Denne artikel er udgivet i det elektroniske tidsskrift
Artikler fra Trafikdage på Aalborg Universitet
(Proceedings from the Annual Transport Conference at Aalborg University)
ISSN 1603-9696
https://journals.aau.dk/index.php/td

trafikdage
NY VIDEN & NETVÆRK

Skibsfartsstatistik baseret på big data

Peter Ottosen, pot@dst.dk
Danmarks Statistik

Abstrakt

Big data passer ofte dårligt inden for rammerne af officiel statistik på grund af deres flygtige og ustrukturerede karakter. AIS-data er en big datakilde præget af stor stabilitet og struktur – til gengæld er datamængden omfattende med et stort behov for filtrering og behandling for at passe ned i den officielle maritime transportstatistiks afgrænsning.

AIS-data er et international aftalt format for udveksling af skibspositioner og er en relevant datakilde for alle lande med kyst. I Danmark overvåger Søfartsstyrelsen fartøjer i dansk søterritorium med en række landbaserede AIS modtagere. Siden 2016 har Danmarks Statistik modtaget et live feed af de indsamlede data og gemt dem og i januar 2020 publicerede Danmarks Statistik sin første månedsstatistik baseret på AIS-data. Da COVID19 ramte bare en god måned senere, kunne AIS-data derfor let omdannes til en høj frekvent indikator.

Artiklen vil vise de processer, Danmarks Statistik bruger for at identificere havne-lignende områder med brug af AIS-data alene og for at konvertere data til overskuelige og mere traditionelle datasæt, som kan bruges som supplement eller erstatning af officiel havnestatistik.

Kernen i databehandlingen er den cluster metode, der ud fra tætheden af fortøjrede fartøjer afgrænser havne. For store havne med afgrænsede kajområder kan der sågar skelnes mellem disse. Afgrænsningen kan dernæst anvendes til at skelne mellem fartøjer, der anløber kaj og dem, der blot passerer havneområderne.

Baggrund

EU-kommisionen har i en årrække medfinansieret projekter, der skulle belyse mulighederne i big data som kilde til officiel statistik. Danmarks Statistik indgik omkring 2016 i et sådan projekt. Som en del af projektet startede en løbende opsamling af AIS[1] data fra et live feed fra Søfartsstyrelsen.[2] Danmarks Statistik har

[1] Automated Information System (AIS) er et informationsudvekslingssystem mellem skibe, som deler oplysninger om fx position, hastighed og retning. I Danmark indsamles AIS data af Søfartsstyrelsen gennem en række landbaserede datamodtagere.
[2] Søfartsstyrelsens hjemmeside for AIS data: https://www.soefartsstyrelsen.dk/sikkerhed-til-soes/sejladsinformation/ais-data

således registreret alle danske AIS data siden marts 2016.Big data karakteriseres ved en eller flere af følgende karakteristika: stor volumen, hurtig tilgængelighed og stor variation eller kompleksitet. AIS-data er velstrukturerede data i faste formater, men der dannes utrolige mængder af information og de er til gængelige i real time. Et yderligere karakteristika ved en del big data kilder er en kort levetid. Det gør sig ikke gældende for AIS data heller. De er bundet op på internationale aftaler gennem International Maritime Organisation.

I 2018 begyndte et arbejde i regi af transportstatistikkerne i Danmarks Statistik med at operationalisere data med henblik på at afdække potentialet i AIS data og eventuelt udarbejde statistik baseret på data – i første omgang som eksperimentel statistik. Arbejdet resulterede i en AIS baseret månedlig havnetrafikindikator, som gik i luften første gang den 12. februar 2020.[3]

Eksperimentel statistik er et begreb, der har vundet indpas hos de officielle statistikproducenter i de senere år. For hurtigt at teste nye produkter, metoder, data kilder og være mere fremme med helt relevante data har man introduceret begrebet for at kunne publicere betaversioner af statistikker og få input fra en bredere brugergruppe end den traditionelle statistikproduktion. Begrebet fik for alvor luft under vingerne internationalt og i Danmark i forbindelse med COVID-19, hvor behovet for helt aktuel statistik opstod. I Danmark var den AIS baserede anløbsstatistik dog allerede i januar 2020 publiceret som den første eksperimentelle statistik i Danmarks Statistik.

Indholdet i datakilden

I IMO (2015) beskrives indholdet i AIS-meddelelserne. De relevante meddelelser for Danmarks Statistiks arbejde sendes fra skibene og består af en statisk del med faste oplysninger om AIS senderen og fartøjet; en dynamisk del, som består af automatisk opdaterede data om position, retning, hastighed mm. og endelige turrelaterede oplysninger, som skal opdateres manuelt og kan kræve opdatering undervejs på en tur.

De vigtigste oplysninger i de statiske data er MMSI-nr (AIS-transponderen nr), IMO nr, skibsnavn og –type, kaldesignal, dybdegang og destination samt forventet ankomst til destinationen.

De vigtigste oplysninger i de dynamiske data er MMSI-nr, som bruges som nøgle; navigationsstatus, som angiver om skibet fx er for anker, ved kaj eller sejler; fart i 1/10 knob; længde- og breddegrad samt kurs.

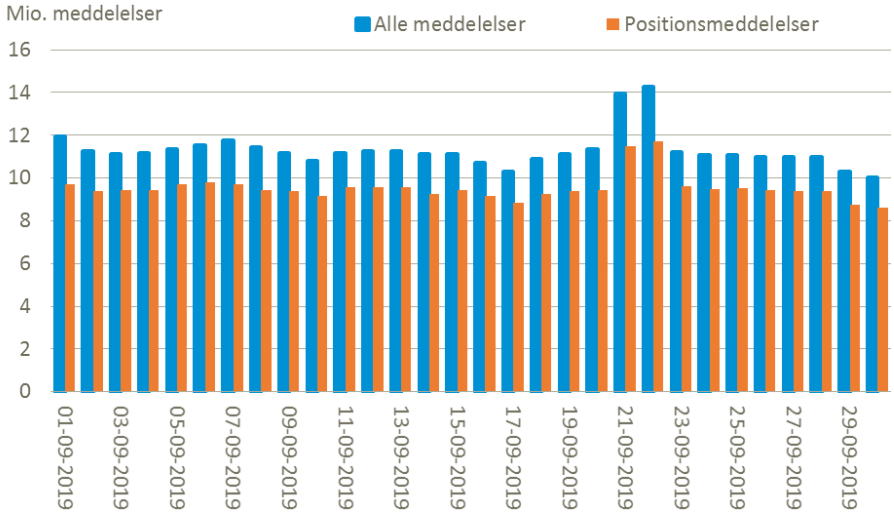
Desuden indeholder alle meddelelsestyper et tidsstempel.

Hyppigheden af afsendelsen af de dynamiske data afhænger af skibets fart, men der er mellem 2 og 10 mellem hver meddelelse – jo hurtigere des hyppigere. D

Danmarks Statistik modtager typisk over 300 mio. meddelelser i live feedet fra Søfartsstyrelsen hver måned, hvoraf næsten 85 pct. er positionsmeddelelser.

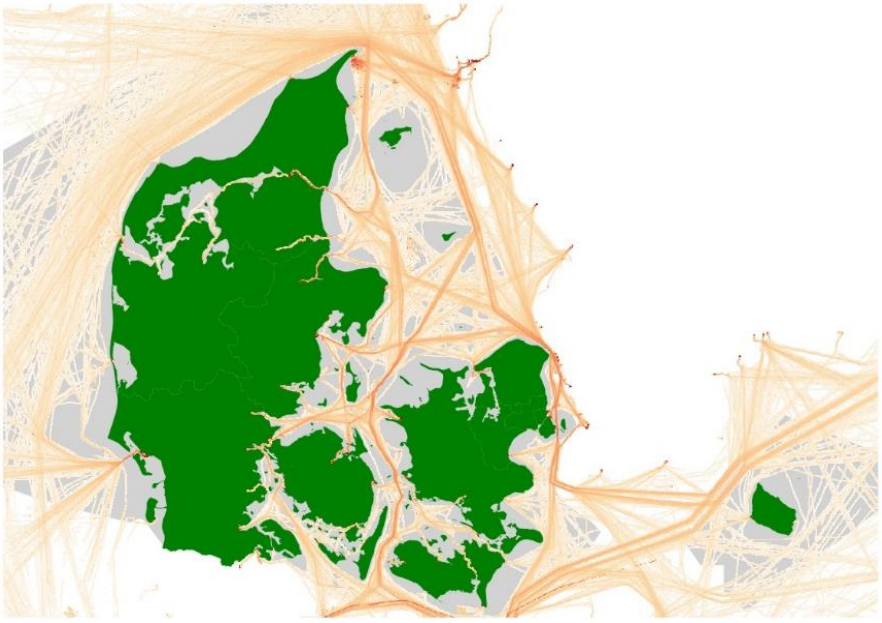
[3] Statistiktabellen i Statistikbanken.dk: https://www.statistikbanken.dk/ais1

Figur 1 Antal AIS meddelelser i september 2019



For at illustrere data vises i nedenstående Figur 2 et heatmap dannet ud fra alle skibspositionsmeddelelser modtaget i juni 2019. Herved tydeliggøres sejlruterne og deres trafikintensitet for skibstrafikken i den vestlige del af Østersøen og danske farvande. Dansk territorial farvand er markeret med grå baggrund.

Figur 2 Tæthed af registrerede skibspositioner juni 2019



AIS data er ikke ufejlbarlige. Data udveksles via et VHF signal, hvilket betyder dels, at det enkelte signal kan påvirkes af vejrforhold og blive fejlbehæftet samtidig med, at der ikke er en kontrolprotokol, hvor transmissionsfejl detekteres, dels at AIS-modtagerne kun kan modtage et mindre antal meddelelser ad gangen, således, at der i stærkt trafikerede områder går signaler tabt, fordi modtageren ikke har plads. På grund af den høje frekvens af dataudvekslinger, skønnes fejlen ikke at kunne påvirke resultaterne i nævneværdig omfang.

Metode til havneafgrænsning

I andre anvendelser af AIS data til at se på skibstrafik i havne har man typisk valgt en metode, der kan kaldes *geofencing*. Der er defineret en linje omkring havnen og skibe, der passerer linjen, betragtes som været anløbet havnen. Metoden har tre primære problemer. For det første er det tidskrævende at definere

linjerne; for det andet betyder ændringer i havnenes størrelse, at grænserne skal opdateres; og endelig kan det være vanskeligt afgrænse en havn, der ligger i et smalt stræde eller sund på en måde, der frasorterer forbipasserende fartøjer.

Metoden, der bruges i Danmarks Statistiks opgørelse, tager udgangspunkt i AIS data til at definere grænsen for havnen. Dermed undgår vi primært de to første problemer, mens problemet med havne i smalle stræder og sunde reduceres markant.

Der er fire primære trin i processen:

1. Datareduktion
2. Udvælgelse af ankomstobservation og afgangsobservation
3. Afgrænsning af havne
4. Placering af ankomst/afgangsobservationer i havne og statistik

Datareduktion

Det første trin er datareduktion. Formålet er udelukkende at reducere mængden af data, der skal behandles og består af tre dele: geografisk reduktion, reduktion i observationshyppighed og reduktion i analyserede skibstyper. Det vigtigste i datareduktionen er, at den ikke påvirker data på en måde, der skævvrider resultatet.

AIS data indeholder alle de informationer fra AIS systemet, som de danske modtagere registrerer. Det inkluderer en lang række observationer, som ikke direkte vedrører Danmark, men snarere Tyskland, Norge eller Sverige og en del transitsejlads. Metoden til geografisk afgrænsning er simpel, idet data afgrænses med en firkant, som således omkranser hele Danmark og dermed også den sydlige del af Sverige for at få Bornholm dækket. De svenske havne bliver frasorteret i senere trin. Her er der mulighed for raffinering og afgrænse data til kun at indeholde observationer, der er inden for det danske søterritorium. Årsagen til at den mere raffinerede metode ikke er valgt, er, at geodata håndtering er væsentlig mere ressourcekrævende end en sammenligning af tal og den valgte løsning kræver blot, at længde- og breddegraden holder sig inden for et givet interval.

Skibene udsender signaler med et interval bestemt af hastighed og aktivitetstype (sejler, ligger for anker mm.) samt type af senderen/transponderen. De hyppigste signaler kommer med 2 sekunders mellemrum. Når formålet er at undersøge anløb i havnene, er så hyppige data slet ikke nødvendige. Data reduceres derfor til den første observation per minut. Også her kunne der ske en yderligere reduktion. Transaktioner, der indebærer godshåndtering, vil typisk vare væsentlig længere end et minut.

Aktiviteten i havnene kan dække over mange typer af aktiviteter. For at understøtte eksisterende havnestatistikker, som fokuserer på godshåndtering i havnene, er data reduceret til de skibstyper, der udelukkende anvendes til godsfragt: fragtskibe og containerskibe.

Ankomst- og afgangsobservation

Det næste trin i processen er at identificere skibsanløbene med ankomst og afgang. Hvert skib efterlader en række af positionsoplysninger og målet er at identificere den observation, der repræsenterer ankomsten til en havn og den tilhørende observation, der repræsenterer afgangen fra havnen.

Processen er principielt simpel: Datasættet sorteres efter skibsidentifikation og tid. Dernæst markeres alle observationer, hvor skibet går fra at sejle (markør sat til *under way* skifter til *moored* samtidig med, at skibet går fra at bevæge sig (mere end 1 knob) til at lægge (næsten) stille). Det definerer potentielle ankomster. Det samme gøres for potentielle afgange, hvor skibet går fra at have markøren sat til *moored* til *under way* og det bevæger sig.

Med alle potentielle ankomster og afgange matches de sammen i par. Langt hovedparten hænger fint sammen, men der er både ankomster med manglende afgange og afgange med manglende ankomster.

Mulige forklaringer på manglende parring er:

- Lange havneophold: Medianopholdstid i havnen er omkring 12-13 timer, men hvis den er flere dage lang, kan der i starten eller enden af den betragtede periode mangle den matchende observation. Det kompenseres delvis for ved at bruge data, der rækker ud over perioden, fx vente 5-6 dage før seneste måneds statistik udarbejdes. Meget lange ophold i havnen er sjældent forbundet med fragtsejlads, men snarere et behov for reparation eller andet og er formentlig ikke væsentlige i forhold til statistikkens formål.
- Slukket transponder: Det er (som regel) ikke ulovligt at slå transponderen fra, og man kan forestille sig, at skibe først får slået transponderen til, når havnen er anløbet eller får den slået fra i havnen og glemmer at tænde den igen, når havnen forlades.
- Dataudfald: Der findes udfald i data, hvor data fra en enkelt modtager ikke kommer ind i en periode, data ikke streames fra Søfartsstyrelsen eller hvor data ikke opsamles og lagres af Danmarks Statistik. Det sidste var særlig i opstartsfasen og udfald forekommer ikke længere. Søfartsstyrelsen gemmer selv data, men ikke med det fulde datasæt, så mistede feeds kan ikke gendannes.

Afgrænsning af havne

Det tredje og i denne sammenhæng afgørende trin er at afgrænse anløb til de observationer, der rent faktisk er anløb i en havn. Der er på dette tidspunkt fortsat en lang række af ankomst/afgange, som ikke er i nærheden af en havn. Ydermere skal data afgrænses til danske havne. Der er fortsat observationer fra særligt svenske havne med i datagrundlaget, herunder aktiviteten i Nordens største godshavn, Göteborg. Hvis data fra starten bliver reduceret til dansk søterritorium, forsvinder denne sidste del også.

Når dette tredje trin er lavet på mindst et par års data, kan en enkelt måneds statistik køres uden at danne en ny havneafgrænsning. Det reducerer den månedlige produktionstid markant og indarbejdelsen af nye data i afgrænsningen kan foretages efter statistikproduktionen.

Selve processen i dette trin består af tre dele:

- Sammenknyt observationer (ankomster bruges) i grupper (clusters) efter den indbyrdes afstand mellem observationerne og antallet af observationer i nærheden.
- Dan polygoner, som omslutter alle observationer i samme cluster
- Forbind de enkelte polygoner til faktiske havne

Resultatet af de tre dele er en spatial opslagstabel, som består af polygoner og tilhørende havn. En havn kan bestå af flere polygoner, mens en polygon kun kan være tilknyttet én havn. Hvis en ankomst ligger inden for en given polygon, ved vi dermed hvilken havn, anløbet hører til. For større havne repræsenterer hver polygon typisk et særskilt kajanlæg, som ofte er dedikeret til specifikke godstyper.

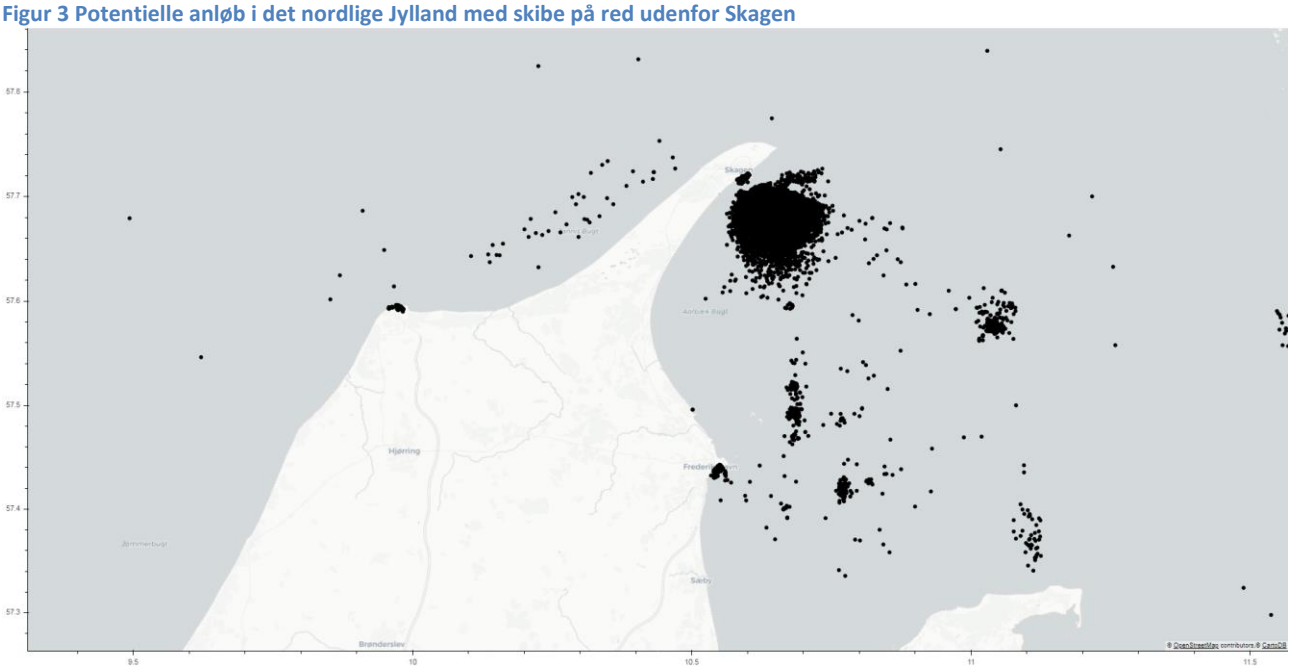
Første del består i at beregne alle indbyrdes afstande mellem ankomstpunkterne, hvorefter ankomsterne samles i grupper. Hvis to observationer (uafhængigt af tid) er mindre end fx 50 meter fra hinanden, knyttes de sammen. Hvis en tredje observation er mindre end 50 meter fra en af de første to observationer, bliver den også en del af gruppen. Alle grupper med flere end fx 5 observationer bliver en endelig gruppe. Antallet af observationer i en gruppe og afstanden mellem observationer i gruppen er parametre, der kan tilpasses. Jo færre observationer, der anvendes des større bør afstanden være og des mindre bør antallet af observationer i en gruppe være. Med tre års data anvendes 70 meter og 5 observationer. Efter første skridt har alle observationer fået tilknyttet en gruppe (cluster). Alle observationer, der ikke opfylder kriterierne får clusternummer -1 og betragtes som falsk positive anløb.

Anden del danner polygoner, der omkranser de enkelte clusters, således at alle observationer i samme cluster ligger indeni eller på kanten af polygonen. Hvis man ser på beliggenheden af disse clusters, får man uden videre bearbejdning et godt billede af, hvor der er havne. I lande, hvor der etableres uofficielle havne, kan man her identificere disse.

Tredje del knytter de enkelte polygoner til en havn. Det gøres i en iterativ proces, hvor der tages udgangspunkt i et havnecentroid (eller tilsvarende – i praksis anvendes oftest koordinatet for havnen, som angivet i UnLocode listen, men det kan også findes ved simpel opslag i fx Google Maps). Den iterative proces sikrer, at ingen clusters knyttes til mere end én havn og at polygonen knyttes til den havn, der ligger tættest på polygonen. Tætliggende havne, som skal kunne adskilles, kan give lidt manuelt arbejde med at tilpasse udgangspunkterne, så de enkelte cluster kommer med i den rette havn. Iterationen foregår ved, at der dannes en gradvis større cirkel omkring udgangspunktet for havnen. Hvis cirklen overlapper en clusterpolygon, tilknyttes denne cluster til den pågældende havn, hvorefter den ikke kan kobles til andre havne.

Den maksimale afstand fra havnens udgangspunkt for clusters må defineres ud fra konkrete data. I de danske data er begrænsningen sat efter, at der ud for Skagen, se Figur 3, er defineret en række clusters for skibe, der ligger på red. De betragtes reelt ikke som værende i havn og grænsen er sat, således, at disse clusters ikke medtages. Da der udelukkende tages udgangspunkt i danske havne, sorterer processen også de primært svenske havne fra.

Figur 3 Potentielle anløb i det nordlige Jylland med skibe på red udenfor Skagen



Ved afslutning af processen har vi en opslagstabel, hvor de øvrige oplysninger om konkrete havne også kan findes, fx kommune, unlocode, navn, CVR nummer og kystzone.

Endelig allokering af anløb til havne

Det sidste trin i hele processen er, at anløbene (defineret som ankomst og tilhørende afgang i trin 2) matches op mod polygonerne, som definerer (dele af) havnene. Resultatet er et datasæt, som indeholder oplysninger om anløbet (primært tid, skibsidentifikation og skibsnavn) og den havn, der er anløbet (primært havnens navn, kode og landsdel). Data kan yderligere beriges med mere detaljerede skibsoplysninger fra skibsregistre, fx størrelse, flagstat, mere præcis skibstype og ejer/operatør.

Herfra er processen helt standard for statistikproduktion. Tabellering på eventuelle på underopdelinger, indeksering og sæsonkorrektion.

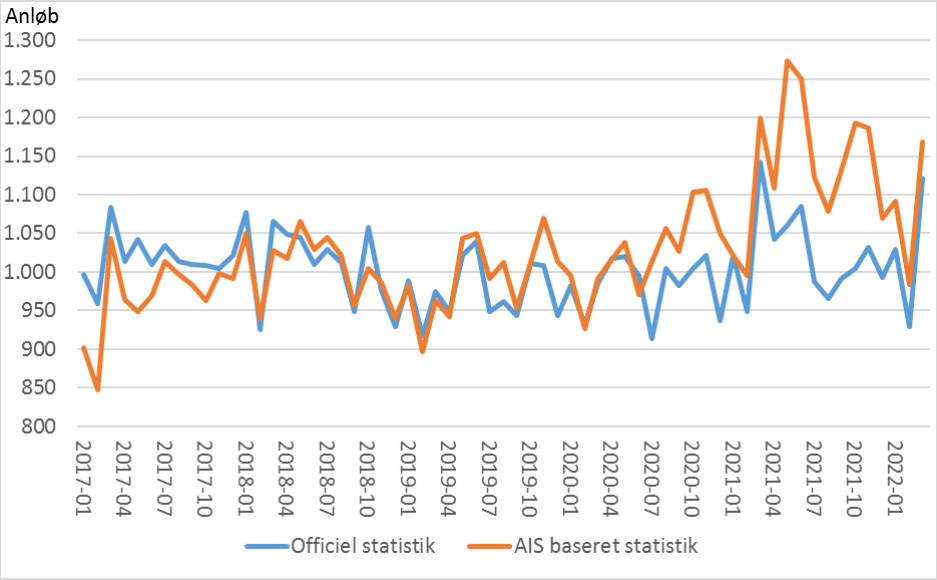
Konklusion

Der er generelt en god overensstemmelse mellem anløbene opgjort med AIS data og de officielle statistikker. Der skal dog understreges, at der er grundlæggende definitoriske forskelle på de to opgørelser, idet AIS baseret statistik ikke skelner mellem forskellige typer af anløb, hvor den officielle havnestatistik udelukkende opgør anløb, hvor der lastes eller losses gods. Omvendt er AIS statistikken baseret på skibe, der er registreret som fragtskibe, mens den officielle statistik ikke begrænses af den registrerede skibstype.

Som det fremgår af Figur 4 har der i det meste af opgørelsesperioden været en pæn overensstemmelse mellem de to opgørelser. Der er en korrelationsfaktor på 0,59.

I 2021 har der dog været en større diskrepans, hvis årsag endnu mangler blive afklaret. Det er forhold som det, der skal afklares før AIS data kan indgå i en mere officiel rolle i statistikproduktionen.

Figur 4 Skibsanløb på større danske havne efter metode



Big data kan være flygtige i deres natur – de opstår og nedlægges. Officiel statistik baseret på flygtige kilder er problematisk, idet konsistens over tid kan vanskeliggøres. En big data kilde som AIS baseret på internationale aftaler er der i mod oplagt at integrere i statistikproduktionen, idet der kan forventes at data indsamles i overskuelig fremtid og at den basale information overlever. Udfordringen bliver således mere af proces-, afgrænsnings- og definitionsmæssig karakter.

Perspektiver

Gennemgangen her beskæftiger sig udelukkende med anløb for at anskueliggøre potentialet i data. Der er flere potentielle anvendelsesmuligheder for AIS data, som skal undersøges yderligere. Her nævnes blot et par stykker.

I selve AIS data findes også data om dybdegangen. Dybdegang er usikkert, da det registreres manuelt, men synes dog relativ systematisk at blive opdateret i forbindelse med anløb i havnene. Dybdegangen på skibet bestemmes af to forhold: godsmængden og ballast. Ved at kombinere enkeltanløb fra AIS data og den officielle statistik, kan der eventuelt etableres en sammenhæng mellem dybdegang og lastning/losning. Som minimum kan det bruges til at selektere godsrelaterede anløb og dermed nærme sig definitionen i den officielle statistik. I bedste fald vil ændringer i dybdegang kunne bruges som en indikator for retning og håndteret godsmængde.

AIS data kan knyttes til de enkelte skibe og dermed til registeroplysninger om fx flag, reder, bruttotonnage mm. En del af de oplysninger indsamles udelukkende fra store havne, men små havne, der indberetter årlig, udelukkende indberetter summariske oversigtstal. AIS data kan muligvis anvendes til at supplere oplysningerne fra de små havne, så der kan laves hurtigere komplette opgørelser af havneaktiviteten. Endvidere kan der muligvis være et forenklingspotentiale i brugen af AIS data.

Referencer

IMO, 2015. Resolution A.1106(29) adopted on 2 December 2015 (Agenda item 10) Revised Guidelines for the Onboard Operational Use of Shipborne Automatic Identification Systems (AIS)