Taming the Tail: Sparse Top Incomes and Inequality Measurement in China, 2012–2018*

Kristina Butaeva[1]

[1]Harris School of Public Policy Stone Center for Research on Wealth Inequality and Mobility, University of Chicago, 1307 E 60th St, Chicago, 60637, IL, United States.

Contributing authors: kbutaeva@uchicago.edu;

Abstract

This paper develops and applies new measurement procedures to the data from the China Household Finance Survey (CHFS) to estimate income inequality in China from 2012 to 2018. The proposed estimation method aims to account for the sparse influential observations of the top income earners in the survey data and involves two main attributes. First, to approximate income distribution, I employ a semiparametric density model that combines a nonparametric kernel density approximating the bulk of the distribution with a Generalized Pareto Distribution (GPD) for the upper tail. Second, to estimate the GPD parameters, I propose the use of a Maximum Penalized Likelihood Estimator (MPLE) with a beta penalty function tuned to model the top income distribution. The results yield estimates of the Gini index in China of 0.616 in 2012, 0.604 in 2014, 0.581 in 2016, and 0.590 in 2018. These estimates are higher than those obtained by applying the same estimation procedures to the data from the China Family Panel Studies (CFPS) in a supplementary analysis in this paper. Nevertheless, the results are consistent with the Gini index estimates from the previous literature that incorporate top income adjustments and are substantially higher than conventional estimates based solely on household survey data. The empirical analysis also examines rural–urban and regional disparities, as well as income inequality in five provinces (Shanghai, Guangdong, Liaoning, Henan, and Gansu).

Keywords: Income inequality, China, Generalized Pareto, penalized likelihood.

JEL Classification: D31 , E01

Funding: This research was supported by the James M. and Cathleen D. Stone Foundation and the Greater Eurasia Research Center (GEAR) at HKUST.

Acknowledgements: The author thanks Grigory Kosenok, Albert F. Park, Shlomo Weber, Steven N. Durlauf, Geoffrey T. Wodtke, Christopher M. Clapp, Li Gan, Jin Wang, Li Han, and Sujata Visaria for their helpful comments and valuable suggestions. The author also acknowledges Grace Kolavo and Ferrán Yagcier-Rodriguez for their instrumental support in preparation of the manuscript.

[0]ORCID: 0000-0001-5865-577X

*An earlier version of this paper was circulated under the title "Income Inequality in China, 2012–2018: A New Measurement Approach."

1 Introduction

China demonstrated spectacular economic growth accompanied by a massive reduction of poverty following the launch of sweeping market reforms in 1978. GDP per capita increased twenty three times from $447 in 1981 to $10,370 in 2020 (Figure 1), in turn, the share of the population living below the poverty line dropped from 92% in 1981 to 0.11% in 2020. Many scholars emphasize that this tremendous economic progress coincided with a "marked" (Luo et al. 2020), "rapid" (Knight 2014), "inexorable" (Kanbur et al. 2021), "substantial" (Jain-Chandra et al. 2018), and even "unprecedented" (Naughton 2006) increase in income inequality. Some scholars argue the rise of inequality in China might be a "price" for such spectacular economic progress (Kanbur et al. 2021), whereas others claim that there might not be a trade-off between them (Ravallion and Chen 2009).

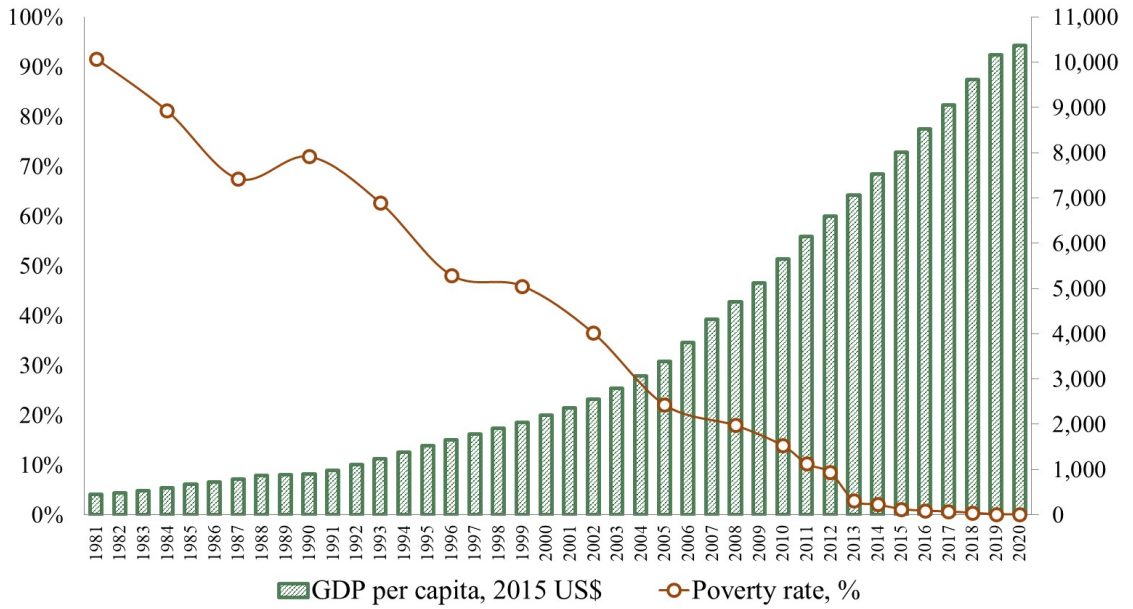


Fig. 1: GDP per capita estimates from Roser et al. (2023) and poverty rates from the "World Bank (2024), Poverty and Inequality Platform" https://pip.worldbank.org/poverty-calculator. Poverty rate is defined as the percentage of the population living in households with consumption or income per person below the poverty line at 2017 international prices ($2.15)

In the last decade, the National Bureau of Statistics (NBS) of China revealed a steady decline in the income inequality Gini index from 0.481 in 2010 to 0.468 in 2020. However, as noted by Li and Sicular (2014), the Chinese government initiated efforts to alleviate income inequality as a social issue in February 2013, when the State Council issued a special circular[1] that introduced new reforms targeted at addressing the problem. In the same year, some of these reforms were discussed at the Third Plenum of the Chinese Communist Party (CCP).[2] Attention remained dedicated to the issue in August 2021, where at the 10th meeting of the Central Financial and Economic Committee President Xi Jinping promoted "common prosperity" as one of the key priorities of current Chinese economic policy.[3] Despite evidence that income inequality in China may have slightly declined since 2007/2008,[4] most researchers share the view that existing high levels of income inequality presents warning signs for future barriers to economic progress and remains an important challenge for Chinese society and policymakers (Yang and Qiu 2016). It is worth noting that income inequality in China is important not only because of its potential negative consequences for the Chinese economy but also because inequality in China is a substantial contributor to global inequality (Milanovic 2002).

Current academic and policy discussions are complicated by the fact that there is still no consensus on accessing levels of and trends in income inequality in China. Along with the official rates

[1] State Council Circular No. 6. 2013. "Several opinions on deepening reform of the income distribution system," 5 February, http://www.gov.cn/zwgk/2013-02/05/content_2327531.htm.

[2] "CCP Central Committee resolution concerning some major issues in comprehensively deepening reform," 15 November 2013, https://chinacopyrightandmedia.wordpress.com/2013/11/15/ccp-central-committee-resolution-concerning-some-major-issues-in-comprehensively-deepening-reform.

[3] "Xi urges financial risk prevention while seeking high-quality growth," 17 August 2021, http://eng.mod.gov.cn/xb/News_213114/TopStories/4892182.html.

[4] See, e.g., Naughton (2006), Jain-Chandra et al. (2018), Luo et al. (2020).

of inequality announced by the NBS, there exist many alternative Gini index estimates[5] produced by researchers using numerous methods and based on a variety of independent household surveys;[6] this reveals both higher and lower levels of income inequality compared to the official statistics. The existing discrepancies result from the data and methodological issues, which persist as the most substantial challenge in the accurate measurement of income distribution. In this paper I employ a unique dataset, the China Household Finance Survey (CHFS), which was specially constructed to adequately represent affluent households.[7] Therefore, my baseline assumption is the reliability of the recorded top income values in the CHFS. I presume high incomes are carefully registered in the survey, but the data may still be sparse and, therefore, contain a small number of highly influential observations in the top tail of the income distribution. It is well-established in the statistical literature that many inequality measures are sensitive to extreme top-income values. I encountered this issue, observing top-income–driven instability in Gini index estimates across different survey years, which was even more pronounced when using disaggregated regional- and provincial-level data with smaller sample sizes. Most of the previous studies have been focused on improving the robustness of inequality measures towards top income contamination or measurement errors (Cowell and Flachaire 2007; Cowell and Victoria-Feser 2007; Davidson and Flachaire 2007), and the typical solutions include downweighting the importance of such observations in the estimation procedure or even fully ignoring them. This contradicts my basic assumption of the reliability of top income values recorded in CHFS, and therefore, this study is motivated to search for new estimation procedures that better account for high-income observations.

The goal of this study is twofold. First, it proposes a new estimation procedure for measuring the income distribution using survey data that efficiently accounts for the presence of extreme high-income values. Even with accurate measurement, top income observations in survey data may remain sparse, particularly in small samples, making them highly influential. The proposed method is designed to reduce the variability of inequality estimates across survey years that is driven by top-income observations. The method employs a semiparametric income distribution model akin to the one in Cowell and Victoria-Feser (2007). The distribution model consists of two parts: nonparametric kernel density approximating the bottom $100 \times p_\tau\%$ of the income distribution and a Generalized Pareto Distribution (GPD) which represents the top $100 \times (1-p_\tau)\%$ of incomes.[8] The key distinctive feature of this paper is that the parameters of the Generalized Pareto top income tail are estimated by applying a maximum penalized likelihood estimator (MPLE) originating from the extreme value models literature (Coles and Dixon 1999; Martins and Stedinger 2000, 2001; Pauli and Coles 2001; Huang et al. 2012). An MPLE is a computationally cheaper, simplified analogous to a Bayesian approach. A prior guess on the distribution of an estimated parameter is formulated as a likelihood penalty function, which aims to penalize "unreasonable" values of a parameter based on the expert's prior knowledge. The penalty function in an MPLE, by construction, does not depend on the sample size. Therefore, the maximum penalized likelihood keeps the large-sample optimality of a classical maximum likelihood estimator (MLE) and all of its asymptotic properties, but at the same time, it helps to substantially improve the MLE's efficiency in small samples and in the presence of sparse high-income observations. To demonstrate that the proposed MPLE outperforms other commonly used estimators of Pareto income tail parameters in the income inequality literature, I conduct a Monte Carlo simulation study. The simulation results show that MPLE not only addresses upper-tail sparsity but also mitigates concerns related to data contamination, underreporting, and tail underrepresentation, exhibiting superior finite-sample efficiency relative to the standard maximum likelihood estimator and the optimal bias-robust estimator commonly used in the income inequality literature, in samples of 50 to 500 observations. At the same time, MPLE achieves numerically equivalent point estimation performance to a fully Bayesian Markov chain Monte Carlo approach at a substantially lower computational cost.

Second, by applying the proposed estimation procedure to the data from the CHFS, a novel set of estimates for income inequality in China is provided. The obtained baseline Gini index values are 0.616 in 2012, 0.604 in 2014, 0.581 in 2016, and 0.590 in 2018. In a supplementary analysis, inequality patterns are also documented separately for urban and rural households, as well as across different regions in China. It is found that households in rural areas have lower

[5] Recent comprehensive literature reviews are presented in Zhang (2021) and Kanbur et al. (2021).

[6] Those include Chinese General Social Surveys (CGSS), Chinese Household Income Project (CHIP), China Health and Retirement Longitudinal Study (CHARLS), China Labor Force Dynamic Survey (CLDS), China Family Panel Studies (CFPS), China Household Finance Survey (CHFS), etc. For a detailed review of the existing data for studying income distribution in China, see Gustafsson et al. (2014).

[7] Piketty et al. (2019) pointed out that household surveys in China especially suffer from massive under-reporting at the top of the income distribution, making the opportunity to use a better-quality dataset both essential and exciting.

[8] In Cowell and Victoria-Feser (2007) the empirical distribution function is used to model the bulk of the distribution and a Pareto I model applied to approximate the top income tail.

incomes and similar levels of income inequality compared to urban households. Regional analysis confirmed a well-established fact that the coastal Eastern region is considerably richer than other Chinese regions. Eastern and Western regions are found to be the most unequal, whereas the least unequal is the Northeastern region. By employing a standard decomposition analysis of the Theil T inequality index, within-group variation both for urban-rural areas and geographic regions is found to explain the major component ($> 87\%$) of nationwide income inequality in China.

Additionally, a supplementary analysis of inequality based on the data from the China Family Panel Study (CFPS 2012, 2014, 2016, and 2018) was conducted. The results from the CFPS in the first two waves estimate inequality in China at relatively lower levels compared to the CHFS and in the latest two waves, and Gini index estimates converge to comparable levels in both datasets. Decomposition of the Theil T index by income source shows that the higher overall inequality in the CHFS is driven mainly by business income, concentrated among affluent households, which are particularly well covered in the CHFS. In contrast, labor income, better represented in the CFPS, contributes most to inequality there, but its lower dispersion produces smaller overall inequality compared to the CHFS. To further demonstrate the performance of the proposed estimation method in small samples, provincial-level Gini index estimates are produced using both CHFS and CFPS datasets. This analysis involved five provinces: Shanghai, Guangdong, Liaoning, Henan, and Gansu. This selection of provinces was driven by the availability of representative samples in the CFPS dataset.

The paper is organized as follows. Section 2 introduces the proposed measurement procedures. Section 3 describes the data. Section 4 presents the obtained income inequality estimates, makes comparisons to the previous literature, as well as demonstrates the robustness checks. Section 5 concludes.

2 Measurement

Table 1: Classification of upper-tail issues in income survey data

Category	**Issue**	**Description**
1. Data Quality *(recorded income ≠ true income)*	A. Contamination	Extreme observations are misreported, miscoded, or otherwise inconsistent with the underlying income distribution.
	B. Underreporting	Respondents intentionally understate high income.
2. Sampling *(reported income is accurate)*	C. Under-representation	Affluent households are systematically underrepresented in survey samples.
	D. Data Sparsity	The upper tail contains sparse observations, rendering statistical inference sensitive to extreme values; this issue is especially acute under heavy-tailed distributions (Davidson and Flachaire 2007).

One of the major measurement challenges in income inequality research refers to the fact that most of the standard inequality indices are substantively influenced by the top income observations. While their statistical sensitivity to measurement issues varies, the level of any index, including the most commonly used Gini index, responds meaningfully to changes at the top of the distribution. The Gini is theoretically more sensitive to transfers around the center of the distribution, and is relatively robust to contamination in the upper tail compared to entropy-based indices like the Theil index (Cowell and Victoria-Feser 1996; Cowell and Flachaire 2007). Nevertheless, its estimated value remains highly responsive to top income values because it assigns substantial weight to large income gaps across the distribution. Consequently, changes in top incomes are associated with significant shifts in aggregate inequality measures (Atkinson 2007; Alvaredo 2011).

Extreme top-income observations may be recorded with error (i.e., may contain data contamination), or they may represent valid observations that nonetheless exert a disproportionate influence on the volatility of estimated inequality measures over time. To consider these issues systematically, Table 1 classifies the most common problems affecting income survey data that are particularly salient in the upper tail of the distribution. The classification distinguishes between two broad categories: data quality issues, which arise when recorded incomes deviate from true

incomes (A and B), and sampling issues, which arise even when reported incomes are accurate (C and D).

To mitigate sample composition and representativeness issues (C and D), the existing literature typically suggests parametrizing the upper tail of the income distribution, or even imposing restrictive parametric assumptions on the entire income distribution. The modeling of income and wealth distributions has a long history, beginning with the use of parametric models,[9] through employing parametric mixture models,[10] and, later, through combining nonparametric models approximating the bulk with the parametric distributional tails.[11] Underreporting (B), under-sampling (C), and sparse representation of the upper income tail (D) are often addressed by incorporating administrative tax data into the estimation of the top part of the income distribution (Atkinson 2007; Atkinson et al. 2011; Alvaredo 2011; Blanchet et al. 2022; Jenkins 2017).

To account for data contamination at the top of the income distribution (A), a distinct but related literature emphasizes statistical testing to assess the influence of outliers on inequality indices. These studies typically propose semi-parametric income density models combined with robust inference procedures (Cowell and Victoria-Feser 1996; Cowell and Flachaire 2007; Davidson and Flachaire 2007). Among modern approaches in the income inequality literature, the Optimal B-Robust Estimator (OBRE) is widely used to reduce the sensitivity of inequality measures to potential data contamination (Victoria-Feser and Ronchetti 1994; Victoria-Feser 2000; Cowell and Victoria-Feser 2007; Van Kerm 2007; Jenkins 2017). Developed within the framework of influence function analysis (Hampel 1974; Hampel et al. 1986), the OBRE minimizes asymptotic variance subject to a bound on the bias induced by arbitrary contaminating distributions. It belongs to the class of M-estimators, a generalization of maximum likelihood estimators (MLE). The key idea is to apply the maximum likelihood score function to most observations while capping the influence of extreme values at an upper limit k. In essence, the OBRE is a weighted maximum likelihood estimator that "downweights" highly influential observations to improve the robustness of inequality estimates.

Leveraging the relatively rare feature of the CHFS survey design that emphasizes the representativeness of affluent households and improves coverage of the top income tail, this paper proceeds under the assumption that observed top incomes are measured accurately. Nevertheless, even with accurate measurement, top income observations may remain sparse, particularly in small samples such as CHFS province-level data. As a result, these observations are highly influential: a small number of extremely high incomes can exert a substantial impact on estimated Gini coefficients. For example, within a given province, the presence of a few extreme observations in one survey year—and their reduction or disappearance in another year due to non-response or genuine income changes—can generate substantial year-to-year variation in estimated Gini coefficients. Existing approaches provide only partial solutions to this problem. Parametric modeling of the upper tail can improve precision, but estimation of Pareto tail parameters entails a bias–variance trade-off. Maximum likelihood estimation is highly sensitive to extreme values, while robust alternatives such as OBRE, which aim to limit the influence of outliers, may incorrectly classify sparse but valid extreme incomes as contamination and consequently downweight or exclude them. However, the top income observations are exactly those that are most informative about tail behavior, and they are treated in this paper as accurate realizations of top incomes rather than as measurement error. These considerations motivate the need for an alternative estimation strategy. Accordingly, this paper proposes the use of a maximum penalized likelihood estimator (MPLE) with a penalty function informed by the existing empirical literature, designed to efficiently accommodate extreme-income observations and primarily address upper-tail sparsity (D) in small samples. As shown in the simulation study, the estimator also mitigates concerns related to data contamination (A), underreporting (B), and tail underrepresentation (C), thereby demonstrating broader finite-sample efficiency than its common alternatives.

2.1 Income distribution model

Building on Cowell and Victoria-Feser (2007), I employ a semiparametric model of the income distribution that differs in two ways. First, a nonparametric kernel density function is used to approximate the bulk of income distribution, instead of the empirical density function used in

[9] E.g., the Pareto distribution in Pareto (1895); the lognormal distribution in Gibrat (1931), Lebergott (1959), and Aitchison and Brown (1969); the gamma distribution in Salem and Mount (1974) and the generalized gamma distribution in Esteban (1986); the Weibull distribution in Bandourian et al. (2002); the Singh-Maddala distribution in Singh and Maddala (2008); and the Dagum distribution in Dagum (2008).

[10] Eg.: mixture of lognormals in Flachaire and Nunez (2007), Chotikapanich and Griffiths (2008), Lubrano and Ndoye (2016); lognormal plus Pareto distribution in Clementi and Gallegati (2016).

[11] See a broad review in Jenkins (2017).

their papers. Kernel density allows control of the smoothness of the distribution function in order to make it less sensitive to noisy individual values that appeared in the realized sample. Second, the upper tail is modeled by a more flexible three-parameter Generalized Pareto density, instead of a two-parameter Pareto I. In combined distribution models, the GPD tail had been proven to be less sensitive to the choice of the threshold compared to the Pareto I distribution (Charpentier and Flachaire 2022).

The proposed semiparametric probability density function (PDF) is defined by:

$$f(x) = \begin{cases} \dfrac{\sum_{i=1}^{n_\tau} w_i K\left(\dfrac{x_i - x}{h}\right)}{\int_0^\tau \sum_{i=1}^{n_\tau} w_i K\left(\dfrac{x_i - t}{h}\right) dt}, & \text{if } x \leq \tau \\ (1 - p_\tau) \times \left[\dfrac{1}{\sigma}\left(1 + \xi \dfrac{x - \tau}{\sigma}\right)^{-\frac{1}{\xi} - 1}\right], & \text{if } x > \tau. \end{cases} \tag{1}$$

where τ is a threshold income level, and p_τ is the corresponding threshold income percentile. The bottom $100 \times p_\tau\%$ of incomes are represented by the weighted nonparametric kernel density estimated on the n_τ observations below the threshold τ, where $w_i = n_\tau \tilde{w}_i / \sum_{j=1}^{n_\tau} \tilde{w}_j$ are the normalized survey weights, h is a smoothing parameter (bandwidth), and $K(\cdot)$ is a kernel function. The kernel density is renormalized to integrate to unity over $(0, \tau]$, ensuring that the semiparametric PDF in eq. (1) integrates to one over its full support. The upper $100 \times (1 - p_\tau)\%$ of incomes are approximated by a GPD model where τ is a location parameter, ξ is a shape, and $\sigma > 0$ is a scale. The shape parameter ξ of the GPD can, in principle, take any real value, with the limiting case $\xi = 0$ corresponding to the exponential distribution. In this paper, however, I restrict the analysis to the range $0 < \xi < 1$. Positive values of ξ imply a heavy-tailed, Pareto-type distribution (Singh and Guo 1995; Jenkins 2017), which has been shown in the economics literature to be particularly well suited for modeling top incomes (Kleiber and Kotz 2003; Atkinson and Piketty 2007). Restricting ξ to be less than one ensures that the mean of the GPD is finite (Coles 2001), a necessary condition for the estimation of the Gini index. More generally, the kth moment of the GPD exists only if $k < 1/\xi$ (Jenkins 2017). The corresponding cumulative distribution (CDF) and quantile functions are defined in the Appendix B.

2.2 Parameters estimation

For the choice of bandwidth h and kernel function $K(\cdot)$ in the nonparametric density, common rules of thumb are applied. By default, I use a Gaussian kernel with a smoothed cross-validation (CV) bandwidth estimator. With a Gaussian kernel, the estimated density can cover negative income values. To address this issue, the kernel density estimate is truncated at zero and then renormalized, ensuring that it integrates to unity over the non-negative domain. This transformation preserves the level of inequality within the positive income domain and yields a valid non-negative income distribution, to which the standard Gini formula can be applied directly. To check robustness (Section 4.4), the results are also re-estimated using skewed kernels with positive support, namely the asymmetric gamma kernel proposed by Abadir and Lawford (2004) and the log-transformation kernel of Charpentier and Flachaire (2015). In addition, I consider alternative bandwidth choices, using half of the cross-validation bandwidth ($0.5 \times CV$) and a bandwidth 1.5 times larger than the cross-validation bandwidth ($1.5 \times CV$).

A threshold percentile p_τ was arbitrarily chosen at 95^{th} percentile for the national-level samples and at the 90^{th} percentile for the provincial-level samples. The threshold choice was determined by balancing a standard bias-variance trade-off: the fewer the observations on the tail considered as a part of a GPD density, the more volatile the GPD parameters estimates, but the lower the potential bias, and vice versa—the lower the chosen threshold, the lower the variability of the GPD estimated parameters, and potentially the larger the bias (Charpentier and Flachaire 2022). The robustness of the baseline results was checked against the threshold choice by employing 85^{th}, 90^{th}, and 95^{th} percentiles for both the national- and provincial-level samples.

To estimate the GPD shape ξ and scale σ, a modified version of a penalized likelihood estimator from Coles and Dixon (1999) is employed. The distinctions are, first, the usage of a different penalty function more suitable for fitting incomes, and second, the application of survey weights. Let's call $\boldsymbol{\theta} = [\xi, \sigma]$ a vector of GPD parameters (shape and scale). For every fixed threshold

income τ, the upper-tail weighted likelihood is defined:

$$L(\boldsymbol{\theta}|\mathbf{x},\tau) = P(\xi) \times \prod_{i=1}^{n_\tau} [f(x_i,\tau|\xi,\sigma)]^{w_i} = \\ = P(\xi) \times \prod_{i=1}^{n_\tau} \left[\frac{(1-p_\tau)}{\sigma} \left(1+\xi\frac{x_i-\tau}{\sigma}\right)^{-\frac{1}{\xi}-1} \right]^{w_i}, \tag{2}$$

where $\mathbf{x} = (x_1, x_2, ..., x_{n_\tau})$ is the observed income data above the threshold ($x_i > \tau$, $i \in [1; n_\tau]$); $f(x_i, \tau)$ is a GPD density with a location parameter τ; w_i are survey weights; and $P(\xi)$ is a penalty function. Survey weights w_i are normalized so that $\sum_{i:x_i>\tau} w_i = n_\tau$, where n_τ denotes the number of observations above the threshold. This normalization ensures that the influence of the penalty term $\log(P(\xi))$ relative to the weighted log-likelihood remains comparable to the unweighted case studied in the simulation, and does not diminish as a function of the survey's population expansion factors. Optimal $\hat{\boldsymbol{\theta}}(\tau)$ could be found by solving:

$$\hat{\boldsymbol{\theta}}(\tau) = \arg\max_{\boldsymbol{\theta}} \left\{ \log(P(\xi)) + \sum_{i=1}^{n_\tau} w_i \times \log\left(\frac{(1-p_\tau)}{\sigma} \left(1+\xi\frac{x_i-\tau}{\sigma}\right)^{-\frac{1}{\xi}-1} \right) \right\}. \tag{3}$$

The objective of incorporating a penalty function $P(\xi)$ is to employ prior knowledge about the distribution of a GPD shape parameter to reduce the variability of parameter estimates. This setting is close to a Bayesian approach, where parameters are treated as random variables with the initial guesses on their true distributions formulated as priors. Only the GPD shape ξ is penalized for two reasons. First, this parameter is the most important for estimating inequality measures because it defines the thickness of the GPD tail. Second, it is presumed that a reasonable guess regarding more or less possible values of ξ might be inferred from the existing income inequality literature.

Coles and Dixon (1999) considered a Generalized Extreme Value (GEV) distribution for fitting data on a sea level. They defined a likelihood penalty as a monotonically decreasing function, increasingly penalizing larger values of the GEV shape. They assumed that lower-level GEV shapes are more probable compared to large shapes in the sea level data.[12] Their penalty is designed for sea-level extremes where large positive shapes are implausible, and is not directly appropriate for income distributions where the shape parameter must be strictly positive to generate the heavy Pareto-type tails. In our case, for fitting the income model, the beta distribution penalty function used in Martins and Stedinger (2000) is utilized for the GEV distribution shape and later in Martins and Stedinger (2001) for the GPD shape. The penalty is formulated as following:

$$P(\xi) = \frac{\xi^{\alpha-1}(1-\xi)^{\beta-1}}{B(\alpha,\beta)}, \tag{4}$$

where $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, and Γ is the Gamma function. The beta distribution is convenient in this setting because for some parameter values beta might transform into an inverted U-shape function, which allows for assigning a higher probability for some medium range of ξ values and penalizing unreasonable extremely large and extremely small ξ values. It is also convenient that the support of the beta distribution ranges from 0 to 1, while the Gini index for the heavy-tailed GPD is well defined only for shape parameters $\xi \in (0, 1)$.

The penalty induces a finite-sample bias proportional to the gradient of $\log P(\xi)$ at the true parameter and inversely related to the curvature of the penalized objective. This reflects the standard bias–variance trade-off in penalized likelihood estimation (Coles and Dixon 1999; Martins and Stedinger 2001): the penalty introduces bias while reducing the variance of $\hat{\xi}$, shrinking estimates toward the prior mode ξ^* (implying positive bias if $\xi < \xi^*$ and negative bias if $\xi > \xi^*$). As $n \to \infty$, the penalty becomes negligible and the MPLE converges to the MLE, sharing the same asymptotic variance. In small samples from heavy-tailed distributions such as the GPD, sparse and highly influential upper-tail realizations arise mechanically (Davidson and Flachaire 2007), even in the absence of measurement error. Since empirical applications rely on a single realized sample rather than repeated draws, this feature can generate substantial instability in inequality measures

[12] Penalty function from Coles and Dixon (1999): $P(\xi) = \begin{cases} 1, & \text{if } \xi \le 0 \\ \exp\left\{-\lambda\left(\frac{1}{1-\xi}-1\right)^{\alpha}\right\}, & \text{if } 0 < \xi < 1 \\ 0, & \text{if } \xi \ge 1 \end{cases}.$

across periods. The MPLE is therefore intended to stabilize estimation in such settings, trading a controlled and predictable bias for reduced variance. The simulation analysis in Section 2.4 is designed to verify that the induced bias remains quantitatively modest.

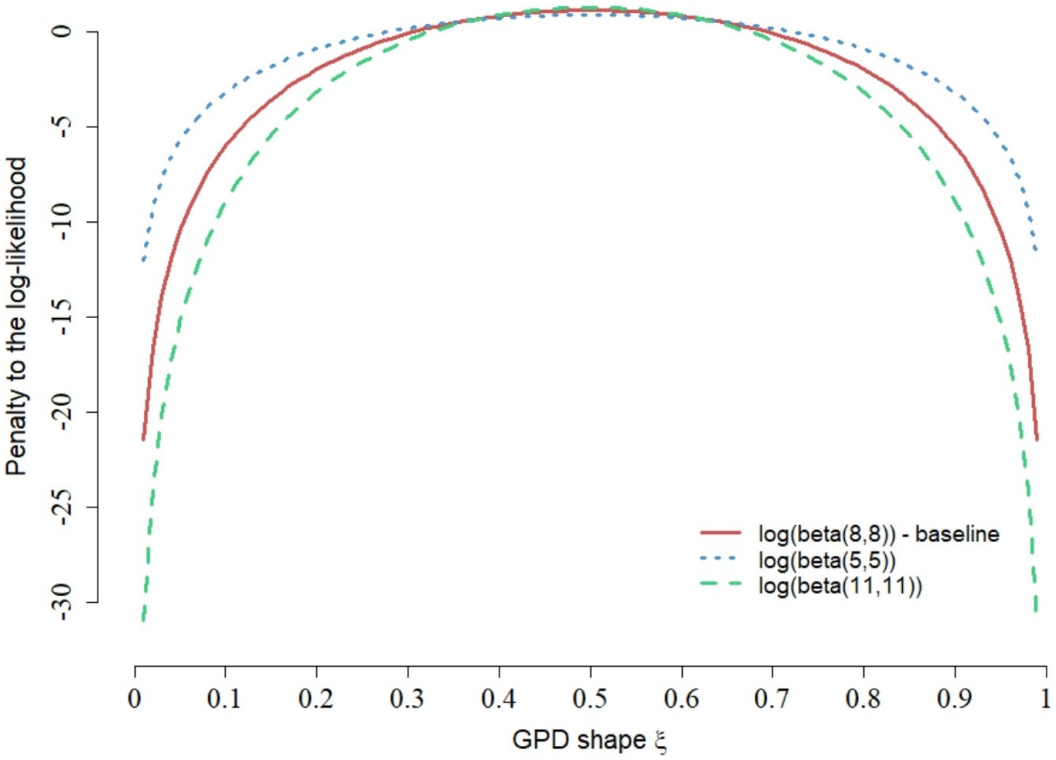


Fig. 2: Different $\log(P(\xi))$ penalties to the log-likelihood function against the GPD shape ξ. $P(\xi) = beta(8, 8)$ is a baseline penalty. $Beta(5, 5)$ and $beta(11, 11)$ are penalties used in the robustness analysis

For a more precise formulation of a penalty, I will use an observation made by Atkinson et al. (2011) in their multi-country study of top incomes. They pointed out that in practice ratios between the average income of individuals above the top income threshold, and the threshold value—which is called the inverted Pareto coefficient "b"—typically vary between 1.5 and 3 where values of b between 1.5 and 1.8 indicate low income inequality and near to or above 2.5 indicate high inequality. In their paper, Atkinson et al. (2011) used a two-parameter Pareto I distribution[13] for approximating the income distribution tail, and the Pareto I shape $\tilde{\alpha}$ has a one-by-one mapping to b: $\tilde{\alpha} = b/(b-1)$. However, the GPD used in this paper's approach is a three-parameter distribution, and the inverted Pareto coefficient b is a function of all its parameters: $b = 1 + \sigma/[\tau(1-\xi)]$. For simplicity, a standard formula communicating GPD's shape with a Pareto I shape is used, $\xi = 1/\tilde{\alpha}$ relying on $\xi \approx (b-1)/b$.[14] This offers an opportunity to assume that the most likely values of ξ are ranges between 0.33 and 0.67.[15] Likewise, low income inequality in these settings corresponds to the GPD with shapes that lie approximately between 0.33 and 0.44, and high inequality with shapes near or larger than 0.6. To specify the parameters of the selected beta penalty function, I assume, roughly, that 90% of the GPD shape parameters measuring the income distribution tail are $\xi \in [0.3; 0.7]$. Therefore, a guess on α and β for a penalty in eq. (4) is formulated by solving a system of two equations with two unknowns: 5^{th} percentile of beta from eq. (4) is equal to 0.3, and 95^{th} percentile is equal to 0.7. The solution of this system implies a chosen penalty function to be a beta distribution with $\hat{\alpha} = 8$ and $\hat{\beta} = 8$.[16] Figure 2 plot different $\log(beta)$ penalties against the

[13]The Pareto I distribution has PDF $f(x) = \frac{\tilde{\alpha} x_m^{\tilde{\alpha}}}{x^{\tilde{\alpha}+1}}$, where x_m is a location parameter and $\tilde{\alpha}$ is a scale. The Pareto I and GPD models are equivalent when $\xi = \frac{1}{\tilde{\alpha}}$, $\sigma = \frac{x_m}{\tilde{\alpha}}$, and $\tau = x_m$.

[14]Approximating the link between b and ξ via $\tilde{\alpha}$ does not imply assuming a Pareto I tail distribution, but only that the tail thickness is comparable to that implied by a Pareto I model. This assumption is sufficient to establish a one-to-one mapping between ξ and b, while avoiding the more restrictive implication of a Pareto I specification and preserving the use of the GPD for all subsequent computations.

[15]Because $(1.5-1)/1.5 = 0.33$ and $(3-1)/3 = 0.67$.

[16]Martins and Stedinger (2001) for penalizing likelihood on the GPD shape estimates from hydrological data using a beta distribution with $\alpha = 6$ and $\beta = 9$.

GPD shape ξ. The choice of $a = 8$ is further supported by a penalty sensitivity analysis reported in Appendix B.3, in which a is varied across a wide grid from 1 to 50. The results show that $a = 1$ — corresponding to a uniform Beta$(1, 1)$ penalty and thus equivalent to MLE — yields the highest variance, and that variance declines rapidly as a increases. The MSE of the tail Gini index $\hat{G}_R$ (introduced formally in Section 2.3) reaches its minimum in the neighborhood of $a = 8$ for the most challenging heavy-tail case of $\xi = 0.7$, and remains near-minimal for all other values of ξ, confirming that the baseline penalty strikes the best attainable bias–variance trade-off across the empirically relevant range of tail thickness.

2.3 Computation of the Gini index

After estimating the income distribution parameters to measure the Gini index, I employ a common decomposition formula (Dagum 1998; Atkinson 2007; Alvaredo 2011; Jenkins 2017):

$$Gini = P_R S_R G_R + P_N S_N G_N + G_B, \tag{5}$$

where P_R is the share of the "Rich" ($P_R = 1 - p_\tau$, in our notations), P_N is the proportion of the "Non-rich" ($P_N = p_\tau$); $S_R = P_R\mu_R/\mu$ and $S_N = P_N\mu_N/\mu$ are the shares of the total income in each group; μ_R and μ_N are the group mean incomes, and $\mu = P_R\mu_R + P_N\mu_N$ is the overall mean; G_R and G_N are Gini indices in each group, and $G_B = S_R - P_R$ refers to a between-group inequality. The mean income and Gini index for the bulk (μ_N and G_N) are estimated from the kernel density estimates and the grid by applying gini.den[17] function in R. The mean income and Gini index for the GPD tail (μ_R and G_R) are derived analytically from the estimated GPD density function. The GPD mean could be estimated from parameters by a standard formula: $\mu_R = \tau + \sigma/(1 - \xi)$. Analogous to Arnold (2015, 2008), I derive a Gini index formula for the three-parameter GPD density.[18] In current notations, the result will be called Gini for rich G_R and it is defined by:

$$G_R = 1 - 2\int_0^1 L(p)dp = 1 - \frac{(\xi - 1)(\sigma - \tau(\xi - 2))}{(\xi - 2)(\sigma - \tau(\xi - 1))}, \tag{6}$$

where $L(p)$ is a Lorenz curve function.[19]

I use a standard bootstrap bias-corrected estimator (Cameron and Trivedi 2005, p. 365) and 95% basic (reverse percentile) bootstrap confidence intervals (Davison and Hinkley 1997, p. 194) for the Gini index and all other income distribution parameters reported.

2.4 Simulation study

To evaluate the finite-sample performance of the proposed MPLE estimator under the Beta$(8, 8)$ penalty, I conduct Monte Carlo simulations. Specifically, I generate 5,000 samples from the GPD with fixed location parameter $\tau = 1$, scale parameter $\sigma = 0.5$, and shape parameters $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. Without loss of generality, the location and scale parameters are held constant, while the shape parameter is varied, as it has the greatest influence on inequality measures derived from the GPD. For each simulated sample, GPD parameters are estimated using the proposed MPLE and three benchmark methods: (a) the standard maximum likelihood estimator (MLE); (b) the optimal bias-robust estimator (OBRE); and (c) a fully Bayesian approach based on Markov chain Monte Carlo (MCMC) Gibbs sampling, employing the same Beta$(8, 8)$ prior for the shape parameter.

As described at the beginning of Section 2, the key idea underlying the OBRE is to limit the influence of extreme observations by imposing an upper bound k on the influence function. This bound must be specified *a priori*. Larger values of k yield higher efficiency, while smaller values increase robustness (Victoria-Feser and Ronchetti 1994). In the simulation study, I implement the OBRE with $k = 3$, which Victoria-Feser and Ronchetti (1994) show to be optimal for fitting Pareto-type distributions, and with $k = 4$ as a more efficient, closely related alternative.

While Bayesian MCMC methods provide richer uncertainty quantification through full posterior distributions in applied settings, the proposed MPLE yields point estimates with approximate standard errors. The present simulation study, however, focuses on point-estimation accuracy,

[17] R package 'acid' - https://cran.r-project.org/web/packages/acid/acid.pdf.

[18] Appendix B provides a complete derivation of Equation (6).

[19] Lorenz curve is derived by a standard formula from the GPD quantile function $L(p) = \frac{\int_0^p Q(t)dt}{\int_0^1 Q(t)dt}$.

given that the true parameter values are known by construction. Accordingly, the inclusion of MCMC is intended solely to assess its relative performance in generating point estimates from a Bayesian approach, compared with the proposed simplified MPLE within a frequentist framework. In empirical applications, MCMC offers a more comprehensive treatment of uncertainty, and its computational costs should therefore be weighed against the value of full posterior inference relative to point estimates with confidence intervals.

Following the extreme value literature, I assess goodness-of-fit by comparing estimates of the GPD shape parameter $\hat{\xi}$ and the 99$^{\text{th}}$ percentile $\hat{q}_{0.99}$ (Coles and Dixon 1999; Brazauskas and Kleefeld 2009). To facilitate interpretation, I present the main simulation results in terms of the Gini index $\hat{G}_R$, with full supporting results for $\hat{\xi}$ and $\hat{q}_{0.99}$ provided in Appendix C.[20] The MLE and OBRE are implemented using fit.gpd[21] function, while the MCMC Gibbs sampler is implemented using bugs.[22]

In this simulation study, I illustrate the performance of the proposed MPLE estimator in small samples across four scenarios reflecting the most common income survey issues classified at the beginning of Section 2: (A) data contamination; (B) underreporting; (C) under-sampling; and (D) sparse upper-tail data.

To simulate *data contamination* (A), I randomly draw samples of size $n = 100$ from a GPD and introduce contamination in two ways: (1) multiplying the largest observation in each sample by 10, following Cowell and Flachaire (2007); and (2) contaminating 15% of observations with a GPD having a larger shape parameter ($\xi = 1.5$), following Brazauskas and Kleefeld (2009). To simulate *underreporting* (B), after drawing from the baseline GPD, I: (1) randomly halve 50% of observations; and (2) apply the same intervention non-randomly. In the non-random case, observations are selected with a U-shaped probability over the normalized support, $p(x) = 4(x - 0.5)^2$, so that extreme values in both tails are more likely to have their incomes halved. To simulate *under-sampling* (C), I repeat the same random and non-random selection procedures as in case (B), but instead of halving the values, I drop 25% of observations. Finally, to simulate *sparse upper-tail data* (D), I reduce the sample size to $n = 50$ and randomly draw observations from the GPD without further intervention. This scenario reflects the main motivation of this paper: mitigating the impact of data sparsity in small samples, assuming all observations are correctly measured but limited in number, as is often the case in CHFS survey.

Figure 3 presents Gini index estimates and Monte Carlo confidence intervals across six scenarios at $n = 100$: the baseline, two data contamination scenarios, two underreporting scenarios, and two under-sampling scenarios. The sparse upper-tail scenario (D), which reduces the sample size to $n = 50$, is illustrated together with results for $n \in \{100, 500, 5000\}$ in Figure 4. Across all panels of Figure 3, MPLE and its Bayesian analogue, the MCMC Gibbs sampler, produce noticeably narrower confidence intervals compared with MLE and OBRE, reflecting their substantially lower variance. The detailed bias, variance, and MSE decomposition for $\hat{G}_R$ is reported in Tables C1, C2 (Appendix C) and in Figures C1–C6 therein.

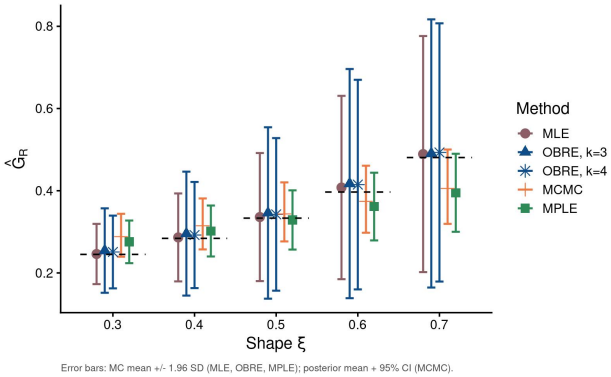
In the baseline scenario (panel a, Figure C2), MPLE and MCMC reduce variance on $\hat{G}_R$ by a factor of roughly 6 to 11 relative to MLE at $\xi = 0.5$ to $\xi = 0.7$, and by a factor of 10 to 14 relative to OBRE. This variance reduction comes at the cost of a moderate, predictable bias: at $\xi = 0.3$ MPLE slightly overestimates the Gini (bias = 0.030), and at $\xi = 0.7$ it underestimates it (bias = -0.086), reflecting the bell-shaped Beta$(8, 8)$ penalty centered at $\xi = 0.5$ (panel a, Figure C1). For moderate values of $\xi \in [0.4, 0.6]$, MPLE Gini bias is at most 0.018 in absolute value, which is small relative to the level of the Gini index itself. The key result is that MPLE achieves the lowest MSE on $\hat{G}_R$ across all values of ξ in the baseline, ranging from roughly one-third to one-half of the MLE's MSE (panel a, Figure C3). The findings are consistent across scenarios and are further corroborated by the results for $\hat{\xi}$ and $\hat{q}_{0.99}$ reported in Tables C3–C6, which show the same pattern of variance reduction at the cost of moderate bias, with MPLE and MCMC uniformly outperforming MLE and OBRE on MSE across all scenarios. Notably, the variance advantage of MPLE is most dramatic for $\hat{q}_{0.99}$, where MLE variance at $\xi = 0.7$ exceeds that of MPLE by a factor of more than ten, underscoring how severely heavy-tailed distributions destabilize likelihood-based tail quantile estimation in small samples.

As expected, MLE exhibits the largest Gini bias under simulated data contamination (panels b and c, Figure C1). In the Brazauskas scenario (panel c), MLE Gini bias reaches 0.167 at $\xi = 0.3$, compared with 0.110 for MPLE and 0.122 for MCMC. In the Cowell scenario (panel b), MLE

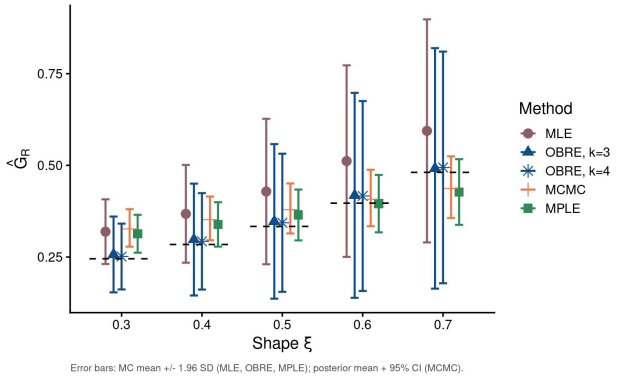
[20] I report the Gini index $\hat{G}_R$ in the main text as a primary expositional choice. While $\hat{G}_R$ is a deterministic function of the GPD parameters (Equation 6), its standardized 0–1 scale is more interpretable in the context of income inequality measurement, allowing for a clearer assessment of the bias–variance trade-off.

[21] R package 'mev' — https://cran.r-project.org/web/packages/mev/mev.jpg.

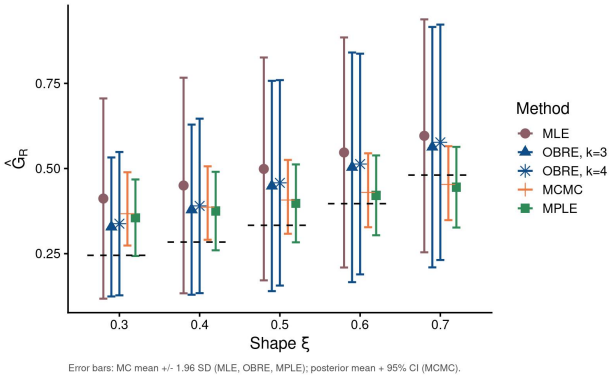
[22] R package 'R2OpenBUGS' — https://cran.r-project.org/web/packages/R2OpenBUGS/R2OpenBUGS.jpg.



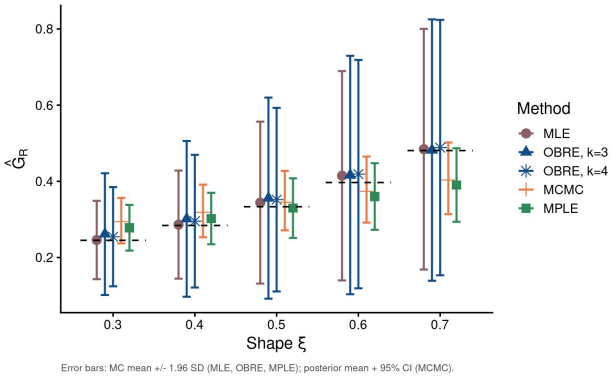
(a) Baseline



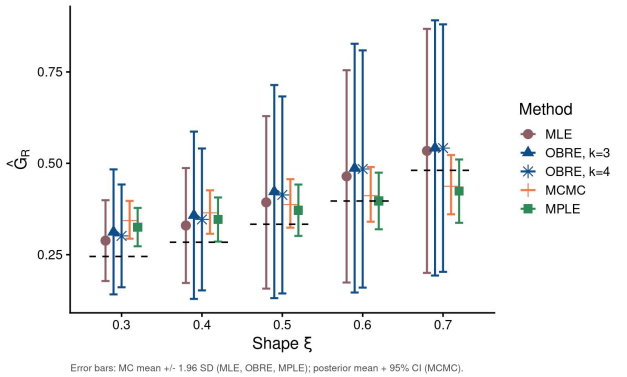
(b) Largest value $\times 10$



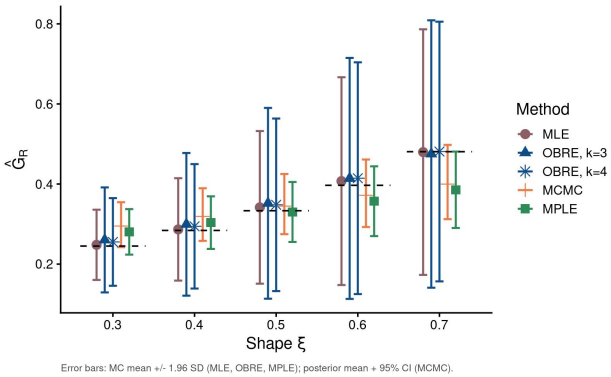
(c) Draw 15% from GPD($\xi = 1.5$)



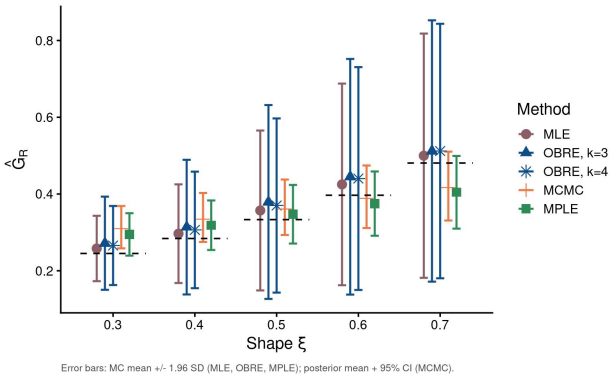
(d) Divide 50% by 2, randomly



(e) Divide 50% by 2, non-randomly



(f) Drop 25%, randomly



(g) Drop 25%, non-randomly

Fig. 3: Gini index $\hat{G}_R$ estimates from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations across replications. For MCMC, points represent the average posterior mean across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ prior Beta(8, 8) and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2). Sample size: $n = 100$.

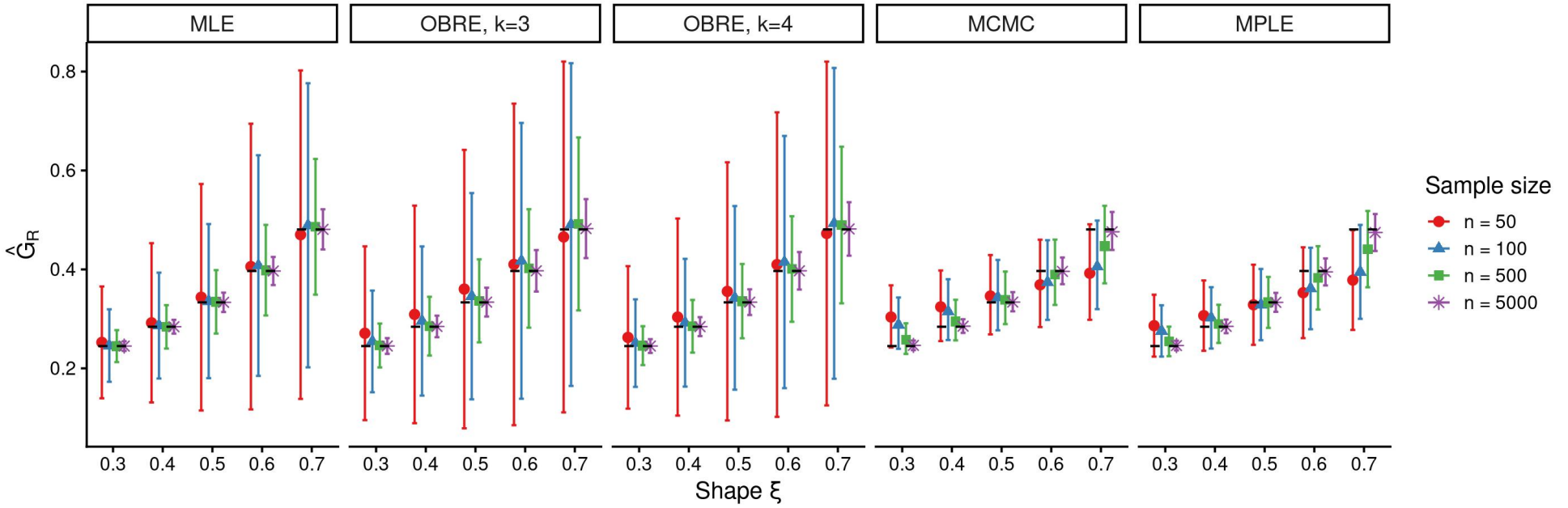


Fig. 4: Gini index $\hat{G}_R$ estimates across sample sizes from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations. For MCMC, points represent posterior means averaged across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ prior Beta$(8, 8)$ and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2).

Gini bias reaches 0.115 at $\xi = 0.6$, while MPLE bias is effectively zero (-0.001). OBRE achieves the smallest Gini bias in the Cowell scenario, as it is specifically designed to downweight a single extreme outlier, but MPLE and MCMC outperform OBRE in the Brazauskas scenario where contamination is distributed across 15% of observations rather than concentrated in one. OBRE produces the largest Gini bias in all other scenarios, including underreporting and under-sampling, because it incorrectly classifies sparse but valid upper-tail observations as contamination and downweights them. For random underreporting (Halve 50% randomly), MLE exhibits the smallest Gini bias, whereas under non-random underreporting (Halve 50% non-randomly), MPLE and MCMC outperform MLE for $\xi > 0.5$, with MPLE bias of 0.000 at $\xi = 0.6$ compared with MLE bias of 0.068 (panels d and e, Figure C1). A similar pattern holds in the random and non-random under-sampling scenarios (panels f and g, Figure C1). The bias patterns for $\hat{\xi}$ across all scenarios (Table C3, Appendix C) closely mirror those for $\hat{G}_R$, as expected given that $\hat{G}_R$ is a monotone function of $\hat{\xi}$. One notable difference arises in the Cowell contamination scenario, where OBRE achieves near-zero bias on $\hat{\xi}$ — confirming its design optimality for single extreme outliers — while MPLE and MCMC absorb a moderate upward bias through prior shrinkage. Despite this, MPLE still achieves lower MSE on $\hat{\xi}$ than MLE in this scenario due to its much smaller variance.

In the very small sample ($n = 50$, sparse upper-tail scenario), MPLE and MCMC again produce a positive Gini bias for $\xi < 0.5$ and a negative bias for $\xi > 0.5$, reflecting the shrinkage induced by the Beta$(8, 8)$ penalty (Figure C4). Despite the somewhat larger bias at $\xi = 0.7$ (MPLE bias $= -0.102$), the MSE of MPLE remains well below that of MLE and OBRE, because variance is reduced so dramatically: MPLE Gini variance is 0.003 compared with 0.029 for MLE at $\xi = 0.7$, a tenfold reduction (Figure C6). The evolution of estimation uncertainty across all sample sizes $n \in \{50, 100, 500, 5000\}$ is illustrated in Figure 4, which shows that the confidence intervals of MLE and OBRE remain wide even at $n = 500$ for large ξ, whereas those of MPLE and MCMC shrink rapidly and are already narrow at $n = 100$. This highlights the central trade-off of the paper. In a Monte Carlo study, bias and variance are evaluated by averaging over thousands of hypothetical repeated samples. In practice, however, a researcher observes only a single dataset — one draw from the distribution. In this single-draw reality, an estimator with lower variance is considerably more reliable: it produces estimates that are closer to the truth more consistently, even if it accepts a moderate and predictable bias. This is precisely what MPLE is designed to deliver.

Results from a large sample ($n = 5,000$) show that Gini bias and MSE become negligible for all estimators. The MSE of MPLE converges to that of MLE, confirming that the relative contribution of the penalty term $\log P(\xi)$ to the objective function becomes asymptotically negligible as the data log-likelihood grows linearly with n_τ. OBRE exhibits slightly greater variance and reduced efficiency even at large samples, as seen in Table C2 (Appendix C). Results for $n = 500$ confirm that the efficiency advantage of MPLE generally holds for moderately sized samples as well. This convergence pattern is confirmed for $\hat{\xi}$ and $\hat{q}_{0.99}$ in Tables C4 and C6 (Appendix C), where MPLE bias and MSE similarly shrink toward those of MLE as n grows, while OBRE retains slightly elevated variance even at $n = 5{,}000$.

Overall, the simulation results suggest that, assuming the penalty function is specified correctly—implying that in 90% of cases the GPD shape parameters of the income distribution tail lie between 0.3 and 0.7, with the distribution of possible shapes following Beta$(8, 8)$—the proposed MPLE estimator offers substantial advantages over commonly used parameter estimators. First, MPLE is more efficient than both MLE and OBRE in small samples, achieving substantially lower MSE on $\hat{G}_R$ across all scenarios. The efficiency gain stems from variance reduction: the penalty prevents the estimator from being destabilized by sparse but highly influential extreme observations that heavy-tailed distributions naturally produce (Davidson and Flachaire 2007), at the cost of a moderate and predictable bias.[23] Importantly, the direction of this bias is conservative with respect to inequality measurement: for small values of ξ, MPLE slightly overestimates the Gini, but the bias is small in magnitude; for large values of ξ, where the true inequality is high, MPLE shrinks estimates downward toward the center of the Beta$(8, 8)$ penalty, producing more conservative Gini estimates. This asymmetry is a direct consequence of the bell-shaped penalty centered at $\xi = 0.5$ and may be viewed as a desirable property in applied work, where overstating inequality is generally considered the more consequential error. Second, unlike OBRE — which is specifically designed to limit the influence of contaminated observations but underperforms relative to MPLE in all non-contamination scenarios considered — MPLE delivers consistent efficiency gains across contamination, underreporting, and under-sampling scenarios alike. Third, MPLE achieves numerically equivalent point estimation performance to the MCMC Gibbs sampler with the same

[23] The issue of large standard errors in many existing empirical studies on inequality is discussed extensively in Maasoumi (1999).

Beta$(8, 8)$ prior at every sample size from $n = 50$ to $n = 5{,}000$ (Table C2), at a substantially lower computational cost. Finally, in large samples the penalty term $\log P(\xi)$ becomes asymptotically negligible relative to the data log-likelihood, so that MPLE converges to MLE and recovers full asymptotic efficiency, while OBRE retains slightly elevated variance even at $n = 5{,}000$.

3 Data

The data for my baseline inequality estimates comes from the four waves of the China Household Finance Survey (CHFS 2013, 2015, 2017, and 2019). The CHFS is a survey conducted by the Research Institute of Economics and Management at the Southwestern University of Finance and Economics in Chengdu, China. The CHFS covers 29 provinces, excluding Hong Kong, Macao, Taiwan, and the Tibet and Xinjiang Autonomous Regions, with sample sizes ranging from approximately 30,000 to 40,000 households across survey waves. Appendix D presents a more detailed description of the CHFS sample. The CHFS is a panel study that contains micro-level information on key household demographic and financial characteristics. The survey employs a stratified multi-stage random sampling procedure and is representative at national and provincial levels.[24]

Together with the baseline inequality estimates, I produce supplementary comparative calculations by utilizing four waves of the China Family Panel Studies (CFPS 2012, 2014, 2016, and 2018). The CFPS is a panel survey conducted by the Institute of Social Science Survey at Peking University. The baseline sample covers 25 mainland provinces (excluding Xinjiang, Tibet, Qinghai, Inner Mongolia, Ningxia, and Hainan) and represents around 95% of the Chinese population. The CFPS employs a three-stage stratified probability proportion to size sampling procedure. The target sample size is about 16,000 households. In the four above-mentioned waves, five provinces (Liaoning, Henan, Shanghai, Guangdong, and Gansu) were oversampled to be representative at the provincial level. I use the data on these five provinces both from CFPS and CHFS, to demonstrate the performance of the proposed estimation method in smaller, compared to national-level, samples.[25]

In the current study, to measuring inequality I use the annual household disposable income per capita.[26] Household disposable income is defined in CHFS in accordance with the NBS standards. It includes the income generated from labor, business, assets, transfers, and other sources over the past twelve months.[27] In detail, CHFS household disposable income comprises: (a) labor income includes total after-tax monetary wages and salaries, bonuses, and subsidies from employer (both monetary and in-kind); (b) business income is the sum of net income from agricultural production and business operations; (c) asset income includes net after-tax income from non-financial assets (rental income from lands, properties, and vehicles) and income from financial assets (interests on deposits, stocks, bonds, dividends from commercial life insurances, etc.); (d) transferred and other income accounts include retirement benefits and pensions; unemployment insurance; other government compensation and subsidies; private transfers; net after-tax income from all other sources including lotteries, sale of properties, vehicles, intellectual property rights, etc.

In the CFPS, household incomes consist of five main categories: wages, business income, transfers, property income, and other income. Wage income is the after-tax wage combined with bonuses and non-cash benefits from agricultural or non-agricultural employment. Business income covers the net income from family production in farming, forestry, fishing, etc.; the value of agricultural production consumed by the family; and net profits from self-employment or operating private enterprises. Transfer income includes various kinds of government transfer payments (e.g., pensions, subsidies, and alms) and social donations. Property income is the income from renting land, real estate, and other means of production. Other income represents financial support from families or friends as well as any type of gifts.

Table D4 in Appendix D provides a detailed comparison of income categories in the CHFS and CFPS. Labor and business income are broadly comparable across the two surveys, capturing post-tax wages, bonuses, and non-cash work-related benefits, as well as net income from agricultural and non-agricultural enterprises, respectively. The largest divergence arises in the definition of

[24] For more details on the CHFS data, please refer to https://chfs.swufe.edu.cn/sjzx/sjsq.htm, and Gan et al. (2014).

[25] For more details on the CFPS data, see https://www.isss.pku.edu.cn/cfps/en.

[26] The Gini coefficients reported in this paper are based on nominal income data. For this descriptive study of income inequality in China, spatial price adjustments are unnecessary, as applying a uniform deflator within a geographic unit is a linear transformation that does not affect the Gini coefficient (Allison 1978). Nonetheless, a substantial body of literature emphasizes that a full understanding of inequality in China requires accounting for spatial price differences, as the cost of living is substantially higher in urban and coastal areas (Brandt and Holz 2006; Sicular et al. 2007; Ravallion and Chen 2009). Consequently, while the Gini indices within geographic units reported here accurately capture income inequality, comparisons of mean incomes across units should be interpreted with caution.

[27] Therefore, I adjust the obtained estimates to capture the year preceding the survey year. That is, from CHFS 2013 I obtain the Gini index estimates for the 2012 year, from CHFS 2014 - for the 2013 year, and so on.

asset income: while the CFPS primarily records rental income, the CHFS additionally includes a wide range of financial income components, such as returns from deposits, stocks, funds, and bonds. The CHFS also offers more comprehensive coverage of transfer income, including both government and private transfers. In contrast, private transfers—such as financial support from family and friends—are classified under the "other income" category in the CFPS rather than under transfers. Finally, "other income" in the CHFS captures irregular income sources, such as lottery winnings and income from the sale of various types of property (e.g., vehicles), whereas these categories are not recorded in the CFPS.

The descriptive statistics of the annual household income per capita from CHFS and CFPS are presented in Tables D2 and D3 (Appendix D), respectively (in both cases, survey weights were applied). In the current study, zero and negative household income observations arise only in the CHFS, as the income measure used in the CFPS (*fincome1_per*, net family income per capita) is strictly positive. I exclude zero and negative CHFS income observations to mitigate potential measurement issues. This decision is motivated by three considerations. First, in the CHFS, missing income information is sometimes recorded as zero, making it difficult to distinguish genuine zero incomes from missing values. Second, zero and negative income observations represent only 1–2.5% of the CHFS sample, suggesting that their impact on the estimates is likely to be limited. Third, following Van Kerm (2007), I conduct robustness checks using alternative treatments of zero and negative incomes, including winsorization (replacing values with 10% of the median income) and imputation by randomly drawing values from the lowest income decile. As shown in Section 4.4 and in the Appendix I.2, these alternative treatments do not yield substantial differences in the estimated Gini coefficients.

4 Results

4.1 National-level income inequality

The resulting national-level income distribution parameters and Gini estimates are summarized in Table F1 (Appendix F).[28] The baseline results from CHFS suggest that Gini index in China decreased from 0.616 in 2012 to 0.581 in 2016, and rose back to 0.590 in 2018. Figure 5 compares the results with the official NBS,[29] the World Bank[30] estimates and the results from the previous literature. Notably, the Gini index reported by the NBS is substantially lower than the current results but exhibits similar dynamics of a subsequent fall then rise corresponding to the same years. Our results, which employ a new low-variance estimator, are also higher than most of the existing estimates based on household income surveys which do not account for the top-income's underrepresentation. However, baseline results from 2012 (0.616) and 2014 (0.604) are consistent with the inequality levels reported in previous studies employing the same CHFS data: Xie and Zhou (2014) the estimated Gini index equaled 0.611 in 2010, and Gan et al. (2014) obtained 0.610 for 2012. The only study that reported higher inequality levels is Li et al. (2020). They estimated the Gini index in 2015 to be equal to 0.646 based on CHIP 2016 and the "Top Incomes in China" (TIC) database, which they compiled from multiple online sources: "Hurun Rich List", "Forbes Rich List", "Online Celebrity", payments of CEOs of listed companies, etc.

Supplementary results from CFPS on Gini equal to 0.540 in 2011 coincide with multiple results from Xie and Zhou (2014) estimated using the CFPS (0.532), CGSS (0.539), and CLDS (0.536). The results from the two latest waves in CHFS and CFPS are converging. This tendency might be attributed to the improvement in representing rich households in CFPS. Notably, from Figure G1 (Appendix G), it can be seen that the upward trend in inequality based on CFPS is mostly driven by the top 5% of observations. However, there is insufficient evidence to conclude that top incomes are better represented in the CFPS.Therefore, the main conclusions regarding inequality levels and trends are drawn from the baseline estimates based on the CHFS data. The raw data weighted Gini estimates are reported in the Appendix G. They are not statistically different from the estimates presented in this section because the sample size for national-level data in both datasets is sufficiently large, and the role of the proposed incorporated penalty in the likelihood function is small.

[28] Figure G2 in Appendix G presents the comparison of the obtained results to the Gini index estimated from the raw weighted CHFS data.

[29] For the 2010 - 2011 Gini indices, NBS used the China Rural Household Survey (CRHS) and China Urban Household Survey (CUHS), and for 2012-2019 they used the China National Integrated Household Survey (CNIHS).

[30] To estimate inequality in China, the World Bank used the grouped (ventile) data on household per capita consumption provided by the NBS. The World Bank's estimates are lower than those of the NBS because they are produced using the household per capita consumption data, as opposed to the per capita disposable income data used by the NBS. An additional factor contributing to the difference might be that the World Bank applies special unified procedures to the data from all countries to facilitate international comparisons.

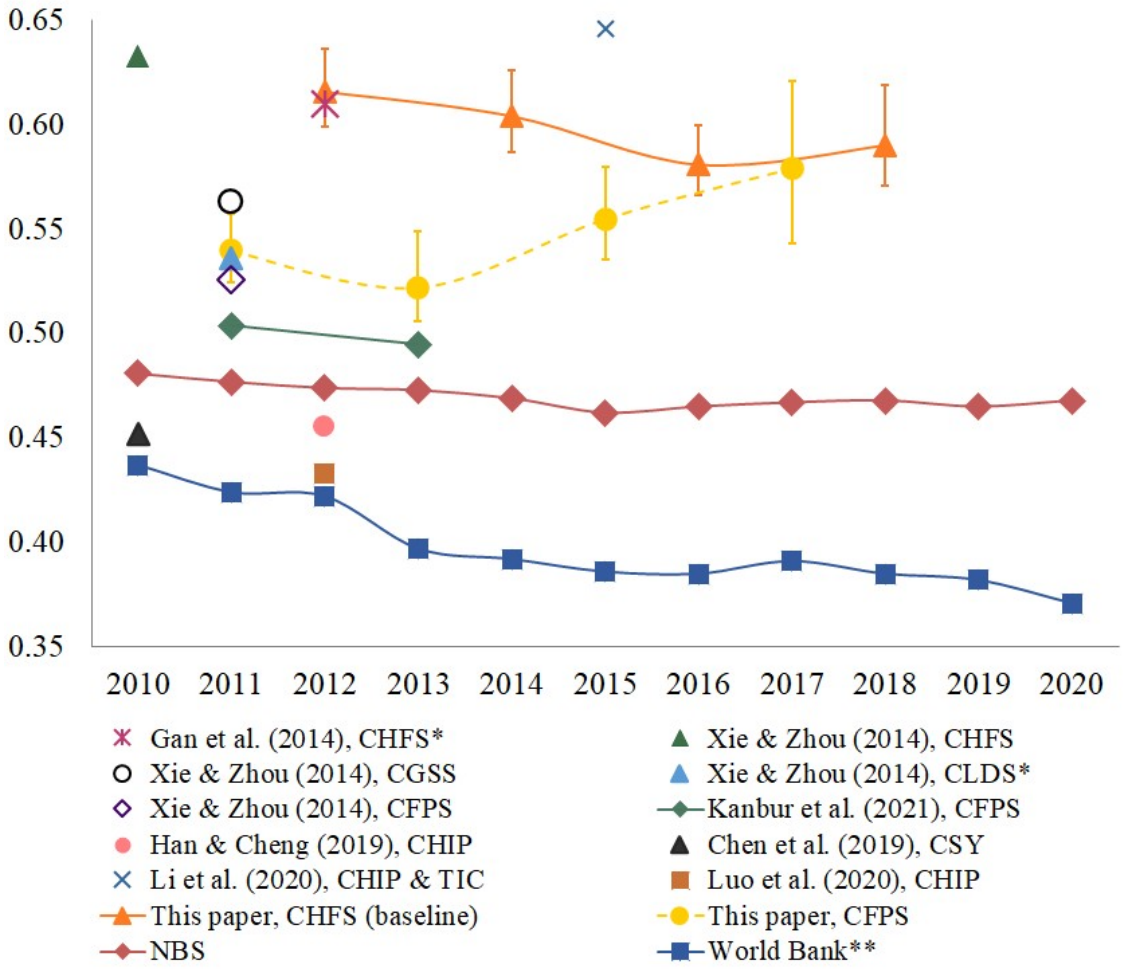
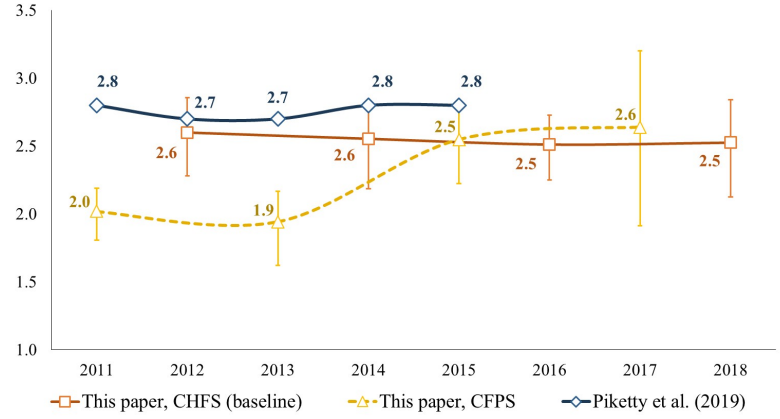


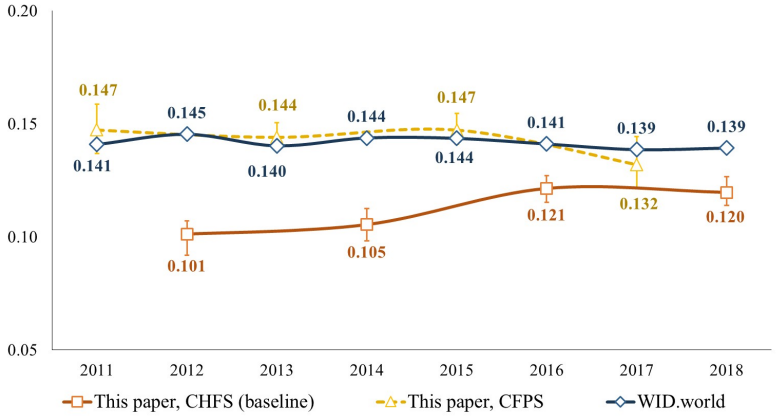
Fig. 5: Gini index estimates in China. The orange solid line refers to the baseline national-level Gini estimates from the CHFS, and the yellow dashed line represent a supplementary set of Gini estimates based on the CFPS data. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated from 1,000 samples. In this figure, the obtained results are placed among the existing Gini estimates from official statistics and previous studies. The corresponding empirical evidence is summarized in Table A1 in the Appendix A. All Gini estimates on this figure produced based on household income per capita, except "*" results produced based on the total household income, and the World Bank's "**" estimates which were assessed based on the per capita consumption data shared by the NBS

In a recent paper Piketty et al. (2019) applied the Pareto-interpolation technique to the combined data from national accounts, household surveys, and tax data, and concluded that there is a significant underestimation of income inequality in China. They do not report the Gini index but provide estimates of several income distribution characteristics. Figure 6 compares my results on estimated income distribution with estimates reported in Piketty et al. (2019). Up to date, Piketty et al. (2019) employed the widest set of data sources studying the dynamics of inequality in China, including national accounts, surveys, wealth rankings, and tax data covering high-income earners. They measured the inverted Pareto coefficient b to be equal to 2.7-2.8 in 2011-2015. As explained in Piketty et al. (2019), the Pareto shape parameter is obtained from the government tax data, in which taxable income is defined as the sum of wage income, business income, and taxable capital income, and doesn't include tax-exempt capital income (which mainly consists of undistributed profits of private corporations and owner-occupied housing rents).[31] Overall, the results, reported in Figure 6, indicate that the distributional characteristics I derived by relying entirely on the data from CHFS are very close to those estimated from tax-data by Piketty et al. (2019). In the latest years of 2015-2018, percentile estimates converge with the supplementary results from the CFPS as well. These facts could serve as an indicator of the credibility of CHFS and late CFPS data in representing top-income observations.

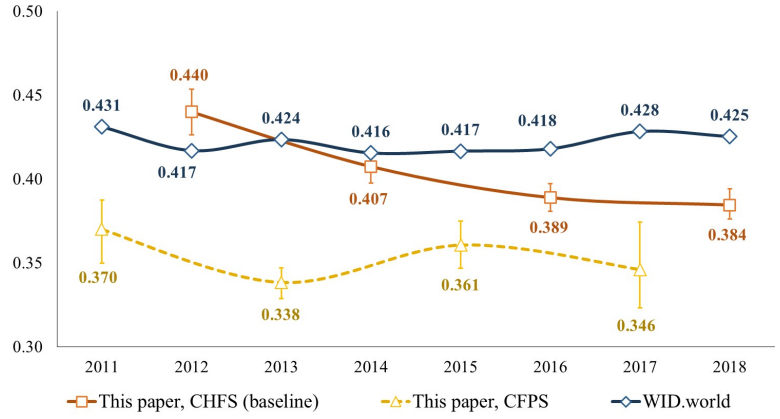
[31] Conditioning on the specific of the data used in Piketty et al. (2019), the current comparison seems to be reasonable because the main components of household income in CHFS (labor, business, and assets income) are also accounted for in the tax data. Tax data does not account for the transfer income. However, it is unlikely that this type of income accounts for a big share of the income of the rich.



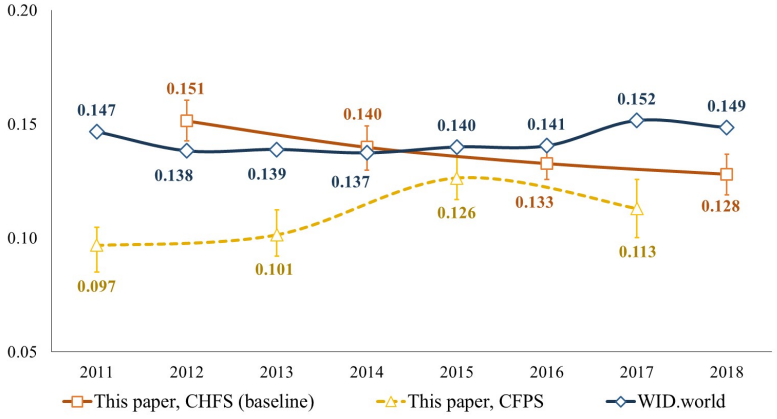
(a) Inverted Pareto coefficient



(b) Bottom 50% income share



(c) Top 10% income share



(d) Top 1% income share

Fig. 6: The comparison of several estimated characteristics of income distribution to the results from Piketty et al. (2019) and the World Inequality Database (WID.world) (2024). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated from 1,000 samples

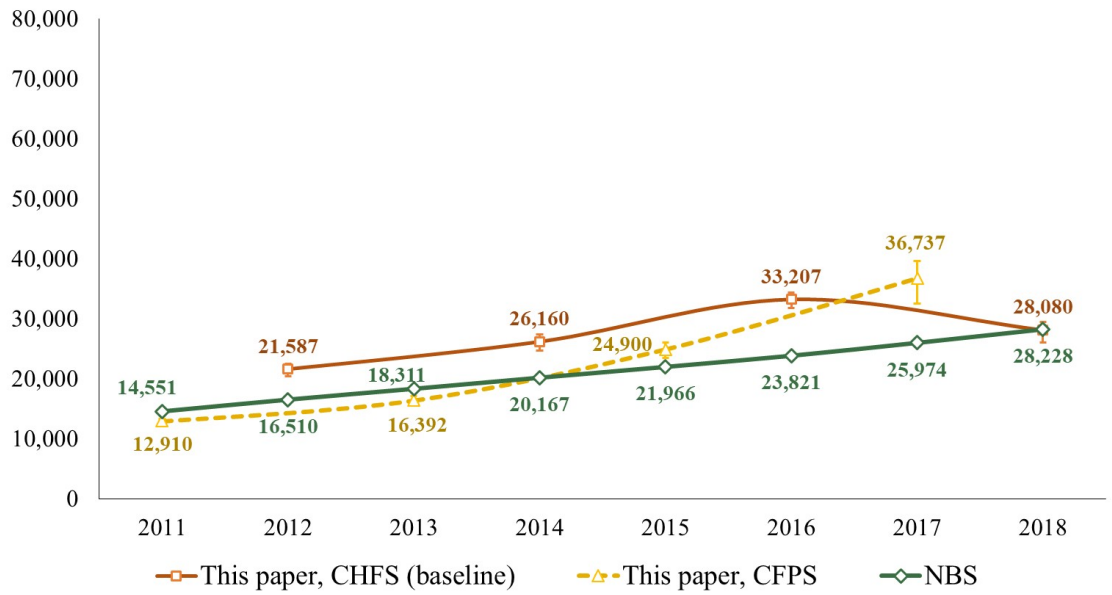


Fig. 7: Comparison of estimated mean household income per capita (CN¥) in China using three data sources: China Household Finance Survey (CHFS), China Family Panel Studies (CFPS), and the National Bureau of Statistics (NBS). For the NBS, values are taken from Per Capita Disposable Income as reported in the China Statistical Yearbook (https://www.stats.gov.cn/english/Statisticaldata/yearbook). The error bars represent the 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

Figure 7 compares the mean household income per capita estimated in the current study to the annual per capita disposable income published in the China Statistical Yearbook. The baseline results from CHFS suggest the increase in mean household income per capita from around 22K CN¥in 2012 to 33K CN¥in 2016, and the subsequent drop to 28K CN¥in 2018. The estimates for 2012-2016 are slightly higher than those released by the NBS but converge to a similar level in 2018. Supplementary estimates from CFPS are slightly lower than the NBS levels in 2011-2013 and larger in 2015-2017.

4.1.1 Theil T decomposition by income source

To assess whether differences in the composition of household income sources in the CHFS and CFPS contribute to the observed disparities in inequality levels, this subsection conducts an additional analysis based on a decomposition of the Theil T index by income source. The Theil T index is defined as

$$T = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{\bar{y}}\ln\left(\frac{y_i}{\bar{y}}\right), \tag{7}$$

where y_i denotes annual household income per capita for household i, $\bar{y}$ is mean household income per capita, and N is the number of households in the data. Unlike the Gini coefficient, the Theil T index is fully decomposable by income source. I apply the decomposition proposed by Shorrocks (1982, 1983):[32]

$$T = \sum_k S_k = \sum_k s_k \cdot T = \sum_k \frac{\sum_i y_{ik}\ln\left(\frac{y_i}{\bar{y}}\right)}{\sum_i y_i \ln\left(\frac{y_i}{\bar{y}}\right)} \cdot T, \tag{8}$$

where S_k denotes the contribution of income source k to overall income inequality, and $s_k = S_k/T$ is the proportional contribution of source k.

As it was discussed in the Section 3, household income is defined slightly different in CHFS and CFPS.[33] Figure 8 present the relative shares of each

According to Figures 8, panels (a) and (b), the decomposition of the national-level Theil T index by income source reveals substantial differences in the primary drivers of income inequality between the CHFS and CFPS survey samples. The largest discrepancy arises from labor and business income. As discussed in Section 3, these two categories are defined similarly across both surveys. However, their contributions to total inequality differ markedly: labor income is the largest

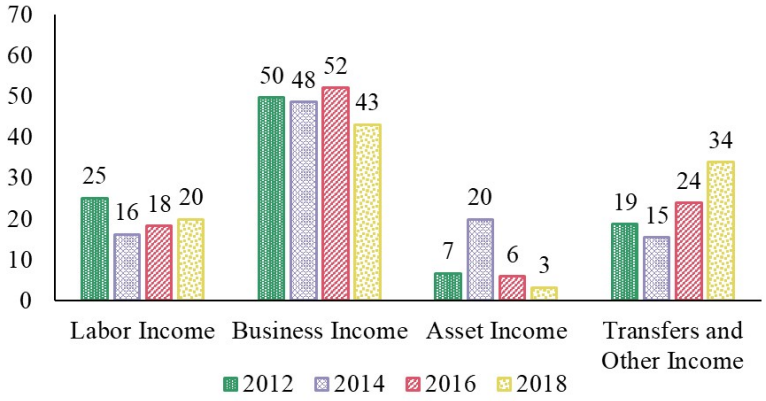
[32] An extended decomposition framework is provided by Jenkins (1995).
[33] Table D4 in Appendix D presents a detailed comparison of income categories across thse two surveys.

contributor in the CFPS, accounting for 49.3% on average across years, whereas business income is the largest contributor in the CHFS, averaging 48.3%. Moreover, business income is typically more dispersed than wages, which increases its contribution to total inequality in the CHFS. Thus, the difference in contributions reflects both the composition of the surveyed populations and the relative inequality within these income types. Differences in the definitions of asset income (with the CHFS including financial income not recorded in the CFPS) and in the aggregated "transfer and other" category (including income from property sales in the CHFS) affect the magnitude of contributions for these categories but do not explain the reversal in the relative importance of labor and business income. Panels (c) and (d) of Figure 8 present the results of the same Theil T decomposition for the population below the 95^{th} percentile. These results provide further evidence that business income in the CHFS is largely concentrated among affluent households, which are particularly well represented in that survey. For both the CFPS and CHFS, labor income remains the dominant contributor to inequality among the bottom 95% of the population. Consequently, the shift toward business income as the largest contributor in the CHFS is primarily driven by top-income observations. This result emphasizes the particular importance of careful treatment of the income distribution's upper tail, a primary focus of the present study.

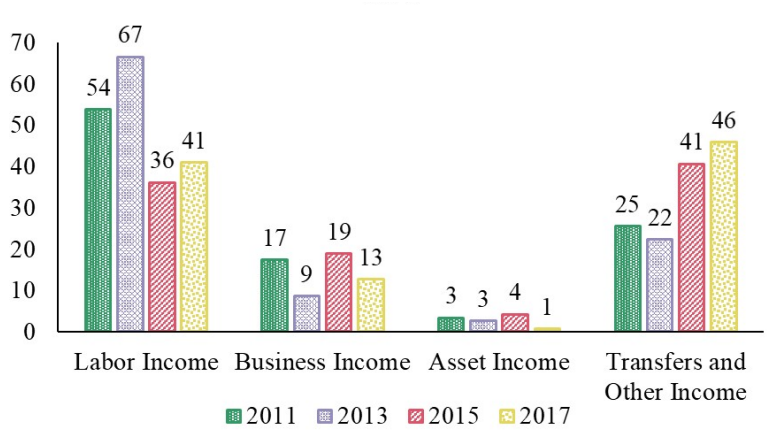
Overall, baseline national-level Gini index estimates for China presented in this paper are higher compared to the NBS-released Gini indices as well as to results from previous non-CHFS-based studies. The interpretation of these inequality estimates requires caution. Gini estimates calculated solely from household surveys are systematically different from those obtained using surveys that oversample wealthy households or incorporate exogenous adjustments for top income values. The former typically yield lower Gini index estimates than the latter. Although the Gini index is statistically less sensitive than entropy measures to measurement error in extreme values (Cowell and Flachaire 2007), its estimated value remains strongly influenced by top-income observations. Moreover, our results from the decomposition of the Theil T index by income source indicate that the higher overall inequality in the CHFS is driven primarily by business income, largely represented by the incomes of affluent households, which are particularly well covered in the CHFS. Labor income, better represented in the CFPS sample, is the largest contributor to overall inequality in that survey, but its lower dispersion results in smaller total inequality compared to the CHFS.

This also affects possibilities for international comparisons, since most estimates for other countries do not account for top incomes. Thus, the results of the current study do not imply that the "true" inequality levels in China are close to those, for example, in South Africa. While in most studies the Gini index is typically evaluated using household survey data, an increasing number of recent economic papers employ top-income adjustments.[34] Comparisons are therefore meaningful only within these two distinct branches of the inequality literature. In Figure 5, these two sets of estimates are clearly seen: (a) Gini values of 0.580 and above are reported in studies paying special attention to top incomes (e.g., Gan et al. (2014), Xie and Zhou (2014), Li et al. (2020), and CHFS-based results from the current study); (b) lower Gini values are reported in studies relying solely on household survey data (including CFPS-based results from the current paper).

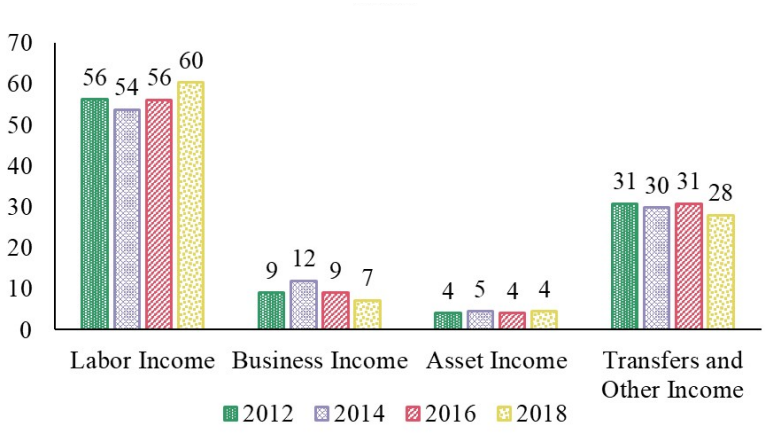
[34] E.g., Novokmet et al. (2018) on inequality in Russia



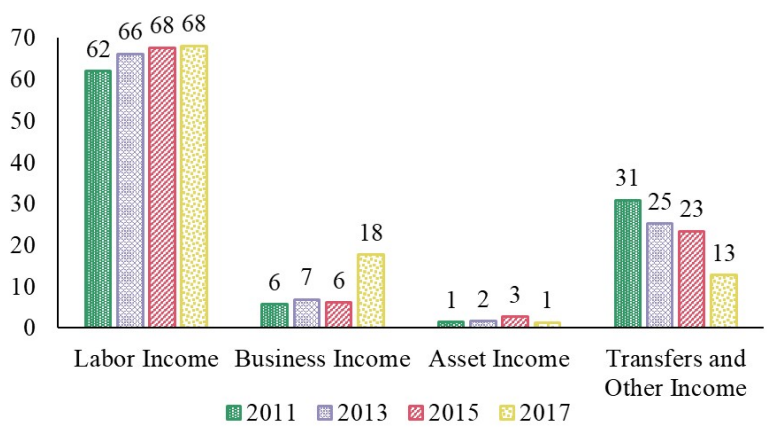
(a) % of Theil T, CHFS (full sample)



(b) % of Theil T, CFPS (full sample)



(c) % of Theil T, CHFS (below the 95^{th} percentile)



(d) % of Theil T, CFPS (below the 95^{th} percentile)

Fig. 8: Decomposition of the Theil T index by income source. The figure displays the proportional contribution s_k of each income source k to total income inequality, as obtained from the Theil T decomposition in equation (8). Transfers and other income are reported together to align with the classification of the transfer variable in the CHFS.

4.2 Regional-level income inequality

In this subsection, I provide a brief summary of the supplementary analysis of regional and rural-urban inequality in China presented in Appendix E. Using data from the CHFS, I document substantial rural-urban and regional disparities: urban households earn significantly more than rural households (2.7–3.4 times), and coastal regions are wealthier than inland regions, with mean household income in the Eastern region approximately 1.5–1.8 times higher than in the Central, Western, and Northeastern regions. These patterns align with prior literature (Wan et al. 2012; Gao et al. 2014; Zhang et al. 2012; Zhang 2021). Contrary to earlier studies reporting higher inequality in rural areas (Zhang et al. 2012; Zhang 2021; Li and Sicular 2014), I find that income inequality across rural and urban households is similar, once confidence intervals are taken into account. This difference likely reflects the more recent period covered in my study and the use of the CHFS survey. Income inequality is higher in the Eastern and Western regions, a pattern broadly consistent with (Luo et al. 2020). Decomposition of the Theil T index indicates that within-group variation—both rural-urban and regional—explains the majority of nationwide income inequality, exceeding 87%, in line with previous literature (Wan et al. 2018; Luo et al. 2020).

4.3 Provincial-level income inequality

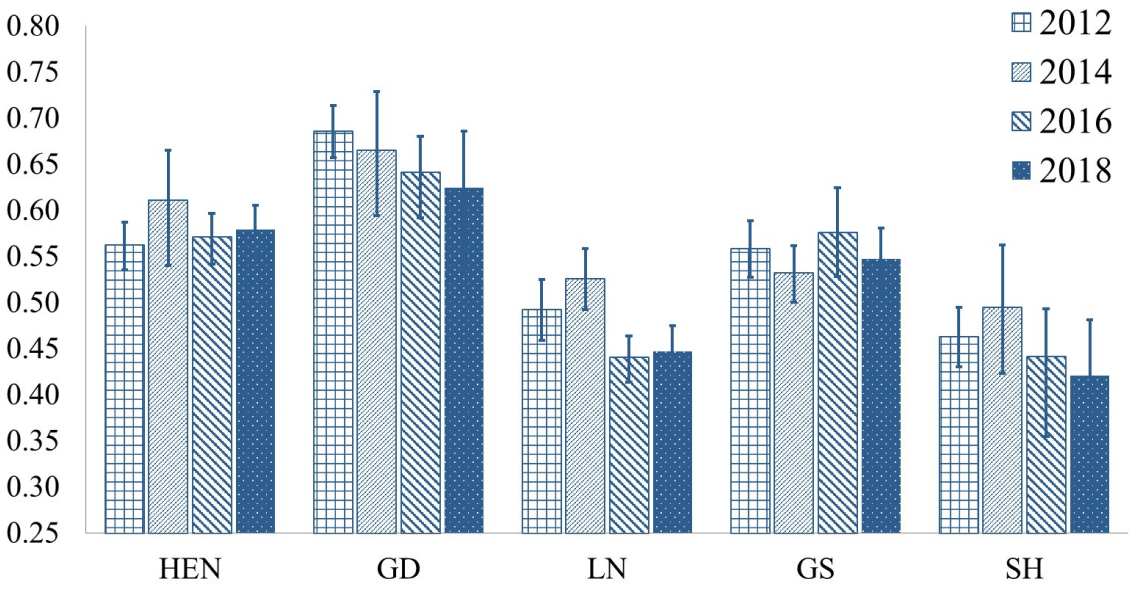


Fig. 9: Provincial-level baseline Gini index estimated from the CHFS data. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

Figures 9 and G6 (in Appendix G) report the results on provincial-level Gini index estimates for five provinces which are representative both in CHFS and CFPS samples: Shanghai, Guangdong, Liaoning, Henan, and Gansu. Sample sizes at the provincial level range from 700 to 3000 observations in the CHFS and from 900 to 1500 observations in the CFPS. Point estimates and confidence intervals on the Gini index and the distributional parameters are reported in Table F4 (Appendix F). Due to the small samples, these results are the most pronounced in terms of the benefits of the estimation procedures introduced in Section 2: the variance of Gini estimates across years dropped by 15% in CHFS and by 37% in CFPS compared to the raw data estimates presented in Table G5 and Table G6 in the Appendix G, respectively. The adjusted inequality estimates are, in most cases, corrected downwards at the provincial-level CFPS and both downwards and upwards in CHFS at different years. The baseline Gini estimates from CHFS suggest that during the period under study, income inequality was the highest in Guangdong province. The Gini index ranges there from 0.62 to 0.69. Henan and Gansu had the second largest level of inequality measured by Gini equal to 0.58 and 0.55, respectively. The Gini index in Liaoning is estimated to be around 0.50. Shanghai is the least unequal among the provinces considered, with a Gini index of around 0.48.

Figure 10 shows that the obtained provincial results on mean household income per capita from CHFS in most of the province/years are higher than per capita disposable income reported by the

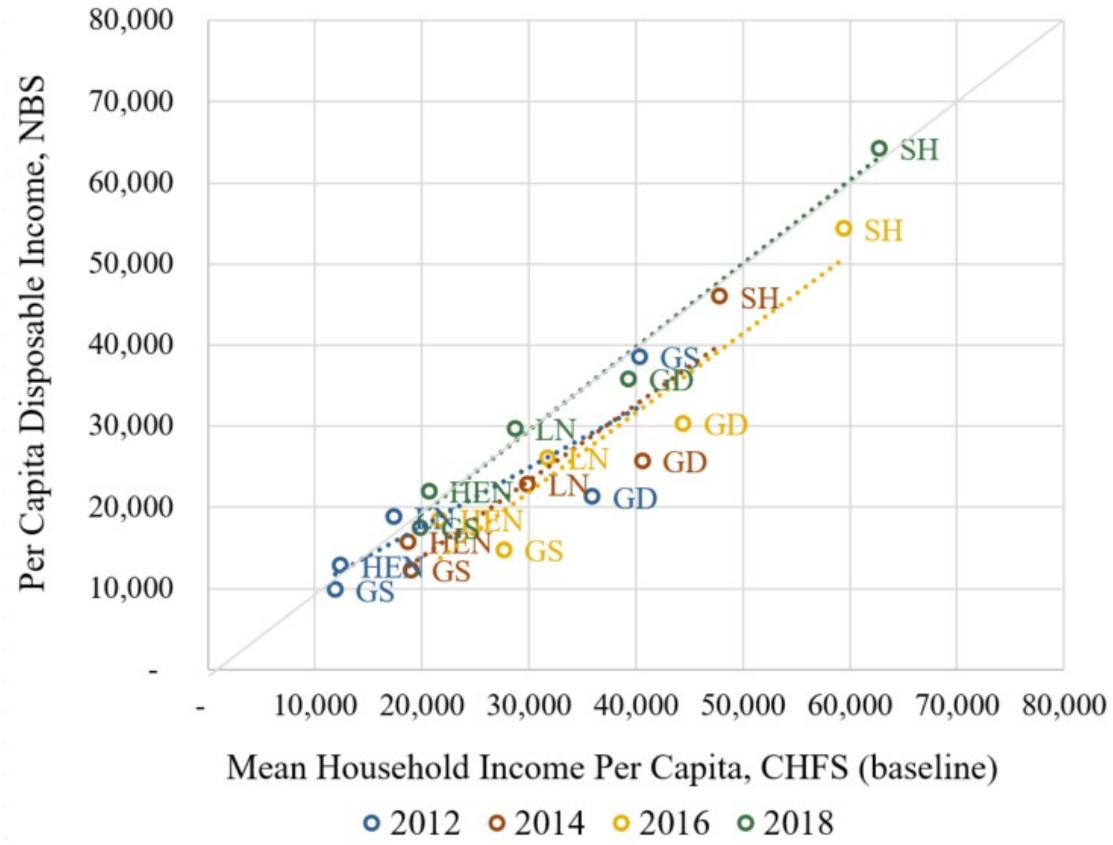


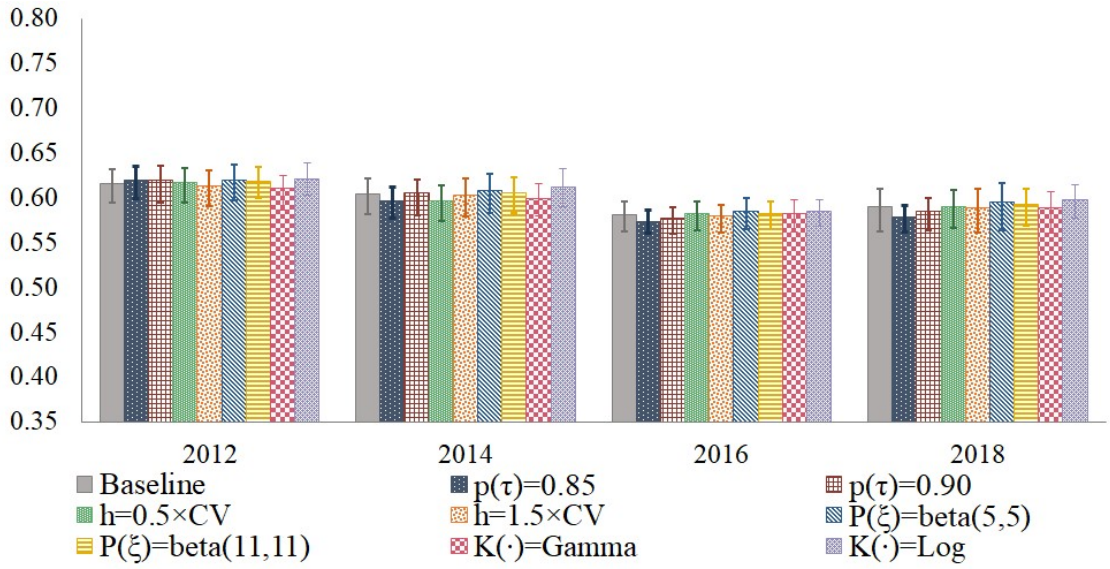
Fig. 10: Baseline estimates of mean household income per capita from the CHFS (X-axis) plotted versus "Per Capita Disposable Income" reported by the NBS (Y-axis) ("China Statistical Yearbook" - https://www.stats.gov.cn/english/Statisticaldata/yearbook)

NBS. Notably, results from the CHFS-2019 wave of the survey for all five provinces are very close to those from the China Statistical Yearbook. The supplementary results from CFPS in 2011 and 2013 are mostly smaller than the per capita disposable income from NBS, and in 2015 and 2017 - larger (Figure F1 in Appendix F). A full set of the results on the provincial mean household income per capita and corresponding NBS per capita disposable income levels are presented in the Appendix H.

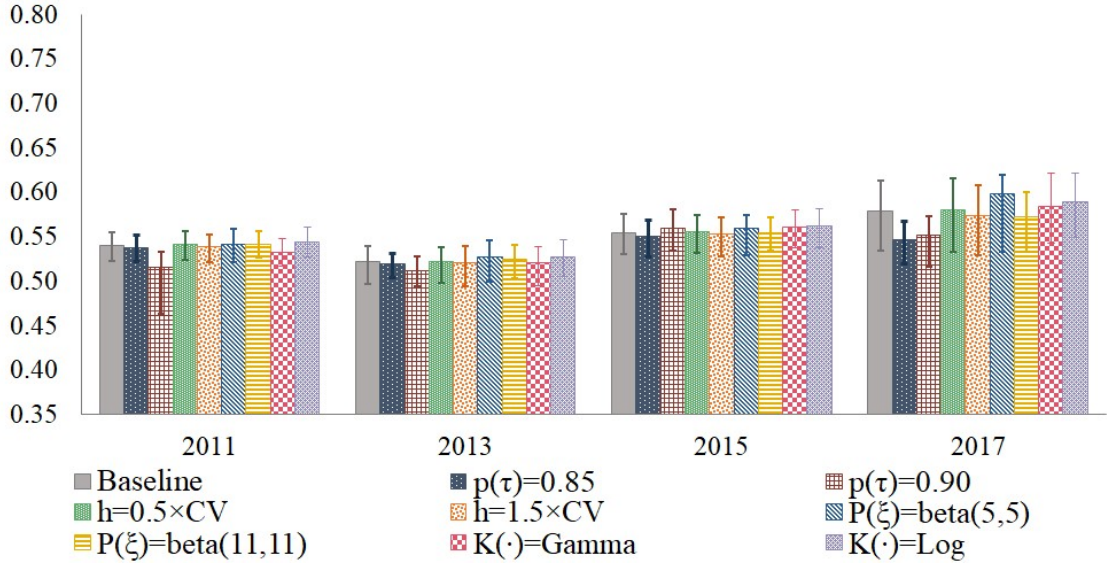
4.4 Robustness analysis

To test the robustness of the results to alternative estimation choices, I repeat the estimation algorithm proposed in the Section 2 under a range of alternative specifications. These include varying the threshold separating the bulk of the semi-parametric density from the GPD tail, alternative kernel functions and bandwidth choices for the kernel density component, different likelihood penalty functions. Robustness checks are conducted using both the largest (national-level) and smallest (provincial-level) samples.

Figure 11 shows the national-level robustness check. Gray bars on Figure 11 present baseline estimates produced using 95^{th} percentile threshold and bandwidth h for the non-parametric part estimated using cross-validation (CV). Blue and red bars represent the results applying 85^{th} and 90^{th} percentile as a threshold (respectively) and the baseline CV bandwidths h. Green and orange bars depict estimates varying bandwidth from half of the standard cross-validation bandwidth (green) to $1.5 \times CV$ bandwidth (orange) and keeping the baseline threshold percentile 95^{th} constant. Light blue and yellow bars represent the results which keep the baseline threshold percentile and bandwidth constant, but vary the penalty function. Light blue results employ a $beta(5,5)$ (smaller) penalty derived from the assumption that in 80% of cases, GPD shape ξ fitting income distribution lies between 0.3 and 0.7. The yellow results test $beta(11,11)$ (larger) penalty which implies that in 95% of cases the GPD shape ξ lies between 0.3 and 0.7. Pink bars represent the asymmetric gamma kernel proposed by Abadir and Lawford (2004) applied to the kernel density part. Asymmetric kernels are known to perform well for skewed and heavy-tailed distributions. In selecting a specific gamma kernel, I follow (Chen 2000, p. 472): the gamma shape parameter is set to $x/h+1$, where x denotes the evaluation point and h is the bandwidth, and the scale parameter



(a) CHFS (baseline)



(b) CFPS

Fig. 11: Gini estimates under alternative parameter choices. Robustness checks are based on different threshold percentiles ($p(\tau) = 0.85$ or 0.90), bandwidths ($h = 0.5 \times \text{CV}$ and $1.5 \times \text{CV}$, where CV denotes the optimal bandwidth selected via cross-validation), alternative penalty functions ($P(\xi) = \text{Beta}(5, 5)$ or $\text{Beta}(11, 11)$), and alternative kernel functions ($K(\cdot)$=Gamma or log-transformed kernel). The baseline parameter choice is $p(\tau) = 0.95$, $h = \text{CV}$, $P(\xi) = \text{Beta}(8, 8)$, and $K(\cdot)$=Gaussian. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

is h. Purple bars in Figure 11 correspond to a log-transform kernel from Charpentier and Flachaire (2015), which addresses the same issue of distributional skewness. Under this approach, a Gaussian kernel density estimator is applied to log-transformed income data. An additional advantage of both the gamma and log-transform kernels is their strictly positive support, which aligns with the support of the income distribution. Consequently, unlike the baseline Gaussian kernel, these alternatives do not require additional truncation or rescaling of the estimated density. Conditioning on confidence intervals, all approaches reveal no drastic changes in Gini estimates compared to the baseline both in CHFS (Figure 11a) and CFPS (Figure 11b) samples. According to Figures I1 — I5 in Appendix I, provincial-level results are also robust to the threshold, bandwidth, penalty, and kernel function choices.

To assess the robustness of the results to alternative treatments of zero and negative incomes in the CHFS, I follow the winsorization and imputation procedures proposed by Van Kerm (2007). Figure 12 presents the resulting national-level Gini index estimates under these alternative treatments. Grey bars correspond to the baseline specification, which excludes zero and negative income observations. Green bars report Gini estimates obtained after winsorizing zero and negative incomes by replacing these values with 10% of the median income. Finally, blue bars show

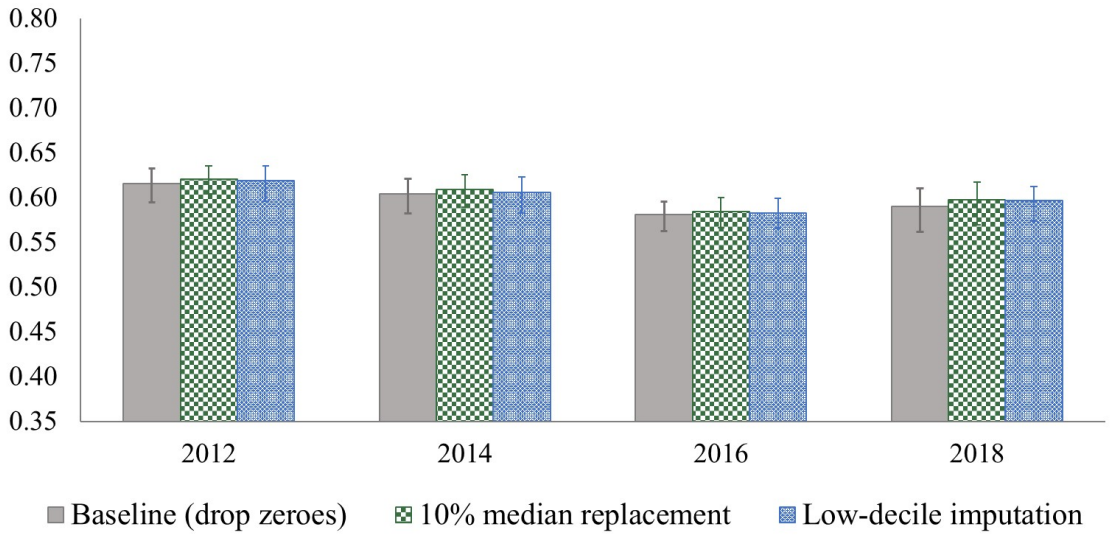


Fig. 12: National-level Gini estimates under alternative treatments of zero and negative income in the CHFS, following Van Kerm (2007). Bars correspond to: (a) the baseline approach, which excludes zero and negative income; (b) replacing zero and negative income with 10 percent of median income; and (c) imputation by randomly drawing values from the lowest income decile. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

Gini estimates based on imputing zero and negative incomes by randomly drawing values from the lowest income decile. To recall, zero and negative income observations account for only 1–2.5% of the CHFS sample. Consistent with this small share, Figure 12 shows that alternative treatments of these observations have a negligible effect on national-level Gini estimates. Figures I6–I10 in Appendix I.2 further confirm that alternative treatments of zero and negative incomes do not substantially affect provincial-level Gini estimates.

5 Conclusion

In recent years, a variety of scholars have measured income inequality in China employing different datasets and methods, yet, there is still an intense debate about the "true" levels and trends of inequality. In this paper, I assess inequality levels in China by applying a new measurement approach, which aims to appropriately account for the presence of top-income outliers in survey data, especially in small samples. The proposed method employs a semiparametric model combining the nonparametric kernel density approximating the bulk and Generalized Pareto density (GPD) fitting the income distribution top tail. Parameters of the GPD tail are suggested to be estimated using a maximum penalized likelihood estimator (MPLE). Conducting a Monte Carlo simulation study, I demonstrate that, conditioning on the assumption about the distribution of the GPD shape parameter, the MPLE estimator outperforms MLE and OBRE estimators commonly used in the previous literature in terms of mean squared error, achieving substantially lower variance at the cost of a moderate and predictable bias that is conservative with respect to inequality measurement. The MPLE achieves numerically equivalent point estimation performance to a fully Bayesian MCMC approach at a substantially lower computational cost, and converges to MLE asymptotically as sample size grows.

The introduced estimation approach has been employed to study income inequality in China based on the data from the China Household Finance Survey (CHFS 2013-2019). Relying on top income observations from the CHFS, the proposed algorithm helped to reduce the top-income-driven variation in Gini indices across years in rural/urban-, and regional-level estimates, as well as among Gini indices for five Chinese provinces (Shanghai, Guangdong, Liaoning, Henan, and Gansu). I also performed supplementary inequality analysis using the data from China Family Panel Studies (CFPS 2012-2018).

The Gini index in China was estimated at 0.616 in 2012, 0.604 in 2014, 0.581 in 2016, and 0.590 in 2018. Consistent with previous literature, I document substantial rural-urban and regional disparities: urban households earn substantially more than rural households, and coastal regions are richer than inland regions (Wan et al. 2012; Gao et al. 2014; Zhang et al. 2012; Zhang 2021). However, unlike earlier studies reporting higher inequality in rural areas (Zhang et al. 2012; Zhang

2021; Li and Sicular 2014), I find similar levels of income inequality across rural and urban households. This difference likely reflects the more recent period covered in my study and the use of the CHFS survey. Income inequality is higher in the Eastern and Western regions, a pattern broadly in line with (Luo et al. 2020). Decomposition of the Theil T index indicates that within-group variation (both rural-urban and regional) accounts for the majority of nationwide income inequality, exceeding 87%, consistent with previous literature (Wan et al. 2018; Luo et al. 2020).

In this study, I emphasize the existence of two distinct branches in the income inequality literature: (a) studies estimating inequality purely based on household survey data; and (b) studies that use survey data plus any top income adjustments. The former typically yields lower income inequality estimates than the latter because inequality indices are highly sensitive to top-income observations. I claim that the comparison of inequality estimates is only appropriate within each of these distinct branches of inequality literature, not across them. Since I rely on the CHFS, as a survey that was especially targeted for adequately representing affluent households, my results should only be compared to studies that also imply top income adjustments. The obtained income distribution parameters and the resulting inequality estimates are consistent with those reported in other studies based on the CHFS data (Gan et al. 2014; Xie and Zhou 2014), as well as with those relying on other top income data for China (Li et al. 2020; Piketty et al. 2019). Gini index estimates produced in the current study are incomparable with the existing international statistics or any research that is entirely based on regular household survey data.

References

Abadir, K.M. and S. Lawford. 2004. Optimal Asymmetric Kernels. *Economics Letters 83*(1): 61–68 .

Aitchison, J. and J.A.C. Brown. 1969. *The Lognormal Distribution*. Cambridge: University Press.

Allison, P.D. 1978. Measures of Inequality. *American sociological review* 43: 865–880 .

Alvaredo, F. 2011. A Note on the Relationship Between Top Income Shares and the Gini Coefficient. *Economics Letters 110*(3): 274–277 .

Arnold, B. 2008. Pareto and Generalized Pareto Distributions, In *Modeling Income Distributions and Lorenz Curves*, ed. Chotikapanich, D., 119–145. New York: Springer.

Arnold, B. 2015. *Pareto Distributions (2nd ed.)*. New York: Chapman and Hall/CRC.

Atkinson, A.B. 2007. Measuring Top Incomes: Methodological Issues, In *Top Incomes Over the Twentieth Century: a Contrast Between Continental European and English-peaking Countries*, eds. Atkinson, A.B. and T. Piketty, 18–42. Oxford: Oxford University Press Oxford.

Atkinson, A.B. and T. Piketty. 2007. *Top Incomes Over the Twentieth Century: a Contrast Between Continental European and English-speaking Countries*. Oxford.

Atkinson, A.B., T. Piketty, and E. Saez. 2011. Top Incomes in the Long Run of History. *Journal of Economic Literature 49*(1): 3–71 .

Bandourian, R., J. McDonald, and R.S. Turley. 2002. A Comparison of Parametric Models of Income Distribution Across Countries and Over Time. Luxembourg Income Study Working Paper No. 305, Available at SSRN: https://ssrn.com/abstract=324900.

Blanchet, T., J. Fournier, and T. Piketty. 2022. Generalized Pareto Curves: Theory and Applications. *Review of Income and Wealth 68*(1): 263–288 .

Bourguignon, F. 1979. Decomposable Income Inequality Measures. *Econometrica: Journal of the Econometric Society 47*(4): 901–920 .

Brandt, L. and C.A. Holz. 2006. Spatial Price Differences in China: Estimates and Implications. *Economic Development and Cultural Change 55*(1): 43–86 .

Brazauskas, V. and A. Kleefeld. 2009. Robust and Efficient Fitting of the Generalized Pareto Distribution with Actuarial Applications in View. *Insurance: Mathematics and Economics 45*(3): 424–435 .

Cameron, A.C. and P.K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge: University Press.

Charpentier, A. and E. Flachaire. 2015. Log-transform Kernel Density Estimation of Income Distribution. *L'Actualité économique 91*(1): 141–159 .

Charpentier, A. and E. Flachaire. 2022. Pareto Models for Top Incomes. *The Journal of Economic Inequality* 20: 1–25 .

Chen, J., M. Pu, and W. Hou. 2019. The Trend of the Gini coefficient of China (1978–2010). *Journal of Chinese Economic and Business Studies 17*(3): 261–285 .

Chen, S.X. 2000. Probability Density Function Estimation Using Gamma Kernels. *Annals of the institute of statistical mathematics 52*(3): 471–480 .

Chotikapanich, D. and W.E. Griffiths. 2008. Estimating Income Distributions Using a Mixture of Gamma Densities, In *Modeling Income Distributions and Lorenz Curves*, ed. Chotikapanich, D., 285–302. New York: Springer.

Clementi, F. and M. Gallegati. 2016. *The Distribution of Income and Wealth*. New York: Springer.

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values*. New York: Springer.

Coles, S.G. and M.J. Dixon. 1999. Likelihood-based Inference for Extreme Value Models. *Extremes* 2: 5–23 .

Cowell, F.A. 1980. On the Structure of Additive Inequality Measures. *The Review of Economic Studies 47*(3): 521–531 .

Cowell, F.A. and E. Flachaire. 2007. Income Distribution and Inequality Measurement: The Problem of Extreme Values. *Journal of Econometrics 141*(2): 1044–1072 .

Cowell, F.A. and M.P. Victoria-Feser. 1996. Robustness Properties of Inequality Measures. *Econometrica 64*(1): 77–101 .

Cowell, F.A. and M.P. Victoria-Feser. 2007. Robust Stochastic Dominance: A Semi-parametric Approach. *The Journal of Economic Inequality 5*(1): 21–37 .

Dagum, C. 1998. A New Approach to the Decomposition of the Gini Income Inequality Ratio. In D. J. Slottje and B. Raj (Eds.), *Income Inequality, Poverty, and Economic Welfare*, Heidelberg, pp. 47–63. Physica-Verlag HD.

Dagum, C. 2008. A New Model of Personal Income Distribution: Specification and Estimation, In *Modeling Income Distributions and Lorenz Curves*, ed. Chotikapanich, D., 3–25. New York: Springer.

Davidson, R. and E. Flachaire. 2007. Asymptotic and Bootstrap Inference for Inequality and Poverty Measures. *Journal of Econometrics 141*(1): 141–166 .

Davison, A.C. and D.V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: University Press.

Dunford, M. and L. Li. 2010. Chinese Spatial Inequalities and Spatial Policies. *Geography Compass 4*(8): 1039–1054 .

Esteban, J. 1986. Income-Share Elasticity and the Size Distribution of Income. *International Economic Review 27*(2): 439–444 .

Fan, S., R. Kanbur, and X. Zhang. 2011. China's Regional Disparities: Experience and Policy. *Review of Development Finance 1*(1): 47–56 .

Flachaire, E. and O. Nunez. 2007. Estimation of the Income Distribution and Detection of Subpopulations: an Explanatory Model. *Computational Statistics & Data Analysis 51*(7): 3368–3380 .

Gan, L., Z. Yin, N. Jia, S. Xu, S. Ma, L. Zheng, et al. 2014. *Data You Need to Know About China*. New York: Springer.

Gao, Y., J. Zheng, and M. Bu. 2014. Rural-urban Income Gap and Agricultural Growth in China. *China Agricultural Economic Review 6*(1): 92–107 .

Gibrat, R. 1931. *Les Inégalites Économiques*. Paris: Libraire du Recueil Sirey.

Gustafsson, B., L. Shi, and H. Sato. 2014. Data for Studying Earnings, the Distribution of Household Income and Poverty in China. *China Economic Review* 30: 419–431 .

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Probability and Statistics Series. Wiley.

Hampel, F.R. 1974. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association 69*(346): 383–393 .

Han, X. and Y. Cheng. 2019. Does the "Missing" High-income Matter? - Income Distribution and Inequality Revisited With Truncated Distribution. *China Economic Review* 57: 101337 .

Huang, C., J.G. Lin, and Y.Y. Ren. 2012. Statistical Inferences for Generalized Pareto Distribution Based on Interior Penalty Function Algorithm and Bootstrap Methods and Applications in Analyzing Stock Data. *Computational Economics* 39: 173–193 .

Jain-Chandra, M.S., N. Khor, R. Mano, J. Schauer, M.P. Wingender, and J. Zhuang. 2018. *Inequality in China – Trends, Drivers and Policy Remedies*. International Monetary Fund.

Jenkins, S.P. 1995. Accounting for Inequality Trends: Decomposition Analyses for the UK, 1971-86. *Economica*: 29–63 .

Jenkins, S.P. 2017. Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. *Economica 84*(334): 261–289 .

Kanbur, R., Y. Wang, and X. Zhang. 2021. The Great Chinese Inequality Turnaround. *Journal of Comparative Economics 49*(2): 467–482 .

Kleiber, C. and S. Kotz. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*, Volume 470. New Jersey: John Wiley & Sons.

Knight, J. 2014. Inequality in China: An Overview. *The World Bank Research Observer 29*(1): 1–19 .

Lebergott, S. 1959. The Shape of the Income Distribution. *The American Economic Review 49*(3): 328–347 .

Li, Q., S. Li, and H. Wan. 2020. Top Incomes in China: Data Collection and the Impact on Income Inequality. *China Economic Review* 62: 101495 .

Li, S. and T. Sicular. 2014. The Distribution of Household Income in China: Inequality, Poverty and Policies. *The China Quarterly* 217: 1–41 .

Lu, D. 2008. China's Regional Income Disparity An Alternative Way to Think of the Sources and Causes 1. *Economics of Transition 16*(1): 31–58 .

Lubrano, M. and A. Ndoye. 2016. Income Inequality Decomposition Using a Finite Mixture of Log-normal Distributions: A Bayesian Approach. *Computational Statistics & Data Analysis* 100: 830–846 .

Luo, C., S. Li, and T. Sicular. 2020. The Long-term Evolution of National Income Inequality and Rural Poverty in China. *China Economic Review* 62: 101465 .

Maasoumi, E. 1999. Empirical Analyses of Inequality and Welfare, In *Handbook of Applied Econometrics, Volume 2: Microeconomics*, eds. Pesaran, M.H. and P. Schmidt, 187–226. Wiley Online Library.

Martins, E.S. and J.R. Stedinger. 2000. Generalized Maximum-likelihood Generalized Extreme-value Quantile Estimators for Hydrologic Data. *Water Resources Research 36*(3): 737–744 .

Martins, E.S. and J.R. Stedinger. 2001. Generalized Maximum Likelihood Pareto-Poisson Estimators for Partial Duration Series. *Water Resources Research 37*(10): 2551–2557 .

Milanovic, B. 2002. True World Income Distribution, 1988 and 1993: First Calculation Based on Household Surveys Alone. *The Economic Journal 112*(476): 51–92 .

Naughton, B.J. 2006. *The Chinese Economy: Transitions and Growth*. Cambridge, MA: MIT press.

Novokmet, F., T. Piketty, and G. Zucman. 2018. From Soviets to Oligarchs: Inequality and Property in Russia 1905-2016. *The Journal of Economic Inequality 16*(2): 189–223 .

Pareto, V. 1895. La Legge Della Domanda. *Giornale Degli Economisti*: 59–68 .

Pauli, F. and S. Coles. 2001. Penalized Likelihood Inference in Extreme Value Analyses. *Journal of Applied Statistics 28*(5): 547–560 .

Piketty, T., L. Yang, and G. Zucman. 2019. Capital Accumulation, Private Property, and Rising Inequality in China, 1978–2015. *American Economic Review 109*(7): 2469–2496 .

Ravallion, M. and S. Chen. 2009. China's (Uneven) Progress Against Poverty, In *Governing rapid growth in China*, eds. Kanbur, R. and X. Zhang, 65–111. London: Routledge.

Roser, M., B. Rohenkohl, P. Arriagada, J. Hasell, H. Ritchie, and E. Ortiz-Ospina. 2023. Economic Growth. Our World in Data, https://ourworldindata.org/economic-growth.

Salem, A.B. and T.D. Mount. 1974. A Convenient Descriptive Model of Income Distribution: the Gamma Density. *Econometrica 42*(6): 1115–1127 .

Shorrocks, A.F. 1982. Inequality Decomposition by Factor Components. *Econometrica: Journal of the Econometric Society*: 193–211 .

Shorrocks, A.F. 1983. The Impact of Income Components on the Distribution of Family Incomes. *The Quarterly Journal of Economics 98*(2): 311–326 .

Shorrocks, A.F. 1984. Inequality Decomposition by Population Subgroups. *Econometrica 52*(6): 1369–1385 .

Sicular, T., Y. Ximing, B. Gustafsson, and L. Shi. 2007. The Urban–Rural Income Gap and Inequality in China. *Review of income and wealth 53*(1): 93–126 .

Singh, S.K. and G.S. Maddala. 2008. A Function for Size Distribution of Incomes, In *Modeling Income Distributions and Lorenz Curves*, ed. Chotikapanich, D., 27–35. New York: Springer.

Singh, V.P. and H. Guo. 1995. Parameter Estimation for 3-parameter Peneralized Pareto Distribution by the Principle of Maximum Entropy (POME). *Hydrological sciences journal 40*(2): 165–181 .

Van Kerm, P. 2007. Extreme Incomes and the Estimation of Poverty and Inequality Indicators from EU-SILC. IRISS Working Papers; No. 2007-01.

Victoria-Feser, M.P. 2000. Robust Methods for the Analysis of Income Distribution, Inequality and Poverty. *International Statistical Review 68*(3): 277–293 .

Victoria-Feser, M.P. and E. Ronchetti. 1994. Robust Methods for Personal-income Distribution Models. *Canadian Journal of Statistics 22*(2): 247–258 .

Wan, G., T. Wu, and Y. Zhang. 2018. The Decline of Income Inequality in China: Assessments and Explanations. *Asian Economic Papers 17*(3): 115–140 .

Wan, G., J. Ye, and J. Zhuang. 2012. On Regional and Inter-household Inequality in China, In *A New Economic Growth Engine for China: Escaping the Middle-income Trap by Not Doing More of the Same*, eds. Woo, W.T., M. Lu, J.D. Sachs, and Z. Chen, 155–182. London: Imperial College Press.

Wang, C., G. Wan, and D. Yang. 2014. Income Inequality in the People's Republic of China: Trends, Determinants, and Proposed Remedies. *Journal of Economic Surveys 28*(4): 686–708 .

World Bank. 2020. World Development Indicators, Gini index.

World Inequality Database (WID.world). 2024.

Xie, Y. and Y. Jin. 2015. Household Wealth in China. *Chinese Sociological Review 47*(3): 203–229 .

Xie, Y. and X. Zhou. 2014. Income Inequality in Today's China. *Proceedings of the National Academy of Sciences 111*(19): 6928–6933 .

Yang, D.T. 1999. Urban-Biased Policies and Rising Income Inequality in China. *American Economic Review 89*(2): 306–310 .

Yang, D.T. and H. Zhou. 1999. Rural-Urban Disparity and Sectoral Labour Allocation in China. *The Journal of Development Studies 35*(3): 105–133 .

Yang, J. and M. Qiu. 2016. The Impact of Education on Income Inequality and Intergenerational Mobility. *China Economic Review* 37: 110–125 .

Yang, X. and L. Gan. 2020. Bequest Motive, Household Portfolio Choice, and Wealth Inequality in Urban China. *China Economic Review* 60: 101399 .

Zhang, J. 2021. A Survey on Income Inequality in China. *Journal of Economic Literature 59*(4): 1191–1239 .

Zhang, X., Y. Liu, and Z. Wang. 2012. *Zhongguo Jumin Shouru Fenpei Niandu Baogao 2011 (Annual Report on Income Distribution of Chinese Residents 2011)*. Beijing: Economic Science Press.

Appendix A Empirical evidence on inequality in China

Table A1: National-level Gini index estimates for China

Source	Data	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
					Income Inequality							
CSY NBS [1]	NBS surveys[2]	0.481	0.477	0.474	0.473	0.469	0.462	0.465	0.467	0.468	0.465	0.468
World Bank (2020)	NBS surveys[3]	0.437**	0.424**	0.422**	0.397**	0.392**	0.386**	0.385**	0.391**	0.385**	0.382**	0.371**
Chen et al. (2019)	CSY NBS[4]	0.452										
Luo et al. (2020)	CHIP 2013			0.433								
Kanbur et al. (2021)	CFPS 2012-2014		0.504		0.495							
Xie and Zhou (2014)	CGSS 2012		0.563									
	CFPS 2012		0.526									
	CHFS 2011	0.633										
	CLDS 2012		0.536*									
Han and Cheng (2019)	CHIP 2013			0.455								
Gan et al. (2014)	CHFS 2013			0.610*								
Li et al. (2020)	CHIP 2016, TIC[5]						0.646					
					Wealth Inequality							
Yang and Gan (2020)	CHFS 2011-2015	0.72		0.74		0.76						
Xie and Jin (2015)	CFPS 2012, HRL[6]		0.73									

Notes: All Gini estimates of income inequality in this table produced based on household income per capita, except "*" results produced based on the total household income, and the World Bank's "**" estimates which were assessed based on the per capita consumption data shared by the National Bureau of Statistics (NBS).

[1]China Statistical Yearbook of the NBS.

[2]Chinese official statistical surveys of the NBS https://www.stats.gov.cn/english/Statisticaldata/yearbook.

[3]The data is based on primary household surveys from NBS adopted for cross-country comparable format.

[4]Grouped income distribution data from the China Statistical Yearbook of 2011.

[5]Top Incomes in China (TIC).

[6]Hurun Rich List (HRL).

Appendix B Technical Appendix

B.1 Distributional formulas

The semiparametric cumulative distribution function (CDF) is defined by:

$$F(x) = \begin{cases} \dfrac{\int_0^x \sum_{i=1}^{n_\tau} w_i K\left(\dfrac{x_i - t}{h}\right) dt}{\int_0^\tau \sum_{i=1}^{n_\tau} w_i K\left(\dfrac{x_i - t}{h}\right) dt}, & \text{if } x \leq \tau \\ \\ p_\tau + (1 - p_\tau)\left[1 - \left(1 + \xi \dfrac{x - \tau}{\sigma}\right)^{-\frac{1}{\xi}}\right], & \text{if } x > \tau. \end{cases} \tag{B1}$$

Denoting the non-parametric part of the CDF by $G(\cdot)$, the corresponding quantile function is:

$$Q(p) = \begin{cases} G^{-1}(p), & \text{if } p \leq p_\tau \\ \tau - \dfrac{\sigma}{\xi} + \dfrac{\sigma}{\xi}\left(\dfrac{1 - p_\tau}{1 - p}\right)^{\xi}, & \text{if } p > p_\tau, \end{cases} \tag{B2}$$

where $0 < p < 1$ is a probability value.

B.2 Derivation of the GPD tail Gini index

For observations above the threshold τ, the GPD quantile function and mean are

$$Q_R(p) \;=\; \tau + \frac{\sigma}{\xi}\big[(1-p)^{-\xi} - 1\big], \qquad \mu_R \;=\; \tau + \frac{\sigma}{1-\xi}, \tag{B3}$$

where μ_R is finite for $\xi < 1$. To simplify notation, write $Q_R(p) = A + B(1-p)^{-\xi}$ with $A = \tau - \sigma/\xi$ and $B = \sigma/\xi$.

1. The **Lorenz curve**. By definition, $L_R(p) = \mu_R^{-1} \int_0^p Q_R(t)\, dt$. Evaluating the integral by direct antidifferentiation:
$$\int_0^p Q_R(t)\, dt \;=\; Ap + \frac{B}{1-\xi}\big[1 - (1-p)^{1-\xi}\big]. \tag{B4}$$
Setting $p = 1$ confirms $\int_0^1 Q_R(t)\, dt = A + B/(1-\xi) = \mu_R$.

2. The **integral of the Lorenz curve**:
$$\int_0^1 L_R(p)\, dp = \frac{1}{\mu_R}\int_0^1 \left[Ap + \frac{B}{1-\xi}\left(1 - (1-p)^{1-\xi}\right)\right] dp.$$
The numerator splits into two parts $I_1 + I_2$, where
$$I_1 = \int_0^1 Ap\, dp = \frac{A}{2}, \qquad I_2 = \frac{B}{1-\xi}\int_0^1 \left[1 - (1-p)^{1-\xi}\right] dp = \frac{B}{2-\xi}.$$
Substituting back:
$$\int_0^1 L_R(p)\, dp = \frac{\dfrac{A}{2} + \dfrac{B}{2-\xi}}{A + \dfrac{B}{1-\xi}}.$$

3. The **Gini index**:

$$G_R = 1 - 2\int_0^1 L_R(p)\,dp = 1 - \frac{A(2-\xi)+2B}{(2-\xi)\left(A+\dfrac{B}{1-\xi}\right)}. \tag{B5}$$

Substituting $A = \tau - \dfrac{\sigma}{\xi}$ and $B = \dfrac{\sigma}{\xi}$:

$$A(2-\xi)+2B = \left(\tau - \frac{\sigma}{\xi}\right)(2-\xi) + \frac{2\sigma}{\xi} = \tau(2-\xi)+\sigma,$$
$$A + \frac{B}{1-\xi} = \tau - \frac{\sigma}{\xi} + \frac{\sigma}{\xi(1-\xi)} = \tau + \frac{\sigma}{1-\xi}.$$

Multiplying numerator and denominator by $(1-\xi)$:

$$G_R = 1 - \frac{(1-\xi)\left[\tau(2-\xi)+\sigma\right]}{(2-\xi)\left[\tau(1-\xi)+\sigma\right]}. \tag{B6}$$

4. To match the form of G_R in Equation (6), apply the identities $1-\xi = -(\xi-1)$ and $2-\xi = -(\xi-2)$:

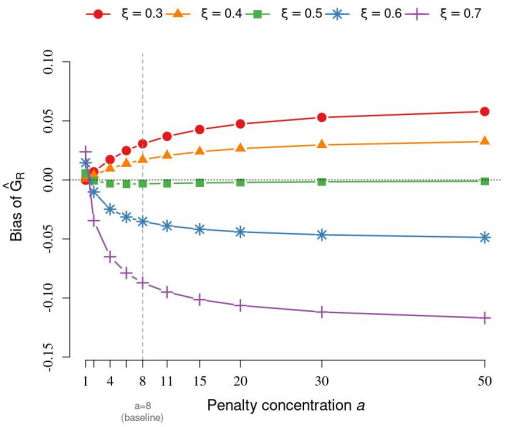
$$(1-\xi)\left[\tau(2-\xi)+\sigma\right] = -(\xi-1)\left[\sigma - \tau(\xi-2)\right],$$
$$(2-\xi)\left[\tau(1-\xi)+\sigma\right] = -(\xi-2)\left[\sigma - \tau(\xi-1)\right].$$

Substituting into Equation (B6) yields the final form of Equation (6):

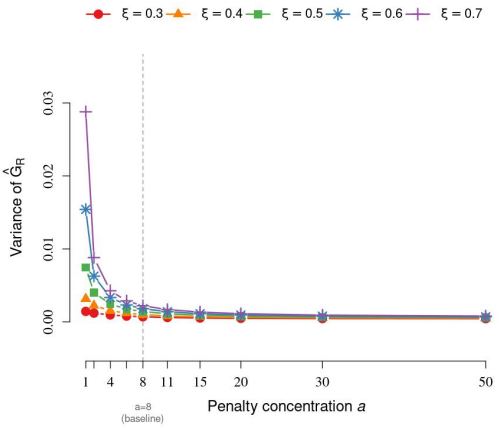
$$G_R = 1 - \frac{(\xi-1)\left(\sigma - \tau(\xi-2)\right)}{(\xi-2)\left(\sigma - \tau(\xi-1)\right)}. \tag{B7}$$

B.3 Penalty sensitivity

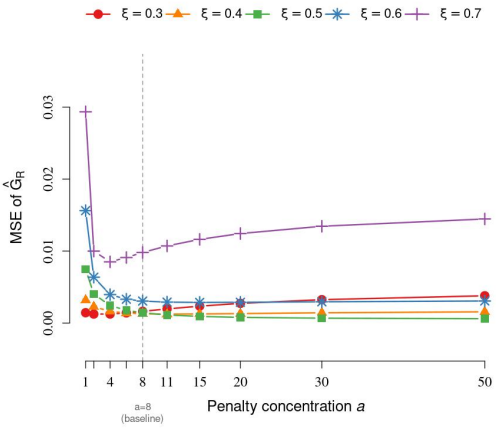
Figure B1 examines the sensitivity of $\hat{G}_R$ to the choice of penalty concentration a, based on 5,000 Monte Carlo replications from samples of size $n = 100$ drawn from the GPD with $\tau = 1$ and $\sigma = 0.5$, varying a across the grid $a \in \{1, 2, 4, 6, 8, 11, 15, 20, 30, 50\}$ with the baseline $a = 8$ indicated by the dashed vertical line.Note that $a = 1$ corresponds to a Beta$(1, 1)$ uniform penalty, which is equivalent to the standard MLE, providing a natural reference point. Three findings are as follows. First, variance decreases monotonically and rapidly as a increases: even a modest concentration of $a = 4$ reduces variance substantially relative to MLE, and by $a = 8$ variance is already near its minimum for all values of ξ. Second, bias increases in magnitude as a grows, pulling estimates upward for $\xi < 0.5$ and downward for $\xi > 0.5$, consistent with shrinkage toward the mode of the Beta(a,a) distribution at 0.5. The bias at $\xi = 0.5$ remains negligible throughout, as the true value coincides with the penalty mode. Third, and most importantly, the MSE figure reveals that the baseline choice of $a = 8$ sits at or very near the MSE minimum across all values of ξ. For the most challenging case of $\xi = 0.7$, where variance is largest and the bias–variance trade-off is most consequential, the MSE minimum is clearly located in the range $a \in [6, 8]$, and MSE rises noticeably for $a > 11$ as increasing bias begins to dominate. For lower values of ξ, MSE is relatively flat beyond $a = 4$, confirming that the penalty choice matters most precisely in the heavy-tail regime where variance reduction is most needed. Taken together, these Monte Carlo results provide simulation-based support for the choice of $a = 8$: it achieves the best bias–variance trade-off for heavy-tailed distributions and remains near-optimal for moderate tails, and the conclusions of the simulation study are not sensitive to moderate perturbations of the penalty concentration around this value.



(a) Bias of $\hat{G}_R$



(b) Variance of $\hat{G}_R$



(c) MSE of $\hat{G}_R$

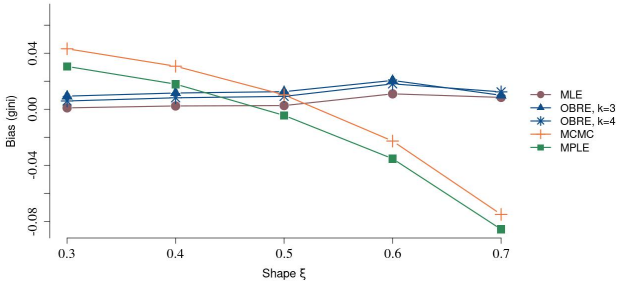
Fig. B1: Sensitivity of $\hat{G}_R$ to the penalty concentration parameter a of the Beta(a, a) penalty function. Simulated from GPD$(\tau = 1,\ \sigma = 0.5)$, $n = 100$, 5,000 Monte Carlo replications, baseline scenario. The dashed vertical line marks the baseline choice $a = 8$. Panel (a) shows bias, panel (b) variance, and panel (c) MSE of the tail Gini index $\hat{G}_R$ as functions of $a \in \{1, 2, 4, 6, 8, 11, 15, 20, 30, 50\}$, separately for each true shape parameter $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. The dotted horizontal line in panel (a) marks zero bias. Increasing a concentrates the penalty toward $\xi = 0.5$, reducing variance at the cost of increasing bias away from $\xi = 0.5$. Error bars represent ± 1.96 Monte Carlo standard deviations across 5,000 replications.

Appendix C Simulation results

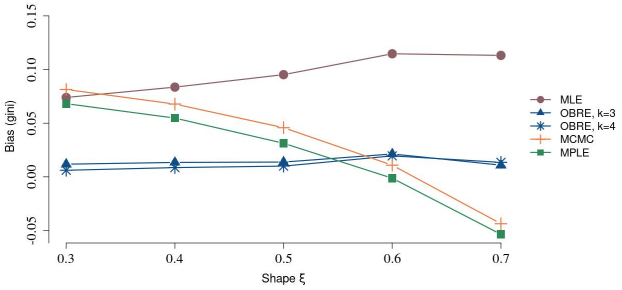
Table C1: Bias (B), Variance (V), and MSE (M) of the tail Gini index $\hat{G}_R$ across seven simulation scenarios. Samples of size $n = 100$ drawn from GPD($\tau = 1$, $\sigma = 0.5$), 5,000 Monte Carlo replications.

ξ	MLE			OBRE, k=3			OBRE, k=4			MCMC			MPLE		
	B	V	M	B	V	M	B	V	M	B	V	M	B	V	M
Baseline															
0.3	0.001	0.001	0.001	0.009	0.003	0.003	0.006	0.002	0.002	0.043	0.001	0.003	0.030	0.001	0.002
0.4	0.002	0.003	0.003	0.012	0.006	0.006	0.008	0.004	0.004	0.031	0.001	0.002	0.018	0.001	0.001
0.5	0.003	0.006	0.006	0.013	0.011	0.011	0.009	0.009	0.009	0.010	0.001	0.001	-0.004	0.001	0.001
0.6	0.011	0.013	0.013	0.021	0.020	0.021	0.018	0.017	0.017	-0.023	0.002	0.002	-0.035	0.002	0.003
0.7	0.008	0.021	0.021	0.010	0.028	0.028	0.012	0.026	0.026	-0.075	0.002	0.008	-0.086	0.002	0.010
Brazauskas (15% from GPD($\xi = 1.5$))															
0.3	0.167	0.022	0.050	0.083	0.011	0.018	0.093	0.011	0.020	0.122	0.003	0.018	0.110	0.003	0.015
0.4	0.166	0.026	0.054	0.095	0.016	0.025	0.106	0.017	0.028	0.102	0.003	0.013	0.091	0.004	0.012
0.5	0.166	0.028	0.055	0.115	0.025	0.038	0.125	0.024	0.039	0.074	0.003	0.009	0.064	0.003	0.007
0.6	0.150	0.030	0.052	0.107	0.030	0.041	0.116	0.027	0.041	0.033	0.003	0.004	0.024	0.004	0.004
0.7	0.115	0.030	0.044	0.082	0.032	0.039	0.096	0.031	0.040	-0.028	0.003	0.004	-0.036	0.004	0.005
Cowell (largest obs. $\times 10$)															
0.3	0.074	0.002	0.007	0.012	0.003	0.003	0.006	0.002	0.002	0.081	0.001	0.007	0.068	0.001	0.005
0.4	0.084	0.005	0.012	0.013	0.006	0.006	0.009	0.004	0.005	0.068	0.001	0.005	0.055	0.001	0.004
0.5	0.095	0.010	0.019	0.014	0.012	0.012	0.010	0.009	0.009	0.046	0.001	0.003	0.031	0.001	0.002
0.6	0.115	0.018	0.031	0.021	0.020	0.021	0.019	0.018	0.018	0.011	0.002	0.002	-0.001	0.002	0.002
0.7	0.113	0.024	0.037	0.011	0.028	0.028	0.013	0.026	0.026	-0.044	0.002	0.004	-0.054	0.002	0.005
Drop 25% non-randomly															
0.3	0.013	0.002	0.002	0.027	0.004	0.004	0.021	0.003	0.003	0.065	0.001	0.005	0.050	0.001	0.003
0.4	0.013	0.004	0.004	0.030	0.008	0.009	0.022	0.006	0.006	0.050	0.001	0.004	0.035	0.001	0.002
0.5	0.024	0.011	0.012	0.046	0.017	0.019	0.037	0.013	0.015	0.028	0.001	0.002	0.014	0.002	0.002
0.6	0.028	0.018	0.019	0.048	0.025	0.027	0.043	0.022	0.024	-0.008	0.002	0.002	-0.022	0.002	0.002
0.7	0.019	0.026	0.027	0.031	0.030	0.031	0.031	0.029	0.030	-0.064	0.002	0.006	-0.076	0.002	0.008
Drop 25% randomly															
0.3	0.003	0.002	0.002	0.015	0.004	0.005	0.010	0.003	0.003	0.050	0.001	0.003	0.035	0.001	0.002
0.4	0.003	0.004	0.004	0.015	0.008	0.009	0.010	0.006	0.006	0.035	0.001	0.002	0.019	0.001	0.002
0.5	0.008	0.009	0.009	0.018	0.015	0.015	0.015	0.012	0.012	0.011	0.002	0.002	-0.003	0.002	0.002
0.6	0.010	0.018	0.018	0.017	0.024	0.024	0.018	0.022	0.022	-0.025	0.002	0.002	-0.040	0.002	0.004
0.7	-0.001	0.025	0.025	-0.006	0.029	0.029	0.000	0.027	0.027	-0.081	0.002	0.009	-0.095	0.002	0.011
Halve 50% non-randomly															
0.3	0.043	0.003	0.005	0.067	0.008	0.012	0.056	0.005	0.008	0.098	0.001	0.010	0.080	0.001	0.007
0.4	0.046	0.006	0.009	0.074	0.014	0.019	0.062	0.010	0.014	0.080	0.001	0.007	0.062	0.001	0.005
0.5	0.060	0.015	0.018	0.089	0.022	0.030	0.080	0.019	0.025	0.054	0.001	0.004	0.038	0.001	0.003
0.6	0.068	0.022	0.026	0.090	0.030	0.038	0.088	0.027	0.035	0.015	0.002	0.002	0.000	0.002	0.002
0.7	0.053	0.029	0.032	0.061	0.032	0.035	0.061	0.030	0.034	-0.044	0.002	0.004	-0.057	0.002	0.005
Halve 50% randomly															
0.3	0.001	0.003	0.003	0.016	0.007	0.007	0.009	0.004	0.004	0.049	0.001	0.003	0.033	0.001	0.002
0.4	0.002	0.005	0.005	0.017	0.011	0.011	0.011	0.008	0.008	0.034	0.001	0.002	0.018	0.001	0.002
0.5	0.011	0.012	0.012	0.022	0.018	0.019	0.018	0.015	0.015	0.012	0.002	0.002	-0.004	0.002	0.002
0.6	0.018	0.020	0.020	0.020	0.025	0.026	0.022	0.023	0.024	-0.023	0.002	0.003	-0.037	0.002	0.003
0.7	0.004	0.026	0.026	0.001	0.031	0.031	0.008	0.029	0.029	-0.077	0.002	0.008	-0.090	0.002	0.011

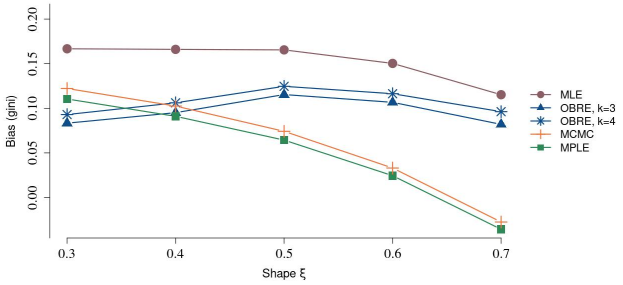
Notes: B = Bias, V = Variance, M = MSE. Bias and MSE for all estimators are computed relative to the Monte Carlo mean estimate across 5,000 replications. For MCMC, the point estimate is the average posterior mean across replications; variance is the sampling variance of the posterior mean estimator across Monte Carlo replications, approximated from the interquartile range of its sampling distribution assuming normality. MLE: maximum likelihood estimator. OBRE: optimal bias-robust estimator (Victoria-Feser and Ronchetti 1994), with sensitivity parameters $k \in \{3, 4\}$. MCMC: Bayesian Gibbs sampler with shape prior $\xi \sim \text{Beta}(8, 8)$ and uninformative prior for scale σ. MPLE: maximum penalized likelihood estimator with Beta$(8, 8)$ penalty on ξ.



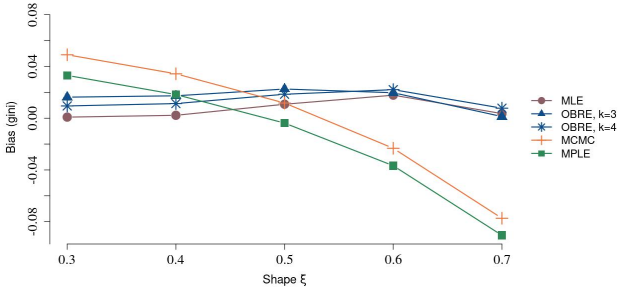
(a) Baseline



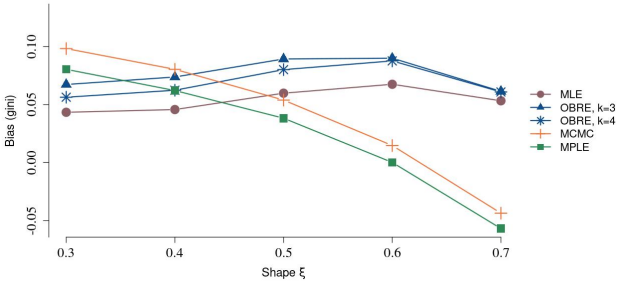
(b) Largest value $\times 10$



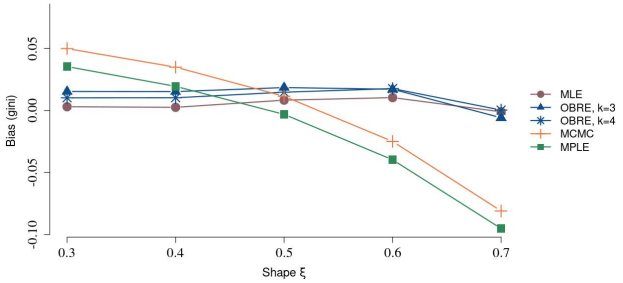
(c) Draw 15% from GPD($\xi = 1.5$)



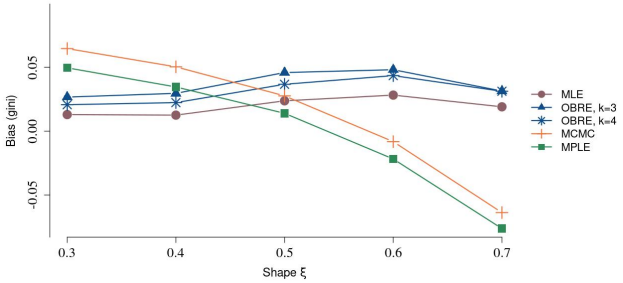
(d) Divide 50% by 2, randomly



(e) Divide 50% by 2, non-randomly

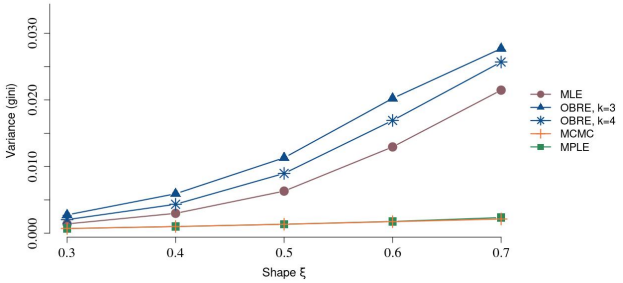


(f) Drop 25%, randomly

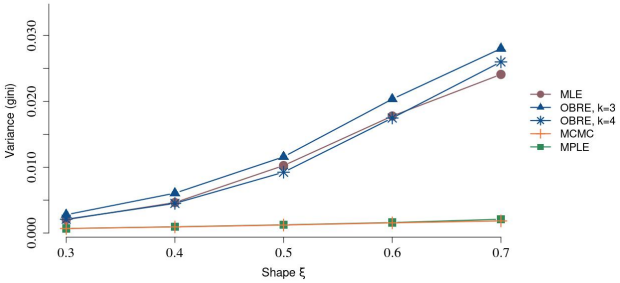


(g) Drop 25%, non-randomly

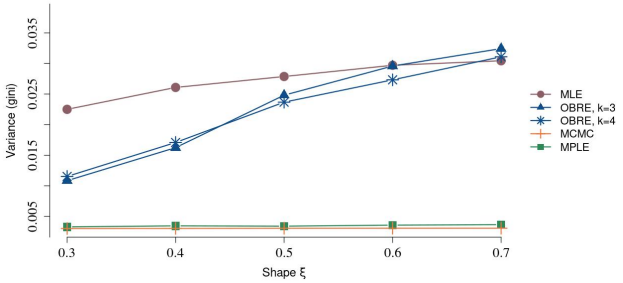
Fig. C1: *Bias* of the Gini index $\hat{G}_R$ estimates from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations across replications. For MCMC, points represent the average posterior mean across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ prior Beta(8, 8) and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2). Sample size: $n = 100$.



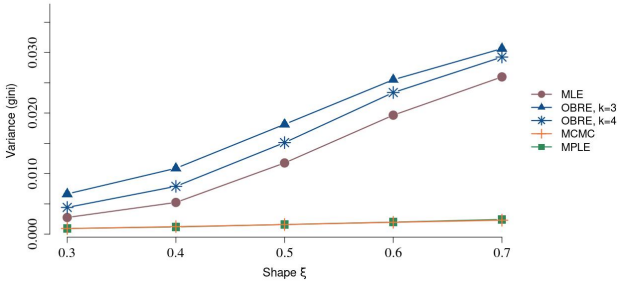
(a) Baseline



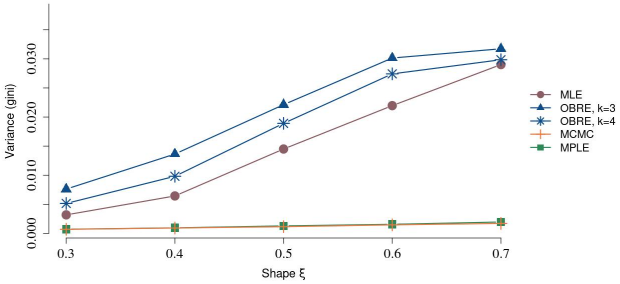
(b) Largest value $\times 10$



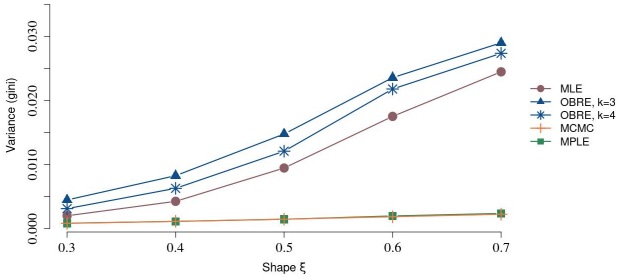
(c) Draw 15% from GPD($\xi = 1.5$)



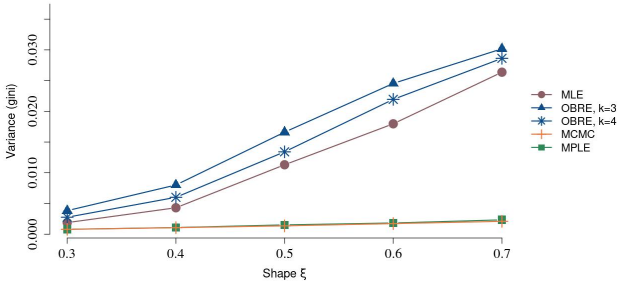
(d) Divide 50% by 2, randomly



(e) Divide 50% by 2, non-randomly

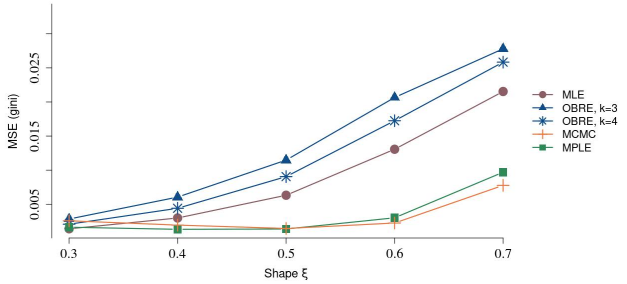


(f) Drop 25%, randomly

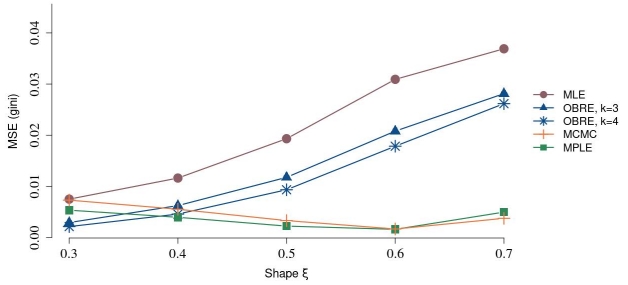


(g) Drop 25%, non-randomly

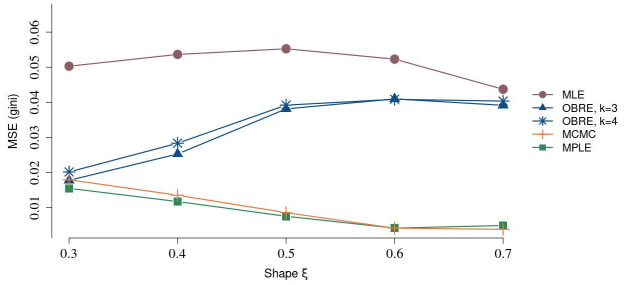
Fig. C2: *Variance* of the Gini index $\hat{G}_R$ estimates from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations across replications. For MCMC, points represent the average posterior mean across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ prior Beta(8, 8) and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2). Sample size: $n = 100$.



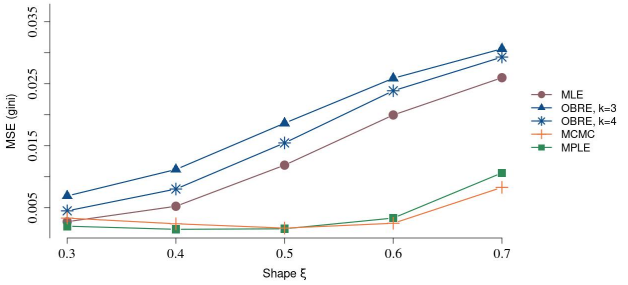
(a) Baseline



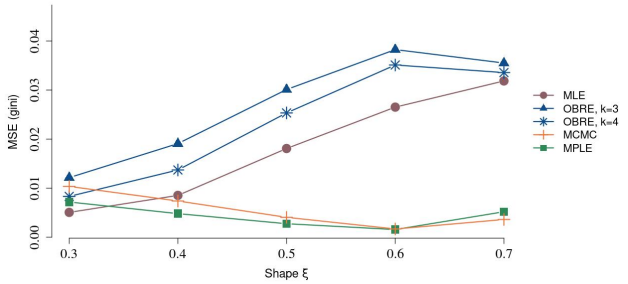
(b) Largest value $\times 10$



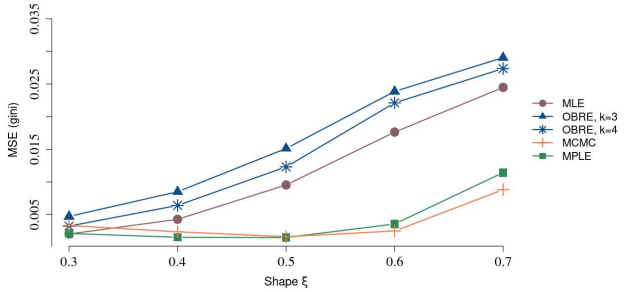
(c) Draw 15% from GPD($\xi = 1.5$)



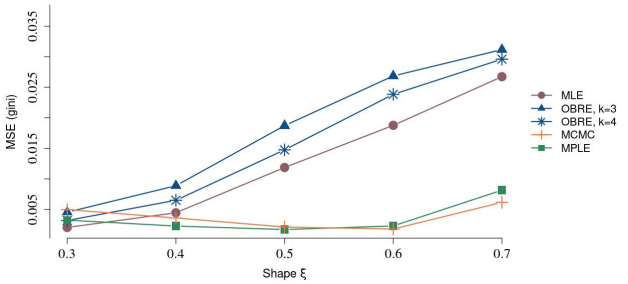
(d) Divide 50% by 2, randomly



(e) Divide 50% by 2, non-randomly



(f) Drop 25%, randomly



(g) Drop 25%, non-randomly

Fig. C3: *MSE* of the Gini index $\hat{G}_R$ estimates from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations across replications. For MCMC, points represent the average posterior mean across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ prior Beta(8, 8) and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2). Sample size: $n = 100$.

Table C2: Bias (B), Variance (V), and MSE (M) of the tail Gini index $\hat{G}_R$ across sample sizes $n \in \{50, 100, 500, 5000\}$. Samples drawn from GPD($\tau = 1$, $\sigma = 0.5$), baseline scenario, 5,000 Monte Carlo replications.

ξ	MLE			OBRE, k=3			OBRE, k=4			MCMC			MPLE		
	B	V	M	B	V	M	B	V	M	B	V	M	B	V	M
$n = 100$															
0.3	0.001	0.001	0.001	0.009	0.003	0.003	0.006	0.002	0.002	0.043	0.001	0.003	0.030	0.001	0.002
0.4	0.002	0.003	0.003	0.012	0.006	0.006	0.008	0.004	0.004	0.031	0.001	0.002	0.018	0.001	0.001
0.5	0.003	0.006	0.006	0.013	0.011	0.011	0.009	0.009	0.009	0.010	0.001	0.001	-0.004	0.001	0.001
0.6	0.011	0.013	0.013	0.021	0.020	0.021	0.018	0.017	0.017	-0.023	0.002	0.002	-0.035	0.002	0.003
0.7	0.008	0.021	0.021	0.010	0.028	0.028	0.012	0.026	0.026	-0.075	0.002	0.008	-0.086	0.002	0.010
$n = 50$															
0.3	0.007	0.003	0.003	0.026	0.008	0.009	0.017	0.005	0.006	0.059	0.001	0.004	0.041	0.001	0.003
0.4	0.008	0.007	0.007	0.025	0.013	0.013	0.020	0.010	0.011	0.040	0.001	0.003	0.022	0.001	0.002
0.5	0.010	0.014	0.014	0.027	0.021	0.021	0.022	0.018	0.018	0.013	0.002	0.002	-0.005	0.002	0.002
0.6	0.009	0.022	0.022	0.013	0.028	0.028	0.013	0.025	0.025	-0.028	0.002	0.003	-0.044	0.002	0.004
0.7	-0.011	0.029	0.029	-0.015	0.033	0.033	-0.008	0.031	0.032	-0.088	0.002	0.010	-0.102	0.003	0.013
$n = 500$															
0.3	-0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.000	0.000	0.013	0.000	0.000	0.009	0.000	0.000
0.4	-0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.011	0.000	0.001	0.006	0.000	0.000
0.5	0.001	0.001	0.001	0.003	0.002	0.002	0.003	0.002	0.002	0.005	0.001	0.001	-0.000	0.001	0.001
0.6	0.002	0.002	0.002	0.005	0.004	0.004	0.004	0.003	0.003	-0.007	0.001	0.001	-0.014	0.001	0.001
0.7	0.005	0.005	0.005	0.011	0.008	0.008	0.009	0.006	0.007	-0.034	0.002	0.003	-0.040	0.002	0.003
$n = 5000$															
0.3	-0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
0.4	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000
0.5	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
0.6	-0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	-0.001	0.000	0.000	-0.002	0.000	0.000
0.7	0.000	0.000	0.000	0.002	0.001	0.001	0.001	0.001	0.001	-0.005	0.000	0.000	-0.006	0.000	0.000

Notes: B = Bias, V = Variance, M = MSE. Bias and MSE for all estimators are computed relative to the Monte Carlo mean estimate across 5,000 replications. For MCMC, the point estimate is the average posterior mean across replications; variance is the sampling variance of the posterior mean estimator across Monte Carlo replications, approximated from the interquartile range of its sampling distribution assuming normality. Unlike the scenario simulations in Table C1, which fix $n = 100$, MCMC here is run separately for each sample size, so that the shrinkage toward the Beta(8, 8) prior diminishes as n grows and both MCMC and MPLE converge to MLE asymptotically. MLE: maximum likelihood estimator. OBRE: optimal bias-robust estimator (Victoria-Feser and Ronchetti 1994), with sensitivity parameters $k \in \{3, 4\}$. MCMC: Bayesian Gibbs sampler with shape prior $\xi \sim$ Beta(8, 8) and uninformative prior for scale σ. MPLE: maximum penalized likelihood estimator with Beta(8, 8) penalty on ξ.

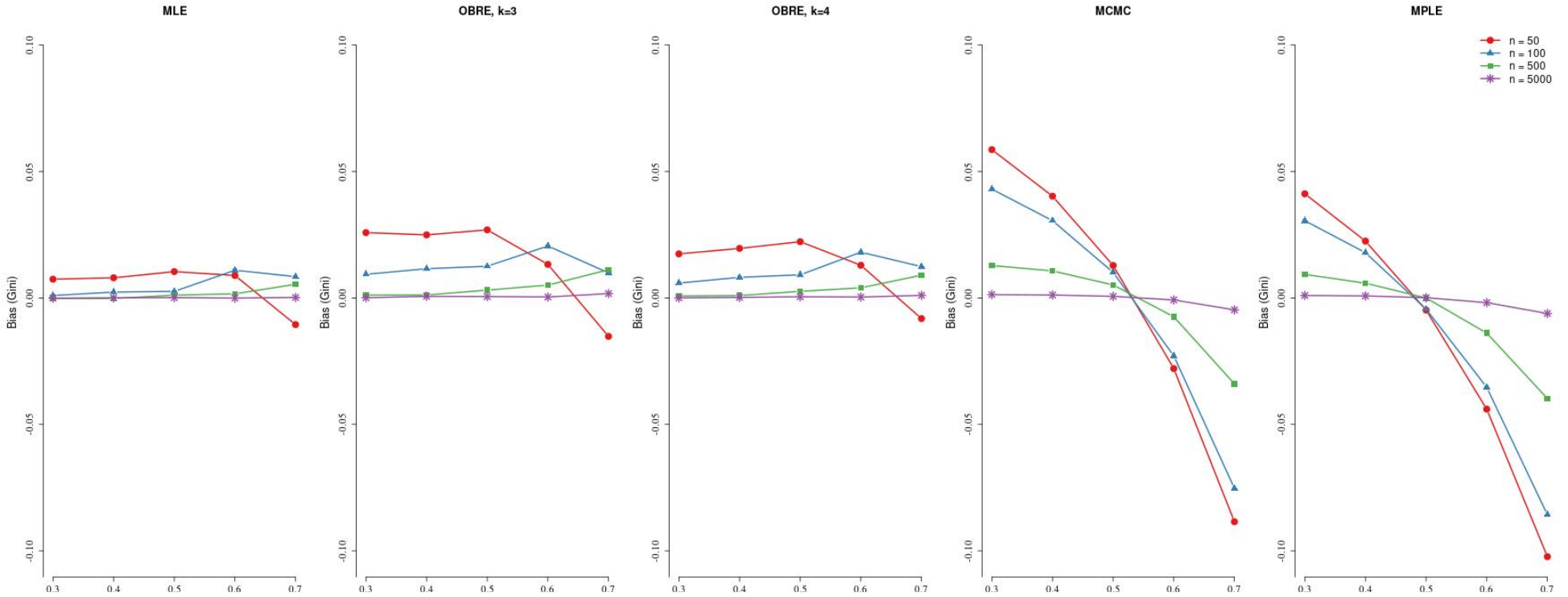


Fig. C4: *Bias* of the Gini index $\hat{G}_R$ estimates across sample sizes from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations. For MCMC, points represent posterior means averaged across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ priorBeta$(8, 8)$ and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2).

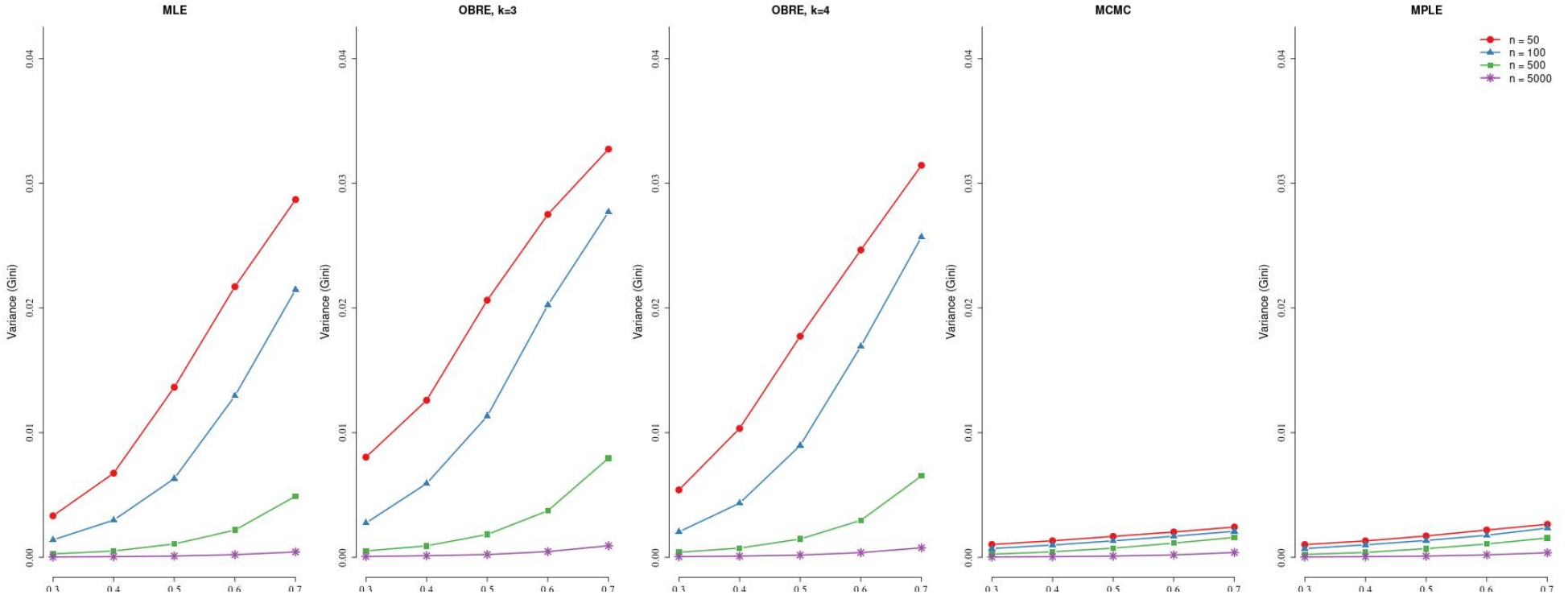


Fig. C5: *Variance* of the Gini index $\hat{G}_R$ estimates across sample sizes from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations. For MCMC, points represent posterior means averaged across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ prior Beta$(8, 8)$ and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2).

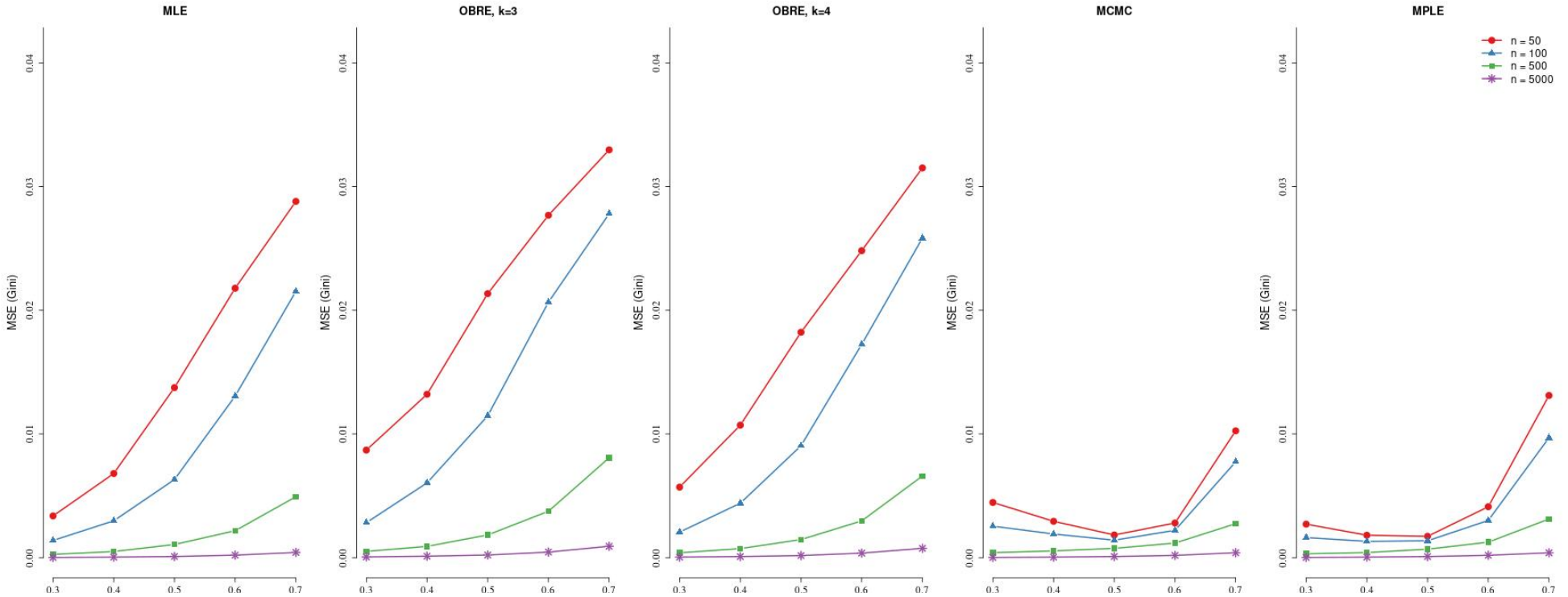


Fig. C6: *MSE* of the Gini index $\hat{G}_R$ estimates across sample sizes from the Monte Carlo simulation randomly drawing 5,000 times from GPD with shapes $\xi \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, location $\tau = 1$ and scale $\sigma = 0.5$. Dashed horizontal lines mark the true Gini value at each shape ξ. For MLE, OBRE, and MPLE, points represent Monte Carlo means and error bars represent ± 1.96 Monte Carlo standard deviations. For MCMC, points represent posterior means averaged across replications and error bars represent the 2.5th and 97.5th percentiles of the posterior mean distribution across replications. MLE is a maximum likelihood estimator. OBRE is an optimal bias-robust estimator (with sensitivity parameters $k \in \{3, 4\}$). MCMC is a Bayesian Markov chain Monte Carlo Gibbs sampler with shape ξ prior Beta$(8, 8)$ and uninformative prior for scale σ. MPLE is the proposed maximum penalized likelihood estimator from eq. (2).

Table C3: Bias (B), Variance (V), and MSE (M) of the $\hat{\xi}$ across seven simulation scenarios. Samples of size $n = 100$ drawn from GPD($\tau = 1$, $\sigma = 0.5$), 5,000 Monte Carlo replications.

ξ	MLE			OBRE, k=3			OBRE, k=4			MCMC			MPLE		
	B	V	M	B	V	M	B	V	M	B	V	M	B	V	M
Baseline															
0.3	-0.024	0.018	0.019	-0.015	0.029	0.029	-0.015	0.024	0.024	0.116	0.003	0.016	0.104	0.003	0.014
0.4	-0.023	0.021	0.022	-0.017	0.033	0.033	-0.015	0.027	0.027	0.056	0.003	0.006	0.047	0.004	0.006
0.5	-0.026	0.023	0.024	-0.023	0.035	0.036	-0.022	0.030	0.030	-0.005	0.004	0.004	-0.012	0.004	0.004
0.6	-0.019	0.027	0.027	-0.014	0.040	0.040	-0.014	0.033	0.033	-0.066	0.004	0.008	-0.068	0.004	0.009
0.7	-0.016	0.031	0.031	-0.011	0.043	0.043	-0.012	0.037	0.037	-0.129	0.004	0.021	-0.127	0.005	0.021
Brazauskas (15% from GPD($\xi = 1.5$))															
0.3	0.301	0.045	0.136	0.154	0.039	0.063	0.178	0.037	0.068	0.246	0.007	0.067	0.243	0.008	0.067
0.4	0.250	0.045	0.108	0.137	0.042	0.060	0.157	0.038	0.063	0.165	0.006	0.033	0.163	0.008	0.034
0.5	0.213	0.042	0.087	0.134	0.042	0.060	0.150	0.038	0.061	0.085	0.005	0.013	0.087	0.006	0.014
0.6	0.175	0.042	0.073	0.111	0.047	0.059	0.126	0.042	0.058	0.006	0.005	0.005	0.009	0.006	0.006
0.7	0.144	0.043	0.063	0.102	0.049	0.059	0.113	0.044	0.057	-0.072	0.004	0.010	-0.067	0.005	0.010
Cowell (largest obs. $\times 10$)															
0.3	0.204	0.007	0.049	-0.013	0.031	0.031	-0.008	0.024	0.024	0.207	0.002	0.044	0.200	0.002	0.042
0.4	0.171	0.010	0.039	-0.014	0.034	0.034	-0.010	0.027	0.027	0.136	0.002	0.020	0.131	0.002	0.019
0.5	0.140	0.013	0.033	-0.020	0.036	0.036	-0.017	0.030	0.030	0.066	0.002	0.006	0.061	0.002	0.006
0.6	0.124	0.017	0.033	-0.011	0.040	0.040	-0.009	0.033	0.033	-0.005	0.002	0.002	-0.005	0.003	0.003
0.7	0.110	0.022	0.034	-0.009	0.043	0.043	-0.008	0.037	0.037	-0.077	0.002	0.008	-0.073	0.003	0.008
Drop 25% non-randomly															
0.3	-0.117	0.039	0.053	-0.091	0.050	0.059	-0.095	0.043	0.052	0.109	0.003	0.015	0.095	0.004	0.013
0.4	-0.093	0.041	0.050	-0.065	0.052	0.056	-0.070	0.044	0.050	0.046	0.004	0.006	0.035	0.004	0.006
0.5	-0.062	0.046	0.050	-0.026	0.054	0.055	-0.035	0.048	0.049	-0.015	0.004	0.004	-0.020	0.005	0.005
0.6	-0.039	0.046	0.048	-0.005	0.057	0.057	-0.014	0.050	0.050	-0.078	0.004	0.010	-0.079	0.005	0.011
0.7	-0.026	0.051	0.052	0.010	0.062	0.062	0.001	0.054	0.054	-0.143	0.004	0.025	-0.142	0.005	0.025
Drop 25% randomly															
0.3	-0.032	0.026	0.027	-0.020	0.040	0.041	-0.020	0.033	0.033	0.130	0.002	0.019	0.119	0.003	0.017
0.4	-0.035	0.028	0.030	-0.026	0.043	0.044	-0.025	0.036	0.036	0.062	0.003	0.006	0.052	0.003	0.006
0.5	-0.029	0.033	0.033	-0.024	0.048	0.049	-0.023	0.040	0.041	-0.006	0.003	0.003	-0.011	0.004	0.004
0.6	-0.030	0.038	0.038	-0.025	0.054	0.054	-0.024	0.045	0.046	-0.074	0.004	0.009	-0.076	0.004	0.010
0.7	-0.030	0.039	0.040	-0.028	0.055	0.056	-0.025	0.046	0.047	-0.142	0.004	0.024	-0.142	0.004	0.025
Halve 50% non-randomly															
0.3	-0.285	0.099	0.181	-0.241	0.121	0.180	-0.248	0.108	0.169	0.103	0.003	0.014	0.086	0.004	0.012
0.4	-0.202	0.104	0.145	-0.144	0.121	0.141	-0.158	0.108	0.133	0.042	0.004	0.006	0.029	0.005	0.006
0.5	-0.118	0.102	0.116	-0.046	0.107	0.110	-0.066	0.100	0.104	-0.020	0.005	0.005	-0.024	0.006	0.007
0.6	-0.055	0.089	0.092	0.009	0.095	0.096	-0.009	0.087	0.087	-0.082	0.005	0.012	-0.082	0.006	0.013
0.7	-0.024	0.086	0.086	0.038	0.092	0.094	0.018	0.084	0.085	-0.147	0.005	0.027	-0.146	0.006	0.027
Halve 50% randomly															
0.3	-0.037	0.034	0.035	-0.022	0.051	0.051	-0.021	0.041	0.042	0.140	0.002	0.021	0.130	0.003	0.019
0.4	-0.038	0.035	0.036	-0.030	0.053	0.054	-0.028	0.043	0.044	0.065	0.003	0.007	0.058	0.003	0.006
0.5	-0.036	0.041	0.043	-0.029	0.059	0.060	-0.027	0.050	0.050	-0.007	0.003	0.003	-0.011	0.004	0.004
0.6	-0.030	0.044	0.045	-0.028	0.063	0.064	-0.025	0.052	0.053	-0.078	0.003	0.009	-0.078	0.004	0.010
0.7	-0.032	0.047	0.048	-0.024	0.064	0.065	-0.023	0.054	0.055	-0.149	0.004	0.026	-0.148	0.004	0.026

Notes: B = Bias, V = Variance, M = MSE. Bias and MSE for all estimators are computed relative to the Monte Carlo mean estimate across 5,000 replications. For MCMC, the point estimate is the average posterior mean across replications; variance is the sampling variance of the posterior mean estimator across Monte Carlo replications, approximated from the interquartile range of its sampling distribution assuming normality. MLE: maximum likelihood estimator. OBRE: optimal bias-robust estimator (Victoria-Feser and Ronchetti 1994), with sensitivity parameters $k \in \{3, 4\}$. MCMC: Bayesian Gibbs sampler with shape prior $\xi \sim \text{Beta}(8, 8)$ and uninformative prior for scale σ. MPLE: maximum penalized likelihood estimator with Beta$(8, 8)$ penalty on ξ.

Table C4: Bias (B), Variance (V), and MSE (M) of $\hat{\xi}$ across sample sizes $n \in \{50, 100, 500, 5000\}$. Samples drawn from GPD($\tau = 1$, $\sigma = 0.5$), baseline scenario, 5,000 Monte Carlo replications.

ξ	MLE			OBRE, k=3			OBRE, k=4			MCMC			MPLE		
	B	V	M	B	V	M	B	V	M	B	V	M	B	V	M
$n = 100$															
0.3	-0.024	0.018	0.019	-0.015	0.029	0.029	-0.015	0.024	0.024	0.116	0.003	0.016	0.104	0.003	0.014
0.4	-0.023	0.021	0.022	-0.017	0.033	0.033	-0.015	0.027	0.027	0.055	0.003	0.006	0.047	0.004	0.006
0.5	-0.026	0.023	0.024	-0.023	0.035	0.036	-0.022	0.030	0.030	-0.005	0.004	0.004	-0.012	0.004	0.004
0.6	-0.019	0.027	0.027	-0.014	0.040	0.040	-0.014	0.033	0.033	-0.066	0.004	0.008	-0.068	0.004	0.009
0.7	-0.016	0.031	0.031	-0.011	0.043	0.043	-0.012	0.037	0.037	-0.129	0.004	0.020	-0.127	0.005	0.021
$n = 50$															
0.3	-0.049	0.041	0.044	-0.032	0.060	0.061	-0.032	0.049	0.050	0.147	0.002	0.023	0.138	0.002	0.021
0.4	-0.047	0.045	0.048	-0.036	0.065	0.067	-0.035	0.054	0.056	0.069	0.002	0.007	0.063	0.003	0.007
0.5	-0.047	0.051	0.053	-0.033	0.072	0.073	-0.032	0.060	0.061	-0.007	0.003	0.003	-0.011	0.003	0.003
0.6	-0.046	0.056	0.058	-0.040	0.079	0.081	-0.037	0.066	0.068	-0.083	0.003	0.010	-0.085	0.004	0.011
0.7	-0.040	0.063	0.065	-0.029	0.084	0.085	-0.028	0.072	0.073	-0.159	0.003	0.029	-0.157	0.004	0.029
$n = 500$															
0.3	-0.004	0.004	0.004	-0.002	0.006	0.006	-0.002	0.005	0.005	0.040	0.002	0.004	0.035	0.002	0.004
0.4	-0.005	0.004	0.004	-0.004	0.007	0.007	-0.003	0.005	0.005	0.021	0.003	0.003	0.015	0.003	0.003
0.5	-0.004	0.005	0.005	-0.003	0.007	0.007	-0.003	0.006	0.006	0.001	0.003	0.003	-0.004	0.003	0.003
0.6	-0.005	0.005	0.005	-0.004	0.008	0.008	-0.003	0.007	0.007	-0.024	0.003	0.004	-0.027	0.003	0.004
0.7	-0.004	0.006	0.006	-0.003	0.009	0.009	-0.003	0.007	0.007	-0.053	0.003	0.005	-0.054	0.003	0.006
$n = 5000$															
0.3	-0.001	0.000	0.000	-0.001	0.001	0.001	-0.001	0.001	0.001	0.005	0.000	0.000	0.004	0.000	0.000
0.4	-0.000	0.000	0.000	0.001	0.001	0.001	-0.000	0.001	0.001	0.002	0.000	0.000	0.002	0.000	0.000
0.5	-0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.001	0.001	0.000	0.000	0.000	-0.000	0.000	0.000
0.6	-0.001	0.001	0.001	-0.001	0.001	0.001	-0.001	0.001	0.001	-0.003	0.001	0.001	-0.004	0.001	0.001
0.7	-0.001	0.001	0.001	0.000	0.001	0.001	-0.000	0.001	0.001	-0.007	0.001	0.001	-0.008	0.001	0.001

Notes: B = Bias, V = Variance, M = MSE. Bias and MSE for all estimators are computed relative to the Monte Carlo mean estimate across 5,000 replications. For MCMC, the point estimate is the average posterior mean across replications; variance is the sampling variance of the posterior mean estimator across Monte Carlo replications, approximated from the interquartile range of its sampling distribution assuming normality. Unlike the scenario simulations in Table C1, which fix $n = 100$, MCMC here is run separately for each sample size, so that the shrinkage toward the Beta$(8, 8)$ prior diminishes as n grows and both MCMC and MPLE converge to MLE asymptotically. MLE: maximum likelihood estimator. OBRE: optimal bias-robust estimator (Victoria-Feser and Ronchetti 1994), with sensitivity parameters $k \in \{3, 4\}$. MCMC: Bayesian Gibbs sampler with shape prior $\xi \sim$ Beta$(8, 8)$ and uninformative prior for scale σ. MPLE: maximum penalized likelihood estimator with Beta$(8, 8)$ penalty on ξ.

Table C5: Bias (B), Variance (V), and MSE (M) of the $\hat{q}_{0.99}$ across seven simulation scenarios. Samples of size $n = 100$ drawn from GPD($\tau = 1$, $\sigma = 0.5$), 5,000 Monte Carlo replications.

ξ	MLE			OBRE, k=3			OBRE, k=4			MCMC			MPLE		
	B	V	M	B	V	M	B	V	M	B	V	M	B	V	M
Baseline															
0.3	0.052	2.513	2.515	0.355	5.224	5.350	0.245	3.772	3.832	2.011	1.458	5.501	1.373	1.290	3.176
0.4	0.174	6.436	6.466	0.597	13.013	13.370	0.443	9.424	9.620	1.613	2.547	5.151	0.902	2.251	3.064
0.5	0.228	15.623	15.675	0.812	30.119	30.778	0.583	22.409	22.749	0.764	4.126	4.710	-0.127	3.652	3.668
0.6	0.839	43.061	43.765	1.872	83.533	87.036	1.413	61.464	63.460	-0.853	6.389	7.117	-1.740	5.829	8.858
0.7	1.929	123.802	127.524	3.485	231.201	243.349	2.771	168.731	176.407	-3.713	9.438	23.223	-4.583	9.317	30.322
Brazauskas (15% from GPD(ξ = 1.5))															
0.3	9.410	159.206	247.751	4.024	31.759	47.950	4.589	38.668	59.727	6.019	9.492	45.714	5.247	9.254	36.783
0.4	10.688	184.462	298.707	5.398	67.307	96.441	5.949	70.021	105.413	5.506	10.688	41.003	4.682	10.847	32.772
0.5	12.527	295.078	452.015	7.790	152.763	213.441	8.295	150.851	219.665	4.484	12.529	32.639	3.681	12.513	26.066
0.6	15.175	557.352	787.636	10.345	354.338	461.367	10.826	324.577	441.785	2.646	14.945	21.948	1.845	15.179	18.582
0.7	18.035	766.456	1091.720	14.563	723.268	935.361	14.793	675.158	893.992	-0.479	17.953	18.182	-1.362	17.994	19.848
Cowell (largest obs. $\times$10)															
0.3	3.316	4.288	15.285	0.350	5.443	5.566	0.256	3.854	3.920	3.785	1.730	16.056	3.062	1.569	10.945
0.4	4.005	10.920	26.957	0.612	13.297	13.672	0.459	9.597	9.808	3.440	2.815	14.649	2.652	2.572	9.604
0.5	4.806	26.357	49.459	0.841	30.581	31.288	0.608	22.829	23.199	2.616	4.399	11.243	1.658	3.999	6.749
0.6	6.626	71.534	115.433	1.916	84.719	88.388	1.460	62.847	64.980	0.997	6.677	7.670	0.064	6.167	6.171
0.7	9.439	201.635	290.737	3.544	234.095	246.654	2.841	172.224	180.293	-1.883	9.671	13.215	-2.777	9.668	17.381
Drop 25% non-randomly															
0.3	0.067	4.182	4.187	0.596	8.661	9.017	0.404	6.191	6.355	3.089	2.140	11.680	2.204	1.812	6.671
0.4	0.371	10.563	10.700	1.258	22.710	24.291	0.898	15.789	16.595	2.756	3.453	11.046	1.745	2.955	5.999
0.5	1.355	33.665	35.500	2.981	66.905	75.790	2.309	47.090	52.420	1.957	5.357	9.188	0.953	5.044	5.952
0.6	2.629	89.972	96.882	5.475	228.404	258.376	4.264	145.851	164.034	0.362	8.383	8.515	-0.765	7.405	7.991
0.7	5.029	250.573	275.863	9.650	573.186	666.305	7.639	386.767	445.120	-2.496	12.284	18.516	-3.681	11.489	25.036
Drop 25% randomly															
0.3	0.093	3.711	3.720	0.512	8.435	8.697	0.361	5.870	6.000	2.367	1.789	7.393	1.606	1.590	4.169
0.4	0.154	9.030	9.053	0.733	19.545	20.082	0.506	13.565	13.821	1.884	2.984	6.533	0.999	2.580	3.577
0.5	0.558	25.438	25.750	1.375	52.415	54.305	1.041	37.389	38.473	0.913	4.736	5.570	-0.034	4.050	4.051
0.6	1.134	68.024	69.309	2.450	138.010	144.010	1.894	99.355	102.944	-0.857	7.161	7.896	-1.925	6.564	10.270
0.7	1.942	166.371	170.141	3.710	318.758	332.521	2.953	231.745	240.464	-3.898	10.686	25.882	-5.063	9.812	35.447
Halve 50% non-randomly															
0.3	0.105	7.550	7.561	0.989	19.271	20.248	0.642	12.790	13.202	5.108	2.894	28.986	3.767	2.312	16.505
0.4	0.902	20.649	21.462	2.741	57.958	65.473	1.951	35.166	38.971	4.794	4.494	27.475	3.322	3.620	14.656
0.5	2.947	69.355	78.039	6.520	190.084	232.592	4.931	114.674	138.990	3.987	6.865	22.763	2.548	5.868	12.362
0.6	5.646	160.460	192.332	11.487	509.514	641.464	8.759	282.956	359.670	2.399	10.300	16.055	0.889	8.610	9.400
0.7	10.158	488.081	591.255	19.420	1459.439	1836.588	14.951	840.707	1064.246	-0.403	14.781	14.943	-1.980	13.091	17.009
Halve 50% randomly															
0.3	-0.037	4.800	4.802	0.497	12.230	12.477	0.295	7.944	8.031	2.364	2.012	7.599	1.528	1.699	4.034
0.4	0.118	11.398	11.412	0.781	26.479	27.088	0.517	17.421	17.689	1.891	3.300	6.876	0.953	2.666	3.574
0.5	0.695	32.574	33.056	1.717	70.799	73.749	1.283	48.592	50.237	0.989	5.344	6.323	-0.044	4.499	4.501
0.6	1.837	91.523	94.900	3.365	222.985	234.306	2.599	131.142	137.894	-0.641	8.284	8.694	-1.699	7.208	10.095
0.7	3.429	234.219	245.977	6.072	507.973	544.844	4.870	339.752	363.472	-3.444	12.473	24.333	-4.643	10.986	32.544

Notes: B = Bias, V = Variance, M = MSE. Bias and MSE for all estimators are computed relative to the Monte Carlo mean estimate across 5,000 replications. For MCMC, the point estimate is the average posterior mean across replications; variance is the sampling variance of the posterior mean estimator across Monte Carlo replications, approximated from the interquartile range of its sampling distribution assuming normality. MLE: maximum likelihood estimator. OBRE: optimal bias-robust estimator (Victoria-Feser and Ronchetti 1994), with sensitivity parameters $k \in \{3, 4\}$. MCMC: Bayesian Gibbs sampler with shape prior $\xi \sim$ Beta(8, 8) and uninformative prior for scale σ. MPLE: maximum penalized likelihood estimator with Beta(8, 8) penalty on ξ.

Table C6: Bias (B), Variance (V), and MSE (M) of $\hat{q}_{0.99}$ across sample sizes $n \in \{50, 100, 500, 5000\}$. Samples drawn from GPD($\tau = 1$, $\sigma = 0.5$), baseline scenario, 5,000 Monte Carlo replications.

ξ	MLE			OBRE, k=3			OBRE, k=4			MCMC			MPLE		
	B	V	M	B	V	M	B	V	M	B	V	M	B	V	M
$n = 100$															
0.3	0.052	2.513	2.515	0.355	5.224	5.350	0.245	3.772	3.832	2.006	1.433	5.458	1.373	1.290	3.176
0.4	0.174	6.436	6.466	0.597	13.013	13.370	0.443	9.424	9.620	1.608	2.534	5.118	0.902	2.251	3.064
0.5	0.228	15.623	15.675	0.812	30.119	30.778	0.583	22.409	22.749	0.756	4.084	4.656	-0.127	3.652	3.668
0.6	0.839	43.061	43.765	1.872	83.533	87.036	1.413	61.464	63.460	-0.862	6.309	7.053	-1.740	5.829	8.858
0.7	1.929	123.802	127.524	3.485	231.201	243.349	2.771	168.731	176.407	-3.724	9.491	23.358	-4.583	9.317	30.322
$n = 50$															
0.3	0.114	5.947	5.960	0.745	15.686	16.241	0.489	9.783	10.022	2.866	2.479	10.691	1.890	1.995	5.567
0.4	0.326	16.000	16.106	1.170	36.821	38.189	0.830	25.177	25.866	2.265	3.833	8.961	1.173	3.137	4.514
0.5	0.804	45.586	46.233	2.270	108.353	113.504	1.670	72.719	75.509	1.139	5.931	7.229	-0.067	4.883	4.887
0.6	1.716	126.136	129.082	3.744	308.110	322.130	2.849	198.714	206.833	-0.824	8.911	9.590	-2.062	7.733	11.985
0.7	3.988	424.029	439.932	7.315	915.481	968.986	5.804	634.975	668.663	-4.087	12.857	29.559	-5.345	11.528	40.092
$n = 500$															
0.3	0.008	0.459	0.459	0.074	0.878	0.884	0.054	0.685	0.688	0.574	0.432	0.761	0.413	0.397	0.568
0.4	0.010	1.046	1.046	0.090	1.930	1.938	0.072	1.550	1.555	0.532	0.928	1.211	0.286	0.815	0.896
0.5	0.080	2.696	2.702	0.198	4.642	4.681	0.164	3.692	3.719	0.308	1.871	1.965	0.022	1.761	1.762
0.6	0.105	6.368	6.379	0.301	10.745	10.836	0.235	8.538	8.594	-0.321	3.471	3.574	-0.711	3.220	3.726
0.7	0.298	15.497	15.586	0.620	24.903	25.286	0.496	20.428	20.674	-1.793	5.790	9.004	-2.173	5.443	10.165
$n = 5000$															
0.3	-0.002	0.042	0.042	0.008	0.111	0.111	0.003	0.085	0.085	0.059	0.042	0.045	0.043	0.041	0.043
0.4	0.005	0.105	0.105	0.034	0.259	0.260	0.014	0.200	0.200	0.057	0.101	0.104	0.038	0.102	0.104
0.5	0.009	0.251	0.251	0.034	0.556	0.557	0.028	0.443	0.443	0.037	0.246	0.248	0.008	0.240	0.240
0.6	-0.002	0.618	0.618	0.025	1.339	1.340	0.022	1.095	1.096	-0.035	0.559	0.561	-0.097	0.570	0.579
0.7	0.008	1.405	1.405	0.095	3.007	3.017	0.055	2.468	2.471	-0.256	1.279	1.344	-0.342	1.205	1.322

Notes: B = Bias, V = Variance, M = MSE. Bias and MSE for all estimators are computed relative to the Monte Carlo mean estimate across 5,000 replications. For MCMC, the point estimate is the average posterior mean across replications; variance is the sampling variance of the posterior mean estimator across Monte Carlo replications, approximated from the interquartile range of its sampling distribution assuming normality. Unlike the scenario simulations in Table C1, which fix $n = 100$, MCMC here is run separately for each sample size, so that the shrinkage toward the Beta(8, 8) prior diminishes as n grows and both MCMC and MPLE converge to MLE asymptotically. MLE: maximum likelihood estimator. OBRE: optimal bias-robust estimator (Victoria-Feser and Ronchetti 1994), with sensitivity parameters $k \in \{3, 4\}$. MCMC: Bayesian Gibbs sampler with shape prior $\xi \sim$ Beta(8, 8) and uninformative prior for scale σ. MPLE: maximum penalized likelihood estimator with Beta(8, 8) penalty on ξ.

Appendix D Data and summary statistics

Table D1: CHFS sample size.

	CHFS 2013	CHFS 2015	CHFS 2017	CHFS 2019
Number of provinces	29	29	29	29
Number of counties	267	351	355	343
Number of communities (villages)	1,048	1,396	1,439	1,360
Number of households	28,143	37,289	40,011	34,643

Table D2: Summary statistics of annual household income per capita (CN¥): CHFS (baseline)

		2012			2014			2016			2018		
		Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD
Shanghai	SH	1,378	40,598	60,536	1,378	49,336	130,484	1,896	60,299	143,178	1,171	63,869	168,210
Guangdong	GD	1,320	35,739	65,614	2,940	42,073	156,265	2,876	44,985	111,214	1,855	40,316	171,977
Liaoning	LN	1,030	17,475	27,995	1,805	29,736	43,146	2,205	31,756	36,580	1,369	28,817	36,192
Henan	HEN	1,026	12,389	17,114	1,009	18,966	63,047	1,122	21,523	29,770	1,327	20,611	29,849
Gansu	GS	764	11,864	14,847	759	18,990	23,041	810	27,676	44,671	851	19,843	40,043
Full sample		27,120	21,735	48,902	35,187	26,157	76,339	39,221	32,766	74,658	33,637	27,497	73,730

Table D3: Summary statistics of annual household income per capita (CN¥): CFPS

		2011			2013			2015			2017		
		Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD	Obs.	Mean	SD
Shanghai	SH	935	30,266	87,582	951	46,442	110,260	919	59,491	149,182	853	70,390	178,837
Guangdong	GD	985	12,966	17,505	1,118	19,274	21,329	1,321	31,362	58,224	1,463	47,772	77,096
Liaoning	LN	1,256	15,045	16,822	1,320	18,044	15,474	1,354	26,216	34,358	1,285	34,009	39,273
Henan	HEN	1,314	10,977	22,359	1,444	13,444	15,474	1,514	15,270	21,681	1,505	22,560	28,658
Gansu	GS	1,328	10,152	21,318	1,371	13,548	30,999	1,507	15,763	15,447	1,592	24,042	32,037
Full sample		11,842	12,251	21,743	12,698	15,926	47,397	13,895	25,039	51,939	14,182	34,491	78,877

Table D4: Income Components in CHFS and CFPS

Income category	CHFS	CFPS
Labor income	Total after-tax wages and salaries, after-tax bonuses, and after-tax work-related subsidies, including both monetary subsidies and in-kind subsidies converted to their monetary equivalents.	Post-tax wages, bonuses, and non-cash benefits from agricultural and non-agricultural employment.
Business income	Income from agricultural production and household business operations. Net agricultural income is defined as the total value of agricultural output plus agricultural subsidies, minus labor and other production costs. Net business income refers to after-tax profits from household business activities after deducting operating costs.	Net income from family production in farming, forestry, animal husbandry, fishing, and sidelines; the value of agricultural products consumed by the household; and net profits from self-employment or privately operated enterprises.
Asset income	Net rental income from land, property, and vehicles; after-tax interest income from deposits, stocks, funds, bonds, financial derivatives and products, foreign currency holdings, and gold; dividends from commercial life, endowment, and property insurance; and interest income from lending.	Income from renting land, real estate, and other means of production.
Transfer income	Retirement benefits and pensions; unemployment insurance; withdrawals from the housing provident fund; transfers to households in extreme poverty; child-related allowances; disability and survivor compensation; relief payments and funds; food subsidies; other government transfers; and private transfers.	Government transfer payments (e.g., pensions, subsidies, and alms) and social donations.
Other income	Lottery winnings and income from the sale of property, intellectual property rights, vehicles, and other assets.	Financial support from family or friends, as well as gifts and cash transfers.

Source: Compiled by the author from the official survey documentation: CHFS (https://chfs.swufe.edu.cn/sjzx/sjsq.htm) and CFPS (https://www.isss.pku.edu.cn/cfps/en)

Appendix E Rural, urban and regional income inequality

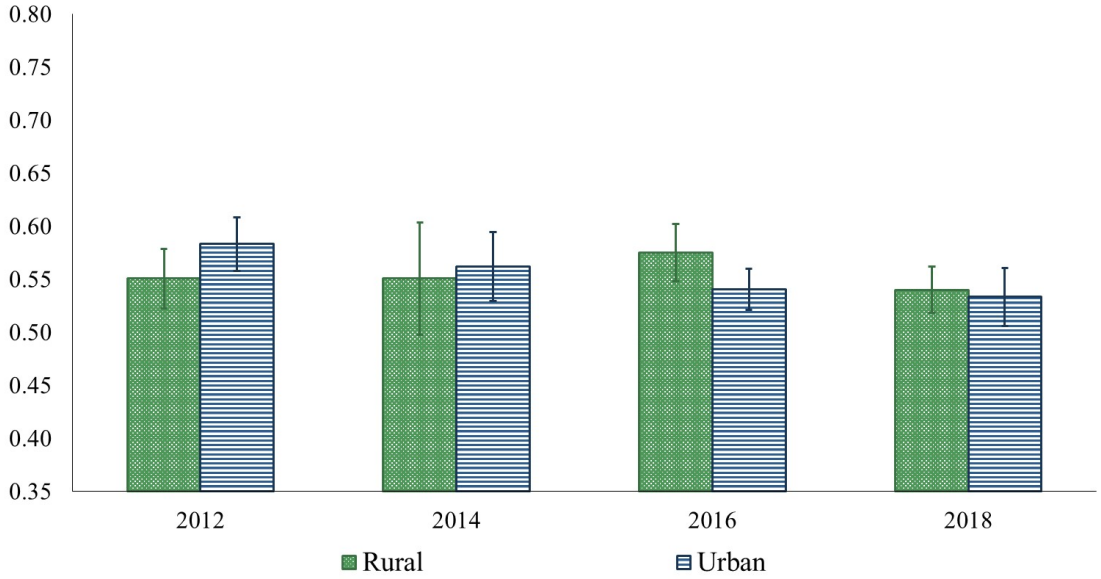


Fig. E1: Rural and urban Gini index estimated from the CHFS data. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

In this section, I analyze income inequality in China at both the rural-urban and regional levels, and assess the relative contributions of within-group versus between-group inequality to overall inequality across these geographic units. Figure E1 shows that, conditioning on confidence intervals, income inequality in urban (0.54–0.58) and rural (0.53–0.58) areas is comparable.[35] Previous studies by Zhang et al. (2012), Zhang (2021), and Li and Sicular (2014) indicate that income inequality in rural areas historically exceeded that in urban areas by 3 to 8 percentage points over extended periods (1978–2010 and 1995–2011, respectively). However, my analysis, which covers more recent years (2012–2018) and uses the CHFS survey, shows different findings. As shown in Figure G3 in Appendix G, both the raw data and baseline point estimates reveal fluctuations in the urban-rural inequality gap, with positive and negative differences that generally do not exceed 5 percentage points. Therefore, the observed similarity is likely due to the use of a different survey (CHFS) and more recent data, rather than the estimation procedure itself. While the proposed method does not substantially alter point estimates at the urban-rural level, it reduces the variance of Gini estimates across survey years by approximately 9%, compared to results from the raw weighted data without any model assumptions on income distribution.

[35] The full set of rural/urban distribution parameters and Gini estimates is reported in Table F2 of Appendix F.

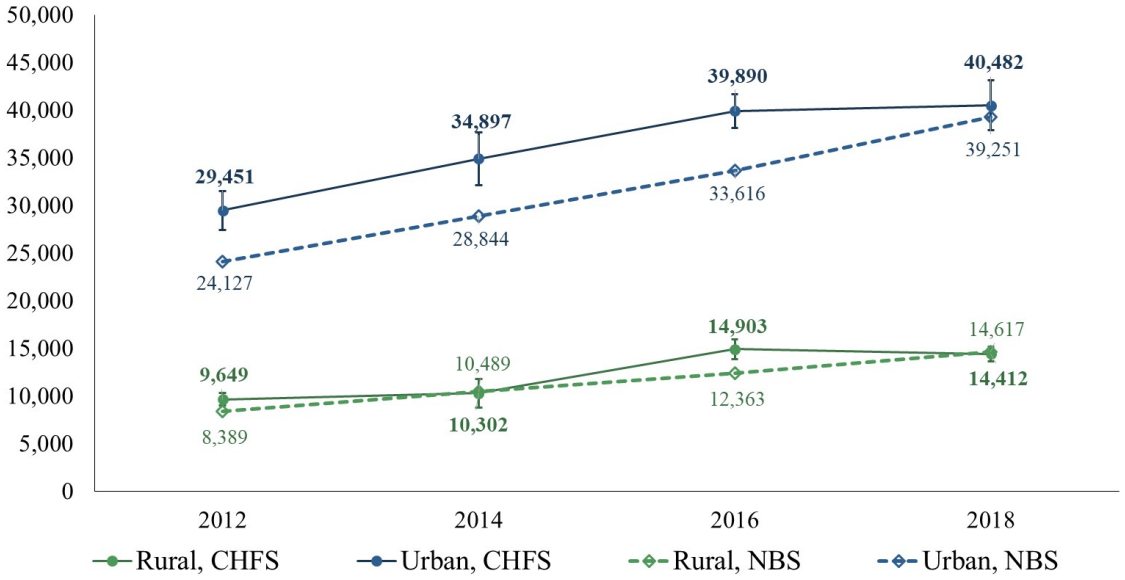


Fig. E2: Comparison of rural and urban estimates of mean household income per capita (CN¥) from the CHFS data with figures from the NBS. For the NBS, values are sourced from the Per Capita Disposable Income data in the China Statistical Yearbook (https://www.stats.gov.cn/english/Statisticaldata/yearbook). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

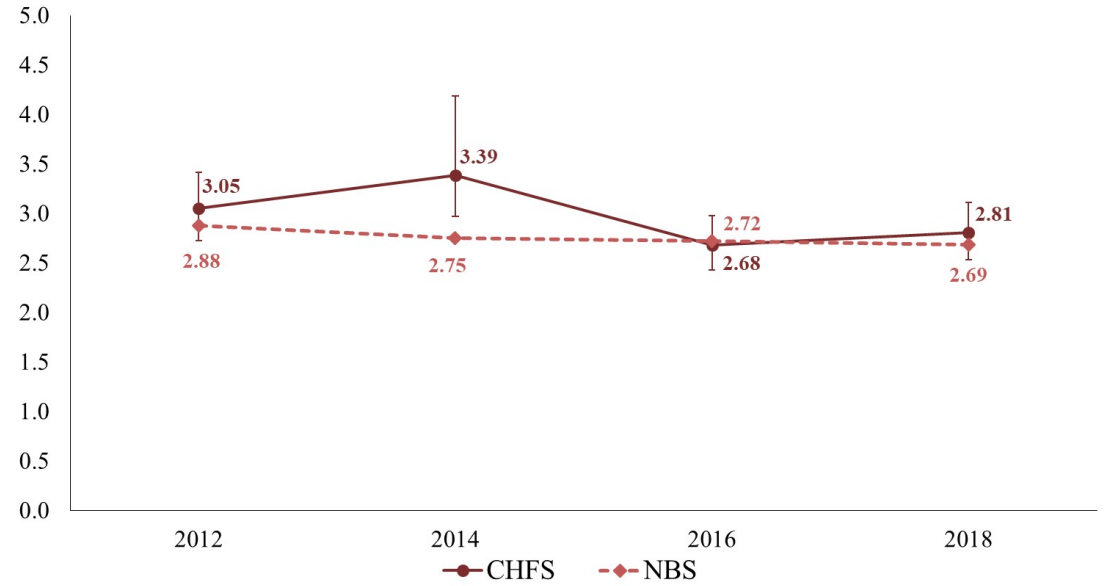


Fig. E3: Comparison of urban-rural income disparity (the ratio of urban to rural mean household income per capita) from the CHFS with the corresponding urban-rural disparity reported by the NBS (China Statistical Yearbook—https://www.stats.gov.cn/english/Statisticaldata/yearbook). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

As shown in Figures E2 and E3, and consistent with official statistics from the NBS, I find that mean income in urban areas 2.7-3.4 times exceeds that in rural areas in China. However, my estimated income levels are slightly higher than those reported by the NBS. The trend in urban-rural income disparity from 2012 to 2018 also aligns closely with the levels reported by the NBS. These findings are consistent with the well-documented rural-urban economic disparities in both income (Wan et al. 2012; Gao et al. 2014; Zhang et al. 2012) and consumption (Yang and Zhou 1999) in the existing literature. Many scholars attribute these rural-urban differences in economic development to the urban bias in government policies that emerged following the initiation of market reforms (Yang 1999; Wang et al. 2014).

It is well-documented that coastal provinces in China have historically been the primary beneficiaries of various government policies, which has contributed to growing regional disparities

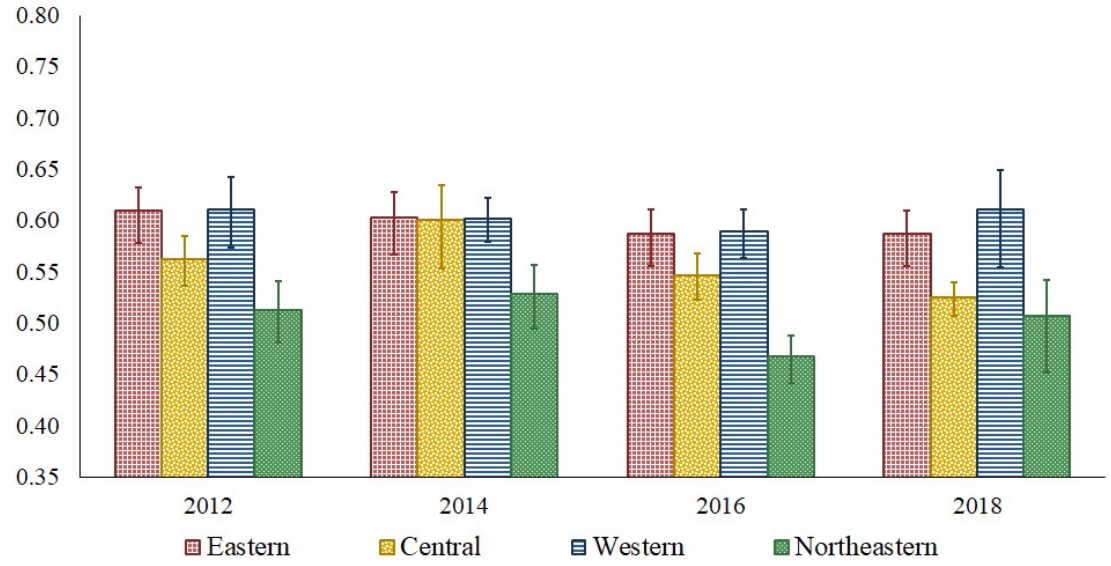


Fig. E4: Regional Gini index estimated from the CHFS data. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals from estimated based on 1,000 samples

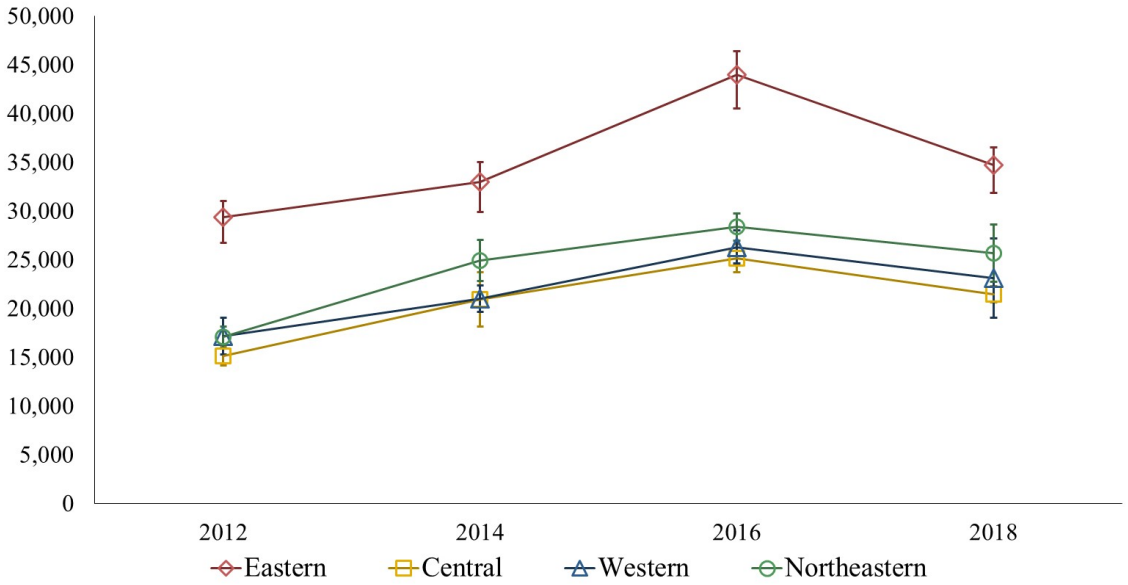


Fig. E5: Regional estimates of mean household income per capita (CN¥) from the CHFS data. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

(Dunford and Li 2010; Fan et al. 2011).[36] The major factors driving these disparities include superior infrastructure, greater human capital, and locational advantages for exports in coastal provinces (Lu 2008), as well as preferential policies in investment, banking, and taxation, which have fostered industrial agglomeration in these regions (Wang et al. 2014). Figure E5 supports this well-established fact that the Eastern region is substantially richer than other Chinese regions (Zhang 2021), with mean income of households being, on average, 1.5-1.8 times higher. However, the results on income disparities should be interpreted with caution. A substantial body of literature underscores that a comprehensive understanding of inequality in China must account for spatial price differences, as the cost of living is notably higher in urban and coastal areas (Brandt

[36]In this section I am employing the official definition of regions in China from NBS (https://www.stats.gov.cn/english/PressRelease/202105/t20210510_1817188.html). According to this classification, the Eastern region includes 10 provinces (Beijing, Tianjin, Hebei, Shanghai, Jiangsu, Zhejiang, Fujian, Shandong, Guangdong, and Hainan), the Central region includes 6 provinces (Shanxi, Anhui, Jiangxi, Henan, Hubei and Hunan), the Western region includes 12 provinces (Inner Mongolia, Guangxi, Chongqing, Sichuan, Guizhou, Yunnan, Tibet, Shaanxi, Gansu, Qinghai, Ningxia and Xinjiang), and Northeastern region includes 3 provinces (Liaoning, Jilin and Heilongjiang). The results on the Western region might be less accurate, given Tibet and Xinjiang are absent in CHFS.

and Holz 2006; Sicular et al. 2007; Ravallion and Chen 2009). Figure E4 shows that the Eastern and Western regions exhibit the highest levels of inequality, while the Northeastern region is the least unequal.[37] These findings are broadly consistent with those of Luo et al. (2020), who, using a Generalized Entropy (GE) decomposition, found the lowest inequality in the Central region (corresponding to Central + Northeastern in the current classification), followed by the Eastern region, and slightly higher inequality in the Western region. However, in contrast to Luo et al. (2020), I cannot conclude that regional income inequality in China is narrowing. Figure G4 in Appendix G compares our results with the raw data estimates. The proposed new estimation procedure helps to reduce outlier-driven variation in Gini across survey years, on 23%, on average.

Finally, to study the relative contribution of within and between urban-rural and regional variation in the national income disparity, I employ a standard Theil T index decomposition formula which goes back to an old economic literature (Bourguignon 1979; Cowell 1980; Shorrocks 1984):

$$T = \sum_{g=1}^{h} s_g T_g + \sum_{g=1}^{h} s_g \ln\left(\frac{s_g}{p_g}\right), \tag{E8}$$

where the first term is the within component (T_g is each subgroup's Theil index, weighted by the subgroup's share in total income $s_g = p_g\bar{y}_g/\bar{y}$), and the second term is the between component. Note that the between-group component is calculated using the mean income of each subgroup $\bar{y}_g$, where $p_g = N_g/N$ is the share of each subgroup's population N_g in the total population N, while the within-group components T_g are computed from individual household incomes.

Table E5: Theil T index decomposition

	CHFS			
	2012	2014	2016	2018
Theil T	0.827	0.868	0.735	0.790
Theil T Rural	0.651	0.778	0.817	0.856
Theil T Urban	0.737	0.756	0.648	0.629
% Within Rural/Urban	87.1%	87.5%	91.2%	87.6%
Theil T Eastern	0.793	0.884	0.764	0.824
Theil T Central	0.665	0.935	0.647	0.561
Theil T Western	0.931	0.746	0.701	0.923
Theil T Northeastern	0.596	0.632	0.483	0.592
% Within Regions	95.5%	97.3%	95.6%	97.2%

The decomposition results are presented in Table E5, which indicate that rural-urban intragroup inequality accounts for approximately 87-91% of the nationwide income inequality. This finding aligns with the conclusions of Wan et al. (2018), based on multiple recent household surveys in China. Additionally, I find that intragroup regional inequality contributes to around 95-97% of nationwide income inequality, a result that is consistent with Luo et al. (2020), who estimate a contribution of 92% in 2013 based on the CHIP data.

[37] The full set of regional distribution parameters and inequality estimates is reported in Table F3, Appendix F.

Appendix F Results: detailed

Table F1: National-level Gini index and estimated distribution parameters

	CHFS (baseline)				CFPS			
	2012	2014	2016	2018	2011	2013	2015	2017
Gini index	0.616	0.604	0.581	0.590	0.540	0.522	0.555	0.579
Lower 95% CI	0.597	0.584	0.565	0.566	0.524	0.501	0.533	0.538
Upper 95% CI	0.634	0.624	0.598	0.614	0.556	0.543	0.577	0.620
Threshold income $\hat{\tau}$	61,788	73,346	90,587	76,443	36,087	45,379	65,800	98,011
GPD shape $\hat{\xi}$	0.622	0.639	0.655	0.687	0.546	0.611	0.665	0.799
GPD scale $\hat{\sigma}$	37,997	41,873	47,577	37,150	16,878	16,881	34,742	34,692
Bandwidth $\hat{h}$	432	392	654	709	411	580	821	898
Mean income (CN¥)	21,587	26,160	33,207	28,080	12,910	16,392	24,900	36,737

Table F2: Gini index and estimated distribution parameters: rural/urban.

	CHFS			
	2012	2014	2016	2018
Gini index				
Rural	0.548	0.574	0.571	0.539
Urban	0.583	0.562	0.543	0.535
Lower 95% CI				
Rural	0.525	0.537	0.546	0.519
Urban	0.560	0.534	0.525	0.510
Upper 95% CI				
Rural	0.571	0.611	0.595	0.560
Urban	0.605	0.590	0.560	0.560
Threshold income $\hat{\tau}$				
Rural	19,964	23,835	29,298	30,029
Urban	52,893	61,928	72,149	70,768
GPD shape $\hat{\xi}$				
Rural	0.523	0.448	0.576	0.491
Urban	0.626	0.653	0.634	0.651
GPD scale $\hat{\sigma}$				
Rural	9,106	12,092	15,001	13,952
Urban	31,800	33,344	36,758	35,321
Bandwidth $\hat{h}$				
Rural	323	271	473	492
Urban	797	861	1,363	1,208
Mean income (CN¥)				
Rural	9,611	10,933	14,812	14,429
Urban	29,475	35,064	40,180	40,657

Table F3: Gini index and estimated distribution parameters: regional-level.

	CHFS			
	2012	2014	2016	2018
Gini index				
Eastern	0.610	0.603	0.587	0.589
Central	0.562	0.602	0.547	0.524
Western	0.629	0.601	0.584	0.621
Northeastern	0.510	0.532	0.469	0.506
Lower 95% CI				
Eastern	0.583	0.571	0.560	0.563
Central	0.537	0.560	0.525	0.508
Western	0.594	0.579	0.561	0.573
Northeastern	0.482	0.501	0.446	0.462
Upper 95% CI				
Eastern	0.637	0.634	0.614	0.615
Central	0.586	0.643	0.570	0.540
Western	0.664	0.622	0.607	0.670
Northeastern	0.539	0.563	0.493	0.550
Threshold income $\hat{\tau}$				
Eastern	56,951	60,724	79,386	66,234
Central	30,957	38,048	48,576	44,532
Western	34,722	43,721	52,214	44,796
Northeastern	31,932	47,527	52,198	48,827
GPD shape $\hat{\xi}$				
Eastern	0.578	0.659	0.675	0.615
Central	0.563	0.652	0.544	0.469
Western	0.692	0.508	0.607	0.681
Northeastern	0.563	0.553	0.550	0.599
GPD scale $\hat{\sigma}$				
Eastern	36,125	32,519	41,130	35,529
Central	14,530	21,128	25,434	19,450
Western	16,776	25,845	24,435	22,462
Northeastern	14,709	24,796	21,206	18,063
Bandwidth $\hat{h}$				
Eastern	854	766	1,174	1,213
Central	567	615	890	875
Western	523	516	778	759
Northeastern	1,075	1,119	1,428	1,367
Mean income (CN¥)				
Eastern	29,534	33,045	44,136	34,953
Central	15,180	21,169	25,192	21,399
Western	18,087	21,015	26,051	24,079
Northeastern	16,975	25,181	28,436	25,743

Table F4: Gini index and estimated distribution parameters: selected provinces.

	CHFS (baseline)				CFPS			
	2012	2014	2016	2018	2011	2013	2015	2017
Gini index								
Shanghai	0.463	0.494	0.441	0.421	0.488	0.477	0.471	0.446
Guangdong	0.686	0.665	0.641	0.624	0.495	0.483	0.529	0.539
Liaoning	0.492	0.526	0.441	0.447	0.476	0.424	0.477	0.460
Henan	0.562	0.611	0.572	0.580	0.518	0.465	0.448	0.504
Gansu	0.559	0.532	0.576	0.548	0.506	0.465	0.447	0.474
Lower 95% CI								
Shanghai	0.430	0.424	0.355	0.333	0.409	0.384	0.371	0.394
Guangdong	0.657	0.594	0.592	0.530	0.469	0.454	0.494	0.488
Liaoning	0.459	0.492	0.414	0.415	0.446	0.402	0.443	0.428
Henan	0.536	0.540	0.542	0.552	0.478	0.437	0.419	0.478
Gansu	0.528	0.501	0.528	0.512	0.451	0.429	0.425	0.424
Upper 95% CI								
Shanghai	0.495	0.563	0.493	0.481	0.556	0.558	0.545	0.490
Guangdong	0.714	0.729	0.681	0.686	0.519	0.513	0.563	0.579
Liaoning	0.525	0.559	0.464	0.475	0.504	0.443	0.510	0.488
Henan	0.587	0.665	0.597	0.605	0.553	0.488	0.472	0.529
Gansu	0.589	0.562	0.624	0.580	0.549	0.494	0.468	0.514
Threshold income $\hat{\tau}$								
Shanghai	71,931	75,478	99,508	103,504	53,247	67,436	100,015	126,185
Guangdong	88,938	74,173	87,128	71,211	28,785	38,950	58,691	98,756
Liaoning	31,877	55,026	57,182	54,335	29,772	35,565	46,307	67,297
Henan	28,115	36,032	44,786	46,388	21,056	26,873	29,303	43,950
Gansu	26,388	39,432	51,237	43,023	19,386	29,171	30,125	48,313
GPD shape $\hat{\xi}$								
Shanghai	0.613	0.771	0.682	0.708	0.720	0.763	0.695	0.538
Guangdong	0.384	0.710	0.620	0.667	0.333	0.496	0.558	0.565
Liaoning	0.546	0.471	0.465	0.655	0.423	0.196	0.513	0.509
Henan	0.428	0.618	0.362	0.417	0.548	0.507	0.358	0.408
Gansu	0.239	0.455	0.505	0.491	0.702	0.412	0.295	0.683
GPD scale $\hat{\sigma}$								
Shanghai	31,591	35,602	43,877	41,641	21,407	31,018	47,397	59,622
Guangdong	66,748	51,909	59,541	52,293	10,158	15,016	33,266	45,866
Liaoning	14,958	35,470	24,659	15,812	12,311	12,976	26,036	23,282
Henan	13,286	22,474	29,748	23,911	9,576	9,109	13,497	25,306
Gansu	13,233	18,584	36,221	16,755	6,671	8,162	15,014	11,484
Bandwidth $\hat{h}$								
Shanghai	4,042	4,158	3,729	4,871	2,791	3,652	3,477	5,313
Guangdong	1,268	1,469	2,053	2,671	1,028	1,544	2,089	2,397
Liaoning	1,527	1,898	2,390	2,281	1,091	1,546	1,882	2,669
Henan	890	1,138	1,295	1,447	708	1,058	1,183	1,565
Gansu	937	1,371	1,717	1,644	565	866	986	1,235
Mean income (CN¥)								
Shanghai	40,410	47,870	59,438	62,767	29,741	44,654	57,915	70,126
Guangdong	35,990	40,641	44,475	39,356	12,980	19,307	31,218	47,463
Liaoning	17,410	29,930	31,773	28,735	15,037	18,058	26,216	33,976
Henan	12,453	18,791	21,690	20,701	10,962	13,479	15,148	22,652
Gansu	11,941	19,062	27,720	19,928	10,054	13,397	15,809	23,780

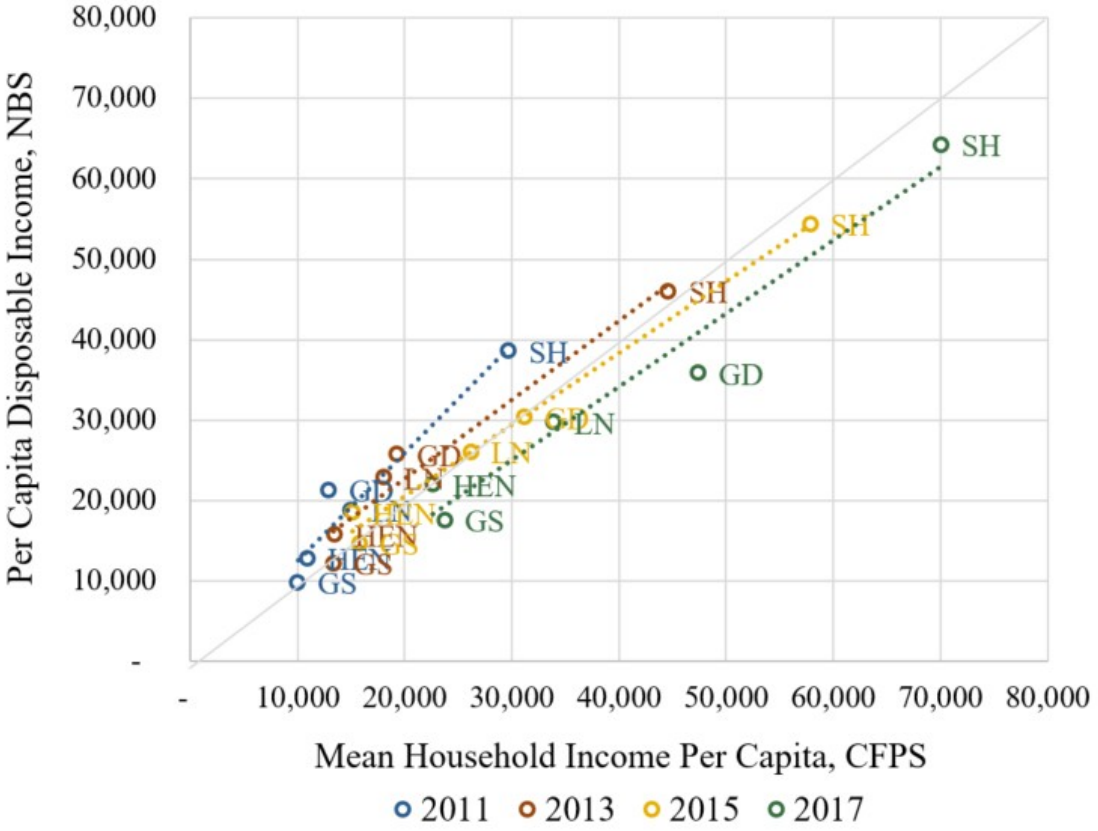


Fig. F1: Baseline estimates of mean household income per capita from the CFPS (X-axis) plotted versus "Per Capita Disposable Income" reported by the NBS (Y-axis) ("China Statistical Yearbook" - https://www.stats.gov.cn/english/Statisticaldata/yearbook)

Appendix G Raw Gini estimates

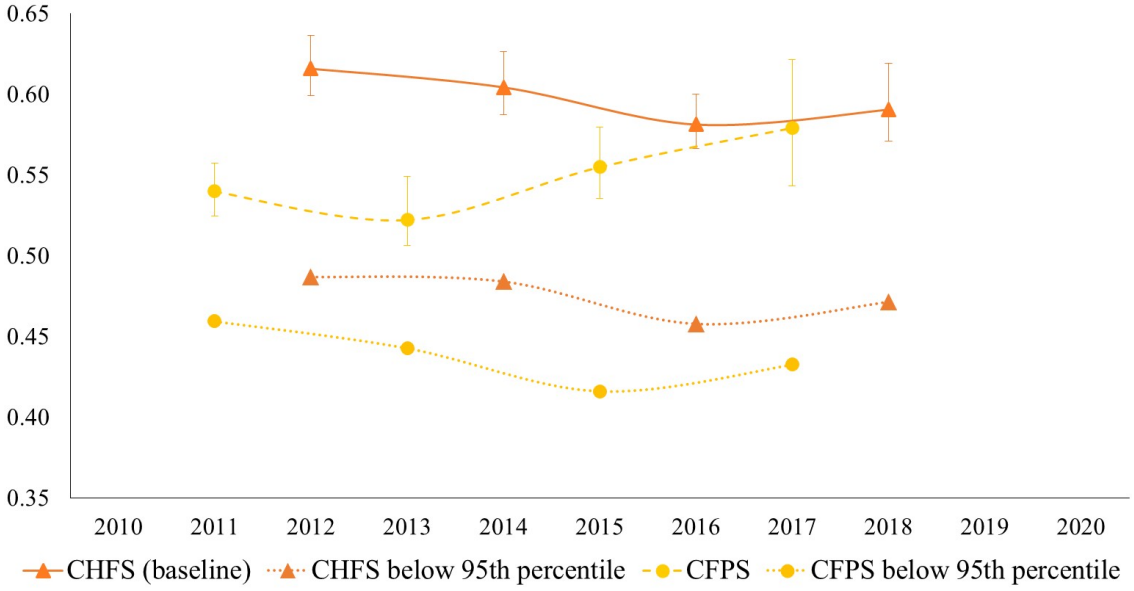


Fig. G1: Comparison of the estimated Gini to the below 95^{th} percentile raw data estimates. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated from 1,000 samples

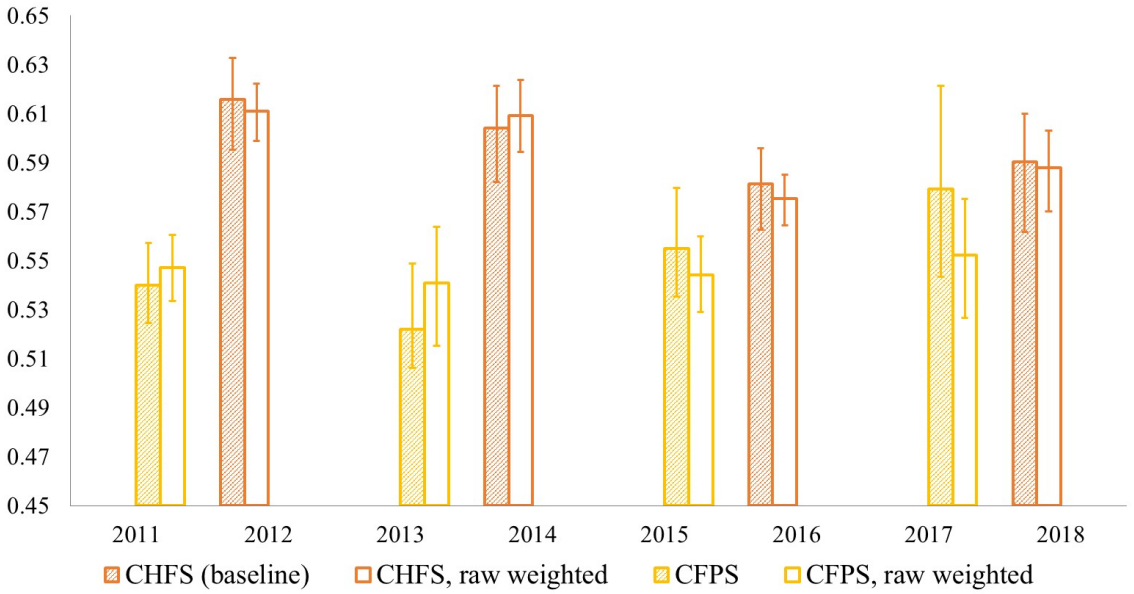
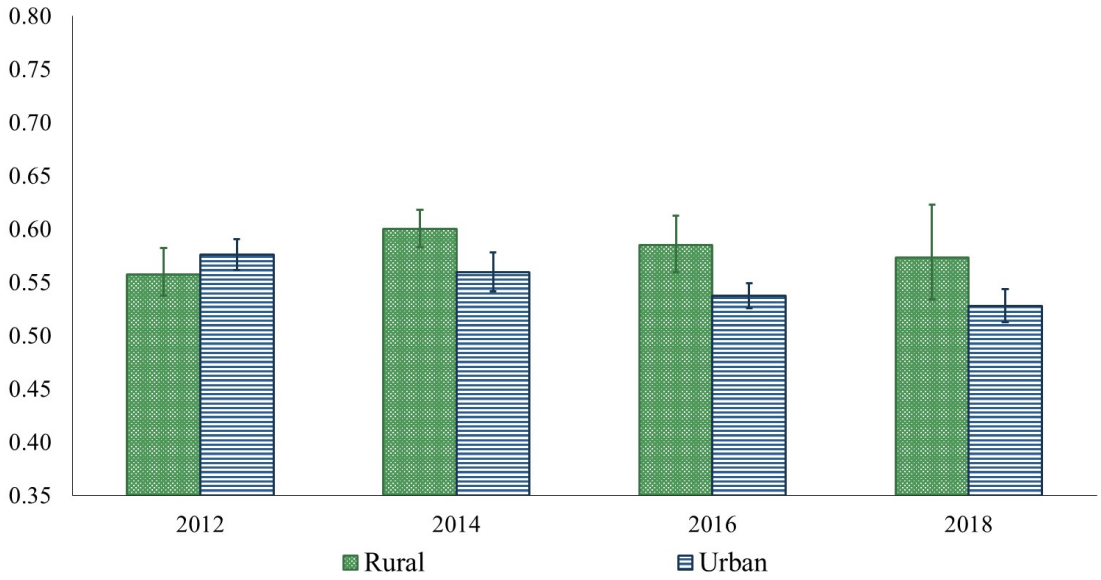
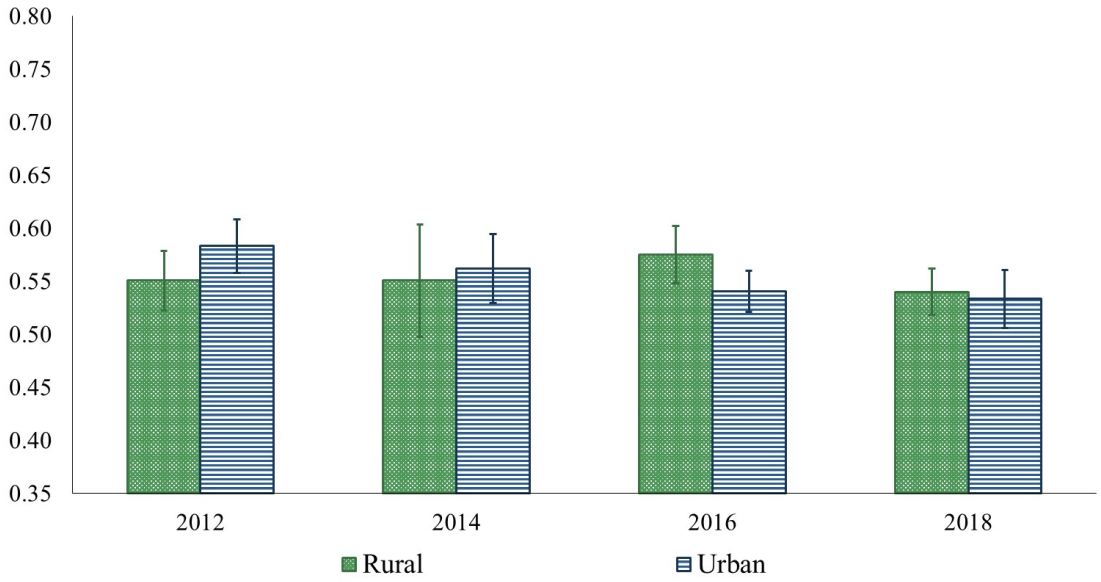


Fig. G2: Comparison between the baseline and raw data Gini index estimates. Error bars represent 95% confidence intervals estimated from 1,000 samples. Basic (reverse percentile) bootstrap confidence intervals from (Davison and Hinkley 1997, p. 194) are used for our estimates, and percentile bootstrap confidence intervals are applied to the raw weighted data results reported here

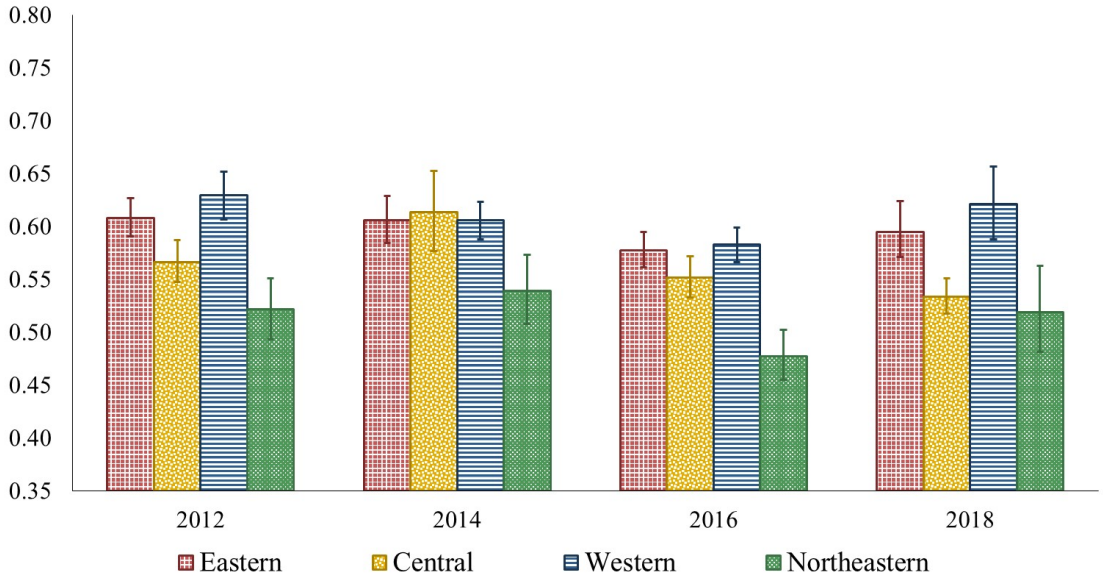


(a) Raw data

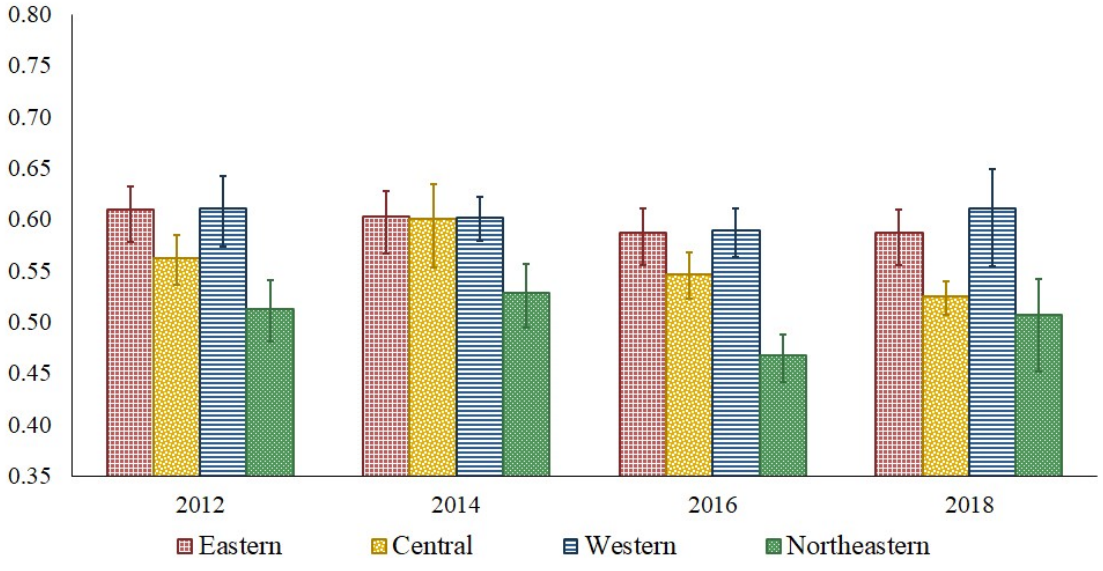


(b) Baseline estimates

Fig. G3: Comparison of our baseline estimated rural/urban Gini indices (b) to the raw weighted data estimates (a). Error bars represent 95% confidence intervals estimated from 1,000 samples. Basic (reverse percentile) bootstrap confidence intervals from (Davison and Hinkley 1997, p. 194) are used for our estimates, and percentile bootstrap confidence intervals are applied to the raw weighted data results reported here

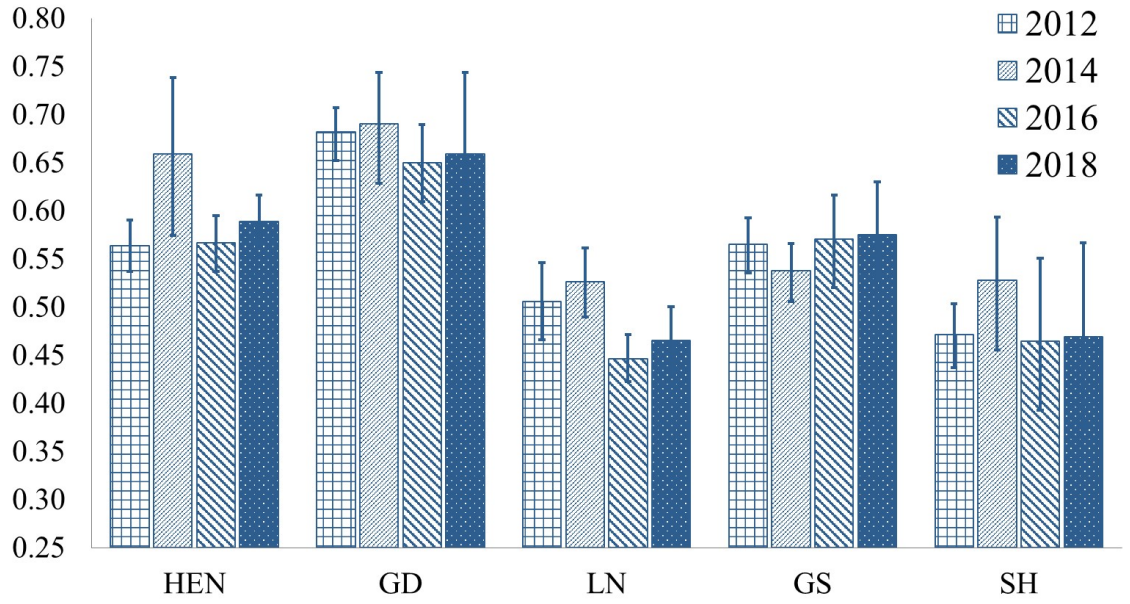


(a) Raw data

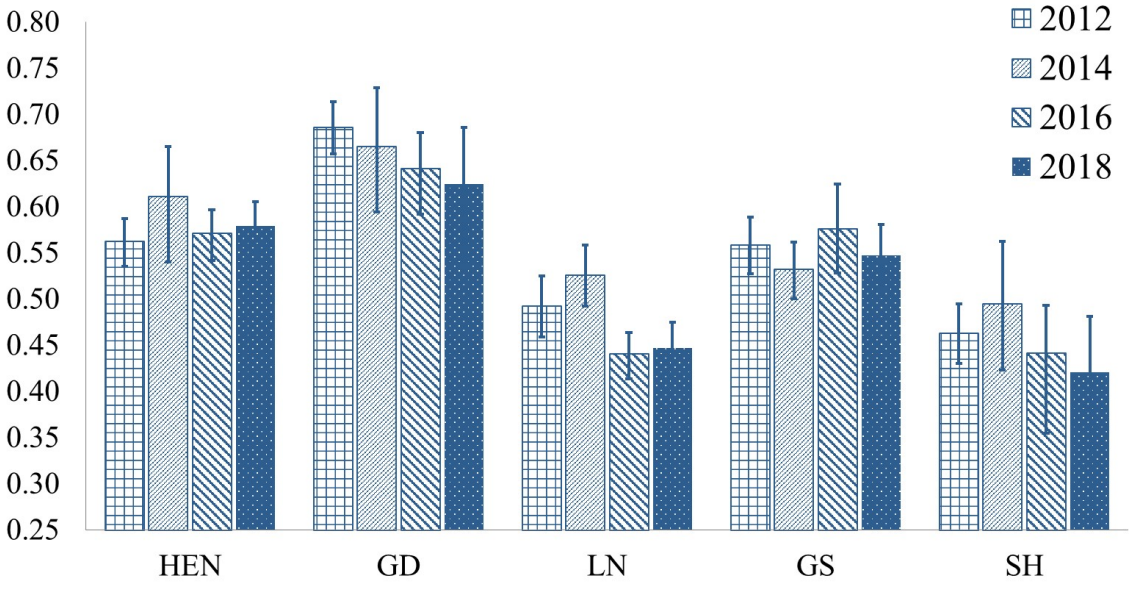


(b) Baseline estimates

Fig. G4: Comparison of our baseline estimated regional Gini indices (b) to the raw weighted data estimates (a). Error bars represent 95% confidence intervals estimated from 1,000 samples. Basic (reverse percentile) bootstrap confidence intervals from (Davison and Hinkley 1997, p. 194) are used for our estimates, and percentile bootstrap confidence intervals are applied to the raw weighted data results reported here

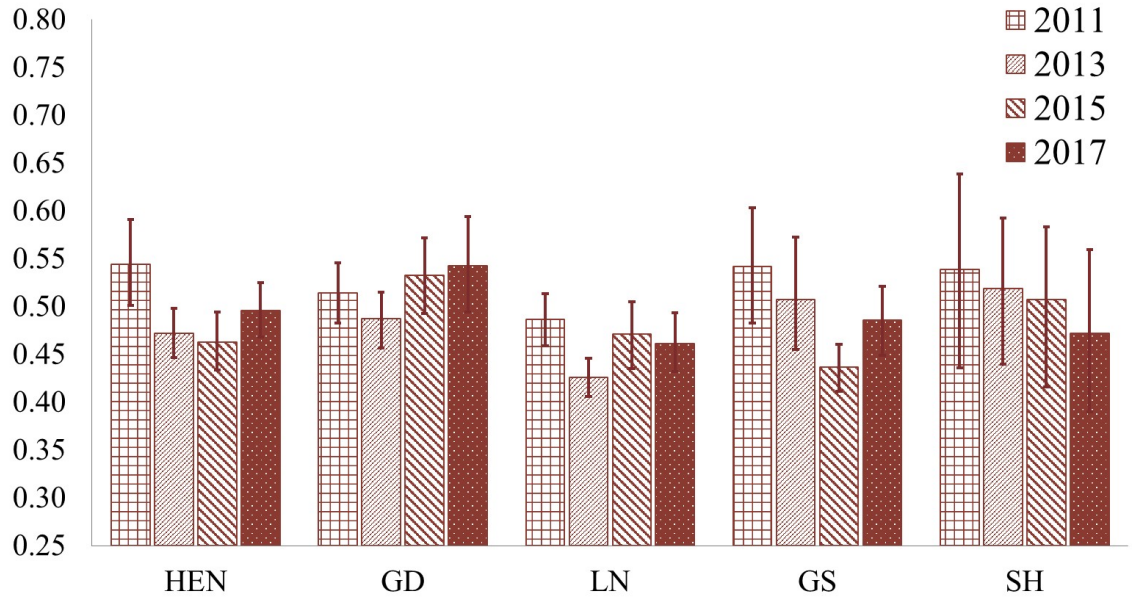


(a) Raw data

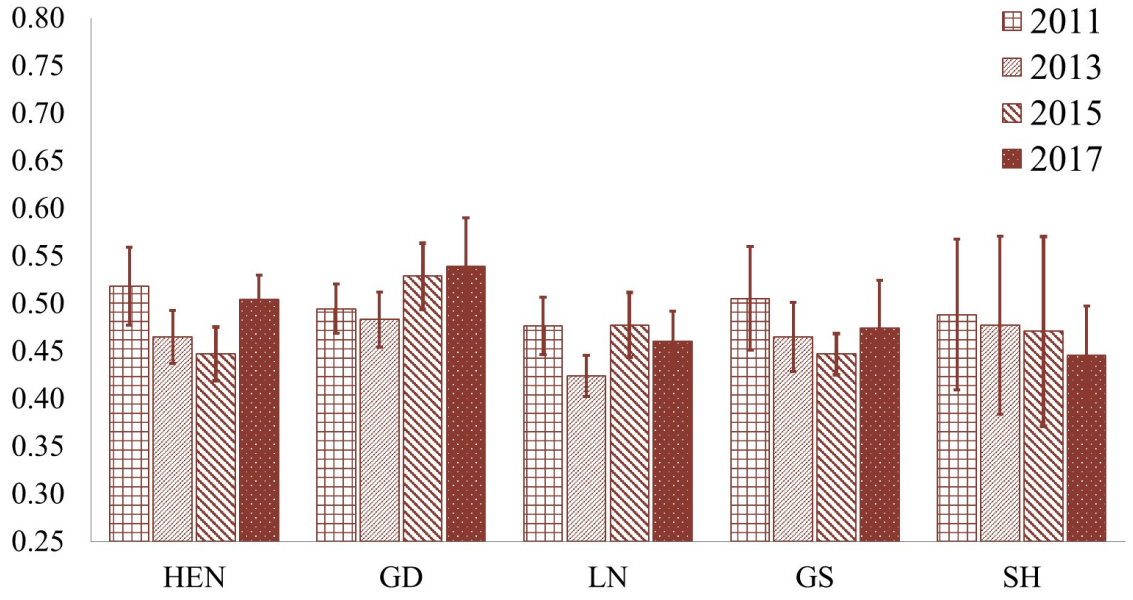


(b) Baseline estimates

Fig. G5: Comparison of our baseline estimated provincial Gini indices (b) to the raw weighted data estimates (a) based on the data from CHFS. Error bars represent 95% confidence intervals estimated from 1,000 samples. Basic (reverse percentile) bootstrap confidence intervals from (Davison and Hinkley 1997, p. 194) are used for our estimates, and percentile bootstrap confidence intervals are applied to the raw weighted data results reported here



(a) Raw data



(b) Baseline estimates

Fig. G6: Comparison of our estimated provincial Gini indices (b) to the raw weighted data estimates (a) based on the data from CFPS. Error bars represent 95% confidence intervals estimated from 1,000 samples. Basic (reverse percentile) bootstrap confidence intervals from (Davison and Hinkley 1997, p. 194) are used for our estimates, and percentile bootstrap confidence intervals are applied to the raw weighted data results reported here

Appendix H Provincial-level mean income

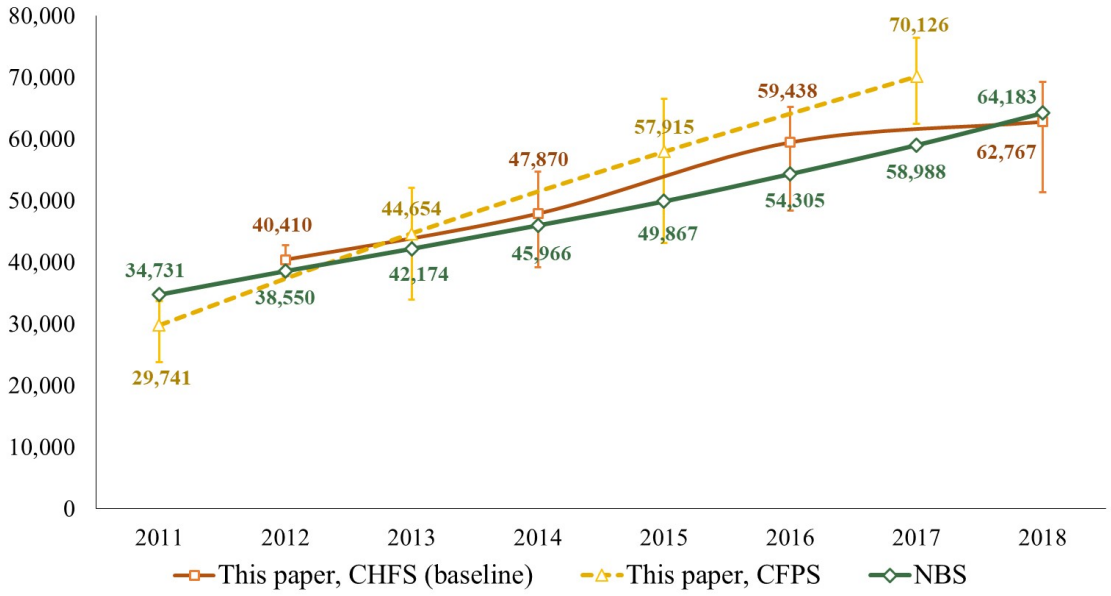


Fig. H1: Comparison of estimated mean household income per capita (CN¥) in Shanghai using three data sources: CHFS (baseline), CFPS, and NBS. For the NBS, values are taken from Per Capita Disposable Income as reported in the China Statistical Yearbook (https://www.stats.gov.cn/english/Statisticaldata/yearbook). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

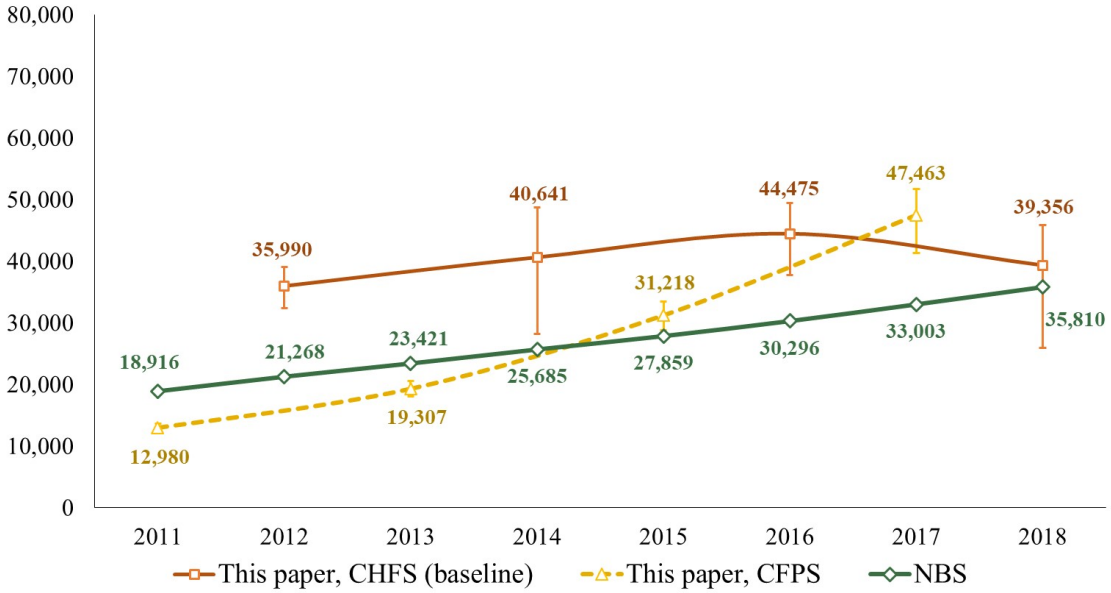


Fig. H2: Comparison of estimated mean household income per capita (CN¥) in Guangdong using three data sources: CHFS (baseline), CFPS, and NBS. For the NBS, values are taken from Per Capita Disposable Income as reported in the China Statistical Yearbook (https://www.stats.gov.cn/english/Statisticaldata/yearbook). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

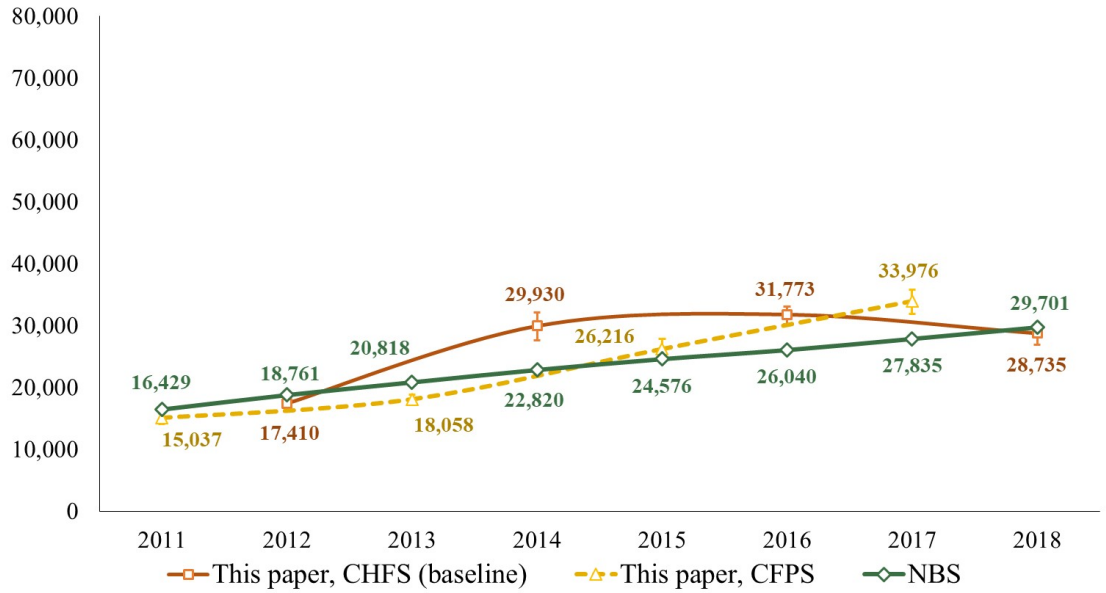


Fig. H3: Comparison of estimated mean household income per capita (CN¥) in Liaoning using three data sources: CHFS (baseline), CFPS, and NBS. For the NBS, values are taken from Per Capita Disposable Income as reported in the China Statistical Yearbook (https://www.stats.gov.cn/english/Statisticaldata/yearbook). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

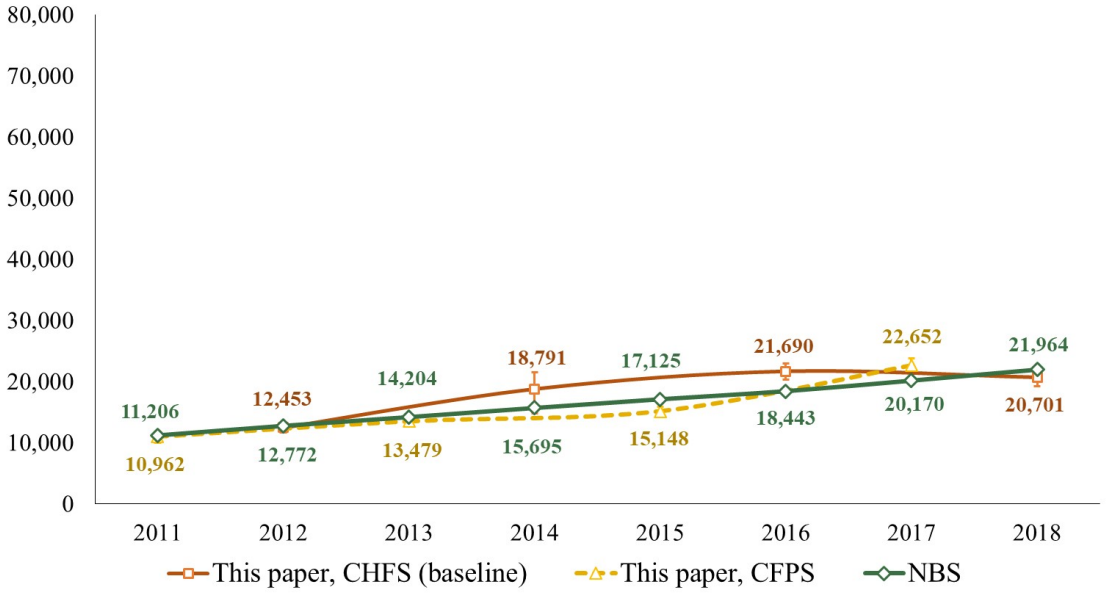


Fig. H4: Comparison of estimated mean household income per capita (CN¥) in Henan using three data sources: CHFS (baseline), CFPS, and NBS. For the NBS, values are taken from Per Capita Disposable Income as reported in the China Statistical Yearbook (https://www.stats.gov.cn/english/Statisticaldata/yearbook). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

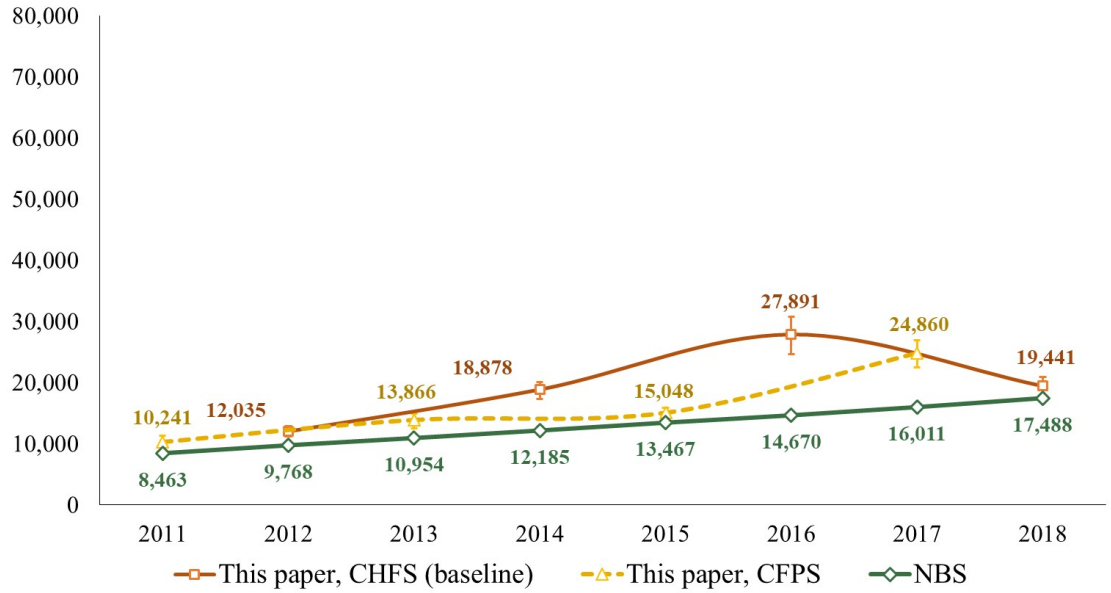
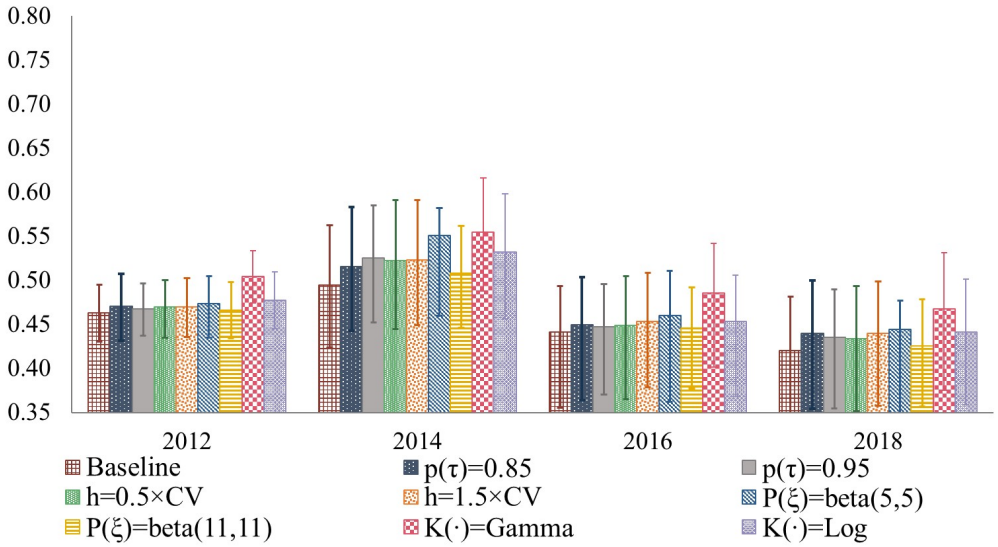


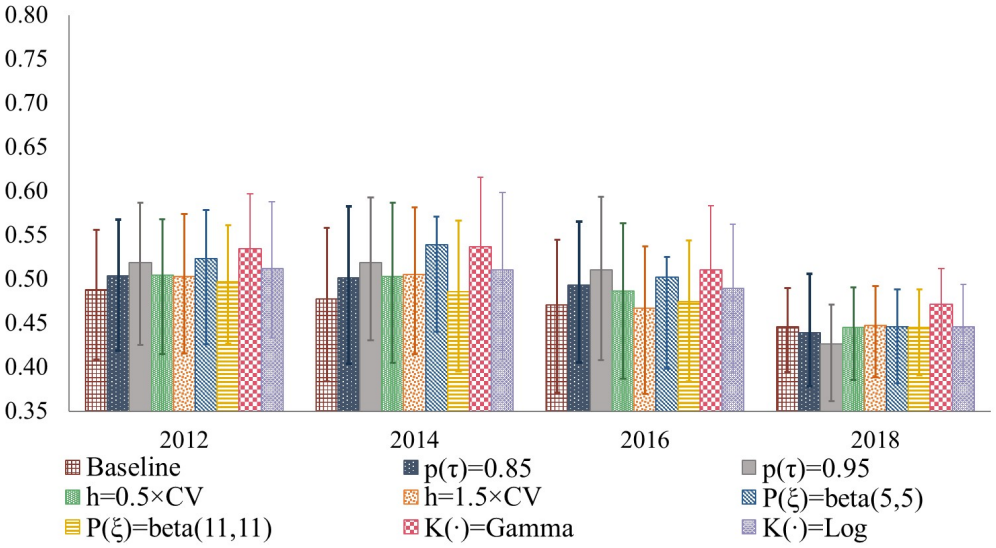
Fig. H5: Comparison of estimated mean household income per capita (CN¥) in Gansu using three data sources: CHFS (baseline), CFPS, and NBS. For the NBS, values are taken from Per Capita Disposable Income as reported in the China Statistical Yearbook (https://www.stats.gov.cn/english/Statisticaldata/yearbook). Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

Appendix I Robustness analysis

I.1 Alternative estimation settings

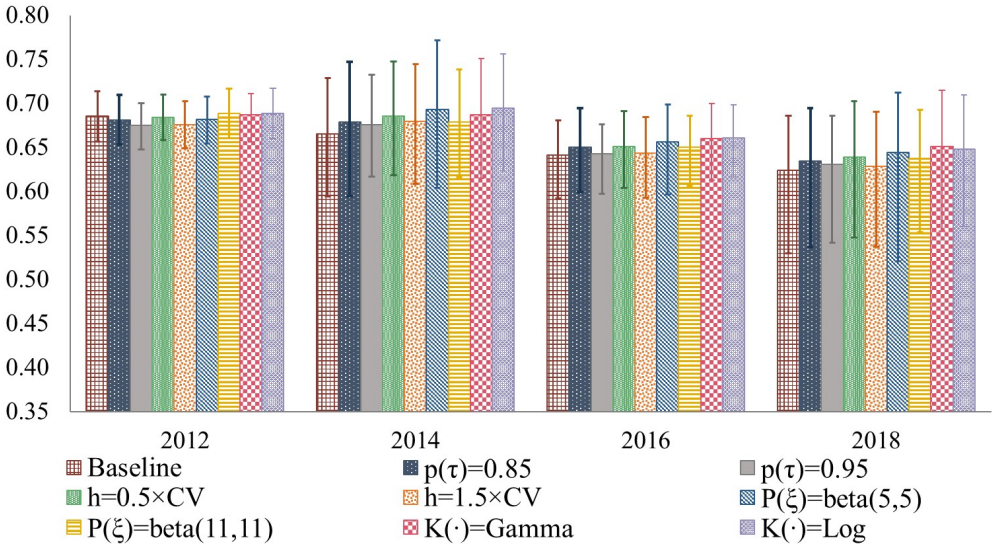


(a) CHFS (baseline), Shanghai

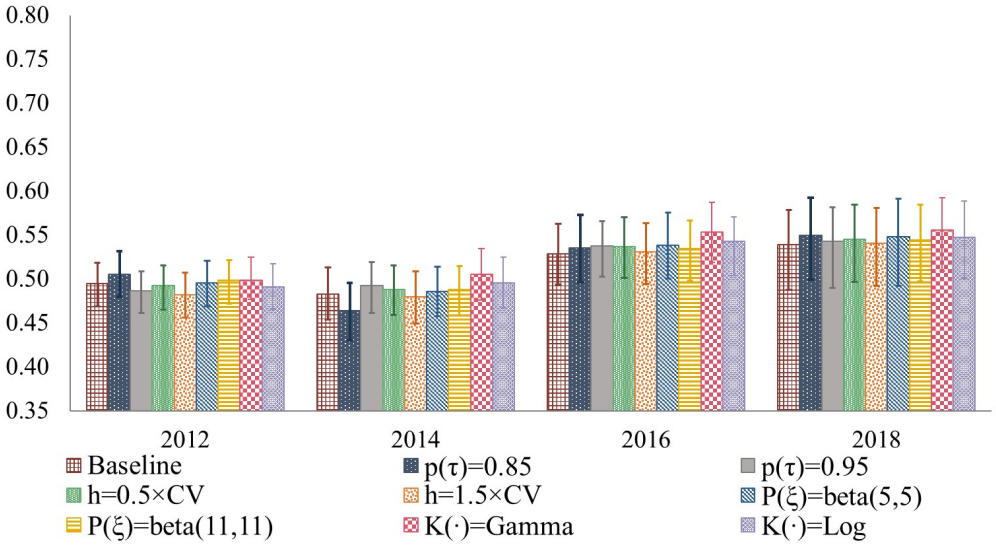


(b) CFPS, Shanghai

Fig. I1: Shanghai Gini estimates under alternative parameter choices. Robustness checks are based on different threshold percentiles ($p(\tau) = 0.85$ or 0.95), bandwidths ($h = 0.5\times$CV and $1.5\times$CV, where CV denotes the optimal bandwidth selected via cross-validation), alternative penalty functions ($P(\xi) =$ Beta$(5, 5)$ or Beta$(11, 11)$), and alternative kernel functions ($K(\cdot)$=Gamma or log-transform kernel). The baseline parameter choice is $p(\tau) = 0.90$, $h =$ CV, $P(\xi) =$ Beta$(8, 8)$, and $K(\cdot)$=Gaussian. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

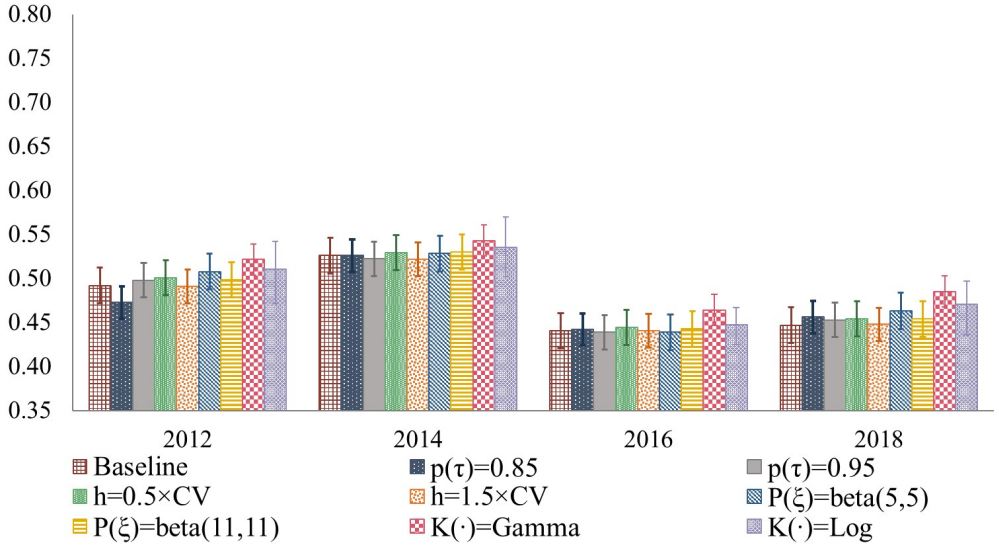


(a) CHFS (baseline), Guangdong

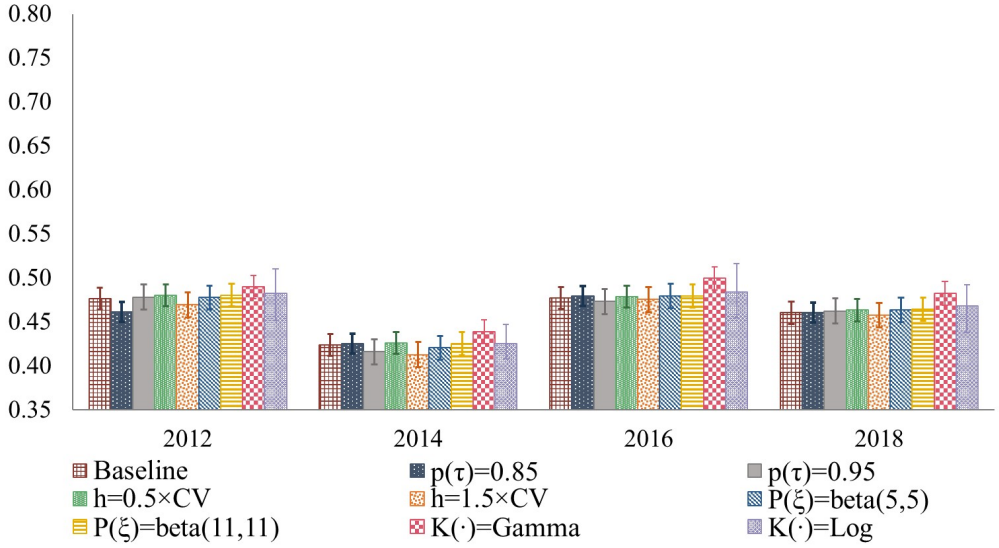


(b) CFPS, Guangdong

Fig. I2: Guangdong Gini estimates under alternative parameter choices. Robustness checks are based on different threshold percentiles ($p(\tau) = 0.85$ or 0.95), bandwidths ($h = 0.5 \times$ CV and $1.5 \times$ CV, where CV denotes the optimal bandwidth selected via cross-validation), alternative penalty functions ($P(\xi) =$ Beta$(5, 5)$ or Beta$(11, 11)$), and alternative kernel functions ($K(\cdot)$=Gamma or log-transform kernel). The baseline parameter choice is $p(\tau) = 0.90$, $h =$ CV, $P(\xi) =$ Beta$(8, 8)$, and $K(\cdot)$=Gaussian. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

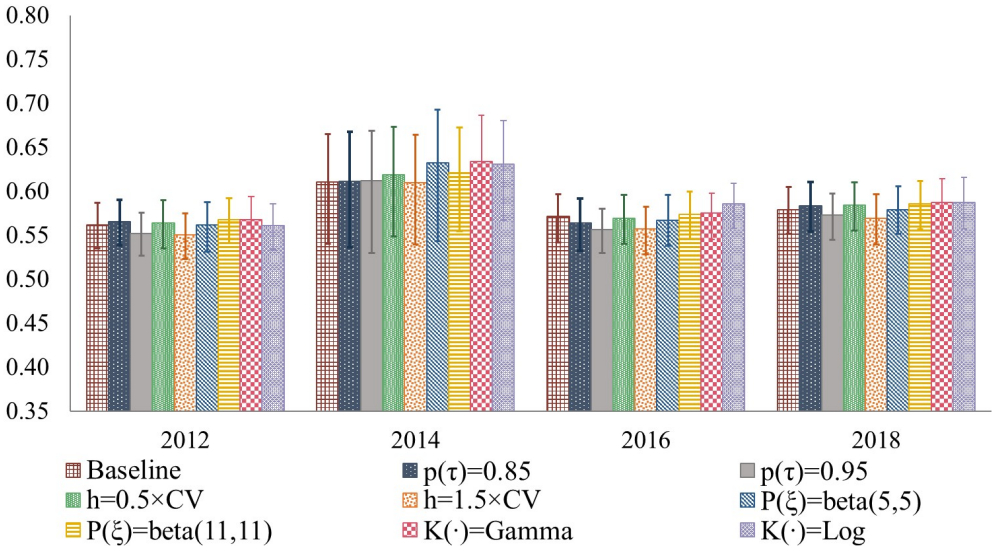


(a) CHFS (baseline), Liaoning

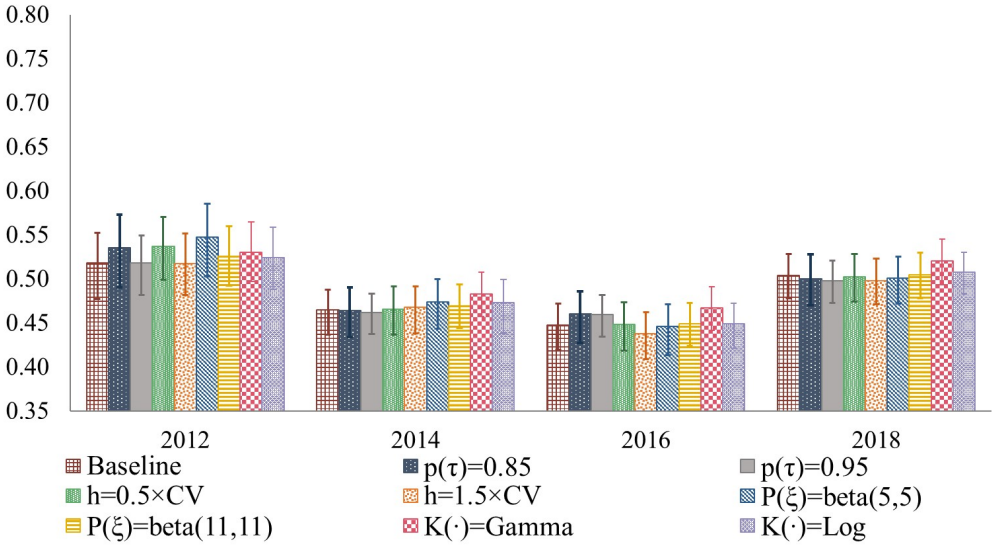


(b) CFPS, Liaoning

Fig. I3: Liaoning Gini estimates under alternative parameter choices. Robustness checks are based on different threshold percentiles ($p(\tau) = 0.85$ or 0.95), bandwidths ($h = 0.5\times$CV and $1.5\times$CV, where CV denotes the optimal bandwidth selected via cross-validation), alternative penalty functions ($P(\xi)$ = Beta$(5, 5)$ or Beta$(11, 11)$), and alternative kernel functions ($K(\cdot)$=Gamma or log-transform kernel). The baseline parameter choice is $p(\tau) = 0.90$, h = CV, $P(\xi)$ = Beta$(8, 8)$, and $K(\cdot)$=Gaussian. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

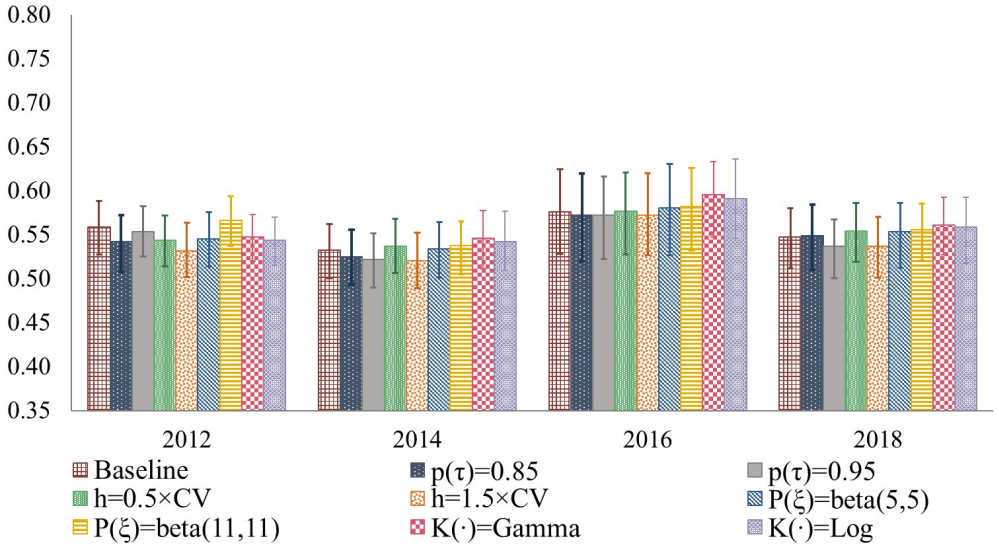


(a) CHFS (baseline), Henan

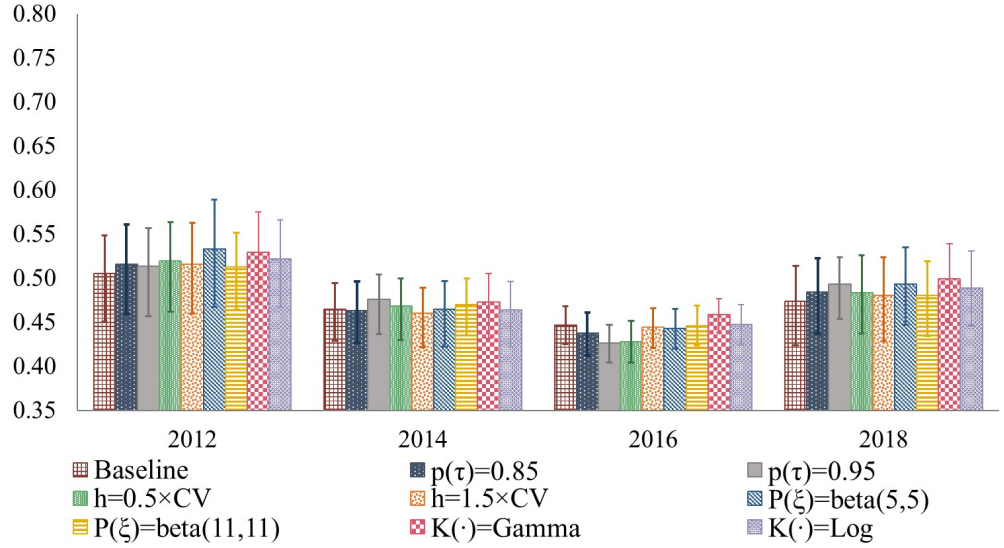


(b) CFPS, Henan

Fig. I4: Henan Gini estimates under alternative parameter choices. Robustness checks are based on different threshold percentiles ($p(\tau) = 0.85$ or 0.95), bandwidths ($h = 0.5\times$CV and $1.5\times$CV, where CV denotes the optimal bandwidth selected via cross-validation), alternative penalty functions ($P(\xi) =$ Beta$(5, 5)$ or Beta$(11, 11)$), and alternative kernel functions ($K(\cdot)$=Gamma or log-transform kernel). The baseline parameter choice is $p(\tau) = 0.90$, $h =$ CV, $P(\xi) =$ Beta$(8, 8)$, and $K(\cdot)$=Gaussian. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated from 1,000 samples



(a) CHFS (baseline), Gansu.



(b) CFPS, Gansu.

Fig. I5: Gansu Gini estimates under alternative parameter choices. Robustness checks are based on different threshold percentiles ($p(\tau) = 0.85$ or 0.95), bandwidths ($h = 0.5\times$CV and $1.5\times$CV, where CV denotes the optimal bandwidth selected via cross-validation), alternative penalty functions ($P(\xi)$ = Beta$(5, 5)$ or Beta$(11, 11)$), and alternative kernel functions ($K(\cdot)$=Gamma or log-transform kernel). The baseline parameter choice is $p(\tau) = 0.90$, h = CV, $P(\xi)$ = Beta$(8, 8)$, and $K(\cdot)$=Gaussian. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

I.2 Alternative treatments of zero and negative income

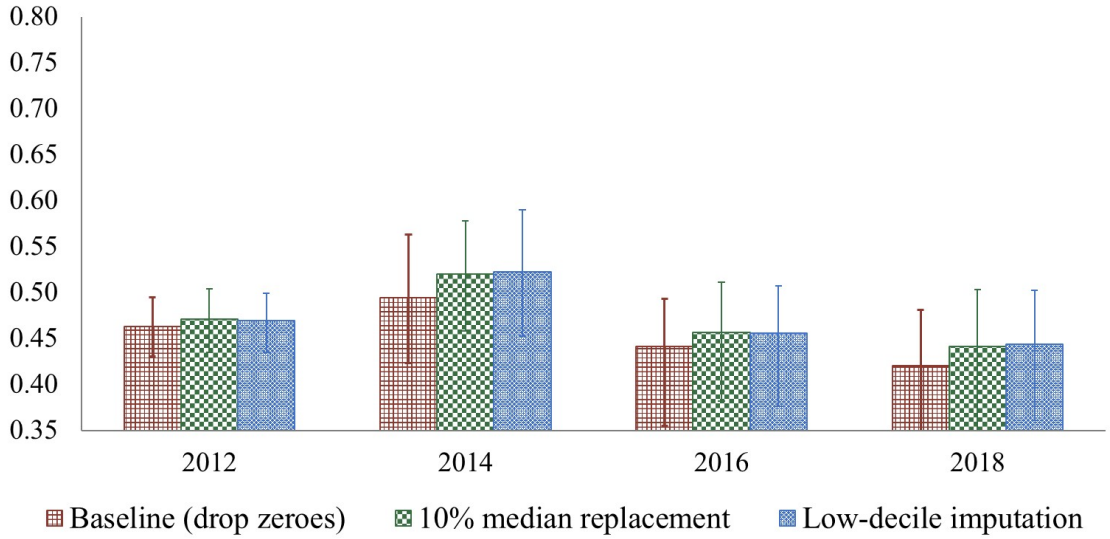


Fig. I6: Shanghai Gini estimates under alternative treatments of zero and negative income in the CHFS, following Van Kerm (2007). Bars correspond to: (a) the baseline approach, which excludes zero and negative income; (b) replacing zero and negative income with 10 percent of median income; and (c) imputation by randomly drawing values from the lowest income decile. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

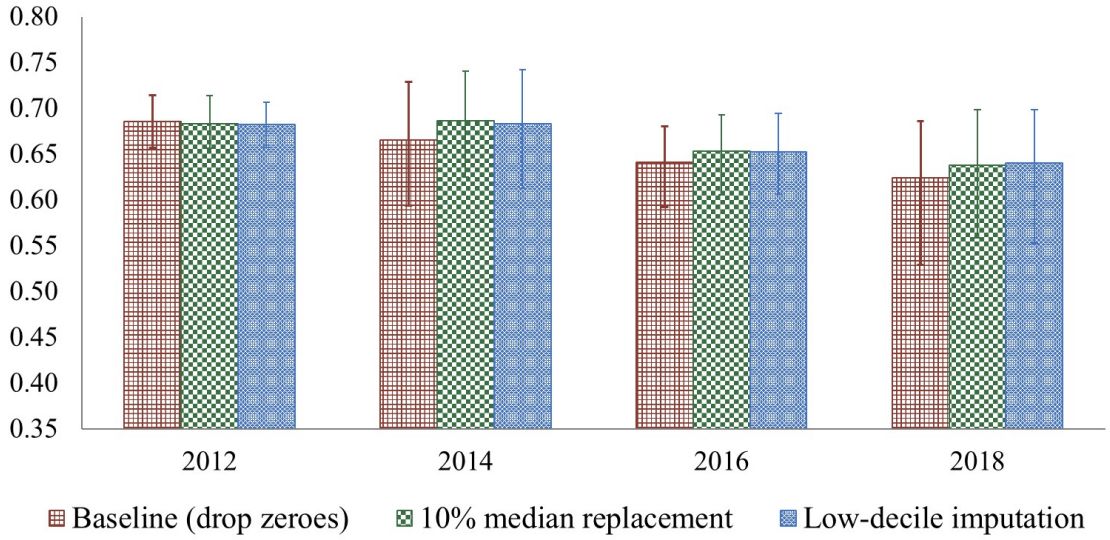


Fig. I7: Guangdong Gini estimates under alternative treatments of zero and negative income in the CHFS, following Van Kerm (2007). Bars correspond to: (a) the baseline approach, which excludes zero and negative income; (b) replacing zero and negative income with 10 percent of median income; and (c) imputation by randomly drawing values from the lowest income decile. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

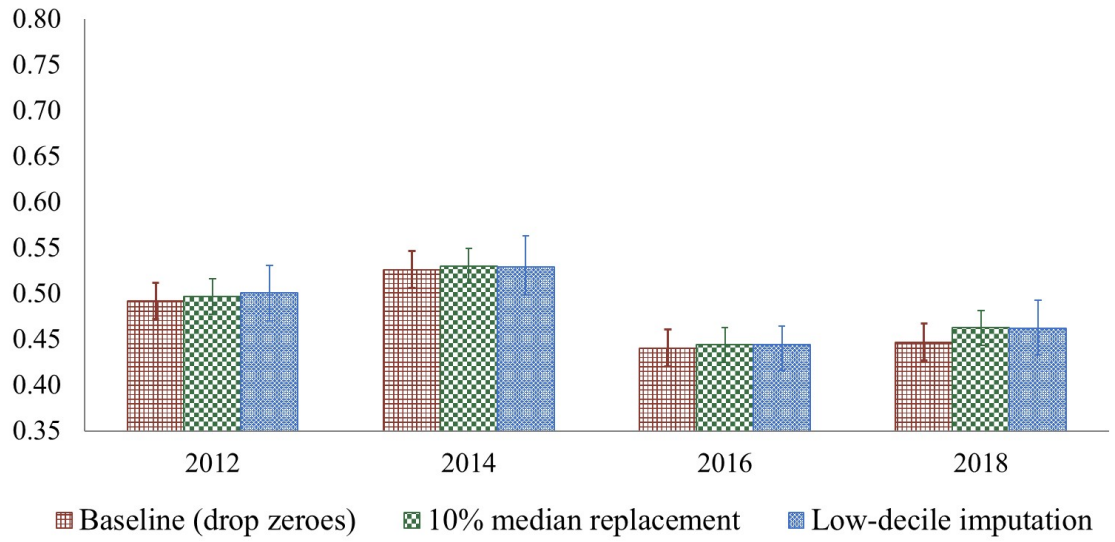


Fig. I8: Liaoning Gini estimates under alternative treatments of zero and negative income in the CHFS, following Van Kerm (2007). Bars correspond to: (a) the baseline approach, which excludes zero and negative income; (b) replacing zero and negative income with 10 percent of median income; and (c) imputation by randomly drawing values from the lowest income decile. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

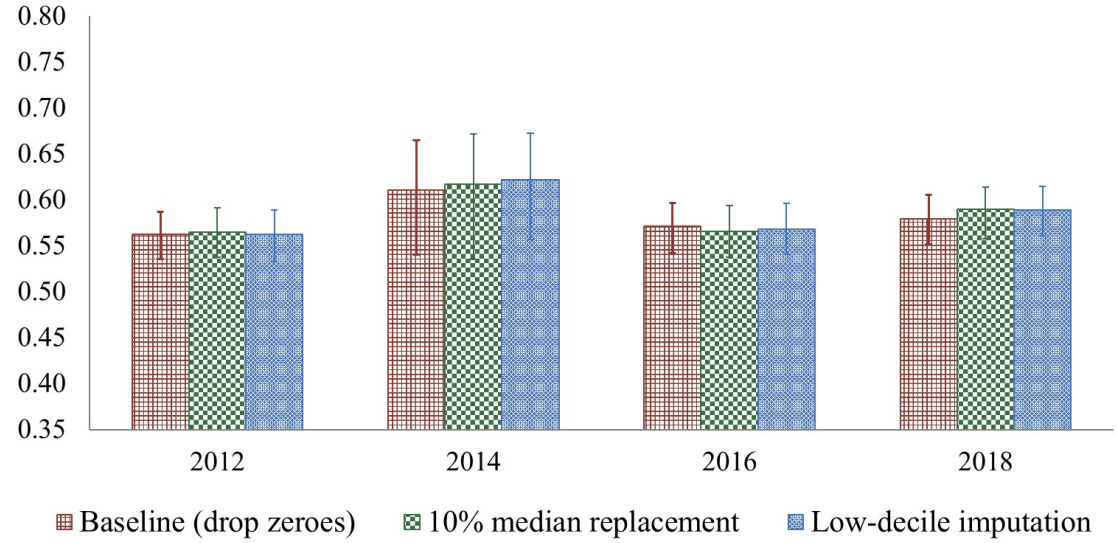


Fig. I9: Henan Gini estimates under alternative treatments of zero and negative income in the CHFS, following Van Kerm (2007). Bars correspond to: (a) the baseline approach, which excludes zero and negative income; (b) replacing zero and negative income with 10 percent of median income; and (c) imputation by randomly drawing values from the lowest income decile. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples

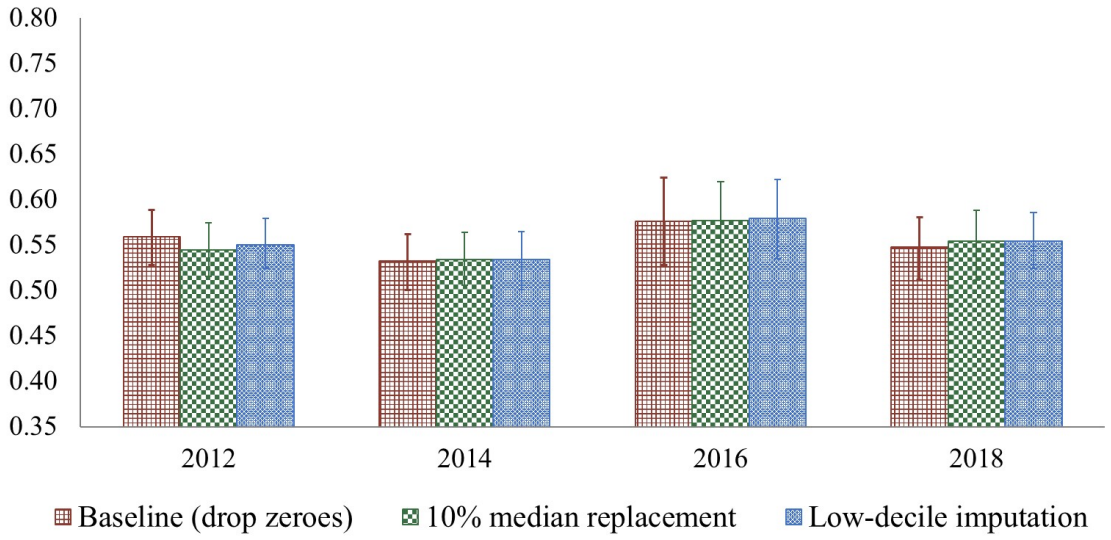


Fig. I10: Gansu Gini estimates under alternative treatments of zero and negative income in the CHFS, following Van Kerm (2007). Bars correspond to: (a) the baseline approach, which excludes zero and negative income; (b) replacing zero and negative income with 10 percent of median income; and (c) imputation by randomly drawing values from the lowest income decile. Error bars represent 95% basic (reverse percentile) bootstrap confidence intervals estimated based on 1,000 samples