KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

Discussion Paper No. 1119

Robustly Estimating Japan's Gini Coefficient for Individual Earned Income Using Household Survey and Tax Agency Data

Naoki TANI and Taro OHNO

2025 年 6 月

KYOTO UNIVERSITY

KYOTO, JAPAN

Robustly Estimating Japan's Gini Coefficient for Individual Earned Income Using Household Survey and Tax Agency Data [†]

Naoki TANI
Institute of Economic Research, Kyoto University and Ministry of Finance

Taro OHNO
Faculty of Economics and Law, Shinshu University

Abstract

This study investigates the benefits and caveats of using tax agency data through a descriptive analysis, and estimates the robust Gini coefficient for individuals' earned income by applying statistical tools combining household survey and tax agency data. We show that the advantage of the tax agency data is that it captures top incomes, while its weak coverage of female non-regular workers can be complemented by combining it with the household survey data. Further, the descriptive results show that the Gini coefficient computed using the household survey is larger than that from using tax agency data, despite not covering top incomes. This indicates that capturing the distribution of the middle and low incomes is more important to estimate the inequality level than capturing the top incomes in Japan. Moreover, the robust estimate of the Gini coefficient indicates that combining top incomes does not substantially affect the overall Gini index computed solely from the household survey data, which is distinct from the results for other countries in the literature. However, when we decompose the Gini coefficient into between- and within-group components of gender and employment status, combining the tax agency and household survey data is important. Although both data show an increase in the between-group component from 2014 to 2019, the integrated data indicate that the between-group contribution actually decreases from 2014 to 2019, reflecting the increases in the incomes of regular female workers.

Keywords: Gini coefficient, Pareto distribution, top incomes, Lorenz curves, tax agency data

JEL classification: C46, C81, D31, H24

[†]We are grateful to Koyo Miyoshi and Junji Ueda for their insightful comments and suggestions at the Policy Research Institute (PRI) of the Ministry of Finance seminar. We also wish to thank Takahisa Omata and Naruhiko Fukumuro in PRI for their support with the application for the use of the micro data. The opinion in the article does not reflect the views of the Ministry of Finance Japan or members of its staff. All remaining errors are our own.

1 Introduction

Income inequality has attracted growing attention globally. With the increasing social awareness of the disparity between employment statuses, especially that between regular and non-regular workers, the Japanese government has extensively promoted wage increases for non-regular workers by enforcing the equal pay for equal work system from the mid-2010s[1].

Indeed, the Gini coefficient of Japanese households' earned income was in a gradual decline in the second half of the 2010s (Kitao and Yamada (2024); Kitao and Yamada (2025); Cabinet Office (2022)). Studies reporting the Gini coefficient for earned income in Japan have used household survey data so far, such as the National Survey of Family Income and Expenditure (NSFIE); National Survey of Family Income, Consumption and Wealth (NSFICW); Family Income and Expenditure Survey; and Comprehensive Survey on Living Conditions (Kohara and Ohtake (2014); Kitao and Yamada (2024); Kitao and Yamada (2025); Cabinet Office (2022)). However, several recent studies report that household-survey-based estimates of inequality underreport the change in income inequality due to the lack of top incomes (Atkinson (2007); Atkinson, Piketty, and Saez (2011); Alvaredo (2011); Jenkins (2017); Li, Yu, and Li (2021); Flachaire, Lustig, and Vigorito (2023)). In Japan, some researchers have recently tackled the issue of estimating Pareto coefficients using the tax filing data, or reporting the top income shares using the household survey and semi-aggregated tax dataset(Kunieda and Yoneta (2023); Mikayama, Imahori, Ohno, Yoneta, and Ueda (2023)). Thus, studying the robust inequality level using a dataset that captures top incomes can be beneficial for researchers and policymakers to understand the income disparity between employment statuses.

Here, we use the micro data from the Statistical Survey of Actual Status for Salary in the Private Sector (SSASSPS; hereinafter, "tax agency data"), which are compiled by Japan's National Tax Agency and has recently been made available to academic researchers. One of the main advantages of the tax agency data is that it requires the sampled establishments to report all employees who earned more than 20 million JPY. Thus, we first investigate the benefits and caveats using the tax agency data through descriptive analysis. Thus, the advantage of the tax agency data is clearly the strong coverage of top incomes. Conversely, the disadvantage is its weak coverage of female non-regular workers. We compare the Gini coefficient of individuals' earned income estimated from the tax agency data with that from the NSFIE and NSFICW (hereinafter, "household survey data"). We find that the Gini coefficient of the household survey data is larger than that of the tax agency data, despite the lack of coverage of the top incomes of the household survey data. This indicates that capturing the distribution of middle and low

[1]The policy was started by Prime Minister Shinzo Abe in 2016.

incomes is more important to estimate inequality level than capturing that of top incomes in Japan.

Then, to complement the disadvantage of the tax agency data, we combine the top incomes of the tax agency data with the household survey data. After checking Paretianity of the tax agency data and finding the splicing points at which we combine two datasets using the Kolmogorov-Smirnov (KS) statistic, we estimate the robust Gini coefficient using the integrated dataset. Clearly, combining top incomes does not substantially affect the overall inequality level of the Gini index computed solely from the household survey data, which is distinct from the results for other countries in the literature.

Finally, we conduct the Gini decomposition analysis based on Dagum (1997). The Gini decompositions from the household survey and tax agency data show that the between-group contribution of gender and employment status increased from 2014 to 2019. This result on the increase in the group-component remains unchanged even when we use the simply-integrated data. However, we show that the simple integration method, which involves combining two datasets at one splicing point following the literature, may overestimate the Gini coefficient and between-group component. When we use our proposed parallel integration method, which identifies the gender-employment-status-specific splicing points and combines two data parallelly by every group-specific distribution, the between-group component of our integrated data decreases from 2014 to 2019. This reflects the increases in the incomes of regular female workers.

The main contributions of this work are as follows: (1) We compare the distributions of individuals' earned income in the household survey and tax agency data through the descriptive analysis, and report the benefits and caveats using the tax agency data. In particular, we show that the Gini coefficient computed using the household survey is larger than that computed using the tax agency data because of the difference in their coverage. Since the micro data from the household survey have been widely used by researchers in Japan, analyzing the difference in the household survey and tax agency data provides evidence of the correctness of the coverage of the household survey. (2) Additionally, we combine the two datasets and estimate the robust Gini coefficient. Again, its value is not substantially different from the coefficient estimated solely from the household survey. This evidence provides a reference for studies on this topic. (3) The parallel integration method corrects the overestimates of the Gini coefficients obtained from the simple integration method in the literature, which is a methodological contribution. (4) We report the decomposition of the Gini coefficient into between- and within-components of gender and employment status after combining the two data sources. This provides evidence of the contribution of the difference in gender and employment status on overall inequality level, thereby contributing to the related literature.

The remainder of the paper is organized as follows. Section 2 presents the data and basic facts. Section 3 shows the descriptive analysis. Section 4 introduces the method to find the splicing point and outlines reason we use parallel integration. In Section 5, we integrate the datasets, re-estimate the Gini coefficient, and conduct the Gini decomposition analysis. Section 6 presents the conclusions of this study.

2 Data and basic facts

The household survey data are from the NSFIE in 2014 and the NSFICW in 2019[2], which are compiled by the Ministry of Health, Labour and Welfare in Japan. We use individual-level pre-tax annual earned income from a main job as a variable of individuals' earned income[3]. Since NSFIE and NSFICW are household surveys, we use the data on the employed household heads and spouses[4]. That is, we exclude members other than the head and spouse in each household because their earned incomes are aggregated. We also drop individuals who are not working since we focus on earned income. The person whose earned income is zero is also excluded. This is because surveyed households report their earned income of the previous year. If they did not work in the previous year and start working in the survey year, they report their income as zero. Finally, we use the gender and employment status of the surveyed person.

Tax agency data are from the SSASSPS in 2014 and 2019, which is compiled by Japan's National Tax Agency. SSASSPS conducts a two-stage sampling to select the surveyed individuals. In particular, in the first stage, establishments are stratified by number of employees and other factors, and the sample establishments are extracted. Then, in the second stage, surveyed employees are extracted from the sample establishments. The sampling rate depends on the number of employees[5]. However, notice that sample establishments must report all employees

[2]We do not adjust the difference in the sampling weights of NSFICW and NSFIE since the adjusted sampling weights are not recorded in our datasets.

[3]Since our study focuses on earned income, we can not discuss the total income. In particular, if we include capital income in our study, the effect of covering top incomes on the inequality levels may increase. However, it is important to study the earned income in order to discuss inequality between genders and employment statuses.

[4]Although NSFIE and NSFICW record the aggregate earned income of household members other than their heads and spouses, it is not possible to identify their individual-level income. In order to focus on the analysis of individual-level earned income, we select data on the employed household heads and spouses. It is notable that excluding household members other than their heads and spouses may negatively affects inequality levels since they tend to earn much less than household heads.

[5]For establishments with 33 $\sim$ 99 employees, the sampling rate of surveyed employees is defined as "All employees who earned more than 20 million JPY and one-sixth of those who earned equal to or less than 20 million JPY."

who earned more than 20 million JPY. Thus, SSASSPS has a strong coverage to top incomes. The individuals' earned income is from the variable of the individual-level annual salary from the sampled establishments. Notice that the annual salary from the sampled establishments is not always equal to the annual earned income because workers may receive salary from other establishments outside the sample. Additionally, SSASSPS only covers the private sector. We use other variables like gender, employment status, and employers'(establishments') size and industry.

We summarize the comparison of household survey and tax agency data in Table 1. As we show in the Section 3, the household survey has (tax agency data have) strong coverage of low-income (high-income) earners but weak coverage of top (low) incomes. Thus, the household survey and tax agency data can complement each other.

Table 1: Household survey and tax agency data

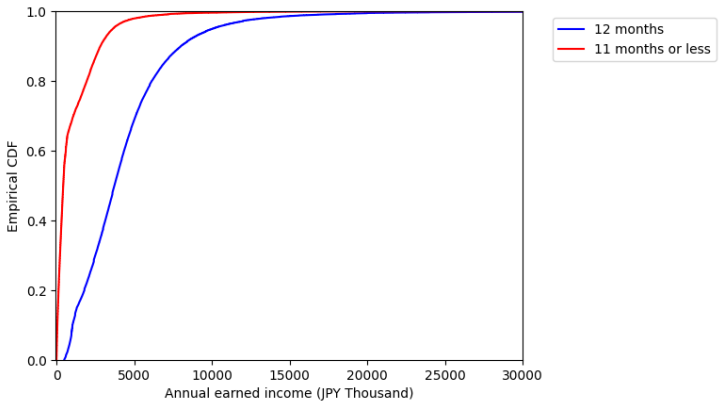
Data	Positive	Negative
Household survey data (NSFIE/NSFICW)	- Strong coverage of low-income earners - Cover public employees	- Weak coverage of top incomes
Tax agency data (SSASSPS)	- Strong coverage of top incomes - Record of employment status of employees and size and industry of employers (not included in Tax Filing data) - More accessible than tax filing data (but used by few researchers so far)	- Weak coverage of low incomes (Female, Non-regular workers) - Not cover public employees

2.1 Potential problems related to the tax agency data

In the tax agency data, employers report earners who received salaries of 12 months and those who earned 11 months or less. The latter may receive a salary from other establishments

outside the sample. Fig. 1 depicts the empirical cumulative distribution function (CDF) of earners who received salaries of 12 months and 11 months or less. The latter have much lower distribution. However, we do not know whether they received salaries from other employers and how many months they received salaries from surveyed establishments.

Figure 1: Empirical cumulative distribution function of the tax agency data (2019)(only incomes below 30 million JPY are shown)



To reduce the negative bias from the low distribution of earners who received salaries of 11 months or less, we drop them from our sample. However, their exclusion causes the loss of non-regular workers, especially those of females as shown in Table 2. Thus, we solve this problem by combining the household survey data in Sections 4 and 5.

Table 2: Statistics for the tax agency data by gender and employment status

	12 months		11 months or less	
	Non-regular worker	Others	Non-regular worker	Others
Male	7.3%	50.4%	23.1%	20.3%
Female	16.5%	25.7%	41.4%	15.1%

Source: SSASSPS in 2019.

Note: The fraction of each group is adjusted by sampling weights.

2.2 Composition of the household survey and tax agency data

The household survey data has some advantages in recording the substantial number of female non-regular workers as shown in Table 3. Leveraging this advantage enables us to complement low-income earners to estimate inequality indices.

Table 3: Statistics for the two data sets by gender and employment status

	Tax agency data		Survey data	
	Non-regular worker	Others	Non-regular worker	Others
Male	7.3%	50.4%	8.2%	47.4%
Female	16.5%	25.7%	23.7%	20.7%

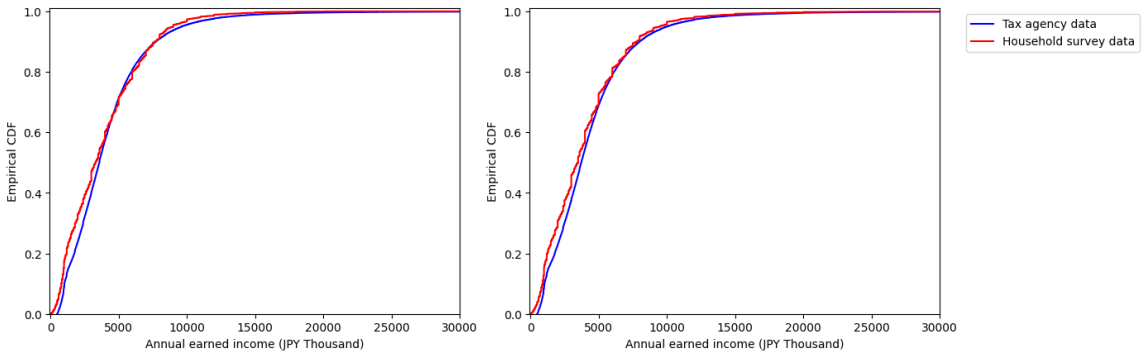
Source: SSASSPS and NSFICW in 2019.

Note: The fraction of each group is adjusted by sampling weights.

3 Descriptive analysis

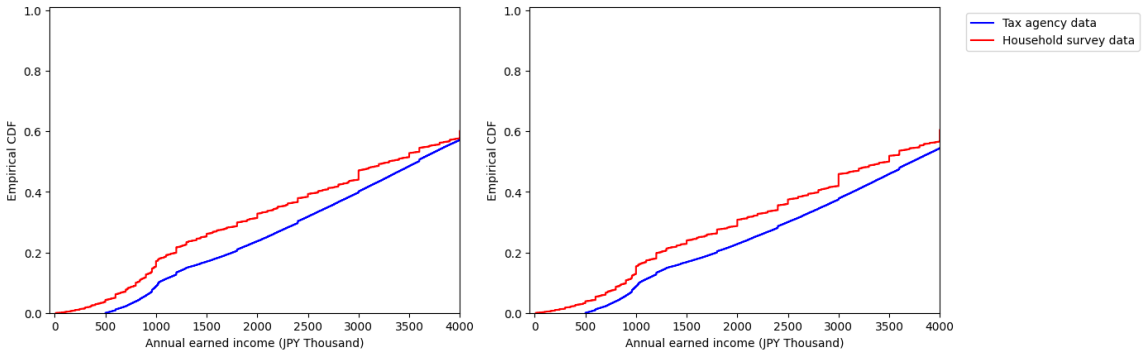
Figs. 2, 3, and 4 show that the household survey and tax agency data have similar empirical CDFs except for the top and bottom incomes. The household survey has higher values of the CDF in the bottom-income brackets than those in the tax agency data in Fig. 3; the opposite holds for the top-income brackets in Fig. 4.

Figure 2: Empirical cumulative distribution function of the two datasets (only incomes below 30 million JPY are shown)



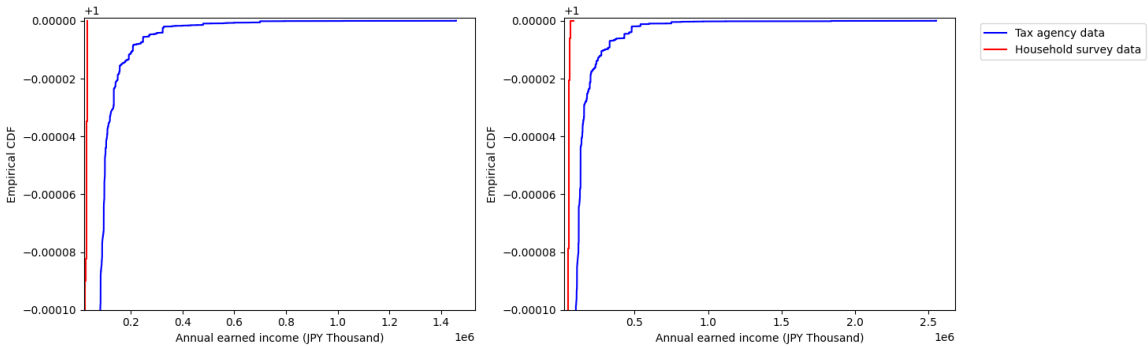
(a) 2014 (b) 2019

Figure 3: Empirical cumulative distribution function of the two datasets (only incomes below 4 million JPY are shown)



(a) 2014 (b) 2019

Figure 4: Empirical cumulative distribution functions of the two datasets (only incomes greater than 20 million JPY are shown)



(a) 2014 (b) 2019

3.1 Gini inequality index and its decomposition

Next, we introduce the decomposition of Gini coefficient considering the sampling weights based on Dagum (1997).

Consider $N = \sum_{k=1}^{K} \sum_{i=1}^{N_k} w_{ik}$ individuals with mean income μ, sampling weight w_{ik}, and $k = 1, 2, ...K$ partitioned groups with $N_k = \sum_{i=1}^{N_k} w_{ik}$ individuals with mean income μ_k: $\mu = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N_k} w_{ik} y_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{N_k} w_{ik}}$ and $\mu_k = \frac{\sum_{i=1}^{N_k} w_{ik} y_{ik}}{\sum_{i=1}^{N_k} w_{ik}}$. Each individual i in group k has y_{ik} income.

The Gini coefficient of the whole population G, within component G_w, and between component G_b are expressed as follows:

$$G = \frac{\sum_{k=1}^{K} \sum_{\ell=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_\ell} |y_{ik} - y_{j\ell}| w_{ik} w_{j\ell}}{2N^2 \mu} = \sum_{k=1}^{K} G_{kk} P_k S_k + \sum_{k=1}^{K} \sum_{\ell \neq k}^{K-1} G_{k\ell} (P_k S_\ell + P_\ell S_k) \tag{1}$$
$$= G_w + G_b$$

where $P_k = \frac{N_k}{N}$, $S_k = \frac{N_k \mu_k}{N \mu}$,

$$G_{kk} = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_k} |y_{ik} - y_{jk}| w_{ik} w_{jk}}{2N_k^2 \mu_k}; \tag{2}$$

$$G_{k\ell} = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N_\ell} |y_{ik} - y_{j\ell}| w_{ik} w_{j\ell}}{N_k N_\ell (\mu_k + \mu_\ell)}. \tag{3}$$

The results are reported in Table 4. Compared to Kitao and Yamada (2025), who report the Gini coefficients at the household level (0.627 in 2014 and 0.559 in 2019), our coefficients from the household survey are smaller because we drop individuals who are not working or whose earned income is zero. The coefficients in both datasets have increased a little from 2014 to 2019. Remarkably, the coefficients from the household survey are larger than those from the tax agency data despite the lack of covering top incomes. Moreover, the between-component has increased a little from 2014 to 2019 in both datasets despite the government's policy.

To show the reason why the Gini index from household survey is larger than that from tax agency data, we depict the Lorenz curves from the two datasets in Fig. 5. The Gini coefficient is known to be twice the area captured between Lorenz curve $L_<(w)$ and the line of perfect equality:$G \equiv 2 \int_0^1 [w - L_<(w)] dy$, where w is cumulative share of workers. Thus, we can visualize why the household survey has a higher Gini coefficient than the tax agency data.

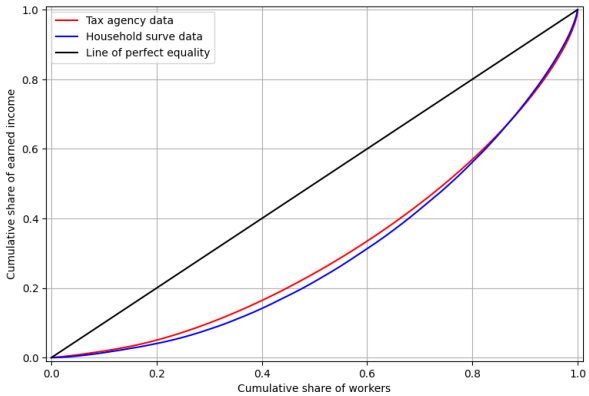
Table 4: The gini coefficient, and its decomposition by gender and employment status

	Gini	Decomposition	
		Between	Within
Household survey data			
2014	0.401	77.8%	22.2%
2019	0.402	78.0%	22.0%
Tax agency data			
2014	0.375	76.8%	23.2%
2019	0.378	77.4%	22.6%

Source: SSASSPS and NSFIE/NSFICW in 2014 and 2019.

Note: Employment status is classified into "Regular worker", "Non-regular worker", and "Others". The coefficients are adjusted by the sampling weights.

Figure 5: Lorenz curves of the tax agency and household survey data, 2019

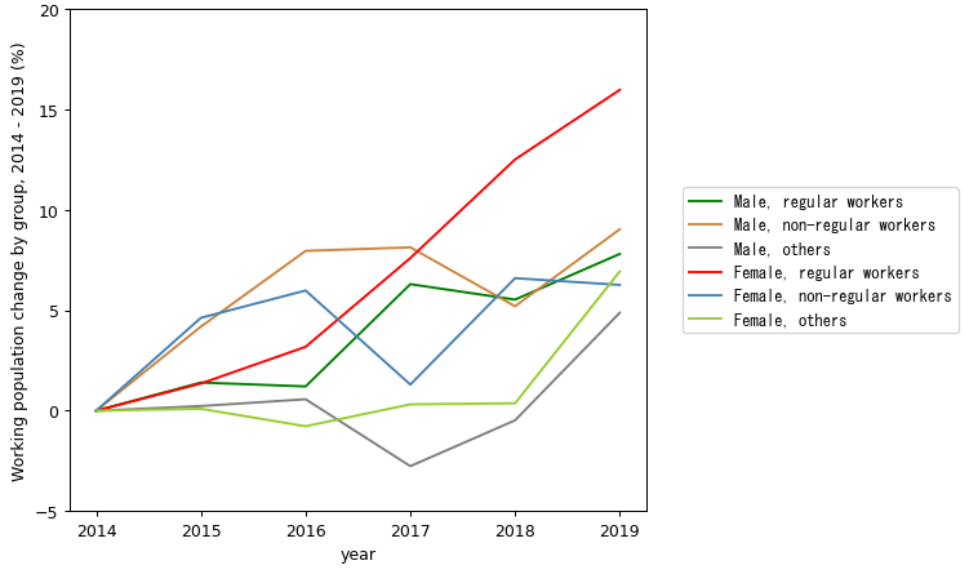


The household survey has lower values of the Lorenz curve for low and middle percentiles than the tax agency data, and vice versa for top percentiles. Therefore, the effect of the household survey's capturing middle and low incomes on Gini coefficient is larger than that of tax agency data's capturing top incomes. Consequently, the coefficients of the former are larger than those of the latter.

Why did the Gini coefficient of both datasets increase from 2014 to 2019? Possibly, the promotion of women's participation in the workplace increases the low-income female earners,

such as non-regular female workers. Since our analysis focuses on employed workers, the Gini coefficient may increase "among workers". Then, the increase in the Gini coefficient does not mean an increase in the inequality in earned income. Rather, it represents the transitional result of the promotion of women's participation in the workplace. However, this is not the case as seen in Fig. 6. Specifically, the increase in working population from 2014 to 2019 was not concentrated on female non-regular workers, but on regular female workers.

Figure 6: Working population change by group from 2014 to 2019



Source: SSASSPS

Additionally, we use the relative income divergence curve to show that the distribution of female workers did not shift to the lower percentiles. The relative income divergence curve is introduced as the "Relative Regional Income Divergence Curve" in Rinz and Voorheis (2023) to visualize the distributional difference between groups. We use it to show the distributional difference between gender and employment status groups.

The relative income divergence curve can be expressed as follows:

$$R(p) = F_N(q_n(p)) - p, \ \forall p \in [0, 1]$$

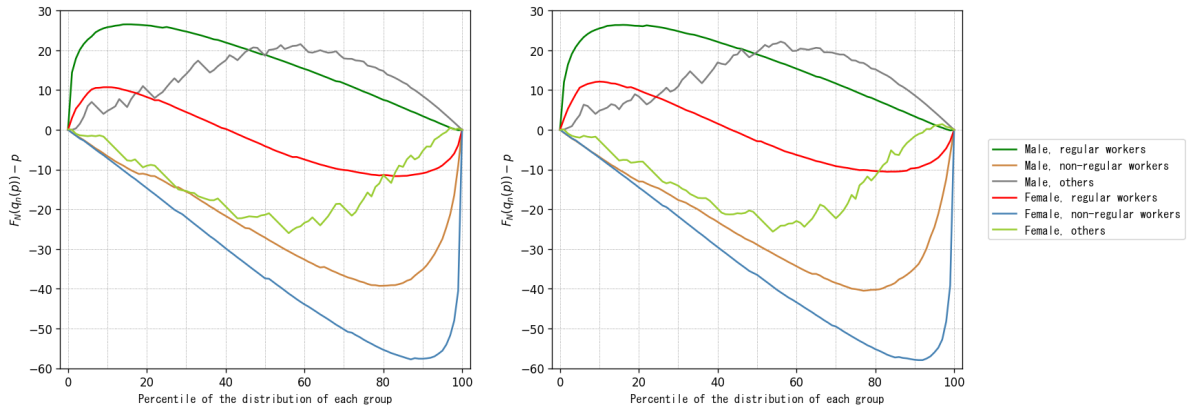
where $F_N(.)$ is the cumulative income distribution function of the national distribution, p is percentile of the group-specific distribution, and $q_n(p)$ is the quantile function of the distribution of group n.

As shown in Fig. 7, the relative income of regular female workers increased at around 10 to 90 percentiles, while that of female non-regular workers remains almost unchanged from 2014 to 2019.

Considering the results of Figs. 6 and 7, increasing women's participation in the workforce did not cause the increase in female low pay earners in the sample. Clearly, this is not the cause of the increase in the Gini coefficient.

Fig. 8 depicts the CDF of the tax agency data in 2014 and 2019. The value and composition ratio of top incomes increased from 2014 to 2019. Thus, the growth in top incomes may cause the larger increase in the Gini coefficient in tax agency data. However, since the Gini coefficient of the household survey also increased despite the different coverage from the tax agency data, further analysis is needed to obtain true changes in inequality by combining the two datasets.

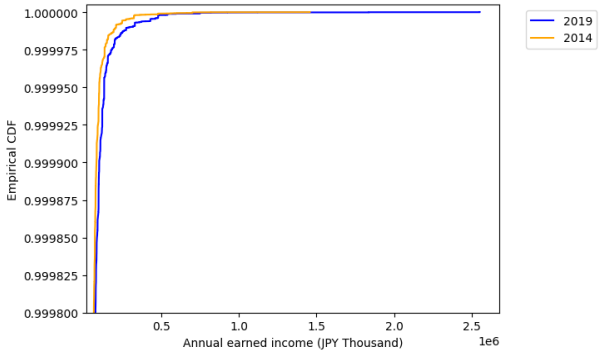
Figure 7: Relative income divergence curves by gender and employment status



(a) 2014 (b) 2019

Source: SSASSPS

Figure 8: Growth of top incomes



Source: SSASSPS

4 Integration methodology

4.1 Splicing point

This section introduces the methodology to find a splicing point, and integrate the tax agency and household survey data at the points. The methodology is based on Jenkins (2017) and Li et al. (2021).

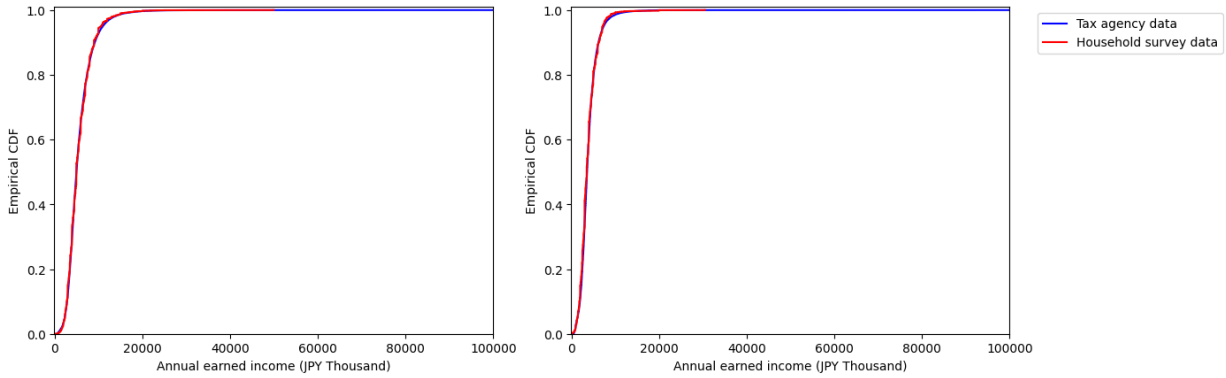
First, we find the splicing point from the distribution of the tax agency data such that the complementary CDF (CCDF) of the left-truncated tax agency data offers the closest fit to the CCDF of the analytical Pareto model. In particular, suppose that the splicing point is an arbitrary income value x_{min}. Following Jenkins (2017) and Li et al. (2021), we estimate the scaling parameter α of the Pareto distribution as $\hat{\alpha} = 1 + n\left[\sum_{i=1}^{n}\left(\ln\frac{x_i}{x_{min}}\right)\right]^{-1}$. Given $\hat{\alpha}$, we obtain the CCDF $P(x)$ as $P(x) = \left(\frac{x}{x_{min}}\right)^{-\alpha+1}$. Then, we calculate the KS statistic by minimizing the distance between the empirical CCDF $S(x)$ and analytical Pareto model CCDF $P(x)$ given x_{min}. We repeat the process above for any value separated from a million JPY to the maximum value of the income distribution by every million JPY to decide the final value of x_{min} such that the KS statistic is minimized: $\min\{\max_{x_{min}\leq x\leq x_{max}} |S(x) - P(x)|\}$.

Second, after finding the splicing point, we integrate records whose earnings are higher than the splicing point in the tax agency data and those lower than the splicing point in the household survey data. We find just one splicing point from the entire distribution of the tax agency data.

However, although the literature decides one splicing point, combining the two datasets at just one splicing point may cause an overestimation of the Gini coefficient. Fig. 9 shows how the maximum values of earned income in the household survey differ by gender and employment status. In particular, the maximum income of non-regular workers in the household survey is much lower than that of other workers. Thus, if the splicing point for the whole distribution is higher than the maximum income of non-regular workers, the integrated data will not include non-regular workers whose income is higher than the maximum value in the household survey but lower than the splicing point. Consequently, non-regular workers with relatively high incomes are dropped from the sample, resulting in the overestimation of the Gini coefficient.

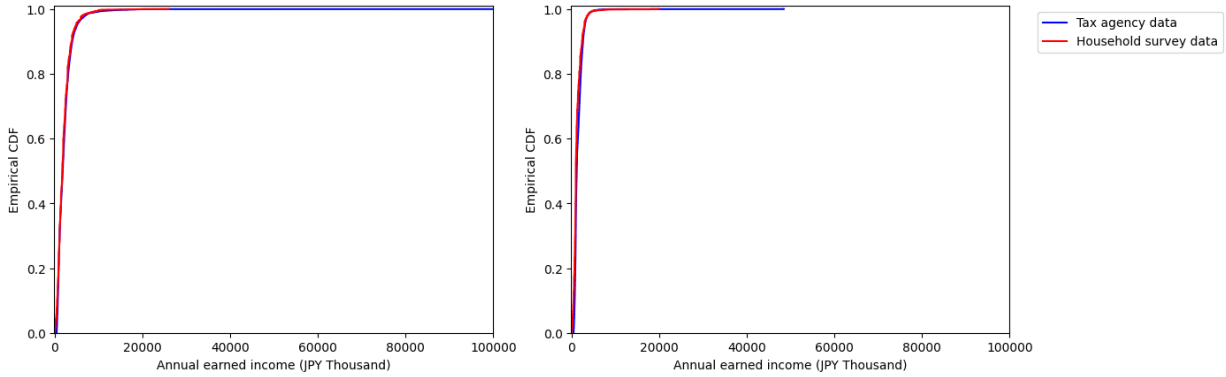
To avoid this overestimation problem, we try the parallel integration method: we apply the integration method to each distribution by gender and employment status parallelly, and then combine all the data. We can make splicing points by gender and employment status thanks to the variables of the tax agency data.

Figure 9: Empirical cumulative distribution function of the two datasets by gender and employment status, 2019 (only incomes below 100 million JPY are shown)



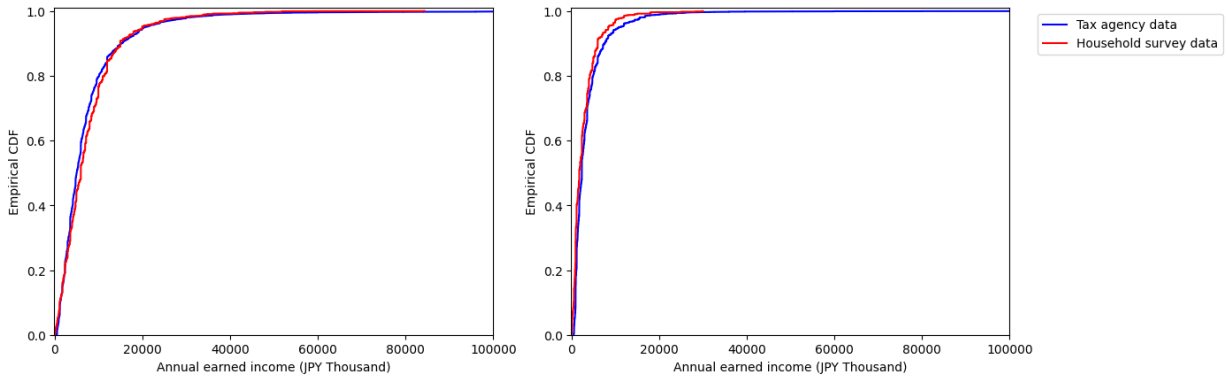
(a) Male, regular workers

(b) Female, regular workers



(c) Male, non-regular workers

(d) Female, non-regular workers



(e) Male, other workers

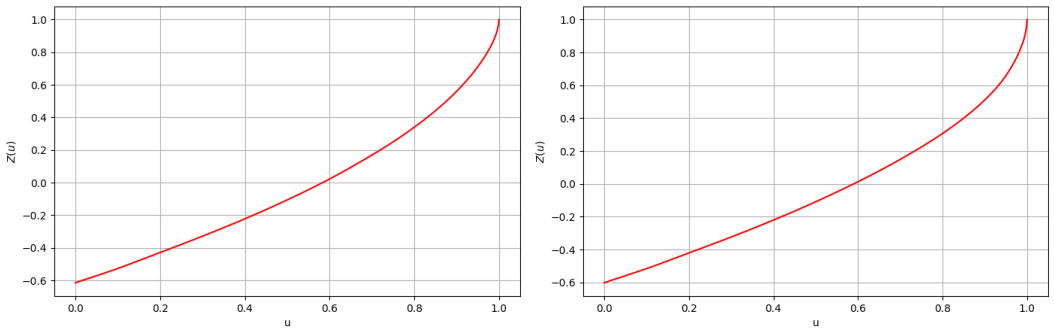
(f) Female, other workers

4.2 Paretianity

Our method explained in Section 4.1 assumes the Paretianity of all distributions by gender and employment status. Before combining the household survey and tax agency data, we check the Paretianity of the distributions by using the Zenga curve. The Zenga curve $Z(w)$ has been

developed by Cirillo (2013) to check Paretianity.: $Z(w) = \frac{w - L>(w)}{w[1 - L>(w)]}$, $0 < w < 1$ where $L > (w) = 1 - L < (w)$. A dataset follows the Pareto distribution when $Z(w)$ is positively-sloped and rises as $w \rightarrow 1$. As shown in Figs. 10 and 11, the Zenga plots provide strong evidence of Paretanity for all distributions by gender and employment status.

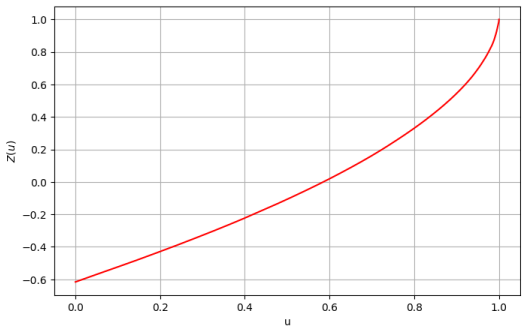
Figure 10: Zenga curves for tax agency data (Threshold = JPY 23 million)



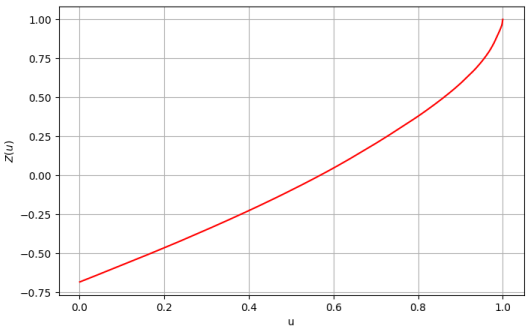
(a) 2014

(b) 2019

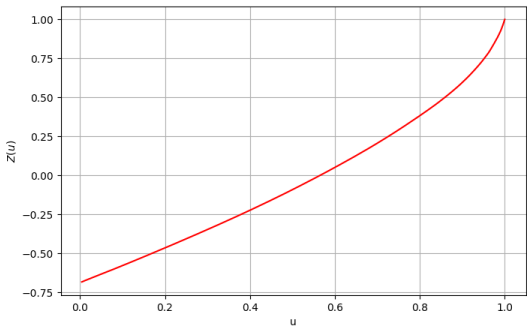
Figure 11: Zenga curves for tax agency data, by gender and employment status, 2019



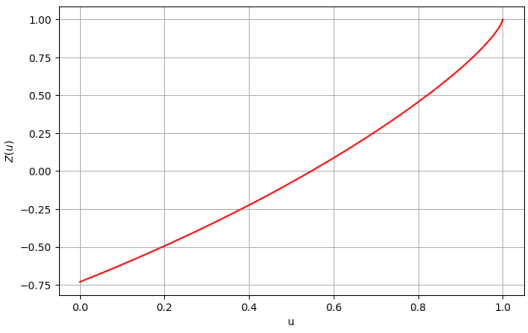
(a) Male, non-regular worker (Threshold = JPY 35 million)



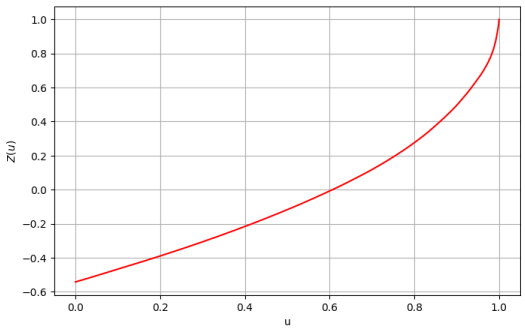
(b) Female, non-regular worker (Threshold = JPY 24 million)



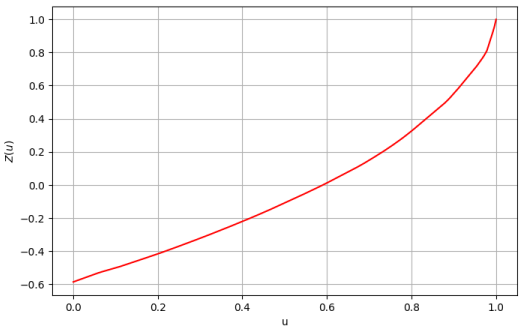
(c) Male, non-regular worker (Threshold = JPY 30 million)



(d) Female, non-regular worker (Threshold = JPY 2 million)



(e) Male, non-regular worker (Threshold = JPY 37 million)



(f) Female, non-regular worker (Threshold = JPY 27 million)

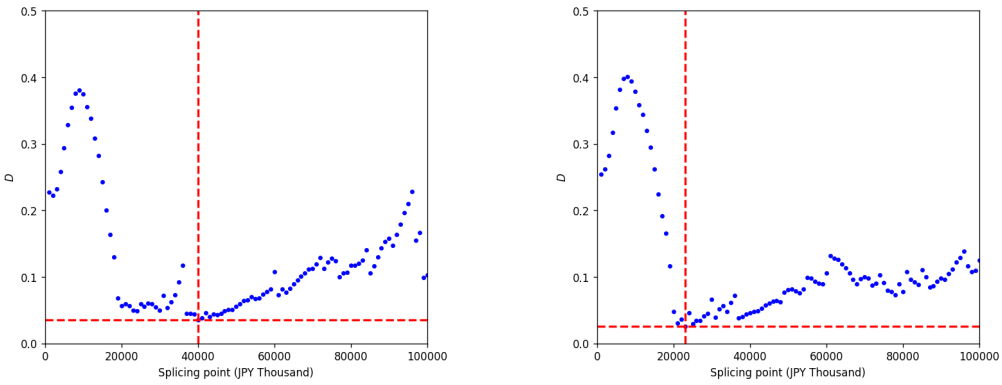
5 Results

5.1 Fitting to the Pareto model

Fig. 12 shows the KS statistic for each splicing point in the distributions of the whole sample in 2014 and 2019. We set 40 and 23 million yen as the splicing points in 2014 and 2019, respectively.

Similarly, we define the splicing points by gender and employment status in Table 5. As shown in Fig. 11, the Zenga curves certify the Paretianity of all distributions by gender and employment status.

Figure 12: Fitting tax agency data to the Pareto model



(a) 2014 (b) 2019

Table 5: Splicing point by gender and employment status (JPY Thousand)

Employment status	2014	2019
Male, regular worker	32000	35000
Male, non-regular worker	21000	30000
Male, other worker	21000	37000
Female, regular worker	3000	24000
Female, non-regular worker	2000	2000
Female, other worker	20000	27000

5.2 Robust Gini coefficient and its decomposition

After integrating the tax agency and household survey data at the splicing point reported in Section 5.1, we calculate the robust Gini coefficients, and decompose them into between- and within- group components of gender and employment status. The results are summarized in Table 6.

The Gini coefficients and between-components of simply-integrated data are larger than those estimated solely from the household survey or tax agency data. However, as stated in Section 4.1, the results of the simply-integrated data may overestimate the coefficients and between-components owing to discard high incomes at around 20–40 million yen.

To avoid the overestimation problem, we apply the parallel integration to the tax agency and household data by gender and employment status. The Gini coefficients of parallelly-integrated data are lower than those of simply-integrated data since the former avoid discarding the data of relatively high incomes. Notably, the coefficients of parallelly-integrated data have increased from 2014 to 2019. Further, the increase is larger than those of household survey and tax agency data. The increase in the coefficients of parallelly integrated data reflects the summation of the change in household survey and tax agency data. Next, the coefficients of the parallelly-integrated data are closer to those of household survey than to those of tax agency data in both 2014 and 2019. In Japan, the contribution of the disparity between top and bottom incomes to Gini inequality index is much lower than that in other countries. This finding which is consistent with Mikayama et al. (2023), who report a small share of top incomes in Japan. This small contribution of top incomes to overall inequality is distinctive from that in other countries(Jenkins (2017); Li et al. (2021)). Next, in contrast to the results of the other three datasets, the between-component of parallelly-integrated data has decreased from 2014 to 2019. Since the parallel integration method enables us to capture the growth of female incomes at around 10 to 90 percentiles (see Fig. (7)), its between-component has decreased from 2014 to 2019.

Table 6: Gini coefficient and its decomposition by gender and employment status after parallel integration

	Gini	Decomposition	
		Between	Within
Household survey data			
2014	0.401	77.8%	22.2%
2019	0.402	78.0%	22.0%
Tax agency data			
2014	0.375	76.8%	23.2%
2019	0.378	77.4%	22.6%
Simply-integrated data			
2014	0.407	78.1%	21.9%
2019	0.413	78.7%	21.3%
Parallelly-integrated data			
2014	0.405	79.4%	20.6%
2019	0.409	78.6%	21.4%

Source: SSASSPS and NSFIE/NSFICW in 2014 and 2019.

Note: The employment status is classified into "Regular worker", "Non-regular worker", and "Others". The coefficients are adjusted by sampling weights.

6 Conclusion

Household survey and tax agency data complement each other to estimate inequality level. This study shows the strong coverage of the top (low) incomes in the tax agency (household survey) data by plotting the empirical CDF and Lorenz curves. Then, we integrate both the data by our proposed parallelly integration method after confirming the Paretianity by Zenga curves and re-estimate the Gini coefficients. Our robust estimation of the Gini coefficients indicates that the change in inequality level and decomposition of the coefficients should be analyzed by combining household survey and tax agency data. Although the difference in the Gini coefficients estimated from widely-used household survey data and our parallelly-integrated

data is not large, the decomposition analysis suffers from an overestimation problem due to the lack of relatively higher incomes, unless we use our proposed parallel integration method.

References

Alvaredo, F. (2011). A note on the relationship between top income shares and the gini coefficient. *Economics Letters*, *110*(3), 274-277. Retrieved from https://www.sciencedirect.com/science/article/pii/S0165176510003460 DOI: https://doi.org/10.1016/j.econlet.2010.10.008

Atkinson, A. B. (2007, 05). Measuring top incomes: Methodological issues. In *Top incomes over the twentieth century: A contrast between continental european and english-speaking countries*. Oxford University Press. Retrieved from https://doi.org/10.1093/oso/9780199286881.003.0002 DOI: 10.1093/oso/9780199286881.003.0002

Atkinson, A. B., Piketty, T., & Saez, E. (2011, March). Top incomes in the long run of history. *Journal of Economic Literature*, *49*(1), 3–71. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jel.49.1.3 DOI: 10.1257/jel.49.1.3

Cabinet Office, G. o. J. (2022). *Japanese Economy 2021-2022* (Tech. Rep.). Retrieved from https://www5.cao.go.jp/keizai3/2021/0207nk/n21_3_3.html

Cirillo, P. (2013). Are your data really pareto distributed? *Physica A: Statistical Mechanics and its Applications*, *392*(23), 5947-5962. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378437113006924 DOI: https://doi.org/10.1016/j.physa.2013.07.061

Dagum, C. (1997). A new approach to the decomposition of the gini income inequality ratio. *Empirical Economics*, *22*, 515–531. Retrieved from https://link.springer.com/article/10.1007/BF01205777#citeas DOI: 10.1007/BF01205777

Flachaire, E., Lustig, N., & Vigorito, A. (2023). Underreporting of top incomes and inequality: A comparison of correction methods using simulations and linked survey and tax data. *Review of Income and Wealth*, *69*(4), 1033-1059. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12618 DOI: https://doi.org/10.1111/roiw.12618

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica*, *84*(334), 261-289. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/ecca.12217 DOI: https://doi.org/10.1111/ecca.12217

Kitao, S., & Yamada, T. (2024). *Inequality Dynamics in Japan, 1981-2021*. Retrieved from https://www.esri.cao.go.jp/en/esri/archive/e_dis/2024/e_dis392-e.html (ESRI Discussion Paper Series. 392)

Kitao, S., & Yamada, T. (2025). Earnings, income, and wealth inequality in japan: a long-term perspective, 1984-2019. *Japanese Economic Review, Special Issue: Heterogeneity and*

Macroeconomics. DOI: https://doi.org/10.1007/s42973-024-00187-0

Kohara, M., & Ohtake, F. (2014, 01). Rising inequality in japan: A challenge caused by population ageing and drastic changes in employment. In *Changing inequalities and societal impacts in rich countries: Thirty countries' experiences.* Oxford University Press. Retrieved from https://doi.org/10.1093/acprof:oso/9780199687428.003.0017 DOI: 10.1093/acprof:oso/9780199687428.003.0017

Kunieda, S., & Yoneta, Y. (2023). *Significance of Analysis Based on Tax Data Related to the Japanese Income Tax System (in Japanese).* Retrieved from https://www.nta.go.jp/about/organization/ntc/kyodokenkyu/kohyo/pdf/230100-01ST.pdf?1741229927967 (National Tax College Discussion Paper Series 230100-01ST)

Li, C., Yu, Y., & Li, Q. (2021). Top-income data and income inequality correction in china. *Economic Modelling*, *97*, 210-219. Retrieved from https://www.sciencedirect.com/science/article/pii/S0264999321000262 DOI: https://doi.org/10.1016/j.econmod.2021.01.018

Mikayama, M., Imahori, T., Ohno, T., Yoneta, Y., & Ueda, J. (2023). *Top Income Shares in Japan from the Survey and Tax Data in 2014 and 2019: Following the Distributional National Accounts Guidelines.* Retrieved from https://www.mof.go.jp/pri/research/discussion_paper/ron371.pdf (PRI Discussion Paper Series 23A-04)

Rinz, K., & Voorheis, J. (2023). *Re-examining Regional Income Convergence: A Distributional Approach.* Retrieved from https://www.census.gov/library/working-papers/2023/adrm/CES-WP-23-05.html (CES Working papers 23-05)