Norges miljø- og biovitenskapelige universitet

Masteroppgave 2022 30 stp
Fakultetet for realfag og teknologi

Mikroorganismer i norske råvannskilder: Statistisk modellering med utgangspunkt i Vannverksregisteret

Microorganisms in Norwegian Raw Water Sources: Statistical Modelling Based on Data from Vannverksregisteret

Rakel Debora Solhaug
Vann- og miljøteknikk

Forord

Denne masteroppgaven er skrevet våren 2022 ved Norges miljø- og biovitenskapelige universitet i Ås som en avslutning av mitt studie i Vann- og miljøteknikk. Jeg valgte denne oppgaven fordi interessen min for risiko knyttet til vannbåren sykdom gjennom studietiden har økt. Jeg har også blitt mer innteressert i hvordan man kan bruke statistiske modeller til å beregne risiko knyttet til vannbåren sykdom. Gjennom arbeidet med denne oppgaven har jeg tilegnet meg mye ny kunnskap om norske råvannskilder og hvordan man kan fortsette å undersøke disse. Jeg har også fått bruk for kunnskap jeg har lært i løpet av studietiden min på tvers av fagfelt.

Jeg ønsker å takke min veileder Vegard Nilsen som har hjulpet til med tilbakemeldinger på idéer, innhold og gitt gode råd når arbeidet ikke har gått etter planen. Jeg vil også takke for fleksibiliteten han har vist under arbeidet med denne oppgaven, særlig når det har oppstått problemer.

I tillegg ønsker jeg å takke Knut Kvaal for at han delte sin kunnskap om statistiske modeller, ideer til hvordan man kunne gjennomføre målsetningen i oppgaven og ikke minst brukt av sin tid til å gjøre modelleringer for å hjelpe til med fremgangsmåten i oppgaven.

Til slutt ønsker jeg å takke familien min som har hjulpet til med lesing og tilbakemeldinger under arbeidet med denne oppgaven. Jeg har satt stor pris på råd, korrekturlesing og idéer for hvordan jeg kunne strukturere informasjonen.

Moss, mai, 2022

Rakel Debora Solhaug

Sammendrag

Det er lite informasjon om sykdom i drikkevann i Norge. For å kunne gjennomføre en QMRA trenger man konsentrasjon av patogene mikroorganismer i råvannet. I Sverige har Chalmers tekniska högskola utviklet et QMRA-verktøy, men mange brukere av verktøyet hadde ikke informasjon om patogenkonsentrasjoner i råvannet. Svenskt Vatten Utveckling ga ut en rapport i 2018 for å hjelpe brukere av QMRA-verktøyet til å velge patogenkonsentrasjoner i råvannet.

Hovedformålet med denne oppgaven er å undersøke om norske råvannskilder kan kategoriseres på en måte som kan benyttes sammen med den svenske rapporten. Det ses også på hvordan man kan bruke den svenske rapporten sammen med norske råvannskilder.

For å undersøke dette ble data fra Vannverksregisteret om råvannsprøver og inntakspunkter for norske råvannskilder hentet inn. Disse dataene ble undersøkt med PCA og clusteranalyse med Ward's metode.

Resultatene fra disse metodene viste at det er vanskelig å kategorisere de norske råvannskildene inn i kildetypene fra den svenske rapporten. Fra clusteranalysen finner man fire grupper blant de norske råvannskildene. PCA gir en indikasjon på at norske grunnvannskilder kanskje ikke er så mikrobielt rene som tidligere antatt. Om man ønsker å bruke den svenske rapporten er et alternativ å plukke tilfeldige statistiske fordelinger for hver av de norske råvannskildene. Man kan deretter gjennomføre en Monte Carlo-simulering for å få et overslag over patogenkonsentrasjoner.

Det mangler en del kunnskap om patogene mikroorganismer i norske råvannskilder, og det er fortsatt behov for flere undersøkelser for å kartlegge dette. Et forslag til videre undersøkelser er å finne en mulig korrelasjon mellom variablene i Vannverksregisteret.

Summary

Information about disease in Norwegian drinking water is scarce. The concentration of pathogen's in raw water is of absolute necessity when conducting a QMRA. Chalmers technical college in Sweden has developed a framework for application of QMRA in Swedish raw water sources. Many users of the framework didn't have the necessary information about pathogen's in their raw water sources. Svenskt Vatten Uteveckling published a report in 2018 to help users of the QMRA framework selecting pathogen concentrations.

The main purpose of this thesis is to examine the possibility to categorize Norwegian raw water sources for usage with the Swedish report. In addition, how to use the Swedish report with Norwegian raw water sources is examined.

Intake point data and results of raw water analysis was collected from Vannverksregistret. A PCA and Ward's method cluster analysis was conducted.

The results imply difficulty categorizing Norwegian raw water sources for usage with the Swedish report. The clustering shows four possible groups of Norwegian raw water sources. The PCA indicate more microbial contamination in ground water sources than previously assumed. Choosing at random statistical distributions for each raw water source is a possible use of the Swedish report. Monte Carlo simulations may yield estimates of concentration of pathogen's.

There's still a lot we do not know about concentration of pathogen's in Norwegian raw water sources. The need for more research in this field of study is urgent. Further studies might examine the correlation of variables in Vannverksregisteret to expand the knowledge about Norwegian raw water sources.

Innhold

Figurer

Tabeller

Forkortelser

C. jejuni	*Campylobacter jejuni*
E. coli	*Escherichia coli*
COD	Kjemisk oksygenforbruk
DALYs	Disability-adjusted life years
f.eks	for eksempel
FHI	Folkehelseinstituttet
hhv	henholdsvis
m.m.	med mer
MATS	Mattilsynests skjematjenester
MBA	Mikrobiell barriereanalyse
PCA	Principal component analysis
pga.	på grunn av
QMRA	Kvantitativ mikrobiell risikoanalyse
STEC	Shigatoksinproduserende
SVU	Svenskt Vatten Utveckling
VREG	Vannverksregisteret
WHO	Verdens helseorganisasjon

1. Introduksjon

1.1 Hvorfor rent drikkevann er viktig

Rent drikkevann er en menneskerett. Ifølge WHO (2022) hadde 74% av verdens befolkning tilgang på trygt drikkevann i 2020, men det er fortsatt 2 milliarder mennesker som ikke har tilgang på trygt drikkevann. (Li og Wu, 2019) Utrygt drikkevann er knyttet til spredning av sykdommer som kolera, diaré, dysenteri, tyfoidfeber, hepatitt A og polio. (WHO, 2022) WHO estimerer at 829 000 mennesker dør hvert år av diaré knyttet til utrygt drikkevann og dårlige sanitære forhold. Mange av disse dødsfallene kunne vært forhindret. Dette gjelder særlig dødsfall blant barn under 5 år. Det antas at man kunne unngått 297 000 av disse barnedødsfallene hvis drikkevannskvaliteten og de sanitære forholdene hadde blitt forbedret. (WHO, 2022)

1.2 Vannkvalitet i Norge

Det antas at drikkevannskvaliteten i Norge er god. I Norge har over 92% av befolkningen det man beskriver som god vannforsyning og 90% av befolkningen får vann fra godkjenningspliktige vannbehandlingsanlegg. (Hyllestad, 2017) Råvannskildene har lite tungmetaller, miljøgifter, plantevernmidler og andre uønskede stoffer. Allikevel finnes det lite informasjon om hvor mange som faktisk blir syke i Norge av drikkevann. Utfordringene i den norske vannforsyningen er at det er mange små vannbehandlingsanlegg fordi befolkningne bor spredt, det brukes mye overflatevann og ledningsnettet er gammelt og lekker mye. (Hyllestad, 2017)

E. coli og leveringsstabilitet er grunnlaget for å kunne si om drikkevannskvaliteten er god. I *Forskrift om vannforsyning og drikkevann (drikkevannsforskriften)* (2016) blir kravet for antall prøver av drikkevannet som må analyseres for *E. coli* gitt. Det kan ikke være påvist *E. coli* i minst 95% av prøvene i løpet av et år for at drikkevannskvaliteten skal være god. I tillegg kan det ikke være mer enn 30 minutter med ikke-planlagte avbrudd i i vannforsyningen i løpet av et år.(Hyllestad, 2017)

Vannbehandlingen i Norge har blitt kraftig forbedret de siste 20-30 årene. Den størs-

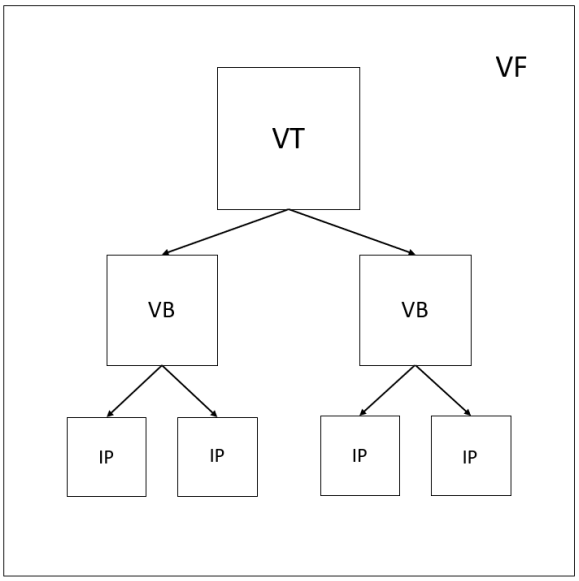
te forbedringen skjedde mellom 1995 og 2001, da et statlig program med økonomisk støtte for å oppgradere vannbehandlingsanlegg ble satt igang. I løpet av de seks årene ble 350 vannbehandlingsanlegg oppgradert. Disse leverte vann til ca. 800 000 personer. (Hyllestad, 2017)

Man kan dele vannforsyningssystemene i Norge inn i to grupper:

- Vannforsyningssystemer som leverer drikkevann til minst 50 personer
- Enkeltforsyninger som leverer drikkevann til færre enn 50 personer

Vannforsyningssystemer som leverer drikkevann til minst 50 personer må godkjennes av Mattilsynet. (Hyllestad, 2017) I 2020 var det over 1300 slike godkjente vannforsyningssystemer i Norge (Steinberg mfl., 2021). Informasjon om disse anleggene kan man finne i VREG.

Informasjonen i VREG blir registrert selv av eierene av vannforsyningssystemene gjennom MATS. Dette er påkrevd av helsemyndighetene. Opprinnelig hadde FHI ansvaret for innhenting av data, men ansvaret ble overført til Mattilsynet i 2009. (FHI, 2022) Utvalgte data fra VREG er tilgjengelig på Mattilsynets nettsider hvor man kan finne flere Excel-filer med informasjon om forskjellige deler av vannforsyningssystemene.



Figur 1.1: Antatt hieriarki for hvordan et vannforsyningssytem er bygd opp. VF er vannforsyningssystem, VT er vanntransportssystem, VB er vannbehandlingsanlegg og IP er inntakspunkt.

De fire hovedgruppene med data på Mattilsynets nettside er:

- Vannforsyningssystem
- Vannbehandlingsanlegg

- Inntakspunkter
- Distribusjonsnett (vanntransportsystem)

Det antatte hieriarkiet for disse fire hovedgruppene med data er vist i figur 1.1. I tillegg finner man på Mattilsynets nettsider historiske data over f.eks. råvannsprøver, drikkevannsprøver, fornying av ledningsnettet m.m.

Enkeltforyninger må ikke godkjennes av Mattilsynet og det finnes derfor lite informajson om disse anleggene. Enkeltforsyninger leverte vann til ca. 525 000 personer i 2017. (Hyllestad, 2017)

De norske vannkildene kan deles inn i to hovedtyper med flere undertyper. Disse er:

- overflatekilder
  - innsjø
  - elv
  - bekk
  - tjern
- grunnvannskilder
  - grunnvann i løsmasser
  - grunnvann i fjell
  - grunnvann fra infiltrert overvann

15% av den norske befokningen får drikkevann fra grunnvannskilder. Dette er en lav andel grunnvannskilder sammenlignet med resten av Norden. 85% får drikkevannet sitt fra overflatekilder. Andelen av vannbehandlingsanlegg som leverer vann fra grunnvannskilder er 40%. Forskjellen mellom denne andelen og hvor mange personer som får drikkevann fra grunnvannskilder kan forklares med at mange vannbehandlingsanlegg som bruker grunnvann er små. De store vannbehandlingsanleggene bruker overflatevann som råvannskilder. (Kvitsand og Fiksdal, 2010)

I VREG registreres indikatorbakterier og *Cryptosporidium* i enkelte grunnvannskilder. Dette stemmer ikke med tidligere oppfatning om kvaliteten på grunnvann. I perioden 2003-2012 ble 43% av registrerte utbrudd i Norge knyttet til grunnvann når råvannskilden ble gitt. Når det ble tatt hensyn til hvor mye vann som ble produsert, sto grunnvann for 32% av utbruddene. Hvis man tar ut utbrudd knyttet til distribusjonsnett ble andelen rundt 15%. (Kvitsand og Fiksdal, 2010)

I enkelte grunnvannskilder er det også høye konsentrasjoner av fluor og radon. Radon er knyttet til lungekreft og store mengder fluor påvirker tannutvikling. Det finnes ingen oversikt over hvor mange grunnvannskilder disse problemene gjelder, men det viser at drikkevannskvalitet ikke bare er forbundet med patogene mikroorganismer. (Hyllestad, 2017)

Det har vært store sykdomsutbrudd knyttet til drikkevann i Norge selv om drikkevannskvaliteten regnes for å være god. I 2004 var det et utbrudd av *Giardia* i Bergen hvor det er estimert at mellom 4000-6000 mennesker ble syke. Man antar at årsaken var en lekkasje av avløp til råvannskilden. Vannbehandlingsanlegget var ikke bygd for desinfisering av parasitter. (Guzman-Herrador mfl., 2016)

I 2007 var det et utbrudd av *Campylobacter* i Røros. Det antas at 1500 mennesker ble syke. Man antar at årsaken var forurensing fra fugler etter at det øvre beskyttende laget over brønnen var blitt fjernet for å grave ut en reservebrønn. Det var også problemer med distribusjonsnettet. Utbruddet i Røros var i en grunnvannskilde. (Guzman-Herrador mfl., 2016)

I 2019 var det et utbrudd av *Campylobacter* i Askøy. Over 2000 personer ble syke. Årsaken er gitt som forurensing i et av høydebassengene som skal sikre trykk i distribusjonsnettet. (FHI, 2019)

Disse tre utbruddene viser at også vannforsyningen i Norge kan være sårbar for mikrobiell forurensing. Samtidig er det også flere mindre utbrudd. I løpet av perioden 2003-2012 ble det rapportert om 28 vannbårne utbrudd med 8060 syke. Mange av disse utbruddene er i små vannforsyningssystemer, men mange av de syke tilhørte utbruddene i Bergen og Røros som nevnt tidligere. (Guzman-Herrador mfl., 2016)

Ved de store vannbehandlingsanleggene var klorering den mest brukte desinfeksjonsmetoden, men etter utbruddet i Bergen har flere vannbehandlingsanlegg tatt i bruk UV-bestråling. Av utbruddene fra overflatekilder i perioden 1984-2007, var 42% av dem knyttet til svikt i desinfiseringen. (Kvitsand og Fiksdal, 2010)

Vannkvaliteten vil også endres ved klimaendringer. Klimaendringene kan gi kraftigere nedbørsperioder som vil øke avrenningen til råvannskilder. Temperaturendringer kan senke barrierevirkningene i vannkildene. Nye og ukjente smittestoffer kan komme inn i vannforsyningssytemet. (Guzman-Herrador mfl., 2016)

1.2.1 Patogene mikroorganimser i norske vannkilder

De fleste mikroorganismer i råvann er ikke patogene (Ødegaard mfl., 2014), men de som er patogene kan i verste fall være dødelige. Sykdommer som er et stort problem i resten

av verden, er ikke så omfattende i Norge. Bekymringen er knyttet til enkelte patogene mikroorganimsmer som ofte gir mage-tarmsykdom.

Med mikroorganismer menes det parasitter, bakterier og virus. Virus er de minste av disse mikroorganismene, bakterier er større og parasitter er de største. Virus er en mikroorganisme som består av arvestoff, enten som RNA eller DNA. (Klein, 2022) De kan ikke formere seg utenfor en vertcelle. Den vanligste årsaken til vannbåren sykdom i Norge av virus er Norovirus. De fleste virus inaktiveres som oftest godt med klor.

Bakterier er prokaryote celler som finnes overalt på jorden. Siden de er prokaryote har de ikke cellekjerne og arvematerialet ligger derfor løst inne i cellen. Størrelsen varier mellom 0,2-10 mikrometer. (Sirevåg, 2022) Den vanligste vannbårne bakterien som gir sykdom i Norge er *Campylobacter*.

Parasitter i norske vannkilder er protozoer. Protozoer er encellete dyreorganismer. De er innvendige parasitter, noe som betyr at de lever i vevet på mennesker som blir smittet. (Tønjum, 2020) Parasittiske protozoer er mye mer resistente mot klor enn bakterier og virus. De to vanligste i vannbehandling er *Giardia* og *Cryptosporidium*.

For å påvise patogene mikroorganismer i vann brukes indikatororganismer. Koliforme bakterier og *E. coli* brukes som indikator på fersk fekal forurensing. *E. coli* er derimot ikke en god indikator for virus, *Cryptosporidium*, *Giardia* eller *Campylobacter*. Det er ingen rutine for å teste for virus og det er heller ikke lovpålagt. (Ødegaard mfl., 2014)

Noen patogene mikroorganismer som finnes i norske råvannskilder er norovirus, *E. coli*, *C. jejuni*, *Cryptosporidium* og *Giardia*. Noen kjennetegn ved disse mikroorganismene vil bli beskrevet litt nærmere.

Norovirus

Norovirus er en betegnelse på en gruppe virus. Viruset gir mage-tarminfeksjon hos mennekesker. (Klein, 2020)

Norovirus finnes i norske grunnvannskilder. En årsak til det kan være at norske grunnvannnskilder har lave temperaturer, noe som øker levetiden til virus. De største utbruddene av Norovirus er i månedene desember og januar. (Kvitsand og Fiksdal, 2010)

E. coli

E. coli er en bakterie som lever naturlig i tarmfloraen til dyr. Den er viktig for fordøyelsen. Den brukes derfor som en indikatororganisme for å kunne påvise fersk fekal forurensing sammen med koliforme bakterier. (Ødegaard mfl., 2014) Bakterien er i seg selv ikke sykdomsfremkallende, men noen typer kan gi mage-tarmsykdom. Det er da

STEC *E. coli* som man er bekymret for. (Sirevåg, 2019) I Norge er sykdom knyttet til *E. coli* vanligere i næringsmidler enn i drikkevann.

C. jejuni

C. jejuni er en termofil bakterie. Den forårsaker mage-tarmsykdom hos mennesker. Sykdomsformløpet kan også være asymptomatisk, men i verste fall kan den gi fulminant sepsis og død. De mest alvorlige sykdomsforløpene er som oftest knyttet til personer med nedsatt immunforsvar. Bakterien finnes f. eks. i fulger siden de har høy kropstemperatur. (Snelling mfl., 2005)

I Norge skjer de fleste utbruddene fra mars til november. Toppen for utbrudd er fra juli til semptember. Dette korrelerer med gjødsling i landbruk og stemer med rapportering fra de andre nordiske landene.(Kvitsand og Fiksdal, 2010) Siden sykdomsforløpet er så forskjellig antas det at antall utbrudd er underrapportert.

Cryptosporidium

Cryptosporidium er en parasitt som kan gi mage- tarmsykdom hos mennesker. Parasitten er den vanligste årsaken til sykdom i drikkevann fra parasitter. Sykdom fra smitte av parasitten kan gi diaré hos personer med godt immunforsvar, men sykdom kan være livstruende for personer med nedsatt immunforsvar. (Sunnotel mfl., 2006)

Cryptosporidium kan fjernes i drikkevann med konvensjonelle rensemetoder. Parasitten kan enten fjernes mekanisk med filtrering, sedimentering eller med kombinasjonen koagulering og flokkulering.(Sunnotel mfl., 2006) Klorering er lite effektivt mot *Cryptosporidium* siden den er en parasitt, men den kan desinfiseres med UV-stråling eller ozonering. (Ødegaard mfl., 2014)

Man kan ikke dyrke *Cryptosporidium*. Dette er fordi den trenger en vert for å overleve og reprodusere seg. For å påvise parasitten må man altså finne den i avføringsprøver hos smittede personer. (Sunnotel mfl., 2006)

Giardia

Giardia er en parasitt som gir mage-tarmsykdom hos mennesker. Smitte skjer som oftest gjennom kontaminert drikkevann. (Otterholt, 2021) Som andre parasitter er klor lite effektivt for å desinfiserer den. (Ødegaard mfl., 2014) Det antas at parasitten er kommet til Norge ved importsmitte siden den oppdages blant personer som har vært på reise i utlandet. (Guzman-Herrador mfl., 2016)

1.3 Tidligere kartlegging av sykdom i drikkevann

Det finnes lite informasjon om tidligere kartlegging av sykdom i drikkevann i Norge. FHI har gjennomført drikkevannsstudien som har som mål å kartlegge sykdom i drikkevann. Studien gjennomføres med selvrapportering av mage-tarmsykdom og drikkevannskonsum over en periode på 12 måneder. Datainnsamlingen ble avsluttet i starten av 2021, men det foreligger ingen resultater fra denne studien når denne oppgaven skrives.

Det finnes enkelte rapporter som ser på utbrudd som har blitt rapportert inn til Vesuv. Et eksempel på dette er rapporten av Guzman-Herrador mfl. (2016) som så på sykodmsutbrudd i drikkevann mellom 2003-2012. Et annet eksempel er artikkelen av Kvitsand og Fiksdal (2010) som så på vannbårne utbrudd i Norge mellom 1984-2007 med et hovedfokus på grunnvannskilder. Siden disse undersøkelsene kun ser på rapporterte utbrudd sier de ikke noe om hvor mange som blir syke av drikkevann.

Petterson mfl. (2016) utførte en studie for å se om det er mulig å lage en modell for Noroviruskonsentrasjoner til eventuell bruk i en QMRA. Noroviruskonsentrasjoner ble estimert ved to avløpsutløp som lå oppstrøms et drikkevannsinntak. De estimerte verdiene av Norovirus skulle komplimentere allerede eksisterende funn av patogene mikroorganismer.

Grøndahl-Rosado mfl. (2014) så på konsentrasjonen av Norovirus og Adenovirus i avløpsvann og råvannskilder til drikkevann. Vannprøver ble gjennomført mellom januar 2011 og april 2012. Det ble funnet betydelige konsentrasjoner av virus i råvannet, men det ble ikke gjort flere undersøkelser for å se hvilken helserisiko disse viruskonsentrasjonene kunne ha for konsumenter.

Lieungh (2021) estimerte i sin masteroppgave at det kan være rundt 300 000 personer som blir syke av drikkevann i året i Norge. Dette estimatet ble funnet med en QMRA. For å kunne gjennomføre QMRA ble *E. coli*-konsentrasjoner i vannprøver som er registrert i VREG sammenlignet med råvannstyper i Åström (2018) og kategorisert. Det ble da mulig å estimere konsentrasjon av patogene mikroorganismer i råvannskildene. 300 000 er et tall som FHI har benyttet tidligere, men som nå er blitt gått bort fra. Det finnes derfor ikke offisielle tall på hvor mange som blir syke av drikkevann i Norge.

1.3.1 Metoder for kartlegging av sykdom i drikkevann

Kartlegging av sykdom i drikkevann er en omfattende prosess. For å kunne finne ut hvor mange som blir syke trenger man å vite risikoen for sykdom. Det finnes ulike metoder for risikoanlyse i drikkevann. Omfanget varier fra visuell kartlegging av mulig risiko for kontaminering av vannkilden til statistiske modeller for å kvantifisere eventuell

mikrobiell risiko.

Inspeksjon av sanitære forhold er den enkleste metoden for risikoanalyse. Metoden består i å kartlegge visuelt for hvilke risikoer som kan oppstå. I små vannforsyningssystemer er dette en enkel og effektiv måte å kartlegge risiko på. Metoden er standardisert med sjekklister. En ulempe med metoden er at den ikke kan ta hensyn til spesifikke forhold rundt en kilde og metoden er kun visuell.(WHO, 2016)

En risikomatrise er en annen metode for risikoanalyse av drikkevann. Denne metoden er en firestegs metode som identifiserer risiko i et vannforsyningssytem og bruker en matrise til å identifisere alvorlighetsgraden til denne risikoen. Fordelen med denne metoden er at den fanger opp flere typer risiko. Ulempen er at det er vanskelig å bruke metoden konsekvent siden det kan bli uenighet i hva som er risiko og hvor alvorlig den er. (WHO, 2016)

QMRA

QMRA er en metode for å estimere sykdom i drikkevann. Et verktøy for QMRA har blitt utarbeidet av WHO (2016) for å standardisere metoden. Metoden bruker en systematisk vitenskapelig tilnærming for risikoanalyse. Dette gjør at metoden kan lettere brukes til å sammenligne mulige måter å forbedre håndtering av mulig risiko i vannforsyningssystemer.

I WHOs QMRA-verktøy er det fire hovedsteg med flere understeg som må gjennomføres. Disse er:

1. Formulering av problem
    (a) Identifiser mulige farer
    (b) Identifiser mulige måter for eksponering
    (c) Identifiser mulige helseutfall
2. Evaluering av eksponering
    (a) Definer måtene for eksponering
    (b) Kvantifiser komponentene i måtene for eksponering
    (c) Karakteriser eksponeringen
3. Evaluering av helseeffektene
    (a) Dose-respons
    (b) Sannsynlighet for sykdom

(c) Sannsynlighet for følgetilstand av sykdom

(d) Sykdomsbyrde

(e) Sekudærsmitte

4. Karakterisering av risiko

(a) Hensikten med evaluering av risiko

(b) Kvantitative målinger for risiko

(c) Variasjon og usikkerhet

(d) Sensivitetsanalyse

En av fordelene med QMRA er at man får kvantitative resultater for risiko. Noen av ulempene med en QMRA er at metoden er veldig ressurskrevende å gjennomføre. Man trenger mye bakgrunnskunskap og man trenger data som kan være vanskelig å finne. Mange vannforsyningssystemer har ikke prøver av patogeninnholdet i råvannskilden som man trenger for å kunne gjennomføre analysen.

I Sverige er det utviklet et QMRA-verktøy som man kan finne på nettsiden til høyskolen Chalmers. For å kunne gjennomføre en QMRA trenger man patogenkonsentrasjoner i råvannet. Dette kan være vanskelig. Rapport SVU 2018-3 av Åström (2018) skal hjelpe brukere av QMRA-verktøyet til å velge patogenkonsentrasjoner i råvannskilden. Rapporten bruker man ved å sammenligne råvannskilden sin med en av de seks kategoriene som er vist i figur 1.2. Disse kategoriene er tilpasset overflatekilder.

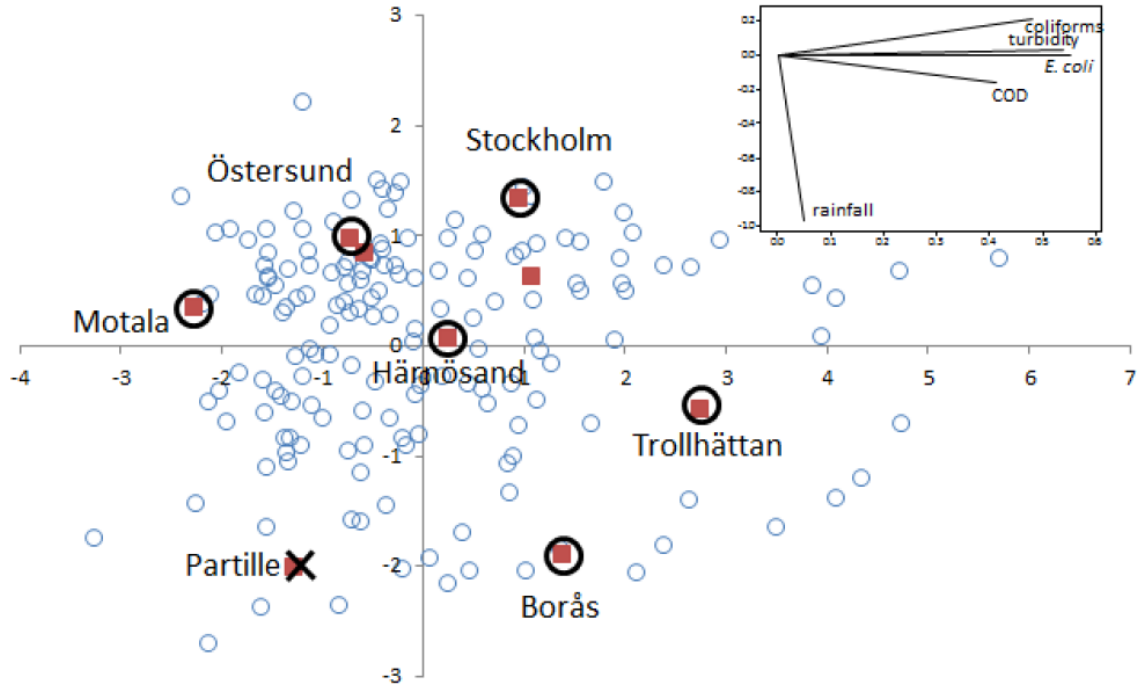
			Medelvärde (spridningsintervall)			
Punkt	Täkttyp	Fekala utsläppskällor uppströms vattenintaget, angivna i en enkät av vattenproducenterna	*E. coli* (antal/100 ml)	Koliforma bakterier (antal/100 ml)	COD (mg/l O_2)	Turbiditet (NTU)
A	Sjö	Reningsverk, avloppsutsläpp, betesdjur, gödsel, badande, dagvatten	1,5 (0–15)	87 (0–2 419)	8,0 (5,8–11)	2,9 (1,2–6)
B	Älv	Reningsverk, avloppsutsläpp, betesdjur, gödsel, enskilda avlopp, dagvatten	31 (0–486)	95 (0–2 400)	4,3 (3,4–6,1)	4,2 (1,0–19)
C	Sjö	Betesdjur, gödsel, enskilda avlopp	0,1 (0–2)	57 (1–530)	8,3 (6,8–11)	1,5 (0,7–3,4)
D	Sjö	Avloppsreningsverk, avloppsutsläpp, betesdjur, gödsel, badvatten, dagvatten	0,6 (0–9)[a]	11,9 (0–360)[a]	1,6 (–2,5)[a]	0,4 (0,2–1,6)[a]
E	Sjö	Avloppsreningsverk, betesdjur, gödsel, badande, enskilda avlopp, dagvatten	1,5 (0–32)	36 (0–1 100)	3 (–7,3)[a]	0,3 (0,12–1)
F	Sjö	Betesdjur, gödsel, fåglar, badande, enskilda avlopp, dagvatten, avloppsutsläpp	5,0 (0–200)	70 (0–13 000)	9,1 (6,6–13,4)	1,2 (0,46–3,2)

[a] Data från Vattentäktsarkivet vid SGU.

Figur 1.2: Tabell 2.1 fra SVU 2018-3 av Åström (2018)

Estimatene for patogenkonsentrasjonene har bakgrunn i en undersøkelse gjennomført

av det svenske Livsmedelverket i perioden 2013-2015. Data fra seks geografisk spredte inntakspunkter ble valgt ut til å være representanter for svenske overflatekilder. Disse inntakspunktene er markert med svarte sirkler i figur 1.3.



Figur 1.3: Biplot for PCA fra Säve-Söderbergh mfl. (2014) som er bakgrunnen for de svenske råvannskategoriene

Det ble gjennomført utvidede tester for patogene mikroorganismer i råvannsprøvene. Samtidig ble prøver tatt etter store nedbørsmengder for å se på om det var økte konsentrasjoner av patogene mikroorganismer (Säve-Söderbergh mfl., 2014).

Når man har funnet hvilken av de seks typene fra figur 1.2 råvannskilden passer inn i ut fra min-, max- og gjennomsnittsverdi for *E. coli*, koliforme bakterier, COD og tubiditet, kan man bruke tabellene gitt i figur 1.4 og figur 1.5 til å finne konsentrasjon av patogene mikroorganismer i råvannskilden. Disse tabellene gir verdier for tilpassede statistiske fordelinger for konsentrasjon av patogene mikroorganismer i hver råvannstype.

Patogene mikroorganismer som det er laget fordelinger for er *Cryptosporidium*, *Giardia*, *Campylobacter*, *Salmonella* og STEC *E. coli*. Statistiske modeller ble brukt til å estimerer patogeninnhold for *Cryptosporidium* og *Giardia* med Poisson-modeller og for *Cryptosporidium*, *Giardia*, *Campylobacter*, *Salmonella* og STEC *E. coli* med Gamma-modeller.

Poisson-modellen brukes når man har data for hvor mange ganger en hendelse inntreffer, men man ikke vet hvor mange ganger hendelsen ikke har inntruffet (Crawley, 2013). Denne modellen brukes derfor for konsentrasjon av patogene mikroorganismer siden den tar hensyn til at de fleste prøvesvarene vil være negative.

	Passning till Poisson-modell				Passning till Gamma-modell				
	Presumtiva		Konfirmerade		Konfirmerade			Förväntad medelhalt	Övre 95-percentil av variationen
Punkt	D[a]	μ	D[a]	μ	D[a]	α	β	Antal/liter	Antal/liter
Crypto-sporidium									
A	172,2	0,045	83,4	0,015	65,0	0,180	0,083	0,015	0,118
B	134,3	0,030	64,4	0,012	53,4	0,216	0,055	0,012	8,63E-02
C	93,0	0,025	42,0	6,89E-03	36,1	0,128	0,053	6,76E-03	6,08E-02
D	39,3	7,55E-03	25,6	4,31E-03	24,2	0,238	0,018	4,34E-03	3,05E-02
E	54,9	6,57E-03	9,1	5,47E-04	9,08	591	9,25E-07	5,47E-04	5,92E-04
F	125,1	0,038	61,9	0,013	61,8	2973	4,49E-06	0,013	0,014
Giardia									
A	44,9	6,56E-03	10,2	8,20E-04	10,2	1,51E+4	5,41E-08	8,20E-04	8,94E-04
B	54,3	9,03E-03	23,0	2,71E-03	20,0	0,073	0,037	2,72E-03	0,029
C	27,5	3,64E-03	-	-	-	-	-		
D	44,5	9,98E-03	8,7	9,98E-04	8,7	877	1,14E-06	9,98E-04	1,07E-03
E	186	0,015	74,1	5,57E-03	32,4	0,040	0,149	5,91E-03	0,068
F	65,6	0,012	10,4	8,83E-04	10,4	818	1,08E-06	8,83E-04	9,44E-04

[a] Värdet D kan sägas motsvara avvikelsen mellan uppmätta halter och den antagna fördelningen (D = –2 × Loglikelihood).

Figur 1.4: Tabell 3.2 fra SVU 2018-3 av Åström (2018)

	Passning till Gamma-modell			Förväntad medelhalt	Övre 95-percentil av variationen
Punkt	D[a]	α	β	Antal/liter	Antal/liter
Campylobacter					
A	72,2	0,403	3,196	1,28	7,14
B	152	0,512	5,366	2,74	13,7
C	-137	1,51	0,046	0,070	0,216
D	-9,6	0,394	1,204	0,475	2,66
E	-119	1,48	0,045	0,066	0,206
F	219	0,375	16,707	6,26	35,9
Salmonella					
A	-	-	-	ED	-
B	-179,2	2,118	0,023	0,050	0,136
C	-	-	-	ED	-
D	-	-	-	ED	-
E	-	-	-	0,14[b]	-
F	-117,2	0,671	0,159	0,106	0,468
STEC					
A	-184,1	2,11	0,022	0,047	0,128
B	7,5	0,255	5,22	1,33	9,06
C	-157,9	1,44	0,037	0,053	0,168
D	-	-	-	0,08[c]	-
E	-96,3	0,642	0,150	0,096	0,433
F	-66,3	0,612	0,327	0,200	0,919

[a] Värdet D kan sägas motsvara avvikelsen mellan uppmätta halter och den antagna fördelningen (D = –2 × Loglikelihood).
[b] *Salmonella typhimurium* påvisad i ett prov.
[c] VT2 påvisad i ett prov (i en av tre brunnar).

Figur 1.5: Tabell 3.3 fra SVU 2018-3 av Åström (2018)

1.4 Hovedmål

Råvannskvalitet har stor påvirkning på drikkevannskvalitet, men det finnes lite informasjon om patogene mikroorganismer i norsk drikkevann. Vi vet heller ikke hvor mange som blir syke av drikkevannet. For å få gjennomført en QMRA som kan gi et estimat på sykdom trenger man konsentrasjoner for patogene mikroorganismer i råvannet. Hovedmålsetningen med denne oppgaven er derfor:

1. Kan norske råvannskilder kategoriseres på en måte som kan benyttes sammen med Tabell 2.1 fra SVU 2018-3 som er vist i figur 1.2?

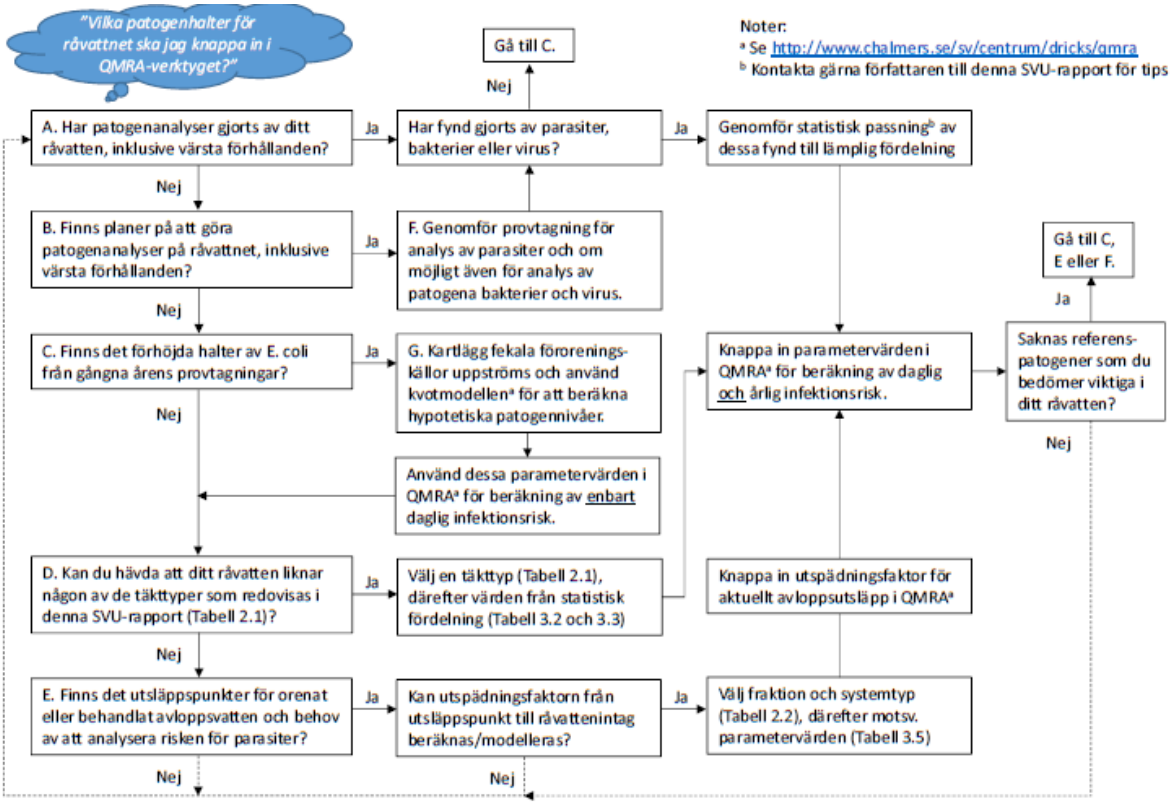
2. Gitt utfall av spørsmål 1: Hvordan kan SVU 2018-3 best brukes for å velge råvannskonsentrasjoner i Monte-Carlo-analyser for en norsk QMRA-studie?

For å svare på disse spørsmålene vil to statistiske metoder bli brukt. For det første vil det bli gjennomført en PCA. Man gjennomfører gjerne en PCA på store datasett hvor variablene er avhengige av hverandre og korrelerte. Ønsket er å redusere dimensjonen i datasettet slik at det blir lettere å analysere. Man beholder dermed kun den viktigste informasjonen. (Abdi og Williams, 2010)

For det andre vil det bli gjennomført en clusteranalyse. Den spesifikke metoden for clusteranalyse er Ward's metode. Denne clustermetoden analyserer for varians istedenfor avstandsberegninger. Den starter med n-antall clustere med størrelse 1 og kombinerer observasjoner til man står igjen med 1 cluster. (*14.7 - Ward's Method* 2022)

2. Metode

Metoden som er brukt i analysen er satt sammen av metoden fra SVU 2018-3 og masteroppgaven til F. T. Lieungh. I figur 2.1 ser man beslutningstreet som er gitt i vedlegg 3 i SVU 2018-3. Dette beslutningstreet beskriver hvordan man skal bruke rammeverket



Figur 2.1: Beslutningstre for bruk av QMRA-verktøyet fra Sverige

ut fra hvilken situasjon man er i. Man starter først i A hvor man svarer på om det er gjort patogenanalyser for våre råvann. Siden man ikke har spesifikk informasjon om hvert enkelt råvann i VREG er svaret nei og man går til punkt B.

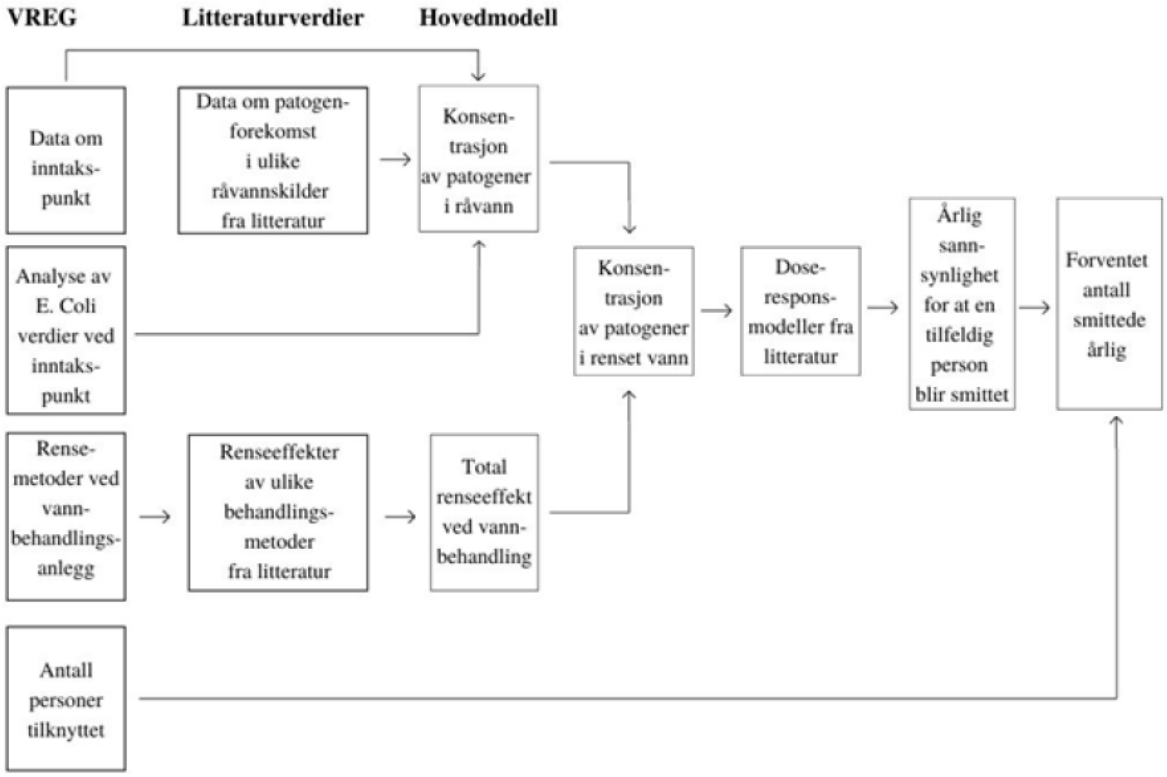
I punkt B spørres det om det finnes planer for å gjøre en analyse av patogeninnhold i råvannet. Igjen så har man ikke informasjon for hvert enkelt råvann og svaret blir derfor nei. Man går da videre til punkt C.

I punkt C blir det spurt om det er forhøyede verdier av *E. coli* de siste årene i råvannet.

Dette har man heller ikke nøyaktig informasjon for og svaret er derfor nei. Man går så videre til punkt D.

I punkt D er spørsmålet om råvannskilden ligner på noen av overflatetypene som finnes i tabell 2.1 i SVU 2018-3. Det er dette som skal undersøkes og det er her man stanser i beslutningstreet for denne analysen.

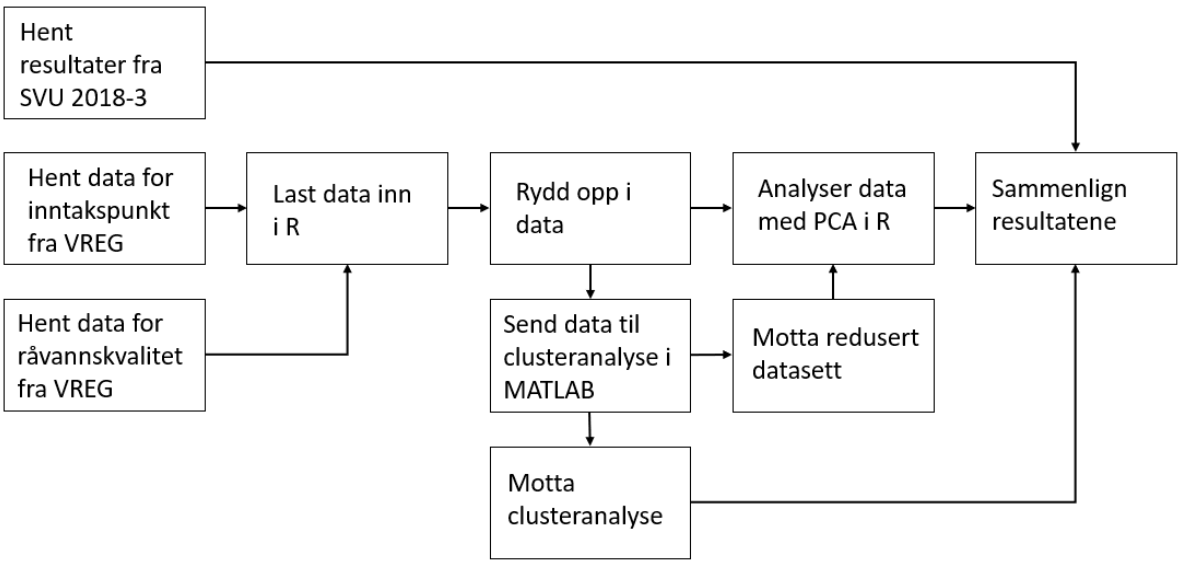
I figur 2.2 er flytskjemaet som ble brukt av F. T. Lieungh. Fra dette flytskjemaet vil data



Figur 2.2: Flytskjema fra masteroppgave til F. T. Lieungh

om inntakspunkt og data om patogenforekomst i ulike råvannskilder fra litteraturen bli fulgt. Istedenfor kun analyser av *E. coli*-verdier ved inntakspunktene vil også verdier for koliforme bakterier og turbiditet bli brukt. Resten av dette flytskjemaet vil ikke bli brukt i denne oppgaven.

I figur 2.3 ser man en oversikt over metoden som er brukt for denne analysen.



Figur 2.3: Flytskjema som viser arbeidsmetode for masteroppgaven

2.1 Data Wrangling

Full kode er tilgjengelig i vedlegg A. Analysen er gjort med Excel og R versjon 4.0.3.

Data for norske råvannskilder ligger tilgjengelig på Mattilsynets sider. (Mattilsynet, 2021) Der ligger et utvalg av variabler fra VREG i flere ulike Excel-filer. Disse filene er lagret som semikolindelte csv-filer. Filene som har blitt brukt i denne oppgaven er filen for inntakspunkt og filen for råvannskvalitet. Disse filene ble lastet ned og lagret i xlsx-format. Disse filene ble så lastet inn i R med readxl-pakken og lagret som to tabeller i formatet tibble.

Kun et utvalg av variablene og datapunktene skal brukes i analysen. Filene må derfor ryddes opp i, altså data wrangling. Ved å bruke de innebygde funksjonenen as.numeric() og as.character() blir kolonner lagret i riktig format. Numeriske verdier lagres som tall og kolonner med tekst lagres som characters.

I datsettet for råvannsverdiene brukes pakken stringr til å endre bokstaver og tegn i celler som ikke er formatert riktig. Her brukes str_replace()-funksjonen til å fjerne parenteser i celler siden R hopper over disse når kommandoer blir utført. Bokstaver som ikke er lastet inn riktig blir også endret slik at de er lesbare både for brukeren, men også i R. Dette gjøres ved å bruke stri_replace_all_regex()-funksjonen fra stringi-pakken. Denne funksjonen tillater å endre flere tegn og bokstaver samtidig.

Råvannsdata reduseres til kun rader som har data om *E. coli*, koliforme bakterier og turbiditet med den innebygde funskjonen subset(). Samtidig i subset()-funksjonen fjernes

Tabell 2.1: Utdrag fra råvannsdata etter variabelreduksjon

	mtid_ip	periode	analysetype	verdi_max	verdi_min	verdi_gjsn
1	Z0804221329513033252NQTQA	2013	Turbiditet	0.81	0.00	0.30
2	Z0804221329513033252NQTQA	2013	E.coli	0.00	0.00	0.00
3	Z0804221329513033252NQTQA	2013	Koliforme_bakterier	39.00	0.00	44635.00
4	Z0804221329513033252NQTQA	2014	Koliforme_bakterier	0.00	0.00	0.00
5	Z0804221329513033252NQTQA	2014	Turbiditet	0.37	0.00	0.20
6	Z0804221329513033252NQTQA	2014	E.coli	0.00	0.00	0.00
7	Z0804221329513033252NQTQA	2015	E.coli	0.00	0.00	0.00
8	Z0804221329513033252NQTQA	2015	Koliforme_bakterier	0.00	0.00	0.00
9	Z0804221329513033252NQTQA	2015	Turbiditet	0.27	0.12	0.00
10	Z0804221329513033252NQTQA	2016	E.coli	0.00	0.00	0.00

variabler som ikke skal brukes slik at man står igjen med variablene mtid_ip, periode, verdi_max, verdi_min og gjennomsnittsverdi.

Med funksjonen pivot_wider fra pakken tidyr blir max.-verdi, min.-verdi og gjennomsnittsverdi utvidet med *E. coli*, koliforme bakterier og turbiditet. Tabell 2.1 viser de 10 første radene i datasettet og de 6 gjenværende variablene i datasettet etter fjerning av unødvendige variabler. Tabell 2.2 viser de 10 første radene i datasettet, samt de 5 første variablene i datasettet etter kolonneutvidelsen.

Tabell 2.2: Utdrag fra råvannsdata etter kolonneutvidelse

	mtid_ip	periode	max_Turbiditet	max_E.coli	max_Koliforme_bakterier
1	Z08042213295130332 52NQTQA	2013	0.81	0.00	39.00
2	Z08042213295130332 52NQTQA	2014	0.37	0.00	0.00
3	Z08042213295130332 52NQTQA	2015	0.27	0.00	0.00
4	Z08042213295130332 52NQTQA	2016	0.16	0.00	0.00
5	Z08042213295130332 52NQTQA	2017	0.27	0.00	0.00
6	Z08042213295130332 52NQTQA	2018	0.17	0.00	0.00
7	Z08042811013014701 58IDBUT	2010	NA	0.00	0.00
8	Z08050713170952201 14EEKWW	2010	45658.00	NA	NA
9	Z08050713170952201 14EEKWW	2011	1.00	2.00	4.00
10	Z08050713170952201 14EEKWW	2012	44743.00	1.00	4.00

Til slutt blir råvanndataene over flere år kombinert. Dette gjøres ved å ta maksimums-

Tabell 2.3: Utdrag fra råvannsdata etter kombinering av max-, min- og gjennomsnittsverdier

	mtid_ip	Gjsn_E.coli	Max_E.coli	Min_E.coli	Gjsn_Koli.b	Max_Koli.b
1	Z08042213295130332 52NQTQA	0.00	0.00	0.00	7439.17	39.00
2	Z08042811013014701 58IDBUT	0.00	0.00	0.00	0.00	0.00
3	Z08050713170952201 14EEKWW	11166.04	6.00	0.00	32146.42	19.00
4	Z08050722582044411 89TLXPI	0.07	1.00	0.00	1.25	8.00
5	Z08050722582044611 89JEJBW	0.10	1.00	0.00	2.33	8.00
6	Z08050722582045111 89CLSAF	0.20	2.00	0.00	7465.26	29.00
7	Z08050722582045311 89ZEGTS	0.00	0.00	0.00	6385.14	25.00
8	Z08050722582158811 90BGXPG	0.28	3.00	0.00	3733.62	200.00
9	Z08050722582345311 94KTQBB	59.10	490.00	0.00	169.77	2400.00
10	Z08050722582536811 94FFYEL	0.00	0.00	0.00	0.00	0.00

verdien av max-verdi for *E. coli*, kolforme bakterier og turbiditet, minimumsverdien for min-verdi av *E. coli*, koliforme bakterier og turbiditet og gjennomsnittsverdien av *E. coli*, koliforme bakterier og turbiditet for hvert inntakspunkt med funksjonen setDT()[] fra pakken data.table. I tabell 2.3 kan man se et utdrag hvor perioden ikke er en variabel lenger, og hvert inntakspunkt har kun én rad. Datatabellen blir så gjort om til en tibble med funskjonen as_tibble() fra pakken tibble.

Tabell 2.4: Utdrag fra ferdig inntaksdatatabell

	mtid_ip	vannkildetype
1	Z08032514073135401 16IYKIB	Innsjø
2	Z08040910155710001 71ZIHTZ	Borebrønn løsmasse
3	Z08050713170952201 14EEKWW	Innsjø
4	Z08050722582158811 90BGXPG	Innsjø
5	Z08050722582345311 94KTQBB	Elv/bekk
6	Z08050722583306711 89ZDEBA	Innsjø
7	Z08050722583899411 95ACMMC	Innsjø
8	Z08050722583938611 95SNOIT	Innsjø
9	Z08050722584437111 95GYTJD	Innsjø
10	Z08050722584526011 95YHHCL	Innsjø

I datasettet for inntakspunkter blir unødvendige variabler fjernet med funksjonen subset() samtidig som kun inntakspunkter som er aktive og tilhører en hovedkilde blir tatt vare på. Deretter blir funksjoenen gsub() som er innebygd i R brukt til å fjerne parenteser fra kolonneverdiene i vannkildetyper. Funksjonen stri_replace_all_regex() blir brukt til å endre flere tegn og bokstaver slik at de blir lesbare. Etter dette fjernes variablene ”aktiv” og ”vannkildefunksjon” siden disse variablene ikke skal brukes videre i

oppgaven. I tabell 2.4 vises et utdrag av de 10 første radene fra dette datasettet.

Tabell 2.5: Utdrag fra ferdig datasett

	mtid_ip	vannkildetype	Gjsn_E.coli	Max_E.coli	Min_E.coli
1	Z08050713170952201l4EEKWW	Innsjø	11166.04	6.00	0.00
2	Z0805072258215881190BGXPG	Innsjø	0.28	3.00	0.00
3	Z0805072258234531194KTQBB	Elv/bekk	59.10	490.00	0.00
4	Z0805072258330671189ZDEBA	Innsjø	0.30	14.00	0.00
5	Z0805072258389941195ACMMC	Innsjø	0.42	4.00	0.00
6	Z0805072258393861195SNOIT	Innsjø	3435.08	8.00	0.00
7	Z0805072258443711195GYTJD	Innsjø	0.02	2.00	0.00
8	Z0805072258452601195YHHCL	Innsjø	0.14	2.00	0.00
9	Z0805072258455811189MNEGQ	Innsjø	3437.12	6.00	0.00
10	Z0805072258455821189KSSML	Innsjø	3437.24	6.00	0.00

Tabellene kombineres med hverandre med funksjonen left_join() fra pakken dplyr. Dette gjøres gjennom nøkkelen mtid_ip. For å kunne gjennomføre enkelte av analysene i R må rader med manglende verdier fjernes. Celler som inneholder verdier som -Inf, Inf og NaN blir gjort om til NA-verdier med dplyr-funksjonen na_if. Rader med manglende verdier blir etter det fjernet med funksjonen na.omit(). Tabell 2.5 viser et utdrag fra det ferdige datasettet.

2.2 Metode for analyse

Det ble utført en PCA på det ferdige datasettet med den innebygde funksjonen prcomp() i R. Resultatene fra anlysen ble hentet ut og plottet med autoplot()-funksjonen.

Datasettet ble sendt til Knut Kvaal for å gjøre en clusteranalyse. Datasettet ble lagt inn i MATLAB og PLS/Toolbox ble brukt til å redusere datasettet til 500 representative datapunkter slik at en analyse ble lettere å gjennomføre. En clusteranalyse med Ward's metode ble gjennomført. Disse resultatene, samt det reduserte datasettet ble sendt tilbake for videre analyse. Det ble gjennomført en PCA på samme måte som for hele datasettet.

3. Resultater

I denne delen vil resultatene fra PCA og clusteranlyse bli presentert. De viktigste funnene fra analysene vil bli presentert og tolkning av resultatene vil bli gitt i kapittel 4.

3.1 Resultater fra analyse av hele datasettet

I figur 3.1 ser man resultatene fra analysen. Her ser man at tre komponenter beskriver

```
> summary(pca_full)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     1.2588 1.2424 1.0924 1.0385 1.0005 0.91827 0.86233 0.75971
Proportion of Variance 0.1761 0.1715 0.1326 0.1198 0.1112 0.09369 0.08262 0.06413
Cumulative Proportion  0.1761 0.3476 0.4802 0.6000 0.7112 0.80491 0.88753 0.95166
                           PC9
Standard deviation     0.65959
Proportion of Variance 0.04834
Cumulative Proportion  1.00000
```

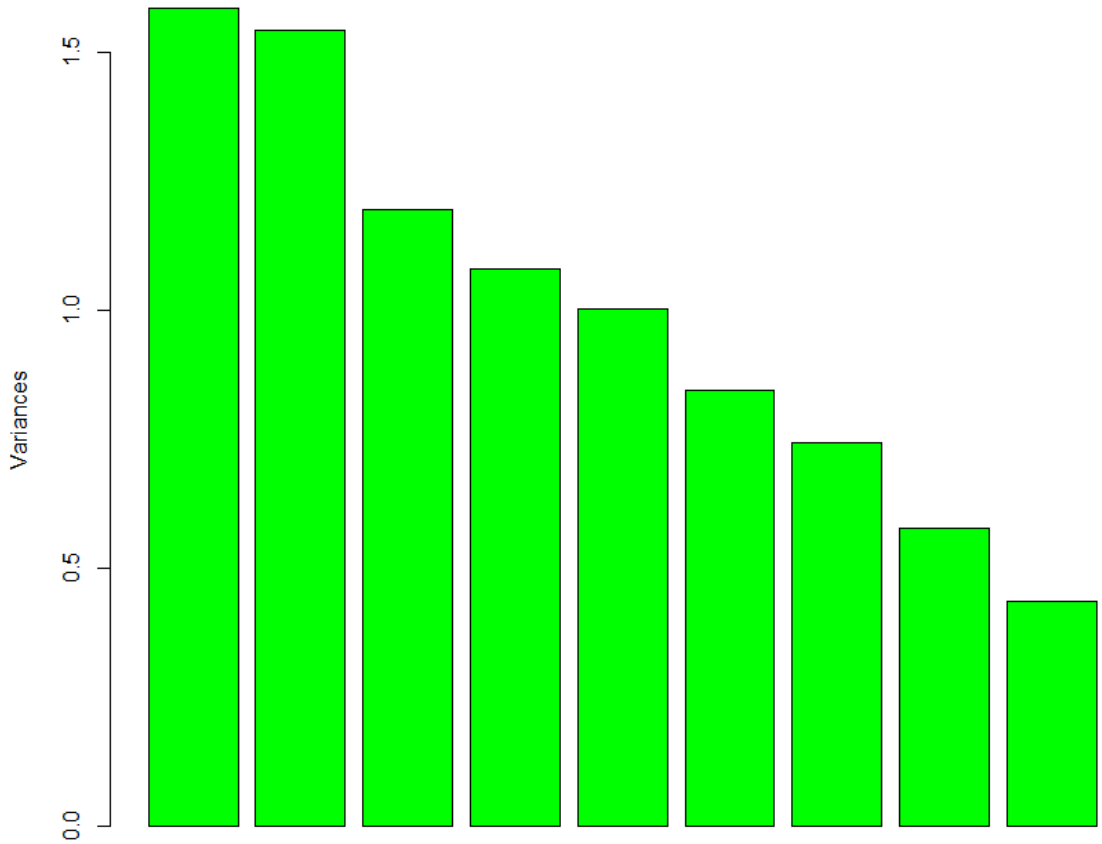
Figur 3.1: R-utskrift av standardavvik, proposjon av varians forklart av de ulike komponentene samt kumulativ proposjon av varians for hele datasettet fra VREG

ca. 48 % av variansen i datasettet. For å beskrive over 90 % av variansen i datasettet trenger man 8 komponenter. Komponent 1 beskriver 17,61 % av variansen og komponent 2 beskriver 17,15 % av variansen. I figur 3.2 ser man en grafisk framstiling av Proportion of Variance fra figur 3.1.

I figur 3.3 ser man hvor mye hver variabel bidrar med til hver komponent. Her ser man at maksimumsverdi av *E. coli* og maksimumsverdi av koliforme bakterier er de to variablene som har størst påvirkning på PC1 med verdiene 0,3962 og 0,3798 hhv. Minimumsverdi av *E. coli* og koliforme bakterier har størst påvirkning i negativ retning med verdiene -0,4811 og -0,4576 hhv.

For PC2 har minimumsverdi av *E. coli* og minimumsverdi av koliforme bakterier størst påvirkning med verdiene 0,5162 og 0,5373 hhv. Det er kun én verdi som er negativ i PC2 og det er verdien for gjennomsnittlig tubiditet, men den er nær 0.

Gjennomsnittsverdi av *E. coli* og gjennomsnittsverdi av koliforme bakterier har størst



Figur 3.2: Plot av fordeling av variasjon som er forklart av hver komponent

påvirkning på PC3 med verdiene 0,4582 og 0,4425 hhv. Maksimumsverdi av *E. coli* og koliforme bakterier har størst påvirkning i negativ retning med verdiene -0,4534 og -0,5002 hhv.

Grafisk framstilling av resultatene fra figur 3.1 og 3.3 for de tre første komponentene er i figur 3.4, figur 3.5 og figur 3.6.

I figur 3.4 ser man hvordan de ulike vannkildetypene ligger langs PC1 og PC2. Pilene med variablene viser hvordan disse påvirker komponentene. Hvis et punkt ligger nærme starten av pilen er verdiene for disse variablene lave. Hvis punktet ligger mot tuppen av pilen er verdiene for disse variablene høye.

Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier på venstre side, og av maksimumsverdi av *E. coli* og koliforme bakterier, samt gjennomsnittsverdi for *E. coli* og koliforme bakterier på høyresiden.

```
> pca_full$rotation
                  PC1         PC2         PC3         PC4         PC5
Gjsn_E.coli  0.303379726  0.319052067  0.458266850  0.23883115  0.029681995
Max_E.coli   0.396212787  0.352285347 -0.453443565 -0.04372280 -0.011319233
Min_E.coli  -0.481192401  0.516254322 -0.018456929 -0.02630954 -0.002545778
Gjsn_Koli.b  0.321643445  0.295100730  0.442527262  0.24037944 -0.030890848
Max_Koli.b   0.379860254  0.329272851 -0.500217495 -0.09571488 -0.019563465
Min_Koli.b  -0.457620000  0.537349616  0.007993112 -0.01082492  0.003820301
Gjsn_Turb   -0.001467717 -0.002813995  0.022530490  0.02239739 -0.996345288
Max_Turb     0.249324525  0.152517708  0.339376376 -0.48799740  0.045612108
Min_Turb    -0.010039221  0.007940818  0.149144267 -0.79665043 -0.053370478
                  PC6         PC7         PC8         PC9
Gjsn_E.coli  0.224261570 -0.60695558  0.332032423  0.096451099
Max_E.coli   0.095245708 -0.29836814 -0.641826730 -0.034967717
Min_E.coli   0.002044468  0.07839690 -0.073067112  0.699576272
Gjsn_Koli.b  0.199154125  0.67001527 -0.245648703 -0.077253110
Max_Koli.b  -0.039542476  0.28517941  0.635093075  0.035914479
Min_Koli.b  -0.080896710 -0.05337758  0.058365808 -0.699173549
Gjsn_Turb   -0.065217090 -0.04396725  0.008174521  0.004849558
Max_Turb    -0.742448725 -0.03230885 -0.067497556  0.048450887
Min_Turb     0.580841798  0.01901044  0.027537382 -0.039936615
```

Figur 3.3: R-utskrift av hvordan variablene påvirker hver komponent i analysen for hele datasettet

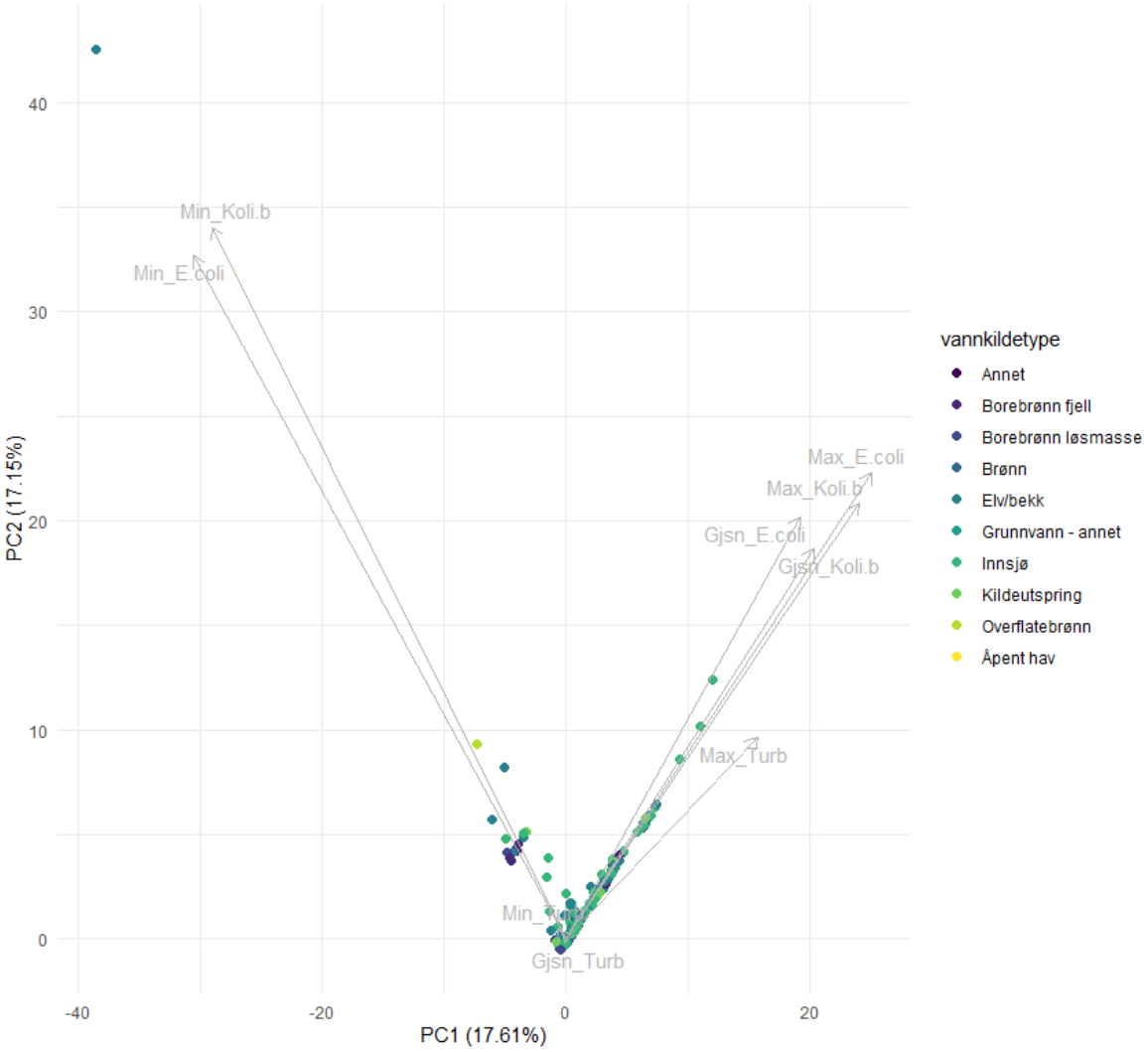
Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier øverst i plottet og av minimumsverdi og gjennomsnittsverdi av turbiditet nederst i plottet.

I figur 3.5 ser man hvordan de ulike vannkildetypene ligger langs PC2 og PC3. Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier på høyre side, og av minimumsverdi og gjennomsnittsverdi av turbiditet på venstresiden.

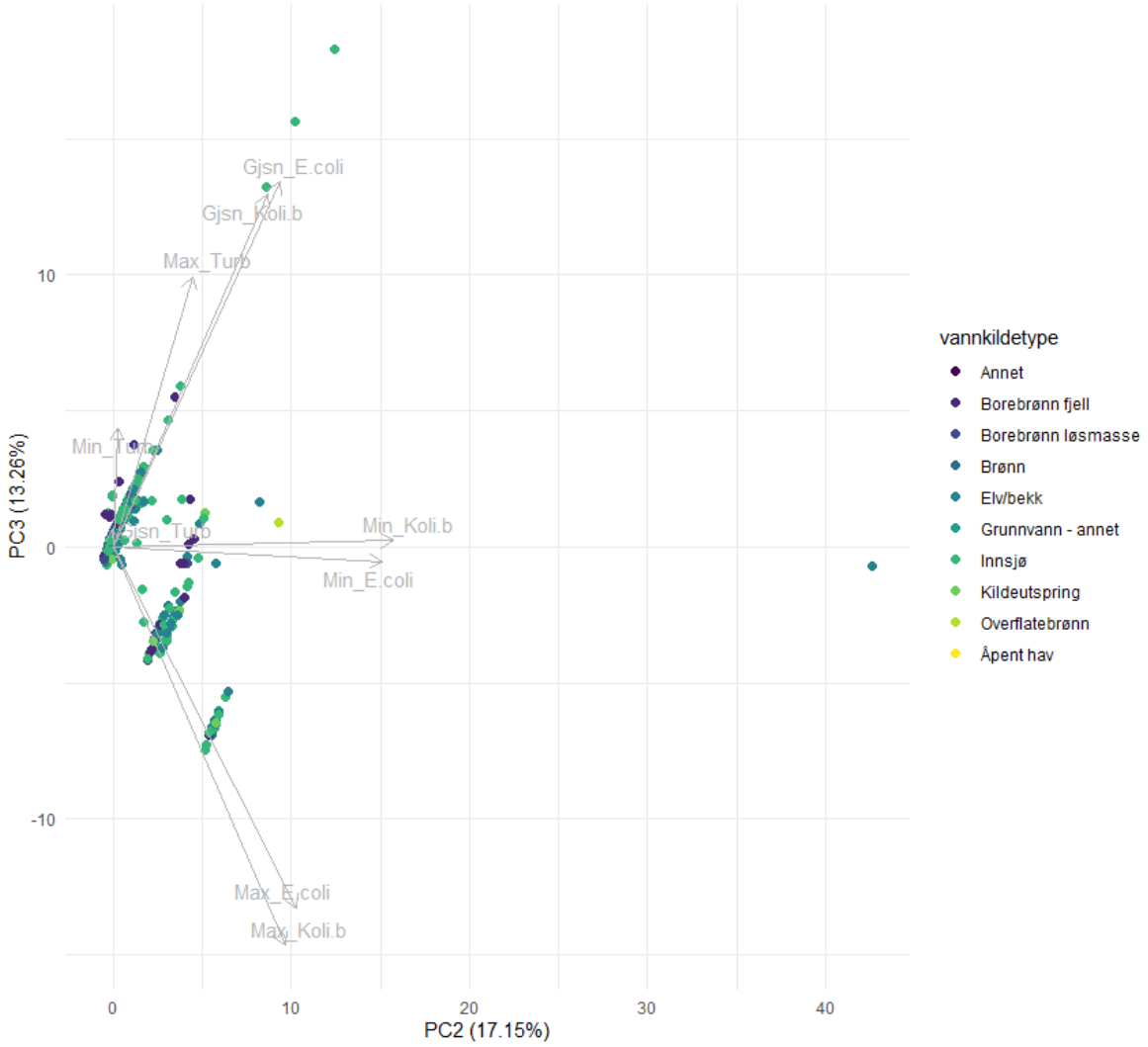
Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av gjennomsnittsverdi for *E. coli* og koliforme bakterier, samt maksimumsverdi for tubiditet øverst i plottet og av maksimumsverdi av *E. coli* og koliforme bakterier nederst i plottet.

I figur 3.6 ser man hvordan de ulike vannkildetypene ligger langs PC1 og PC3. Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier på venstre side, og av maksimumsverdi for *E. coli* og koliforme bakterier på høyresiden.

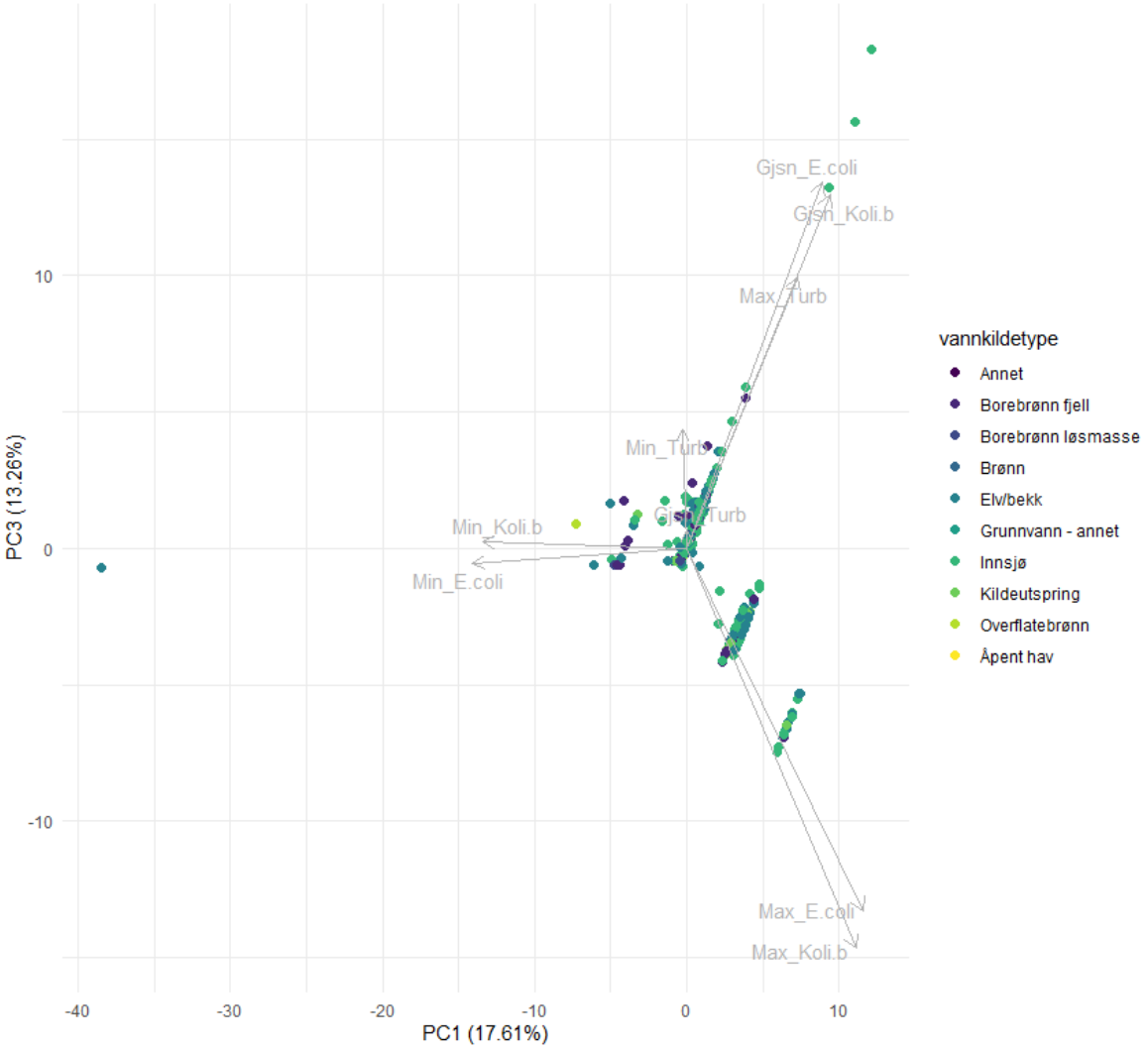
Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av gjennomsnittsverdi for *E. coli* og koliforme bakterier, samt maksimumsverdi for tubiditet øverst i plottet og av maksimumsverdi av *E. coli* og koliforme bakterier nederst i plottet.



Figur 3.4: Biplot av PC1 og PC2 for hele datasettet. PC1 langs x-aksen og PC2 langs y-aksen og vannkildetyper for hvert datapunkt er farget inn



Figur 3.5: Biplot av PC2 og PC3 for hele datasettet. PC2 langs x-aksen og PC3 langs y-aksen og vannkildetype for hvert datapunkt er farget inn



Figur 3.6: Biplot av PC1 og PC3 for hele datasettet. PC1 langs x-aksen og PC3 langs y-aksen og vannkildetype for hvert datapunkt er farget inn

3.2 Resultater av analyse for redusert datasett

I figur 3.7 ser man resultatene fra analysen. Her ser man at tre komponenter beskriver

```
> summary(pca_500)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6     PC7     PC8
Standard deviation     1.3223 1.1868 1.0673 1.0381 0.9839 0.9599 0.91943 0.77399
Proportion of Variance 0.1943 0.1565 0.1266 0.1197 0.1076 0.1024 0.09393 0.06656
Cumulative Proportion  0.1943 0.3508 0.4774 0.5971 0.7047 0.8070 0.90096 0.96752
                           PC9
Standard deviation     0.54063
Proportion of Variance 0.03248
Cumulative Proportion  1.00000
```

Figur 3.7: R-utskrift av standardavvik, proposjon av variasjon forklart av de ulike komponentene samt kumulativ proposjon av variasjon for redusert datasett

nesten 48 % av variasjonen i datasettet. For å beskrive over 90 % av variasjonen i datasettet trenger man 7 komponenter. De to første komponentene beskriver 19,43 % og 15,65 % av variasjonen hhv. I figur 3.8 ser man en grafisk framstiling av Proportion of Variance fra figur 3.7.

I figur 3.9 ser man hvor mye hver variabel bidrar med til hver komponent. Her ser man at maksimumsverdi av turbiditet og *E. coli* er de to variablene som har størst påvirkning på PC1 med verdiene 0,1682 og 0,1344 hhv. Minimumsverdi av *E. coli* og koliforme bakterier har størst påvirkning i negativ retning med verdiene -0,6833 og -0,6700 hhv.

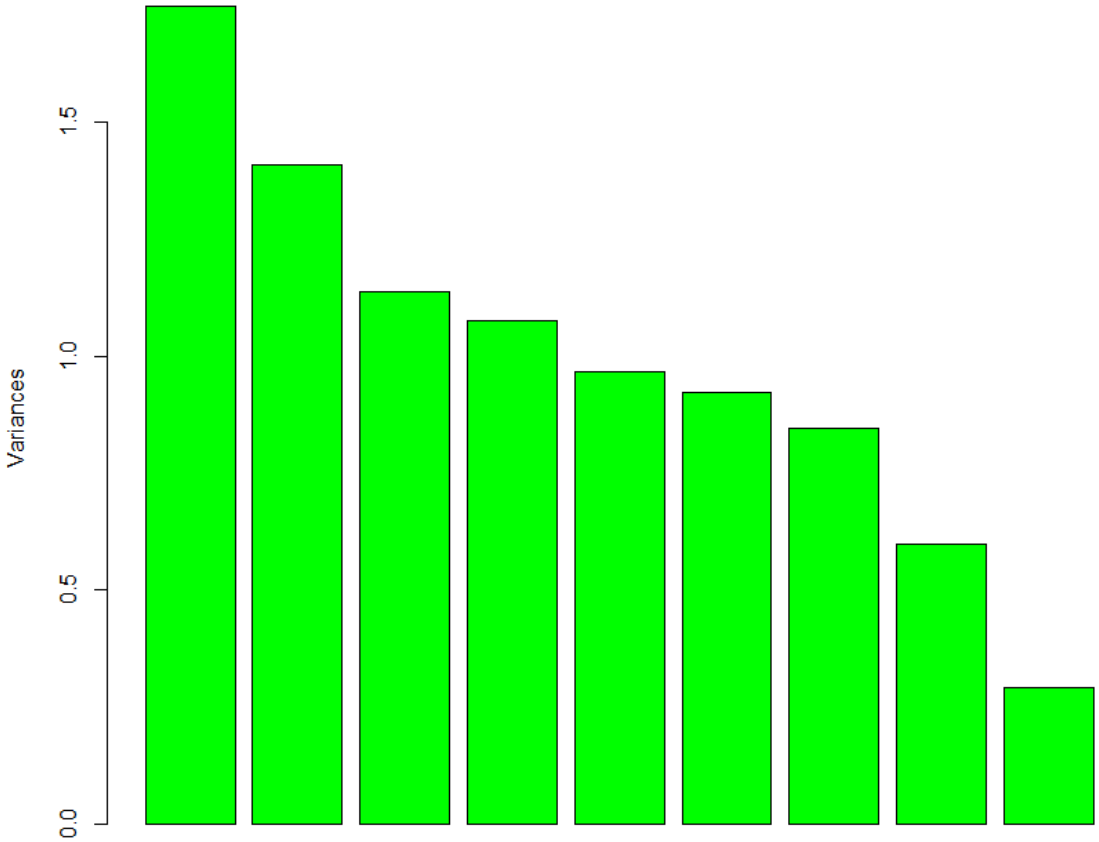
For PC2 har maksimumsverdiene av *E. coli* og koliforme bakterier størst påvirkning med verdiene 0,6419 og 0,6711 hhv. Gjennomsnittsverdien for *E. coli* og koliforme bakterier har størst påvirkning i negativ retning med verdiene -0,2705 og -0,1979 hhv.

Gjennomsnittsverdi av *E. coli* og gjennomsnittsverdi av koliforme bakterier har størst påvirkning på PC3 med verdiene 0,6180 og 0,5369 hhv. De to variablene med påvirkning i negativ retning er gjennomsnittsverdi av turbiditet og minimumsverdi av turbiditet med verdiene -0,2226 og -0,4122 hhv.

Grafisk framstilling av resultatene fra figur 3.7 og figur 3.9 for de tre første komponentene er i figur 3.10, figur 3.11 og figur 3.12.

I figur 3.10 ser man hvordan de ulike vannkidletypene ligger langs PC1 og PC2. Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier på venstre side, og av resten av variablene på høyre side.

Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av maksimumsverdi av *E. coli* og koliforme bakterier øverst i plottet og av gjennomsnittsverdi av *E. coli* og koliforme bakterier nedersti plottet.



Figur 3.8: Plot av fordeling av variasjon som er forklart av hver komponent

I figur 3.11 ser man hvordan de ulike vannkildetypene ligger langs PC2 og PC3. Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av gjennomsnittsverdi av *E. coli* og koliforme bakterier på venstre side, og av maksimumsverdi av *E. coli* og koliforme bakterier på høyresiden.

Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av gjennomsnittsverdi for *E. coli* og koliforme bakterier øverst i plottet og av minimumsverdi og gjennomsnittsverdi av turbiditet nederst i plottet.

I figur 3.12 ser man hvordan de ulike vannkildetypene ligger langs PC1 og PC3. Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier på venstre side, og av resten av variablene på høyresiden.

Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av gjennomsnittsverdi for *E. coli* og koliforme bakterier øverst i plottet og av minimumsverdi og gjennomsnittsverdi av turbiditet nederst i plottet.

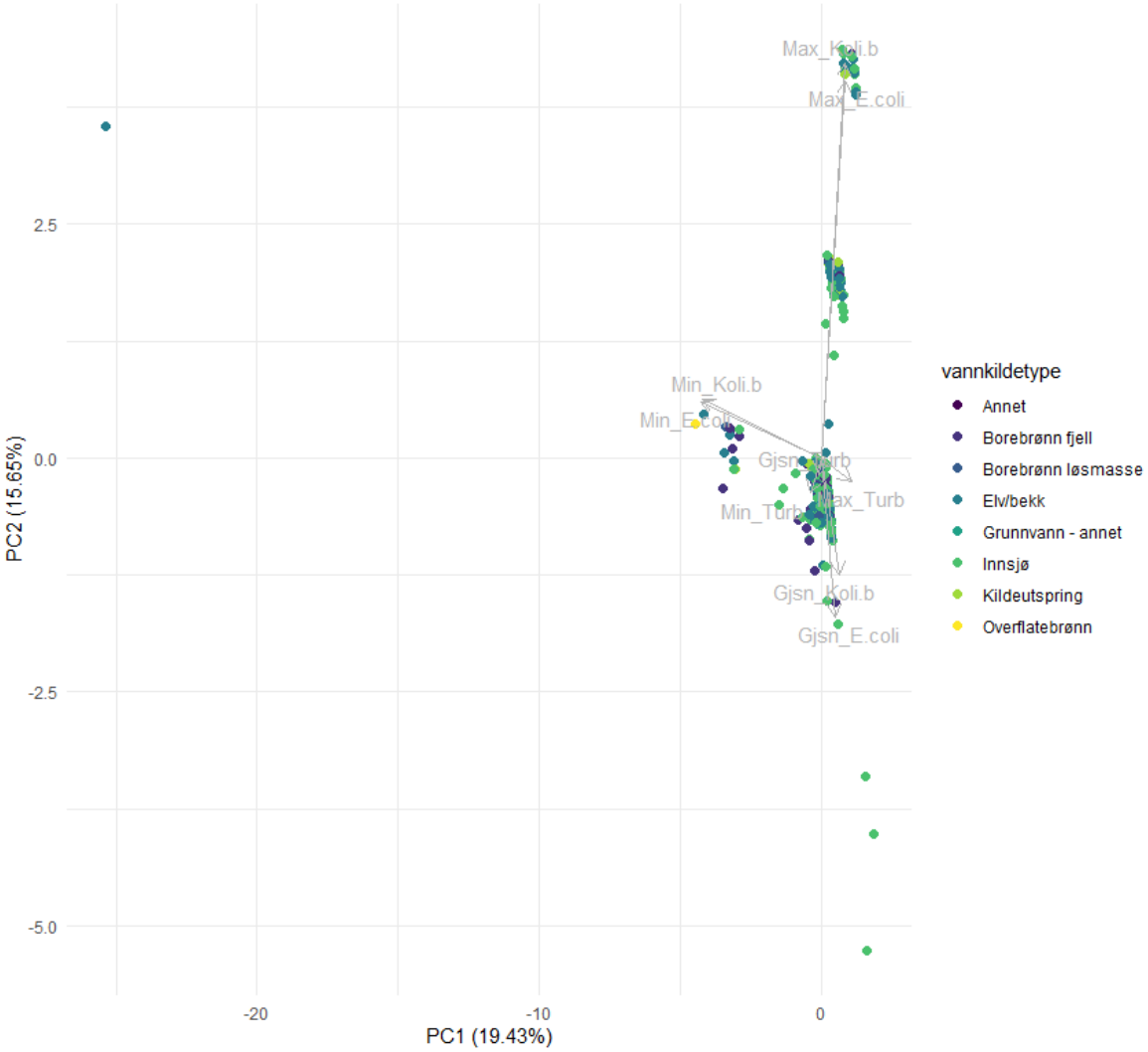
```
> pca_500$rotation
                    PC1         PC2         PC3         PC4         PC5         PC6
Gjsn_E.coli  0.07737193 -0.27052114  0.61802376 -0.09294852 -0.15334986 -0.24260340
Max_E.coli   0.13448182  0.64195384  0.21623372 -0.04466849 -0.13652077 -0.01659398
Min_E.coli  -0.68334565  0.09596309  0.10224487 -0.06401421 -0.07046664  0.02864075
Gjsn_Koli.b  0.10003144 -0.19794765  0.53692539  0.08599394 -0.34161818  0.58338425
Max_Koli.b   0.13287458  0.67117660  0.11849814 -0.01280544 -0.10115107  0.09937135
Min_Koli.b  -0.67000754  0.09942324  0.17812385 -0.07180333 -0.04586269 -0.11610039
Gjsn_Turb   -0.01171273 -0.02876982 -0.22266012  0.51300736 -0.75237382 -0.34434467
Max_Turb     0.16822131 -0.03970715  0.08507907 -0.66580781 -0.22823215 -0.50015335
Min_Turb    -0.06292300 -0.05961567 -0.41226306 -0.51582578 -0.45352637  0.45566453
                    PC7         PC8          PC9
Gjsn_E.coli -0.63000251 -0.21819164  0.057928838
Max_E.coli  -0.23316155  0.66913025 -0.006687671
Min_E.coli   0.04937301  0.01856848  0.707608409
Gjsn_Koli.b  0.43276180  0.09723235 -0.036736100
Max_Koli.b   0.07701239 -0.70092908  0.017940060
Min_Koli.b   0.05965132 -0.02149130 -0.696217096
Gjsn_Turb    0.03765871 -0.01268726  0.007891235
Max_Turb     0.46105742  0.03463053  0.059748941
Min_Turb    -0.36814500 -0.04164736 -0.076603036
```

Figur 3.9: R-utskrift av hvordan variablene påvirker hver komponent i analysen for redusert datasett

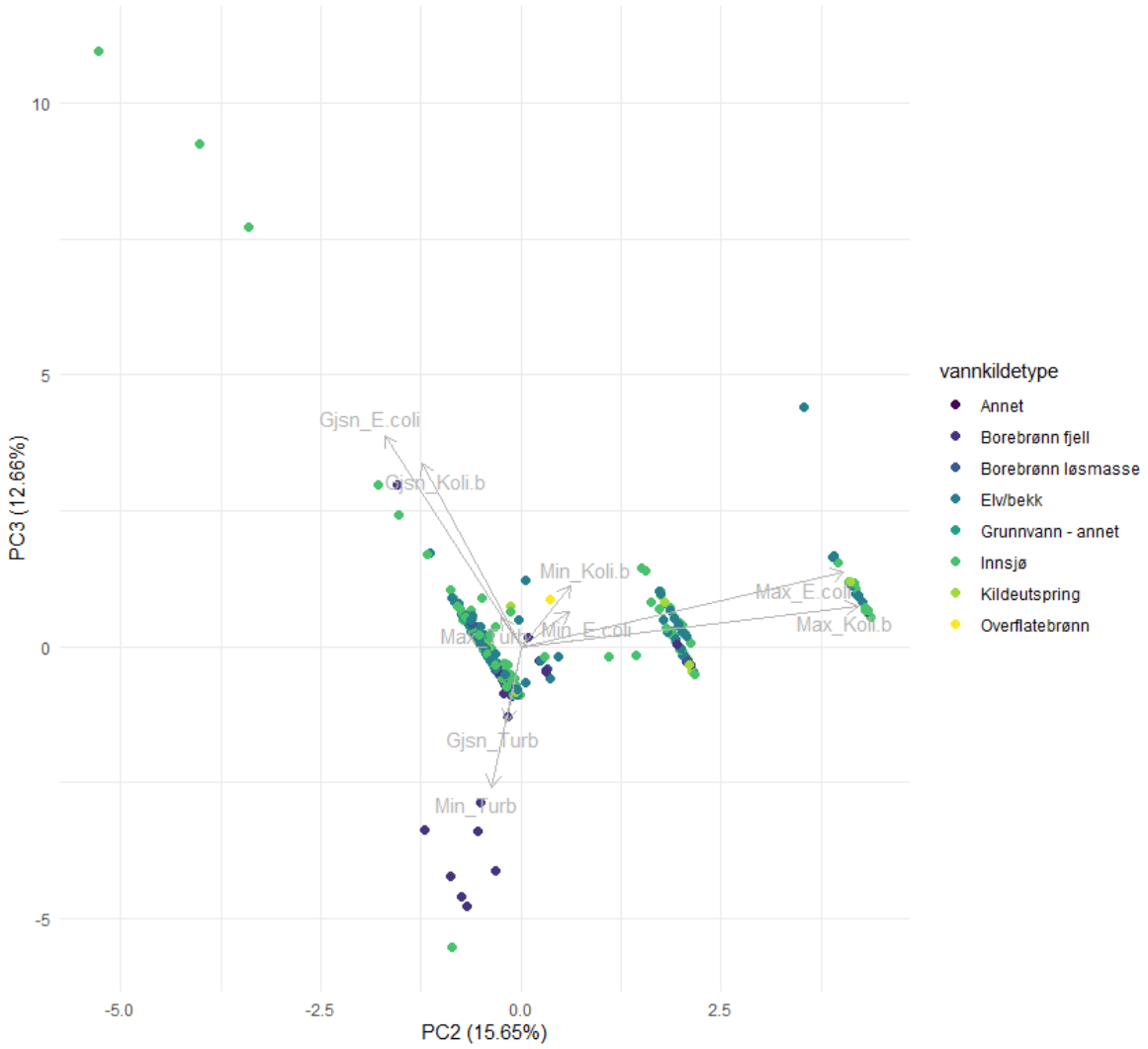
I figur 3.13 ser man at det gjennom clusteranalysen er blitt funnet fire grupper av vannkidletyper i det reduserte datasettet basert på maksimums-, minimums- og gjennomsnittsverdier for *E. coli*, koliforme bakterier og turbiditet. Disse fire gruppene kommer fra to hovedgrupper.

I figur 3.14 ser man hvordan de fire gruppene ligger i et scoreplot. Her ser det ut til at vannkildetypene som ligger i gruppen til høyre består i hovedsak av elv/bekk og innsjøer. Gruppen som ligger øverst i plottet består i hovedsak av borebrønner i fjell.

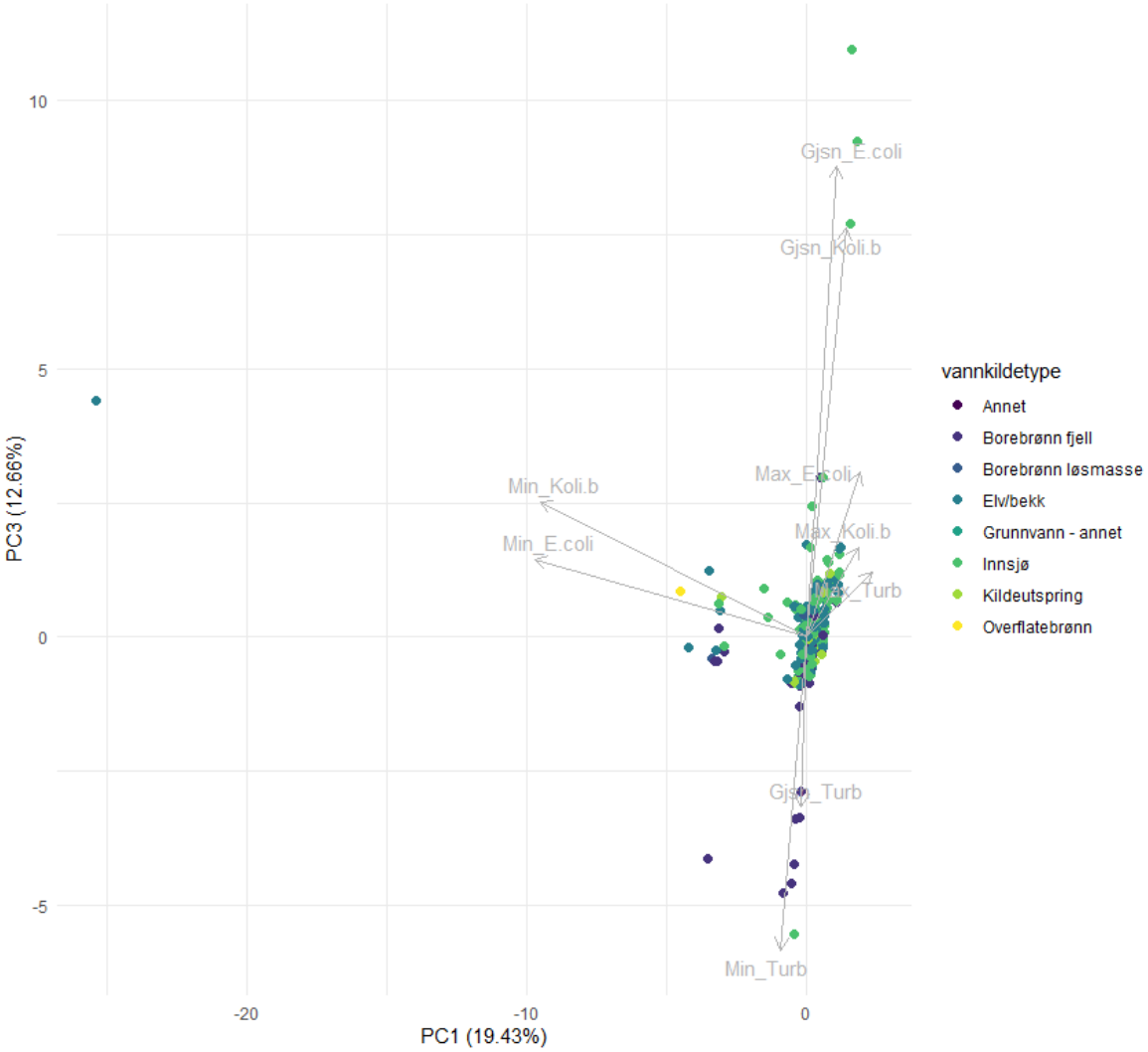
I figur 3.15 ser man hvordan variablene påvirker de ulike gruppene. F.eks. har gruppen til høyre i figur 3.14 høye verdier av maksimumsverdier av *E. coli* og koliforme bakterier og lave verdier av maksimumsverdier for turbiditet. Gruppen øverst i figur 3.14 har høye verdier av minimumsverdier for *E. coli* og lave verdier for gjennomsnittsverdier for *E. coli* og koliforme bakterier.



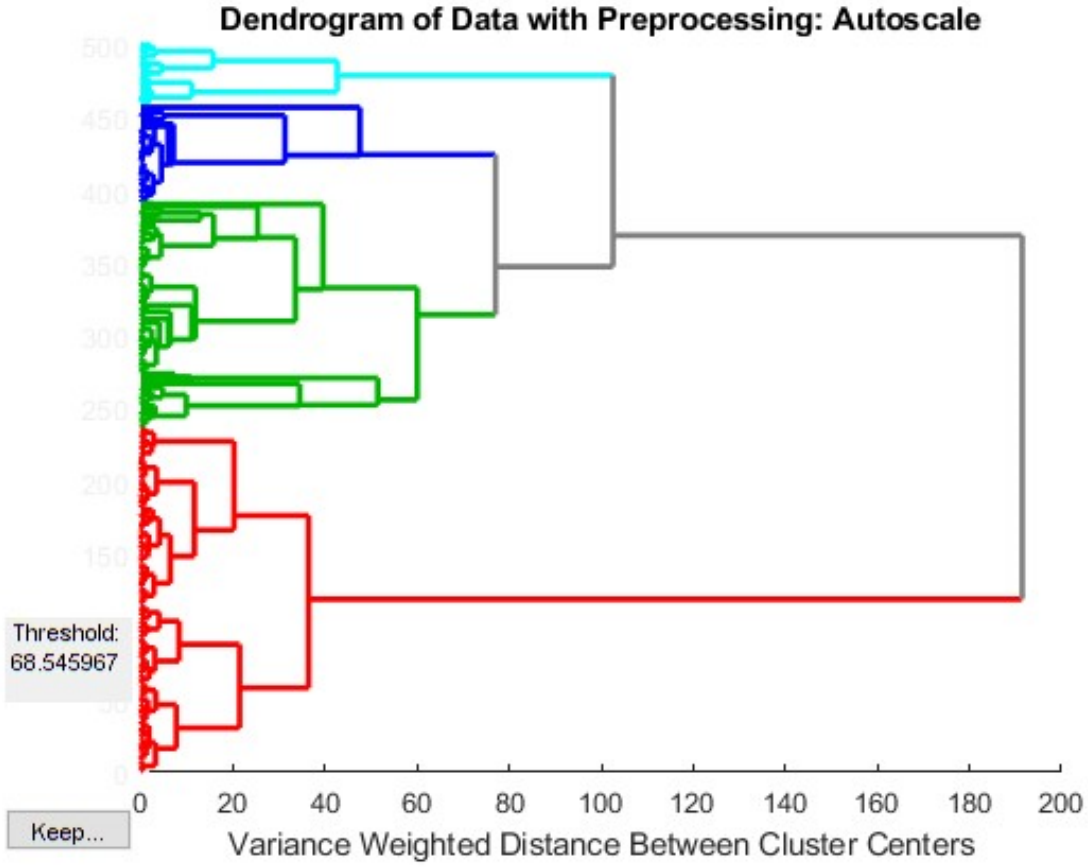
Figur 3.10: Biplot av PC1 og PC2 for redusert datasett. PC1 langs x-aksen og PC2 langs y-aksen og vannkildetyper for hvert datapunkt er farget inn



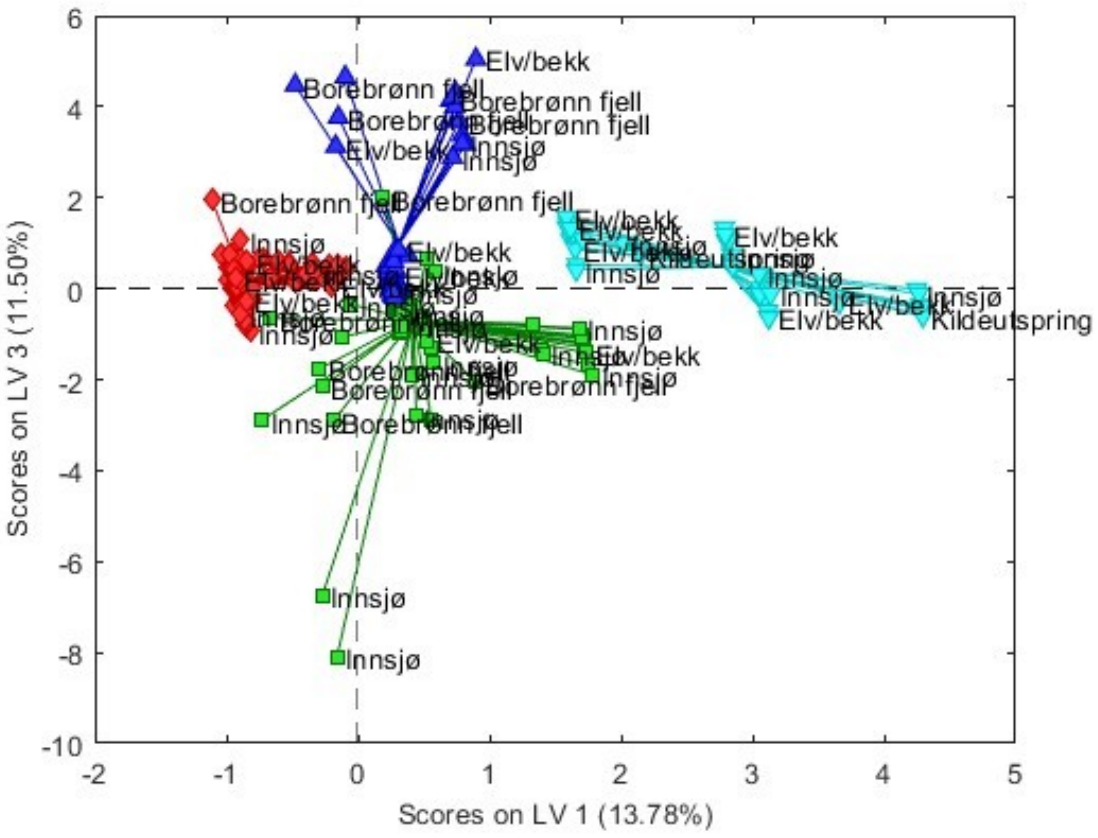
Figur 3.11: Biplot av PC2 og PC3 for redusert datasett. PC2 langs x-aksen og PC3 langs y-aksen og vannkildetyper for hvert datapunkt er farget inn



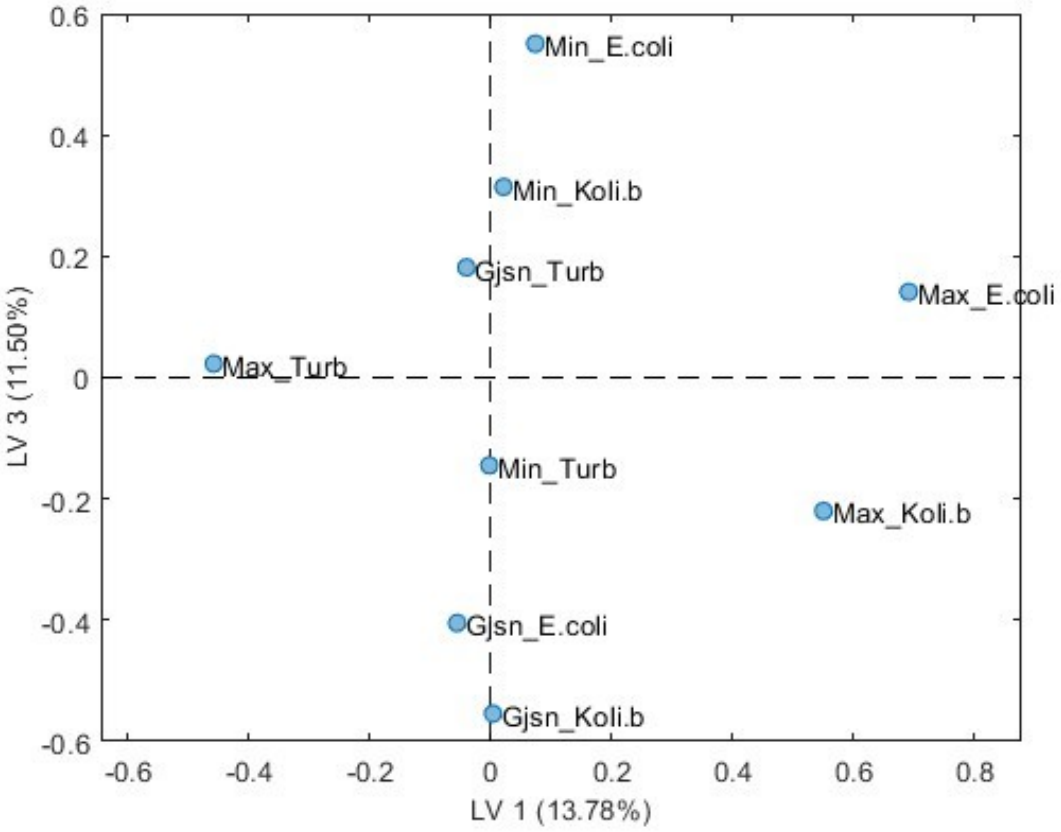
Figur 3.12: Biplot av PC1 og PC3 for redusert datasett. PC1 langs x-aksen og PC3 langs y-aksen og vannkildetyper for hvert datapunkt er farget inn



Figur 3.13: Dendrogram for clusteranalyse gjort på det reduserte datasettet med Ward's metode. Fargene representerer de ulike grupperingene i datasettet. Figuren er laget av Knut Kvaal.



Figur 3.14: Scoreplot av clusteranalysen for det reduserte datasettet som viser hvordan de fire gruppene ligger langs nivå 1 og nivå 3. Figuren er laget av Knut Kvaal.



Figur 3.15: Vekting for clusteranalysen som viser hvordan variablene påvirker gruppene. Figuren er laget av Knut Kvaal.

3.3 Resultater for data kun bestående av innsjøer og elver

I figur 3.16 ser man resultatene fra analysen. Her ser man at tre komponenter beskriver

```
> summary(pca_inel)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6     PC7     PC8
Standard deviation     1.345 1.2125 1.1113 1.0179 0.9997 0.9501 0.88934 0.74327
Proportion of Variance 0.201 0.1633 0.1372 0.1151 0.1110 0.1003 0.08788 0.06138
Cumulative Proportion  0.201 0.3644 0.5016 0.6167 0.7278 0.8281 0.91595 0.97734
                           PC9
Standard deviation     0.45164
Proportion of Variance 0.02266
Cumulative Proportion  1.00000
```

Figur 3.16: R-utskrift av standardavvik, proposjon av variasjon forklart av de ulike komponentene samt kumulativ proposjon av variasjon for kun innsjøer og elver

omtrent 50 % av variasjonen i datasettet. For å beskrive over 90 % av variasjonen i datasettet trenger man 7 komponenter. De to første komponentene beskriver 20,1 % og 16,33 % av variasjonen hhv. I figur 3.17 ser man en grafisk framstiling av Proportion of Variance fra figur 3.16.

I figur 3.18 ser man hvor mye hver variabel bidrar med til hver komponent. Her ser man at minimumsverdi av *E. coli* og koliforme bakterier er de to variablene som har størst påvirkning på PC1 med verdiene 0,6969 og 0,6946 hhv. Verdien som har størst påvirkning i negativ retning tilhører maksimumsverdi av turbiditet og er på -0,1203.

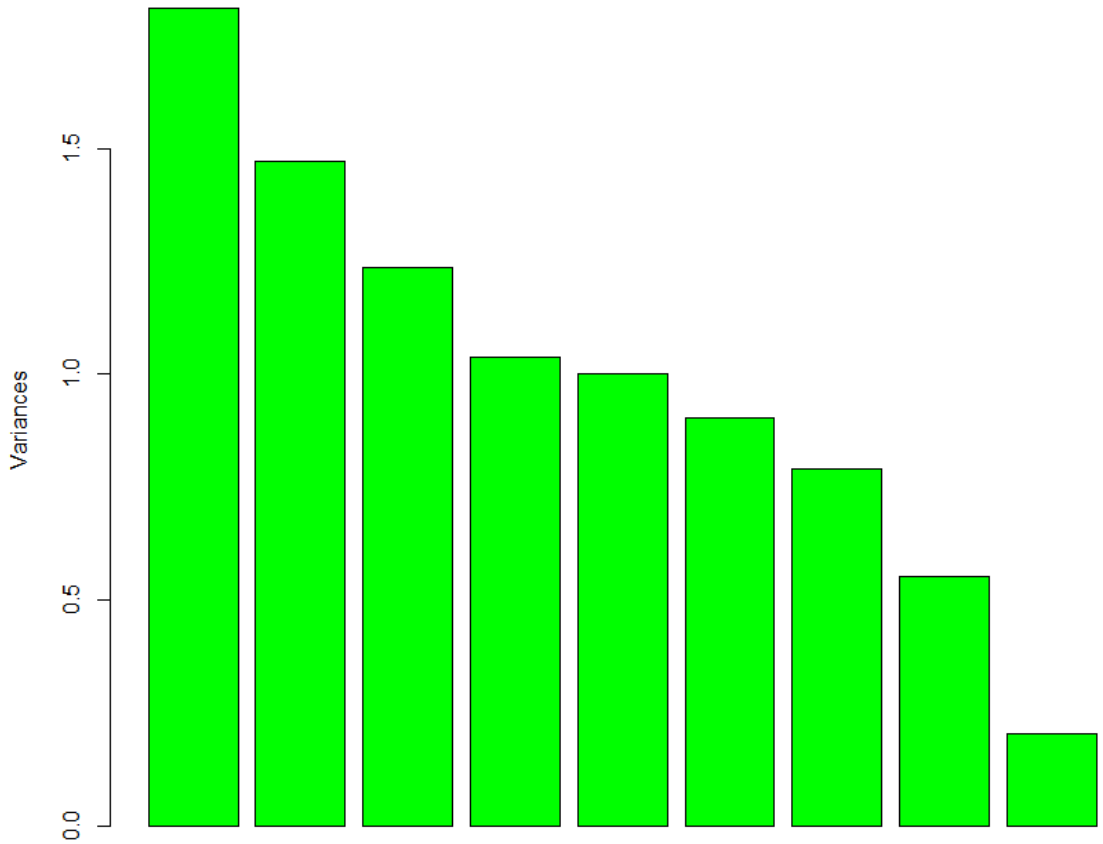
For PC2 har maksimumsverdi av *E. coli* og koliforme bakterier størst påvirkning med verdiene 0,6438 og 0,6437 hhv. Det er to verdier som påvirker PC2 negativt. Det er gjennomsnittsverdi og minimumsverdi av turbiditet, men disse verdiene er nær 0.

Maksimumsverdi av *E. coli* og koliforme bakterier har størst påvirkning på PC3 med verdiene 0,2706 og 0,2856 hhv. Gjennomsnittsverdi av *E. coli* og koliforme bakterier har størst påvirkning i negativ retning med verdiene -0,6278 og -0,5818 hhv. Grafisk framstilling av resultatene fra figur 3.16 og figur 3.18 for de tre første komponentene er i figur 3.19, figur 3.20 og figur 3.21.

I figur 3.19 ser man hvordan innsjøer og elver/bekker ligger langs PC1 og PC2.

Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier på høyre side, og av resten av variablene på venstresiden.

Vannkildetypene som ligger langt fra hverandre på y-aksen skilles av maksimumsverdi av *E. coli* og koliforme bakterier øverst i plottet og av minimumsverdi og gjennomsnittsverdi



Figur 3.17: Plot av fordeling av variasjon som er forklart av hver komponent

av turbiditet nederst i plottet.

I figur 3.20 ser man hvordan innsjøer og elver/bekker ligger langs PC2 og PC3. Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av maksimumsverdier av *E. coli* og koliforme bakterier på høyre side, og av minimumsverdi og gjennomsnittsverdi av turbiditet på venstresiden.

Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av maksimumsverier for *E. coli* og koliforme bakterier øverst i plottet og av gjennomsnittsverdi av *E. coli* og koliforme bakterier, samt maksimumsverdi av turbiditet nederst i plottet.

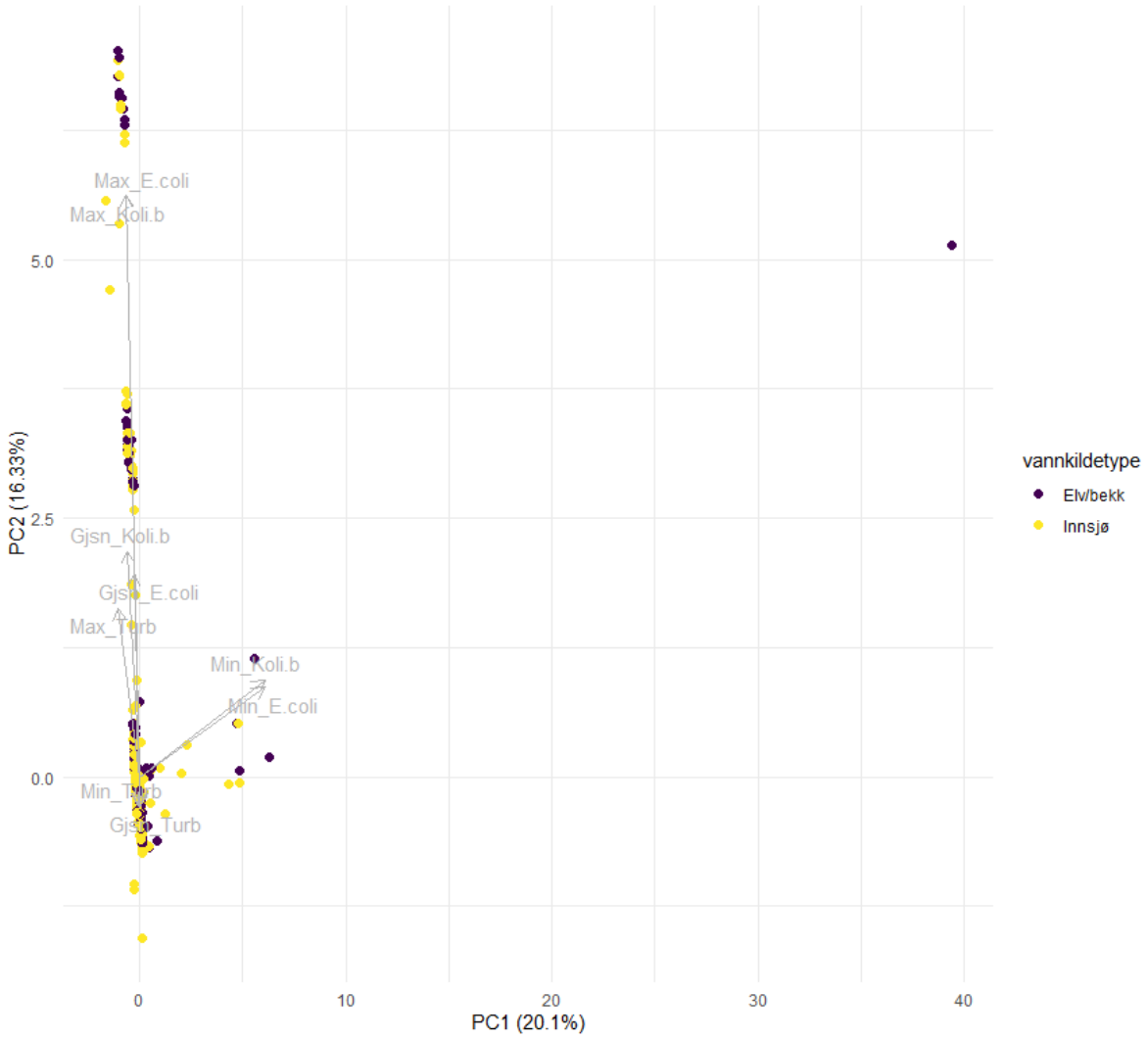
I figur 3.21 ser man hvordan de ulike vannkildetypene ligger langs PC1 og PC3. Vannkildetyper som ligger langt fra hverandre på x-aksen skilles av minimumsverdi av *E. coli* og koliforme bakterier på høyre side, og av resten av variablene på venstre side.

Vannkildetypene som ligger lagt fra hverandre på y-aksen skilles av maksimumsverdier

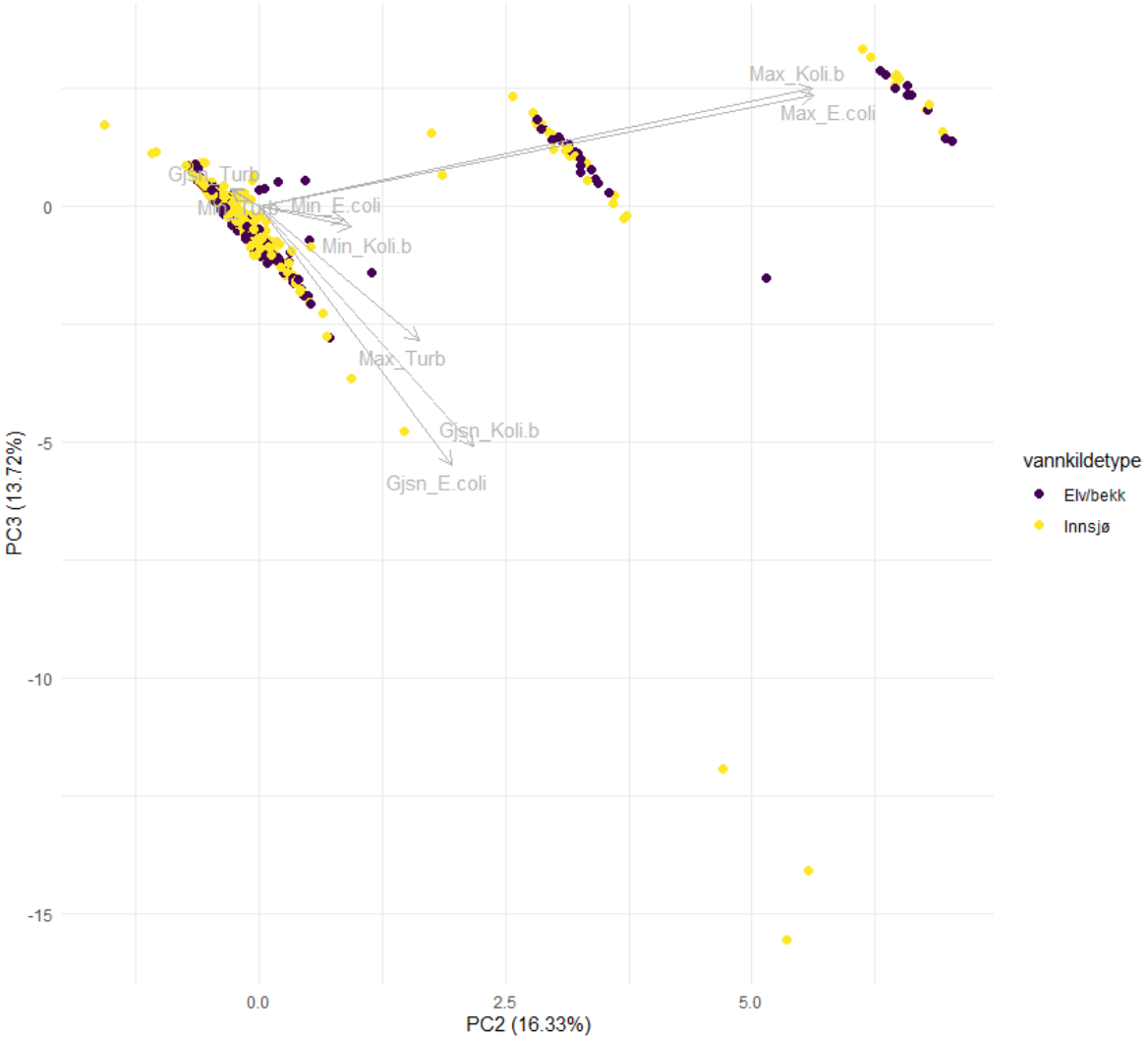
```
> pca_inel$rotation
                     PC1         PC2         PC3         PC4         PC5          PC6
Gjsn_E.coli -0.033636036  0.22455861 -0.62788831 -0.08253999 0.014498649 -0.212086755
Max_E.coli  -0.073891383  0.64388735  0.27068150 -0.02963679 0.008014225 -0.073575628
Min_E.coli   0.696915964  0.09911347 -0.03349957  0.02450433 0.008048996  0.027093042
Gjsn_Koli.b -0.069044831  0.24985670 -0.58183185 -0.17627070 0.070646257 -0.262191483
Max_Koli.b  -0.076526544  0.64371843  0.28569110 -0.00540194 0.014187864 -0.001059916
Min_Koli.b   0.694666122  0.10774274 -0.05031968  0.04301556 0.008630737  0.057545709
Gjsn_Turb    0.002781182 -0.03288857  0.04253099 -0.37261690 0.897619390  0.222705630
Max_Turb    -0.120329990  0.18614120 -0.32435870  0.48473071 0.034448136  0.776347891
Min_Turb    -0.006727262 -0.03191462  0.04084543  0.76482041 0.432997520 -0.473815314
                     PC7          PC8          PC9
Gjsn_E.coli -0.68569081  0.175182750  0.036518262
Max_E.coli  -0.18179531 -0.683554807 -0.005293341
Min_E.coli   0.05078703 -0.018101351  0.706439081
Gjsn_Koli.b  0.68609980 -0.133112091 -0.031902234
Max_Koli.b   0.13171477  0.693195680  0.007087939
Min_Koli.b  -0.01130297  0.006984024 -0.705610049
Gjsn_Turb   -0.05412002  0.002647436  0.002002659
Max_Turb     0.05252965 -0.058436291  0.024956736
Min_Turb     0.01736224  0.005148311 -0.001356501
```

Figur 3.18: R-utskrift av hvordan variablene påvirker hver komponent i analysen for kun innsjøer og elver

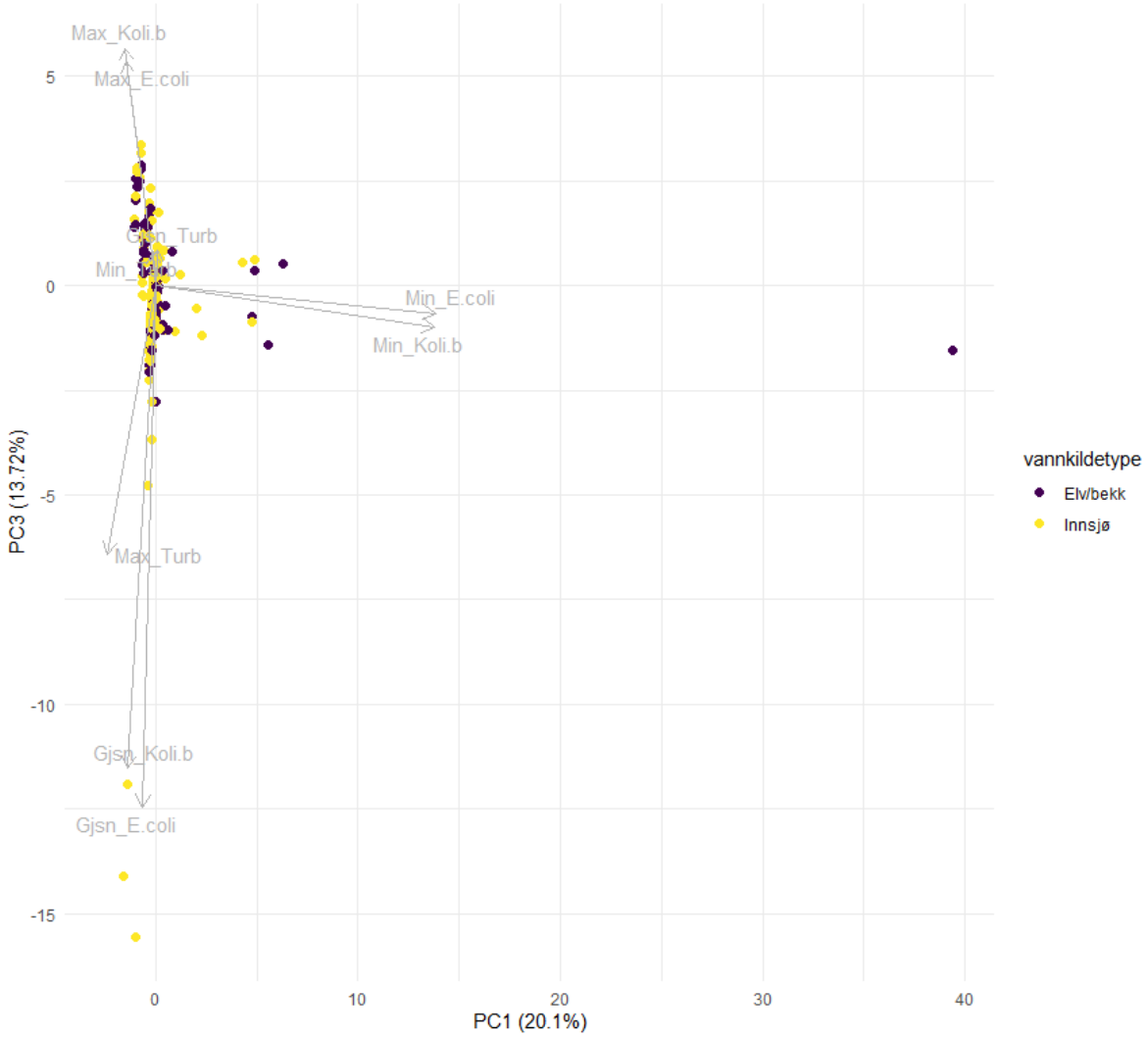
for *E. coli* og koliforme bakterier øverst i plottet og av gjennomsnittsverdier av *E. coli* og koliforme bakterier nederst i plottet.



Figur 3.19: Biplot av PC1 og PC2 for kun innsjøer og elver. PC1 langs x-aksen og PC2 langs y-aksen og vannkildetypene innsjø og elv/bekk er farget inn for hvert datapunkt



Figur 3.20: Biplot av PC2 og PC3 for kun innsjøer og elver. PC2 langs x-aksen og PC3 langs y-aksen og vannkildetyper for hvert datapunkt er farget inn



Figur 3.21: Biplot av PC1 og PC3 for kun innsjøer og elver. PC1 langs x-aksen og PC3 langs y-aksen og vannkildetyper for hvert datapunkt er farget inn

4. Diskusjon

Det ser ut til at det er vanskelig å kategorisere de norske råvannskildene inn i de svenske kildetypene ut fra grafer og utskrifter i kapittel 3. Man kan se seks ulike kategorier i tabell 2.1 fra SVU 2018-3 (Åström, 2018) som er vist i figur 1.2. I figur 3.13 er det fire grupper for det norske datasettet med råvannskilder.

Ved å bruke figur 3.14 sammen med figur 3.15 kan man se på kjennetegnene for hver enkelt gruppe. Deretter kan man sammenligne disse kjennetegnene med kjennetegnene for de svenske råvannskildene som man finner i figur 1.2. Ut fra resultatene av cluster-analysen kan man beskrive de fire gruppene som:

Turkis Høye maksimumsverdier for *E. coli* og koliforme bakterier, men lave maximums-verdier for turbiditet.

Blå Høye minimumsverdier for *E. coli*, men lave gjennomsnittsverdier for *E. coli* og koliforme bakterier.

Grønn Høye gjennomsnittsverdier for *E. coli* og koliforme bakterier, men lave mini-mumsverdier for *E. coli*.

Rød Høye maksimumsverdier for tubiditet, men lave maksimumsverdier for *E. coli* og koliforme bakterier.

Hvor mange grupperinger man får i clusteranalysen er avhengig av hvor man bestemmer at terskelen for inkludering skal gå. I denne oppgaven er den satt ved ca. 68,5. Dette gir fire clustere/grupper. Det som er sikkert er at det er minimum to grupper i det norske datasettet. Det er derfor en klar forskjell mellom gruppen som er markert som rød og de andre gruppene, men det er mulig å se en forskjell mellom de fire gruppene.

Beskrivelsen av de fire gruppene som er nevnt tidligere og kategoriene i den svenske tabellen har ikke noen tydelig sammenheng. Verdiene i figur 1.2 stemmer ikke godt med verdiene fra clusteranalysen. Det er allikevel ikke helt sikkert at det ikke er mulig å kategorisere de norske råvannskildene inn i de svenske kateogoriene. Dette er fordi det ikke er kvantitativt bestemt hva som er høye og lave verdier for hver variabel.

SVU 2018-3 (Åström, 2018) er kun beregnet på overflatekilder, men i denne oppgaven er grunnvannskilder også tatt med i analysen. Det ble derfor også gjennomført en PCA med kun overflatekilder. Resultatene fra figur 3.19, figur 3.20 og figur 3.21 gir ikke noe klart skille mellom innsjø og elv/bekk. Ved å se på kategoriene i figur 1.2 fra SVU 2018-3 (Åström, 2018) er det mulig å skille mellom elven som er kategori B og innsjøene som er resten av kategoriene.

En grunn til forskjellen mellom de norske råvannskildene og de svenske råvannskildene kan være pga. hvilke variabler som er brukt til å kategorisere. *E. coli*, koliforme bakterier, turbiditet, COD og nedbør er brukt for å klassifisere de svenske råvannskildene. Fra figur 1.3 kan man også se at *E. coli*, koliforme bakterier, COD og turbiditet trekker i samme retining, mens nedbør trekker i en annen.

I analysen for de norske råvannskildene er kun *E. coli*, koliforme bakterier og turbiditet brukt. Dette er fordi det ikke er pålagt å teste COD for drikkevann, noe som gir mangelfulle opplysninger i VREG om COD. Det er også vanskelig å koble nedbørsdata mot VREG siden man ikke har kunnskap om hvilke vanntilsigsområder inntakspunktene ligger i. Siden man ikke har nedbørsdata for de norske inntakspunktene er det vanskelig å vite om dette hadde hatt mye å si for resultatene.

De seks kildene som brukes i Åström (2018) er markert i figur 1.3. Det ser ut til at disse råvannskildene er valgt ut for å dekke de fleste svenske overflatekildene, men det ser ikke ut til at det er gjort noen klassifisering. Det er derfor vanskelig å vite om det ligger noen statistisk metode for valg av de spesifikke råvannskildene som skulle bli brukt.

PCA-resultatene gir ikke noen klare grupperinger for vannkildetypene. Tendensen er at overflatekidler har høye gjennomsnittsverdier og maksimumsverdier for *E. coli* og koliforme bakterier. Dette kan man se i alle tre PCA-plottene hele datasettet vist i figur 3.4, figur 3.5 og figur 3.6. I de samme figurene er det også en tendens til at enkelte grunnvannskilder har høye minimumsverdier av *E. coli* og koliforme bakterier.

PCA-resultatene for det reduserte datasettet viser heller ikke noen klare grupperinger ut fra vannkildetyper. Tendensen i det reduserte datasettet er at noen av grunnvannskildene har høye minimumsverdier og gjennomsnittsverdier for turbiditet. Overflatekilder ser ut til å ha høye gjennomsnittsverdier og maksimumsverdier for *E. coli* og koliforme bakterier.

Andelen forklart varians er forskjellig mellom R og MATLAB for det reduserte datasettet. En forklaring på dette kan være at R og MATLAB bruker forskjellige formler i koden som kan resultere i litt avvikende resultater.

Tidligere har det vært antatt at grunnvannskilder har lav mikrobiell forurensing. Fra

PCA-resultatene ser det ut til at forskjellen mellom mikrobiell forurensing i overflatekilder og grunnvanskilder ikke er så klar som disse tidligere antagelsene. Dette nevnes av Kvitsand og Fiksdal (2010) som også har brukt VREG som datagrunnlag. Innrapporteringen viser at norske grunnvannskilder har høyere konsentrasjon av mikroorganismer enn tidligere antatt.

All data i VREG kommer fra selvrapportering. Det er defor ikke noen kvalitetskontroll på dataene, og man vet ikke hvor godt dataene stemmer. Om man undersøker datasettet ser det ut til at ikke alle dataene er rapportert inn riktig. En av variablene i datasettet for råvannsanalyser er pH. Noen av disse pH-verdiene ligger utenfor 0-14 på pH-skalaen. Man kan også se at enkelte verdier av *E. coli* er veldig høye. Det er mulig at disse verdiene også er feilrapporteringer. I PCAene er det ikke tatt hensyn til at noen av verdiene kan være statistiske utliggere. Disse verdiene kan derfor ha hatt påvirkning på resultatet. I PCA-plottene kan man se at noen råvannskilder ligger langt unna de andre råvannskildene.

For clusteranalysen ble tre antatte statistiske utliggere fjernet. Begrunnelsen for ikke å inkludere disse ble tatt visuelt. Det er derfor ikke noen formell statistisk metode som ligger bak avgjørelsen.

Dataene i VREG kan være uoversiktlige. Det er derfor vanskelig å gjøre analyser med datasettene. Dette problemet er også nevnt av Steinberg mfl. (2021) i en rapport om vannforsyningssystemer i Norge på oppdrag fra Mattilsynet.

Screeplottene i figur 3.2, figur 3.8 og figur 3.17 viser visuelt hvor mye av variansen som er beskrevet av hver komponent. Disse screeplottene har en jevn reduksjon av forklart varians. Vanligvis ser man et drastisk fall i i forklart varians i disse screeplottene. Dette kan være et tegn på at en av antagelsene for å gjennomføre PCA kanskje ikke stemmer.

Grunnen til å gjennomføre en PCA er for å redusere dimensjonene i et datasett. Man ønsker derfor at hver komponent skal forklare så mye av variansen som mulig. Blant de norske råvannskildene må man ta med nesten alle komponentene for å beskrive 90% av variansen. Dette er kan være et tegn på at modellen ikke passer, eller at en antagelse om datasettet ikke stemmer. Dette ser man også ved at hver komponent forklarer lite av variansen.

Man kan sjekke om funnene fra analysen stemmer ved å teste dem på et eget datasett, f.eks. plukke ut et år og se om grupperingene fortsatt passer. Det er gjort analyser på både hele datasettet og det reduserte datasettet, men det er ikke gjort noen egne tester for å se om resultatene stemmer med hverandre.

En mulig måte å bruke SVU 2018-3 (Åström, 2018) på for å velge råvannskonsentrasjo-

ner er å trekke en tilfeldig statistisk fordeling for patogene mikroorganismer fra figur 1.4 og figur 1.5. Med denne metoden antar man at sannsynligheten er lik for hvilken råvannstype den norske kilden tilhører. Man utfører en Monte Carlo-simulering med en tilfeldig statistisk fordeling for hver råvannskilde. Monte Carlo-simuleringene gir intervaller for patogenkonsentrasjoner som brukes videre i en QMRA. Til slutt står man igjen med et intervall over hvor mange som blir syke av drikkevann i Norge. Dette vil ikke være en nøyaktig metode for å kunne gjennomføre en QMRA, men man kan få et overslag.

Et alternativ til å bruke tilfeldige statistiske fordelinger fra SVU 2018-3, er å lage disse fordelingene selv. Dette vil være ressurs- og tidkrevende. Siden det er lite informasjon om konsentrasjon av patogene mikroorganismer i norske råvannskilder vil dette ta tid. Det må også velges ut råvannskilder som kan gjennomføre analyser for patogene mikroorganismer.

En mulig måte å bedre kategoriseringen av norske råvannskilder på er ved å se nøyere på korrelasjonen mellom variablene fra råvannsanalysene i VREG. Det er antatt at det er en korrelasjon mellom variablene i datasettet. Dette er ikke sjekket ordentlig. Dette gjør det også mulig å systematisk teste forskjellige hypoteser med ulike statistiske modeller.

Geografisk beliggenhet kan påvirke råvannskvaliteten. Informasjonen i VREG er ikke blitt knyttet opp mot geografisk tilhørighet. Det har derfor ikke blitt tatt hensyn til om eventuelle forskjeller kan skyldes geografisk beliggenhet. Dette er også noe det er mulig å se nærmere på.

I et eventuelt videre arbeid er det også mulig å knytte mer av dataene fra VREG sammen, slik at det f.eks. er mulig å ta med lekkasjer og risiko knyttet til uventede avbrudd i vanndistribusjonen. Feil i distribusjonsnettet kan gi sykdom. Flere av utbruddene i Norge er fra distribusjonsnettet. Utbruddet i Askøy er et eksempel på dette.

Noe som også er mulig å se videre på er om råvannskvaliteten i Norge har endret seg i løpet av årene. I denne oppgaven ble det ikke tatt hensyn til årlige variasjoner. Verdiene som er valgt er gjennomsnittet av alle gjennomsnittsverdiene, den største maksimumsverdien fra alle årene og den minste minimumsverdien for alle årene. Dette er for å bevare dimensjonene til de høye og lave verdiene.

Det er uvisst hvor mange inntakspunkter som tilhører samme vannbehandlingsanlegg. Det er derfor mulig at noen av vannkildetypene fra resultatene er samme råvannskilde. I et videre arbeid er det mulig å bruke mer av VREG til å redusere antall inntak ned til antall råvannskilder. Sammenligning av inntakspunkter fra samme råvannskilde kan også gi et bilde på om det er store variasjoner i råvannskvalitet.

Det er også mulig å skille dataene mellom kommunale og private vannforsyninssyste-

mer. Dette er ikke gjort i denne oppgaven, men det er en mulighet i et senere arbeid. Datasettene for inntakspunkter og råvannskvalitet er kun brukt i denne oppgaven. Inntakspunkter som ble brukt er aktive.

5. Konklusjon

Det er vanskelig å kategorisere norske råvannskilder på en måte som passer til kildetypene i SVU 2018-3 (Åström, 2018). En mulig måte å kunne bruke SVU 2018-3 på er ved å trekke tilfeldige statistiske fordelinger fra tabellene vist i figur 1.4 og figur 1.5. Clusteranalysen kategoriserer de norske råvannskildene inn i fire grupper. Resultatene fra PCAene viser at grunnvannskildene er mer mikrobielt forurenset enn tidligere antatt. Det er derfor ikke mulig å kategorisere norske råvannskilder ut fra vannkildetyper.

En mulig måte å bruke SVU 2018-3 på (Åström, 2018) er ved å trekke tilfeldige statistiske fordelinger for hver råvannskilde og bruke disse til en Monte Carlo-simulering. Man kan da få et overslag for hvor mange som blir syke av drikkevann i Norge.

Det mangler mye kunnskap om norsk råvannskvalitet og da særlig konsentrasjoner av patogene mikroorganismer. For å kunne gjennomføre en QMRA trenger man konsentrasjonene for patogene mikroorganismer. Det er utbrudd i norske råvannskilder selv om de fleste vannforsyningssystemene i Norge leverer hygienisk trygt drikkevann. Hvor mange som blir syke er vanskelig å vite, men det antas at mange utbrudd ikke blir rapportert. Det hender også at det blir store utbrudd i norske vannforsyningssystemer.

For å kunne estimere sykdom i drikkevann må man få mer kunnskap om forholdene i norske råvannskilder. Forslag til videre arbeid er:

- Studere korrelasjonen mellom variablene i VREG og systematisk teste statistiske modeller på datasettene.
- Teste funnene fra denne oppgaven på et uavhengig datasett.
- Finne ut om råvannskvaliteten i Norge har forandret seg gjennom årene

Referanser

14.7 - Ward's Method (2022). Web Page. URL: https://online.stat.psu.edu/stat505/lesson/14/14.7.

Abdi, H. og Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (4): 433–459. DOI: 10.1002/wics.101.

Crawley, M. J. (2013). *The R book*. 2. utg. John Wiley Sons.

FHI (jan. 2019). *Utbrudd av Campylobacter, Askøy, juni 2019*. Report. URL: https://www.fhi.no/globalassets/dokumenterfiler/tema/utbrudd/utbrudd_askoy_web.pdf.

FHI (des. 2022). *Vannverksregisteret (VREG)*. Web Page. URL: https://www.fhi.no/ml/drikkevann/ovrige-artikler/om-vannverksregisteret-vreg/.

Forskrift om vannforsyning og drikkevann (drikkevannsforskriften) (2016). Statute. URL: https://lovdata.no/pro/SF/forskrift/2016-12-22-1868.

Grøndahl-Rosado, R. C., Yarovitsyna, E., Trettenes, E., Myrmel, M. og Robertson, L. J. (2014). A one year study on the concentrations of norovirus and enteric adenoviruses in wastewater and a surface drinking water source in Norway. *Food and environmental virology* 6 (4): 232–245. DOI: https://doi.org/10.1007/s12560-014-9161-5.

Guzman-Herrador, B., de Blasio, B. F., Lund, V., MacDonald, E., Vold, L., Wahl, E. og Nygård, K. (2016). Vannbårne utbrudd i Norge i perioden 2003–12. *Tidsskrift for Den norske legeforening*. DOI: 10.4045/tidsskr.15.0114.

Hyllestad, S. (22.02.2017 2017). *Drikkevatn*. Web Page. URL: https://www.fhi.no/nettpub/hin/smitte/drikkevann/.

Klein, J. (des. 2020). *Norovirus*. Web Page. URL: https://sml.snl.no/norovirus.

Klein, J. (24.01.2022 2022). *Virus*. Web Page. URL: https://sml.snl.no/virus.

Kvitsand, H. M. og Fiksdal, L. (2010). Waterborne disease in Norway: emphasizing outbreaks in groundwater systems. *Water Science and Technology* 61 (3): 563–571. DOI: https://doi.org/10.2166/wst.2010.863.

Li, P. og Wu, J. (2019). Drinking Water Quality and Public Health. *Exposure and Health* 11 (2): 73–79. DOI: 10.1007/s12403-019-00299-8.

Lieungh, F. T. (2021). Nasjonalt estimat over sykdom fra drikkevann i Norge basert på Kvantitativ Mikrobiell Barriereanalyse (QMRA). Masteroppgave.

Mattilsynet (2021). *Vannforsyningssystemer til lands*. Web Page. URL: https://www.mattilsynet.no/mat_og_vann/drikkevann/opplysninger_om_vannforsyningssystemer/vannforsyningssystemer_til_lands.36094.

Otterholt, E. (apr. 2021). *Giardia*. Web Page. URL: https://sml.snl.no/Giardia.

Petterson, S. R., Stenström, T. A. og Ottoson, J. (2016). A theoretical approach to using faecal indicator data to model norovirus concentration in surface water for QMRA: Glomma River, Norway. *Water research* 91: 31–37.

Sirevåg, R. (sep. 2019). *E. coli*. Web Page. URL: https://sml.snl.no/E._coli.

Sirevåg, R. (nov. 2022). *Bakterier*. Web Page. URL: https://sml.snl.no/bakterier.

Snelling, W. J., Matsuda, M., Moore, J. E. og Dooley, J. S. G. (2005). Campylobacter jejuni. *Letters in Applied Microbiology* 41 (4): 297–302. DOI: 10.1111/j.1472-765x.2005.01788.x.

Steinberg, M., Lyngstad, T. M. og Nordheim, C. F. (2021). Rapportering av data for vannforsyningssystemer i Norge for 2020. *Folkehelseinstituttet: Oslo, Norway*.

Sunnotel, O., Lowery, C. J., Moore, J. E., Dooley, J. S. G., Xiao, L., Millar, B. C., Rooney, P. J. og Snelling, W. J. (2006). Cryptosporidium. *Letters in Applied Microbiology* 43 (1): 7–16. DOI: 10.1111/j.1472-765x.2006.01936.x.

Säve-Söderbergh, M., Dryselius, R., Jacobsson, K., Simonsson, M. og Tjolander, J. (2014). *Microbiological risks in public drinking water production and distribution in Sweden: Raw water quality, source tracking and public health effects*. Conference Paper.

Tønjum, T. (21.02.2020 2020). *Parasitter*. Web Page. URL: https://sml.snl.no/parasitter.

WHO (2016). Quantitative microbial risk assessment: application for water safety management.

WHO (2022). *Drinking-water*. Web Page. URL: https://www.who.int/news-room/fact-sheets/detail/drinking-water.

Ødegaard, H., Lindholm, O., Mosevoll, G., Sægrov, S., Thorolfsson, S., Heistad, A. og Østerhus, S. (2014). *Vann-og avløpsteknikk, b. 2: Norsk vann*.

Åström, J. (2018). *Patogenhalter i svenska ytvattentäkter för QMRA Statistisk modellering och utvärdering av ett hypotesbaserat angrepssätt*. Report 3. URL: https://www.svensktvatten.se/contentassets/18ba7182a3024feca5655f94fbaa4cf6/svu-rapport-2018-03.pdf.

Vedlegg A. R-kode

Datasett

```
library(tidyverse)
library(readxl)
library(stringi)
library(data.table)
library(mixlm)
library(ggfortify)
```

```
anraa <- read_excel("D://Dokumenter/2022/Masteroppgave/Masteroppgave/data/inntakspunkt_analyse_excel.xl
```

```
inntak <- read_excel("D://Dokumenter/2022/Masteroppgave/Masteroppgave/data/inntakspunkt_excel.xlsx")
```

```
anraa$analysetype <- str_replace(anraa$analysetype, "\\s*\\([^\\)]+\\)",
    "")

anraa$analysetype <- stri_replace_all_regex(anraa$analysetype,
    pattern = c("Kimtall 22Â° C", "E. Coli", "Koliforme bakterier",
        "Intestinale enterokokker", "Clostridium perfringens",
        "Totalt organisk karbon", "Kjemisk oksygenforbruk, COD-Mn",
        "Trihalometaner - total", "Plantevernmidler - total",
        "1,2-dikloroetan", "Tetrakloreten og trikloreten", "KvikksÃ¸lv",
        "Polyaromatiske hydrokarboner", "Giardia lamblia", "Cryptosporidium parvum",
        "Microcystin-LR", "UV-transmisjon 1 cm", "Total P", "Total N",
        "Total indikativ dose", "UV-transmisjon 5 cm", "Temperatur - UTGÃ...TT KODE",
        "Kimtall 36Â° C - UTGÃ...TT KODE", "Plantevernmidler - enkeltvis - UTGÃ...TT KODE",
        "Klorofyll - UTGÃ...TT KODE", "Bacillus cereus", "Bacillus - UTGÃ...TT KODE",
        "Hydrokarboner, mineraloljer", "Patogene E.coli", "COâ‚‚-fritt",
        "PFAS-totalt", "Strontium-90", "Pseudomonas aeruginosa"),
    replacement = c("Kimtall", "E.coli", "Koliforme_bakterier",
        "Intestinale_enterokokker", "Clostridium_perfringens",
        "TOC", "COD", "Trihalometaner_tot", "Plantevernmidler_tot",
        "1_2_Dikloretan", "Tetrakloreten_trikloreten", "Kvikksølv",
        "Polyaromatiske_hydrokarboner", "Giardia_lamblia", "Cryptosporidium_parvum",
        "Microcystin_LR", "UV_transmisjon_1cm", "Tot_P", "Tot_N",
        "Tot_indikativ_dose", "UV_transmisjon_5cm", "Temp_UTGÅTT",
        "Kimtall_UTGÅTT", "Plantevernmidler_UTGÅTT", "Klorofyll_UTGÅTT",
        "Bacillus_cereus", "Bacillus_UTGÅTT", "Hydrokarboner_mineraloljer",
        "Patogene_E.coli", "CO2_fritt", "PFAS_tot", "Strontium_90",
        "Pseudomonas_aeruginosa"), vectorize = FALSE)
```

```
anraa <- anraa %>%
    mutate_at(c(7:10), as.numeric)
anraa$periode <- as.integer(anraa$periode)
```

```
ana1 <- subset(anraa, analysetype == "Turbiditet" | analysetype ==
    "E.coli" | analysetype == "Koliforme_bakterier", select = -c(mtid_vf,
    ant_krav, ant_analyser, verdi_median))

ana2 <- ana1 %>%
    pivot_wider(names_from = analysetype, values_from = c(verdi_max,
```

```
        verdi_min, verdi_gjsn))

ana3 <- setDT(ana2)[, list(Gjsn_E.coli = mean(verdi_gjsn_E.coli,
    na.rm = TRUE), Max_E.coli = max(verdi_max_E.coli, na.rm = TRUE),
    Min_E.coli = min(verdi_min_E.coli, na.rm = TRUE), Gjsn_Koli.b = mean(verdi_gjsn_Koliforme_bakterier
        na.rm = TRUE), Max_Koli.b = max(verdi_max_Koliforme_bakterier,
        na.rm = TRUE), Min_Koli.b = min(verdi_min_Koliforme_bakterier,
        na.rm = TRUE), Gjsn_Turb = mean(verdi_gjsn_Turbiditet,
        na.rm = TRUE), Max_Turb = max(verdi_max_Turbiditet, na.rm = TRUE),
    Min_Turb = min(verdi_min_Turbiditet, na.rm = TRUE)), by = mtid_ip]

as_tibble(ana3)
```

```
inn1 <- subset(inntak, aktiv == "ja" & vannkildefunksjon == "Hovedkilde",
    select = -c(mtid_vf, mtid_parent, vannforsyningssystem, inntakspunkt,
        kommunenr, kommune, periode, ant_dogn_egen_noedvannskilde,
        noedvann_inggaar_alt_kilde, ant_fastb_egen_noedvannskilde))

inn1$vannkildetype <- gsub("[()]", "", inn1$vannkildetype)

inn1$vannkildetype <- stri_replace_all_regex(inn1$vannkildetype,
    pattern = c("Ã¸", "Ã¦", "Ã¥", "Ã\230", "Ã©", "Ã...", "Â´S"),
    replacement = c("ø", "æ", "å", "Ø", "é", "Å", "'s"), vectorize = FALSE)
inn2 <- subset(inn1, select = -c(aktiv, vannkildefunksjon))
```

```
komb <- inn2 %>%
    left_join(ana3, by = "mtid_ip")

komb2 <- komb %>%
    dplyr::na_if(-Inf)

komb3 <- komb2 %>%
    dplyr::na_if(Inf)

komb4 <- komb3 %>%
    dplyr::na_if(NaN)

komb5 <- na.omit(komb4)  # For analyse med alle vannkildetyper uten utliggere
```

```
komb6 <- subset(komb5, vannkildetype == "Innsjø" | vannkildetype ==
    "Elv/bekk", )  # For analyse med kun innsjøer og elver uten utliggere
```

Analyse for alle vannkildetyper

```
pca_full <- prcomp(komb5[, 3:11], scale = TRUE)

summary(pca_full)

pca_full$rotation
```

```r
autoplot(pca_full,
         x = 1,                                # plot PC1 on the x-axis
         y = 2,                                # plot PC2 on the y-axis
         data = komb5,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```r
autoplot(pca_full,
         x = 2,                                # plot PC1 on the x-axis
         y = 3,                                # plot PC2 on the y-axis
         data = komb5,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```r
autoplot(pca_full,
         x = 1,                                # plot PC1 on the x-axis
         y = 3,                                # plot PC2 on the y-axis
         data = komb5,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```r
plot(pca_full, main = "", col = "green")
```

Analyse for innsjø og elv

```
pca_inel <- prcomp(komb6[, 3:11], scale = TRUE)

summary(pca_inel)

pca_inel$rotation
```

```
autoplot(pca_inel,
         x = 1,                                # plot PC1 on the x-axis
         y = 2,                                # plot PC2 on the y-axis
         data = komb6,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```
autoplot(pca_inel,
         x = 2,                                # plot PC1 on the x-axis
         y = 3,                                # plot PC2 on the y-axis
         data = komb6,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```
autoplot(pca_inel,
         x = 1,                                # plot PC1 on the x-axis
         y = 3,                                # plot PC3 on the y-axis
         data = komb6,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
```

```
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```
plot(pca_inel, main = "", col = "green")
```

Reduserte data

```
data500 <- read_excel("D://Dokumenter/2022/Masteroppgave/Masteroppgave/data/norsk_datasett_uten_na_500_
```

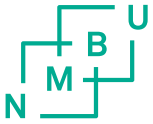
```
pca_500 <- prcomp(data500[, 5:13], scale = TRUE)

summary(pca_500)

pca_500$rotation
```

```
autoplot(pca_500,
         x = 1,                              # plot PC1 on the x-axis
         y = 2,                              # plot PC2 on the y-axis
         data = data500,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```
autoplot(pca_500,
         x = 2,                              # plot PC2 on the x-axis
         y = 3,                              # plot PC3 on the y-axis
         data = data500,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```r
autoplot(pca_500,
         x = 1,                                # plot PC1 on the x-axis
         y = 3,                                # plot PC3 on the y-axis
         data = data500,
         scale = 0,
         size = 2,
         loadings = TRUE,
         loadings.label = TRUE,
         loadings.label.size = 4,
         loadings.label.repel = TRUE,
         loadings.label.colour = "grey70",
         loadings.colour = "grey70",
         colour = "vannkildetype") +
  scale_colour_viridis_d() +
  theme_minimal()
```

```r
plot(pca_500, main = "", col = "green")
```



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway