VIDENCENTRET FOR LANDBRUG
Kvæg

Agro Food Park 15
DK 8200 Aarhus N

+45 8740 5000
vfl@vfl.dk
vfl.dk

CVR 32346987

22. december 2014

# Afrapportering af projekt 3277 – Nye værktøjer til analyse af komplekse data i besætningen

*Specialkonsulent Søs Ancker, Team Sundhed, Velfærd og Reproduktion*

I det moderne kvægbrug er indsamlingen af produktionsdata eksploderet de senere år med indførelsen af ny teknologi på bedrifterne. De stadig større mængder af data nærmer sig i volumen, hastighed, kompleksitet og variabilitet, det man kalder "big data". Overblik over disse stadigt stigende datamængder er vanskeligt at fastholde for den enkelte landmand og rådgiverne.

**Projektets formål** er, at udvikle et let tilgængeligt og praktisk anvendeligt værktøj, til vurdering af komplicerede årsagssammenhænge på besætningsniveau. Værktøjet skal anvendes af landmænd og forskellige typer af rådgivere i fællesskab. Projektet skal danne grundlag for øget konsensus omkring prioritering af indsatsområder i den enkelte besætning. Dermed fremmes en helhedsorienteret rådgivning, der sigter efter samme mål og et økonomisk optimalt udbytte.

Projektet løber frem til medio 2016 og der er fastlagt følgende indsatsområder:

I1. Afdækning af relevante metoder til analyse af 'big data'
I2. Etablering af datasæt, statistisk model og analyse
I3. Præsentation af data
I4. Funktionalitetsbeskrivelse

**I indeværende projekt-år** er hovedvægten lagt på følgende:

I1. Afdækning af relevante metoder til analyse af 'big data' – herunder
- Hvilke analysemetoder har potentiale til at estimere årsagssammenhænge, og visualisere indbyrdes afhængighed mellem data

I2. Etablering af datasæt, statistisk model og analyse – herunder
- Udpegning af datakilder, der skal kunne levere input til den analytiske platform indenfor produktion, reproduktion, fodring og sundhed

Som en del af projektet er desuden arbejdet på en central udpegning af besætninger med markant ydelsesfald. De besætninger, der har et markant ydelsesfald, kan med fordel gøre brug af det værktøj/den analyse, vi ønsker at udvikle – et værktøj der viser hvad der har haft effekt på ydelsesændringen i besætningen, og dermed hvilke indsatsområder man fremadrettet skal have fokus på. Andre projekter, har ligeledes fokus på ydelsesfald på driftsenhedsniveau (projekt 2365, Multifaktorielle besætningsproblemer og projekt 2345, Værktøjer til overvågning). Når fokus i år har været, at blive så skarpe som muligt, på anvendelsen af værktøjet og målgruppen herfor, har det derfor været nødvendigt at lægge en indsats på ydelses-delen.

På de efterfølgende sider laves en opsummering af aktiviteterne i projektet. Opsummeringen laves ved hjælp af udpluk fra mødereferater og Notater der er udarbejdet undervejs i projektet.

**Første vidensopsamling**

- I januar blev afholdt møde med Michael van Straten og Shmuel Bruckstein fra Hachaklait Veterinærservice i Israel for at få deres indspil til projektet. De har erfaringerne fra mange års arbejde med at videreudvikle en besætningsrapport, som netop hviler på resultater fra statistiske analyser på besætningsniveau.
- Der blev taget kontakt til Luke O'Grady i Irland, som har erfaringer med forskellige analysemetoder til håndtering af 'big data'
- Der blev taget kontakt til statistiker Flemming Skjøth, Aalborg sygehus, for at gøre brug af hans kombinerede viden dels fra analyser udført på baggrund af data fra malkekvægsbesætninger, og dels fra analyser gennemført indenfor humanregistreringer i sygehussektoren
- Anders Ringgaard Kristensen, KU og Søren Højsgaard, AAU blev kontaktet, dels for at få deres umiddelbare indspil og dels for at bede deltage i en workshop senere på året
- Møde om Big Data med folk fra DTU.

En opsamling på første vidensopsamling ses her.

**Fokus på mulige analysemetoder**

- I de to artikler Caraviello et al (2006)[1] og især McQueen et al (1995)[2] inddrages Alternating Decision Tree Algorithm (ADTree) i deres analyse, implementereret i Weka. Weka er et Machine Learning software, som frit kan downloades og anvendes fra en hjemmeside[3] under University of Waikato, New Zealand. Statistiker Jørgen Nielsen undersøgte hvilke muligheder der lå i dette værktøj, og opsamlede viden i notatet her.
- Der blev afholdt en workshop med deltagelse af den interne projektgruppe, ekstern projektdeltager Søren Østergaard fra Foulum samt 5 eksterne eksperter på området (Søren Højsgaard + 3 andre statistikere fra Institut for matematiske fag, Aalborg Universitet samt Anders Ringgaard Kristensen, Sektion for Produktion og Sundhed, Københavns Universitet).
  Workshoppen blev afholdt d. 19. september på Videncentret for Landbrug, med følgende punkter på dagsorden:

  - Projektbeskrivelse og strategi for statistiske analyser i DMS (MLA)
  - Eksempel på mulige data og tidligere implementeret analyse: ReproAnalyse (JNI)
  - Analyser til kombination af informationer på enkeltdyr og bedriftsniveau (ARK)
  - Input til mulige analysemetoder set fra AAU's synspunkt (SH)
  - Beskrivelse af analysetyper i relation til tabellen 'Karakteristika'* (alle)
  - Opsamling på workshop og fremtidige samarbejdsmuligheder

[1] Analysis of Reproductive Performance of Lactating Cows on Large Dairy Farms Using Machine Learning Algorithms. D. Z. Caraviello, K. A. Weigel, M. Craven, D. Gianola, N. B. Cook, K. V. Nordlund, P. M. Fricke, and M. C. Wiltbank. J. Dairy Sci. 89 (2006), 4703-4722.
[2] Applying Machine Learning to Agricultural Data. Robert McQueen, Stephen R. Garner, Craig G. Nevill-manning and Ian H. Witten. Computers and Electronics in Agriculture 12 (1995), 275-293.
[3] http://www.cs.waikato.ac.nz/ml/weka/

* **Karakteristika, som analysemetoderne skal vurderes på baggrund af**

| **Karakteristika** | **Forklaring / kommentar** |
|---|---|
| Relation til én responsvariabel eller sammenhæng mellem flere? | Betragter analysemetoden én responsvariabel, som skal forklares af en eller flere variable, eller beskriver analysemetoden sammenhængen mellem en gruppe af ligeværdige variable? Eller…? |
| Varierende antal variable | At analysemetoden kan håndtere, at det antal variable, som indgår i analysen, er forskellig fra besætning til besætning. |
| Missing values | Kan analysemetoden håndtere at der i større eller mindre grad kan være manglende værdier, f.eks. for enkelte køer? |
| Afhængighed mellem variable | Hvor robust er analysemetoden til at håndtere tilfælde, hvor forklarende variable er afhængige? |
| Vekselvirkninger mellem variable | Hvordan kan vekselvirkninger håndteres? – altså at effekten af to variable sammen, kan være mere end summen af de to hver for sig. |
| Varierende typer af variable | Kan analysemetoden håndtere 0/1-variable, kategoriske, kontinuerte osv.? Både i evt. respons og forklarende variable? |
| | |
| Tidseffektiv | Forventer vi, at analysemetoden tidsmæssigt kan afvikles på et datasæt i løbet af få sekunder, så resultatet kan vises til landmanden inden for et minut efter, at han har anmodet om analysen. Eller kan vi finde andre løsninger? |
| Robust / selvkørende | Forventer vi, at analysemetoden kan afvikles og give et svar uden at der skal menneskelig assistance til at køre analysen? Hvis ikke – kan vi så gøre noget andet? |
| Formidling af resultat | Kan resultatet af analysen omsættes til en overskuelig oversigt? Grafisk form? Et handlingsanvisende resultat? En operationel eller strategisk beslutning for landmanden? |
| Validitet af analyseresultat | Når analysens resultat studeres ude på staldgangen – virker det så troværdigt? |
| Ressourcer til implementering | Kan VFL, Kvæg selv stå for implementeringen? Eller skal vi have ekspertbistand? Som konsulenthjælp eller som en del af et nyt projekt? |
| Ressourcer til drift/overvågning | Kan VFL, Kvæg selv stå for overvågningen og håndtere evt. fejl i den løbende drift? Eller skal vi have ekspertbistand? Hvordan kunne det etableres? |

Resultatet af workshoppen var, at især to metoder kunne være interessante at gå videre med:

1. Benytte en populations-prior til Bayesiansk inferens om en besætnings parametre, som en enkel og oplagt forbedring af f.eks. en logistisk regression
2. Benytte en træ-teknik (klassifikations-træ / statistical learning) til analyse af en besætnings data.

I forbindelse med metode 1) tilbød Anders R. Kristensen (ARK) et R-program der kunne demonstrere metoden.

- Opsamling på workshop blev efterfølgende gennemført af Søs Ancker (mla), Jørgen Nielsen (JNI) og Søren Østergaard (SOO), med henblik på
  - At fastlægge hvilken konkret analysemetode vi skal gå videre med i projektet
  - At aftale hvordan vi bedst muligt anvender de resterende resurser i projektet

Det blev besluttet, at tage imod ARK's tilbud om at stille en R-kode til rådighed for os, som illustrerer hvordan metode 1 virker. JNI skulle efterfølgende bruge besætningsdata, responser og forklarende variable fra ReproAnalyse, som grundlag for R-koden vi fik af ARK, og arbejde videre med denne metode med henblik på at føre eksemplet videre, så det støtter op om udpegning af faktorer med betydning for en besætnings ydelsestab.

**Arbejdet med metoderne efter workshop**
Det videre arbejde med metode 1 fremgår her. Sigtet var, at nå så langt som muligt inden det afsluttende projektmøde, hvor denne del-opgave skulle præsenteres for projektgruppen.

Ovennævnte metode 2, kan være et godt værktøj, når relevante forklarende variable skal udpeges. Ved hjælp af klassifikations træer, styrkes grundlaget for udpegning af variable, idet metoden synliggør vekselvirkninger i data vi bør tage højde for. I et første step i den retning, er fokus rettet mod responsen 'Ydelsestab' på ko-niveau. Definition af respons samt oversigt over forklarende variable forventes endeligt gennemført i starten af 2015. Som grundlag for arbejdet i 2015 udarbejdes et dokument, hvor muligheder for at dette kan lade sig gøre / hvilke udfordringer vi evt. står overfor, skitseres her.

**Identifikation af ydelsesfald på besætningsniveau**
I regi af projekt 2365, Multifaktorielle besætningsproblermer er der lavet en indikator for ydelsesfald i mælk på driftsenhedsniveau, hvor en 'alarm' afhænger af hvor stor en procentvis afvigelse den enkelte landmand vil acceptere. Der ligger således ikke en statistisk analyse til grund for denne alarm. I nærværende projekt har vi interesse i at få udpeget besætninger/driftsenheder med ydelsesfald på baggrund en statistisk sikker vurdering, hvorfor der er gennemført et stykke arbejde i samarbejde mellem VFL Kvæg og AgroTech på dette punkt, som beskrevet her. Den 18/12 ' 14 afholdes møde mellem de to parter, hvor AgroTech's leverance overdrages. Udviklingsarbejdet af en model til udpegning af et ydelsesfald på driftsenhedsniveau, er således gennemført, og kan bruges i 2015 på et projekt-datasæt.

I forhold til en statistisk sikker udpegning af ydelsesfald på driftsenhedsniveau er der et sammenfald af interesser mellem nærværende projekt og projekt 2345, Værktøjer til overvågning'. Implementeringen af den statistiske model i Kvægdatabasen med henblik på en generel overvågning af driftsenhederne, ligger derfor i regi af projekt 2345 efter 2014.

**Afrunding af projektet**
- Projektåret blev afsluttet med et møde for projektgruppen d. 18. december, med følgende deltagere

  Søs Ancker (projektleder, VFL Kvæg)
  Søren Østergaard (ekstern projektdeltager, AU)
  Jørgen Nielsen, Erik Rattenborg, Dorte Bossen og Lars Arne Hjort Nielsen (projektdeltagere, VFL Kvæg).

  Dagsorden:
  1. Præsentation af arbejdet med at benytte en populations-prior til Bayesiansk inferens om en besætnings parametre, som en enkel og oplagt forbedring af f.eks. en logistisk regression (JNI)
  2. Udkast vedr. definition af responsen 'Ydelsestab på ko-niveau' (muligheder, begrænsninger og udfordringer) og udkast over forklarende variable (LAN og DOB)
  3. Næste projekt-år – resurser og mål

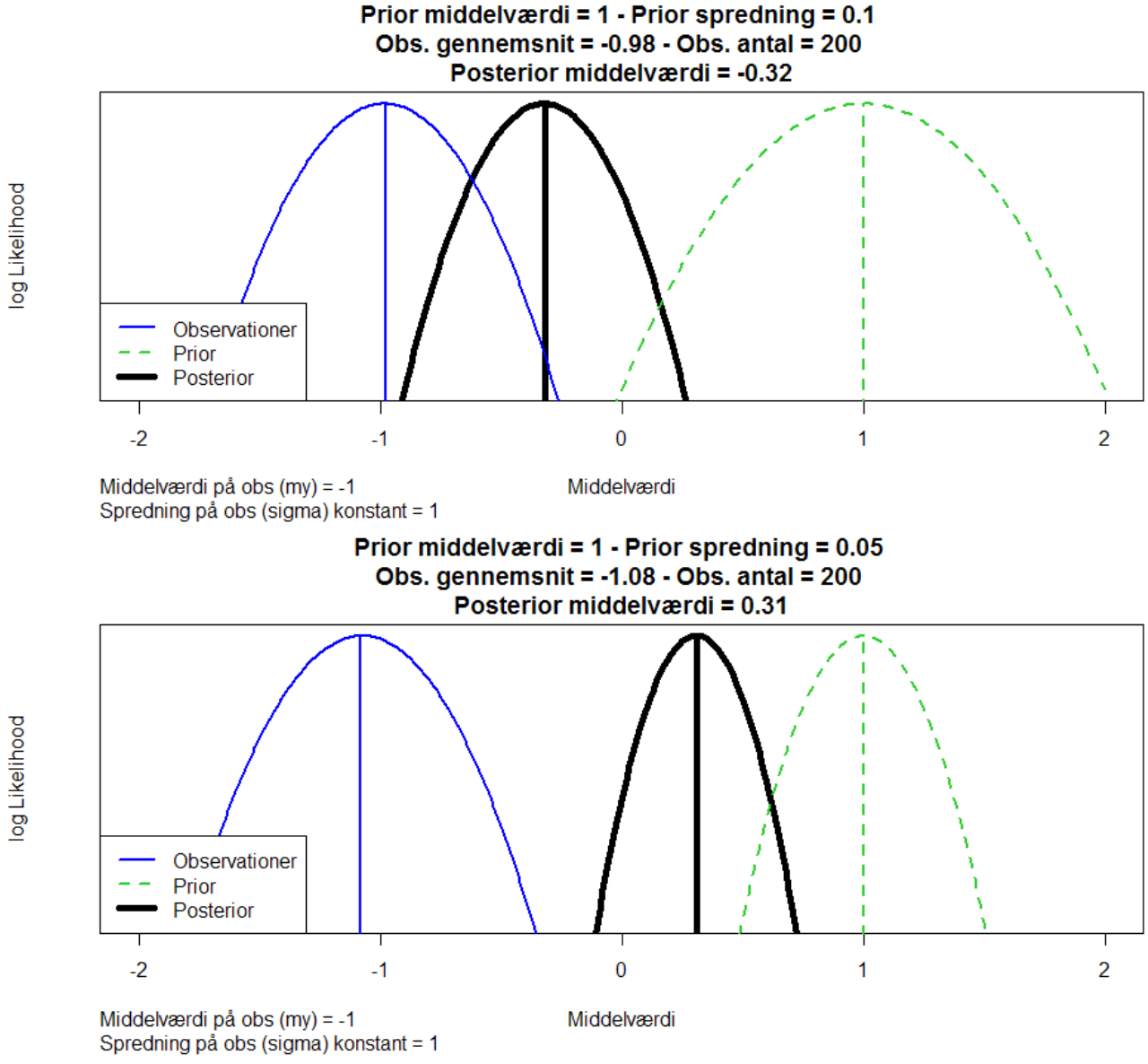
Præsentationer fra afsluttende møde ses her

## Konklusioner fra projektmøde 18/12 2014

**Vedr. valg af prior**

Det blev diskuteret, hvilken 'prior' der skal vælges, til korrektion af 'den lokale sandhed' – om den skal være populations- eller besætningsspecifik.

SOO foreslog at man laver en prior på baggrund af f.eks. 200 tilfældigt udvalgte besætninger fra Kvægdatabasen. Dette datagrundlag vil give en prior med en relativt stor spredning (f.eks. som skitseret i den øverste figur), men det er jo 'den lokale sandhed' vi er interesseret i, så derfor giver det udmærket mening at vi ikke vælger / definerer en prior der har en meget lille spredning, som dermed vil 'trække'/påvirke den observerede situation i besætningen markant (f.eks. som skitseret i den nederste figur).




*Figur 1 Eksempler på prior + data = posterior*

**Vedr. variable**

- Hvad er det der påvirker om en ko ligger over eller under dens målydelse ved sidste ydelseskontrol?
- Hvordan ser de sidste 3 mdr. ud, når man holder målydelsen op imod laktationsdage?
- Mulige responsvariable
  - Opnået minus målydelse ved seneste ydelseskontrol (ved mælkemålere måske 7 dages gns.)
  - Hvilket fald har koen haft de seneste 3 mdr. når der er korrigeret for ydelseskurvens form?
  - Estimeret topydelse og estimeret persistens for 1.kalvs, 2.kalvs og ældre køer