Robust definition af nøgletal

Baggrund
Det bliver mere og mere udbredt at leje aktiver i landbruget. Det kan f.eks. være leasing af maskiner, leje af driftsbygninger og forpagtning af jord. Det er forskelligt, hvordan denne leje og leasing påvirker sammenligneligheden af nøgletal på tværs af landbrugsvirksomheder. Det kommer vi tilbage til senere.

Først skal det slås fast, at problemstillingen alene vedrører nøgletal, hvor summen af aktiver (bogførte værdier af aktiver) indgår i beregningen af nøgletallet. Det skyldes, at når aktiverne lejes bliver aktivmassen alt andet lige mindre end hvis man ejer aktiver. Derimod indgår der omkostninger til leje mv. i resultatopgørelsen på samme vis, som hvis der havde været renter og afskrivninger på et ejet aktiv. Der kan dog være udfordringer med proportionaliteten mellem omkostninger og værdi af aktivet.

Det er kun afkastningsgraden, der indeholder summen af aktiver, hvis man ser på de nøgletal, der anvendes i dag. Det er derfor udelukkende afkastningsgraden, som analysen vedrører.

Det er vigtigt at pointere, at der alene er tale om en ændring af definition i analyser og modeller, hvor der anvendes data fra ejendomme på tværs af hele landbrugssektoren eller på tværs af de enkelte sektorer i landbruget – f.eks. kvæg, svin osv. Der er IKKE tale om at ændre definition i de værktøjer, der viser den enkelte landbrugsvirksomheds nøgletal. De beregnede nøgletals definitioner er korrekte for den enkelte bedrift, hvis SEGES' definitioner anvendes, da de tager udgangspunkt i den forretningsmodel, som landmanden har valgt for sin virksomhed. Det er alene et spørgsmål om sammenlignelighed på tværs af forretningsmodeller. Derfor er det alene i sektoranalyser og modeller, at der anbefales en anden definition af afkastningsgraden.

Når der er leaset aktiver, er det typisk det, som man i regnskabsmæssig sammenhæng kalder finansiel leasing. Ved den type leasing skal aktivet indregnes med en bogført værdi på balancen, og leasingydelsen skal fordeles mellem renter og afskrivninger – ligesom hvis aktivet havde været ejet. Dette er bestemt i årsregnskabsloven. Selv om landbrugsbedrifter (de enkeltmandsejede) ikke er underlagt Årsregnskabslovens bestemmelser, følger langt de fleste alligevel Årsregnskabsloven. Der er derfor i denne sammenhæng ikke udfordringer med leasede aktiver i forhold til definition af nøgletal.

De lejede aktiver, der er undersøgt i projektet, er derfor leje af driftsbygninger og forpagtning af jord. Driftsbygninger og især jord udgør ofte en betydelig del af aktivmassen, når den er ejet. Derfor påvirkes summen af aktiverne i væsentlig grad, hvis aktiverne lejes og dermed ikke er indregnet som aktiv på balancen. Det er i sig selv heller ikke en udfordring, hvis lejen (påvirkning af tælleren) fuldstændig modsvarer den lavere balancesum.

Her adskiller driftsbygninger og jord sig væsentlig fra hinanden. Jord er et såkaldt varigt gode, som generelt set aldrig mister sin produktionsværdi fuldstændigt over tid[1]. Jord bliver derfor heller ikke afskrevet ligesom alle andre aktiver. I og med at jorden er et varigt gode, er forpagtningsafgiften (lejen) for jorden relativt lav i forhold til jordens værdi sammenlignet med forholdet mellem leje for en driftsbygning og bygningens værdi. Det forhold forventes at skævvride afkastningsgraden mere for dem, der har en stor andel forpagtet jord end for dem, der har en stor andel leje af driftsbygninger.

[1] Der er selvfølgelig arealer, der mister sin produktionsværdi. Det skyldes ofte politiske bestemmelser, der gør et areal udyrkbart. Denne problemstilling diskuteres ikke yderligere i dette notat. Det vil være uden for notatets formål.

Det er i denne artikel beskrevet, hvilken betydning det har for afkastningsgraden, hvis værdi på forpagtning jord inkluderes i aktivmassen.

For at kunne kompensere for denne skævvridning, er det nødvendig at beregne en værdi af de lejede og forpagtede aktiver. Hvis beregningen skal være sikker nok, er det nødvendigt, at beregningerne er tilstrækkeligt præcise i forhold til den faktisk værdi på aktiverne. Ellers gør beregningen mere skade end gavn.

I projektet er det derfor undersøgt, om vi ud fra oplysninger i SEGES' Økonomidatabase kan beregne en værdi af lejede driftsbygninger og forpagtet jord med tilstrækkelig sikkerhed.

Analyse vedr. forpagtning af jord
En analytisk værdi vil for nogle bedrifter kunne udregnes direkte fra den i regnskabet opgjorte hektarværdi, under antagelse at den forpagtede jord tilnærmelsesvis har samme værdi som den ejede jord. Men for ejendomme, som har en stor andel forpagtet jord, kan den opgjorte hektarværdi være misvisende. Her kan det give bedre mening at udregne en analytisk hektarværdi ud fra en række variable der giver et mere korrekt estimat af, hvad hektarværdien for en sammenlignelig bedrift bør være. Derfor er det ønskværdigt at have en præcis model til udregning af analytiske jordværdier på forpagtet jord.

Dataudvælgelse

For at kunne opstille en model, der kan forudsige analytiske jordværdier for den forpagtede jord, skal forklarende variable udvælges. Til at forklare jordværdien udvælges følgende variable:

Bedrift oplysninger: Jordtype, kommunenummer, postnummer, driftsbygninger, beboelse landbrug.
Dyreenheder: Dyreenheder i alt.
Arealer og andele: Ejet landbrugsareal, andel sandjord, andel forpagtet, gennemsnitlig markstørrelse.
Udbytter (pr. ha): Korn, græs, majs.
Forpagtningsinformation: Forpagtningsindtægt, forpagtningsudgift.
Andet: Betalingsrettigheder EU.

Databehandling

For at gøre analysen bedre fjernes ekstreme ejendomme. Analysen består derfor af ejendomme der opfylder følgende kriterier:

- Bogført jordværdi er mellem $[80.000, 400.000]$.
- Andel af totalt landbrugsareal der er forpagtet, er under (eller lig med) 50%.
- Totalt landbrugsareal er over (eller lig med) 50 hektar.

Det forudsættes yderligere også, at ejendommen har et kommunenummer, således finanstilsynets værdi kendes for ejendommen.

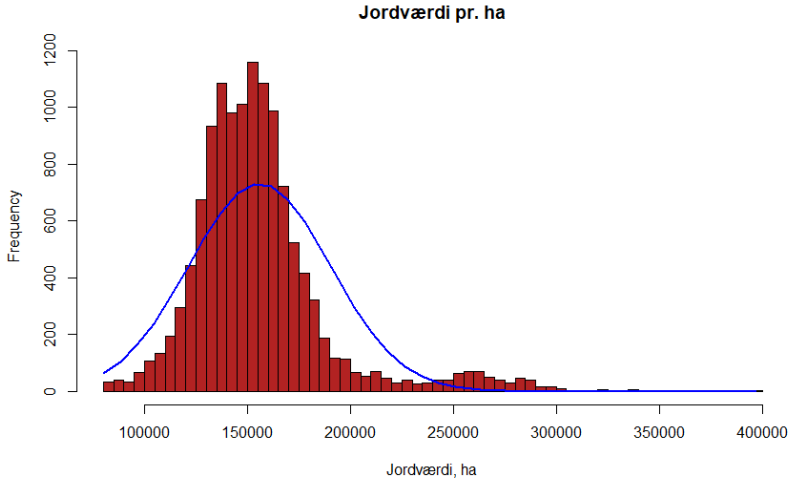
Årene 2016 – 2017 og 2019 bruges i analysen. 2016 - 2017 bruges til at estimere de valgte modeller og 2019 bruges til at vurdere, hvordan modellerne klarer sig.

Man vælger typisk separat data til at vurdere, hvor gode modellerne er relativt til hinanden. Det gøres, da man gerne vil have, at modellerne har en stærk generaliseringsevne til nyt data. Derfor skal data brugt til estimering af modellerne ikke bruges til at vurdere modellerne efterfølgende.

Eksplorativ dataundersøgelse

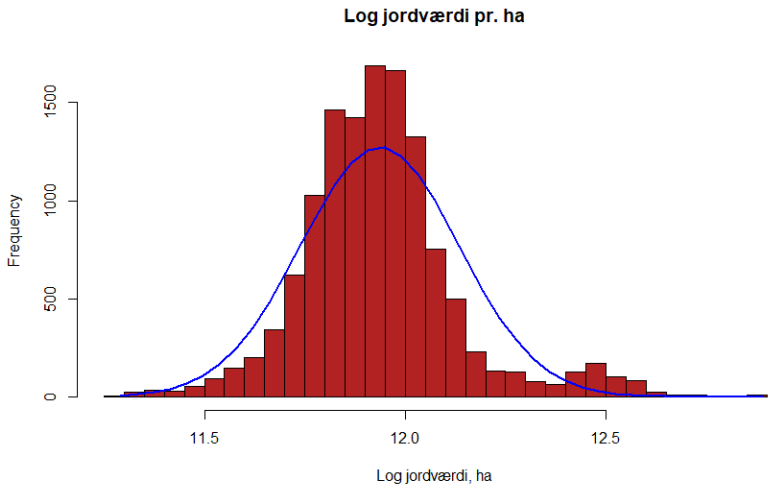
Den gennemsnitlige jordværdi pr. ha i datasættet er 156.528 kr. Den mest hyppige jordbundstype er sandjord (cirka 55%), efterfulgt af lerjord (cirka 42%). Den mest hyppige Gtype er 121, som er kvægbedrifter. En gennemsnitlig bedrift ejer 143,37 hektar land og forpagter i gennemsnit 73,05 hektar. En gennemsnitsbedrift har i gennemsnit 118,06 hektar sandjord og 92,71 hektar lerjord. Derudover har en gennemsnitsbedrift 244,26 dyreenheder.

For at danne et overblik over, hvordan jordpriserne pr. hektar ser ud, er nedenfor i Figur 1 lavet et histogram over jordpriserne pr. hektar.



Figur 1 - Jordværdi pr. ha.

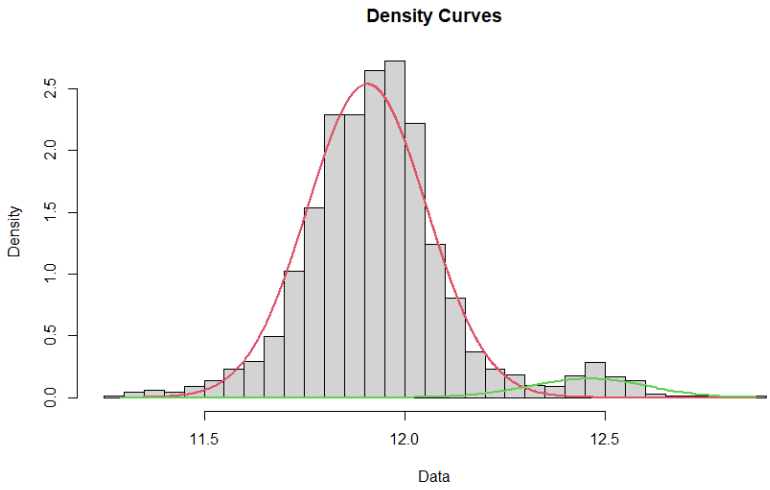
Her kan det ses, at data er forholdsvis normalfordelt, dog en smule skævt fordelt til den positive side. Derfor tages gerne log værdien af jordværdien pr. ha. Denne kan ses i Figur 2, hvor det fremgår, at log-data er meget normalfordelt.



Figur 2 - Log jordværdi pr. ha.

Fra både Figur 1 og Figur 2 kan det ses, at normalfordelingen passer forholdsvis præcist, men at der en lille del med høj ejendomsværdi, som ikke forklares af den tilpassede normalfordeling. Det ligner, at der er en stor del, som passer på normalfordelingen og har en log jordværdi på cirka 11,9 – 12,0, mens der er en lille normalfordeling, som har en højere log-jordværdi på 12,5. Her kan man bruge en EM-algoritme til at identificere de to normalfordelinger og sandsynligheden for, at en observation er i den ene eller den anden fordeling.

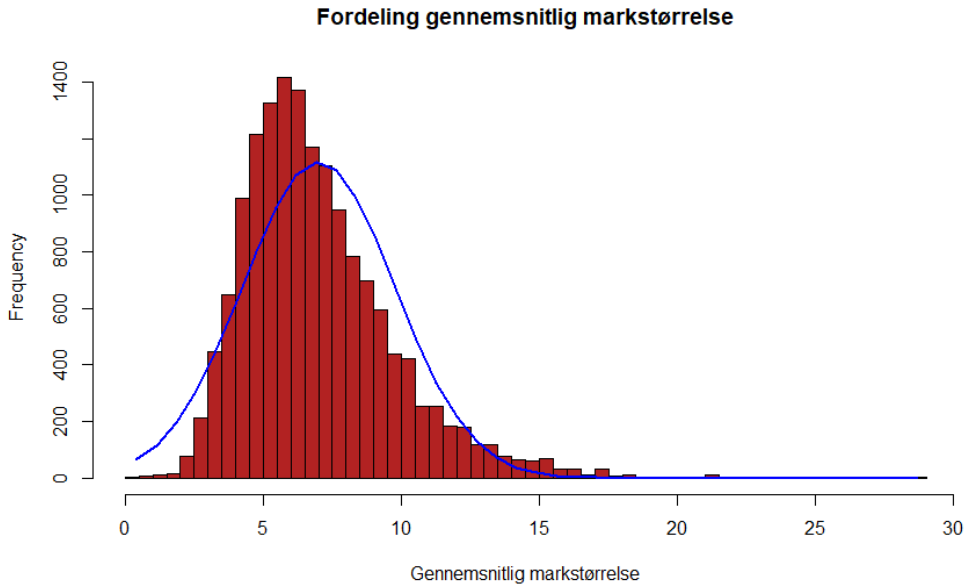
Med en tilpasset EM-algoritme vil følgende mønster vise sig:



Figur 3 - Log-jordværdi med EM fordelinger

Her kan det ses, at de tilpassede normalfordelinger passer bedre, når vi tillader, at der er to normalfordelinger til stede i data.

Fordelingen for den gennemsnitlige markstørrelse er givet ved:



Empirisk analyse

Performance bliver her målt i den gennemsnitlige absolutte afvigelse fra den bogførte jordværdi pr. ha. Dette mål bliver i dette sammenhæng udregnet som følger:

$$MAE_{test} = \frac{\sum_{i=1}^{N_{test}} \left| Y_{jordværdi\ pr.\ ha} - \exp\left(\hat{Y}_{log\ jordværdi\ pr.ha}\right) \right|}{N_{test}}$$

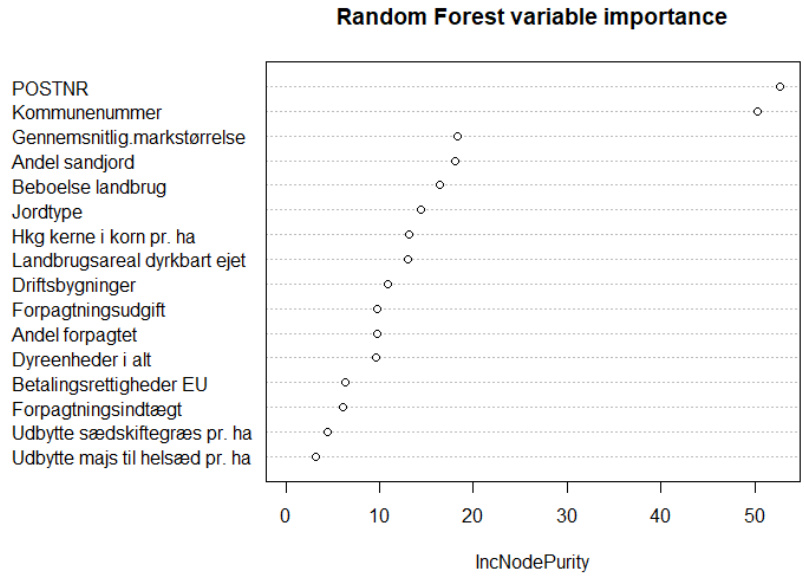
Hver afvigelse måles også i procent, hvor den gennemsnitlige absolutte procentvise afvigelse også udregnes.

hvor $|x|$ indikerer den absolutte værdi af x.

Model:	*MAE*	*MAE %*
Baseline model	18.074	11,3%
Regressionsmodel	14.064	9,2%
Random Forest	11.477	7,5%
Bagging	11.242	7,3%
Boosting	12.252	8,0
Neural network	22.167	14,0%
Random forest korrektion	11.461	7,6%
KNN (1 nabo)	13.007	8,5%

Læs mere om de enkelte modeller i bilag 1.

Man kan studere vigtigheden af de forklarende variable i Random Forest modellen. Det vil sige, at hvor meget nedsættes residual sum of squares (RSS) hver gang et beslutningstræ laver et split baseret på den forklarende variabel:



Figur 4 - Random Forest variabel betydning

Her kan det ses, at postnummer har stor betydning for modellen. Det samme har kommunenummeret, som er lidt mere bredt. Disse variable kan identificere højværdijord, som oftest er placeret i større områder.

Diskussion af resultater

Det, som vi med en gennemsnitlig afvigelse på 12.000 kr. pr. ha. unøjagtighed kan prædiktere, er jordværdien pr. ha. ejet jord, ikke logaritmisk transformeret. Modellen laver forudsigelser på logaritmiske værdier,

fordi den har det en smule nemmere med normalfordelte Y variable, men værdierne transformeres tilbage til reelle tal, inden fejlværdierne udregnes.

Derfor med hensyn til hvad modellen kan fortælle noget om, bliver det bliver en mellemting.
Y variablen i modellen er den bogførte værdi pr. ha, som er hektarværdien af det ejede + bortforpagtede.
Men X, som er de forklarende variable, er jo i regnskabet opgjort for det areal, der anvendt i årets drift, dvs. ejet + forpagtet – bortforpagtet.

Det vil sige, at det modellen kan fortælle noget om, er hektarværdien af det ejede på baggrund af tal fra det anvendte areal.
Under antagelse af, at det forpagtede jord ofte minder om den jord, som er ejet, vil analysen kunne overføres til bedrifter med høj andel af dyrkbar jord, der er forpagtet.

Ved bedrifter med høj andel forpagtet jord kan den bogførte værdi af den ejede jord være betydeligt forskellig fra den som driften foregår på, da bestanddelen af driftens jord er forpagtet. Derfor kan det her give mening at udregne en analytisk jordværdi af den forpagtede jord, som kan bruges til udregning af analytiske nøgletal. Det vil her være mere korrekt end at udregne en analytisk jordværdi for driften ud fra den i regnskabet opgjorte hektar værdi af den ejede jord.

Hvorimod for ejendomme, som har meget få hektar forpagtet jord, kan antagelsen om en sammenlignelig værdi med den i regnskabet opgjorte hektarværdi give bedre mening, og her vil en mere korrekt analytisk værdi sandsynligvis beregnes fra den opgjorte hektarværdi på ejendommen selv.

Konklusion og fremtidsudsigter

Analysen udført i dette projekt viser, at der er muligheder for at forbedre prædiktionsevnen over benchmark modellen, som bruger finanstilsynets værdi. Machine learning modellerne kan yderligere tunes, hvilket kan gøre dem stærkere til den givne opgave.

Det findes i den empiriske analyse, at random forest og bagging modellerne er de mest præcise. Her findes for den bedste model en gennemsnitlig absolut afvigelse på 11.242 kr. pr. ha, som svarer til en gennemsnitlig afvigelse på 7,3% fra den opgjorte hektarværdi. Denne vil dermed kunne overføres til at kunne udregne den analytiske hektarværdi for ejendomme med stor andel forpagtet jord, hvor den opgjorte hektarværdi kan være misvisende.

Der kan være flere grunde til, at en given model ikke laver mere præcise prædiktioner, hvor den væsentligste årsag ofte datakvaliteten. I forhold til grundværdier kan det være svært at få præcise resultater hvis der ikke er noget mønster i grundværdierne og de andre variable i regnskabet. Jordværdier kan både være bogført for højt og bogført for lavt, og der er ofte forskellige incitamenter til, hvorfor man skal eller ikke skal nedskrive/opskrive den bogførte værdi af sin jord. Dette vil gøre, at mønsteret mellem den bogførte værdi og de andre regnskabsmæssige variable ikke er klart. Derfor vil der ikke være noget forklarligt i regnskabet, der indikerer, hvis en ejendom har en jordværdi, der er betydeligt højere end hvad finanstilsynet har vurderet den bør være. For disse særejendomme vil en model derfor få dårlige resultater, og disse kan sågar forstyrre ydeevnen for de andre bedrifter.

Modellen kan potentielt styrkes ved at inkludere grundigere variable til forklaring af jordværdien, som ikke findes i Økonomidatabasen. Den gennemsnitlige markstørrelse for en bedrift kan øge forklaringsgraden, da større marker ofte er forbundet med en større værdi pr. ha. Yderligere er marker med få hjørner også

forbundet med større værdi, da mest muligt jord kan dyrkes i tilfælde med få hjørner. Når markerne bliver store (50-100 ha pr. mark), gør det ikke så meget, om der er lidt ekstra hjørner, da det bliver få hjørner pr. ha. Viden om markform og markstørrelse vil gøre modellen mere robust.

Der ses altså et potentiale i at kunne udregne en analytisk jordværdi pr. ha. for ejendomme, hvor den opgjorte værdi i regnskabet er forskellig for, hvad der må antages at være rimeligt. Dette kan være tilfældet hos bedrifter, som har forpagtet størstedelen af deres dyrket jord. Yderligere kan en model også bruges til at vurdere en bedrift med lignende bedrifter, da den foretager sine prædiktioner ud fra andre lignende ejendomme. I denne analyse er random forest og bagging modellen blevet brugt til at udregne de analytiske jordværdier pr. ha. da netop denne model viste sig stærkest i den empiriske analyse.

Analyse vedr. leje af driftsbygninger
Modellerne i denne analyse bruger nedenstående forklarende variable til at forudsige log-værdien af driftsbygninger pr. enhed. Herfra bliver log-værdierne transformeret tilbage til almindelige værdier, som bruges til at udregne afvigelserne fra de ønskede værdier. I denne analyse er der estimeret en separat model for smågrise-, slagtesvin- og kvægbedrifter.

Analyserne for de tre driftsgrene arbejder med følgende variable, som modellen forsøger at forklare:
Smågrise: Driftsbygninger pr. årsso.
Slagtesvin: Driftsbygninger pr. slagtesvin.
Kvæg: Driftsbygninger pr. årsko.

Databeskrivelse

I dette projekt arbejdes med data udtrukket fra Økonomidatabasen. For at gøre evaluering af modellerne mere korrekt anvendes data fra 2016 – 2017 til estimering og data fra 2019 bruges til evaluering af modellernes generaliseringsevne. En model skal være baseret på historisk data og have en stærk generaliseringsevne til nyt data for at kunne anvendes.

Følgende variable er udvalgt som forklarende variable til den empiriske analyse

Beskrivende variable: Gtype, landbrugsareal (ha)
Maskinstation og vedligehold: Vedligeholdsomkostninger
Besætningstype og størrelse: Antal dyreenheder i alt

Et effektivitetsmål er også inkluderet for de specifikke driftsgrene

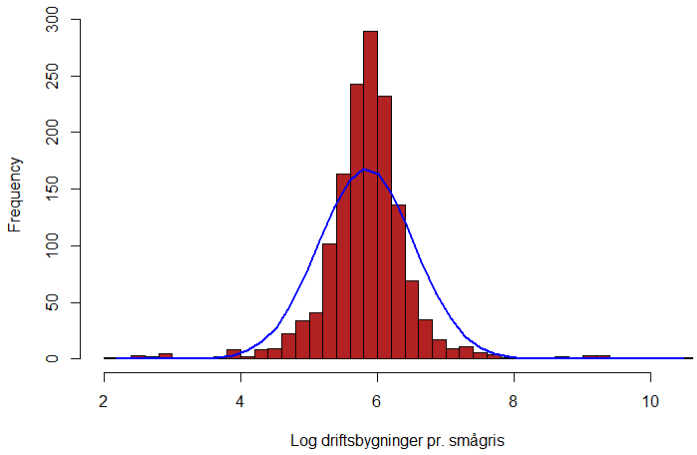
Smågrise: Producerede smågrise pr. årsso
Slagtesvin: Foderomkostning pr. kg slagtesvin
Kvæg: Kg EKM pr. årsko

En bedrift er kun taget med i analysen, hvis den er konventionel, heltid, har et gyldigt kommunenummer, den bogførte værdi af driftsbygningerne er over 200.000 kroner og landbrugsareal er over 50 hektar. Det er derudover antaget, at antal dyreenheder er større end nul.

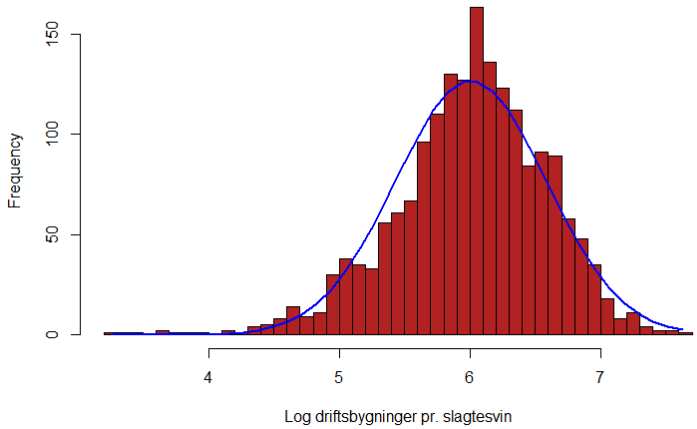
Analysen er adskilt på de tre driftsgrene kvæg, smågrise og slagtesvin.

Dataundersøgelse

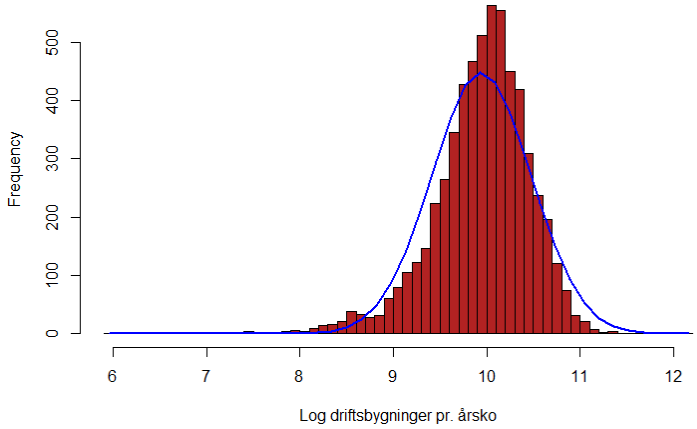
Fordelingen for log værdierne af driftsbygninger pr. smågris kan ses i figur 2, log-driftsbygninger pr. slagtesvin i figur 3 og log-driftsbygninger pr. årsko i figur 4.



Figur 5 - Histogram over log-driftsbygninger for smågrise



Figur 3 - Histogram over log-driftsbygninger for slagtesvin



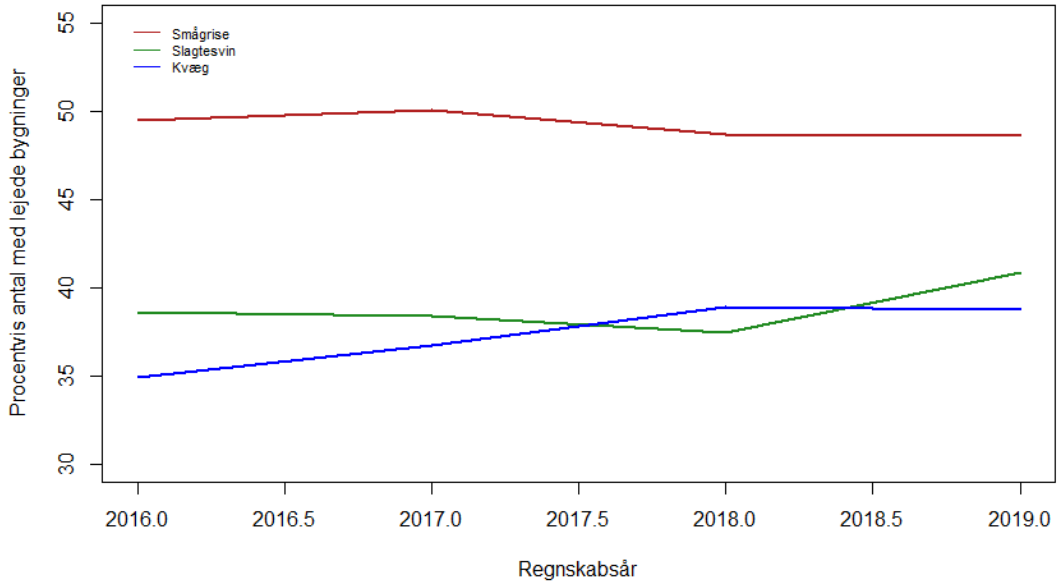
Figur 4 - Histogram over log-driftsbygninger for kvæg

I figurerne kan det ses, at de log-fordelte værdier er forholdsvis normalfordelte. Derfor vil log værdierne passe bedst ind i modeller, der bygger på normalfordelte data, såsom lineære regressionsmodeller estimeret med OLS estimationsmetoden.

For årene 2016-2019 ses følgende:

Driftsgren:	År	Har lejede bygninger	Har ikke lejede bygninger:
Smågrise	2016	300 (49,5%)	306 (50,5%)
	2017	311 (50,08%)	310 (49,92%)
	2018	294 (48,68%)	310 (51,32%)
	2019	266 (48,63%)	281 (51,37%)
Slagtesvin	2016	233 (38,64%)	370 (61,36%)
	2017	219 (38,42%)	351 (61,58%)
	2018	213 (37,43%)	356 (62,57%)
	2019	188 (40,87%)	272 (59,13%)
Kvæg	2016	639 (34,9%)	1192 (65,1%)
	2017	637 (36,71%)	1092 (63,29%)
	2018	644 (38,94%)	1010 (61,06%)
	2019	579 (38,73%)	916 (61,27%)

Den procentvise andel der har lejede bygninger (bygningsleje > 0) kan plottes som:



Empirisk analyse

Performance bliver her målt i den gennemsnitlige absolutte afvigelse fra den bogførte jordværdi. Dette mål bliver i dette sammenhæng udregnet som følger:

$$MAE_{test} = \frac{\sum_{i=1}^{N_{test}} \left| Y_{driftsbygninger} - \exp\left(\hat{Y}_{\log driftsbygninger}\right) \right|}{N_{test}}$$

hvor $|x|$ indikerer den absolutte værdi af x.

Resultater for smågrisebedrifter

Den gennemsnitlige driftsbygningsværdi pr. smågris er 11.112,44 kr.

Model:	MAE	MAE %
Regression	3661,86	65,36%
Random Forest	3620,58	62,1%
Bagging	3670,01	62,06
Boosting	4266,53	68,45%
Neural network	3806,71	59,39%
SVR	3780,84	62,17%
Multivariate splines	3593,06	64,33%

Læs mere om de enkelte modeller i bilag 1.

Resultater for slagtesvinebedrifter

Den gennemsnitlige driftsbygningsværdi pr. slagtesvin er 473,79 kr.

Model:	MAE	MAE %
Regression	191,22	60,77%
Random Forest	186,17	58,48%
Bagging	189,91	60,18%
Boosting	209,17	65,33%
Neural network	200,25	66,31%
SVR	195,68	61,32%
Multivariate splines	194,03	61,56%

Resultater for kvægbedrifter

Den gennemsnitlige driftsbygningsværdi pr. årsko er 23.507,3 kr.

Model:	MAE	MAE %
Regression	7879,86	57,73%
Random Forest	7604	54,81%
Bagging	7827,72	54,61%
Boosting	8614,54	59,06%
Neural network	7906,56	59,18%
SVR	7904,14	55,28%
Multivariate splines	7872,09	57,95%

Konklusion og fremtidsudsigter

Det kan ses, at modellerne til udregning af analytiske bygningsværdier er upræcise (ses ved de meget høje MAE %-værdier), hvilket skyldes datakvaliteten. I Økonomidatabasen er alderen af driftsbygningerne ikke opgivet. Derfor kan to bedrifter, som i sin drift ligner hinanden, have vidt forskelligt opgjorte driftsbygningsværdier, fordi alderen på driftsbygningerne varierer mellem dem. Værdierne på disse driftsbygninger vil afhænge af alderen, da den afskrevne mængde vil afvige fra hinanden. Derfor vil information om alder på ejede og lejede driftsbygninger være vigtig for denne analyse.

Yderligere er det nødvendigt med et mere præcist effektivitetsmål for slagtesvin. Det er ønskværdigt at kunne måle den daglige tilvækst i gram for slagtesvin, som er et bedre mål for effektiviteten. Dette nøgletal er på nuværende tidspunkt frivilligt, om det indtastes, og derfor er datamængden på nuværende tidspunkt ikke fyldestgørende for dette effektivitetsmål.

Med mere fyldestgørende data vil modellerne sandsynligvis være i stand til at prædiktere bygningernes værdig betydeligt mere præcist.

Konklusion og perspektivering

Ud fra de analyser, der er beskrevet ovenfor, vurderes det, at man med udgangspunkt i de oplysninger, der pt. er til rådighed i Økonomidatabasen, kan beregne en analytisk afkastningsgrad, hvor man anvender en beregnet jordværdi for den forpagtede jord. Det vurderes derimod ikke, at modellen for beregning af bygningsværdi har en tilstrækkelig sikkerhed. Denne del indregnes derfor ikke.

Hvis der på et tidspunkt er yderligere oplysninger om f.eks. bygningernes alder, kan man gentage analysen med henblik på igen at vurdere, om det tilføjer tilstrækkelig sikkerhed i beregningen af lejede bygningers værdi til, at det kan indgå i en beregning af analytisk afkastningsgrad. Beregning af jordværdien på forpagtet areal kan ligeledes gøres mere sikker med oplysninger om f.eks. markernes størrelse og form.

Der er i nedenstående beregninger alene indregnet værdi af forpagtet jord i beregningen af analytisk afkastningsgrad. Beregningen er vores forslag til den definition af den afkastningsgrad, der bør anvendes i analyser og modeller, hvor afkastningsgrad analyseres på tværs af landbrugsvirksomheder.

Beregning af analytisk afkastningsgrad

Bagging modellen, som er den bedste model i den empiriske analyse, kan nu bruges til at udregne en analytisk jordværdi pr. ha. for alle ejendomme. Her studeres kun konventionelle ejendomme, ejendomme med en jordværdi pr. ha. bogført mellem [80.000; 400.000] og et landbrugsareal på mindst 50 ha.

Derved kan en analytisk afkastningsgrad nu udregnes ud fra modellens prædiktioner omkring den analytiske jordværdi pr. ha. Formlen for afkastningsgrad er originalt givet som:

$$Afkastningsgrad$$

$$= \frac{Resultat\ før\ finansiering - ejerløn + forpagtningsindtægt - forpagningsudgift}{Aktiver\ i\ alt - private\ aktiver - finansielle\ aktiver} * 100$$

Den analytiske afkastningsgrad udregnes derimod som

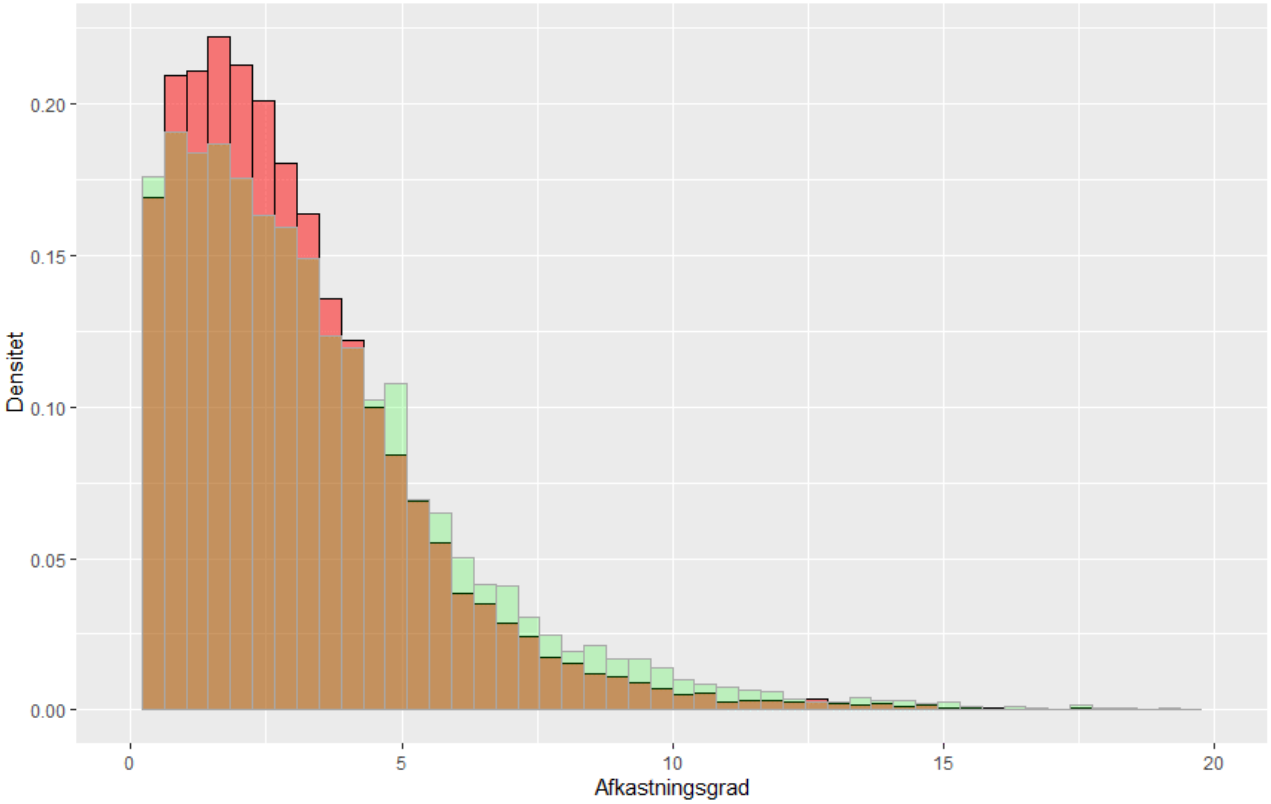
$$Analytisk\ afkastningsgrad$$

$$= \frac{Resultat\ før\ finansiering - ejerløn + forpagtningsindtægt}{Aktiver\ i\ alt - private\ aktiver - finansielle\ aktiver + Analytisk\ jordværdi * Antal\ ha\ forpagtet} * 100$$

Beskrivende statistikker for både opgjorte og analytiske afkastningsgrader kan ses herunder:

Opgjort afkastningsgrad:	
Minimum:	-45,7
25% percentil	0
Median	1,8
Middelværdi	2,036
75% percentil	3,9
Maksimum	96,3
Analytisk afkastningsgrad:	
Minimum	-26,72
25% percentil	0,59
Median	2,01
Middelværdi	2,24
75% percentil	3,67
Maksimum	55,35

De fordeler sig således:



Figur 6 - Fordeling for afkastningsgrad (rød = analytisk, grøn = opgjort)

Fra Figur 5 er den røde serie de analytisk udregnede afkastningsgrader, og den grønne serie er de opgjorte afkastningsgrader.

Antallet af bedrifter, der har over 50% andel af deres dyrkbare jord, som er forpagtet, har været cirka konstant over de seneste år og er omkring 20% af det samlede antal bedrifter.

Bilag 1 Modeller

Regressionsmodel

En multiple regressionsmodel er her brugt for at modellere forholdet mellem de forklarende variable og den afhængige Y variabel. Regressionsmodeller er meget simple og har høj fortolkningsevne. Fortolkningen af en stigning i en forklarende variabel giver en stigning på regressionskoefficienten for den afhængige variabel.

Random Forest model

Random forest modeller er en del af den klasse af modeller der bruger beslutningstræer til at lave prædiktioner. Beslutningstræer kan ses som en regressionsmodel der automatisk laver udvælgelsen af hvilke interaktionsvariable der skal inkluderes. Det gør at prædiktionsudfaldsrummet bliver ikke-lineært. Yderligere er random forest en ensemblemodel der bruger en forsamling af beslutningstræer til at lave en enkelt prædiktionen. Resultatet af dette er at der kommer til at være en lavere varians i prædiktionerne og derved bliver de ofte mere præcise.

I en random forest model tillader man kun en brøkdel af alle de forklarende variable at være split variabel hver gang der i et beslutningstræ foretages et split. Formålet med kun at tillade en brøkdel af variablene er at skabe forskellighed mellem beslutningstræerne, hvilket skaber en større variansreduktion.

Bagging

Bagging er den samme model som Random forest men hvor man altid tillader alle variable som splitvariable, i stedet for kun at tillade en del af variablene som splitvariable ved hvert split.

Boosting model

Ligesom random forest modellen, så bruger boosting modellen også beslutningstræer og det er også en ensemblemodel. Men den adskiller sig ret meget i fra random forest som bruger uafhængige beslutningstræer. I en boosting model er beslutningstræerne afhængige og de estimeres sekventielt. Hvert beslutningstræ er afhængig af de forudgående beslutningstræer da det estimeres ud fra residualerne af det forrige beslutningstræ. Det gør at det næste beslutningstræ vil sætte ekstra fokus på observationer hvor modellen klarer sig dårligt, da residualerne her vil være store.

Dybt neuralt netværk med embeddings

Et neural netværk er en af de mest kraftige modeller der findes. Et neuralt netværk er meget fleksibelt og tvinger ikke en datastruktur på data. Formålet i et neural netværk at lave ikke-lineære transformationer af data indtil man i det transformerede variabelrum kan skabe lineær separation af den afhængige variabel. Yderligere har neurale netværk også adskillelige tilføjelser der kan bruges til at få endnu stærkere ydeevne. En af disse er de såkaldte embeddings, som gør det samme som multinomial PCA for at skabe en kontinuerlig repræsentation af de diskrete variable. Forskellen fra normal multinomial PCA er her at de i et neural netværk bliver specialtilpasset til det givne problem.
Da neurale netværk har så stor fleksibilitet, så kræver det også at man kan udvinde mere information fra data, og derfor kræver disse modeller ofte mere data for at opnå en stærk performance. Da der arbejdes med små datamængder i dette projekt vil man derfor forvente at disse modeller har begrænset performance.

Support vektor regression (SVR)

SVR modellen er en ikke lineær regressionsmodel der ligesom det neurale netværk forsøger at skabe lineær adskillelse i et ikke-lineært transformeret inputrum. Her gøres inputrummet ikke-lineært ved hjælp af en

kernel funktion. Modsat almindelige regressioner er det i en SVR-model ikke alle observationer der bruges til at estimere parametrene. Kun de observationer man med en absolut fejlværdi større end ε bruges til at estimere modellens parametre. Det gør at modellen derfor er opbygget omkring de observationer som er svære at forklare.