Humboldt University of Berlin
School of Business and Economics

Master's Thesis

A Bayesian Workflow for Poverty Estimation

Flavio Mejía Morelli

Email: flavio-morelli@outlook.com
First supervisor: Prof. Dr. Natalia Rojas Perilla
Second supervisor: Prof. Dr. Timo Schmid
Student Nr.: 608727

November 15, 2021

Abstract

Reliable poverty estimates are the basis for public policy decision-making. However, the estimation of poverty indicators is challenging not only due to the unimodal, leptokurtic and skewed nature of income data, but also because of small sample sizes in income surveys. This paper applies the ideas of the Bayesian workflow proposed by Gelman et al. (2020) to develop a hierarchical Bayesian (HB) model iteratively in the *small area estimation* (SAE) context. The main advantages of the Bayesian paradigm are modelling flexibility and a high-degree of model interpretability. After comparing plausible modelling approaches, the Bayesian model is developed by defining adequate priors on regression coefficients and variance parameters, doing variable selection, redefining the specification of the random effect and including area-level correlation. The resulting Bayesian model is then compared with the frequentist EBP approach under data-driven transformations (Rojas Perilla et al., 2020).

Die verlässliche Schätzung von Armutsindikatoren ist eine notwendige Grundlage für den politischen Entscheidungsfindungsprozess. Die Schätzung von Armutsindikatoren ist jedoch nicht nur aufgrund der unimodalen, leptokurtischen und schiefen Verteilung von Einkommensdaten eine Herausforderung, sondern auch wegen der kleinen Stichprobengrößen bei Einkommenserhebungen. Die vorliegende Arbeit wendet die Grundprinzipien des Bayesianischen Workflows aus Gelman et al. (2020) an, um ein hierarchisches Bayesianisches (HB) Modell im Kontext von *small area estimation* (SAE) iterativ zu entwickeln. Die Vorteile des Bayesianischen Paradigmas sind zum einen die Flexibilität in der Modellierungsphase und zum anderen die Modellinterpretierbarkeit. Unter Berücksichtigung verschiedener Modellierungsansätze werden das Bayesianische Modell entwickelt, passende A-Priori-Verteilungen für Regressionskoeffizienten und Varianzparameter definiert, eine Variablenselektion vorgenommen, die Spezifikation des *random effect* geändert und die räumliche Korrelation einbezogen. Das resultierende Bayesianische Modell wird im Abschluss mit dem frequentistischen EBP-Ansatz unter datengetriebenen Transformationen (Rojas Perilla et al., 2020) verglichen.

Contents

List of Figures

List of Tables

1. Introduction

In recent years, there has been an increased interest in the distinction between *inference* and *workflow* in Bayesian statistics (Gelman et al., 2020). *Inference* describes the process by which models are fitted using data, with an emphasis on the correct quantification of estimate uncertainty. With the increased accessibility of Bayesian statistics, the focus has shifted away from developing one single model or method to taking an iterative approach – also called *workflow* (Gelman et al., 2020). Doing inference on a single model is not enough, because uncertainty not only impacts estimated parameters, but also model choice. It is not uncommon to find multiple models that are compatible with the problem at hand. Moreover, the Bayesian paradigm lets certain aspects of the model be changed in a modular way. This modularity is an advantage, as it allows a higher degree of flexibility when developing a model and also provides the tools to develop a series of models iteratively.

This paper aims to develop a hierarchical Bayesian model (Molina et al., 2014) iteratively to estimate poverty indicators (Foster et al., 1984) based on the workflow presented by Gelman et al. (2020) in a *small area estimation* (Rao & Molina, 2015) context. Poverty is chosen as a use case for a wide variety of reasons. Firstly, ending poverty is one of the Sustainable Development Goals (SDG) of the United Nations (United Nations, 2015). Providing reliable estimates is a necessary condition to track progress on this particular goal. Secondly, income data (the basis for poverty indicators) presents particular challenges due to its unimodel, leptokurtic and skewed nature, which can be approached from different perspectives in the context of a Bayesian workflow. Thirdly, due to privacy reasons, questions on income are usually not part of the census. Often, income data is collected in surveys, which have a sample size several orders of magnitude lower than the census population. When disaggregating the data according to categories such as municipalities, gender and ethnicity, the resulting subgroups can be sparse. In such cases, small area estimation (SAE) methods are necessary to provide reliable estimators. This paper compares the results of the Bayesian model developed in the workflow with the frequentist SAE approach (EBP with data-driven transformations) of Rojas Perilla et al. (2020).

The structure of this paper deviates from the statistical literature, which usually first presents a method that is then applied to real data and compared to existing benchmarks.

This deviation stems from the specific characteristics of the workflow developed by Gelman et al. (2020): the focus on iterative model development requires a structure that first introduces the general theoretical ideas and the problem to be captured by the model and then builds up the model in a modular way. Chapter 2 provides the theoretical background on Bayesian statistics and SAE used in the rest of the paper. The data from the Mexican state of Guerrero is introduced in chapter 3 together with three simulation scenarios used as a benchmark. In chapter 4, the Bayesian workflow is presented as defined by Gelman et al. (2020). Starting with different approaches to modelling income, the model is improved iteratively by considering different priors on regression coefficients and variance, selecting variables with the highest predictive power, redefining the specification of the random effect and modelling correlations at the area level. A comparison between the Box-Cox EBP (Rojas Perilla et al., 2020) and the model developed with the Bayesian workflow is included in chapter 5, both in terms of estimates and uncertainty quantification. Chapter 6 critically discusses the advantages and disadvantages of a Bayesian workflow for poverty estimation in the SAE context. The paper finishes with concluding remarks.

2. Theoretical Background

This chapter is an overview of the key theoretical concepts used in the rest of this paper. The main aim is to provide a short review to the reader, but not to discuss every theoretical aspect in detail. Therefore, additional literature references are mentioned throughout the chapter. Note that more specific theoretical ideas will be introduced in the relevant step of the workflow in chapter 4.

2.1. Overview of Bayesian statistics

This section provides a short introduction to the main concepts of Bayesian statistics. An exhaustive exposition of Bayesian statistics can be found in sources such as Gelman et al. (2014) or McElreath (2020). The first part presents the notation of Bayes' theorem and introduces the components of a Bayesian model, while the second part summarizes different ways of evaluating Bayesian models that will be used throughout this paper to compare multiple models. The more technical topic of Bayesian computation is introduced in appendix A – specifically, the Markov Chain Monte Carlo (MCMC) and Hamiltonian

Monte Carlo (HMC) algorithms. In general, the reader should assume that there were no major computational problems in the models presented, unless otherwise specified.

2.1.1. Bayesian inference and terminology

The concepts presented in this section are based on chapters 1 and 2 from Gelman et al. (2014). Inference is the process of analyzing population parameters based on samples from the population. Because the sample is an imperfect representation of the population, there is always uncertainty associated with parameter inference. In frequentist statistics, the common approach is to calculate point estimates of the true parameters from the data. Uncertainty about the estimate comes from the randomness of the data in the sample and it is usually quantified by confidence intervals, which are based on distributional assumptions such as normality. Bayesian inference takes a different approach: the true parameter is not a constant but a random variable, while the data is assumed to be fixed. The aim is to estimate the distribution of the population parameter θ given the data y, i.e. $p(\theta|y)$. The derivation of the distribution is based on Bayes' theorem

$$p(\theta|y) \overset{\text{(a)}}{=} \frac{p(\theta, y)}{p(y)} \overset{\text{(b)}}{=} \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta}.$$

Here, θ is a vector of parameters and vector y represents the data, which can be anything from a single variable to a multivariate outcome variable. Inference is based on the distribution $p(\theta|y)$ – the **posterior distribution**. The first equality (a) is the definition of conditional probability. By using the definition of conditional probability again, it is possible to decompose the joint distribution $p(\theta, y)$ into $p(\theta)$ and $p(y|\theta)$ in the second inequality (b). $p(\theta)$ is the **prior distribution** which encodes knowledge about the parameter vector θ prior to collecting the data. One example of prior information in the context of a socioeconomic survey would be knowing the past average per capita income or the Gini coefficient of a country from sources such as a textbook or existing research papers before even collecting current data. Prior information also includes well-known facts about distribution parameters – e.g., the variance cannot be negative. This information can be used to parametrize the prior distribution. The **likelihood** is given by $p(y|\theta)$ and it reflects the modelling of the data generating process. A simple example is a binomial distribution that models the number of sucesses y in n tosses of a coin, given the probability θ. Both the prior and the likelihood are the main ingredients of Bayesian inference.

The term in the denominator $p(y)$ is the **marginal likelihood**, which gets its name from marginalizing θ out of the expression in the numerator by integration: $\int p(\theta)p(y|\theta)d\theta$.

The marginal likelihood is a normalizing constant that ensures that $p(\theta|y)$ is a proper distribution, i.e., integrates to 1. The integral is intractable even for very simple models and analytic solutions exist mostly for simple cases of conjugate priors. Therefore, it is necessary to use Monte Carlo methods or variational inference to estimate the posterior. The popularity of Bayesian methods has increased with the availabilty of computational power that makes such estimation methods more accessible (see appendix A). Nevertheless, the marginal likelihood is more than a simple normalizing constant. By sampling from $p(y)$, the marginal likelihood can be used to generate data from the model, even before observing any data. Therefore, it is also called the **prior predictive distribution**. To generate new data $\tilde{y}$ from the posterior model, replace the prior $p(\theta)$ with the posterior $p(\theta|y)$ in the definition of $p(y)$ and the result is $p(\tilde{y}|y) = \int p(y|\theta)p(\theta|y)d\theta$, where $p(\tilde{y}|y)$ is the **posterior predictive distribution**. Data generated from the prior predictive and posterior predictive distributions can be used to check model quality.

2.1.2. Evaluation of Bayesian models

There are numerous ways to evaluate Bayesian models. Piironen & Vehtari (2017a) provide an overview of evaluation methods that quantify predictive power, which can be used to compare different models. Moreover, Bayesian inference provides two additional checking tools: prior and posterior predictive checks. The main idea is to generate numerous samples either from the prior predictive or posterior predictive distributions described in the previous section. Thus, it is possible to check how the model behaves before and after fitting the data and whether the generated samples are in a plausible range compared to the dependent variable. This section focuses on Pareto Smoothed Importance Sampling Leave-One-Out Cross-Validation (PSIS-LOO) as a measure of predictive power and includes a brief discussion of prior and posterior predictive checks (Vehtari et al., 2017).

Given some data y and future observations $\tilde{y}_i, i = 1, ..., N$, where N is the original sample size, the quality of the predictive distribution can be defined in terms of a utility function as a logarithmic score (Piironen & Vehtari, 2017a)

$$u(\tilde{y}_i) = \sum_{i=1}^{N} \log p(\tilde{y}_i|y).$$

However, as $\tilde{y}$ is unobserved, it is necessary to marginalize it out of the expression, thus

getting the expected utility

$$\text{elpd} = \sum_{i=1}^{N} E[\log p(\tilde{y}_i|y)] = \sum_{i=1}^{N} \int p_t(\tilde{y}_i) \log p(\tilde{y}_i|y) d\tilde{y}_i,$$

where p_t is the true data generating distribution and elpd stands for expected log pointwise predictive density for a new data set. As p_t is unknown, it is not possible to calculate the elpd directly. An unbiased estimate for the elpd is given by

$$\text{elpd}_{\text{loo}} = \sum_{i=1}^{N} \log p(\tilde{y}_i|y_{-i}), \quad p(\tilde{y}_i|y_{-i}) = \int p(\tilde{y}_i|\theta) p(\theta|y_{-i}) d\theta. \tag{2.1}$$

Here, $p(\tilde{y}_i|y_{-i})$ is the leave-one-out predictive distribution given y_{-i}, i.e. the data without the i-th observation. In practice, elpd_{loo} is estimated efficiently by Pareto-smoothed importance sampling without having to refit the model N times and is therefore referred to as PSIS-LOO. A higher elpd_{loo} indicates a model with a higher predictive power (Vehtari et al., 2017). Pareto smoothing is not only useful for the efficient estimation of elpd_{loo}, but it also provides a diagnostic to whether the estimated PSIS-LOO can be trusted. Specifically if the shape parameter k of the Pareto distribution is higher than 0.7, then PSIS-LOO is not reliable (Vehtari et al., 2017). Problems with very high k values can be alleviated with the moment match method Paananen et al. (2021), but they can also indicate that the model does not deal well with outliers. Note that a utility function that quantifies predictive power takes into account the distributional characteristics of the model. In contrast, loss function approaches such as the root mean squared error (RMSE) or the mean absolute error (MAE) measure the distance between the predicted values and the true values, but are not directly related to the distribution.

Beyond PSIS-LOO, prior and posterior predictive checks are useful to assess model quality. The main idea is to generate multiple samples from the prior predictive distribution $\int p(y|\theta)p(\theta)d\theta$ or from the posterior predictive distribution $\int p(y|\theta)p(\theta|y)d\theta$. With these samples, it is possible to check how similar the simulated distribution is to the original distribution and also whether the simulations are within a reasonable range. This similarity check can be done for example with histograms, KDE plots or scatter plots when using the full distribution. As Bayesian models produce a full distribution, it is also possible to check how certain summary statistics (median, IQR, variance, quantiles etc.) vary between the simulated samples and compare them to the summary statistics in the original distributions (Gelman et al., 2014, Chapter 6). However, these statistics should be ancillary in the sense that they should test something different than parameter fit. For example, while a linear

model fits both the mean and the variance quite well, this is not the case for a Poisson regression, where the variance of the dependent variable is assumed to be equal to its expected value. Therefore, checking the mean and the variance in the Poisson case will reveal potential problems in the model (Stan Development Team, 2021, Chapter 27.3). A practical application of prior and posterior checks will be shown in sections 4.2 and 4.3.

Another tool that can be used to evaluate and compare models is *stacking* (Yao et al., 2018). The details of this method are highly technical and beyond the scope of this paper. Therefore, only a short summary is provided based on the explanation in Gelman et al. (2020). Cross-validation can be used to compare the performance of multiple models, but there is also uncertainty in model comparison. Choosing only one model discards the information contained in all other models. Stacking makes it possible to combine inferences from different models using weights that minimize the cross-validation error. In a nutshell, if there are two models and the first one outperforms the second one 70% of the time, then the weights should be roughly 0.7 and 0.3. Moreover, stacking is robust to the presence of multiple similar models. If there is a group of three nearly identical models, the weights of two models will be almost zero while the third one gets a weight that represents the whole group of models. In this sense, stacking is also a measure of model heterogeneity. Further details can be found in Yao et al. (2018).

2.2. Poverty indicators and small area estimation

After the introduction to the basics of Bayesian inference and model evaluation, this section describes how hierarchical Bayesian models are used in small area estimation – with a special focus on unit-level models. It then defines the poverty indicators that are at the center of this paper and presents an algorithm to estimate those indicators with hierarchical Bayesian models.

2.2.1. Bayesian unit-level model for small area estimation

Small area estimation is a field of survey statistics that deals with prediction in areas for which there is little or no information. The terms *area* or *domain* – usually used as synonyms – do not necessarily imply a geographic area. More generally, it can denote any subgroup of a population arising from disaggregation – by gender, region, ethnicity, etc. A typical scenario for small area estimation arises in surveys, where the representative sample is small compared to the whole population. If indicators are needed at a finer disaggregated level (e.g., by municipality), the sample size in each area can become extremely small (under 20

observations) and there might be areas with no observations at all. To improve predictions for these small areas, the models borrow strength from additional data sources such as a census or a register. Depending on the data available, there are two types of small area models. Area-level models such as the Fay-Herriot use aggregated data for each area to improve direct estimators. On the other hand, unit-level models need information at the individual or household level to generate predictions (Rao & Molina, 2015, Chapter 1 and 2).

Molina et al. (2014) formulate a Bayesian unit-level model, which is referred to as the Hierarchical Bayes (HB) model and is based on the Battese-Harter-Fuller (BHF) model (Battese et al., 1988):

$$\begin{aligned} y_{di}|\boldsymbol{\beta}, u_d, \sigma_e &\sim \mathcal{N}(\boldsymbol{x'}_{di}\boldsymbol{\beta} + u_d, \sigma_e),\ d = 1, ..., D,\ i = 1, ..., N_d \\ u_d|\sigma_u &\sim \mathcal{N}(0, \sigma_u), d = 1, ..., D \\ p(\boldsymbol{\beta}, \sigma_u, \sigma_e) &= p(\boldsymbol{\beta})p(\sigma_u)p(\sigma_e) \propto p(\sigma_u)p(\sigma_e). \end{aligned} \tag{2.2}$$

The first distribution defines the likelihood and the last two lines define the prior, taking into account conditional dependencies. D is the number of domains and N_d is the number of observations for domain $d = 1, ..., D$, whereas y_{di} and $\boldsymbol{x}_{di}$ are respectively the dependent and independent variables for area d and observation i in that area. $\boldsymbol{\beta}$ is a vector of regressor coefficients common to all areas. The effect for area d is given by u_d and the common variance parameter for all area effects is σ_u. The variance parameter at the individual level is given by σ_e. Note that $\boldsymbol{\beta}, \sigma_u$ and σ_e are assumed to be independent. Their priors are $p(\boldsymbol{\beta}), p(\sigma_u)$ and $p(\sigma_e)$ respectively.

However, model 2.2 does not take full advantage of Bayesian modelling. By taking the normal distribution as the likelihood, the model faces the same limitations of a frequentist linear regression and will not be able to deal with heavy-tailed data.[1] Moreover, the prior distributions in Molina et al. (2014) are non-informative (flat), i.e., they are proportional to a constant and not a proper distribution. This poses two problems. Firstly, a Bayesian model with flat priors is not a generative model in the sense that it is not possible to simulate new observations from the prior predictive distribution. Secondly, it does not take advantage of the extra control that priors provide over the model when modelling relation between parameters. Therefore, Morelli (2021) reformulates the model 2.2 as follows with a

[1] Morelli (2021) dealt with heavy tails by using a Student's t-distribution as the likelihood.

Student's t-likelihood:

$$\begin{aligned}
y_{di}|\boldsymbol{\beta}, u_d, \sigma_e &\sim \text{Student}(\boldsymbol{x'}_{di}\boldsymbol{\beta} + u_d,\ \sigma_e\ , \nu),\ d = 1, ..., D,\ i = 1, ..., N_d, \\
u_d|\sigma_u &\sim \mathcal{N}(0, \sigma_u),\ d = 1, ..., D, \\
\beta_k &\sim \mathcal{N}(\mu_k, \sigma_k),\ k = 1, ..., K, \\
\sigma_u &\sim Ga(2, a), \\
\sigma_e &\sim Ga(2, b), \\
\nu &\sim Ga(2, 0.1).
\end{aligned} \tag{2.3}$$

The t-distribution has an extra parameter ν, the degrees of freedom, that has an impact on its excess kurtosis and consequently also on the variance. Note that in this case there is only one ν for all areas. Because the excess kurtosis converges to zero as $\nu \to \infty$ (for $\nu \approx 50$ the excess kurtosis is just 0.1), a gamma distribution with shape 2 and 0.1 as the rate parameter is used. This forces ν to be positive and places more weight on the areas of ν for which the t-distribution is leptokurtic, while still allowing values where the likelihood is close to normal (the 95% quantile is just below 50, and for $\nu \geq 50$ there is little difference between a Gaussian and a Student's t-distribution). Note that σ_e is a scale parameter of the Student's t-distribution and is not equal to its standard deviation. The standard deviation of the likelihood is given by the relation $\sigma = \sigma_e \cdot \frac{\nu}{\nu-2}$, which makes clear that the variance is only finite for $\nu > 2$. The gamma distribution is chosen for the scale parameters σ_u and σ_e, as the standard deviation cannot be negative. The shape parameter is set to 2 in line with Gelman (2020), which makes the distribution clearly skewed to the right. a and b are positive constants that define the rate parameters of the gamma distributions. These have to be chosen according to the scale of the dependent variable y_{di} for each specific data set. K is the number of coefficients in the regression, and $k = 1$ is the intercept. Thus, according to 2.3 the prior can in theory be set independently for each coefficient in $\boldsymbol{\beta}$. The exact prior parameters for the coefficients will be discussed in the next sections based on the variables from the data set.

Molina et al. (2014) also present a reparametrized version of model 2.2 with $\rho = \sigma_u(\sigma_u + \sigma_e)^{-1}$ to avoid using MCMC methods[2]. Avoiding MCMC should not be a major concern due to the many developments in the field of Bayesian computation since 2014. While the reparametrized model may simplify estimation under certain circumstances, it comes at the cost of model flexibility. It is not straightforward to imagine how their reparametrized version could be estimated without MCMC after changing the likelihood or prior distributions.

[2]An in-depth explanation of MCMC can be found in appendix A.

Further information on Bayesian computation can be found in appendix A. The next section defines poverty indicators based on income and describes an algorithm to calculate them based on Bayesian models such as 2.3.

2.2.2. Estimating poverty indicators with Bayesian models

As the aim of this paper is to calculate poverty indicators based on income data, this section gives a formal definition of poverty indicators and describes how to estimate them with an HB model. The two main indicators in this paper are the head count ratio (HCR) and poverty gap (PGAP), which are based on the Foster-Greer-Thorbecke (FGT) indicators (Foster et al., 1984):

$$F_d(\alpha, t) = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\frac{t - y_{di}}{t} \right)^{\alpha} I(y_{di} \leq t), \qquad \alpha = 0, 1, 2,$$

where t is the poverty line (60% of median income), y_{di} is the income for the i-th household in area d and $I(\cdot)$ is the indicator function. If $\alpha = 0$, then F_d is the HCR – i.e., the proportion of households below the poverty line in area d. On the other hand, F_d quantifies poverty intensity (PGAP) when $\alpha = 1$, i.e. it measures the amount by which poor people are below the poverty line on average. $\alpha = 2$ defines poverty severity, which will not be considered in this paper.

Based on model 2.3, it is possible to generate synthetic income data form the posterior predictive distribution. The poverty indicators are based on theses synthetic income simulations. This is procedure is similar to Rojas Perilla et al. (2020), but in a Bayesian context. The HB estimator for the poverty indicator F_d is given by the following algorithm, where S denotes the total number of MCMC samples[3]:

[3]For details on MCMC, see appendix A.

Algorithm 1: Estimate FGT-indicators with HB model

Input: A model $p(\theta, y)$, some data y and $\alpha \in \{0, 1, 2\}$
Output: $\hat{F}_d^{HB}, \hat{\sigma}_d^{HB}$, for $d = 1, ..., D$
for $s \in \{1, ..., S\}$ **do**
  for $d \in \{1, ..., D\}$ **do**
    $\tilde{y}_d^{(s)}|y = (\tilde{y}_{d1}^{(s)}, ..., \tilde{y}_{dN_d}^{(s)})'$;
    Sample $\hat{\boldsymbol{\beta}}^{(s)}, \hat{u}_d^{(s)}, \hat{\sigma}_e^{(s)}, \hat{\sigma}_u^{(s)}, \hat{\nu}^{(s)}$ from $p(\beta, u_d, \sigma_e, \sigma_u, \nu|y)$;
    if *d is in-sample* **then**
      Sample $\tilde{y}_d^{(s)}|y$ from Student$(\boldsymbol{x'}_d\hat{\boldsymbol{\beta}}^{(s)} + \hat{u}_d^{(s)}, \hat{\sigma}_e^{(s)}, \hat{\nu}^{(s)})$
    else
      if *d is out-of-sample* **then**
        Sample $\tilde{u}_d^{(s)}$ from $\mathcal{N}(0, \hat{\sigma}_u^{(s)})$;
        Sample $\tilde{y}_d^{(s)}|y$ from Student$(\boldsymbol{x'}_d\hat{\boldsymbol{\beta}}^{(s)} + \tilde{u}_d^{(s)}, \hat{\sigma}_e^{(s)}, \hat{\nu}^{(s)})$
      end
    end
  end
  $\tilde{y}^{(s)} = (y_1^{(s)}, ..., y_D^{(s)})'$;
  $t^{(s)} = 0.6 \cdot median(\tilde{y}^{(s)})$;
  for $d \in \{1, ..., D\}$ **do**
    $$F_d^{(s)}(\alpha, t^{(s)}) = \frac{1}{N_d}\sum_{i=1}^{N_d}\left(\frac{t^{(s)} - \tilde{y}_{di}^{(s)}}{t^{(s)}}\right)^{\alpha} I(\tilde{y}_{di}^{(s)} \leq t^{(s)})$$
  end
end
for $d \in \{1, ..., D\}$ **do**
  $$\hat{F}_d^{HB} = \frac{1}{S}\sum_{s=1}^{S} F_d^{(s)};$$
  $$\hat{\sigma}_d^{HB} = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}\left(F_d^{(s)} - \hat{F}_d^{HB}\right)^2}$$
end

For clarity, the simulated income $\tilde{y}_d^{(s)}|y$ from the posterior predictive distributions is displayed at the area level to avoid an additional loop at the unit level. Strictly speaking, $\tilde{y}_d|y^{(s)} = (\tilde{y}_{d1}^{(s)}, ..., \tilde{y}_{dN_d}^{(s)})'$ is a vector of size N_d, where each component is sampled independently from the corresponding Student's t-distribution. Moreover, the poverty line $t^{(s)}$ is based on $\tilde{y}^{(s)} = (y_1^{(s)}|y, ..., y_d^{(s)}|y)$ – the generated income for *all* areas – not just for the income of area d given by $\tilde{y}_d^{(s)}$. Therefore, there is only one poverty line for each MCMC sample s.

There are two limitations to the algorithm presented in this section. Firstly, it is valid

only for model 2.3. Choosing another likelihood or priors is likely to change the number and meaning of parameters included in θ. Nevertheless, the algorithm provides a blueprint that can be adapted to the corresponding model. Secondly, it does not take into consideration whether the dependent variable is transformed. The assumption is that $\tilde{y}^{(s)}$ is already in the desired scale. In such cases, it is straightforward to add an additional step that applies the backtransformation to the simulations before calculating the FGT-indicators. For brevity, this is not included in the algorithm.

After the discussion on poverty indicators and their estimation with Bayesian models, the next chapter describes in detail the challenges faced when modelling income data and presents the data set and simulated scenarios that will be used throught the paper.

3. Income Data and its Challenges

Before discussing Bayesian workflow in detail, it is necessary to analyze the problem at the core of this paper: how to adequately model income data, which is the basis for estimating poverty indicators. The first section presents reasons why modelling income data is challenging. It also briefly discusses available modelling possibilities and their drawbacks. In the second section, the data set from the Mexican state of Guerrero used throughout this paper is introduced together with the relevant variables. The survey design of the Mexican data is summarized in the third section, which is useful for certain modelling decisions, as will become clear in section 4.5. The last section formulates simulation scenarios that capture the main difficulties of income data, based on the scenarios in Rojas Perilla et al. (2020). The simulations have the advantage that the true data generating process is known, so they can be used to check the results from different models without having to use the data multiple times.

3.1. Difficulties of modelling income data

The estimation of poverty indicators poses two main challenges. First, indicators such as the head count ratio and the poverty gap are non-linear transforms of income. This makes it hard to model them directly as a dependent variable in a regression, because it is necessary to model income first. This leads to the second challenge. Empirically, income is characterized by a right-skewed, unimodel and leptokurtic distribution: most people earn

an average or lower-than-average income, while only a few have very high incomes. Using a linear model as in models 2.2 or 2.3 naively will lead to poor results. There are two possible ways to deal with this characteristic shape of income data. Firstly, a GLM-regression with a skewed distribution (e.g., gamma, skew-normal, lognormal, etc.) can be used to capture the main features of income. Secondly, data-driven transformation can be applied to make income more symmetric or even closer to a normal distribution as in in Rojas Perilla et al. (2020).

Each approach presents its own challenges. Choosing a skewed likelihood for a GLM model is all but straightforward. First, common skewed distributions present the researcher with clear restrictions: the skew-normal distribution has a maximum skewness of 1 and allows negative values, the lognormal implies that its logarithm must follow a normal distribution, the gamma likelihood assumes a fixed ratio between the expected value and the variance, exponential or Pareto distributions have the mode at their minimum. This problem might in principle be alleviated by choosing a distribution with more parameters that allows a higher degree of flexibility, such as a generalized beta distribution. However, this additional flexibility comes at the cost of interpretability and ease of parametrization. In cases where the use of a highly complex likelihood distribution seems necessary, the question arises whether it might be better to use a non-parametric method instead. Moreover, skewed likelihoods can cause problems for the sampling algorithm. HMC[4] can have trouble with the heavy tail to the right of the distribution mode, which can increase the sampling time considerably (Betancourt, 2017).

On the other hand, the approach using data-driven transformations has to choose an appropriate transformation and consider the uncertainty generated by the estimation of transformation parameters. It is not guaranteed that after the transformation the dependent variable will display more desirable properties such as symmetry or closeness to a normal distribution. Moreover, the complexity of backtransformation functions varies depending on the chosen transformation. For a log or log-shift transformation, the backtransformation can be done simply with the exponential function. Nevertheless, more complex and piecewise transformations such as the Box-Cox have less straightforward backtransformation functions. Finally, due to backtransformation it is usually not clear what the impact of the model parameters is in the original scale. For example, a normal distribution in the transformed scale with parameters μ and σ becomes a lognormal distribution when backtransforming with the exponential function. In the backtransformed scale, the mean is a function of both μ and σ given by $e^{\mu+\frac{\sigma^2}{2}}$. However, if the likelihood in the transformed scale is a Student's

[4]Hamiltonian Monte Carlo (HMC) is explained in appendix A.

t-distribution and a data-driven transformation is used, it is difficult to know the exact effect of the likelihood parameters μ, σ and ν on the mean in the backtransformed scale.

There is a last difficulty related to both the skewed likelihood and the data-driven transformations approach. Choosing a likelihood or transformation not only entails an assumption on the dependent variable, but also on the type of model. A skew-normal distribution assumes an additive model, as the mean can be parametrized directly in the original scale. On the other hand, a gamma distribution with the commonly used log-link implies a multiplicative model in the original scale. Similarly, a transformation like a log-shift with an additive likelihood in the transformed scale (e.g., normal distribution) causes the covariates to have a multiplicative effect in the backtransformed scale. In practice, it is not possible to know whether the true data generating process is additive or multiplicative so it is necessary to check the models against both additive and multiplicative simulation scenarios.

To explore how income can be modelled in a Bayesian context, data from the Mexican state of Guerrero as well as simulated scenarios are used to develop and showcase the Bayesian model in chapter 4. The next section introduces the Mexican data set.

3.2. Data from the Mexican state of Guerrero

To explore the estimation of poverty indicators for small areas, this paper uses the 2010 Household Income and Expenditure Survey (*Encuesta Nacional de Ingresos y Gastos de los Hogares – ENIGH*) and the 2010 National Population and Housing Census from Mexico by the National Institute of Statistics and Geography (INEGI). Specifically, both data sets will be used to estimate the head count ratio (HCR) and poverty gap (PGAP) presented in chapter 2.2.2. The state of Guerrero is divided into 81 municipalities of which 40 are in-sample and 41 out-of-sample in the income survey. The census contains 148083 observations after cleaning the data. Income data is provided at the household level and there are 1801 households in the income survey. The Acapulco municipality contains 511 households, almost a third of the survey sample, and the smallest municipality contains only 13 households. The median sample size at the municipality level is 26.5. Most of the state's population consists of impoverished Indians and mestizos. However, there are important tourist destinations in the municipalities of Acapulco and Zihuatanejo on the pacific coast as well as in Taxco de Alarcón in the highlands. Guerrero's economy is based mainly on the primary sector and two fifths of the population live in rural areas (Encyclopaedia Britannica, 2019).

Table 1.: Variables related to head of household.

Variable	Description	Source
Occupation type	Occupation in the primary, secondary, tertiary sector or not employed	`jsector`
Gender	male or female	`jsexo`
Work experience	Years of work experience	`jexp`
Age	Age of head of household	`jedad`

Table 2.: Variables related to household demographics.

Variable	Description	Source
Minors under 16	Presence of minors under 16 years old in household	`id_men`
Percentage of women	Percentage of women in household	`muj_hog / tam_hog`
Literacy	Percentage of literate memebers of household	`nalfab / tam_hog`
Indigenous population	Presence of indigenous population in household	`pob_ind`
Geography	Household in urban or rural area	`rururb`

In an applied setting, the variables that can actually be chosen is limited by a number of factors. For SAE methods, it is necessary to have auxiliary data to borrow strength and improve the estimations, which is only possible if both the main and the auxiliary data sources contain the same variables. Moreover, the amount of missings in a variable can severly undermine its usefulness for prediction. A large number of missings (20% or more) in one or multiple variables can lead to a high number of observations being discarded. While imputation is possible in principle, it is not straightforward to take the clustered and stratified structure of survey data into account in the imputation process. Finally, some variables include missing values due to structural reasons. For example, if the head of household is single then there will be missing values in variables that concern the partner. Imputation in such cases is unrealistic. Therefore, variables with a low amount of missings (<5 %) both in the survey and in the census are chosen as candidates for the final model.

The income variable y_{di} for a given municipality and individual corresponds to `icptc` in the survey and it is not available in the census. `icptc` measures equivalized total household per capita income in Mexican pesos and is used as a proxy for the living standard. To generate high-quality predictions, variables present in both data sets that are plausibly related to income are used as regressors. These variables are grouped into three categories.

Table 3.: Variables related to economic situation.

Variable	Description	Source
Working members	Percentage of working members of household	`pcocup`
Income-receiving members	Percentage of members who receive an income	`pcpering`
Unusual work	Presence of child or senior work in household	`trabinf || trabadulmay`
External income	Household receives remittances or financial help from other households	`remesas || ayuotr`
Communication goods	Number of communication goods per capita in household	`actcom / tam_hog`
General goods	Number of goods per capita in household	`actcom / tam_hog`

Variables related to head of household (Table 1[5]) are likely to have high predictive power, because the head of household is usually the main breadwinner of the household and his/her situation has a large impact on the economic situation of the whole household. Additionally, variables related to household demographics (Table 2) are relevant, because they provide information about socio-demographic patterns that impact household income – e.g., rural areas tend to have lower income than urban areas, or indigenous people are more likely to be marginalized in former colonies. Finally, variables about the economic situation of the household (Table 3) represent economic circumstances, which reflect household income and wealth more directly[6]. These variable are only a starting point and a more careful variable selection is done in section 4.4.

There is a last group of binary indicator variables that are based on the presence of certain structural disadvantages in the household concerning four different areas (Table 4): education, health care, housing quality and access to public utilities. These variables are not used as predictors in the regression, but they will be considered in chapter 4.5 when discussing whether it is possible to have an alternative definition of the random effect in the model.

[5]For practical purposes, the *not employed* category summarizes both the unemployed and individuals out of the labor force.

[6]In the tables, code notation is used for some variables. The symbol `/` indicates division and `||` indicates the Boolean OR. The variable `tam_hog` is the number of members in the household and is used to compute certain quantities per capita to make them more comparable among households of different sizes.

Table 4.: Indicators of structural disadvantages.

Variable	Description	Source
Education	Adequate access to education	`ic_rezedu`
Health care	Adequate access to health care	`ic_asalud`
Housing quality	Adequate housing quality	`ic_cv`
Public utilities	Adequate acces to public utilities (electricity, running water, sewer system)	`ic_sbv`

3.3. Introduction to the survey design

Before presenting the methodology used in this paper, it is crucial to have a rough understanding of the survey design of the data. The following contains a high-level summary of this topic and more details can be found in INEGI (2011), which will become relevant in section 4.5. The MCS module of the ENIGH has a stratified, two-stage, clustered survey design. The main unit in the survey design is called the primary survey unit (PSU), which is a cluster of households. The exact definition of the context varies depending on the context: while in an urban setting the PSU is defined as a grouping of street blocks, in a rural context such a definition is not possible. In general, each PSU can contain between 80 and 300 households.

The stratification is done in multiple stages. The first stage consists of four strata based on socio-economic indicators taken from the census.[7] These four strata contain all PSUs in the country. The second stage corresponds to a geographic stratification. In each federal state, PSUs are stratified according to whether they are in a rural, urban or highly urban area (*ámbito*) with even finer groups inside each one of these categories (*zona*). Note that the stratification *does not* consider any municipal divisions. This fact will play a key role in chapter 4.5.

Two-stage sampling indicates that inside each stratum multiple PSUs are sampled and then, in a second stage, a fixed number of households inside each selected PSU is sampled. This sampling strategy implies that the households are clustered depending on the PSU they are sampled from. There might be deviations from the sampling scheme presented in order to guarantee that certain global properties of the whole sample still hold – e.g., a balanced ratio between the number of women and men – or that there is enough differentiation in a given area. The effect of nonresponse is taken into account in the survey weights. However, in this paper the weights are not taken into account when calculating the final poverty

[7] A complete list of the indicators can be found in the appendix of the MCS documentation (INEGI, 2011).

indicators.[8]

While the clustered structure arising from two-stage sampling can be accounted for when building the model, this paper focuses instead on how to integrate the stratification into the model. Thus, the adequate consideration of the clustered structure in the model is left to future research. The last section of this chapter introduces simulation scenarios that capture the main features of income data.

3.4. Simulation scenarios

Working with simulations is one key step of iterative model improvement in a Bayesian workflow (Gelman et al., 2020). Testing models against synthetic data lets the researcher check and understand potential problems with the methods. While a model that works well with simulated data is not guaranteed to work well with real-world data, a model that does not work with simulated data is certain to fail in a real application. This will become clearer in the next chapter on Bayesian workflow.

As described in section 3.1, income data is characterized by a unimodal, right-skewed and leptokurtic distribution. To mimic these characteristics, three simulation scenarios based on Rojas Perilla et al. (2020) are proposed. The first one – the *log-scale* scenario – is defined so that the logarithm of simulated income is roughly normal:

$$
\begin{aligned}
u_d &\sim \mathcal{N}(0, 0.4), \quad d = 1, ..., D, \\
\varepsilon_{di} &\sim \mathcal{N}(0, 0.3), i = 1, ..., N, \\
\mu_{dk} &\sim \mathcal{U}(2, 3), k = 1, ..., K, \\
\Sigma_{mn} &= \begin{cases} 1, & m = n, \quad m = 1, ..., K, n = 1, ..., K, \\ \rho, & \text{otherwise}, \end{cases} \\
\boldsymbol{x}_{di} &\sim \mathcal{N}(\boldsymbol{\mu}_d, \Sigma), \quad \boldsymbol{\mu}_d = (\mu_{d1}, ..., \mu_{dK}), \\
y_{di} &= \exp(5 + 0.1 \cdot \boldsymbol{x}_{di} + u_d + \varepsilon_{di}),
\end{aligned}
\tag{3.1}
$$

where K is the number of regressors, D is the number of domains, and $N = N_1 + ... + N_D$ is the total number of observations. Σ is the covariance matrix and it assumes that all covariates have unit variance. ρ controls the correlation between independent variables and is set to 0.2. A certain varation among areas is achieved through the vector $\boldsymbol{\mu}_d$, which contains the means for all covariates in area d. Thus, some covariates are consistently higher

[8] The reason for this is that common frequentist SAE packages in `R`, which will be used to compare the EBP to the Bayesian model, do not allow for the inclusion of survey weights in the estimation procedure.

or lower in a given area, creating a clear profile for each area. Because the focus is on prediction, the regression coefficients are fixed and equal for all covariates (e.g., 0.1 in the log-scale scenario), but this is only done for simplicity. There are two additional scenarios – the *Pareto* and the *GB2*. The GB2 scenario can be formulated as

$$\begin{aligned}
u_d &\sim \mathcal{N}(0, 500), \quad d = 1, ..., D, \\
\varepsilon_{di} &\sim \mathcal{GB}2(2.5, 18, 1.46, 1700), i = 1, ..., N, \\
\mu_{dk} &\sim \mathcal{U}(-1, 1), k = 1, ..., K, \\
\boldsymbol{x}_{di} &\sim \mathcal{N}(\boldsymbol{\mu}_d, \Sigma), \quad \boldsymbol{\mu}_d = (\mu_{d1}, ..., \mu_{dK}), \\
\tilde{\varepsilon}_{di} &= \varepsilon_{di} - \bar{\varepsilon}, \\
y_{di} &= 9000 - 250 \cdot \boldsymbol{x}_{di} + u_d + \tilde{\varepsilon}_{di},
\end{aligned} \tag{3.2}$$

where $\mathcal{GB}2$ is the generalized beta distribution of the second kind with four parameters usually referred to as a, b, p, q. Σ is defined as in scenario 3.1 with $\rho = 0.2$. Note that ε_{di} needs to be centered, as they have a non-zero mean. The third and last scenario has a Pareto error term:

$$\begin{aligned}
u_d &\sim \mathcal{N}(0, 500), \quad d = 1, ..., D, \\
\varepsilon_{di} &\sim \text{Pareto}(3, 2000), i = 1, ..., N, \\
\mu_{dk} &\sim \mathcal{U}(-3, 3), k = 1, ..., K, \\
\boldsymbol{x}_{di} &\sim \mathcal{N}(\boldsymbol{\mu}_d, \Sigma), \quad \boldsymbol{\mu}_d = (\mu_{d1}, ..., \mu_{dK}), \\
\tilde{\varepsilon}_{di} &= \varepsilon_{di} - \bar{\varepsilon}, \\
y_{di} &= 12000 - 350 \cdot \boldsymbol{x}_{di} + u_d + \tilde{\varepsilon}_{di}.
\end{aligned} \tag{3.3}$$

Here again, $\tilde{\varepsilon}_{di}$ are the centered unit-level residuals. Note that all fixed parameters in the three scenarios (e.g., the standard deviation of the area-level random effect u_d) are chosen so that the simulations are in a realistic range in line with the income variable `icptc` from the Mexican survey. This means roughly that no simulation should be above the tens of thousands.

The three scenarios provide a variety of characteristics against which to test the models in the workflow. First, the log-scale scenario is multiplicative, while the GB2 and Pareto scenarios are additive. This is useful to check whether a model with a multiplicative likelihood also can approximate a model whose generating process is additive, and vice versa. Moreover, the logarithmic transformation of y_{di} should be roughly normal, while the GB2 and Pareto scenarios should display a higher excess kurtosis after a logarithmic

transform, which is often observed in real-world income data.

Nevertheless, there are two main limitations with the simulation scenarios. Firstly, the area-level intercepts u_d are assumed to be independent, by defining the standard deviation as a constant. This is unrealistic, as there are similarities and differences between areas that are likely to manifest themselves as correlations. A clear example would be two neighboring geographic areas with highly intertwined economies. Secondly, some covariates might have different effect sizes depending on the area. For example, a higher level of education could have a much stronger effect on income in an urban region, where there is a higher demand for specialized, well-paying jobs. In contrast, most jobs in a rural region are less likely to require high skills and therefore offer lower pay. Here, income not only depends on the quality of labor supply, but also on the job market demands in a given area. For the present paper, these two limitations are acceptable. On the one hand, there are many ways of including spatial correlation in the simulation scenarios (see section 4.6) and it is not clear whether a decision for one type of correlation might turn out to be unlike the spatial dependencies present in real-world data. On the other hand, random slopes are not considered in the Bayesian workflow of the next chapter. Extensions to the simulation scenarios such as spatial correlation and random slopes might play an important role in developing a model for poverty estimation, but this question is left for further research.

4. Iterative Bayesian Modelling for Poverty Estimation

This chapter shows how to develop a Bayesian model iteratively to estimate poverty indicators. The main ideas of Bayesian workflow according to Gelman et al. (2020) are summarized in the first section. As the workflow is highly non-linear, the sections afterwards do not correspond exactly to single steps from the workflow. Instead, each section is a model improvement iteration, in which many (but not necessarily all) of the steps in the workflow are followed. The second section presents two plausible alternatives to model skewed, leptokurtic and unimodal variables such as income. GLM models with skewed likelihoods are compared with a data-driven approach similar to Morelli (2021). In the third section, the regularized horseshoe prior is explained, which is then used to select variables with adequate predictive power. The rest of the chapter focuses on the impact of priors on model perfomance. Section four discusses how to define coefficient and variance priors with

the help of prior predictive checks. In section five, the stratification described in chapter 3.3 is used to redefine the areas and consequently the random effect. Section six explores two alternatives to model correlations at the area level: the LKJ prior and the SAR prior. Finally, the last section compares the models using stacking weights.

4.1. An introduction to Bayesian workflow

The idea of a statistical workflow is not new. One early example of a statistical workflow can be found in Box (1976). In the Bayesian context, Gabry et al. (2019) and Betancourt (2020) have referred to the idea of a Bayesian workflow. This section summarizes the ideas behind Bayesian workflow as presented by Gelman et al. (2020).

While Bayesian *inference* deals with the formulation and computation of (conditional) probability densities, Bayesian *workflow* consists of three steps: model building, inference and model checking/improvement. In Bayesian workflow, it is inevitable that a series of models are fit iteratively. Flawed models are a necessary step towards improving the model and finding models that are useful in practice. At the same time, model improvement is not limited to finding the best model, but it also lets us better understand the models used; specifically, why they fail or lead to different results under certain conditions. There are several reasons for considering a workflow and not just plain inference. Bayesian computation is challenging and it is often necessary to iterate through simpler and alternative models, sometimes using faster but less precise approximation algorithms. Moreover, it might not be clear ahead of time which model is adequate and how it can be modified or extended. The relation between fitted models and data can be best understood by comparing inferences from different models. For example, it is possible to gain valuable insights by comparing a model to simpler or more complex models. Finally, there is uncertainty associated with model choice, as different models might lead to diverging but realistic results for the same application.

The graphical representation of the workflow is included in Figure 1. The workflow contains many steps, but the authors emphasize that there are some steps that might be skipped or changed depending on the data and the use case. For this reason, Gelman et al. (2020) argue that a workflow is more general than an example, but less clearly specified than a methodology. While the exact steps are likely to vary depending on the specific application, a workflow provides a framework to develop statistical models. Note that this workflow is mostly focused on data modeling. Other steps such as data collection are not taken into account.

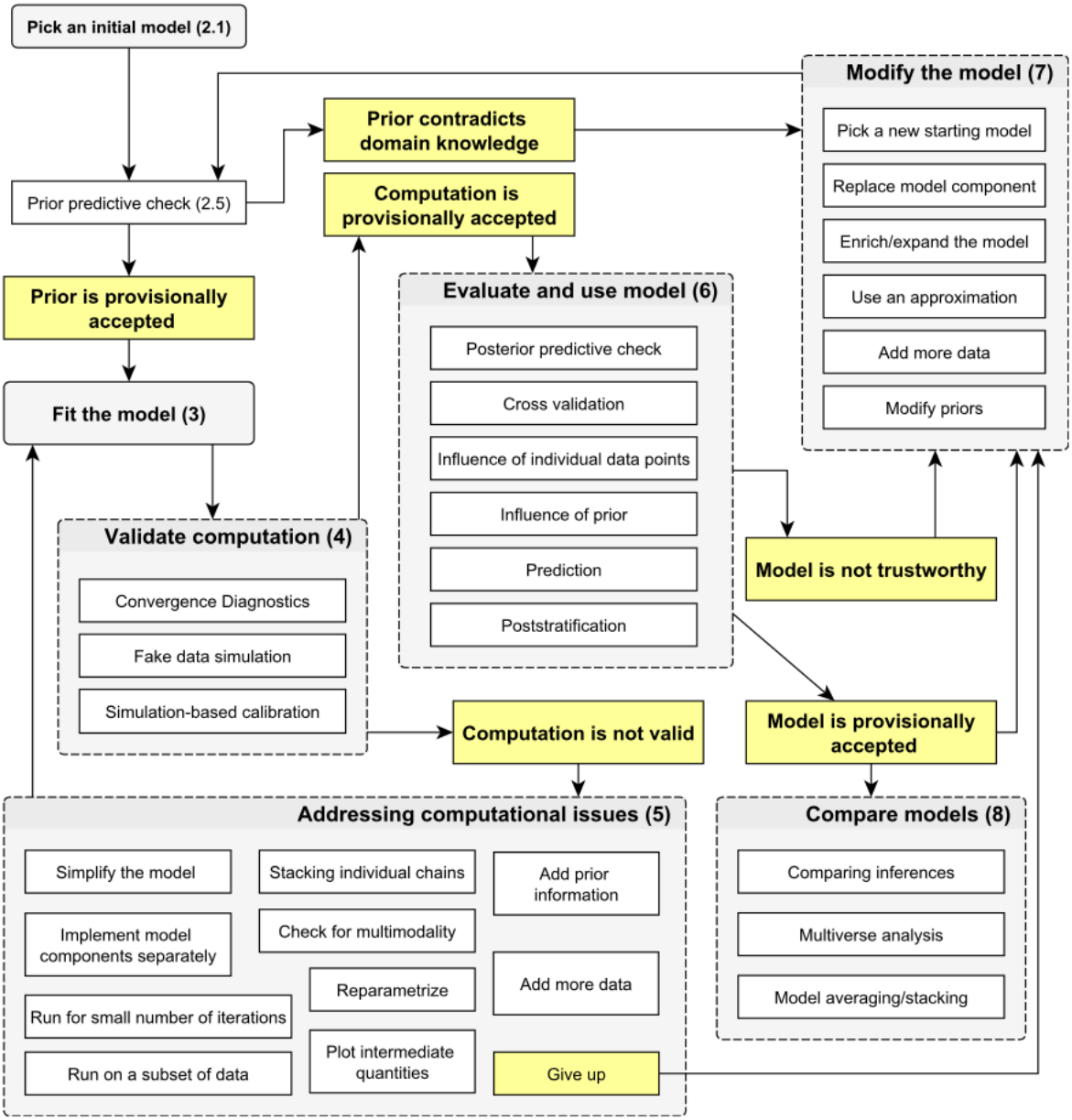


Figure 1.: Workflow from Gelman et al. (2020).

An exhaustive discussion of the whole workflow is beyond the scope of this paper. Instead, the focus is on how the ideas in Gelman et al. (2020) can be applied to the estimation of poverty indicators in the small area estimation context. Due to the non-linear nature of the workflow, the sections and subsections in the current chapter are not strictly named after single steps of the workflow. On the contrary, each section in this chapter reflect specific aspects of a model that are investigated separately. However, throughout the rest of the paper there are explicit references to the corresponding step and an explanation of why it is necessary at a given stage. For completeness, the rest of this section summarizes the seven workflow steps in Gelman et al. (2020) that are also represented in Figure 1.

4.1.1. Step 1: pick an initial model

Usually, the starting point is to adapt an idea that already exists in the literature. This adaptation can be done in different ways: (i) start with a simple model and add layers of complexity, (ii) simplify a complex model, so that it is more understandable or easier to fit while still delivering a similar performance, (iii) consider different starting models with diverging assumptions and follow multiple paths. Bayesian models are highly modular, as priors and likelihood can be replaced with other distributions if necessary. Moreover, there is flexibility regarding how parameter priors interact with each other, which allows for a high degree of model complexity.

For example, in this paper, one initial model is the HB model by Molina et al. (2014) with its extensions in Morelli (2021), where the effect of different likelihoods for log-shift transformed income was explored. The t-distribution provides the best performance over many different scenarios. However, in Morelli (2021) the shift parameter was found through a heuristic whereas the initial model in the present paper does full inference on the shift parameter from the start. Additionally, an alternative group of initial models that use skewed likelihoods instead of transformations will be considered in section 4.2.2.

4.1.2. Step 2: fit the model

Appendix A discusses the two most common Bayesian estimation approaches: MCMC and variational inference. When fitting a model, the user is confronted with decisions on which algorithm to run and under which conditions. To make an adequate choice, it is necessary to be aware of the modelling stage. MCMC provides the most exact approximation, which becomes more robust with more samples and more chains. However, if a model was just modified or the user is at an early stage of model development, it is not efficient to fit

the model the most exact algorithms and a high number of samples with multiple chain. Instead, the fit of a bad model should fail fast, as computational issues usually indicate a deeper problem in the model. Such problems can already be clear with just two Markov chains and a drastically lower number of samples than what would be chosen for the final model. Even a check with an approximate method such as variational inference might be enough to notice problems with the model, while being drastically faster than MCMC. In this paper, variational algorithms were often used to do a first fit of the model. While iterating through the models only two Markov chains were used and, depending on fitting time, the number of iterations was reduced compared to the default 2000 iterations in `Stan` Stan Development Team (2021).

4.1.3. Step 3: validate computation

After fitting the model it is crucial to check whether the computation results are reliable. Diagnostics for Bayesian computation methods are discussed in detail in appendix A. When using an MCMC algorithm such as NUTS, the two main aims are to have no divergences and to have an $\hat{R}$ lower than 1.01. All the models presented in this paper fulfill at least these two conditions, unless otherwise specified. However, adequate diagnostic values are a necessary but not sufficient condition for a reliable model. To assess model quality, the model has to be fitted to some data set. The use of real data can be challenging, as there is no way to distinguish modeling issues from computational issues. This problem can be avoided by using simulated data, ideally from multiple realistic scenarios. Section 3.4, already presented three scenarios used in this paper. A model that can fit fake data, is not necessarily correct, but a model that fails when fitting simulating data will also fail with real data. This paper deals primarily with prediction, so it is enough if samples from the posterior predictive distribution capture the main characteristics of the data. There will be less emphasis on correctly recovering other model parameters from the simulated data.

4.1.4. Step 4: address computational issues

A model that leads to computational problems often has some underlying modelling issues. Usually, it is best to start with a simple model and make it more complex one step at a time, making it easier to determine what part of the model is causing problems. On the other hand, if the model is already complex and shows signs of computational problems, then it is useful to simplify it step by step until the computation is successful. For example, in a model with numerous groups of random intercepts or random slopes it makes sense to

start with just one group and then add each additional group one step at a time, so as to know whether the estimation is working. A common source of computational problems is prior choice. Tightening moderately informative priors can help by pushing the sampler towards certain regions of the parameter space. However, the priors should be adjusted in line with available knowledge and not only to solve fitting problems. One such example is discussed in section 4.2.1.

4.1.5. Step 5: evaluate and use the model

After ensuring that the computation results are reliable, it is necessary to evaluate the quality of the estimated model. Posterior predictive checks are a useful tool when diagnosing fitting problems to the data. This can be seen as a safeguard against misspecification. Moreover, such checks might reveal which aspects of the data are not captured well by the model. Cross-validation is an alternative to posterior predictive checks and has the advantage that part of the data is left out of the fitting process. Thus, it is less optimistic than posterior predictive checks, which uses the data for both model fitting and evaluation. While refittig a model multiple times to do cross-validation can be computationally expensive, there are efficients approximations such as PSIS-LOO (Vehtari et al., 2017). To check how informative the data is with respect to a parameter, one can compare the standard deviation of prior and posterior parameters. A higher shrinkage in uncertainty indicates that the data is more informative. If the model is good enough it can be used in an applied setting. In the small area estimation context, this means generating predictions from the model and calculating poverty indicators.

4.1.6. Step 6: modify the model

Bayesian statistics provides a modular approach in which models can be expanded or reduced in response to new data or failures to fit the model to the data. This is usually done by changing certain aspects of the prior distribution. The prior determines what kind of available information is integrated into the model and acts as a constraint on the fitting procedure. There are various levels of priors from completely non-informative to highly informative. However, the way prior information acts on this information depends on the type of parameter. Parameters controlling central quantities like a mean are less sensitive to weak priors than scale parameters such as variance. In turn, scale parameters are less sensitive to weak priors than shape parameters, which control the tails of a distribution (e.g., the degrees of freedom in a t-distribution). When expanding the model with additional

parameters (e.g., introducing random intercepts or random slopes), it should be considered whether the priors should be tightened to stabilize the estimates, as the amount of data has not changed. Nevertheless, the priors should not be tightened beyond a range compatible with prior information. Note that even if a model is trustworthy, it can still make sense to modify it, e.g., because it is an intermediate model that leads to a more complex model.

4.1.7. Step 7: compare models

Models are fitted many times, for multiple reasons. It might be easier to start with simple models, before getting to a more complex model. There are often bugs in the code and in the models. A model might be well-specified, but it could be improved by being expanded. The priors might only be placeholders, which will be replaced at a later stage. Moreover, multiple models might provide acceptable results. Here, model comparison plays an important role.

Comparing different models is always tied to a certain degree of uncertainty. Instead of choosing the model with the best cross-validation results, using model stacking can give an insight into model differences. Stacking combines inferences using a weighting that minimizes cross-validation error, as discussed in section 2.1.2. At the same time, care must be taken when comparing a large number of models to reduce the risk of overfitting. Therefore, it is useful to select only a couple of models at each stage of the workflow for a comparison.

4.2. Approaches to modelling income

This section compares two different categories of initial models, which corresponds to the first step of the workflow in Gelman et al. (2020). The first one is an extension of model 2.3 from Morelli (2021) that does full Bayesian inference on the shift parameter of the log-shift transform. The second category includes a series of skewed likelihoods, which are tested against the simulation scenarios from section 3.4 through posterior predictive tests. At the end of this section, there is a short discussion on the adequacy of each type of model when dealing with unimodal, skewed, leptokurtic data.

4.2.1. Alternative 1: Data-driven transformations

As discussed by Rojas Perilla et al. (2020), data-driven transformations can change the distribution shape of the dependent variable so that it is more convenient to use simple methods such as a linear mixed model with a Gaussian error term. While Rojas Perilla et

al. (2020) includes a discussion of multiple data-driven transformations such as the Box-Cox or the Yeo-Johnson, the focus here lies on the log-shift transform, as it is the closest to the conceptually simple logarithmic transform. This section starts with a general discussion of the log-shift transform and then shows how to include the transform in the Bayesian model.

Log-shift transformation and skewness

A common transformation for income in economic applications is the natural logarithm. However, there might still be some skewness left, which is exacerbated by very low incomes. If there are a lot of units with incomes between zero and one, there will likely be a long tail to the left side of the transformed distribution. As Rojas Perilla et al. (2020) point out, it is possible to add a fixed term s inside the logarithm so that $y + s \geq 1$ to avoid problems when $0 \leq y \leq 1$, but the transformed variable might still be highly skewed. Moreover, log-income is usually heavy-tailed. This is not surprising, as a variable has to be distributed according to a log-normal distribution for its logarithm to be normally distributed.

To bring the dependent variable closer to a normal distribution, Rojas Perilla et al. (2020) explore different types of data-driven transformations such as Box-Cox or Yeo-Johnson. While effective, these transformations are piecewise functions, which adds an additional layer of complexity when backtransforming to the original scale. Another more simple data-driven transformation described by Rojas Perilla et al. (2020) is the log-shift defined as $y^* = \log(y + \lambda)$, where y is the original variable and λ is the shift term. Although s and λ fulfill similar purposes, they have different meanings: s is a fixed term chosen in advance, whereas λ is a parameter to be estimated. A key advantage is that the backtransformation is straighforward: $y = e^{y^*} - \lambda$. By adjusting λ, it is possible to make the transformed variable more symmetric. However, as Morelli (2021) points out, the transformed variable might still have considerable excess kurtosis, even when it is almost symmetric. Rojas Perilla et al. (2020) point out that minimizing skewness is just one approach to estimating λ. One can also aim to minimize the distance (e.g. Kolomogorov-Smirnov or Cramér-von Mises) to another distribution, usually the Gaussian. Their preferred approach is to maximize the REML of the model under data-transformations. For the method proposed in this section, it is only necessary to minimize skewness to better fit the symmetric likelihood chosen for the model in the transformed scale.

There is a further question related to dependent variable skewness in the linear mixed model. Rojas Perilla et al. (2020) propose to a pooled skewness measure that weights the skewness of the unit and area-level error according to their variances. In principle, it is possible that skewness not only affects the unit-level error ε_{di}, but also the area-level

error u_d. While right-skewness is a common pattern of income at the unit-level (a few individuals/households earn much more than the rest), the picture is less clear at the area-level, as the areas can be defined in very different ways. For example, if areas are defined as municipalities, there is a certain degree of arbitrariness to geographic boundaries: although the Mexican states of Guerrero and Baja California have roughly the same area, the former has 81 municipalities while the latter only has six. Any distribution of u_d will reflect primarily the arbitrary subdivision rather than an underlying economic phenomenon. Therefore, only the skewness at the unit-level errors is considered in this paper. The Bayesian model proposed assumes that u_d follows a normal distribution and is therefore symmetric by definition.

Estimating the shift parameter

A key question is how to estimate the shift term λ. There are two options: estimate it from the data in an empirical Bayes way or do full Bayesian inference. Both approaches have their advantages and disadvantages. When estimating λ from the data, the aim is to reduce skewness as much as possible. This can be done by minimizing the absolute empirical skewness of y^*, or at least bringing it below a predetermined threshold. This approach has the advantage that it is straightforward to control the reduction of skewness in y^*. However, the uncertainty of estimating the shift parameter is not taken into account, as the shift parameter is not included into the model. In practice, the empirical Bayes that minimizes skewness should be in the same range of the full Bayesian estimate, so it can be used as an additional check for the Bayesian model.

Integrating the shift parameter λ naively into the model might lead to estimation problems. Defining the prior directly on λ is not straightforward, as each data set might lead to different results and with no further prior constraints the Markov chains might get stuck and not mix well. As discussed in the previous section, minimizing the skewness of the transformed variable close to zero is likely to improve the performance of a symmetric likelihood distribution. In the Bayesian context, this can be done by including a very tight prior on skewness centered around zero

$$S(y^*) \sim \mathcal{N}(0, \delta).$$

Here, $\delta > 0$ is a small positive constant and S is skewness defined as

$$S(y^*) = \frac{\frac{1}{N}\sum_{i=0}^{N}(y_i^* - \bar{y}^*)^3}{\left[\frac{1}{N-1}\sum_{i=0}^{N}(y_i^* - \bar{y}^*)^2\right]^{3/2}},$$

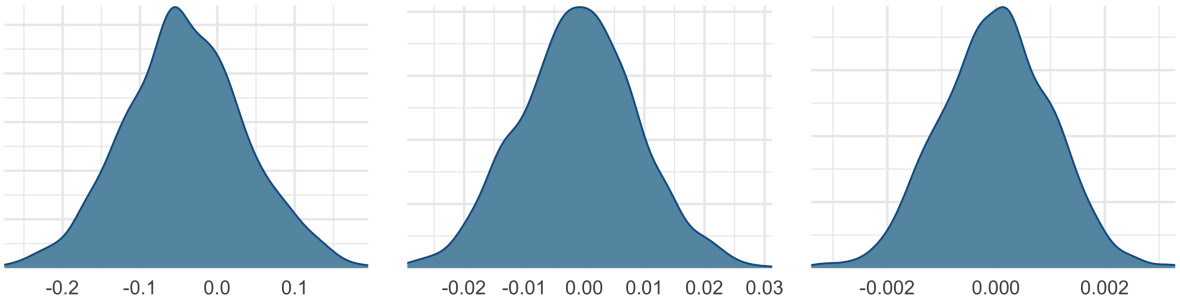
where $y^* = \log(y + \lambda)$ is the transformed dependent variable. Thus, the prior on S indirectly defines a prior on λ. While it is not necessary to formulate a prior directly on λ, it still recommended to define the lower bound of λ as $-\min(y)$ in the programming framework as a safety check, to avoid negative values inside the logarithm function. In practice, the Markov chain steers clear of regions too close to the minimum for λ after warmup iterations.

Because of the transformation, the Jacobian of the likelihood is adjusted by the multiplicative factor $(y_{di} + \lambda)^{-1}$, derived in appendix B. Together with the prior on skewness, this leads to a reformulation of model 2.3 as follows:

$$\begin{aligned}
p(\log(y_{di} + \lambda)|\boldsymbol{\beta}, u_d, \sigma_e, \nu) &= \text{Student}(\log(y_{di} + \lambda)|\boldsymbol{x'}_{di}\boldsymbol{\beta} + u_d,\ \sigma_e\ , \nu) \cdot (y_{di} + \lambda)^{-1}, \\
u_d|\sigma_u &\sim \mathcal{N}(0, \sigma_u),\ d = 1, ..., D, \\
\beta_k &\sim \mathcal{N}(0, 0.5),\ k = 1, ..., K, \\
\sigma_u &\sim Ga(2, 0.75), \\
\sigma_e &\sim Ga(2, 0.75), \\
\nu &\sim Ga(2, 0.1), \\
S(\log(y + \lambda)) &\sim \mathcal{N}(0, \delta),\ \delta > 0
\end{aligned} \tag{4.1}$$

where $d = 1, ..., D,\ i = 1, ..., N_d$. The transformation $\log(y_{di} + \lambda)$ is left explicitly in the model to remind the reader that the prior on skewness directly affects lambda. The prior parameterization of β_k, σ_u and σ_e are taken from Morelli (2021). The lack of index for y inside of S reflects that the skewness function $S(\cdot)$ takes into account all the observations of the target variable – independent of the area.

After a log-shift transformation, the regression coefficients can be interpreted as an approximate percentage change in the original scale. A detailed justification for this interpretation can be found in appendix C. Note that the use of the exponential distribution in the backtransformation together with the heavy-tailed Student's t-distribution can lead to very extreme prediction in the backtransformed scale. Empirically, the percentage of predictions above the dependent variable maximum is between 0.03% and 0.5%. This proportion is very small and should not have a large impact on the poverty indicators, as they depend on the median, which is robust to outliers. Nevertheless, it can cause



(a) $S(y^*) \sim \mathcal{N}(0, 0.1)$ (b) $S(y^*) \sim \mathcal{N}(0, 0.01)$ (c) $S(y^*) \sim \mathcal{N}(0, 0.001)$

Figure 2.: Marginal posterior distribution of skewness $S(y^*)$ for different values of δ in the GB2 scenario: 0.1, 0.01 and 0.001. Note the different scaling of the x-axes.

problems when examining the samples from the posterior predictive distribution. Therefore, all samples that are above the data maximum are taken as missing and are imputed. The imputation procedure is explained in detail in appendix D.

The hyperparameter δ controls the deviation from the zero skewness constraint. Values around 10^{-2} worked well in simulation experiments and the posterior values for $S(y^*)$ are usually very close to zero. Nevertheless, there are two potential problems to be aware of. If δ is too small, then skewness might be reduced too much. This would be equivalent to overfitting the skewness of the training data, which by no means reflects the skewness of out-of-sample data. However, if δ is too large, then many unrealistic values for λ will be allowed, which might lead to poorly mixing Markov chains with an $\hat{R}$ higher than the recommended 1.01. By plotting the posterior density of $S(y^*)$ and using Bayesian diagnostic tools such as posterior checks, it is possible to assess whether δ is too small or too large.

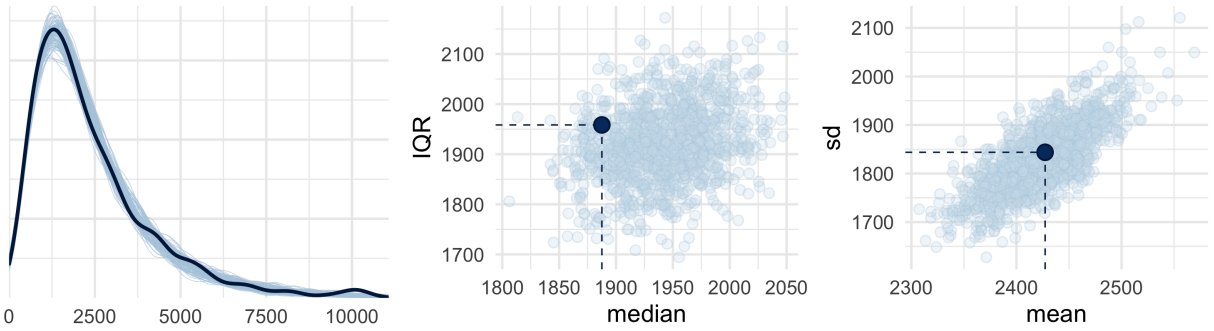
Figure 2 shows the posterior density plots of $S(y^*)$ for the GB2 scenario. The other scenarios displayed similar results. For $\delta = 0.1$, the mode of the posterior is below zero, which indicates that there is a systematic bias towards negative skewness. This is problematic when using a symmetric likelihood. On the other extreme, $\delta = 0.001$ produces an extremely tight distribution around zero. The range of posterior values for λ is very small in this specification, which increases the risk of overfitting the training data. The middle density graph ($\delta = 0.01$) is a compromise between the two extreme scenarios. The posterior distribution for $S(y^*)$ is clearly centered around zero, but the posterior is not as tight as with $\delta = 0.001$. The standard deviation of the shift parameter λ is around an order of magnitude larger than in the model with the tighter prior, which allows more possible models and prevents overfitting. Moreover, the elpd_{loo} has the highest value with the prior

$S(y^*) \sim \mathcal{N}(0, 0.01)$, which indicates that this is the most adequate specification for the skewness.

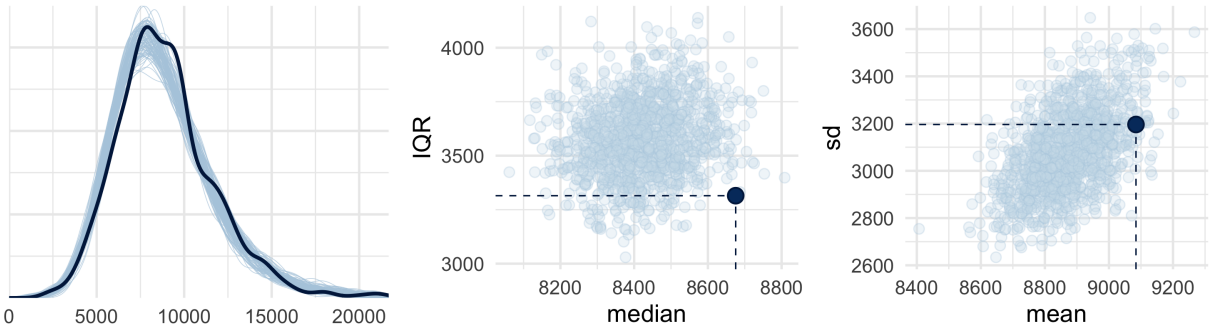
Posterior predictive checks with simulated data

The results from fitting the modified model 4.1 with the log-shift transformation are presented in Figure 3. In the first column, the density of the dependent variable is overlaid with the densities from 100 samples from the posterior predictive distribution. From this first visual check, it is clear that the model is able to fit the log-scale and GB2 scenarios reasonably well. The Pareto scenario is still quite close, but the predictions from the model are slightly lower than the dependent variable. This is an indication that the Pareto scenario contains elements that are particularly challenging to capture.

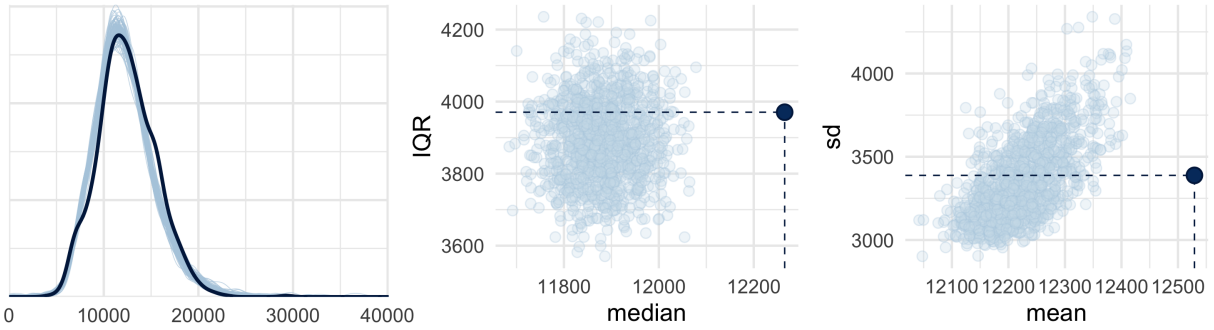
The other two columns show two pairs of descriptive statistics as scatterplots: IQR against median and standard deviation against the mean. The scatterplots also display the respective descriptive statistic value for the dependent variable. In the log-scale and GB2 scenarios, the median of the simulations is in a similar range to the median of the dependent variable. In contrast, the predictions from the model have a median that is around 400 units lower than the dependent variable of the Pareto scenario. This would mean that the simulated poverty lines (ca. 60% of 11.900) are systematically lower than the poverty line from the data (ca. 60% of 12.300) – approximately 3%. Thus, certain poverty indicators are likely to be underestimated. On the other hand, all samples from the posterior predictive distribution in the three scenarios are in a plausible range compared to the IQR of the dependent variable. The mean and the standard deviation are presented as an additional check, but they are somewhat less trustworthy due to their lack of robustness. In the logscale scenario, both the mean and the standard deviation are captured well, whereas in the GB2 scenario the mean and standard deviation of the predictions are a little lower than the dependent variable. The standard deviation Pareto scenario is modelled well, but the mean of the dependent variable is higher than the predictions. A noteworthy pattern is that in all three scenarios there is a positive correlation between the mean and the standard deviation. Although the probability density for the backtransformation is not available in its analytical form, it is a clear indication that the mean of the implied distribution is coupled to the variance in the backtransformed scale – a pattern also found in distributions like the lognormal. The next section explores the performance of models with skewed likelihoods.



(a) Log-scale



(b) GB2

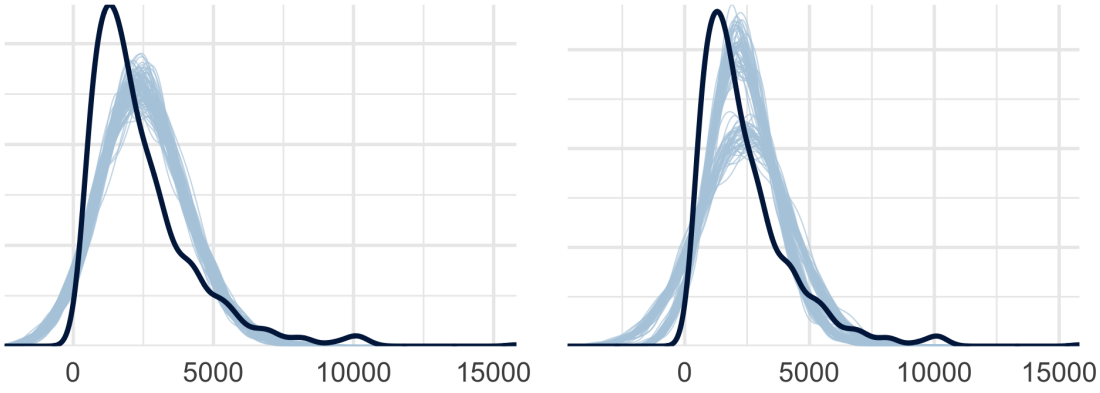


(c) Pareto

Figure 3.: Posterior predictive check for the log-shift model with all three simulation scenarios. *Left:* density of the dependent variable (black) against the density of a 100 backtransformed predictions (light blue). *Middle:* scatterplot of IQR against median for 1000 samples. *Right:* scatterplot of standard deviation against mean for 1000 samples. In the middle and right columns, the dark point represents the respective values for the dependent variable in the original data set.

4.2.2. Alternative 2: Skewed likelihoods

A natural question raised by income data is whether a skewed likelihood might provide the best results. In line with the Bayesian workflow proposed by Gelman et al. (2020), this paper is not limited to a single initial model based on data-driven transformations, but also explores the impact of using skewed likelihoods. The following skewed likelihoods were taken into account: gamma with logarithmic link, gamma with softplus link, lognormal, skew-normal and exponentially modified Gaussian (exGaussian). This variety goes beyond the distribution shape, but it also affects the assumed relation between dependent variable and predictors. The gamma with a log link and lognormal distributions implies a multiplicative model, whereas the rest assume that the predictors have an additive impact on the dependent variable.



(a) Skew-normal (b) (b) exGaussian

Figure 4.: Posterior predictive check for skew-normal and exGaussian likelihoods in the log-scale scenario. The black line is the density of the dependent variable, while the blue lines represent 100 draws from the predictive posterior distribution. Neither the skew-normal nor the exGaussian likelihood provide an adequate approximation. Moreover, the chains in the exGaussian model have trouble mixing, which is clearly seen in the two diverging densities of posterior predictive samples.

The results of these alternative models are presented in this section. The package `brms` (Bürkner, 2017), which provides a user-friendly interface to `Stan`, is used to do a first check of the likelihoods. This follows a principle from Gelman et al. (2020): when working with preliminary models, it is better to use tools that allow for quick checks. `brms` already has a number of likelihoods implemented, and if modifications are needed, it is always possible to take the `Stan` code and modify it. But starting coding from scratch with `Stan` would take longer, especially because working with different likelihoods, requires different parametrizations.

Some of these likelihoods proved to be very poor fits. The skew-normal likelihood captures the main characteristics of the GB2 scenario well. On the other hand, it is not adequate for the logscale scenario, which is characterized by a much higher skewness, as shown in Figure 4a. This is due to the limitations of the skew-normal distribution, as it has a maximum skewness of ± 1. Moreover, in this scenario the model produces some negative predictions. Because this likelihood is unreliable for the log-scale scenario, the model with a skew-normal likelihood is not further developed. The exGaussian likelihood also produces negative predictions in some scenarios and it is not skewed enough for the log-scale scenario (Figure 4b). The graph shows that the chains do not mix well, with two different posterior predictive distributions as a result. Neither placing a more restrictive prior on the coefficients nor decreasing the step size of the NUTS algortihm did alleviate the convergence problem. Nevertheless, even if the MCMC chains mix well, Figure 4b suggests that the exGaussian function is unlikely to be skewed enough to capture the dependent variable from the log-scale scenario. Therefore, this version of the likelihood is also discarded.

The gamma likelihood with a logarithmic link (Figure 11 in appendix E) performs quite well on all three scenarios. Only in the GB2 scenario, the posterior predictive draws are somewhat flatter than the dependent variable. While the median is captured well in all three scenarios, the IQR is clearly higher in the GB2 scenario and somewhat higher in the Pareto scenario. The mean and the standard deviation of the dependent variable and the posterior predictive samples are quite similar, which is not a suprise, due to the parameterization of the gamma likelihood. Note the correlation between mean and standard deviation, due to the fact that the coefficient of variation (σ^2/μ) is always equal to the shape parameter of the gamma distribution. More details can be found in appendix E. A gamma likelihood with a softplus link function (not shown) produced extremely small predictions, far away from the original order of magnitude and is therefore not taken into consideration. An additional check is provided by a lognormal likelihood (Figure 12 in appendix E), which performs well in all three scenarios according to the posterior predictive check. It is not suprising that the posterior predictive checks for the lognormal likelihood and the gamma with a logarithmic link are almost identical, due to the similarity of their shapes.

There are other skewed distributions, for which the mode is equal to the minimum. Some examples include the exponential, the Pareto, the chi-squared and the half-normal distributions. Both the exponential and the chi-squared are special cases of the gamma distribution, which makes them less flexible as likelihood than using a plain gamma distribution. The half normal has very thin tails and a very low skewness that does not correspond to the long tails observed empirically in income data. Therefore, they are not taken into account

as a likelihood.

In summary, based on the posterior predictive checks, only the lognormal likelihood and the gamma with log link worked well for all three scenarios. The next section briefly discusses the advantages and disadvantages of the two types of initial models considered: a model with a log-shift data-driven transformation and a model with a skewed likelihood.

4.2.3. Initial model comparison

In principle, it is possible to keep developing the two alternative models in parallel and then compare them at the end of the workflow, but this implies exploring additional dimension of complexity in the model space. Therefore, one of both alternatives should be chosen based on the flexibility and limitations of each approach.

The log-shift scenario has two main disadvantages. By using a log-shift transformation, the user does not exactly know the analytical form of the density in the backtransformed scale. Moreover, the medians of the posterior predicitve samples in the Pareto scenario are relatively close to the data, but still systematically lower. Another problem is represented by the very extreme predictions that can arise due to backtransforming samples from the heavy-tailed Student's t-distribution.

With respect to the skewed likelihoods, two similar distributions performed well in all three scenarios: the lognormal and the gamma with a log link. However, these two distribution have theoretical limitations. On the one hand, the logarithm of a lognormal variable has to be normally distributed, which is rarely the case for income data. For this reason, it is not deemed to be a realistic alternative. On the other hand, a gamma distribution has a constant coefficient of variation, which depends on the shape parameter. Additionally, the logarithm of a gamma distribution is either left-skewed or symmetric, which might not necessarily be realistic for income data.

Whereas the gamma and lognormal distributions only have two parameters (shape/rate or mean/variance, respectively), the main advantage of the log-shift model 4.1 is that it offers more flexibility due to the additional parameters. The Student's t-distribution is not only parametrized in terms of mean and scale, but also of degrees of freedom. Besides, the log-shift distribution adds another parameter λ to the likelihood through the transformed dependent variable. This flexibility does not come at a high cost of interpretability that characterizes more complex distributions. While the less than ideal performance in the Pareto scenario is problematic, there are two factors which should be considered. Firstly, although the prior parametrization was taken mostly from Morelli (2021), this can be improved through the use of prior predictive checks (see chapter 4.3). Secondly, in this

scenario the degrees of freedom were below two, which is not ideal, due to the fact that the Student's t-distribution has an infinite variance for $\nu \leq 2$ and an undefined variance for $\nu \leq 1$. This extreme behavior might cause problems in the Pareto scenario. There are two simple solutions to this problem. Either ν is set to a constant (e.g., a value between 2 and 3) when the posterior of ν is clearly below 2, or the model is reparametrized so that ν does not take values below zero.

In light of this discussion, the log-shift model 4.1 has been chosen for the rest of this paper, as it provides a good balance between interpretability, ease of parametrization and flexibility. Additionally, it extends the literature on data-driven transformation in the line of Rojas Perilla et al. (2020) to the Bayesian paradigm. This does not mean that the use of skewed likelihoods is an inferior approach, only that its use with income data is left for future research. A feature of the gamma likelihood that can be useful is that its mean can be parametrized in the original scale even when using a log link as done in the `brms` package. This is particularly relevant, if benchmarking (Pfeffermann, 2013) should be included as a module in the Bayesian model. For example, a somewhat narrow prior can be placed on the mean of all domain estimates so that it roughly matches the benchmarking mean based on the direct estimates. The integration of benchmarking into the model is not further considered in this paper.

4.3. Specification of coefficient and variance priors

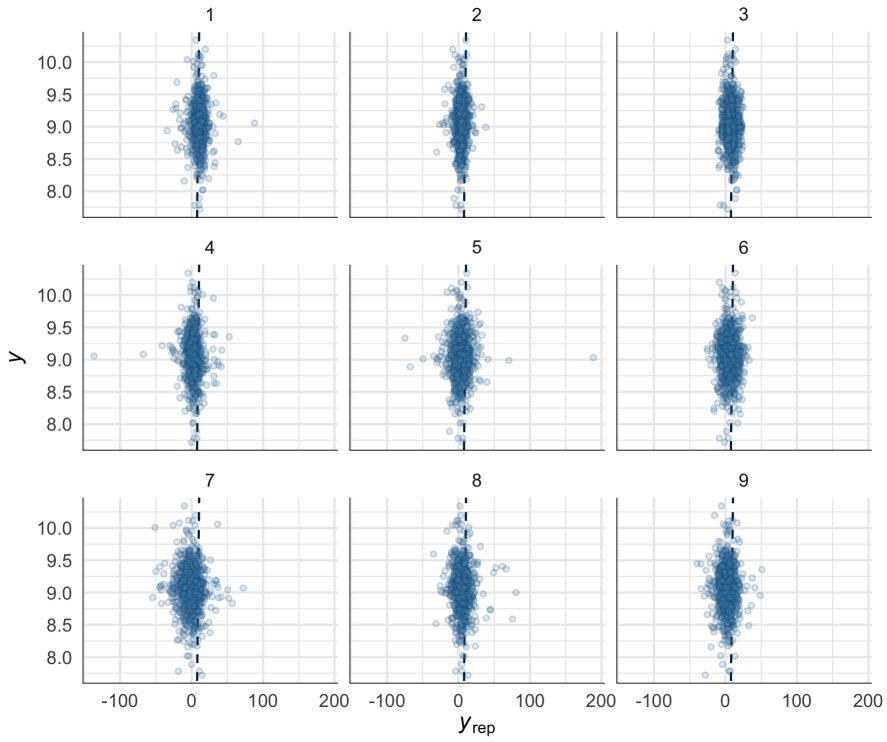
After choosing the Student's t-likelihood with a log-shift transformation, the next step reevaluates the plausibility of the priors introduced in model 4.1 – the extension from the model in Morelli (2021). This section presents the first iterative improvement of the model. For this modification of the model (step 6 in the workflow), prior predictive checks are made with the simulation scenarios to avoid using the data multiple times.

Appendix C shows that even with a shift term in the logarithmic transformation, the coefficients can be interpreted as the approximate percentage change of the dependent variable in the original scale. This approximation holds mostly for coefficients between -0.3 and 0.3. However, a change of around 30% in the original scale for each additional unit is itself very high and this observation can be included in the prior distribution. Moreover, note that through the use of the logarithmic transform the covariates have a multiplicative effect on the original scale. Thus, somewhat large coefficients in the logarithmic scale can have a huge impact when backtransforming to the original scale. For the independent priors, two different types of distribution can be considered, either a normal distribution

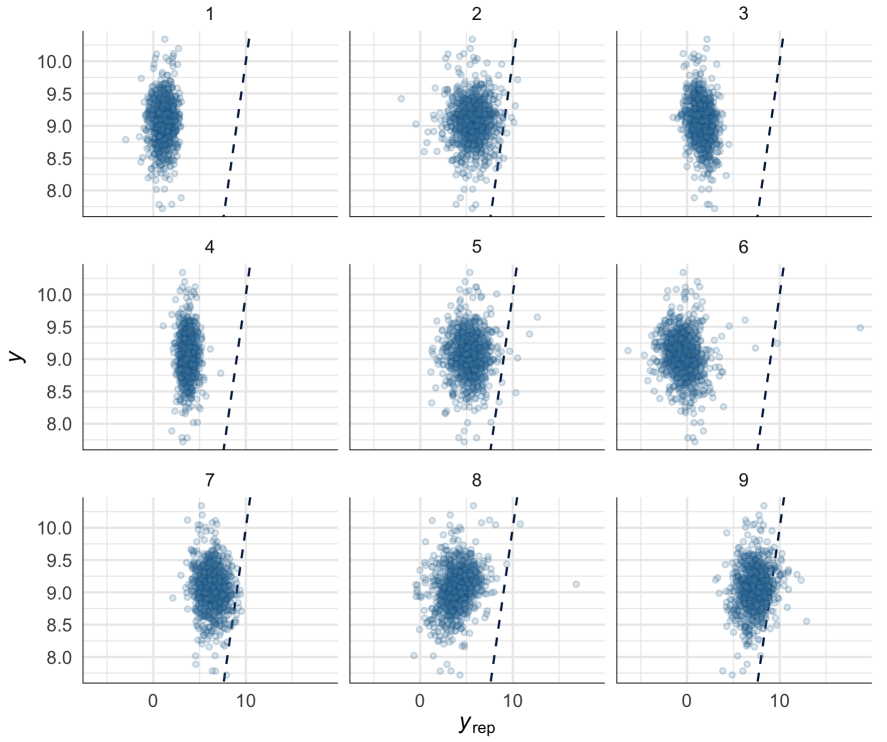
or a heavy-tailed distribution such as the Student's t with 3 degrees of freedom. In this case, a prior distribution with heavy tails is not as desirable, because even if most of the probability mass is contained in the interval between -0.3 and 0.3, the prior would allow more extreme values than a thinner-tailed distribution. This can be avoided by choosing a Gaussian distribution, which does not place as much probability mass on extreme values. The coefficient prior is parametrized to have a mean of zero and a standard deviation of 0.2, which implies that the 5% and 95% quantiles are around -0.3 and 0.3 respectively. To ensure that the samples from the prior predictive distribution are in a realistic range, the prior for the intercept is set to a relatively $\mathcal{N}(4, 3)$. In the final model, the intercept is fit to the data and the prior can be less informative, e.g., $\mathcal{N}(0, 5)$.

Another key element of the models is how the priors are defined for the standard deviations at the unit-level (σ_e) and area-level (σ_u), as these have a large impact on the implied dispersion in the dependent variable. As the likelihood is a generalized Student's t-distribution, the parameter σ_e cannot be interpreted directly as the standard deviation of the distribution. Given a random variable Y that follows this distribution, the variance is given by $Var(Y) = \sigma_e^2 \frac{\nu}{\nu-2}$. To reason more easily about the unit-level variance a new parameter $\sigma = \sqrt{Var(Y)}$ is introduced so that $\sigma_e = \sigma\sqrt{\frac{\nu-2}{\nu}}$. The effect of different values for the rate parameter of the gamma prior on σ and σ_e is explored in the prior predictive checks.

For the prior predictive checks, the shift term is dropped. The shift term depends on the dependent variable so it is not possible to sample from its prior distribution. Besides, it amounts to an additive constant in the backtransformed scale ($e^{y^*} - \lambda$). The impact of the exponential function is therefore much larger than the effect of the shift term λ. The range of Mexican income data is in the tens of thousands, which is also captured in the simulation scenarios. Therefore, the prior predictive simulations should not be much higher than 12 in the logarithmic scale (which corresponds to around 160000 pesos in the original scale). Conversely, as the real data has almost no observations below 1, the values in the logarithmic scale should not go far below zero. Finally, note that the degrees of freedom in the Student's t-distribution can lead to very extreme simulations when ν is low. As a consequence, the prior for ν is changed to $Ga(2, 2)$ *only* for the prior predictive check to focus on the impact of heavier tails. The model is reparametrized so that the minimum value for ν is 2. This ensures that the variance of the distribution is still finite. The assumption here is that if the prior predictive simulations are not too extreme for low degrees of freedom, they will also be in a reasonable range for high values of ν, for which the likelihood has thinner tails.



(a) Scale prior: $Ga(2, 0.75)$



(b) Scale prior: $Ga(2, 7)$

Figure 5.: Prior predictive checks for scale parameters σ and σ_u in the GB2 scenario. In the scatterplots, the logarithm dependent variable is plotted against the corresponding sample from the prior predictive distribution. Note the different scaling of the x-axes.

The results for the prior predictive check in the GB2 scenario can be seen in Figure 5. Other scenarios produced very similar results. The prior predictive checks are shown in form of scatterplots, where the dependent variable is plotted against samples from the prior predictive distributions. These samples do not have to fit perfectly the dependent variable, but they should be in a realistic range. The data is shown in the logarithmic scale. The aim is to determine which value of the rate parameter is best for both scale parameters σ and σ_u. Figure 5a shows the scatterplots for a rate parameter value equal to 0.75, while Figure 5b displays the results for a tighter rate value of 7. Note the different scaling of the x-axes. For a rate parameter of 0.75, there are samples that are extremely high, above 50 in the logarithmic scale. On the other hand, a rate parameter of 7, which implies a much tighter prior is much more realistic – especially because only very few samples are much higher than 10 and even samples with higher values are not as high as in the first specification. Prior predictive checks with a very wide scale prior can be found in appendix F.

After the prior predictive check, it is possible to conclude that the tighter scale prior provides the most realistic results, while still allowing for a few higher than expected but not extreme predictions. Model 4.1 can thus be reformulated as:

$$
\begin{aligned}
p(\log(y_{di}+\lambda)|\boldsymbol{\beta}, u_d, \sigma_e, \nu) = \text{Student}(\log(y_{di}+\lambda)|\boldsymbol{x'}_{di}\boldsymbol{\beta} + u_d,\ \sigma_e\ , \nu) \cdot (y_{di}+\lambda)^{-1}, \\
u_d|\sigma_u \sim \mathcal{N}(0, \sigma_u),\ d = 1, ..., D, \\
\beta_0 \sim \mathcal{N}(0, 5), \\
\beta_k \sim \mathcal{N}(0, 0.2),\ k = 1, ..., K, \\
\tilde{\nu} \sim Ga(2, 0.1), \\
\nu = \tilde{\nu} + 2, \\
\sigma_u \sim Ga(2, 7), \\
\sigma \sim Ga(2, 7), \\
\sigma_e = \sigma\sqrt{\frac{\nu-2}{\nu}}, \\
S(\log(y+\lambda)) \sim \mathcal{N}(0, 0.01),
\end{aligned}
\tag{4.2}
$$

where $d = 1, ..., D,\ i = 1, ..., N_d$. The parameter σ is now explicitly included into the model and a new parameter $\tilde{\nu}$ is introduced to enforce that ν does not take values below 2. The intercept is now explicitly included into the model as β_0.

4.4. Variable selection

A key step in specifying the model is deciding which variables to use. However, this is at the same time one of the more challenging steps. The ideal solution would be to estimate models with all possible variable combinations and then compare the predictive quality of the models – e.g., through cross-validaton. However, this solution is not feasible computationally due to the very large number of possible models. Another common approach is to use shrinkage methods like the Lasso (Tibshirani, 1996) to determine which variables can be left out without significantly compromising the predictive power of the model. If a coefficient is shrunk to zero, it is a sign that the respective variable might be removed. In this section, the regularized horseshoe prior (Piironen & Vehtari, 2017b), which follows a shrinkage principle similar to the Lasso or the ridge, is presented and then used to select relevant variables. This section shows how priors can be used inside a Bayesian workflow to do variable selection. The selection is done by comparing the elpd_{loo} of models with different groups of less meaningful variables removed. An alternative to the horseshoe prior is to use the projective prediction approach developed by Piironen et al. (2020). Unfortunately, this only works with distributions from the exponential family, which does not include the Student's t-distribution used in this paper. Therefore, the focus lies on the regularized horseshoe prior.

4.4.1. Regularized horseshoe prior

In the Bayesian framework, shrinkage can be best understood in relation to the prior. The general idea is to have a narrow region of high density around zero that shrinks coefficients, while at the same time including fat tails to allow some coefficients to deviate from zero over a wider range. In its most extreme formulation, it corresponds to the spike-and-slab prior introduced originally by Mitchell & Beauchamp (1988). While this prior has a high theoretical relevance, it can be computationally demanding for a large number of variables and is sensitive to specification details such as slab width. On the other hand, continuous shrinkage priors provide similar results while also being easier to compute (Piironen & Vehtari, 2017b). The horseshoe prior (Carvalho et al., 2010) is a continuous shrinkage prior with a similar performance to the spike-and-slab prior and can be formulated as

$$\begin{aligned} \theta_j|\lambda_j, \tau &\sim \mathcal{N}(0, \tau^2\lambda_j^2), \\ \lambda_j &\sim \mathcal{C}^+(0, 1), \quad j = 1, ..., J, \end{aligned} \tag{4.3}$$

where θ_j are the parameters to be shrunk and $\mathcal{C}^+$ is the half-Cauchy distribution. In a regression setting, θ_j corresponds to the coefficients β_j. This prior is characterized by a global parameter τ that shrinks all parameters towards zero as $\tau \to 0$ and a local parameter λ_j for each θ_j that pushes against the global shrinkage. Due to the extremely heavy tails of $\mathcal{C}^+(0,1)$, some λ_j will be large enough to counteract even small values of τ.

However, there is no consensus on how to do inference for τ and parameters far from zero are not shrunk at all. Piironen & Vehtari (2017b) address this issue by introducing the regularized horseshoe prior, which is given by

$$\begin{gathered} \theta_j|\lambda_j,\tau \sim \mathcal{N}(0,\tau^2\tilde{\lambda}_j^2), \quad \tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{c^2+\tau^2\lambda_j^2} \\ \lambda_j \sim \mathcal{C}^+(0,1), \quad j=1,...,J, \end{gathered} \tag{4.4}$$

where $c > 0$ is a constant. If $\tau^2\lambda_j^2 \ll c^2$, then θ_j is shrunk strongly towards zero and $\tilde{\lambda}_j^2 \to \lambda_j^2$, which approximates the original horseshoe in expression 4.3. For the case $\tau^2\lambda_j^2 \gg c^2$, then $\tilde{\lambda}_j^2 \to c^2/\tau^2$ and the prior reduces to $\mathcal{N}(0,c^2)$, which corresponds to a Gaussian slab with variance c^2 that regularizes θ_j. When $c \to \infty$ the regularizing effect of the prior disappears, just as in the original horseshoe. Although c can be chosen as a constant, Piironen & Vehtari (2017b) recommend doing inference with the prior

$$c^2 \sim \text{Inv-Gamma}(\alpha,\beta), \quad \alpha = \nu/2, \quad \beta = \nu s^2/2,$$

which leads to a Student$(0,s^2,\nu)$ slab for θ_j far away from zero. This is desirable as the heavier tails of the Student's t distribution prevent too much probability mass accumulating near zero.

The last parameter in the specification of the regularized horseshoe is τ. For a linear model, Piironen & Vehtari (2017b) explore different specifications for τ. Defining p_0 as the prior guess of the number of non-zero variables, the prior for τ should have most of its mass near

$$\tau_0 = \frac{p_0}{J-p_0}\frac{\sigma}{\sqrt{n}}, \tag{4.5}$$

where σ is the variance of the error term[9]. However, there is no best choice for the prior of τ. The three main options are setting $\tau = \tau_0$, or assuming either $\tau \sim \mathcal{C}^+(0,\tau_0^2)$ or $\tau \sim \mathcal{N}^+(0,\tau_0^2)$, where $\mathcal{N}^+$ is the half-normal distribution. In practice, it is common to set

[9]In Piironen & Vehtari (2017b), the assumed linear model does not have a random intercept. Thus the randomness comes mostly from the unit-level error term $\varepsilon \sim \mathcal{N}(0,\sigma)$.

the ratio of relevant and irrelevant parameters, – $p_0/(J - p_0)$ –, and use $\tau = \tau_0$. In this paper, a ratio of 0.2 is chosen, which assumes that for every 10 irrelevant regressors, there are 2 relevant variables. Note that the definition of τ_0 in equation 4.5 is derived under the assumption of a Gaussian model. Therefore, its meaning is only an approximation when using the Student's t-distribution.

Certain shrinkage methods like Lasso force certain coefficients to be zero, which provides an easy selection criterion. However, a disadvantage of the horseshoe is that it does not have an integrated criterion to select variables. There are two possible workarounds for this problem. Firstly, one could leave out variables whose mean coefficient is below a certain threshold (Piironen & Vehtari, 2017b). Secondly, one can leave out variables if a significant portion of the coefficient distribution is in both a positive and a negative area. To check that the predictive power of the model does not deteriorate by removing variables, PSIS-LOO can be used to compare the full model with the smaller model.

4.4.2. Results of variable selection and further considerations

The package `brms` was used to avoid implementing the horseshoe prior from scratch. As the package does not provide data-driven transformations, the model is estimated with a simple logarithmic transformation of income and a Student's t-likelihood. While this does not correspond exactly to the log-shift model presented in the previous section, it can be seen as an approximation. The shift term is unlikely to have a major impact on the predictive power of the regressors. This can be clearly seen in the backtransformation ($e^{\eta} - \lambda$, with η the linear predictor), where the shift term is just an additive constant shifting e^{η}. The original horseshoe prior does not take into account the grouping of categorical variables. In this case, if any of the categories has a clear effect on the dependent variable, the whole variable is considered to be relevant.

After estimating the full model with the horseshoe prior, the following variables presented in section 3.2 have either very small effects or effects that are not clearly positive or negative: job experience of head of household, gender of head of household, presence of minors under the age of 16, percentage of working household members, percentage of women in household and percentage of literate members of household. The density plots of the coefficients can be found in Figure 14 in appendix G. A reduced model without these six variables but with the same prior and likelihood specification as the full model was estimated in a second step, but the difference in elpd_{loo} to the full model is -17.8 with a standard error for the difference of 7.3. The difference is therefore more than twice as large as its standard error, so it reduces the predictive power of the model. A second model with only gender of head

of household, percentage of women in household and percentage of literate members of household removed is indistinguishable from the full model with an elpd_{loo} difference of -2.7 and a difference standard error of 4.2. Therefore, the variables selected in the second model are used for the rest of the paper.

Before moving to the next section, two issues need to be addressed. First, the effectiveness of the horseshoe prior depends crucially on the number of variables included in the full model. The shrinkage will be most noticeable with a large number of variables. Otherwise, the shrinkage from the horseshoe is very similar to placing a prior centered around zero for the coefficients. In this project, ca. 20 variables were chosen from the survey and census data. Therefore, the shrinkage of the regularized horseshoe is not much different from the shrinkage of a non-structured prior (e.g., independent normal priors on the regression coefficients). In any case, this section showed that it is possible to select variables using the horseshoe prior and that its shrinkage properties are particularly useful in high-dimensional data sets.

Finally, finding an adequate subset of variables is a field of active research in the context of Bayesian workflow. The projective predictive distribution approach provides a clear selection of variables that keeps the predictive power of the full model, but at the moment of writing only available for the exponential family (Piironen & Vehtari, 2017b). As the t-distribution is one of the likelihoods used in this paper, no projective predicitve methods are used in this paper. It is very likely that in the near future there will be new tools to do this step of the workflow, especially concerning the projective predictive method to distributions outside the exponential family.

4.5. Specification of the random effect

In applications of small area estimation methods, the random effect u_d is usually defined in terms of the domain for which the indicators should be calculated. For example, Rojas Perilla et al. (2020) use a random effect at the municipality level in the model, as the poverty indicators of interest have to be calculated for each municipality. However, this approach poses two problems. Firstly, the information contained in the survey design might be overlooked. There is no guarantee that the random effect is meaningful from the perspective of how the sample was constructed. Secondly, it might lead to very sparse areas with lots of out-of-sample areas. In this section, the traditional way of defining the random effect is compared to an approach that approximates the stratification in the survey design as an

additional iteration of the model[10].

Section 3.3 included a short introduction to the survey design of the data from the Mexican state of Guerrero. On closer inspection, municipalities play no role whatsoever in the sampling process of the households. Instead, the population is stratified by federal state, by geographic region (urban or rural) and also by socio-economic indicators taken from the previous census. Changing the definition of the random intercept does not have any adverse impact on the calculation of FGT indicators. As the data is at the unit level and each observation has information on which municipality it belongs to, it is not necessary that the random effect coincides with the municipality. Predictions for income at the unit level can be generated from any model as discussed in section 2.2.2 and the FGT indicator can be calculated simply by using the municipality as a grouping variable.

A major challenge is to find a structure that can be found both in the survey and the census. While the survey includes a variable that indicates the stratum to which a given observation belongs, this cannot be matched with the stratum contained in the census. Therefore, the stratification variable in the data cannot be used directly. However, there are some variables in both the survey and the census that can be used to approximate the stratification procedure. In addition to the `rururb` variable that indicates whether the observation corresponds to a rural or urban area, there are four additional binary variables (Table 4), which contain information on disadvantage in areas such as education, health care, housing quality and access to public utilities. As all five variables are binary, there are $2^5 = 32$ possible combinations. Each one of these 32 combinations is now considered as a domain that is used to define the varying intercept in the model.[11] The variable `rururb` is no longer used as a regressor in this specification. This definition of the domain results in sample sizes between 2 and 368 with a median of 37. No domains are out-of-sample.

The results from PSIS-LOO using model 4.2 show that the stratified specification is clearly better. The elpd_{loo} for this specification is around 34.8 higher than when using municipalities as the domains. The standard error for the difference in elpd_{loo} between both specifications is 12.8, which is around one third of the difference itself. Even if the models were not substantially different in predictive terms, there is one substantial advantage of

[10]The alternative specification of the random effect in this section is included only as an improvement of the model in the Bayesian workflow context. However, there is nothing in the definition of the random effect itself exclusive to the Bayesian paradigm. The same ideas could also be applied to the frequentist EBP with transformations (Rojas Perilla et al., 2020), as it also uses a Monte Carlo approach.

[11]The strata domains are coded as a binary number with five digits. Each digit represents one variable in the following order: `rururb`, `ic_rezedu`, `ic_asalud`, `ic_sbv`, `ic_cv`. For the `rururb` variable 1 represents "rural" and 0 "urban". For all the other variables, 1 represents the presence of a disadvantage measured by each indicator as defined in Table 4.

using the alternative specification: there are no out-of-sample domains. This drastically reduces the uncertainty in the predictions for out-of-sample municipalities. Therefore, the stratified specification of the domain is used in the rest of this paper.

There are some limitations to this approach. Firstly, the stratified structure is only an approximation. Secondly, there are still other dimensions that are not taken into account that might still be problematic, e.g., the clustering that stems from structuring the sample process with PSUs. Thirdly, this is an example that is limited to the data from Mexico. Other statistical institutes might provide more detailed variables for the strata or none at all and depending on the survey design the approach developed in this section might not even be applicable. Nevertheless, the main insight from this section is that for unit-level data the random effect does not have to follow the level at which indicators are estimated and that these improvements can be evaluated in a Bayesian workflow. Looking for alternative specifications for the random effect might lead to better predicitive power and to fewer out-of-sample domains.

4.6. Modelling correlations at the area-level

Until now, the assumption has been that all random effects at the area-level are independent from each other, i.e., $u_d|\sigma_u \sim \mathcal{N}(0, \sigma_u)$. However, this assumption seems inplausible. For example, if the areas are defined as municipalities, it is likely that neighboring regions have similar characteristics and are therefore correlated. If on the other hand the areas are the strata considered in the previous section, the 16 strata that correspond to a rural area have more in common with each other than with urban strata. In this section, a new prior that captures correlations at the area level is introduced to the model. The result is then compared to the model that was defined in the previous section. Thus, the domains are not defined as the municipalities, but as the strata presented in the last section.

There are multiple options to model correlation between areas. The simpler LKJ prior based on the work by Lewandowski et al. (2009) places a restriction over permissible correlation matrices. Autoregressive approaches such as IAR, CAR and SAR (Chung & Datta, 2020) represent another way of capturing spatial correlation and takes into account how similar or close the different domains are. Moreover, other priors such as the random walk prior used in Gao et al. (2021) offers an alternative to modelling dependencies between areas.

The resulting model will be compared with model 3.1 that assumes independence between areas, but with the modified random effect. A comparison between all possible priors that

capture area correlation is beyond the scope of this paper and left for future research.

4.6.1. LKJ prior

Let Σ be a $D \times D$ covariance matrix, so that $u \sim \mathcal{N}(\mathbf{0}_D, \Sigma)$, where $\mathbf{0}_D$ and u are D-dimensional vectors. The density of u can be written as:

$$p(u|\Sigma) = (2\pi)^{-\frac{D}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{(-\frac{1}{2}u'\Sigma^{-1}u)}.$$

The covariance matrix Σ can be decomposed as $\Sigma = \text{diag}(\tau)\Omega\text{diag}(\tau)$, where τ is a D-dimensional vector of scale factor, $\text{diag}(\tau)$ is a $D \times D$ matrix with τ as its main diagonal and Ω is a $D \times D$ correlation matrix. Up to this point, the assumption has been that there is only one scale parameter σ_u for all area-level effects u_d and this assumption is still kept. Because $\tau_d = \sqrt{\Sigma_{d,d}}, d = 1, ..., D$, the decomposition is simplified to $\Sigma = \sigma_u^2\Omega$.

The standard deviation for the random effect σ_u^2, already has the prior $Ga(2, 7)$. However, the correlation between domains is captured by defining a LKJ prior over the correlation matrix Ω[12]:

$$\Omega \sim \text{LKJ}(\eta), \quad \eta \geq 1.$$

The LKJ correlation distribution is defined so that $p(\Omega|\eta) \propto \det(\Omega)^{\eta-1}$ (Stan Development Team, 2021, Chapter 1.13). Note that the determinant of Ω increases as the correlation between components decreases: an identity matrix has a determinant of 1, while a correlation matrix consisting only of ones (perfect correlation between components) has a determinant of 0. For $\eta = 1$, the LKJ prior is a uniform distribution over all possible correlation matrices. However, due to the fact that the function $f(\eta) = k^{\eta-1}$ for a fixed $k \in (0, 1)$ does not converge uniformly towards zero as $\eta \to \infty$, a higher η puts more probability mass on matrices with a higher determinant, i.e., on matrices with a lower correlation between components. In the model, η is set to 5, which strongly favors matrices with determinants closer to 1, i.e., correlation matrices with a moderate correlation between domains. Figure 6a shows four posterior draws for Ω as heatmaps. There are correlation values in the range from -0.5 to 0.5, which indicates that there is some correlation between areas. However, the heatmaps do not reveal any clear correlation pattern. This will be discussed more in detail in section 4.6.3. In the next step, the SAR prior is presented and included into the model

[12] In practice, it is common to use the Cholesky decomposition of Ω to avoid numerical issues when estimating the model. For clarity, the traditional matrix notation is kept in the paper, but note that the decomposition is used in the accompanying code.

as an alternative to the LKJ prior.

4.6.2. SAR prior

The simultaneous autoregression (SAR) prior as described in Chung & Datta (2020) is defined on a precision matrix Π. The starting point is the $D \times D$ proximity matrix W, which contains information on how close two domains are. If $w_d, d = 1, ..., D$ is defined as the sum of row d of matrix W – i.e., $w_d = \sum_{i=1}^{D} w_{di}$ –, then the $D \times D$ matrix L is defined as $L = \text{diag}\{w_d\}_{d=1}^{D}$. Thus, it is possible to calculate the row-normalized matrix $\widetilde{W} = L^{-1}W$, in which all rows sum to 1. The SAR prior in the context of model 4.2 is then given by

$$\begin{aligned} \Pi(\rho) &= (I_D - \rho\widetilde{W})'(I_D - \rho\widetilde{W}), \quad \rho \in (-1, 1), \\ u &\sim \mathcal{N}(\mathbf{0}_D, \sigma_u^2 \Pi^{-1}), \\ \rho &\sim \mathcal{U}(-1, 1), \end{aligned}$$

where $\mathbf{0}_D$ is a D-dimensional zero vector. Again, note that in this model Π is *precision* and not a correlation matrix. Because the single strata used as domains are coded as a 5-digit binary string (see Section 4.5), the distance matrix W is defined using the Hamming distance. In short, the Hamming distance takes two strings of the same length and counts how many digits are different, e.g., 00111 and 10100 have a Hamming distance of 3. This also ensures that the distance of each stratum from itself is 0. The maximum distance in this case is 5, because each stratum is coded with five binary digits. Note that when $\rho = 0$ the precision matrix Π is equal to the identity matrix, which means that there is no correlation between the domains. As ρ deviates from 0, the effect of the spatial increases. The uniform prior is chosen for ρ, because it is not possible to know how strong the correlation values are before fitting the model to the data. The limits $\rho \in (-1, 1)$ are needed so that Π is positive definite. Similarly to the LKJ prior, Figure 6b shows four posterior draws for the correlation matrix. Here, the actual correlation matrix is depicted, *not* the precision matrix Π.[13] The pattern created by the distance matrix is very clear and will be discussed more in detail in the next section. Finally, although the prior on ρ is uniform, its posterior distribution (not shown) is right skewed, with 70% of the probability mass between -1 and 0 and the rest between 0 and 1. It is not straightforward to interpret the effect of this parameter for large matrices, as the precision matrix Π has to be inverted at a later stage. However, a clear

[13] Π^{-1} is a covariance matrix, which is rescaled with the `cov2cor` function in `R` to get the correlation matrix. The scaling parameter σ_u^2 is ignored in the rescaling as it is only a multiplicative constant of the covariance matrix.

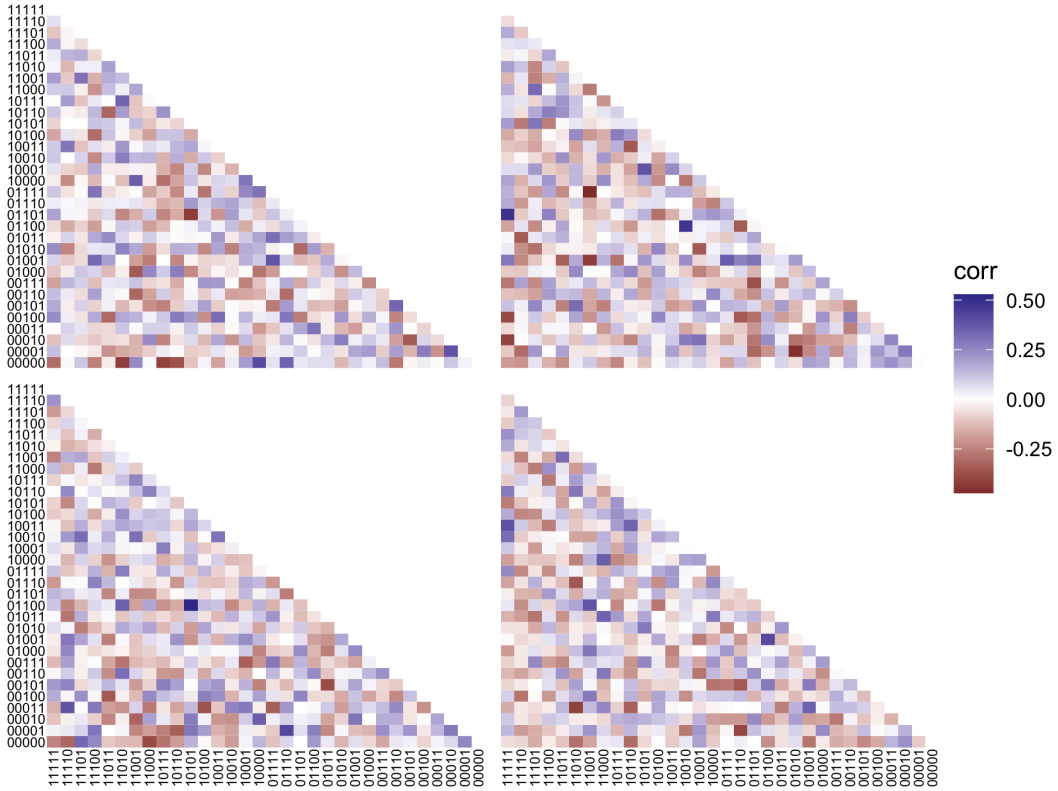
deviation from 0 (as in this case) indicates area-level correlation.

4.6.3. Comparison of LKJ and SAR priors

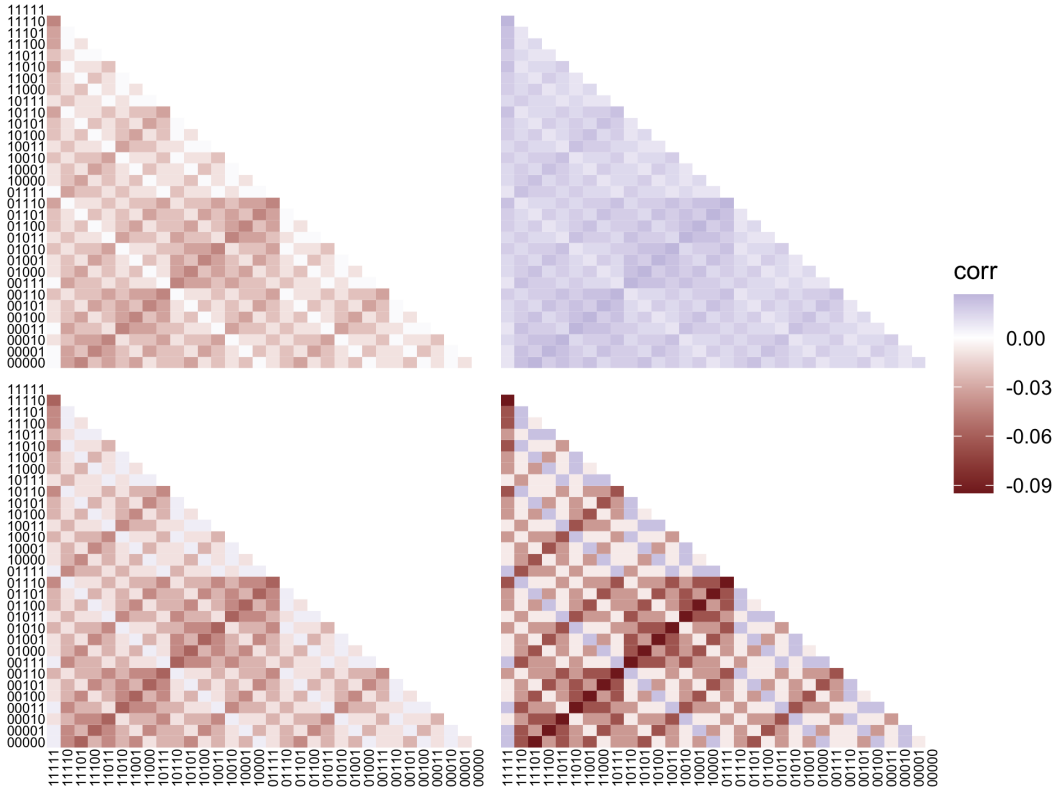
The main difference between both correlation priors is that the SAR prior uses actual information on how similar or close the domains are. In contrast, the LKJ prior is only a prior on the admissible correlation matrices. The difference becomes clear in Figure 6, which shows four posterior samples of the correlation matrices from the models with the LKJ and SAR priors. The axes display the respective domains, whose notation as a 5-digit binary string (e.g., 01100) was explained in section 4.5. The correlation matrices produced by the LKJ prior display a completely random pattern with respect to the correlations. On the other side, with the SAR prior there is a clear structure in the correlation matrices stemming from the distance matrix W. Even though in Figure 6b the correlation values vary from sample to sample, a very clear symmetric structure is visible for each sample, which is related to the way the Hamming distance is used to define W[14]. The nature of the pattern can be further understood in Figure 7, which shows the distributions of the correlation values for each one of the posterior samples in Figure 6. The LKJ prior does not force any particular correlation pattern on the data, but it guarantees that the distribution of the correlation values is roughly equal for each sample. With an increasing η, the LKJ prior puts more weight on correlation matrices closer to the identity matrix, which leads to tighter distributions for the correlation values. On the other hand, there is a much clearer pattern in the densities of the SAR prior specifications. The spikes in the densities are caused by the Hamming distance, which in this paper takes only six values, from 0 to 5. With continuous distances, the densities would be much smoother. Moreover, the correlation values densities do not overlap as in the LKJ prior.

The LKJ and SAR specification are compared with model 4.2 without domain-level correlation (base model) using PSIS-LOO. All models use the stratified random effect specification. The results are shown in Table 5. The elpd_{loo} of the SAR model is slightly worse than the base model (-0.07), but the standard error of the difference (second column) is more than three times the estimated elpd_{loo} difference itself (0.24). According to PSIS-LOO, the base and SAR models are undistinguishable. The elpd_{loo} difference for the LKJ specification (-0.72) is more than three times its standard error (0.22). Neverteheless, it is important to notice that the elpd_{loo} (third column) are in an almost identical range and

[14]Figure 6 depicts only four posterior samples from more than 1000 MCMC samples. The correlation pattern in the SAR prior is present in all samples, but the range of correlation values from just four posterior samples is not representative.

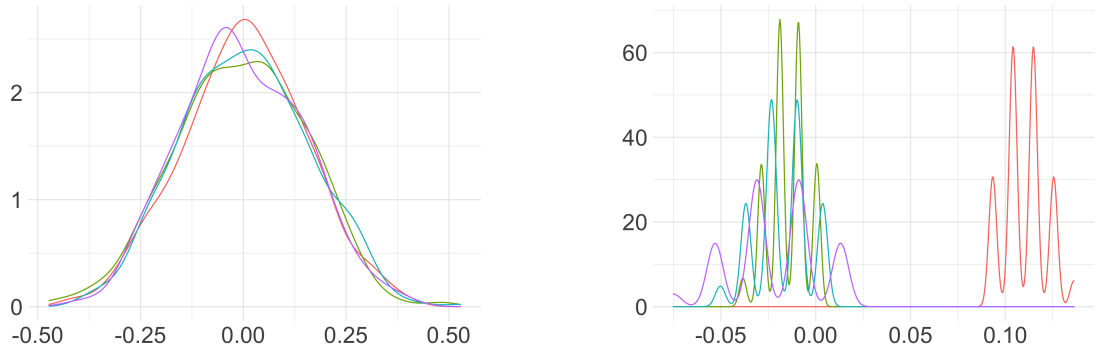


(a) LKJ(5) prior



(b) SAR prior

Figure 6.: Four samples of the posterior correlation matrices for the models with LKJ and SAR priors. The domain coding in the axes is explained in section 4.5.



(a) LKJ(5) prior

(b) SAR prior

Figure 7.: Correlation density for the four posterior samples of the LKJ and SAR priors shown in Figure 6. Each sample is displayed in a different color.

that the standard errors are quite large. Therefore, when looking at the larger picture, the differences between models seem to be marginal.

The main question at this point is why there is no clearer difference between these three specifications. There are three possible explanations. Firstly, as some of the areas are quite small, it is difficult to indentify correlations between areas. It is possible that with more data the correlation could become more visible. Secondly, the PSIS-LOO quantifies leave-one-out cross-validation. However, leaving out only one observation might not make a large difference for the correlations at the area-level. PSIS-LOO might not be the right tool to quantify the difference and alternatives such as leave-one-group-out cross-validation could provide a clearer result. Thirdly, the predictors might have a good predictive power, meaning that area-level correlation does not have such a large impact. Finally, this result points to a blind spot in the simulation scenarios, because they do not take area-level correlation into account. Therefore, it is not possible to know for certain whether the similar results for all three models in Table 5 is due to lack of data, to the limitations of PSIS-LOO or simply because the area-level correlation is very small. Simulated data could have been used to check whether the diagnostic tools can capture a difference in predictive power due to area-level correlations. Unfortunately, due to this limitation of the simulation scenarios from section 3.4 it is not possible to know the exact reason for the similarity between the three model specifications. A critical discussion of the simulations scenarios can be found in section 6.3.

In the current model specification, the random intercept was redefined so that there are no out-of-sample areas. Nevertheless, the LKJ and SAR priors are not as useful when there are out-of-sample regions, as the correlation matrix is assumed to have as many dimensions

Table 5.: Comparison of LKJ, SAR and base specification with PSIS-LOO.

	elpd_{loo} diff.	S.E. diff.	elpd_{loo}	S.E. elpd_{loo}
Base	0.00	0.00	-14799.32	54.40
SAR	-0.07	0.24	-14799.39	54.45
LKJ	-0.72	0.22	-14800.05	54.40

as in the training set. In such cases, an autoregressive or a random walk prior on u_d such as in Gao et al. (2021) can be more appropriate, as it does not depend strictly on the dimensions of the correlation matrix. In any case, it is implausible that the area-level effects are completely independent. Therefore, the SAR specification is chosen as the final model. The next section compares all models fitted until now using stacking weights.

4.7. Comparison of Bayesian models with stacking weights

Model selection is an uncertain procedure, especially when there are multiple plausible models that are consistent with the data (Gelman et al., 2020). While in section 4.6 the SAR model was selected as the final model, it was also clear in Table 5 that the model was indistinguishable from or only marginally better than the other models. Stacking, which was discussed in section 2.1.2, provides a method to combine predictions from multiple heterogeneous models. Weights are calculated jointly for all models based on their elpd_{loo}, which can be used to estimate a weighted average of the predictions from the different models. This paper does not investigate whether stacking provides better results than a single model. However, the weights (shown in Table 6) provide valuable insights into the predictive power of the different models. First, stacking is relatively insensitive to similar models. This can be seen in the first three models, which are identical with the exception of the tightness of the prior on the skewness. The first three models share their weights in the sense that only the weight of the third model is non-zero. The last three entries in Table 6 correspond to the three models used to introduce area-level correlation: a base model with no correlation, the LKJ model and the SAR model. These three models offer a different picture: all of their weights are non-zero, which indicates that they are heterogenous, or else some of the weights would be equal to zero. Additionally, their weights (0.25, 0.31, 0.24) are in a very similar range. This indicates that they would contribute almost equally to a weighted average of predictions. This confirms the results in Table 5 that showed a very similar elpd_{loo} for all three models. However, their weights are higher than the first three models, which indicates that the initial models were improved through the workflow

Table 6.: Stacking weights of models in the workflow.

Mid skew.	Low skew.	High skew.	Base	LKJ	SAR
0.00	0.00	0.19	0.25	0.31	0.24

The first three models correspond to the three different priors on skewness from section 4.2.1. The *Base* model corresponds to the model with no area-level correlation developed in section 4.3. The *LKJ* and *SAR* models correspond to the models with area-level correlation from section 4.6. All models used the stratified random effect discussed in section 4.5.

In summary, the decision to use only the SAR model is not necessarily the only possible one. It is likely that a combined prediction with methods such as stacking will outperform the predictions from a single model, but this question is not further considered in this paper. The estimated poverty indicators are presented in the next chapter and compared to the Box-Cox EBP from Rojas Perilla et al. (2020).

5. Comparing EBP and HB Models

This chapter presents the results of the model developed in the workflow. First, the FGT estimates for the final SAR model are presented and compared with the estimators from the Box-Cox EBP. In the second section, the RMSE of the Bayesian model is estimated and compared with the EBP model. As there is no RMSE estimator available for the HB model without flat priors, this comparison is made with the simulation scenarios from section 3.4.

5.1. Estimated poverty indicators with Bayesian model and EBP

Figure 8 shows the HCR and estimates for both the SAR HB model and the Box-Cox EBP. The EBP was estimated with the `emdi` package (Kreutzmann et al., 2019) with the Wild bootstrap option. As `emdi` does not include the survey weights in the estimation process, neither the EBP nor the HB model use the weights when estimating the poverty indicators. Unfortunately, at the time of writing `emdi` does not allow the random effect to be redefined as in chapter 4.5. Because extending or changing the `emdi` package is beyond the scope of this paper, the municipalities are used to define the random effect in the EBP. Despite this limitation, the comparison between the final Bayesian model and the EBP is still helpful to

evaluate the differences between both methods. A more exact comparison is left for future research[15]. For the Bayesian model, the standard deviation as defined in section 2.2.2 is depicted. The RMSE is used for the EBP estimates. While the RMSE and the standard deviation are related, they are not directly comparable, as the RMSE also measures bias.[16]

The pattern in the estimates is consistent with the economic structure of the state. Municipalities with a strong touristic sector such as Acapulco and Zihuatanejo on the Pacific coast as well as Taxco de Alarcón in the north have the lowest values for both indicators. This is also true for the municipal seat region Chipalcingo de los Bravo. On the other hand, poverty indicators are higher in the rural southeastern municipalities near the neighboring state of Oaxaca. However, it is clear from the maps that the HB estimates are lower than the EBP for both the HCR and the PGAP. A more precise picture is given by Figure 9, which shows the difference between the EBP and the HB indicators for each municipality. The EBP indicators are all larger than the HB indicators, both for the HCR and the PGAP[17]. Moreover, the pattern in the differences is similar for both indicators: if the difference is large for the HCR it is also large for the PGAP[18]. Additionally, the HB and EBP estimates were benchmarked against the direct estimate as described by Pfeffermann (2013)[19]. The benchmarked estimate for the EBP (HCR: 0.248, PGAP: 0.112) were closer to the benchmarked direct estimate (HCR: 0.280, PGAP: 0.130) than the benchmarked HB estimates (HCR: 0.151, PGAP: 0.059). As the EBP estimates are higher than the HB estimates, this result is not surprising. The reasons for this systematic deviation from the EBP are further considered in the discussion.

The uncertainty maps (standard deviation and RMSE) provide a more detailed picture into estimate quality. Note that the HB uses the stratified random effect from section 4.5, whereas the EBP uses the municipality as the random effect like in the original BHF specification. For the EBP, the uncertainty maps for the HCR and PGAP display a clear pattern of out-of-sample municipalities that have a higher RMSE[20]. This is not the case

[15]Morelli (2021) uses an HB model with the classical random effect to predict HCR and PGAP for the state of Guerrero. The model is not exactly the SAR HB model developed in this paper, but it can be used as an additional point of comparison without the stratified random effect.

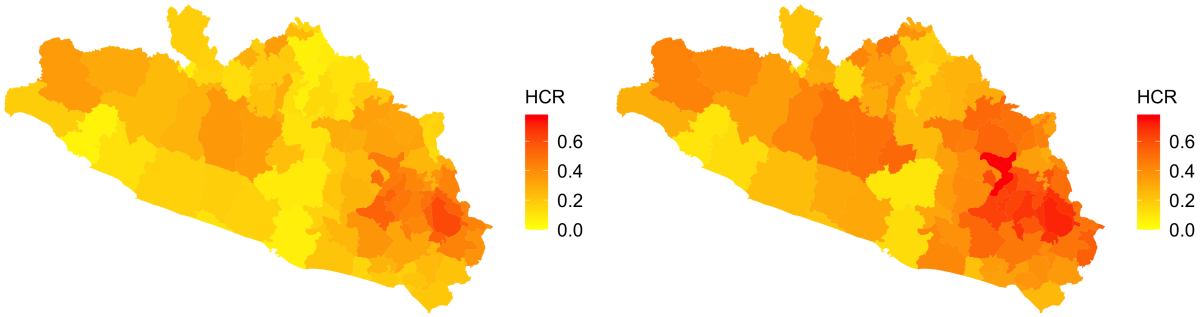
[16]$MSE(\hat{\theta}) = bias^2(\hat{\theta}) + V(\hat{\theta})$.

[17]The HB estimates are also lower than the EBP in Morelli (2021), which uses the classical BHF random effect specification. This will be further discussed in chapter 6.1.

[18]This pattern does not change when taking the relative difference to account for the magnitude of the EBP indicator, i.e., $(\theta_d^{EBP} - \theta_d^{HB})/\theta_d^{EBP}$.

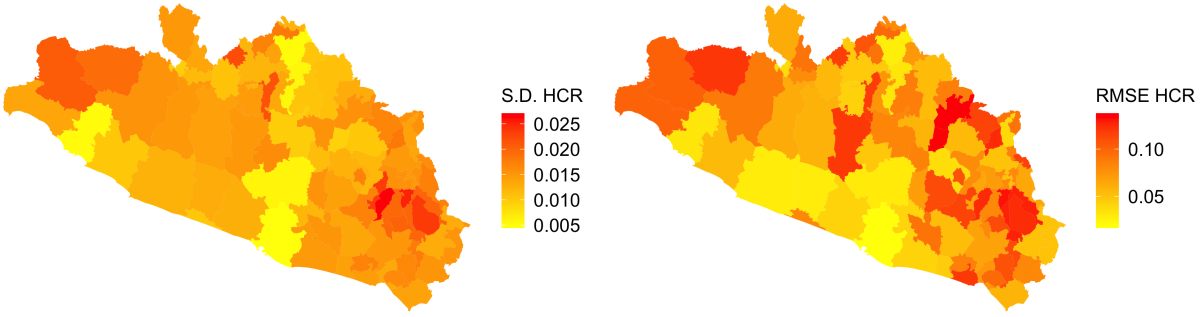
[19]If w_d are weights given relative population size at the municipality level, then the benchmarked direct estimate $\sum_{d=1}^{D} w_d \hat{\theta}_d^{Dir}$ should be roughly equal to the benchmarked estimate from an EBP or HB model $\sum_{d=1}^{D} w_d \hat{\theta}_d^{Model}$. Note that by definiton $\sum_{d=1}^{D} w_d = 1$.

[20]A map of the state of Guerrero with in and out-of-sample areas can be found in appendix H.



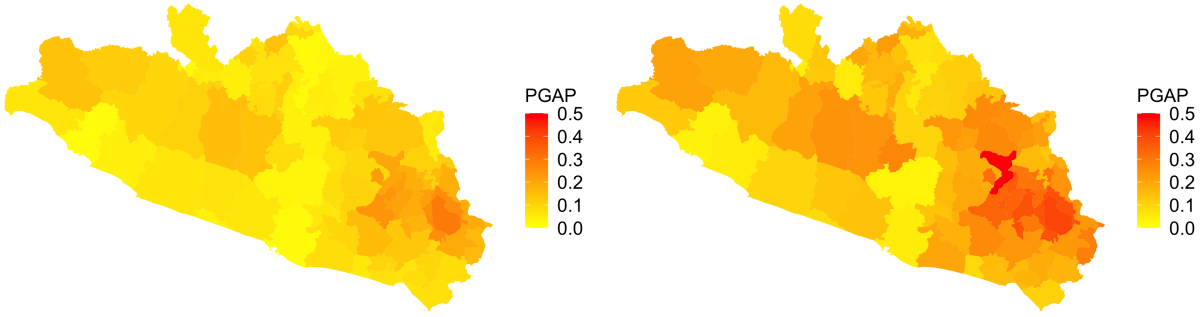
(a) HCR estimate (HB)

(b) HCR estimate (EBP)



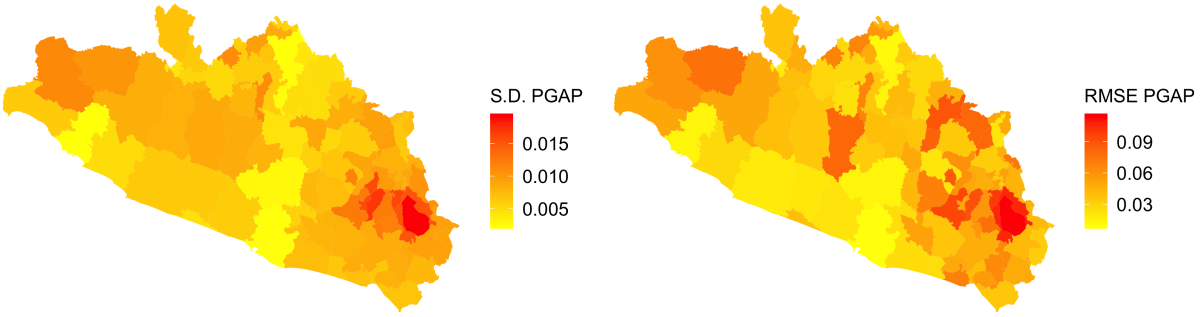
(c) HCR standard deviation (HB)

(d) HCR RMSE (EBP)



(e) PGAP estimate (HB)

(f) PGAP estimate (EBP)



(g) PGAP standard deviation (HB)

(h) PGAP RMSE (EBP)

Figure 8.: Mean and uncertainty estimation for the HCR and PGAP indicators.

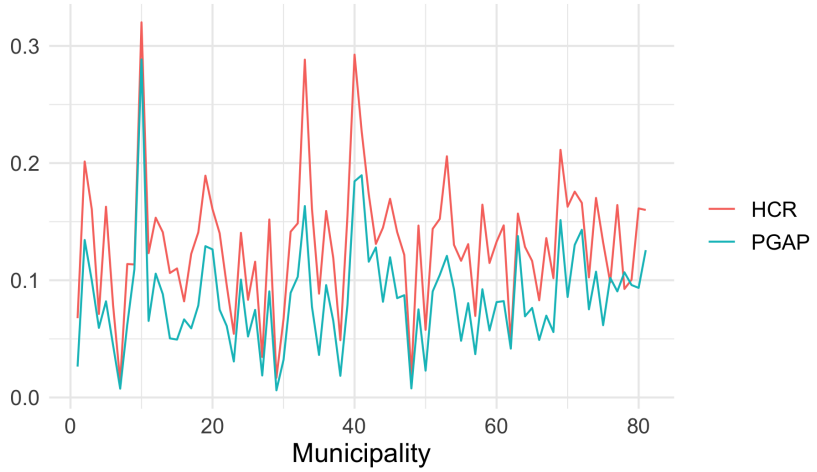


Figure 9.: Difference between EBP and HB indicators, calculated by subtracting the HB estimates from the EBP estimates for each municipality. All EBP estimates are higher for both the HCR and the PGAP.

for the HB model, as the stratified random effect specification has no out-of-sample areas. Moreover, in the HB model some municipalities have a standard deviation that is just one tenth of their corresponding standard deviation in Morelli (2021), which is a drastic reduction in uncertainty. Additional maps of the coefficients of variation can be found in appendix H.

5.2. RMSE of EBP and Bayesian model

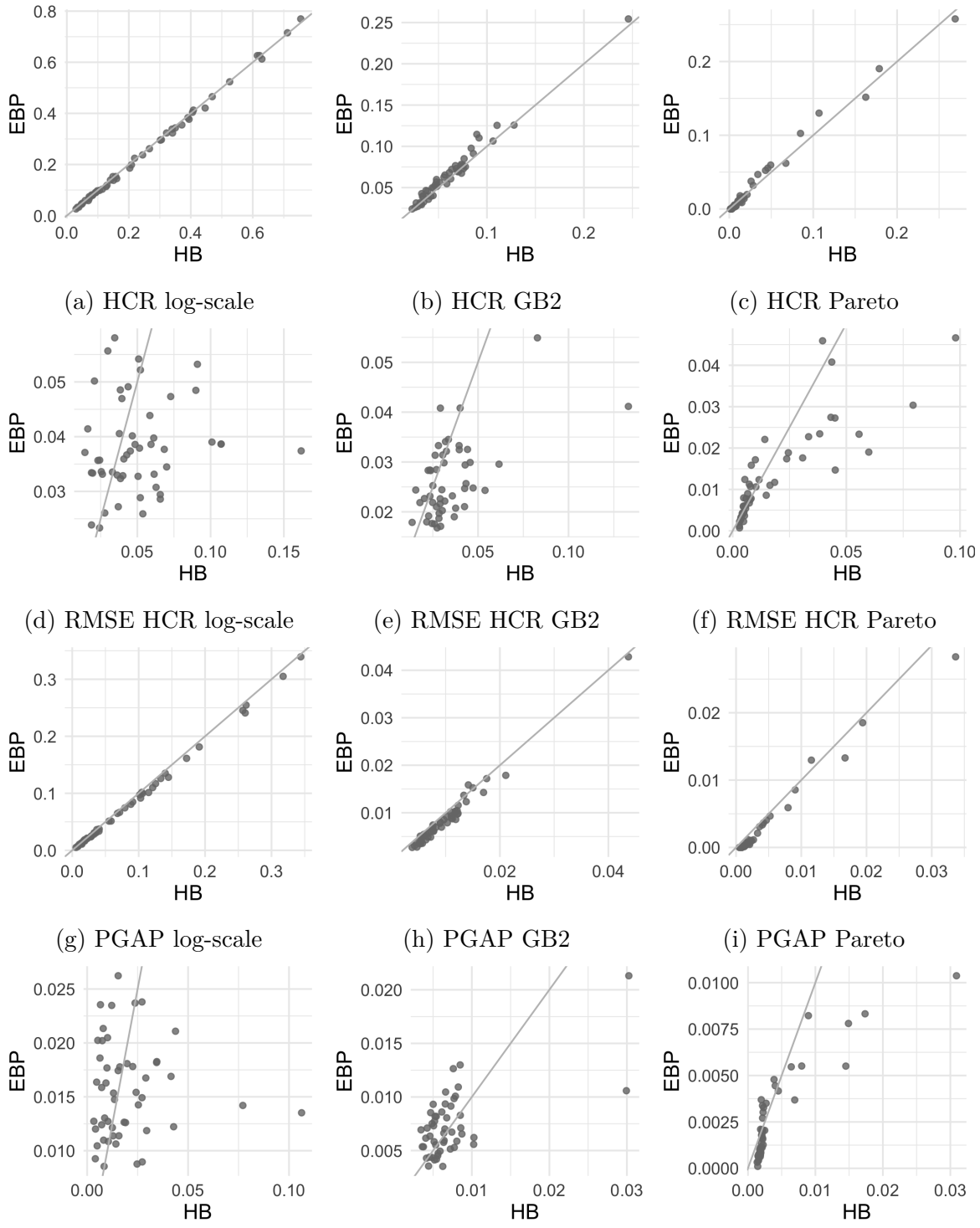
The main issue when comparing HB and EBP approaches is RMSE estimation. While Rojas Perilla et al. (2020) develop two bootstrapping algorithms to estimate the RMSE of the indicators, it is unclear whether a similar approach could work with a Bayesian model. Molina et al. (2014) use the standard deviation of the estimators to quantify prediction quality. However, the RMSE can only be approximated by the standard deviation when using flat priors (Rao & Molina, 2015, Chapter 10.3.2). To avoid this problem, the simulated data from section 3.4 are used to estimate the empitical RMSE of the HB estimates. This is straightforward, as both the population data set and the small area samples drawn from the population are available. First, the FGT indicators F_d^{pop} are calculated for the population. With S draws from the posterior predictive distribution, the FGT indicator $F_d^{(s)}, s = 1, ..., S$,

can be calculated for each sample s. The RMSE is given by

$$RMSE_d^{HB} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(F_d^{(s)} - F_d^{pop})^2}.$$

The results are presented in Figure 10. Note that in this chapter the model with no area-level correlation is used, as the simulation scenarios do not include spatial correlation, nor a distance measure for the SAR prior. The first and third rows show the HCR and PGAP estimates, with the corresponding RMSE in the second and fourth rows. The figures are presented as scatterplots of the EBP against the HB model where each one of the 50 domains in the simulated data is one observation. The 45-degree line shows the area where the values of the EBP and HB models are equal. In other words, if a point in the scatterplot is below this line it means that its value is larger for the HB model than for the EBP. In general terms, the HCR and PGAP estimates are almost identical in both models for all scenarios. Only for the PGAP the estimates are slighlty higher for the HB model compared to the EBP in the GB2 and Pareto scenarios.

The RMSE plots paint a different picture. While the RMSE of the HB is better than the EBP in some areas (above the 45-degree line), most of the areas have a higher RMSE in the HB model. An exception is the PGAP for the GB2 scenario, in which most of the RMSE estimates seem to be lower in the HB model, but this can be an artifact of randomness. This is hardly surprising, as the EBP is guaranteed to have the smallest possible RMSE of all estimators under certain conditions – i.e., it is approximately the *best predictor* (Molina et al., 2014). Another pattern found in the RMSE graphs is the presence of outliers in the HB model (x-axis). While most RMSE estimates are in a limited region of the graph, there are a few observations in all RMSE scatterplots that are clearly higher than the rest. However, it is important to remember that this outliers might be artifacts that arise due to the different ways of estimating the RMSE. For the HB scenarios, it is calculated analytically using the simulated population, while the EBP uses a bootstrapping estimate of the RMSE that does not take the simulated population into account. A possible solution to this problem would be to use the single Monte Carlo samples of the EBP directly together with the simulated population as a reference, which at the moment of writing is not possible with `emdi`. Finally, the scenarios used included no area-level correlation. Further research is needed to evaluate how a comparison as in Figure 10 would change with a higher correlation between domains.



(a) HCR log-scale (b) HCR GB2 (c) HCR Pareto

(d) RMSE HCR log-scale (e) RMSE HCR GB2 (f) RMSE HCR Pareto

(g) PGAP log-scale (h) PGAP GB2 (i) PGAP Pareto

(j) RMSE PGAP log-scale (k) RMSE PGAP GB2 (l) RMSE PGAP Pareto

Figure 10.: Comparison of the Box-Cox EBP and the HB approach developed in this paper. Each row shows scatterplots for all three simulation scenarios of the EBP against the HB approach. The *first row* shows the comparison between the HCR estimates, while the *second row* shows the comparison of the corresponding RMSE estimates. The *third row* presents the PGAP estimates and the *fourth row* the RMSE of the PGAP estimates.

6. Discussion

After presenting the results, it is necessary to evaluate them critically and also to revisit certain steps of the workflow that could potentially be improved. In the first section, the Bayesian and EBP are compared in their advantages and disadvantages. The second section discusses whether a Bayesian workflow is useful in the context of survey statistics and more specifically of small area estimation. Lastly, the simulation scenarios are evaluated with respect to their limitations and extensions to these scenarios are proposed.

6.1. Critical evaluation of the Bayesian approach against the EBP

The aim of this paper was to develop a Bayesian model for poverty estimation iteratively. However, it is puzzling that the estimated indicators from the final HB model are lower than the EBP results. There are a number of possible reasons for the discrepancy:

- *Regularization*: prior distributions have a regularizing effect on the model, while the EBP does not contain any form of shrinkage.
- *Random effect specification*: the stratified random effect specification from section 4.5 might lead to systematically lower estimates than the traditional BHF specification.
- *Survey weights*: by not taking into account the survey weights – through which one observation can count as multiple observations – a Bayesian model can have trouble capturing the true distributional characteristics of the data.

It is noteworthy that the EBP and HB estimates are almost identical, when using the survey weights with the HB model in the calculation of FGT indicators (not shown). *Area-level correlation* was excluded as a cause for the discrepancy between the two models, as the estimated indicators from the base model without spatial correlation and from the SAR model are almost identical.

The HB estimates were already lower than the EBP estimates in Morelli (2021), which used the traditional random effect based on municipalities. Therefore, the discrepancy is unlikely to be caused only by the stratified random effect specification. Moreover, the simulation study in section 5.2 showed that even when using a Bayesian model with proper

priors that induce shrinkage, the estimates for the indicators are almost identical. These observations lead us to the following conclusion. On the one hand, the simulation scenarios are constructed in a way that is unlikely to lead to overfitting. However, shrinkage in the Bayesian model might be more noticeable when using more complex real-world data, which might include covariates that increase the risk of overfitting. On the other hand, survey weights bring the HB model in line with the EBP model and this indicates that they might play a key role in adjusting the distributional characteristics of the Bayesian model. Further research is needed to answer this question conclusively. For this, it is necessary to take into account both overfitting and survey weights in the simulation scenario. An additional check could include estimates from a skewed distribution such as the gamma with a log link to see whether the HB estimates change substantially. Finally, as the EBP and HB models use different uncertainty measures and have different random effect specifications, it is more difficult to compare them with respect to estimate uncertainty. Nevertheless, avoiding out-of-sample regions by redefining the random effect was shown to reduce estimate uncertainty drastically and it is plausible that it would have a similar effect on the EBP.

An additional dimension to consider is ease of use. The `emdi` package already provides a series of implemented models out of the box. In contrast, the models developed in this paper were all coded using `Stan`, which requires more specialized knowledge of how Bayesian models are computed. This can be overwhelming for some users. Many modelling decisions such as prior choice and estimation of FGT-indicators from backtransformed samples from the posterior predictive distribution can be automatized in a package like `brms`. Moreover, it is important to keep computation time in mind. The EBP model took around 8 minutes to fit *and* produce various socio-economic indicators such as HCR and PGAP. The final Bayesian model with the SAR prior also takes around 8 minutes to fit, but an additional 5 to 10 minutes are needed to calculate the indicators, depending on the number of Monte Carlo samples. While this still is an acceptable and realistic time, it is important to underline that Bayesian models usually take longer to estimate and that model complexity can have a huge impact in this respect. Moreover, the `Stan` code was optimized for speed by taking full advantage of vectorization and slight changes can lead to a large increase in computational time.

6.2. Using a Bayesian workflow for SAE

This section discusses the benefits and disadvantages of Bayesian workflow. A key advantage of using a Bayesian workflow is that any model changes can be presented and analyzed in a

transparent way. There are a variety of tools that can be used to check model assumptions (prior predictive checks), model fit (posterior predictive checks), or predictive power (PSIS-LOO and stacking). Moreover, as the model parameters are distributions and not point estimates it is always possible to look at each parameter distribution independently to check if the values are realistic and also to better understand how the model works. This can be done both for unidimensional parameters such as standard deviation and multidimensional parameters such as correlation matrices (e.g., Figure 6). In summary, there are tools to understand how the model works and the iterative development of the model allows for new insights into the problem at hand and makes the uncertainty in model choice visible. This transparency is well-suited for survey statistics – a field which often provides information for public policy decision-making.

Nevertheless, there are still disadvantages in using a Bayesian workflow. As Bayesian statistics has become more widespread only in the last few years, there are still some tools in need of development. Variable selection is one of those areas, as the horseshoe prior is still not ideal for this task due to the lack of a clear selection criterion and due to its tendency to create computation problems like HMC divergences. A better alternative would be the projective prediction approach from Piironen et al. (2020), but which at the time of writing only works for a few likelihoods from the exponential family. This is an area of active research and new developments are to be expected in the following years. Another difficult aspect of Bayesian workflow is prior choice, especially for groups of similar parameters such as regression coefficients. While prior predictive checks are useful to determine the impact of prior choice, there might be too many possibilities to be considered. A current area of research is joint priors such as the R2D2 (Zhang et al., 2020) or even the regularized horseshoe (Piironen & Vehtari, 2017b), which affects multiple parameters simultaneously. Joint priors are simpler to parametrize and are therefore more user-friendly. They can also have advantages in avoiding overfitting, as they are usually controlled by parameters such as R^2 or number of relevant parameters. Lastly, as Gelman et al. (2020) point out, iterative model fitting might cause problems with respect to inference validity. Again, this is an area of current research.

6.3. Adequacy of simulation scenarios

The use of simulated data is a central part of Bayesian workflow according to Gelman et al. (2020). However, it is also a time-consuming step. In general, it is hard to create a scenario or multiple scenarios that capture the main features of the data, while at the same time not

defining overly complicated scenarios or generating so many scenarios that it is a burden for the researcher to work with them.

The models used in this paper are an extension of the models in Rojas Perilla et al. (2020), mainly through the addition of correlated covariates. Nevertheless, there are some critical aspects that can be reconsidered. Firstly, there might be room for improving the assumed distributions in the scenarios. On the one hand, the GB2 and Pareto scenarios add a skewed error term to covariates and random effects that are normal. The result is a relatively symmetric distribution with a long right tail. To make both scenarios more realistic, the distributions of the regressors could be more varied, so as to avoid a strong symmetry before adding the error term. On the other hand, the log-scale scenario assumes normality in the logarithmic scale. While this is a good sanity check for the results, it is unlikely to reflect the difficulty of fitting real-world data. A simple modification to avoid this strong assumption, would be again to change the normal distribution of the regressors to a heavier tailed Student's t with different degrees of freedom for each covariate. Secondly, a majority of variables found in a survey or a census are categorical. Neither Rojas Perilla et al. (2020) nor the present paper include categorical variables in the simulations. Therefore, there are probably some blind spots in the analysis of this paper, regarding difficulties that might arise from the use of categorical covariates. Thirdly, the area-level effects are assumed to be independent – an unrealistic assumption. Many different ways of including correlation were already mentioned in section 4.6: an LKJ prior, an SAR prior or using a random walk between areas. While choosing any one of those procedures can place a strong assumption on the simulations, it should at least help the researcher to show whether the model can capture correlations at the area level even with small areas. Finally, this paper ignored the issue of problematic regressors. The correlation between covariates is moderate, only 0.2. Fitting a model that does not overfit is more challenging with a higher correlation between regressors and also with covariates that are highly correlated with the dependent variable. Including these suggestions into the simulations scenarios, can provide a more realistic setting to test the models.

At the same time, it is crucial not to overcomplicate the models and not to have too many scenarios. In this paper, the GB2 and the Pareto scenarios are both additive and generate dependent variables that are similar in shape. To simplify model analysis, it might be easier to limit the scenarios to one multiplicative scenario and one additive scenario, either with a GB2 or a Pareto error.

One last aspect to take into account is the sampling procedure used with the simulated data. In this paper, the survey data set was created by sampling from the simulated

population with no out-of-sample domains. However, there are cases where many domains are out-of-sample. For example, in many states of the Mexican data set, there are no observations for 50%-70% of domains in the survey. This is an area of future research, especially when comparing the differences between EBP and an HB model with informative priors.

7. Conclusion

This paper applied the Bayesian workflow from Gelman et al. (2020) to the estimation of poverty indicators. The workflow developed in this paper showcases how a Bayesian model can be improved iteratively in the small area estimation context and how the model can be made more explainable by using different diagnostic tools. This explainability is particularly important for public policy decision-making. Nevertheless, there is still an open question regarding the systematic differences between the final EBP and HB estimates. Further research is needed to answer this question. A simple extension to the workflow would be to check the poverty estimates after each iterative improvement. This would allow a comparison of the HB results with other methods such as the EBP or a direct estimate after each step to identify point at which the results from the methods start diverging. Additionally, only point estimates were taken into account in the present paper. It is worthwhile to explore new ways of presenting the whole distribution of the estimates.

A more complete workflow would also include estimates from other types of models – e.g., with skewed likelihoods. Moreover, due to the modular nature of Bayesian models it is possible to include benchmarking (Pfeffermann, 2013) into the model itself. For example, the weighted average of expected values for each area can be defined to match a weighted average of direct estimates. This would guarantee that the results are in line with the unbiased direct estimator.

References

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error components model for

prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*, 28–36.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. Retrieved 2021-01-21, from `http://arxiv.org/abs/1701.02434`

Betancourt, M. (2020, April). *Towards A Principled Bayesian Workflow.* Retrieved 2021-08-05, from `https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html`

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Box, G. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1).

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Chung, H. C., & Datta, G. S. (2020). *Bayesian hierarchical spatial models for small area estimation* (Tech. Rep.). Washington, D.C. (US): U.S. Census Bureau.

Encyclopaedia Britannica. (2019). *Guerrero.* Retrieved 2021-01-21, from `https://www.britannica.com/place/Guerrero`

Foster, J., Greer, J., & Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica*, *52*(3), 761–766.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402.

Gao, Y., Kennedy, L., Simpson, D., & Gelman, A. (2021). Improving multilevel regression and poststratification with structured priors. *Bayesian Analysis*, *16*(3), 719–744.

Gelman, A. (2020). *Prior Choice Recommendations.* Retrieved 2021-08-06, from `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations`

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). Boca Raton, FL (US): CRC Press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ... Modrák, M. (2020). Bayesian Workflow. Retrieved 2021-09-10, from `https://arxiv.org/abs/2011.01808`

Gómez-Rubio, V. (2020). *Bayesian inference with INLA.* Boca Raton, FL (US): CRC Press.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

INEGI. (2011). *Módulo de Condiciones Socioeconómicas. Encuesta Nacional de Ingresos y Gastos de los Hogares 2010. Diseño muestral.* Retrieved 2021-08-09, from `https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825002426`

Jacod, J., & Protter, P. (2004). *Probability Essentials* (2nd ed.). Berlin, Heidelberg: Springer-Verlag.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators. *Journal of Statistical Software*, *91*(7).

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). Boca Raton, FL (US): CRC Press.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Molina, I., Nandram, B., & Rao, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, *8*(2), 852–885.

Morelli, F. (2021). *Hierarchical Bayesian Models: An Application to Small Area Estimation* (Tech. Rep.). Free University Berlin.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 113–162). Boca Raton, FL (US): CRC Press.

Paananen, T., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2021). Implicitly Adaptive Importance Sampling. *Statistics and Computing*, *31*(16).

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science*, *28*(1), 40–68.

Piironen, J., Paasiniemi, M., & Vehtari, A. (2020). Projective Inference in High-dimensional Problems: Prediction and Feature Selection. *Electronic Journal of Statistics*, *14*(1), 2155 – 2197.

Piironen, J., & Vehtari, A. (2017a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735.

Piironen, J., & Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051.

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation* (2nd ed.). Hoboken, NJ (USA): John Wiley & Sons, Inc.

Rojas Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121–148.

Stan Development Team. (2021). *Stan Modeling Language Users Guide and Reference Manual, 2.27.* Retrieved 2021-08-05, from `https://mc-stan.org`

Stock, J. H., & Watson, M. W. (2015). *Introduction to Econometrics* (3rd ed.). Harlow, England: Pearson Education Limited.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288.

United Nations. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development.* Retrieved 2021-10-03, from `https://sdgs.un.org/2030agenda`

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved R-hat for Assessing Convergence of MCMC. *Bayesian Analysis*, *16*(2), 667–718.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions. *Bayesian Analysis*, *13*(3), 917–1007.

Zhang, Y. D., Naughton, B. P., Bondell, H. D., & Reich, B. J. (2020). Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior. *Journal of the American Statistical Association*.

A. Appendix: An Introduction to Bayesian Computation

A central challenge in Bayesian statistics is model estimation. The integral of the marginal likelihood $p(y) = \int p(\theta)p(y|\theta)d\theta$ is usually intractable. Analytical solutions exist only for a limited group of models where prior and likelihood are conjugate distributions. A thorough review of Bayesian computation methods is beyond the scope of this paper. This section provides a short overview of algortihms and concepts relevant to this paper. The focus is on Markov Chain Monte Carlo (MCMC) and variational inference, but there are other alternatives such as Laplace Approximations (Gómez-Rubio, 2020).

As the normalizing term $p(y)$ in Bayes' theorem is intractable, the posterior distribution has to be approximated. Simple Monte Carlo algorithms such as rejection or importance sampling are effective only for low dimensional parameter spaces. Monte Carlo Markov Chain (MCMC) is a widely used family of algorithms which builds upon the concepts of Monte Carlo approximations. The following overview on MCMC is largely based on Gelman et al. (2014) Chapters 10 to 12. The main idea of MCMC is to have a Markov Chain, which moves through the parameter space according to some previously defined transition distribution. Each step of the Markov Chain is a proposed sample to approximate the *target* distribution – i.e. the distribution which has to be approximated. A proposal can be accepted or rejected according to a pre-defined rule, which is based on the acceptance ratio r of the target probability at the current step and the target probability at the last step.

Because of the ratio, any normalizing constant (the marginal likelihood) cancels out, which allows to find an approximation to the target distribution using only the unnormalized target distribution. Generally speaking, a proposal from the Markov chain that increases the probability of the unnormalized target distribution is always accepted. Otherwise, r is smaller than 1 and the proposal is accepted with probability r. Under certain conditons, the Markov chain is guaranteed to converge to a stationary distribution as the number of steps goes to infinity (Gelman et al., 2014, Chapter 11). As any computation has only a finite number of steps in the real world, it is necessary to check whether the accepted proposals from the Markov chain are reliable. A common approach is to start multiple Markov chains (usually between two and four) with different initial values at random. After discarding the warm-up iterations needed to get from the initial value to the stationary distribution, it is possible to check whether the chains have mixed well – i.e., converged to a similar distribution. While it is sometimes possible to recognize mixing problems with the Markov chains traceplots, it is better to use a numeric diagnostic such as $\hat{R}$ (Vehtari et al., 2021). An $\hat{R}$ higher than 1.01 indicates either that there is a trend in any given chain so that it cannot be stationary or that the chains have converged to different values.

The Metropolis-Hastings algorithm (with its special case, the Gibbs sampler) falls under the MCMC category and is characterized by exploring the parameter space through a random walk. However, there are two problems with this approach. Firstly, the Markov chain can get stuck in certain regions of the parameter space while exploring posteriors where the parameters are highly correlated. Secondly, the sparsity induced in a parameter space with increasing dimensionality makes the random walk approach very inefficient. Hamiltonian Monte Carlo (HMC) is an MCMC algorithm that can explore high dimensional spaces much more efficiently than the Metropolis-Hastings algorithm. The main idea behind HMC is to add a momentum parameter for each one of the model parameters and use Hamiltonian dynamics to explore the parameter space and find a proposal. HMC is a mix of random and deterministic elements. While the momentum parameters are sampled at random, the parameter space exploration which ultimately leads to a proposal is based entirely on a deterministic set of Hamiltonian differential equations. The use of differential equations is an unexpected advantage of HMC, because their discretization with symplectic integrator offers a diagnostic unique to HMC: *divergences*. A divergence occurs when there is a region of high curvature in the parameter-momentum space, which pushes the symplectic integrator towards the edges of such a space. While the exact explanation is somewhat technical, a divergence indicates that there might be areas in the parameter space from which the sampler cannot explore well enough. This is a sign that the resulting Monte

Carlo approximation is not reliable, because of model misspecification or because there is a better parametrization for the model (Betancourt, 2017; Neal, 2011). A more efficient version of the HMC algorithm is NUTS (Hoffman & Gelman, 2014), which has a more complex heuristic for the deterministic exploration. This is the most commonly used MCMC algorithm in Bayesian frameworks such as `Stan`.

Variational inference is an alternative approximation method, which aims to find a distribution that minimizes the divergence (usually Kullback-Leibler) to the target distribution. While variational inference is very fast compared to MCMC, it lacks its theoretical guarantees. Therefore, the resulting variational approximation might be highly misleading (Blei et al., 2017). However, variational inference is a useful tool to do a first check of the model. Because it is faster, it also makes it possible to recognize potential problems sooner. Moreover, there are some types of models which would take a very long time to estimate using MCMC, e.g., topic models (Blei et al., 2003).

B. Appendix: Jacobian Adjustment after Log-Shift Transformation

The transform of the dependent variable introduces a distortion that makes it necessary to adjust the Jacobian of the likelihood using Jacobi's transformation formula (Jacod & Protter, 2004, Theorem 12.6). Let X be a univariate random variable defined over $I \subseteq \mathbb{R}$ with density f^X. For a continuous differentiable function $\varphi : I \to \mathbb{R}$ and $\varphi' \neq 0$ for all $x \in I$, define $Y := \varphi(X)$. The density of the random variable Y can then be defined as

$$f^Y(y) = f^X(\varphi^{-1}(y)) \left| \frac{d\varphi^{-1}(y)}{dy} \right| \mathbb{I}(y \in \varphi(I)), \quad y \in \mathbb{R},$$

where $\mathbb{I}(\cdot)$ is the indicator function. Assuming that y represents the dependent variable in the original scale, then $\varphi^{-1}(y) = \log(y + \lambda)$ and the Jacobian adjustment is $\dfrac{d\varphi^{-1}(y)}{dy} = \dfrac{1}{y + \lambda}$, which is positive due to $y + \lambda > 0$. As the transformation parameter λ is chosen to drastically reduce the skewness of $\log(y + \lambda)$, the density f^X is chosen to be from a symmetric distribution – e.g., a Student's t-distribution. The likelihood of y under the

log-shift transformation is therefore

$$f^Y(y) = \text{Student}(\log(y+\lambda)|\mu,\sigma,\nu) \cdot \frac{1}{y+\lambda} \cdot \mathbb{I}(y \geq \lambda), \quad y \in \mathbb{R}.$$

C. Appendix: Coefficient Interpretation after Log-Shift Transformation

The transformation of the dependent variable complicates coefficient interpretation. In a prediction context, coefficients are of secondary importance, as the estimated values might be biased due to complex interactions between variables, e.g., omitted variable bias, post-treatment bias, collider bias. Models with biased coefficients can offer even better predictive performance than non-biased models (McElreath, 2020, Chapter 7). However, in a Bayesian context it is still useful to understand how regressors affect the dependent variable to choose adequate priors.

For simplicity, consider a regression with a transformed dependent variable and two independent variables:

$$\log(y+\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

By taking the exponential function on both sides, the expression becomes

$$y + \lambda = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$$

Assuming that λ is a fixed constant, it can be analyzed how y changes when increasing x_1 by one unit

$$\frac{y^{new} + \lambda}{y^{old} + \lambda} = \frac{e^{\beta_0 + \beta_1 (x_1+1) + \beta_2 x_2 + \varepsilon}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}} = e^{\beta_1}. \tag{C.1}$$

When $\lambda = 0$, this derivation corresponds to the coefficient interpretation in a log-linear model as an approximate percentage change (Stock & Watson, 2015, Chapter 8). This stems from the approximate relation $\beta_1 \approx e^{\beta_1} - 1$, which is accurate for small β_1 ($e^{0.04} - 1 \approx 0.0408$), but worsens as β_1 increases ($e^{0.6} - 1 \approx 0.822$). Nevertheless, this inaccuracy has to be put into perspective. If x_1 is scaled so that an increase of one unit is comparatively small, an increase

of 4% in income for each additional unit of x_1 seems plausible and even high. On the other hand, $\beta_1 = 0.6$ would imply an increase of around 82%, which seems extraordinarily high for a comparatively small change in x_1. Therefore, it is safe to expect that most coefficients in a log-linear should be rather close to zero – assuming that there is no confounding that might radically change the magnitude of coefficients.

Finally, the effect of the shift term λ on coefficient interpretation has to be considered. There are two main factors. Firstly, the distortion of the interpretation of β_1 is smaller as λ gets closer to zero or is comparatively small compared to the whole range of y. As λ is usually below the 2% quantile of variable y, it is reasonable to assume that the interpretation as an approximate percentage change can still be used. Secondly, it is crucial to remember that a linear regression is mainly related to the expected value of the dependent variable. β_1 has to be interpreted with this in mind. Focusing on extreme regions of y, shows a larger effect of λ. For example, assume that $y \in (0, 100)$ and $\lambda = 1$. For values near the minimum of y, λ has a very large influence: $\frac{y^{new}+1}{y^{old}+1} = \frac{3+1}{2+1} = 1.33$, while with no shift term λ the ratio between y^{new} and y^{old} would be 1.5. With larger values of y the difference between the ratios of shifted and non-shifted dependent variables is much smaller: $\frac{30+1}{20+1} = 1.48$, which is much closer to $\frac{30}{20} = 1.5$.

In summary, the traditional coefficient interpretation as approximate percentage changes in a log-linear model can also be used with a log-shift transform. The interpretation has some caveats that assume that λ is small compared to the range of y and that the focus is not on extreme values of y. However, the interpretation when $\lambda = 0$ is itself an approximation and it is therefore possible to use it as a rule of thumb when $\lambda \neq 0$.

D. Appendix: Imputation of Extreme Predictions

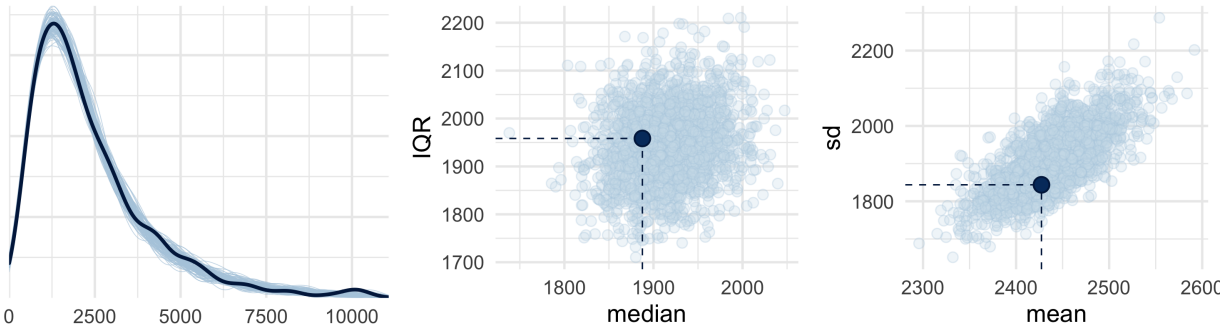
Fitting skewed data is challenging due to the long right tail of the distribution, which can lead to very high predictions. This is especially the case when using the log-shift transformation combined with the Student's t-distribution, which has very heavy tails depending on the degrees of freedom. In practice, these extreme predictions represent only a small proportion (under 0.5%) of the samples from the posterior predictive distribution. Still, to avoid problems with the posterior predictive checks a simple imputation procedure is implemented. The first step is to check which predictions are 10% or more above the

dependent variable y maximum and to mark them as missing. Because the prediction has a very high value, the assumption is that it would still have a higher value otherwise, just not in such an extreme range. Therefore, those predictions are then drawn from $\mathcal{U}(q_{0.99}(y), \max(y))$, i.e., a uniform distribution between the 99%-quantile of y and its maximum. As an example, in the log-scale scenario the maximum of y is around 30.000 while its 99%-quantile is approximately 10.000. While a uniform distribution might seem a crude choice for the imputation, the probability mass above the 99% quantile is almost zero and it should therefore be an adequate approximation. Nonetheless, it is certainly possible to use distributions that place less probability mass on higher values. Finally, it is important to underline that this imputation procedure is a heuristic and that depending on the application it might make sense to use another procedure.

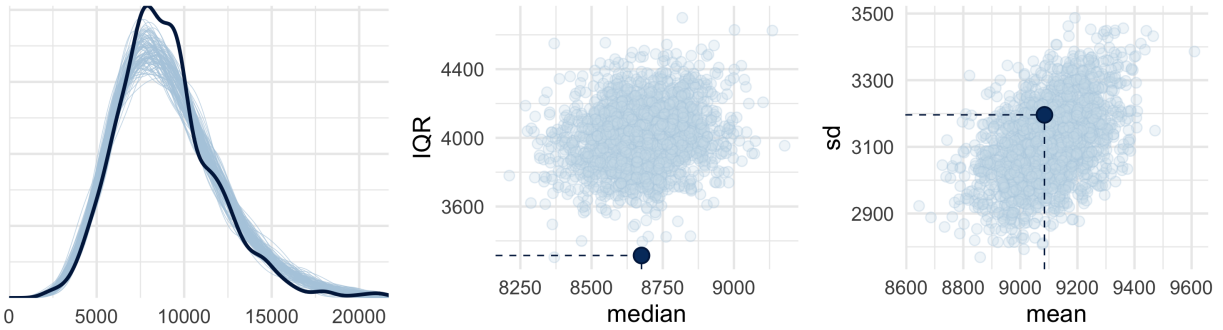
E. Appendix: Additional Posterior Predictive Checks for Skewed Likelihoods

This appendix contains the posterior predictive checks for two models: one with a gamma likelihood and a logarithmic link, and another one with a logscale distribution. The results for the gamma likelihood are shown in Figure 11. The density plots in the first column show that the model can approximate the dependent variable well. There is some divergence around the mode in the GB2 scenario only. In the middle column, the median from the predictions is in a range close to the dependent variable. However, the IQR is too high for the predictions in the GB2 scenario and somewhat high in the Pareto scenario compared to the data. It is not surprising to see in the third column that the model captures the mean and standard deviation better than the log-shift model 4.1, as the mean in the original scale is used directly to parametrize the likelihood. As the gamma distribution has a fixed ratio between expected value and variance, it follows that the standard deviation is also well approximated. Surprisingly, the results for the lognormal distribution in Figure 12 are very similar to the results for the gamma likelihood: some divergence around the mode in the GB2 scenario, a somewhat high IQR for the prediction in the last two scenarios and also a mean and standard deviation that are well captured by the model.

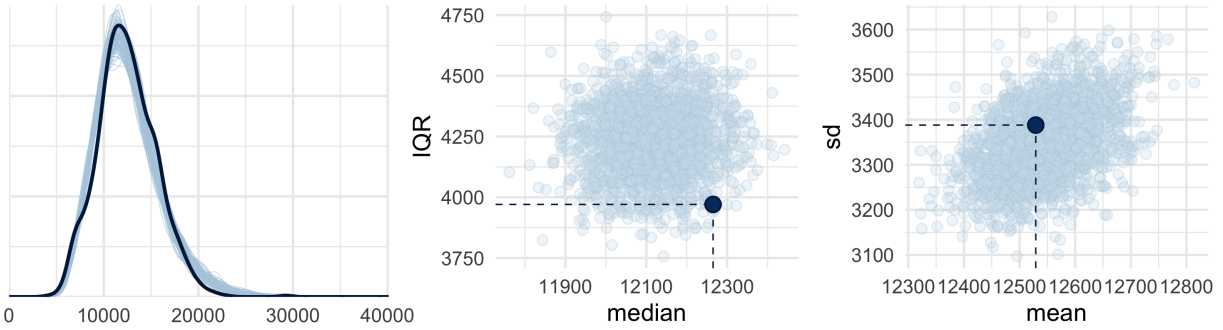
These results suggest that both likelihoods work in a very similar way. On the one hand, both likelihoods imply a multiplicative model through the use of the log link. On the other hand, this might be related to the fact that both distributions have two parameters and also that they both have clear theoretical restrictions: the logarithm of a lognormal distribution must be normal and the ratio between variance and expected value is equal to the shape parameter in the gamma distribution. Although both distributions provide good results, these theoretical restrictions are problematic, because they place strong assumptions on the dependent variable. It is up to the researcher to decide whether these assumptions are acceptable.



(a) Log-scale

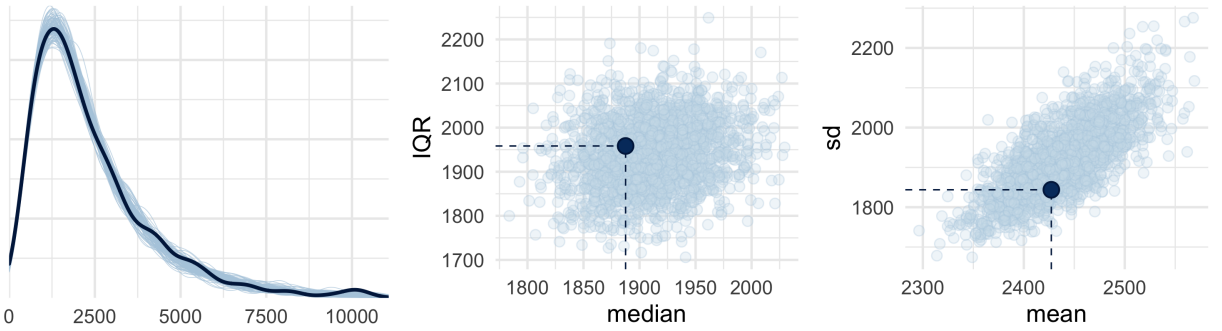


(b) GB2

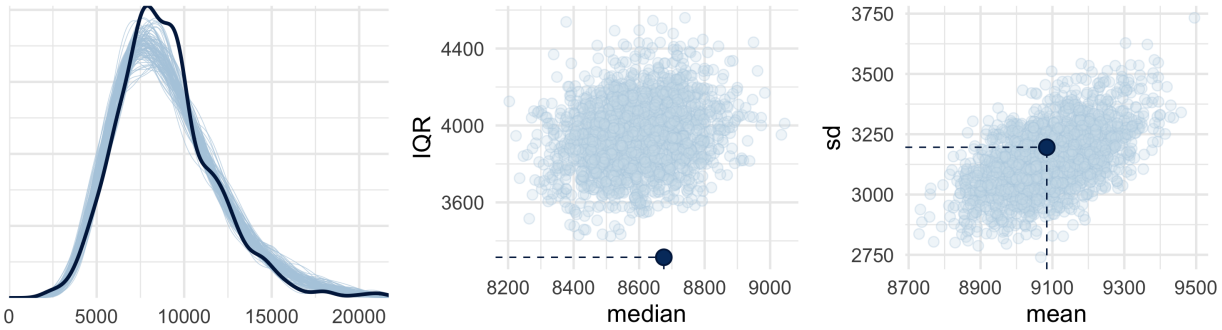


(c) Pareto

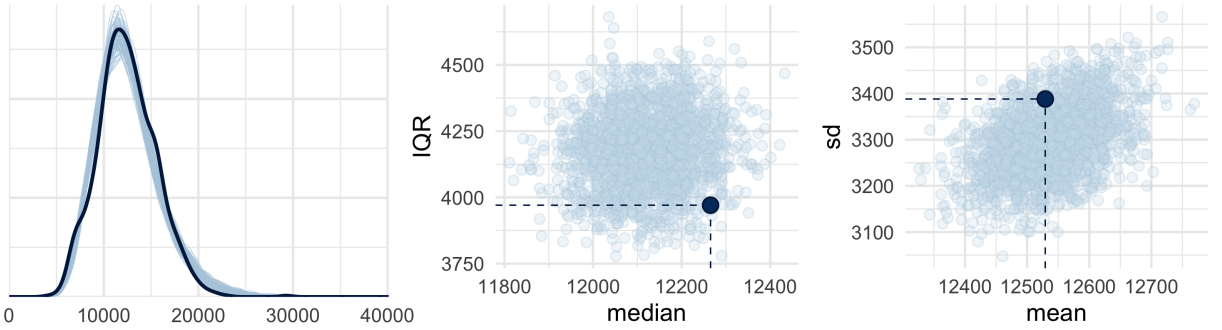
Figure 11.: Posterior predictive check for the gamma likelihood with logarithmic link in all three simulation scenarios. *Left:* density of the dependent variable (black) against the density of 100 backtransformed predictions (light blue). *Middle:* scatterplot of IQR against median for 1000 samples. *Right:* scatterplot of standard deviation against mean for 1000 samples. In the middle and right columns, the dark point represents the respective values for the dependent variable in the original data set.



(a) Log-scale



(b) GB2



(c) Pareto

Figure 12.: Posterior predictive check for the lognormal likelihood in all three simulation scenarios. *Left:* density of the dependent variable (black) against the density of 100 backtransformed predictions (light blue). *Middle:* scatterplot of IQR against median for 1000 samples. *Right:* scatterplot of standard deviation against mean for 1000 samples. In the middle and right columns, the dark point represents the respective values for the dependent variable in the original data set.

F. Appendix: Prior predictive check for Wide Scale Prior

Figure 13 presents the prior predictive checks for a prior $Ga(2, 0.01)$ on the scale parameters σ and σ_u. The samples from the prior predictive distribution have a very extreme range for the logarithmic scale, even when compared to Figure 5a. Note that there is an unnatural accumulation of observations at the extremes of the x-axis of almost all scatterplots. This indicates very extreme samples from $p(y)$ that take the values $\pm\infty$. In this case, the prior on the scale parameters is too wide to produce realistic values.

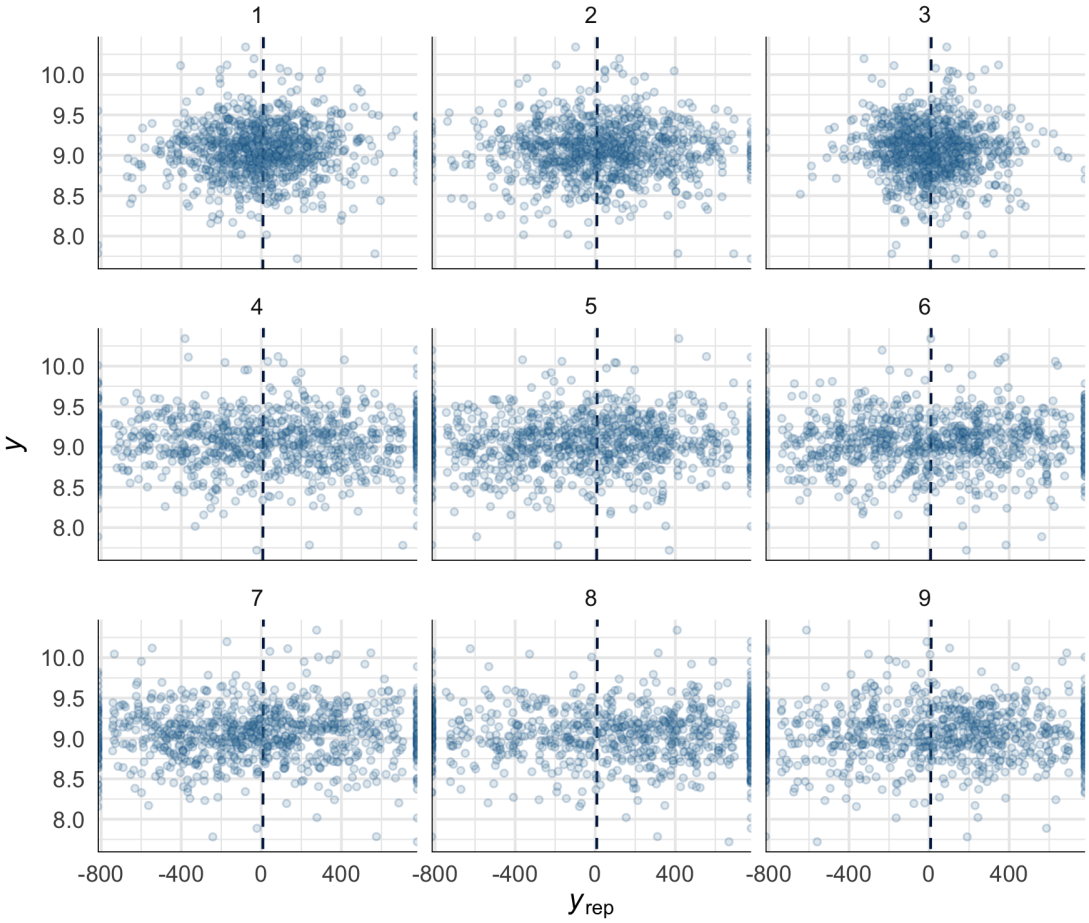


Figure 13.: Prior predictive check for a wide prior $Ga(2, 0.01)$ on the scale parameters σ and σ_u. All values are in the logarithmic scale.

G. Appendix: Density Plots from the Full Horseshoe Model

Figure 14 shows the density plots for the full horseshoe model with the variables from section 3.2. There are two types of variable that can be considered irrelevant to the model. On the one hand, variables such as `jexp` (work experience of head of household) with very small coefficients. On the other hand, the effect of variables such as `jsexo` (gender of head of household) is not clearly positive or negative. Section 4.4 describes how the horseshoe prior is used to choose the relevant predictors.

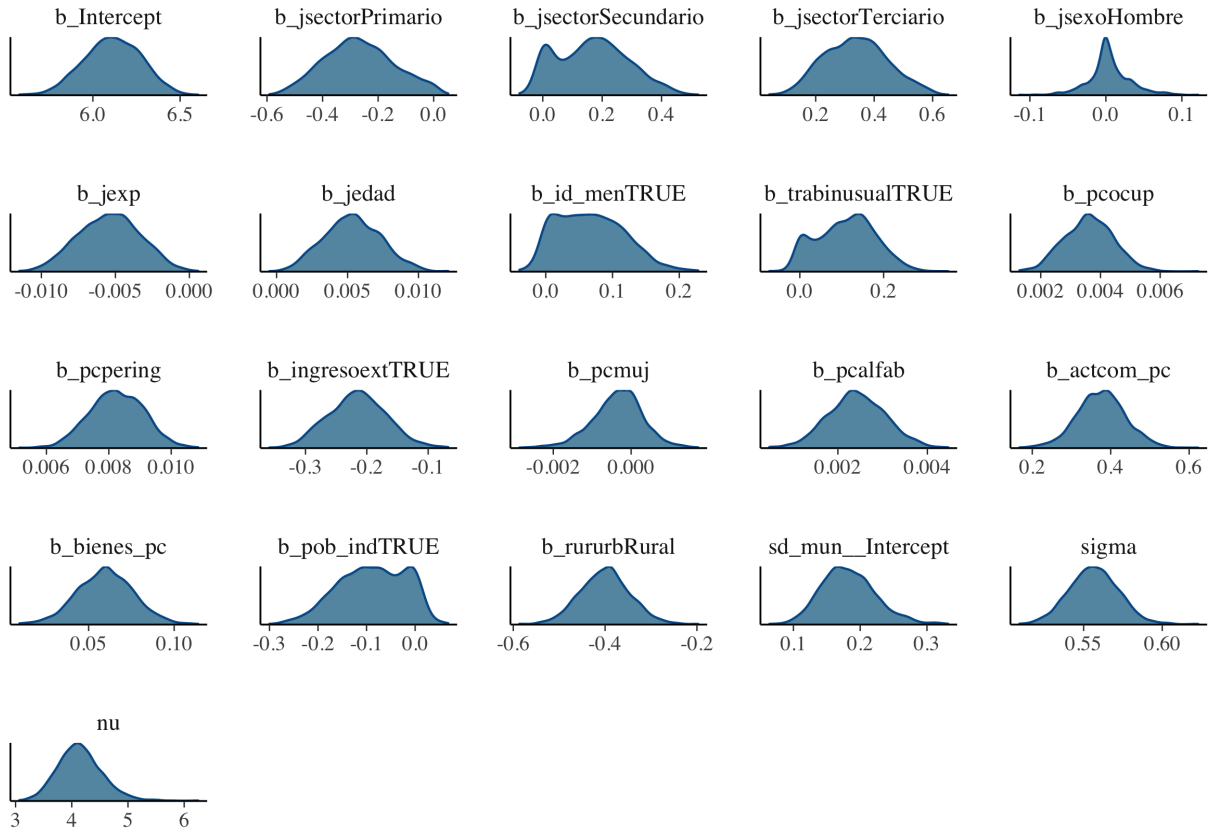
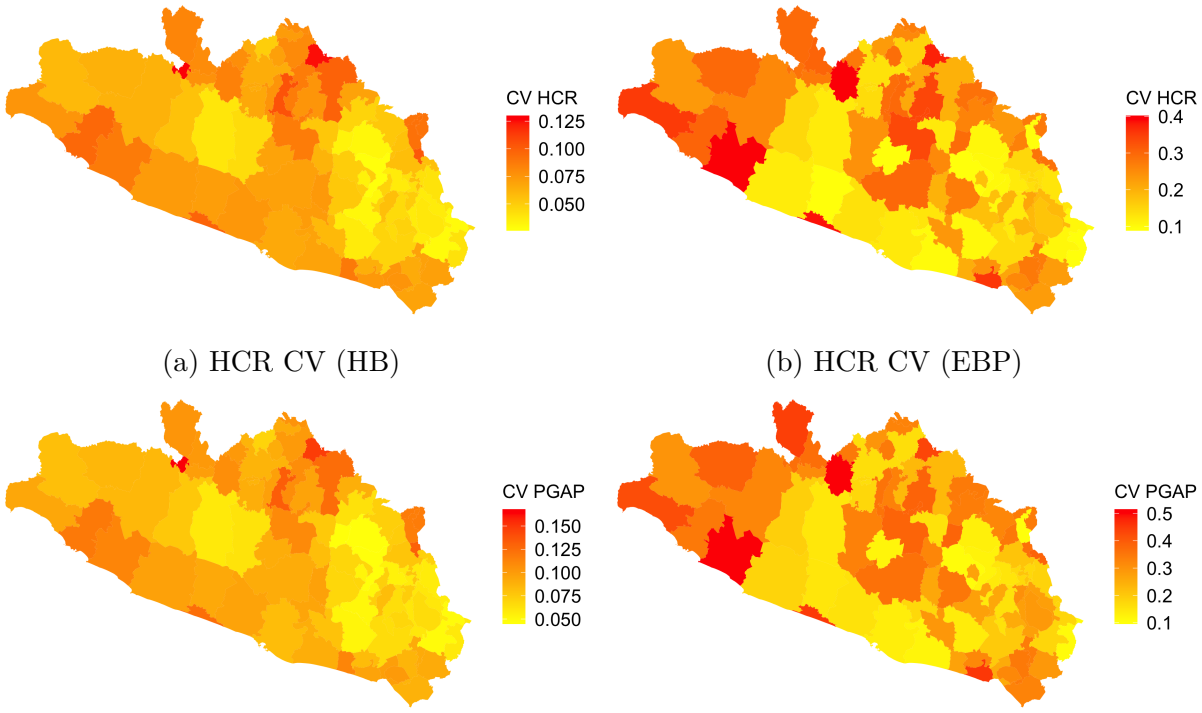


Figure 14.: Density plots of coefficients from full horseshoe model.

H. Appendix: Additional Maps

This appendix presents additional maps with the coefficients of variation (CV). For the HB model, the CV is defined as the standard deviation of the estimate over the estimate itself. In the EBP, the RMSE of the estimate is used instead of the standard deviation. The results are shown in Figure 15. A striking feature of this map is how different the patterns are in the HB and EBP. Regions that have a higher CV in the HB map compared to other regions show the opposite behavior in the EBP map. Also note that in general terms the CV of the EBP model has higher values, which is likely caused by the differences between RMSE and standard deviation.



(a) HCR CV (HB)

(b) HCR CV (EBP)

(c) PGAP CV (HB)

(d) PGAP CV (EBP)

Figure 15.: Coefficient of variation for the HCR and PGAP indicators.

To better understand this pattern, it is useful to take a look at the out-of-sample map in Figure 16. The out-of-sample pattern matches the regions with the highest CV in the

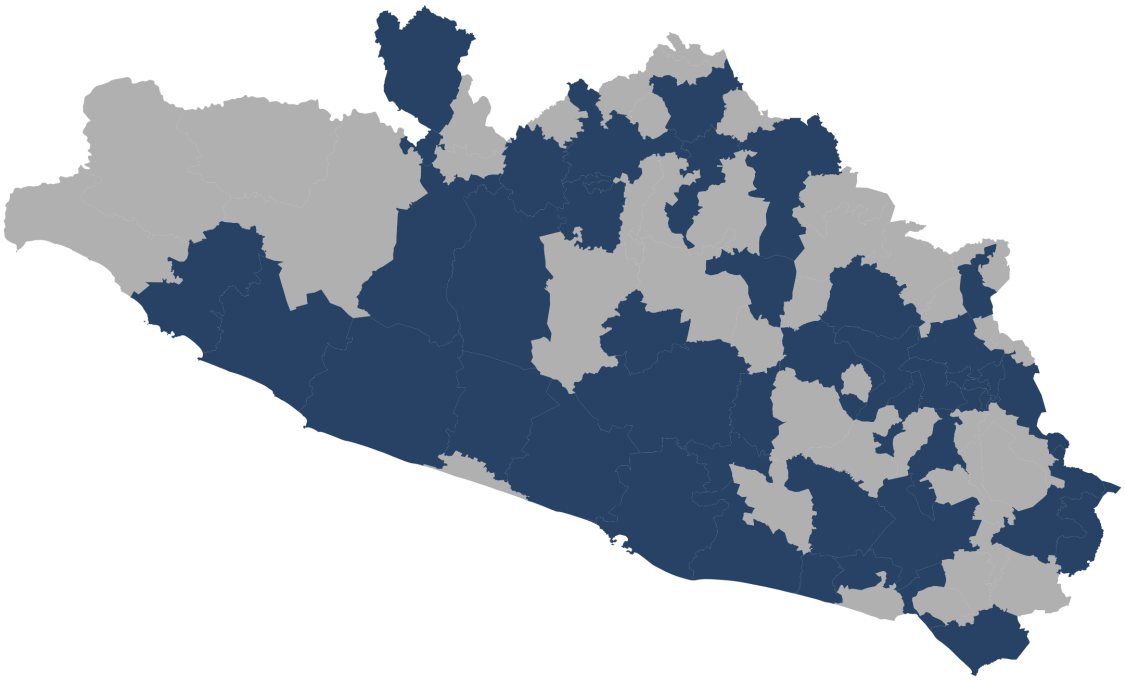


Figure 16.: Map of Guerrero. Light gray areas are out of sample, dark areas in sample.

EBP model, which reflects a higher uncertainty. On the other hand, some regions that have the lowest CV in the HB model are out-of-sample. This shows that redefining the random effect to avoid out-of-sample areas drastically reduces estimate uncertainty.