Universität Trier

Estimation of Regional Transition Probabilities for Spatial Dynamic Microsimulations from Survey Data Lacking in Regional Detail

Jan Pablo Burgard
Joscha Krause
Simon Schmaus

Research Papers in Economics
No. 12/19

Estimation of Regional Transition Probabilities for Spatial Dynamic Microsimulations from Survey Data Lacking in Regional Detail

Jan Pablo Burgard, Joscha Krause, Simon Schmaus

Department of Economic and Social Statistics, Trier University

Abstract

Spatial dynamic microsimulations allow for the multivariate analysis of complex socio-economic systems with geographic segmentation. For this, a synthetic replica of the system as base population is stochastically projected into future periods. Thereby, the projection is based on micro-level transition probabilities. They need to accurately represent the characteristic dynamics of the system to allow for reliable simulation outcomes. In practice, transition probabilities are unknown and must be estimated from suitable survey data. This can be challenging when the characteristic dynamics vary locally. Survey data often lacks in regional detail due to confidentiality restrictions and limited sampling resources. In that case, transition probability estimates may misrepresent local dynamics as a result of insufficient local observations and coverage problems. The simulation process then fails to provide an authentic evolution. We present two transition probability estimation techniques that account for regional heterogeneity when the survey data lacks in regional detail. Using methods of constrained optimization and ex-post alignment, we show that local micro level transition dynamics can be accurately recovered from aggregated regional benchmarks. The techniques are compared in theory and subsequently tested in a simulation study.

Keywords: Constrained Maximum Likelihood, Logit Scaling, Spatiotemporal Modelling, Regional Benchmark

1 Introduction

Microsimulations are powerful tools for the multivariate analysis of complex systems, such as economic markets or medical care infrastructures. They differ from the more established macrosimulations in terms of the objects that are considered in the simulation process. While in macrosimulations the behaviour of aggregated system-intrinsic entities is modelled, microsimulations target the smallest entities of the system (units) directly. This allows for the investigation of multidimensional interactions and nonlinear dependencies within the system that cannot be studied by macrosimulations. Examples for microsimulation models can be found in Klevmarken (2010), Lawson (2011), O'Donoghue et al. (2011), as well as Markham et al. (2017).

Microsimulations are often conducted according to a basic procedure. First, a base population as synthetic replica of the system of interest is constructed. In practice, this may be either artificially constructed data, or real-world observations from administrative records and surveys (Li and O'Donoghue, 2014). Next, multiple parameters that characterize the system in its initial state are altered in scenarios. Thereby, the alterations are designed to target properties of the system in the light of the research objectives. The effects of the alterations are then projected into future periods and construct individual branches in the system's evolution. After a given number of periods (simulation horizon), the branches are compared and give insights on important dynamics and interdependencies within the system (Burgard et al., 2019).

There are different types of microsimulations. They mainly differ in the manner in which the mentioned alterations are projected. An important distinction is between static and dynamic microsimulations (Li and O'Donoghue, 2013). Static microsimulations are characterized by the constancy of unit characteristics over time. When constructing the synthetic replica, every unit is provided with a set of characteristics that determines its behaviour and interaction with other units. In static microsimulations, these characteristics don't change over the simulation horizon. Only specific simulation inputs are altered, depending on the research objectives. Examples for static microsimulation models can be found in Peichl et al. (2010), as well as Sutherland and Figari (2013).

Dynamic microsimulations, on the other hand, are characterized by stochastic changes of unit characteristics (state transitions) over time. The evolution of units, as well as their interactions, are determined by frequently changing base datasets. Examples for dynamic microsimulation models can be found in O'Donoghue et al. (2009), as well as Fialka et al. (2011). If the dynamic microsimulation is time-discrete, state transitions can only appear

periodically at distinct points of time. If the simulation is time-continuous, they can appear at any given time and thus are modelled via survival functions (Willekens, 2009).

In the following, we focus on dynamic microsimulations with discrete time. More precisely, we look at dynamic microsimulation models in socio-economic research where primarily polytomous variables are of interest. This conceptual delimitation differentiates the topic from other fields where corresponding simulation methods are also relevant, such as particle physics or cancer research. In order to initialize a corresponding simulation, every unit in the synthetic replica must be provided with an individual set of transition probabilities. They define the conditional likelihood of a state transition for some unit characteristic given its current state as well as other characteristics. The probabilities constitute stochastic processes within the synthetic replica over the simulation horizon. Thereby, they need to represent elementary dynamics of the real system as genuine as possible to obtain valid simulation results. Transition probabilities are usually unknown in practice and thus must be estimated. This is done via parametric statistical models using suitable survey data.

Transition probability estimation can be challenging if the system of interest is geographically segmented into regions. In the literature, a microsimulation that accounts for regional data structures is often referred to a small area or spatial microsimulation (Rahman et al., 2010; Rahman and Harding, 2016; Tanton et al., 2018). In such a setting, there may be heterogeneity across regions with respect to transition dynamics. The statistical approach used for transition probability estimation must explicitly account for these local differences in order to adequately reflect the system's dynamics. However, in practice, we often encounter the problem that the survey data used for transition probability estimation lacks in regional detail. Due to confidentiality restrictions, regional identifiers that would allow for spatial localization of the sample elements may be censored. Regional heterogeneity in transition dynamics then cannot be observed as spatial aggregates are indistinguishable.

Further, even if regional identifiers are available, the majority of survey samples often contain only a few observations per region due to limited resources. In that case, observed regional transition frequencies may be inaccurate or even biased as a result of coverage problems. Ignoring these issues may cause only small deviations in the initial phase of the simulation. But due to the complex interactions between units, the inaccuracies accumulate and self-reinforce over the simulation horizon. Hence, local transition dynamics are misrepresented over time and the simulation fails to provide an authentic evolution of the synthetic replica with respect to the real system. The simulation outcomes are subsequently not reliable anymore (Chin and Harding, 2006; Tanton, 2014). Accordingly, if the survey

data used for transition probability estimation lacks in regional detail, methodological adjustments are required.

In this paper, we investigate two extensions of the multinomial logit model (McCullagh and Nelder, 1989; Greene, 2002) to recover local heterogeneity in transition dynamics when the primary database lacks in regional detail. The extensions are addressed in the context of spatial dynamic microsimulations where it is required to provide micro-level transition probabilities for every unit of the synthetic replica. We consider a situation in which external benchmarks on regional transition dynamics are available (e.g. from census data). The general idea is to incorporate this regional information in the multinomial logit model and modify the estimation process such that resulting micro level probability estimates are consistent with the benchmarks when aggregated. Benchmark consistency can be either perfect in the sense that all values are reproduced exactly, or approximately by allowing for deviations in terms of box constraints. With this, we seek to recover the previously unobservable regional heterogeneity in transition dynamics on the micro level.

The first extension is called logit scaling and was originally proposed by Stephensen (2016). It is an ex-post alignment method based on iterative proportional fitting (Bishop et al., 1970). After the initial estimation process, the transition probability estimates obtained from multinomial logit are adjusted sequentially until they are consistent with the external benchmarks. Thereby, the Kullback-Leibler divergence between original and adjusted estimates is minimized. The second extension was developed in this study and draws from constrained maximum likelihood theory (e.g. Dong and Wets, 2000; Chatterjee et al., 2016). The external regional benchmarks are used to directly modify parameter estimation in the multinomial logit model. This done by imposing box constraints on model predictions and thus transition probability estimates. Constrained parameter estimation is performed by a sequential quadratic programming approach (Kraft, 1994).

The methods are first described and discussed in theory. Afterwards, they are applied and tested under different settings in an extended simulation study. For this, survey data from the German Microcensus 2013. We find that the inclusion of aggregated regional benchmarks allows for the recovery of local micro level transition dynamics despite a lack in regional detail. The remainder of the paper is organized as follows. In Chapter 2, the required technical framework and the multinomial logit model are described. In Chapter 3, the two extensions to the model are presented. Chapter 4 contains the simulation study. Chapter 5 closes with some conclusive remarks.

2 Basic Methodology

2.1 Technical Framework

We introduce a technical framework that is required to describe the estimation methods within this paper. For simplicity, assume that the system of interest is an arbitrary population of individuals. In the following, three representations of this population are considered. First, let $\mathcal{U}$ denote the real population that contains $|\mathcal{U}| = N$ individuals. It marks the system that the researcher seeks to analyze. With respect to the spatial segmentation discussed in Chapter 1, assume that $\mathcal{U} = \bigcup_{r=1}^{R} \mathcal{U}_r$ consists of R areas indexed by $r = 1, ..., R$ with $\mathcal{U}_r, \mathcal{U}_v \in \mathcal{U}$ pairwise disjoint for $r \neq v$. The number of individuals per area is $|\mathcal{U}_r| = N_r$ with $\sum_{r=1}^{R} N_r = N$. Second, denote $\tilde{\mathcal{U}}$ as the synthetic replica of $\mathcal{U}$ containing $|\tilde{\mathcal{U}}| = \tilde{N}$ units indexed by $u = 1, ..., \tilde{N}$. It is projected over simulation horizon $\mathcal{S} = \{1, ..., S\}$ with simulation periods s to provide essential insights on $\mathcal{U}$. Note that formally $\mathcal{S}$ is an index set. For the projection, every $u \in \tilde{\mathcal{U}}$ must be associated with a set of transition probabilities that accurately represent relevant dynamics of $\mathcal{U}$. Since $\tilde{\mathcal{U}}$ is designed to be a close-to-reality representation of $\mathcal{U}$, it must reflect the spatial segmentation of the real population. Accordingly, we have $\tilde{\mathcal{U}} = \bigcup_{r=1}^{R} \tilde{\mathcal{U}}_r$ with $\tilde{\mathcal{U}}_r, \tilde{\mathcal{U}}_v \in \tilde{\mathcal{U}}$ pairwise disjoint for $r \neq v$. The number of units per synthetic area is $|\tilde{\mathcal{U}}_r| = \tilde{N}_r$ with $\sum_{r=1}^{R} \tilde{N}_r = \tilde{N}$. And third, let $\mathcal{D} \subset \mathcal{U}$ be a survey sample that is drawn from $\mathcal{U}$ with observations of $|\mathcal{D}| = n$ unique individuals for T time periods, where $i = 1, ..., n$ and $t = 1, ..., T$. It marks the data basis from which transition probability estimates are derived in order to apply them to $\tilde{\mathcal{U}}$. In an ideal sampling situation, $\mathcal{D}$ contains area-specific subsamples $\mathcal{D}_r \subset \mathcal{U}_r$ with $|\mathcal{D}_r| = n_r$ and n_r sufficiently large for all r. In practice, the regional index r might be unknown, or n_r may be small.

As stated previously, every unit is associated with a set of characteristics that determines its behaviour and interaction with other units. Since we consider dynamic microsimulations, the unit-specific values of this set may change in any $s \in \mathcal{S}$. Let $Y : \mathcal{Y} \rightarrow \mathbf{E}$ be a polytomous random variable representing a unit characteristic for which transition probabilities are desired. It can have a finite number J of possible outcomes that form the state space $\mathcal{Y} = \{Y_1, ..., Y_J\}$ whose elements Y_j are unordered, mutually exclusive, and indexed by $j = 1, ..., J$. $\mathbf{E}$ is some measurable space. Let $\mathcal{F}$ be the σ-algebra of subsets of $\mathcal{Y}$ and $P : \mathcal{F} \rightarrow [0, 1]$ be a probability measure. In order to understand the following definitions, we briefly recall the concept of discrete stochastic processes.

Definition 1 *Let $(\mathcal{Y}, \mathcal{F}, P)$ be a probability space, $\mathbf{E}$ be a measurable space, and $\mathcal{S} = \mathbb{Z}^+$ be an index set. Suppose that for every $s \in \mathcal{S}$, there is a $Y^{(s)} : \mathcal{Y} \rightarrow \mathbf{E}$ defined on $(\mathcal{Y}, \mathcal{F}, P)$.*

Then, the function $Y : \mathcal{S} \times \mathcal{Y} \rightarrow \mathbf{E}$ *is called discrete stochastic process.*

Thereafter, we need to distinguish between the concepts of state occurrence and state transition for the estimation methods in the subsequent chapters.

Definition 2 *Let* $y_{ur}^{(s)}$ *be the realized value of* Y *for some* $u \in \tilde{\mathcal{U}}_r$ *in* $s \in \mathcal{S}$*. A state occurrence is the outcome of a discrete stochastic process where* $y_{ur}^{(s)} = Y_j$ *for a given* $Y_j \in \mathcal{Y}$*. Its probability is given by* $\pi_{ur}^{(s)j} := P(y_{ur}^{(s)} = Y_j)$*.*

Definition 3 *A state transition is the outcome of a discrete stochastic process where* $y_{ur}^{(s)} = Y_j$ *and* $y_{ur}^{(s+1)} = Y_k$ *with* $Y_j, Y_k \in \mathcal{Y}$ *and* $Y_j \neq Y_k$ *for* $u \in \tilde{\mathcal{U}}_r$ *and* $s \in \mathcal{S}$*. Its probability is given by* $\pi_{ur}^{(s+1)jk} := P(y_{ur}^{(s+1)} = Y_k | y_{ur}^{(s)} = Y_j)$*.*

Please note that $Y_j = Y_k$ is also allowed in the simulation process. However, we don't refer to it as a state transition since the unit-specific value of Y has not changed between periods. Let $\mathbf{X} = (X_1, ..., X_p)$ be a set of random variables with $X_\iota : \Omega \rightarrow \mathbb{R}$ for $\iota = 1, ..., p$ that represent other unit characteristics statistically related to Y. Denote the value of $\mathbf{X}$ for $u \in \tilde{\mathcal{U}}_r$ in simulation period s as $\mathbf{x}_{ur}^{(s)}$. The conditional probability of a state transition from Y_j to Y_k for $u \in \tilde{\mathcal{U}}_r$ in $s+1$ is defined according to:

$$\pi_{ur}^{(s+1)jk}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}, s) := P(y_{ur}^{s+1} = Y_k | y_{ur}^{(s)} = Y_j, \mathbf{X} = \mathbf{x}_{ur}^{(s+1)}, \mathcal{S} = s), \tag{1}$$

where $Y_j, Y_k \in \mathcal{Y}$ and $0 \leq \pi_{ur}^{(s+1)jk}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}, s) \leq 1$. For notational convenience, assume that it is sufficient to only consider the last period when modelling the state transition. Further, for simplicity, assume that the conditional transition probabilities are time-invariant and only vary across units as well as simulation scenarios. Hence,

$$\pi_{ur}^{(s+1)jk}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}, s) = \pi_{ur}^{(s+1)jk}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}). \tag{2}$$

However, the methods discussed in this paper can be adjusted to provide conditional time-variant transition probabilities, e.g. by including time variable in $\mathbf{X}$. Note that $\pi_{ur}^{(s+1)k}$ is driven by the last state $y_{ur}^{(s)}$ and the additional characteristics $\mathbf{x}_{ur}^{(s+1)}$. On the contrary, $\pi_{ur}^{(s+1)jk}$ varies exclusively with $\mathbf{x}_{ur}^{(s+1)}$. Generally, in the light of all simulation periods and potential states, transition dynamics can be summarized in a right stochastic matrix

$$\mathbf{P}_{ur} = \begin{pmatrix} \pi_{ur}^{(s+1)11}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}) & \pi_{ur}^{(s+1)12}(y_{ur}^{s}, \mathbf{x}_{ur}^{s+1}) & ... & \pi_{u}^{(s+1)1J}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}) \\ \vdots & \vdots & \ddots & \\ \pi_{ur}^{(s+1)J1}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}) & \pi_{ur}^{(s+1)J2}(y_{ur}^{s}, \mathbf{x}_{ur}^{s+1}) & ... & \pi_{ur}^{(s+1)JJ}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}) \end{pmatrix}, \tag{3}$$

where $\sum_{k=1}^{J} \pi_{ur}^{(s+1)jk}(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)}) = 1$. However, note that for a given $u \in \tilde{\mathcal{U}}_r$ with known $y_{ur}^{(s)} = Y_j$ and $\mathbf{x}_{ur}^{(s+1)}$, we have $\pi_{ur}^{(s+1)jk} = \pi_{ur}^{(s+1)k}$. In that case, the probabilities of state occurrences are equal to the state transition probabilities. Subsequently, they can be summarized for a given simulation period $s+1$ in a vector $\boldsymbol{\pi}_{ur}^{(s+1)} = (\pi_{ur}^{(s+1)j1}, ..., \pi_{ur}^{(s+1)jJ})$. Next, the concept of regional heterogeneity in transition dynamics in our setting is introduced.

Definition 4 *Let* $\pi_{ur}^{(s+1)jk} = P(y_{ur}^{s+1} = Y_k | y_{ur}^{(s)} = Y_j)$ *with* $Y_j, Y_k \in \mathcal{Y}$ *and* $s \in \mathcal{S}$. *Regional heterogeneity in transition dynamics is a situation where* $\tilde{N}_r^{-1} \sum_{u \in \tilde{\mathcal{U}}_r} \pi_{ur}^{(s+1)jk} \neq \tilde{N}_q^{-1} \sum_{u \in \tilde{\mathcal{U}}_q} \pi_{uq}^{(s+1)jk}$ *for* $\tilde{\mathcal{U}}_r, \tilde{\mathcal{U}}_q \in \tilde{\mathcal{U}}$ *and* $r \neq q$.

Accordingly, the term corresponds to regional differences in the mean of probabilities for a given state transition. Within this paper, we provide an overview of methods to obtain transition probability estimates $\widehat{\pi}_{ur}^{(s+1)jk}$ from the sample elements $i \in \mathcal{D}$ to obtain $\boldsymbol{\pi}_{ur}^{(s+1)}$ for every $u \in \tilde{\mathcal{U}}$ and $s = 1, ..., S-1$ under regional heterogeneity. It is argued that if $\mathcal{D}$ does not allow for the empirical observation of local transition dynamics, the resulting estimates $\widehat{\pi}_{ur}^{(s+1)jk}$ are inaccurate with respect to transition dynamics in $\mathcal{U}_r$.

2.2 Multinomial Logit

We use the well-established multinomial logit model (McCullagh and Nelder, 1989; Greene, 2002) as basic methodology for transition probability estimation. All descriptions are with respect to the survey-based micro data obtained from $\mathcal{D}$. Assume that the regional index r is not observed for the sample elements. Let the pair $(Y, \mathbf{X})$ be observed for the sampled individuals $i \in \mathcal{D}$ with time- and individual-specific values $(y_i^{(t)}, \mathbf{x}_i^{(t)})$ in t. Let $\pi_i^{(t)j} = P(y_i^{(t)} = Y_j)$ be the occurrence probability of Y_j for individual i in t with $\sum_{j=1}^{J} \pi_i^{(t)j} = 1$. Define $Y_i^{(t)j}$ as a binary random variable that takes value 1, if $y_i^{(t)} = Y_j$, and 0 else. Its realization is denoted by $y_i^{(t)j}$, with $\sum_{j=1}^{J} y_{ir}^{(t)j} = 1$. The probability distribution of $y_i^{(t)j}$ is given by

$$P(Y_i^{(t)1} = y_i^{(t)1}, ..., Y_i^{(t)J} = y_i^{(t)J}) = \begin{pmatrix} 1 \\ y_i^{(t)1}, ..., y_i^{(t)J} \end{pmatrix} (\pi_i^{(t)1})^{y_i^{(t)1}} \cdot ... \cdot (\pi_i^{(t)J})^{y_i^{(t)J}}. \quad (4)$$

In order to model the probabilities $\pi_i^{(t)j}$ dependent on the time- and individual-specific covariate values $\mathbf{x}_i^{(t)}$ as well as the last state value $y_i^{(t-1)}$, it is common to determine one state as reference outcome. Since our basic setting is to estimate the probability of a transition from Y_j to Y_k, we use Y_j as reference. Recall from Chapter 2.1 that the probability of occurrence is equal to the transition probability when conditioned on the previous period. The log-odds for all feasible states relative to this reference outcome are calculated as a

linear function of the predictors:

$$\eta_i^{(t)k} = \eta_i^{(t)k}(\alpha_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k) = \log\left(\frac{\pi_i^{(t)k}}{\pi_i^{(t)j}}\right) = \alpha_k + (\mathbf{x}_i^{(t)})'\boldsymbol{\beta}_k + (\mathbf{y}_i^{(t-1)})'\boldsymbol{\gamma}_k, \tag{5}$$

for $k \neq j$, where $\alpha_k \in \mathbb{R}$ is a state-specific constant, $\boldsymbol{\beta}_k \in \mathbb{R}^p$ is the vector of regression coefficients associated with $\mathbf{x}_i^{(t+1)}$ and $\boldsymbol{\gamma}_k \in \mathbb{R}^J$ is a coefficient vector quantifying the influence of $\mathbf{y}_i^{(t-1)} = (y_i^{(t-1)1}, ..., y_i^{(t-1)J})$. Note that $\boldsymbol{\beta}_k$ also varies across log-odds. One obtains a set of $J-1$ independent binary regression models, in which all other states are separately regressed against the reference outcome. From (5), the individual probabilities can be obtained from (Böhning, 1992)

$$\pi_i^{(t)k} = \frac{\exp\left(\eta_i^{(t)k}(\alpha_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k)\right)}{1 + \sum_{l \in \{1,...,J\}\setminus j} \exp\left[\eta_i^{(t)l}(\alpha_l, \boldsymbol{\beta}_l, \boldsymbol{\gamma}_l)\right]} \tag{6}$$

for $k \neq j$, and for $k = j$

$$\pi_i^{(t)k} = \frac{1}{1 + \sum_{l \in \{1,...,J\}\setminus j} \exp\left[\eta_i^{(t)l}(\alpha_l, \boldsymbol{\beta}_l, \boldsymbol{\gamma}_l]\right)}. \tag{7}$$

The parameters of the multinomial logit model are then estimated via maximum likelihood. Define $\boldsymbol{\theta}_k := (\alpha_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k)$. For notational convenience, we display the log-likelihood function for a single individual $i \in \mathcal{D}$, which is given by (Böhning, 1992)

$$\begin{aligned} l_i^{(t)}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J) &= \log\left(\textstyle\prod_{l=1}^J (\pi_i^{(t)l})^{y_i^{(t)l}}\right) \\ &= \textstyle\sum_{l \in \{1,...,J\}\setminus j} y_i^{(t)l} \eta_i^{(t)l}(\boldsymbol{\theta}_l) - \log\left(1 + \sum_{l \in \{1,...,J\}\setminus j} \exp\left[\eta_i^{(t)l}(\boldsymbol{\theta}_l)\right]\right). \end{aligned} \tag{8}$$

Assuming the sample observations are independent, model parameter estimates are obtained from minimizing the sum of negative individual log-likelihoods

$$(\widehat{\boldsymbol{\theta}}_1, ..., \widehat{\boldsymbol{\theta}}_J) = \underset{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_J}{\operatorname{argmin}} \left\{ - \left[\sum_{t=2}^{T} \sum_{i \in \mathcal{D}} l_i^{(t)}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J) \right] \right\}. \tag{9}$$

Maximization can be performed by various numerical procedures, such as generalized iterative scaling (Darroch and Ratcliff, 1972), iteratively reweighted least squares (Bishop, 2006), or the Newton-Raphson (Böhning, 1992). Theoretically, once the model parameter estimates are obtained from $\mathcal{D}$, they can be used to estimate $\widehat{\boldsymbol{\pi}}_{ur}^{(s+1)}$ for all $u \in \tilde{\mathcal{U}}_r$. This is

then achieved by combining them with the unit-specific values $(y_{ur}^{(s)}, \mathbf{x}_{ur}^{(s+1)})$ according to

$$\widehat{\pi}_{ur}^{(s+1)jk} = \widehat{\pi}_{ur}^{(s+1)k} = \frac{\exp[\widehat{\alpha}_k + (\mathbf{x}_{ur}^{(s+1)})'\widehat{\boldsymbol{\beta}}_k + \widehat{\gamma}_k^j]}{1 + \sum_{l \in \{1,...,J\} \setminus j}^{J} \exp[\widehat{\alpha}_l + (\mathbf{x}_{ur}^{(s+1)})'\widehat{\boldsymbol{\beta}}_l + \widehat{\gamma}_l^j]} \quad \forall\ k = 1, ..., J, \tag{10}$$

where $\widehat{\gamma}_k^j \in \widehat{\boldsymbol{\gamma}}_k$ is the coefficient resulting from $y_{ur}^{(s)} = Y_j$. However, we argue that if the survey data $\mathcal{D}$ does not allow for the observation of regional heterogeneity according to Definition 4, the resulting transition probability estimates $\widehat{\pi}_{ur}^{(s+1)jk}$ for a given $u \in \tilde{\mathcal{U}}_r$ misrepresent the local transition dynamics in $\mathcal{U}_r$.

3 Extensions for Regional Heterogeneity

We now show how to account for regional heterogeneity despite $\mathcal{D}$ lacking in regional detail. For this, assume that benchmarks

$$\tau_r^{(t)k} := \sum_{i \in \mathcal{U}_r} \sum_{k=1}^{J} y_{ir}^{(t)k} \quad \text{with} \quad y_{ir}^{(t)k} = \begin{cases} 1 & \text{if } y_{ir}^{(t)} = Y_k \\ 0 & \text{else} \end{cases} \tag{11}$$

are known for all $Y_k \in \mathcal{Y}$ as well as all $r = 1, ..., R$ and some t corresponding to $s+1$. At this point, we postpone the discussion whether corresponding knowledge is realistic and pick it up again in Section 5. Due to the lack in regional detail in $\mathcal{D}$ and the consequences for the resulting transition probability estimates (10), it may be that

$$\Delta_r^{(s)k} := \left| \sum_{u \in \tilde{\mathcal{U}}_r} \sum_{j=1}^{J} \widehat{\pi}_{ur}^{(s+1)jk} - \tau_r^{(t)k} \right| > \epsilon_r^k, \tag{12}$$

for some $\tilde{\mathcal{U}}_r \in \tilde{\mathcal{U}}$ and $Y_k \in \mathcal{Y}$. $\epsilon_r^k \in \mathbb{R}_{\geq 0}$ denotes a predefined critical deviation value for category Y_k resulting from box constraints with respect to $\mathcal{U}_r$. To ensure consistency with respect to the regional benchmarks, we would like to find new sets $\{\widehat{\boldsymbol{\pi}}_{1r}^{(s+1)*}, ..., \widehat{\boldsymbol{\pi}}_{N_r r}^{(s+1)*}\}$ that satisfy the system of inequality constraints

$$\begin{array}{cccc} \Delta_1^{(s)1} \leq \epsilon_1^1, & \Delta_1^{(s)2} \leq \epsilon_1^2, & \dots & \Delta_1^{(s)J} \leq \epsilon_1^J, \\ \Delta_2^{(s)1} \leq \epsilon_2^1, & \Delta_2^{(s)2} \leq \epsilon_2^2, & \dots & \Delta_2^{(s)J} \leq \epsilon_2^J, \\ \vdots & \vdots & \ddots & \\ \Delta_R^{(s)1} \leq \epsilon_R^1, & \Delta_R^{(s)2} \leq \epsilon_R^2, & \dots & \Delta_R^{(s)J} \leq \epsilon_R^J. \end{array} \tag{13}$$

For this, the methodology for transition probability estimation must be extended. Hereafter, two corresponding extensions are described.

3.1 Logit Scaling

The first extension is logit scaling and has been proposed by Li and O'Donoghue (2014), as well as Stephensen (2016). It is a simple multivariate ex-post alignment method that manipulates the estimated probability distribution resulting from the multinomial logit model. For some $\tilde{\mathcal{U}}_r \in \tilde{\mathcal{U}}$ and $Y_j \in \mathcal{Y}$, this is achieved by sequentially adjusting the mean

$$q_{ur}^{(s+1)j} := \frac{1}{\tilde{N}_r} \sum_{u \in \tilde{\mathcal{U}}_r} \widehat{\pi}_{ur}^{(s+1)j} \tag{14}$$

until the regional constraint is satisfied. Thereby, the adjustment is performed via iterative proportional fitting Bishop et al. (1970). Hereafter, we sketch the methodology based on Stephensen (2016). Let $\mathcal{Q}_r = \{q_{ur}^{(s+1)j} | u = 1, ..., \tilde{N}_r; j = 1, ..., J\}$ be the joint discrete probability distribution of states and units for $\tilde{\mathcal{U}}_r$. Denote $\mathcal{Q}_r^{[0]}$ as the initial probability distribution estimated from the multinomial logit model in Chapter 2.2. On that note, let $q_{ur}^{[0](s+1)j}$ and $\widehat{\pi}_{ur}^{[0](s+1)j}$ be the initial versions of $q_{ur}^{(s+1)j}$ and $\widehat{\pi}_{ur}^{(s+1)j}$ after estimating the multinomial logit model. The objective is to find a set of new probability distributions $\{\mathcal{Q}_1^*, ..., \mathcal{Q}_R^*\}$ that satisfy (13) while restricting the adjustment of the $\widehat{\pi}_{ur}^{[0](s+1)j}$ for all $Y_j \in \mathcal{Y}$ within $\tilde{\mathcal{U}}_r \in \tilde{\mathcal{U}}$ to a minimum. This is achieved by minimizing the Kullback-Leibler divergence between $\mathcal{Q}_r^{[0]}$ and $\mathcal{Q}_r$

$$\begin{aligned} D_r^{KL}\left(\mathcal{Q}_r \,||\, \mathcal{Q}_r^{[0]}\right) &= \sum_{u \in \tilde{\mathcal{U}}_r} \sum_{j=1}^{J} q_{ur}^{(s+1)j} \log\left(\frac{q_{ur}^{(s+1)j}}{q_{ur}^{[0](s+1)j}}\right) \\ &= \frac{1}{N_r} \sum_{u \in \tilde{\mathcal{U}}_r} \sum_{j=1}^{J} \widehat{\pi}_{ur}^{(s+1)j} \log\left(\frac{\widehat{\pi}_{ur}^{(s+1)j}}{\widehat{\pi}_{ur}^{[0](s+1)j}}\right) \end{aligned} \tag{15}$$

subject to the system of inequality constraints. Hence, for some $\tilde{\mathcal{U}}_r \in \tilde{\mathcal{U}}$, we obtain the minimization problem

$$\min_{\widehat{\boldsymbol{\pi}}_{1r}^{(s+1)}, ..., \widehat{\boldsymbol{\pi}}_{N_r r}^{(s+1)}} \left\{ D_r^{KL}\left(\mathcal{Q}_r \,||\, \mathcal{Q}_r^{[0]}\right) \right\} \quad \text{s.t.} \quad \Delta_r^{(s)1} \leq \epsilon_r^1, \ldots, \Delta_r^{(s)J} \leq \epsilon_r^J. \tag{16}$$

Solving it for $r = 1, ..., R$ individually then obtains sets $(\widehat{\boldsymbol{\pi}}_{1r}^{(s+1)*}, ..., \widehat{\boldsymbol{\pi}}_{N_r r}^{(s+1)*})$ with the desired properties. Stephensen (2016) showed that the minimization problem can be solved by a bi-proportionate scaling algorithm Define matrices $\widehat{\boldsymbol{\Pi}}_r^{[0](s+1)} := (\widehat{\boldsymbol{\pi}}_{1r}^{[0](s+1)}, ..., \widehat{\boldsymbol{\pi}}_{\tilde{N}_r r}^{[0](s+1)})'$

for all $\tilde{\mathcal{U}}_r \in \tilde{\mathcal{U}}$. Let $\ell = 1, 2, ...$ be the index of iterations. The algorithm is performed on $\widehat{\mathbf{\Pi}}_1^{(s+1)}, ..., \widehat{\mathbf{\Pi}}_R^{(s+1)}$ as described hereafter.

Algorithm 1 Bi-Proportionate Scaling

1: set $\widehat{\pi}_{ur}^{[\ell](s+1)j} = \widehat{\pi}_{ur}^{[0](s+1)j}$ for $j = 1, ..., J$ and $r = 1, ..., R$

2: **for** $r = 1, ..., R$ **do**

3: **while** $\Delta_r^{(s)j} > \epsilon_r^j$ for any $Y_j \in \mathcal{Y}$ **do**

4: scale columns with $\omega_k^{[\ell]}$: set $\widehat{\pi}_{ur}^{[\ell+1](s+1)j} = \omega_k^{[\ell]} \widehat{\pi}_{ur}^{[\ell](s+1)j}$ such that

$$\left| \sum_{u \in \tilde{\mathcal{U}}_r} \widehat{\pi}_{ur}^{[\ell+1](s+1)j} - \tau_r^{(t)j} \right| < \epsilon_r^k$$

5: scale rows with $\zeta_{ur}^{[\ell]}$: set $\widehat{\pi}_{ur}^{[\ell+1](s+1)j} = \zeta_{ur}^{[\ell]} \widehat{\pi}_{ur}^{[\ell](s+1)j}$ such that

$$\sum_{j=1}^{J} \widehat{\pi}_{ur}^{[\ell+1](s+1)j} = 1$$

6: **end while**

7: **end**

Note that logit scaling can also be seen as an ex-post adjustment of the intercept estimated in the multinomial logit model. For further details, we refer to Stephensen (2016).

3.2 Constrained Maximum Likelihood

The second extension was developed in this study and can be viewed as a special case of constrained maximum likelihood estimation (Dong and Wets, 2000; Chatterjee et al., 2016). Unlike logit scaling, it is a direct alignment method where the consistency to the regional benchmarks is achieved within model parameter estimation of the multinomial logit model. This is achieved by solving the constrained minimization problem

$$\min_{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J} \left\{ - \left[\sum_{t=2}^{T} \sum_{i \in \mathcal{D}} l_i^{(t)}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J) \right] \right\} \quad \text{s.t.} \quad \Delta_r^{(s)1} \leq \epsilon_r^1, \ldots, \Delta_r^{(s)J} \leq \epsilon_r^J \tag{17}$$

for all $\tilde{\mathcal{U}}_r \in \tilde{\mathcal{U}}$ individually. The solutions are obtained from a sequential quadratic programming approach (Kraft, 1994). At this point, providing a technical description of the computational details is beyond the scope of this paper. Therefore, we only briefly sketch

the method and refer to Kraft (1994) for deeper insights. Sequential quadratic programming allows for the inclusion of nonlinear constraints within a given minimization problem. In the process, each of the R original constraint minimization problems in (17) is substituted by a series of constrained least squares problems. The algorithm optimizes successive second-order approximations of the objective function with first-order affine approximations of the constraints. Starting point of the method is the minimization of the negative log-likelihood over the sample observations in $\mathcal{D}$. Let $\ell = 1, 2, ...$ denote the index of iterations required for model parameter estimation. Within every iteration and for a given $\tilde{\mathcal{U}}_r \in \tilde{\mathcal{U}}$, predictions $\widehat{\pi}_{ur}^{[\ell](s+1)jk}$ for all $u \in \tilde{\mathcal{U}}_r$ are produced simultaneously to model parameter estimation. Thereby, the predictions are based on the current model parameter estimates $\widehat{\boldsymbol{\theta}}_1^{[\ell]}, ..., \widehat{\boldsymbol{\theta}}_J^{[\ell]}$. For each region, the (potentially) resulting deviations $\Delta_r^{[\ell](s)1}, ..., \Delta_r^{[\ell](s)J}$ are penalized by adding them from the regional Lagrangian. Define the negative log-likelihood

$$L(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J) := -\left[\sum_{t=2}^{T}\sum_{i\in\mathcal{D}} l_i^{(t)}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J)\right], \tag{18}$$

regional constraint functions

$$\mathcal{C}_r(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J) := -\sqrt{\left(\sum_{u\in\tilde{\mathcal{U}}_r}\sum_{j=1}^{J}\widehat{\pi}_{ur}^{(s+1)jk} - \tau_r^{(t)k}\right)^2} \quad \forall\ r = 1, ..., R, \tag{19}$$

as well as the regional Lagrangians

$$\mathcal{L}_r(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J, \boldsymbol{\lambda}) := L(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J) - \boldsymbol{\lambda}'\mathcal{C}_r(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J) \quad \forall\ r = 1, ..., R, \tag{20}$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier. In a given iteration ℓ of the algorithm, a descent direction $\mathbf{d}^{[\ell]}$ is defined as a solution to the constrained least squares subproblem

$$\begin{aligned}\mathbf{d}^{[\ell]} = \underset{\mathbf{d}}{\operatorname{argmin}} &\left\{L(\boldsymbol{\theta}_1^{[\ell]}, ..., \boldsymbol{\theta}_J^{[\ell]}) + \nabla L(\boldsymbol{\theta}_1^{[\ell]}, ..., \boldsymbol{\theta}_J^{[\ell]})'\mathbf{d} + \frac{1}{2}\mathbf{d}'\nabla^2\mathcal{L}_r(\boldsymbol{\theta}_1^{[\ell]}, ..., \boldsymbol{\theta}_J^{[\ell]}, \boldsymbol{\lambda}^{[\ell]})\mathbf{d}\right\} \\ \text{s.t.} \quad &\mathcal{C}_r(\boldsymbol{\theta}_1^{[\ell]}, ..., \boldsymbol{\theta}_J^{[\ell]}) + \nabla\mathcal{C}_r(\boldsymbol{\theta}_1^{[\ell]}, ..., \boldsymbol{\theta}_J^{[\ell]})'\mathbf{d} \geq 0,\end{aligned} \tag{21}$$

which can be solved efficiently according to Kraft (1988). Afterwards, an appropriate step size is determined and a BFGS update is used to obtain new parameter estimates $\widehat{\boldsymbol{\theta}}_1^{[\ell+1]}, ..., \widehat{\boldsymbol{\theta}}_J^{[\ell+1]}$. The procedure is repeated until the constraints are satisfied. For more details, see Kraft (1994).

4 Simulation Study

4.1 Setup

The presented methods are tested within a Monte Carlo simulation study with $S = 500$ runs. Thereby, we focus on two particular questions that are relevant in the context of transition probability estimation from survey data for dynamic microsimulations:

1. Can the methods obtain decent estimates when the survey sample is subject to coverage problems (i.e. when specific population groups are underrepresented)?
2. Can the methods recover regional transition dynamics when there are no regional identifiers for the survey observations?

Although both aspects can emerge from a lack in regional detail, we evaluate them individually. This allows for a clear separation of the potential adjustment effects in different scenarios. Further, we conduct two distinct simulations within the study: a model-based and a design-based. They are briefly sketched hereafter.

Model-Based Setup

In the model-based simulation, the data used for estimation and evaluation is created artificially. This includes all population representations $\mathcal{U}, \tilde{\mathcal{U}}, \mathcal{D}$ described in Section 2.1. Note that for the study $\mathcal{U} = \tilde{\mathcal{U}}$ is defined. It would be redundant to create an artificial population $\mathcal{U}$ first, and then generate a synthetic replica $\tilde{\mathcal{U}}$ that is meant to be an artificial representation of $\mathcal{U}$. The advantage of the model-based approach is that the purely artificial creation allows for a controlled environment with respect to modelling and estimation. Therefore, it serves as a proof of concept in this study. The population $\mathcal{U}$ is created in each simulation run individually with a size of $N = 100\,000$. Transition probability estimates are produced for the employment status (employed, unemployed) of any $i \in \mathcal{U}$. Thus, we choose Y as binary variable rather than a polytomous one. This is, on the one hand, for illustrative purposes as adjustment effects are much easier to visualize in the binary case. On the other hand, it is due to the heavy computational burden of the simulation study. However, note that the simulation results are still meaningful with respect to the multinomial estimation techniques. It is well-known that a multinomial logit for J categories can be restated as a set of $J - 1$ binary logits.

Regarding X, every individual is associated with iid variables age (15-80), gender (male, female) and ISCED level (1-5). The functional relation between Y and X on the micro level is retrieved from a binary logit model on data from the German Microcensus 2013.

The values used for artificial employment probability definition can be found in Table 2 of the appendix. The employment status is added by drawing a state regarding the calculated probabilities. Subsequently, samples of 1%, 0.05% and 0.025% are drawn from $\mathcal{U}$ to calculate a logit model for the estimation of the employment status on the population data. In the simulation, we assume that the set of auxiliary variables and the regional population totals $\tau_r := \sum_{i \in \mathcal{U}_r} y_{ir}$ are known as external benchmarks.

Design-Based Setup

The design-based simulation is largely similar to the model-based setup. However, the important difference is that it is conducted on real-world observations obtained from the German Microcensus 2013. There is no synthetic data generation in the original sense. The advantage of this approach is that the methods are tested on actual data structures that occur in practice, where there are no ideal distribution characteristics. The sampled individuals of the Microcensus are taken as fixed population $\mathcal{U}$. In every simulation run, random samples of 0.1%, 0.05% and 0.025% are drawn and subsequently used for transition probability estimation. The inference is then analysed with respect to the fixed population rather than the hyper parameters as in the model-based study. The number of employed persons per region is again defined as a known benchmark. In order to examine the alignment to regional benchmarks, the sampling model is used to predict the employment status in each German federal state. In this case, the true employment rate of each federal states is assumed to be known.

Scenarios and Implementation

With respect to the first aspect of missing regional detail, different sampling scenarios are implemented to mimic coverage problems in both simulation types. In practice, a coverage problem occurs when the sample proportions of essential characteristics don't fit the corresponding proportions in the regional population. Table 1 shows the different scenarios where disproportional sampling probabilities are used to differ in the observation proportions of Y and X. The scenarios are repeated for all three sample sizes.

Scen. 1	Simple Random Sampling
Scen. 2	Reduction of the drawing probability of employed persons by 10%
Scen. 3	Reduction of the drawing probability of employed persons by 20%
Scen. 4	Reduction of the drawing probability of employed persons by 30%
Scen. 5	Reduction of the drawing probability of unemployed persons by 10%
Scen. 6	Reduction of the drawing probability of unemployed persons by 20%
Scen. 7	Reduction of the drawing probability of unemployed persons by 30%
Scen. 8	Reduction of the drawing probability of males by 40%
Scen. 9	Reduction of the drawing probability of males by 60%
Scen. 10	Reduction of the drawing probability of males by 80%
Scen. 11	Reduction of the drawing probability education level 1 by 40%
Scen. 12	Reduction of the drawing probability education level 1 by 60%
Scen. 13	Reduction of the drawing probability education level 1 by 80%
Scen. 14	Reduction of the drawing probability education level 2 by 40%
Scen. 15	Reduction of the drawing probability education level 2 by 60%
Scen. 16	Reduction of the drawing probability education level 2 by 80%
Scen. 17	Reduction of the drawing probability age $\geq$50 by 80%
Scen. 18	Reduction of the drawing probability age $<$50 by 80%

Table 1: Sampling scenarios

Regarding the second aspect of missing regional details, in every simulation run, a random sample is drawn from $\mathcal{U}$ via simple random sampling. The sample observations are drawn from all regions of the population with equal sampling probability (SRS). Thereby, all regional identifiers of the sample observations are deleted. Transition probability estimation is then conducted for one specific region at a time.

Performance Measures

To evaluate the discussed methods with respect to transition probability estimation and predictive inference, we look at several performance measures. The first is the mean squared

deviation of the predicted state occurrence probabilities and the actual state occurrences,

$$\frac{1}{S \cdot N} \sum_{s=1}^{S} \sum_{i \in \mathcal{U}} \left(\widehat{\pi}_i^{(s)} - y_i^{(s)} \right)^2, \tag{22}$$

where $y_i^{(s)} = 1$ if the state has occurred in simulation run s, and 0 else. Since transition probability estimation is performed by minimizing the negative log-likelihood of the logit model, this measure is suitable for assessing goodness of fit with respect to the estimated parameters $\widehat{\boldsymbol{\theta}}^{(s)}$ on the population data. It is given by

$$-\sum_{i \in \mathcal{U}} \left(y_i^{(s)} \log \left[\frac{\exp\left((\boldsymbol{x}_i^{(s)})' \widehat{\boldsymbol{\beta}}^{(s)} \right)}{1 + \exp\left((\boldsymbol{x}_i^{(s)})' \widehat{\boldsymbol{\beta}}^{(s)} \right)} \right] + (1 - y_i^{(s)}) \log \left[\frac{1}{1 + \exp\left((\boldsymbol{x}_i^{(s)})' \widehat{\boldsymbol{\beta}}^{(s)} \right)} \right] \right), \tag{23}$$

where $\widehat{\boldsymbol{\beta}}^{(s)}$ are obtained from the sample data in the s-th run of the simulation. The smaller the negative log-likelihood value, the better the estimated parameters fit the population. To further evaluate the model parameter estimates, we look at the (squared) differences

$$\boldsymbol{\beta}^{(s)} - \widehat{\boldsymbol{\beta}}^{(s)} \quad \text{and} \quad \left(\boldsymbol{\beta}^{(s)} - \widehat{\boldsymbol{\beta}}^{(s)} \right)^2 \tag{24}$$

of the parameters estimated on the sampling data in simulation run s to the parameters of the population model. In the design-based framework, the population remains coefficients $\boldsymbol{\beta_s}$ remain constant. Note that in the design-based setup, the population remains constant. Hence, $y_i^{(s)} = y_i$ and $\boldsymbol{x}_i^{(s)} = \boldsymbol{x}_i$ for all $i \in \mathcal{U}$, as well as $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}$.

4.2 Model-Based Results

Although both aspects of missing regional detail have been studied in both simulation types, we focus hereafter on coverage problems to avoid repetitions. The recovery of regional transition dynamics is then discussed in the design-based setup. All additional results are included in the appendix. Logit scaling is referred to as *LS* while constrained maximum likelihood is called *Opt*. The standard logit model without adjustments is denoted as *Mod*.

To evaluate the goodness of fit with respect to the parameters on the population data, we compare the negative log-likelihood values (21) in Figure 1. It can be seen that both Opt and LS are capable of improving the likelihood relative to Mod. Especially in case of smaller sample sizes, Opt consistently leads to better likelihood values. Accordingly, despite the absence of coverage problems in Scenario 1, the additional information on the

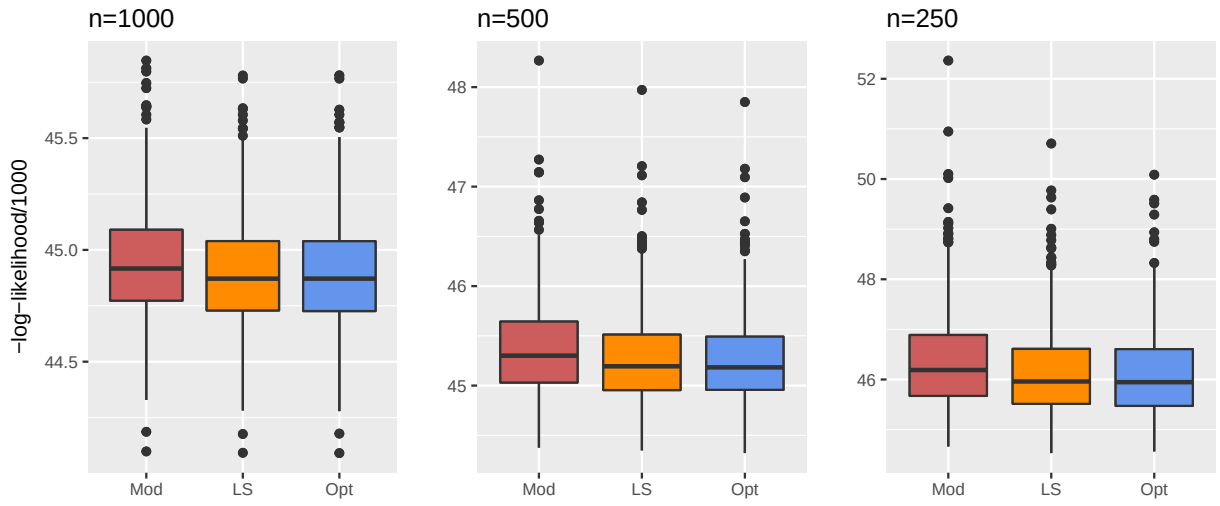

Figure 1: Scenario 1: negative log-likelihood values

benchmarks slightly improves the probability estimates. The mean value over 500 sampling runs is for $n = 500$ ($n = 1000$, $n = 250$) 45.643 (45.090, 46.888), for Mod, 45.412 (45.039, 45.611) for LS and 45.492 (45.039, 45.606) for Opt.

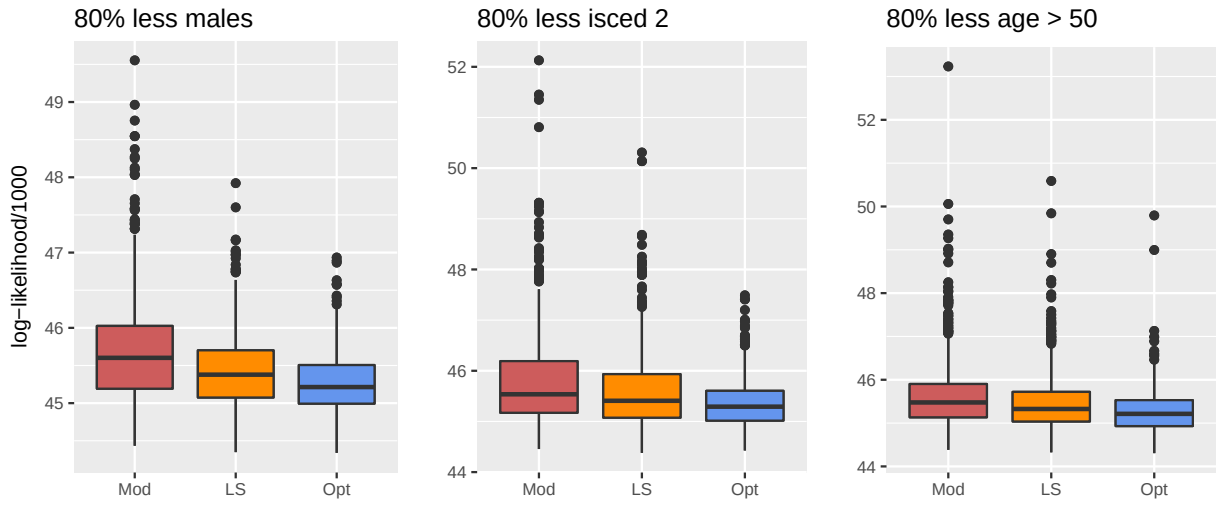

Figure 2: Scenario 10, 16, 18: negative log-likelihood values, n=500

Now, coverage problems with respect to the auxiliary variables are introduced via disproportional sampling. Here, the performance differences are more evident. We look at Scenario 10, 16, and 18. Figure 2 shows the strongest disproportionality scenarios for the variables sex, ISCED level and age. In all three scenarios, Opt leads to the smallest negative log-likelihood values on average. With less disproportional sampling, the results show

the same tendencies, although less pronounced. Further details can be found in Tables 4, 5 and 16 in the appendix. There is no systematic bias detectable for the estimates under Opt and LS, and they improve the probability estimates relative to Mod. The consideration of the standard error and impact of all parameters in Opt provides a more efficient and targeted adjustment relative to LS. The increasing standard deviation of the coefficients of the undercovered variable leads to the adjustment of this parameter having less impact on the overall log-likelihood of the model estimated on the sample. Additionally, the disproportionality causes a primarily small adjustment of a parameter to have much influence on the constraint of the population data.

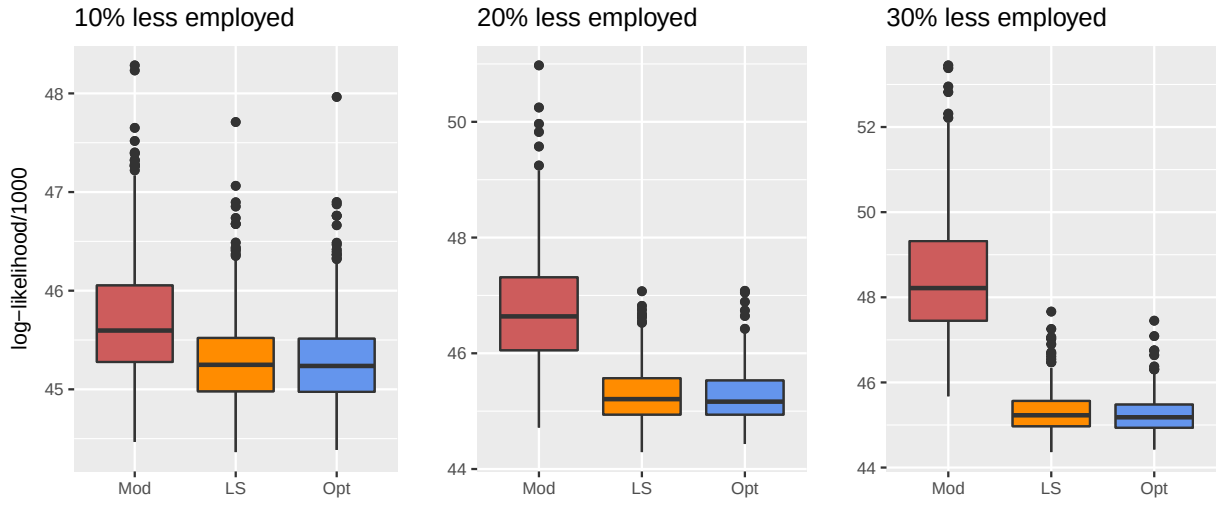

Figure 3: Scenario 2 - 4: negative log-likelihood values, n=500

When coverage problems are with respect to the target variable, the efficiency discrepancies between adjusted and unadjusted estimates become very evident. Unlike the previously discussed disproportional sampling regarding to auxiliary variables, the underlying missing data mechanism now directly depends on the dependent variable Y. Without adjustment, this introduces a considerable bias to transition probability estimation. Figure 3 shows boxplots of the likelihood values based on samples with approx. 10%, 20% and 30% less employed persons (Scenario 2 - 4). These scenarios directly affect the intercept of the logit model, allowing LS to counteract the distortion quite well. Nevertheless, Opt (10%: 45,513, 20%: 45,530, 30% 45,479) displays slightly lower mean values than LS (10%: 45,521, 20%: 45,569, 30% 45,564). The analysis of the squared deviations of the predicted from the actual value is very similar. It can also be shown that the sum of squared errors can be reduced after using Opt and LS. Detailed results can be found in the Tables 7, 8 and 9.

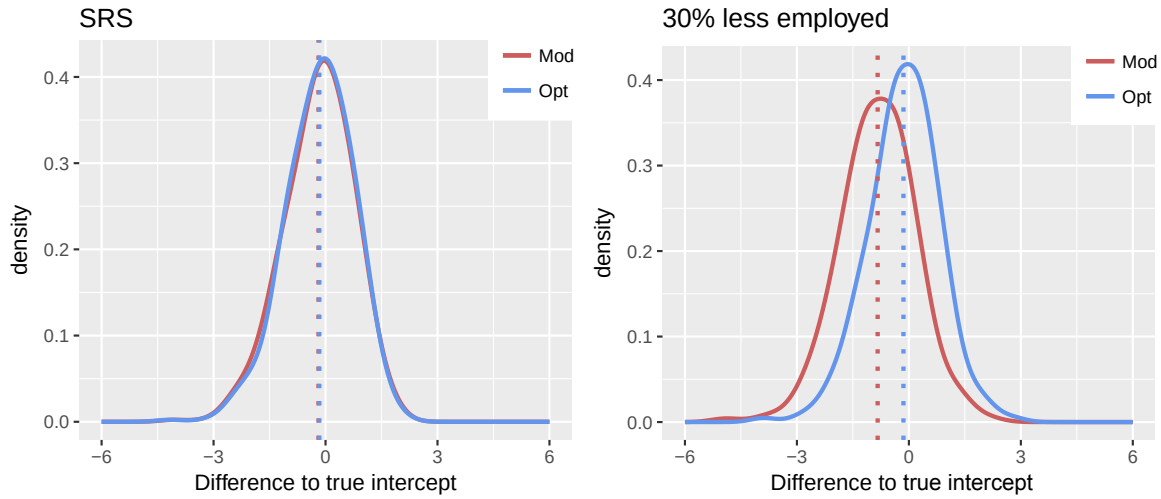

Figure 4: Scenario 1 and 4: intercept estimation, n=500

To consider the distribution of the estimated parameters, the densities of deviations (22) are superimposed for the intercept in Figure 4. On the left density plot, the sample was drawn without restriction (Scenario 1), in the right side employed persons were underdrawn by 30% (Scenario 4). While in the case of SRS no differences can be identified, Opt is able to reliably counteract any distortion of a biased intercept in Scenario 4.

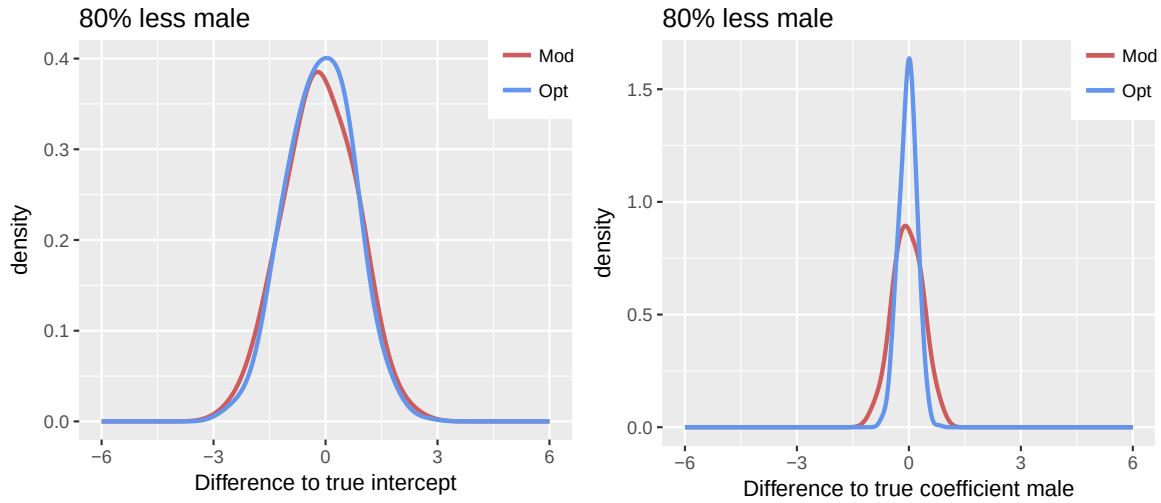

Figure 5: Scenario 10: intercept estimation, n=500

A slightly different picture occurs when undercoverage is with respect to auxiliary variables, as in Figure 5. On the left, we have the intercept estimation under Scenario 4, and on the right, the regression coefficient for males is depicted. We see that both methods are

unbiased in terms of intercept estimation. However, the estimation efficiency with respect to the regression coefficient is considerably increased when accounting for the regional benchmarks. The estimation density of Opt is much more concentrated around 0 compared to Mod in that case.

4.3 Design-Based Results

We now focus on recovering regional transition dynamics from a sample lacking regional identifiers. The probability estimates are produced for the German federal states. In the following the results for Rhineland-Palatinate, Baden-Wuerttemberg, and Bavaria are presented. The results of the remaining federal states for different sample sizes can be found in the Tables 18, 19 and 20 of the appendix. Note that unlike the previous results, no disproportional sampling is conducted. Evaluation of results is again with respect to the log-likelihood values (21).

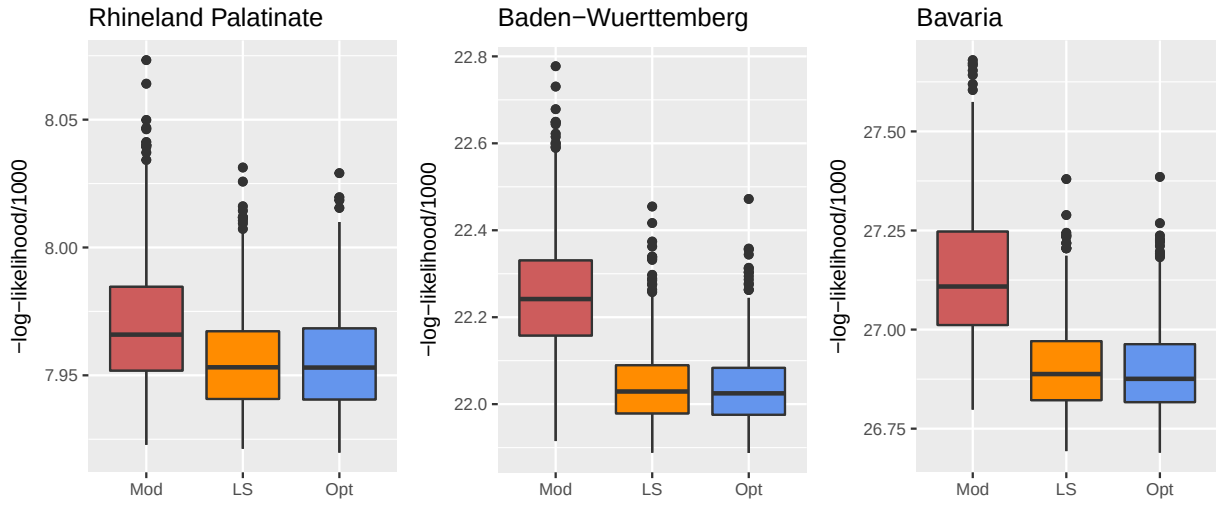

Figure 6: -Log-likelihood Model based simulation, n=2000

In Figure 6, we again see the improvements of LS and Opt relative to Mod. Both methods allow for considerably better goodness of fit with respect to model parameters on the regional population - despite the fact that model parameter estimation was performed on observations from all federal states. Thereby, it is not clear whether LS or Opt obtain the best results. In the model-based setup, Opt was slightly more efficient and generally had a smaller probability of producing estimation outliers (Figure 2). However, in practice, this tendency is not observable anymore. Regarding the design-based results of the undercoverage scenarios of Table 1, see the appendix. The results are essentially identical to the

model-based setup. Therefore, we omitted a dedicated interpretation at this point.

5 Discussion

The estimation of regional transition probabilities from survey data lacking in regional detail has been investigated. Missing regional observations can either lead to coverage problems in local samples or prevent a spatial localization of the sample observations. It could be shown that common estimation methods obtain inefficient or even biased results in these cases. We discussed two methods that are able to account for regional heterogeneity by aligning micro level transition probability estimates to regional benchmarks. Both methods allowed for considerably more efficient estimates in all scenarios. In the light of the coverage problems, the most significant efficiency gains were realized when undercoverage was with respect to the variable transition probability estimates were required for. Regarding the missing regional identifiers, both adjustments were able to recover the transition dynamics more efficiently relative to standard methods.

The findings of this study have major implications for future research conducted by means of dynamic microsimulations. The basic principle of this simulation type is the projection of system units into future periods with micro-level transition probabilities. If these probabilities don't accurately represent regional transition dynamics of the real world, the obtained simulation results are not reliable. In general, both logit scaling and constrained maximum likelihood are suitable to solve this problem. However, from a practical perspective, the researcher may want to choose one of them. One that note, Stephensen (2016) pointed out several criteria that determine a good alignment method in the context of dynamic microsimulations. The subsequent discussion is partially guided by these aspects.

Both techniques adjust the probability estimates such that they are consistent with the regional benchmarks. Thereby, both allow for multinomial alignment in a logit environment, are formulated symmetrically, and retain zero probabilities in the process. A difference is that logit scaling retains the original shape of the probability distribution. In constrained maximum likelihood, this is not the case due to the adjustment of multiple parameters simultaneously. However, our point of view is that depending on the context, it may not be desirable to retain the original shape of the distribution. The simulation study showed that in the case of an undercoverage regarding the target variable, the model parameter estimates are considerably biased. Subsequently, the resulting probability distribution gets distorted and may misrepresent real-world transition dynamics. Another point is the easy and efficient implementation of the methods in the simulation process. In general, the

algorithm used for logit scaling is certainly easier and faster to apply than the sequential quadratic programming approach we propose. But it should be recalled that logit scaling is an ex-post alignment method. Thus, it has to be performed for each period of the simulation horizon individually. The constrained maximum likelihood approach allows for direct alignment in the process of model parameter estimation. Accordingly, it has to be performed only once in the initial phase of the simulation. Further, there are many software packages in which sequential quadratic programming is implemented such that they can be easily applied. Further research should be conducted on the behaviour of adjustment methods when regional benchmarks are not known, but also estimated from survey data. This setting is more likely in the light of public reporting and data sources of official statistics. In principal, the box constraints discussed for the regional benchmarks can be constructed to account for estimation errors, for example in terms of confidence intervals. However, this might not be the optimal choice since the optimization algorithms for alignment will stop when the estimates are consistent with the interval boundaries. Since regional estimates of population characteristics are often subject to high uncertainty due to a lack in regional detail as well, boxes may be very large and the efficiency gains from adjustment decrease.

Acknowledgements

This study was conducted within the research project *Regionale Mikrosimulationen und Indikatorensysteme (REMIKIS)* funded the Nikolaus Koch Foundation, as well as the research group *Sektorenübergreifendes kleinräumiges Mikrosimulationsmodell (MikroSim)* funded by the German Research Foundation. We kindly thank all involved parties for the financial support.

References

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics 44*(1), 197–200.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bishop, Y., S. Fienberg, and P. Holland (1970). *Discrete multivariate analysis: Theory and practice*. MIT Press.

Burgard, J., J. Krause, H. Merkle, R. Münnich, and S. Schmaus (2019). Conducting a dynamic microsimulation for care research: Data generation, transition probabilities and sensitivity analysis. *Springer Proceedings in Mathematics and Statistics*.

Chatterjee, N., Y.-H. Chen, P. Maas, and R. Carroll (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association 111*(513), 107–117.

Chin, S.-F. and A. Harding (2006). Regional dimensions: Creating synthetic small-area micro data and spatial microsimulation models. Technical report, National Centre for Social and Economic Modelling, Canberra.

Darroch, J. N. and D. Ratcliff (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics 43*(5), 1470–1480.

Dong, M. and R.-B. Wets (2000). Estimating density functions: A constrained maximum likelihood approach. *Journal of Nonparametric Statistics 12*(4), 549–595.

Fialka, J., A. Krejdl, and P. Bednarik (2011). Summary based on the final project report of the dynamic microsimulation model of the czech republic. Deloitte.

Greene, W. H. (2002). *Econometric analysis* (5 ed.). Prentice Hall. Upper Saddle River, New Jersey.

Klevmarken, N. A. (2010). Microsimulation for public policy: Experiences from the swedish model sesim. Economics and Social Research Institute Discussion Paper 242. Cabinet Office, Tokyo, Japan.

Kraft, D. (1988). A software package for sequential quadratic programming. Technical Report DVFLR-FB 88-28, Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt.

Kraft, D. (1994). Algorithm 733: Tomp–fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software 20*(3), 262–281.

Lawson, T. (2011). An agent-based model of household spending using a random assignment scheme. Conference Paper, IMA Conference, Stockholm, Sweden.

Li, J. and C. O'Donoghue (2013). A survey of dynamic microsimulation models: Uses, model structure and methodology. *International Journal of Microsimulation 6*(2), 3–55.

Li, J. and C. O'Donoghue (2014). Evaluating binary alignment methods in microsimulation models. *Journal of Artificial Societies and Social Simulation 17*(1), 15.

Markham, F., M. Young, and B. Doran (2017). Improving spatial microsimulation estimates of health outcomes by including geographic indicators of health behaviour: The example of problem gambling. *Health & Place 46*, 29–36.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (2 ed.), Volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall. London, New York.

O'Donoghue, C., J. Lennon, and S. Hynes (2009). The life-cycle income analysis models (liam): A study of flexible dynamic microsimulation modelling computing framework. *International Journal of Microsimulation 2*, 16–31.

O'Donoghue, C., J. Loughrey, and K. Morrissey (2011). Modelling the impact of the economic crisis on inequality in ireland. Conference Paper, IMA Conference, Stockholm, Sweden.

Peichl, A., H. Schneider, and S. Siegloch (2010). Documentation iza mod: The iza policy simulation model. IZA Discussion Paper No. 4865.

Rahman, A. and A. Harding (2016). *Small area estimation and microsimulation modeling*. Chapman and Hall/CRC.

Rahman, A., A. Harding, R. Tanton, S. Liu, et al. (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation 3*(2), 3–22.

Stephensen, P. (2016). Logit scaling: A general method for alignment in microsimulation models. *International Journal of Microsimulation 9*(3), 89–102.

Sutherland, H. and F. Figari (2013). Euromod: The european union tax-benefit microsimulation model. *International Journal of Microsimulation 6*(1), 4–26.

Tanton, R. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation 7*(1), 4–25.

Tanton, R. et al. (2018). Spatial microsimulation: Developments and potential future directions. *International Journal of Microsimulation 11*(1), 143–161.

Willekens, F. (2009). Continuous-time microsimulation in longitudinal analysis. In A. Zaidi, A. Harding, and P. Williamson (Eds.), *New Frontiers in Microsimulation Modelling*, pp. 413–436. Ashgate.

A Appendix

Table 2: Logit Regression for employment status

	Emplyment status: employed
Age	0.386 (0.002)***
Age^	−0.005 (0.00002)***
Male	−0.546 (0.011)***
ISCED 2	1.017 (0.032)***
ISCED 3	1.489 (0.029)***
ISCED 4	1.759 (0.036)***
ISCED 5	2.151 (0.031) ***
Intercept	−6.811 (0.053)***
Observations	271,288
Log Likelihood	−107,899.400
Akaike Inf. Crit.	215,814.900
Note:	*p<0.1; **p<0.05; ***p<0.01

A.1 Model based simulation results

Table 3: -Log-likelihood by scenarios for $n = 1000$, model based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Szen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	44,773	44,729	44,726	44,917	44,871	44,871	44,941	44,887	44,885	45,090	45,039	45,039
2	45,018	44,713	44,711	45,250	44,874	44,875	45,270	44,888	44,884	45,511	45,040	45,037
3	45,819	44,709	44,689	46,235	44,881	44,866	46,271	44,889	44,873	46,606	45,043	45,032
4	47,435	44,707	44,698	47,972	44,886	44,847	48,011	44,902	44,869	48,615	45,059	45,021
5	44,941	44,705	44,710	45,131	44,849	44,848	45,185	44,875	44,873	45,391	45,040	45,039
6	45,673	44,734	44,721	46,005	44,909	44,885	46,037	44,903	44,887	46,355	45,056	45,031
7	46,920	44,746	44,731	47,478	44,932	44,899	47,525	44,943	44,913	48,065	45,095	45,057
8	44,755	44,713	44,702	44,917	44,855	44,845	44,952	44,894	44,881	45,115	45,045	45,037
9	44,780	44,714	44,707	44,961	44,904	44,887	45,001	44,924	44,893	45,185	45,096	45,050
10	44,824	44,753	44,709	45,029	44,931	44,863	45,125	44,978	44,887	45,323	45,151	45,031
11	44,744	44,711	44,710	44,913	44,864	44,864	44,948	44,894	44,890	45,118	45,045	45,036
12	44,788	44,732	44,720	44,962	44,909	44,887	44,987	44,923	44,906	45,159	45,091	45,078
13	44,800	44,748	44,713	45,024	44,964	44,904	45,134	45,043	44,925	45,324	45,220	45,105
14	44,781	44,731	44,735	44,939	44,891	44,875	44,968	44,910	44,902	45,126	45,070	45,055
15	44,773	44,729	44,734	44,956	44,909	44,891	44,982	44,925	44,907	45,152	45,079	45,056
16	44,844	44,787	44,748	45,048	44,986	44,903	45,140	45,051	44,942	45,320	45,207	45,100
17	44,796	44,752	44,727	44,980	44,897	44,853	45,063	44,963	44,884	45,245	45,119	45,018
18	44,876	44,798	44,757	45,073	44,985	44,921	45,142	45,025	44,954	45,337	45,199	45,119

Table 4: -Log-likelihood by scenarios for $n = 500$, model based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Szen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	45,031	44,955	44,957	45,301	45,194	45,183	45,385	45,272	45,266	45,643	45,512	45,492
2	45,278	44,980	44,975	45,596	45,249	45,237	45,709	45,294	45,283	46,053	45,521	45,513
3	46,053	44,941	44,940	46,638	45,208	45,164	46,730	45,287	45,254	47,312	45,569	45,530
4	47,450	44,972	44,937	48,216	45,228	45,181	48,433	45,299	45,236	49,319	45,564	45,479
5	45,246	44,989	44,977	45,607	45,246	45,231	45,668	45,316	45,305	46,063	45,586	45,580
6	45,888	44,955	44,942	46,457	45,201	45,201	46,561	45,289	45,258	47,006	45,550	45,500
7	47,144	44,997	44,975	47,986	45,282	45,203	48,109	45,345	45,284	48,809	45,624	45,543
8	45,070	44,953	44,936	45,368	45,235	45,231	45,424	45,309	45,290	45,713	45,588	45,577
9	45,107	45,007	44,962	45,407	45,269	45,204	45,515	45,343	45,270	45,794	45,596	45,504
10	45,192	45,075	44,993	45,602	45,378	45,214	45,711	45,447	45,260	46,026	45,701	45,506
11	45,080	44,995	44,988	45,344	45,248	45,230	45,409	45,301	45,285	45,675	45,550	45,537
12	45,080	44,986	44,959	45,386	45,250	45,198	45,482	45,345	45,283	45,754	45,593	45,513
13	45,148	45,071	44,973	45,590	45,431	45,258	46,121	45,918	45,352	46,166	45,943	45,613
14	45,053	44,958	44,960	45,353	45,239	45,222	45,429	45,307	45,286	45,724	45,581	45,537
15	45,097	45,007	44,977	45,369	45,259	45,208	45,495	45,364	45,318	45,770	45,618	45,603
16	45,171	45,074	45,016	45,536	45,408	45,291	45,831	45,635	45,371	46,191	45,934	45,604
17	45,134	45,037	44,931	45,477	45,326	45,213	45,684	45,486	45,272	45,902	45,720	45,529
18	45,191	45,060	44,972	45,606	45,396	45,243	45,831	45,541	45,345	46,190	45,833	45,597

Table 5: -Log-likelihood by scenarios for $n = 250$, model based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	45,673	45,515	45,474	46,189	45,961	45,946	46,379	46,132	46,106	46,888	46,611	46,606
2	45,845	45,444	45,416	46,475	45,920	45,898	46,676	46,124	46,075	47,282	46,651	46,549
3	46,599	45,483	45,425	47,516	45,981	45,922	47,688	46,133	46,040	48,470	46,531	46,484
4	48,058	45,472	45,357	49,372	45,973	45,875	49,627	46,173	45,994	50,687	46,640	46,410
5	45,806	45,426	45,427	46,365	45,889	45,861	46,584	46,089	46,041	47,104	46,525	46,468
6	46,427	45,480	45,449	47,206	45,972	45,882	47,494	46,163	46,064	48,322	46,677	46,538
7	47,710	45,561	45,464	48,865	46,056	45,940	49,156	46,341	46,164	50,171	46,877	46,664
8	45,689	45,468	45,419	46,212	46,033	45,961	46,435	46,192	46,174	46,992	46,665	46,578
9	45,727	45,559	45,499	46,329	46,099	45,973	46,664	46,336	46,160	47,311	46,929	46,625
10	45,879	45,611	45,399	46,594	46,203	45,876	47,164	46,542	46,028	47,774	47,040	46,495
11	45,662	45,454	45,415	46,210	45,978	45,928	46,437	46,188	46,109	47,030	46,707	46,611
12	45,712	45,535	45,474	46,356	46,124	46,043	46,717	46,447	46,183	47,136	46,793	46,694
13	45,902	45,663	45,483	46,700	46,354	46,009	51,970	51,498	46,431	47,947	47,581	46,783
14	45,666	45,482	45,459	46,229	45,971	45,926	46,460	46,221	46,153	46,891	46,616	46,530
15	45,648	45,499	45,428	46,265	45,997	45,979	46,642	46,360	46,207	47,113	46,790	46,685
16	45,829	45,691	45,504	46,717	46,392	46,109	49,766	49,346	46,404	47,929	47,545	46,984
17	45,908	45,626	45,443	46,598	46,280	45,879	47,104	46,589	46,078	47,639	47,114	46,563
18	45,992	45,704	45,486	46,918	46,522	46,069	47,632	46,919	46,282	48,389	47,429	46,862

Table 6: Squared prediction error for $n = 1000$, model based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	14,590	14,574	14,572	14,647	14,627	14,625	14,653	14,633	14,632	14,711	14,689	14,689
2	14,684	14,571	14,570	14,770	14,626	14,626	14,781	14,633	14,632	14,873	14,692	14,691
3	14,991	14,569	14,562	15,151	14,631	14,625	15,166	14,634	14,629	15,302	14,690	14,686
4	15,605	14,569	14,562	15,813	14,631	14,624	15,835	14,638	14,627	16,069	14,698	14,684
5	14,656	14,568	14,570	14,718	14,620	14,620	14,736	14,628	14,627	14,811	14,691	14,690
6	14,911	14,579	14,574	15,008	14,637	14,629	15,037	14,639	14,634	15,162	14,695	14,689
7	15,335	14,586	14,578	15,528	14,648	14,637	15,546	14,652	14,642	15,748	14,708	14,698
8	14,584	14,572	14,569	14,647	14,624	14,621	14,657	14,636	14,632	14,713	14,691	14,688
9	14,594	14,570	14,567	14,661	14,637	14,632	14,673	14,645	14,634	14,742	14,709	14,687
10	14,608	14,582	14,567	14,685	14,648	14,625	14,717	14,666	14,632	14,789	14,733	14,684
11	14,584	14,564	14,565	14,642	14,625	14,627	14,658	14,638	14,636	14,714	14,691	14,685
12	14,593	14,575	14,574	14,664	14,641	14,633	14,675	14,650	14,643	14,734	14,711	14,705
13	14,606	14,585	14,567	14,690	14,667	14,640	14,732	14,696	14,650	14,802	14,758	14,716
14	14,596	14,577	14,575	14,648	14,635	14,630	14,663	14,641	14,639	14,722	14,700	14,697
15	14,588	14,574	14,579	14,662	14,644	14,637	14,669	14,647	14,641	14,730	14,701	14,696
16	14,618	14,597	14,585	14,692	14,668	14,641	14,727	14,694	14,654	14,792	14,748	14,712
17	14,597	14,581	14,573	14,665	14,638	14,620	14,687	14,653	14,628	14,747	14,714	14,680
18	14,633	14,605	14,585	14,711	14,675	14,649	14,735	14,689	14,662	14,815	14,755	14,720

Table 7: Squared prediction error for $n = 500$, model based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	14,682	14,656	14,657	14,781	14,743	14,742	14,805	14,764	14,763	14,904	14,854	14,849
2	14,775	14,660	14,662	14,896	14,752	14,753	14,934	14,771	14,768	15,069	14,858	14,851
3	15,072	14,654	14,646	15,291	14,744	14,728	15,325	14,769	14,758	15,524	14,863	14,846
4	15,611	14,662	14,643	15,903	14,754	14,733	15,973	14,774	14,753	16,302	14,872	14,842
5	14,750	14,665	14,665	14,875	14,756	14,753	14,897	14,781	14,778	15,017	14,878	14,876
6	14,979	14,653	14,652	15,160	14,744	14,741	15,193	14,768	14,758	15,350	14,866	14,840
7	15,399	14,672	14,657	15,675	14,764	14,743	15,712	14,789	14,769	15,940	14,885	14,855
8	14,697	14,653	14,649	14,801	14,759	14,754	14,820	14,778	14,772	14,923	14,876	14,876
9	14,716	14,678	14,659	14,822	14,776	14,747	14,851	14,791	14,765	14,953	14,876	14,849
10	14,737	14,696	14,660	14,883	14,797	14,744	14,920	14,828	14,761	15,028	14,930	14,845
11	14,706	14,675	14,672	14,790	14,763	14,755	14,818	14,778	14,772	14,922	14,863	14,846
12	14,703	14,664	14,662	14,810	14,757	14,743	14,847	14,795	14,773	14,942	14,875	14,850
13	14,737	14,700	14,665	14,904	14,832	14,771	14,975	14,905	14,801	15,117	15,009	14,904
14	14,692	14,663	14,654	14,797	14,752	14,748	14,820	14,776	14,769	14,929	14,872	14,864
15	14,709	14,673	14,661	14,801	14,765	14,748	14,844	14,796	14,780	14,946	14,886	14,885
16	14,731	14,696	14,677	14,863	14,816	14,770	14,965	14,897	14,803	15,072	15,002	14,900
17	14,714	14,676	14,649	14,832	14,778	14,742	14,883	14,819	14,758	14,975	14,907	14,849
18	14,741	14,696	14,660	14,896	14,810	14,759	14,975	14,867	14,794	15,117	14,989	14,898

Table 8: Squared prediction error for $n = 250$, model based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	14,895	14,830	14,827	15,096	15,006	14,997	15,129	15,043	15,037	15,315	15,221	15,208
2	14,960	14,827	14,810	15,189	14,987	14,973	15,254	15,040	15,027	15,467	15,207	15,189
3	15,260	14,847	14,816	15,580	14,991	14,973	15,643	15,045	15,018	15,937	15,190	15,189
4	15,795	14,827	14,805	16,290	14,997	14,947	16,354	15,050	14,997	16,749	15,207	15,140
5	14,924	14,812	14,809	15,124	14,974	14,966	15,182	15,030	15,016	15,361	15,161	15,162
6	15,133	14,827	14,806	15,381	15,005	14,980	15,452	15,046	15,018	15,710	15,214	15,178
7	15,559	14,856	14,834	15,869	15,032	14,994	15,963	15,099	15,046	16,336	15,300	15,217
8	14,890	14,824	14,812	15,073	15,013	14,986	15,145	15,061	15,062	15,305	15,241	15,203
9	14,914	14,856	14,831	15,135	15,043	15,004	15,228	15,116	15,057	15,441	15,329	15,249
10	14,979	14,883	14,806	15,209	15,086	14,973	15,375	15,176	15,009	15,600	15,368	15,156
11	14,903	14,832	14,819	15,096	15,006	14,991	15,159	15,072	15,046	15,342	15,249	15,228
12	14,922	14,849	14,838	15,132	15,037	15,016	15,205	15,110	15,072	15,408	15,299	15,253
13	15,006	14,912	14,850	15,283	15,169	15,018	15,531	15,441	15,130	15,718	15,589	15,311
14	14,909	14,833	14,832	15,080	15,008	14,991	15,155	15,072	15,052	15,317	15,209	15,178
15	14,897	14,836	14,825	15,103	15,020	14,996	15,208	15,115	15,067	15,370	15,260	15,233
16	14,960	14,911	14,850	15,267	15,146	15,043	15,501	15,434	15,143	15,632	15,506	15,323
17	14,974	14,890	14,812	15,185	15,077	14,967	15,313	15,150	15,019	15,505	15,314	15,182
18	15,009	14,919	14,849	15,351	15,198	15,055	15,547	15,327	15,110	15,788	15,522	15,308

Table 9: Squared differences to true parameter, n=1000

	Intercept			Age			Age^2			Sex		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	6.994	6.790	6.698	0.332	0.332	0.327	0.004	0.004	0.004	1.787	1.787	1.774
2	8.496	6.838	6.717	0.344	0.344	0.338	0.004	0.004	0.004	1.645	1.645	1.631
3	11.686	6.814	6.357	0.324	0.324	0.305	0.004	0.004	0.004	1.692	1.692	1.596
4	15.592	6.998	6.264	0.351	0.351	0.316	0.004	0.004	0.004	1.420	1.420	1.269
5	7.047	6.769	6.638	0.326	0.326	0.320	0.004	0.004	0.004	1.504	1.504	1.500
6	8.256	6.848	6.575	0.333	0.333	0.320	0.004	0.004	0.004	1.886	1.886	1.788
7	11.925	7.664	6.960	0.392	0.392	0.355	0.005	0.005	0.004	2.086	2.086	2.019
8	6.556	6.353	6.318	0.309	0.309	0.307	0.004	0.004	0.004	1.963	1.963	1.593
9	7.470	7.073	6.962	0.353	0.353	0.350	0.004	0.004	0.004	2.790	2.790	1.820
10	7.879	6.969	6.651	0.343	0.343	0.340	0.004	0.004	0.004	4.762	4.762	1.757
11	6.648	6.539	6.522	0.313	0.313	0.310	0.004	0.004	0.004	1.575	1.575	1.575
12	6.820	6.409	6.279	0.294	0.294	0.292	0.003	0.003	0.003	1.969	1.969	1.952
13	8.436	7.455	6.631	0.295	0.295	0.294	0.003	0.003	0.003	1.774	1.774	1.755
14	6.877	6.819	6.723	0.321	0.321	0.317	0.004	0.004	0.004	1.728	1.728	1.717
15	6.770	6.600	6.586	0.332	0.332	0.329	0.004	0.004	0.004	1.572	1.572	1.579
16	7.043	7.087	6.903	0.353	0.353	0.347	0.004	0.004	0.004	1.686	1.686	1.686
17	5.891	5.524	5.460	0.343	0.343	0.307	0.005	0.005	0.004	1.537	1.537	1.541
18	14.807	13.049	10.792	0.563	0.563	0.509	0.005	0.005	0.005	1.819	1.819	1.806

	Educ2			Educ3			Educ4			Educ5		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	2.513	2.513	2.491	2.619	2.619	2.601	2.644	2.644	2.620	2.921	2.921	2.890
2	2.584	2.584	2.539	2.371	2.371	2.358	2.772	2.772	2.716	2.681	2.681	2.631
3	2.746	2.746	2.612	2.717	2.717	2.589	2.751	2.751	2.628	2.962	2.962	2.767
4	2.867	2.867	2.577	2.613	2.613	2.294	2.677	2.677	2.363	3.082	3.082	2.802
5	2.236	2.236	2.239	2.745	2.745	2.696	2.894	2.894	2.858	3.206	3.206	3.159
6	2.558	2.558	2.430	2.494	2.494	2.401	2.741	2.741	2.655	2.893	2.893	2.737
7	3.261	3.261	2.985	3.211	3.211	3.002	2.989	2.989	2.804	3.475	3.475	3.143
8	2.421	2.421	2.402	2.728	2.728	2.727	2.950	2.950	2.932	2.854	2.854	2.839
9	2.487	2.487	2.464	2.698	2.698	2.660	2.864	2.864	2.837	2.887	2.887	2.850
10	2.699	2.699	2.680	2.766	2.766	2.751	2.659	2.659	2.642	2.869	2.869	2.857
11	3.329	3.329	3.224	2.985	2.985	2.898	3.501	3.501	3.424	3.167	3.167	3.019
12	4.125	4.125	3.649	4.145	4.145	3.746	3.926	3.926	3.477	3.837	3.837	3.368
13	8.360	8.360	4.650	8.118	8.118	4.582	8.695	8.695	4.797	7.803	7.803	4.374
14	3.463	3.463	3.211	2.641	2.641	2.638	3.083	3.083	3.065	2.566	2.566	2.546
15	4.116	4.116	3.664	2.096	2.096	2.093	2.207	2.207	2.192	2.496	2.496	2.488
16	8.439	8.439	5.061	2.210	2.210	2.202	2.688	2.688	2.678	2.412	2.412	2.392
17	2.094	2.094	2.088	2.390	2.390	2.369	2.395	2.395	2.387	2.926	2.926	2.914
18	3.329	3.329	3.285	3.572	3.572	3.519	3.349	3.349	3.339	3.373	3.373	3.348

Table 10: Squared differences to true parameter, n=500

	Intercept			Age			Age^2			Sex		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
19	13.963	13.529	13.232	0.667	0.667	0.650	0.008	0.008	0.008	3.598	3.598	3.573
20	16.888	14.114	13.467	0.728	0.728	0.702	0.009	0.009	0.008	3.683	3.683	3.640
21	20.006	13.084	12.365	0.683	0.683	0.644	0.008	0.008	0.008	3.365	3.365	3.265
22	27.461	16.089	14.663	0.801	0.801	0.721	0.009	0.009	0.008	3.613	3.613	3.299
23	13.601	14.566	14.249	0.741	0.741	0.724	0.008	0.008	0.008	3.809	3.809	3.784
24	13.166	13.723	12.977	0.677	0.677	0.649	0.008	0.008	0.008	3.865	3.865	3.761
25	16.947	14.323	13.798	0.712	0.712	0.673	0.008	0.008	0.008	4.461	4.461	4.056
26	14.308	13.645	13.316	0.681	0.681	0.672	0.008	0.008	0.008	4.310	4.310	3.822
27	14.721	13.767	13.239	0.668	0.668	0.650	0.008	0.008	0.008	5.823	5.823	3.569
28	14.771	13.319	12.759	0.669	0.669	0.651	0.008	0.008	0.008	9.629	9.629	3.366
29	14.706	14.244	13.934	0.682	0.682	0.669	0.008	0.008	0.008	3.875	3.875	3.848
30	16.794	15.409	14.206	0.651	0.651	0.636	0.008	0.008	0.007	3.569	3.569	3.531
31	34.312	30.979	14.351	0.623	0.623	0.613	0.007	0.007	0.007	3.778	3.778	3.721
32	14.582	14.598	14.355	0.763	0.763	0.751	0.009	0.009	0.009	3.761	3.761	3.690
33	14.454	14.164	13.820	0.768	0.768	0.746	0.009	0.009	0.009	3.442	3.442	3.393
34	14.135	14.093	13.664	0.720	0.720	0.701	0.008	0.008	0.008	3.657	3.657	3.650
35	13.625	12.045	11.176	0.804	0.804	0.612	0.012	0.012	0.008	3.142	3.142	3.139
36	30.716	26.108	19.661	1.122	1.122	0.951	0.011	0.011	0.010	3.945	3.945	3.873

	Educ2			Educ3			Educ4			Educ5		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	5.142	5.142	5.100	5.840	5.840	5.811	5.702	5.702	5.610	6.516	6.516	6.400
2	5.509	5.509	5.476	6.054	6.054	5.952	6.005	6.005	5.915	6.522	6.522	6.385
3	5.588	5.588	5.142	6.527	6.527	6.168	6.256	6.256	5.987	6.260	6.260	5.947
4	5.576	5.576	5.251	5.238	5.238	4.857	5.787	5.787	5.351	5.769	5.769	5.267
5	6.211	6.211	6.187	5.838	5.838	5.670	5.671	5.671	5.528	6.746	6.746	6.608
6	5.201	5.201	4.859	6.311	6.311	6.013	6.613	6.613	5.951	6.467	6.467	6.159
7	6.325	6.325	5.989	5.426	5.426	5.107	6.225	6.225	5.893	6.440	6.440	6.149
8	5.601	5.601	5.595	6.233	6.233	6.139	6.257	6.257	6.215	6.205	6.205	6.192
9	5.918	5.918	5.876	5.714	5.714	5.682	5.585	5.585	5.501	5.095	5.095	5.055
10	5.836	5.836	5.777	6.050	6.050	6.055	4.965	4.965	4.927	5.871	5.871	5.879
11	7.444	7.444	7.046	7.565	7.565	7.107	7.443	7.443	7.113	7.760	7.760	7.413
12	10.023	10.023	7.743	9.841	9.841	7.873	10.195	10.195	8.232	10.733	10.733	8.627
13	53.902	53.902	9.967	52.460	52.460	9.302	52.960	52.960	10.560	50.946	50.946	10.068
14	6.995	6.995	6.396	4.662	4.662	4.618	4.927	4.927	4.864	6.040	6.040	5.903
15	9.555	9.555	8.192	4.812	4.812	4.767	5.035	5.035	4.985	5.481	5.481	5.428
16	18.707	18.707	9.705	5.074	5.074	5.041	5.253	5.253	5.196	5.290	5.290	5.223
17	4.665	4.665	4.681	4.696	4.696	4.649	5.909	5.909	5.889	5.372	5.372	5.322
18	6.596	6.596	6.521	6.781	6.781	6.758	6.166	6.166	6.116	6.697	6.697	6.658

Table 11: Squared differences to true parameter, n=250

	Intercept			Age			Age^2			Sex		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
39	44.264	34.040	31.199	1.742	1.742	1.562	0.021	0.021	0.018	7.476	7.476	6.887
40	53.000	33.953	28.593	1.859	1.859	1.532	0.023	0.023	0.019	7.889	7.889	6.991
41	29.803	30.562	29.873	1.586	1.586	1.542	0.019	0.019	0.018	7.987	7.987	7.884
42	31.805	36.103	33.784	1.815	1.815	1.702	0.022	0.022	0.020	7.539	7.539	6.996
43	35.634	39.544	36.429	2.023	2.023	1.855	0.024	0.024	0.022	8.117	8.117	7.249
44	37.921	36.714	36.551	1.851	1.851	1.868	0.022	0.022	0.022	9.798	9.798	8.172
45	38.497	36.148	34.462	1.720	1.720	1.666	0.020	0.020	0.019	12.814	12.814	7.765
46	37.077	32.958	31.044	1.588	1.588	1.512	0.019	0.019	0.018	26.226	26.226	8.169
47	34.677	33.049	31.648	1.549	1.549	1.502	0.018	0.018	0.018	7.649	7.649	7.440
48	47.533	45.144	34.359	1.713	1.713	1.643	0.020	0.020	0.019	8.135	8.135	7.959
49	262.721	242.277	34.845	1.503	1.503	1.472	0.018	0.018	0.017	7.474	7.474	7.372
50	32.781	32.393	31.089	1.658	1.658	1.585	0.020	0.020	0.019	7.752	7.752	7.471
51	34.754	35.621	33.482	1.756	1.756	1.668	0.021	0.021	0.020	8.034	8.034	7.870
52	32.935	34.476	32.065	1.610	1.610	1.543	0.019	0.019	0.019	8.000	8.000	7.843
53	30.238	26.582	24.575	1.823	1.823	1.372	0.028	0.028	0.018	7.093	7.093	7.069
54	84.069	69.243	42.466	2.969	2.969	2.130	0.028	0.028	0.023	8.484	8.484	8.390

	Educ2			Educ3			Educ4			Educ5		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
39	10.968	10.968	10.502	12.490	12.490	12.204	12.307	12.307	11.713	15.358	15.358	14.497
40	12.599	12.599	10.873	12.421	12.421	10.617	12.397	12.397	11.075	14.169	14.169	12.657
41	10.934	10.934	10.568	11.594	11.594	11.316	12.633	12.633	12.047	12.401	12.401	11.930
42	12.135	12.135	11.260	13.013	13.013	12.084	11.929	11.929	11.059	13.832	13.832	12.526
43	12.589	12.589	11.755	16.426	16.426	14.016	15.422	15.422	13.992	16.932	16.932	15.301
44	11.245	11.245	10.976	11.928	11.928	11.701	13.479	13.479	13.503	15.235	15.235	15.177
45	11.895	11.895	11.522	11.674	11.674	11.462	12.748	12.748	12.462	13.885	13.885	13.459
46	11.283	11.283	10.882	11.741	11.741	11.428	12.618	12.618	12.262	13.667	13.667	13.042
47	16.279	16.279	14.318	17.216	17.216	15.116	18.145	18.145	16.284	17.658	17.658	15.598
48	41.024	41.024	17.007	42.612	42.612	17.650	39.018	39.018	16.390	37.072	37.072	17.532
49	590.656	590.656	34.446	564.785	564.785	33.270	554.181	554.181	34.586	541.848	541.848	33.366
50	16.269	16.269	14.963	11.139	11.139	10.739	12.453	12.453	12.119	13.000	13.000	12.743
51	23.342	23.342	18.013	11.144	11.144	10.881	10.977	10.977	10.739	13.314	13.314	12.948
52	231.263	231.263	24.985	10.282	10.282	10.182	11.802	11.802	11.598	11.547	11.547	11.367
53	9.607	9.607	9.389	10.314	10.314	10.225	11.180	11.180	10.957	13.206	13.206	13.029
54	15.082	15.082	14.577	15.968	15.968	15.337	16.107	16.107	15.467	15.231	15.231	14.628

A.2 Design based simulation results

Table 12: -Log-likelihood by scenarios for $n = 4000$, design based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	169,268	169,233	169,234	169,399	169,354	169,354	169,423	169,377	169,378	169,542	169,496	169,497
2	170,390	169,228	169,227	170,725	169,324	169,317	170,756	169,369	169,367	171,094	169,468	169,469
3	174,227	169,232	169,221	174,897	169,339	169,327	174,987	169,375	169,357	175,716	169,475	169,457
4	180,999	169,246	169,230	182,101	169,373	169,350	182,237	169,416	169,376	183,446	169,539	169,483
5	169,957	169,236	169,233	170,242	169,352	169,352	170,294	169,381	169,380	170,604	169,486	169,488
6	172,389	169,249	169,249	173,025	169,366	169,361	173,078	169,407	169,400	173,688	169,511	169,503
7	176,743	169,249	169,241	177,752	169,382	169,381	177,782	169,439	169,420	178,788	169,579	169,543
8	169,291	169,236	169,230	169,427	169,372	169,361	169,467	169,409	169,405	169,604	169,530	169,530
9	169,331	169,282	169,273	169,472	169,390	169,375	169,518	169,445	169,427	169,655	169,558	169,541
10	169,419	169,363	169,332	169,630	169,538	169,456	169,719	169,587	169,502	169,891	169,779	169,650
11	169,277	169,234	169,239	169,413	169,362	169,355	169,456	169,403	169,401	169,576	169,515	169,517
12	169,311	169,263	169,260	169,454	169,400	169,391	169,505	169,449	169,447	169,631	169,553	169,565
13	169,362	169,321	169,300	169,517	169,470	169,436	169,645	169,584	169,543	169,815	169,739	169,684
14	169,304	169,263	169,259	169,447	169,402	169,381	169,496	169,445	169,438	169,648	169,582	169,572
15	169,363	169,308	169,296	169,525	169,466	169,460	169,588	169,524	169,493	169,761	169,683	169,625
16	169,459	169,414	169,365	169,693	169,638	169,542	169,837	169,750	169,621	170,051	169,941	169,782
17	169,698	169,578	169,650	169,990	169,870	169,924	170,059	169,943	169,983	170,340	170,209	170,258
18	170,171	169,871	170,021	170,619	170,243	170,384	170,702	170,358	170,518	171,130	170,694	170,906

Table 13: -Log-likelihood by scenarios for $n = 2000$, design based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
19	169,538	169,451	169,454	169,779	169,682	169,682	169,851	169,752	169,751	170,095	169,968	169,981
20	170,613	169,474	169,466	171,098	169,694	169,674	171,200	169,763	169,754	171,703	169,959	169,961
21	174,313	169,504	169,468	175,248	169,698	169,673	175,444	169,804	169,777	176,525	170,050	170,016
22	181,112	169,472	169,432	182,814	169,723	169,681	182,902	169,780	169,720	184,592	170,013	169,927
23	170,207	169,451	169,446	170,624	169,678	169,682	170,739	169,740	169,736	171,184	169,968	169,955
24	172,528	169,526	169,530	173,329	169,761	169,732	173,487	169,825	169,799	174,286	170,038	169,988
25	176,998	169,545	169,507	178,161	169,826	169,770	178,298	169,891	169,849	179,437	170,119	170,105
26	169,597	169,526	169,510	169,876	169,757	169,737	169,932	169,819	169,799	170,169	170,017	169,981
27	169,640	169,544	169,495	169,936	169,794	169,725	170,013	169,870	169,811	170,247	170,100	170,044
28	169,769	169,648	169,567	170,181	169,995	169,835	170,349	170,084	169,908	170,655	170,386	170,156
29	169,592	169,481	169,485	169,871	169,773	169,765	169,925	169,816	169,818	170,167	170,040	170,043
30	169,614	169,544	169,522	169,948	169,823	169,815	170,048	169,937	169,923	170,301	170,172	170,153
31	169,715	169,602	169,595	170,075	169,904	169,848	170,643	170,508	170,189	170,665	170,492	170,391
32	169,621	169,523	169,502	169,885	169,767	169,745	169,998	169,873	169,856	170,271	170,119	170,082
33	169,700	169,587	169,544	169,973	169,831	169,817	170,152	170,008	169,949	170,438	170,269	170,228
34	169,822	169,694	169,607	170,243	170,119	169,988	170,581	170,384	170,096	171,053	170,713	170,429
35	170,007	169,815	169,841	170,448	170,178	170,261	170,629	170,397	170,379	170,963	170,745	170,744
36	170,471	170,134	170,175	171,102	170,756	170,783	171,391	170,945	170,975	172,048	171,551	171,455

Table 14: Log-likelihood by scenarios for $n = 1000$, design based simulation

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
37	170,179	169,978	169,976	170,683	170,420	170,413	170,813	170,579	170,569	171,318	171,028	171,053
38	171,127	169,986	169,962	171,928	170,404	170,413	172,145	170,568	170,551	172,884	171,001	170,980
39	174,772	169,924	169,875	176,369	170,333	170,313	176,507	170,609	170,530	177,908	171,120	171,001
40	181,433	170,042	169,966	183,687	170,471	170,332	183,974	170,671	170,503	186,434	171,127	170,873
41	170,723	169,959	169,977	171,471	170,444	170,425	171,670	170,606	170,603	172,383	171,065	171,060
42	173,021	170,023	169,954	174,169	170,464	170,429	174,475	170,659	170,627	175,722	171,164	171,173
43	177,364	170,032	170,009	179,191	170,583	170,541	179,433	170,793	170,712	181,307	171,314	171,192
44	170,107	169,916	169,914	170,711	170,464	170,416	170,886	170,617	170,558	171,430	171,065	170,982
45	170,175	169,976	169,970	170,790	170,533	170,420	170,954	170,656	170,563	171,510	171,184	170,990
46	170,420	170,202	170,011	171,229	170,886	170,455	171,646	171,082	170,698	172,364	171,551	171,150
47	170,143	169,997	169,979	170,771	170,571	170,553	170,981	170,770	170,773	171,459	171,242	171,195
48	170,253	170,063	170,033	170,942	170,675	170,661	171,655	171,403	170,993	171,688	171,400	171,408
49	170,401	170,192	170,177	171,298	171,054	170,923	176,931	176,682	171,695	172,655	172,395	172,131
50	170,186	170,026	169,989	170,731	170,473	170,465	170,930	170,710	170,682	171,331	171,160	171,094
51	170,280	170,071	170,071	170,955	170,734	170,596	171,167	170,914	170,803	171,867	171,572	171,358
52	170,589	170,327	170,134	171,425	171,058	170,758	172,046	171,642	171,100	172,812	172,328	171,719
53	170,511	170,236	170,141	171,326	170,922	170,815	171,770	171,250	171,012	172,487	171,773	171,602
54	171,412	170,951	170,830	172,401	171,820	171,658	172,830	172,157	172,015	173,815	173,053	172,867

Table 15: Squared prediction Error, design based simulation, n=4000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	53,560	53,549	53,550	53,613	53,597	53,597	53,630	53,615	53,615	53,685	53,664	53,664
2	54,069	53,561	53,561	54,205	53,607	53,603	54,232	53,623	53,621	54,372	53,665	53,663
3	55,611	53,566	53,564	55,912	53,621	53,611	55,920	53,634	53,627	56,213	53,683	53,676
4	58,260	53,582	53,572	58,736	53,640	53,627	58,778	53,657	53,639	59,236	53,716	53,688
5	53,719	53,553	53,553	53,818	53,593	53,593	53,831	53,606	53,605	53,926	53,646	53,645
6	54,462	53,548	53,548	54,643	53,592	53,589	54,664	53,604	53,602	54,863	53,646	53,641
7	55,801	53,548	53,547	56,094	53,594	53,590	56,100	53,611	53,602	56,364	53,652	53,640
8	53,579	53,562	53,561	53,632	53,612	53,611	53,647	53,627	53,626	53,698	53,674	53,670
9	53,591	53,575	53,570	53,655	53,625	53,621	53,669	53,645	53,637	53,727	53,697	53,685
10	53,625	53,605	53,595	53,701	53,674	53,646	53,738	53,693	53,660	53,813	53,769	53,710
11	53,566	53,556	53,555	53,623	53,604	53,603	53,632	53,615	53,615	53,680	53,661	53,662
12	53,585	53,570	53,569	53,639	53,617	53,617	53,655	53,635	53,634	53,705	53,678	53,678
13	53,593	53,582	53,573	53,663	53,641	53,628	53,700	53,679	53,665	53,768	53,732	53,720
14	53,576	53,561	53,561	53,630	53,614	53,609	53,641	53,625	53,623	53,686	53,669	53,666
15	53,586	53,571	53,569	53,644	53,631	53,624	53,669	53,648	53,639	53,724	53,700	53,691
16	53,615	53,605	53,586	53,695	53,674	53,646	53,747	53,718	53,676	53,830	53,788	53,745
17	53,789	53,730	53,759	53,897	53,840	53,872	53,926	53,872	53,894	54,037	53,977	53,999
18	53,739	53,698	53,784	53,852	53,809	53,930	53,898	53,859	53,971	54,006	53,975	54,128

Table 16: Squared prediction error, design based simulation, n=2000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	53,653	53,632	53,633	53,745	53,714	53,714	53,771	53,734	53,734	53,860	53,814	53,811
2	54,137	53,634	53,635	54,338	53,716	53,713	54,384	53,749	53,745	54,574	53,833	53,829
3	55,612	53,658	53,638	55,993	53,734	53,727	56,081	53,770	53,760	56,531	53,851	53,849
4	58,335	53,654	53,636	59,019	53,745	53,736	59,036	53,778	53,752	59,689	53,875	53,841
5	53,795	53,619	53,618	53,932	53,704	53,700	53,966	53,720	53,719	54,100	53,802	53,798
6	54,478	53,641	53,638	54,740	53,720	53,715	54,777	53,743	53,735	55,035	53,813	53,805
7	55,793	53,644	53,631	56,221	53,733	53,716	56,233	53,754	53,739	56,622	53,833	53,814
8	53,674	53,653	53,648	53,778	53,736	53,729	53,800	53,761	53,754	53,902	53,839	53,829
9	53,690	53,660	53,642	53,794	53,744	53,728	53,829	53,781	53,758	53,925	53,878	53,845
10	53,745	53,701	53,660	53,893	53,817	53,775	53,950	53,855	53,795	54,071	53,970	53,891
11	53,666	53,634	53,636	53,760	53,724	53,731	53,787	53,751	53,752	53,885	53,843	53,843
12	53,684	53,660	53,648	53,788	53,754	53,745	53,835	53,797	53,793	53,942	53,888	53,883
13	53,709	53,669	53,660	53,832	53,784	53,768	53,919	53,870	53,852	54,041	53,996	53,954
14	53,673	53,646	53,639	53,764	53,730	53,723	53,806	53,764	53,759	53,901	53,846	53,842
15	53,683	53,652	53,644	53,790	53,751	53,739	53,844	53,796	53,778	53,940	53,887	53,873
16	53,727	53,696	53,678	53,872	53,820	53,780	53,976	53,915	53,822	54,112	54,032	53,920
17	53,891	53,792	53,816	54,077	53,957	53,991	54,135	54,031	54,035	54,280	54,169	54,187
18	53,833	53,784	53,851	54,041	53,983	54,029	54,119	54,044	54,100	54,283	54,211	54,278

Table 17: Squared prediction error, design based simulation, n=1000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Scen	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	53,840	53,783	53,779	54,010	53,930	53,924	54,056	53,980	53,978	54,240	54,156	54,151
2	54,330	53,810	53,804	54,625	53,961	53,964	54,710	54,008	54,004	54,984	54,144	54,160
3	55,730	53,793	53,779	56,356	53,950	53,919	56,433	54,023	53,996	57,019	54,177	54,145
4	58,366	53,834	53,811	59,285	54,012	53,952	59,374	54,062	54,005	60,266	54,226	54,154
5	53,973	53,788	53,791	54,186	53,929	53,926	54,242	53,990	53,990	54,439	54,137	54,127
6	54,609	53,769	53,758	54,997	53,944	53,942	55,040	53,992	53,984	55,390	54,152	54,146
7	55,856	53,807	53,790	56,420	53,975	53,963	56,497	54,034	54,008	57,012	54,188	54,158
8	53,835	53,776	53,769	54,062	53,960	53,951	54,114	54,023	54,002	54,307	54,201	54,182
9	53,877	53,802	53,800	54,088	53,989	53,964	54,140	54,038	54,003	54,346	54,210	54,158
10	53,982	53,881	53,821	54,259	54,095	53,986	54,391	54,192	54,059	54,638	54,393	54,215
11	53,843	53,798	53,795	54,051	53,982	53,984	54,127	54,054	54,055	54,325	54,231	54,218
12	53,898	53,831	53,819	54,131	54,035	54,028	54,199	54,112	54,101	54,379	54,304	54,295
13	53,955	53,875	53,866	54,250	54,173	54,121	54,443	54,374	54,250	54,679	54,588	54,483
14	53,858	53,800	53,788	54,042	53,973	53,962	54,098	54,025	54,017	54,264	54,170	54,158
15	53,867	53,811	53,809	54,099	54,018	53,998	54,171	54,092	54,056	54,366	54,286	54,232
16	53,966	53,893	53,823	54,238	54,138	54,041	54,438	54,310	54,138	54,698	54,538	54,348
17	54,045	53,911	53,899	54,371	54,177	54,151	54,536	54,310	54,246	54,829	54,532	54,501
18	54,129	54,029	54,032	54,402	54,303	54,289	54,541	54,412	54,403	54,812	54,662	54,685

Table 18: -Log-likelihood by federal states, design based simulation, n=4000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Fed. State	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	5,846	5,844	5,844	5,852	5,850	5,850	5,862	5,860	5,859	5,855	5,853	5,852
2	3,357	3,356	3,357	3,365	3,364	3,365	3,376	3,373	3,375	3,367	3,366	3,367
3	16,586	16,580	16,581	16,599	16,593	16,592	16,620	16,609	16,608	16,604	16,596	16,596
4	1,543	1,533	1,532	1,548	1,537	1,536	1,554	1,542	1,541	1,549	1,538	1,537
5	35,358	35,266	35,263	35,426	35,310	35,304	35,497	35,368	35,361	35,437	35,321	35,318
6	13,107	13,105	13,105	13,120	13,115	13,115	13,133	13,127	13,127	13,122	13,118	13,118
7	7,938	7,930	7,930	7,946	7,937	7,936	7,957	7,945	7,944	7,949	7,938	7,938
8	22,127	21,955	21,954	22,177	21,987	21,983	22,237	22,024	22,021	22,187	21,994	21,991
9	26,983	26,802	26,797	27,054	26,850	26,841	27,135	26,899	26,892	27,066	26,857	26,848
10	2,169	2,167	2,166	2,173	2,170	2,170	2,178	2,175	2,174	2,174	2,171	2,171
11	7,811	7,660	7,678	7,842	7,676	7,696	7,873	7,692	7,718	7,844	7,678	7,700
12	4,950	4,941	4,944	4,966	4,957	4,960	4,981	4,973	4,976	4,966	4,958	4,960
13	3,276	3,242	3,244	3,287	3,248	3,250	3,296	3,254	3,256	3,287	3,249	3,251
14	8,443	8,410	8,416	8,463	8,427	8,435	8,486	8,446	8,453	8,467	8,430	8,436
15	4,861	4,826	4,826	4,877	4,838	4,838	4,894	4,850	4,850	4,878	4,839	4,839
16	4,660	4,656	4,655	4,670	4,666	4,665	4,683	4,678	4,676	4,672	4,667	4,666

Table 19: -Log-likelihood by federal states, design based simulation, n=2000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Fed. State	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	5,856	5,853	5,853	5,867	5,862	5,862	5,883	5,879	5,879	5,870	5,866	5,866
2	3,362	3,360	3,361	3,374	3,372	3,373	3,390	3,387	3,388	3,377	3,375	3,376
3	16,614	16,601	16,601	16,639	16,625	16,626	16,675	16,662	16,659	16,648	16,634	16,633
4	1,545	1,535	1,534	1,552	1,542	1,540	1,561	1,548	1,547	1,553	1,542	1,541
5	35,417	35,317	35,321	35,510	35,394	35,393	35,621	35,491	35,484	35,533	35,414	35,409
6	13,129	13,122	13,122	13,148	13,141	13,141	13,175	13,166	13,167	13,156	13,148	13,148
7	7,952	7,941	7,941	7,966	7,953	7,953	7,985	7,967	7,968	7,970	7,956	7,956
8	22,158	21,979	21,976	22,242	22,029	22,025	22,331	22,089	22,083	22,254	22,044	22,039
9	27,012	26,822	26,817	27,109	26,888	26,876	27,247	26,971	26,963	27,136	26,907	26,897
10	2,172	2,169	2,168	2,178	2,175	2,174	2,185	2,181	2,180	2,179	2,176	2,175
11	7,810	7,670	7,685	7,856	7,694	7,715	7,899	7,719	7,746	7,860	7,697	7,718
12	4,954	4,945	4,948	4,974	4,965	4,967	4,994	4,986	4,990	4,976	4,967	4,970
13	3,276	3,244	3,246	3,290	3,253	3,255	3,305	3,263	3,264	3,292	3,254	3,256
14	8,448	8,417	8,423	8,477	8,441	8,448	8,508	8,466	8,475	8,481	8,444	8,451
15	4,863	4,831	4,832	4,882	4,845	4,846	4,905	4,863	4,862	4,886	4,847	4,848
16	4,665	4,660	4,660	4,677	4,672	4,672	4,693	4,688	4,687	4,680	4,675	4,674

Table 20: -Log-likelihood by federal states, design based simulation, n=1000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Fed. State	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	5,877	5,869	5,869	5,899	5,890	5,890	5,928	5,919	5,918	5,905	5,897	5,896
2	3,375	3,369	3,368	3,394	3,389	3,390	3,418	3,411	3,413	3,398	3,393	3,394
3	16,675	16,655	16,653	16,729	16,702	16,700	16,794	16,767	16,766	16,741	16,716	16,714
4	1,550	1,539	1,539	1,563	1,549	1,548	1,577	1,562	1,559	1,565	1,551	1,549
5	35,543	35,435	35,432	35,736	35,571	35,559	35,948	35,757	35,731	35,780	35,608	35,596
6	13,180	13,167	13,166	13,221	13,204	13,202	13,272	13,251	13,249	13,233	13,214	13,213
7	7,978	7,964	7,963	8,005	7,988	7,988	8,042	8,017	8,018	8,013	7,995	7,994
8	22,215	22,058	22,049	22,345	22,134	22,124	22,480	22,233	22,224	22,364	22,157	22,151
9	27,073	26,901	26,893	27,220	27,003	26,981	27,395	27,118	27,102	27,255	27,027	27,016
10	2,178	2,174	2,174	2,190	2,184	2,183	2,202	2,195	2,193	2,192	2,186	2,185
11	7,839	7,691	7,709	7,910	7,726	7,749	7,985	7,778	7,803	7,920	7,738	7,759
12	4,963	4,950	4,953	4,995	4,982	4,985	5,031	5,014	5,017	5,000	4,985	4,988
13	3,285	3,251	3,253	3,305	3,264	3,266	3,331	3,279	3,281	3,310	3,267	3,269
14	8,472	8,429	8,437	8,512	8,468	8,475	8,568	8,514	8,520	8,524	8,476	8,484
15	4,874	4,835	4,839	4,905	4,861	4,862	4,941	4,887	4,887	4,911	4,865	4,866
16	4,673	4,667	4,666	4,696	4,686	4,686	4,724	4,711	4,709	4,701	4,691	4,690

Table 21: Squared prediction error by federal states, design based simulation, n=4000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Fed. State	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	1,850	1,849	1,849	1,853	1,851	1,851	1,856	1,855	1,855	1,853	1,852	1,852
2	1,056	1,056	1,056	1,059	1,059	1,059	1,062	1,062	1,062	1,059	1,059	1,060
3	5,240	5,237	5,237	5,246	5,242	5,242	5,255	5,248	5,248	5,248	5,243	5,243
4	498	496	496	500	497	497	502	498	499	500	497	497
5	11,336	11,317	11,316	11,354	11,331	11,331	11,377	11,350	11,347	11,359	11,335	11,335
6	4,127	4,127	4,126	4,131	4,131	4,131	4,137	4,136	4,136	4,133	4,132	4,132
7	2,519	2,516	2,516	2,523	2,519	2,518	2,527	2,522	2,521	2,524	2,519	2,519
8	6,947	6,880	6,879	6,969	6,891	6,890	6,992	6,906	6,904	6,972	6,894	6,892
9	8,443	8,372	8,369	8,474	8,390	8,387	8,505	8,416	8,415	8,477	8,395	8,393
10	695	694	694	696	695	695	698	697	696	697	696	695
11	2,520	2,482	2,489	2,530	2,488	2,495	2,539	2,495	2,504	2,530	2,489	2,497
12	1,541	1,539	1,540	1,545	1,544	1,544	1,550	1,548	1,549	1,546	1,544	1,545
13	1,052	1,045	1,045	1,055	1,047	1,047	1,059	1,049	1,049	1,056	1,047	1,047
14	2,661	2,655	2,657	2,667	2,660	2,662	2,673	2,665	2,667	2,668	2,661	2,662
15	1,541	1,533	1,533	1,545	1,536	1,536	1,550	1,539	1,540	1,546	1,536	1,537
16	1,459	1,459	1,458	1,463	1,462	1,461	1,467	1,466	1,465	1,463	1,462	1,462

Table 22: Squared prediction error by federal states, design based simulation, n=2000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Fed. State	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	1,853	1,852	1,852	1,858	1,855	1,855	1,863	1,860	1,860	1,858	1,856	1,856
2	1,057	1,057	1,057	1,062	1,061	1,062	1,068	1,067	1,067	1,063	1,062	1,063
3	5,249	5,244	5,244	5,260	5,252	5,252	5,274	5,263	5,263	5,263	5,255	5,255
4	499	497	496	501	498	498	504	500	500	501	499	499
5	11,354	11,334	11,333	11,383	11,359	11,359	11,419	11,392	11,391	11,391	11,366	11,366
6	4,134	4,132	4,133	4,141	4,139	4,139	4,151	4,148	4,148	4,144	4,141	4,141
7	2,524	2,519	2,519	2,529	2,524	2,524	2,536	2,529	2,528	2,531	2,525	2,524
8	6,956	6,886	6,885	6,987	6,903	6,901	7,024	6,924	6,923	6,993	6,908	6,907
9	8,450	8,374	8,371	8,488	8,402	8,400	8,545	8,442	8,438	8,499	8,409	8,406
10	696	695	695	698	697	697	700	699	698	698	697	697
11	2,521	2,487	2,491	2,534	2,495	2,502	2,547	2,504	2,513	2,535	2,496	2,503
12	1,543	1,541	1,542	1,548	1,547	1,548	1,555	1,553	1,553	1,549	1,547	1,548
13	1,053	1,045	1,046	1,057	1,048	1,049	1,061	1,051	1,052	1,057	1,049	1,049
14	2,663	2,657	2,659	2,671	2,665	2,667	2,681	2,673	2,674	2,673	2,666	2,667
15	1,541	1,534	1,535	1,547	1,539	1,539	1,554	1,543	1,544	1,548	1,539	1,540
16	1,461	1,460	1,460	1,466	1,464	1,464	1,471	1,469	1,469	1,466	1,465	1,465

Table 23: Squared prediction error by federal states, design based simulation, n=1000

	0.25 quantile			0.5 quantile			0.75 quantile			mean		
Fed. State	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt	Mod	LS	Opt
1	1,859	1,857	1,857	1,867	1,864	1,864	1,876	1,872	1,872	1,869	1,865	1,865
2	1,061	1,059	1,060	1,067	1,066	1,067	1,076	1,074	1,075	1,069	1,068	1,068
3	5,266	5,259	5,258	5,286	5,273	5,273	5,310	5,296	5,297	5,290	5,279	5,278
4	501	498	498	504	501	501	508	505	504	505	501	501
5	11,394	11,371	11,372	11,449	11,422	11,417	11,518	11,470	11,469	11,466	11,428	11,426
6	4,149	4,144	4,144	4,162	4,157	4,157	4,181	4,173	4,173	4,167	4,161	4,161
7	2,532	2,526	2,526	2,541	2,534	2,533	2,554	2,544	2,544	2,544	2,537	2,536
8	6,971	6,906	6,906	7,013	6,930	6,928	7,065	6,963	6,962	7,023	6,938	6,936
9	8,462	8,389	8,388	8,515	8,427	8,427	8,591	8,481	8,479	8,532	8,441	8,438
10	698	697	697	702	700	699	705	703	702	702	700	700
11	2,531	2,493	2,499	2,549	2,507	2,514	2,572	2,524	2,532	2,552	2,510	2,517
12	1,546	1,544	1,544	1,555	1,553	1,554	1,566	1,562	1,563	1,557	1,554	1,555
13	1,055	1,048	1,049	1,061	1,052	1,052	1,069	1,057	1,057	1,063	1,053	1,054
14	2,670	2,663	2,665	2,682	2,674	2,676	2,701	2,686	2,689	2,686	2,676	2,678
15	1,545	1,538	1,538	1,554	1,544	1,545	1,567	1,552	1,552	1,557	1,546	1,546
16	1,464	1,462	1,462	1,471	1,469	1,468	1,480	1,477	1,476	1,473	1,470	1,470