Foreløbig

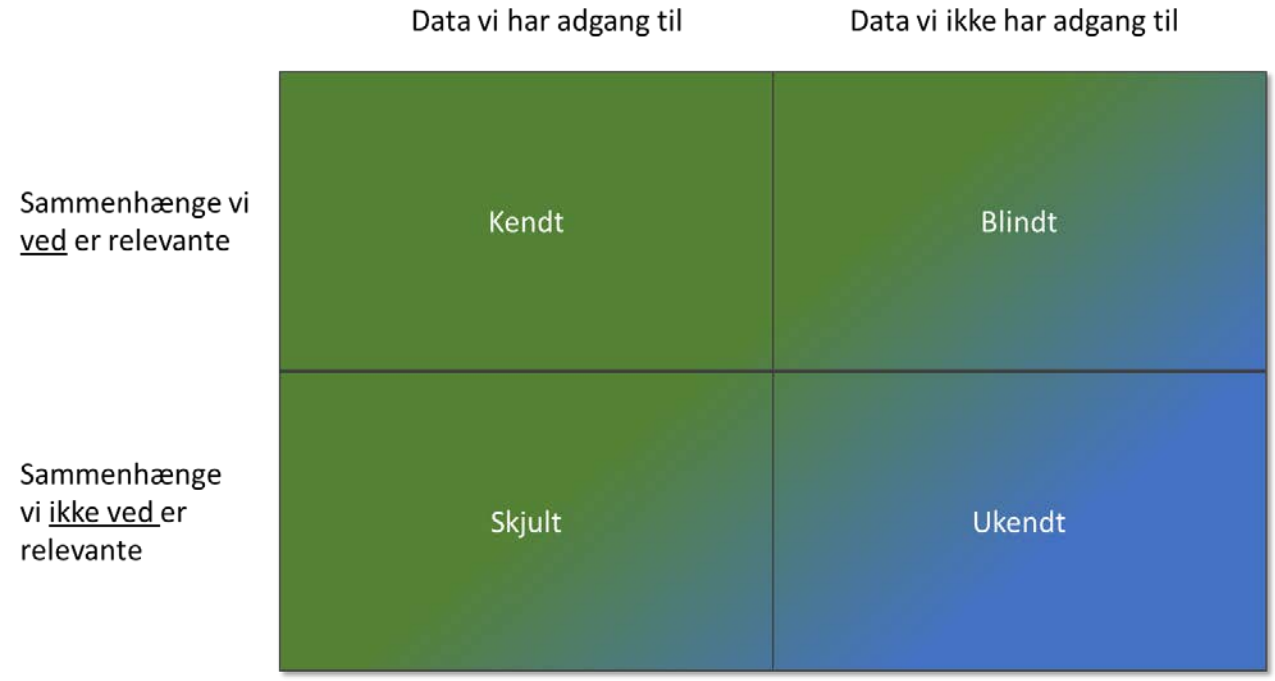
Nyttiggørelse af Landbrugets Datagrundlag – Fælles Datasæt

Baggrund.

Arbejdet med analyse af landbrugets data er historisk set en aktivitet som primært er pågået indenfor de enkelte fagretninger. Der udveksles i et vist omfang en mindre mængde forud definerede data mellem disse, feks afleveres data fra Kvægdatabasen i faste kørsler til Ø90 Økonomisystemet.

Med de stadigt mere avancerede og tilgængelige værktøjer til analyse opstår imidlertid også ønsket om på mere ad-hoc præget vis at kunne foretage analyser som ikke bare rapporterer men også på eksplorativt søger årsagssammenhænge i data. Det stiller krav om en tværgående adgang til data, også til data som ikke umiddelbart kan ses at have en indbyrdes sammenhæng.

Det kan illustreres på følgende måde:



For at kunne flytte vores viden fra det blinde og det ukendte felt må vi altså gøre så megen data som muligt tilgængelig så vi med analyser kan komme fra til endnu mere"Kendt" viden.

I projektet er arbejdet på en case der har kunnet illustrere nogle af de tekniske og organisatoriske overvejelser samt egenskaber ved at skabe et fælles datasæt af landbrugsdata som kan bruges mere bredt end de enkelte systemer i dag giver mulighed for.

Teknik.

Teknisk set er det en relativt ukompliceret opgave at skabe et sådant fælles datasæt. Det kan baseres på kendt teknologi som et såkaldt "Data Warehouse" hvortil fagsystemer og andre kilder føder data ind

via processer kendt som Extract/Transform/Load (ETL). Der findes allerede en række værktøjer og metoder til rådighed både til datalagring og ETL. Indenfor Kvæg- og Økonomi-området findes i dag dedikerede datawarehouses som kan integreres ind i det samlede datasæt, og integration af Mark-data er også umiddelbart muligt.

Svineområdet, som i dag ikke har en egentlig central dataløsning, vil udgøre en lidt mere teknisk kompleks løsning for at integrere til de decentrale datalagre. Det vil dog svare til forretningsløsninger som er kendt i andre sammenhænge, feks aggregering fra in-store systemer i detailhandelen.

Selve den tekniske implementering kan baseres på en række anerkendte datalagringsteknologier som vælges ud fra den sædvanlige definition af "Big Data": Volume, Velocity, Variety. Definitionen kan suppleres med Variability, Veracity, Visualization, and Value, hvoraf de tre sidste dog mere har indflydelse på den organisatoriske vinkel jvf næste afsnit.

Det fælles datasæt kan beskrives som en intern data service for analytikere, på linie med feks et mail- eller dokument-system. Som en intern service kan også de eksisterende systemer til legalisering og sikkerhed bringes i anvendelse. Dette står i modsætning til den nuværende situation hvor data er placeret vanskeligt tilgængeligt i forbindelse med de operationelle systemer udenfor de interne netværksadgange.

Servicen kan anvendes analytisk af kendte værktøjer som SAS, R. Excel, PowerBI mfl i og med at disse systemer har en integreret adgang til mange forskelige typer af datakilder.

Til brug for case-studiet er der i projektet etableret en intern test-dataserver, baseret på Microsoft Sql-server, med et fælles datasæt baseret på Økonomidatabasen samt Kvægdatabasens nøgletal om produktionen. Analysen, inkl grafer, er foretaget med R som primært værktøj.

Organisatoriske forhold.

Adgangen til data kan gøre det hurtigere og billigere at foretage analyser, men selv om fri adgang til al data kan synes som en optimal vision er det imidlertid vigtigt for at sikre værdi af data at der etableres en organisering omkring dataanvendelsen eller med andre ord at der er tale om "Managed Self Service".

Det indebærer at data skal være "kuraterede", dvs forvaltes i en proces af selektering, evt aggregering og anonymisering, beskrivelse og kvalitetskontrol. Ligeledes vil sikkerheden være en del af kurateringen.

Brugerens erfaring med dataanalyse og præsentation spiller også ind og det er særdeles vigtigt med den lette adgang til data at konklusioner og præsentationer hviler på et korrekt, valideret statistisk fundament, feks via peer reviews.

Af den gennemførte case fremgik det klart at eftersom de involverede systemer hver især er komplekse, kan det være vanskeligt – eller direkte føre til misvisende konklusioner - blot at have data til rådighed uden et dybere kendskab til hvorledes disse faktisk opstår i kildesystemerne. Tværfaglige analyser må således altid også bygge på tværfagligt system- og datakendskab og tværfaglig kvalitetssikring for at kunne udrede validiteten af de foretagne analyser. Det vil hyppigt være nødvendigt at gå tilbage til oprindelige dataleverandører for at udrede sammenhænge og det vil derfor være en kontinuert proces at fagsystemet supporterer dataanvendelsen.

Løbende opdateret datadokumentation, også kaldet metadata, i struktureret form vil skulle være en del af det fælles datasæt som baggrundsviden for analytikeren der ønsker at arbejde med de lagrede data.

Med andre ord er det nødvendigt for et "Fælles Datasæt" service at være klart forankret organisatorisk, både hvad angår den daglige drift, opdatering og opfølgning og hvad angår de data der afleveres af kildesystemerne. Det indebærer også en forpligtelse fra kildesystemerne til at informere om nye data og definitioner idet disse nu ikke længere blot har betydning internt i systemet, men kan influere på de konklusioner andre måtte træffe på baggrund af data der er til rådighed.

Case-studie: Fremstillingsprisen.

Baggrund.

Med gennemgangen af nedenstående case har vi ønsket en praktisk tilgang til hvordan en analyse ved hjælp af det fælles datasæt kan gennemføres.

Problemformulering.

Som case for evaluering af protoypen på et fælles datasæt blev udvalgt en analyse af fremstillingsprisen for mælk som er en kombination af produktionstekniske og økonomiske faktorer, her med kilde i Kvægdatabasen og i Ø90.

Fremstillingsprisen er et nøgletal i opgørelser på tværs af erhvervet og i rådgivningen af den enkelte landmand. Det samler omkostningen på input-faktorer og opgøres feks kr pr kg mælk, pr kg byg osv.
I landbruget kan salgsprisen i vidt omfang ikke påvirkes af den enkelte producent og derfor er der fokus på fremstillingsprisen.

Opgørelse af fremstillingsprisen er kompleks og debatteres løbende i det faglige miljø. Den fremkommer primært på basis af tal i Økonomidatabasen som dels rummer regnskabsoplysninger, dels generelle tekniske produktionsoplysninger som fødes i fagsystemerne og overføres via Ø90.

For at få en større gennemsigtighed og et mere detaljeret syn på fremstillingsprisen, vil det derfor være nyttigt at kunne analysere sammenhængen mellem denne og de tekniske forhold, via det fælles datasæt. Herunder kan det analyseres hvilke signifikante sammenhænge der måtte være til mange forskellige tekniske forhold for at evaluere sammenhænge og evt at finde sammenhænge der i dag ikke er kendte. Dette vil være en eksplorativ og mere agil tilgang end det er muligt i dag.

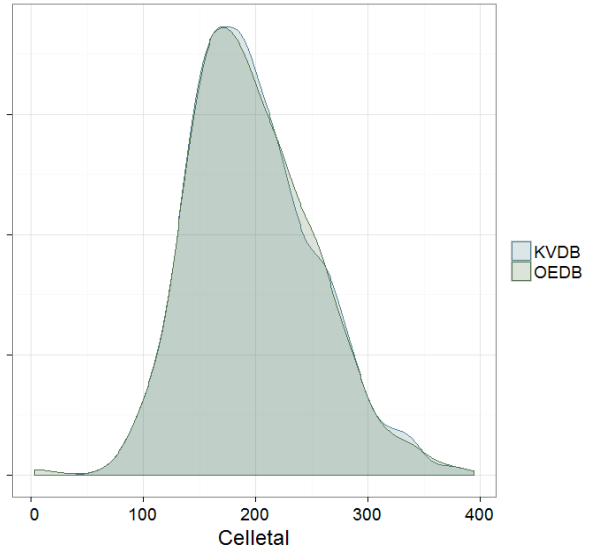
Det vil indledningsvist være vigtigt at fastslå kvaliteten af de data der kan udtrækkes i forhold til anvendelsen. Det vil være en analyseopgave i sig selv.

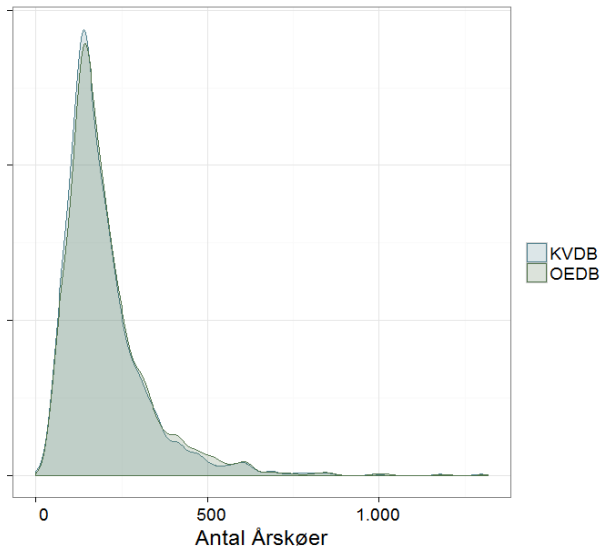
Derpå skal selve den samlede analyse bearbejdes. Resultatet vi først og fremmest være en større viden om sammenhænge generelt i erhvervet og der er således tale om overordnet ad-hoc analyse.

Indledende analyse.

Indledningsvis er foretaget en analyse for at afdække kvaliteten (her defineret som sammenhængen) i data ved at se på om oplysninger som kendes i begge datasæt stemmer overens.

Visuelt konstateres en høj grad af overensstemmelse mellem de to involverede datasæt, bla eksemplificeret ved celletalsopgørelse og antalårskøer i hhv Økonomidatabasen (OEDB) og Kvægdatabasen (KVDB), idet kurverne ses at være stort set identiske.





Analysen.

Via en lineær regressionsmodel (OLS) har vi undersøgt, om vi kan få en dybere forståelse for, hvilke faktorer der forklarer variationen i fremstillingsprisen på mælk.

Når der alene laves en regression på baggrund af variable fra kvægdatabasen til at forklare variationen i fremstillingsprisen fra Økonomidatabasen, fås en høj forklaringsgrad (R^2) på ca. 42 pct., når vi inkluderer data vedrørende foderforbrug og foderomkostninger. Hvis vi ser bort fra disse data, får vi en R^2 på ca. 28 pct. Selv om forklaringsgraden er lavere, når vi ser bort fra foderdata, er denne alligevel taget med i analysen, da der er en del ejendomme, der mangler disse oplysninger. Ved at koble oplysninger fra kvægdatabasen og nogle få variable fra økonomidatabasen kan vi forklare over 65 pct. af variationen i frem-

stillingsprisen. Hvis man alene ser på det rent statistiske, er der tale om høje forklaringsgrader i alle tre tilfælde, hvor sidstnævnte jo især skiller sig ud.

I analysen er det valgt at fokusere på regressionen, som indeholder data vedrørende foder, da foder er den største omkostning i fremstillingsprisen. I regressionen indgår 798 ejendomme.

Analysen viser bl.a., at variationen i celletal er med til at forklare variationen i fremstillingsprisen, og regressionen viser, at celletallet påvirker fremstillingsprisen positivt. Det betyder altså, at et lavt celletal er med til at holde fremstillingsprisen nede. Ud fra en rent faglig betragtning kan det måske virke mærkeligt, at celletallet i sig selv har en positiv indflydelse på omkostningerne. Hvis man ser det i et lidt bredere perspektiv, kan celletallet være et udtryk for det ofte anvendte begreb "management", som vi i økonomiopgørelser traditionelt har haft svært ved at få greb om. Hvis vi kan sige, at celletallet er udtryk for "management", giver celletallets forklaringsgrad god mening, fordi det betyder, at god management (= lavt celletal) medfører lav fremstillingspris.

Et produktionsnøgletal som energiudnyttelse har en negativ koefficient, hvilket populært sagt betyder, at en høj energiudnyttelse af foderet betyder lav fremstillingspris. Det vil sige, at jo bedre køerne er til at udnytte den energi, der er i foderet, jo lavere er fremstillingsprisen. Det hænger godt sammen med, at koefficienten for foderomkostninger pr. kg EKM er positiv, hvilket betyder, at en lav foderomkostning giver lav fremstillingspris.

Det er vigtigt at huske, at regressionen bygger på alt andet lige. Det vil sige, at hvis f.eks. foderomkostningen pr kg EKM sænkes, så viser analysen, at fremstillingsprisen alt andet lige vil falde. Det virker logisk. Modellen tager dog ikke højde for, om eventuelle andre omkostninger stiger eller om det har konsekvenser for mængden af mælk. Lavere mælkemængde vil alt andet lige få fremstillingsprisen pr. kg EKM til at stige. Og så er spørgsmålet, hvilket effekt der er stærkest. Det svarer denne analyse ikke på, og det afhænger naturligvis også af, hvilket udgangspunkt den enkelte ejendom har.

Konklusion.

Med case-studiet er der, ud over det interessante i den konkrete kontekst, indhentet værdifuld viden om forudsætningerne for hvorledes et fælles datasæt vil kunne bringes i spil for at skabe en øget viden og værdi på tværs af faggrene når der skal arbejdes holistisk og eksplorativt med analyser. Det vil værdifuldt for analytikere på tværs at have adgang til data som hurtigt og effektivt kan be- og afkræfte hypoteser uden at skulle igennem komplekse ad-hoc udtræk fra fagsystemer.