Ludwig-Maximilians-Universität München

Department of Statistics

Master Thesis

with the title

Estimation of Voter Transitions in Multi-Party Systems

- Quality of Credible Intervals in (hybrid) Multinomial-Dirichlet Models -

---

Author: Jan Moritz Klein

Supervisors: Prof. Dr. Helmut Küchenhoff (LMU)
Dr. André Klima (LMU)
Prof. Myriam Vimond (ENSAI)

Date: 08.04.2019

Acknowledgments

At this point, I would like to thank all those who contributed to the success of this Master's thesis with their professional and personal support.

Firstly, I want to thank Dr. André Klima for his helpful and instructive support as well as his uncomplicated aid over the telephone which made it possible for me to write this Master's thesis in the course of my Erasmus semester in France.

I thank Prof. Dr. Helmut Küchenhoff for the acceptance of the topic and the opportunity to write this Master's thesis in France.

I want to thank Prof. Myriam Vimond for being my supervisor at ENSAI.

Furthermore, I would like to thank all of those who read my work and corrected it, thus helping me to put the finishing touches on it.

Last but not least, I thank my family and friends who supported me personally and helped me a lot over the course of writing this thesis.

Abstract

Voter transitions in multi-party systems can be estimated with the help of the (hybrid) Multinomial-Dirichlet models. Both models are based on Bayesian inference. This thesis aims to investigate the quality of the resulting credible intervals. The origin of this objective lies in the current state of research around this topic as previous research has shown that the resulting credible intervals have a low quality. In these contributions simulation studies were conducted, which simulated voter transitions under different assumptions. As the "true" voter migrations are known in a simulation study, the coverage of the true values by the credible intervals can be evaluated. In some cases, this coverage is well below the target credible level. Due to the complexity of the previous simulation studies, however, it has been difficult to evaluate the reasons for the low quality of the credible intervals.

In the course of this thesis, the quality of credible intervals will be evaluated within the framework of a separate simulation study. The underlying hypothesis of the simulation study is the assumption that the construction of valid credible intervals works in simple data scenarios and that only more complex data scenarios lead to poor quality of the credible intervals. Assuredly, this hypothesis was confirmed in the simulation study. Also, it was shown that above all the presence of large differences in the transition probabilities leads to a deterioration of the quality of the credible intervals. Reduced interval widths, an increased estimation error and assumptions in the simulation were identified as reasons for this.

A further objective of this thesis was the formulation and evaluation of possible correction approaches for the construction of valid credible intervals. Two concrete and one theoretical solution were formulated. The evaluation has shown that the resulting corrected credible intervals pose a substantial improvement compared to the initial ones. However, it has also become clear that there is still a need for further research in this area.

In a concluding application example, the voter migrations between the Bavarian state elections in 2013 and 2018 are calculated with the help of the hybrid Multinomial-Dirichlet model. Subsequently, corrected credible intervals are calculated according to the proposed correction approaches.

Contents

List of Figures

List of Tables

List of Abbreviations

AfD Alternative for Germany

AD Average Distance

CDU Christian Democratic Union

CSU Christian Social Union

EI Ecological Inference

FDP Free Democratic Party

FV Free Voters

HMDM Hybrid Multinomial-Dirichlet-Model

HPD Highest Porsterior Density

MCMC Markov Chain Monte Carlo

SPD Social Democratic Party

USBW University Study Bavarian Elections 18

1. Introduction

Electoral research plays an integral part in modern politics, and its results have long since been a factor in the formation of voters' opinions and the orientation of election campaigns by political players. For each election, polls are used to conduct forecasts in order to predict the outcome of the election as accurately as possible. In addition, the analysis of the election results post-election has also gained vital importance. It is without a doubt then that voter transition analyses play a particularly important and prevalent role. Through this analysis, voter transitions between two elections, i.e. which voters have transitioned where, is studied. In times of drastically changing party landscapes and decreasing voter loyalty, this can be an essential tool to explain the dynamics of a political system.

Currently, there are two different approaches to estimating voter transitions. On the one hand, aggregated election results can be used to infer individual voting behavior and on the other, voter transitions can be extrapolated from individual data. This thesis concentrates on the former, in which the estimation of voter transitions is based on aggregate data and is generally referred to as Ecological Inference. It is also applied in other fields such as sociology or medicine in addition to political science. The methods suitable for the estimation can be divided into two model classes. The first being the Ecological Inference model, in which the estimation of voter transitions is based solely on aggregate data in the form of official election results and the second being the relatively new so-called hybrid models, which combine aggregate data with individual data in the estimation. The individual data is collected separately from the election results and can, for example, come from a telephone survey or exit poll.

Previous evaluations of these two model classes have shown that both models are capable of estimating voter transitions without major estimation errors. Besides the accurate estimation of voter migration, a correct quantification of the uncertainty of the estimation is crucial to evaluate the accuracy of the estimation. Here, however, both model classes reveal their weaknesses. The previous evaluations have shown that the interval estimates in this case credible intervals, which indicate the uncertainty of the estimation, have a low quality. This aspect will be further investigated in the course of this Master's thesis. Therefore, two objectives are formulated. On the one hand, the reasons for the low quality of the credible intervals are to be investigated and analyzed. Additionally, possible solutions for the construction of valid credible will be presented and also evaluated.

The thesis can be divided into three sections. The first section consists of chapters two and three. In these chapters the theoretical basics necessary for the master thesis are discussed. Chapter 2 deals with the statistical fundamentals. Since the estimations with the two model classes are based on the Bayesian inference and the associated Markov-Chain-Monte-Carlo methods, an introduction to the fundamentals of these methods is necessary. Chapter 3 presents the models for estimating Ecological Inferences whereby its beginnings are discussed and models for the 2x2 case, a situation with only two parties, are presented. Based on this, the extensions of these models for the RxC case are then introduced. With these models, voter transitions between any number of parties can be estimated. Afterward, the extension of the models with individual data creating the hybrid models is then analyzed. The last part of this chapter gives a literature overview of the current state of the research regarding the quantification of uncertainty in ecological inference models.

The second section, consisting of chapters four and five, focuses on investigating the reasons for the low quality of the credible intervals. These will be investigated through a simulation study, which will be presented in chapter four. In a simulation study voter transitions are simulated, which are thus known. Following this, the voter transitions are then estimated based on the simulated data basis. Since the true values are known, the quality of the estimated point and interval estimates can then be evaluated. At the beginning of the chapter, the different variables of the simulation study, as well as their specifications, are presented in addition to the technical aspects of the simulation study. Evaluation scenarios are also considered at the end of the chapter. Hypothesis behind the simulation study is the assumption that the construction of valid credible intervals works in simple data scenarios and that only more complex data scenarios lead to poor quality of the credible intervals. The first set of scenarios examines the influence of individual variable specifications on the quality of the credible intervals. In two further sets, the effects of more complex and realistic data scenarios on the quality of the interval estimators are examined. In the following chapter 5, the results are evaluated, and possible reasons for the low quality of the interval estimators are discussed.

The last section deals with possible solution approaches with which valid credible intervals can be constructed. Chapter 6 presents two concrete as well as one theoretical approach and evaluates the performance of the concrete ones using an evaluation scenario presented in chapter 4. In order to get an impression of the resulting credible intervals, the approaches are applied to a real data scenario. For this purpose, voter transitions between the Bavarian state elections in 2013 and 2018 are calculated, and the credible intervals are constructed on the basis of the proposed solution approaches. In the last chapter (Chapter 7), the results of the Master's thesis are summarised and an outlook on further open research questions is given.

2. Statistical Fundamentals

Chapter 2 describes and explains the statistical fundamentals required in this master thesis. Chapter 2.1. deals with the basics of Bayesian inference. The following sub-chapters chapter 2.1.1. - chapter 2.1.3 introduce the idea behind hierarchical models and point as well as interval estimates in the context of Bayesian inference. Chapter 2.2. describes the methodology of a Monte Carlo integration, while chapter 2.3. introduces the basics of Markov-Chain-Monte-Carlo (MCMC) techniques. The subsequent chapter 2.3.1. and chapter 2.3.2. discuss the relevant MCMC algorithms and convergence diagnostics, respectively.

2.1. Bayesian Inference

The two models used to estimate voter transitions in this thesis, namely the hybrid Multinomial-Dirichlet-Model (HMDM) as well as the Ecological Inference (EI) Model [1], both rely on Bayesian inference for the estimation. Therefore, this chapter will give a short introduction to the concept of Bayesian inference and its terminologies.

The objective of any statistical inference is to draw conclusions about an unknown quantity using the information available. There are two different approaches to statistical inference, namely the frequentist and the Bayesian approach. While the frequentist approach only incorporates information from one data source the so-called sample information, Bayesian inference also uses information obtained independently from or prior to that data. The latter is called prior information and sources could be subjective views and opinions, theories about the behavior of the unknown quantity or information derived from another data source, which is not closely related to the data under investigation. These two data sources are utilized by Bayesian inference in order to make probability statements about unknown quantities conditional on the sample and prior information.[2]

The parameter θ represents k values of the unknown quantity of interest:

$$\theta = (\theta_1, \theta_2, ..., \theta_k). \tag{1}$$

Contrary to the assumption in frequentist inference, where θ is regarded as fixed, θ is regarded as a random variable in the Bayesian inference. In addition to θ, prior information about the unknown quantity

[1] See chapter 3.3.2. and chapter 3.2.2., respectively.

[2] Peter E. Rossi, Greg M. Allenby, and Robert McCulloch, *Bayesian Statistics and Marketing* (Chichester, UK: John Wiley & Sons, Ltd, 2005). doi: 10.1002/0470863692, 13–4.

exist, which can be expressed as a probability density function

$$p(\theta). \tag{2}$$

$p(\theta)$ is the so called prior distribution. Additionally, the observed data $X = x$ follow a probability distribution, which in conditional on the k unknown parameters,

$$p(x|\theta). \tag{3}$$

This probability distribution is also called the likelihood $l(\theta)$. While frequentists regard the data X as a random variable, is assumed to be fixed in the Bayesian approach. To achieve the goal of Bayesian inference, the prior distribution $p(\theta)$ and the likelihood $l(\theta)$ have to be combined in order to compute the posterior distribution. This distribution is the most important quantity in Bayesian Inference, as it contains all the information about the unknown quantity θ after observing the data $X = x$. The calculation of the posterior density is based on Bayes Theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}. \tag{4}$$

Consequently, the posterior distribution can be defined as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}. \tag{5}$$

If the parameter θ is discrete, the denominator has to be replaced by a sum. In the case of a continuous parameter θ, the denominator can be rewritten as

$$\int p(x|\theta)p(\theta)d\theta = \int p(x,\theta)d\theta = p(x), \tag{6}$$

which is important because it emphasizes that the denominator is independent of θ. This implies that $1/f(x)$ is the normalizing constant and thus can be dropped. Hence, the posterior distribution can be approximated as follows

$$p(\theta|x) \propto p(x|\theta)p(\theta). \tag{7}$$

This means that the density of the posterior distribution is proportional to the product of the prior distribution $p(\theta)$ and the likelihood $p(x|\theta)$.[3][4][5]

2.1.1. Hierarchical Models

As seen in the previous chapter, the classical non-hierarchical approach to Bayesian inference assumes that the data $X = x$ is explained by a parameter (vector) θ, e.g. $X_i \sim p(x \mid \theta)$. However, this approach can be problematic as the specification of the distribution $p(x \mid \theta)$ can be difficult in case of complex multidimensional densities.[6][7]

A hierarchical model assumes that the distribution of $X \mid \theta$ has multiple levels. For example it might be assumed that $X_i \mid \beta \sim p_1(x \mid \beta_i)$. If β is not constant over i and follows its own distribution, one can construct a second level to the model by assuming $\beta_i \sim p_2(\beta \mid \theta)$. Thus, it is assumed that β_i follows a distribution depending on the parameter θ, which in turn is independent from i. Using these two levels it is possible to construct the distribution of interest for $X \mid \theta$, which is given by

$$p(x \mid \theta) = \int_{-\infty}^{\infty} p_1(x \mid \beta)p_2(\beta \mid \theta)d\beta = \int_{-\infty}^{\infty} p_3(x, \beta \mid \theta). \tag{8}$$

The advantage of the hierarchical setup is that it allows more complex interrelations compared to the non-hierarchical setup. However, the computation of the integral seen in equation 8 constitutes a problem as it is not always possible to computational solve the given integral. Therefore, MCMC methods are commonly used to estimate hierarchical models.[8][9]

The extension of the posterior distribution from equation 7 to a hierarchical setup is straightforward. As is it now assumed that the parameter β is conditional on another unknown parameter θ, which follows its

[3]Peter M. Lee, *Bayesian statistics: An introduction* (4. ed., 1. publ, Chichester: Wiley, 2012), 36–8.

[4]Leonhard Held and Daniel Sabanés Bové, *Applied Statistical Inference* (Berlin, Heidelberg: Springer Berlin Heidelberg, 2014). doi: 10.1007/978-3-642-37887-4, 170–1.

[5]Rossi, Allenby, and McCulloch, op. cit., 14–5.

[6]Gary King, O. R.I. Rosen, and Martin A. Tanner, "Binomial-Beta Hierarchical Models for Ecological Inference", *Sociological Methods & Research*, 28/1 (1999), 61–90. doi: 10.1177/0049124199028001004 at 69–70.

[7]André Klima, "Ökologische Inferenz und hybride Modelle: Schätzung der Wählerwanderung in Mehrparteiensystemen", Dissertation (München: Ludwig-Maximilians-Universität, 2016), 33–4.

[8]Please refer to chapter 2.3. for more information on MCMC methods.

[9]King, Rosen, and Tanner, op. cit. at 69.

own prior distribution called *hyperprior*, the posterior distribution becomes[10]

$$p(\beta, \theta \mid x) \propto p(x \mid \beta)p(\beta \mid \theta)p(\theta). \tag{9}$$

2.1.2. Point Estimates

In the Bayesian context statistical inference about the parameter θ is solely based on the posterior distribution $p(\theta|x)$. Therefore, suitable point estimators are the mean, median and mode of the posterior distribution. These three point estimators are defined as follows[11]:

The posterior mean $E(\theta|x)$ is the expected value of the posterior distribution:

$$E(\theta|x) = \int \theta p(\theta|x)dx. \tag{10}$$

The posterior mode $Mod(\theta|x)$ reflects the mode of the posterior distribution:

$$Mod(\theta|x) = \underset{x}{\text{argmax}}\, p(\theta|x). \tag{11}$$

The posterior median $Med(\theta|x)$ is the median of the posterior distribution and as such every number a that satisfies

$$\int_{-\infty}^{a} p(\theta|x)d\theta = 0.5 \quad \text{and} \quad \int_{a}^{\infty} p(\theta|x)d\theta = 0.5. \tag{12}$$

2.1.3. Interval Estimates

Interval estimates quantify the uncertainty of statistical inference. In the context of Bayesian inference, credible intervals represent such interval estimates. These credible intervals are also derived from the posterior distribution and can be seen as the Bayesian counterpart to the frequentist confidence intervals, although they have different interpretations. A $100\times(1-\alpha)$% confidence interval of the unknown quantity θ reflects the percentage of intervals, which would correctly cover θ if such intervals are repeatedly calculated from many independent random samples. On the contrary, a $100 \times (1 - \alpha)$% credible interval reflects the region where the probability of covering θ is equal to $1-\alpha$. Thus, credible intervals allow for the probability statement that the unknown parameter θ lies within the credible interval with a probability of $(1 - \alpha)$.

[10] Jeff Gill, *Bayesian Methods: A Social and Behavioral Sciences Approach, Third Edition* (Chapman & Hall / CRC Statistics in the Social and Behavioral Sciences; 3rd ed., Hoboken: CRC Press, 2015), `http://gbv.eblib.com/patron/FullRecord.aspx?p=1598081`, 425.

[11] Held and Sabanés Bové, op. cit., 171.

There a two commonly used approaches to the calculation of credible intervals: *Equal-Tail Credible Intervals* and *Highest Posteriori Credible (HPD) Intervals*.[12]

A $100 \times (1 - \alpha)$% Equal-Tail credible interval is defined using two real numbers t_l and t_u, which satisfy

$$\int_{t_l}^{t_u} p(\theta|x)d\theta = (1 - \alpha). \tag{13}$$

This credible interval is constructed in a way that the areas to the left and right of the interval each represent $\alpha/2$ of the density. These credible intervals can be obtained by choosing t_l as the $\alpha/2$-quantile and t_u as the $(1 - \alpha)/2$-quantile of the sample draws from the posterior distribution.[13][14] An example of an Equal-Tail credible interval is given in the upper graph of Figure 1. It depicts the 80% Equal-Tail credible interval from an estimation of voter transitions from and to a fictive party between two simulated elections with the HMDM. In this case, 80% of the density can be found within the credible interval in the dark blue area, while the two light blue areas in the tails to the left and right each represent 10% of the density. The two red dotted lines display t_l and t_u, respectively.

HPD Intervals are more complex in their computation but therefore allow for more flexibility. They combine two main properties, which set them apart from the Equal-Tail Credible Intervals. Firstly, the density of every point inside the HPD interval is greater compared to every point outside the HPD interval and secondly, the HPD approach produces the credible intervals with the shortest length. A $100 \times (1 - \alpha)$% credible interval is the subset C of the parameter space, which satisfies

$$C = \{\theta : p(\theta|x) \geq k\} \tag{14}$$

where k represents the largest number such that:

$$1 - \alpha = \int_{\theta : p(\theta|x) \geq k} p(\theta|x)d\theta. \tag{15}$$

In this case, k can also be thought of as a horizontal line, which separates the HPD regions with the rest. Multiple HPD regions can occur if the density function is multimodal.[15][16] An example of an HPD interval is shown in the lower graph of Figure 1, where k is displayed as a red dotted line and the HPD region is

[12] Gill, op. cit., 42–3.
[13] Held and Sabanés Bové, op. cit., 172.
[14] William Q. Meeker, JR, Gerald J. Hahn, and Luis A. Escobar, *Statistical intervals: A guide for practitioners and researchers* (2nd ed., Hoboken: John Wiley et sons, 2017), 305.
[15] Gill, op. cit., 46–7.
[16] Held and Sabanés Bové, op. cit., 176–7.

displayed in dark blue. In addition to k, the two main properties of the HPD can be observed. Compared to the credible interval in the upper graph of Figure 1 the HPD interval is smaller. Furthermore, all points inside of the HPD interval have a higher density compared to those outside the interval.

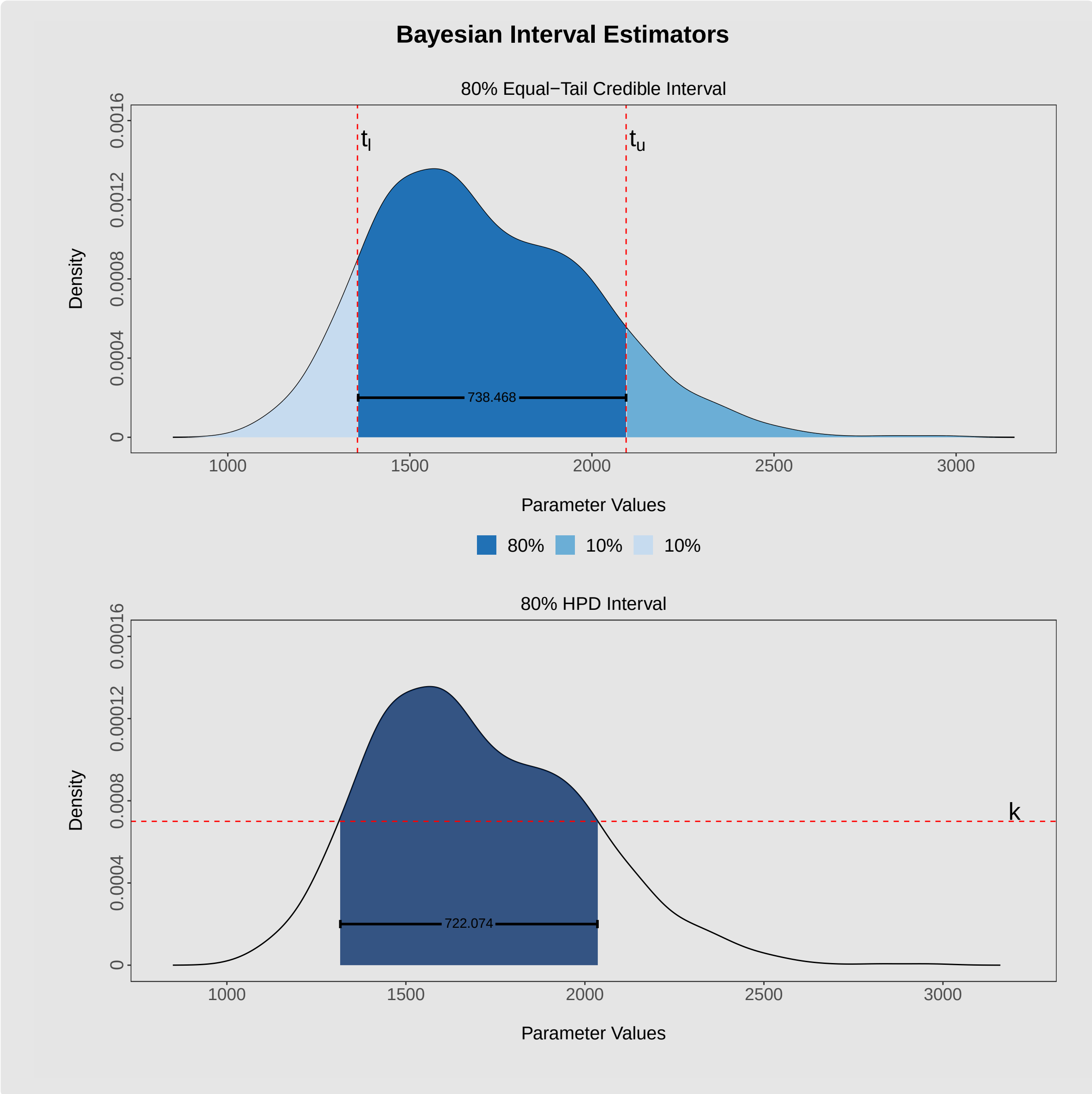


Figure 1: *Example for Bayesian interval estimators with intervals calculated for voter transitions from and to a fictive party between two simulated elections. The upper graph shows an Equal-Tail credible interval. The dark blue area displays 80% of the density, while the two light blue areas each stand for 10% of the density.* t_l *and the left red dotted line stand for the* $(alpha)/2$*-quantile, while* t_u *and the right dotted line stand for the* $(1-\alpha)/2$*-quantile. The lower graph depicts a Highest Posterior Density (HPD) credible interval, with the HPD region of the credible interval shown in dark blue. In this graph the red dotted line represents the value k, which separates the HPD region with the rest. The numbers in the middle of both graphs display the interval widths.*

2.2. Monte Carlo Integration

The previous two sections have shown the particular interest of summarizing the posterior distribution in the Bayesian inference. The general problem of summarizing the posterior distribution can be written as finding the posterior expectation of a function h of θ:

$$E_{\theta|x}[h(\theta)] = \int h(\theta)p(\theta|x)d\theta = \frac{\int h(\theta)l(\theta)p(\theta)d\theta}{\int l(\theta)p(\theta)d\theta}. \tag{16}$$

If properly defined $h(\theta)$ can represent marginal posteriors, moments, quantiles and probability intervals. Problems can arise in the calculation of certain characteristics as the posterior distribution might not be analytically solvable due to the integrals in equation 16. A simple solution is only to utilize a set of priors and likelihoods, which produce a posterior of known distributional form and for which the integrals can be solved analytically. However, modern computational tools allow for the use of numerical methods in the approximation of these integrals.[17][18][19]

This thesis will be focused on the Monte Carlo Integration and MCMC-Methods. This focus is set because the two packages used to estimate voter transitions, namely the eiwild-package[20] and the eiPack-Package[21], both rely on these methods.[22][23] Further information about other numerical methods used to approximate integrals can be found in *Monte Carlo Statistical Methods*[24] and *Bayesian computational methods*[25].

Monte Carlo integration relies on the assumption that independent samples $\theta^{(1)}, ..., \theta^{(M)}$ can be drawn

[17] Rossi, Allenby, and McCulloch, op. cit., 16–7.
[18] Christian P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (Springer texts in statistics; New York, NY: Springer, 2007). doi: 10.1007/0-387-71599-1, http://dx.doi.org/10.1007/0-387-71599-1, 285–6.
[19] Held and Sabanés Bové, op. cit., 269–70.
[20] Thomas Schlesinger, *eiwild: Ecological Inference with individual and aggregate data*, 2014, https://CRAN.R-project.org/package=eiwild.
[21] Olivia Lau, Ryan T. Moore, and Kellermann Kellermann, *eiPack: eiPack: Ecological Inference and Higher-Dimension Data*, 2012, https://CRAN.R-project.org/package=eiPack.
[22] Thomas Schlesinger, “Kombination von Aggregat- und Individualdaten bei der Analyse von RxC-Tafeln: Neue Implementierung in R”, Masterarbeit (München: Ludwig-Maximilians-Universität, 2013), 37–49.
[23] Olivia Lau, Ryan T. Moore, and Michael Kellermann, “eiPack: Ecological Inference and Higher-Dimension Data Management”, *R News*, 7(2) (2007), 43–7 at 45.
[24] Christian P. Robert and George Casella, *Monte Carlo Statistical Methods* (New York, NY: Springer New York, 2004). doi: 10.1007/978-1-4757-4145-2, Chapter 2-5.
[25] Adrian Frederick Melhuish Smith, David Roxbee Cox, and D. M. Titterington, “Bayesian computational methods”, *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337/1647 (1991), 369–86. doi: 10.1098/rsta.1991.0130.

from the posterior distribution $p(\theta|X)$. A Monte Carlo estimation of equation 16 is then given by

$$\hat{E}_{\theta|x}(h(\theta)) = \frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)}). \tag{17}$$

Accordingly, the Monte Carlo estimate of the posterior mean is given by

$$\hat{E}(\theta|x) = \frac{1}{M}\sum_{m=1}^{M} \theta^{(m)}. \tag{18}$$

This holds true because the law of large numbers ensures that the estimate converges to the true posterior mean as $M \to \infty$.[26][27]

2.3. Markov-Chain-Monte-Carlo

MCMC methods are used to evaluate integral quantities and as such represent a method to approximate the integrals in equation 16. While the foundation of MCMC methods was already laid by the works of Metropolis et al.[28] as well as Geman and Geman[29], their application has been growing exponentially with modern computational advances. As the name already reveals MCMC methods are a combination of two parts: Markov Chains and Monte Carlo Integration. As discussed in the previous chapter, Monte Carlo Integration represents a method to calculate characteristics of the posterior distribution if they cannot be solved analytically. However, this method relies on the assumption that independent samples $\theta^{(1)}, ..., \theta^{(M)}$ can be drawn from the posterior distribution $p(\theta|x)$. MCMC methods can be used to generate such samples from the posterior distribution in the form of a Markov Chain.

A Markov Chain is a stochastic process with the property that the movement probability to any new state θ^{t+1} only depends on the current state of the process θ^{t}. Thus, θ^{t+1} is conditionally independent from the former states $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(t-1)}$:

$$p(\theta^{(t+1)} \in A|\theta^{(1)}, ..., \theta^{(t-1)}, \theta^{(t)}) = p(\theta^{(t+1)} \in A|\theta^{(t)}), \tag{19}$$

where A represents any identified set on the complete state space which represents the allowable range of values for the random vector of interest. Therefore, a Markov Chain can be thought of as a system which

[26] Held and Sabanés Bové, op. cit., 258–65.
[27] Robert, op. cit., 294–8.
[28] Nicholas Metropolis et al., "Equation of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, 21/6 (1953), 1087–92. doi: 10.1063/1.1699114.
[29] Stuart Geman and Donald Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6/6 (1984), 721–41. doi: 10.1109/TPAMI.1984.4767596.

randomly moves through a series of states without having any memory of where it has been. Only the current state influences the next step of the system.[30]

An important aspect of a Markov Chain is the transition probability. The transition probability indicates the probability of moving to state $\theta^{(t)}$ at time point t from the state $\theta^{(t-1)}$ at time point $t-1$. Robert and Casella[31] have defined this transition probability in terms of a transition Kernel K. This definition has the advantage of grasping both the continuous state space as well as the discrete state space. In the case of a discrete state space, K represents a $k \times k$-Matrix for the k discrete elements in A:

$$P_A = \begin{bmatrix} p(\theta_1, \theta_1) & \cdots & p(\theta_1, \theta_k) \\ \vdots & & \vdots \\ p(\theta_k, \theta_1) & \cdots & p(\theta_k, \theta_k) \end{bmatrix}.$$

The rows of the matrix represent where the process currently is, while the columns show where the process is going to be in the next period. The rows of the transition matrix A are a probability distribution with the elements $p(\theta_i, \theta_j)$ being transition probabilities. The transition probabilities must satisfy:

$$\sum_{j=1}^{k} p(\theta_i, \theta_j) = 1 \tag{20}$$

and

$$p(\theta_i, \theta_j) \geq 0. \tag{21}$$

In the case of a continuous state space, K represents a probability density function $p(\theta|\theta_i)$.[32][33]

An important aspect of the transition Kernel is that all further states of a Markov chain can theoretically be developed from the knowledge of the first value $\theta^{(1)}$ and K. For example, the transition probability from the state $\theta_i = x$ at time point 1 to state $\theta_j = y$ in m steps is given by:

$$p^m(\theta_j^{(m)} = y|\theta_i^{(1)} = x) = \underbrace{\sum_{\theta_2} \sum_{\theta_3} \cdots \sum_{\theta_{(m-1)}}}_{\text{all possible paths}} \underbrace{p(\theta_i, \theta_2) p(\theta_2, \theta_3) \cdots p(\theta_{(m-1)}, \theta_j)}_{\text{transition products}}. \tag{22}$$

[30] Gill, op. cit., 333–4.
[31] Robert and Casella, op. cit., 208.
[32] Gill, op. cit., 333–5.
[33] Peter W. Jones and Peter Smith, *Stochastic processes: An introduction* (Chapman & Hall/CRC Texts in Statistical Science Series; Third edition, Boca Raton, London, and New York: CRC Press, Taylor & Francis Group, 2018), 65–6.

In the case of a continuous Markov chain the summations need to be replaced with integrals.[34]

A Markov Chain is said to be *irreducible* if for each pair of states $a_i, a_j \in A$ there is an n such that $P(\theta^{(m+n)} = a_j|\theta^{(n)} = a_i) > 0$. That means that every point or collection of points A can be reached from every other point or collection of points. An irreducible Markov Chain is called *recurrent* if at a given state A the probability that the chain returns to this state infinitely often is nonzero: $P(\theta_n \in$ A *infinitely often*$) > 0$. One speaks of *positive recurrence* if $\mathbb{E}(T_i|\theta^{(0)} = A) < \infty$, meaning if the mean time to return to A, $T_i = min\{n \geq 1|\theta^{(n)} = i\}$, is finite. In case the probability of visiting A infinitely often is 1, $P(\theta_n \in$ A *infinitely often*$) = 1$, one speaks of *Harris recurrence*. The linkage of *irreducibility* and *recurrence* is important because it ensures the creation of a subspace in which every point can be reached by the Markov Chain, while ensuring that the chain will explore all of the subspace.[35][36]

A further important aspect of Markov Chains is the concept of *stationarity*. Given are a Markov Chain with the state space A, a transition probability $p(\theta_i, \theta_j)$ and the stationary distribution of the Markov Chain $\pi(\theta)$. In this case $\pi(\theta)$ represents the posterior distribution of interest. Based on the definition of this stationary distribution it satisfies

$$\sum_{\theta_i} \pi^t(\theta_i)p(\theta_i, \theta_j) = \pi^{t+1}(\theta_j) \tag{23}$$

$$\int \pi^t(\theta_i)p(\theta_i, \theta_j)d\theta = \pi^{t+1}(\theta_j) \tag{24}$$

for the discrete and continuous case, respectively. In Matrix notation this can be written as $P\pi = \pi$. The above equations state that if the marginal distribution at any given step n reaches the stationary distribution then the distribution at the next step is $P\pi = \pi$. Thus, if the stationary distribution has been reached, the chain stays within this distribution for all subsequent steps. It can be shown that if a stationary distribution exists and $\lim_{t\to\infty} p^t(\theta_i, \theta_j) = \pi(\theta_j)$ the distribution of the chain will reach the stationary distribution as $n \to \infty$[37]. In this case, one also speaks of a limiting distribution. Reaching the stationary distribution is precisely what is expected when employing MCMC-Methods. If the stationary distribution and thus the posterior distribution is reached, it can be summarized by the Monte Carlo Integration using the empirical samples from the Markov Chain.[38][39]

[34] Gill, op. cit., 335–5.

[35] James R. Norris, *Markov chains* (Cambridge series on statistical and probabilistic methods; 15. printing, Cambridge: Cambridge Univ. Press, 2009), 24–9.

[36] Gill, op. cit., 340–1.

[37] Dani Gamerman and Hedibert Freitas Lopes, *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference* (Texts in statistical science series; 2. ed., Boca Raton Fla. u.a.: Chapman & Hall/CRC, 2006), 121.

[38] Gill, op. cit., 341–2.

[39] Gamerman and Lopes, op. cit., 121–4.

However, situations also exist in which a stationary distribution exists while a limiting one does not. [40] Therefore one more characterization of the states needs to be introduced in order to ensure the limiting results. This further characterization is called *periodicity*. A state is called *aperiodic* if the only common denominator for the pair $\{t \geq 1 : P(\theta, \theta) > 0\}$ is equal to one. This means that the only length of time in which the chain repeats a cycle of values is equal to one. In the discrete case an irreducible, positive recurrent and aperiodic Markov Chain is called *ergodic*. In the continuous case the characteristic of positive recurrence is replaced with Harris recurrence. Ergodic Markov Chains satisfy

$$\lim_{t \to \infty} p^t(\theta_i, \theta_j) = \pi(\theta_j) \tag{25}$$

for all θ_i and θ_j in the subspace. Thus, in the limit the marginal distribution at one step is identical to the marginal distributions at all other steps and the posterior distribution of interest can be examined with samples from the Markov Chain.[41][42]

2.3.1. MCMC-Algorithms

When it comes to algorithms implementing the MCMC methods, the Metropolis-Hastings algorithm introduced by Metropolis[43] and Hastings[44] as well as the Gibbs Sampler by Geman and Geman[45] are the most popular ones. While many more advanced and specialized extensions of these algorithms have been developed [46], this subsection will concentrate on the two most popular methods as these are the ones applied in the eiPack-package[47] and the eiwild-package[48].[49]

The groundwork for the Metropolis-Hastings algorithm was laid by Metropolis[50] in a paper dealing with the calculation of properties of chemical substances. Only later the proposed method was generalized by Hastings[51] and thus became the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm starts with a target distribution $\pi(\theta)$, the posterior distribution of

[40] see Gamerman and Lopes (ibid., 121) for an example.
[41] Gill, op. cit., 342–3.
[42] Gamerman and Lopes, op. cit., 124–32.
[43] Metropolis et al., op. cit.
[44] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, 57/1 (1970), 97. doi: 10.2307/2334940.
[45] Geman and Geman, op. cit.
[46] See for example (Gamerman and Lopes, op. cit.) and (Robert and Casella, op. cit.) for an overview.
[47] Lau, Ryan T. Moore, and Michael Kellermann, loc. cit.
[48] Schlesinger, op. cit., 27.
[49] Ioannis Ntzoufras, *Bayesian modeling using WinBUGS* (Wiley series in computational statistics; Hoboken, NJ: Wiley, 2009). doi: 10.1002/9780470434567, http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10296519, 42.
[50] Metropolis et al., op. cit.
[51] Hastings, op. cit.

interest, from which a sample of the size T is to be generated. The algorithm then iterativly goes through the steps shown in Figure 2, where θ is the vector of generated values and i represents the iteration:

1. Initialize the iteration at counter $i = 1$ and set an arbitrary initial value $\theta^{(0)}$.
2. Move the chain to a new value θ^* generated from the proposal density $q(\theta^{(i)}, \cdot)$.
3. Calculate the acceptance probability $\alpha(\theta, \theta^*) = min\left\{1, \frac{\pi(\theta^*)q(\theta^*,\theta)}{\pi(\theta)q(\theta,\theta^*)}\right\}$.
4. Generate an independent uniform quantity u from a continuous (0-1) - uniform distribution $U(0,1)$.
5. If $u \leq \alpha$ the move is accepted and $\theta^{i+1} = \theta^*$. Otherwise the move is not allowed and $\theta^{i+1} = \theta^i$.
6. Change the counter form i to $i + 1$ and return to step 2 until convergence is reached.

Figure 2: *Metropolis-Hastings algorithm in pseudo code. (Gamerman and Lopes, op. cit., 195)*

At the i-th step of the Markov Chain, when the chain is at θ^i, θ^* is drawn from the proposal density. In this case, θ can represent a vector as well as a scalar. The proposal density can be chosen freely because the algorithm will converge to its equilibrium distribution regardless of the proposal density. However, a poorly chosen proposal density can delay the convergence of the algorithm considerably. Afterward the acceptance probability α has to be calculated. Keeping the acceptance ratio within a reasonable range is essential. In case of a very high acceptance ratio, most of the moves would be accepted. This leads to the chain making many minuscule steps, which would require a large number of iterations to achieve convergence of the chain. The long convergence time results due to the fact, that the chain must be able to move through the whole parameter space in order to converge to the equilibrium distribution. On the other hand, a low acceptance ratio can lead to the chain being stuck at $\theta^{(i+1)} = \theta^{(i)}$ as the next steps are being rejected. Thus, the probability distribution $q(\theta^{(i)}, \cdot)$ must be chosen in a way that the chain moves with considerable displacements while having a substantial probability of being accepted. Given the ratio in the acceptance probability, the target distribution $\pi(\theta)$ only needs to be known until a constant of proportionality. Finally, after comparing α to the independent uniform quantity u the decision can be made whether the chain stays at the current location or moves to the next one.[52][53][54]

A special case of the Metropolis-Hastings algorithm is given when the acceptance ratio is always 1. This is

[52]Gill, op. cit., 354–5.
[53]Ntzoufras, op. cit., 42–3.
[54]Gamerman and Lopes, op. cit., 193–6.

the case if the proposal density is equal to the target density, $q(\theta, \theta^*) = \pi(\theta^*)$. The Gibbs-Sampler, which was introduced by Geman and Geman[55], makes this assumption. Compared to the Metropolis-Hastings Algorithm the Gibbs-Sampler is more restrictive as it requires the availability of the full conditionals of the target distribution, $\pi(\theta)$ with $\theta = (\theta_1, ..., \theta_d)$, of the form

$$\pi_i(\theta_1) = \pi(\theta_i|\theta_{-i}), i = 1, ..., d. \tag{26}$$

This means that they have to be fully known and can be sampled from. It facilitates the iterative nature of the Gibbs Sampler, which cycles through the full conditionals and iterately updates the elements of θ by drawing parameters from the full conditionals based on the most recent values for the previous parameters. In Figure 3 the procedure is shown in the form of pseudo code:

1. Initialize the iteration counter of the chain $i = 1$ and set initial values $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_P^{(0)})$.
2. Obtain a new value $\theta^{(j)} = (\theta_1^{(j)}, ..., \theta_d^{(j)})$ from $\theta^{(j-1)}$ through successive generation of values
   - $\theta_1^{(j)} \sim \pi(\theta_1|\theta_2^{(j-1)}, ..., \theta_d^{(j-1)})$,
   - $\theta_2^{(j)} \sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, ..., \theta_d^{(j-1)})$,

   $\vdots$
   - $\theta_d^{(j)} \sim \pi(\theta_d|\theta_1^{(j)}, ..., \theta_{d-1}^{(j-1)})$.
3. Change counter i to $i + 1$ and return to step 2 until convergence is reached.

Figure 3: *Gibbs-Sampler in pseudo code. (Gamerman and Lopes, op. cit., 142-3)*

In some complex cases a direct sampling from the full conditionals might not be possible for all components of θ. For this scenario Müller[56] proposed what came to be known as the Metropolis-within-Gibbs-Sampler. In this sampler the components, which cannot be directly sampled from the full conditionals, are sampled from proposal density q using the Metropolis-Hastings algorithm.[57][58] This approach is used in both the eiPack-ackage[59] as well as the eiwild-ackage[60].

[55] Geman and Geman, op. cit.
[56] Peter Müller, *Alternatives to the Gibbs Sampling Scheme* (1992).
[57] Refer to appendix 1 for a pseudo code of the Metropolis-within-Gibbs-Sampler
[58] Gamerman and Lopes, op. cit., 213.
[59] Lau, Ryan T. Moore, and Michael Kellermann, loc. cit.
[60] Schlesinger, loc. cit.

2.3.2. Convergence Diagnostics

When utilizing MCMC methods, it is important to generate enough iterations for the Markov Chain to converge. Only if a convergence takes place the samples from the Markov Chain can be used in a Monte Carlo integration. Thus, in order to ensure convergence, a few practical considerations have to be made before implementing an MCMC algorithm. These considerations include decisions regarding the starting points used to initialize the chains, the burn-in period, the thinning parameter as well as the sample size.

The choice of starting points can influence the posterior summaries if they are far away from the stationary distribution because the chain might need longer to converge to said distribution. There are several approaches used to choose the right starting points. A first rudimentary approach is to run several chains with different starting values. If they all converge to different regions of the state space, it could be a sign for non-convergence and other starting values should be used. Unfortunately, the opposite does not hold, if all chains converge to the same region one cannot conclude that the chain properly converged as they could all have been attracted by the same local maxima. A more sophisticated approach is the use of so-called *points of theoretical interest*. These could be points used in other studies, previous works with the same data set or chosen by subject experts. In a third approach called *overdispersion* the starting points are chosen relative to the expected modal point. If it is not possible or highly complicated to evaluate the mode one can also use starting values which are randomly distributed through the state space.[61][62] The last approach is used in the eiwild-package and the eiPack-package to choose the starting points of the chains.[63][64]

In order to ensure that the posterior summaries only consists out of values from the chain after it converged to the stationary distribution, the first N iterations of the chain $(\theta_1, ..., \theta_t)$ are discarded. This is called the burn-in period. The thinning parameter refers to a practice used to reduce the autocorrelation in the chain. Since every value of the chain depends on its preceding values, the resulting chain is not independent. Thus, the chain is thinned out by only using every $t - th$ value of the chain. The sample size refers to the length of the chain. The longer the chain runs, the more stable the posterior summaries becomes.[65][66]

[61] Gill, op. cit., 476–7.
[62] Ntzoufras, op. cit., 38.
[63] Thomas Schlesinger, *Ecological Inference with individual and aggregate data*, 2014, `https://cran.r-project.org/web/packages/eiwild/eiwild.pdf`, 9.
[64] Olivia Lau, Ryan T. Moore, and Michael Kellermann, *eiPack: Ecological Inference and Higher-Dimension Data Management*, 2012, `https://cran.r-project.org/web/packages/eiPack/eiPack.pdf`, 5.
[65] Ntzoufras, loc. cit.
[66] Robert, op. cit., 303.

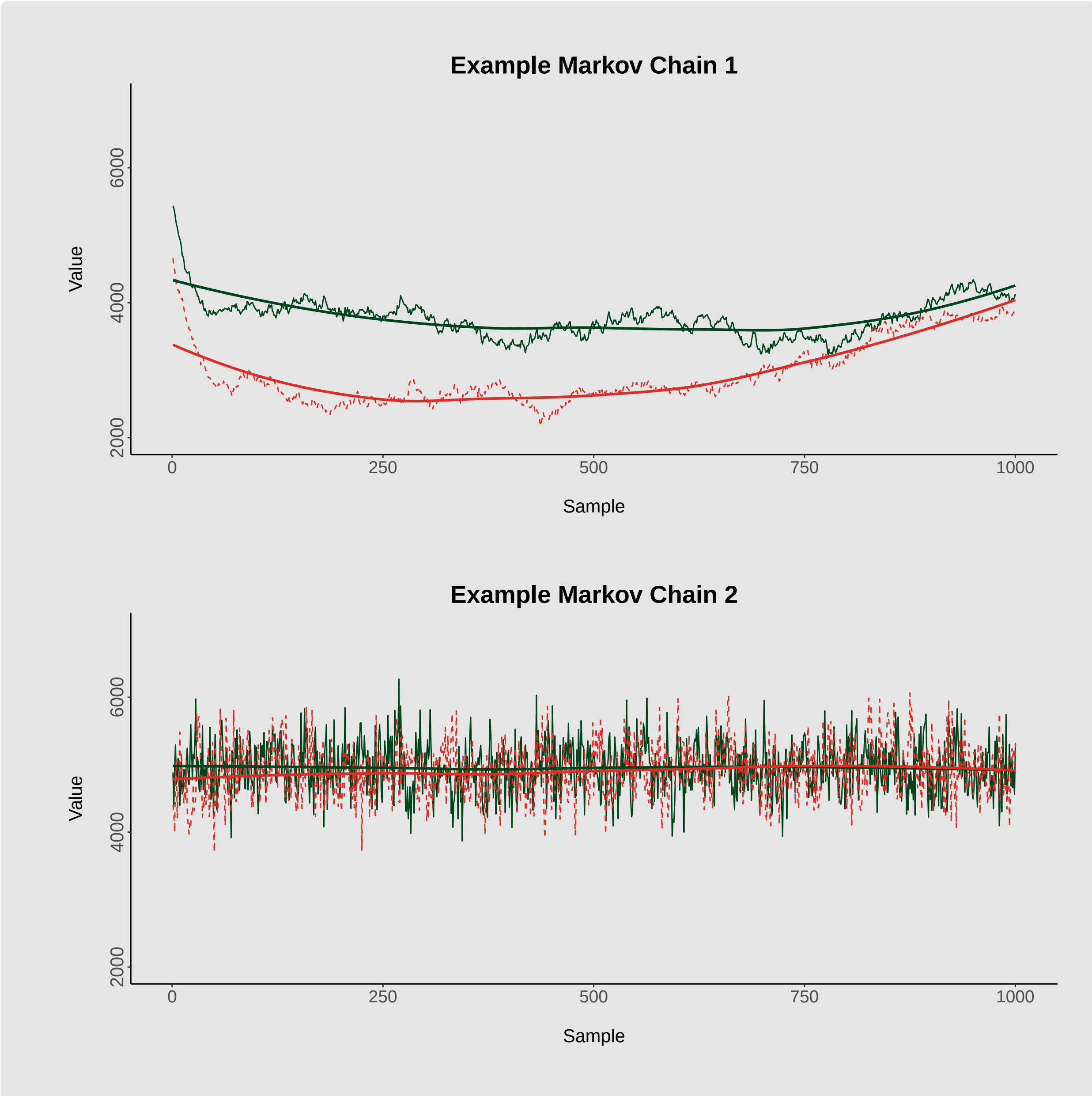


Figure 4: *Example of Markov Chains with different parameters from an estimation of voter transitions with the hybrid Multinomial-Dirichlet Model. For the chains in the upper graph, the parameters burn-in=0, thinning=1 and sample=1000 were chosen. The parameters for the chains in the lower graph were set to burn-in=250.000, thinning=1000 and sample=1000. While the chains in the upper graph show no convergence, the chains in the lower graph converge.*

In this master thesis, a burn-in parameter of 250.000, as well as thinning and sample parameters of 1000, were used. These parameter values were chosen because they led to a reasonable convergence of the chains. The convergence can be checked by generating and comparing several chains $(\theta_1, ..., \theta_t)$. Among other things, the similarity of the location measures, the autocorrelation between the realizations of the chains or the presence of a trend in the chains can be checked. Further methods for verifying convergence can be found in chapter 12 of Robert and Casella[67].

Figure 4 illustrates the influence of the parameters on the convergence with two Markov chains from models which were estimated in the context of this master thesis. The upper graph shows an example of two Markov chains where the burn-in parameter was set to 0 and the thinning and sample parameters to 1 and 1000, respectively. These chains have not converged yet, mainly due to the missing burn-in and a high autocorrelation between the realizations. To achieve convergence higher values should be chosen for the burn-in and thinning parameters. The lower graph of Figure 4 displays two Markov Chains from the same models which were estimated using the earlier specified values for the burn-in, thinning and sample parameter. The higher parameter values have led to a convergence of the chains.

3. Ecological Inference

The problem of ecological inference made its first appearances in the literature in the early 20th century in works by Ogburn and Goltra[68], Allport[69] as well as Gehlke and Biehl[70]. It describes a situation in which information is only available on an aggregated level, but the interest lies in the behavior on a less aggregated, mostly individual level. The goal is to infer information about individual-level behavior based on the information reported on an aggregated level. The name, ecological inference, comes from the fact that most of the aggregated information used in the inference is reported in geographical/ecological units such as counties, precincts or municipalities.[71][72]

The scientific fields which employ ecological inference in their research are manifold. Political Scientists for example use it to estimate voter transitions (Klima[73] and Klima et al.[74]; King[75]), epidemiologists to investigate associations between risk and exposure (Wakefield[76]) and social scientists to research sociological

[67] Robert and Casella, op. cit., Chapter 12.

[68] William F. Ogburn and Inez Goltra, "How Women Vote", *Political Science Quarterly*, 34/3 (1919), 413. doi: 10.2307/2141684.

[69] Floyd H. Allport, "The Group Fallacy in Relation to Social Science", *American Journal of Sociology*, 29/6 (1924), 688–706. doi: 10.1086/213647.

[70] C. E. Gehlke and Katherine Biehl, "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material", *Journal of the American Statistical Association*, 29/185 (1934), 169. doi: 10.2307/2277827.

[71] Adam Glynn and Jon Wakefield, "Ecological Inference in the Social Sciences", *Statistical methodology*, 7/3 (2010), 307–22. doi: 10.1016/j.stamet.2009.09.003. at 1–3.

[72] Wendy Tam Cho and Charles F. Manski, *Cross–Level/Ecological Inference* (The Oxford Handbook of Political Methodology; Oxford University Press, 2009). doi: 10.1093/oxfordhb/9780199286546.003.0024, 1–2.

[73] André Klima et al., "Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches", *AStA Advances in Statistical Analysis*, 100/2 (2016), 133–59. doi: 10.1007/s10182-015-0254-8.

[74] André Klima et al., "Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions", *Sociological Methods & Research*, 1 (2017), 004912411770147. doi: 10.1177/0049124117701477; André Klima et al., *Exit Polls und Hybrid-Modelle* (Wiesbaden: Springer Fachmedien Wiesbaden, 2017). doi: 10.1007/978-3-658-15674-9.

[75] Gary King, *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data* (Princeton, NJ: Princeton University Press, 1997). doi: 10.2307/j.ctt46n43p, http://www.jstor.org/stable/10.2307/j.ctt46n43p.

[76] Jonathan Wakefield, "Ecologic studies revisited", *Annual review of public health*, 29 (2008), 75–90. doi: 10.1146/annurev.publhealth.29.020907.090821.

issues such as the connection between the ethnic background and illiteracy (Robinson[77]). Further fields of application can be found in economics (Kramer[78]) and geography (Openshaw[79]). This master thesis deals with the estimation of voter transitions. Thus, the following examples will be focused on this area.

In general, the problem ecological inference wants to solve is one of known margins and unknown inner cells. In the case of voter transitions, the election results are known, but it cannot unequivocally be concluded how many voters stayed loyal to their party and how many switched votes between elections. The notation for this problem is displayed in Table 1.

Situation in absolute values

Election 1 \ Election 2	Party 1	Party2	Election Result 1
Party 1	$Y_i^{1,1}$	$Y_i^{1,2}$	$N_{1,i}$
Party 2	$Y_i^{2,1}$	$Y_i^{2,2}$	$N_{2,i}$
Election Result 2	$T_{1,i}$	$T_{2,i}$	N_i

Situation in relative values

Election 1 \ Election 2	Party 1	Party2	Election Result 1
Party 1	$\beta_i^{1,1}$	$\beta_i^{1,2}$	$n_{1,i}$
Party 2	$\beta_i^{2,1}$	$\beta_i^{2,2}$	$n_{2,i}$
Election Result 2	$t_{1,i}$	$t_{2,i}$	n_i

Note: $N_{r,i}$ ($n_{r,i}$) display the results of the first election, $T_{c,i}$ ($t_{c,i}$) represent the results of the second election and $Y_i^{r,c}$ ($\beta_i^{r,c}$) display the voter transition. In all three cases the values refer to absolute (relative) values. i represents the district.

Table 1: *Basic notation of a 2x2 voter transition table with two fictitious parties in the case of absolute values (upper graph) and relative values (lower graph).*

The known elements of the problem are the election results given by the margins of the contingency table N_r (n_r) and T_r (t_r). These margins display the election results in absolute (relative) numbers from the first

[77] W. S. Robinson, "Ecological Correlations and the Behavior of Individuals", *American Sociological Review*, 15/3 (1950), 351. doi: 10.2307/2087176.

[78] Gerald H. Kramer, "The Ecological Fallacy Revisited: Aggregate- versus Individual-level Findings on Economics and Elections, and Sociotropic Voting", *American Political Science Review*, 77/01 (1983), 92–111. doi: 10.2307/1956013.

[79] Stan Openshaw, *The modifiable areal unit problem* (Concepts and techniques in modern geography; Norwich: Geo Abstracts Univ. of East Anglia, 1984).

and second election, respectively. The quantities of interest are given by the interior cells $Y^{r,c}$ and $\beta^{r,c}$, which represent the voter transitions in absolute and relative numbers. $\beta^{r,c}$ is also called the transitions probability. It depicts the probability of staying loyal to one party ($\beta^{1,1}$ and $\beta^{2,2}$) or switching votes between elections ($\beta^{1,2}$ and $\beta^{2,1}$). Thus, it can also be defined as the conditional probability of voting for party c in the second election given the person voted for party r in the first election, $\beta_i^{r,c} = Y_i^{r,c}/N_{r,i}$. The inner cells remain unobserved because of the secret ballot. The fundamental difficulty of ecological inference is that one is interested in the internal counts yet only overserves the sums of these internal counts.

In the case of the 2x2 contingency table shown in Table 1 , the following equation applies for every district i,

$$T_{c,i} = \beta_i^{1,c} \times N_{1,i} + \beta_i^{2,c} \times N_{2,i}. \tag{27}$$

Thus, the election results of Party c in the second election can be calculated by multiplying the first election results with the respective transition probabilities.

Election 1 \ Election 2	Party 1	Party 2	Election Result 1
(A)			
Party 1	80	0	80
Party 2	0	20	20
(B)			
Party 1	60	20	80
Party 2	20	0	20
Election Result 2	80	20	100

Note:(A) represents a scenario in which all voters stay loyal to the party, which they voted for in the first election, while (B) displays a more dispersed voting patter. The right margins represent $N_{r,i}$, the lower margins T, c, i and the inner cells $Y_i^{r,c}$.

Table 2: *Hypothetic example of two valid voter transition tables given the same margins.*

However, applying this to every of the I districts yields a system of I equations with $2I$ unknown quantities, namely $\beta_i^{1,c}$ and $\beta_i^{2,c}$. This reflects the problem of a fundamental indeterminacy the ecological inference is facing. As shown in Table 2 one can construct multiple values for these parameters, which satisfy the restrictions formulated in equation 27 and result in valid voter transitions given the same margins. This phenomenon of making wrong inferences about individual behavior from aggregated data was first described by Robinson[80]. Later Selvin[81] coined the term *ecological fallacy*.[82][83][84]

In order to solve the problem of ecological inference further assumptions regarding the parameters to be estimated need to be made. Numerous approaches to this problem can be found in the literature. Some of these approaches are discussed in the following chapters. Chapter 3.1 deals with the beginnings of the ecological inference and presents Goodman's *Ecological Regression* as well as the *Method of Bounds* according to Duncan and Davis. In chapter 3.2 hierarchical models for the estimation of voter transitions are presented. With Kings *Ecological Inference* model, chapter 3.2.1 discusses an approach for the 2x2 case. Chapter 3.2.2 presents Rosen et al.'s *Multinomial-Dirichlet-Model*, an extension of King's approach for the RxC case. With the goal of improving the estimation, *hybrid models* incorporate individual data into the estimation process. These models are presented in chapter 3.3. While chapter 3.3.1 discusses the first approach to a hybrid model for the 2x2 case by Wakefield, chapter 3.3.2 presents Klima et al. extension of Wakefield's to the RxC case, the *hybrid Multinomial-Dirichlet Model*. The latter represents the second model used to estimation voter transitions in this thesis. Due to this thesis focus on the quality of credible intervals when estimating voter transitions with (hybrid) Multinomial-Dirichlet Models, chapter 3.4 gives a literature overview of the estimation of uncertainty with ecological inference models.

3.1. The Beginnings of Ecological Inference

With the *Method of Bounds* Duncan and Davis[85] proposed one of the first approaches to ecological inference. The focus of this method lies in extracting deterministic information from the data. In the same year Goodman[86] offered a statistical approach to the problem with the *Ecological Regression*. For nearly half a century these were the only two approaches proposed to solve the problem of ecological inference.[87]

Duncan and Davis developed a method to extract deterministic information, which is known with abso-

[80] Robinson, op. cit.

[81] Hanan C. Selvin, "Durkheim's Suicide and Problems of Empirical Research", *American Journal of Sociology*, 63/6 (1958), 607–19. doi: 10.1086/222356.

[82] Klima et al., "Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions" at 4–5.

[83] Glynn and Wakefield, op. cit. at 3–4.

[84] Cho and Manski, op. cit., 2–4.

[85] Otis Dudley Duncan and Beverly Davis, "An Alternative to Ecological Correlation", *American Sociological Review*, 18/6 (1953), 665. doi: 10.2307/2088122.

[86] Leo A. Goodman, "Ecological Regressions and Behavior of Individuals", *American Sociological Review*, 18/6 (1953), 663. doi: 10.2307/2088121; Leo A. Goodman, "Some Alternatives to Ecological Correlation", *American Journal of Sociology*, 64/6 (1959), 610–25. doi: 10.1086/222597.

[87] Gary King, Ori Rosen, and Martin A. Tanner, *Ecological Inference* (Cambridge: Cambridge University Press, 2004). doi: 10.1017/CBO9780511510595, 2–3.

lute certainty. The intuition behind this method is rather easy and does not require any further assumptions to be made. Going back to the situation formulated in Table 1 and assuming, that only the margins of the contingency table are known, ranges between 0% and 100% containing the true transition probabilities can be constructed from these margins. These ranges are calculated according to the following equations:

$$\beta_i^{1,c} \in \left[max\left(0, \frac{t_{c,i} - n_{2,i}}{n_{1,i}}\right), min\left(\frac{t_{c,i}}{n_{1,i}}, 1\right)\right], \quad (28)$$

$$\beta_i^{2,c} \in \left[max\left(0, \frac{t_{c,i} - n_{1,i}}{n_{2,i}}\right), min\left(\frac{t_{c,i}}{n_{2,i}}, 1\right)\right]. \quad (29)$$

Bounds that range over several districts are simply weighted averages of the district level bounds.[88][89] Interestingly, the resulting ranges also reflect 100% confidence intervals, as the true value always lies within them.[90] However, they might be too wide and not precise enough to be of any practical use. In fact, this has been the main criticism of this method.[91]

The Ecological Regression, sometimes also called Goodman Regression, was introduced by Goodman[92] based on an idea discussed in *How Women Vote* by Ogburn[93]. In contrast to the Method of Bounds, the approach extracts statistical information and at the same time uses information from all the districts together rather than independently. Goodman formalized a regression model between ecological variables in order to infer information about individual behavior:

$$t_{c,i} = n_{1,i} \times \beta_i^{1,c} + n_{2,i} \times \beta_i^{2,c}. \quad (30)$$

However, applying this regression model to the situation formulated in Table 1 would lead to the same problem of indeterminacy described at the beginning of chapter 3. Thus, Goodman pointed out that this approach only works if the so-called *constancy assumption* holds. This assumption states that the parameters to be estimated $\beta_i^{1,c}$ and $\beta_i^{2,c}$, as well as the margins $n_{r,i}$ must be uncorrelated: $Cov(\beta_i^{1,c}, n_{1,i}) = Cov(\beta_i^{2,c}, n_{2,i}) = 0$.[94] Applying the constancy assumption leads to a situation with constant β's. This would solve the problem of indeterminacy because it would lead to a system of I equations with only I un-

[88] Ibid., 3.
[89] Cho and Manski, op. cit., 6–7.
[90] King, Rosen, and Tanner, loc. cit.
[91] Klima et al., "Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches" at 135.
[92] Goodman, "Ecological Regressions and Behavior of Individuals"; id., "Some Alternatives to Ecological Correlation".
[93] Ogburn and Goltra, op. cit.
[94] Later King, op. cit., Chapter 3 showed, that along with the constancy assumption formulated by Goodman, the number of voting aged people N_i and the parameters also need to be uncorrelated in order to get unbiased results.

knowns rather than $2I$ unknowns. However, Goodman already recognized that this approach yields highly biased results, sometimes even outside the unit interval, in the case that his assumptions do not hold.

3.2. Hierarchical Models

3.2.1. King's Ecological Inference Model

The ecological inference model proposed by King[95] combines the two approaches introduced in chapter 3.1. He realized that the two sources of information could supplement each other in order to improve the estimation. Thus, King combines the determinist information from the bounds with a statistical approach for extracting information from within the bounds. This model is less dependent on assumptions and can be seen as the first approach to use a hierarchical model for ecological inferences. King's notation, which has been adapted to the estimation of voter transitions, is displayed in Table 3.[96][97]

Election 1 \ Election 2	Party 1	Party2	Election Result 1
Party 1	$\beta_i^{1,1}$	$1-\beta_i^{1,1}$	$n_{1,i}$
Party 2	$\beta_i^{2,1}$	$1-\beta_i^{2,1}$	$1-n_{1,i}$
Election Result 2	$t_{1,i}$	$1-t_{1,i}$	

Note: $n_{1,i}$ displays the results of the first election of Party 1, $t_{1,i}$ represents the result of Party 1 in the second election and $\beta_i^{r,c}$ display the transition probabilities. In all three cases the values refer to relative values. i represents the district.

Table 3: *Basic notation of a 2x2 voter transition table with two parties in the case of relative values according to King. (King, op. cit.)*

The starting point of King's EI model is Goodman's Ecological Regression formula given in equation 30, which King solves for one of the unknown parameters:

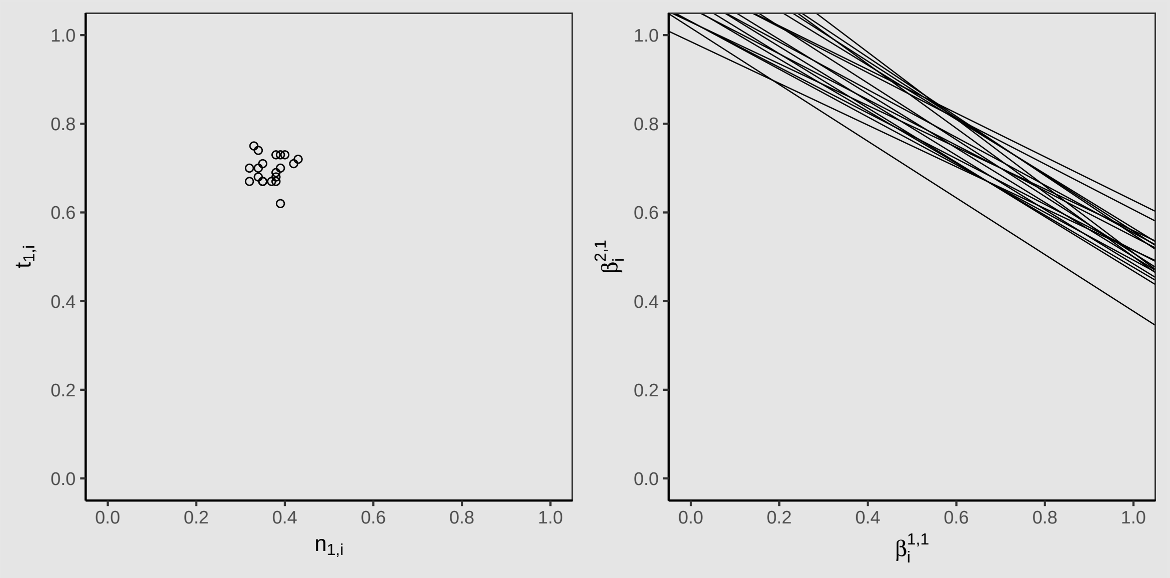
$$\beta_i^{2,1} = \frac{t_{1,i}}{1-n_{1,i}} - \frac{n_{1,i}}{1-n_{1,i}}\beta_i^{1,1}. \tag{31}$$

Equation 31 reflects the linear connection between the two unknown parameters $\beta_i^{2,1}$ and $\beta_i^{1,1}$. The slope and the intercept are known, since they are functions of the known margins $n_{1,i}$ and $t_{1,i}$. Thus, it is

[95]Ibid.
[96]Klima, op. cit., 33.
[97]King, Rosen, and Tanner, op. cit., 4.

possible to construct a line for each district using the known slope and intercept on which the true values of $\beta_i^{2,1}$ and $\beta_i^{1,1}$ have to lie. King called these lines *tomography lines*.

Extending on the example given in Table 3, the connection between the aggregated values and the unknown parameters of interest can be shown with the tomography lines. The left plot Figure 5 displays the connection between the election results $n_{1,i}$ and $t_{1,i}$ for one fictitious party in 20 districts. For each of the dots, one can construct a tomography line. These are displayed in the right graph of Figure 5.



Note: Based on simulated election results in 20 districts for one Party.

Figure 5: *Displayed it the relationship between the aggregated values and the parameters of interest. The left graph displays simulated election results in the first ($n_{1,i}$) and second election ($t_{1,i}$) in relative numbers in 20 districts for one fictitious party. The graph on the right shows one tomography line for each dot in the left graph. Each line represents the range of values in which the true parameter of interest has to lie, in this case this is the voter transition in relative numbers ($\beta_i^{1,1}$ and $\beta_i^{2,1}$).*

The tomography lines on the right reflect what information is left after aggregation. Namely, the linear connection between the unknown parameters $\beta_i^{2,1}$ and $\beta_i^{1,1}$, as well as the deterministic bounds from Duncan and David's Method of Bounds. These can be read of the x and y-axis when projecting the tomography lines on them.

In order to be able to infer information about the unknown parameters, King makes three assumptions, which bridge over the information lost in the aggregation process. As a first assumption King postulates a dependency structure between the unknown parameters $\beta_i^{2,1}$ and $\beta_i^{1,1}$. Namely, that they follow a truncated bivariate normal distribution. This means that the normal distribution will be limited to the unit square $[0,1] \times [0,1]$. Estimating this distribution requires two further assumptions. Firstly, an absence of an aggregation bias is assumed. This assumptions postulates that $\beta_i^{1,1}$ and $\beta_i^{2,1}$ are independent from $n_{1,i}$ in the i different districts. Secondly, King assumes an absence of spatial autocorrelation, meaning that $n_{1,i}$ and $t_{1,i}$ are independent from each other.[98]

In the first step of the model five parameters of the truncated bivariate normal distribution μ_{β_1},μ_{β_2},$\sigma^2_{\beta_1}$,$\sigma^2_{\beta_2}$ and $\sigma_{\beta_1\beta_2}$ are estimated using Maximum Likelihood. The covariance is only estimated once because it is assumed that $\sigma_{\beta_1\beta_2} = \sigma_{\beta_2\beta_1}$. In a second step the parameters of interest $\beta_i^{1,1}$ and $\beta_i^{2,1}$ will be estimated via Bayesian estimation from the before estimated truncated bivariate normal distribution.[99]

Based on Kings ecological inference model King, Rosen and Tanner[100] presented their hierarchical Binomial-Beta Model for the 2x2-case. This model can be seen as a true hierarchical Bayesian model and explicitly states the use of MCMC methods for its estimation.

On the first level of the hierarchical Binomial-Beta Model they postulate a binomial distribution for the second election results $t_{1,i}$. The parameters of the binomial distribution are the probability $\theta_i = n_{1,i}\beta_i^{1,1} + (1-n_{1,i})\beta_i^{2,1}$, corresponding to a weighted sum of the transition probabilities and the first election results, as well as the number of voters N_i as the count. On second level of the hierarchical model the transition probabilities are independently drawn from a beta distribution with the parameters $c_{r,c}$ and $d_{r,c}$. The third level of the model is made up of an exponential distribution from which the parameters $c_{r,c}$ and $d_{r,c}$ are drawn.[101][102]

3.2.2. The Mutinomial-Dirichlet Model (Ecological Inference Model)

The Multinomial-Dirichlet Model by Rosen et al.[103] extends the Binomial-Beta Model to the RxC-case. Similar to the Binomial-Beta model it is a hierarchical model, which relies on MCMC methods for its estimation. In the further course of this thesis, the model will also be referred to as the Ecological Inference (EI) model.

For the introduction of the EI Model the 2x2 example given in Table 3 will be extended to the RxC-case. The notation of this approach is shown in Table 4. Given are R parties in the first and C parties in the second election. The margins $n_{1,i}, ..., 1 - \sum_{r=1}^{R-1} n_{r,i}$ represent the relative election results of party r in the first election. Analogously, the margins $t_{1,i}, ..., 1 - \sum_{c=1}^{C-1} t_{c,i}$ reflect the relative election results of party c in the second election. The unknown inner cells $\beta_i^{r,c}$ constitute the probability of voting party r in the first

[98] Ibid., 4.
[99] Thomas Gschwend, "Ökologische Inferenz", *Methoden der Politikwissenschaft: neuere qualitative und quantitative Analyseverfahren* (2006), 227–37 at 231.
[100] King, Rosen, and Tanner, op. cit.
[101] Ibid. at 70–2.
[102] Klima et al., op. cit. at 34–5.
[103] Ori Rosen et al., "Bayesian and Frequentist Inference for Ecological Inference: The RxC Case", *Statistica Neerlandica*, 55 (2001), 134–56.

and party c in the second election.

Election 1 \ Election 2	Party 1	Party 2	...	Party C	Election Result 1
Party 1	$\beta_i^{1,1}$	$\beta_i^{1,2}$	...	$1 - \sum_{c=1}^{C-1} \beta_1^{1,c}$	$n_{1,i}$
Party 2	$\beta_i^{2,1}$	$\beta_i^{2,2}$	...	$1 - \sum_{c=1}^{C-1} \beta_1^{2,c}$	$n_{2,i}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Party R	$\beta_i^{R,1}$	$\beta_i^{R,2}$	...	$1 - \sum_{c=1}^{C-1} \beta_1^{R,c}$	$1 - \sum_{r=1}^{R-1} n_{r,i}$
Election Result 2	$t_{1,i}$	$t_{2,i}$	...	$1 - \sum_{c=1}^{C-1} t_{c,i}$	

Note: $n_{r,i}$ displays the results of the first election, $t_{c,i}$ represents the results of the second election and $\beta_i^{r,c}$ displays the voter transition. In all three cases the values refer to relative numbers. i represents the district.

Table 4: *Basic notation of a RxC voter transition table with R parties in the first and C parties in the second election according to Rosen et al.. (Rosen et al., op. cit., 137)*

Since all inner rows as well as the margins must add up to 1 in order to fulfill the probability definition, Rosen et al.[104] chose the notations $1 - \sum_{c=1}^{C-1} \beta_1^{r,C}$, $1 - \sum_{r=1}^{R-1} n_{r,i}$ and $1 - \sum_{c=1}^{C-1} t_{c,i}$ for the respective last cells. For the sake of clarity this will be assumed in the further course of the work and the notation $\beta_i^{r,C}$, $n_{R,i}$ and $t_{C,i}$ will be used instead.

On the first level of the model, Rosen et al. assume $T_{1,i}, T_{2,i}, ..., T_{C,i}$ to be the results of the second election in absolute numbers. It is postulated that the results of the second election follow a Multinomial distribution in each district i

$$(T_{1,i}, T_{2,i}, ..., T_{C,i}) \sim Multi(\theta_{1,i}, \theta_{2,i}, ..., \theta_{C,i}, N_i) \tag{32}$$

with

$$\theta_{c,i} = \sum_{r=1}^{R} \beta_i^{r,c} n_{r,i}. \tag{33}$$

Analogous to the Binomial-Beta model it is assumed, that $\theta_{c,i}$ is the weighted sum of the transition probabilities. These are weighted with the relative results of the first election $n_{r,i}$, thus creating a connection between the first and the second election.

[104] Ibid.

On the second level of the model a row wise Dirichlet distribution is assumed for the transition probabilities $\beta_i^{r,c}$. Furthermore, it is assumed that *a priori* every row is independent from each other. Thus it follows that

$$\beta_i^{r,c} \sim Dirichlet(\alpha_{r,1}, \alpha_{r,2}, ..., \alpha_{r,C}). \tag{34}$$

The use of the Dirichlet distribution ensures that the transition probabilities sum up to 1.[105] In the original model $\theta_{c,i}$ and $\beta_i^{r,c}$ are defined to be dependent on a covariate Z_i. Such incorporation of covariates allows the distributions to be more flexible.[106] However, the approach of modeling with covariates is not pursued in this master thesis, which is why a notation without them was chosen.

On the third level of the model Rosen et al. assume an Exponential distribution as hyperpriori for the parameter of the Dirichlet distribution:

$$\alpha_{r,c} \sim Exp(\lambda). \tag{35}$$

The parameters of the Dirichlet distribution $\alpha_{r,c}$ are, in contrast to the district-specific transition probabilities, globally defined. Thus, every district uses the same set of row-wise defined $\alpha_{r,c}$'s.

The two software implementations of the Multinomial-Dirichlet Model, *eiPack*[107] and *eiwild*[108], allow a more flexible modelling of the hyperprior $\alpha_{r,c}$. Lau et al.[109] implemented a Gamma distribution instead of the Exponential distribution assumed by Rosen et al.,

$$\alpha_{r,c} \sim Gamma(\lambda_1, \lambda_2). \tag{36}$$

The eiwild-package allows even further flexibility. Instead of using one distribution for all parameters, it is possible to model cell-specific hyperpriors and thus include available prior knowledge about the cells.

[105] Klima et al., op. cit.
[106] Rosen et al., op. cit. at 136–8.
[107] Lau, Moore, and Kellermann, op. cit.
[108] Schlesinger, op. cit.
[109] Lau, Ryan T. Moore, and Michael Kellermann, op. cit.

According to the Bayes theorem the posterior distribution is proportional to

$$f(\beta_i^{r,c}, \alpha_{r,c} \mid n_{r,i}, T_{c,i}, (\lambda_1, \lambda_2) \text{ or } \lambda) \propto \tag{37}$$

$$\left.\prod_{i=1}^{P}\prod_{c=1}^{C}(\theta_{c,i})^{T_{c,i}} \quad \right] \textbf{Aggregate Data} \tag{38}$$

$$\left.\times \prod_{i=1}^{P}\prod_{r=1}^{R}\frac{1}{B(\alpha)}\prod_{c=1}^{C}(\beta_i^{r,c})^{\alpha^{r,c}-1} \quad \right] \textbf{Prior Distribution} \tag{39}$$

$$\left.\times \prod_{r=1}^{R}\prod_{c=1}^{C}\frac{1}{\lambda_2^{\lambda_1}\Gamma(\lambda_1)}\alpha_{r,c}^{\lambda_1-1}exp(-\alpha_{r,c}\lambda_2) \quad \right] \textbf{Hyperprior Distribution} \tag{40}$$

or

$$\left.\times \prod_{r=1}^{R}\prod_{c=1}^{C}\lambda \exp(-\lambda\alpha_{r,c}) \quad \right] \textbf{Hyperprior Distribution} \tag{41}$$

$$\text{with } B(\alpha) = \frac{\prod_{c'=1}^{C}\Gamma(\alpha_{rc'})}{\Gamma(\Sigma_{c'=1}^{C}\alpha_{r,c'})}. \tag{42}$$

3.3. Hybrid Models

3.3.1. Wakefield's Hybrid Model

The model proposed by Wakefield[110] was one of the first to integrate individual data into an ecological inference model. The initial model for the 2x2 case is solely based on aggregate data and later expanded with individual data. Wakefield's notational approach is shown in Table 5. Applied to the problem of estimating voter transitions, $Y_{1,i}$ and $Y_{2,i}$ represent the unknown number of people who voted for Party 2 in the second election given they voted for Party 1 or Party 2 in the first election. The respective row and column sums are noted with $N_{1,i}$ and $N_{2,i}$ as well as $N_i - T_i$ and T_i and represent the first and second elections results, respectively.

[110] Jon Wakefield, "Ecological inference for 2 x 2 tables", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167/3 (2004), 385–425. doi: 10.1111/j.1467-985x.2004.02046_1.x.

Election 1 \ Election 2	Party 1	Party 2	Election Result 1
Party 1		$Y_{1,i}$	$N_{1,i}$
Party 2		$Y_{2,i}$	$N_{2,i}$
Election Result 2	$N_i - T_i$	T_i	N_i

Note: $N_{r,i}$ displays the results of the first election, T_i represents the results of the second election for the second Party and $Y_{r,i}$ displays the voter transition. N_i displays the number of voters in district i In all three cases the values refer to absolute numbers. i represents the district.

Table 5: *Basic notation of a 2x2 voter transition table with 2 parties in the first and 2 parties in the second election according to Wakefield. (Wakefield, op. cit., 389)*

The transition probabilities $\beta_{1,i}$ and $\beta_{2,i}$ reflect the probability that a voter switches from Party 1 to Party 2 in the second election or stays loyal to Party 2. Wakefield defines these probabilities as follows

$$\beta_{1,i} = Pr(\text{Party 2} \mid \text{Party 1}, i), \tag{43}$$

$$\beta_{2,i} = Pr(\text{Party 2} \mid \text{Party 2}, i). \tag{44}$$

In order to avoid the problem of ecological fallacy described in chapter 3 further assumptions are required to achieve an unambiguous solution. Thus, Wakefield[111] assumes the voter transitions $Y_{r,i}$ to follow an independent binomial distribution

$$Y_{r,i} \mid \beta_{r,i} \sim Bin(N_{ri}, \beta_{ri}), \tag{45}$$

where $r = 0, 1$ is the row indicator and i represents the districts.

Based on this assumption and the assumption, that the row totals $N_{1,i}$ and $N_{2,i}$ are fixed, Wakefield defines the so called *Convolution Likelihood*,

$$L(\beta_{1,i}, \beta_{2,i}) = \sum_{Y_{1,i}=l_i}^{u_i} \binom{N_{1,i}}{Y_{1,i}} \binom{N_{2,i}}{T_i - Y_{1,i}} \beta_{1,i}^{Y_{1,i}} (1-\beta_{1,i})^{N_{1,i}-Y_{1,i}} \beta_{2,i}^{T_i - Y_{1,i}} (1-\beta_{2,i})^{N_{2,i}-T_i+Y_{1,i}} \tag{46}$$

[111] Ibid. at 390.

where

$$l_i = max(0, T_i - N_{2,i}) \tag{47}$$

and (48)

$$u_i = min(N_{1,i}, T_i) \tag{49}$$

define the admissible range for $Y_{1,i}$. This Convolution Likelihood creates the first level of the model.[112] Because the Convolution Likelihood is computationally intensive to calculate, Wakefield adopts a normal approximation of said likelihood on the first level of the model.[113]

An alternative approach discussed by Wakefield would be the assumption of

$$T_i \mid \beta_{1,i}, \beta_{2,i} \sim Bin(N_i, q_i) \tag{50}$$

where

$$q_i = \beta_{1,i}\frac{N_{1,i}}{N_i} + \beta_{2,i}\frac{N_{2,i}}{N_i}. \tag{51}$$

This assumption corresponds to the Binomial-Beta Model discussed in chapter 3.2.1. However, Wakefield advised against this assumption as it would lead to a higher variance of the sampled values.[114] Therefore, in the following only the case with two independent binomial distributions is discussed.

On the second level Wakefield proposes the use of either a (truncated) normal distribution or a beta distribution as a prior distribution for the probabilities $\beta_{1,i}$ and $\beta_{2,i}$. In the case of the beta distribution, independent exponential hyperpriors $Exp(\lambda_i)$ are assumed.[115]

Based on this model Wakefield now integrates individual data sourced from a survey into the estimation process. Table 6 displays the new notation and the underlying logic of the new approach, which has been adapted to the context of estimating voter transitions. $Z_{1,i}$ and $Z_{2,i}$ reflect the available individual data for voters in district i, who voted for Party 2 in the second election and for Party 1 or 2 in the first election, respectively. The column and row totals $M_{1,i}$ and $M_{2,i}$ as well $Z_{r,i}$ are known and subtracted from the available aggregate data. Consequently, the model only uses aggregate data, which cannot be explained by

[112] Ibid. at 390.
[113] Ibid. at 401–2.
[114] Ibid. at 391.
[115] Ibid. at 405–6.

the collected individual data.[116]

Election 1 \ Election 2		Party 1	Party 2	Row Total
(A)	Individual Data			
	Party 1		$Z_{1,i}$	$M_{1,i}$
	Party 2		$Z_{2,i}$	$M_{2,i}$
	Column Total	$M_i - Z_i$	Z_i	M_i
(B)	Aggregate Data			
	Party 1			$N_{1,i} - M_{1,i}$
	Party 2			$N_{2,i} - M_{2,i}$
	Column Total	$N_i - M_i - (T_i - Z_i)$	$T_i - Z_i$	$N_i - M_i$

Note: $Z_{r,i}$ represents the available individual data and $M_{r,i}$ the row sums of the individual data. $N_{r,i}$ and $T_{c,i}$ display the election results of the first and second election respectively. i represents the district.

Table 6: *Basic notation of a 2x2 voting transition table with (A) individual data and (B) aggregate data for 2 parties according to Wakefield. (Wakefield, op. cit., 418)*

Wakefield assumes the individual data to be distributed according to two row wise Binomial distributions

$$Z_{r,i} \sim Bin(M_{r,i}, \beta_{r,i}), \tag{52}$$

where $\beta_{r,i}$ still represents the transition probabilities defined in equation 43 and 44. These are then combined with the likelihood for the aggregate data in order to create an overall likelihood for the given data:

$$L(\beta_{1,i}, \beta_{2,i}) = p(Z_{1,i} \mid \beta_{1,i})p(Z_{2,1} \mid \beta_{2,i})p(T_i - Z_i \mid \beta_{1,i}, \beta_{2,i}), \tag{53}$$

where the first two terms represent a binomial distribution and the third term the convolution likelihood given in equation 46. The only difference being that the available individual data has been subtracted from the aggregate data.[117]

The first level of this hybrid hierarchical model is made up of combined binomial and normal approximations of the Convolution Likelihoods. While on the second level a beta distribution is assumed for the probabilities $\beta_{r,i}$.[118]

First analyses of this approach conducted by Wakefield show that the introduction of individual data can improve the estimation of the inner cells and that already a limited amount of individual data can have a

[116] Ibid. at 418.
[117] Ibid. at 418.
[118] Ibid. at 418.

positive effect on the estimation quality.[119]

3.3.2. Hybrid Multinomial-Dirichlet Model

Wakefield[120] proposed the first hybrid model for the estimation of ecological inference problems. However, their proposed approach was constrained to the 2x2 case and elections rarely only have two parties contesting in them. Thus, Klima et al.[121] took Rosen's Multinomial-Dirichlet Model and expanded it with Wakefields's approach for the integration of individual data, forming the *hybrid Multinomial-Dirichlet Model*.

Election 1 \ Election 2		Party 1	Party 2	...	Party C	Row Total
(A)	Individual Data					
	Party 1	$Z_i^{1,1}$	$Z_i^{1,2}$	...	$Z_i^{1,C}$	$M_{1,i}$
	Party 2	$Z_i^{2,1}$	$Z_i^{2,2}$	...	$Z_i^{2,C}$	$M_{2,i}$
	⋮	⋮	⋮	⋱	⋮	⋮
	Party R	$Z_i^{R,1}$	$Z_i^{R,2}$	...	$Z_i^{R,C}$	$M_{R,i}$
	Column Total	$Z_{1,i}$	$Z_{2,i}$	...	$Z_{C,i}$	M_i
(B)	Aggregate Data					
	Party 1					$N_{1,i} - M_{1,i}$
	Party 2					$N_{2,i} - M_{2,i}$
	⋮					⋮
	Party R					$N_{R,i} - M_{R,i}$
	Column Total	$T_{1,i} - Z_{1,i}$	$T_{2,i} - Z_{2,i}$	...	$T_{C,i} - Z_{C,i}$	$N_i - M_i$

Note: $Z_i^{r,c}$ represents the available individual data and $Z_{r,i}$ the row sums of the individual data. $N_{r,i}$ and $T_{c,i}$ display the election results of the first and second election respectively. i represents the district.

Table 7: *Basic notation of a RxC voting transition table with (A) individual data and (B) aggregate data for R parties in the first and C parties in the second election according to Klima et al.. (Klima et al., 2017, 7)*

Table 7 introduces a possible data scenario, which could be used in the Hybrid Multinomial-Dirichlet Model. The notation is analogous to the Multinomial-Dirichlet Model, but has been supplemented with individual data according to Wakefield's approach. Corresponding to Wakefield's Model, data from two sources may be available: individual data and aggregate data. $Z_i^{r,c}$ represents the available individual data of voters from district i, who voted for Party R in the first election and for Party C in the second one. In extension of Wakefield's assumed row wise binomial distribution Klima et al. assume that the rows of the

[119] Ibid. at 418–21.
[120] Ibid.
[121] Klima et al., "Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions".

individual data in district i follow a Multinomial distribution with the parameters $M_{r,i}$ and the transition probabilities $\beta_i^{r,1}, ..., \beta_i^{r,C}$:

$$(Z_i^{r1}, ..., Z_i^{rC}) \sim Multi(M_{r,i}, \beta_i^{r1}, ..., \beta_i^{rC}) \tag{54}$$

where

$$\sum_c^C \beta_i^{rc} = 1. \tag{55}$$

Following the assumptions made by Rosen et al., a Multinomial distribution for the aggregate data, which is assumed to be available in every district, is presumed:

$$(T_{1,i} - Z_{1,i}, ..., T_{C,i} - Z_{C,i}) \sim Multi(N_i - M_i, \Theta_{1,i}, ..., \Theta_{C,i}) \tag{56}$$

with

$$\Theta_{c,i} = \sum_{r=1}^R \beta_i^{rc} X_{r,i} \tag{57}$$

and

$$X_{r,i} = \frac{N_{r,i} - M_{r,i}}{N_i - M_i}. \tag{58}$$

Parameters of the Multinomial distribution are $N_i - M_i$ as the number of trials and $\Theta_{c,i}$ as the event probabilities. $\Theta_{c,i}$ is defined as the weighted sum of the transition probabilities $\beta_i^{r,c}$ multiplied with the row fractions $X_{r,i}$. As proposed by Wakefield, this model only utilizes the available aggregate data after subtracting the individual data, $T_{r,i} - Z_{r,i}$. If no individual data is available only aggregate data is used in the estimation. On the other hand, if a full census is carried out in one district, no aggregate data would be used in the estimation.

The second level of the hybrid Multinomial-Dirichlet Model is made up of the Dirichlet distributed transition probabilities $\beta_i^{r,c}$:

$$(\beta_i^{r,1}, ..., \beta_i^{r,C}) \sim Dirichlet(\alpha_{r1}, ..., \alpha_{rC}). \tag{59}$$

The above equation reflects an important assumption of the model. Namely, that all transition probabilities result from the same distribution. This is reflected by the missing district specific index in the parameters α_{rc} of the Dirichlet distribution.

The third level of the hybrid Multinomial-Dirichlet Model is constructed with either a Gamma or an Exponential distribution from which the parameters of the Dirichlet distribution are drawn:

$$\alpha_{rc} \sim Gamma(\lambda_1^{r,c}, \lambda_2^{r,c}) \tag{60}$$

or

$$\alpha_{rc} \sim Exp(\lambda). \tag{61}$$

Besides the choice of distribution for the hyperprior parameters, it is also possible to define a global hyperprior instead of a cell-specific one.

The posterior distribution of the model is defined as follows:

$$f(\beta_i^{rc}, \alpha_{rc} | X_{r,i}, T_{c,i}, Z_i^{r,c}, Z_{c,i}, \lambda) \sim \tag{62}$$

$$\prod_{i=1}^{P} \prod_{r=1}^{R} \prod_{c=1}^{C} (\beta_i^{rc})^{Z_i^{rc}} \quad \Big] \textbf{Individual Data} \tag{63}$$

$$\times \prod_{i=1}^{P} \prod_{c=1}^{C} (\Theta_{c,i}^{T_{c,i} - Z_{c,i}}) \quad \Big] \textbf{Aggregate Data} \tag{64}$$

$$\times\, p(\beta_i^{rc} | \alpha_{rc}) \times p(\alpha_{rc} | \lambda) \quad \Big] \textbf{Prior-Distributions} \tag{65}$$

As the estimation of the parameters of the hybrid Multinomial-Dirichlet Model is only possible via complex high-dimensional integration, this model relies on MCMC methods such as the Metropolis-within-Gibbs-Sampler for the estimation of voter transitions.

3.4. Literature Review: Estimation of Uncertainty with Ecological Inference Models

While the calculation of point estimates has attracted much attention in the literature on ecological inference, the calculation of interval estimates remains a little discussed topic in this field. However, the correct quantification of the uncertainty of the estimated quantities by interval estimates is just as important as the evaluation of the estimation quality. In most cases, uncertainty is quantified by frequentist confidence intervals or Bayesian credible intervals. Although there is no lack of possibilities to calculate such interval estimates within the various frameworks of the ecological inference, it has been shown in various papers that the quality of the calculated intervals, i.e. the coverage of the true values by the interval estimates, is in most cases poor. Thus, the quality of the calculated interval estimators must be considered with care. It is therefore surprising that detailed discussions about the reasons of and possible solutions for the poor quality are scarce in the available literature. The goal of the following chapter is to summarize the previous contributions regarding the calculation of interval estimates in the ecological inference. The first part of the chapter summarizes the different models and ways to calculate interval estimates. This will be followed by a discussion of the quality of the resulting interval estimates. Lastly, the existing discussion about possible reasons for the poor quality of the calculated credible intervals and possible solutions will be summarized.

First attempts to quantify uncertainty in ecological inference models can already be found in the fundamental beginnings of these models. Duncan and Davis Method of Bounds [122] results in 100% confidence intervals, which always cover the true value. However, the interval estimates are too wide and imprecise to be of practical use for quantifying uncertainty. Goodman's Ecological Regression [123] also allows for the construction of interval estimates. However, the problem arises that it is possible to observe point estimates outside of the unit interval. If these estimates are corrected, it is no longer clear how the confidence intervals are to be calculated.[124] For the case that the interval estimates are not corrected Plescia et al.[125] carried out a comprehensive evaluation of the quality of the confidence intervals resulting from Goodman's Ecological Inference. On the basis of election data from New Zealand and Scotland, in which the true transition probabilities are known, they check the coverage of the true values by the 95% confidence or credible intervals of the approaches by Goodman[126], Greiner and Quinn[127] as well as Rosen et al.[128]. The results for the latter two models will be discussed later in this chapter. Plescia et al. were able to show that Goodman's 95% confidence intervals only cover the true value in about 30% of cases. Thomson's Probit/Logit Model[129] offers another approach for the estimation of voter transitions. The initial model does not allow for the calculation of interval estimates. This feature was later added in an extension of the model by Park[130]. However, Park also points out that the interval estimates calculated according to his extension

[122] Please refer to chapter 3.1.
[123] Please refer to chapter 3.1.
[124] Klima, op. cit., 103.
[125] Carolina Plescia and Lorenzo de Sio, "An evaluation of the performance and suitability of R × C methods for ecological inference with known true values", *Quality & quantity*, 52/2 (2018), 669–83. doi: 10.1007/s11135-017-0481-z.
[126] Goodman, "Ecological Regressions and Behavior of Individuals"; id., "Some Alternatives to Ecological Correlation".
[127] D. James Greiner and Kevin M. Quinn, "R × C ecological inference: Bounds, correlations, flexibility and transparency of assumptions", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172/1 (2009), 67–81. doi: 10.1111/j.1467-985X.2008.00551.x.
[128] Rosen et al., op. cit.
[129] Søren R. Thomsen, *Danish elections 1920 - 1979: A logit approach to ecolog. analysis and inference: Zugl.: Åarhus, Univ., Diss., 1987* (1. udg., 1. opl, Åarhus: Politica, 1987).
[130] Won-ho Park, "Ecological Inference and Aggregate Analysis of Elections", Dissertation (Michigan: The University of Michigan, 2008), 63–5.

have a low coverage rate of the true values and thus, do not offer a solution to the problem of the poor quality of interval estimates.

In theory, the calculation of interval estimates for models utilizing Bayesian estimation methods such as MCMC is relatively simple, e.g. using the quantiles of the resulting Markov chains. However, even these models do not guarantee sufficient coverage of the true values by the resulting credible intervals. King's ecological inference model [131] is one of these models which use Bayesian estimation methods. Yet, in a commentary on King's solution Freedman et al.[132] show, with data from the 1980 Los Angeles Census including known individual data, that the 80% credible intervals according to King's model only cover the true values in about 20% of the cases. As extensions of King's ecological inference model, the Binomial-Beta model [133] and the Multinomial-Dirichlet Model [134] also offers the possibility to calculate interval estimates. King et al.[135] and Rosen et al.[136] make use of this possibility without discussing the quality of the resulting credible intervals. For the Multinomial-Dirichlet Model, this discussion was provided by Klima[137] as well as Plescia et al.. Based on simulated voter transition Klima was able to show that the coverage rates of the true values by the credible intervals are far from the targeted 80%. In the best case, Klima determines an average coverage rate of 60%, while in the worst case the average coverage rate is only 30%. These results are consistent with those of Plescia et al., who estimate voter transitions in New Zealand with the Multinomial-Dirichlet Model and only achieve coverage rates between 30% and 40%. Greiner and Quinn[138] offer an additional approach to drawing inferences about individual behavior. Their first paper on the model also briefly discusses the coverage rates of the 95% credible intervals. Based on a simulation study of the electoral behavior of different ethnic groups, they achieve coverage rates between 89% and 98%. Due to the stochasticity in the model, a certain variance around 95% is to be expected and the results can be regarded as very good. However, Plescia et al. were not able to replicate these good results in their performance study. They observed substantially worse coverage rates between 10% and 40%.

Since hybrid Models also use MCMC methods for their estimation, the intervals are calculated analogously to the previous models. Even though the quality of the credible intervals of the hybrid Multinomial-Dirichlet Model was not explicitly investigated, according to Klima[139], the results from the investigation of

[131] Please refer to chapter 3.2.1.
[132] D. A. Freedman et al., "A Solution to the Ecological Inference Problem", *Journal of the American Statistical Association*, 93/444 (1998), 1518. doi: 10.2307/2670067.
[133] Please refer to chapter 3.2.1.
[134] Please refer to chapter 3.2.2.
[135] King, Rosen, and Tanner, op. cit.
[136] Rosen et al., op. cit.
[137] Klima, op. cit.
[138] James Greiner and Quinn, op. cit.
[139] Klima, op. cit., 104.

the Multinomial-Dirichlet Model can mostly be transferred to the hybrid version due to the similar structure of the two models. In the hybrid version of their model, which Greiner and Quinn presented in 2010[140], they show that the coverage rates of the 95 % credible intervals deteriorate substantially compared to their original model. Again, based on a simulation study of the voting behavior of different ethnic groups, coverage rates between 68% and 85% were achieved.

While the poor quality of credible intervals resulting from ecological inference models has found some recognition in the available literature, a wide discussion about the reasons and possible solutions has not taken place. Only Plescia et al.[141] and Klima[142] further investigate the reasons for the poor performance and try to provide solutions. Plescia et al. define three so-called *predictors of unreliability*, which have a negative impact on the coverage rates in their performance study. Their first predictor is the size of the contingency tables. The larger the contingency table, i.e. the more rows and columns, the worse are the coverage rates of the confidence or credible intervals. However, this correlation is statistically only significant for the rows of a reduced form of the contingency table Plescia et al. used in their performance study. The reduced form refers to a contingency table which contains only about 5% of the total votes at the district level. A second predictor is the number of polling stations, which is positively correlated with a higher coverage rate. The final predictor discussed by Plescia et al. is the variance between polling stations on a district level. Here they found a negative correlation between the variance and the coverage rate.

In his dissertation Klima concludes that the problem lies in a non-bias-free estimation. Some cells are strongly underestimated by the Multinomial-Dirichlet Model, which in turn leads to an overestimation of the remaining cells. Since this bias is not considered in the chains of the model, a poor coverage rate is the result. In the following, Klima tries to correct the bias and tests four different approaches. The first method uses the idea of empirical bootstrapping and tries to improve the credible intervals by combining several Markov Chains. The three following methods are loosely based on the idea of parametric bootstrapping. Klima samples new election results for the second election and new voter transitions with them. The variance between the new models is then used to quantify a bias component, which is used to increase the variance of the Markov Chains. At first glance, the coverage rates of the chains improve substantially. However, when considering the coverage rates per cell, Klima notices, that such high coverage rates cannot be maintained at this level. Therefore, Klima concludes that while the approaches show some potential, they do not represent a solution for the problem of the poor coverage rates.

[140] D. James Greiner and Kevin M. Quinn, "Exit polling and racial bloc voting: Combining individual-level and R $\times$ C ecological data", *The Annals of Applied Statistics*, 4/4 (2010), 1774–96. doi: 10.1214/10-AOAS353.

[141] Plescia and Sio, op. cit.

[142] Klima, op. cit.

Two main conclusions can be drawn from the literature review. Firstly, all ecological inference models have a problem with the exact quantification of uncertainty. While Duncan and Davis Method of Bounds produces too conservative interval estimates to be of practical use, the remaining approaches suffer from too low coverage rates. Secondly, only very few explanations and solution approaches are discussed in the literature. This is particularly surprising since the exact quantification of uncertainty is an immensely important aspect of statistical inference.

4. Simulation Study

When trying to evaluate the results of an ecological inference one encounters a fundamental problem. Namely, the fact that in most cases the individual data of interest is unknown. Thus, no comparison between the estimation and the truth can be drawn, which makes it difficult to evaluate the quality of the estimation. The only available option is to check the results for their plausibility. However, such an approach is very objective and does not provide a meaningful evaluation of the results. Therefore, when evaluating the quality of ecological inferences, one falls back on data scenarios in which the individual data, i.e. the truth, is known. These can be simulated data scenarios, e.g. Klima[143] and Greiner and Quinn[144], or real data scenarios, e.g. Plescia et al.[145]. This thesis follows the approach of simulating a data scenario.

As discussed in the literature review, previous analysis has revealed the insufficient quality of interval estimates. In most cases, the coverage of the true value by the interval estimate was insufficient. While this phenomenon has been shown in many cases, it is challenging to identify the causes of the poor coverage rates due to the complexity of the data scenarios used. Hence, the first goal of this simulation study is to understand how expected variations and complexity in the data affect the quality of the interval estimates. Therefore, six different variables are used to model variations in the data and complexity of it. The six different variables are *number of parties*, *number of districts*, *average population size*, *vote shares*, *transition probabilities* and the *sampling scheme for individual data*, which are also displayed in Table 8. The goal of identifying their impact on the quality of the credible intervals is achieved by assuming a basic scenario with sufficient coverage rates and in a second step, departing from that scenario by gradually adding different variations and complexity to the data. The underlying hypothesis is that while the coverage rate in simple scenarios is sufficient, the quality of the interval estimates deteriorates with added complexity to the data.

[143] Ibid.
[144] Greiner and Quinn, op. cit.
[145] Plescia and Sio, op. cit.

Overview of the Variables of the Simulation Study	
Number of Parties	Number of Districts
Average Population Size	Vote Shares
Transition Probabilities	Sampling Scheme for individual data

Table 8: *Overview of the six variables varied in the process of simulating data for the simulation study.*

It is important to mention that the main focus of the simulation study does not lie in the complete and most realistic representation of each real use case, but instead on understanding the effect small variations in the data have on the quality of the credible intervals.

The models used to estimate voter transitions rely on assumptions regarding the distribution of voter transitions, election results, and individual data. If the simulation study violates these assumptions, it can affect the quality of the resulting credible intervals. Therefore, in addition to the impact of variations and complexity in the data, this thesis investigates the impact of three different data simulation approaches on the quality of interval estimates. The approaches vary in the way they simulate voter transitions and second election results. The first and second approach, hereafter called *Approach 1* and *Approach 2*, regard the voter transition $Y_i^{r,1}, ..., Y_i^{r,C}$ as the stochastic component and the second election results $T_{1,i}, ..., T_{C,i}$ as deterministic. In Approach 1 transition probabilities $\beta_i^{r,1}, ..., \beta_i^{r,C}$ are drawn from a row-wise Dirichlet distribution. The voter transitions are then calculated by multiplying the transition probabilities with the respective first election results $N_{r,i}$. Approach 2 follows the assumptions of the (hybrid) Multinomial-Dirichlet Model after which the voter transitions follow a row-wise Multinomial distribution. In both cases, the second election results are calculated by summing over the columns of the voter transition matrix. However, the (hybrid) Multinomial-Dirichlet Model assumes the second election results to follow a Multinomial distribution. Because in the first two elections they are calculated by summing over the simulated voter transitions, it could lead to a situation in which the second election results follow a convolutional distribution rather than a simple one. This would constitute a violation of the model assumptions. Therefore, the third approach, hereafter called *Approach 3*, reverses the point of view and regards the second election results as the stochastic component and as such, they are drawn from a Multinomial distribution. As the deterministic component, the voter transitions are then calculated based on the results of the first and second election.

The chapter is structured as follows. Chapter 4.1 - chapter 4.4 present and motivate the various variables of the simulation study displayed in Table 8. Chapter 4.5 discusses the technical aspects of the simulation study, while in chapter 4.5.1 the variations in the data simulation are presented.Lastly, in chapter 4.6 the different scenarios are introduced.

4.1. The Choice of Number of Parties and Districts as well as the Average Population Size

The starting point for the simulation study is the choice of the number of parties and districts as well as the average population size per district. For the sake of simplicity, it is assumed that the population consists only of eligible voters. Since the number of relevant parties can vary greatly between elections in different areas, two variations model different levels of complexity in the party landscapes. In the first variation seven parties plus non-voters are assumed, while the second only assumes four parties plus non-voters.

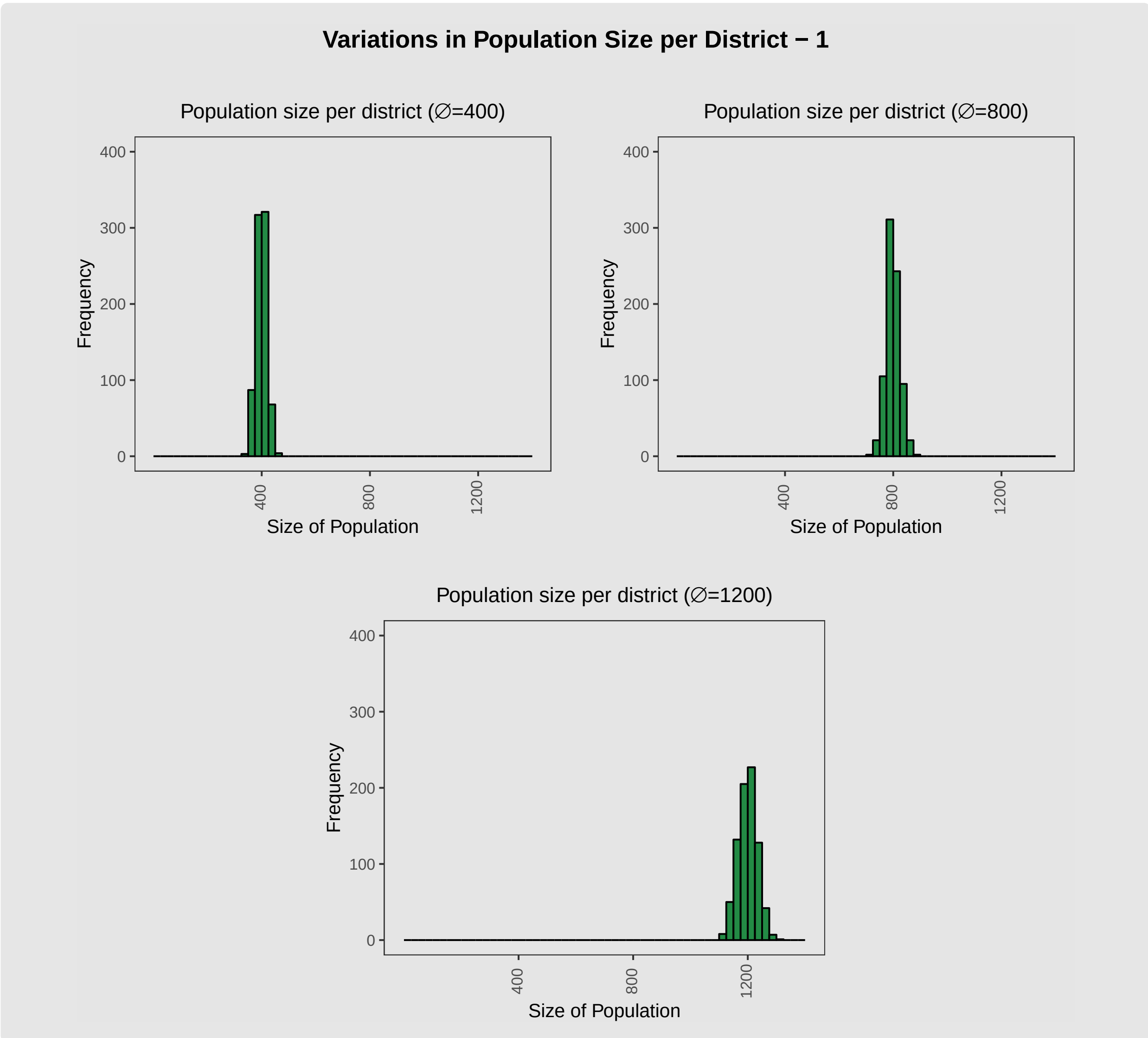


Figure 6: *Example of the distribution of population size per district over 800 municipalities from one simulation in the case of an average population size of 400 (top left), 800 (top right) and 1200 (bottom).*

Similarly, the number of districts can vary considerably depending on different administrative levels, e.g.

city, state or country level. Therefore, variations of 200, 400, 800, 1600 and 2000 districts are assumed. As in the case of the number of districts, the average population per district differs depending on the use case. While an application in a city would suggest a high average population per district, an application on a federal or state level suggests a lower average population due to less densely populated rural areas. Therefore, average populations of 400, 800 and 1200 people per district were simulated. An additional variation models the existence of structural differences between the average population sizes. Structural differences refer to a situation in which the estimation of voter transitions takes place in a region where cities and municipalities with substantially different average population sizes are present. To cover such a situation, two groups of districts with different average population sizes are modeled.

Examples for the first three variations are displayed in Figure 6. The graph on the top left shows the distribution with an average population of 400 voters per district. The number of voters varies between about 340 and 460 per district, while the majority has around 400. In the case of an average population of 800 voters per district, which is shown in the graph on the top right, the population ranges between about 720 and 890. As expected, the majority is centered around 800. The lower graph displays the situation with an average population of 1200 per district. In this case, the population fluctuates between about 1100 and 1300. It also applies that a majority of the districts has a population, which corresponds approximately to the average.

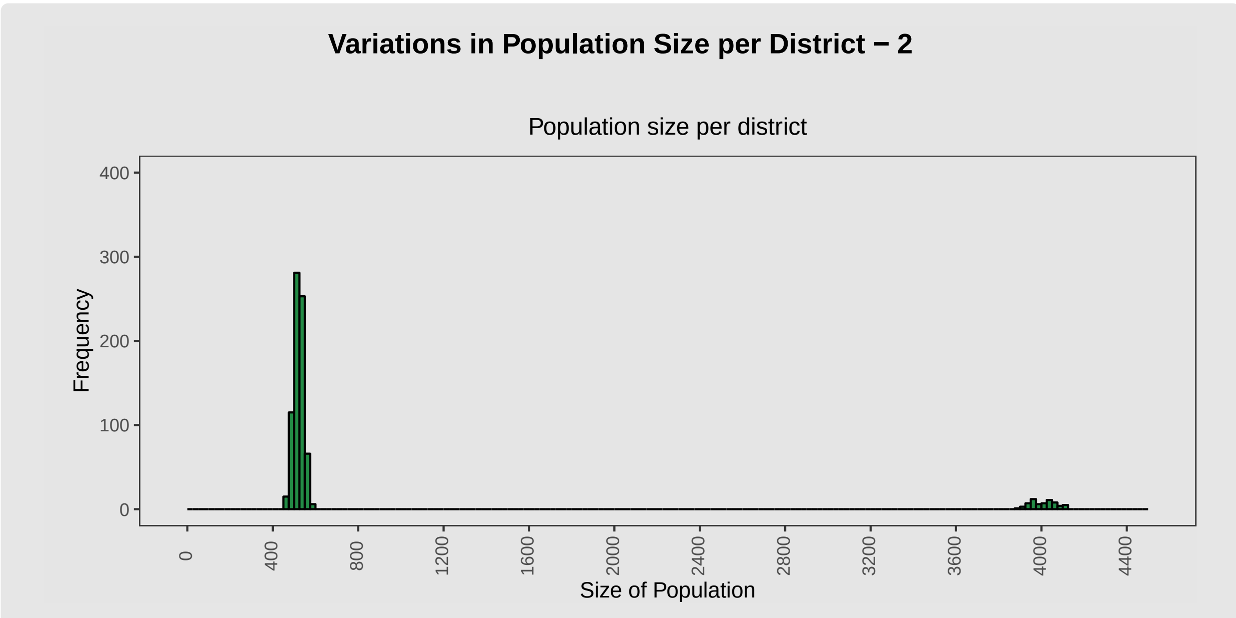


Figure 7: *Example of the distribution of population size per district over 800 municipalities from one simulation in case of structural differences between the municipalities.*

Figure 7 displays the last variation, which models structural differences in population sizes. Here the population size per district varies between approximately 450 and 4100 inhabitants. This broad range arises due to the assumption of structural difference. As Figure 7 shows, the districts are grouped into two different groups. In the first group, the number of voters per district fluctuates between approximately 450 and 600, while the number of voters per district in the second group varies between approximately 3880 and 4120.

4.2. Vote Shares in the First Election

Four different variations are assumed for the results of the first election. The simplest case consists of fixed vote shares, which means that each party and the non-voters have equal vote shares across all districts. The second case adds variance while maintaining overall averages similar to those assumed in the case of fixed vote shares. The introduction of these two cases, although not being realistic, is justified by the objective of the simulation study. It creates the possibility to test the influence of growing complexity in the election results of the first election on the quality of the interval estimates. The third and fourth case model more realistic vote shares for the non-voters and the presence of one or two major parties with larger vote shares compared to the rest.

Figure 8 shows the distribution and overall averages of the first election results in all four cases. The data is an example taken from one simulation with 800 districts and seven parties plus non-voters. The graph at the top left shows the first case with fixed vote shares. As depicted in the graph, the average vote share of each party and the non-voters is exactly 12.5%. The second graph at the top right displays how the introduced variance between the districts affects the distribution of the vote shares. They vary between approximately 0% and 40% and in all cases upward outliers are present. While the fluctuations between the districts are apparent, the averages across the districts remain approximately at the level assumed in the first case.

The distribution of vote shares in the more realistic scenarios are shown in the graphs on the bottom left (two major parties) and bottom right (one major party). In the case of two major parties, the vote shares of these two parties and the non-voters vary between about 15% and 35%, while the shares of the rest fluctuate between approximately 0% and 10%. The major parties and the non-voters have a relatively high average vote share of about 25%. In contrast to that, the other parties have an average vote share of only about 5%. There are isolated outliers upwards. In the case one major party is assumed, the proportion of votes of this party fluctuates between about 20% and 50%, while the proportion of non-voters remains at a similar level compared to the previous case. The major party has on average 35% of the votes and the non-voters about 25%. The shares of the remaining parties vary between about 0% and 15% with an average of about 6.5%. In this case, upward outliers are also present.

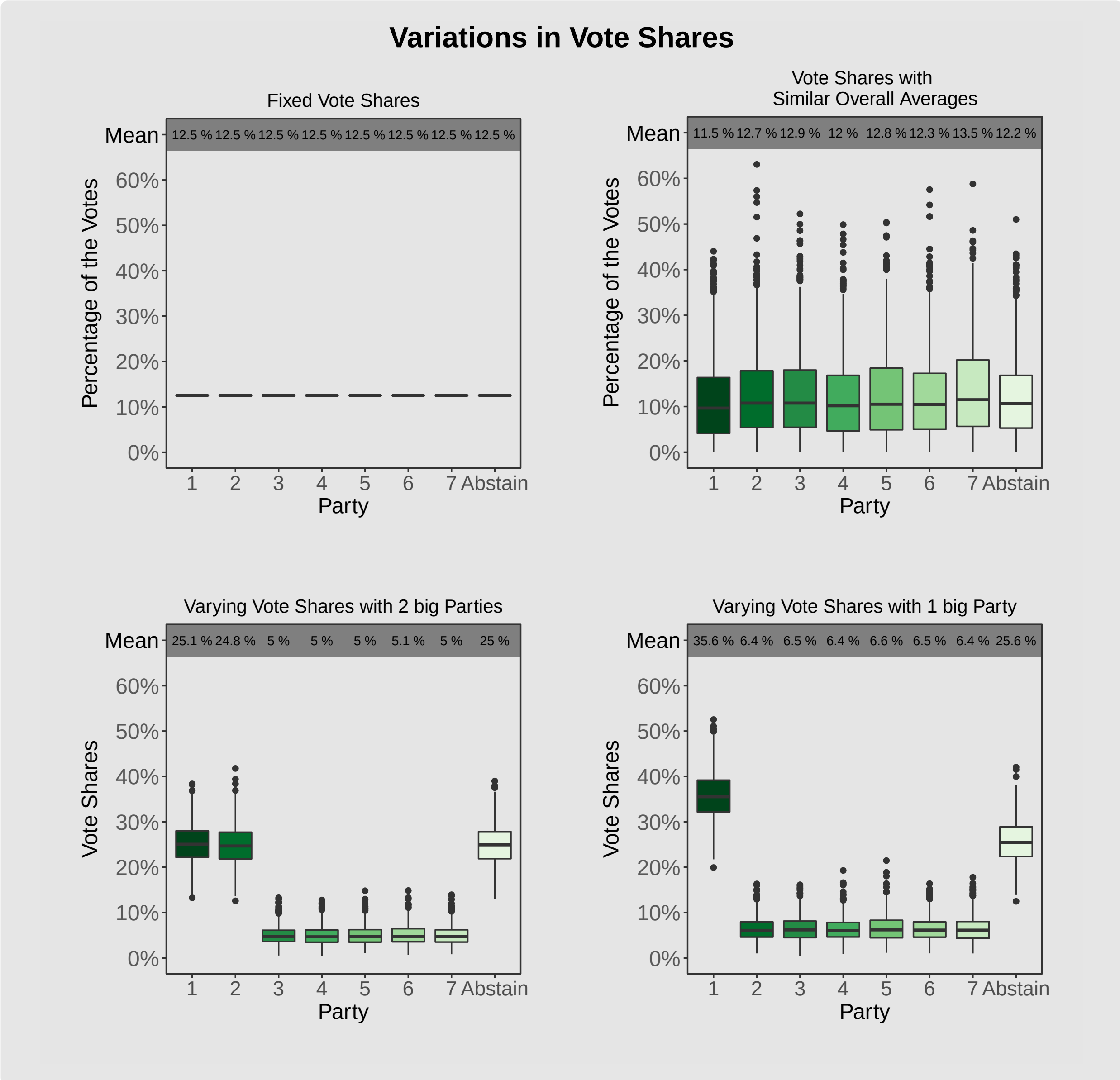


Figure 8: *Example of the distribution of vote shares in the first election of seven parties and non-voters over 800 municipalities in the case of identical vote shares (upper left), varying vote shares and similar overall averages (upper right), two major parties (lower left) and one major party (lower right) from one simulation. In the grey bar on top of the graphs the averages of the distributions over the 800 municipalities for each party and the non-voters are displayed.*

4.3. Transition Probabilities

The transition probabilities are modeled in five different ways. The first two cases resemble those of the simulation of the first election results described in the previous chapter. Here the first case assumes identical transition probabilities and the second case varying transition probabilities with a similar overall average.

In order to add more complexity, the next two cases assume low and high loyalty rates. Loyalty rates are also transition probabilities, namely those giving the probability that a voter stays loyal to one party in both elections. Although all of these do not seem very likely to occur in reality, they serve the purpose of investigating the impact of added complexity in the data on the quality of credible intervals. In order to understand the impact of complex transition probabilities, a fifth case models large differences between the expected transition probabilities. These have been constructed in a way that they could resemble plausible transition probabilities observed in a real election.

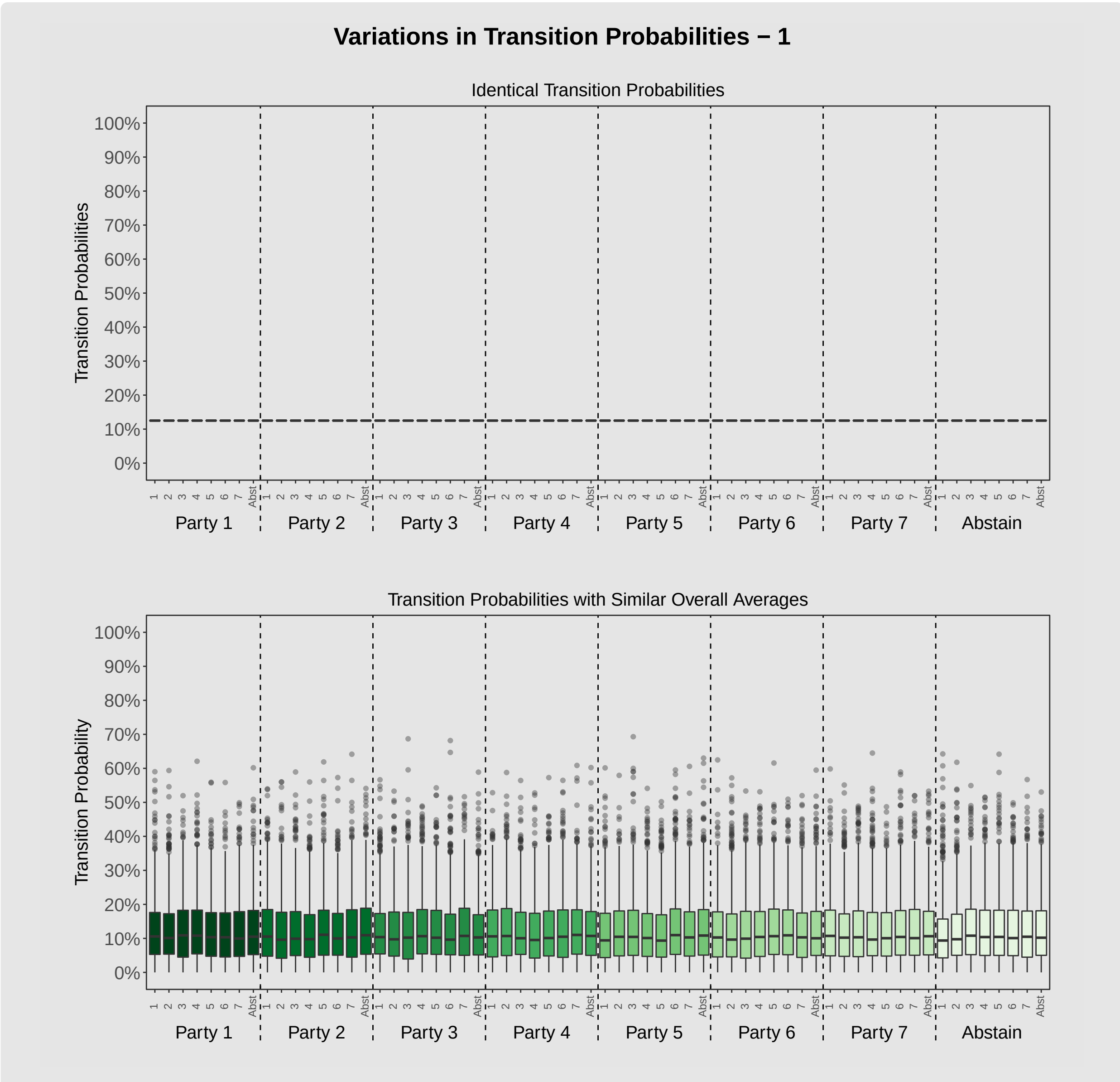


Figure 9: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of identical loyalty and transition probabilities (upper graph) and varying loyalty and transition probabilities with similar overall averages (lower graph) from one simulation.*

Figure 9 gives an impression of the distribution of transition probabilities in the first two cases. The plotted data comes from simulations with 800 districts and seven parties plus non-voters. Transition probabilities assumed to be identical for all parties and the non-voters are displayed in the upper graph. Hence, the transition probabilities for all parties across all districts are exactly 12.5%. The second case assumes variations between the districts. The lower graph from Figure 9 depicts the transition probabilities simulated under this assumption. In this case, the transition probabilities vary between about 0% and 35 %, while the overall averages are similar to those in the first case. For all transitions, upward outliers can be observed.

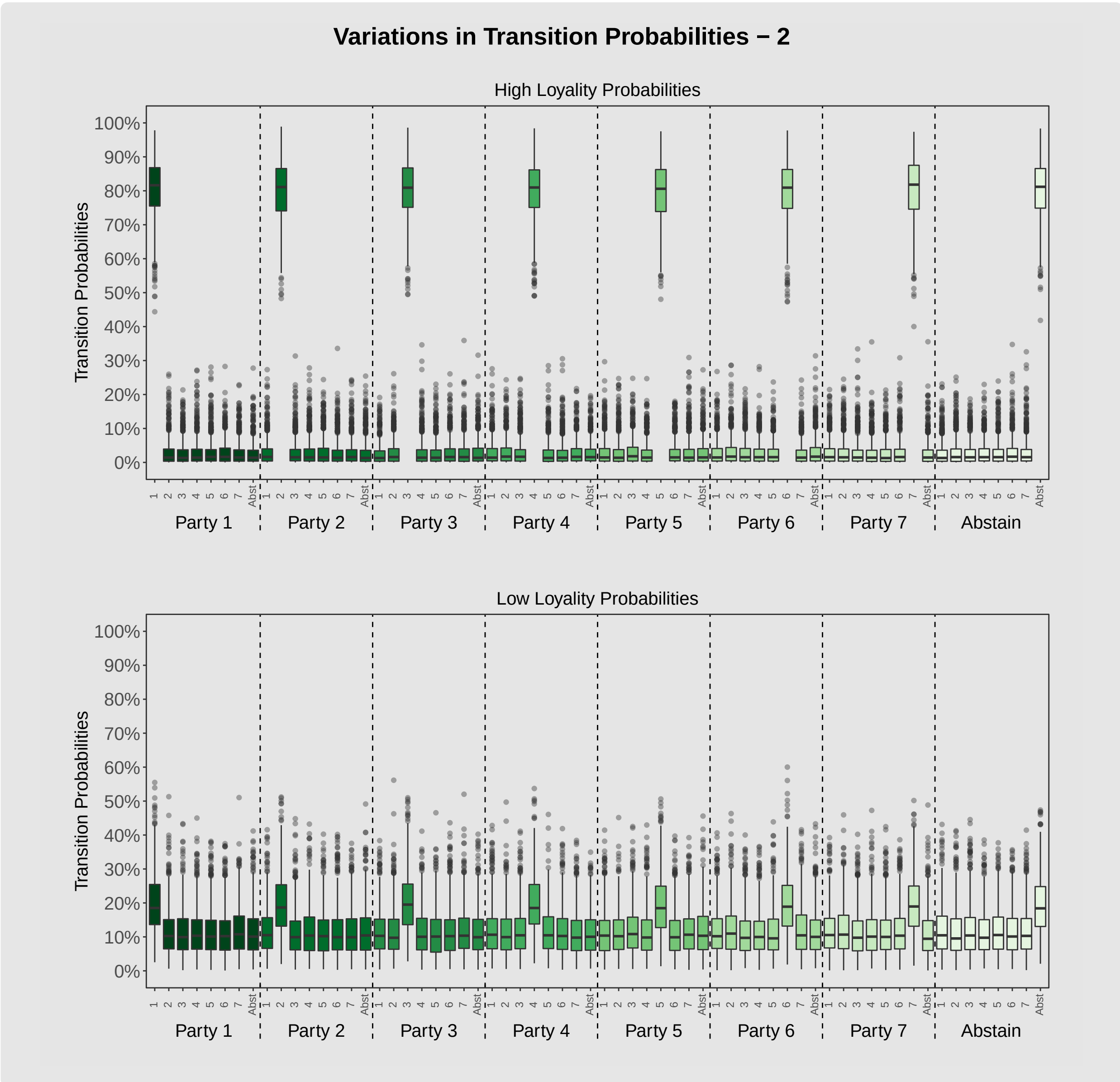


Figure 10: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of high loyalty probabilities (upper graph) and low loyalty probabilities (lower graph) from one simulation.*

The upper graph of Figure 10 shows the case of high loyalty rates. Here the loyalty rates vary between about 60% and 95% with some outliers downwards. The remaining transition probabilities fluctuate between 0% and 10% with some upward outliers. On average the loyalty rates are at about 80%, while the other transition probabilities reach an average of about 3%. In the case of low loyalty rates, shown in the lower graph of Figure 10, the loyalty rates fluctuate between approximately 5% and 40%, while the other transition probabilities fluctuate between 0% and 30%. The loyalty rates have an overall average of about 20% and the remaining transition probabilities of about 11%. In both cases, outliers upwards can be observed.

An example of transition probabilities in case large differences between them are assumed, is given in Figure 11. It is evident that the loyalty rates show a high difference among themselves. Accordingly, the average of the loyalty rates fluctuates between approximately 45 % and 92 %, while it was almost identical in all the previous cases. The loyalty rates are ranging between 20 % and up to 100%. Furthermore, isolated outliers can be observed both upwards and downwards. A substantially higher variation can also be observed in the other transition probabilities. This variation results from the restriction that the transitions probabilities from one Party to the others need to add up to one[146]. Thus, lower loyalty rates lead to higher transition probabilities and vice versa. Hence, the transition probabilities apart from the loyalty rates vary between approximately 0% and up to about 40%. Here too, upwards outliers can be observed.

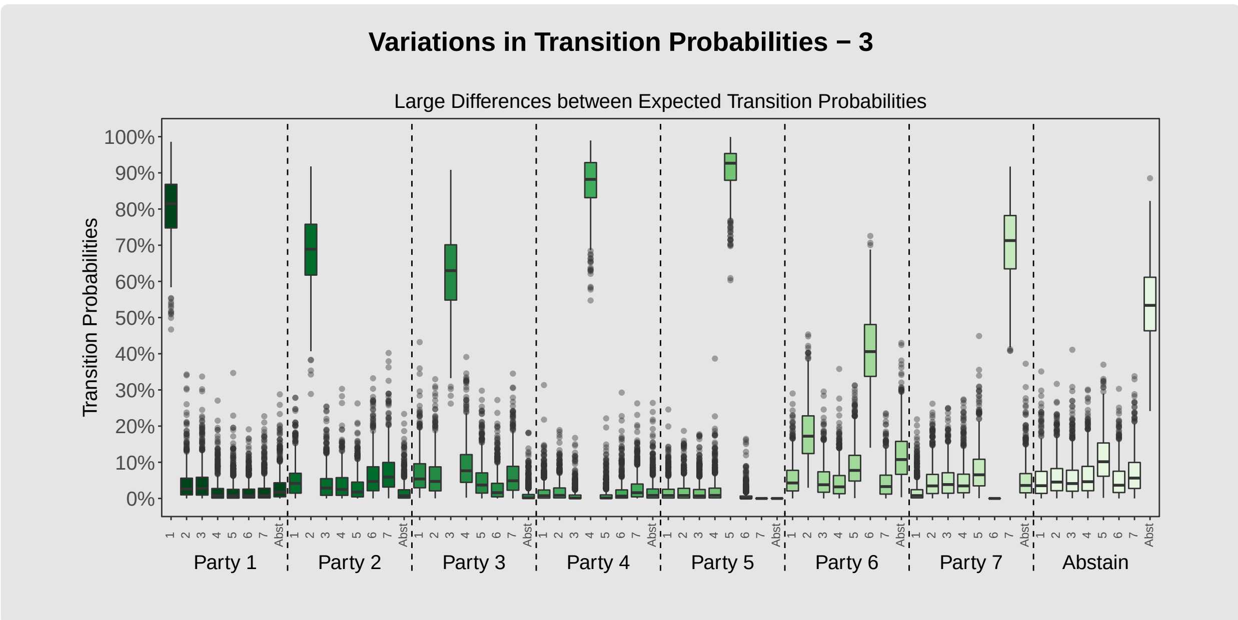


Figure 11: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of large differences between the expected transition probabilities from one simulation.*

[146] Please refer to equation 55.

4.3.1. Impact of Approach 2 and Approach 3 on the Transition Probabilities

As discussed at the beginning of the chapter there are three different approaches used to simulate data for the simulation study. This chapter will shortly discuss the impact of Approach 2 and 3 on the distribution of the transition probabilities on the basis of two examples, since the impact of the additional approaches in similar in all cases. All variations in transition probabilities for the two additional approaches can be found in appendix 2.

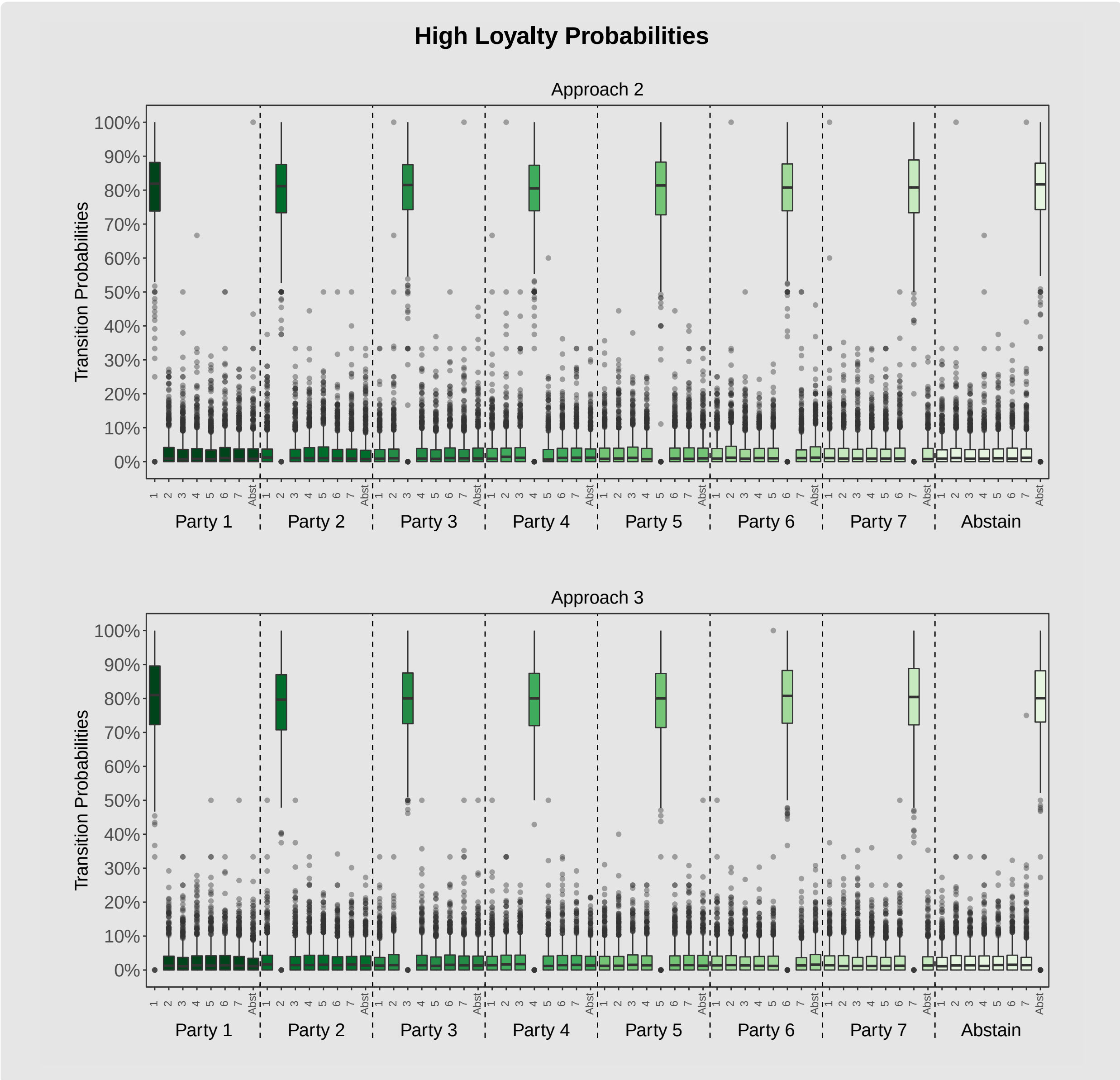


Figure 12: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of high loyalty probabilities from one simulation with Approach 2 (upper graph) and Approach 3 (lower graph).*

Approach 1 draws the transition probabilities directly from a Dirichlet distribution and thus in relative numbers. In contrast, Approach 2 and 3 simulate voter transition in absolute numbers. In order to compare the impact of the assumptions made in the additional approaches on the transition probabilities, they are calculated by dividing the voter transitions in absolute numbers $Y_i^{r,1}, ..., Y_i^{r,C}$ with the first election results $N_{r,i}$.

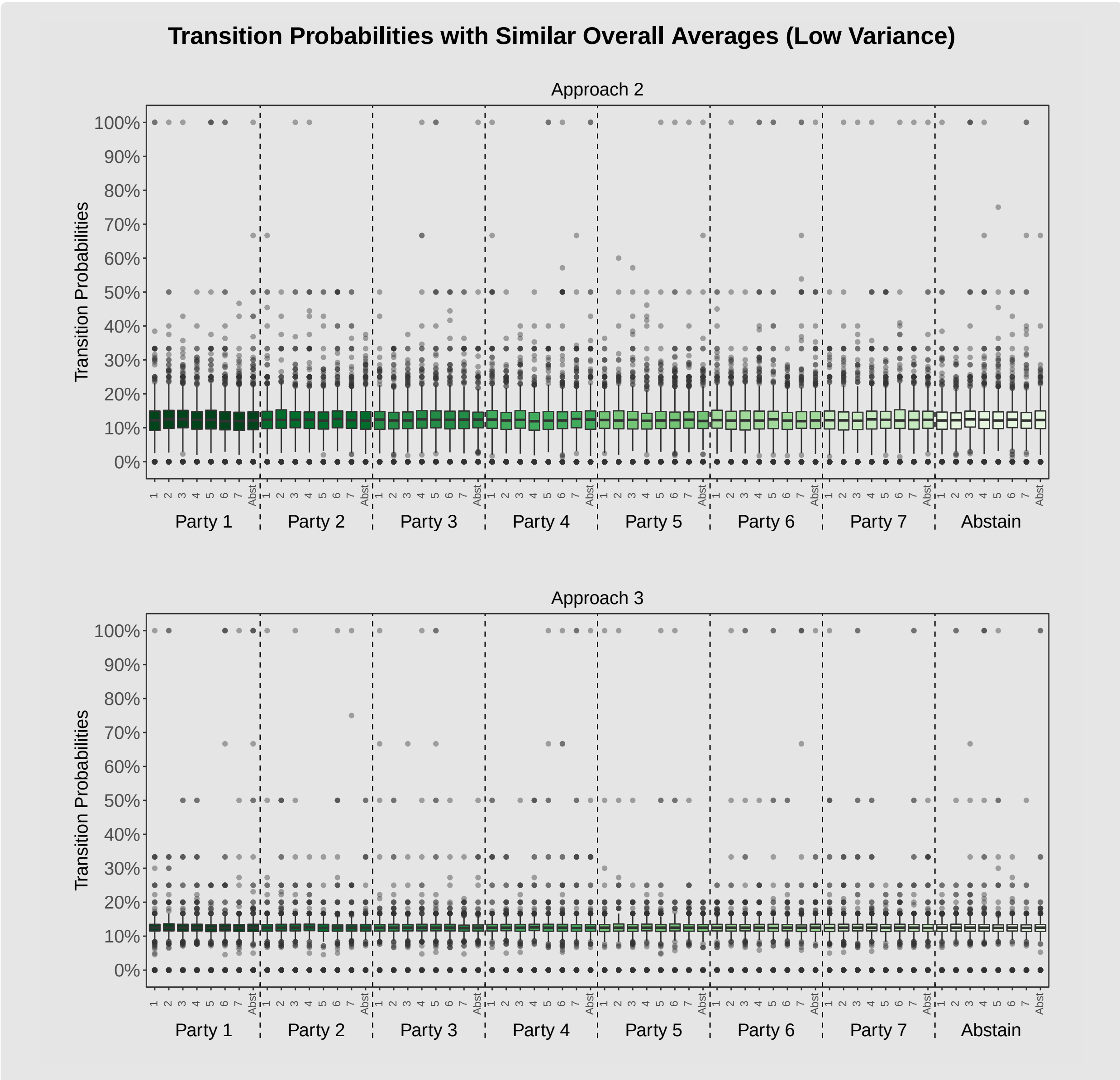


Figure 13: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of varying loyalty and transition probabilities with similar overall averages with low variance in transition probabilities. The data comes from one simulation with Approach 2 (upper graph) and Approach (lower graph).*

Fundamentally, both approaches increase the variance in the transition probabilities, while the overall averages stay at a similar level compared to the previous approach. Figure 12 illustrates the changes in the case of high loyalty rates. The two new approaches increase the range of the loyalty rates to about 50 % and 100 %, compared to 60 % and 95 % in the initial approach. However, similar to the initial approach the loyalty are on average still at about 80%. Downward outliers of up to 25 % occur in both cases. In contrast to the loyalty rates, the other transition probabilities do not change much. They fluctuate between 0 % and 10%, which is similar to the initial approach. Also, the overall average of the other transition probabilities stays with about 3% at a similar level. Only the upward outliers show a higher fluctuation.

A further change results from the fact that the case with fixed transition probabilities breaks away because it would lead to the same results as in the initial approach. Therefore, the scenario of fixed transition probabilities was replaced by a scenario with varying loyalty and transition probabilities with similar overall averages but low variance. The decision to introduce this assumption instead of the fixed transition probabilities was motivated by the results, which are discussed in Chapter 5.1. The resulting transition probabilities are displayed in Figure 13. With an average of about 12.5% the overall averages of the transition probabilities stay at a similar level. However, while they are fixed in the initial approach, in this case, they fluctuate between about 3% and 25 % (Approach 2) and between about 10% and 15 % (Approach 3). Even though the ranges differ, similar upwards outliers of up to 100% occur in both cases.

4.4. Individual Data

The hybrid Multinomial-Dirichlet Model allows for the integration of individual data. Thus, individual data from two different sampling schemes were simulated. Based on the simulation study by Klima et al.[147] these were a telephone survey and an exit poll.

The telephone survey was assumed to be conducted with approximately 2.5% of the overall population. In each district between 1% and 5% of the voter transitions are known and can be used as individual data in the hybrid Multinomial-Dirichlet Model. The telephone survey was modeled as a “perfect” sample in which every person has the same probability of being included in the survey. Furthermore, no bias was assumed and voters of all parties, as well as non-voters, are included in the sample. The expected overall number of respondents depends on the assumptions in the respective scenario and can be calculated using the following formula:

Number of Districts $\times$Number of Voters $\times$0.025.

[147] Klima et al., “Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches” at p.12-15.

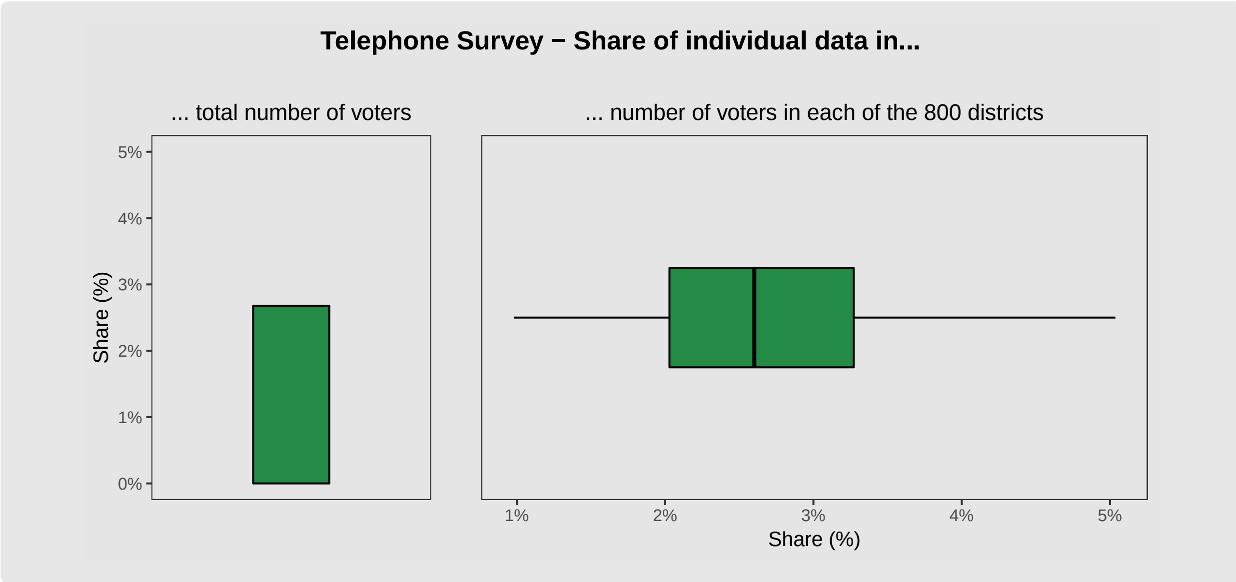


Figure 14: *Share of individual data from a telephone survey in the total number of voters (left graph) and in the number of voters in the 800 districts in which the telephone survey was conducted (right graph). Example from one simulation.*

The second sampling scheme for the individual data is the Exit Poll, which was assumed to be conducted with roughly the same share in the overall population as the telephone survey. An Exit Poll represents a survey in which the voters are asked for their decision on whom they voted for right after they cast their vote in the voting station. Such a survey cannot be realized in every district. Therefore, 5% of the districts were drawn with the same probability and an exit poll was simulated in each of them. With regards to the exit poll, two aspects are essential to take into consideration. Firstly, not all voters take part in such a survey. Hence, a response quote of 70% was assumed. Secondly, non-voters do not go to the voting stations and thus cannot be interviewed in the context of an exit poll. Therefore, non-voters are not represented in the individual data available from an exit poll. The share of individual data in the population of the selected districts is slightly lower than the response rate because the non-voters are not represented. As in the case of the telephone survey, the expected number of respondents depends on the particular scenario and can be calculated using the following formula:

Number of Districts $\times$0.05 $\times$Number of Voters $\times$0.7 $\times$Share of Non-Voters.

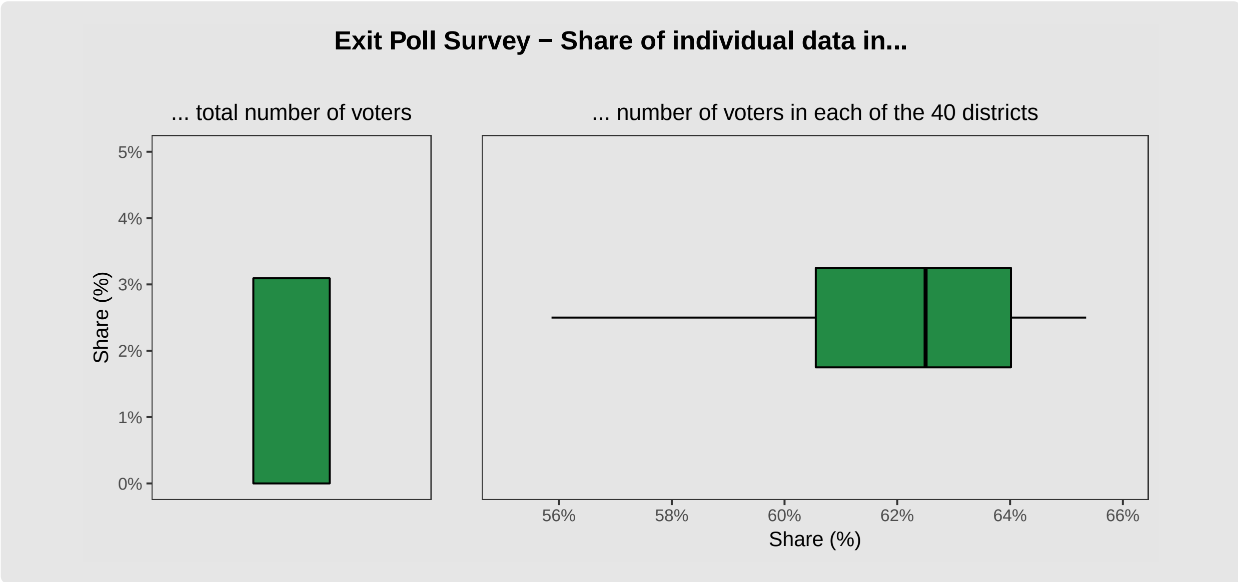


Figure 15: *Share of individual data from an exit poll in the total number of voters (left graph) and in the number of voters in the 40 districts in which the Exit Poll was conducted (right graph). Example from one simulation.*

4.5. Technical Aspects of the Data Simulation

This chapter is intended to give an overview of the technical aspects of the simulation study. Accordingly, the approach for the simulation of the data and the different variations will be explained. The methodology for the simulation study is based on the approach used by Klima et al.[148][149]. The data simulation begins with simulating the population size of each district. The population in each district N_i is drawn from a Poisson distribution:

$$N_i \sim Po(\lambda = N_{Voters}), \tag{66}$$

where N_{Voters} represents the expected average number of eligible voters per district. In the case structural differences in average population size between groups of districts, e.g. cities and municipalities, are assumed, the population for the districts in each group is drawn from two separate Poisson distributions:

$$N_i^1 \sim Po(\lambda = N_{Voters}^1), \tag{67}$$

$$N_i^2 \sim Po(\lambda = N_{Voters}^2), \tag{68}$$

where N_{Voters}^1 and N_{Voters}^2 reflect the expected average number of voters in group one and two, respec-

[148]Klima et al., "Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions".
[149]Their approach is explained in the online appendix of the paper.

tively.

In a second step, the results of the first election in the form of vote shares are simulated. For that purpose, three different assumptions about the vote shares of the parties are made: fixed vote shares, variable vote shares with similar overall averages and varying vote shares depending on party size. In the first case vote shares are simply given by:

$$(s_1, ..., s_n) \sim (1/Parties, ..., 1/Parties), \tag{69}$$

where $Parties$ and n represent the assumed number of parties plus non-voters. In the second case, the vote shares for each party in each district are drawn from a Dirichlet distribution, while the parameters of the Dirichlet distribution are drawn from a district specific Gamma distribution:

$$(\alpha_{1,i}, ..., \alpha_{n,i}) \sim Gamma(Parties, \lambda_1, \lambda_2), \tag{70}$$

$$(s_{1,i}, ..., s_{n,i}) \sim Dir(q \times (\alpha_{1,i}, ..., \alpha_{n,i})), \tag{71}$$

where $q = 50$ has been chosen to ensure a slight variation between the districts. The parameters of the Gamma distribution are chosen to be $\lambda_1 = 4$ and $\lambda_2 = 2$ in order to ensure similar overall averages. In the third case, the vote shares are simply drawn from a Dirichlet distribution with different shape vectors based on the number of parties.[150]

Subsequently, the results of the first election in each district in absolute numbers are calculated by multiplying the vote shares of the different parties $s_{1,i}, ..., s_{n,i}$ in disctrict i with the total number of voters in that district N_i :

$$(N_{1,i}, ..., N_{n,i}) = (s_{1,i}, ..., s_{n,i}) \times N_i. \tag{72}$$

The simulation of the transition probabilities for the parties in each district i constitutes the next step of the data simulation process. In this step five different cases are assumed: fixed transition probabilities, variable transition probabilities with a similar overall average, cases with high and low loyalty rates as well as a case with high differences between the expected transition probabilities. The approaches for simulating the transition probabilities resemble those for the simulation of the vote shares. However, instead of only simulating one vector for each district the approach is used to generate a $R \times C$-matrix containing the

[150] The shape vectors used can be found in the R-Code *functions_masterthesis.R* in the digital appendix.

transition probabilities where R stands for the number of Parties participating in the first election and C for the number of Parties participating in the second election.

In case fixed transition probabilities are assumed they each cell of the $R \times C$-matrix containing the transition probabilities is simply given by:

$$\beta_i^{r,c} = 1/Parties. \tag{73}$$

Variable transition probabilities with a similar global average are drawn from a row-wise Dirichlet distribution, while the parameters of the Dirichlet distribution are generated from a district specific Gamma distribution:

$$(\alpha_{1,i}, ..., \alpha_{n,i}) \sim Gamma(Parties, \lambda_1, \lambda_2), \tag{74}$$

$$(\beta_i^{r,c}, ..., \beta_i^{r,c}) \sim Dir(\alpha_{1,i}, ..., \alpha_{n,i}), \tag{75}$$

where the parameters of the Gamma distribution have been set to $\lambda_1 = 4$ and $\lambda_2 = 2$ in order to ensure similar overall averages. The method for simulating the last three cases closely resembles this approach.

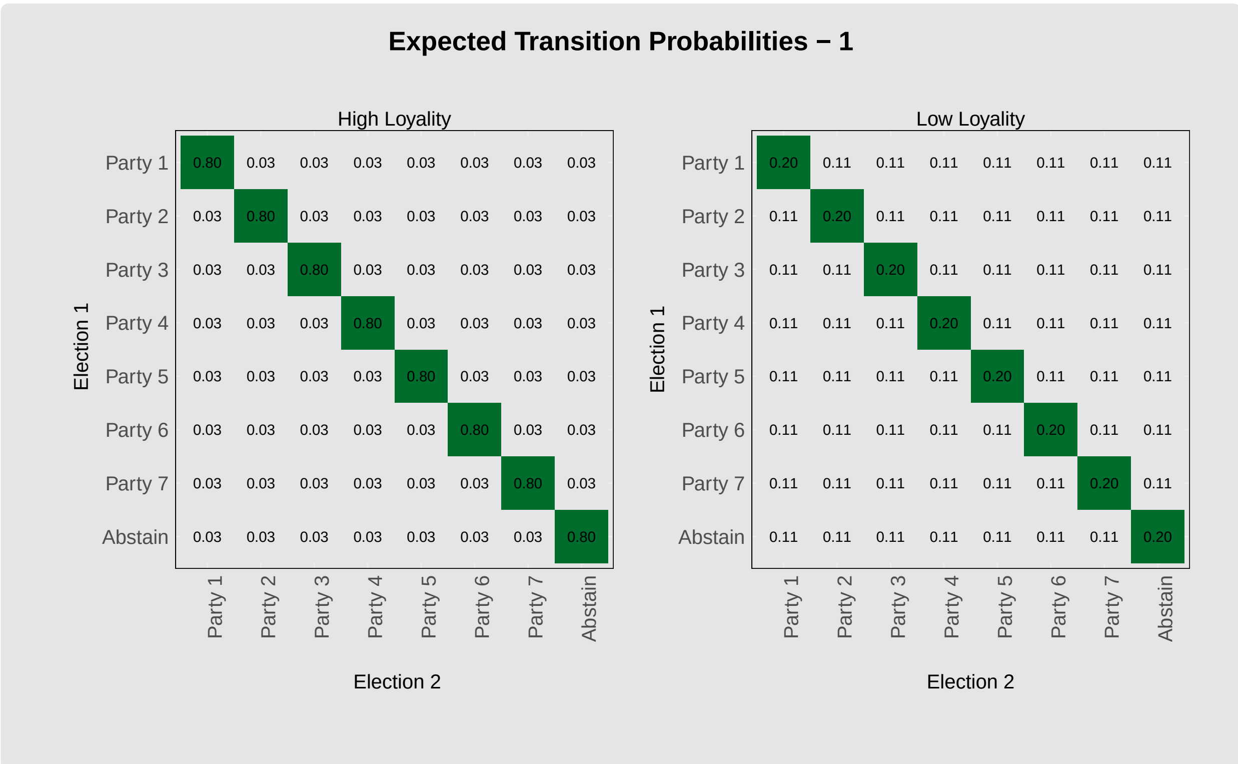


Figure 16: *Expected transition probabilities in the case of high loyalty rates (left) and in the case of low loyalty rates (right). The displayed expected transition probabilities were used in the case of seven assumed parties and non-voters. For this presentation some cell values had to be rounded to the second decimal place.*

However, instead of generating the parameters of the Dirichlet distribution from a Gamma distribution they correspond to one row in the matrices shown in Figure 16 and Figure 17. Figure 16 displays the expected transition probabilities in case either high (left graph) or low (right graph) loyalty rates are assumed, while Figure 17 displays the case in which large differences between the expected transition probabilities are assumed. Both matrices are displayed for the case of 7 parties and non-voters. [151]These expected transition probabilities are assumed for every district. The actual transition probabilities are drawn from a row-wise Dirichlet distribution:

$$(\beta_i^{r,c}, ..., \beta_i^{r,c}) \sim Dir(k \times (p_{r,1}, ..., p_{r,C})), \tag{76}$$

where $(p_{r,1}, ..., p_{r,C})$ refers to one row in one of the tables shown in Figure 16 and Figure 17. $k = 20$ has been chosen to ensure homogeneous transition probabilities between the districts.

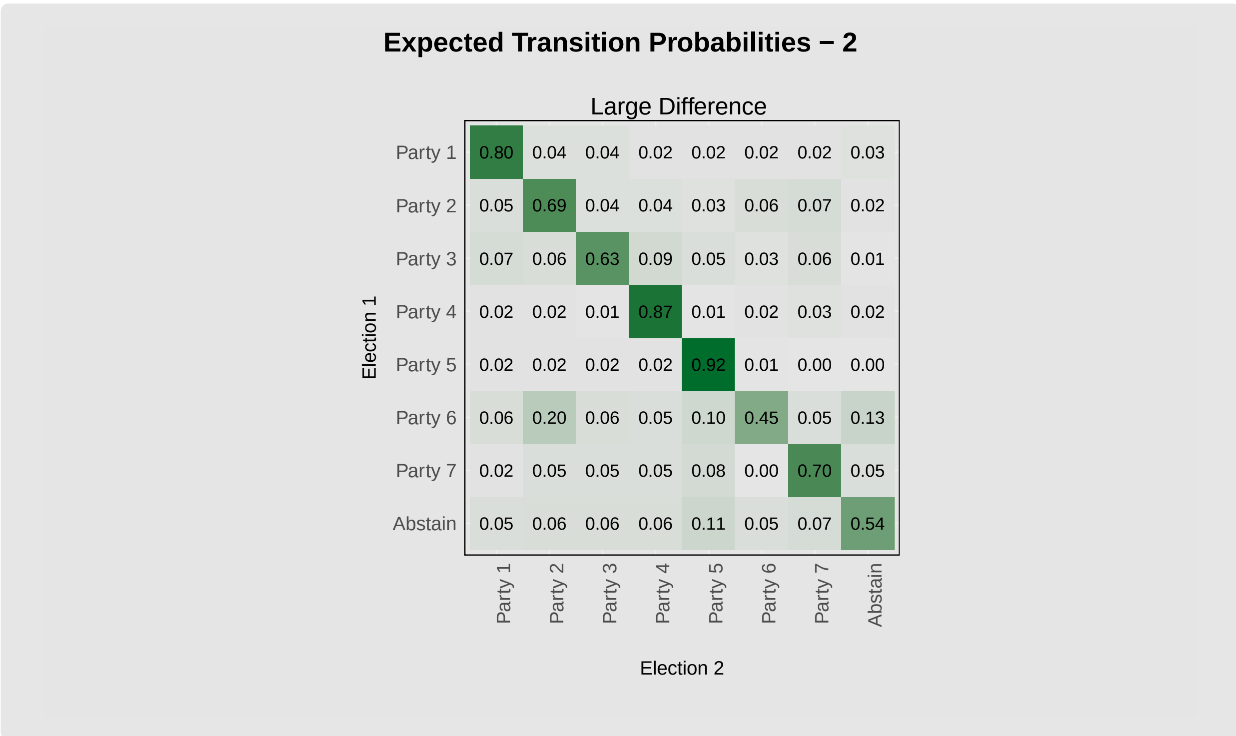


Figure 17: *Expected transition probabilities in the case of assumed large differences between the transition probabilities. The displayed expected transition probabilities were used in the case of seven assumed parties and non-voters. For this presentation some cell values had to be rounded to the second decimal place.*

The results of the second election are simply calculated by multiplying the election results of the first election in each district i with the transition probabilities. Summing over the columns of the resulting

[151] Please refer to appendix 3 for the matrix used in the case of five parties and large differences in the expected transition probab ilities.

matrix results in the outcome of the second election.

4.5.1. Variations in the Data Simulation

As discussed in the introduction to Chapter 4 two additional data simulation approaches were modeled. Both change the way the results of the second election are simulated but are motivated by different hypotheses. The following chapter explains the variations and discusses the underlying hypothesis.

The first variation, hereafter called *Approach 2*, extends the data simulation by an assumption made in the (hybrid) Multinomial-Dirichlet Model. As explained in chapter 3.2.2 and chapter 3.3.2 these models assume that the election results of the second election $T_{1,i}, T_{2,i}, ..., T_{C,i}$ follow a Multinomial distribution. Since this is not considered in the original data simulation approach it is now implemented in Approach 2. Instead of multiplying the transition probabilities $\beta_i^{r,1}, ..., \beta_i^{r,C}$ with the results of the first election $N_{r,i}$, they are now used as parameters in a row-wise Multinomial distribution:

$$Y_i^{r,1}, ..., Y_i^{r,C} \sim Multi(\beta_i^{r,1}, ..., \beta_i^{r,C}, N_{r,i}), \tag{77}$$

with $N_{r,i}$ being the first election results of Party r in district i. The results of the second election in district i are given by the column sums of the resulting matrix for that district.

The second variation, hereafter called *Approach 3*, fundamentally changes the way the second election results are simulated. In the first two approaches, the transition probabilities were regarded as stochastic, while the second election results were deterministic. Approach 3 reverses this point of view. This third approach is based on the consideration that the previous two approaches might violate assumptions made in the (hybrid) Multinomial-Dirichlet Model. These models assume the results of the second election to follow a Multinomial distribution. However, in the first two approaches the election results of the second election are deterministic and are a result of summing over the columns of the transition matrix. They might not follow a Multinomial distribution but a convolutional distribution, which would constitute a violation of the model assumptions. Therefore, in Approach 3 the results of the second election are simulated according to the respective assumption [152] in both models. Consequently, the second election results are simulated as follows:

$$(T_{1,i}, ..., T_{C,i}) \sim Multi(N_i, \Theta_{1,i}, ..., \Theta_{C,i}), \tag{78}$$

[152] Please refer to chapter 3.2.2 and chapter 3.3.2 for reference.

with

$$\Theta_{c,i} = \sum_{r=1}^{R} \beta_i^{rc} X_{r,i}, \tag{79}$$

and

$$X_{r,i} = \frac{N_{r,i}}{N_i}. \tag{80}$$

In a previous step the β's used in the calculation of $\Theta_{c,i}$ are drawn from a Dirichlet distribution.[153] Afterward, the now deterministic voter transitions are calculated. This step includes distributing the voters of the second election in the individual cells of the voter transition matrix. Therefore, for each cell a share of voters is calculated, which shows the share that cell has in the second election results of party c:

$$Share\_Voters_{r,c,i} = \frac{\beta_i^{r,c} N_{r,i}}{\Theta_{c,i}}. \tag{81}$$

To calculate the voter transitions $Share\_Voters_{r,c,i}$ is multiplied with the results of the second election $T_{c,i}$. This step, however, reveals a potential problem of this approach. Firstly, the resulting voter transitions are not integers, but decimals and while the column sums fit the results of the second election the row sums do not fit the results of the first election. Thus, in a further step, the vote transitions are rounded and adapted to the results of the first election, while keeping the column sums fixed.[154] The main problem with this approach however is, that there is no unambiguous solution for the voter transitions. This problem makes the evaluation of the estimation quality less clear, as a deterioration of the estimation quality could in part be due to the unambiguous nature of the simulated "true" voter transitions.

4.6. The Evaluation Scenarios

After the previous chapters introduced the variables of the simulation study, the following chapter will introduce and motivate the evaluation scenarios. All in all, there are six different variables with 23 different specifications, which are displayed in Table 9. In order to create the scenarios, these specifications are combined in different ways to highlight certain aspects of the data. Two groups of scenarios are tested: the *Baseline Scenarios* and the *Extended Scenarios*.

[153] For the sampling of the β's the approaches described in the previous Chapter 4.1 are used. The choice depends on the assumption regarding the transition probabilities.

[154] The rounding algorithm is displayed in appendix A.4

Variable Specifications
Number of Parties
7 Parties and Non-Voters
4 Parties and Non-Voters
Vote Shares
Fixed Vote Shares
Varying Vote Shares with Similar Overall Averages
1 Major Party
2 Major Parties
Transition Probabilities
Identical Transition Probabilities
Varying Transition Probabilities with Similar Overall Averages
Low Loyalty Rates
High Loyalty Rates
Large Differences between Transition Probabilities
Number of Districts
200
400
800
1600
2000
Average Population Size
400
800
1200
Structural Differences
Individual Data
None
Exit Poll
Telephone Survey

Table 9: *Overview of the different variable specifications, whose influences on the quality of the credible intervals were examined with the Baseline scenarios. The six variables used in the data simulation are displayed with a grey background. Under them, their specifications are presented. The specifications used in the control scenario are displayed in cursive.*

The Baseline Scenarios

The goal of the Baseline scenarios is to gain an understanding of how the different specifications of the variables shown in Table 9 impact the quality of the credible intervals. Therefore, their impact on the quality of the interval estimates was tested independently. The starting point for the Baseline scenarios were the specifications in Table 9, which are written in cursive: 7 parties plus non-voters, varying vote shares with similar overall averages, varying transition probabilities with similar overall averages, 800 districts, an average of 800 voters per district and no individual data. These were chosen because they represent a relatively simple scenario, which produces valid 80% credible intervals.[155] Chapter 5.1 will discuss this aspect in more detail. The impact of the other specifications was tested by successively adding them to the initial scenario while keeping the other specifications of the first scenario fixed. This procedure results in 18 different Baseline scenarios, including the initial one. An overview of all scenarios and their specifications can be found in appendix 6. For each scenario, 15 iterations were simulated with each of the three approaches discussed in chapter 4 and voter transitions were estimated for these iterations using the (hybrid) Multinomial-Dirichlet Model.

The Extended Scenarios

The Baseline scenarios investigate the impact of the variable specifications shown in Table 9 individually. However, it is unlikely that these specifications only appear individually. Therefore, it is important to evaluate the impact of compound data scenarios on the interval estimates. For this purpose, two sets of scenarios have been created, which are loosely based on the general circumstances found in Bavarian state elections and German federal elections. The intention lies not in a perfect replication of the circumstances found in these elections, but to create more complex data scenarios. The scenarios are created by gradually adding further complexity to a base scenario. They are hereafter called *Bavaria Scnearios* and *Germany Scenarios*.

As mentioned above the *Bavaria Scenearios* are loosely based on the circumstances found in a state election in Bavaria. Traditionally, with the Christian Social Union (CSU) Bavaria has one major party capable of achieving a high vote share. In 2013 state elections they attained 47.7% of the popular vote, which dropped to 37.2% in the 2018 state elections. The number of relevant smaller parties fluctuates between elections. For example, in the 2013 state elections the Social Democratic Party (SPD), the Free Voters (FV) and the Greens leaped into parliament with vote shares between 8.6% and 20.6%. In 2018 with the Alternative for Germany (AfD) and the Free Democratic Party (FDP) two additional parties joined them. The range of vote shares of these five parties in the 2018 elections was between 5.1% and 17.6%, while the proportion of non-voters in both elections fluctuates between 36.4% (2013) and 27.7% (2018). Thus, the Bavarian state elections have on average one major party, three smaller ones and non-voters. An overview of the election results of the two elections can be found in appendix 5. Based on the results of these elections in the *Bavaria Scnearios* a situation with four parties plus non-voters is assumed. Additionally, the existence of one major party and three smaller parties is postulated. Accordingly, the first scenario, *Bavaria 1*, is configured as follows: 4 parties and non-voters, one major party, varying transition probabilities with similar

[155] Initially, an even simpler scenario with fixed vote shares and fixed transition probabilities was assumed. However, this led to poor performance of the point and consequently the interval estimates. Appendix 9 discusses these results in more detail.

overall averages, 800 districts, an average population of 800 and no individual data.

Further variable specifications are gradually added to this first scenario. The federal state of Bavaria has 2056 municipalities and cities. Naturally, the average population sizes between municipalities and cities differ substantially. This was taken into consideration by assuming 2000 districts and structural differences in population sizes. To create even more realistic and complex scenarios it was assumed that individual data from a telephone survey and an exit poll were available and that substantial differences between the expected transition probabilities were present. An overview of the scenario schedule can be found in appendix 6.

A second extended scenario is loosely based on the circumstances found in the German federal elections. In these elections, the two so-called “Volksparteien”, people’s parties, Christian Democratic Party (CDU) and the SPD accumulate about 50% of the votes. In 2013 they achieved vote shares of 34.1% and 25.7%, respectively. However, their respective vote shares dropped to 26.8% and 20.5% in the 2017 elections. In addition to the CDU and SPD, the Left Party, the Greens and the CSU leaped into parliament in 2013. The AfD and the FDP joined them after the federal election in 2017. In 2017 the vote shares of the smaller parties ranged between 8.9% and 12.6%. The share of non-voters in these two elections was with 28.5% in 2013 and 23.8% in 2018 slightly smaller compared to Bavaria. Therefore, in the *Germany Scenarios* a situation with seven parties and non-voters is assumed. Also, the existence of two major parties and four smaller parties is postulated. Accordingly, the scenario *Germany 1* is configured as follows: 7 parties and non-voters, two major parties, varying transition probabilities with similar overall averages, 800 districts, an average population of 800 and no individual data.

Similar to the Bavaria scenarios, further variable specifications are gradually added to this first scenario. The population sizes between the constituencies differ substantially, which was considered with the assumption of structural differences in population sizes. As in the Bavaria Scenarios, it was assumed that individual data from a telephone survey and an exit poll were available and that substantial differences between the expected transition probabilities were present. An overview of the scenario schedule can be found in appendix 6.

5. Evaluation of the Quality of the Credible Intervals

The following chapter evaluates the quality of credible intervals resulting from Ecological Inference and hybrid Multinomial-Dirichlet Models. The basis for the evaluation is the simulation study described in chapter 4. The evaluated scenarios are introduced in chapter 4.6, and a tabular overview of the scenario schedule can be found in appendix 6. All scenarios have been calculated with data from the three different simulation approaches introduced in chapter 4. The results of the models estimated with data simulated with the second approach will not be presented, due to their similarity with results of the models estimated with data from the first approach.

Since the data basis is simulated, the voter transitions are known. Thus, it can be evaluated in how many cases the cell value is covered by the $(1-\alpha)$ credible interval. The evaluations in this chapter are based on credible intervals calculated using the Equal-Tail approach introduced in chapter 2.1.3. The HPD intervals have also been evaluated, but because the analyzed Markov Chains had unimodal density plots the intervals were very similar compared to those resulting from the Equal-Tail approach. Thus, this thesis refrains from presenting the results of the HPD intervals. To determine the quality of the credible intervals the so-called coverage rate is calculated for each model. This rate indicates the share of the 80% credible intervals that cover the true value, aggregated over one voter transition table. For example, a coverage rate of 0.5 would indicate that 50% of the credible intervals cover the true value of the simulation.

For some selected scenarios, the cell-wise coverage rates are also considered. In this case, the coverage rates are calculated for each cell across all models rather than for each model across all cells. Since 15 or 30 models are not sufficient for a meaningful evaluation of the cell-wise coverage rates, 100 models were estimated for these selected scenarios. Due to the limited time and computing capacities, this could not be carried out for all scenarios. The selected scenarios are the Control scenario and scenario with large differences between transition probabilities (both *Baseline Scenarios*) as well as the scenarios Germany 4.1 and 4.2.

The choice of $\alpha = 20$% is justified by the number of calculated models for each scenario. Apart from the selected scenario described above, only 15 models are calculated for the baseline scenarios and 30 for the extended ones. Thus, the choice of $\alpha = 5$% would have been too restrictive. Depending on the number of simulated parties $5 \times 5 = 25$ or $8 \times 8 = 64$ credible intervals are calculated for each model. In the case of $\alpha = 5$%, the number of expected errors, in this case, credible intervals that do not cover the true value, would only be about one in the case of five parties and three in the case of eight parties. Since a certain level of stochasticity can be expected in every model, the low error tolerance could have led to a misinterpretation of the results, because small variations of one or two errors would already drastically change the interpretation of the results. With 15 or 30 calculated models, it was not valid to assume that the expected mean value of errors would be reached. The choice of $\alpha = 20$% allows a higher error tolerance due to the higher number of expected errors. Hence, individual variations do not influence the interpretation of the results so drastically.

In the case increasing or decreasing coverage rates are observed, two possible reasons will be discussed, namely the width of the credible intervals and the estimation error. Firstly, an increased width of the credible interval can, ceteris paribus, lead to a higher coverage rate and vice versa. This relationship exists due to the fact, that an increased width of the credible interval also increases the probability that it covers the true value. Secondly, a smaller estimation error can, ceteris paribus, lead to an improved coverage rate and vice versa. A smaller estimation error means that the values of the Markov Chain are on average closer to the true value. Thus, the credible intervals automatically move closer to the true value as well. The Average Distance (AD) will be used to quantify the estimation error. It is defined as follows:

$$AD(A_1, A_2) = \sum_{r=1}^{R} \sum_{c=1}^{C} \mid A_1(r,c) - A_2(r,c) \mid, \tag{82}$$

where A_1 and A_2 represent two voter transition tables with relative values. The AD represents the sum of the absolute cell-wise differences between the two tables. The resulting value ranges between 0 and 2. Half of that value represents the share of voters that has to be redistributed in one table in order to get identical voter transition tables.

The following two chapters will discuss the results of the evaluation of the credible interval quality. While chapter 5.1 discusses the results of the Baseline scenarios, chapter 5.2 discusses the results of the extended Germany and Bavaria scenarios.

5.1. Baseline Scenarios

The following chapter will evaluate the 80% credible intervals from the Baseline scenarios introduced in chapter 4.6. For each scenario 15 datasets were simulated according to the first and the third Approach.

Displayed in Figure 18 are the coverage rates of the 80 % credible intervals from the control scenario. In this case, the coverage rates fluctuate around 80%. Due to the stochasticity in the estimation of the voter transition a certain level of fluctuation of the coverage rates around the targeted 80 % is to be expected. However, on average the coverage rates are about 80%, which is the level that is expected from valid credible intervals. Furthermore, it is evident that there is no considerable difference between the coverage rates. Hence, it can be concluded that in this case neither the model used to calculate the voter transitions nor the approach used to simulate the data seems to have an impact on the quality of the credible intervals.

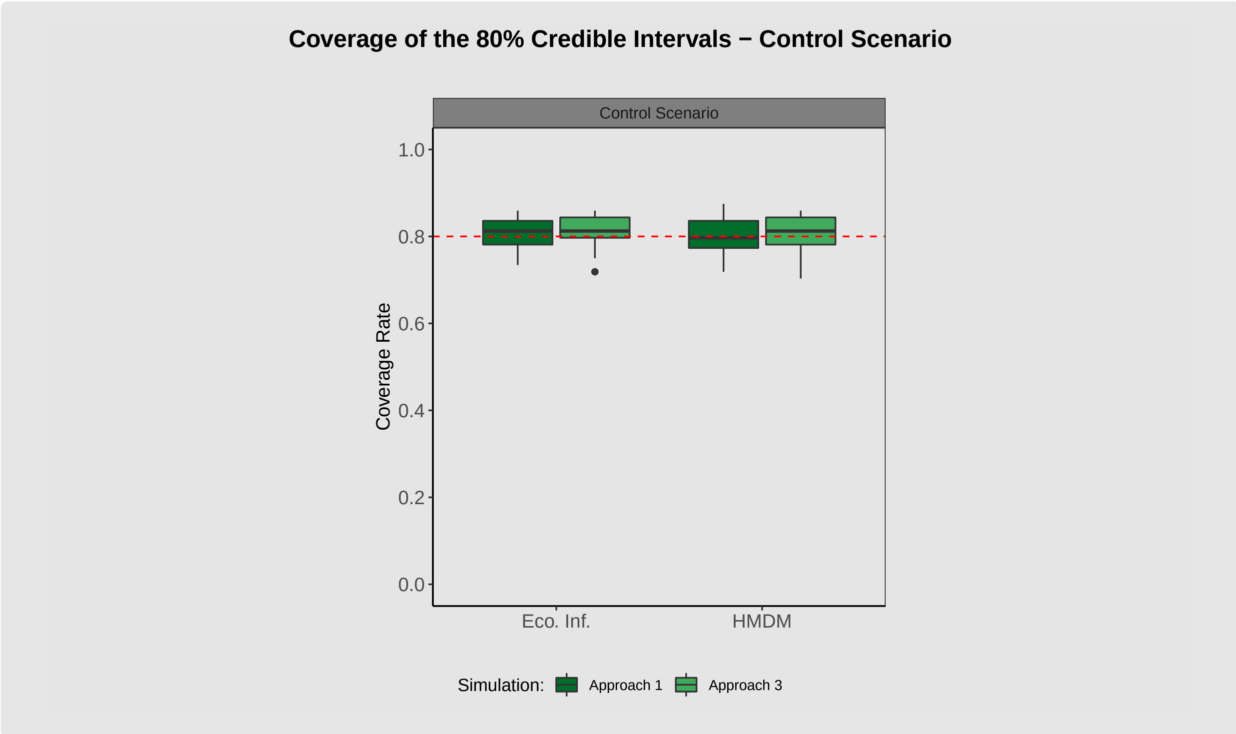


Figure 18: *Detailed evaluation of the 80%-credible intervals for the Baseline Model of the Control scenarios. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

Furthermore, the control scenario was among the four scenarios selected for the cell-wise evaluation of the credible intervals. The results of this evaluation for HMDMs estimated with data from Approach 1 and Approach 3 are displayed in Figure 19. The tables display coverage rates between 70% and 90% in green. Due to the stochasticity in the model, a certain degree of fluctuation around the 80% mark is expected and thus credible intervals with coverage rates within this range can be regarded as valid. Coverage rates above 90% are displayed in blue, while coverage rates below 70% are shown in red. The cell-wise coverage rates of Ecological Inference Models have also been evaluated, which led to similar results. They can be found in appendix 8.1. In most cases, the coverage rates fluctuate relatively closely around the targeted 80% mark. Only in two cases the coverage rates deviate more than 10% from that mark. Hence, the graphs reveal that the data simulation approach does not seem to have a substantial impact on the cell-wise coverage rates as the results are very similar. Nonetheless, the cell-wise coverage rates, as well as the coverage rates per model, lead to the conclusion that the control scenario represents an adequate baseline from which the impact of variations in the data can be investigated.

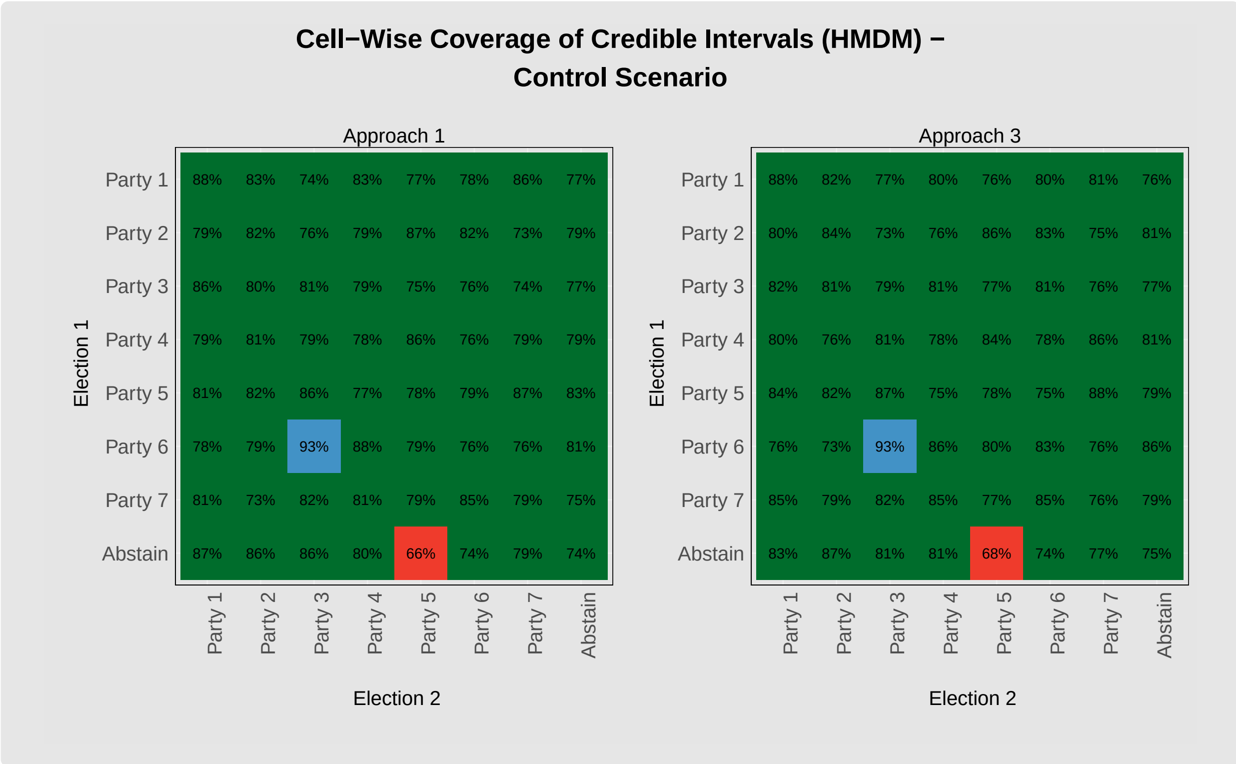


Figure 19: *Detailed cell-wise evaluation of the 80%-credible intervals for the Control Scenario of the Baseline scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graphs) and Approach 3 (right graphs). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach.*

The first variable investigated for its impact on the quality of the credible intervals was the number of parties. Here a change from eight to five parties was assumed. The resulting coverage rates are displayed in Figure 20. The results reveal that the change from eight to five parties does not negatively impact the quality of the credible intervals. The coverage rates are still scattered around about 80%. Besides a few more downward outliers from the models estimated with data from Approach 1, the data simulation approach did not impact the resulting credible intervals. Thus, it can be concluded that in this case the number of parties does not impact the quality of the credible intervals negatively.

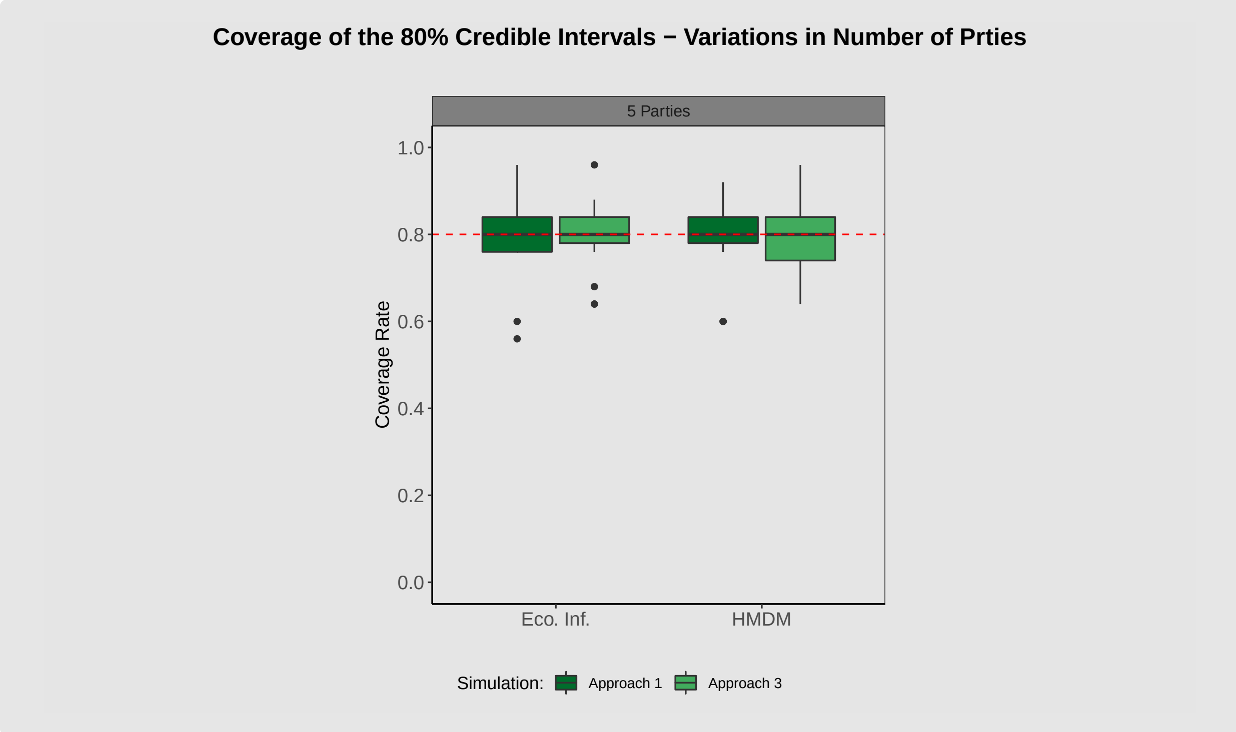


Figure 20: *Detailed evaluation of the 80%-credible intervals for Baseline scenario 1. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

Figure 21 displays the coverage rates in the case of different specifications of the vote shares. This is the first set of scenarios in which the coverages rates deviate from the expected 80%. The assumption of fixed vote shares leads to coverage rates that are too conservative. Besides two outliers all coverage rates are at 100%. Interestingly, the assumption of varying vote shares with one or two major parties also leads to conservative credible intervals with coverage rates well above 80%, but only in case the models were estimated with data from Approach 1. The coverage rates of models, which used data from Approach 3 still fluctuate around the expected 80%. Contrary to the data simulation approach, the choice of model used to estimate the voter transitions does not have an impact on the credible intervals.

Table 10 displays the average credible width and average AD of the HMDMs for the control scenario and all scenarios with different vote share specifications. The values of the EI model can be found in appendix 7. Because the values of the HMDMs and the EI models are so similar the following conclusions equally apply to the results from the EI models as well. The very conservative coverage rates in the case of fixed vote shares can be explained by both the interval widths and the AD. While the interval widths are almost quintuple compared to the control scenario, the AD decreases from 0.8 to 0.5. Both lead to an increase in the coverage rates.

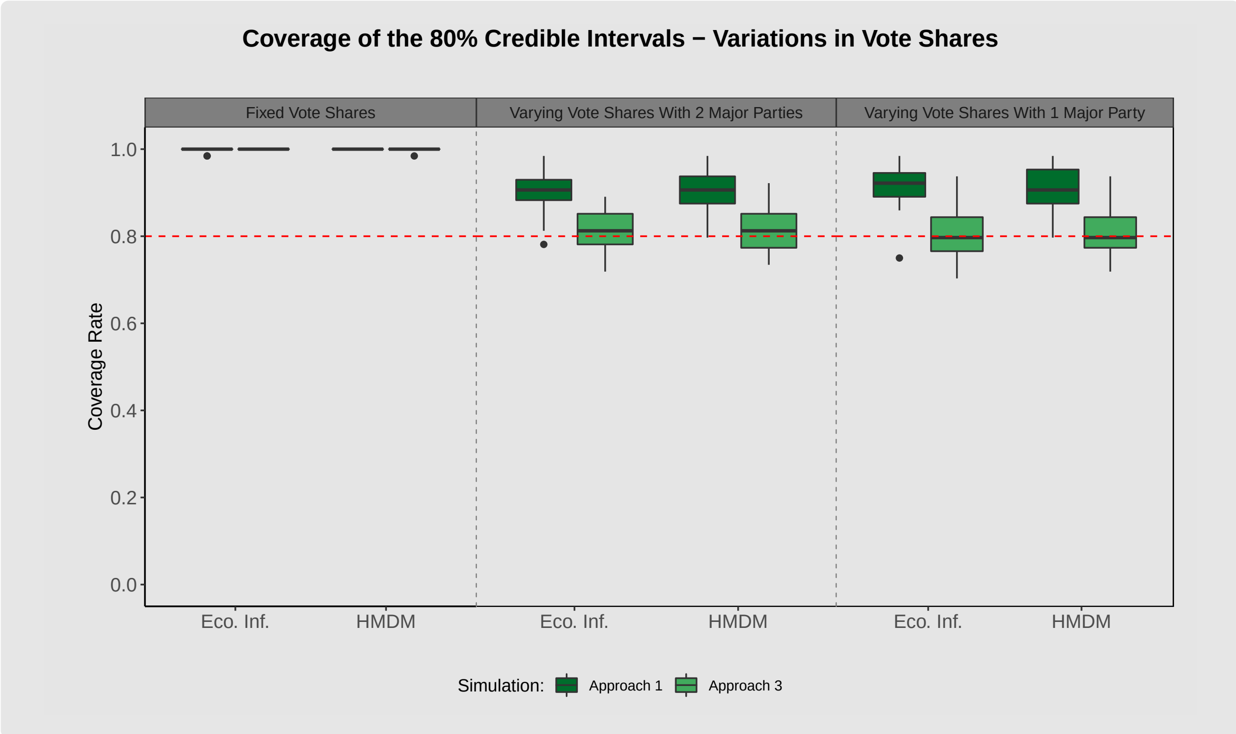


Figure 21: *Detailed evaluation of the 80%-credible intervals for Baseline scenario 2-4. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

Average Widths Credible Intervals and Average AD (HMDM) - Variations in Vote Share

	ø Interval Width		ø AD	
	Approach 1	Approach 3	Approach 1	Approach 3
Control	2522	2605	0.08	0.08
Fixed Vote Shares	10238	10027	0.05	0.05
Varying Vote Shares with 1 major Party	5510	3178	0.14	0.09
Varying Vote Shares with 2 major Parties	4955	2861	0.12	0.09

Table 10: *Detailed evaluation of the average interval width and average AD for Baseline scenarios 2-4 in contrast to those from the control scenario. Displayed are average interval widths and AD's resulting from estimations of voter transitions with the HMDM using data simulated with Approach 1 and 3. Results from estimations with the Ecological Inference model can be found in appendix 7.*

The assumptions of variable vote shares with one or two major parties also lead to an increase in the average interval widths. In the case of estimates with data from Approach 1, an increased AD of 0.14 or 0.12 does not offset the effect of the increased interval widths on the coverage rates. If the estimations were performed with data from Approach 3, the average interval widths increase only marginally by approximately 200-500. This slight increase in interval widths, in conjunction with a marginally increased AD, results in coverage rates which continue to scatter around the expected 80%.

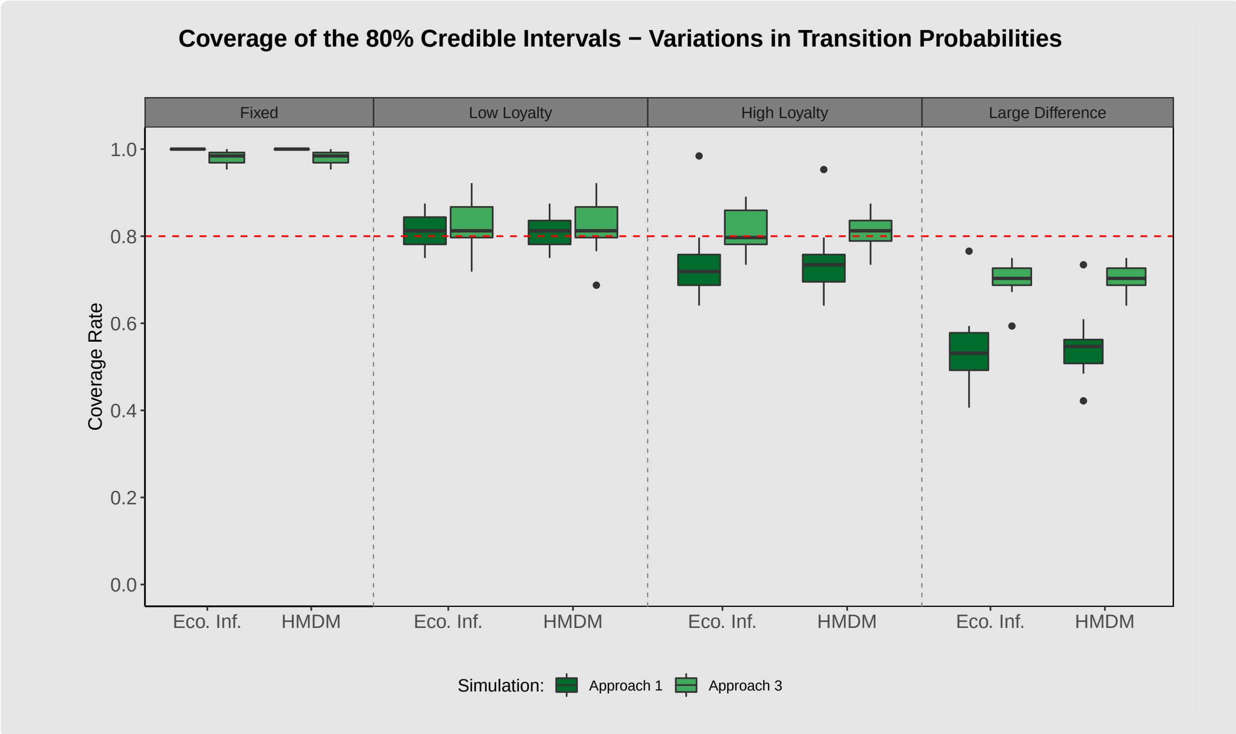


Figure 22: *Detailed evaluation of the 80%-credible intervals for Baseline scenario 5-8. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

Displayed in Figure 22 is the impact of variations in the transition probabilities have on the quality of the credible intervals. In the case of fixed transition probabilities, the coverage rates are too conservative. Models estimated with data from Approach 1 produce credible intervals covering 100% of the true values, while the coverage rates of models estimated with data from Approach 3 are on average slightly below 100%.A possible explanation could be the slight different implementations of this scenario discussed in hapter 4.5.1]. Because fixed transition probabilities would have led to the same results, varying transition probabilities with a low variance were assumed in Approach 3.

Postulating low loyalty rates leads to valid credible intervals or slightly too conservative in the case data from Approach 3 was used in the estimation. The next two assumptions, namely high loyalty rates and substantial differences between the expected transition probabilities, lead to a substantial impact on the quality of the credible intervals. The extent of the impact is highly dependent on the data used in the estimation. If models were estimated with data simulated with Approach 1, the quality of the interval estimates deteriorates considerably. In the case of high loyalty rates, the coverage rates are about 70%, while the assumption of substantial differences between the expected transition probabilities leads to coverages rates slightly above 50%. This is a level that corresponds to the results observed in other studies, which have shown the poor quality of interval estimates from ecological inference models.[156] At the same time, models estimated with data from Approach 3 produce valid credible intervals in the case of high loyalty rates and coverage rates, which deteriorate slightly to about 70% in the case of large differences between the expected transition probabilities. Nonetheless, in both cases the credible intervals have a higher quality if the data used in the estimation originated from Approach 3.

A look at the average interval width and average AD can again shed some light on the results displayed in Figure 22. The interval widths halve compared to those in the control scenario if high loyalty rates are postulated. While this would point towards lower coverage rates, the combination with an AD of 0 and 0.2 could offset the effect and explain the very conservative credible intervals. The assumption of low loyalty rates leads to interval widths and ADs similar to those in the control scenario, which could explain why the coverage rates still fluctuate around 80%. Furthermore, the deterioration of the coverages rates in the case of high loyalty rates and data simulated with Approach 1 can be explained by the small average interval width of around 800. The same applies to the scenario with large differences between the expected transition probabilities where the average interval widths are less than half compared to the control scenario. The even larger deterioration could be explained by the higher AD.

Average Widths Credible Intervals and Average AD (HMDM) - Variations in Transition Probabilities

	ø Interval Width		ø AD	
	Approach 1	Approach 3	Approach 1	Approach 3
Control	2522	2605	0.08	0.08
Fixed	1125	1216	0.00	0.02
Low Loyality	2657	2028	0.08	0.06
High Loyality	843	1028	0.03	0.03
Large Difference	1170	1130	0.07	0.04

Table 11: *Detailed evaluation of the average interval width and average AD for Baseline scenarios 5-8 in contrast to those from the control scenario. Displayed are average interval widths and AD's resulting from estimations of voter transitions with the HMDM using data simulated with Approach 1 and 3. Results from estimations with the Ecological Inference model can be found in appendix 7.*

The last scenario with large differences between the expected transition probabilities is also among the four selected for the cell-wise evaluation of the credible intervals. The results can be found in appendix 8.1. Two points stand out from the results. Firstly, the quality of the cell-wise credible intervals drops

[156] e.g. Klima et al., "Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches" at 106–7

considerably, with some coverage rates as low as 0%. Secondly, the overall quality of the credible intervals is substantially lower in the case of models estimated with data from Approach 1 compared to models estimated with data from Approach 3. Both observations correspond to the results of the model-wise coverage rates.

Figure 23 displays the coverages rates resulting from the models with variations in the number of districts. The results show that diverging from the assumption of 800 districts made in the control scenario only leads to a slight deterioration in the quality of the interval estimates. A lower number of districts leads to too conservative interval estimates and thus, on average slightly too high coverage rates. While the assumption of 200 districts results in an average coverage rate of about 85 %, assuming 400 districts leads to an average coverage rate of about 83%. On the contrary, a situation with 2000 districts leads to slightly lower coverage rates, if the models were estimated with data from Approach 1. However, the results displayed in Figure 23 show that the number of districts does not have a substantial impact on the quality of the credible intervals.

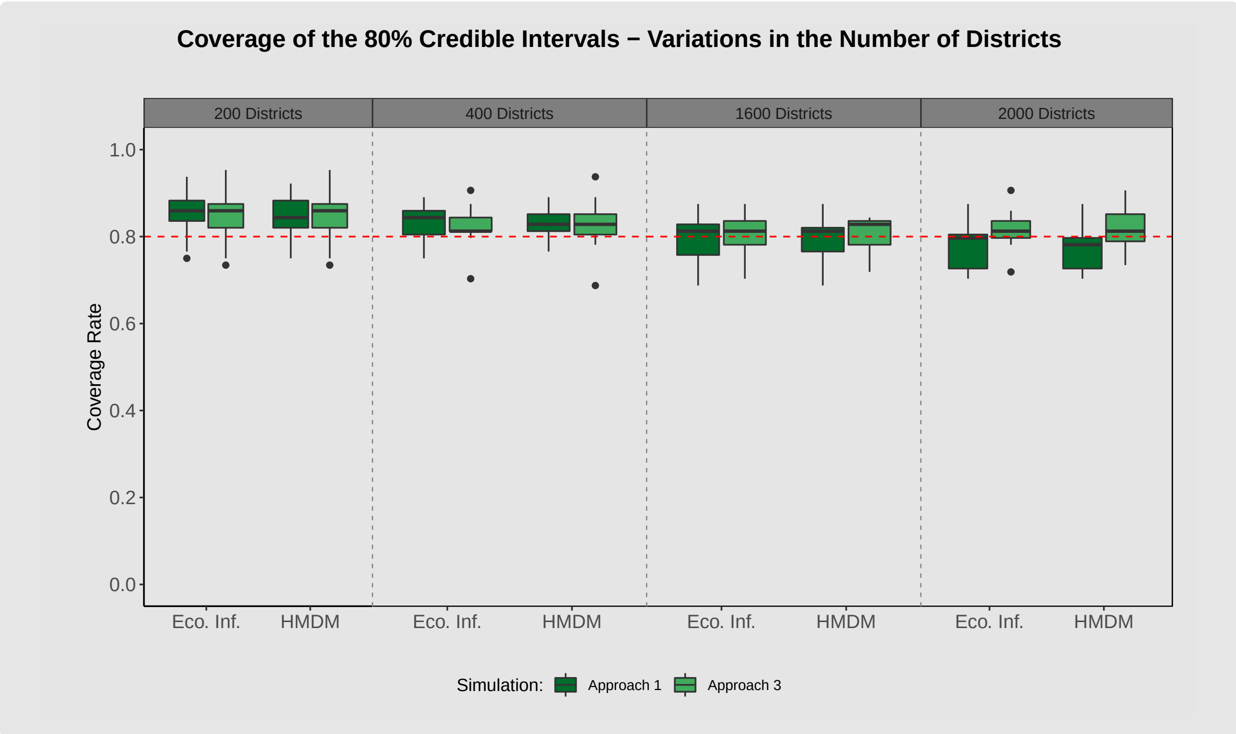


Figure 23: *Detailed evaluation of the 80%-credible intervals for Baseline scenario 9-12. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

Similar to the variations in the number of districts, the variations in the population size do not have a substantial impact on the quality of the credible intervals. The coverage rates for scenarios with variations in population size are displayed in Figure 24. Only in the case of structural differences, the quality of the credible intervals deteriorates resulting in average coverage rates slightly above the expected 80%.

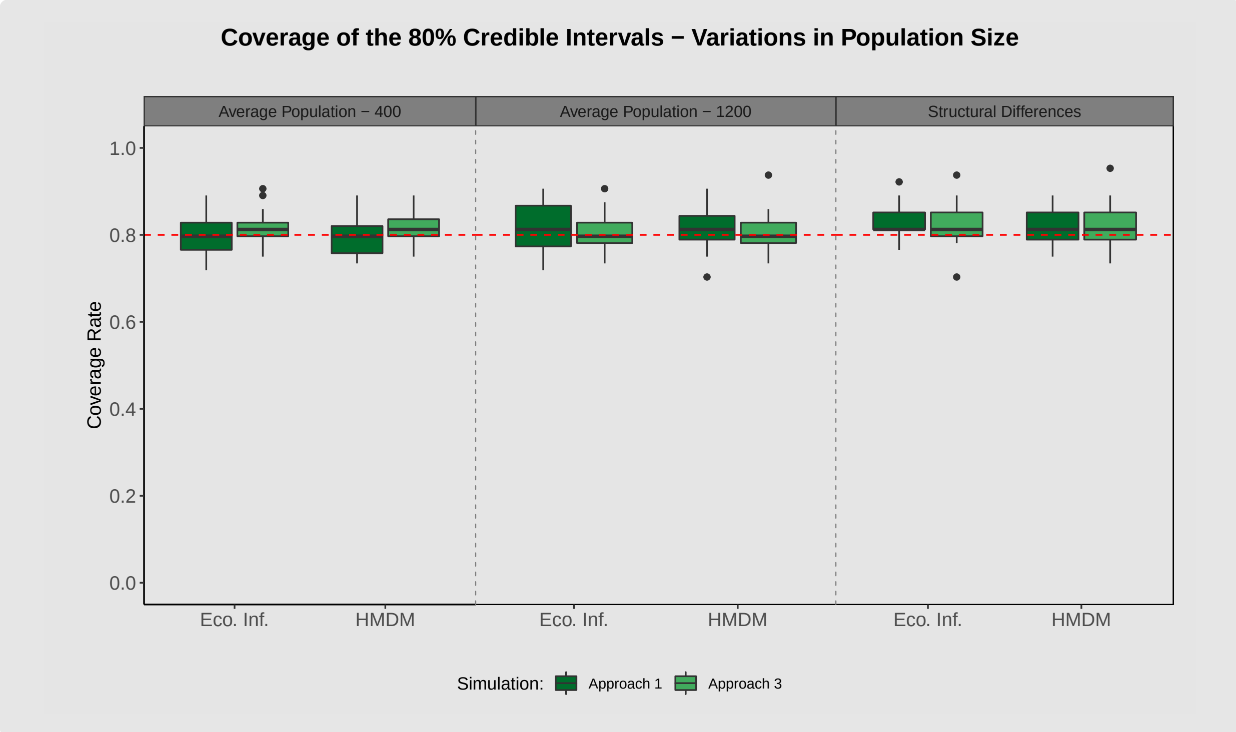


Figure 24: *Detailed evaluation of the 80%-credible intervals for Baseline scenario 13-15. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

The last two scenarios investigate the impact the inclusion of individual data in the estimation has on the quality of the credible intervals.[157] Figure 25 shows the resulting coverage rates only for the HMDMs, because individual data cannot be incorporated in an estimation with the EI model. The inclusion of individual data from a simulated exit poll does not hurt the quality of the credible intervals. This conclusion can be drawn from coverage rates, which still fluctuate around 80%. However, in the case of models estimated with individual data from a simulated telephone survey the quality of the credible intervals deteriorates. They are too conservative and fluctuate around 90%. The reason for that seems to be a very low estimation error. In the case of the telephone survey, the AD's are at 0.03 compared to 0.08 in the control scenario. Table 12 displays the average interval widths and average AD's for the control scenarios as well as those including individual data.

[157] An initial implementation of the telephone survey simulation produced an implicit bias in the outcome of the survey. The implications of such a bias are shortly discussed in appendix 10.

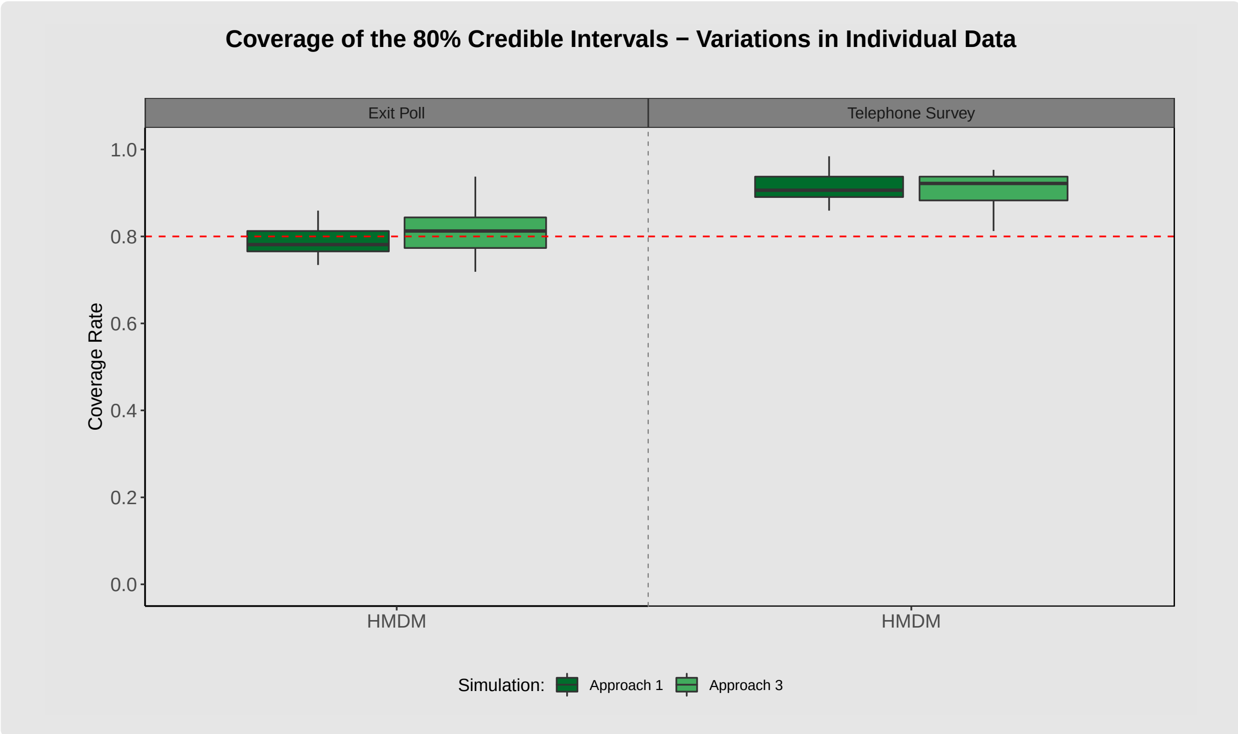


Figure 25: *Detailed evaluation of the 80%-credible intervals for Baseline scenario 16-17. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

Average Widths Credible Intervals and Average AD (HMDM) - Variations in Individual Data

	ø Interval Width		ø AD	
	Approach 1	Approach 3	Approach 1	Approach 3
Control	2522	2605	0.08	0.08
Exit Poll	1873	1922	0.06	0.06
Telephone Survey	1092	1097	0.03	0.03

Table 12: *Detailed evaluation of the average interval width and average AD for Baseline scenarios 16-17 in contrast to those from the control scenario. Displayed are average interval widths and AD's resulting from estimations of voter transitions with the HMDM using data simulated with Approach 1 and 3.*

5.2. Extended Scenarios

The following chapter will evaluate the 80% credible intervals of the extended scenarios Germany and Bavaria. For each scenario 30 datasets were simulated according to the first and third Approach.

Figure 26 displays the coverage rates of the 80% credible intervals resulting from the Germany scenarios. The first scenario has the following variable specifications: seven parties and non-voters, two major parties, varying transition probabilities with similar overall averages, 800 districts, an average population per district of 800 and no individual data. As can be seen in the left column of Figure 26, the coverage rates of the first scenario differ depending on the approach used to simulate the data. While the coverage rates of the models estimated with data from Approach 1 are with an average of about 90% too high, the models estimated with data simulated with Approach 3 produce coverage rates, which are with an average of about 75% slightly too low. The second scenario assumes structural differences in the district populations. This added complexity does not impact the coverage rates of the models simulated with data from Approach 1 substantially. Similar to the first scenario they are still too high and fluctuate around about 90%.On the other hand, the coverage rates of models estimated with data simulated with Approach 3 increase to an average level slightly above the targeted 80%.

The next two scenarios, Germany 3.1 and Germany 3.2, introduce individual data from an exit poll and a telephone survey, respectively. Because the EI model cannot incorporate individual data into the estimation, the results should on average not change compared to those from the Germany 2 scenario. This is precisely what happens, which can be observed in column three and four of Figure 25. Contrary to the EI models, the coverage rates of the estimations with HMDMs are impacted substantially. Including individual data from an exit poll results in coverage rates, which are on average slightly below 80% in the case of models estimated with data from Approach 1 and coverage rates fluctuating around 80% in the case models are estimated with data from Approach 3. The inclusion of individual data from a telephone survey leads to conservative credible intervals with coverage rates, which are fluctuating around 90%. Interestingly, so far the added complexity in the data did not seem to have a substantial impact on the quality of the interval estimates, as the results are very similar to those from the Baseline Scenarios.

For scenario Germany 4.1 and Germany 4.2 the assumption of large differences between the expected transition probabilities is added. The models with data from Approach 3 still produce valid credible intervals, which are fluctuating around about 80%. Only the combination of large differences in expected transition probabilities and individual data from a telephone survey leads to a substantial decrease in average coverage rates. On the contrary, models estimated with data simulated with Approach 1 are all heavily impacted by the addition of substantial differences between the expected transition probabilities. Resulting coverage rates range on average between 55% and 60%. Only the combination with individual data from a telephone survey leads to an even more drastic deterioration in the quality of the credible intervals. In this case, the average coverage rates fluctuate slightly above 20%.

Three interesting aspects of these results should be highlighted. Firstly, increasing complexity in the data does not necessarily seem to influence the quality of the interval estimates. Secondly, the assumption of large differences in the expected transition probabilities seems to be the only variable that leads to coverage rates well below the expected 80 %. Lastly, the coverage rates are substantially better, if the models were estimated with data from Approach 3 compared to data from Approach 1.

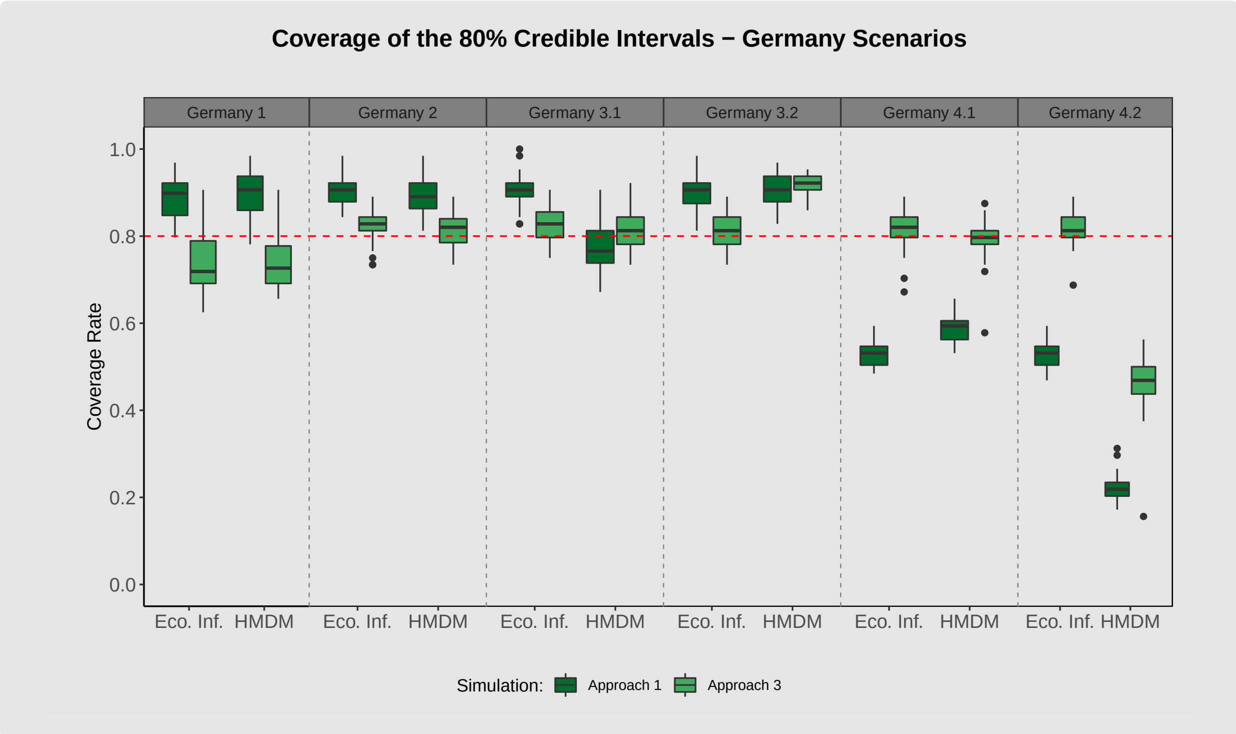


Figure 26: *Detailed evaluation of the 80%-credible intervals for the Germany scenarios. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 30 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

For the scenarios Germany 4.1 and 4.2 the cell-wise coverage rates of the credible intervals were also evaluated. The results can be found in appendix 8.2. The cell-wise coverage rates reveal that in most cases the mark of 80% is not achieved. Either the credible intervals are too conservative or too slim. This is leading to coverage rates, which are either fluctuating between 90% and 100% or between 0% and 10%. Only a few others can be found outside of these ranges.

A look at the interval widths and the average AD provides possible explanations for the results of the Germany scenarios. For the extended scenarios, the interval widths are represented with the help of violin plots in order to highlight their distribution. The interval widths, as well as the average AD for the Germany scenarios, are shown in Figure 27. Taking the interval widths and ADs of the first two scenarios into consideration, it is noticeable that they practically do not change. Besides, it is noticeable that the intervals of the models estimated with data from Approach 1 are on average considerably wider. This could explain the higher average coverage rates of the models estimated with data from Approach 1. However, if individual data is incorporated into the estimation as it is done in scenario Germany 3.1 and 3.2, the interval widths shrink substantially. This effect is particularly noticeable in the case of individual data from a telephone survey as integrated into scenario Germany 3.2. The fact that the average coverage rates do not decrease or even increase, as in the case of Germany 3.2, is due to the improved estimation quality. This improvement is expected when individual data is integrated into the estimation. In the case of individual data from an exit poll, the average AD drops from approximately 0.15 (Approach 1) or 0.11 (Approach 3) to 0.09 or 0.07, respectively. In the case of individual data from a telephone survey, the average AD drops even further to 0.03 or 0.02, respectively.

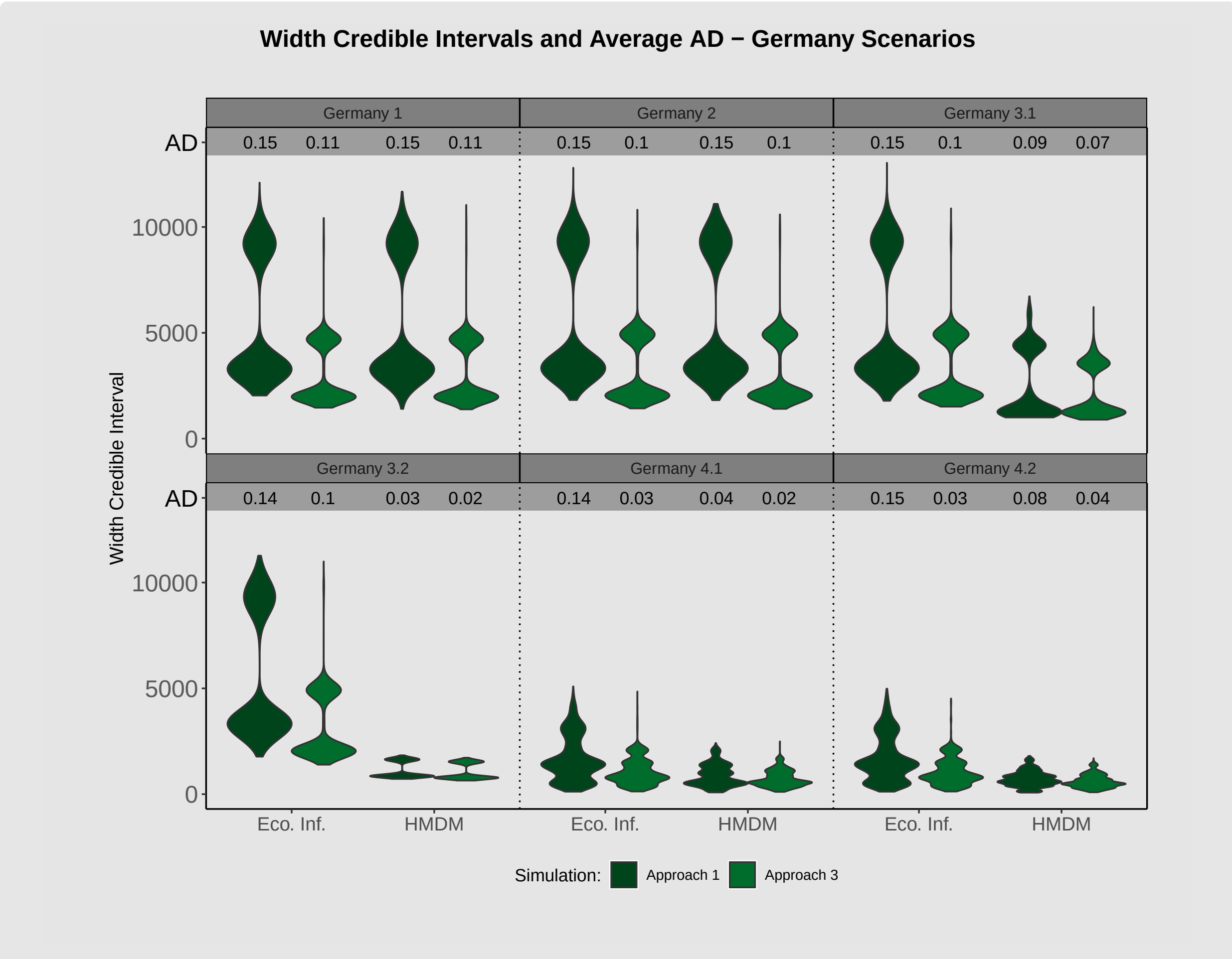


Figure 27: *Detailed evaluation of the interval widths of the 80% credible intervals for the Germany Scenarios with violin plots. Displayed is, for each estimated voter transition, the width of the credible interval. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 30 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The grey bar above the violin plots displays the average AD for each scenario.*

The assumption of large differences in the expected transition probabilities also leads to smaller average interval widths. Furthermore, for models simulated with data from Approach 1 the average AD increases substantially. The combination of these two factors could explain the low quality of the coverage rates. In the case of models estimated with data from Approach 3, the AD remains at a low level. Only for the HMDM in Scenario Germany 4.2, it increases slightly to 0.04. Since in this case, the average interval widths are minimal it could explain the drastic deteriorating quality of the interval estimates in this scenario.

Figure 28 displays the coverage rates of the 80% credible intervals resulting from the Bavaria scenarios. The first scenario has the following variable specifications: four parties and non-voters, one major party, varying transition probabilities with similar overall averages, 800 districts, an average population of 800 and no individual data.

Generally, the results from the Germany Scenarios and the Bavaria scenarios do not differ much. The EI models produce similar results in the first four scenarios. While the coverage rates of the models estimated with data from Approach 1 are on average slightly above 80%, the coverage rates from models estimated with data from Approach 3 fluctuate on average around the 80% mark. The results of the HMDMs for the first two scenarios are comparable. However, similar to the Germany Scenarios the introduction of individual data in scenario Bavaria 3.1 and 3.2 has a substantial impact on the coverage rates. In case of individual data from an exit poll, the average coverage rates are slightly below 80%. On the other hand, introducing individual data from a telephone survey yields coverage rates, which are on average almost at 100%. The assumption of 2000 districts instead of 800, which is added in the scenarios Bavaria 4.1 and 4.2, does not have a substantial impact on the coverage rates, other than decreasing the differences resulting from the different simulation approaches.

Again, similar to the Germany scenarios, it is the addition of large differences between the expected transition probabilities in the scenarios Bavaria 5.1 and 5.2, which has a substantial negative impact. Two points stand out. Firstly, models estimated with data from Approach 3 produce coverage rates with a higher quality, which are in this case on average between 80% and 90%. At the same, the average coverage rates of the models estimated with data from Approach 1 are much lower. They fluctuate around 20% and only the HMDM in the Bavaria 5.1 scenario produces higher coverage rates fluctuating around 40%. However, and this is the second point, when both individual data from a telephone survey and large differences between the expected transition probabilities are assumed the quality of the interval estimates vastly deteriorates. This is especially the case for the HMDM in scenario Bavaria 5.2, where the average coverage rates drop to about 20% (Approach 1) and 10% (Approach 3). This is also the only case in which the coverage rates of the models estimated with data from Approach 3 are considerably worse compared to models estimated with data from Approach 1.

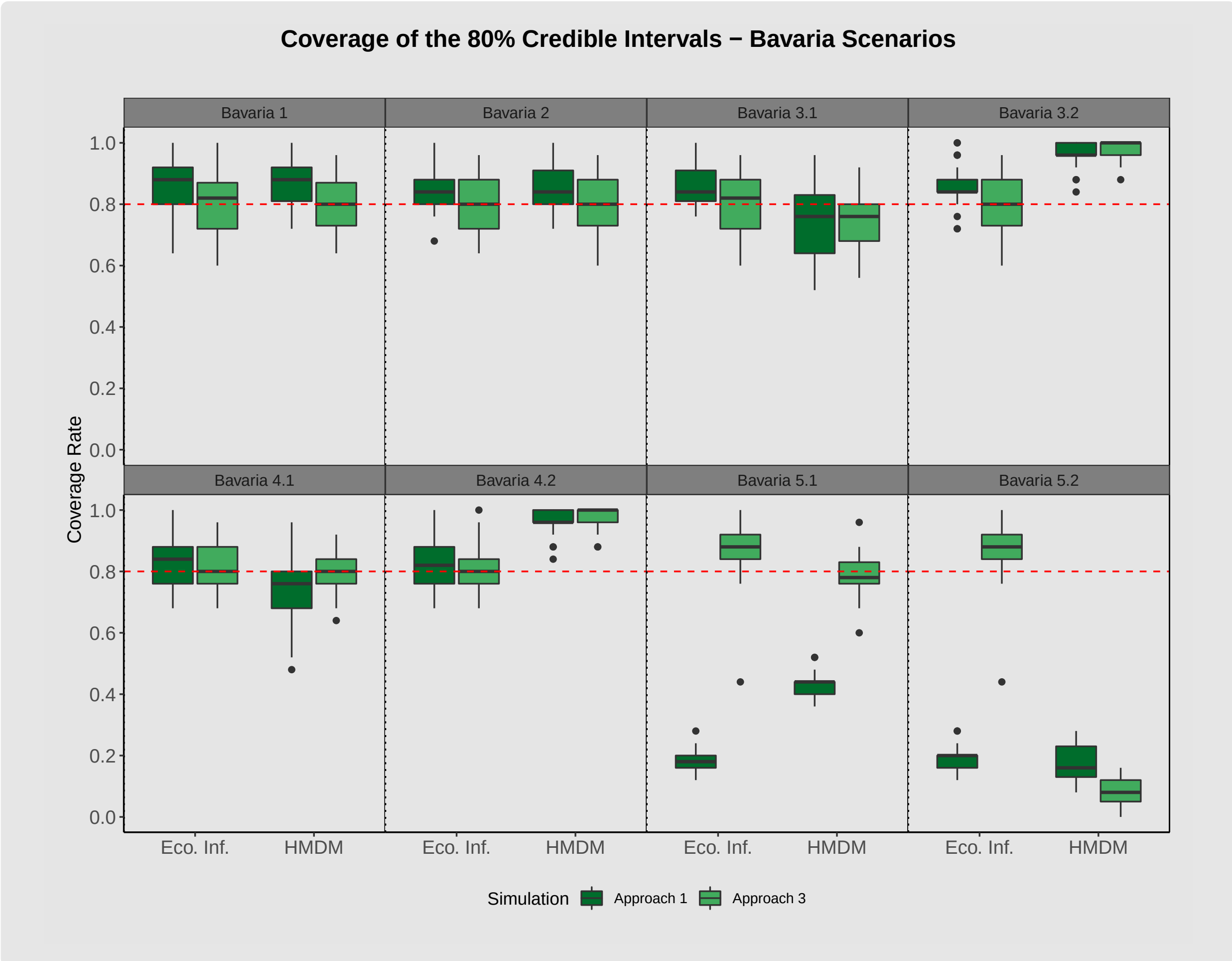


Figure 28: *Detailed evaluation of the 80%-credible intervals for the Bavaria scenarios. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 30 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

Figure 29 displays the interval widths and the average ADs for the Bavaria scenarios. The conclusions that can be drawn from these results are again similar to those from the Germany scenarios. Accordingly, for the first four scenarios the interval widths of the EI Models practically do not change, which can explain that the average coverage rates remain unchanged as well. A similar result can be observed for the HMDMs in the first two scenarios. Furthermore, the substantially wider intervals of the models estimated with data from Approach 1 could explain the on average higher coverage rates.

Analogously to the German scenarios, the interval widths drop drastically, if individual data is introduced into the estimation. This process can be observed for the HMDMs in the scenarios Bavaria 3.1 and 3.2 in which individual data is introduced in the form of an exit poll and a telephone survey, respectively. The slight decrease in coverage rates in the case of Germany 3.1 could be explained by the smaller interval widths, while this effect seems to be overcompensated for by the extreme decrease in the average AD in scenario Germany 3.2.

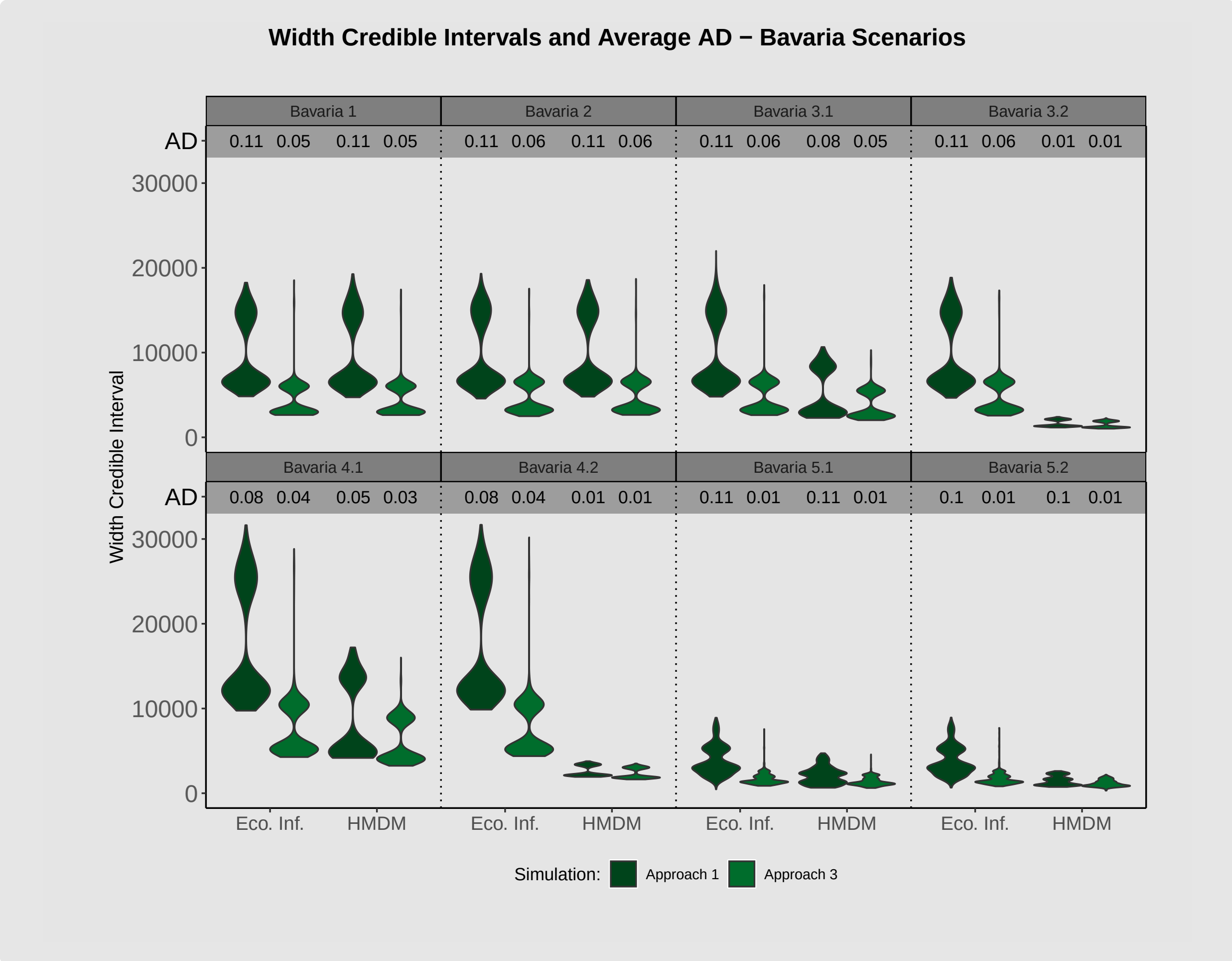


Figure 29: *Detailed evaluation of the interval widths of the 80% credible intervals for the Bavaria Scenarios with violin plots. Displayed is, for each estimated voter transition, the width of the credible interval. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 30 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The grey bar above the violin plots displays the average AD for each scenario.*

The assumption of 2000 districts instead of 800 districts in the scenarios Bavaria 4.1 and 4.2 leads to wider credible intervals. Due to the higher number of districts the overall population increases, which in turn leads to larger cell values in the estimated voter transition table. The connection between larger cell values and wider credible intervals indicates that the interval width depends up to a certain degree on the size of the cell value. Interestingly, even though the credible widths increase and the average AD decrease the coverage rates do not change substantially.

Lastly, the assumption of large differences between expected transition probabilities leads to a drastic drop in the interval widths. At the same time, the AD for models estimated with data from Approach 1 increase, which could explain the drastic drop in the coverage rates. The average AD for models estimated with data from Approach 3 even decreases to an average of 0.01, which could explain that the coverage rates are ranging slightly above or around 80%. The only exception is, again comparable to the Germany scenarios, the combination of individual data from a telephone survey and large differences between the expected transition probabilities, which is present in the HMDMs for scenario Bavaria 5.2. In this case, the interval widths drop to minimal levels leading to the low coverage rates for this scenario. Even the low average ADs of the models estimated with data from Approach 3 in the Bavaria 5.2 scenario does not seem to offset this effect.

6. Solution Proposals for the Correction of the Credible Intervals

The evaluation of the quality of the credible intervals in chapter 5 has shown that there are multiple cases in which the quality of the coverage rates is insufficient. Especially in the more complex scenarios Germany 4.1 and 4.2 as well as Bavaria 5.1 and 5.2 the coverage rates were inadequate. Thus, in the current form they do not represent an adequate representation of the estimation uncertainty. Therefore, it seems necessary to correct the given credible intervals. This chapter will present and if possible evaluate three different approaches for the correction of credible intervals on the basis of the Germany scenarios. These approaches tackle the three factors influencing the coverage rates, which were identified in chapter 5. Namely, the width of the credible intervals, the estimation error, and the data simulation approach.

Chapter 6.1 discusses an approach trying to improve the coverage rates by increasing the width of the credible intervals. Afterward, chapter 6.2 proposes the extension of Klima's parametric bootstrapping approach to the HMDM, and lastly, chapter 6.3 discusses a possible modification of the assumptions made in the (hybrid) Multinomial-Dirichlet Model.

6.1. Extention of Credible Interval Width

As discussed in the onset of Chapter 5, a decreasing width of the credible intervals can lead to a lower coverage rate. This is due to the decreasing probability that the credible interval covers the true value. The evaluation of the quality of the credible intervals in the previous chapter has shown that the average width of the credible intervals in scenarios, which produced coverages rates substantially below 80%, decreased considerably. Examples of this connection are the scenarios Germany 4.1 and 4.2, which show substantial decreases in both the width of the credible intervals as well as the coverage rates. At the same time, the evaluation has shown that the widths of the credible intervals in the first two Germany scenarios stay relatively constant while producing coverage rates scattered around the 80% mark.

It is noticeable that in the scenarios Germany 4.1 and 4.2 the uncertainty is substantially lower while the amount of data stays the same. Thus, in a first approach to the correction of the credible intervals it is assumed that the estimations have the same degree of uncertainty. This is achieved by postulating that all credible intervals have the same width, namely the width of the credible intervals of the first scenarios. The new intervals were constructed by taking the point estimate as the midpoint and applying half of the new width to each side:

$$CI_{new}^{r,c} = [\hat{E}(\theta \mid X)_{r,c} - \tfrac{1}{2}I; \hat{E}(\theta \mid X)_{r,c} + \tfrac{1}{2}I], \tag{83}$$

where $\hat{E}(\theta \mid X)_{r,c}$ stands for the Monte Carlo estimate of the posterior mean as described in chapter 2.2. In this case, it represents the points estimate for a voter transition from Party r in the first to Party c in the second election. I represents the average credible interval width across all $R{\cdot}C$ cells and E estimations of the respective first scenario:

$$I = \frac{1}{R \cdot C \cdot M} \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{m=1}^{M} I_{r,c,m} \quad , \tag{84}$$

where $I_{r,c,m}$ represents the width of the credible interval estimated for a voter transition from Party r in the first to Party c in the second election from model m. The resulting average credible interval widths used for the correction of the credible intervals from the Germany scenarios are displayed in Table 13.

Average Widths Credible Intervals - Germany 1

Approach 1		Approach 3	
Eco. Inf.	HMDM	Eco. Inf.	HMDM
5506.25	5518.094	3085.744	3083.349

Table 13: *Average widths of the credible intervals from scenario Germany 1. The credible intervals result from estimations of voter transitions, which have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference model (Eco. Inf.). 30 estimations were carried out with both models each with datasets simulated according to Approach 1 and Approach 3. The averages were taken across estimations carried out with the same model and data simulated with the same approach. A brief overview of the scenario Germany 1 can be found in appendix 6.*

It is entirely possible that this approach produces corrected credible intervals, which contain values lower than 0. In such a case the lower limit of the credible interval was set to 0, as voter transition lower than 0 are not possible. This approach hoped that the increased width of the credible intervals would lead to sufficient coverage rates scattered around the 80% mark. The resulting coverage rates are displayed in igure

30].

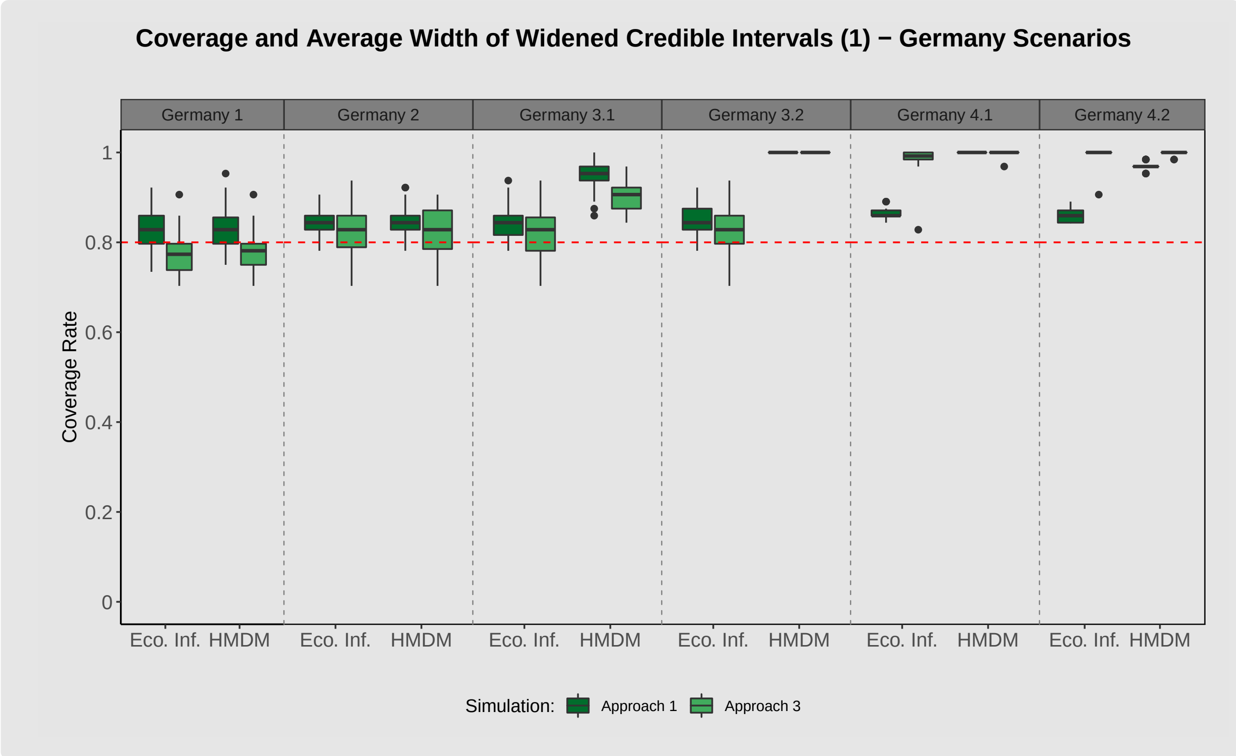


Figure 30: *Detailed evaluation of the average width credible intervals for the Germany scenarios. Displayed is, for each estimated voter transition, the share of the average width credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the average width approach introduced in chapter 6.1. The dotted red line marks a share of 80%.*

Except for the models estimated with data from Approach 3 in the scenario Germany 1 the resulting credible intervals are too conservative. In general, for the first two scenarios no major changes can be observed when comparing the coverage rates resulting from the corrected credible intervals to the initial ones. The only slight change is that the coverage rates on average move slightly closer to the 80% mark. The same can be observed for the corrected credible intervals of the EI models in the scenarios Germany 3.1 and 3.2. However, this was to be expected, since the average interval widths used for the correction roughly coincide with those from these scenarios. On the other hand, another picture emerges for the HMDMs in the scenarios Germany 3.1 and 3.2, as well as for all models in the scenarios Germany 4.1 and 4.2. Here, the widths of the initial credible intervals had shrunk substantially, and thus the correction of the credible intervals has a stronger influence in these cases. The correction leads to very conservative credible intervals with coverage rates fluctuating around the 100%. The resulting credible intervals have average widths ranging between 4600 and 5500 for models estimated with data from Approach 1 and between

2900 and 3100 for models estimated with data from Approach 3.

For the scenarios Germany 4.1 and 4.2 the cell-wise coverage rates of the corrected credible intervals were also evaluated. The results can be found in appendix 8.3. Similar to the coverage rates displayed in Figure 30 the cell-wise coverage rates are mostly too conservative. In most cases, they are at 100% or slightly below. Too conservative coverage rates could be seen as an improvement compared to the initial ones because the true value is covered 80% of the time or more but not less. Unfortunately, this level is not maintained across all cells, with some cell-wise coverage rates dropping as low as 0%.

Figure 27 provides a possible explanation for the too conservative credible intervals resulting from the first correction approach. For example, already in scenario Germany 1 the large differences between the individual interval widths are evident. In this scenario, it is assumed that there are two large parties and the non-voters, who each have a voting share of about 25% in the first election. Also, the assumption of transition probabilities with similar overall averages applies here. As a result, one-third of the cells in the estimated voter transition tables are substantially larger than the rest. This explains the apparent differences in the interval widths because the width of the credible intervals seems to depends on the size of the cell value. In order to take this aspect into account, the intervals in the next approach were not expanded across the board by the average interval width from the first scenario. Instead, the degree of expansion depended on the size of the cell value. For this purpose, the following weight was calculated for the first scenarios:

$$w = \frac{1}{R \cdot C \cdot M} \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{m=1}^{M} \left(\frac{I_{r,c,m}}{\hat{E}(\theta \mid X)_{r,c,m}} \right), \tag{85}$$

where $I_{r,c,m}$ again represents the width of the credible interval and $\hat{E}(\theta \mid X)_{r,c,m}$ the point estimate, both for an estimated voter transition from Party r in the first to Party c in the second election from model m. The resulting weights used for the correction of the credible intervals of the Germany scenarios are displayed in Table 14.

Weights - Germany 1

Approach 1		Approach 3	
Eco. Inf.	HMDM	Eco. Inf.	HMDM
0.703	0.704	0.417	0.416

Table 14: *Average weights from scenario Germany 1. The weights are calculated by weighing the width of the credible intervals with the respective point estimates. Both credible intervals and point estimates result from estimations of voter transitions, which have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference model (Eco. Inf.). 30 estimations were carried out with both models each with datasets simulated according to Approach 1 and Approach 3. The averages were taken across estimations carried out with the same model and data simulated with the same approach. A brief overview of the scenario Germany 1 can be found in appendix 6. The displayed values were rounded to the third decimal place.*

In order to calculate the corrected credible interval using the weight from equation 85 two steps are necessary. In a first step new credible interval widths are calculated by multiplying the point estimate with the respective weight in Table 14:

$$I_{new}^{r,c} = \hat{E}(\theta \mid X)_{r,c} \cdot w. \tag{86}$$

The sizes of the resulting widths are depending on the size of the point estimate. In a second step, the new credible intervals are calculated according to the following formula:

$$CI_{new}^{r,c} = [\hat{E}(\theta \mid X)_{r,c} - \frac{1}{2} I_{new}^{r,c}; \hat{E}(\theta \mid X)_{r,c} + \frac{1}{2} I_{new}^{r,c}], \tag{87}$$

where $\hat{E}(\theta \mid X)_{r,c}$ is the point estimate for an estimated voter transition from Party r in the first to Party c in the second election.

Analogous to the first correction approach, it is perfectly possible that this approach produces corrected credible intervals which contain values lower than 0. In such a case the lower limit of the credible interval was set to 0, as voter transition lower than 0 are not possible. This approach hoped that the increased width of the credible intervals would lead to less conservative ones, which still have sufficient coverage rates scattered around the 80% mark. The resulting coverage rates are displayed in Figure 31.

For the first two scenarios, the corrected credible intervals do not lead to substantially different coverage rates compared to the initial credible intervals. The same applies to the coverage rates from the EI Models in the scenarios Germany 3.1 and 3.2. Thus, similar to the first correction approach discussed in this chapter, only the coverage rates from scenarios which showed a substantial drop in the widths of the initial credible intervals are considerably affected by the correction. In the cases of the HMDMs in Scenario Germany 3.1 and 3.2 the correction leads to too conservative credible intervals with average coverage rates between

90% and 100%. Interestingly, the correction works surprisingly well in the case of the last two scenarios. While the initial credible intervals produced average coverage rates as low as 20%, in this case they move substantially closer towards the targeted 80% mark. At the same time, they are not too conservative, which was the problem with the average width credible intervals discussed earlier in this chapter. However, coverage rates for EI models, which were estimated with data from Approach 1, are with an average of about 60% still too low. The average credible width for models estimated with data from Approach 1 was 7000 and for models estimated with data from Approach 3 was 4200.

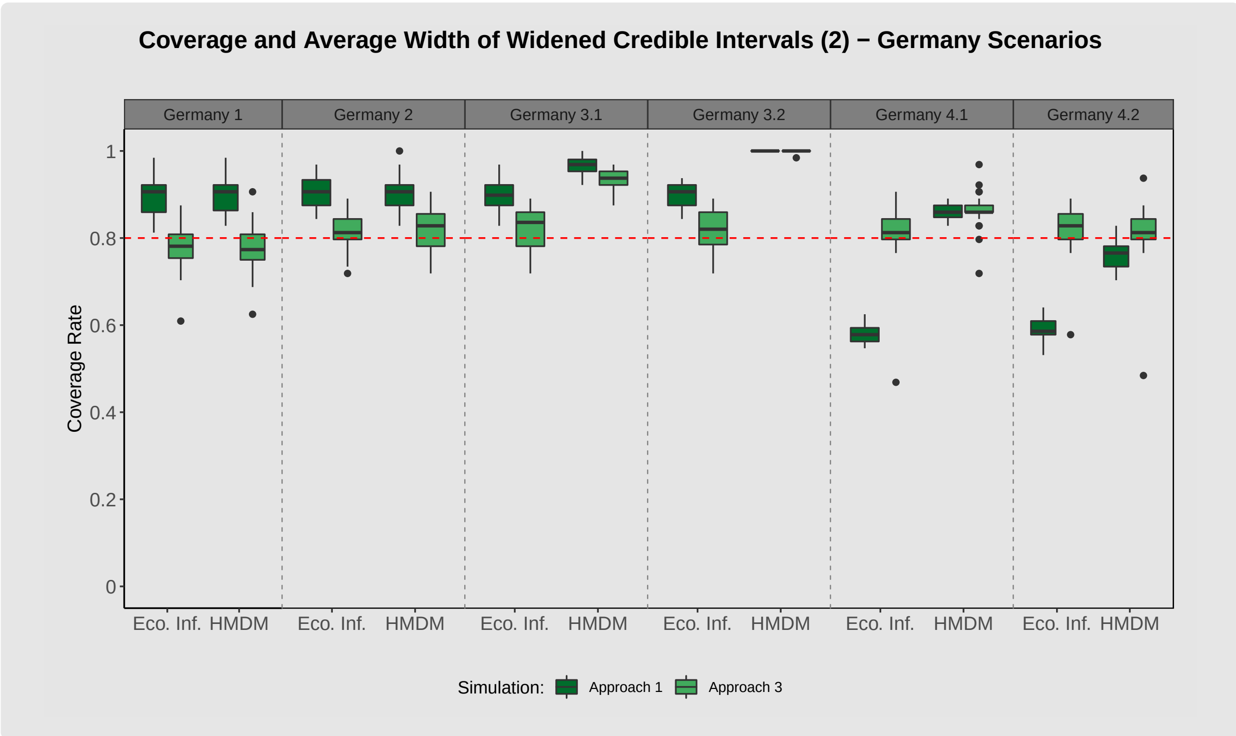


Figure 31: *Detailed evaluation of the weighted credible intervals for the Germany scenarios. Displayed is, for each estimated voter transition, the share of the weighted credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were calculated according to the weighted approach introduced in chapter 6.1. The dotted red line marks a share of 80%.*

For the scenarios Germany 4.1 and 4.2 the cell-wise coverage rates of the corrected credible intervals were also evaluated. The results can be found in appendix 8.4. As expected the cell-wise coverage rates are on average less conservative compared to those from the initial correction approach discussed in this chapter. Unfortunately, that does not mean that they are fluctuating around 80%, but that there are more cases in which the cell-wise coverage rates are considerably below 80%. The rest of the cell-wise coverage rates are still fluctuating between 90% and 100%.

In summary, it can be said that approaches have a certain potential. Besides the case of an estimation with the EI model, data from Approach 1 and a correction according to the second approacj, the problem of coverage rates substantially below 80% can be resolved. However, in most cases this results in credible intervals which are too conservative. Unfortunately, the problem still occurs that the cell-wise coverage rates do not maintain the targeted level of 80% or at least higher across all cells. Hence, while the corrected credible intervals discussed in this chapter can be seen as an improvement compared to the initial ones, they do not pose a complete solution to the problem of an inadequate quantification of the estimation uncertainty.

6.2. Bias Correction

Another factor which could play a role in the depreciating quality of the credible intervals is an increasing estimation error. This possible connection has been shown in the scenario evaluation in chapter 5. An example can be found in the Baseline scenarios when comparing the scenario with high loyalty rates and the scenario with large differences in the expected transition probabilities. While the average interval width does not show a big change, the average AD increases substantially in the case of models estimated with data from Approach 1 and marginally in the case of models estimated with data from Approach 3. In both cases the average coverage rates drop noticeably below 80%.

In his dissertation, Klima discusses the non bias-free estimation, which could explain the higher estimation error and consequently the drop in average coverage rate.[158] The same phenomenon appears in the estimations that were conducted in the course of this master thesis. Figure 32 gives an example of this bias for estimations of voter transitions with the HMDMs in the case of Scenario Germany 4.2. The graphs display the cell-wise deviations of the estimated voter transitions from the simulated ones. In the upper graph, voter transitions were estimated with data from Approach 1, while in the lower graph voter transitions were estimated with data from Approach3. The estimations of the inner cells of a voter transition table are not independent because the marginal sums are fixed. Thus, an overestimation of one cell would lead to an underestimation of the other cells and vice versa, which leads to an estimation error. This connection can be observed in the graphs of Figure 32, where for example the underestimation of loyal voters of Party 4 leads to an overestimation of the other voter transitions. In contrast to the results of Klima, where loyal voters were always underestimated, no obvious pattern can be detected. Both under- and overestimation of loyal voters can be observed. While the direction of the deviation does not differ, the size of the bias in estimations conducted with data from Approach 1 is much larger.

[158] Klima, op. cit., 107–9.

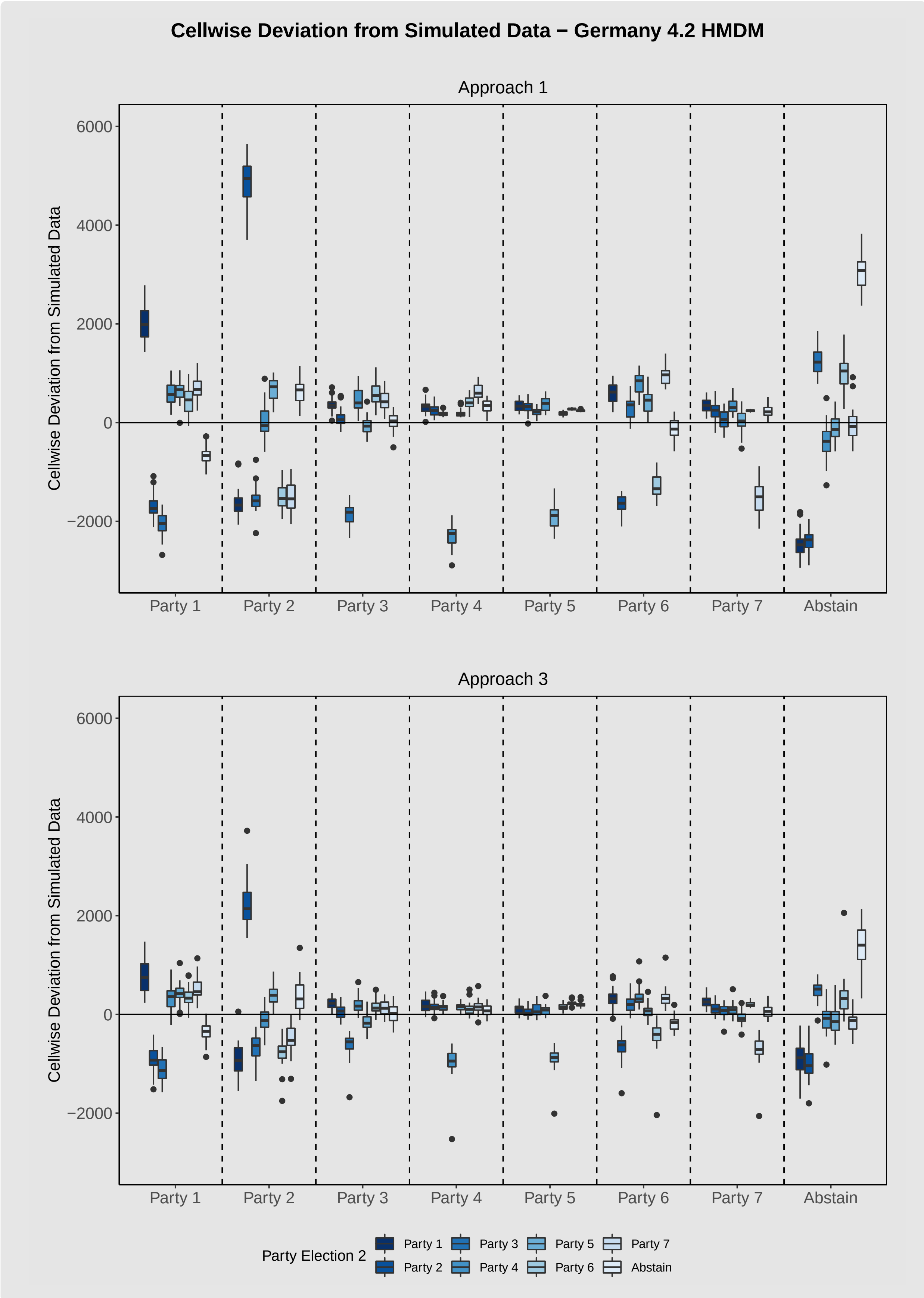


Figure 32: *Displayed are the cell-specific deviations of the estimated voter transitions from simulated voter transitions according to Approach 1 (upper graph) and Approach 3 (lower graph). The estimated voter transitions were calculated with the hybrid Multinomial-Dirichlet Model. 30 models were evaluated and the displayed values are averages across all models.*

The bias in the estimation leads to a situation in which the variability in the chains is not sufficient to construct valid credible intervals. Thus, Klima[159] proposes three different approaches for the EI model. All of them aim to estimate the bias component and then increase the variance of the chains by said component. All three approaches are loosely based on the idea of parametric bootstrapping. As discussed in the literature review, the three approaches lead to improvements in the coverage rates. However, the cell-wise coverage cannot maintain this level. In an attempt to improve the results, the following chapter will extent Klima's correction approach to the HMDMs. For this purpose, the correction on the level of the transition probabilities will be used because it has shown the most promising results in Klima's analysis.[160]

For the correction B new voter transitions are simulated. In this case, the number of bootstrap samples B was set to 10. Starting point are the Dirichlet distributions of the transitions probabilities, which have been estimated with the HMDM. From those Dirichlet distributions new voters transition probabilities are drawn for each district. Afterward, using the newly drawn voter transition probabilities new voter transitions are simulated with the help of a Multinomial distribution. In this step, the newly proposed approach will diverge from Klimas proposition. Instead of simulating entirely new voter transitions, only those not explained by the available individual data are drawn from the Multinomial distributions. Subsequently, the new aggregate data is calculated and used alongside the individual data in the estimation of B voter transitions using the HMDM.

The correction of the chains is conducted on the level of the transition probabilities because this approach has shown the most promising results in Klimas analysis. For the correction the absolute values of the differences between the estimated transition probabilities $\beta^{'}_{r,c,k}$ and those from the k-th bootstrap sample $\beta_{r,c,k}$ are calculated:

$$B_{r,c,k} = \mid \beta^{'}_{r,c,k} - \beta_{r,c,k} \mid . \tag{88}$$

For the correction of the chain variance the maximum absolute difference is taken:

$$B_{max} = \underset{r,c,k}{\text{argmax}}(B_{r,c,k}). \tag{89}$$

This value is used to transform the variance of the chains so that the resulting chain variance is equal to the intial chain variance plus the squared maximum absolute difference:

$$Var(\theta^{''}_{r,c}) = Var(\theta^{'}_{r,c}) + B^2_{max}, \tag{90}$$

[159] Ibid.
[160] Ibid., 112.

where $\theta_{r,c}$ stands for a transformed Markov Chain, which does not contain absolute voter transitions but transition probabilities. In order to adjust the variance of the transformed Markov Chains, the chain values are standardized and multiplied with the new variance. Afterward, the initial mean is added. This results in a transformed Markov Chain, which has the initial mean but the new variance.

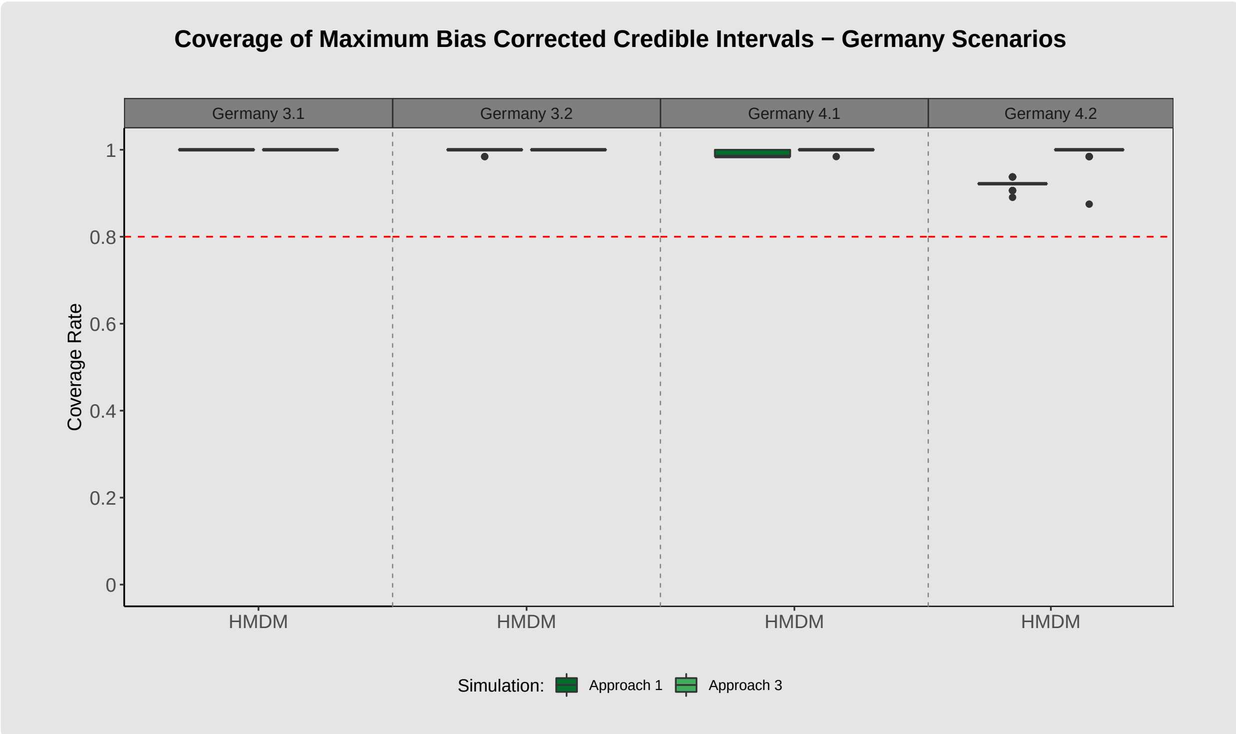


Figure 33: *Detailed evaluation of the maximum bias corrected credible intervals for the Germany scenarios. Displayed is, for each estimated voter transition, the share of the maximum bias corrected credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were corrected with the maximum bias according to the approach introduced in chapter 6.2. The dotted red line marks a share of 80%.*

Figure 33 shows the coverage rates of the maximum bias corrected credible intervals. Because the correction approach requires the presence of individual data, only the credible intervals for HMDMs in scenarios Germany 3.1-4.2 have been corrected. It stands out that the corrected credible intervals are in all cases too conservative. Apart from a few downwards outliers, models estimated with data from Approach 3 have coverage rates of 100%. The same is true for the scenarios Germany 3.1 and 3.2 and models estimated with data from Approach 1. Only the added assumption of large differences between the expected transition probabilities in the scenarios Germany 4.1 and 4.2 leads to credible intervals which are on average slightly below 100%. The resulting credible intervals have average widths ranging between 4000 and 19100 for models estimated with data from Approach 1 and between 5000 and 16800 for models estimated with data from Approach 3.

For the scenarios Germany 4.1 and 4.2 the cell-wise coverage rates of the maximum bias corrected credible intervals have also been evaluated. The results can be found in appendix 8.5. Similar to model-wise coverage rates the cell-wise ones are too conservative. In most cases the resulting coverage rates are at 100%. However, the cell-wise coverage rates only hold this level if the model has been estimated with data from Approach 3. In case the model was estimated with data from Approach 1 some cells have coverage rates considerably below the 80% mark. Nonetheless, the high coverage rates and large interval widths indicate that the correction with B_{max} is too drastic.

Therefore, in a second approach the cell-specific maximum absolute difference between the estimated transition probabilities and those from the bootstrap sample is taken to adjust the variance of the respective chains:

$$B^{r,c}_{max} = \underset{r,c,k}{\text{argmax}}(B_{r,c,k}). \tag{91}$$

The correction of the chains works analogous to equation 90, but instead of correcting each chain variance with the same value, each chains variance is corrected with a specific value. The hope behind this approach is that the correction is less drastic and results in slimmer credible intervals and coverage rates fluctuating around 80%.

Figure 34 displays the coverage rates of the credible intervals, which have been corrected according to the cell-specific correction approach. Even though the credible intervals are in most cases still too conservative, the cell-specific approach has indeed lead to less conservative credible intervals. However, the credible intervals from scenario Germany 4.2 pose a problem. In the presence of individual data from a telephone survey and large differences between the expected transition probabilities the corrected coverage rates are well below the targeted 80% mark. The resulting credible intervals have average widths ranging between 1100 and 4600 for models estimated with data from Approach 1 and between 1000 and 4100 for models estimated with data from Approach3. Compared to the initial approach these credible intervals are very thin, which could explain the drop in coverage rates.

Again, the cell-wise coverage rates have been evaluated for the scenarios Germany 4.1 and 4.2. As expected they are on average less conservative compared to the initial approach. However, there are also more cases in which the cell-wise coverage rates drop considerably below the 80% mark.

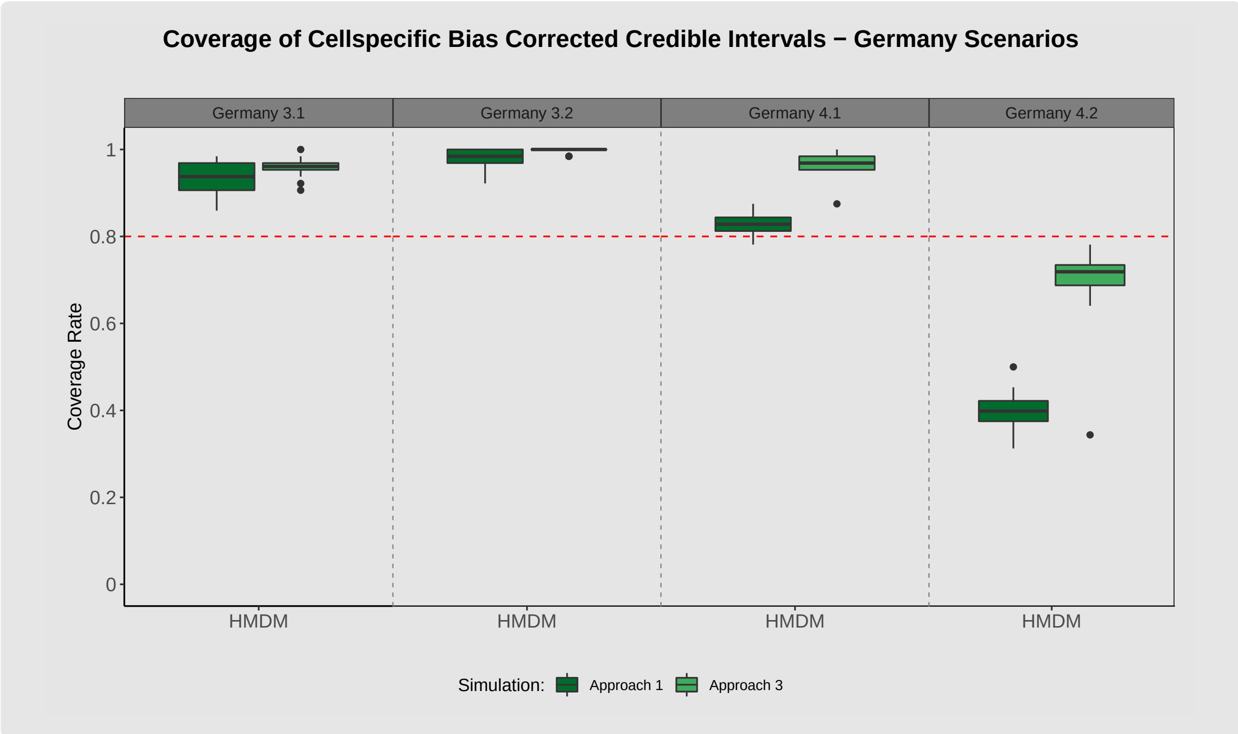


Figure 34: *Detailed evaluation of the cell-specific bias corrected credible intervals for the Germany scenarios. Displayed is, for each estimated voter transition, the share of the cell-specific bias corrected credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 15 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (dark green) and Approach 3 (light green). A brief overview of the scenarios can be found in appendix 6. The credible intervals were corrected with the cell-specific bias according to the approach introduced in chapter 6.2. The dotted red line marks a share of 80%.*

In conclusion, it can be said that the proposed correction approach does not pose an improvement compared to the one proposed by Klima. The coverage rates of the corrected credible intervals are in most too conservative, while Klima achieved coverage rates fluctuating around 80%. At the same, time the problem of cell-wise coverage rates not holding the targeted level still exists.

6.3. Modification of the Model Assumptions of the (Hybrid) Multinomial-Dirichlet Model

The evaluation in chapter 5 has shown that models estimated with data from Approach 3 generally produce higher quality credible intervals compared to models estimated with data from Approach 1. The third approach was based on the consideration that the first two approaches might violate assumptions made in the utilized models, namely the assumption that the second election results follow a Multinomial distribution. This consideration originates from the fact that the first two approaches calculate the second elections results by taking the column sums of the simulated voter transitions. In Approach 2 for example, these voter transitions are drawn from a row-wise Multinomial distribution. Therefore, it might be the case that the column sums follow a convolutional Multinomial distribution rather than a simple one. This would constitute a violation of the model assumptions. Furthermore, the evaluation in chapter 5 has shown that the quality of the credible intervals is higher if the assumption of multinomial distributed second elections results is fulfilled. The following approach aims at combining these two points.

The approach is basically a further development of Wakefield's idea of a convolutional likelihood[161], which is applied to the (hybrid-) Multinomial-Dirichlet Model. In his paper "Ecological inference for 2x2 tables" Wakefield criticizes King ecological inference model for the assumption that the second election results follow a Binomial distribution. Instead, Wakefield proposes a convolution of two binomial distributions, which has a underdispersed variance compared to the simple Binomial distribution.[162] Applied to the (hybrid-) Multinomial-Dirichlet Model this would mean that the second election results are assumed to follow a convolutional Multinomial distribution. In the context of the model, this assumption would make sense. Since the individual data is already assumed to follow a Multinomial distribution and the second election results are a sum of said individual data, it might be too restrictive to assume that the second election results also follow a Multinomial distribution. Furthermore, this approach would fulfill the two points discussed in the first paragraph.

However, Wakefield already noticed that the convolution is computationally intensive to calculate which is why he falls back on a normal approximation. The same would apply to a convolution of Multinomial distributions, which could be approximated with a multivariate Normal distribution. Even though this might be an interesting approach it was not possible to implement this in the course of this thesis due to the limited time.

7. Applied Example

In chapter 7 the solution proposals discussed in chapter 6 will be applied to a real data situation. Voter transitions between the 2013 and 2018 Bavarian state elections will be estimated with the hybrid Multinomial-Dirichlet-Model. Afterwards, credible intervals will be calculated according to the approaches discussed in chapter 6.1 and 6.2. The goal is to gain a better understanding of the resulting credible intervals in a situation with actual election data.

Data basis for the estimation are the official election results of the Bavarian state elections in 2013

[161] Please refer to chapter 3.3.1
[162] Wakefield, op. cit. at 390–1.

and 2018.[163] Additionally, individual data from the *University Study Bavarian Elections 18* (USBW) by the Ludwig-Maximilians-University Munich, University Passau and University Regensburg was used in the estimation (chapter 7.1).[164] Chapter 7.2 will shortly discuss the convergence of the model, while chapter 7.3 will present the results of the applied example.

7.1. Databasis

For the estimation of voter transitions the parties CSU, SPD, the Greens, FDP and the Left were considered for the Bavarian state elections in 2013. The 2018 Bavarian state elections will also include the AfD, which did not take part in the 2013 elections. All other parties will be grouped under the category *Others*. Furthermore, the non-voters are also considered as an alternative. This results in the estimation of an 8x9 voter migration table.

The data basis for the estimation of voter migrations are the official election results of the Bavarian state elections in 2013 and 2018. These are aggregated election results at the level of the 2056 municipalities in Bavaria. Voter migrations were only estimated on the basis of second votes, i.e. party votes. Due to e.g. migration and/or mortality, differences in the number of eligible voters in the municipalities between the two elections may occur. These differences were resolved by a proportional adjustment of the election results of the 2013 Bavarian state elections to the number of eligible voters in the 2018 Bavarian state elections.

All in all, the voter transitions are calculated for 9.479.428 eligible voters. On average each municipality has about 4611 voters. However, large structural differences between the municipalities are present and therefore large upwards outliers of up to 910.459 eligible voters (Munich) and downwards outliers as low as 149 eligible voters (Balderschwang) exist.

Table 15 displays the elections results of the 2013 and 2018 Bavarian state elections. The largest vote shares in 2013 belong to the CSU and the non-voters with over 30% each. They are followed by the SPD with about 13%, while the rest of the parties have about 5% or less of the vote shares. The vote shares of the largest three and others drop in the 2018 elections, while those of all other parties increase. The most substantial increases can be observed for the Greens and the AfD.

The second data source for the hybrid Multinomial-Dirichlet model is individual data from the university study of the Ludwig-Maximilians-University of Munich, the University of Passau and the University of Re-

[163] Statistisches Landesamt Bayern, *Landtagswahlen in Bayern*, `https://www.wahlen.bayern.de/landtagswahlen/`.
[164] Helmut Küchenhoff et al., *Universitätsstudie Bayernwahl USBW 18*, München, 2018, `https://www.stablab.stat.uni-muenchen.de/projekte/wahlforschung/usbw18/vorl_ergebnisse.pdf`.

Election Results Bavarian State Election 2013 and 2018

	CSU	SPD	FW	The Greens	FDP	AfD	The Left	Others	Non-Voters
2013	30.6%	13%	5.1%	5.3%	2.1%	/	1.3%	5.2%	37.5%
2018	26.9%	6.6%	8.1%	12.6%	3.6%	7.2%	2.3%	4.1%	28.6%

Note: This representation of the election results includes the non-voters. Therefore, the results differ from the election results discussed in chapter 4.6.

Table 15: *Election results of the Bavarian state elections in 2013 and 2018.*

gensburg. Within the course of this study, individual data were collected via telephone surveys before and after the election, as well as via an exit poll.

The selection of the interviewees for the telephone survey was conducted via a two-step random sample. Within the constituencies, so-called pre-election areas were drawn according to the *Probability Proportional to Size Design*. This ensured that the probability of a pre-election area being included in the sample was proportional to the size of the district. In a second step, telephone numbers were randomly dialed via *Random Digit Dialing* until the pre-determined sample size was reached.[165] The Exit Polls were conducted in Munich, Regensburg and Passau. Constituencies for Munich were taken from a similar survey in 2013, while those for Regensburg and Passau were randomly sampled.[166]

Election 1 \ Election 2	CSU	SPD	FW	Gruene	FDP	AfD	Die Linke	Others	Abstain	$\sum$
CSU	1094	36	197	171	61	79	5	35	88	1766
SPD	43	295	63	169	9	9	25	12	35	660
FW	17	5	74	12	3	3	0	2	6	122
Gruene	22	35	29	430	5	1	13	12	17	564
FDP	37	6	38	26	76	3	3	5	8	202
Die Linke	3	0	3	18	1	5	44	6	11	91
Others	11	3	7	16	4	11	6	51	8	117
Abstain	61	21	24	70	19	25	11	17	165	413
$\sum$	1288	401	435	912	178	136	107	140	338	3935

Table 16: *Available individual data from 304 municipalities covering voter transitions between the Bavarian state elections in 2013 and 2018. The data was collected by the Ludwig-Maximilians-University Munich, University Passau and University Regensburg in the context of the University Study Bavarian Elections USBW 18.*

All in all, the telephone survey covered approximately 3800 interviews, while approximately 9800 interviews were conducted in the context of the Exit Poll. However, only a fraction of the results could be used in the estimation. The reasons for this are manifold and include among others not usable location information collected in the telephone survey or refusal to disclose information concerning the election decision. Thus

[165] Ibid., 3.
[166] Ibid., 11.

only 3935 data points from 304 districts were available for the individual data.[167]

In each of the 304 districts on average 0.3% of the eligible voters are available in the form of individual data. However, large outliers of up to 4% are also present. This is due to the fact that the individual data includes data from a telephone survey as well as an Exit Poll. Generally, during an Exit Poll it is easier to interview more people from one particular district compared to a telephone survey. Table 17 displays the available individual data.

7.2. Model Parameters and Convergence Diagnostics

For the estimation of the voter transitions with the hybrid Multinomial-Dirichlet Model the burn-in parameter was set to 250.000, while the thinning and sample parameters were set to 1000. Those are the same parameters discussed in chapter 2.3.2 and used for the models in the course of the simulation study, which have proven to lead to a good convergence of the chains. In order to be able to evaluate the stability of the estimators, the voter transitions were estimated twice.

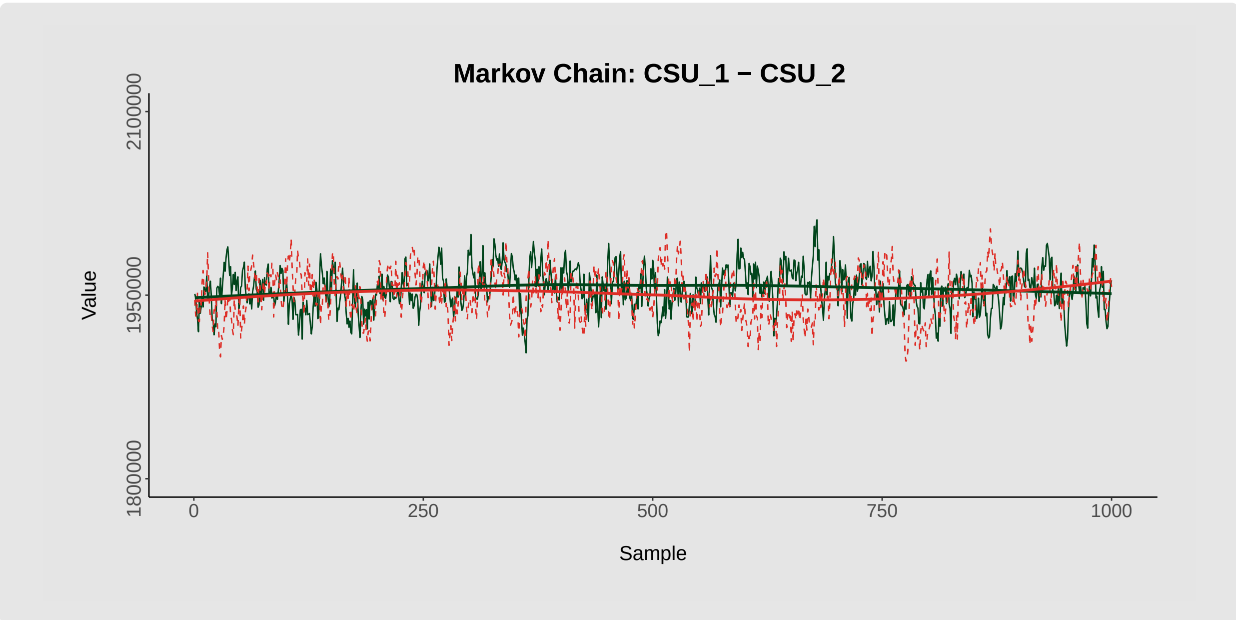


Figure 35: *Example of the evaluation of the chain convergence for the estimation of voter transitions between the Bavarian state elections in 2013 and 2018. The voter transitions have been estimated with the hybrid Multinomial-Dirichlet Model. Displayed in red and green are the Markov Chains from two separate estimations of loyal CSU voters.*

In a first step of the convergence diagnostic, the Markov Chains of the two models were compared. In this comparison it is important that the chain show no particular trend and converge to the same region. An example of this comparison is given in Figure 35. Here the chains of estimations for the number of voters

[167] Ibid., 1.

who stayed loyal to the CSU between the two elections are compared. Because the chains are very similar and do not show any particular trend it can be assumed that these chains are converged. The comparisons of the other chains show similar results.

Another aspect that can help evaluate convergence is the consideration of the AD between the two models. A small AD is seen as an indicator that the two chains have converged to the same region. The AD between the two models is 0.006. This can be seen as a further indicator that the estimators are stable. Thus, it can be assumed that the selected parameters have led to a convergence of the chains.

7.3. Results

The following chapter will discuss the results of the estimated voter transitions between the Bavarian state elections in 2013 and 2018. Furthermore, the corrected credible intervals will be presented and compared to the initial credible intervals, which have been calculated according to the Equal-Tail Approach[168]. Here the focus lies on the average width credible interval as well as the maximum bias corrected credible interval introduced in chapter 6.2 and chapter 6.3, respectively. They were chosen because they produced better results compared to the other approaches.

The results of the estimated voter transition between the Bavarian state elections in 2013 and 2018 are taken from the first estimated model. The estimated transition probabilities and voter transitions are displayed in Figure 36. The model estimates the highest loyalty rate for the voters of the Greens, who are followed by voters of the CSU and FW. In 2018 79.5% (the Greens), 67.5% (CSU) and 66.3% (FW) decided to vote for the same party again. Furthermore, the group of non-voters also has a relatively high loyalty rate with 71.1%. The Left and the FPD both have estimated loyalty rates of 61.8%. In the case of the SPD and the other parties (Others) are the estimated loyalty rates considerably lower. For voters of the SPD, the model estimated a loyalty rate of 44.1 % and for voters of other parties 35.3%. Relevant voter movements can be observed from the CSU and SPD to the Greens, from the CSU to the FW and from the non-voter to the AfD, CSU and the Greens. Furthermore, larger movements can also be observed from the other parties to the AfD and the CSU.

[168] Please refer to chapter 2.1.3.

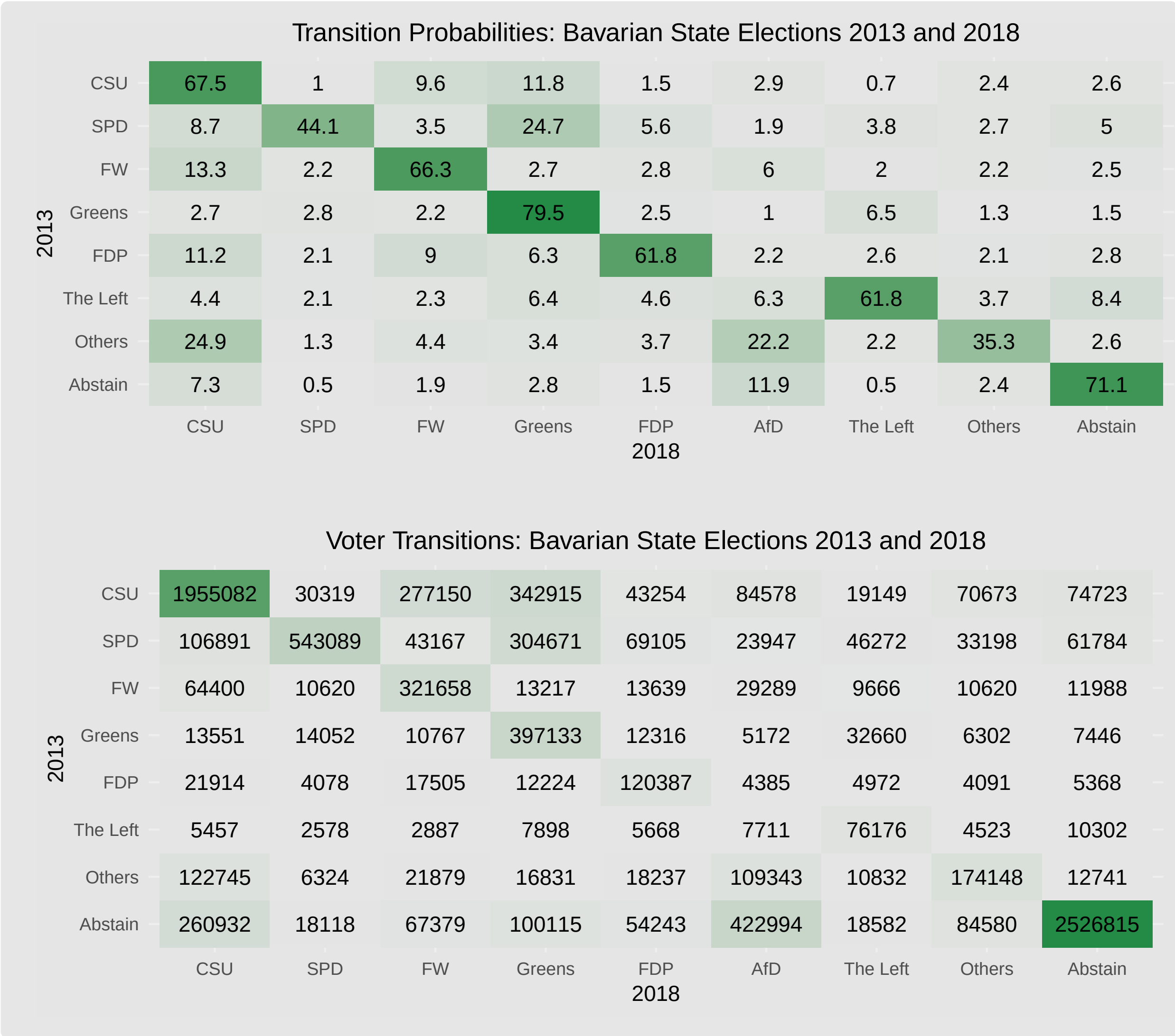


Figure 36: *Estimated voter transitions and transitions probabilities between the Bavarian state election in 2013 and 2018. The voter transitions and transition probabilities were estimated with the hybrid Multinomial-Dirichlet-Model Estimated transition probabilities are displayed in the upper graph and estimated voter transitions in the lower one. Presented are the results of the first model.*

The credible intervals shown in Figure 37 represent the status quo before this thesis. They were calculated using the Equal-Tail Approach presented in chapter 2.1.3. Based on the results discussed in chapter 5, as well as those from the evaluations by Klima[169] and Plescia et al.[170], it can be assumed that these credible intervals do not represent valid interval estimators. However, the presentation is helpful in assessing the extent of the corrections presented in this chapter.

[169] Klima, op. cit.
[170] Plescia and Sio, op. cit.

Equal Tail Credible Intervals: Lower Bounds

2013 \ 2018	CSU	SPD	FW	Greens	FDP	AfD	The Left	Others	Abstain
CSU	1933056	26555	265651	324492	35559	76231	14803	62749	63763
SPD	93119	535410	36791	285919	58381	18228	37665	25654	51616
FW	55836	8992	313317	10683	11502	21988	7995	9110	9821
Greens	10280	10992	8365	388696	8138	4123	26413	4990	5721
FDP	17480	3103	14054	9085	114488	3230	3593	3128	3725
The Left	3407	1702	2033	5836	3013	4896	71893	3080	7155
Others	110291	4863	15432	12825	12556	96798	7594	163800	8661
Abstain	239547	14619	56276	84459	42974	406620	14370	71716	2509788

Equal Tail Credible Intervals: Upper Bounds

2013 \ 2018	CSU	SPD	FW	Greens	FDP	AfD	The Left	Others	Abstain
CSU	1976586	34689	288448	362471	52006	93132	24883	78916	86215
SPD	121149	549769	49874	322781	80012	30032	54864	41329	72901
FW	73306	12370	330214	16098	15954	36214	11728	12210	14343
Greens	17060	17456	13325	405186	16923	6315	38897	8055	9425
FDP	26500	5259	21250	15810	125943	5819	6619	5212	7320
The Left	7624	3713	3859	10093	8718	10767	80389	6239	13786
Others	134954	7871	28328	21301	24055	121636	14309	183945	17426
Abstain	282292	22426	78675	116435	65897	439157	24264	97819	2543401

Figure 37: *80% credible intervals resulting from the estimation of voter transitions between the Bavarian state elections in 2013 and 2018. The voter transitions have been estimated with the hybrid Multinomial-Dirichlet Model and the credible intervals were calculated according to the Equal-Tail approach introduced in chapter 2.1.3. The upper table displays the lower bounds and the lower table displays the upper bounds of the credible intervals. Presented are the results from the first model.*

In order to construct the average width credible intervals for the Germany scenarios, the average widths of the credible intervals resulting from the first scenario were taken. The key difference between the situation found in the first scenario and the situation of the applied example lies in the transition probabilities. While the varying transition probabilities with similar overall averages are assumed in the scenario Germany 1, the results in the upper graph of Figure 36 show that large differences between the transition probabilities exist in the applied example. Hence, new data, which replicated the situation found in the scenario Germany 1, was simulated. That means that transition probabilities with similar overall averages were assumed and new voter transitions, as well as second election results, were simulated based on the election results of the Bavarian state elections in 2013. Afterward, voter transitions were estimated with the HMDM. The average widths of the resulting credible intervals were then taken to construct the average width credible intervals for the estimated voter transitions between the Bavarian state elections in 2013 and 2018.

Average Width Credible Intervals: Lower Bounds

2013 \ 2018	CSU	SPD	FW	Greens	FDP	AfD	The Left	Others	Abstain
CSU	1940565	15802	262633	328398	28737	70061	4632	56156	60206
SPD	92374	528572	28650	290154	54588	9430	31755	18681	47267
FW	49883	0	307141	0	0	14772	0	0	0
Greens	0	0	0	382616	0	0	18143	0	0
FDP	7397	0	2988	0	105870	0	0	0	0
The Left	0	0	0	0	0	0	61659	0	0
Others	108228	0	7362	2314	3720	94826	0	159631	0
Abstain	246415	3601	52862	85598	39726	408477	4065	70063	2512298

Average Width Credible Intervals: Upper Bounds

2013 \ 2018	CSU	SPD	FW	Greens	FDP	AfD	The Left	Others	Abstain
CSU	1969599	44836	291667	357432	57771	99095	33666	85190	89240
SPD	121408	557606	57684	319188	83622	38464	60789	47715	76301
FW	78917	25137	336175	27734	28156	43806	24183	25137	26505
Greens	28068	28569	25284	411650	26833	19689	47177	20819	21963
FDP	36431	18595	32022	26741	134904	18902	19489	18608	19885
The Left	19974	17095	17404	22415	20185	22228	90693	19040	24819
Others	137262	20841	36396	31348	32754	123860	25349	188665	27258
Abstain	275449	32635	81896	114632	68760	437511	33099	99097	2541332

Figure 38: *Corrected 80% credible intervals resulting from the estimation of voter transitions between the Bavarian state elections in 2013 and 2018. The voter transitions have been estimated with the hybrid Multinomial-Dirichlet Model and the corrected credible intervals were calculated according to the Average Width approach introduced in chapter 6.1. The upper table displays the lower bounds and the lower table displays the upper bounds of the corrected credible intervals. Values lower than 0 are set to 0. Presented are the results from the first model.*

The resulting credible intervals are displayed in Figure 38. In general, the average width credible intervals are wider compared to those resulting from the Equal-Tail Approach. This is particularly noticeable in that in almost one-third of the cases the lower limit of the corrected credible intervals is 0, while this does not occur once in the case of the initial credible intervals.

Maximum Bias Corrected Credible Intervals: Lower Bounds

2013 \ 2018	CSU	SPD	FW	Greens	FDP	AfD	The Left	Others	Abstain
CSU	1500195	0	0	0	0	0	0	0	0
SPD	0	348964	0	113692	0	0	0	0	0
FW	0	0	246760	0	0	0	0	0	0
Greens	0	0	0	320855	0	0	0	0	0
FDP	0	0	0	0	88437	0	0	0	0
The Left	0	0	0	0	0	0	56665	0	0
Others	47173	0	0	0	0	32269	0	95326	0
Abstain	0	0	0	0	0	0	0	0	1971848

Maximum Bias Corrected Credible Intervals: Upper Bounds

2013 \ 2018	CSU	SPD	FW	Greens	FDP	AfD	The Left	Others	Abstain
CSU	2399209	440904	715234	794436	517893	512334	471375	500429	509100
SPD	295797	711927	234468	489109	254159	210247	230713	224359	252038
FW	141115	83287	398483	90243	86357	103318	85190	80691	84650
Greens	90198	92126	88182	469937	91936	78842	108836	86717	82041
FDP	52184	31531	47950	44181	150477	36091	32235	30215	34782
The Left	24330	22055	21472	27414	25941	27512	95371	24726	30225
Others	196822	77296	95890	95799	92116	184866	86100	248778	91131
Abstain	774506	557567	602277	626347	595169	962657	622512	650599	3067383

Figure 39: *Corrected 80% credible intervals resulting from the estimation of voter transitions between the Bavarian state elections in 2013 and 2018. The voter transitions have been estimated with the hybrid Multinomial-Dirichlet Model and the corrected credible intervals were calculated according to the Maximum Bias correction approach introduced in chapter 6.2. The upper table displays the lower bounds and the lower table displays the upper bounds of the corrected credible intervals. Values lower than 0 are set to 0. Presented are the results of the first model.*

Interestingly, not all corrected credible intervals are wider. In the case of loyal voters of the CSU and voters who did not vote in both elections, the corrected credible intervals are smaller. In general, it can be observed that the smaller the point estimate, the wider the corrected credible interval becomes compared to the initial interval. This was to be expected and is related to the phenomenon discussed in chapter 5.2 according to which the width of the credible interval depends on the size of the point estimates. Hence, since in this case the average interval width is used for correcting the resulting credible intervals are substantially wider.[171] The wide corrected credible intervals implicate substantial insecurity in the estimation.

[171] For this reason the Approach of the weighted credible intervals was formulated. However, the evaluation showed that the resulting credible intervals were not valid in all cases. Thus, they will not be applied here

However, the resulting credible intervals are not entirely uninformative. For example, it can be deduced that a six figured number of voters who abstained in the 2013 elections voted for the AfD in 2018.

The maximum bias corrected credible intervals are displayed in Figure 39. Similar to the results by Klima the maximum bias corrected credible intervals are extremely large. The comparison with the initial credible intervals as well as the average width credible intervals shows that the maximum bias credible intervals are substantially larger. In almost all cases the lower bound is 0. Exceptions are the credible intervals of the loyal voters as well as cases in which substantial voter movements can be observed. Examples for this are from other parties to the CSU and AfD as well as from the SPD to the Greens. The upper bounds are at least five and in most cases six-figure numbers, which further exemplifies how wide the corrected credible intervals are.

8. Conclusion and Outlook

Two objectives were pursued in the context of this Master's thesis. The first being to analyze the reasons for the poor quality of the credible intervals resulting from the (hybrid) Multinomial-Dirichlet model. At the beginning of this master thesis, the following hypothesis was formulated - the coverage rates in simple data scenarios are valid and that only an increased complexity in the data leads to a quality deterioration of the credible intervals. A second objective was to formulate possible solution proposals for the correction and hence, construction of valid credible intervals.

In order to analyze the reasons for the poor quality of the credible intervals, a simulation study was presented in chapter 4. The primary aim of it was to investigate how expected variations and complexity in the data affect the quality of the credible intervals. For this purpose, the six variables: number of parties, number of districts, population size, vote shares, transition probabilities and the sampling scheme for individual data were defined and different variable characteristics specified. Within the context of the Baseline scenarios, their effects on the quality of the credible intervals were then examined. In the Germany and Bavaria scenarios, the effects of combinations of some variable characteristics on the quality of the credible intervals were investigated.

A further goal was to investigate how possible violations of the model assumptions of the (hybrid) Multinomial-Dirichlet model during the simulation study affect the quality of the credible intervals. Hence, three different simulation approaches were formulated. In Approach 1 and Approach 2 voter transitions are regarded as a statistical component. In the former, transition probabilities are drawn from a Dirichlet distribution and multiplied with the election results of the first election to calculate voter transitions. In order to follow the model assumptions more closely, the second approach in which the voter transitions are drawn directly from a multinomial distribution was created. In both cases, the election results of the second election are calculated by adding up the voter transitions. The summation could, however, lead to a situation in which the results of the second choice follow a convolutional distribution. Since this would violate the assumption of multinomial distributed election results of the second election made in the (hybrid) Multinomial-Dirichlet model, a third simulation approach was constructed. In this approach, the election results of the second election are considered as the statistical component and therefore, are directly drawn from a Multinomial distribution. After which, the now deterministic voter transitions are calculated.

In chapter 5, the credible intervals resulting from the model estimated in the context of the Baseline, as well as the Germany and Bavaria scenarios were evaluated. One result of the evaluation can be seen in the confirmation of the hypothesis that valid credible intervals can be calculated in simple data scenarios. Also, it is shown that the addition of large differences between the expected transition probabilities is the only case that leads to coverage rates of well below 80%. This evaluation also exhibited that the assumptions of fixed election shares and transition probabilities along with the presence of one or two large parties and the addition of individual data from a telephone survey, lead to conservative credible intervals with coverage rates well above 80%. At the same time, the evaluation of the Germany and Bavaria scenarios showcased that these effects are more than offset if the assumption of large differences between the expected transition probabilities is added.

The interval widths and the estimation error could be identified as reasons for the poor quality of the credible intervals. For example, the credible intervals shrunk substantially due to the addition of the assumption of large differences between the expected transition probabilities. Simultaneously, the estimation error in the form of the AD also increases. This combination considerably reduces the coverage rates. In the cases of conservative credible intervals, opposing effects can be observed. The evaluation in chapter 5 has also shown that an explicit adherence to the assumption of multinomially distributed second election results, as is the case in Approach 3, leads to substantially better but not perfect results.

In chapter 6, possible correction approaches for the credible intervals are discussed and evaluated on the basis of the Germany scenarios. They originate in the three identified factors that influence the quality of the credible intervals: the width of the credible intervals, the estimation error and the utilized simulation approach. Since the interval widths become very small in scenarios with coverage rates below 80%, the first proposed solutions revolve around extending them. In a first approach, the average interval widths of the first Germany scenario, which produces valid credible intervals, are used to construct corrected credible intervals. This approach leads to very conservative intervals, especially in the more complex scenarios. The reason for this is the fact that the interval widths depend on the size of the point estimates. Hence, in many cases an overcorrection occurs. Therefore, in a second approach, the interval widths of the first Germany scenarios were weighted with the point estimates. After that, the average of the resulting weights is used to form corrected credible intervals. This led to less conservative interval which in some cases, however, had coverage rates well below 80%.

Similar to Klima, the model estimations conducted in this thesis are not bias free, leading to an estimation error. Thus, another innovation is the extension of Klima's bias correction approach to the hybrid Multinomial-Dirichlet model. Subsequently, the credible intervals are corrected with the maximum global bias and the maximum cell-specific bias. The former leads to very conservative credible intervals with coverage rates that are in almost all cases 100%. The second approach is less conservative but leads to coverage rates that are in some cases well below 80%.

However, the problem of cell-wise coverage rates, which could not maintain the level of the model-wise ones, still exists in all proposed correction approaches. Hence, further research in this area is necessary. A first approach could be to attempt to identify noticeable patterns and use this knowledge for cell-specific correction approaches. The third proposed approach also requires further research. The coverage rates improve substantially if the distribution assumption of the second election results is adhered to. Also, considering that the second election results in case of multinomially distributed voter migrations could follow a convolution Multinomial distribution, an extension of Wakefields's approach of a convolutional likelihood for the (hybrid) Multinomial-Dirichlet model is proposed. Since the calculation of a convolutional Multinomial likelihood is computationally demanding and therefore difficult to implement, a first approach could be the approximation with a Multinomial Normal distribution.

In order to get a gauge for the proposed correction procedures, the procedure based on the average interval widths and the maximum global bias correction was applied to an applied example in the final chapter. For this purpose, real voter transitions between the Bavarian state elections of 2013 and 2018 were estimated using the hybrid Multinomial-Dirichlet model. The official election results are freely available, while the used individual data was taken from a telephone survey and an exit poll conducted in the course of the USBW. The application example shows that the corrected credible intervals are much wider than the initial ones, which implies a large uncertainty in the estimation.

References

Allport, Floyd H., "The Group Fallacy in Relation to Social Science", *American Journal of Sociology*, 29/6 (1924), 688–706. doi: 10.1086/213647.

Cho, Wendy Tam and Manski, Charles F., *Cross–Level/Ecological Inference* (The Oxford Handbook of Political Methodology; Oxford University Press, 2009). doi: 10.1093/oxfordhb/9780199286546.003.0024.

Duncan, Otis Dudley and Davis, Beverly, "An Alternative to Ecological Correlation", *American Sociological Review*, 18/6 (1953), 665. doi: 10.2307/2088122.

Freedman, D. A. et al., "A Solution to the Ecological Inference Problem", *Journal of the American Statistical Association*, 93/444 (1998), 1518. doi: 10.2307/2670067.

Gamerman, Dani and Lopes, Hedibert Freitas, *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference* (Texts in statistical science series; 2. ed., Boca Raton Fla. u.a.: Chapman & Hall/CRC, 2006).

Gehlke, C. E. and Biehl, Katherine, "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material", *Journal of the American Statistical Association*, 29/185 (1934), 169. doi: 10.2307/2277827.

Geman, Stuart and Geman, Donald, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6/6 (1984), 721–41. doi: 10.1109/TPAMI.1984.4767596.

Gill, Jeff, *Bayesian Methods: A Social and Behavioral Sciences Approach, Third Edition* (Chapman & Hall / CRC Statistics in the Social and Behavioral Sciences; 3rd ed., Hoboken: CRC Press, 2015), http://gbv.eblib.com/patron/FullRecord.aspx?p=1598081.

Glynn, Adam and Wakefield, Jon, "Ecological Inference in the Social Sciences", *Statistical methodology*, 7/3 (2010), 307–22. doi: 10.1016/j.stamet.2009.09.003..

Goodman, Leo A., "Ecological Regressions and Behavior of Individuals", *American Sociological Review*, 18/6 (1953), 663. doi: 10.2307/2088121.

Goodman, Leo A., "Some Alternatives to Ecological Correlation", *American Journal of Sociology*, 64/6 (1959), 610–25. doi: 10.1086/222597.

Greiner, D. James and Quinn, Kevin M., "Exit polling and racial bloc voting: Combining individual-level and R×C ecological data", *The Annals of Applied Statistics*, 4/4 (2010), 1774–96. doi: 10.1214/10-AOAS353.

Gschwend, Thomas, "Ökologische Inferenz", *Methoden der Politikwissenschaft: neuere qualitative und quantitative Analyseverfahren* (2006), 227–37.

Hastings, W. K., "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, 57/1 (1970), 97. doi: 10.2307/2334940.

Held, Leonhard and Sabanés Bové, Daniel, *Applied Statistical Inference* (Berlin, Heidelberg: Springer Berlin Heidelberg, 2014). doi: 10.1007/978-3-642-37887-4.

James Greiner, D. and Quinn, Kevin M., "R $\times$ C ecological inference: Bounds, correlations, flexibility and transparency of assumptions", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172/1 (2009), 67–81. doi: 10.1111/j.1467-985X.2008.00551.x.

Jones, Peter W. and Smith, Peter, *Stochastic processes: An introduction* (Chapman & Hall/CRC Texts in Statistical Science Series; Third edition, Boca Raton, London, and New York: CRC Press, Taylor & Francis Group, 2018).

King, Gary, *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data* (Princeton, NJ: Princeton University Press, 1997). doi: 10.2307/j.ctt46n43p, http://www.jstor.org/stable/10.2307/j.ctt46n43p.

King, Gary, Rosen, O. R.I., and Tanner, Martin A., "Binomial-Beta Hierarchical Models for Ecological Inference", *Sociological Methods & Research*, 28/1 (1999), 61–90. doi: 10.1177/0049124199028001004.

King, Gary, Rosen, Ori, and Tanner, Martin A., *Ecological Inference* (Cambridge: Cambridge University Press, 2004). doi: 10.1017/CBO9780511510595.

Klima, André, "Ökologische Inferenz und hybride Modelle: Schätzung der Wählerwanderung in Mehrparteiensystemen", Dissertation (München: Ludwig-Maximilians-Universität, 2016).

Klima, André et al., "Combining Aggregate Data and Exit Polls for the Estimation of Voter Transitions", *Sociological Methods & Research*, 1 (2017), 004912411770147. doi: 10.1177/0049124117701477.

Klima, André et al., "Estimation of voter transitions based on ecological inference: An empirical assessment of different approaches", *AStA Advances in Statistical Analysis*, 100/2 (2016), 133–59. doi: 10.1007/s10182-015-0254-8.

Klima, André et al., *Exit Polls und Hybrid-Modelle* (Wiesbaden: Springer Fachmedien Wiesbaden, 2017). doi: 10.1007/978-3-658-15674-9.

Kramer, Gerald H., “The Ecological Fallacy Revisited: Aggregate- versus Individual-level Findings on Economics and Elections, and Sociotropic Voting”, *American Political Science Review*, 77/01 (1983), 92–111. doi: 10.2307/1956013.

Küchenhoff, Helmut et al., *Universitätsstudie Bayernwahl USBW 18*, München, 2018, https://www.stablab.stat.uni-muenchen.de/projekte/wahlforschung/usbw18/vorl_ergebnisse.pdf.

Lau, Olivia, Moore, Ryan T., and Kellermann, Kellermann, *eiPack: eiPack: Ecological Inference and Higher-Dimension Data*, 2012, https://CRAN.R-project.org/package=eiPack.

Lau, Olivia, Moore, Ryan T., and Kellermann, Michael, *eiPack: Ecological Inference and Higher-Dimension Data Management*, 2012, https://cran.r-project.org/web/packages/eiPack/eiPack.pdf.

Lau, Olivia, Ryan T. Moore, and Michael Kellermann, “eiPack: Ecological Inference and Higher-Dimension Data Management”, *R News*, 7(2) (2007), 43–7.

Lee, Peter M., *Bayesian statistics: An introduction* (4. ed., 1. publ, Chichester: Wiley, 2012).

Meeker JR, William Q., Hahn, Gerald J., and Escobar, Luis A., *Statistical intervals: A guide for practitioners and researchers* (2nd ed., Hoboken: John Wiley et sons, 2017).

Metropolis, Nicholas et al., “Equation of State Calculations by Fast Computing Machines”, *The Journal of Chemical Physics*, 21/6 (1953), 1087–92. doi: 10.1063/1.1699114.

Müller, Peter, *Alternatives to the Gibbs Sampling Scheme* (1992).

Norris, James R., *Markov chains* (Cambridge series on statistical and probabilistic methods; 15. printing, Cambridge: Cambridge Univ. Press, 2009).

Ntzoufras, Ioannis, *Bayesian modeling using WinBUGS* (Wiley series in computational statistics; Hoboken, NJ: Wiley, 2009). doi: 10.1002/9780470434567, http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10296519.

Ogburn, William F. and Goltra, Inez, “How Women Vote”, *Political Science Quarterly*, 34/3 (1919), 413. doi: 10.2307/2141684.

Openshaw, Stan, *The modifiable areal unit problem* (Concepts and techniques in modern geography; Norwich: Geo Abstracts Univ. of East Anglia, 1984).

Plescia, Carolina and Sio, Lorenzo de, “An evaluation of the performance and suitability of R $\times$ C methods for ecological inference with known true values”, *Quality & quantity*, 52/2 (2018), 669–83. doi: 10.1007/s11135-017-0481-z.

Robert, Christian P., *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (Springer texts in statistics; New York, NY: Springer, 2007). doi: 10.1007/0-387-71599-1, http://dx.doi.org/10.1007/0-387-71599-1.

Robert, Christian P. and Casella, George, *Monte Carlo Statistical Methods* (New York, NY: Springer New York, 2004). doi: 10.1007/978-1-4757-4145-2.

Robinson, W. S., "Ecological Correlations and the Behavior of Individuals", *American Sociological Review*, 15/3 (1950), 351. doi: 10.2307/2087176.

Rosen, Ori et al., "Bayesian and Frequentist Inference for Ecological Inference: The RxC Case", *Statistica Neerlandica*, 55 (2001), 134–56.

Rossi, Peter E., Allenby, Greg M., and McCulloch, Robert, *Bayesian Statistics and Marketing* (Chichester, UK: John Wiley & Sons, Ltd, 2005). doi: 10.1002/0470863692.

Schlesinger, Thomas, *Ecological Inference with individual and aggregate data*, 2014, https://cran.r-project.org/web/packages/eiwild/eiwild.pdf.

Schlesinger, Thomas, *eiwild: Ecological Inference with individual and aggregate data*, 2014, https://CRAN.R-project.org/package=eiwild.

Schlesinger, Thomas, "Kombination von Aggregat- und Individualdaten bei der Analyse von RxC-Tafeln: Neue Implementierung in R", Masterarbeit (München: Ludwig-Maximilians-Universität, 2013).

Selvin, Hanan C., "Durkheim's Suicide and Problems of Empirical Research", *American Journal of Sociology*, 63/6 (1958), 607–19. doi: 10.1086/222356.

Smith, Adrian Frederick Melhuish, Cox, David Roxbee, and Titterington, D. M., "Bayesian computational methods", *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337/1647 (1991), 369–86. doi: 10.1098/rsta.1991.0130.

Statistisches Landesamt Bayern, *Landtagswahlen in Bayern*, https://www.wahlen.bayern.de/landtagswahlen/.

Thomsen, Søren R., *Danish elections 1920 - 1979: A logit approach to ecolog. analysis and inference: Zugl.: Åarhus, Univ., Diss., 1987* (1. udg., 1. opl, Åarhus: Politica, 1987).

Wakefield, Jon, "Ecological inference for 2 x 2 tables", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167/3 (2004), 385–425. doi: 10.1111/j.1467-985x.2004.02046_1.x.

Wakefield, Jonathan, "Ecologic studies revisited", *Annual review of public health*, 29 (2008), 75–90. doi: 10.1146/annurev.publhealth.29.020907.090821.

Won-ho Park, “Ecological Inference and Aggregate Analysis of Elections”, Dissertation (Michigan: The University of Michigan, 2008).

A. Appendix

A.1. Metropolis-within-Gibbs Sampler

1. Chose number of iterations by defining Burn-In, Sample-Size and Thinning
2. Choose starting values (β_i^{rc},$\alpha_{rc})^{(0)}$ with ($i = 1, ..., P$), ($r = 1, ..., R$) and ($c = 1, ..., C$)
3. Start of the algorithm with iteration $j = 1$:
   a) Successively simulate $(\beta_j^{rc})^{(j)}$ with Full Conditional f_β for β_j^{rc}:
   $f_\beta(\beta_j^{rc}) \propto (\beta_j^{rc})^{Z_i^{rc}} \times (\theta_{c,i})^{(T_{c,i}-Z_{c,i})} \times (\beta_i^{rc})^{\alpha_{rc}-1}$
      i. Draw $(\beta_i^{rc})^*$ from $h_\beta \sim N\left((\beta_i^{rc})^{(j-1)}, \sigma^2_{\beta_i^{rc}}\right)$
      ii. Calculate $\gamma = min\left\{1, \frac{f_\beta\left((\beta_i^{rc})^*\right)}{f_\beta\left((\beta_i^{rc})^{(j-1)}\right)}\right\}$
      iii. Set $(\beta_i^{rc})^{(j)} = (\beta_i^{rc})^*$ with probability γ
   b) Successively simulate $(\alpha_{rc})^{(j)}$ with Full Conditional f_α of α_{rc}:
   $f_\alpha(\alpha_{rc}) \propto \frac{\Gamma(\sum_{c'=1}^{C} \alpha_{rc'})}{\Gamma(\alpha_{rc})} \times \prod_{i=1}^{P} (\beta_i^{rc})^{(\alpha_{rc}-1} \times \alpha_{rc}^{\lambda_1^{rc}-1} \times exp(-\lambda_2^{rc}\alpha_{rc})$
      i. Draw $(\alpha_{rc})^*$ from $h_\alpha \sim N\left((\alpha_{rc})^{(j-1)}, \sigma^2_{\alpha_{rc}}\right)$
      ii. Calculate $\gamma = min\left\{1, \frac{f_\alpha((\alpha_{rc})^*)}{f_\alpha\left((\alpha_{rc})^{(j-1)}\right)}\right\}$
      iii. Set $\alpha_{rc}^{(j)} = (\alpha_{rc})^*$ with probability γ
   c) $(\beta_i^{rc}, \alpha_{rc})^{rc}$ will be saved with following conditions:
   - j > Burn-In
   - Rest of $\frac{j}{Thinning} = 0$
4. Result is Markov Chain $\left((\beta_i^{rc}, \alpha_{rc})^{(1)}, ..., (\beta_i^{rc}, \alpha_{rc})^{(S)}\right)$

Figure 40: *Metropolis-within-Gibbs-Sampler in Pseudocode. (Klima 2017, op. cit., Digital Appendix)*

A.2. Transition Probabilities of Approach 2 and Approach 3

A.2.1. Approach 2

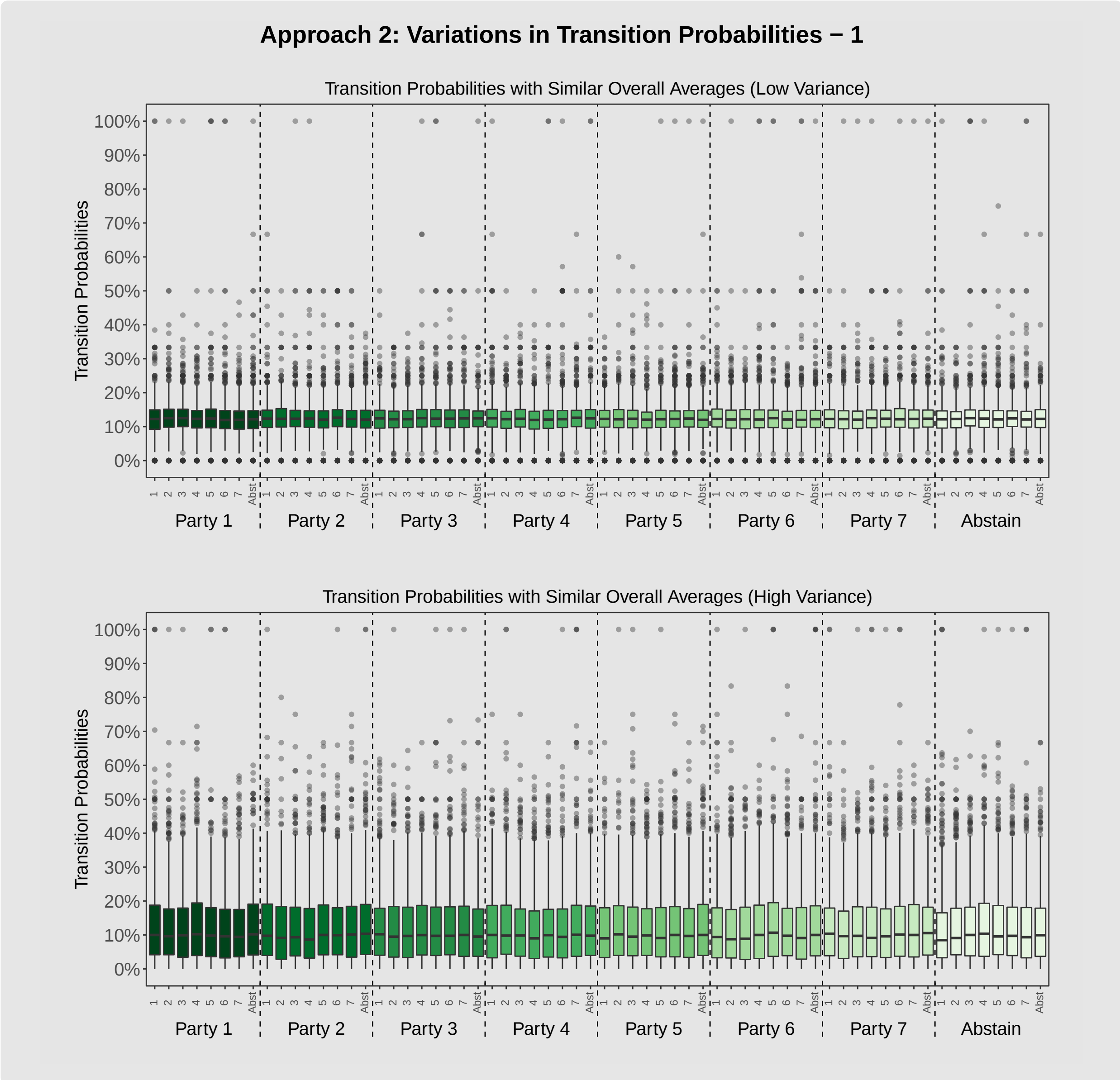


Figure 41: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of varying loyalty and transition probabilities with similar overall averages with a low variance (upper graph) and a high variance probabilities (lower graph) from one simulation with Approach 2.*

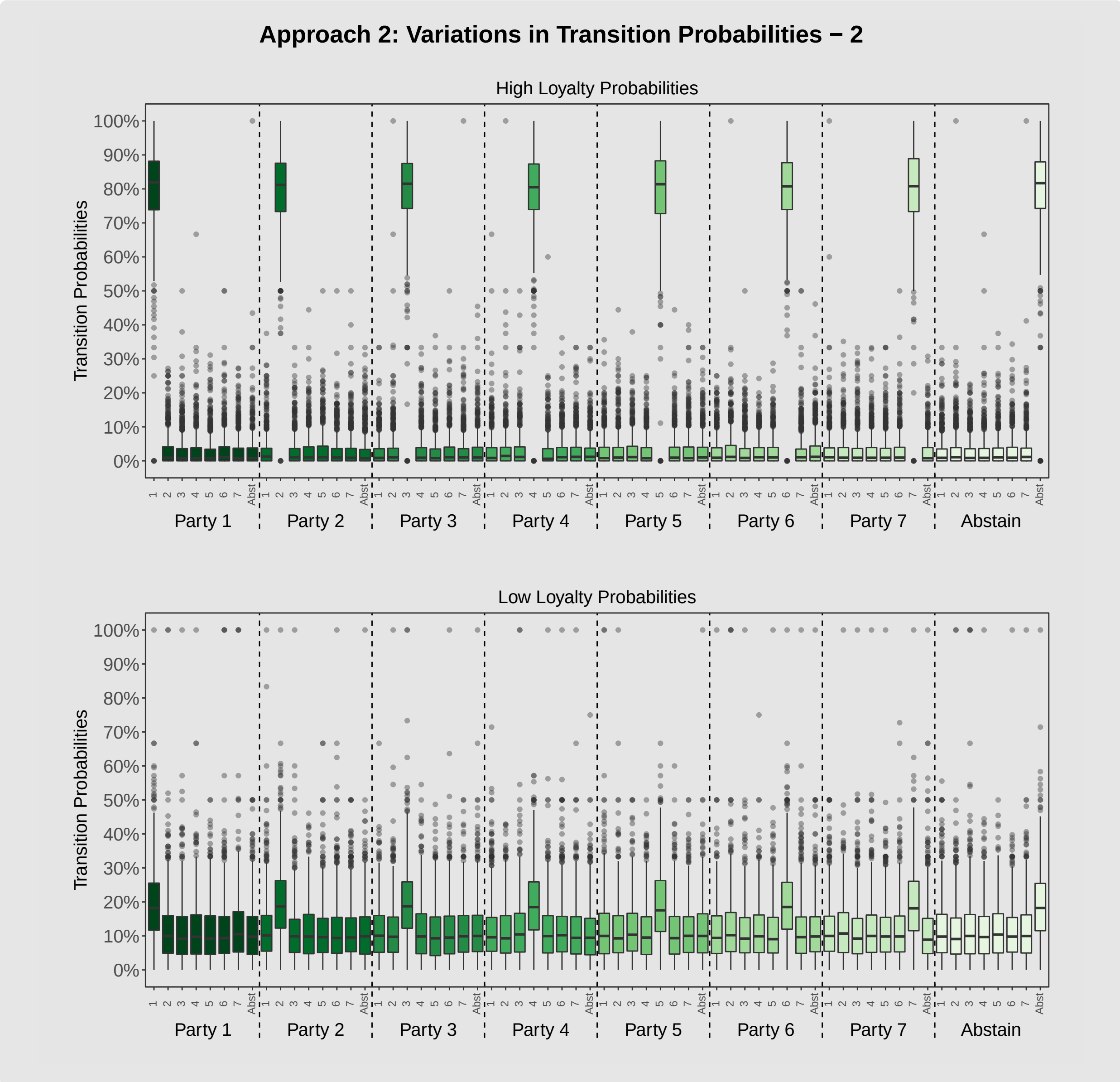


Figure 42: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of high loyalty probabilities (upper graph) and low loyalty probabilities (lower graph) from one simulation with Approach 2.*

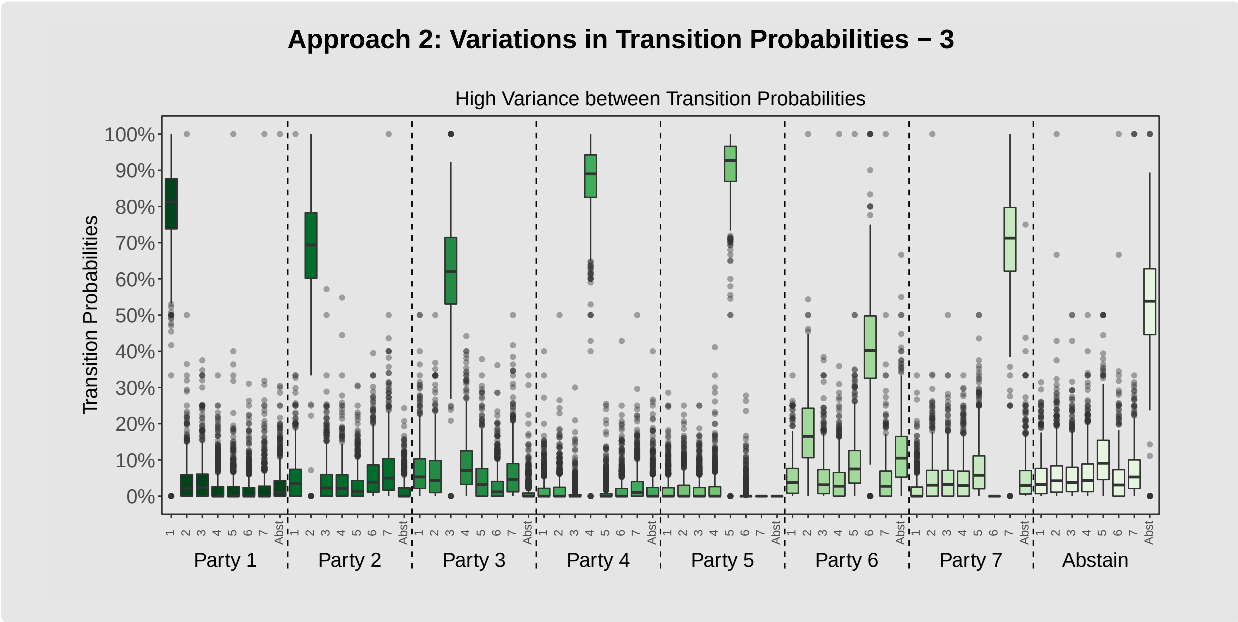


Figure 43: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of large differences between the expected transition probabilities from one simulation with Approach 2.*

A.2.2. Approach 3

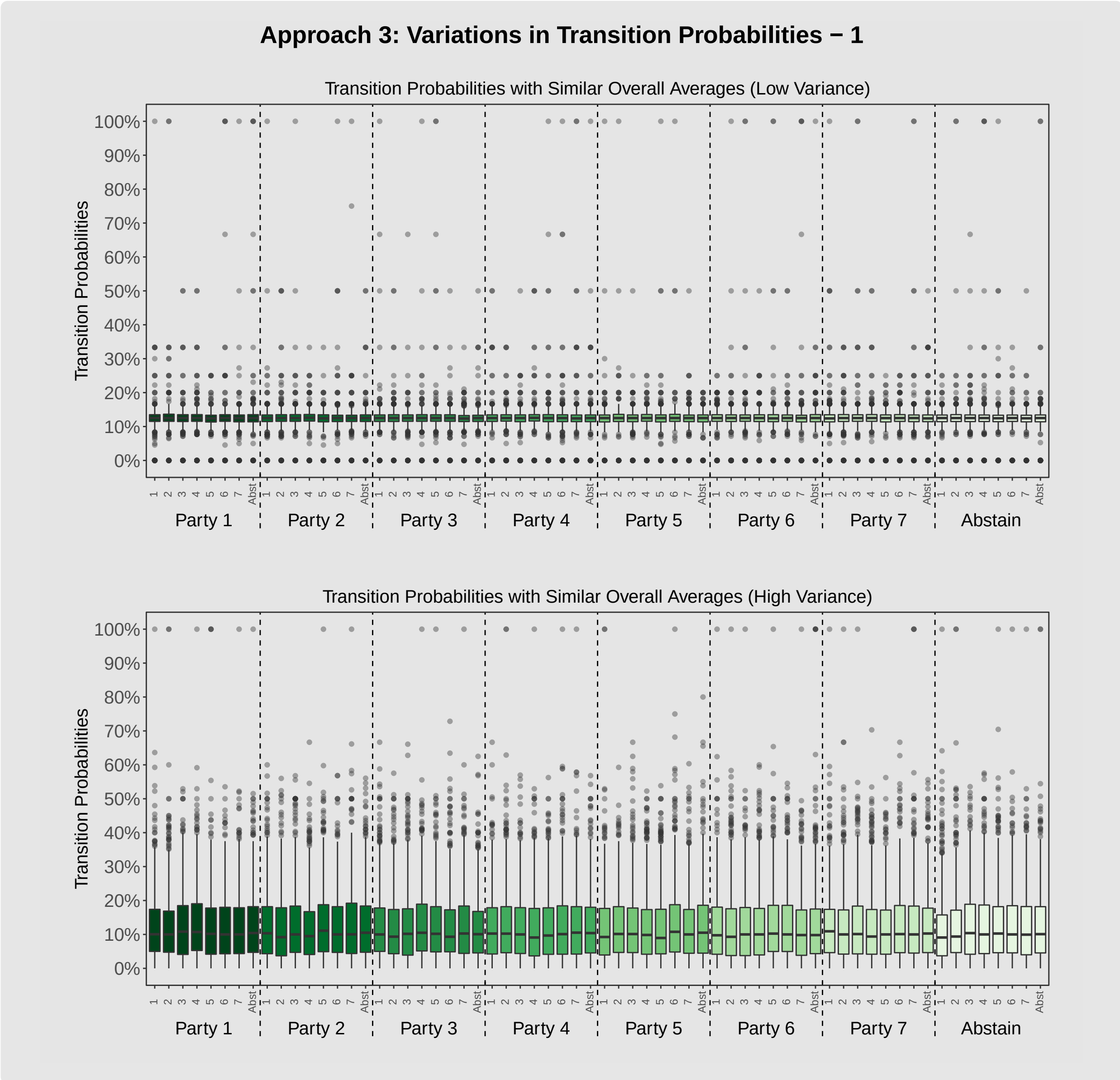


Figure 44: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of varying loyalty and transition probabilities with similar overall averages with a low variance (upper graph) and a high variance probabilities (lower graph) from one simulation with Approach 3.*

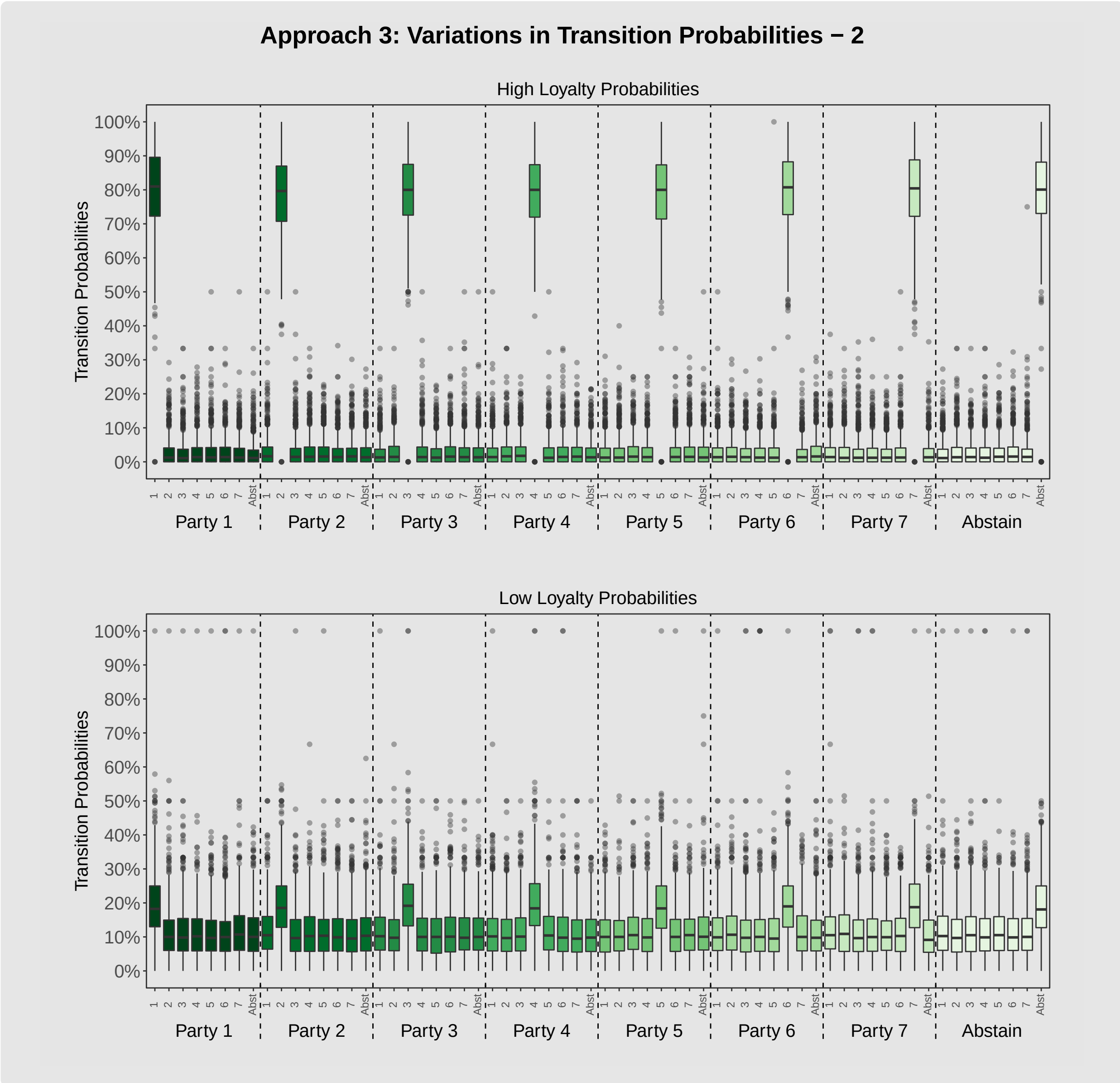


Figure 45: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of high loyalty probabilities (upper graph) and low loyalty probabilities (lower graph) from one simulation with Approach 3.*

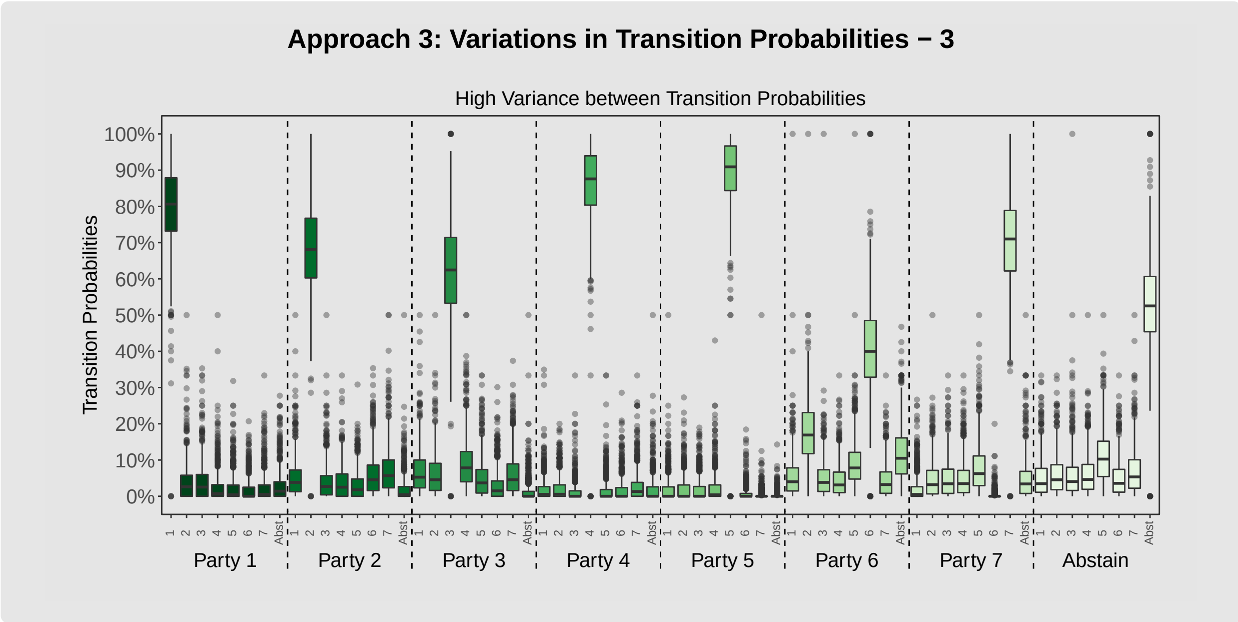


Figure 46: *Example of the distribution of loyalty and transition probabilities over 800 municipalities for seven parties and non-voters in the case of large differences between the expected transition probabilities from one simulation with Approach 3.*

A.3. Expected Transition Probabilities: High Variance in the Case of 5 Parties

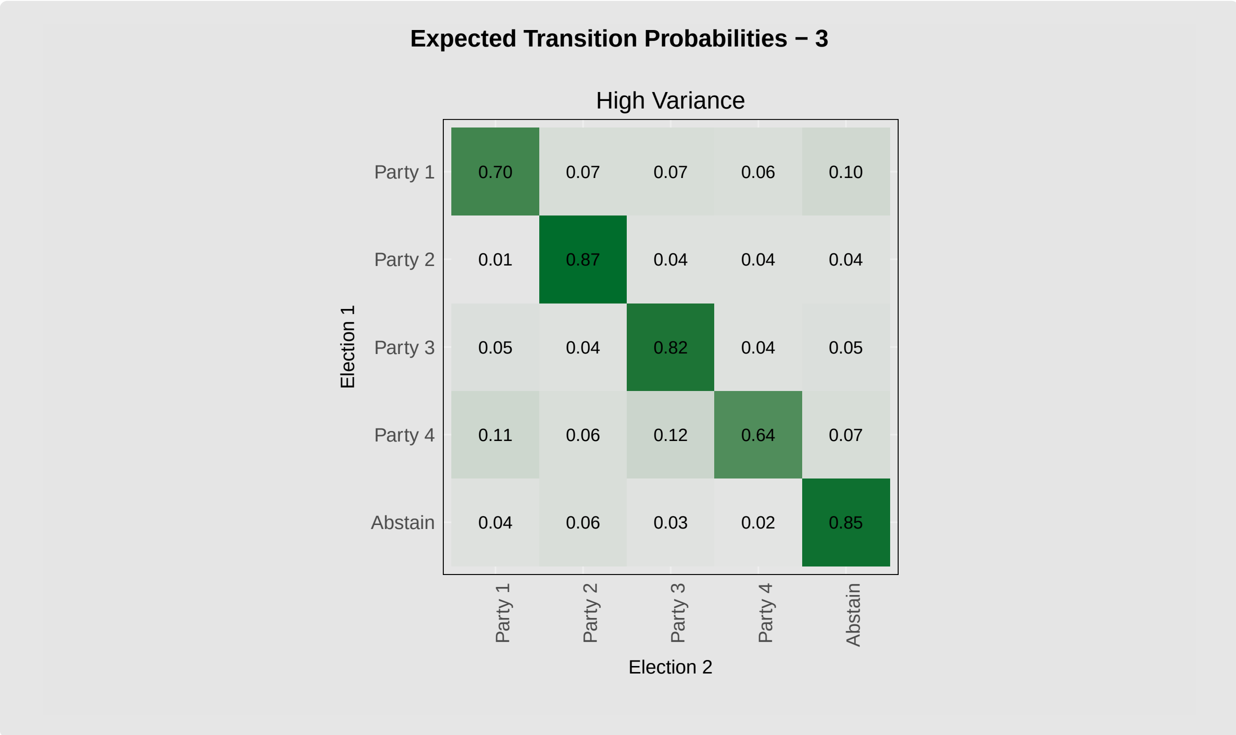


Figure 47: *Expected transition probabilities in the case of assumed large differences between the expected transition probabilities. The displayed expected transition probabilities were used in the case of four assumed parties and non-voters. For this presentation some cell values had to be rounded to the second decimal place.*

A.4. Rounding Algorithm Approach 3

```
### Rounding Algorithm Approach 3 ###

# x: simulated voter transition table
# districts: number of districts
# election1: simulated first election results

approach3_rounding <- function(x,districts,election1){
  ### rounding loop: round decimal values while keeping the column sums fixed
  for (j in 1:districts){
    # pick simulated voter transitions from the j-th district
    temp <- x[,,j]
    # calculate column sums of numbers after decimal point
    sums <- colSums(temp-floor(temp))
    # save matrix with only number after decimal point
    temp2 <- temp-floor(temp)
    # if values in sums are larger than 0:
    # order values in column i of temp2 in decreasing order
    # pick the n largest values from the i-th column in temp2
    # n corresponds to the respictive column sum from sums
    # round n largest values to 1 and the rest to 0
    for (i in 1:ncol(temp2)){
      if(sums[i]>0){od <- order(-(temp2[,i]))
      indices <- od[1:sums[i]]
      temp2[,i][indices] <-  1}
    }
    temp2[temp2 !=1] <- 0
    # add matrix with rounded decimal values (temp2) and matrix
    # with initial voter transitions without decimals (temp) together
    x[,,j] <- temp2+(floor(temp))}
```

```
require(svMisc)
### fitting loop: fit values in rounded matrix to the first election results
for(j in 1:districts){
  # observe progress
  progress(j,max.value = districts)
  # pick rounded simulated voter transitions from the j-th
  # district
  temp <- x[,,j]
  # while row sums of simulated voter transitions do not
  # fit the first election results:
  while(sum(abs(rowSums(temp)-election1[j,2:9]))!=0){
    # take the difference between the row sums and first election results
    diff <- rowSums(temp)-election1[j,2:9]
    # pick the maximum difference
    max <- max(diff)
    # take the indice of the maximum difference
    max.ind <- which(diff == max)
    # if lenght(max.ind) larger than 1 sample one
    if(length(max.ind)!=1){
      max.ind <- sample(max.ind,1)}
    # pick the minimum difference
    min <- min(diff)
    # take the indice of the minimum difference
    min.ind <- which(diff == min)
    # if lenght(min.ind) larger than 1 sample one
    if(length(min.ind)!=1){
      min.ind <- sample(min.ind,1)}
    # pick indices of values in row with maximum difference
    # that are larger than 0. If there are multiple sample one.
    sample.row <- which(temp[max.ind,]>0)
    if(length(sample.row)!=1){
```

```
        sample.row <- sample(sample.row,1)}
      # deduct one from a cell in row with maximum difference,
      # which is larger than 0. Add one to a cell in row with
      # minimum difference.
      temp[max.ind,sample.row] <- temp[max.ind,sample.row]-1
      temp[min.ind,sample.row] <- temp[min.ind,sample.row]+1
    }
    x[,,j] <- temp
  }
  return(x)
}
```

A.5. Elections Results

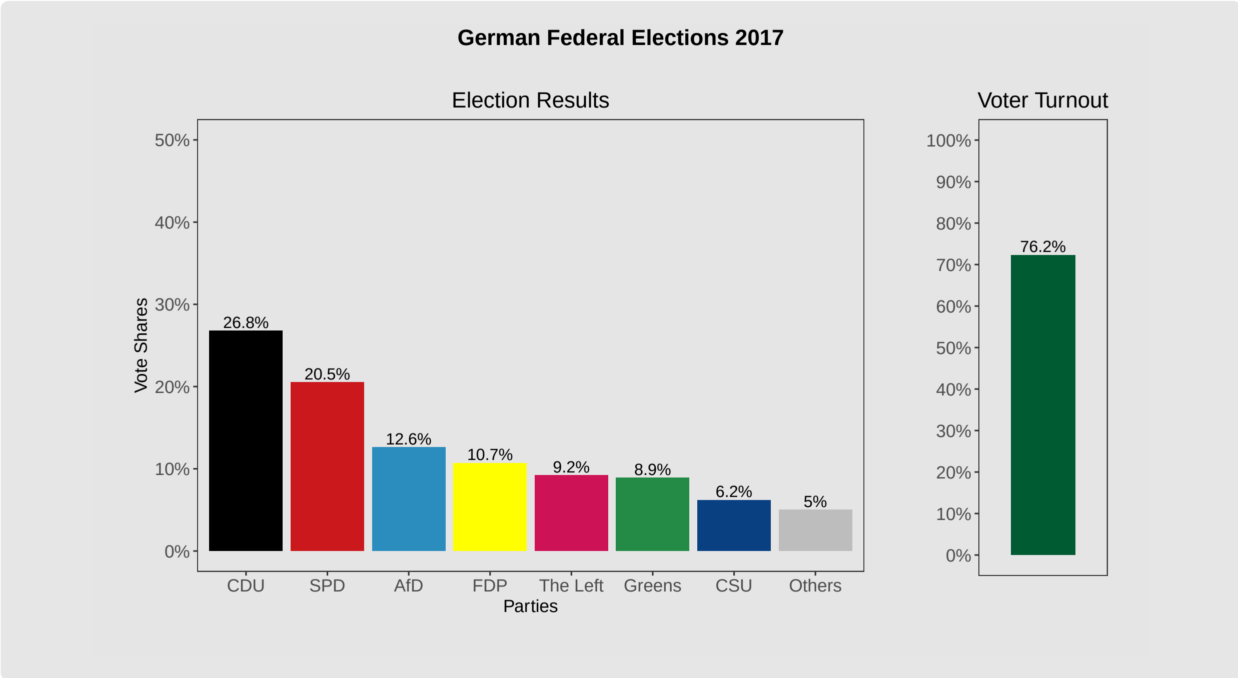


Figure 48: *Left Graph: Election results of the German federal elections 2017. Displayed are all parties which achieved more than 5%. All other parties are summarized in the category Others. Right Graph: Voter turnout at the German federal elections 2017.*

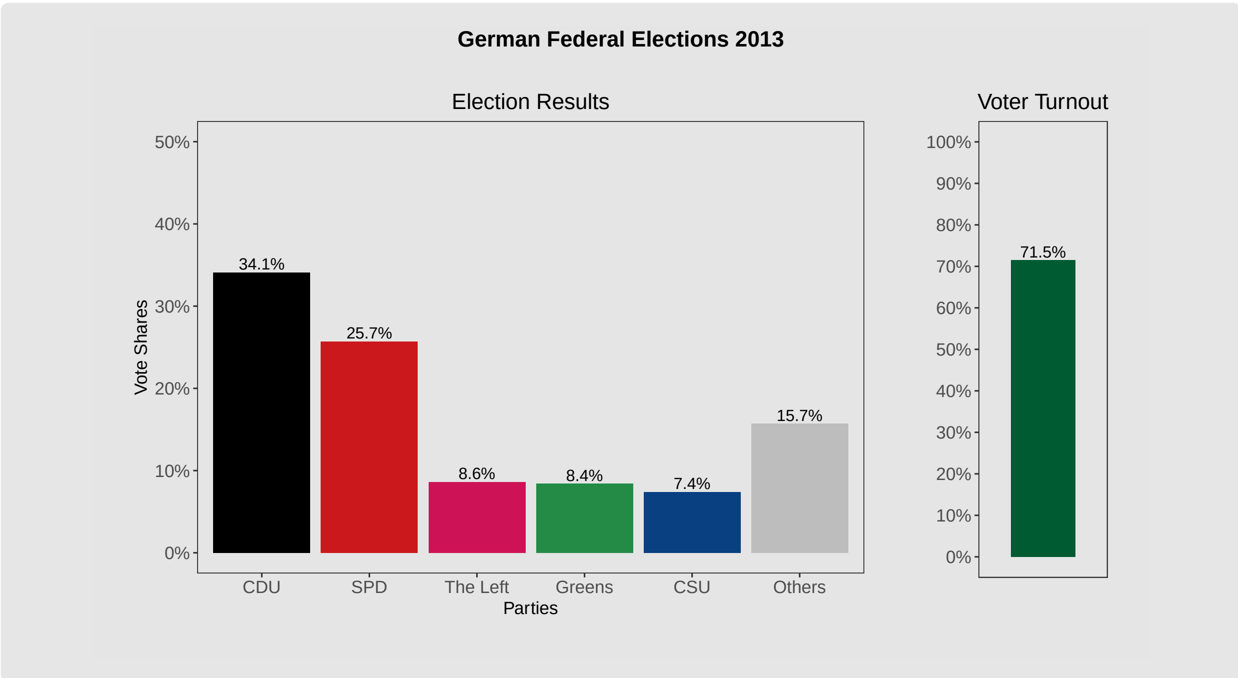


Figure 49: *Left Graph: Election results of the German federal elections 2013. Displayed are all parties which achieved more than 5%. All other parties are summarized in the category Others. Right Graph: Voter turnout at the German federal elections 2013.*

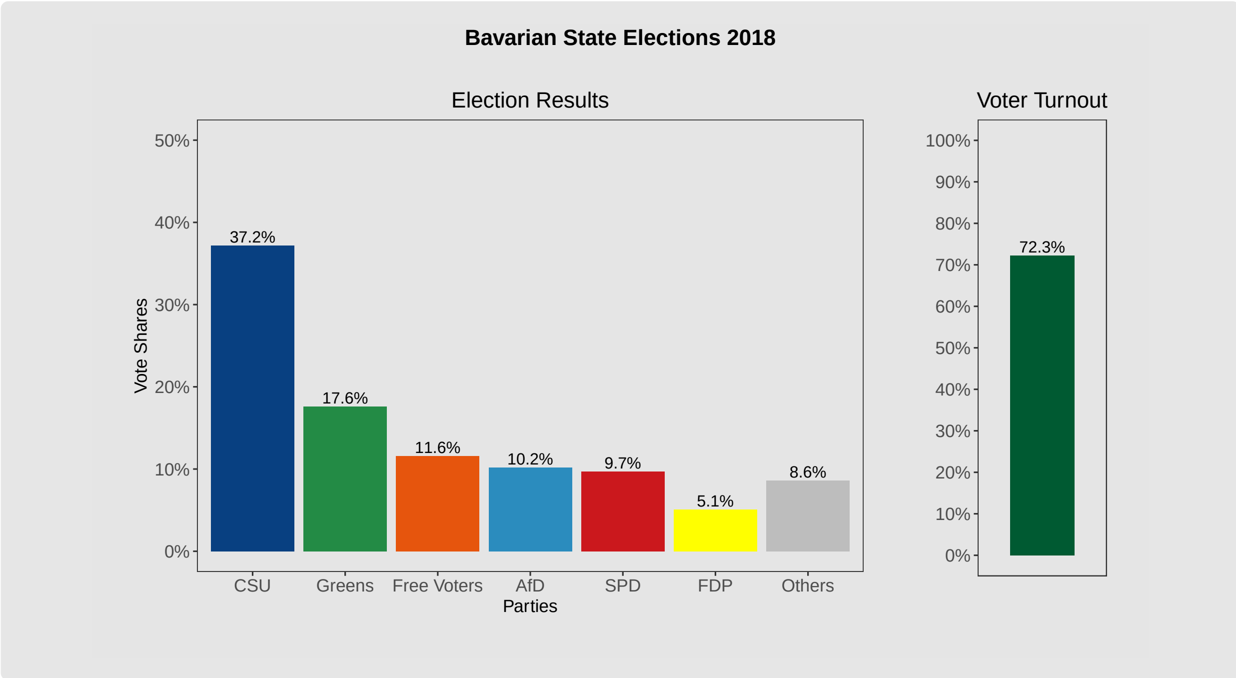


Figure 50: *Left Graph: Election results of the Bavarian state elections 2018. Displayed are all parties which achieved more than 5%. All other parties are summarized in the category Others. Right Graph: Voter turnout at the Bavarian state elections 2018.*

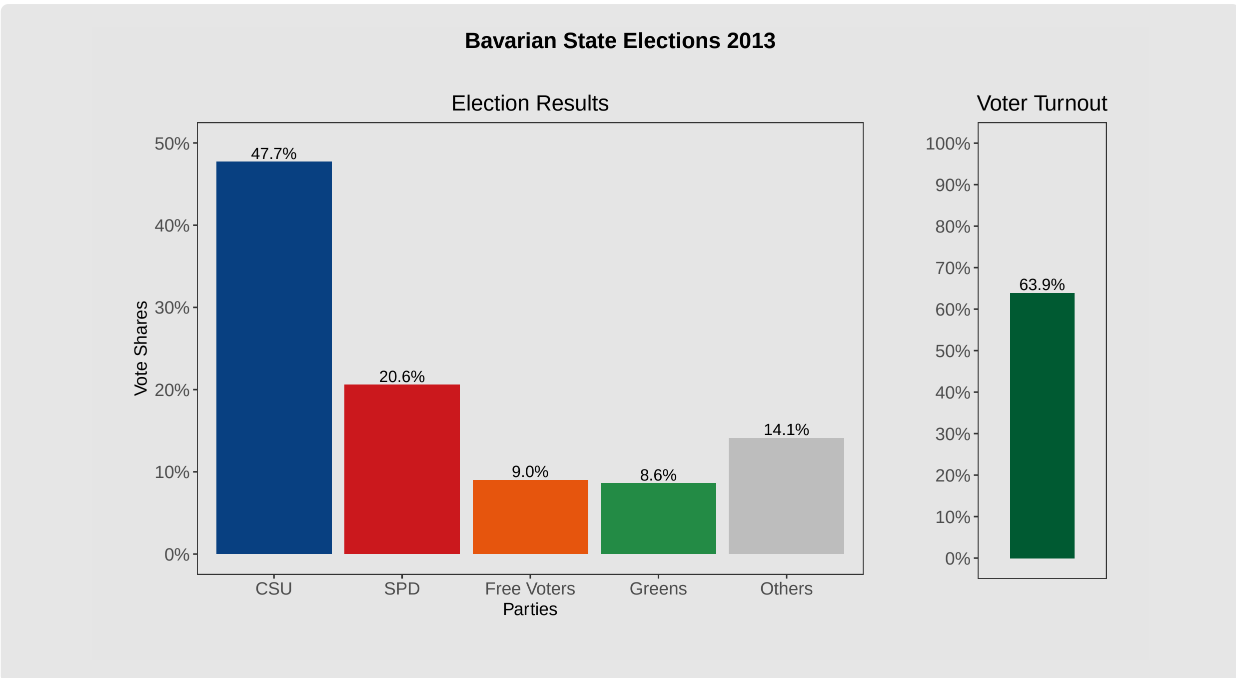


Figure 51: *Left Graph: Election results of the Bavarian state elections 2013. Displayed are all parties which achieved more than 5%. All other parties are summarized in the category Others. Right Graph: Voter turnout at the Bavarian state elections 2013.*

A.6. Scenario Schedules

Schedule Baseline Scenarios

Name	Number of Parties	Count of Districts	Average Population Size	Voting Shares	Transition Probabilities	Sampling Schemes for Individualdata
Control	8	800	800	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 1	5	800	800	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 2	8	800	800	Fixed	Varying with Similar Overall Average	None
Baseline 3	8	800	800	2 Major Parties	Varying with Similar Overall Average	None
Baseline 4	8	800	800	1 Major Party	Varying with Similar Overall Average	None
Baseline 5	8	800	800	Varying with Similar Overall Average	Fixed	None
Baseline 6	8	800	800	Varying with Similar Overall Average	High Loyalty	None
Baseline 7	8	800	800	Varying with Similar Overall Average	Low Loyalty	None
Baseline 8	8	800	800	Varying with Similar Overall Average	Large Differences	None
Baseline 9	8	200	800	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 10	8	400	800	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 11	8	1600	800	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 12	8	2000	800	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 13	8	800	400	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 14	8	800	1200	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 15	8	800	Clustered	Varying with Similar Overall Average	Varying with Similar Overall Average	None
Baseline 16	8	800	800	Varying with Similar Overall Average	Varying with Similar Overall Average	Exit Poll
Baseline 17	8	800	800	Varying with Similar Overall Average	Varying with Similar Overall Average	Telephone Survey

Table 17: *Scenario Schedule of the Baseline scenarios. Displayed are the variable specifications used in each of the Baseline scenarios.*

Schedule Germany Scenarios

Name	Number of Parties	Count of Districts	Average Population Size	Voting Shares	Transition Probabilities	Sampling Schemes for Individualdata
Germany 1	8	800	800	2 Major Parties	Varying with Similar Overall Average	None
Germany 2	8	800	Structural Differences	2 Major Parties	Varying with Similar Overall Average	None
Germany 3.1	8	800	Structural Differences	2 Major Parties	Varying with Similar Overall Average	Exit Poll
Germany 3.2	8	800	Structural Differences	2 Major Parties	Varying with Similar Overall Average	Telephone Survey
Germany 4.1	8	800	Structural Differences	2 Major Parties	Large Differences	Exit Poll
Germany 4.2	8	800	Structural Differences	2 Major Parties	Large Differences	Telephone Survey

Table 18: *Scenario Schedule of the Germany scenarios. Displayed are the variable specifications used in each of the Germany scenarios.*

Schedule Bavaria Scenarios

Name	Number of Parties	Count of Districts	Average Population Size	Voting Shares	Transition Probabilities	Sampling Schemes for Individualdata
Bavaria 1	5	800	800	1 Major Party	Varying with Similar Overall Average	None
Bavaria 2	5	800	Structural Differences	1 Major Party	Varying with Similar Overall Average	None
Bavaria 3.1	5	800	Structural Differences	1 Major Party	Varying with Similar Overall Average	Exit Poll
Bavaria 3.2	5	800	Structural Differences	1 Major Party	Varying with Similar Overall Average	Telephone Survey
Bavaria 4.1	5	2000	Structural Differences	1 Major Party	Varying with Similar Overall Average	Exit Poll
Bavaria 4.2	5	2000	Structural Differences	1 Major Party	Varying with Similar Overall Average	Telephone Survey
Bavaria 5.1	5	2000	Structural Differences	1 Major Party	Large Differences	Exit Poll
Bavaria 5.2	5	2000	Structural Differences	1 Major Party	Large Differences	Telephone Survey

Table 19: *Scenario Schedule of the Bavaria scenarios. Displayed are the variable specifications used in each of the Bavaria scenarios.*

A.7. Interval Width and AD

Average Widths Credible Intervals and Average AD (Eco. Inf.) - Variations in Vote Shares

	ø Interval Width		ø AD	
	Approach 1	Approach 3	Approach 1	Approach 3
Control	2522	2605	0.08	0.08
Fixed Vote Shares	10220	10068	0.05	0.05
Varying Vote Shares with One Major Party	5497	3183	0.15	0.09
Varying Vote Shares with Two Major Parties	4951	2868	0.12	0.08

Table 20: *Detailed evaluation of the average interval width and average AD for Baseline scenarios 2-4 in contrast to those from the control scenario. Displayed are average interval widths and AD's resulting from estimations of voter transitions with the Ecological Inference (Eco. Inf.) model using data simulated with Approach 1 and 3.*

Average Widths Credible Intervals and Average AD (Eco. Inf.) - Variations in Transition Probabilities

	ø Interval Width		ø AD	
	Approach 1	Approach 3	Approach 1	Approach 3
Control	2522	2605	0.08	0.08
Fixed	1126	1215	0.00	0.02
Low Loyalty	2662	2032	0.08	0.06
High Loyalty	827	1028	0.03	0.03
Large Differences	1171	1129	0.07	0.04

Table 21: *Detailed evaluation of the average interval width and average AD for Baseline scenarios 6-9 in contrast to those from the control scenario. Displayed are average interval widths and AD's resulting from estimations of voter transitions with the Ecological Inference (Eco. Inf.) using data simulated with Approach 1 and 3.*

A.8. Cell-Wise Coverage Rates

A.8.1. Selected Baseline Scenarios

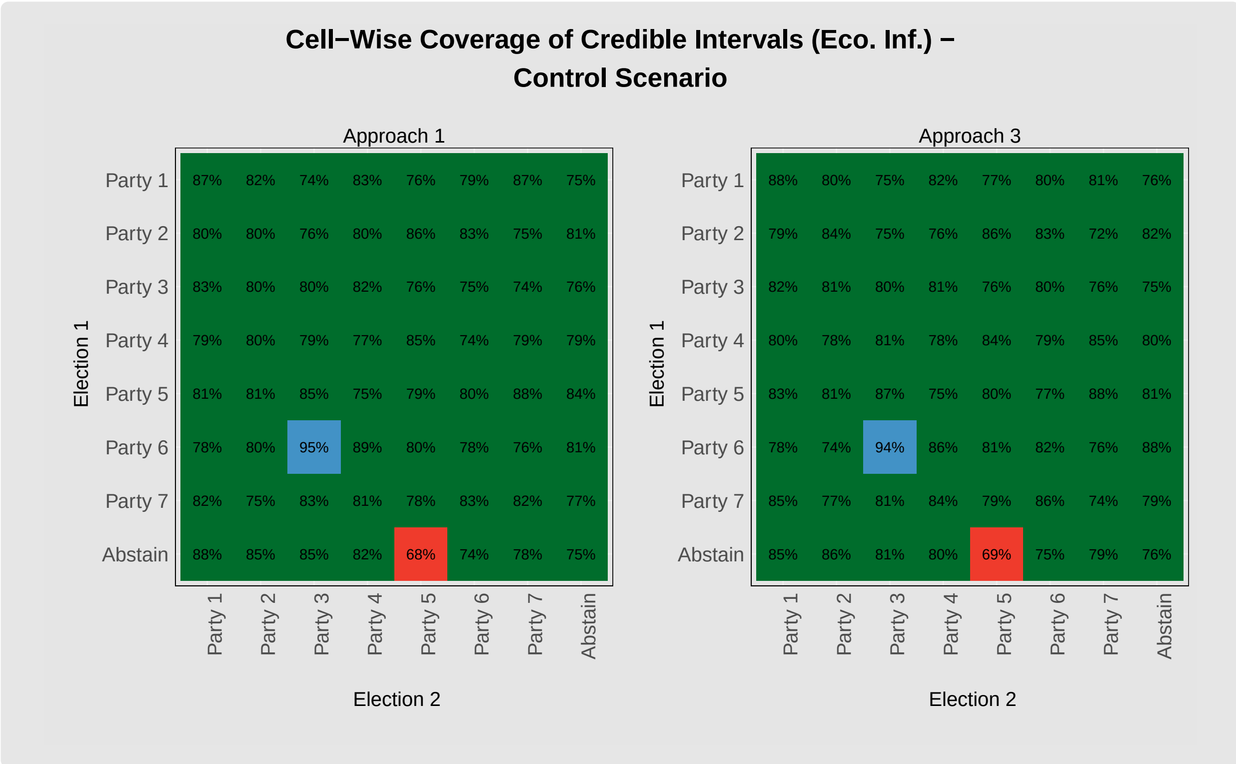


Figure 52: *Detailed cell-wise evaluation of the 80%-credible intervals for the Control scenario of the Baseline scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the Ecological Inference Model (Eco. Inf.). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach.*

Figure 53: *Detailed cell-wise evaluation of the 80%-credible intervals for Scenario 9 of the Baseline scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM [upper Graph]) and the Ecological Inference Model (Eco. Inf. [lower Graph]). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graphs) and Approach 3 (right graphs). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach.*

A.8.2. Selected Germany Scenarios

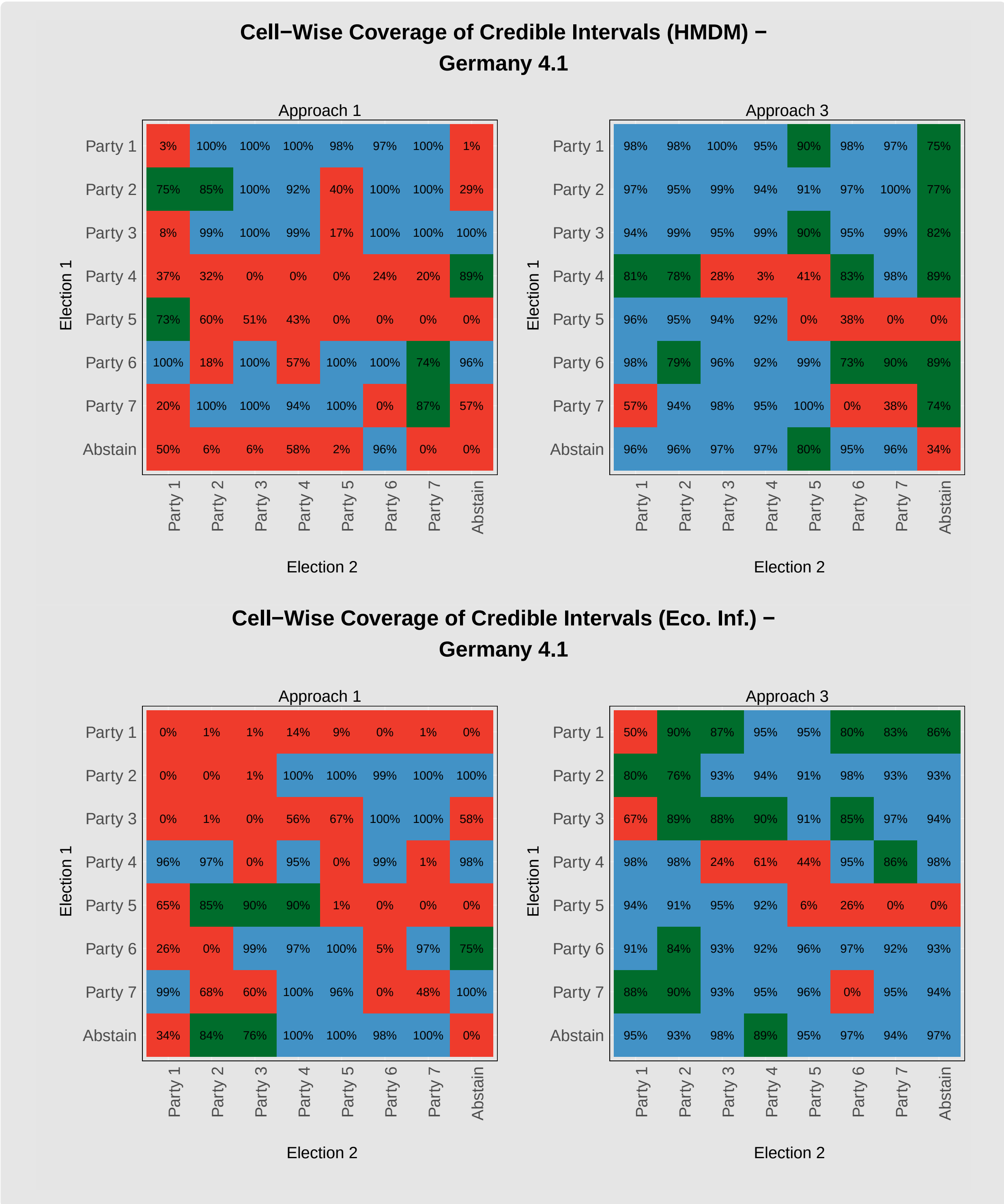


Figure 54: *Detailed cell-wise evaluation of the 80%-credible intervals for Scenario 4.1 of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM [upper Graph]) and the Ecological Inference Model (Eco. Inf. [lower Graph]). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach.*

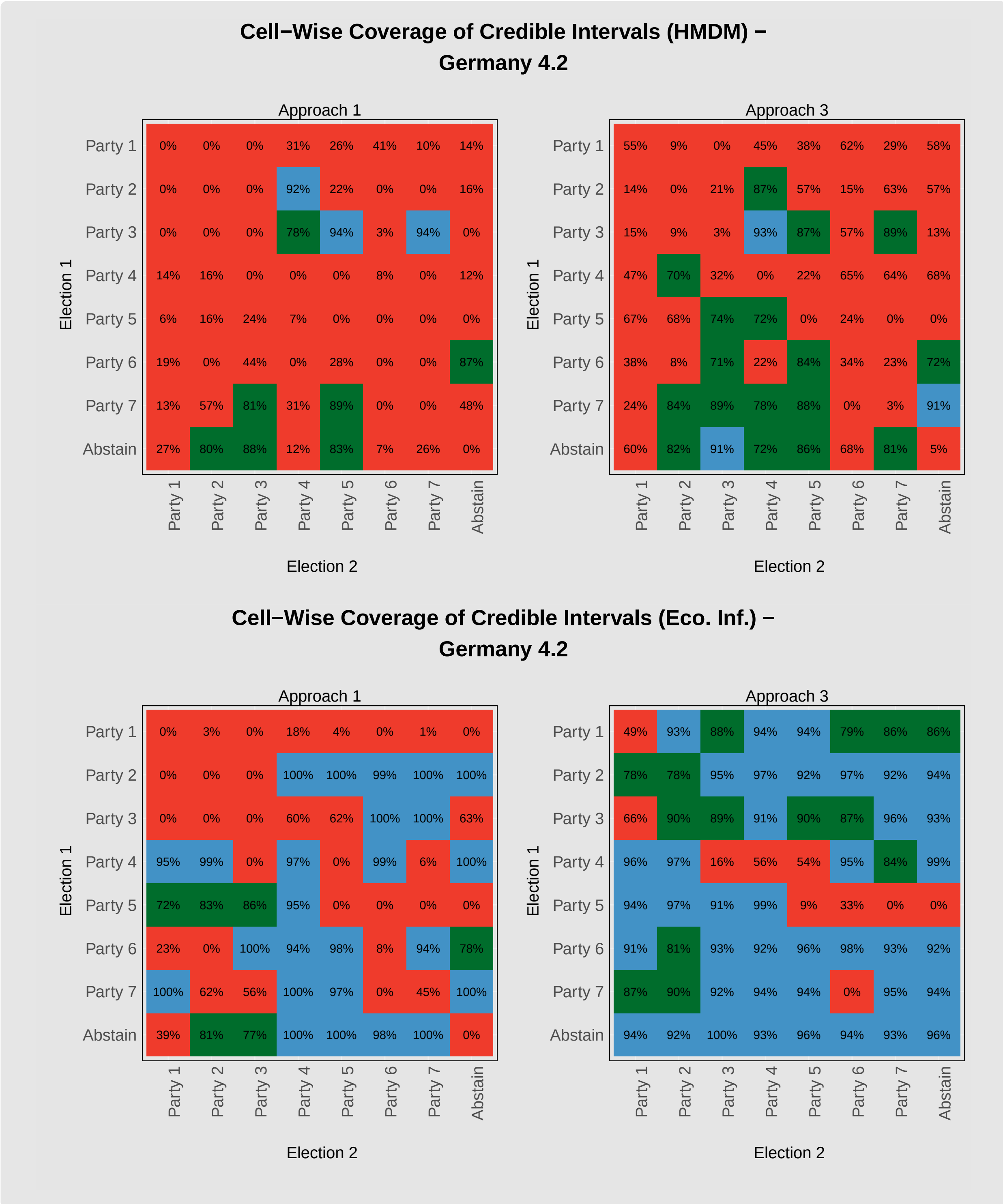


Figure 55: *Detailed cell-wise evaluation of the 80%-credible intervals for Scenario 4.2 of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM [upper Graph]) and the Ecological Inference Model (Eco. Inf. [lower Graph]). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach.*

A.8.3. Selected Germany Scenarios: Average Width Credible Intervals

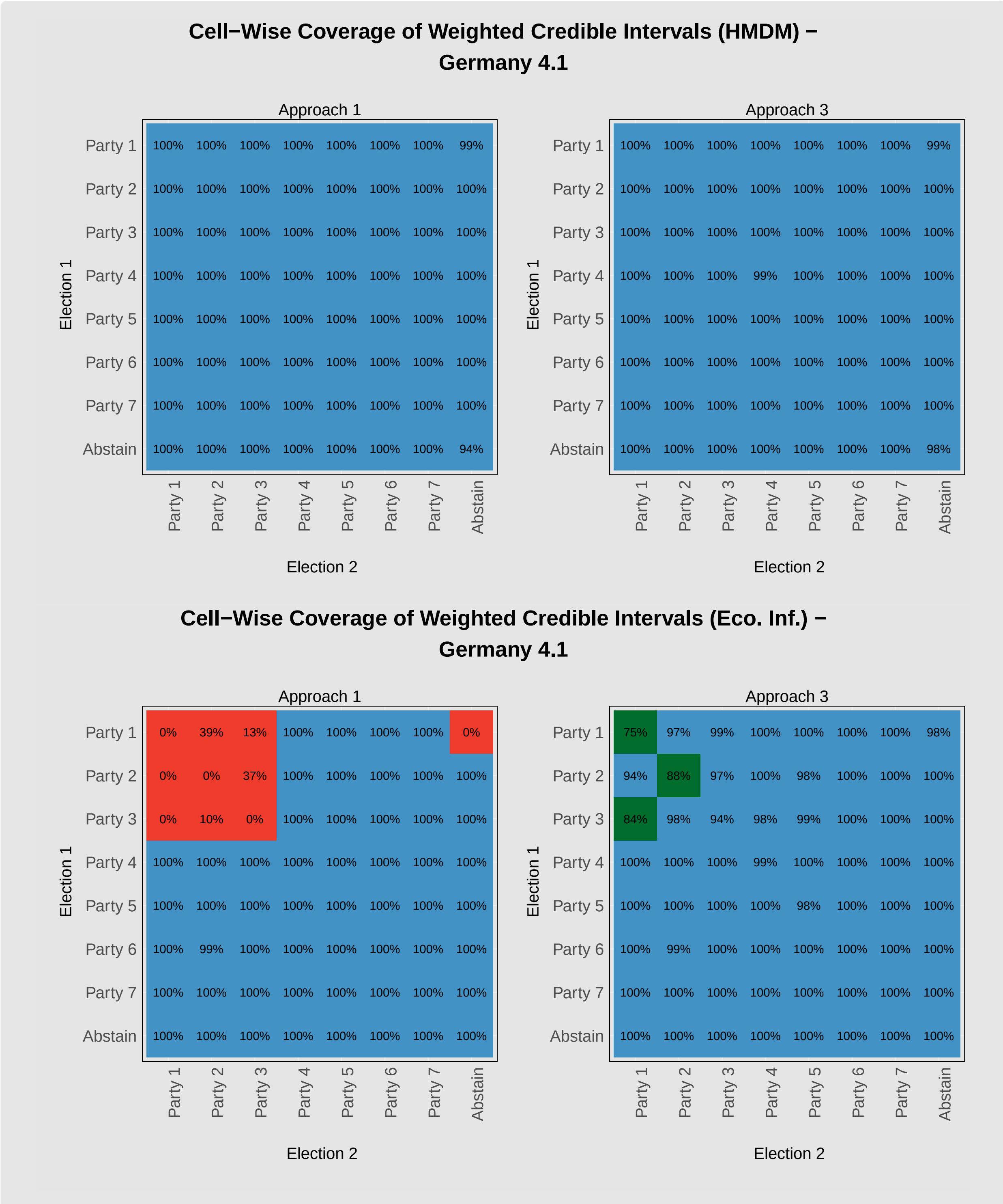


Figure 56: *Detailed cell-wise evaluation of the average width credible intervals for Scenario 4.1 of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM [upper Graph]) and the Ecological Inference Model (Eco. Inf. [lower Graph]). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the average width approach discussed in chapter 6.1.*

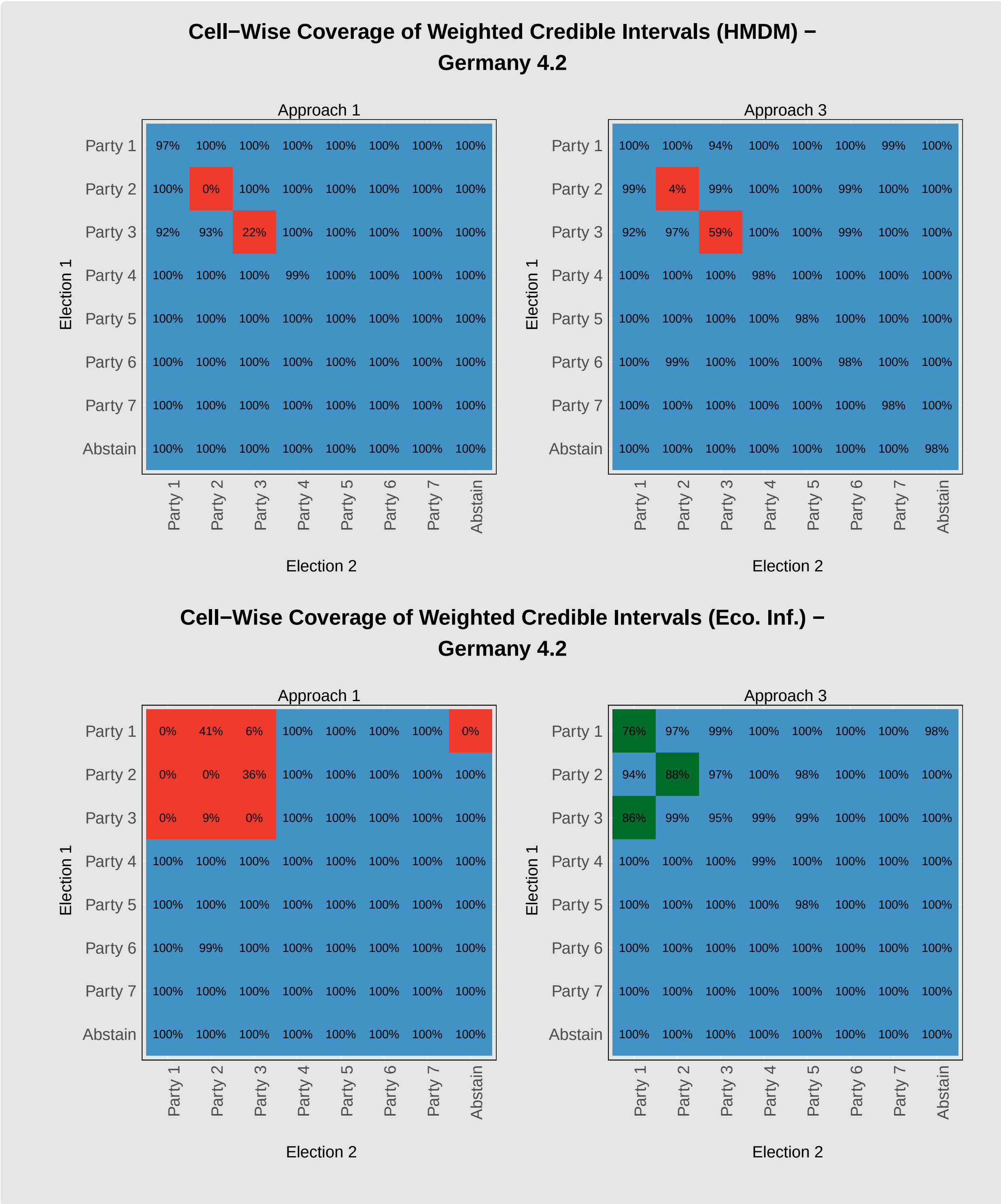


Figure 57: *Detailed cell-wise evaluation of the average width credible intervals for Scenario 4.2 of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM [upper Graph]) and the Ecological Inference Model (Eco. Inf. [lower Graph]). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the average width approach discussed in chapter 6.1.*

A.8.4. Selected Germany Scenarios: Weighted Credible Intervals

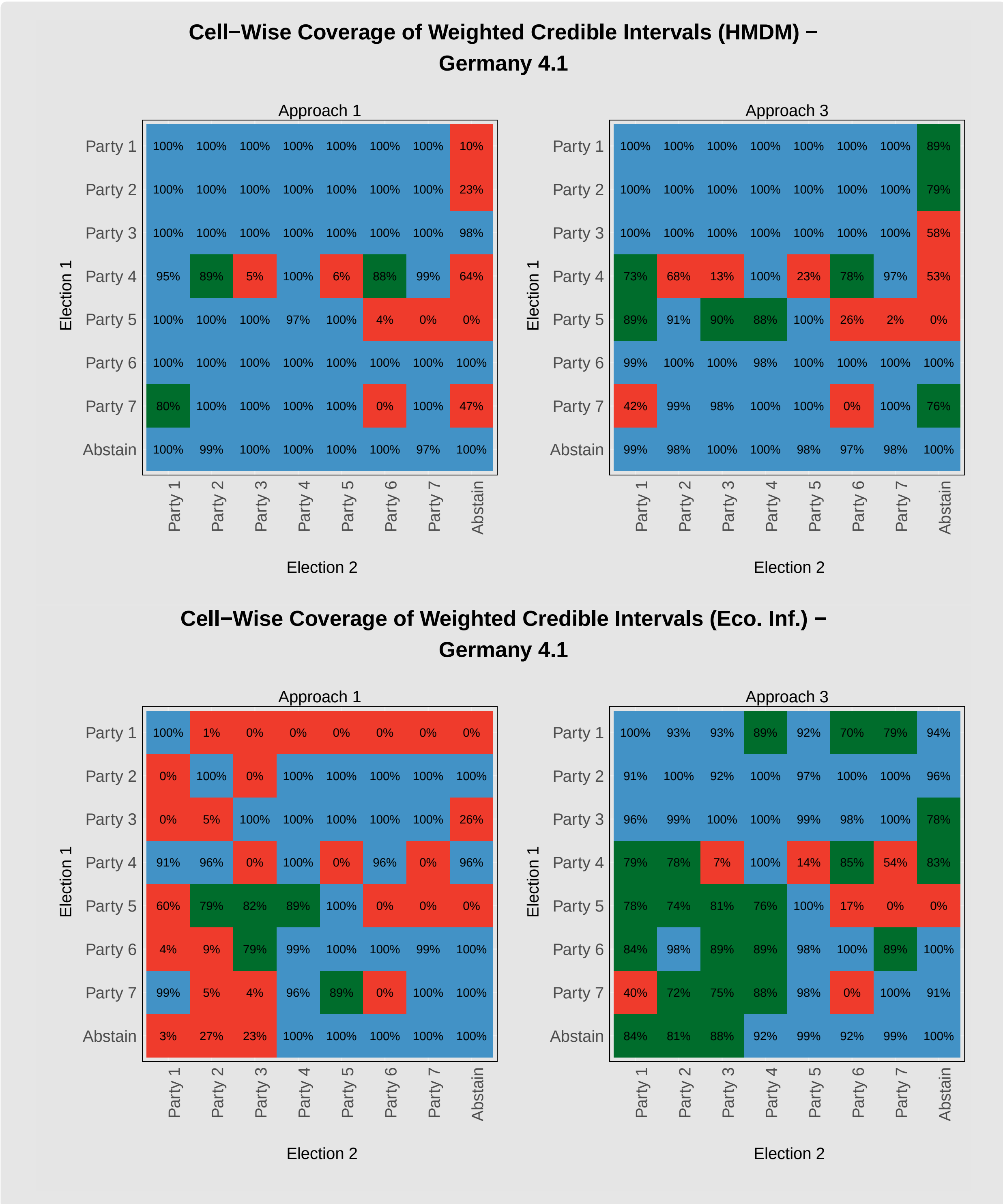


Figure 58: *Detailed cell-wise evaluation of the weighted credible intervals for Scenario 4.1 of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM [upper Graph]) and the Ecological Inference Model (Eco. Inf. [lower Graph]). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the weighted approach discussed in chapter 6.1.*

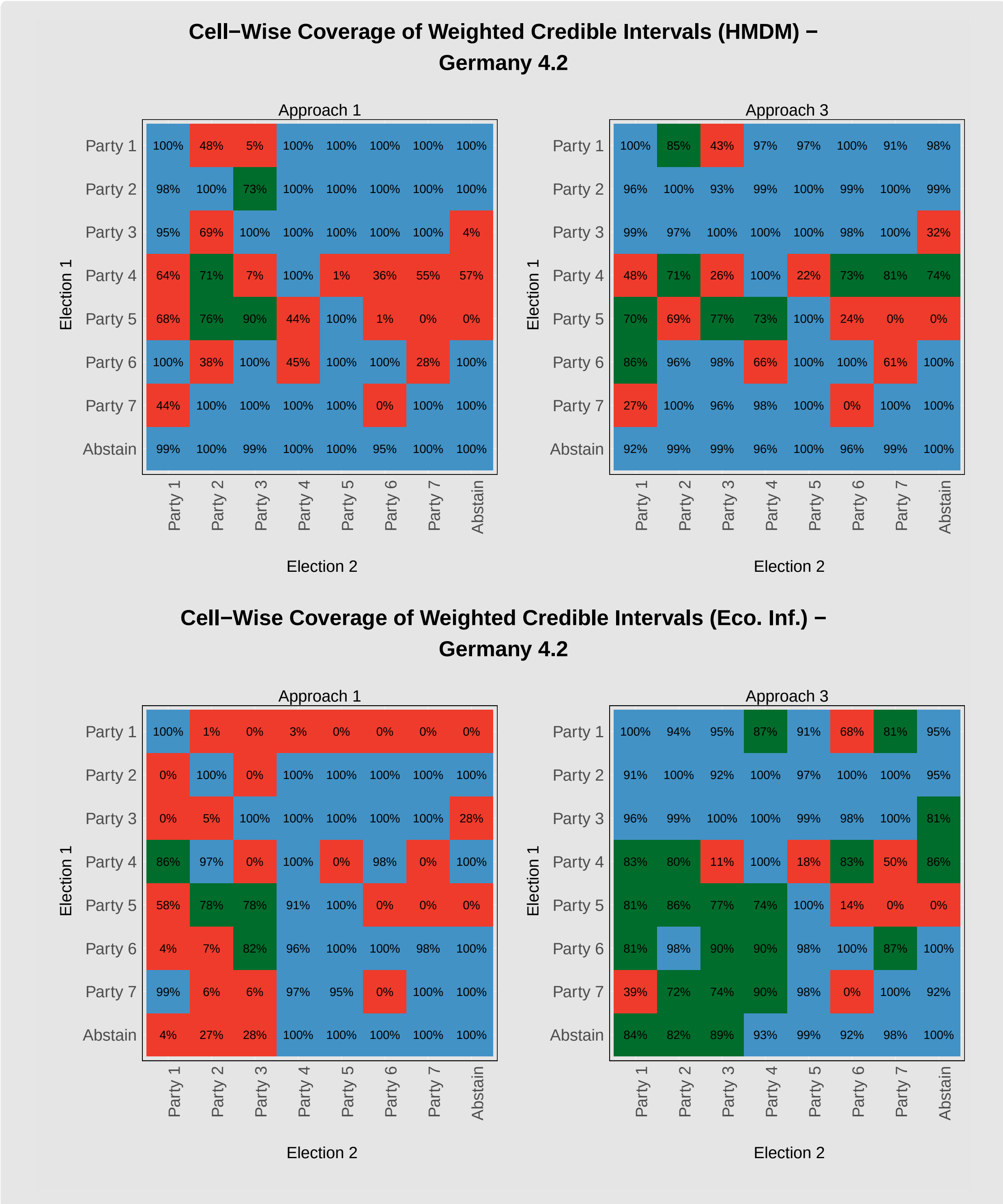


Figure 59: *Detailed cell-wise evaluation of the weighted credible intervals for Scenario 4.2 of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM [upper Graph]) and the Ecological Inference Model (Eco. Inf. [lower Graph]). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the weighted approach discussed in chapter 6.1.*

A.8.5. Selected Germany Scenarios: Bias Corrected Credible Intervals

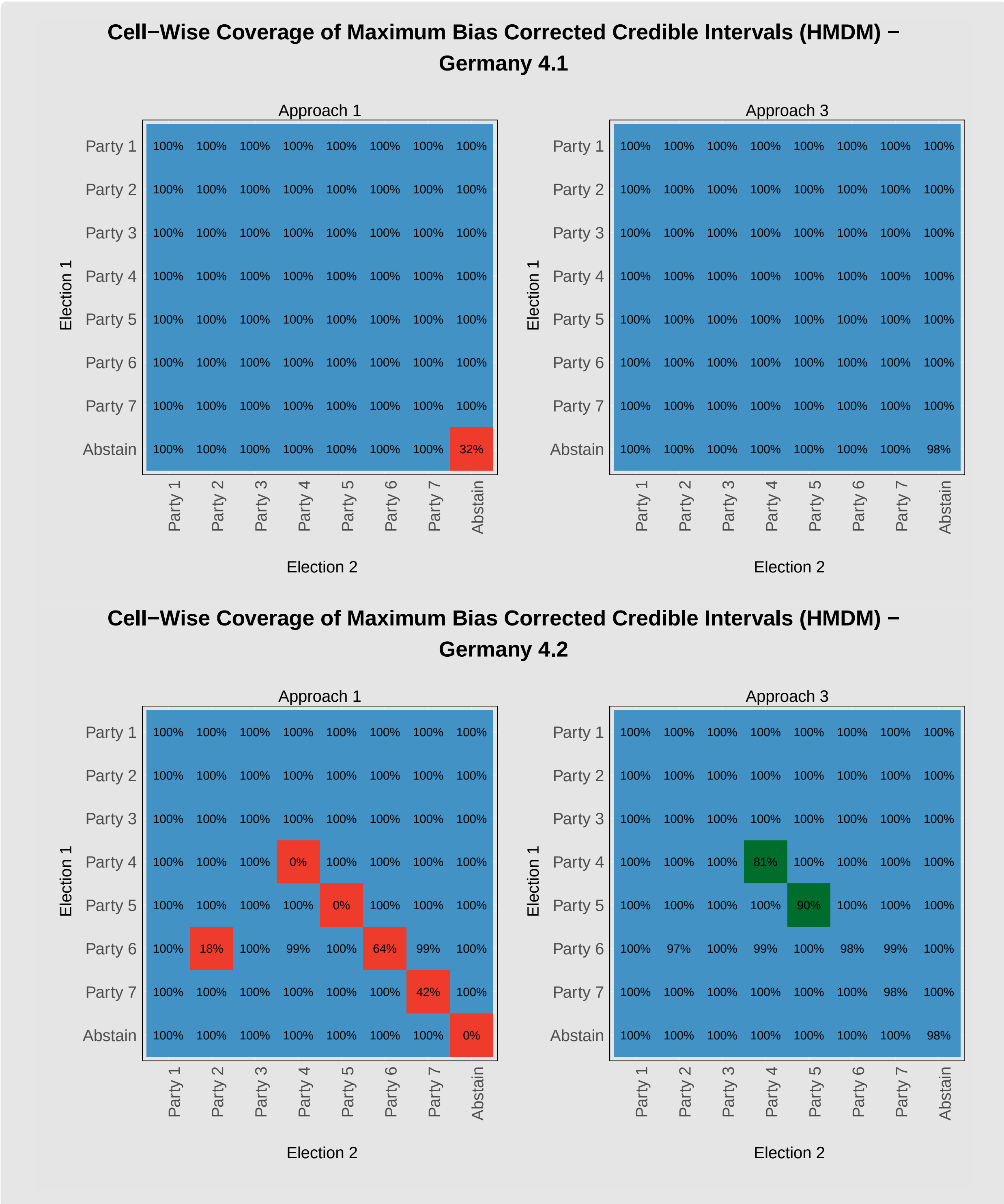


Figure 60: *Detailed cell-wise evaluation of the maximum bias corrected credible intervals for scenario 4.1 (upper graph) and 4.2 (lower graph) of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the maximum bias correction approach discussed in chapter 6.2.*

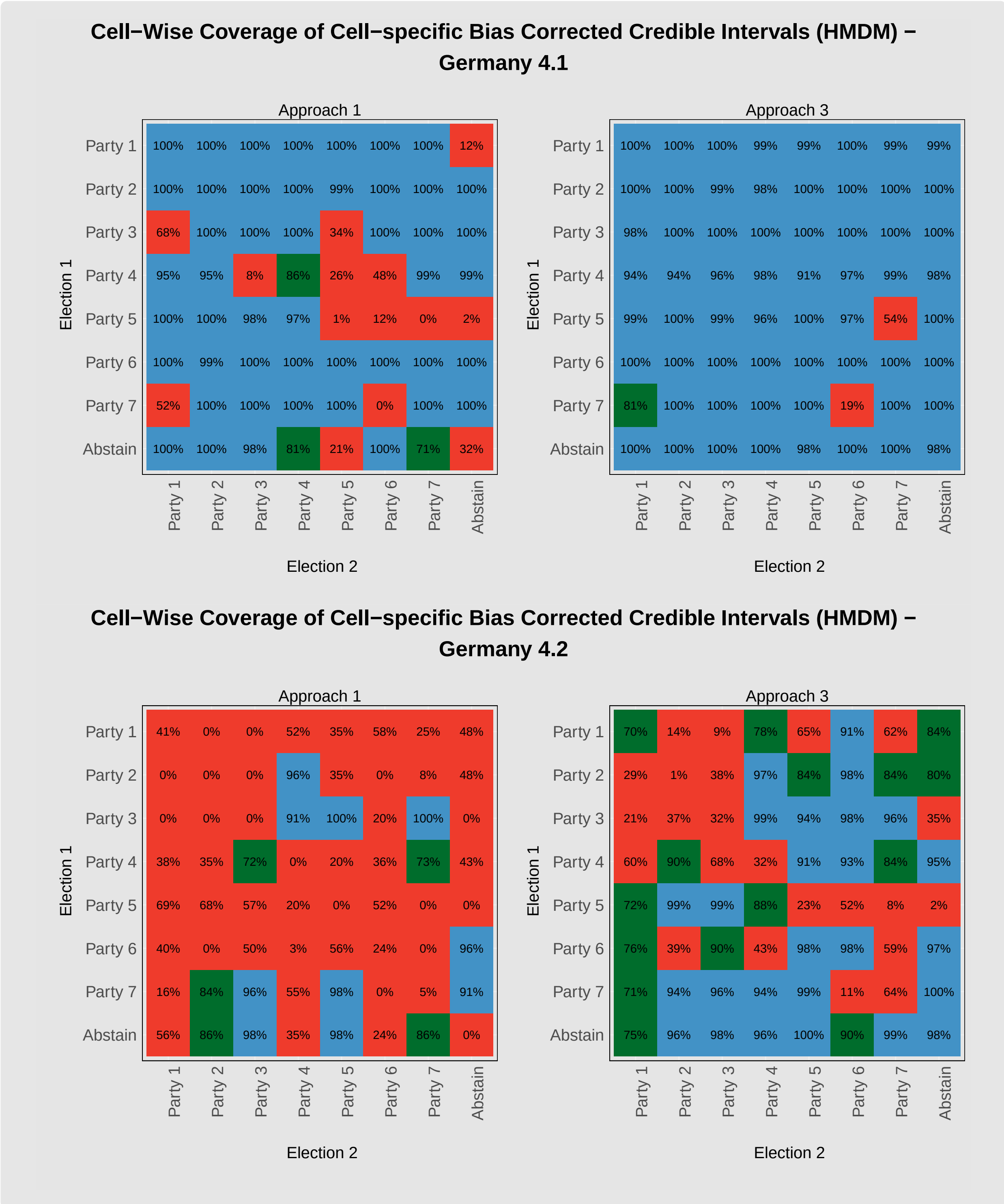


Figure 61: *Detailed cell-wise evaluation of the cell specific bias corrected credible intervals for scenario 4.1 (upper graph) and 4.2 (lower graph) of the Germany scenarios. Displayed is, for each cell, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. Coverage rates between 70% and 90% are colored in green, above 90% in blue and below 70% in red. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM). 100 Datasets are evaluated for each scenario. These datasets were simulated according to Approach 1 (left graph) and Approach 3 (right graph). A brief overview of the scenario can be found in appendix 6. The credible intervals were calculated according to the maximum bias correction approach discussed in chapter 6.2.*

A.9. Excursus: Fixed Transition Probabilities and Fixed Vote Shares

As already mentioned in chapter 5, an initial approach to the Control scenario assumed fixed vote shares as well as fixed transition probabilities. Further specifications of the model were: 8 parties, 800 districts, an average of 800 voters per district and no individual data. However, due to the poor performance of the model this approach was discarded and varying transition probabilities with similar overall averages were assumed instead of fixed ones.

Figure 62: *Simulated voter transitions and transitions probabilities between two fictitious elections in the case fixed vote shares and fixed transitions probabilities are assumed. The voter transitions and transitions probabilities were simulated according to the first approach presented in chapter 4. The specifications of the scenario are: fixed vote shares, fixed transition probabilities, 8 parties, 800 districts, an average of 800 voters per district and no individual data. Simulated transition probabilities are displayed in the upper graph and estimated voter transitions in the lower one. Presented are the results from the first simulation.*

Nonetheless, the following excursus will discuss the results of that first attempt because it portrays the limits of the (hybrid-) Multinomial-Dirichlet model and shows what happens if the assumptions made in the models are violated. In this particular case, the assumption of Dirichlet distributed transition probabilities is violated by postulating fixed transition probabilities.

Figure 61 displays the simulated transition probabilities and voter transitions from one simulation. In the context of this example they constitute the true values, which are supposed to be estimated with the (hybrid-) Multinomial-Dirichlet model. As one can see, the simulated transition probabilities and voter transitions are almost equal for all cells. Due to a mistake in the initial programming of a rounding function, the values gradually increase. However, since the differences are not substantial it should not have an impact on the estimation. The mistake was corrected in the final implementation of the function.

Figure 63: *Estimated voter transitions and transitions probabilities for a scenario, which assumes fixed vote shares and fixed transitions probabilities. The voter transitions and transitions probabilities were estimated with the hybrid Multinomial-Dirichlet Model. Estimated transition probabilities are displayed in the upper graph and estimated voter transitions in the lower one. Presented are the results from the first model.*

The resulting transition probabilities and voter transitions from one estimation with the HMDM can be found in Figure 62. Estimations with the EI model led to similar results which is why they will not be presented. Interestingly, in this particular case the HMDM severely overestimates one cell per row, while all other cells are underestimated. The model completely fails at accurately estimating the simulated voter transitions. However, as the evaluation in chapter 5 has shown, if variation between the transition probabilities in the different districts is allowed, as expected by the model, voter transitions are estimated accurately.

Even though the postulated scenario with fixed vote shares and fixed transition probabilities does not seem realistic it is important as it emphasizes the importance of the assumptions made in the (hybrid-) Multinomial-Dirichlet model. As the given example shows, a violation of these assumptions can have a drastic and very negative impact on the estimation.

A.10. Excursus: Biased Individual Data

An initial implementation of the telephone survey simulation produced an implicit bias in the outcome of the survey. This chapter will shortly discuss the bias in the telephone survey and its effect of the coverage rates as it yielded very interesting results.

The results of the unbiased and biased telephone survey are displayed in Table 22 in the upper and lower contingency table, respectively. A short comparison of these two telephone surveys shows the impact of the implicit bias in the initial implementation. In the initial implementation of the telephone survey, the probability of the loyal voters to be represented in the survey was assumed to be higher compared to all other voters. This leads to an overrepresentation of the loyal voters in the telephone survey. Thus, the diagonal cell values are higher in the biased telephone survey while all other cell values decrease compared to the unbiased telephone survey. Furthermore, for parties with high loyalty rates, in this case Party 3 and Party 4, only voter transitions of loyal voters are available.

The impact of such a bias in the individual data has on the coverages rates is shown in Figure 64. Scenario Germany 4.2 was estimated with a biased and unbiased telephone survey. Figure 64 presents the coverage rates of the models estimated with a biased telephone survey in red and those from models estimated with an unbiased telephone survey in green. As discussed in chapter 5.2 the combination of individual data from a telephone survey and large differences in transition probabilities, as it is present in scenario Germany 4.2, leads to a drastic drop in coverage rates. Interestingly, a bias in the individual data leads to another substantial drop. In the case of an unbiased telephone survey, the coverage rates of models estimated with data from Approach 1 are on average about 20%, while the coverage rates of models estimated with data

Telephone Survey without Bias

Election 1 \ Election 2	Party 1	Party 2	Party 3	Party 4	Party 5	Party 6	Party 7	Abstain	$\sum$
Party 1	3500	127	153	106	107	88	105	120	4306
Party 2	155	2952	127	134	120	165	200	88	3941
Party 3	33	35	793	14	40	70	40	6	1031
Party 4	40	41	29	816	9	0	0	29	964
Party 5	75	120	70	98	275	56	113	58	865
Party 6	39	74	64	72	0	610	70	70	999
Party 7	56	62	85	102	60	75	398	70	908
Abstain	190	166	272	158	207	163	64	2695	3915
$\sum$	4088	3577	1593	1500	818	1227	990	3136	16929

Telephone Survey with Bias

Election 1 \ Election 2	Party 1	Party 2	Party 3	Party 4	Party 5	Party 6	Party 7	Abstain	$\sum$
Party 1	4429	70	73	37	36	31	35	51	4762
Party 2	87	3680	67	71	56	115	165	30	4271
Party 3	0	0	1226	0	0	0	0	0	1226
Party 4	0	0	0	1268	0	0	0	0	1268
Party 5	10	58	4	30	311	5	40	5	463
Party 6	0	6	5	19	0	884	6	4	924
Party 7	6	13	15	13	40	5	535	17	644
Abstain	159	105	334	84	53	112	4	3421	4272
$\sum$	4691	3932	1724	1522	496	1152	785	3528	17830

Table 22: *Example of the available individual data from a telephone survey in 800 districts. The given example is taken from one simulation. The upper graph illustrates the available individual data from an unbiased telephone survey and the lower graph illustrates the available individual data from a biased telephone survey.*

from Approach 3 are on average about 45%. However, if the individual data is biased the average coverage rates drop to about 5% and 20%, respectively.

A possible reason for the drop in coverage rates can be found in the bias of the estimation. The problem of a non-bias free estimation has already been discussed in chapter 6.2 but seems to be especially relevant in the case of a bias in the individual data. Figure 65 displays the cell-wise deviation of the estimated voter transitions from the simulated voter transitions. In all cases the number of loyal voters are severely overestimated, which leads to an underestimation of almost all other voter transitions. This overestimation of the number of loyal voters corresponds to the overrepresentation of said voters in the individual data. It seems as if a bias in the individual data is also reflected in the estimation of the voter transitions and thus, leads to a high estimation error. Therefore, this phenomenon could explain the substantial drop in coverage rates if a bias in the individual data is present.

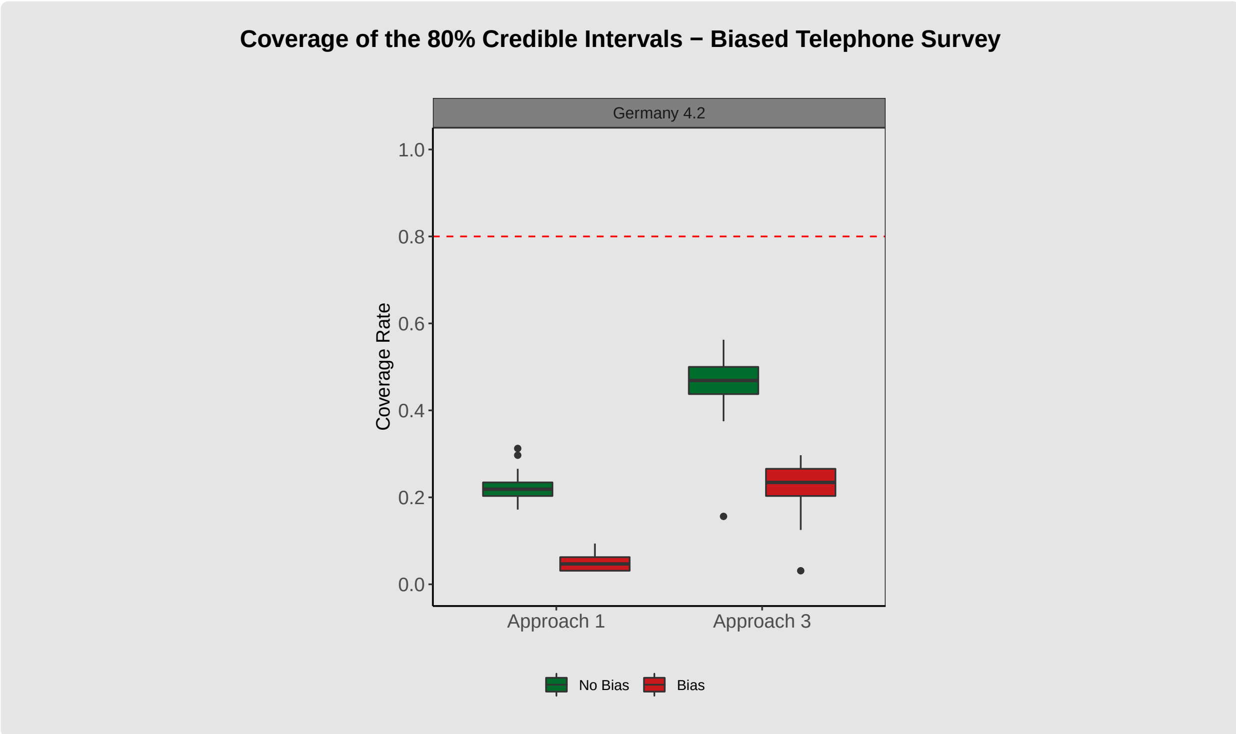


Figure 64: *Detailed evaluation of the credible intervals for the Scenario Germany 4.2 with (red) and without (green) a biased telephone survey. Displayed is, for each estimated voter transition, the share of the 80%-credible intervals, which cover the true value from the simulated dataset. The voter transitions have been calculated with the hybrid Multinomial-Dirichlet Model (HMDM) and the Ecological Inference (Eco. Inf.). 30 Datasets with a biased telephone survey and 30 datasets without a biased telephone survey are evaluated for each scenario. These datasets were simulated according to Approach 1 and Approach 3. A brief overview of the Germany 4.2 scenario can be found in appendix 6. The credible intervals were calculated according to the Equal-Tail Approach. The dotted red line marks a share of 80%.*

As already mentioned the results were only briefly discussed as they seemed very interesting. However, in order to better understand the effect of a bias in the individual data, further research would be necessary. For example, it would be interesting to see whether the same results can be observed if voter transitions are estimated with a biased Exit Poll. However, as this is not the main focus of this master thesis, a further investigation of this aspect will be omitted.

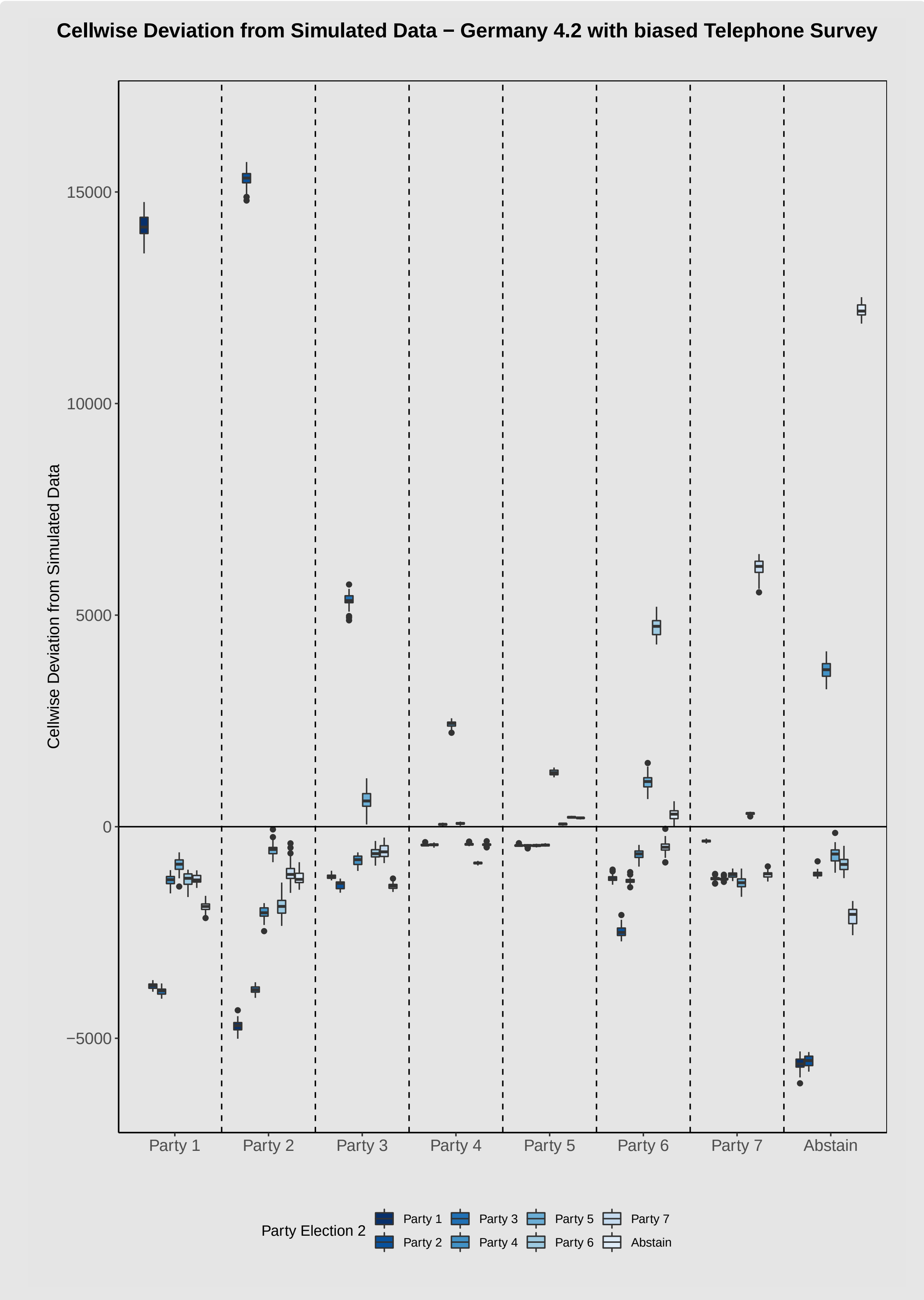


Figure 65: *Example cell specific deviation from simulated data for scenario Germany 4.2 in case of a biased telephone survey. Displayed are the cell specific deviations of the estimated voter transitions from simulated voter transitions. The voter transitions were simulated according to Approach 1. The voter transitions were estimated with the hybrid Multinomial-Dirichlet Model. 30 models were evaluated and the displayed values are averages across all models.*

A.11. Digital Appendix

Chapter 2	
example_approaches_credible_intervals.R	R-Code for the plot showing approaches to credible intervals
plot_example_convergence_diagnostics.R	R-Code for the convergence diagnostic plot
Data-Folder	Folder containing the required data for the plots
Chapter 3	
example_tomography_lines.R	R-Code for the plot displaying the tomography lines
Chapter 4	
plots_example_data.R	R-Code for the plots displaying variable variations
Data-Folder	Folder with data required for the plots displaying variable variations
Chapter 5	
evaluation_cellwise_coverage_rates.R	R-Code for the evaluation of the cell-wise coverage rates
plot_cellwise_coverage_rates.R	R-Code for the plot displaying the cellwise coverage rates
plot_coverage_rates.R	R-Code for the plot displaying the coverage rates
evaluation_MSE_interval_width_AD.R	R-Code for the evaluation of the MSE, AD and interval widths of the models
plot_interval_width_AD.R	R-Code for the plot displaying interval widths and ADs
Data-Folder	Folder containing the required data for the plots and calculations
Chapter 6	
evaluation_parametric_bootstrap.R	R-Code for the evaluation of parametric boostrap models

bias_corrected_credible_intervals.R	R-Code for the calculation and evaluation of bias corrected credible intervals
plot_coverage_rates_bias_corrected_credible_intervals.R	R-Code for the plots displaying the results of the bias corrected credible intervals
plot_cellwise_coverage_rates_bias_correction.R	R-Code for the plot of the cell-wise coverage rates for the bias corrected credible intervals
calculate_average.width_weights.R	R-Code for the calculation of average width and weights used in correction
evaluation_avg.width_weighted_credible_intervals.R	R-Code for the evaluation of average width and weighted credible intervals
plot_coverage_rates_avg.width_weighted_credible_intervals.R	R-Code for the calculation of average width and weighted credible intervals
plot_cellwise_coverage_rates_avg.width_weighted_CI.R	R-Code for the plot of the cell-wise coverage rates for the average width and weighted credible intervals
evaluation_cellwise_deviation.R	R-Code for the evaluation of the cell-wise deviation
plot_example_cellwise_deviation.R	R-Code for the example plot displaying the cell-wise deviation
Data-Folder	Folder containing the required data for the plots and calculations
Chapter 7	
convergence_diagnostics_example_application.R	R-Code for trace plots
plot_results_example_application.R	R-Code for the plots displaying the results of the example application
model_average_width_credible_intervals.R	R-Code for the model required to create average width credible intervals
parametric_bootstrap_example_application.r	R-Code required for the creation of bias corrected credible intervals
calculate+plot_credible_intervals.R	R-Code for the calculation and plots for all discussed credible intervals

Data-Folder	Folder containing the required data for the plots and calculations
Appendix	
plot_excursus_bias_telephone_coverage_rates.R	R-Code for the plot showing the coverage rates for the excursus regarding the biased telephone survey
plot_excursus_bias_telephone_example_cw_deviation.R	R-Code for the plot showing an example of the cellwise deviations of a model estimated with data from the biased telephone survey
plot_excursus_fixed_vs_fixedaverage.R	R-Code for the plots required for the excursus discussing the combination of fixed transition probabilities and fixed vote shares
Data-Folder	Folder containing the required data for the plots and calculations
Plots	
.PDF-Files	Plots used in the Masterthesis
Models	
.R-Files	R-Code for the ecological inference models
Summaries	
.RData-Files	Summaries of the ecological inference models calculated for the master thesis
Functions	
functions_masterthesis.R	R-Code with functions required for the simulation (Approach 1) and estimation of voter transitions as well as for the evaluation of credible intervals
functions_masterthesis_approach2.R	R-Code with functions required for the simulation (Approach 2), estimation and estimation of voter transitions as well as for the evaluation of credible intervals

functions_masterthesis_approach3.R	R-Code with functions required for the simulation (Approach 3), estimation and estimation of voter transitions as well as for the evaluation of credible intervals
function_cellwise_deviation.R	R-Code with function required for the evaluation of the cell-wise deviation
functions_MSE_interval_width_AD.R	R-Code with functions required for the calculation of the MSE, AD and interval width
functions_avg.width_weighted_credible_intervals.R	R-Code with functions required for the calculation and evaluation of average width and weighted credible intervals
functions_bias_corrected_credible_intervals.R	R-Code with functions required for the calculation and evaluation bias corrected credible intervals

A.12. Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Jan Moritz Klein

Place, Date