SEGES INNOVATION

RAPPORT

Udgivet 31. december 2022

Overvejelser omkring model- og metodeanvendelse samt deres muligheder og afgrænsninger

Forfattere: Mikkel Krabbe, Jørgen Nielsen, Katarina Nilsen Dominiak, Morten Nyland Christensen, Lars Arne Hjort Nielsen & Lene Bruun Siriwadhananuraks

Seges Innovation P/S

STØTTET AF

Promilleafgiftsfonden for landbrug

Indledning

Der er i "Prognoser for vurdering af bedriftens fremtidige økonomiske situation" arbejdet med, hvilke modeller og metoder, der kan udgøre den bagvedliggende motor i at kunne give landmanden og vedkommendes rådgivere et stærkt grundlag til at tage beslutninger for landbrugsbedriften. Det kræver forskellige teknikker, som er forsøgt indarbejdet som en del af prototypen. Notatet her viser de model- og metodeanvendelser, der er arbejdet med og diskuterer, hvilke muligheder og begrænsninger de giver.

Indhold

Model- og metodeanvendelser

I de næste afsnit er der redegjort for de model- og metodeanvendelser, der er arbejdet med i dette projekt. Afsnittene kan læses uafhængig af hinanden, men har hver især været anvendt i forskellige faser af arbejdet med prognoser i projektet.

Machine learning teknikker til opkvalificering af data

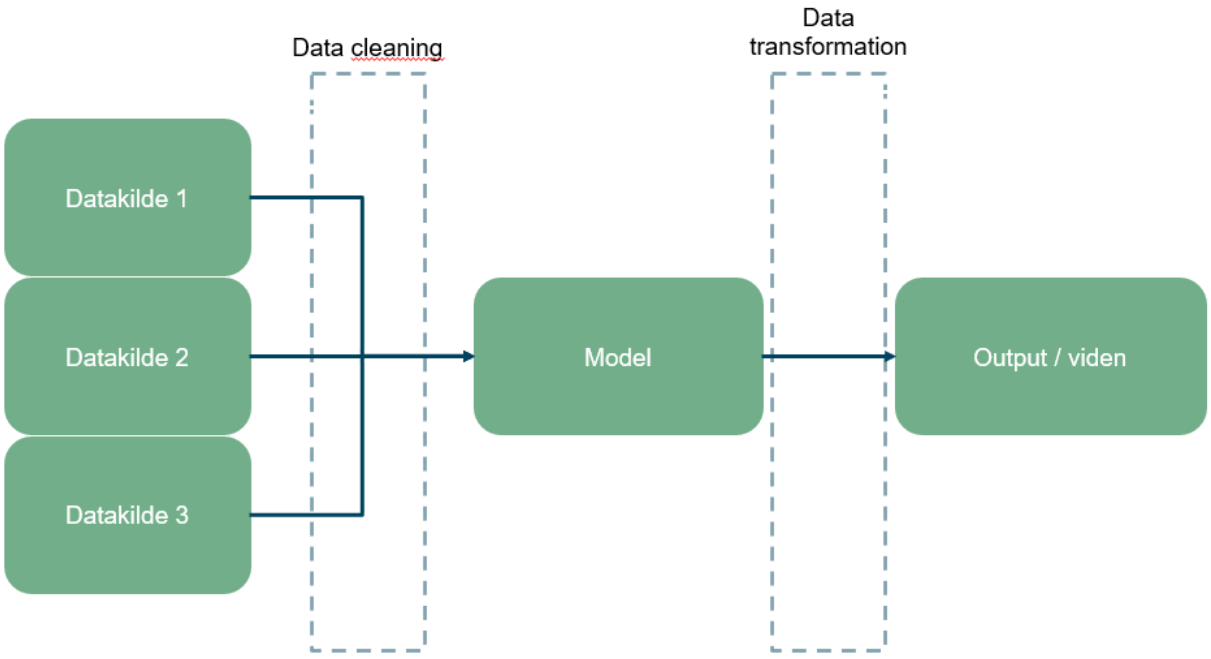
Dette afsnit omhandler brugen af Machine Learning til kategorisering af ø90 bilagsdata. Det har til formål at afdække problemstillingerne forbundet med kategorisering og Machine Learning.

Afsnittet vil afdække:

- Afgræsninger af prognoseværktøj
- Hvad er formålet med kategorisering af bilagsdata
- Gennemgang af metodevalg
- Udfordringer ved data og metode
- Det næste trin i "data til viden"

Formål med teknikken

"*Fra data til viden*". En del af formålet er at trække viden ud af data, hvor denne viden bidrager til et fremadrettet sigte, en prognose. Hvordan kan vi bruge historisk data til at hjælpe landmændene med at tage beslutninger?



Figur 1 – visualisering af en mulig data til viden proces

Figur 1 viser en af de mange mulige processer data kan gå igennem for at nå til et stadie, hvor det er muligt at danne viden. De analyser, der er lavet og beskrevet i dette notat, har fokus på datakilder, datarensning(cleaning), modellering og outputtet.

Afgrænsninger

I dette afsnit afgrænses undersøgelsesområdet til at omhandle:

- Kvægbrug

- Foderdata

Ved at være meget specifikke i, hvilke typer af data, der analyseres, så bliver outputtet af analyserne des stærkere. Der er ofte et trade-off i form at manglende ekstern validitet og/eller ekstrapoleringsmuligheder. Det vil ikke nødvendigvis være muligt at bruge et værktøj, der er specialbygget til kvægbrug, på slagtesvin. Der er forskellige underliggende mekanismer og informationer i data på tværs af f.eks. forskellige driftsformer.

I det efterfølgende er der set på foderindkøb, med henblik på at kunne give prognose på udvikling og optimering af foderomkostninger.

Der er findes en Nordisk foderdatabase (Norfor), hvor forskellige fodermidler og deres sammensætning af næringsstoffer er samlet til brug i foderoptimering hos nordiske mælkeproducenter. Der er cirka 8.900 forskellige fodermidler i databasen.

Datakilder

1. Bilagsdata fra ø90. Det er en database af information, hvor informationerne bliver lagret helt ned på de enkelte fakturalinjer.
    a. Databasen indeholder bilagsdata med informationer om:
        i. Landmanden / gården (ident)
        ii. Posteringstekst (den fritekst som udsteder / modtageren af bilaget har skrevet)
        iii. Posteringsnumre – det er helt ned på kontonummer niveau.
        iv. Beløb på bilaget
        v. Enhed (kg. Hkg, stk. osv.)
        vi. Flere oplysninger, som ikke er direkte relevante for analysen, derfor undlades en yderligere beskrivelse.
2. Norfordatabasen. Der er cirka 8.900 fodermidler gemt i Norfordatabasen.
    a. Indeholder information om:
        i. Navn på fodermidlet.
        ii. Foderkode, som også beskriver typen af fodermiddel
        iii. Hvilken sammensætning af næringsstoffer, der er indeholdt i fodermidlet.
3. Fodermiddelsgruppe.
    a. Der er fortaget en faglig vurdering af, hvor mange fodermidler en bedrift i hovedtræk fodrer med samtidig. Hvilke er fodermidler, der komplementerer hinanden, men ikke erstatter hinanden i fodringen. I en given fodergruppe må der gerne være fodermidler med forskellige nærringsparametre, men det må ikke være sådan, at fodermidlerne i gruppen i en konkret bedrift bruges samtidig. Dette vil give en estimeret fejl ved opgørelse af foderbrug senere. Der blev defineret 22 fodergrupper til første test, baseret på de midler med største forbrug.

Analyse

Analysen er lavet ud fra succeskriteriet:

- Det skal være muligt at kategorisere posteringsdata på kvægbrugsfodermidler ind i en af de fagligt forudbestemte fodermiddelsgrupper, så der opnås struktur på posteringsdata.
    - Det vil give en tidslig struktur på data på bilagsniveau.
    - Struktur på posteringsdata vil være vidensberigende i sig selv
    - Det vil give adgang til et bedre datagrundlag for fremtidige analyser.
- Modellen skal kunne håndtere nye fodermidler / posteringstekster.

Analysen er opdelt i 3 elementer, ikke rangeret efter vigtighed, som alle indgår i en samlet løsning:

1. String matching
2. Annotering af norfordata i kategorier
3. Kategoriseringsalgoritme

String matching
Formålet har været at få skabt sammenhæng mellem navnene skrevet i posteringsteksterne (tekststrengene –"string") i bilagsdata og navnene på fodermidlerne i Norfordatabasen. Begge disse navnevariable i deres respektive datasæt er strengvariable. Der er ikke forudbestemt en struktur på strengene, hvorved der er meget stor forskel på, hvad der er skrevet, og hvordan det er skrevet.
Et eksempel på, hvordan navnene kan være skrevet i posteringsdatasættet, der er 230 unikke posteringsnavne som indeholder ordet "rapskage", og der er 23 unikke Norfor-fodermidler som indeholder ordet "rapskage". Kobling herimellem er ikke ligetil, da det er igennem posteringsnavnene vi skal finde informationen til at vide, hvilket fodermiddel, der refereres til i Norfor. Der er for store datamængder på tværs af hhv. posteringsdatasættet og Norfordatabasen til at kunne lave forbindelsen manuelt for hvert fodermiddel.

Posteringsdatasættet vil blive benævnt "foderdata" i den efterfølgende beskrivelse, og datasættet fra Norfordatabasen vil blive benævnt "Norfordata".

Data cleaning
Den første del i denne analyse omhandler rensning af foderdatanavnene og norfordatanavnene. Rensningen består i 2 dele. Først laves der en reduktion af informationerne i strengene, hvor strengene bliver reduceret ned til de 2 første ord, specialtegn bliver fjernet og tal bliver fjernet. Dette er en indgribende operation på informationen i strengene, men et nødvendigt trin i processen for at bruge informationen.
Den anden del af datarensningen involverer en reduktion i norfordatanavnene, hvor der er lavet specifikke indgreb på navene, med henblik på at konkretisere informationen.

Begge dele i datarensningen kan / vil have en indflydelse på, hvor godt der bliver matchet mellem foderdatanavnene og norfordatanavnene. Det er en meget omfangsrig opgave at rense strenge på den rigtige måde. Det er et udfordrende område i analysen, og kvaliteten af outputtet afhænger af kvaliteten af rensningen.

Matching strategi
Det er fundamental forskelligt at sammenligne tal med hinanden ift. at sammenligne ord med hinanden. Det er ikke muligt at bruge algebraikke operationer på ord. Ord kan matche perfekt ("Ord" og "Ord" matcher en-til-en, men "Ord" og "ord" matcher ikke en-til-en, da der er forskel i store og små bogstaver), men de kan også matche ikke-perfekt. Når der er tale om ikke-perfekte match(kobling) mellem ord, så bliver der ofte arbejdet med et afstandsmål. Et afstandsmål kunne være et beregnet tal for, hvor tæt på hinanden 2 ord eller 2 sætninger er. Når der er valgt et mål (metric) for at måle afstand mellem strengene, så vil det være muligt at fastsætte en grænseværdi, så 2 strenge vil matche med hinanden, hvis deres afstand er mindre end grænseværdien. Jo lavere grænseværdien er, jo stærkere (og derfor færre) matches vil opnås. Ved at hæve grænseværdien vil det være muligt at skabe flere matches, men der er større afstand mellem disse matches, og derved risiko for at koblingen ikke er korrekt.

Metoden der er valgt til at skabe match mellem foderdata og norfordata hedder "fuzzy join", hvor fuzzy benævner, hvordan det ikke er eksakte matches, der skabes.

Outputtet af denne fuzzy join metode giver derved for hver bilagsposteringstekst i foderdata et eller flere matches i norfordata. Muligheden for flere matches sker, hvis der er multiple navne i norfordata, hvis afstand til posteringsteksten er under den valgte grænseværdi.
Formålet med analysen var blandt andet at skabe struktur i posteringsdata, ved at skabe præ-defineret kategorier blandt norfordata. Multiple matches mellem foderdata og norfordata er forventelig inden for samme kategori.
Det ønskede output af string matching er, at hver posteringstekst i foderdata får tilkoblet en kategori fra norfordata – hvorved der er opnået struktur af posteringsdata.

Annotering af norfordata i kategorier
Norfordata er en datakilde, hvor der for hvert fodermiddel er et navn og tilhørende mængder af næringsstoffer. Der er ikke lavet struktur i norfordata i forhold til at være meget specifik på, at visse foderkoder kun refererer til en bestemt dyregruppe (kvier, slagtekalve, malkekøer), da de fleste fodermidler kan bruges bredt. Struktur kunne have været i form af at forskellige fodermidler som alle er navngivet ”goldko mineraler” er kategoriseret i en kategori ”goldko”, som så er forskellige fra kvie mineraler. Processen af at annotere data er vigtig, hvis vi ønsker at bruge supervised machine learning til at skabe en kategoriseringsalgoritme.

Formålet med annoteringen af norfordata i kategorier er at skabe struktur på det store datasæt. Det vil muliggøre at bruge supervised machine learning til at skabe en kategoriseringsalgoritme. En kategoriseringsalgoritme vil være bindeleddet imellem posteringsdata og struktureret kategorisering. Derved vil det være muligt at lave analyser på det strukturerede data. Et eksempel på fordelen ved en kategoriseringsalgoritme er, når en landmand skifter fodermiddel, så vil det være muligt at vide, hvad fodermidlet skal bruges til (hvilket hidtidigt fodermiddel det f.eks. erstatter).

Selve annoteringen er sket igennem 2 niveauer:
- Faglig ekspertviden af at skabe kategorier og inddele et sæt af 240 observationer i norfordata i kategorier.
- Struktur i navngivningen til at videreannotere cirka 500 yderligere observationer.

Det er en manuel opgave at annotere data ind i kategorierne, som er grundstenen for kategoriseringsalgoritmen.

Kategoriseringsalgoritme
Formålet med kategoriseringsalgoritmen er at skabe struktur på alt posteringsdata, både det, der er tilgængeligt nu og fremtidigt data. Strukturen skal muliggøre prognoseanalyser i fremtiden. Der er mange måder at skabe en kategoriseringsalgoritme på, nogle af disse vil blive beskrevet.
- Unsupervised machine learning algoritmer:
    - Algoritmer, hvor der ikke er specificeret en target-variabel. En target-variabel er en variabel, der indeholder den rigtige kategori. Data kan indeholde forskellige informationer, som algoritmen kan bruge til at gruppere data ind i homogene klynger.
    - Eksempler herpå kunne være:
        - Hierachical clusters (hierakiske grupperinger
        - K-means neighbours
    - Fordele ved unsupervised algoritmer:
        - Brugeren behøver ikke have kendskab til den rigtige gruppering
        - Der er ikke behov for at annotere data
        - De lader data tale selv
    - Ulemper ved unsupervised algoritmer:
        - De bliver ikke kontrolleret af brugeren i kategoriseringen
        - Ekspertviden vil ikke indgå i konstruktionen af kategorier

- Supervised machine learning algoritmer:
    - Kræver at en delpopulation af datasættet er annoteret, som kan bruges til at træne modellen til mønstregenkendelse for kategoriseringen.
    - Eksempler herpå kunne være:
        - Multiklassificeringsmodeller som Multinominal model eller Random Forest
    - Fordele ved supervised algoritmer:
        - De bliver kontrolleret ved stikprøver at kategoriseringen virker
        - Ekspertviden indgår i konstruktionen af kategorier
    - Ulemper ved supervised algoritmer:
        - Brugeren behøver kendskab til den rigtige gruppering,
        - Der er tidskrævende at annotere data

Der er valgt at bruge supervised machine learning algoritmer til kategoriseringsalgoritmen. Det er valgt på baggrund af en eksplorativ analyse af unsupervised algoritmer og behovet for at bruge ekspertviden i konstruktionen af kategorierne.

Valget af supervised machine learning algoritmer skabte behovet for annoteringen af et repræsentativ udsnit af datasættet, som kunne bruges til træningen af algoritmen.

Azure ML

Til at skabe kategoriseringsalgoritmen er Azure ML blevet taget i brug. Azure ML er Microsofts online machine learning portal. Det er et miljø skabt til at træne og afprøve forskellige modelspecifikationer og bruge computerkraft i skyen.
Der er blevet brugt 3 forskellige features i Azure ML:

- Designer:
    - Det er en drag&drop platform til at skabe en pipeline af data, hvor der er præ-specificeret et udvælg af forskellige kategoriseringsalgoritmer.
- Automated ML
    - Det en platform, hvor det er muligt at lade Azure ML træne mange forskellige modeller og vurdere, hvilken model, der har den bedste performance. Således at koblingen mellem norfordata og foderdata bliver bedst mulig.
- Deploy endpoints:
    - Den / de trænede modeller kan deployes (anvendes) som endpoint der er mulige at kalde og få respons fra. Det gør det muligt at give modellen nye data og få en kategori returneret (med tilhørende sikkerhed).

Udfordringer ved metodevalget

Processen med at skabe struktur på posteringsdata på bilagsniveau involverer udfordringer. Nogle af disse udfordringer er blevet beskrevet ovenfor, men vil blive opsummeret her:

- Data cleaning af tekst strenge fra posteringstekster:
    - Processen med at afskære information fra posteringsteksten på en systematisk måde, så det bliver muligt at finde partielle matches imellem posteringsteksten og norfordata, involverer mulige fejlkilder i:
        - Afskæres for meget af tekststrengen, så mistes brugbar information
        - Afskæres for lidt af tekststrengen, er det ikke muligt at skabe et fuzzy match
    - En metode til at overkomme nogle af udfordringerne er at kombinere data cleaning med kategorierne, så det bliver muligt at se på andelen af multiple matches, der omhandler forskellige kategorier.
    - Det vil forbedre data cleaning processen at inddrage ekspertviden til vurdering af afskæring af information.
- Annoteringen af norfordata i kategorier:

    - Processen af at manuelt annotere data med kategorier er lang, og der er mulighed for menneskelige fejl
    - Ved at præ-definere antallet af kategorier, så vil kategoriseringen være sårbar overfor fremtidige strukturelle ændringer i data, som kræver opdateringer af kategorierne.
- Kategoriseringsalgoritme
    - Det er en proces, der afhænger af annoteringen:
        - Er der annoteret ”nok” træningsdata til at fange de korrekte mønstre
    - Det er en kategorisering af eksisterende data og fremtidigt data, der er mulighed for klassificeringsfejl.

Outputprognose (DMS prognose)

Dette afsnit undersøger mulighederne for at tage eksisterende prognosemodeller og anvende dem i en større kontekst end anvendt i dag. Der findes i DMS et værktøj som oftest bruges til at bestemme fremtidigt salg af mælk og omsætning af dyr til anvendelse i et budget. Dette afsnit vurderer, om DMS prognose kan supplere en mere dynamisk økonomiprognose, hvor forskellige scenarier skal indgå.

Når der oprettes en ny mælkeprognose i DMS, tager denne prognose udgangspunkt i et nyt datasæt fra kvægdatabasen, som indeholder de dyr, som findes på bedriften i dag samt nøgletal på de seneste 12 måneders leverance af mælk samt 12 mdr. nøgletal for dødelighed, reproduktion og omsætning af dyr.
Udarbejdes der en prognose, som skal bruges i det kommende års budget, er det normalt, at der tages stilling til de enkelte nøgletal i forhold til ændringer, som forventes frem i tiden. Det er dog ikke nødvendigt at ændre i forudsætningerne for at afvikle en prognose. Prognosen kan beregnes uden at brugeren behøver at lave supplerende indtastninger. I dette tilfælde vil det være en passiv fremskrivning, baseret på de seneste 12 måneders information. Således forventes de næste 12 måneders produktion at bibeholde samme niveau for de fleste parametre, men for ydelsen er der indarbejdet en ydelsesstigning eller ydelsesfald ud fra de 6 seneste ydelseskontrolleringer i besætningen. Ligeså indgår afkommet fra de kommende kælvninger.

Det er også muligt at kopiere og genberegne en eksisterende prognose. I dette tilfælde bibeholdes alle egne indtastninger, men der trækkes ikke noget nyt datasæt, så en sådan beregning vil ikke give et frisk billede af produktionen på bedriften. Og derfor kan det ikke anbefales, at det er denne model, som anvendes til fremskrivning.

Afgrænsning af virkningsmetode
Det er i det nuværende DMS program ikke muligt at starte DMS prognoseberegningen automatisk, og det vil kræve udvikling af koden bag programmet at udvide DMS prognosen til denne mulighed. Hvis det i fremtiden bliver muligt at starte prognosen automatisk, vil der blive forhold som kræver opmærksomhed, men som dog ikke anses at give betydelig usikkerhed undtagen i situationer, hvor der planlægges store ændringer i antallet af dyr- eller produktionsmetoder.
Prognosen vil fx aldrig øge ko-antallet i forhold til det startgrundlag, som den henter. Står der fx 203 køer i besætningen, vil prognosen bibeholde dette antal af køer, med mindre at udsætningen bliver for lav og prognosen vil slagte køer, og dermed reducere ko-antallet. Udsætningen skal som minimum være 10% lavere end udsætningen sidste år. Prognosen vil ikke øge ko-antallet, uden at der laves en manuel indtastning.
Selvom prognosen baserer sig på de seneste data om leveret mælkeproduktion, kan der stadig opstå en pludselig stigning eller et fald i mælkeproduktionen, der hvor prognosen starter. Dette skyldes at prognosen forholder sig til antallet af malkende køer og deres ydelsesniveau. Ydelsesniveauet bestemmes ud fra de seneste 12 måneders mælkeproduktion, og ydelsesændringen ved de seneste 6 ydelseskontroller, hvor den seneste vægter mest. Det betyder, at drastiske udsving indenfor fx den sidste måned i mælkeproduktionen ikke fremskrives lige drastisk med en prognose.

Ydermere vil der være flere forhold, som gør nogle bedrifter mere egnet til en automatisk beregnet prognose end andre. At de er tilknyttet 11 gange ydelseskontrol giver mere sikkert estimat på mælkeproduktionen, at registreringer er gode (inseminering og drægtighed), og at deres praksis er konstant og de følger deres grundoplysninger fx med antallet af golddage. Desuden er det vigtigt at alle besætninger, som der går dyr ud og ind fra, er inkluderet i den driftsenhed som prognosen baserer sig på. Er dette ikke i orden, kan der ske manglende indsætning af nye 1. kalvskøer og slagtninger, koantal og mælkeproduktion bliver mere og mere forkert beregnet, des længere frem i tiden det beregnes.

Til sidst er der enkelte nøgletal, hvor det ikke er bedriftens egne tal som bruges. Det gælder anvendelsen af kønssorteret sæd, men der går min. 10 mdr. frem i tid, før det får nogen betydning (drægtighedslængde). Der anvendes også landstotaler på dødelighed, men det har også minimal betydning på nær hvis bedriften har en høj dødelighed.

Bidrag fra DMS-prognosen til den økonomiske fremskrivning

Prognosen kan give vigtige input omkring det fremtidige antal af dyr, både malkende og golde samt opdræt. Ligeså giver DMS prognosen den vigtige information om producerede mælkemængder og antallet af solgte og slagtede dyr. Derved kan prognosen være væsentlig i forhold til at bestemme fremtidige indtægter, men også til at fremskrive forbruget af foder m.m. til mælkeproduktion. I forhold til det fremtidige forbrug af foder vil der være stor sammenhæng til mælkeproduktionen og antallet af dyr.

Statistisk behandling af indkøbsmønstre til prognose

Dette afsnit ser på mulighederne for at behandle foderindkøbsdata, når de er kategoriseret efter anvendelsestype. Hvilket er med henblik på at fremskrive forbruget af foder til kvæg (og evt. andre husdyr) baseret på tidligere realisere forbrug.

Data omfatter bl.a.

- Mælkelevering
  - Faktisk leveret mængde mælk pr. dag, kg EKM (EnergiKorrigeret Mælk)
  - Budgetteret mængde mælk pr. dag, kg EKM
- Foderdata for 8-20 fodermidler pr. bedrift:
  - Indkøb af fodermidler – med angivelse af dato, kvantum og pris
  - Foderbudget – med angivelse af måned og forventet kg foder i alt
  - Status = lagerbeholdning – men kun få gange på én af bedrifterne.

Budgettallene dækker både 2020, 2021 og 2022 – typisk dannet i slutningen af året forud.

Indkøb af fodermidler

Nogle fodermidler indkøbes fx hver anden uge, mens andre kun indkøbes et par gange om året. Der vil derfor være stor forskel på, hvor detaljerede oplysninger vi har om de forskellige fodermidler.

Antagelser om foderforbrug mellem indkøb

Ofte vil vi antage, at foderforbruget er jævnt mellem to indkøb – enten ved at det er fast i forhold til leveret kg EKM, antal dyr, eller endnu simplere: fast pr. dag. Således kan ét indkøb deles ud over dagene frem til næste indkøb. Det kan også være, at vi udglatter det anslåede forbrug over to eller flere indkøbsperioder.

Men uanset det, så vil det anslåede forbrug som udgangspunkt være baseret på en antagelse om, at *lagerbeholdning ved alle indkøb er det samme*. Det vil dog betyde, at hvis lagerbeholdningen umiddelbart inden indkøb nr. 2 reelt er væsentlig højere (lavere) end lagerbeholdningen umiddelbart inden indkøb nr. 1, så overvurderer (undervurderer) vi foderforbruget i perioden mellem indkøb nr. 1 og indkøb nr. 2.

Denne antagelse om stabil lagerbeholdning ved indkøb er på sin vis essentiel for nogle af beregningerne. Hvis antagelse ikke holder, så kan en del af beregningerne være behæftet med fejl. En sådan fejl kan korrigeres helt eller delvist på forskellig vis:

- At vi kender lagerbeholdningen for forskellige tidspunkter (lagerstatus) og korrigerer med den viden.
- At vi kigger på flere indkøbsperioder end blot én periode. På den måde bliver fejlen ofte mindre. Se senere afsnit om udglatning lag 1, lag 3 osv.

Fodermidler og beregnede størrelser

I dette afsnit kigger vi på forskellige beregnede størrelser, som siger noget om foderforbruget. Disse størrelser og definitioner af fodermængder kan teoretisk set bruges for hvert fodermiddel. Så afsnittet her har til formål at håndterede nogle beregnede størrelser systematisk. Og med en mere præcis forståelse af disse størrelser er det tanken, at vi kan komme tættere på en prognosemodel.

For foder (og sådan set også for mælk) kan vi regne med både faktisk forbrug og budgetteret forbrug. Desuden i flere niveauer:

- Kg for en periode
- Kg pr. EKM, eller kg pr. dyr

- Kg pr. dag

Forbrugt værdi	Budgetteret værdi	Forklaring og enhed
PF_F(X)	PF_B(X)	Periodens foder, kg
FR_F(X)	FR_B(X)	Foder ratio, kg foder pr. kg EKM eller pr. dyr
DF_F(X)	DF_B(X)	Foder pr. dag, kg pr. dag

X ovenfor i disse udtryk angiver perioden, som foder er angivet for, eller beregnet ud fra.
Den simple version er 1 indkøbsperiode, altså en antagelse om, at landmanden fra et indkøb til næste forbruger præcis alt det indkøbte foder (dette kalder vi lag 1). En anden variation er, at vi udglatter over to af sådanne indkøbsperioder (dette kalder vi lag 2).

X kan være forskellige værdier:

X	Forklaring
1	Lag 1 = beregning inden for 1 indkøbsperiode
2	Lag 2 = beregning over 2 efterfølgende indkøbsperioder
3	Lag 3 = …
Kvt	Beregning over de indkøb, som har været inden for et kvartal
År	Beregning over de indkøb, som har været inden for et år
Tot	Beregning over alle registrerede indkøb
…	…

At benytte sig af X i forskellige versioner er nok mest relevant, når vi skal forholde os til det *forbrugte* foder, for det angiver hvilken periode, som vi har benyttet til at estimere den forbrugte mængde foder.

I det følgende angiver vi

DWi = daglig vægtning (fx mælkemængde i kg EKM, eller antal dyr)
∑DWi = periodens sum af daglige vægtninger (fx kg EKM, eller ko-dage)

Beregning af foder ratio
Foder ratio i en periode kan sættes til at være lig med periodens foderindkøb, divideret med periodens sum af daglige vægtninger. Dvs.

FR_F(X) = PF_F(X) / ∑DWi

Det kan således betegne foderforbrug, enten pr. kg EKM, eller pr. dyr pr. dag.

Formlen gælder for både _F og _B, altså hhv. faktisk og budgetteret mængde.

Beregning af kg foder pr. dag
Foder pr. dag i en periode kan sættes til at være lig med periodens foderindkøb, vægtet med den daglige vægtning (fx kg EKM eller antal dyr). Dvs.

DF_F(X) = PF_F(X) / ∑DWi * DWi
= FR_F(X) * DWi

Det er derfor også det samme som Foder Ratio * daglig vægtning.

Formlen gælder for både _F og _B, altså hhv. faktisk og budgetteret mængde.

Yderligere bemærkninger

Når det gælder disse forbrugte fodermængder, afhænger det altså hele tiden af, hvilken periode, som vi opsummerer foderet for.

Hvis man ønsker en meget simpel model, hvor man ikke vægter med hverken kg EKM eller antal dyr, kan man blot sætte DWi = 1 konstant, og dermed bliver ∑DWi lig med antal dage i perioden. På den måde opnår man, at DF_F(X) = FR_F(X), altså at foder pr. dag, kan beregnes præcis som foder ratio, FR_F(X). På den måde kan formler og beregner genbruges uanset, om man inddrager en daglig vægtning eller ej i sin model.

Bemærk, at lige som fodermængder kan være forbrugte eller budgetterede, så kan DWi (altså kg EKM eller antal dyr) også være enten faktiske eller budgetterede. På den måde kan notationen ovenfor i høj grad bruges for forskellige scenarier. Men det er muligt, at der stadig er basis for at præcisere ting ved den her valgte notation.

Daglig vægtning med kg EKM

Der er arbejdet med en vægtning, hvor der er anvendt kg EKM (EnergiKorrigeret Mælk). I dette afsnit ses der på dette arbejde. Når man har 2 kurver, som virker meget parallelle, kan det være muligt at den ene kurve kan bruges til at fremskrive den anden.

Inden vi her viser de to kurver, så kigger vi lidt på hvordan størrelserne for foderforbrug ser ud, når vi bruger kg EKM, som den daglige vægtning.

I det følgende angiver vi

DWi = EKMi = daglig mælkemængde, i kg EKM

∑DWi = ∑EKMi = periodens samlede mælkemængde, i kg EKM

Dermed er foderratio lig med foder pr. kg EKM. Dvs.

FR_F(X) = PF_F(X) / ∑EKMi

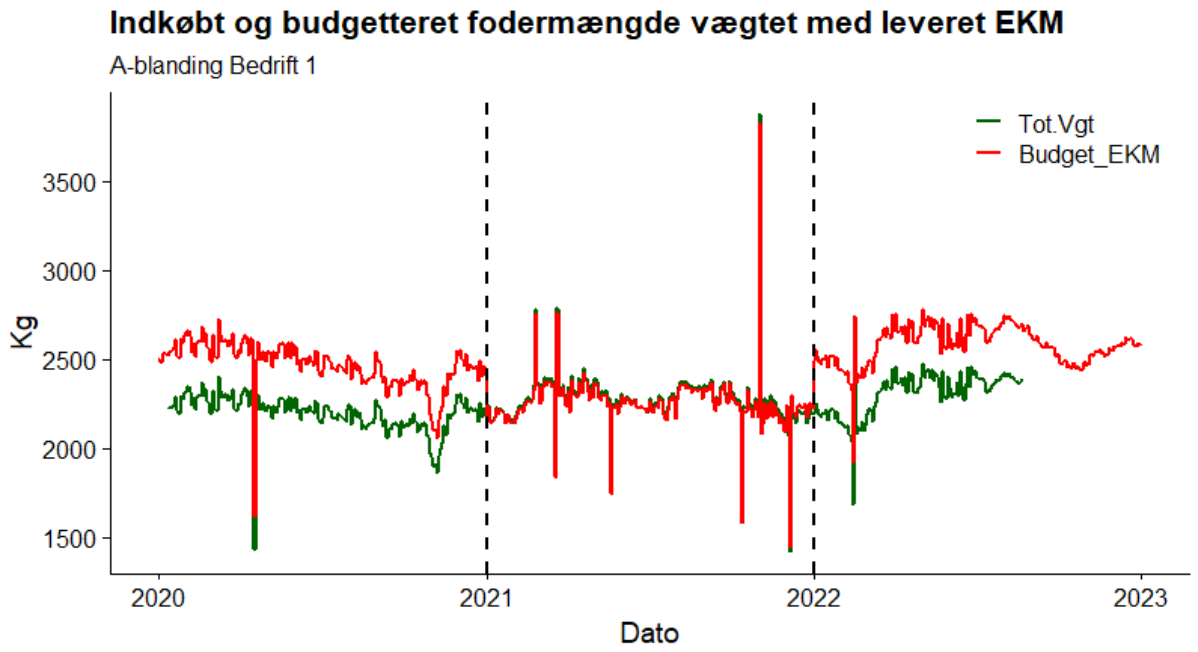
Og kg foder pr. dag er lig med foder ratio * daglig kg EKM. Dvs.

DF_F(X) = FR_F(X) * EKMi

Foderbudgettet er beregnet som budgetteret kg foder i løbet af måneden, altså PF_B(måned). Og når vi dividerer dette med måneds samlede kg EKM, får vi

PF_B(måned) / ∑EKMi = FR_B(måned)

Altså budgetteret foder ratio for måneden, kg foder pr. kg EKM.
De to kurver, som var tilnærmelsesvist parallelle, er de to kurver, som i *figur 3* er hhv. Budget_EKM (Rød kurve) og Tot.Vgt (Grøn kurve)



Figur 2 Total mængde indkøbt foder pr. år vægtet med daglig EKM (grøn kurve) og budgetteret fodermængde pr. år vægtet med daglig EKM (rød kurve) for det primære kraftfodermiddel 'A-blanding' i Bedrift 1. Indenfor år er kurverne tilnærmelsesvist parallelle. I 2020 og 2022 er der budgetteret med en højere fodermængde, end der er indkøbt. I 2021 er der stort sammenfald mellem indkøbt og budgetteret fodermængde. De lodrette, stiplede linjer markerer årsskifte.

Disse to størrelser kan vi nu med ovenstående notation regne lidt yderligere på:

Budget_EKM (kg pr. dag, Rød kurve)
= DF_B(måned)
= FR_B(måned) * EKMi

Tot.Vgt (kg pr. dag, Grøn kurve)
= DF_F(Tot)
= FR_F(Tot) * EKMi

Dvs. forholdet mellem de to størrelser er Budget_EKM (Rød kurve) divideret med Tot.Vgt (Grøn kurve)
= FR_B(måned) / FR_F(Tot)

Altså relationen mellem foderbudget pr. kg EKM (beregnet for måneden) og forbrugt foder pr. kg EKM (beregnet for den totale periode). Da forbrugt foder i dette tilfælde er beregnet for hele perioden, er størrelsen konstant i hele perioden. Det eneste, som svinger i perioden, er foderbudgettet, som skifter (lidt) niveau hver måned.

I forbindelsen med "Rød kurve" og "Grøn kurve" kan man også definere en størrelse K, som der kan regne lidt yderligere på og dermed få en mere præcis forståelse af:
K = Totalmængde foder / ∑ EKMi
= PF_F(Tot) / ∑EKMi
= FR_F(Tot)
Altså forbrug af foder pr. kg EKM beregnet for den totale periode. Og det er jo netop en konstant for hele perioden.

Valg af prognosevariabel
Ud fra ovenstående noter, og de overvejelser, som ligger bag, vurderer vi, at en egnet størrelse, når vi skal estimere og prediktere foderforbrug, er en størrelse som FR_F(X), hvor X så både kan være lag 1, lag 2 osv. eller år eller lignende.
Vi arbejder altså med størrelsen kg foder pr. kg EKM, og en prognose for foder pr. dag kan så fås ved at regne FR_F(X) * EKMi, hvor EKMi er forventet mælkemængde pr. dag.

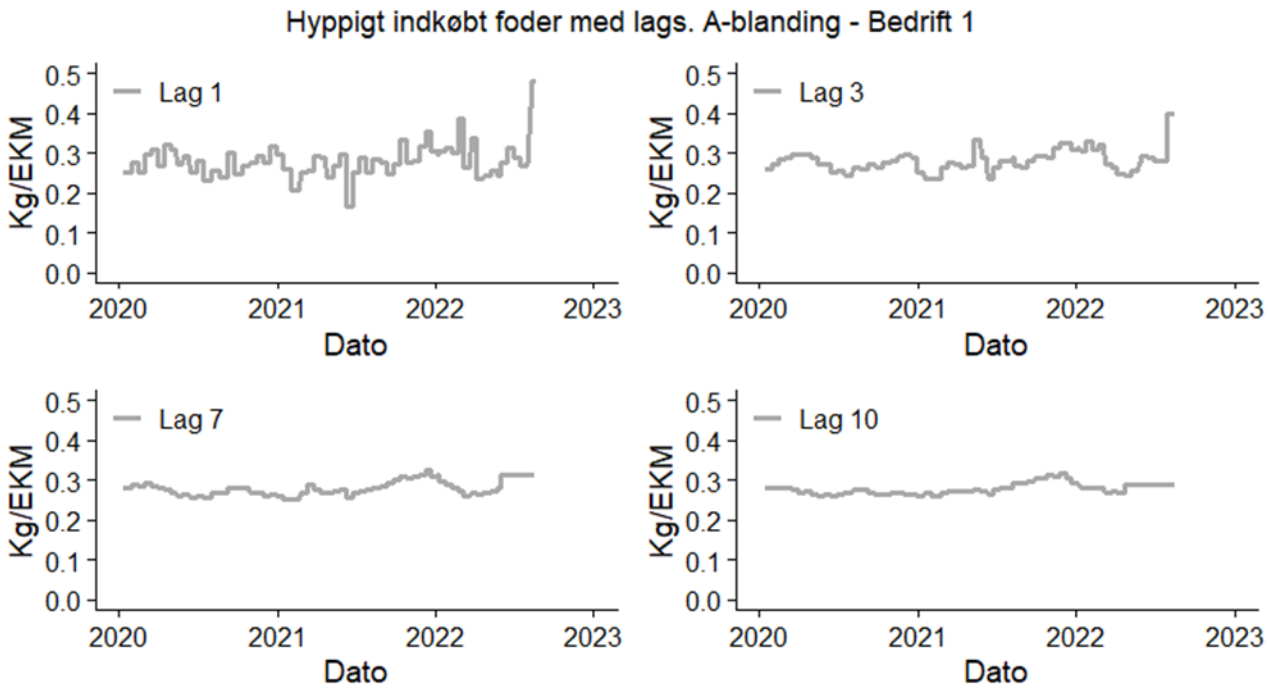
Hvis man af forskellige grunde ikke ønsker at benytte EKMi som den daglige vægtning – fx fordi fodermidlet ikke er til køer – så kan man evt. benytte antal dyr som daglig vægtning. Eller blot regne helt uden vægtning, DWi = 1, så det også fremgår af næste afsnit.

Udglatning af foderforbrug
Da både intervallet mellem foderindkøb, den indkøbte fodermængde og den leverede mælkemængde (EKM) kan variere over tid, vil foderforbrug pr. kg EKM variere relativt meget, hvis der ses på et kortere tidsrum. Tilsvarende vil foderforbruget pr. kg EKM være mere stabilt, når der ses på et længere tidsrum.
Dette er illustreret i *Figur 2* og *Figur 3* nedenfor. Figurerne viser foderforbrug (kg) pr. registreret kg EKM beregnet for to fodermidler over fire forskellige tidsrum.

De to figurer viser den samlede mængde af fodermidlerne 'A-blanding' (*Figur 2*) og 'Rapskagefoder' (*Figur 3*) delt med den samlede mængde leveret EKM registreret for følgende fire tidsrum:
i) perioden fra ét foderindkøb til dagen inden det næste (Lag 1),
ii) perioden fra ét foderindkøb til dagen inden det tredje indkøb herefter (Lag 3),
iii) perioden fra ét foderindkøb til dagen inden det syvende indkøb herefter (Lag 7) og
iv) perioden fra ét foderindkøb til dagen inden det tiende indkøb herefter (Lag 10).

Når der anvendes *lag,* kan beregningen kun udføres så længe der kan medtages det antal observationer, *laget* definerer (her 1, 3, 7 og 10). Det betyder, at de sidste to registrerede indkøb i Lag 3, de sidste seks indkøb i Lag 7 og de sidste 9 indkøb i Lag 10 fremstår med samme værdi. Nemlig den sidst mulige beregnede værdi.
Dette har større betydning for prognoser for sjældnere indkøbte fodermidler end for hyppigere indkøbte fodermidler, fordi hver *lag*-værdi repræsenterer en længere tidsperiode ved sjældnere indkøb.

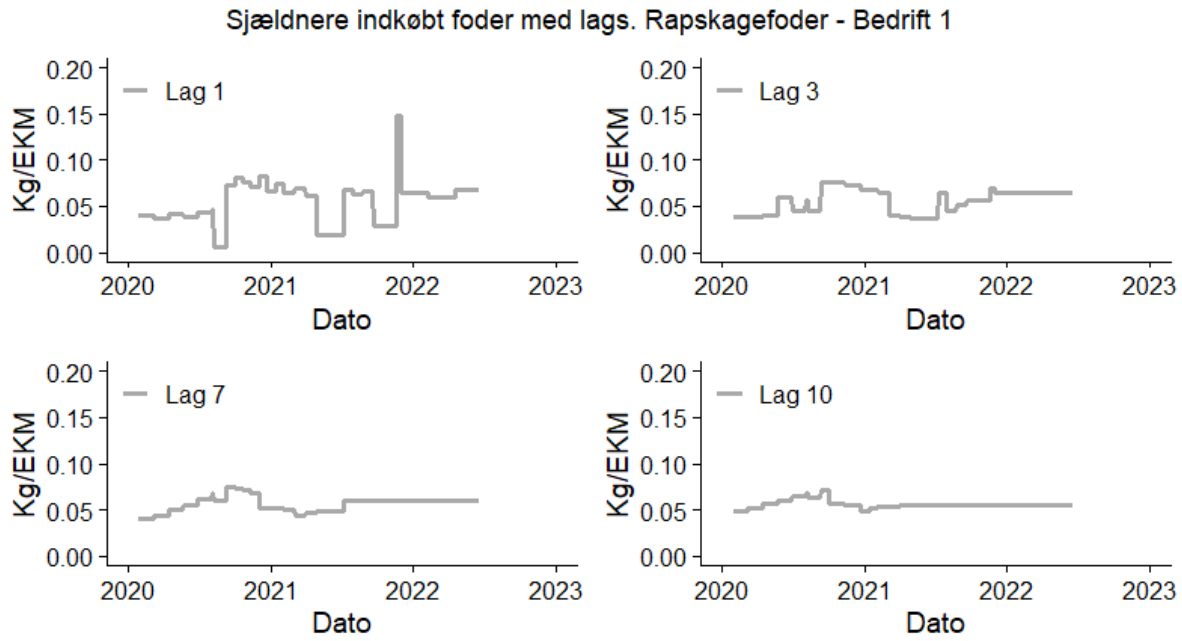
Figur 3 viser, at fodermængden pr. kg EKM er relativt stabil for A-blanding allerede ved Lag 3, men endnu mere stabil ved Lag 7. Der er ikke stor forskel på Lag 7 og Lag 10.



Figur 3 For fodermidlet 'A-blanding' er der på Bedrift 1 foretaget 68 indkøb over cirka 2½ år. Dette svarer til et køb cirka hver anden uge. Lag 1 er derfor foder pr. EKM for 14 dage, Lag 3 for 1½ måned, Lag 7 for 3½ måned og Lag 10 for 5 måneder.

A-blandingen er et af de primære kraftfodermidler i besætningen, og der sker indkøb cirka hver anden uge. Hyppige indkøb betyder, at de fire *laggede* perioder hver især repræsenterer relativt korte tidsrum sammenlignet med samme *lag* for et fodermiddel, der købes sjældnere. Derfor kan et længere *lag*, som Lag 7, med fordel anvendes i foderprognoser for dette fodermiddel.

Figur 4 viser den samlede mængde af fodermidlet 'Rapskagefoder' delt med den samlede mængde leveret EKM registreret for samme antal indkøb som A-blandingen beskrevet ovenfor. Da Rapskagefoder indkøbes sjældnere (cirka hver femte uge), repræsenterer de fire *laggede* perioder hver især et relativt længere tidsrum sammenlignet med A-blandingen. For Rapskagefoder ligger den sidst mulige beregnede værdi ret langt tilbage i tiden, og der mistes, med andre ord, for meget information ved Lag 7 og Lag 10. Derfor vil et kortere *lag*, som Lag 3, være at foretrække i foderprognoser for dette fodermiddel.



Figur 4 For fodermidlet 'Rapskagefoder' er der på Bedrift 1 foretaget 28 indkøb over cirka 2½ år. Dette svarer til et køb cirka hver femte uge. Lag 1 er derfor foder pr. EKM for fem uger, Lag 3 for næsten fire måneder, Lag 7 for cirka ni måneder og Lag 10 for cirka et år.

Udglatning af EKM ved forskudt leverance

Under normale omstændigheder afhentes den producerede mælk hver anden dag hos mælkeproducenten. Det kan dog hænde, at afhentningen forskydes, hvilket giver udsving i den leverede mælkemængde for de pågældende dage.

Sådanne udsving påvirker estimeringen af foderforbruget, når EKM indgår i beregningen, som beskrevet ovenfor. Herunder beskrives en metode, til udglatning af sådanne udsving. De udglattede EKM-kurver er ikke anvendt til estimering af foderforbrug i dette notat, men det kan gøres fremadrettet.

Metode

Beregninger foretages i udgangspunktet indenfor hver måned i et kalenderår, men dette kan ændres til den tidsperiode, der ønskes.

For hver måned laves en regression over leveret EKM som:

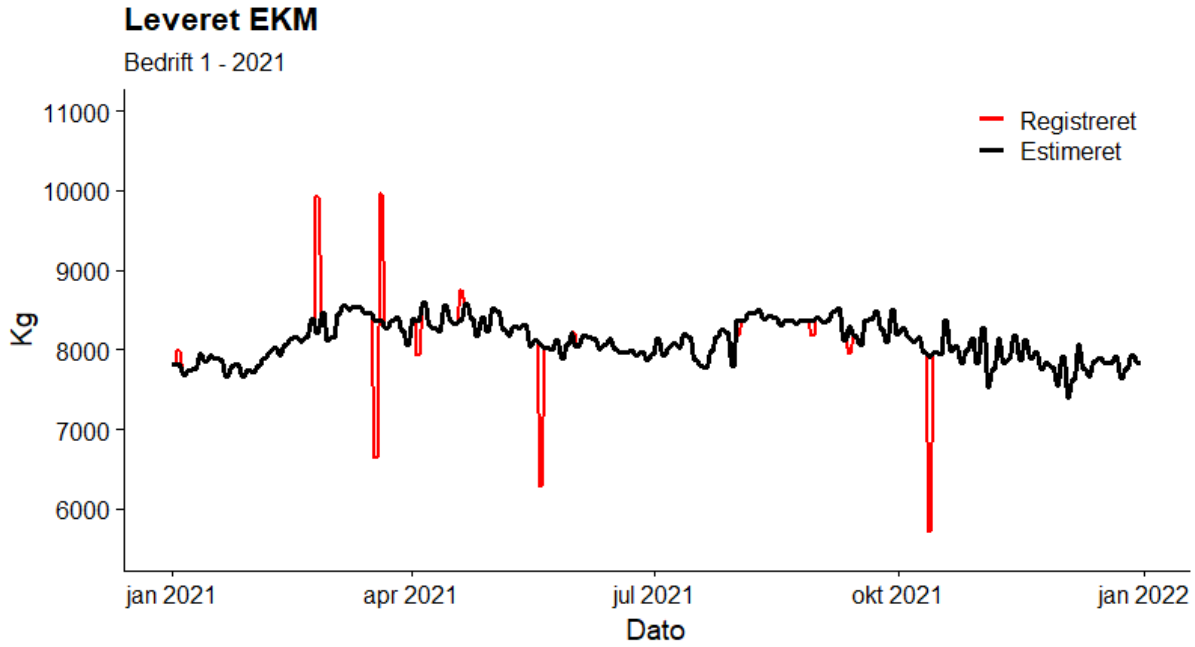
model = lm(leveret_EKM ~ 1, måned)

Konfidensniveauet angives som 0.9999999999, for kun at fjerne leverede mælkemængder langt over, eller under, bedriftens normale niveau for måneden.

For hver dag i måneden evalueres den leverede mængde EKM. Hvis mængden ligger over månedens øvre konfidensgrænse, eller under månedens nedre konfidensgrænse, erstattes dagens mængde med gennemsnittet af værdien for øvre og nedre konfidensgrænse:

udglattet EKM = (øvre konfidensgrænse + nedre konfidensgrænse)/2

Figur 5 viser, hvordan den registrerede mængde leveret EKM pr. dag i 2021 ("Registreret", rød kurve) har større afvigelser flere gange i løbet af året, og hvordan den estimerede mængde ("Estimeret", sort kurve) ligger oveni den leverede mængde, når denne ligger indenfor konfidensintervallet, og erstattes af valide værdier, når den leverede mængde ligger udenfor konfidensintervallet



Figur 5 Kg registreret leveret EKM (Rå data, rød kurve) og Kg estimeret leveret EKM (Estimeret, sort kurve) i 2021 for Bedrift 1. Sort kurve kan anvendes som vægtningsparameter til beregning af forventet foderforbrug.

Overvejelser om modeller for prognoser

Man kan lave modeller på forskellige niveauer – forstået på den måde, at man kan inddrage eller udelade forskellige mere eller mindre relevante størrelser, afhængigt af hvilke data (størrelser), som er tilgængelig for bedriften.

Vi har set i overvejelserne ovenfor, at det er naturligt, at en prognose for foderforbrug kan hægtes op på de størrelser, som kommer fra et *foderbudget*. For visse bedrifter kan der også være *foderkontroller*. Og de vil måske kunne tilføre endnu mere viden end blot foderbudgettet.

Vi har også set, at en prognose for foderforbruget kan hægtes op på kg EKM. Hvis vi skal bruge fx daglig kg EKM i en prognose, skal vi som udgangspunkt også have en prognose for kg EKM for at kunne beregne en prognose for foderforbruget. Men hvis en bedrift ikke har et mælkebudget, og dermed ikke en prognose for kg EKM, så kan vi som en simpel mulighed blot sætte den daglig mælkemængde, EKMi = 1, og ud fra dette regne på formlerne som ellers beskrevet.

Ovenfor har vi set en parallelitet mellem nogle kurver. Denne parallelitet kan måske udnyttes til en fremskrivning. Den er dog fremkommet, når vi bruger EKMi som den daglige vægtning. Det kan med fordel undersøges om ikke denne parallelitet også opstår, når man bruger antal dyr som daglig vægtning. På den måde vil metoden også kunne anvendes for dyregrupper, som ikke indbefatter køer.

Ud fra disse overvejelser kan vi liste en række forskellige forudsætninger op for diverse prognosemodeller:

A. Med brug af data fra både foderkontroller, foderbudget og mælkebudget
B. Med brug af data fra foderbudget og mælkebudget (ikke foderkontrol)
C. Med brug af data fra mælkebudget, men ikke foderbudget eller kontrol
D. Uden brug af de tre ovenstående størrelser, men dog, som for alle mulighederne: brug af data vedr. indkøb af foder.

Modeltype	Typer af data, som kan anvendes som grundlag for fremskrivning			
	Foderkontrol	Foderbudget	Mælkebudget	Indkøb af fodermidler
A	X	X	X	X
B	--	X	X	X
C	--	--	X	X
D	--	--	--	X

Modellerne for hver af disse fire muligheder kan dog vise sig at være meget forskellige, når det er varierende, hvilke data som skal indgå i dem.

I alle disse muligheder vil det kunne være en fordel af have en status på lagerbeholdningen på forskellige tidspunkter, jf. beskrivelsen af datagrundlaget om antagelser om foderforbrug mellem indkøb.

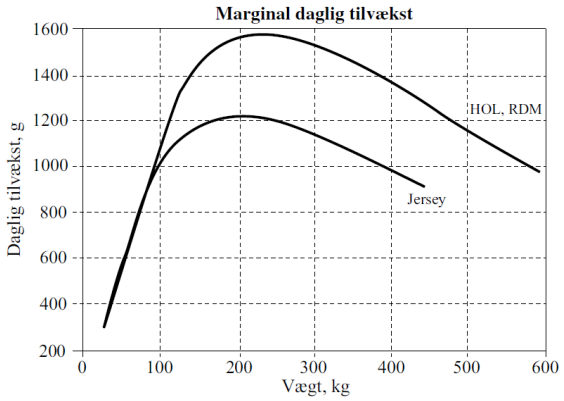
Dynamisk modellering og kalmanfilter
I processen med at finde et godt prognoseestimat har brugen af dynamiske modelleringsmetoder og kalmanfilter været diskuteret. I princippet kan sådanne metoder implementeres, men i praksis vil det kræve hyppigere observationer, gerne reelle tidsserier af foderindkøb, for at metoderne giver værdi. Et kalmanfilter vil kunne anvendes på EKM, og muligvis på et fodermiddel som 'A-blanding', der indkøbes hver anden uge. Det vil dog ikke nødvendigvis være den bedst egnede metode. Særligt hvis der stræbes efter at anvende samme metode til prognoseberegning for alle fodermidler, vil det være relevant fortsat at afklare relationen mellem indkøbt/forbrugt foder og leveret kg EKM eller en anden daglig vægtning.

Multivariat modellering
Udgangspunktet vil være at kunne modellere hvert fodermiddel for sig indenfor bedriften. Men multivariat modellering kan vise sig at være gavnligt ved, at vi kan gøre en prognose bedre, hvis vi modellerer alle de anvendte fodermidler på én gang. For på den måde vil vi måske kunne hente noget information fra ét fodermiddel og anvende i prognoser for de andre fodermidler. Men også her vil kompleksiteten stige, hvis vi skal håndtere forskellige relationer af fodermidler på forskellige bedrifter. Og måske ville en sådan multivariat modellering også kun kunne lade sig gøre, hvis vi har hyppigere observationer, som nævnt i afsnittet ovenfor om dynamisk modellering.

Regresionsanalyse på forsøgsdata til interpolation af tilvækst

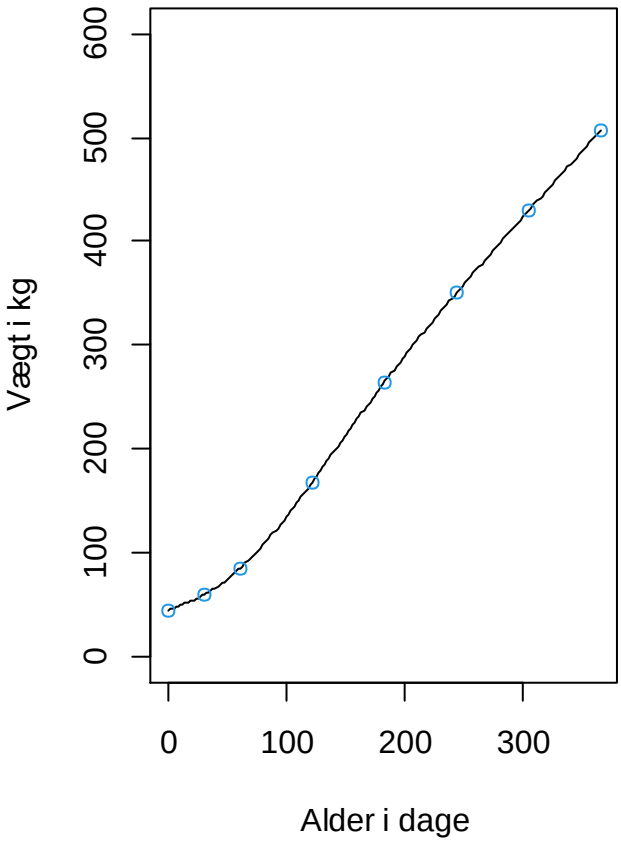
Landmænd vil ofte stå i situation, hvor der er behov for en marginalbetragtning på, om det har en økonomisk værdi at øge tilvæksten, holde et dyr længere i besætningen eller forbedre kvaliteten af kødet. Det kræver, at der er en kontinuerlig beregningsmodel, der kan arbejde med marginaler. I arbejdet med beslutningsprocesser i slagtekalveproduktionen er det fundet nødvendigt at kunne evaluere til hver dag, en aktuel vægt og foderforbrug, for at kunne give en økonomisk vurdering af om produktionen er optimeret til de aktuelle produktionsresultater og økonomiske forhold.

Der findes tilvækst kurver i blandt andet ”Håndbog i kvæghold 2013” (Landbrugsforlaget, Videncenteret for landbrug 2013), hvilken kan ses i figur 6. Den første udfordring er, at der i dag anvendes krydsningssæd, således at der opnås kalve, der har en anden tilvækst end i det, der ses i figur 6. Derudover er det er baseret på ældre data, så tilvæksten hos kalvene kan have ændret sig, eftersom at malkekøerne har udviklet sig avlsmæssigt.



Figur 6 Marginal daglig tilvækst for slagtekalve, Håndbog i kvæghold 2013 s.110

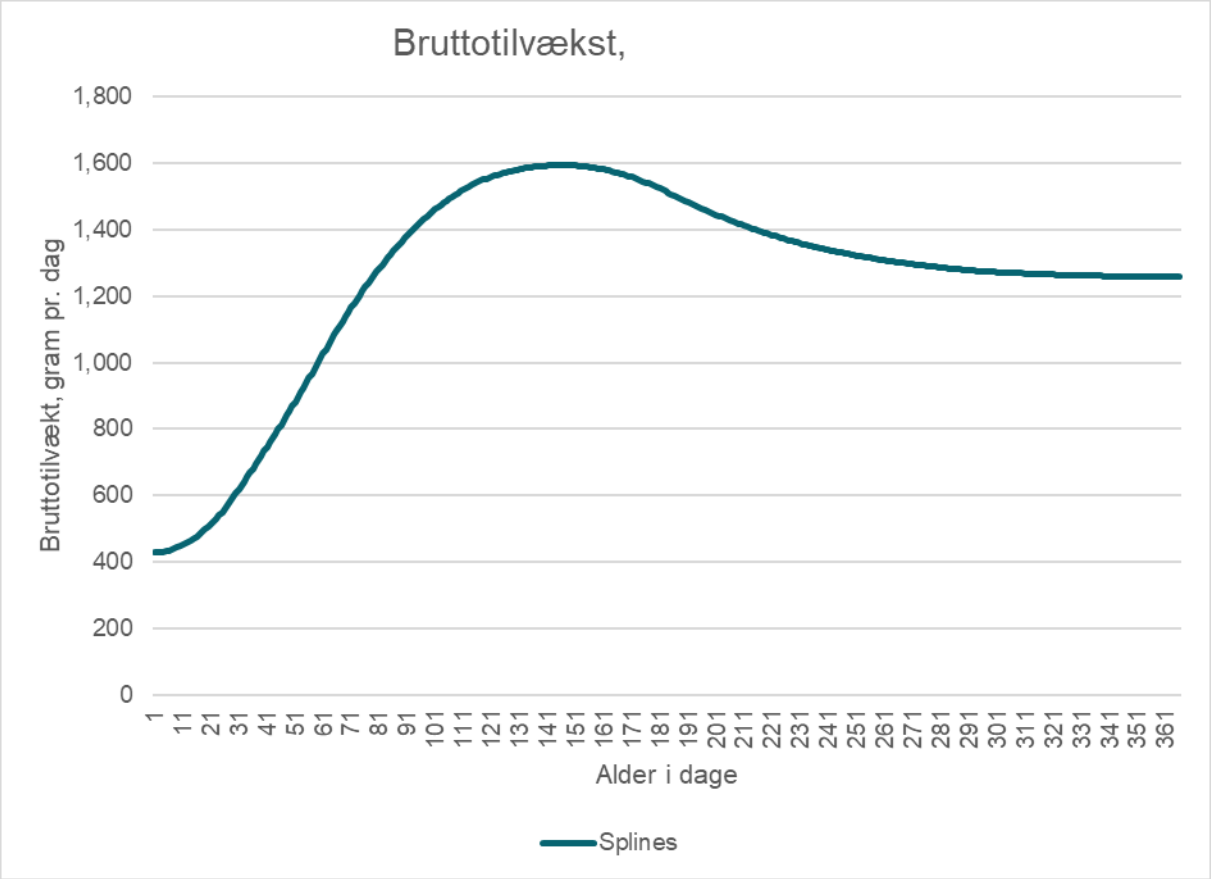
Det er undersøgt, om der er nyere og relevante tilvækstkurver for de dyregrupper, der anvendes i dag i Danmark, men det har ikke været tilfældet. I stedet har det været muligt at få adgang til forsøgsdata fra Aarhus Universitet, der giver punktestimater af vejninger fra deres seneste slagtekalveforsøg. De har testet de nye krydsningskalve, der anvendes i dag som er kødkvæg/holstein-krydsninger. Punktestimaterne fremgår for en Holstein tyr i figur 7, hvor de er markeret med en blå cirkel.



Figur 7 Plot af forsøgsdag med punk vejninger af Krydsningskalve Holstein/kødkvæg. De blå cirkler er data fra forsøget, og stregen imellem punkterne er en Splines funktion, der interpolerer datapunkterne

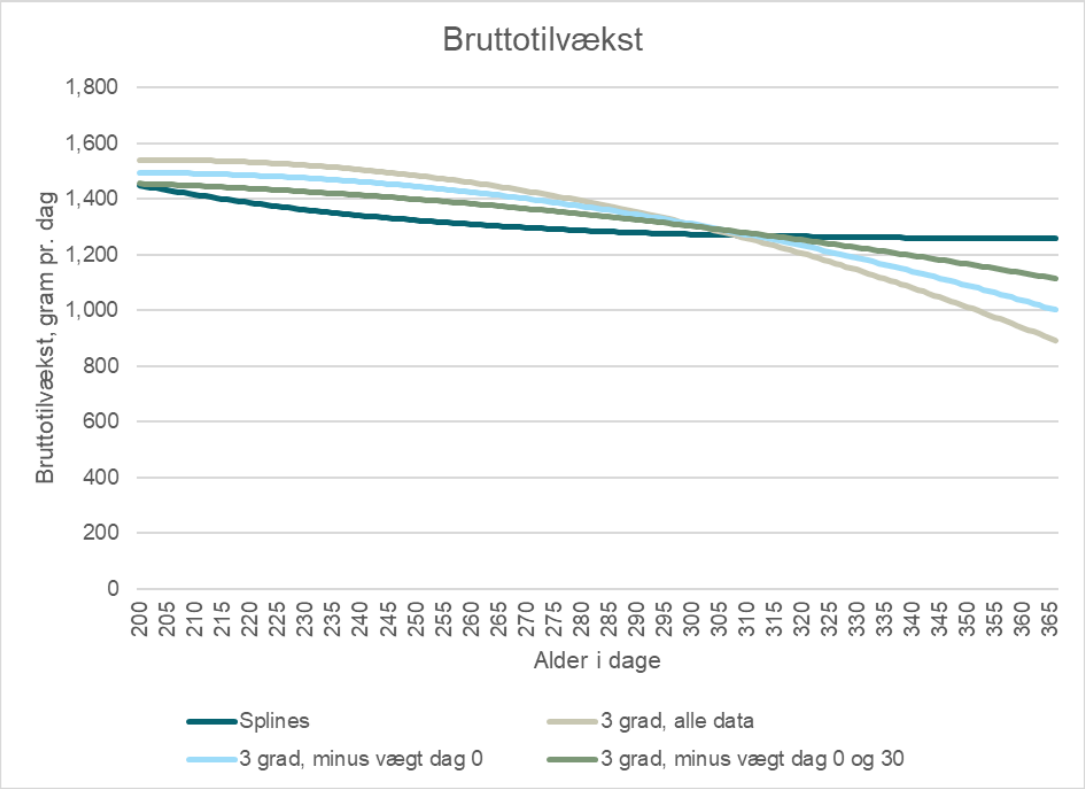
Der er i projektet arbejdet med at kunne skabe en tilvækstkurve, der er baseret på nyeste viden. Dette kræver, at de vægtpunker, der er vist i figur 7, interpoleres imellem punkterne. Dette kan gøres med regresionsmodeller, hvor der ofte anvendes et n'te gradspolynomium. Men dette har den udfordring, at den valgte funktionelle form af polynomiumet har stor betydning for resultatet. Derfor er der arbejdet med en mere fleksibel model, hvor der ikke er behov for at specificere en bestemt funktion på forhånd, men i stedet at lade dataet selv fitte en tilvækstkurve bedst muligt.

Det er valgt at arbejde med Splines, hvilket resulterer i den sorte streg mellem punkterne i figur 7. Funktionen tilpasser en glidende kurve, der giver en meget glat tilvækstkurve, som kan ses i figur 8. Grunden til at Splines har dette resultat er, at den udregner mange stykvise polynomier, og minimerer derfor store hak. Dette er vist med alt tydelighed i figur 8, som viser den første afledte af Spilnes funktionen i figur 7. Her ses det at hele justeringen foregår glidende.



Figur 8 Marginal tilvækstkurve for slagtetyrekalv af krydsningsrace Hol/Kød baseret på splines funktion på data i figur 7

Der er vigtigt at vælge den rigtige metode, da det kan påvirke de beslutninger landmanden skal tage, og det er ikke uden betydning, hvilken funktion der vælges, kan ses i figur 8. I forbindelse med en fagligdrøftelse af den rigtige metode, blev det foreslået af produktionseksperter at forsøge at anvende et 3 gradspolynomium og evt. tage enkelte datapunkter i starten ud, da de ligger forholdsvis tæt, og dermed ville have for stor vægt. Resultatet kan ses i figur 9, hvor der er vist forskellige bruttotilvækst i det intervaller, hvor det kan være relevant at slagte krydsningskalvene, og her ændrer det sig meget hvilken form for estimering, der vælges.



Figur 9 Effekt på bruttotilvækst for slagtetyrekalv af krydsningsrace Hol/Kød ved forskellige estimeringer af tilvækst funktion.

Det er vigtigt at vælge en metode til at estimere tilvækstkurver, der er robust, men som figur 9 viser, så har det stor betydning, hvordan den estimeres. Derfor er det vigtigt, at der inddrages eksperter til at vurdere, om de tilvækstkurver, der generes, er biologiske plausible, og de kan støtte op om resultatet, således de ikke underkender en fremtidig beslutningsmodel.

Grafisk visning og tilgængelighed beslutningsmodel

Det er ikke i sig selv nok at have en sikker beslutningsstøttemodel, hvis ikke den kan give en forståelig præsentation af beslutningsgrundlaget overfor landmanden. Det er i den sidste ende landmanden, der skal tage beslutningen, og her kan metoden hvormed beregningen præsenteres være afgørende. Der er flere forhold, som kan være afgørende for at beslutningsstøttemodellen når helt ud, og bliver anvendt i en aktuel beslutningssituation. Først er det vigtigt med tilgængeligheden af metoden – kan beslutningstageren/landmanden selv tilgå beslutningsstøttemodellen og modificere den med sine egne input? - eller skal der inddrages en anden part som for eksempel en rådgiver, og hvor hurtigt (og med hvilke omkostninger) kan beslutningen så drages. Dernæst kommer præcisionen af metoden, hvor den formodentlig har størst interesse og gennemslagskraft, hvor den er mest nøjagtig. Og metoden har muligvis knap så meget interesse i de situationer, hvor den blot er et første overblik, som der bør regnes videre på.
Til sidst er selve brugeroplevelsen af det digitale produkt vigtig, hvor beslutningsmodellen bliver præsenteret. Er der lavet en opstilling, som er let at overskue og forstå? Og hvor beslutningstageren forstår outputtet. Samtidig er det også vigtigt, at brugeren forstår, hvilket input, der er anvendt, og hvorledes input kan redigeres.

Der er arbejdet i projektet med at teste en form for prototyper af beslutningsmodeller overfor landmænd, som vi her kalder preto-typer. Selve ideen har været at teste en metode, hvor vi grafisk har en model, som indeholder landmandens egne data, der er tastet ind på forhånd, og at landmanden kan få mulighed for at interagere med det input, der findes data om på bedriften. Der er fokuseret på, at relevante problemstillinger kan behandles, således at der er et reelt beslutningsbehov. Der er fokus på en visning, der er selvforklarende og forståelig, og samtidig baseret på relevante/egne data. Til det formål er der valgt at arbejde med et program, der hedder Calconic. Hvilket er en webbaseret interaktive model, der har en regnefunktionalitet. Programmet er ikke udviklet til driftsøkonomi på landbrugsbedrifter, men kan forholdsvis let tilpasses, så det passer til landmandsrettede beslutningssituationer. Samtidig giver programmet mulighed for implementering på eksisterende hjemmesider.

Et eksempel er preto-typen vist i figur 10. Det er en fuld funktionsdygtig model, der kan give et udkast til en handelspris på grovfoder, som landmand efterfølgende hurtigt og nemt kan tilpasse til egne forudsætninger. Der er også afprøvet en preto-type til beregning af fremstillingspris på mælk, hvor man kan se, hvilke forskelle der opnås, når der ændres i forudsætningerne bag fremstillingsprisen. I det tilfælde kan det fx være, at der ændres i fodermidler og fodringsomkostninger, som følge af andre leverandører og valg af en anden fodring, og hvilken betydning det har på fremstillingsprisen af mælk Metoden og afprøvningen har fokuseret på, at brugeren let kan komme fra at udføre en ændring til at se resultatet af ændringen i sit program.

Figur 10 Eksempel på grafisk visning af beslutningsmodel for handelspris på majshelsæd