

Promilleafgiftsfonden for landbrug

Notat

SEGES Innovation
Ledelse & Økonomi

Modellering af tidsmæssig udvikling i antal markdage	Ansvarlig	MATT/JKG
	Oprettet	12-07-2023
Projekt: 7891:2312 Vand væk fra værdifulde landbrugsjorde	Side	1 af 3

Modellering af tidsmæssig udvikling i antal markdage

I den følgende tekst beskrives formålet med den sidste del af nedbørsværktøjet (Model D, regression fit). Først forklares hvad formålet med værktøjet er. Dertil hvordan værktøjet belyser dette formål. Til sidst diskuteres hvorfor præcist denne løsning er valgt.

Formålet med værktøjet kan beskrives på følgende måde:

At kunne lave et værktøj der viser udviklingen i en givet kommunes antal markdage på tværs af år for et givet tidspunkt på året. Modellen er udelukkende baseret på nedbørsdata i dagene (ugerne) op til den valgte dato. Man skal gerne kunne se en tendens samt spredningsudviklingen på årlig basis. Derudover skal modellen give et skøn på udviklingen de kommende år (frem i tiden).

Der skal altså laves et værktøj som kan beskrive a) hvad har tendensudviklingen været og b) hvordan har spredningen ændret sig over tid. Vi belyser dette ved hjælp af to statistiske mål: et mål for udviklingen i middelværdien og et mål for udviklingen i standardafvigelsen. Det første mål – middelværdien – beregnes som et eksponentielt vægget rullende gennemsnit (EWMA). Metoden tager hensyn til både tidligere observationer og de nyeste observationer, men giver større vægt til de nyere observationer. Dette betyder, at den resulterende værdi vil reagere hurtigere på ændringer i de seneste observationer end de ældre. Altså vil perioder i den nære fortid vægtes højere end perioder langt tilbage i tiden. Metoden er rigtig god til at beskrive tendensen i data, især hvis denne ændrer sig over tid. For at metoden virker skal man vælge en værdi for vægtnings-parameteren beta. Hvis vi vælger en høj værdi for beta (tæt på 1) så er nyere observationer meget vigtige og hvis vi vælger en lav værdi for beta (tæt på 0) så er ældre observationer også vigtige. Der er ikke noget regel ift. at vælge beta, så dette er baseret på empirisk at prøve sig frem. Det andet mål er en rullende standardafvigelse. Værdien i dag beregnes som standardafvigelsen for de sidste K perioder (år). Standardafvigelsen er et mål for hvor meget de enkelte år afviger for middelværdien i standard enheder. Det vil sige at en standard afvigelse på 2 betyder at observationerne forventes at afvige 2 markdage fra middelværdien inden for tidsintervaller K-perioder tilbage. Dette giver netop et mål for spredningen i antal markdage. Vi beregner et rullende mål, fordi vi gerne vil se udviklingen over tid. Igen skal K-vælges og dette er også en empirisk proces som kræver at man "prøver sig frem".

I modellen inkluderer vi også en simpel lineær regressions linje som virker som en baseline for trenden. Her antager vi at sammenhængen mellem data er konstant over tid. Denne metode er meget rigid og fanger ikke fluktuationer i tendensen over tid. Man kunne forestille sig at der i perioden 1995-2005 er en anden udvikling i tendensen end fra 1987-2020 (hele perioden) og dette afspejles ikke i den simple metode. Vi oplyser dette mål som sammenligningsgrundlag til EWMA.

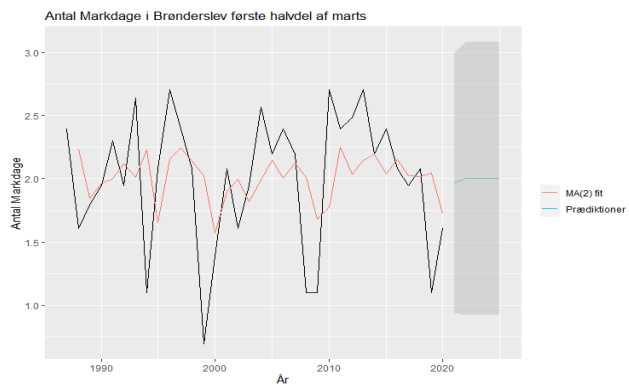
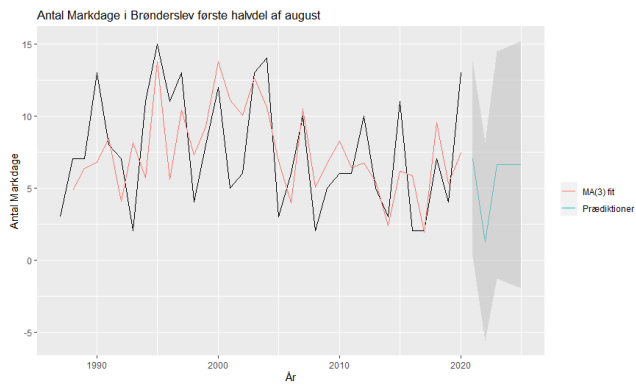
De ovennævnte metoder er relativt simple, men belyser netop hvad vi gerne vil besvare – udviklingen i tendens og spredning. I det følgende afsnit diskuteres det kort hvorfor vi ikke har valgt meget sofistikerede modeller samt problemerne ved at prædiktere ud i fremtiden.

Det første argument for ikke at vælge mere komplekse metoder er at ovenstående virker godt. Metoderne er simple at gennemskue og lette at fortolke på. Mere komplekse metoder kan måske fange en mere præcis udvikling, men

dette vil være på baggrund af en ringere fortolkning. Med henblik på hvad værktøjet skal anvendes til, er det mere relevant at kunne præsentere et gennemskueligt og håndgribeligt mål. Det næste argument, som i høj grad understøtter valget af en mere simpel løsning, er mangel på data. Mere komplekse modeller er baseret på parameter-estimationer som forringes når datagrundlaget er tyndt. Sagt på en anden måde så indeholder komplekse modeller en usikkerhed som normalt forsvinder når de præsenteres for meget data – dette opnås ikke. Når vi har lidt data, har mere sofistikerede modeller også større risiko for at 'overfitte' data. Det vil sige at modellerne tager højde for ligegyldige udsving (også kaldt støj), som er unikt for enkelte år, men som ikke siger noget om den generelle udvikling. I sådan et tilfælde vil vi altså præsentere et mål, hvor det er svært at fortolke på den generelle udvikling, fordi målet inkluderer irrelevant støj.

Foruden at beskrive tendensen og spredningen, så ville værktøjet ideelt også indeholde en prædiction / et skøn på udviklingen de næste år frem. Først og fremmest argumenterer vi for at dette ikke giver mening baseret på de simple mål vi benytter ovenfor. Hvis vi skulle bruge EWMA til at prædikere ud i fremtiden, så skal vi antage at tendensen udvikler i samme retning som den har gjort de sidste år op til 2020. Vi har intet belæg for at dette er en plausibel antagelse og prædictionen bliver derfor lige så usikker som hvis man bare tegnede en tilfældig streg ud i fremtiden. Man kan selvfølgelig under denne antagelse se hvad vi skal forvente de næste par år. Dertil kan man spørge sig selv: Givet at jeg ikke ved om denne antagelse holder, hvor meget kan jeg så egentligt bruge denne prædiction til? Når nu disse simple mål ikke er relevante for en prædiction, hvorfor så ikke benytte sig af mere komplekse modeller. Her benyttes samme argumentation som ovenfor – mangel på data. Med få datapunkter er tidsserien ikke repræsentativ for den underliggende proces og vi kan derfor ikke lave pålidelige prædictioner (uanset modellen). Desuden er disse mere komplekse modeller sensitive overfor udsving når vi har få datapunkter – overfitting – og dette øger usikkerheden endnu mere. Hvis man skal lave en valid prædiction i dette tilfælde, så bør det være på baggrund af en forståelse for den underliggende proces – har vi det? (nej). Derudover vil prædictionen også være på baggrund af historisk data, da ingen forklarende variable er inkluderet. Det er urealistisk at antal markdage kun afhænger af antal markdage i tidligere år. F.eks. vil man forvente at klimaforandringer har en betydelig påvirkning på nedbørsmængderne. Alt i alt er der ikke inkluderet en prædiction af fremtidige års antal markdage, da vi vurderer at dette skøn ikke er informativt når usikkerheden tages i betragtning.

Som et illustrativt eksempel er det skitseret to modelleringer samt prædictioner for antal markdage i hhv. første halvdel af august og første halvdel af marts for Brønderslev kommune. Konfidensintervallerne viser tydeligt at prædictionerne ikke har nogen relevant information.



Resultatet af modellens anvendelse er at finde i notatet med økonomiske konsekvenser af det vildere vejr.