

Riktig velkommen til deg som leser dette:

Du har mest sannsynlig tenkt å ta en mastergrad ved BIAS. Her kan du lese litt om masteroppgaver du kan velge dersom jeg skal være veilederen din.

Jeg har en svært anvendt profil med stor interesse for undervisning (les evt [mer om det her](#)), og med en bred bakgrunn fra medisinsk forskning på OUS Rikshospitalet og Sunnaas sykehus.

Tar du en oppgave hos meg, vil det alltid være med utgangspunkt i en anvendelse, altså et problem vi håper å løse med statistisk analyse. De neste sidene viser prosjekter som kan være aktuelle. Som du vil se, innebærer de fleste prosjektene samarbeid med andre forskere eller forskergrupper, enten ved NMBU eller eksternt.

Hvis du har egne ideer, kan vi gjerne diskutere dem også.

OBS: Jeg har forskningstermin H23/V24, og det betyr at jeg ikke har mulighet til å inngå avtaler med nye masterstudenter før det ordinære opptaket er over i juni/juli 2024.

Ta kontakt i august 2024 dersom du ønsker en masteroppgave med meg som veileder.

Hvis noen av de skisserte masteroppgavene skulle være aktuelle før den tid, er det mulig å ta kontakt for å høre om en av mine BIAS-kolleger kan bistå i stedet for meg.

Inspirert av tidligere deltakelse i det populærvitenskapelige radioprogrammet Abels tårn

Spørsmålet fra en lytter var:

Er det større eller mindre sannsynlighet for å bli oldemor i dag?

Forventet levealder øker jevnt og trutt. Samtidig øker gjennomsnittlig alder for førstegangsfødende. Mitt spørsmål er hvilken av disse tendensene som går raskest, og om det dermed er større eller mindre sannsynlighet for å bli oldemor i dag i forhold til for hundre år siden.

Spørsmålet var vanskeligere å svare på enn først antatt. I notatene mine står det blant annet

- *Hundre år siden: 1923 (etter spanskesyken og første verdenskrig, få år før antibiotika ble oppdaget)*
- *Riktig at forventet levealder øker. Gjennomsnittsalderen for førstegangsfødende, øker den i hele hundreårsperioden?*
- *Det er andre ting som har endret seg også, så vi kan ikke bare se på gjennomsnittet.*
- *Ser vi på befolkningspyramider, ser vi også at aldersfordelingen i en befolkning endrer seg. Gjennomsnittlig færre barn per person. Flere som ikke får barn?*
- *Feilregistrering av barn født utenfor ekteskap (vanligere før?)*
- *Oldemor vs oldefar?*

[Hør episoden her.](#)

Her er det mye å utforske for en kandidat i anvendt statistikk. Kanskje kan det svaret kandidaten kommer frem til, til og med selges inn til NRK?

Samarbeid med Veterinærhøgskolen:

Planned master project in applied statistics:

Analysing horses' movements

Kathrine Frey Frøslie

At the movement laboratory at Veterinærhøgskolen, horses' movements are captured by infrared cameras that track how the joints, head and back of the horses move while the horse is trotting.

Such data can be analysed by functional data analysis, developed by Jim Ramsay in the early 2000s.

This craves attention from a statistician with programming skills, and the ability to navigate in large data sets.

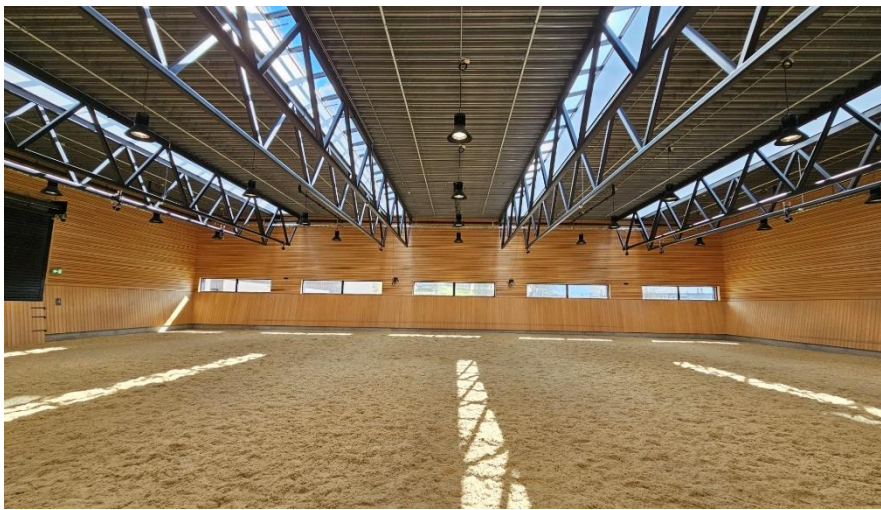


Fig: The movement laboratory for horses at NMBU, with cameras to detect the reflection of IR light from reflective markers.



Fig: Reflective markers are attached to the horse.

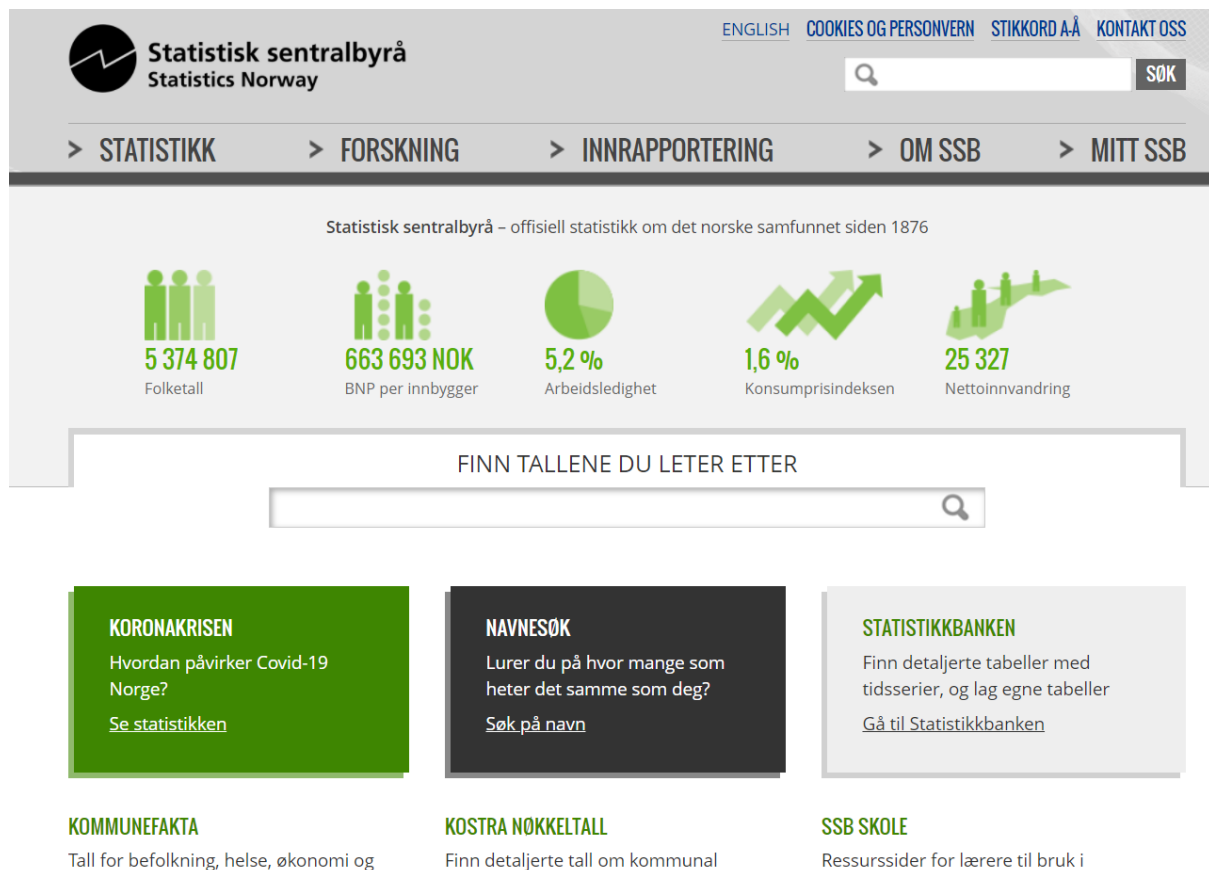
[Les mer om en nylig doktorgradsavhandling om bevegelsesanalyse](#) av hester.

Statistisk sentralbyrå (SSB): Master projects in applied statistics, in collaboration with the Statistics Norway

Kathrine Frey Frøslie (internal NMBU supervisor)

Statistics Norway is the provider of all official statistics in Norway. Several projects are available from SSB's Department for methodology and data collection.

You will find an elaboration of possible project below.



Research and student project ideas using AI/ML in SSB

May 2024

1. Master thesis projects

The division for methods has been contacted by NTNU, NMBU in addition has collaboration with UiO regarding master projects. They are generally interested in projects around IT-security (NTNU) and AI/ML themes. Students write 30 or 60 point theses and are therefore limited in scope. The following are ideas for possible master projects at SSB.

1.1. Grønnboks (NTNU)

Background: (Ongoing with NTNU)

1.2. New NACE classifications

Background: The industry classification of establishments is an important variable for the majority of SSB's business statistics. A new NACE classification standard is being introduced soon, (potentially from 2025) for industry activity. This introduces several challenges but also provides an opportunities to access the way in which we approaches classification tasks. There are several areas where ML/LLM may be able contribute to improved quality and provide more efficient processes:

- Automating classification of NACE based on descriptive texts and other sources.
- Using ML models to detect incorrect classifications.
- Aiding and automating the transition to new NACE classifications

The emergence and popularity of LLMs in the past few years provides opportunities to perform these types of tasks in a novel way. There is potential for finetuning these models in a way to perform predictive classification tasks with much less training data than more traditional supervised learning methods

Project suggestion: Look at ways to use LLMs for predicting new NACE codes (or using current NACE as a backup depending on timing/access to new data)

1.3. Innovation prevalence among Norwegian companies

Background: There have been a number of studies in other countries on predicting innovative companies based on data from their website. This has not been tested in Norway to our knowledge. Norway is a highly digital society and the coverage of websites for companies is likely to be high, making Norway a good candidate for this type of data collection. Many have used DNN for prediction after data extraction but newer LLMs may be a good approach for this task.

Source: Maria, L. E. (2019). Case Study of Big Data Methods for Innovation Detection in the Netherlands

Project suggestion: Use data from webscraping of company websites combined with text mining and prediction algorithms to predict innovation prevalence in Norway. Test different models and the use of LLMs in this context.

1.4. Extracting data from pdf/tiff documents

Background: Manual tasks to find specific data in tiff documents is common in SSB. This is particularly common with economic statistics needing additional information from annual accounts, which have previously only been available in picture tiff format. A survey among the statistical division in April 2024 indicates that around 4.6 man-years is used on manually searching these files for information. SSB has recently gained access to these files in more computer readable formats, however the format is not standardized (many are pdf but also xls, doc, jpeg, png). NLP opens options for automating this process in a way to significantly reduce this time-consuming task.

In one such case study, around 1300 tiff picture files are manually scanned for information on pension costs. NLP could most likely help in this task of extracting the correct information.

Project suggestion: Investigate the use of NLP to extract specific information from pdf/and or picture files (tiff, jpeg, png). A specific case could be to look at extract pensjon costs from annual accounts.

2. Research project suggestions

<https://www.forskningssentret.no/utlysninger/2025/forskningssentre-kunstig-intelligens/>

2.1. Microsimulation in official statistics

Background: Microsimulation is a modelling technique that operates at the level of individual units such as persons, households, vehicles or firms, which strives to allow for interactions among the individual units, and to capture relevant spatial-temporal variations statistically.

Econometric microsimulation is commonly practiced for taxation, pension, health targets and policies... digital twin presentation using of synthetic population data...

See also LifePaths programme at Statistics Canada/Wolfson, and MIKROSIM project at University of Trier...

Project suggestion: There are three broad topic areas that need to be covered.

- Data fusion, which brings together separate data across time and sources, in order to create/capture plausible 'joint information' of a range of associated variables.
- Representation learning of graph-structured data, which also facilitates the modelling of necessary spatial/subpopulation variations or temporal transitions.
- Dynamic (i.e. over time) simulation which uses the learned model for prediction and analysis of socio-political and demographic developments.

2.2. Using AI to enhance information exchange

Background: At the input end of statistical data, survey respondents frequently experience frustrations at answering questionnaires despite the effort and resource invested in them by the Statistical Agency. Interviewers can provide useful assistance in face-to-face or telephone interviews, but intelligent support is lacking for other survey modes such as post, web, mobile.

At the output end, online dissemination of statistical information are structured and labelled according to the standard and tradition of the Statistical Agency, which is unfamiliar to many non-expert public users. Finding the right information may be complicated in such situations, and one may fail to do so despite considerable patience and perserverance.

Similar difficulties and challenges of information exchange exist between many other institutions on the one hand and the general public on the other.

Project suggestion: Using AI to enhance and improve information exchange both ways. The issue is not generative but semantic AI, i.e. interactive tools and technologies that support intelligent information retrieval in either direction, rather than content synthesis or creation.

Analyse av slettsnokdata, samarbeid med Naturmuseum og botanisk hage, UiA, i Kristiansand, ved Pål Sørensen og Beate Strøm Johansen.

Forslag til MASTEROPPGAVER i STATISTIKK Nmbu, Ås

Statistiske problemstillinger knyttet til nasjonal slettsnokforskning



SLETTSNOK (*Coronella austriaca*) er Norges sjeldneste slange, og den eneste reptilarten som er på Rødlista over truede og sårbare arter. Det foregår innsamling av forskningsdata på slettsnok fra 3 hovedregioner i Norge; Kristiansand, Oslo/Akershus, Moss. Disse regionene representerer den sørligste, nordligste og østligste utbredelsen av arten i landet, og vi er interessert i å finne ut om det forekommer geografiske variasjoner, kjønnsforskjeller, samt arvelighet. Datamaterialet består av 943 individer fra perioden 1982-2020. Etter feltsesongen 2021 vil materialet bestå av nærmere 1000 individer.

Problemstilling 1: Forekomst av nakkestriper hos slettsnok i Norge

Slettsnoken har en karakteristisk, mørk hodeflekk som har gitt grunnlaget for artens vitenskapelige slektsnavn, *Coronella*, som betyr liten krone. Mange individer har en eller to utløpere ut fra hodeflekken, disse utløperne kaller vi nakkestriper.

Mønsteret som nakkeflekken danner, er spesielt for det enkelte individ. Sammen med variasjoner i skjellkledningen på hodet kan dette brukes til sikker identifisering av slettsnokindivider.

Fra Oslo-området er det undersøkt 38 ungekull fra kjente mødre, og disse er samlet i fra 8 ulike isolerte bestander. Vi ønsker å finne ut om nakkestriper er arvelig hvis mor har nakkestriper. NB! Noen av hunnene er representert ved flere kull. Hvilken betydning har det at enkelthunner er representert flere ganger?

-Forskjell i forekomst av nakkestriper mellom regioner, lokaliteter og kjønn. Eventuelt også mellom adulte og juvenile individer.

- Forskjell i forekomst av ulike typer nakkestripe (begge, høyre, venstre, ingen)
- Analysere arvelighet av nakkestripe for ungekull med kjent mor.
- Vurdere datasettenes størrelse og anvendbarhet i den statistiske behandlingen. Hvilken betydning har ulik størrelse på datasettene?
- Vurdere signifikans i de ulike sammenlikningene.
- På hvilken måte har «bakgrunssituasjonen», forekomsten av nakkestriper på lokaliteten hunnen hører til, betydning for beregningen?
- Mens data fra Kristiansand- og Østfold-regionen er innsamlet i løpet av få år, er data fra Oslofjord-regionen innsamlet gjennom en periode på nesten 40 år. Dataene fra Kristiansand og Østfold gir derfor et «øyeblikksbilde» av samtidige individer, mens dette ikke gjelder for Oslofjord-regionen. Forekomsten av nakkestriper kan da ha endret seg over tid, avhengig av hvilke individer som til enhver tid har befolket lokalitetene. Har dette noen betydning for behandlingen av dataene?

Problemstilling 2: Forekomst av «anomalier» hos slettsnok i Norge



Individer av slanger innen samme art kan ha ulikt antall ryggvirvler og halevirvler. Skjellkledningen på kroppen har ofte sammenheng med ryggvirvlene inni kroppen. Iblant kan slanger ha ekstra halve skjell (=anomalier), og hvis dette er knyttet til en deformitet i ryggsøylen så kan det påvirke slangens bevegelse – og dermed evnen til overlevelse og reproduksjon. Vi har samlet inn data på anomalier ved kloakkåpningen i det samme datasettet med 943 slettsnokindivider. I tillegg har vi samlet inn data på anomalier i leppeskjellene på slettsnokens hode, men dette er et mer begrenset datamateriale. Etter feltsesongen 2021 vil vi ha samlet inn et mer omfattende materiale på leppe-anomalier.

Spørsmålene vi ønsker svar på er de samme som nevnt over med nakkestripene: geografiske forskjeller, kjønnsforskjeller, arvelighet i ungekull-materialet.

Problemstilling 3: Slange-plater og herptilkartlegging

I feltsesongen 2019 ble det gjort en bacheloroppgave i biologi i Kristiansand, som var 2-delt:

A) Undersøkelse av reptiler under 30 slangeplater + 3 temperaturmålinger + lysmålinger.

B) Visuell kartlegging av herptilarter fra vår til høst: artsfunn, dato, klokkeslett, antall timer i felt. Det var 6 personer som drev med kartlegging av amfibier og reptiler (=herptiler) i 2019 på dette stedet. Det ble derfor mange timer i felt og samlet inn mye data.

For å kunne publisere dette materialet, trengs statistiske analyser.

Til masterstudenter i statistikk: Du/dere er selvfølgelig hjertelig velkommen til å bli med på slettsnok-feltarbeid, men det er ikke nødvendig fordi alt datamaterialet foreligger i ferdige excelfiler. Det foregår feltarbeid i Moss, Oslo-området, og Kristiansand. Pål Sørensen har tatt masteroppgave på slettsnok ved Nmbu, og han bor på Nesodden. Pål har forsket på slettsnok siden 1982, er i dag pensjonist, og jobber med å publisere slettsnokmateriale. Kontaktinfo Pål: Epost: pal_sor@hotmail.com, mobiltilf: 91551426.

Beate Strøm Johansen er zoolog/etolog fra UiO+NTNU, og er ansatt som forsker ved Naturmuseum og botanisk hage, UiA, i Kristiansand, der hun driver slettsnokforskning.

Kontaktinfo Beate: Epost: beate.johansen@uia.no, mobiltilf: 93218653, www.naturmuseum.no

Vi ønsker selvsagt at eventuelle studenter og forskere i statistikk som får arbeide med slettsnokmaterialet, blir medforfatter på publikasjonene sammen med Pål Sørensen og Beate Strøm Johansen.