

Henrik Bråten

Befolkningsdivergens i Norge

En empirisk studie av perioden 1920-2017

Masteroppgave i samfunnsøkonomi

Veileder: Hildegunn Ekroll Stokke

Trondheim, juni 2019

Norges teknisk-naturvitenskapelige universitet

Fakultet for Økonomi

Institutt for Samfunnsøkonomi

NTNU

Norges teknisk-naturvitenskapelige universitet

Masteroppgave i samfunnsøkonomi

Fakultet for Økonomi
Institutt for Samfunnsøkonomi

Sammendrag

I denne oppgaven studeres forskjellene i regional befolkningsvekst i Norge. Oppgaven tar utgangspunkt i et paneldatasett for norske regioner i perioden 1920 til 2017. Resultatene viser både absolutt og betinget divergens i befolkningsnivået for regionene. Dersom man fokuserer på kortere perioder viser analysen tegn til at befolkningsdivergensen har vært tiltagende med tiden. Markov-kjedeanalysen gir en ergodisk fordeling som viser utvikling mot at noen få regioner ender opp med den største delen av befolkningen. Dette sammen med at Kernel-tetthetene viser bevegelse mot flere små regioner, forsterker inntrykket av befolkningsdivergens mellom regionene. Ved å studere dynamikken for ulike nivåer av befolkningen, ved hjelp av Spline-regresjon, virker divergensen mellom regionene å være drevet av divergens mellom regioner som er mellomstore i befolkningsnivå.

Abstract

The objective of this thesis is to analyse the disparities in regional population growth in Norway. To this end, a panel data set for Norwegian regions during the period 1920 to 2017 are used. The results show both absolute and conditional divergence in the population level for the regions. When focusing on shorter periods, the analyses show signs that population divergence has been increasing with time. The Markov chain analysis offers an ergodic distribution that shows development towards a few number of regions ending up with the largest part of the population. This, together with the kernel densities showing movement towards smaller regions, strengthen the impression of population divergence between regions. Dividing regions based on their population size and analysing the dynamics using spline regression, shows that divergence between the regions are driven by divergence between medium sized regions.

Forord

Denne masteroppgaven er en del av min mastergrad i samfunnsøkonomi ved NTNU. Jeg ønsker å rette en stor takk til min veileder Hildegunn Ekroll Stokke for verdifull og utrolig god veiledning, hun har alltid hatt en åpen dør både i tide og utide. Videre rettes en stor takk til Helén Stenshol og Karsten Mykleset for gjennomlesning av oppgaven, til tross for en antatt lav interesse for innholdet. Pareto fotballklubb og boms og president gjengen fortjener også en hyllest for å ha gjort oppholdet i Trondheim til en fryktelig minnerik tid.

Dessverre kan ingen andre enn meg selv klandres for eventuelle feil eller mangler i denne oppgaven.

Grinder, Mai 2019

Henrik Bråten

Innhold

1	Introduksjon	1
1.1	Problemstilling	1
1.2	Disposisjon	2
2	Litteratur	3
2.1	Teoretisk rammeverk	3
2.2	Tidligere empiriske studier	6
3	Metode	9
3.1	Barroregresjoner	9
3.2	Økonometriske utfordringer	10
3.2.1	Utelatte variabler	10
3.2.2	Simultanitet	11
3.2.3	Målefeil	11
3.2.4	Problemer knyttet til inferens	12
3.3	Spline-regresjoner	12
3.4	Kernel-tettheter	13
3.5	Lokal polynomglatting	14
3.6	Markov-kjeder	14
3.7	Q-og LR-tester	16
4	Data	19
4.1	Avhengig variabel	19
4.2	Forklaringsvariabel	20
4.3	Kontrollvariabler	20
4.3.1	Areal	21
4.3.2	Andel fjell	21
4.3.3	Klima	21
4.3.4	Universitet	21
4.3.5	Tollhavn	22
4.4	Deskriptiv statistikk	22
5	Analyse	25

5.1	Befolkningsutvikling	25
5.2	Barr regresjoner	26
5.3	Robusthet av tidsperioder	30
5.4	Robusthet region	31
5.5	Lokal polynomglatting og Spline-regresjon	32
5.6	Kernel-tettheter	33
5.7	Markov-kjeder for etterkrigstida (1946-2016)	34
5.7.1	Regioninndelingens betydning	37
5.7.2	Forskjeller over tid	38
6	Konklusjon	43
	Referanser	45
A	Tabeller	49

1 Introduksjon

Sentralisering er et stadig aktuelt ord i nyhetsbildet. Til tross for en aktiv distriktspolitikk har man sett at sentraliseringen er sterk i Norge. Livet i sentrale og rurale strøk arter seg forskjellig, og mange mennesker virker å være mer tiltrukket av livstilen i urbane strøk. Høyere lønninger som følge av høyere økonomisk vekst i sentrale strøk kan samtidig bidra til jobbmotivert flytting. Sentraliseringen har negative virkninger på rurale områder, færre skattebetaler som igjen gir lavere inntekter og dermed dårligere tilbud for kommunene i rurale områder, er noe som igjen kan bidra til en selvforsterkende effekt.

I denne oppgaven vil det bli gjort rede for en teori som kan forklare divergens mellom regioners befolkningsnivå. Hovedfokuset vil være å se om det faktisk har vært divergens mellom befolkningsnivå på regionalt nivå og hvordan denne i så fall har utspilt seg. Dette gjøres ved å analysere tall for økonomiske regioner i Norge. Regionene er delt opp i henhold til Hustoft et al. (1999). Det er ingen tvil om at enkelte kommuner har store utfordringer knyttet til synkende befolkningsnivå. Kan det tenkes at det samme problemet er gjeldene for hele regioner? Resultatene i denne oppgaven kan være av interesse for den pågående debatten om kommunesammenslåing. Kanskje er politikerne nødt til å tenke større enheter for å være godt nok rustet til å håndtere utfordringene som synkende befolkningsnivå gir i forhold kommunale tilbud innen eksempelvis helse og skole.

1.1 Problemstilling

I denne analysen blir det undersøkt for divergens/konvergens i befolkningsnivå mellom regioner i perioden 1920 til 2017. Analysen utvides også til å se på regionene sin dynamikk seg i mellom i forhold til befolkningsnivå. Den konkrete problemstillingen for oppgaven er: “Har Norge vært preget av divergens i regionalt befolkningsnivå i perioden 1920-2017”. Videre undersøkes det også om regional befolkningsvekst kan forklares av naturlige egenskaper.

1.2 Disposisjon

Oppgaven er strukturert i fem påfølgende kapitler. Kapittel 2 tar for seg litteratur, her forklares det teoretiske rammeverket oppgaven støtter seg på, samt tidligere studier av befolkningsdynamikk og hvilke funn disse studiene har gjort. I kapittel 3 redegjøres det for hvilken empirisk strategi som benyttes i oppgaven og hvilke problemer det er sannsynlig å møte. Datamaterialet analysen baserer seg på vil bli redegjort for i kapittel 4. I dette kapitlet presenteres og argumenteres det også for forklaringsvariablene som inkluderes i analysene. Selve analysen og diskusjonen rundt denne gjøres i kapittel 5. Her introduseres også en rekke robusthetsanalyser. En kort konklusjon vil bli presentert i kapittel 6.

2 Litteratur

Denne oppgaven har mest til felles med arbeid innen urban økonomi og økonomisk geografi. I dette kapitlet vil det bli utledet en teoretisk modell som viser effekten et høyt befolkningsnivå kan ha på befolkningsvekst. Tidligere studer som har utført lignende forskning vil også bli gjennomgått.

2.1 Teoretisk rammeverk

Det presenteres her en modell i henhold til Duranton (2008). Denne modellen benyttes normalt til å se på størrelse på byer. Modellen vil derimot i denne oppgaven bli benyttet i et forsøk på å forklare befolkningsdivergens mellom regioner. Det antas at alle regionene står overfor en Cobb-Douglas produktfunksjon der produksjon er gitt som X :

$$X = N^\alpha K^\beta L^\delta$$

Antall arbeidere er gitt som N . Det er antatt at disse jobber og bor i samme region. Kapital er gitt ved K og areal L . Det er videre antatt at kapital er mobil mens areal er immobil mellom regionene. Profittmaksimering gir:

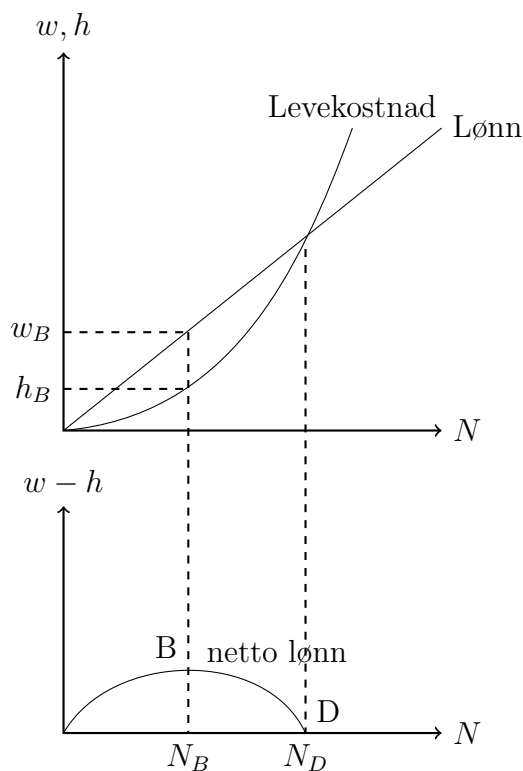
$$\frac{\partial X}{\partial N} = w = \alpha N^{\alpha-1} K^\beta L^\delta$$

Dette er den inverse av funksjonen for arbeidskraftsetterspørsel, der w er lønn. Lønna antas å være stigende i forhold til antall arbeidere i regionen. Dette forklares med agglomerasjonseffekter, som betyr $\alpha > 1$.

$$\frac{\partial w}{\partial N} = \alpha(\alpha - 1)N^{\alpha-2} K^\beta L^\delta > 0$$

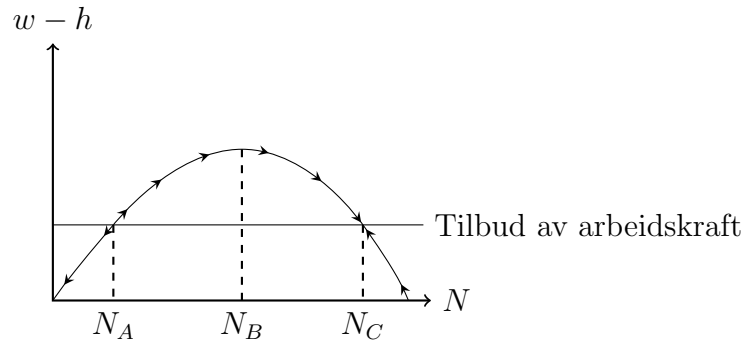
Jo større agglomerasjonseffekter regionen har desto mer vil lønna øke når antall arbeidere i regionen går opp. Når antall arbeidere i en region øker, stiger også trengselseffektene (h). Dette gir en kurve for levekostnader som også er stigende i befolkning. Ved å se på forskjellen mellom lønnskurven og levekostnadskurven, finner man

kurven for netto lønn. Det er antatt at agglomerasjonseffektene dominerer trengsels-effektene for små befolkningsnivå, og motsatt for store befolkningsnivå. Dette gjør at netto lønnskurven har en invers u-form.



Figur 1: Netto lønnskurve i en region.

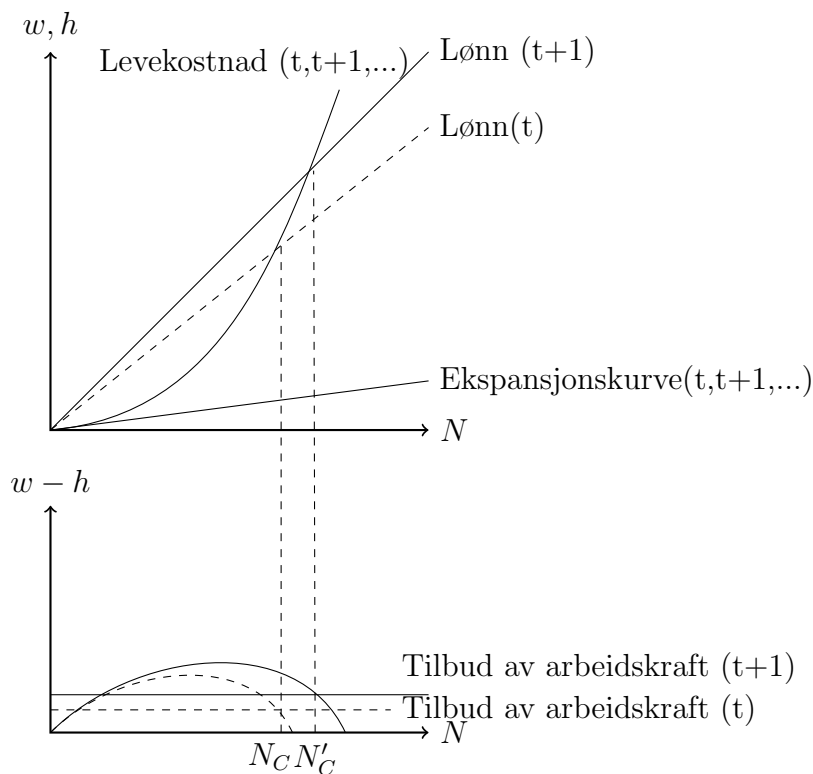
Det antas videre at arbeiderne er perfekt mobile slik at de kan flytte til en ny region uten kostnad. Dette fører til at kurven for tilbud av arbeidskraft er horisontal. Plasseringen til kurven for tilbud av arbeidskraft avgjøres av arbeiderens preferanser. Attraktive regioner vil ha en lavere tilbudskurve for arbeidskraft enn andre regioner, siden arbeiderne er villig til å bo i regionen for lavere netto lønn. Ved hjelp av kurvene for tilbud av arbeidskraft og netto lønn, kan man finne hvor likevekten av befolkningsnivået i en region er.



Figur 2: Likevekt for befolkningsnivå for en region.

Likevekt vil være der kurven for netto lønn og kurven for tilbud av arbeidskraft krysser hverandre. Dersom en region har høyere nettolønn enn lønnen arbeidere er villig til å akseptere vil regionen ha tilflytting. Dette gjør at N øker og det blir en bevegelse mot høyre i figur 2. I motsatt tilfelle, der netto lønn er lavere enn lønna arbeidere er villig til å akseptere, vil regionen ha fraflytting, og N vil da reduseres. N_A er en ustabil likevekt der man ved å øke antall arbeidere i regionen kan få høyere netto lønn. N_B er det optimale antall arbeidere i en region, men N_C er den eneste stabile likevekten som kan oppnås.

Duranton (2008) argumenterer også for å inkludere en ekspansjonskurve. Denne skal fange opp de positive effektene størrelsen på regionen har på humankapital i regionen. Det antas at alle arbeidere lærer noe av hverandre i løpet av en periode. Jo flere arbeidere det er i en region desto mer lærer arbeiderne. I et dynamisk rammeverk starter arbeiderne neste periode med mer humankapital. Dette vil skifte lønnskurven opp i hver periode.



Figur 3: Dynamisk befolkningsvekst i en region.

Den dynamiske modellen vil skifte befolkningsnivået i likevekten utover i hver periode som følge av den økte lønnen økt humankapital fører med seg. I figur 3 er skiftet i befolkningsnivå representert med skiftet fra N_C til N'_C . Ser at når humankapitalen går opp vil arbeiderne øke minimumslønnen de er villige til akseptere i alle regioner. Dette skjer fordi alle regioner er utsatt for en form av læring. Humankapitalnivået i landet vil da øke, som igjen øker minimumslønnen arbeiderene er villige til å akseptere. Regioner som nå har nettolønn lavere enn den minste lønnen arbeiderne vil akseptere, vil oppleve fraflytting. På denne måten kan Duranton (2008) sin dynamiske modell bidra med en forklaring til hvorfor befolkningsnivået i regionene divergerer.

2.2 Tidligere empiriske studier

Det meste av forskning innen urban økonomi og økonomisk geografi prøver å forklare størrelse på byer, og å finne grunner til at noen byer blir store mens andre forblir små. Denne oppgaven ønsker å studere divergens i befolkningsnivået, dette gjør det

mer hensiktsmessig å benytte regioner. Fordelen med å se på regioner er at disse er definert for alle områder, ikke bare de som opplevde vekst og ble til byer, men også de områdene som ikke ble til byer. Ved å bruke data på regionnivå unngår man dermed et suksessutvalg, noe De Long (1988) kritiserer Baumol (1986) for i hans analyse av inntektskonvergens for OECD land. Studier av befolkningsmobilitet på regionnivå er noe som gjøres i blant annet Ayuda et al. (2010), P.E. Beeson et al. (2001) og Rattsø og Stokke (2014a).

P.E. Beeson et al. (2001) benytter seg av befolkningsdata på county nivå i USA, for perioden 1840 til 1990. Studien undersøker om naturlige karakteristikk påvirker hvor man valgte å bosette seg i 1840. Videre ser de om naturlige eller sosiale karakteristikk kan ha påvirket befolkningsveksten i perioden. De finner at naturlige karakteristikk har påvirket hvor man valgte å bosette seg i 1840, samt befolkningsveksten i perioden. Sosiale karakteristikk har også hatt en signifikant effekt på befolkningsveksten i perioden. Til slutt ser de på om det har vært befolkningskonvergens eller befolkningsdivergens ved bruk av barroregresjoner. De finner tegn på befolkningsdivergens for amerikanske county.

Ayuda et al. (2010) finner befolkningsdivergens mellom regioner i 9 europeiske land i perioden 1850 til 2000. Basert på data for befolkningstetthet finner de ved hjelp av barroregresjoner, Kernel-tettheter og Markov-kjeder at alle landene de ser på har opplevd befolkningsdivergens på regionsnivå i perioden 1850–2000. Ayuda et al. (2010) ser også en signifikant negativ sammenheng mellom areal og befolkningsvekst i Belgia, Frankrike, Spania og England. Samtidig viser også Nederland tegn på negativ sammenheng mellom areal og befolkningsvekst.

I Norge undersøker Rattsø og Stokke (2014a) divergens i inntekt og befolkningsnivå på regionnivå i perioden 1972-2008. De finner tegn på at Norge har vært preget av regional befolkningsdivergens i perioden, og at denne divergensen har vært drevet av mellomstore regioner. Denne positive sammenhengen for mellomstore regioner er også noe Michaels et al. (2012) finner når de ser på befolkningsvekst for regioner i USA fra 1880 til 2000. De finner en sammenheng mellom befolkningsvekst og initial befolkningstetthet.

Desmet og Rappaport (2017) forklarer sammenhengen mellom befolkningsvekst og

initielt befolkningsnivå for ulike grupper av befolkningsnivået ved å utvikle en modell med en sektor. De bruker modellen for å se hvordan befolkningsnivået tilpasser seg i regionene. Studien finner at trengselseffektene som følge av at tilgang på land er konstant, er avtagende i forhold til agglomerasjonseffekter. Et resultat av dette er at mellomstore og store regioner divergerer i forhold til hverandre. Rattsø og Stokke (2014a) og Michaels et al. (2012) sine funn tyder derimot på at avtagende trengselseffekter kun gjelder for mellomstore regioner. Desmet og Rappaport (2017) finner at de minste regionene konvergerer, men dette blir forklart med at modellen tillater nyetablering av regioner. En stund etter nyetablering divergerer også befolkningsnivået i de minste regionene.

Kim (1995) finner tegn til økonomisk spesialisering på tvers av amerikanske regioner i perioden 1860 til 1990. Dette er et resultat som passer bedre inn i et rammeverk av neoklassiske modeller. I slike modeller antas produktivitet å være eksogent gitt. Forskjeller mellom regionene vil stamme fra ulik tilgang på mobile ressurser som kapital og arbeidskraft, samt tilgangen på umobile naturressurser (Borts & Stein, 1964). Forskjeller i befolkningsvekst mellom regioner vil da komme av at regionene tilpasser seg til langtidslikevekt. Ulik initiell fordeling av ressurser gjør at avkastning på innsatsfaktorer vil variere på tvers av regioner. Uten hindringer på flyt av ressurser, vil utligning av avkastning på innsatsfaktorer føre til betinget befolkningskonvergens.

3 Metode

3.1 Barroregresjoner

Ved å benytte regresjoner av paneldata, kan man undersøke om det har vært konvergens eller divergens i underliggende data. Barro og Sala-i Martin (1992) introduserte en lineær modell for å gjøre dette:

$$\frac{1}{T} \ln \left(\frac{p_{iT}}{p_{i0}} \right) = \beta_0 + \beta_1 \ln(p_{i0}) + u_i \quad (1)$$

Denne estimeres ved hjelp av minste kvadraters metode (MKM). Gauss-Markov forutsetningene for multipl regresjon er derfor sentrale for at estimatoren skal være den beste som kan oppnås (Wooldridge, 2015, s. 92). Venstresidevariabelen i ligning 1 er årlig vekstrate, der $T = 97$ og p_{it} er region i sin befolkning i år t ($t=T$ gir 2017 og $t=0$ gir 1920). u_i er restleddet i modellen. Logaritmiske verdier benyttes for å gjøre analysen mindre sensitiv for ekstremverdier (Wooldridge, 2015, s. 172). En slik modell omtales ofte som en ubetinget modell for konvergens fordi den ikke inneholder andre kontrollvariabler. Denne undersøker hvilken effekt logaritmisk initielt befolkningsnivå har på befolkningsveksten i perioden. En negativ β_1 gir indikasjon på absolutt konvergens. Absolutt konvergens betyr at uavhengig av initielt befolkningsnivå vil regionene konvergere mot det samme befolkningsnivået. Dette skjer ved at regioner som opprinnelig hadde et høyt befolkningsnivå får en lav befolkningsvekst i perioden, mens regioner med opprinnelig lavt befolkningsnivå får en høy befolkningsvekst i perioden.

Utvider man denne modellen med flere forklaringsvariabler vil en negativ β_1 indikere betinget konvergens. I motsetning til absolutt konvergens betyr betinget konvergens at regioner vil konvergere dersom de har samme karakteristikk. Betingede barroregresjoner har møtt kritikk fra blant annet D.T. Quah (1996) samt Bernard og Durlauf (1996). De viser hvordan en slik modell vil ha problemer dersom det er konvergens mot ulike langsiktige likevekter. P.E. Beeson et al. (2001) mener man i alle fall kan slå fast at en negativ β_1 viser en tendens der store initielle populasjoner vokser sakte, dog må regresjonene tolkes med et kritisk blikk. Denne kritikken

mot Barroregresjoner gjør at Kernel-tetthetsfunksjoner samt Markov-kjeder også estimeres for å bedre kunne studere dynamikken i data.

3.2 Økonometriske utfordringer

Denne økonometriske analysen gir flere utfordringer som kan gi forventningskjevhet i estimatene, eller problemer med å trekke inferens. Her vil de største utfordringene modellen i ligning 1 gir diskuteres.

En svært sentral forutsetning for at MKM skal gi en kausal sammenheng er at ingen av forklaringsvariablene er korrelerte med restleddet (Wooldridge, 2015, s. 76-77). Brudd på denne forutsetningen vil føre til at variabelen tolkes som endogen i en økonometrisk forstand. Det er hovedsakelig tre kilder til endogenitet: utelatte variabler, simultanitet og målefeil. Disse problemene samt aktualiteten i forhold til oppgaven vil bli diskutert her.

3.2.1 Utelatte variabler

Problemet med utelatte variabler oppstår dersom det finnes en eller flere relevante variabler som er utelatt fra modellen, men er korrelert med en eller flere av forklaringsvariablene i regresjonen. Det vil da oppstå en skjevhet i estimatoren fordi den nå er en bruttokoeffisient, hvilket innebærer at den fanger opp både effekten til befolkningsnivået i 1920 samt eventuelle utelatte variabler. Dersom vi tenker oss at den sanne modellen er $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, men kun regresserer $y = \hat{\beta}_0 + \hat{\beta}_1$ vil forventningsverdien til $\hat{\beta}_1$ være:

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \hat{\delta} \quad (2)$$

Hvor $\hat{\delta}$ er regresjonskoeffisienten i en regresjon mellom x_1 og x_2 . β_1 og β_2 er de faktiske verdiene fra den sanne modellen.

$$\hat{\delta} = \frac{\sum(x_2 - \bar{x}_2)(x_1 - \bar{x}_1)}{\sum(x_1 - \bar{x}_1)^2} \quad (3)$$

\bar{x}_1 og \bar{x}_2 er gjennomsnittsverdiene til henholdsvis x_1 og x_2 .

Forventet skjevhet ved en utelatt variabel vil da være:

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \hat{\delta} \quad (4)$$

MKM-estimatet vil være forventningsrett dersom x_2 ikke har noen effekt på venstresidevariabelen ($\beta_2 = 0$), eller dersom korrelasjonen mellom x_1 og x_2 er null ($\delta=0$) (Wooldridge, 2015, s. 78-81). Dette er et svært relevant problem for oppgaven. Det er derfor viktig å inkludere alle variabler som kan påvirke befolkningsvekst og samtidig være korrelert med befolkningsnivået i 1920.

3.2.2 Simultanitet

Når en eller flere forklaringsvariabler bestemmes samtidig som den avhengige variabelen, har man simultanitet. Wooldridge (2015, s. 499-504) kaller dette toveis kausalitet. Dette er mindre aktuelt i denne oppgaven siden alle forklaringsvariablene enten er antatt konstante eller hentet fra år null. De to eneste variablene som hentes under perioden befolkningsveksten studeres, er variablene for klima. Dette trekker i retning av at simultanitetsskjevhet ikke er et stort problem for analysen.

3.2.3 Målefeil

Målefeil oppstår når den faktiske verdien til en variabel avviker fra den observerte verdien (Wooldridge, 2015, s. 287-292). Det er ikke spesielt alvorlig med målefeil i venstresidevariabelen så lenge dette ikke er systematisk korrelert med forklaringsvariabelen. Dersom målefeilen er i en av forklaringsvariablene vil den aktuelle variabelen gi en skjevhet mot null ved bruk av MKM-estimatoren. Wooldridge (2015, s. 290) kaller denne skjevheten for klassisk målefeilskjevhet.

Alle data er hentet fra SSB, SNL, NSD eller Meteorologisk institutt. Disse er alle anerkjente og seriøse miljøer, dette reduserer sannsynligheten for systematisk målefeil. Målefeil antas derfor å påvirke oppgaven i liten grad.

3.2.4 Problemer knyttet til inferens

Dersom forutsetningen om homoskedastisitet brytes vil dette føre til problemer med å trekke inferens. Estimatorenns varians og standardavvik vil være basert på feil formel og dette kan gi villedende og ugyldige resultater (Verbeek, 2012, s. 95-96).

Heteroskedastisitet innebærer at restleddet avhenger av verdien til forklaringsvariabelen. Wooldridge (2015, s. 245) forklarer hvordan White sin formel for å estimere variansen til estimatoren under heteroskedastisitet er gyldig for alle former for heteroskedastisitet også homoskedastisitet, så lenge utvalget er stort nok. Det er derfor estimert heteroskedastiske robuste standardavvik i denne oppgaven.

3.3 Spline-regresjoner

Spline-regresjoner er en form for utvidelse av Barroregresjoner, der man kan studere om det er forskjell i sammenhengen mellom vekstrate og initielt befolkningsnivå når man tar hensyn til størrelsen på det initielle nivået. Dette gjøres ved å dele logaritmisk befolkningsnivå inn i tre intervaller som har like mange observasjoner. Det kan da estimeres en koeffisient for hver av disse tre undergruppene. Fra StataCorp (2013) har vi følgende ligning:

$$\frac{1}{T} \ln \left(\frac{p_{iT}}{p_{i0}} \right) = \beta_0 + \beta_1 \ln(p_{i0}) + \beta_2 \ln(p_{i0}) + \beta_3 \ln(p_{i0}) + u_i \quad (5)$$

der:

$$\frac{\partial \frac{1}{T} \ln \left(\frac{p_{iT}}{p_{i0}} \right)}{\partial p_{i0}} = \begin{cases} \beta_1 & \text{om } p_{i0} < x_1 \\ \beta_2 & \text{om } x_1 \leq p_{i0} < x_2 \\ \beta_3 & \text{om } p_{i0} \geq x_2 \end{cases} \quad (6)$$

Her vil x_1 og x_2 være grensene som deler logaritmen til initielt befolkningsnivå i ulike grupper. β_1 ser på konvergens for gruppen med lavest befolkning. Tilsvarende viser β_2 konvergens for de mellomstore regionene, og β_3 konvergens for de største regionene.

3.4 Kernel-tettheter

Kernel-tettheter er en metode for å beskrive fordelingen datamaterialet har. Utledningen i denne oppgaven støtter seg på Parzen (1962). Dersom man har et utvalg n med observasjoner X_1, X_2, \dots, X_n , der X_i har en tetthetsfunksjon $f(x)$ vil estimatoren for Kernel-tettheten til X_i være ($\widehat{f_k}(x)$). Denne er gitt som:

$$\widehat{f_k}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X - X_i}{h}\right) \quad (7)$$

Hvor K er en Kernel-funksjon. En sentral forutsetning for at Kernel-funksjonen er gyldig er at arealet under kurven er lik 1, altså: $\int k(x)dx = 1$. Jeg vil her benytte en fordeling basert på normalfordeling (Gaussisk) i henhold til Zucchini et al. (2003):

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (8)$$

Båndbredden h påvirker hvor glatt estimatet for Kernel-tettheten til X blir. En høy verdi gir glattere, og dermed mindre nøyaktig kurve. Zucchini et al. (2003) sier at man risikerer å overse informasjon dersom båndbredden settes for stor. I denne oppgaven benyttes en båndbredde i henhold til "Silvermans rule of thumb" (Silverman, 2018):

$$h = 1,06 \cdot \sigma \cdot n^{-\frac{1}{5}} \quad (9)$$

σ er standardavviket til data. Denne formelen vil gi et optimalt estimat dersom data er normalfordelt.

Praktisk forklart lager man den estimerte Kernel-tettheten ved å tillegge alle X_i -verdiene den Gaussiske fordelingen for så å addere alle disse fordelingene sammen. Der hvor det er overlapping mellom tetthetene blir Kernel-kurven høyere. Til slutt normaliseres estimatet til 1 ved å dividere med antallet observasjoner.

3.5 Lokal polynomglatting

Ved å benytte lokal polynomglatting prøver man uten en bestemt funksjonsform å la sammenhengene i data komme frem. Jianqing og Gijbels (1996) viser at dersom man har et spredningsplott $\{(X_i, Y_i)\}_{i=1}^n$ fra en modell $y_i = m(x_i) + \sigma(x_i)\epsilon_i$ hvor restleddet(ϵ) har $E(\epsilon_i) = 0$ og $Var(\epsilon_i) = 1$, hvor m og ϵ er glattingsparametere, kan $m(x_0)$ estimeres ved hjelp av et polynom av orden P^1 på observasjonene (X_i, Y_i) . Der X_i er vektet av Epanechnikov Kernel-funksjoner. Denne kernelen er kjent for å være optimal når man estimerer m og den deriverte av m (Fan et al., 1995). Her benyttes også et glattingsparameter med verdi fra “Silvermans rule of thumb”.

3.6 Markov-kjeder

Markov-kjedeanalyse benyttes for å studere hvordan fordelingen av befolkningen endres over tid, og hvordan dynamikken mellom regionene er. Grunnprinsippene for analysen ble presentert av Shorrocks (1978) og videreført av D. Quah (1993a, 1993b). Markov-kjedeanalyse benyttes normalt til å studere konvergens av inntekt, men Rattsø og Stokke (2014b) benytter også metoden til å studere distribusjonen av utdanning mellom regioner. De poengterer at man må benytte lange nok overgangsperioder dersom variabelen endres sakte. Hovedsakelig vil analysen benytte overgangsperioder på fem år, men det tas også hensyn til at variablene endres sakte ved å sjekke robustheten med overgangsperioder på ti år.

Først genereres en variabel med befolkningsnivå i en region relativt til gjennomsnittlig befolkningsnivå for alle regioner det aktuelle året. Deretter deler man disse i N gjensidig utelukkende grupper for relativt befolkningsnivå. Normal praksis er å benytte $N = 5$. I hver tidsperiode som studeres plasseres regionene i en gruppe.

I oppgaven benyttes en førsteordens Markov-kjedeprosess. Dette betyr at regionene sin sannsynlighet for å bevege seg mellom gruppene kun avhenger av hvilken gruppe regionen tilhørte i forrige periode. Sannsynligheten for at en region beveger seg fra gruppe i til j fra periode $t-1$ til t , måles ved $p_{ij}(t)$. Denne sannsynligheten estimeres ved å benytte “maximum likelihood-method”. Analysen ser på bevegelsene for en

¹I oppgaven benyttes orden av null, dette innebærer lokal gjennomsnittsglatting.

gruppe av regioner i løpet av perioden. Antall regioner som beveger seg fra gruppe i i perioden $t - 1$ til t er gitt ved:

$$n_i(t - 1) = \sum_j n_{ij}(t) \quad (10)$$

Markov-kjeden blir redusert til et produkt av fem gjensidig uavhengige multinomiale fordelinger, en fordeling for hver rad i i overgangssannsynlighetsmatrisen. For hver tidsperiode t er fordelingsfunksjonen:

$$f(n_{ij}(t)) = \prod_{i=1}^N f_i(n_{ij}(t)) = \prod_{i=1}^N \left[\frac{n_i(t - 1)!}{\prod_{j=1}^N n_{ij}(t)!} \prod_{j=1}^N p_{ij}^{n_{ij}(t)} \right] \quad (11)$$

Overgangssannsynlighetene estimeres ved å maksimere “log likelihood funksjonen” av ligning 11, for alle T , med hensyn på p_{ij} :

$$f(n_{ij}) = \prod_{t=1}^T f(n_{ij}(t)) \quad (12)$$

Siden en region alltid må være i en av gruppene, vil summen av p_{ij} for alle j være 1. Dette gjør at estimatoren for “maximum likelihood” er den relative frekvensen av overganger:

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_i} = \frac{\sum_{t=1}^T n_{ij}(t)}{\sum_{t=1}^T n_i(t - 1)} \quad (13)$$

n_{ij} er antallet observasjoner som har gått fra i til j mellom periode $t - 1$ og t . n_i er antall observasjoner som var i i i periode $t - 1$. Den initielle fordelingen av regionalt befolkningsnivå antas uniform og er definert som:

$$h(0) = [h_1(0), h_2(0), h_3(0), h_4(0), h_5(0)] \quad (14)$$

der:

$$\sum_i h_i(0) = 1 \quad (15)$$

Fordelingen etter første overgangsperiode er gitt som:

$$h(1) = h(0)\Omega \quad (16)$$

Ω er den estimerte 5×5 Markov-overgangssannsynlighetsmatrisen. Etter k perioder har vi:

$$h(k) = h(0)\Omega^k \quad (17)$$

Gitt at matrisen er regulær vil fordelingen konvergere til den begrensede fordelinga:

$$h^* = \lim_{k \rightarrow \infty} h(0)\Omega^k \quad (18)$$

Dette gir den langsiktige, ergodiske fordelingen, gitt at overgangsdynamikken forblir konstant. Det er derfor viktig å være klar over at Markov-kjedene gir tendens på hva som har skjedd i perioden vi undersøker for, slik at man er forsiktig med å tolke den ergodiske fordelingen som noe som kommer til å skje i fremtiden (D. Quah, 1993a).

3.7 Q-og LR-tester

Pearson (Q) tester og Likelihood Ratio (LR) tester kan undersøke om befolkningsovergangene er tidsuavhengige (Bickenbach & Bode, 2003). Dette gjøres ved å dele hele tidsperioden i M delutvalg som er gjensidig utelukkende, men samtidig fanger opp alle observasjonene. Overgangsmatrisene for hver av delutvalgene sammenlignes så med overgangsmatrisen for hele utvalget. Begge testene har en asymptotisk χ^2 -fordeling der frihetsgradene bestemmes av antall uavhengig parvise sammenligninger.

$$Q = \sum_{m=1}^M \sum_{i=1}^5 \sum_{j \in A_i} n_{i|m} \frac{(\hat{p}_{ij|m} - \hat{p}_{ij})^2}{\hat{p}_{ij}} \sim \text{asy } \chi^2 \left(\sum_{i=1}^5 (a_i - 1)(b_i - 1) \right) \quad (19)$$

$$LR = 2 \sum_{m=1}^M \sum_{i=1}^5 \sum_{j \in A_i|m} n_{ij|m} \ln \frac{\hat{p}_{ij|m}}{\hat{p}_{ij}} \sim \text{asy } \chi^2 \left(\sum_{i=1}^5 (a_i - 1)(b_i - 1) \right) \quad (20)$$

Her er A_i settet med overgangssannsynligheter større enn null i den i 'te raden for hele utvalget. $A_{i|m}$ er tilsvarende for delutvalg m . Antall overganger fra gruppe i for delutvalg m gis som $n_{i|m}$, og antall overganger fra gruppe i til gruppe j for

delutvalget er gitt som $n_{ij|m}$. Antall frihetsgrader gis i siste parentes. a_i er antall elementer i A_i mens b_i er antall delutvalg med observasjoner som er forskjellig fra null i rad i . Nullhypotesen blir her at overgangssannsynlighetene er uavhengige av tid.

4 Data

Datamaterialet som brukes i analysen er i hovedsak hentet fra Norsk Samfunnsvitenskapelig Datatjeneste (NSD)², samt Statistisk Sentralbyrå (SSB). Unntakene er temperaturnormal, nedbørnormal og dummyen for universitet. Nedbør- og temperaturnormal er tilsendt av Meteorologisk institutt, mens universitetsdummyen er kodet etter Norsk utdanningshistorie (Thune, 2019).

All data er hentet på kommunenivå, men aggregert opp i henhold til økonomiske regioner fra SSB (Hustoft et al., 1999). Dette er en inndeling som er laget for å tilfredsstille EU sine NUTS-4 krav. En av fordelene med å addere opp til regionnivå fra kommunenivå, er at mye økonomisk aktivitet krysser kommunegrensene. Et eksempel på dette er arbeidskraft. Mange arbeidere bor i en annen kommune enn de jobber i. Økonomiske regioner fra SSB er definert ut fra pendlingsdata, og fungerer derfor bedre enn kommuner for å avgrense lokale økonomier. Et sentralt krav for inndelingen er dog at regioner ikke kan gå på tvers av fylkesgrenser. Dette er noe Bhuller (2009) kritiserer. I robusthets analysen er det derfor inkludert en alternativ regioninndeling foreslått av Bhuller (2009). Denne inndelingen tillater regioner å gå på tvers av fylkesgrenser. Kommuner som da samhandler med hverandre kan ses på som en økonomi og vil da være i samme region. Hovedsakelig er den største forskjellen at størrelsen på by-regionene øker.

Et sentralt problem med å addere kommunedata over en tidsperiode på nesten 100 år er forandringer i kommunegrensene. Denne utfordringen er løst ved å skape konstante geografiske områder. Regionene er derfor etter beste evne kodet slik de så ut i 2017.

4.1 Avhengig variabel

Avhengig variabel vil være gjennomsnittlig årlig vekstrate i populasjonen på logaritmisk form. Denne er konstruert på følgende måte:

$$\frac{1}{T} \cdot \ln \left(\frac{P_T}{P_o} \right) \quad (21)$$

²Data er tilgjengelig med spesiell tillatelse fra NSD.

Her er T antall år, P_T er populasjonsstørrelsen i siste periode og P_0 er populasjonsstørrelsen i første periode. I utgangspunktet studeres perioden 1920-2017, men også kortere perioder vil bli studert for å se på robustheten til analysen.

Tallene for populasjonen i 1920 kommer fra folketellingen i 1920 som ble avholdt "1. desember og påfølgende virkedager". Disse er basert på hjemmehørende folkemengde, dette innebærer følgende: "Hjemmehørende folkemengde skal forstås som det antall personer som på tellingstidspunktet hadde fast bosted innenfor de enkelte tellingsenheter. Midlertidig fraværende personer er medregnet, derimot ikke midlertidig nærværende." Valget falt på dette fordi det er svært likt måten folkeregistrering fungerer i dag. Befolkningsnivået etter 2. verdenskrig er målt som registrerte innbyggere i kommunen 1. januar, med unntak av årene 1947-1950 hvor det ikke er registrert data.

4.2 Forklaringsvariabel

Poenget med oppgaven er å beskrive hvordan befolkningsstrømmene i Norge har beveget seg. Dette gir to aktuelle muligheter for hovedforklaringsvariabel, logaritmisk tilnærming til befolkningsnivå eller populasjon per kvadratkilometer. Det vil her bli benyttet befolkningsnivå siden by-regioner med store utkantareal vil få lav tetthet, og regioner som er geografisk små vil kunne få høy tetthet.

4.3 Kontrollvariabler

Hensikten med kontrollvariabler er å fange opp effekten andre variabler enn befolkningsnivå kan ha på veksten. Utelatelse av relevante variabler åpner for at regresjonen kan ha et utelatt variabelproblem. Jo flere signifikante variabler modellen har, desto høyere forklaringskraft vil modellen ha. Følgende kontrollvariabler vil bli benyttet:

4.3.1 Areal

Siden befolkningsnivå benyttes som forklaringsvariabel inkluderes areal som kontrollvariabel. Denne representeres ved bruk av logaritmen til arealet.

4.3.2 Andel fjell

Denne variabelen viser hvor stor andel av arealet i en region som består av fjell. Generelt er dette en naturtype med lav produktivitet og med liten økonomisk interesse. Fjellområder er også med på å gjøre ferdsel vanskeligere. Det er derfor naturlig å tro at regioner med stor andel av fjell vil ha lavere befolkningsvekst enn områder uten. Variabelen for andel fjell er konstruert på følgende måte:

$$\text{Andel fjell} = \frac{\text{Fjell km}^2}{\text{Totalt areal km}^2} \quad (22)$$

4.3.3 Klima

Det kontrolleres for to klimafaktorer. I denne oppgaven benyttes normaltemperatur og nedbørnormal. Normalt vær beregnes ved å ta gjennomsnittsverdi over en bestemt 30-års periode. Normalperiodene har vært 1901-1930, 1931-1960 og 1961-1990. Her vil normalperioden 1961 til 1990 benyttes. Grunnen til at denne perioden ble valgt og ikke normalperioden 1931 til 1960 er at den inneholdt data for alle regionene. P.E. Beeson et al. (2001) finner at høyere temperatur og mindre nedbør gir en positiv effekt på befolkningsveksten.

4.3.4 Universitet

Universitetsvariabelen er en dummyvariabel som tar verdien 1 dersom det var et universitet i regionen i 1920. Universiteter er med på å øke humankapitalen. Dette er noe flere studier (e.g., P. Beeson & Montgomery, 1993; Glaeser et al., 1995; Simon & Nardinelli, 2002) finner bevis for at påvirker befolkningsveksten positivt.

4.3.5 Tollhavn

Havet har i Norge vært en viktig transportåre. Ved hjelp av en dummyvariabel for om regionen hadde en havn med tollstasjon i 1920, blir det forsøkt å representere om regionen har kontakt med utlandet. Muligheten til å bytte ideer og varer er noe det er naturlig å tenke seg kan bidra til å øke befolkningsveksten. (Sokoloff, 1988)

4.4 Deskriptiv statistikk

Den deskriptive statistikken tar først for seg tall for befolkningsvekst og befolkningsnivå. Veksten er kalkulert både for perioden 1920-2017 og kortere delperioder. Tabell 1 viser gjennomsnitt, standardavvik og den minste og største observasjonen. Fra tabellen kan man se at gjennomsnittlig årlig befolkningsvekst er på 0.45 prosent. 13 av de 90 regionene har hatt negativ befolkningsvekst, mens den raskest voksende regionen har hatt en årlig gjennomsnittlig befolkningsvekst på 2 %. De tre regionene med lavest befolkningsvekst er Rjukan, Høyanger og Frøya/Hitra. På motsatt ende av skalaen finner vi de tre regionene Bærum/Asker, Follo og Jæren med høyest befolkningsvekst. Tabell 1 viser også at gjennomsnittlig befolkningsnivå i regionene var nesten 30 000 med et standardavvik på nesten 40 000 i 1920. Dette vitner om at det var stor variasjon mellom regionene. Lavest befolkningsnivå hadde Grong med 4 442 innbyggere mens Oslo som største regionen hadde 312 046 innbyggere. Mye variasjon i forklaringsvariabelen legger et solid grunnlag for beregning av gode estimater. Areal er også en variabel med mye variasjon. Regionene har i gjennomsnitt et areal på 3 403 kvadratkilometer, med et standardavvik på nesten 3 233. Den minste regionen i areal er Sande/Svelvik med 230 kvadratkilometer. Hammerfest er den største regionen etter areal med 18 998 kvadratkilometer.

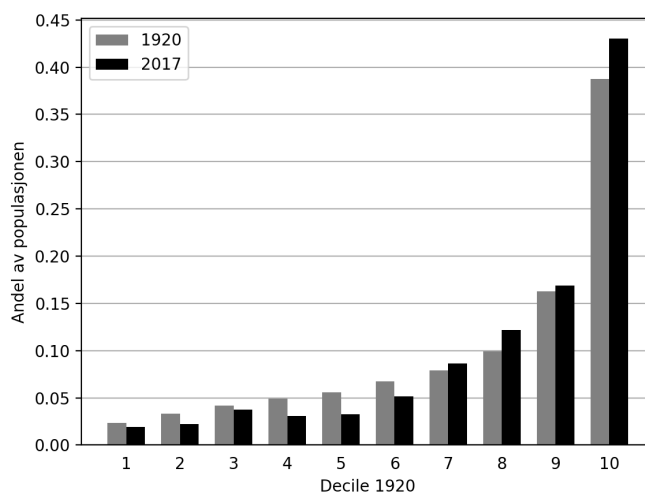
Tabell 1: Deskriptiv statistikk for 90 økonomiske regioner.

Variabler	Gjennomsnitt	Standardavvik	min	max
årlig_popg_full	0.00450	0.00455	-0.00747	0.0197
pop_g_1920_1946	0.00545	0.0051	-0.01216	0.0196
pop_g_1946_1982	0.00548	0.0064	-0.00897	0.0268
pop_g_1982_2017	0.002799	0.00574	-0.01227	0.0162
pop_1920	29399.67	39535.35	4442	312046
pop_1946	35077.3	51451.51	6408	417225
pop_1982	45634.02	63165.04	6592	450386
pop_2017	57887.12	92747.25	5197	666692
Kontrollvariabler				
km2	3402.967	3232.876	230	18998
Universitet (=1)	0.0333	0.181	0	1
Anedelfjell	0.0563	0.0653	0.0001	0.216
Tollhavn1920 (=1)	0.467	0.502	0	1
Noramlttemp (°C)	4.515	1.942	-0.543	7.395
Nedbørsnomal (mm)	1 136	448.3	443.5	2 312

5 Analyse

5.1 Befolkningsutvikling

Starter med å se på fordelingsegenskapene til regionene. Figur 4 viser befolkningsandel for 1920 og 2017, der regionene er delt inn i konstante desiler etter befolkningsnivået i 1920. I de 60% lavest befolkede regionene har andelen innbyggere sunket fra 23.8% i 1920 til 16.3% i 2017. På andre siden ser vi at de regionene som var blant de 40% største i 1920 har økt sin andel av befolkningen. Topp-desilen har gått fra nesten 38% i 1920 til over 43% i 2017.



Figur 4: Befolkningsandel fordelt i desiler for 1920.

Figur 4 viser at andelen av befolkningen har økt i regioner som allerede hadde en høy andel av befolkningen. Dette er et resultat ganske likt det Rattsø og Stokke (2014a) fant for perioden 1972-2008. De fant dog at det kun var de tre øverste desilene som har hatt en økning i populasjonsandelen i løpet av perioden. Her studeres en lengre periode og da har også den fjerde desilen økt sin andel av befolkningen. Rattsø og Stokke (2014a) fant at de tre største desilene økte fra 69.8% til 74%, altså en økning på 4.2 prosentpoeng. Figur 4 viser at de tre største desilene har økt fra 64.9% til 72.1% altså en økning på 7.1 prosentpoeng. Resultatene i denne oppgaven viser tegn til en sterkere urbanisering enn hva Rattsø og Stokke (2014a) fant. Dette kan forklares med at det er en lengre periode som studeres. Det viser

seg at andelen i den største desilen er lavere enn Rattsø og Stokke (2014a) sine resultater. Dette kan forklares med at man i 1972 hadde en annen inndeling i desiler enn hva som var tilfellet i 1920. Rattsø og Stokke (2014a) sammenlignet sine funn med P.E. Beeson et al. (2001) sine funn for amerikanske county. Beeson fant at andelen av befolkningen i desil 4-9 reduseres i perioden 1840-1990 i Amerika, altså økes andelen av befolkningen i desil 1-3. I likhet med Rattsø og Stokke (2014a) viser denne oppgaven at Norge har hatt en raskere og bredere urbanisering enn USA. Dette resultat peker i retning av divergens i befolkningsnivået mellom regionene i Norge.

5.2 Barroregresjoner

Tabell 2 viser barroregresjonene for de 90 økonomiske regionene over hele tidsperioden. Fra kolonne 1 i tabellen ser man at det er en signifikant positiv sammenheng mellom befolkningsnivået i 1920 og befolkningsveksten i perioden 1920 til 2017. Regresjonen sier at en dobling av befolkningsnivået i 1920 vil øke den årlige befolkningsveksten med 0.2 prosentpoeng. Den positive koeffisienten for initielt befolkningsnivå indikerer ubetinget divergens for norske regioner i perioden 1920 til 2017. Det betyr at regioner som hadde et høyt befolkningsnivå i 1920, har hatt en høyere befolkningsvekst enn regioner som hadde et lavere befolkningsnivå i 1920. MKM-regresjonslinjen er illustrert i figur 5. Sammenlignet med andre studier har modellen samme fortegn, men en lavere effekt enn hva Rattsø og Stokke (2014a) finner for Norge i perioden 1972-2008. Det samme er tilfellet for studien til Ayuda et al. (2010) der de ser på populasjonsdivergens i 9 europeiske land.

En regresjon som i tabell 2 kolonne 1 viser ubetinget divergens, men den kontrollerer ikke for andre relevante faktorer som kan påvirke befolkningsveksten. Dersom utelatte variabler også er korrelert med befolkningsnivået i 1920, vil regresjonen gi en skjev koeffisient som forklart i kapittel 3.2.1.

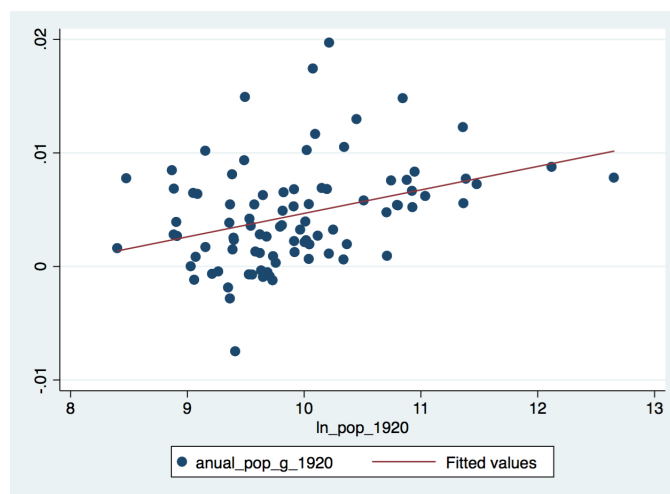
For å gjøre problemet med utelatt variabel mindre sannsynlig inkluderes også sosiale og naturlige faktorer i tabell 2 kolonne 2-4. Disse regresjonene vil også kunne gi en indikasjon på om det er noen sosiale eller naturlige karakteristikk som har innvirkning på hvor folk ønsker å bosette seg.

Tabell 2: Barroregresjoner for 90 regioner.

	(1)	(2)	(3)	(4)
Variabler	Årlig vekst	Årlig vekst	Årlig vekst	Årlig vekst
ln_pop_1920	0.00207*** (0.000462)	0.00149** (0.000642)	0.00179*** (0.000665)	0.00206*** (0.000517)
ln_km2			-0.00113 (0.000956)	-0.00143** (0.000580)
Anedelfjell			-0.000899 (0.00615)	
Noramltemp			0.000368 (0.000669)	
Nedbørsmal			-1.99e-06 (1.77e-06)	
Universitet(=1)		0.00361 (0.00390)		
Tollhavn1920 (=1)		0.00119 (0.000960)		
Konstant	-0.0160*** (0.00467)	-0.0110* (0.00613)	-0.00386 (0.00708)	-0.00486 (0.00617)
N	90	90	90	90
R ²	0.121	0.157	0.222	0.206

Robust standardavvik i parentes

*** p<0.01, ** p<0.05, * p<0.1



Figur 5: Ubetinget befolkningsdivergens for 90 regioner.

Estimatet for befolkningsdivergens er i kolonne 2 justert for sosiale faktorer. Tabellen viser at koeffisienten for befolkningsnivået i 1920 fortsatt er positiv og statistisk signifikant når det kontrolleres for sosiale faktorer. Det at interessevariabelen fortsatt har et positivt fortegn er konsistent med resultatet for absolutt divergens. De to sosiale faktorene har begge et positivt fortegn. En positiv sammenheng mellom utdanningsinstitusjoner og befolkningsvekst stemmer med funnene til P. Beeson og Montgomery (1993). De finner en positiv sammenheng mellom kvaliteten til universitetet og netto migrasjon på 70- og 80-tallet. Det har også vist seg viktig å ha en utdannet befolkning for å skape vekst i både befolkning og inntekt (Glaeser et al., 1995). Tilgang på en tollhavn i regionen vil ha gitt en enklere tilgang til importerte varer og tjenester, samtidig vil det ha gitt muligheten til å få påvirkning og ideer utenfra. Tilgang til en tollhavn ville også gjort eksport av varer mer tilgjengelig. Fra tabellen ser vi dog at disse to ikke har en statistisk signifikant effekt hver for seg. Dersom forklaringsvariablene har høy korrelasjon kan dette skape et multikollinearitet problem. Dette vil kunne føre til at variansen blir estimert for høy og på sånn måte gjør at koeffisienten fremstår som ikke signifikant når den i realiteten er det. Foretar derfor en F-test med nullhypotesen om at disse to ikke har en samlet signifikant effekt. Testen sin P-verdi er på 0.34 og dette gir ikke grunnlag til å forkaste nullhypotesen. Jeg velger derfor ikke å inkludere sosiale effekter videre i analysene, da de ikke virker å ha en statistisk signifikant effekt.

Ved å justere for naturlige karakteristikker oppnås resultatet i kolonne 3 i tabell 2. Naturlige karakteristikker fører ikke til store endringer i koeffisienten til interessevariabelen. Den har fortsatt samme fortegn og er signifikant. Fra denne regresjonen fremgår heller ingen av de inkluderte variablene som signifikante. En F-test med nullhypotese om at alle kontrollvariablene samlet ikke har en effekt gir en p-verdi på 0.12. Dette gir heller ikke grunnlag for å forkaste nullhypotesen. Nedbør virker å ha en negativ effekt på befolkningsveksten. P.E. Beeson et al. (2001) finner også at områder med mindre regn hadde høyere befolkningsvekst. For Norge virker ikke effekten å være signifikant. Ved å studere datasettet viser det seg at det er sterk korrelasjon mellom areal, andel fjell og normaltemperatur. I Norge er stort sett de største regionene, rangert etter areal, høyere liggende områder med en del fjellområder. Disse har også lavere temperaturer enn resten av regionene. På bak-

grunn av den sterke korrelasjonen mellom de naturlige karakteristikene estimeres en regresjon der bare areal inkluderes. Fra kolonne 4 i tabell 2 ser man at selv om areal inkluderes påvirker ikke dette koeffisienten til forklaringsvariabelen, men den totale forklaringskraft øker. Dette gjør at mer av variasjonen i modellen forklares. Denne regresjonen forteller at effekten av en dobling av befolkningsnivå fortsatt vil øke den årlige veksten med 0.2 prosentpoeng. I tillegg viser regresjonen at en dobling av antall kvadratkilometer gir en redusert årlig vekst på 0.14 prosentpoeng. At dette er en kausal sammenheng er ikke opplagt, det tolkes heller i retning av at de regionene med mindre areal har hatt en større vekst enn de med mye areal.

Denne modellen kan potensielt fortsatt være utsatt for et utelatt variabelproblem. En relevant metode for å unngå dette ville vært å benytte instrumentvariabelmetoden (IV-metoden). IV-metoden går ut på å benytte en instrumentvariabel til å estimere forklaringsvariabelen ved bruk av MKM, ofte kalt første steget. Andre steget går ut på å benytte estimatet for forklaringsvariabelen til å estimere effekten på den avhengige variabelen, også her benyttets MKM. I dette tilfelle er det vanskelig å benytte denne metoden på grunn av et av kravene som stilles til instrumentvariabelen. Kravet som skaper problemer er at instrumentvariabelen kun kan påvirke venstresidevariabelen gjennom effekten den har på forklaringsvariabelen. Dette tilsier at instrumentvariabelen ikke er korrelert med restleddet, $cov(U_i, Z_i) = 0$, der Z er instrumentvariabelen. Dette er ikke noe man kan teste empirisk, og man må derfor kunne argumentere for dette. Dersom man har brudd på denne forutsetningen vil den estimerte effekten ved bruk av IV-metoden bli mer forventningsskjev enn MKM er i utgangspunktet. I dette tilfellet er det ikke noe opplagt instrument. På bakgrunn av dette vil det ikke bli gjennomført estimering ved bruk av IV-metoden, på tross av at dette skaper en usikkerhet om estimatet i tabell 2 kolonne 4 er til å stole på.

5.3 Robusthet av tidsperioder

Ved å dele opp perioden 1920-2017 i kortere delperioder undersøkes det om divergensen i populasjonsnivået mellom regionene har hatt ulik hastighet. Første periode slutter i 1946. Dette gjør at denne perioden blir 10 år kortere enn de to påfølgende, men bidrar til å isolere mer av effekten før/under krigen kontra etter.

Tabell 3: Oppdeling i tidsperioder.

Variabler	Koeffisient
1920-1946	
ln_pop_1920	0.000841 (0.000685)
ln_km2	0.00130** (0.000568)
Konstant	-0.0130 (0.00812)
R^2	0.072
1946-1982	
ln_pop_1946	0.00245*** (0.000790)
ln_km2	-0.00184*** (0.000678)
Konstant	-0.00494 (0.00938)
R^2	0.159
1982-2017	
ln_pop_1982	0.00326*** (0.000484)
ln_km2	-0.00303*** (0.000465)
Konstant	-0.00721 (0.00624)
R^2	0.511

Robust standardavvik i parentes

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Tabell 3 viser at det er først etter krigen det har vært en signifikant divergenseffekt mellom regionene. I perioden før og under krigen er det en positiv sammenheng

mellom befolkningsnivå og befolkningsvekst, men denne er ikke statistisk signifikant. Etter krigen har det derimot vært to perioder preget av divergens i befolkningsnivå mellom regionene. Ut i fra tabell 3 kan man se tendenser på at urbaniseringen er tiltagende i og med at siste periode har en større positiv koeffisient enn andre periode. I siste periode viser tabellen at en dobling i befolkningsnivå øker befolkningsveksten med 0.33 prosentpoeng. Dette er en effekt som er lik effekten Rattsø og Stokke (2014a) finner for perioden 1972 til 2008.

5.4 Robusthet region

I tabell 4 studeres hovedmodellen, men regionene er endret i henhold til Bhuller (2009). Her tillates det at regioner strekker seg på tvers av fylkesgrenser. Dette fører til at byregionene blir vesentlig større enn under de 90 økonomiske regionene. Dette vil ta hensyn til den positive effekten byregionene kan ha på omkringliggende regioner. Et eksempel på dette er regionen Asker/Bærum, som er den regionen som har hatt høyest befolkningsvekst av samtlige regioner i perioden. Denne veksten er sannsynligvis sterkt preget av nærheten til Oslo.

Tabell 4: Alternativ regioninndeling.

Variabler	Årlig vekst
\ln_pop_1920	0.00174*** (0.000580)
\ln_areal	0.000331 (0.000579)
Konstant	-0.0167** (0.00734)
N	45
R^2	0.192

Robust standardavvik i parentes

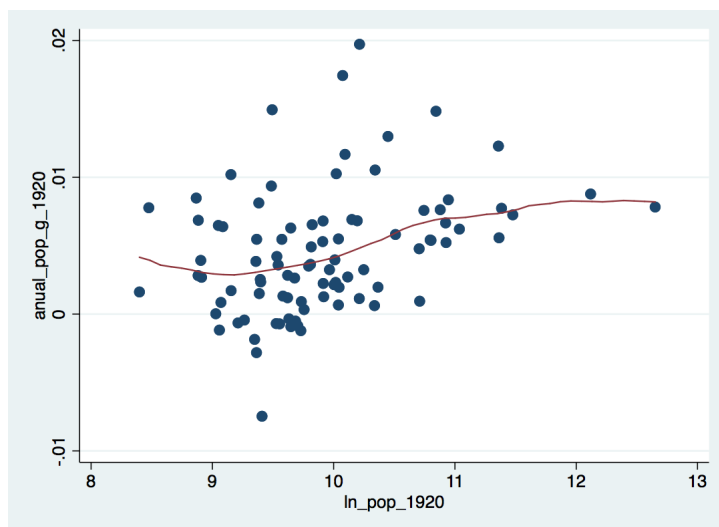
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Endringen av regioninndeling gir ikke empiriske resultater som endrer konklusjonene som trekkes ved opprinnelig inndeling. Tabell 4 viser at \ln_pop_1920 fortsatt er positiv og signifikant ned til 1 % signifikansnivå, men variabelen har her lavere effekt. Dobling av befolkningsnivået vil øke veksten med 0.17 prosentpoeng mot

0.2 prosentpoeng ved opprinnelig regioninndeling. Den estimerte effekten regionenes størrelse har på befolkningsvekst har byttet fortegn og er ikke lengre statistisk signifikant som følge av den alternative regioninndelingen. Dette kan forklares med at man i hovedsak slår sammen regioner med lite areal rundt de store byene, slik at regionene blir mer like i areal.

5.5 Lokal polynomglatting og Spline-regresjon

For å studere sammenhengen mellom befolkningsvekst og initielt befolkningsnivå ved ulike grupperinger av befolkningsnivået benyttes et ikke-lineært kernel plot og en Spline-regresjon. Figur 6 viser en svakt avtakende sammenheng for den gruppen med lavest befolkningsnivå. For den mellomstore gruppen er linjen markant stigende. Regionene med det høyeste befolkningsnivået virker å ha en flat linje som indikerer lite sammenheng mellom befolkningsvekst og initielt befolkningsnivå for den største gruppen. Dette er et resultat som samsvarer med funnene i tabell 5.



Figur 6: Lokal polynomglatting.

Som beskrevet i delkapittel 3.3 deles regionene inn i tre grupper med like mange observasjoner i hver gruppe i tabell 5. Dette gjør at grensen for den minste gruppen er satt slik at logaritmisk befolkningsnivå i 1920 kan være opp til 9.55. Gruppe to har logaritmisk befolkningsnivå i intervallet mellom 9.55 og 10.06. Regionene med høyest befolkningsnivå er de regionene som i 1920 hadde logaritmisk befolkningsnivå

over 10.06.

Tabell 5: Piecewise-linear spline regresjon.

Årlig populasjons vekst 1920-2017	
$\ln_pop_1920 < 9.55$	-0.00338 (0.00223)
$\ln_pop_1920 \text{ } 9.55\text{-}10.06$	0.00824*** (0.00269)
$\ln_pop_1920 > 10.06$	0.00127 (0.00110)
Konstant	0.0339 (0.0206)
N	90
R^2	0.191

Robust standardavvik i parentes

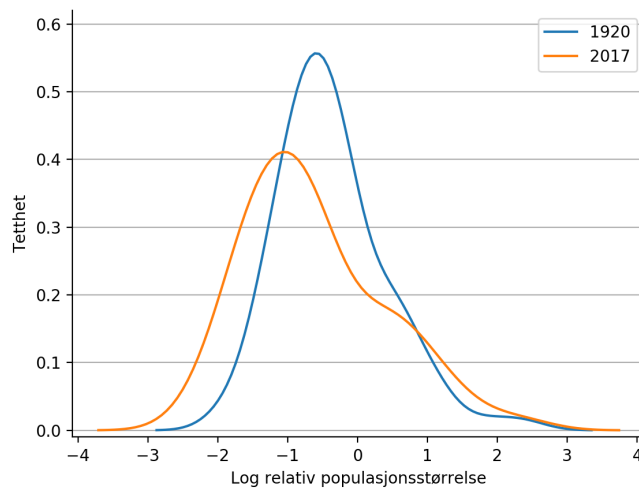
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Fra tabell 5 kan man lese at de lavest og høyest befolkede regionene ikke har en statistisk signifikant effekt mellom befolkningsvekst og befolkningsnivå. Fortegnene samsvarer imidlertid godt med linjen fra figur 6. For de mellomstore regionene derimot er det en statistisk signifikant effekt. Denne effekten er positiv og koeffisienten har en verdi på 0.0082. Denne effekten er større enn for regresjonen for absolutt divergens. Dette er resultater som stemmer overens med Rattsø og Stokke (2014a) og Michaels et al. (2012). Begge finner at den positive sammenhengen mellom befolkningsvekst og befolkningsnivå er størst for mellomstore regioner.

5.6 Kernel-tettheter

Kernel tetthetsfunksjoner estimeres for relativt befolkningsnivå i 1920 og 2017, som vist i figur 7. Den vannrette aksen er befolkningen i en region relativ til gjennomsnittet for alle regionene, målt i logaritmisk skala. Den vertikale aksen viser tettheten for de ulike relative befolkningsstørrelsene. Figur 7 viser at de to funksjonene har en entoppet fordeling, og det er størst tetthet av regioner med befolkning litt lavere enn befolkningsgjennomsnittet i regionene. Over perioden 1920-2017 flytter tettheten seg fra toppunktet og ut på begge sidene. Dette kan vi forstå med at det er

blitt flere lavt og høyt befolkede regioner i forhold til gjennomsnittlig befolkning i regionene. Dette stemmer overens med resultatene i Spline-regresjonen der vi så at de mellomstore regionene varierte mye i vekst slik at de største av disse regionene vil vokse raskt. De minste av de mellomstore regionene på sin side vil vokse noe saktere. Figur 7 indikerer at forskjellene mellom regionene sin befolkningsstørrelse har økt, som igjen er et tegn på populasjonsdivergens.



Figur 7: Kernel-tetthet for utvalget.

5.7 Markov-kjeder for etterkrigstida (1946-2016)

Markov-kjedene gir mer informasjon om utviklingen til regionene. Disse kan derfor øke forståelsen for divergensmønstret i befolkningsnivået. Datasettet som brukes i denne delen av analysen består av befolkningsnivå relativt til gjennomsnittet for 90 regioner i perioden 1946-2016, med overgangsperioder på fem år. Perioden velges fordi det kun er her det er sammenhengende data, men også fordi barroregresjonene viste at det er i denne perioden vi har sett signifikant befolkningsdivergens.

Tabell 6 viser at det ikke er noen bevegelse mot en ergodisk fordeling. Dette kommer av at det ikke er noen bevegelse ut av den minste gruppen og man kan ikke ha en likevekt der alle regionene er mindre enn gjennomsnittet. Dette krever at dynamikken i selve tabellen studeres nøye for å forstå resultatene. Dynamikken i tabellen viser at det for alle gruppene er mer sannsynlig å bevege seg en gruppe ned enn det

er å bevege seg en gruppe opp. Regioner med befolkningsnivå mellom 0.672 og 1.40 ganger gjennomsnittet befinner seg i gruppe 4. For disse regionene er sannsynligheten nesten 6% for å bevege seg til gruppe 3 i neste periode, mens det bare er litt lavere enn 2% sannsynlighet for at de beveger seg til den største gruppen i neste periode. Den samme dynamikken finner vi også i gruppe 2, hvor regionene har befolkningsnivå mellom 0.29 og 0.44 ganger gjennomsnittet. Disse regionene har over 7% sannsynlighet for å bevege seg til gruppen med lavest befolkningsnivå i neste periode, mens det bare er 1.59% sannsynlig at de klarer å komme seg opp til gruppe 3. Dette legger grunnlaget for en betydelig opphopning i den miste gruppen, hvor regionene har et befolkningsnivå lavere enn 0.29 ganger gjennomsnittlig regionalt befolkningsnivå. I tabell A.1 studeres en Markov-kjede der det brukes tiårige overgangsperioder. Dersom en variabel endres sakte vil det være mer sannsynlig å få en endring dersom overgangsperioden økes. Tabell A.1 viser den samme dynamikken. Det er fortsatt mer sannsynlig å gå ned en gruppe enn å gå opp en gruppe selv om overgangsperiodene er på ti år. Forskjellene er at sannsynlighetene for å endre gruppe er større ved tiårige overgangsperioder, men dette er å forvente når man øker overgangsperiodene. Resultatet er konsistent med kernel-tetthetsanalysen, hvor det også var klare tegn på at det var bevegelse mot flere regioner med lavt befolkningsnivå i forhold til gjennomsnittlig befolkningsnivå. Dette styrker mistanken om at perioden etter krigen har vært preget av divergens i regionalt befolkningsnivå.

Tabell 6: Overgangsperioder på fem år for befolkningsnivået i 90 regioner.

Periode t-1	Intervall	Overgangssannsynlighet					Obs.
		Periode t					
		1	2	3	4	5	
1	≤ 0.2936	100.00					252
2	$\langle 0.2936, 0.435 \rangle$	7.14	91.27	1.59			252
3	$\langle 0.435, 0.672 \rangle$		7.14	91.27	1.59		252
4	$\langle 0.672, 1.40 \rangle$			5.95	92.06	1.98	252
5	> 1.40				1.19	98.81	252
Initiell fordeling		20.0	20.0	20.0	20.0	20.0	

En svakhet med denne måten å gjøre analysen på er at dersom det har vært ekstraordinære hendelser i noen av de utvalgte årene vil disse kunne prege analysen. Dette kan unngås ved å benytte gjennomsnittets verdien for femårsperioden. Kostnaden ved dette vil være to overganger færre siden det ikke finnes data for årene 1947-1950. Denne kostnaden av to færre observasjoner er betydelig i flere av analysene. Robustheten til alle analysene blir derfor ikke sjekket ved bruk av gjennomsnittsperioder, men i tabell 6 er det derimot et betydelig antall overgangsperioder. Resultatene for Markov-kjedeanalysen med gjennomsnittsperioder finnes i tabell A.2. Når denne sammenlignes med tabell 6 viser det seg å ikke være store forskjeller på de to tabellene. Dette tolkes i retning av at enkelthendelser ikke har preget analysen i veldig stor grad. Det er likevel viktig å være klar over effekten enkelthendelser kan spille på resultatene der robustheten ikke er undersøkt.

I et forsøk på å få en Ergodisk fordeling, deles regionene inn i 3 grupper i stedet for de tradisjonelle 5 gruppene. Tabell 7 viser en ergodisk fordeling som styrker inntrykket fra tidligere analyser. Tabellen indikerer at dersom tendensen vi har sett i perioden 1946 til 2016 fortsetter, vil 73% av regionene ha et befolkningsnivå som er lavere enn 0.4 ganger snittet. Bare 6% av regionene vil ha et befolkningsnivå som er over 0.75 ganger av snittet. Dette vil bety at dersom utviklingen fortsetter, ender

vi opp med veldig få regioner som inneholder størsteparten av Norges befolkning.

Tabell 7: Femårsoverganger med tredeling av befolkningen i 90 regioner.

Periode t-1	Intervall	Overgangssannsynlighet			Obs.
		Periode t			
		1	2	3	
1	≤ 0.396	98.57	1.43		420
2	$(0.396, 0.753]$	5.00	94.52	0.48	420
3	> 0.753		1.67	98.33	420
Initiell fordeling		33.3	33.3	33.3	
Ergodisk fordeling		73.13	20.90	5.97	

5.7.1 Regioninndelingens betydning

Tabell 8 viser resultatene dersom inndelingen for regionene endres til inndelingen forslått av Bhuller (2009). Her er dynamikken annerledes enn i tabell 6. Mobiliteten oppover er mye lavere med Bhuller sin regioninndeling. Det er nå ingen regioner som har klart å gå fra gruppe 4 hvor regionene har befolkningsnivå mellom 0.60 og 0.99 ganger snittet, til den største gruppen hvor regionene har befolkningsnivå over 0.99 ganger snittet. Den eneste gruppen som har høyere sannsynlighet for å komme seg en gruppe opp er den minste gruppen. Regioner med befolkningsnivå lavere enn 0.3 ganger gjennomsnittet, har nå 0.79 % sannsynlighet for å komme seg opp til gruppe 2. Dynamikken i tabell 8 tyder på at områdene med befolkningsvekst større en gjennomsnittet nå er inkludert i regioner som allerede er store. Samtidig ser vi også med denne regioninndelingen at det er en sterk mobilitet nedover i regionstørrelse relativt til gjennomsnittlig befolkningsnivå. Denne måten å dele inn regionene på virker å gi enda sterkere indikasjoner på divergens i befolkningsutviklingen. Dette tolkes som at områdene rundt de store byene, som opprinnelig hadde relativt lavt befolkningsnivå har hatt en sterk befolkningsvekst. Når disse inkluderes i regioner der befolkningsnivået var høyt forsvinner deres bidrag til mobilitet oppover blant gruppene.

Tabell 8: Femårsoverganger med 45 regioner.

Periode t-1	Intervall	Overgangssannsynlighet					Obs.
		Periode t					
		1	2	3	4	5	
1	≤ 0.30	99.21	0.79				126
2	$\langle 0.30, 0.3905]$	10.32	88.89	0.79			126
3	$\langle 0.3905, 0.60]$		7.94	92.06			126
4	$\langle 0.60, 0.99]$			3.17	96.83		126
5	> 0.99					100.00	126
Initiell fordeling		20.0	20.0	20.0	20.0	20.0	

5.7.2 Forskjeller over tid

Barr regresjonene viste tegn på at det var forskjeller i divergens mellom tidsperiodene. Det er derfor interessant å sjekke om det også er signifikante forskjeller mellom periodene 1946-1981 og 1981-2016 når man benytter Markov-kjeder. En slik undersøkelse gir indikasjon på om divergensen har vært en ensartet prosess over tid. Dersom dette ikke er tilfelle, slik at det er signifikant forskjell på de to periodene, gir det ikke mening å beregne overgangssannsynligheter som skal gjelde for hele perioden. Analysen utføres ved å holde grensene for de forskjellige gruppene av befolkningsnivå konstante i begge periodene. Dette tillater at antallet i hver gruppe varierer i delutvalgene, men summen av observasjoner i delutvalgene vil være lik som i tabell 6.

Tabell 9 og 10 gir indikasjoner på at det har blitt vanskeligere å bevege seg oppover i relativ størrelse i den siste perioden. Samtidig virker det som det er blitt mer sannsynlig å bevege seg nedover. Det kan altså virke som det i siste periode har blitt større og større forskjell i innbyggertall mellom regionene. Dette tyder i likhet med barr regresjonene på at divergensen mellom regionene er tiltagende i styrke.

Tabell 9: Femårsoverganger for 90 regioner i perioden 1946-1981.

Periode t-1	Intervall	Overgangssannsynlighet					Obs.
		Periode t					
		1	2	3	4	5	
1	≤ 0.2936	100					105
2	$\langle 0.2936, 0.435 \rangle$	3.82	93.89	2.29			131
3	$\langle 0.435, 0.672 \rangle$		8.73	88.10	3.17		126
4	$\langle 0.672, 1.40 \rangle$			4.05	93.92	2.03	148
5	> 1.40				1.67	98.33	120
Initiell fordeling		16.67	20.79	20.00	23.49	19.05	

Tabell 10: Femårsoverganger for 90 regioner i perioden 1981-2016

Periode t-1	Intervall	Overgangssannsynlighet					Obs.
		Periode t					
		1	2	3	4	5	
1	≤ 0.2936	100.00					147
2	$\langle 0.2936, 0.435 \rangle$	10.74	88.43	0.83			121
3	$\langle 0.435, 0.672 \rangle$		5.56	94.44			126
4	$\langle 0.672, 1.40 \rangle$			8.65	89.42	1.92	104
5	> 1.40				0.76	99.24	132
Initiell fordeling		23.33	19.21	20.00	16.51	20.95	

For å teste om forskjellene mellom delutvalgene er signifikante, benyttes Q- og LR-tester. Nullhypotesen er her at dynamikken er lik i begge periodene. Resultatene til disse to testene er $Q = 13.17$ og $LR = 14.90$. Med syv frihetsgrader ved 5% signifikansnivå er kritisk verdi gitt som 14.07. Her vil altså LR-testen forkaste nullhypotesen mens Q-testen ikke vil forkaste nullhypotesen. Ved å studere bidragene til Q- og

LR-testene fra tabell 11 og 12 finner man at størst bidrag kommer fra spesielt to overganger. Overgangen fra gruppe 2 til 1 samt fra 3 til 4 står for over halvparten av verdien i Q-testen. For LR-testen ser vi at gruppe 2 og 3 bidrar mest også her. Det er altså de regionene med et mellomstort befolkningsnivå som har bidratt til forskjellene mellom de to periodene. Q- og LR-testen peker i retning av at divergensen mellom regioner har vært tiltagende over perioden etter krigen dersom man godtar et signifikansnivå litt over 5 % når man skal forkaste nullhypotesen.

Tabell 11: Person-test på tidsperioder.

		Bidrag til Person-test						
		Periode t						
	Periode t-1	Obs.	1	2	3	4	5	Sum
1946-1982	1	105	0					0
	2	131	2.02	0.10	0.40			2.52
	3	126		0.45	0.14	1.98		2.57
	4	148			0.90	0.06	0.001	0.961
	5	120				0.23	0.003	0.233
	Sum							6.284
1982-2017	1	147	0					0
	2	121	2.20	0.11	0.44			2.75
	3	126		0.44	0.14	2.00		2.58
	4	104			1.27	0.08	0.001	1.351
	5	132				0.21	0.002	0.212
	Sum							6.893
Person-teststatistikk (Q)								13.177
Kritisk verdi 5% signifikansnivå (syv grader av frihet)								14.07

Tabell 12: LR-test på tidsperioder

		Bidrag til LR-test						
		Periode t						
	Periode t-1	Obs.	1	2	3	4	5	Sum
1946-1982	1	105	0					0
	2	131	-3.13	3.48	1.09			1.44
	3	126		2.21	-3.92	2.76		1.05
	4	148			-2.31	2.78	0.07	0.54
	5	120				0.68	-0.57	0.11
	Sum							3.14
1982-2017	1	147	0					0
	2	121	5.31	-3.38	-0.65			1.28
	3	126		-1.75	4.06			2.31
	4	104			3.37	-2.71	-0.06	0.60
	5	132				-0.45	0.57	0.12
	Sum							4.31
LR-teststatistikk								14.90
Kritisk verdi 5% signifikansnivå (syv grader av frihet)								14.07

6 Konklusjon

Denne oppgaven har tatt for seg flere metoder for å studere migrasjonsstrømmene mellom norske regioner i løpet av perioden 1920-2017. Ved å studere hvordan andelen av befolkningen som bor i ulike regioner har endret seg ser man at de regionene som hadde en stor andel av befolkningen i 1920 har økt andelen i løpet av perioden.

Analysens hovedvekt legges på barroregresjoner. Disse viser tegn til at data er preget av absolutt divergens. Dette er et resultat som virker å være relativt uavhengig av valg av regioninndeling. Betingende barroregresjoner blir også benyttet, disse viser også tegn til divergens. Når det kontrolleres for naturlige og sosiale karakteristikk virker fortsatt data å være preget av divergens mellom regionene. Ingen av de faktorene det kontrolleres for virker å ha en signifikant effekt på befolkningsvekst. Denne oppgaven finner derfor ingen tegn til at det er noen karakteristikk som har hatt en positiv effekt på befolkningsvekst. Barroregresjonene utføres på en relativ lang tidsperiode. Perioden deles derfor opp i mindre delperioder. Det viser seg at det er perioden etter andre verdenskrig som har vært preget av divergens. Det kan også virke som divergensen er tiltagende da regresjonen for perioden 1982-2017 har større koeffisient og høyere forklaringskraft enn perioden 1946-1982.

Siden det ikke blir funnet signifikante variabler i de betingede barroregresjonene, og det heller ikke virker å være noen åpenbare instrument for befolkningsnivået i 1920, utvides analysen med Kernel-tettheter og Markov-kjeder. Disse støtter opp om at regionalt befolkningsnivå har vært preget av divergens. De finner begge tegn til dannelsen av en entoppet fordeling hvor største del av regionene ender opp med et lavt befolkningsnivå i forhold til gjennomsnittlig regionalt befolkningsnivå. Q- og LR-testene viser også at det er signifikante forskjeller over tid. Overgangssannsynlighetene viser at det har blitt mindre sannsynlig å bevege seg fra en gruppe av befolkningsnivå til en gruppe med høyere befolkningsnivå i perioden 1981 til 2016 i forhold til perioden 1946-1981. Dette støtter opp under funnet om divergens fra barroregresjonene.

Spline-regresjonen viser at det er ulik effekt for ulikt opprinnelig befolkningsnivå. Regresjonen indikerer at det er innad i gruppen for de mellomstore regionene det har

vært størst divergens. Dette er et resultat som også stemmer med funnene Rattsø og Stokke (2014a) gjør for norske regioner og Michaels et al. (2012) gjør for amerikanske regioner.

Duranton (2008) sin teori for bystørrelse gir støtte til analysen sine empiriske resultater. Teorien sier at regioner med høyt befolkningsnivå drar nytte av befolkningsnivået gjennom læring av hverandre, dette gjør at disse regionene har et høyere humankapitalnivå enn regioner med mindre befolkningsnivå. Derfor er lønnen høyere og følgelig befolkningsveksten høyere i disse regionene. At det er de mellomstore regionene som virker å være mest preget av divergens trenger ikke nødvendigvis å være en motsetning til Duranton (2008) sin teori. Bryter man ned teorien sier den at det som avgjør befolkningsvekst er forskjellen mellom agglomerasjonseffekter og trengselseffekter i en region. Dersom det i de største regionene nå har blitt så mange innbyggere at trengselseffektene har begynt å dominere agglomerasjonseffektene er det naturlig at befolkningsveksten avtar. Flytter vi oss fra teori til den virkelige verden har vi sett at Oslo og andre store norske byer har vært preget av en enorm boligprisvekst. Dette har nødvendigvis gjort levekostnadene tilsvarende høye i disse regionene. I den virkelige verden er det heller ikke restriksjoner på pendling, man kan altså bosette seg i en annen region enn der man arbeider. Dette kan ha gjort regionene rundt de største byene mer attraktive. Disse regionene kan dra nytte av agglomerasjonseffektene til naboregionen uten at de i like stor grad er preget av trengselseffektene. Altså virker alle empiriske resultater i denne analysene å være tilknyttet Duranton (2008) sin teori om at befolkningsrike områder har større befolkningsvekst enn mindre befolkningsrike områder.

Forslag til videre forskning vil være å studere det som virker å være en tiltagende grad av urbanisering. Det kunne også vært interessant og studert perioder lengre tilbake i tid, da manglende divergens for årene 1920-1946 potensielt kan være sterkt påvirket av 2. verdenskrig.

Referanser

- Ayuda, M.I., Collantes, F. & Pinilla, V. (2010). Long-run regional population disparities in europe during modern economic growth: a case study of spain. *The Annals of regional science*, 44(2), 273–295.
- Barro, R.J. & Sala-i Martin, X. (1992). Convergence. *Journal of political Economy*, 100(2), 223–251.
- Baumol, W.J. (1986). Productivity growth, convergence, and welfare: what the long-run data show. *The American Economic Review*, 1072–1085.
- Beeson, P. & Montgomery, E.B. (1993). The effects of colleges and universities on local labor markets. *Review of Economics and Statistics*, 75(4), 753–761.
- Beeson, P.E., DeJong, D.N. & Troesken, W. (2001). Population growth in us counties, 1840–1990. *Regional Science and Urban Economics*, 31(6), 669–699.
- Bernard, A.B. & Durlauf, S.N. (1996). Interpreting tests of the convergence hypothesis. *Journal of econometrics*, 71(1-2), 161–173.
- Bhuller, M. (2009). Inndeling av norge i arbeidsmarkedsregioner. *Statistisk Sentralbyrå (SSB). Notater*, 24.
- Bickenbach, F. & Bode, E. (2003). Evaluating the markov property in studies of economic convergence. *International Regional Science Review*, 26(3), 363–392.
- Borts, G.H. & Stein, J.L. (1964). Economic growth in a free market.
- De Long, J.B. (1988). Productivity growth, convergence, and welfare: comment. *The American Economic Review*, 78(5), 1138–1154.
- Desmet, K. & Rappaport, J. (2017). The settlement of the united states, 1800–2000: the long transition towards gibrat’s law. *Journal of Urban Economics*, 98, 50–68.
- Duranton, G. (2008). Are cities engines of growth and prosperity for developing countries? *Urbanization and growth*(3), 67–114.
- Fan, J., Gasser, T., Gijbels, I., Brockmann, M. & Engel, J. (1995). *On nonparametric estimation via local polynomial regression*. Institute of Statistics, Catholic University of Louvain.

- Glaeser, E.L., Scheinkman, J. & Shleifer, A. (1995). Economic growth in a cross-section of cities. *Journal of monetary economics*, 36(1), 117–143.
- Hustoft, A.G., Hartvedt, H., Nymoene, E., Stålnacke, M. & Utne, H. (1999). Standard for økonomiske regioner. *Statistics Norway, Report*, 99(6).
- Jianqing, F. & Gijbels, I. (1996). Local polynomial modelling and its applications. *Monographs on Statistics and Applied Probability. Chapman & Hall/CRC*.
- Kim, S. (1995). Expansion of markets and the geographic distribution of economic activities: the trends in us regional manufacturing structure, 1860–1987. *The Quarterly Journal of Economics*, 110(4), 881–908.
- Michaels, G., Rauch, F. & Redding, S. (2012). Urbanization and structural transformation. *The Quarterly Journal of Economics*, 127(2), 535–586.
- Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3), 1065–1076.
- Quah, D. (1993a). Empirical cross-section dynamics in economic growth. *European Economic Review*, 37, 426–434.
- Quah, D. (1993b). Galton’s fallacy and tests of the convergence hypothesis. *The Scandinavian journal of Economics*, 95(4), 427–443.
- Quah, D.T. (1996). Twin peaks: growth and convergence in models of distribution dynamics. *The economic journal*, 106(437), 1045–1055.
- Rattsø, J. & Stokke, H. (2014b). Regional convergence of income and education: Investigation of distribution dynamics. *Urban Studies*, 51(8), 1672–1685.
- Rattsø, J. & Stokke, H.E. (2014a). Population divergence and income convergence: Regional distribution dynamics for norway. *Regional Studies*, 48(11), 1884–1895.
- Shorrocks, A.F. (1978). The measurement of mobility. *Econometrica: Journal of the Econometric Society*, 1013–1024.
- Silverman, B.W. (2018). *Density estimation for statistics and data analysis*. Routledge.
- Simon, C.J. & Nardinelli, C. (2002). Human capital and the rise of american cities, 1900–1990. *Regional Science and Urban Economics*, 32(1), 59–96.
- Sokoloff, K.L. (1988). Inventive activity in early industrial america: evidence from patent records, 1790–1846. *The Journal of Economic History*, 48(4),

813–850.

StataCorp, L. (2013). mkspline—linear and restricted cubic spline construction.

Stata Glossary and Index release, 13.

Thune, T. (2019). *Norsk utdanningshistorie*.

https://snl.no/Norsk_utdanningshistorie. (hentet 01.02.2019)

Verbeek, M. (2012). *A guide to modern econometrics*. John Wiley & Sons.

Wooldridge, J.M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Zucchini, W., Berzel, A. & Nenadic, O. (2003). Applied smoothing techniques.

Part I: Kernel Density Estimation, 15.

A Tabeller

Tabell A.1: Tiårsoverganger for befolkningsnivået i 90 regioner.

Periode t-1	Intervall	Overgangssannsynlighet					Obs.
		Periode t					
		1	2	3	4	5	
1	≤ 0.2986	100.00					126
2	$\langle 0.2986, 0.44 \rangle$	14.29	82.54	3.17			126
3	$\langle 0.44, 0.679 \rangle$		15.87	80.95	3.17		126
4	$\langle 0.679, 1.40 \rangle$			11.11	84.92	3.97	126
5	> 1.40				2.38	97.62	126
Initiell fordeling		20.0	20.0	20.0	20.0	20.0	

Tabell A.2: Femårsoverganger med gjennomsnitt.

Periode t-1	Intervall	Overgangssannsynlighet					Obs.
		Periode t					
		1	2	3	4	5	
1	≤ 0.3025	100.00					216
2	$\langle 0.3025, 0.4295 \rangle$	8.33	89.81	1.85			216
3	$\langle 0.4295, 0.6615 \rangle$		6.94	91.67	1.39		216
4	$\langle 0.6615, 1.408665549 \rangle$			5.09	93.06	1.85	216
5	> 1.408665549				0.46	99.54	216
Initiell fordeling		20.0	20.0	20.0	20.0	20.0	