Oprensning af udbyttedata fra markdatabasen

Baggrund
I forbindelse med udvikling af udbytteprognosen i vinterhvede er det vigtigt at anvende troværdige udbytteregistreringer til modeludviklingen. Dyrkningspraksis (sorter m.m.) og udbytteniveau ændrer sig løbende gennem årene. Hvis prognosen skal blive ved med at forudsige udbyttet så præcist som muligt, er der behov for at modellen løbende gentrænes på nye, relevante data. Det giver mening at automatisere denne proces, så der ikke er behov for at søge nye midler til at optimere udbytteprognosen om f.eks. 5-10 år.

Derfor undersøges det, hvordan et feedbacksystem på udbytteregistreringer i vinterhvede kan se ud, for at udbyttedata fra markdatabasen fremover kan inkluderes i modellen automatisk. Formålet med øvelsen er at sikre, at udbytteprognosen fremover forudsiger udbyttet i vinterhvede med størst mulig nøjagtighed, så danske landmænd kan anvende en opdateret udbytteprognose til at regulere kvælstoftildelingen til deres vinterhvedemarker.

Tidligere er udbytteprognosen i vinterhvede udviklet på baggrund af positionsbestemte udbyttedata fra mejetærskere, men analyser fra 2022 viser, at modellerne præsterer bedst, hvis modellen trænes på markniveau i stedet for på pixelniveau (10 x 10 meter). Derudover er antallet af valide udbytteregistreringer på markniveau i Dansk Markdatabase stigende, hvorfor datagrundlaget for modellen skifter til udbytter registreret på markniveau i markdatabasen.

POC på automatisk oprensning af udbyttedata fra markdatabasen
For at finde frem til troværdige udbytteregistreringer i markdatabasen skal data igennem følgende oprensning.

1. Fjernelse af dubletter

Først fjernes dubletter, så der kun er ét udbytte for en given mark i et givet år.

1.1 Fjernelse af dubletter, hvor alle kolonner er ens

1.2 Fjernelse af dubletter, hvor alle vigtige kolonner er ens (ID-kolonner, udbytte, år, areal)

1.3 Fjernelse af dubletter, hvor ID-kolonner og areal er ens, men hvor udbytte varierer

Her anvendes følgende algoritme for dubletterne for en given mark i et givet år:

1.3.1 Undersøg om et eller flere udbytter er et decimaltal.
- Hvis kun et enkelt udbytte er et decimaltal, bruges dette, og algoritmen stopper.
- Hvis der er flere udbytter, som er et decimaltal, sorteres heltal fra, og algoritmen fortsætter.

1.3.2 Undersøg om et eller flere udbytter er forskellig fra normudbyttet
- Hvis kun et enkelt udbytte er forskellig fra normudbyttet, bruges dette, og algoritmen stopper.
- Hvis der er flere udbytter, som er forskellig fra normudbyttet, sorteres de udbytter, som er lig med normudbyttet fra, og algoritmen fortsætter.

1.3.3 Undersøg om et eller flere udbytter ligger inden for grænserne på 20 hkg/ha og 150 hkg/ha

- Hvis kun et enkelt udbytte ligger inden for grænserne, bruges dette, og algoritmen stopper.
- Hvis der er flere udbytter, som ligger inden for grænserne, sorteres udbytter uden for grænserne fra, og algoritmen fortsætter.
- Hvis der er ingen udbytter, som ligger inden for grænserne, forsøges der at gange og dividere udbytterne med arealet for at se, om de derefter ligger inden for grænserne.

1.3.4 Undersøg om et eller flere udbytter ender på et andet tal end 0 eller 5.
- Hvis kun et enkelt udbytte ikke ender på 0 eller 5, bruges dette, og algoritmen stopper.
- Hvis der er flere udbytter som ikke ender på 0 eller 5, sorteres de udbytter, som ender på 0 eller 5 fra, og algoritmen fortsætter.

1.3.5 Undersøg om et eller flere udbytter har 'registered'=1
- Hvis kun et enkelt udbytte har 'registered'=1, bruges dette.
- Hvis der er flere udbytter, som har 'registered'=1, sorteres de udbytter, som har 'registered'=0 fra, og algoritmen fortsætter.

1.3.6 Hvis der stadig ikke er fundet et enkelt udbytte, bruges det udbytte med den seneste registreringsdato.

For langt de fleste af dubletterne stoppes algoritmen ved de første par skridt. Hvis der ikke er fundet et enkelt udbytte ved skridt 1.3.4, markeres udbyttet som ' utroværdigt' og bliver sorteret fra senere i skridt 2.10.

1.4 Fjernelse af dubletter, hvor ID-kolonner er ens, men hvor areal varierer

Her anvendes følgende algoritme for dubletterne for en given mark i et givet år:

1.4.1 Undersøg om et eller flere udbytter er et decimaltal.

1.4.1.1 Hvis et eller flere udbytter er et decimaltal, undersøges der, om en eller flere af registreringerne har et areal, som er maks. 3 ha mindre end markens samlede areal.
- Hvis der er en eller flere af registreringerne, som har et areal, som er maks. 3. ha mindre end markens samlede areal, bruges udbyttet fra den registrering hvis areal er tættest på markens samlede areal, og algoritmen stopper.
- Hvis ingen af registreringerne har et areal, som er maks. 3 ha mindre end markens samlede areal, fortsættes algoritmen.

1.4.1.2 Undersøg om summen af registreringernes areal er tæt på markens samlede areal (maks. 5 ha fra)
- Hvis summen af registreringernes areal er tæt på markens samlede areal, konstrueres et vægtet gennemsnit af registreringernes udbytte vægtet med arealet, og algoritmen stopper.
- Hvis summen af registreringernes areal ikke er tæt på markens samlede areal, fortsættes algoritmen.

1.4.2 Undersøg om det totale udbytte er ens for alle registreringer
- Hvis det totale udbytte er ens for alle registreringer, vælges den registrering med den mindste arealforskel fra det samlede markareal, og udbyttet bliver markeret som 'utroværdig' (og bliver sorteret fra senere).

1.4.3 Undersøg om udbyttet er ens for alle registreringer
- Hvis udbyttet er ens for alle registreringer, vælges den registrering med den mindste arealforskel fra det samlede markareal, hvis denne er mindre end 1 ha, ellers laves der et vægtet gennemsnit (som blot bliver samme udbytteværdi), men hvor arealet af registreringerne summeres.

1.4.4 Undersøg om udbyttet er ens for alle registreringer

- Hvis udbyttet er ens for alle registreringer, vælges den registrering med den mindste arealforskel fra det samlede markareal, hvis denne er mindre end 1 ha, ellers laves der et vægtet gennemsnit (som blot bliver samme udbytteværdi), men hvor arealet af registreringerne summeres.

1.4.5 Undersøg om et eller flere udbytter er forskellig fra normudbyttet

- Hvis et eller flere udbytter er forskellig fra normudbyttet, vælges den registrering med den mindste arealforskel fra det samlede markareal, hvis denne er mindre end 1 ha, ellers laves der et vægtet gennemsnit.

For langt de fleste af dubletterne stoppes algoritmen ved skridt 1.4.1. Hvis algoritmen stoppes ved skridt 1.4.2 eller frem, markeres udbyttet som ' utroværdigt' og bliver sorteret fra senere i skridt 2.10.

2. Fjernelse af utroværdige udbytter

Efter fjernelse af dubletter, er der for en given mark i et givet år kun ét udbytte. Næste skridt er at fjerne utroværdige udbytter.

2.1 Fjernelse af udbytter, hvor det samme udbytte går igen i et givet år for det samme cvr

2.2 Fjernelse af udbytter, hvor det samme udbytte går igen på tværs af år for det samme cvr

2.3 Fjernelse af udbytter, hvor det samme totale udbytte går igen i et givet år for det samme cvr

2.4 Fjernelse af udbytter, hvor det samme udbytte, totale udbytte og areal går igen i et givet år for det samme postnummer

2.5 Fjernelse af udbytter, hvor det samme udbytte, totale udbytte og areal går igen i et givet år for den samme registreringsdato og samme editor

2.6 Fjernelse af udbytter, hvor udbyttet er lig med normudbyttet

2.7 Fjernelse af udbytter, hvor afgrøden ikke er vinterhvede

2.8 Fjernelse af udbytter, hvor udbyttet ligger uden for grænserne på 20 hkg/ha og 150 hkg/ha

Ved udbytter der ligger uden for grænserne, forsøges der først at se, om udbyttet ligger inden for grænserne, hvis der hhv. ganges eller deles med arealet. Det resulterende udbytte må ikke være for langt fra det gennemsnitlige udbytte for det pågældende cvr for at sikre, at det ikke blot var en tilfældighed, at udbyttet endte med at ligge inden for grænserne.

2.9 Fjernelse af udbytter, som er et heltal.

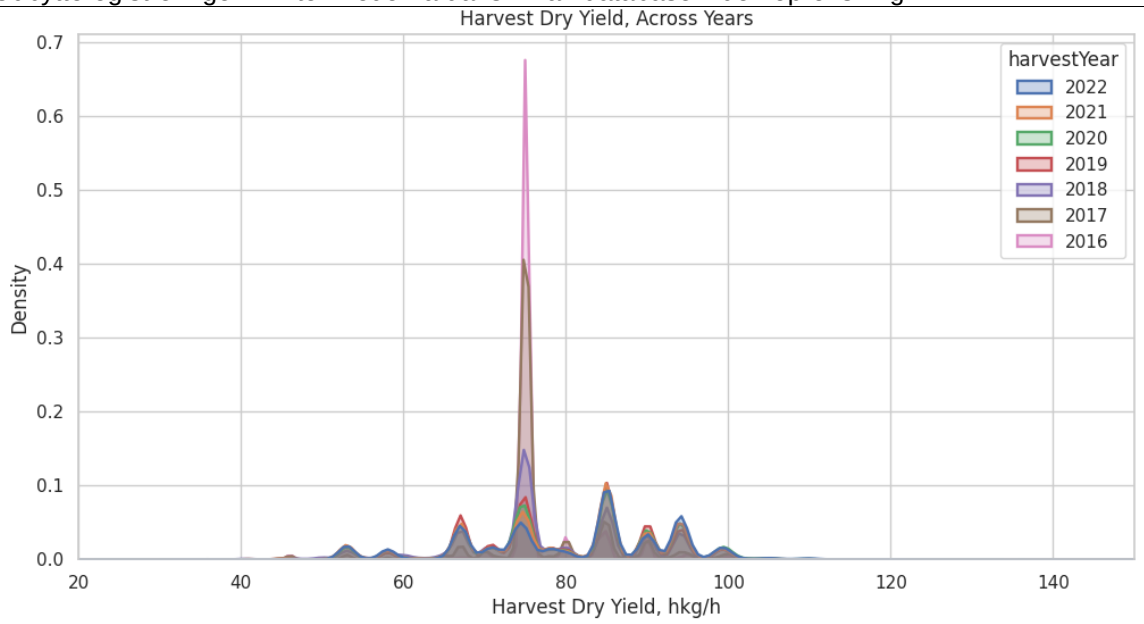
Trods alt tidligere oprensning, er der stadig alt for mange gengangere ved heltal - især heltal som slutter på 0 og 5, f.eks. 75, 80, 85 osv. Men også generelt heltal som f.eks. 53, 67, 74, 94. Derfor beholdes kun udbytter, som er et decimaltal.

2.10 Fjernelse af udbytter, som er markeret som 'utroværdig' under tidligere oprensningsskridt (bl.a. skridt 1.4.2 og 1.4.3)

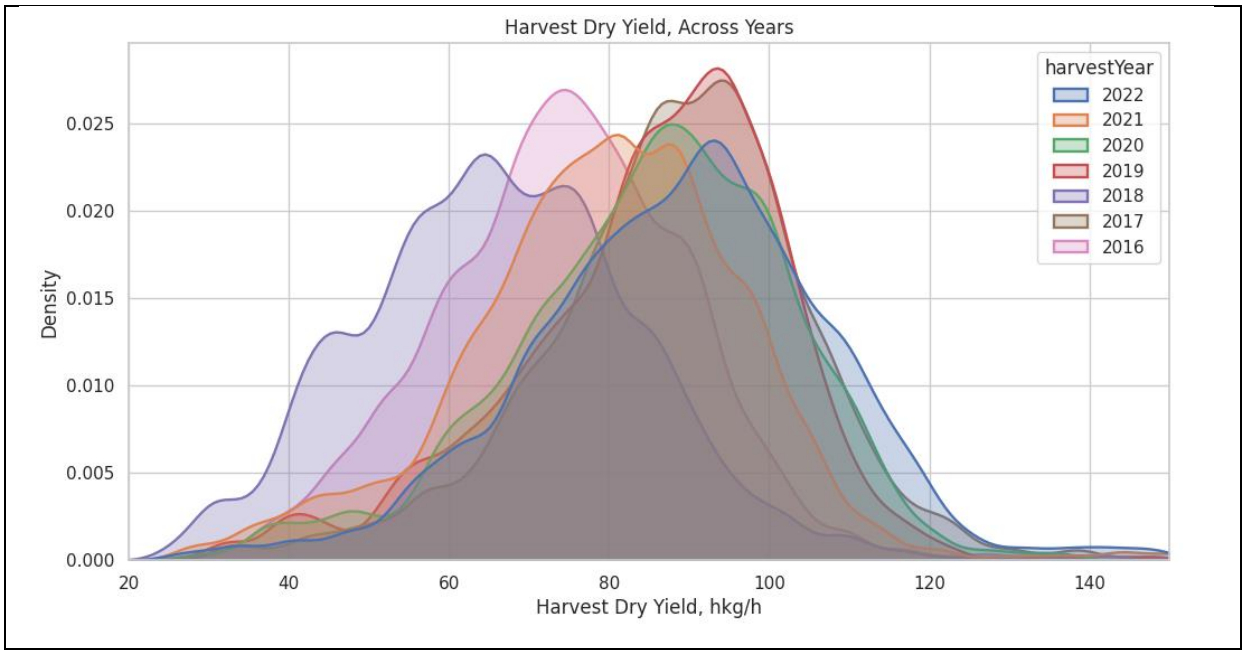
2.11 Fjernelse af marker, hvortil der ikke eksisterer et markpolygon

I projektet *"Lær af verdens største forsøgsareal"* støttet af Promilleafgiftsfonden er ovennævnte oprensning benyttet i forbindelse med udvikling af udbytteprognosen i vinterhvede. Inden oprensning af udbyttedata fra markdatabasen var der 423.574 marker med udbytteregistreringer fra 2016-2022, og efter oprensning var der 10.758 marker med valide udbytteregistreringer, som kunne indgå i modeludviklingen. Nedenfor i figur 1 ses fordelingen af udbyttedata fra 2016 til 2022 før oprensning og efter oprensning. Efter oprensning kan årsspecifikke trend i udbytteniveauerne genfindes. I 2018 lå udbytterne i vinterhvede lavere end gennemsnittet og i 2022 højre end gennemsnittet.

Udbytteregistreringer i vinterhvede fra dansk markdatabase inden oprensning



Udbytteregistreringer i vinterhvede fra dansk markdatabase efter oprensning



Figur 1. Fordelingen af udbyttedata fra 2016 til 2022 før (øverst) og efter oprensning (nederst).