# Semantiske Teknologier for Distribusjon og Visualisering av Statistikk

Resultatene av et sommerprosjekt for Semicolon-prosjektet ved Universitetet i Oslo, Institutt for Informatikk

**Håvard Mikkelsen Ottestad og Lars-Erik Bruce**

**16.09.2010**

## Forord

Vi har i sommer sett på mulighetene for å anvende W3C-standarder for å visualisere og distribuere innholdet i statistikkbanken til Statistisk Sentralbyrå (SSB). Med utgangspunkt i programpakkene *Anzo for Excel* og *Anzo on the Web*, utviklet av Cambridge Semantics, har vi brukt enkle ontologier og mange RDF-grafer for å representere statistikk. Under dette arbeidet har vi utarbeidet en enkel prosedyre for hvordan man kan hente inn, merke og laste opp statistisk materiale fra SSB og andre byråer og undersøkelser. Dette kan igjen benyttes for å lage visuelle representasjoner. Under hele arbeidet har hovedfokuset vært å benytte standardiserte teknologier, da spesielt OWL, RDF og TRIG.

Å benytte standardiserte teknologier, spesielt med tanke på dataformatet, gjør det betydelig lettere å distribuere data til ulike aktører, og benytte data for ulike typer representasjoner. Med verktøy som kan visualisere data som er semantisk lagret, som *Anzo on the Web*, blir det lett å for eksempel visualisere data for ulike fylker over forskjellige år, mens data i statistikkbanken er lagret på kommune og måned. Man trenger ikke lenger instruere datamaskinen om hva som skal summeres, med litt semantikk og mye data gjøres dette automatisk.

## Sammendrag

Formålet med oppgaven var å utforske nye teknologier og fremgangsmåter for å finne mer effektive måter å distribuere og visualisere innholdet i Statistisk Sentralbyrås statistikkbank. Denne rapporten forklarer hvordan man kan lage ontologier som beskriver forholdet mellom ulike statistikk-variabler, merke statistikken med vokabular fra ontologien, lagre den som RDF-tripler og visualisere statistikken i et web-grensesnitt, hvor sluttbrukeren selv kan bestemme hvilke data som skal sammenlignes.

Semantiske teknologier og Anzo fra Cambridge Semantics tillater brukeren å genere grafer som inneholder statistikk fra flere kilder så lenge denne statistikken har en variabel til felles, for eksempel kommuner. Brukeren kan selv velge hvilket nivå han ønsker å visualisere dataene på, for eksempel ved å snitte alle kommunene sammen til fylker. Dataene må kun være lastet opp på det laveste nivået. Anzo kan ikke ta høyde for at dataene er i prosent, og ved å snitte disse dataene blir resultatet galt.

Vi ser at Anzo har stort potensial for å gjøre tilgjengelig offentlig statistikk for sluttbrukere. Anzo vil gjøre det lettere å samarbeide om statistikk og metadata. Muligheten for å laste opp egen statistikk vil gjøre det enklere for blant annet forskere og bedrifter å bedrive forskning.

# Innhold

# 1 Innledning – problemstilling

Utgangspunktet for oppgaven var å visualisere sykefravær, med tilsvarende funksjonalitet lik den som allerede finnes hos SSB. Deretter skulle vi supplere dette med nye datasett om skilsmisser og separasjoner samt barnehagedekning, og vise hvordan man kan sammenligne disse tre datasettene på en enkel måte for å lete etter korrelasjoner. Videre lastet vi opp data om vannkvalitet, hvor vi også benyttet tall fra Folkehelseinstituttet, hvor også disse kunne sammenlignes med ovennevnte variabler. Til slutt så vi på muligheten for å benytte andre datakilder enn SSB, hvor vi har brukt både DIFIs befolkningsundersøkelse og data fra The World Bank. For flere detaljer, se Vedlegg 1: Studentoppgave for IFI, sommer 2010.

Vi mener å ha fått løst de essensielle momenter som ble etterspurt i studentoppgaven. Enkelte utfordringer har vi derimot ikke tatt oss tid til, som for eksempel visualisering av statistiske data på kart. Slike visualiseringer skal være forholdsvis enkelt å implementere i *Anzo on the Web*, uten at man trenger å gjøre endringer i datasettene som er lastet opp og vil dermed være en naturlig utvidelse av løsningen vi her representerer.

# 2 Hoveddel

Hoveddelen i denne sluttrapporten er delt opp i fire underkategorier, som et motstykke til de fire av de fem oppgavene vi har blitt bedt om å løse (vedlegg 1)[1]. Vi vil her gi en kort introduksjon til oppgaven, beskrive de trinnene som ble brukt for å løse den, og gi noen eksempler på visualisering fra *Anzo on the Web*. Vi vil ikke gå så langt i å beskrive de tekniske fremgangsmåtene, disse vil bli lagt ved som vedlegg til denne rapporten (Vedlegg 3).

## 2.1 Oppgave 1

### 2.1.1 Innledning

Første oppgave gikk ut på å visualisere sykefravær for arbeidstakere i prosent, basert på kommune, kjønn, alder og tid. Funksjonaliteten på vår løsning skulle være så lik som mulig den funksjonaliteten som vi finner hos SSBs hjemmesider[2]. Av de visualiseringene vi ikke har fått til enda kan man nevne kart og befolkningspyramide. Disse visualiseringsfunksjonene finnes ikke som standard ved *Anzo on the Web* og må utvikles for seg.

Når vi har fått lastet opp alle relevante data med hensyn til sykefravær har vi derimot en rekke visualiseringsmuligheter, mange som heller ikke finnes hos SSB. I tillegg har vi fått til en viktig funksjon som er helt fraværende hos SSB; aggregering. Hos SSB kan man vise søyler både for kommuner og fylker, noe vi også har fått til. Men SSB kan kun vise ett kvartal av gangen, vi kan vise gjennomsnittet av sykefravær for et år, eller alle år. Vi kan også summere opp hele landet, for deretter å filtrere ut uønskete fylker. Vi kan også definere egne geografiske soner, som landsdeler, og la Anzo regne ut verdiene basert på data som ligger på kommune-nivå. Anzo-serveren (*Anzo Collaboration Server*) summerer disse verdiene automatisk i det brukeren angir hva hun vil ha visualisert.

Et stort problem har vi derimot ved at Anzo-serveren bruker meget lang tid på å regne ut alle disse verdiene. Med et ikke så alt for stort datasett tok det 6 minutter å tegne en graf. Dette er selvfølgelig ikke ønskelig i et slutt-produkt, men mye kan gjøres for å effektivisere dette. Man

[1] Vi valgte å ikke løse oppgaven om å visualisere statistikk for kriminalitet (oppgave 3). Både fordi oppgaven var underspesifisert, og fordi vi ikke ville avdekket mer funksjonalitet og dermed ikke bidratt til noe mer for prosjektet som helhet.

[2] SSBs løsning kan man finne ved å søke etter tabell 03218 (Legemeldt sykefravær for arbeidstagere, etter kjønn og alder i prosent per kommune) på SSBs hjemmesider (www.ssb.no).

kan benytte seg av en kraftigere server og database, og versjon 2 av Anzo bruker en trippel database – i stedet for den opprinnelige SQL-databasen – som skal gi en hastighetsøkning på 50%.

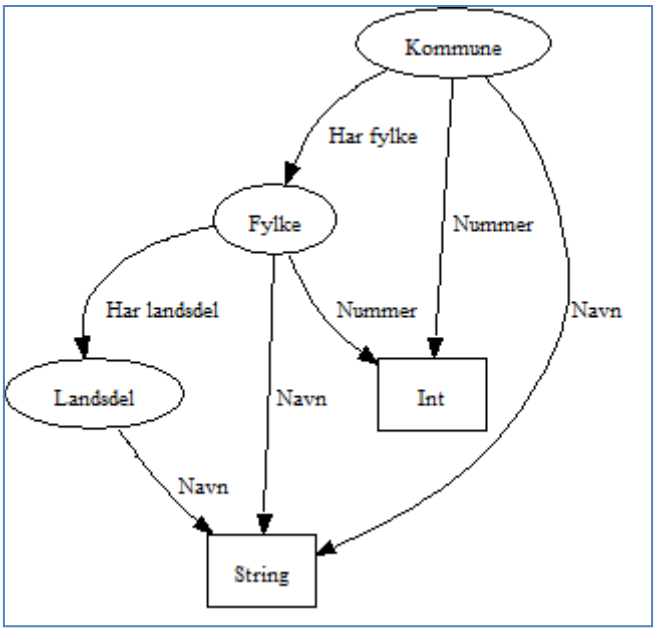
Anzo kan ikke regne ut prosent/promille/rate av reelle verdier, som f.eks. fødselsrate basert på antall fødsler og befolkningen i den tilhørende kommunen. Det beste Anzo kan gjøre er å snitte eller summere verdier. Ved å snitte prosentverdier blir den utregnede verdien feil hvis ikke alle verdiene allerede er vektet likt (for eksempel at alle kommuner har like mange innbyggere). Dette problemet diskuteres videre i del 3.1 av denne rapporten.

### 2.1.2 Gjennomføring av oppgaven

Første trinn for å gjennomføre oppgave 1 gikk ut på å anskaffe de dataene vi hadde bruk for. I og med at vi bruker *Anzo for Excel* til å laste data opp til Anzo-serveren er det også helt essensielt å få tak i dataene i Excel-formatet. Heldigvis har SSBs statistikkbank mulighet til å eksportere tabellene til Excel. Anzo har ikke mulighet for å tolke regneark med flere dimensjoner, så oppsettet av regnearket er også et viktig moment. Vi vil ha selve statistikkvariablene i én kolonne, mens kriterier som kommune, kjønn, alder og kvartal/år også burde ha hver sin kolonne. Også her har SSBs statistikkbank gode muligheter for å vri på tabellen som ønsket, ved hjelp av brukervalget Roter fritt. Vi kan her velge å ha statistikkvariablene under Hode, og alle andre kriterier under Forspalte, og få gode Excel-ark man kan laste opp i Anzo. En mer detaljert beskrivelse av hvordan man burde hente inn data fra SSBs hjemmesider finner du i vedlegg 3.

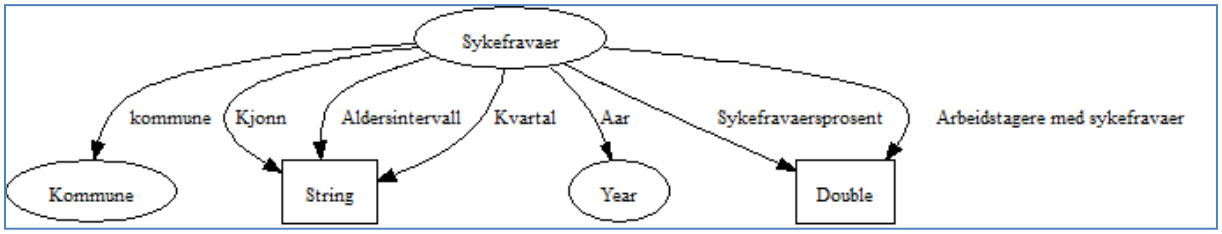
Før vi kan laste opp dataene til Anzo-serveren trenger vi å lage ontologier, som kan benyttes til å merke dataene. (For å lese mer om ontologier henviser vi til vedlegg 2.) Vi har her valgt å lage en enkel Geografi-ontologi, som forteller hvordan ulike geografiske begreper forholder seg til hverandre. De mindre geografiske enheter peker oppover mot de større, for eksempel tilhører en kommune ett fylke og ett politidistrikt. Ett fylke tilhører en landsdel. Se Figur 2.1-1[3].

[3] Grafene som viser oppbyggingen av ontologier i denne rapporten er kun ment veiledende, og reflekterer ikke nødvendigvis hvordan de ser ut den faktiske implementasjonen. Vi har forsøkt å holde oss til notasjonen hvor sirkel er klasse og firkant literal.




Figur 2.1-1 Geografi-ontologien, som beskriver kommune, fylke og landsdel

I tillegg trenger vi å beskrive selve regnearket som skal lastes opp. Alle regneark blir beskrevet i samme ontologi, som vi har kalt "SSB Data", og hvert regneark har hver sin klasse i ontologien. For sykefravær har vi en klasse som beskriver at én verdi tilhører en kommune, aldersintervall, kvartal, år og kjønn. Se Figur 2.1-2.



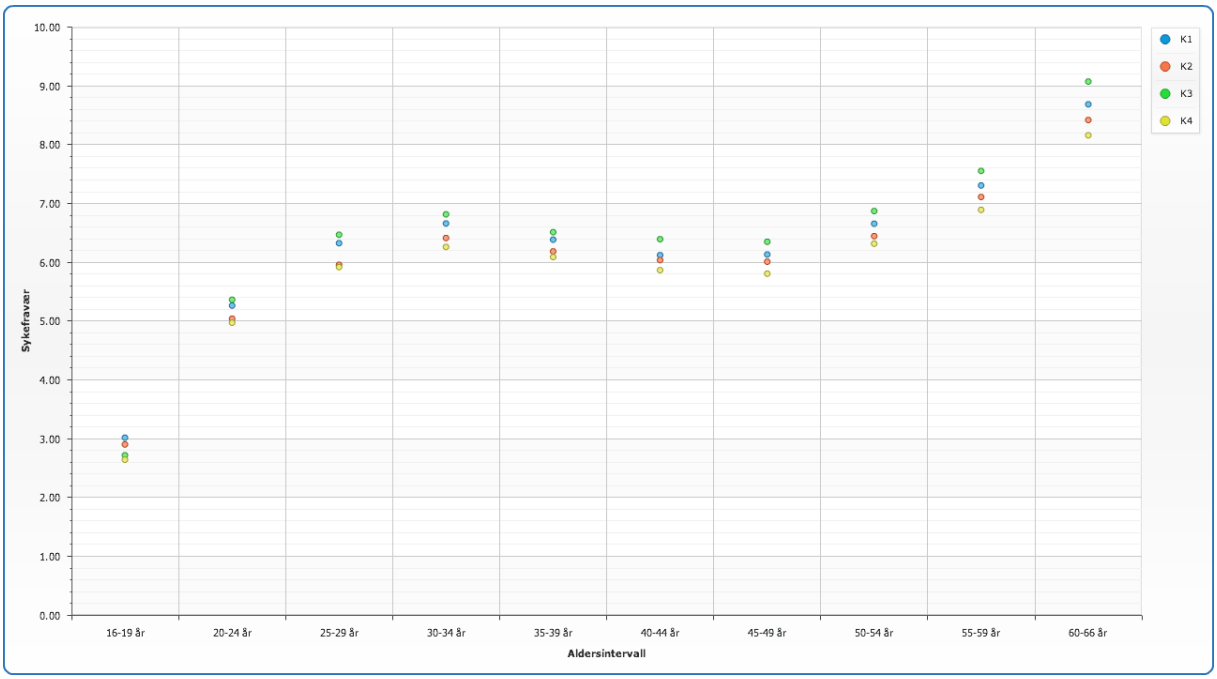

Figur 2.1-2 Klassen Sykefraværs relasjoner til andre klasser og verdier

Når vi har lagd ferdig disse ontologiene, er det klart for å laste dataene opp til serveren. I Excel-regnearket markerer vi de ulike cellene med de ulike verdiene fra Sykefravær-klassen. Noter spesielt at kommune-relasjonen i Sykefravær-klassen peker på Kommune-klassen som tilhører Geografi-ontologien.

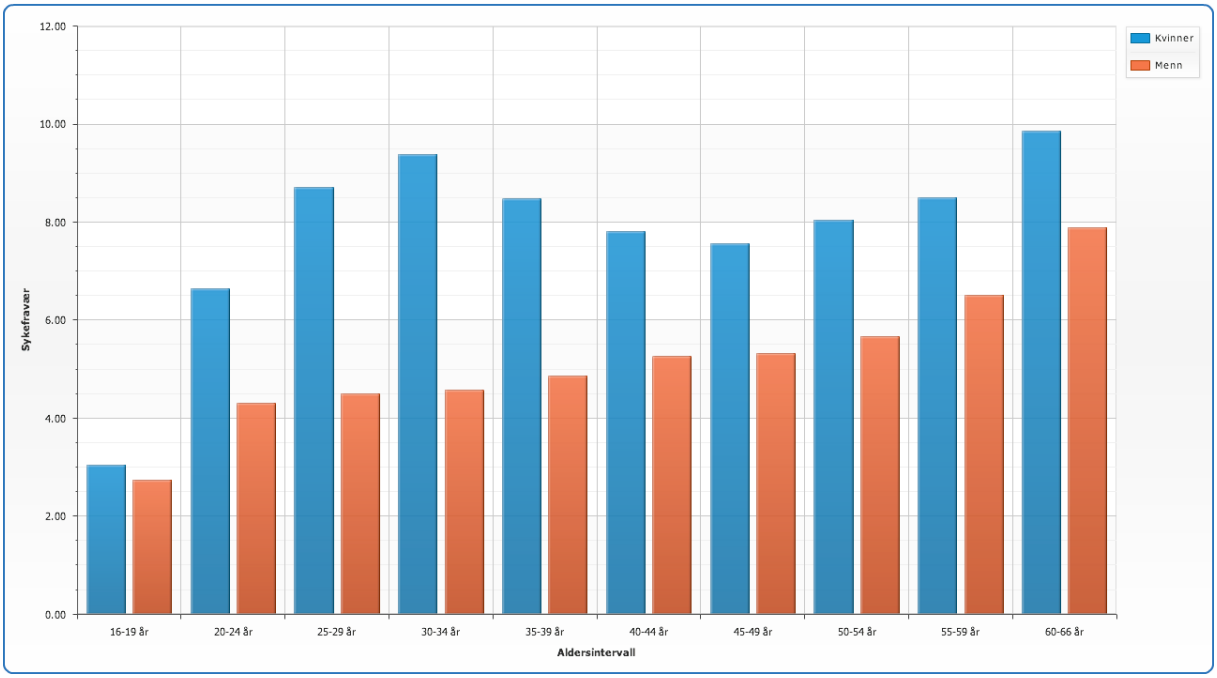
I tillegg må man laste opp en matching-tabell som forklarer hvilke landsdeler de forskjellige fylker tilhører, og hvilke fylker de forskjellige kommuner tilhører. Når alt dette er gjort, er det klart for å generere visualiseringer i *Anzo on the Web*.

### 2.1.3 Presentasjon av oppgaven

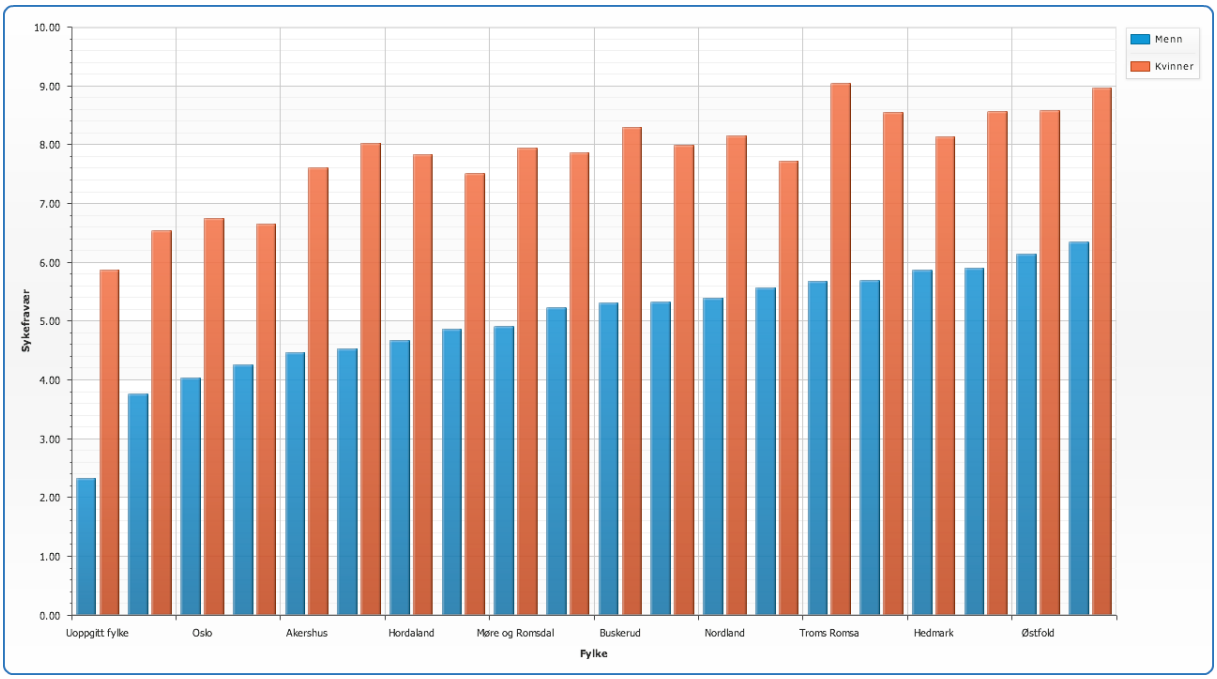
Vi vil ikke komme med noen redegjørelse for hvordan man setter opp ulike visualiseringer i *Anzo on the Web*, men henviser leseren til vedlegg 3. Vi vil derimot presentere noen av de diagrammene man kan generere når man har lastet opp data for sykefravær. Figur 2.1-3 viser hvordan sykefravær i de ulike kvartalene fordeler seg i forhold til aldersintervall. Vi ser at det er betydelig mer sykefravær på høsten enn på vinteren. Her er tallene aggregert for alle årene som er lastet opp og begge kjønn, altså snittet for alle år og kjønn. Figur 2.1-4viser hvordan sykefravær er fordelt på kjønn og aldersintervall, her er kommuner og år aggregert på gjennomsnitt. Figur 2.1-5 viser sykefravær basert på fylke og kjønn.




**Figur 2.1-3 Fordeling av sykefravær i de ulike kvartaler, basert på aldersintervall**




**Figur 2.1-4 Fordeling av sykefravær for kjønn, basert på aldersintervall**




**Figur 2.1-5 Fordeling av sykefravær for kjønn, basert på fylker**

## 2.2 Oppgave 2

### 2.2.1 Innledning

En av de viktigste egenskapene ved den løsningen vi har utarbeidet er å kunne lage visuelle sammenligninger av vidt forskjellige statistiske variabler. Spesielt at sluttbrukeren selv kan bestemme hvilke data som skal sammenlignes. I oppgave 2 tilnærmer vi oss nettopp dette, ved å supplere datasettet i Anzo-serveren med data om skilsmisser og separasjoner og om barnehagedekning.

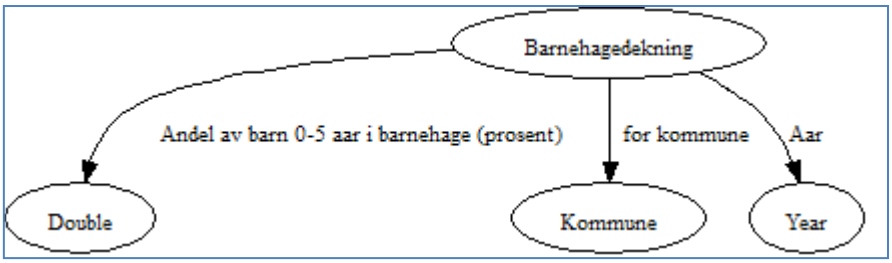
### 2.2.2 Gjennomføring av oppgaven

Innhentingen av data skjer på samme vis som ved innhenting av data om sykefravær. Et problem oppstod derimot i forbindelse med data for skilsmisser og separasjoner, disse er i statistikkbanken til SSB kun oppført som antall per kommune, det vil si at eventuelle visuelle grafer mer vil vise størrelsen på kommunen snarere enn tendensen til å skille seg i kommunen. Vi måtte altså regne ut promille-verdien selv, og gjorde dette med følgende formel for skilsmisser (tilsvarende for separasjoner):

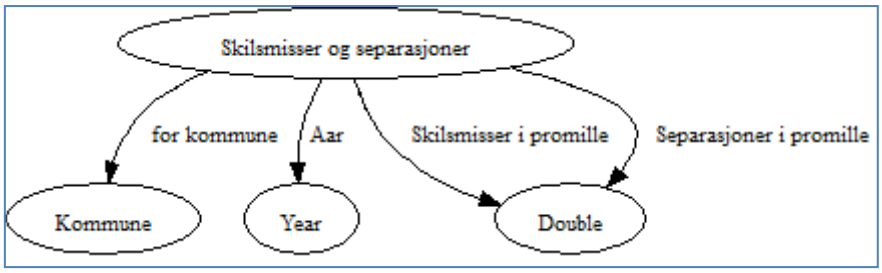
(1000 / Folkemengden i kommunen) * antall skilsmisser i kommunen[4]

Klassene som beskriver regnearket er også tilsvarende som ved Oppgave 1, men disse er noe mindre da man ikke har verdier for kjønn og aldersintervall. For en oversikt over klassen Barnehagedekning, se Figur 2.2-1, tilsvarende for Skilsmisser og separasjoner se Figur 2.2-2.

[4] Det er sikkert bedre måter å gjøre dette på, vi har ikke etterstrebet statistisk nøyaktighet, men fokusert på nye teknologiske løsninger på å distribuere statistikk.



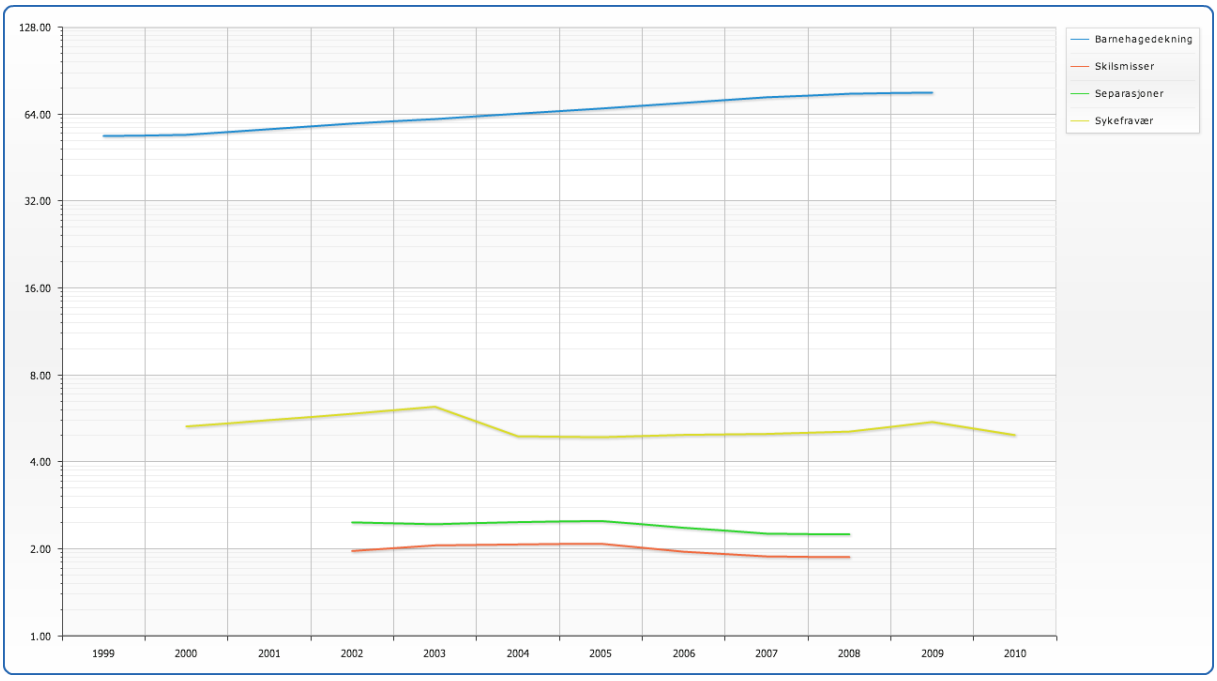

Figur 2.2-1 Klassen som beskriver data for barnehagedekning




Figur 2.2-2 Klassen som beskriver data for skilsmisser og separasjoner

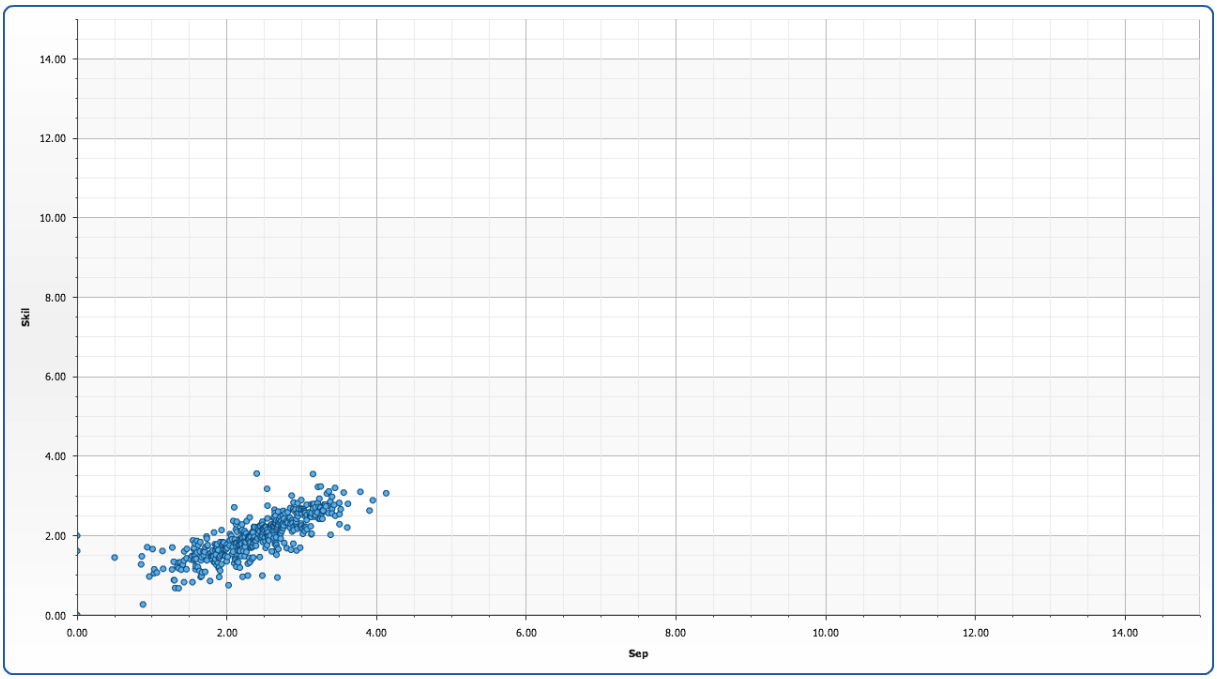
### 2.2.3 Presentasjon av oppgaven

I Figur 2.2-3 kan vi se verdier for barnehagedekning, skilsmisser, separasjoner og sykefravær som tidslinjer, og sammenligne de med hverandre. Legg merke til at ikke alle variablene har data for de samme årene – dette forstyrrer ikke visningen av grafen nevneverdig.




Figur 2.2-3 Barnehagedekning, sykefravær, separasjoner og skilsmisser fra 1999 til 2010

For en se en virkelig korrelasjon, dog ikke så overraskende, kan man se Figur 2.2-4, sammenheng mellom separasjoner og skilsmisser. Denne er vist på kommunenivå, og er et gjennomsnitt av alle årene.




Figur 2.2-4 Separasjoner og skilsmisser i en scattergraf

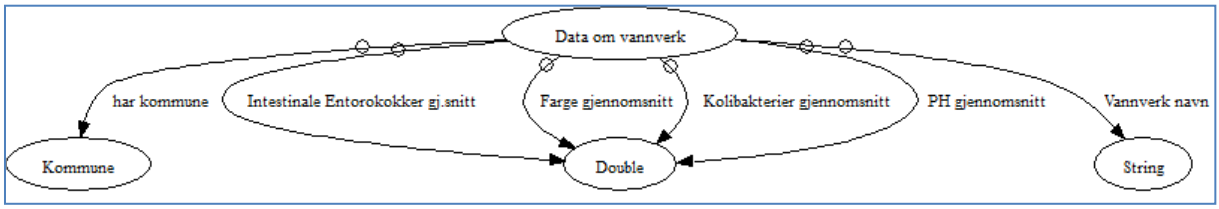
## 2.3 Oppgave 4

### 2.3.1 Innledning

I oppgave 4 ser vi på opplysninger om vannkvalitet og driftskvalitet for kommunenes vanntjenester, og har samlet inn data både fra SSB og Folkehelseinstitutett (FHI). Målet er også her å sette nye data opp mot opplysninger om sykefravær. I denne oppgaven har vi kun konsentrert oss om 2008, fordi det er tall fra dette året vi finner hos FHI. Vi har også begrenset oss til data omkring test av e.coli/koli, PH, farge og intestinale enterokokker, samt ulike økonomiske data. Vi har lett etter korrelasjon mellom disse dataene og sykefravær.

### 2.3.2 Gjennomføring av oppgaven

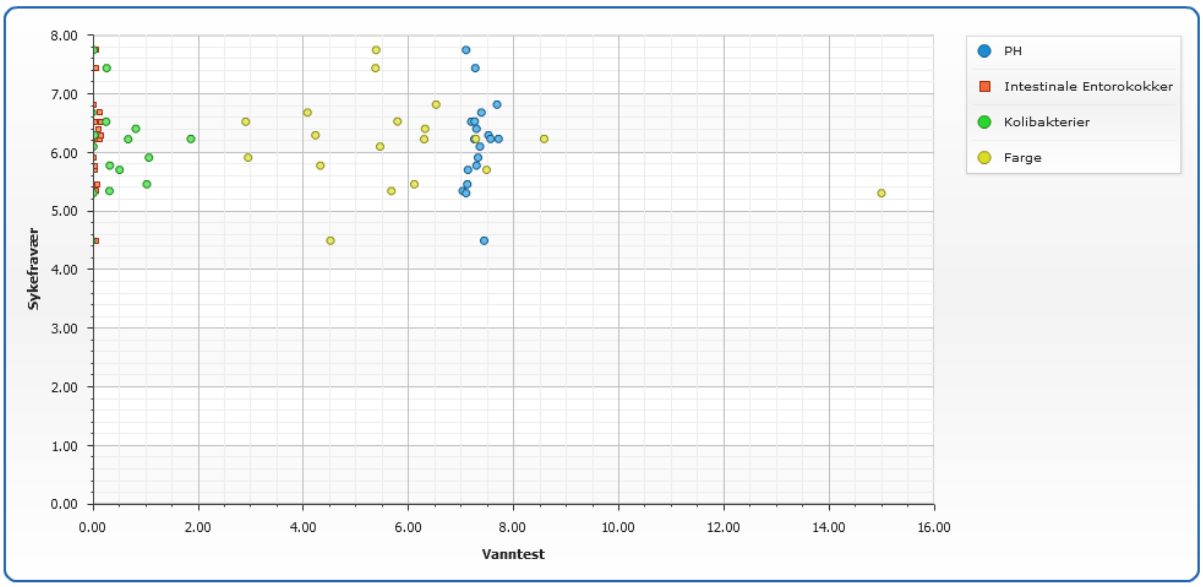
Data fra FHI har vi lastet opp på vannverksnivå, og data fra SSB har vi lastet opp på kommunenivå. Det er ulike løsninger for å få dette til å samkjøre ved visning i ulike grafer. Her har vi valgt å bruke en kommune-peker i ontologien for data fra FHI (som ikke peker på noen kommune!), og peke oss videre inn til fylke når vi laster opp dataene. Denne manøveren gjør oss i stand til å aggregere både vannverk og kommuner til riktig fylke, slik at vi kan sammenligne dataene på fylkesnivå. I teorien kunne vi også lastet opp vannverk på kommunenivå, men da enkelte vannverk er interkommunale, ville vi uansett ha avvik for grafer over kommunenivå. Ontologi-klassen som inneholder data om ulike vannverk, fra FHI, ser man på Figur 2.3-1. Som man ser har den en kommune-peker, men kommune-ressursene som blir pekt på inneholder ikke noe annet enn en peker videre til fylke. Ontologien for data fra SSB er tilsvarende, men går da på de ulike kommuner, ikke vannverk.




Figur 2.3-1 Klassen som beskriver data om vannverk

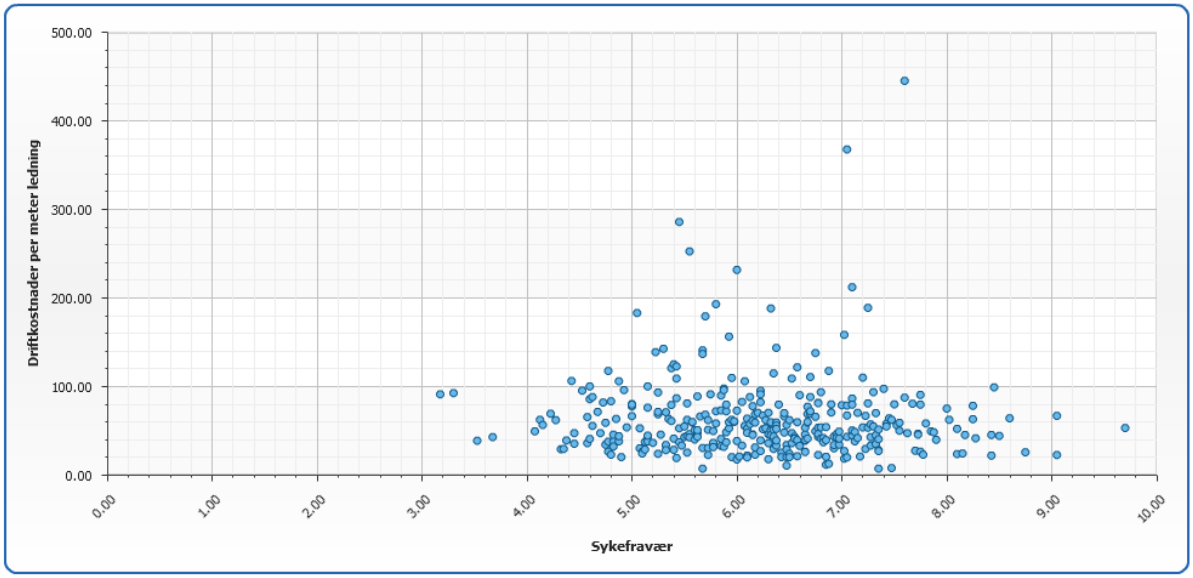
### 2.3.3 Presentasjon av oppgaven

Det første vi gjorde var å sjekke FHIs statistikk om vannkvalitet (Farge, PH, Kolibakterier og Intestinale Entorokokker) mot sykefravær. Her finner man ingen stor korrelasjon, men vi ser et eksempel på hvor lett det er å lete etter korrelasjoner mellom flere variabler samtidig. Med sykefravær i den ene aksen, og de ulike testene, med forskjellige farger, i den andre, kan man lett få et fint overblikk over dataene. Se Figur 2.3-2.




Figur 2.3-2 Sykefravær koblet mot vannkvalitet

Figur 2.1-1 viser sykefravær koblet opp mot driftskostnader per meter ledning med vann. Dette er vist for kommuner. Vi ser at alle kommuner som har høye driftskostnader (over 100) har under 8 prosent sykefravær.




Figur 2.3-3 Sykefravær koblet til driftskostnader per meter ledning, med fargekvalitet, kommuner

I teorien kan oppgave 4 utvides med enda mer data, men vi føler at vi har vist med dette at det er mulig å sette opp fine grafer som viser eventuelle korrelasjoner mellom vannkvalitet, driftskostnader, med mer. Vi har også klart å sammenligne data fra forskjellige kilder, noe vi vil gjøre mer av i oppgave 5.

## 2.4 Oppgave 5

### 2.4.1 Innledning

I oppgave 5 utforsket vi videre muligheten til å sammenligne data fra SSB med andre kilder. I utgangspunktet benyttet vi DIFIs folkeundersøkelse. Vi fikk tak i individuelle data fra undersøkelsen og benyttet IBM PASW Statistics 18 for å utregne statistikk for de ulike kommuner, basert på kjønn. Vi genererte statistikk for hvordan folk opplevde barnehagedekning, barnehagekvalitet, luftkvalitet, vannkvalitet, vannsikkerhet, fastlegekvalitet, med mer.

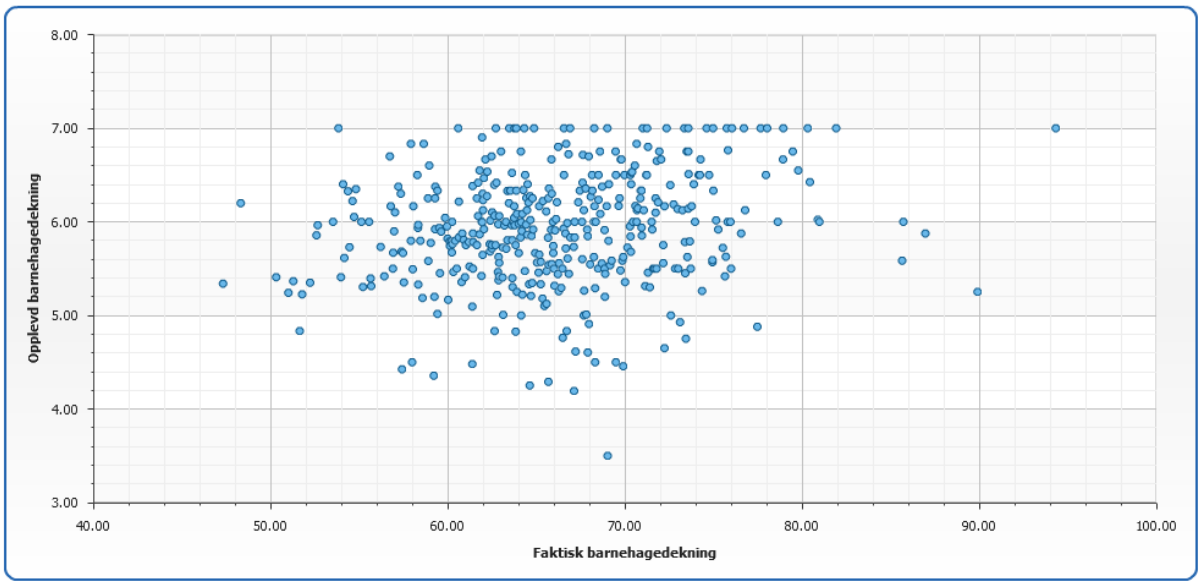
### 2.4.2 Gjennomføring av oppgaven

I undersøkelsen ble subjektene spurt om hva de syntes om kvalitet og dekning på en skala fra svært dårlig (-3) gjennom middels (0) til svært godt (+3). Dette har vi så oversatt til en verdi fra 1 til 7. Deretter regnet vi ut gjennomsnittet av hva folk har svart (basert på kommune og kjønn), slik at vi endte opp med en desimalverdi som går fra 1.0 til 7.0.

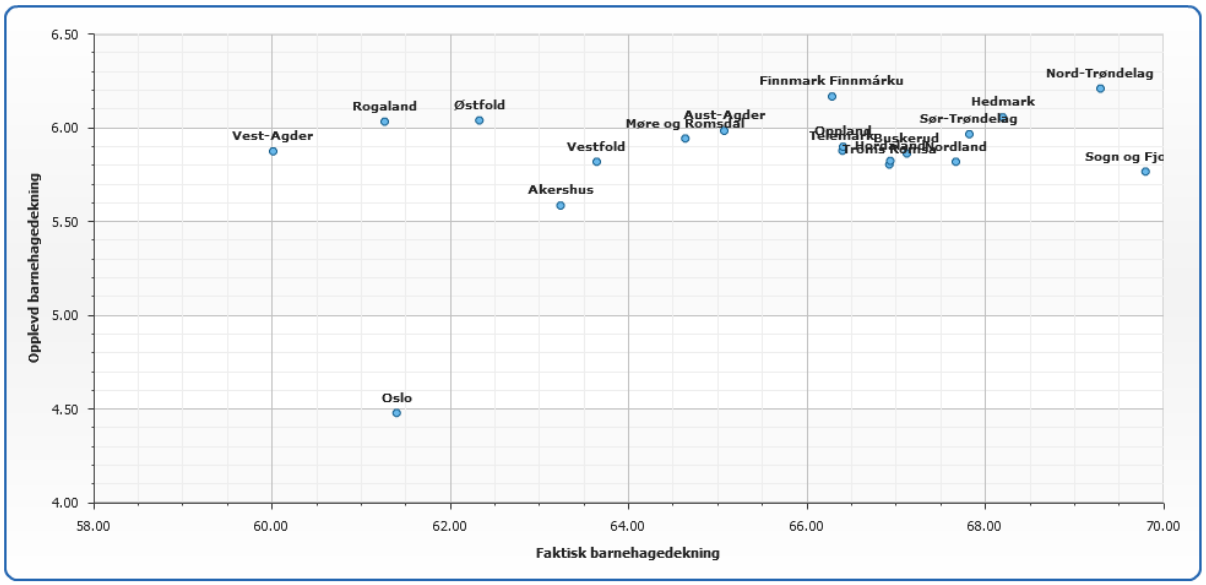
Resten av løsningen er ellers identisk med de tre tidligere oppgaver.

### 2.4.3 Presentasjon av oppgaven

Den første grafen vi lagde var en scatter-graf over hvordan folk opplevde barnehagedekningen (ifølge DIFIs undersøkelse) og den faktiske barnehagedekningen (fra SSBs statistikkbank), basert på kommuner Figur 2.4-1. Deretter benyttet vi aggregeringsmulighetene i *Anzo on the Web* for å vise samme statistikk på fylkesnivå, Figur 2.4-2. Legg merke til at skalaen på denne er moderert for å få markørene til å flyte vekk fra hverandre.

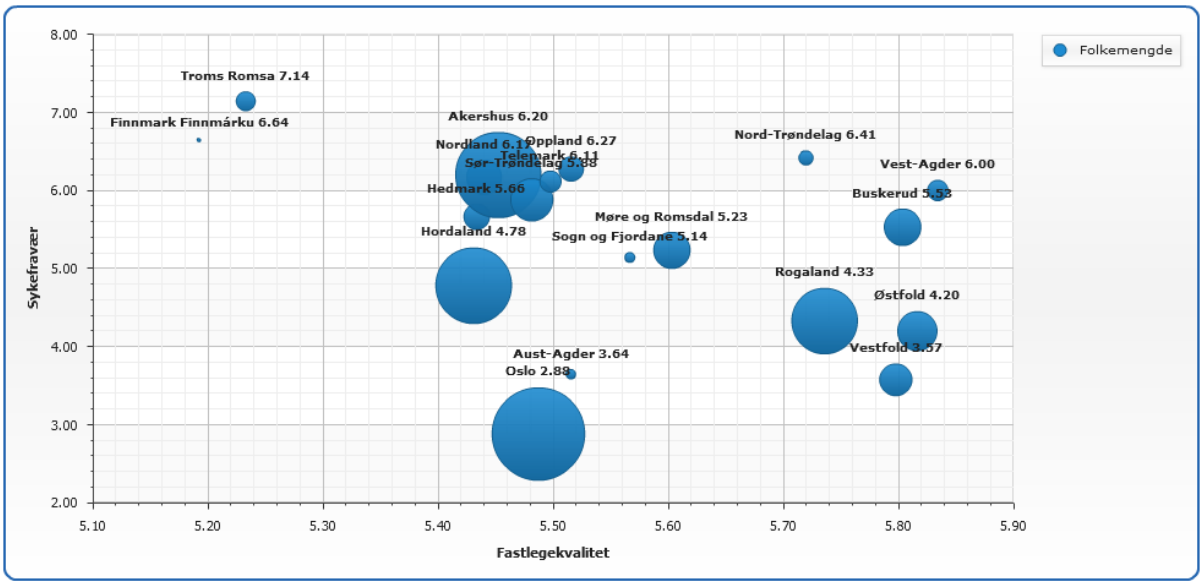



**Figur 2.4-1 Barnehagedekning, opplevd og faktisk, kommuner**




Figur 2.4-2 Barnehagedekning, opplevd og faktisk, fylker

En annen interessant korrelasjon å se etter er hvordan folk opplever kvaliteten på fastlegen sin, i forhold til hvor mange sykemeldinger som blir skrevet ut i fylket. Det ser ikke ut til å være særlig korrelasjon her, som vi ser på Figur 2.4-3. Legg merke til at vi her viser 3 dimensjoner, hvor størrelsen på boblen bestemmes av antall innbyggere i fylket. (Det er litt uoversiktlig med hensyn til hvilken boble som tilhører hvilket fylke. I web-grensesnittet kan man derimot holde musepekeren over boblen, og se hvilket fylke den tilhører.)




Figur 2.4-3 Opplevd kvalitet på fastlege i henhold til sykefravær (gjennomsnitt i fylke)

Vi har også forsøkt å visualisere data fra et verdensperspektiv, hvor man kan grave seg fra kontinentnivå helt ned til kommunenivå i Norge. Vi har her benyttet data for fødselsrate, hvor dataene om de forskjellige land i verden er hentet fra World Bank og data fra kommuner i Norge er hentet fra SSB. Vi har her konsentrert oss om årene 2006-2008.

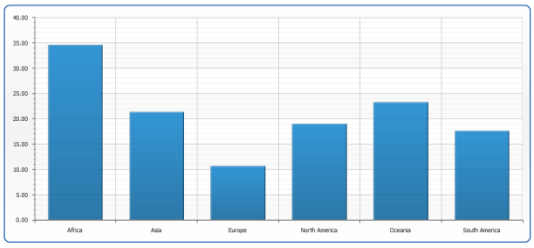
Først måtte vi lage en tabell som kobler land til verdensdel. Vi regner her da Afrika, Asia, Europa, Nord og Sør-Amerika samt Oseania som verdensdeler. Informasjon om hvilke land

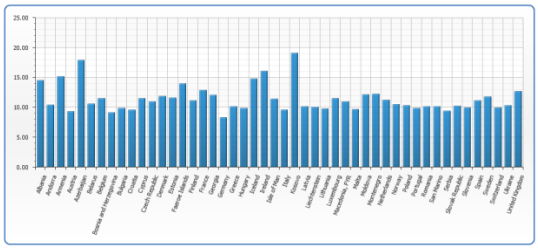
som tilhører hvilken verdensdel ble hentet fra www.worldatlas.com. Deretter utvidet vi tabellen som kobler fylker til landsdeler, og la på informasjon om at landsdelene tilhører landet Norge. Ontologien blir ikke ulikt Figur 2.1-1 på side 5, men da med land og verdensdel i tillegg.

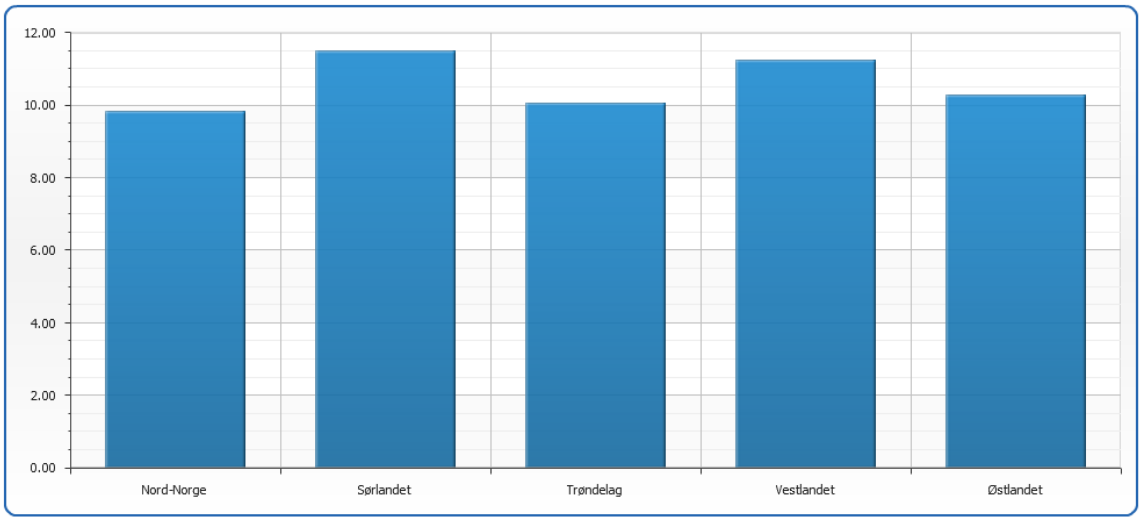
Klassen som representerer regnearket for data om fødselsrate har kun en kommune-peker. Deretter må vi grave oss opp til landsnivå når vi skal laste opp data for land, fra kommunepekeren. Når vi merker at dataene er tilknyttet et land, må vi altså i ontologien si at den tilhører en kommune, for deretter å klikke oss oppover (via fylke og landsdel) til land når dataene lastes opp. Dette er nødvendig for at *Anzo on the Web* kan sidestille data om kommuner og land i en graf som viser hvert enkelt land.

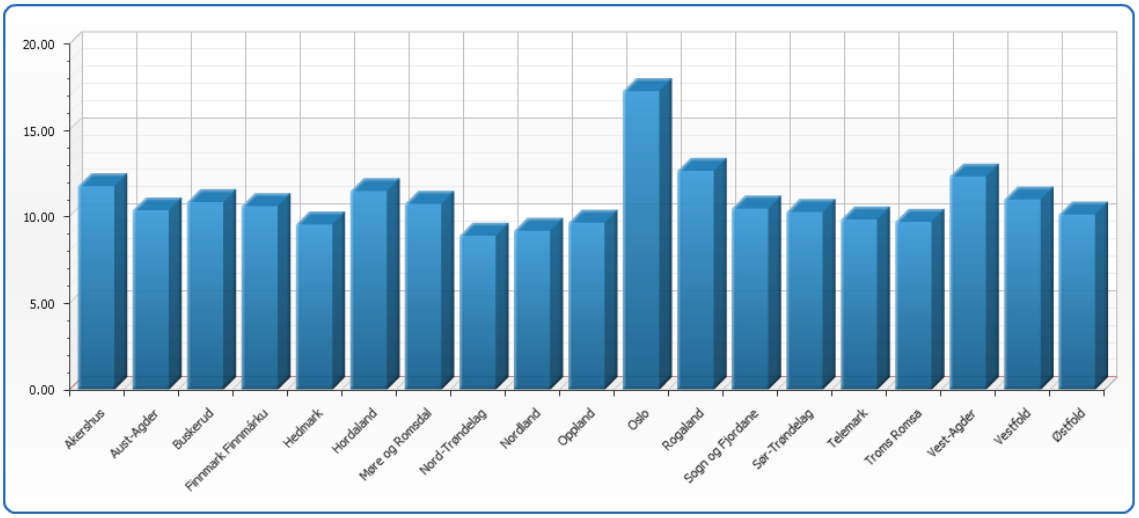
På neste side ser vi først fødselsrate for de ulike kontinenter, dette er da snittet av alle land. Deretter ser vi alle land i Europa, her er verdiene for Norge regnet ut med snittet for alle kommuner. Deretter ser vi landsdeler i Norge, så fylker i Norge og til slutt kommuner i Akershus.

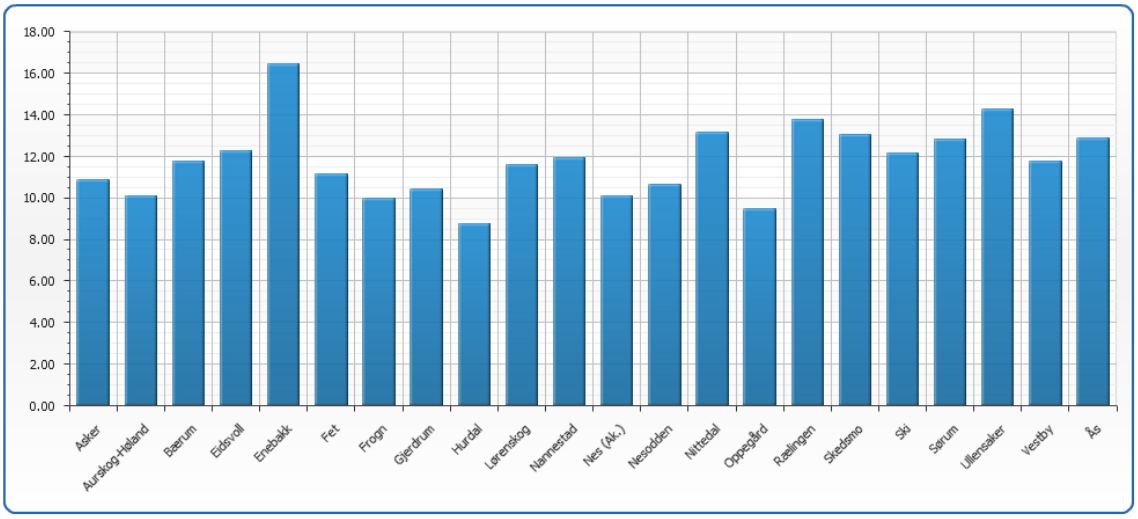
I denne oppgaven kom vi over et stort problem: Fødselsratesnitt for alle kommuner 2008 er 10.4, men regner man ut fødselsrate for Norge blir den 12.6. Samme problem gjelder da når man regner snitt for alle land, og viser kontinent, etc. Grunnen til dette er at alle land i europa blir vektet likt med alle kommuner i Norge. Problemene rundt utregning av prosentverdier i Anzo diskuteres videre i del 3.1.


40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Africa
Asia
Europe
North America
Oceania
South America


25.00
20.00
15.00
10.00
5.00
0.00


12.00
10.00
8.00
6.00
4.00
2.00
0.00
Nord-Norge
Sørlandet
Trøndelag
Vestlandet
Østlandet


20.00
15.00
10.00
5.00
0.00
Akershus
Aust-Agder
Buskerud
Finnmark Finnmárku
Hedmark
Hordaland
Møre og Romsdal
Nord-Trøndelag
Nordland
Oppland
Oslo
Rogaland
Sogn og Fjordane
Sør-Trøndelag
Telemark
Troms Romsa
Vest-Agder
Vestfold
Østfold


18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Asker
Aurskog-Høland
Bærum
Eidsvoll
Enebakk
Fet
Frogn
Gjerdrum
Hurdal
Lørenskog
Nannestad
Nes (Ak.)
Nesodden
Nittedal
Oppegård
Rælingen
Skedsmo
Ski
Sørum
Ullensaker
Vestby
Ås

# 3 Avslutning

## 3.1 Problemer med rater og prosentverdier

Å beskrive og sammenligne prosent og rater er vanskelig i seg selv, men nødvendig når man skal presentere statistikk. Problemene vi har møtt på i disse oppgavene er som følger:

- Si at vi har tall for sykefravær i alle kommunene i Akershus oppgitt i prosent, hvordan finner vi prosenttallet for Akershus?
- Si at vi har tall for sykefravær for alle kvartalene i 2009, hvordan finner prosenttallet for 2009?

Har man bare prosenttallene å gå ut i fra er dette en umulig oppgave. Det vi har gjort er å bruke gjennomsnitt av alle kommuner, eller alle kvartaler, noe som åpenbart gir et feil resultat. Noen bedre løsning finner man dessverre ikke i Anzo.

For nå finnes det to måter å løse dette problemet, men vi er ikke fornøyde med noen av dem. Man kan enten regne ut prosentverdiene for alle nivåer på forhånd, og lagre disse til serveren. Eller man kan gi SPARQL CONSTRUCT-spørringer som tar utgangspunkt i tallene som blir lastet opp, og regner ut verdier for større områder selv. Begge løsninger krever større kompleksitet og detaljnivå i ontologiene. For hver variabel trenger man nå en egenskap for alle de ulike geografiske nivåene, og alle de ulike tidsaspektene en bruker kunne tenke seg å etterspørre. For å ta Sykefravær som et eksempel:

- Kommuner hvert kvartal
- Kommuner hvert år
- Kommuner hvert tiår
- Kommuner for 2009
- Fylker hvert kvartal
- Fylker hvert år
- Fylker hvert tiår
- Politidistrikt hvert kvartal
- …
- Politidistrikt hvert enkelt år

Å forutse alle kombinasjoner av geografiske områder og tidsaspekter en bruker kan etterspørre har en enorm kompleksitet O(n!)[5], både hva angår beregning og lagringsplass, og vil ikke være løsbart for større datasett.

Vårt forslag til en løsning er å lagre formelen for utregning av prosent. Sykefraværsprosent regnes blant annet ut slik: (sykefraværsdagsverk / avtalte dagsverk) * 100. Kunne man lagret data for sykefraværsdagsverk og avtalte dagsverk, samt formelen for prosent tilknyttet klassen som representerer Sykefraværsprosent, kan Anzo selv beregne prosentverdien etter å ha summert sykefraværsdagsverk og avtalte dagsverk for de aktuelle geografiske og tidsmessige områdene.[6] (Se også vedlegg 4, fra side 25.)

Når brukeren da velger å se på sykefraværsprosent for de ulike fylker for 2010, blir data for de korrekte kommuner og kvartaler summert, for deretter å benytte formelen som regner ut

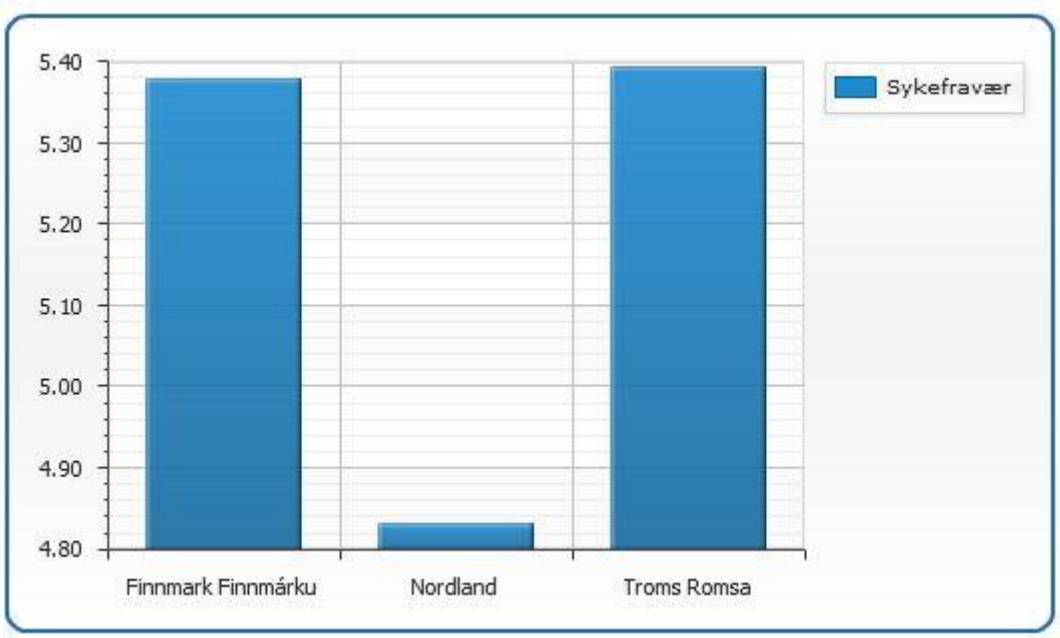
[5] På samme måte som ved traveling salesman må man regne ut alle mulige kombinasjoner.

[6] Vi vil gjøre oppmerksom på at OWL ikke innehar egenskaper for å beskrive slike komplekse matematiske størrelser. Man må altså utenfor ontologienes sfærer for å løse dette problemet på den måten vi har forespeilet. Man kan dog lagre formlene som RDF-tripler og gi dem en entydig semantisk tolkning.

prosentverdien. Rent ingeniørmessig skulle dette være en forholdsvis enkel egenskap å implementere.

## 3.2 Øvrige problemer

Enkelte problemstillinger er verdt å få med seg ved bruk av verktøy som visualiserer statistikk automatisk. Et av problemene er at *Anzo on the Web* stiller verdiene på aksene automatisk, noe som kan gi et feil inntrykk. Et ekstremt eksempel kan man se på Figur 2.1-1. Her ser det ut til at folk i nordland nesten aldri er syke, mens innbyggerne i Troms og Finnmark er dårligere stilt. Sannheten er selvfølgelig den at forskjellen mellom Nordland og Troms i virkeligheten kun er 0,6 prosent (som vi jo ser om vi leser av tallene på grafen).




**Figur 3.2-1 Hvorfor er folk i Nordland så friske i forhold til de i nabofylkene?**

## 3.3 Videre arbeid

Første problem man må se på videre, er aggregering av rater og prosent. Slik Anzo fungerer nå, er det ikke tilfredsstillende for visning av statistikk. Det er også et spørsmål om hvor vidt vårt løsningsforslag er plausibel. Owl har ingen muligheter for å definere matematiske komplekse størrelser, og løsningsforslaget vil på den måten bryte med W3C-standardens ånd. Tilsvarende, men andre, problemer kan også dukke opp i det man begynner å anvende tilsvarende teknologier for visualisering av statistikk på større skala.

Neste forbedringspotensial er GUI. Brukeren må enkelt kunne velge de kategorier som han vil sammenligne, uten å måtte vite hvordan Anzo er skrudd sammen. Det skal for eksempel ikke være nødvendig å bla seg opp på landsdelsnivå (for å sammenligne data for ulike landsdeler) gjennom klassen Kommune sin Fylke-egenskap, og Fylke sin Landsdel-egenskap. Velger man Landsdel fra GUI burde dette være implisitt.

Det burde også være mulig å la flere collaboration-servere samarbeide. Hvis SSB har noe data på sin server, og for eksempel DIFI annen statistikk hos seg, burde man kunne sammenligne disse hvis ontologiene stemmer overens. SSB og DIFI kan da dele generelle ontologier som beskriver verden, og ha private ontologier som beskriver datasettet.

For å få til dette, burde man også kunne utveksle RDF-tripler direkte, for eksempel ved hjelp av SPARQL-spørringer. Her er det viktig å også tenke på beskyttelse av privatliv, man vil

ikke gjøre RDF-tripler som gir opplysninger på individnivå tilgjengelig for offentligheten. Samtidig burde disse triplene kunne benyttes for å beregne data på makro-nivå.

Vi vil foreslå at én aktør står ansvarlig for utforming av generelle ontologier, og rettlingslinjer for hvordan ontologier som beskriver data internt burde utformes. En naturlig kandidat til dette vil være SERES II-prosjektet (Semantikkregisteret for elektronisk samhandling). Det er mulig en slik aktør må ha en viss forståelse for hvordan Anzo bruker ontologier for generering av data.

Man burde også forsøke å gjøre et oppskalert prosjekt med utgangspunkt i SSBs statistikkbank, med dedikerte servere og databaser. En naturlig aktør vil her være Semicolon 2. Nå har vi sett at det lar seg gjøre å anvende semantiske teknologier for distribuering og visualisering av statistikk. Eventuelle problemer rundt praktisk gjennomførbarhet, med tanke på momenter som hurtighet, brukervennlighet og offentlig tilgang til interaktive visualiseringer må forskes på videre.

Ytelsen er også en utfordring. Det er store mengder data som skal hentes ut fra trippeldatabasen, og som må bearbeides matematisk før programmet kan starte med å tegne grafer. *Anzo Collaboration Server* versjon 2 har allerede tatt i bruk en bedre databaseløsning. Cambridge Semantics rapporterer at visuelle diagrammer som før tok minutter å generere, nå tar sekunder (Se Vedlegg 4, side 31). I tillegg trenger man dedikerte servere og databaser som kan ta seg av forespørsler fra sluttbrukere. Vi anser allikevel at dette er en overkommelig utfordring hvis man har tenkt å lage en løsning basert på de teknologiene vi har sett på i vårt prosjekt.

## 3.4 Konklusjon

Det er vår oppfatning at produktsuiten til Cambridge Semantics har et stort potensiale som et rammeverk for visualisering av statistikk. Den generelle fremgangsmåten gjør det som tidligere nevnt mulig å stille opp ulike data opp mot hverandre for å se etter korrelasjoner. Med en bedre GUI-løsning på toppen, er funksjonaliteten tilgjengelig for alt fra skoleelever til forskere og media. Når man kan generalisere dataene og sammenligne disse både på et overordnet nivå, samt drille seg ned alle de stedene det er mulig, har man et kraftig verktøy for å generere de grafene man måtte trenge.

Mer generelt vil vi si at semantiske teknologier bidrar til et kraftig verktøy for å behandle statistikk. Ved å bygge på W3C-standarder kan data som er lagret lett benyttes i andre settinger. Også statistisk data kan unngå å havne i siloer, hvor statistikk vanskelig kan byttes mellom ulike aktører. I stedet kan man med semantiske teknologier tilby løsninger hvor sluttbrukere kan sammenligne statistikk publisert forskjellige steder. Alt som trengs er felles ontologier, som beskriver hvordan statistikken er relatert til verden.

Vi mener den største fordelen med en standardisert og generalisert måte å lagre statistikk, er alle mulighetene som åpner seg for å sammenligne ulike data. Man kan bla opp hvilke som helst type data og lete etter korrelasjoner. Semantiske teknologier gjør det enkelt for sluttbrukere å generere statistikk for de områder man er interessert i å sammenligne. I tillegg er det verdt å nevne at den offentlige tilgjengeligheten på data blir betraktelig økt. En web-applikasjon som selv setter opp helt nye, aldri før genererte, visualiseringer basert på hva som helst brukeren ønsker å sammenligne gir muligheter man knapt kan forestille seg!

Fremtidige utfordringer for å utvikle en slik løsning vil være automatisk utregning av rater, ytelse ved visualisering basert på store mengder data og oppsett av ontologier som er tilpasset Anzos visualiseringsmuligheter.

## 3.5 Referanser

Cambridge Semantics, hjemmeside, http://www.cambridgesemantics.com/

DIFI, Innbyggerundersøkelse, http://www.difi.no/artikkel/2010/01/alle-resultater

Semicolon, hjemmeside, http://semicolon.no/

Seres II-prosjektet, http://www.brreg.no/samordning/semantikk/SERES_II_prosjektet.pdf

SSB, statistikkbanken, http://statbank.ssb.no/statistikkbanken/

Statistisk Sentralbyrå, hjemmeside, http://www.ssb.no

Video-demonstrasjon ved bruk av Anzo: http://www.youtube.com/watch?v=G913TGTXhK8

World Atlas, Countries listed by Continent, http://www.worldatlas.com/cntycont.htm

## 3.6 Liste over vedlegg

Vedlegg 1: Studentoppgave for IFI, sommer 2010

Vedlegg 2: Ontologier og datasett

Vedlegg 3: Veiledninger for bruk av Anzo og innsamling av data

Vedlegg 4: Mail-korrespondanse

Vedlegg 5: Beskrivelse av demonstratorer

Vedlegg 6: Presentasjonen gitt for SSB