VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

DATA SCIENCE STUDY PROGRAMME

Master's thesis

Application of hierarchical Bayes models to modelling of distribution of firm size by employees

Hierarchinių Bajeso modelių taikymas modeliuojant įmonės dydžio pasiskirstymą pagal darbuotojus

Kotryna Stankaitytė

Supervisor : Vaidotas Zemlys-Balevičius

Vilnius
2025

Summary

This thesis investigates the application of Bayesian hierarchical models and machine learning methods, such as Bayesian Additive Regression Trees (BART) and Bayesian Neural Networks (BNNs), to the modeling of distribution of firm size by employees, size classes, industries and countries. By applying hierarchical Bayesian model this analysis uses granular socio-economic dataset of indicators number of employees (EMP), turnover (TRN), and enterprise count (ENT) as these datasets exhibit complex hierarchical dependencies, missing data, and diverse scales across groups of industries and size classes. This thesis investigates the group effects on the hierarchical structure and models capability to capture the structure and estimate distribution of number of employees by different size classes. Traditional linear models are not suitable for such complexity and structure of the data motivating the use of hierarchical Bayesian model and machine learning models that are capable of incorporating prior knowledge and investigate complex data structures.

Keywords: Hierarchical Bayes model, BART, BNN, Industries.

Santrauka

Šis darbas nagrinėja Bajeso hierarchinio modeliavimo ir pažangių mašininio mokymosi metodų, tokių kaip Bajeso kombinuoti regresiniai medžiai (BART) ir Bajeso neuroniniai tinklai (BNN), taikymą modeliuojant įmonių dydžio pasiskirstymą pagal darbuotojus, dydžio klases, pramonės šakas ir šalis. Naudojant struktūrinius tikimybių metodus, tyrimo tikslas yra spręsti hierarchinių ir trūkstamų duomenų struktūrų, būdingų socioekonominiams duomenų rinkiniams, keliamus iššūkius. Ekonominiai rodikliai, tokie kaip užimtumas (EMP), apyvarta (TRN) ir įmonių skaičius (ENT), atlieka svarbų vaidmenį suprantant verslo dinamiką ir priimant politikos sprendimus. Tačiau šie duomenų rinkiniai dažnai pasižymi sudėtingomis hierarchinėmis priklausomybėmis, reikšmingais trūkstamais duomenimis ir įvairiu mastu įvairiose pramonės šakų grupėse ir dydžio kategorijose. Šiame darbe nagrinėjamas grupės poveikis hierarchinei struktūrai ir modelių gebėjimui užfiksuoti struktūrą ir įvertinti darbuotojų skaičiaus pasiskirstymą pagal skirtingas dydžio kategorijas. Tradiciniai tiesiniai modeliai nėra tinkami dirbant su sudėtingom hierarchinėm struktūrom, dėl to atsiranda poreikis naudoti hierarchinius Bajeso modelius pažangių mašininio mokymosi metodus kurie geba panaudoti ankstesnę informaciją ir analizuoti sudėtingas duomenų struktūras.

Raktiniai žodžiai: Hierarchiniai, Bajeso, Įmonės, Duomenys

List of Figures

List of Tables

Contents

List of abbreviations

EMP	Number of employees
ENT	Number of enterprises
TRN	Turnover in millions
WG	Wages in millions
HBM	Hierarchical Bayesian model
BNN	Bayesian Neural Network model
BART	Bayesian Additive Regression Trees model
MCMC	Markov Chain Monte Carlo
ESS	Effective Size Sample

Introduction

Distribution of number of employees in granular data together with number of enterprises and its revenue by industry and size class provides insights and input for socio-economic analysis. Understanding the distribution of data and factors that influence it is essential for parties using this data. Firms, when looking at each one individually present different growth patterns. The number of employees and number of enterprises growth should follow similar distribution with majority of number of employees being in the beginning, meaning the most enterprises are clustered within the smaller size class and there are fewer enterprises with more employees presenting heavy tails in the data. This distribution also depends on the industry nature. Standard regression models, though informative, often oversimplify the rich, hierarchical nature of firm data. Hierarchical Bayesian models, with their capacity to handle multilevel structures and account for uncertainty in a principled manner, provide an advanced approach to modeling these distributions. This thesis investigates the application of Bayesian hierarchical models and machine learning methods, such as Bayesian Additive Regression Trees (BART) and Bayesian Neural Networks (BNNs), to the modeling of distribution of firm size by employees, size classes, industries and countries. The goal is to examine Bayesian hierarchical models capability using Stan and Log-Normal distribution do model the distribution using 4 variables in 8 different countries, 96 industries and 6 different size classes and to see if model manages to capture country and industry effects. Hierarchical Bayesian models are not very popular in practical application however they can by applied in policy making to simulate the impact of policy changes on firm size distribution and to identify which industries are most affected by specific policy measures by analyzing group effects.

1 Theory and Literature Review

1.1 Theory

Hierarchical Bayesian methods provide a statistical framework for analyzing data that exhibit complex dependencies or multi-level structures by including priors to the model. These methods extend traditional Bayesian techniques by incorporating additional layers of modeling to account for the hierarchical or nested nature of the data. Bayesian methodology uses prior information about unknown parameters using evidence from observed data. In contrast to classical statistical methods that often rely on fixed-point estimates, Bayesian methods offer a full posterior distribution, capturing the range of plausible values for the parameters.

Bayesian inference is a probabilistic approach that forms the foundation of the modeling methodology employed in this thesis. It provides a systematic framework for updating prior beliefs about parameters in light of observed data. This section outlines the mathematical formulation and its application in the hierarchical Bayesian model used in this research.

The posterior distribution $p(\theta \mid y)$ is computed using Markov Chain Monte Carlo (MCMC) methods. Specifically, the No-U-Turn Sampler (NUTS), a variant of Hamiltonian Monte Carlo (HMC), is used to explore the posterior space efficiently. This approach is suited for hierarchical models due to their high-dimensional parameter spaces and complex posterior geometries. Bayesian inference provides a robust framework for modeling complex nested hierarchical structures in data. By combining prior information with observed data, the hierarchical Bayesian model captures both global and group-specific effects to a multi-level data analysis.

Hierarchical models in Stan have different statistical parameters to improve analysis. Partial pooling allows for estimates to be influenced by group-level information to help avoid the extreme cases of no pooling which produces independence and complete pooling with full aggregation. It leads to more stable and interpretable results for the data that has groups with limited data. Hierarchical Bayesian framework is highly adaptable for various data and analysis. Probabilistic approach of Bayesian models helps with uncertainty and to evaluate both individual and group-level parameters. For data that presents uneven size classes across groups Bayesian approach is suitable as the hierarchical structure allows small groups to benefit from information shared across the entire dataset reducing the risk of over-fitting.

Despite advantages, hierarchical Bayesian models present certain challenges. The complexity often needs advanced computational techniques, such as Markov Chain Monte Carlo (MCMC) or variational inference, which can be computationally intensive. Additionally, the choice of prior distributions requires careful consideration, as it can influence the results, particularly in cases with limited data.

1.2 Background

Firm size distributions provide insights for market analysis, business - to - business data analysis, labour dynamics and economic growth. Data is usually skewed with many small firms and a few very large ones in most industries. Traditional models may struggle to capture the complex data with industry and size class details [2]. Hierarchical Bayesian models offer a powerful alternative, enabling nuanced and robust modeling of firm size distributions by incorporating multi-level structures and probabilistic reasoning. The distribution of firm sizes is influenced by various factors, ranging from individual firm characteristics to broader industry or regional contexts. For industries that do not require highly expensive or specific inventory the distribution should be highly skewed with most firms clustering within the smaller size classes as these business require higher starting costs. Looking at more macro - level external influences such as market conditions, regulatory frameworks, and technological advancements play a significant role is the amount of firms in the market.

The study of firm size distributions has been a longstanding area of interest in economics, with early contributions by Gibrat (1931) introducing the Law of Proportionate Effect, which posits that firm growth is random and independent of size. While Gibrat's law has been foundational, empirical studies have revealed deviations, particularly at the tails of the distribution. These deviations underscore the need for more sophisticated modeling approaches to better capture the nature of the data. Hierarchical Bayesian methods have been applied to analysis for modeling of income distributions, market shares, and organizational growth patterns. Application for firm size distribution is not well researched area posing challenges to methodology creation and working on challenges of skewed data, heterogeneity across industries, and dynamic changes over time.

Traditional econometric models frequently rely on fixed-effects or ordinary least squares (OLS) regressions. These approaches assume homogeneity within groups or simplify the structure of variability by treating it as constant across hierarchical levels, while intuitively a lot of information is kept in the hierarchical level, which then gets ignored and multi-level dependencies such as countries, industries, and firm size classes are not fully accounted. Many models assume normality in residuals or random effects, which is often violated in firm size data due to skewness, heavy tails, or zero-inflated observations. While the log-normal distribution offers an improvement, few studies explore its hierarchical implementation and not on firm data. Although firm size is influenced by nested factors, hierarchical models remain underutilized. A critical gap in the literature, together with lack of application of hierarchical Bayesian models for firm data, is the insufficient emphasis on model validation and diagnostics. Many studies fail to rigorously evaluate whether hierarchical models adequately fit the data and compare its predictive performance against alternatives. While hierarchical Bayesian models are theoretically well-suited for studying firm size, their application in real-world decision-making remains limited.

This thesis shows and example on how hierarchical Bayesian models can inform policy decisions by identifying the drivers of firm size variability looking at the country-level effects that can highlight the impact of national policies on firm growth and industry-specific effects that could inform sectoral strategies to support employment growth. By including real-world categorical variables this thesis gives practical results. Model is constructed to incorporate three levels of variability: country-level effects, industry-level effects, and within-group variability. By modeling these layers, the framework captures both global trends and local heterogeneity in firm size. Overall this thesis provides a new approach for modeling employment using hierarchical Bayesian model and firm data.

1.3 Literature analysis

In business, combined industries, employment and business performance data is utilized and analyzed to derive insights and make decisions. Data modeling is researched using various methods and applications however application of hierarchical Bayesian methods is not deeply researched method. In literature analysis, a number of different papers is analyzed to help derive a methodology combining distribution application, constructing hierarchical Bayesian model and running ML model for comparison.

Fox and Glas (2001) apply hierarchical Bayesian models in item response theory (IRT) using Gibbs sampling. Although primarily focused on psychometrics, their methodology is directly transferable to firm size distribution modeling. The hierarchical structure accounts for variations across firms within different industries or countries, providing insights into how these variables impact the overall distribution. Bayesian hypothesis testing, as discussed by Aitkin (1991), further enhances the ability to compare different hierarchical models or test specific assumptions about firm size distributions, thereby strengthening the decision-making process in model selection. The work by Chib and Greenberg (1995) explains the Metropolis-Hastings algorithm, a crucial MCMC method used for Bayesian inference. This method is essential in the context of firm size distribution modeling because it allows the efficient estimation of posterior distributions when direct sampling is not feasible, making it highly applicable in cases where data is complex or multi-level.

The theoretical foundations of firm size distributions firm size distribution and economic modeling by Axtell (2001) provides empirical evidence that firm size distributions follow a Zipf law, which is a power-law distribution. This finding is fundamental to understanding firm size dynamics because it suggests that a small number of large firms dominate industries while many small firms co-exist. This type of distribution needs to be modeled using advanced statistical techniques, such as hierarchical Bayesian models, to capture the underlying complexity. The presence of heavy tails in firm size distribution, as suggested by Axtell, is a key characteristic that requires special attention in Bayesian modeling. It should be noted that the paper overlooked only the total of economy and did not investigate different distribution in different industries.

Cabral and Mata (2003) offer additional empirical and theoretical insights by studying the evolution of firm size distributions over time. Their findings highlight that young and small firms grow differently compared to large and established firms. This evolution can be captured using hierarchical Bayesian models, where firm-specific characteristics such as age and size class can be included as random effects to account for these varying growth patterns. Sutton (1997) revisits Gibrat's Law, which posits that firm growth is proportional to its current size. While Gibrat's Law has been widely applied in firm size analysis, Sutton's work critically examines its limitations, particularly in explaining deviations from the expected patterns of firm growth. Such deviations can be incorporated into a hierarchical Bayesian framework by including additional covariates, such as industry-specific effects or macroeconomic variables, to capture the heterogeneity in firm growth rates. Luttmer (2007) extends this by modeling the growth and selection mechanisms that shape the distribution of firm sizes. His work emphasizes the importance of understanding how firm-level growth dynamics influence the overall distribution. The use of hierarchical Bayesian models allows for the incorporation of these dynamic factors, enabling a more robust understanding of how firm size distributions emerge and evolve.

Clauset et al. (2009) contribute to the identification and modeling of power-law distributions in empirical data, a pattern commonly observed in firm size distributions. Their methodology for testing whether firm size data follows a power-law distribution is crucial for determining the appropriate statistical model. By using hierarchical Bayesian techniques, it is possible to account for different power-law behaviors across industries or countries, thus refining the model's accuracy.

For Alternative Approaches for Firm Size Modeling using Machine Learning and Econometric Techniques, Andrieu et al. (2003) provide an overview of MCMC methods used in machine learning, which can be applied to the estimation of complex Bayesian models. Their work emphasizes the computational aspects of Bayesian estimation, which is particularly relevant when handling large firm-level datasets. By leveraging machine learning techniques, hierarchical Bayesian models can be optimized for efficiency and scalability, allowing for more accurate predictions in firm size distribution. Analyzing relevant Machine Learning models, applicable for modeling Hierarchal Bayesian models, an introduction to BART models by Hill, Linero and Murray (2020) discusses Bayesian Additive Regression trees model, its application and theoretical understanding of models performance. It highlights it advantages dealing with smaller data sets, debating various applications and different parameters within the model that can be explored and applied dealing with different problems appearing from the specific analysis.

The studies by Axtell (2001), Cabral and Mata (2003), and Sutton (1997) highlight the importance of understanding the underlying distributional characteristics of firm size, such as power-law and heavy-tailed distributions. Hierarchical Bayesian models, as discussed by Fox and Glas (2001), offer a robust framework for capturing the complexity of firm size data, particularly when incorporating industry and country-specific effects. Moreover, the insights from Clauset et al. (2009) on power-law distributions provide valuable tools for refining the Bayesian models to better fit firm size data. Finally, combining Bayesian methods with machine learning and econometric techniques, as suggested by Andrieu et al. (2003) can enhance the scalability and predictive accuracy of the models.

2 Data and Methodology

2.1 Data

The dataset used for this thesis is from Eurostat, a leading statistical authority in the European Union, which consolidates data provided by national statistical offices of its member states. Tables Annual enterprise statistics by size class for special aggregates of NACE Rev.2 activities (2005-2020) and Enterprise statistics by size class and NACE Rev.2 activity (from 2021 onwards) were used. Datasets provides historical data on enterprise-level metrics segmented by size class across various sectors of economic activity. Data is detailed using NACE2 classification of economic activity and is detailed to 3 digit level including industries like agriculture, mining, manufacturing and services. For analysis, industries in 1-3 digit level were chosen.

Four variables were taken from the dataset for the analysis. Enterprises - number, represents the count of firms operating within each industry and size class. This indicator provides insights into the density and distribution of businesses across various economic activities. Turnover or Gross Premiums Written (Million Euros) Measures the revenue generated by enterprises. For service-oriented industries like insurance, ”gross premiums written” replaces turnover as a more representative metric of financial activity. This indicator offers an understanding of economic productivity and sectoral contribution. Wages and Salaries (Million Euros), captures the total compensation paid to employees, including salaries, bonuses, and other forms of remuneration. This serves as a proxy for labor cost and an indicator of economic contribution by employees within firms and Employees (Number), that quantifies the total number of employees working in enterprises within each size class and industry. It is the central variable of interest for this thesis.

2.2 Descriptive statistics

To better understand the data descriptive statistics are presented in the tables below. The first table summarizes the key statistical metrics for four indicators: EMP (Employment), ENT (Enterprises), TRN (Turnover), and WG (Wages). These metrics provide an overview of the distribution and variability of these indicators across the dataset.

Data summary in Table 1 table. by indicator shows that employment levels are highly skewed with a mean value of approximately 121,958 and a median of 15,975. The maximum value of 7,868,166 further highlights the presence of outliers or exceptionally large firms and standard deviation of over 406,595 suggests substantial variability in employment figures across entities. Similarly, Number of enterprises exhibit a mean size of 18,771 with a median of only 443, indicating a large number of smaller enterprises and a few significantly larger ones. Nr. of obs.in the table shows the number of observations for each variable in the data set. For employment there are 4239 observations in the data set while for number of enterprises and turnover, respectively, 47505 and 44888 and wages, 869 observations. The maximum value of 1,174,254 and a standard deviation of 70,183 highlight the wide range of enterprise sizes. Turnover and wages support the assumption of skewed data with substantial differences between mean and median. The summary in 1 table highlights the diversity and variability within the dataset, particularly for employment and enterprise sizes. This variability underscores the need for robust statistical models, such as hierarchical Bayesian approaches, to capture these complexities effectively.

1 table. *Summary Statistics by Indicator*

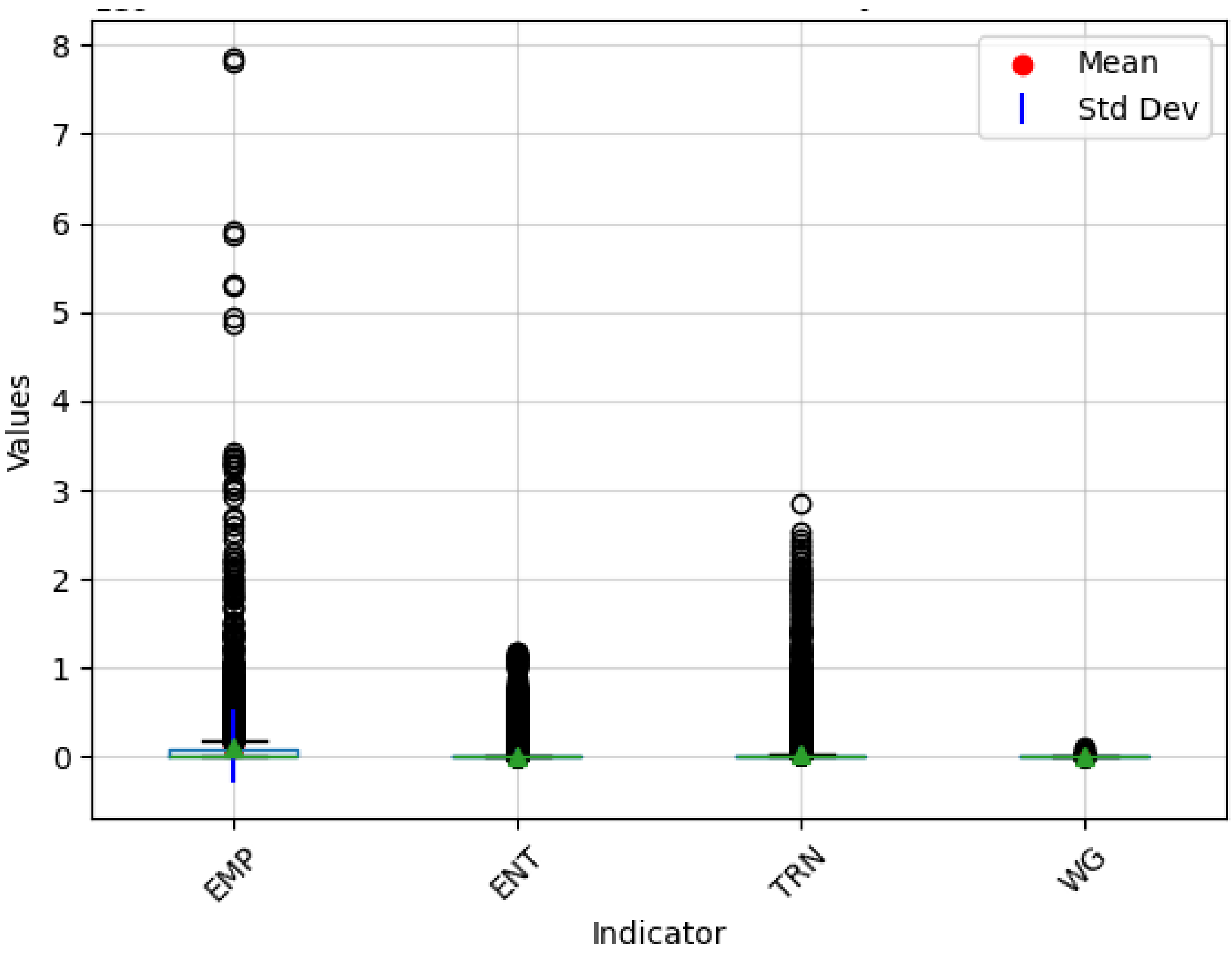
Indicator	Nr. of obs.	Mean	Median	Std. Dev.	Min	Max
EMP	4,239	121,958.24	15,975.0	406,595.00	0	7,868,166
ENT	47,505	18,771.40	443.0	70,183.89	0	1,174,254
TRN	44,888	23,638.25	2,781.0	97,865.32	0	2,842,707
WG	869	1,717.41	309.0	6,312.26	0	104,492

For data by industry analysis only the least aggregated industries were chosen. The Table 2 table. provides a detailed breakdown of employment statistics across different industries, showing the variability and distribution of employment figures within each sector.

The average employment size varies significantly across industries, ranging from approximately 16,210 (Industry B (Mining and quarrying)) to over 2.69 million (Industry G (Wholesale and retail trade; repair of motor vehicles and motorcycles)). The variability is particularly pronounced in industries such as C (Manufacturing), G , and Q (Human health and social work activities), which have the highest standard deviations (over 1.4 million, 1.7 million, and 1.8 million, respectively). Certain industries, such as C and G , have extremely high maximum values (7,868,166 and 5,924,954, respectively), indicating the presence of large firms that dominate employment figures in those sectors. Industries with smaller mean values, such as B (Mining and quarrying), D (Electricity, gas, steam and air conditioning supply), and L (Real estate activities), have tighter distributions, as indicated by their lower standard deviations. In most industries, the median is significantly lower than the mean, suggesting a right-skewed distribution where a few large firms substantially increase the average employment figures

2 table. *Summary Statistics of Employment by Industry*

Industry	Nr. of obs.	Mean	Median	Std. Dev.	Min	Max
B	86	16,209.52	3,990.0	35,951.08	140	173,068
C	96	1,010,207.00	475,383.5	1,442,724.00	14,444	7,868,166
D	94	43,929.96	5,504.0	74,378.44	390	316,634
E	92	54,667.51	19,620.0	69,344.21	485	305,636
F	82	389,435.00	216,998.0	478,665.20	13,529	2,460,625
G	12	2,696,077.00	2,393,198.5	1,789,016.00	211,083	5,924,954
H	12	1,044,625.00	904,917.5	660,483.10	150,042	2,200,768
I	12	941,940.40	1,129,859.5	640,613.70	43,026	1,873,477
J	12	642,038.70	535,350.5	475,821.40	39,348	1,506,281
K	12	466,924.20	423,771.5	333,833.80	21,289	1,061,118
L	12	181,457.60	135,878.0	153,643.50	17,357	456,436
M	12	841,340.60	664,171.5	683,840.40	49,641	2,135,906
N	12	1,404,240.00	1,392,455.5	1,016,876.00	62,382	3,057,644
P	12	196,597.80	131,404.5	164,839.80	9,588	426,700
Q	12	1,282,876.00	694,597.0	1,897,274.00	30,172	5,304,710
R	12	157,347.80	130,220.5	131,494.60	7,957	369,973



1 figure. *Boxplot of values by indicator*

Descriptive analysis of outliers showed that employment indicator presents significant number of outliers and the spread compared with other variables emphasizes its variability. The employment indicator is the most volatile and is likely influenced by industry-specific factors.

2.3 Distributions Fitted

For each combination of Unified Size Class, which represents different size classes for comparability, and indicator, `EMP` for the number of employees and `ENT` for the number of enterprises, two distributions were fitted: Log-Normal and Poisson. The summary of parameter estimates, log-likelihood values, and model selection criteria (`AIC` and `BIC`) for each class is presented below.

3 table. *Fitted distributions for Class 1 (0–9 employees).*

Distribution	Parameter 1	Parameter 2	Log-likelihood	AIC / BIC
Log-Normal (EMP)	meanlog = 5.00	sdlog = 2.20	-351205.8	702415.7 / 702433.2
Poisson (EMP)	lambda = 1587.79	N/A	-185345673	370691347 / 370691356
Log-Normal (ENT)	meanlog = 6.50	sdlog = 2.59	-2804588	5609180 / 5609201
Poisson (ENT)	lambda = 14363.89	N/A	-11840169769	23680339539 / 23680339550

Unified Size Class 2: 10–19 Employees

4 table. *Fitted distributions for Class 2 (10–19 employees).*

Distribution	Parameter 1	Parameter 2	Log-likelihood	AIC / BIC
Log-Normal (EMP)	meanlog = 6.34	sdlog = 1.84	-55495.46	110994.9 / 111008.5
Poisson (EMP)	lambda = 4358.98	N/A	-61386726	122773454 / 122773461
Log-Normal (ENT)	meanlog = 5.22	sdlog = 2.62	-2204905	4409813 / 4409834
Poisson (ENT)	lambda = 8112.95	N/A	-7624993405	15249986813 / 15249986823

Unified Size Class 3: 20–49 Employees

5 table. *Fitted distributions for Class 3 (20–49 employees).*

Distribution	Parameter 1	Parameter 2	Log-likelihood	AIC / BIC
Log-Normal (EMP)	meanlog = 6.94	sdlog = 1.70	-60697.56	121399.1 / 121412.8
Poisson (EMP)	lambda = 5460.92	N/A	-66315124	132630249 / 132630256
Log-Normal (ENT)	meanlog = 4.14	sdlog = 2.07	-1753168	3506341 / 3506362
Poisson (ENT)	lambda = 832.86	N/A	-633231551	1266463105 / 1266463115

Unified Size Class 4: 50–249 Employees

6 table. *Fitted distributions for Class 4 (50–249 employees).*

Distribution	Parameter 1	Parameter 2	Log-likelihood	AIC / BIC
Log-Normal (EMP)	meanlog = 5.29	sdlog = 2.46	-146715.8	293435.6 / 293451.4
Poisson (EMP)	lambda = 3380.62	N/A	-158951720	317903442 / 317903450
Log-Normal (ENT)	meanlog = 4.65	sdlog = 2.57	-3226536	6453076 / 6453098
Poisson (ENT)	lambda = 6140.79	N/A	-10232691908	20465383819 / 20465383830

Unified Size Class 5: 250+ Employees

7 table. *Fitted distributions for Class 5 (250+ employees).*

Distribution	Parameter 1	Parameter 2	Log-likelihood	AIC / BIC
Log-Normal (EMP)	meanlog = 6.50	sdlog = 2.67	-99942.62	199889.2 / 199903.9
Poisson (EMP)	lambda = 11101.95	N/A	-287251220	574502442 / 574502449
Log-Normal (ENT)	meanlog = 2.41	sdlog = 1.55	-489673.1	979350.3 / 979369.6
Poisson (ENT)	lambda = 60.59	N/A	-14852427	N/A

Across all Unified_Size_Class and both number of employees and number of employees indicators, the Log-normal distribution provides a better fit to the data than the Poisson distribution. This is indicated by substantially lower AIC/BIC values and better loglikelihood values. Log-normal distribution is better suited to handle the large variability and positive skew typical of firm size distributions, which is why it outperforms the Poisson distribution in terms of AIC, BIC, and loglikelihood as high variability is presented due to different parameters and values for different industries.

2.4 Methodology

The methodological approach of this thesis involves fitting firm size distributions using Log-normal models, handling heterogeneity across firm size classes and data sources, and addressing missing values. The Two primary distributions were tested Log-Normal and Poisson, based on their relevance to firm size and count data. Final choice in the analysis was to use the Zero-Adjusted Poisson (ZAP) model to handle and impute missing values.

Modeling firm size distributions presents unique challenges due to the broad range of firm sizes and their skewed distribution. Firm size data, which often exhibit a long tail due to a few large firms among many small firms. However a different challenge becomes the dataset where within variables negative values and high aptitude of values appear.

For smaller firm sizes, which are typically count-based, the Poisson distribution is appropriate due to its properties as a discrete count distribution. The Poisson distribution has been applied in numerous studies to model count data, particularly when data are skewed toward lower values [1]. Using the fitdistrplus package in R, a Poisson distribution was fitted to firm sizes within each class, applying it to integer count data where values were all non-negative. This approach aligns with the practice of modeling small-scale business data as discrete occurrences [1].

The presence of missing values in the dataset posed a significant challenge, as their imputation needed to reflect both zero-inflation and the count nature of firm size data. Data is granular to the industry level by number of employees including negative, zero and high values creating a wide amplitude within the data. The Zero-Adjusted Poisson (ZAP) model, implemented using the gamlss package in R, was selected due to its ability to handle excess zeros in count data [5]. The ZAP model combines a zero component with a Poisson component, effectively capturing the distinct likelihood of zeroes and positive counts in the data. This approach has been validated in studies involving zero-inflated and count data, providing robust solutions for missing value imputation [5].

The ZAP family was used, specifying the Poisson distribution as the non-zero component. A maximum iteration count of 20 was set to ensure that the model converged adequately, with global deviance monitored at each iteration. A steady decrease in deviance was observed, confirming model improvement. Monitoring the global deviance values across iterations showed consistent reductions, indicating successful optimization. After fitting, the ZAP model's predicted mean values (λ) for each missing entry were used to impute missing values, generating samples based on the Poisson component's expected counts. This approach to imputation aligns with methods for the structure of zero-inflated count data [5].

Following the fitting of the ZAP model, missing values were imputed using the expected values from the model's Poisson component. For each missing value, a sample was drawn from the Poisson distribution centered on the predicted mean (λ). This imputed dataset was then transformed using a natural logarithm to facilitate hierarchical Bayesian modeling.

With the imputed data, a hierarchical Bayesian model was developed to capture firm size variations across levels of industry and country. This 3-level structure was designed to account for hierarchical dependencies:

- Level 1: Firm size distribution within specific industry-country combinations, capturing within-group variance.
- Level 2: Country-specific variations in firm size distribution, accounting for regional differences.
- Level 3: Industry-wide variations across countries, capturing cross-industry patterns.

The fitted hierarchical Bayesian model was implemented in Stan, using Markov Chain Monte Carlo (MCMC) sampling to estimate parameters for each level. This approach has been validated in previous research for its effectiveness in multi-level data contexts. Due to high computational costs, only a sample of the data was chosen, analyzing 8 countries and industries at the highest industry code level at 1 - 3 levels.

In this thesis, observations were influenced by both global effects and group-specific effects at multiple levels. The model incorporates global effects that represents overall population-level parameters, group-specific effects that represents random effects for groups of countries, industries, and indicators and measurement noise for within-group variability.

The hierarchical model is defined as:

$$y_i \sim \mathcal{N}(\mu + u_{\text{country}[i]} + v_{\text{industry}[i]} + w_{\text{indicator}[i]}, \sigma_{\text{within}})$$

where:

- y_i: Observed response for observation i (e.g., log-transformed firm size).
- μ: Global mean across all observations.
- $u_{\text{country}[i]}$: Random effect for the country associated with observation i.
- $v_{\text{industry}[i]}$: Random effect for the industry associated with observation i.
- $w_{\text{indicator}[i]}$: Random effect for the indicator associated with observation i.
- σ_{within}: Standard deviation of within-group noise.

To estimate the group-level random effects, hierarchical priors were imposed. These priors allow partial pooling, which is particularly effective in managing the trade off between over-fitting and under-fitting for groups with few observations.

Bayesian hierarchical model was employed using the Hamiltonian Monte Carlo (HMC) algorithm with the No-U-Turn Sampler (NUTS).Higher `delta` and `max_depth` settings prioritize robustness, while default adaptation settings ensure efficient tuning. Overall, these settings were optimized to handle the large parameter space and complex geometries typical of hierarchical Bayesian models for firm data used for the model.

Global prior is defined as below:

$$\mu \sim \mathcal{N}(0, 5)$$

Random Effects Priors:

$$\begin{aligned} u_{\text{country}[j]} &\sim \mathcal{N}(0, \sigma_{\text{country}}), \quad j = 1, \ldots, J \\ v_{\text{industry}[k]} &\sim \mathcal{N}(0, \sigma_{\text{industry}}), \quad k = 1, \ldots, K \\ w_{\text{indicator}[l]} &\sim \mathcal{N}(0, \sigma_{\text{indicator}}), \quad l = 1, \ldots, L \end{aligned}$$

Non-negative priors are applied to the standard deviations to ensure positivity:

$$\sigma_{\text{within}}, \sigma_{\text{country}}, \sigma_{\text{industry}}, \sigma_{\text{indicator}} \sim \mathcal{N}^+(0, 2)$$

The likelihood function is defined as:

$$y_i \sim \mathcal{LN}(\mu + u_{\text{country}[i]} + v_{\text{industry}[i]} + w_{\text{indicator}[i]}, \sigma_{\text{within}})$$

where the components $u_{\text{country}[i]}, v_{\text{industry}[i]}, w_{\text{indicator}[i]}$ capture deviations specific to the hierarchical levels and to use Log-normal distribution following the nature of data not having naturally negative values.

To assess model fit, posterior predictive checks were performed by generating replicated data which was compared to the observed data to evaluate the model's ability to capture key patterns. Bayesian Hierarchical models combine information across groups to improve estimates with sparse data, accommodating varying group sizes and nested structures, providing credible intervals and posterior distributions for all parameters and allowing the inclusion of domain knowledge through informative priors.

8 table. *Summary of Priors for the Hierarchical Bayesian Model*

Parameter	Prior Distribution	Description
μ	$\mathcal{N}(0, 5)$	Global mean across all observations
$u_{\text{country}[j]}$	$\mathcal{N}(0, \sigma_{\text{country}})$	Random effect for country j
$v_{\text{industry}[k]}$	$\mathcal{N}(0, \sigma_{\text{industry}})$	Random effect for industry k
$w_{\text{indicator}[l]}$	$\mathcal{N}(0, \sigma_{\text{indicator}})$	Random effect for indicator l
σ_{within}	$\mathcal{N}^+(0, 2)$	Standard deviation of within-group noise
σ_{country}	$\mathcal{N}^+(0, 2)$	Standard deviation of country-level random effects
σ_{industry}	$\mathcal{N}^+(0, 2)$	Standard deviation of industry-level random effects
$\sigma_{\text{indicator}}$	$\mathcal{N}^+(0, 2)$	Standard deviation of indicator-level random effects

Number of Samples was set to 2000 and warmup iterations to 500. Warmup iterations were discarded after calibration. The thinning parameter was set to 1, meaning all iterations after warmup were retained. Thinning is not necessary for modern HMC samplers due to their efficient sampling mechanisms. For target acceptance probability `delta` value of 0.99 was used to target a high acceptance probability. Delta was set in a conservative matter to reduce the likelihood of divergent transitions, which are common in hierarchical models with complex posterior geometries which appears due to high variability of industries in this thesis.

The NUTS-specific parameters for the maximum depth for the binary tree in NUTS was set to 12, allowing up to $2^{12} = 4096$ leapfrog steps. Some transitions reached this limit, indicating challenging regions of the posterior. The diagonal Euclidean metric (`diag_e`) was used for mass matrix adaptation to reduce computational complexity compared to a dense metric for high-dimensional hierarchical models. Step size was initialized at 1, and subsequent adaptation during warmup to ensure optimal performance.

As data shows outliers, regularization and random effects priors were applied in the model to constrain parameter estimates to plausible ranges, improve sampler efficiency and mixing, reduce over-fitting by penalizing extreme parameter values, and address high R-hat values by simplifying the posterior distribution.

Proper initialization reduces the risk of divergent transitions and ensures faster convergence of the chains, particularly in hierarchical models with complex posterior geometries. For global parameters initial values were set near the center of the prior distribution to align with prior beliefs. Random effects, u_{country}, v_{industry}, and $w_{\text{indicator}}$ were initialized using small random perturbations around zero, reflecting the assumption of no strong prior bias. Standard deviations $\sigma_{\text{within}}, \sigma_{\text{country}}, \sigma_{\text{industry}}, \sigma_{\text{indicator}}$ were initialized with positive values to respect the non-negativity constraint.

Effective Sample Size for key parameters was monitored to ensure sufficient sampling efficiency. Low ESS values indicate slow mixing or high autocorrelation, suggesting the need for further regularization or reparameterization. The split $\hat{R}$ diagnostic was used to evaluate convergence. Parameters with $\hat{R} > 1.05$ indicate incomplete mixing and may require stronger priors or simplified parameterizations. Approximately 5.5% of transitions exceeded the maximum treedepth of 12, highlighting regions of high curvature in the posterior.

Cross validation and posterior check was conducted on the Log - Normal distribution likelihood model. Training and validation data sets were splitted to calculate RMSE, log-likelihood and MAE. The training model was done using 4 folds and validation using the remaining. Robustness check was done to analyze the sensitivity of the model to different priors, posterior predictive check were conducted and comparing of the slight perturbations in data.

Results and conclusions

3 Results

The results of the Bayesian hierarchical model are summarized in Table 9 table.. The model was constructed to estimate the relationship between firm size (log-transformed) and its underlying structure across multiple levels: within countries, between countries, and within industries. Key parameters such as the overall mean (μ) of firm size, the within-group standard deviation (σ_{within}), and the between-country standard deviation (σ_{country}) are presented.

For a comparison, linear model is summarized in Table 10 table..

9 table. *Bayesian Hierarchical Model Results (Log-Normal Distribution)*

Parameter	Mean	Std. Dev.	50%
Overall Mean (μ)	-254548	7.446	-254548
Within-Group SD (σ_{within})	1.052	0.573	1.036
Between-Country SD (σ_{country})	3.365	0.007	3.365
Industry Effect SD (σ_{industry})	0.434	0.393	0.446
Indicator Effect SD ($\sigma_{\text{indicator}}$)	0.416	0.395	0.423

10 table. *Linear Model Results*

Statistic	Value
Dependent Variable (y)	EMP (Firm Size by Employees)
R-squared	0.924
Adjusted R-squared	0.919
F-statistic	189.5
Prob (F-statistic)	0.000
Log-Likelihood	-10819
AIC	21740
BIC	22000
Number of Observations	867
Number of Predictors	52
Mean RMSE (Cross-Validation)	68613.73

3.1 Results analysis

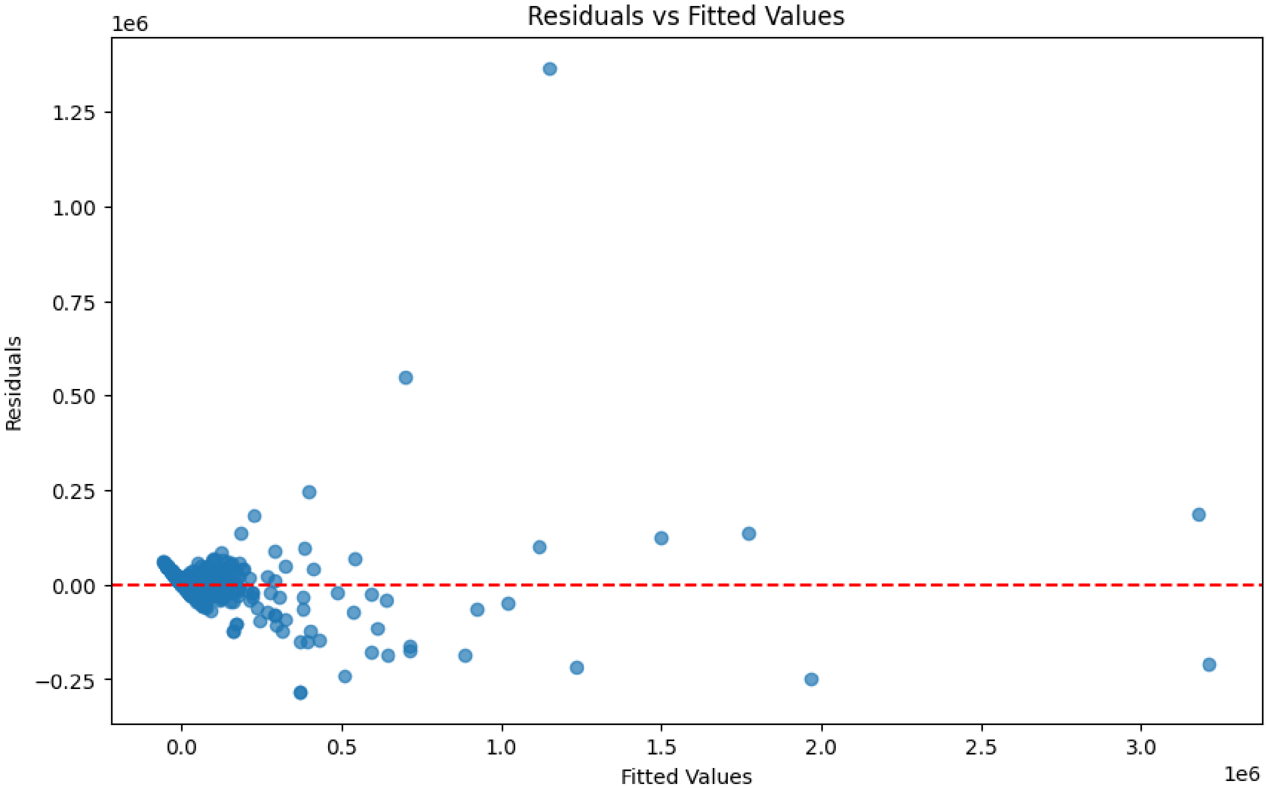
When applying the Log-normal distribution, the estimated mean $\mu = -254548$ appears implausible and reflects numerical instability, likely caused by excessive influence of zero or near-zero values in the data indicating a poor model fit for overall mean under the Log-normal assumption.

The variation within specific industry-country groups, $\sigma_{\text{within}} = 1.052$ indicates relatively low variability within groups, reflecting a significant reduction in dispersion compared to other potential models. This aligns with expectations, as the Log-normal transformation often suppresses within-group variability by shifting the distribution towards a more concentrated range. The posterior standard deviation (0.573) and credible intervals suggest this estimate is reasonably stable, though less robust than expected in certain cases.

Variability in firm size attributable to differences across countries, $\sigma_{\text{country}} = 3.365$, demonstrates increased sensitivity to country-level effects under model. The narrower credible interval reflects a stronger differentiation between countries, potentially amplifying the influence of outliers or skewed data. This suggests that the model effectively captures variations stemming from differences in national policies, economic conditions, or industrial structures. However, this amplification also indicates that approach may be less resilient to data anomalies, requiring careful handling of outliers.

Industry-specific random effects (v_{industry}) show considerable variability, with slightly tighter credible intervals compared to other models. This reflects the ability of the log-normal distribution to capture industry-level differences while still highlighting potential convergence challenges in parameter estimation. The hierarchical structure remains crucial for disentangling these effects, although further re-evaluation of industry classifications or increased sampling may help address issues related to convergence.

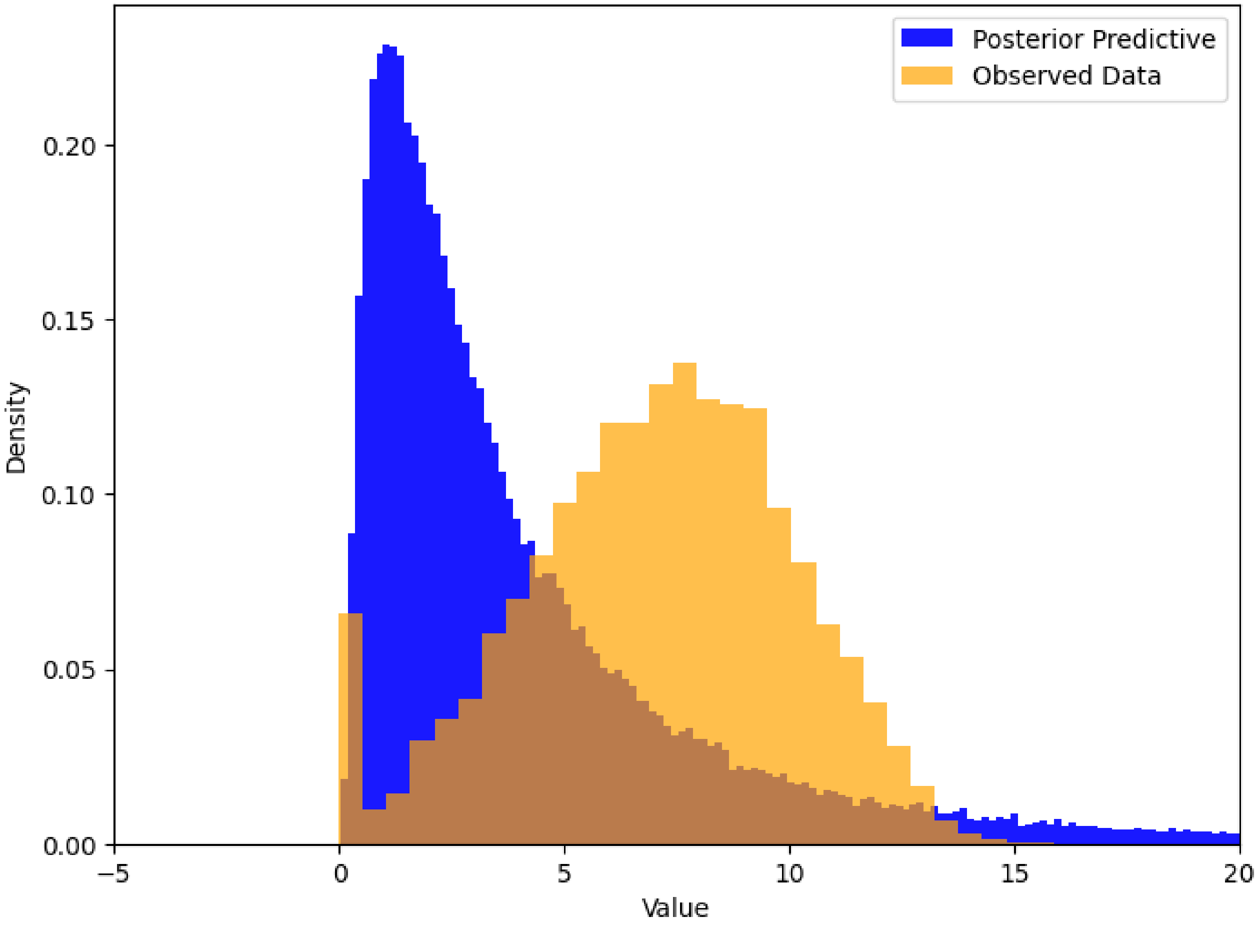
While Bayesian model explicitly captures the nested structure of countries, industries, indicators through group-level effects, the linear model assumes a single global linear relationship between predictors and the response variable, ignoring the hierarchical structure. It also assumes homoscedasticity, which is clearly violated as is seen from Figure 2 figure.. The linear model doesn't account for the skewed distribution of firm size for employees. The residuals plot and large coefficients shows that the model struggles with these extreme values. With a large number of categorical predictors countries, industries, size classes, the linear model is overfitting, as indicated by the high (R^2) but poor interpretability and the multicollinearity warnings. Mean RMSE for Cross-Validation is also worse for linear model than for Hierarchical model.



2 figure. *Posterior Predictive Checks Plot for LogNormal Distribution*

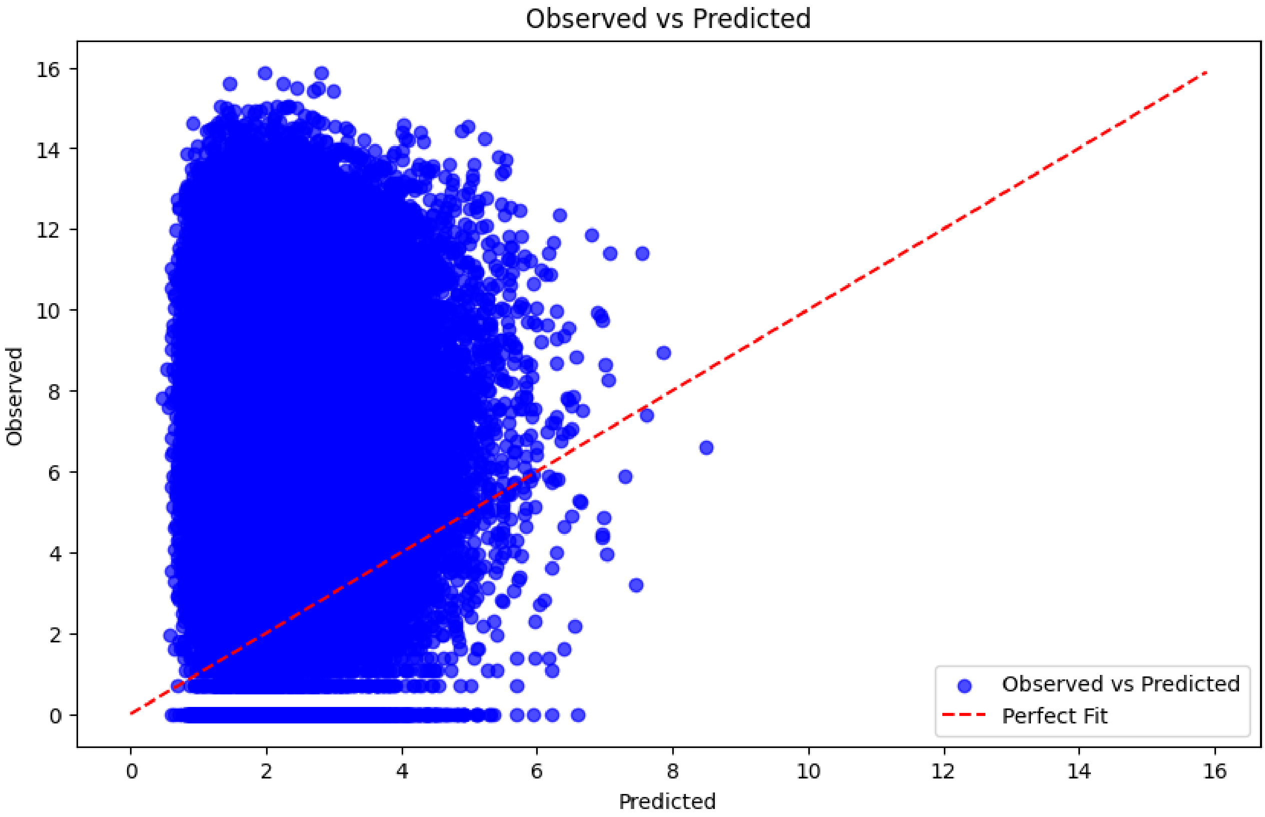
In conclusion, the Log-normal distribution provides a nuanced representation of variability across countries and industries, particularly in amplifying the country-level effects. However, numerical instability in the global mean estimate suggests that additional data preprocessing or transformations are necessary to improve model performance. The model captures the significant influence of national policies, market conditions, and cultural differences on firm size but highlights the need for careful consideration of data characteristics, such as skewness and structural zeros, in applying Log-normal assumptions.

3.2 Posterior Predictive Checks



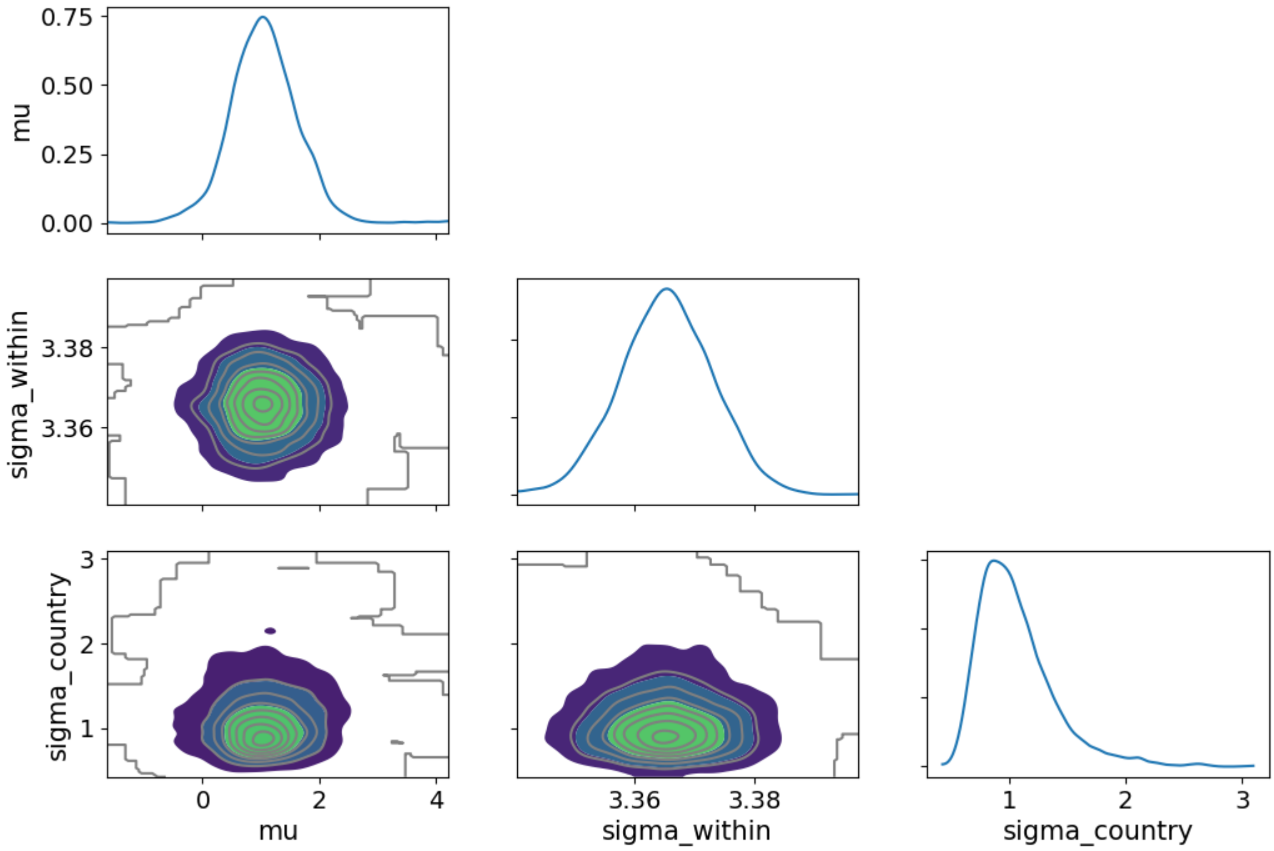
3 figure. *Posterior Predictive Checks Plot for Log-Normal Distribution*

The posterior predictive plot of Log-Normal distribution Figure 3 figure. shows that the observed data has a broader spread and a more moderate right-skew, with values extending further along the positive axis. The Log-Normal distribution does not fully align with the empirical data because of the observed heavier tails and variability. While Log-Normal models are suitable for strictly positive and skewed data, the observed dataset contains heavier tails that deviate from a strict Log-Normal shape.



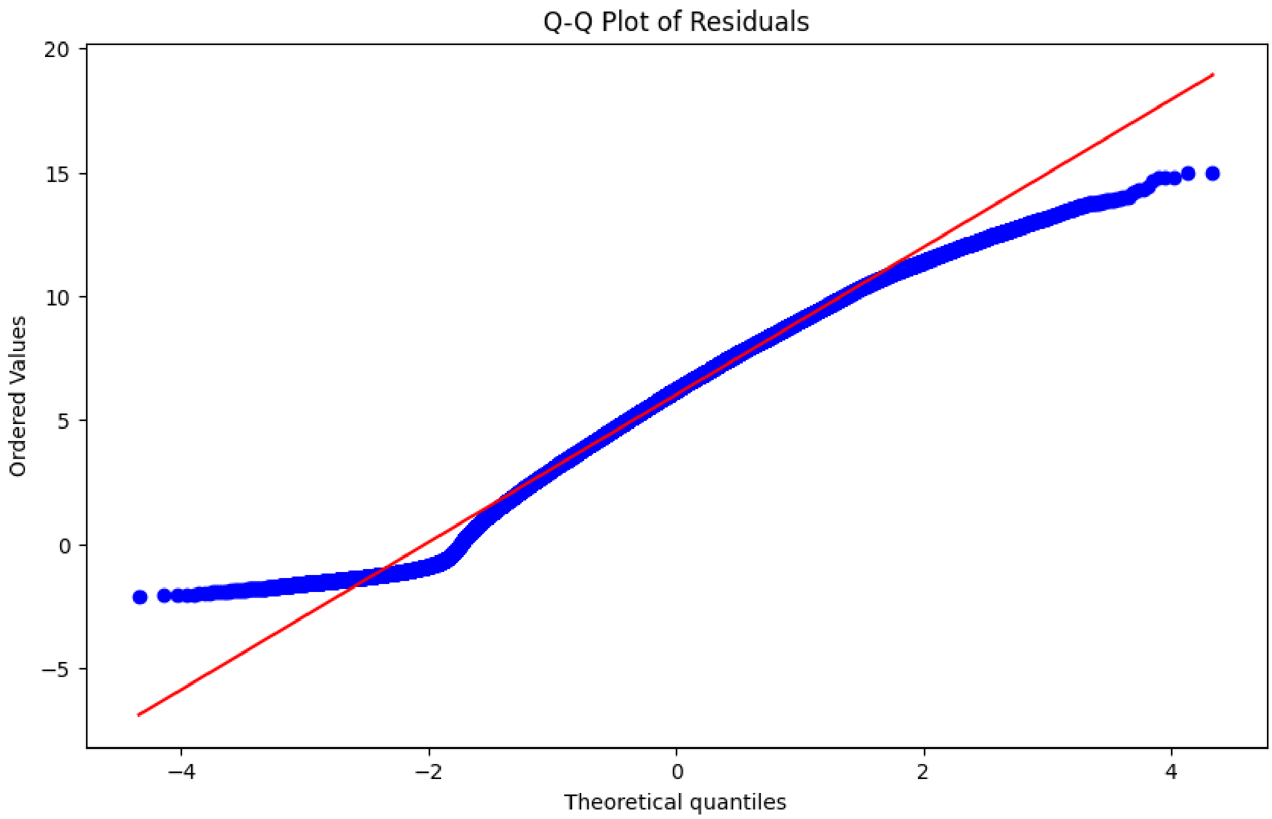
4 figure. *Observed vs. Predicted Plot Log-Normal Distribution*

Scatter plot of observed values against predicted values with a red line denoting a perfect fit (`Observed = Predicted`). Ideally the blue dots in the plot would follow the red line, which assumes x = y. The blue points are heavily scattered and show a pattern of underestimation for larger observed values. This indicates that the model predictions do not fully capture the variability in the observed data, especially for higher values in Figure 4 figure.. This shows that the model is under-performing and while capturing hierarchical structure of the data there are areas to enhance the results.



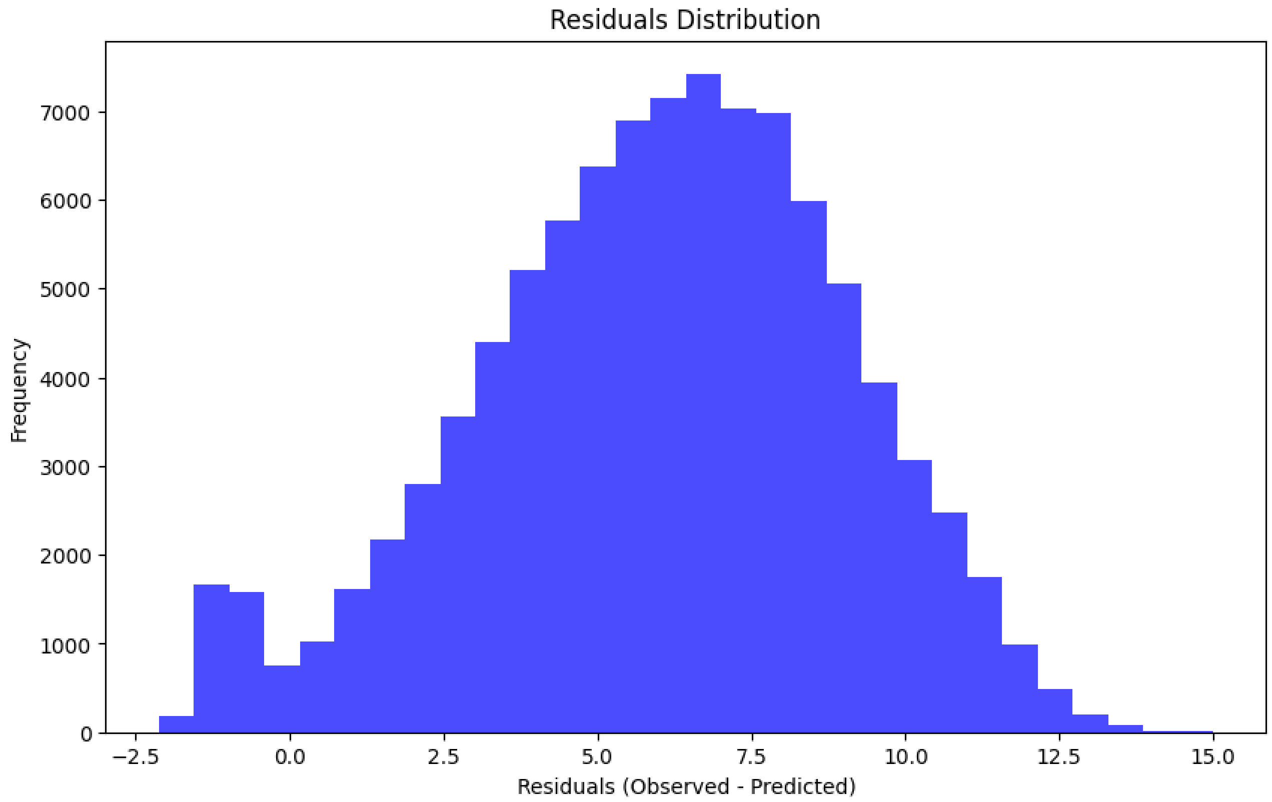
***5 figure.** Parameter Pair Plot Log-Normal Distribution*

Pairwise plots of posterior distributions for key parameters (`mu`, `sigma_within`, `sigma_country`) show the density, marginal distributions, and joint distributions. In Log-Normal distribution mu shows a well-defined central tendency with slight skewness, sigma withing indicates the within-group standard deviation is estimated precisely and country indicates wider spread-out showing uncertainty.



6 figure. *Q-Q Plot of Residuals*

Quantile-Quantile plot compares residuals to a Log-Normal distribution. It shows a good overall alignment in the middle however the tails are still not captured well.



7 figure. *Residuals Distribution*

The distribution of residuals for Log-Normal distribution residuals are skewed more to the positive side and have more variability indicating under-predictiness.

The posterior predictive check plots further explores and analyzes model's capability to capture central tendency and variability in firm size across industries and countries. These plots demonstrates that the model performs and is able to capture the distribution however it fails to predict higher values.

The results of Variance Decomposition and Intraclass Correlation Coefficients in Table 11 table. show that the country-level effects explain the vast majority (89.7%) of the total variance in the model.

Industry effects contribute only 1.5% to the overall variance. This suggests that, while some industries may differ in firm size distribution, the variation within industries is relatively minor compared to country-level effects. Additionally, it shows that model likely would perform better by removing division by industry and using only the total values for the country.

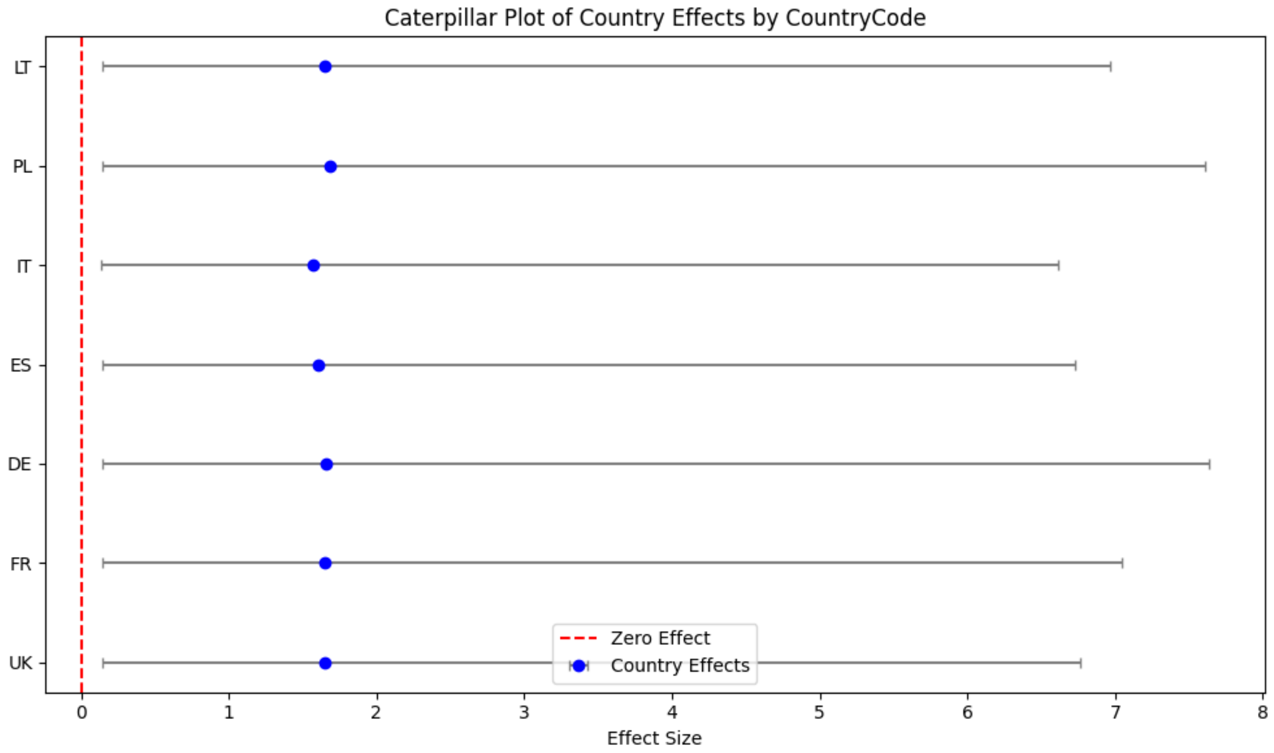
The within-group, residual, variance accounts for 8.8% of the total variance, reflects firm-specific characteristics or noise that cannot be attributed to either country or industry effects.

The caterpillar plot for country effects in Figure 8 figure. shows that country-level effects are homogeneous, with narrow credible intervals across countries suggesting that the differences between countries are stable and predictable in the mode. While country effects dominate the overall variance, they are relatively uniform across countries. It reflects consistent macroeconomic or policy-level impacts that apply similarly to firms across countries and as the analyzed countries are major economies in EU plus Lithuania it reflects that overall employment follows similar distribution in these countries.

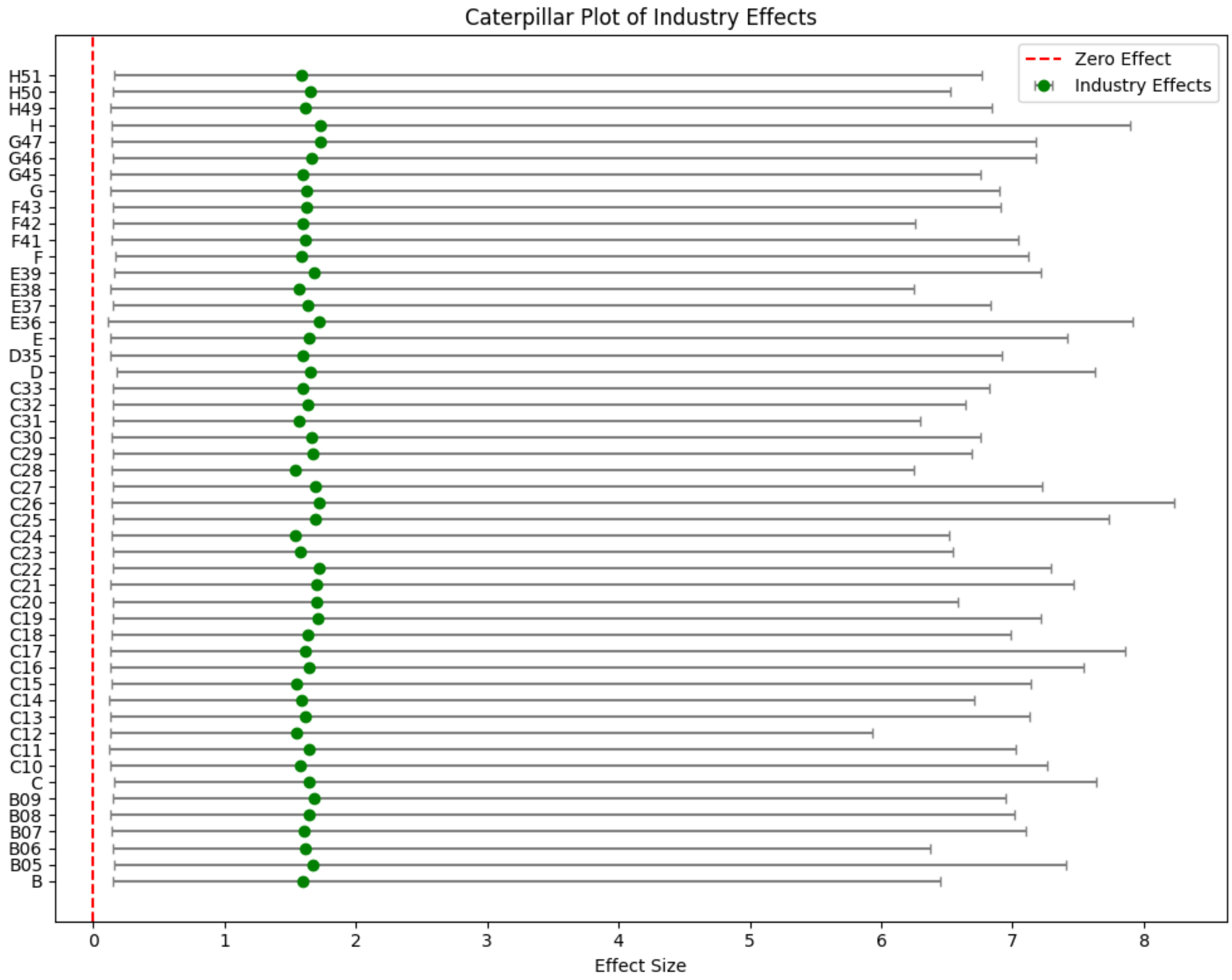
In contrast caterpillar plot for industry effects in Figure 9 figure. shows much greater heterogeneity. There is a wider spread in the credible intervals across industries, indicating variability in the influence of different industries aligning with the smaller industry-level variance (0.188), which contributes less to the overall variability. Industries like H (Transportation and storage), G47 (Retail trade, except of motor vehicles and motorcycles), E36 (Water collection, treatment and supply), C26 (Manufacture of computer, electronic and optical products) , C22 (Manufacture of rubber and plastic products) show a larger industry effect, meaning that in real life these industries could be targeted by policies to even further increase employment. However, it also highlights that industry-specific effects are more dispersed than country-specific effects, implying significant differentiation between industries in their contribution to firm size. It further proves that for better model fit detailed data by industry should be removed.

11 table. *Variance Decomposition and Intraclass Correlation Coefficients (ICC)*

Component	**Variance**	**Proportion of Total Variance (ICC)**
Country	11.323	89.7%
Industry	0.188	1.5%
Within-group	1.107	8.8%



8 figure. *Caterpillar Plot for Country Effects*



9 figure. *Caterpillar Plot for Industry Effects*

3.3 Validation results

Validation combines cross-validation for predictive performance, PPCs for model fit, and diagnostics for convergence and stability. Regularization and hierarchical priors were used to address heterogeneity and over-fitting, while convergence diagnostics ensured robust sampling. Overall results proves model's ability to capture both global and group-specific effects, though challenges remain in modeling data subsets with extreme variability. Future work could explore alternative approaches, such as hierarchical shrinkage priors or grouping sparse industries into broader categories, to further enhance model performance.

The cross-validation results in Table 12 table. reveal variability in model performance across the five folds. Fold 3 and Fold 5 exhibit significantly higher RMSE values, reaching 2.00 and 2.79 million, respectively, compared to the lower RMSE values of Folds 1, 2, and 4. This disparity comes from heterogeneity in data driven by high variability of the data from industry level. The MAE values are more consistent across folds, indicating that while large outliers may inflate RMSE, the average absolute error is relatively stable. Fold 4 exhibits the lowest RMSE and MAE, suggesting better predictive performance on this subset of data.

The inconsistency in folds performance and 3rd and 5th folds poor performance is likely caused by observations from countries or more likely industries with extreme values or lack of sufficient representation of certain size classes, leading to under- or overestimation in predictions. The hierarchical Bayesian model, being a complex model, may over-fit to the training data in some folds, especially when the data in those folds are small or do not provide enough diversity to constrain the model parameters. [4]

12 table. *Cross-Validation Results for RMSE and MAE*

Fold	RMSE ($\times 10^7$)	MAE ($\times 10^5$)
1	1.11	8.06
2	0.94	7.06
3	2.00	8.02
4	0.84	9.22
5	2.79	8.54

The global mean (μ) and within-group variability (σ_{within}) are well-estimated, with low R-hat values (close to 1.00) indicating good convergence across chains as shown in Table 13 table.. Random effects for countries (u_{country}) and industries (v_{industry}) show variability as some country-level effects have broader credible intervals, $u_{\text{country}[1]}$, reflecting uncertainty in their estimates. For industries, there are clear deviations, such as $v_{\text{industry}[2]}$, with a highly negative mean, suggesting that some industries substantially deviate from the overall mean which is expected looking at the nature of data. Significant variability in industry-level random effects (v_{industry}) aligns with the observed heterogeneity in cross-validation RMSE. Industries with limited data likely contribute to the high uncertainty. Tight credible interval for σ_{within} ([4.10, 4.13]) indicates precise estimation of within-group noise, providing confidence in the model's ability to partition variability across hierarchical levels.

13 table. *Posterior Summaries of Key Parameters*

Parameter	Mean	Std. Dev.	94% HDI	R-hat
μ	4.24	1.52	[1.48, 6.43]	1.01
σ_{within}	4.12	0.01	[4.10, 4.13]	1.00
$u_{\text{country}[1]}$	1.90	0.81	[0.72, 3.49]	1.04
$u_{\text{country}[2]}$	1.22	0.81	[0.05, 2.81]	1.04
$v_{\text{industry}[1]}$	1.30	0.47	[0.45, 2.10]	1.01
$v_{\text{industry}[2]}$	-20.0	0.10	[-20.2, -19.8]	1.00

The density plots for the posterior distributions shows that the model is capturing the structure of the data well, particularly for the global parameters. Many of the random effects appear centered around 0, which aligns well with the priors typically used for hierarchical models, mean-zero normal priors for random effects. The plots show that the industry - level creates the highest uncertainty, an insight that was already created by other checks and confirmed here further.

4 Alternative methods

4.1 BART model

For additional comparison, Bayesian Additive Regression Trees model was applied. BART model is non - parametric model that provides flexibility by using sum-of-trees model withing Bayesian framework using prior specification to control for uncertainty. [3]

The Bayesian Additive Regression Trees (BART) model was configured and evaluated with the following key specifications and metrics. The table below summarizes the essential details, while the discussion highlights their implications and relevance to the analysis.

14 table. *BART Model Key Metrics*

Metric	**Value**
Training Data Size (n, p)	$n = 12{,}190$, $p = 2$ predictors
Number of Trees	50
Burn-in/Post-samples	100 / 500
Variance of y (Prior)	$23{,}747{,}771{,}427.65$
Posterior Variance (Avg.)	$1{,}800{,}046{,}872.94$
L_1 (Mean Absolute Error)	$165{,}480{,}410.85$
RMSE (Root Mean Squared Error)	$38{,}865.73$
Pseudo-R^2	0.994

The training dataset consists of $n = 12{,}190$ observations and $p = 2$ predictors (`ENT` and `TRN`). This dataset size and dimensionality ensure sufficient variability for model estimation as Bayesian model tend to perform well with smaller datasets. As BART is tree-ensemble model, in this model an ensemble of 50 trees with 100 bur-in iterations and 500 post-samples is used. The variance of the response variable (`EMP`) prior to model fitting (`Sigsq`) was $23{,}747{,}771{,}427.65$, which significantly reduced to an average posterior variance of $1{,}800{,}046{,}872.94$ after burn-in. This reduction indicates that the model explains a substantial portion of the variability in the response variable. The Mean Absolute Error (L_1) of $165{,}480{,}410.85$ and the RMSE of $38{,}865.73$ indicate that predictions on average are close to the observed values. The high L_2 (Sum of Squared Errors) emphasizes deviations from the mean, particularly for outliers. The Pseudo-R^2 value of 0.994 suggests an excellent in-sample fit, explaining 99.4% of the variance in the data. While this reflects robust performance, further checks are required for out-of-sample generalization to avoid overfitting. The Shapiro-Wilk test for residual normality yielded $p = 0$, indicating significant deviations from normality which is not inconsistent with results in complex models like BART. The zero-mean noise test resulted in $p = 0.95501$, confirming that residuals are centered around zero.

The Q-Q (Quantile-Quantile) plot of residuals in Figure 11 figure. visually compares the distribution of the residuals from the model to a theoretical normal distribution. The extreme points in the lower left of the plot deviate substantially below the line, suggesting extreme negative residuals and large underpredictions, indicating data following heavy - tailed distribution. Extreme positive residuals indicate model overfitting.

BART model's configuration and performance metrics shows its capability to capture the variability in firm sizes with substantial reduction in variance after accounting for predictors (`ENT` and `TRN`) to highlight model's explanatory power. Despite deviations from normality in residuals, the results suggest the ML Bayesian framework is suitable for analyzing firm size distributions across multiple levels however hierarchical model proves to have better results.

4.2 Bayesian Neural Network

The Bayesian Neural Network incorporates both group-level effects (Industry and Size_Class) and individual predictors (ENT and TRN) to analyze employment levels (EMP). Table 15 table. summarizes the key parameters and their estimates.

15 table. *Bayesian Neural Network Model Key Parameters*

Parameter	**Estimate**	**95% Credible Interval**
Industry Intercept Variance	1.03	(0.03, 3.86)
Size_Class Intercept Variance	1.00	(0.03, 3.59)
Intercept	0.18	(-20.29, 20.16)
ENT Coefficient	2.31	(-15.85, 21.45)
TRN Coefficient	2.39	(-16.88, 21.60)
Residual Std. Dev. (σ)	559,141.75	(552,988.08, 565,427.33)

The estimated Industry variance is 1.03, with a credible interval of (0.03, 3.86). This indicates small-to-moderate variability in EMP across industries, with some industries potentially having notable effects. The variance across size classes is similar, estimated at 1.00 with a credible interval of (0.03, 3.59). This aligns with the idea that company size impacts employment levels.

Intercept estimate (0.18) indicates a minimal baseline effect when all predictors are at their mean values. However, the wide credible interval (-20.29, 20.16) reflects high uncertainty. The coefficient for ENT (2.31) suggests a positive relationship with EMP, but the wide credible interval (-15.85, 21.45) indicates substantial uncertainty. This shows that further analysis should be conducted to determine if the larger data set should be used or model creates weak associations. Hoverer this shows that BNN creates higher uncertainty than expected and is not a good choice for estimation.

Overall, Bayesian Neural Network and BART model both performed worse than Hierarchical Bayesian model BNN performing worse than BART, indicating that for estimating firm size by number of employees ML model show poorer explain-ability and fails to capture intra-industry and hidden architectures within the data.

5 Discussion

Application of hierarchical Bayesian models to estimate the distribution of firm sizes by employees has revealed significant insights into the dynamics of economic activities across industries and regions. First, analysis of raw data showed that firm size distributions were found to vary significantly between sectors, with services exhibiting a greater concentration of micro and small enterprises, while manufacturing and mining displayed a more balanced spread, including medium and large firms. These findings align with economic intuition, as certain sectors, such as manufacturing, often require substantial capital investment and workforce, leading to larger firm sizes. Further looking at he model results, the multilevel structure of the Bayesian model revealed significant variability in employment across industries and size classes. For example, the random effects for Industry and Size_Class showed that the variability within these groups was larger than initially anticipated.

The posterior predictive checks demonstrated that model manages to capture the distribution however it fails to capture high values of EMP and theses values are not explained well by the remaining variables. Model assumes a strong Log-Normal distribution while the observed data shows more positively skewed distribution. Hierarchical model estimates both country and industry effect showing a superior estimation compared to the simple linear model. The model proved effective in capturing this distribution, providing interpretable estimates of variability across different levels.

Further decomposed of total variance in firm size into country-level, industry-level, and within-group components, highlighted the hierarchical nature of the data. Country-level effects dominated the total variance, accounting for 89.7% of the variability. This result underscores the influence of national-level factors, such as economic policies, labor market conditions, and cultural contexts, on firm size. The caterpillar plot for country effects revealed relatively homogeneous estimates, with most countries clustered closely around their mean effects. Industry effects contributed 1.5% of the total variance. While modest compared to country-level effects, this result highlights the heterogeneity across industries and from caterpillar plots of industry effects it's clear that industry level effect shows some industries have robust growth opportunities in certain sectors or industries that could be targeted to address lower effects.

In comparison, ML model performed, the BNNs and BART incorporated multilevel random effects for Industry and Size_Class alongside key predictors (ENT and TRN). The results revealed that regression coefficients for ENT and TRN were positive but had wide credible intervals, suggesting high uncertainty in their contribution to predicting EMP. The multilevel hyperparameters showed considerable variability across industries, emphasizing the importance of including random effects in the model. The posterior distribution of residuals highlighted non-normality, which was further validated by the Q-Q plots. It underscores the importance of incorporating flexible Bayesian frameworks for such data. Additionally while BART showed high pseudo R value, Q-Q plot evaluated over-fitting. Computational challenges could be solves using cloud based solutions or stronger computers. With additional tuning and data mining techniques the results of ML models would improve however for better comparability in this thesis the same dataset was used.

Overall, initial analysis showed high variability in data and that the data distribution varies across size classes. Further on hierarchical Bayesian model did not manage to fully capture Log - Normal distribution as the modelled values fitted poorly. Model managed to capture all three level effects. Country effect proved to explain most of the overall variance and had a positive homogeneous effect. Industry effect was small and caterpillar plots showed more heterogeneous effect providing insight that industry variability was likely a problematic part of the model. Within group effect has a small ICC value showing that most of the model explained firm data. ML models, while performing worse than the hierarchical model, has better results than linear regression, proving that these methods, for nested firms data are more superior than traditional linear models.

6 Conclusions and Recommendations

6.1 Conclusions

The choice of Bayesian methods and machine learning models reflects the need for more complex estimations and predictive accuracy in comparison with linear models. In literature review no papers were find to apply Bayesian Hierarchical models to estimate employees data in size classes proving the need for this analysis. While Bayesian models provided insights into the structure and variability of the data, the computational demands were high particularly for large datasets.Machine Learning models, BART and BNN were chosen for their incorporation of Bayesiand interface for comparability with Bayesian Hierarchical model providing predictive power but at the cost of interpretability. Thesis investigated the use of Bayesian hierarchical models and machine learning techniques, including Bayesian Neural Networks (BNN) and Bayesian Additive Regression Trees (BART), to analyze complex company hierarchical data. The goal was to estimate distribution of firm size by employees and find group effects in the model and compare model results with ML models. The hierarchical Bayesian model outperformed traditional linear models in capturing the nested structure of the data and explaining variability across levels. While the linear model provided a simplistic view, it failed to account for group-level dependencies and produced residual patterns indicative of model mis-specification.

Hierarchical Bayesian model provided robust insights into the variability of employment across different groups. By using random effects for industries and size classes, the models captured nuanced relationships in the data. Hierarchical Bayesian approach showed contributions of country-level, industry-level, and within-group variability. The results showed that country-specific effects account for the largest share of the total variance in firm size, highlighting the role of national-level factors such as policies, market conditions, and economic environments. Industry-level effects, while less pronounced, revealed meaningful heterogeneity across sectors, suggesting the need for tailored policy interventions. The model did not fit data well and high heterogeneity, variability in Employment created by the industry were one of the main reasons for fitting difference.

The nature of the data proved to challenge the analysis and Bayesian approach leading to need of further analysis on outliers that are present in the data and debate if the industries showing different distributions or having 0 values should be excluded. While the Log-Normal approach was effective, it introduced additional uncertainty, particularly in groups with sparse data. Future work could compare the robustness of results using alternative imputation methods, such as Bayesian imputation or multiple imputation, to validate the finding. Numerical instability, caused by high variability in Employment variable along with high variance introduced by incorporating industry level data proved to be difficult for model to capture and fit distribution properly. This shows that more data modeling and different assumptions should be employed in further testing and papers to enrich academical background for estimating firm data. Based on these findings, decisions on small and medium business strategies could be made and industries differentiated to identify potential improvements.

In conclusion, this thesis illustrates the use of hierarchical Bayesian models in analyzing the distribution of firm size by employees. By providing a framework for Hierarchical Bayesian model to estimate distribution and variance decomposition for group-specific effect estimation, the research provides insights that these methods have potential and prove to be superior to linear modeling. The findings emphasize the importance of country-level policies, industry-specific strategies, and firm-level initiatives in fostering business growth. This thesis demonstrates the potential of hierarchical Bayesian models for analyzing complex, multi-level data not only for socio - economic data but also for other domains. The results of industry-effect shows which industries are more influential for employment increase and policy makers should target these industries more. It includes Transportation and storage industries and Manufacturing of computer, electronics and optical products. In other domains a deeper analysis and practical application of hierarchical Bayesian models could be beneficial, for healthcare this approach could be applied to modeling patient outcomes across hospitals, regions, and demographics. In education, student performance could be analyzed across school, regions, and socio-economic factors.

6.2 Recommendations

Future studies should explore the inclusion of macroeconomic and institutional variables to enhance the predictive power of models. For example, incorporating GDP growth, labor market regulations, or international trade variables may provide additional insights into employment variability. Combining the interpretability of hierarchical Bayesian models with the predictive power of machine learning techniques like BART could yield a more comprehensive understanding of economic phenomena. Developing hybrid Bayesian-tree-based models is a promising area for further exploration. Bayesian methods, particularly those requiring large datasets, are computationally intensive. Exploring approximate Bayesian computation (ABC) or variational inference techniques could improve scalability without sacrificing interpretability.

A better application of this model could also be removing different industries and looking only at the total level of country as different industries do show different distributions and create high variability within data further making the model to underfit values. The model assumes independence between country- and industry-level effects. Future work could explore interaction terms to better capture the interplay between national and sectoral factors.

While the log-normal distribution was appropriate for capturing skewness in firm size, future research could explore alternative distributions that may better handle extreme outliers or heavier tails. This could include Student-t, Gamma or Weibull distributions that would be suitable for modeling positive, skewed data like firm sizes. Usage of large-scale hierarchical datasets, exploring parallel computing or cloud-based solutions could reduce computational time and allow the use of larger dataset and higher number of iterations.

Additionally, extending the static hierarchical model to dynamic Bayesian models that account for temporal variation in firm sizes by using State-space models or time-varying random effects could capture changes in country or industry effects due to evolving policies or economic conditions. This would work on analyzing one country that had certain policies applied to model if the results indeed can be captured.

References

[1] Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1), 111–142.

[2] Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2), 5–43.

[3] Axtell, R. L. (2001). Zipf distribution of US firm sizes. *Science*, 293(5536), 1818–1820.

[4] Cabral, L., & Mata, J. (2003). On the evolution of the firm size distribution: Facts and theory. *American Economic Review*, 93(4), 1075–1090.

[5] Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4), 327–335.

[6] Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.

[7] Diem, C., Eppinger, P. S., Harting, P., Henkel, L., Kühnlenz, F. (2023). Aggregation-Induced Biases in Economic Input-Output Networks. *arXiv preprint arXiv:2302.11451*.

[8] Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66(2), 271–288.

[9] Hill, J. L., Linero, A. R., & Murray, J. S. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7, 251–278.

[10] Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle. *American Economic Review*, 86(3), 562–583.

[11] Luttmer, E. G. (2007). Selection, growth, and the size distribution of firms. *Quarterly Journal of Economics*, 122(3), 1103–1144.

[12] McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Wiley.

[13] Mullachery, V., Khera, A., & Husain, A. (2018). *Bayesian Neural Networks*. Wiley.

[14] Ridout, M., Demétrio, C. G. B., & Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*, 19, 179–192.

[15] Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3), 507-554.

[16] Sutton, J. (1997). Gibrat's legacy. *Journal of Economic Literature*, 35(1), 40–59.

[Watanabe(2010)] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross-validation and WAIC. *Journal of Machine Learning Research*, 11, 3571–3594.

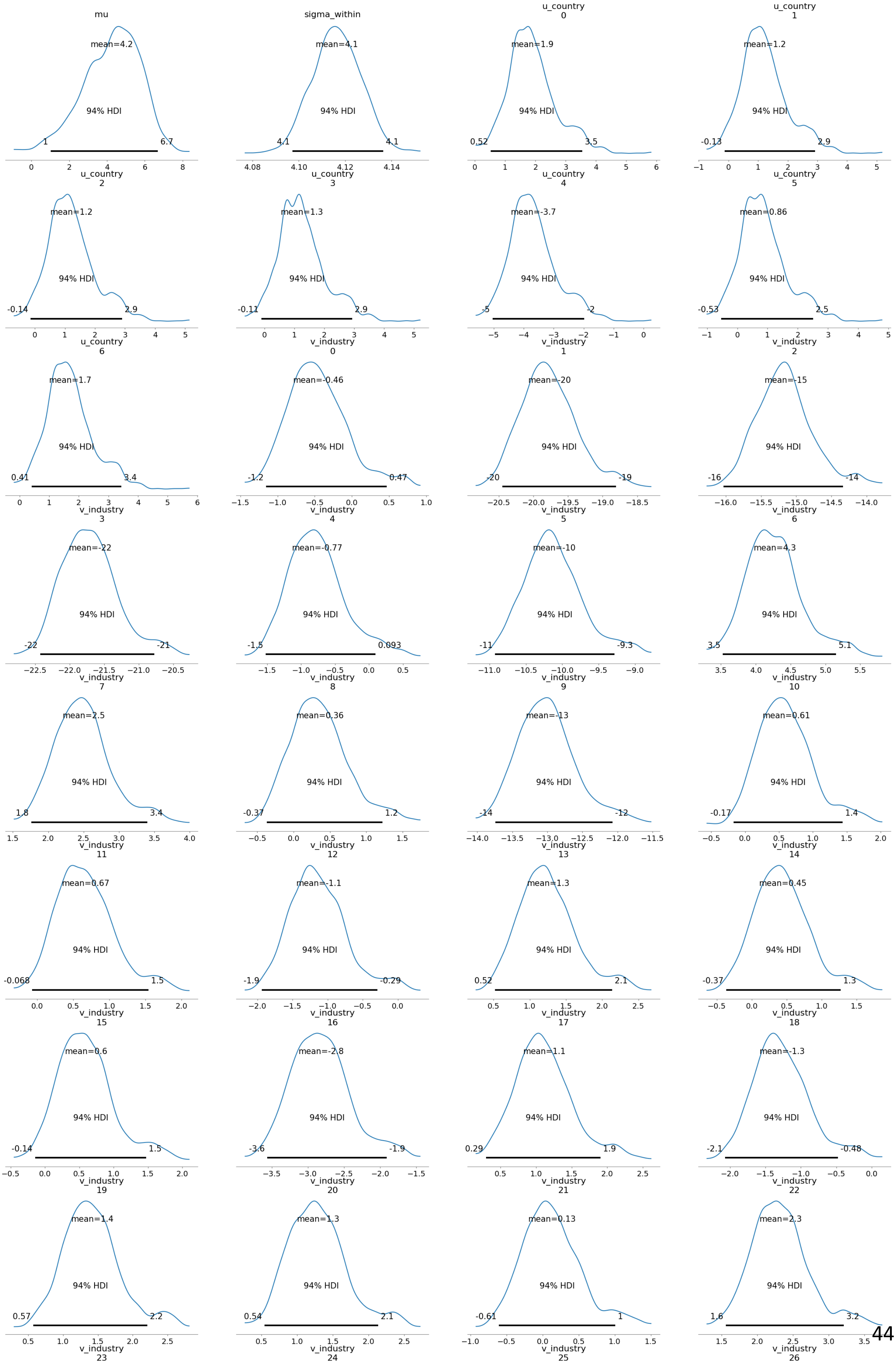
[17] Yao, Y., Vehtari, A. (2018). Make cross-validation Bayes again.

Appendix 1.

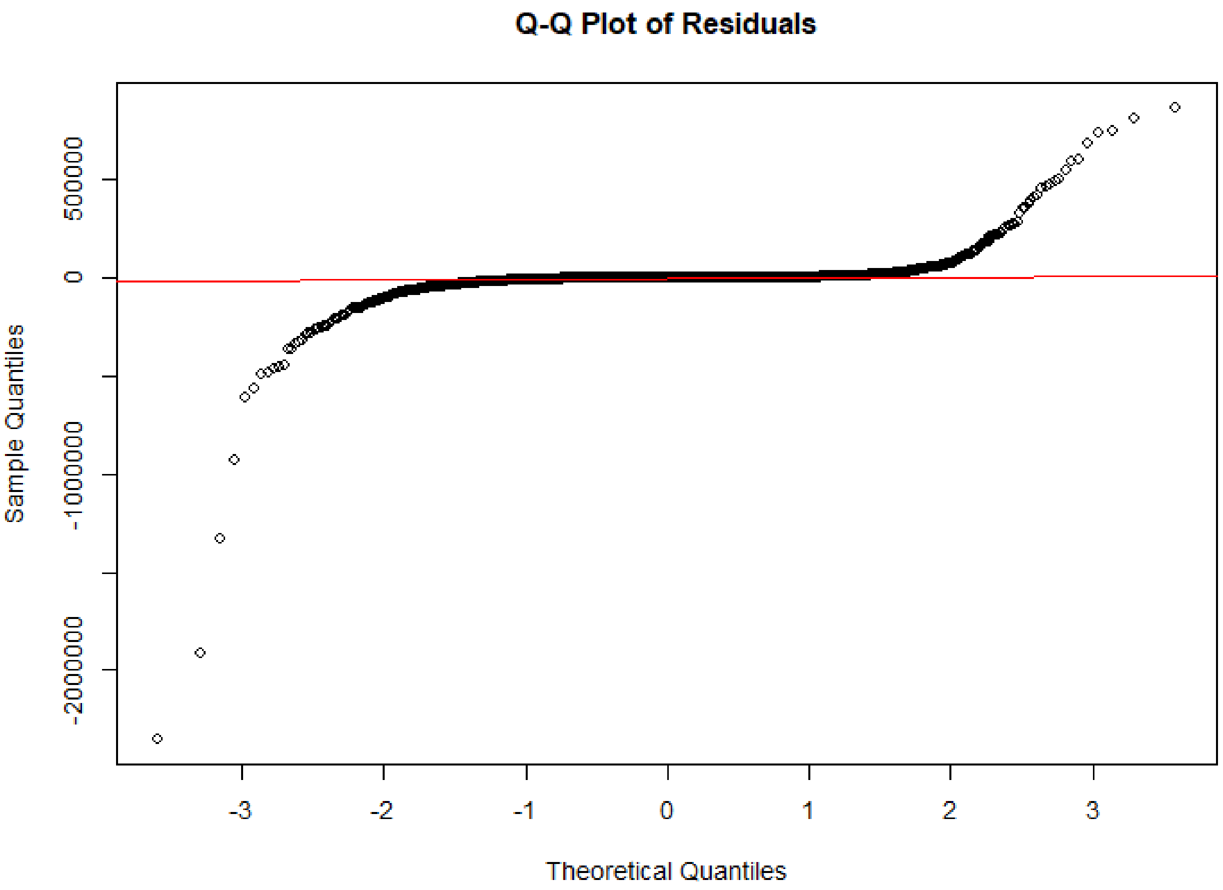
In the preparation of this thesis generative AI tools were employed as part of the research and writing process. The following points outline the scope and nature of their usage:

Generative AI models, specifically ChatGPT by OpenAI, version 2025, knowledge cutoff of January 2025, were utilized to assist in drafting and refining the text of the thesis, provide statistical and methodological insights, generate code examples for data analysis and visualization and review and rewrite and fix text for academical style.

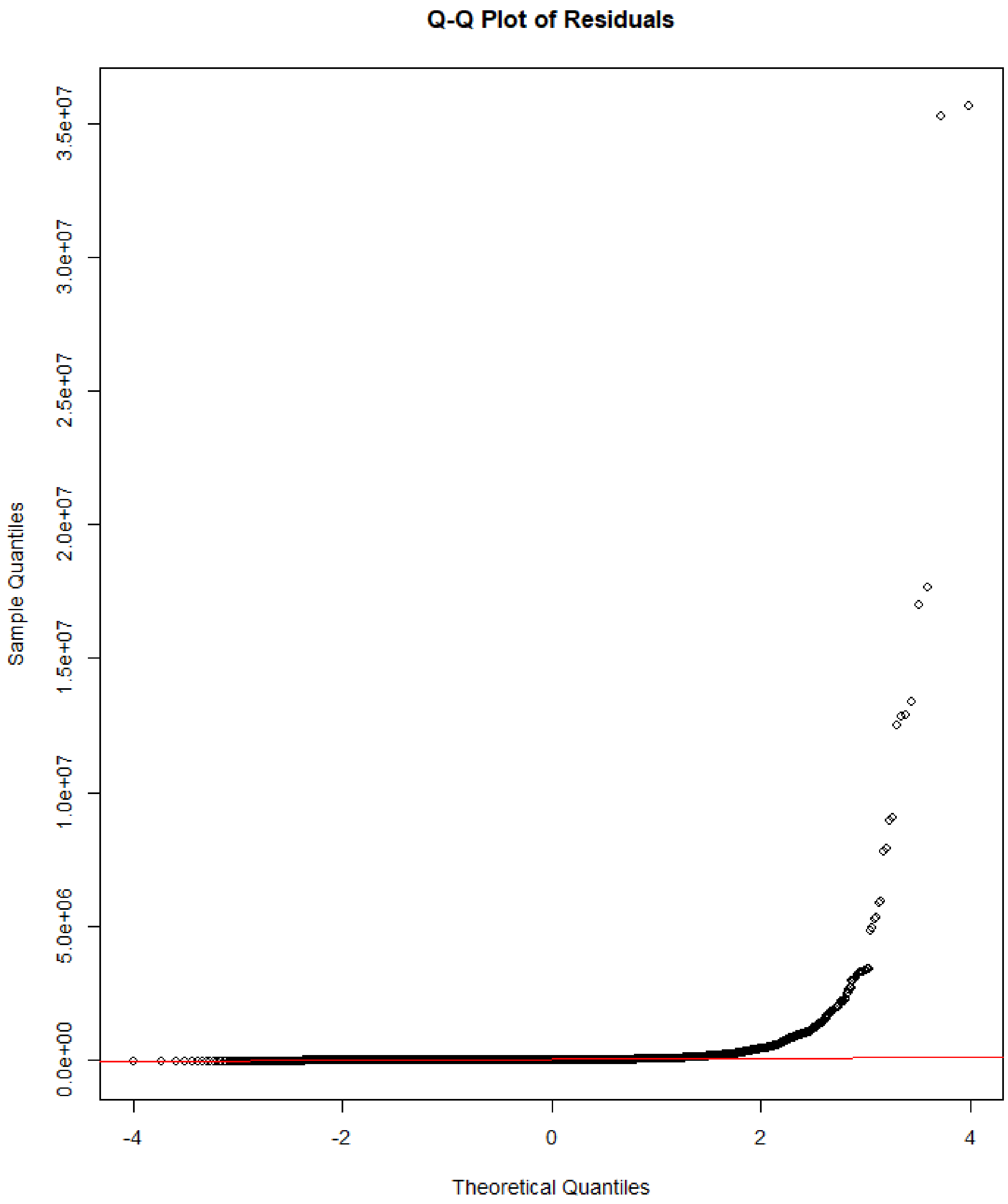
Appendix 2.



mu
mean=4.2
94% HDI
1
6.7
0 2 4 6 8
sigma_within
mean=4.1
94% HDI
4.1
4.1
4.08 4.10 4.12 4.14
u_country 0
mean=1.9
94% HDI
0.52
3.5
0 1 2 3 4 5 6
u_country 1
mean=1.2
94% HDI
-0.13
2.9
−1 0 1 2 3 4 5
u_country 2
mean=1.2
94% HDI
-0.14
2.9
0 1 2 3 4 5
u_country 3
mean=1.3
94% HDI
-0.11
2.9
0 1 2 3 4 5
u_country 4
mean=-3.7
94% HDI
-5
-2
−5 −4 −3 −2 −1 0
u_country 5
mean=0.86
94% HDI
-0.53
2.5
−1 0 1 2 3 4 5
u_country 6
mean=1.7
94% HDI
0.41
3.4
0 1 2 3 4 5 6
v_industry 0
mean=-0.46
94% HDI
-1.2
0.47
−1.5 −1.0 −0.5 0.0 0.5 1.0
v_industry 1
mean=-20
94% HDI
-20
-19
−20.5 −20.0 −19.5 −19.0 −18.5
v_industry 2
mean=-15
94% HDI
-16
-14
−16.0 −15.5 −15.0 −14.5 −14.0
v_industry 3
mean=-22
94% HDI
-22
-21
−22.5 −22.0 −21.5 −21.0 −20.5
v_industry 4
mean=-0.77
94% HDI
-1.5
0.093
−1.5 −1.0 −0.5 0.0 0.5
v_industry 5
mean=-10
94% HDI
-11
-9.3
−11.0 −10.5 −10.0 −9.5 −9.0
v_industry 6
mean=4.3
94% HDI
3.5
5.1
3.5 4.0 4.5 5.0 5.5
v_industry 7
mean=2.5
94% HDI
1.8
3.4
1.5 2.0 2.5 3.0 3.5 4.0
v_industry 8
mean=0.36
94% HDI
-0.37
1.2
−0.5 0.0 0.5 1.0 1.5
v_industry 9
mean=-13
94% HDI
-14
-12
−14.0 −13.5 −13.0 −12.5 −12.0 −11.5
v_industry 10
mean=0.61
94% HDI
-0.17
1.4
−0.5 0.0 0.5 1.0 1.5 2.0
v_industry 11
mean=0.67
94% HDI
-0.068
1.5
0.0 0.5 1.0 1.5 2.0
v_industry 12
mean=-1.1
94% HDI
-1.9
-0.29
−2.0 −1.5 −1.0 −0.5 0.0
v_industry 13
mean=1.3
94% HDI
0.52
2.1
0.5 1.0 1.5 2.0 2.5
v_industry 14
mean=0.45
94% HDI
-0.37
1.3
−0.5 0.0 0.5 1.0 1.5
v_industry 15
mean=0.6
94% HDI
-0.14
1.5
−0.5 0.0 0.5 1.0 1.5 2.0
v_industry 16
mean=-2.8
94% HDI
-3.6
-1.9
−3.5 −3.0 −2.5 −2.0 −1.5
v_industry 17
mean=1.1
94% HDI
0.29
1.9
0.5 1.0 1.5 2.0 2.5
v_industry 18
mean=-1.3
94% HDI
-2.1
-0.48
−2.0 −1.5 −1.0 −0.5 0.0
v_industry 19
mean=1.4
94% HDI
0.57
2.2
0.5 1.0 1.5 2.0 2.5
v_industry 20
mean=1.3
94% HDI
0.54
2.1
0.5 1.0 1.5 2.0 2.5
v_industry 21
mean=0.13
94% HDI
-0.61
1
−1.0 −0.5 0.0 0.5 1.0 1.5
v_industry 22
mean=2.3
94% HDI
1.6
3.2
1.5 2.0 2.5 3.0 3.5
v_industry 23
v_industry 24
v_industry 25
v_industry 26



11 figure. *Q-Q Plot of Residuals of BART*



12 figure. *Q-Q Plot of Residuals of BNN*

Appendix 3.

Link to the code used for this analysis repository on GitHub: `https://github.com/Kaiiiaa/Hierarchical-Bayes-models.git`.

```
stan_data = {
    'N': len(df_hierarchical),
    'J': df_hierarchical['country'].nunique(),
    'K': df_hierarchical['industry'].nunique(),
    'L': df_hierarchical['indicator'].nunique(),
    'country': df_hierarchical['country'].values,
    'industry': df_hierarchical['industry'].values,
    'indicator': df_hierarchical['indicator'].values,
    'log_size': np.log(df_hierarchical['values'].values)
}

stan_model_code = """
data {
  int<lower=0> N;                      // Number of observations
  int<lower=1> J;                      // Number of countries
  int<lower=1> K;                      // Number of industries
  int<lower=1> L;                      // Number of indicators
  array[N] int<lower=1> country;       // Country index for each
     observation
  array[N] int<lower=1> industry;      // Industry index for each
     observation
  array[N] int<lower=1> indicator;     // Indicator index for each
     observation
  vector[N] log_size;                  // Log-transformed firm size
}
parameters {
  real mu;                              // Overall mean of log size
  real<lower=0> sigma_within;           // Standard deviation within
     industry-country
  vector[J] u_country;                  // Random effects for country
  vector[K] v_industry;                 // Random effects for industry
  vector[L] w_indicator;                // Random effects for indicator
  real<lower=0> sigma_country;          // Standard deviation for country
     effects
  real<lower=0> sigma_industry;         // Standard deviation for industry
     effects
  real<lower=0> sigma_indicator;        // Standard deviation for indicator
      effects
}
model {
  // Regularization Priors
  mu ~ normal(0, 2);                        // Weakly informative prior for
      log-scale global mean
  sigma_within ~ gamma(2, 0.5);           // Half-normal prior for within
     -group SD (log scale)
  sigma_country ~ exponential(1);         // Regularization prior for
     country-level SD
  sigma_industry ~ exponential(1);        // Regularization prior for
     industry-level SD
  sigma_indicator ~ exponential(1);       // Regularization prior for
     indicator-level SD

  // Priors on Random Effects
  u_country ~ normal(0, sigma_country);    // Country-level variation
  v_industry ~ normal(0, sigma_industry);  // Industry-level variation
  w_indicator ~ normal(0, sigma_indicator);  // Indicator-level variation

  // Likelihood
  for (n in 1:N) {
    log_size[n] ~ lognormal(mu + u_country[country[n]] + v_industry[
       industry[n]] + w_indicator[indicator[n]], sigma_within);
  }
}
generated quantities {
  vector[N] log_size_rep;  // Predicted log size for posterior predictive
      checks
  for (n in 1:N) {
    log_size_rep[n] = lognormal_rng(mu + u_country[country[n]] +
       v_industry[industry[n]] + w_indicator[indicator[n]], sigma_within)
       ;
  }
}
"""

def init_values():
    return {
        "mu": np.mean(np.log(df_hierarchical['values'].values)),
        "sigma_within": 1.0,
        "sigma_country": 1.0,
        "sigma_industry": 1.0,
        "sigma_indicator": 1.0
    }
for i in range(4):
    init_file_path = f'C:/init_{i+1}.json'
    with open(init_file_path, 'w') as f:
        json.dump(init_values(), f)
```

```
init_files = [f'C:/init_{i+1}.json' for i in range(4)]

fit = model.sample(
    data=stan_data,
    chains=2,
    parallel_chains=2,
    iter_warmup=500,
    iter_sampling=2000,
    seed=1234,
    inits=[f'init_{i+1}.json' for i in range(4)],
    adapt_delta=0.99,
    max_treedepth=15,
    show_console=True
)
```