ERASMUS UNIVERSITEIT ROTTERDAM

Forecasting Household Income Distribution using a Bayesian Non-Homogeneous Infinite Switching State Space Model

Lisa Meijer

Student ID: 431906
Supervisor: Dr. Andreas Pick
Second assessor: Prof. Dr. Richard Paap

A Master Thesis Econometrics submitted in partial fulfilment for the degree of
Master of Science in Econometrics

at the

Erasmus School of Economics
Erasmus University Rotterdam

Saturday 22nd April, 2023

Abstract

In this study we develop a Bayesian non-homogeneous switching state space model to forecast the annual income distribution of households living in the Netherlands. This Bayesian model captures the time-varying hidden group structure among households by letting the households' hidden group membership follow a non-homogeneous hidden Markov model. This transition model uses either the multinomial logistic regression model or the logistic stick-breaking process to link the household characteristics at a certain point in time to the hidden clusters, and sample efficiently from their posterior distribution by using the the Pólya-gamma data augmentation technique. Moreover, our model uses a separate state space model for each cluster to capture gradually changing income processes within a cluster, and to deal with households entering the dataset in the future by setting their previous earnings to the general income of the cluster. To estimate the model parameters and to compute the h-step ahead distribution forecast, we have developed and implemented a posterior Markov chain Monte Carlo sampling algorithm, the block Gibbs sampling, that iterates between sampling the hidden clusters, the Gaussian states, and the model parameters. Our results demonstrate that the use of these time-varying clusters and time varying parameters notably improves the forecasting performance. Our models are more accurate at forecasting the one-step-ahead income distribution than a single state space model or a non-homogeneous hidden Markov model, even with households entering the dataset in the forecasting year itself. Especially, our models with ten clusters obtain better results than the models with fewer clusters.

Contents

Chapter 1

Introduction

Forecasting the annual income distribution plays an important role in statistically summarising future earnings, in studying economic inequality, and in being an input into studies of for instance forecasting consumption and savings (Altonji et al., 2022). Therefore, in this thesis we develop a Bayesian non-homogeneous switching state space model to forecast the annual income distribution of households living in the Netherlands. This model tries to capture the time-varying hidden group structure among the households by either using a parametric or non-parametric non-homogeneous Hidden Markov Model and uses a state space model for each individual cluster to model the clusters' overall income and income change over the years. As a result, we are able to model highly non-linear patterns in the income distribution; to account for a large increase or decrease in household earnings due to ageing, divorce or job loss; and to capture gradually changing income processes by using time-varying model parameters. Furthermore, our model corrects the income distribution forecasts for future changes in the Dutch population by utilizing a population forecasting model, as these changes also affect the future income distributions. For example, if we expect to have more low-income elderly people relative to other groups (e.g. young adults), the average income in the Netherlands will decline.

Many existing Bayesian approaches for income dynamics assume a unique fixed effect to incorporate heterogeneity among households (Gu and Koenker, 2017; Arellano et al., 2017). However, we cannot compute a unique fixed effect for each household individually, because households can leave and enter our panel dataset at any time. They might even leave forever due to death or emigration, or they might enter to the dataset sometime in the future. Moreover, our household earnings panel features a large cross-sectional dimension N but short time series T. A unique fixed effect will introduce a tremendous number of parameters, and in short panel data models these fixed effect estimators suffer from the "incidental parameters" problem (Neyman and Scott, 1948) leading to inaccurate estimates and unreliable forecasts. To address these problems, we divide the households into a finite number of groups with an unknown group structure in advance, and we let households within a group share the same parameters, as we expect that these households will have equivalent unit-specific parameters. Hence, our method tries to optimise the unknown time-varying group structure between households as well as the parameters associated with these groups simultaneously.

Still, much research on these grouped fixed-effects models assumes that individual group membership does not vary over time (Bonhomme and Manresa, 2015), while households tend to transition between groups. They might divorce, lose their jobs, or become older. Thus, instead of modeling cluster memberships as being time independent, we model them as first order Markov process by using a discrete-state Hidden Markov model (HMM). A HMM models a sequence of observations that are drawn conditionally on a fixed number of discrete hidden states, which were generated by a first-order Markov process (Rabiner, 1989). The HMM can thus be seen as a time-dependent clustering technique where the hidden states represent the cluster memberships and the Markov process the temporal dependence. However, the time-homogeneity of the standard HMM can be limiting, as the transition probabilities may vary over time and might differ between individual households. Therefore, we relax this assumption and allow the transition probabilities to be dependent on exogenous variables (e.g. age, education level), resulting in a non-homogeneous Hidden Markov model (NHMM). In a NHMM, we connect the

transition probabilities with the coefficients of the exogenous variables through a link function, typically the multinomial logistic (MNL) or the multinomial probit (MNP). We have chosen to use MNL over MNP because MNL has a more straight-forward interpretation of the parameters than MNP. Moreover, it has an efficient conjugate sampling scheme of the MNL parameters, the Pólya-Gamma data augmentation method, introduced by Polson et al. (2013). This enables handling much larger datasets. It also enables adding a global local shrinkage prior, in particular the horseshoe prior (Carvalho et al., 2009), on the coefficients to deal with sparsity in our exogenous variables (Uddin and Gaskins, 2023).

Nevertheless, the overall household earnings tend to slightly change every year, and thus a change in income should not always lead to a change in cluster membership. Otherwise, low-income households will slowly shift towards high-income clusters resulting in empty low-income clusters. Besides, current earnings will generally depend on previous earnings when a household has not experienced any life event, such as a divorce or job loss. Consequently, we model the household earnings as a local level trend model. And, instead of only using a HMM we now use a switching state space model (SSSM) (i.e. switching linear dynamical system) to model the household earnings. In a SSSM, the parameters of the state space model (SSM) switch according to the discrete hidden states of a HMM (Hamilton, 1990). It has an advantage over auto-regressive panel data models, since we can straightforwardly cope with missing observations by letting their previous earnings follow the general income of the cluster. Additionally, a SSM can be regarded as a regression model with time-varying regression coefficients, and hence, it enables us to capture income processes which gradually change over time. Despite these advantages the exact inference of the SSSM is intractable, and we require approximate inference algorithms. Therefore, we use blocked Gibbs sampling to sample the Gaussian states and hidden states sequentially. Still, a SSSM typically assumes that the number of groups is a known and fixed quantity, and specifying this number in advance has some significant drawbacks. Therefore, we also propose a non-parametric switching state space model that uses a hierarchical logistic stick-breaking process HMM (Teh et al., 2006; Ren et al., 2011) to switch between the infinite hidden states, and that models each row of the infinite state transition matrix with a logistic stick-breaking hierarchical Dirichlet process prior (Rigon and Durante, 2021; Linderman et al., 2015).

In summary, our study examines various approaches to forecast the h-step-ahead annual household income distribution. We look at a non-homogeneous HMM, a non-homogeoneous SSSM, and a non-parametric non-homogeoneous SSSM, for which their hidden states follow a non-homogeneous first-order Markov process using either the multinomial logistic regression model or the logistic stick-breaking process. Hence, households can transition between a countably (in)finite number of clusters according to their characteristics at that particular point in time, and the household earnings evolve by letting the cluster parameters follow a local level trend model. Thus, the common income of the clusters might alter over time and may thereby prevent certain clusters from emptying out. To estimate the model parameters and to compute the h-step ahead distribution forecast, we have developed and implemented a posterior Markov chain Monte Carlo (MCMC) sampling algorithm, the block Gibbs sampling, that iterates between sampling the hidden states, the Gaussian states, and the model parameters. However, due to time restrictions, we use a "weak limit" approximation of the non-parametric SSSM that truncates the number of clusters to a fixed truncation level (Fox et al., 2011a).

Our results show that non-homogeneous SSSMs (NSSSMs) are more accurate at forecasting the income distribution than a single state space model according to the Anderson Darling test statistic. This result demonstrates that the use of clusters improves the forecasting performance.

In addition, our results show that our NSSSMs are better at forecasting the distribution than non-homogeneous hidden Markov Models, even with households entering the dataset in the forecasting year itself, so their previous earnings were unknown. Although we truncate the non-parametric NSSSM to 10 clusters, our weak limit non-parametric models still perform comparably and sometimes even better than parametric NSSSMs with 10 clusters. The models forecast the distribution less accurately if they use fewer clusters. Six clusters can still obtain good results, but in that case it is necessary to pre-define the households based on their household composition and age.

1.1 Prior Work

In this section we give a concise overview of some prior work related to our thesis, as our study relates to several disciplines in the literature. Firstly, it relates to the literature on univariate models of earnings dynamics, which already dates back to early contributions such as Lillard and Weiss (1979) and MaCurdy (1982). Although prior literature has primarily focused on linear ARMA-type time series models, recent literature also looks at non-linear earnings processes, and examines the degree of heterogeneity in the parameters across individuals, mainly to allocate the total error variances into transitory and permanent components (Fernández-Val et al., 2022). Geweke and Keane (2000) developed a Bayesian model to consider non-normal shocks using a mixture of three normal distributions. Arellano et al. (2017) specified an age-dependent non-linear earnings process that separately identifies the distributions of the persistent and the transitory components by using a quantile-based panel data method. Hu et al. (2019) developed a semi-parametric state space model by modelling the persistent component through a unit root process and the transitory component through a semi-parametric model of a higher-order ARMA process. However, prior work on for example consumption and savings uses family earnings (or household earnings) and not individual earnings as their input (Altonji et al., 2022). Only a few studies consider household earnings. Altonji et al. (2021) described an econometric model of earnings, marriage, and family income by modeling marital transitions. Additionally, most of these models mentioned primarily focused on directly modeling the transitory and permanent components of individual income shocks, while our model indirectly models the transitory component through cluster transitions and permanent component through a local linear trend. Besides, they are aimed at providing statistical insights about the annual earnings, and are not designed for forecasting an overall income distribution.

Consequently, this thesis also contributes to the sparse panel forecast literature. Wang et al. (2019) estimated the slope parameters of panel data regressions by using a pooling averaging method. They achieved an optimal bias-variance trade-off by combining the estimators from different pooling specifications with appropriate weights. But, their approach assumed regression with a long-time dimension, while we need a method for a very short time dimension. Liu et al. (2020), for instance, suggested an empirical Bayes predictor that uses the cross-sectional information in the panel and Tweedie's formula (Robbins, 1992) to construct a prior distribution either parametrically or non-parametrically. This distribution can subsequently be used to form a posterior mean predictor for each cross-sectional unit. Liu (2022) proposed a dynamic linear panel data model that draws the individual effect from a Bayesian semi-parametric distribution. This distribution allows for correlation between the heterogeneous parameters and the initial conditions utilising the probit stick-breaking process prior (Rodriguez and Dunson, 2011). However, many of these panel data models assume that the households are observable the full time span, and they do not deal with observations entering the dataset at a later point

in time. Therefore, our work also relates to the literature on clustering in panel data models, as we assign such unobserved observations to a matching cluster. One strand of methods to estimate these latent groups addresses adaptations of the k-means clustering technique (Lin and Ng, 2012; Bonhomme and Manresa, 2015; Bonhomme et al., 2022). These methods iterate between estimating the group membership and the model parameters. Su et al. (2016) proposed an alternative approach, a lasso-type estimator. C-Lasso is a penalized technique that shrinks the individual level coefficients to the unknown group specific coefficients. This parametric approach have been further expanded to a non-parametric approach by Su et al. (2019) to allow for time varying coefficients. Similar to our approach is the use of model based clustering techniques, such as mixture or hidden Markov models. Whereas Fröhwirth-Schnatter and Kaufmann (2008) proposed a Bayesian finite-mixture model to group the time series of as panel, Kim and Wang (2019) proposed an non-parametric mixture model by adopting the Dirichlet process prior. However, many of these previous studies do not consider individual cluster membership that could change over time; they do not study time varying model parameters; and they only predict the parameters for a known time span.

Lastly, we build upon the literature of (non-)parametric Bayesian Switching State Space Models. While plentiful earlier studies have worked on Bayesian methods for SSSMs (Frühwirth-Schnatter, 2001; Kim et al., 1999; Ghahramani and Hinton, 2000), there has been little work on Bayesian non-parametric or non-homogeneous SSSMs (NSSSMs). Recent studies mainly initiated (non-)parametric Bayesian non-homogeneous HHMs to forecast (multivariate) time series (Hoskovec et al., 2022; Holsclaw et al., 2017; Koki et al., 2022). Fox et al. (2011a), for instance, developed a Bayesian Switching Linear Dynamic System with a Gibbs sampling inference scheme that utilizes a variant of the hierarchical Dirichlet process HMM (HDP-HMM), the sticky HDP-HMM, to improve control over the number of states. Linderman et al. (2017) proposed a non-parametric non-homogeneous SSSM for which the transition probabilities depend on the continuous latent states using the Pólya-gamma auxiliary variable technique. Nassar et al. (2019) builds on the recurrent SLDS (Linderman et al., 2017) by introducing the tree-structured stick breaking that generalizes the sequential logistic stick breaking process. Nonetheless, our work is closely related to the work of Linderman et al. (2017), as they also utilised the Pólya-gamma data augmentation to model the hidden states via the hierarchical logistic stick breaking HMM. Therefore, we have applied a similar Gibbs sampling inference scheme, but we have modified the model in order that households follow the clusters' variance as well as their mean, and we have let our transition probabilities depend on exogenous variables instead of the Gaussian states.

1.2 Thesis Structure

We proceed as follows. In Chapter 2 we give an description of the household earnings dataset and the population forecasts we use in our models. Chapter 3 describes the local level trend model and the multinomial logistic model that respectively models the clusters' general income and the cluster membership of the households. In Chapter 4, we specify the Markov Chain Monte Carlo steps we have constructed and implemented to estimate the model parameters and to obtain the h-step ahead forecasting of the household earnings distribution. The results that show how well our models forecast are given in Chapter 5. Finally, we conclude our study with an discussion of our work in Chapter 6.

Chapter 2

Households' Earnings

In this chapter we give a description of the household earnings data and the exogenous covariates, which we have used in this study. We also briefly describe the results of the population forecasting model as these results are used to correct the income distribution forecasts for changes in the future population.

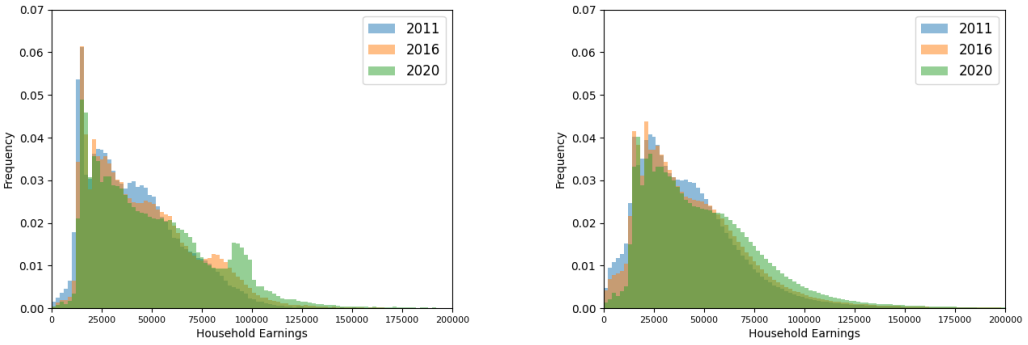
2.1 Description of the dataset

To forecast the household income distribution we use the earnings data of households living in the Netherlands collected by the non-public micro-data catalogue of statistics Netherlands (CBS). The household earnings data corresponds to the annual real disposable household income from December 2011 till December 2020 with December 2020 as reference year. A household's disposable income consists of the gross income minus current transfers such as alimony from the ex-spouse, income insurance contributions, health insurance contributions, and taxes on income and assets. These data files contain the household earnings of more than eight million households. However, because of the time and memory restrictions of our models and of the closed CBS environment we only use 106149 randomly selected households to fit our models. We randomly select the households, since we assume that a large enough dataset will be representative of the entire dataset. Moreover, each year we select some new households, so some households may have an unknown income for previous years. Nevertheless, we assume that the group of households with a very low or large income may be difficult to predict, as their income is often temporary. Therefore, we have not selected the households which had an income lower than 10000 and higher than 100000 euros in the year they are selected. Table 1 gives an overview of the annual real disposable household income by representing its 10th, 30th, 50th, 70th, and 90th percentile and its mean ($\hat{\mu}$) for both our training dataset of 106149 households and a dataset of 5 million households in order to demonstrate our dataset its representativeness. First of all, Table 1 the second column illustrates that each year the total number of households varies, because households enter the dataset at a later point in time. In addition, it shows that the percentiles of our dataset slightly differ from the population of 5 million households. Our dataset has a smaller 10% and 30% percentile and mean ($\hat{\mu}$), and it has a larger 50%, 70%, and 90% percentile. Moreover, Table 1 shows that the households earnings increases over the years, especially for the high-income groups. To illustrates how well the income distributions of our dataset correspond to distributions of the dataset of 5 million households, Figure 1 presents the histograms of the households earnings for both datasets. Figure 1a demonstrates that our dataset has some odd peaks around an income of 20000 and 100000 euros, but generally it follows a similar pattern as in Figure 1b: The overall income has slightly been increased; the distribution has got a fatter tail due to the dispersion of the high-income households; and the low-income group has slightly be declined.

Whether households belong to a certain income group can potentially depend on certain household characteristics. Single-earners might have lower household earnings than dual-earners, and young households might earn less than middle-age households due to the absence of experience. Therefore, changes in these characteristics may affect households' transitions to another income

	106149 households							5 million households					
	N	10%	30%	50%	70%	90%	$\hat{\mu}$	10%	30%	50%	70%	90%	$\hat{\mu}$
2011	96323	14.25	24.72	36.99	51.10	74.92	41.23	14.60	25.10	36.83	50.67	74.11	42.12
2012	98357	14.00	24.43	36.59	50.86	75.04	41.07	14.26	24.52	35.98	49.97	73.43	41.48
2013	100188	14.28	24.49	36.62	51.32	75.91	41.79	14.39	24.36	35.71	50.13	73.85	42.08
2014	102144	14.51	24.85	37.14	52.12	77.09	42.16	14.53	24.44	35.76	50.43	74.41	41.89
2015	103460	15.01	25.81	38.57	54.46	80.47	43.93	14.95	25.23	36.90	52.45	77.09	43.42
2016	103901	15.63	26.63	39.81	56.30	83.42	45.47	15.65	26.25	38.38	54.22	79.43	45.18
2017	104383	15.96	27.12	40.78	57.60	85.38	46.55	16.13	27.03	39.68	55.71	81.26	46.47
2018	105032	16.34	27.99	42.24	59.74	88.70	48.81	16.74	28.19	41.41	57.91	84.36	49.07
2019	105535	16.53	28.38	42.97	61.19	90.91	49.39	17.13	28.82	42.50	59.45	86.15	49.72
2020	106149	16.71	28.94	44.10	63.22	94.00	50.96	17.58	29.75	43.91	61.52	89.28	51.53

Table 1: 10th, 30th, 50th, 70th, and 90th percentile, and mean ($\hat{\mu}$) of the household earnings in euros for our training dataset of 106149 households and a dataset of 5 million households ($\times$1000).



(a) Our training dataset of 106149 households

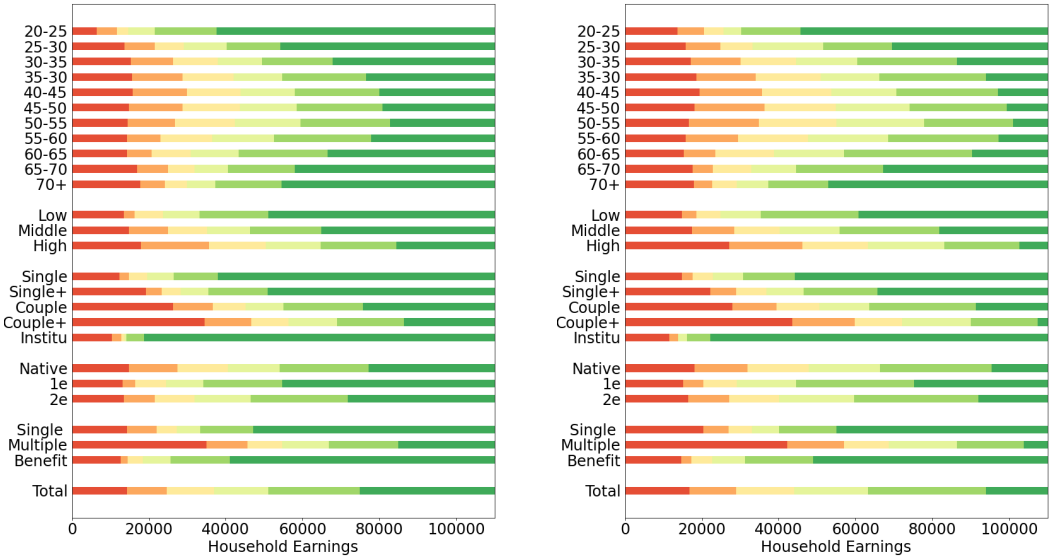
(b) Dataset of 5 million households

Fig. 1: Histograms of the household earnings from our training dataset of 106149 households and from the dataset of 5 million households for the years 2011, 2016, 2020.

group. In this study, we only look at a small number of household characteristics in order to limit our computational time and to enable matching these characteristics with those from the population forecast. The first two columns of Table 2 summarise the number of households for each characteristic in 2011 and 2020: age, education level, generation and the household composition. The main income source is given by 26.6% single-earners, 46.4% multiple-earners, and 27% benefit recipients in 2011 and in 2020 20.4% single-earners, 46.4% multiple-earners, and 33.2% benefit recipients. The characteristics 'age', 'generation', and 'education level' respectively represent the age, generation, and education level of the household's main breadwinner, which is specified as the parent or adult with the highest personal income. We have categorised the education level into three levels according to the standard educational classification[1] given by CBS. We have used generation instead of the ethnicity, as we had no access to ethnicity because of its sensitivity. We have grouped the main income source into single-earner, multiple-earners, and benefit recipients to limit the number of categories. The group of benefit recipients consists of unemployment, social assistance, social welfare, incapacity, pension, and student benefit recipients. The household composition is categorised into single-person (single), single-parent (single+), couple without children (couple), couple with children (couple+), and people living in institutions, facilities and homes (institu(tional)). Figure 2 illustrates the 10th, 30th, 50th, 70th and 90th income percentile of 2011 and 2020 for the five categories over the households

[1] https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/onderwijs-en-beroepen/standaard-onderwijsindeling--soi--/standaard-onderwijsindeling-2021

earnings from our training dataset. For clarification, the color transitions imply the percentile limits. Thus, the boundary between red and orange symbolises the 10th percentile. Figure 2 demonstrates that household earnings significantly varies across age groups, education levels, and main income sources. It also show that the 50%, 70%, and 90% percentile have increased over the years, except for benefits recipients and elderly.



(a) 2011 (b) 2020

Fig. 2: The 10%, 30%, 50%, 70%, and 90% income percentile over our training dataset of 106194 households disaggregated across the different categories (age, education level, household composition, generation, and main income source) for the years 2011 and 2020. The color transitions denote the income percentiles.

In addition to small annual income shifts, the population structure of the Netherlands also evolves. For instance, due to an ageing population the number of households above 70+ will grow. Such a development could affect the overall future household earnings, and therefore, we must correct our predicted income densities for these demographic changes. We thereby use a population forecasting model developed by ABF Research[2]. This model predicts the number of households from 2020 to 2050 for a different age, household composition, ethnicity, and education level. The last three columns of Table 2 present the percentages of the number of households having these characteristics for 2020, 2024, and 2029. It illustrates that the number of households above 70+ will grow; that more households will live as a single-person household; and that in the future the Netherlands will have more households with a migration background. As demonstrated in Table 2, these forecast percentages differ from the percentages of our training dataset. Our dataset consists of fewer elderly people (9.4% over 22.0%) and more couples with children (33.1% over 25.4%). As be noted, this forecasting model does not incorporate income source. To adjust for this category we use the distribution of the main income source from 2020 to determine how many households are single-earners, multiple-earners, or benefit recipients given the other categories.

[2] https://abfresearch.nl

		Training dataset		Population forecast		
		2011	2020	2020	2024	2029
Age	20-25	4.4	0.1	3.4	3.2	3.0
	25-30	9.7	2.2	6.7	6.4	6.3
	30-35	11.5	7.5	7.9	8.0	7.8
	35-40	12.7	10.6	7.6	7.9	8.2
	40-45	14.9	11.3	7.6	7.6	7.9
	45-50	13.8	12.7	8.5	7.5	7.5
	50-55	11.7	14.3	9.8	8.8	7.4
	55-60	9.6	12.7	9.7	9.5	8.4
	60-65	6.7	10.7	8.8	9.2	9.0
	65-70	3.4	8.5	8.0	8.2	8.7
	70+	1.6	9.4	22.0	23.7	25.8
	Total	100%	100%	100%	100%	100%

		Training dataset		Population forecast		
		2011	2020	2020	2024	2029
Education Level	Low	25.0	24.6	29.9	27.9	25.3
	Middle	34.8	34.9	30.8	32.2	34.1
	High	40.2	40.5	39.3	39.9	40.7
	Total	100%	100%	100%	100%	100%
Generation	Native	77.3	76.9	75.8	73.9	71.4
	1e	14.7	14.8	16.4	17.9	19.7
	2e	8.0	8.3	7.8	8.3	8.9
	Total	100%	100%	100%	100%	100%
Composition	Single	35.8	34.4	38.3	39.3	40.6
	Single+	8.8	8.6	7.4	7.3	7.1
	Couple	20.3	22.5	28.4	28.3	27.9
	Couple+	34.0	33.1	25.4	24.6	23.9
	Institutional	1.1	1.4	0.5	0.5	0.5
	Total	100%	100%	100%	100%	100%

Table 2: Percentages of households (%) having a certain age, education level, generation, and household composition in our training dataset of 106149 households for the years 2011 and 2020 and in the population forecasts of 2020, 2024, and 2029.

Chapter 3

Model Specification

In this chapter, we discuss our local level trend model (Section 3.1) which has a time-varying latent group structure to introduce non-linearity in the earnings distribution forecasts. This hidden group structure depends on both the households earnings and their characteristics, and therefore we model the transition probabilities using a multinomial logistic (MNL) model regarding an either known or unknown number of groups, as described in Section 3.2.

3.1 The Local Level Trend Model

We consider a panel of household observations $\{(y_{it}, x_{it})\}$ with $i = 1, \ldots, N$ households in periods $t = 1, \ldots, T$. y_{it} denotes the logarithm of household earnings; and x_{it} is a $P \times 1$ vector of exogenous variables. A household can enter and leave the dataset at anytime, and therefore some y_{it} and x_{it} may be missing over the given time span. To forecast the household earnings, our model first divide the households into a number of clusters with an unknown group structure in advance. Such a division aims to ensure that households with an identical income profile and similar characteristics end up in the same cluster, and thus, they will share the same model components. Our model consists of a group-specific intercept and trend component, which respectively represent the average income and income change of a particular group. We assume that these model components (intercept and trend) vary over time; may differ between the various clusters; and are auto-correlated within a cluster through time, as current household earnings heavily depend on previous earnings. Therefore, we model their movement as a local level trend model with a time-varying hidden group structure:

$$y_{it} = \alpha_{it}^{(g_{it})} + \varepsilon_{it} \qquad \varepsilon_{it} \sim N(0, \sigma_{\varepsilon}^2) \tag{3.1}$$

$$\alpha_{it}^{(k)} = \alpha_{it-1}^{(k)} + \theta_{it-1}^{(k)} + \eta_{it}^{(k)} \qquad \eta_{it}^{(k)} \sim N(0, \sigma_{\eta^{(k)}}^2) \tag{3.2}$$

$$\theta_{it}^{(k)} = \theta_{it-1}^{(k)} + \zeta_{it}^{(k)} \qquad \zeta_{it}^{(k)} \sim N(0, \sigma_{\zeta^{(k)}}^2) \tag{3.3}$$

where ϵ_{it} is the observation error term modeled with a normal distribution and featured by a zero mean and homoskedastic variance (σ_{ε}^2) regardless of time, household, or group membership. $\alpha_{it}^{(g_{it})}$ denotes the time-varying group-specific intercept, and $\theta_{it}^{(k)}$ identifies the time-varying group-specific trend. $\alpha_{it}^{(k)}$ follows a random walk with an extra slope component $\theta_{it-1}^{(k)}$ and state disturbance $\sigma_{\eta^{(k)}}^2$. $\theta_{it}^{(k)}$ follows a regular random walk with state disturbance $\sigma_{\zeta^{(k)}}^2$. These state disturbances are also modeled with a normal distribution and introduce some individual-level heterogeneity within each group, even over time. As some y_{it} may be unknown at $t = 1$, we specify an initialisation equation to initialise $\alpha_{i1}^{(k)}$ and $\theta_{i1}^{(k)}$ at $t = 1$:

$$\alpha_{i1}^{(k)} \sim N(\mu_{\alpha^{(k)}}, \sigma_{\alpha^{(k)}}^2) \qquad \theta_{i1}^{(k)} \sim N(\mu_{\theta^{(k)}}, \sigma_{\theta^{(k)}}) \tag{3.4}$$

Hence, we no longer define $\mu_{\alpha^{(k)}}$ and $\mu_{\theta^{(k)}}$ for each household individually in order to reduce the number of parameters in our model, and to enable income forecasting of households entering the dataset somewhere in the future. The superscript $g_{it} \in \{1, \ldots, K\}$ refers to the group membership of household i at time period t with either a known or an unknown number of

groups K. This group membership depends on the household characteristics (x_{it}) at time point t, and therefore may vary over time ($g_{it-1} \neq g_{it}$).
Since the model parameters follow a continuous state space model and the cluster transitions follow a discrete HMM, we can represent our model as a time-invariant switching linear state space model (SSSM). Which can be represented respectively by an observation and a state equation:

$$y_{it} = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{it}^{g_{it}} \\ \theta_{it}^{g_{it}} \end{bmatrix} + \begin{bmatrix} \sigma_{\varepsilon}^2 \end{bmatrix} = H a_{it}^{g_{it}} + R \tag{3.5}$$

$$a_{it}^{(k)} = \begin{bmatrix} \alpha_{it}^{(k)} \\ \theta_{it}^{(k)} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{it-1}^{(k)} \\ \theta_{it-1}^{(k)} \end{bmatrix} + \begin{bmatrix} \sigma_{\eta^{(k)}}^2 & 0 \\ 0 & \sigma_{\zeta^{(k)}}^2 \end{bmatrix} = F a_{it-1}^{(k)} + Q^{(k)} \tag{3.6}$$

where F and H are fixed matrices identical to all households and $Q^{(k)}$ and R respectively contain the state and observation disturbance(s). The state disturbances in $Q^{(k)}$ are uncorrelated. As a_{i1} is an unknown quantity beforehand, we initialise a_{i1} by sampling from a normal distribution, and we assume that the initial state disturbances in $V_0^{(k)}$ are uncorrelated:

$$a_{i1}^{(k)} \sim N\left(\begin{bmatrix} \mu_{\alpha^{(k)}} \\ \mu_{\theta^{(k)}} \end{bmatrix}, \begin{bmatrix} \sigma_{\alpha^{(k)}}^2 & 0 \\ 0 & \sigma_{\theta^{(k)}}^2 \end{bmatrix} \right) = N(\mu_0^{(k)}, V_0^{(k)}) \tag{3.7}$$

The main goal of this study is to forecast the future distribution in $T + h$ by estimating the model parameters (Λ) and the transition distribution parameters (Section 3.2) for $k = 1, \ldots, K$ by using the sample from period 1 to period T:

$$\Lambda = [\mu_{\alpha^{(k)}}, \mu_{\theta^{(k)}}, \sigma_{\alpha^{(k)}}^2, \sigma_{\theta^{(k)}}^2, \sigma_{\eta^{(k)}}^2, \sigma_{\zeta^{(k)}}^2, \sigma_{\varepsilon}^2] \qquad \beta_k = [\lambda_k, \gamma_{1k}, \ldots, \gamma_{Kk}]$$

We do not aim to forecast the earnings of an individual household. Our panel is rather short, and these forecasts may be influenced by many unknown factors. Instead of household-specific parameters we estimate the parameters of distinct clusters that characterise several group of households in our population, such as elderly or young adults. Households can transition between these clusters such that they move to the cluster that best represents them at that specific point in time. We thereby assume that the identified clusters characterise both today's households and future households. For instance, the future elderly have a similar income as the elderly of today adjusted for an annual income change. Figure 3 presents a schematic illustration of our model, in which the circles symbolises the parameters and the squares the data. Figure 3 illustrates that we compute a Gaussian states $a_{it}^{(k)}$ for each $k = 1, \ldots, K$. However, only the Gaussian state belonging to the household's cluster $a_{it}^{(g_{it})}$ affects the household's predicted income y_{it} at time t. In addition, we only update a Gaussian state of a cluster if a household belongs to that cluster. The other Gaussian states will follow the clusters' general income, since we have no knowledge about what their earnings would have been if the households had been in a different cluster at that particular time. We only have this information, if a household had belonged to this cluster somewhere in the past, as demonstrated by $a_{i2}^{(2)}$ in Figure 3. Moreover, Figure 3 shows that the Gaussian states and the hidden states respectively depend on the model parameters (Λ) and the multinomial logistic model parameters (β_k). Whereas we have discussed the model parameters in this section, we will describe the latter in the next section.

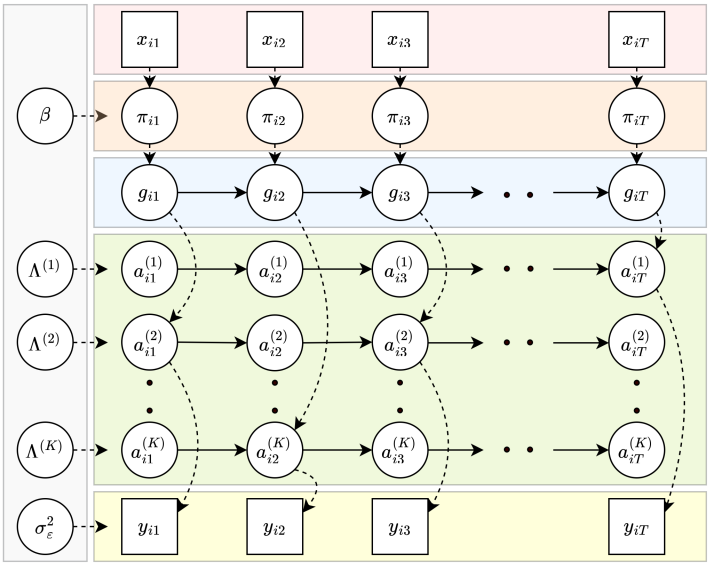


Fig. 3: A schematic illustration of our local level trend model with a time-varying hidden group structure. The circles symbolises the parameters, the squares the data, and the arrows the dependence.

3.2 Group Membership Transition Model

As group membership varies over time due to changing households' characteristics, we model their membership as a K-state non-homogeneous first order Hidden Markov process (NHMM). In a NHMM, the transition probabilities vary over time as a function of the covariates, as demonstrated by Figure 3. Additionally, these probabilities depend upon the previous hidden states by including a set of $k = 1, \ldots, K$ transition regressors (γ_{jk}). Nonetheless, to allow for a more parsimonious approach the $p = 1, \ldots, P$ coefficients of the exogenous variables do not dependent on the previous state, but only on the current state (λ_{kp}). Hence, we describe the dynamics of the hidden clusters by the time-varying transition probabilities $\pi_{jk}(x_{it})$ for $j, k = 1, \ldots, K$. These probabilities depend on the exogenous variables x_{it} and previous state (g_{it-1}) of household i at time point t, and are given by the following multinomial logistic relationship:

$$\pi_{jk}(x_{it}) = P(g_{it} = k | g_{it-1} = j, x_{it}, \beta) = \frac{e^{\gamma_{jk} + x_{it}\lambda_k}}{\sum_{l=1}^{K} e^{\gamma_{jl} + x_{it}\lambda_l}} = \frac{e^{x_{it}\beta_{jk}}}{\sum_{l=1}^{K} e^{x_{it}\beta_{jl}}} \tag{3.8}$$

where λ_k is a P-dimensional vector of coefficients corresponding to the P components of $x_{it} = (x_{it1}, \cdots x_{itP})$, and γ_{jk} denotes a k transition regressor representing the intercept from cluster j to cluster k. We let $\beta_{jk} = (\lambda_k, \gamma_{jk})$ for all $j, k \in 1, \cdots, K$, and we assign $\beta_{\cdot K+P}$ to zero for identifiability. Still, in a non-homogeneous hidden Markov model (NHMM), we do not directly observe the multinomial data as in multinomial logistic regression. Therefore, we arrange the sampled hidden cluster memberships g_{it} in matrix form Y. Y is a TN by K matrix having g_{it} as entries, and its columns contain the binary representation of the hidden clusters: $Y_{it} = [0, 0, 1]$ if $g_{it} = 2$. To determine the transition regressor γ_{jk} and the coefficients of the exogenous variables λ_k, we include the previous group membership in addition to the exogenous variables in matrix X. As these exogenous variables are not continuous but discrete without an inherent order, we cannot directly use them in our transition model (Eq 3.8). Therefore, we encode these nominal data into a binary format using an one-hot-encoding, which results in a total of twenty-five exogenous variables. By using an one-hot-encoding we can set the variables of unknown household

to zeros. Thus, the first K columns of X encode the previous group membership in a binary form, and the second P columns contains the one-hot encoding of the exogenous variables. If an observation is unknown in the dataset, we set both Y_{it} and X_{it} to a row of zeros. Therefore, this data point does not affect the MNL parameters, and its transition probability will be the same for each cluster.

However, the sampling scheme for the multinomial logistic (MNL) model coefficients β_k can be challenging due to the intractable form of the likelihood and the lack of a conjugate prior for the coefficients (Wang et al., 2023). Therefore, we adopt the efficient data augmentation approach proposed by Polson et al. (2013) that enables conjugate sampling of the MNL coefficients via Pólya-Gamma distribution. Polson et al. (2013) proved two useful properties of Pólya-gamma variables. First,

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} p_{PG}(\omega|b,0) d\omega \tag{3.9}$$

where $\kappa = a - b/2$ and $p_{PG}(\omega|b,0)$ is the density of the Pólya-Gamma distribution, and second, $p(\omega|\psi) \sim PG(b,\psi)$, for $b > 0$. Based on these main results, we can now compute the full conditional posterior of β_k using a normal distribution, which results in a two-step sampling scheme: a Pólya-Gamma update for the latent variables ω_{itk} and a joint Gaussian update for MNL coefficients $\beta_k \sim N(m_k, V_k)$. In respectively Section 3.2.1 and 3.2.2, we describe how we compute m_k for either a fixed number of clusters (parametric) or infinite number of clusters (non-parametric). In addition, in Section 3.2.3 we explain how we have added a shrinkage prior to the coefficients to handle the many zeros in the exogenous variables.

3.2.1 Known Number of Groups (Parametric)

When the number of clusters is a fixed quantity, we can specify the full conditional posterior for β_k given the other parameters of the model. Thereby, we use the multinomial logistic representation of the transition probabilities in Eq. 3.8, and we follow the steps of Held and Holmes (2006):

$$\begin{aligned}
p(\beta_k|\beta_{-k}, x_{it}) &= p(\beta_k|\beta_{-k}) \prod_{t=1}^{T}\prod_{i=1}^{N}\prod_{k=1}^{K-1} (\pi_{jk}(x_{it}))^{\mathbb{1}[g_{it}=k]} \\
&= p(\beta_k|\beta_{-k}) \prod_{t=1}^{T}\prod_{i=1}^{N}\prod_{k=1}^{K-1} \left(\frac{e^{x_{it}\beta_k}}{\sum_{l=1}^{K} e^{x_{it}\beta l}}\right)^{\mathbb{1}[g_{it}=k]} \\
&= p(\beta_k|\beta_{-k}) \prod_{t=1}^{T}\prod_{i=1}^{N} \left(\frac{e^{\eta_{itk}}}{1+e^{\eta_{itk}}}\right)^{\mathbb{1}[g_{it}=k]} \left(\frac{1}{1+e^{\eta_{itk}}}\right)^{\mathbb{1}[g_{it}\neq k]} \\
&= p(\beta_k|\beta_{-k}) \prod_{t=1}^{T}\prod_{i=1}^{N} \frac{(e^{\eta_{itk}})^{\mathbb{1}[g_{it}=k]}}{1+e^{\eta_{itk}}}
\end{aligned} \tag{3.10}$$

where $\eta_{itk} = x_{it}\beta_k - c_{itk} = x_{it}\beta_k - \ln(\sum_{l\neq k} x_{it}\beta_l)$. By combining the property of the Pólya-Gamma distribution in Eq. 3.9 and the full conditional posterior (Eq. 3.10), we assume conditionally conjugate priors for the MNL coefficients $\beta_k \sim N(m_0, V_0)$, and hence, conditioning on the Pólya-Gamma random variables ω_{itk}, the posterior transforms into a single Gaussian distribution:

$$\beta_k|\Omega_k \sim N(m_k, V_k) \qquad\qquad \omega_{itk}|\beta_k \sim PG(1, \eta_{itk}) \tag{3.11}$$

where scalars $V_k = (X'\Omega_k X + V_0^{-1})^{-1}$ and $m_k = V_k(X'((Y_k - 1/2) + \Omega_k C_k) + b_0^{-1} m_0)$. Ω_k is a TN by TN diagonal matrix containing ω_{itk} along the diagonal. m_0 and V_0 are parameters of the conjugate prior of the of the form $\beta_k \sim N(m_0, V_0)$. Once we have sampled the MNL coefficients (β_k), we can easily obtain the transition probabilities $\pi_{it}(x_{it})$ through the logistic relationship given in Equation 3.8. This leads to a K by K transition matrix for each household i at time t, where each rows sums to one.

3.2.2 Unknown Number of Groups (Non-Parametric)

Nevertheless, selecting the best number of hidden states is a serious problem for statistical modeling, and therefore we extend the finite-state HMM to an Hierarchical Dirichlet process HMM (HDP-HMM) allowing for potentially countably infinite number of hidden states. Instead of imposing a multinomial logit prior on the rows of the finite state transition matrix, we use a hierarchical logistic stick-breaking prior. A stick-breaking prior divides a unit-length stick into infinitely many segments by iteratively breaking off a proportion from the remainder of the stick, and enables direct sampling despite its infinitely (Sethuraman, 1994). Thereby, the logistic stick-breaking prior relates each stick-breaking weight to a function of the covariates, the logistic link function (Rigon and Durante, 2021), and thus, we can define the transition probability from state j to state k as follows:

$$p(g_{it} = k | g_{it-1} = j, x_{it}) = \pi_{jk}(x_{it}) = \begin{cases} \xi_{jk}(x_{it}) \prod_{l<k} (1 - \xi_{jl}(x_{it})) & k > 1 \\ \xi_{jk}(x_{it}) & k = 1 \end{cases} \tag{3.12}$$

where $\xi_{jk}(x_{it})$ is defined as a logistic link function $\xi_{jk}(x_{it}) = \frac{e^{x_{it}\beta_k}}{1+e^{x_{it}\beta_k}}$. And similar to equation 3.10, we can now define the full conditional posterior for β_k as follows:

$$\begin{aligned} p(\beta_k | \beta_{-k}, x_{it}) &= p(\beta_k | \beta_{-k}) \prod_{t=1}^{T} \prod_{i=1}^{N} \prod_{k=1}^{K} (\pi_{jk}(x_{it}))^{\mathbb{1}[g_{it}=k]} \\ &= p(\beta_k | \beta_{-k}) \prod_{t=1}^{T} \prod_{i=1}^{N} (\xi_{jk}(x_{it}))^{\mathbb{1}[g_{it}=k]} (1 - \xi_{jk}(x_{it}))^{\mathbb{1}[g_{it}>k]} \\ &= p(\beta_k | \beta_{-k}) \prod_{t=1}^{T} \prod_{i=1}^{N} \left(\frac{e^{\beta_{jk} x_{it}}}{1 + e^{\beta_{jk} x_{it}}} \right)^{\mathbb{1}[g_{it}=k]} \left(\frac{1}{1 + e^{\beta_{jk} x_{it}}} \right)^{\mathbb{1}[g_{it}>k]} \\ &= p(\beta_k | \beta_{-k}) \prod_{t=1}^{T} \prod_{i=1}^{N} \frac{(e^{\beta_{jk} x_{it}})^{\mathbb{1}[g_{it}=k]}}{(1 + e^{\beta_{jk} x_{it}})^{\mathbb{1}[g_{it} \geq k]}} \end{aligned} \tag{3.13}$$

And again by applying the Pólya-Gamma data augmentation technique, we can now specify the full posterior for β_k as follows:

$$\beta_k | \Omega_k \sim N(m_k, V_k) \qquad \omega_{itk} | \beta_k, x_{it} \sim PG(I[g_{it} \geq k], \beta_k x_{it}) \tag{3.14}$$

where $m_k = V_k(X'\kappa + b_0^{-1} m_0)$, $\kappa = \mathbb{1}[g_{it} = k] - \mathbb{1}[g_{it} \geq k]/2$, and $V_k = (X'\Omega_k X + b_0^{-1})^{-1}$. Ω_k is again a TN by TN diagonal matrix containing ω_{itk} along the diagonal, and m_0 and V_0 are parameters of the conjugate prior of the form $\beta_k \sim N(m_0, V_0)$. After we have sampled the MNL coefficients (β_k), we can obtain the transition probabilities $\pi_{it}(x_{it})$ through the logistic stick-breaking construction given in Equation 3.12. This leads to a K by K transition matrix for the current occupied clusters, and we calculate the transition probabilities for a potential new cluster $K + 1$ as $\pi_{j,K+1} = 1 - \sum_{k=1}^{K} \pi_{jk}(x_{it})$.

3.2.3 Global-Local Shrinkage Prior

Nevertheless, our exogenous variables x primarily contain zeros, and possibly some predictors might not have any effect in determining the cluster membership of households. Therefore, we use a Bayesian variable selection technique in our MNL model, the global-local shrinkage prior, to cope with this sparseness. As the MNL model has $(K-1)$ logistic regressions equation, we express the prior distribution of a regression coefficient as $\beta_{kp} \sim N(0, \delta_{kp}^2\phi^2)$. The local parameter δ_{kp} is the shrinkage parameter of the p predictor for cluster $k = 1, \ldots, K-1$, and the global parameter ϕ^2 determines the overall level of shrinkage to all coefficients. We have chosen to implement the horseshoe prior to sample these global and the local parameters $\delta_{kp}, \phi^2 \sim C^+(0,1)$ (Carvalho et al., 2009). And we use the data augmentation strategy proposed by Makalic and Schmidt (2016) to achieve conjugacy of the hyper-parameters. Therefore, we now sample β_k from $N(m_k, (X'\Omega_k X + \Delta_k)^{-1})$. We already defined m_k in the previous sections, and Δ_k is the $(P+K)\times(P+K)$ diagonal matrix with $[(\delta_{1k})^{-1}, \ldots, (\delta_{Kk})^{-1}, (\delta_{k1}\phi^2)^{-1}, \ldots, (\delta_{kP}\phi^2)^{-1}]$ on the diagonal. We set δ_{jk} for $j = 1, \ldots, K$ to a fixed value ($\delta_{jk} = 0.01$) to account for the prior variance of the transition regressors (γ_{jk}). Namely, households tend to transition towards the same cluster. As a result, the transition model mainly considers the transition regressors as the most important coefficients, and it lets the coefficients of the exogenous variables λ_k shrink to zero. We sample ϕ^2 and δ_{kp} for $k = 1, \ldots, K-1$ clusters and $p = 1, \ldots, P$ predictors from the following conditional posteriors:

$$\delta_{kp}^2|\eta_{kp}, \beta_{kp}, \phi^2 \sim IG\left(\frac{1}{2}, \frac{1}{\eta_{kp}} + \frac{\beta_{kp}^2}{2\phi^2}\right) \tag{3.15}$$

$$\phi^2|\beta, \delta, \xi \sim IG\left(\frac{(K-1)P+1}{2}, \frac{1}{\xi} + \sum_{k=1}^{K-1}\sum_{p=1}^{P}\frac{\beta_{kp}^2}{2\delta_{kp}}\right) \tag{3.16}$$

$$\eta_{kp}|\delta_{kp} \sim IG(1, 1 + \frac{1}{\delta_{kp}}) \tag{3.17}$$

$$\xi|\phi^2 \sim IG(1, 1 + \frac{1}{\phi^2}) \tag{3.18}$$

Here, IG stands for the inverse gamma distribution with probability density of $f(x|a,b) \propto x^{-a-1}e^{-b/x}$, $x > 0$

Chapter 4

Bayesian Inference and Forecasting

In this chapter, we describe how we perform a fully Bayesian inference via a Markov chain Monte Carlo (MCMC) sampling scheme to obtain samples from the posterior distribution and use them to approximate the true posterior distribution. In addition, in Section 4.2 we explain how we compute the h-step-ahead forecast distribution of the household earnings, and how we evaluate the one-step-ahead forecast distribution against the true income distribution.

4.1 MCMC Sampling from Posterior

For inference in our (non-)parametric non-homogeneous SSSMs, we use a block Gibbs sampler (Roberts and Sahu, 1997) that groups two or more variables together in a block and update this block by sampling from the joint distribution of these variables conditioned on all of the variables, so we can write the joint distribution of observations and hidden states over the model parameters as follows:

$$p(g_{1:T}, y_{1:T}, a_{1:T}) = p(g_1) \prod_{t=1}^{T} p(g_{t+1}|g_t) \cdot \prod_{k=1}^{K} p(a_1^{(k)}|g_1) \prod_{t=2}^{T} p(a_t^{(k)}|a_{t-1}^{(k)}, g_t) \cdot \prod_{t=1}^{T} p(y_t|a_t) \quad (4.1)$$

where $y_{1:T}$; $a_{1:T}$, and $g_{1:T}$ denote the sequences (of length T) of the observations and hidden state variables. As the structure of the model allows for closed form conditional posterior distributions, we obtain the MCMC samples from the conditional posterior distribution of the parameters by cycling through the following steps:

1. Calculate the probabilities of the time-varying transition matrix $(\pi_{jk}(x_{it}))$ given the transition distribution parameters β (Section 3.2).

2. Given the model parameters, the Gaussian states, the transition probabilities, and the data, simulate the hidden states $p(g_{1:T}|\Lambda, a_{1:T}, \pi, y_{1:T})$ using the the forward filtering, backward sampling algorithm (Section 4.1.1).

3. Given the hidden states, the model parameters, and the data, simulate the Gaussian states $p(a_{1:T}|g_{1:T}, y_{1:T}, \Lambda)$ using the Kalman Filter and backward sampler (Section 4.1.2).

4. Given the hidden states, the Gaussian states, and the data, sample the model parameters $p(\Lambda|g_{1:T}, a_{1:T}, y_{1:T})$ (Section 4.1.3)

5. Given the hidden states and the exogenous covariates, sample the transition distribution parameters $p(\beta|g_{1:T}, x_{1:T})$ using the Pólya-Gamma representation by Polson et al. (2013) (Section 3.2).

Figure 4 gives a schematic representation of our Gibbs sampling inference scheme. As illustrated, to iterate between the sampling states we first require an initialisation of Gibbs sampler. Therefore, we initialise the MNL parameters β_0 and the model parameters Λ_0 using their prior distributions, and we base our initialisation of the hidden states fully on the transition probabilities without knowledge of the Gaussian states. Figure 4 also illustrates that we can update some parameters several times in succession within a single block, which improves mixing if there is a strong dependence in the state-process.

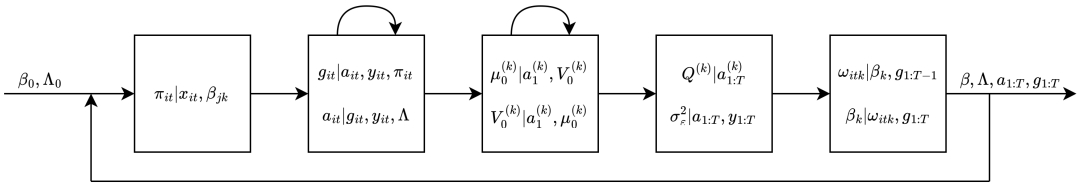


Fig. 4: A schematic representation of our block Gibbs sampler hat iterates between sampling the hidden clusters $g_{1:T}$, the Gaussian states $a_{1:T}$, and the model parameters (Λ,β).

4.1.1 Updating the Hidden States

Conditioned on the Gaussian states, the transition matrix, and the data we can jointly update the hidden states using the forward filtering-backward sampling (FFBS) algorithm (Scott, 2002). We jointly update these hidden states, as it allows us to avoid mixing problems when individual sampling states are strongly dependent on one another. We first compute the filtering step to get the marginal distribution over g_{it} given the observations and Gaussian states up to time t as follows:

$$p(g_{it} = k|y_{1:t}, a_{1:t}, \Lambda^{(k)}) \propto p(y_{it}, a_{it}|a_{it-1}, \Lambda^{(k)}) \sum_{l=1}^{K} \pi_{lk}(x_{it}) p(g_{it-1} = l|a_{1:t-1}, y_{1:t-1}, \Lambda^{(l)}) \quad (4.2)$$

We normalize $p(g_{it}|y_{1:t}, a_{1:t}\Lambda^{(k)})$ by dividing it by $\sum_{k=1}^{K} p(g_{it} = k|y_{1:t}, a_{1:t}\Lambda^{(k)})$. Once we have computed the filtered distributions, we can sample from the joint distribution over $g_{1:T}$ by applying the chain rule (Linderman, 2016):

$$p(g_{1:T}|y_{1:T}, a_{1:T}, \Lambda) \propto p(g_T|y_{1:T}, a_{1:T}, \Lambda) \prod_{t=1}^{T-1} p(g_t|y_{1:t}, \Lambda, a_{1:t}) p(g_{t+1}|g_t, \beta) \quad (4.3)$$

Thus, we can sample g_{it} in reverse order. We first jointly sample $g_{iT} \sim p(g_{iT}|y_{1:T}, a_{1:T})$, and subsequently we sample g_{it} given g_{it+1} backwards through:

$$p(g_{it} = k|g_{it+1:T}, y_{1:T}, a_{1:T}) \propto p(g_{it} = k|y_{1:t}, a_{1:t}, \Lambda^{(k)}) \pi_{k,g_{it+1}}(x_{it}) \quad (4.4)$$

Where g_{it+1} is the most recent sampled cluster at $t+1$ for household i. We repeat this until we obtain a value for g_{i1} for each household $i = 1, \ldots, N$.
As specified in Equation 4.2, this approach requires a value for $p(y_{it}, a_{it}|a_{it-1}, g_{it} = k, \Lambda^{(k)})$. But we will face a problem, if we fully follow the computation steps defined in Linderman (2016) that specified $p(y_{it}, a_{it}|a_{it-1}, g_{it} = k, \Lambda^{(k)})$ as:

$$p(y_{it}, a_{it}|a_{it-1}, \Lambda^{(k)}) = \begin{cases} p(y_{it}|a_{it}, \sigma_\varepsilon^2) p(a_{it}|a_{it-1}, Q^{(k)}) & t > 1 \\ p(y_{i1}|a_{i1}, \sigma_\varepsilon^2) p(a_{i1}|\mu_0^{(k)}, V_0^{(k)}) & t = 1 \end{cases} \quad (4.5)$$

Namely, for $t > 1$ households tend to transition to a cluster for which their Gaussian states best match the state disturbances $Q^{(k)}$. As a results, a household does not transition to the cluster that best corresponds to its income at that point in time. Moreover, the difference between a_{it} and a_{it-1} can only be calculated if the households belong to the same cluster at t and $t-1$. Otherwise, we calculate the difference between two Gaussian states while not actually succeeding each other. Thus, we can only compute $p(a_{it}|a_{it-1}, Q^{(k)})$, if a_{it} and a_{it-1} belong to k. Nevertheless, a household can transition between clusters, so their Gaussian states might not

belong to the same cluster ($a_{it}^{g_{it}}, a_{it-1}^{g_{it-1}}, g_{it} \neq g_{it-1}$). Therefore, we define a heuristic approach to compute $p(y_{it}, a_{it}|a_{it-1}, g_{it} = k, \Lambda^{(k)})$.

$$p(y_{it}, a_{it}|\Lambda^{(k)}) = p(y_{it}|a_{it,t}, \sigma_\varepsilon^2)p(a_{it,1}|\mu_0^{(k)}, V_0^{(k)}) \prod_{q=1}^{t-1} p(a_{it,q+1}|a_{it,q}, Q^{(k)}) \tag{4.6}$$

where $a_{it} = [a_{it,1}, \ldots, a_{it,t}]$ are the Gaussian states up to time t. This heuristic approach works well in practise, but we refer to Section 6.1.2 for a more time efficient approach.

4.1.1.1 Non-Parametric

The forward filtering-backward sampler can not directly applied to HMMs with infinite states, because we need to compute the sum over an infinite number of clusters. Therefore, we have implement the beam sampling algorithm that combines a slice sampler with the forward-backward algorithm (Gael et al., 2008). It restricts the number of reachable clusters at each MCMC iterations to a finite number by introducing an auxiliary slice variable $u_{it} \sim U(0, \pi_{g_{it-1}, g_{it}})$. This slice variable truncates the sum over the infinite number of clusters to a finite number of clusters by imposing an restriction on the transition probabilities, $0 < u_{it} < \pi_{g_{i,t-1}, g_{it}}$, and we can compute the filtering step of the FFBS algorithm as follows:

$$p(g_{it} = k|y_{1:t}, a_{1:t}, \Lambda^{(k)}) \propto p(y_{it}, a_{it}|a_{it-1}, \Lambda^{(k)}) \sum_{l=1}^{\infty} \mathbb{1}[u_{it} < \pi_{lk}(x_{it})] p(g_{it-1} = l|y_{i1:t-1}, u_{i1:t-1}, \Lambda^{(l)}) \tag{4.7}$$

and in the backward step we sample g_{it} for $t = T, \ldots, 1$ as:

$$p(g_{it} = k|g_{i,t+1:T}, y_{1:T}, a_{1:T}, \Lambda^{(k)}) \propto (g_{it} = k|y_{it}, u_{i,1:T}, a_{1:t}, \Lambda^{(k)}) \mathbb{1}[u_{it} < \pi_{k, g_{it+1}}(x_{it})] \tag{4.8}$$

However, Fox et al. (2011b) showed that applications of the beam sampler result in slower mixing rates compared to a truncated approximation of the forward filtering-backward sampler, for which we fix a truncation level for the logistic stick-breaking process (LSBP) such that the sum in Equation 4.7 is no longer infinite. Besides, we also experienced that the probability of moving from one cluster to another cluster decreases drastically, so that most households remain in the first cluster. Therefore, we mainly examine a weak limit truncation of LSBP-HMM that fixes the truncation level to a given number, so that we can continue to use the forward filtering-backward sampler.

4.1.1.2 Label Switching

A potential problem in Bayesian analysis of HMMs is the Label Switching Problem (Jasra et al., 2005), especially, when we use exchangeable priors for the state specific parameters, as no prior information is available for the hidden states. This problem arises if the posterior distribution is invariant to permutations of the state labels, as this leads to identical marginal posterior distributions of the state specific parameters. Any pair of states could swap labels, while the likelihood remain identical, which makes the generated MCMC samples non-identifiable. According to Meligkotsidou and Dellaportas (2011) and Holsclaw et al. (2017) NHMMs are less likely to be affected by this issue than HMMs, as the hidden states are depended on exogenous covariates. Therefore, we decided not to address this problem. We refer to Section 6.1.2 for some solutions to this problem and to future research to implement them.

4.1.2 Updating the Gaussian states

Conditioned on the discrete hidden states and the model parameters, our dynamical process simplifies to a time-varying linear state space model. We can then perform a blocked Gibbs update

for the entire Gaussian state sequence, $a_{1:T}$, using forward filtering-backward sampling algorithm (Carter and Kohn, 1994; Frühwirth-Schnatter, 2004), similar as we used in Section 4.1.1. We can compute the marginal 'filtered' distribution $p(a_{it}^{(k)}|y_{i1:t}, g_{i1:t}, \Lambda) = N(a_{it}^{(k)}|\mu_{it}^{(k)}, V_t^{(k)})$ using the Kalman filter, where $\mu_{it}^{(k)}$ and $V_t^{(k)}$ are the filtered mean and covariance, respectively for $k = 1, \ldots, K$. The Kalman filter consists of iterating forward in time, starting from $t = 1$ by using $\mu_0^{(k)}$, $V_0^{(k)}$, F, $Q^{(k)}$, H, and R as its input:

$$
\begin{aligned}
K_1^{(k)} &= V_0^{(k)} H^{\mathrm{T}} (H V_0^{(k)} H^{\mathrm{T}} + R)^{-1} \\
\mu_{i1}^{(k)} &= \mu_0^{(k)} + K_1^{(k)} (y_{i1} - H\mu_0^{(k)}) \\
V_1^{(k)} &= (I - K_1^{(k)} H) V_0^{(k)} \\
P_1^{(k)} &= F V_1^{(k)} F^{\mathrm{T}} + Q^{(k)}
\end{aligned}
$$

We proceed until we have computed $\mu_{it}^{(k)}$ and $V_t^{(k)}$ for $t = 2, \ldots, T$:

$$
\begin{aligned}
K_t^{(k)} &= P_{t-1}^{(k)} H^{\mathrm{T}} (H P_{t-1}^{(k)} H^{\mathrm{T}} + R)^{-1} \\
\mu_{it}^{(k)} &= F\mu_{it-1}^{(k)} + K_t^{(k)} (y_{it} - HF\mu_{it-1}^{(k)}) \\
V_t^{(k)} &= (I - K_t^{(k)} H) P_{t-1}^{(k)} \\
P_t^{(k)} &= F V_t^{(k)} F^{\mathrm{T}} + Q^{(k)}
\end{aligned}
$$

After we have computed the filtering densities $p(a_t|y_{1:t}, g_{1:t})$ for $t = 1, \ldots, T$, we proceed backward in time to draw a joint sample from the backward kernel $p(a_t|a_{t+1}, y_{1:t}, g_{1:t}) = N(a_t|m_{it}, L_t)$ with:

$$
\begin{aligned}
C_t^{(k)} &= V_t F^{\mathrm{T}} (F V_t F^{\mathrm{T}} + Q^{(k)})^{-1} \\
m_{it}^{(k)} &= \mu_{it}^{(k)} + C_t^{(k)} (a_{it+1}^{(k)} - F\mu_{it}^{(k)}) \\
L_t^{(k)} &= (I - C_t^{(k)} F) P_t^{(k)}
\end{aligned}
$$

Initially, we generate a sample from the filtering density at time T, $a_{iT}^{(k)} \sim N(\mu_{iT}^{(k)}, V_T^{(k)})$, and then continue to use $a_{it+1}^{(k)}$ for sampling $a_{it}^{(k)} \sim N(m_{it}^{(k)}, L_t^{(k)})$ until we reach $a_{i1}^{(k)}$ for $i = 1, \ldots, N$ households and $k = 1, \ldots, K$ clusters. However, y_{it} is not always known. Households might not be in the dataset at time t, or they may not belong to cluster k at time t ($g_{it} \neq k$). Therefore, we cannot calculate $y_{it} - HF\mu_{it-1}^{(k)}$, and we compute $\mu_{it}^{(k)}$ by $F\mu_{it-1}^{(k)}$ in the forward filtering step (Durbin and Koopman, 2012). In the backward sampling step we consider two approaches. We compute $m_{it}^{(k)}$ either by $\mu_{it}^{(k)}$ (1) or by $\mu_{it}^{(k)} + C_t^{(k)}(a_{it+1}^{(k)} - F\mu_{it}^{(k)})$ (2). In option (1) compared to option (2), the Gaussian states of the missing observations will follow the general income of the cluster without being influenced by the Gaussian states forward in time, similarly to h-step-ahead Gaussian states a_{iT+h}. Therefore, the households' group membership will be determined more by the cluster's general income than by the vagaries of the households' income. Moreover, by using option (1) we assume that the current households earnings only depend on previous earnings and not on upcoming earnings, and therefore the current earnings of a household are only affected by the earnings we actually know at the time t. Nonetheless, with option (2) we assume that the future household earnings resemble the past earnings, and therefore, we include information about the future earnings households into the past earnings.

Now that we have obtained a Gaussian state $a_{it}^{(k)}$ for all $k = 1, \ldots, K$, we can define the Gaussian state given g_{it} (a_{it}). However, as we already discussed in Section 4.1.1, we cannot determine the hidden state g_{it} of the households on $p(a_{it}^{g_{it}}|a_{it-1}^{g_{it-1}})$. These Gaussian states may be far apart if

$g_{it} \neq g_{it-1}$, making $Q^{(k)}$ no longer correspond. Besides, we cannot directly compare $a_{it}^{(g_{it})}$ for $t = 2, \ldots, T$ with $\mu_0^{(k)}$ for $k = 1, \ldots, K$, as the clusters' overall income may have shifted in the upcoming years. Therefore, we require to define the Gaussian states as a vector both containing the previous and current Gaussian states of cluster g_{it} up to time t. Hence, given the hidden state g_{it} we define the Gaussian state $a_{it} = [a_{it,1}^{g_{it}}, \ldots, a_{it,t}^{g_{it}}]$ for $t = 1 \ldots, T$. Suppose household i belongs to cluster l at $t = 1$, and cluster m at $t = 2$, its Gaussian states are $a_{i1} = [a_{i1,1}^{(l)}]$ and $a_{i2} = [a_{i2,1}^{(m)}, a_{i2,2}^{(m)}]$.

4.1.3 Updating the Parameters of the Local Level Trend Model

Conditioned on the Gaussian states, the hidden states, and the data, we can sample the model parameters $\Lambda^{(k)}$ for $k = 1 \ldots, K$ from their posterior densities. Since we assume that the initial state disturbances $V_0^{(k)}$ are uncorrelated, we place a separate inverse gamma prior $IG(\nu_0, \delta_0)$ on $V_0^{(k)}$ which results in the following posterior distribution given $\mu_0^{(k)}$:

$$
\begin{aligned}
p(V_0^{(k)} | \mu_0^{(k)}, a_{i1,1}) &= IG(\nu^{(k)}, \delta^{(k)}) \\
\nu^{(k)} &= \nu_0 + \frac{|C_1^{(k)}|}{2} \\
\delta^{(k)} &= \delta_0 + \sum_{i \in C_1^{(k)}} (a_{i1,1} - \mu_0^{(k)})(a_{i1,1} - \mu_0^{(k)})^{\mathrm{T}}
\end{aligned}
$$

where the set $C_1^{(k)}$ implies the set of households belonging to cluster k on time $t = 1$, $|C_1^{(k)}|$ defines the number of household in k at $t = 1$, and $a_{i1,1}$ denotes the Gaussian state at $t = 1$ for household i. We model μ_0 using a Normal prior $N(m_{a^{(k)}}, b_{a^{(k)}})$ given the initial state noise $V_0^{(k)}$. Thus, we sample from the posterior given $V_0^{(k)}$:

$$
\begin{aligned}
p(\mu_0^{(k)} | V_0^{(k)}, a_{i1,1}) &= N(\mu_{a_0}^{(k)}, \Sigma_{a_0}^{(k)}) \\
\Sigma_{a_0}^{(k)} &= (b_{a^{(k)}}^{-1} + |C_{1^{(k)}}| (V_0^{(k)})^{-1})^{-1} \\
\mu_{a_0}^{(k)} &= \Sigma_{a_0}^{(k)} (b_{a^{(k)}}^{-1} m_{a^{(k)}} + (V_0^{(k)})^{-1} \sum_{i \in C_1^{(k)}} a_{i1,1})
\end{aligned}
$$

where the set $C_{1^{(k)}}$ and $a_{i1,1}$ are already defined above. To improve parameter inference, in practise we iterate multiple times between sampling $\mu_0^{(k)}$ given $V_0^{(k)}$ and $V_0^{(k)}$ given $\mu_0^{(k)}$ before moving to the next sampling stage, as illustrated in Figure 4. Similar to initial state noises $V_0^{(k)}$, we also assume that the state disturbances in $Q^{(k)}$ are uncorrelated. Therefore, we place an inverse gamma prior $IG(\nu_0, \delta_0)$ on the variances in $Q^{(k)}$, and we define its posterior distribution as follows:

$$
\begin{aligned}
Q^{(k)} &= IG(\nu^{(k)}, \delta^{(k)}) \\
\nu^{(k)} &= \nu_0 + \frac{|C_t^{(k)}|}{2} \\
\delta^{(k)} &= \delta_0 + \sum_{t=2}^{T} \sum_{i \in C_t^{(k)}} (a_{it,t} - F a_{it-1,t-1})(a_{it,t} - F a_{it-1,t-1})^{\mathrm{T}}
\end{aligned}
$$

Where the set $C_t^{(k)}$ contains the households i belonging to cluster k at t and $t-1$ for $t = 2, \ldots, T$, $C_t^{(k)}$ represents its cardinality, and $a_{it,t}$ and $a_{it-1,t-1}$ imply the Gaussian states at t and $t-1$ for household i. Lastly, we additionally place an inverse gamma prior $IG(\nu_0, \delta_0)$ on the measurement

noise covariance R, for which we assume it is shared by all clusters. The posterior distribution is given by $\sigma_\varepsilon^2 \sim IG(\nu_\varepsilon, \delta_\varepsilon)$ where

$$\nu_\varepsilon = \nu_0 + \frac{TN}{2}$$
$$\delta_\varepsilon = \delta_0 + \sum_{t=1}^{T}\sum_{i=1}^{N}(y_{it} - Ha_{it,t})(y_{it} - Ha_{it,t})^{\mathrm{T}}$$

where $a_{it,t}$ denotes the Gaussian states and y_{it} represents the log household earnings for household $i = 1, \ldots, N$ at time points $t = 1, \ldots, T$.

4.2 Forecast Evaluation

In this section, we describe how we compute the h-step ahead annual household income distribution and how we compare this distribution forecast with the true distribution from the data.

4.2.1 Forecasting the h-step-head Income Distribution

As our posterior predictive density cannot be found in closed form, we sample from the posterior distribution numerically by following the iterative procedure of our MCMC algorithm described in Section 4.1. At the r-th iteration of our algorithm, the Gibbs sampler has computed model M_r having model parameters Λ_r and the transition distribution parameters β_r. In order to predict the h-step ahead forecasting distribution, we sample the households earnings y_{iT+h}^r for $i = 1, \ldots, N$ households and for $r = 1, \ldots, R$ different MCMC draws by following the steps listed below:

1. Sequentially sample the hidden states $g_{1:T}^r$ from $p(g_{1:T}^r | y_{1:T}, \beta_r, a_{1:T}^r, x_{it})$ and the Gaussian states from $p(a_{1:T}^r | y_{1:T}, \Lambda_r, g_{1:T}^r)$ for a given number of iterations up to time T, as described in Section 4.1.

2. Determine the hidden states $g_{T:T+h}^r | x_{T:T+h}, \beta_r, g_T$ by $g_{it}^r = \arg\max_k \pi_{g_{it-1},k}^r(x_{it})$ for $t = T+1, \ldots, T+h$.

3. Run the Kalman filter to compute $\mu_{it}^{(k)}$ and $V_t^{(k)}$ for $t = 1, \ldots, T+h$, $i = 1 \ldots, N$, and $k = 1, \ldots, K$ by using the model parameters Λ_r, the hidden states $g_{1:T}^r$ and the data $y_{1:T}$ up to time T.

4. Sample the Gaussian states $a_{it}^r \sim N\left(F\mu_{it-1}^{(g_{it}^r)}, FV_{t-1}^{(g_{it}^r)}F^{\mathrm{T}} + Q_r^{(g_{it}^r)}\right)$ for $t = T+1, \ldots, T+h$ given the hidden states $g_{T+1:T+h}^r$.

5. Sample the predicted household income $y_{it}^r \sim N(a_{it}^r, \sigma_{\varepsilon^r}^2)$ for $t = T+1, \ldots, T+h$ and $i = 1, \ldots, N$

Now, the h-step ahead forecasting distribution for MCMC draw r is given by all N predicted household earnings for $T+h$ (y_{T+h}^r).

4.2.2 Forecasting Criteria

To evaluate our forecasting performance, we look at four different test statistics. We first look at two individual forecast performance statistics that compare the one-step ahead forecast $\hat{y}_{iT+1}^r$ for models $r = 1, \ldots, R$ against its true realisation y_{it}. Although, we are not particularly interested

in forecasting individual households, we look at these test statistics to still give an indication of how well our model perform for households individually. To evaluate the point forecasts, we use the Root Mean Square Forecast Error (RMSFE), which we calculate for each model r apart:

$$L_r(\hat{y}_{1:N,T+1}, y_{1:N,T+1}|\lambda_r, \beta_r) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{iT+1}^r - y_{iT+1})^2 \tag{4.9}$$

In addition, we use the continuous ranked probability score (CRPS) to compare the forecasting performance of individual household. CRPS examines the performance of the density forecast by computing the squared difference between the individual forecast cumulative distribution function, $F_i^{T+1}(y)$, over the R models and the empirical CDF of the observation:

$$CRPS_{T+1} = \frac{1}{N}\sum_{i=1}^{N}\int_0^{\infty}(F_i^{T+1}(y) - \mathbb{1}[y_{iT+1} \leq y])^2 dy \tag{4.10}$$

Secondly, we analyse two distribution forecast performance statistics that compare the one-step ahead empirical forecast distribution $\hat{F}_{T+1}^r$ for models $r = 1, \ldots, R$ against the true distribution F_{T+1} (hypothesized distribution): the Cramer-von Mises test (CVM) and Anderson darling test (AD). Both criteria are used for judging the goodness of fit of a hypothesized cumulative distribution function compared to a given empirical distribution function. Since Anderson-Darling test is a modification of the Cramer-von Mises test, we can define their quadratic EDF statistics measure as follows:

$$Q^2 = n\int[F_n(x) - F(x)]^2\phi(x)dF(x) \tag{4.11}$$

where $\phi(x)$ is a weight function, F_n is the empirical cumulative distribution function (CDF), and F is the hypothesized CDF. When $\phi(x) = 1$, we consider the CVM statistic, and when $\phi(x) = [F(x)(1 - F(x))]^{-1}$, we consider the AD statistic. Thus, CVM has more power against deviations in the middle, as it measures the mean squared difference between the empirical distribution and the hypothetical CDF. AD places more weight on observations in the tails of the distribution. As our hypothesized distribution is also an empirical cumulative distribution function, we use for both tests their 2-sample test variant to compare $\hat{F}_{T+1}^r(x)$ against the empirical F_{T+1}.

Chapter 5

Evaluation

In this chapter we discuss the forecasting results for the different models we described in Chapter 3. We start this chapter by a short description about our implementation. Then in Section 5.2 we describe the initialisation of our models, and we explain our baseline models. In Section 5.3 we give an overview of the out-of-sample one-step -ahead forecasting performance of all models we consider in this study. In Section 5.4 we look in more depth to one model in particular to illustrate how the clusters are formed. In the last section, we try to forecast the future household earnings distribution from 2020 to 2029 and uses the population forecasts to correct these distribution for changes in the population.

5.1 Model Implementation

We have implemented our models entirely ourselves. It is written in Python, and it only uses Numpy[1] and Scipy[2] for fast matrix computations. Our programming code can be found on https://github.com/lhmeijer/NSSSM. Our implementation is programmed in the object oriented way such that we can use similar function without writing a lot of duplicate code. Our implementation contains a class for hidden Markov models, switching state space models and also mixture models. These classes are set up very flexibly so that one can easily add additional variables to the model; switch to another transition model; and use various distributions (e.g normal inverse Wishart) to model the model parameters.

5.2 Model Initialisation

As the initialisation of prior values can have a strong effect on the results, we use informative priors to estimate the model parameters, primarily based on the data and not on some prior knowledge. Therefore, we initialise the state disturbances using $Q^{(k)} \sim IG(\nu_0, \delta_0)$, where $\alpha = 10, \beta = [10, 10]$ for $k = 1, \ldots, K$, and the observation distribution using $R \sim IG(10, 10)$. For the initial distribution $a_{i1} \sim (\mu_0^{(k)}, V_0^{(k)})$ for $i = 1, \ldots, N$, we use some information from the dataset of 5 million households. Hence, we initialise $V_0 \sim IG(\nu_0, \delta_0)$, where $\alpha = 10, \beta = [10, 10]$ for $k = 1, \ldots, K$, and $\mu_0 \sim N(m_a^{(k)}, b_a^{(k)})$, where $b_a^{(k)} = [1, 0.1]$ for $k = 1, \ldots, K$, and $m_a^{(k)} = [x, 0]$. x is the $100/K$ cumulative log percentile over the dataset of 5 million households. Hence, for $K = 4$, $m_a^{(1)} = [9.612, 0]$, $m_a^{(2)} = [10.356, 0]$, $m_a^{(3)} = [10.706, 0]$, and $m_a^{(4)} = [11.231, 0]$. We have initialised the coefficients of the transition model $\beta_k \sim N(m_0, V_0)$, where $m_0 = [0, \ldots, 0]$ and V_0 is a $B \times B$ diagonal matrix with 0.1 on the diagonal.

Due to time restrictions our posterior results are maximally based on 1550 MCMC draws. For the different section below we use a different number of draws which we discarded as burn-in, as for some models we only need a few iterations to already achieve good results. In order to reproduce our results, we set the seed value to 1 (Numpy). Furthermore, due to our time constraints we have only considered a switching state space model with $K = 4$ (SSSM-4-NT), $K = 6$ (SSSM-6-NT), $K = 8$ (SSSM-8-NT), and $K = 10$ (SSSM-10-NT), and we have truncated the logistic

[1] https://numpy.org/
[2] ttps://scipy.org

stick-breaking process to $K = 10$ (TN-SSSM-10-NT) and $K = 20$ (TN-SSSM-20-NT) clusters. We also look at a model that uses beam sampling instead of the forward filtering-backward sampler (N-SSSM-20-NT). In addition, from our experience we concluded that the models began to perform poorly when we added the transition regressors γ_{jk} to the transition distribution (Eq. 3.8). Therefore, we only added these regressors to a model that considers six clusters (SSSM-6-T). Besides, we also examine whether the models perform differently by adding a regularisation term to the transition distribution (SSSM-4-R-NT, SSSM-6-R-NT, SSSM-8-R-NT, SSSM-10-R-NT, etc.). Finally, we study how the models will perform when the missing observation follow option 2 ($m_{it}^{(k)} = \mu_{it}^{(k)} + C_t^{(k)}(a_{it+1}^{(k)} - F\mu_{it}^{(k)})$ instead of option 1 ($m_{it}^{(k)} = \mu_{it}^{(k)}$) in the backwards sampling step of forward filtering-backward sampling step. As the previous Gaussian states are always modified based on earnings in future, we call these models the modified backwards (MB) models (MB-SSSM-6-T, MB-SSSM-6-NT, etc.). To compare the forecasting performance of our models, we propose the following baseline models:

- There is only one discrete hidden state ($K = 1$), which reduces our model to a standard state space model.
- There are six predefined fixed discrete hidden clusters, which reduces our model to six separated standard state space models (Fixed $K = 6$). We have based these predefined clusters on household composition and age. Cluster one consists of singles under 65; cluster two of single-parents under 65; cluster three of couples under 65; cluster four of couples with children under 65; cluster five of singles above 65; and cluster six of couples above 65.
- The state disturbances $\sigma^2_{\eta^{(k)}}$ and $\sigma^2_{\zeta^{(k)}}$ are set to zero, which reduces our model to a Hidden Markov model (HMM): $y_{it} = \mu_\alpha^{(g_{it})} + \mu_\theta^{(g_{it})} t + \varepsilon_{it}$, where $\varepsilon_{it} \sim N(0, \sigma^2_{\varepsilon^{(k)}})$

5.3 Out-of-Sample Forecasting Performance

Table 3 represents the one-step ahead (y_{T+1}) forecasting performance for the different models stated in Section 5.2 over an out-of-sample dataset of 25000 randomly selected households. It presents the performance for both known $y_{1:T}$ and unknown $y_{1:T}$ to show how well our models perform if we only cluster our households based on their household characteristics and if we determine their income based on the general income trend of the cluster they belong to. It mainly shows how well the models are able to predict the household earnings of newcomers. Table 5.2 also presents the performance for a higher (1300-1550) and lower (250-500) number of MCMC iterations to give a indication how well our models already perform despite being trained for less time. For the baseline models $K = 1$ and $K = 6$ we experienced that they converge even with a small number of MCMC draws. Therefore, their results in Table 3 are based on respectively 30 to 80 and 85 to 135 MCMC iterations. As we have described in Section 4.1.3, we evaluate our individual one-step ahead forecasts using the Root Mean Squared Forecast Error (RMSFE) and the continuous ranked probability score (CRPS) to respectively examine the point forecast and the density forecast. And to assess our one-step ahead forecast distribution, we use the Cramer-von Mises (CVM) and the Anderson Darling (AD) test statistic.

From Table 3 we conclude that the models have more accurate results when $y_{1:T}$ are known than when $y_{1:T}$ are unknown. Which is reasonable as the current income is mainly based on the previous income, if a household does not transition to another cluster. From Table 3 we can see that the baseline model (Fixed K=6), MB-SSSM-10-NT, and MB-TN-SSSM-10-NT are

	Known y_T 1300-1550				Unknown y_T 1300-1550				Unknown y_T 250-500			
	RMSFE	**CRPS**	**CVM**	**AD**	**RMSFE**	**CRPS**	**CVM**	**AD**	**RMSFE**	**CRPS**	**CVM**	**AD**
K=1	0.480	0.144	4.198	55.46	0.728	0.369	198.4	1869				
Fixed K=6	0.464	**0.136**	3.024	38.28	0.570	0.254	11.28	121.7				
HMM-6-NT	0.778	0.399	102.3	662.4	0.778	0.399	102.1	662.6	0.778	0.399	102.3	663.1
HMM-6-R-NT	0.772	0.275	156.6	1079	0.772	0.274	156.7	1080	0.772	0.275	157.0	1081
SSSM-6-T	0.547	0.152	7.501	104.0	0.724	0.302	110.3	746.7	0.718	0.296	87.78	601.3
SSSM-6-R-T	0.512	0.150	6.302	75.54	0.723	0.305	156.8	1041	0.723	0.302	162.8	1075
SSSM-6-NT	0.519	0.157	1.941	25.89	0.583	0.255	24.37	238.4	0.579	0.253	22.30	229.4
SSSM-6-R-NT	0.531	0.160	2.024	28.93	0.578	0.249	20.95	191.5	0.581	0.251	21.85	194.1
SSSM-4-NT	0.535	0.166	4.659	52.37	0.586	0.257	36.37	295.2	0.590	0.259	38.86	305.1
SSSM-4-R-NT	0.535	0.167	4.808	53.26	0.587	0.257	37.01	296.2	0.588	0.258	38.60	312.9
SSSM-8-NT	0.538	0.163	2.038	30.29	0.590	0.257	15.42	153.0	0.566	0.240	10.24	114.2
SSSM-8-R-NT	0.527	0.159	1.810	24.66	0.580	0.254	16.35	164.0	0.601	0.266	24.29	241.0
SSSM-10-NT	0.534	0.158	1.873	27.17	0.571	0.242	7.872	94.56	0.573	0.240	**7.925**	**74.50**
SSSM-10-R-NT	0.532	0.161	2.777	33.81	0.571	0.246	**6.475**	**74.00**	0.577	0.241	11.80	100.3
TN-SSSM-10-NT	0.523	0.156	3.113	33.97	0.553	0.242	12.78	120.9	0.569	0.238	11.70	136.6
TN-SSSM-10-R-NT	0.524	0.164	1.554	18.16	0.560	0.241	10.13	118.1	0.561	0.243	11.92	108.9
TN-SSSM-20-NT	0.586	0.159	7.153	71.90	0.582	0.237	18.67	129.2	0.598	0.239	24.08	172.4
TN-SSSM-20-R-NT	0.535	0.161	5.852	43.07	0.564	0.243	46.79	387.6	0.621	0.276	65.29	509.0
N-SSSM-20-NT	0.713	0.272	73.98	611.9	0.734	0.368	136.9	1326	0.729	0.357	123.6	1131
MB-SSSM-6-T	0.552	0.153	8.086	103.4	0.721	0.300	74.96	522.2	0.716	0.289	55.66	392.4
MB-SSSM-6-NT	0.519	0.142	1.727	19.56	0.572	0.235	40.07	401.0	0.569	0.234	45.02	400.4
MB-SSSM-10-NT	0.485	0.141	**1.350**	15.30	0.531	**0.229**	17.70	190.3	0.550	0.252	33.75	326.8
MB-SSSM-10-R-NT	0.502	0.140	1.899	23.41	0.575	0.241	69.85	525.0	**0.535**	**0.221**	20.01	218.2
MB-TN-SSSM-10-NT	**0.463**	0.142	1.394	**11.81**	**0.507**	0.230	17.37	196.7	0.540	0.231	21.01	248.0

Table 3: Forecasting performance of various models based on the root mean squared forecasting error (RMSFE), the continuous ranked probability score (CRPS), the Cramer-von Mises (CVM) and the Anderson Darling (AD) test statistic for known and unknown previous earnings $y_{1:T}$ and for a low and high number of MCMC iterations. The best performing model according to each metric has been display using bold text.

performing best. Probably the baseline model performs better than the other models, because it has less parameters to optimise than the other models, and therefore it makes a better parameter estimate. MB-SSSM-10-NT and MB-TN-SSSM-10-NT are performing better than the models without backward sampling for unknown data points, as they make better use of all income information both past and future earnings.

From Table 3 we see that SSSM-10-NT, and SSSM-10-R-NT and the baseline model (Fixed K=6) are performing rather well, although the previous earnings $y_{1:T}$ are unknown, especially on the Cramer-von Mises (CVM) and the Anderson Darling (AD) test statistic. However, these models are less good at predicting individual earnings than the models using option 2 in the backward sampling step. Table 3 also illustrates that for most models the performance increases when we consider more MCMC iterations, only for some models the performance decreases, such as the models MB-SSSM-10-R-NT, SSSM-8-NT, and SSSM-10-NT. We may have optimized the parameters that they too closely correspond to the training dataset. Hence, overfitting makes the models less capable of predicting the earnings of the households in the test dataset.

Moreover, Table 3 shows that the baseline model (Fixed $K = 6$) is performing quite well with only six clusters compare to model SSSM-6-NT. It additionally forecasts the earnings much better than the other baseline models. A hidden Markov model (HMM) is less accurate at estimating the earnings than a switching state space model, but a HMM performs better than the state space model ($K = 1$) if the previous earnings are unknown. Hence we conclude that household clustering definitely affects the model performance, especially for households entering the dataset or a cluster some time in the future. From Table 3 we can also conclude that beam sampling

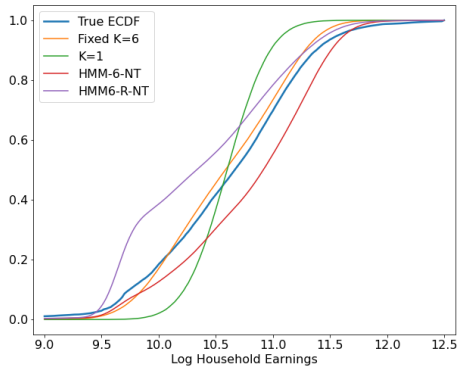
(Section 4.1.1.1) worsen the prediction accuracy significantly, and as Fox et al. (2011b) already suggested, we can better use a "weak limit" truncation of the logistic stick breaking process, particularly for 10 clusters. Namely, TN-SSSM-10-NT and TN-SSSM-10-R-NT have better results than TN-SSSM-20-NT and TN-SSSM-20-R-NT. It is possible that 1550 MCMC iterations is insufficient for $K = 20$. Nevertheless, between TN-SSSM-10-NT and SSSM-10-NT we do not see major result differences. SSSM-10-NT and SSSM-10-R-NT just forecast a bit better than TN-SSSM-10-NT and TN-SSSM-10-R-NT. However, TN-SSSM-10-NT is performing best over all models on the CVM and AD test statistic if we only use the 100 to 150 MCMC iterations: CVM of 1.005 and AD of 11.230.

In general, from Table 3 we conclude that the models with 10 clusters are performing better than the models with a lower number of clusters. Hence, it seems that with more clusters we can better distinguish the different households and income groups. We also conclude that the models without transition regressors (γ_{jk}) (SSSM-6-NT, SSSM-6-R-NT) have lower forecasting statistics than model with these regressors (SSSM-6-T, SSSM-6-R-T). Because of these regressors households may tend to stay in the same cluster even more, and whether they transition to another cluster is primarily determined by which cluster it came from rather than what characteristics the households have. Nonetheless, we cannot directly conclude from Table 3 whether the models are better able to identify the groups when we add a regularisation term to the transition equation. Their results are almost comparable. Possibly 25 exogenous covariates is too few to actually see a significant difference.

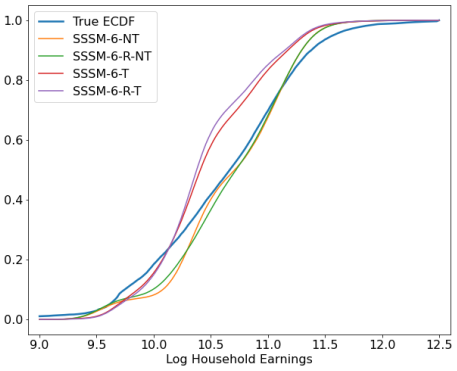
Table 3 illustrates that some models are performing rather well despite unknown previous earnings $y_{1:T}$. To obtain an even better understanding in how well these model perform, Figure 5 presents their empirical cumulative distributions (ECDFs) against the true ECDF of the test dataset of 25000 households. Figure 5a presents the ECDFs of the baseline models with a single cluster, six fixed clusters, a HMM, and a regularized HMM. It shows that the ECDF of fixed $K = 6$ follows the true ECDF fairly well. It also shows that without clustering the model is unable to match the true distribution properly. Figure 5b presents the ECDFs of the four models considering six clusters. It shows that with six clusters we are still quite capable of matching the true distribution, but we overestimate the low-income groups. Figure 5c displays the ECDFs of the models considering 10 clusters. It illustrates that indeed these models are best at matching the true distribution when the previous earnings are unknown. Lastly, Figure 5c shows the ECDFs of the models considering 10 clusters with option 2 in the backward sampling step. It illustrates that just as with six clusters we overestimate the low-income groups.

5.4 Group Membership

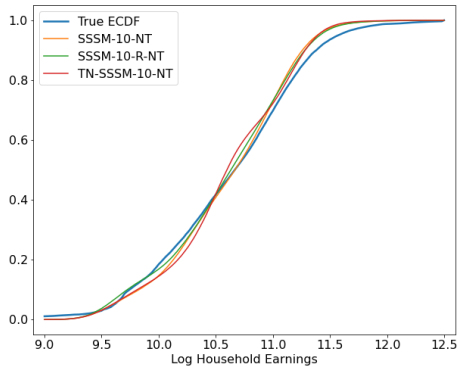
In this section we look in more depth to model SSSM-10-R-NT. As demonstrated in Table 3 SSSM-10-R-NT is performing relatively well for both known $y_{1:T}$ and unknown $y_{1:T}$. We included similar tables and figures for model SSSM-10-NT, MB-SSSM-10-NT, and TN-SSSM-10-NT respectively in Appendix 1, but we will not elaborate on their comparison. The figures and tables which we present in this section are based on the predicted cluster memberships and predicted household earnings over 250 MCMC iterations (1300-1550) for which the previous household earnings were unknown. In most figures we only present the 50% percentile over these MCMC iterations, because we hardly see any difference between the 5% and 50% percentile and



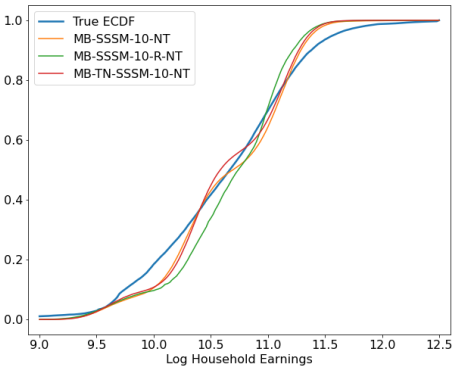
(a) Baseline Models



(b) SSSMs 6 Clusters



(c) SSSMs 10 Clusters



(d) MB-SSSMs 10 Clusters

Fig. 5: Empirical Cumulative Distribution Functions (ECDFs) of various models against the true ECDF of the test dataset.

the 50% and 95% percentile. We have included the 5% and 95% percentile in the appendix for comparison.

5.4.1 Model Parameters

We start this section by presenting the estimated model parameters for $k = 1, \ldots, K$ in Table 4 and Table 5. These tables give the 5%, 50%, and 95% over the $r = 1300, \ldots, 1550$ model parameters (Λ_r) in order to show their deviation between the MCMC iterations. Table 4 displays the estimates of the initialisation distributions for the different clusters $k = 1, \ldots, 10$ specified in Equation 3.4, and Table 5 shows the estimated state and observation disturbances for $k = 1, \ldots, 10$ specified in Equation 3.1. We can see from Table 4 that cluster 5 and 6 and clusters 7 and 8 have a similar distribution for the group-specific intercept, but their group-specific trend differ. Clusters 5 and 7 have a negative trend, while clusters 6 and 8 have a positive trend. Table 4 also shows that the model parameters are rather similar across the MCMC iterations and that the variance for both intercept and trend are relatively small, except for cluster 4. In addition, the trend of cluster 4 is much larger than the trend of other clusters. From Table 5 we can see that the state disturbances are quite small, except for cluster 4, and again they are similar over the MCMC iterations.

	μ_α			σ^2_α			μ_θ			σ^2_θ		
K	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
1	9.5515	9.5541	9.5568	0.0109	0.0113	0.0118	-0.0022	0.0004	0.0031	0.0045	0.0047	0.0048
2	9.7600	9.7730	9.7850	0.0530	0.0560	0.0600	-0.0110	-0.0050	0.0010	0.0080	0.0080	0.0090
3	10.0370	10.0410	10.0460	0.0380	0.0400	0.0410	0.0110	0.0140	0.0170	0.0050	0.0050	0.0050
4	10.174	10.1880	10.2000	0.6590	0.6710	0.6860	0.0870	0.0900	0.0920	0.0030	0.0030	0.0030
5	10.3362	10.3439	10.3521	0.0651	0.0683	0.0717	-0.0106	-0.0066	-0.0035	0.0069	0.0072	0.0074
6	10.3763	10.3836	10.3928	0.0615	0.0644	0.0677	0.0115	0.0139	0.0168	0.0055	0.0057	0.0059
7	10.7343	10.7454	10.7574	0.0876	0.0913	0.0955	-0.0187	-0.0163	-0.0130	0.0039	0.0040	0.0042
8	10.7932	10.7988	10.8044	0.0603	0.0626	0.0653	0.0063	0.0094	0.0119	0.0042	0.0043	0.0045
9	10.9782	10.9854	10.9924	0.0657	0.0679	0.0703	0.0092	0.0120	0.0143	0.0044	0.0045	0.0047
10	11.1836	11.1886	11.1933	0.0399	0.0412	0.0426	0.0053	0.0093	0.0143	0.0045	0.0046	0.0047

Table 4: The 5%, 50%, and 95% percentile of the estimates of the model parameters ($\mu_\alpha^{(k)}$, $\sigma^2_{\alpha(k)}$, $\mu_\theta^{(k)}$, $\sigma^2_{\theta(k)}$) for $k = 1, \ldots, 10$ over 250 MCMC iterations for model SSSM-10-R-NT.

	K	1	2	3	4	5	6	7	8	9	10	σ^2_ε
$\sigma^2_{\eta(k)}$	5%	0.0106	0.0139	0.0140	0.1902	0.0167	0.0159	0.0229	0.0153	0.0171	0.0151	0.0041
	50%	0.0108	0.0143	0.0143	0.1925	0.0170	0.0162	0.0236	0.0156	0.0175	0.0156	0.0042
	95%	0.0110	0.0147	0.0147	0.1948	0.0174	0.0166	0.0243	0.0160	0.0179	0.0159	0.0043
$\sigma^2_{\zeta(k)}$	5%	0.0029	0.0038	0.0029	0.0020	0.0035	0.0030	0.0029	0.0029	0.0032	0.0033	
	50%	0.0030	0.0040	0.0030	0.0021	0.0036	0.0031	0.0030	0.0030	0.0032	0.0034	
	95%	0.0030	0.0041	0.0030	0.0021	0.0037	0.0032	0.0030	0.0031	0.0033	0.0036	

Table 5: The 5%, 50%, and 95% percentile of the state disturbances estimates (σ^2_η, σ^2_ζ) for $k = 1, \ldots, 10$ and the 5%, 50%, and 95% percentile of the observation disturbance estimates over 250 MCMC iterations for model SSSM-10-R-NT.

5.4.2 Cluster Transitions

We also look at how the households in our dataset transition from one cluster to another cluster over the years. Figure 6 illustrates a Sankey diagram of these cluster flows. The numbers in the diagram correspond to the numbers in Table 4 and 5. Thus, cluster 1 denotes the lowest-income group and cluster 10 indicates the highest-income group. The size of the vertical bars indicates the size of the clusters. Thus, we see that some clusters are larger than other, in particular cluster 4. The thickness of the horizontal illustrates the number of households flowing from one cluster to another cluster. We see that in general households remain in the same income group over time and only some households transition to another cluster. From cluster 1 households primarily transition to cluster 2 and 4. From cluster 2 households move to cluster 4. From cluster 3 they shift to cluster 4, 6, and 7. From cluster 4 they transition to all other clusters evenly. From cluster 5 households mainly transition to cluster 4 and 7. From cluster 6 they shift to cluster 4 and 7. From cluster 7 they move to cluster 3, 4, 5, 8, 9, and 10. From cluster 8 households shift to 4, 7, and 9; from cluster 9 to 4, 7, 8 and 10; and from cluster 10 to 4, 7 and 9. Hence, in general high-income groups may still transition to other high- or middle-income-groups but hardly to low-income-groups, and low-income groups shift to other low- or middle-income groups but barely to high-income groups. Nevertheless, notable is cluster 4, as many households from all clusters regularly transition to it and many also leave it. Cluster 4 seems to be a remaining income group into which households end up if they are poorly identifiable or do not match another income group.

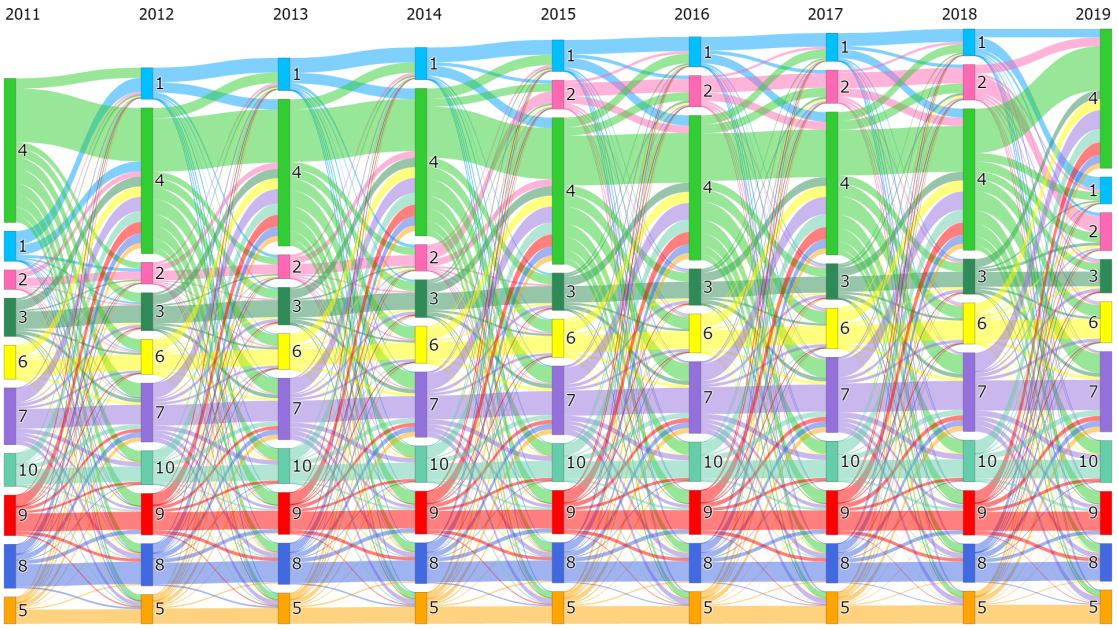
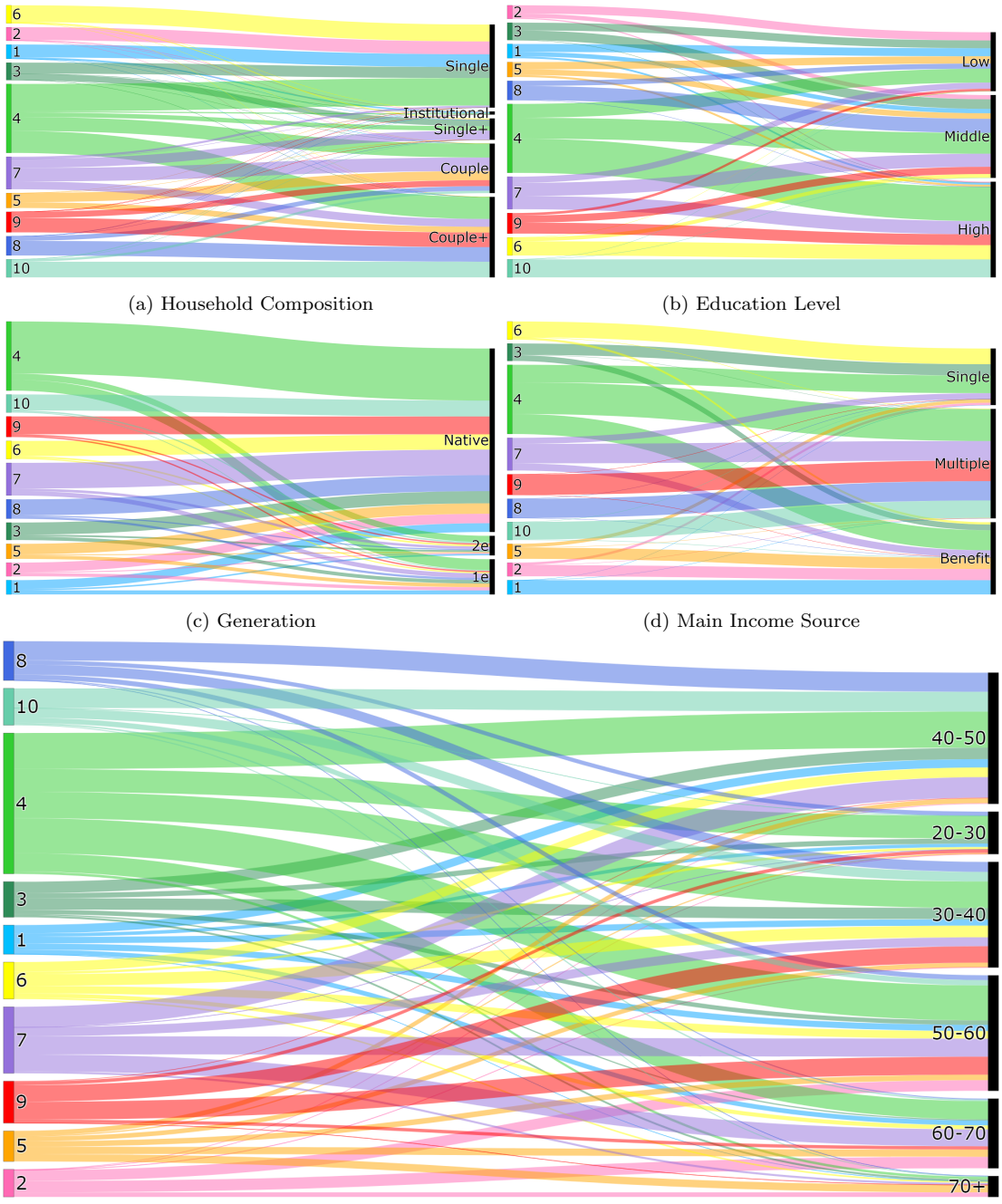


Fig. 6: A Sankey Diagram of the cluster transitions over the years 2011 to 2019 for model SSSM-10-R-NT. The clusters are ordered: the lowest cluster number represents the lowest-income cluster, and the highest cluster numbers implies the highest-income cluster.

5.4.3 Cluster Characteristics

In this sub-section we examine how the household characteristics are distributed among the clusters. Figure 7 illustrates 5 diagrams that illustrate the number of households in a cluster summed over all years (2011-2019) that has a certain household characteristic: household composition, education level, generation, main income source, and age. Again, the numbers in the diagram correspond to the numbers in Table 4 and 5. Figure 7a illustrates that cluster 1, 2, 3 and 6 mainly consist of single households; that cluster 5, 8, 9 and 10 primarily contain couples; and cluster 4 and 7 consists of all different household compositions. Figure 7b shows that cluster 1, 2, 3, and 8 mainly contain low and middle educated households; cluster 6 and 10 consists of high educated households; and cluster 4, 7, and 9 are not necessarily identifiable by education level. From Figure 7c we cannot clearly identify which clusters represent a specific generation. Only cluster 1 has relatively more first generation households compared to the other clusters. Figure 7d illustrates that cluster 1, 2 and 5 primarily consists of benefit recipients; that cluster 6 and 3 contain single earners; and that cluster 8, 9 and 10 consists of multiple earners. Again cluster 4 and 7 are difficult to identify based on the main income source. For Figure 7e we have merged some age categories in order to simplify and clarify the diagram. Figure 7e shows that in our dataset 70+ households mainly transition to cluster 2, 4 and 5; that young households (20-30) shift to cluster 4; and that cluster 3, 8 and 10 mainly consists of households between the age (30-50). In conclusion, cluster 1 consists of single benefit recipients between 25 to 70 with a low education level. Cluster 2 contains single 50+ benefit recipients. Cluster 3 consists of single households between 20 and 50 with a low or middle education level. Cluster 4 is not identifiable based on these categories. Cluster 5 consists of 50+ couples receiving a benefit with low or middle education level. Cluster 6 contains high educated single households without children. Cluster 7 mainly consists of multiple earners between the ages of 30 to 60. Cluster 8 contains low or middle educated multiple-earners between the ages of 20 to 50 with children. Cluster 9 consists of middle or high educated multiple-earners with or without children. And cluster 10

contains high educated multiple-earners with children. Hence, this group division reasonably corresponds to the results we presented in Figure 2, such as couples earning more than single households, or high educated households earning more than low educated households.



(a) Household Composition

(b) Education Level

(c) Generation

(d) Main Income Source

(e) Age

Fig. 7: 5 diagrams representing the number of households with a certain household characteristic in a particular cluster $k = 1, \ldots, 10$ for model SSSM-10-R-NT, respectively, household composition, education level, generation, main income source, and age.

5.4.4 Matching True Earnings

Lastly, we look at how well a cluster represents its true earnings. Therefore, we compare the median calculated over the true earnings with the median computed over the predicted earnings for each cluster individually and the years 2011, 2015, and 2019. These results are presented

in Table 6. In Appendix 1 we also look at the 5% and 95% percentile. Table 6 shows that in 2011 the predicted median of cluster 1, 2, 3, 5, 7, and 8 closely matches the true median of that respective cluster. In 2015 the predicted median of cluster 1, 3, 4, 8, 9, and 10 closely matches the true median. In 2019 only for cluster 3 and 6 we are still getting close to the true median. For the other clusters, our predictions are often lower than the true earnings. For cluster 2, 5 and 8 we even estimate that the earnings will decline, while in reality the earnings increase. Probably because for these clusters the household earnings declined between the years 2011 and 2012. Table 6 also shows that the earnings of cluster 4 households will increase at an unrealistic rate. This rate mainly represents the rate of an individual household in cluster 4 rather than the group rate. Which might be due to the selection of the dataset that some groups are unrepresented in some years. Nevertheless, eventually Table 6 shows that the model SSSM-10-R-NT is reasonable good at predicating the median household earnings without knowing a household's previous income ($y_{1:T}$). Mainly for the year 2019, we underestimate the median income of most clusters.

		1	2	3	4	5	6	7	8	9	10
2011	median($\hat{y}$)	14.10	17.54	22.96	26.57	31.07	32.33	46.43	48.97	59.0	72.29
	median(y)	14.32	17.64	23.01	34.51	32.3	30.45	44.05	47.66	55.24	68.43
2015	median($\hat{y}$)	14.13	17.17	24.29	37.98	30.24	34.21	43.5	50.86	61.94	74.94
	median(y)	14.64	18.32	23.92	37.83	32.17	32.04	45.16	51.49	59.59	73.04
2019	median($\hat{y}$)	14.16	16.84	25.7	54.41	29.46	36.17	40.74	52.8	65.03	77.75
	median(y)	15.49	19.58	25.45	44.68	34.17	34.02	49.35	58.1	68.0	80.21

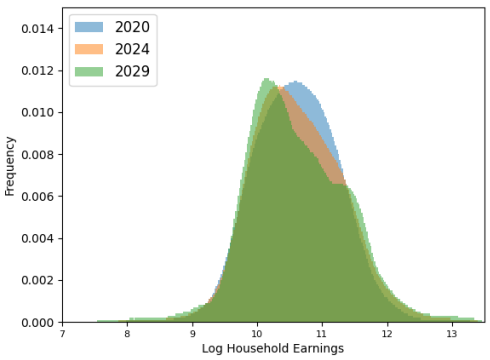
Table 6: Median calculated over the predicted households earnings $\hat{y}$ and median computed over the true household earnings for cluster $k = 1, \ldots, 10$, the years 2011, 2015, and 2019, and model SSSM-10-R-NT using the 50% percentile over the 250 MCMC draws.

5.5 Forecasting Household Earnings Distribution

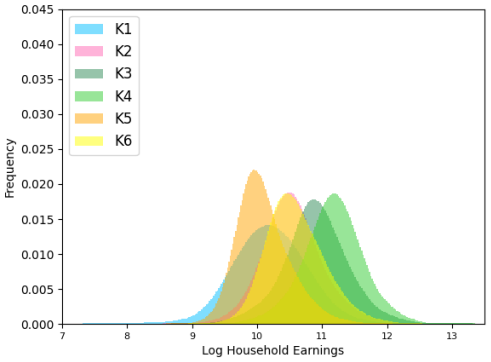
In this last section, we try to forecast the household earnings from 2020 to 2029, and use the population forecasts to correct the income distribution forecasts for a changing population. We first apply the steps described in Section 4.2.1 to obtain a prediction for 1.6 million households y_{iT+h} $h = 1, \ldots, 10$ (1). This dataset consists mainly of existing households which we extrapolate through time, but also consists of imaginary households that enter the dataset in the future. We assumed that households leave the dataset when they are older than 85, and we randomly select a small number of household (5%) that lose their jobs, get divorced, get married, or get a job. We then select all households having a certain age, household composition, generation, education level, and main income source (2). We determine the distribution for this group by grouping the household in particular income bins (3). We divide these bins by the total number of households (4), and multiply these frequency bins by the number of households given by the population forecast (5). If we now sum up these bins across all 831 groups and divide them by the total number of households in the population forecast, we obtain a new income distribution corrected for changes in the population. Unfortunately, these distributions were rather erratic, and therefore we use a smooth spline approximation (Dierckx, 1975) to smooth the households earnings distributions.

Figure 8 and 9 represents these forecasting distributions for 2020, 2024, and 2029; for respectively the baseline model (Fixed $K = 6$) and model SSSM-10-R-NT; and for different clusters

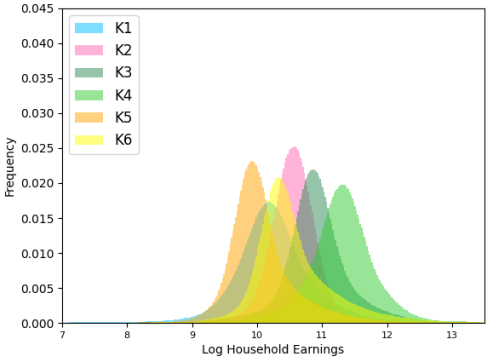
$K = 1, \ldots, K$. We cannot compare the distribution of 2020 against the distribution of 2020 presented in Figure 1b, as the x-axis of these distributions is in log scale and their population structures differ. In Appendix 1, we have presented similar figure for the models SSSM-10-NT, and TN-SSSM-10-NT, but we will not discuss these figures further. We have trained SSSM-10-R-NT and the baseline model respectively on 1550 and 135 MCMC draws with a burn-in sample of 1300 and 85 iterations. Looking at Figure 8a, it shows that the peak of 10.7 log earnings will move towards 10.1 log earnings. This is probably because we get more elderly in the future population and since their household earnings will decline (see Fig. 8b and 8d ($K5$ and $K6$)), the overall income will also decrease. Moreover, Figure 8a demonstrates that the household earnings above 10.3 will become more dispersed over a larger income range. This is consistent with the trend we already saw between the years 2011 and 2020 in Figure 1. In Figure 9a we also see a similar trend, although in this figure its tail even gets fatter, primarily, because of cluster 4. As presented in Figure 9b, 9c, and 9d, cluster 4 tends to shift quickly to the right. Despite the fact that a fatter tail might be realistically in the future, the size of this shift is rather unlikely. Figure 9a also has a similar peak around 10.5 as Figure 8a, but it only shifts a little to the left. Additionally, we can see a small peak around 9.8 log earnings in 2029, which is probably due to the smooth spline approximation. We do not detect this peak clearly in earlier years. In both Figure 8 and 9, we notice that the distributions become taller and narrower over time. Which might be due to households shifting to another clusters over time, and therefore, their income is only based on the general cluster income trend instead of their previous known earnings.



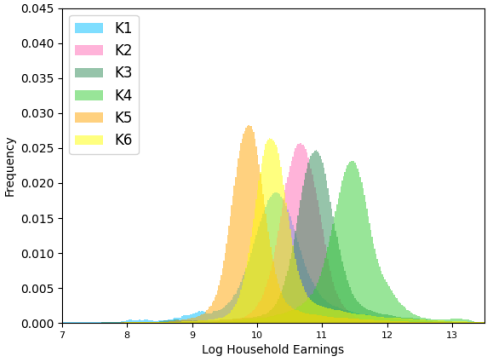
(a) Total forecasted distributions by baseline model for 2020, 2024, and 2029



(b) Forcasted distributions by baseline model for the clusters $K = 1, \ldots, 6$ and 2020

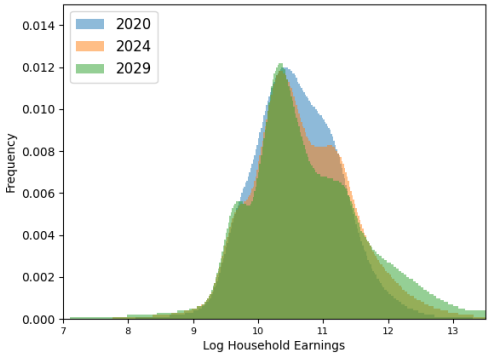


(c) Forcasted distributions by baseline model for the clusters $K = 1, \ldots, 6$ and 2024

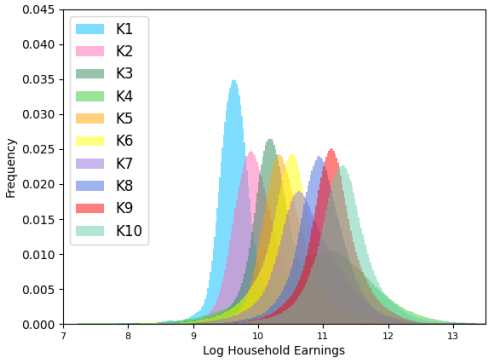


(d) Forcasted distributions by baseline model for the clusters $K = 1, \ldots, 6$ and 2029

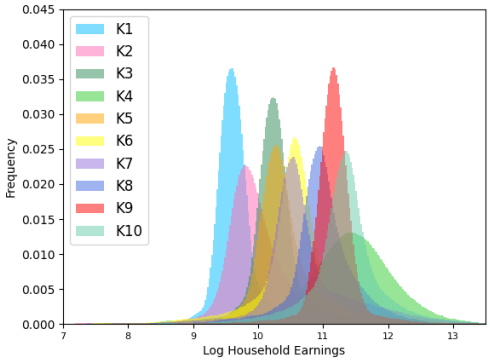
Fig. 8: Forecasted Household Earnings Distributions for the baseline model with fixed six clusters $K = 1, \ldots, 6$.



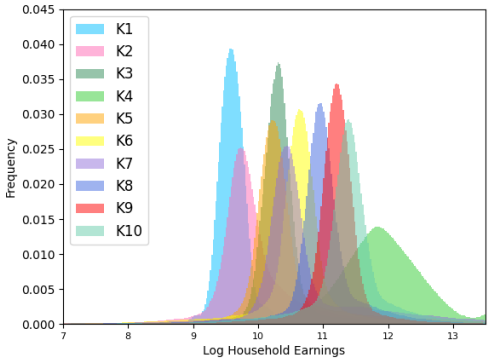
(a) Total forecasted distributions by SSSM-10-R-NT for 2020, 2024, and 2029



(b) Forcasted distributions by SSSM-10-R-NT for the clusters $K = 1, \ldots, 10$ and 2020



(c) Forcasted distributions by SSSM-10-R-NT for the clusters $K = 1, \ldots, 10$ and 2024

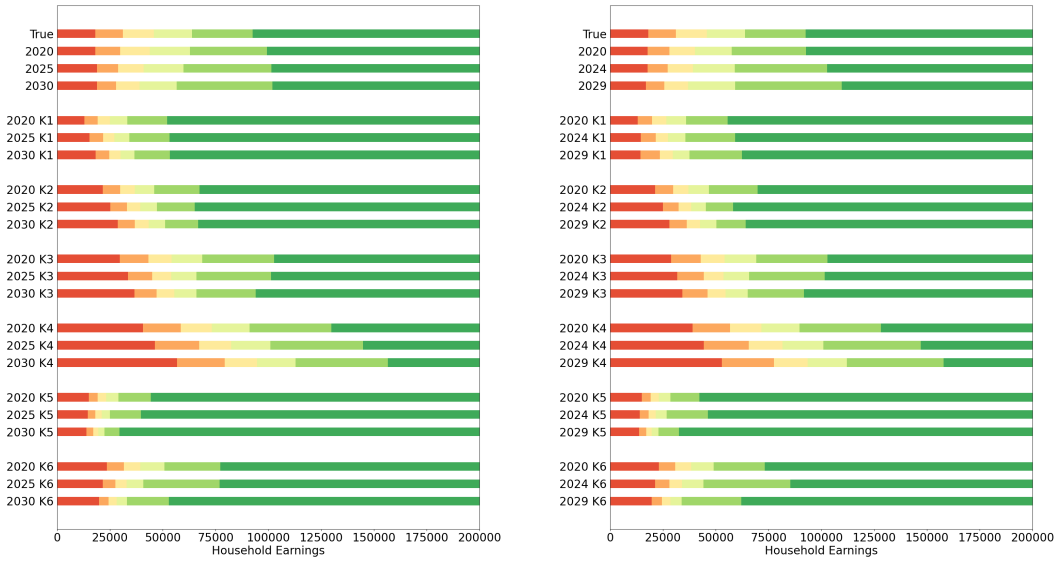


(d) Forcasted distributions by SSSM-10-R-NT for the clusters $K = 1, \ldots, 10$ and 2029

Fig. 9: Forecasted Household Earnings Distributions for model SSSM-10-R-NT with $K = 1, \ldots, 10$ clusters.

To get a better idea of the percentiles of these distributions, Figure 10 and 11 presents the 10%, 30%, 50%, 70%, and 90% percentile of respectively the baseline model and the model SSSM-10-R-NT, in a similar fashion as in Figure 2. For clarification, the color transitions imply the percentile limits. In appendix 1 we have included similar figures for model SSSM-10-NT and TN-SSSM-10-NT. The right figures (Fig. 10b and 11b) display the percentiles computed over the distribution given in Figure 8 and 9, and the left figures (Fig. 10a and 11a) show the percentiles calculated over the predicted household earnings of 1.6 million households uncorrected for the population changes. With these figures we demonstrate the influence of the population forecasts. The 'True' percentiles imply the percentiles calculated over 1.6 million households in 2020. From Figure 10a and 11a we conclude that both models are relatively well in predicting the 10%, 30%, 50%, and 70% percentiles of 2020, but overestimate the 90% percentile. However, from Figure 10b and 11b we see that the percentiles will be slightly lower if we correct the distributions for population changes, primarily, because the population forecasts expect to contain more elderly than our dataset of 1.6 million households (Appendix 1). Figure 8 demonstrates that the 10%, 30%, 50%, and 70% percentile predicted by the baseline model will decline or remain roughly the same, and the 90% percentile will increase, but looking at the clusters in particular the 10%, 30% and 50% percentile of cluster 1 to 4 will increase, especially the earnings of household belonging to cluster 4 (couples with children under 65). For cluster 5 and 6 all percentiles decreases and for cluster 1, 2, and 3 only the 90% percentile does. Figure 11 illustrates a similar pattern as in Figure 8. The 10%, 30% and 50% decreases over time, but the 70% and 90% increases. The

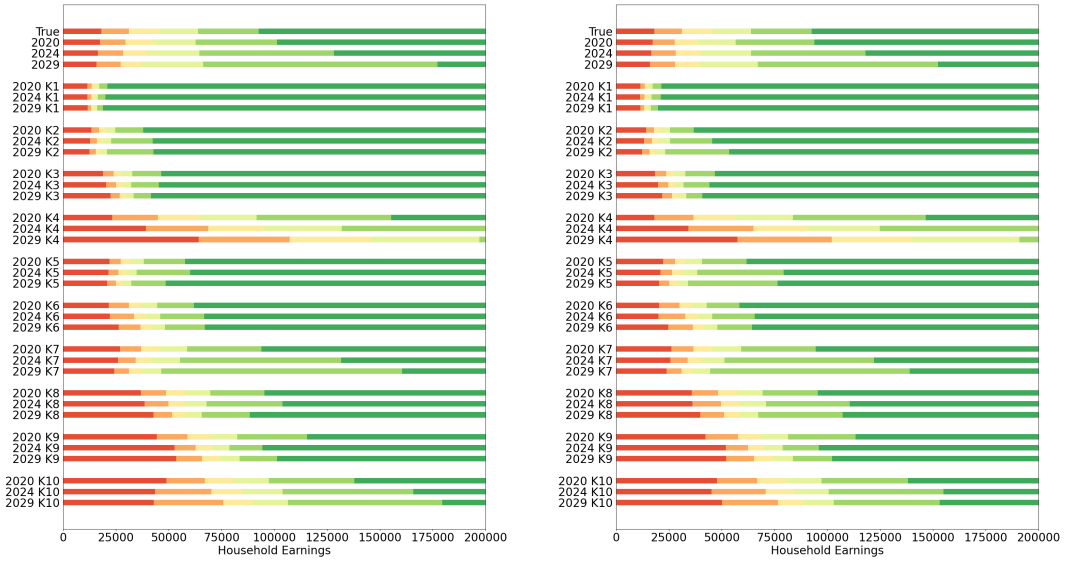
90% percentile even grows at an unrealistic rate, which is primarily due to the unreal growth of cluster 4 and somewhat cluster 7. The percentiles of the other clusters only slightly change over time, which is also fairly similar to the percentiles representing in Table 1. The percentiles presented in Table 1 slightly changes over time, but for low-income households (10%/30%) less than for high-income households (70%/90%).



(a) Uncorrected for population forecasts

(b) Corrected for population forecasts

Fig. 10: 10%, 30%, 50%, 70%, and 90% percentiles for 2020, 2024, 2029 and clusters $K = 1, \ldots, 6$ forecasted by the baseline model with six fixed clusters. The color transitions denote the income percentiles.



(a) Uncorrected for population forecasts

(b) Corrected for population forecasts

Fig. 11: 10%, 30%, 50%, 70%, and 90% percentiles for 2020, 2024, 2029 and clusters $K = 1, \ldots, 10$ forecasted by model SSSM-10-R-NT. The color transitions denote the income percentiles.

Chapter 6

Discussion & Conclusion

In this chapter we briefly discuss some possible improvements and future work for both our experiments and our models (respectively Section 6.1.1 and 6.1.2), and we end this chapter with a conclusion in Section 6.2.

6.1 Discussion

6.1.1 Experiments

Even though we have presented the forecasting performance of many different models, our evaluation approach has some limitations. We did not use one or multiple synthetic dataset(s) to compare the performance of our model. Such an experimental set up would allow us to assess the performance of our models in a controlled environment and hence obtain insight in the models' working on predefined aspects, such as finding the correct clusters and accurately estimating the model parameters. Now we have applied our models directly on household earnings data, for which the clusters and parameters are fully unknown, and therefore, we are not sure to what extent our models are able to estimate the correct parameters and groups. Another limitation of our approach is that we did not use a structured experiment to assess whether the parameters really converged. Although, we have looked at the performance for both 250-500 and 1300-1550 MCMC iterations, it is still a small indication. For some models we even get a better result for a small number than for a large number of MCMC draws. A line plot over the 1550 iterations would have given more information about the convergence of our models. We were limited by how much data we were allowed to get out of the secure CBS micro data environment, and therefore we have chosen not to take this out of the environment. In addition to model convergence, we also did not examine the model initialisation in a structured way. The initialisation of the clusters and prior distributions may have had a lot of influence on the results. Therefore, for future research we recommend looking at different initialisation processes using different seed values, to be able to distinguish the effect of the initialization from other factors such as the number of the clusters or the regularisation term. Another point of discussion is our use of frequentist evaluation metrics instead of Bayesian evaluation metrics, such as Expected log-predictive density or Watanabe-Akaike Information Criteria. We were mainly interested in evaluating the predicted earnings distribution and not the performance of individual household forecasts. Therefore, we have chosen to use the Cramer-von Mises and the Anderson Darling test statistics. And lastly it would have been interesting to compare our model results against a baseline from the literature instead of our predefined baselines, so we could show whether our models match or even exceed the state-of-the-art.

6.1.2 Models

Although some of our models are performing quite well, they still have a number of limitations. First of all we truncate the non-parametric SSSMs to a fixed truncation level, $K = 10$ or $K = 20$, which is a rather low number for a dataset of 106149 households. It would have been better to set this number K to 1000 or higher, so that the clusters later in the stick breaking process will get such a small probability that only a few households will end up in

these clusters. However, due to time and memory limitations we were not able to test such a model. All experiments had to be performed in the CBS micro data environment where relatively little memory and computational power is available. Moreover, Gibbs sampling can suffer from slow convergence in high dimensional spaces leading to an overall high computational cost. We already limited the computation time by using a small number of MCMC iterations, and saw no real improvements by iterating much longer. As an alternative to Gibbs sampling Variational Bayesian (VB) inference methods may be used to lower the computation time. Rather than collecting a set of samples, these methods attempt to find a tractable distribution that most closely matches the true posterior and subsequently try to solve an optimization problem over this tractable distribution to approximate the true posterior. Nevertheless, Gibbs sampling is easy to implement and may have higher accuracy.

Moreover, Variational Bayesian (VB) methods avoid the label-switching problem. Because our hidden states depend on exogenous covariates, the label-switching problem is less likely to occur. Still, it would be better to cope with the label-switching problem and to rule out that this problem leads to inaccurate results. For computing the earnings distribution this might not be so much of an issue, but for computing the individual household earnings it does matter, because we may cluster the households wrongly. Another approach to deal with the label switching problem would be to enforce an ordering among the initial state parameters as $\mu_0^{(1)} < \ldots < \mu_0^{(K)}$, or to use informative priors, which both lead to better identification of the MCMC samples. The latter approach might even improve our model performance, since we could define the clusters in such a way that they are more distinctive.

Another change that may improve computation time would be to compute $p(y, a)$ in a different way. Instead of iterating over t, we could calculate $p(y_{it}, a_{it}|\Lambda^{(k)})$ as $p(a_{it}|F^{t-1}\mu_0^{(k)})p(y_{it}|a_{it,t})$ $p(a_{it,t}|a_{it,t-1}, Q^{(k)})$. Rather than comparing $a_{it,1}$ with $\mu_0^{(k)}$, we could compare $a_{it,t}$ with $F^{t-1}\mu_0^{(k)}$. This approach could also improve cluster allocations, as $a_{it,t}$ always contains information about the actual households' income, while $a_{it,1}$ does not. In Section 5.4 we saw that some clusters have a relative large income growth. This growth represents the growth of an individual household instead of the general change of the cluster. As a result, the predicted income of new entrants is relatively high compared to their actual income. For instance, the income of young working households often grow faster than the earnings of older working households. We may improve this by splitting the trend component into a permanent and transitory trend component: $\theta_{it}^{(k)} = \theta_{it-1}^{(k)} + \rho^{(k)} c_{it}^{(k)} + \zeta_{it}^{(k)}$. Where $c_{it}^{(k)}$ represents the time a household belongs to cluster k in binary form, and where $\rho^{(k)}$ corresponds to the transitory trend component.

Another point of discussion is our dataset of 106149 households we used for training our models. This dataset has some odd peaks in its income distributions and has a rather low number of households for some categories (e.g. age 20-25, 70+). We would have been better off using a larger dataset that would represent all groups well. However, due to time and memory constraints we were not able to use a larger dataset. Hence, instead of randomly selecting some households, we should have chosen our dataset more precisely so that it properly represented all groups.

It would probably be better to have the income trend component (θ_{it}) not vary over time: $\alpha_{it}^{(k)} = \alpha_{it-1}^{(k)} + \theta^{(k)} t + \eta_{it}^{(k)}$. Often the initial value of θ_{it} adopts the difference between 2011 and 2012, which is negative in some clusters. In subsequent years it remains negative, while these years have a positive income growth. In addition, a longer time series would capture this trend even better.

Finally, it would have been interesting to generate a dataset that exactly matches the population forecast instead of correcting the earnings distribution afterwards (Section 5.5). With

such an approach it might not be needed to smooth the distributions and we may obtain more accurate results.

6.2 Conclusion

In this study we have developed a Bayesian (non-)parametric non-homogeneous switching state space model to forecast the earnings distribution of households living in the Netherlands. Our models try to simultaneously capture a time-varying hidden group structure among the households and optimise a separate state space model associated with each group. As a result, we are able to model highly non-linear patterns in the annual income distribution; to account for a large increase or decrease in household earnings due to ageing, marriage, or job loss; and to capture gradually changing income processes by using time-varying model parameters. Our results demonstrate that the use of these time-varying clusters and time varying parameters notably improve the forecasting performance. Our models are more accurate at forecasting the one-step-ahead income distribution than a single state space model or a non-homogeneous hidden Markov model, even with households entering the dataset in the forecasting year itself. Especially, our models with ten clusters obtain better results than the models with fewer clusters. Moreover, our results show that the clusters, which our models identify, match the prior knowledge about income groups, such as couples generally earning more than singles or young households commonly earning less than middle-age households. Nevertheless, we do not see a significant performance difference between models using a regularisation term or the models which do not, and we cannot strongly conclude whether our models using the logistic stick-breaking process are performing more accurately than the models using a multinomial logistic model as cluster transition model. We recommend future research to study these differences in more depth. Furthermore, we conclude that the modified backward models (option 2 in backward sampling step) forecast the individual households earnings better than the models using option 1 ($m_{it}^{(k)} = \mu_{it}^{(k)}$), since these models make use of the previous earnings to a greater degree. However, the models using option 1 are better at predicting the households earnings of newcomers, since these models are better at identifying the various clusters. Finally, notable is the performance of the baseline model with six fixed clusters. This baseline model is performing relatively well. Besides this model is much faster in practise, as it does not need to determine the hidden clusters.

Bibliography

Altonji, J. G., Hynsjo, D. M., and Vidangos, I. (2021). Marriage dynamics, earnings dynamics, and lifetime family income. Working Paper 28400, National Bureau of Economic Research.

Altonji, J. G., Hynsjö, D. M., and Vidangos, I. (2022). Individual earnings and family income: Dynamics and distribution. Working Paper 30095, National Bureau of Economic Research.

Arellano, M., Blundell, R., and Bonhomme, S. (2017). Earnings and consumption dynamics: A nonlinear panel data framework. *Econometrica*, 85(3):693–734.

Bonhomme, S., Lamadon, T., and Manresa, E. (2022). Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, volume 5 of *JMLR Proceedings*, pages 73–80. JMLR.org.

Dierckx, P. (1975). An algorithm for smoothing, differentiation and integration of experimental data using spline functions. *Journal of Computational and Applied Mathematics*, 1(3):165–184.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. OUP Oxford.

Fernández-Val, I., Gao, W. Y., Liao, Y., and Vella, F. (2022). Dynamic heterogeneous distribution regression panel models, with an application to labor income processes. *arXiv preprint arXiv:2202.04154*.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011a). Bayesian nonparametric inference of switching dynamic linear models. *IEEE Trans. Signal Process.*, 59(4):1569–1585.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011b). A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056.

Fröhwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89.

Frühwirth-Schnatter, S. (2001). Fully bayesian analysis of switching gaussian state space models. *Annals of the Institute of Statistical Mathematics*, 53(1):31–49.

Frühwirth-Schnatter, S. (2004). Efficient bayesian parameter estimation. *State Space and Unobserved Component Models: Theory and Applications*.

Gael, J. V., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden markov model. In Cohen, W. W., McCallum, A., and Roweis, S. T., editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1088–1095. ACM.

Geweke, J. and Keane, M. (2000). An empirical analysis of earnings dynamics among men in the psid: 1968–1989. *Journal of Econometrics*, 96(2):293–356.

Ghahramani, Z. and Hinton, G. (2000). Variational learning for switching state-space models. *Neural computation*, 12:831–64.

Gu, J. and Koenker, R. (2017). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics*, 35(1):1–16.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1):39–70.

Held, L. and Holmes, C. C. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145 – 168.

Holsclaw, T., Greene, A., Robertson, A., and Smyth, P. (2017). Bayesian non-homogeneous markov models via polya-gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics*, 11.

Hoskovec, L., Koslovsky, M. D., Koehler, K., Good, N., Peel, J. L., Volckens, J., and Wilson, A. (2022). Infinite hidden markov models for multiple multivariate time series with missing data. *Biometrics*.

Hu, Y., Moffitt, R., and Sasaki, Y. (2019). Semiparametric estimation of the canonical permanent-transitory model of earnings dynamics. *Quantitative Economics*, 10(4):1495–1536.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1):50 – 67.

Kim, C.-J., Nelson, C. R., et al. (1999). State-space models with regime switching: classical and gibbs-sampling approaches with applications. *MIT Press Books*, 1.

Kim, J. and Wang, L. (2019). Hidden group patterns in democracy developments: Bayesian inference for grouped heterogeneity. *Journal of Applied Econometrics*, 34(6):1016–1028.

Koki, C., Leonardos, S., and Piliouras, G. (2022). Exploring the predictability of cryptocurrencies via bayesian hidden markov models. *Research in International Business and Finance*, 59:101554.

Lillard, L. and Weiss, Y. (1979). Components of variation in panel earnings data: American scientists, 1960-70. *Econometrica*, 47(2):437–54.

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1):42–55.

Linderman, S. W. (2016). *Bayesian Methods for Discovering Structure in Neural Spike Trains*. PhD thesis, Harvard University.

Linderman, S. W., Johnson, M. J., and Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3456–3464.

Linderman, S. W., Johnson, M. J., Miller, A. C., Adams, R. P., Blei, D. M., and Paninski, L. (2017). Bayesian learning and inference in recurrent switching linear dynamical systems. In Singh, A. and Zhu, X. J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 914–922. PMLR.

Liu, L. (2022). Density forecasts in panel data models: A semiparametric bayesian perspective. *Journal of Business & Economic Statistics*, 0(0):1–15.

Liu, L., Moon, H. R., and Schorfheide, F. (2020). Forecasting with dynamic panel data models. *Econometrica*, 88(1):171–201.

MaCurdy, T. E. (1982). The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of Econometrics*, 18(1):83–114.

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Meligkotsidou, L. and Dellaportas, P. (2011). Forecasting with non-homogeneous hidden markov models. *Statistics and Computing*, 21:439–449.

Nassar, J., Linderman, S. W., Bugallo, M. F., and Park, I. M. (2019). Tree-structured recurrent switching linear dynamical systems for multi-scale modeling. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286.

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). Logistic stick-breaking process. *J. Mach. Learn. Res.*, 12:203–239.

Rigon, T. and Durante, D. (2021). Tractable bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference*, 211:131–142.

Robbins, H. E. (1992). *An empirical Bayes approach to statistics*. Springer.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2):291–317.

Rodriguez, A. and Dunson, D. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6:145–178.

Scott, S. L. (2002). Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American statistical Association*, 97(457):337–351.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Su, L., Shi, Z., and Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Su, L., Wang, X., and Jin, S. (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics*, 37(2):334–349.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Uddin, M. N. and Gaskins, J. T. (2023). Shared bayesian variable shrinkage in multinomial logistic regression. *Computational Statistics Data Analysis*, 177:107568.

Wang, E., Chiang, S., Haneef, Z., Rao, V., Moss, R., and Vannucci, M. (2023). Bayesian non-homogeneous hidden markov model with variable selection for investigating drivers of seizure risk cycling. *The Annals of Applied Statistics*, 17.

Wang, W., Zhang, X., and Paap, R. (2019). To pool or not to pool: What is a good strategy for parameter estimation and forecasting in panel regressions? *Journal of Applied Econometrics*, 34(5):724–745.

Appendix

Group Membership

In this section we give a more in depth insight in how the models SSSM-10-NT, TN-SSSM-10-NT and MB-SSSM-10-NT perform for the different clusters and years. We have trained the SSSM-10-NT on 1550 MCMC with a burn-in sample of 1300 MCMC iterations. We have used 150 MCMC draws and a burn-in sample of 100 MCMC draws for training model TN-SSSM-10-NT. We trained MB-SSSM-10-NT on 1550 MCMC samples with a burn-in sample of 1300 iterations.

	μ_α			σ^2_α			μ_θ			σ^2_θ		
K	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
1	9.5489	9.5518	9.5545	0.0108	0.0112	0.0116	-0.0042	-0.0009	0.002	0.0046	0.0048	0.0050
2	9.7455	9.7576	9.7694	0.0472	0.0507	0.054	0.0026	0.0071	0.0123	0.0073	0.0077	0.0081
3	10.0333	10.04	10.0456	0.0403	0.0426	0.0456	0.0134	0.0167	0.0197	0.0051	0.0053	0.0055
4	10.1367	10.1576	10.1708	0.6123	0.6264	0.6392	0.0856	0.0893	0.0931	0.0031	0.0032	0.0033
5	10.3661	10.3937	10.4059	0.0601	0.0622	0.0649	-0.0132	-0.0103	-0.0081	0.0041	0.0042	0.0044
6	10.422	10.4329	10.4422	0.0675	0.072	0.0758	0.0074	0.01	0.0132	0.0053	0.0055	0.0057
7	10.6959	10.7014	10.7097	0.034	0.0402	0.0453	0.0048	0.0081	0.0122	0.0055	0.006	0.0063
8	10.934	10.9982	11.0386	0.0655	0.0703	0.0732	-0.0102	-0.0006	0.0067	0.004	0.0049	0.0052
9	11.0493	11.0657	11.0976	0.0507	0.0524	0.0738	-0.0095	0.0186	0.0242	0.004	0.0047	0.0051
10	11.07	11.0949	11.1288	0.0513	0.0643	0.0712	-0.0134	-0.005	0.024	0.004	0.0044	0.0048

Table 7: The 5%, 50%, and 95% percentile of the estimates of the model parameters ($\mu_\alpha^{(k)}$, $\sigma^2_{\alpha(k)}$, $\mu_\theta^{(k)}$, $\sigma^2_{\theta(k)}$) for $k = 1, \ldots, 10$ over 250 MCMC iterations for model SSSM-10-NT.

	K	1	2	3	4	5	6	7	8	9	10	σ^2_ε
$\sigma^2_{\eta(k)}$	5%	0.0101	0.0144	0.0129	0.1995	0.0189	0.0162	0.0158	0.0165	0.0139	0.0143	0.0038
	50%	0.0105	0.0149	0.0135	0.2027	0.0210	0.0165	0.0170	0.0172	0.0146	0.0183	0.0040
	95%	0.0107	0.0152	0.0139	0.2052	0.0219	0.0169	0.0175	0.0179	0.0194	0.02	0.0041
$\sigma^2_{\zeta(k)}$	5%	0.003	0.0037	0.0031	0.0021	0.0026	0.0029	0.0035	0.0033	0.0032	0.0032	
	50%	0.0031	0.0039	0.0032	0.0021	0.0027	0.0029	0.0036	0.0035	0.0034	0.0033	
	95%	0.0033	0.004	0.0033	0.0022	0.0027	0.003	0.0038	0.0037	0.0037	0.0035	

Table 8: The 5%, 50%, and 95% percentile of the state disturbances estimates (σ^2_η, σ^2_ζ) for $k = 1, \ldots, 10$ and the 5%, 50%, and 95% percentile of the observation disturbance estimates over 250 MCMC iterations for model SSSM-10-NT.

Forecasting Household Earnings Distribution

In this section we give the forecasting results over the years 2020 to 2029 of the models SSSM-10-NT and TN-SSSM-10-NT. We have trained the SSSM-10-NT on 1550 MCMC with a burn-in sample of 1300 MCMC iterations. We have used 150 MCMC draws and a burn-in sample of 100 MCMC draws for training model TN-SSSM-10-NT.

	μ_α			σ^2_α			μ_θ			σ^2_θ		
K	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
1	9.5455	9.5517	9.5545	0.0095	0.0103	0.011	0.0008	0.003	0.0062	0.0047	0.0049	0.0056
2	9.746	9.7571	9.8124	0.0357	0.0473	0.0534	0.0101	0.0138	0.0182	0.0046	0.0069	0.0075
3	10.0308	10.0885	10.1227	0.0444	0.7912	0.815	-0.0004	0.1205	0.1274	0.0042	0.0045	0.0051
4	10.1226	10.1503	10.2499	0.0382	0.0462	0.7785	-0.0149	0.0097	0.12	0.0039	0.0046	0.0053
5	10.2237	10.2503	10.296	0.0388	0.0421	0.0455	-0.0194	0.0074	0.0104	0.0041	0.0042	0.0049
6	10.5923	10.6703	10.6853	0.0268	0.0595	0.0729	-0.0277	-0.02	0.0078	0.0054	0.0062	0.0069
7	10.6883	10.6957	10.8512	0.0324	0.0349	0.0734	-0.0297	0.0058	0.0199	0.005	0.0052	0.0062
8	10.8545	10.8652	10.8826	0.0486	0.0586	0.0679	-0.0295	0.0068	0.0161	0.0039	0.0042	0.0058
9	11.0481	11.0563	11.0669	0.0401	0.0424	0.0474	0.0082	0.0109	0.0147	0.0045	0.0048	0.0054
10	11.2207	11.2377	11.2692	0.0345	0.0388	0.0436	-0.0093	-0.0036	0.0036	0.0052	0.0054	0.0062

Table 9: The 5%, 50%, and 95% percentile of the estimates of the model parameters ($\mu_\alpha^{(k)}$, $\sigma^2_{\alpha^{(k)}}$, $\mu_\theta^{(k)}$, $\sigma^2_{\theta^{(k)}}$) for $k = 1, \ldots, 10$ over 250 MCMC iterations for model TN-SSSM-10-NT.

	K	1	2	3	4	5	6	7	8	9	10	σ^2_ε
$\sigma^2_{\eta^{(k)}}$	5%	0.014	0.0202	0.0173	0.0197	0.0222	0.02	0.019	0.0217	0.0183	0.0204	0.0058
	50%	0.0159	0.0216	0.2356	0.0283	0.0255	0.0225	0.022	0.0237	0.0209	0.0237	0.0069
	95%	0.0186	0.0256	0.2437	0.2294	0.0296	0.0241	0.0292	0.0278	0.0228	0.0263	0.0078
$\sigma^2_{\zeta^{(k)}}$	5%	0.0034	0.0036	0.0026	0.0025	0.0031	0.004	0.0037	0.0036	0.0038	0.0038	
	50%	0.0037	0.0044	0.0027	0.0036	0.0033	0.0041	0.004	0.004	0.0042	0.0043	
	95%	0.0053	0.0056	0.0038	0.0044	0.0042	0.0047	0.005	0.005	0.005	0.0053	

Table 10: The 5%, 50%, and 95% percentile of the state disturbances estimates (σ^2_η, σ^2_ζ) for $k = 1, \ldots, 10$ and the 5%, 50%, and 95% percentile of the observation disturbance estimates over 250 MCMC iterations for model TN-SSSM-10-NT.

	μ_α			σ^2_α			μ_θ			σ^2_θ		
K	5%	50%	95%	5%	50%	95%	5%	50%	95%	5%	50%	95%
1	8.863	9.0492	9.1784	0.2907	2.796	3.2394	0.0795	0.4589	0.509	0.0547	0.1302	0.1549
2	9.067	9.2631	9.4272	0.3423	0.4162	1.8373	-0.1345	0.16	0.448	0.0224	0.0302	0.3803
3	9.3168	9.519	9.5229	0.0127	0.0138	1.815	-0.1615	0.0029	0.1323	0.0092	0.0104	0.3769
4	9.5184	9.5606	9.6607	0.0126	0.228	0.8661	-0.15	0.0271	0.1915	0.0089	0.0656	0.3159
5	9.5909	9.7216	10.0187	0.1087	0.3068	0.7298	-0.0869	0.0758	0.167	0.0026	0.0584	0.324
6	10.0104	10.0298	10.174	0.1082	0.1105	0.6591	-0.0523	0.0124	0.1032	0.0025	0.0026	0.2975
7	10.38	10.3894	10.4031	0.0453	0.0487	0.0522	0.0338	0.0384	0.0442	0.0159	0.0169	0.0177
8	10.4519	10.4913	10.5309	0.0421	0.0454	1.0244	-0.0634	-0.0476	-0.0375	0.0141	0.0149	0.0564
9	10.5262	10.8212	10.9218	0.061	0.5215	1.0636	-0.0673	-0.0396	-0.0147	0.0157	0.0388	0.0562
10	10.9362	10.9421	10.9543	0.0867	0.0878	0.0892	0.0148	0.0155	0.0161	0.0021	0.0022	0.0022

Table 11: The 5%, 50%, and 95% percentile of the estimates of the model parameters ($\mu_\alpha^{(k)}$, $\sigma^2_{\alpha^{(k)}}$, $\mu_\theta^{(k)}$, $\sigma^2_{\theta^{(k)}}$) for $k = 1, \ldots, 10$ over 250 MCMC iterations for model MB-SSSM-10-NT.

	K	1	2	3	4	5	6	7	8	9	10	σ^2_ε
$\sigma^2_{\eta^{(k)}}$	5%	0.435	0.1587	0.0267	0.0254	0.0198	0.02	0.0613	0.0521	0.051	0.0231	0.0059
	50%	1.5464	0.2507	0.0299	0.2128	0.2257	0.0207	0.0656	0.0554	0.1918	0.0236	0.0061
	95%	1.804	2.7933	2.5115	0.8754	0.846	0.7673	0.0733	0.1522	0.213	0.0252	0.0067
$\sigma^2_{\zeta^{(k)}}$	5%	0.0603	0.0197	0.0124	0.0119	0.0013	0.0012	0.0161	0.0118	0.0126	0.0011	
	50%	0.1034	0.0301	0.0143	0.0843	0.0838	0.0013	0.0186	0.0127	0.0323	0.0011	
	95%	0.1153	0.5256	0.4487	0.4661	0.4906	0.5013	0.0208	0.0582	0.061	0.0011	

Table 12: The 5%, 50%, and 95% percentile of the state disturbances estimates (σ^2_η, σ^2_ζ) for $k = 1, \ldots, 10$ and the 5%, 50%, and 95% percentile of the observation disturbance estimates over 250 MCMC iterations for model MB-SSSM-10-NT.

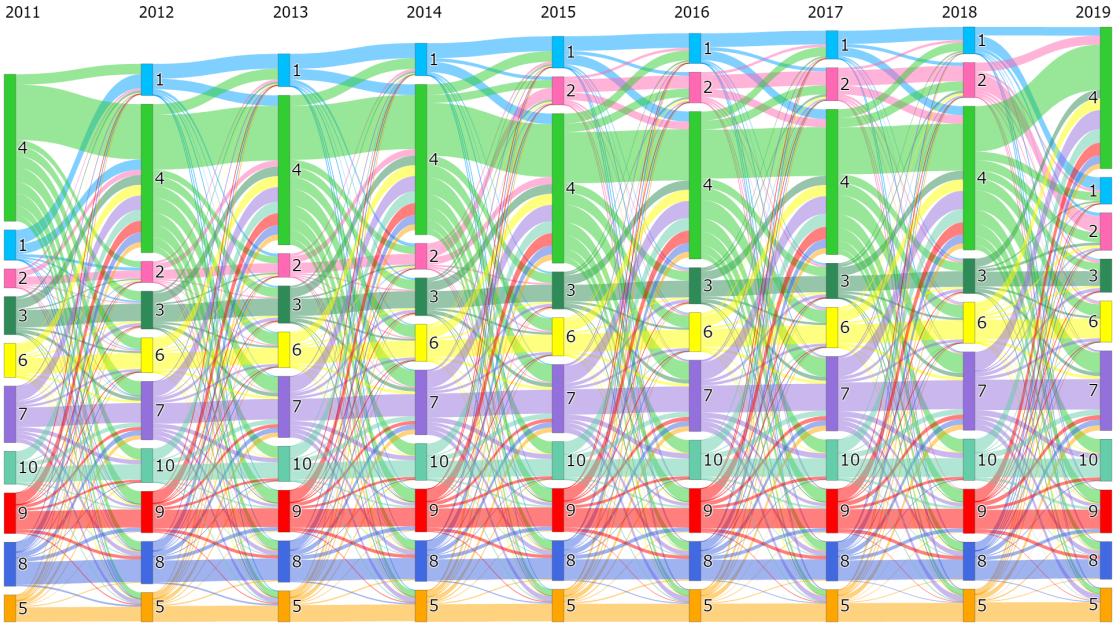


Fig. 12: A Sankey Diagram of the cluster transitions over the years 2011 to 2019 for model SSSM-10-R-NT 5% percentile.

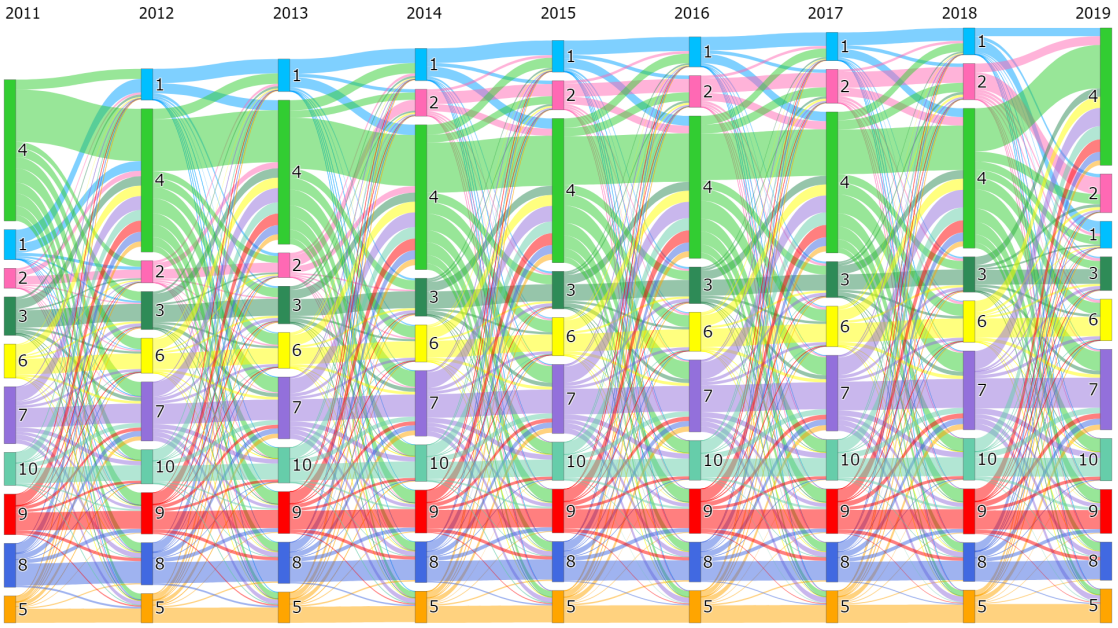


Fig. 13: A Sankey Diagram of the cluster transitions over the years 2011 to 2019 for model SSSM-10-R-NT 95% percentile.

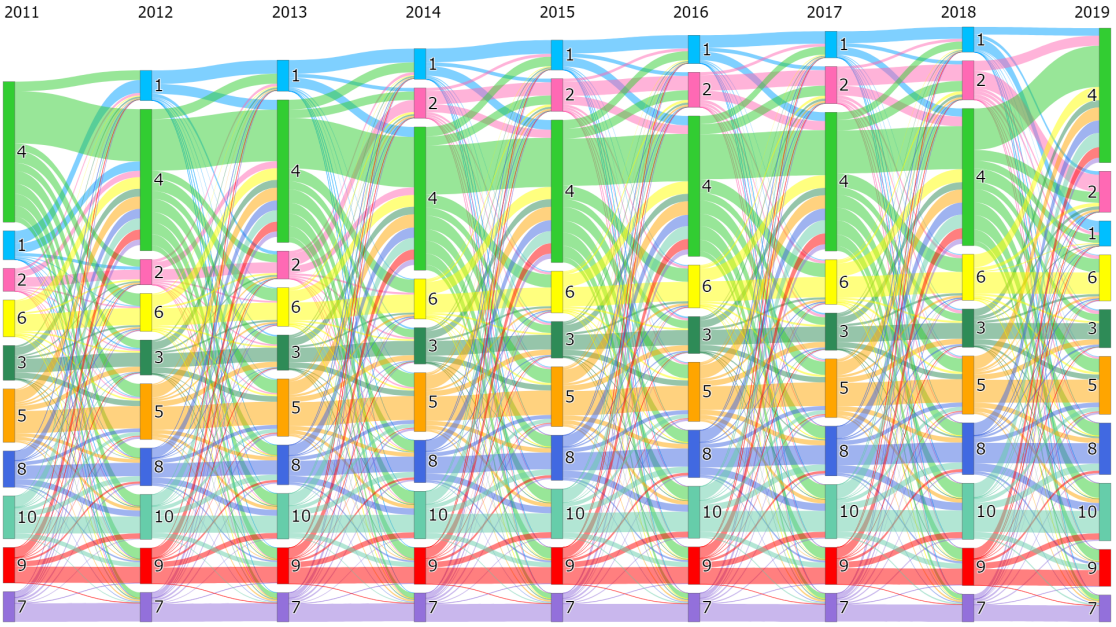


Fig. 14: A Sankey Diagram of the cluster transitions over the years 2011 to 2019 for model SSSM-10-NT.

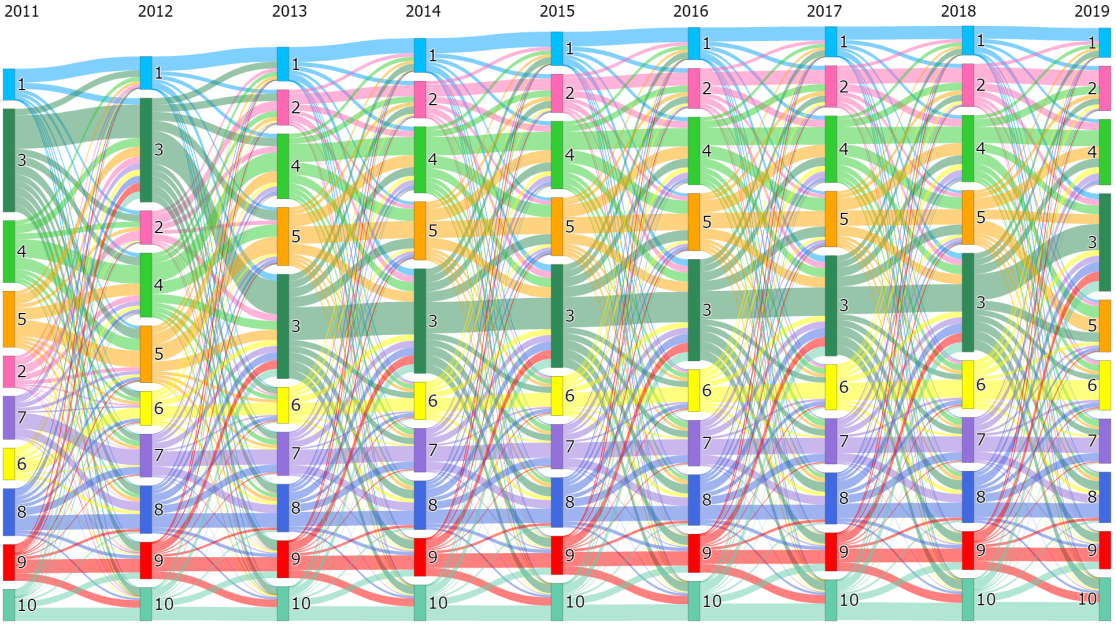


Fig. 15: A Sankey Diagram of the cluster transitions over the years 2011 to 2019 for model TN-SSSM-10-NT.

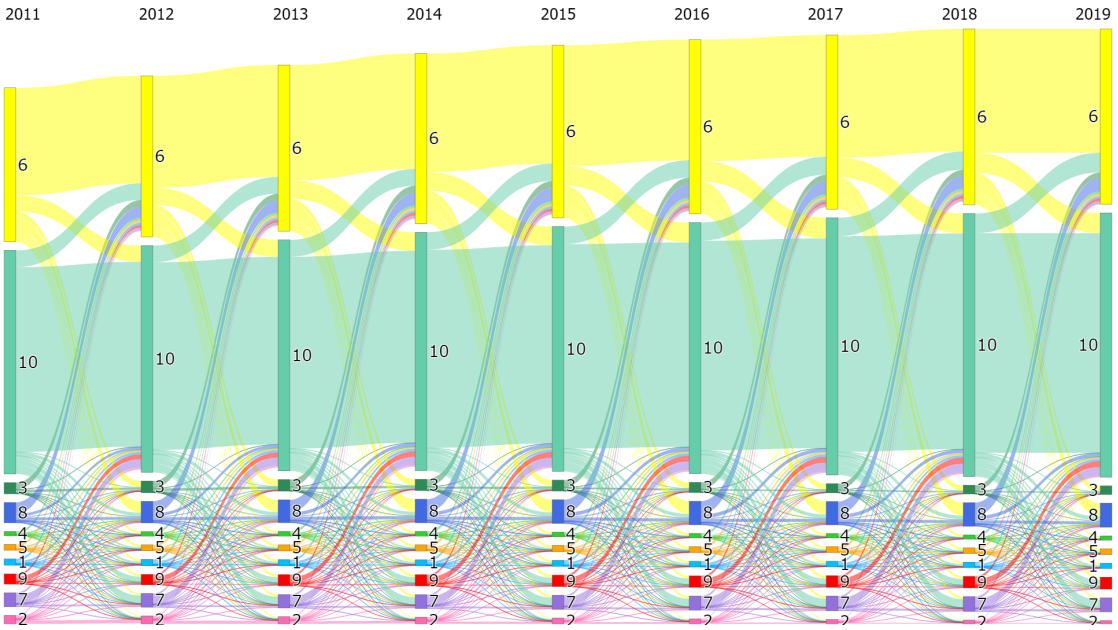


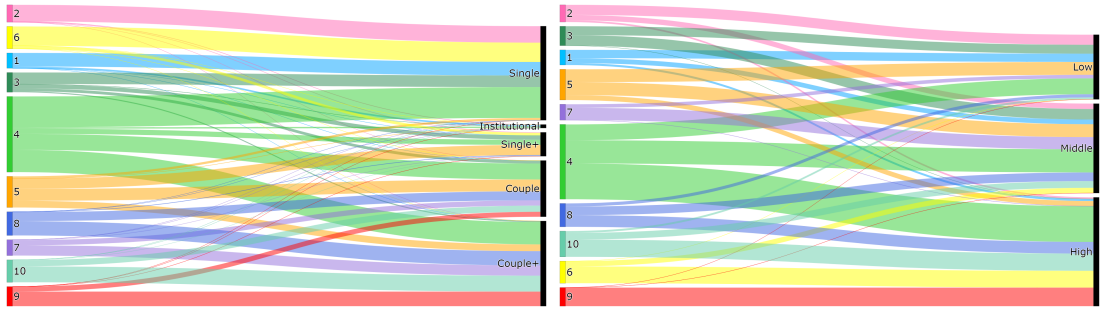
Fig. 16: A Sankey Diagram of the cluster transitions over the years 2011 to 2019 for model MB-SSSM-10-NT.

		1	2	3	4	5	6	7	8	9	10
2011	median($\hat{y}$)	14.06	17.32	22.83	26.22	30.83	32.07	45.9	48.66	58.61	71.93
	median(y)	14.29	17.43	22.92	34.17	32.09	30.22	43.62	47.47	54.95	68.08
2015	median($\hat{y}$)	13.96	16.74	23.95	37.44	29.77	33.75	42.9	50.3	61.0	73.83
	median(y)	14.61	18.18	23.8	37.38	31.95	31.84	44.76	51.28	59.17	72.62
2019	median($\hat{y}$)	13.85	16.02	25.06	53.23	28.57	35.33	39.85	51.62	63.34	75.61
	median(y)	15.45	19.39	25.23	44.17	33.9	33.78	48.97	57.86	67.52	79.66

Table 13: Median calculated over the predicted households earnings $\hat{y}$ and median computed over the true household earnings for cluster $k = 1, \ldots, 10$, the years 2011, 2015, and 2019, and model SSSM-10-R-NT using the 5% percentile over the 250 MCMC draws.

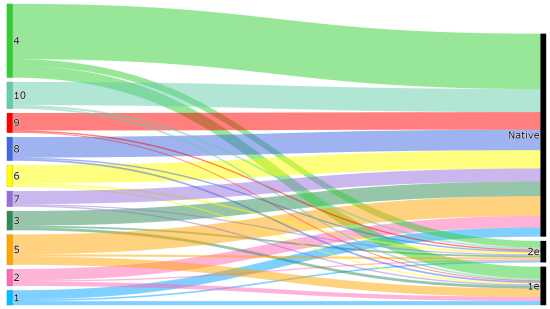
		1	2	3	4	5	6	7	8	9	10
2011	median($\hat{y}$)	14.15	17.75	23.06	26.89	31.34	32.62	47.0	49.24	59.47	72.69
	median(y)	14.35	17.91	23.13	34.86	32.54	30.72	44.51	47.82	55.6	68.81
2015	median($\hat{y}$)	14.28	17.59	24.6	38.51	30.75	34.66	44.29	51.36	62.66	76.36
	median(y)	14.67	18.5	24.09	38.23	32.37	32.33	45.61	51.69	59.95	73.44
2019	median($\hat{y}$)	14.47	17.64	26.3	55.4	30.27	37.01	41.95	53.8	66.27	80.77
	median(y)	15.53	19.74	25.63	45.32	34.41	34.35	49.79	58.37	68.58	80.69

Table 14: Median calculated over the predicted households earnings $\hat{y}$ and median computed over the true household earnings for cluster $k = 1, \ldots, 10$, the years 2011, 2015, and 2019, and model SSSM-10-R-NT using the 95% percentile over the 250 MCMC draws.

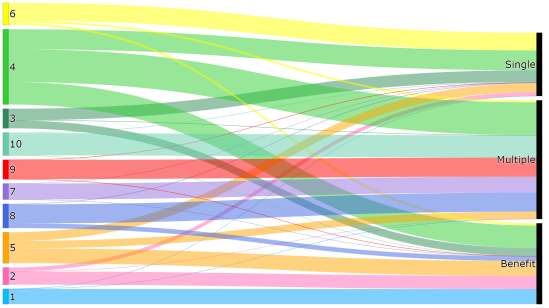


(a) Household Composition

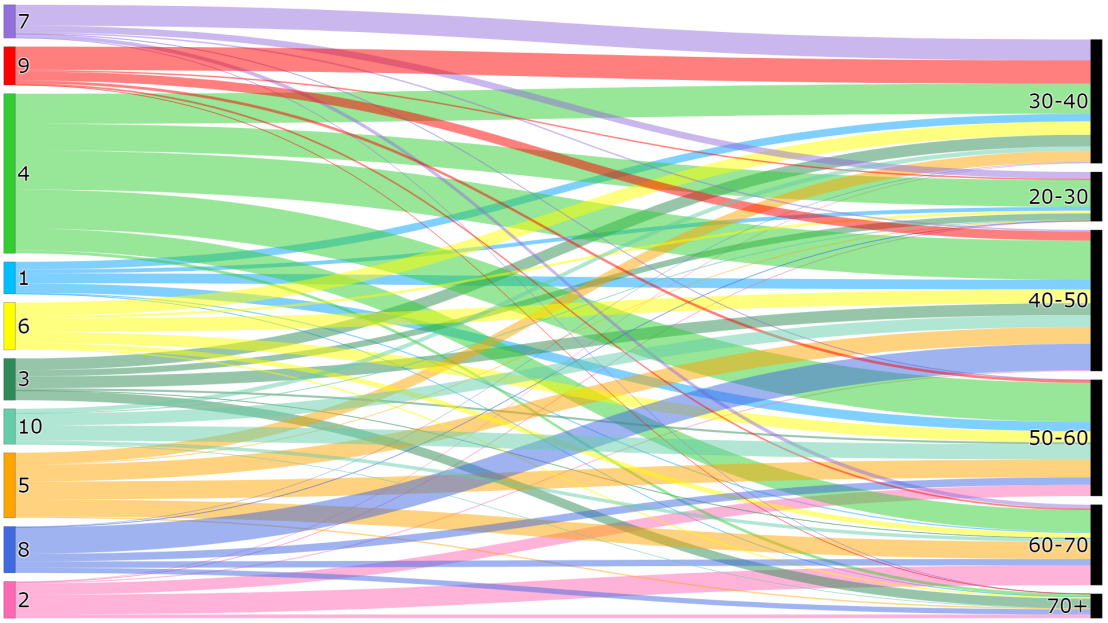
(b) Education Level



(c) Generation

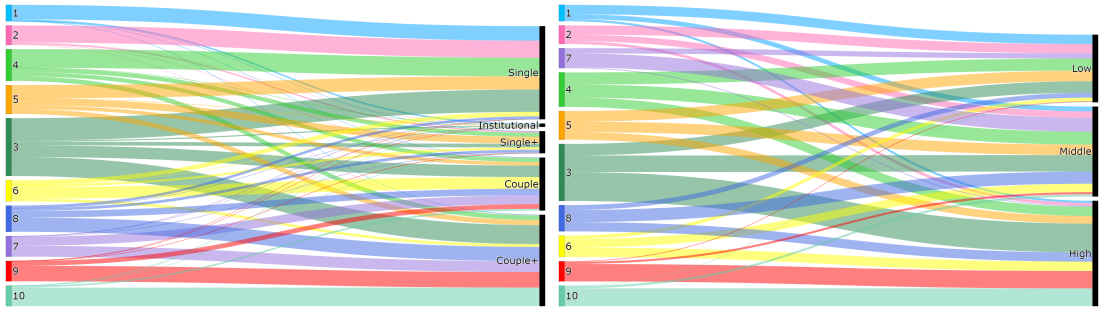


(d) Main Income Source



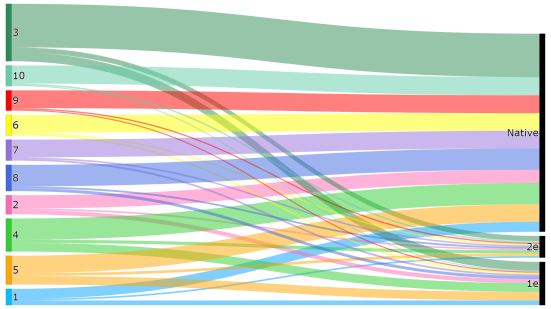
(e) Age

Fig. 17: 5 diagrams representing the number of households with a certain household characteristic in a particular cluster $k = 1, \ldots, 10$ for model SSSM-10-NT, respectively, household composition, education level, generation, main income source, and age.

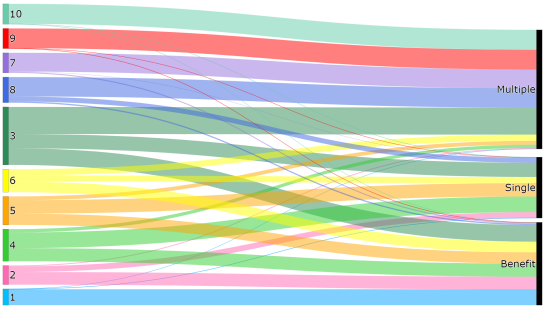


(a) Household Composition

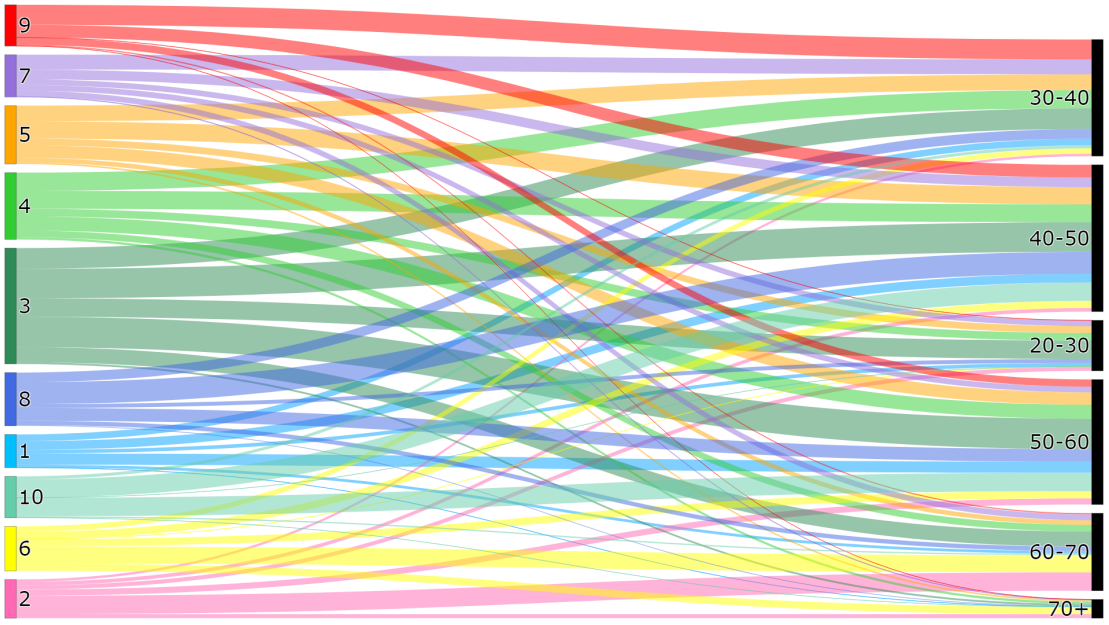
(b) Education Level



(c) Generation

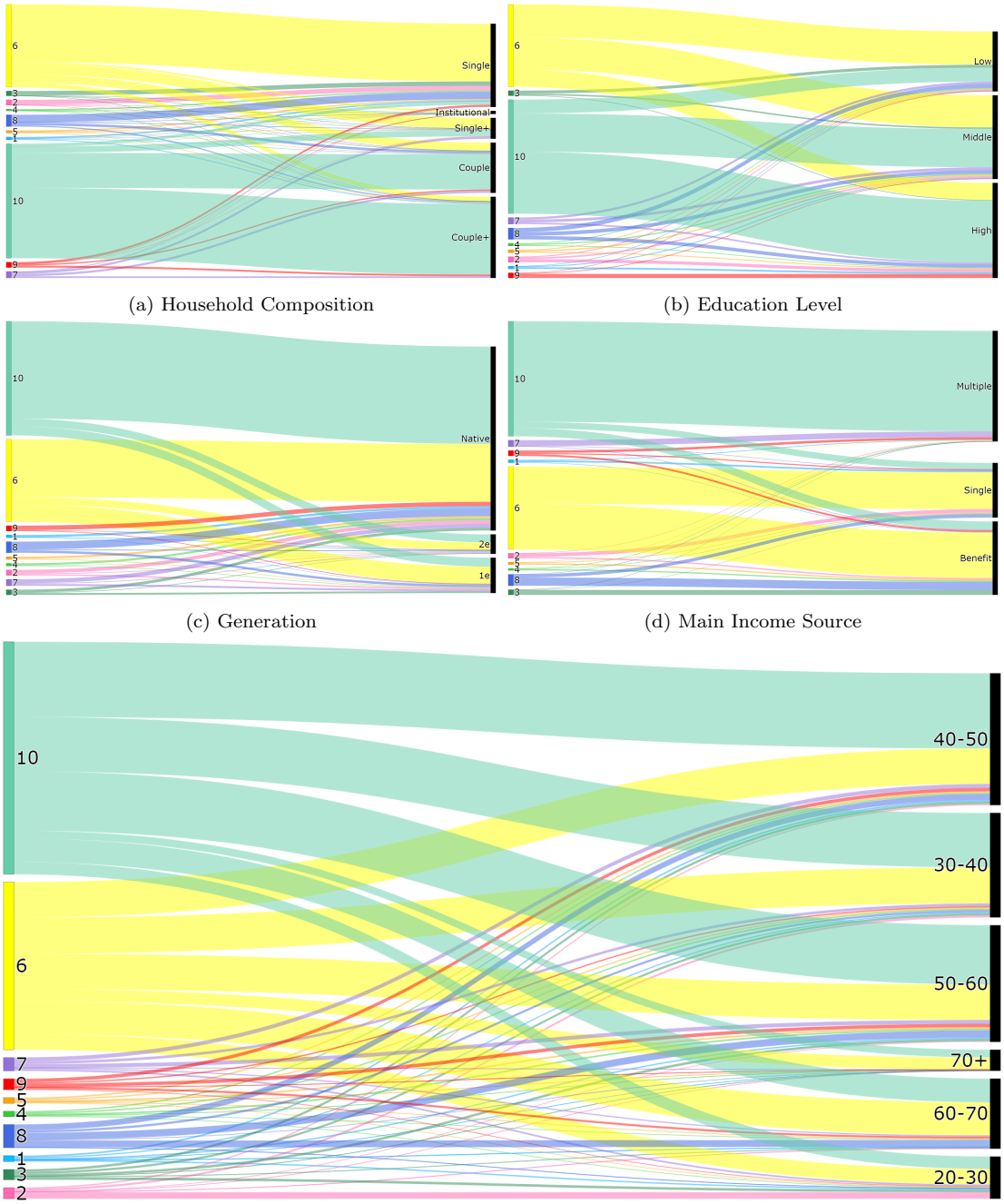


(d) Main Income Source



(e) Age

Fig. 18: 5 diagrams representing the number of households with a certain household characteristic in a particular cluster $k = 1, \ldots, 10$ for model TN-SSSM-10-NT, respectively, household composition, education level, generation, main income source, and age.



(a) Household Composition

(b) Education Level

(c) Generation

(d) Main Income Source

(e) Age

Fig. 19: 5 diagrams representing the number of households with a certain household characteristic in a particular cluster $k = 1, \ldots, 10$ for model MB-SSSM-10-NT, respectively, household composition, education level, generation, main income source, and age.

		1	2	3	4	5	6	7	8	9	10
2011	median($\hat{y}$)	14.07	17.29	22.92	25.78	32.66	33.97	44.41	59.72	63.95	65.82
	median(y)	14.28	17.71	23.72	33.51	32.01	31.32	44.23	56.57	62.23	61.02
2015	median($\hat{y}$)	14.02	17.8	24.53	36.79	31.31	35.36	45.85	59.35	69.13	65.25
	median(y)	14.61	18.06	25.31	36.34	32.82	32.95	48.13	59.25	66.9	64.51
2019	median($\hat{y}$)	13.96	18.33	26.2	52.65	30.0	36.83	47.38	59.25	74.87	63.96
	median(y)	15.52	19.38	27.13	42.39	35.74	35.2	54.8	63.15	73.84	72.46

Table 15: Median calculated over the predicted households earnings $\hat{y}$ and median computed over the true household earnings for cluster $k = 1, \ldots, 10$, the years 2011, 2015, and 2019, and model SSSM-10-NT.

		1	2	3	4	5	6	7	8	9	10
2011	median($\hat{y}$)	14.07	17.29	24.08	25.63	28.28	43.11	44.21	52.32	63.35	76.02
	median(y)	14.4	19.47	33.99	27.33	27.04	40.22	45.18	47.49	62.87	68.32
2015	median($\hat{y}$)	14.24	18.32	37.78	29.29	29.17	39.26	45.09	53.9	66.49	75.41
	median(y)	14.71	20.13	37.9	28.22	27.89	40.46	48.55	49.69	66.96	72.95
2019	median($\hat{y}$)	14.41	19.37	62.55	30.38	30.07	35.63	46.01	55.32	69.57	74.35
	median(y)	15.64	21.25	45.56	29.9	29.97	42.54	54.57	54.39	73.29	82.40

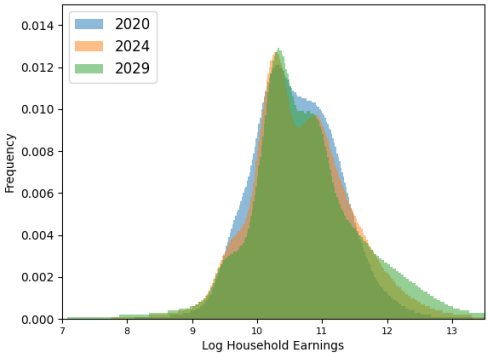
Table 16: Median calculated over the predicted households earnings $\hat{y}$ and median computed over the true household earnings for cluster $k = 1, \ldots, 10$, the years 2011, 2015, and 2019, and model TN-SSSM-10-NT.

		1	2	3	4	5	6	7	8	9	10
2011	median($\hat{y}$)	8.51	10.61	13.62	14.18	16.65	22.68	32.56	36.02	50.06	56.5
	median(y)	33.24	22.1	14.23	27.53	27.45	22.92	41.8	24.47	42.38	54.07
2015	median($\hat{y}$)	52.37	19.58	13.78	16.38	23.22	23.67	38.07	29.75	42.92	60.13
	median(y)	36.3	24.91	14.79	27.59	27.97	23.45	43.94	24.55	43.55	57.49
2019	median($\hat{y}$)	326.05	37.06	13.95	18.36	27.91	24.93	44.22	24.41	36.28	64.01
	median(y)	40.34	28.94	15.94	30.49	30.82	24.64	48.07	25.65	48.3	63.33

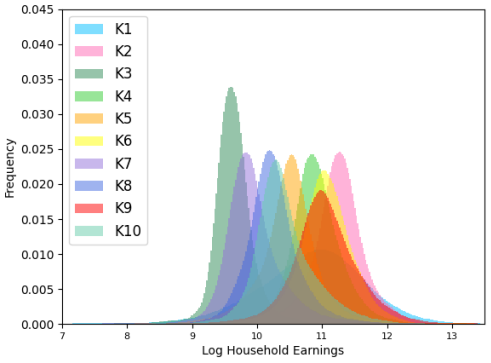
Table 17: Median calculated over the predicted households earnings $\hat{y}$ and median computed over the true household earnings for cluster $k = 1, \ldots, 10$, the years 2011, 2015, and 2019, and model MB-SSSM-10-NT.

Age	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70	70+
100%	0.90	2.02	7.18	10.9	11.7	12.9	14.1	12.8	11.0	7.99	8.55
Household Composition	Single+	Single+	Couple	Couple+	Institutional						
100%	30.7	9.35	24.4	33.1	2.4						
Education Level	Low	Middle	High								
100%	23.3	35.0	41.7								
Generation	Native	1e	2e								
100%	77.3	14.0	8.70								
Main Income Source	Single	Multiple	Benefit								
100%	21.6	47.0	31.3								

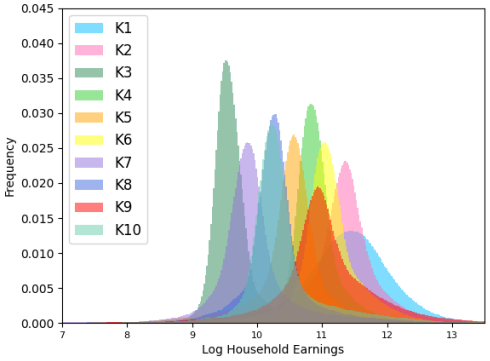
Table 18: Percentages of households (%) having a certain age, education level, generation, household composition, and main income source in our forecasting dataset of 1.6 million households for the year 2020.



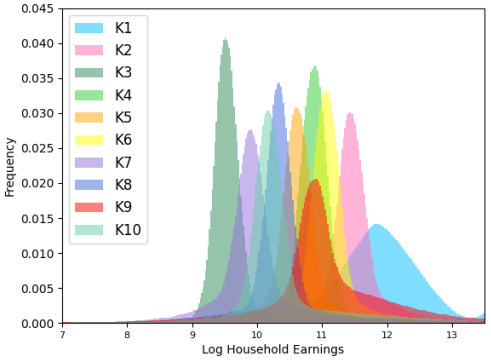
(a) Total forecasted distributions by SSSM-10-NT for 2020, 2024, and 2029



(b) Forcasted distributions by SSSM-10-NT for the clusters $K = 1, \ldots, 10$ and 2020

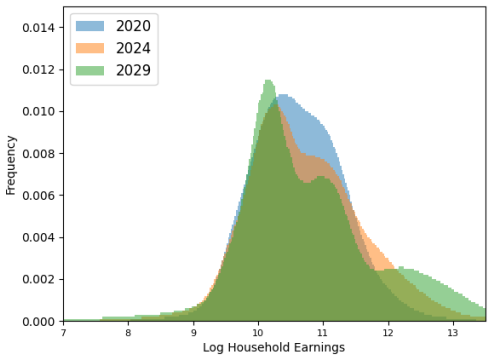


(c) Forcasted distributions by SSSM-10-NT for the clusters $K = 1, \ldots, 10$ and 2024

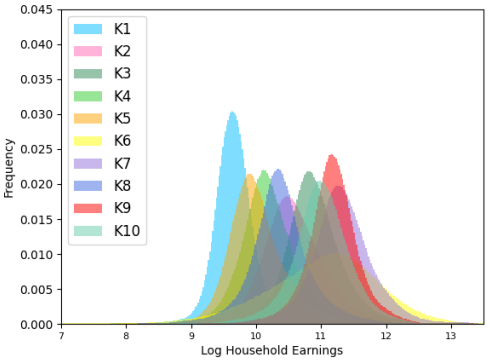


(d) Forcasted distributions by SSSM-10-NT for the clusters $K = 1, \ldots, 10$ and 2029

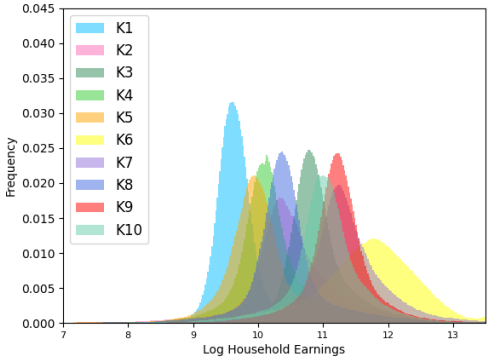
Fig. 20: Forecasted Household Earnings Distributions for model SSSM-10-NT with $K = 1, \ldots, 10$ clusters.



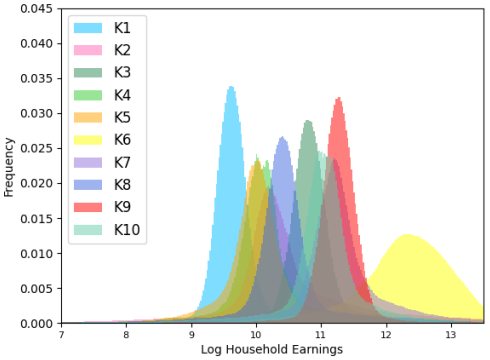
(a) Total forecasted distributions by TN-SSSM-10-NT for 2020, 2024, and 2029



(b) Forcasted distributions by TN-SSSM-10-NT for the clusters $K = 1, \ldots, 10$ and 2020

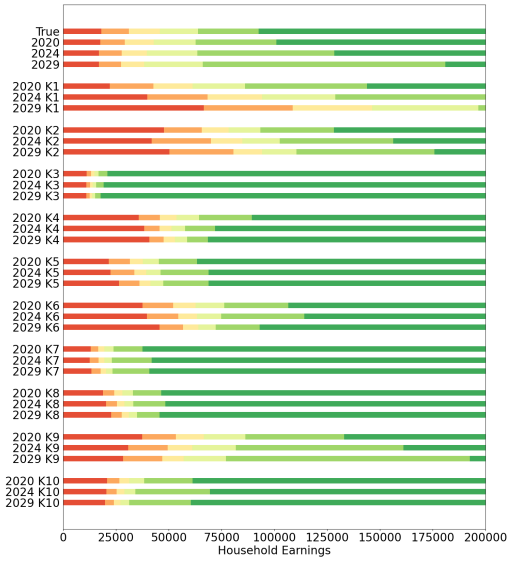


(c) Forcasted distributions by TN-SSSM-10-NT for the clusters $K = 1, \ldots, 10$ and 2024

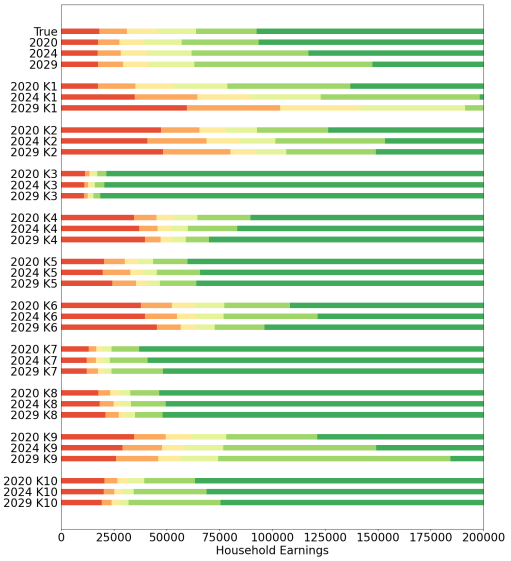


(d) Forcasted distributions by TN-SSSM-10-NT for the clusters $K = 1, \ldots, 10$ and 2029

Fig. 21: Forecasted Household Earnings Distributions for model SSSM-10-NT with $K = 1, \ldots, 10$ clusters.

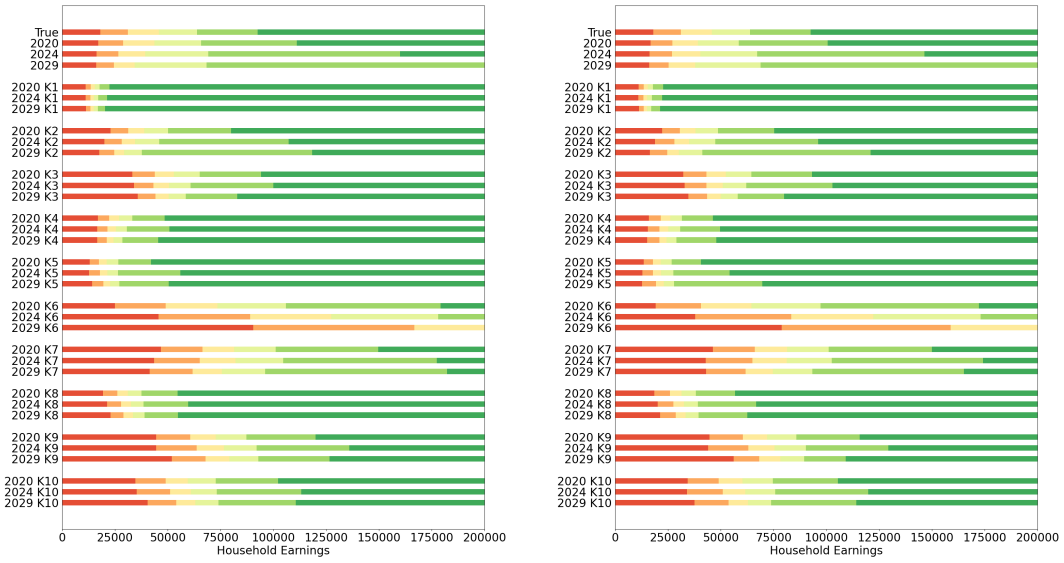


(a) Uncorrected for population forecasts



(b) Corrected for population forecasts

Fig. 22: 10%, 30%, 50%, 70%, and 90% percentiles for 2020, 2024, 2029 and clusters $K = 1, \ldots, 10$ forecasted by model SSSM-10-NT. The color transitions denote the income percentiles.



(a) Uncorrected for population forecasts

(b) Corrected for population forecasts

Fig. 23: 10%, 30%, 50%, 70%, and 90% percentiles for 2020, 2024, 2029 and clusters $K = 1, \ldots, 10$ forecasted by model TN-SSSM-10-NT. The color transitions denote the income percentiles.