Biewen, Martin; Glaisner, Stefan

Working Paper

Using Distributional Random Forests for the Analysis of the Income Distribution

IZA Discussion Papers, No. 17774

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Biewen, Martin; Glaisner, Stefan (2025) : Using Distributional Random Forests for the Analysis of the Income Distribution, IZA Discussion Papers, No. 17774, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/316729

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17774

Using Distributional Random Forests for the Analysis of the Income Distribution

Martin Biewen
Stefan Glaisner

MARCH 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17774

Using Distributional Random Forests for the Analysis of the Income Distribution

Martin Biewen
University of Tübingen and IZA

Stefan Glaisner
University of Tübingen

MARCH 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Using Distributional Random Forests for the Analysis of the Income Distribution

This paper utilises distributional random forests as a flexible machine learning method for analysing income distributions. Distributional random forests avoid parametric assumptions, capture complex interactions among covariates, and, once trained, provide full estimates of conditional income distributions. From these, any type of distributional index such as measures of location, inequality and poverty risk can be readily computed. They can also efficiently process grouped income data and be used as inputs for distributional decomposition methods. We consider four types of applications: (i) estimating income distributions for granular population subgroups, (ii) analysing distributional change over time, (iii) spatial smoothing of income distributions, and (iv) purging spatial income distributions of differences in spatial characteristics. Our application based on the German Microcensus provides new results on the socio-economic and spatial structure of the German income distribution.

JEL Classification: D31, C55, I3
Keywords: inequality, poverty, small area estimation, grouped income data

Corresponding author:
Martin Biewen
School of Business and Economics
University of Tübingen
Mohlstr. 36,
72074 Tübingen
Germany
E-mail: martin.biewen@uni-tuebingen.de

1. Introduction

Measuring distributions of income and wealth is a central concern of both statistics and the social sciences. A large number of statistical techniques have been developed to measure such distributions and to investigate their structure (e.g., Jenkins and Van Kerm, 2009; Fortin et al., 2011; Chernozhukov et al., 2013; Cowell and Flachaire, 2015; Chotikapanich et al., 2018; Molina et al., 2022). In many cases, these techniques involve strong parametric assumptions on distributional shapes or the structure of regression models representing the dependence of distributions on covariates. A key advantage of modern machine learning methods such as random forests (Breiman, 2001; Athey et al., 2019) is their ability to avoid such assumptions. As a recursive partitioning algorithm, the random forest is based on sequentially splitting the covariate space into cells of observations that are similar with respect to a target criterion, and by overlaying independent repetitions of this procedure. It has a non-parametric structure, allows for complex interactions and potentially non-smooth relationships, and implicitly solves model-selection problems. Random forests have demonstrated remarkable success across a wide range of applications (Biau and Scornet, 2016).

Breiman (2001)'s original random forest was designed for non-parametric mean estimation. Subsequent extensions included survival analysis (Hothorn et al., 2004), conditional quantile estimation (Meinshausen, 2006) and estimators defined by local moment conditions (Athey et al., 2019). More recently, Cevid et al. (2022) and Näf et al. (2023) proposed a highly general variant of the random forest aimed at estimating full distributions conditional on covariates (distributional random forest, DRF).

As pointed out by Cevid et al. (2022), building random forests for full distributions – rather than for individual target objects such as means, quantiles or other distributional indices – has a number of advantages. These advantages particularly apply to analyses of the income distribution. First, forest building has to be carried out only once to obtain estimates for arbitrarily many targets. For example, if one is interested in distributional indices such as median income, the at-risk-of-poverty rate, quantile ratios, the Gini index etc. for small population subgroups, one has to fit the random forest only once and then obtain estimates of these targets from the conditional distribution. Second, since the estimates for different targets are obtained from the same forest, they have the advantage of being mutually compatible. This is not necessarily the case if a new forest is fit for each target. For example, it is well-known that conditional quantiles may cross if they are estimated separately. Similarly, fitting separate forests may produce values of the at-risk-of-poverty rate and the Gini coefficient for individual subgroups that are difficult to reconcile. Third, fitting separate forests for different target objects requires suitable target-specific splitting criteria. For many targets these are unknown or could be hard to derive. By contrast, the DRF directly uses a powerful distributional criterion for splitting, the maximum mean discrepancy statistic (MMD) (Gretton et al., 2007).

As a statistical method, the distributional random forest follows the same estimation goal as a number of other estimators of conditional distributions. These typically have a parametric or semi-parametric structure, see, e.g., Donald et al. (2000), Biewen and Jenkins (2005), Rigby and Stasinopoulos (2005), Hothorn et al. (2013). Conditional quantile models (Koenker, 2005) and binary models for distributional

thresholds (Chernozhukov et al., 2013) can also be used to construct conditional distribution functions, but they require fitting a large number of quantiles or distributional thresholds. However, in all of these models, it is not easy to deal with issues such as non-smooth dependencies, complex interaction effects and automatic variable selection, which are automatically handled by the random forest. Before the development of the fully non-parametric distributional random forest, Schlosser et al. (2019) and Hothorn and Zeileis (2021) proposed parametric variants based on fitting predefined distributional forms. This can be an attractive option if the number of training observations is limited. By contrast, this paper uses Cevid et al. (2022)'s non-parametric version of the distributional random forest as a fully flexible device to estimate the relationship between outcome distributions and covariates based on a large data set.

The purpose of this paper is to apply distributional random forests to various estimation problems in the analysis of the income distribution. We consider the following applications: (i) estimating income distributions for granular population subgroups, (ii) analysing distributional change over time, (iii) spatial smoothing of income distributions, and (iv) purging spatial income distributions of differences in spatial characteristics. Application (i) is commonly used by governments and statistical agencies to monitor the well-being of population subgroups and to inform policy measures (e.g., poverty alleviation). Task (ii) decomposes changes in the aggregate distribution over time, separating changes in the distribution that stem from changes in the composition of the population from those caused by income changes in population subgroups. Application (iii) is also a common task of governments and statistical agencies aimed at constructing maps of statistical information on quantities such as median income, at-risk-of-poverty indices or income inequality across geographic areas with potentially sparse observations. This question has been addressed by a large literature on small area estimation, see Tzavidis et al. (2018) and Molina et al. (2022) for overviews. Given the inherent smoothing properties of random forests (Lin and Jeon, 2006), this method is well-suited for small area estimation. Indeed, Krenmair and Schmid (2022) have recently incorporated a random forest component into a small-area mixed effects model for estimating area-level means. In this paper, we use the DRF to estimate area-level distributions with the goal of constructing area-level statistical indices (means, at-risk-of-poverty rates, inequality indices etc.). In a final application (iv), we consider the problem of purging spatial income distributions of differences in spatial characteristics. This isolates the 'pure' spatial structure of income levels and inequality, independent of variations in age, employment, education, etc., across spatial units. To the best of our knowledge, this application is novel in the literature.

Our empirical analysis is based on the German Microcensus, an annual survey conducted by the Federal Statistical Office of Germany. (Federal Statistical Office, 2024). The Microcensus is the largest sample survey in Germany and Europe. Despite its large sample size and exceptional representativeness, it has rarely been used for income distribution analysis due to its grouped income data. While grouped income information always represents a limitation of information content, we demonstrate in this paper how the distributional random forest can effectively deal with this issue.

In addition to demonstrating the usefulness of the distributional random forest approach for analysing the income distribution, this paper contributes a number of substantive results on the German income distribution based on the Microcensus for the years 2005 and 2019. Specifically, we provide new evidence on the incomes and poverty risk of granular population subgroups and analyse distributional change over time. We show that inequality and poverty risk increased between 2005 and 2019, but that this was the result of changes in the composition of the population rather than of income changes. Finally, we provide distributional maps of household income and inequality for Germany at a much higher geographical resolution than previous analyses (Immel and Peichl, 2020; Walter et al., 2022).

The remainder of this paper is structured as follows. Section 2 outlines the method of distributional random forests due to Cevid et al. (2022). In section 3, we provide basic information about the data used by us. Section 4 presents our random forests estimates and the analyses derived from them. Section 5 concludes.

2. Distributional random forest

We outline the main properties of the distributional random forest (DRF) as introduced by Cevid et al. (2022) and Näf et al. (2023). Let $\mathbf{Y} = (Y_1, \ldots, Y_d) \in \mathbb{R}^d$ be a potentially multivariate outcome vector and $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ a vector of covariates. The goal of the DRF is to estimate the conditional distribution $\mathbb{P}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ based on a random sample $(\mathbf{y}_i, \mathbf{x}_i), i = 1, \ldots n$.

The DRF produces an estimate $\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ of the conditional distribution by repeating a recursive partitioning algorithm (= tree building) $k = 1, \ldots N$ times on random variations of the data and by averaging the results (= random forest). For each tree k, the sample is successively partitioned into groups of observations (= leaves). The partitioning proceeds greedily by splitting a parent node P into two children nodes $C_L = \{X_j \leq l\}$ and $C_R = \{X_j > l\}$ based on candidate splitting variables X_j that are chosen randomly (see below). The split is chosen such that the resulting nodes C_L and C_R are as different as possible with respect to an objective function.

In Breiman (2001)'s original random forest for mean outcomes, splits were performed so that (in the case of an univariate outcome), the resulting mean outcomes in C_L and C_R differed the most, i.e.,

$$\max \frac{n_L n_R}{n_P^2} \left(\frac{1}{n_L} \sum_{i \in C_L} y_i - \frac{1}{n_R} \sum_{i \in C_R} y_i \right)^2, \tag{1}$$

where n_L, n_R and n_P are the number of observations in the children and parent nodes, respectively.

In the DRF, splits are performed to maximize *distributional differences* between the resulting children nodes C_L and C_R. Distributional differences are measured by the Maximum Mean Discrepancy (MMD) statistic (Gretton et al., 2007). The MMD statistic is based on the theory of distributional embeddings in Reproducing Kernel Hilbert Spaces (RKHS) (Muandet et al., 2017).

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ denote a RKHS of real valued functions on $\mathbb{R}^d$ induced by a positive-definite kernel $k(\cdot,\cdot)$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, norm $\|\cdot\|_{\mathcal{H}}$, and implicit feature map $\varphi : \mathbb{R}^d \to \mathcal{H}$ satisfying $k(\mathbf{y}, \mathbf{y}') = \langle\varphi(\mathbf{y}), \varphi(\mathbf{y}')\rangle_{\mathcal{H}}$. The feature map $\varphi(\mathbf{y})$ can be interpreted as a (possibly infinite-dimensional) collection of aspects of $\mathbf{y}$. The term $k(\mathbf{y}, \mathbf{y}')$ is then a measure of similarity between points $\mathbf{y}$ and $\mathbf{y}'$ in terms of all their aspects described by the feature map. This similarity measure is linear in the feature space, but may be very nonlinear in the original space $\mathbb{R}^d$, depending on the richness of the feature map ('kernel trick').

Let $\mathcal{P}$ be a distribution and define

$$\mu(\mathcal{P}) = \mathbb{E}_{\mathbf{Y}\sim\mathcal{P}}\left[\varphi(\mathbf{Y})\right] \tag{2}$$

as its mean embedding into the Hilbert space $\mathcal{H}$ (i.e., every distribution $\mathcal{P}$ is represented as an element of $\mathcal{H}$). It turns out that, for certain choices of the kernel (i.e., characteristic kernels), this mapping is one-to-one, so that each distribution is uniquely represented by one element in the RKHS. Differences between two distributions $\mathcal{P}$ and $\mathcal{Q}$ can thus be measured by the distance function in the corresponding Hilbert space, i.e., $d(\mathcal{P}, \mathcal{Q}) = \|\mu(\mathcal{P}) - \mu(\mathcal{Q})\|^2_{\mathcal{H}}$ (i.e., the distance between their mean embeddings in the Hilbert space).

The distributional random forest uses this distance measure between the distributions of outcomes in two children nodes C_L and C_R to find splits that make distributions in C_L and C_R as different as possible. In this case, the MMD statistic is defined as

$$\begin{aligned}\mathcal{D}_{\text{MMD}}(C_L, C_R) &= \|\mu(\mathcal{P}_{C_L}) - \mu(\mathcal{P}_{C_R})\|^2_{\mathcal{H}} \\ &= \left\| \frac{1}{n_L} \sum_{i\in C_L} \varphi(\mathbf{y}_i) - \frac{1}{n_R} \sum_{i\in C_R} \varphi(\mathbf{y}_i) \right\|^2_{\mathcal{H}}.\end{aligned} \tag{3}$$

Note the similarity to Breiman (2001)'s original splitting criterion (1), which results when the feature map only consists of the value $\mathbf{y}$ itself (i.e., $\varphi(\mathbf{y}) = \mathbf{y}$). In this case, the statistic only measures average differences in the *levels* of $\mathbf{y}$. By contrast, if the feature map is richer, it measures average differences *in all features* described by the feature map $\varphi(\cdot)$ (see (3)). For example, if $\varphi(\mathbf{y})$ includes higher-order terms of $\mathbf{y}$, it will not only measure differences in means between C_L and C_R but in higher-order moments. It can be shown that the implicit feature maps of characteristic kernels is infinite-dimensional and powerful enough to detect *any* differences between distributions (Gretton et al., 2007).[1] The MMD statistic can be equivalently written as

$$\begin{aligned}\mathcal{D}_{\text{MMD}}(C_L, C_R) = &\frac{1}{n_L^2} \sum_{i,j\in C_L} k(\mathbf{y}_i, \mathbf{y}_j) \; + \frac{1}{n_R^2} \sum_{i,j\in C_R} k(\mathbf{y}_i, \mathbf{y}_j) \\ &- \frac{2}{n_L n_R} \sum_{i\in C_L} \sum_{j\in C_R} k(\mathbf{y}_i, \mathbf{y}_j).\end{aligned} \tag{4}$$

[1] In our empirical application, we use the Gaussian kernel as in Cevid et al. (2022).

This formulation provides an intuitive interpretation: the statistic measures how similar – as described by the kernel – observations are *within* each sample, as compared to *between* the two samples.

As in Breiman (2001)'s original splitting criterion (1), the distributional random forest uses a version of $\mathcal{D}_{\text{MMD}}$ that is rescaled by the factor $n_L n_R / n_P^2$. In order for the forest to be consistent for the true conditional distribution $\mathbb{P}(\mathbf{Y}|\mathbf{X} = \mathbf{x})$, forest construction has to comply with a number of rules (Athey et al., 2019; Cevid et al., 2022):

1. *Honesty:* Splits are determined on one half of the data, distributional predictions are computed on the other half of the data.
2. *Random-split:* The probability that the split occurs along feature X_j is bounded from below by π/p for some $\pi > 0$ (p is the number of covariates).
3. *Symmetry:* The tree output does not depend on the ordering of the training samples.
4. *Regularity:* Each child contains at least a fraction $\alpha \leq 0.2$ of the parent node. Trees are grown until each leaf contains between κ and $2\kappa - 1$ observations.
5. *Subsampling:* Trees are grown on subsamples of size $s_n = n^\beta$ out of the original n sample observations, where β has to be chosen within particular bounds depending on p, π and α (Cevid et al., 2022).

The distributional random forest is based on N trees $\mathcal{T}_1, \ldots, \mathcal{T}_N$ that are grown according to the rules above. Define $\mathcal{L}_k(\mathbf{x})$ as the set of training data observations that end up in the same leaf as $\mathbf{x}$ in tree $k = 1, \ldots, N$. The main output of the distributional random forest is a set of observation and test-point-specific weights

$$\hat{w}_i(\mathbf{x}) = \frac{1}{N} \sum_{k=1}^{N} \frac{1(\mathbf{x}_i \in \mathcal{L}_k(\mathbf{x}))}{|\mathcal{L}_k(\mathbf{x})|}, \tag{5}$$

measuring the proportion with which training observation $i = 1, \ldots, n$ ended up in the same leaf as a test point with $\mathbf{X} = \mathbf{x}$. The weights quantify the importance of each training data point $(\mathbf{y}_i, \mathbf{x}_i), i = 1, \ldots, n$ for predicting the conditional distribution of $\mathbf{Y}$ at test-point $\mathbf{X} = \mathbf{x}$. Formally, the resulting estimate is given by

$$\hat{\mathbb{P}}(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} \hat{w}_i(\mathbf{x}) \cdot \delta_{\mathbf{y_i}}, \tag{6}$$

where $\delta_{\mathbf{y}_i}$ denotes the point mass at $\mathbf{y}_i$. The weights (5) characterize the distributional random forest as a locally-adaptive nearest-neighbour method which smoothes observations across the covariate space (Lin and Jeon, 2006).

Cevid et al. (2022) have shown that (6) is consistent in the sense that its estimate of the conditional distribution function converges in probability to the true conditional distribution function. This implies that smooth functionals of this distribution also converge to their population counterparts. In practice, this means that the random forest weights $\hat{w}_i(\mathbf{x})$ can be used to compute any statistic of interest based on the plug-in principle.

Näf et al. (2023) have shown that, under suitable conditions, the mean embeddings of the distributional random forest estimates are asymptotically normal, implying that

sufficiently smooth functionals based the random forest weights are also asymptotically normal. Moreover, these sampling distributions can be practically simulated by a bootstrap half-sampling procedure. To this end, $b = 1, \ldots, B$ half-samples $\mathcal{S}_b$ are drawn from the original training observations. In each of these half-samples, L trees are grown to build 'mini forests' $\mathcal{T}_b, b = 1, \ldots, B$. The weights $\hat{w}_i^b(\mathbf{x})$ of these mini forests then serve as bootstrap versions of the original weights $\hat{w}_i(\mathbf{x})$ to compute bootstrap versions of the statistics of interest. The procedure can be efficiently used to construct the overall forest consisting of $N = L \cdot B$ trees by combining the L mini forests to form the total forest ('bootstrap of little bags', Athey et al., 2019). In our empirical analysis, we use Näf et al. (2023)'s bootstrap procedure to compute confidence intervals for our statistics of interest.

3. Data

Our analysis is based on the German Microcensus for the years 2005 and 2019 (Federal Statistical Office, 2024). The Microcensus is conducted annually and provides a 1 % random sample of the German population, including information on income and socioeconomic characteristics of all persons in the households surveyed. It is the largest sample survey in Germany and in Europe. Data quality is high, and non-response is low due to mandatory participation. Most parts of our analysis rely on the Scientific Use File (SUF) of the Microcensus (Federal Statistical Office, 2024). For analyses requiring local identifiers at the municipality level, we use a restricted version of the Microcensus, accessible only onsite at the Research Data Centers (RDC) of the Federal Statistical Offices.

Although the Microcensus is the largest and most representative sample survey for Germany, it has rarely been used for income distribution analysis (Boehle, 2015; Hochgürtel, 2019; Walter et al., 2022). One reason for this is its grouped income information. In the two survey years analysed by us, respondents were asked to provide information on monthly household net income in income brackets of increasing width. The income brackets used for grouped income data are given in table 1. Note that the last income group is open-ended.

Table 1. Income brackets household net income (euros)

$(0; 150]$	$(150; 300]$	$(300; 500]$	$(500; 700]$	$(700; 900]$
$(900; 1,100]$	$(1,100; 1,300]$	$(1,300; 1,500]$	$(1,500; 1,700]$	$(1,700; 2000]$
$(2,000; 2,300]$	$(2,300; 2,600]$	$(2,600; 2,900]$	$(2,900; 3,200]$	$(3,200; 3,600]$
$(3,600; 4,000]$	$(4,000; 4,500]$	$(4,500; 5,000]$	$(5,000; 5,500]$	$(5,500; 6,000]$
$(6,000; 7,500]$	$(7,500; 10,000]$	$(10,000; 18,000]$	$(18,000; \infty)$	

Source: German Microcensus, 2005, 2019

Following standard practice, we adjust income data using the OECD equivalence scale. This scale assigns a weight of one to the first person in the household, and weights of 0.5 to each additional person aged over 14 years, as well as 0.3 to each additional person aged up to 14 years. For example, if household's net income falls within the interval $(4000; 4500]$, the equivalised income for a household with two

adults and two children (equivalence weight $= 1+0.5+0.3+0.3 = 2.1$) is transformed into the interval $(1904; 2143]$. As the distributional random forest can handle multivariate outcomes, we define as its dependent outcomes the upper and lower limits of these intervals, i.e., $\mathbf{Y} = (y_{\text{lower}}, y_{\text{upper}})$.

To ensure applicability across all income groups, we impose an upper limit on the highest income bracket, which is open-ended in the data. Following Walter et al. (2022), we define this limit as $3 \cdot 18,000 = 54,000$, resulting in a top interval of $(18,000; 54,000]$. Walter et al. (2022) did not provide a formal justification for their choice. However, a reasonable rationale is that household incomes in the Microcensus follow an approximate Pareto tail with $\alpha = 2$, implying that the midpoint of the interval $(18,000; 54,000]$ aligns with the expected income of this group, i.e., $\mathbb{E}(\text{household income}|\text{household income} > 18,000) = \alpha/(1-\alpha) \cdot 18,000 = 36,000$ (Blanchet and Piketty, 2022, p. 275). This approach is consistent with practices for grouped data, where interval midpoints are commonly used as approximations for group means. We found that our results are fairly robust to different choices of the upper limit, as only a small fraction of observations fall into the top income interval (0.21% in 2005 and 0.45% in 2019).

The equivalisation procedure produces overlapping income intervals for our observations, which is not an issue for the distributional random forest. In order to calculate a proper distribution function $F^r(y|\mathbf{X})$ for equivalised incomes y given charactersistics $\mathbf{X}$, we cumulate up probability masses across upper interval limits, i.e., we ask what fraction of observations have equivalent income up to A_1, up to $A_2, \ldots$, etc., where $A_1, A_2, \ldots$ represent the ordered upper interval limits for equivalised incomes appearing in the data. This produces the conditional cumulative distribution function for equivalised incomes representing all the available information in the data.

We use the resulting income groups $(A_1, A_2], (A_2, A_3], \ldots$, along with their implied frequencies for calculating statistics of interest (quantiles, means, Gini-coefficients) based on the formulae for grouped income data developed by Tille and Langel (2012). When calculating and aggregating distributions, we take into account the sampling weights of the Microcensus. By contrast, it is at present not possible to fully incorporate sampling weights into the training of the random forest. We do not expect this to influence our estimation results in substantial ways as the variation of the Microcensus weights is very limited. As a sensitivity check, we re-estimated some models using a reweighted sample based on the original sample weights. This led to results that were in most cases nearly identical to those from the original sample.

4. Empirical analysis

We now present our set of applications and elaborate on our implementation of distributional random forests for the German income distribution.

4.1. Estimating income distributions for granular population subgroups

Our first goal is to estimate distributions of equivalised net incomes for narrowly defined population subgroups. This is a relevant task for monitoring the well-being of specific groups, especially those at risk of poverty or social exclusion. To define population subgroups, we leverage the rich set of socio-economic characteristics at the

Table 2. Covariates for distributional random forest

	2005		2019	
Variable	Mean	Std.dev.	Mean	Std.dev.
# adults in hh	2.032	0.802	1.981	0.792
# adults 18-29 years	0.380	0.668	0.340	0.645
# adults 30-49 years	0.866	0.869	0.717	0.842
# adults 50-64 years	0.434	0.717	0.536	0.766
# adults 65+ years	0.351	0.668	0.386	0.698
# children in hh	0.691	1.028	0.642	0.642
# children 0-3 years	0.123	0.382	0.136	0.403
# children 4-6 years	0.304	0.651	0.403	0.637
# children 7-17 years	0.262	0.581	0.217	0.217
# adults foreign nationality	0.159	0.556	0.239	0.652
Share foreign adults > 0.5	0.090	0.286	0.134	0.341
# adults male	0.996	0.601	0.975	0.587
# adults female	1.035	0.496	1.006	0.508
0 FT[1], 0 PT, 0 MPT	0.294	0.454	0.253	0.408
0 FT, 0 PT, ≥ 1 MPT	0.028	0.164	0.029	0.167
0 FT, ≥ 1 PT, ≥ 0 MPT	0.046	0.207	0.072	0.258
1 FT, 0 PT, 0 MPT	0,240	0.426	0.202	0.401
1 FT, 0 PT, ≥ 1 MPT	0.059	0.235	0.049	0.215
1 FT, ≥ 1 PT, ≥ 0 MPT	0.117	0.320	0.171	0.376
≥ 2 FT, ≥ 0 PT, ≥ 0 MPT	0.212	0.408	0.221	0.414
# registered unemployed in hh	0.153	0.426	0.059	0.280
# unemployment benefits in hh	0.136	0.404	0.106	0.528
# adults tertiary education[2]	0.253	0.559	0.511	0.728
# adults higher secondary	0.174	0.452	0.196	0.476
# adults vocational training	1.097	0.894	0.919	0.865
# adults low education	0.507	0.775	0.775	0.689
East Germany	0.218	0.413	0.194	0.395
Indicators for 16 federal states		(details omitted)		
# observations		440,268		506,615

Source: Microcensus, 2005, 2019.

[1]FT = Full-time, PT = Part-time, MPT = Marginal part-time.

[2]Highest educational qualification.

individual and household level provided in the Microcensus. Since equivalised income is based on the assumption of income pooling within households, all covariates are constructed at the household level (for the equivalised income of a given individual,

it matters in what household she lives). A summary of the covariates $\mathbf{X}$ used in our analysis is shown in table 2.

Table 3. Tuning parameters of distributional random forest

Tuning parameter	Range	Description
num.trees	100, 200, 500, 1000	Number of trees
sample.fraction	0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1	Subsampling fraction ($=\beta$)
mtry	2, 3, 8, 12, 15, 20, 30	# variables tried for each split
min.node.size	2, 5, 10,15, 20, 25	Targeted minimal leaf size
alpha	0, 0.01, 0.05, 0.1, 0.25	Maximum imbalance of split ($=\alpha$)
imbalance.penalty	0, 0.05, 0.1	Imbalance of splits

Note: See Cevid et al. (2022) for more details.

Our first step is to fit and tune the distributional random forest. Cevid et al. (2022) did not discuss tuning of the distributional random forest. To arrive at a practically feasible procedure, we carry out the following steps. These are based on a training sample (40% of the original 2019 sample) and a test sample (30% of the original 2019 sample). The steps are as follows:

1. *Random parameter selection:* We randomly varied the tuning parameters within the ranges given in table 3, generating 300 random combinations.
2. *Training:* For each parameter combination, we fit the distributional random forest $F^r(y|\mathbf{x})$ on the training dataset.
3. *Testing:* We then computed the model's predicted aggregate distribution of equivalised incomes,
$$F^r(y) = \int_{\mathbf{x}} F^r(y|\mathbf{x})\, dF_{\mathbf{x}}(\mathbf{x}) \tag{7}$$
in the test sample, and compared it with the observed distribution of equivalised incomes $F^e(y)$ in the test sample.
4. *Evaluation:* We assessed goodness-of-fit using several statistical distance measures between $F^r(y)$ and $F^e(y)$, including Anderson-Darling, Cramer–von Mises, Kolmogorov-Smirnov, and Chi-square tests.

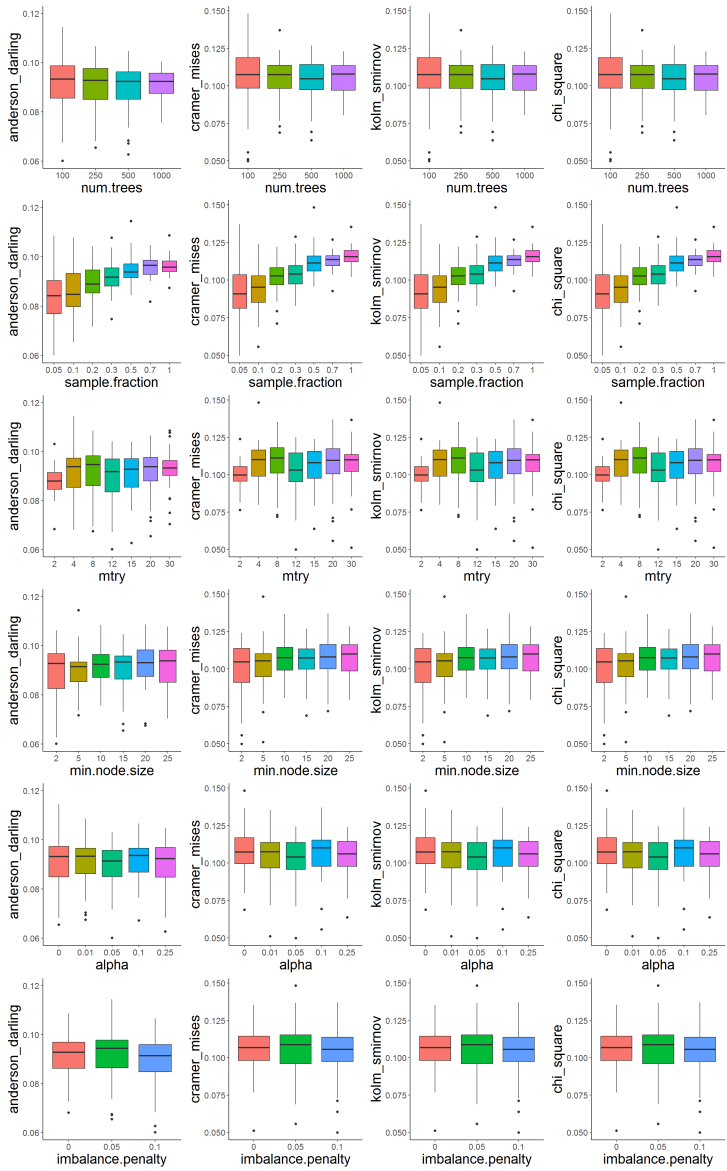
The results of this exercise are shown in figure 1. Minimizing discrepancy statistics between predicted and observed outcome distributions within reasonable ranges, we chose our final tuning parameters as num.trees=500, sample.fraction=0.1, mtry=12, min.node.size=5, alpha=0.05, imbalance.panelty=0.1. With these, we fit our final random forest based on the full sample. We found that our random forest results typically did not vary much across different specifications of tuning parameters. This is reflected in the small differences of goodness-of-fit between alternative choices (figure 1). Our final random forest model produced an aggregate income distribution function that was practically indistinguishable from the empirical distribution in the test set. This was generally true even for suboptimal tuning parameters.

Table 4. Ten examples of population subgroups.

Label	Description
Two pensioners	2 adults aged 65+ years (m+f), 2 vocational training, no employment, North-Rhine Westfalia
Single mother	1 female 18-29 years, PT, low education, 2 children (0-3, 4-11 years), Berlin
DINK	m+f 30-49 years, 2 FT, 2 tertiary education, Hamburg
5-person family BW	m+f 30-49 years, 1 FT, 1 PT, 1 tertiary, 1 voc. tr., 3 children (4-11, 2×12-17 years), Baden-Württemberg
Single immigrant	male 18-29 years, MPT, low education, foreign nationality, Lower Saxony
Single unemployed	male 30-49 years, registered unemployed, unemployment benefits, low education, Saxony
Immigrant family	m+f 30-49 years, 1 FT, 1 voc. training, 3 children (0-3, 2×4-11 years), foreign, North-Rhine Westfalia
Young professional	male 18-29 years, 1 FT, tertiary education, Berlin
Elderly widow	female 65+ years, no employment, Saxony-Anhalt
5-person family MW	m+f 30-49 years, 1 FT, 1 PT, 1 tertiary, 1 voc. tr., 3 children (4-11, 2×12-17 years), Meck-Westpomerania

Notes: m=male, f=female, FT=full-time, PT=part-time, MPT=marginal part-time, DINK=Double income no kids.

Fig. 1: Tuning of distributional random forest

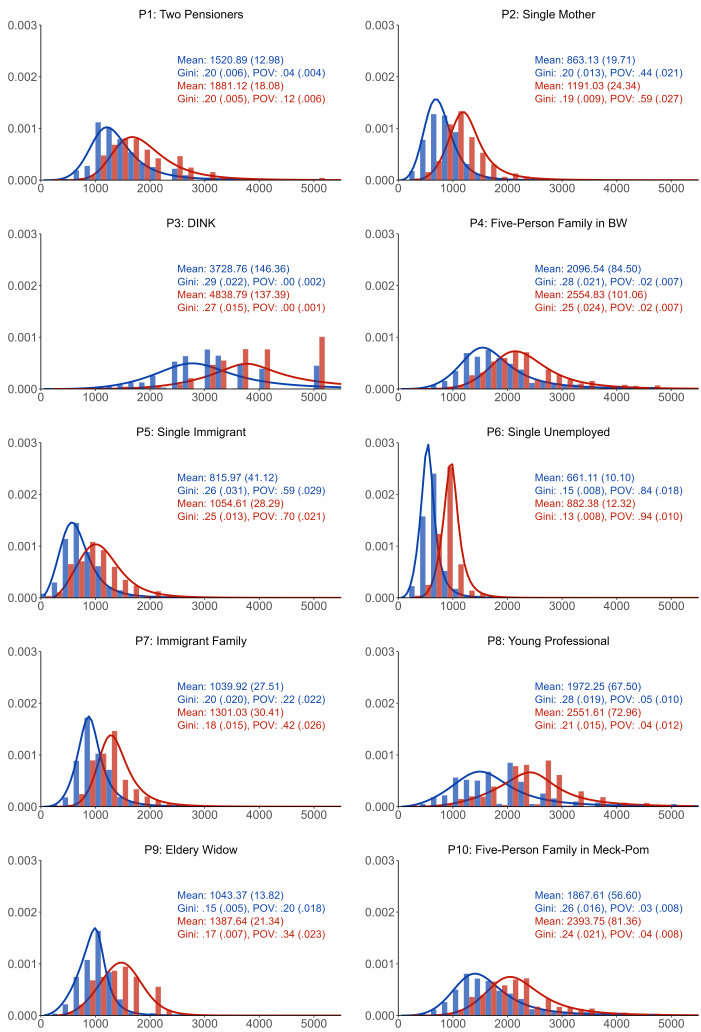


Note: The figure shows the distribution of goodness-of-fit indicators across 300 specification variants for tuning parameters on a test sample (low value = good fit).

Table 4 defines ten examples of narrowly defined population subgroups for which we estimate equivalised net income distributions using DRFs for 2005 and 2019. These subgroups consist of individuals in households with specific socio-economic

characteristics, as listed in table 2. In most cases, the number of observations for the particular type of individual would be much too low to estimate meaningful income statistics for the given group. Like in regression analysis, the random forest estimates group-specific outcomes by leveraging observations with similar characteristics (i.e., observations from other regions and households that are similar in terms of age, education and employment behaviours of their members).

Fig. 2: Distribution of equivalised net incomes in narrow population subgropus



Notes: See table 4 for the definition of types. Blue = 2005, red = 2019. POV = at-risk-of poverty rate (fraction of incomes below 60 % of population median). The figures include a GB2 density fit for illustration. Bootstrap standard errors in parentheses.

Figure 2 presents the results for different subgroups. For illustration, the graphs include a GB2 density fit. This is done purely for visualization purposes to highlight distributional shapes. The generalised beta distribution of the second kind (GB2) has been shown to provide a good fit to aggregate income distributions (Chotikapanich et al., 2018), but appears to fit less favourably in some of our group-specific distributions.

For each subgroup, we compute key statistics such as mean equivalised income, the at-risk-of-poverty rate, and the Gini coefficient, using the formulas from Tille and Langel (2012).[2]

The results shown in figure 2 exhibit highly plausible patterns. There is general income growth from 2005 to 2019, but gains in mean income are heterogenous. They range from 21.8% for individuals from the five-person family in BW to 37.9% for individuals in the single-mother household with two children as defined in table 4. As expected, the at-risk-of-poverty rate varies significantly across household types, from zero percent in double-income-no-kids households to approximately 90% in the single unemployed household (the characteristics of the latter were intentionally set to unfavourable values to produce an extreme result). Finally, there a large differences in within-group inequality across subgroups as measured by the Gini coefficient. Some groups are extremely homogenous (individuals in single unemployed households with a Gini coefficient of around 0.14), while others are highly heterogenous even within the narrow type definitions considered by us (individuals in the double-income-no-kids household with a Gini of around 0.28). Also note a general trend of decreasing within-group inequality for most types between 2005 and 2019 (an exception are elderly widows). In summary, once fitted, the distributional random forest enables policy makers and statistical agencies to flexibly monitor multiple aspects of economic welfare for finely defined population subgroups.

4.2. Analysis of distributional change over time

The distributional random forest captures information on income distributions for finely defined population subgroups. It can, therefore, be used to decompose aggregate distributional change into compositional and structural factors (Fortin et al., 2011). To this end, consider the counterfactual distribution

$$F^c_{\langle 19,05\rangle}(y) = \int_{\mathbf{x}} F^r_{2019}(y|\mathbf{x})\, dF_{\mathbf{x},2005}(\mathbf{x}), \tag{8}$$

which is the distribution of equivalised incomes that would have prevailed in 2019, if the distribution of household characteristics $F_{\mathbf{x}}(\mathbf{x})$ had still been as in 2005. This gives rise to the decomposition

$$F_{\langle 19,19\rangle}(y) - F_{\langle 05,05\rangle}(y) = \underbrace{F_{\langle 19,19\rangle}(y) - F^c_{\langle 19,05\rangle}(y)}_{\text{Composition effect}} + \underbrace{F^c_{\langle 19,05\rangle}(y) - F_{\langle 05,05\rangle}(y)}_{\text{Structural effect}}, \tag{9}$$

i.e., changes in the distribution of equivalised incomes between 2005 and 2019 are decomposed into effects explained by changes in the population composition $F_{\mathbf{x}}$, and

[2] Following standard practice in European countries, the at-risk-of-poverty rate is defined as the proportion of the population with equivalised income below 60% of population median income.

by changes in income structures $F^{\tau}(y|\mathbf{x})$ as described by the distributional random forest.

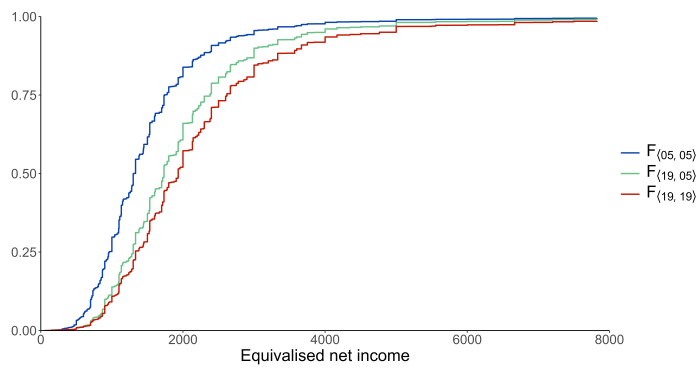
Table 5. Decomposition of distributional change, 2005-2019

	2005	Counterfactual[1]	Std.err.	2019
Mean	1,554.85	2,022.35	3.95	2,297.03
Gini	0.304	0.297	0.001	0.321
P10	707.74	915.97	8.76	986.13
P50	1,306.76	1,731.65	0.07	1,927.91
P90	2,396.28	3,034.72	15.70	3,616.29
P90/P10	3.386	3.313	0.038	3.667
P90/P50	1.834	1.753	0.009	1.876
P50/P10	1.846	1.891	0.018	1.955
At-risk-of-poverty rate	0.136	0.140	0.001	0.171

Source: Microcensus 2005, 2019. Own computations.

[1]Population composition from 2005, income structure from 2019.

Fig. 3: Aggregate decomposition, 2005-2019



Source: Microcensus 2005, 2019. Distribution functions. Own computations.

Results for this decomposition are presented in table 5 and figure 3. Between 2005 and 2019, we observe general income growth, but also an increase in inequality and poverty risk: mean equivalised income increased from 1,555 to 2,292 euros, the median rose from 1,307 to 1,927, while the Gini coefficient increased from 0.304 to 0.321 and the at-risk-of-poverty rate rose from 0.136 to 0.171. Inequality increased only in the lower half of the distribution: the P90/P10-ratio rose from 3.386 to 3.667, but this was entirely driven by an increase of the P50/P10-ratio from 1.846 to 1.955, while the P90/P50-ratio only changed very little from 1.834 to 1.876.

The counterfactual results in the middle column of table 5 suggests that holding the population composition fixed at its 2005 level while updating income structures to their 2019 levels accounts for most of the observed income growth between 2005 and 2019. However, this shift has little impact on inequality and poverty levels. Indeed, when only income structures $F^r(y|\mathbf{x})$ are updated while keeping composition constant, inequality as measured by the Gini coefficient and the P90/P10 ratio even slightly declines (from 0.304 to 0.297, and from 3.386 to 3.313, rows 2 and 6 of table 5). Changing only income structures but holding composition fixed slightly increase the at-risk-of-poverty rate and the P50/P10 ratio (from 0.136 to 0.140, and from 1.846 to 1.890, respectively). However, these effects are very small. On the other hand, changing income structures mitigated inequality in the upper half of the distribution as indicated by the counterfactual fall of the P90/P50-ratio from 1.834 to 1.753.

In contrast, adjusting the population composition to its 2019 level implies large increases in inequality and poverty risk (middle vs. last column of table 5). These largely account for the observed increase in inequality and poverty between 2005 and 2019, suggesting that the rise in inequality over this period can be fully explained by compositional changes in the population.

How did the composition of the population change between 2005 and 2019? Table 2 presents these shifts. We observe significant population aging, an increasing share of households with foreign nationals, greater heterogeneity in employment outcomes, and a growing polarization in educational qualifications. All of these changes increased population heterogeneity, which in turn amplified income inequality. In some cases, the observed shifts also increased the proportion of low-income households, thereby raising the aggregate at-risk-of-poverty rate.

Figure 3 provides a graphical summary of changes between 2005 and 2019. Changing income structures significantly shifted the distribution upwards, with no apparent impact on inequality. Adjusting in addition population composition to its 2019 level provides further – much weaker – income growth but contributes to stretching the distribution to the right, indicating higher inequality.

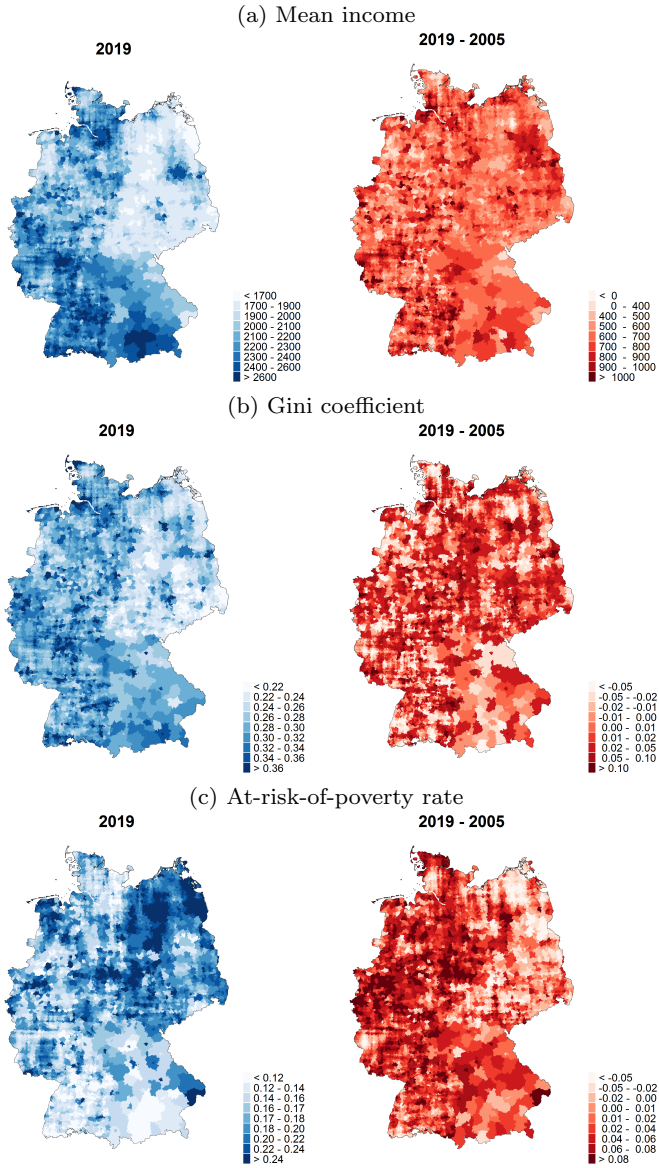
4.3. Spatial smoothing of income distributions

In this section, we leverage the smoothing property of the distributional random forest to estimate local income distributions. To this end, we utilise detailed geographical data from the Microcensus down to the municipality level. Germany has approximately 10,000 municipalities, including 2,000 towns and cities and around 8,000 smaller administrative entities that combine multiple geographic units. Due to stricter data protection rules, geographical data for Bavaria are only available at the county level, which represents the next administrative tier above municipalities. Our analysis therefore uses county-level data for Bavaria, while retaining municipality-level data for all other German regions.

In order to estimate local distributions of net equivalised income, we fit a distributional random forest based on the latitude and longitude of geographical units, i.e., we estimate

$$F^r(y|(\text{latitude}, \text{longitude})), \tag{10}$$

Fig. 4: Distributional indices and their change between 2005 and 2019



Source: Microcensus 2005, 2019. Computations are based on estimated distributions of equivalised net incomes at the municipality level (county level for Bavaria).

where (latitude, longitude) refer to the center of a geographical unit. Our approach produces estimates of local income distributions from which we compute measures of location and inequality as in the previous sections. Our approach is conceptually

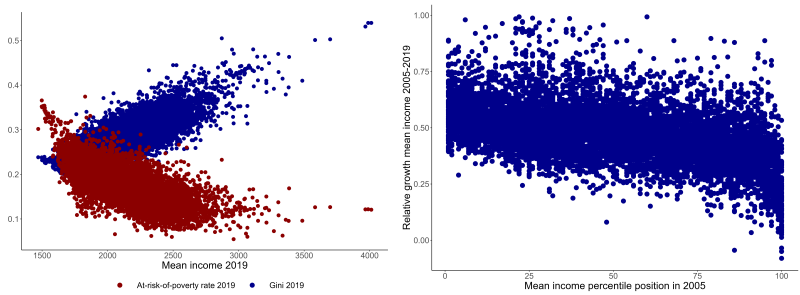
similar to Sugasawa et al. (2020), who smooth local income distributions based on a latent spatial correlation structure. Additionally, it relates to small-area estimation methods, which share the same goal but typically rely on explicit area-level models absent in our approach (Fabrizi et al., 2020; Gardini et al., 2022; Molina et al., 2022; De Nicolo et al., 2024).

Figure 4 presents maps of distributional indices for Germany. To the best of our knowledge, these are the first maps for Germany providing distributional indices for net incomes at the municipality level. Net incomes are widely regarded as the most informative indicators for personal financial well-being as they represent net disposable incomes after government transfers, taxes and social security deductions. Frieden et al. (2023) and Garbasevschi et al. (2023) have presented maps at the municipality level but for pre-tax incomes. Immel and Peichl (2020) and Walter et al. (2022) analysed regional differences in household net incomes, but at much higher level than municipalities. Schluter and Trede (2024) present a spatial analysis of wage incomes across regional labour markets, which are also defined at a higher level than municipalities.

The local distributional indices presented in the maps have several important applications. First, they allow statistical agencies and policy makers to monitor local levels of well-being and to identify areas with high and low levels of income or inequality. Second, the high degree of spatial heterogeneity is interesting in its own right, providing useful variation for studying relationships between different aspects of the distribution. For example, the left-hand graph in figure 5 plots the Gini coefficient and the at-risk-of-poverty rate against the mean income of geographical units. Mean income and inequality as measured by the Gini coefficient turn out to be positively related, i.e., geographical units with high mean equivalised incomes also tend to exhibit higher income inequality. In contrast, there is a weakly negative relationship between mean income and the at-risk-of-poverty rate, which is a likely consequence of the fact that the poverty threshold is defined at the national level (60 % of national median income). The right-hand graph of figure 5 relates the relative change of mean income in a geographical unit to the original relative position of the unit in the base year 2005. The results indicate that units with relatively low mean income in 2005 experienced higher relative income growth than those with a higher initial income level, suggesting convergence of mean incomes across regions. However, growth rates exhibit considerable variation, suggesting that this relationship is only approximate.

An important additional application of the data in figure 4 is its potential as explanatory variables in microeconomic or spatial analyses. Local measures of income, inequality, or poverty can serve as covariates in studies of individual behaviour (e.g., the effect of local inequality on individual consumption behaviour), or local outcomes (e.g., the impact of poverty rates on local election outcomes). In order to support such applications, we will make our estimates of distributional indices for the around 9,000 geographical units considered by us available in the supplementary material to this study.

Fig. 5: Associations between distributional aspects across geographical units



Source: Microcensus 2005, 2019. The graph on the left plots Gini coefficient and at-risk-of-poverty rate against mean equivalised income for 2019. The graph on the right plots growth rate of mean income 2005 to 2019 against percentile position of mean income in distribution of geographical units of 2005.

4.4. Purging spatial income distributions of differences in spatial characteristics

As our final application, we address the problem of correcting spatial income distributions for differences in spatial characteristics to obtain a *pure* spatial income structure - one that is independent of the fact that individuals in different regions tend to have different characteristics. To achieve this, we fit a distributional random forest conditional on location *and* characteristics, i.e.,

$$F^r(y|(\text{latitude}, \text{longitude}), \mathbf{x}). \tag{11}$$

Here, (latitude, longitude) represent the coordinates of a geographical unit as before, and $\mathbf{x}$ includes all household characteristics listed in table 2 (except the regional indicators, whose information is now captured by latitude and longitude).

In order to construct local income distributions that do not depend on the local composition of household characteristics, we consider

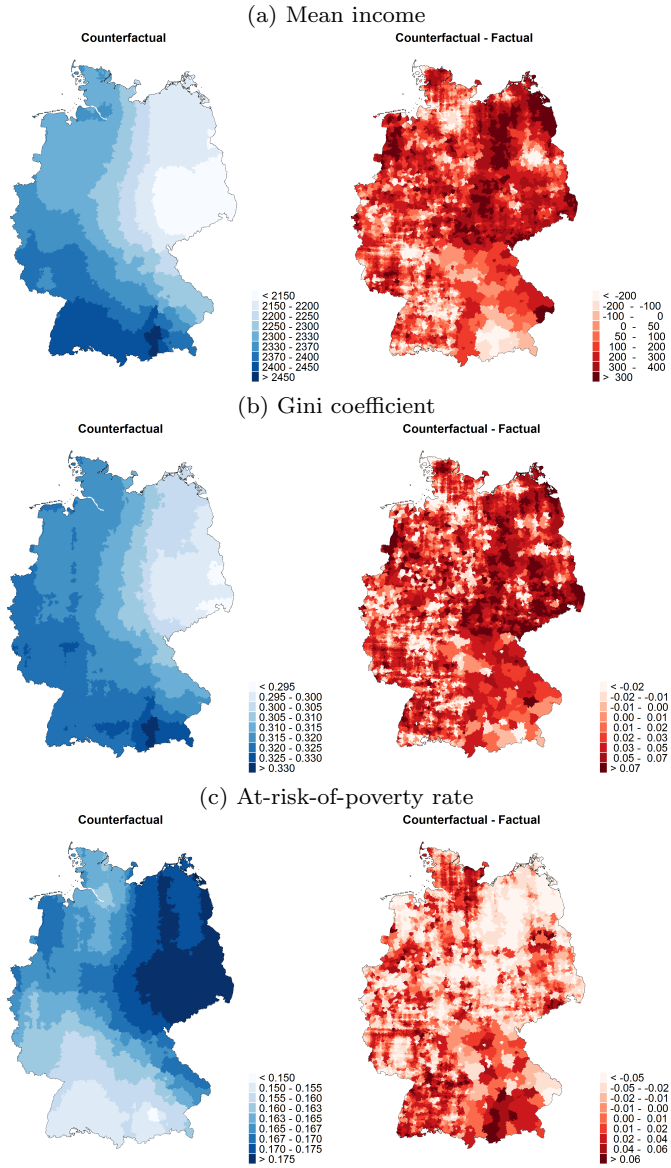
$$F^c(y|(\text{latitude}, \text{longitude})) = \int_{\mathbf{x}} F^r(y|(\text{latitude}, \text{longitude}), \mathbf{x})\, dF_{\mathbf{x}, \text{Germany}}(\mathbf{x}), \tag{12}$$

i.e., the local income distribution that would prevail if the distribution of household characteristics in region (latitude, longitude) were the same as in the whole of Germany.

This results in informative maps, as shown in figure 6. The results reveal a divide in mean income, inequality and poverty risk between East and West Germany, as well as between North and South. Assuming equal composition in all regions indicates both lower mean income and lower inequality in the East than in the West, as well as a higher degree of poverty risk in the East. To some extent, similar disparities persist between northern and southern Germany.

However, these regional differences are much smaller than when regional characteristics are allowed to vary. For example, under the assumption of equal composition

Fig. 6: Purged spatial income distributions



Source: Microcensus 2005, 2019. The graphs on the right show difference between counterfactual and factual maps.

across regions, the range of mean equivalised incomes is between 2,150 and 2,460 euros (figure 6), compared to a much wider range of 1,700 to 2,600 euros when composition is allowed to vary across regions (figure 4). Similarly, the range of Gini coefficients

across geographical units is 0.295 to 0.330 under the equal composition assumption (figure 6), whereas it spans from 0.220 to 0.360 when allowing for variation in household characteristics (figure 4). A similar pattern is observed for the at-risk-of-poverty rate. These findings suggest that variations in household characteristics across regions significantly contribute to disparities in income, inequality, and poverty risk.

5. Conclusion

Our analysis demonstrates that distributional random forests are a powerful and versatile tool for analysing income distributions with minimal parametric assumptions. Once trained, they allow for the estimation of any distributional index – quantiles, means, Gini coefficients, poverty rates, etc. – without requiring separate model specifications. They also easily handle grouped income information as present in our application. By applying this technique to the German Microcensus data, we illustrated four key applications relevant to both researchers and policymakers: (i) estimating granular subgroup distributions, (ii) analysing temporal changes in inequality and poverty, (iii) spatial smoothing of local income distributions, and (iv) purging spatial distributions of differing household characteristics.

From these analyses, we derived several insights about the German income distribution. First, the shape and location of income distributions vary dramatically across granular population subgroups, and income growth exhibits considerable heterogeneity. Second, while average incomes grew between 2005 and 2019, so did income inequality and the at-risk-of-poverty rate. However, the rise in inequality and poverty risk was almost entirely driven by compositional shifts (population aging, changes in educational attainment, a rising share of immigrants) rather than by diverging income trajectories for fixed population subgroups. Our geographical analysis provides new insights into the spatial structure of the German income distribution. We characterise regions with high or low income and inequality, showing that geographical units with higher mean incomes also tend to exhibit higher inequality. We show that income growth was uneven across regions, with poorer regions experiencing faster relative growth than wealthier ones, suggesting a degree of income convergence across space. Finally, we find that much of the observed regional variation in income and inequality is attributable to differences in household characteristics. After accounting for these compositional differences, residual 'pure' spatial disparities remain. These still follow clear geographical patterns but are less pronounced than the raw disparities observed in the raw data.

6. Acknowledgments

We are grateful to Sarah Bohnensteffen, Michael Knaus, Niko Muffler, Jeffrey Näf, Henri Pfleiderer, Klaus Pforr, Julie Schnaitmann, and participants of the 11th Mikrocensus User Conference at GESIS Mannheim for valuable comments and discussions. We thank Kerstin Stockmayer and Kristin Nowak at the Research Data Center of the Statistical Office of Baden Württemberg for their continuous support. Martin Biewen

is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2064/1 - Project 390727645.

References

Athey, S., J. Tibshirani, S. Wager (2019). Generalized Random Forests. *Annals of Statistics*, *47*, 1148-1178.

Boehle, M.. Armutsmessung mit dem Mikrozensus: Methodische Aspekte und Umsetzung für Querschnitts- und Trendanalysen. *GESIS Papers*, *2015/16*.

Biau, G., E. Scornet (2016). A random forest guided tour. *TEST*, *25*, 197-227.

Biewen, M., S. Jenkins (2005). A framework for the decomposition of poverty differences with an application to poverty differences between countries. *Empirical Economics*, *30*, 331-358.

Blanchet, T., T. Piketty (2022). Generalized Pareto Curves: Theory and Applications. *Review of Income and Wealth*, *68*, 263-288.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32.

Cevid, D., L. Michel, J, Näf, P. Bühlmann, N. Meinshausen (2022). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression (2022). *Journal of Machine Learning Research*, *23*, 1-79.

Chernozhukov, V., I. Fernandez-Val, B. Melly (2013). Inference on counterfactual distributions. *Econometrica*, *81*, 2205-2268.

Chotikapanich, D., W.E. Griffiths, G. Hajargasht, W. Karunarathne, D.S. Prasada Rao (2018). Using the GB2 Income Distribution. *Econometrics*, *6*, 1-24.

Cowell, F., E. Flachaire (2015). Statistical Methods for Distributional Analysis. In *Handbook of income distribution* (eds. A.B. Atkinson, F. Bourguignon), vol. 2., ch. 6, pp. 359-460. Amsterdam: Elsevier.

DeNicolo, S., M.R. Ferrante, and S. Pacei (2024). Small area estimation of inequality measures using mixtures of Beta. *Journal of the Royal Statistical Society, Series A*, *187*, 85-109.

Donald, S.G., D.A. Green, H. Paarsch (2000). Differences in wage distributions between Canada and the United States: An application of a flexible estimator of distribution functions in the presence of covariates. *Review of Economic Studies*, *67*, 609-633.

Fabrizi, E., M.R. Ferrante, C. Trivisano (2020). A functional approach to small area estimation of the relative median poverty gap. *Journal of the Royal Statistical Society, Series A*, *183*, 1273-1291.

Federal Statistical Office of Germany (2024). `https://www.forschungsdatenzentrum.de/de/haushalte/mikrozensus` (accessed Dec 2024).

Fortin, N., T. Lemieux, S. Firpo (2011). Decomposition Methods in Economics In *Handbook of Labor Economics* (eds. D. Card, O. Ashenfelter), vol. 4a, ch. 1, pp. 1-102. Amsterdam: Elsevier.

Frieden, I., A. Peichl, P. Schüle (2023). Regional Income Inequality in Germany. *EconPol Forum*, *24*, pp. 50-55.

Garbasevschi, A., H. Tabenböck, P. Schüle, J. Baarck, P. Hufe, M. Wurm, A. Peichl (2023). Learning Income Levels and Inequality from Spatial and Sociodemographic

Data in Germany. *Applied Geography*, *159*, 103058.

Gardini, A., E. Fabrizi, and C. Trivisano (2021). Poverty and inequality mapping based on a unit-level log-normal mixture model. *Journal of the Royal Statistical Society, Series A*, *185*, 2073-2096.

Gretton, A., K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola (2007). A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems*, vol. 19., pp. 513-520.

Hochgürtel, T. (2019). Einkommensanalysen mit dem Mikrozensus. *Wirtschaft und Statistik*, *3*, pp. 53-64.

Hothorn, T., B. Lausen, A. Benner, M. Radespiel-Tröger (2004). Bagging Survival Trees. *Statistics in Medicine*, *23*, pp. 77-91.

Hothorn, T., T. Kneib, P. Bühlmann (2013). Conditional transformation models. *Journal of the Royal Statistical Society, Series B*, *75*, pp. 1-24.

Hothorn, T., A. Zeileis (2021). Predictive Distribution Modeling Using Transformation Forests. *Journal of Computational and Graphical Statistics*, *30*, pp. 1181-1196.

Immel, L., A. Peichl (2020). Regionale Ungleichheit in Deutschland: Wo leben die Reichen und wo die Armen? *Ifo Schnelldienst*, *73*, 43-47.

Jenkins, S.P., P. Van Kerm (2009). The measurement of economic inequality. In *The Oxford Handbook on Economic Inequality* (eds. W. Salverda, B. Nolan, T. Smeeding), ch. 3, pp. 40-67. Oxford University Press.

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

Krenmair, P. T. Schmid (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society, Series C*, *71*, pp. 1865-1894.

Lin, Y., Y. Jeon (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, *101*, pp. 578-590.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, pp. 983-999.

Muandet, K., K. Fukumizu, B. Sriperumbudur, B. Schölkopf (2017). Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning*, *10*, pp. 1-144.

Molina, I., P. Corral, M. Nguyen (2022). Estimation of poverty and inequality in small areas: review and discussion. *TEST*, *81*, pp. 1143-1166.

Näf, J., C. Emmenegger, P. Bühlmann, N. Meinshausen (2023). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, *23*, 1-79.

Rigby, R.A., D.M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, *54*, pp. 507-554.

Schluter, C., M. Trede (2024). Spatial earnings inequality. *Journal of Economic Inequality*, *22*, pp. 531-550.

Schlosser, L., T. Hothorn, R. Stauffer, A. Zeileis (2019). Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain. *Annals of Applied Statistics*, *13*, pp. 1564-1589.

Sugasawa, S., G. Kobayashi, Y. Kawakubo (2020). Estimation and inference for area-wise spatial income distributions from grouped data. *Computational Statistics and*

Data Analysis, *145*, 106904.

Tille, Y., M. Langel (2012). Histogram-based interpolation of the Lorenz curve and Gini index for grouped data. *The American Statistician*, *66*, pp. 225-231.

Tzavidis, N., L.C. Zhang, A. Luna (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society, Series A*, *181*, pp. 927-979.

Walter, P., M. Groß, T. Schmid, K. Weimer (2022). Iterative Kernel Density Estimation Applied to Grouped Data: Estimating Poverty and Inequality Indicators from the German Microcensus. *Journal of Official Statistics*, *38*, pp. 599-635.