Analysis of large data sets: Bayesian methods and applications in energy and health economics

Dissertation

zur

Erlangung des Doktorgrades

Dr. rer. pol.

der Fakultät für Wirtschaftswissenschaften der Universität Duisburg-Essen

vorgelegt von

Matthias Kaeding

aus Göttingen

Betreuer:
Prof. Dr. Christoph Hanck
Lehrstuhl für Ökonometrie

Essen, Juli 2020

DuEPublico

Duisburg-Essen Publications online

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

ub | universitäts bibliothek

Diese Dissertation wird via DuEPublico, dem Dokumenten- und Publikationsserver der Universität Duisburg-Essen, zur Verfügung gestellt und liegt auch als Print-Version vor.

DOI: 10.17185/duepublico/74250
URN: urn:nbn:de:hbz:464-20210615-105920-4

Dieses Werk kann unter einer Creative Commons Namensnennung - Nicht kommerziell - Keine Bearbeitungen 4.0 Lizenz (CC BY-NC-ND 4.0) genutzt werden.

Gutachter:
Prof. Dr. Christoph Hanck
Prof. Dr. Dr. h. c. Christoph M. Schmidt

Tag der mündlichen Prüfung: 01.02.2021

If if if... doesn‘t exist.

RAFAEL NADAL

Preface

I wrote this thesis while working at the Research Data Center at the RWI – Leibniz Institute for Economic Research, half of the essays are written in conjunction with colleagues from the RWI. I want to thank several people essential in writing of this thesis:

- I owe my deepest gratitude to my Ph.D. advisor Christoph Hanck for his excellent supervision, his very helpful, thorough and fast feedback and the freedom in choosing my research topics.
- I am very grateful to Christoph M. Schmidt for being my second supervisor and for creating a stimulating and open research environment at the RWI, from which I benefited extensively.
- I would like to thank everyone at the Research Data Center for helpful discussions, in particular Philipp Breidenbach and Sandra Schaffner.
- Furthermore, I would like to thank my office mates Fabian Dehos and Lea Eilers for helpful comments and lots of great talks.
- Thank you to all my coauthors: Manuel Frondel, Alexander Haering, Stephan Sommer and Anna Werbeck, who taught me a lot about their respective fields and who allowed me to benefit from their expertise and ideas.
- Special thanks to my parents Ingeburg and Jürgen, and to my sister Laura.
- Most importantly, I thank Rilana. For everything.

Contents

Chapter 1

Introduction

The availability of large data sets is increasing dramatically, reshaping decision-making in many domains, such as energy, education and health. Data sets may be large in two dimensions: in the number of observations and in the number of variables. This thesis mainly deals with the first case. For the purpose of this dissertation a data set is large when its size causes problems for statistical inference. Such data sets may consist of structured data, where the distinction between observation and variable is clear, such as insurance data or high resolution population data. Large data sets may also consist of unstructured data, such as data from social networks sites or news articles.

Often, large data sets arise as a byproduct of emerging technologies, possibly allowing very detailed measurements in space or time. For instance, current continuous glucose monitoring systems measure blood sugar levels every five minutes, while smartphone data may provide precise information on the location of persons. It is common that large data sets provide information on every element of interest, instead of a subsample of elements. For instance, German gas stations are required by law to report gas prices immediately, resultant in a data set giving nearly complete price information.

While existing statistical methods were not developed for small data sets per se, their direct application to large data sets is often problematic, even though many methods are justified by large sample asymptotics. These problems may be inferential, for instance common testing procedures may break down in practice. This thesis is mainly concerned with computational problems, as common estimation algorithms are often too time or memory consuming to use with large data sets.

The analysis of large data sets using the appropriate methodology allows researchers to ask new kinds of research questions or to recast old ones, benefiting from the resultant statistical precision. Furthermore, large data methods offer useful tools to solve applied problems.

This thesis aims to contribute to the statistical analysis of large data sets. These contributions are twofold: First, this thesis lightens the computational demands of existing methods, improving applicability to large-scale problems. Second, this thesis uses large data methods to solve problems with important policy implications in energy and health economics. Consequently, this thesis is divided into a methodological and an econometric part, where each part consists of two essays.

The first part consists of two single-authored essays developing statistical methods for Bayesian analysis of large data sets: Somewhat paradoxically, large-scale inference often involves including additional information to obtain useful

estimates. Consider the estimation of a flexible regression model with nonlinear or spatial components. Such a model may involve a high dimensional parameter vector and may require a large data set. Here, imposing some structure on the estimates via regularization improves or enables inference. For instance, shrinking regression coefficients towards zero, or setting some exactly to zero, allows estimation of regression coefficients when the number of covariates exceeds the number of observations. In this case, an unadjusted least squares estimator is unusable. Another important example is function estimation via splines, where reducing deviations of neighboring spline coefficients leads to more realistic function estimates. Bayesian inference offers a natural way to incorporate regularization via the prior distribution, representing prior knowledge. As such, Bayesian inference is well-equipped to handle large-scale problems in theory. However, existing Bayesian methods are often computationally too demanding for such problems. The essays in the first part of this thesis aim to decrease computational demands of Bayesian methods: Chapter 3 streamlines Bayesian hazard regression, Chapter 4 accelerates a simulation algorithm of central importance for Bayesian inference.

The second part of this thesis consists of two co-authored essays in health and energy economics. The studies apply large data methods to research questions with important policy implications and thus demonstrate the potential of applying these new developments to economic policy debates. In particular, the second part answers questions that could not be answered without access to new data and methods developed for large data sets: Chapter 5 requires precise prediction of negative electricity prices, which occur very rarely. To achieve this, the chapter uses a machine learning model. Chapter 6 involves the use of a fast approximation algorithm to determine an optimal allocation of general practitioners on a small scale in Germany. The remainder of this introduction summarizes each essay, Chapter 2 gives technical preliminaries necessary for the essays.

Chapter 3 accelerates and simplifies inference for Bayesian hazard regression. The hazard rate gives the instantaneous rate of failure at t, conditional on survival until t and covariate values of zero. Important applications of hazard regression are the analysis of unemployment durations or the modeling of time until death. Covariate effects for hazard regression are hard to interpret, Bayesian approaches are computationally expensive. Usually no closed form expression for the hazard regression likelihood is available, requiring an approximation which may be expensive to compute. This chapter aims to alleviate these problems by modeling the integrated baseline hazard via monotonic penalized B-splines (P-splines). B-splines are functions, formed from connected local polynomials. The basic idea of Bayesian P-splines is to model an unknown function via a weighted sum of B-splines, controlling smoothness via an estimated penalty parameter.

The presented modeling strategy gives a closed form expression for the likelihood, allows accounting for arbitrary censorship and the inclusion of further nonparametric components modeled by P-splines. Because of the good numerical properties of B-splines, involved computations are fast and stable. Furthermore, the computational advantages allow easy effect interpretation by combination of two concepts: Partial dependence plots show the relationship between a covariate and an estimand, marginalizing over all covariates except the one of interest. Here, the estimand is the restricted mean survival time, easy to compute in the presented framework. This allows effect interpretation in terms of survival times instead of the hazard rate.

Monte Carlo simulations show that the modeling strategy works well, if the sample size is large enough. An application using a large real estate data set shows that the presented methods give useful results in practice.

Chapter 4 develops a fast, approximate Metropolis-Hastings algorithm for Bayesian analysis of large data sets. Because the Metropolis-Hastings algorithm allows sampling from the posterior distribution if no complete analytic expression is available, the algorithm is of central importance in Bayesian statistics. Running the algorithm involves computation of loglikelihood-ratio sums, so that the run time of the algorithm is usually linear in the sample size. This makes inference with large data sets impractical. The chapter proposes a fast approximation embedded in a sampling design framework, which is concerned with the estimation of finite population sums. Here, the outcome of interest is the sum of loglikelihood-ratio differences.

The presented algorithm is based on a single subsample, bypassing the need to store the complete data set. The subsample is taken via the cube method, a balanced sampling method where a random sample is drawn so that the mean of a set of auxiliary variables is close to the population mean. This reduces the variance of the sample mean, if the auxiliary variables are correlated with the outcome variable. Here, the auxiliary variables satisfy this condition by design. Furthermore, presence of the auxiliary variables allows using variants of the regression estimator: This estimator uses the auxiliary information to predict the estimation error, which is subtracted from the estimate.

An application using a data set consisting of 31 million rows and simulation studies show that the presented approach can lead to a strong reduction in computation time. Results are very close to those obtained using the complete data set.

Chapter 5, co-authored by Manuel Frondel and Stephan Sommer, models the effect of the introduction and amendment of a market premia scheme on the occurrence of negative electricity prices. In Germany, plant operators using renewable technologies could obtain guaranteed fixed payments. As this so-called feed-in-tariff is paid independent of the market price of electricity, it may result in negative prices when high electricity supply coincides with low demand, for instance in the morning hours or during national holidays.

Energy oversupply causes negative electricity prices, in turn causing instability of the electricity grid, increasing grid maintenance cost. Furthermore, negative electricity prices induce welfare losses: Because of substantial ramp-up costs, providers of conventional plants usually do not halt production when revenues are negative. The increasing share of electricity from renewable energy sources magnifies these problems. To counteract, a market premia scheme was introduced, aiming to align electricity production with demand. The market premia scheme creates incentives for plant operators to sell electricity directly to customers at variable wholesale prices, instead of a fixed price.

Because of the small share of negative prices overall (less than one percent), and due to the presence of nonlinear interactions, common statistical methods may fail in predicting negative prices. Consequently, our modeling strategy is based on Bayesian Additive Regression Trees (BART), a machine learning technique known to perform well in such settings. We find that BART precisely identifies potential negative price spikes, unlike common methods such as linear regression. We use the BART model to simulate electricity prices under coun-

terfactual settings, finding that the introduction of a market premia scheme is associated with a reduction of negative electricity prices by some 70%.

Chapter 6, co-authored by Alexander Haering and Anna Werbeck, creates a benchmark for the spatial allocation of general practitioners (GPs) in Germany. Unlike existing approaches, we use a large data set with detailed population information for $1km^2$ grids with information on the current allocation of GPs. Furthermore, we use realistic estimates of driving times, incorporating detailed road information. Based on German regulation, an optimal allocation minimizes the number of GPs under two constraints: Every person reaches at least one GP by car in less than 15 minutes, and no GP services more than 13,000 patient visits per year. Because of the high dimensionality of the solution set, determination of the optimal allocation is not feasible. Instead, we use a fast approximation algorithm.

For all Germany, we find a deficit in GPs of some 6%. Using a spatial linear regression model, we analyze regional patterns by comparing the optimal allocation with the status quo. We use two indices defined on a municipality level: number of excess GPs (supply side) and percentage of uncovered cases (demand side). We find a surplus of GPs in cities, compared to rural and suburban municipalities. However, this does not translate to the demand side, where we do not find an association of healthcare demand with density.

The remainder of this thesis is structured as follows: Chapter 2 gives technical preliminaries, Chapters 3 to 6 give the main results. Chapter 7 concludes.

Chapter 2

Technical preliminaries

2.1 Bayesian inference

Let $\boldsymbol{\theta} \in \Theta$ denote the parameter vector of interest and $\mathcal{D}$ denote the data. The probability density function of $\boldsymbol{\theta}$, conditional on $\mathcal{D}$, the so-called posterior distribution, is the basis for Bayesian inference. The posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ arises by combining the prior distribution $p(\boldsymbol{\theta})$, representing available information about $\boldsymbol{\theta}$ before seeing the data, with the likelihood $L(\boldsymbol{\theta}|\mathcal{D})$. This gives

$$p(\boldsymbol{\theta}|\mathcal{D}) = kL(\boldsymbol{\theta}|\mathcal{D})p(\boldsymbol{\theta}),$$

with normalizing constant

$$k \coloneqq \left(\int L(\boldsymbol{\theta}|\mathcal{D})p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right)^{-1}.$$

The crucial steps for Bayesian inference are the choice of prior and likelihood, and the summary of the resultant posterior distribution. For a comprehensive overview of Bayesian inference see Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2013) or Robert (2001).

2.1.1 Metropolis-Hastings algorithm

In practice, the proportionality constant k is usually unknown except for very simple models, so that the posterior distribution is not fully available. As an alternative, Markov chain Monte Carlo (MCMC) methods create a sequence of parameter draws $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$, which mimic samples drawn from the posterior distribution. Inference is then based on those parameter draws, which form a Markov chain: the distribution of $\boldsymbol{\theta}^{(t)}$ given $\boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(1)}$ depends only on $\boldsymbol{\theta}^{(t-1)}$. It holds that

$$P(\boldsymbol{\theta}^{(t)} \in A|\boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(1)}) = P(\boldsymbol{\theta}^{(t)} \in A|\boldsymbol{\theta}^{(t-1)}) = \int_A K(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})d\boldsymbol{\theta}^{(t)},$$

where the transition kernel $K(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})$ denotes the conditional density of $\boldsymbol{\theta}^{(t)}$, given $\boldsymbol{\theta}^{(t-1)}$. The transition kernel is a representation of a Markov chain, it is chosen so that output of the chain mimics output from the posterior distribution. To achieve this, the invariant distribution of the Markov chain must be the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$, then

$$p(\boldsymbol{\theta}^{(t)}|\mathcal{D}) = \int K(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})p(\boldsymbol{\theta}^{(t-1)}|\mathcal{D})d\boldsymbol{\theta}^{(t-1)} \tag{2.1}$$

holds. It follows from equation (2.1) that, if $\boldsymbol{\theta}^{(t)}$ is sampled from the posterior distribution, all subsequent samples are as well. Detailed balance

$$p(\boldsymbol{\theta}^{(t)}|\mathcal{D})K(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(t)}) = p(\boldsymbol{\theta}^{(t-1)}|\mathcal{D})K(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}) \tag{2.2}$$

is a sufficient, but not necessary condition to ensure the posterior distribution as invariant distribution of a Markov chain. This follows by integrating equation (2.2) over $\boldsymbol{\theta}^{(t-1)}$. Then, the left hand side becomes

$$\int p(\boldsymbol{\theta}^{(t)}|\mathcal{D})K(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(t)})d\boldsymbol{\theta}^{(t-1)} = p(\boldsymbol{\theta}^{(t)}|\mathcal{D}) \int K(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(t)})d\boldsymbol{\theta}^{(t-1)} = p(\boldsymbol{\theta}^{(t)}|\mathcal{D}), \tag{2.3}$$

since K is a density, so that the condition in (2.1) is satisfied. See Andrieu, De Freitas, Doucet, and Jordan (2003) or Robert and Casella (2004) for an introduction to Markov chain Monte Carlo methods.

This thesis uses as main MCMC method the Metropolis-Hastings algorithm, or variants thereof. The algorithm allows obtaining draws from a density only known up to a proportionality, such as the posterior distribution. The main ingredient is the proposal distribution g, from which one can sample. When used for the simulation of the posterior distribution, the Metropolis-Hastings algorithm consists of the following steps:

1. Choose start values $\boldsymbol{\theta}^{(0)}$ and proposal distribution g
2. For $t = 1, \ldots, T$:
   - Draw a proposal $\boldsymbol{\theta}^{\star}$ from $g(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t-1)})$
   - Compute

     $$\alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{\star}) = \min\left(1, \frac{p(\boldsymbol{\theta}^{\star}|\mathcal{D})g(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{\star})}{p(\boldsymbol{\theta}^{(t-1)}|\mathcal{D})g(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t-1)})}\right),$$

     Accept the proposal with probability $\alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{\star})$; set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{\star}$. Otherwise, reject the proposal; set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
3. Return $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$

The normalizing constant k, cancels out in the acceptance ratio $\alpha(\cdot, \cdot)$, so that k may be unknown, for instance $k = (2\sqrt{\pi})^{-1}$ for a standard normal distribution. The transition kernel of the Metropolis-Hastings algorithm is

$$K_{MH}(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}) = g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})\alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{(t)}) + \tag{2.4}$$

$$I[\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}] \int g(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t-1)})(1 - \alpha(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{\star}))d\boldsymbol{\theta}^{\star}, \tag{2.5}$$

where the right hand side of equation (2.4) corresponds to acceptance, and equation (2.5) corresponds to rejection of a proposal. The kernel satisfies the detailed balance condition, so that the posterior distribution is the invariant distribution of the chain.

Chapter 5 uses the popular Gibbs sampler to draw from the posterior distribution. Let

$$\boldsymbol{\theta}^{(t)}_{-i} \coloneqq (\theta^{(t)}_1, \theta^{(t)}_2, \ldots, \theta^{(t)}_{i-1}, \theta^{(t-1)}_{i+1}, \ldots, \theta^{(t-1)}_I)^{\top}$$

denote the parameter vector, without the subvector i, at iteration t while running a Gibbs-sampler. The main requirement for use of the Gibbs sampler is the

ability to draw from the conditional density $p(\boldsymbol{\theta}_i^{(t)}|\boldsymbol{\theta}_{-i}^{(t)}, \mathcal{D})$ for $i = 1, \ldots, I$, i.e., the posterior distribution of the subvector $\boldsymbol{\theta}_i^{(t)}$, conditional on all other parameters. The sampler iterates the following steps:

1. Choose start values $\boldsymbol{\theta}^{(0)}$
2. For $t = 1, \ldots, T$:
   - For $i = 1, \ldots, I$:
     - Draw $\boldsymbol{\theta}_i^{(t)}$ from $p(\boldsymbol{\theta}_i^{(t)}|\boldsymbol{\theta}_{-i}^{(t)}, \mathcal{D})$
3. Return $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm with proposal distribution $p(\boldsymbol{\theta}_i^{(t)}|\boldsymbol{\theta}_{-i}^{(t)}, \mathcal{D})$.

2.1.2 Point and interval estimation

Under the Bayesian decision theoretical approach, point estimators arise by defining a loss function $l(\gamma, \boldsymbol{\theta})$, and minimizing the posterior expected loss

$$\int_{\Theta} l(\boldsymbol{\gamma}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}, \tag{2.6}$$

where $\boldsymbol{\gamma}$ is the true parameter vector. Because $\boldsymbol{\gamma}$ is unknown, the Bayesian approach integrates over the parameter space Θ, weighting the loss $l(\boldsymbol{\gamma}, \boldsymbol{\theta})$ via the posterior distribution of the unknown parameters, conditional on the known data. See Robert (2001) or Lehmann and Casella (1998) for decision-theoretic background of Bayesian statistics. The choice of loss function determines the estimator as minimizer of (2.6). This thesis uses mainly the posterior mean as point estimate, corresponding to quadratic loss $l(\boldsymbol{\gamma}, \boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\gamma})^2$. Another popular choice is the posterior median, corresponding to absolute loss.

Using MCMC output, an estimator is computed from its empirical equivalent, e.g. the average

$$T^{-1} \sum_{t=1}^{T} h(\boldsymbol{\theta}^{(t)}) \tag{2.7}$$

is used to estimate the posterior mean of some function $h(\boldsymbol{\theta})$. The most common case is $h(\boldsymbol{\theta}) = \boldsymbol{\theta}$, but sometimes h is more complex: in Chapter 3, one quantity of interest is $h(\boldsymbol{\theta}) = \exp(-\exp(\boldsymbol{b}^\top \boldsymbol{\theta}))$. If some regulatory conditions are satisfied, averages such as (2.7) converge to the expectation $E_{p(\boldsymbol{\theta}|\mathcal{D})}[h(\boldsymbol{\theta})]$.

Posterior intervals are used to communicate uncertainty. Interval estimates are derived in analogy to point estimators from simulation output: We can estimate the interval $[L, U]$, so that $P(h(\boldsymbol{\theta}) > L) = P(h(\boldsymbol{\theta}) < U) = \alpha$, via the interval formed by the α and $1-\alpha$ quantile from the simulated values $h(\boldsymbol{\theta}^{(1)}), \ldots, h(\boldsymbol{\theta}^{(T)})$.

2.2 Set cover problem

The set cover problem is a problem from combinatorics with wide ranging applications such as personnel shift planning, virus detection or location planning. It is of central importance in the field of approximation algorithms, because many problems can be cast as set cover problem instance or relate directly to the problem. This is the case for Chapter 6, where a variant of the set cover problem is applied to finding an allocation of general practitioners in Germany.

The set cover problem presents as follows: The input is a collection of sets $\mathcal{S} = \{S_1, \ldots, S_n\}$ with cost (or weight) function $c : \mathcal{S} \to \mathbb{R}_{\geq 0}$. The collection $\mathcal{S}$ covers a universe $\mathcal{U}$, i.e.,

$$S_1 \cup S_2 \cup \cdots \cup S_n = \mathcal{U}.$$

The aim is to find the minimal cost subcollection $\mathcal{X} \subseteq \mathcal{S}$, so that $\mathcal{X}$ covers $\mathcal{U}$. For the unweighted set cover problem, it holds that

$$c(S_1) = c(S_2) = \cdots = c(S_n),$$

so $\mathcal{X}$ is the subcollection with the smallest number of sets, covering $\mathcal{U}$. For an overview of the problem see Korte and Vygen (2018) or Vazirani (2001). Because the solution set is finite, one can always find the optimal solution to the problem in theory, for instance by brute force or linear programming. However, in practice the size of the solution often makes this infeasible due to the resultant computational demands. It turns out that a greedy approximation algorithm gives a solution, which is hard to improve. The basic idea of the approximation is to find the set with highest cost-effectiveness in each iteration. The algorithm consists of the following steps:

1. Set $C = \{\}$ and $\mathcal{X} = \{\}$
2. While $\mathcal{X}$ does not cover $\mathcal{U}$:
   - Find the set R with the least cost per uncovered element, i.e. the set with smallest value of
     $$\frac{c(R)}{|R - C|}, \text{ given that } |R - C| > 0,$$
     where $|R - C|$ denotes the cardinality of the set $R - C$.
   - Set $C = C \cup R$ and $\mathcal{X} = \mathcal{X} \cup \{R\}$
3. Return $\mathcal{X}$

For the unweighted case, the set with the highest cost-effectiveness is always the set with the largest number of uncovered elements. Let $\mathrm{OPT}(\mathcal{U}, \mathcal{S}, c)$ be the cost of the optimal solution for a problem instance with universe $\mathcal{U}$, collection $\mathcal{S}$ and cost function c. Chvatal (1979) shows that the cost of the solution from the greedy approximation is bounded from above by the factor

$$\mathrm{OPT}(\mathcal{U}, \mathcal{S}, c)\Big[1 + \frac{1}{2} + \ldots \frac{1}{R}\Big]$$

where $R \coloneqq \max(|S_1|, |S_2|, \ldots, |S_n|)$ and $|S_i|$ denotes the cardinality of S_i. Vazirani (2001, chap. 29) shows rigorously why it is very hard to improve upon the somewhat obvious greedy approximation algorithm.

Part I

Bayesian methods

Chapter 3

Efficient Bayesian nonparametric hazard regression[1]

Abstract

We model the log-cumulative baseline hazard for the Cox model via Bayesian, monotonic P-splines. This approach permits fast computation, accounting for arbitrary censorship and the inclusion of nonparametric effects. We leverage the computational efficiency to simplify effect interpretation for metric and non-metric variables by combining the restricted mean survival time approach with partial dependence plots. This allows effect interpretation in terms of survival times. Monte Carlo simulations indicate that the proposed methods work well. We illustrate our approach using a large data set of real estate data advertisements.

Keywords: Bayesian survival analysis; Nonparametric modeling; Penalized spline; Restricted mean survival time

[1]This chapter has been published as: Kaeding, M. (2020). Efficient Bayesian nonparametric hazard regression. *Ruhr Economic Papers, 850*, 1–21. doi:10.4419/86788985

3.1 Introduction

In economic, epidemiological and engineering applications, the Cox proportional hazards model is the benchmark for survival analysis. However, nonparametric modeling strategies for the Cox model do not scale up to large data sets. This paper aims to alleviate this problem by speeding up computation. The baseline hazard $h_0(t)$ is the key concept for the Cox model. It gives the instantaneous rate of failure at t, conditional on survival until t and covariate values of zero. We propose to model the log-integrated baseline hazard via Bayesian, monotonic penalized B-splines. As we can evaluate the likelihood analytically, and due to the benefits of Bayesian P-splines, our approach holds five key advantages: (1) Fast, automatic computation. (2) Exact likelihood calculation. (3) Accounting for arbitrary censoring. (4) Inclusion of nonparametric components. (5) Easier effect interpretation in regards to survival times, not hazard rates.

Most Bayesian non- or semiparametric approaches use a flexible model for some functional of the baseline hazard: Hennerfeind, Brezger, and Fahrmeir (2006) use P-splines for the (log) baseline hazard, Fernandez, Rivera, and Teh (2016) use a Gaussian process. Because the likelihood is usually not analytically available under this strategy, numerical integration is necessary, introducing approximation error and slowing down inference. There are approaches where this does not apply, as the likelihood is analytically available: Kalbfleisch (1978) uses the gamma process prior, Dykstra and Laud (1981) use the extended gamma process prior. Gelfand and Mallick (1995) use a mixture of Beta densities, Nieto-Barajas and Walker (2002) use a Markov increment prior. Cai, Lin, and Wang (2011) and Lin, Cai, Wang, and Zhang (2015) use monotone regression splines for left- or right censored data. Zhou and Hanson (2018) use a Bernstein polynomial prior for arbitrary censored data. In a frequentist context, Royston and Parmar (2002) use natural cubic splines for left- or right censored data, Zhang, Hua, and Huang (2010) use monotone B-splines for interval censored data,

However, these approaches are either computationally expensive, not flexible enough or only cover special cases, which does not apply to the estimation strategy proposed here.

The paper is structured as follows: section 3.2 gives the modeling approach, section 3.3 details inference. Section 3.4 shows a simulation study, section 3.5 applies the methods to real estate data. Section 3.6 concludes.

3.2 Hazard regression model

In hazard regression, the modeling of survival times is of interest, for instance unemployment durations or time until death. A non-negative random variable T with density $s(t)$ and survival function $S(t) = P(T > t)$ represents survival time. The hazard rate $h(\cdot)$ is the conditional density of T, given that $T > t$, so that

$$h(t) = s(t|T > t) = s(t)/S(t).$$

It holds that

$$S(t) = \exp(-H(t)), \text{ where } H(t) \coloneqq \int_0^t h(u)du$$

is the cumulative (or integrated) hazard, so that h uniquely determines T.

Under interval censoring, we observe data

$$\mathcal{D} = \{(y_i, \boldsymbol{x}_i), i = 1, \ldots, N\},$$

where $\boldsymbol{x}_i$ is a covariate vector and $y_i = [t_i^-, t_i^+)$ denotes the interval containing the true survival time. Left censoring is a special case with lower bound $t_i^- = 0$, right censoring is a special case with upper bound $t_i^+ = \infty$. By convention, we write $t_i^- = t_i^+ = t_i$, for an uncensored survival time.

The benchmark model for survival times is the semiparametric Cox model (Cox, 1972) with conditional survival function

$$S(t_i|\boldsymbol{x}_i, \boldsymbol{\alpha}) = \exp(-\exp\left(\log(H_0(t_i)) + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right)),$$

where $H_0(t_i)$ is the unspecified cumulative baseline hazard

$$H_0(t_i) = \int_0^{t_i} h_0(u)du,$$

with baseline hazard h_0. In a nonparametric setting, the model includes nonlinear effects. We partition each $\boldsymbol{x}_i$ into vectors $\boldsymbol{z}_i$ (linear effects) and $\boldsymbol{v}_i$ (nonlinear effects). Define the linear predictor

$$\xi_i = \xi(\boldsymbol{x}_i, t_i) \coloneqq f_0(t_i) + f_1(v_{i,1}) + \cdots + f_R(v_{i,R}) + \boldsymbol{z}_i^\top \boldsymbol{\alpha},$$

where $\boldsymbol{\alpha}$ is a vector of regression coefficients, $f_0(t_i) \coloneqq \log H_0(t_i)$ is the log-cumulative hazard and $f_1, \ldots, f_R$ are functions. Let $\boldsymbol{f}_r = (f_r(v_{1,r}), \ldots, f_r(v_{N,r}))^\top$ denote the vector of function evaluations for f_r and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^\top$ denote the linear predictor vector, then we can write

$$\boldsymbol{\xi} = \boldsymbol{Z}\boldsymbol{\alpha} + \sum_{r=0}^{R} \boldsymbol{f}_r,$$

where $\boldsymbol{Z}$ is a design matrix. We can write the survival function as

$$S(t_i|\xi_i) = \exp(-\exp(\xi_i)).$$

We model the log-cumulative baseline hazard f_0 via monotonic, penalized B-splines (P-splines). As Hennerfeind, Brezger, and Fahrmeir (2006), we model $f_1, \ldots, f_R$ via (non monotonic) P-splines. The basic idea of P-splines is to model a function f_r by a weighted sum of B-spline basis functions $B_{r,1}, \ldots, B_{r,J_r}$, augmenting the loss function with a penalty controlling the smoothness of the estimated function. Hence,

$$f_r(v) = \sum_{j=1}^{J_r} \beta_{r,j} B_{r,j}(v) \text{ for } r = 0, \ldots, R,$$

where $\beta_{r,1}, \ldots, \beta_{r,J_r}$ are regression coefficients associated with the function f_r, see figure 3.1. Given a knot vector $\boldsymbol{k}_r \in \mathbb{R}^m$, a B-spline $B_{r,j}(v) = B_{r,j}^{l_r}(v)$ of order $l_r = 1$ is the function

$$B_{r,j}^1(v) \coloneqq I[v \in [k_{r,j-1}, k_{r,j})],$$

where $I[\text{condition}]$ equals one if the condition is met and zero otherwise. See De Boor (1978) for a rigorous introduction to B-splines. We assume that $\boldsymbol{k}_r$ is equally spaced from the minimum to the maximum of a covariate v_r. Then B-splines of order $l_r > 1$ are defined recursively[2] as

$$B_{r,j}^{l_r}(v) = w_{r,j} B_{r,j}^{l_r-1}(v) + (1 - w_{r,j+1}) B_{r,j+1}^{l_r-1}(v), \text{ for } j = 1, \ldots, J_r,$$

[2]Due to this recursive definition, for $l_r > 1$, the knot vector needs to be extended with additional outer knots defined analogous to $\boldsymbol{k}_r$.

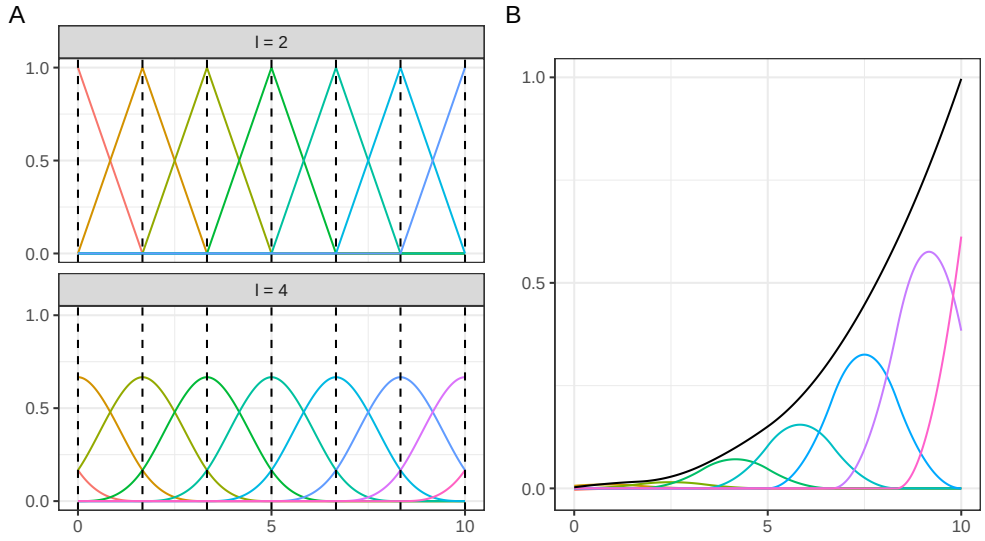


Figure 3.1: A: B-splines for varying order. Dashed, vertical lines mark the knots. B: Function obtained by weighted sum of B-splines. The red line is the estimated function, given by the sum of the scaled basis functions below, here giving a monotone estimate.

with $J_r = m + l_r - 2$ and

$$w_{r,j} := \frac{x - k_{r,j}}{(l_r - 1)h},$$

where h is the spacing between the knots. If function specific domain knowledge is available for function f_r, it can be used to set l_r, i.e., if one function is known to be a step function. Otherwise, one usually sets l_r to the smallest value where the smoothness of the estimated function is satisfactory, a good default is $l_r = 4$ for $r = 0, \ldots, R$. Define the design matrix $\boldsymbol{B}_r \in \mathbb{R}^{N \times J}$ with i, jth element $B_{r,j}(v_{i,r})$. Then we can write each vector of function evaluation as $\boldsymbol{f}_r = \boldsymbol{B}_r \boldsymbol{\beta}_r$ and represent the linear predictor vector $\boldsymbol{\xi}$ compactly as

$$\boldsymbol{\xi} = \boldsymbol{Z}\boldsymbol{\alpha} + \sum_{r=0}^{R} \boldsymbol{B}_r \boldsymbol{\beta}_r.$$

Because B-splines vanish outside a domain spanned by $l_r - 1 + 2$ knots (see figure 3.1), the matrices $\boldsymbol{B}_0, \ldots, \boldsymbol{B}_R$ are sparse. Hence the computation of the linear predictor vector $\boldsymbol{\xi}$ is fast.

3.2.1 Priors

The flexibility of the B-splines basis increases with m, the number of knots, which determines J_r, the number of basis functions for function f_r. For a large number of knots, the B-spline fit approaches a rough interpolation of the data which is usually undesired behaviour. Varying m on the fly changes the number of parameters, complicating inference. Using penalization, one can use fixed, large m, say $m = 30$ (so that we obtain a flexible fit) and control the smoothness of the estimated function by a single parameter penalizing unsmooth function estimates. See figure 3.2 for a demonstration. In a Bayesian context, this is handled by the prior distribution of $\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_R$ and the associated penalty parameters. As a result, we can directly obtain precision measures for function estimates from the

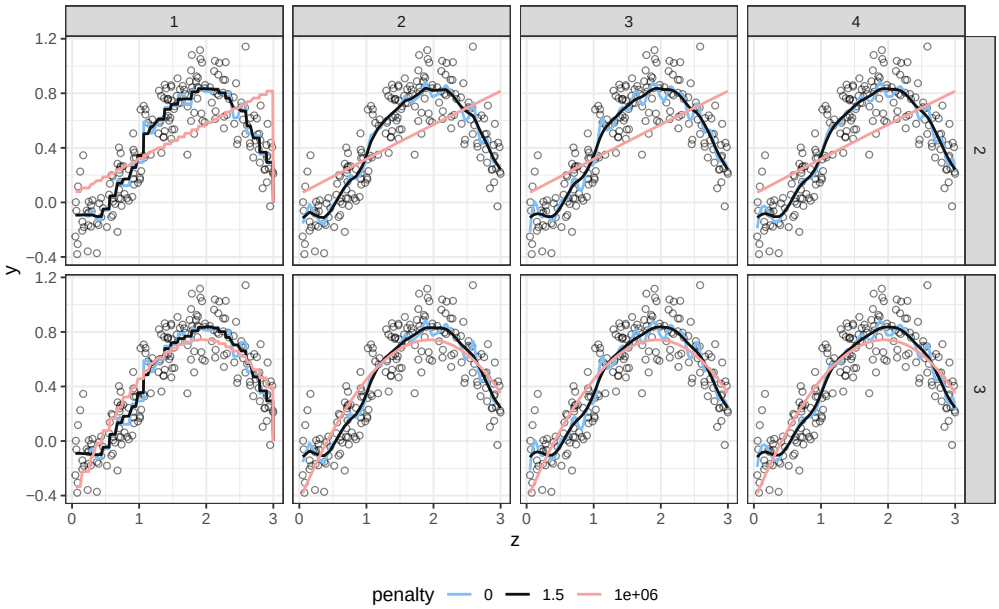


Figure 3.2: Influence of order l, difference order d and penalty parameter on function estimates. Data is simulated via $y_i \sim N(\sin(z_i)\log(z_i + 0.5), 0.15^2)$. Rows are varying values of d, columns are varying values of l. Lines are the estimated function under varying penalty parameter. Without a penalty, the estimated function is very unsmooth, for a large penalty the function estimate approaches a polynomial of degree $d - 1$.

posterior distribution. Furthermore, inference is automatic, in the sense that no post-processing such as cross validation is necessary. Let Δ^d be the difference operator of order d, defined recursively by

$$\begin{aligned}\Delta^1 \beta_{r,j} &:= \beta_{r,j} - \beta_{r,j-1}, \\ \Delta^d &:= \Delta^1 \Delta^{d-1} \text{ for } d > 1.\end{aligned}$$

The curve fitting literature uses the squared dth derivative of the estimated function as smoothness penalty. Eilers and Marx (1996) show that

$$\lambda_r \sum_{j=d}^{J} (\Delta^d \beta_{r,j})^2, \tag{3.1}$$

approximates this smoothness penalty. As such, λ_r controls the smoothness of the estimated function $\hat{f}_r$. For $\lambda_r \to \infty$ the estimated function approaches a polynomial of degree $d - 1$. Increasing d results in smoother estimates. A value of $d > 3$ is rarely used. We use $d = 3$ as default option and assume that d is the same for $\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_R$. We use the prior distribution from Lang and Brezger (2004), who base their prior on (3.1). Here

$$\Delta^d \beta_{r,j} \sim N(0, \tau_r^2) \text{ for } j > d,$$

so that for instance

$$\begin{aligned}\beta_{r,j} &= \beta_{r,j-1} + e_{r,j}, \text{for a difference of order } d = 1, \\ \beta_{r,j} &= 2\beta_{r,j-1} - \beta_{r,j-2} + e_{r,j} \text{ for a difference of order } d = 2, \\ \beta_{r,j} &= 3\beta_{r,j-1} - 3\beta_{r,j-2} + \beta_{r,j-3} + e_{r,j} \text{ for a difference of order } d = 3.\end{aligned}$$

with $e_{r,j} \sim N(0, \tau_r^2)$ for $j = d, \ldots, J_r$. A high standard deviation τ_r, indicating an unsmooth function, is associated with a low λ_r. Parameters $\beta_{0,1}, \ldots, \beta_{0,d}, \beta_{1,1}, \ldots, \beta_{1,d} \ldots \beta_{R,d}$ are assigned a flat prior $p(\cdot) \propto 1$. Let $\boldsymbol{D}_r^d \in \mathbb{R}^{(J_r-d)\times J_r}$ denote a matrix representation of the difference operator for function f_r. Then, element j of $\boldsymbol{D}_r^d \boldsymbol{\beta}_r$ is $\Delta^d \beta_{r,j}$ and

$$\boldsymbol{\beta}_r^\top \boldsymbol{K}_d \boldsymbol{\beta}_r = \sum_{j=d+1}^{J_r} (\Delta^d \beta_{r,j})^2,$$

where $\boldsymbol{K}_d = \boldsymbol{D}_d^\top \boldsymbol{D}_d$ is the penalty matrix. For instance, for $d = 2$, we have

$$\boldsymbol{D}_2 = \begin{bmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{bmatrix}.$$

and

$$\boldsymbol{K}_2 = \begin{bmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix}.$$

We can write the prior for $\boldsymbol{\beta}_r$ as

$$p(\boldsymbol{\beta}_r|\tau_r) \propto \exp(-\frac{1}{2\tau_r^2} \boldsymbol{\beta}_r^\top \boldsymbol{K}_d \boldsymbol{\beta}_r), \text{ for } r = 1, \ldots, R. \tag{3.2}$$

Because $\boldsymbol{K}_d$ is a sparse band matrix with range $d + 1$, we can exploit sparse matrix operations to compute the quadratic form $\boldsymbol{\beta}_r^\top \boldsymbol{K}_d \boldsymbol{\beta}_r$ in (3.2).

Some adjustments are necessary for modeling the log-cumulative baseline hazard f_0 via P-splines. Because the cumulative baseline hazard is defined on $[0, t_i]$, the knot vector is a sequence from 0 to the largest $t_i^+ < \infty$. To achieve a monotonic function estimate for the log-cumulative baseline hazard[3], we restrict the prior (3.2) to non-decreasing vectors, resultant in a monotonic function estimate as Brezger and Steiner (2008) show:

$$p_0(\boldsymbol{\beta}_0|\tau_0) = p(\boldsymbol{\beta}_0|\tau_0) I[\beta_{0,1} \leq \beta_{0,2} \leq \cdots \leq \beta_{0,J_0}]. \tag{3.3}$$

We can extend this to further nonlinear effects if a monotonic function estimate for $f_1, \ldots, f_R$ is desired. We assign a flat prior to regression coefficients $\boldsymbol{\alpha}$ associated with linear effects. For positive scale parameters such as τ_r, a popular[4] choice is an inverted gamma prior, see for instance Hennerfeind, Brezger, and Fahrmeir (2006) or Kneib and Fahrmeir (2007). We follow Gelman (2006), who recommends a half Cauchy prior instead:

$$p(\tau_r|\phi_r) \propto I[\tau_r > 0](1 + (\tau_r/\phi_r)^2)^{-1} \text{ for } r = 0, \ldots, R,$$

[3] One might also model the cumulative baseline hazard, this involves an additional positivity restriction on $\boldsymbol{\beta}_0$. We tried this but the Hamiltonian Monte Carlo sampler converged slowly.

[4] The popularity of the inverted gamma prior is probably due to its convenience under a Gaussian model, so that it has become somewhat of a default.

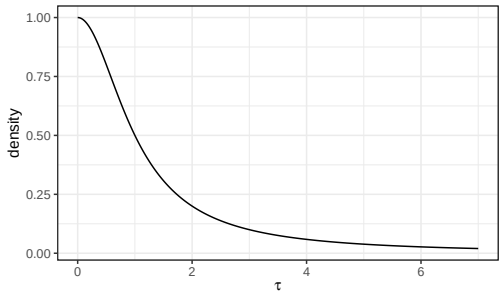


Figure 3.3: Half-Cauchy prior for τ_r with $\phi_r = 1$.

with low scale parameter $\phi_r = 1$ as default option. This puts most prior mass on smooth functions, i.e., those with low τ_r. Due to the heavy tails of the Cauchy distribution, τ_r may be large, resultant in less smooth function estimates if the data demands it. We estimate τ_r from the data, so that the parameter adjusts to the number of B-splines.

3.2.2 Likelihood construction

We use P-splines to model the log-cumulative baseline hazard, so that

$$\log H_0(t) = f_0(t) = \sum_{j=1}^{J_0} \beta_{0,j} B_{0,j}(t) \text{ and}$$

$$h_0(t) = \frac{dH_0(t)}{dt} = \exp\left(f_0(t)\right) df_0(t)/dt.$$

Because the derivative of a weighted sum of B-splines is

$$\frac{d\sum_{j=1}^{J_r} \beta_{r,j} B_{r,j}^{l_r}(v)}{dv} = h^{-1} \sum_{j=2}^{J_r} B_{r,j}^{l_r-1}(v) \Delta^1 \beta_{r,j}, \tag{3.4}$$

the baseline hazard is analytically available, resulting in a tractable likelihood. Because the computation of (3.4) involves lower order B-splines, the computational advantages of B-splines carry over to the computation of the baseline hazard. The likelihood is $L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{N} L_i$, with likelihood contributions $L_1, \ldots, L_N$ accounting for censoring. Each likelihood contribution is the probability $P(t_i \in y_i|\xi_i)$, except for uncensored survival times. Here the likelihood contribution is the density $h(t_i|\xi_i)S(t_i|\xi_i)$, so that:

$$L_i = \begin{cases} S(t_i^-|\xi_i) - S(t_i^+|\xi_i) & \text{if } t_i \text{ is interval censored,} \\ 1 - S(t_i^+|\xi_i) & \text{if } t_i \text{ is left censored,} \\ S(t_i^-|\xi_i) & \text{if } t_i \text{ is right censored,} \\ h(t_i|\xi_i)S(t_i|\xi_i) & \text{if } t_i \text{ is uncensored.} \end{cases}$$

We need to compute the log-likelihood $\mathcal{L}$ for model evaluation and to sample from the posterior distribution via Hamiltonian Monte Carlo. There are some convenient shortcuts for the computation of $\mathcal{L}$. Let $\mathcal{S}$ denote the set of all uncensored observations. Define the vectors of totals $\boldsymbol{z}^{\mathcal{S}} \coloneqq \sum_{i \in \mathcal{S}} \boldsymbol{z}_i$ and $\boldsymbol{b}_j^{\mathcal{S}} \coloneqq \sum_{i \in \mathcal{S}} \boldsymbol{b}_r(v_{i,r})$, which we have to compute only once. Then we can write the log-likelihood for the uncensored observations as the sum $\xi^{\mathcal{S}} + \sum_{i \in \mathcal{S}} \eta_i$, where

$$\xi^{\mathcal{S}} \coloneqq \boldsymbol{\alpha}^\top \boldsymbol{z}^{\mathcal{S}} + \sum_{r=0}^{R} \boldsymbol{\beta}_r^\top \boldsymbol{b}_r^{\mathcal{S}} \tag{3.5}$$

is the sum of the linear predictor vector and η_i is defined as

$$\eta_i \coloneqq \log\left(\frac{df_0(t_i)}{dt_i}\right) - \exp(\xi_i).$$

The computational cost of $\xi^{\mathcal{S}}$ does not grow with the cardinality of $\mathcal{S}$. However, this does not hold for the computation of $\sum \eta_i$, but computation of ξ_i is fast due to the sparsity of the involved vectors. This applies to the contributions of censored observations as well: Here, the likelihood contributions depend on the value of the linear predictor, with aforementioned computational advantages. For instance, the likelihood contribution is $\log\left(S(t_i^-|\xi_i)\right) = -\exp(\xi_i)$ for a right censored survival time.

3.3 Inference

We use the probabilistic programming language Stan (Carpenter, Gelman, Hoffman, Lee, Goodrich, Betancourt, Brubaker, Guo, Li, & Riddell, 2017) to sample from the posterior distribution

$$p(\boldsymbol{\theta}|\mathcal{D}) = L(\boldsymbol{\theta}|\mathcal{D})p_0(\boldsymbol{\beta}_0|\tau_0)\prod_{r=1}^{R} p(\boldsymbol{\beta}_r|\tau_r)p(\tau_r|\phi_r). \tag{3.6}$$

For point estimation we use the posterior mean, for interval estimation we use the 0.025 and 0.975 quantile. A nice feature of simulation based Bayesian inference is the option to obtain uncertainty measures directly for functions of parameters from the samples of the parameters, e.g. for $\exp(f_0) = H_0$.

Stan implements the No-U-Turn sampler for Hamiltonian Monte Carlo (Hoffman & Gelman, 2014). This sampler converges quickly for high dimensional posterior distribution of correlated parameters as for the problem at hand. It is almost fully automatic[5] and allows easy use of non-conjugate prior distributions such as the Cauchy prior for τ_r, unlike a Gibbs sampler. Furthermore, Stan supports sparse matrix operations[6].

For models with nonparametric components the means of the function $f_0, \ldots, f_r$ are not identified. For instance

$$\xi_i = f_0(t_i) + f_1(v_{1,i})$$

is equivalent to

$$\xi_i^\star = f_0^\star(t_i) + f_1^\star(v_{1,i})$$

with $f_0^\star(t_i) = f_0(t_i) + c$ and $f_1^\star(v_{1,i}) = f_1(v_{1,i}) - c$, so that the mean of f_0 and f_1 is not identifiable. As such, we need to impose constraints or the sampler would not converge. To that end, we use the decomposition from Kneib and Fahrmeir (2007) of the P-spline regression coefficients into a unpenalized and a penalized part:

$$\boldsymbol{\beta}_r = \boldsymbol{U}_r\boldsymbol{\beta}_r^u + \boldsymbol{P}_r\boldsymbol{\beta}_r^p, \text{ for } r = 1\ldots, R,$$

[5]There are some tuning parameters, but the sampler worked fine with default values except the targeted acceptance rate, which we set to 0.99. This increases the effective sample size per iteration for the cost of increased time per iteration (Stan Development Team, 2021).

[6]Stan also includes optimizing routines based on the automatic differentiation, so that the posterior mode (equivalent to penalized maximum likelihood) is also an option for point estimation, for example for frequentist inference.

with priors $p(\boldsymbol{\beta}_r^u) \propto 1$ and $\boldsymbol{\beta}_r^p \sim N(0, \tau_r^2)$. The matrix $\boldsymbol{U}_r$ contains the basis for the null space of $\boldsymbol{K}_r$, so that

$$(\boldsymbol{U}_r \boldsymbol{\beta}_r^u)^\top \boldsymbol{K}_r (\boldsymbol{U}_r \boldsymbol{\beta}_r^u) = 0.$$

As such, $\boldsymbol{\beta}_r^p$ captures the unpenalized polynomial of degree $d-1$ in f_r, see figure 3.2 for an illustration. For the penalized part, it holds that

$$\boldsymbol{P}_r^\top \boldsymbol{K}_r \boldsymbol{P}_r = \boldsymbol{I},$$

where $\boldsymbol{I}$ denotes the identity matrix, resulting in a Gaussian prior for $\boldsymbol{\beta}_r^p$. $\boldsymbol{P}_r$ can be taken as $\boldsymbol{L}_r(\boldsymbol{L}_r^\top \boldsymbol{L}_r)^{-1}$, where $\boldsymbol{L}_r$ comes from the factorized penalty matrix $\boldsymbol{K}_r = \boldsymbol{L}_r \boldsymbol{L}_r^\top$. The first column of $\boldsymbol{P}_r$ is a vector of ones, so that the parameter $\beta_{r,1}^u$ represents the mean of f_r. Then, the estimate for f_r is constrained by deleting the first column of $\boldsymbol{P}_r$, which has a similar effects as imposing a zero mean constraint. We let the log-cumulative baseline hazard f_0 sets the global mean, so that we can sample $\boldsymbol{\beta}_0$ without further restrictions. The vector $\boldsymbol{\beta}_r^p$ is equivalent to a vector of random effects, allowing the use of specialized Stan routines such as the non-centered parameterization.

3.3.1 Model choice

Model choice in a Bayesian framework is an ongoing research area with several competing approaches. We use expected log predictive density criterion (henceforth *elpd*), because it is a measure for the generalizability of a model to unknown data, which is usually the pertinent task. Vehtari and Ojanen (2012) derive the criterion in a Bayesian decision theoretic approach: Here we choose some model M which maximizes an utility function of our choice. Using the log score results in the *elpd*:

$$elpd_M \coloneqq \int \pi(\dot{y}) \log p_M(\dot{y}|\dot{\boldsymbol{x}}, \mathcal{D}) \; d\dot{y},$$

where

$$p_M(\dot{y}|\dot{\boldsymbol{x}}, \mathcal{D}) \coloneqq \int p_M(\dot{y}|\dot{\boldsymbol{x}}, \mathcal{D}, \boldsymbol{\theta}) p_M(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

is the posterior predictive distribution for a new observation $\dot{y}$ under model M, with covariates $\dot{\boldsymbol{x}}$. π denotes the distribution associated with the unknown data generating process. Using leave-one-out cross-validation, an estimator for the *elpd* is given by:

$$\widehat{elpd}_M = N^{-1} \sum_{i=1}^{N} \log p_M(y_i|\boldsymbol{x}_i, \mathcal{D}_{-i}), \tag{3.7}$$

where $\mathcal{D}_{-i}$ is the data without observation i. The *elpd* measures predictive performance of a model, it does not contain an explicit penalty for the number of parameters, however this can be estimated as shown by Vehtari, Gelman, and Gabry (2017). They also show how to compute (3.7) efficiently via Pareto smoothed importance sampling. Their method bypasses the need to compute N models, instead using log-likelihood evaluations from a single MCMC run. Magnusson, Andersen, Jonasson, and Vehtari (2019) present a method to further speed up computation for large data sets based on subsampling.

3.3.2 Covariate effects

Let σ denote the follow-up time and $\mu = \mu(\xi)$ denote the conditional expectation $E[T|\xi]$. If σ and the sample size are large enough so that we can precisely

estimate the survival time where the baseline survivor function tends to zero, we can estimate μ via the identity

$$\mu = \int_0^\infty S(u|\xi)du. \tag{3.8}$$

In practice, this is rarely the case so that estimating the integral in equation (3.8) necessitates extrapolation. The restricted mean survival time (rmst), defined as

$$\mu^\sigma \coloneqq E[\min(T, \sigma)|\xi] = \int_0^\sigma S(u|\xi)du$$

is an alternative which avoids extrapolation beyond the follow up time and bypasses the need to interpret effects in terms of the hazard rate (Stensrud, Aalen, Aalen, & Valberg, 2018). Because researchers can interpret the restricted mean survival time as the average survival time until σ, the rmst has attracted much attention as a measure for covariate effects, see for instance Chen and Tsiatis (2001), Royston and Parmar (2011) and Zhao, Tian, Uno, Solomon, Pfeffer, Schindler, and Wei (2012). We use numerical integration with the trapezoid rule to estimate the integral in (3.8), where we split up the integral at $\min(t_1^-, \dots, t_N^-)$, to avoid extrapolation[7]:

$$\widehat{\mu}^\sigma(\xi) = \frac{t_1^\star}{2}(1 + S(t_1|\xi) + \frac{h}{2}(S(t_1|\xi) + S(\sigma|\xi)) + h\sum_{k=2}^{K-1}(S(t_k^\star|\xi) + S(t_{k+1}^\star|\xi)),$$

with spacing $h = \sigma/(K-1)$ and K control points $t_k^\star = \min(t_1^-, \dots, t_N^-) + (k-1)h$. We can estimate the restricted mean survival time of one observation via

$$\widehat{\mu}^\sigma(\xi_i) = Q^{-1}\sum_{q=1}^{Q}\widehat{\mu}^\sigma(\xi_i^q),$$

where the superscript $q = 1 \dots Q$ denotes the qth draw from the posterior distribution.

The most important application of the restricted survival time is the estimation of a binary treatment effect, the comparison between outcomes μ_1^σ (treatment) and μ_0^σ (control). Most commonly this is the simple difference $\mu_1^\sigma - \mu_0^\sigma$. However, inference for other forms such as ratios can easily be done in a Bayesian framework (Imbens & Rubin, 2015). A unit level treatment effect W_i is the comparison of μ_1^σ and μ_0^σ for unit i. We estimate W_i via $\widehat{W}_i = Q^{-1}\sum_{q=1}^{Q} W_i^q$. An easy-to-interpret scalar measure is the average treatment effect (ATE), which we estimate via

$$\widehat{ATE} = N^{-1}\sum_{i=1}^{N}\widehat{W}_i.$$

We propose to combine partial dependence plots (Friedman, 2001) with the restricted mean survival to simplify effect interpretation for metric covariates: Say we are interested in the effect of the metric covariate v_k. The basic idea of partial dependence plots is to compute the restricted mean survival time, marginalizing over all parameters and covariates except v_k. Let $\dot{\xi}_i$ denote the linear predictor for unit i where we set the value of v_{ik} to $\dot{v}_k$. Then we create a partial dependence plot for the covariate v_k by computing

$$(NQ)^{-1}\sum_{i=1}^{N}\sum_{q=1}^{Q}\hat{\mu}^\sigma(\dot{\xi}_i^q)$$

[7] This might be problematic if the observed minimum is large. In this case extrapolating H_0 or $\log H_0$ might be preferable.

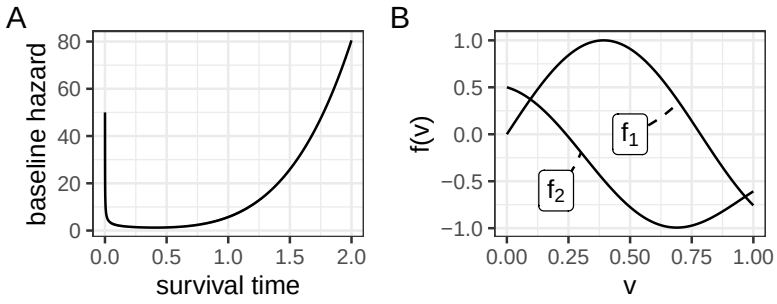


Figure 3.4: Functions involved in the simulation study. A: Additive Weibull baseline hazard. B: Functions f_1 and f_2.

over a grid of control points $\dot{v}_{k,1}, \ldots, \dot{v}_{k,C}$, plotting the result with the associated posterior interval. We can do this for variables which we model by a linear or a nonlinear component. Partial dependence plots may also be used for non-metric variables.

A related concept is the marginal survival function. Often one is interested in a global average of the survival function. For the Cox model, a simple solution is the baseline survival function, which is the survival function conditional on all covariates taking the value zero. However, this might be nonsensical or require an unwanted transformation of the covariates to achieve interpretability. The marginal survival function allows averaging over the covariates and parameters. We compute it via

$$\hat{S}_{average}(t) = (NQ)^{-1} \sum_{i=1}^{N} \sum_{q=1}^{Q} S(t|\xi_i^q).$$

We can furthermore condition on specific covariates values, for instance a subgroup indicator for group differences.

For longer chain lengths, one may use thinning to speed up computations, i.e., the use of every nth sample from the posterior distribution.

3.4 Simulation study

We investigate the performance of the presented methods under varying censoring mechanisms and sample sizes. For each censoring mechanism, we simulate 50 data sets each with sample size $N = 100, 200, 500, 1000, 2000$. Survival times are additive Weibull distributed with hazard where

$$\begin{aligned} h(t_i|\boldsymbol{x}_i, \boldsymbol{\alpha}) &= h_0(t_i) \exp(\boldsymbol{z}_i^\top \boldsymbol{\alpha} + f_1(v_{1i}) + f_2(v_{2i})), \\ h_0(t_i) &= t_i^5 + 2\sqrt{t_i}, \\ f_1(v_{i1}) &= \sin(4v_{i1}) \text{ and } f_2(v_{i2}) = 0.5[\cos(5v_{i2}) - 1.5v_{i2}]. \end{aligned}$$

Figure 3.4 shows the baseline hazard and involved functions. The baseline hazard is bathtub shaped, exemplifying a shape that is hard to capture by common parametric methods yet highly relevant in practice. An example for such a mechanism is human mortality: here the hazard rate is high immediately after birth, followed by a period with low hazard, while rising in later years.

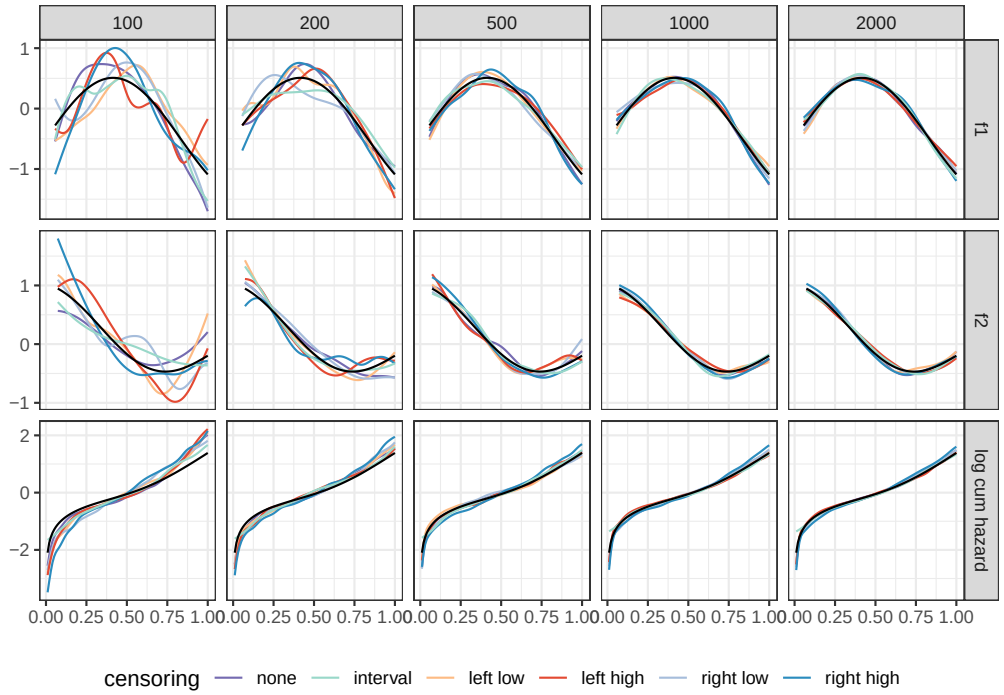


Figure 3.5: Representative examples by sample size (columns), variable (rows). The color of the lines is mapped to the combination of the type of censoring with the fraction of censored survival times. For instance, right low means 20% of the survival times are right censored. The x-axis is scaled to the interval $[0, 1]$ for visualization purposes. Each example is chosen to be closest to the overall mean squared error over all iterations.

Regression coefficients $\boldsymbol{\alpha}$ are equally spaced between -0.3 and 0.3, covariates $z_1, \ldots, z_5$ are standard normal, v_1 and v_2 are standard uniform. There are four variants regarding the fraction of censored observations: no censoring, low (20%) and high (40%) percent censored observations for all three censoring types, and 100% for interval censored. For all censoring types, we draw a simple random sample of the failure times and censor afterwards.

To evaluate estimates, we compute the mse of f_0, f_1 and f_2 on a grid of C control points $e_{r,1}, \ldots, e_{r,C}$ as

$$\text{mse}(\hat{f}_r^q) = C^{-1} \sum_{c=1}^{C} \left(\hat{f}_r^q(e_{r,c}) - f_r(e_{r,c}) \right)^2$$

and the mse for $\boldsymbol{\alpha}$ as

$$\text{mse}(\hat{\boldsymbol{\alpha}}^q) = 5^{-1} (\hat{\boldsymbol{\alpha}}^q - \boldsymbol{\alpha})^\top (\hat{\boldsymbol{\alpha}}^q - \boldsymbol{\alpha}).$$

Figure 3.5 shows representative examples of estimates, figure 3.6 shows the results for the log mean squared error, henceforth log-mse. As might be expected for a complex model, estimates for f_0, f_1 and f_2 are quite imprecise for small sample sizes ($N \leq 200$). The log-mse decreases with increasing sample size. For the estimation of $\boldsymbol{\alpha}$, f_0 and f_1 the censoring mechanism does not seem to cause large differences. However, the log-mse is usually highest under a high fraction of right censored observations and lowest under no censoring and interval censoring. For estimation of the log-cumulative baseline hazard, there is a clear negative effect

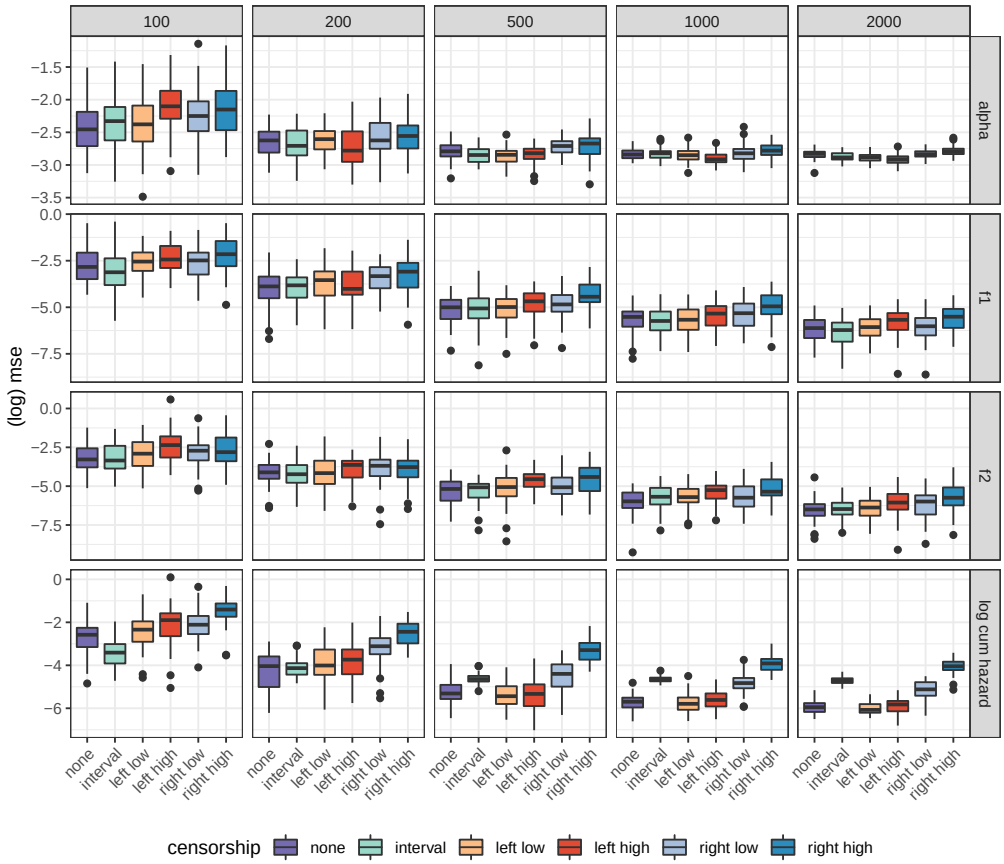


Figure 3.6: Boxplots showing log mean squared error by sample size (columns), variable (rows). The horizontal axis within each box corresponds to the combination of the type of censoring with the fraction of censored survival times, for instance right low means 20% of the survival times are right censored, which is furthermore mapped to the color of each boxplot.

associated with the information loss from censoring. This effect is strongest for right censoring, where more information about the survival times is lost compared to left- and interval-censoring, as the interval around the true survival time is wider. Under interval censoring, the log-mse for estimates of f_0 is lowest for small sample sizes, however it does not improve much for $N > 500$. Overall, the estimation strategy works as desired, given a large enough sample size. However, large fractions of right censored observations may be problematic.

3.5 Application: Real estate data

The data set in this application consists of survival times of real estate advertisements for flat rents. The advertisements were published on the website ImmobilienScout24 in the year 2017. For a description of the data set and data access see Boelmann and Schaffner (2018). The survival time is the number of days an advertisement is online. While there are several reasons why an advertisement may be taken offline, we assume that in the majority of cases someone rented the flat. To illustrate inference for a treatment effect, we estimate a hazard model for the city states Bremen and Hamburg, where we create a balanced data set using coarsened exact matching (Iacus, King, & Porro, 2012). We chose

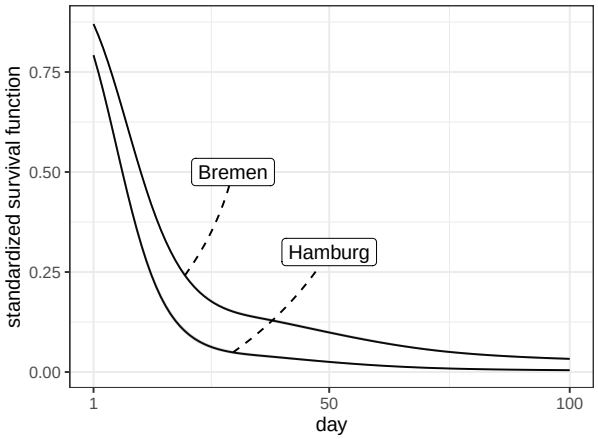


Figure 3.7: Standardized survival curves for Hamburg and Bremen with 0.025 and 0.975 posterior quantile. Note that the posterior interval is very tight due to the sample size and might be only visible when zooming into the figure.

Bremen and Hamburg because these are the only cities which are city states that are located within the same region (Northern Germany). There are 16480 observations in Hamburg, of which 953 are right censored. In Bremen, the are 7356 observations, of which 466 are right censored. We include several covariates

Table 3.1: Descriptive statistics

	Mean		Std. deviation	
Variable	Bremen	Hamburg	Bremen	Hamburg
days online (uncens.)	21.239	16.189	36.595	26.778
days online (cens.)	42.790	23.939	89.599	45.876
censoring indicator	0.063	0.058	0.244	0.233
commission	0.018	0.023	0.132	0.151
missing entries	5.232	4.130	2.432	2.098
rent residual	-0.480	1.443	1.942	2.494

in the model, see Table 3.1 for descriptive statistics: (1) The residual from a hedonic regression of rent price per square metre on a set of control variables such as age of the house. A positive residual indicates an overpriced flat. We use a similar set of variables as Eilers (2017), extending the model for the hedonic regression by a spatial effect. The residual then gives the relative price for a flat conditional covariates and the location. (2) The number of missing fields in the advertisement. (3) Binary indicators for flats requiring a broker commission and a binary indicator for Hamburg, representing a treatment effect in our analysis. We define the individual treatment W_i effect as difference between restricted mean survival times with follow up time $\sigma = 100$ days.

Figure 3.7 shows the marginal survival function for Hamburg and Bremen, indicating that flats in Hamburg are rented out quicker than in Bremen. Figure 3.8 shows the partial dependence plots. An overly high price is associated with an increase in the restricted mean survival time. This effect is approximately linear. An increase in the number of missing entries is associated with a decrease in restricted mean survival time. This might be due to advertisement for unattractive flats, where the supplier hopes to increase attractiveness by providing more

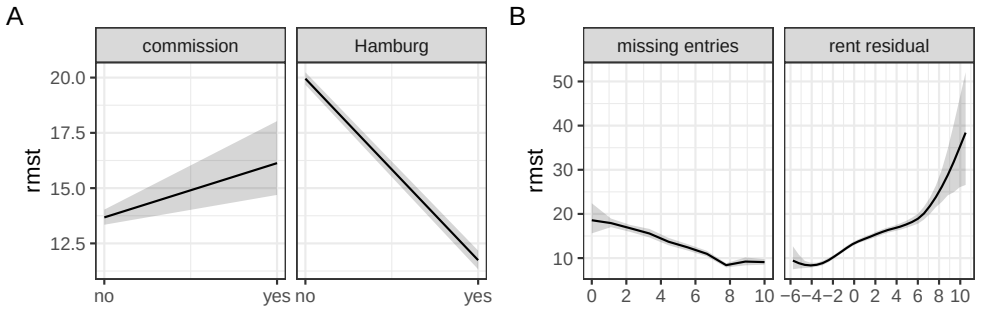


Figure 3.8: Partial dependence plots for restricted mean survival time, for $\sigma = 100$ days, with 0.025 and 0.975 posterior quantiles. A: Partial dependence plots for binary covariates. B: Metric covariates.

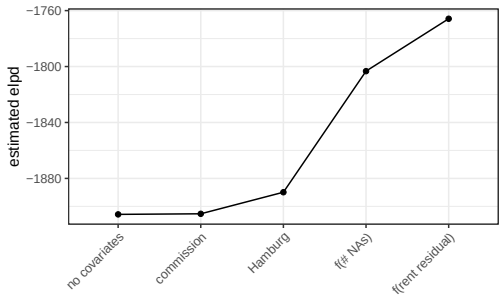


Figure 3.9: Estimated *elpd* for the chosen models. From left to right, the x-axis gives the element which is added to the model, starting with a model without covariates, so that the second model for the cumulative baseline hazard is $H_0(t)\exp(\beta_1 \text{commission})$, the third model $H_0(t)\exp(\beta_1 \text{commission} + \beta_2 \text{Hamburg})$ and so forth.

information.

For Hamburg, the estimated average treatment effect $\widehat{ATE}$ is -8.62, probably due to higher demand. The associated posterior interval is $[-8.21, -7.77]$, so that the effect is precisely estimated. Requiring a broker commission is associated with an increase in restricted mean survival time by 2.45 days. Because the share of advertisements requiring a broker commission in the data set is low (2.1 %), the associated posterior interval $[1.03, 4.19]$ is much wider.

Figure 3.9 shows the estimated expected log predictive density values, working up from a model containing no covariates. While the *elpd* does not contain an explicit penalty term for the number of parameters, including another parameter may decrease the *elpd*. The effect of the covariates on the *elpd* varies strongly between the variables. However, the order of the covariates influences this effect. For instance, including the broker commission hardly increases the *elpd* compared to the inclusion of the number of missing entries.

3.6 Conclusion

This paper presents an approach for fast Bayesian hazard regression using monotonic P-splines. Because involved quantities are analytically available and we can exploit sparsity for the involved computations, this estimation strategy is computationally more efficient than existing approaches. We leverage this effi-

ciency to simplify effect interpretation by combining partial dependence plots with the restricted mean survival time approach. We tested the proposed strategy with numerical examples: Simulations show that the approach works well, an application shows that the approach gives useful results for a large data set.

There are several extensions of this work. It might be fruitful to relax the proportional hazards assumption. This may be done by allowing interactions with survival or by allowing the baseline hazard to vary by subgroup. While using P-splines for $f_1, \ldots, f_R$ is one (very good) choice among many, we argue that monotonic P-splines are useful for $\log H_0$. For instance, one might use a Gaussian process instead, which would be feasible in our framework.

Chapter 4

Fast, approximate MCMC for Bayesian analysis of large data sets: A design based approach[1]

Abstract

We propose a fast approximate Metropolis-Hastings algorithm for large data sets embedded in a design based approach. Here, the loglikelihood ratios involved in the Metropolis-Hastings acceptance step are considered as data. The building block is one single subsample from the complete data set, so that the necessity to store the complete data set is bypassed. The subsample is taken via the cube method, a balanced sampling design, which is defined by the property that the sample mean of some auxiliary variables is close to the sample mean of the complete data set. We develop several computationally and statistically efficient estimators for the Metropolis-Hastings acceptance probability. Our simulation studies show that the approach works well and can lead to results which are close to the use of the complete data set, while being much faster. The methods are applied on a large data set consisting of all German diesel prices for the first quarter of 2015.

Keywords: Bayesian inference; Big Data; Approximate MCMC; Survey sampling

[1]This chapter has been published as: Kaeding, M. (2016). Fast, approximate MCMC for Bayesian analysis of large data sets: A design based approach. *Ruhr Economic Papers, 660*, 1–23. doi:10.4419/86788766

4.1 Introduction

Consider the update step in the Metropolis-Hastings algorithm for the simulation of the posterior distribution

$$\pi(\boldsymbol{\theta}|\mathcal{D}) \propto \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})\pi(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is the parameter vector, $\pi(\boldsymbol{\theta})$ is the prior, $\mathcal{D}$ is the available data and $\mathcal{L}$ is the likelihood

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \prod_{k=1}^{N} \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}_k).$$

We restrict attention to the case where the observations are independent given $\boldsymbol{\theta}$, most common under a regression framework. The update step consists of the following steps: Given the current value $\boldsymbol{\theta}^c$, a proposal $\boldsymbol{\theta}^\star$ is drawn from the proposal distribution $G(\cdot|\boldsymbol{\theta}^c)$. The current value is set to the proposal with probability

$$\alpha(\boldsymbol{\theta}^c, \boldsymbol{\theta}^\star) = 1 \wedge \frac{\mathcal{L}(\boldsymbol{\theta}^\star|\mathcal{D})\pi(\boldsymbol{\theta}^\star)G(\boldsymbol{\theta}^c|\boldsymbol{\theta}^\star)}{\mathcal{L}(\boldsymbol{\theta}^c|\mathcal{D})\pi(\boldsymbol{\theta}^c)G(\boldsymbol{\theta}^\star|\boldsymbol{\theta}^c)}. \tag{4.1}$$

The computational cost for the computation of $\alpha = \alpha(\boldsymbol{\theta}^c, \boldsymbol{\theta}^\star)$ is linear in N. As a result, large data sets pose a problem for the algorithm. Due to the onset of large data sets, accelerating the algorithm is a highly relevant issue (Green, Łatuszyński, Pereyra, & Robert, 2015). The aim of this paper is to accelerate the update step by using a computationally efficient and precise design based estimator for α using a single *fixed* subsample of the data. Compared to repeatedly subsampling the data, this approach avoids two types of overhead: (1) Subsampling the data, (2) storing and accessing the data. This is especially relevant for data sets which are too large to be read in memory.

Several proposals have been made to accelerate the algorithm by drawing a subsample for every update step: Korattikara, Chen, and Welling (2014) and Bardenet, Doucet, and Holmes (2014) represent the update step as a hypothesis test. Maclaurin and Adams (2014) use a completion of the posterior distribution via auxiliary variables. For their approach, a computationally cheap lower bound must be available for all likelihood contributions. Quiroz, Kohn, Villani, and Tran (2019) use a pseudo-marginal argument where an unbiased estimator for the likelihood is needed: They construct this estimator by debiasing an exponentiated estimator for the loglikelihood, relying on a normality assumption. To the best of our knowledge, Maire, Friel, and Alquier (2015) is the only article considering the use of subsamples which stay fixed during several iterations. Bardenet, Doucet, and Holmes (2017) provide a good overview over existing approaches.

We take the same perspective as these authors: Our algorithm belongs to the class of *noisy* Monte Carlo algorithms: Here, we obtain deviates from a (close) approximation to the posterior distribution using the complete data set.

The paper is structured as follows: Section 2 gives background on noisy MCMC, Section 3 on sampling theory. Section 4 describes the proposed algorithm, Section 5 reports a simulation study. Section 6 applies the methods to a large data set on gas prices. Section 7 concludes with a discussion.

4.2 Noisy Metropolis-Hastings

It is convenient to use the following representation of the update step of the Metropolis-Hastings algorithm, used by Korattikara, Chen, and Welling (2014)

Algorithm 1 Metropolis Hastings update step for posterior simulation

Draw u from $U(0,1)$
Set ρ to $\log[u\frac{\pi(\boldsymbol{\theta}^c)G(\boldsymbol{\theta}^\star|\boldsymbol{\theta}^c)}{\pi(\boldsymbol{\theta}^\star)G(\boldsymbol{\theta}^c|\boldsymbol{\theta}^\star)}]$
if $\rho > \dot{\phi}$ **then**
  Set $\boldsymbol{\theta}^c$ to $\boldsymbol{\theta}^\star$
end if
return $\boldsymbol{\theta}^c$

Algorithm 2 Metropolis Hastings algorithm for posterior simulation, generic version

for $r = 1$ **to** $R-1$ **do**
  Draw $\boldsymbol{\theta}^\star$ from $G(\cdot|\boldsymbol{\theta}^c)$
  Set $\dot{\phi}$ to $\sum_{i=1}^N L_i(\boldsymbol{\theta}^\star|\mathcal{D}_i) - L_i(\boldsymbol{\theta}^c|\mathcal{D}_i)$
  Do Metropolis-Hastings update step for posterior simulation, using $\boldsymbol{\theta}^\star, \dot{\phi}$
  Set $\boldsymbol{\theta}^{(r)}$ to $\boldsymbol{\theta}^c$
end for
return $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(R-1)}$

and Bardenet, Doucet, and Holmes (2014): Draw $\boldsymbol{\theta}^\star \sim G(\cdot|\boldsymbol{\theta}^c)$, $u \sim Unif(0,1)$, and accept the proposal if

$$\log\left[u\frac{\pi(\boldsymbol{\theta}^c)G(\boldsymbol{\theta}^\star|\boldsymbol{\theta}^c)}{\pi(\boldsymbol{\theta}^\star)G(\boldsymbol{\theta}^c|\boldsymbol{\theta}^\star)}\right] < \dot{\phi}, \tag{4.2}$$

where $\dot{\phi} = \sum_{k=1}^N \phi_k$, is the sum of loglikelihood ratios:

$$\phi_k = \phi(\boldsymbol{\theta}^\star, \boldsymbol{\theta}^c|\mathcal{D}_k) = \log L(\boldsymbol{\theta}^\star|\mathcal{D}_k) - \log L(\boldsymbol{\theta}^c|\mathcal{D}_k).$$

The MH algorithm simulates a Markov chain with transition kernel K which is invariant under $\pi(\boldsymbol{\theta}|\mathcal{D})$. Replacing the left hand side or the right hand side in (4.2) by an estimator implies the use of an approximate transition kernel $\hat{K}$ which in general is not invariant under $\pi(\boldsymbol{\theta}|\mathcal{D})$. One exception is given by the pseudo-marginal approach by Andrieu and Roberts (2009), where an unbiased estimator of the likelihood is available, used by Quiroz, Kohn, Villani, and Tran (2019). However, getting an unbiased estimator of the likelihood in a fixed subsample context is not feasible without questionable assumptions and might lead to trading small bias with high variance.

Although the theoretical properties of noisy MCMC are not completely understood, there are some encouraging results: Alquier, Friel, Everitt, and Boland (2014) have shown that the stationary distribution of the Markov chain obtained using an estimator $\hat{\alpha} = \hat{\alpha}(\boldsymbol{\theta}^c, \boldsymbol{\theta}^\star)$ for α is a useful approximation to $\pi(\boldsymbol{\theta}|\mathcal{D})$, provided that $|\hat{\alpha} - \alpha|$ is bounded. Nicholls, Fox, and Watt (2012) show that, if a Gaussian unbiased estimator for $\log \pi(\boldsymbol{\theta}^\star|\mathcal{D}) - \log \pi(\boldsymbol{\theta}^c|\mathcal{D})$ with known variance is available, the update step of the Metropolis-Hastings algorithm can be adjusted via a method called *the penalty method*, so that the chain targets $\pi(\boldsymbol{\theta}|\mathcal{D})$ as desired. Usually, such an estimator is not available. Nicholls, Fox, and Watt (2012) show that plugging an estimate of $\log \pi(\boldsymbol{\theta}^\star|\mathcal{D}) - \log \pi(\boldsymbol{\theta}^c|\mathcal{D})$ into α leads to a chain which is very close to the penalty method, provided that the expected absolute error $E|\hat{\alpha} - \alpha|$ is small. As such, the main objective here is to find a computationally cheap estimator for $\dot{\phi}$, so that $|\hat{\alpha} - \alpha|$ is small.

4.3 Some sampling theory

Let $\mathcal{U} = \{1, \ldots, N\}$ be a set of labels associated with a finite population of N units, here given by the complete data set. The aim of survey-sampling is to estimate a finite population total, $\dot{y} = \sum_{k \in \mathcal{U}} y_k$ of a variable y, using a sample $\mathcal{S} \subseteq \mathcal{U}$. Totals are denoted by dots. The set $\mathcal{S}$ is selected via a stochastic selection scheme, called the *sampling design*. Design based sampling is used when the cost of obtaining the value y_k for all units $k \in \mathcal{U}$ is too expensive, so the sample size $n := |\mathcal{S}|$ is usually much smaller than N. For this article, cost is given by computation time. In a design based approach, all randomization is due to the sampling design, while the values $y_1, \ldots, y_N$ of the study variable are unknown constants.

A sample can be represented by a random vector

$$\boldsymbol{i} = (i_1, \ldots, i_k, \ldots, i_N)^\top,$$

where i_k takes the value 1 if unit k is in the sample $\mathcal{S}$ and 0 otherwise. The sampling design $f(\cdot)$ is a probability distribution on a support $\mathcal{Q}$. Here, attention is restricted to without-replacement sampling designs with fixed sample size, so that

$$\mathcal{Q} = \Big\{ \boldsymbol{i} \in \{0,1\}^N | \sum_{k \in \mathcal{U}} i_k = n \Big\}.$$

It holds that

$$E_f[\boldsymbol{i}] = \sum_{\boldsymbol{i} \in \mathcal{Q}} \boldsymbol{i} f(\boldsymbol{i}) =: \boldsymbol{\eta},$$

where the *inclusion probability* $P(k \in \mathcal{S}) = E_f[i_k] = P(i_k = 1)$ of an element k is given by η_k. Note that the empty sample $\boldsymbol{i} = (0, \ldots, 0)^\top$ and the census $\boldsymbol{i} = (1, \ldots, 1)^\top$ are in $\mathcal{Q}$, corresponding to the cases $n = 0$ and $n = N$. For more on sampling theory see Tillé (2006) or Särndal, Swensson, and Wretman (2003).

4.3.1 Sampling design: Cube sampling

Denoting sums of the form $\sum_{k \in \mathcal{S}} k$ as $\sum_{\mathcal{S}} k$, the benchmark estimator for a total $\dot{y}$ is the Horvitz-Thompson estimator:

$$\hat{\dot{y}}^{HT} = \sum_{\mathcal{U}} \frac{i_k y_k}{\eta_k} = \sum_{\mathcal{S}} y_k d_k,$$

where $d_k = 1/\eta_k$ is called the design weight. Here, only sampling designs are used where all inclusion probabilities are equal, so that without loss of generality $\eta_k = n/N$ for $k \in \mathcal{U}$ and

$$\hat{\dot{y}}^{HT} = \frac{N}{n} \sum_{\mathcal{S}} y_k = N \bar{y},$$

where $\bar{y}$ is the sample mean of y.

The variance of the Horwitz-Thompson estimator can be lowered by exploiting additional information: Assume $\boldsymbol{z}_{q,k}$, $q \in \{cube, greg\}$ is known for all elements in the population, where $\boldsymbol{z}_{q,k} = (z_{q,k,1}, \ldots, z_{q,k,p_q})^\top$ is a vector of values of p_q auxiliary variables for unit k. The auxiliary variables $z_{cube,1}, \ldots, z_{cube,p_{cube}}$ and $z_{greg,1}, \ldots, z_{greg,p_{greg}}$ may intersect and may be identical or distinct. Let $\dot{\boldsymbol{z}}_{cube} = \sum_{\mathcal{U}} \boldsymbol{z}_{cube,k}$ denote the known total of the auxiliary variables used for cube

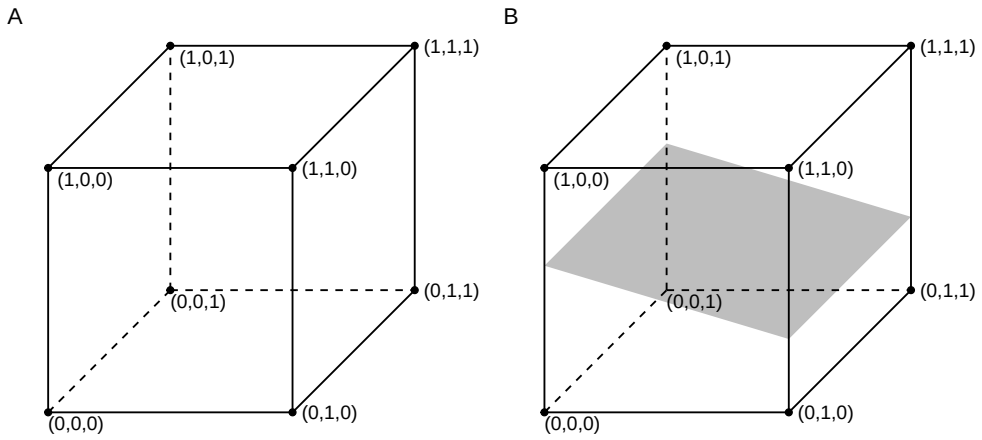


Figure 4.1: A: Set of all samples as represented as cube, for the case $N = 3$, giving $2^3 = 8$ possible samples. B: Set of all samples with constraint space. In this case, there are no samples in the constraint space.

sampling. The objective of balanced sampling is to obtain a sample, respecting the vector of inclusion probabilities $\boldsymbol{\eta}$, so that the constraint

$$\hat{\dot{\boldsymbol{z}}}_{cube}^{HT} = \dot{\boldsymbol{z}}_{cube} \tag{4.3}$$

holds. For equal inclusion probability designs as used here, this is equivalent to the constraint

$$n^{-1} \sum_{\mathcal{S}} \boldsymbol{z}_{cube,k} = N^{-1} \sum_{\mathcal{U}} \boldsymbol{z}_{cube,k},$$

so that the sample mean of the auxiliary variables used for cube sampling is equal to their population mean. Usually, an exact solution is not possible for all auxiliary variables, so that samples are found which satisfy (4.3) approximately. Here, the role of the balanced sampling algorithm is to find a sample which is "close" to the complete data set. The choice of auxiliary variables is discussed in section 4.4. We will use the cube method Deville and Tillê (2004), which is based on a geometrical representation of a sampling design: Let $\mathcal{C} = [0, 1]^N$ denote a cube equipped with 2^N vertices. $\mathcal{C}$ represents the set of all 2^N possible samples from a population of size N, where every vertex of $\mathcal{C}$ is associated with a sample, see figure 4.1.

Define the vector $\boldsymbol{a}_k \coloneqq \boldsymbol{z}_{cube,k}/\eta_k$ and the $p_{cube} \times N$ matrix

$$\boldsymbol{A} \coloneqq (\boldsymbol{a}_1, .., \boldsymbol{a}_k, \ldots, \boldsymbol{a}_N),$$

then the balancing equations (4.3) can be written as

$$\boldsymbol{A}\boldsymbol{i} = \boldsymbol{A}\boldsymbol{\eta}. \tag{4.4}$$

The system of equations (4.4) defines the hyperplane

$$\mathcal{P} = \boldsymbol{\eta} + K(\boldsymbol{A})$$

in $\mathcal{R}^N$, where $K(\boldsymbol{A})$ is the kernel of $\boldsymbol{A}$. The basic idea of cube sampling is to choose a sample as a vertex of $\mathcal{C}$ in $\mathcal{P}$ or, if that is not possible, near $\mathcal{P}$. Cube sampling consists of two phases: The *flight phase* is a martingale with initial

value $\boldsymbol{\eta}$ in the constraint space $\mathcal{K} = \mathcal{C} \cap \mathcal{P}$, where the constraint (4.3) is satisfied. At the end of the flight phase a vector $\boldsymbol{i}^\star$ is obtained. If all elements of $\boldsymbol{i}^\star$ are either 1 or 0, the *landing phase* is not necessary as a vertex of $\mathcal{C}$ is reached. If this is not the case, the balancing equations are relaxed and the elements of $\boldsymbol{i}^\star$ are randomly rounded so that

$$E[\boldsymbol{i}] = \boldsymbol{\eta},$$

hence, given inclusion probabilities are respected. Deville and Tillê (2004) show that it is always possible to satisfy one balancing equation. It holds that $\sum_{\mathcal{U}} \eta_k = n$. As such, setting $\boldsymbol{z}_{cube,1} = \boldsymbol{\eta}$ guarantees a fixed sample size, as the balancing equations are relaxed starting with the last auxiliary variable $z_{cube,p_{cube}}$. The authors show that the cube method achieves the bound

$$\frac{|\dot{z}_{cube,j} - \hat{\dot{z}}^{HT}_{cube,j}|}{N} = O(p_{cube}/n), \text{ for } j = 1, \ldots, p_{cube}, \tag{4.5}$$

where $f = O(g)$ if there is an upper bound of $|f|$ which is proportional to g. Due to the bound (4.5), the error becomes small if the sample size is large, compared to the number of auxiliary variables. We will use the implementation by Chauvet and Tillê (2006) with computational cost linear in N.

4.3.2 Regression estimator

The Horvitz-Thompson estimator is the only estimator which is unbiased for all sampling designs. However, there are better estimators in an MSE-sense if this condition is relaxed. A large class is given by the generalized regression estimator (greg). Model the study variable y via

$$y_k = \boldsymbol{\beta}^\top \boldsymbol{z}_{greg,k} + \epsilon_k, \ \epsilon_k \sim N(0, \omega^2).$$

The generalized regression estimator $\hat{\dot{y}}^{greg}$ for a total $\dot{y}$ is given by the sum of fitted values plus the Horvitz-Thompson estimator for the prediction error:

$$\begin{aligned} \hat{\dot{y}}^{greg} =& \sum_{\mathcal{U}} \hat{y}_k + \hat{\dot{e}}^{HT} && (4.6)\\ =& \hat{\boldsymbol{\beta}}^\top \dot{\boldsymbol{z}}_{greg} + \frac{N}{n} \sum_{\mathcal{S}} (y_k - \hat{y}_k), && (4.7) \end{aligned}$$

where $\hat{y}_k = \boldsymbol{z}_k^\top \hat{\boldsymbol{\beta}}$ is the fitted value of element k, $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{\mathcal{S}} \boldsymbol{z}_{greg,k} \boldsymbol{z}_{greg,k}^\top \right)^{-1} \sum_{\mathcal{S}} y_k \boldsymbol{z}_{greg,k},$$

and e_k is the residual $e_k = y_k - \hat{y}_k$. The generalized regression estimator can also be written as weighted sum $\sum_{\mathcal{S}} w_k y_k$, where the weights

$$\begin{aligned} w_k = N/n \Big[1 + (N/n)(\dot{\boldsymbol{z}}_{greg} - \hat{\dot{\boldsymbol{z}}}^{HT}_{greg})^\top \times \\ \left(\sum_{\mathcal{S}} \boldsymbol{z}_{greg,k} \boldsymbol{z}_{greg,k}^\top \right) \boldsymbol{z}_{greg,k} \Big], \text{ for } k \in \mathcal{S}, \end{aligned} \tag{4.8}$$

depend on the sample but not on the study variable. Hence, the same weight vector can be used for several variables, so that the estimator is cheap to compute. The greg weights (4.8) satisfy the calibration equation

Algorithm 3 Approximate Metropolis-Hastings - phase 1

Select subsect $\mathcal{S} \subset \mathcal{U}$ via cube method, using auxiliary variables $z_{cube,1}, ..., z_{cube,p_{cube}}$
Set j to 1
for $r = 1$ to $p_{\text{greg}} \times 100$ **do**
  Draw $\boldsymbol{\theta}^\star$ from $G(\cdot|\boldsymbol{\theta}^c)$
  Set $\dot{\phi}$ to $\sum_{\mathcal{S}} \phi_k$
  Do Metropolis-Hastings update step for posterior simulation, using $\boldsymbol{\theta}^\star, \dot{\phi}$
  Set $\boldsymbol{\theta}^{(r)}$ to $\boldsymbol{\theta}^c$
  if r modulo $100 = 0$ **then**
    for $k \in \mathcal{U}$ **do**
      Set $z_{greg,k,j}$ to ϕ_k
    end for
    Set j to $j + 1$
  end if
end for

Algorithm 4 Approximate Metropolis-Hastings - greg, basic version

Do approximate Metropolis-Hastings - phase 1
for $k \in \mathcal{S}$ **do**
  Compute w_k
end for
for $r = 1$ to R **do**
  Draw $\boldsymbol{\theta}^\star$ from $G(\cdot|\boldsymbol{\theta}^c)$
  for $k \in \mathcal{S}$ **do**
    Compute ϕ_k
  end for
  Set $\hat{\dot{\phi}}$ to $\sum_{\mathcal{S}} w_k \phi_k$
  Do Metropolis-Hastings update step for posterior simulation, using $\boldsymbol{\theta}^\star, \hat{\dot{\phi}}$
  Set $\boldsymbol{\theta}^r$ to $\boldsymbol{\theta}^c$
end for
return $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(R-1)}$

$$\sum_{\mathcal{S}} w_k \boldsymbol{z}_{greg,k} = \dot{\boldsymbol{z}}_{greg}.$$

For equal inclusion probabilities, the approximate variance of the generalized regression estimator is, up to a constant, given by the sum $\sum_{\mathcal{U}}(y_i - \boldsymbol{z}_{greg,k}^\top \boldsymbol{b})^2$, where

$$\boldsymbol{b} = \left(\sum_{\mathcal{U}} \boldsymbol{z}_{greg,k} \boldsymbol{z}_{greg,k}^\top \right)^{-1} \sum_{\mathcal{U}} y_k \boldsymbol{z}_{greg,k}.$$

As such, predictive power for y is the main requirement for the auxiliary variables.

4.4 Description of algorithm

The objective is to estimate the total $\dot{\phi} = \sum_{\mathcal{U}} \phi_k$. The basic idea is to combine a sample obtained by cube-sampling with a regression estimator. Denote the posterior distribution associated with a subsample $\mathcal{S}$ by $\pi_{\mathcal{S}}$. The algorithm consists of two phases. In the first phase, a sample is chosen via cube sampling,

Algorithm 5 Approximate Metropolis-Hastings - greg, ridge variant

Do approximate Metropolis-Hastings - phase 1
for $r = 1$ to R **do**
  Draw $\boldsymbol{\theta}^\star$ from $G(\cdot|\boldsymbol{\theta}^c)$
  for $k \in \mathcal{S}$ **do**
    Compute ϕ_k
  end for
  Set $\boldsymbol{c}$ to $\boldsymbol{M}^\top \boldsymbol{\phi}$
  for $i = 1$ **to** p_{greg} **do**
    $\hat{\alpha}_i \leftarrow c_i/\lambda_i$
  end for
  Set $\hat{\sigma}^2$ to $(n - p_{greg})^{-1}(\boldsymbol{\phi} - \boldsymbol{M}\hat{\boldsymbol{\alpha}})^\top(\boldsymbol{\phi} - \boldsymbol{M}\hat{\boldsymbol{\alpha}})$
  Set $\hat{\kappa}$ to $\frac{p_{greg}\hat{\sigma}^2}{\hat{\boldsymbol{\xi}}^\top\hat{\boldsymbol{\xi}}}$
  for $i = 1$ **to** p_{greg} **do**
    Set $\hat{\xi}_i(\hat{\kappa})$ to $c_i/(\lambda_i + \hat{\kappa})$
  end for
  Set $\hat{\dot{\phi}}$ to $\boldsymbol{\xi}(\hat{\kappa})^\top \tilde{\tilde{\boldsymbol{z}}}_{greg} + (N/n)\left((\sum_{\mathcal{S}} \phi_k) - \boldsymbol{\xi}(\hat{\kappa})^\top \tilde{\tilde{\boldsymbol{z}}}_{greg,\mathcal{S}}\right)$
  Do Metropolis-Hastings update step for posterior simulation, using $\boldsymbol{\theta}^\star$, $\hat{\dot{\phi}}$
end for
return $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(R-1)}$

so that the posterior distribution from this subsample is close to π. To find a measure for closeness, the result of Maire, Friel, and Alquier (2015) is used. These authors have shown that the minimal Kullback-Leibler distance from $\pi_{\mathcal{S}}$ to $\pi_{\mathcal{U}}$ is minimized if the sufficient statistics for $\boldsymbol{\theta}$ in $\mathcal{S}$ are equal to the sufficient statistics in the population. For many nontrivial cases, there exists no such sufficient statistic which can also be written as sum (as would be necessary for the use of cube sampling). However, the result serves as a general guide, in that statistics which summarize the complete data $\mathcal{D}$ set are used. Such statistics can be derived from inspection of the posterior distribution. Note that the computation time of cube sampling algorithm scales badly with the number of auxiliary variables, so that a low number of auxiliary variables for the cube sampling algorithm is preferable. Also during the first phase, auxiliary variables for the regression estimator are computed. A good predictor for the value of $\phi(\boldsymbol{\theta}^\star, \boldsymbol{\theta}^c|\mathcal{D}_k)$ is given by $\phi(\cdot, \cdot|\mathcal{D}_k)$, for arguments near $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}^c$. This is used to obtain the auxiliary variables for the regression estimator: If the subsample is obtained, the posterior distribution $\pi_{\mathcal{S}}$ is an overdispersed approximation to π so that the support of $\boldsymbol{\theta}$ is covered by $\pi_{\mathcal{S}}$. The auxiliary variables for the regression estimator are subsequently computed as followed: A regular Metropolis-Hastings algorithm is run using the subsample for $100p_{cube}$ iterations. Every 100th iteration, the loglikelihood ratio obtained from the current and proposed value of $\boldsymbol{\theta}$ is used as auxiliary variables, so that the auxiliary variables are computed as

$$z_{greg,k,j} = \phi(\boldsymbol{\theta}^\star, \boldsymbol{\theta}^{(r)}|\mathcal{D}_k),$$

for $k \in \mathcal{S}, r = 100, 200, \ldots, 100p_{greg}$, where $j = r/100$ varies with r.

The complete data set only has to be available for the first phase. For the second phase, the Metropolis-Hastings algorithm proceeds using the subsample with $\dot{\phi}$ replaced by an estimator using the derived auxiliary variables $z_{greg,1}, ..., z_{greg,p_{greg}}$. We use the posterior mode from the trial run of the first phase as starting values.

4.4.1 Ridge variant

The greg estimator can be improved by optimizing the predictions for the values of ϕ_k. We use the ridge estimator to achieve better prediction performance. This estimator has the following advantages: (1) It is available in closed form[2], as

$$\hat{\boldsymbol{\beta}}(\kappa) = \left(\sum_{\mathcal{S}} \boldsymbol{z}_{greg,k} \boldsymbol{z}_{greg,k}^\top + \boldsymbol{I}\kappa \right)^{-1} \sum_{\mathcal{S}} y_k \boldsymbol{z}_{greg,k}.$$

(2) Shrinkage is controlled by a single parameter κ. (3) The complete data posterior distribution is better covered with a higher number of auxiliary variables. This results in multicollinearity, as the auxiliary variables $z_{greg,1}, ..., z_{greg,p_{greg}}$ are strongly correlated, so that the matrix $\sum_{\mathcal{S}} \boldsymbol{z}_{greg,k} \boldsymbol{z}_{greg,k}^\top$ is ill conditioned for large p_{greg}. However, this is a desired property, as it indicates strong correlation with future values of $\phi(\cdot, \cdot|\mathcal{D}_k), k \in \mathcal{U}$. Hence, multicollinearity is a consequence of a set of useful auxiliary variables, which in turn is solved by the ridge estimator.

The third point suggests a simple heuristic to set p_{greg}: It is set at least so large that the matrix $\sum_{\mathcal{S}} \boldsymbol{z}_{greg,k} \boldsymbol{z}_{greg,k}^\top$ is ill conditioned. Following Hoerl, Kannard, and Baldwin (1975), the shrinkage parameter κ is set as

$$\kappa \coloneqq \frac{k\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}}},$$

where $\hat{\sigma}^2$ is an estimate for $var(\phi|z_{greg,1}, ..., z_{greg,p_{greg}})$. While determining κ via cross validation might be preferable, using a closed form expression is much faster. In addition, cross-validation is used to estimate the out-of sample prediction error. However, the procedure used here is based on the assumption that the subsample is similar to the full data set. Given κ, the ridge variant of the greg estimator can be quickly computed using precomputed entities. For details see appendix 4.A.

4.5 Simulation study

4.5.1 Setup

Here, results of a simulation study will be reported. The simulation was carried out in R and C++ via the Rcpp interface (Eddelbuettel & François, 2011). We generated $M = 100$ data sets from the model

$$y_k = \boldsymbol{x}_k^\top \boldsymbol{\beta} + \epsilon_k, \epsilon_k \sim \mathcal{N}(0, \sigma^2), k = 1, \ldots, N = 100000,$$

where $\boldsymbol{x}_1 = (1, \ldots, 1)^\top$ and the elements of $\boldsymbol{x}_2, ..., \boldsymbol{x}_5$ are draws from a uniform distribution with minimum and maximum given by -2 and 2. The elements of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \log \sigma)^\top$ are deviates from a standard normal distribution. The posterior distribution is simulated using the algorithm described in section 4.4. We use a random walk Metropolis-Hastings sampler, where the variance of the proposal distribution is scaled during burnin so that the acceptance probability is around 0.234. We use $R = 100000$ iterations, taking the first half of the chain as burnin. For reference, we also simulate the posterior distribution using the complete data set and use a variant where no auxiliary variables are used, so that $\dot{\phi}$ is estimated via the sample mean. To study the gain of the cube

[2]Unlike e.g., the LASSO or the elastic net estimator, which might be preferable from a prediction perspective.

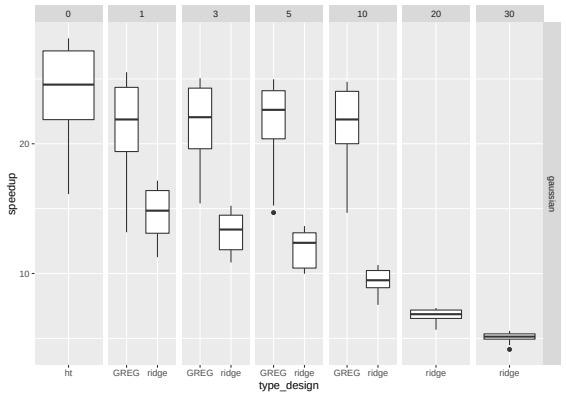


Figure 4.2: Speed up relative to the use of the complete data set for the gaussian likelihood. Columns: Varying values of of p_{greg}.

sampling procedure, the sample was additionally chosen using simple random sampling (srs); which is equivalent to cube sampling with a single auxiliary variable given by the inclusion probabilities. We use $p_{greg} = 1, 3, 5, 10, 20, 30$ auxiliary variables for the regression estimator. For the basic greg estimator, it was not possible to set $p_{greg} \geq 20$ due to multicollinearity. In addition, we generated data sets from the model $y_k \sim Poisson(\exp(\boldsymbol{x}_k^\top \boldsymbol{\beta}))$, with the same configurations as above otherwise. For both models, the auxiliary variables used for cube sampling are derived from the likelihood: For the gaussian likelihood, we take the residuals and the squared residuals obtained from the maximum likelihood estimator $\boldsymbol{\beta}_{ml} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ as $z_{cube,1}$ and $z_{cube,2}$. In addition, we generate three draws from $N(\boldsymbol{\beta}_{ml}, 3\sigma_{ml}^2 \boldsymbol{X}^\top \boldsymbol{X})^{-1})$, where $\sigma_{ml}^2 = N^{-1} \sum_{\mathcal{U}} (y_k - \boldsymbol{x}_k^\top \boldsymbol{\beta}_{ml})^2$ and use the associated residuals (and the squared residuals) as auxiliary variables. For the poisson likelihood, we use the loglikelihood kernel $y_k \boldsymbol{x}_k^\top \boldsymbol{\beta} - \exp(\boldsymbol{x}_k^\top \boldsymbol{\beta})$ as auxiliary variables. The maximum likelihood estimator is obtained via numerical methods, otherwise values of $\boldsymbol{\beta}$ were obtained as above.

4.5.2 Results

The main criterion is the root mean squared error (rmse) for the true parameter vector $\boldsymbol{\theta}$:

$$rmse(\hat{\boldsymbol{\theta}}) = \sqrt{\frac{1}{6} \sum_{i=1}^{6} (\hat{\theta}_i - \theta_i)^2},$$

where $\hat{\theta}_i$ is estimated using the posterior mean. We report the relative root mean squared error, in reference to the use of the complete data set: $rmse(\hat{\boldsymbol{\theta}})/rmse(\hat{\boldsymbol{\theta}}^{\mathcal{U}})$, where $rmse(\hat{\boldsymbol{\theta}}^{\mathcal{U}})$ is the rmse obtained by the use of the complete data set.
Speedup: The speedup, defined as computation time using the given method, divided by the computation time using the complete data set, is shown in figure 4.2. The basic greg variant is by far fastest, while the speedup for the ridge variant drops for larger number of auxiliary variables.
Parameter vector: Figure 4.3 shows the relative root mean squared error for the true parameter vector $\boldsymbol{\theta}$. It can be seen that the error mainly depends on p_{greg}: If there are enough auxiliary variables for the estimation (in this case, around 20), the relative rmse can approach 1. This holds for both likelihoods. Cube sampling reduces the error, although the effect is larger for the gaussian likelihood. We interpret this as a sign that the auxiliary variables are better chosen for the gaussian likelihood. For a large number of auxiliary variables used

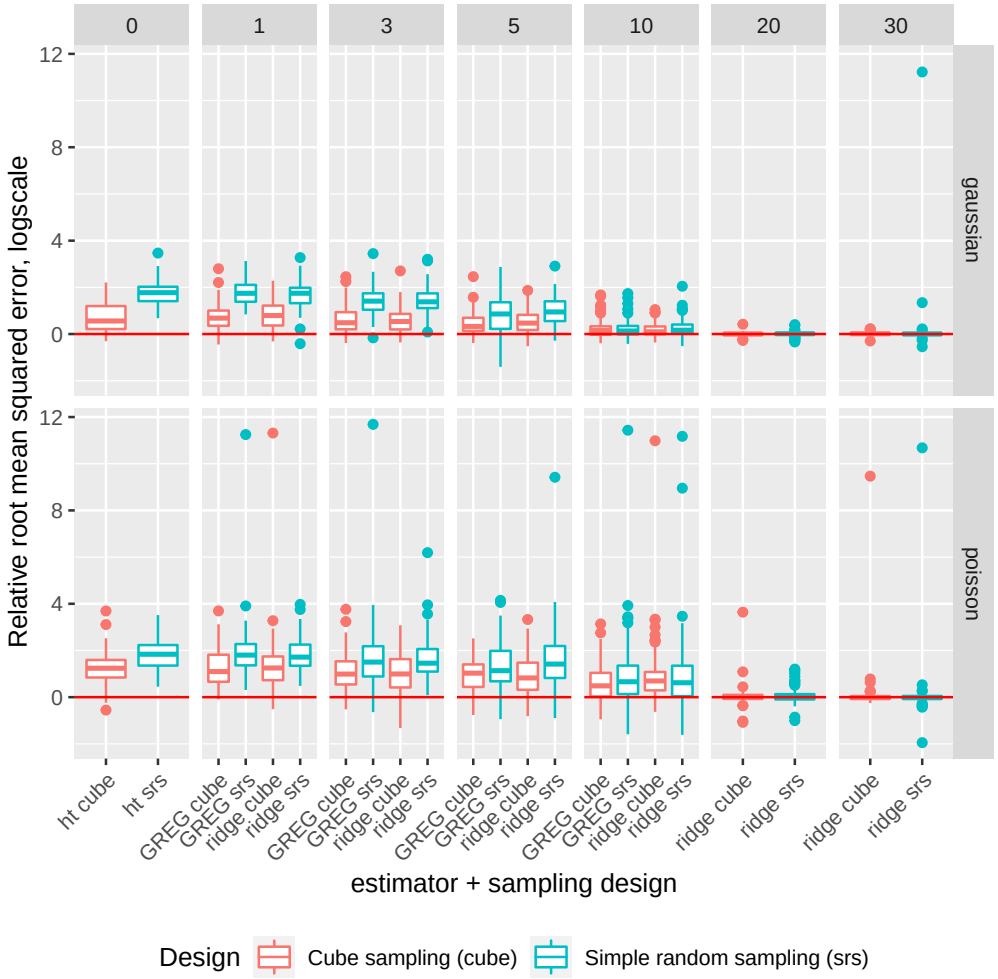


Figure 4.3: Root mean squared error for $\boldsymbol{\theta}$, relative to root mean squared error obtained with complete data set, plotted on log scale for visualization. Rows: Gaussian and poisson likelihood. Columns: Varying values of of p_{greg}.

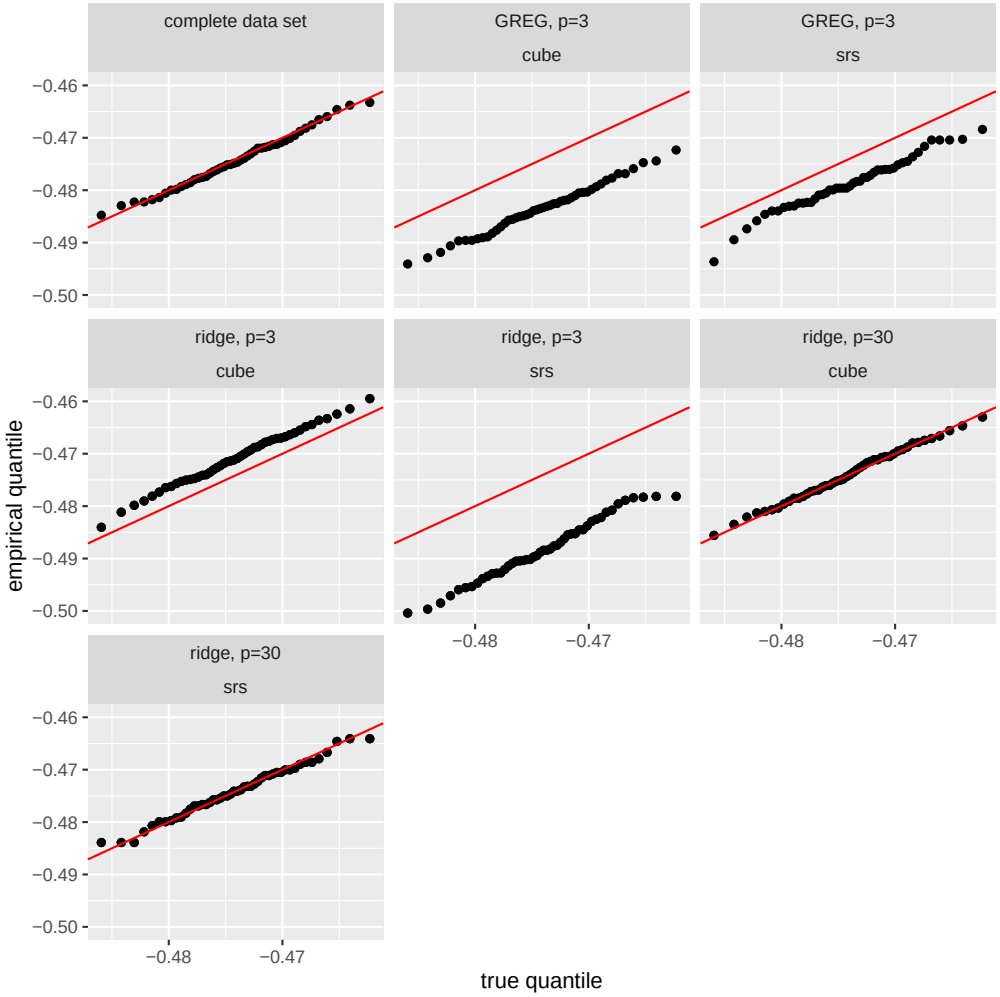


Figure 4.4: Exemplary quantile-quantile plots for a single parameter.

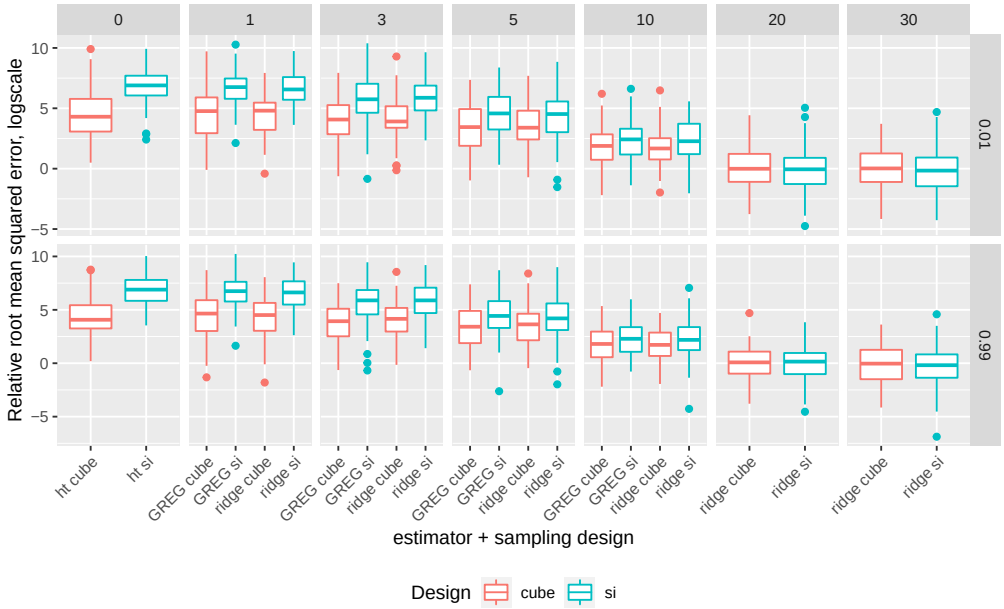


Figure 4.5: Gaussian likelihood with known variance: Root mean squared error for quantiles of $\boldsymbol{\theta}$, relative to root mean squared error obtained with complete data set, for quantile $q = 0.01, 0.99$.

for estimation, this difference becomes small. Given a fixed value of p_{greg}, the difference in error between the basic greg estimator and the ridge estimator are small. However, the ridge variant allows the use of a larger number of auxiliary variables and is therefore to be preferred.

Quantiles: To study the accuracy of the simulated posterior distribution, we set σ equal to one and assume it to be known, so that the posterior distribution under a non-informative prior $\beta \propto constant$ is given by a multivariate normal distribution with variance $\boldsymbol{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and mean $\boldsymbol{\Sigma} \boldsymbol{X}^\top \boldsymbol{y}$. As such, we can evaluate how well the true posterior distribution is simulated as whole, beyond point estimation. Exemplary quantile-quantile plots for a single parameter are shown in figure 4.4. The results are similar to the ones regarding $\boldsymbol{\theta}$: More auxiliary variables lead to a better approximation of the results of the complete data set, cube sampling improves over simple random sampling. Figure 4.5 shows the relative root mean squared error for the marginal quantiles of $\theta_1, \ldots, \theta_5$. The root mean squared error of the quantile vector is computed as:

$$rmse(Q_q(\boldsymbol{\theta})) = \sqrt{\frac{1}{5}\sum_{i=1}^{5}(Q_q(\theta_i) - \hat{Q}_q(\theta_i))^2},$$

for $q = 0.01, 0.99$, where $Q_q(x)$ is the q-Quantile of x. Here, the results are similar to the ones regarding $\boldsymbol{\theta}$ as well, however the relative root mean squared error does not approach 1 as fast as for point estimation.

4.6 Application

The methods presented here will be applied on a large data set consisting of the most recent gas price for every hour for every gas station in Germany for the first quarter of 2015. The data set consists of 31 million rows. We estimate the

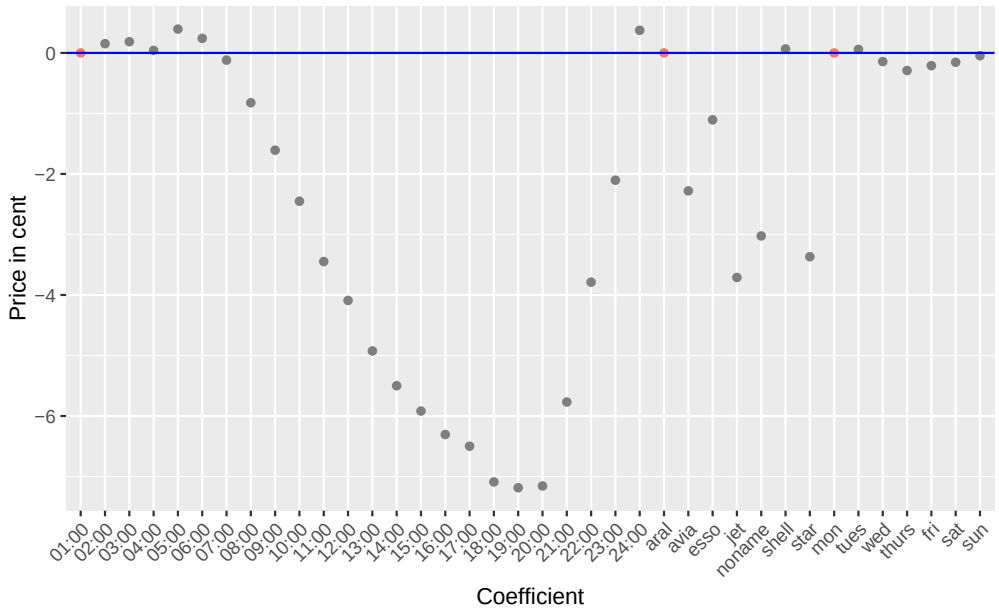


Figure 4.6: Coefficient plot (posterior means). Reference categories are added as red points. Posterior intervals for 0.01-quantile and 0.99-quantile are barely visible and omitted.

regression model:

$$price_{i,t} = \beta_1 + f_{periodic}(weekday) + f_{periodic}(hour) + \sum_{j=2}^{6} \beta_j I[brand_i = j] + \epsilon_{i,t}, \epsilon_{i,t} \sim N(0, \sigma^2),$$

where the periodic effects are estimated using dummies, setting one category equal to zero, so that $\boldsymbol{\theta} = (\beta_1, \boldsymbol{\beta}_{day}^\top, \boldsymbol{\beta}_{hour}^\top, \boldsymbol{\beta}_{brand}^\top, \sigma)^\top$. We use the ridge variant with a sample size of $n = 10000$ and $p_{greg} = 50$ auxiliary variables. For the cube sampling, 3 partial least squares latent variables are used. Partial least-squares linear combinations are fast to compute, and summarize the design matrix while preserving the correlation with y. The algorithm takes about 11 minutes on a Intel xeon cpu e5 with 2.4GHZ. Figure 4.6 reports the estimated posterior means of the regression coefficents. Most variation is explained by the hourly effects, diesel price is most expensive in the morning, least expensive in the evening around 19:00 with a difference of around 7 cents. The day of the week effect is very small, compared to the hour effect: Gas prices are highest on mondays, with an estimated price difference of less than 0.5 cent. The most expensive brand is Shell, followed by Aral. The smaller brands (Star and Jet) are more than 3 cents cheaper than Aral. These results are country-wide and do not take local effects into account, mainly the effect of the location of the gas station. Furthermore, there might be an interaction between the day of the week and the hour of the day. These topics should be the subject of a separate study.

4.7 Discussion

We developed an approximate version of the Metropolis-Hastings using a single subsample. Our simulation study shows that the algorithm can produce posterior simulations which are almost identical to the use of the full data while being several

orders of magnitudes faster. The methods were validated using a simulation comparing the inferences obtained from the approximate posterior to the complete data posterior. The results suggest that the error obtained from using the algorithm is small enough for practical purposes for standard models. However, further work regarding more complex models is necessary.

We propose to use this approximate version for the simulation of a posterior distribution which is used directly for inference. However, there are further possible applications for the algorithm; the algorithm might be used for the case when a fast approximation to the posterior distribution is of use, e.g., during the burnin of an adaptive Metropolis-Hastings algorithm.

4.A Appendix

Computation of Ridge estimator

For the model

$$\boldsymbol{\phi} = \boldsymbol{Z}_{greg}\boldsymbol{\beta} + \boldsymbol{e},$$

the ridge estimator for a given smoothing parameter κ is given by $\boldsymbol{\beta} = (\boldsymbol{Z}_{greg}^\top \boldsymbol{Z}_{greg} + \kappa \boldsymbol{I})^{-1}\boldsymbol{Z}_{greg}\boldsymbol{\phi}$, where the rows of $\boldsymbol{Z}_{greg}$ are given by $\boldsymbol{z}_{greg,k}$, for $k \in \mathcal{S}$. Define $\boldsymbol{M} \coloneqq \boldsymbol{Z}_{greg}\boldsymbol{G}$, $\boldsymbol{\xi} \coloneqq \boldsymbol{G}^\top\boldsymbol{\beta}$, then

$$\boldsymbol{\phi} = \boldsymbol{Z}_{greg}\boldsymbol{G}\boldsymbol{G}^\top\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{M}\boldsymbol{\xi} + \boldsymbol{e};$$

where $\boldsymbol{Z}_{greg}^\top \boldsymbol{Z}_{greg} = \boldsymbol{G}\boldsymbol{\Lambda}\boldsymbol{G}^\top$ is the eigen-decomposition of $\boldsymbol{Z}_{greg}^\top \boldsymbol{Z}_{greg}$ with orthonormal $\boldsymbol{G}$, i.e., $\boldsymbol{G}^\top\boldsymbol{G} = \boldsymbol{G}\boldsymbol{G}^\top = \boldsymbol{I}$, so that $\Lambda = diag(\lambda_1, \ldots, \lambda_k)$,

$$\lambda_1 \geq \lambda_2, \geq \cdots \geq \lambda_p$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{Z}_{greg}^\top \boldsymbol{Z}_{greg}$. It holds that $\boldsymbol{Z}_{greg}^\top \boldsymbol{Z}_{greg} + c\boldsymbol{I} = \boldsymbol{G}(\boldsymbol{\Lambda} + c\boldsymbol{I})\boldsymbol{G}^\top$, so that

$$(\boldsymbol{Z}_{greg}^\top \boldsymbol{Z}_{greg} + \kappa\boldsymbol{I})^{-1} = \boldsymbol{G}(\Lambda + \kappa\boldsymbol{I})^{-1}\boldsymbol{G}^\top,$$

and

$$\boldsymbol{\beta}(\kappa) = \boldsymbol{G}(\Lambda + \kappa\boldsymbol{I})^{-1}\boldsymbol{M}^\top\boldsymbol{\phi} \rightarrow$$
$$\boldsymbol{\xi}(\kappa) = (\Lambda + \kappa\boldsymbol{I})^{-1}\boldsymbol{M}^\top\boldsymbol{\phi}.$$

Define $\boldsymbol{c} \coloneqq \boldsymbol{M}^\top\boldsymbol{\phi}$. Than, the elements of $\hat{\boldsymbol{\xi}}(\kappa)$ are

$$\hat{\xi}_i = \frac{c_i}{\lambda_i + \kappa}, i = 1, \ldots, p_{greg}.$$

Following Hoerl, Kannard, and Baldwin (1975), the ridge parameter is set as:

$$\hat{\kappa} = \frac{p_{greg}\hat{\sigma}^2}{\hat{\boldsymbol{\xi}}(0)^\top\hat{\boldsymbol{\xi}}(0)}. \tag{4.9}$$

The residual variance is estimated using the usual unbiased estimator

$$\begin{aligned}\hat{\sigma}^2 &= (n - p_{greg})^{-1}\sum_{i\in\mathcal{S}}(\phi_i - \boldsymbol{z}_i^\top\hat{\boldsymbol{\beta}}(0))^2 \\ &= (n - p_{greg})^{-1}(\boldsymbol{\phi} - \boldsymbol{M}\hat{\boldsymbol{\xi}}(0))^\top(\boldsymbol{\phi} - \boldsymbol{M}\hat{\boldsymbol{\xi}}(0)).\end{aligned}$$

The vector $\hat{\boldsymbol{\beta}}(\hat{\kappa})$ does not have be computed, using representation (4.7) of the greg estimator: Define $\tilde{\dot{\boldsymbol{z}}}_{greg} \coloneqq \dot{\boldsymbol{z}}_{greg}^\top\boldsymbol{G}$ and $\tilde{\dot{\boldsymbol{z}}}_{greg,\mathcal{S}} \coloneqq (\sum_{\mathcal{S}}\boldsymbol{z}_i)^\top\boldsymbol{G}$, then $\dot{\phi}$ is estimated via:

$$\boldsymbol{\xi}(\hat{\kappa})^\top\tilde{\dot{\boldsymbol{z}}}_{greg} + (N/n)\left((\sum_{\mathcal{S}}\phi_k) - \boldsymbol{\xi}(\hat{\kappa})^\top\tilde{\dot{\boldsymbol{z}}}_{greg,\mathcal{S}}\right).$$

The vectors $\tilde{\dot{\boldsymbol{z}}}_{greg}$ and $\tilde{\dot{\boldsymbol{z}}}_{greg,\mathcal{S}}$ only have to be computed once.

Part II

Applications in energy and health economics

Chapter 5

Market premia for renewables in Germany: The effect on electricity prices[1]

Abstract

Due to the growing share of "green" electricity generated by renewable energy technologies, the frequency of negative price spikes has substantially increased in Germany. To reduce such events, in 2012, a market premium scheme (MPS) was introduced as an alternative to feed-in tariffs for the promotion of green electricity. Drawing on hourly day-ahead spot prices for the time period spanning 2009 to 2016 and employing a nonparametric modeling strategy called Bayesian Additive Regression Trees, this paper empirically evaluates the efficacy of Germany's MPS. Via counterfactual analyses, we demonstrate that the introduction of the MPS decreased the number of hours with negative prices by some 70%.

Keywords: Negative electricity prices; Merit order effect; Bayesian Additive Regression Trees

[1]This paper is jointly written with Manuel Frondel and Stephan Sommer and has been published as: Frondel, M. and Kaeding, M. and Sommer, S. (2020). Market premia for renewables in Germany: The effect on electricity prices. *SFB 823 Discussion paper, 13*, 1–31. doi:10.2139/10.17877/DE290R-21016 and Frondel, M. and Kaeding, M. and Sommer, S. (2020). Market Market premia for renewables in Germany: The effect on electricity prices. *USAEE Working Paper, 456(20)*, 1–31. doi:10.2139/ssrn.3643762

5.1 Introduction

Almost all over the world, policy-makers foster the deployment of renewable energy technologies, such as solar and wind power, as a means to reduce carbon emissions. The European Union (EU), for instance, aims at raising the share of "green" electricity, produced by renewable energy technologies, in electricity consumption from about 29% in 2015 to 56% by 2030 (Agora, 2019). To achieve this target, the majority of Member States has established promotion schemes for renewable energy sources (RES) that are based on subsidies (IEA & IRENA, 2018).

In Germany, for example, since the beginning of the new millennium, green electricity has been promoted via technology-specific feed-in tariffs (FITs) that guarantee fixed payments per kilowatthour (kWh) to the plant operators for up to 21 years. Moreover, grid operators are obliged to give priority dispatch to RES (Andor, Frondel, & Vance, 2017). In the aftermath of the introduction of the feed-in tariff system, Germany has been very successful in increasing RES capacities, yet their massive increase has come at high cost: Currently, the annual promotion costs amount to more than 25 billion euros, equaling about 1% of the German GDP (Andor, Frondel, & Sommer, 2018).

In addition to the substantial cost due to the unconditonal payment of fixed tariffs irrespective of demand levels, the steadily growing amount of green electricity has further adverse economic effects: it tends to increase both grid balancing costs and the frequency of negative prices (Nicolosi, 2010; Weber, 2010). In fact, while RES plants are not obliged to contribute to grid stability, abundant green electricity production may lead to negative prices when a high electricity supply coincides with a low demand, thereby inducing welfare losses (Andor, Flinkerbusch, Janssen, Liebau, & Wobben, 2010).

Gerster (2016), for instance, demonstrates that the growing feed-in of green electricity raises the probability of negative price spikes and thus threatens the financial viability of conventional plants. Negative prices primarily arise because conventional power plants face substantial ramp-up costs (Gerster, 2016) and, hence, are most often not shut down despite negative revenues due to negative prices. Not least, numerous studies argue that there is a negative link between green electricity generation and the wholesale price, commonly referred to as the merit-order effect (Cludius, Hermann, Matthes, & Graichen, 2014; de Lagarde & Lantz, 2018; Paschen, 2016; Praktiknjo & Erdmann, 2016; Würzburg, Labandeira, & Linares, 2013).

To avoid such adverse effects and to align the production of green electricity with market signals, many countries have implemented market premium schemes that pay operators of renewable plants a variable bonus on top of the wholesale electricity price, rather than guaranteeing fixed feed-in-tariffs (RES, 2018). Germany, for example, introduced such a market premium scheme (MPS) in January 2012 and revised it in August 2014.

Using hourly day-ahead spot prices for the time period spanning January 2009 to December 2016, this paper empirically evaluates the efficacy of the German MPS, particularly with respect to reducing the frequency of negative price spikes. To this end, we employ a nonparametric modeling strategy called Bayesian Additive Regression Trees (BART), which is highly appropriate for identifying nonlinear interactions between covariates (Hill, 2011) and for predicting those electricity prices that would have emerged in the counterfactual scenario without the MPS. Our results indicate that the BART method is very successful in

replicating past electricity prices, as we are able to replicate those roughly 470 hours in which negative prices occurred within the sample period of 2009 to 2016.

Moreover, our results suggest that the introduction of the MPS led to fewer hours with negative prices, particularly in the morning, and fewer positive price peaks relative to the counterfactual situation of an absent MPS. By comparison with this counterfactual, we demonstrate that negative electricity prices were avoided by some 70%, in more than 560 hours. Based on these empirical results, we conclude that the MPS turned out to be effective in increasing the market integration of renewable energy technologies and, hence, in reducing the costs of their promotion.

The subsequent sections describe Germany's electricity market and its MPS, as well as the data base underlying our research. Section 5.4 briefly introduces the BART modeling method, while Section 5.5 presents our empirical results. The last section summarizes and concludes.

5.2 Germany's Market Premium for RES

Germany stipulated ambitious targets for the expansion of RES, aiming at increasing the share of RES-based electricity in consumption to 35% by 2020 and to 80% by 2050. As more than 35% of electricity consumption was covered by RES technologies in 2019, compared to less than 7% in 2000 (BMWi, 2017), Germany has already reached its 2020 target. Clearly, the key driver of this rapid expansion is the Renewable Energy Sources Act (Erneuerbare-Energien-Gesetz, EEG), which came into force in 2000. Its key characteristic is a set of technology-specific FITs that are granted for up to 21 years in an intertemporally fixed amount. While exceeding the average generation cost of conventional electricity, in many cases substantially, these technology-specific tariffs are paid for each kWh fed into the grid irrespective of the level of demand for electricity. In addition, electricity based on RES enjoys preferential access to the grid. Both features, FITs and priority of green over conventional electricity, were established to shield plant operators from adverse market signals (Andor & Voss, 2016).

In terms of RES capacity expansion, Germany's FIT system proved highly successful: Between 2000 and 2016, RES capacities increased ten-fold to reach 104 Gigawatt (GW), thereby exceeding the capacity of conventional power plants for the first time (BMWi, 2017). Above all, photovoltaic (PV) capacities skyrocketed, most notably in the boom years 2010 to 2012: In each of these years, almost 8 GW were newly installed, with the sum of PV capacities installed in this period being equivalent to almost 60% of the total PV capacity in 2016 (Table 5.1). With an increment of about 4 GW in 2016, prior to Germany's introduction of auctioning systems for the installation of new renewable energy plants in 2017, onshore wind has experienced a strong push as well.

The expansion of RES capacities causes numerous problems for the electricity system, in particular with respect to grid stability. For instance, because grid operators are obliged by law to treat green electricity preferentially, conventional power plants have to adjust their production downward when demand is low (Römer, Reichhart, Kranz, & Picot, 2012). In contrast, RES technologies are not obliged to contribute to the balancing of supply and demand, thereby undermining grid stability due to their intermittent electricity generation. In fact, prior to the introduction of the MPS in January 2012, operators of renewable power plants, such as wind turbines and PV systems, had no monetary incentive to cease their electricity generation even when prices were negative, as they earned a fixed

Table 5.1: Electricity Generation Capacity (in MW) in Germany Prior to the Introduction of a Tendering System in 2017

	Hydro, Bio, Geo	Wind Onshore	Wind Offshore	Photo-voltaic	Total Renewables	Total Conventional
2000	5,534	6,097	0	114	11,745	107,500
2001	5,658	8,738	0	176	14,572	106,800
2002	5,967	11,976	0	296	18,239	100,900
2003	6,381	14,381	0	435	21,197	99,400
2004	6,873	16,419	0	1,105	24,397	100,900
2005	7,562	18,248	0	2,056	27,866	98,900
2006	8,203	20,474	0	2,899	31,576	98,400
2007	8,532	22,116	0	4,170	34,818	99,800
2008	8,848	22,794	0	6,120	37,762	101,800
2009	10,219	25,697	35	10,566	46,517	101,200
2010	10,878	26,823	80	18,006	55,787	104,100
2011	12,057	28,524	188	25,916	66,685	98,100
2012	12,379	30,711	268	34,077	77,435	97,300
2013	12,656	32,969	508	36,710	82,843	94,600
2014	12,873	37,620	994	37,900	89,387	100,200
2015	13,090	41,297	3,283	39,224	96,894	97,600
2016	13,308	45,460	4,132	40,716	103,616	96,800

Source: BMWi (2017).

FIT for each kWh of green electricity that was fed into the grid. This producer behavior is often described by the formula "produce and forget", indicating that producers of green electricity can be completely oblivious to market signals and electricity demand.

To better align green electricity production with demand, Germany's FIT system for RES promotion was supplemented in 2012 by the MPS to increase the incentives for a demand-orientated generation of green electricity. The aim of this scheme was to encourage operators of RES plants to sell their green electricity to the market, rather than feeding it into the grid at any time, thereby obtaining fixed FITs per kWh. While taking part in the MPS was voluntary and operators of RES plants could switch back and forth between the MPS and the FIT system on a monthly basis, risk-averse RES operators could always stay in the traditional FIT system.[2]

Yet, switching to the MPS is attractive, as the remuneration per kWh may be higher than under the FIT system (Figure 5.1). Under the MPS, in addition to the spot market price SP, plant operators receive a variable premium, called market premium (MPR). It is calculated ex post on a monthly basis as the difference between the feed-in-tariff FIT and the electricity's technology-specific average market value MV in the previous month.[3] Not least, RES operators receive a fixed management premium MMP that is intended to cover the costs arising from participating in the market. Thus, under the MPS, total revenues R are given by the sum of three components: $R := SP + MPR + MMP$, the spot

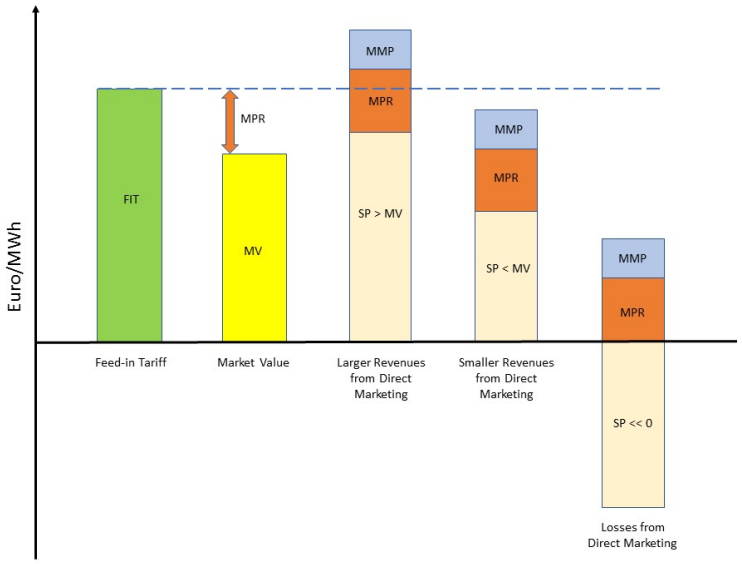
[2] A prerequisite for participating in the MPS is that plants can be curtailed via remote control.

[3] The market values of electricity generated by various RES technologies differ because of distinct production peaks. For instance, solar electricity production peaks at noon in the summer months, thereby coinciding with high electricity demand and, thus, high prices, whereas wind power frequently peaks at winter nights when demand and prices tend to be lower.

price, the market and the management premium.

Unlike the FIT system, revenues are not guaranteed under the MPS, but are primarily determined by the market. Hence, RES plant operators are exposed to price risks, which in times of high demand may turn out to be positive: If the spot market price SP exceeds the average market value MV of the previous month, the revenues emerging from the MPS are higher than those from the FIT system (see Figure 5.1). However, while in the FIT system green electricity production always yields positive revenues, under the MPS, the generation of green electricity may not only yield lower revenues than under the FIT system, but might even turn out to be unprofitable. In fact, recognizing that the marginal cost of production is negligible for renewable technologies such as wind and solar power, if spot market prices are negative, green electricity production is only profitable under the MPS as long as the sum of the market premium MPR and the management premium MMP exceeds the magnitude of the negative spot market price SP.

Figure 5.1: Revenues under Germany's Feed-in Tariff (FIT) System and its Market Premium Scheme (MPS)



Hence, operators stop feeding green electricity into the grid when the spot market price is notably below zero and the losses from direct marketing exceed the premia paid to the operators: $SP + MPR + MMP < 0$. Therefore, as was intended with the introduction of the MPS, plant operators would curtail their production in response to significantly negative prices. In case of moderately negative prices, however, even under the MPS, producers still have an incentive to feed their green electricity into the grid. In other words, negative electricity prices cannot be entirely avoided by the MPS.

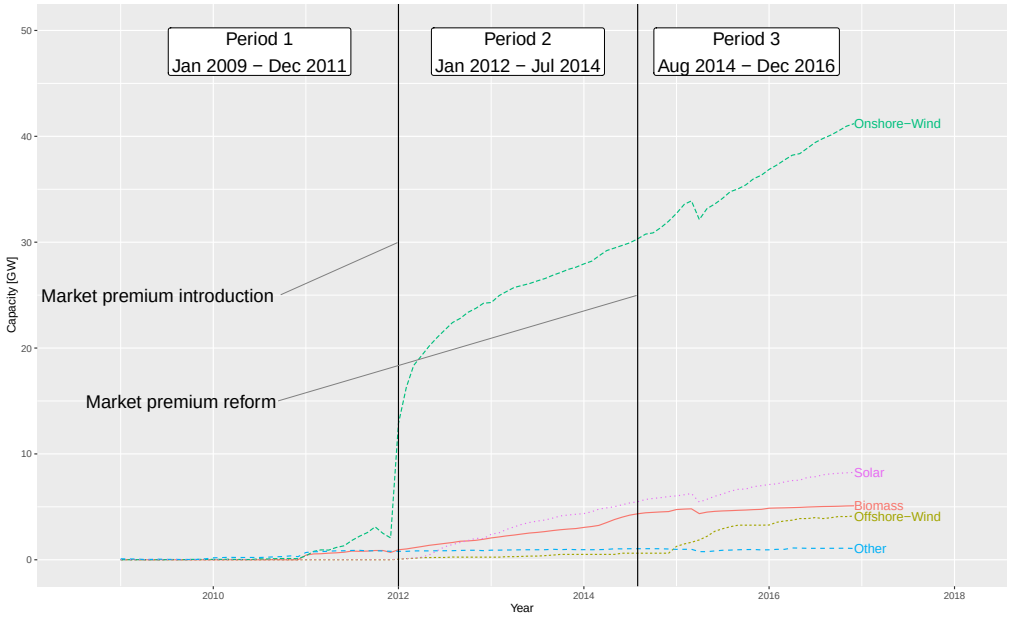
In August 2014, Germany revised the MPS by mandating participation of all new RES plants with a capacity of more than 500 kilowatt (kW). Still, though, operators of RES plants that were installed prior to August 2014 remained free to choose between the MPS and the FIT system. In 2016, the capacity limit of obligatory participation of 500 kW was further reduced: Since January 1, 2016, all new RES plants with a capacity of more than 100 kW must participate in

the MPS. To further diminish the frequency of negative prices at the wholesale market, another market incentive was established in the 2014 reform: If electricity prices are negative for at least six consecutive hours, the market premium for RES plants that came into operation after January 1, 2016, and whose capacity exceeds 500 kW (3,000 kW in the case of wind power) is defined to equal zero during the entire period with negative prices.

With another reform of its feed-in tariff system that came into force at the outset of 2017, the German government aimed at increasing the economic efficiency of the RES promotion by implementing technology-specific tender schemes, where participants bid on the subsidy amount per kWh they receive when feeding-in green electricity into the grid. Still, though, RES operators can opt for selling their electricity at the market via the MPS. While the reform of 2017 thus should not have affected the occurrence of negative electricity prices, we nevertheless have deliberately restricted our sample to observations prior to this new promotion regime.

As Figure 5.2 reveals, prior to 2012, just a few operators sold their green electricity at the market, instead of enjoying fixed FITs. In the aftermath of the introduction of the MPS, though, the share of green electricity that was sold at the market and remunerated via the MPS rapidly increased. Most notably, many operators of onshore wind farms and biomass plants switched immediately to the MPS in 2012, most likely because of the generous management premium MMP. By contrast, only a minority of the PV plant operators opted for the MPS, which is due to the fact that the overwhelming majority of PV systems are installed on the roofs of private houses.[4]

Figure 5.2: Capacities Participating in the Market Premium Scheme (MPS)



Source: TSO (2018).

[4]The kink in the uptake of the direct selling scheme in early 2015 can be explained by the obligation to dispose of a remote control. Hence, plants operators who did not install such a device were not allowed to sell their electricity directly. Obviously, these operators reacted quickly and equipped their plants with the required technology.

5.3 Data

To gauge the impact of the MPS on electricity prices, our research draws on data that covers the period from January 2009 to December 2016 and originates from two sources: the websites of Fraunhofer Institute for Solar Energy Systems (2018) and the European Energy Exchange (EEX, 2018). From the EEX, we retrieved the price of the EU emission allowances, which represents an important control variable, as well as the Physical Electricity Index (Phelix). This index serves as the dependent variable in our analysis and measures the hourly electricity spot market price for the joint German-Austrian market area.

The key explanatory variables of interest, the hourly electricity generation of wind and PV power plants, as well as other control variables, such as the hourly demand load and electricity generation from conventional sources, are taken from Fraunhofer Institute for Solar Energy Systems (2018). Additionally, we control for different patterns of seasonality by including the hour of the day, the day of the week, the day of the year, and the month.

In line with the timing of the introduction of the German MPS in 2012 and its reform in August 2014, we divide our data base into three disjunct periods of quite similar length: Period 1 spans from January 2009 to December 2011, that is, the time period prior to the introduction of the MPS. Period 2 ranges from January 2012 to July 2014, that is, the period between the introduction and the reform of the MPS, and Period 3 spans from August 2014 to December 2016, that is, the time period after the reform of the MPS and the launch of the tendering system for PV and wind power installations in Germany in 2017.

Over the entire period of January 2009 to December 2016, about 17% of Germany's electricity demand load could be attributed to wind and solar power plants, averaging 6,412 Megawatt (MW) and 3,174 MW, respectively (Table 5.2). Yet, there was substantial inter-temporal variation due to the intermittent character of both technologies. For instance, 25,643 hours elapsed without any feed-in from solar power plants, while another 665 hours saw solar electricity production exceeding that of lignite and hard coal fired plants.

Table 5.2: Summary Statistics for the Time Period January 2009 to December 2016

Variable	Mean	Std. Dev.	Min	Max	# Obs.	Frequency
Phelix (EUR/MWh)	38.54	16.67	−500.0	210.0	69,527	hourly
Allowance price (EUR/t CO2)	9.10	4.06	3.02	16.84	2,511	daily
Solar energy (MW)	3,174	5,171	0	28,323	60,434	hourly
Wind energy (MW)	6,412	5,798	0	34,078	60,434	hourly
Lignite (MW)	15,760	2,169	0	20,940	60,434	hourly
Hard coal (MW)	11,539	5,436	0	22,298	60,434	hourly
Nuclear energy (MW)	11,096	2,640	0	18,506	60,765	hourly
Load (MW)	56,095	10,383	29,201	81,109	60,765	hourly

Between January 2009 and December 2016, hourly electricity prices averaged EUR 38.5 per Megawatthour (MWh), but exhibited substantial volatility, as they ranged from EUR -500 to EUR 210. For 464 hours, negative prices emerged, 98 hours in Period 1, 146 hours in Period 2, and 220 hours in Period 3. Given the total number of 61,320 hours in the period spanning 2009 to 2016, the occurrence of negative prices was rather seldom, yet clearly increasing over time (Figure 5.3). Negative prices most likely occur when electricity demand is low, primarily at holidays (Figure 5.3) and in the early morning hours (Figure 5.5). In fact,

around half of the number of hours with negative prices fell between midnight and 5 a.m.

Figure 5.3: Hourly Spot Market Electricity Prices for Germany between 2009 and 2016 Reflected by the Physical Electricity Index (Phelix)

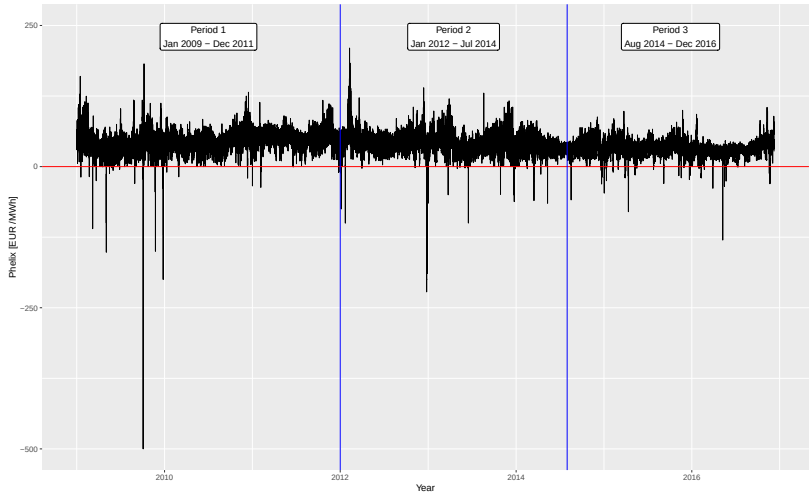
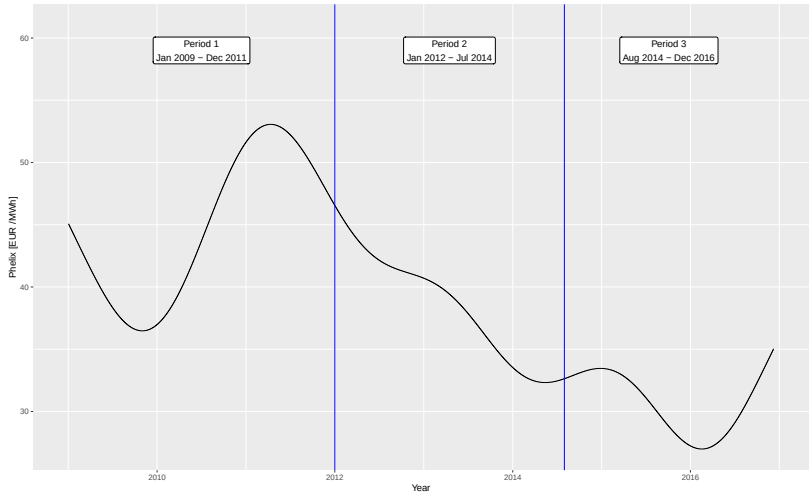


Figure 5.4: Daily Spot Market Electricity Prices for Germany between 2009 and 2016 given by the Physical Electricity Index (Phelix)



From Period 1 to Period 3, there was a declining trend in electricity prices (Figure 5.4): Prices peaked at about EUR 55 per MWh in mid-2011 due to the sudden shutdown of about half of Germany's nuclear power plants in the aftermath of the accident in Japan's Fukushima, then fell to EUR 25 per MWh at the outset of 2016, but recovered afterwards. In fact, the period after the introduction of the MPS, but before its reform in August 2014 (Period 2), is characterized by a downward drift in prices. Overall, in Period 3, electricity prices were lowest and their distribution was tightest (Figure 5.6).

5.4 Method

The effect of RES generation on electricity prices has been typically estimated on the basis of time series models – see, for instance, de Lagarde and Lantz (2018),

Figure 5.5: Percent Negative Prices - Within Day Distribution

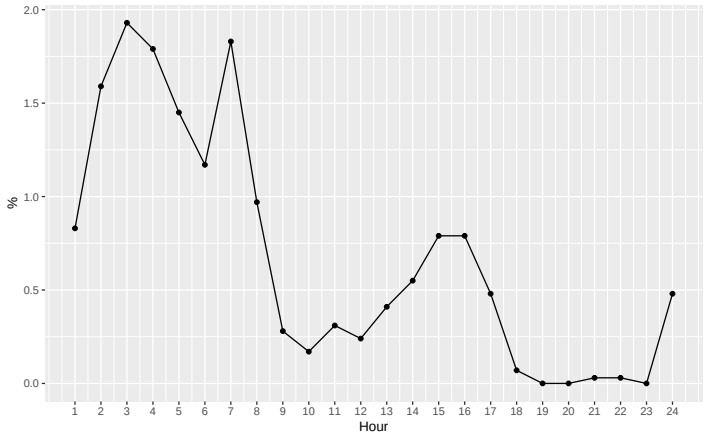
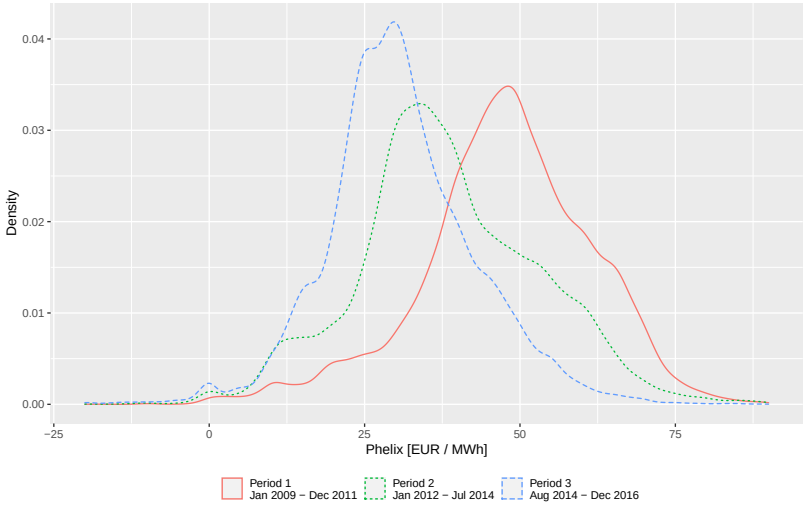


Figure 5.6: Distribution of the Physical Electricity Index (Phelix, in Euro per Megawatthour) by Periods



Fanone, Gamba, and Prokopczuk (2013), Gerster (2016), Ketterer (2014), and Paschen (2016). This approach necessitates the specification of a functional form that models the relationship between prices and key variables, such as hour of the day, as well as a set of covariates. In contrast, we apply a flexible nonparametric method called Bayesian Additive Regression Trees (BART) that requires less guess-work in model fitting. Actually, a key advantage of the BART method is that there is no necessity to specify a parametric form for the conditional expectation $E[p_t|\boldsymbol{z}_t]$ of the dependent variable, where in our example p_t denotes the price of electricity in hour t and $\boldsymbol{z}_t$ designates the vector of covariates.

Furthermore, BART methods are particularly appropriate for identifying complex interactions between covariates and, hence, effect heterogeneity (Hill, 2011). This is most relevant in the context of determining the effect of RES generation on electricity prices given that electricity demand load, as well as green electricity generation, highly vary over time, making dummies for the day of the week and the hour of the day, as well as interaction terms of these variables, indispensable model ingredients. Another advantage is that, owing to the flexibility of the model, predictions from BART models are highly precise – see for instance the empirical examples provided by Hill (2011), as well as our own

results presented below. Not least, rather than estimating numerous separate models, for instance for each hour of a day individually, for both peak- and off-peak hours (see Gerster, 2016), a single BART model suffices for our analysis.

The basic idea of the BART modeling strategy is to explain the outcome variable, in our cases electricity prices p_t, by a large number L, say $L = 200$, of regression trees:

$$p_t = \sum_{l=1}^{L} f(\boldsymbol{z}_t|\boldsymbol{\theta}_l) + u_t, \tag{5.1}$$

where $u_t \sim N(0, \sigma^2)$, $f(\boldsymbol{z}_t|\boldsymbol{\theta}_l)$ is called tree function and $\boldsymbol{\theta}_l$ denotes a parameter vector that describes tree l – see Breiman (2001) for an introduction to tree- and forest-based methods. Each tree l explains a part of the variation in the dependent variable p_t using only a subset of the covariates. If a tree refers to a single covariate, it captures a main effect, otherwise a tree reflects the effects of two or more covariates and their interaction. The high flexibility of BART results from the potentially large number of trees with which any sort of interaction between covariates can be modeled.

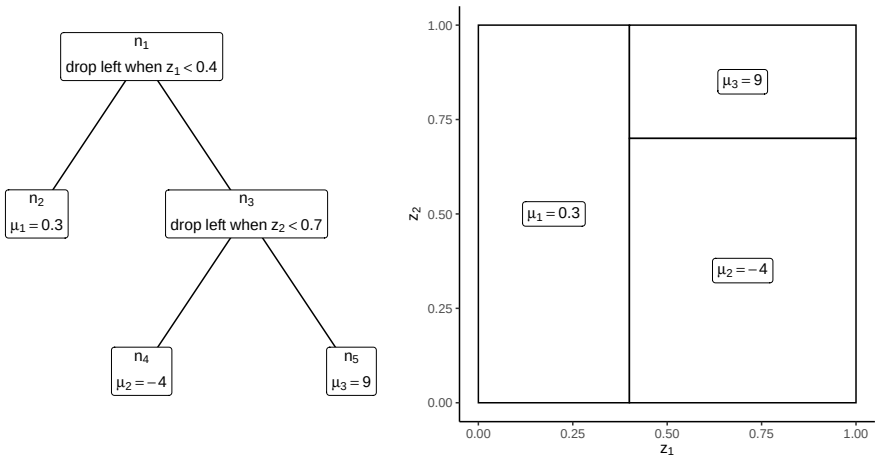
As an illustration, Figure 5.7 presents an example of a single tree that at the first interior node n_1 sends all observations with $z_1 < 0.4$ to the left branch, ending in terminal node n_2. The right branch divides observations further according to the decision rule $z_2 < 0.7$, resulting in terminal nodes n_4 and n_5. Formally, a tree of size s is defined by a node vector $\boldsymbol{n} \coloneqq (n_1, ..., n_j, ..., n_s)^\top$ consisting of both interior and terminal nodes. To each interior node, a threshold c_j is associated that implies a decision rule on how to traverse the tree: If $z_{it} < c_j$, observation z_{it} is sent left, if $z_{it} \geq c_j$, it is sent right in the tree. Thresholds are determined *a posteriori* on the basis of the observations on the covariates. Each terminal node is associated with a parameter μ_k, representing the mean outcome of the subgroup of observations that are sent to that terminal node. Gathering mean outcomes in vector $\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_f)^\top$, where f is the number of terminal nodes, and defining $\boldsymbol{c} \coloneqq (c_1, ..., c_g)^\top$, where g indicates the number of internal nodes, the parameter vector $\boldsymbol{\theta}_l \coloneqq (\boldsymbol{c}_l^\top, \boldsymbol{n}_l^\top, \boldsymbol{\mu}_l^\top)^\top$ determines the tree function $f(\boldsymbol{z}_t|\boldsymbol{\theta}_l)$ of tree l.

To provide intuition for the BART modeling approach, consider taking the fit $f(\boldsymbol{z}_t|\boldsymbol{\theta}_1)$ from the first tree and subtracting it off from the observed outcome p_1 to form residuals. Then imagine fitting the next tree to these residuals, each tree contributing to the overall fit. This process would be performed many times, thereby ever improving the overall fit. While improving the overall fit by adding tree by tree, however, one wants to avoid overfitting. To counter overfitting, Chipman, George, and McCulloch (2010) propose regularization priors, such as prior distributions to penalize the number of nodes and the means of each component tree. For instance, the prior for the number of nodes puts most probability mass on trees with two interior nodes, while still allowing large trees if the data demands it. The typical prior for the thresholds assigns each observed value the same probability.

To estimate the effect of the introduction of the MPS in 2012, as well as of its reform in August 2014, we split the data set into the three time periods described in the previous section and define the following potential outcomes in terms of prices:

$$p_t^s(r) = \begin{cases} p_t^s(1) & \text{Regime 1: Absence of the MPS,} \\ p_t^s(2) & \text{Regime 2: Prevalence of the MPS,} \\ p_t^s(3) & \text{Regime 3: Prevalence of the Revised MPS,} \end{cases} \tag{5.2}$$

Figure 5.7: Illustration of a Regression Tree



The illustration depicts a tree of size 3 for function $f(z_1, z_2) = 0.3I[z_1 < 0.4] + 9I[z_1 \geq 0.4]I[z_2 < 0.7] - 4I[z_1 \geq 0.4]I[z_2 \geq 0.7]$, where $I[.]$ equals one if the condition in brackets is met. The left panel shows the tree in terms of nodes, where with each interior node n_1 and n_3, a decision rule is associated. The terminal nodes n_2, n_4, and n_5 are associated with mean outcomes, displayed in the partition in the right panel.

where $p_t^s(r)$ denotes the price of electricity in hour t under Regime r in Period s. As presented in Definition (5.2), there are three regimes that emerge in a natural way due to the introduction and revision of the MPS. Given our focus on the frequency with which negative prices occur, to estimate the effect of the MPS, in principle, we would like to calculate the difference between the actual frequency of negative prices in Period 2 and the hypothetical frequency in the counterfactual situation of an absence of the MPS (Regime 1).

By employing BART to estimate the effect of the MPS, we follow Imbens and Rubin (2015) and, for each draw $\boldsymbol{\theta}_m$ from the posterior distribution of tree parameter vector $\boldsymbol{\theta}$, we simulate potential outcomes for the prices $p_t^s(r)$ that pertain to all nine combinations of regimes and periods. Altogether, we draw $M = 10,000$ parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ from the posterior distribution, after discarding M parameter draws during the burnin phase, in which the sampler reaches its equilibrium distribution. We note that our empirical results remain largely unchanged when M is larger than 10,000. Precisely, we estimated our model with $M = 20,000$ and $M = 40,000$ and observe that our results do not change. We opted for $M = 10,000$ as it is computationally less demanding. Based on these price simulations, we then compute the number of hours with negative prices in Period s under Regime r:

$$\sum_{t \in \text{Period } s} I\left[p_t^s(r) < 0\right],$$

where indicator function $I\left[p_t^s(r) < 0\right]$ indicates the occurrence of a negative price in hour t. Taking respective differences in the number of negative hours across regimes yields the effects τ of introducing and revising the MPS. For instance,

$$\tau^2(1) := \sum_{t \in \text{Period } 2} I\left[p_t^2(1) < 0\right] - \sum_{t \in \text{Period } 2} I\left[p_t^2(2) < 0\right]$$

reflects the effect of the MPS (Regime 2) relative to Regime 1 (absence of MPS). A positive value of $\tau^2(1)$ would indicate that the introduction of the MPS is associated with less hours with negative prices. Repeating this exercise $M = 3,000$ times, we ultimately get an estimate $\widehat{\tau^2(1)} := M^{-1} \sum_{m=1}^{M} \tau_m^2(1)$ of the effect of the introduction of the MPS in Period 2 by taking the mean of the differences $\tau^2(1)$. Finally, the lower and upper bounds of the corresponding posterior intervals, reflecting the uncertainty of the BART estimator, are given by the 2.5th and 97.5th percentiles of the sequence of $M = 3,000$ estimates $\tau_1^2(1), \ldots, \tau_M^2(1)$. We can extend this procedure to estimate heterogeneous group effects by computing $\tau^s(r)$ for each group, for instance the hour within a day.

Generally, for treatment effects to be causal, two standard identification assumptions must hold: (1) the unconfoundedness assumption and (2) the common support (overlap) assumption – see, for example, Hill (2011). The unconfoundedness assumption requires that potential outcomes are (conditionally) independent of the treatment assignment, conditional on a vector of confounding covariates $\boldsymbol{z}_t$. The overlap assumption requires that the probability of assignment to treatment is strictly between zero and one, implying that there is the possibility of being assigned to both the treatment and the control group. This is ensured when the distributions of the covariates overlap across regimes. While the unconfoundedness assumption is not testable, Figure 5.12 of the appendix provides empirical evidence for the validity of the overlap assumption.

5.5 Results

Figure 5.8 indicates that our empirical approach, for which we have used the `R` package BART (McCulloch, Sparapani, Gramacy, Spanbauer, & Pratola, 2018), is highly appropriate: the within-sample predictions of electricity spot prices on the basis of BART are dramatically better than those based on a ARMA(24,0)-GARCH(1,1) time series model[5], most notably with respect to extremely high prices and negative price spikes (see Figure 5.8). Note that the time series model conditions both on covariates and observed values of the electricity price, while the BART model only conditions on covariates.

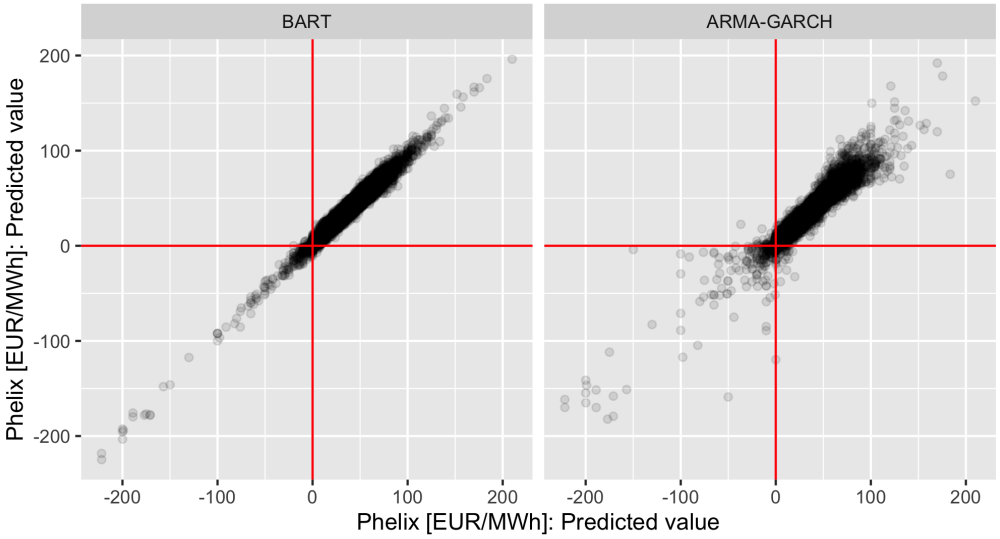
This comparison highlights the value added of using BART for our analysis, where the primary interest is on the occurrence of negative prices. Along with the quite high fit of $R^2 = 0.94$, this prediction performance is all the more remarkable, as the share of hours with negative prices of less than 1% is low in our database.

Given the time trends appearing from Figure 5.4, in what follows, we eliminate such trends by subtracting a smooth time trend $\widehat{f(t)}$, estimated using regression splines, from the observed values w_t: $\tilde{w}_t = w_t - \widehat{f(t)}$, where w stands for the dependent or any explanatory variable. Figure 5.13 of the appendix illustrates that the detrended covariates overlap much more than the original covariates (see Figure 5.12). Moreover, in accord with Figure 5.8, Figure 5.14 of the appendix demonstrates that the BART model with the detrended variables also outperforms a ARMA(24-1)-GARCH(1,1) time series model with respect to the within-sample prediction of negative prices.

Figure 5.9 presents further evidence on the high accuracy of BART predictions. The predicted number of hours with negative prices amounts to some 153 and 226 for Period 2 and 3, respectively, while the actual values amount to 146 and 220. The efficacy of the MPS with respect to avoiding negative prices is strongly

[5] The parameters of the AR-GARCH Model are chosen in analogy to Ketterer (2014)

Figure 5.8: Comparison of the Within-sample Predictions of Electricity Spot Prices of the Bayesian Additive Regression Tree (BART) Method with Static Predictions from a ARMA(24,0)-GARCH(1,1) Model



corroborated by the effect estimate of $\widehat{\tau^3(1)} = 561$ hours, which is accompanied by a relatively tight 95% posterior interval of [512, 609]. In other words, Germany's MPS helped to avoid 561 hours with negative prices. This corresponds to about 70% of the almost 790 hours with negative prices that would have occurred in Period 3 in the absence of the MPS – see the prediction for Regime 1 in the right panel of Figure 5.9.

Likewise, the left panel of Figure 5.9 illustrates that the introduction of the MPS in 2012 helped to reduce the number of hours with negative prices in Period 2 by $\widehat{\tau^2(1)} = 227$, with a posterior interval of [182, 266], given that the counterfactual number of hours with negative prices under Regime 1 is estimated at about 380 and the predicted number for Period 2 is 153. In other words, due to the introduction of the MPS, around half of the number of hours with negative prices could be avoided in Period 2. Lastly, the effect of reforming the MPS in August 2014 is estimated at $\widehat{\tau^3(2)} = 246$ hours in which negative prices could be avoided, with a posterior interval of [213, 279].

It also bears noting that the MPS helped to avoid extremely negative prices. For instance, Figure 5.10 indicates that the reform of the MPS was successful in avoiding about 200 negative prices of less than EUR -20 per MWh. In this case, the effect of the reform, denoted by $\xi^3(2) := \sum_t I\left[p_t^3(3) < -20\right] - \sum_t I\left[p_t^3(2) < -20\right]$, amounts to $\widehat{\xi^3(2)} = 196$ hours, with a posterior interval of [163, 228]. Accordingly, for the introduction of the MPS, we estimate an effect of $\widehat{\xi^2(1)} = 32$ hours, with [19, 35] being the posterior interval.

For illustration purposes, Figure 5.11 provides a detailed picture of the frequency of hours with avoided negative prices by hours of the day and across periods. This figure demonstrates that at 5 a.m., negative prices have been avoided about 70 times in Period 3 by the introduction and reform of the MPS (see the right panel of Figure 5.11). Furthermore, the panel on the left-hand side of Figure 5.11 suggests that if the MPS had not been implemented, Period 2

Figure 5.9: Number of Hours with Negative Prices in the Aftermath of the Introduction of the German Market Premium System in 2012 and its Reform of August 2014

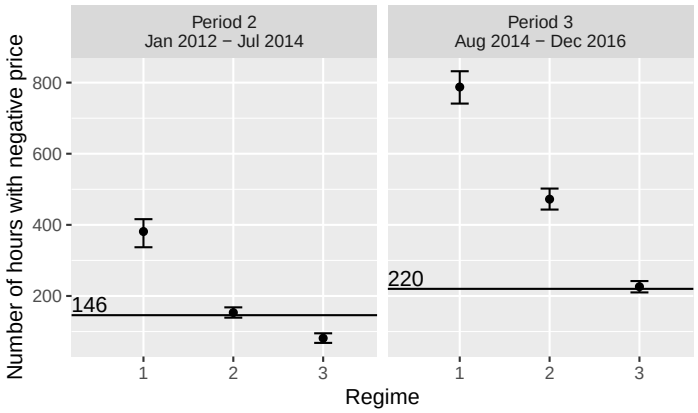
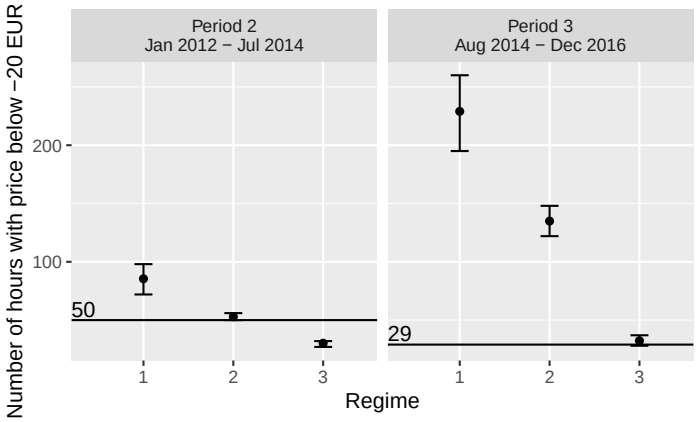


Figure 5.10: Number of Hours with Prices below EUR -20 per Megawatthour in the Aftermath of the Introduction of the German Market Premium System in 2012 and its Reform of August 2014



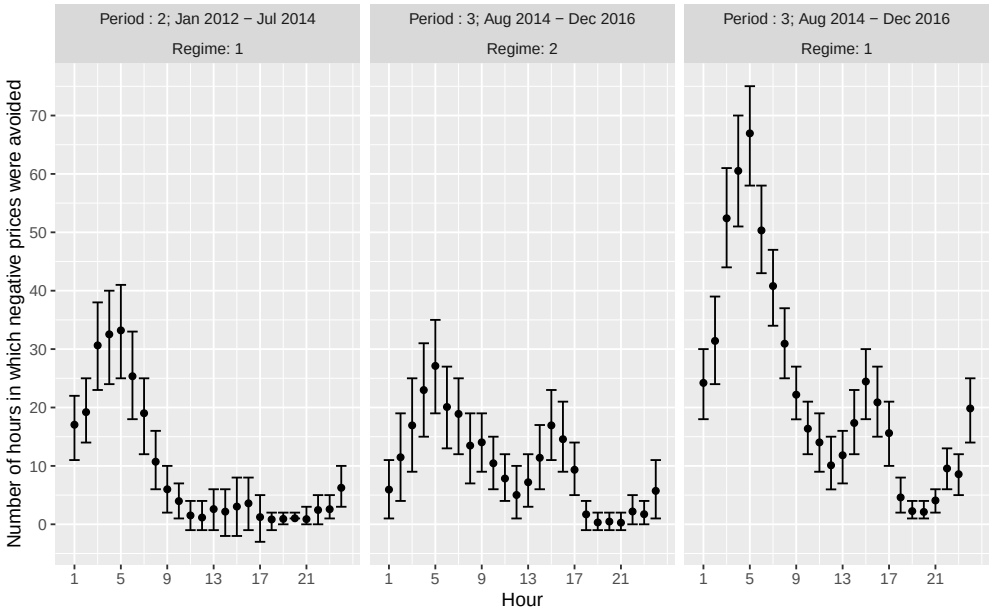
would have experienced around 30 additional days with negative prices at 5 a.m. The panel in the middle of Figure 5.11 indicates that the reform of the MPS helped to avoid 27 additional hours with negative prices at 5 a.m. This finding substantiates the above result that the MPS reform has been effective in further aligning the production of green electricity with market signals.

Finally, we conduct two robustness checks whose results are reported in the appendix. First, instead of detrending the variables by using regression splines as in our previous analysis, we take first differences. Instead of changes in the frequency of negative prices, the results from this exercise, displayed in Figure 5.15, reflect the difference in short-term price changes. These results confirm our findings displayed in Figure 5.11: The differences in short-term prices changes are among the highest in the early hours of a day. As this is the period when most negative prices occur (Figure 5.5), we conclude that the MPS was successful in avoiding negative prices.

Second, we take into account an important market coupling event, which may have had a bearing on the occurrence of negative electricity prices: the integration of Germany into the European electricity market in November 2010.[6]

[6]The coupling of European electricity markets started in 2006 when Belgium, France, and

Figure 5.11: Number of Hours with Avoided Negative Prices, by Hour and Period



To investigate this issue, we split Period 1 into two subperiods: Period 1a (January 2009 – October 2010) and Period 1b (November 2010 – December 2011), as Germany was connected to the joint European electricity market in November 2010.

The results of this robustness check indicate that if the MPS had not been in operation in Period 2 (January 2012 – July 2014), negative prices would have occurred in about 450 hours, while the actual value amounts to 146 hours (lower left panel of Figure 5.16). Even though the point estimates of the number of negative prices is somewhat higher compared to Figure 5.9, the absence of differences between Period 1a and Period 1b indicates that this market coupling event had no impact on the frequency of negative prices.

In addition, if the reform of the MPS had not taken place, we would have observed about even more hours with negative prices in Period 3 (see lower right panel of 5.16). As our estimate from Figure 5.9 of around 800 hours with negative prices roughly lies in the middle of the two estimates for Regimes 1a and 1b, we conclude that Germany's integration into the European electricity market did not have a bearing on the occurrence of negative prices. This conclusion is substantiated by the evidence displayed in Figure 5.17, where we focus on highly negative prices.

5.6 Conclusion

To reduce greenhouse gas emissions, an overwhelming number of countries promote renewable energy technologies, most often, as in Germany, in the form of FITs. These tariffs are paid for each kWh of green electricity produced by renewable

the Netherlands connected their electricity markets. Germany and Luxembourg joined this common market area in November 2010. In the meantime, more countries were connected to the common electricity market, such that the coupled market area covers more than 20 countries, standing for about 90% of Europe's electricity consumption (Ringler, Keles, & Fichtner, 2017). The idea behind market coupling is to increase welfare by allowing electricity to flow from low cost areas to high cost areas, resulting in price convergence (Keppler, Phan, & Le Pen, 2016).

energy technologies irrespective of whether the demand for electricity is low. In terms of RES capacity expansion, Germany's FIT system, established in 2000, proved highly successful and is thus widely seen as a role model for the promotion of renewable energy technologies. Between 2000 and 2016, RES capacities increased ten-fold to reach 104 GW, thereby exceeding the capacity of conventional power plants in 2016 for the first time (BMWi, 2017). As a consequence of the increasing RES capacities, prices on the wholesale electricity market decreased, which is commonly referred to as merit-order effect (Praktiknjo & Erdmann, 2016). In times of low demand, such as on Sundays or on public holidays, the pressure on electricity prices due to a large supply of green electricity may be so strong that prices turn out to be even negative.

To counteract these adverse effects, numerous countries have implemented premium schemes with the aim of aligning green electricity generation with market signals. Under premium schemes, rather than getting fixed FITs (RES, 2018), operators of renewable plants are commonly paid a bonus on top of the wholesale electricity price, thereby providing incentives to reduce production in times of low electricity demand.

In this paper, using hourly day-ahead spot market prices from 2009 to 2016 and the nonparametric method of Bayesian Additive Regression Trees, we have analyzed the efficacy of the German MPS in terms of reducing the frequency of hours with negative prices. This method allows for the straightforward construction of counterfactual situations and predictions of hypothetical outcomes, such as the electricity prices in the absence of the MPS. Along with a very high predictive power, this constitutes a distinctive feature of this method (Hill, 2011).

Based on such counterfactual analyses, we find that the implementation of the MPS in Germany in 2012 was quite effective in reducing the prevalence of negative electricity prices. Altogether, the introduction of the MPS and its reform in August 2014 helped to avoid some 560 hours with negative prices in the period spanning January 2012 to December 2016, particularly in the morning hours. Without the MPS, negative prices would have occurred over about 790 hours, indicating that about 70% of potential hours with negative prices were avoided. Moreover, the MPS was successful in avoiding about 200 negative prices of less than EUR -20 per MWh. Given these results, we conclude that, compared to FIT systems with guaranteed payments, MPS are a more cost-effective measure to promote renewable energy technologies.

It bears noting, though, that a comprehensive cost-benefit analysis would also take into account the implementation and transaction costs that arise due to a MPS. Moreover, while it is beyond the scope of our analysis to provide any guidance on how to design optimal promotion systems, Andor and Voss (2016) suggest that if positive externalities arise from the installation of renewable capacities, capacity-based instruments, such as tax cuts, should be employed, whereas generation-based instruments, such as FITs, would be preferable when positive externalities arise from the generation of green electricity. Yet, although capacity-based subsidies are particularly suited to push new technologies, such promotion schemes have not been implemented in many countries so far. While this also holds true for Germany, the country has recently improved the cost efficiency of the promotion of renewable technologies by introducing a tendering scheme in 2017, in which participants bid on a fixed remuneration per kilowatthour of green electricity, but are allowed to gain additional profits by direct marketing.

5.A Appendix

Figure 5.12: Overlap in the Covariates across Time Periods

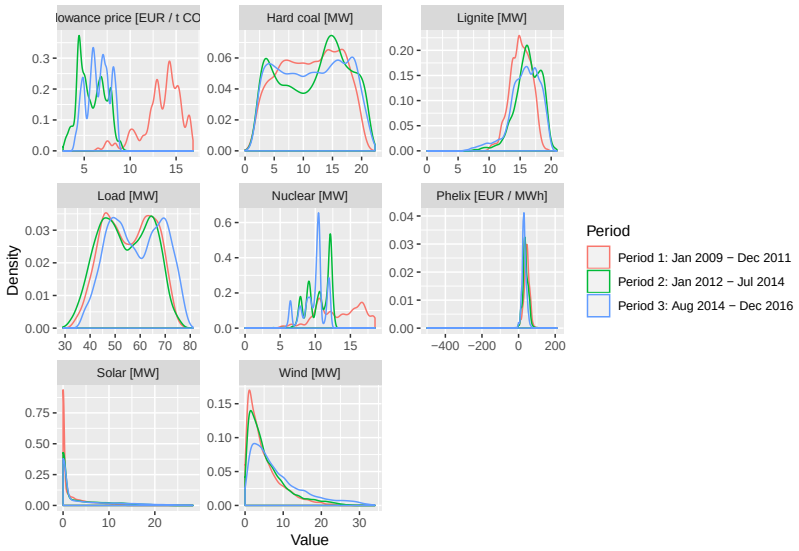


Figure 5.13: Overlap in the Detrended Covariates across Time Periods

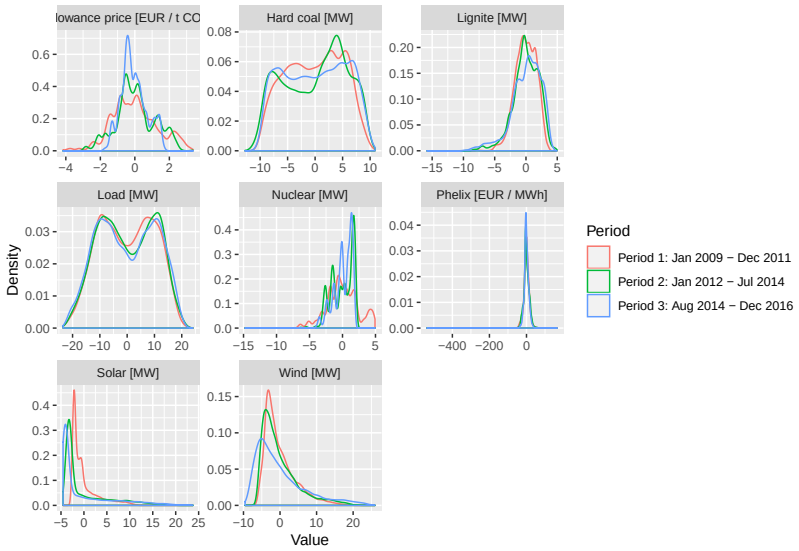


Figure 5.14: Comparison of the Within-sample Predictions of Electricity Spot Prices of the Bayesian Additive Regression Tree (BART) Method with Static Predictions from a AR(24)-GARCH(1,1) Model

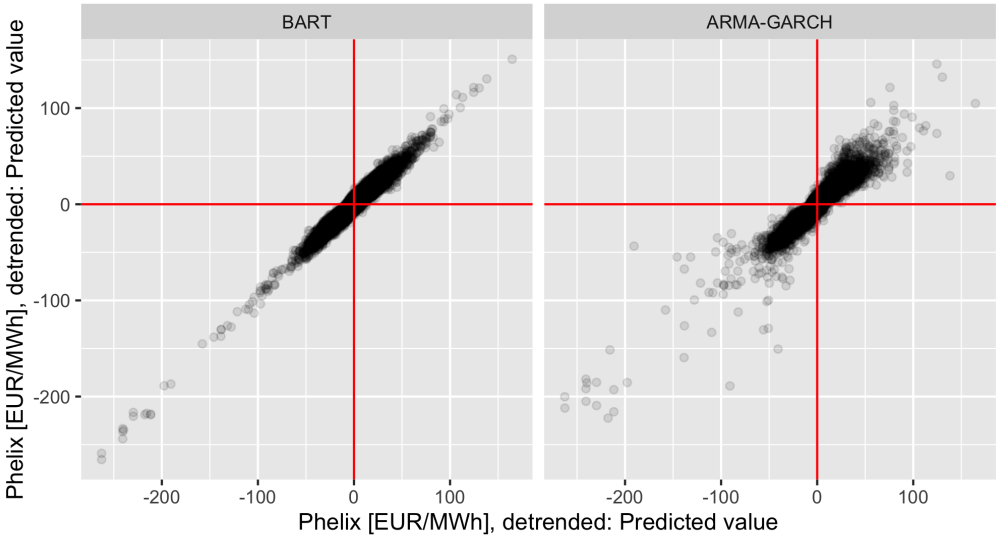


Figure 5.15: Difference in Price Changes, by Hour and Period using Differenced Variables

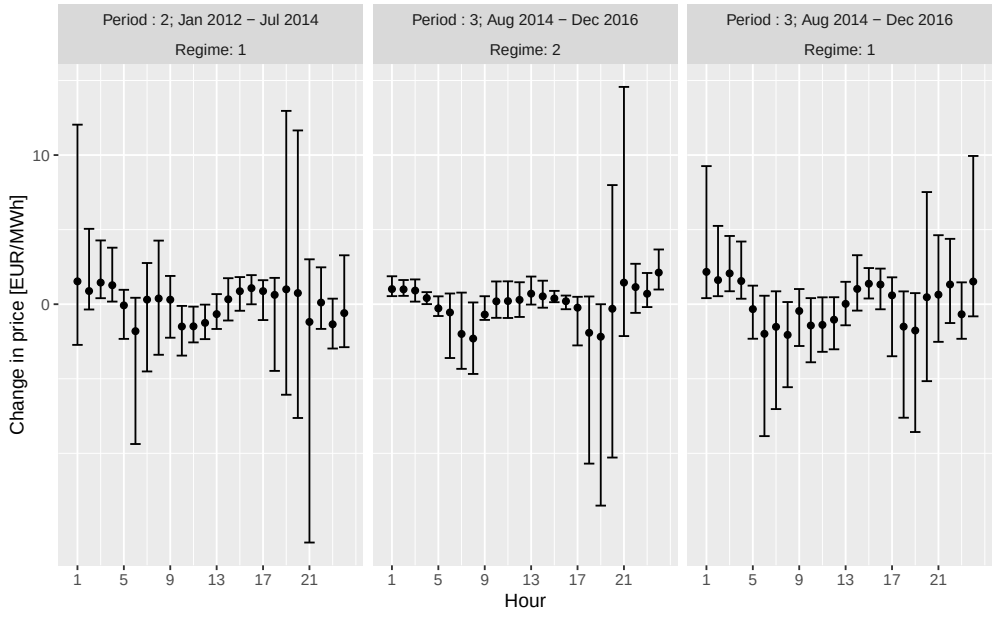


Figure 5.16: Number of Hours with Negative Prices, by Hour and Period – Market Coupling

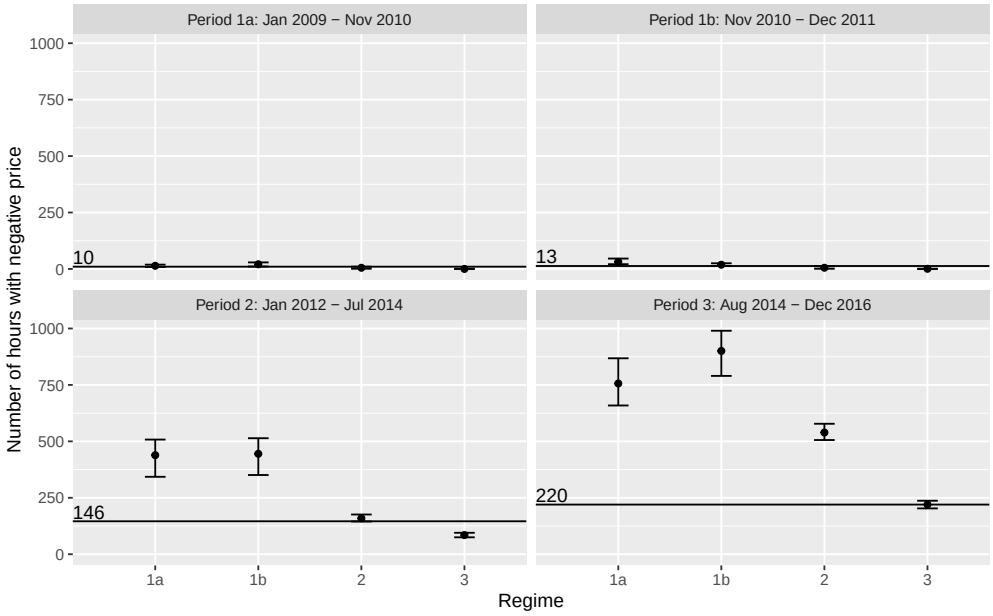
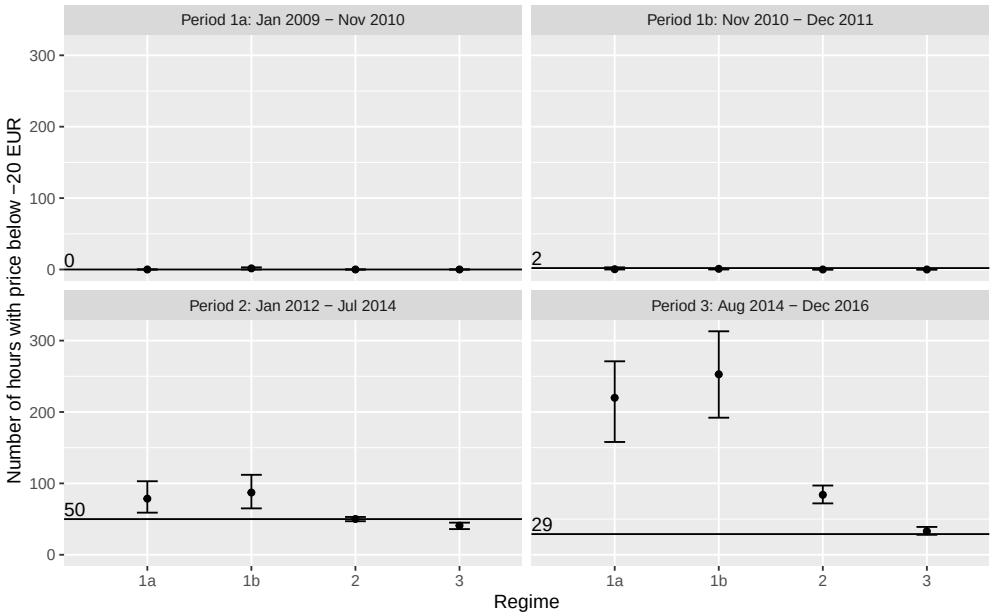


Figure 5.17: Number of Hours with Prices below EUR -20 per Megawatthour, by Hour and Period – Market Coupling



Chapter 6

Equal access to primary care: A benchmark for spatial allocation[1]

Abstract

Using a greedy algorithm together with very fine spatial data, we calculate an optimal allocation of general practitioners in Germany. This benchmark allows us to not only identify regions experiencing over- and undersupply but also to propose how general practitioners should locate to relax existing uneven regional distribution. Our results suggest that overall there were 6% too few GPs in Germany to provide equal primary care for every person in 2019 and that more GPs should be placed in rural regions. However, we do not observe that residents of rural regions experience stronger undersupply than residents of urban regions.

Keywords: Optimization, Regional data, Health, Medical access, Redistribution

[1]This paper is jointly written with Alexander Haering and Anna Werbeck.

6.1 Introduction

Equal access to health care is fundamental for a reliable health care system. Yet, uneven geographic distribution of physicians causes concern across the member states of the Organisation for Economic Co-operation and Development (OECD) and, although it has been in the centre of health care discussion for many decades, it still remains the single most commonly named issue in the health sector (OECD, 2014). Various policies have been installed to tackle this problem. They aim (i) at medical students to recruit new physicians to serve in undersupplied areas, (ii) at current physicians to redistribute the physician workforce by financial incentives or regulations, or (iii) at an integration of innovative forms of care (OECD, 2014). To clearly evaluate these policies, we need a tool proposing efficient allocations of physicians.

In this paper, we calculate a benchmark of smallest quantity and optimal allocation of general practitioners (GPs) for Germany in 2019. We first calculate the demand for GP visits on a small regional scale based on age and gender. We then optimize the number of GPs by allocation under the two constraints that each resident has access to at least one GP within reasonable driving time and that each GP has limited capacity of patient contacts. Because there usually too few GPs overall, or the GPs are not optimally distributed, these constraints are usually not satisfied in practice.

Germany is particularly suited as a first testing ground. Equal access to health care is manifested in the very first paragraph in Book I of the Social Code (SGB I, *Sozialgesetzbuch I*). It determines that all social rights, including health care, apply uniformly to all residents. In the German health care system, general practitioners[2] constitute the centre of the ambulatory care sector (§73 SGB V) and are generally the first medical point of reference for patients. Therefore, equal access to health care requires, besides quality, a sufficient number of general practitioners combined with an adequate distribution among different regions. However, the problem cannot be solved by massively increasing the number of GPs, as Book V of the Social Code also demands cost-effectiveness of health care services (§12 SGB V). Incorporating these legislative demands, we use guidelines and data on the current workload of GPs to determine the area and demand a single physician can cover. Using these constraints, we formulate our German-specific optimization problem as follows: What is the optimal number and regional allocation of general practitioners such that each resident can reach a GP within 15 minutes driving time (Federal Institute for Research on Building, Urban Affairs and Spatial Development, 2005) and that GPs serve within their limited capacity of $13,000$ annual cases?[3]

The political discussion in Germany is centered on two major issues. First, an overall shortage of general practitioners and second, an especially severe undersupply in rural regions. Our benchmark allocation suggests that there were indeed 6% too few GPs in Germany in 2019. Also, we find uneven geographic distribution of GPs that is more pronounced in less densely populated areas. However, the maldistribution does not translate into an especially severe under-

[2] In this study, we follow the specification of general pracitioners as specified in §73 SGB V, excludinig paediatricians. We include general physicians, internists without a specialization who have chosen to participate in GP care, doctors who are listed in the register of doctors in accordance with §95a IV/V 1 and those who were participating in GP care up to December 31st, 2000.

[3] We obtained this number from (National Associations of Statutory Health Insurance Physicians' & Verband der niedergelassenen Ärzte Deutschlands, 2018).

supply of rural dwellers' health care demand. Put differently, even though our benchmark solution would place more GPs in rural regions, we do not see that residents of less densely populated areas experience structurally more undersupply than those in more densely populated ones. However, there is a general level of undersupply which might be relaxed with a more efficient GP-allocation.

We bridge the gap between two main strands of literature: operations research and regional variation in health care. We do so by generating a benchmark allocation for GPs in Germany using the methodology developed in the operations research literature to inform our view on regional differences in health care access.

Researchers in the field of operations research in health care developed an algorithmic toolkit for location-allocation problems (Daskin & Dean, 2005) and applied it to several health care facilities (see Rais & Viana, 2011, for a well-structured overview). Related to our research, some studies calculate the optimal allocation of health care facilities to maximize accessibility under different constraints. Marianov and Taborga (2001), for example, set up an allocation plan for health-care centers maximizing coverage for a low-income population within pre-specified distance. Harper, Shahani, Gallagher, and Bowie (2005) develop a simulation model accounting for patients traveling to the providers, resource capacities, heterogeneous patient needs and different transport modes. We contribute to this research field by presenting a benchmark for the optimal allocation of GP practices in Germany, accounting for local geography, resource capacities of GPs, and heterogenous patient needs. In addition, we account for correlations between regional indicators and health care supply and demand.

In health economics, several studies investigate regional disparities in health care and their sources (see, e.g., Finkelstein, Gentzkow, and Williams, 2016; Johansson, Jakobsson, and Svensson, 2018; Salm and Wübker, 2020). Other studies develop solutions to this global problem. Sutton and Lock (2000), for example, derive a resource allocation formula incorporating regional variation in health care using data from 15 regions in Scotland. Others evaluate policies that have been implemented to tackle the challenge of undersupply of general health care (see, e.g., Zhang, Baik, Fendrick, & Baicker, 2012). Yet, what is missing is a reference for an optimal number of GPs and their practice locations ensuring everyone's health care demand is satisfied on a small-scale. Without a benchmark it remains unclear whether or not the resultant quantity and practice locations of GPs meet the respective needs sufficiently and efficiently. Our contribution is to fill this gap and propose such a benchmark for Germany.

The rest of the paper is structured as follows: the next section describes our data, followed by the methodology. In the fourth section, we present our findings, section five shows sensitivity analyses, and section six concludes.

6.2 Data

Driving times

Our computed driving times are based on grid cells from the RWI-GEO-GRID (Breidenbach & Eilers, 2018). The RWI-GEO-GRID database contains socioeconomic data for Germany on 218,875 populated 1km^2 grid cells.[4] We determine driving time by car from each populated grid cell to all populated grid cells within 40 km radius using the Open Source Routing Machine program (Luxen & Vetter, 2011), based on OpenStreetMap (Haklay & Weber, 2008) road data,

[4]The grid cells are aligned to the EU-wide INSPIRE Directive (Bartha & Kocsis, 2011).

which contains nearly the complete German road network (Barrington-Leigh & Millard-Ball, 2017). A radius of 40 km ensures that we include all grid cells that can be reached within 15 minutes.

This procedure leaves us with around 600,000,000 driving times. Next, we discard driving times exceeding 15 minutes to avoid unnecessary calculations and because Federal Institute for Research on Building, Urban Affairs and Spatial Development (2005) demands that every person reaches one GP within a maximum of 15 minutes. This results in around 170,000,000 driving times.

Current GP Allocation: the Status Quo

The Regional Associations of Statutory Health Insurance Physicians' (RASHIP) contains information on all contract physicians and contract psychotherapists in their respective region. NASHIP is their national umbrella organization. RASHIP and NASHIP provide aggregated numbers of GPs on federal state levels. However, for our study we need the exact practice location. Therefore, we utilize data from the official on-line search engines of both organizations.[5] The search engines are accessible to the German population to search for doctors in all fields. Yet, physicians have to agree to be listed in the NASHIP and/or RASHIP online database. This potential self-selection into the search-engines may lead to two problems: Either, when focusing only on one database, we are left with an undercount as physicians that do not agree to be listed in the chosen search-engine are not part of our set; or, when using the two databases, we are left with an overcount when a physician is listed in both. To avoid these problems, we used both data sources. In addition, we also created a third data-source, consisting of the intersection of both data sets. To verify our approach, we aggregated our GP-count on federal state level and compared it to the official number provided by NASHIP (2018). For every federal state, we chose the data source with the GP-count closest to the official number. Table 6.1 summarizes the number of GPs obtained from the three sources. In total, there are 55,006 GPs in the data set, representing the status-quo allocation.

Table 6.1: Number of GPs by Source

Data source	N
NASHIP	33,031
RASHIP	15,140
Intersection	6,835
Total	55,006

Note: State of data as of 2019

GP visits

Furthermore, we use data from the SOEP (Goebel, Grabka, Liebig, Kroh, Richter, Schröder, & Schupp, 2019), which is the largest representative longitudinal panel of private households in Germany. We utilize wave 2016 because, besides age and gender, it contains self-reported number of annual visits to the doctor. This

[5] See appendix 6.A for all 18 webpages.

allows us to calculate the expected number of cases in each grid cell. Our resulting data set contains 24,418 observations. Table 6.2 shows the mean and standard deviation for our variables of interest.

Table 6.2: Summary Statistics Regression Tree

	Mean	SD
Age	48.663	17.617
Female	0.542	0.498
# visits doctor	9.092	14.624

Source: SOEP wave 2016 (Goebel, Grabka, Liebig, Kroh, Richter, Schröder, & Schupp, 2019)

Characteristics of municipalities

To examine whether a shortage of GPs only appears in less densely populated regions and which regional patterns are correlated with over- or undersupply, we use additional municipality characteristics.[6]

For ease of interpretation, we generate dummy variables for rural and suburban areas following the classification of Dijkstra and Poelman (2014) to assort all German municipalities with respect to their population density. Sensitivity analyses (see section 6.5) show that this aggregation of grids does not influence our results. We calculate a *commuter index* by

$$\frac{(\text{in commuter} - \text{out commuter})}{\text{population}} \tag{6.1}$$

using data from the German Federal Agency of Employment (BA, 2008). We also calculate a *migration index* in the same way:

$$\frac{(\text{in migration} - \text{out migration})}{\text{population}}. \tag{6.2}$$

Furthermore, we add employment-rate, self-employment rate and the business tax from (Federal Office for Building and Regional Planning, 2020) as measures of economic power.[7]

In addition, we use house and rent prices, as derived from the FDZ data set (see Boelmann & Schaffner, 2018, for a description). Table 6.3 summarizes the mean and standard deviation of our variables.

[6] Because the area of municipalities may vary in Germany over time, e.g. due to mergers, and our data sets are not from the exact date in 2019, there is some mismatch and we do not have information on all 10,848 municipalities.

[7] Employment rate and business tax are only available on the level of municipality association, self-employment rate is on district level. Since municipalities are sharply defined in both levels of aggregation, it enables us to merge the respective value of employment rate and business tax on municipality level.

Table 6.3: Municipality Characteristics

	Mean	SD
Rural	0.757	0.429
Suburban	0.240	0.427
Urban	0.004	0.060
Business tax (in € per resident)	414.637	744.131
Commuter index	0.086	0.152
Employment rate (in employees per 100 potential workers)	61.141	5.546
Migration index	0.00	0.01
House price (in € per km^2)	1781.441	855.253
Rent price (in € per km^2)	6.811	1.548
Self-employment rate (in self-employed per 100 workers)	115.053	17.362

Note: Rural, suburban and urban are dummy variables.

6.3 Methodology

Set cover problem

Our unique combination of data sets helps us to determine the optimal number and regional allocation of GPs in Germany under the following constraints: All residents can reach at least one GP within 15 minutes driving time and GPs serve within their limited capacity of 13,000 annual cases. That is, we face a capacitated set cover problem. The set cover problem is a central problem in combinatorics with wide ranging applications, see, as examples, Vazirani (2001) or Korte and Vygen (2018).

Initially, we ignore the limited capacity. We set up a general uncapacitated set cover problem, where one determines the smallest sub-collection from a collection of sets $\mathcal{S} = \{S_1, S_2, \ldots, S_l\}$ that covers a universe of elements $\mathcal{U}$, i.e., every element in $\mathcal{U}$ appears in at least one of the subsets of $\mathcal{S}$. Here, the universe is $\mathcal{U} = \{g_1, \ldots, g_n\}$, where g_i is the ith 1km^2 grid. We define S_i as

$$S_i \coloneqq \{g_j \mid d(g_j, g_i) \leq 15\text{min}\},$$

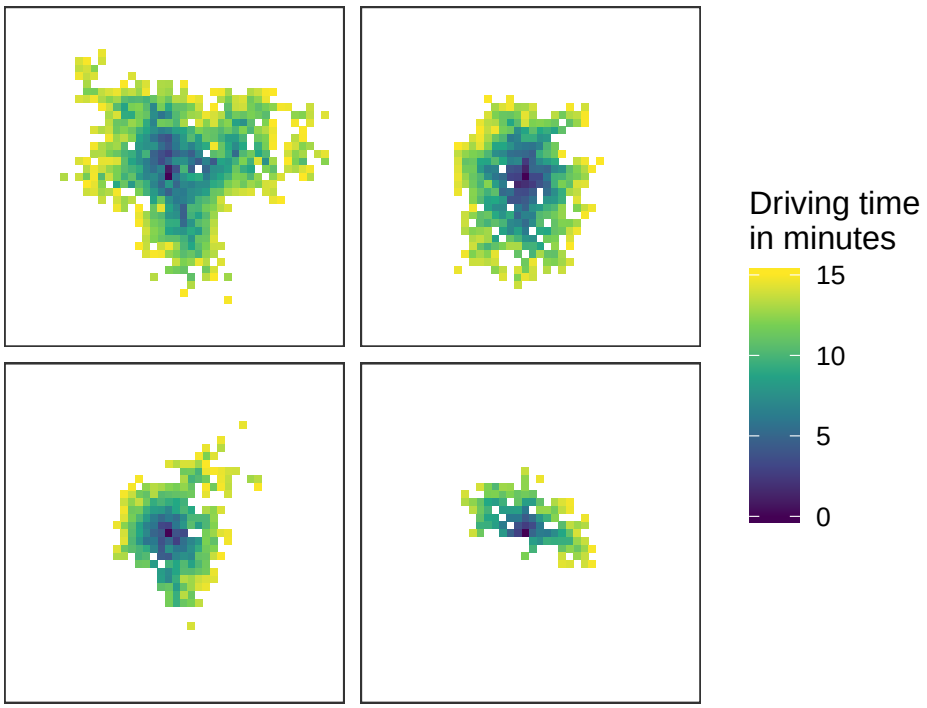
where $d(g_i, g_j) \geq 0$ is the asymmetric driving time from g_i to g_j. This implies that S_i consists of all grid cells from where grid g_i is reachable within 15 minutes. As $d(g_i, g_i) = 0$, S_i always includes g_i.

Figure 6.1 shows four exemplary sets from our data. In all four figures, the dark blue grid cell at the center depicts g_i, the potential location of a general practitioner. The differently colored cells in the surrounding show all grid cells for which g_i is reachable within 15 minutes. We see that the shapes of the sets vary. This is driven by street infrastructure, classified by 15 speed types and geographical conditions.

In the capacitated case, we restrict the demand to be covered by a single set to be below 13,000 patient-physician contacts, i.e., cases, per year, so that each set S_i consists of the grids containing the cases serviced by the ith GP. This might require placing multiple GPs in one grid to satisfy the demand.[8]

[8]Because we operate within a resolution of 1km^2, we do not differentiate between GPs working in a joint practice and those who do not.

Figure 6.1: Grids from where the Center Grid can be reached within 15 Minutes.



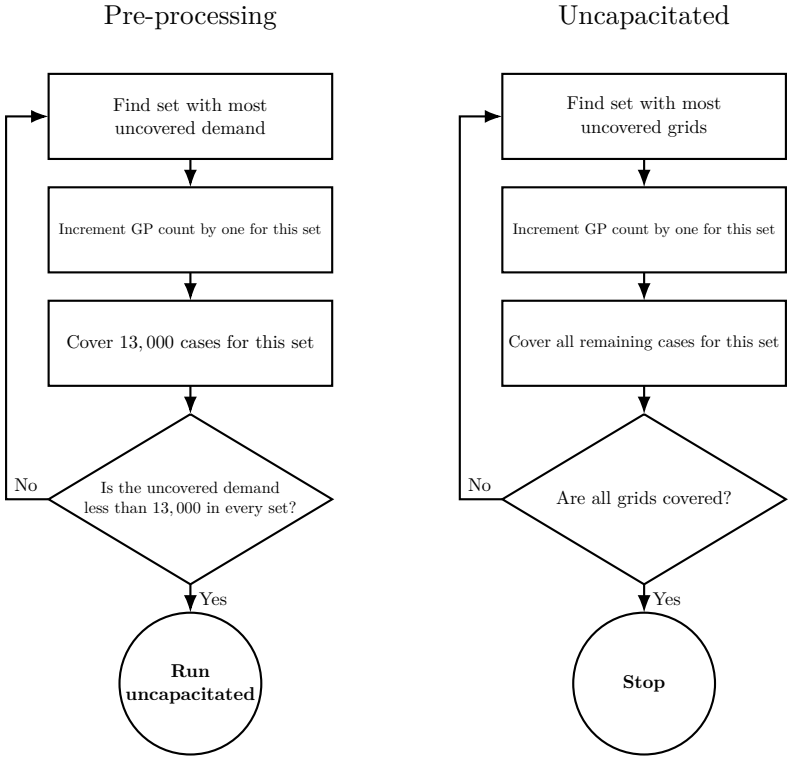
Greedy approximation

Because the set of possible solutions grows exponentially with the number of sets in $\mathcal{S}$, the problem becomes too complex to be solved exactly. Therefore, we use the capacitated variant of the greedy approximation algorithm (Chvatal, 1979; Lovász, 1975). Feige (1998) shows that the basic greedy approximation gives reasonably good approximations to the true solution while scaling well to large data sets. Vazirani (2001, chap. 29) shows why the approximation is very hard to improve.

The basic idea of our greedy approximation is to iteratively distribute doctors across the grid. Given that in Germany most GPs serve in private practices with not more than a few general practitioners working together, we introduce a penalty to avoid placing more than 10 GPs per grid. We start from a green field without any doctors. In every iteration, we choose the set with the most uncovered demand, until every case is covered. Here, covering demand means associating cases with a GP. Parka and Honga (2009) add a pre-processing step to the greedy approximation for the capacitated case, where one iteratively covers demand (here, $13,000$ cases), until no sets exceeding the capacity remain. Figure 6.2 gives a short illustration of the two algorithms.

For our implementation of the uncapacitated greedy algorithm in R we build on the *RcppGreedySetCover* package (Kaeding, 2018). For all pre-processing steps in case of the capacitated set cover algorithm, we create a fast `C++` implementation.

Figure 6.2: Set Cover Algorithm Uncapacitated



Feeding the greedy algorithm: calculating the number of cases

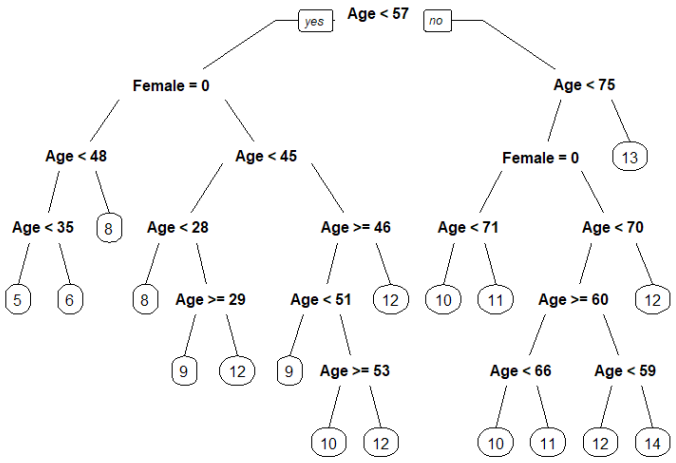
To calculate the adequate distribution of general practitioners, we account for the expected number of cases in each grid cell. The SOEP 2016 data on number of visits to the doctor's office provides us with a proxy for medical demand. We assume patient subgroups with an approximately constant number of visits. As such, we use a regression tree to forecast demand, which is suitable for identifying subgroups. We calculate a regression tree with age and gender as explanatory variables to forecast demand (see Table 6.2). While additional variables might improve prediction, we can only use those variables which are available in both the SOEP 2016 data and the RWI-GEO-GRID. However, we assume that age and gender are the most relevant variables and capture health demand fairly well.

For our estimation, we use the R package *rpart* (Therneau & Atkinson, 2019). Figure 6.3 shows our final tree[9]. The interpretation is straightforward. For example, a female subject ($Female = 0$, right branch) $Age < 28$ is expected to visit the doctor's office 8 times a year.[10] We then use our regression tree to compute the expected number of GP visits for each 1km^2 grid cell using the demographic data provided by the RWI-GEO-GRID database. As an example, the city of Essen has a population of some 0.6 million persons with an approximate demand of 5 million cases per year.

[9] Because we only have two exogenous variables, we use a simple model.

[10] Interestingly: at $Age > 75$ the number of expected visits to the doctor is independent of the gender with 13 visits a year.

Figure 6.3: Regression Tree Number of Visits to the Doctor per Year



Note: Calculation based on SOEP wave 2016, endogenous variable is self-reported number of visits to the doctor per year, $R^2 = 0.023$.

Regressions

To examine whether our variables of interest, excess GPs and uncovered demand, are correlated with municipality characteristics, e.g. population density, we run additional regressions. We aggregate our grid data on municipality level, assigning each grid to the municipality with which it has the largest intersecting area.[11] In addition, we look at a commuter index as well as a migration index for municipalities. First, we define our dependent variable as

$$d_i := N_i^{observed} - N_i^{greedy}, \tag{6.3}$$

i.e. the difference between $N_i^{observed}$, the observed number of GPs in the ith municipality, and N_i^{greedy} , the number of GPs needed to achieve full support in the ith municipality. Therefore, we can interpret d_i as residual, measuring excess number of GPs in a municipality. On these terms, an optimal allocation of GPs would result in d_i equal to zero for all i and our coefficients should not pick up correlation with the excess number of GPs. Figure 6.4 shows a map of the dependent variable.

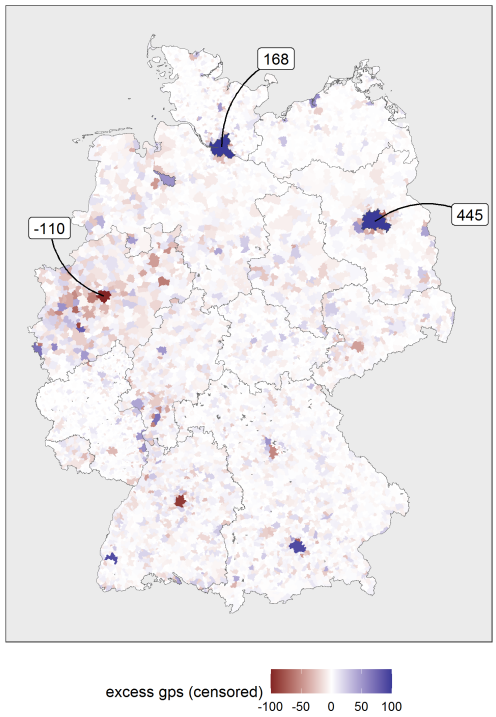
We use d_i from (6.3) and specify our regression as:

$$\begin{aligned} d_i = \alpha_0 + \alpha_1 rural_i + \alpha_2 suburban_i + \alpha_3 commuter_i + \\ \alpha_4 migration_i + \boldsymbol{\beta}^\top \boldsymbol{econpower}_i + f_{\text{spatial}}(\boldsymbol{s}_i) + \boldsymbol{\gamma}^\top \boldsymbol{x}_i + \epsilon_i. \end{aligned} \tag{6.4}$$

In equation (6.4), $rural_i$ and $suburban_i$ are dummy variables indicating rural or suburban municipalities, where *urban* is the reference category. $\alpha_0, \ldots, \alpha_4$ are regression coefficients, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients. The variable $commuter_i$ represents our commuter index, $migration_i$ the migration index. $\boldsymbol{econpower}_i$ incorporates the variables unemployment rate, self-employment rate and business tax that we report separately. Depending on our model specification, we add additional controls: $f_{\text{spatial}}(\boldsymbol{s}_i)$ is a spatial effect accounting for regional dependencies and vector $\boldsymbol{x}_i$ contains rent and housing prices of

[11] In Section 6.5 we provide evidence that our results are not influenced by this aggregation.

Figure 6.4: Excess GPs in Germany by municipality



Note: Because outliers would otherwise dominate the color scale, values outside the interval $[-100, 100]$ are shown via annotation.

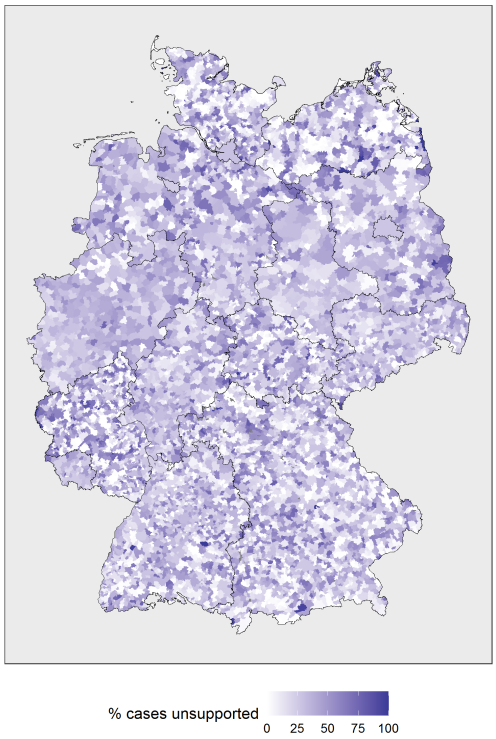
municipality i. We additionally add dummies for the federal states.[12] The smooth spatial effect f_{spatial} accounts for spatial dependence across neighboring municipalities unexplained by the covariates. We model f_{spatial} via thin-plate regression splines (cf. Wood, 2003, for further details).

We estimate the model using the R package *mgcv* (Wood, 2011). The fixed effects on federal state level account for heterogeneity in the institutional framework of Germany, for instance varying incentives for GPs to practice in smaller towns. All metric independent variables are scaled by one standard deviation to achieve effect comparability.

We shall also investigate the German population generating health care demand by running additional regressions to study regional patterns correlated with uncovered cases. We again focus on the three classifications of municipalities with regards to population density as our main variable and a commuter as well as a migration index. To this end, we create an index measuring the percentage of demand which, given our constraints, cannot be covered, i.e., maximal drivetime of 15 minutes and a GP-capacity of 13,000 cases per year. Notably, these uncovered cases only exist in case of the status quo, in the benchmark situation all health care demand is covered perfectly. We assume that each person frequents the closest GP. Then we classify each case as uncovered if it exceeds the capacity of the providing GP. We define our index as the percentage of uncovered cases in

[12] We add two dummies for North Rhine-Westphalia, because it is divided into two RASHIPs. In addition, we use a single dummy for all city-states, which only have one municipality each. Therefore, we add 15 federal state dummies to our regression analysis.

Figure 6.5: Uncovered cases in Germany by municipality



each municipality. This is an approximation because of the following two reasons: (i) In reality, GPs probably still provide care when above capacity. (ii) Persons might switch to a GP further away who is not overcapacitated. However, these limitations mainly concern the overall level of undersupply, while our index allows us to compare the level of potential undersupply across regions.

More precisely, we assume the number of cases a GP covers in a grid g is proportional to the overall demand, i.e.,

$$\dot{c}_k = \frac{\lambda c_k}{\sum_j I[\text{grids } j \text{ and } k \text{ share same GP}]c_j}, \tag{6.5}$$

where $\dot{c}_k$ denotes number of cases covered, i.e., treated patients, in grid k, λ is the capacity all GPs providing care for this grid have. c_k is the number of cases, i.e., sick patients, in grid k. We define the dependent variable for every municipality i as the percentage of uncovered cases by

$$u_i := 100 \frac{\sum_{j \in i} I[c_j > \dot{c}_j](c_j - \dot{c}_j)}{\sum_{j \in i} c_j}. \tag{6.6}$$

This gives us the percentage of uncovered cases, i.e., unsatisfied health care demand generated by each municipality. As an illustration, assume that municipality i has a demand of $20,000$ cases, and a single GP that is closest to all patients in i. Assume further that the GP does not cover any cases of patients from other municipalities. Then, this municipality i is left with $7,000$ uncovered cases. This equals a share of $100 \times \frac{7{,}000}{20{,}000} = 35\%$. Figure 6.5 shows a map of variable u. Note that our approach assumes that each patient chooses her nearest GP within 15 minutes driving time. Therefore, it is not necessary that the GP

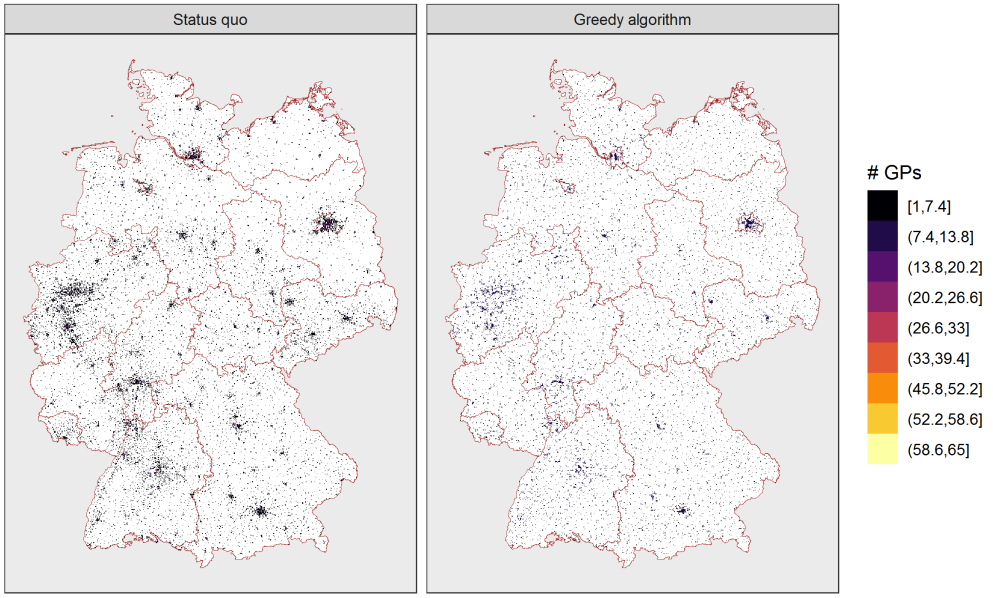
practice is located in the patient's municipality. For our second regression, we use the same controls as in (6.4) and u as the dependent variable.

6.4 Results

This section discusses our results. The current political discussion is centered on two major concerns: First, an overall shortage of general practitioners, and second, an especially severe undersupply in rural regions. We now investigate these claims through the lens of our analysis.

Figure 6.6 shows the status-quo allocation compared to our benchmark allocation on a grid level. The left panel shows the distribution of general practitioners in Germany as of 2019. The overall number of active general practitioners was $55,006$. As expected, densely populated areas like the city-states of Berlin and Hamburg as well as other large regions like Munich or the Ruhr area have a high density of GPs. The right hand panel of Figure 6.6 shows our benchmark distribution of GPs in Germany. We determine the optimal number of GPs needed in 2019 to be $58,144$. This is fairly similar to that of the actual distribution. Remarkably, the optimal number of GPs is fairly close to the actual number of active GPs in 2019. We see a shortage of $3,138$ GPs ($\approx 6\%$). Still, we take this difference as noteworthy. In our benchmark, GPs are more evenly spread out across Germany, resulting in complete coverage. The distributions differ mainly in densely populated areas such as Munich or the Ruhr area. Here, our benchmark solution places fewer GPs than there are in the status quo. Furthermore, more GPs are located in the city-centers, which are easy to reach [13].

Figure 6.6: Distribution of General Practitioners in Germany



Comparison of GPs allocation under status-quo and our benchmark allocation - grid level.

[13] It should be noted that our model does not account for the costs of the location, so that it might be more cost-efficient to place GPs elsewhere even though this would mean that more GPs are needed.

We now turn to the claim that Germany faces an especially severe undersupply in rural regions. We start off by examining whether a shortage of GPs only appears in a small number of cities or regions, or if it is a broader issue. For this analysis, we consider the excess number of GPs in a municipality i calculated by $N_i^{observed} - N_i^{greedy}$ (see equation (6.3)). Thereby, we take the allocation of GP practices suggested by the greedy algorithm as the benchmark. Compared to this optimal solution, we assess whether or not there are too many or too few GP practices in one region.

Table 6.4 summarizes the excess number of GPs in German municipalities. It reports the total number as well as numbers split into three categories as suggested by EUROSTAT (see Dijkstra & Poelman, 2014, for further details): cities (densely populated), towns and suburbs (intermediate density) and rural areas (thinly populated). This gives us the opportunity to gain more insights regarding the claim of undersupply being primarily a problem of rural regions.

The overall picture reveals that of the $11,004$ municipalities $3,688$ face an excess of GPs compared to the benchmark allocation, $3,084$ experience a shortage of GPs, and $4,232$ are served optimally.

When split into the three regional categories, the data show that of the 125 densely populated areas, 77 face an excess of general practitioners, while 45 face a shortage of GPs. Of the intermediately populated regions $1,485$ have an excess of GPs, 864 a shortage. And in the low populated regions $2,126$ face an excess, $2,175$ a shortage of GPs and in $3,953$ rural regions the benchmark solution would not change the number of GPs from the current situation. Considering relative numbers, 36% of cities, 33% of towns and suburbs and only 25% of rural areas face an undersupply of GPs. This does not show a clear-cut picture of the situation. We therefore regress our calculated excess number of GPs on the three region categories.

Table 6.4: Municipality with excess/shortage of GPs by Population Density

Population Density	Number of municipalities			
	Excess GPs	Shortage GPs	Optimal GPs	Total
Cities	77	45	3	125
	(66%)	(36%)	(2%)	(100%)
Towns/Suburbs	1,485	864	276	2,625
	(57%)	(33%)	(10%)	(100%)
Rural	2,126	2,175	3,953	8,254
	(26%)	(26%)	(48%)	(100%)
Total	3,688	3,084	4,232	11,004
	(34%)	(28%)	(68%)	(100%)

Notes: Share in parenthesis, classification based on Dijkstra and Poelman (2014)

Table 6.5 summarizes our regression results for excess number of GPs (see equation (6.4)). In specification (1), we run a simple OLS regression with dummies for regional categories as explanatory variables. In (2), we additionally include the commuter and migration index. In (3) we add the three measures for economic power. Then, starting in (4), we include our smooth spatial effect to account for partial dependence across neighboring municipalities. In specification (5), we add federal state fixed effects. Lastly, in (6) we include rent and house prices as

controls for costs of living.

In all specifications, we observe a stable negative effect of suburban areas compared to cities as our baseline. This effect is even more pronounced when looking at rural areas ($p < 0.001$, two-sided χ^2 test). This suggests that there is indeed a disparity of rural and suburban areas after controlling for other factors in terms of GP practices. Furthermore, the results show a negative correlation between excess number of GPs and a region's commuter index ($p < 0.001$), implying that regions with higher in-commuting have fewer excess GPs. The picture is reversed for the migration index. We find a positive correlation, indicating that regions with higher in-migration have more excess GPs. In terms of our measures of economic power the results are mixed. Business taxes ($p < 0.001$) and employment rate ($p < 0.050$) are negatively correlated with excess number of GPs, whereas self-employment rate ($p > 0.050$) does not correlate with excess number of GPs. The results suggests that there might be other factors driving the uneven geographic distribution of GPs.

Let us take a step back. If we focus solely on the location of GP practices, we miss the actual question of concern: does a structural lack of rural GPs lead to structural failure to satisfy health care demand of rural dwellers' health care demand? A crucial point why this question may not be addressed by looking only at the location of GP practices: doctors do not necessarily provide care exclusively to the residents of the municipality in which their practice is located.

To answer this question, we start by looking at what proportion of health care demand of residents in one specific municipality remains unsatisfied. Flipping this indicator around, this translates into the question: how much more would the GPs accountable for dwellers of one municipality have to work such that all health care demand is satisfied? This results in the proportion of unsatisfied health care demand being equal to zero, just like in case of our benchmark scenario.

We run a regression for the linkage between unsatisfied health care demand and population density as well as the commuter and migration indices and our economic power measures. Table 6.6 reports the regression results for the percentage of uncovered cases, i.e., unsatisfied health care demand (see equation (6.6)), with the same specifications as in our model for excess number of GPs.

Table 6.5: Regression: Excess Number of GPs and Municipality Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	21.62^{***}	21.24^{***}	20.87^{***}	23.29^{***}	8.80^{***}	10.62^{***}
	(1.27)	(1.28)	(1.63)	(1.81)	(1.61)	(1.67)
Rural	-22.31^{***}	-21.70^{***}	-21.96^{***}	-22.24^{***}	-7.27^{***}	-7.74^{***}
	(1.28)	(1.28)	(1.29)	(1.29)	(1.14)	(1.15)
Suburban	-21.31^{***}	-20.83^{***}	-20.98^{***}	-21.04^{***}	-5.92^{***}	-6.17^{***}
	(1.28)	(1.29)	(1.29)	(1.29)	(1.14)	(1.14)
Commuter index		-0.36^{***}	-0.39^{***}	-0.37^{***}	-0.36^{***}	-0.36^{***}
		(0.08)	(0.08)	(0.08)	(0.07)	(0.07)
Migration index		0.12	0.10	0.13	0.11	0.16^{*}
		(0.08)	(0.08)	(0.08)	(0.07)	(0.07)
Business tax			-0.27^{***}	-0.24^{**}	-0.26^{***}	-0.24^{***}
			(0.08)	(0.08)	(0.07)	(0.07)
Employment rate			-0.06	-0.18	-0.20^{*}	-0.20^{*}
			(0.08)	(0.10)	(0.09)	(0.08)
Self-employment rate			0.22^{**}	0.08	0.02	0.02
			(0.08)	(0.10)	(0.09)	(0.09)
Spatial effect				✓	✓	✓
State FE					✓	✓
Housing prices						✓
AIC	74,659.72	74,640.22	74,623.80	74602.10	71,414.99	71,404.30
N	10,688	10,688	10,688	10688	10,688	10,688
Adjusted R^2	0.03	0.03	0.03	0.04	0.29	0.29

Note: All independent variables are scaled by standard deviation, the baseline is category urban, the three German city-states are represented by one fixed effect, standard errors reported in parentheses, $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$, the mean of excess GPs is -0.37 with a standard deviation of 8.01.

Notably, we do not observe a correlation between unsatisfied health care demand and rural regions. It contrasts the clear association between less densely populated areas and lower levels of excess GPs (see Table 6.5). This might be due to two factors: (1) GPs might be distributed near the outskirts so that they can cover demand outside the city as well; (2) by design, our index for uncovered cases is not symmetric in oversupply: It might be the case that parts of a city are strongly oversupplied, while other parts are strongly undersupplied (i.e., local oversupply does not reduce overall undersupply). If we look at the results regarding the commuter index, we find a positive correlation with uncovered cases ($p < 0.001$). This is a reversed picture compared to our results for the correlation with excess number of GPs: Regions with higher in-commuting have more excess GPs. Yet, the correlations between uncovered cases and migration index or economic power measures are not significant.

In summary, we find mixed results regarding the two main issues of an overall shortage of GPs in Germany and an especially severe undersupply of rural regions. We find only a small discrepancy in the overall number of GPs, which we regard as noteworthy but not necessarily as a major cause for concern. With regard to the inequity in regional health care markets, our results are inconclusive. When we look at the supply side, i.e., the allocation of GPs, our benchmark solution would increase the number of rural doctors. However, we do not find evidence for structural failure to satisfy rural health care demand. Our additional analysis suggests that there are likely other factors capturing a region's unmet health care demand.

Table 6.6: Regression: Percentage of Uncovered Cases and Municipality Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	29.68***	30.43***	44.20***	30.03***	28.56***	30.75***
	(3.57)	(3.57)	(4.56)	(5.13)	(5.57)	(5.86)
Rural	−0.30	−1.48	−0.07	−0.35	−1.03	−1.57
	(3.58)	(3.59)	(3.60)	(3.59)	(3.69)	(3.72)
Suburban	−1.98	−2.89	−1.58	−1.98	−2.38	−2.69
	(3.59)	(3.60)	(3.61)	(3.59)	(3.68)	(3.69)
Commuter index		0.67**	0.76***	0.79***	0.90***	0.89***
		(0.22)	(0.22)	(0.22)	(0.23)	(0.23)
Migration index		−0.30	−0.34	−0.34	−0.30	−0.25
		(0.22)	(0.23)	(0.23)	(0.24)	(0.24)
Business tax			−0.11	−0.02	0.02	0.03
			(0.22)	(0.22)	(0.22)	(0.22)
Employment rate			−0.95***	−0.29	−0.35	−0.34
			(0.22)	(0.30)	(0.31)	(0.31)
Self-employment rate			−0.70**	0.37	−0.08	−0.08
			(0.08)	(0.10)	(0.09)	(0.09)
Spatial effect				✓	✓	✓
State FE					✓	✓
Housing prices						✓
AIC	96,683.20	96,676.09	96,658.11	96,483.05	96,456.40	96,458.37
N	10,688	10,688	10,688	10,688	10,688	10,688
Adjusted R^2	0.00	0.00	0.00	0.02	0.03	0.03

Note: All independent variables are scaled by standard deviation, the baseline is the urban category, the three German city-states are represented by one fixed effect, standard errors reported in parentheses, $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$, the mean of uncovered cases is 28.28 with a standard deviation of 21.97.

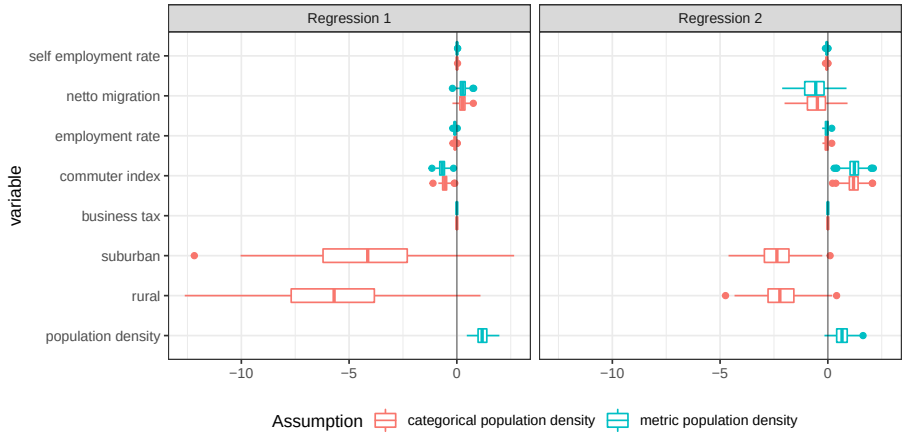
6.5 Sensitivity analysis

Influence of aggregation on regression results

Our regression results may depend on the aggregation of grids to the municipality level. To test this, we carry out an experiment where we create random partitions, by iteratively merging each municipality with a random number of neighbors, until each municipality is merged with some neighbor. See appendix 6.A for details.

For each partition, we aggregate all covariates and re-estimate the models. Binary variables are aggregated by choosing the value where the majority of the populations lives, e.g., we categorize an aggregation as "urban" if the majority of people lives in urban municipalities. We aggregate metric covariates by sum if appropriate, otherwise we take an average, weighted by population. Figure 6.7 shows a boxplot of the estimates under 200 random partitions.

Figure 6.7: Placebo Test: Aggregation



Note: Boxplots of coefficients under varying aggregations. The colors of the boxplots are mapped to the specification of population density.

For both regressions, results seem robust towards the choice of aggregation. Estimates of the dummies for population density exhibit higher variation than estimates of the metric variables. A reason for this is the lower level of precision when aggregating the variables, which increases variance of estimates. Therefore, to analyze this effect, we fit the model with a metric specification of population density and find a robust effect. This shows that the underlying correlation is robust towards change in aggregation. For both regressions, the sign and order of effect size are overall robust.

Influence of input parameters on number of GPs

We test the influence of the two constraints for the capacitated set cover algorithm, i.e., limited capacity of GPs and driving time less than 15 minutes, on the required number of GPs. To do so we vary each parameter, keeping the other parameter constant. We analyze the effect of GP capacity and the threshold given the maximal driving time to the nearest GP. Figure 6.8 shows the results. The effect of the capacity is approximately linear, while the effect of the threshold seems exponential.

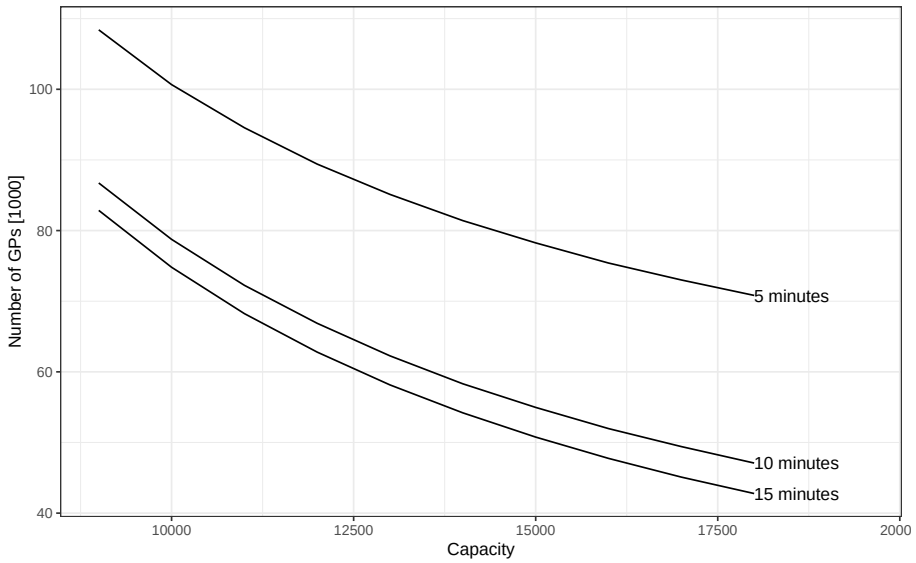
6.6 Conclusion

The main objective of this study was to implement a benchmark for GP allocation to shed light on the roots of regional inequalities in unmet health care demand and offer a way to approach this issue.

We base a greedy capacitated algorithm on health care demand calculated on a very fine scale using sociodemographic characteristics from survey-based panel data. This allows us to optimize the number of general practitioners by allocation on 1km^2 grids given that each resident has access to at least one general practitioner within a driving time of 15 minutes, and that each general practitioner has a limited annual capacity of $13,000$ patient contacts.

Our findings suggest that focusing solely on uneven regional distribution in favor of urban areas might be misleading. Especially when this is the only

Figure 6.8: Placebo Test: Input Parameters



Note: Sensitivity of required number of GPs towards input parameters

criterion guiding incentives for physician allocation. A technical benchmark solution, as suggested in our study, can provide remedy.

There are some limitations to our approach. For one, our estimates of uncovered cases represent an upper bound. There are three reasons for this: first, we assume that all patients always choose the closest GP. This is potentially a greater issue in cities, because they have a higher GP density within a 15-minute driving time radius so that patients have more alternatives when the capacities of the closest doctor are exhausted. We address this point by including spatial effects as controls. Second, doctors may have more than 25 hours of weekly patient contact that they are required to have by law (TSVG). Therefore, we most probably find an upper bound of uncovered cases. Because we aim to construct a benchmark allocation, we rely on this law as it implements the 25 hours as a guideline for the weekly workload of GPs. Third, the SOEP 2016 wave on which we base our estimation of health care demand only asks for doctors' visits in general and not explicitly for GP visits.

Another point is that our demand is based only on age and gender due to lack of other parameters that could be connected between survey and RWI-GEO-GRID data. Therefore, we are not able to include individual preconditions influencing health care demand.

Last, in this paper we only analyze the situation as of 2019 and do not include a forecast of the future. Demographic change will probably have a large influence on the discussed situation. Therefore, we do not claim that the issues of a shortage of GPs and the especially severe undersupply of rural regions cannot become true concerns. This surely is worth investigating in the future. Nevertheless, we believe that an efficient German-wide benchmark of GP allocation, rather than solely recruiting rural GPs, might help to reduce the probability of these two issues to become inevitable.

6.A Appendix

Summary of websites NASHIP and RASHIP

Table 6.7 give the website for each health insurance association.

Table 6.7: Scraped Websites

Association	URL
NASHIP Germany (2019)	https://arztsuche.kbv.de/
RASHIP Baden-Württermberg (2019)	https://www.kvbawue.de/
RASHIP Bavaria (2019)	https://www.kvb.de/
RASHIP Berlin (2019)	https://www.kvberlin.de/
RASHIP Berlin (2019)	http://kvbb.de/
RASHIP Bremen (2019)	https://www.kvhb.de/
RASHIP Hamburg (2019)	http://www.kvhh.net/
RASHIP Hessen (2019)	https://www.kvhessen.de/
RASHIP Nordrhein (2019)	https://www.kvn.de/
RASHIP Mecklenburg-Vorpommern (2019)	https://www.kvmv.de/
RASHIP Nordrhein (2019)	https://www.kvno.de/
RASHIP Rhineland-Palatinate (2019)	https://www.kv-rlp.de/
RASHIP Saarland (2019)	https://www.kvsaarland.de/
RASHIP Saxony (2019)	https://www.kvs-sachsen.de/
RASHIP Saxony-Anhalt (2019)	https://www.kvsa.de/
RASHIP Schleswig-Holstein (2019)	https://www.kvsh.de/
RASHIP Thuringia (2019)	https://www.kv-thueringen.de/
RASHIP Westfalen-Lippe (2019)	https://www.kvwl.de/

Description of steps for robustness checks

- Sample a municipality i
- Determine ∂, the set consisting of ith neighbours
- Draw u from a discrete uniform distribution with support

  $$\{1, \ldots, \min(3, N_\partial)\},$$

  where N_∂ denotes the cardinality of ∂
- Take a simple random sample with size u from ∂ and merge each sampled municipalities with i

Chapter 7

Conclusion

The first part of this thesis developed methods to scale up Bayesian inference to large scale problems. The second part tackled relevant applied large-scale problems, using methods from approximation algorithms and machine learning. As Chapter 1 summarizes all chapters in detail, this conclusion presents several extensions of the presented work.

Chapter 3 covers the Cox proportional hazard model, which is the benchmark model for survival analysis. As such it is a natural starting point. However, other models, for instance non-proportional hazard models or joint models for longitudinal and survival data, might also be covered in the given framework. Furthermore, the inclusion of other basis functions apart from P-splines is possible, for instance to cover periodic or spatial effects.

One might further speed up the approximative Metropolis-Hastings algorithm given in Chapter 4, an interesting variant would be the use of parallelization: One might use several subsamples in parallel, each producing an estimator for the loglikehood-ratio sum, which could be combined to form a composite estimator. Furthermore, an interesting approach is to debias the estimator. Another approach might be to combine the methods with those of Chapter 3, this could furthermore accelerate inference for Bayesian hazard regression.

Chapter 5 deals with the effect of a market premia scheme on negative electricity prices. The chapter might be extended by use of other data sets not available at publication time: this might allow the comparison with a control group by looking at comparable time series for other countries.

The analysis of the distribution of general practitioners (GPs) in Chapter 6 might be extended in several ways. One extension is the use of further data, for example by a survey, either covering the behavior of patients or the behavior of GPs. This would allow us to answer further empirical questions: For instance, we assume that patients always choose the nearest GP, is this assumption justified? Furthermore, it would be preferable to obtain official data on general practitioners, which was not available for our analysis. This would allow us to estimate the surplus and deficit of GPs in a given region more precisely. Additionally, including a forecast into the analysis would allow us to identify regions where a deficit in the supply of GPs might occur in the future.

By improving methods for the analysis of large data sets, this thesis lays some groundwork for future use, while the application of large data methodology shows its usefulness for relevant policy questions.

Bibliography

Agora. (2019). European energy transition 2030: The big picture [Agora Energiewende].

Alquier, P., Friel, N., Everitt, R., & Boland, A. (2014). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, *29*, 1–19. https://doi.org/10.1007/s11222-014-9521-x

Andor, M., Flinkerbusch, K., Janssen, M., Liebau, B., & Wobben, M. (2010). Negative Strompreise und der Vorrang Erneuerbarer Energien. *Zeitschrift für Energiewirtschaft*, *34*(2), 91–99. https://doi.org/10.1007/s12398-010-0015-z

Andor, M., Frondel, M., & Vance, C. (2017). Germany's Energiewende: A tale of increasing costs and decreasing willingness-to-pay. *The Energy Journal*, *38*(2), 91–99. https://doi.org/10.5547/01956574.38.SI1.mand

Andor, M., & Voss, A. (2016). Optimal renewable-energy promotion: Capacity subsidies vs. generation subsidies. *Resource and Energy Economics*, *45*, 144–158. https://doi.org/10.1016/j.reseneeco.2016.06.002

Andor, M. A., Frondel, M., & Sommer, S. (2018). Equity and the willingness to pay for green electricity in Germany. *Nature Energy*, *3*(10), 876–881. https://doi.org/10.1038/s41560-018-0233-x

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, *50*(1-2), 5–43. https://doi.org/10.1023/A:1020281327116

Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, *37*(2), 697–725. https://doi.org/10.1214/07-AOS574

BA. (2008). Bundesagentur für Arbeit: Der Arbeitsmarkt in Deutschland: Monatsbericht Dezember und das Jahr 2007.

Bardenet, R., Doucet, A., & Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 405–413.

Bardenet, R., Doucet, A., & Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, *18*, 1515–1557.

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PloS one*, *12*(8), 1–20. https://doi.org/https://doi.org/10.1371/journal.pone.0180698

Bartha, G., & Kocsis, S. (2011). Standardization of geographic data: The European inspire directive. *European Journal of Geography*, *2*, 79–89.

BMWi. (2017). Erneuerbare Energien in Zahlen – Nationale und internationale Entwicklung im Jahr 2016.

Boelmann, B., & Schaffner, S. (2018). FDZ data description: Real-estate data for Germany (RWI-GEO-RED): Advertisements on the internet platform ImmobilienScout24. *RWI Projektberichte*.

Breidenbach, P., & Eilers, L. (2018). RWI-GEO-GRID: Socio-economic data on grid level. *Jahrbücher für Nationalökonomie und Statistik*, *238*(6), 609–616. https://doi.org/https://doi.org/10.1515/jbnst-2017-0171.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brezger, A., & Steiner, W. J. (2008). Monotonic regression based on Bayesian P–Splines: An application to estimating price response functions from store-level scanner data. *Journal of Business & Economic Statistics*, *26*, 90–104.

Cai, B., Lin, X., & Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics and Data Analysis*, *55*(9), 2644–2651. https://doi.org/10.1016/j.csda.2011.03.013

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chauvet, G., & Tillé, Y. (2006). A fast algorithm for balanced sampling. *Computational Statistics*, *21*(1), 53–62. https://doi.org/10.1007/s00180-006-0250-2

Chen, P.-Y., & Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, *57*(4), 1030–1038. https://doi.org/10.1111/j.0006-341X.2001.01030.x

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-AOAS285

Chvatal, V. (1979). A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, *4*(3), 233–235. https://doi.org/10.1287/moor.4.3.233

Cludius, J., Hermann, H., Matthes, F. C., & Graichen, V. (2014). The merit order effect of wind and photovoltaic electricity generation in Germany 2008 – 2016: Estimation and distributional implications. *Energy Economics*, *44*, 302–313. https://doi.org/10.1016/j.eneco.2014.04.020

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Daskin, M. S., & Dean, L. K. (2005). Location of health care facilities. In M. L. Brandeau, S. Francois, & P. William P (Eds.), *Operations research and health care* (pp. 43–76). Springer. https://doi.org/10.1007/b106574

De Boor, C. (1978). *A practical guide to splines* (Rev. ed.). Springer.

de Lagarde, C. M., & Lantz, F. (2018). How renewable production depresses electricity prices: Evidence from the German market. *Energy Policy*, *117*, 263–277. https://doi.org/10.1016/j.enpol.2018.02.048

Deville, J.-C., & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, *91*(4), 893–912. https://doi.org/10.1093/biomet/91.4.893

Dijkstra, L., & Poelman, H. (2014). A harmonised definition of cities and rural areas: The new degree of urbanisation. *European Commission Working Paper*.

Dykstra, R. L., & Laud, P. (1981). A Bayesian nonparametric approach to reliability. *The Annals of Statistics*, *9*(2), 356–367. https://doi.org/10.1214/aos/1176345401

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

EEX. (2018). Retrieved December 8, 2016, from https://www.eex.com/de/

Eilers, L. (2017). Is my rental price overestimated? A small area index for Germany. *Ruhr Economic Papers*, *734*, 1–25. https://doi.org/10.4419/86788854

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 89–102.

Fanone, E., Gamba, A., & Prokopczuk, M. (2013). The case of negative day-ahead electricity prices. *Energy Economics*, *35*, 22–34. https://doi.org/10.1016/j.eneco.2011.12.006

Federal Institute for Research on Building, Urban Affairs and Spatial Development. (2005). *Bundesamt für Bauwesen und Raumordnung, Raumordnungsbericht 2005*. Selbstverlag des Bundesamtes für Bauwesen und Raumordnung.

Federal Office for Building and Regional Planning. (2020). Bundesinstitut für Bau-, Stadt- und Raumforschung.

Feige, U. (1998). A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, *45*(4), 634–652. https://doi.org/10.1145/285055.285059

Fernandez, T., Rivera, N., & Teh, Y. W. (2016). Gaussian processes for survival analysis. In M. D. Lee, U. Sugiyama, I. Luxburg, & R. G. Guyon (Eds.), *Advances in neural information processing systems 29* (pp. 5021–5029). Curran Associates, Inc.

Finkelstein, A., Gentzkow, M., & Williams, H. (2016). Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics*, *131*(4), 1681–1726. https://doi.org/10.1093/qje/qjw023

Fraunhofer Institute for Solar Energy Systems. (2018). Retrieved May 24, 2018, from https://energy-charts.de/

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Gelfand, A. E., & Mallick, B. K. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics*, *51*(3), 843–852. https://doi.org/10.2307/2532986

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). CRC press.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gerster, A. (2016). Negative price spikes at power markets: The role of energy policy. *Journal of Regulatory Economics*, *50*(3), 271–289. https://doi.org/10.1007/s11149-016-9311-9

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German socio-economic panel (soep). *Jahrbücher für Nationalökonomie und Statistik*, *239*(2), 345–360. https://doi.org/https://doi.org/10.1515/jbnst-2018-0022.

Green, P. J., Łatuszyński, K., Pereyra, M., & Robert, C. P. (2015). Bayesian computation: A summary of the current state, and samples backwards

and forwards. *Statistics and Computing*, *25*(4), 835–862. https://doi.org/10.1007/s11222-015-9574-5

Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Harper, P. R., Shahani, A., Gallagher, J., & Bowie, C. (2005). Planning health services with explicit geographical considerations: A stochastic location–allocation approach. *Omega*, *33*(2), 141–152. https://doi.org/10.1016/j.omega.2004.03.011

Hennerfeind, A., Brezger, A., & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, *101*(475), 1065–1075. https://doi.org/10.1198/016214506000000348

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics*, *4*(2), 105–123. https://doi.org/10.1080/03610927508827232

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

IEA, & IRENA. (2018). IEA / IRENA joint policies and measures database [International Energy Agency and International Renewable Energy Agency]. http://www.iea.org/policiesandmeasures/renewableenergy/

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Johansson, N., Jakobsson, N., & Svensson, M. (2018). Regional variation in health care utilization in Sweden – the importance of demand-side factors. *BMC Health Services Research*, *18*(1), 403. https://doi.org/10.1186/s12913-018-3210-y

Kaeding, M. (2018). *RcppGreedySetCover: Greedy set cover* [R package version 0.1.0]. https://CRAN.R-project.org/package=RcppGreedySetCover

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *40*(2), 214–221. https://doi.org/10.1111/j.2517-6161.1978.tb01666.x

Keppler, J. H., Phan, S., & Le Pen, Y. (2016). The impacts of variable renewable production and market coupling on the convergence of French and German electricity prices. *The Energy Journal*, *37*(3), 343–360. https://doi.org/10.5547/01956574.37.3.jkep

Ketterer, J. C. (2014). The impact of wind power generation on the electricity price in Germany. *Energy Economics*, *44*, 270–280. https://doi.org/10.1016/j.eneco.2014.04.003

Kneib, T., & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, *34*(1), 207–228. https://doi.org/10.1111/j.1467-9469.2006.00524.x

Korattikara, A., Chen, Y., & Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. *Proceedings of The 31st International Conference on Machine Learning*, 181–189.

Korte, B., & Vygen, J. (2018). *Combinatorial optimization* (6th ed.). Springer. https://doi.org/10.1007/978-3-662-56039-6

Lang, S., & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, *13*(1), 183–212. https://doi.org/10.1198/1061860043010

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer.

Lin, X., Cai, B., Wang, L., & Zhang, Z. (2015). A Bayesian proportional hazards model for general interval-censored data. *Lifetime Data Analysis*, *21*(3), 470–490. https://doi.org/10.1016/j.csda.2011.03.013

Lovász, L. (1975). On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, *13*(4), 383–390. https://doi.org/10.1016/0012-365X(75)90058-8

Luxen, D., & Vetter, C. (2011). Real-time routing with OpenStreetMap data. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 513–516. https://doi.org/10.1145/2093973.2094062

Maclaurin, D., & Adams, R. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. *Proceedings of the Thirtieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-14)*, 543–552.

Magnusson, M., Andersen, M., Jonasson, J., & Vehtari, A. (2019). Bayesian leave-one-out cross-validation for large data. *International Conference on Machine Learning*, *97*, 4244–4253.

Maire, F., Friel, N., & Alquier, P. (2015). Light and widely applicable MCMC: Approximate Bayesian inference for large datasets. *arXiv preprint arXiv:1503.04178*.

Marianov, V., & Taborga, P. (2001). Optimal location of public health centres which provide free and paid services. *Journal of the Operational Research Society*, *52*(4), 391–400. https://doi.org/10.1057/palgrave.jors.2601103

McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer, C., & Pratola, M. (2018). *BART: Bayesian additive regression trees* [R package version 2.1]. https://CRAN.R-project.org/package=BART

NASHIP. (2018). Regionale Verteilung der Ärzte in der vertragsärztlichen Versorgung. https://gesundheitsdaten.kbv.de/cms/html/16402.php

NASHIP Germany. (2019). Retrieved July 30, 2019, from https://arztsuche.kbv.de/

National Associations of Statutory Health Insurance Physicians', & Verband der niedergelassenen Ärzte Deutschlands. (2018). Tabellenband. Ärztemonitor 2018 Ergebnisse für Haus- und Fachärzte.

Nicholls, G. K., Fox, C., & Watt, A. M. (2012). Coupled MCMC with a randomized acceptance probability. *arXiv preprint arXiv:1205.6857*.

Nicolosi, M. (2010). Wind power integration and power system flexibility – An empirical analysis of extreme events in Germany under the new negative price regime. *Energy Policy*, *38*(11), 7257–7268. https://doi.org/10.1016/j.enpol.2010.08.002

Nieto-Barajas, L. E., & Walker, S. G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics*, *29*(3), 413–424. https://doi.org/10.1111/1467-9469.00298

OECD. (2014). Health at a glance: Europe 2014.

Parka, M.-J., & Honga, S.-P. (2009). Approximation of the capacitated set cover [Manuscript submitted for publication].

Paschen, M. (2016). Dynamic analysis of the German day-ahead electricity spot market. *Energy Economics*, *59*, 118–128. https://doi.org/10.1016/j.eneco.2016.07.019

Praktiknjo, A., & Erdmann, G. (2016). Renewable electricity and backup capacities: An (un-)resolvable problem? *The Energy Journal*, *37*, 89–106. https://doi.org/10.5547/01956574.37.SI2.apra

Quiroz, M., Kohn, R., Villani, M., & Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, *114*(526), 831–843. https://doi.org/10.1080/01621459.2018.1448827

Rais, A., & Viana, A. (2011). Operations research in healthcare: A survey. *International Transactions in Operational Research*, *18*(1), 1–31. https://doi.org/10.1111/j.1475-3995.2010.00767.x

RASHIP Baden-Württermberg. (2019). Retrieved August 20, 2019, from https://www.arztsuche-bw.de

RASHIP Bavaria. (2019). Retrieved August 22, 2019, from https://dienste.kvb.de/arztsuche/app/einfacheSuche.htm

RASHIP Berlin. (2019). Retrieved August 12, 2019, from https://www.kvberlin.de/60arztsuche/index.html

RASHIP Bremen. (2019). Retrieved August 12, 2019, from https://www.kvhb.de/arztsuche/

RASHIP Hamburg. (2019). Retrieved August 12, 2019, from https://www.kvhh.net/kvhh/arztsuche/index/p/274

RASHIP Hessen. (2019). Retrieved August 12, 2019, from https://arztsuchehessen.de/arztsuche

RASHIP Mecklenburg-Vorpommern. (2019). Retrieved August 12, 2019, from https://www.kvmv.de/ases-kvmv/ases.jsf

RASHIP Nordrhein. (2019). Retrieved August 12, 2019, from https://www.kvno.de/20patienten/10arztsuche/index.html

RASHIP Rhineland-Palatinate. (2019). Retrieved August 12, 2019, from https://www.kv-rlp.de/patienten/arztfinder/

RASHIP Saarland. (2019). Retrieved August 13, 2019, from http://arztsuche.kvsaarland.de/

RASHIP Saxony. (2019). Retrieved July 2, 2019, from https://asu.kvs-sachsen.de/arztsuche/pages/search.jsf

RASHIP Saxony-Anhalt. (2019). Retrieved August 13, 2019, from https://www.kvsa.de/service/arzt_und_therapeutensuche_in_sachsen_anhalt.html

RASHIP Schleswig-Holstein. (2019). Retrieved August 13, 2019, from https://arztsuche.kvsh.de/suche.do#

RASHIP Thuringia. (2019). Retrieved August 13, 2019, from https://www.kv-thueringen.de/arztsuche/

RASHIP Westfalen-Lippe. (2019). Retrieved August 13, 2019, from https://www.kvwl.de/earzt/index.htm

RES. (2018). Legal sources on renewable energy [RES LEGAL Europe].

Ringler, P., Keles, D., & Fichtner, W. (2017). How to benefit from a common European electricity market design. *Energy Policy*, *101*, 629–643. https://doi.org/10.1016/j.enpol.2016.11.011

Robert, C. P. (2001). *The Bayesian choice: From decision-theoretic foundations to computational implementation*. Springer.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). Springer.

Römer, B., Reichhart, P., Kranz, J., & Picot, A. (2012). The role of smart metering and decentralized electricity storage for smart grids: The importance of positive externalities. *Energy Policy*, *50*, 486–495. https://doi.org/10.1016/j.enpol.2012.07.047

Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, *21*(15), 2175–2197. https://doi.org/10.1002/sim.1203

Royston, P., & Parmar, M. K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, *30*(19), 2409–2421. https://doi.org/10.1002/sim.4274

Salm, M., & Wübker, A. (2020). Sources of regional variation in healthcare utilization in Germany. *Journal of Health Economics*, *69*, 102271–102286. https://doi.org/10.1016/j.jhealeco.2019.102271

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer.

Stan Development Team. (2021). *Stan modeling language users guide and reference manual, 2.26*. Retrieved February 6, 2021, from https://mc-stan.org

Stensrud, M. J., Aalen, J. M., Aalen, O. O., & Valberg, M. (2018). Limitations of hazard ratios in clinical trials. *European Heart Journal*, *40*(17), 1378–1383. https://doi.org/10.1093/eurheartj/ehy770

Sutton, M., & Lock, P. (2000). Regional differences in health care delivery: Implications for a national resource allocation formula. *Health Economics*, *9*(6), 547–559. https://doi.org/10.1002/1099-1050(200009)9:6⟨547::AID-HEC543⟩3.0.CO;2-E

Therneau, T., & Atkinson, B. (2019). *Rpart: Recursive partitioning and regression trees* [R package version 4.1-15]. https://CRAN.R-project.org/package=rpart

Tillê, Y. (2006). *Sampling algorithms*. Springer.

TSO. (2018). Informationen zur Direktvermarktung [German Transmission System Operators].

Vazirani, V. V. (2001). *Approximation algorithms*. Springer.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. https://doi.org/10.1214/12-SS102

Weber, C. (2010). Adequate intraday market design to enable the integration of wind energy into the european power systems. *Energy Policy*, *38*(7), 3155–3163. https://doi.org/10.1016/j.enpol.2009.07.040

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Würzburg, K., Labandeira, X., & Linares, P. (2013). Renewable generation and electricity prices: Taking stock and new evidence for Germany and Austria.

Energy Economics, *40*, 159–171. https://doi.org/10.1016/j.eneco.2013.09.011

Zhang, Y., Hua, L. E. I., & Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics*, *37*(2004), 338–354. https://doi.org/10.1111/j.1467-9469.2009.00680.x

Zhang, Y., Baik, S. H., Fendrick, A. M., & Baicker, K. (2012). Comparing local and regional variation in health care spending. *New England Journal of Medicine*, *367*(18), 1724–1731. https://doi.org/10.1056/NEJMsa1203980

Zhao, L., Tian, L., Uno, H., Solomon, S. D., Pfeffer, M. A., Schindler, J. S., & Wei, L. J. (2012). Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*, *9*(5), 570–577. https://doi.org/10.1177/1740774512455464

Zhou, H., & Hanson, T. (2018). A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. *Journal of the American Statistical Association*, *113*(522), 571–581. https://doi.org/10.1080/01621459.2017.1356316

Eidesstattliche Erklärung

Ich gebe folgende eidesstattliche Erklärung ab: Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig ohne unzulässige Hilfe Dritter verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen unter der Angabe der Quelle als solche gekennzeichnet habe. Die Grundsätze für die Sicherung guter wissenschaftlicher Praxis an der Universität Duisburg-Essen sind beachtet worden. Ich habe die Arbeit keiner anderen Stelle zu Prüfungszwecken vorgelegt.

Essen, 24.07.2020

Matthias Kaeding