KAN VI GOOGLE DET?

Bruk av stordata til prognoser for arbeidsledigheten

Av Malin Charlotte Engel Jensen

Sammendrag

Formålet med denne analysen er å undersøke om bruk av stordata fra Google Søketrender kan benyttes til å lage kortsiktige prognoser for arbeidsledigheten i Norge. Modellene og rammeverket som presenteres gir treffsikre anslag på de kortsiktige svingningene i konjunktursyklusen. Modellene kan, for eksempel, fange opp et brått og uventet hopp i ledigheten. Slik hyppig og pålitelig informasjon om forventet utvikling i arbeidsledigheten vil gi NAV bedre styringsinformasjon og muligheten til å agere raskere og med mer effektiv ressursbruk.

Et typisk eksempel på et verktøy som samler inn stordata, er Google Søketrender. Google Søketrender er en internettbasert tjeneste som lager statistikk og systematiserer hva folk søker etter på Google. Datagrunnlaget kan antas å reflektere sanntidsinformasjon om søkemotorbrukerens intensjoner (til å blant annet gjennomføre en økonomisk beslutning). I denne artikkelen undersøkes hvorvidt Google Søketrender kan benyttes til å lage kortsiktige prognoser, såkalt «nowcasting», for arbeidsledigheten.

Resultatene fra analysen viser at prognosemodeller basert på Google søketrender gir statistisk signifikante og presise anslag for arbeidsledigheten. Dette finner jeg ved å sammenligne Søketrend-modellenes anslag med anslag laget av to kjente referansemodeller. De empiriske resultatene stemmer overens med tidligere forskning på området og indikerer at rammeverket som benyttes her er stabilt på tvers av utfallsvariabler.

Nøkkelord: nowcasting, Google søketrender, prognose, stordata, arbeidsledighet, kortsiktige prognoser, Google trends, konjunktur analyse

Innledning

NAV publiserer månedlig ledighetsstatistikk. Disse tallene er blant de makroøkonomiske størrelsene det er knyttet størst oppmerksomhet til i norsk økonomi. Ledighetstallene er en av de viktigste temperaturmålerne på tilstanden i økonomien og følges med stor interesse av en rekke offentlige institusjoner, finanssektoren, media og befolkningen for øvrig. I likhet med mange andre makroøkonomiske nøkkelvariabler, blir ikke ledighetstallene publisert fullt så hyppig og med et relativt stort tidsetterslep[1]. Store avvik mellom når lediggang faktisk inntreffer og når det blir målt og tilgjengeliggjort, gjør det vanskeligere å fatte treffsikker økonomisk politikk. Med dette som bakteppe, er det i den senere tid blitt diskutert hvorvidt kortsiktige prognoser, heretter omtalt som «nowcasting», av slike nøkkelvariabler kan gi beslutningstakere et bedre informasjonsgrunnlag for å gjennomføre økonomisk politikk.

Det engelske ordet nowcasting har sitt opphav fra de to ordene «now» (nå) og «forecasting» (prognose) og er ment å henvise til svært kortsiktige prognoser. Prinsippet med nowcasting er enkelt. Vi bruker data, eller mer spesifikt indikatorer, som blir publisert tidligere og mer hyppig enn den variabelen vi ønsker å predikere. Hensikten er å finne indikatorer som er tilgjengelige *før* offisiell statistikk blir publisert, se Banbura mfl. (2013).

I denne artikkelen er tanken å bruke nowcasting til å gi oss et mer oppdatert temperaturmål på arbeidsledigheten. Dette er særlig nyttig for NAV: Jo tidligere indikasjoner vi har på at ledighetsutviklingen er på vei til å snu, desto bedre kan NAV tilpasse seg svingningene i den norske økonomien. Store og uforutsigbare hopp i ledigheten medfører, blant annet, lengre vente- og behandlingstid for brukere som trenger bistand fra NAV ved lediggang. Hvis NAV-kontorene har et mer oppdatert informasjonsgrunnlag idet slike sjokk inntreffer, kan de forberede seg og møte situasjonen mer effektivt. Videre gir kortsiktige prognoser med presise estimater på den nåværende tilstanden i det norske arbeidsmarkedet, et bedre utgangpunkt for mer langsiktige prognoser. Dermed kan kvaliteten på informasjonen som NAV gir til beslutningstakerne styrkes slik at den økonomiske politikken blir mer treffsikker.

Det er flere aktuelle variabler som kan benyttes som indikatorer på de kortsiktige svingningene i arbeidsledigheten. Finansielle variabler, slik som valutakursen eller renter, er særlig interessante indikatorer ettersom disse oppdateres i sanntid og til en viss grad kan betraktes som tett sammenvevd med resten av økonomien. Thorsrud (2018) fremhever imidlertid to problemer ved å bruke slike variabler som beslutningsgrunnlag. For det første er sammenhengen mellom indikatorene og aggregerte makrovariabler svært ustabil. For det andre er det vanskelig å identifisere hva slags type sjokk som får finansielle variabler til å svinge. Se eksempelvis på den norske kronen: Tradisjonell økonomisk teori predikerer at når differansen mellom den norske styringsrenten og internasjonale styringsrenter øker, vil den norske kronen styrke seg. I praksis har imidlertid den norske valutakursen vært rekordsvak til tross for at denne differansen har økt som følge av at styringsrenten her hjemme har blitt hevet tre ganger det siste året. Når man ikke vet hvorfor variablene fluktuerer slik som de gjør, blir det vanskelig for en beslutningstaker å vite hva man skal basere avgjørelsene sine på. Kanskje er det fruktbart å undersøke hvorvidt andre datagrunnlag, som vi både kjenner de bakenforliggende drivkreftene til og som mest sannsynlig produserer lave prognosefeil, kan tas i bruk?

I denne artikkelen vil jeg bruke data fra Google Søketrender til å forklare de kortsiktige fluktuasjonene i den registrerte arbeidsledigheten. Ved å bruke Google Søketrender som datagrunnlag, unngår jeg problemene Thorsrud (2018) peker på. Ideen er at søketrendene samlet kan tjene som en god indikator på de kortsiktige fluktuasjonene i den registrerte ledigheten, ettersom et søk i Googles søkemotor reflekterer søkemotorbrukerens *intensjon* om å ta en (økonomisk) beslutning. Slike intensjoner vil bli målt av Google Søketrender **før** resultatene av de samme intensjonene blir målt og tilgjengeliggjort

[1] Et tidsetterslep (eller et «lag» på engelsk) beskriver en variabel som har sin verdi fra én tidsperiode tidligere.

gjennom offisiell statistikk. Vi kan lett tenke oss til hva vi selv ville søkt etter hvis vi trodde vi sto i fare for å miste jobben. Typiske søk kan inkludere ord som [dagpenger NAV], [finn jobb], [hvordan skrive jobbsøknad] og så videre. Denne informasjonen vil Google ha tilgang til før personen eventuelt registrerer seg hos NAV. Google Søketrend-variablene danner derfor et godt grunnlag for å kunne si noe om utviklingen i ledighetstallene den nærmeste tiden framover.

Hva er Google Søketrender, og hva har det blitt brukt til?

Google Søketrender er en nettbasert tjenesteplattform levert av Google som tilbyr høyfrekvent, disaggregert stordata om brukernes søkeadferd.[2] De siste ti årene har Google Søketrender gjort seg svært gjeldende som indikator på en mengde forskjellige variabler anvendt i forskningslitteraturen. Blant de første brukerne av Google Søketrender var Polgreen mfl. (2009) som brukte verktøyet for å overvåke og avdekke sykdomsutbrudd i USA. Siden har Google Søketrender blitt brukt til å måle alt fra hvor klima- og miljøbevisste vi er, til endringer i selvmordsrater, til å gi anslag på hvem som vinner presidentvalget i USA - og med hvor stor margin.

Choi og Varian (2009) var de første som utforsket bruken av søkedata fra Google i en samfunnsøkonomisk kontekst og argumenterte for at Google Søketrender var en relativt treffsikker indikator på den kortsiktige variasjonen i den amerikanske arbeidsledigheten. Resultatet finner også støtte i den norske litteraturen om Google Søketrender. Blant annet viser Anvik og Gjelstad (2010) til prognosemodeller som anslår den registrerte ledigheten med opptil 18 prosent høyere presisjon enn standard referansemodeller i over tolv måneder i strekk. I likhet med Anvik og Gjelstad (2010) og Ellingsen (2017) undersøker jeg hvordan vi kan benytte Google Søketrender som grunnlag for å danne kortsiktige prognoser av den registrerte ledigheten i Norge. Det som skiller denne artikkelen fra de sistnevnte, og som er artikkelens viktigste bidrag til litteraturen, er at det her tas i bruk et mer formelt rammeverk for å håndtere store datamengder. Metodikken er lånt fra Jensen (2019) som viste at Google Søketrend-modeller var signifikant bedre til å anslå de kortsiktige svingningene i kvartalsvis BNP enn to naive referansemodeller.

Google Søketrender for nybegynnere

Google Søketrender[3] rapporterer en indeks som beskriver interessen for ett enkelt ord som er tastet inn i Googles søkemotor over tid. Her får man hovedsakelig tilgang til et ufiltrert utvalg av søk som gjøres i Googles søkemotor. Utvalget er anonymisert, kategorisert og gruppert sammen. På denne måten kan Google Søketrender gi en oversikt over interessen for bestemte emner eller søkeord både nasjonalt og lokalt i Norge. Mer konkret rapporterer Google et *relativt mål på søkeinteressen*. Søkeinteressen viser hvor ofte ett gitt søkeord blir utforsket relativt til det totale søkevolumet. Volumet av søkeinteresse blir normalisert og deretter skalert. Normaliseringen skjer ved at Google deler søkeordet på et urelatert alminnelig søkeord. For eksempel kan vi tenke oss at indeksen for søkeordet [Champions League] blir normalisert ved å dele det på det urelaterte, alminnelige søkeordet [svimmel]. Videre blir indeksen skalert slik at den varierer mellom 0 og 100. På denne måten kan vi måle den relative endringen i interessen for et spesifikt søkeord over tid. Indeksen oppdateres daglig og spenner fra januar 2004 til i dag.

Av forskjellige grunner filtrerer Google Søketrender ut enkelte typer søk. Dette gjelder spesielt søk som gjøres av svært få personer, gjentatte søk som gjennomføres av de samme personene i løpet av korte tidsperioder og søk som inneholder apostrofer eller andre spesialtegn.

[3] https://trends.google.com/trends/?geo=NO

Data og Metode

I artikkelen vil jeg predikere endringen i bruttoledigheten. Dette betyr altså at bruttoledigheten er den avhengige variabelen, også kalt utfallsvariabelen, i prognosemodellene. Bruttoledighet defineres som summen av antallet helt ledige og antallet arbeidssøkere på tiltak, heretter omtalt som «ledigheten» eller «den registrerte ledigheten». Den registrerte ledigheten må ikke forveksles med «AKU-ledigheten» som er estimerte ledighetstall basert på intervjuundersø-

[2] Lesere som vil lære mer om hva Google Søketrender egentlig er, henvises til faktaboksen «Google Søketrender for nybegynnere».

kelser i regi av Statistisk Sentralbyrå. AKU-ledigheten er ikke aktuell som utfallsvariabel i denne artikkelen da det underliggende datagrunnlaget er mindre kompatibelt med volumet av søkedata fra Google.

Hva er forskjellen mellom nowcasting prognoser og de vi lager i «Utviklingen på arbeidsmarkedet»?

Det er kanskje ikke helt opplagt at det er behov for denne typen prognoser ettersom NAV allerede produserer arbeidsmarkedsprognoser, som i artikkelen «Utviklingen på arbeidsmarkedet» (UPA). Prognosemodellene som omtales her fyller imidlertid en helt annen rolle enn prognosene i UPA. For det første måler de to metodene arbeidsledigheten over svært forskjellige prognoseperioder. Med dette rammeverket ønsker vi å måle arbeidsledigheten i sanntid, mens i UPA er formålet å anslå arbeidsledigheten inntil tre år frem i tid. De store forskjellene i hvor langt fremover prognosen er ment å skue, innebærer også at prognosene beror på svært forskjellige metodologiske rammeverk. I denne artikkelen brukes for eksempel en algoritme til å velge en treffsikker nowcasting-modell, mens man i UPA bruker KVARTS-modellen til å anslå veksten i ledigheten. Videre bruker de to prognosemetodene ulike datagrunnlag. I UPA er man avhengig av data fra eksempelvis nasjonalregnskapet og internasjonale statistikkbyråer for å oppdatere KVARTS-modellen, slik at vi kan estimere ledigheten frem i tid. I denne artikkelen brukes derimot høyfrekvent stordata fra Google Trends for å estimere ledigheten. Metodene kan slik sett betraktes som komplementære, der det typisk er mulig å basere de langsiktige prognosene på oppdatert informasjon som kommer fra nowcasting-modeller. Spesielt nyttig er nowcasting-modellene for de langsiktige prognosemodellene dersom de kan forutse vendepunkter i konjunktursyklusen, noe som ofte er vanskelig å anslå ved hjelp av store strukturelle makromodeller, slik som blant annet KVARTS-modellen.

Den registrerte ledigheten oppdateres siste fredag i måneden. Google Søketrend-variablene måles i sanntid noe som innebærer at modeller basert på slike søketrender kan lage ledighetsprognoser minst en måned i forkant av publisering av bruttoledigheten. Mange arbeidstakere har for eksempel tre måneders oppsigelsestid og vil mest sannsynlig bruke Google til å fremskaffe informasjon om dagpenger, ledige stillinger o.l. før oppsigelsestiden løper ut og de registrerer seg som arbeidssøkere. Det er derfor sannsynlig at Søketrend-variablene fanger opp informasjon som ikke blir tilgjengelig gjennom ledighetsstatistikken før opptil tre måneder etter oppsigelsen ble registrert.

Den registrerte ledigheten er ikke egnet til å brukes som utfallsvariabel. Dette skyldes at variabelens statistiske egenskaper (som gjennomsnitt og standardavvik) ikke er konstante over tid[4]. For å ta høyde for dette ser vi heller på endringen[5] i bruttoledigheten. Utfallsvariabelen blir dermed målt som *prosentvis endring i den registrerte bruttoledigheten* fra måned til måned. Transformasjonen av variabelen er illustrert med før- og etter -bilde (se Figur 1 og 2).

Videre har jeg valgt å bruke det ujusterte målet på den registrerte ledigheten, som dermed ikke korrigerer for sesongmønster eller brudd i serien. Valget er basert på en langvarig økonometrisk tradisjon for å bruke rådata fremfor brudd- og sesongjusterte tall. En klar fordel med dette er at vi blant annet slipper å være prisgitt sesongjusteringsmetoder. Dette kan enkelt justeres for i etterkant. En utvilsom ulempe ved å la være å justere for sesongvariasjon er at modellene typisk vil inneholde Søketrend-variabler med store sesongkomponenter som korrelerer sterkt med sesongkomponentene i utfallsvariabelen. Dermed mister vi muligheten til å undersøke hvorvidt enkelte variabler er fundamentalt viktigere for å anslå svingningene i ledigheten enn andre variabler.

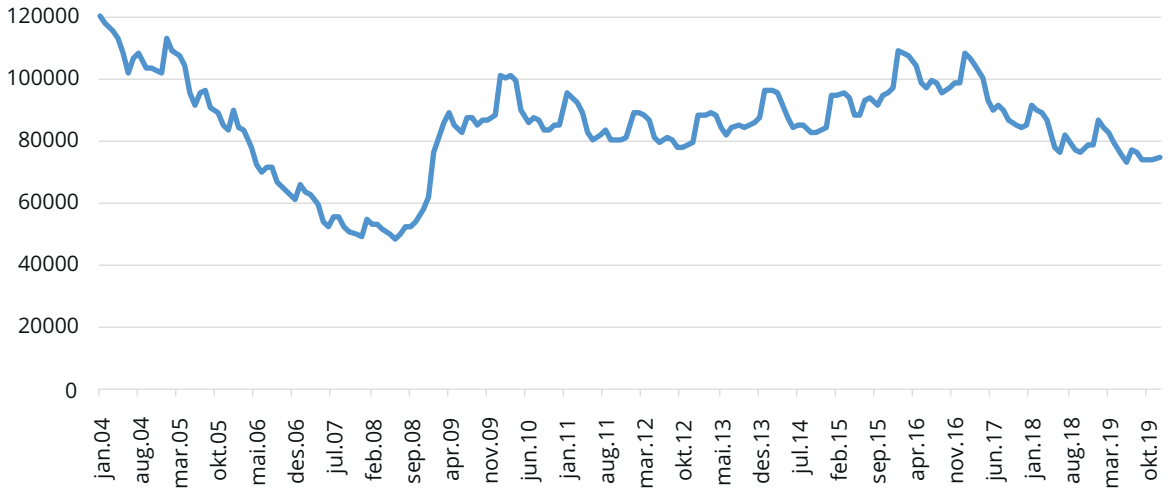
Metode for å velge typiske «Google»-søkeord

Et springende punkt i et opplegg hvor vi til slutt skal bestemme oss for en modell som skal predikere arbeidsledigheten i sanntid, gjelder utvelgelsen av Google Søketrend-variabler, eller mer bestemt: søkeord. Vilkårlig utvelgelse av søkeord kan nemlig føre til skjevhet i utvalget og vi blir nødt til å holde oss til et metodisk rammeverk for å kunne sikre et balansert utvalg. Som en praktisk løsning har jeg valgt å benytte Store Norske Leksikon (SNL) for å finne søkeord som er relatert til aktiviteten i arbeidsmarkedet. Ordene blir valgt fra kategorien «samfunn» som inneholder 11 nye underkategorier. Søkeordene velges fra underkategorien «arbeid og

[4] På fagspråket kalles slike variabler for «ikke-stasjonære» variabler. Den Augmenterte Dickey Fuller-testen blir brukt for å undersøke nullhypotesen om at variabelen ikke er stasjonær og hypotesen kan ikke forkastes, se Tabell V1.

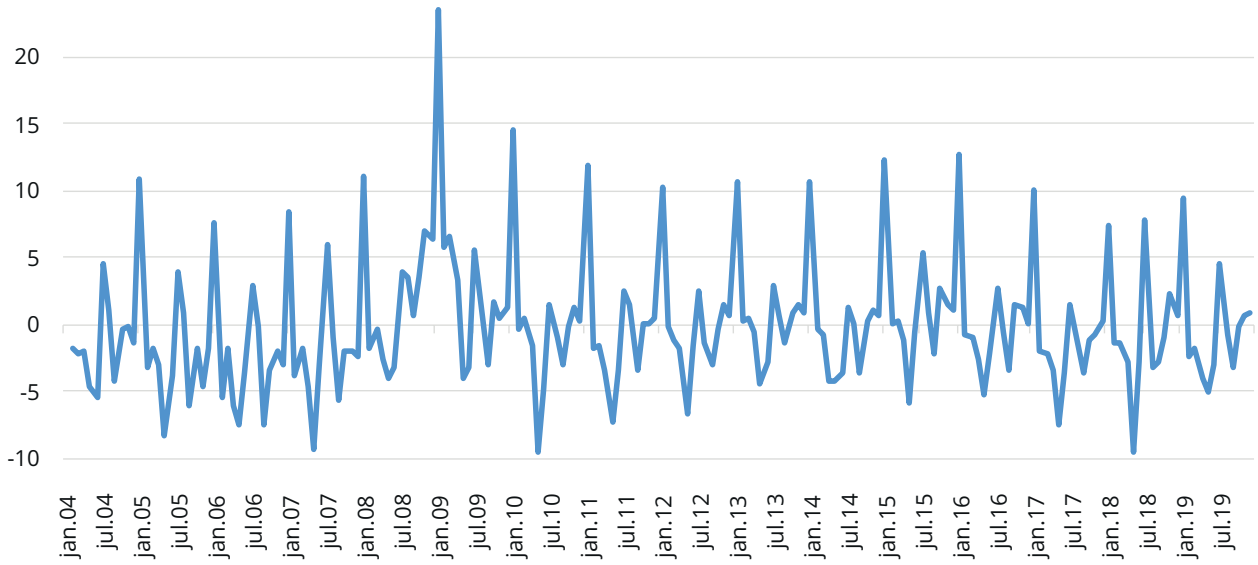
[5] Endringen defineres som differansen mellom variabelens verdi i dag og variabelens verdi i går (på engelsk omtaler man transformasjonen som «first difference»).

Figur 1. Antall «bruttoledige», Jan. 2004 – Des. 2019.



Kilde: **NAV**

Figur 2. Den prosentvise endringen i bruttoledigheten, Feb. 2004 – Des. 2019.



Kilde: **NAV**

velferd», som inneholder i overkant av 200 artikler. Artikkelnavn som eksempelvis [arbeidsledighet], [sosialhjelp] og [dagpenger] blir gjenstand for utvelgelse. En andel av artiklene blir ikke med i utvalget fordi det ikke er nok søkeinteresse i Googles søkemotor for det utvalgte ordet, slik som [lønnsmottager], eller fordi de anses som uhensiktsmessige å inkludere, slik som [levealdersjustering] og [blå resept].

Ettersom SNL ikke er tilpasset denne typen formål, legges det til noen subjektivt valgte ord, slik som eksempelvis [søknadstekst] og [CV]. I litteraturen finner vi få eksempler på at utvalget balanseres av søkeord som reflekterer tilbudet av arbeidskraft så vel som etterspørselen etter arbeidskraft. Dette blir til en viss grad tatt høyde for her ved å inkludere ord som [rekruttering hjelp], [bemanne], [hvordan ansette] o.l. Jeg bruker også et program som

observerer hva slags søkeord brukerne av www.nav.no og www.arbeidsplassen.no taster inn i Googles søkemotor for å nå frem til de to nettsidene[6]. De ti mest brukte søkeordene tastet inn i Googles søkemotor før man når de to nettsidene mellom 1. Oktober og 1. Januar inkluderes også i utvalget. Totalt gir dette et utvalg på 144 Google Søketrend-variabler som måles over 193 måneder fra januar 2004 til januar 2020.

Detaljer om variablene

I likhet med den registrerte ledigheten er flere av Google Søketrend-variablene også ikke-stasjonære. Alle variablene blir undersøkt ved hjelp av Augmenterte Dickey Fuller (ADF)-testen og ti av variablene blir målt på differanseform som følge av at heller ikke disse variablene har statistiske egenskaper som er konstante over tid.

Jeg korrigerer for sesongvariasjonen i datasettet ved å inkludere 12 dummy-variabler, en for hver måned. I løpet av perioden vi overvåker bruttoledigheten er datagrunnlaget «endret» to ganger som følge av to forskjellige brudd. Det første bruddet oppstod som følge av at en ny registreringsløsning for de som registrerer seg som arbeidssøkere på nav.no ble innført i slutten av 2018. Vi korrigerer for dette bruddet ved å inkludere en dummy-variabel som er lik 1 for november og desember 2018 og lik 0 resten av perioden. Den andre dummy-variabelen er lik 1 i mars 2010 og lik 0 resten av perioden og korrigerer for bruddet i statistikken som følge av store regelverksendringer, særlig knyttet til innføringen av AAP-ordningen. Til slutt velger jeg også å inkludere interaksjonsledd mellom de variablene som korrelerer med mer enn 80% med hverandre. Dette utvider datasettet med 19 variabler og inneholder nå 163 variabler til sammen.

Hvordan redusere antall variabler og sitte igjen med de viktigste

En utfordring med å anvende stordata til prognoseformål er at vi må finne en måte å velge noen få variabler av svært mange. I tillegg ønsker vi å spore opp de variablene som sammen har betydning for utfallet. Et viktig spørsmål blir hva slags kriterier eller metoder man skal bruke for å velge et hensiktsmessig antall variabler, blant et utvalg på totalt 163 variabler, til en endelig prognosemodell. Som tidligere nevnt kan vi benytte en automatisk søkealgoritme, Autometrics, til å trekke ut relevante variabler til den endelige modellen. Tidligere forsking (Epprecht mfl. (2019)) viser derimot at dersom antall variabler i datasettet reduseres vil Autometrics gjør en mer effektiv jobb i å hente frem den endelige prognosemodellen. Slike modeller er både «riktigere» og gir lavere prognosefeil, viser Monte Carlo simuleringer. En mulig forklaring på dette kan være at Autometrics bruker hypotesetesting for å finne frem til den endelige prognosemodellen. Når datasettet inneholder veldig mange variabler, fører dette til en akkumulasjon av type-I feil. Det betyr at maskinen forkaster nullhypoteser som i realiteten er sanne. For å forhindre dette tilføyer jeg noen kriterier i tillegg til de som allerede er bygd inn i maskinvaren. Dette bidrar til å redusere akkumulasjonen av type-I feil ettersom de nye kriteriene bistår i å kutte ned på antall variabler i datasettet. Samtidig håper jeg at akkurat disse utvalgte kriteriene er tilstrekkelige midler for å luke vekk variabler som forvirrer algoritmen eller oppfattes som støy (se avsnittet «valg av algoritme i jakten på den endelige modellen»)

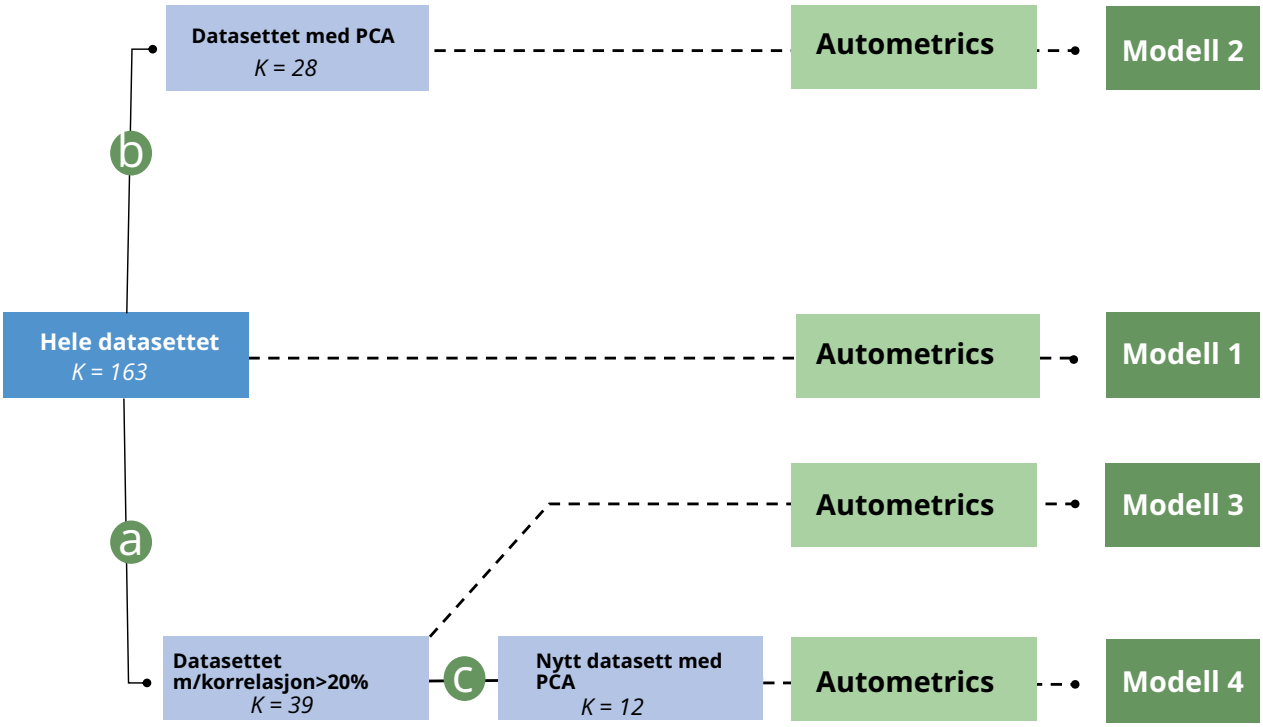
Det er altså av interesse å teste hvorvidt små datasett med færre variabler fungerer som et bedre grunnlag for å danne prognosemodeller enn større datasett med mange variabler. Ved bruk av systematiske kriterier, som korrelasjon med utfallsvariabel og prinsipal komponent analyse, lager jeg tre nye datasett som inneholder færre variabler enn det opprinnelige datasettet. Fordelen ved å bruke akkurat disse metodene for å dele variablene inn i nye datasett er at de sikrer at flere av de viktigste og mest signifikante variablene blir med videre i analysen, se Jensen (2019). Etter vi har delt datasettene inn i mindre datasett legges hvert datasett inn i Autometrics, som videre trekker ut kombinasjonen av variabler med størst forklaringskraft som sammen danner den endelige prognosemodellen. Figur 3 gir en forenklet framstilling av prosedyren foreslått av Jensen (2019).

To metoder for å koke ned datasettet

Det første kriteriet som blir benyttet for å redusere datasettets omfang er å se etter enkle parvise korrelasjoner mellom hver uavhengige Google Søketrend-variabel og utfallsvariabelen, som foreslått av Boivin mfl. (2006). Jeg bestemmer meg for å beholde kun de regressorene som har en korrelasjonskoeffisient som er over 20 prosent med bruttoledigheten. Denne gren-

[6] Jeg vil gjerne benytte anledningen til å takke Tobias Mcvey fra NAV designseksjon for introduksjonen og tillatelsen til å ta i bruk dette programmet.

Figur 3. Kart over datasettene.



Kartet leses fra venstre til høyre. Til venstre, har vi «Hele datasettet» med alle 144 Google Søketrend-variablene og 19 dummy-variabler og interaksjonsledd. Når vi beveger oss mot høyre, blir datasettet delt inn i tre nye datasett med færre forklaringsvariabler; «PCA», «Korrelasjon» og «Nytt datasett med PCA». I siste steg, bruker vi en automatisk algoritme (Autometrics) for å trekke ut de variablene, i hvert enkelt datasett, som skal inkluderes i den endelige prognosemodellen helt til høyre i fremstillingen.

Kilde: **NAV**

sen er delvis satt på bakgrunn av at det er svært få (kun 6) variabler som korrelerer med bruttoledigheten med mer enn 30 prosent, samtidig som det er svært mange (mer enn 80) variabler som korrelerer med ledigheten med mer enn 10 prosent. Ved å sette grensen på 20 prosent kvitter vi oss med nesten 100 variabler og står igjen med et datasett på 39 Søketrend-variabler. Et datasett i denne størrelsesordenen er passende da det er nok variabler til at det lønner seg å bruke Autometrics samtidig som det er et lavt nok antall variabler til at sannsynligheten for type-I feil tvinges ned. Prosedyren gjenspeiles i reiserute (a) til venstre i Figur 3.

Metoden jeg henviser til som «PCA» er en prinsipal komponent analyse der man forvandler et gitt antall (muligens) korrelerende variabler til et mindre antall ikke-korrelerende variabler, som vi kaller «prinsipale komponenter». Kort fortalt går metoden ut på å oppsummere et spredningsplott ved å lage en lineær kombinasjon av alle variablene i datasettet. Her vil de variablene som er viktigst i å forklare hele datasettet få høyere vekt i komponenten enn de variablene som, i større grad, kan betraktes som støy. PCA blir anvendt på et sett med indikator-variabler etterfulgt av en lineær regresjon. Deretter kjøres en ordinary least squares (OLS) regresjonsanalyse for å trekke ut de «prinsipale komponentene» som sammen er mest signifikante i å forklare arbeidsledigheten over tid. I denne regresjonsanalysen blir nesten 20 prosent av variasjonen i datasettet oppsummert i den første prinsipale komponenten. De neste to komponenter står for henholdsvis 9 og 7 prosent av variasjonen i datasettet. Til sammen forklarerer de første 28 prinsipale komponentene 75 prosent av variasjonen og jeg bestemmer meg for å inkludere kun disse i videre analyser.

Antallet variabler har gjennom denne prosedyren blitt redusert fra 163 til 28, se reiserute (b) i Figur 3.

Boivin og Ng (2006) var blant de første til å stille spørsmål ved om store datamengder alltid danner det beste grunnlaget for å trekke ut prinsipale komponenter. For å lage mer presise prognosemodeller, basert på stordata, foreslo de isteden å bruke mindre datasett som grunnlag for å trekke de prinsipale komponentene man senere skulle benytte for å lage prognosemodeller. Sammen med Jensen (2019) finner de at modeller som baseres på denne metoden gir signifikant lavere prognosefeil enn tradisjonelle referansemodeller. I tråd med hypotesen over velger jeg å kombinere de to metodene jeg introduserte overfor. Det betyr at jeg gjennomfører PCA på det datasettet som inneholder de 39 variablene som korrelerer med utfallsvariabelen med mer enn 20 prosent. I dette datasettet forklarer den første komponenten opptil en fjerdedel av variasjonen i datasettet mens de to neste komponentene står for henholdsvis 13 og 10 prosent av variasjonen i datasettet. Jeg velger å inkludere de første tolv komponentene ettersom disse også bidrar til å forklare opptil 75 prosent av variasjonen i datasettet. I Figur 3 er de metodologiske stegene visuelt fremstilt ved at vi først tar reiserute (a) etterfulgt av reiserute (c), før vi til slutt ender opp i å legge datasettet inn i Autometrics.

Valg av algoritme i jakten på den endelige modellen

Valg av algoritme er et viktig steg i prosessen med å søke etter den endelige prognosemodellen. Til tross for at det i teorien er mulig å bruke kjente metoder slik som OLS for å predikere arbeidsledigheten, er dette litt upraktisk da det er en tidkrevende øvelse å estimere om lag 2 kvadragintillion modeller. For å kunne skille ut de variablene som sammen gir en tilfredsstillende og presis prognose av arbeidsledigheten ved hjelp av enkle kriterier, kan et verktøy som «Autometrics» benyttes. Autometrics er en maskinvare basert på «General to Specific» (Gets) – prinsippet, der en empirisk analyse begynner i en generell statistisk modell som blir redusert i kompleksitet ved å eliminere ikke-signifikante variabler. Litt uformelt kan vi tenke på Autometrics som en lottomaskin der ballene som roterer rundt i maskinen representerer variablene i datasettet. I motsetning til en lottomaskin der ballene trekkes på tilfeldig grunnlag velger Autometrics ut en rekke med baller basert på enkle kriterier som sammen danner den endelige prognosemodellen. Et av kriteriene er eksempelvis at alle variablene i modellen skal ha et bestemt statistisk signifikansnivå. Dette nivået blir satt av forskeren.

Autometrics har en akilleshæl. Gjentatt hypotesetesting fører til en akkumulering av type-I feil. Det betyr at jo flere hypoteser som testes desto høyere er sannsynligheten for at noen av hypotesene som ikke skal forkastes, blir forkastet på basis av rene tilfeldigheter. En måte man enkelt kan moderere dette problemet på er å pålegge algoritmen et lavt signifikansnivå. Det betyr at variablene som blir inkludert i den endelige modellen må være signifikante på, for eksempel, et 0,1 prosent signifikansnivå. Dette bidrar til å fjerne irrelevante variabler som kun er inkludert i modellen basert på rene tilfeldigheter, men gjør også at relevante variabler som tilfeldigvis ikke er signifikante på dette nivået blir ekskludert. Med et lavt signifikansnivå vil dermed modellen inneholde færre irrelevante variabler, men dessverre også færre relevante variabler. De mer konkrete detaljene rundt modellspesifikasjonene går jeg nærmere inn på i neste avsnitt.

Spesifisering av modellene

De fire modellene blir *estimert* fra februar 2004 til desember 2015, mens *prognoseperioden* løper fra januar 2016 til januar 2020, i 48 perioder. Inn i Autometrics-maskineriet legges hver Søketrend-variabel med to tidsetterslep, syv interaksjonsledd, 12 dummy-variabler for hver måned i året og to dummy-variabler som i hensyntar bruddene i 2010 og 2018.

Ettersom de fire datasettene inkluderer såpass ulikt antall variabler bør jeg velge forskjellige signifikansnivåer i hver av de fire modellestimeringene. Hele datasettet (det lysegrå rektangelet helt til venstre i Figur 3) inneholder totalt 510 variabler noe som (antagelig) er for mange variabler for å kunne lage en god prognosemodell. Et høyt antall variabler taler for å ta i bruk et lavt eller strengt signifikansnivå for å unngå en for høy akkumulering av type-I feil. Jeg bestem-

Tabell 1: Modellspesifikasjoner.

	Datasett	Antall variabler[7] inn i Autometrics	Signifikansnivå
Modell 1 (M1)	Fullstendig datasett	510	0,01 %
Modell 2 (M2)	Korrelasjon > 20%	138	0,1 %
Modell 3 (M3)	PCA	81	1 %
Modell 4 (M4)	PCA ← Korrelasjon>20%	57	1 %

Datasett beskriver hvilket datasett modellen springer ut av. *Antall variabler i Autometrics* spesifiserer hvor mange variabler som blir lagt inn i Autometrics. *Signifikansnivå* spesifiserer hvilket signifikansnivå variablene minst må ha for at de skal inkluderes i den endelige modellen.

[7] Dette inkluderer to lags av hver variabel, alle dummy-variabler og interaksjonsledd.

mer derfor at alle variabler i den endelige modellen minst skal ha et signifikansnivå på 0,01 prosent. Med dette signifikansnivået viser Hendry og Nielsen (2007, Kap 19.3) at den endelige modellen vil inkludere 0,0001*510 = 0,051 irrelevante variabler og sannsynligheten for type-I feil er 4,9 prosent, som er relativt lavt.

Datasettet som inkluderer kun de variablene som korrelerer med mer enn 20 prosent med bruttoledigheten inneholder totalt 138 variabler. Dette er også et relativt høyt antall variabler. Med et signifikansnivå på 0,1 prosent vil den endelige modellen i gjennomsnitt inkludere 0,138 irrelevante variabler og sannsynligheten for type-I feil er i dette tilfellet rett under 10 prosent noe som er litt høyt, men akseptabelt. Modell 3 og 4 blir estimert med henholdsvis 81 og 57 variabler hver og jeg legger til grunn at hvis variablene er signifikante på et 1 prosent signifikansnivå, holder vi både akkumulasjonen av type-I feil og antallet irrelevante variabler nede på et hensiktsmessig nivå. Modellspesifikasjonene er oppsummert i Tabell 1 over.

Empiriske resultater:

Estimering av Google Søketrend-modellene

Modell 1 er gjengitt i Tabell 2 under og gjenspeiler et eksempel på en typisk Google Søketrend-modell. De fire modellene har to viktige fellesnevnere. For det første er alle modellene relativt komplekse og inkluderer et overaskende høyt antall forklaringsvariabler. Den minste modellen (M1) har 6 forklaringsvariabler mens den største modellen (M2) inkluderer hele 19 forklaringsvariabler. Videre viser resultatene fra modellestimeringen at alle modellene inkluderer minst 3 måneds-dummyer. Den eneste variabelen som går igjen i alle fire modellene er dummy-variabelen for januar. Dette er ikke spesielt overaskende ettersom den ujusterte ledigheten nesten utelukkende er høyest

Statistiske antagelser: Test av restledd og parameterstabilitet

Før man går i gang med å evaluere presisjonen i anslagene er det viktig å vurdere hvorvidt de underliggende statistiske antagelsene er støttet av datagrunnlaget. Mer konkret har jeg antatt at (i) restleddene ikke er feilspesifisert (ikke-korrelerende over tid og normale) og (ii) at parameterne i modellene er stabile over tid. Vi tester først antagelse (i) ved hjelp av enkle tester[8] og finner at alle prognosemodellene vurdert i denne analysen har normale og ikke-korrelerte restledd på et 5 prosent signifikansnivå. Jeg kan dermed konkludere at restleddene ser ut til å være riktig spesifisert og ikke utgjør noen trussel for inferensen i analysen videre.

Antagelsen (ii) om at parameterne i modellene er stabile blir testet ved hjelp av rekursiv estimering. Rekursiv estimering gir oss et visuelt inntrykk av hvor stabile parameterne i modellen er over tid ved å kontinuerlig endre estimeringsperioden. Modellen blir først estimert over en kort periode før perioden gradvis øker samtidig som man re-estimerer modellen. På denne måten gir estimeringsmetoden oss et kontinuerlig bilde av modellens stabilitet over en lengre tidsperiode. «Break-point Chow»-testen blir benyttet til å teste hypotesen om at modellene er stabile i perioden modellen blir estimert (februar 2004 til desember 2015). Testen viser at modellene er stabile over tid og at det mest sannsynlig ikke har forekommet noen strukturelle brudd i perioden modellene ble estimert over. Den visuelle fremstillingen av testen finner du i vedlegget, se Figur V1-V4.

[8] Slik som, for eksempel, White-testen som tester hvorvidt restleddene er homoskedastiske og RESET-testen som tester modellspesifikasjonene.

i januar hvert år noe som hovedsakelig skyldes at mange kontrakter går ut ved årsskiftet. Vi ser også at tre av fire modeller inkluderer dummy-variabler for mai og juli. Dummy-variabelen for mai er mest sannsynlig inkludert ettersom den ujusterte ledigheten er lavest i mai hvert år. Dette kan forklares med at mai er den måneden i året der det er færrest nyutdannete studenter som søker jobber mens i (juni og) juli derimot, begynner denne gruppen å registrere seg som ledige. Samtidig løper kontrakter også ofte ut på denne tiden av året, noe som gjerne kan forårsake en brå økning i ledigheten.

Mer generelt inneholder modell 1 og 2 en relativt bred portefølje av variabler som beskriver fluktuasjonene i arbeidsmarkedet. Søkeordene reflekterer typiske ord en *jobbsøker* ville tatt i bruk, slik som [jobbsøknad] og [manpower] og søkeord en *arbeidsledig* ville tatt i bruk, slik som [dagpenger] og [nav]. En av modellene inkluderer også Google Søketrend-variabelen [ansette], et søkeord som kan tenkes å være typisk for *arbeidsgivere* som vurderer å utvide virksomheten.

Tabell 2: Modell 1 (M1).

Variabler	Koeffisient	Standard feil
Jobber	0,04**	0,01
Jobbsøknad	-0,07**	0,01
Samordna opptak	-0,06**	0,01
Permittering lønn	0,07**	0,02
Nav arbeid	0,07**	0,01
Januar	9,4**	0,6
Mai	-4,3**	0,5
Juli	4,5**	0,5

Koeffisienten til hver respektive variabel som er inkludert i modellen av Autometrics er gjengitt i kolonnen «koeffisient». Heteroskedastiske robuste standardfeil er gitt til høyre for koeffisientene, som er statistisk signifikante på et *5 % og **1 % signifikansnivå ved bruk av en tosidig test. Modellen er estimert fra Feb. 2004 – Des. 2015.

Vurdering av prognosene

For å vurdere prognoseevnen til nowcasting-modellene deler jeg tidsserien i to perioder: en *treningsperiode* og en *testperiode*. Med *treningsperiode* mener jeg den perioden vi bruker for å estimere hver enkelt prognosemodell. Etter at en modell er estimert ønsker vi å teste hvor gode prognoser modellen lager. Under *testperioden* sammenligner vi hver månedlige arbeidsledighetsprognose med «fasiten» eller rettere sagt den realiserte arbeidsledigheten. På denne måten finner vi ut hvor mye og ofte modellen «bommer» i sine anslag på de kortsiktige variasjonene i ledigheten. Dessverre finner vi ikke ut om de modellene vi er interesserte i lager bedre prognoser enn allerede etablerte modeller. Derfor sammenligner vi modellenes anslag med anslagene til to referansemodeller. En autoregressiv modell med ett tidsetterslep (AR(1)) og en «random walk»-modell (RW) tjener som referansemodeller i denne analysen. AR(1)-modellen spesifiseres ved at utfallsvariabelen, den prosentvise endringen i arbeidsledigheten, blir lineært bestemt av sitt eget tidsetterslep og et konstantledd. RW-modellen referer til en modell der verdien av ledigheten i dag er lik verdien av ledigheten i går pluss et uforutsigbart restledd. Forskning viser at de to referansemodellene er vanskelige å slå av mer komplekse modeller, se blant annet D'Agostino, Giannone og Surico (2006).

Modellenes treffsikkerhet blir vurdert ut ifra kriteriet «Root Mean Square Error» (RMSE). RMSE måler modellens gjennomsnittlige prognosefeil, som er differansen mellom realisert arbeidsledighet og modellens prognose for arbeidsledigheten. Jo lavere RSME desto bedre er modellen til å predikere fremtidige fluktuasjoner i ledigheten. Testperioden løper fra januar 2016 til januar 2020. Dette gir meg 48 perioder som kan brukes til å sammenligne anslagene til hver enkelt modell med den realiserte veksten i ledigheten.

Tabell 3: Root Mean Squared Error (RMSE).

	Referansemodeller		Google Søketrend-modeller			
	AR(1)	RW	M1	M2	M3	M4
RMSE	4,23	4,24	3,2	2,53	2,67	1,64

Lav RMSE innebærer mer treffsikre anslag. Prognosene av den prosentvise endringen i bruttoledigheten er 1-stegsprognoser (nowcasts). Modellenes prognoseperiode løper fra Jan. 2016 – Jan. 2020.

Lavere prognosefeil enn referansemodellene

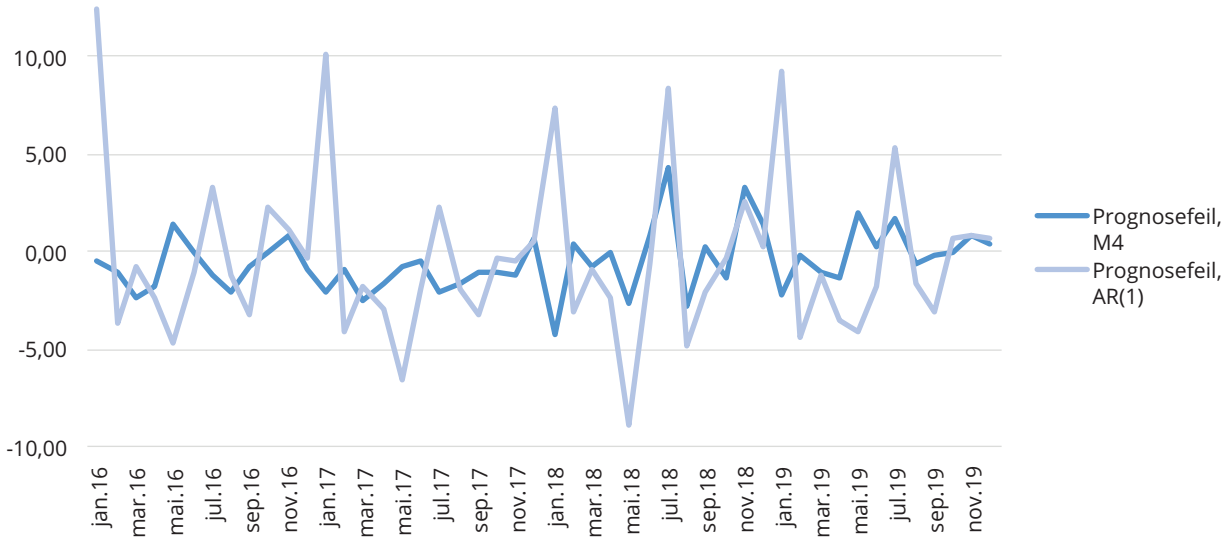
Samlet sett viser de empiriske funnene i Tabell 3 at alle Google Søketrend-modellene har lavere prognosefeil enn de to referansemodellene. Vi ser imidlertid at det er stor variasjon mellom de fire prognosemodellene. Blant annet kan vi lese av tabellen at modell 1 (M1) har høyere prognosefeil enn for eksempel modell 4 (M4). Modell 4 danner altså et bedre grunnlag for å lage kortsiktige prognoser for arbeidsledigheten enn modell 1. Dette resultatet styrkes av at metodene har samme resultat som i en tidligere analyse med en annen utfallsvariabel. Også i den tidligere analysen hadde modeller klart lavest prognosefeil når PCA og korrelasjon med utfallsvariabel ble brukt samlet som kriterier for å kutte antall variabler i datasettet.

Videre viser Tabell 3 at både modell 2 og modell 3 gir relativt gode anslag for ledigheten med en lav prognosefeil, på henholdsvis 2.53 og 2.67, sammenlignet med de to referansemodellene. Modell 2 (M2) springer ut av datasettet som inneholder kun de variablene som korrelerer med utfallsvariabelen med mer enn 20 prosent. Den lave prognosefeilen til modell 2 signaliserer at bruk av enkle kriterier (som korrelasjon) for å kutte antall variabler i datasettet, virker å ha en god effekt på modellens prediksjonskraft. Modell 3 (M3) inneholder prinsipale komponenter som derimot er dannet med hele datasettet, bestående av 144 variabler, som grunnlag. Sammenlignet med modell 4 som også er basert på prinsipal komponent analyse, gir modell 3 derimot noe upresise anslag. I tråd med funnene til Boivin og Ng (2006) ser vi altså at prinsipale komponenter som er basert på få (potensielt signaltunge) variabler legger forutsetningen for å produsere mer presise prognosemodeller enn modeller som baserer seg på prinsipale komponenter som er trukket ut av store og til dels støyfulle datasett.

Figur 4 viser prognosefeilen til to modeller grafisk over tid. Her representerer den mørkeblå grafen avviket mellom den realiserte veksten i arbeidsledigheten og prognosen til modell 4. Grafen svinger balansert rundt x-aksen uten store og uregelmessige hopp. Sammenligner vi dette med avviket mellom realisert arbeidsledighet og prognosen til AR(1)-prosessen viser figuren større og mer uregelmessige svingninger rundt x-aksen. AR(1)-modellens prognoser bommer med andre ord klart kraftigere og oftere enn prognosene laget av modell 4.

Dessuten kan vi si at alle de fire Google Søketrend-modellene presenterer signifikant lavere prognosefeil i sine anslag for de kortsiktige fluktuasjonene i ledigheten sammenlignet med de to referansemodellene. Dette er analysens viktigste funn fordi det understreker både at modellene i gjennomsnitt har lavere prog-

Figur 4. 1-stegprognoser (nowcasts) av Modell 4 (M4) og AR(1)-prosessen..



Et lite avvik mellom den realiserte ledigheten og modellens prognose innebærer lav prognosefeil og verdier nær x-aksen. Modellenes prognoseperiode løper fra Jan. 2016 – Jan. 2020.

nosefeil, men viser også at Google Søketrend-baserte modeller utkonkurrerer hyppig brukte referansemodeller på et signifikant nivå. Usikkerheten rundt disse funnene er heller ikke spesielt høy ettersom vi har såpass mange observasjoner av avviket mellom realisert arbeidsledighet og anslagene til hver Google Søketrend-modell. For å kunne avgjøre hvorvidt modellene har signifikant lavere prediksjonsfeil enn de to referansemodellene har vi benyttet DM-testen (se faktaboks).

Er Google Søketrend-modellene signifikant bedre til å anslå ledigheten sammenlignet med to referansemodeller?

Vi finner at alle Google Søketrend-modellene utkonkurrerer de to referansemodellene på et 10 prosent signifikansnivå. Mer spesifikt viser resultatene (i Tabell 4) at prognosefeilen til Modell 4 (M4) er signifikant lavere enn prognosefeilen produsert av de to referansemodellene på et 1 prosent signifikansnivå. Videre finner vi at prognosemodellene M2 og M3 utkonkurrerer begge referansemodellene på et 5 prosent signifikansnivå.

Vi må benytte oss av en test for å kunne vurdere hvorvidt én prognosemodell gir signifikant lavere prognosefeil enn en annen modell. Denne testen kalles Diebold-Mariano (DM)–testen og brukes for å måle den komparative presisjonskraften til hver enkelt Google Søketrend-modell sammenlignet med de to referansemodellene, som anbefalt av Clements (2005, pp.12–14). Med DM-testen kan vi regne ut hvorvidt to modeller er signifikant ulike fra hverandre. Dette er tilfellet dersom teststatistikken er større enn den kritiske verdien i normalfordelingen. Teststatistikkene er oppsummert i Tabell 4 under. Her måles hvorvidt hver respektive Google Søketrend-modell har signifikant lavere prognosefeil enn de to referansemodellene.

Tabell 4: Test for komparativ presisjon.

	M1	M2	M3	M4
AR(1)	-1,79	-2,55	-2,40	-3,52
	(0,074)	(0,011)	(0,016)	(0,0004)
RW	-1,73	-2,51	-2,31	-3,37
	(0,084)	(0,012)	(0,021)	(0,0008)

På øverste rad er teststatistikken gjennomført med AR(1)-modellen som referanse og på nederste rad med Random walk-modellen som referanse. I kolonnene rapporteres teststatistikken. Negativ teststatistikk indikerer at den gjennomsnittlige prognosefeilen til hver respektive Google Søketrend-modell er lavere enn referansemodellen. I parentes under teststatistikken rapporteres testens p-verdi.

Avsluttende kommentarer

De empiriske resultatene jeg har presentert over viser at prognosemodeller med Google Søketrend-variabler som datagrunnlag, gir treffsikre anslag på de kortsiktige fluktuasjonene i arbeidsledigheten. Mer konkret ser vi at tre av fire søketrend-modeller utkonkurrerer to tradisjonelle referansemodeller på et 5 prosent signifikansnivå. Felles for de tre modellene er at de beror på mengder med data som har vært gjenstand for utvelgelse basert på bestemte kriterier og metoder. Funnene viser også at rammeverket som er tatt i bruk her, er stabilt på tvers av utfallsvariabler. Dette kan komme til nytte for andre som skal gjøre lignende analyser med bruk av stordata, der det ennå ikke er etablert en klar metode for utvelgelse av data, algoritmer og modeller.

Begrensningene ved å bruke Google Søketrender

I avsnittene under vil jeg presentere noen betraktninger om begrensningene ved å bruke Google Søketrender som datagrunnlag. De tre viktigste begrensningene blir oppsummert under.

Google Søketrend-variablene er trolig korrelert med alder

For det første er søketrend-variablene trolig korrelert med alder, se for eksempel Dommes (2010). Denne hypotesen har opphav i at utbredt internettbruk er et relativt nytt fenomen, noe som fører til at i alle fall begynnelsen av tidsseriene vil være preget av en overvekt av unge søkemotorbrukere. Dessuten er det tenkelig at terskelen for å benytte seg av søkemotorer for å finne frem til en nettside er langt lavere for yngre brukere enn for eldre. Det er altså mulig at eldre brukere i større grad henvender seg til Google for å finne frem til nettsider de ikke har besøkt tidligere, og som dermed ikke ligger lagret i hukommelsen til brukerens nettleser. Dersom brukeren har vært inne på nettsiden før, dukker den som regel opp som forslag i utforsker-feltet og brukeren vil kunne gå direkte til nettsiden istedenfor å ta omveien om Google. Yngre brukere antas å være mer vant til å ta veien innom Google før de når frem til det tiltenkte nettstedet enn det den eldre generasjonen er. Dette kan, for eksempel, komme av at yngre brukere er mer komfortable med eller mer vant til Googles grensesnitt eller ulik oppfatning av hvilken fremgangsmåte som er mest tidseffektiv.

Teorien får til en viss grad støtte fra en spørreundersøkelse gjennomført av Princeton survey Research Associates, der det fremheves at yngre brukere oftere og med høyere sannsynlighet henvender seg til Google enn det eldre brukere gjør. Dette indikerer at det kan være større forskjeller i adferden til yngre og eldre internettbrukere. Dette er særlig problematisk fordi det kan føre til at Google Søketrend-variablene ikke er representative for hele populasjonen, og dermed vil modellenes prediksjonsfeil sannsynligvis være systematisk skjev. Når det først og fremst er unge internettbrukere som benytter seg av Google, vil modellene kanskje kunne predikere en økning i ledigheten blant unge, men vil ikke nødvendigvis klare å predikere det som faktisk er utfallsvariabelen, nemlig ledigheten på tvers av alder. En større amerikansk studie viste derimot at populasjonen av søkemotorbrukere ga et relativt representativt speilbilde av den amerikanske populasjonen (Weber mfl. 2010). Lignende studier finnes ikke for Norge, men det er tenkelig at de norske internettbrukerne ikke skiller seg spesielt fra amerikanske brukere og at problemet med skjeve Google Søketrend-variabler ikke gjelder i særlig grad for denne analysen.

Google Søketrend-variablene er trolig korrelert med internetterfaring

For det andre er det også mulig at bruken av søkemotorer er korrelert med (internett)erfaring. For eksempel kan det tenkes at jo mer erfaring du har som arbeidsledig, permittert, eller på annen måte midlertidig utenfor arbeidslivet, jo større erfaring har du også med å orientere deg i arbeidsmarkedstemaer på internett. For eksempel vil en som tidligere har vært arbeidsledig trolig vite mer om hvordan man skal registrere seg eller sende inn dagpengekrav enn en som er «nybegynner». Det er nærliggende å tro at de som søker dagpenger for første gang vil bruke lengre tid på Google for å finne ut hvor nærmeste NAV-kontor ligger eller hvordan man søker etter dagpenger, enn det en mer «erfaren arbeidsledig» trenger. Dette kan bidra til at Google Søketrend-variablene i større grad reflekterer søkeadferden til dem som ikke har vært arbeidsledige før, enn det de gjør for brukere som er mer erfarne. Hvis dette er tilfellet, er det en risiko for at variablene ikke fanger opp søkeadferden til dem som blir hyppig ledige. Dette kan typisk gjelde arbeidere innen yrker som er ekstra eksponert for sesongledighet eller konjunktursvingninger. Hvis dette er tilfellet vil vi få et datagrunnlag som ikke gir grunnlag for å fange opp endringer og vendepunkter i ledigheten. Denne bekymringen støttes imidlertid ikke i litteraturen. Tvert imot er det belegg for at Google Søketrender klarer å fange opp vendepunkter i konjunktursyklusen. Blant annet fant Ellingsen (2017) at modeller som brukte Google Søketrender til å predikere den registrerte ledigheten i Norge under finanskrisen ved hjelp av nowcasting, utkonkurrerte standard referansemodeller, slik som AR(1)-modellen.

Mye støy gjør utvelgelse av søketrend-variablene viktig

For det tredje kan selve bruken av søkemotoren også være et diskusjonstema. McLaren mfl. (2011) peker på at forskjellige brukere som er interesserte i det samme emnet, kan angi vidt forskjellige søkeord. Samtidig kan brukere med svært forskjellige intensjoner med sitt søk, angi svært like søkeord. For eksempel angir vi mange søkeord av ren nysgjerrighet. Slik søkeatferd kan resultere i at noen Google Søketrend-variabler inneholder signifikante mengder med støy. Dette kan gjøre det vanskelig å lage presise prognoser. Mye støy innebærer blant annet at variablene ikke nødvendigvis fanger opp det vi tror de gjør. Kunnskapen vi har om at søkemotorer brukes på denne måten er noe av bakgrunnen for at vi også velger å ta i bruk kriterier og metoder for å sortere vekk de variablene som er mindre viktige for å forklare utviklingen i arbeidsledigheten. For eksempel er [trygd] en variabel som har blitt ekskludert i et av datasettene fordi den har for lav korrelasjon med utfallsvariabelen. Dette er et søkeord man kan se for seg at like gjerne blir angitt basert på nysgjerrighet som at noen er interessert i å lære mer om hvilken type trygd man kan gjøre krav på. Når intensjonen bak et søk i Googles søkemotor ikke er tydelig nok kan det gjøre at variabelens svingninger i større grad er drevet av støy enn av fundamentale forhold. Slike variabler blir derfor vanskeligere å bruke for å anslå fremtidige svingninger i arbeidsledigheten. Til tross for at det er argumenter for at variablene kan inneholde mye støy, viser de empiriske resultatene i denne artikkelen at datagrunnlaget fungerer svært godt til å lage kortsiktige prognoser for arbeidsledigheten. Likevel er det tydelig at prognosene blir langt mer presise når vi eks-

kluderer noen variabler basert på at de ikke når opp til ulike sorteringskriterier, slik som lav korrelasjon med utfallsvariabelen.

Nytteverdien for NAV

Hyppig oppdatert og treffsikker informasjon danner grunnlaget for gode økonomiske og organisatoriske beslutninger og politikk. Dette er kanskje spesielt viktig i situasjoner hvor endringer og svingninger skjer brått. Dagpenger er mest sannsynlig den av NAVs ytelser som svinger sterkest. Brå og uventede vendinger i arbeidsledigheten medfører at også veksten i dagpengekostnader blir relativt uforutsigbar. Det er særlig i slike situasjoner at nowcasting-modeller, som tar temperaturen på arbeidsmarkedet i (tilnærmet) sanntid, er nyttige. I litteraturen er det relativt bred enighet om at Google Søketrender fungerer spesielt godt til å lage anslag på vendepunkter i konjunktursyklusen. Dette er testet både for konjunkturutviklingen under finanskrisen og under den oljedrevne nedgangskonjunkturen, se for eksempel Ellingsen (2017) og Jensen (2019). Bedre og mer oppdatert informasjonsgrunnlag om en av NAVs nøkkelvariabler er også viktig styringsinformasjon for NAV som organisasjon. Det kan bidra til at NAV i større grad makter å fange opp vendepunkter i konjunktursyklusen tidligere og dermed kan sette inn ekstra ressurser til, for eksempel, raskere behandling av dagpengesaker og større veiledningskapasitet overfor arbeidssøkerne. Metodene som introduseres i denne artikkelen er også enkle å tilpasse til lokale forhold, som gjør det mulig å se om ledigheten i visse regioner kommer til å øke i løpet av de neste månedene. Dersom oljeprisen igjen faller mye og holder seg lav over lengre tid, kan vi legge til søketrend-variabler, slik som [jobb Stavanger kommune], [jobb ingeniør] eller [jobb Rogaland] i utvalget, for å se om bruken av slike søkeord har økt.

Litteraturliste

Anvik, Christian og Kristoffer Gjelstad (2010). «Just Google it»: Forecasting Norwegian unemployment figures with web queries. *Working papers 11. Centre for Research in Economics and Management (CREAM), BI Norwegian Business School.*

Banbura, Marta, Giannone, Domenico, Modugno, Michele og Reichlin, Lucrezia (2013), «Nowcasting and the real-time data flow», *Working Paper Series* 1564. European Central Bank.

Boivin, J. og Ng, S. (2006), «Are more data always better for factor analysis?» *Journal of Econometrics*, 132(1), 169–194.

Carrière-Swallow, Yan og Felipe Labbé (2013), «Nowcasting with Google Trends in an Emerging Market: Nowcasting with Google Trends in an Emerging Market». *Journal of Forecasting* 32, nr. 4: 289–98.

Choi, H. og H. Varian (2009), «Predicting initial claims for unemployment benefits», Google Inc, 15.

Ellingsen, J. (2017). «Let's google it. Can google search indices nowcast Norwegian retail sales and unemployment rate?» (Master oppgave, UiO).

Epprecht, C, D. Veiga, J. Correa da Rosa (2019) «Variable selection and forecasting via automated methods for linear models: LASSO/adaLASSO and Autometrics», *Communications in Statistics-Simulation and Computation*, 1–20.»

Gunn III, John F., og David Lester. (2013) «Using google searches on the internet to monitor suicidal behaviour.» *Journal of affective disorders* 148, no. 2–3: 411–412.

Hendry, D. F. & Nielsen, B. (2007). «Econometric modelling: A likelihood approach», *Princeton University Press*.

Jensen, M. (2019), «In search of the present. An indicator comparison: Nowcasting quarterly GDP using Google search data and monthly accounts of GDP» (Master oppgave, UiO).

Kassraie, Parnian, Alireza Modirshanechi, og Hamid K. Aghajan (2017) «Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online Polls.» *In DATA*, pp. 363–370.

Lineman, Maurice, Yuno Do, Ji Yoon Kim og Gea-Jae Joo. (2015) «Talking about climate change and global warming.» *PloS one* 10, no. 9.

McLaren, Nick, og Rachana Shanbhogue (2011) «Using Internet Search Data as Economic Indicators», *Electronic Journal*, Q2.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., og Weinstein, R. A. (2008), «Using internet searches for influenza surveillance», *Clinical infectious diseases*, 47(11), 1443–1448.

Purcell, K, Rainie, L og Brenner, J. (2012), «Search engine use». Pew internet and American life project.

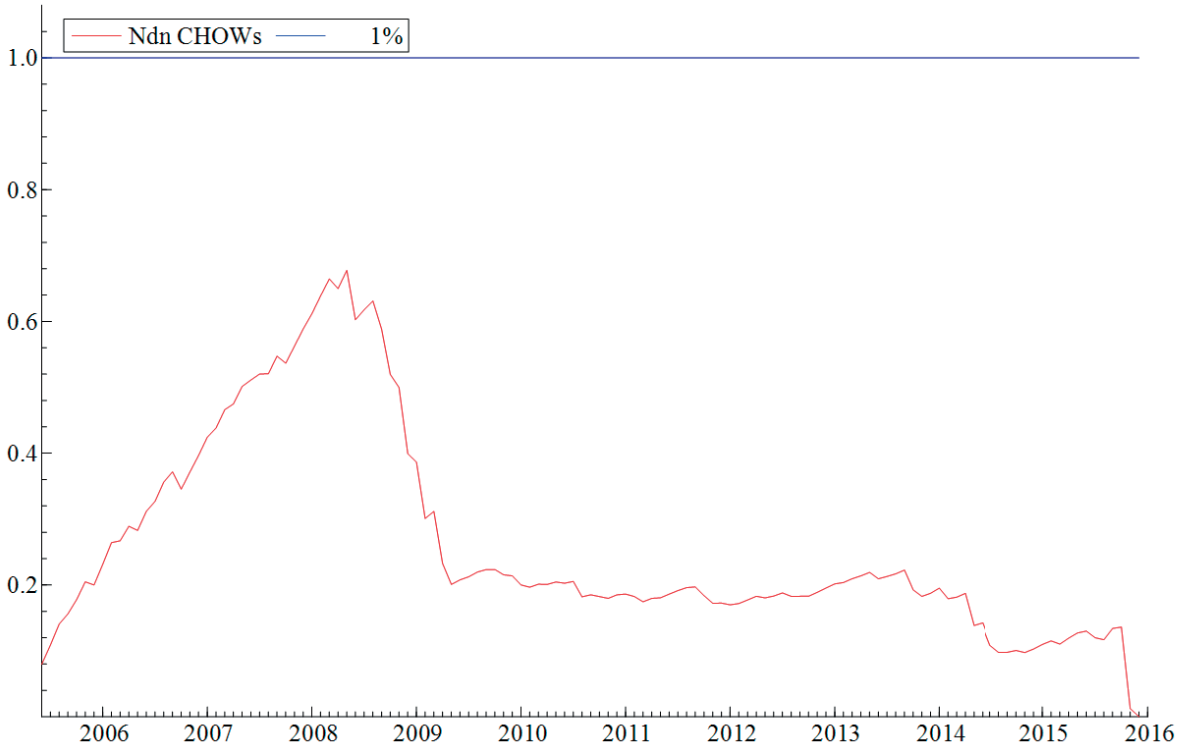
Thorsrud, Leif Anders (2018) «Words Are the New Numbers: A Newsy Coincident Index of Business Cycles», *Journal of Business & Economic Statistics*, 1–17.

Weber, I. og Castillo, C., (2010), «The demographics of web search». *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 523–530).

Wu, L. og Brynjolfsson, E. (2015) «The future of prediction: How google searches foreshadow housing prices and sales», *In Economic analysis of the digital economy* (pp. 89–118). University of Chicago Press.

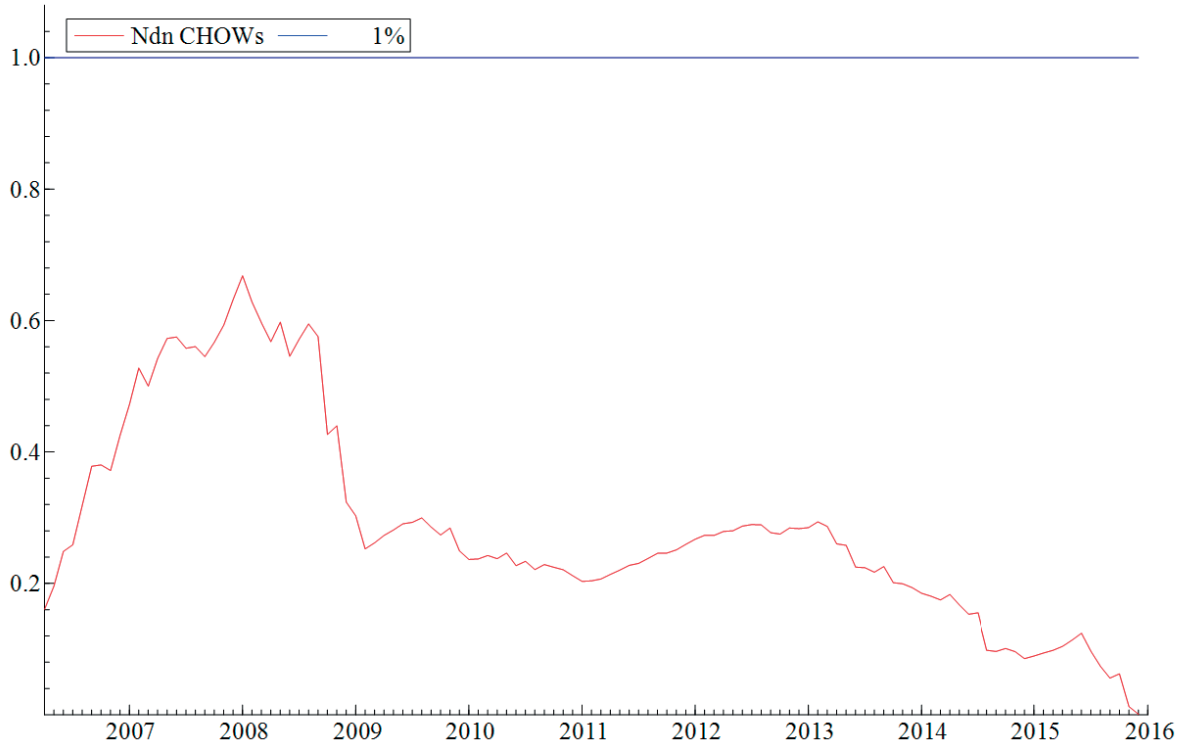
Vedlegg

Figur V1. Break-point Chow test av modell 1, M1.



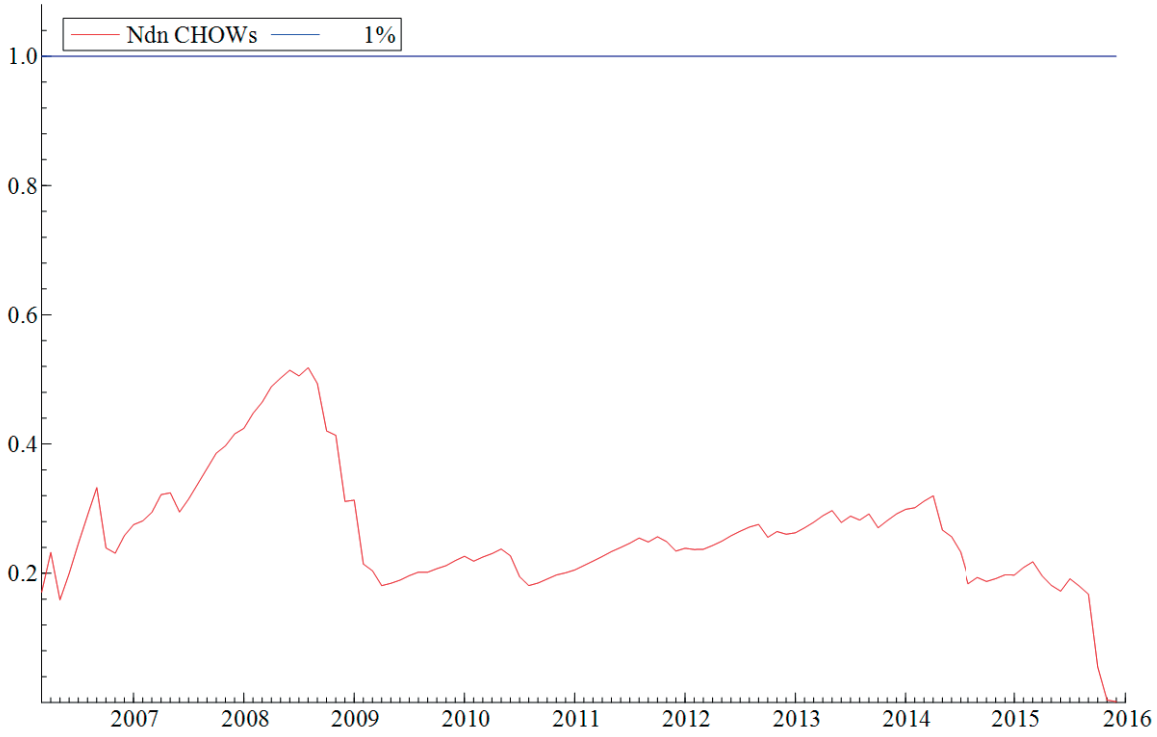
Modellen er estimert ved hjelp av rekursiv estimering fra Jun. 2005 – Des. 2015. Nullhypotesen om at alle parameterne samlet er konstante på et 1 prosent signifikansnivå, gjenspeiles i den blå grafen på toppen av figuren. Den røde grafen gjengir modellens numeriske variasjon etter hvert som tiden tiltar.

Figur V2. Break-point Chow test av modell 2, M2.



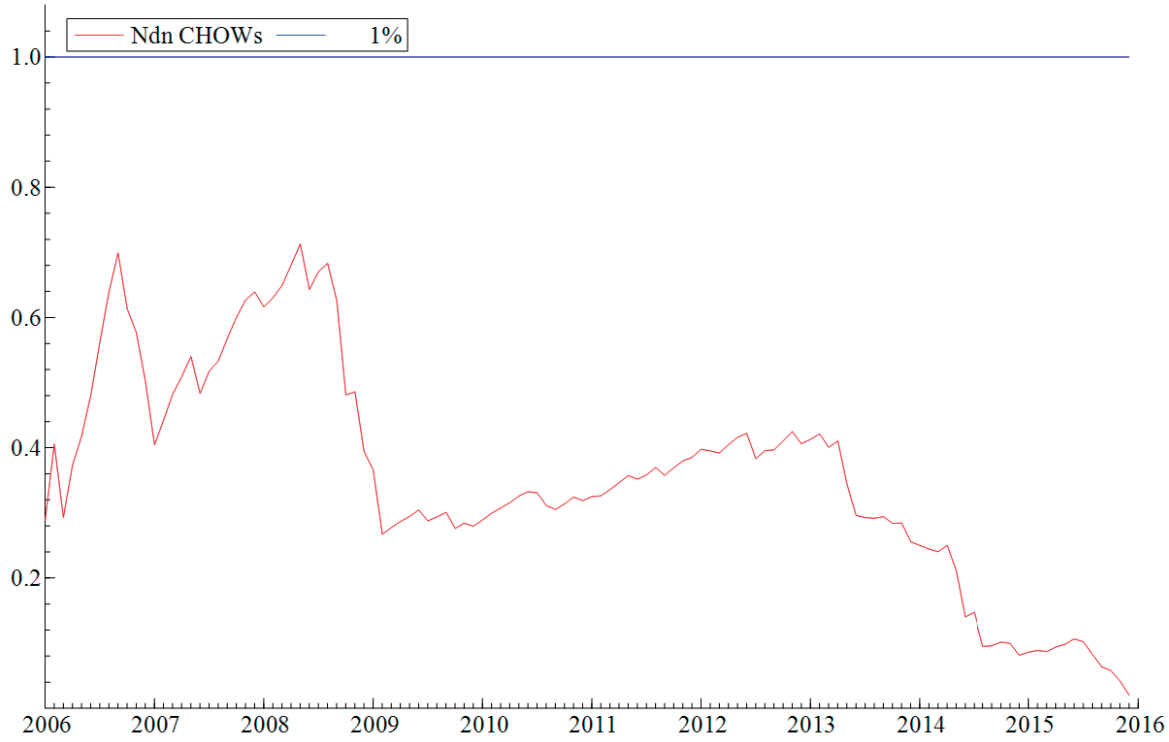
Modellen er estimert ved hjelp av rekursiv estimering fra Apr. 2006 – Des 2015. Nullhypotesen om at alle parameterne samlet er konstante på et 1 prosent signifikansnivå, gjenspeiles i den blå grafen på toppen av figuren. Den røde grafen gjengir modellens numeriske variasjon etter hvert som tiden tiltar.

Figur V3. Break-point Chow test av modell 3, M3.



Modellen er estimert ved hjelp av rekursiv estimering fra Mar. 2006 – Des 2015. Nullhypotesen om at alle parameterne samlet er konstante på et 1 prosent signifikansnivå, gjenspeiles i den blå grafen på toppen av figuren. Den røde grafen gjengir modellens numeriske variasjon etter hvert som tiden tiltar.

Figur V4. Break-point Chow test av modell 4, M4.



Modellen er estimert ved hjelp av rekursiv estimering fra Jan. 2006 – Des 2015. Nullhypotesen om at alle parameterne samlet er konstante på et 1 prosent signifikansnivå, gjenspeiles i den blå grafen på toppen av figuren. Den røde grafen gjengir modellens numeriske variasjon etter hvert som tiden tiltar.

Tabell V1: Den augmenterte Dickey-Fuller (ADF) testen.

D-lag	t-ADF	t-statistikk
1	-9.06**	0.495
1	-9.06**	0.495
0	-11.41**	

Nullhypotesen er at den prosentvise endring i bruttoledigheten ikke er stasjonær. Stjernene indikerer hvorvidt vi kan forkaste nullhypotesen på et *5% eller **1% signifikansnivå. Modellen er estimert fra Feb. 2004 – Jan. 2019.