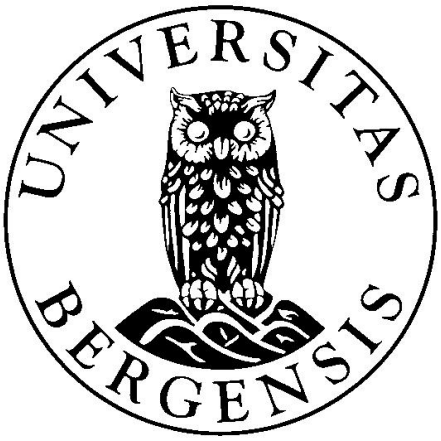
Modellering og estimering av romlig avhengighet i forsikring

Nikolai Sellereite

Masteroppgave i statistikk

Finansteori og forsikringsmatematikk



Universitetet i Bergen

Matematisk institutt

1.juni 2015

Sammendrag

De seneste årene har det blitt publisert flere studier hvor bayesianske hierarkiske modeller, med gitte spatiale avhengighetsstrukturer, blir foreslått som potensielle verktøy i forsikringsselskapers arbeid rettet mot geografisk prisdifferensiering. I denne oppgaven blir problemstillingen angrepet fra et frekventisk ståsted, hvor fokuset er begrenset til modeller for antall skader. Skadestørrelse, levetid o.l. er andre eksempler på responsvariabler hvor det teoretiske rammeverket kan anvendes. Parameterestimeringen blir utført ved hjelp av maksimum likelihood, og som følge av høydimensjonale integral introduseres Laplace-approksimasjon og automatisk derivasjon, hvor estimeringsprosedyren automatiseres ved hjelp av pakken Template Model Builder (TMB). Den latente spatiale effekten modelleres som Gaussian Markov random fields (GMRFs) med ulike valg av spatial avhengighetsstruktur. Modeller med og uten latente spatiale variabler tilpasses en simulert forsikringsportefølje, hvor modellene uten latente spatiale variabler tilsvarer generaliserte lineære modeller. Valideringen av den prediktive evnen til modellene blir utført ved å simulere 1000 nye forsikringsporteføljer.

Takk

Jeg ønsker først og fremst å takke min veileder, Hans J. Skaug, for all hjelp underveis med oppgaven. Jeg vil også rette en stor takk til de resterende foreleserne ved Matematisk institutt, mine medstudenter, og særlig Steffen Bjørgum Pedersen som jeg har delt kontor med de siste to årene. Det har vært fem utfordrende og hyggelige år ved Universitetet i Bergen.

Til slutt vil jeg også takke min familie for oppmuntrende ord i perioder hvor det var tungt å motivere seg.

Nikolai Sellereite

Innhold

Tabeller

Figurer

Kapittel 1

Introduksjon

De seneste årene har tilgangen på data hos forsikringsselskaper økt drastisk, samtidlig som datamaskinene har blitt langt kraftigere. Kombinasjonen av de to faktorene gjør at man er i stand til å anvende langt mer komplekse modeller enn tidligere, og i en svært konkurransepreget bransje vil dette kunne gi fordeler og konkurransemessige fortrinn ovenfor konkurrenter. Dette gjelder ikke bare det analytiske arbeidet knyttet til tariffanalyser innen skadeforsikring, men også områder som kundesegmentering, sensitivitetsanalyser vedrørende pris og ikke minst i forbindelse med oppgaver knyttet til reservering.

En meget aktuell problemstilling innenfor forsikring er hvordan man kan og bør håndtere geografisk lokalitet som risikofaktor. Det er ikke urimelig å tenke seg at dette er et område hvor potensialet er uforløst, og i langt større grad bør fokuseres på. Enkle søk viser at det er klare forskjeller mellom ulike forsikringsselskaper i hvordan regioner tildeles en risikoprofil (se tabell 1.1). Dette er trolig en kombinasjon av data tilgjengelig hos forsikringsselskapene, diverse markedstilpasninger og selve metodikken som blir anvendt i det analytiske arbeidet knyttet opp mot tariffanalyser.

I denne oppgaven blir problemstillingen vedrørende geografisk lokalitet som risikofaktor i forsikring angrepet fra et frekventisk ståsted, hvor fokuset er begrenset til modeller for antall skader. Dersom det endelige målet er en fullverdig tariff for et forsikringsprodukt ville det også være nødvendig å introdusere modeller for gjennomsnittlig skadestørrelse. Det teoretiske rammeverket som presenteres i oppgaven vil forøvrig også kunne anvendes på denne typen responsvariabler.

Oppgaven er strukturert på følgende vis. Delkapittel 1.1 tar for seg en generell metodikk forsikringsselskaper tar i bruk for å håndtere geografisk lokalitet som risikofaktor, og delkapittel 1.2 gir en kort oppsummering av tidligere studier knyttet opp mot geografisk

prisdifferensiering. Felles for de fleste studiene er at bayesianske hierarkiske modeller, hvor inferens utføres ved hjelp av MCMC-metoder, blir foreslått som potensielle modeller. Det teoretiske grunnlaget for oppgaven presenteres i kapittel 2, 3 og 4. Kapittel 2 gir en kort introduksjon til generaliserte lineære modeller (GLM), kapittel 3 tar for seg teorien rundt Gaussian Markov random fields (GMRFs), mens kapittel 4 presenterer aktuelle modeller for antall skader. Modellene i kapittel 4 tilpasses en simulert forsikringsportefølje, som forøvrig presenteres i kapittel 5. Estimering av parametere utføres ved hjelp av maksimum likelihood, sammenfattet i kapittel 6. Selve estimeringsprosedyren automatiseres ved hjelp av pakken Template Model Builder (TMB). I kapittel 7 tilpasses et knippe ulike modeller den simulerte forsikringsporteføljen i kapittel 5, med et klart fokus på hvilken effekt latente variabler har dersom man mangler forklaringsvariabler fra den sanne modellen. Problemstillingen ble også tatt for seg i Dimakos og Di Rattalma (2002). Kapittel 8 gir en kort oppsummering av oppgaven og resultater, samt mulighetene for videre studier, potensielle feilkilder og eventuelle begrensninger ved oppgaven.

I vedlegg A presenteres estimeringsprosedyren forbundet med GLM. Vedlegg B presenteres en rekke simuleringer fra GMRFs, med et klart fokus på parameteren δ. Vedlegg C presenterer R-koder for simulering av forsikringsporteføljen i kapittel 5, samt nødvendig kode knyttet til estimering av parametere ved hjelp av TMB. Til slutt er datagrunnlaget, hentet fra Statistisk Sentralbyrå (SSB), inkludert i vedlegg D.

1.1 Geografi som risikofaktor i forsikring

Geografisk lokalitet er en meget sentral risikofaktor ved flere typer forsikringsprodukter, og for et gitt forsikringsprodukt er det ofte store prisforskjeller knyttet til geografisk lokalitet blant forsikringsselskaper . Generaliserte lineære modeller (GLM) er hyppig brukt for å estimere effekten fra faktorer som alder, bilmerke, kjørelengde etc. på forventet risiko. Risiko tolkes her som en fellesbetegnelse for responsvariabler som gjennomsnittlig skadestørrelse, skadefrekvens, sannsynligheten for dødsfall o.l.

Hos finansportalen[1] er det mulig å få prisoverslag for alle slags typer forsikringer fra forskjellige forsikringsselskaper i Norge. Dersom man endrer adressen for et gitt tilbud, og holder alle andre faktorer like, ser man at geografisk lokalitet har ulik effekt på prisoverslagene hos forsikringsselskapene. I tabell 1.1 presenteres prisoverslag for en bilforsikring fra fem ulike forsikringsselskap i Norge, hvor område A og B tilsvarer adresser i henholdsvis Bergen og

[1] https://www.finansportalen.no/

Aalesund. Ikke overraskende er det forskjeller mellom de ulike selskapene, og selskap 5 skiller seg ut ved at prisoverslaget er lavere i Aalesund enn i Bergen. Det er verdt å merke seg at forskjellene trolig er en kombinasjon av markedstilpasninger og ulik estimert risiko. Det kan være at forsikringsselskap 5 ønsker å øke porteføljen sin i Møre og Romsdal, og dermed reduserer prisen flatt i denne regionen, i håp om at dette kan øke antall kunder.

Selskap	Bil	Egenandel	Kjørelengde	Område	Pris	Faktor
1	Golf 1.4 80hk	4000 kr	12000 km	A	kr 15.228	1.084
	Trendline			B	kr 16.502	
2	VW Golf 1.4 80hk,	4000 kr	12000 km	A	kr 15.550	1.051
	Bensin, 2WD, Manuell			B	kr 16.348	
3	Golf 1.4	4000 kr	12000 km	A	kr 16.306	1.105
	80/90 hk			B	kr 18.023	
4	Golf 1.4 80hk Trendline	4000 kr	12000 km	A	kr 16.990	1.029
	(Combi-Coupe)			B	kr 17.491	
5	Volkswagen Golf 1.4	3000 kr	12000 km	A	kr 21.677	0.911
	80hk Trendline			B	kr 19.752	

Tabell 1.1: Prisoversikt for bilforsikring (kasko), hvor område A tilsvarer Michael Kronhs gate 8, 5057 Bergen og område B tilsvarer Spjelkavikvegen 66A, 5010 Aalesund. Pristilbudene er hentet fra https://www.finansportalen.no/Forsikring/Bilforsikring 21.mars 2015, og er lagt ved for å illustrere at forsikringsselskap i Norge differensierer priser geografisk med ulike utfall.

Det vil ofte være tilfeller hvor man har lite eller ingen historikk hos enkelte geografiske regioner, noe som medfører at man har vanskeligheter med å tildele den geografiske regionen en fornuftig risikoprofil. I slike tilfeller er det vanlig å bruke såkalte glattingsmetoder for å tildele de geografiske regionene en passende risikoprofil. Metodene er basert på at man antar at områder som ligger nær hverandre har like egenskaper, og dermed lignende risikoprofiler. Dette er åpenbart en antagelse som slett ikke alltid er tilfredsstillende. Det kan være områder som ligger nær hverandre hvor kvaliteten på veisystemene er ulik, og det kan være egenskaper som byggeskikk, nedbør o.l. som varierer til tross for at områdene ligger nær hverandre. Selve topografien i Norge gjør dette til en svært aktuell problemstilling. Uavhengig av dette vil det være tilfeller hvor glattingsmetodene gir bedre resultater, til tross for at antagelsene

nødvendigvis ikke er helt passende, sammenlignet med metoder uten noen form for glatting.[2]

Før man introduserer glattingsmetodene er det nødvendig å fjerne effekten fra kjente faktorer som alder, kjørelengde, biltype, bilmerke o.l. Dette blir vanligvis gjort ved hjelp av GLM, hvor ulike faktorer (bortsett fra geografisk lokalitet) inkluderes, og analysegrunnlaget tilpasses modellen. I dette steget vil man få en indikasjon på hvilke faktorer som har en signifikant effekt på forventet risiko. Naturlig nok vil flere av de signifikante effektene allerede være kjent basert på tidligere analyser og erfaring. I det neste steget blir en valgt glattingsmetode introdusert og benyttet på de aggregerte residualene (forskjellen mellom faktiske data og estimerte utfall) innenfor valgte regioner. I delkapittel 1.2 er enkelte av glattingsmetodene nevnt.

Når residualene er glattet tildeles regionene en gitt risikoprofil basert på resultatene. Regionene kan være alt fra postkoder, kommuner, arbeidsmarkedsregioner o.l. Eksempelvis kan det tenkes at man bare ønsker tre ulike nivåer, gitt ved «lav risiko», «middels risiko» og «høy risiko». Dersom analysen er utført på kommunenivå, vil hver enkelt kommune bli tildelt en av de tre nevnte nivåene basert på resultatene fra glattingsmetoden. Kategoriseringen blir deretter inkludert i en GLM, sammen med de ikke-spatiale risikofaktorene, og man undersøker om de ulike nivåene har en signifikant effekt på responsvariabelen. Denne delen av analysen bør bli utført på et annet datasett enn det som ble brukt til å estimere de ikke-spatiale risikoparameterne. Årsaken til dette er at de kategoriske forklaringsvariablene knyttet til geografisk lokalitet vil kunne ha en overdrevet sterk forklaringskraft på det opprinnelige datasettet.

Forklaringskraften til modellen kan økes ytterligere ved å ta i bruk geodemografiske- og geofysiske data. Slike data er som oftest lett tilgjengelig og kan bli sett på som gjennomsnittlige egenskaper hos regioner. Det har vist seg at denne type data ofte har en signifikant og prediktiv forklaringskraft i seg. Dataene kan inkluderes i den opprinnelige modellen som en hvilken som helst forklaringsvariabel, og til slutt bli brukt som en selvstendig risikofaktor eller inkluderes som en del av den geografiske risikofaktoren. Dersom man velger å inkludere denne type informasjon i modellen er man nødt til å ta stilling til om hvorvidt dataene skal bli inkludert som forklaringsvariabler i det første steget av analysen. Rent intuitivt kan det tenkes at dette vil medføre at de aggregerte residualene vil få en tilnærmet glatt fordeling. Dersom dette ikke er tilfellet vil ikke glattingsmetodene være like aktuelle, og det vil være mer naturlig å vente med å inkludere dataene. Dette vil bli gjort i det siste steget av analysen, hvor residualene

[2] http://www.theactuary.com/archive/old-articles/part-5/geographical-spatial-analysis-in-general-insurance-pricing/

allerede er glattet og det geografiske området er inndelt i ulike soner med lik risikoprofil.

I en realistisk situasjon vil modellene ovenfor bare være en del av selve prisingsprosessen. Det vil også være lønnsomt å kunne forutse hvilken effekt en eventuell prisendring vil ha på porteføljen. Vil en økning i prisen resultere i at kunder forlater forsikringsselskapet? Vil en nedgang i prisen føre til nye kunder? Dette er svært sentrale og aktuelle spørsmål hos forsikringsselskaper. Det er også verdt å merke seg at forsikringsselskaper ofte benytter seg av markedstilpasninger i områder hvor konkurransen om kundene er stor.

1.2 Tidligere studier

De seneste årene har det blitt publisert flere studier som tar stilling til hvordan man kan håndtere geografisk lokalitet som risikofaktor i en forsikringsrelatert problemstilling. Hovedformålet med studiene er å analysere hvilke statistiske metoder som er best egnet til å avdekke hvordan den «sanne» risikoen fordeler seg innenfor et geografisk område. Det geografiske området deles inn i mindre regioner, og hovedtanken er at regioner som ligger nær hverandre har like egenskaper. Ved å inkludere en avhengighetsstruktur mellom regionene er man ofte i stand til å estimere risikoen innenfor det geografiske området på en mer tilfredsstillende måte. Sett fra et forsikringsrepresentativt ståsted vil dette kunne avdekke interessante resultater, og de statistiske metodene kan være nyttig både i tariff- og kundeanalyser.

En av de første publiserte studiene er Taylor (1989). Studien baseres på en portefølje bestående av husholdningsforsikringer, og det geografiske området deles inn ved hjelp av postnummer. Postnummer j identifiseres ved koordinatene (x_j, y_j). Hovedformålet med studien er å estimere risikoen som antas å være proporsjonal med en ukjent funksjon $I(x, y)$. Analysen blir utført ved hjelp av bivariate splines, som blir tilpasset ved hjelp av regresjon. Postnumrene blir til slutt kategorisert på grunnlag av den estimerte risikoen, og i en bestemt kategori vil ikke risikopremien varierer mht. postnummer gitt at de resterende forklaringsvariablene er like. Resultatene blir, ved hjelp av kartrepresentasjon, sammenlignet med hvordan et spesifikt forsikringsselskap har estimert risikoen innenfor det geografiske området.

Et alternativt løsningsforslag til problemstillingen gitt i Taylor (1989) er presentert av Boskov og Verrall (1994). Modellen er utviklet basert på teorien rundt bayesianske spatiale modeller, og parameterne blir estimert ved hjelp av Gibbs samplingsalgoritme. Det antas at datasettet er standardisert, og man analyserer hvilken effekt forklaringsvariabelen

postnummmer har på responsvariabelen skadeprosent[3]. Datasettet fra Taylor (1989) blir reanalysert og resultatene bekrefter langt på vei at valget av statistisk modell vil være avgjørende for det endelige resultatet. Boskov og Verrall (1994) argumenterer for at man bør analysere effekten av geografisk lokalitet på skadeantall og skadeerstatninger hver for seg, som følge av at den estimerte effekten hos geografiske regioner som har få poliser vil være svært sensitiv for enkeltskader med store erstatningsbeløp.

Dimakos og Di Rattalma (2002) tar også utgangspunkt i bayesiansk statistikk og bygger statistiske modeller, hvor responsvariablene tilsvarer skadeantall og skadeerstatninger, som kan kompensere for manglende informasjon om ulike forklaringsvariabler. Modellene er basert på teorien om bayesianske hierarkiske modeller, hvor inferens blir utført ved Markov-Chain Monte Carlo (MCMC). Modellene tilpasses både et simulert datasett og en porteføljе bestående av bilforsikringer begrenset til biltyveri. Resultatene viser at samspillet mellom spatiale latente variabler kan øke prediksjonskraften betraktelig, men dersom det er få observasjoner vil ikke prediksjonene avvike spesielt sammenlignet med generaliserte lineære modeller. Det argumenteres også for at dersom latente variabler skal kompensere for manglende forklaringsvariabler må de manglende forklaringsvariablene ha en relativt glatt geografisk fordeling.

Márkus et al. (2010) utvikler en modell hvor det nok en gang blir tatt utgangspunkt i bayesianske hierarkiske modeller. Formålet med studien er å predikere antall skader innenfor hvert enkelt postnummer i Ungarn for en gitt type skade som dekkes av en husholdningsforsikring. Fordelingen til antall skader antas å tilfredsstille egenskapene til en ikke-homogen Poissonprosess, og den spatiale avhengigheten mellom postnumrene modelleres ved Gaussian Markov random fields (GMRFs). Parameterne i modellen estimeres ved hjelp MCMC, og det utvikles i tillegg en prosedyre for å optimalisere akseptraten når man oppdaterer Markov-kjeden.

En aktuell problemstilling i forsikringsbransjen er hvordan man på best mulig måte kan håndtere de pågående klimaendringene. Scheel et al. (2013) utvikler en bayesiansk hierarkisk modell hvor målet er å kunne forklare og predikere frekvensen av forsikringsutbetalinger som er et direkte resultat av klimarelaterte hendelser i Norge. Frekvensen av forsikringsutbetalinger blir modellert ved å kombinere generaliserte lineære modeller med geografisk glattet variabel seleksjon. Modellen tilpasses reelle data ved hjelp av MCMC og Gibbs samplingsalgoritme, og valideres ved presise out-of-sample prediksjoner.

[3]Skadeprosent: Inntrufne skader i prosent av opptjent premie.

Thorsen (2012) analyserte hvordan ulike hydrologiske og meteorologiske forhold påvirker skadefrekvenser for vannskader på norske boliger ved hjelp av generaliserte lineære modeller (GLM). I startfasen ble geografiske kjennetegn ignorert for å avdekke en rimelig modellformulering for en del ikke-romlige sammenhenger. Videre ble det påvist at forklaringskraften økte betraktelig når geografiske kjennetegn ble inkludert i modeller med regionspesifikke konstantledd. Det ble også påvist at responsen på variasjoner i klimarelaterte risikofaktorer varierer systematisk både mellom fylker og arbeidsmarkedsregioner, og internt mellom kommunene i slike regioner. Hovedformålet var å teste om det var romlige avhengigheter i responsen på variasjoner i risikofaktorene. Dette ble gjort ved hjelp av en betinget autoregressiv romlig respons (CAR-modell). En slik modell korrigerer for muligheten for at nabosoner har tilsvarende respons på blant annet store nedbørsmengder, for eksempel som følge av ensartet jordsmonn, topografi og byggeskikk. Estimeringen dokumenterte klart signifikante nabolagseffekter, samt at enkelte deler av landet har en markert større evne enn andre til å absorbere store nedbørsmengder.

Kapittel 2

Generaliserte lineære modeller

2.1 Introduksjon

Generaliserte lineære modeller ble introdusert av Nelder og Wedderburn (1972), og er en fleksibel generalisering av ordinær lineær regresjon. Rammeverket gjør oss i stand til å modellere forholdet mellom variabler hvor responsvariablene ikke nødvendigvis er normalfordelte, og sammenhengen mellom responsvariablene og forklaringsvariablene er ikke begrenset til å være lineær. Generaliserte lineære modeller blir anvendt innefor flere områder, og er et viktig verktøy for forsikringsselskapers arbeid rettet mot prising av forsikringsprodukter.

2.2 Generaliserte lineære modeller

Den ordinære lineære regresjonsmodellen er definert ved

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{2.1}$$

hvor $\mathbf{y} = [Y_1, \ldots, Y_N]^T$, $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$, $\mathbf{I}$ er en $N \times N$ identitetsmatrise og $\sigma^2\mathbf{I}$ er kovariansmatrisen til $\boldsymbol{\epsilon}$ som er multivariat normalfordelt. $\mathbf{X}$ er en kjent $N \times p$ matrise hvor $\mathbf{x}_i^T$ tilsvarer rad i hos $\mathbf{X}$ og $\boldsymbol{\beta}$ er en $p \times 1$ vektor av ukjente parametere slik at $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$. Den lineære prediktoren i (2.1) er gitt ved

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{2.2}$$

hvor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$. Årsaken til at det kalles en lineær prediktor er nettopp fordi $\mathbf{x}_i^T\boldsymbol{\beta}$ er en lineær kombinasjon av de ukjente parameterne $\beta_1, \ldots, \beta_p$.

Generaliserte lineære modeller er en generalisering av ordinær lineær regresjon basert på følgende to punkter:

1. I den klassiske lineære modellen antas det at $Y_1, \ldots, Y_N$ er uavhengige og normalfordelt med identisk varians lik σ^2. I en generalisert lineær modell er det tilstrekkelig å anta at fordelingen til $Y_1, \ldots, Y_N$ tilhører eksponentielle familier. Dette medfører at man kan modellere responsvariabler hvor gammafordeling, poissonfordeling og binomialfordeling er mer naturlige valg.

2. Fremfor å modellere den lineære sammenhengen $E[\mathbf{y}] = \boldsymbol{\mu}$ direkte som en funksjon av den lineære prediktoren $\boldsymbol{\eta}$ kan man modellere en funksjon $g(\boldsymbol{\mu})$, slik at

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \tag{2.3}$$

Som et resultat av antagelsene behøver ikke responsvariablene å ha identisk varians, effekten av de ukjente parameterne kan være ikke-lineær og utfallsrommet av de predikerte verdiene kan begrenses.

Generaliserte lineære modeller er, som tidligere nevnt, et viktig verktøy i det analytiske arbeidet hos forsikringsselskaper. En av årsakene til dette er at antagelsen om normalfordelte responsvariabler svært ofte ikke tilfredsstillende når man analyserer forsikringsdata. Statistisk modellering av skadestørrelse, skadefrekvens og sannsynligheten for at det inntreffer en skade hos en enkelt polise er eksempler på hvor antagelsen om normalfordelte responsvariabler ikke er rimelig.

2.3 Eksponentielle familier

Eksponentielle familier er en samling av fordelinger hvor sannsynlighetstettheten til Y kan skrives på formen

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{2.4}$$

hvor $a(\cdot)$, $b(\cdot)$ og $c(\cdot)$ er kjente funksjoner. θ tilsvarer den kanoniske parameteren i fordelingen, og er en funksjon av lokasjonsparameteren, og ϕ kalles dispersjonsparameteren. I de neste avsnittene blir det vist at poissonfordelingen, normalfordelingen og binomialfordelingen hører til eksponentielle familier.

2.3.1 Poissonfordeling

Sannsynlighetstettheten til en poissonfordelt variabel Y er gitt ved:

$$f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\left\{y\log(\lambda) - \lambda - \log(y!)\right\}, \tag{2.5}$$

hvor $y = 0, 1, \ldots,$ og $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$. Sammenlignet med (2.4) kan man slå fast at $\theta = \log(\lambda)$, og dermed er $\lambda = \mathrm{e}^{\theta}$. Innsatt i (2.4) får man følgelig:

$$f(y;\theta) = \exp\Big\{y\theta - \mathrm{e}^{\theta} - \log(y!)\Big\}. \tag{2.6}$$

Poissonfordelingen er altså et spesialtilfelle av (2.4), hvor $\theta = \log(\lambda)$, $b(\theta) = \exp(\theta)$, $a(\phi) = 1$ og $c(y,\phi) = -\log(y!)$, og tilhører dermed eksponentielle familier.

2.3.2 Binomisk fordeling

Sannsynlighetstettheten til en variabel Y som er binomisk fordelt med parameter π og antall forsøk lik n er gitt ved:

$$\begin{aligned} f(y;\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= \binom{n}{y}\left(\frac{\pi}{1-\pi}\right)^y(1-\pi)^n \\ &= \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}\right\} \end{aligned} \tag{2.7}$$

hvor $y = 0, 1, \ldots, n$, $\mathrm{E}(Y) = n\pi$ og $\mathrm{Var}(Y) = n\pi(1-\pi)$. Sammenlignet med (2.4) kan man slå fast at $\theta = \log\left(\frac{\pi}{1-\pi}\right)$. Dette medfører at $\pi = \frac{\mathrm{e}^{\theta}}{1+\mathrm{e}^{\theta}}$, og innsatt i (2.4) får man:

$$f(y;\pi) = \exp\left\{y\theta - n\log(1+\mathrm{e}^{\theta}) + \log\binom{n}{y}\right\} \tag{2.8}$$

Binomialfordelingen er et spesialtilfelle av (2.4), hvor $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = n\log(1+\mathrm{e}^{\theta})$, $c(y,\phi) = \log\binom{n}{y}$ og $a(\phi) = 1$, og tilhører dermed eksponentielle familier.

2.3.3 Normalfordeling

Sannsynlighetstettheten til en variabel Y som er normalfordelt med parametere μ og σ er gitt ved:

$$\begin{aligned} f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\} \\ &= \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\} \end{aligned} \tag{2.9}$$

hvor $y \in \mathbb{R}$, $\mathrm{E}(Y) = \mu$ og $\mathrm{Var}(Y) = \sigma^2$. Sammenlignet med (2.4) kan man slå fast at $\theta = \mu$ og $\phi = \sigma^2$, og konkludere med at normalfordelingen er et spesialtilfelle av (2.4) med $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\phi} + \log\left(2\pi\phi\right)\right]$ og $a(\phi) = \phi$.

2.3.4 Egenskaper hos eksponentielle familier

Det er en rekke gunstige egenskaper hos fordelinger i eksponentielle familier. Funksjonen $b(\theta)$ har en sentral rolle, og bestemmer sammenhengen mellom varians og forventning. Dersom tetthetsfunksjonen til Y kan skrives på formen (2.4) er følgende tilfredsstilt:

- $\mathrm{E}(Y) = b'(\theta)$
- $\mathrm{Var}(Y) = a(\phi)b''(\theta)$

For å vise dette er nødvendig med noen generelle resultater basert på generell matematisk og statistisk teori. Fremgangsmåten er tilsvarende for både kontinuerlige og diskrete stokastiske variabler, men i dette tilfellet antas det at Y er kontinuerlig. Integralene skal tolkes slik at man integrerer over y hvor $f(y;\mathbf{\Theta}) > 0$.

For en kontinuerlig, stokastisk variabel Y med parametervektor $\mathbf{\Theta}$ er følgende tilfredsstilt per definisjon:

$$\int f(y;\mathbf{\Theta})\,\mathrm{d}y = 1. \tag{2.10}$$

Dersom $f(y;\mathbf{\Theta})$ i tillegg kan skrives på formen (2.4) vil

$$\frac{\partial}{\partial\theta}f(y;\theta,\phi) = \frac{y - \mathrm{b}'(\theta)}{a(\phi)}f(y;\theta,\phi) \tag{2.11}$$

$$\frac{\partial^2}{\partial\theta^2}f(y;\theta,\phi) = \left[\frac{y - \mathrm{b}'(\theta)}{a(\phi)}\right]^2 f(y;\theta,\phi) - \frac{\mathrm{b}''(\theta)}{a(\phi)}f(y;\theta,\phi) \tag{2.12}$$

Som et direkte resultat av (2.10) vil følgende være tilfredsstilt:

$$\frac{\partial}{\partial\theta}\int f(y;\theta)\,\mathrm{d}y = 0 \tag{2.13}$$

$$\frac{\partial^2}{\partial\theta^2}\int f(y;\theta)\,\mathrm{d}y = 0. \tag{2.14}$$

Videre antas det at det er lov å flytte $\frac{\partial}{\partial\theta}$ og $\frac{\partial^2}{\partial\theta^2}$ innenfor integraltegnet og bevisene for de to punktene er derfor som følger:

Bevis: $E(Y) = b'(\theta)$.

$$\begin{aligned}
0 &= \int \frac{\partial}{\partial\theta} f(y;\theta)\,\mathrm{d}y \\
&= \int \frac{y - \mathrm{b}'(\theta)}{a(\phi)} f(y;\theta)\,\mathrm{d}y \\
&= \frac{1}{a(\phi)}\left[\int y f(y;\theta)\,\mathrm{d}y - \mathrm{b}'(\theta)\int f(y;\theta)\,\mathrm{d}y\right] \\
&= \frac{1}{a(\phi)}\left[\mathrm{E}(Y) - b'(\theta)\right] \\
&\Longrightarrow \mathrm{E}(Y) = b'(\theta)
\end{aligned} \tag{2.15}$$

□

Bevis: $Var(Y) = a(\phi)b''(\theta)$.

$$\begin{aligned}
0 &= \int \frac{\partial^2}{\partial\theta^2} f(y;\theta)\,\mathrm{d}y \\
&= \int \left[\frac{y - \mathrm{b}'(\theta)}{a(\phi)}\right]^2 f(y;\theta)\,\mathrm{d}y - \int \frac{\mathrm{b}''(\theta)}{a(\phi)} f(y;\theta)\,\mathrm{d}y \\
&= \frac{1}{a(\phi)^2}\int \{y - \mathrm{E}(Y)\}^2 f(y;\theta)\,\mathrm{d}y - \frac{\mathrm{b}''(\theta)}{a(\phi)}\int f(y;\theta)\,\mathrm{d}y \\
&= \frac{\mathrm{Var}(Y)}{a(\phi)^2} - \frac{\mathrm{b}''(\theta)}{a(\phi)} \\
&\Longrightarrow \mathrm{Var}(Y) = a(\phi)b''(\theta)
\end{aligned} \tag{2.16}$$

□

Variansen til Y er et produkt av faktorene $a(\phi)$ og $b''(\theta)$. Parameteren ϕ tilsvarer, som tidligere nevnt, dispersjonsparameteren og funksjonen $b''(\theta)$ kalles variansfunksjonen. Det er også vanlig å skrive $b''(\theta) = V(\mu)$, der hensikten er å vise hvordan forventningen påvirker variansen direkte. Dette kommer som et resultat av (2.15) som viser at forventningen til Y er en funksjon av θ.

2.4 Den lineære prediktoren

Den lineære prediktoren, $\boldsymbol{\eta}$, er en lineær kombinasjon av de ukjente parameterne gitt ved $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$. Koeffisientene tilsvarer $\mathbf{X}$, slik at

$$\eta_i = x_i^T \boldsymbol{\beta} = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \tag{2.17}$$

Matrisen $\mathbf{X}$ blir kalt designmatrisen og inneholder kjente forklaringsvariabler til de realiserte observasjonene gitt ved $\mathbf{y} = [y_1, \ldots, y_N]^T$. Forklaringsvariablene kan være både kontinuerlige, diskrete og kategoriske. Det er typisk med dummyvariabler som forklaringsvariabler i variansanalyser (ANOVA) hvor man, eksempelvis, ønsker å undersøke om det er forskjeller mellom ulike populasjoner.

2.5 Linkfunksjonen

Linkfunksjonen $g(\cdot)$ er en funksjon som relaterer forventningen til responsvariablene $\mathbf{y}$ til designmatrisen $\mathbf{X}$ slik at $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. Linkfunksjonen må være monoton og differensierbar. Dette medfører at den inverse funksjonen $g^{-1}(\cdot)$ eksisterer, slik at $g^{-1}(\mathbf{X}\boldsymbol{\beta}) = g^{-1}(g(\boldsymbol{\mu})) = \boldsymbol{\mu}$. Valget av linkfunksjon avhenger av hvilke antagelser man gjør når det gjelder responsvariablenes fordeling. Dersom man modellerer responsvariabler som er begrenset til å være positive, for eksempel skadestørrelse og skadefrekvens, vil $g(\cdot) = \log(\cdot)$ være et fornuftig og naturlig valg. Tilsvarende vil linkfunksjonen $\log(\frac{\mu}{1-\mu})$ være passende dersom responsvariablene er begrenset til verdier i intervallet $[0, 1]$.

Enkelte linkfunksjoner er naturlige valg for gitte fordelinger, og går under navnet kanoniske linkfunksjoner. Dette er funksjoner som transformerer forventningen slik at $g(\mu) = \theta$. For en variabel Y som er poissonfordelt har man fra (2.5) at $\theta = \log(\mu)$, og den kanoniske linkfunksjonen er dermed lik $\log(\cdot)$. Dersom Y er normalfordelt forteller (2.9) at $\theta = \mu$ og den kanoniske linkfunksjonen er lik identitetsfunksjonen. Tilsvarende vil den kanoniske linkfunksjonen være gitt ved $\log(\frac{\mu}{1-\mu})$ dersom Y er binomisk fordelt som følge av (2.7).

2.6 Estimering av parametere

Antar at man har en sekvens med uavhengige variabler $Y_1, \ldots, Y_N$ som tilfredsstiller betingelsene til en generalisert lineær modell gitt i delkapittel 2.2. Målet er å estimere de ukjente parameterne $\beta_1, \ldots, \beta_p$ som er relatert til Y_i-ene gjennom $\mathrm{E}(Y_i) = \mu_i$ og $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$. Dette blir gjort ved hjelp av den iterative ligningen gitt ved:

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}. \tag{2.18}$$

$\mathbf{z}$ er en $N \times 1$ vektor og elementene i $\mathbf{z}$ er gitt ved

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right), \tag{2.19}$$

hvor μ_i og $\frac{\partial \eta_i}{\partial \mu_i}$ er evaluert ved $\mathbf{b}^{(m-1)}$. $\mathbf{W}$ er en $N \times N$ diagonalmatrise med elementer lik

$$w_{ii} = \frac{1}{\mathrm{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right). \tag{2.20}$$

Dobson og Barnett (2011) konkluderer med at den iterative ligningen gitt ved (2.18) er på samme form som normalligningene for en modell gitt ved (2.1), med unntak at den må løses iterativt som følge av $\mathbf{z}$ og $\mathbf{W}$ blir bestemt av $\mathbf{b}$. Som et resultat av dette blir metoden for å finne maksimum likelihood estimatorer for generaliserte lineære modeller ofte kalt iterative vektet minste kvadraters metode. Estimatene for parameterne blir dermed funnet numerisk ved å velge initialverdier lik $\mathbf{b}^{(0)}$, evaluere $\mathbf{z}$ og $\mathbf{W}$, og dersom estimatene konvergerer, avslutte den iterative ligningen når forskjellen mellom $\mathbf{b}^{(m)}$ og $\mathbf{b}^{(m-1)}$ er tilstrekkelig liten.

For nærmere beskrivelse og utledelse av den iterative ligningen se vedlegg A. Der vil også en annen velkjent metode, vektet minste kvadraters metode for multippel lineær regresjon, bli presentert i sin helhet.

Kapittel 3

Modellering av romlig avhengighet

Dette kapittelet tar for seg teori og anvendelse av Gaussian Markov random fields (GMRFs). Rue og Held (2005) er en meget god introduksjon til GMRFs og for en mer detaljert gjennomgang anbefales den på det sterkeste. Som Rue og Held (2005) introduserer teorien med:

> «*A GMRF is really a simple construct: It is just a (finite-dimensional) random vector following a multivariate normal (or Gaussian) distribution.*»

3.1 Gaussian Markov random fields (GMRFs)

For å kunne forstå og utnytte fordelene til GMRFs er det nødvendig å ha en god forståelse for betinget uavhengighet mellom stokastiske variabler. Lar $\mathbf{x} = [X_1, X_2, X_3]^T$ representere en stokastisk vektor, hvor simultantettheten er gitt ved $\pi(\mathbf{x})$. X_1 og X_2 er betinget uavhengig av hverandre gitt X_3 hvis og bare hvis $\pi(\mathbf{x})$ tilfredsstiller

$$\pi(\mathbf{x}) = \pi(x_1|x_3)\pi(x_2|x_3)\pi(x_3). \tag{3.1}$$

Ligning (3.1) er en forenkling av det generelle uttrykket $\pi(\mathbf{x}) = \pi(x_1|x_2, x_3)\pi(x_2|x_3)\pi(x_3)$, og tolkningen er at dersom X_3 allerede er kjent bidrar ikke X_2 med noe mer informasjon i forhold til X_1. Definisjonen om betinget uavhengighet kan enkelt overføres til multivariate fordelinger, hvor $\mathbf{x}$ og $\mathbf{y}$ er betinget uavhengig gitt $\mathbf{z}$ hvis og bare hvis

$$\pi(\mathbf{x}, \mathbf{y}|\mathbf{z}) = \pi(\mathbf{x}|\mathbf{z})\pi(\mathbf{y}|\mathbf{z}). \tag{3.2}$$

Betinget uavhengighet vil heretter bli presisert ved notasjonen $\mathbf{x} \perp \mathbf{y}|\mathbf{z}$. Det er ellers viktig å være klar over at selv om to variabler er betinget uavhengig, er det fullt mulig at de er marginalt avhengige.

3.1.1 Eksempler

Et enkelt eksempel på GMRFs er gitt ved en førsteordens autoregressiv prosess med uavhengige standard normalfordelte støyledd uttrykt ved

$$\phi_t = \zeta\phi_{t-1} + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad |\zeta| < 1 \tag{3.3}$$

hvor t representerer tid. I dette tilfellet er ikke antagelsene om betinget uavhengighet eksplisitt definert, men kan enkelt uttrykkes ved

$$\phi_t|\phi_1,\ldots,\phi_{t-1} \sim \mathcal{N}(\zeta\phi_{t-1}, 1) \tag{3.4}$$

for $t = 2,\ldots,n$. Dette medfører at ϕ_s og ϕ_t for $1 \leq s < t \leq n$ er betinget uavhengige gitt $\{\phi_{s+1},\ldots,\phi_{t-1}\}$ dersom $t - s > 1$. Dersom det i tillegg antas at $\phi_1 \sim \mathcal{N}\left(0, \frac{1}{1-\zeta^2}\right)$ vil simultantettheten til $\boldsymbol{\phi} = [\phi_1,\ldots,\phi_n]^T$ tilfredsstille

$$\pi(\boldsymbol{\phi}) = \frac{1}{(2\pi)^{n/2}}|\mathbf{Q}|^{1/2}\exp\left\{-\frac{1}{2}\boldsymbol{\phi}^T\mathbf{Q}\,\boldsymbol{\phi}\right\}, \tag{3.5}$$

som et resultat av $\pi(\boldsymbol{\phi}) = \pi(\phi_1)\pi(\phi_2|\phi_1)\ldots\pi(\phi_n|\phi_{n-1})$.

Dersom man sammenligner (3.5) med tetthetsfunksjonen til en multivariat normalfordeling ser man at $\boldsymbol{\phi} \sim \mathcal{N}\left(\boldsymbol{\mu}, \sum_{\boldsymbol{\phi}}\right)$, hvor $\boldsymbol{\mu} = \mathbf{0}$ og $\sum_{\boldsymbol{\phi}} = \mathbf{Q}^{-1}$. Matrisen $\mathbf{Q}$ kalles presisjonsmatrisen, og er i dette tilfellet lik den tridiagonale matrisen

$$\mathbf{Q} = \begin{pmatrix} 1 & -\zeta & & & & \\ -\zeta & 1+\zeta^2 & -\zeta & & & \\ & \ddots & \ddots & \ddots & & \\ & & & -\zeta & 1+\zeta^2 & -\zeta \\ & & & & -\zeta & 1 \end{pmatrix} \tag{3.6}$$

Årsaken til at matrisen er tridiagonal er at ϕ_i og ϕ_j er betinget uavhengig gitt $\boldsymbol{\phi}_{-ij}$ dersom $|i-j| > 1$. Presisjonsmatrisen $\mathbf{Q}$ representerer den betingede avhengighetsstrukturen mellom variablene, mens kovariansmatrisen $\sum_{\boldsymbol{\phi}} = \mathbf{Q}^{-1}$ gir tilsvarende informasjon om den marginale avhengigheten mellom variablene. Dette forklares nærmere i seksjon 3.1.2.

Et annet og mer generelt eksempel på GMRFs er gitt ved

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \mathcal{N}\left(\sum_{j:j\neq i}\beta_{ij}\phi_j, \kappa_i^{-1}\right), \quad i = 1,\ldots,n, \tag{3.7}$$

hvor $\boldsymbol{\phi}_{-i}$ tilsvarer alle elementer i $\boldsymbol{\phi}$ bortsett fra ϕ_i.

I dette tilfellet er det ikke en naturlig rekkefølge mellom variablene slik som i (3.3), og den vanlige fremgangsmåten er å definere en multivariat normalfordeling, hvor $\boldsymbol{\mu} = \mathbf{0}$ og $\sum = \mathbf{Q}^{-1}$, som tilfredsstiller ligningene gitt ved (3.7). Denne modellen ble introdusert av Besag (1974), og går under navnet betingede autoregressive prosesser (CAR). Parameterne i (3.7) må også tilfredsstille enkelte krav for at $\pi(\boldsymbol{\phi})$ skal eksistere. Dette blir forklart nærmere i delkapittel 3.2, hvor det også blir presentert spesialtilfeller av den generelle modellen.

3.1.2 Definisjon og egenskaper

Denne seksjonen presenterer en rekke egenskaper ved GMRFs. Teoremene og definisjonene, samt bevisene, er hovedsaklig hentet fra Rue og Held (2005). Enkelte deler har blitt noe forenklet, og bevisene er blitt utført med noen små modifikasjoner. Illustrasjonene baseres på kartdata fra SSB, og hensikten er å relatere teorien om GMRFs til selve oppgaven.

Den betingede avhengighetsstrukturen hos GMRFs blir presentert ved notasjonen $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, hvor $\mathcal{G}$ tilsvarer en graf og $\mathcal{V} = \{1, \ldots, n\}$ er de tilhørende nodene. $\mathcal{E}$ representerer den betingede avhengighetsstrukturen, hvor $\{i, j\} \in \mathcal{E}$ hvis og bare hvis node i og j er betinget avhengig (også omtalt som naboer). Figur 3.1 illustrerer $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, hvor det antas at en bestemt node tilsvarer et unikt administrativt punkt for en kommune i Sogn og Fjordane, og observasjonene hos to kommuner er betinget avhengige hvis og bare hvis grensene til kommunene møtes i minst ett punkt.

Definisjonen på GMRFs er gitt ved:

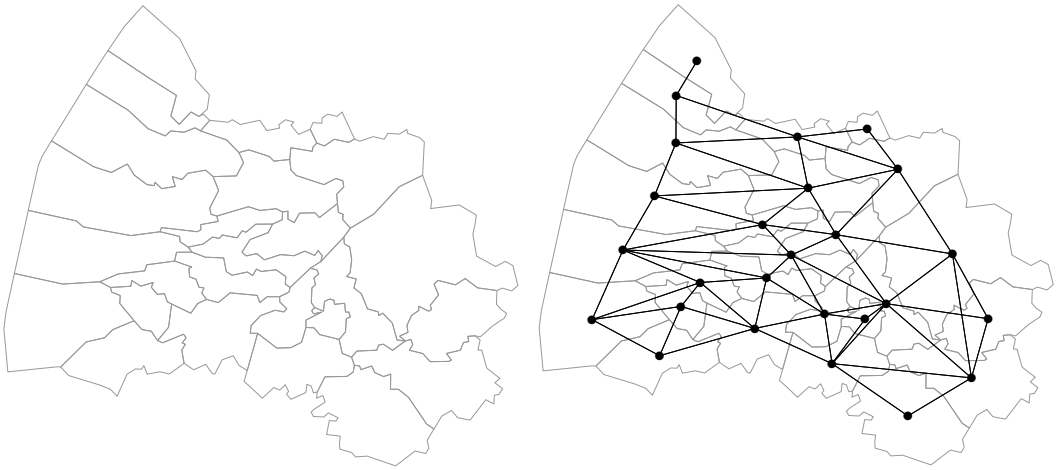
Definisjon 3.1. *En stokastisk vektor* $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]^T \in \mathbb{R}^n$ *blir kalt en GMRF mht* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *med forventning* $\boldsymbol{\mu}$ *og presisjonsmatrise* $\mathbf{Q} > 0$ *hvis og bare hvis simultantettheten kan uttrykkes ved*

$$\pi(\boldsymbol{\phi}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\phi} - \boldsymbol{\mu})^T \mathbf{Q} (\boldsymbol{\phi} - \boldsymbol{\mu}) \right\} \tag{3.8}$$

og

$$Q_{ij} \neq 0 \quad \Longleftrightarrow \quad \{i, j\} \in \mathcal{E} \quad \textit{for} \quad i \neq j.$$

I definisjon 3.1, og videre i kapittelet, vil $\mathbf{Q} > 0$ være ekvivalent med at $\mathbf{Q}$ er positiv-definit. Dette vil medføre at $\mathbf{Q}$ er invertibel, og det vil sikre at kovariansmatrisen $\mathbf{Q}^{-1}$ eksisterer.



(a) **(b)**

Figur 3.1: a) Kommuneinndelingen i Sogn og Fjordane. (b) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, hvor $\mathcal{V}$ tilsvarer de administrative punktene hos kommunene i Sogn og Fjordane, og $\mathcal{E}$ tilsvarer alle kombinasjoner av kommuner som deler grenser.

Teorem 3.1. *La $\boldsymbol{\phi}$ være multivariat normalfordelt med forventningsvektor $\boldsymbol{\mu}$ og presisjonsmatrise $\mathbf{Q} > 0$. For $i \neq j$ vil*

$$\phi_i \perp \phi_j \,|\, \boldsymbol{\phi}_{-ij} \quad \Longleftrightarrow \quad Q_{ij} = 0 \tag{3.9}$$

være sant.

Ved hjelp av teorem 3.1 kan man slå fast at dersom $\mathbf{Q}$ er gitt vil også $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ være bestemt, og man kan umiddelbart lese fra $\mathbf{Q}$ hvilke variabler som er betinget uavhengige.

Bevis Teorem 3.1, ligning (3.9). Antar at $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]^T \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\right)$. For bestemte verdier av i og j, hvor $i \neq j$, vil simultantettheten kunne uttrykkes ved

$$\begin{aligned} \pi\left(\phi_i, \phi_j, \boldsymbol{\phi}_{-ij}\right) &\propto \exp\left\{-\frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}\phi_k Q_{kl}\phi_l\right\} \\ &\propto \exp\left\{-\frac{1}{2}\phi_i\phi_j(Q_{ij}+Q_{ji}) - \frac{1}{2}\sum_{\{k,l\}\neq\{i,j\}}\phi_k Q_{kl}\phi_l\right\} \end{aligned} \tag{3.10}$$

En generalisering av (3.2) forteller at $\phi_i \perp \phi_j | \boldsymbol{\phi_{-ij}}$ hvis og bare hvis $\pi(\boldsymbol{\phi}) = f\left(\phi_i|\boldsymbol{\phi}_{-ij}\right) g\left(\phi_j|\boldsymbol{\phi}_{-ij}\right)$. Fra (3.10) ser man at dette er tilfelle hvis og bare hvis $Q_{ij} = Q_{ji} = 0$. Det ble for enkelhetsskyld antatt at $\boldsymbol{\mu} = \mathbf{0}$, men beviset kan enkelt overføres til en vilkårlig $\boldsymbol{\mu}$. □

Ellers er følgende teorem viktig for forståelsen av presisjonsmatrisen $\mathbf{Q}$ og egenskapene hos GMRFs.

Teorem 3.2. *Lar $\boldsymbol{\phi}$ være en GMRF mht $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ med forventning $\boldsymbol{\mu}$ og presisjonsmatrise $\mathbf{Q} > 0$, da er*

$$\mathrm{E}\left(\phi_i|\boldsymbol{\phi}_{-i}\right) = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(\phi_j - \mu_j) \tag{3.11}$$

$$\mathrm{Prec}\left(\phi_i|\boldsymbol{\phi}_{-i}\right) = Q_{ii} \tag{3.12}$$

$$\mathrm{Corr}\left(\phi_i, \phi_j|\boldsymbol{\phi}_{-ij}\right) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}} \tag{3.13}$$

Teorem 3.2 illustrer blant annet hvilken effekt diagonalelementene i $\mathbf{Q}$ har på de betingede egenskapene hos GMRFs. Dersom man legger til $\delta > 0$ hos diagonalelementene, vil dette medføre at $\mathrm{Corr}\left(\phi_i, \phi_j|\boldsymbol{\phi}_{-ij}\right)$ reduseres og $\mathrm{E}\left(\phi_i|\boldsymbol{\phi}_{-i}\right)$ i mindre grad legger vekt på $\boldsymbol{\phi}_{-i}$.

Bevis Teorem 3.2, ligning (3.11) *og* (3.12). Dersom $\phi \sim \mathcal{N}(\gamma, 1/\kappa)$ vil tetthetsfunksjonen til ϕ tilfredsstille

$$f(\phi) \propto \exp\left\{-\frac{1}{2}\kappa\phi^2 + \phi\kappa\gamma\right\}. \tag{3.14}$$

Videre har man at dersom $\boldsymbol{\phi} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\right)$, hvor $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]^T$, vil $\pi(\phi_i|\boldsymbol{\phi}_{-i})$ tilfredsstille

$$\begin{aligned}\pi(\phi_i|\boldsymbol{\phi}_{-i}) &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\phi}^T\mathbf{Q}\,\boldsymbol{\phi}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\phi_i^2 Q_{ii} - \phi_i \sum_{j:j\sim i} Q_{ij}\phi_j\right\}\end{aligned} \tag{3.15}$$

som et resultat av at $\pi(\phi_i|\boldsymbol{\phi}_{-i}) = \pi(\boldsymbol{\phi})/\pi(\boldsymbol{\phi}_{-i}) \propto \pi(\boldsymbol{\phi})$. Sammenlignes dette med (3.14) ser man at $\pi(\phi_i|\boldsymbol{\phi}_{-i})$ tilsvarer en univariat normalfordeling med forventning og presisjon gitt ved:

$$\begin{aligned}\mathrm{E}\left(\phi_i|\boldsymbol{\phi}_{-i}\right) &= -\frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}\phi_j \\ \mathrm{Prec}\left(\phi_i|\boldsymbol{\phi}_{-i}\right) &= Q_{ii}\end{aligned}$$

Dersom $\boldsymbol{\phi}$ har forventning $\boldsymbol{\mu}$ vil $\boldsymbol{\phi} - \boldsymbol{\mu}$ ha forventning $\mathbf{0}$. Som følge av dette vil beviset være fullstendig dersom ϕ_i og ϕ_j erstattes med henholdsvis $\phi_i - \mu_i$ og $\phi_j - \mu_j$. □

Se Rue og Held (2005)[2.2] for resterende bevis av teoremet.

3.2 Betinget autoregressiv prosess (CAR)

GMRFs blir som oftest spesifisert med en forventningsvektor $\boldsymbol{\mu}$ og en presisjonsmatrise $\mathbf{Q} > 0$. Besag (1974) introduserte modeller som går under navnet betingede autoregressive modeller, som er en alternativ fremgangsmåte for å spesifisere GMRFs. Dette blir gjort ved å definere de normalfordelte betingede sannsynlighetsfordelingene slik at

$$\mathrm{E}(\phi_i|\boldsymbol{\phi}_{-i}) = \mu_i - \sum_{j:j\sim i} \beta_{ij}(\phi_j - \mu_j) \tag{3.16}$$

$$\mathrm{Prec}(\phi_i|\boldsymbol{\phi}_{-i}) = \kappa_i > 0 \tag{3.17}$$

for $i = 1, \ldots, n$, $\{\beta_{ij}, i \neq j\}$, og vektorene $\boldsymbol{\mu}$ og $\boldsymbol{\kappa}$. For at det skal eksistere en gyldig simultantetthet $\pi(\boldsymbol{\phi})$ må de betingede tetthetsfunksjonene tilfredsstille en rekke krav. Blant annet må $\beta_{ij} \neq 0$ dersom $\beta_{ji} \neq 0$ som et resultat av at $\sim$ er symmetrisk . Sammenligner man (3.16) og (3.17) med (3.11) og (3.12) ser man at dersom man velger

$$Q_{ii} = \kappa_i \qquad \text{og} \qquad Q_{ij} = \kappa_i \beta_{ij}$$

og i tillegg krever at $\mathbf{Q}$ er symmetrisk ved å velge

$$\kappa_i \beta_{ij} = \kappa_j \beta_{ji}$$

vil $\pi(\boldsymbol{\phi})$ eksistere gitt at $\mathbf{Q} > 0$. Teorem 3.3 viser også at dette valget av $\pi(\boldsymbol{\phi})$ er unikt. Beviset er presentert av Rue og Held (2005)[2.2.4] ved hjelp av Brook's lemma.

Teorem 3.3. *Gitt de n normalfordelte betingede tetthetsfunksjonene med betinget forventning og presisjon gitt ved* (3.16) *og* (3.17)*, er $\boldsymbol{\phi}$ en GMRF mht $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ med forventning $\boldsymbol{\mu}$ og presisjonsmatrise $\mathbf{Q}$, hvor*

$$Q_{ij} = \begin{cases} \kappa_i \beta_{ij} & i \neq j \\ \kappa_i & i = j \end{cases} \tag{3.18}$$

gitt at $\kappa_i \beta_{ij} = \kappa_j \beta_{ji}$, $i \neq j$ og $\mathbf{Q} > 0$.

CAR-modeller er spesialtilfeller av GMRFs, og Lee (2013) gir en fin introduksjon til ulike CAR-modeller (se tabell 3.1) som ofte blir brukt innenfor spatial modellering. Som tidligere nevnt blir modellene ofte tatt i bruk innenfor bayesianske hierarkiske modeller og CAR-modellene blir typisk valgt som aprioriforfordelinger, slik at inferens utføres ved hjelp av MCMC-metoder.

Simultantettheten til modellene i tabell 3.1 kan skrives på formen $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{Q}_c^{-1})$, hvor $\mathbf{Q}_c/\sigma^2$ tilsvarer presisjonsmatrisen $\mathbf{Q}$ i seksjonene over. Den betingede avhengighetsstrukturen

		Fordeling	$\mathrm{E}(\phi_i\|\boldsymbol{\phi}_{-i})$	$\mathrm{Var}(\phi_i\|\boldsymbol{\phi}_{-i})$
(i)	Besag et al. (1991)	$\phi_i\|\boldsymbol{\phi}_{-i} \sim \mathcal{N}$	$\frac{\sum_{j:j\sim i} w_{ij}\phi_j}{\sum_{j:j\sim i} w_{ij}}$	$\frac{\sigma^2}{\sum_{j:j\sim i} w_{ij}}$
(ii)	Stern og Cressie (1999)	$\phi_i\|\boldsymbol{\phi}_{-i} \sim \mathcal{N}$	$\frac{\rho\sum_{j:j\sim i} w_{ij}\phi_j}{\sum_{j:j\sim i} w_{ij}}$	$\frac{\sigma^2}{\sum_{j:j\sim i} w_{ij}}$
(iii)	Leroux et al. (2000)	$\phi_i\|\boldsymbol{\phi}_{-i} \sim \mathcal{N}$	$\frac{\rho\sum_{j:j\sim i} w_{ij}\phi_j}{\rho\sum_{j:j\sim i} w_{ij}+1-\rho}$	$\frac{\sigma^2}{\rho\sum_{j:j\sim i} w_{ij}+1-\rho}$

Tabell 3.1: Oversikt over ulike CAR-modeller som ofte blir tatt i bruk dersom det er behov for å modellere spatial avhengighet.

blir bestemt ved $\mathbf{Q}_c$, og σ er en skaleringsparameter som kontroller den marginale variansen til variablene og tilhørende kovarianser (se vedlegg B). Ved enkle beregninger er det enkelt å vise at de tre modellene i tabell 3.1 er spesialtilfeller av ligning (3.16) og (3.17).

I denne oppgaven blir det tatt utgangspunkt i et spesialtilfelle av modellen til Besag et al. (1991), hvor $w_{ij} = 1$ dersom i og j er betinget avhengige. Dette vil medføre at presisjonsmatrisen $\mathbf{Q}_c$ ikke tilfredsstiller kravet om postiv-definit, siden $\mathbf{Q}_c\mathbf{1} = \mathbf{0}$. En slik modell går under navnet *intrinsic conditional autoregressions* (ICAR). For å unngå problemet legges det til en $\delta > 0$ hos diagonalelementene, slik at $\mathbf{Q}_c\mathbf{1} > \mathbf{0}$. Dette er en suffisient, men ikke nødvendig, betingelse for at matrisen $\mathbf{Q}_c$ er positiv-definit. Se Rue og Held (2005)[1.1.2] for nærmere diskusjon, samt egenskaper hos matriser som er positiv-definit.

Presisjonsmatrisen $\mathbf{Q}_c$ er følgelig gitt ved:

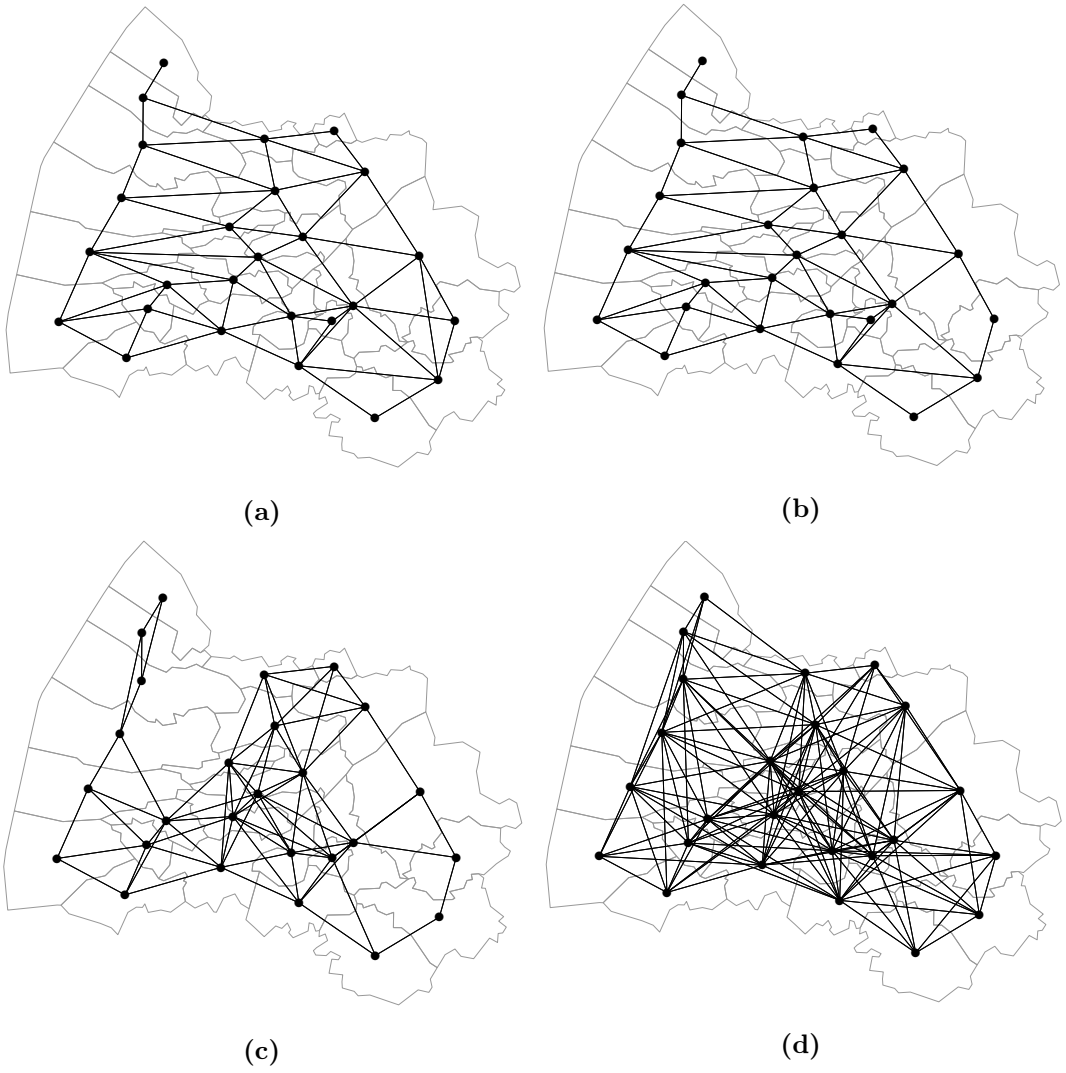
$$Q_{ij} = -1 \quad \text{hvis og bare hvis kommune } i \text{ og } j \text{ er naboer, ellers lik 0.}$$
$$Q_{ii} = n_i + \delta \quad \text{hvor } \delta > 0 \text{ og } n_i \text{ er antall naboer til kommune } i.$$

Modellene til Stern og Cressie (1999) og Leroux et al. (2000) introduserer også parameteren ρ, hvor begge modellene tilsvarer modellen til Besag et al. (1991) dersom $\rho = 1$. Årsaken til at ρ ble introdusert er for å kunne modellere situasjoner med ulik grad av spatial avhengighet i dataene, hvor økende ρ blir tolket som sterkere spatial korrelasjon i dataene.

For nærmere beskrivelse og diskusjon rundt fordeler og ulemper ved de ulike modellene henvises leseren til Lee (2013).

3.3 Valgmuligheter for betinget avhengighetsstruktur

Et av valgene som må tas når man analyserer data hvor det er naturlig å anta at det eksisterer en form for geografisk avhengighet er hvordan man skal definere to regioner som naboer. Som Bivand et al. (2008) poengterer er det mange muligheter, og det er en problemstilling som bør bli behandlet med varsomhet. Det er vanlig å definere to regioner som naboer dersom grensene deres møtes i minst ett punkt. En annen mulighet er å definere to regioner som naboer dersom avstanden mellom deres representative punkter er mindre enn en gitt verdi.



Figur 3.2: To kommuner er naboer dersom: (a) grensene møtes i minst ett punkt (b) grensene møtes i minst to punkter (c) avstanden mellom de administrative punktene er mindre enn 50 km (d) avstanden mellom de administrative punktene er mindre enn 75 km.

3.4 Simulering

Tabell 3.2 presenterer en algoritme for å simulere fra en vilkårlig GMRF $\boldsymbol{\phi}$ med presisjonsmatrise $\mathbf{Q} > 0$. Rue og Held (2005) presenterer også algoritmer for simulering fra blant annet $\pi(\boldsymbol{\phi}_A|\boldsymbol{\phi}_{-A})$ og $\pi(\boldsymbol{\phi}|\mathbf{A}\boldsymbol{\phi} = \boldsymbol{e})$, hvor $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$. For nærmere beskrivelse se Rue og Held (2005). Formålet med seksjonen er å illustrere hvilken effekt presisjonsmatrisen $\mathbf{Q}$ har, og spesielt hvordan resultatene påvirkes dersom man endrer diagonalen i $\mathbf{Q}$.

Sampling $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$
1. Utfør Cholesky dekomponeringen, $\mathbf{Q} = \boldsymbol{L}\boldsymbol{L}^T$
2. Sample $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
3. Løs $\boldsymbol{L}^T\boldsymbol{v} = \boldsymbol{z}$
4. Utfør $\boldsymbol{\phi} = \boldsymbol{\mu} + \boldsymbol{v}$
5. **Returner** $\boldsymbol{\phi}$

Tabell 3.2: Algoritme for å simulere et tilfeldig utvalg fra $\boldsymbol{\phi}$, hvor $\boldsymbol{\phi}$ er en GMRF med presisjonsmatrise $\mathbf{Q} > 0$.

Det presenteres simuleringer fra GMRFs $\boldsymbol{\phi}$ med følgende egenskaper:

- En enkelt node tilsvarer et unikt administrativt punkt hos en kommune i Norge.
- Kommuner defineres som naboer dersom grensene deres møtes i minst ett punkt.
- $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_{429}]^T \sim \mathcal{N}(\boldsymbol{\mu}, \sum) \quad \boldsymbol{\mu} = \mathbf{0} \quad \sum = \mathbf{Q}^{-1}$
- $Q_{ij} = -1$ hvis og bare hvis kommune i og j er naboer, ellers lik 0.
- $Q_{ii} = n_i + \delta$, hvor $\delta > 0$ og n_i er antall naboer for kommune i.

Merk at σ for enkelhetsskyld er satt lik 1. Resultatene er presentert i figur 3.3, hvor $\delta = \{10, 1, 0.01\}$ i henholdsvis (b), (c) og (d). I figur 3.3(a) illustreres et tilfeldig utvalg fra $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, hvor $\boldsymbol{I}$ er en 429×429 identitetsmatrise.

Gitt teorem (3.2) er de betingede egenskapene til $\boldsymbol{\phi}$ lik:

$$\mathrm{E}\left(\phi_i|\boldsymbol{\phi}_{-i}\right) = \frac{1}{n_i+\delta}\sum_{j:j\sim i}\phi_j \tag{3.19}$$

$$\mathrm{Prec}\left(\phi_i|\boldsymbol{\phi}_{-i}\right) = n_i+\delta \tag{3.20}$$

$$\mathrm{Corr}\left(\phi_i,\phi_j|\boldsymbol{\phi}_{-ij}\right) = \begin{cases} \frac{1}{\sqrt{n_i+\delta}\sqrt{n_j+\delta}} & \text{hvis og bare hvis } i\sim j \\ 0 & \text{ellers} \end{cases} \tag{3.21}$$

Dette vil medføre at $\mathrm{Corr}\left(\phi_i,\phi_j|\boldsymbol{\phi}_{-ij}\right)$ minker for økende δ, og $\mathrm{E}\left(\phi_i|\boldsymbol{\phi}_{-i}\right)$ i mindre grad legger vekt på $\boldsymbol{\phi}_{-i}$. I hvilken grad de betingede fordelingene påvirkes av δ vil også avhenge av antall naboer hos de ulike kommunene.

Banerjee et al. (2014) presenterer to standardmetoder, *Moran's I* og *Geary's C*, som blir anvendt for å måle den spatiale avhengigheten mellom observerte verdier hos regioner og dermed kan være med å underbygge en eventuell påstand om at det eksisterer en romlig avhengighet i dataene.

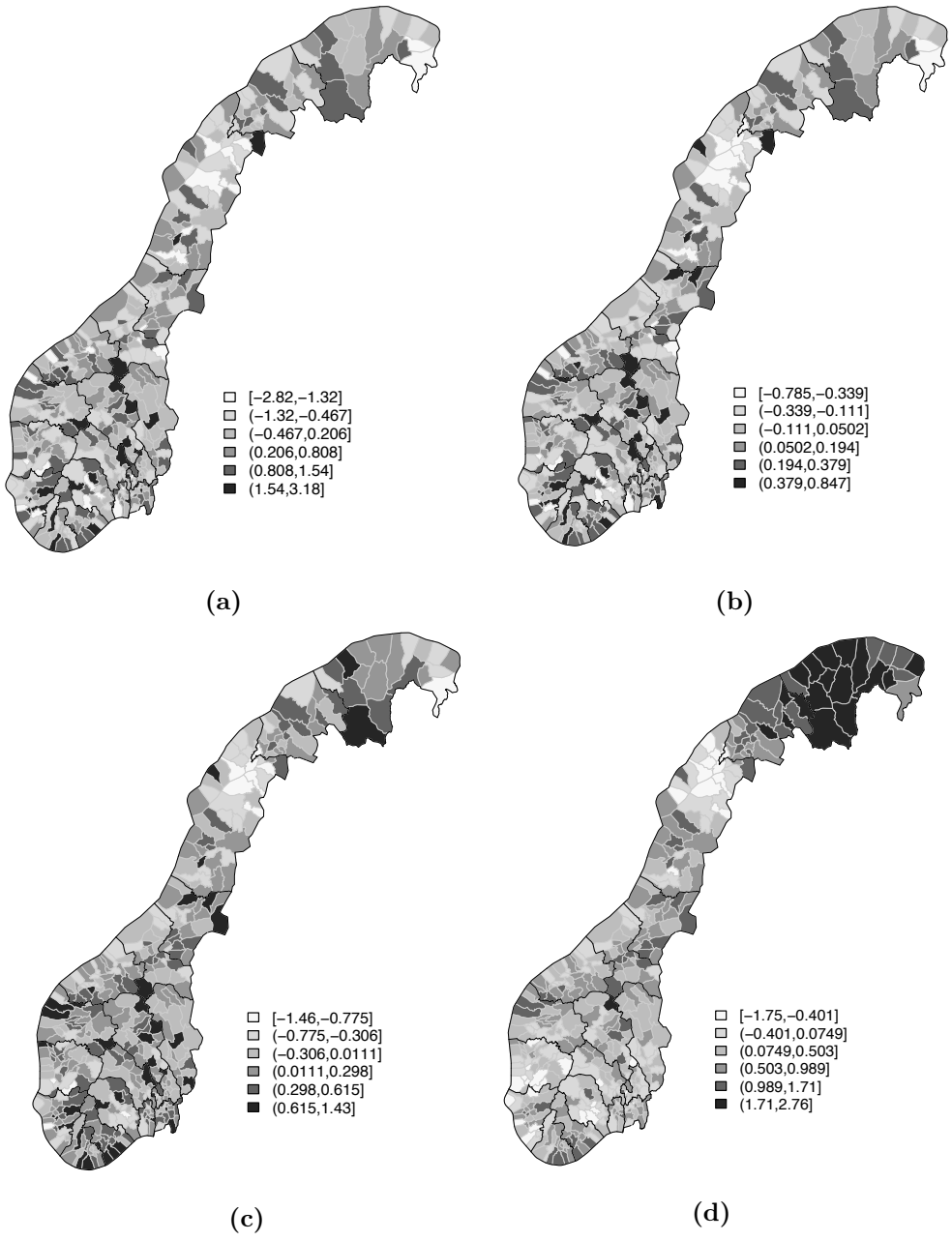
For å illustrere effekten av δ utføres Moran's I test på de fire datasettene, og resultatene oppsummeres som følgende:

	I	E(I)	Var(I)	st.deviate	p-verdi
(a)	-0.04135	-0.00234	0.00089	-1.30947	0.90481
(b)	0.03374	-0.00234	0.00089	1.21106	0.11294
(c)	0.23762	-0.00234	0.00089	8.05618	3.9e-16
(d)	0.58721	-0.00234	0.00089	19.80276	1.4e-87

Tabell 3.3: Moran's I test for de fire tilfeldige utvalgene i figur 3.3

Ved første øyekast er det «fristende» å anta at det ikke er noen romlig avhengighet hos dataene i figur 3.3(a) og 3.3(b), mens det på samme tid ser ut til å være romlig avhengighet hos dataene i figur 3.3(c) og 3.3(d). Moran's I testene i tabell 3.3 underbygger hypotesen om fordelingene hos (a), (c) og (d), men er i strid med (b). Det er også verdt å nevne at I og de tilhørende p-verdiene av I henholdsvis øker og synker når δ øker noe som er fornuftig med tanke på diskusjonen rundt δ i seksjonen over.

I vedlegg B presenteres effekten av å inkludere δ fra et mer teoretisk ståsted, hvor det i større grad fokuseres på kovariansmatrisen $\sum_{\boldsymbol{\phi}} = \boldsymbol{Q}^{-1}$.



Figur 3.3: (a) Tilfeldig utvalg fra $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. (b), (c) og (d) er tilfeldige utvalg fra $\boldsymbol{\phi}$, hvor $\boldsymbol{\phi}$ tilfredsstiller egenskapene i starten av delkapittel 3.4 og δ er henholdsvis lik 10, 1 og 0.01. Det tilfeldige utvalget i (a) tilsvarer vektoren $\mathbf{z}$ i tabell 3.2 for det tre tilfeldige utvalgene i (b), (c) og (d). Merk her at inndelingen er unik for hvert enkelt utvalg, og er funnet ved metoden Fisher-Jenks natural breaks (Bivand et al. (2008), s.77).

Kapittel 4

Modeller for antall skader

Innenfor forsikring er det viktig med gode modeller for skadeantall, og valget av modeller avhenger blant annet av hvilke type forsikringsprodukter man arbeider med. I noen tilfeller kan poissonmodeller for modellering av skadeantall være tilfredsstillende, mens i andre tilfeller er det nødvendig med modeller med tyngre haler. De Jong og Heller (2008) nevner blant annet dødelighetsanalyser, helseforsikring og bilforsikring som eksempler på situasjoner hvor det er behov for frekvensmodeller.

4.1 Antall skader

Ved modellering av skadeantall er det ofte naturlig å ta utgangspunkt i at responsvariablene, gitt ved $\mathbf{y} = [y_1, \ldots, y_n]^T$, er poissonfordelt. Det antas at forventet skadeantall er en funksjon av en gitt mengde kjente forklaringsvariabler slik at modellen kan uttrykkes ved:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i), \\ g(\mu_i) &= \mathbf{x}_i^T \boldsymbol{\beta}, \end{aligned} \tag{4.1}$$

hvor $\mathbf{x}_i^T = [1, x_{i1}, \ldots, x_{ip}]^T$ og $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]$. Et naturlig valg for linkfunksjonen $g(\cdot)$ er den naturlige logaritmen. Dette vil resultere i at den estimerte forventningen alltid er positiv. Det er også mulig å velge identitetsfunksjonen som linkfunksjon, men man vil da kunne risikere negative estimater for forventningen, noe som er i strid med at antall skader konsekvent er større eller lik 0.

Dersom man velger den naturlige logaritmen eller identitetsfunksjonen som linkfunksjon vil man henholdsvis få en multiplikativ- og additiv modell. For å illustrere dette kan man ta utgangspunkt i det enkle tilfellet hvor man bare har én forklaringsvariabel slik at $\mathbf{x}_i^T = [1, x_i]$

og $\mu_i = e^{\beta_0 + x_i\beta_1}$. En marginaløkning i x_i vil medføre at $\mu_i = e^{\beta_0+(x_i+1)\beta_1} = e^{\beta_0+x_i\beta_1}e^{\beta_1}$, og den multiplikative effekten er lik e^{β_1}. Tilsvarende vil en marginaløkning i x_{is}, for $s \in \{1, \ldots, p\}$, øke forventningen med en faktor lik e^{β_s} i den mer generelle modellen. Se De Jong og Heller (2008)[6.1, s.82] for en tilsvarende forklaring for den additive modellen.

4.1.1 Eksponeringstid

Når man modellerer antall skader er det viktig å korrigere for eksponeringstiden til de enkelte forsikringsavtalene, det vil si, hvor lenge hver enkelt forsikringsavtale var gjeldende i den aktuelle perioden. Dersom analysegrunnlaget baseres på data fra 01.januar 2012 til 31.desember 2012 vil forsikringsavtale i ha en eksponeringstid lik $t_i \in (0, 1]$, hvor $t_i = 1$ tilsvarer at forsikringsavtale i har vært gjeldende i hele perioden. I en realistisk situasjon vil en stor andel av forsikringsavtalene i porteføljen ha eksponeringstid lik 1. Forsikringsavtaler hvor eksponeringstiden er mindre enn 1 er enten avtaler som blir tegnet eller avsluttet i løpet av perioden.

Korrigeringen for eksponeringstiden løses ofte ved å inkludere et såkalt offset e_i for observasjon i, og modellen utvides slik at:

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i), \\ g(\mu_i) &= \log(e_i) + \mathbf{x}_i^T\boldsymbol{\beta}, \end{aligned} \tag{4.2}$$

hvor $g(\cdot)$ tilsvarer den naturlige logaritmen.

En annen problemstilling som dukker opp i forbindelse med eksponeringstiden er hvordan man skal definere e_i-ene. I eksempelet med bilforsikring kan man, eksempelvis, velge $e_i = t_i$ eller $e_i = t_i^2$.

4.1.2 Overdispersjon

En av de sentrale egenskapene til Poissonfordelingen er at forventningen, per definisjon, er lik variansen. Dersom man har et datasett bestående av n observasjoner slik at:

$$\mathbf{y} = [y_1, \ldots, y_n]^T \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{4.3}$$

kan det være ugunstig å anta at responsvariablene er poissonfordelt dersom $\hat{\sigma}^2$ er betydelig større enn $\overline{y}$. Dette kalles overdispersjon, og er et velkjent problem ved poissonregresjon.

Det er ulike måter å håndtere eventuell overdispersjon på, og De Jong og Heller (2008) presenterer den sammensatte Poisson-gamma modellen gitt ved

$$\begin{aligned} &Y_i|\eta_i \overset{iid}{\sim} \text{Poisson}\,(\eta_i) \\ &\text{hvor} \quad \eta_i \overset{iid}{\sim} \text{G}(a,b), \quad i = 1, \cdots, n, \end{aligned} \tag{4.4}$$

som en mulig løsning på problemet. En slik modell vil medføre at at $Y_i \sim \text{NEGBIN}(r,p)$, hvor $r = a$ og $p = \frac{1}{1+b}$. Forventningen og variansen til Y_i er gitt ved

$$\text{E}(Y_i) = \frac{pr}{1-p} \qquad \text{Var}(Y_i) = \frac{pr}{(1-p)^2} = \frac{\text{E}(Y_i)}{1-p} > \text{E}(Y_i). \tag{4.5}$$

og problemet med varians som er betydelig større enn forventningen kan være tatt hånd om. Det vil likevel være tilfeller hvor denne utvidelsen av modellen ikke er tilstrekkelig, og det bør undersøkes om andre modeller kan være mer passende. Selve beviset for at $Y_i \sim \text{NEGBIN}(r = a, p = \frac{1}{1+b})$ dersom $Y_i|\eta_i \sim \text{Poisson}(\eta_i)$ og $\eta_i \sim \text{Gamma}(a,b)$ er presentert hos De Jong og Heller (2008)[2.9, s.32-33].

De Jong og Heller (2008) presenterer også to andre metoder som blir brukt for å håndtere overdispersjon i dataene. Begge har tilsvarende struktur som Poisson-gamma modellen, men istedenfor å anta at $\eta \sim \text{Gamma}(a,b)$ antar man at η enten følger en generalisert invers gaussisk fordeling eller en invers gaussisk fordeling. Den generaliserte inverse gaussiske fordelingen har større fleksibilitet som følge av tre parametere, men selve anvendelsen kan vise seg å være utfordrende ved numeriske metoder. Den inverse gaussiske fordelingen har to parametere, og er således enklere å benytte seg av ved numeriske metoder. Årsaken til at de to fordelingene kan være gunstig å bruke fremfor gammafordelingen er at fordelingene har en større grad av skjevhet og kan fange opp eventuelle situasjoner hvor det er en enda større variasjon i dataene. Se De Jong og Heller (2008) for en nærmere diskusjon og referanser til artikler som har sammenlignet de ulike modellene.

4.2 Hierarkiske modeller

Banerjee et al. (2014) argumenterer for at dersom målet er å finne den sanne risikofordelingen over et geografisk område er det fornuftig å introdusere modeller som inneholder randomiserte effekter, siden det er rimelig å anta at den sanne risikoen i de ulike regionene kommer fra en felles underliggende fordeling. For å relatere dette til problemstillinger innenfor forsikring kan man tenke på risiko som, eksempelvis, antall skader eller antall dødsfall. Heretter vil risiko

være ekvivalent med antall skader, siden det er antall skader som gjennomgående er fokus i oppgaven.

Modeller som inneholder randomiserte effekter tillater at regionene «låner» informasjon fra hverandre, og håpet er at en slik fleksibilitet i modellen vil resultere i at risikoen i regionene estimeres med større presisjon. De randomiserte effektene er ofte høydimensjonale, som et resultat av det geografiske området er delt inn i mange mindre regioner, og inferens og estimering blir ofte utført ved hjelp av bayesianske hierarkiske modeller og MCMC.

Banerjee et al. (2014) gir en fin introduksjon til oppbyggingen av slike modeller, hvor utgangspunktet er en standard poissonmodell gitt ved (4.1). I noen tilfeller kan dette være en passende modell, men det vil også være tifeller hvor det er nødvendig å introdusere hierarkiske modeller for større fleksibilitet.

I dette delkapittelet vil det være snakk om modeller hvor man bare har en observasjon per region, mens det i delkapittel 4.3 introduseres modeller hvor man har flere observasjoner for hver region. I denne oppgaven vil sistnevnte være aktuelt, siden responsvariablene tilsvarer skadeantall hos en enkelt forsikringsavtale, og ikke det totale antall skader innefor de ulike regionene.

Det enkleste eksempelet er gitt ved en Poisson-gamma modell i ligning (4.4). Dersom det i tillegg inkluderes offset er modellen gitt ved

$$\begin{aligned} Y_i|\eta_i &\overset{iid}{\sim} \text{Poisson}\,(e_i\eta_i) \\ \eta_i &\overset{iid}{\sim} \text{G}(a,b), \quad i = 1, \cdots, n, \end{aligned} \tag{4.6}$$

hvor $a = \mu^2/\sigma^2$ og $b = \mu/\sigma^2$. Her tilsvarer μ og σ^2 henholdsvis forventning og varians til η_i-ene.

For å relatere modellen til et konkret eksempel kan man tenke seg at e_i tilsvarer forventet antall skader i region i, mens η_i opptrer som en slags justeringsfaktor. Ved å ta utgangspunkt i at μ er lik 1, vil det være naturlig å undersøke om η_i er forskjellig fra 1. Dersom det viser seg at η_i er statistisk signifikant større enn 1, vil konklusjonen være at risikoen innenfor region i er større enn hva som var forventet. For en nærmere diskusjon se Banerjee et al. (2014)[5.4, s.158-167].

Poisson-gamma modellen gitt ved (4.6) er relativt enkel å arbeide med, men lider av visse mangler. Blant annet vil det være vanskelig å inkludere en korrelasjon mellom η_i-ene, noe som kan være en passende antagelse i mange tilfeller. For å inkludere en slik antagelse i modellen gitt ved (4.6) vil det være nødvendig å introdusere en multivariat gammafordeling, noe som er både er vanskelig og lite gunstig med tanke på beregningene. Løsningen på problemet er å

omformulere modellen slik at

$$\begin{aligned} &Y_i|\psi_i \sim \text{Poisson}\left(e_i \mathrm{e}^{\psi_i}\right) \\ &\text{hvor} \quad \psi_i = \mathbf{x}_i^T\boldsymbol{\beta} + \theta_i + \phi_i, \quad i = 1, \cdots, n \end{aligned} \tag{4.7}$$

Det blir antatt at ψ_i-ene, hvor $\psi_i = \log(\eta_i)$, er multivariat normalfordelt. $\mathbf{x}_i^T$ er kjente forklaringsvariabler hos region i, mens $\boldsymbol{\beta}$ er en vektor med ukjente parametere. I mange tilfeller vil det være tilstrekkelig å bare inkludere $\mathbf{x}_i^T\boldsymbol{\beta}$, og utføre inferens og estimering av parametere ved tradisjonelle metoder presentert i delkapittel 2.6. Dette vil typisk være situasjoner hvor $\mathbf{x}_i^T$ alene forklarer den spatiale variasjonen hos responsvariablene, og en eventuell inkludering $\boldsymbol{\theta}$ og $\boldsymbol{\phi}$ vil være overflødig.

Formålet med å inkludere θ_i-ene er å fange opp regional heterogenitet. θ_i-ene antas å være normalfordelt og uavhengige slik at

$$\theta_i \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{\tau_h}\right), \tag{4.8}$$

hvor τ_h tilsvarer den inverse variansen til θ_i-ene og er en presisjonsparameter som kontrollerer effekten av θ_i-ene. De randomiserte effektene inkluderes i modellen for å fange opp en eventuell ekstra-variasjon i ψ_i-ene. Dette er altså en metode for å håndtere eventuell overdispersjon i dataene som varierer globalt, det vil si, over hele det geografiske området.

Inkluderingen av ϕ_i-ene gjør modellen til en spatial modell, og det er ϕ_i-ene som modellerer avhengighetsstrukturen i dataene og fanger opp eventuell ekstra-variasjon i dataene som varierer lokalt. Dette bygger på antagelsen om at regioner som ligger nær hverandre har like egenskaper, som resulterer i en korrelasjon mellom risikoen i regionene.

I kapittel 3 ble teorien om GMRFs og ulike spesialtilfeller av GMRFs (CAR-modeller) presentert, og det er vanlig å anta at $\boldsymbol{\phi}$ følger en av de nevnte fordelingene. De er spesielt aktuelle i tilfeller hvor det geografiske området er delt inn i irregulære regioner, noe som er tilfellet i denne oppgaven (Banerjee et al. (2014)[5.4, p.163]).

I denne oppgaven vil det bli tatt utgangspunkt i at $\boldsymbol{\phi}$ tilfredsstiller

$$\pi(\phi_1, \ldots, \phi_n) \propto \exp\left\{-\frac{\tau_c}{2}\boldsymbol{\phi}^T\boldsymbol{Q}\boldsymbol{\phi}\right\}, \tag{4.9}$$

hvor $Q_{ij} = -1$ hvis og bare hvis kommune i og j er naboer. Diagonalelementene er gitt ved $Q_{ii} = n_i + \delta$, hvor $\delta > 0$ og n_i er antall naboer for kommune i.

4.3 Modeller med flere observasjoner innenfor regioner

I denne oppgaven vil all simulering av data utføres slik at det vil være tilfeller hvor man har flere observasjoner for en bestemt kommune, samtidlig som det vil være tilfeller hvor man ikke har noen observasjoner for andre kommuner.

Et av hovedmålene er å undersøke effekten av å inkludere latente variabler med en bestemt avhengighetsstruktur. I motsetning til modellene som ble presentert i delkapittel 4.2, vil det ikke inkluderes en unik latent variabel for hver enkelt responsvariabel, men en latent variabel for hver enkelt kommune.

Utgangspunktet for analysene vil være modeller på formen

$$\begin{aligned} Y_i|\psi_i &\sim \text{Poisson}\left(e_i \mathrm{e}^{\psi_i}\right) \\ \psi_i &= \mathbf{x}_i^T\boldsymbol{\beta} + \theta_{K(i)} + \phi_{K(i)}, \end{aligned} \tag{4.10}$$

hvor $\mathbf{x}_i^T = [1, x_{i1}, \ldots, x_{ip}]^T$, $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_p]$ og $K(i)=j$ hvis og bare hvis observasjon i stammer fra kommune j. Her er det valgt å ikke skille mellom individuelle og kommunespesifikke forklaringsvariabler. Fordelingen til $\boldsymbol{\phi}$ og $\boldsymbol{\theta}$ er identisk med fordelingene beskrevet i delkapittel 4.2.

Márkus et al. (2010) tok i bruk en lignende fordeling, men valgte å skille mellom individuelle og kommunespesifikke forklaringsvariabler. Det ble antatt at antall skader hos poliseholder i var poissonfordelt med forventning $\lambda_i = \lambda \cdot e_i \cdot \kappa_i \cdot \mathrm{e}^{\theta_{K(i)}}$, slik at

$$Y_i|\lambda_i \sim \text{Poisson}\left(\lambda \cdot e_i \cdot \kappa_i \cdot \mathrm{e}^{\theta_{K(i)}}\right). \tag{4.11}$$

Her tilsvarer e_i eksponeringstiden til polise i, og κ_i er den totale effekten fra de individuelle forklaringsvariablene. Den spatiale effekten er gitt ved faktoren $\lambda \mathrm{e}^{\theta_{K(i)}}$, hvor λ fungerer som en skaleringsparameter. Dersom man antar at Y_i-ene er uavhengige gitt λ_i vil dette medføre at antall skader innenfor region j, gitt ved $N_j = \sum_{i:K(i)=j} Y_i$, er poissonfordelt, slik at

$$N_j|\lambda_j \sim \text{Poisson}\left(\lambda \mathrm{e}^{\theta_j}\left\{\sum_{i:K(i)=j} e_i \cdot \kappa_i\right\}\right). \tag{4.12}$$

Márkus et al. (2010) estimerte de ulike parameterne i en iterativ prosedyre. I det første steget ble det antatt at den spatiale effekten var lik 1, og effekten fra de individuelle forklaringsvariablene ble estimert på tradisjonelt vis. I det neste steget ble parameterne hos den spatiale effekten estimert. I dette steget ble det antatt at de effekten fra de individuelle forklaringsvariablene var kjent, og gitt ved de estimerte effektene i det første steget. Denne prosedyren ble gjentatt til de ulike estimatene konvergerte.

Kapittel 5

Simulering av forsikringsportefølje

Som tidligere nevnt vil resultatene baseres på simulerte data og ikke reelle forsikringsdata. Dimakos og Di Rattalma (2002) simulerte en forsikringsportefølje bestående av 5000 poliser, hvor hver enkelt polise ble tildelt et fylkesnummer. Jeg har hentet inspirasjon fra deres fremgangsmåte, men i mitt tilfelle vil hver enkelt polise bli tildelt et kommunenummer fremfor fylkesnummer. For å gjøre simuleringen så realistisk som mulig er informasjon om antall innbyggere, alder, inntekt, urbaniseringsgrad (figur 5.1a) og kriminalitetsrate (figur 5.1b) hentet fra Statistisk Sentralbyrå (SSB) sine hjemmesider[1].

Det har vært nødvendig å gjøre enkelte forenklinger som følge av at det kontinuerlig skjer endringer når det gjelder kommunegrenser og antall kommuner i Norge[2]. Datagrunnlaget, fra både SSB og Kartverket, forholder seg til at det fra 1.januar 2012 til 31.desember 2012 var 429 kommuner i Norge, og det faller seg naturlig å anta at dette ikke endrer seg.

5.1 Beskrivelse av simulert forsikringsportefølje

Porteføljen består av 10000 poliser. Hver enkelt polise blir tildelt et kommunenummer basert på et tilfeldig utvalg fra en fordeling hvor utfallsrommet tilsvarer de unike kommunenumrene og punktsannsynlighetene er lik befolkningsandelene. Dette vil kunne medføre at enkelte kommuner ikke blir tildelt noen poliser, og kommuner med svært lave innbyggertall vil være utsatt for dette. Dimakos og Di Rattalma (2002) antok at hver enkelt polise har en eksponeringstid som er uniformfordelt og mindre enn to år. For å gjøre simuleringen mer realistisk antas det at 70% av polisene har en eksponeringstid lik ett år, og de resterende

[1] http://www.ssb.no/

[2] http://www.statkart.no/Kunnskap/Fakta-om-Norge/Fylker-og-kommuner/Tabell/

vil følge en diskret uniformfordeling med utfallsrom lik $\{1, 2, \ldots, 365\}$. Forklaringsvariablene til polise i lokalisert i kommune $K(i)$, hvor $K(i) = j$ dersom forsikringstaker i er bosatt i kommune j, er som følgende:

Urbanisering:	$x_{K(i),1} = \log(\text{antall innbyggere per km}^2 \text{ i kommune } K(i))$
Kriminalitetsrate:	$x_{K(i),2} = \text{antall forbrytelser per 1000 bosatte i kommune } K(i)$
Inntekt poliseholder:	$x_{i3} = I\{\text{lav}\} \quad x_{i4} = I\{\text{medium}\} \quad x_{i5} = I\{\text{høy}\}$
Alder poliseholder:	$x_{i6} = I\{1\}\ x_{i7} = I\{2\}\ x_{i8} = I\{3\}\ x_{i9} = I\{4\}\ x_{i10} = I\{5\}$

Tabell 5.1: Forklaringsvariabler for forsikringstaker i

Aldersgruppe	1	2	3	4	5
Alder	20-29	30-39	40-49	50-69	>69

Tabell 5.2: Inndeling av aldersgrupper

Inndelingen av aldersgruppene er presentert i tabell 5.2. Funksjonen $I(\cdot)$ er en indikatorfunksjon slik at $x_{i5} = 1$ dersom inntekten til forsikringstaker i blir karakterisert som «høy», og null ellers. Inntekten til forsikringstakerne i kommune $j = 1, \ldots, 429$ tilsvarer et tilfeldig utvalg fra en gammafordeling med forventning og standardavvik lik henholdsvis medianinntekten i kommune j og differansen mellom den største og minste medianinntekten. Inntekten er karakterisert som «lav» dersom den er mindre enn første kvartil, «høy» dersom den er større enn tredje kvartil, og «medium» ellers. Den lineære prediktoren til polise i er lik $\eta_i = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}$, hvor $\mathbf{x}_i^T = [x_{K(i),1}, x_{K(i),2}, x_{i3}, \ldots, x_{i10}]$, $\boldsymbol{\beta} = [0.1, 0.005, 0, -0.1, -0.7, 0.3, 0.1, 0, -0.1, 0.2]^T$ og $\beta_0 = -2.3$. De to parameterne som er satt lik null indikerer at en person med lav inntekt i aldersgruppe 3 er brukt som referansenivå. Til slutt antar man at antall skader inntreffer uavhengig av hverandre og er poissonfordelt slik at

$$N_i \sim \text{Poisson}(e_i \exp\{\eta_i\}), \quad i = 1, 2, \ldots, 10000. \tag{5.1}$$

Eksponeringstiden til polise i er gitt ved e_i. I likhet med Dimakos og Di Rattalma (2002) er parameterverdiene valgt slik at det er en fornuftig differanse mellom forventet antall skader hos poliseholdere i antatte høyrisiko- og lavrisikogrupper. Det er også tatt hensyn til at den prosentvise andelen av poliser hvor det inntreffer skader ikke er for stor.

5.2 Datagrunnlag

Dataene er basert på årene 2011-2012 og ble hentet fra SSB sine hjemmesider. Tabell 5.3 og 5.4 tilsvarer ikke de eksakte tabellene fra SSB, men er resultatet etter bearbeidelse av rådata. Bosatte innenfor de ulike aldersgruppene er gitt ved fylkesnivå som følge av at bearbeidelse av tilsvarende datamateriale gitt ved kommunenivå var for omfattende og tidkrevende. Merk her at første kolonne i tabell 5.4 ikke tilsvarer kommunenummer. Årsaken til at kommunene ikke er koblet mot kommunenummer er at kartdataene fra Kartverket brukte tallene i nevnte tabell som koblingsnøkkel.

fylke	20-29	30-39	40-49	50-59	60-69	70-79	80-89	>89
Østfold	31589	35093	41766	36520	33250	18539	11052	2160
Akershus	60578	74215	90697	70969	57623	31076	17628	3189
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Troms	20659	20225	23109	20452	18023	9835	5511	1036
Finnmark	9486	8914	11233	9473	8520	4656	2395	371

Tabell 5.3: Antall bosatte i ulike aldersgrupper i fylkene i Norge.

nr	kommune	fylke	antall_pers	areal	krim_rate	inntekt
1	Bjugn	Sør-Trøndelag	4584	355.83	19.9	402000
2	Rissa	Sør-Trøndelag	6543	588.05	19.7	430000
3	Leksvik	Nord-Trøndelag	3527	399.68	15.6	448000
⋮	⋮	⋮	⋮	⋮	⋮	⋮
428	Volda	Møre og Romsdal	8693	524.89	26.5	426000
429	Meløy	Nordland	6657	798.45	25.4	446000

Tabell 5.4: Oversikt over antall bosatte, kriminalitetsrate, areal og inntekt hos kommunene i Norge. Inntekt tilsvarer medianen til registrerte husholdningsinntekter, kriminalitetsrate er antall registrerte forbrytelser per 1000 bosatte og areal er landarealet gitt i km^2.

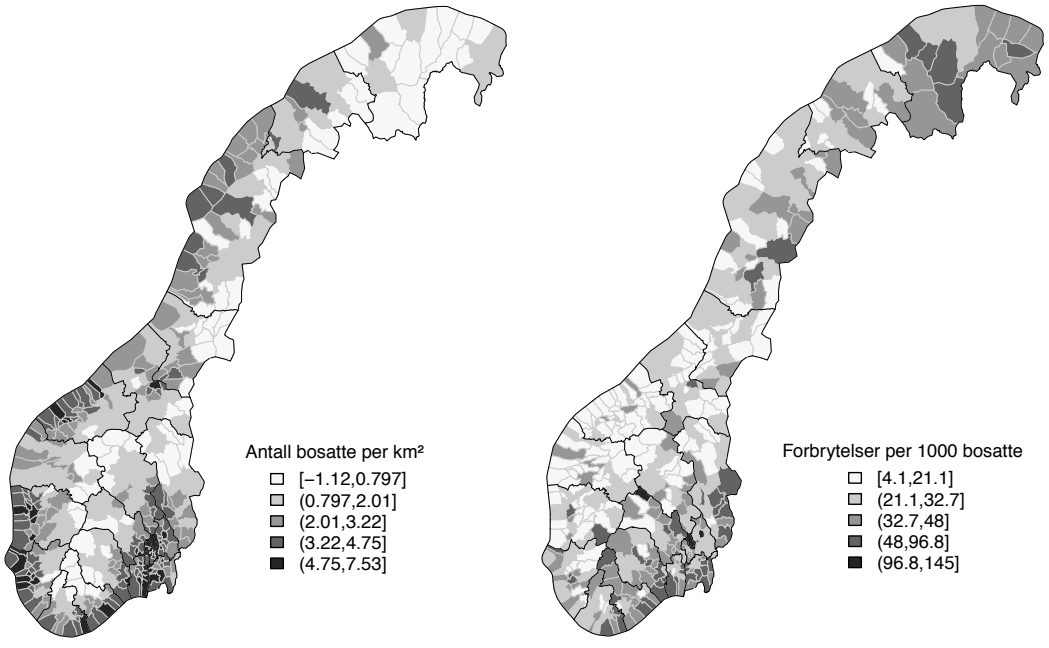
5.3 Deskriptiv statistikk

5.3.1 Forklaringsvariabler

Forklaringsvariablene i delkapittel 5.1 er en kombinasjon av individuelle egenskaper hos poliseholder og egenskaper hos kommunen poliseholder er bosatt i.

Kommunenivå

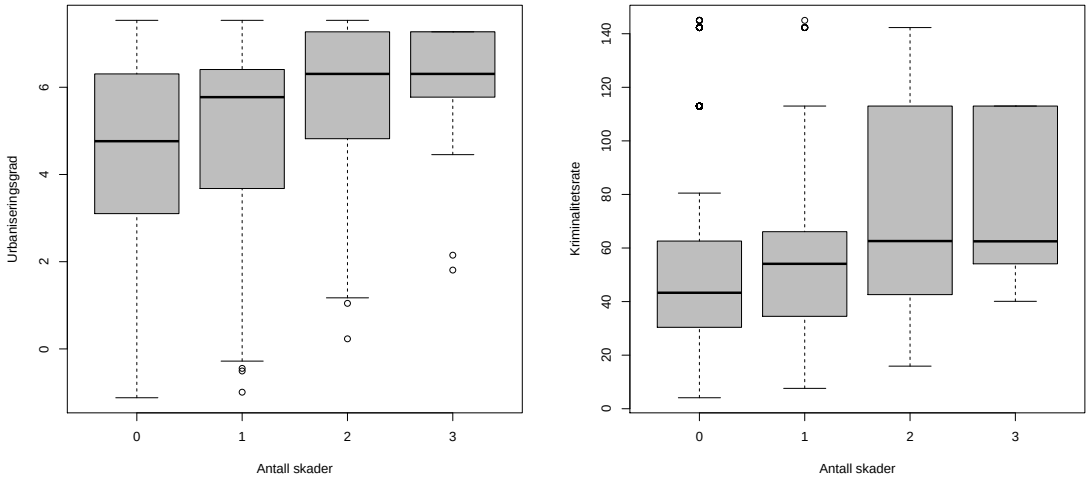
Forklaringsvariablene knyttet til urbaniseringsgrad og kriminalitetsrate, gitt ved $x_{K(i),1}$ og $x_{K(i),2}$, beskriver egenskaper hos kommune $K(i)$ hvor poliseholder i er bosatt. Urbaniseringsgrad og kriminalitetsrate er illustrert i figur 5.1, og viser blant annet at kommuner langs kystlinjen som strekker seg fra Hordaland til Vestfold har et høyt antall bosatte per km^2, i tillegg til kommuner i de ytre områdene i Møre og Romsdal og kommuner rundt Oslo. Ellers er det et høyt antall registrerte forbrytelser per 1000 bosatte i nord, og tilsvarende for kommuner i de sørøstlige områdene.



(a) **(b)**

Figur 5.1: Urbaniseringsgrad og kriminalitetsrate varierer for kommuner i Norge. Figur (a) og figur (b) tilsvarer henholdsvis logaritmen til antall bosatte per km^2 og antall registrerte forbrytelser per 1000 bosatte. Grenser for fylker er illustrert ved tykke, svarte linjer.

Figur 5.2 illustrerer sammenhengen mellom urbaniseringsgrad og kriminalitetsrate og antall skader, ved boksplott. For enkelthetsskyld omtales logaritmen til urbaniseringsgrad som urbaniseringsgrad. Det er ingen overraskelse at antall skader ser ut til å være økende med hensyn til både urbaniseringsgrad og kriminalitetsrate, da dette er en antagelse som har blitt gjort i simuleringen gjennom faktorene $0.1x_{K(i),1}$ og $0.005x_{K(i),2}$. Figur 5.2 illustrerer også tilfeller av uteliggere, som kan være et resultat av flere faktorer. Uteliggerne i figur 5.2b kan, for eksempel, være poliser med svært liten eksponeringstid, poliser i såkalte lavrisikogrupper med unntak av kriminalitetsrate eller en kombinasjon av de to. Lignende konklusjoner kan også trekkes for de andre tilfellene.



(a) **(b)**

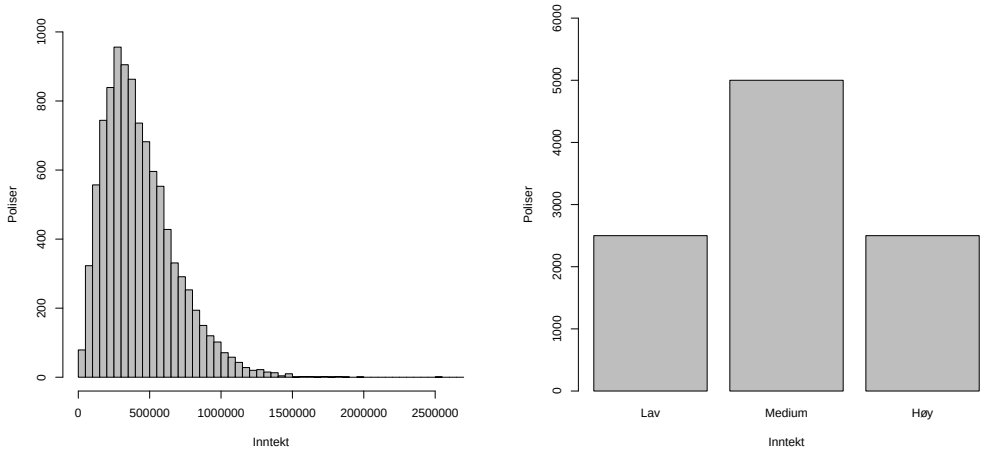
Figur 5.2: Boksplott av urbaniseringsgrad og kriminalitetsrate mot antall skader. I tråd med (5.1) øker medianen til urbaniseringsgraden når antall skader øker. Lignende resultater for kriminalitetsrate, men ikke like tydelig.

Individnivå

Forklaringsvariablene inntekt og alder er individuelle egenskaper hos poliseholderne. Det er likevel verdt å merke seg at også de vil være tilknyttet til hvor poliseholderen er bosatt gjennom medianinntekt hos kommuner og fordeling av aldersgrupper hos fylker.

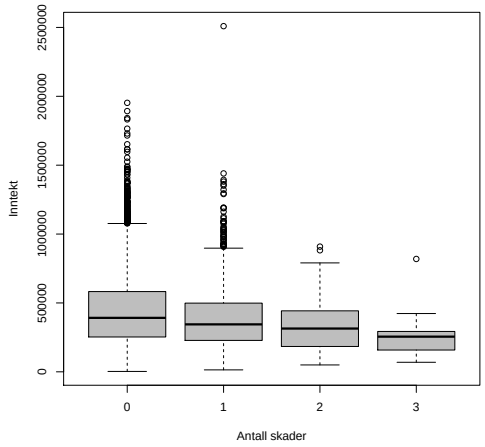
Inntekt

Som tidligere nevnt antas det at inntekten til forsikringstakerne i kommune $j = 1, \ldots, 429$ tilsvarer et tilfeldig utvalg fra en gammafordeling med forventning og standardavvik lik henholdsvis medianinntekten i kommune j og differansen mellom den største og minste medianinntekten. Resultatene fra simuleringen er presentert i figur 5.3a.



(a) **(b)**

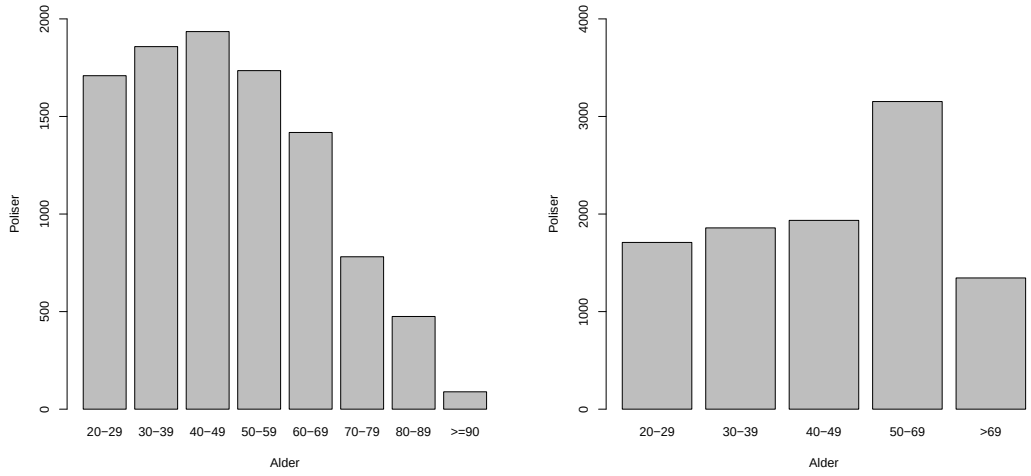
Figur 5.3: Simulert inntektsfordeling, samt resultater etter at inntekt er kategorisert gitt ved x_{i3},x_{i4} og x_{i5}.



Figur 5.4: Sammenhengen mellom inntekt og antall skader. I tråd med fordelingen til N_i, gitt ved (5.1), er antall skader økende med synkende inntekt.

Alder

Alder blir simulert ved hjelp av aldersfordelingen innenfor fylkesgrensene. For poliseholdere innenfor et gitt fylke, bestemt av kommunenummer, tilsvarer alderen et tilfeldig utvalg fra en fordeling med utfallsrom lik $\{20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, >89\}$ og punktsannsynligheter lik befolkningsandelen i de ulike aldersgruppene innenfor fylket. Aldersfordelingen til poliseholderne er illustrert i figur 5.5. Resultatet fra simuleringen bestemmer dermed verdien av forklaringsvariablene $x_{i6}, x_{i7},\ x_{i8},\ x_{i9}$ og x_{i10} definert i tabell 5.1.



(a) **(b)**

Figur 5.5: Fordelingen til polisene hos de ulike aldersgruppene og forklaringsvariablene $x_{i6}, x_{i7},\ x_{i8},\ x_{i9}$ og x_{i10}.

5.3.2 Responsvariabler

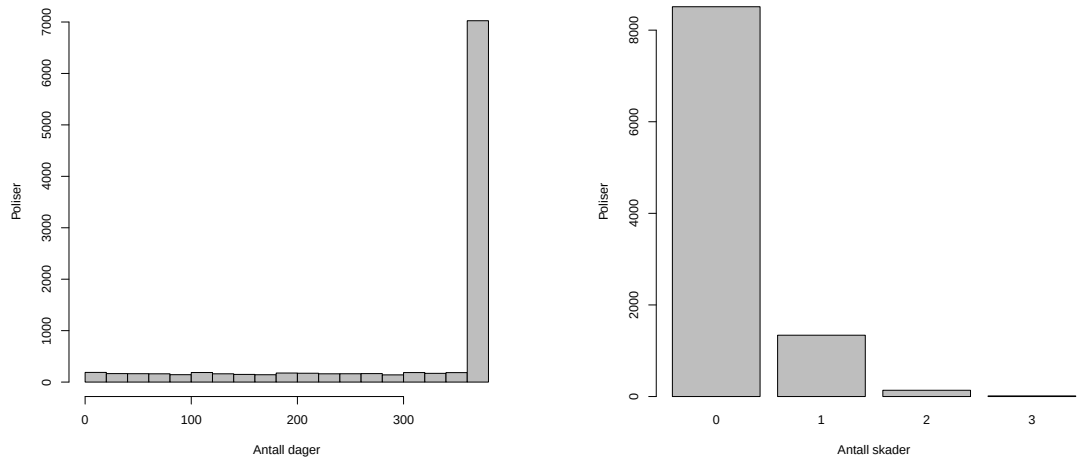
Antall skader og eksponeringstid

For hver enkelt polise er det registrert et skadeantall. Ca 85% av polisene har ingen skader, og det høyeste antall skader hos en polise er tre. Tabell 5.5 oppsummerer en del deskriptive statistikker hos den simulerte forsikringsporteføljen. Som forventet ser man at gjennomsnittlig antall skader hos poliser hvor varigheten er mindre enn ett år er lavere enn hos poliser som har en varighet på et helt år. Dette er ikke overraskende og er et resultat av fordelingen til N_i-ene gitt ved (5.1).

Beskrivelse	Alle poliser	Poliser hvor e_i =365	Poliser hvor e_i <365
Antall poliser	10000	6972	3028
Snitt	0.1652	0.1932014	0.1007266
Median	0	0	0
Empirisk varians	0.1731263	0.1992193	0.1071286

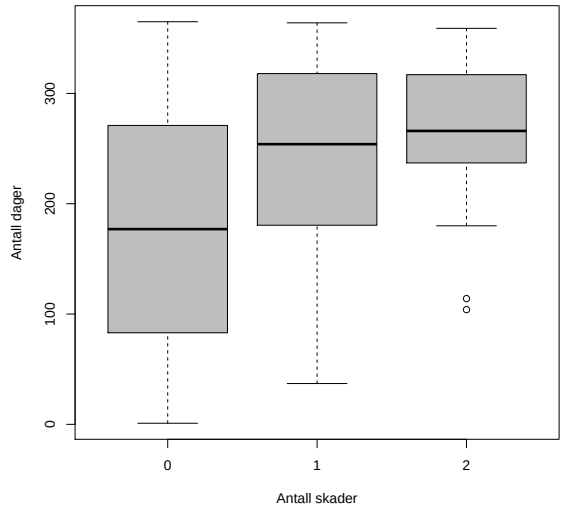
Tabell 5.5: Sammendrag av enkel deskriptiv statistikk for forsikringsporteføljen. Skiller mellom poliser som har en varighet på ett år og poliser som har varighet mindre enn ett år.

Både figur 5.6a og tabell 5.5 viser at poliser med en varighet på ett år dominerer forsikringsporteføljen. Årsaken til at dette inkluderes i simuleringen er for å gjøre den simulerte forsikringsporteføljen så realistisk som mulig. Dette simuleres ved et tilfeldig utvalg fra en fordeling med utfallsrom 0 og 1 og punktsannsynligheter lik 0.3 og 0.7, hvor 1 indikerer at polisen har en varighet på ett år. Eksponeringstiden til poliser hvor utfallet er lik 0 blir trukket fra en diskret uniform fordeling. Polisene kan tolkes som nye avtaler som registreres i løpet av et år, og avtaler som avsluttes og ikke fornyes.



(a) **(b)**

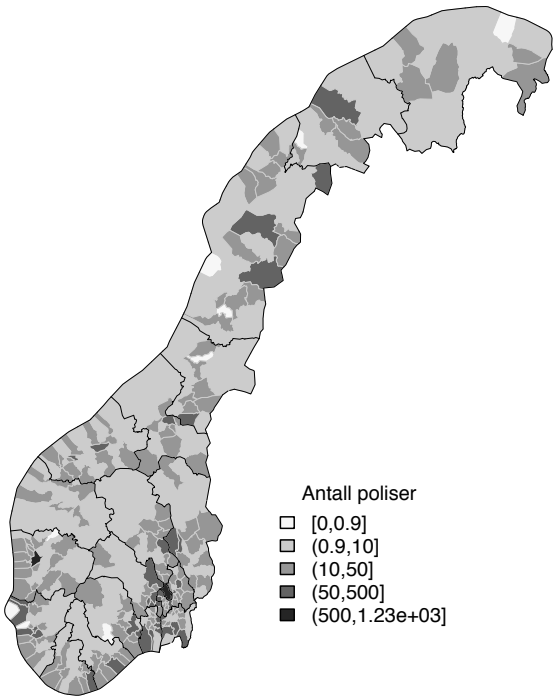
Figur 5.6: Fordelingen til eksponeringstid og antall skader hos den simulerte forsikringsporteføljen. Poliser med en varighet på ett år dominerer, i likhet med poliser hvor skadeantallet er lik null.



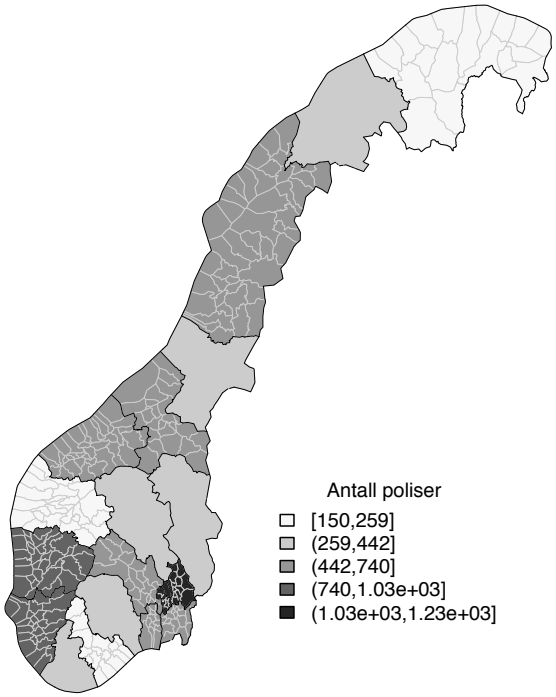
Figur 5.7: Sammenheng mellom eksponeringstid og antall skader. Medianen til eksponeringstiden er økende opp til to skader. Det er også to tilfeller hvor eksponeringstiden er lav og det har inntruffet to skader. Merk at poliser med en varighet på ett år ikke er inkludert.

Geografisk fordeling av poliser

Den geografiske fordelingen av poliser er basert på befolkningsandelene i de ulike kommunene. Ca 10% av polisene er tildelt Oslo. Her kunne det vært aktuelt å endre simuleringen når det gjelder tildeling av kommunenummer som følge av at Oslo er såpass sterkt representert. Ellers er det 9 kommuner som ikke har blitt tildelt noen poliser, illustrert ved hvit farge i figur 5.8a. Fordelingen av poliser ved kommunenivå og fylkesnivå er illustrert i henholdsvis figur 5.8a og 5.8b.



(a)



(b)

Figur 5.8: Geografisk fordeling av poliser, aggregert ved: (a) kommuner (b) fylker.

Kapittel 6

Inferens og estimering av parametere

6.1 Maksimum likelihood

Maksimum likelihood er en veletablert og meget populær metode for estimering av parametere. Dobson og Barnett (2011) gir en kort introduksjon til teorien ved å ta utgangspunkt i at man har en stokastisk vektor $\mathbf{y} = [Y_1, \ldots, Y_n]^T$, hvor simultantettheten gitt ved

$$f(\mathbf{y}; \boldsymbol{\theta}) \tag{6.1}$$

avhenger av parametervektoren $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$.[1]

Likelihoodfunksjonen $L(\boldsymbol{\theta}; \mathbf{y})$ er rent algebraisk identisk med $f(\mathbf{y}; \boldsymbol{\theta})$. Forskjellen mellom de to funksjonene er at man antar at $L(\boldsymbol{\theta}; \mathbf{y})$ og $f(\mathbf{y}; \boldsymbol{\theta})$ henholdsvis er funksjoner av $\boldsymbol{\theta}$ og $\mathbf{y}$. Videre blir alle gyldige verdier av $\boldsymbol{\theta}$, kalt parameterrommet til $\boldsymbol{\theta}$, definert ved Ω.

Maksimum likelihood estimatoren til $\boldsymbol{\theta}$ er verdien $\hat{\boldsymbol{\theta}}$ som maksimerer $L(\boldsymbol{\theta}; \mathbf{y})$ slik at

$$L(\hat{\boldsymbol{\theta}}; \mathbf{y}) \geq L(\boldsymbol{\theta}; \mathbf{y}) \qquad \text{for alle } \boldsymbol{\theta} \in \Omega. \tag{6.2}$$

$\hat{\boldsymbol{\theta}}$ vil også være verdien som maksimerer log-likehood-funksjonen, gitt ved $l(\boldsymbol{\theta}; \mathbf{y}) = \log L(\boldsymbol{\theta}; \mathbf{y})$, som et direkte resultat av at $\log(\cdot)$ er en monotont økende funksjon. Hovedårsaken til at log-likelihood-funksjonen introduseres er at det ofte viser seg å være enklere å maksimere $l(\boldsymbol{\theta}; \mathbf{y})$, sammenlignet med $L(\boldsymbol{\theta}; \mathbf{y})$, mht. $\boldsymbol{\theta}$.

For å finne $\hat{\boldsymbol{\theta}}$ deriverer man $l(\boldsymbol{\theta}; \mathbf{y})$ mht. hvert enkelt element i $\boldsymbol{\theta}$, og løser ligningene gitt ved:

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_j} = 0 \qquad j = 1, \ldots, p. \tag{6.3}$$

[1] Merk at parametervektoren $\boldsymbol{\theta}$ ikke skal forveksles med de randomiserte effektene gitt ved (4.8).

Betingelsene gitt ved (6.3) er nødvendige, men ikke suffisiente, betingelser for at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ tilfredsstiller (6.2) Det er nødvendig å bekrefte at løsningene på ligningene faktisk maksimerer funksjonen $l(\boldsymbol{\theta}; \mathbf{y})$. Dette blir gjort ved å verifisere at $p \times p$ matrisen, kalt Hessematrisen, bestående av elementene

$$\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_j \theta_k} \tag{6.4}$$

er negativ-definit evaluert ved $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Dersom dette er tilfellet kan man slå fast at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ er et lokalt maksimum. I det enkleste tilfellet, hvor $\boldsymbol{\theta}$ er en skalar, vil løsningen på ligningen være et lokalt maksimum hvis og bare hvis

$$\left[\frac{\mathrm{d}^2 l(\theta; \mathbf{y})}{\mathrm{d}\theta^2}\right]_{\theta=\hat{\theta}} < 0 \tag{6.5}$$

er tilfredsstilt. I tillegg til dette er det også nødvendig å undersøke om det eksisterer lokale maksimum på kanten av parameterrommet Ω. Når alle lokale maksimum er identifisert vil den verdien av $\hat{\boldsymbol{\theta}}$ som maksimerer $l(\boldsymbol{\theta}; \mathbf{y})$ også være et globalt maksimum, og følgelig tilsvare maksimum likelihood estimatoren til $\boldsymbol{\theta}$.

En viktig egenskap ved maksimum likelihood estimatorer er at dersom $\hat{\boldsymbol{\theta}}$ er maksimum likelihood estimatoren til $\boldsymbol{\theta}$ og $g(\boldsymbol{\theta})$ er hvilken som helst funksjon av $\boldsymbol{\theta}$, da er maksimum likelihood estimatoren til $g(\boldsymbol{\theta})$ gitt ved $g(\hat{\boldsymbol{\theta}})$. For en nærmere beskrivelse av flere egenskaper og tilhørende beviser henvises det til Hastie et al. (2009)[8.2.2] og Casella og Berger (2002)[7.2.2]

Rent praktisk vil denne fremgangsmåten ofte være svært omfattende og vanskelig, og man velger heller numeriske metoder for å finne maksimum likelihood estimatorene. Estimeringsalgoritmen for maksimum likelihood estimatorer hos generaliserte lineære modeller er presentert i sin helhet i vedlegg A, og vil bli anvendt på et utvalg av modellene i kapittel 7.

6.2 Maksimum likelihood estimering av modeller med latente variabler

Modellene som ble presentert i delkapittel 4.2 og 4.3 inneholder latente variabler, noe som medfører at det ikke er mulig å bruke den tradisjonelle iterative ligningen gitt ved (2.18) for estimering av tilhørende parametere. Her presenteres fremgangsmåten for å estimere parametere i modeller som inneholder latente variabler, heretter kalt hierarkiske modeller.

Det blir tatt utgangspunkt i at man har en vektor $\mathbf{y} = [y_1, \ldots, y_n]^T$ bestående av observasjoner og en stokastisk vektor $\mathbf{u} = [u_1, \ldots, u_q]^T$ bestående av latente variabler. Videre antar man at den betingede tetthetsfunksjonen til $\mathbf{y}$ gitt $\mathbf{u}$ er lik $f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u})$ og den marginale

tetthetsfunksjonen til $\mathbf{u}$ er gitt ved $h_{\boldsymbol{\theta}}(\mathbf{u})$. Tolkningen av indekseringen $\boldsymbol{\theta}$, gitt ved $f_{\boldsymbol{\theta}}$ og $h_{\boldsymbol{\theta}}$, er at både f og h avhenger av en ukjent parametervektor $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]^T$.

Gitt antagelsene over er den marginale likelihoodfunksjonen til $\boldsymbol{\theta}$, hvor $\mathbf{u}$ er integrert ut av $f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u})h_{\boldsymbol{\theta}}(\mathbf{u})$, gitt ved:

$$L(\boldsymbol{\theta}) = \int f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u})h_{\boldsymbol{\theta}}(\mathbf{u})\,\mathrm{d}\mathbf{u}. \tag{6.6}$$

Dette integralet vil typisk være høydimensjonalt og umulig å finne på lukket form. Løsningen på dette vil være å ta i bruk numeriske metoder for å finne en tilnærmet løsning på integralet.

I den neste seksjonen blir det gitt en generell introduksjon til Laplace-approksimasjon[2], en mye anvendt for å finne tilnærmede løsninger på høydimensjonale integral. Deretter blir det vist hvordan det er mulig å finne en tilnærmet løsningen på integralet gitt ved (6.6) ved å ta i bruk Laplace-approksimasjon, samt en beskrivelse av estimeringsprosedyren forbundet med parametervektoren $\boldsymbol{\theta}$.

Laplace-approksimasjon

Laplace-approksimasjon blir benyttet for å finne tilnærmede løsninger til integraler på formen

$$\int \mathrm{e}^{M\cdot g(\mathbf{u})}\,\mathrm{d}\mathbf{u} \tag{6.7}$$

hvor $\mathbf{u} = [u_1, \ldots, u_n]^T$ og $g(\cdot)$ er en skalarfunksjon av $\mathbf{u}$. Det første steget for å finne en approksimert løsning til (6.7) er å anvende en andre ordens Taylorutvikling på funksjonen $g(\mathbf{u})$ rundt $\mathbf{u}_0$, slik at

$$g(\mathbf{u}) \approx g(\mathbf{u}_0) + \nabla g(\mathbf{u}_0)(\mathbf{u} - \mathbf{u}_0) + \frac{1}{2}(\mathbf{u} - \mathbf{u}_0)^T \nabla^2 g(\mathbf{u}_0)(\mathbf{u} - \mathbf{u}_0), \tag{6.8}$$

hvor $\nabla g(\mathbf{u}_0)$ og $\nabla^2 g(\mathbf{u}_0)$ er henholdsvis gradienten og Hessematrisen til $g(\mathbf{u})$ evaluert i $\mathbf{u}_0$. Dersom $g(\mathbf{u})$ har et unikt globalt maksimum i $\hat{\mathbf{u}}$ vil dette medføre at $\nabla g(\hat{\mathbf{u}})$ er lik $\mathbf{0}$ og $\nabla^2 g(\hat{\mathbf{u}})$ er negativ-definit. Som et resultat av dette vil en Taylorutvikling av $g(\mathbf{u})$ rundt $\hat{\mathbf{u}}$ være gitt ved:

$$g(\mathbf{u}) \approx g(\hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}})^T \nabla^2 g(\hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}}) \tag{6.9}$$

Innsatt i (6.7) har man at den approksimerte løsningen er gitt ved:

$$\int \mathrm{e}^{M\cdot g(\mathbf{u})}\,\mathrm{d}\mathbf{u} \approx \mathrm{e}^{M\cdot g(\hat{\mathbf{u}})} \int \mathrm{e}^{\frac{1}{2}(\mathbf{u}-\hat{\mathbf{u}})^T M\cdot\nabla^2 g(\hat{\mathbf{u}})(\mathbf{u}-\hat{\mathbf{u}})}\,\mathrm{d}\mathbf{u}. \tag{6.10}$$

[2] http://en.wikipedia.org/wiki/Laplace%27s_method

Videre ser man at integranden tilsvarer kjernetettheten til en multivariat normalfordeling, med forventningsvektor $\hat{\mathbf{u}}$ og kovariansmatrise lik $[-M \cdot \nabla^2 g(\hat{\mathbf{u}})]^{-1}$. Dette medfører at den approksimerte løsningen på integralet er gitt ved:

$$\int \mathrm{e}^{M \cdot g(\mathbf{u})} \, \mathrm{d}\mathbf{u} \approx \left(\frac{2\pi}{M}\right)^{n/2} | - \nabla^2 g(\hat{\mathbf{u}})|^{-1/2} \mathrm{e}^{M \cdot g(\hat{\mathbf{u}})}, \tag{6.11}$$

hvor $| - \nabla^2 g(\hat{\mathbf{u}})|$ tilsvarer determinanten til matrisen $-\nabla^2 g(\hat{\mathbf{u}})$. Dette kommer som et resultat av at integralet over hele definisjonsområdet, per definisjon, er lik 1.

Skaug og Fournier (2006) presenterte en metode for å automatisere approksimasjonen av den marginale likelihoodfunksjonen hos ikke-gaussiske hierarkiske modeller ved hjelp av Laplace-approksimasjon og automatisk derivasjon. Følgende seksjon gir en introduksjon til nevnte artikkel.

Approksimert likelihoodfunksjon i hierarkiske modeller

Som tidligere nevnt blir det tatt utgangspunkt i at man har en vektor $\mathbf{y} = [y_1, \ldots, y_n]^T$ bestående av observasjoner og en stokastisk vektor $\mathbf{u} = [u_1, \ldots, u_q]^T$ bestående av latente variabler. De tilhørende tetthetsfunksjonene er definert $f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u})$ og $h_{\boldsymbol{\theta}}(\mathbf{u})$, slik at den marginale likelihoodfunksjonen gitt ved

$$L(\boldsymbol{\theta}) = \int f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u}) h_{\boldsymbol{\theta}}(\mathbf{u}) \, \mathrm{d}\mathbf{u} = \int \exp\{g(\mathbf{u}, \boldsymbol{\theta})\} \, \mathrm{d}\mathbf{u}, \tag{6.12}$$

hvor $g(\mathbf{u}, \boldsymbol{\theta}) = \log f_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{u}) + \log h_{\boldsymbol{\theta}}(\mathbf{u})$. I tillegg antas det at $g(\mathbf{u}, \boldsymbol{\theta})$ er slik at

$$\hat{\mathbf{u}}(\boldsymbol{\theta}) = \underset{\mathbf{u}}{\operatorname{argmax}} \; g(\mathbf{u}, \boldsymbol{\theta}) \tag{6.13}$$

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \mathbf{u}^2} g(\mathbf{u}, \boldsymbol{\theta})|_{\mathbf{u}=\hat{\mathbf{u}}(\boldsymbol{\theta})} \tag{6.14}$$

er definerbare gitt eventuelle begrensninger hos parameterrommet til $\boldsymbol{\theta}$.

Anvendelsen av Laplace-approksimasjon vil her være et spesialtilfelle av den mer generelle beskrivelsen gitt ved (6.11). Ved å sette $M = 1$, og deretter anvende en andre ordens Taylor-utvikling for $g(\mathbf{u}(\boldsymbol{\theta}), \boldsymbol{\theta})$ rundt $\hat{\mathbf{u}}(\boldsymbol{\theta})$ er følgende tilfredsstilt:

$$L(\boldsymbol{\theta}) \approx L^*(\boldsymbol{\theta}) = \mathrm{e}^{g(\hat{\mathbf{u}}(\boldsymbol{\theta}), \boldsymbol{\theta})} \int \mathrm{e}^{\frac{1}{2}(\mathbf{u}(\boldsymbol{\theta}) - \hat{\mathbf{u}}(\boldsymbol{\theta}))^T \mathbf{H}(\boldsymbol{\theta})(\mathbf{u}(\boldsymbol{\theta}) - \hat{\mathbf{u}}(\boldsymbol{\theta}))} \, \mathrm{d}\mathbf{u}. \tag{6.15}$$

Gitt antagelsene i (6.13) vil matrisen $\mathbf{H}(\boldsymbol{\theta})$, gitt ved (6.14), være negativ-definit, som igjen medfører at matrisen $-\mathbf{H}(\boldsymbol{\theta})$ er positiv-definit, og dermed invertibel. Dette medfører at approksimasjonen er gitt ved:

$$L^*(\boldsymbol{\theta}) = (2\pi)^{q/2} \det\{-\mathbf{H}(\boldsymbol{\theta})^{-1}\}^{1/2} \mathrm{e}^{g(\hat{\mathbf{u}}(\boldsymbol{\theta}), \boldsymbol{\theta})} \tag{6.16}$$

Den approksimerte marginale likelihoodfunksjonen, gitt ved (6.16), forenkles ytterligere ved å ta i bruk kjente regneregler for determinanter, slik at $\det\{-\mathbf{H}(\boldsymbol{\theta})^{-1}\} = \det\{-\mathbf{H}(\boldsymbol{\theta})\}^{-1}$ og $\det\{-\mathbf{H}(\boldsymbol{\theta})\} = (-1)^q \det\{\mathbf{H}(\boldsymbol{\theta})\}$. I tillegg kan man observere at $\det\{-\mathbf{H}(\boldsymbol{\theta})\} = |\det\{\mathbf{H}(\boldsymbol{\theta})\}|$.

Den approksimerte log-likelihood-funksjonen er gitt ved

$$l^*(\boldsymbol{\theta}) = -\frac{1}{2}|\det\{\mathbf{H}(\boldsymbol{\theta})\}| + g(\hat{u}(\boldsymbol{\theta}), \boldsymbol{\theta}) \tag{6.17}$$

hvor konstantleddet er utelatt.

I selve maksimeringsprosessen vil ikke $\hat{\mathbf{u}}(\boldsymbol{\theta})$ og $\mathbf{H}(\boldsymbol{\theta})$, gitt ved henholdsvis (6.13) og (6.14), være eksplisitt gitt. $\hat{\mathbf{u}}(\boldsymbol{\theta})$ vil typisk bli funnet ved en standard ikke-lineær optimeringsalgoritme (*quasi-Newton algorithm* eller *limited memory Newton method*), mens $\mathbf{H}(\boldsymbol{\theta})$ evalueres ved hjelp av automatisk derivasjon.

Som et resultat av at $-\mathbf{H}(\boldsymbol{\theta})$ er positiv-definit er man i stand til å utføre en Cholesky dekomponering av $-\mathbf{H}(\boldsymbol{\theta})$, slik at $-\mathbf{H}(\boldsymbol{\theta}) = \boldsymbol{L}\boldsymbol{L}^T$, hvor $\boldsymbol{L}$ er en nedre triangulær matrise med strengt positive diagonalelementer. Determinanten til $-\mathbf{H}(\boldsymbol{\theta})$ er gitt ved $\det(\boldsymbol{L})^2$, som et resultat av at $\det\{-\mathbf{H}(\boldsymbol{\theta})\} = \det(\boldsymbol{L})\det(\boldsymbol{L}^T) = \det(\boldsymbol{L})^2 > 0$. $\boldsymbol{L}$ er som nevnt en nedre triangulær matrise og $\det(\boldsymbol{L})$ tilsvarer produktet av diagonalelementene.

Det er naturlig å se på optimeringsprosedyren som et nøstet optimeringsproblem, hvor optimeringsproblemet i (6.13) og evalueringen av (6.14) tilsvarer det «indre» problemet, og det «ytre» problemet er selve maksimeringen av Laplace-approksimasjonen gitt ved (6.17) mht. $\boldsymbol{\theta}$. I det første steget antas det at $\mathbf{u} = \mathbf{0}$, slik at man står ovenfor et standard optimeringsproblem hvor man ikke benytter seg av Laplace-approksimasjon. Dette resulterer i estimater for en delmengde av $\boldsymbol{\theta}$, som i det neste steget blir benyttet for å løse optimeringsproblemet gitt i (6.13). Dette vil være en suksessiv prosedyre hvor estimatene for $\boldsymbol{\theta}$ og $\mathbf{u}(\boldsymbol{\theta})$ oppdateres når man løser det «indre» og «ytre» problemet. Estimeringsprosedyren stoppes når man har oppnådd de valgte konvergenskriteriene beskrevet i Skaug og Fournier (2006).

6.3 Implementering

I denne oppgaven anvendes Template Model Builder (TMB) for å estimere parametere i de hierarkiske modellene. Pakken er utviklet av Kristensen (2014), og bygger i stor grad på en lignende pakke, utviklet av Fournier et al. (2012), kalt Automatic Differentiation Model Builder (ADMB).

TMB er en R-pakke som er utviklet for å tilpasse modeller med latente variabler til data. Dette blir gjort ved å anvende Laplace-approksimasjon og automatisk derivasjon, samt

algoritmer fra R vedrørende Cholesky dekomponering. Pakken er formulert ved programspråket C++, noe som medfører større fleksibilitet.

Det første steget, når man benytter seg av TMB, er å lage en cpp-fil, hvor programspråket C++ blir tatt i bruk. I denne filen deklareres alle parametere, samt datamaterialet som skal tilpasses modellen og likelihoodfunksjonen til modellen. I denne oppgaven vil det være mindre forskjeller hos cpp-filene som blir benyttet for de ulike modellene. Årsaken til dette er at det hovedsaklig er input for data som varierer blant de ulike modellene, og formen på likelihoodfunksjonen er relativt lik.

Dersom man skal estimere parametere som har naturlige begrensninger, eksempelvis standardavviket i en normalfordeling gitt ved σ, har man to muligheter for hvordan man implementere dette. Den første muligheten er å spesifisere dette eksplisitt når man optimerer funksjonen i R. Den andre muligheten vil være å implementere dette direkte i cpp.filen, ved hjelp av enkle transformasjoner. Istedenfor å estimere parameteren σ velger man heller å estimere parameteren $\tau = \log(\sigma)$, slik at $\sigma = \exp(\tau)$. Dette vil sikre at estimatet av σ er større eller lik 0. Her vil det faktumet at maksimum likelihood estimatorer er invariante bli tatt i bruk, slik at $\hat{\sigma} = \exp(\hat{\tau})$. Ved hjelp av enkle kommandoer vil TMB levere estimater for både τ og σ, alt etter behov.

TMB vil også produsere standardavvik for både det faste parametere og de latente variablene. Dersom modellen ikke inneholder latente variabler finner man standardavvikene for parameterne ved hjelp av delta-metoden, og dersom modellen inneholder latente variabler vil standardavvikene for både de faste parameterne og de latente variablene bli funnet ved en generalisert delta-metode.

Vedlegg B.2 presenterer et eksempel på bruk av TMB, samt tilhørende R-koder.

Kapittel 7

Resultater

I dette kapittelet presenteres resultater for ulike modeller som blir tilpasset den simulerte forsikringsporteføljen i kapittel 5. Hovedmålet er å analysere effekten av å inkludere latente variabler, og om hvorvidt inkluderingen av latente variabler vil kunne kompensere for manglende forklaringsvariabler. Det blir først gitt en kort oppsummering av metoder og resultater hos Dimakos og Di Rattalma (2002), samt forskjeller mellom nevnte artikkel og denne oppgaven. Videre blir det gitt en oversikt over de ulike modellene som blir anvendt i denne oppgaven, og resultatene etter at modellene har blitt tilpasset forsikringsporteføljen. Valideringen av den prediktive evnen til modellene følger deretter, og til slutt blir det gitt en kort oppsummering og konklusjon.

7.1 Innledning

Dimakos og Di Rattalma (2002) undersøkte, ved hjelp av en simulert forsikringsportefølje, om latente variabler kunne kompensere for manglende forklaringsvariabler. Den simulerte forsikringsporteføljen bestod av forklaringsvariablene urbaniseringsgrad, kriminalitetsrate, inntekt og kjønn. Hver enkelt polise ble tildelt et fylkesnummer, og det ble utført to ulike eksperiment. I det første eksperimentet ble forklaringsvariablene kriminalitetsrate og urbaniseringsgrad fjernet, og i det andre eksperimentet ble inntekt fjernet. For hvert eksperiment ble det tilpasset totalt 14 modeller, hvor en av modellene tilsvarte den sanne modellen som ble brukt til å simulere forsikringsporteføljen.

I de påfølgende avsnittene blir de ulike aprioriforde lingene til de latente variablene som anvendes hos Dimakos og Di Rattalma (2002) presentert, samt hvilke konklusjoner som ble trukket basert på resultatene.

En av modellene tilsvarte en generalisert lineær blanda modell (GLMM), hvor apriori-fordelingen til de latente variablene ble spesifisert ved:

$$\begin{aligned} \gamma_j|\mu,\sigma^2 &\sim \mathcal{N}(\mu,\sigma^2), \quad j=1,\ldots,19 \\ \mu &\sim \mathcal{N}(a,b) \\ 1/\sigma^2 &\sim \text{Gamma}(c,d) \end{aligned} \tag{7.1}$$

For de resterende modellene tok Dimakos og Di Rattalma (2002) utgangspunkt i et spesialtilfelle av modell (i) presentert i tabell 3.1, hvor $w_{ij} = 1$ dersom fylke i og j defineres som naboer. Simultantettheten til de latente variablene er følgelig gitt ved:

$$\boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma}) \propto \exp\left\{-\frac{\kappa}{2}\sum_j \sum_{k\in\delta_j}(\gamma_j-\gamma_k)^2\right\}, \tag{7.2}$$

hvor $\kappa = 1/\sigma^2$ og $k \in \delta_j$ hvis og bare hvis fylke k og j er naboer.

Dimakos og Di Rattalma (2002) estimerte ikke parameteren κ, men tilpasset modeller hvor man antok at κ var konstant og lik $\{2,4,6,8,10,20,30,40,50,60\}$. I tillegg til dette ble det inkludert to generaliserte lineære modeller, hvor den ene hadde et felles skjæringspunkt for alle fylkene og den andre tok utgangspunkt i at de ulike fylkene kunne ha unike skjæringspunkter.

Dimakos og Di Rattalma (2002) simulerte tusen nye porteføljer for å trekke slutninger om hvorvidt inkluderingen av latente variabler kunne kompensere for manglende forklaringsvariabler. Metodikken som ble anvendt i valideringen av den prediktive evnen til modellene er presentert i delkapittel 7.4.

Resultatene i det første eksperimentet viste at modellene hvor $\kappa = \{30,40,50\}$ var signifikant bedre enn de resterende modellene, med unntak av $\kappa = \{20,60\}$, når det gjaldt underestimering av forventet antall skader. Tilsvarende var modellene hvor $\kappa = \{20,30\}$ signifikant bedre enn de resterende modellene, med unntak av $\kappa = \{6,8,10,40,50,60\}$, når det gjaldt overestimering. I det andre eksperimentet viste det seg at modellene med korrelerte latente variabler ikke bidrog med noe mer forklaringskraft, og de mer tradisjonelle modellene var å foretrekke.

Konklusjonen som ble trukket basert på de to eksperimentene var at dersom forklaringsvariablene som ble utelatt hadde en relativt glatt geografisk fordeling, noe som var tilfellet hos kriminalitetsrate og urbaniseringsgrad, ville inkluderingen av latente variabler øke prediksjonskraften til modellen.

I denne oppgaven utføres lignende eksperimenter, men med enkelte modifikasjoner og endringer. De utføres tre ulike eksperiment, og de viktigste forskjellene kan sammenfattes i følgende fem punkter:

1. Simulering av forsikringsportefølje

Hovedforskjellen mellom simuleringene er at hver enkelt polise ble tildelt et fylkesnummer hos Dimakos og Di Rattalma (2002), mens polisene i denne oppgaven blir tildelt et kommunenummer. Dette resulterer i at dimensjonen til vektoren som representerer de latente variablene øker fra 19 til 429. Kommunenummeret vil dermed bestemme verdien av forklaringsvariablene knyttet til urbaniseringsgrad og kriminalitetsrate. I tillegg til dette er forklaringsvariabelen kjønn fjernet, samtidlig som alder er inkludert.

2. Valg av fordeling for latente variabler

Dimakos og Di Rattalma (2002) tok utgangspunkt i aprioriforde lingene gitt ved (7.1) og (7.2). Fordelingene til de latente variablene i denne oppgaven er gitt ved (4.8) og (4.9), og vil ikke bli omtalt som aprioriforde linger. Dimakos og Di Rattalma (2002) inkluderte én latent variabel per fylke, mens det i denne oppgaven vil bli inkludert to latente variabler per kommune.

3. Estimering av parametere

Dimakos og Di Rattalma (2002) estimerte parametere og latente variabler ved MCMC-metoder. I denne oppgaven vil problemstilling bli sett på fra et frekventisk ståsted, hvor parametere og latente variabler estimeres ved maksimum likelihood. Parameterne τ_h og τ_c i henholdsvis (4.8) og (4.9), samt δ i (4.9), estimeres direkte. Dimakos og Di Rattalma (2002) estimerte ikke κ i (7.2), men utførte en sensitivitetsanalyse vedrørende parameteren. Merk at det er naturlig å sammenligne τ_c med κ.

4. Valg av nabostruktur

Dimakos og Di Rattalma (2002) valgte å definere to fylker som naboer dersom grensene til fylkene møttes i minimum ett punkt. I denne oppgaven blir det anvendt fire ulike former for nabostrukturer, presentert i seksjon 7.2.1.

5. Generaliserte lineære modeller

Dimakos og Di Rattalma (2002) inkluderte to modeller som tilfredsstilte egenskapene til GLM i delkapittel 2.2. Forskjellen mellom de to modellene var at den ene hadde et unikt skjæringspunkt for hvert enkelt fylke, mens den andre hadde et felles skjæringspunkt for alle fylkene. Som følge av at det er 429 kommuner vil det i denne oppgaven ikke bli tilpasset en modell med unike skjæringspunkt for de ulike kommune.

Ellers er fokuset begrenset mot følgende tre situasjoner:

Eksperiment 1: Urbaniseringsgrad og kriminalitetsrate mangler.

Eksperiment 2: Urbaniseringsgrad mangler.

Eksperiment 3: Kriminalitetsrate mangler.

For de tre eksperimentene over vil ulike modeller bli foreslått og tilpasset den simulerte forsikringsporteføljen. Den sanne modellen, gitt ved (5.1), vil også tilpasses den simulerte forsikringsporteføljen, og vil kunne fungere som et slags sammenligningsgrunnlag.

Totalt vil det være fem modeller for hver av de tre eksperimentene. Den første modellen er en generalisert lineær modell, og tilpasses ved hjelp av den tradisjonelle iterative ligningen gitt ved (2.18). De resterende modellene, hvor latente variabler inkluderes, blir tilpasset ved hjelp av TMB (delkapittel 6.3). Resultatene for modellene i de tre eksperimentene er presentert i tabell 7.1, 7.2 og 7.3.

I de neste seksjonene presenteres de ulike nabostrukturene og modellene, samt resultatene for de tilpassede modellene. Valideringen av prediksjonskraften til modellene, samt tilhørende beskrivelse av metodikk, blir først presentert i delkapittel 7.4.

7.2 Modeller

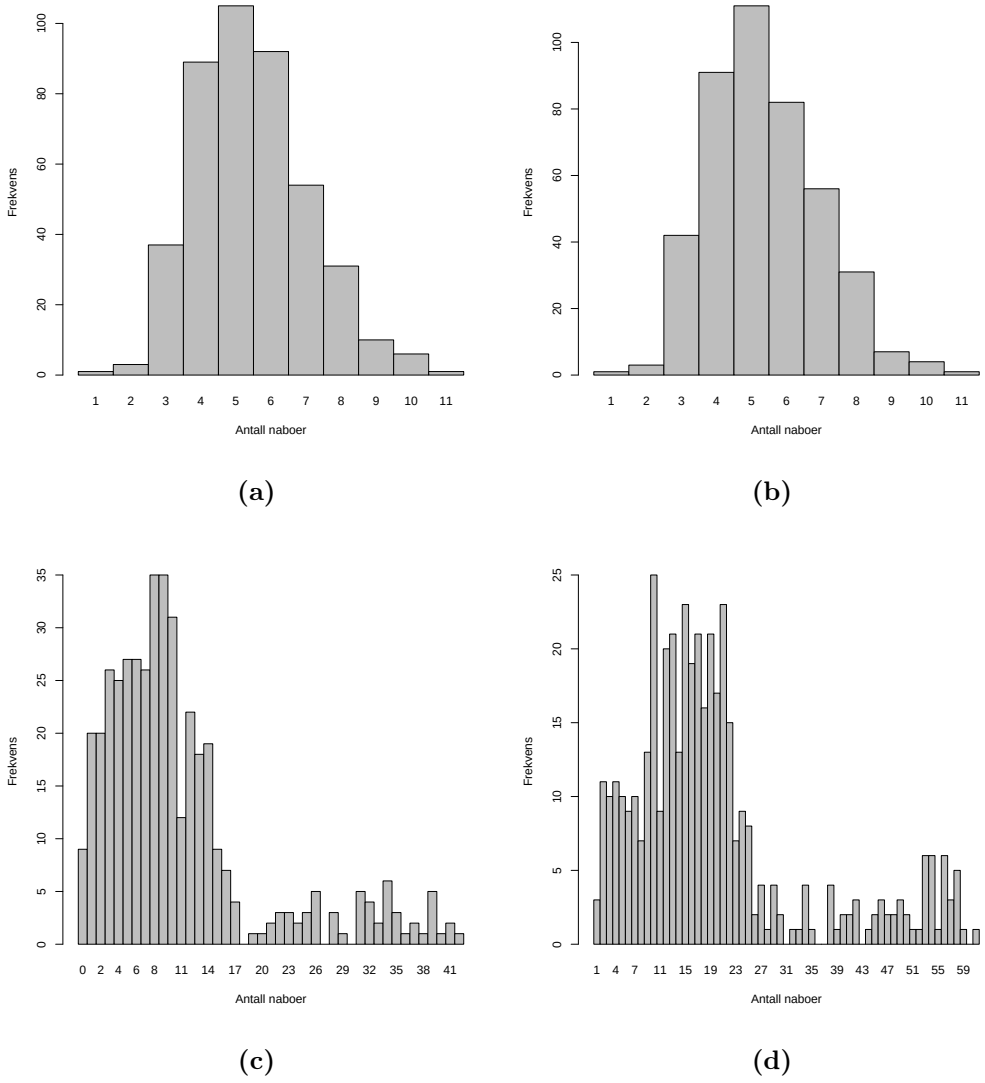
Datagrunnlaget er presentert i sin helhet i kapittel 5, og består av 10000 observasjoner og tilhørende forklaringsvariabler. I modellene vil personer med lav inntekt i aldersgruppe 3 bli brukt som referansenivå. De øvrige forklaringsvariablene som ble brukt i simuleringen er kriminalitetsrate og urbaniseringsgrad.

7.2.1 Valg av nabostruktur

Det blir tatt i bruk fire ulike definisjoner for nabokommuner. De to første er velkjente og mye brukt, og de to sistnevnte inkluderes for å illustrere hvilke valgmuligheter man har og hvilken effekt valgene har på resultatene. Definisjonene er som følger.

To kommuner er naboer hvis og bare hvis:

1. Grensene møtes i minimum ett punkt.
2. Grensene møtes i minimum to punkter.
3. Avstanden mellom de administrative punktene er mindre enn 50 km.
4. Avstanden mellom de administrative punktene er mindre enn 70 km.



Figur 7.1: Fordelingen til antall naboer hos de 429 kommunene i Norge ved de ulike definisjonene.

7.2.2 Modellbeskrivelse

I samtlige modeller antar man at antall skader, gitt ved $\mathbf{y} = [y_1, \ldots, y_n]^T$, er poissonfordelt. Videre vil man skille mellom modeller med og uten latente variabler, slik at modeller uten latente variabler er gitt ved

$$\begin{aligned} N_i|\eta_i &\sim \text{Poisson}\left(e_i \mathrm{e}^{\eta_i}\right) \\ \text{hvor } \eta_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}. \end{aligned} \tag{7.3}$$

Modeller med latente variabler er på formen:

$$\begin{aligned} N_i|\psi_i &\sim \text{Poisson}\left(e_i \mathrm{e}^{\psi_i}\right) \\ \text{hvor } \psi_i &= \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \theta_{K(i)} + \phi_{K(i)}. \end{aligned} \tag{7.4}$$

For nærmere beskrivelse av fordelingen til $\boldsymbol{\theta}$ og $\boldsymbol{\phi}$, og deres funksjon, refereres det til delkapittel 4.3.

Hos de ulike eksperimentene vil $\mathbf{x}_i^T$ og $\boldsymbol{\beta}$ være gitt ved:

Eksperiment 1: $\mathbf{x}_i^T = [x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i9}, x_{i10}] \quad \boldsymbol{\beta} = [\beta_4, \beta_5, \beta_6, \beta_7, \beta_9, \beta_{10}]^T$

Eksperiment 2: $\mathbf{x}_i^T = [x_{i2}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i9}, x_{i10}] \quad \boldsymbol{\beta} = [\beta_2, \beta_4, \beta_5, \beta_6, \beta_7, \beta_9, \beta_{10}]^T$

Eksperiment 3: $\mathbf{x}_i^T = [x_{i1}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i9}, x_{i10}] \quad \boldsymbol{\beta} = [\beta_1, \beta_4, \beta_5, \beta_6, \beta_7, \beta_9, \beta_{10}]^T$

Selve beskrivelsen av de ulike forklaringsvariablene er forøvrig presentert i tabell 5.1.

Hver enkelt modell blir tildelt et navn gitt ved MOD-$u.v$, hvor $u = \{1, 2, 3\}$ og $v = \{0, 1, 2, 3, 4\}$. Her vil u representere hvilket eksperiment modellen tilhører, og v representerer hvilken definisjon som har blitt benyttet for å definere nabostrukturen. Eksempelvis vil MOD-1.2 tilsvare modellen fra eksperiment 1, hvor nabostrukturen blir bestemt basert på definisjon 2. Dersom $v = 0$ indikerer dette at modellen er en standard GLM.

7.3 Resultater

Resultatene for de ulike modellene er gitt i tabell 7.1, 7.2 og 7.3. Følgende steg utføres for å sammenligne de ulike modellene innenfor hvert enkelt eksperiment:

1. Drøfter eventuelle forskjeller mellom de estimerte bidragene fra $\boldsymbol{\theta}$ og $\boldsymbol{\phi}$ for modellene med latente variabler, med spesielt fokus rettet mot den estimerte korrelasjonsstrukturen hos $\boldsymbol{\phi}$ og tilhørende usikkerhet.

2. Velg en av modellene med latente variabler, basert på *Akaikes informasjonskriterium* (AIC), som den antatt beste modellen. For en nærmere beskrivelse av AIC refereres det til Akaike (1974).

3. Sammenligner den antatt beste modellen med latente variabler med tilsvarende GLM basert på følgende punkter:

    - AIC
    - Estimerte parametere
    - Samlet estimert spatial effekt

Den samlede estimerte spatiale effekten hos kommune j er gitt ved $e^{\hat{s}_j}$, hvor $\hat{s}_j$ tilsvarer summen av estimerte effekter fra eventuelle kommunespesifikke forklaringsvariabler og latente variabler. I MOD-1.0 vil den samlede estimerte spatiale effekten være lik 1, som et direkte resultat av at modellen hverken inneholder forklaringsvariabler som er kommunespesifikke eller latente variabler. I MOD-2.0 og MOD-3.0 er den samlede estimerte spatiale effekten for kommune j gitt ved henholdsvis $e^{\hat{\beta}_2 x_{j,2}}$ og $e^{\hat{\beta}_1 x_{j,1}}$. Tilsvarende vil den samlede estimerte spatiale effekten hos kommune j i MOD-1.v, MOD-2.v og MOD-3.v være gitt ved henholdsvis $e^{\hat{\phi}_j + \hat{\theta}_j}$, $e^{\hat{\beta}_2 x_{j,2} + \hat{\phi}_j + \hat{\theta}_j}$ og $e^{\hat{\beta}_1 x_{j,1} + \hat{\phi}_j + \hat{\theta}_j}$, hvor $v = 1, 2, 3, 4$.

7.3.1 Eksperiment 1

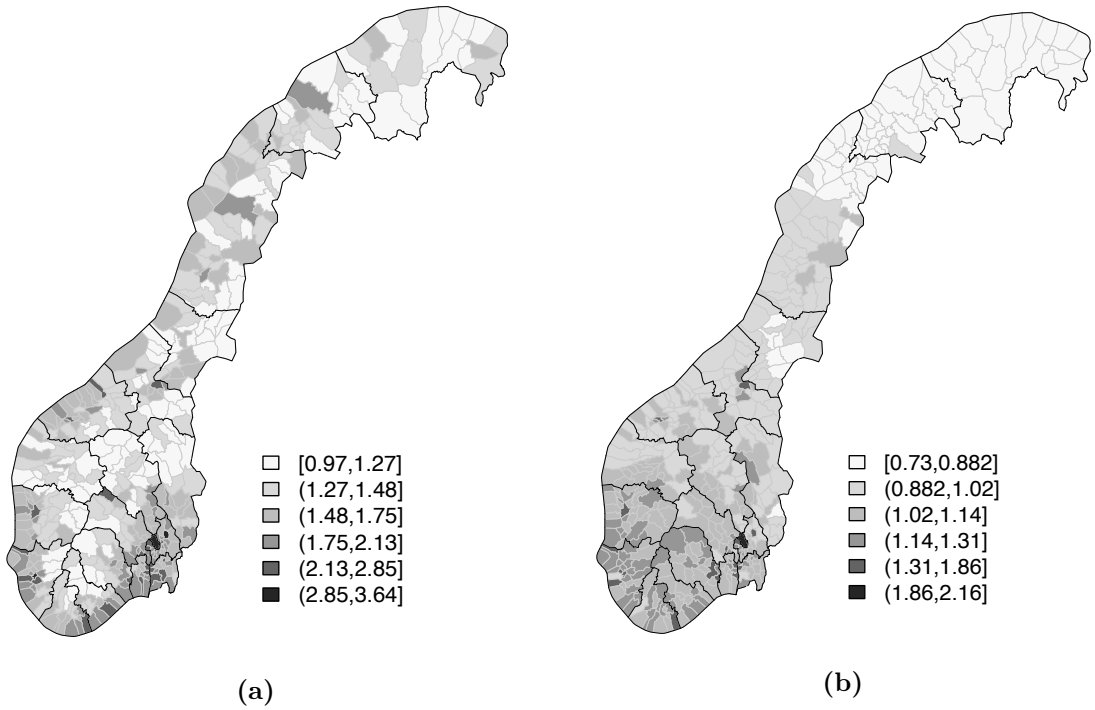
Ikke overraskende er det minimale forskjeller mellom MOD-1.1 og MOD-1.2, som følge av at den eneste forskjellen mellom modellene er at man hos MOD-1.2 ikke definerer to kommuner som naboer dersom grensene deres bare møtes i et enkelt punkt.

Nabostrukturen som danner grunnlaget for MOD-1.3 resulterer i at bidraget fra $\boldsymbol{\phi}$ er neglisjerbart som et direkte resultat av parameterestimatet til τ_c. Det er også verdt å merke seg at usikkerheten hos parameterestimatet til τ_c skiller seg ut hos MOD-1.3 sammenlignet med de

andre modellene. MOD-1.4 skiller seg fra MOD-1.1 og MOD-1.2 gjennom parameterestimatet til τ_c. Ellers er det verdt å merke seg at estimatene for τ_h ikke avviker spesielt fra hverandre. Dette kan være et resultat av at fordelingen til $\boldsymbol{\theta}$ ikke påvirkes direkte av hvilken nabostruktur man velger å bruke. Basert på AIC faller det seg naturlig å velge MOD.1-2 som den antatt beste modellen med latente variabler.

Sammenligner man de estimerte parameterverdiene for de individuelle forklaringsvariablene hos MOD-1.0 og MOD-1.2 med de sanne parameterverdiene i 5 viser det seg at MOD-1.2 presterer noe bedre enn MOD-1.0.

Figur 7.2a illustrerer den samlede effekten fra geografisk lokalitet på forventet antall skader fra fordelingen til N_i-ene gitt ved (5.1), mens figur 7.2b illustrer den estimerte effekten fra MOD-1.1. Resultatene viser at det er svært mange kommuner som har blitt tildelt lik risikoprofil, og det er en sterk form for glatting til stede. Blant annet ser man at øvre grensen er betydelig lavere i MOD.1-1 sammenlignet med den sanne fordelingen.



Figur 7.2: (a) Geografisk effekt på forventet antall skader. (b) Estimert geografisk effekt på forventet antall skader hos MOD-1.1.

	MOD-1.0	MOD-1.1	MOD-1.2	MOD-1.3	MOD-1.4
Intercept	-1.514***	-1.784***	-1.784***	-1.759***	-1.779***
	(0.069)	(0.125)	(0.125)	(0.079)	(0.114)
Urbaniseringsgrad	-	-	-	-	-
	-	-	-	-	-
Kriminalitetsrate	-	-	-	-	-
	-	-	-	-	-
Medium inntekt	-0.116**	-0.071	-0.071	-0.065	-0.069
	(0.055)	(0.056)	(0.056)	(0.056)	(0.056)
Høy inntekt	-0.723***	-0.671***	-0.671***	-0.663***	-0.667***
	(0.078)	(0.079)	(0.079)	(0.079)	(0.079)
Alder: 20-29	0.415***	0.375***	0.375***	0.375***	0.374***
	(0.077)	(0.077)	(0.077)	(0.077)	(0.077)
Alder: 30-39	0.111	0.068	0.068	0.071	0.068
	(0.081)	(0.081)	(0.081)	(0.081)	(0.081)
Alder: 50-69	-0.161**	-0.149*	-0.149*	-0.151**	-0.15*
	(0.077)	(0.077)	(0.077)	(0.077)	(0.077)
Alder: >69	0.108	0.14	0.14	0.14	0.139
	(0.088)	(0.088)	(0.088)	(0.088)	(0.088)
τ_c	-	37.571	37.198	476.995	17.689
	-	(37.168)	(37.42)	(5497.119)	(21.805)
τ_h	-	18.173***	18.246***	13.991***	16.738***
	-	(6.396)	(6.463)	(4.637)	(5.473)
δ	-	0.007	0.007	0.056	0.019
	-	(0.012)	(0.013)	(0.386)	(0.039)
Observasjoner	10000	10000	10000	10000	10000
Log likelihood	-4537.941	-4473.586	-4473.565	-4476.237	-4474.761
Akaike Inf.Crit	9082.883	8967.173	8967.13	8972.474	8969.522

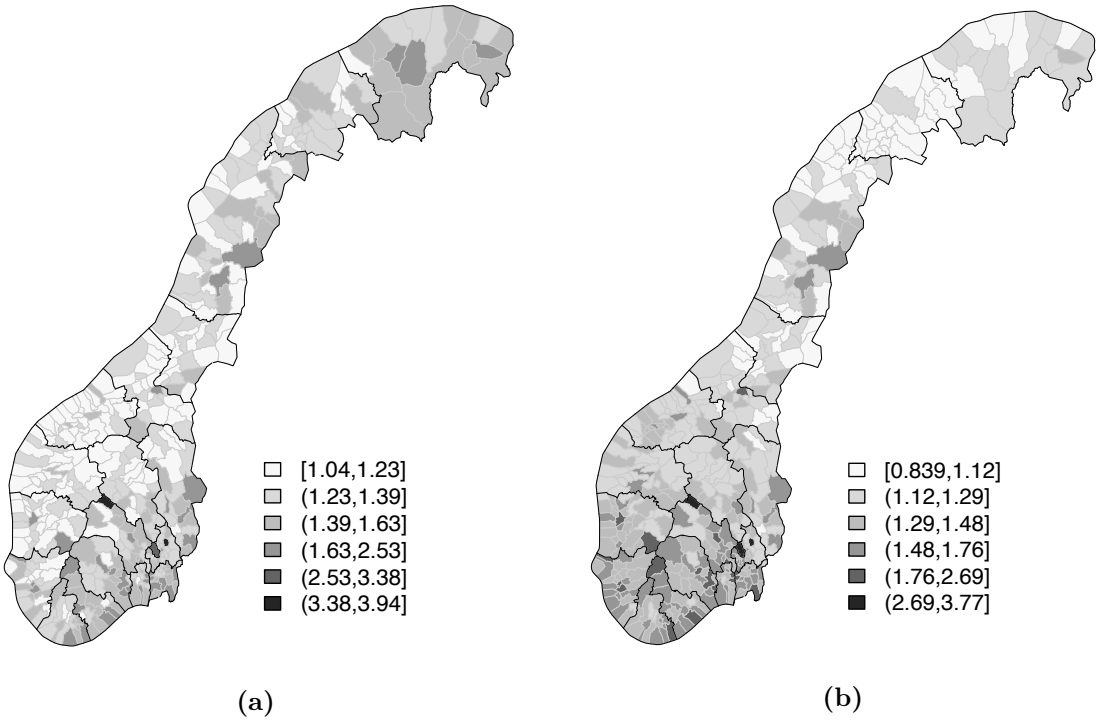
Note: *p<0.1; **p<0.05; ***p<0.01

Tabell 7.1: Resultater for modeller i eksperiment 1. Estimeringen av parametere i MOD-1.0 blir utført ved hjelp av den iterative ligningen gitt ved (2.18), mens de resterende modellene blir tilpasset ved hjelp av TMB. Verdier i parentes tilsvarer standardavvik.

7.3.2 Eksperiment 2

Også i dette eksperimentet er det minimale forskjeller mellom MOD-2.1 og MOD-2.2. De to modellene skiller seg i stor grad fra tilsvarende modeller i det første eksperimentet gjennom parameterestimatene til τ_h. I selve estimeringsprosedyren ble det valgt en øvre begrensning for τ_h, som også ble nådd, og det er valgt å definere dette som ∞. Rent praktisk vil dette si at det estimerte bidraget fra $\boldsymbol{\theta}$ er lik 0 , og i realiteten kunne være utelatt.

Nabostrukturen i MOD-2.3 viser seg, som i MOD-1.3, å resultere i et neglisjerbart bidrag fra $\boldsymbol{\phi}$. MOD-2.4 ser ut til å være en bedre tilpasning, og modellerer også et positivt bidrag fra $\boldsymbol{\theta}$. Basert på AIC antas det likevel at MOD-2.2 er det beste forslaget.



Figur 7.3: (a) Estimert geografisk effekt på forventet antall skader hos MOD-2.0. (b) Estimert geografisk effekt på forventet antall skader hos MOD-2.2

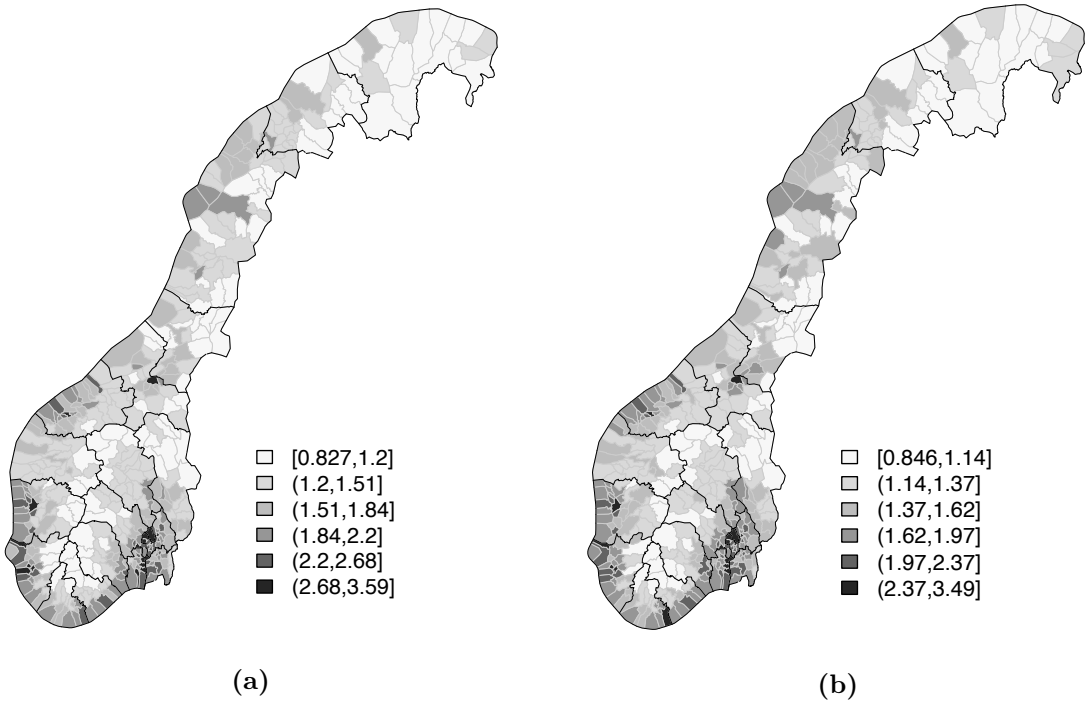
	MOD-2.0	MOD-2.1	MOD-2.2	MOD-2.3	MOD-2.4
Intercept	-2.091***	-2.108***	-2.109***	-2.086***	-2.094***
	(0.085)	(0.119)	(0.118)	(0.094)	(0.118)
Urbaniseringsgrad	-	-	-	-	-
	-	-	-	-	-
Kriminalitetsrate	0.009***	0.009***	0.009***	0.009***	0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Medium inntekt	-0.046	-0.056	-0.056	-0.05	-0.052
	(0.056)	(0.056)	(0.056)	(0.056)	(0.056)
Høy inntekt	-0.639***	-0.654***	-0.654***	-0.644***	-0.647***
	(0.078)	(0.079)	(0.079)	(0.079)	(0.079)
Alder: 20-29	0.376***	0.374***	0.374***	0.374***	0.373***
	(0.077)	(0.077)	(0.077)	(0.077)	(0.077)
Alder: 30-39	0.077	0.066	0.066	0.072	0.069
	(0.081)	(0.081)	(0.081)	(0.081)	(0.081)
Alder: 50-69	-0.161**	-0.156**	-0.156**	-0.158**	-0.158**
	(0.077)	(0.077)	(0.077)	(0.077)	(0.077)
Alder: >69	0.135	0.137	0.137	0.136	0.135
	(0.088)	(0.088)	(0.088)	(0.088)	(0.088)
τ_c	-	24.339	23.841	310.656	27.374
	-	(23.385)	(22.796)	(5449.768)	(36.314)
τ_h	-	∞	∞	57.156	105.836
	-	(∞)	(∞)	(61.225)	(169.455)
δ	-	0.016	0.017	0.31	0.015
	-	(0.035)	(0.037)	(4.71)	(0.032)
Observasjoner	10000	10000	10000	10000	10000
Log likelihood	-4464.94	-4457.234	-4457.192	-4460.075	-4458.912
Akaike Inf.Crit	8937.88	8936.469	8936.385	8942.15	8939.825

Note: *p<0.1; **p<0.05; ***p<0.01

Tabell 7.2: Resultater for modeller i eksperiment 2. Estimeringen av parametere i MOD-2.0 blir utført ved hjelp av den iterative ligningen gitt ved (2.18), mens de resterende modellene blir tilpasset ved hjelp av TMB. Verdier i parentes tilsvarer standardavvik.

7.3.3 Eksperiment 3

I dette eksperimentet viser det seg at bidraget fra $\boldsymbol{\phi}$ i modellene er neglisjerbart, og de ulike modellene med latente variabler blir identiske. Dette kommer som et direkte resultat av at fordelingen til $\boldsymbol{\theta}$ ikke påvirkes av hvordan man definerer kommuner som naboer. Siden hovedformålet med oppgaven er å illustrere modellering av romlig avhengighet, vil ikke dette eksperimentet bli vektlagt i like stor grad som de to andre.



Figur 7.4: (a) Estimert geografisk effekt på forventet antall skader hos MOD-3.0. (b) Estimert geografisk effekt på forventet antall skader hos MOD-3.1

	MOD-3.0	MOD-3.1	MOD-3.2	MOD-3.3	MOD-3.4
Intercept	-2.364***	-2.234***	-2.234***	-2.234***	-2.234***
	(0.1)	(0.099)	(0.099)	(0.099)	(0.099)
Urbaniseringsgrad	0.17***	0.148***	0.148***	0.148***	0.148***
	(0.014)	(0.017)	(0.017)	(0.017)	(0.017)
Kriminalitetsrate	-	-	-	-	-
	-	-	-	-	-
Medium inntekt	-0.094*	-0.078	-0.078	-0.078	-0.078
	(0.055)	(0.056)	(0.056)	(0.056)	(0.056)
Høy inntekt	-0.716***	-0.693***	-0.693***	-0.693***	-0.693***
	(0.078)	(0.079)	(0.079)	(0.079)	(0.079)
Alder: 20-29	0.386***	0.376***	0.376***	0.376***	0.376***
	(0.077)	(0.077)	(0.077)	(0.077)	(0.077)
Alder: 30-39	0.068	0.062	0.062	0.062	0.062
	(0.081)	(0.081)	(0.081)	(0.081)	(0.081)
Alder: 50-69	-0.152**	-0.15*	-0.15*	-0.15*	-0.15*
	(0.077)	(0.077)	(0.077)	(0.077)	(0.077)
Alder: >69	0.13	0.14	0.141	0.141	0.141
	(0.088)	(0.088)	(0.088)	(0.088)	(0.088)
τ_c	-	∞	∞	∞	∞
	-	(∞)	(∞)	(∞)	(∞)
τ_h	-	93.017	93.011	93.016	93.011
	-	(71.37)	(71.363)	(71.367)	(71.363)
δ	-	0.651	0.542	0.09	0.683
	-	(4495.815)	(2976.185)	(261.574)	(4195.956)
Observasjoner	10000	10000	10000	10000	10000
Log likelihood	-4455.487	-4446.489	-4446.489	-4446.489	-4446.489
Akaike Inf.Crit	8918.974	8914.977	8914.977	8914.977	8914.977

Note: *p<0.1; **p<0.05; ***p<0.01

Tabell 7.3: Resultater for modeller i eksperiment 3. Estimeringen av parametere i MOD-3.0 blir utført ved hjelp av den iterative ligningen gitt ved (2.18), mens de resterende modellene blir tilpasset ved hjelp av TMB. Verdier i parentes tilsvarer standardavvik.

7.4 Validering av modeller

Som tidligere nevnt vil det simulert tusen nye porteføljer for å validere og sammenligne den prediktive kraften til de ulike modellene. Modellene som ble tilpasset den opprinnelige porteføljen i kapittel 5 blir brukt til å estimere forventet antall skader hos hver enkelt polise i de nye porteføljene. Som et resultat av at de sanne parameterverdiene er kjent, er det enkelt å måle graden av prediksjonskraft hos de ulike modellene.

Dette blir gjort ved å definere den totale feilen som $\mathrm{TE}^{\mu} = \sum_{i=1}^{M}(\mu_i - \hat{\mu}_i)$, hvor M tilsvarer antall poliser i porteføljen, μ_i er forventet antall skader hos polise i og $\hat{\mu}_i$ er punktestimatet til μ_i. Videre blir den totale feilen dekomponert ved

$$\mathrm{TE}^{\mu} = \sum_{i=1}^{M}(\mu_i - \hat{\mu}_i)I(\mu_i \geq \hat{\mu}_i) + \sum_{i=1}^{M}(\mu_i - \hat{\mu}_i)I(\mu_i < \hat{\mu}_i) = \mathrm{TE}^{\mu}_u + \mathrm{TE}^{\mu}_o. \tag{7.5}$$

Årsaken til at dette blir gjort er for å fange opp eventuelle modeller som konsekvent overestimerer og underestimerer. To modeller kan ha den samme prediktive kraften i form av den totale feilen, men det kan samtidlig være store forskjeller mellom modellene i form av at den ene overestimerer og underestimerer i langt større grad enn den andre.

For portefølje $j = \{1, \ldots, 1000\}$ finner man de tre summene TE^{μ}_j, $\mathrm{TE}^{\mu}_{u,j}$ og $\mathrm{TE}^{\mu}_{o,j}$ for hver enkelt modell i seksjon 7.2.2. For å unngå at resultatene i for stor grad påvirkes av tilfeldig variasjon hos de simulerte porteføljene blir

$$\begin{aligned}
\overline{\mathrm{TE}}^{\mu} &= \frac{1}{N}\sum_{j=1}^{N}\mathrm{TE}^{\mu}_j \qquad \frac{1}{N-1}\sum_{j=1}^{N}\left(\mathrm{TE}^{\mu}_j - \overline{\mathrm{TE}}^{\mu}\right)^2 \\
\overline{\mathrm{TE}}^{\mu}_u &= \frac{1}{N}\sum_{j=1}^{N}\mathrm{TE}^{\mu}_{u,j} \qquad \frac{1}{N-1}\sum_{j=1}^{N}\left(\mathrm{TE}^{\mu}_{u,j} - \overline{\mathrm{TE}}^{\mu}_u\right)^2 \\
\overline{\mathrm{TE}}^{\mu}_o &= \frac{1}{N}\sum_{j=1}^{N}\mathrm{TE}^{\mu}_{o,j} \qquad \frac{1}{N-1}\sum_{j=1}^{N}\left(\mathrm{TE}^{\mu}_{o,j} - \overline{\mathrm{TE}}^{\mu}_o\right)^2
\end{aligned} \tag{7.6}$$

brukt som sammenligningsgrunnlag mellom modellene. Dette er tilsvarende fremgangsmåte som hos Dimakos og Di Rattalma (2002). Resultatene for den sanne modellen gitt ved ligning (5.1) er også inkludert.

7.4.1 Eksperiment 1

Ikke overraskende viser resultatene at den sanne modellen er den modellen med størst prediksjonskraft. I tillegg kommer det tydelig frem at MOD-1.0 er lite passende, som følge av at modellen i liten grad er i stand til å tildele de ulike polisene en korrekt risikoprofil. MOD-1.3 skiller seg også negativt ut, og er den modellen som predikerer det totale antall skader med klart lavest presisjon. Dette er med på å underbygge påstanden om at MOD-1.3 er lite passende basert på diskusjonen vedrørende resultatene i tabell 7.1.

		Modeller:				
	Sann	MOD-1.0	MOD-1.1	MOD-1.2	MOD-1.3	MOD-1.4
$\overline{\text{TE}}_u^{\mu}$	46.761	226.779	106.689	106.604	119.776	110.396
	(0.475)	(4.049)	(1.62)	(1.619)	(1.77)	(1.67)
$\overline{\text{TE}}_o^{\mu}$	-35.441	-216.834	-94.68	-94.583	-91.821	-97.205
	(0.878)	(3.044)	(1.733)	(1.732)	(1.683)	(1.762)
$\overline{\text{TE}}^{\mu}$	11.32	9.945	12.009	12.022	27.955	13.191
	(1.145)	(5.601)	(2.838)	(2.835)	(2.901)	(2.912)

Tabell 7.4: Resultater fra eksperiment 1, hvor urbaniseringsgrad og kriminalitetsrate er fjernet som forklaringsvariabler.

Forskjellene mellom MOD-1.1 og MOD-1.2 er neglisjerbare, og de to modellene ser ut til å være noe bedre enn MOD-1.4. Det vil dermed være naturlig å anbefale en av de mer tradisjonelle nabostrukturene i MOD-1.1 og MOD-1.2 fremfor nabostrukturen i MOD-1.4 for videre analyser. På samme tid kunne det også vært interessant å teste flere modeller med ulike nabostrukturer basert på avstanden mellom de administrative punktene til kommunene.

MOD-1.1 og MOD-1.2 reduserer over- og underestimering med over 50% sammenlignet med MOD-1.0. Dette understreker at modellene i mye større grad klarer å tildele de ulike polisene en riktig risikoprofil. Som tidligere nevnt kommer dette hovedsaklig som et resultat av at risikoen hos polisene i Oslo, som utgjør ca 10% av porteføljen, blir estimert med liten presisjon. Det er også interessant å se at forskjellene mellom $\overline{\text{TE}}^{\mu}$ for de ulike modellene er liten, noe som også var tilfellet hos Dimakos og Di Rattalma (2002). Tolkningen av dette

kan være at modeller med latente variabler ikke er like passende for oppgaver vedrørende reservering.

7.4.2 Eksperiment 2

Tilsvarende som i eksperiment 1 viser resultatene at den sanne modellen er den modellen med størst prediksjonskraft. Resultatene viser også at når man inkluderer forklaringsvariabelen kriminalitetsrate reduseres forskjellen mellom modellene betraktelig. Dette er et klart tegn på at forklaringsvariabelen har en signifikant effekt på responsvariablene, og er i tråd med modellen som ble brukt til å simulere forsikringsporteføljene.

		Modeller:				
	Sann	MOD-2.0	MOD-2.1	MOD-2.2	MOD-2.3	MOD-2.4
$\overline{\text{TE}}_u^{\mu}$	46.761	98.569	79.165	78.811	88.966	84.234
	(0.475)	(1.41)	(1.123)	(1.119)	(1.259)	(1.194)
$\overline{\text{TE}}_o^{\mu}$	-35.441	-87.548	-72.241	-71.82	-75.293	-76.648
	(0.878)	(1.765)	(1.437)	(1.43)	(1.489)	(1.511)
$\overline{\text{TE}}^{\mu}$	11.32	11.021	6.924	6.991	13.673	7.586
	(1.145)	(2.604)	(2.144)	(2.135)	(2.26)	(2.258)

Tabell 7.5: Resultater fra eksperiment 2, hvor urbaniseringsgrad er fjernet som forklaringsvariabel.

Resultatene viser også at valget av nabostruktur har en noe mindre betydning sammenlignet med eksperiment 1. MOD-2.3 skiller seg likevel ut som den dårligste modellen, og er også den modellen som predikerer det totale antall skader med klart lavest presisjon. Tilsvarende som i eksperiment 1 er dette med på å underbygge påstanden om at man bør unngå å definere to kommuner som naboer dersom avstanden mellom deres administrative punkt er mindre enn 50 km.

Forskjellene mellom MOD-2.1 og MOD-2.2 er også her neglisjerbare, og de to modellene ser ut til å være noe bedre enn MOD-2.4. I likhet med tolkningen rundt resultatene i tabell 7.2 anbefales en av de mer tradisjonelle nabostrukturene i MOD-2.1 og MOD-2.2 fremfor nabostrukturen i MOD-2.4.

MOD-2.1 og MOD-2.2 reduserer over- og underestimering med over omlag 20% sammenlignet med MOD-2.0. Dette tyder på at inkluderingen av de latente variablene fremdeles har en klar positiv effekt på sluttresultatene. Forskjellene mellom $\overline{\text{TE}}^{\mu}$ for de ulike modellene er liten, men i motsetning til eksperiment 1 leverer MOD-2.1 og MOD-2.2 bedre resultater enn både den sanne modellen og MOD-2.0.

7.4.3 Eksperiment 3

Som tidligere nevnt er bidraget fra $\boldsymbol{\phi}$ neglisjerbart dersom man utelater forklaringsvariabelen urbaniseringsgrad, og de ulike modellene med latente variabler ble identiske som følge av at fordelingen til $\boldsymbol{\theta}$ ikke påvirkes av valget av nabostruktur. Dette gjenspeiles også i resultatene i tabell 7.6. Modellene med latente variabler reduserer underestimeringen med over omlag

		Modeller:				
	Sann	MOD-3.0	MOD-3.1	MOD-3.2	MOD-3.3	MOD-3.4
$\overline{\text{TE}}_u^{\mu}$	46.761	88.184	22.939	22.938	22.938	22.938
	(0.475)	(1.503)	(0.763)	(0.763)	(0.763)	(0.763)
$\overline{\text{TE}}_o^{\mu}$	-35.441	-76.339	-112.295	-112.29	-112.29	-112.29
	(0.878)	(1.495)	(1.623)	(1.623)	(1.623)	(1.623)
$\overline{\text{TE}}^{\mu}$	11.32	11.845	-89.356	-89.352	-89.352	-89.352
	(1.145)	(2.428)	(1.951)	(1.951)	(1.951)	(1.951)

Tabell 7.6: Resultater fra eksperiment 3, hvor kriminalitetsrate er fjernet som forklaringsvariabel.

70% sammenlignet med MOD-3.0, men på samme tid øker modellene overestimeringen med omlag 50%. Dette tyder på at modellene med latente variabler er lite passende for reserveringsoppgaver sammenlignet med MOD-3.0. Dersom man ser nærmere på resultatene i tabell 7.3 ser man at det er små forskjeller mellom parameterestimatene hos MOD-3.0 og de resterende modellene. I tillegg til dette har man også understreket at bidraget fra $\boldsymbol{\phi}$ er tilnærmet lik null. Dette forteller at det er estimeringen av $\boldsymbol{\theta}$ som resulterer i den store forskjellen mellom modellene. Ved å undersøke de aggregerte geografiske residualene og den estimerte effekten av $\boldsymbol{\theta}$ viser det seg at forventet antall skader hos poliser i kommunen Oslo

(som utgjør omlag 10% av polisene) blir justert opp med omlag 20%. Dette ser ut til å være hovedgrunnen til at over- og underestimeringen henholdsvis økes og reduseres hos modellene med latente variabler.

7.5 Oppsummering

Resultatene viser at inkluderingen av latente variabler med en bestemt avhengighetsstruktur kan øke prediksjonskraften hos en modell. Det ble også påvist at valg av nabostruktur er avgjørende for sluttresultat, og dette er muligens en problemstilling som ikke blir tatt stilling i tilstrekkelig grad. Selv om det i dette tilfellet viste seg at de mer tradisjonelle nabostrukturene var å foretrekke, kan det være andre tilfeller hvor nabostrukturer basert på avstand er det beste valget.

Resultatene i tabell 7.1 var noe overdrevet som følge av at MOD-1.0 ikke hadde noen forklaringsvariabler som fanget noe som helst form for geografisk effekt på responsvariablene. Resultatene fra modellene med latente variabler gir likevel en klar indikasjon på at MOD-1.0 mangler nettopp dette, og dette i seg selv kan være svært nyttig når man ønsker å komme frem til en god modell.

Videre viser resultatene i tabell 7.2 og 7.3 at selv om MOD-2.0 og MOD-3.0 har kommunespesifikke forklaringsvariabler, klarer ikke modellene å fange opp den samlede effekten fra de ulike kommunene. Som følge av dette estimeres det et postitivt bidrag fra henholdsvis $\boldsymbol{\phi}$ og $\boldsymbol{\theta}$ i MOD-2.1 og MOD-3.1. Det er en klar korrelasjon mellom urbaniseringsgrad og kriminalitetsrate, men på samme tid eksisterer det kommuner hvor dette ikke er tilfellet. Enkelte av modellene med latente variabler er i stand til å fange opp den sanne effekten fra det geografiske området på en mer tilfredsstillende måte.

Det er også verdt å merke seg at forskjellen mellom AIC-verdiene hos MOD-2.0 og de resterende modellene er betydelige lavere enn ved eksperiment 1. Dette kommer hovedsaklig som et resultat av at MOD-2.0 fanger opp en sterk effekt for poliser i Oslo. Ved å inkludere en indikatorvariabel i MOD-1.0, som tar verdien 1 dersom polisen tilhører Oslo, og 0 ellers, faller AIC-verdien fra 9082.883 til under 9000.

Kapittel 8

Avslutning

I denne oppgaven har det blitt presentert en alternativ metode for arbeidet knyttet mot geografisk prisdifferensiering hos forsikringsselskaper, hvor tidligere studier stort sett baseres på bayesiansk metodikk. Her har problemstillingen blitt angrepet fra et frekventisk ståsted, hvor fokuset har begrenset seg til aktuelle modeller for antall skader. Estimering av parametere ble utført ved hjelp av maksimum likelihood, og estimeringsprosedyren ble automatisert ved hjelp av pakken Template Model Builder (TMB).

I kapittel 1 ble det gitt en kort introduksjon til hvordan forsikringsselskaper tar i bruk glattingsmetoder som verktøy i arbeidet knyttet til geografisk prisdifferensiering, samt et sammendrag av publisert forskning som har hatt fokus på geografisk lokalitet som risikofaktor innenfor forsikring. I kapittel 2 ble generaliserte lineære modeller (GLM) introdusert, og i kapittel 3 ble teorien om Gaussian Markov random fields (GMRFs) presentert. Kapittel 4 tok for seg ulike modeller for antall skader, med et ekstra fokus på hvorfor det i mange tilfeller kan være gunstig å introdusere korrelerte latente variabler, samt problemstillingen vedrørende overdispersjon. Videre ble simuleringen av en forsikringsporteføljе presentert i kapittel 5, hvor porteføljen ble begrenset til å innholde informasjon om antall skader innefor et tidsrom på ett år. Selve estimeringsprosedyren av de valgte modellene ble tatt for seg i kapittel 6, og resultatene for modellene ble sammenfattet i kapittel 7.

Resultatene viste at korrelerte latente variabler kan ha en positiv effekt dersom det eksisterer en underliggende faktor, med en relativt glatt geografisk fordeling, som de tilgjengelige forklaringsvariablene ikke er i stand til å beskrive. Resultatene indikerte også at valget av nabostruktur vil være avgjørende for det endelige resultatet. I denne oppgaven var de mer tradisjonelle nabostrukturene å foretrekke, men det er likevel rimelig å anta at andre nabostrukturer, eksempelvis avstandsbaserte, også kan avdekke interessante sammenhenger.

I de neste seksjonene presenteres potensielle feilkilder og begrensninger ved oppgaven, samt mulighetene for videre studier vedrørende problemstillingen i oppgaven.

Potensielle feilkilder

Simuleringen av forsikringsporteføljen i kapittel 5 baserte seg på en rekke data innhentet fra SSB sine hjemmesider. Dataene var ikke bearbeidet i tilfredsstillende grad etter mine behov, og det var noe utfordrende at SSB og Kartverket tok i bruk ulike identifikasjoner for de ulike kommunene i Norge. Dette medførte en del manuell koding som åpenbart kan være potensielle feilkilder, og som et resultat av dette ble også selve arbeidet med dataene noe mer tidkrevende enn planlagt. Det har likevel vært en lærerik prosess å arbeide med større datamengder enn hva som har vært vanlig tidligere i studieløpet, og underveis blitt tryggere på hvordan man kan sikre seg at dataene blir strukturert på en ønskelig måte.

Begrensninger

Det har gjennom hele oppgaven vært et ønske å knytte teorien og de valgte modellene opp mot forsikringsrelaterte problemstillinger. Den simulerte forsikringsporteføljen i kapittel 5 har åpenbart klare svakheter, og det er rimelig å stille spørsmålstegn om hvorvidt simuleringen er representativ for en portefølje for et gitt forsikringsprodukt hos et forsikringsselskap i Norge. Her er det viktig å presisere at formålet til oppgaven ikke er å komme frem til en endelig prisingsmodell for et forsikringsprodukt, men å introdusere et alternativt syn på problemstillingen rundt geografisk prisdifferensiering , som det i liten grad har blitt forsket på. Simuleringen av forsikringsporteføljen bør i dette tilfellet heller bli sett på som et slags verktøy i forbindelse med modellene og ikke minst anvendelse av pakken TMB. Det gjenstår å undersøke om dette kan faktisk kan være et nyttig bidrag inn mot forsikringsselskapers arbeid knyttet til geografisk prisdifferensiering.

Det er også verdt å nevne usikkerheten rundt estimatene til de faste parameterne og ikke minst de latente variablene. Usikkerheten hos de faste parameterne er åpenbart enklere å ta stilling til, og trenger for så vidt ikke noe nærmere forklaring, da kunnskapen rundt dette er velkjent for personer som til daglig arbeider med slike problemstillinger hos forsikringsselskaper. Usikkerheten til estimatene hos de latente variablene er derimot en annen sak, og bør vektlegges i langt større grad enn det som har vært tilfellet i denne oppgaven. Ellers er det verdt å merke seg at i tilfeller hvor denne type modeller kan være aktuelle vil det typisk være stor tilgang til data, noe som kan resultere i mindre usikkerhet hos estimatene til de

latente variablene. Det kan også tenkes at tilgangen til data er såpass stor at modellene som er presentert i oppgaven ikke bidrar med noe ytterligere forklaringskraft sammenlignet med de mer tradisjonelle modellene.

Videre studier

Jeg ser ellers på mulighetene for videre studier som veldig stor. Det aller første jeg ville gjort var å få tilgang til et reelt datasett fra et forsikringsselskap i Norge, og utarbeidet en god modell (GLM) for antall skader. Denne modellen ville blitt brukt som et utgangspunkt for videre analyser for modeller hvor latente variabler ble introdusert. Her ville valgmulighetene vært mange, både når det gjelder aggregeringen av de geografiske regionene og hvordan man definerer to soner som naboer. Basert på analysene i denne oppgaven vil det vært naturlig å ta utgangspunkt i de mer tradisjonelle definisjonene for naboer, og heller introdusere avstandsbaserte naboer dersom det viser seg at det estimerte bidraget fra førstnevnte er neglisjerbart. Ellers vil det også være interessant å utarbeide modeller for gjennomsnittlig skadepris.

Videre vil det også være naturlig å introdusere modeller som ikke bare tar stilling til den romlige dimensjonen, men også introduserer tidsaspektet. Dette vil ikke medføre alt for store endringer i selve programmeringsfasen. Her vil det også være naturlig å sammenligne modellene med de tradisjonelle modellene ved hjelp av out-of-sample tester. I tillegg til dette vil det være rimelig å diskutere potensielle utvidelser av modellene som kan fange opp eventuelle korrelasjoner mellom de ulike poliseholderne fra år til år. Jeg har fått inntrykk av at det ikke er vanlig å ta hensyn til dette i modeller som blir anvendt i forsikringsbransjen.

Ellers har jeg også brukt noe tid på å utarbeide noen forslag for modellering av δ, hovedsaklig som en funksjon av x- og y-koordinater. Resultatene var noe vanskelig å fortolke, og det har derfor blitt utelatt fra selve oppgaven. Jeg ser likevel på dette som en interessant idé å bringe videre.

Referanser

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

S. Banerjee, B. P. Carlin og A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC Press, 2014.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, side 192–236, 1974.

J. Besag, J. York og A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

R. S. Bivand, E. J. Pebesma og V. Gómez-Rubio. *Applied spatial data analysis with R*, volum 747248717. Springer, 2008.

M. Boskov og R. Verrall. Premium rating by geographic area using spatial models. *Astin Bulletin*, 24(01):131–143, 1994.

G. Casella og R. L. Berger. *Statistical inference*, volum 2. Duxbury Pacific Grove, CA, 2002.

P. De Jong og G. Z. Heller. *Generalized linear models for insurance data*, volum 136. Cambridge University Press Cambridge, 2008.

X. K. Dimakos og A. F. Di Rattalma. Bayesian premium rating with latent structure. *Scandinavian Actuarial Journal*, 2002(3):162–184, 2002.

A. J. Dobson og A. Barnett. *An introduction to generalized linear models*. CRC press, 2011.

D. A. Fournier, H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen og J. Sibert. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*,

27(2):233–249, 2012. doi: 10.1080/10556788.2011.597854. URL `http://dx.doi.org/10.1080/10556788.2011.597854`.

T. Hastie, R. Tibshirani og J. Friedman. *The elements of statistical learning*, volum 2. Springer, 2009.

K. Kristensen. *TMB: General random effect model builder tool inspired by ADMB.*, 2014. R package version 1.1.

D. Lee. CARBayes: an R package for bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24, 2013.

B. G. Leroux, X. Lei og N. Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. I *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, side 179–191. Springer, 2000.

L. Márkus, N. M. Arató og V. Prokaj. Hierarchical bayesian modelling of geographic dependence of risk in household insurance. I *Advances in Data Analysis*, side 219–227. Springer, 2010.

J. A. Nelder og R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384, 1972.

U. Olsson. Generalized linear models: an applied approach. *Studentlitteratur, Lund*, 18, 2002.

H. Rue og L. Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.

I. Scheel, E. Ferkingstad, A. Frigessi, O. Haug, M. Hinnerichsen og E. Meze-Hausken. A bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):85–100, 2013.

H. J. Skaug og D. A. Fournier. Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. *Computational Statistics & Data Analysis*, 51(2):699–709, 2006.

H. Stern og N. A. Cressie. Inference for extremes in disease mapping. I *Disease mapping and risk assessment for public health*. John Wiley & Sons, 1999.

G. C. Taylor. Use of spline functions for premium rating by geographic area. *Astin Bulletin*, 19(01):91–122, 1989.

I. Thorsen. Modellering av romlig variasjon i frekvenser av vannskader på boliger. Master's thesis, University of Bergen, 2012.

Vedlegg A: Estimering av parametere

A.1 Vektet minste kvadraters metode

Lar $Y_1, \ldots, Y_N$ være en sekvens med uavhengige variabler med tilhørende forventning gitt ved $\mu_1, \ldots, \mu_N$. Videre antar man at forventningen er en funksjon av parametervektoren $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$, hvor $p < N$. Målet er å estimere de ukjente parameterne $\beta_1, \ldots, \beta_p$. I denne seksjonen presenteres estimeringsprosedyren kalt vektet minste kvadraters metode på tilsvarende måte som hos Dobson og Barnett (2011).

Basert på antagelsene har man at

$$E(Y_i) = \mu_i(\boldsymbol{\beta}). \tag{A.1}$$

Den enkleste fremgangsmåten er å finne estimatoren $\hat{\boldsymbol{\beta}}$ som minimerer summen gitt ved

$$S = \sum_{i=1}^{N} \left(Y_i - \mu_i(\boldsymbol{\beta})\right)^2. \tag{A.2}$$

Dette blir gjort ved å derivere S med hensyn på hvert enkelt element i $\boldsymbol{\beta}$, og deretter løse ligningene gitt ved

$$\frac{\partial S}{\partial \beta_j} = 0, \qquad j = 1, \ldots, p. \tag{A.3}$$

Videre er det viktig å undersøke om $\hat{\boldsymbol{\beta}}$ er et globalt minimum og dermed en korrekt løsning på problemet. Dette er tilfellet dersom Hessematrisen er positiv-definit, og eventuelle lokale minimum, eksempelvis langs parameterrommet, er tatt hensyn til.

Dersom det er rimelig å anta at variablene har ulik varians er det vanlig å heller minimere summen

$$S = \sum_{i=1}^{N} w_i \left[Y_i - \mu_i(\boldsymbol{\beta})\right]^2, \tag{A.4}$$

hvor $w_i = 1/\mathrm{Var}(Y_i)$. Dette vil medføre at variablene som er mindre troverdige (eksempelvis Y_i-ene med høy varians) blir lagt mindre vekt på, og dermed påvirker estimatene i en mindre grad sammenlignet med tilfellet hvor w_i-ene er lik 1.

Mer generelt kan det tenkes at man har en vektor $\mathbf{y} = [y_1, \ldots, y_N]^T$ med tilhørende forventningsvektor $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^T$ og kovariansmatrise $\mathbf{V}$. Ved hjelp av vektet minste kvadraters metode vil parameterestimatene tilsvare verdiene som minimerer

$$S = (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}). \tag{A.5}$$

Legg merke til at dersom variablene er uavhengige vil metoden være identisk med minimeringsproblemet gitt ved (A.4). Dersom variablene også antas å ha identisk varians vil metoden tilsvare minimeringsproblemet gitt ved (A.2).

A.2 Maksimum likelihood estimering

Antar at man har en sekvens med uavhengige variabler $Y_1, \ldots, Y_N$ som tilfredsstiller betingelsene til en generalisert lineær modell i delkapittel 2.2. Målet er å estimere de ukjente parameterne $\beta_1, \ldots, \beta_p$, som er relatert til Y_i-ene gjennom $\mathrm{E}(Y_i) = \mu_i$ og $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. I denne seksjonen fokuseres det på en mye anvendt metode, nemlig maksimum likelihood. Fremgangsmåten under er en kombinasjon av estimeringsprosedyrene presentert hos Dobson og Barnett (2011) og Olsson (2002).

For hver enkelt Y_i er log-likelihood-funksjonen gitt ved

$$l_i = \log\left[L(\theta_i, \phi; y_i)\right] = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi), \tag{A.6}$$

hvor $a(\cdot)$, $b(\cdot)$ og $c(\cdot)$ er kjente funksjoner. I tillegg har man at $\mathrm{E}(Y_i) = b'(\theta_i)$, $\mathrm{Var}(Y_i) = a(\phi)b''(\theta_i)$ og $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, fra henholdsvis (2.15), (2.16) og (2.17). Som et resultat av antagelsen om uavhengige responsvariabler er log-likehood-funksjonen, l, for alle responsvariablene gitt ved følgende uttrykk:

$$l = \sum_{i=1}^{N} l_i = \sum_{i=1}^{N} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]. \tag{A.7}$$

Det neste steget innebærer å derivere log-likelihood-funksjonen med hensyn på hver enkelt element i $\boldsymbol{\beta}$, slik at

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \left[\frac{\partial l_i}{\partial \beta_j}\right] = \sum_{i=1}^{N} \left[\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j}\right]. \tag{A.8}$$

Som et resultat av (A.6) er følgende tilfredsstilt:

$$\frac{\partial l_i}{\partial \theta_i} = \big(y_i - b'(\theta_i)\big)/a(\phi) = (y_i - \mu_i)/a(\phi). \tag{A.9}$$

Videre kan $\frac{\partial \theta_i}{\partial \mu_i}$ uttrykkes ved

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\left(\frac{\partial \mu_i}{\partial \theta_i}\right)} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{\text{Var}(Y_i)}, \tag{A.10}$$

som et resultat av (2.15) og (2.16). For å finne den siste faktoren i (A.8) tar man i bruk kjerneregelen ytterligere en gang, slik at:

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}. \tag{A.11}$$

U_j, ofte kalt scorefunksjonen, kan følgelig uttrykkes ved:

$$U_j = \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)}{\text{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \tag{A.12}$$

Kovariansmatrisen til U_j-ene, $\mathfrak{F}$, er en $p \times p$ matrise kalt informasjonsmatrisen og er per definisjon gitt ved:

$$\Im_{jk} = E[U_j U_k] = E\left\{ \sum_{i=1}^{N} \left[\frac{(Y_i - \mu_i)}{\text{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{l=1}^{N} \left[\frac{(Y_l - \mu_l)}{\text{Var}(Y_l)} x_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right\}. \tag{A.13}$$

Som et resultat av at Y_i-ene er uavhengige vil $E\left[(Y_i - \mu_i)(Y_l - \mu_l)\right] = 0$ for $i \neq l$. I tillegg er $E\left[(Y_i - \mu_i)^2\right] = \text{Var}(Y_i)$ og (A.13) forenkles til

$$\Im_{jk} = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \tag{A.14}$$

Dobson og Barnett (2011) utleder en estimeringsprosedyre kalt «method of scoring» for et gitt tilfelle hvor bare en parameter skal estimeres. En generalisering er gitt ved:

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + \left[\mathfrak{F}^{(m-1)} \right]^{-1} \mathbf{U}^{(m-1)}, \tag{A.15}$$

hvor $\mathbf{b}^{(m)}$ er de m-te estimatene for $\boldsymbol{\beta}$, $[\mathfrak{F}^{(m-1)}]^{-1}$ er den inverse matrisen til $\mathfrak{F}$ gitt ved (A.14) og $\mathbf{U}^{(m-1)}$ er en $p \times 1$ vektor definert ved (A.12). $\mathfrak{F}^{(m-1)}$ og $\mathbf{U}^{(m-1)}$ er ikke kjent og er derfor evaluert ved $\mathbf{b}^{(m-1)}$. Videre kan man multiplisere begge sider med $\mathfrak{F}^{(m-1)}$ slik at

$$\mathfrak{F}^{(m-1)} \mathbf{b}^{(m)} = \mathfrak{F}^{(m-1)} \mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}, \tag{A.16}$$

og deretter observere at $\mathfrak{F}$ kan uttrykkes som

$$\mathfrak{F} = \mathbf{X}^T \mathbf{W} \mathbf{X}, \tag{A.17}$$

hvor $\mathbf{W}$ er en $N \times N$ diagonalmatrise med elementer lik

$$w_{ii} = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right). \tag{A.18}$$

Uttrykket på høyre side i (A.16) kan uttrykkes ved

$$\mathbf{X}^T\mathbf{W}\mathbf{z}, \tag{A.19}$$

hvor $\mathbf{z}$ er en $N \times 1$ vektor. Elementene i $\mathbf{z}$ er gitt ved

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right), \tag{A.20}$$

hvor μ_i og $\frac{\partial \eta_i}{\partial \mu_i}$ er evaluert ved $\mathbf{b}^{(m-1)}$. Den iterative ligningen kan følgelig uttrykkes ved

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}. \tag{A.21}$$

Estimatene for parameterne blir dermed funnet numerisk ved å velge initialverdier lik $\mathbf{b}^{(0)}$, evaluere $\mathbf{z}$ og $\mathbf{W}$, og dersom estimatene konvergerer, avslutte den iterative ligningen når forskjellen mellom $\mathbf{b}^{(m)}$ og $\mathbf{b}^{(m-1)}$ er tilstrekkelig liten.

A.3 Kommentar

En viktig forskjell mellom vektet minste kvadraters metode og maksimum likelihood er at man ved førstnevnte ikke trenger å gjøre noen antagelser angående fordelingen til de stokastiske variablene, sett bort i fra forventning, varians og eventuelt en kovariansmatrise. I mange tilfeller vil likevel parameterestimatene være identiske for de to metodene, og det kan også nevnes at det svært ofte er nødvendig med numeriske metoder for å kunne løse problemstillingene i de ulike metodene.

Vedlegg B: Simulering fra ulike GMRFs

I delkapittel 3.4 ble det simulert tilfeldige utvalg fra GMRFs med ulike verdier for δ. Rent empirisk ga resultatene en indikasjon på at δ påvirket korrelasjonen mellom variablene. I dette vedlegget er målet å illustrere hvilken effekt δ har på fordelingen fra et mer teoretisk ståsted.

For enkelhetsskyld begrenses det geografiske området til Sogn og Fjordane, som består av 26 kommuner. I likhet med simuleringen i delkapittel 3.4 antas det at to kommuner defineres som naboer hvis og bare hvis grensene til de to kommunene møtes i minst ett punkt. Tilhørende graf, gitt ved $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, er illustrert ved figur 3.1b s.20.

Den multivariate fordelingen er gitt ved

$$\boldsymbol{\phi} = [\phi_1, \ldots, \phi_{26}]^T \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{-1}/\tau\right), \tag{B.1}$$

hvor $Q_{ij} = -1$ hvis og bare hvis kommune i og j er naboer, ellers lik 0. Diagonalelementene til $\mathbf{Q}$ er gitt ved $Q_{ii} = n_i + \delta$, hvor $\delta > 0$ og n_i er antall naboer til kommune i. For enkelhetsskyld er τ lik 1.

Matrisen som det blir tatt utgangspunkt i, før δ legges til hos diagonalelementene, er gitt ved

$$\mathbf{Q}^* = \begin{pmatrix} 6 & . & -1 & . & . & \ldots \\ . & 6 & . & -1 & . & \ldots \\ -1 & . & 4 & . & . & \ldots \\ . & -1 & . & 3 & . & \ldots \\ . & . & . & . & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{B.2}$$

hvor $Q_{ij} = .$ tilsvarer $Q_{ij} = 0$.

De presenteres resultater for tre ulike fordelinger, hvor $\delta = \{10, 1, 0.01\}$. For hver av de tre fordelingene simuleres det tusen utvalg som blir brukt for å lage empiriske korrelasjonsmatriser.

Dette blir hovedsaklig gjort for å bekrefte at simuleringene har blitt utført på en korrekt måte. I tillegg til dette presenteres $\mathbf{Q}$, $\mathbf{Q}^{-1}$ og $\text{corr}(\boldsymbol{\phi})$, hvor elementene i de ulike matrisene er avrundet til to desimaler.

Resultater for $\delta = 10$

$$\mathbf{Q} = \begin{pmatrix} 16 & . & -1 & . & . & \dots \\ . & 16 & . & -1 & . & \dots \\ -1 & . & 14 & . & . & \dots \\ . & -1 & . & 13 & . & \dots \\ . & . & . & . & 11 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \qquad \mathbf{Q}^{-1} = \begin{pmatrix} 0.06 & 0.00 & 0.01 & 0.00 & 0.00 & \dots \\ 0.00 & 0.06 & 0.00 & 0.01 & 0.00 & \dots \\ 0.01 & 0.00 & 0.07 & 0.00 & 0.00 & \dots \\ 0.00 & 0.01 & 0.00 & 0.08 & 0.00 & \dots \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.09 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$\text{corr}(\boldsymbol{\phi}) = \begin{pmatrix} 1.00 & 0.01 & 0.07 & 0.00 & 0.00 & \dots \\ 0.01 & 1.00 & 0.00 & 0.08 & 0.00 & \dots \\ 0.07 & 0.00 & 1.00 & 0.00 & 0.00 & \dots \\ 0.00 & 0.08 & 0.00 & 1.00 & 0.00 & \dots \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \hat{\text{corr}}(\boldsymbol{\phi}) = \begin{pmatrix} 1.00 & -0.01 & 0.00 & -0.08 & 0.02 & \dots \\ -0.01 & 1.00 & -0.02 & 0.06 & 0.05 & \dots \\ 0.00 & -0.02 & 1.00 & -0.05 & 0.04 & \dots \\ -0.08 & 0.06 & -0.05 & 1.00 & 0.05 & \dots \\ 0.02 & 0.05 & 0.04 & 0.05 & 1.00 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Resultater for $\delta = 1$

$$\mathbf{Q} = \begin{pmatrix} 7 & . & -1 & . & . & \dots \\ . & 7 & . & -1 & . & \dots \\ -1 & . & 5 & . & . & \dots \\ . & -1 & . & 4 & . & \dots \\ . & . & . & . & 2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \mathbf{Q}^{-1} = \begin{pmatrix} 0.20 & 0.03 & 0.07 & 0.02 & 0.01 & \dots \\ 0.03 & 0.20 & 0.03 & 0.08 & 0.00 & \dots \\ 0.07 & 0.03 & 0.28 & 0.01 & 0.00 & \dots \\ 0.02 & 0.08 & 0.01 & 0.31 & 0.00 & \dots \\ 0.01 & 0.00 & 0.00 & 0.00 & 0.59 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$\text{corr}(\boldsymbol{\phi}) = \begin{pmatrix} 1.00 & 0.16 & 0.30 & 0.07 & 0.02 & \dots \\ 0.16 & 1.00 & 0.11 & 0.33 & 0.01 & \dots \\ 0.30 & 0.11 & 1.00 & 0.05 & 0.01 & \dots \\ 0.07 & 0.33 & 0.05 & 1.00 & 0.00 & \dots \\ 0.02 & 0.01 & 0.01 & 0.00 & 1.00 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \hat{\text{corr}}(\boldsymbol{\phi}) = \begin{pmatrix} 1.00 & 0.18 & 0.32 & 0.07 & 0.01 & \dots \\ 0.18 & 1.00 & 0.10 & 0.34 & 0.01 & \dots \\ 0.32 & 0.10 & 1.00 & 0.09 & -0.03 & \dots \\ 0.07 & 0.34 & 0.09 & 1.00 & 0.03 & \dots \\ 0.01 & 0.01 & -0.03 & 0.03 & 1.00 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Resultater for $\delta = 0.01$

$$\mathbf{Q} = \begin{pmatrix} 6.01 & . & -1.00 & . & . & \dots \\ . & 6.01 & . & -1.00 & . & \dots \\ -1.00 & . & 4.01 & . & . & \dots \\ . & -1.00 & . & 3.01 & . & \dots \\ . & . & . & . & 1.01 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \mathbf{Q}^{-1} = \begin{pmatrix} 4.07 & 3.85 & 3.94 & 3.82 & 3.72 & \dots \\ 3.85 & 4.08 & 3.86 & 3.96 & 3.67 & \dots \\ 3.94 & 3.86 & 4.27 & 3.82 & 3.66 & \dots \\ 3.82 & 3.96 & 3.82 & 4.28 & 3.66 & \dots \\ 3.72 & 3.67 & 3.66 & 3.66 & 5.45 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$$\text{corr}(\boldsymbol{\phi}) = \begin{pmatrix} 1.00 & 0.95 & 0.95 & 0.92 & 0.79 & \dots \\ 0.95 & 1.00 & 0.92 & 0.95 & 0.78 & \dots \\ 0.95 & 0.92 & 1.00 & 0.89 & 0.76 & \dots \\ 0.92 & 0.95 & 0.89 & 1.00 & 0.76 & \dots \\ 0.79 & 0.78 & 0.76 & 0.76 & 1.00 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad \hat{\text{corr}}(\boldsymbol{\phi}) = \begin{pmatrix} 1.00 & 0.94 & 0.95 & 0.91 & 0.77 & \dots \\ 0.94 & 1.00 & 0.92 & 0.95 & 0.77 & \dots \\ 0.95 & 0.92 & 1.00 & 0.89 & 0.75 & \dots \\ 0.91 & 0.95 & 0.89 & 1.00 & 0.75 & \dots \\ 0.77 & 0.77 & 0.75 & 0.75 & 1.00 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Kommentar

De ulike matrisene bekrefter antagelsene fra simuleringen i delkapittel 3.4. Dersom δ synker vil det resultere i en sterkere marginal korrelasjon mellom variablene, samtidlig som den marginale variansen øker.

En annen egenskap som er verdt å merke seg er hvilken effekt presisjonsparameteren τ har på fordelingen. Dersom $\tau > 1$ vil dette medføre at den marginale variansen hos variablene, samt tilhørende kovarianser, reduseres som et resultat av at kovariansmatrisen er gitt ved $\mathbf{Q}^{-1}/\tau$. Parameteren vil forøvrig ikke påvirke den marginale korrelasjonen mellom variablene som følge av at elementene i korrelasjonsmatrisen, for $\tau > 0$, kan uttrykkes ved

$$\text{corr}(\boldsymbol{\phi})_{ij} = \frac{\frac{\sum_{ij}}{\tau}}{\sqrt{\frac{\sum_{ii}}{\tau}}\sqrt{\frac{\sum_{jj}}{\tau}}} = \frac{\sum_{ij}}{\sqrt{\sum_{ii}\sum_{jj}}}, \tag{B.3}$$

hvor $\sum = \mathbf{Q}^{-1}$.

Vedlegg C: R-koder og eksempel fra TMB

C.1 R-koder: Simulering av forsikringsporteføljе

I denne seksjonen presenteres fremgangsmåten og tilhørende R-koder for simuleringen av forsikringsporteføljen i kapittel 5. Selve forarbeidet som ble gjort i forbindelse med den eksterne informasjonen, hentet fra SSB sine hjemmesider, er ikke presentert.

Laster først inn all nødvendig informasjon om kommunene i Norge, gitt ved tabellene «info_kommune_korr» og «alders_grupper_andel».

```
load("informasjon_kommuner.RData")
load("alders_grupper_andel.RData")
```

Merk at andelen av de ulike aldersgruppene innenfor en gitt kommune tilsvarer den respektive fordelingen i fylket kommunen tilhører. Tabellene er som følger:

knr	kom_navn	antall_pers	areal_km	krim_rate	inntekt_med	fylke	urb_grad	log_urb_grad
1	Bjugn	4,584	355.830	19.900	402,000	Sør-Trøndelag	12.883	2.556
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
429	Meløy	6,657	798.450	25.400	446,000	Nordland	8.337	2.121

Nr.	Kommune	20-29	30-39	40-49	50-59	60-69	70-79	80-89	>=90
1	Bjugn	0.198	0.181	0.189	0.162	0.139	0.076	0.047	0.009
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
429	Meløy	0.159	0.149	0.191	0.180	0.159	0.094	0.056	0.012

Definerer hvor mange poliser som skal simuleres.

```
antall_poliser<-10000
```

Definerer deretter en rekke faste variabler. Dette er variabler som ikke avhenger av antall poliser som blir simulert, og det vil være tilstrekkelig å gjøre de ulike beregningene en enkelt gang.

```
befolk_and<-(info_kommune_korr$antall_pers)/
  sum(info_kommune_korr$antall_pers)
inntekt<-info_kommune_korr$inntekt_med
max_inntekt<-max(inntekt)
min_inntekt<-min(inntekt)
gamma_shape<-(inntekt/(max_inntekt-min_inntekt))^2
gamma_scale<-((max_inntekt-min_inntekt)^2)/inntekt
alfa<-c(-2.3,0.1,0.005,0,-0.1,-0.7,0.3,0.1,0,-0.1,0.2)
alders_gruppe<-c("20-29", "30-39", "40-49", "50-59",
                 "60-69", "70-79", "80-89", ">=90")
```

Distribuerer polisene til de ulike kommunene ved random sampling beskrevet i kapittel 5. For enkelhetsskyld har hver enkelt kommune blitt tildelt et unikt nummer fra 1 til 429. Vanligvis ville det vært naturlig å bruke de reelle kommunenumrene, men kartdataene fra Kartverket hadde identifisert hver enkelt kommune med numrene fra 1 til 429, og derfor ble det naturlig å ta utgangspunkt i nettopp dette.

```
x_kom<-sample(x=info_kommune_korr$knr,
              size=antall_poliser,
              replace=TRUE,
              prob=befolk_and)
```

I det neste steget blir de ulike polisene tildelt en aldersgruppe. Punktsannsynlighetene til fordelingen som alder blir trukket fra blir bestemt på grunnlag av hvilken kommune polisen tilhører. All nødvendig informasjon er gitt i tabellen «alders_gruppe_andel».

```
alder<-character(length=antall_poliser)
for(i in 1:antall_poliser){
  alder[i]<-sample(x=alders_gruppe,
                   size=1,
                   prob=alders_grupper_andel[x_kom[i],])
}
```

Antar at omlag 70% av polisene er gjeldende i hele perioden (365 dager). Hver enkelt polise blir tildelt verdien 1 dersom den er gjeldende i hele perioden, og 0 ellers. Videre blir varigheten

til polisene som ikke er gjeldende hele perioden trukket fra en uniformfordeling med utfallsrom lik $\{1, 2, \ldots, 365\}$. Helt til slutt blir varigheten til polisene transformert om til år.

```
t_polise_dag<-rep(0,antall_poliser)
helaars_polise<-sample(x=c(0,1),
                       size=antall_poliser,
                       replace=TRUE,
                       prob=c(0.3, 0.7))
for(i in 1:antall_poliser){
  if(helaars_polise[i]==1){
    t_polise_dag[i]=365
  }
  else t_polise_dag[i]<-sample(x=seq(1:365),replace=TRUE, 1)
}
E<-t_polise_dag/365
```

Antar at inntekten er gammafordelt, hvor forventningen avhenger av hvilken kommune polisen tilhører. Forventningen tilsvarer medianinntekten i tilhørende kommune, og standardavviket tilsvarer differansen mellom største og minste medianinntekt.

```
inntekt_polise<-rgamma(antall_poliser,
                       shape=gamma_shape[x_kom],
                       scale =gamma_scale[x_kom])
inntekt_kvant<-quantile(inntekt_polise, prob=c(0.25,0.5,0.75))
```

Definerer tre indikatorvariabler som blir bestemt på grunnlag av inntekten til poliseholderne og kvartilene til den simulerte inntektsfordelingen.

```
x3<-as.integer(inntekt_polise < inntekt_kvant[1])
x4<-as.integer(inntekt_kvant[1]<=inntekt_polise &
                 inntekt_polise < inntekt_kvant[3])
x5<-as.integer(inntekt_polise >=inntekt_kvant[3])
```

Definerer fem indikatorvariabler knyttet opp mot forklaringsvariabelen alder.

```
x6<-as.integer(alder==alders_gruppe[1])
x7<-as.integer(alder==alders_gruppe[2])
x8<-as.integer(alder==alders_gruppe[3])
x9<-as.integer(alder %in% alders_gruppe[4:5])
x10<-as.integer(alder %in% alders_gruppe[6:8])
```

Hver enkelt polise blir tildelt en kriminalitetsrate og urbaniseringsgrad bestemt ut i fra hvilken kommune polisen tilhører.

```
urb_grad_polise<-c(info_kommune_korr$urb_grad[x_kom])
x1<-log(urb_grad_polise)
x2<-c(info_kommune_korr$krim_rate[x_kom])
```

Samler all informasjon hos de ulike polisene. Dette kan bli sett på som en designmatrise beskrevet i kapittel 2.

```
info_polise<-cbind(as.integer(1), x1, x2, x3, x4,
                   x5, x6, x7, x8, x9, x10)
```

Finner forventet antall skader hos hver enkelt polise. Her er det verdt å merke seg at varigheten hos polisene blir inkludert som offset i modellen, og påvirker dermed forventet antall skader.

```
forventning<-E*exp(info_polise %*% alfa)
```

Antar at antall skader hos polisene er poissonfordelt.

```
counts<-rpois(n=antall_poliser, lambda=forventning)
```

C.2 Eksempel fra TMB

I denne seksjonen presenteres programmeringen i forbindelse med inferens og estimering hos modeller med latente variabler. Som tidligere nevnt er dette bli gjort ved hjelp av pakken Template Model Builder (TMB), utviklet av Kristensen (2014).

Det første som blir gjort er å lage en cpp-fil (C++ template file).

```
//Fil som skriver cpp-filer
#include <TMB.hpp>

template<class Type>
Type objective_function<Type>::operator() ()
{
  DATA_INTEGER(n); DATA_INTEGER(N); DATA_SPARSE_MATRIX(Q);
  DATA_VECTOR(counts); DATA_IVECTOR(x_kom); DATA_VECTOR(x0);
  DATA_IVECTOR(x2); DATA_IVECTOR(x4); DATA_IVECTOR(x5);
  DATA_IVECTOR(x6);DATA_IVECTOR(x7); DATA_IVECTOR(x9);
```

```
  DATA_IVECTOR(x10); DATA_VECTOR(E);

  PARAMETER(b_0); PARAMETER(b_2); PARAMETER(b_4);
  PARAMETER(b_5); PARAMETER(b_6); PARAMETER(b_7);
  PARAMETER(b_9); PARAMETER(b_10);
  PARAMETER(sigma); PARAMETER(tau); PARAMETER(delta);
  PARAMETER_VECTOR(phi); PARAMETER_VECTOR(theta);

  Type f=0.0;

  Type tauh = Type(1.0)/(tau*tau);
  ADREPORT(tauh);

  Type tauc = Type(1.0)/(sigma*sigma);
  ADREPORT(tauc);

  using namespace density;
  using namespace Eigen;

  SparseMatrix<Type> QQ = Q;

for(int i=0;i<n;i++)
    QQ.coeffRef(i,i) += delta;
  f += GMRF(QQ)(phi);

  // Random effects distribution on theta
  f -= dnorm(theta,Type(0.0),Type(1.0),true).sum();

  // Contribution from likelihood
  for(int j=0;j<N;j++){
    Type eta = b_0*x0(j)+b_2*x2(j)+b_4*x4(j)+
               b_5*x5(j)+b_6*x6(j)+b_7*x7(j)+
               b_9*x9(j)+b_10*x10(j)+
               tau*theta(x_kom(j)-1)+
               sigma*phi(x_kom(j)-1);
    f -= dpois(counts(j),E(j)*exp(eta),true);
  }
  return f;

}
```

Laster inn dataene som blir benyttet i tilpasningen av modellen. Dataene tilsvarer all informasjon om den simulerte forsikringsporteføljen i kapittel 5.

```
load("analyse_grunnlag.RData")
```

Laster inn TMB.

```
require(TMB)
```

Kompilerer cpp-filen.

```
compile("begr_mod1.cpp")
```

I denne delen lastes modellobjektet fra cpp-filen inn.

```
dyn.load(dynlib("begr_mod1"))
```

Definerer en rekke variabler som er en del av modellen. Her tilsvarer n antall kommuner, Q er presisjonsmatrisen og N er antall poliser. De resterende variablene, med unntak av E, er respons- og forklaringsvariabler. Vektoren E blir brukt som offset, og tilsvarer hvor lenge polisene er gjeldende i den aktuelle perioden.

```
data_grunnlag<-list(n=n,Q=Q,N=N, counts=counts,x_kom=x_kom,
                    x0=x0, x2=x2, x4=x4, x5=x5, x6=x6, x7=x7,
                    x9=x9, x10=x10,E=E)
```

Definerer de aktuelle parameterne med tilhørende startverdi for algoritmen.

```
parameter_startverdi<-list(b0=0.1, b2=0.1, b4=0.1,b5=0.1,
                           b6=0.1, b7=0.1, b9=0.1, b10=0.1,
                           sigma=1, tau=1, delta=.001,
                           phi=numeric(n),theta=numeric(n))
```

Spesifiserer hvilke av parameterne som er randomiserte.

```
parameter_random<-c("phi","theta")
```

I dette steget blir det laget et objekt, med tilhørende data og parametere, som blir sendt til TMB gjennom kommandoen «MakeADFun».

```
obj_begr_mod1<- MakeADFun(data=data_grunnlag,
                          parameters=parameter_startverdi,
                          random=parameter_random)
```

Minimerer objektfunksjonen som ble definert i steget over. Modellen i cpp-filen er bygget slik at simultantettheten har blitt tildelt et negativt fortegn, og steget her vil følgelig tilsvare en maksimering av log-likelihood-funksjonen.

```
opt_begr_mod1<- nlminb(obj_begr_mod1$par,
                       obj_begr_mod1$fn,
                       obj_begr_mod1$gr,
                       lower=c(rep(x=-Inf,times=10), 0.001),
                       upper=c(rep(x=Inf,times=10), 100))
```

Ved hjelp av kommandoen «sdreport» får man informasjon om standardavvik og p-verdier hos de ulike estimatene. Her er det mulig å skille mellom ikke-randomiserte og randomiserte effekter.

```
sdreport_begr_mod1<-sdreport(obj_begr_mod1)
```

Eksempel på utskrift av resultater fra steget over. I dette tilfellet har man gjennom kommandoen «fixed» begrenset utskriften til ikke-randomiserte effekter.

```
round(summary(sdreport_begr_mod1, "fixed", p.value=TRUE), 4)
```

	Estimate	Std. Error	p.value
b0	−2.1081	0.1185	0.0000
b2	0.0090	0.0011	0.0000
b4	−0.0558	0.0560	0.3185
b5	−0.6540	0.0788	0.0000
b6	0.3737	0.0769	0.0000
b7	0.0663	0.0809	0.4122
b9	−0.1562	0.0766	0.0415
b10	0.1371	0.0882	0.1199
sigma	0.2027	0.0974	0.0374
tau	0.0000	0.1168	1.0000
delta	0.0164	0.0351	0.6409

Vedlegg D: Tabeller

Tabell D.1: Oversikt over data brukt til simulering (kommune).

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
1	Bjugn	Sør-Trøndelag	4.584	355, 83	19, 90	402.000	12, 88
2	Rissa	Sør-Trøndelag	6.543	588, 05	19, 70	430.000	11, 13
3	Leksvik	Nord-Trøndelag	3.527	399, 68	15, 60	448.000	8, 82
4	Frosta	Nord-Trøndelag	2.618	74, 30	24, 10	418.000	35, 24
5	Aure	Møre og Romsdal	3.511	624, 25	13, 40	402.000	5, 62
6	Aukra	Møre og Romsdal	3.289	58, 82	14, 30	499.000	55, 92
7	Sandøy	Møre og Romsdal	1.315	20, 35	6, 80	475.000	64, 62
8	Tydal	Sør-Trøndelag	870	1.217, 05	10, 30	430.000	0, 71
9	Sande	Møre og Romsdal	2.588	90, 10	15, 10	413.000	28, 72
10	Ørskog	Møre og Romsdal	2.202	128, 85	23, 60	482.000	17, 09
11	Ørsta	Møre og Romsdal	10.398	785, 50	20, 90	463.000	13, 24
12	Alvdal	Hedmark	2.431	918, 71	17, 30	414.000	2, 65
13	Askvoll	Sogn og Fjordane	3.010	313, 47	12, 30	435.000	9, 60
14	Balestrand	Sogn og Fjordane	1.338	411, 21	13, 50	392.000	3, 25
15	Solund	Sogn og Fjordane	851	219, 19	5, 90	423.000	3, 88
16	Leikanger	Sogn og Fjordane	2.236	177, 38	7, 60	502.000	12, 61
17	Fedje	Hordaland	576	8, 93	8, 70	434.000	64, 50
18	Hemsedal	Buskerud	2.228	711, 76	145	417.000	3, 13
19	Askøy	Hordaland	26.210	93, 98	22, 10	535.000	278, 89
20	Meland	Hordaland	7.036	87, 54	21, 70	529.000	80, 37
21	Granvin	Hordaland	923	205, 01	13	402.000	4, 50
22	Sund	Hordaland	6.409	94, 85	22, 60	530.000	67, 57
23	Hurdal	Akershus	2.657	260, 98	34, 20	412.000	10, 18
24	Austevoll	Hordaland	4.792	114, 26	14, 40	548.000	41, 94
25	Nannestad	Akershus	11.362	324, 33	29, 50	481.000	35, 03
26	Fitjar	Hordaland	2.944	134, 45	19, 40	506.000	21, 90
27	Nordkapp	Finnmark	3.228	890, 76	31, 60	403.000	3, 62
28	Måsøy	Finnmark	1.243	1.066, 56	32, 20	352.000	1, 17
29	Skjervøy	Troms	2.880	464, 27	22, 20	429.000	6, 20
30	Berg	Troms	887	276, 54	15, 80	396.000	3, 21
31	Harstad	Troms	23.640	355, 62	26, 90	433.000	66, 48
32	Ibestad	Troms	1.410	234, 37	15, 60	369.000	6, 02

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
33	Tjeldsund	Nordland	1.284	309, 34	23, 40	425.000	4, 15
34	Moskenes	Nordland	1.116	110, 26	17, 90	372.000	10, 12
35	Røst	Nordland	595	10, 13	21, 80	421.000	58, 74
36	Dønna	Nordland	1.433	187, 64	23	402.000	7, 64
37	Alstahaug	Nordland	7.372	186, 29	41, 40	447.000	39, 57
38	Vikna	Nord-Trøndelag	4.241	310, 21	29, 90	441.000	13, 67
39	Torsken	Troms	892	235, 08	14, 60	341.000	3, 79
40	Tranøy	Troms	1.524	499, 17	27, 60	379.000	3, 05
41	Øksnes	Nordland	4.467	310, 15	29, 10	414.000	14, 40
42	Salangen	Troms	2.214	438, 05	32, 50	405.000	5, 05
43	Skånland	Troms	2.972	464, 84	23, 20	426.000	6, 39
44	Vestvågøy	Nordland	10.848	405, 58	27, 40	409.000	26, 75
45	Herøy	Nordland	1.711	64, 22	16, 90	370.000	26, 64
46	Vega	Nordland	1.256	161, 05	25, 50	396.000	7, 80
47	Sømna	Nordland	2.038	191, 19	17, 70	413.000	10, 66
48	Leka	Nord-Trøndelag	573	107, 97	21, 70	401.000	5, 31
49	Fosnes	Nord-Trøndelag	668	473, 39	19, 50	425.000	1, 41
50	Lenvik	Troms	11.345	848, 74	37, 90	431.000	13, 37
51	Andøy	Nordland	5.032	616, 04	26, 80	416.000	8, 17
52	Balsfjord	Troms	5.502	1.440, 69	33, 80	405.000	3, 82
53	Målselv	Troms	6.599	3.206, 97	33, 90	446.000	2, 06
54	Dyrøy	Troms	1.188	276, 89	22, 70	374.000	4, 29
55	Sortland	Nordland	9.983	698, 22	26, 80	417.000	14, 30
56	Kvæfjord	Troms	3.025	497, 59	16, 20	406.000	6, 08
57	Hadsel	Nordland	7.937	551, 12	22, 40	405.000	14, 40
58	Steigen	Nordland	2.609	962, 87	21, 10	384.000	2, 71
59	Saltdal	Nordland	4.710	2.085, 86	40, 30	417.000	2, 26
60	Leirfjord	Nordland	2.107	451, 21	20, 90	413.000	4, 67
61	Hattfjelldal	Nordland	1.456	2.414, 88	16, 50	401.000	0, 60
62	Høylandet	Nord-Trøndelag	1.264	702, 44	10, 30	444.000	1, 80
63	Frøya	Sør-Trøndelag	4.369	229, 81	23, 30	393.000	19, 01
64	Åfjord	Sør-Trøndelag	3.257	895, 72	13, 20	411.000	3, 64
65	Smøla	Møre og Romsdal	2.182	270, 44	11	405.000	8, 07
66	Ørland	Sør-Trøndelag	5.119	73, 48	27, 20	427.000	69, 67
67	Trondheim	Sør-Trøndelag	176.348	321, 81	62, 50	410.000	547, 99
68	Halsa	Møre og Romsdal	1.641	292, 77	18, 30	401.000	5, 61
69	Selje	Sogn og Fjordane	2.831	219, 90	15, 20	449.000	12, 87
70	Hareid	Møre og Romsdal	5.000	76, 87	24, 20	453.000	65, 04
71	Hornindal	Sogn og Fjordane	1.220	178, 44	8, 20	480.000	6, 84
72	Suldal	Rogaland	3.845	1.588, 80	18, 70	444.000	2, 42
73	Nedre Eiker	Buskerud	23.262	114, 04	35, 30	463.000	203, 98
74	Karmøy	Rogaland	40.536	219, 24	35, 60	475.000	184, 89
75	Utsira	Rogaland	218	6, 27	21, 70	538.000	34, 77
76	Rømskog	Østfold	688	158, 88	29, 10	440.000	4, 33
77	Sauherad	Telemark	4.314	289, 59	35, 50	435.000	14, 90

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
78	Valle	Aust-Agder	1.293	1.131, 91	36, 30	404.000	1, 14
79	Lardal	Vestfold	2.413	270, 64	23, 20	423.000	8, 92
80	Rennesøy	Rogaland	4.388	64, 58	20, 70	579.000	67, 95
81	Kvitsøy	Rogaland	519	6, 16	21, 70	475.000	84, 25
82	Drangedal	Telemark	4.126	996, 45	35, 60	398.000	4, 14
83	Porsgrunn	Telemark	35.219	160, 90	64, 30	419.000	218, 89
84	Kviteseid	Telemark	2.498	623, 88	28	404.000	4, 00
85	Marker	Østfold	3.518	367, 65	52, 30	392.000	9, 57
86	Eidsberg	Østfold	11.049	228, 76	49, 10	401.000	48, 30
87	Hjelmeland	Rogaland	2.807	974, 46	18, 90	469.000	2, 88
88	Skien	Telemark	52.509	718, 91	59, 60	413.000	73, 04
89	Fyresdal	Telemark	1.335	1.107, 81	23, 20	420.000	1, 21
90	Nome	Telemark	6.579	385, 83	43, 90	395.000	17, 05
91	Forsand	Rogaland	1.190	701, 38	23, 50	539.000	1, 70
92	Sirdal	Vest-Agder	1.816	1.375, 16	41, 90	478.000	1, 32
93	Aremark	Østfold	1.423	281, 94	35, 80	439.000	5, 05
94	Bygland	Aust-Agder	1.219	1.150, 18	41, 80	410.000	1, 06
95	Gjesdal	Rogaland	10.778	557, 72	35, 10	554.000	19, 33
96	Hå	Rogaland	17.244	247, 73	32, 60	504.000	69, 61
97	Risør	Aust-Agder	6.899	178, 98	37	395.000	38, 55
98	Eigersund	Rogaland	14.475	386, 61	36, 70	467.000	37, 44
99	Hægebostad	Vest-Agder	1.665	424, 42	18, 60	491.000	3, 92
100	Froland	Aust-Agder	5.257	601, 02	37, 10	473.000	8, 75
101	Birkenes	Aust-Agder	4.828	629, 91	25, 10	463.000	7, 66
102	Arendal	Aust-Agder	42.801	254, 99	62, 90	423.000	167, 85
103	Audnedal	Vest-Agder	1.689	236, 05	15, 40	493.000	7, 16
104	Vennesla	Vest-Agder	13.583	362, 04	49, 50	450.000	37, 52
105	Farsund	Vest-Agder	9.433	251, 74	29, 80	443.000	37, 47
106	Lindesnes	Vest-Agder	4.753	297, 26	28, 40	425.000	15, 99
107	Mandal	Vest-Agder	15.149	210, 37	52, 10	429.000	72, 01
108	Klepp	Rogaland	17.746	102, 36	32, 20	531.000	173, 37
109	Åmli	Aust-Agder	1.825	1.058, 42	36, 20	393.000	1, 72
110	Åseral	Vest-Agder	912	797, 75	31, 80	411.000	1, 14
111	Lebesby	Finnmark	1.356	3.231, 92	27, 30	364.000	0, 42
112	Storfjord	Troms	1.909	1.477, 70	28, 30	443.000	1, 29
113	Lødingen	Nordland	2.179	504, 71	21, 10	393.000	4, 32
114	Vågan	Nordland	9.086	459, 17	27, 80	402.000	19, 79
115	Hamarøy	Nordland	1.783	920, 31	38, 70	363.000	1, 94
116	Vestre Slidre	Oppland	2.232	420, 96	30	376.000	5, 30
117	Voss	Hordaland	13.978	1.732	24, 10	443.000	8, 07
118	Gjerstad	Aust-Agder	2.478	307, 65	31, 90	407.000	8, 05
119	Vegårshei	Aust-Agder	1.933	321, 77	20, 70	459.000	6, 01
120	Evje og Hornnes	Aust-Agder	3.496	514, 63	51, 80	420.000	6, 79
121	Tvedestrand	Aust-Agder	6.019	203, 98	53	416.000	29, 51
122	Flekkefjord	Vest-Agder	9.046	481, 56	26, 30	448.000	18, 78

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
123	Iveland	Aust-Agder	1.298	246, 31	18, 50	477.000	5, 27
124	Marnardal	Vest-Agder	2.286	375, 89	24, 10	458.000	6, 08
125	Songdalen	Vest-Agder	6.165	206, 29	37, 30	476.000	29, 89
126	Kristiansand	Vest-Agder	83.243	258, 79	69, 50	422.000	321, 66
127	Søgne	Vest-Agder	10.855	143, 76	38	481.000	75, 51
128	Time	Rogaland	16.769	171, 10	47, 30	527.000	98, 01
129	Bjerkreim	Rogaland	2.739	577, 29	20, 10	545.000	4, 74
130	Lavangen	Troms	1.016	296, 20	23, 60	430.000	3, 43
131	Lærdal	Sogn og Fjordane	2.205	1.277, 42	30, 40	470.000	1, 73
132	Austrheim	Hordaland	2.776	56, 70	28, 50	459.000	48, 96
133	Lindås	Hordaland	14.668	457, 24	31, 20	507.000	32, 08
134	Fjell	Hordaland	22.720	141, 15	26, 90	554.000	160, 96
135	Østre Toten	Oppland	14.747	485, 60	23, 50	412.000	30, 37
136	Fusa	Hordaland	3.811	354, 90	16, 80	498.000	10, 74
137	Larvik	Vestfold	42.947	500, 36	50, 20	416.000	85, 83
138	Bindal	Nordland	1.562	1.193, 26	19, 80	383.000	1, 31
139	Røyrvik	Nord-Trøndelag	494	1.332, 40	14, 20	424.000	0, 37
140	Flatanger	Nord-Trøndelag	1.141	434, 17	10, 50	429.000	2, 63
141	Roan	Sør-Trøndelag	987	355, 44	11, 10	398.000	2, 78
142	Hitra	Sør-Trøndelag	4.399	645, 63	21, 40	371.000	6, 81
143	Snillfjord	Sør-Trøndelag	981	489, 12	18, 30	395.000	2, 01
144	Skaun	Sør-Trøndelag	6.941	213, 05	15, 40	511.000	32, 58
145	Meldal	Sør-Trøndelag	3.924	592, 54	21, 20	408.000	6, 62
146	Sula	Møre og Romsdal	8.256	57, 27	15, 50	491.000	144, 16
147	Stordal	Møre og Romsdal	1.026	243, 62	12, 70	443.000	4, 21
148	Vanylven	Møre og Romsdal	3.388	365, 41	11, 80	455.000	9, 27
149	Dovre	Oppland	2.742	1.348, 45	32, 80	386.000	2, 03
150	Sør-Fron	Oppland	3.208	711, 51	12, 50	389.000	4, 51
151	Rælingen	Akershus	16.170	56, 39	28, 30	492.000	286, 75
152	Asker	Akershus	56.447	96, 82	33, 40	547.000	583, 01
153	Seljord	Telemark	2.959	670, 22	54, 40	408.000	4, 41
154	Drammen	Buskerud	64.597	134, 73	60, 70	402.000	479, 46
155	Tysvær	Rogaland	10.320	399, 39	24, 40	517.000	25, 84
156	Malvik	Sør-Trøndelag	12.785	162	25, 20	533.000	78, 92
157	Fræna	Møre og Romsdal	9.484	361, 80	19, 60	469.000	26, 21
158	Tingvoll	Møre og Romsdal	3.101	321, 78	14, 80	403.000	9, 64
159	Holtålen	Sør-Trøndelag	2.013	1.172	7, 90	439.000	1, 72
160	Herøy	Møre og Romsdal	8.727	118, 25	14, 20	494.000	73, 80
161	Os	Hedmark	2.040	1.008, 51	18, 10	414.000	2, 02
162	Sykkylven	Møre og Romsdal	7.664	328, 49	15, 90	471.000	23, 33
163	Vågsøy	Sogn og Fjordane	6.129	171, 14	23, 70	446.000	35, 81
164	Bremanger	Sogn og Fjordane	3.891	786, 55	14, 90	431.000	4, 95
165	Flora	Sogn og Fjordane	11.654	646, 51	34, 50	460.000	18, 03
166	Naustdal	Sogn og Fjordane	2.710	355, 85	13, 30	488.000	7, 62
167	Førde	Sogn og Fjordane	12.307	553, 12	41, 20	466.000	22, 25

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
168	Sogndal	Sogn og Fjordane	7.348	736, 13	26, 30	416.000	9, 98
169	Hyllestad	Sogn og Fjordane	1.461	247, 98	4, 10	394.000	5, 89
170	Vang	Oppland	1.617	1.311, 53	27, 20	403.000	1, 23
171	Vik	Sogn og Fjordane	2.748	797, 33	17, 80	426.000	3, 45
172	Øygarden	Hordaland	4.419	64, 21	19, 20	504.000	68, 82
173	Osterøy	Hordaland	7.521	243, 76	16, 40	471.000	30, 85
174	Søndre Land	Oppland	5.761	659, 52	36, 80	384.000	8, 74
175	Flå	Buskerud	1.034	670, 97	42, 60	370.000	1, 54
176	Os	Hordaland	17.726	133, 72	33, 60	511.000	132, 56
177	Gran	Oppland	13.493	658, 54	27, 60	427.000	20, 49
178	Jevnaker	Oppland	6.483	194, 73	35, 80	432.000	33, 29
179	Krødsherad	Buskerud	2.186	339, 79	72, 30	407.000	6, 43
180	Bømlo	Hordaland	11.503	234, 96	17	500.000	48, 96
181	Gamvik	Finnmark	1.008	1.353, 64	43, 70	373.000	0, 74
182	Berlevåg	Finnmark	1.015	1.082, 43	41, 40	366.000	0, 94
183	Hasvik	Finnmark	995	534, 46	46, 20	365.000	1, 86
184	Porsanger	Finnmark	3.946	4.640, 95	57, 30	423.000	0, 85
185	Kåfjord	Troms	2.210	950, 24	17, 60	379.000	2, 33
186	Bjarkøy	Troms	455	73, 44	21, 70	408.000	6, 20
187	Sørreisa	Troms	3.381	346, 94	26, 30	456.000	9, 75
188	Bø	Nordland	2.720	235, 84	13, 20	353.000	11, 53
189	Gratangen	Troms	1.136	305, 41	16, 70	354.000	3, 72
190	Evenes	Nordland	1.359	241, 23	36, 10	417.000	5, 63
191	Flakstad	Nordland	1.383	168, 81	15, 20	431.000	8, 19
192	Værøy	Nordland	751	18, 49	24	364.000	40, 62
193	Træna	Nordland	497	16, 25	36, 20	400.000	30, 58
194	Nesna	Nordland	1.813	181, 34	36, 40	330.000	10, 00
195	Høyanger	Sogn og Fjordane	4.216	838, 34	34, 40	426.000	5, 03
196	Vaksdal	Hordaland	4.138	683, 80	12, 30	436.000	6, 05
197	Kvam	Hordaland	8.522	580, 97	21	452.000	14, 67
198	Tysnes	Hordaland	2.766	245, 27	23, 50	413.000	11, 28
199	Nore og Uvdal	Buskerud	2.540	2.273, 89	18, 90	396.000	1, 12
200	Nord-Odal	Hedmark	5.141	475, 42	19, 80	400.000	10, 81
201	Grue	Hedmark	5.003	777, 10	33, 60	369.000	6, 44
202	Kongsvinger	Hedmark	17.522	952, 71	65, 30	383.000	18, 39
203	Sør-Odal	Hedmark	7.859	479, 08	30, 30	427.000	16, 40
204	Nes	Akershus	19.462	608, 97	31, 20	480.000	31, 96
205	Modum	Buskerud	13.116	462, 96	37, 70	411.000	28, 33
206	Eidskog	Hedmark	6.288	603, 34	38, 60	369.000	10, 42
207	Flesberg	Buskerud	2.638	537, 84	23, 50	456.000	4, 90
208	Skedsmo	Akershus	49.698	75, 02	61, 20	462.000	662, 46
209	Aurskog-Høland	Akershus	14.905	893, 62	25, 20	435.000	16, 68
210	Fet	Akershus	10.626	137, 45	31, 80	529.000	77, 31
211	Notodden	Telemark	12.469	852, 32	47, 50	391.000	14, 63
212	Nesodden	Akershus	17.809	60, 76	26, 80	507.000	293, 10

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
213	Kongsberg	Buskerud	25.479	754, 16	42, 60	466.000	33, 78
214	Trøgstad	Østfold	5.219	187, 68	36	418.000	27, 81
215	Hurum	Buskerud	9.185	156, 40	30, 60	481.000	58, 73
216	Finnøy	Rogaland	2.955	102, 75	14, 60	491.000	28, 76
217	Skiptvet	Østfold	3.631	93, 04	31, 40	451.000	39, 03
218	Re	Vestfold	8.936	222, 26	28, 10	457.000	40, 21
219	Andebu	Vestfold	5.448	182, 79	19, 80	437.000	29, 80
220	Randaberg	Rogaland	10.265	24, 11	20, 90	569.000	425, 76
221	Sandnes	Rogaland	67.814	285, 55	53, 20	511.000	237, 49
222	Tjøme	Vestfold	4.813	39, 29	39, 70	439.000	122, 50
223	Lillesand	Aust-Agder	9.878	180, 32	39, 50	486.000	54, 78
224	Årdal	Sogn og Fjordane	5.572	930, 90	18, 30	436.000	5, 99
225	Rana	Nordland	25.652	4.205, 93	54, 10	436.000	6, 10
226	Narvik	Nordland	18.473	1.905, 72	43, 50	415.000	9, 69
227	Loppa	Finnmark	1.087	669, 35	11	366.000	1, 62
228	Kvænangen	Troms	1.284	2.007, 67	24, 10	400.000	0, 64
229	Åsnes	Hedmark	7.606	1.003, 93	49, 60	373.000	7, 58
230	Hjartdal	Telemark	1.602	738, 76	19, 40	425.000	2, 17
231	Frogn	Akershus	15.154	84, 64	36	516.000	179, 04
232	Bø	Telemark	5.766	258, 43	69	376.000	22, 31
233	Siljan	Telemark	2.432	202, 20	24, 70	500.000	12, 03
234	Bamble	Telemark	14.106	282, 06	53, 40	469.000	50, 01
235	Lund	Rogaland	3.183	353, 89	17	466.000	8, 99
236	Surnadal	Møre og Romsdal	5.952	1.314, 22	20, 80	420.000	4, 53
237	Båtsfjord	Finnmark	2.089	1.415, 36	33	391.000	1, 48
238	Vadsø	Finnmark	6.125	1.233, 90	54, 90	449.000	4, 96
239	Nesseby	Finnmark	901	1.366, 89	42, 20	375.000	0, 66
240	Hemne	Sør-Trøndelag	4.221	635, 71	23, 70	440.000	6, 64
241	Norddal	Møre og Romsdal	1.738	900, 51	15	435.000	1, 93
242	Aurland	Sogn og Fjordane	1.712	1.382, 59	27, 50	422.000	1, 24
243	Stord	Hordaland	17.957	137, 16	33	482.000	130, 92
244	Ulstein	Møre og Romsdal	7.828	95, 08	20, 20	462.000	82, 33
245	Beiarn	Nordland	1.097	1.181, 11	14, 60	386.000	0, 93
246	Rødøy	Nordland	1.320	686, 54	8, 30	394.000	1, 92
247	Lurøy	Nordland	1.937	257, 63	23, 20	401.000	7, 52
248	Levanger	Nord-Trøndelag	18.922	610	30, 10	441.000	31, 02
249	Melhus	Sør-Trøndelag	15.392	653, 76	24, 40	481.000	23, 54
250	Midtre Gauldal	Sør-Trøndelag	6.153	1.807, 48	17, 40	397.000	3, 40
251	Oppdal	Sør-Trøndelag	6.755	2.201, 32	35, 20	422.000	3, 07
252	Vefsn	Nordland	13.258	1.839, 96	60, 60	419.000	7, 21
253	Gulen	Sogn og Fjordane	2.310	575, 41	15, 20	415.000	4, 01
254	Sande	Vestfold	8.680	174, 20	30, 10	479.000	49, 83
255	Holmestrand	Vestfold	10.251	84, 23	37, 80	435.000	121, 70
256	Tønsberg	Vestfold	40.677	106, 07	69, 30	406.000	383, 49
257	Kristiansund	Møre og Romsdal	23.813	86, 28	42, 30	419.000	276, 00

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
258	Gjøvik	Oppland	29.202	630, 42	43, 30	397.000	46, 32
259	Ålesund	Møre og Romsdal	44.416	93, 33	43, 80	435.000	475, 90
260	Rakkestad	Østfold	7.698	420, 65	50, 80	418.000	18, 30
261	Fredrikstad	Østfold	75.583	283, 45	54, 90	415.000	266, 65
262	Sola	Rogaland	23.877	68, 88	40	562.000	346, 65
263	Kvinnherad	Hordaland	13.318	1.079, 78	19, 40	455.000	12, 33
264	Nord-Fron	Oppland	5.830	1.092, 21	17	398.000	5, 34
265	Jølster	Sogn og Fjordane	3.052	620, 12	12, 10	463.000	4, 92
266	Oslo	Oslo	613.285	426, 29	113	377.000	1.438, 66
267	Halden	Østfold	29.543	595, 50	69, 40	400.000	49, 61
268	Stange	Hedmark	19.190	640, 90	32, 10	426.000	29, 94
269	Ullensaker	Akershus	31.044	250, 33	142, 30	477.000	124, 01
270	Gildeskål	Nordland	2.000	619, 09	25	410.000	3, 23
271	Luster	Sogn og Fjordane	5.026	2.601, 57	13, 50	445.000	1, 93
272	Namsskogan	Nord-Trøndelag	916	1.353, 90	22, 90	405.000	0, 68
273	Osen	Sør-Trøndelag	1.020	369, 99	11, 80	407.000	2, 76
274	Namdalseid	Nord-Trøndelag	1.694	735, 01	18, 90	408.000	2, 30
275	Skjåk	Oppland	2.307	1.968, 47	16	401.000	1, 17
276	Orkdal	Sør-Trøndelag	11.429	564, 18	30, 40	445.000	20, 26
277	Eide	Møre og Romsdal	3.442	145, 75	22, 10	463.000	23, 62
278	Stryn	Sogn og Fjordane	7.065	1.321, 94	19, 70	428.000	5, 34
279	Nesset	Møre og Romsdal	3.004	985, 87	12	459.000	3, 05
280	Røros	Sør-Trøndelag	5.604	1.757, 84	28, 60	413.000	3, 19
281	Stranda	Møre og Romsdal	4.602	844, 90	16, 90	451.000	5, 45
282	Fauske	Nordland	9.513	1.108, 42	40, 10	433.000	8, 58
283	Eid	Sogn og Fjordane	5.950	420, 22	23	469.000	14, 16
284	Sørfold	Nordland	2.003	1.472, 36	29, 50	417.000	1, 36
285	Bardu	Troms	3.875	2.515, 73	28, 10	465.000	1, 54
286	Øystre Slidre	Oppland	3.174	880, 75	25, 20	411.000	3, 60
287	Modalen	Hordaland	370	380, 86	21, 70	387.000	0, 97
288	Lom	Oppland	2.382	1.888, 98	15, 50	404.000	1, 26
289	Gol	Buskerud	4.581	514, 15	43, 20	396.000	8, 91
290	Bergen	Hordaland	263.762	444, 98	62, 60	419.000	592, 75
291	Nes	Buskerud	3.452	772, 44	35, 60	382.000	4, 47
292	Våler	Hedmark	3.844	677, 57	68, 90	390.000	5, 67
293	Radøy	Hordaland	4.952	107, 17	16, 40	472.000	46, 21
294	Ål	Buskerud	4.741	1.083, 35	18, 40	408.000	4, 38
295	Sør-Aurdal	Oppland	3.154	1.070, 53	20	385.000	2, 95
296	Samnanger	Hordaland	2.417	257, 77	17, 40	490.000	9, 38
297	Ringerike	Buskerud	29.236	1.423, 20	48, 70	403.000	20, 54
298	Sigdal	Buskerud	3.535	811, 20	22, 90	423.000	4, 36
299	Lunner	Oppland	8.776	272, 24	26, 50	486.000	32, 24
300	Rollag	Buskerud	1.383	430, 22	28, 20	413.000	3, 21
301	Gjerdrum	Akershus	6.152	82, 16	26, 50	539.000	74, 88
302	Sørum	Akershus	16.091	199, 88	28, 50	517.000	80, 50

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
303	Sauda	Rogaland	4.754	508, 56	26, 30	410.000	9, 35
304	Sveio	Hordaland	5.228	223, 90	27	478.000	23, 35
305	Lier	Buskerud	24.177	281, 80	34, 20	492.000	85, 79
306	Lørenskog	Akershus	33.709	67, 26	43	483.000	501, 17
307	Haugesund	Rogaland	35.099	68, 40	80, 50	403.000	513, 14
308	Enebakk	Akershus	10.487	195, 45	28, 50	513.000	53, 66
309	Røyken	Buskerud	19.594	111, 18	23, 70	532.000	176, 24
310	Svelvik	Vestfold	6.581	56, 19	24, 90	462.000	117, 12
311	Hof	Vestfold	3.048	148, 19	34, 40	456.000	20, 57
312	Bokn	Rogaland	851	44, 57	23, 50	487.000	19, 09
313	Horten	Vestfold	26.307	68, 36	59, 90	417.000	384, 83
314	Rygge	Østfold	14.691	69, 70	55, 30	459.000	210, 77
315	Strand	Rogaland	11.533	195, 24	33, 50	506.000	59, 07
316	Nøtterøy	Vestfold	20.995	60, 74	27, 90	462.000	345, 65
317	Sandefjord	Vestfold	44.150	118, 60	57	408.000	372, 26
318	Hole	Buskerud	6.322	134, 22	25, 50	493.000	47, 10
319	Rennebu	Sør-Trøndelag	2.569	924, 96	24, 10	426.000	2, 78
320	Sel	Oppland	5.992	888, 22	21, 90	393.000	6, 75
321	Gloppen	Sogn og Fjordane	5.679	964, 23	21	459.000	5, 89
322	Vestnes	Møre og Romsdal	6.539	347, 62	19, 30	429.000	18, 81
323	Grimstad	Aust-Agder	21.301	272, 26	42, 60	440.000	78, 24
324	Eidsvoll	Akershus	21.621	385, 50	43, 50	450.000	56, 09
325	Oppegård	Akershus	25.520	34, 28	29, 70	551.000	744, 46
326	Nærøy	Nord-Trøndelag	5.069	1.011, 80	25, 80	407.000	5, 01
327	Agdenes	Sør-Trøndelag	1.715	296, 87	11, 10	430.000	5, 78
328	Meråker	Nord-Trøndelag	2.513	1.188, 27	21, 50	370.000	2, 11
329	Haram	Møre og Romsdal	8.973	253, 75	20, 30	465.000	35, 36
330	Midsund	Møre og Romsdal	1.988	93, 99	6, 50	459.000	21, 15
331	Skodje	Møre og Romsdal	4.184	110, 97	17, 90	515.000	37, 70
332	Folldal	Hedmark	1.641	1.259, 02	15, 20	400.000	1, 30
333	Bodø	Nordland	48.422	1.308, 57	42, 70	457.000	37, 00
334	Tana	Finnmark	2.896	3.831, 02	39, 70	403.000	0, 76
335	Karasjok	Finnmark	2.763	5.209, 45	48, 50	418.000	0, 53
336	Nordreisa	Troms	4.807	3.335, 37	37, 70	419.000	1, 44
337	Kautokeino	Finnmark	2.927	8.970, 28	44, 80	430.000	0, 33
338	Lierne	Nord-Trøndelag	1.410	2.631, 23	6, 40	412.000	0, 54
339	Snåsa	Nord-Trøndelag	2.164	2.150, 34	19, 40	427.000	1, 01
340	Verdal	Nord-Trøndelag	14.387	1.479, 49	36, 80	422.000	9, 72
341	Stjørdal	Nord-Trøndelag	22.058	913, 26	38, 10	459.000	24, 15
342	Klæbu	Sør-Trøndelag	5.930	175, 06	25, 10	526.000	33, 87
343	Selbu	Sør-Trøndelag	4.042	1.140, 84	20, 50	440.000	3, 54
344	Tynset	Hedmark	5.564	1.822, 60	32, 20	420.000	3, 05
345	Tolga	Hedmark	1.681	1.097, 57	14, 30	399.000	1, 53
346	Sunndal	Møre og Romsdal	7.196	1.647, 90	30, 70	419.000	4, 37
347	Gaular	Sogn og Fjordane	2.848	539, 03	8, 40	471.000	5, 28

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
348	Fjaler	Sogn og Fjordane	2.832	390, 10	12	375.000	7, 26
349	Tinn	Telemark	5.982	1.854, 28	27, 20	386.000	3, 23
350	Vevelstad	Nordland	511	517, 14	21, 70	394.000	0, 99
351	Grane	Nordland	1.455	1.883, 44	39, 90	395.000	0, 77
352	Brønnøy	Nordland	7.778	1.000, 59	30, 20	422.000	7, 77
353	Masfjorden	Hordaland	1.683	511, 23	10, 70	502.000	3, 29
354	Vestby	Akershus	15.143	133, 61	45, 90	518.000	113, 34
355	Moss	Østfold	30.723	57, 90	66, 10	396.000	530, 62
356	Stokke	Vestfold	11.178	115, 52	31, 80	447.000	96, 76
357	Hvaler	Østfold	4.206	89, 49	39	459.000	47, 00
358	Averøy	Møre og Romsdal	5.593	173, 10	16, 80	479.000	32, 31
359	Vestre Toten	Oppland	12.928	231, 37	31, 80	421.000	55, 88
360	Sør-Varanger	Finnmark	9.860	3.467, 24	39, 20	442.000	2, 84
361	Åmot	Hedmark	4.337	1.292, 95	32, 30	367.000	3, 35
362	Elverum	Hedmark	20.152	1.209, 16	49	410.000	16, 67
363	Ullensvang	Hordaland	3.417	1.287, 45	12	449.000	2, 65
364	Inderøy	Nord-Trøndelag	6.682	350, 67	12, 40	464.000	19, 05
365	Tysfjord	Nordland	1.956	1.359, 89	23	368.000	1, 44
366	Råde	Østfold	6.987	105, 28	45, 50	474.000	66, 37
367	Sarpsborg	Østfold	53.333	369, 85	54	402.000	144, 20
368	Stavanger	Rogaland	127.506	68, 11	62, 70	476.000	1.872, 06
369	Grong	Nord-Trøndelag	2.409	1.097, 65	29, 50	390.000	2, 19
370	Overhalla	Nord-Trøndelag	3.679	688, 79	11, 70	461.000	5, 34
371	Askim	Østfold	15.096	66, 07	50, 90	415.000	228, 48
372	Ski	Akershus	28.970	161, 72	45, 60	525.000	179, 14
373	Ås	Akershus	17.284	101, 29	41, 70	467.000	170, 64
374	Vinje	Telemark	3.700	2.731, 65	35, 40	419.000	1, 35
375	Rendalen	Hedmark	1.959	3.061, 14	20, 40	361.000	0, 64
376	Engerdal	Hedmark	1.390	1.916, 03	30, 20	394.000	0, 73
377	Gausdal	Oppland	6.160	1.146, 28	12, 50	418.000	5, 37
378	Nord-Aurdal	Oppland	6.428	850, 16	40, 60	389.000	7, 56
379	Nordre Land	Oppland	6.768	920, 87	26, 40	389.000	7, 35
380	Etnedal	Oppland	1.408	443, 37	16, 30	329.000	3, 18
381	Jondal	Hordaland	1.050	199, 14	14, 30	421.000	5, 27
382	Verran	Nord-Trøndelag	2.705	558, 06	14	336.000	4, 85
383	Namsos	Nord-Trøndelag	12.953	751, 75	31	437.000	17, 23
384	Steinkjer	Nord-Trøndelag	21.303	1.423, 30	29, 60	421.000	14, 97
385	Stor-Elvdal	Hedmark	2.678	2.124, 81	37, 70	338.000	1, 26
386	Ringebu	Oppland	4.561	1.221, 45	18	393.000	3, 73
387	Etne	Hordaland	3.963	692, 45	28, 50	463.000	5, 72
388	Øyer	Oppland	5.095	616, 17	33, 40	423.000	8, 27
389	Ringsaker	Hedmark	33.191	1.122, 94	31, 20	425.000	29, 56
390	Lillehammer	Oppland	26.765	450, 77	37, 40	409.000	59, 38
391	Hamar	Hedmark	29.045	337, 60	51, 90	403.000	86, 03
392	Ballangen	Nordland	2.616	846, 84	19, 10	392.000	3, 09

Fortsetter neste side

Tabell D.1 – *Fortsetter fra forrige side*

ID	Kommune	Fylke	Bosatte	Areal	Krim.rate	Inntekt	Urb.grad
393	Løten	Hedmark	7.477	362, 14	29, 70	418.000	20, 65
394	Trysil	Hedmark	6.752	2.940, 27	60, 90	365.000	2, 30
395	Eidfjord	Hordaland	957	1.387, 41	24	433.000	0, 69
396	Giske	Møre og Romsdal	7.312	39, 53	19, 10	523.000	184, 97
397	Odda	Hordaland	6.946	1.478, 30	52, 30	410.000	4, 70
398	Hobøl	Østfold	4.911	139, 39	31, 20	468.000	35, 23
399	Spydeberg	Østfold	5.348	133, 48	31, 40	458.000	40, 07
400	Gjemnes	Møre og Romsdal	2.579	371, 42	13, 60	456.000	6, 94
401	Nittedal	Akershus	21.454	179, 48	30, 60	531.000	119, 53
402	Bærum	Akershus	114.489	188, 86	37, 40	526.000	606, 21
403	Molde	Møre og Romsdal	25.488	355, 93	39, 60	441.000	71, 61
404	Øvre Eiker	Buskerud	17.421	417, 89	40, 20	435.000	41, 69
405	Bykle	Aust-Agder	970	1.261, 41	62, 90	420.000	0, 77
406	Tokke	Telemark	2.287	904, 81	30, 20	406.000	2, 53
407	Våler	Østfold	4.705	238, 88	28, 50	494.000	19, 70
408	Nissedal	Telemark	1.430	787, 15	33, 60	433.000	1, 82
409	Kragerø	Telemark	10.710	288, 76	50, 10	395.000	37, 09
410	Kvinesdal	Vest-Agder	5.834	886, 41	28, 30	433.000	6, 58
411	Sokndal	Rogaland	3.257	267, 12	23, 30	444.000	12, 19
412	Lyngdal	Vest-Agder	7.895	369, 95	40, 40	442.000	21, 34
413	Lyngen	Troms	3.028	796, 28	17, 80	422.000	3, 80
414	Hammerfest	Finnmark	9.934	819, 80	49, 90	457.000	12, 12
415	Vardø	Finnmark	2.122	585, 45	40, 10	360.000	3, 62
416	Kvalsund	Finnmark	1.010	1.739, 28	55, 40	400.000	0, 58
417	Karlsøy	Troms	2.355	1.049, 22	30, 60	405.000	2, 24
418	Alta	Finnmark	19.282	3.653, 36	39, 80	460.000	5, 28
419	Tromsø	Troms	69.116	2.473, 36	45, 50	419.000	27, 94
420	Rindal	Møre og Romsdal	2.088	611, 10	9, 10	426.000	3, 42
421	Rauma	Møre og Romsdal	7.428	1.442, 12	18, 40	435.000	5, 15
422	Lesja	Oppland	2.195	2.168, 91	18, 70	406.000	1, 01
423	Vågå	Oppland	3.739	1.252, 30	23	403.000	2, 99
424	Hol	Buskerud	4.457	1.660, 18	40, 60	390.000	2, 68
425	Ulvik	Hordaland	1.112	670, 13	13, 50	407.000	1, 66
426	Vindafjord	Rogaland	8.447	598, 78	31, 10	477.000	14, 11
427	Hemnes	Nordland	4.585	1.432, 32	21, 80	417.000	3, 20
428	Volda	Møre og Romsdal	8.693	524, 89	26, 50	426.000	16, 56
429	Meløy	Nordland	6.657	798, 45	25, 40	446.000	8, 34

Tabell D.2: Oversikt over data brukt til simulering (fylke).

Fylke	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80-89**	**>=90**
Østfold	31.589	35.093	41.766	36.520	33.250	18.539	11.052	2.160
Akershus	60.578	74.215	90.697	70.969	57.623	31.076	17.628	3.189
Oslo	108.825	116.263	85.737	66.518	52.778	26.401	17.910	4.782
Hedmark	21.366	21.591	27.772	26.714	25.006	14.492	9.432	1.860
Oppland	21.104	21.490	27.250	25.556	23.606	13.872	8.729	1.804
Buskerud	31.313	35.344	39.781	34.427	30.721	16.582	9.951	2.190
Vestfold	27.500	29.364	34.888	31.937	27.324	15.959	9.034	2.035
Telemark	20.107	20.104	24.662	23.175	20.947	11.403	7.500	1.643
Aust-Agder	13.487	14.288	15.741	14.698	12.812	6.815	3.889	935
Vest-Agder	23.074	22.862	24.720	21.343	17.702	10.345	6.112	1.374
Rogaland	61.489	64.159	63.576	52.922	40.528	22.082	12.910	3.092
Hordaland	69.198	68.168	68.406	58.993	49.237	28.309	17.493	4.250
Sogn og Fjordane	12.848	12.036	14.796	14.430	12.134	7.311	4.865	1.269
Møre og Romsdal	31.606	30.891	35.503	34.161	29.332	16.511	11.034	2.539
Sør-Trøndelag	44.474	40.623	42.349	36.273	31.135	17.101	10.460	2.089
Nord-Trøndelag	15.559	14.885	18.839	17.391	16.027	9.179	5.632	1.115
Nordland	28.525	26.659	34.182	32.224	28.548	16.857	10.105	2.208
Troms	20.659	20.225	23.109	20.452	18.023	9.835	5.511	1.036
Finnmark	9.486	8.914	11.233	9.473	8.520	4.656	2.395	371