UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of science and technology

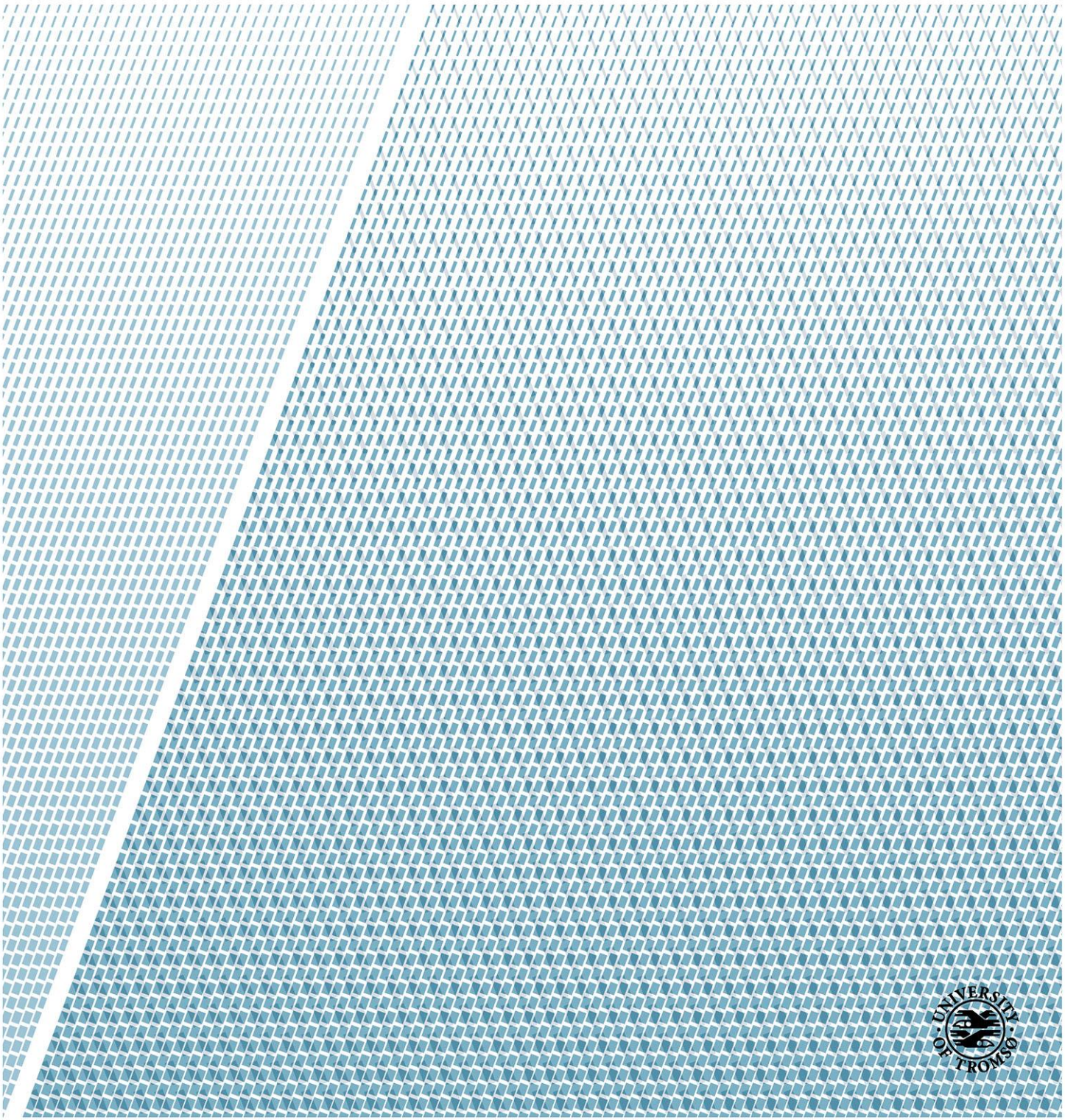
Department of Mathematics and Statistics

Bayesian analysis of temporal and spatial trends of house prices in Norway

—

George Sasha Tendai Mushore

STA-3941: Master's thesis in applied physics and mathematics
September 2018



Abstract

The goal of this thesis is to analyse the temporal and spatial trends of house prices in Norway in a Bayesian setting. We will perform regression analysis of the data which will be modelled using structured additive regression models. This choice was made because structured additive regression models can be put into a computational framework of latent Gaussian models that can be analysed using integrated nested Laplace approximation (INLA). In addition, in a Bayesian setting each of the model parameters have their own posterior distributions from which we can get posterior means and credible intervals.

The main findings were that after applying simple linear regression, new houses have both higher prices and higher price growths than used houses for all counties. Prices in Oslo grow much faster than in any other county. Including a spatially structured effect in the model, large geographical differences between counties were revealed. We conclude that the price differences between counties are reduced, taking the different population sizes into account.

Acknowledgements

I would like to thank my supervisor Sigrunn Sørbye for being patient with me and guiding me through this thesis. You have been motivating and a beacon of hope.

I would also like to thank my mom and my family, for the unconditional love and support. I hope this thesis will make you proud.

A big thanks to my friends at the university of Tromsø and in Oslo for just being my friends. The moments we shared have been a blessing.

Contents

Chapter 1

Introduction

This thesis will introduce and apply Bayesian methodology to analyse housing prices in Norway. We will focus on using Bayesian inference in a spatio-temporal setting. Spatio-temporal models require the use of hierarchical models, see Ghosh et al. (2006) for an introduction. Structured additive regression models (Fahrmeir and Tutz, 2001), which can be used for prediction and analysing relationships between variables, will be introduced. These models can be analysed as three-stage hierarchical models using the computational framework of latent Gaussian models (Rue et al., 2009).

Bayesian inference became popular in the 1990s due to possibilities of using computers to write algorithms for complex models and performing inference for large datasets. This could be done with the help of Gibbs sampling and other Markov chain Monte Carlo (MCMC) methods, see Gilks et al. (1995) for a comprehensive introduction to MCMC-methods. One of the first freely available software for Bayesian computation was Bayesian inference Using Gibbs Sampling (BUGS), launched in 1999 (Lunn et al., 2000).

This software attracted many fields of applications such as epidemiology, astrology, social science, engineering and medicine to Bayesian modelling. MCMC methods allowed for Bayesian analysis of complex hierarchical models. In particular, Bayesian inference is commonly used to analyse time series models, spatial models and a combination of the two, see for example Blangiardo and Cameletti (2015). However, due to the sampling-based nature of MCMC-methods, these can be very time-consuming. In 2009, an alternative to MCMC methods was introduced called integrated nested Laplace approximations (INLA) (Rue et al., 2009). INLA was based on numerical integration and approximation and it greatly improved the computational efficiency in analysing latent Gaussian models.

The data sets used in this thesis are acquired from Statistisk sentralbyrå and is openly available on-line at `http://data.ssb.no/api/v0/dataset/25138?lang=no`. The data set shows average housing prices per square meter in Norway for the years 1999-2017 for 19 counties. It includes a variable that separates the average prices of new house versus second-hand houses. Here the goal is to use both temporal and spatial models to see how the prices develop over time as well as how they differ from location to location. An expected result would be that the counties with the big cities such as Oslo, Bergen in Hordaland and Trondheim in Sør Trøndelag should have some of the highest average housing prices. The number of inhabitants in each county for each year will be included in the analysis and can be used to see whether population can be a factor that explains the variation in prices. We should expect counties with big populations to have higher averages prices as well.

The structure of this thesis is as follows. It will start by introducing Bayesian

inference in general in chapter 2. In chapter 3, we will discuss structured additive regression models and subclasses of these models such as generalized linear models and generalized additive models. We will also describe specific model components used to reflect spatially structured effects and temporal trends. These models are referred to as intrinsic conditional auto-regressive (CAR) models. Chapter 4 will describe how structured additive regression models can be analysed using the computational framework of latent Gaussian models, including the INLA methodology. This chapter also introduces penalized complexity (PC) priors (Simpson et al., 2017) that are used for the precision parameters of the intrinsic CAR models. In chapter 5, we will analyse the data by first using simple linear regression for each county. We also investigate whether the price growth for new and used houses is the same through a test of parallelism. Finally we analyse the data jointly including a spatial effect for each county and also a random effect for the population sizes. This is done for new and used houses separately.

In chapter 6 we give a brief discussion on the work we have done and possible future work. We will also give some concluding remarks.The R-code used in this thesis is given in the appendix.

Chapter 2

Methodology

2.1 Background on Bayesian inference

The two main ways to perform statistical inference include using either a frequentist or a Bayesian approach. Frequentist inference, which has been a widely popular form of statistics from a historical point of view, bases its deduction on the sample data using known experiments (Hoijtink et al., 2008). These experiments are assumed to give the same result if repeated an infinite number of times. The strength of evidence supporting a hypothesis is measured by a p-value or by calculating confidence intervals. Hypothesis testing result in finite conclusions, such as either reject or not rejected, and parameters such as the mean and variance in a frequentist model are fixed. The main difference between frequentist and Bayesian inference is that parameters in a Bayesian setting are not fixed. They are considered to be stochastic variables. Parameters are assigned probability distributions before one knows about the data, and they get updated when more informa-

tion becomes available from the data.

The Bayesian modelling framework can be described in terms of three basic parts which are the likelihood function, the prior and the posterior. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ be unknown parameters and $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the data. Given that $\boldsymbol{\theta}$ and $\boldsymbol{y}$ are random variables and $\pi(.)$ denotes the probability distribution or the density function for a random variable, the likelihood function which is a function of $\boldsymbol{\theta}$ is the sample data's density function,

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i|\boldsymbol{\theta}). \tag{2.1}$$

Here the observations $y_1, \ldots, y_n$ are assumed to be independent given the unknown parameters $\boldsymbol{\theta}$ and therefore the likelihood can be written as the product in the equation. The prior probability distribution or just the prior $\pi(\boldsymbol{\theta})$ gives a subjective belief on $\boldsymbol{\theta}$. It is the first assumption on how the uncertainty of $\boldsymbol{\theta}$ might be. The posterior distribution reflects the uncertainty of the unknown parameter $\boldsymbol{\theta}$ after observing the data $\boldsymbol{y}$. The posterior is defined by

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{2.2}$$

where $\pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})$ represents the joint density of $\boldsymbol{\theta}$ and $\boldsymbol{y}$. The denominator $\int \pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the marginal distribution for $\boldsymbol{y}$. It is a normalizing constant which ensures a proper posterior density. Often, the normalizing constant does not have to be calculated and we can express the posterior as just being proportional to the product of the prior and likelihood,

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta}). \tag{2.3}$$

The posterior represents a compromise between our subjective belief on $\boldsymbol{\theta}$ and the given data from the likelihood function. It is typically used to find summary statistics like the posterior mean, variance and quantiles. It can also be used to find credible regions for $\boldsymbol{\theta}$. The posterior marginals can be used to find credible intervals for the elements of $\boldsymbol{\theta}$. In the univariate case the posterior mean is defined by

$$E(\theta|\boldsymbol{y}) = \int_{-\infty}^{\infty} \theta\pi(\theta|\boldsymbol{y})d\theta \tag{2.4}$$

and the variance is

$$\text{Var}(\theta|\boldsymbol{y}) = E[(\theta - E(\theta|\boldsymbol{y}))^2|\boldsymbol{y}] = \int_{-\infty}^{\infty} (\theta - E(\theta|\boldsymbol{y}))^2\pi(\theta|\boldsymbol{y})d\theta \tag{2.5}$$

Credible intervals specify the range in which a parameter lies between two limits with a given probability. They are comparable to confidence intervals in a frequentist setting. Confidence intervals are given as random variables for fixed parameters and depend only on the data, where as credible intervals are quantiles for the density of the parameter of interest which depend on the data and the prior. We can define a $100(1-\alpha)\%$ credible interval by

$$\int_{c_l}^{c_u} \pi(\theta|\boldsymbol{y})d\theta = 1 - \alpha, \quad \alpha \in (0,1) \tag{2.6}$$

where c_u and c_l are the relevant quantiles of the posterior giving the specified probability. This implies that there exists an infinite number of different credible intervals. The most commonly used credible intervals are the equi-tailed and the highest posterior density (HPD) intervals. In the case of a

equi-tailed credible interval, we choose $c_l = \alpha/2$ and $c_u = 1 - \alpha/2$. The HPD approach finds the sample space of θ that make up a $100(1-\alpha)\%$ interval beginning from the highest point or peak of the density function. This interval is defined by the region

$$R(c) = \{\theta : \pi(\theta|\boldsymbol{y}) \geq c\} \tag{2.7}$$

where c is the largest constant such that

$$\int_{\theta \in R(c)} \pi(\theta|\boldsymbol{y}) = 1 - \alpha \tag{2.8}$$

The HPD and the equi-tailed intervals are equal when the posterior density function is symmetric. In general the HPD-interval is optimal in the sense that it has the shortest length of all credible intervals. To introduce these concepts we will take a look at a simple example.

Example 1: Let $Y \sim \text{bin}(n, \theta)$ where n is the number of experiments, while $\theta \in [0, 1]$ represents the success probability in Bernoulli trials. We assign a Beta(α, β) prior to θ, where the shape parameters α and β are considered to be fixed i.e.

$$\pi(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \alpha, \beta > 0. \tag{2.9}$$

The likelihood is

$$\pi(y|\theta) \propto \theta^y (1-\theta)^{n-y}. \tag{2.10}$$

To find the posterior we use equation (2.3).

$$\begin{aligned}\pi(\theta|y,\alpha,\beta) &\propto \theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{\beta-1+n-y}\end{aligned} \tag{2.11}$$

We can see that the posterior becomes a Beta distribution with $\alpha^* = \alpha + y$ and $\beta^* = \beta + n - y$ giving the posterior $\pi(\theta|y,\alpha,\beta) = \text{Beta}(\alpha^*, \beta^*)$. To find the posterior mean we can just use the known mean for Beta distributions which gives

$$E(\theta|y,\alpha,\beta) = \int_0^1 \theta\pi(\theta|y,\alpha,\beta) = \frac{\alpha^*}{\alpha^*+\beta^*} = \frac{\alpha+y}{\alpha+\beta+n}. \tag{2.12}$$

The result can be written as

$$\begin{aligned}\frac{\alpha+y}{\alpha+\beta+n} &= \frac{\alpha+\beta}{\alpha+\beta+n}\cdot\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\cdot\frac{y}{n} \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\cdot\hat{\theta}_{apriori} + \frac{n}{\alpha+\beta+n}\cdot\hat{\theta}_{MLE}\end{aligned} \tag{2.13}$$

where $\hat{\theta}_{apriori}$ is the prior estimate and $\hat{\theta}_{MLE}$ is the maximum likelihood estimate of the success. This probability shows that the posterior is a weight of the two. When n gets large the weight of the prior estimate gets smaller. This tells us that the prior's influence on the posterior is minimal when we have a lot of data and the choice of prior is important when we have little data. The variance can be found by

$$\text{Var}(\theta|y,\alpha,\beta) = \frac{\alpha^*\beta^*}{(\alpha^*+\beta^*)^2(\alpha^*+\beta^*+1)} = \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}. \tag{2.14}$$

To find the credible intervals we can find the quantiles of the posterior distribution. To do so we have to first give values for n, y α and β. Table 2.1 shows how the posterior mean and variance including the credible intervals of θ would look like for different values of α and β at $n = 15$ $y = 10$ successes. The different parameters change the density function a lot. In figure 2.1 we have chosen $n = 15, y = 10, \alpha = 2, \beta = 3$ to illustrate how the credible intervals would look like for the HPD-interval (blue) and the equi-tailed (red).

Prior	$E(\theta\|y,\alpha,\beta)$	$Var(\theta\|y,\alpha,\beta)$	CI_l	CI_u	HPD_l	HPD_u
$\alpha = 1, \beta = 2$	0.611	0.013	0.383	0.816	0.392	0.823
$\alpha = 4, \beta = 1$	0.700	0.010	0.488	0.874	0.503	0.886
$\alpha = 2, \beta = 6$	0.522	0.010	0.322	0.718	0.323	0.719
$\alpha = 1, \beta = 8$	0.458	0.010	0.268	0.655	0.266	0.653

Table 2.1: A list of differents values for the posterior mean and variance using different prior parameters. The list also includes the 95% equi-tailed credible intervals and the corresponding HPD-intervals.

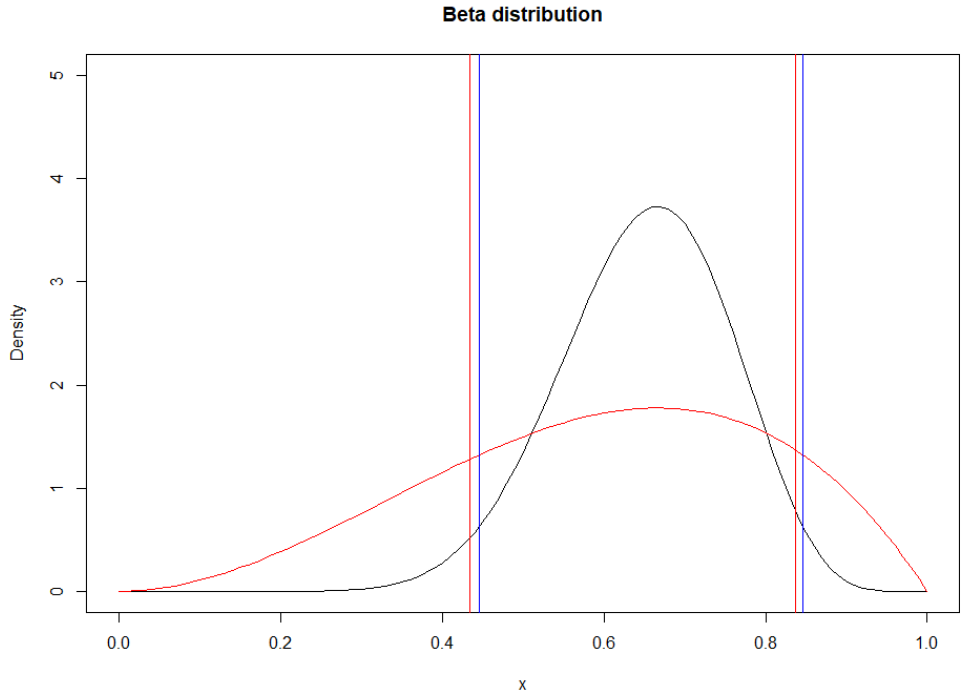


Figure 2.1: The figure shows plot of the beta(2,3) prior (red) and the resulting posterior density and its credible intervals where blue is the HPD-interval and red is the 95% equi-tailed intervals.

.

This example is a very simple example where we have a conjugate prior. This means that the posterior will have the same distribution as the prior, just with different parameters.

2.2 Issues in performing Bayesian inference

In general calculation of the posterior can be very challenging as this cannot be expressed in an analytical form. We therefore have to turn to approximation methods. A commonly applied class of methods is MCMC. These methods provide algorithms to generate irreducible and aperiodic Markov chains which can be regarded as a sample from a stationary target posterior distribution. The longer the generated chain is, the closer the sampling distribution gets to being an exact approximation of the target distribution. Subclasses of MCMC methods include Gibbs sampling, rejection sampling, the Metroplis-Hastings algorithm and others (Givens and Hoeting, 2012). An alternative to MCMC methods is using INLA which uses numerical approximations and integration to find the posterior marginals. In this thesis we will use the INLA methodology and details will be given in Section 4.2.

Another issue in Bayesian inference is to choose prior distributions. The choice of priors depends on if one wants the prior to be informative or non-informative (Gelman et al., 2003). An informative prior influences a parameter by assuming some information of the parameter. An example of an informative prior is assigning a normal prior with a small variance. This is a conjugate prior for data the have a normal distribution. However, if the

data is not normal the posterior might come out as leaning towards the prior therefore giving a wrong reflection of the data. The idea of non-informative priors is to let the data speak for itself such that the inference is not affected much by the prior. Non-informative priors can be difficult to create. A popular class of non-informative priors is Jeffreys' priors (Jeffreys, 1946). These priors are invariant to transformations. This means that if $\pi(\boldsymbol{\theta})$ is a prior for $\boldsymbol{\theta}$, then $\pi(f(\boldsymbol{\theta})$ is a prior for $f(\boldsymbol{\theta})$ (Jeffreys, 1946). In this thesis we will apply a recently suggested class of priors called penalised complexity (PC) priors (Simpson et al., 2017). These are weakly informative and will be described in Section 4.3.

Chapter 3

Structured additive regression models

In this thesis, we will focus on performing Bayesian inference for specific regression models. These models can be seen as subclasses of general structured additive regression models. This class of models is very flexible and includes among others, the linear regression models, generalized linear model and generalized additive model. Also this class of models can be used for time series and spatial analysis.

3.1 Subclasses of structured additive regression models

3.1.1 Linear regression

Linear regression is a popular statistical tool in data analysis. It assumes a linear relationship between the response and the predictor variables. Such a model is described as

$$Y_i = \alpha + \sum_{m=1}^{M} \beta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent and normal distributed with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \tau^{-1}$. We will use a precision parameter τ instead of variance σ^2 and they are related as $\tau = 1/\sigma^2$. For fixed covariates the mean is described as

$$\mu_i = E(Y_i|\alpha, \beta_1, \ldots, \beta_m, z_{i1}, \ldots, z_{iM}) = \alpha + \sum_{m=1}^{M} \beta_m z_{im}, \quad i = 1, \ldots, n \tag{3.2}$$

which implies that the response variable $Y_i \sim N(\mu_i, \tau^{-1})$, α is the intercept and β_m is a regression parameter giving the linear effect of the predictor variable z_m. When $m = 1$ we have a simple linear regression model which has only one predictor variable. We will use this model in Section 5.1.

3.1.2 Generalized linear models

In a generalized linear model (GLM), which was introduced in Nelder and Wedderburn (1972), we extend the linear models so that the response can be drawn from other distributions than the Gaussian. GLMs have a general link between the response and predictor. This makes GLMs a broad class which includes for example models for binary data, categorical data, log-linear data or data from many well-known distributions. GLMs can be specified in stages:

1. The linear predictor is defined as $\eta_i = \sum_{j=1}^{J} \beta_j z_{ji}$ where β_j measures the linear effect of the covariates z_j.

2. The GLM uses a link function $g(.)$ to relate the linear functions of the predictors to the mean of the response variable,

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) \tag{3.3}$$

   where η_i is the linear predictor. Examples of different link functions include: the logit link $g(\mu) = \log(\frac{\mu}{1-\mu})$, log link $g(\mu) = \log(\mu)$, and the identity link $g(\mu) = \mu$.

3. The response Y_i is assumed to be drawn from the exponential family and the density is defined as

$$\pi(y|\theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(Y, \phi)\right\}, \tag{3.4}$$

   where θ is the natural parameter which is related to the mean of the

distribution and ϕ is the dispersion parameter. $b(\theta), a(\theta)$ and $c(y, \phi)$ are given functions. Many well-known distributions are included in the exponential family such as the Poisson, the binomial, the Gaussian and the gamma distribution.

3.1.3 Generalized additive models

Generalized additive models (GAM) are an extension of GLMs in which the predictor is modelled using the linear dependence of smooth functions of the predictor variables (Hastie and Tibshirani, 1990). The additive form of the model is described as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} f_k(c_{ki}), \tag{3.5}$$

where f_k are non-parametric functions called smooth functions. These can be of many types, but the most common ones are splines such as cubic regression splines, thin plate regression splines and p-splines. Inferences can be made about these smooth functions. GAMs have the same properties as GLMs, but are a broader model class since smooth functions are a more flexible category.

3.2 Structured additive regression models in general

Structured additive regression models make up a flexible class of regression models introduced in Fahrmeir and Tutz (2001). This class provides a unified and flexible framework for a wide range of models including the well estab-

lished models mentioned in Section 3.1. The distribution of the response variable is still assumed to belong to the exponential family and the mean of the response variable is linked to a structured additive predictor η_i. Following Rue et al. (2017) the structured additive predictor η_i is defined as

$$\eta_i = \alpha + \sum_{j=1}^{J} \beta_j z_{ji} + \sum_{k=1}^{K} f_k(c_{ki}), \quad i = 1, \ldots, n. \tag{3.6}$$

The predictor includes linear effects in the first sum like in a GLM. In addition, the predictor includes smooth effects of covariates like in GAMs. However, the function effects of covariates f_k in structured additive regression models are not restricted to smooth models. These can also include time trends and seasonal effects making it possible to analyse time series. Also the functions f_k can denote spatially correlated random effects used for example in geographically weighted regression. Simple linear regression is a special case of structured additive models where $g(.)$ is an identity link, $K = 0$ and $J = 1$.

3.3 CAR-models

A Gaussian Markov random field (GMRF) is a random vector with a multivariate Gaussian distribution. What characterises GMRF is that it has Markov properties which imply conditional independence between its variables. Formally, a GMRF is defined by a vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ with the distribution

$$\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}). \tag{3.7}$$

This vector can be defined on a graph with nodes and edges, where the nodes represent the variables x_i and the edges give the relationship between neighbouring variables. We say that a graph is connected when all nodes connect to at least one other node. Due to Markov properties the precision matrix $\boldsymbol{Q}$ will typically be sparse. GMRFs are specified by the precision matrix that can be expressed as $\boldsymbol{Q} = \tau \boldsymbol{R}$ where τ is the random precision parameter and $\boldsymbol{R}$ is a matrix that reflects the neighbourhood structure of the graph. GMRFs can also be formulated as conditional auto-regression (CAR) models described in Besag and Kooperberg (1995). They were introduced as a way to account for spatial correlation between regions in spatial models, and have been extended to a broader usage in statistics (Rue and Held, 2005). A version of GMRFs called intrinsic Gaussian Markov random field (IGMRF) is specified as

$$\pi(x) = (2\pi)^{-(n-k)/2}(|\boldsymbol{Q}|^*)^{1/2}\exp\{-\frac{1}{2}\boldsymbol{x'Qx}\}, \tag{3.8}$$

where $\boldsymbol{Q}$ is an $n \times n$ precision matrix with rank $n-k$. The vector $\boldsymbol{x}$ is then an improper GMRF in which we use additional constraints to get a proper model.

In this thesis we will use two examples of IGMRFs, also referred to as ICAR models. To model a smooth function we will use a second order random walk. This model is defined by having independent second-order increments:

$$\Delta^2 x_i = x_i - 2x_{i+1} + x_{i+2} \sim N(0, \tau^{-1}) \tag{3.9}$$

such that the density becomes

$$\begin{aligned}\pi(\boldsymbol{x}|\tau) &\propto \tau^{(n-2)/2}\exp(-\frac{\tau}{2}\sum(\Delta^2 x_i)^2) \\ &= \tau^{(n-2)/2}\exp(-\frac{\tau}{2}\boldsymbol{x'Rx})\end{aligned} \tag{3.10}$$

where $\boldsymbol{R}$ has the bandwidth 5. This model will capture local deviation from a line.

The other IGMRF that we will use will account for a spatially structured effect. The graph of the model represents the spatial neighbourhood of an area. This model is defined as

$$x_i|x_j,\tau \sim N(\frac{1}{n_i}\sum_{i\sim j} x_i, \frac{1}{n_i\tau}), \quad i \neq j \tag{3.11}$$

where n_i is the number of neighbours of node i. The neighbourhood of node i is denoted by $i \sim j$ and τ is the precision parameter which determines the smoothness of the estimated effects. The mean of x_i accounts for the overall neighbourly effect, where the precision is proportional to the number of neighbours. The density is then defined as

$$\pi(\boldsymbol{x}|\tau) \propto \tau^{(n-1)/2}\exp(-\frac{\tau}{2}\sum_{i\sim j} w_{ij}(x_i - x_j)^2), \tag{3.12}$$

where w_{ij} are the weights for all pairs of adjacent nodes. This model is also referred to as the Besag model. When interpreting our model, we are interested in how much the effects vary from the mean value which is chosen equal to zero. The precision matrix needs to be scaled so that when the marginal variance is 1 the precision parameter τ has a unified interpretation.

(Sørbye and Rue, 2014).

Chapter 4

The computational framework

In this chapter we will describe the computational framework used to perform Bayesian inference on the structured additive regression models that have been described. We will describe the INLA methodology and the class of PC priors that is used in this thesis.

4.1 Latent Gaussian models

Structured additive regression models can be analysed in a unified way using the computational framework of latent Gaussian models. Latent Gaussian models are a hierarchical model that have three layers. These models are useful to model simple as well as complex models with multiple parameters. Joint probability models are required and we need to infer the relationships that may exist between these parameters. The first layer in the hierarchical

model is the prior described as

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) \tag{4.1}$$

where $\boldsymbol{\theta}$ are the hyper-parameters. These hyper-parameters can be for example a variance, correlation parameter or an auto-regression coefficient. The number of hyper-parameters is typically assumed to be small.

The second layer is the latent Gaussian field described as

$$\boldsymbol{x}|\boldsymbol{\theta} \sim N(0, \boldsymbol{Q}(\boldsymbol{\theta})^{-1}) \tag{4.2}$$

where the unobserved variables $\boldsymbol{x}$ describe the latent dependency structure of the data. The latent field given the hyper-parameters are multivariate-normal. Especially, the latent field $\boldsymbol{x}$ is assumed to be a GMRF and might have a large dimension. It is important to note that all parameters in the structured additive model in equation (3.6) can be placed into a latent field $\boldsymbol{x}$ so that it becomes $\boldsymbol{x} = \{\alpha, \boldsymbol{\beta}, \{f_i(.)\}, \boldsymbol{\eta}\}$.

The third layer is the likelihood described as

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} \sim \prod_i \pi(y_i|x_i, \boldsymbol{\theta}) \tag{4.3}$$

where the observations $\boldsymbol{y}$ are assumed to be conditionally independent, given $\boldsymbol{\theta}$ and $\boldsymbol{x}$.

Combining the layers together the joint posterior density of latent variables

$\boldsymbol{x}$ and the hyper-parameters $\boldsymbol{\theta}$ is obtained:

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_i \pi(y_i|x_i, \boldsymbol{\theta}) \tag{4.4}$$

where we want to estimate marginal distributions from the joint distribution by intergration. Both MCMC methods and INLA can be used to approximate the marginals and in this thesis we will use the INLA-methodology which will be described in the next section.

4.2 INLA

Integrated nested Laplace approximations (INLA) is a method used to analyse latent Gaussian models as an alternative to inference with MCMC. INLA's main advantage is it's computational speed compared to MCMC methods. The idea is to estimate the marginals of the hyper-parameters and the latent field of the LGMs through Laplace approximations, and take advantage of numerical algorithms for sparse matrices.

The main aim in analysing LGMs is to estimate the marginals for each hyper-parameter θ_j and each component of the latent field x_i. These marginals can be written as

$$\pi(\theta_j|\boldsymbol{y} = \int \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j} \quad j = 1, \ldots, |\boldsymbol{\theta}| \tag{4.5}$$

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \quad i = 1, \ldots, n \tag{4.6}$$

The INLA methodology achieves this by several computational steps. The first step is to find a numerical approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ in (4.5). To do this

a Laplace approximation for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is used given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\bigg|_{x=x^*(\theta)} = \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}), \tag{4.7}$$

where $x^*(\theta)$ is the mode. The denominator can be rewritten as a Gaussian approximation

$$\begin{aligned}\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}) &\propto \exp(-\frac{1}{2}\boldsymbol{x'Q(\theta)x} + \sum \log\pi(y_i|x_i,\boldsymbol{\theta})) \\ &= (2\pi)^{n/2}|\boldsymbol{P(\theta)}|^{\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu(\theta)})'\boldsymbol{P(\theta)}(\boldsymbol{x}-\boldsymbol{\mu(\theta)}))\end{aligned} \tag{4.8}$$

where $\boldsymbol{P(\theta)} = \boldsymbol{Q(\theta)} + \text{diag}c(\boldsymbol{\theta})$ and $\boldsymbol{\mu(\theta)}$ is the location of the mode. $c(\boldsymbol{\theta})$ is a vector with the negative second derivatives of the log-likelihood of x_i at the mode. This form is used for computer efficiency. The Laplace approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ can now be numerically integrated at a low computational cost to find the marginal posterior of the hyper-parameter of interest.

The next step is to find the approximation of the latent field $\pi(x_i|\boldsymbol{y})$. It requires to find approximations for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$ from the intergral in (4.6). For the first approximation it has already been done in (4.7) and for the latter the standard method is to use the simplified Laplace approximation. To do this we fit a skew-normal density to a Taylor series expansion of the Laplace approximation.

$$\log\tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y}) = bx_i - \frac{1}{2}x_i^2 + \frac{1}{6}dx_i^3 +, \ldots. \tag{4.9}$$

Two other alternative methods to simplified Laplace approximation are Gaussian approximations or Laplace approximations. Now to find the marginals

for the components of the latent field, the approximations for $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ can be numerically integrated with respect to $\boldsymbol{\theta}$

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}^k, \boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}^k|\boldsymbol{y})\Delta\boldsymbol{\theta}^k \tag{4.10}$$

where $\Delta\boldsymbol{\theta}^k$ denotes area-weights that sum over values of $\boldsymbol{\theta}$

4.3 PC priors

In Simpson et al. (2017) a unified approach for constructing weakly informative priors for different hyper-parameters was introduced called penalized complexity (PC) priors. They are invariant to reparameterisations. These priors are computed based on four principles

1. Occam's razor says that a model should be kept simple until there is enough support for a complex model. A flexible model can be defined as

   $$f = \pi(\boldsymbol{x}|\xi) \tag{4.11}$$

   where ξ is the flexibility parameter. f is a flexible version of a base model

   $$g = \pi(\boldsymbol{x}|\xi = \xi_0). \tag{4.12}$$

   An example is the Student T distribution, where its base model is the normal distribution and its flexibility parameter is the degrees of freedom.

2. The Kullback-Leibler divergence (KLD) can be used to measure the

complexity of model and is defined as:

$$KLD(f||g) = \int f(x)log(\frac{f(x)}{g(x)})dx \tag{4.13}$$

where $g(x)$ is the base model of the prior $f(x)$.

3. The 3rd principle assigns a prior to the measure of complexity which penalizes deviation from the base model. This measure is given by the unidirectional distance $d(f||g) = \sqrt{\text{KLD}(f||g)}$ which measures the complexity of the model $f(x)$ when compared to the base model $g(x)$. The distance is a assigned an exponential prior

$$\pi(d(\xi)) = \lambda exp(-\lambda(d(\xi))), \quad \lambda > 0 \tag{4.14}$$

The mode at $d = 0$ is the base model. The prior for the parameter of interest can be found through the transformation.

$$\pi(\xi) = \lambda exp(-\lambda d(\xi))|\frac{\delta d(\xi)}{\delta \xi}| \tag{4.15}$$

4. User-defined scaling: Determining λ is based on the user knowledge of the model. λ can be selected by adjusting the broadness of the tail by the probability statement

$$Prob(Q(\xi) > U) = \alpha, \tag{4.16}$$

where $Q(\xi)$ is a transformation of the flexibility parameter and U specifies the upper limit of the standard deviation. α is a small probability.

The user-defined scaling influences how informative a PC prior becomes and the magnitude of random effects.

In this thesis we will use PC priors on the CAR models which have a precision parameter τ. The prior for these models is defined using $\xi = 1/\tau$. The base of this model is given by $\xi = 0$. A criterion for IGMRFs is allowing the the transformation of the prior to be $Q(\xi) = \frac{1}{\sqrt{\tau}}$ such that

$$P(\frac{1}{\sqrt{\tau}} > U) = \alpha. \tag{4.17}$$

We will choose that $U = 1$ and $\alpha = 0.01$

Chapter 5

Application: Analysis of housing prices in Norway

A main aim of this chapter is to apply Bayesian inference to real data. The data represents the average price for houses in Norway per square metre for every county. The data is measured annually from 1999 until 2017. There are 19 counties in total. The data also distinguishes between new houses and second hand houses. Methods chosen for inference include simple linear regression and spatial analysis using the CAR model and the random walk model of the second order. All of the analysis will be done with the programming software and language R.

5.1 Simple linear regression

In this section we fit a simple linear regression model to the house prices for each county for the period 1999-2017. We present years as $\boldsymbol{z} = (z_1, \ldots, z_n)$

and the housing prices as $\boldsymbol{Y} = (y_1, \ldots, y_n)$ and we assume the prices to be normal distributed as

$$Y_i \sim N(\mu_i, \tau^{-1}) \tag{5.1}$$

The linear predictor is given as

$$E(Y_i|\alpha, \beta, z_i) = \alpha + \beta z_i \tag{5.2}$$

We want to estimate the parameters α and β. We assign normal priors such that $\alpha \sim N(0, 0.001)$, $\beta \sim N(0, 0.001)$, and a gamma prior for τ^{-1} such that $log(\tau) \sim \text{logGamma}(1, 5 \cdot 10^{-5})$.

5.1.1 Results of the simple linear regression

To summarize the results we have listed the posterior means and standard deviations of each county and type of house in table 5.1. In 1999 the average prices where lowest in Sogn og Fjordane for both new houses and used houses, and they were highest in Akershus and Oslo for both new and used houses. Oslo's slope parameter is twice as steep as almost all of the other counties with square meter prices increasing at almost 3000kr every year for new houses and 2200kr for used houses. The standard deviations for the regression parameter are largest for Oslo. Figures 5.1, 5.2 and 5.3 show the estimated mean plotted against the data. We can see that the points follow the line quite well. In general the deviations between line and the observation points are very smal implying that the increase in prices during the given time period is well explained by a linear trend.

	α.new	β.new	α.sd.new	β.sd.new	α.used	β.used	α.sd.used	β.sd.used
Østfold	7.215	1.304	0.394	0.035	6.259	0.799	0.350	0.031
Akershus	**8.197**	1.751	0.493	0.043	8.716	1.252	0.515	0.045
Aust-Agder	6.531	1.248	0.550	0.048	6.075	0.736	0.421	0.037
Buskerud	7.088	1.459	0.457	0.040	6.131	0.922	0.340	0.030
Finnmark	6.938	1.087	0.466	0.041	5.378	0.803	0.639	0.056
Hedmark	7.453	1.220	0.388	0.034	5.395	0.610	0.305	0.027
Hordaland	5.842	1.462	0.392	0.034	7.119	1.074	0.571	0.050
Møre og Romsdal	5.993	1.317	0.415	0.036	5.082	0.709	0.230	0.020
Nord-Trøndelag	6.171	1.226	0.596	0.052	3.922	0.654	0.180	0.016
Nordland	5.597	1.415	0.670	0.059	4.847	0.764	0.242	0.021
Oppland	6.210	1.245	0.453	0.040	5.327	0.637	0.228	0.020
Oslo	5.799	**2.931**	**1.220**	**0.107**	**10.238**	**2.209**	**0.915**	**0.080**
Rogaland	5.594	1.605	0.631	0.055	6.757	1.081	0.888	0.078
Sør-Trøndelag	6.905	1.356	0.553	0.048	6.304	1.007	0.378	0.033
Sogn og Fjordane	5.271	1.219	0.524	0.046	3.732	0.761	0.326	0.029
Telemark	6.277	1.242	0.507	0.044	5.463	0.656	0.299	0.026
Troms	5.822	1.527	0.777	0.068	6.679	0.954	0.567	0.050
Vest-Agder	6.668	1.311	0.718	0.063	6.630	0.808	0.648	0.057
Vestfold	7.691	1.498	0.371	0.033	7.386	0.862	0.311	0.027

Table 5.1: A table showing the posterior mean and standard deviation of the parameters α and β for each county and each type of house. The values are given in thousands of kr.

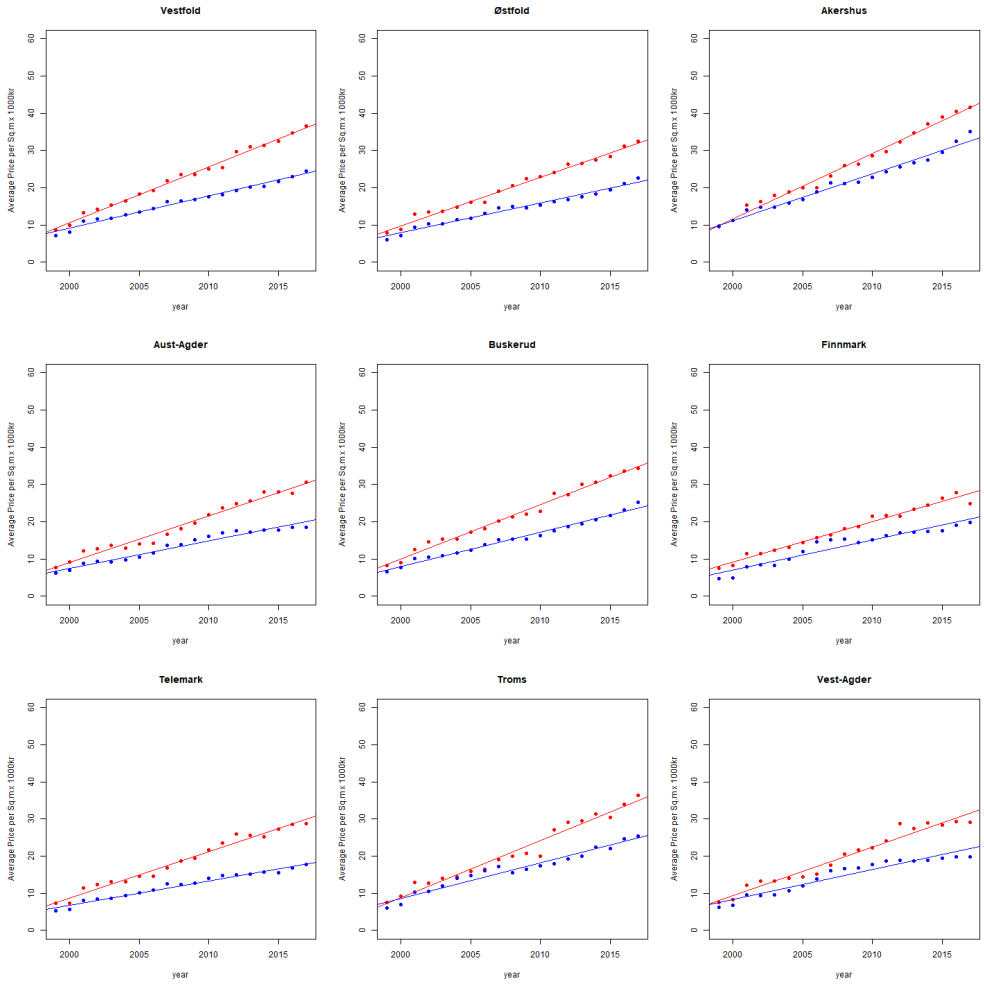


Figure 5.1: Estimated annual square meter prices for used and new houses, where red is the new and blue is the used

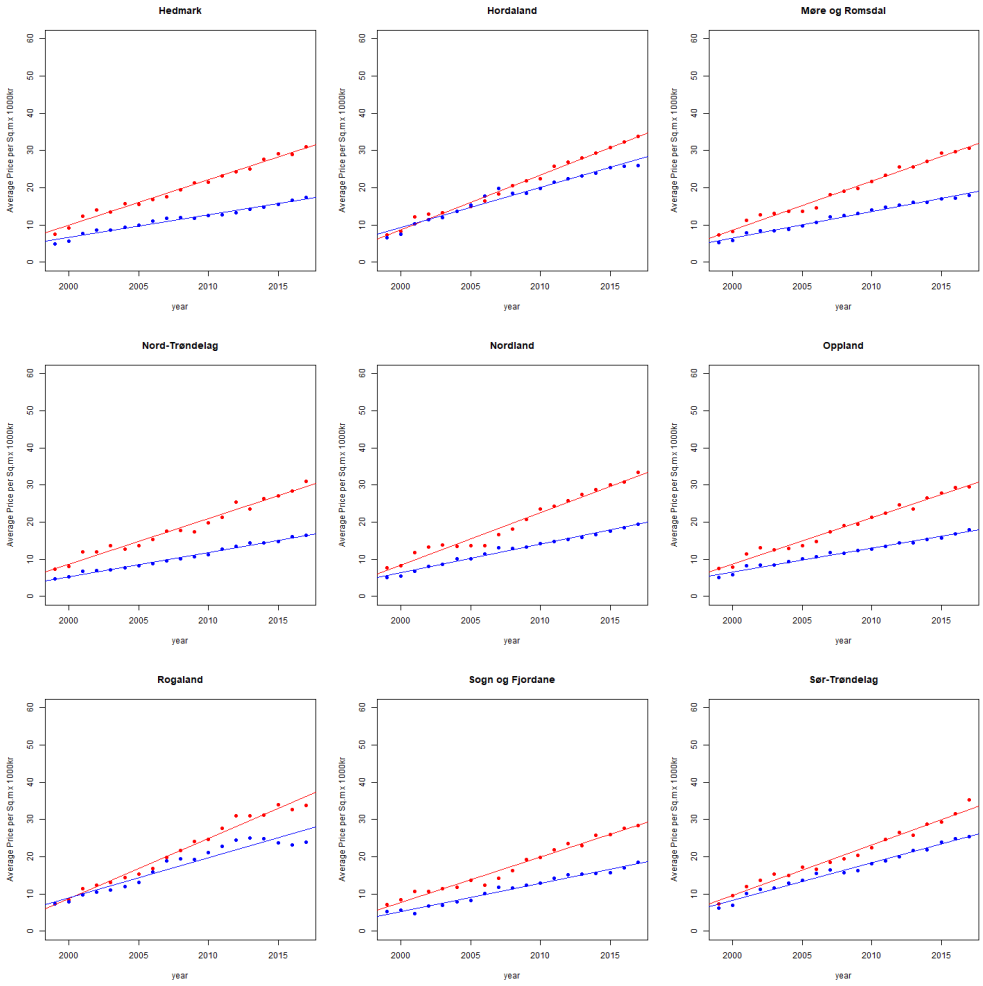


Figure 5.2: Estimated annual square meter prices for used and new houses, where red is the new and blue is the used

Figure 5.4 shows the posterior marginals for the intercept α and the slope β parameters for both new and used houses in Oslo. The 95% credible intervals for the for the new houses are $(2072kr, 3140kr)$ and for used houses are $(2050kr, 2370kr)$. Take note that Oslo has the largest variance, so that means the estimates for the other counties give much more narrow credible intervals.

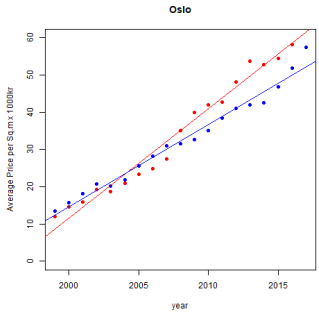


Figure 5.3: Estimated annual square meter prices for used and new houses, where red is the new and blue is the used

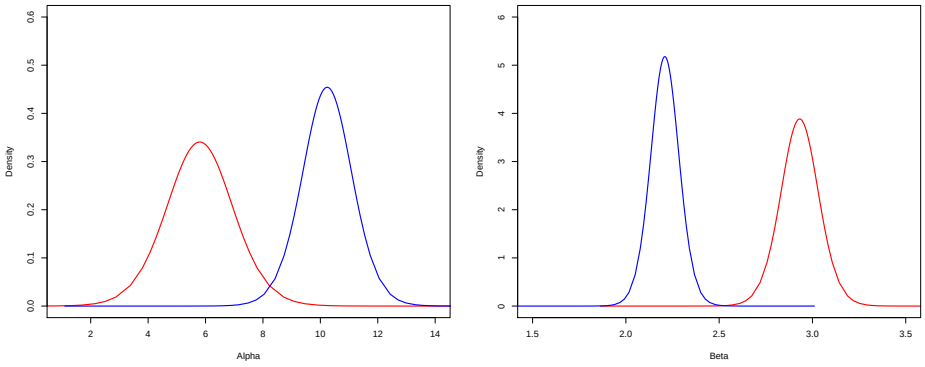


Figure 5.4: The posterior marginals for the parameters α and β for new houses(red) and used houses(blue) in Oslo

5.1.2 Test of parallelism

We are interested in whether the slopes of new houses and for used houses are parallel. Parallel slopes means that the parameter β is the same for new and used. One way to check if this is true is to create a model with categorical variables. A categorical variable d_i is a variable that helps distinguish between two categories. In our case our category variable distinguishes between new and used houses by assigning 0 as an indicator for new houses and 1 as the indicator for used houses. This model can be defined as

$$Y_i = \alpha + \beta_1 z_i + \beta_2 d_i + \beta_3 z_i d_i + \epsilon_i \tag{5.3}$$

To implement his model we stall all the prices for new and used houses in one vector. We also make corresponding vectors for the years and the categorical variable. Each vector has the length of 38. When $d = 1$ we should get

$$E(Y_i|\alpha, \beta_0, \beta_1, \beta_2, \beta_3) = (\alpha + \beta_2) + (\beta_1 + \beta_3)z_i \tag{5.4}$$

and when $d = 0$ we get

$$E(Y_i|\alpha, \beta_0, \beta_1, \beta_2, \beta_3) = \alpha + \beta_1 z_i. \tag{5.5}$$

We assign normal priors of $N(0, 1000)$ to $\alpha, \beta_0, \beta_1, \beta_2$ and β_3. We assign a gamma prior for τ, $log(\tau) \sim \text{logGamma}(1, 5 \cdot 10^{-5})$. We can then apply INLA to find the posterior for our parameters. The parameter of interest is β_3. We can see that if $\beta_3 = 0$, the slope will be the same for both categories. To

perform the tests, we use the 95% credible intervals for β_3 and check whether the contain 0. The tests concludes that none of the slopes are parallel. This means that prices for new houses do not have the same growth rate as the used houses in any of the counties.

5.2 Introducing a spatial effect in the model

Simple linear regression is a good way to detect linear trends in the prices for each individual county. In this section we introduce a spatial model component in the linear regression model. This allows us to analyse all counties simultaneously. To infer spatial modelling we created a connected graph of all the counties. This is achieved by numbering all the counties, and then for each county specify all neighbouring counties. We will define two models. In the first model the linear predictor is defined by

$$\eta_i = \alpha + \beta z_i + f(c_i) \tag{5.6}$$

where we have assumed an identity link. This means that $E(Y_i) = \eta_i$. β represents the linear effect of the years z_i. The function $f(\cdot)$ is an intrinsic CAR model of the first order which represents the spatially structured random effects of the connected graph of the counties described in equation (3.12). The estimated spatial random effect for each county can be positive, 0 or negative. The effects from all the counties sum up to 0. This implies that counties with a positive effect have a larger overall increase in prices than what can be explained by a linear trend for all counties.

5.2.1 Results

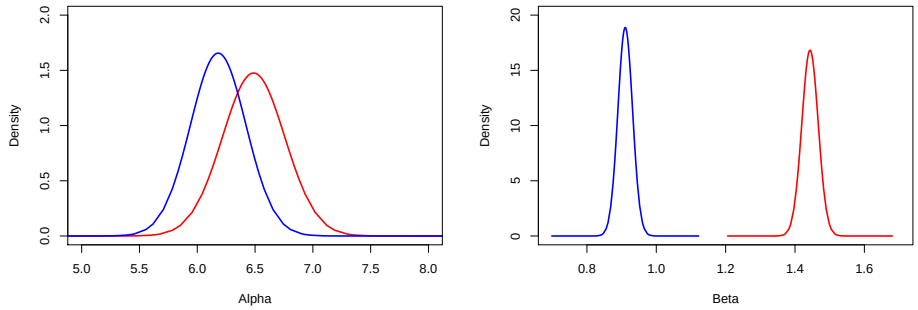


Figure 5.5: Parameters for the linear trend for all counties for new (red)and used (blue) houses

The linear trend of all the counties is given by the new estimates for α and β. The posterior marginals for these parameters are seen in figure 5.5. The parameters β for new and used houses have small variance compared to the α alpha parameters. The posterior mean of of β_{new} is $1443kr$ with a 95% credible intervals $(1397kr, 1490kr)$.The posterior mean of of $\beta_{used} = 910kr$ with credible intervals $(869kr, 952kr)$. The parameter α_{new} has the credible intervals $(5995kr, 7019kr)$ with mean 6487kr. α_{used} has a mean of $6181kr$ with a credible interval $(5707kr, 6655kr)$.

To show the estimated effects of the spatial model, we have plotted this using a map of Norway with the different counties. Figure 5.6 shows the posterior mean of the random spatial effects of each county from the intrinsic CAR model component in equation (5.6). Red is the largest value and dark blue is the lowest value. It is important to note that the two maps do not have

the same scale, so similar colors do not mean the same value. The values for random effects cannot be quantified in a meaningful way except for the fact that the sum of all random effects is 0. Most of Norway is different shades of blue because Oslo drastically stands out with its high prices and raises the mean value. In the northern part of Norway the lightest shade of blue is Troms which has the one of the large cities in Norway, Tromsø. We also notice that other counties with large cities are coloured with a light shade of blue which implies that the prices are high in these counties. Figure 5.7 shows the same values as the maps but just as a bar-plot for better visual understanding. It is easier to see which counties have negative or positive effects. Sør Trøndelag, and Hordaland have negative effects for new houses and positive effects for used houses.

Table 5.2 displays the 95% credible intervals for the estimated spatially structured effects. Credible intervals that do not include 0 represent counties that have prices that are significantly different from the linear trend based on all counties. Aust-Agder, Finnmark, Hedmark, Møre og Romsdal, Nord-Trøndelag, Nordland, Oppland, Sogn og Fjordane and Telemark have significantly lower prices for both types of houses. Akerhus and Oslo have significantly higher prices for both types of houses, whereas Hordaland and Rogaland have significantly higher prices only for used houses and Vestfold for just new houses.

In figure 5.8 we see the posterior marginals for the precision parameter of theintrinsic CAR model component acquired from (4.5). The posterior marginals show a slimmer density and smaller precision for used houses than for new houses.

	$CI.new_l$	$CI.new_u$	$CI.used_l$	$CI.used_u$
Østfold	-1.632	0.515	-1.934	-0.011
Akershus	3.498	5.566	4.803	6.695
Aust-Agder	-2.898	-0.798	-2.762	-0.858
Buskerud	-0.122	1.897	-0.729	1.140
Finnmark	-4.115	-1.969	-2.808	-0.885
Hedmark	-2.201	-0.100	-4.604	-2.697
Hordaland	-1.540	0.538	1.481	3.378
Møre og Romsdal	-2.817	-0.718	-4.041	-2.138
Nord-Trøndelag	-3.483	-1.360	-5.706	-3.792
Nordland	-2.238	-0.116	-3.736	-1.823
Oppland	-3.053	-1.007	-4.361	-2.480
Oslo	12.632	14.799	15.802	17.730
Rogaland	-0.483	1.600	1.217	3.114
Sør-Trøndelag	-1.607	0.473	-0.041	1.859
Sogn og Fjordane	-4.297	-2.211	-4.778	-2.882
Telemark	-3.029	-0.963	-4.021	-2.130
Troms	-0.984	1.140	-0.083	1.831
Vest-Agder	-2.181	-0.059	-1.522	0.391
Vestfold	0.589	2.712	-0.275	1.639

Table 5.2: 95% Credible intervals for the posterior mean of the intrinsic CAR model component for both types of houses

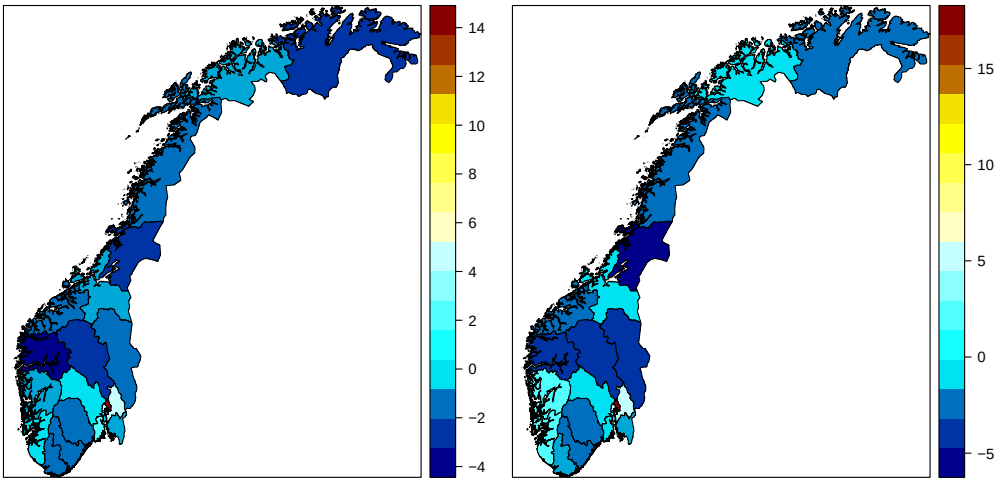


Figure 5.6: Spatial effect for prices in Norway for new houses in the left and used on the right.

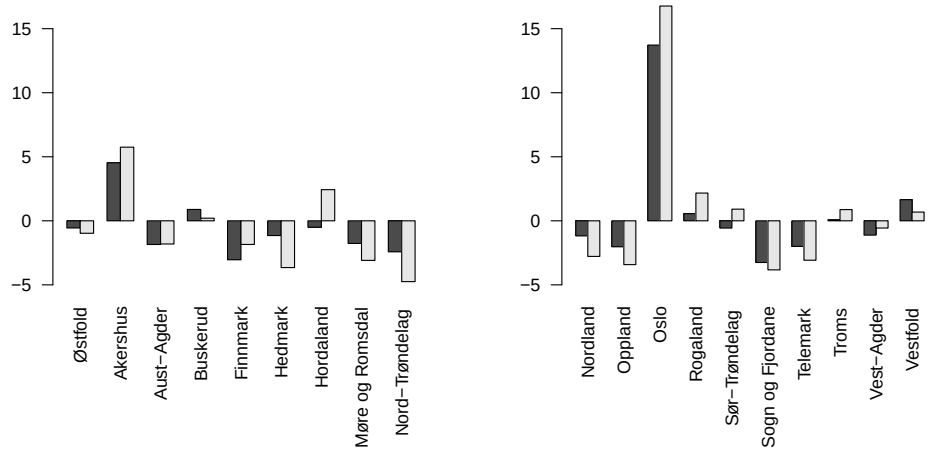


Figure 5.7: Barplot that shows how different the same counties are for new and houses.

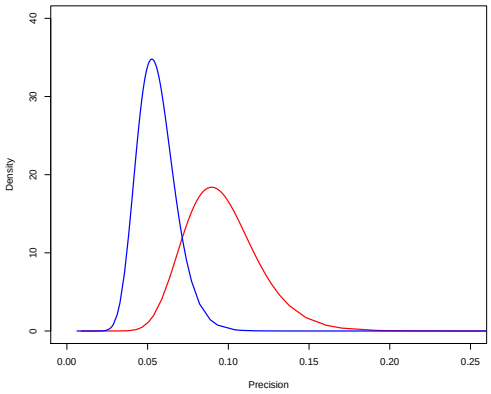


Figure 5.8: The posterior marginals for the precison of the besag model for new houses(red) and used houses(blue).

5.3 Introducing population sizes in the model

In the second model we also take into account potential non-linear effects of the population in each county. The second model's linear predictor also has an identity link and is defined as

$$\eta_i = \alpha + \beta z_i + f_1(c_{1i}) + f_2(log(c_{2i})) \tag{5.7}$$

which is the same as as in equation (5.6) with an addition of a population function $f_2(\cdot)$ that follows a random walk model of the second order described in (3.10). We choose the log of the population to reduce the large variation of the population.

The model in equation (5.7) which includes population should reduce the estimated spatial effects. We know that population is an important explanatory variable in terms of giving higher prices for higher populated counties. The reverse applies as well. To illustrate this, we have calculated the average prices and population for each county within the given time period. Figure 5.9 shows how the log population is spread across the counties as well as how the prices changed as a function of the log of the population for both types of houses. Akershus, Hordaland, Rogaland and Oslo have the highest populations. Finnmark has the lowest population. In the scatter plot we notice a non-linear trend in how prices increase with population. We have two observations that stand as being different from other observations. These observations are of Oslo. Notice that in fitting model (5.7) we do not use the average population sizes, but the registered population sizes for each year. `http://data.ssb.no/api/v0/dataset/49623?lang=no`

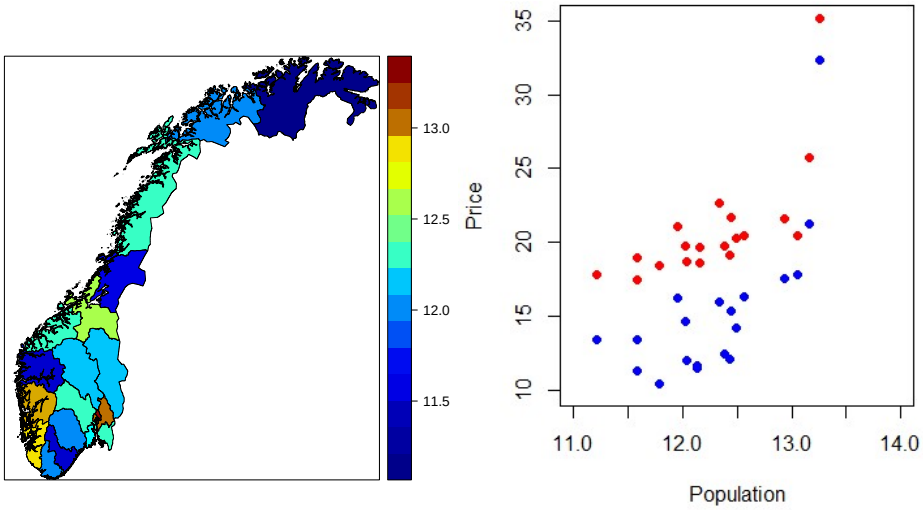


Figure 5.9: Average population for each county represented on a map of Norway and a graph that shows the log of the population plotted against the average price

5.3.1 Results

In figure 5.10 we plot the posterior means of the spatial effects for the intrinsic CAR model component on a map. These effects are now typically decreased as we have accounted for population in the model. We can see that even though Oslo has the highest posterior mean, it has dropped a lot compared to using the previous model in (5.6). We also have a lot of changes in the other counties also. For example, in Tromsø and Vestfold we have high positive estimated spatial effects. Hordaland and Møre og Romsdal have the lowest negative estimated spatial effects. Hordaland had positive effects for used houses, but now has negative estimated effects for both new and used houses. This means that the high prices in Hordaland can be partly explained by the high population. We can conclude the same about Akershus as well. We also notice that the low prices in Finnmark can be explained by the low population size.

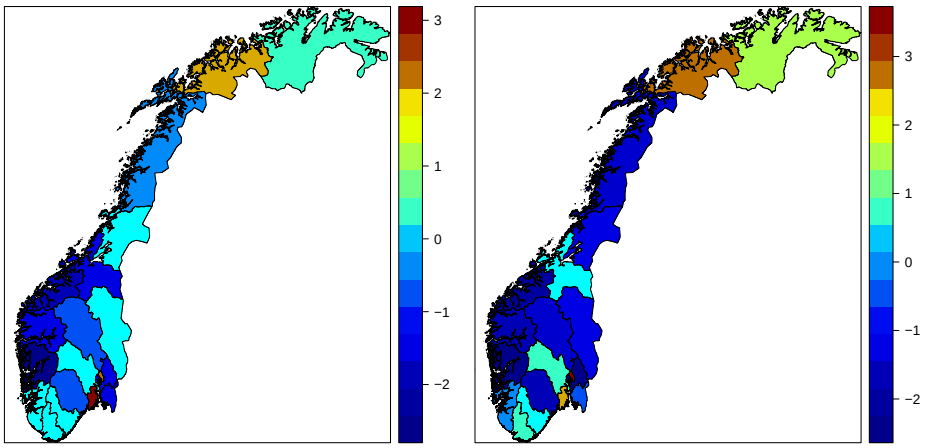


Figure 5.10: Spatial effect after accounting for population for new houses and used housed

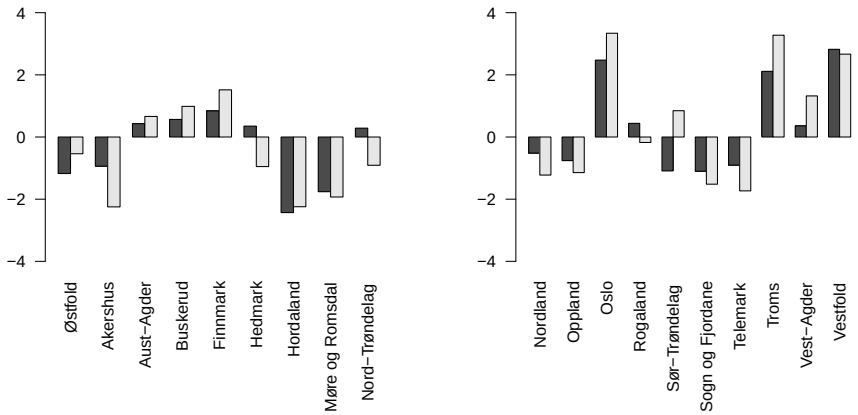


Figure 5.11: barplot that shows how different the same counties are for new and houses accounting for population

Table 5.3 shows the new credible intervals for the random effects of the intrinsic CAR model component. Østfold, Hordaland and Møre og Romsdal are counties that have significantly lower prices for new houses. For used houses, Akershus, Hordaland, Møre og Romsdal, Nordland, Oppland and Telemark have significantly negative effect. Only Oslo, Troms and Vestfold have a significant positive effect for both types of houses. Vest-Agder has significantly higher prices only for used houses. We can see that many of the counties prices are now explained by the linear trend for all counties when we take into account the population as expected. The posterior mean of the second order random walk model is visualized in figure 5.12. The population is in logarithmic scale and we can see that the population has a slowly increasing effect on the prices until the population gets large. When the population is very large the prices get higher. This steep curve comes from the high prices of Akerhus and Oslo. The effects are quite similar for new and used houses.

In figure 5.13 we can see the posterior marginals of the precision parameter for the second order random walk component. The posterior marginals are quite similar for the different types of houses.

	$CI.new_l$	$CI.new_u$	$CI.used_l$	$CI.used_u$
Østfold	-2.202	-0.142	-1.683	0.710
Akershus	-2.208	0.241	-3.757	-0.861
Aust-Agder	-0.915	1.824	-0.944	2.265
Buskerud	-0.293	1.462	-0.036	2.131
Finnmark	-3.153	4.633	-3.353	6.116
Hedmark	-0.673	1.404	-2.032	0.142
Hordaland	-3.686	-1.281	-3.702	-0.904
Møre og Romsdal	-2.703	-0.805	-2.990	-0.774
Nord-Trøndelag	-1.244	1.884	-2.579	0.796
Nordland	-1.500	0.486	-2.306	-0.064
Oppland	-1.697	0.190	-2.171	-0.122
Oslo	1.063	3.850	1.745	4.864
Rogaland	-0.797	1.648	-1.634	1.201
Sør-Trøndelag	-2.160	-0.028	-0.356	2.153
Sogn og Fjordane	-2.452	0.295	-3.156	0.112
Telemark	-1.912	0.074	-2.887	-0.621
Troms	1.006	3.241	2.098	4.443
Vest-Agder	-0.599	1.317	0.262	2.363
Vestfold	1.785	3.906	1.564	3.838

Table 5.3: 95% Credible intervals for the posterior mean of the intrinsic CAR model component after accounting for population

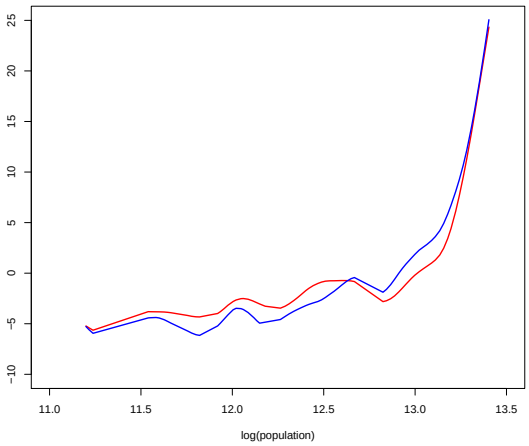


Figure 5.12: Population plotted against the mean of the second order random walk effects for both types of houses

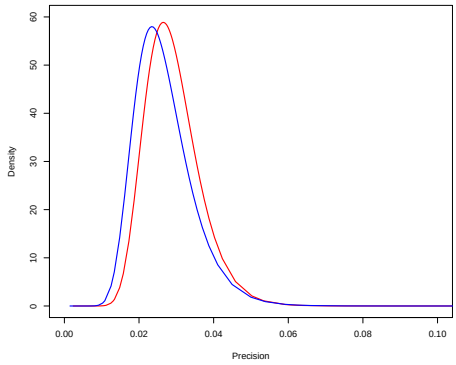


Figure 5.13: The posterior marginals for the second order random walk model for new houses(red) and used houses(blue)

Chapter 6

Discussion and concluding remarks

In section 5.1 the application of simple linear regression analysis led to finding that new houses have higher prices than used houses for all counties. We also found out that for all counties the gap between the prices seem to be increasing. In 1991 the prices for old and new houses were almost the same. One could argue that 19 years of annual observations is a small number and if there were more observations the difference between the price growths would be less. The linear models for the prices are only true for the given time period.

The model in section 5.2 showed that most prices are significantly lower than what can be explained by the linear trend for all counties. This is because the geographical differences between different counties are quite large. The visual interpretation of the spatial effects illustrated in figure 5.6 are obscured by the high prices in Oslo. Visually, the differences between the other countis

then seem quite small. An interesting find was that Rogaland and Hordaland have significantly higher prices for used houses, but not for new houses.

When we accounted for population in the model in section 5.3, this reduced the geographical differences between the counties. The intrinsic CAR model component shows that the mean of the random effects of the the high populated counties, such as Oslo, Hordaland and Akershus dropped. Even though the random effects dropped, the estimated spatial effect in Oslo was significantly higher. However the estimated effects became signifincantly lower in Hordaland and Akerhus.

There are limitations in this thesis we wish to highlight such as the complexity of the data. It would have been interesting to analyse spatial effects for the municipalities. For the municipalities we would have a larger graph of 428 municipalities instead of the 19 counties we have. We have time and population as explanatory variables for the prices. Inflation and the housing market could be other explanatory variables. The annual time intervals could have been monthly for possible detection of seasonal trends. Unfortunately such data was not openly available.

House prices interest people who wish to sell or buy houses such as real estate agents, and families . Future work on this thesis is to apply the analysis on houses prices in order to estimate the optimal time or location to buy or sell houses. This means combining the knowledge of statistics and the housing market in general.

This thesis helps to give a light understanding of the progression of house prices in Norway since 1999. We have seen a linear positive price growth for all counties, which is partly explained by population for most counties.

The structured additive regression model and INLA have made it possible to easily analysis the geographical and temporal differences in prices and providing posterior densities for many of the interesting parameters. Since the data was small, the computations were instant.

Bibliography

Julian Besag and Charles Kooperberg. On conditional and intrinsic autoregression. *Biometrika*, 82(4):733–746, 1995. doi: 10.2307/2337341.

M. Blangiardo and M. Cameletti. *Spatial and Spatio-temporal Bayesian models with R-INLA*. Wiley, 2015.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer New York, 2001.

A. Gelman, J. B. Carlin, H.S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Texts in statistical science. Chapman & Hall/CRC, 2003.

J. K. Ghosh, M. Delampady, and T. Sampanta. *An introduction to Bayesian analysis theory and methods*. Springer, 2006.

W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995.

G.H. Givens and J.A. Hoeting. *Computational Statistics*. Wiley Series in Computational Statistics. Wiley, 2012.

T. Hastie and R. Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

H. Hoijtink, I. Klugkist, and P. Boelen. *Bayesian Evaluation of Informative Hypotheses*. Statistics for Social and Behavioral Sciences. Springer New York, 2008.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186(1007):453–461, 1946. doi: 10.1098/rspa.1946.0056.

D.J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

J.A. Nelder and R.W.M Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society:Series A (General)*, 135(5):370–384, 1972. doi: 10.2307/2344614.

H. Rue and L. Held. *Gaussian Markov random fields: Theory and applications*. Chapman & Hall. Boca Raton, 2005.

H. Rue, S. Martino, and N.Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. doi: 10.1111/j.1467-9868.2008.00700.x.

H. Rue, A.Riebler, S. H. Sørbye, J. Illian, D. P. Simpson, and F. Lindgren. Bayesian computing with inla: A review. *Annual Review of Statistics and its Application*, 4(1), 3 2017. doi: 10.1146/annurev-statistics-060116-054045.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1–28, 02 2017. doi: 10.1214/16-STS576.

S. H. Sørbye and H. Rue. Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39 – 51, 2014. doi: 10.1016/j.spasta.2013.06.004.

Appendix

For the plots and tables in section 2.1

```
[fontsize=\small]
x=rbeta(10,23,3)
a<-curve( dbeta(x,2,3), xlim=c(0,1), ylim=c(0,4) ,main="Beta distribution",
ylab="Density")
lines( qbeta(p=c(0.025, 0.975), shape1=1, shape2=4), add=T, col='red' )
g=qbeta(p=c(0.025, 0.975), shape1=alfah+k, shape2=betah+n-k)
 f = function(x) qbeta(x, shape1=alfah+k, shape2=betah+n-k)
  aa<-hdi(f,credMass=0.95
g[1]
abline(v=g[1],col='red')
abline(v=g[2],col='red')
abline(v=aa[1],col='blue')
abline(v=aa[2] ,col='blue')
betahv=c(2,1,6,8)
alfahv=c(1,4,2,1)
betadata<-matrix(0, 4, 6)
for (i in 1:4){
```

```
  betadata[i,1] <-(alfahv[i]+k)/(alfahv[i]+betahv[i]+n)
  betadata[i,2] <-(alfahv[i]+k)*(betahv[i]+n-k)/((alfahv[i]
  +betahv[i]+n)^2  *(alfahv[i]+betahv[i]+1+n))
  g=qbeta(p=c(0.025, 0.975), shape1=alfahv[i]+k,
  shape2=betahv[i]+n-k)
  betadata[i,3] <-g[1]
  betadata[i,4] <-g[2]
  f = function(x) qbeta(x, shape1=alfahv[i]+k, shape2=betahv[i]+n-k)
  aa<-hdi(f,credMass=0.95)
  betadata[i,5]<-aa[1]
  betadata[i,6]<-aa[2]
}
ckdata<-betadata
row.names(ckdata)<- c("\alpha=1,\beta=2","\alpha=4,\beta=1",
"\alpha=2,\beta=6","\alpha=1,\beta=8")
colnames(ckdata)<-c("E(\theta|y,\alpha,\beta)",
"Var(\theta|y,\alpha,\beta).",
"CI_l","CI_u","HPD_l","HPD_u")

xtable(ckdata,digits = 3)
```

For drawing the map in sections 5.2 and 5.3

```
[fontsize=\small]
#map function#
```

```
library(rgdal)
library(rgeos)        # For geometric operations
library(plyr)         # For simple data manipulation
library(spatstat)
library(spdep)        # For  setting up polygon-neighbours
library(maptools)
library(RColorBrewer)
library(lattice)
mapping<-function(e,p){
  # the map files for the counties
  fylke=readOGR(dsn="C:/Users/george/Documents/house/NOR_adm_shp",
  layer="NOR_adm2")

  fylkedata <- fylke@data
  str(fylkedata); head(fylkedata)
  #list of names of the counties
  d <- c("Østfold" ,"Akershus" , "Aust-Agder", "Buskerud", "Finnmark",
    "Hedmark",    "Hordaland",  "Møre og Romsdal" , "Nord-Trøndelag" ,
      "Nordland" ,  "Oppland",     "Oslo", "Rogaland",
      "Sør-Trøndelag","Sogn og Fjordane",  "Telemark",
       "Troms", "Vest-Agder" , "Vestfold" )

  name3 <- c("NAME_1", "Input.variable");
   dt2 <- as.data.frame(cbind(d, e),
   stringsAsFactors=FALSE)
```

```
  dt2$e <- as.numeric(dt2$e); colnames(dt2) <- name3;
   Input.variable <- dt2
  # We plot the Norwegian regions using the unionSpatialPolygons
  # function from the 'maptools' package
  IDs <- fylkedata$ID_1
  # We merge Polygons
  norway3_new <- unionSpatialPolygons(fylke, IDs)
  # We build the new SpatialPolygonsDataFrame with the Input.variable
  norway4_new <- SpatialPolygonsDataFrame(norway3_new, Input.variable)
  pal2 <- colorRampPalette(c("blue4", "cyan","white", "yellow", "red4"))
# Plot the regions with Lattice
  spplot(norway4_new, "Input.variable", main=p,
         lwd=.2, col="black", col.regions=pal2(19),
          colorkey = list(space = "right"),
          bty="n")
}
```

For preprocessing the house data

```
#fixing dataset#
########
library(tidyr)
#####
bolig<-read.csv("data25138.txt",sep=";",stringsAsFactors=FALSE)
head(bolig)
index<-seq(1,2166,3)
```

```
dim(bolig)
kvdmpris=bolig[index,]
nyebolig<-kvdmpris[kvdmpris$type.enebolig %in% "01 Nye eneboliger",]
bruktbolig<-kvdmpris[kvdmpris$type.enebolig %in% "02 Brukte eneboliger",]
oslony<-nyebolig[nyebolig$region %in% "03 Oslo",]
oslony<-oslony[-c(10),]
sr<-summary(lm(oslony$Gjennomsnittlig.kvadratmeterpris..etter.region.
.type.enebolig.
.år.og.statistikkvariabel~oslony$år,data=oslony))
value2008<-sr$coefficients[1]+sr$coefficients[2]*2008
#which(is.nan(nyebolig$Gjennomsnittlig.kvadratmeterpris..etter.region.
.type.enebolig..år.og.statistikkvariabel))
nyebolig$Gjennomsnittlig.kvadratmeterpris..etter.region.
.type.enebolig..år.og.statistikkvariabel[48]<-value2008
finny<-nyebolig[nyebolig$region %in% "20 Finnmark - Finnmárku",]
finny<-finny[-c(8),]
fr<-summary(lm(finny$Gjennomsnittlig.kvadratmeterpris..etter.region.
.type.enebolig..år.og.statistikkvariabel~finny$år,data=finny))
value2006<-fr$coefficients[1]+fr$coefficients[2]*2006
region.id<-c(rep(1,38),rep(2,38),rep(12,38),rep(6,38),rep(11,38),
rep(4,38),rep(19,38),rep(16,38),rep(3,38),rep(18,38),rep(13,38),
rep(7,38),rep(15,38),rep(8,38),rep(14,38),
rep(9,38),rep(10,38),rep(17,38),rep(5,38))
cac<-c(rep(1,19),rep(2,19))
cac1<-c(rep(1,19),rep(0,19))
```

```
boligtype.id1<-c(rep(cac1,19))
# boligtype<-cbind(boligtype.id,boligtype.id1)
#   region<-gsub("[[:digit:]]", "", kvdmpris$region)
year<-kvdmpris$år
ave.price<-kvdmpris$Gjennomsnittlig.kvadratmeterpris..etter.region.
.type.enebolig..år.og.statistikkvariabel
value2008<-round(value2008)
value2006<-round(value2006)
ave.price[692]<-value2006
ave.price[86]<-value2008

ave.pricey<-as.numeric(ave.price)
ave.pricey<-ave.pricey/1000
house.data<-data.frame(region.id,year,boligtype.id1,ave.pricey)
year2=seq(1,19,1)
year2=rep(year2,38)
length(boligtype.id1)
house.data<-data.frame(region.id,year2,boligtype.id1,ave.pricey)
```

For the tables in 5.1

```
lmdata<-matrix(0, 19, 6)
for (i in 1:19){
  y1=house.data$ave.pricey[house.data$region.id==i &
   house.data$boligtype.id1==1]
  x1= house.data$year2[house.data$region.id==i &
```

```
 house.data$boligtype.id1==1]
y2=house.data$ave.pricey[house.data$region.id==i &
house.data$boligtype.id1==0]
x2=house.data$year2[house.data$region.id==i &
 house.data$boligtype.id1==0]

rgnew<-lm(y1~x1)
rgused<-lm(y2~x2)
simple1<-y1~x1
simple.result1<-inla(simple1,data=list(y=y1,x=x1))
#simple.result1<-inla(simple1,data=house.data[(1+19*2*(i-1))
:(19+19*2*(i-1)),])
simple2<-y2~x2
#simple.result2<-inla(simple2,data=house.data[(19*2*i-18):(19*2*i),])
simple.result2<-inla(simple2,data=list(y=y2,x=x2))
lmdata[i,1] <-simple.result1$summary.fixed$mean[2]
lmdata[i,2] <-simple.result1$summary.fixed$`0.025quant`[2]
lmdata[i,3] <-simple.result1$summary.fixed$`0.975quant`[2]
lmdata[i,4] <-simple.result2$summary.fixed$mean[2]
lmdata[i,5] <-simple.result2$summary.fixed$`0.025quant`[2]
lmdata[i,6] <-simple.result2$summary.fixed$`0.975quant`[2]

}
```

For figures in section 5.1

```
dog <- c("Østfold" ,"Akershus" , "Aust-Agder", "Buskerud", "Finnmark",
  "Hedmark" , "Hordaland",  "Møre og Romsdal" , "Nord-Trøndelag" ,
    "Nordland" ,  "Oppland", "Oslo", "Rogaland","Sør-Trøndelag",
    "Sogn og Fjordane",  "Telemark",  "Troms",  "Vest-Agder" ,
    "Vestfold" )
for (i in 1:19){
  png(filename = paste(dog[i],sep="",".png"))
  y1=house.data$ave.pricey[house.data$region.id==i &
   house.data$boligtype.id1==1]
  x1= house.data$year2[house.data$region.id==i &
   house.data$boligtype.id1==1]
  y2=house.data$ave.pricey[house.data$region.id==i &
   house.data$boligtype.id1==0]
  x2=house.data$year2[house.data$region.id==i &
   house.data$boligtype.id1==0]
  simple1<-y1~x1
  simple.result1<-inla(simple1,data=list(y=y1,x=x1))
  simple2<-y2~x2
  simple.result2<-inla(simple2,data=list(y=y2,x=x2))
  plot(house.data$year[house.data$region.id==i & house.data$boligtype.id1==1]
  house.data$ave.pricey[house.data$region.id==i & house.data$boligtype.id1==1
  ylab="Average Price per Sq.m x 1000kr",ylim=c(0,60))
  points(house.data$year[house.data$region.id==i & house.data$boligtype.id1==
  ,house.data$ave.pricey[house.data$region.id==i & house.data$boligtype.id1==
  pch=19,col="blue")
```

```
  abline(a=simple.result1$summary.fixed$mean[1],
  b=simple.result1$summary.fixed$mean[2],col="red")
  abline(a=simple.result2$summary.fixed$mean[1],
  b=simple.result2$summary.fixed$mean[2],col="blue")
  dev.off()
}
```

For plots and tables in 5.2

```
hus1<-house.data[!(house.data$boligtype.id1==0 %in% house.data),]
hus2<-house.data[(house.data$boligtype.id1==1 %in% house.data),]
library(INLA)
g = inla.read.graph("norwa.graph")
u=1 # For eksempel
alpha=0.01

formula2 = hus1$ave.pricey ~ f(hus1$region.id,model="besag",
graph=g,scale.model=T, hyper=list(prec=list(prior="pc.prec"
,param=c(u,alpha))))+hus1$year
result  =inla(formula2,data=hus1)
echonew<-result$summary.random$‘hus1|S|region.id‘$mean
plot(result$marginals.hyperpar[[1]], type = "l",ylab="Density",
xlab = "Precision")
charlie<-result$summary.random$‘hus1|S|region.id‘$sd
formula3 = hus2$ave.pricey ~ f(hus2$region.id,model="besag",
graph=g,scale.model=T, hyper=list(prec=list(prior="pc.prec",
```

```
param=c(u,alpha))))+hus2$year
result3  =inla(formula3,data=hus2)
plot(result3$marginals.fixed[1],type = "l",ylab="Density",xlab = "Alpha",
xlim=c(0,8),ylim=c(0,20),lwd=2)
plot(result$marginals.fixed[[2]],type ="l", col=2,ylab="Density",xlab = "Beta
xlim=c(0.7,1.7),ylim=c(0,20),lwd=2)
lines(result3$marginals.fixed[[2]],col=4,lwd=2)
echoold<-result3$summary.random$‘hus2|S|region.id‘$mean
mapping(echonew,p=NULL)
mapping(echoold,p=NULL)
mapping(charlie,p=NULL)
test2 <- rbind(echonew[1:9],echoold[1:9])
names(test2)<-dog[1:9]
par(mar=c(8, 4.1, 4.1, 2.1))
barplot(test2,beside=T,ylim = c(-5,16),names.arg = dog[1:9],las=2)
text( -3.7, srt = 60, adj= 1, xpd = TRUE, labels = names(dog[1:9]) ,
 cex=1.2)
cidata<-matrix(0, 19, 4)
for (i in 1:19){
  cidata[i,1]<-resultp$summary.random$‘hus1|S|region.id‘$‘0.025quant‘[i]
  cidata[i,2]<-resultp$summary.random$‘hus1|S|region.id‘$‘0.975quant‘[i]
  cidata[i,3]<-result3p$summary.random$‘hus2|S|region.id‘$‘0.025quant‘[i]
  cidata[i,4]<-result3p$summary.random$‘hus2|S|region.id‘$‘0.975quant‘[i]
}
ckdata<-cidata
```

```
row.names(ckdata)<- c("Østfold" ,"Akershus" , "Aust-Agder", "Buskerud",
 "Finnmark",  "Hedmark" ,  "Hordaland",  "Møre og Romsdal" ,
  "Nord-Trøndelag" ,
    "Nordland" ,  "Oppland", "Oslo", "Rogaland",    "Sør-Trøndelag",
    "Sogn og Fjordane",
      "Telemark",  "Troms", "Vest-Agder" ,
  "Vestfold" )
colnames(ckdata)<-c("CIlnew","CIunew","CIlusd","CIuused")
xtable(ckdata,digits = 3)


test23 <- rbind(echonew[10:19],echoold[10:19])
names(test23)<-dog[10:19]
barplot(test23,beside=T,ylim = c(-5,16),names.arg = dog[10:19],las=2)
```

For plots and tables in 5.3f

```
folk<-read.csv("folkemengde.csv",sep=";")
folk$region <- NULL
colnames(folk)<-c()
folkli<-as.vector(t(folk))
af<-c(rep(1,19),rep(2,19),rep(3,19),rep(4,19),rep(5,19),rep(6,19),rep(7,19),
rep(8,19),rep(9,19),rep(10,19),rep(11,19),rep(12,19),rep(13,19),rep(14,19),
rep(15,19),rep(16,19),rep(17,19),rep(18,19),rep(19,19))
afk<-cbind(af,folkli)
afk<- as.data.frame(afk)
kk<-NULL
```

```
for (i in 1:19) {
  jj<-append(kk,afk$folkli[af==i])
  kk<-append(jj,afk$folkli[af==i])
}
kk<-log(kk)
hd1<-cbind(house.data,kk)
huse1<-hd1[!(hd1$boligtype.id1==0 %in% hd1),]
huse2<-hd1[(hd1$boligtype.id1==1 %in% hd1),]

library(INLA)
g = inla.read.graph("norwa.graph")
# standard BYM model (without covariates)
formula2p= huse1$ave.pricey ~ f(huse1$region.id,model="besag",
graph=g,scale.model=T, hyper=list(prec=list(prior="pc.prec",
param=c(u,alpha))))
+huse1$year+f(inla.group(huse1$kk,100,),model="rw2",scale.model = T,
hyper=list(prec=list(prior="pc.prec",param=c(u,alpha))))
resultp  =inla(formula2p,data=huse1)
echonewp<-resultp$summary.random$‘huse1|S|region.id‘$mean
formula3p = huse2$ave.pricey ~ f(huse2$region.id,model="besag",
graph=g,scale.model=T, hyper=list(prec=list(prior="pc.prec",
param=c(u,alpha))))+huse2$year
+f(inla.group(huse2$kk,100,),model="rw2",scale.model = T,
hyper=list(prec=list(prior="pc.prec",param=c(u,alpha))))
result3p  =inla(formula3p,data=huse2)
```

```
echooldp<-result3p$summary.random$‘huse2|S|region.id‘$mean
mapping(echonewp,p=NULL)
mapping(echooldp,p=NULL)
test2p <- rbind(echonewp[1:9],echooldp[1:9])
names(test2p)<-dog[1:9]
par(mar=c(8, 4.1, 4.1, 2.1))
barplot(test2p,beside=T,ylim = c(-4,4),names.arg = dog[1:9],las=2)
text( -3.7, srt = 60, adj= 1, xpd = TRUE, labels = names(dog[1:9]),
 cex=1.2)
cidata<-matrix(0, 19, 4)
for (i in 1:19){
  cidata[i,1]<-resultp$summary.random$‘huse1|S|region.id‘$‘0.025quant‘[i]
  cidata[i,2]<-resultp$summary.random$‘huse1|S|region.id‘$‘0.975quant‘[i]
  cidata[i,3]<-result3p$summary.random$‘huse2|S|region.id‘$‘0.025quant‘[i]
  cidata[i,4]<-result3p$summary.random$‘huse2|S|region.id‘$‘0.975quant‘[i]
}
ckdata<-cidata
row.names(ckdata)<- c("Østfold" ,"Akershus" , "Aust-Agder", "Buskerud",
 "Finnmark",  "Hedmark" ,  "Hordaland",  "Møre og Romsdal" ,
  "Nord-Trøndelag" ,   "Nordland" ,   "Oppland", "Oslo", "Rogaland",
    "Sør-Trøndelag","Sogn og Fjordane",  "Telemark",    "Troms",
     "Vest-Agder" , "Vestfold" )
colnames(ckdata)<-c("CIlnew","CIunew","CIlusd","CIuused")
xtable(ckdata,digits = 3)
```

```
test23p <- rbind(echonewp[10:19],echooldp[10:19])
names(test23)<-dog[10:19]
par(mar=c(8, 4.1, 4.1, 2.1))
barplot(test23p,beside=T,ylim = c(-4,4),names.arg = dog[10:19],las=2)
plot(resultp$summary.random$‘inla.group(huse1|S|kk, 100, )‘$ID,
resultp$summary.random$‘inla.group(huse1|S|kk, 100, )‘$mean,type="l",
xlab="log(population)", ylab="",xlim=c(11,13.5),ylim=c(-10,25),lwd=2,col=2)
lines(result3p$summary.random$‘inla.group(huse2|S|kk, 100, )‘$ID,
result3p$summary.random$‘inla.group(huse2|S|kk, 100, )‘$mean,type="l",
lwd=2,col=4,ylab=NULL)
```