Revisiting income inequality among households: New evidence from the Chinese Household Income Project

Zheng-Xin Wang, Yue-Qi Jv

PII: S1043-951X(23)00124-4

DOI: https://doi.org/10.1016/j.chieco.2023.102039

Reference: CHIECO 102039

To appear in: *China Economic Review*

Received date: 29 June 2022

Revised date: 20 January 2023

Accepted date: 10 August 2023

Please cite this article as: Z.-X. Wang and Y.-Q. Jv, Revisiting income inequality among households: New evidence from the Chinese Household Income Project, *China Economic Review* (2023), https://doi.org/10.1016/j.chieco.2023.102039

Revisiting income inequality among households: new evidence from the Chinese Household Income Project

Zheng-Xin Wang[1], Yue-Qi Jv[2,*]

[1]School of Economics, Zhejiang University of Finance & Economics, Hangzhou 310018, China

[2]School of Economics, Shanghai University of Finance & Economics, Shanghai 200433, China

Abstract: The Gini coefficient has been widely used as a key indicator to measure income inequality. However, differences in the measurement methods and information in the sample are the main reasons for the bias in the Gini coefficient in China. In order to improve the accuracy of the measurement, we revisit income inequality among Chinese families and propose a multi-group Gini coefficient method from the perspective of optimizing the income distribution function. Based on the disposable income of households in the Chinese Household Income Project (CHIP), a generalized logistic distribution function is used to measure national, urban and rural Gini coefficients and their contribution rates. The results indicate that: The multi-group Gini coefficient method based on the particle swarm optimization (PSO) algorithm makes full use of valid microdata-related information, improves the accuracy of traditional methods of fitting urban or rural income distribution and reduces measurement bias based on the realities of China's binary economic structure and the large size of the population. Overall, the income inequality in China has widened over the five-year period from 2013 to 2018. On the one hand, it has been consistently found that the urban-rural income gap is the most important source of income inequality in China (making a contribution exceeding 50%); on the other hand, the contribution of income inequality within urban areas has increased significantly. Education and industry of urban and rural households as well as the difference in their rates of return are the main causes of the income gap between the urban and rural areas in China. Addressing the root causes of income inequality warrants the creation of institutional conditions for equitable access and points of departure in education and industry.

Keywords: income inequality; Gini coefficient; distribution function; Chinese Household Income Project

JEL codes: C61; D30; R20

E-mail: zxwang@zufe.edu.cn (Zheng-Xin Wang);
jujujujuyq@163.sufe.edu.cn (Yue-Qi Jv).

1 Introduction

* Corresponding author. Tel.: +86 15869142165. E-mail address: jujujujuyq@163.sufe.edu.cn (Y.-Q. Jv).

The sixth plenary session of the 19^{th} Central Committee of the Communist Party of China (CPC) in 2021, has pointed out that promoting common prosperity requires the construction of an income distribution system that reflects efficiency and promotes equity. The core of common prosperity is to build an olive-shaped social structure, in which "expanding middle-income groups and raising incomes of low-income groups" is the specific path to achieve this goal, that is, expanding the size of the middle-income group and raising the income of the low-income group. The report of the 20^{th} National Congress of the Communist Party of China makes a profound point: "Modernization of the Chinese style is the modernization of the common prosperity of the entire population. Common prosperity is the essential requirement for socialism with Chinese characteristics, and it is also a historical process over a prolonged period of time." The Gini coefficient (GC) is a key economic indicator of inequality in income distribution. Scientific measurement methods and representative sample information, especially from low-and middle-income groups, will directly affect the accuracy and efficiency of our income distribution policies to expand middle-income groups and raise the incomes of low-income groups. It is difficult to calculate China's Gini coefficient because of its vast land area, large population, and unbalanced levels of regional and urban-rural development. Differences in measurement methodologies and sample information are the primary reasons for the current bias in the measurement of the Gini coefficient in China.

First, it is important to choose a method of measuring the Gini coefficient applicable to the actual distribution of income in China, the fit of the income distribution function plays a crucial role in accurately measuring the Gini coefficient (Ryu *et al.*, 2019).

Existing studies, however, suffer from subjective problems in setting the basic parameters of commonly used income distribution functions (Logistic distribution, Lognormal distribution, *etc*.), thereby departing from the true position of the middle and ends of the income distribution of Chinese families, which will ultimately influence the Gini coefficient value (Li *et al*., 2021). Due to the large differences in the urban and rural pattern of China's income distribution, an effective analysis of the overall income inequality of China's families is complicated by the lack of differentiation in the income distribution of urban and rural families in China. Thus, the measurement method must incorporate the impact of within-and between-group inequality between urban and rural areas in China, and the multi-group Gini coefficient measurement method accounts for the urban-rural income gap as well as the interurban-rural income gap, which coincides with the prevailing situation in China.

Second, the survey sample data should be representative, that is, the data reflect the overall situation of income inequality in China. Since the National Bureau of Statistics of China only publishes macro-scale grouped data of average income in the early years, many scholars could only estimate China's Gini coefficient using limited data (Hu, 2004; Cheng, 2006). However, the use of small sample information is likely to cause bias in the estimation of China's Gini coefficient. In recent years, due to the in-depth work of micro-scale investigations, scholars have made full use of sample information when using micro survey databases. For example, previous research collects a sample of low-income people while getting a sample of high-income people as far as possible (Li *et al*., 2021). The importance of survey sample information for accurate estimation of the Gini

coefficient cannot be overemphasized. Our research finds that household income inequality in China can be measured more accurately using the household disposable income data from the Chinese Household Income Project (CHIP). Furthermore, the data-processing of micro *per capita* disposable income indicators is conducted to achieve a one-to-one mapping of micro household *per capita* disposable income to familial quartiles and to fit the income distribution in China thereto, which in turn can reduce the bias in the estimation of the Gini coefficient in China. Therefore, how better to measure the Gini coefficient in China is an important issue worth studying.

Since both of these factors may induce significant biases in the measurement of the Gini coefficient, by using data from the Chinese Household Income Project micro survey (CHIP2013 and CHIP2018), this research derives a one-to-one mapping from the *per capita* disposable income values of households to the quantiles of the number of households. In addition, to improve the accuracy of income distribution fit, the particle swarm optimization (PSO) algorithm is employed to fit urban and rural disposable income distribution functions with high accuracy. On this basis, the national Gini coefficient is measured and the extent to which between-and within-group inequality contributes to the national Gini coefficient is analyzed. Finally, this research briefly discusses two major sources of income inequality: education and industry. The main contributions of this paper are drawn as follows:

(1) A generalized logistic distribution function in the multi-group Gini coefficient measurement method is proposed, and its core parameters are solved by PSO and the non-linear programming method. The new method expands the original method to a more

general situation, which makes the new method more consistent with the actual situation of the estimation method and provides a scientific method for the Chinese government and academic circles to measure China's income inequality;

(2) A detailed micro-survey of *per capita* disposable income data in the database is used and it maps *per capita* disposable income values to quantiles of the number of households, to depict China's income distribution between urban and rural areas, thus realizing a Gini coefficient estimate that is closer to the real value. This will formulate reasonable and effective income distribution policies of "expanding middle-income groups and raising incomes of low-income groups" and promote common prosperity;

(3) Comparative analysis of the data in 2013 and 2018 shows that there are important structural income differences between urban and rural areas. The main source of the urban-rural income gap is the wage structure effect, *i.e.*, the influence of skewed policies on urbanization. Education and industry between urban and rural areas as well as the difference in their rates of return are the main sources of the urban-rural income gap. Exploring the root causes of income inequality requires the creation of institutional conditions for equitable access and points of departure in education and industry.

The remainder of the research is organized as follows: Section 2 summarizes related literature. Section 3 proposes the multi-group Gini coefficient method. Section 4 includes the empirical analysis based on data processing, mapping, fitting of the income distribution function, error analysis, and urban-rural decomposition. Section 5 further analyzes the sources of income inequality from the perspective of itemized income and impact factors. The conclusions and insights are summarized in Section 6.

2 Review of the related literature

Existing literature offers a variety of methods for measuring income inequality, which can be broadly categorized into normative and empirical methods. In the canonical methods, the Atkinson Index (AKS) is adopted to choose a social welfare function based on value judgments to measure income inequality (Atkinson, 1970). However, the choice of social welfare function shows some limitations in measuring the actual income gap in China by the ratio of the income of the lowest income group to the average income thought. For this reason, most studies have selected empirical methods for measuring income inequality in China, such as the Gini coefficient (1912), the Theil index (Theil, 1967; Knight and Gunatilaka, 2022), as well as the coefficient of variation (Williamson, 1965). Specifically, the Gini coefficient provides a visual measure of inequality across the entire income distribution. Lorenz (1905) proposed the Lorenz curve on a two-dimensional coordinate system, considering the proportion of the area enclosed on a diagonal to the total area to reflect the size of the Gini coefficient. A large body of literature uses the Gini coefficient as a mainstream method of measuring income inequality in China (Luo *et al.*, 2021).

There are two kinds of methods to calculate the Gini coefficient, one is discrete estimation, which can compute the Gini coefficient based on the data on differences in the sample. Discrete estimation methods are proposed by Gini (1912).

$$G = \frac{1}{2n(n-1)\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right| \tag{1}$$

Where x_i is *per capita* disposable income of household i, μ denotes the *per*

capita disposable income of average households, and n is family size. Based on the ideas of Gini (1912) and Lorenz (1905), the researchers conducted a series of studies to measure the Gini coefficient using discrete and continuous estimation methods from algebraic and geometric perspectives. For example, the discrete method was first widely used in academic circles because of its simple calculation. Xu (2003) summarized the discrete Gini coefficient measurement method.

Based on the ideas of Gini (1912) and Lorenz (1905), many researchers proposed a series of extension methods for measuring Gini coefficient from discrete or continuous estimation perspectives.

$$G = \frac{1}{2n^2\bar{x}} \sum_{i=1}^{n} \sum_{j=1} |x_i - x_j|$$

$$= \frac{2\sum_{i=1}^{n} i x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n} = \sum^{n-1} (F_{i+1}P_i - F_i P_{i+1}) = \frac{n+1}{n} - \frac{2}{n^2\bar{x}} \sum_{i=1}^{n} (n+1-i)x_i \quad (2)$$

Continuous estimation of the Gini coefficient can be approached in two ways: by fitting income distribution function (Rothe, 2010) or fitting the Lorenz curve (Wang and Smyth, 2015; Wang, 2019). Lorenz curve fitting has many economic implications, but there may be an estimation error problem. Most scholars at home and abroad measure the Gini coefficient from the angle of the income distribution function. Some researchers used non-parametric or semi-parametric methods to fit the income distribution to improve estimation accuracy (Grazia Pittau and Zelli, 2004). Since the core of the non-parametric method is data-driven, which is not conducive to converting income distribution functions, we tend to choose the parametric method to fit the income distribution. Many researchers

have chosen the Lognormal (Cheng, 2005), Pareto (Wang and Zhou, 2006), Logistic (Cheng, 2006), and Beta II (Han and Cheng, 2019) distributions in empirical studies. However, the income distribution function proposed in the existing literature takes different forms, which is necessary to analyze the real situation. For example, Cheng (2005) believed that the lognormal distribution is suitable for measuring indicators of socio-economic size. However, Wang and Zhou (2006) thought that the tail of the Pareto distribution is thicker than that of the normal distribution and better suited for fitting income distribution. Furthermore, Cheng (2006) found that the logistic distribution is more accurate than the lognormal distribution and the Pareto distribution in fitting the income distribution of Chinese households. However, the logistic distribution is insensitive at both ends of the income distribution, despite a better overall fit (Molero-Simarro, 2017). To improve the fitting accuracy, we propose a generalized logistic distribution, which is a generalization of the traditional logistic distribution, which can hold more information and further improve the fitting accuracy by using optimization algorithms.

It is reasonable to use the income distribution function method in the continuous estimation method to estimate the Gini coefficient of urban and rural families from the perspective of Gini coefficient geometry. First, it is economically important that measure Gini coefficient by fitting continuous distribution or Lorenz curve. The essence of the income gap measurement is based on the measurement of the degree to which the distribution of income deviates from the state of absolute income equality (Lorenz, 1905).

Second, there is the problem of under-representation of a sample of high-income households in data collection (Atkinson, 1970), and it is therefore necessary to predict the

status of both high-and low-income households by means of income distribution functions or curve-fitting. Gan (2012) considered that the Gini coefficient can fluctuate significantly as a result of improvements in data collection and estimation procedures. The goal of many researchers is to improve measurement accuracy, bias correction based on heavy-tailed data (Fontanari *et al.*, 2018), as well as incremental inference techniques based on complex survey data from stratification and aggregation (Bhattacharya, 2007; Davidson, 2009).

The key to measuring the Gini coefficient is also how to use representative data to measure the equity of the income distribution in China. Due to limitations associated with the data, much of the early literature has been adjusted for subgroup data such as sample sizes as small as seven (Cheng, 2006), the reliability of which has yet to be tested. Small sample data cannot reflect the degree of deviation of the distribution function, and then some deviation in the calculation of the Gini coefficients in China. For nearly a decade, some researchers have used micro-survey databases to measure the Chinese Gini coefficient, such as CFPS (Zhang and Churchill, 2020). Based on the household disposable income data of CHIP survey database, the Gini coefficient, urban and rural Gini coefficient and their proportional contributions in China are calculated by using large samples.

It is worth noting that since the dual economic structure in China causes the income distribution of urban and rural families to differ, one effective method for improving the accuracy of Gini coefficient measurement is to use the hybrid model. Ogwang (2000) compared individual models to mixed models but for the purpose of studying binary

economic structures, measuring the Gini coefficient of urban-rural mixtures at both the technical and theoretical levels remains a challenge. The inability to distinguish between urban and rural incomes is not a reflection of the structural problems of income distribution in China, namely, the inability to decompose the national Gini coefficient to assess inequality in the national income structure (Sundrum, 2003). In contrast, it is contrary to practical significance to have a strongly constrained assumption of non-overlapping rural-urban income distribution when computing the national Gini coefficient. For example, Sundrum (2003) proposed national Gini coefficient on the composition of the poor and wealthy, but based on a strong binding assumption that the income distribution of the poor and rich is non-overlapping.

$$G = p_1^2 \frac{\mu_1}{\mu} G_1 + p_2^2 \frac{\mu_2}{\mu} G_2 + p_1 p_2 \frac{|\mu_1 - \mu_2|}{\mu} \tag{3}$$

where, p_1, p_2 denotes the population shares of Groups 1 and 2; μ_1, μ_2, μ denote the *per capita* income of Group 1, Group 2, and overall, respectively. Li (2002) found that the overlap of urban and rural income distribution in China is not suited for analysis using Sundrum's (2003) method. In view of this overlap, many scholars have adapted and proposed extension methods, but with certain assumptions (Cheng, 2007). Cheng (2007) proposed that the national Gini coefficient method can measure the overlap in the underlying income distribution without additional assumptions. Lin (2013) though that the highest income in Cheng (2007) produces a large error in the calculation of the Gini coefficient, and proposes to calculate the Gini coefficient in the form of the sum of the indirect Lorenz curves in urban and rural areas. This also solves the problem of overlapping income distribution, although the error caused by the grouping's critical points

will affect the Gini coefficient to some extent. In fact, Cheng (2007) conducted a sensitivity analysis of the Gini coefficient calculation for the highest income, suggesting that the Gini coefficient is less sensitive to the highest income calculation error. Lin (2013) further noted that Cheng (2007) used grouping data in the validation of the algorithm to replace the highest income in the group with the mean average income in the group, which in turn influenced the calculation of the Gini coefficient nationally. Ai (2015) argued that Cheng (2007) underestimated the urban-rural income gap without assuming a uniform income distribution.

We conclude that: based on the combination of algorithm, data, and application, a multi-group Gini coefficient measurement method has been proposed. The advantages of this method are more comprehensive than those of previous methods with single advantages. Microdata on household disposable income from the CHIP database can be used to map the index of disposable household income and the quantile of household number after data-processing with the goal of improving the accuracy of fitting the income distribution function, using PSO algorithm to fit the urban and rural income distribution function, to measure the urban or rural Gini coefficient, and on that basis to measure the national Gini coefficient, and to estimate the contribution of the inter-group and intra-group inequality. The multi-group Gini coefficient measurement method is proposed, which overcomes the shortcoming of the continuous estimation method and has the characteristics of being consistent with the income distribution in China.

Apart from the difficult problem of measuring income disparity in China, the study of income inequality in China from the perspective of the source of income or the

influencing factors represents a new direction for such research. To study income inequality in China, it is important to explore factors such as educational expansion (Yang and Gao, 2018), and industry differences (Chen and Wan, 2011), which are closely linked to income inequality.

In addition, domestic and international analysis of the impact of itemized income on income inequality in China from an income source perspective is relatively small, and many studies in recent years have focused on regression decomposition (Wan and Zhou, 2005). Oaxaca-Blinder decomposition is widely used for its simplicity and policy implications, which can decompose the mean difference between groups in each explanatory variable and into composition effect and wage structure effect. However, this method can only be decomposed at the inter-group mean, and the Oaxaca-Blinder Re-centered Influence Function (Firpo *et al.*, 2009) can estimate the effect of changes in explanatory variables on any distributional statistic of the variable (*e.g.* Gini coefficient) based on the re-centered influence function (Firpo *et al.*, 2018). Yang and Gao (2018), for example, based on CHIP 2002 and CHIP 2013, estimated the influence of educational expansion on the wage gap. Policy tilts (Mendoza, 2016) and industries (Gittleman and Wolff, 1993; Chen and Wan, 2011) also have a significant effect on income inequality.

3 Measurement of the national Gini coefficient

3.1 A single measure of the Gini coefficient

In an economy, household disposable income *per capita* can be considered as a continuous random variable $\Theta \in \left[\underline{\theta}, \overline{\theta}\right]$, where $\underline{\theta}$ and $\overline{\theta}$ denote the lower and upper

bounds of household disposable income *per capita* in the economy, respectively. It can be assumed that the distribution function of household disposable income *per capita* in an economy $F(\theta)$ represents number of households with income Θ not greater than θ as a proportion of total households in the economy. The definition of the Lorenz curve shows that the Lorenz curve is explicitly related to the distribution function of household disposable income *per capita*. The horizontal axis of the two-dimensional coordinate system (x, y) is the cumulative proportion of the number of households and the vertical axis represents the cumulative proportion of income distribution. The following is the parametric equation of the Lorenz curve (Cheng, 2006).

$$y = L(x) \begin{cases} x = F(\theta) \\ y = \dfrac{\int_{\underline{\theta}}^{\bar{\theta}} P_0 \theta dF(\theta)}{I_0} \end{cases} \tag{4}$$

where, $y = L(x)$ is the Lorenz curve, P_0 is the total number of households in the economy, and I_0 is the sum of household disposable income *per capita* in the economy. Assuming function $L:[0,1] \to [0,1]$; a denotes the area enclosed by the curve $y = L(x)$, the y-axis and $y = 1$. The parametric equation based on the Lorenz curve is given by:

$$\begin{aligned} a &= \int_0^1 x dy = \int_{\underline{\theta}}^{\bar{\theta}} F(\theta) \frac{P_0}{I_0} \theta dF(\theta) \\ &= \frac{P_0}{I_0} \left[\frac{1}{2} \theta F^2(\theta) \Big|_{\underline{\theta}}^{\bar{\theta}} - \frac{1}{2} \int_{\underline{\theta}}^{\bar{\theta}} F^2(\theta) d\theta \right] \end{aligned} \tag{5}$$

where, $dy = \dfrac{P_0 \theta dF(\theta)}{I_0}$, $s = \begin{cases} s = \underline{\theta}, y \to 0 \\ s = \bar{\theta}, y \to 1 \end{cases}$.

According to the formula for the Gini coefficient $G_s = 2a - 1$, then

$$G_s = \frac{P_0}{I_0}\left[\theta F^2(\theta)\Big|_{\underline{\theta}}^{\bar{\theta}} - \int_{\underline{\theta}}^{\bar{\theta}} F^2(\theta)d\theta\right] - 1 \tag{6}$$

The number of households in the whole economy is large enough, thus satisfying $F(\underline{\theta}^+) \approx F(\underline{\theta}^-) = 0$, $F(\bar{\theta}^-) \approx F(\bar{\theta}^+) = 1$.

$$G_s = \frac{P_0}{I_0}\left[\bar{\theta} - \int_{\underline{\theta}}^{\bar{\theta}} F^2(\theta)d\theta\right] - 1 \tag{7}$$

Then

$$I_0 = P_0\int_{\underline{\theta}}^{\bar{\theta}} \theta dF(\theta) = P_0\left[\bar{\theta}F(\bar{\theta}^-) - \underline{\theta}F(\underline{\theta}^+) - \int_{\underline{\theta}}^{\bar{\theta}} F(\theta)d\theta\right] = P_0\left[\bar{\theta} - \int_{\underline{\theta}}^{\bar{\theta}} F(\theta)d\theta\right] \tag{8}$$

Therefore, the corresponding Gini coefficient is:

$$G_s = \frac{\bar{\theta} - \int_{\underline{\theta}}^{\bar{\theta}} F^2(\theta)d\theta}{\bar{\theta} - \int_{\underline{\theta}}^{\bar{\theta}} F(\theta)d\theta} - 1 \tag{9}$$

3.2 A multi-group measurement of the Gini coefficient

In time, the Gini coefficient within the economy changes accordingly. There are N of different groupings within the economy, representing different economic levels. The number of clusters of the category n is denoted as M_n. The household disposable income *per capita* of the economy of category n in year t is represented by the random variable Θ_{nt}, $\Theta_{nt} \in [\underline{\theta}_{nt}, \bar{\theta}_{nt}]$ and the corresponding distribution function of household disposable income *per capita* is denoted as $F_{nt}(\theta)$, $n = 1,\ldots,N$, $t = 1,\ldots,T$. The total population within the economy is $P_{0t} = \sum_{n=1}^{N} M_{nt}$. The household disposable income *per capita* of the economy of year t without differentiation of clusters is represented by the random variable Θ_t and the distribution function $F_t(\theta)$, interval $\left[\underline{\theta}_t, \bar{\theta}_t\right]$, where

$\underline{\theta}_t = \min_{1\leqslant n\leqslant N}(\underline{\theta}_{nt})$, $\overline{\theta}_t = \max_{1\leqslant n\leqslant N}(\overline{\theta}_{nt})$. The definition of the distribution function can be obtained in year t , for a category with n clusters. The number of units with household disposable income *per capita* less than θ within the group is:

$$M_{nt}P\{\Theta_{nt} < \theta\} = M_{nt}F_{nt}(\theta) \tag{10}$$

Then the percentage of the population in the economy with Θ_{nt} less than θ is

$$\frac{\overline{M}_t}{P_{0t}} = \frac{\sum_{n=1}^{N} M_{nt}F_{nt}(\theta)}{P_{0t}} \tag{11}$$

A composite income distribution function expressed as

$$F_t(\theta) = P\{\Theta_t < \theta\} = \sum_{n=1}^{N} \alpha_{nt}F_{nt}(\theta) \tag{12}$$

where, $\alpha_{nt} = \frac{M_{nt}}{P_{0t}}$, $\sum_{n=1}^{N}\alpha_{nt} = 1$

Considering the heterogeneity among different clusters and substituting the composite income distribution function into the Gini coefficient calculation formula (6), the formula for the multi-group composite Gini coefficient in year t is obtained:

$$G_t = \frac{\overline{\theta}_t - \int_{\underline{\theta}_t}^{\overline{\theta}_t}\left(\sum_{n=1}^{N}\alpha_{nt}F_{nt}(\theta)\right)^2 d\theta}{\overline{\theta}_t - \sum_{n=1}^{N}\int_{\underline{\theta}_t}^{\overline{\theta}_t}\alpha_{nt}F_{nt}(\theta)d\theta} - 1 \tag{13}$$

When $N = 2$, Equation (13) is simplified by solving for the national Gini coefficient:

$$G_t = \frac{\overline{\theta}_t - \int_{\underline{\theta}_t}^{\overline{\theta}_t}\left(\alpha_{1t}F_{1t}(\theta) + \alpha_{2t}F_{2t}(\theta)\right)^2 d\theta}{\overline{\theta}_t - \int_{\underline{\theta}_t}^{\overline{\theta}_t}\left(\alpha_{1t}F_{1t}(\theta) + \alpha_{2t}F_{2t}(\theta)\right)d\theta} - 1 \tag{14}$$

where urban population $\alpha_{1t} = \frac{M_{1t}}{P_{0t}}$ is the proportion of total population and rural population $\alpha_{2t} = \frac{M_{2t}}{P_{0t}}$ is the proportion of total population, assuming $A_{1t} = \int_{\underline{\theta}_{1t}}^{\overline{\theta}_{1t}} F_{1t}^2(\theta)d\theta$, $A_{2t} = \int_{\underline{\theta}_{2t}}^{\overline{\theta}_{2t}} F_{2t}^2(\theta)d\theta$, $B_{1t} = \int_{\underline{\theta}_{1t}}^{\overline{\theta}_{1t}} F_{1t}(\theta)d\theta$, $B_{2t} = \int_{\underline{\theta}_{2t}}^{\overline{\theta}_{2t}} F_{2t}(\theta)d\theta$, $C_t = \int_{\underline{\theta}_{1t}}^{\overline{\theta}_{1t}}\left(F_{2t}(\theta)\cdot F_{1t}(\theta)\right)d\theta$, $\overline{\theta}_t = max(\overline{\theta}_{1t}, \overline{\theta}_{2t}) = \overline{\theta}_{1t}$, $\underline{\theta}_t = \underline{\theta}_{1t} = \underline{\theta}_{2t}$.

Since $F_{2t}(\theta)=1$ when $\theta > \bar{\theta}_{2t}$, then

$$\int_{\underline{\theta}_{2t}}^{\bar{\theta}_{1t}} F_{2t}^2(\theta)d\theta = \int_{\underline{\theta}_{2t}}^{\bar{\theta}_{2t}} F_{1t}^2(\theta)d\theta + (\bar{\theta}_{1t} - \bar{\theta}_{2t}) = A_{2t} + (\bar{\theta}_{1t} - \bar{\theta}_{2t}); \tag{15}$$

$$\int_{\underline{\theta}_{2t}}^{\bar{\theta}_{1t}} F_{2t}(\theta)d\theta = \int_{\underline{\theta}_{2t}}^{\bar{\theta}_{2t}} F_{2t}(\theta)d\theta + (\bar{\theta}_{1t} - \bar{\theta}_{2t}) = B_{2t} + (\bar{\theta}_{1t} - \bar{\theta}_{2t}). \tag{16}$$

Finally,

$$G_t = \frac{\bar{\theta}_t - \left[\alpha_{1t}^2 A_{1t} + \alpha_{2t}^2 A_{2t} + \alpha_{2t}^2(\bar{\theta}_{1t} - \bar{\theta}_{2t}) + 2\alpha_{1t}\alpha_{2t}C_t\right]}{\bar{\theta}_t - \left[\alpha_{1t}B_{1t} + \alpha_{2t}B_{2t} + \alpha_{2t}(\bar{\theta}_{1t} - \bar{\theta}_{2t})\right]} - 1. \tag{17}$$

3.3 Fitting and optimization of distribution functions

The key to the multi-group Gini coefficient method we propose is how to fit the income distribution function with high accuracy. Compared with income distribution functions such as Normal, Lognormal, or the logistic distribution function may be more consistent with the income distribution in China (Cheng, 2006). Besides, Cheng (2006) corrected the original exponent (set to 1) of the dependent variable to 0.05; however, we find that there is a certain problematic subjectivity associated with taking the value of 0.05, and there is some variability in the characteristics of the distribution function across different years and different groups. Therefore, a general logistic distribution function is proposed (Equation (18)). Based on the PSO algorithm, the optimal index is determined:

$$F(\theta) = \frac{1}{1 + ae^{-b\theta^{\omega}}} \tag{18}$$

where, a, b and ω are the parameters of interest. Since the income distribution function is non-linear, there exists an excessive number of iterations in the fitting process, which brings inconvenience to the measured results. To reduce the difficulty of estimation, the non-linear form of Equation (18) is transformed into a linear

form. Let $y = \ln\left(\frac{1}{F(\theta)} - 1\right)$, $x = \theta^{\omega}$, $p = -b$, $q = \ln a$, then

$$y = \ln\left(\frac{1}{F(\theta)} - 1\right) = \ln a - b\theta^{\omega} = px + q \tag{19}$$

This may be abbreviated to

$$y = \hat{p}x + \hat{q} + \varepsilon \tag{20}$$

where, ε is the error term. When the value of ω is determined, the parameters $\hat{p}$ and $\hat{q}$ can be estimated by the ordinary least squares method. The problem is how to determine the optimal parameters ω. The parameters a, b, and ω are found using the PSO algorithm in conjunction with a non-linear programming method.

$$\begin{aligned} &\min_{\omega} MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, i = 1, 2, \ldots, N \\ &s.t. \begin{cases} y_i = \ln\left(\frac{1}{F(\theta_i)} - 1\right) \\ \hat{y}_i = \hat{p}x_i + \hat{q} \\ x_i = \theta_i^{\omega} \end{cases} \end{aligned} \tag{21}$$

Particle swarm optimization (PSO) algorithm is a type of population intelligence optimization algorithm in the field of computational intelligence, the core of which is to use the sharing of information by individuals in the population to make the motion of the whole population evolve from disorder to order in the problem solution space, to obtain the optimal solution of the problem (Shi and Eberhart, 1998).

Assume a population $X = (X_1, X_2, \ldots, X_N)$ of N particles in a D-dimensional space, where the i^{th} particle represents a D-dimensional vector $X_i = (x_i^1, x_i^2, \ldots, x_i^D)$ and the adaptation value corresponding to each particle position X_i is calculated according to the objective function. The velocity of the i^{th} particle is $V_i = (V_{i1}, V_{i2}, \ldots, V_{iD})^T$, the individual extremum is $P_i = (P_{i1}, P_{i2}, \ldots, P_{iD})^T$, and the

population extremum of the population is $P_g = (P_{g1}, P_{g2}, \ldots, P_{gD})^T$. The particle updates its velocity and position by the individual and population extremums:

$$V_{id}(k+1) = \psi V_{id}(k) + c_1 r_1 \left(P_{id}(k) - X_{id}(k)\right) + c_2 r_2 \left(P_{gd}(k) - X_{id}(k)\right) \tag{22}$$

$$X_{id}(k+1) = X_{id}(k) + V_{id}(k+1) \tag{23}$$

where, ψ is the inertia weight, $d = 1, 2, \ldots, D$, $i = 1, 2, \ldots, N$, k is the number of iterations, V_{id} is the velocity of the particle, c_1 and c_2 refer to a non-negative constant acceleration factor ($c_1 = c_2 = 2$). r_1 and r_2 denote a random number distributed in the interval $[0,1]$.

The PSO algorithm solves for the core parameters in the following steps: (1) Determining the number of initialized particle populations, the individual optimal solution, and the global optimal solution; (2) Updating the positions and velocities of the particles according to the iterative formulae (22) and (23); (3) Calculating the fitness value of each particle, and the fitness function is set as the mean square error (MSE); (4) Comparing each particle fitness value with the individual optimum, and updating it if it is better than the individual optimum; (5) Comparing each particle fitness value with the global optimum and updating the solution if it is better than the global optimum; (6) The end of the iteration is marked by reaching the maximum number of iterations or higher than the accuracy of the objective function, then the iteration stops and the search are performed to obtain the optimal index. The process of estimating a more general Logistic income distribution function is shown in Fig.1.

Fig. 1 Flowchart of particle swarm optimization (PSO)-based multi-group Gini coefficient

method.

3.4 The decomposition and the urban-rural income inequality in China

The optimal income distribution function based on the PSO algorithm can be expressed as follows:

$$G_t = \frac{\bar{\theta}_t - \left[\sum_{i=1}^{2} \alpha_{it}^2 \int_{\underline{\theta}_{it}}^{\bar{\theta}_{it}} \left(\frac{1}{1+\hat{a}_{it}e^{-\hat{b}_{it}\theta^{\hat{\omega}_{it}}}}\right)^2 d\theta + \alpha_{2t}^2(\bar{\theta}_{1t} - \bar{\theta}_{2t}) + 2\alpha_{1t}\alpha \right.}{\bar{\theta}_t - \left[\sum_{i=1}^{2} \alpha_{it} \int_{\underline{\theta}_{it}}^{\bar{\theta}_{it}} \left(\frac{1}{1+\hat{a}_{it}e^{-\hat{b}_{it}\theta^{\hat{\omega}_{it}}}}\right) d\theta + \alpha \right.}$$

where, urban households as a share of total households $\alpha_{1t} = \frac{M_{1t}}{P_{0t}}$, rural households as a share of total households $\alpha_{2t} = \frac{M_{2t}}{P_{0t}}$, $\bar{\theta} = \max\left(F_{1t}^{-1}(1-\frac{1}{P_{0t}}), F_{2t}^{-1}(1-\frac{1}{P_{0t}})\right)$.

It is essential to obtain the specific form of the household disposable income *per capita* distribution. The sample data and the income distribution function directly affect the accuracy of measuring the Gini coefficient in China. In terms of sample data, we choose household disposable income *per capita* data from the CHIP database are chosen and processed to achieve a one-to-one mapping between the values of household disposable income *per capita* and household quantile indicators. In terms of selection of income distribution functions, a general Logistic income distribution function is proposed, and the fitted income distribution function is fitted based on the PSO algorithm. In addition, the model results are compared by three indicators: adjusted goodness of fit $Adj - R^2$, mean absolute parentage error ($MAPE$) and root mean square error ($RMSE$).

$$Adj-R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)} \tag{25}$$

$$MAPE=\sum_{i=1}^{N}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{26}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{27}$$

Furthermore, the contribution of intra-urban, intra-rural income inequality and inter-urban-rural income inequality to the degree of fairness of income distribution in China is evaluated. The decomposition of the Gini coefficient for urban and two rural groups is given (Ai, 2015) by:

$$G_{two}=w_1\alpha_1G_1+w_2\alpha_2G_2+\alpha_1\alpha_2\cdot\frac{\vartheta}{u} \tag{28}$$

where, w_1 and w_2 denote the proportion of urban total household income and rural total household income to the total household income in China, respectively; $w_1+w_2=1$. α_1, α_2 denote the total number of urban and rural households as a proportion of the total number of households in China, respectively. $\vartheta=\int_{\theta}^{\bar{\theta}_1}\left(F_1(\theta)+F_2(\theta)-F_1(\theta)F_2(\theta)\right)d\theta$, ϑ can be used as a quantitative indicator to measure the absolute urban-rural gap. In addition, this research obtains the relative urban-rural gap indicator in dimensionless form $\psi=\frac{\vartheta}{u}=\frac{\int_{\theta}^{\bar{\theta}_1}(F_1(\theta)+F_2(\theta)-F_1(\theta)F_2(\theta))d\theta}{u}$.

4 Measurement of income inequality in China

4.1 Data sources, data processing and descriptive analysis

The micro-data were sourced from the Chinese Household Income Project (CHIP)

[1]. CHIP data are currently authoritative and represent accurate micro-data pertaining to income. The two most recent surveys CHIP2013 and CHIP2018 are obtained by using the scientific stratified systematic sampling method in eastern, middle, and western China. To measure the national Gini coefficient accurately and reasonably, household disposable income data are used.

It is worth noting that, the measurement of Gini coefficient in China is complicated because of the large population size and imbalance in the level of development across China, which greatly increases the difficulty of measurement (Li *et al*., 2020). Simple measurement methods and general data could not measure income inequality and standard of living inequality in China to the standard required. We need not only representative microdata, but also the processing aspects of the data and the suitability of measurement methods in practice. (i) the specific data from CHIP 2013 and CHIP 2018 are processed to calculate household disposable income *per capita* in urban and rural China in 2011-2013 and 2018. Table 1 shows the basic information of household disposable income *per capita* in China. (ii) There is a significant income imbalance between the various provinces of China, and the availability of high-quality data on income makes it almost impossible to undertake a scientifically stratified random sample thereof. The income-based databases, CHIP 2013 and CHIP 2018, cover a sample of 15 provinces, and sample information needs to be weighted to allow comprehensive measurement of income inequality in China (Solon and Haider, 2015). We correct the

[1] http: //www.ciidbnu.org/chip/.

sample information of regional and provincial weights. The Gini coefficient of China's households is measured by mapping the income and household percentiles one-by-one to describe the distribution of household income in urban and rural China as accurately as possible.

Table 1 Descriptive statistics of per-capita disposable income for households

Year	Region	Sample size (family)	Per-capita disposable income (yuan)	Standard deviation (yuan)	Coefficient of variation	Max.	Min.
2011	Rural	10,244	10,257	9838	0.959	125	350,000
2011	Urban	7255	21970	19159	0.872	504	580000
2012	Rural	10,244	11,384	10,852	0.953	125	460,000
2012	Urban	7255	24,372	20,366	0.836	520	600,000
2013	Rural	10,244	12,877	12,136	0.942	117	533,333
2013	Urban	7255	27,640	23,492	0.85	540	840,860
2018	Rural	8961	14,867	14,367	0.966	15	296,196
2018	Urban	11,344	41,238	32,913	0.798	281	1,053,561

The *per capita* urban disposable income of households in urban areas has risen from 21,970 yuan in 2011 to 41,238 yuan in 2018. From 2011 to 2018, household disposable rural income *per capita* increased from 10,257 yuan to 14,867 yuan, implying that rural growth is smaller than urban growth. The coefficients of variation (CV) for

households in 2011, 2012, 2013, and 2018 for urban disposable income *per capita* are 0.872, 0.836, 0.85, and 0.798, respectively, whereas the CV for households in rural areas is 0.959, 0.953, 0.942, and 0.966. Income inequality within urban households fell substantially in 2018 from 2011 to 2013. There is a small increase in the income gap within rural households in 2018 compared to 2011. In addition, urban CV is lower than rural CV throughout the year, perhaps indicating a greater intra-rural income differential than urban CV. Compared to macro-level proportional data and discrete time estimation methods, PSO algorithms and non-linear programming are used to match the income distribution in China based on micro-level data. Finally, it is worth mentioning that full use is made of the information contained in the micro data and that it is more efficient to fit the two extreme cases of the income distribution function. However, the CHIP data must be treated in some way to ensure that the value of household disposable income *per capita* achieves an effective one-to-one mapping to the quantile of families and to exclude $F(\theta)=1$ from the sample to avoid going beyond the realm of definition.

4.2 Fitting urban-rural income distribution functions based on PSO algorithm

A generalized logistic distribution function is constructed to fit the distribution of urban and rural incomes, respectively, the non-linear function is transformed to a linear function, and the exponent is determined by use of the PSO algorithm. Given the parameters p, q, and ω, the income distribution function is known to vary across years, and is not a fixed distribution function. One way to predict aggregate high and low

incomes is to fit the income distribution function to microdata in an intelligent manner. The precision of income distribution fitting such as $MAPE$, $RMSE$, and goodness of the fit index $Adj-R^2$ is shown in Table 2.

Table 2 Parameters and precision of income distribution fitting based on PSO algorithm

Data	N	pair	p	q	ω	R^2	$Adj-R^2$	$MAPE$	$RMSE$
2011-r	7255	1813	-11.73***	27.40***	0.092	0.9978	0.9978	0.077	0.1136
2011-u	10,244	1478	-7.56***	23.34***	0.1156	0.9957	0.9957	0.0646	0.1554
2012-r	7255	1818	-13.09***	28.89***	0.087	0.998	0.998	0.0727	0.107
2012-u	10,244	1575	-10.14***	27.53***	0.1013	0.997	0.997	0.0546	0.1294
2013-r	7255	3032	-11.73***	27.40***	0.092	0.9974	0.9974	0.0731	0.1117
2013-u	10,244	2668	-13.19***	32.13***	0.0891	0.997	0.997	0.057	0.1202
2018-r	11,344	8957	-3.24***	14.35***	0.1588	0.9982	0.9982	0.0372	0.0772
2018-	8961	11,25	-17.05*	36.64*	0.073	0.998	0.9982	0.059	0.078

u		8	**	**	8	2		9	7

Note: *** indicates $p < 0.01$.

Based on the PSO algorithm, this research searches for the index value with a minimum error by setting 300 evolutionary processes and 100 population sizes. The values of n in urban and rural areas vary greatly in different years, with the urban income index taking the value of 0.1588 and the rural income index taking the value of 0.0738 in 2018, and the other years also vary greatly, indicating that, unlike the traditional way of setting fixed values for the Logistic function, the Logistic function constructed by the PSO algorithm has better adaptability. In addition, the traditional logistic distribution fitting urban and rural income distribution functions are prone to rise faster than the actual distribution, which causes some errors in the measurement of the Gini coefficient. Using the generalized logistic distribution, the fit of the distribution function of *per capita* urban and rural household disposable income in 2011-2013 and 2018 in Fig. 2 is aligned with the distribution of the data in the sample and the overall effect of the fit is good.

Fig. 2 Fitted income distribution functions

4.3 Results of the measurement of urban-rural Gini coefficients

Based on the known income distribution function, the highest income is introduced through its inverse function, which is simplified by the following formulae:

$$\bar{\theta} = F^{-1}\left(1-\frac{1}{P}\right) = \left[\frac{\ln a + \ln(P-1)}{b}\right]^{\frac{1}{\omega}} \tag{29}$$

The Gini coefficients are measured for urban and rural areas as well as urban-rural household disposable income *per capita*, which are shown in Figs. 3 and 4. The urban income (39,445 yuan) is much higher than the *per capita* income in rural areas (15,165 yuan) in 2018. From 2011 to 2013, the difference between urban and rural living standards was also evident in household disposable income on a *per capita* basis. The urban income increased from 20,826 yuan in 2011, to 39,445 yuan in 2018, and rural incomes had risen from 10,189 yuan to 15,165 yuan between 2011 and 2018. However, from the point of view of the inequity of the income distribution, the urban Gini coefficient was 0.354 in 2011, which belongs to a relatively reasonable income gap. The Gini coefficient for urban area declined slightly from 0.354 to 0.338 per annum from 2011 to 2013, which shows that over these three years, the gap in urban incomes was steadily narrowing and was relatively equitable. Whereas the rural Gini coefficient stabilized around 0.38-0.388 in 2011 to 2013 but was higher than the urban income gap. Urban and rural Gini coefficients were both higher in 2018 than in 2011 to 2013, at 0.359 and 0.405 respectively. The rural Gini coefficient in 2018 exceeded 0.4 therein, which is in the high-income inequality category that should be noted. In summary, the accuracy of the optimal multi-group Gini coefficient measurement method is superior.

Fig. 3 Estimates of *per capita* disposable income of urban or rural households in 2011-2013 and 2018

Fig. 4 Urban or rural Gini coefficients in 2011-2013 and 2018

4.4 Comparison of errors

To measure China's Gini coefficient as realistically as possible, it is essential to fit the distribution of urban and rural incomes in China. It is therefore necessary to use the PSO algorithm to seek optimal index values, the way to map microsamples one-by-one, adjust the regional and provincial weights of samples, and the binary structure of urban and rural income. Error differences are compared between one condition and fixated on the other conditions. Taking CHIP2018 as an illustrative example, column (1) of Table 3 is based on the PSO algorithm, adjusted weights and satisfies the multi-group Gini coefficient measurement method for the treatment of micro-survey mapping data sample-by-sample; column (2) and column (3) list an index of subjective choice power; column (4) is an unweighted coefficient; column (5) is a subset of the data for 17 sample sizes; column (6) lists a single group which does not distinguish between urban and rural areas. Herein, this research proposes, in column (1), for example, a lower $MAPE$ and $RMSE$ and a higher $Adj-R^2$ than methods that do not use PSO algorithms (columns (2) and (3)). Furthermore, the error associated with the traditional logistic distribution (power index 1) is too large to recommend.

Intuitively, Table 3 shows that the Gini coefficient for cities or villages is overestimated as a direct result of the power index adjustment issues (column (3) compared to column (1)): in 2018, the Gini coefficients in urban and rural areas with a generalized logistic distribution were 0.359 and 0.405, respectively, compared with 0.426

and 0.467 for the traditional logistic measures based on PSO algorithms. In addition, the urban-rural Gini coefficient was also overestimated, resulting in an overall Gini coefficient of 0.435 as compared to 0.491 in the absence of the optimization algorithm. It is shown that PSO algorithm plays a very important role in the process of calculating Gini coefficient in China and directly affects the goodness of fit of the income distribution function and the accuracy of Gini coefficient measurement in China. The results in columns (4), (5), (6), and (1) also point to the need for precision and rigor in the measurement procedure, which is of great importance in reflecting the urban, intra-rural, rural-urban income gap as well as the proportional contributions thereof.

Table 3 Precision of power exponential change and other approaches in 2018

Data	*PSO*	$N = 0.5$	$N = 1$	Unweighted	Group data	Single-group
	(1)	(2)	(3)	(4)	(5)	(6)
Urban						
MAPE	0.0599	0.0609	0.5831	0.0571	0.0402	0.0965
$Adj - R^2$	0.9982	0.9979	0.7496	0.9981	0.9995	0.9955
RMSE	0.0797	0.0853	0.9324	0.08	0.0456	0.126
Gini	0.359	0.362	0.426	0.367	0.375	
Contribution rate	39.00%	38.40%	40.50%	39.40%	40.00 %	
Rural						

MAPE	0.0372	0.0859	0.6688	0.0383	0.0418	
$Adj-R^2$	0.9982	0.9877	0.6834	0.9983	0.9998	
RMSE	0.0772	0.2002	1.0161	0.0756	0.0496	
Gini	0.405	0.436	0.467	0.407	0.4	
Contribution rate	7.80%	8.50%	8.30%	7.20 %	7.30 %	
Urban-rural/ Overall						
$U-RGini$	0.96	0.977	1.043	0.932	0.973	
Contribution rate	53.20%	53.20%	51.10%	53.4%	52.60%	
Overall	0.435	0.443	0.491	0.447	0.445	0.451

4.5 Results of the national Gini coefficient measurement

The national Gini coefficient is calculated using Equation (14) and this is based on the measurement of urban and rural Gini coefficients. In Fig. 5, the indicator for China's rural-urban relative income gap gradually decreased from 0.914 to 0.88 from 2011 to 2013, reflecting a slight trend in the Chinese income inequality between rural and urban incomes between 2011 and 2013 (the trend was also reflected in the literature), which is due to the high priority given by China to the eradication of poverty in rural areas in general, including implementing a range of fiscal policies to support agriculture and closing the income gap between families in urban and rural areas. When comparing 2013 to 2018, it is noteworthy that the relative income gap between urban and rural areas grew significantly, reaching 0.96 in 2018. The overall Gini coefficient also shows a consistent pattern across countries.

The resulting Gini coefficients were 0.412, 0.404, 0.399, and 0.435 in 2011, 2012, 2013, and 2018, respectively. In 2011-2013, national Gini income inequality has narrowed slightly, but in 2018, the income gap widened. The comparability, deviation, and precision of the Gini coefficient are affected by inconsistency in research caliber, representativeness of the data, and scientific method of measurement. In the present work, the CHIP survey data that are most representative of income are selected, taking the family unit as a unit, discussing the family income gap in China to evaluate the equity of the standard of living in China. Using data from the China Household Finance Survey as a guide, Gan *et al.* (2012) found that the 2012 Gini coefficient of China is 0.61, which is much larger than the results arising from the present research (and indeed that arising from work undertaken by the National Bureau of Statistics of China). Some researchers (Yue and Li, 2013a; Yue and Li, 2013b; Li and Wan, 2013) pointed out that inadequate sampling methods, a skewed data structure, and underestimation of incomes among low-income groups are the primary reasons leading to overestimation of the Gini coefficient.

Fig. 5 Gini coefficient of multiple groups and urban-rural income gap.

Table 4 shows a comparison of the multi-group Gini coefficient measurement method discussed above, the single group Gini coefficient method of measurement, the method relying on group data, and other simple methods (the discrete Gini coefficient method, Theil entropy index, Mean log definition, Entropy index, and so on). Despite the

complexity of the multi-group Gini coefficient method in the present research, it has two important advantages when measuring the income inequality in China. First, it is consistent with the unbalanced development of dual economic structure and unbalanced provinces in China, and measures the overall income inequality, urban or rural income inequality, urban-rural income gap, and the proportional contributions thereof. Secondly, the high accuracy is of benefit: the PSO algorithm gives a good fit to the income distribution of urban and rural areas in China. The different entropy methods, the discrete Gini coefficients, and simple statistical indicators (CV, Standard definition of logs, *etc.*) do not fully enshrine these advantages.

Table 4 Comparative analysis of income gap indicators

Methods	Economic implications	Conforming to China	Sample represents population	Fitting income distribution	Fitting accuracy	Gini
1. Multi-group Gini coefficient method	Great: Lorenz curves, decompositions do not require	Great: Urban/rural duality, micro-survey data, mapping,	Great: Regional and provincial weighted, large	Great: Micro-survey data Mapping pairs: 20,215	Great: PSO algorithm, great precision	0.435

	overlapping constraints	weighting	sample size, high quality micro-survey database CHIP			
2. Single-group Gini coefficient method	Great: Same as Method 1	Little: Urban/rural duality not considered	Great: Same as Method 1	Mapping pairs: 20,190	Good precision	0.451
3. Using little data	Great: Same as Method 1	Great: Urban/rural duality	Little: Information on poverty	Grouping data: 17	Good precision	0.445
4. Discrete Gini coefficient	The difference between any one sample pair	Little: sample representation problem	None	None	0.450	
5. Theil	Entropy	Great:	Little:	None	None	0.35

entropy index ($GE(a), a=1$)	theory	Urban/rural duality, decomposability	same as Method 4			2
6.Relative mean deviation	Simple statistical theory	Little: Failure to decompose	Little: same as Method 4	None	None	0.327
7.Cofficient of variation	Simple statistical theory	Little: Failure to decompose	Little: same as Method 4	None	None	0.995
8. Standard deviation of logs	Simple statistical theory	Little: Failure to decompose	Little: same as Method 4	None	None	0.927

Note: Multi-group Gini coefficient method represents China's income inequality (binary structure, weighting, full use of information without grouping data, generalized logistic distribution) and high precision (PSO).

4.6 Decomposition of the national Gini coefficient and its contribution

The decompositions of the national Gini coefficient are shown in Table 5 and Fig. 6. Equation (31) estimates the national Gini coefficient in the weighted urban-rural income

gap as well as the proportional contributions thereof. In total, Gini coefficients in China are 0.412, 0.404, 0.399, and 0.435 in 2011-2013 and 2018, respectively, almost all of which are in excess of 0.4 (the International Alert Level). Income inequality in China increased slightly from 2011 to 2018. This can be seen by the fact that the rural-urban relative income gap has fallen from 0.914 in 2011 to 0.880 in 2013, while in 2018 it reached 0.960. It is notable that the urban-rural income gap rose substantially in 2018 compared to 2011, whereafter it increased to 2013. Similarly, the urban-rural income gap exhibits the same trend as the urban-rural gap. Intra-city income differences of 0.354, 0.345, 0.338 and 0.359 are observed. In the 2011-2013 and 2018 intra-rural income differentials were 0.388, 0.382, 0.380 and 0.405 respectively. From a contribution rate perspective, it is useful to note that the following three main results can be derived: first, the contribution of the rural-urban income gap to China's overall income gap exceeded 50% in each of the four years from a low of 55.4% in 2011 to a high of 53.2% in 2018. Second, from 2011 to 2018, the contribution rate of income inequality within towns and cities increased substantially (30.1% to 39.6%), gradually becoming a major source of income inequality in China as a whole. Third, there is a progressive decline in the share of income inequality in rural areas (from 14.5% in 2011 to 7.80% in 2018). These results are consistent with Zhang Tao (2016) on the average log deviation of China's income distribution from 1985 to 2012 as well as the decomposition of the Theil index. This allows us to illustrate the effectiveness of the algorithm through practical applications. On the one hand, the urban-rural income gap is still the largest source of income inequality in China at the present time. On the other hand, the proportional contribution of the urban income gap to

China's overall income inequality has increased substantially, making it essential to pay heed to the effects of the urban income gap.

Table 5 National Gini coefficient decomposition matrix in 2011-2013 and 2018

Matrix	$G_s = 0.412$ in 2011		$G_s = 0.404$ in 2012		$G_s = 0.399$ in 2013		$G_s = 0.435$ in 2018	
	Urban	Rural	Urban	Rural	Urban	Rural	Urban	Rural
Urban	0.354	0.914	0.345	0.895	0.338	0.88	0.359	0.96
	(30.10 %)	(55.40 %)	(31.10 %)	(55.20 %)	(32.00 %)	(54.90 %)	(39.00 %)	(53.20 %)
Rural		0.388		0.382		0.38		0.405
		(14.50 %)		(13.70 %)		(13.10 %)		(7.80 %)

Fig. 6 Structural contribution of National Gini coefficient in 2011-2013 and 2018.

5 Further analysis: sources of income inequality

5.1 Itemized income: Income structure

The income decomposition method we used simply explores the sources of income disparity in China (Fei *et al.*, 1978). Table 6 shows that wage income, transfer income, is a major source of income inequality in China. Wage income contributes more than 60% to the overall income gap in China, up from 63.19% in 2013, and up to 60.52% in 2018. Over five years, the contribution of net business income to the overall income gap

has become more and more important, rising from a low of 5.98% in 2013 to a high of 13.97% in 2018. It is important to note that there are important differences in the income structure between urban and rural areas: wage income and transfer income are two important sources of intra-urban income inequality, although wage income and business income are the two major sources of intra-rural income inequality. Net business income played an important role in 2018, increasing to 40.84% compared to 18.41% in 2013. The contribution of rural transfer income to the intra-rural income gap has also become increasingly apparent, rising from 3.51% in 2013 to 15.8% among rural transfer income. This phenomenon reflects the change of income structure caused by the change of employment structure of rural households, which affects the income distribution of rural households.

Table 6 Contribution of itemized income to the income gap in China (%)

Source of income	The national		Urban		Rural	
	2013	2018	2013	2018	2013	2018
Wage Income	63.19	60.52	59.3	60.13	32.39	40.33
Net business income	5.98	13.97	6.34	13.94	18.41	40.84
Net property income	11.43	11.62	10.53	12.22	6.45	3.89
Net transfer income	17.04	18.59	19.29	18.71	3.51	15.8

5.2 Influencing factors: education and industry

Even households with similar objective conditions can have large inequality in income, and moreover, unlike ¡°income inequality” due to factors such as increasing returns to

education. Income inequality is apparent in some industries (such as monopolies) where incomes are too high. The results indicate that income inequality is caused not only by household-level factors, but also by institutional and structural factors which cannot be overlooked or even more significant. There is both a characteristic difference between urban and rural household income and a remarkable characteristic difference in return. We use the decomposition method (Firpo *et al.*, 2018), which explores the influences of educational inequality, industry and provincial differences on income inequality between urban and rural areas. In addition, since factors such as age structure also play a significant role in income inequality (Zhang, 2019), we use age, gender, health, and marital status as control variables in Tables 7 and 8.

Based on data from both 2013 and 2018, Oaxaca-Blinder Re-centered Influence function decomposition results are shown in Tables 7 and 8. The composition effect was -141.4% and the wage-structure effect was 241.4% in 2013, respectively, while composition effect was -95.9% and wage-structure effect was 195.9% in 2018, indicating that over a five-year period the wage-structure effect is the primary source of the urban-rural income gap.

Educational differences had a strong positive contribution to the 2013 income gap (99.7%), indicating that household educational attainment widened the urban-rural income gap in 2013. In 2018, however, there was a negative contribution (-16.1%), and household educational attainment was found to significantly reduce the urban-rural income gap. The larger change in the time dimension is driven by the large wage structure effect of educational inequality on income inequality in 2013 and the larger inequality in

educational resources between urban and rural areas: by 2018 the composite effect of educational disparities on earnings disparities dominated, with rural and urban families benefiting from relatively equal access to education and the sharing of resources for education.

Industry differences were the greatest contributor to the income gap in 2013 and 2018, accounting for 257.2% and 121.7% of the total influence, respectively. Industrial disparity between urban and rural areas substantially widens income inequality. Of these, the industry difference wage structure effect is stronger than the composition effect, because the development of primary, secondary and tertiary industries can induce a significant difference between urban and rural areas, the urban-rural difference in tertiary industry has greatly widened the income gap. Thus, differences in education and industry are two important factors affecting the inequality of urban-rural incomes. The skewed nature of educational resources in China has been mitigated in the last five years, and families in urban and rural areas have enjoyed fairer access to educational resources. Disparities between industries, especially in the tertiary sector, continue to widen the income gap significantly. Most rural households are engaged in agriculture, and average annual household income has improved over the last five years due to more backward production methods being eliminated, but the gap is substantial relative to urban households. There has always been a large gap between urban families in the financial sector, real estate sector, and other tertiary occupations.

Table 7 RIF decomposition-I

2013	Total variance Value	Contribution rate	Total composition effect Value	Contribution rate	Total wage structure Value	Contribution rate
Primary	0.009	17.70%	-0.007	9.60%	0.016	13.00%
High school	0.021	40.60%	-0.002	2.60%	0.022	18.40%
Undergraduate	0.021	42.40%	-0.047	65.80%	0.068	56.10%
Graduate students	-0.001	-1.00%	-0.008	11.20%	0.008	6.10%
Total: education		99.70%		89.30%		93.60%
Secondary industry	0.03	58.20%	0.004	-6.20%	0.025	20.50%
Tertiary industry	0.101	199.10%	-0.007	10.40%	0.108	88.50%
Total: industry		257.20%		4.20%		109.00%
Midlands	-0.021	-42.00%	-0.002	2.80%	-0.019	-15.80%
Western	0	0.33%	-0.001	1.30%	0.001	0.90%
Total: region		-41.70%		4.20%		-14.80%
Total:	-0.109	-215.30%	-0.002	2.30%	-0.107	-87.80%

controls						
Total	0.051	100%	-0.071	100%	0.122	100%
	(100%)		(141.4%)		(241.40%)	

Notes: The education dummy variables are the reference group based on non-school educational attainment, primary industry is the reference group for industrial dummy variables and the eastern region is the reference group for provincial dummy variables.

Table 8 RIF decomposition-II

2018	Total variance Value	Contribution rate	Total composition effect Value	Contribution rate	Total wage structure Value	Contribution rate
Primary	-0.021	-47.20%	-0.012	28.90%	-0.009	-10.00%
High school	0.007	15.30%	0.004	-8.40%	0.003	3.70%
Undergraduate	0.009	19.80%	-0.023	55.60%	0.032	37.40%
Graduate students	-0.002	-4.00%	0.004	-9.90%	-0.006	-6.90%
Total: education		-16.10%		66.10%		24.10%
Secondary	0	-0.40%	-0.005	11.70%	0.005	5.50%

industry						
Tertiary industry	0.054	122.10%	0	0.10%	0.054	62.40%
Total: industry		121.70%		11.80%		67.90%
Midlands	-0.005	-12.30%	-0.001	2.40%	-0.004	-5.10%
Western	0.011	25.60%	0	-0.30%	0.011	12.90%
Total: region		13.40%		2.10%		7.90%
Total: controls	-0.008	-19.10%	-0.008	2.00%	0	0.10%
Total	0.044	100%	-0.042	100%	0.086	100%
	(100%)		(95.9%)		(195.90%)	

6 Conclusions and Insights

Measuring the Gini coefficient is complicated by China's large population and uneven economic development at the regional and urban and rural levels. It is necessary to find a method of measuring the Gini coefficient that is suitable for China's income distribution. The optimal solution has been proposed using a multi-group Gini coefficient method based on the income distribution function, which has the advantages of making full use of the microdata information, being consistent with the actual situation in China, and being accurate. Although this method entails a certain complexity, it addresses the main shortcomings inherent to the calculation of the Gini coefficient in the literature. The

conclusions are drawn as follows:

(1) By introducing PSO algorithm, the generalized logistic distribution function is modified to improve the accuracy of fitting urban and rural income distribution, which is important when trying to reduce the bias of Gini coefficient data. In the present work, different exponent values are compared with those chosen by PSO algorithms. The accuracy of PSO is better than that of the other values, and the Gini coefficient is different before and after optimization. The empirical results of error analysis confirm the advantages of the PSO algorithm;

(2) Based on the characteristics of China's dual economic structure and the imbalance in the level of development across provinces, the multi-group Gini coefficient method is proposed based on CHIP survey data. Be it the manner of urban and rural decomposition, the adjustment of sample information weights, or the treatment of mapping data, it accords with China's actual situation, effectively measuring overall income inequality, urban or rural income inequality, urban-rural income inequality and their proportional contributions, thus solving the problem of the deviation of Gini coefficient;

(3) China's Gini coefficient was 0.412, 0.404, 0.399, and 0.435 between 2011 and 2013 and 2018, showing an overall upward trend in the latter year: on the one hand, the urban-rural gap contributes more than 50% of the national income gap, and the urban-rural income gap remains China's main source of income inequality. Therefore, due attention should be paid to the phenomenon of widening urban-rural income inequality, and a policy of "raising incomes of low-income groups" should be formulated to ensure

productive and living conditions in the rural areas and to help rural families achieve a common prosperity. On the other hand, the proportional contribution of the urban income gap to China's overall income gap has increased substantially, so the effect of urban income inequality warrants close attention. Thus, we should follow the "expanding middle-income groups" direction, taking full advantage of the social and economic benefits that accrue from urbanization, constantly enlarging the size of middle-income groups, and eventually forming an "olive society" with middle income as its primary organ;

(4) In terms of itemized income, on the one hand, wage and transfer income are the main sources of itemized income for the households in China. Although there are significant differences in the structure of income between urban and rural areas. On the other hand, from the point of view of influencing factors of the rural-urban income inequality. The wage-structure effect is the main source of the urban-rural income inequality, *i.e.* the policy-led tendency to greater urbanization. The education and industry of urban and rural households as well as the difference in their returns are the main causes of income inequality between the households in urban and rural areas. This has alleviated the problem of the biased allocation of educational resources in China, and allows for urban and rural families to have relatively fair access to equal resources for education. Differences between industries, especially in the tertiary sector, continue to widen the income gap substantially. Most rural households are involved in agriculture, and there has been a large gap developed with the secondary and tertiary industries.

References

[1] Ai, X. Q. (2015). A new method of calculation and decomposition of overall Gini coefficient. Statistical Research, 32(09), 91-96. (in Chinese)

[2] Atkinson, A. B. (1970). On the measurement of inequality. Journal of Economic Theory, 2(3), 244-263.

[3] Bhattacharya, D. (2007). Inference on inequality from household surveydata. Journal of Econometrics, 137(2), 674-707.

[4] Chen Z, & Wan G. H. (2011). Inter-industry Inequality: An Important Source of the Urban Income Gap-Regression-based Decomposition. Social Sciences in China, 2011, 32 (02): 159-177.

[5] Cheng, B. W. (2005). Lorenz curve and Gini coefficient under condition of lognormal distribution. The Journal of Quantitative & Technical Economics, (02), 127-135. (in Chinese)

[6] Cheng, Y. H. (2006). Calculation and decomposition of the overall Gini coefficient in dual economics. Economic Research Journal, (01), 109-120. (in Chinese)

[7] Cheng, Y. H. (2007). China's overall Gini coefficient since reform and its decomposition by rural and urban areas since reform and opening up. Social Sciences in China, (04), 45-60+205. (in Chinese)

[8] Davidson, R. (2009). Reliable inference for the Gini index. Journal of Econometrics, 150(1), 30-40.

[9] Fei, J. C. H, Ranis, G., & Kuo, S. W. Y. (1978). Growth and the family distribution of income by factor components. The Quarterly Journal of Economics, 92(1): 17-53.

[10] Fontanari, A., Taleb, N. N., & Cirillo, P. (2018). Gini estimation under infinite variance. Physica A: Statistical Mechanics and its Applications, 502, 256-269.

[11] Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. Econometrica, 2009, 77(3): 953-973.

[12] Firpo, S. P., Fortin, N. M., & Lemieux, T. (2018). Decomposing wage distributions using recentered influence function regressions. Econometrics, 6(2): 28.

[13] Gan, L. (2012). China Household Finance Survey Report: 2012. Southwestern University of Finance and Economics Press. (in Chinese).

[14] Gittleman, M., Wolff, E. N. (1993). International comparisons of inter industry wage differentials. Review of Income and Wealth, 39(3): 295-312.

[15] Gini, C. (1912). Variabilit¤ e mutabilit¤. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E.

[16] Grazia Pittau, M., & Zelli, R. (2004). Testing for changing shapes of income distribution: Italian evidence in the 1990s from kernel density estimates. Empirical Economics, 29(2), 415-430.

[17] Han, X., & Cheng, Y. (2019). Does the" missing" high-income matter?-Income distribution and inequality revisited with truncated distribution? China Economic Review, 57, 101337.

[18] Hu, Z. G. (2004). A Study of the Best Theoretical Value of Gini Coefficient and Its Concise Calculation Formula. Economic Research Journal, (09), 60-69. (in Chinese)

[19] Knight, J., & Gunatilaka, R. (2022). Income inequality and happiness: Which

inequalities matter in China? China Economic Review, 72, 101765.

[20] Li, C., Yu, Y., & Li, Q. (2021). Top-income data and income inequality correction in China. Economic Modelling, 97, 210-219.

[21] Li, Q. H., Li, S., & Wan, H. (2020). Top incomes in China: Data collection and the impact on income inequality. China Economic Review, 62: 101495.

[22] Li, S., & Wan, H. Y. (2013). Improve the reliability of Gini coefficient estimation in China. Economic Perspectives, (02), 43-49. (in Chinese)

[23] Li, S. (2002). Further explanation on the Estimation and decomposition of Gini coefficient - a reply to professor Chen Zongsheng's comment. Economic Research Journal, (05), 84-87. (in Chinese)

[24] Lin, P., Guo J. Q., & Fei S. L. (2013). A new improvement of China's urban and rural Comprehensive Gini coefficient based on indirect Lorenz curve summation. The Journal of Quantitative & Technical Economics, 30(11), 108-124. (in Chinese)

[25] Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. Publications of the American statistical association, 9(70), 209-219.

[26] Luo, C. L., Li, S, & Yue, X. M. (2021). An analysis of changes in the extent of income disparity in China (2013-2018). Social Sciences in China, (01), 33-54+204-205. (in Chinese)

[27] Molero-Simarro, R. (2017). Inequality in China revisited. The effect of functional distribution of income on urban top incomes, the urban-rural gap and the Gini index, 1978¨C2015. China Economic Review, 42: 101-117.

[28] Ogwang, T. (2000). A convenient method of computing the Gini index and its

standard error. Oxford Bulletin of Economics and Statistics, 62(1), 123-123.

[29] Rothe, C. (2010). Nonparametric estimation of distributional policy effects. Journal of Econometrics, 155(1): 56-70.

[30] Ryu, H. K., Slottje, D. J., & Kwon, H. Y. (2019). A New Logit-Based Gini coefficient. Entropy, 2019, 21(5): 488.

[31] Shi, Y., & Eberhart, R. (1998). A modified particle swarm optimizer//1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360). IEEE, 69-73.

[32] Solon, G., & Haider, S. J, (2015). Wooldridge J M. What are we weighting for?[J]. Journal of Human resources, 50(2): 301-316.

[33] Sundrum, R. M. (2003). Income distribution in less developed countries. Routledge.

[34] Theil, H., & Uribe, P. (1967). The information approach to the aggregation of input-output tables. The Review of Economics and Statistics, 451-462.

[35] Wan, G., Zhou, Z. (2005). Income inequality in rural China: Regression based decomposition using household data. Review of development economics, 9(1): 107-120.

[36] Wang, Z. X., & Smyth, R. (2015). A hybrid method for creating Lorenz curves. Economics Letters, 133: 59-63.

[37] Wang, Z. X., Zhang, H. L., & Zheng, H. H. (2019). Estimation of Lorenz curves based on dummy variable regression. Economics Letters, 177, 69-75.

[38] Wang, H. G., & Zhou, H. G. (2006). Is the inequality of income distribution

between Urban and rural China underestimated? -A test based on the Pareto distribution. Statistical Research, (04): 8-15. (in Chinese)

[39] Williamson, J. G. (1965). Regional inequality and the process of national development: a description of the patterns. Economic development and cultural change, 13(4, Part 2), 1-84.

[40] Xu, K. (2003). How has the literature on Gini coefficients expanded over the past 80 years? China Economic Quarterly, (03), 757-778. (in Chinese)

[41] Yang, J., & Gao, M. (2018). The impact of education expansion on wage inequality. Applied Economics, 50(12), 1309-1323.

[42] Yue, X. M., Li, S. (2013a). Lack of convincing response-requestioning the Gini coefficient published by Southwest University of Finance and Economics household survey project. (in Chinese)

[43] Yue, X. M., Li S. (2013b). Whose Gini coefficient should we trust more? (in Chinese)

[44] Zhang, H. F., Zhang H. L., & Zhang, J. S. (2015). Demographic age structure and economic development: Evidence from Chinese provinces. Journal of Comparative Economics, 43(1), 170-185.

[45] Zhang, Q., & Churchill, S. A. (2020). Income inequality and subjective wellbeing: Panel data evidence from China. China economic review, 60: 101392.

[46] Zhang, T. (2016). Fluctuation and causal analysis of China's income gap: 1985 2012. The Journal of Quantitative & Technical Economics, 33(12), 3-22. (in Chinese)

Highlights

- A multi-group Gini coefficient optimization method based on PSO algorithm is proposed.
- We use the general Logistic distribution function to measure the Gini coefficients.
- The new method makes full use of the valid information of microdata.
- Based on CHIP, we measure urban, rural Gini coefficients and contribution rates.
- Education, industry and their rates of return are the main causes of urban-rural income gap.

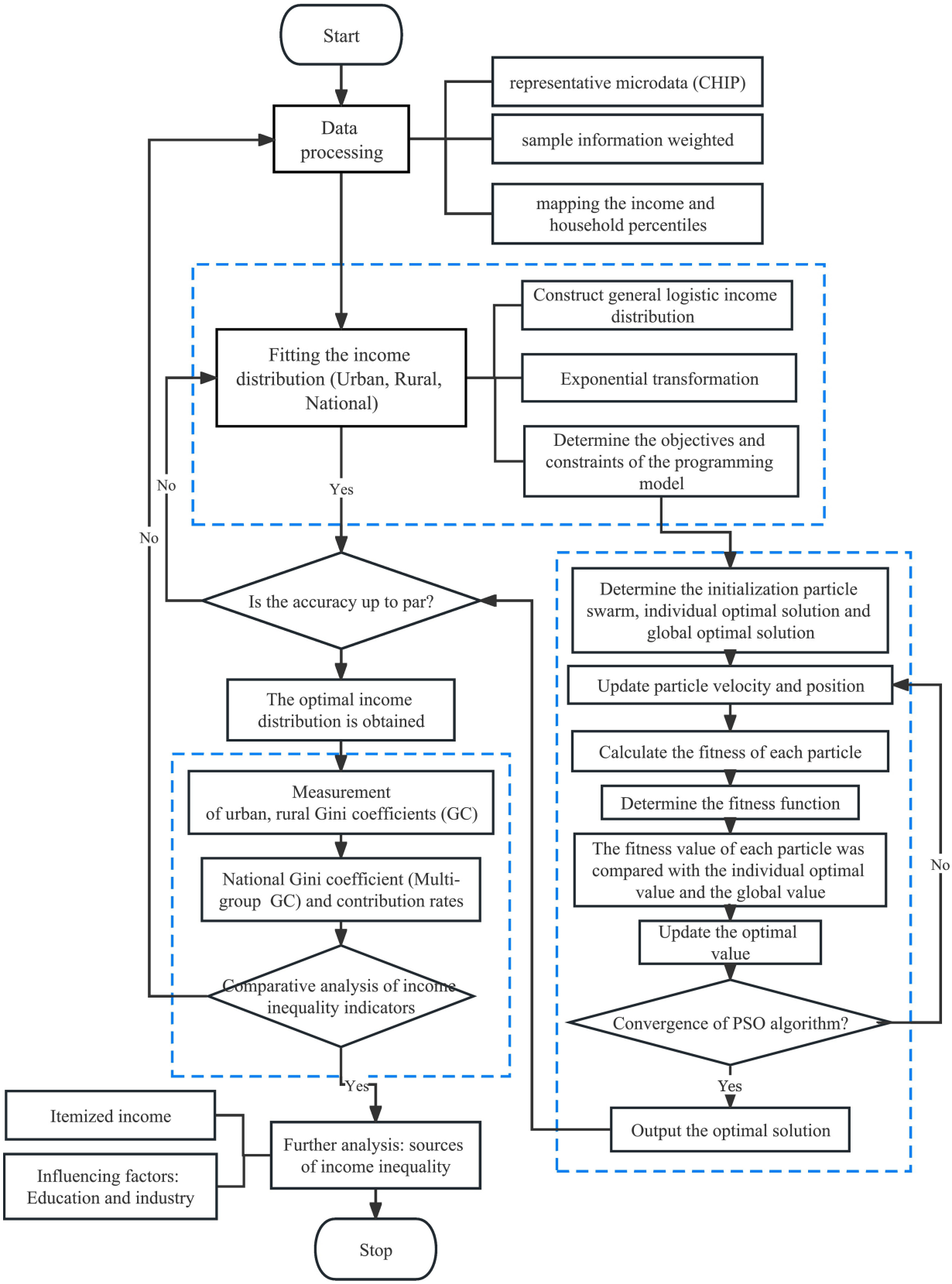
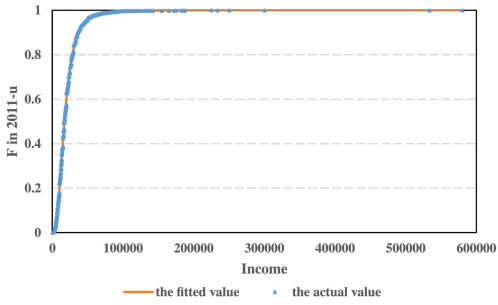
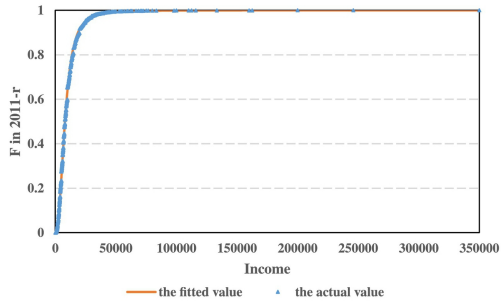

Start
Data processing
representative microdata (CHIP)
sample information weighted
mapping the income and household percentiles
Fitting the income distribution (Urban, Rural, National)
Construct general logistic income distribution
Exponential transformation
Determine the objectives and constraints of the programming model
No
Yes
No
Is the accuracy up to par?
The optimal income distribution is obtained
Measurement of urban, rural Gini coefficients (GC)
National Gini coefficient (Multi-group GC) and contribution rates
Comparative analysis of income inequality indicators
Yes
Itemized income
Influencing factors: Education and industry
Further analysis: sources of income inequality
Stop
Determine the initialization particle swarm, individual optimal solution and global optimal solution
Update particle velocity and position
Calculate the fitness of each particle
Determine the fitness function
The fitness value of each particle was compared with the individual optimal value and the global value
Update the optimal value
Convergence of PSO algorithm?
No
Yes
Output the optimal solution

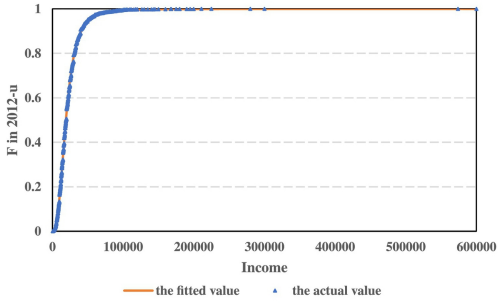
Figure 1



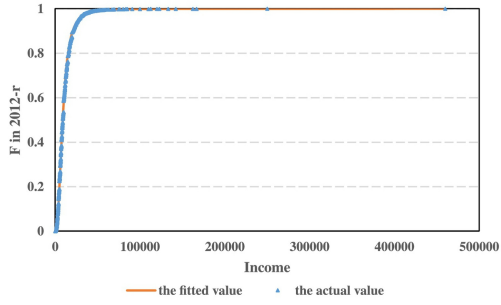
(a) 2011-u



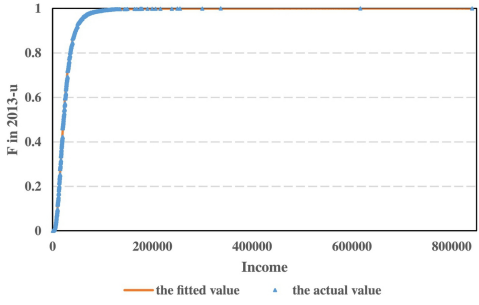
(b) 2011-r



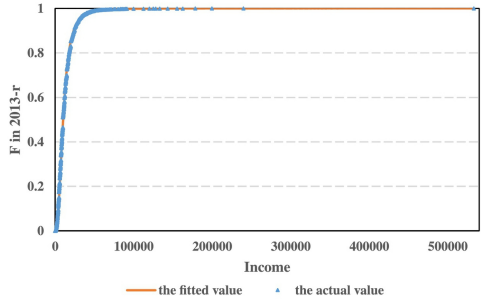
(c) 2012-u



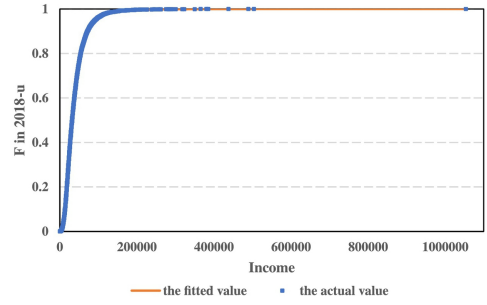
(d) 2012-r



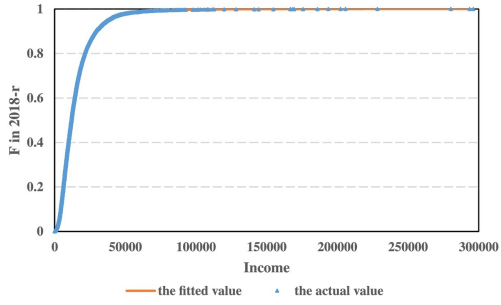
(e) 2013-u



(f) 2013-r



(g) 2018-u



(h) 2018-r

Figure 2

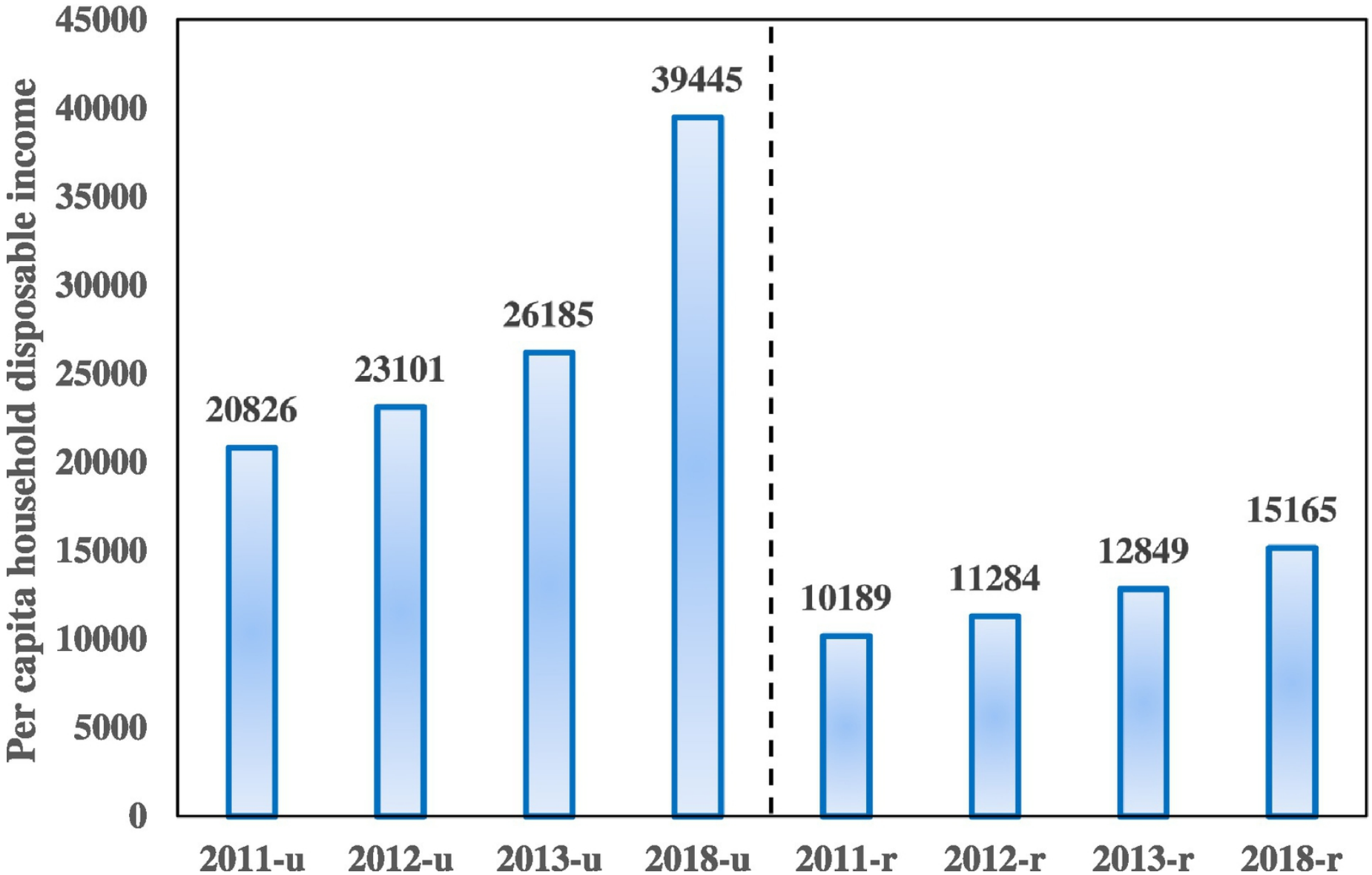

Per capita household disposable income
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
20826
23101
26185
39445
10189
11284
12849
15165
2011-u
2012-u
2013-u
2018-u
2011-r
2012-r
2013-r
2018-r

Figure 3

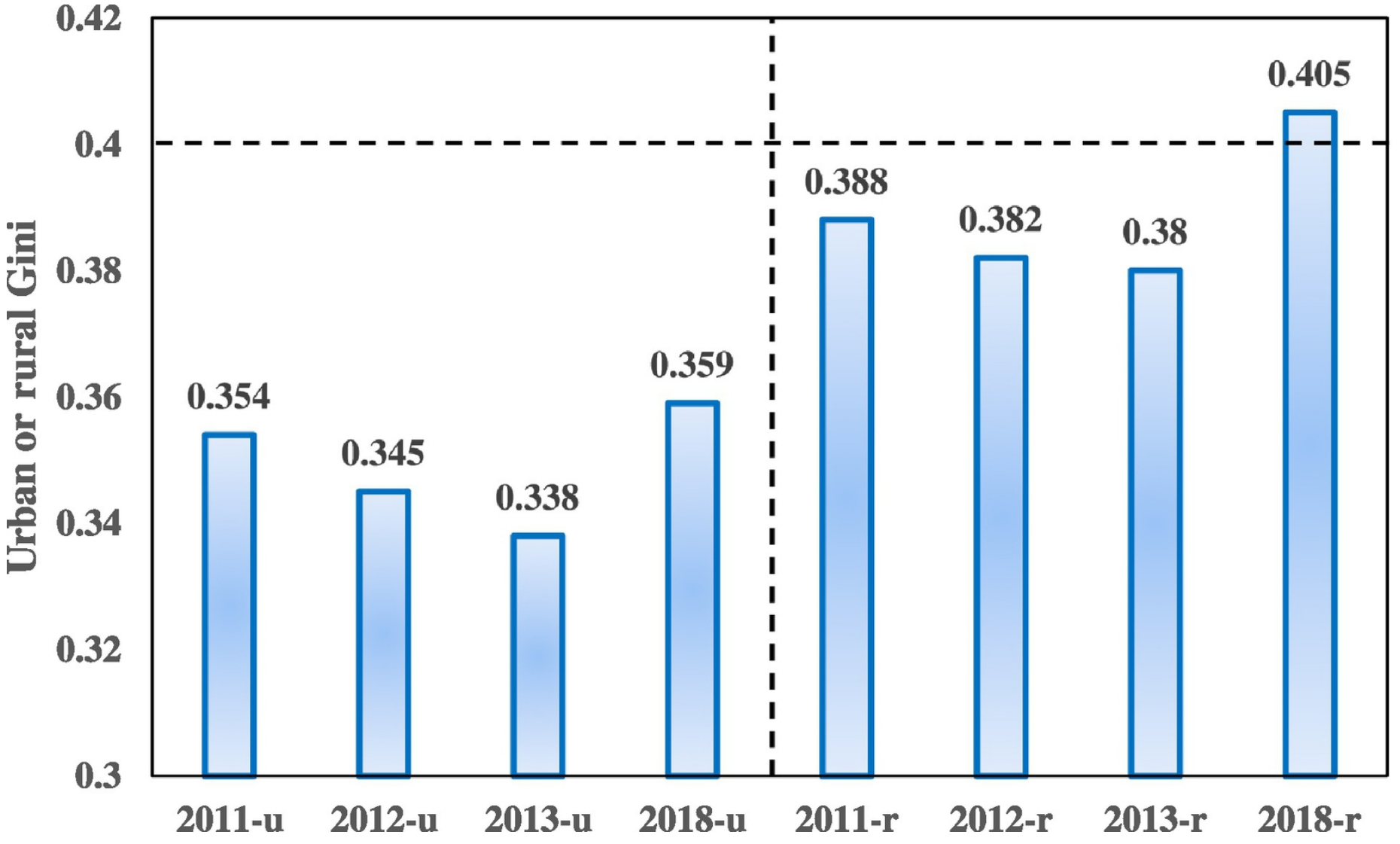

0.42
0.4
0.38
0.36
0.34
0.32
0.3
Urban or rural Gini
0.354
0.345
0.338
0.359
0.388
0.382
0.38
0.405
2011-u
2012-u
2013-u
2018-u
2011-r
2012-r
2013-r
2018-r

Figure 4

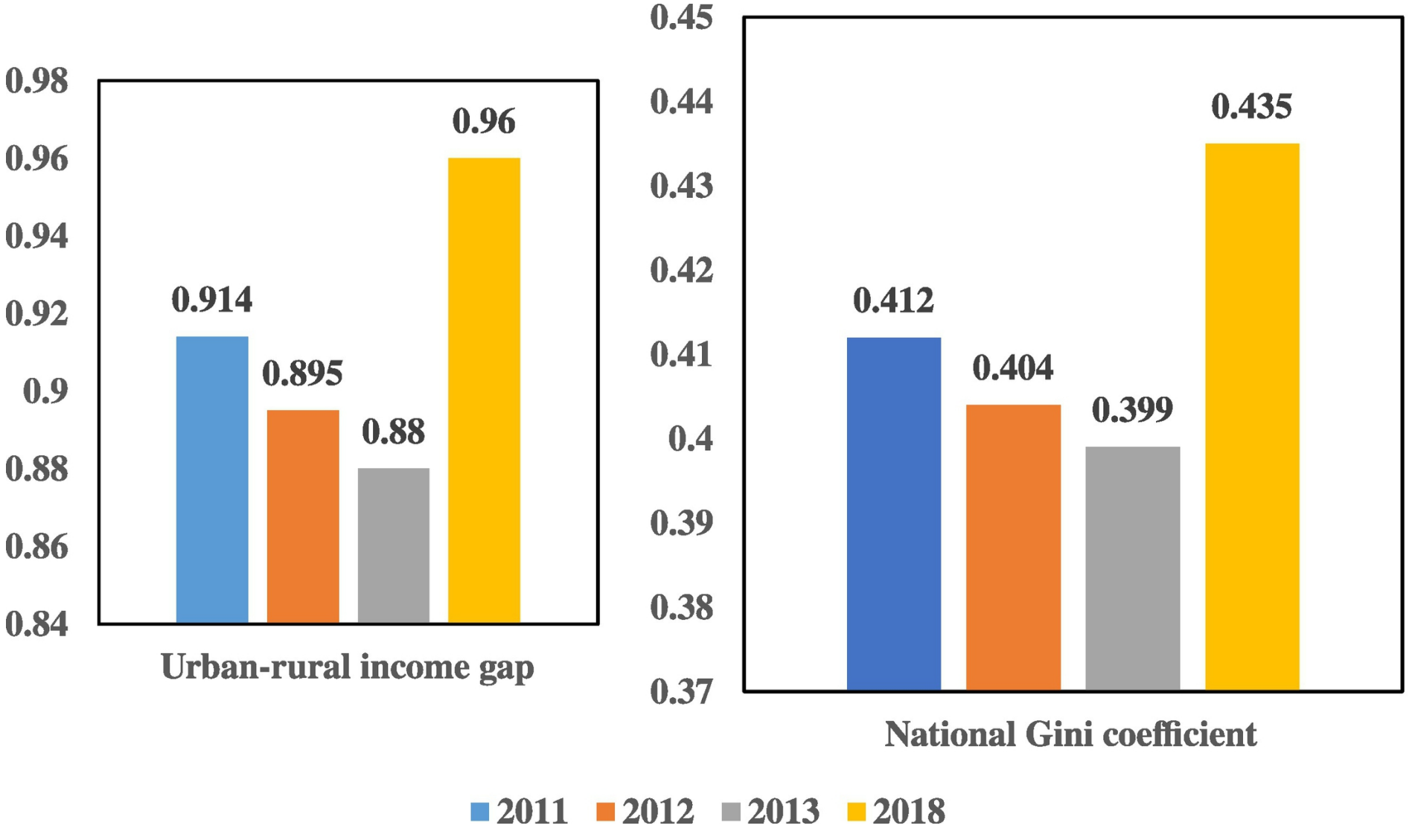

0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.914
0.895
0.88
0.96
Urban-rural income gap
0.45
0.44
0.43
0.42
0.41
0.4
0.39
0.38
0.37
0.412
0.404
0.399
0.435
National Gini coefficient
2011
2012
2013
2018

Figure 5

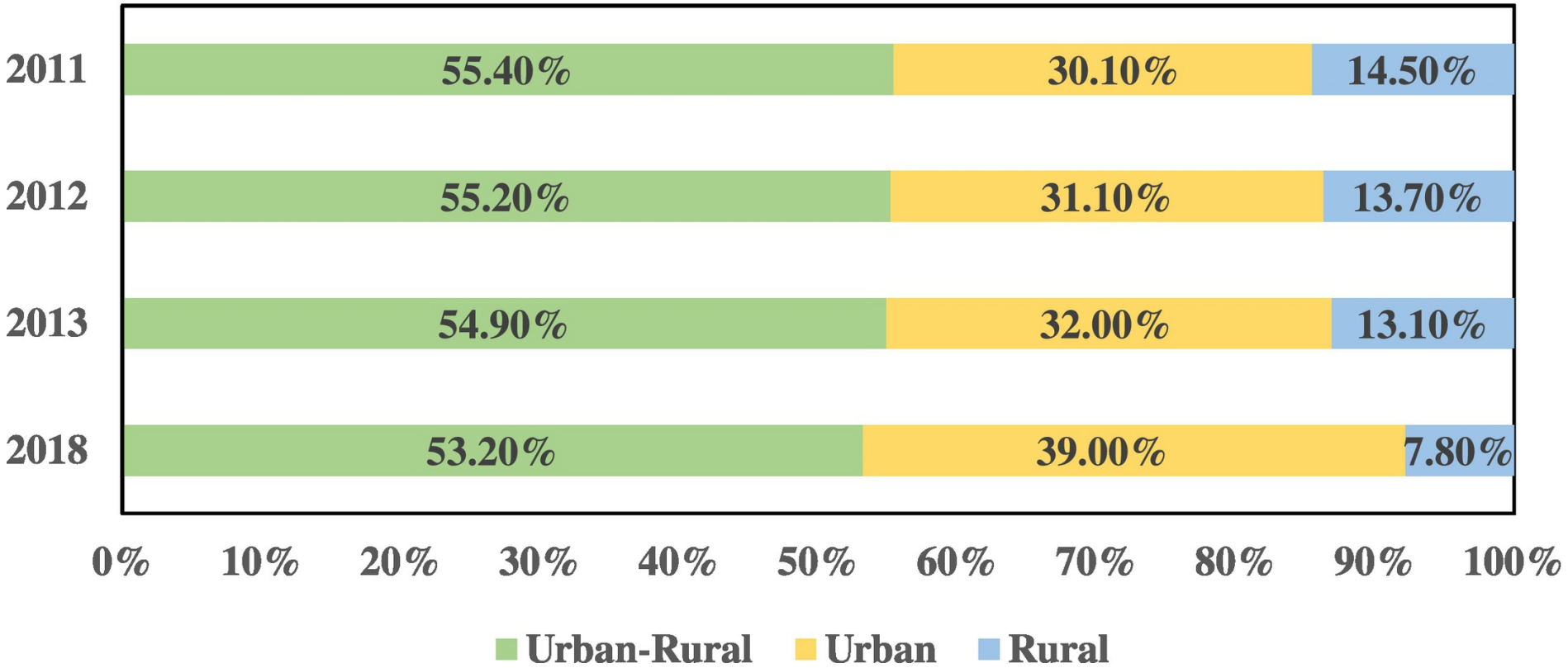

2011
2012
2013
2018
55.40%
30.10%
14.50%
55.20%
31.10%
13.70%
54.90%
32.00%
13.10%
53.20%
39.00%
7.80%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Urban-Rural
Urban
Rural

Figure 6