystbådehavnen. Dette stemmer overens med resultaterne fra

undersøgelsen af de tre perioder i november og december 2018.

Årsagen til, at disse vindretninger giver lave ledningsevner, findes i den fysiske udformning af havnen. Vind fra sydøst medvirker til, at det lette ferskvand, der ligger i overfladen, bliver ledt ind i Bassin 7, da bassinets åbning vender i denne retning. På samme måde medvirker vind fra øst, at det ferskvand, der ligger i overfladen i Bassin 1, bliver ledt ind igennem Bassin 2 og den nye kanal ind til Lystbådehavnen. De lave ledningsevner er altså formentlig et resultat af en lagdeling mellem det lette ferskvand og tunge saltvand, der herefter bliver flyttet af vinden.

Af figur 3.12 fremgår en konceptuel skitse, der viser vindretninger, der medvirker til en lav ledningsevne i Bassin 7, Lystbådehavnen og ved Mellemarmen.



Figur 3.12. Konceptuel skitse af vindretninger der resulterer i en lav ledningsevne for områderne Mellemarmen, Lystbådehavnen og Bassin 7.

Overordnet set konkluderes det, ud fra undersøgelserne af sammenhængen mellem vindpåvirkning og ledningsevne, at vindretningen har stor indflydelse på, hvor det ferske vand fra Aarhus Å bliver ledt hen. Hvilke vindretninger, der medfører en lav ledningsevne, har en tydelig sammenhæng med havnens fysiske udformning.

I kapitel 2.2 er det vist, at overskridelse af badevandskvaliteten i Bassin 7 kan korreleres med vind fra sydlige retninger. Dette stemmer ikke fuldstændig overens med, hvornår der opstår lave ledningsevner i Bassin 7. Dette kan skyldes, at der for undersøgelsen af korrelationen mellem ledningsevne og vindpåvirkning kun er benyttet data fra en mindre periode i løbet af efteråret og vinteren 2018.

3.8 Opsummering

I dette kapitel er variationerne i ledningsevne og årsagerne hertil undersøgt. Det konkluderes, at der ikke er nogen særlige variationer indenfor Bassin 7, mens variationerne mellem Bassin 7, Lystbådehavnen og Mellemarmen er tydeligere. Årsagerne til disse variationer vurderes at skyldes vindens påvirkning på strømningsmønsteret for det ferskvand, der ledes ud fra Aarhus Å.

Da ledningsevnen kun falder kraftigt ved bestemte vindretninger, er det højst sandsynligt, at det er tilstrømning af forurenet åvand, der forårsager faldet i ledningsevne. Ledningsevnen falder altså ikke på grund af andre ferskvandsudledninger som eksempelvis regnvandsudløb, der ikke er forurenet med spildevand, og dermed ikke udgør en risiko for overskridelse af badevandskriteret. Dette understøtter teorien om, at en lav ledningsevne vil medføre dårlig badevandskvalitet.

Kapitel 4

MODELLERING AF BADEVANDSKVALITET PÅ BAGGRUND AF LEDNINGSEVNE

I dette kapitel opbygges og evalueres en model til varsling af badevandskvalitet på baggrund af den elektriske ledningsevne.

I kapitel 2 er sammenhængen mellem ledningsevnen og koncentrationen af fækale bakterier undersøgt. Her blev det vurderet, at en eksponentiel regressionsmodel kan benyttes til at beskrive sammenhængen mellem ledningsevnen og henholdsvis Enterokokker og E. Coli.

4.1 Eksponentiel model til vurdering af badevandskvaliteten

Sammenhængen mellem bakteriekoncentration og ledningsevne kan beskrives med den eksponentielle model, der fremgår af formel 4.1.

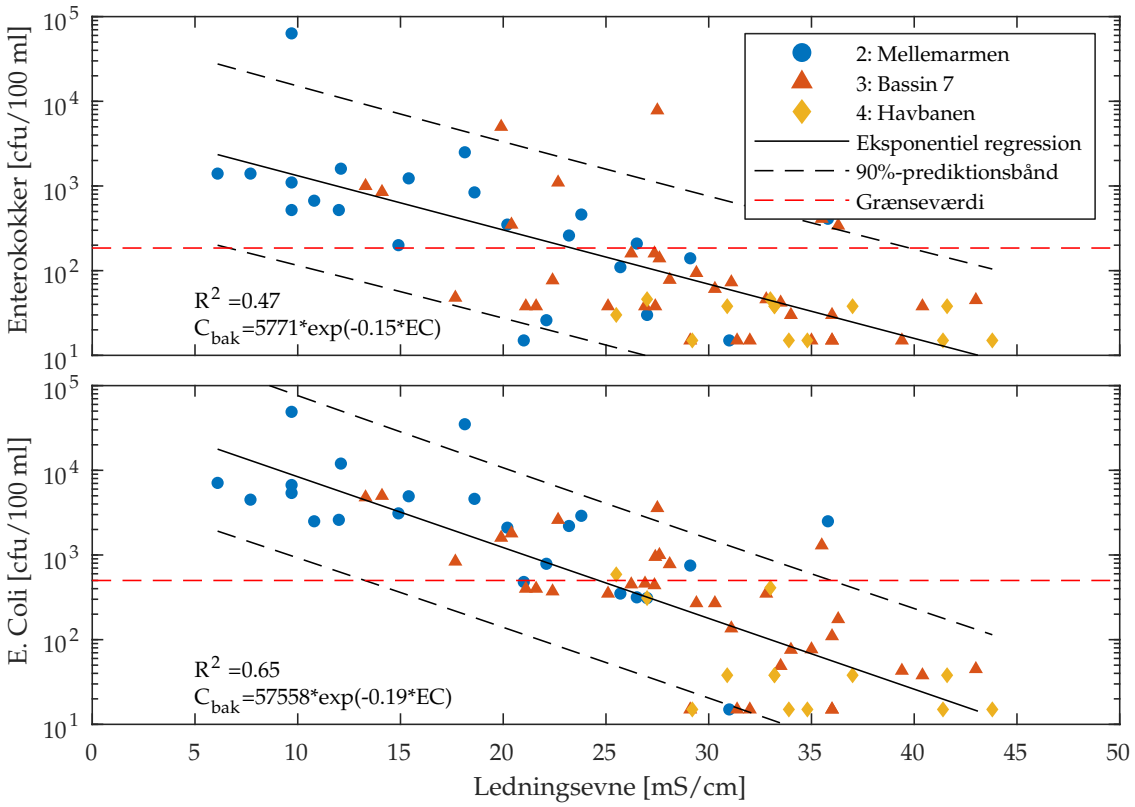
$$C_{bak} = \alpha \cdot \exp(\beta \cdot EC) \tag{4.1}$$

Hvor:

C_{bak}	Bakteriekoncentration [cfu/100ml]
α, β	Modelparametre [-]
EC	Elektrisk ledningsevne [$\frac{\text{mS}}{\text{cm}}$]

Denne model er særligt egnet til at beskrive bakteriekoncentrationens sammenhæng med ledningsevnen, da bakteriekoncentration falder eksponentielt med saliniteten og dermed med ledningsevnen. Dette skyldes, at intestinale bakterier ikke trives i saltvand, men derimod dør hurtigere [Greenberg, 1956].

Den eksponentielle model er tilpasset til koncentration af henholdsvis Enterokokker og E. Coli, og resultatet fremgår af figur 4.1. Modellen beskriver sammenhængen mellem ledningsevnen og bakteriekoncentrationen for enterokokker og E. Coli ved målepunkt 2, 3 og 4. Målepunkt 1 er ikke medtaget, da vandet her altid er ferskt. De to R^2-værdier viser, at der er en god korrelation mellem modellen og de målte værdier for ledningsevne og bakteriekoncentration.



Figur 4.1. Sammenhæng mellem elektrisk ledningsevne og bakteriekoncentrationer ved målepunkt 2, 3 og 4. R^2-værdi angiver korrelationen mellem de målte værdier og den eksponentielle regression. Data fra Aarhus Kommune [2018a].

Modellen på figur 4.1 kan benyttes til at vurdere bakteriekoncentrationen ud fra ledningsevnen. Prædiktionsbåndet beskriver desuden, at der er en spredning på målingerne i forhold til modellen. Dette kan skyldes, at der er svingende bakteriekoncentration i ferskvandet i Aarhus Å, hvilket medfører, at der er en svingende bakteriekoncentration i forhold til ledningsevnen.

Den eksponentielle model kan benyttes til at estimere en numerisk værdi for bakteriekoncentrationen. For at vurdere badevandskvaliteten kræves derimod en grænseværdi for at klassificere badevandskvaliteten som værende tilfredsstillende eller dårlig. Denne grænseværdi kan dog bestemmes ud fra modellen og er i det følgende benævnt kritisk ledningsevne.

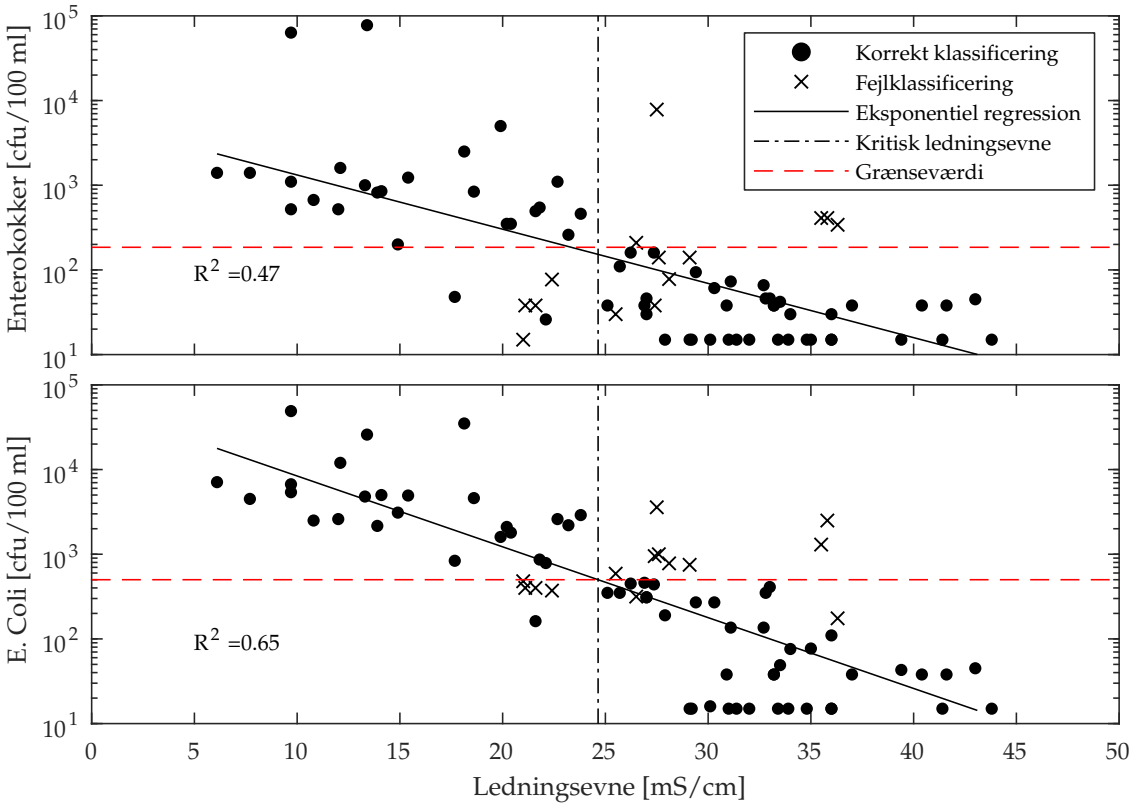
Til at undersøge modellens performance bestemmes skæringspunktet mellem badevandskriteret og modellen som værende den kritiske ledningsevne. Denne ledningsevne benyttes herefter til at modellere, hvorvidt badevandskvaliteten er tilstrækkelig. Er den målte ledningsevne under den kritiske, vil modellens resultat være dårlig badevandskvalitet og omvendt. Modellens respons sammenholdes herefter med den målte badevandskvalitet. Ud fra grænseværdierne for badevandskvalitet kan observationerne omsættes til positiv respons, hvis den overskrider grænseværdierne og negativ respons, hvis den ikke overskrider grænseværdierne. Da der for hver måling er analyseret for både enterokokker og

E. Coli klassifiseres prøven som positiv, hvis en af analyserne viser en overskridelse af den pågældende grænseværdi. Tabel 4.1 viser princippet bag klassificeringen af målingerne.

Enterokokker [cfu/100ml]	**E. Coli** [cfu/100ml]	**Klassifisering** [−]
850	5000	Positiv
26	787	Positiv
341	175	Positiv
30	76	Negativ

Tabel 4.1. Eksempler på klassificering af bakteriemålinger.

Af figur 4.2 fremgår korrekte og forkerte resultater af modellen, hvor den kritiske ledningsevne er benyttet til at bestemme kvaliteten. På figuren angiver krydser de målinger, hvor modellen har fejlklassificeret i forhold til den målte badevandskvalitet. Som kritisk ledningsevne er det valgt at benytte den højeste værdi af skæringspunktet for henholdsvis modellen for enterokokker og modellen for E. Coli.



Figur 4.2. Eksponentiel model til beskrivelse af sammenhængen mellem ledningsevne og badevandskvalitet. Krydser angiver forkerte modelresultater ved en grænseværdi for ledningsevne svarende til modellens skæring med grænseværdien for E. Coli.

Som mål for modellens performance benyttes andelen af punkter, hvor modelresultatet stemmer overens med den målte badevandskvalitet, herefter benævnt hitrate. For at nuancere modellens performance er den opdelt i positive observationer og negative observationer. Yderligere information omkring hitrate fremgår af appendiks G.

Af tabel 4.2 fremgår model-performance ved en kritisk ledningsevne valgt ud fra den eksponentielle models skæring med grænseværdien.

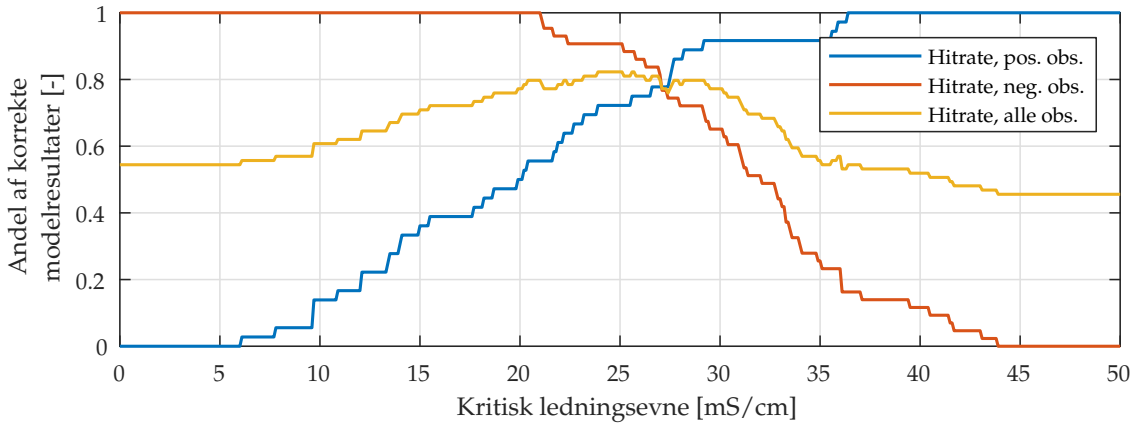
Model	**Kritisk ledningsevne** [mS/cm]	**Hitrate positive obs.** [%]	**Hitrate negative obs.** [%]	**Hitrate alle obs.** [%]
Eksponentiel	24,6	72	91	82

Tabel 4.2. Model-performance for eksponentiel regressionsmodel for ledningsevne og bakteriekoncentration.

Som det fremgår af tabel 4.2, kan 82 % af observationerne klassificeres korrekt ved valg af kritisk ledningsevne ud fra den eksponentielle models skæringspunkt med grænseværdien for bakteriekoncentration. Det fremgår dog også, at det kun er 72 % af de observationer med dårlig badevandskvalitet, der bliver klassificeret korrekt, hvilket muligvis kan optimeres ved at vælge en anden kritisk ledningsevne.

4.2 Optimering af kritisk ledningevne

Hitraterne, der fremgår af tabel 4.2, kan beregnes for alle ledningsevner og analyseres for et optimum. Hitraterne ved alle kritiske ledningsevner fremgår af figur 4.3.



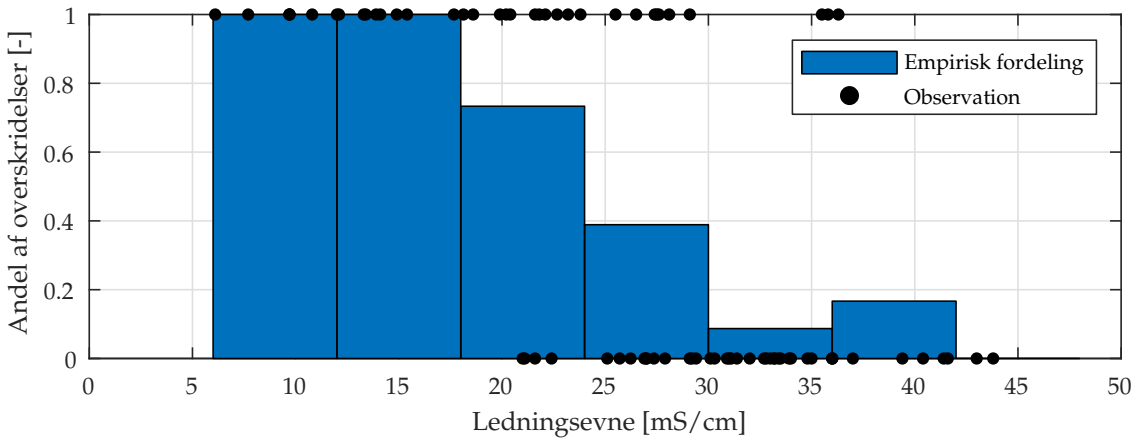
Figur 4.3. Hitrate for kritisk ledningsevne.

Af figur 4.3 fremgår det, at den samlede performance har et toppunkt ved en kritisk ledningsevne mellem 20 og 30 mS/cm. Den samlede performance er altså lige stor, hvis der vælges en kritisk ledningsevne på 20 mS/cm som, hvis der vælges 30 mS/cm. Dog fremgår det, at der ved at vælge 20 mS/cm, kun opnås en hitrate på ca. 55 % for de positive observationer mod ca. 90 %, hvis der vælges en kritisk ledningsevne på 30 mS/cm. Valget af den høje ledningsevne vil dog medføre, at badeområdet er lukket i en større andel af tiden. Dermed er fastsættelse af den kritiske ledningsevne en opvejning mellem at vælge en høj værdi, så sundhedsrisikoen er lav og vælge en lav værdi, så lukketiden bliver mindst.

4.3 Sandsynlighedsmodel for badevandskvalitet

Den eksponentielle regressionsmodel, der er præsenteret i afsnit 4.1, estimerer bakteriekoncentrationen ud fra ledningsevnen. Sammenhængen kan benyttes til at fastsætte en kritisk ledningsevne, der benyttes til at vurdere badevandskvaliteten. I dette afsnit opbygges en model, der estimerer sandsynligheden for overskridelse af badevandskriteriet ud fra ledningsevnen.

Til at modellere sandsynligheden for overskridelse af badevandskriteriet benyttes de binære responser fra observationer af badevandskvaliteten, der er beskrevet i tabel 4.1. Der benyttes observationer fra målepunkt 2, 3 og 4, da badevandskriteriet stort set altid er overskredet ved målepunkt 1. Først opstilles binomial-fordelingen for sammenhængen mellem ledningsevne og badevandskvalitet. Dette gøres ved at inddele ledningsevnen i lige store intervaller og undersøge, hvor stor en andel af målingerne med det givne interval, der overskrider badevandskriteriet. Den empiriske binomial-fordeling fremgår af 4.4.



Figur 4.4. Empirisk binomial-fordeling af badevandskvaliteten afhængig af ledningsevnen.

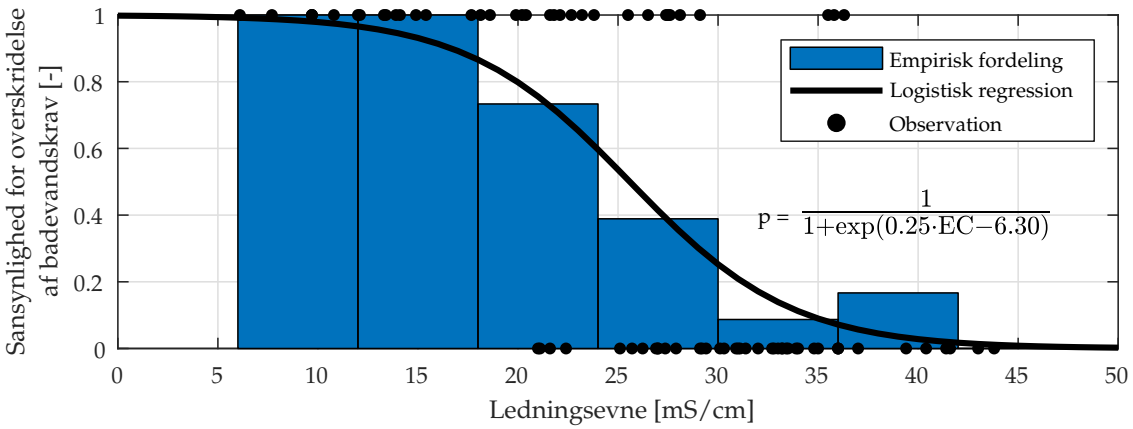
Som det fremgår af 4.4 er der et tydeligt fald i andelen af overskridelser, des højere ledningsevnen er. For at kunne vurdere sandsynligheden for overskridelse af badevandskriteriet, ud fra ledningsevnen, kræves en model for den empiriske binomial-fordeling. Til at opbygge denne model benyttes logistisk regression. Metoden til at opbygge den logistiske regressionsmodel er beskrevet i appendiks E. Den logisitske model fremgår af formel (4.2) [Ngufor og Wojtusiak, 2016].

$$p = \frac{1}{1 + \exp(-a \cdot EC - b)} \tag{4.2}$$

Hvor:

p	Sandsynlighed for overskridelse af badevandskriterie [-]
a, b	Modelparametre [-]
EC	Elektisk ledningsevne [mS/cm]

Den logistiske regressionsmodel for sandsynligheden for overskridelse af badevandskvalitetskravet fremgår af figur 4.5.

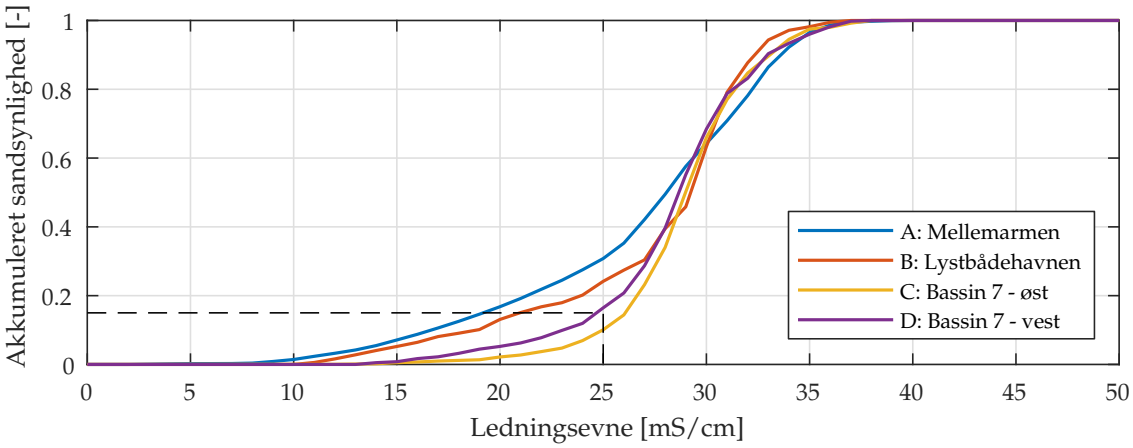


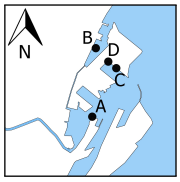
Figur 4.5. Logistisk regression for sammenhængen mellem ledningsevnen og sandsynligheden for overskridelse af badevandskvalitetskravet.

Som det fremgår af figur 4.5 er der en overenstemmelse mellem den logistiske regression og den empiriske fordeling af datasættet. Der er dog en afvigelse i modellen, der eksempelvis kommer til syne ved, at der ved en ledningsevne under 21 mS/cm ikke er observeret tilfredsstillende badevandskvalitet, men alligevel viser modellen, at der er 25 % sandsynlighed for, at der er tilfredsstillende badevandskvalitet.

4.4 Kritisk ledningsevnes betydning for lukketid

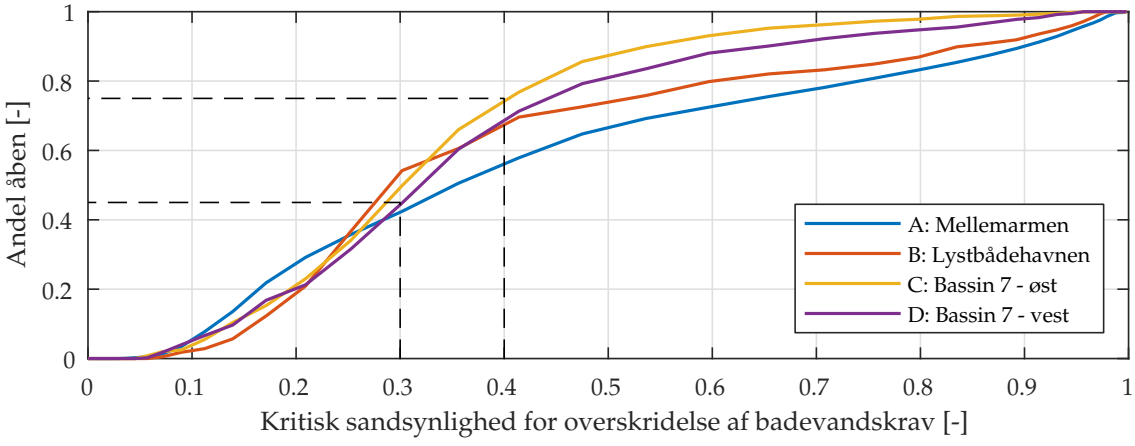
Når varslingsmodellen baseres på en simpel model, som en kritisk ledningsevne til afskæring mellem tilfredsstillende og dårlig badevandskvalitet, er det klart, at denne værdi er meget vigtig for modellens performance. Det giver derfor anledning til at undersøge hvilken værdi, der er optimal. Der er flere aspekter i, hvad en optimal afskæringsværdi er blandt andet sandsynlighed for overskridelse af kravet holdt op i mod badeområdets lukketid. Til at undersøge badeområdets lukketid benyttes den akkumulerede sandsynlighed for ledningsevnen. Denne fremgår af figur 4.6 og er lavet på baggrund af tidsserier for de fire målestationer. Figuren viser sandsynligheden for, at ledningsevnen af en vilkårlig observation er under en vis værdi. Dette kan benyttes til at vurdere, hvor længe badeområdet er lukket ved bestemte kritiske ledningsevner. Eksempelvis vil en grænseværdi for ledningsevnen på 25 mS/cm ved Bassin 7 svare til, at badeområdet vil være lukket i ca. 15 % af tiden, da dette svarer til sandsynligheden, for at ledningsevnen vil være under denne værdi.

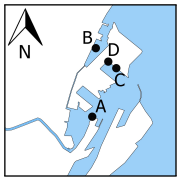




Figur 4.6. Akkumuleret sandsynlighed for ledningsevnen ved målestation A, B, C og D.

I afsnit 4.3 er der opstillet en model, der beskriver sandsynligheden for overskridelse af badevandskvaliteten ud fra ledningsevnen. Dette kan kombineres med den akkumulerede sandsynlighedskurve for ledningsevnen i Aarhus Havn for at estimere, hvor stor en andel af tiden badeområdet skal holde lukket ved en given sandsynlighed for overskridelse af badevandskvalitet. Dette giver mulighed for at fastsætte en grænseværdi ud fra en kritisk sandsynlighed for overskridelse i stedet for kritisk ledningsevne. Af figur 4.7 fremgår sammenhængen mellem den kritiske sandsynlighed for overskridelse af badevandskritere og andel af tiden, hvor badeområdet er åbent. Figuren viser eksempelvis, at hvis badeområdet lukkes, når sandsynligheden for overskridelse af badevandskriteriet er over 30 %, vil Bassin 7 - vest være åbent i ca. 45 % af tiden.





Figur 4.7. Sammenhæng mellem den kritiske sandsynlighed for overskridelse af badevandskritere og andel af tiden hvor badeområdet er åbent.

Figur 4.7 kan benyttes som værktøj til at fastsætte en grænseværdi for, hvornår

badeområdet bør lukkes. Den viser, at man, for at mindske risikoen for dårlig badevandskvalitet, er nødt til at gå på kompromis med, hvor lang tid badeområderne er åbent. For Bassin 7 - øst flader kurven ud ved ca. 75 % andel åben. Det vil derfor være fordelagtigt at lukke badeområdet, når der er 40 % sandsynlighed for dårlig badevandskvalitet, hvilket svarer til 27,4 mS/cm, da fordelen ved at sætte den højere ikke giver så stort afkast på andelen af tid, hvor bassin er åbent. Modsat falder kurven kraftigere, hvis risikoen for overskridelse sættes lavere. Samme tendens kan ses for Bassin 7 - øst og Lystbådehavnen, dog forskudt mod højre. For Mellemarmen er kurven meget lineær, hvorfor det er svært at fastsætte et optimum mellem åben andel og sandsynlighed for overskridelse af badevandskvalitet.

4.5 Opsummering

I dette kapitel er det undersøgt, hvorledes badevandskvalitet kan modelleres med ledningsevnen. Ved at vælge en kritisk ledningsevne for hvornår badevandskvaliteten vurderes at være forringet, kan der opnås en total performance på 82 % korrekte modelresultater. Dette resultat er uden hensyntagen til hvorvidt, det er målinger med tilfredsstillende eller dårlig badevandskvalitet, der modelleres korrekt. Ved hjælp af en logistisk regression, er sammenhængen mellem sandsynligheden for overskridelse af badevandskravet og ledningsevnen undersøgt. Ud fra denne analyse kan det konkluderes, at der ikke er en tydelig afskæring mellem tilfredsstillende og dårlig badevandskvalitet ved brug af ledningsevnen som indikator, da kurven på sandsynlighedsfunktionen er flad. Dog er der en god indikation af, at sandsynligheden for overskridelse af badevandskriteriet stiger med lavere ledningsevne.

Ved varsling af badevandskvalitet på baggrund af ledningsevne er valget af den kritiske ledningsevne afgørende for evnen til at varsle korrekt. En høj kritisk ledningsevne vil kunne varsle om mange af hændelserne, hvor badevandkravet er overskredet, men vil så også varsle om forringet badevandskvalitet i nogle hændelser, hvor kravet ikke er overskredet. Dette har den konsekvens, at der bliver varslet om dårlig badevandskvalitet, selvom det ikke er nødvendigt. Valget af kritisk ledningsevne skal derfor være en samlet betragtning af, hvor ofte badeområdet ønskes lukket sammenholdt med sandsynligheden for, at badevandskravet er overskredet, når det ikke er lukket. Varsling af badevandskvalitet på baggrund af udelukkende ledningsevne vil medføre, at badeområdet er lukket i en stor andel af tiden, hvis sandsynligheden for overskridelse af badevandskravet skal være lille, når badeområdet er åbent.

Kapitel 5

MODELLERING AF BADEVANDSKVALITET MED MACHINE LEARNING

I de foregående tre kapitler er det konkluderet, at den bedste parameter til modellering af badevandskvalitet er ledningsevnen. Ledningsevnen er blevet brugt i en eksponentiel model og som tærskelværdi. Der er dermed mange modeltyper, der ikke er undersøgt. Derudover er der set tendenser til at andre parametre, som nedbør og vind, har en påvirking på badevandskvaliteten, selvom disse ikke kan beskrive sammenhængen alene. Nogle af disse ikke nærmere undersøgte sammenhænge og modeltyper bliver i dette kapitel undersøgt ved brug af machine learning.

I dette kapitel undersøges, om det er muligt at opnå en bedre modelperformence, end ved eksponentiel regression, ved at estimere badevandskvaliteten på baggrund af flere parametre. Når der skal opbygges en model på baggrund af flere parametre, er det ikke længere muligt at benytte en eksponentiel model med én uafhængig variabel. Dette giver mange flere frihedsgrader i modelopbygningen, da der er mange flere mulige kombinationer mellem inputparametre og algoritmer til at opbygge modellen. Her kan machine learning være et godt værktøj til at finde sammenhænge mellem flere parametre, som for det menneskelige øje kan være svære at gennemskue.

5.1 Hvad er machine learning?

Machine learning er den overordnede betegnelse for modeller, der ved at blive trænet på observationsdata, kan optimere sin evne til at forudsige responsen fra nye data [Russel og Norvig, 2010]. Til at opbygge en model til varsling af badevandskvalitet ved brug af machine learning, benyttes den type, der kaldes *supervised learning*. Supervised learning er optimering af regressions- og klassifikationsmodeller ved at benytte data med kendte responsvariable. I dette tilfælde vil det svare til, at man træner sin model på en række inputparametre og den tilhørende målte bakteriekoncentration eller badevandsklassifikation. Ved hjælp af forskellige algoritmer opbygges der en model, der kan genkende mønstre i de data, den trænes på. Forskellen på regressions- og klassifikationsmodeller er, at regressionsmodeller bruges til modellere en numerisk værdi, i denne undersøgelse en bakteriekoncentration, mens klassifikationsmodeller benyttes til at bestemme den klasse et datasæt tilhører. I denne undersøgelse er klasserne tilfredsstillende og dårlig badevandskvalitet.

5.2 Oversigt over udvalgte modelalgoritmer

Til at opbygge en machine learning-model benyttes en algoritme, der på baggrund af inputdata, kan prædiktere et output. Der findes en lang række af forskellige klassifikations- og regressionsalgoritmer til at opbygge modeller. Til udvælgelse af algoritmer er der taget udgangspunkt i de tilgængelige algoritmer i MATLABs applikationer til opbygning af regressions- og klassifikationsmodeller. Applikationerne inkluderer syv grundlæggende metoder, hvor nogle kan benyttes til både regression og

klassifikation. Indenfor de forskellige algoritmer er der yderligere en række frihedsgrader i forhold til opbygning af modeller. Af tabel 5.1 fremgår en beskrivelse af de tilgængelige algoritmer samt hvilke muligheder, der er for at variere de enkelte typer af algoritmer.

Algortime	**Variationer**	**Fordele og ulemper**
Lineær regression *Regression*	Simpel LR Robust LR (Fjerner outliers) Stepwise LR (Optimerer på antal features)	+ Simpel og intuitiv ÷ Sensitiv overfor outliers ÷ Kræver linearitet
Support vektor maskiner *Regression/klassifikation*	Kernel-funktioner: - Lineær - Polynomier af højere grad - Fordelingsfunktioner	+ Multi-dimensionale problemer + Numeriske features ÷ Kræver store datasæt
Beslutnings-træer *Regression/klassifikation*	Højere/lavere diskretisering Ensemble-træer	+ Diskrete problemer + Nem at fortolke ÷ Kan nemt blive ustabil
Gaussiansk proces regression *Regression*	Kernel-funktionener: - Eksponentiel - Kvadreret eksponentiel - Matern - 5/2-kernel	+ God til mindre datasæt + Probabalistisk metode, konfidensintervaller kan bestemmes ÷ Store datasæt giver lang træningstid ÷ Mindre effektiv i højdimensionale problemer
Diskriminant analyse *Klassifikation*	Diskriminanter: - Lineær diskriminant - Kvadratisk diskriminant	+ Simpel og intuitiv ÷ Sensitiv i forhold til irrelevante input ÷ Kræver store datasæt
Nærmeste nabo *Klassifikation*	Antal af naboer der betragtes Afstandsvægtning: - Euclidiansk afstand - Kvadreret afstand	+ Minimal træningstid (lazy-learner) ÷ Skal evaluere hele datasættet ved prædiktion ÷ Sensitiv i forhold til irrelevante input
Neurale Netværk *Regression/klassifikation*	Antal neuroner Antal skjulte lag Optimeringsalgoritmer Transferfunktioner	+ Multi-dimensionale problemer + Kan håndtere ikke-linearitet ÷ Sensitiv overfor valg af transferfunktion ÷ Kræver store datasæt ÷ Mange frihedsgrader, risiko for overfitting

Tabel 5.1. Oversigt over machine learning algoritmer og muligheder for at variere indenfor de enkelte typer. Fordele og ulemper er baseret på Kotsiantis [2007]; Naganna og Deka [2014] og C. E. Rasmussen og Williams [2005].

På baggrund af fordelene og ulemperne, der er opstillet i tabel 5.1, er det valgt, at afgrænse

fra neurale netværk. Dette skyldes, at det tilgængelige datasæt er begrænset, hvilket besværliggør opbygningen af et neuralt netværk, der ikke er overfittet.

På baggrund af redegørelsen i tabel 5.1 er det udfordrende at pege på én algoritme, der vil egne sig bedst til at løse problemet med bestemmelse af badevandskvaliteten. Det vurderes derfor, at det er nødvendigt at screene de resterende seks typer af algoritmer for at vurdere, hvilken der er bedst til regression af bakteriekoncentration, og hvilken der er bedst til klassifikation af badevandskvalitet.

5.3 Fremgangsmåde

Machine learning er et værktøj, der kan benyttes til at løse flere typer af problemer med en række af forskellige algoritmer. Derudover kan der inddrages en ubegrænset mængde af parametre. Det kan derfor hurtigt blive meget kompliceret og tidskrævende at skulle afprøve alle regressions- og klassifikationsalgoritmer og alle kombinationer af parametre. For at begrænse mængden af beregninger følges derfor følgende fremgangsmåde; først screenes udvalgte parametre, således at de parametre, der kan forbedre modellen, kan identificeres. Dette gøres ved at introducere én parameter af gangen, indtil modellens performance ikke forbedres. Herefter bruges de udvalgte parametre i de udvalgte klassifikations- og regressionsalgoritmer for at identificere den bedste algoritme. Dermed kan både de bedste parametere og den bedste algoritme identificeres. Machine-learning modellerne opbygges i dette projekt i MATLABs applikationer til opbygning af regressions- og klassifikationsmodeller. I de digitale Bilag 1 er der vedlagt et script til træning og validering af machine learning modellerne. En beskrivelse af scriptet fremgår af appendiks A.

5.3.1 Modelevaluering

Klassifikations- og regressionsmodellerne har hver deres type af output, hvorfor de vil blive evalueret på forskellige performance-parametre. Til at evaluere klassifikationsmodellerne benyttes modellens hitrate. Hitraten er et mål for hvor mange af modellens forudsigelser, der stemmer overens med det målte. Hitraten beregnes som antal korrekte modelresultater over det samlede antal modelresultater. Regressionsmodellernes evne til at modellere bakteriekoncentrationen evalueres på baggrund af to performance-parametre, RMSE og Nash-Sutcliffe Efficiency (NSE). Det er valgt at benytte disse to parametre, da NSE kan beskrive hvor godt de modellerede værdier passer til de observerede, mens RMSE er et mål for modellens usikkerhed. Derudover evalueres regressionsmodellerne også på hitraten. Yderligere beskrivelse og formler for performance-parametrene fremgår af appendiks G.

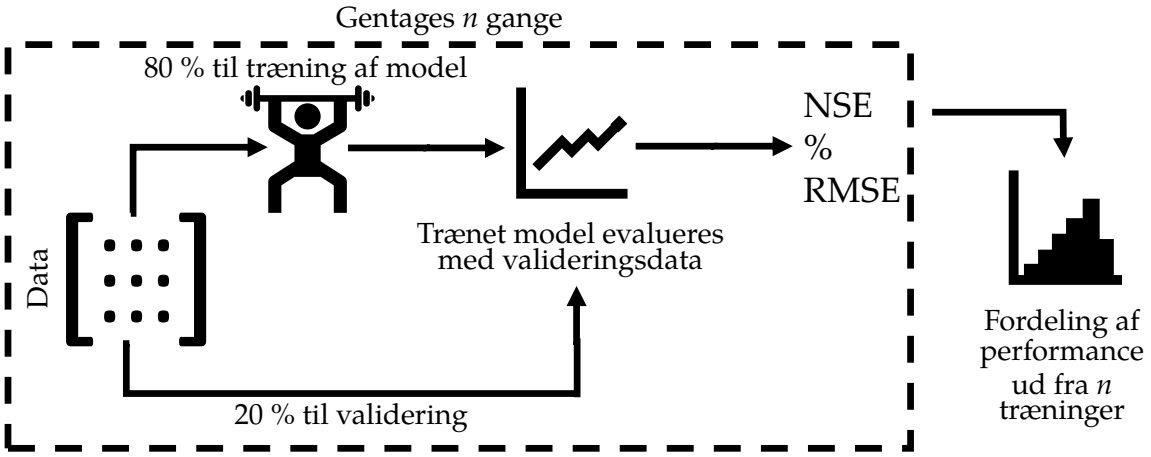
Modellerne, der opbygges ved brug af machine learning, vil blive evalueret op i mod den eksponentielle model, der er beskrevet i kapitel 4, da det ønskes at opbygge en model, der er bedre end denne. Performance-parametrene for den eksponentielle model fremgår af tabel 5.2.

Model	Hitrate [%]	NSE [-]	RMSE [log(cfu/100 ml)]
Eksponentiel regression	82	0,65	1,27

Tabel 5.2. Performance-parametre for eksponentiel regression mellem ledningsevne og E. Coli.

5.3.2 Krydsvalidering

I machine learning algoritmer kan der være mange frihedsgrader for modellen til at tilpasse sig datasættet. Dette kan være et stort problem, da modellen kan blive overfittet til datasættet og dermed ikke kan forudsige responsen til nyt data. Til opbygning af machine learning modellerne benyttes derfor kryds-validering til at sikre, at modellerne ikke overfittes. Det er valgt at bruge krydsvalidering af typen hold-out validering, da dette giver en god indikation om, hvilken performance modellen vil give på data som modellen ikke er trænet på. Princippet bag hold-out validering fremgår af figur 5.1.



Figur 5.1. Oversigt over hold-out validering.

Ved hold-out validering udelades en andel af datasættet til udelukkende at validere modellen. Det er valgt at benytte 80 % af datasættet til at træne modellerne, mens de resterende 20 % af datasættet benyttes til validering. Da datasættet er relativt lille, vil opdelingen i validerings- og træningssæt have en betydning for hvilken performance modellen får. Derfor gentages træning og validering n gange og resultatet er en fordeling af performance-parameteren.

5.4 Udvælgelse af inputparametre

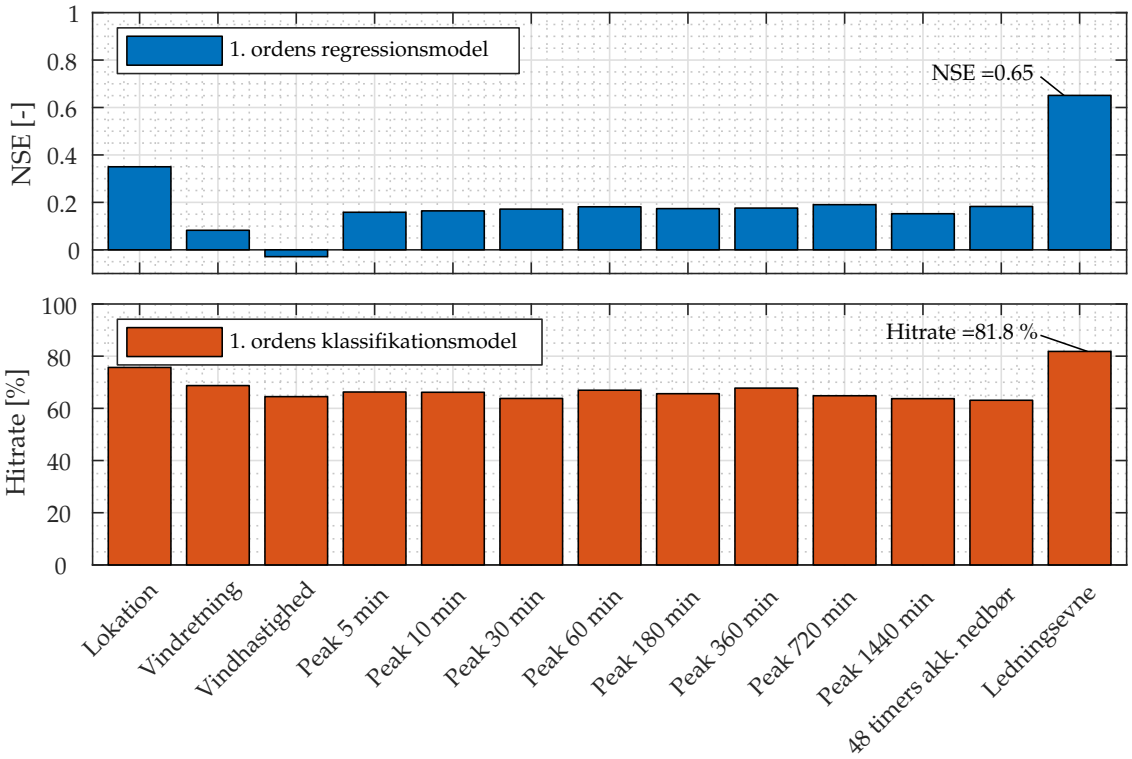
For at udvælge de inputparametre, der kan beskrive badevandskvaliteten bedst, skal der først opbygges et katalog af parametre. De udvalgte parametre er vandets elektriske ledningsevne, lokation for udtagning af vandprøve, vindretning og -hastighed, samt akkumuleret nedbør og peakintensiteter inden for de foregående 48 timer. Peakintensiteterne beregnes for varigheder på 5, 10, 30, 60, 180, 360, 720 og 1440 minutter. Dette giver i alt 13 parametre. I datasættet med bakterieanalyser er der 79 observationer med tilhørende ledningsevnemåling, når observationer

der er taget ved udløbet af Aarhus Å sorteres fra grundet den konstante lave ledningsevne og høje bakteriekoncentration. De udvalgte parametre beregnes for disse 79 datapunkter. Som responsvariabel til regressionsalgoritmerne bruges logaritmen til bakteriekoncentrationen. Til klassifikationsalgoritmerne bruges 0 for tilfredsstillende badevandskvalitet og 1 for dårlig badevandskvalitet.

Udvælgelsen af parametre udføres i ordner, hvor der udvælges en parameter per orden. Først vil der udarbejdes modeller med en parameter. Ved anden orden udarbejdes modeller med to parametre; den bedste parameter fra første orden samt én mere. Der tilføjes flere parametre indtil modellerne ikke forbedres. I udvælgelsen af parametre anvendes de udvalgte klassifikations- og regressionsalgoritmer til at evaluere parametrene.

5.4.1 Udvælgelse af første parameter

Ved udvælgelse af første parameter er alle 13 parametre til rådighed. Performance med de forskellige parametre fremgår af figur 5.2.



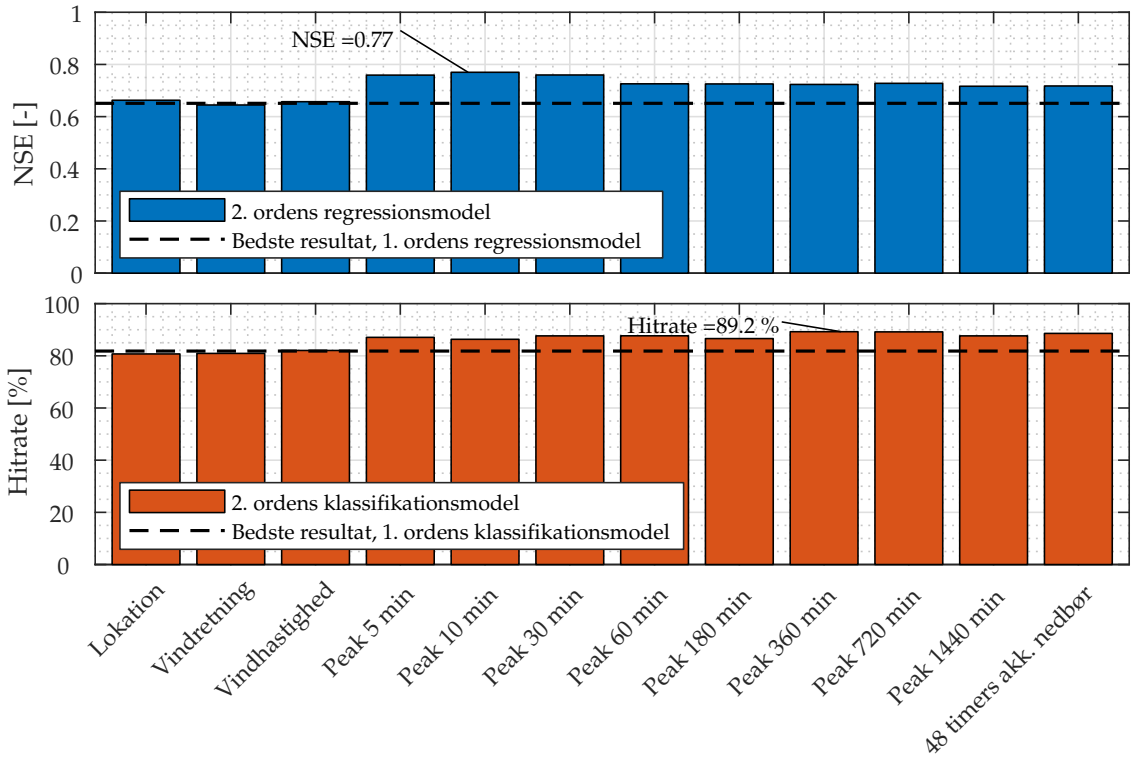
Figur 5.2. Modelperformance for modeller med én inputparameter. Den angivede performance er for den bedste algoritme inden for henholdsvis klassifikations- og regressionsalgoritmer. NSE (Nash–Sutcliffe efficiency) er modelperformance for regressionsalgoritmerne. Hitrate er modelperformance for klassifikationsalgoritmerne.

Af figur 5.2 fremgår det, at det som forventet er ledningsevnen, der er den bedste parameter til at beskrive badevandskvaliteten. Dette er gældende for både den bedste regressions- og klassifikationsalgoritme. Den anden bedste parameter er lokationen. Dette skyldes, at det gennemsnitlige bakterieniveau kan beskrives udemærket med lokationen,

da koncentrationen generelt bliver mindre, når afstanden til Aarhus Å bliver større. Ledningsevnen er dermed den første parameter, der er udvalgt til at indgå i modellerne.

5.4.2 Udvælgelse af anden parameter

Næste generation af modeller bygger på ledningsevnen samt én af de 12 resterende parametre. Performance ved at indrage en af de 12 parametre i modellen fremgår af figur 5.3.

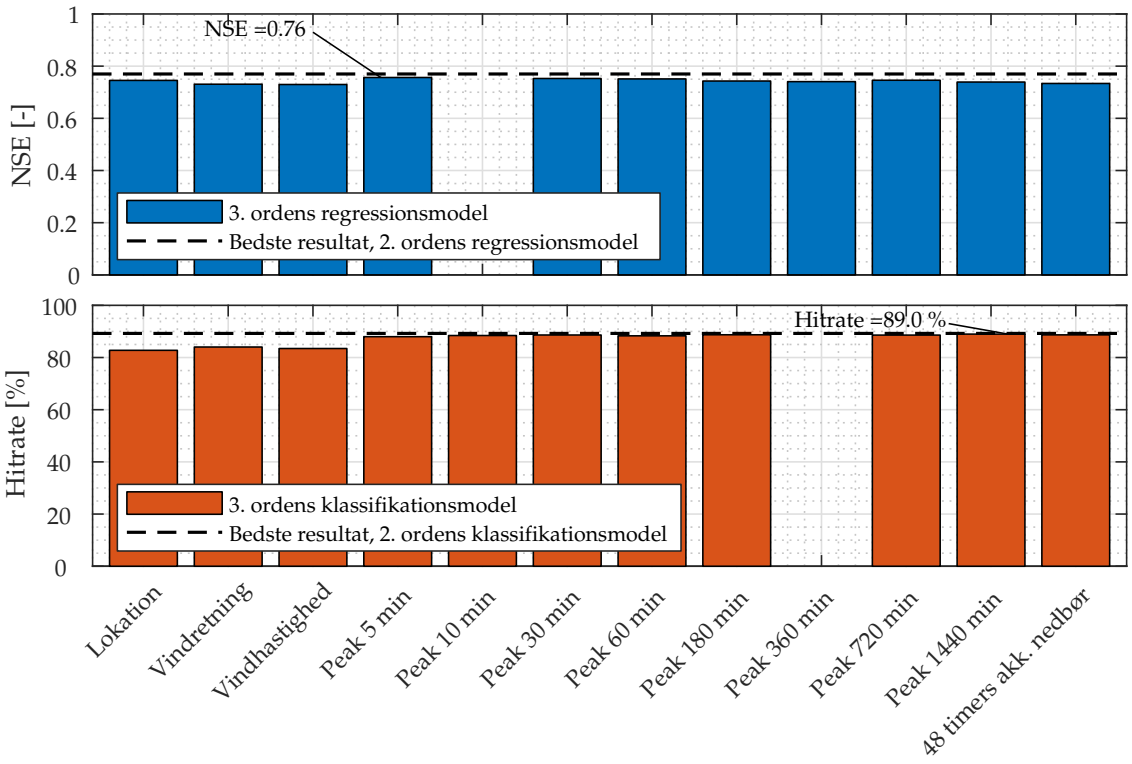


Figur 5.3. Modelperformance ved at indrage ledningsevnen, samt en ekstra parameter. Den angivede performance er for den bedste algoritme inden for henholdsvis klassifikations- og regressionsalgoritmer.

Af figur 5.3 fremgår det, at modelperformance for både regressions- og klassifikationsalgoritmerne forbedres ved at indrage en af parametrene med peakintensiteter. For klassifikationssalgoritmerne opnås den bedste performance ved at inddrage 6 timers peakintensitet. For regressionsalgoritmerne opnås den bedste performance ved at indrage 10 minutters peakintensitet. Årsagen til, at performance er meget ens for alle parametre, der indeholder information om nedbør, er sandsynligvis, at parametrene er korreleret med hinanden. Altså at kraftige korte nedbørshændelser ofte falder i samme perioder som kraftige længere nedbørshændelser. Da lokationen var den anden bedste parameter i 1. ordens modellerne, var det forventet, at denne kunne forbedre modellen yderligere. Dette er dog ikke tilfældet, da effekten af lokationen er inddraget ved brug af ledningsevnen, da denne er stærkt korreleret med lokationen. Den anden parameter til at indgå i modellerne er dermed 6 timers peakintensitet for klassifikations-modellerne og 10 minutters peakintensitet for regressions-modellerne.

5.4.3 Udvælgelse af tredje parameter

Næste generation af modeller bygger på ledningsevnen, den bedste parameter fra anden orden, samt en af de restende 11 parametre. Performance for tredje generation fremgår af figur 5.4.



Figur 5.4. Modelperformance ved at indrage ledningsevnen, den bedste parameter fra anden generation samt en ekstra parameter. Den angivede performance er for den bedste algoritme inden for henholdsvis klassifikations- og regressionsalgoritmer.

Af figur 5.4 fremgår det, at modellerne ikke kan forbedres ved at indrage en tredje parameter. Det skyldes blandt andet, at indflydelsen fra nedbør er indraget i form af parameteren fra anden generation. Tidligere i dette projekt er det vist, at vindretningen har en indflydelse på badevandskvaliteten og ledningsevnen i Aarhus Havn. Det er dog ikke lykkedes at forbedre machine learning-modellernes performance ved at inddrage vindretningen. Dette kan skyldes, at vindretningen ikke følger en lineær-akse, da $0° \approx 359°$, og det kan derfor være en udfordring for algoritmerne at tilpasse sig inputtet. Det er i stedet forsøgt at dele vindretning ind i 12 vindretningsklassser uden mærkbare forbedringer. Årsagen til at vinden ikke forbedrer modellerne kan være, at en del af effekten fra vinden allerede er indregnet i ledningsevnen, og ledningsevnen derfor sandsynligvis kan beskrive effekten af vinden bedre end vinden selv. Dermed forbliver det bedste input til modellerne to parametre, der er henholdsvis ledningsevne og 10 minutters peakintensitet for regressionsalgoritmerne samt ledningsevne og 360 minutters peakintensitet for klassifikationsalgoritmerne.

5.5 Opbygning af regressionsmodel

Formålet med en regressionsmodel er at kunne modellere bakteriekoncentrationen ud fra en række inputparametre. For at undersøge hvilken algoritme, der egner sig bedst til at modellere badevandskvalitet, undersøges performancen for de udvalgte algoritmer fra tabel 5.1.

5.5.1 Udvælgelse af algoritmer til opbyggelse af model

I afsnit 5.4 er det vist, at de parametre der bedst kan benyttes til at estimere bakteriekoncentrationen i Aarhus Havn, er ledningsevnen og den maksimale regnintensitet for en varighed på 10 minutter indenfor de seneste 48 timer. I undersøgelsen, der ligger til grund for valget af netop disse inputparametre, er der benyttet en række af forskellige typer af regressionsmodeller til at finde den bedste kombination af parametre. I tabel 5.3 fremgår rangeringen af de 5 bedste modeller med disse 2 input-parametre, når de evalueres på henholdsvis NSE og RMSE.

Algoritme	**Variant**	**Middel NSE [-]**
Gausiansk proces regression	Kernel: Eksponentiel	0,77
Gausiansk proces regression	Kernel: Rationel kvadratisk	0,72
Gausiansk proces regression	Kernel: Matern	0,72
Lineær regression	Simpel LR	0,71
Support vektor maskine	Kernel: Lineær	0,70
Algoritme	**Variant**	**Middel RMSE [log(cfu/100 ml)]**
Gausiansk proces regression	Kernel: Eksponentiel	1,07
Gausiansk proces regression	Kernel: Rationel kvadratisk	1,14
Lineær regression	Trinvis LR	1,15
Gausiansk proces regression	Kernel: Matern	1,15
Lineær regression	Robust LR	1,16

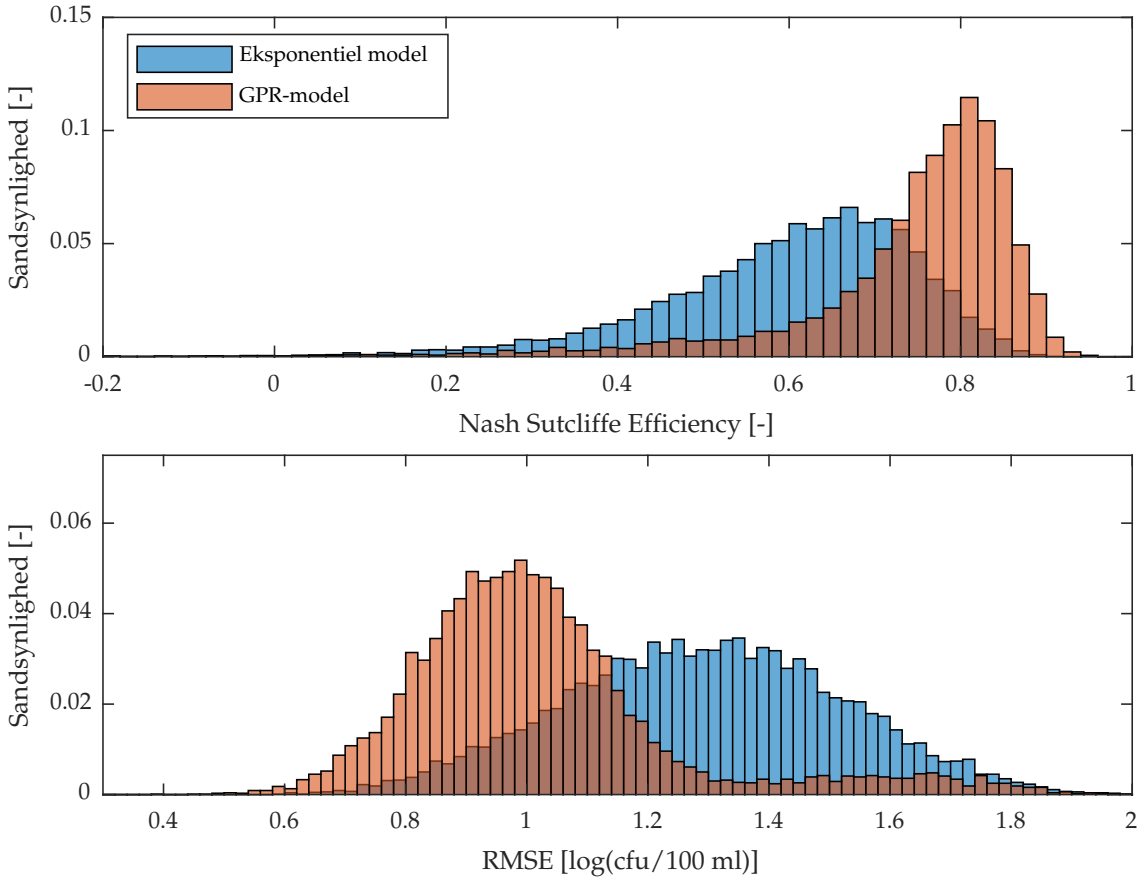
Tabel 5.3. Rangering af de 5 bedste regressionsmodeller målt på henholdsvis NSE og RMSE. Performance-parametrene er beregnet ved kryds-validering og er et gennemsnit af 1000 træninger.

Af tabel 5.3 fremgår det, at der er tre typer af modeller, der ligger blandt de bedst rangerende. Det er Gaussiansk Proces Regression (GPR), lineær regression samt én support vektor maskine (SVM) algoritme. Det fremgår, at den bedste model er eksponentiel GPR, både når den evalueres på NSE og RMSE. Denne vil derfor være grundlag for videre undersøgelser. Gaussian Process Regression er baseret på den multivariate normal-fordeling mellem de variable, der indgår i modellen. Yderligere beskrivelse af GPR-algoritmerne fremgår af appendiks F.

5.5.2 Evaluering af GPR-model

Til at evaluere GPR-modellen, udføres 10.000 træninger med hold-out validering. Ud fra de 10.000 træninger opnås en fordeling af modellens performance-parametre, som kan sammenlignes med fordelingen af performance-parametrene for den eksponentielle

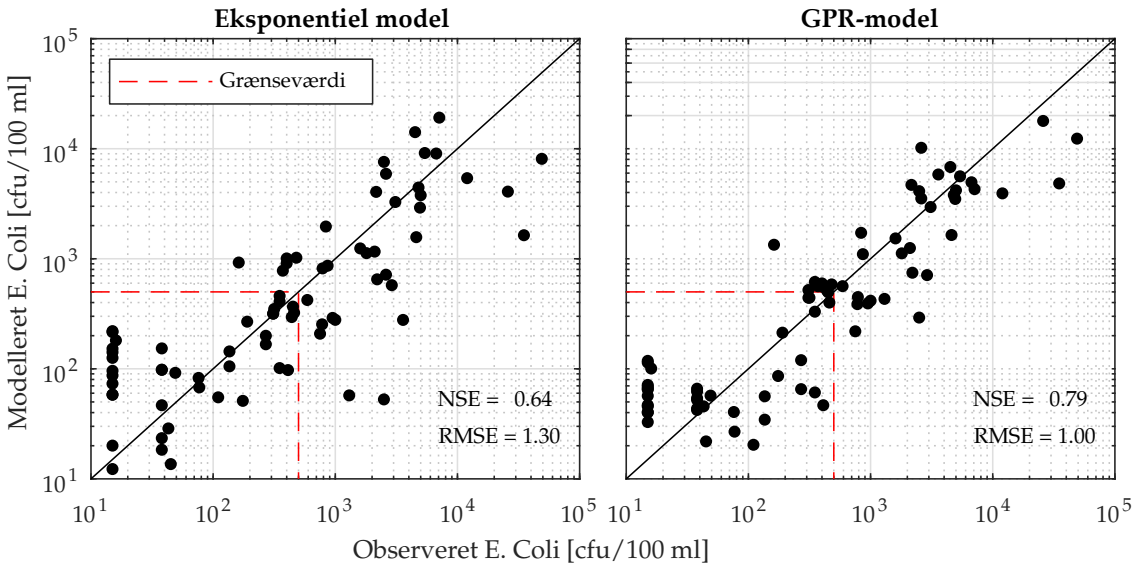
model. Af figur 5.5 fremgår fordelingen af NSE og RMSE for den eksponentielle model, udarbejdet i kapitel 4, samt den eksponentielle GPR-model for 10.000 træninger.



Figur 5.5. Fordeling af performance-parametre for eksponentiel regression og eksponentiel GPR. Fordelingen er baseret på 10.000 kombinationer mellem trænings- og valideringsdata.

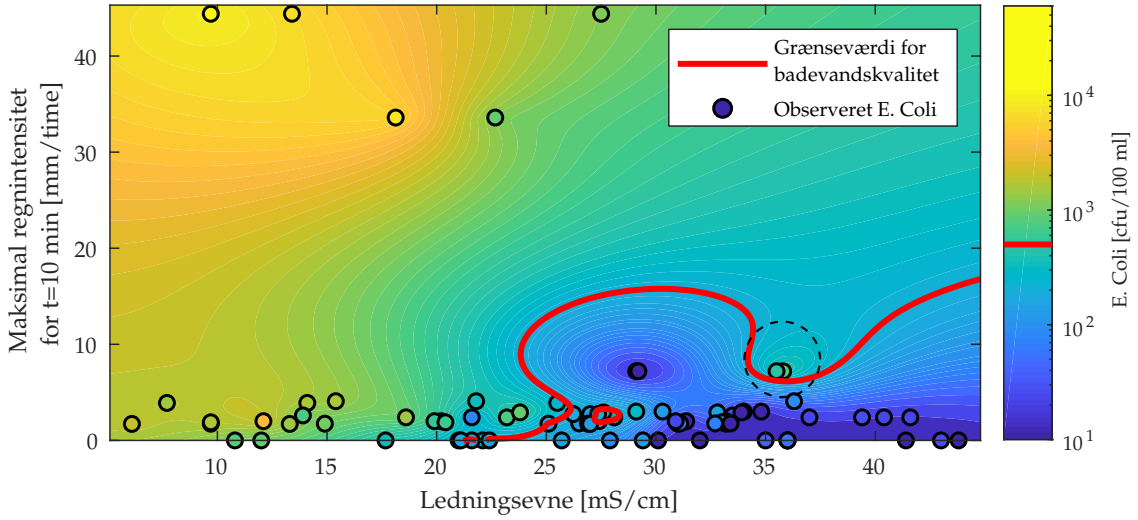
Af figur 5.5 fremgår det, at fordelingen af performance-parametre for den eksponentielle GPR har en højere middelværdi for NSE og lavere for RMSE samt mindre spredning end den eksponentielle model. Dette viser, at GPR-modellen giver et væsentligt bedre estimat af bakterie-koncentrationen end den eksponentielle model.

Af figur 5.6 fremgår et responsplot, der viser middelværdierne for de modellerede værdier af E. Coli sammenholdt med den observerede værdi af E. Coli. Det fremgår, at der er en tydelig forbedring i den modellerede værdi af E. Coli. Det kan i særdeleshed bemærkes, at de tre højeste observerede værdier af E. Coli modelleres væsentligt bedre med GPR-modellen end med den eksponentielle model. Derudover er der en generel tendens til, at punkterne ligger nærmere vinkelhalveringslinjen på respons-plottet, hvilket indikerer en bedre model.



Figur 5.6. Responsplot mellem den modellerede og observerede koncentration af E. Coli for eksponentiel regression og GPR-model. Den modellerede koncentration er middel af forudsigelser fra 10.000 træninger, med forskellige kombinationer af trænings- og valideringsdata.

Da modellen kun har to input-parametre, kan der ud fra modellen genereres en responsflade, der kan visualisere forholdet mellem input-parametrene og bakteriekoncentrationen. Dette gøres ved at bruge modellen til at estimere bakteriekoncentraitonen i alle punkter på fladen. Responsfladen fremgår af figur 5.7.

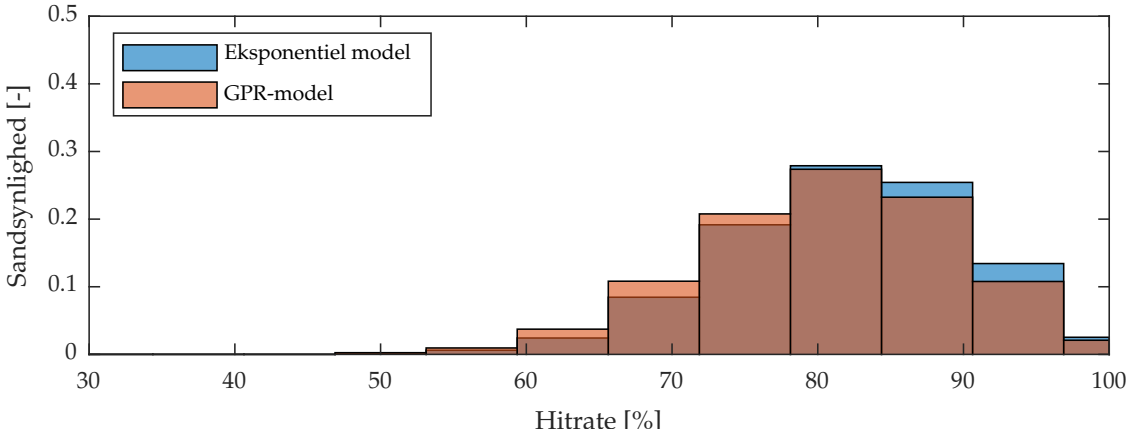


Figur 5.7. Sammenhæng mellem maksimal regnintensitet for varighed på 10 minutter inden for de seneste 48 timer, ledningsevne og E. Coli-koncentration, estimeret ved GPR-model.

Af responsfladen på figur 5.7 fremgår det, at de meget høje koncentrationer af E. Coli opstår i forbindelse med kraftige regnhændelser. Dette er årsagen til, at GPR-modellen er bedre til at modellere de høje værdier i forhold til den eksponentielle model, der ikke tager højde for nedbør. Af responsfladen fremgår det, at modellen har tilpasset sig omkring en gruppe af observationer ved en regnintensitet på ca. 7 mm/time og en ledningsevne

på 36 mS/cm, markeret med en sort stiplet cirkel på figur 5.7. Ved disse observationer er der konstateret et højt indhold af E.Coli, og det har modellen tilpasset sig efter. Det fremgår dog, at der er observationer med samme nedbør og lavere ledningsevne, der har et lavere indhold af E. Coli, hvilket strider imod den fysiske forståelse af systemet. Dette er formentlig et udtryk for, at det ikke er alle situationer, der er repræsenteret i datasættet, og at der kan være observationer, der er gjort på dage, der ikke ligger inden for normalen.

Af figur 5.7 fremgår desuden grænseværdien for tilstrækkelig badevandskvalitet. Ud fra denne grænseværdi kan modellen benyttes til at vurdere, hvorvidt badevandskvaliteten overskrider kriteriet eller ej. Ud fra 10.000 træninger med forskellige kombinationer af trænings- og valideringsdata undersøges modellens evne til at estimere hvorvidt, badevandskriteriet er overskredet. Til hvert af de 10.000 sæt udvælges 80 % af de 79 datapunkter til træning, og de restende 20 % udvælges til validering. Dette giver 16 punkter til hver validering og diskretiseringen af hitraten bliver dermed 6,25 %. Da GPR-modellen kun er opbygget til at kunne modellere E. Coli-koncentrationen, er det kun E. Coli, der bruges til at vurdere, om badevandskriteriert er overskredet. Dette gøres, da der i langt de fleste tilfælde er en overskridelse af grænseværdien for E. Coli, hvis der er overskridelse af grænseværdien for enterokokker. Af figur 5.8 fremgår hitraten af GPR-modellen og den eksponentielle model.



Figur 5.8. Fordeling af hitraten, beregnet ud fra den modellerede koncentration for eksponentiel regression og eksponentiel GPR. Fordelingen er baseret på 10.000 kombinationer mellem trænings- og valideringsdata.

Af figur 5.8 fremgår det, at den eksponentielle model er marginalt bedre til at vurdere overskridelser af badevandskriteriet. Årsagen til at GPR-modellen ikke er bedre, på trods af den bedre NSE og RMSE, kan skyldes, at den er bedre til at estimere de høje bakteriekoncentrationer, men at værdierne omkring grænseværdien ikke modelleres bedre i GPR-modellen end i den eksponentielle model.

5.6 Opbygning af klassifikationsmodel

Formålet med klassifikationsmodeller er at kunne opdele datasæt i de korrekte klasser. I dette projekt er formålet med klassifikationsmodellerner at kunne opdele datasæt i de observationer, der vil resultere i dårlig eller tilfredsstillende badevandskvalitet.

I afsnit 5.4 er det vist, at de bedst egnede parametre til klassifikationsmodellerne er ledningsevnen samt den maksimale nedbørsintensitet for en regnvarighed på 360 minutter indenfor de sidste 48 timer. Disse to parametre benyttes nu sammen med de udvalgte klassifikationsalgoritmer for at identificere de algoritmer, der giver den bedste performance. De fem bedste algoritmer fremgår af tabel 5.4.

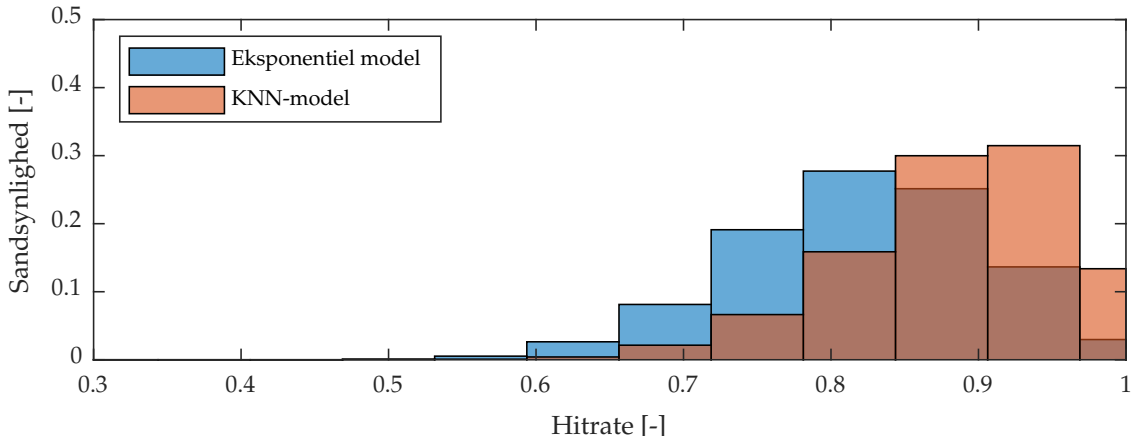
Algoritme	**Variant**	**Middel hitrate [%]**
Nærmeste nabo	Antal naboer: 1	89,2
Nærmeste nabo	Antal naboer: 10	88,7
Diskrimant analyse	Lineær	83,7
Support vektor maskine	Kernel: Gaussiansk fordeling	83,0
Support vektor maskine	Kernel: 3. grads polynomie	82,7

Tabel 5.4. Hitrate for de fem bedste klassifikationsalgoritmer. Hitraten er et gennemsnit af 1000 træninger.

De to bedste algoritmer til at klassificere badevandskvaliteten er begge af typen nærmeste nabo, også kaldet KNN. Klassificering med KNN gøres ved at plotte en ny observation i samme koordinatsystem som de målte data og fastsætte afstanden og klassifikationen af de K nærmeste observationer. Derefter kan den afstandsvægtede klasse beregnes, og klassen for den nye observation kan bestemmes. Dokumentation omkring KNN-, liniær diskriminant- og SVM-algoritmer fremgår af appendiks F.

5.6.1 Evaluering af KNN-model

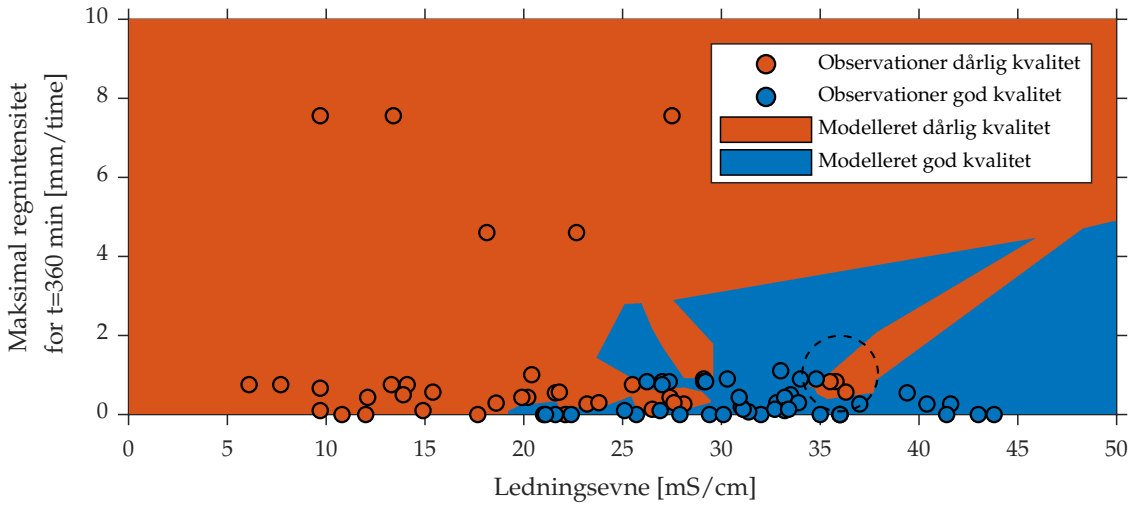
For undersøge nøjagtigheden af den bedste algoritme på nye data trænes den nu med 10.000 trænings- og valideringsdatasæt. Fordelingen af hitraten for KNN-modellen fremgår af figur 5.9. Her fremgår også hitraten for den eksponentielle model, der er trænet og valideret på samme datasæt.



Figur 5.9. Fordeling af modelperformance på valideringsdata for eksponentiel regression og KNN-algoritmen. Fordelingen er baseret på 10.000 kombinationer mellem trænings- og valideringsdata.

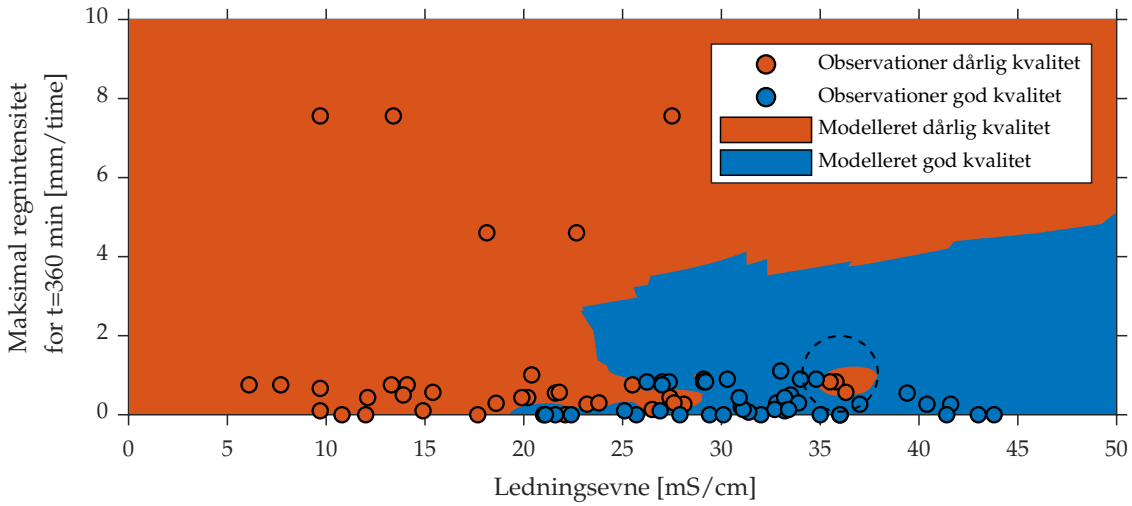
På figur 5.9 fremgår det, at nærmeste nabo-algoritmen har en væsentligt højere hitrate

end den eksponentielle model. Dette kan både skyldes, at der er inddraget en ekstra parameter, samt at KNN er en bedre algoritme til at beskrive sammenhængen end en eksponentiel regression. En visualisering af modelresultaterne fra én nærmeste nabo-algoritmen fremgår af figur 5.10.



Figur 5.10. Modelresultater for én nærmeste nabo algoritmen, samt de observerede datapunkter.

Af figur 5.10 fremgår det tydeligt, at høj ledningsevne og lave regnintensiteter indikerer tilfredsstillende badevandskvalitet. Dette stemmer godt overens med de tidligere undersøgelser. Det fremgår dog af figuren, at der er områder med høj ledningsevne og lav regnintensitet, der modelleres som værende dårlig kvalitet. Området er markeret med en stiplet cirkel på figur 5.10. Da der ikke er nogen fysisk forklaring til disse områder, undersøges det om områdernes størrelse kan reduceres ved at benytte 10 nærmeste naboer-algoritmen. 10 nærmeste naboer-algoritmen har stort set samme performance som én nærmeste nabo. Modelresultaterne fra denne algoritme fremgår af figur 5.11.

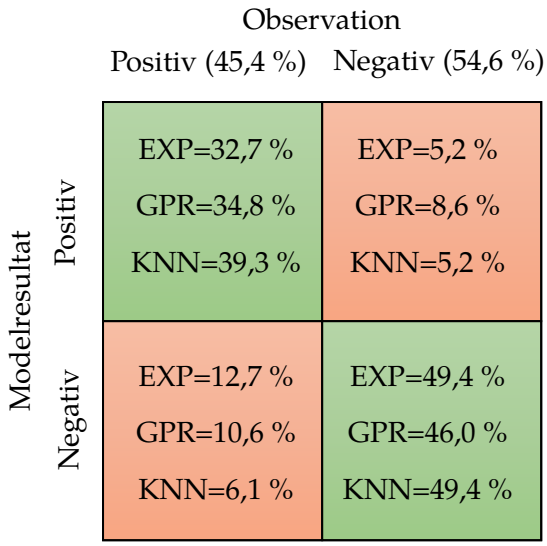


Figur 5.11. Modelresultater for 10 nærmeste naboer algoritmen, samt de observerede datapunkter.

Af figur 5.10 og 5.11 fremgår det, at områderne med høj ledningsevne og lav regnintensitet, der modelleres som værende dårlig kvalitet, er størst, når der modelleres ud fra én nærmeste nabo. Dette skyldes, at en enkelt observation betyder mindre, da der midles over flere målinger, hvilket mindsker fejlen ved outliers i datasættet. Områderne forsvinder dog ikke, når de 10 nærmeste naboer inddrages. Områderne forventes at være et udtryk for usikkerheden i datasættet og ikke at have en fysisk forklaring. Områderne bør derfor også ignoreres, når modellen skal bruges til at modellere badevandskvaliteten på nye datasæt. Til dette formål kan der laves en forsimplet model, der bygger på erfaringerne fra machine learning-modellerne.

5.7 Sammenligning af modeller

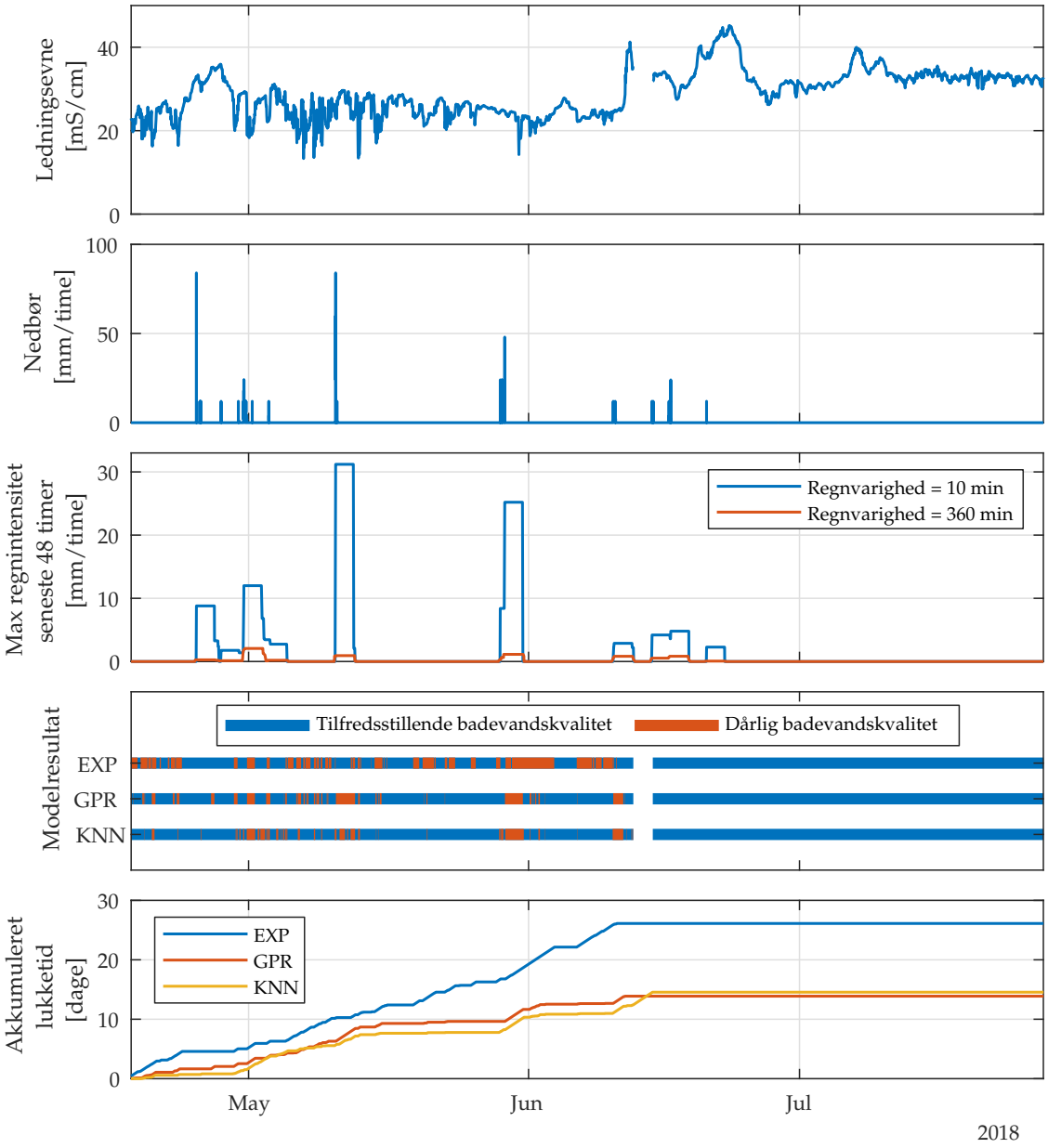
Der er tidligere i kapitlet konkluderet, at GPR-modellen opnår en lidt dårligere samlet hitrate end den eksponentielle model, mens KNN-modellen opnår en væsentlig bedre samlet hitrate. For at opnå et mere nuanceret billede af modellernes performance, opdeles modelresultaterne i en confusion-matrix. En confusion-matrix viser fordelingen af korrekte og forkerte modelresultater i forhold til, om de er positive (dårlig badevandskvalitet) eller negative (tilfredsstillende badevandskvalitet). De tre modellers confusion matrix fremgår af figur 5.12.



Figur 5.12. Confusion-matrix for den eksponentielle model (EXP), GPR-modellen samt KNN-modellen. Figuren viser middelresultatet for 10.000 træninger.

Af figur 5.12 fremgår det, at både GPR-modellen og KNN-modellen er bedre til at forudsige positive observationer med dårlig badevandskvalitet end den eksponentielle model. Samtidig er GPR-modellen dårligere til at forudsige negative observationer, mens KNN-modellen her performer lige så godt som den eksponentielle model på negative observationer. Dette betyder, at badning vil frarådes flere gange, hvor det ikke er nødvendigt ved brug af GPR-modellen. Modellens hovedformål er at varsle om forringet badevandskvalitet, hvorfor en god performance på positive observationer er vigtigere end på negative observationer.

For at undersøge hvordan modellerne fungerer i praksis, samt hvilken påvirkning de har på lukketiden, er det undersøgt hvilke resultater, de tre modeller ville give, hvis de var blevet benyttet i badesæsonen for år 2018. Af figur 5.13 fremgår modelinput og -resultat samt den akkumulerede lukketid fra midt april 2018 til slut juli 2018.



Figur 5.13. Modelinput, -resultater og akkumuleret lukketid for den eksponentielle model (EXP), GPR-modellen samt KNN-modellen i sommeren 2018.

Som det fremgår af figur 5.13, har den eksponentielle model væsentlig flere dage, hvor den modellerer dårlig badevandskvalitet end både GPR- og KNN-modellen. For de ca. 100 dage, der er vist på figur 5.13, er der ved brug af den eksponentielle model forudsagt 26 dage med dårlig badevandskvalitet, mens GPR- og KNN-modellen har forudsagt henholdsvis 14 og 14,5 dage med dårlig badevandskvalitet. Derudover er antallet af gange, hvor der skal udsendes et varsel, også reduceret væsentligt. Ved at benytte disse modeller med flere input-parametre, opnås der dermed ikke blot en mere præcis varsling,

men lukketiden vil ligeledes kunne reduceres.

5.8 Opsummering

I dette kapitel er det undersøgt, hvorvidt inddragelse af flere inputparametre kan bidrage til en mere præcis model. Inputparametrene er testet i en lang række machine learning-algoritmer. De bedste inputparametre er vurderet at være elektrisk ledningsevne og nedbør. Algoritmerne med den bedste performance er KNN og GPR for henholdsvis klassifikation og regression. En oversigt over modellernes performance og lukketid fremgår af tabel 5.5.

Model	Hitrate [%]	RMSE [log(cfu/100 ml)]	NSE [-]	Lukketid sommer 2018 [dage]
EXP	82	1,29	0,60	26
GPR	81	1,02	0,74	14
KNN	89	-	-	15

Tabel 5.5. Performance-parametre og lukketid for den eksponentielle model (EXP), GPR-modellen samt KNN-modellen. Performance er et gennemsnit for valideringsdata på 10.000 træninger.

Af tabel 5.5 fremgår det, at performance og lukketid er væsentligt forbedret ved at indrage nedbør i modellen samt at benytte machine learning-algoritmer. Selvom der er benyttet andre og mere komplicerede model-algoritmer til opbyggelse af GPR- og KNN-modellerne, vurderes det, at det er regninputtet til de to modeller, der udgør den store forskel i forhold til den eksponentielle model. Det, at nedbør medtages som input, har den fordel, at der ved regnhændelser er større sandsynlighed for, at der sker overløb fra fælleskloaker, hvor der er en høj koncentration af E. Coli. Regninputtet kan dermed give en bedre forudsigelse af bakteriekoncentrationen i Aarhus Å, mens ledningsevnen bruges til at vurdere, hvorvidt vandet fra Aarhus Å strømmer der hen, hvor der varsles om badevandskvaliteten.

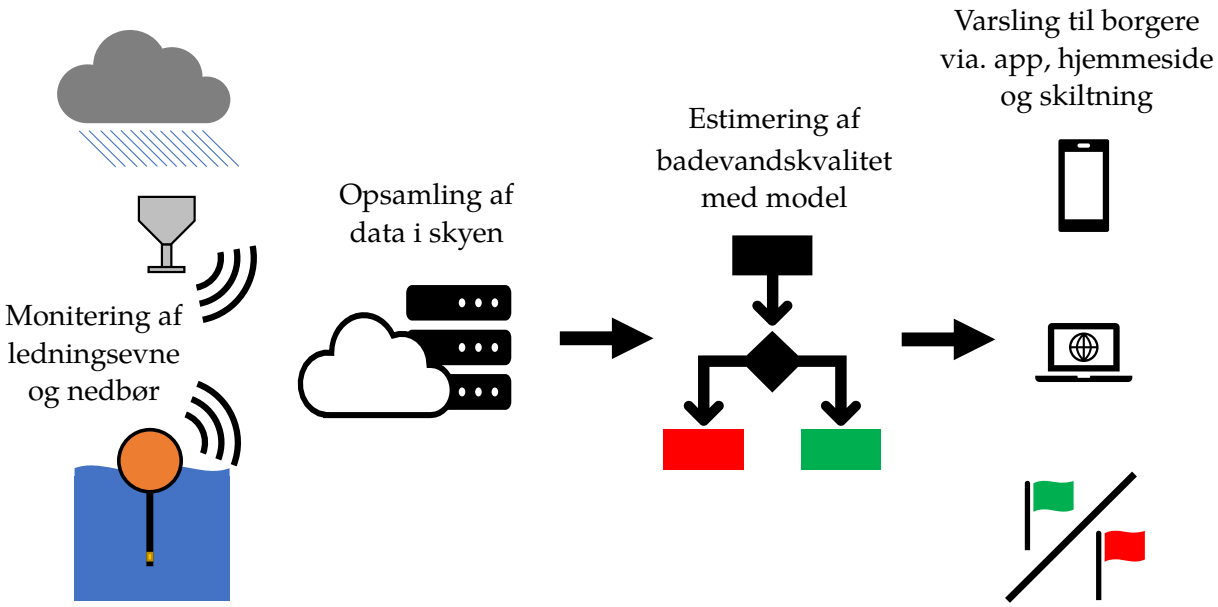
Som det fremgår af responsfladerne for GPR- og KNN-modellerne på figur 5.7 og 5.10, er der store områder, der ikke er repræsenteret af observationer. For at forbedre modellerne vil det derfor give mening at indsamle data, i forbindelse med regnhændelser, der har intensiteter, der ikke er repræsenteret på responsfladerne.

Kapitel 6

IMPLEMENTERING AF VARSLINGSMETODE

I dette kapitel beskrives det, hvordan den udviklede metode til varsling af badevandskvaliteten kan implementeres.

I kapitel 5 er det vist, at badevandskvaliteten kan varsles på baggrund af vandets ledningsevne og nedbør. Den opstillede model er dog kun én komponent i et komplet varslingssystem. Et komplet varslingssystem består, udover selve modellen, også af monitering af de nødvendige data samt udsendelse af information om badevandskvaliteten til borgerne. Af figur 6.1 fremgår informationsstrømmen fra måler til borger.



Figur 6.1. Princip over komplet varslingssytem; fra indsamling af data til varsling af borgere.

Første trin i varslingen af badevandskvalitet er indsamling af ledningsevne- og nedbørsdata. For at kunne varsle i realtid er det nødvendigt med et online system, der kan måle og sende data hjem. Til dette formål benyttes nedbørsdata fra SVK's målere, der kan tilgås online og bliver opdateret flere gange i timen. Til online indsamling af ledningsevnedata er det nødvendigt at opbygge et nyt system, der udover at måle ledningsevnen også kan sende data hjem. Desuden kan der opstå problemer med algevækst på ledningsevnesensoren, hvis den er under vand konstant. Derfor udvikles der en målestation, der udover at kunne monitere ledningsevnen online, skal løse problemet med algevækst. Dernæst opbygges et system, der behandler disse data og viser resultatet på en hjemmeside.

6.1 Forbedring af metode til måling af ledningsevne

I løbet af projektperioden har der været opstillet fire målestationer i Aarhus fra 8. november 2018 til 19. marts 2019. I perioden er målestationerne tilset, og sensoren er rengjort flere gange med tre til syv ugers mellemrum. Ved tilsyn er det konstateret, at der er en væsentlig vækst af alger på sensoren. Algevæksten har vist sig at være værre, når der er længere mellem tilsynene og i solrige perioder. På figur 6.2 fremgår fire fotos af den værste algevækst, der er konstateret i perioden.

Figur 6.2. Algevækst på ledningsevne-loggere

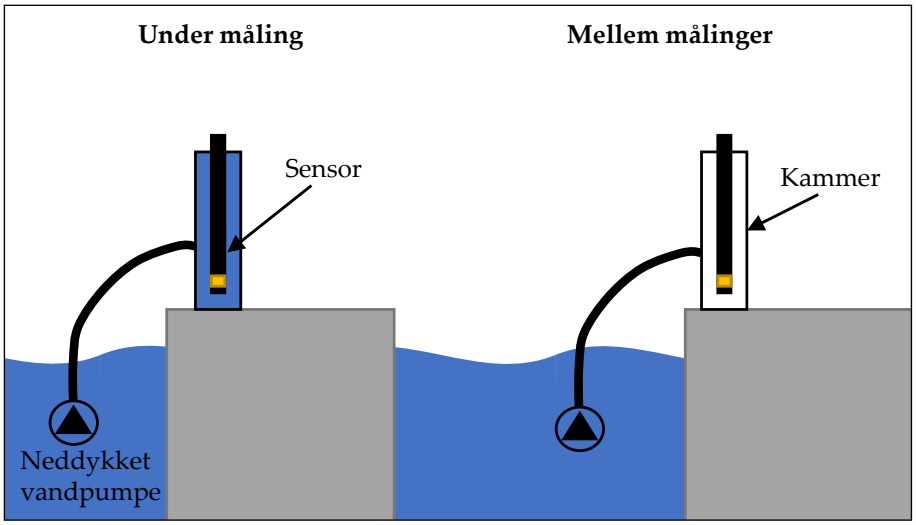
Væksten af biofilm har betydning for måleresultaterne, da biofilmen dækker sensoren, og det derfor er ledningsevnen af biofilmen og ikke af vandet, der måles. Derfor skal algevæksten på sensoren undgås.

6.1.1 Idéer til målestation med begrænset algevækst

Algevækst på neddykkede materialer i næringsrige havnemiljøer er et velkendt problem, der eksempelvis for joller afhjælpes ved at bundmale med algebekæmpelsesmiddel eller tage jollen op mellem brug. For ledningsevneloggeren er der ikke samme muligheder, da sensoren ikke må overmales eller kan tages manuelt op af vandet mellem hver måling. Det betyder, at målestationen skal rengøres hyppigt for at sikre korrekte resultater, hvilket ikke er hensigtsmæssigt, da det på sigt kan ødelægge sensoren og er en driftsmæssig omkostning. Derfor er der brainstormet for at finde frem til en ny type målestation.

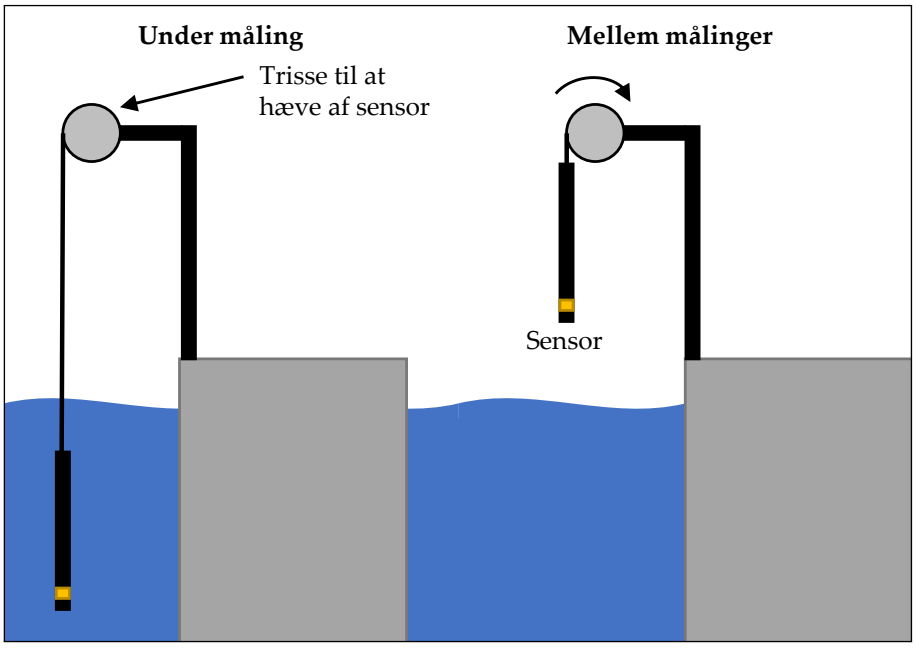
Første idé til en ny type målestation fremgår på figur 6.3. Målestationen fungerer ved at pumpe vandet op i et kammer, hvor sensoren kan placeres. Mellem målingerne kan kammeret tømmes, således at sensoren kan holdes tør og ikke gror til. Udfordringen ved denne målestation er, at det nu er pumpen, der risikere at gro til, da den hele tiden er neddykket. Derudover kan energiforbruget være en udfordring da vandet skal løftes op

til kammeret.



Figur 6.3. Første princip bag målestation med begrænset algevækst på sensoren.

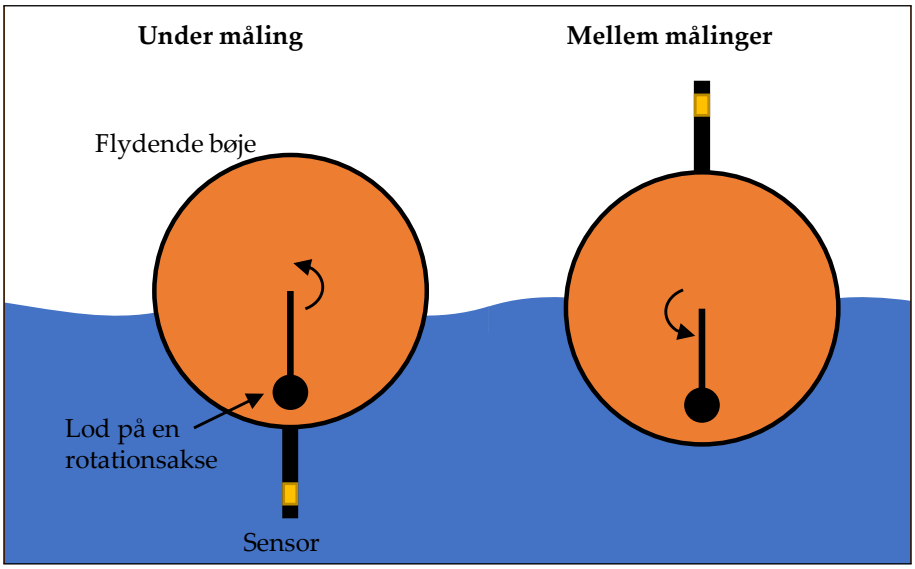
Anden idé til en ny målestationen fremgår af figur 6.4. Her bliver sensoren ført op og ned af vandet via en trisse der kan rotere ved hjælp af en motor. Fordelen ved denne målestation er, at ingen dele er under vand mellem målingerne, og der er derfor ikke noget, der gror til. Udfordringen ved denne målestation, er at der kan sætte sig tang og andre flydene ting på ledningen og sensoren, som kan blive ført op til trissen og motoren, og sætte sig fast. Desuden kan det være vanskeligt at sikre sig, at sensoren altid er i den rigtige dybde, når der måles, da højden af vandspejlet varierer.



Figur 6.4. Andet princip bag målestation med begrænset algevækst på sensoren.

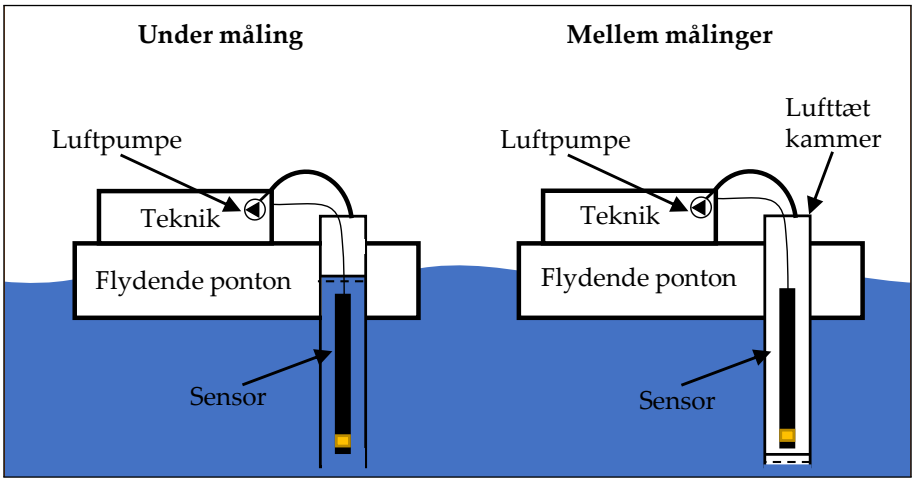
Tredje idé til en ny målestation fremgår af figur 6.5. Her placeres sensoren uden på en rund bøje. Hvilken side af bøjen, der vender ned mod vandet, kan indstilles ved at rotere

et lod inde i bøjen. Dermed kan den side af bøjen med sensoren på vendes ned i vandet under målingen og omvendt, når der ikke måles. Udfordringen ved denne målestation kan være at få plads til alt teknik og mekanik inde i bøjen samtidig med, at det skal være loddet, der styrer massemidtpunktet i bøjen. Desuden kan det være en udfordring at fastmontere strømkabler og fortøjringer til bøjen, når den skal kunne rotere frit.



Figur 6.5. Tredje princip bag målestation med begrænset algevækst på sensoren.

Fjerde og sidste idé til en ny type målestation fremgår af figur 6.6. Her placeres sensoren i et cylinderformet kammer i midten af en flydende ponton. Kammeret stikker ned i vandet og er lukket i toppen og åben i bunden. Sensoren holdes ude af vandet mellem målinger ved at pumpe luft ind i kammeret og derved fortrænge vandet. Mellem målingerne holdes trykket, og når der skal måles, åbnes en ventil der lukker luften ud, således at vandet kommer op i kammeret og op til sensoren. Den flydende ponton sikrer, at den følger tidevandet og altid måler 30 cm under overfladen. På pontonen placeres alt elektronik i en vandtæt beholder. Den flydende ponton fortøjres, og der trækkes en strømforsyning fra kajkanten til den vandtætte beholder.



Figur 6.6. Fjerde princip bag målestation med begrænset algevækst på sensoren.

For målestationen på figur 6.6 er der ingen mekaniske dele, der har kontakt til vandet, hvilket er en stor fordel. Desuden har målestationen den fordel, at hvis der opstår tekniske problemer med pumpen, så vil sensoren fortsat være i vandet, og det vil derfor fortsat være muligt at måle ledningsevnen, indtil fejlen på pumpen kan udbedres. Det er dog en udfordring med denne målestation, at systemet skal være tæt for at kunne opretholde trykket i kammeret mellem målinger. På baggrund af de ovenstående overvejelser vælges det at gå videre med den sidste idé til en ny målestation.

6.1.2 Opbygning af målestation

Den fysiske opbygning af målestationen fremgår af figur 6.7 og består af tre dele. En ponton opbygget af en skumplade belagt med hård plast. Denne har til formål at danne en platform til teknikken og give opdrift. Teknikken placeres i det orange rør med en prop i begge ender. Dette har den fordel, at det er nemt at tilgå teknikken, da proppen nemt kan tages af og på. Fra det store rør føres en ledning til ledningsevnesensoren, en ledning til temperatursensoren samt slange til luft. Disse føres via vandtætte koblinger til det mindre grå rør. Dette rør er så småt som muligt for at gøre mængden af vand, der skal fortrænges, så lille som mulig. I røret hænger temperatursensoren og ledningsevnesensoren frit halvejs nede i røret.

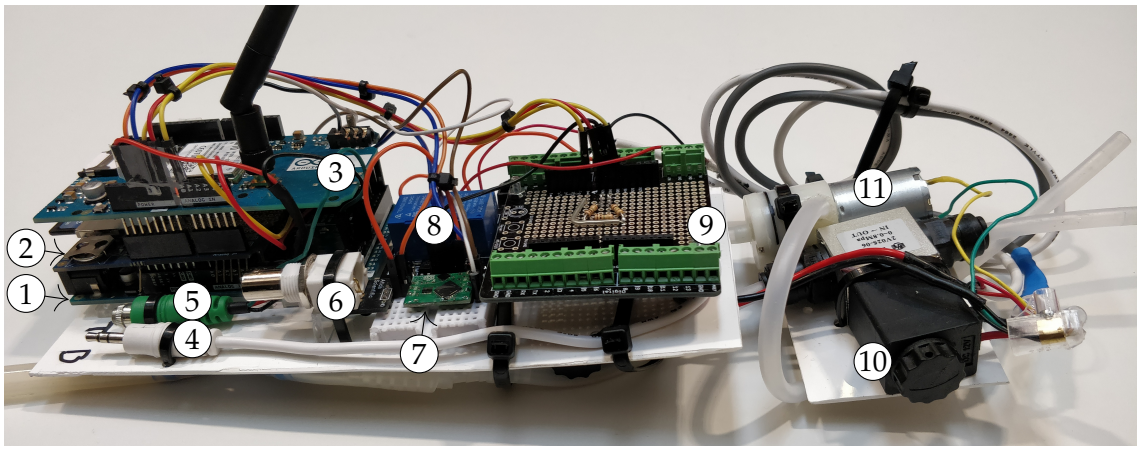
Figur 6.7. Fysisk opbygning af målestation.

Teknikken består af en række elektroniske komponenter, der er placeret i det orange rør. De elektroniske komponenter fremgår af nedstående oversigt:

1. Ardunio MEGA til styring af signaler fra og til øvrige komponenter
2. SD-kort-modul med indbygget ur til at logge data fra sensorerne
3. GSM-modul til upload af filer fra SD-kort til FTP-server
4. Stik til tilslutning af 12V strømforsyning
5. Stik til tilslutning af temperatursensor
6. Stik til tilslutning af Atlas Scientific ledningsevnesensor
7. Kredsløb til seriel kommunikation mellem Atlas Scientific ledningsevnesensor og Arduino MEGA
8. Relæer til styring af pumpe og ventiler
9. Skrueterminal til tilslutning af diverse ledninger, inkl. loddeplade med spændingsdeler til monitering af batterispænding

10. Magnetventil til at lukke luft ud af kammeret
11. Membranpumpe til at pumpe luft ind i kammeret

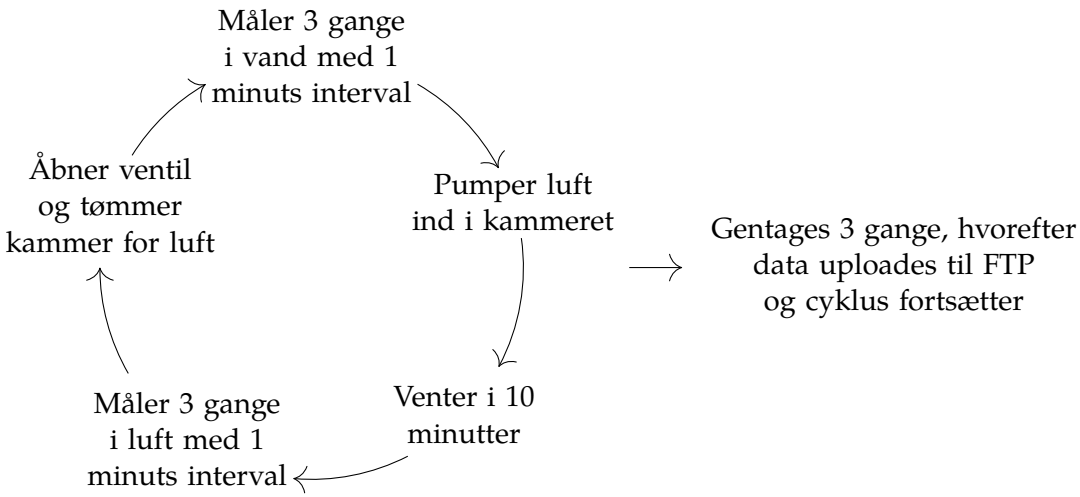
En oversigt over komponenterne fremgår af figur 6.8.



Figur 6.8. Oversigt over elektroniske komponenter i målestation. Tal referer til ovenstående forklaring.

Alle elektroniske komponenter i målestationen er styret af en Arduino MEGA, der er en programmerbar microcontroller. Microcontrolleren fungerer ved at sende og modtage analoge og digitale signaler til og fra de øvrige komponenter. Softwaren til microcontrolleren består af to dele; en opstarts del, der kører en gang ved opstart af microcontrolleren og initialiserer nødvendige indstillinger, og en cyklus der herefter kører kontinuert.

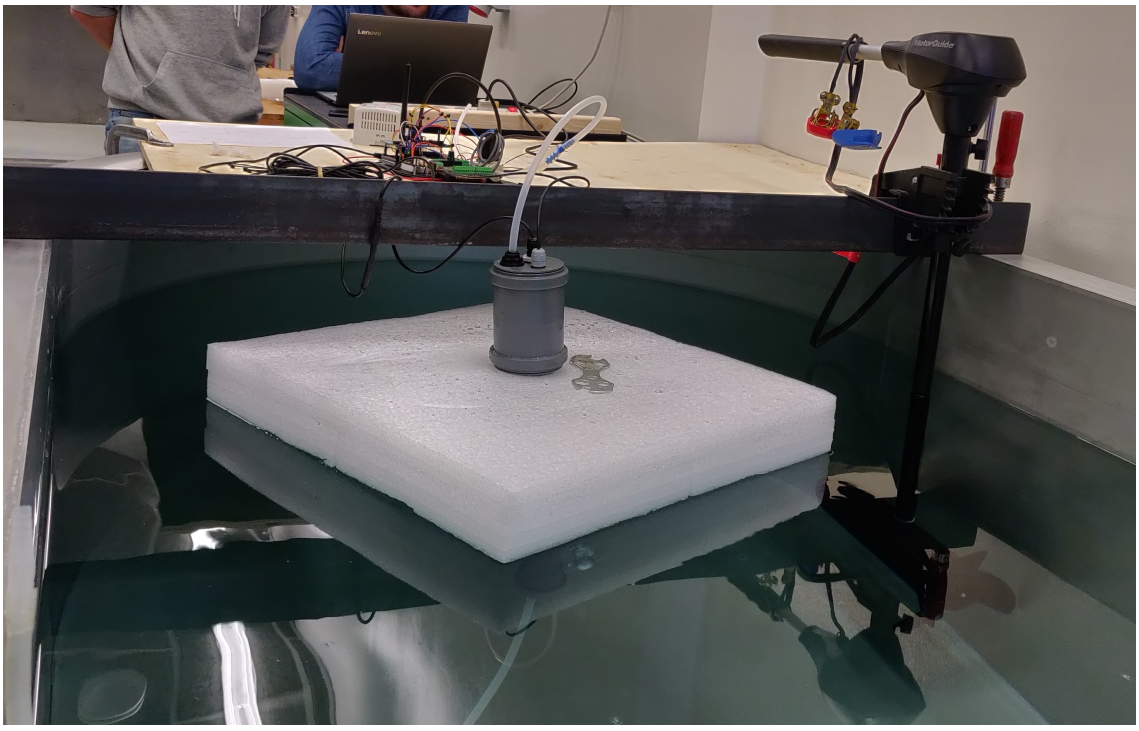
Cyklussen består af fem delelementer; måling af ledningsevne og temperatur i vandet, herefter fortrænges vandet i kammeret ved at aktivere luftpumpen, dernæst er der ti minutters pause, hvorefter ledningsevnen og temperaturen måles i luften, til sidst åbnes ventilen, så der igen kommer vand i kammeret. Elementerne i cyklussen fremgår af figur 6.9. Når denne cyklus har kørt tre gange, uploades en tekst-fil med de loggede data fra SD-kortet til en FTP-server, og cyklussen starter forfra. Årsagen til at der også logges data, når sensorerne er ude af vandet, er for at kunne kontrollere, at der ikke er vand i kammeret og det dermed er tæt og har den ønskede effekt.



Figur 6.9. Cyklus til måling, logning og upload af data fra målestation.

6.1.3 Test af målestation

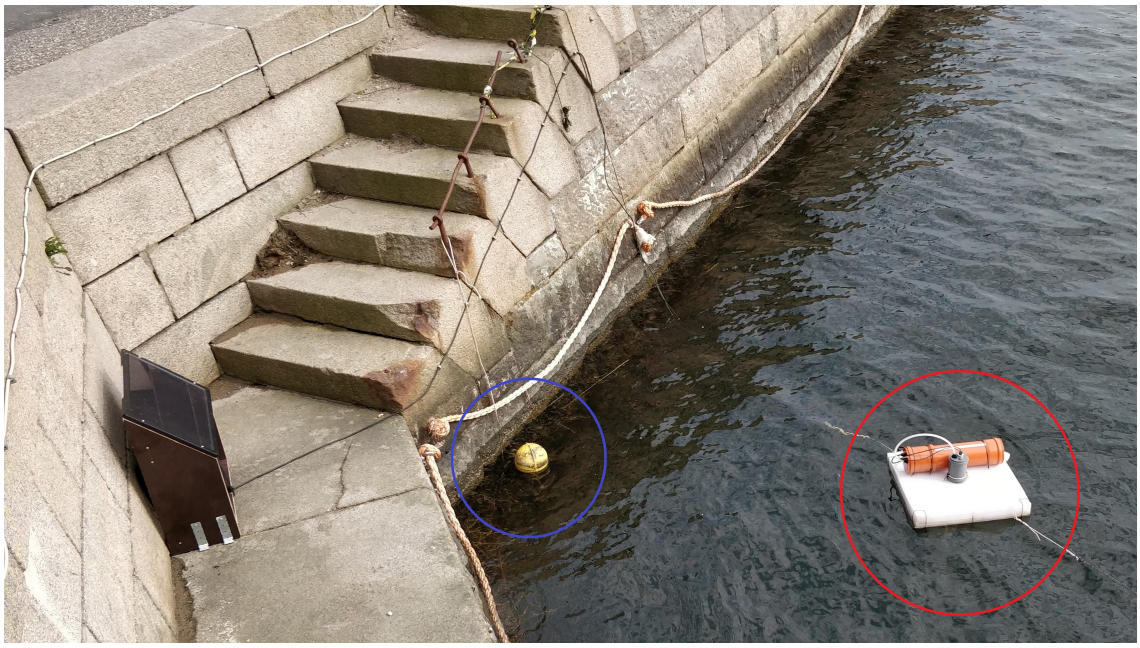
Målestationen er først testet i et kontrolleret miljø i laboratoriet, som det fremgår af figur 6.10. Her var det primære formål at teste konceptet, samt at fastslå den tid, det tager at fylde og tømme cylinderen for luft. Desuden er stabiliteten af målestationen blevet testet, hvilket resulterede i, at der blev sat jernlodder i bunden af cylinderen for at sikre en stabil målestation.



Figur 6.10. Første test af målestation i kar.

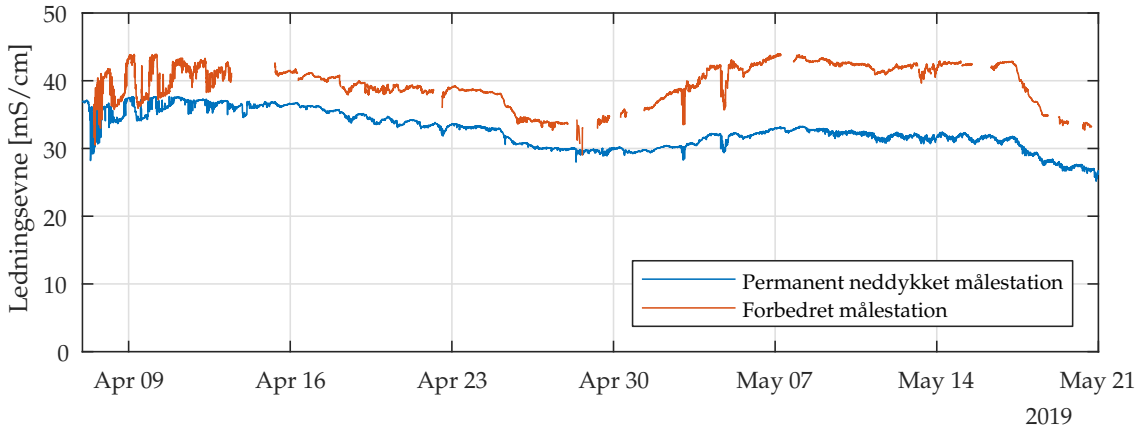
Målestationen er herefter testet i et naturligt miljø i Østre Havn i Aalborg i perioden start april til slut maj 2019. Målestationen er opsat sammen med den gamle type af

målestationer, således at måleresultater og algevækst på sensoren kan sammenlignes. På figur 6.11 fremgår målestationerne.



Figur 6.11. Forbedret målestation med begrænset algevækst (rød cirkel). Original målestation (blå cirkel).

Af figur 6.12 fremgår måleresultaterne for de to forskellige målestationer.



Figur 6.12. Ledningsevne i Østre havn målt med to forskellige målestationer.

Af måleresultaterne på figur 6.12 fremgår det, at der er en væsentlig afvigelse mellem målestationerne på ca. 5-10 mS/cm. Dette kan skyldes, at sensorerne ikke er interkalibreret, at der ikke måles i præcis samme dybde eller placering, eller at den permanent neddykkede målestation gror til. Dog viser måleresultaterne, at variationen i ledningevnen er ens for de to målestationer, og derfor vurderes den forbedrede målestation at kunne måle ledningsevnen.

Algevækst på sensoren i de to målestationer efter tre og syv ugers målinger fremgår af figur 6.13.

Figur 6.13. Algevækst på ledningsevne-sensoren ved de to forskellige målestationer. Øverst fremgår sensoren i den forbedrede målestation til venstre og den konstant neddykkede sensoren til højre efter tre uger. Nederst fremgår de to sensorer efter syv uger.

Det fremgår af figur 6.2, at algevæksten på sensoren er væsentligt reduceret ved at benytte den forbedrede målestation. Dette kan skyldes, at sensoren holdes ude af vandet mellem målingerne, og at det derfor er vanskeligt for alger at gro på sensoren. Samtidig er sensoren aldrig eksponeret for sollys, hvilket også gør det vanskeligt for alger, at kunne gro. Den forbedrede målestation kan derfor være en god metode til at undgå algevækst og sikre kvalitetsdata til varsling af badevandskvalitet samtidig med, at driftsomkostningerne kan reduceres.

6.2 Valg af model til implementering

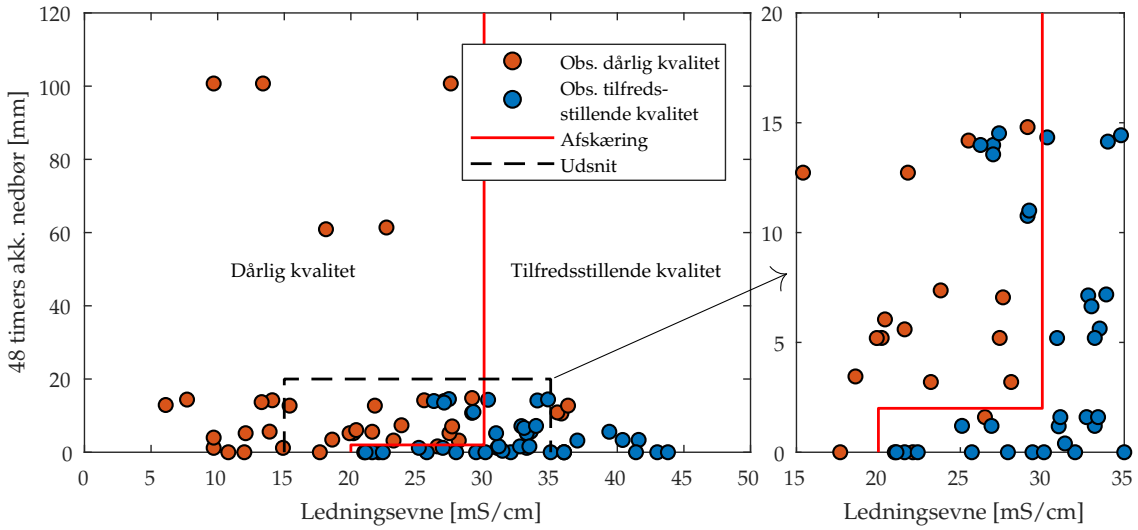
Igennem dette projekt er der udarbejdet forskellige modeller til varsling af badevandskvalitet. Modellerne varierer i, hvor avancerede de er, samt hvor stor en del af modelresultaterne der har fysiske forklaringer. Når der skal vælges en model til implementeringen, kan der lægges vægt på flere ting, herunder nøjagtigheden, den resulterende lukketid samt at den er simpel og nem at forklare. Fordelen ved en simpel model er, at den er sværere at overfitte til datasættet, samt at den er nem at opdatere, når der kommer nye datapunkter i datasættet. Desuden er en simpel model mindre sensitiv over for outliers.

Modellen med den højeste hitrate og laveste lukketid er KNN-modellen, der fremgår af figur 5.11. Det fremgår dog, at der er nogle områder, der strider imod den fysiske

forståelse af systemet. For at undgå disse områder udarbejdes der en simpel model der tager udgangspunkt i resultaterne fra machine learning-modellerne fra kapitel 5. Modellerne viser en sammenhæng mellem nedbør, ledningsevne og badevandskvalitet, der benyttes som udgangspunkt for at opbygge den simple model.

6.2.1 Forsimplet model

Undersøgelserne af KNN-algoritmerne viser, at nedbør har stor betydning for modellens performance i intervallet, hvor ledningsevnen er mellem 20 og 30 mS/cm, da den her kan hjælpe med at skelne datapunkterne. Dette interval har hidtil været det sværeste at beskrive. Til opbygning af en forsimplet model tages der udgangspunkt i formen på modelresultatet på figur 5.11, der viser responsfladen for KNN-modellen. På baggrund af dette opbygges en model med følgende betingelser: Ved en ledningsevne over 30 mS/cm klassificeres badevandet som tilfredsstillende kvalitet. Ved en ledningsevne mellem 20 mS/cm og 30 mS/cm klassificeres badevandskvaliteten kun som tilfredsstillende, hvis der er faldet under 2 mm nedbør de seneste to døgn. Ved en ledningsevne under 20 mS/cm klassificeres badevandskvaliteten altid som dårlig. En illustration af denne model fremgår af figur 6.14.



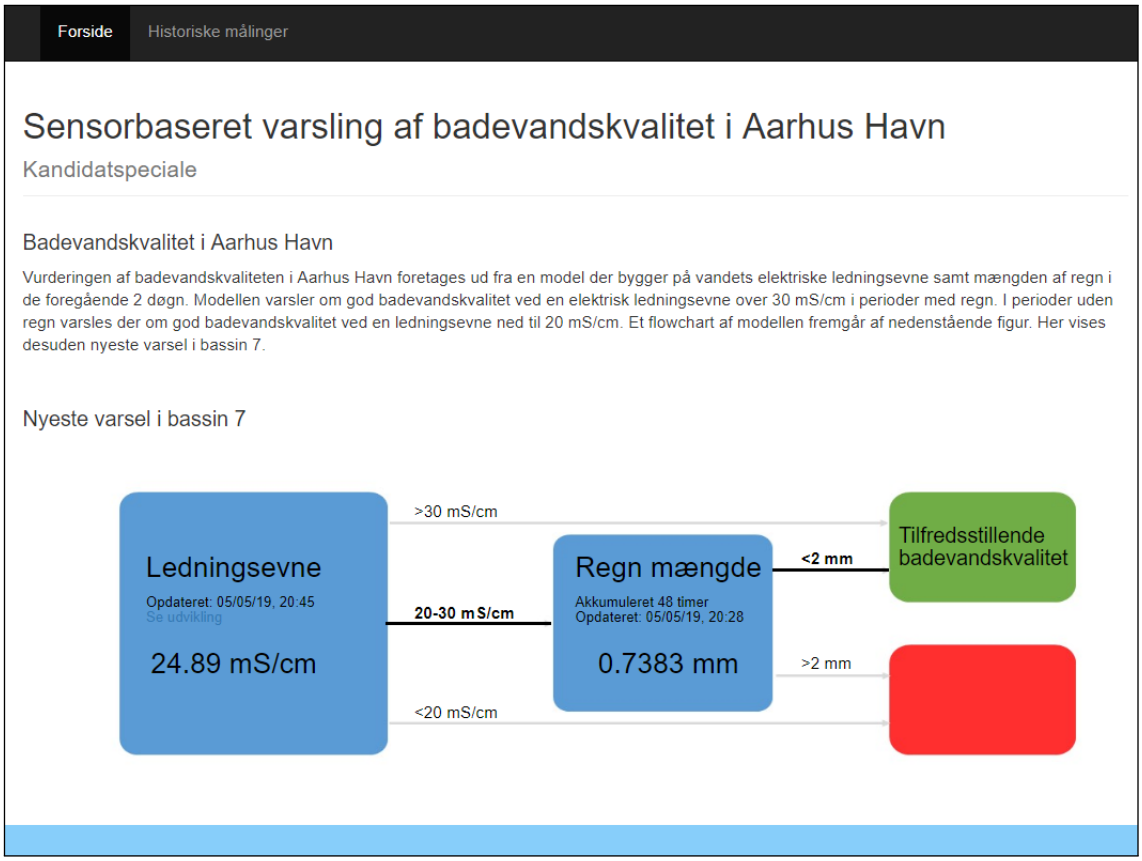
Figur 6.14. Forsimplet model til bestemmelse af badevandsklassifikation.

Den forsimplede model viser en hitrate på 86 % og er derfor stort set lige så nøjagtig som machine learning modellerne. Lukketiden for sommeren 2018 er beregnet til at være 12 dage, hvilket er 2-3 dage mindre end de to bedste machine learning-modeller og 14 dage mindre end den eksponentielle regression. Den simple model viser dermed en høj hitrate, lav lukketid, er simpel og har en fysisk forklaring og er derfor et oplagt bud til at implementere i Aarhus Havn.

6.3 Hjemmeside til varsling af badevandskvalitet

For fuldt at implementere varslingssystemet skal der vægles en metode til varsling af brugerne. Varsling af brugerne om badevandskvaliteten kan ske på flere medier

som eksempelvis sms, e-mail, varslingstavler eller via en hjemmeside. I dette projekt er der valgt at udarbejde en hjemmeside til varslingen. Hjemmesiden kan tilgåes via webadressen: `http://www.badihavnen.dk`. Et stilbillede af hjemmesiden fremgår af figur 6.15.



Figur 6.15. Hjemmeside til varsling af badevandskvalitet.

Ved udarbejdelse af hjemmesiden er der lagt vægt på, at den er simpel og overskuelig. Samtidig skal brugerne have mulighed for at se de parametre, der ligger til grund for varslingen. Dette er gjort ved at opstille modellen som et flowdiagram, der giver en bedre forståelse for varslingen, så brugerne dermed kan benytte badeområdet på et oplyst grundlag.

6.4 Opsummering

Ved udarbejdelse af delelementerne til implementering af varslingssystemet er der lagt vægt på, at systemet skal være simpelt og kræve så lidt vedligehold som muligt. Dette sikrer, at omkostningerne holdes nede samt at flest mulige kan forstå varslingssystemet. Dette er sikret ved at udvikle en ny type målestation. Målestationen bruger trykluft til at fortrænge vandet ud af et kammer hvori sensoren placeres. Målestationen har vist sig at kunne måle variationerne i ledningsevnen og samtidig have en meget begrænset algevækst på sensoren. Det er desuden sikret at systemet er enkelt ved, på baggrund af machine-learning modellerne, at udarbejde en simpel model, der er nem at vedligeholde og forstå.

Kapitel 7

KONKLUSION

Der er i dette projekt udarbejdet en løsning på problemet vedrørende varsling af badevandskvalitet i Aarhus Havn ud fra nedstående problemformulering:

Hvordan kan badevandskvaliteten i Aarhus Havn varsles på baggrund af sensordata, der kan måles i realtid?

Det er konkluderet, at den primære kilde til fækal forurening i Aarhus Havn er Aarhus Å, der er stærkt påvirket af udledninger af opspædet og renset spildevand fra henholdsvis fælleskloak og renseanlæg. Dette medfører, at der er en høj korrelation mellem koncentrationen af fækale bakterier og den elektriske ledningsevne i havnen, da ledningsevnen kan indikere opblandingsgraden mellem det ferske vand fra Aarhus Å og det salte havvand fra Aarhus Bugt. Sammenhængen mellem E. Coli og ledningsevne kan beskrives med en eksponentiel model. Denne model kan klassificere 82 % af alle de observerede badevandskvaliteter korrekt.

Ved anvendelse af flere parametre til varsling af badevandskvaliteten konkluderes det, at den bedste kombination af parametre er ledningsevne og nedbør. Ved anvendelse af disse to input til en klassifikationsmodel, der bygger på nærmeste nabo-algoritmen, kan der opnås en hitrate på 89 %. Det er vurderet, at forklaringen på forbedringen ikke er, at der benyttes andre algoritmer, men kombinationen af parametrene. Nedbør har en særlig indvirkning på modelresultatet i tilfælde, hvor ledningsevnen er mellem 20 og 30 mS/cm. Dette skyldes, at en opblanding mellem åvand og havvand, der er ækvivalent til en ledningsevne på mellem 20 og 30 mS/cm, ikke i sig selv er nok til at medføre dårlig badevandskvalitet. Tilgengæld vil denne opblandingsgrad medføre dårlig badevandskvalitet, når det har regnet, da koncentrationen i åvandet dermed vil være højere.

Ud over forbedret performance i varsling af badevandskvalitet har modellen, der inddrager nedbør, den effekt, at lukketiden bliver reduceret markant. For sommeren 2018 vil der ske en reduktion fra en akkumuleret lukketid på 26 dage til under 15 dage, hvis nedbør inkluderes i varslingsmodellen.

For at kunne monitere ledningsevnen i realtid har det vist sig nødvendigt at opbygge et system, der kan afhjælpe problemer med forekomst af algevækst på sensoren. Der er i dette projekt udviklet et system, hvor sensoren placeres i et kammer, der mellem hver måling tømmes for vand ved brug af trykluft. Den udviklede målestation har vist sig at være meget effektiv i forhold til at undgå algevækst på sensoren.

Aarhus Havn er et badeområde der er eksponeret for en stor risiko for fækal forurening. Dette er et problem, som på kort sigt ikke kan løses. Derfor har sensorbaseret varsling af badevandskvalitet vist sig at være en løsning, der er særdeles brugbar til varsling af badevandskvalitet, hvilket kan medvirke til, at Aarhus Havn alligevel kan blive et attraktivt og ikke mindst sikkert sted at bade.

Kapitel 8

PERSPEKTIVERING: ER METODEN GENEREL?

I dette kapitel undersøges det, hvorvidt metoden til opbygning af varslingsmodel på baggrund af ledningsevne og nedbør kan generaliseres og benyttes i andre systemer.

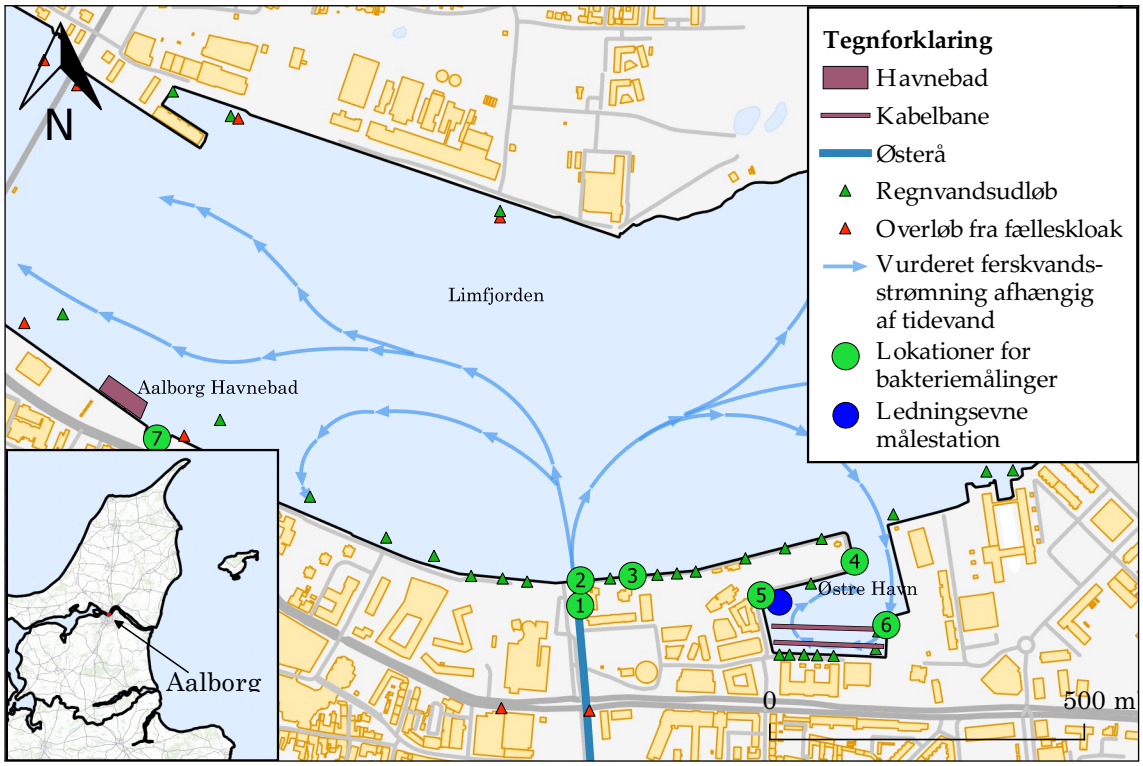
Det er i denne rapport konkluderet, at sensorbaseret varsling af badevandskvalitet, på baggrund af ledningsevne og nedbør, er en særdeles velegnet metode til systemet i Aarhus Havn. Det ønskes at undersøge, hvorvidt nedbør og ledningsevnen kan benyttes til at varsle om forringet badevandskvalitet andre steder end blot i Aarhus Havn. Til dette formål er havnefronten i Aalborg udvalgt til undersøgelsesområde for en generalisering af varslingsmetoden.

Undersøgelsens hypotese er at der er en sammenhæng mellem E. Coli-koncentrationen, vandets ledningsevne samt mængden af nedbør. Hypotesen bygger på, at det, ligesom i Aarhus, er ferskvandet, der er forurenet, og at bakteriekoncentrationen derfor blandt andet afhænger af andelen af ferskvand, hvilket ledningsevnen kan indikere.

Hypotesen vil blive undersøgt ved at indsamle data i form af bakteriekoncetration og ledningsevne ved Aalborg Havnefront.

8.1 Undersøgelsesområde

Det er valgt at benytte Aalborg Havnefront som område til at undersøge, hvorvidt varslingsmetoden er generel. Ligesom Aarhus Havn har Aalborg Havnefront gennemgået en transformation fra industrihavn til et rekreativt element i byens centrum. Langs Havnefronten er Aalborg Havnebad placeret, og mod øst ligger Østre Havn, hvor Aalborg Kabelpark er beliggende. Af figur 8.1 fremgår en oversigt over Aalborg Havnefront. Det fremgår desuden, at der langs hele havnefronten er en række regnvandsudløb og enkelte overløb fra fælleskloak. Mellem Østre Havn og havnebadet har Østerå sit udløb.



Figur 8.1. Oversigt over Aalborg Havn omkring Østerås udløb i Limfjorden.

For Aalborg Havnebad er de primære kilder til forurening de to overløb fra fælleskloak, der er placeret på hver side af havnebadet. I området omkring Østre Havn er der derimod ikke nogen overløbsbygværker, og kilderne til eventuel forurening forventes derfor at være fra Østerå, der er påvirket af fællesvand fra overløbsbygværker fra en stor del af Aalborg. Modsat i Aarhus er der dog ingen permanente udløb, fra eksempelvis renseanlæg, hvorfor det kun forventes at være høje koncentrationer af fækale bakterier i forbindelse med regnhændelser.

Strømningerne i Limfjorden er anderledes end strømningerne i Aarhus Havn. Den primære strømningsmekanisme i Limfjorden er tidevandet, der presser vandet skiftevis fra øst mod vest og omvendt. Ferskvandsstrømmen fra Østerå forventes derfor ligeledes at skifte fra øst- til vestgående sammen med tidevandet. På baggrund af erfaringerne fra Aarhus Havn forventes det, at ferskvandsstrømmen i overfladen ligeledes vil være påvirket af vindretningen.

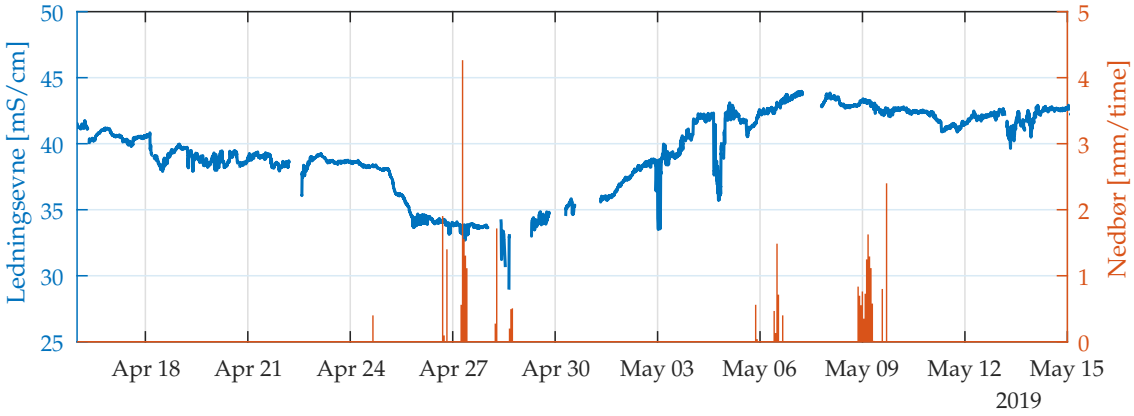
Aalborg Universitet har tidligere udarbejdet en varslingsmodel for badevandskvaliteten i Aalborg Havnebad. Modellen er en enkelt opbygget lineær reservoir-model med afløbssystemet som ét reservoir og havnen som ét reservoir, der modelleres som fuldt opblandet med gennemstrømning. I havne-reservoiret modelleres henfald og fortynding som 1. ordens proces. Modellens eneste input er nedbør, hvilket giver god mening for Aalborg Havnebad, da det uanset strømningsretning vil blive påvirket af aflastninger grundet havnebadets placering mellem to udløb. [M. R. Rasmussen, 2011]

For Aalborg Kabelpark i Østre Havn er der ingen overløbsbygværker fra fælleskloakerede områder, hvorfor Østerå forventes at være den primære kilde til eventuel dårlig

badevandskvalitet i Østre Havn. Dette betyder, at nedbør i sig selv ikke er tilstrækkeligt for at vurdere, hvorvidt der er dårlig badevandskvalitet, i det at ferskvandsstrømningens retning også vil have betydning. Dette giver anledning til at undersøge, hvorvidt ledningsevnen kan benyttes som indikator for forringet badevandskvalitet i kombination med nedbør, da ledningsevnen kan indikere hvor meget ferskvand, der strømmer fra Østerå og ind i Østre Havn.

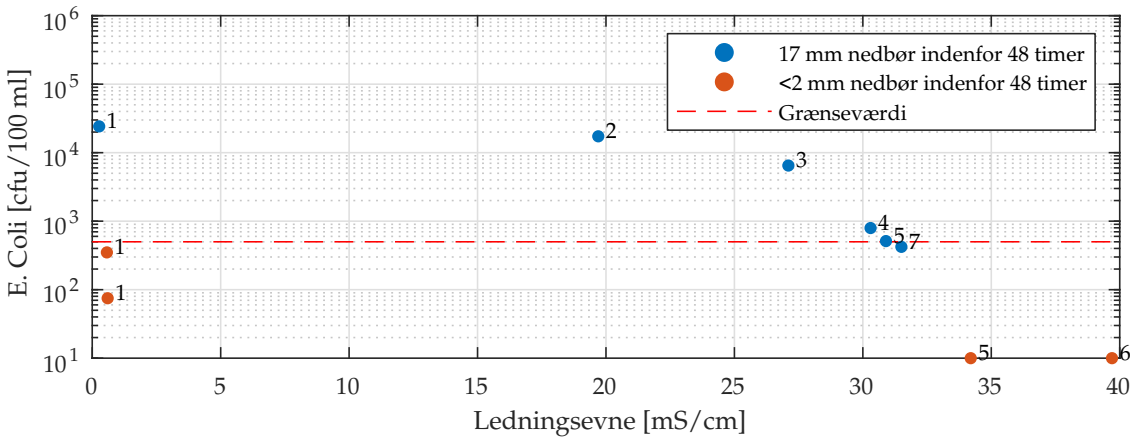
8.2 Dataindsamling i Østre Havn og ved Aalborg Havnefront

Til at undersøge hvorvidt ledningsevne og nedbør kan indikere, om der er en forhøjet risiko for dårlig badevandskvalitet, ønskes det at indsamle data, der repræsenterer et spænd af ledningsevneværdier ved både regnhændelser og ved tørvejr. For at få et datasæt, der repræsenterer både høje og lave ledningsevneværdier, opstilles derfor en online ledningsevnemåler, så bakteriemålingerne kan foretages ved ønskede værdier. Til dette formål benyttes den nyudviklede ledningsevnemåler, beskrevet i afsnit 6.1. Som det fremgår af figur 8.2, er der registreret nogle dyk i ledningsevnen, blandt andet 3. og 5. maj. Det fremgår desuden, at der i denne periode ikke er registreret noget nedbør. Det forventes derfor, at disse dyk skyldes, at ferskvand fra Østerå strømmer ind i Østre Havn og dermed kan ledningsevnen muligvis benyttes til at indikere mængden af åvand.



Figur 8.2. Variation i ledningsevne i Østre Havn, samt nedbør. Placering af måler fremgår af figur 8.1. Nedbør stammer fra Spildevandskomiteens regnmålersystem [2019], målestation 5052

I perioden marts til juni 2019 er der indsamlet 10 vandprøver ved udløbet fra Østerå, langs Aalborg Havnefront og i Østre Havn. Lokationerne for målingerne fremgår af figur 8.1. Vandprøverne er analyseret for E. Coli, jævnfør analysemetoden beskrevet i appendiks H, derudover er der målt ledningsevne. Resultaterne fra analyserne fremgår af figur 8.3 og af det digitale Bilag 2.



Figur 8.3. Sammenhæng mellem ledningsevne, nedbør indenfor seneste 48 timer og E. Coli-koncentration. Nummerering i henhold til lokationer angivet på figur 8.1. For det blå målepunkt 1 er der tale om en minimumskoncentration, da bakterieanalysen viste den maksimale målbare værdi ved den valgte fortynding.

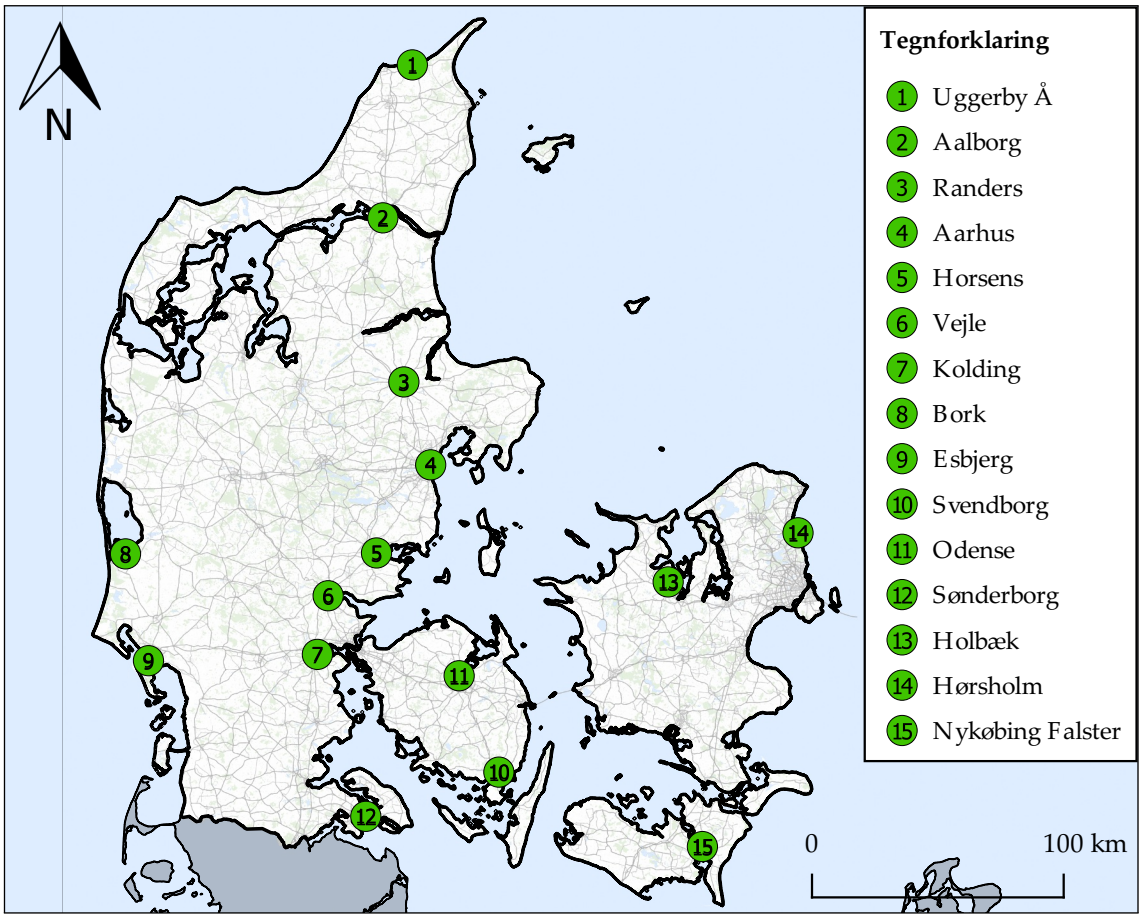
Af figur 8.3 fremgår det, at der i tilfælde, hvor der er registreret under 2 millimeter nedbør indenfor de foregående 48 timer, ikke er observeret bakteriekoncentrationer, der medfører overskridelse af badevandskriteriet. Både ved høje ledningsevner, i Østre Havn, og lave ledningsevner i Østerå, er der målt under 500 cfu/100 ml. I forbindelse med kraftig nedbør er det observeret, at det forekommer meget høje koncentrationer af E. Coli i Østerå. I takt med faldende ledningsevne langs havnefronten falder også bakteriekoncentrationen. Dette indikerer, at ledningsevnen kan benyttes til at vurdere opblandingsgraden mellem ferskvand og havvand, og nedbør kan indikere hvor høj bakteriekoncentrationen er i ferskvandet.

For at kunne modellere badevandskvaliteten i Aalborg, på baggrund af ledningsevne og nedbør, er det nødvendigt med væsentligt flere datapunkter til at dække et bredt spektrum af situationer, blandt andet flere forskellige nedbørsmænger, ledningsevner, men også ved forskellige tidevands og vindforhold.

8.3 Potentiale for udbredelse af sensorbaseret varsling

Det er i projektet vist, at der for Aarhus Havn kan opbygges en model af badevandskvaliteten baseret på vandets ledningsevne og nedbøren i den foregående periode. Modellen er specifik for Aarhus Havn, da den bygger på data, der stammer fra lokationen. Selvom modellen er lokations-specifik, kan metoden med at varsle på baggrund af ledningsevne og nedbør muligvis benyttes i andre systemer. Undersøgelserne af sammenhængen i Aalborg viser, at nedbøren her er af meget større betydning end i Aarhus, hvilket skyldes, at baggrundskoncentration af bakterier i Aalborg er meget lavere. Dog har ledningsevnen i Aalborg vist sig at have en god sammenhæng med bakteriekoncentrationen i nedbørsrige perioder. Det vurderes på denne baggrund, at badevandskvaliteten i områder, der er ferskvandspåvirket, og hvor bakteriekilder kan kobles til ferskvandsudledningen, kan varsles ved at finde den rigtige vægtning mellem nedbør og ledningsevne.

Af figur 8.4 fremgår et udpluk af en række lokationer i hele Danmark, hvor varsling på baggrund af ledningsevne og nedbør vurderes at være mulig. Områderne er alle i nærheden af et udløb fra en ferskvandskilde, der er påvirket af udledninger fra renseanlæg eller fælleskloak. Lokationerne er både kystnære områder, samt bynære områder, hvor det kan blive aktuelt at anlægge et havnebad.



Figur 8.4. Oversigt over udvalgte lokationer med potentiale for sensorbaseret varsling af badevandskvalitet.

Dette projekt viser, hvordan det er muligt at benytte sensorbaserede løsninger, til at vurdere badevandskvaliteten på lokationer, hvor ferskvand og saltvand opblandes. Løsningen anbefales blandt andet til badeområder der grundet manglende varsling er klassificeret som ringe. Løsningen kan også åbne op for at mange fleres steder, med midlertidige forureninger, kan benyttes som badevandsområde, da sensorbaseret varsling i mange tilfælde kan benyttes som et godt alternativ til dyre deterministiske modeller.

LITTERATUR

Aarhus Kommune, 2016, *Spildevandsplan 2017-2020*. URL: https://aarhus.dk/media/4969/spildevandsplan-2017-2020.pdf.

Aarhus Kommune, 2018a. Data udleveret fra Aarhus Kommune, Teknik og Miljø.

Aarhus Kommune, 2018b, *Aarhus Ø - et levende centrum for byliv*. URL: https://aarhusoe.dk/.

Aarhus Kommune, 2019, *Vil du på stranden?* URL: https://bit.ly/2NvnXNX.

Aarhus Vand, 2018, *Renere vand i havn, å og søer*. URL: https://www.aarhusvand.dk/projekter/spildevand/overordnede-planer/soerne-aen-og-havnen/.

Allegue, Laura Bailón, 2017, *Kemisk hygiejnisering på Viby og Åby Renseanlæg*. Udleveret af Aarhus Kommune, Teknik og Miljø.

Atlas Scientific, 2019, *E.C. Probe K 1.0*. URL: https://bit.ly/2Dajxra.

COWI, 2017, *Modellering af strømningsforhold i Bassin 7, Aarhus Havn*. Udleveret af Aarhus Kommune, Teknik og Miljø.

Dalhoff, Marie, 2017, „Havnebad i Aarhus: Byrådet accepterer ekstraregning på 18 millioner kroner". I: *Lokalavisen Aarhus*.

Danmarks Miljøportal, 2018, *Danmarks Miljøportal*. URL: www.miljoeportal.dk.

De Brauwere, Anouk, Nouho Koffi Ouattara og Pierre Servais, 2014, „Modeling fecal indicator bacteria concentrations in natural surface waters: A review". I: *Critical Reviews in Environmental Science and Technology* 44, s. 2380–2453.

DHI, 2019, *Badevandsudsigten*. URL: http://badevand.dk/.

Europa-parlamentet og rådet, 2006, *Direktiv 2006/7/EF af 15. februar 2006 om forvaltning af badevandskvalitet*. Det Europæiske Unions Tidende.

Greenberg, Arnold E., 1956, „Survival of Enteric Organisms in Sea Water: A Review of the Literature". I: *Public Health Reports* 71, s. 77–86. DOI: 10.2307/4589366.

Hao, Ning, Bin Dong og Jianqing Fan, 2015, „Sparsifying the Fisher linear discriminant by rotation". I: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77, s. 827–851. DOI: 10.1111/rssb.12092.

Hayashi, Masaki, 2004, „Temperature-Electrical Conductivity Relation of Water for Environmental Monitoring and Geophysical Data Inversion". I: *Environmental monitoring and assessment* 96, s. 119–28. DOI: 10.1023/B:EMAS.0000031719.83065.68.

Hydrometri, 2018. Data aflæst fra www.hydrometri.dk, for Aarhus Vands målestation nr. 26.20.

IDEXX, 2019a, *Colilert*. URL: https://www.idexx.com/files/colilert-procedure-en.pdf.

IDEXX, 2019b, *IDEXX Quanti-Tray/2000 MPN Table*. URL: https://www.idexx.com/files/qt97mpntable.pdf.

ISO, 1999, *ISO 9308-3:1999*. International Organization for Standardization.

Kay, D. m.fl., 1994, „Predicting likelihood of gastroenteritis from sea bathing: results from randomised exposure". I: *The Lancet* 344, s. 905–909.

Kotsiantis, S. B., 2007, „Supervised Machine Learning: A Review of Classification Techniques". I: *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, s. 3–24. ISBN: 978-1-58603-780-2.

Krause, P., D. P. Boyle og F. Bäse, 2005, „Comparison of different efficiency criteria for hydrological model assessment". I: *Advances in Geosciences* 5, s. 89–97. DOI: 10.5194/adgeo-5-89-2005.

Kulturarvsstyrelsen, 2009, *Brug havnen - Industrihavnens kulturarv*. ISBN: 978-87-91298-09-7.

Luna, Gian Marco m.fl., 2018, „Status of faecal pollution in ports: A basin-wide investigation in the Adriatic Sea". I: *Marine Pollution Bulletin*. DOI: 10.1016/j.marpolbul.2018.03.050.

MathWorks, 2019, *Support Vector Machines for Binary Classification*. URL: https://bit.ly/2WRhXTE.

Miljø- og Fødevareministeriet, 2016, *Bekendtgørelse nr 917 af 27/06/2016 om badevand og badeområder*.

Miljø- og Fødevareministeriet, 2018, *Bekendtgørelse nr 1146 af 24/10/2017 om kvalitetskrav til miljømålinger*.

Miljøministieret, 2010, „Forslag til Vandplan. Hovedopland 1.7 Århus Bugt, bilag 3". I: 978-87-92668-48-6.

Miljøstyrelsen, 2019, *MiljøGIS for Vandområdeplanerne 2015-2021*. URL: http://miljoegis.mim.dk/cbkort?profile=vandrammedirektiv2-2016.

Naganna, Sujay Raghavendra og Paresh Chandra Deka, 2014, „Support vector machine applications in the field of hydrology: A review". I: *Applied Soft Computing* 19, s. 372–386. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2014.02.002.

Ngufor, Che og Janusz Wojtusiak, 2016, „Extreme logistic regression". I: *Advances in Data Analysis and Classification* 10.1, s. 27–52. ISSN: 1862-5355. DOI: 10.1007/s11634-014-0194-2.

Nielsen, Jesper m.fl., 2018, „Can the Volume Ratio of Coarse to Fine Particles Explain the Hydraulic Properties of Sandy Soil?" I: *Soil Science Society of America Journal* 82. DOI: 10.2136/sssaj2018.02.0083.

Nordea Fonden, 2018, *Aktive unge på vandet*. URL: https://nordeafonden.dk/projekter/aktive-unge-paa-vandet.

Oblinger, J. L. og J. A. Koburger, 1975, „Understanding and Teaching the Most Probable Number Technique“. I: *Journal of Milk and Food Technology* 38. DOI: 10.4315/0022-2747-38.9.540.

Onset Computer Corporation, 2019, *Onset HOBO*. URL: https://bit.ly/2SuYx85.

OTT HydroMet, 2019, *OTT ecoLog 800*. URL: https://bit.ly/2SefWCV.

Rasmussen, Carl Edward og Christopher K. I. Williams, 2005, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press. ISBN: 026218253X.

Rasmussen, Michael R., 2011, *Erfaring med at forudsige badevandskvaliteten for Aalborg Havnebad*. Slideshow fra Dansk Vand Konference 2011.

Russel, Stuart J. og Peter Norvig, 2010, *Artificial Intelligence A Mordern Approach*, 3. udgave. ISBN: 978-0-13-604259-4.

Sorensen, John A. og Gary E. Glass, 1987, „Ion and Temperature Dependence of Electrical Conductance for Natural Waters“. I: *Analyt. Chem.* 59, s. 1594–1597. DOI: 10.1021/ac00140a003.

Spildevandskomitéen, 1980, *Spildevandskomiténs skrift nr. 17: Spildevandskomitéens regnmålersystem*.

Spildevandskomiteens regnmålersystem, 2019. URL: http://svk.dmi.dk/.

Styrelsen for Dataforsyning og Effektivisering, 2019, *Kortforsyningen*. URL: www.kortforsyningen.dk.

The Weather Company, 2018, *Weather Underground - Station: I82CHRIS9*. URL: https://www.wunderground.com/wundermap.

Webb, Andrew R. og Keith D. Copsey, 2011, *Statistical pattern recognition*. Wiley. ISBN: 9780470682272.

Wenner, Frank, Edward H. Smith og Floyd M. Soule, 1930, „Apparatus for the determination aboard ship of the Salinity of sea water by the electrical conductivity method“. I: *U.S. Bureau of Standards Journal of Research* 5, s. 711–732. DOI: 10.6028/jres.005.041.

WTW, 2019, *ProfiLine Multi 3320*. URL: https://bit.ly/2GdIWEu.

Appendiks A

OVERSIGT OVER DIGITALE BILAG

Bilag 1: MATLAB-script til træning og validering af tre modeller

I de digitale bilag er filen *"Traening_og_evaluering_af_modeller.m"* vedlagt. Filen er et MATLAB-script til træning og validering af tre forskellige modeller til modellering af badevandskvaliteten; en KNN-model, en GPR-model samt en eksponentiel model. I scriptet er der en statistisk tilgang til validering af modellerene. Derfor trænes og valideres modellerne *n* antal gange og fordelingen af deres performance plottes i et histogram. Til scriptet tilhører tre funktioner samt et datasæt. Disse er beskrevet herunder.

MATLAB-funktion til træning af GPR-model

I de digitale bilag er filen *"GPR_reg_2features.m"* vedlagt. Filen er en MATLAB-funktion til træning af en GPR-model med to input. Funktionen er eksporteret fra MATLAB-applikationen *Regression Learner*.

MATLAB-funktion til træning af KNN-model

I de digitale bilag er filen *"KNN_classi_2features.m"* vedlagt. Filen er en MATLAB-funktion til træning af en KNN-model med to input. Funktionen er eksporteret fra MATLAB-applikationen *Classification Learner*.

MATLAB-funktion til udregning af Nash–Sutcliffe koefficienten

I de digitale bilag er filen *"nse.m"* vedlagt. Filen er en MATLAB-funktionen til udregning af Nash–Sutcliffe koefficienten.

MATLAB-data til træning af modeller

I de digitale bilag er filen *"Data_features.mat"* vedlagt. Filen er en MATLAB datafil, der indeholder data fra de 303 observationer af bakteriekoncentrationen i Aarhus samt de dertilhørende værdier for nedbør, vind, lokation og ledningsevne.

Bilag 2: Resultat af E. Coli målinger

Af bilag 2 i de digitale bilag fremgår et Excel-ark med resultater fra de E. Coli målinger er lavet i dette projekt. Til hver måling er tidspunktet, lokationen, den korrigerede ledningsevne samt den akkumulerede nedbør inden for 2 døgn angivet.

Appendiks B

DATAOVERBLIK

Til analyser og udarbejdelse af modeller er der benyttet en række forskellige typer af hydrologiske og biologiske data. I dette appendiks gennemgås ophavet af de forskellige data. En oversigt over målepunkterne og klimastationernes placering fremgår af figur B.1.

Figur B.1. Placering af klimastationer til måling af nedbør, vandføring og vindforhold samt målepunkter til indsamling af vandprøver.

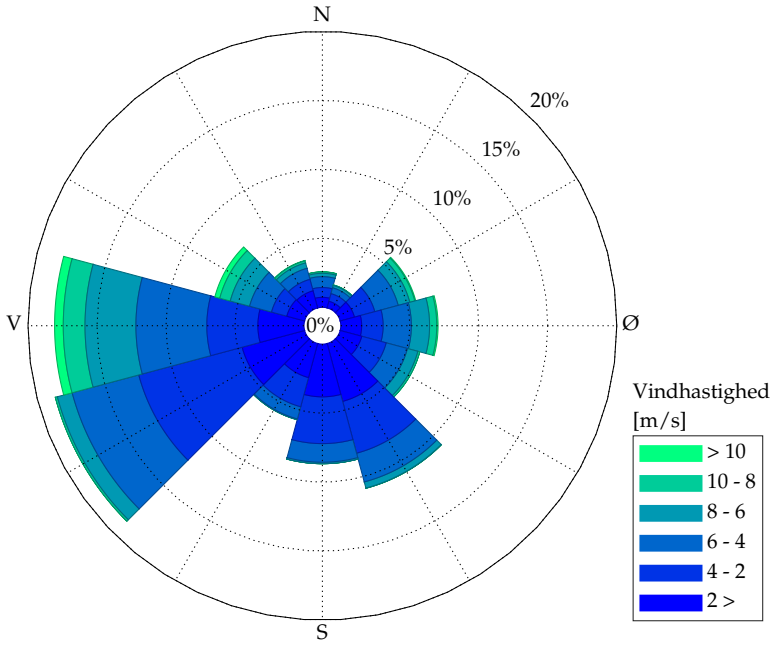
B.1 Nedbørsdata

Nedbørsdata anvendt i dette projekt stammer fra spildevandskommitéens landsdækkende regnmålersystem [Spildevandskomiteens regnmålersystem, 2019]. Det er valgt, at benytte måler nr. 5177, placeret ved Viby J. Renseanlæg, som det fremgår af figur B.1. Denne måler er benyttet, da det er den målestation, der har målt i hele perioden, hvor bakteriemålinger er indsamlet og er mest centralt placeret i Aarhus. Regnserien indeholder data fra 1979 og er stadig i drift. SVKs regnmålersystem består af en række vippekar, der måler nedbør ved at registrerer et vip, hver gang der falder 0,2 mm nedbør [Spildevandskomitéen, 1980].

B.2 Vinddata

Vinddata anvendt i dette projekt stammer fra *Weather Underground - Station: I82CHRIS9* [The Weather Company, 2018]. Weather Underground er et netværk, der samler data fra private klimastationer. På Weather Underground har alle mulighed for at offentliggøre sine klimaobservationer, hvilket også medfører en risiko for, at dataene

ikke er kvalitetssikrede og dermed usikre. *Weather Underground - Station: I82CHRIS9* er placeret, som det fremgår af figur B.1. Fordelingen af vindretninger og -hastigheder fremgår af figur B.2.



Figur B.2. Fordeling af vindretning og -hastigheder i Aarhus. Målestationens placering fremgår af figur B.1. Data fra The Weather Company [2018].

B.3 Vandføringsdata

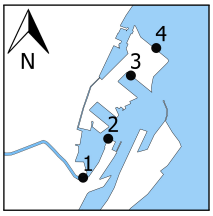
I dette projekt er der benyttet to typer af vandføringsdata; døgnmiddel-vandføring fra Miljøstyrelsen og kontinuerte målinger aflæst fra Hydrometri måler nr. 26.20. Tidsserien med døgnmiddel-vandføringer indeholder data fra 2007 til og med 2017 og er placeret i Aarhus midtby [Danmarks Miljøportal, 2018]. Data fra denne måler er benyttet til at undersøge korrelationen mellem bakteriekoncentrationer i Aarhus Havn og vandføringen.

Hydrometri måler nr. 26.20 er drevet af Aarhus Vand og er placeret ved Aarhus Ås udløb [Hydrometri, 2018]. Dataene er aflæst fra `www.hydrometri.dk` og er benyttet til at undersøge vandføringens påvirkning på ledningsevnen i Aarhus Havn, hvorfor en højere opløsning end døgnværdier er nødvendig.

B.4 Bakterieanalyser

Bakterieanalyser fra Aarhus Havn benyttet i dette projekt er udleveret af Aarhus Kommune, Teknik og Miljø [Aarhus Kommune, 2018a]. Målingerne er foretaget en række steder i havnen, der fremgår af figur B.1, i perioden maj 2012 til november 2018. Datasættet består af 333 observationer fordelt mellem målepunkter og hvorvidt, der er målt ledningsevne som vist i tabel B.3.

Målepunkt	Med ledningsevne	Uden ledningsevne
1: Mindet	52	34
2: Mellemarmen	23	18
3: Bassin 7	35	116
4: Havbanen	13	42



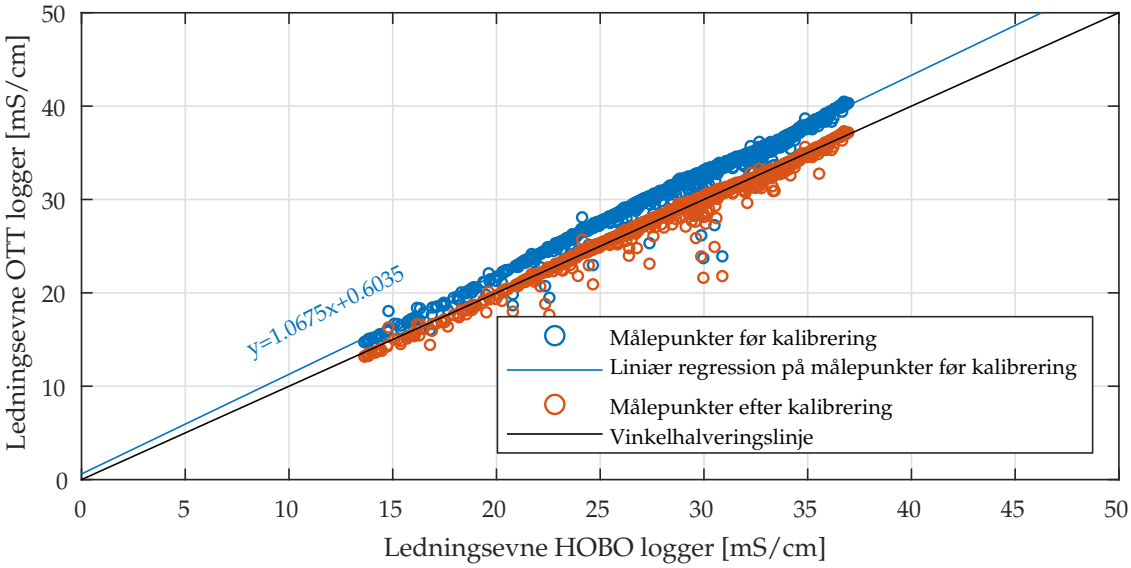
Figur B.3. Oversigt over fordelingen af 333 bakterieobservationer i Aarhus Havn fordelt på lokation og hvorvidt der er målt ledningsevne.

Appendiks C

INTERKALIBRERING AF ELEKTRISK LEDNINGSEVNEMÅLERE

I det følgende appendiks beskrives den kalibrering, der er foretaget på de målte ledningsevner. I projektet er der benyttet ledningsevnemålere af to forskellige fabrikater; OTT ecolog og Onset HOBO-logger. Disse målere er dermed muligvis kalibreret forskelligt og skal derfor interkalibreres. Til dette formål er en ledningsevnemåler af hvert fabrikat placeret med en meters afstand på samme lokation i Aarhus havn. Herefter sammenlignes målingerne, og OTT ecolog målingerne kan korrigeres lineært. Data fra OTT loggeren korrigeres, da der kun er benyttet én måler af dette fabrikat.

På figur C.1 fremgår en sammenligning mellem de målte elektriske ledningsevner for de to målere til samme tidspunkt.



Figur C.1. Sammenhæng mellem ledningsevnen for de to målere.

Som det fremgår af figur C.1 afviger data fra OTT måleren med 6,75 % og 0,6035 mS/cm fra Hobo måleren. På denne baggrund korrigeres data fra OTT måleren med samme.

Appendiks D

TEMPERATURKORRIGERING AF ELEKTRISK LEDNINGSEVNE

Den elektriske ledningsevne afhænger af temperaturen. For at kunne benytte ledningsevnen som et udtryk for opblandingsgraden mellem å- og havvand er det derfor nødvendigt at omskrive ledningsevnen til ledningsevnen ved en standardtemperatur. Dermed vil ændringer i den korrigerede ledningsevne ikke skyldes temperaturændringer. Ved store temperaturændringer er sammenhængen mellem ledningsevnen og temperatur ikke lineær, men ved normalt svingende havtemperatur kan sammenhængen antages at være lineær. Ledningsevnen kan temperaturkorrigeres lineært med følgende ligning [Sorensen og Glass, 1987]:

$$EC_{25} = \frac{EC_t}{1 + a \cdot (t - 25)} \tag{D.1}$$

Hvor:

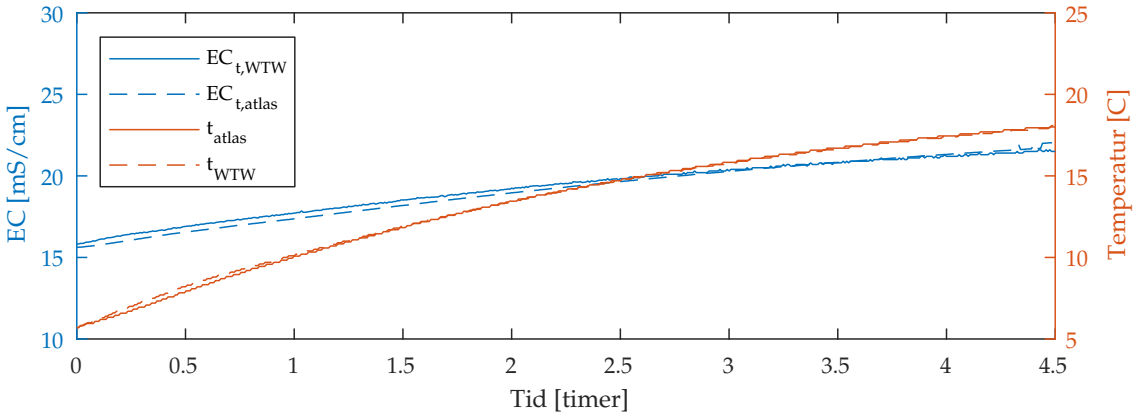
EC_{25}	Elektrisk ledningsevne ved 25 °C [$\frac{\mu S}{cm}$]
EC_t	Elektrisk ledningsevne ved temperaturen t [$\frac{\mu S}{cm}$]
a	Temperaturkoefficient [$°C^{-1}$]
t	Temperatur [°C]

Temperaturkoefficieten angives i litteraturen til cirka 0,02 pr. °C [Hayashi, 2004]. Da temperaturkoefficieten har stor betydning for resultatet af dataindsamlingen bestemmes denne eksperimentelt.

D.1 Eksperimentel bestemmelse af temperaturkoefficient

For at bestemme temperaturkoefficieten udføres et eksperiment, hvor ledningsevnen og temperaturen måles i en væske. Væsken tilsættes salt inden eksperimentet, således at EC_{25} er som i brakvand. Under eksperimentet varieres temperaturen mellem 5 og 20 °C. Temperaturen og ledningsevnen måles løbende under eksperimentet. Herefter korrigeres ledningsevnen til 25 °C ved at tilpasse temperaturkoefficieten, således at EC_{25} bliver konstant.

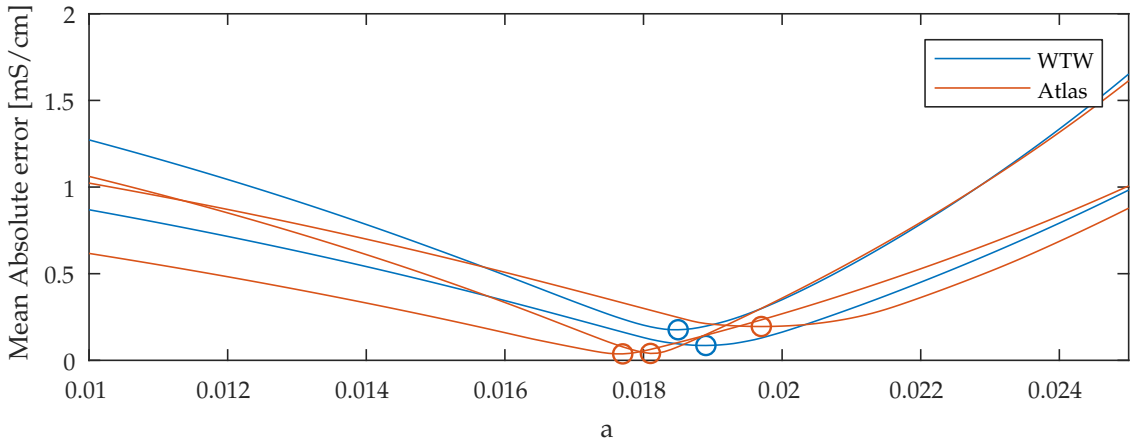
Eksperimentet udføres med to forskellige ledningsevne-sensorer; en sensor fra Atlas Scientific [2019] samt en sensor fra WTW [WTW, 2019]. WTW-sensoren har indbygget temperatursensor, hvorimod Atlas-sensoren skal korrigeres efter en ekstern temperatursensor. Af figur D.1 fremgår forsøgsresultatet fra et af tre forsøg.



Figur D.1. Sammenhæng mellem elektrisk ledningsevne og temperatur.

Eksperimentet viser, som forventet, en stigende ledningsevne ved stigende temperatur. Eksperimentet gentages tre gange med forskellige mængder salt.

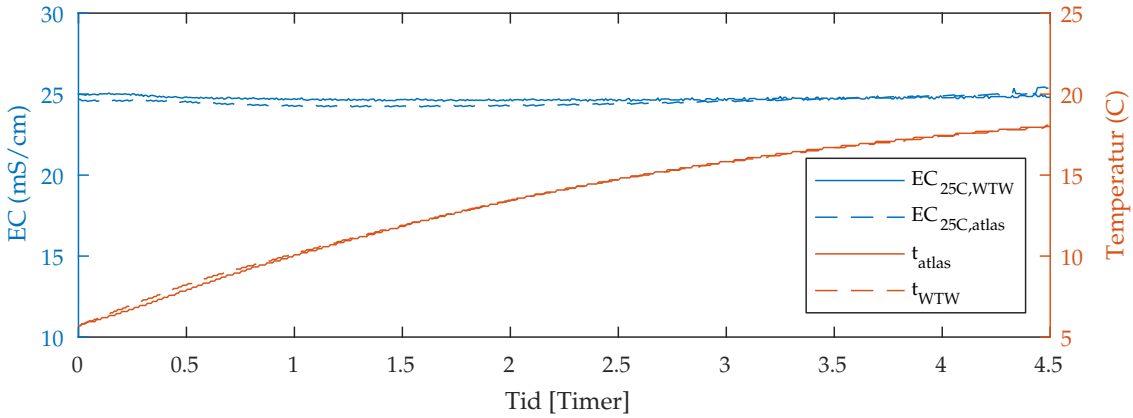
For alle forsøg korrigeres ledningsevnen op til 25 °C. Temperaturkoefficienten tilpasses således, at EC_{25} bliver konstant. På figur D.2 fremgår fejlen på korrektionen ved forskellige temperaturkoefficienter. Fejlen beregnes som den gennemsnitlige absolutte fejl mellem den korrigerede EC_{25} og gennemsnittet af EC_{25}. Fejlen bliver dermed mindre, når EC_{25} bliver mere konstant.



Figur D.2. Optimering af temperaturkoefficient til korrektion af ledningsevne.

Det fremgår af figur D.2, at optimum for temperaturkoefficienten ligger mellem 0.0175 og 0.02. Det fremgår, at der ikke er forskel på WTW-sensoren og sensoren fra Atlas. Dermed vurderes det at være fornuftigt at benytte 0,019 som temperaturkoefficient.

På figur D.3 fremgår EC_{25} for forsøget fra figur D.1, når der korrigeres med en temperaturkoefficient på 0,019.



Figur D.3. EC korrigeret med en temperaturkoefficient på 0.019.

Appendiks E

OPBYGNING AF LOGISTISK REGRESSIONSMODEL

En logistisk regressionsmodel benyttes til at analysere en binær responsvariabel ud fra en forklarende variabel. I dette projekt er den binære responsvariabel hvorvidt, badevandskvaliteten er tilfredstillende eller dårlig, og den forklarende variabel er ledningsevnen.

Til at opbygge en logistisk model benyttes en generaliseret lineær model (GLM). I modsætning til en lineær mode, kan en generaliseret lineær model beskrive sammenhænge mellem data, der ikke er normalfordelt, hvilket er tilfældet for de binære klassifikationer af badevandskvaliteten, der er fordelt efter en Bernoulli fordeling.

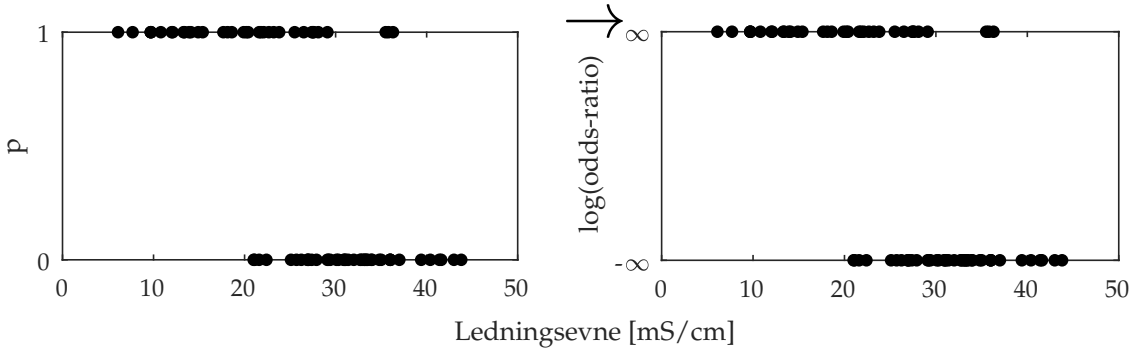
Klassifikationen af målingerne skal ses som sandsynligheden for, at målingen viser overskridelse af badevandskvalitetskravet. Altså en klassificering på 1 giver 100 % sandsynlighed for overskridelse. Hvis en almindelig lineær model tilpasses til sandsynligheden for overskridelse og ledningsevnen, vil det matematisk være muligt at opnå negative sandsynligheder, hvilket ikke er logisk. Sandsynlighederne omskrives derfor til logaritmen af odds ratioen for overskridelse af badevandskravet. Sandsynligheden omskrives, som det fremgår af formel (E.1).

$$p \rightarrow log\left(\frac{p}{1-p}\right) \tag{E.1}$$

Hvor:

p | Sandsynlighed for overskridelse af badevandskriterie [-]

I stedet for at benytte en sandsynlighed mellem 0 og 1, benytter vi nu logaritmen til odds-ratioen der for $p = 1$ og $p = 0$ tilsvarer ∞ og -∞. Sansynligheden, p, og logaritmen til odds-ratioen for målingerne fremgår af figur E.1.



Figur E.1. Sandsynligheden og log(odds-ratio) for overskridelse af badevandskvalitet afhængig af ledningsevnen.

Årsagen til at sandsynligheden transformeres til log(odds-ratio) er, at det nu er muligt at tilpasse en lineær model til datasættet, da responsen ikke kan være en negativ sandsynlighed. Der tilpasses en generaliseret lineær model ved at benytte maximum-likelihood. Denne model beskrives ved formel (E.2).

$$\log\left(\frac{p}{1-p}\right) = ax + b \tag{E.2}$$

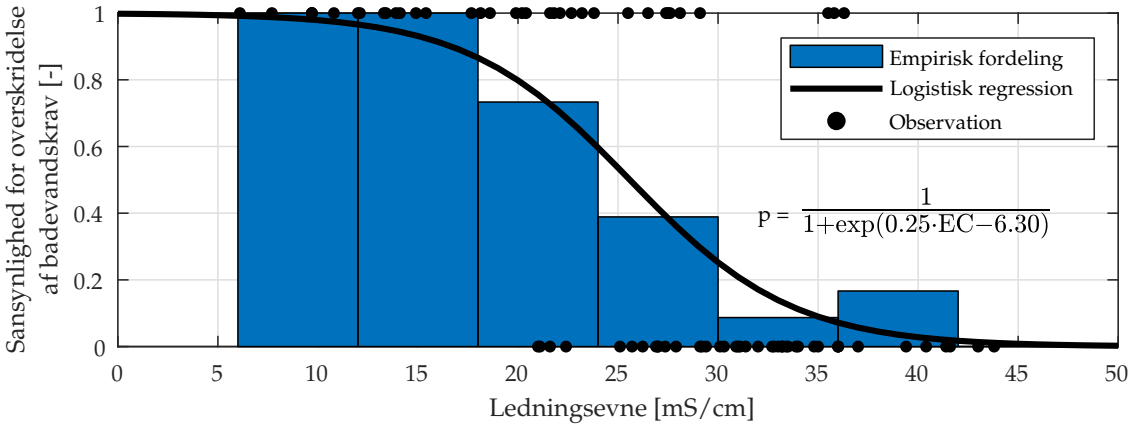
Hvor:

a, b	Modelparametre til GLM [-]
x	Uafhængig variabel til GLM, ledningsevne [mS/cm]

Udtrykket for logaritmen til odds-ratioen kan herefter omskrives til et udtryk for sandsynligheden ved hjælp af logit-funktionen. Logit funktionen fremgår af formel (E.3).

$$p = \frac{1}{1 + \exp(-ax - b)} \tag{E.3}$$

Den logistiske regression for sandsynligheden for overskridelse af badevandskvalitets-kravet ved formel (E.3) fremgår af figur E.2.



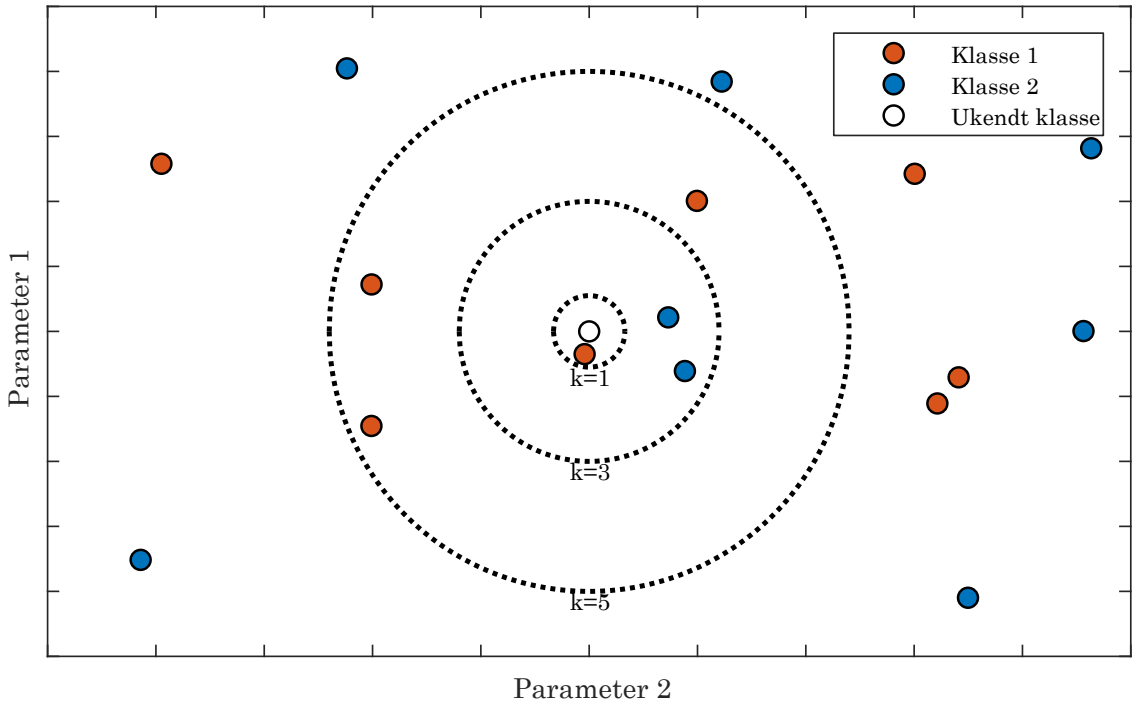
Figur E.2. Logistisk regression for sammenhængen mellem ledningsevnen og sandsynligheden for overskridelse af badevandskvalitetskravet.

Appendiks F

BESKRIVELSE AF MACHINE LEARNING ALGORITMER

F.1 K-Nærmeste naboer

K-nærmeste nabo-algoritmen (KNN) er en algoritme, der typisk anvendes som klassifikationsalgoritme. En illustration af princippet bag nærmeste nabo fremgår af figur F.1.

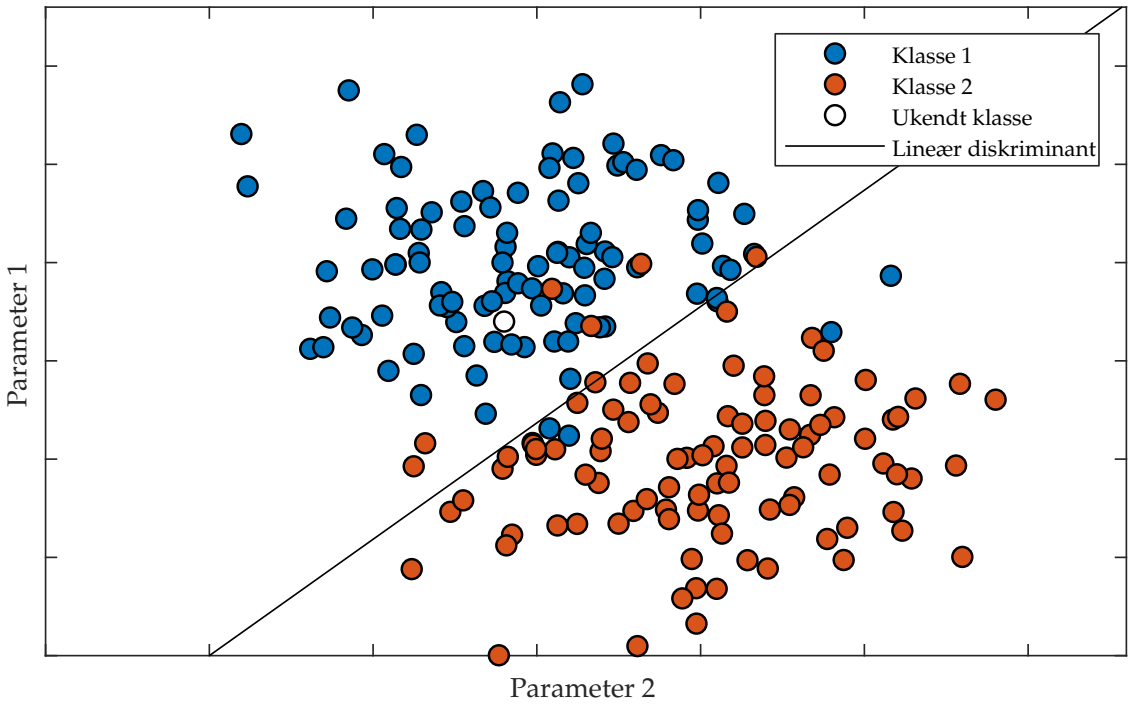


Figur F.1. Princip bag KNN algoritmer til klassificering.

Nærmeste nabo-algoritmer fungerer ved at plotte alle datapunkter i et koordinatsystem, der har én dimension for hver inputparameter. Klassifikation af nye input kan derefter gøres ved at plotte dem i samme koordinatsystem og fastsætte afstanden og klassifikationen af de K nærmeste punkter. Derefter kan den afstandsvægtede klasse beregnes, og det ukendte punkt klassificeres. [Webb og Copsey, 2011]

F.2 Lineær diskriminant

Lineær diskriminant-algoritmer går grundlæggende ud på at dele et datasæt op i klasser med en lineær kombination af inputparametrene. Eksempelvis en linje der deler et datasæt bestående af to input parametre op i to klasser, som det fremgår af figur F.2.

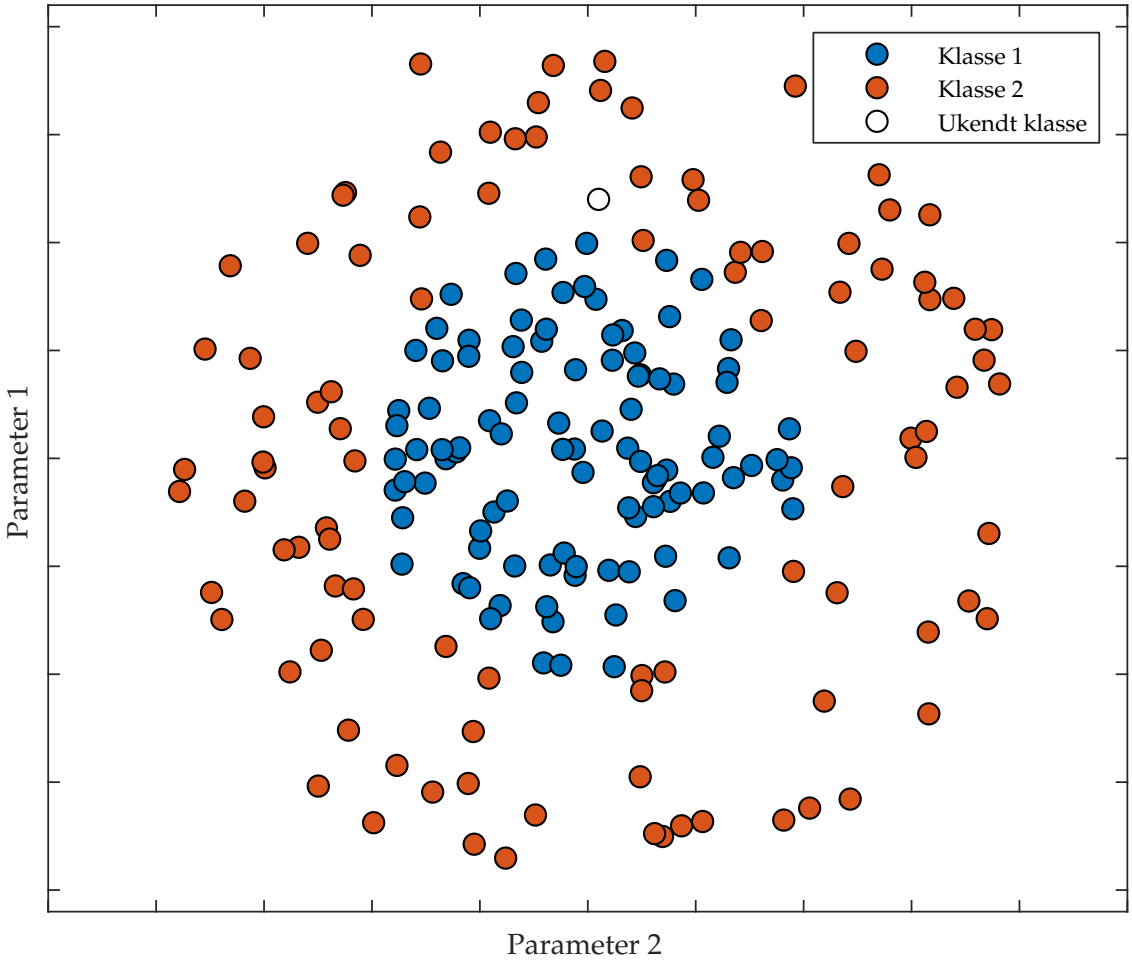


Figur F.2. Lineær diskriminant analyse til opdeling af to klasser

Linjen på figur F.2 kan, efter træning på et kendt datasæt, bruges til at kategorisere nye punkter ved at afgøre hvilken side af linjen, de ligger på. Algorimten fungerer ved at projicere data ned på x-aksen og bestemme fordelingsfunktionen for de to klasser. Derefter roterers datasættet, og det projiceres igen ned på x-asken og fordelingsfunktionerne bestemmes. Ved at finde frem til den rotation, hvor fordelingerne har den mindste spredning, og hvor middelværdierne er længst fra hinanden, kan den mest præcise klassifikation bestemmes. [Hao, Dong og Fan, 2015]

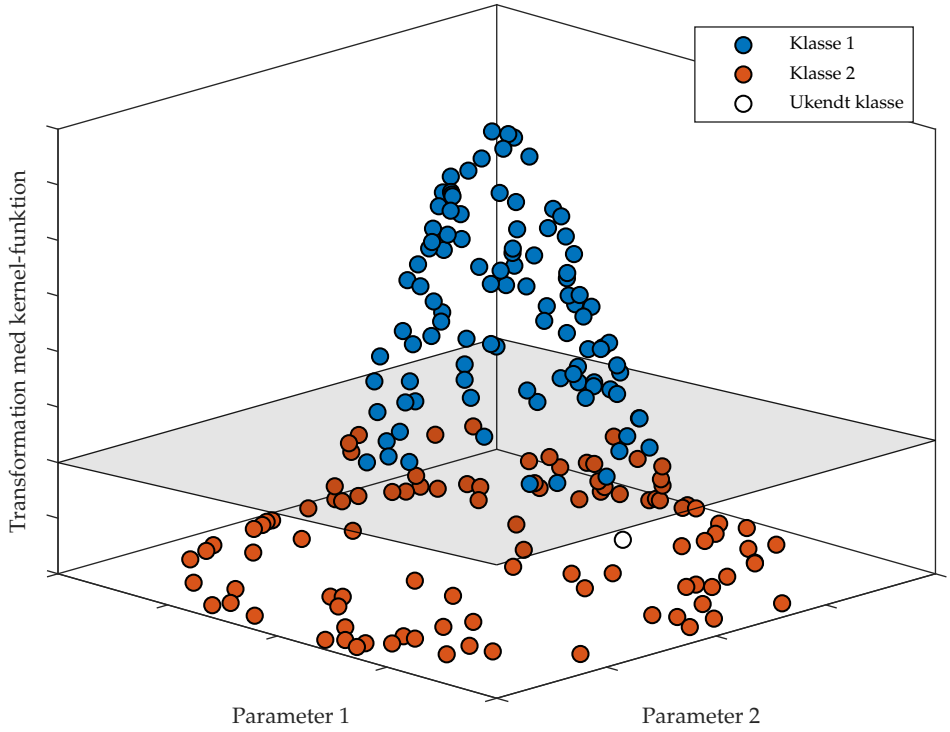
F.3 Support vektor maskiner til klassifikation

Support vektor maskiner (SVM) er en algoritme, der kan bruges til at klassificere observationer af ukendt klasse. Den fungerer ved at tilføje en ekstra dimension til inputdata ved hjælp af en kernel-funktion. At tilføje en ekstra dimension ved hjælp af en kernel-funktion giver mulighed for at separere data fra forskellige klasser med et plan. På figur F.3 fremgår et eksempel på hvordan et to-dimensionelt datasæt umiddelbart kan være svært at adskille for en algoritme. [MathWorks, 2019]



Figur F.3. Originale data, før transformation med en kernel-funktion

Udfordringen ved SVM er at identificere og tilpasse den rigtige kernel-funktion, således at datasættet kan transformeres til en ekstra dimension og afgrænses af et plan. På figur F.4 er data transformeret med en to-dimensionel normalfordelingsfunktion med middel i centrum af datapunkterne og samme varians i begge dimensioner.



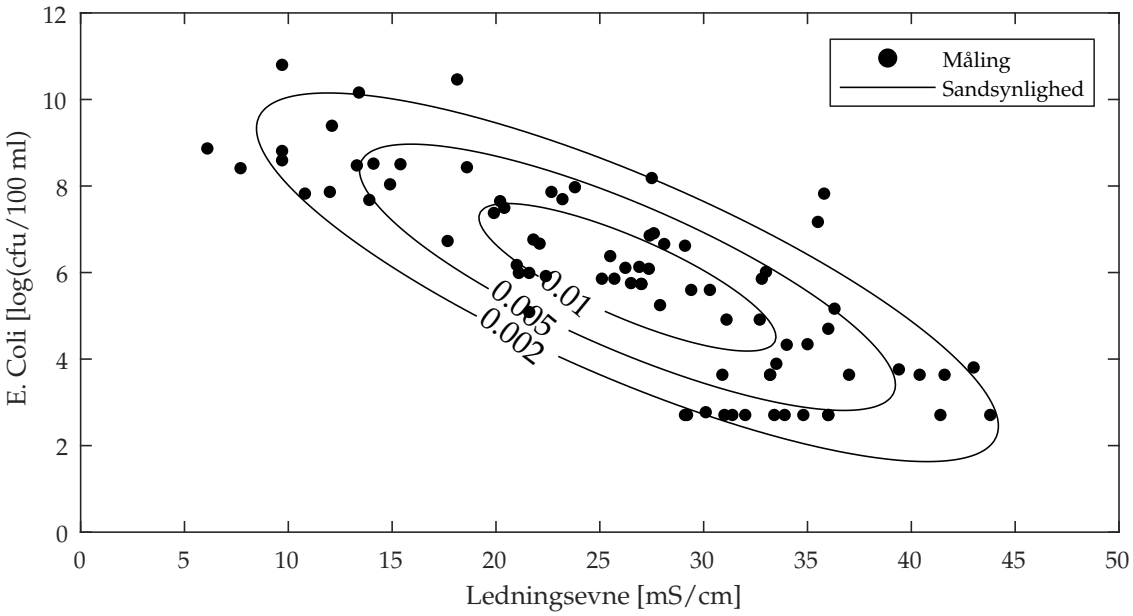
Figur F.4. Data efter transformation med kernel-funktion.

Som det fremgår af figur F.4, er det efter transformation muligt at finde et plan, der adskiller de to klasser med en rimelig nøjagtighed. Derefter kan klassifikationen af ukendte punkter gøres ved at afgøre, om de er på den ene eller anden side af planet.

Til at klassificere badevandskvaliteten, på baggrund af ledningsevne og regninput, er der opnået de bedste resultater ved at benytte kernelfunktionen radial basis function, der minder meget om den funktion, der er benyttet i eksemplet på figur F.4.

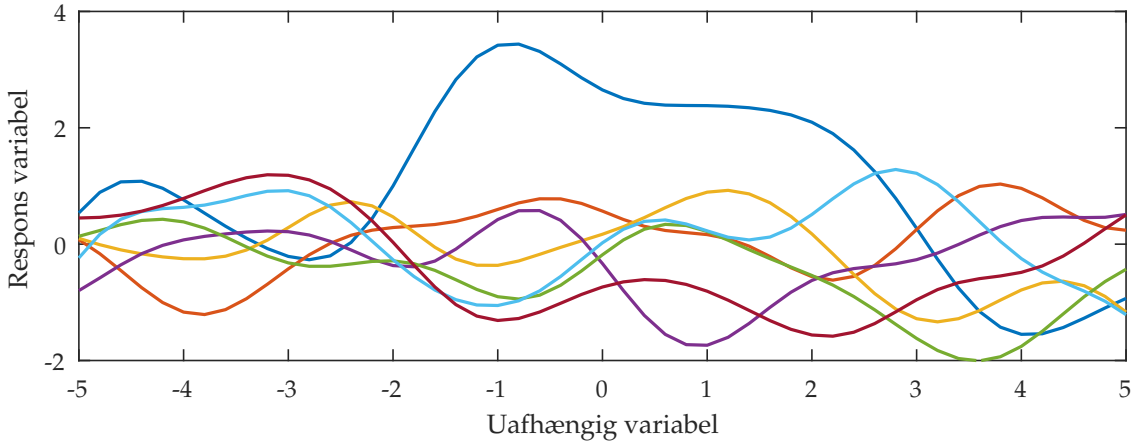
F.4 Gaussiansk proces regression

Gaussisansk proces regression (GPR) er et statistisk værktøj, der bygger på bayesiansk inferens og gausianske processer. Det fundamentale i bayesiansk inferens er, at der samples en tilfældig prior fordeling, eksempelvis en normalfordeling, baseret på en valgt middelværdi og spredning. Herefter tilpasses fordelingen til observationerne, hvilket resulterer i en posterior fordeling med en ny middelværdi og spredning. Dette kan ligeledes udføres for multivariate fordelinger, hvor der benyttes en middelværdi for hver variabel samt en kovarians-matrix. Af figur F.5 fremgår et eksempel på en multivariat normal-fordeling mellem ledningsevne og E. Coli.



Figur F.5. Multivariat normal-fordeling mellem ledningsevne og E. Coli.

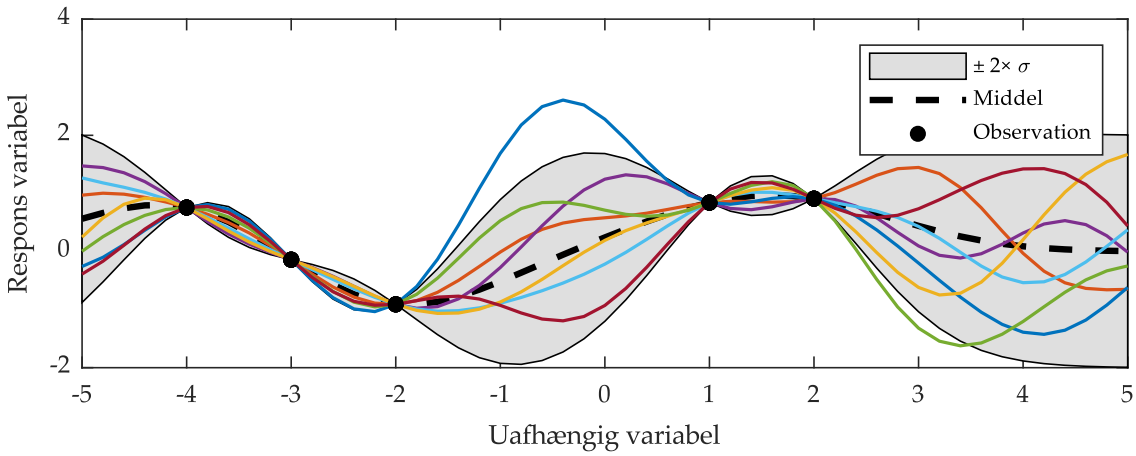
I gaussiansk proces regression benyttes bayesiansk inferens til at optimere en gausiansk proces. En gausiansk proces er en fordeling over funktioner i stedet for en fordeling over vektorer, som en multivariat normal-fordeling er. En gausiansk proces er beskrevet på baggrund af hyperparametrene; middel-funktionen og kovarians-funktionen (også benævnt kernel-funktionen). Ligesom i bayesian inferens opstilles der en prior gaussiansk proces, der er baseret på en prior middel-funktion og kovarians-funktion. Fra denne guasisanske process kan der herefter samples tilfældige funktioner, der kan beskrive den multivariate fordeling mellem data som funktion. Dette gøres ved at bruge kernel-funktionen, der er grundstammen i de funktioner, der samples. Af figur F.6 fremgår et eksempel på en række udtræk fra en tilfældig prior fordeling. [C. E. Rasmussen og Williams, 2005]



Figur F.6. Princip for prior fordeling mellem uafhængig og responsvariabel. Eksemplet viser princippet for en 2-dimensionel sammenhæng. Eksemplet er baseret på fiktive tal.

Herefter tilpasses prior fordelingen ved brug af observationsdata. Dette gøres ved at

optimere middel- og kernel-funktionen ved brug af maksimeret marginal log likelihood. Ud fra den nye middel- og kernel-funktion kan der nu samples en række funktioner igen, hvorfra der kan opstilles en resulterende middel af de samplede funktioner, samt konfidensinterval. Af figur F.7 fremgår de optimerede funktioner fra figur F.6. Funktionerne er optimeret ud fra observationerne på figuren. [C. E. Rasmussen og Williams, 2005]



Figur F.7. Princip for posterior fordeling mellem uafhængig og responsvariabel. Eksemplet viser princippet for en 2-dimensionel sammenhæng.

Ud fra midlen af funktionerne på figur F.7 kan den gaussianske proces regression benyttes til at estimere værdier ud fra et sæt uafhængige variable. Der kan ligeledes bestemmes et konfidensinterval for modellen, hvilket er en særlig egenskab for gaussiansk proces regression.

Appendiks G

BESKRIVELSE AF PERFORMANCE-PARAMETRE

G.1 Hitrate

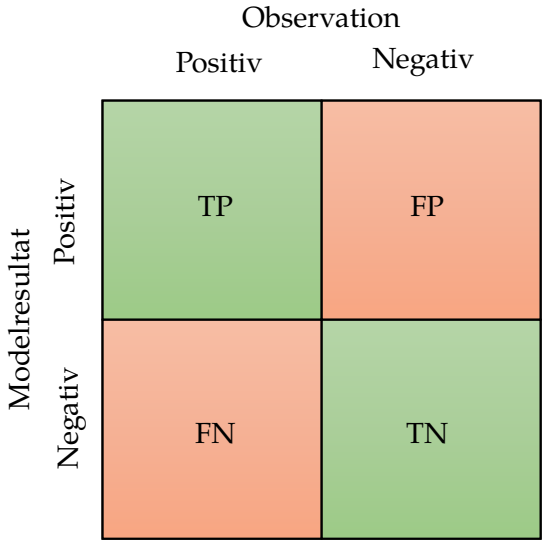
Hitraten af en model benyttes til at beskrive, hvor god modellen er til at klassificere positive og negative observationer. Hitraten beregnes ud fra formel (G.1)

$$Hitrate = \frac{TP + TN}{TP + TN + FP + FN} \tag{G.1}$$

Hvor:

$Hitrate$	Modellens hitrate [-]
TP	Antal korrekte positive modelresultater [-]
TN	Antal korrekte negative modelresultater [-]
FP	Antal forkerte positive modelresultater [-]
FN	Antal forkerte negative modelresultater [-]

Hitraten kan desuden visualiseres ved hjælp af en confusion matrix, der viser fordelingen mellem TP, TN, FP og FN, hvilket kan benyttes til at vurdere, hvorvidt modellen er bedst til at forudsige positive eller negative observationer. Princippet for en confusion matrix fremgår af figur G.1.



Figur G.1. Princip for confusion matrix.

G.2 Nash-Sutcliffe Efficiency

NSE beregnes, som det fremgår af formel (G.2) [Krause, Boyle og Bäse, 2005].

$$NSE = \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)^2} \tag{G.2}$$

Hvor:

NSE	Nash-Sutcliffe Efficiency [-]
n	Antal observationer [-]
O	Observeret bakteriekoncentration [log(cfu/100 ml)]
P	Modelleret bakteriekoncentration [log(cfu(100 ml)]

NSE kan variere mellem $-\infty$ og 1, hvor 1 angiver en perfekt korrelation mellem observerede og målte data. En NSE på 0 svarer til, at modellen beskriver de observerede data lige godt som middel-værdien af de observerede data. NSE beskriver modellens performance uden information om modellens numeriske usikkerhed. [Nielsen m.fl., 2018]

G.3 RMSE

RMSE beregnes, som det fremgår af formel (G.3).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}} \tag{G.3}$$

Hvor:

$RMSE$	Root-mean-square error [log(cfu/100 ml)]
n	Antal observationer [-]
O	Observeret bakteriekoncentration [log(cfu/100 ml)]
P	Modelleret bakteriekoncentration [log(cfu(100 ml)]

RMSE er et mål for usikkerheden på modellen. RMSE på 0 svarer til en perfekt korrelation mellem observerede og modellerede data. [Nielsen m.fl., 2018]

Appendiks H

ANALYSE AF E. COLI I VANDPRØVER

Der har i dette projekt været et behov for at bestemme antallet af E. Coli bakterier i en række vandprøver. Til dette formål er most probable number-metoden (MPN) blevet anvendt. Metoden udnytter det faktum, at det er væsentligt nemmere at fastslå om en prøve indeholder E. Coli eller ej, istedet for at fastslå hvor mange E. Coli en prøve indeholder. Metoden fungerer ved at opdele en prøve i et antal delprøver og så tælle hvor mange delprøver, der indeholder E. Coli. Ud fra antallet af delprøver, der indeholder E. Coli, kan antallet af E. Coli i den originale prøve beregnes eller estimeres ved tabelopslag. [Oblinger og Koburger, 1975]

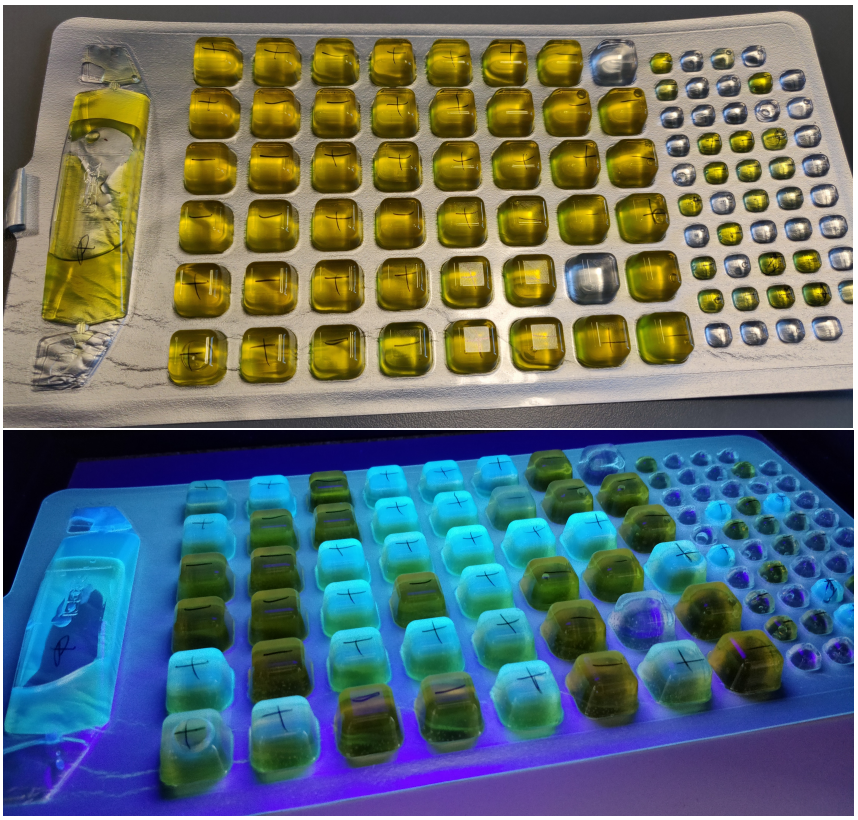
H.1 Fremgangsmåde

Til at bestemme antallet af E. Coli er der benyttet prøvekits fra IDEXX med colilert-reagenspulver samt bakke med 97 brønde til opformering af bakterier [IDEXX, 2019a]. Med den udvalgte metode og udstyr er det muligt at måle mellem 1 og 2419,6 E. Coli per 100 milliliter. Grænsen for for tilfredsstillende badevandskvalitet er 500 E. Coli per 100 milliliter. Derfor fortyndes prøven i forholdet 1:10 således, at det er muligt at måle mellem 10-24.196 E. Coli per 100 milliliter, hvilket passer bedre til koncentrationerne i badevand.

Der hældes 100 milliliter af den fortyndede prøve i bakken, og bakken forsejles. På figur H.1 fremgår bakken inden inkubation.

Figur H.1. 100 milliliter af vandprøven fordelt i bakken inden inkubation. Bemærk at ingen af brøndene er farvede.

Prøven inkuberes ved 37 °C i 24-26 timer. Herefter fremkommer der en gul farve i brønde, der indeholder coliforme bakterier og en fluorescerende farve i UV-lys for brønde, der indeholder E. Coli.



Figur H.2. Prøve efter inkubation. Øverst er prøven i normalt dagslys. Nederst er prøven under UV-lys.

Til venstre på figur H.2 fremgår det, at 21 små brønde og 47 store brønde er gule, og at 6 små brønde og 25 store brønde er fluorescerende. Ved opslag i MPN-tabel estimeres indholdet af coliforme bakterier at være 2400 cfu/100ml og indholdet af E. Coli at være 422 cfu/100ml for den viste prøve [IDEXX, 2019b].