

Li, Qinghai; Li, Shi; Wan, Haiyuan

Working Paper

Top incomes in China: Data collection and the impact on income inequality

WIDER Working Paper, No. 2018/183

Provided in Cooperation with:

United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Li, Qinghai; Li, Shi; Wan, Haiyuan (2018) : Top incomes in China: Data collection and the impact on income inequality, WIDER Working Paper, No. 2018/183, ISBN 978-92-9256-625-8, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki,
<https://doi.org/10.35188/UNU-WIDER/2018/625-8>

This Version is available at:

<https://hdl.handle.net/10419/211221>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



WIDER Working Paper 2018/183

Top incomes in China

Data collection and the impact on income inequality

Qinghai Li,¹ Shi Li,² and Haiyuan Wan³

December 2018

Abstract: With the data on the top incomes collected from different sources, we combine the samples of the top incomes with a household survey to investigate changes in the income distribution with and without the top incomes. The Gini coefficient of income inequality using household survey data is 0.464 for 2016, and it jumps to 0.646 after including the samples of the top incomes, which demonstrates the great importance of the top incomes in estimating income inequality.

Keywords: top incomes, Pareto distribution, income inequality

JEL classification: C46; D31; D63; O15.

Acknowledgements: In the process of data collection, we have been helped by Gao Minghua, Fang Fang, Lv peng, Yu yangcheng, Jiangnan, Qian weijuan, Liuchencheng, Tian miaoqing, Yanxu, Li yueyue, Liu xiaoting, and others. Here we express our sincere thanks.

¹ Nanjing University of Finance and Economics, Nanjing, China; ² Beijing Normal University, Beijing, China, corresponding author: lishi@bnu.edu.cn; ³ Beijing Normal University, Beijing, China.

This study has been prepared within the UNU-WIDER project on '[Inequality in the Giants](#)'.

Copyright © UNU-WIDER 2018

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-625-8

Typescript prepared by Ans Vehmaanperä.

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

The views expressed in this paper are those of the author(s), and do not necessarily reflect the views of the Institute or the United Nations University, nor the programme/project donors.

1 Introduction

Estimating the exact income inequality for one country would face several challenges, among which the loss of data on top-income individuals and income under-reporting in the survey data are the most challenging for the researchers. For the case in China, much evidence supports the view that the group with top incomes is increasingly important due to the rapid economic growth in recent decades. In the past five years there has been considerable controversy regarding the number of billionaires, their income level, and their distribution in China¹. However, the traditional method of using household surveys cannot address these issues. Therefore, we have to seek other ways to collect information about income or wealth of the top incomes. Recent research such as that conducted by Thomas Piketty (2013) seeks out the possibilities in which administrative heritage tax data can be used to complement data on the top incomes. However, it is impossible to do this in China because we do not have detailed information on tax categories such as the heritage tax or estate tax.

It is well known that a method using a household survey to estimate income inequality could suffer from underestimation. Nevertheless, in the absence of better data sources, we must rely on ourselves to collect all kinds of data on top incomes from different sources, and then, we match several data sources together to try to capture as much data on the top incomes as possible. Using this way to capture the top incomes is an imperfect but practical solution², and is the first attempt that we are aware of to correct for the top incomes. In this paper, we try to collect data on the top incomes in different industries or fields one by one using all possible sources and then try to mix them together into an organized data set called “Top Incomes in China (TIC)”; then we estimate the income distribution including the top incomes using the TIC data.

2 Literature Dealing with Top Incomes

2.1 Why correcting top incomes?

Correcting for the top incomes in income distribution research has been a long story for economists, and the literature can be classified into two stages. In the beginning, the research employs methodology, such as the Pareto, to correct, while in recent years, many attempts have been made to determine the quality of the top income data. As stated by Davies et al. (2010), the underestimation of the top incomes has a larger impact on the overall income distribution, as the top incomes may not be included in household surveys, while the under-reporting rate is even higher for this group than for other groups. Piketty (2014) demonstrates that income inequality would be higher than 20 percent if correcting the top incomes in the US in 2011, while this figure is even higher according to Xie and Jin (2014), as the number reaches as high as 30% for the Chinese case in 2012. Meanwhile, the mean income for the overall sample would also increase by 25 percent when including the top incomes for China in 2012, and for the same year in China, this figure reaches 32% according to Gan et al. (2012). Regardless of whether we use a household survey together with the Pareto function, as in Xie and Jin (2014), or only use the household survey

¹ According to the estimation of income inequality published by the National Bureau of Statistical (2013), the national Gini coefficient is 0.481 for 2012, which is also confirmed by Li et.al (2015). However, Xie and Zhou (2014) provide evidence that the Gini should range from 0.52 to 0.55, while a significant higher Gini that reaches as high as 0.61 is also proposed by Gan et.al (2014) for the same year.

² Moreover, this issue could become less important when we consider the dynamics of income distribution.

while highlighting the top incomes during the sampling, as in Gan et al. (2012), both obtain a unanimous result about the importance of the top incomes.

Table 2.1 Empirical evidence for correcting top incomes

	Research context	Data source	Year	Correcting way	Gini before correcting	Gini after correcting	summary
Li and Luo (2011)	income inequality	National CHIP	2007	Forbs rich list to supplement top samples	0.481	0.530	Adding top incomes would apparently raise up the income inequality
Gan et al (2012)	income growth	Urban CHFS	2009	More sampling for the high-income groups	0.478	0.561	Mean value of income has three times increase if use the new way to correct top samples
Xie and Jin (2014)	wealth inequality	National CHIP	2012	Hurun rich list to supplement top samples	0.641	0.734	Wealth and income's growth and distribution are distorted if excluding the top incomes
Li and Wan(2015)	wealth growth	National CHIP/CHIP	2012	Using income tax to supplement top samples	0.538	0.739	Huge difference before and after correction, importance of correcting is rising
Piketty(2017)	income share for the top 10%	Aggregate data	2014	Pareto estimation to supplement top income samples	0.524	0.685	National income account is serious underestimated without top incomes
Knight, Li and Wan (2018)	Wealth distribution	CHIP	2012	Expanding top income samples based on survey	0.617	0.715	Top incomes' activity pattern and returns rate of income is different with other groups
This paper	income inequality	TIC/CHIP	2012	Collect top income data from the public	0.464	0.646	Huge income distribution difference for the top incomes compared to household survey

Source: authors' collection.

Apart from the two attempts above, Li and Wan (2015) also compare the trend of importance for top incomes while estimating income inequality. Using Chinese Household Income Project (hereinafter referred to as CHIP) 2002 and CHIP 2013, it finds that correcting top incomes would push up the Gini coefficient of income inequality to 20% in 2002 and 28% in 2012, while the mean value of income would also increase by 21% in 2002 and 34% in 2012 respectively, so it concludes that the issue of under-representation or under-reporting for the top incomes are more and more important. Knight et al. (2018) simulate the results before and after correcting top incomes with an expansion in the 1% or 5% highest deciles for wealth inequality using CHIP 2013 for the case in China and find that there is a very large difference before and after correcting the top incomes for current China. It found that the basic Gini coefficient without any correction was 0.497, the consideration of the under-representative issue itself would raise the Gini coefficient to 0.613, while the second step correcting both of under-representation and under-reporting would push up the Gini coefficient to 0.718.

Piketty (2018) also joined the research to correct China's top incomes using data from national accounts. Though it is impossible to provide the Gini coefficient, they also find that there is a serious underestimation of income inequality if using deciles or percentiles.

2.2 Dealing with top incomes

For all the literature on income distribution, data quality is most important and would cause many arguments among researchers worldwide. Because of the use of different data sources and different estimation procedures, the top income estimation results in particular are also very different, the results are not easily comparable across years and countries. In general, as the number of top incomes is very limited, it is difficult to obtain enough effective samples using the methodology of simple random sampling with a household survey. Meanwhile, in practice, top incomes are not willing to cooperate with household surveys, so the number of top incomes tends to be

inadequately represented. In addition, top incomes tend to hide their actual income so that the income information for high-income groups is more difficult to obtain accurately in surveys.

In general, it is very difficult to capture the issues of top incomes by using the random sampling method; therefore, the research of top income distribution using household surveys is usually biased because of right truncation. Therefore, global scholars in the field of income distribution are discussing how to supplement the high-income sample.

2.2.1 Using Pareto distribution function

Because of the lack of high-quality income data, household survey data are often preferred by researchers and are indeed most widely used in practice (Davies et al., 2010). However, simply using household survey data to study the income distribution has many shortcomings. They believe that in the context of the increasing concentration of income accumulation, the problem of under-representation of the top income groups is more obvious than before, therefore simply using household survey data to study income distribution is very biased. As a result, many researchers, through technical means, modify traditional household surveys. At the end of the 19th century, an Italian economist named Pareto found that the distribution of top incomes obeys a law such that 20% of the people own 80% of the overall income in the society (80/20 rule), so afterwards we called it a Pareto distribution for top incomes. In practice, there are two ways to estimate the top income distribution while using the Pareto function.

The first way is to estimate the actual income distribution while using top income data from the public. For example, Wang and Zhou (2006) and Xie and Jin (2014) use very limited top income data from public rich lists published by Forbes and Hurun and employ the Pareto function to estimate the parameter for the top income distribution. Afterwards, these studies use the Pareto parameter to derive the remaining part of the top incomes and then merge the derived top incomes with the traditional household survey income data to obtain the final income distribution, including the overall top incomes. The second way is to derive the top incomes from a household survey. Although the top incomes are missing in the survey, the highest-income individuals in the survey can also be used to estimate the sensitivity of the top incomes. In general, using Pareto distribution approximation is an efficient way to estimate income inequality, as we find that the method of using Pareto distribution to approximate the top income distribution has a methodological advantage and thus has been widely used in the field.

2.2.2 Inferring from inheritance tax data

The estimation of income distribution becomes very difficult due to omissions of the most affluent group and the concealment of real income information. In recent years, experts from various countries have begun to use inheritance tax data to infer the distribution of top incomes and have achieved great success. For example, Atkinson (2013) use the actual tax data of many western countries, especially data on the property tax and the inheritance tax, to infer top income data. Thereby, they obtain more accurate and complete top income data and achieve the aim of obtaining the income data of the social elite.

Using this method, Piketty (2014) estimated the level of income gaps in more than 20 countries that have well-established inheritance tax systems, such as the United States, the United Kingdom, and Japan. They found that the highest 10% of people held more than 60% of the total social income in the 18th century. After more than 100 years of a declining trend, it has started to rise sharply in recent years and may continue to return to the social curing state of the middle ages. Moreover, for countries in which the estate tax data are available, researchers often prefer to use estate tax data to estimate the income inequality, because this method is straightforward and more

reliable. Unfortunately, China does not have a heritage tax system, so past inheritance tax data cannot be obtained in China.

2.2.3 Correcting based on household survey data

It is generally believed that although there is a problem of insufficient representation of high-income people in household survey data, the capture of low- and middle-income groups for the survey is still very accurate. Therefore, some studies have proposed that the household survey data can be revised by indirectly adjusting high-income samples. Although high-income samples cannot be guaranteed to be randomly sampled, resulting in insufficient high-income samples, there is also the problem of random errors in the data investigation process, which often leads to a certain proportion of abnormally high-income samples. Because both problems are random, if there is a reasonable sample of abnormally high income after rigorous inspection, the simultaneous existence of underestimation and overestimation may partially offset some of the original estimation error.

In a study of the income distribution of households using CHIP2013 data, it was found that 0.3% of the samples had abnormally high values of income, and there was a significant abrupt change in the distribution curve. For these anomalous observations, Li & Wan (2015) retain abnormal high-income samples for which problems cannot be identified based on the exclusion of problematic outliers. Using the same data, Xie and Jin (2014) adopt another method, which is simply dropping all income outliers and using the Pareto function to supplement the extreme values. However, Chen et al. (2009) note that removing these samples to avoid the effect of extreme values on the income distribution involves a certain degree of subjectivity, and it is easy to underestimate the degree of income inequality. Of course, regardless of which method is adopted, the final estimation of the two is basically the same.

2.2.4 Adjusting with property income tax for the top groups

Due to the absence of the top sample in income distribution studies, there are studies that try to use property income instead of income data to indirectly supplement the top-income sample. Torche and Spilerman (2008) try to use capital income tax data collected by the taxation department to supplement the conventional sample of household income surveys. Using this method, they obtain the relationship between the missing sample and the survey sample and use this parameter relationship in the study of income distribution. Based on this, it is possible to infer the top income data and thus supplement the data in the income distribution curve. Of course, this method also has some problems. For example, previous studies basically calculate income data from wealth assets with a fixed return rate instead of using property income data to infer income data. This process is used because the income flow is implicit and dynamic. Additionally, wealth is a relatively stable stock, so wealth accounting is relatively easy for researchers.

In addition, the relationship between the high-income and low-income parameters in the income distribution may not be consistent with this same relationship in the property income distribution. Therefore, it may not be appropriate to use an analogy method to directly put income into property income research. Of course, based on conscientiously solving these problems, using the property income sample of the tax department to infer the overall income distribution is still a research direction worth trying in the future.

2.2.5 Using administrative management data

As the composition of household income is becoming increasingly complex, many countries try to establish a cross-departmental, multi-level information linkage platform to realize the full sharing of personal income among administrative departments. This practice widely introduces

administrative information data from various departments, such as the people's geographical distribution data in the public security departments, data on financial assets from banks, automobile assets data from traffic and vehicle departments, housing information from the Housing Construction Bureau, and data on enterprises and operating assets from the Commerce and Industry Bureau. Tax payment data from the local tax department can effectively provide information regarding the income of a household. The authenticity and authority of the data make it possible to estimate the real income of a family. On the one hand, these data can be added directly to estimate the household income distribution; on the other hand, these data can also be used for a comparison with self-reported income data from household surveys. Thus, the under-reporting rate of households with different income levels is estimated, and real income data for the top incomes are inferred. Overall, because the administrative data held by government departments are accurate and reliable, it will become one of the effective methods used to estimate the distribution of income in the future.

Generally, we think that using administrative data has great potential for estimating top-income inequality. The administrative databases could also be valuable sources for research. Because of these advantages, the sources of administrative data seem to be promising sources that will be increasingly used in the field.

2.2.6 Collecting behavioral data

We can also discover residents' income by assessing the close relationship between individual behavior and income; for example, the higher the income level is, the lower the proportion of basic consumption to the income, according to Engel's law. Some studies use the consumption tendencies of the highest-income groups to simply discover income levels using Engel's law (Wang Xiaolu 2010). However, there are significant problems with this approach, such as consumption and income, which may not have a simple and stable linear relationship (Luo et al. 2011).

In addition, there are behavioral economics studies that have determined that charitable donations have a relation with donors' assets, which can be used to discover an individual's income from data on charitable giving. Therefore, the donation propensity coefficient is estimated to predict the income level of the highest group. In addition to using behavior data regarding consumption and charitable donations, some practitioners actually obtain data through practice, for example, through providing legal advice to practitioners. These data are used to estimate the top incomes.

3 Data Collection for Top Incomes

3.1 Top income database

China has seen an increasing number of billionaires in the last two decades, which influences measures of income inequality. If they are included or under-representatively included in a household survey, it is not surprising that the measure of income inequality using the household survey data would be underestimated. It is very common that income inequality is underestimated due to an under-representative sample of top income people in almost all the countries, particularly in developing countries. China as one developing country is not exceptional. It is worth exploring alternative ways for collecting more accurate income information on the Chinese super-rich to make more reliable estimates of income inequality in the country. To attain this objective, we search and collect different data sources of data on the super-rich from already published information in the media and on the Internet.

Our research team constructed several databases about the Chinese super-rich, and the details are as follows. The income or wealth information of the rich come from the following databases, Hurun Rich List (2003-2017), Forbes Rich List (2003-2017), Online Celebrity (2016-2018), Payment of CEOs of Listed Companies (2016), Forbes China Celebrity List (2006-2017), Hurun Artist List (2003-2017), Chinese Private Enterprise Owner Survey (2016), Rich Writer List (2006-2016) and so on. Considering the sample size, this paper concentrates on the income of the rich in 2016, which is the latest information we can obtain. It is noted that all the databases contain income information on the rich except the Hurun Rich List, which only provides information on the wealth of the rich. We converted wealth to income using the fixed return rate of 5% proposed by Li and Luo (2011).

There is no uniform and commonly accepted criterion for identifying the high income or super-rich in China so far, while the starting point for individuals to declare personal income tax is 0.12 million according to the China's State Administration of Taxation, so we think that 0.12 million should be a reasonable threshold for the top incomes; thus, we truncate the data at 0.12 million for top incomes³. In addition, the Pareto function is used to simulate the distribution characteristics for the top incomes.

3.2 Hurun Rich list data

We summed the Hurun Rich List and the Forbes Rich List data (hereafter referred to as the Two Lists) from 2003 to 2017 in Table 3.1. The Two Lists were collected and compiled from publicly released data. In the 2003~2017 Hurun Rich List, 4,918 individuals in total had been listed in each of the years. We manually added the missing information of top-income individuals, including their privately owned companies, location of residence, year of birth, and educational attainment. Table 3.1 shows the samples of top incomes and the threshold value of the Two Lists.

Table 3.1 Samples and threshold of Hurun and Forbes Rich List (unit: yuan)

Year	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003
Threshold of Hurun (100 million)	20	20	20	20	20	18	20	15	10	7	8	8	5	12.5	9
Samples in the Hurun List	2128	2055	1875	1257	1017	1024	1004	1000	1015	1012	813	500	400	100	100
Threshold of Forbes(100 million)	67	67	54	42.8	36.6	29.6	32	28.5	20.5	12.2	15	8	5	6.5	8.3
Samples in the Forbes List	400	400	400	400	400	400	400	400	400	400	400	400	400	200	100

Source: authors' calculation based on the Hurun Rich list and Forbes Rich list.

Compared with Forbes Rich list, the sample size of Hurun Rich list is larger and its authority is more recognized by the public. Therefore, this paper takes Hurun Rich List as the benchmark data, and Forbes Rich List as a supplement to merge into Hurun List, which is called Hurun Rich List for the sake of convenience. To compile the data to be used for this paper, we took the following steps.

Step1: Dropped unreliable obs. There are some errors about the names of individuals, families and companies on the two Lists, so we made several rounds of data cleaning and corrected them

³ The mean disposable income for the highest 20% of the population in the year of 2016 is 60,000 according to the NBS, and the mean value of the highest 10% and 5% group's income is 79,000 yuan and 130,000 yuan, so we argue that the use of 120,000 should be a reasonable starting point for the top incomes, as the highest 5% average income from the household survey is already higher than 120,000 yuan.

up based on the authoritative documents like ID information⁴, and finally get 3,495 individuals with correct names on the Lists.

Step2: Modified the incorrect information. Some individuals on the Lists are from the same companies or same families. We clarified each of these cases separately in each year, by dividing equally the family wealth to each of family members. After the clarification, the individuals on the Lists decreased from 3,495 to 3,415 who are on the Lists at least in one of these years. Some individuals have been on the Lists from the first to the last year and some only in one year.

Step3: Created supplementary data for the missing data. There still existed many missing values for rich individuals on the Hurun List after the first-round cleaning. As we chose 2016 as the base year, for those individuals not listed in 2016, we re-estimated their income/wealth by an interpolation method using the average growth rate of production assets in the same industry. At the same time, we replace some obviously unrealistic data of individual wealth by 90% of the threshold or zero. For example, if the individual did not show on the two Lists in 2016, and the calculated value based on industrial average growth rate is bigger than threshold value, then we replaced their wealth value by 90% of the threshold; if an individual was believed to be bankrupt, we replaced their wealth value by zero.

We generally think these should be the wealthiest people in this country; it is not easy to extend the list to include more people, so the sample weight for this kind of database is defined as 1.

3.3 Online celebrities' income data

For the Online celebrities' database, we adopted crawler technology and downloaded the data from the most popular live broadcast platforms in China, which included Yizhibo, Momo, Huajia, Laifeng, Yingke, Meipai, Quanmin, and Huoshan. We also inquired how the reward money on the platform is distributed and find that a platform (company) shares 70% of the total reward while Online celebrities share the remaining 30%. Then, we calculated Online celebrities' personal income based on the distributional shares between the platform and the Online celebrities. The data crawling process covers a period from March 8, 2015, to April 7, 2018, for a total of 37 months. On the basis of the monthly income, we can also infer the yearly income in 2016.

According to the report on the Online celebrities' economy in 2016⁵, the number of the Online celebrities in 2016 had exceeded 1 million, while their annual income was approximately 250,000 yuan per person. Therefore, we assume that the number of Online celebrities in 2016 was 1 million, and the threshold value of their income was 250,000 yuan. We suppose that 500,000 Online celebrities earned more than the threshold value, so 2,536 samples were chosen from the database. Therefore, the weight of every Online celebrity is 197 ($50 \times 1000 / 2536$), and the weight is set to 200 for convenience.

3.4 Large company CEOs' income data

We also collected data on the salaries of CEOs of listed companies in 2016; these were obtained from listed companies' annual reports or China Wind Economic and Financial Database (hereinafter referred to as WIND database)⁶. Our database contains almost all their pay and

⁴ (1) If individual or family names of the lists are wrong or have been changed in different years, or the names of the companies they owned are inconsistent, we collected the right names from the State Administration of Industry and Commerce. (2) We deleted foreigners appearing in the list directly during the data cleaning.

⁵ <https://www.jianshu.com/p/43fe788b4f65>

⁶ We thank Professor Gao Minghua and Fang Fang of Beijing Normal University for their data support.

personal information. By the end of 2015, there were 2,827 companies and 60,540 individuals in the database. Among them, those with annual pay above 120,000 yuan are 31,963, which are kept for our analysis. However, there were many companies of the same size as the listed ones, but these were not listed in the stock market, so the information on the CEOs' pay for these companies is not available. We assume that those companies have the same level and distribution of their CEOs' pay as the listed companies do. Therefore, we gave each of the CEOs in the listed companies a weight of 17; the details are given below in section 5.1.

3.5 Rich Writer data

The database is combined by two sub-databases: the Famous writer, for which the data were collected from the public media, covering the richest traditional writers, online writers, cartoonists and screenwriters etc. in China. From the public media, the number of rich writers in 2016 reported by the Writer Rich List, Cartoonist Writer Rich List, Online Writer Rich List, and Screenwriter Writer Rich List was 60, 10, 20, and 30, respectively, and the threshold value was 1.8 million, 2.35 million, 10 million, 5.4 million yuan, and the highest value was 30 million, 10 million, 122 million and 14 million yuan, respectively. Finally, the overall sample size was 120, the threshold value was 1.8 million, and the weight was 1.

In addition, we also collected related data and prepared the Network writer list. The Qi-dian was a Chinese website (www.qidian.com), China's largest online literature platform, which covers most online writers. Our research team used crawler technology to obtain the following three types of important information about VIP writers' completed novels from this website: the number of words, the authors, and the time of completion⁷. Note that uncompleted novels are not included. Then, we can estimate how many words a VIP writer could write per day and how much he/she could be paid for every ten thousand words according to the platform rules. Moreover, we can calculate each VIP writer's total revenue in 2016. The following information needs to be explained carefully. First, this paper introduces VIP writers only and excludes non-VIP writers because VIP writers usually sign up with the platform and writing is their major occupation. Second, to avoid controversy, this paper calculates only the remuneration for completed novels, while that for the uncompleted novels is not included. Furthermore, these online writers have not appeared on the lists of writers above, so there is no intersection between the two types of writers. Finally, we found that 472 online writers were eligible for our analysis, with a threshold of 120,000 yuan. There is evidence that there are approximately 5,000 online writers whose monthly income is more than 10,000 yuan⁸, so the weight of the online writers is 10 for simplicity.

3.6 Signed-up actors list data

China's movie actors and actresses are in high-income groups that have been gradually expanding and attracting increasing attention in recent years. We collected data on actors' and actresses' payments using crawler technology from the two aspects of TV series and films. For TV series, we collected the following data about all the TV series screened in 2016: the titles of the TV series, the number of the main performers (up to 16 people), episodes, and the types of TV series. By referring to the investment in the same types of TV series and the remuneration of performers of the same grade, we deduced all the main performers' remuneration through manual search and analogy inference. Note that the remuneration of non-main performers is no longer included. For the movies, the same information as the TV series was collected about all the movies released in

⁷ <https://www.qidian.com/finish?action=hidden&orderId=3&page=1&vip=1&sign=1&style=2&pageSize=50&siteid=1&pubflag=0&hiddenField=2>

⁸ http://www.360doc.com/content/17/0117/09/7863900_622985882.shtml

2016. We used a similar method to obtain all the main performers' remuneration. In particular, the same actor or actress may act in several TV series or films, and we merged the TV series and film remuneration to obtain the performer's total remuneration. Obviously, performers who were not Chinese nationals were excluded. Finally, the number of eligible performers was 4,092.

Actually, no fixed definition is used to describe movie and TV-series actors; performers who have ever performed in movies, TV shows, song performance, dramas and even online dramas are all actors. What's more, all extras and the performers who have only a few lines can be called actors too. However, this paper only analyses actors for whom performance is a major occupation. The problem is that there are no relevant figures showing how many performers for whom performance is their major occupation. After multiple verifications, we decided to measure the number of film and TV-series actors by the number of contracted actors. In fact, the movies and TV series listed above were released nationwide, so it is reasonable to believe that all the main performers acting in them are contracted actors and actresses. There is evidence showing that the number of contracted actors in China exceeds 100,000⁹. At the same time, the China Statistical Yearbook reveals that the number of employees in China's culture, sports and entertainment industry was 1.508 million in 2015, so the number of contracted actors has strong credibility. Thus, we included data on 4,092 movie and TV-series actors and actress, so the weight for this group is 25 for simplicity.

3.7 Private enterprise owner data

The data for this group come from the Chinese Private Enterprise Survey (CPES), which has been conducted for many years and is organized by the Chinese Academy of Social Sciences and other organizations. The survey asks the sampled private owners detailed information about their assets and personal and enterprise income. Considering that it is most likely that their income is under-reported, we use information on their assets, by which we impute their income. The method assumes that the return rate of their assets is 5% annually, which is multiplied by their net assets. To derive individual income from family income, we assume that each individual has a family with three members on average. Finally, we have 3,000 individual owners with an annual income of more than 120,000 yuan in our data. Because there are approximately 3.44 million private enterprises in 2016, we gave each of the observations in the data a weight of 1,000.

3.8 Top Athletes

These data are compiled by including information from 4 sub-databases, which we introduce as follows: We collect the income of rich sportsmen and sportswomen from public sources, and the income information for the richest E-sports mobilization and sports athletes comes from professional industry associations. In each year, information on the top 100 income leaders is provided to the public by the international E-sports athlete association; we only keep the 28 Chinese income pioneers from the list in 2016, and then we manually search for demographic information on them including age, gender and education¹⁰. For the traditional sports athletes, the most authoritative sports media "Sports Weekly" releases a "most earned athletes list" each year¹¹, and there are 10 major names on this list in 2016. In general, we have 38 top-income individuals for athletics; their income range is 2.8 million to 69 million yuan, and the weight is 1.

⁹ <https://zhidao.baidu.com/question/363190417174966292.html>

¹⁰ <http://www.zhuobufan.com/Note/79f82f80-801b-470f-8ed7-9ed4cd1cae8c/>,
<http://news.gamedog.cn/a/20170117/2014367.html>

¹¹ <http://baijiahao.baidu.com/s?id=1581885354574463382&wfr=spider&for=pc>

Moreover, we also collected income data on football and basketball professional athletes. We do this because football and basketball were the first and best professionalized sports in China. The Chinese Super League and China Basketball Association (CBA) League are China's highest-level professional leagues, so the players participating in these leagues tend to have high incomes, and the number of teams participating in these leagues are 16 and 20, respectively. Through multiple visits and using our judgment, we obtained salary data for the Chinese Super League team—Shanghai Shanggang, as well as for the CBA team—Guangdong Dongguan. Then, we weighted the salary by weights of 16 and 20. In view of the large differences in the winning and attendance bonuses between the different clubs, only the basic salary was concluded, while other income was not considered in the calculation. Note that foreign players were excluded from the samples. In addition, one sample of the top 10 athletes was repeated and was removed from the data as a result.

In summary, the weights of E-sports, Top 10 Athletes, football players and basketball players are 1, 1, 16 and 20, respectively.

3.9 Databases tabulation without weighting

Generally, we have seven categories of the top incomes, which were distributed in the Signed-up actors List database, Hurun Rich List database, Executive Pay of Listed Companies database, Private Entrepreneur database, Online Celebrity database, Rich Writer database, and Top Athletes database. Table 3.3 provides a simple description of the seven databases before weighting and indicates that the individuals on the Hurun Rich List represent the richest in the population. Meanwhile, the density distribution of the mixed top-income databases before weighting can be found in Figure 3.1, and we find that the samples are not smoothly distributed.

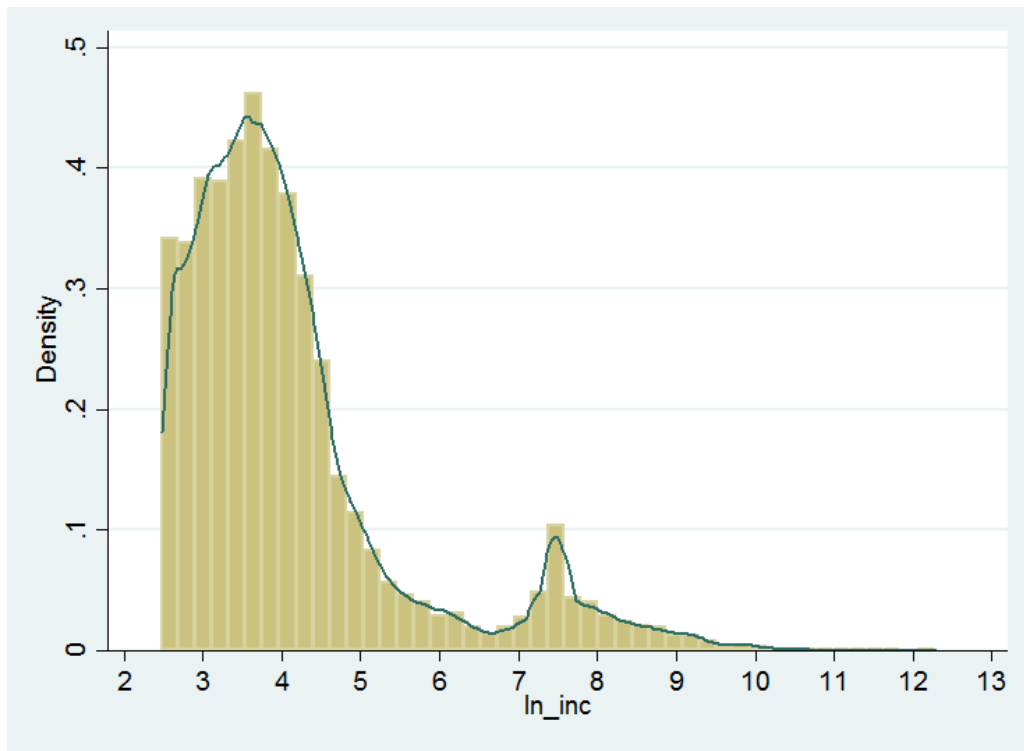
Table 3.3 Sample distribution by income interval for TIC data without weighting

	Signed-up actors	Hurn Rich list	CEO of Listed Companies	Private Entrepreneur	Online Celebrity	Rich Writer	Top Athletes
12-25	459	0	8338	389	2315	401	2
25-50	710	0	11266	307	1128	71	6
50-100	658	0	8710	191	498	0	7
100-1000	1660	14	3636	200	241	81	48
1000-1500	177	227	12	15	0	15	8
1500-2000	131	1218	1	3	1	10	0
2000-5000	240	1090	0	6	0	11	7
5000-10000	48	487	0	2	0	2	2
10000-20000	9	207	0	3	0	1	0
20000-50000	0	71	0	0	0	0	0
50000-100000	0	16	0	0	0	0	0
More than 100000	0	3	0	0	0	0	0
Maximum	17559.60	215000.00	1549.40	14514.30	1528.35	12200.00	6900.00
Minimum	12.00	818.18	12.00	12.00	12.00	21.90	20.00
Average	549.94	4571.86	58.04	156.07	38.59	238.36	782.25
Samples	4092	3333	31963	1116	4183	592	80

Note: unit in this table is 10 thousand.

Source: authors' calculation based on the TIC data.

Figure 3.1 TIC kernel density before weighting



Source: authors' calculation based on the CHIP and TIC data.

In general, we collected information on several types of top incomes based on the seven top-income categories, but we are worried about the representation of the collected top incomes. In other words, the over- or under-representation issue for the collected data creates a great challenge for us, and the seriousness of the over- or under-representation issue for the data is still not clear. Therefore, we think there is a problem of significant loss of income range among the overall income distribution. Generally, all of them point to the direction of the weighting.

4 Main Ideas and Steps to Obtain the Mixed Data

4.1 Basic assumption

Generally, our target is to estimate the exact income distribution using the collected top incomes with the combination of household survey (HS) data. Although we are confident about the ability of our HS data to capture the low, medium and even high incomes, we are not so certain about the representatives of the relatively low-income individuals for the TIC data. Meanwhile, we are also very cautious when matching the two data sources together, as they are from very different sources, as is known.

Using the well-known assumption and practical simulation based on our HS data, we find that the logarithm of the income at the beginning of the lowest salaries usually satisfy the normal distribution; however, afterwards, at a specific point from the threshold, the income distribution usually satisfies the Pareto distribution.

Therefore, in this paper, we first assume that the medium- and low-income households in HS satisfy the normal distribution, while the top incomes in the TIC satisfy the characteristics using the Pareto function. Although the HS data are not able to capture all of the top incomes, it is still

possible to capture some high income, although there are very few in the HS data. Most of the TIC data include information on the super-rich, but there is still some of the sample's income level that is not as high as we expected, and their income is not high enough to be classified as top incomes. In other words, the income variance or heterogeneity within the TIC is still very large. On the basis of the two facts mentioned above, therefore, we must confront the issue of how to link or match the two data sources to avoid individuals being double counted during the intersection or connected region.

4.2 The Introduction of Pareto function

Atkinson et al. (2011), using income tax data for major economies, wanted to measure the income share of the top 5 or 1 or 0.1%. They identified these precise top shares by imputing them using the Pareto distribution. They used the actual distribution below 5 (or 1 or 0.1) % but applied the Pareto distribution for the top incomes. Their justification was that 'a number of top income studies conclude that the Pareto approximation works remarkably well'.

Many studies suggest that the distribution characteristics of top income people can be fitted by Pareto distribution (Li and Luo, 2011). The basic form of the Pareto distribution is

$$\ln N(x) = \ln K - \alpha \ln x, k > 0, \alpha \geq 1 \quad (1)$$

Among them, $N(x)$ is the population whose income is x (that is threshold value) or above. K and α are parameters, and $1/(2\alpha - 1)$ is Gini coefficient. In fact, the greater the value of α , the fairer the distribution is, in other words, the smaller the α is, the more unfair the distribution is. We can calculate the estimated values of the parameters of the Pareto distribution by using the obtained income data of the high-income population. Rewrite the equation (1) in the form of cumulative distribution function:

$$\ln[1-F(x)] = \ln A - \alpha \ln x \quad (2)$$

Among them, $F(x)$ is the proportion of population whose income is less than or equal to x . Naturally, $1-F(x)$ indicates the proportion of population whose income is more than or equal to x . Assume $\ln A = \alpha \ln x_0$, x_0 is the threshold value. Then, we can rewrite Pareto distribution function into the following:

$$F(x) = 1 - (x_0/x)^\alpha \quad (3)$$

Whose density function is $f(x) = \alpha x_0^\alpha / x^{\alpha+1}$ and the corresponding mean value is $\alpha x_0 / (\alpha - 1)$. Because the parameters of Pareto distribution are strictly positive, and the distribution has only one right tail which is thicker than normal distribution, we can fit income distribution well in the empirical study. Furthermore, Pareto distribution is most suitable for fitting the income distribution which is larger than a specific threshold value (initial value) due to $x \geq x_0$. In the following part, we calculate the relevant results by setting the different threshold values. Note that in the estimation, it is necessary to ensure that the value of α is greater than 1, so that Pareto distribution can be used for simulation.

4.3 Steps to link and obtain the mixed data

Here, we introduce how we link the household survey data with the top income data to obtain mixed data. We separate this process into six steps:

Step 1: We collect only the top incomes starting with 0.12 million yuan from different fields, industries and sources, and then, we organize and call this data the original top income data (TIC in short).

Step 2: We find that the original TIC data satisfy the Pareto distribution at 7.7 million yuan. In other words, the samples below 7.7 million cannot satisfy the Pareto function distribution, or the collected samples below 7.7 million in the original TIC are still underestimated, so we have to simulate the samples below 7.7 million.

Step 3: For the samples higher than 7.7 million yuan in the original TIC, we can estimate the basic Pareto distribution parameter, such as the alpha coefficient.

Step 4: The income intervals from 0.65 to 7.7 million yuan cannot be collected by the household survey, and it was also underestimated according to step 2, so we have to simulate them by using other ways. According to step 3, we assume that both income intervals (0.65-7.7 million and 7.7 million+) satisfy the same Pareto distribution with the same alpha coefficient, so we also assume that the income interval from 0.65 to 7.7 million also satisfies the same Pareto distribution. Therefore, we simulate the samples from 0.65 to 7.7 million using the same Pareto function with the same alpha coefficient.

Step 5: We have several sections on the overall income distribution, while combining the household survey with the top incomes, and then we called it mixed data, which is shown in Table 4.1. The first section is the income interval from the minimum value to 0.65 million yuan, which is data from the household survey, which mainly satisfies the log normal distribution as low and median incomes. For the second income group of 0.65 to 7.7 million, we name it high incomes as it cannot be captured by the ordinary household survey, and we obtain those samples using the simulation method based on the same Pareto parameter that is used for the other income intervals of more than 7.7 million. Finally, the third income interval is the group with income greater than 7.7 million. We use the actually collected top incomes from the TIC data, and they also satisfy the Pareto function according to our estimation.

Step 6: Generally, the low and medium incomes are captured by the HS data, and the high and top incomes are captured by the TIC data. If we connect all the income intervals based on the three groups shown in Table 4.1, the different income parts of the datasets can be linked very well, as there is no overlap, and the overall income distribution for the combined dataset also goes smoothly from the low, medium, high to top-income groups. In particular, the overall samples change from the log normal distribution to the Pareto function distribution at the threshold of 0.65 million yuan. Therefore, using this way, we can capture all of the actual income distribution with not only the household survey but also data on the top incomes, so we can re-estimate income inequality or other indicators based on the new mixed databases.

Table 4.1 Income intervals for the Mixed data

Income intervals(ten thousand)	Data source	Income category	Distribution function	Note
0-65	HS	Low and Medium	log normal	The HS should be more accurate based on subjective judgment, while the objective fact finds it satisfy log normal distribution
65-770	Pareto simulation	High	pareto	The TIC should be more accurate based on subjective judgment, while the we simulate the samples based on the pareto distribution
>=770	TIC	Top	pareto	The TIC should be more accurate based on subjective judgment, while the objective fact finds it satisfy pareto distribution

Note: From here and after, we call 0-0.65 million yuan as low and median income, 0.65-7.7 million as high income, while after 7.7 million is named as top incomes.

Source: authors' calculation based on the CHIP and TIC data.

5 Matching top incomes with household survey

5.1 Weighting issues for the TIC

Before the formal estimation of income inequality, we must seriously address the weighting issue that was also noted in the sections before.

Regarding the names listed on the Hurun Rich List, we generally think these should be the wealthiest people in this country; it is not easy to add more people to the list, so the sample weight for this kind of database is defined as 1. After final cleaning, we obtained 3,333 rich families and 19,020 rich individuals using the information in Table 5.1; the details are presented in section 6.2. It is worth noting that the families' assets should be divided by the number of family members to obtain the per capita assets, and then, the per capita annual income of each family member can be calculated at a 5% return rate from their per capital assets.

Table 5.1 The distribution with Family numbers of Hurun rich list

Family members	Frequency	Family population	Family population/Overall population
3	5	15	0.08%
4	28	112	0.59%
5	1803	9015	47.40%
6	1003	6018	31.64%
7	290	2030	10.67%
8	80	640	3.36%
9	70	630	3.31%
10	34	340	1.79%
11	20	220	1.16%

Source: authors' calculation based on the public network information.

For the film and television actors and actress, the weight is set to 25, which is supported by evidence on different aspects: first, there is evidence that the number of actors signed in China is more than 100,000, while the total number of actors and actress after weighting is slightly larger than 100,000; second, China's statistical yearbook shows that China's cultural, sports and entertainment industry in 2016 included approximately 1.5 million employees, so the total number

of people who signed a contract after weighting in line with people's expectations. Finally, more than 30 persons' incomes are over 0.1 billion yuan¹²; however, our signed-up actors database captures 9 persons' incomes that are higher than 0.1 billion. Considering the rapid increase in movie and TV actor's salaries and the serious tax evasion of recent years, we have expanded the actor database.

Meanwhile, for the group of Online celebrities, according to the report on the Online celebrities' economy in 2016¹³, the number of Online celebrities in 2016 exceeded 1 million, while the annual income reached 250,000 yuan per person. Therefore, we assume that the number of Online celebrities in 2016 was 1 million, and the threshold value of their income was 250,000 yuan. Suppose that 500,000 Online celebrities earned more money than the threshold value, so 2,536 samples were chosen from the database, and the weight of each Online celebrity is 200.

For the top managers of the listed companies, we also adjust the weight according to the share of stated-owned companies in China. In general, Chinese enterprises can be divided into state- and privately-owned enterprises and can also be separated into listed and not listed companies. Data on the top managers of the stated-owned or privately-owned listed companies are fully captured by the database of the "top managers in the listed companies", and information on the top managers that belong to the not listed companies was captured by the database of "Private Entrepreneur data". Therefore, the only group that cannot be captured is the top managers in stated-owned and not listed companies. In 2016, 2% of the stated-owned companies were listed in the A-share or Hongkong-share list, while 98% of the state-owned companies were not listed in the capital market. Meanwhile, there are 944 stated-owned companies and 1,881 non-state-owned companies within the listed companies for the original "top manager database". Therefore, we could expand the top manager samples using the weight of 17 in this paper¹⁴; we named the database the "top managers' database" with the weighting.

For the group of the rich list of famous writers, we set its weighting to 1. In fact, whether the writer is in the National Writers Association of China or provincial and municipal associations, the main income of the association members is still the income coming from financial allocation or wages within the bureaucratic system, and their market income or copyright revenue is not so high because of illegal copies or small circulation. As a result, their incomes are not higher than even some white collars¹⁵, so in this paper, we think the list of the famous writers could include all of the top incomes, and the weighting for famous writers is also fixed to be 1. For online writers, there are 472 samples with annual salaries of more than 120,000 yuan and approximately 5,000 online writers have a monthly income of more than 10,000 yuan. For simplicity, the weight of the online writers is 10. Therefore, the weights of famous writers and online writers are 1 and 10, respectively.

For the group of private entrepreneurs, our original sample only captures 3,400 samples of families; however, there are approximately 34,400,000 privately owned companies according to the data provided by the China Private Entrepreneur Survey Report (2016), so we have to expand our samples using the weight of 1,000. Then, we could extend it to as many as 17,440 individuals using the household scale provided by the survey data. For the database of the Top Athletes, for E-sports and Top 10 Athletics, the sample weight is 1. For football and basketball players, the weights are 16 and 20, respectively.

¹² <http://ent.huanqiu.com/star/mingxing-neidi/2015-05/6501988.html>

¹³ <https://www.jianshu.com/p/43fe788b4f65>

¹⁴ $(50*944+1881)/(944+1881)$

¹⁵ http://news.ifeng.com/gundong/detail_2012_10/18/18350253_0.shtml

Finally, we sum up all seven top income databases with more than 7,868,120 individuals after weighting and then use this information to estimate the descriptive indicator and the income distribution. Compared to that of Figure 3.1, we find that the kernel function is much more smoothly distributed in Figure 5.2, so it also demonstrates the considerable necessity of using the weighting adjustment.

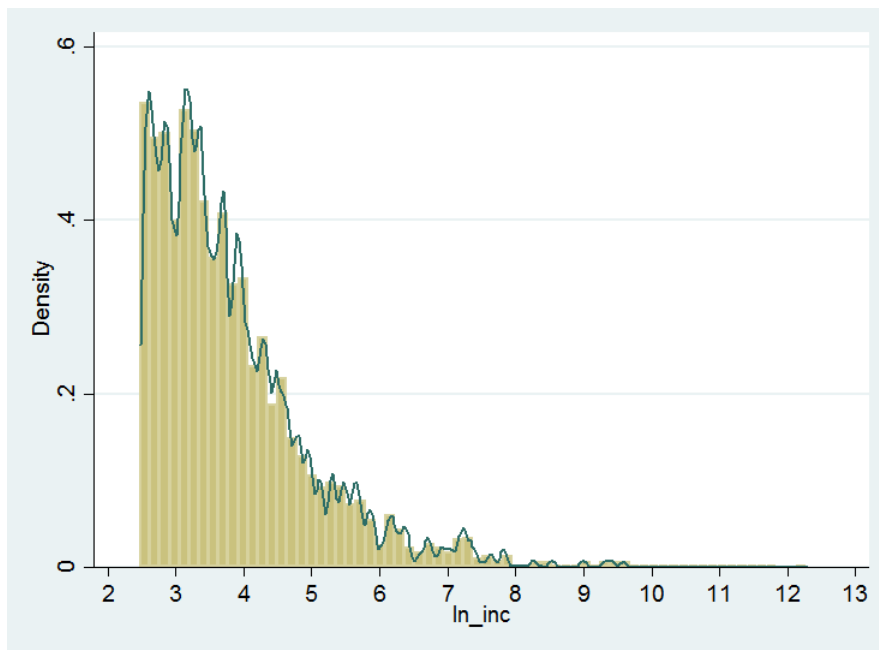
Table 5.2 Sample distribution by income interval for TIC data with weighting

	Signed-up actors	Hurun Rich list	Top manager	Private entrepreneu	Online celebrities	Rich Writer	Top athletics
12-25	11475	0	141746	2217000	463000	4010	40
25-50	17750	0	191522	1750000	225600	710	116
50-100	16450	0	148070	1089000	99600	0	120
100-1000	41500	140	61812	1140000	48200	81	477
1000-1500	4425	1727	204	85500	0	15	27
1500-2000	3275	6684	17	17100	200	10	0
2000-5000	6000	6130	0	34200	0	11	7
5000-10000	1200	2616	0	11400	0	2	2
10000-20000	225	1188	0	17100	0	1	0
20000-50000	0	416	0	0	0	0	0
50000-100000	0	99	0	0	0	0	0
more than 100000	0	20	0	0	0	0	0
Maximum	17559.60	215000.00	1549.40	14514.30	1528.35	12200.	6900.00
Minimum	12.00	818.18	12.00	12.00	12.00	21.90	20.00
Average	829.91	4564.29	128.56	450.45	139.27	1083.1	381.29
Samples	102300	19020	543371	6361200	836600	4840	789

Note: unit in this table is 10 thousand.

Source: authors' calculation based on the TIC data.

Figure 5.1 The kernel density function with TIC data after weighting



Source: authors' calculation based on the TIC data.

5.2 The Pareto parameter for TIC data

As a foundation, we think the samples satisfy the Pareto distribution from one specific threshold if the threshold is high enough, so we begin to find that threshold in the TIC database. In other words, the sample cannot satisfy the Pareto function below the threshold as a result of underestimation, underweighting or other reasons. However, the TIC after data collection and after the threshold is applied, should be good enough to capture all of the top incomes in China.

Thus, the next central issue is how we find the threshold or how we judge whether it satisfies the Pareto distribution or not for the samples beyond the threshold. We use the approach provided by Wang and Zhou (2006) who used the Lorenz curve to calculate the ideal Gini coefficient, which also obtained the actual Gini coefficient based on the OLS regression method. Then, we compare the gap of the Gini coefficient between the ideal and the actual one, where the gap is smaller and the Pareto simulation is better. We begin the simulation and estimation of the gap from the threshold of 12 million yuan and find that the smallest gap occurs around the interval of 7.7 million (see Table 5.3). Furthermore, we also try to obtain a more exact estimation of the Gini gap with more attempts and find that the point of 7.7 million yuan is the smallest gap among the thousands and millions of exercises.

Table 5.3 Search for the Pareto threshold for the TIC data

Iteration point	GINI Gap between two approaches	Actual Gini based on the OLS regression	Ideal Gini based on the Lorenz curve
700	0.764	0.500	0.263811
710	0.764	0.500	0.263697
720	0.764	0.500	0.263664
730	0.763	0.500	0.263544
740	0.763	0.500	0.263527
750	0.761	0.498	0.263188
760	0.761	0.497	0.263164
770	0.761	0.497	0.263161
780	0.759	0.495	0.263584
790	0.759	0.495	0.263586
800	0.759	0.495	0.263610

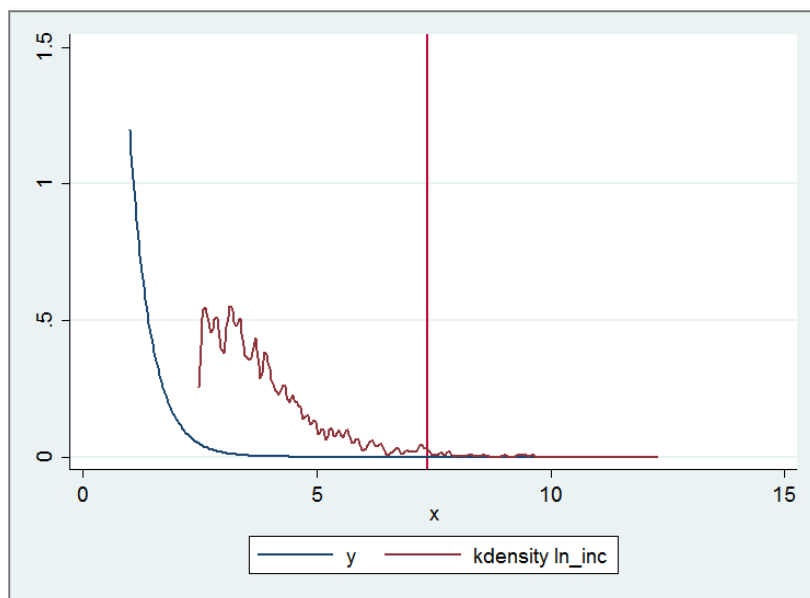
Source: authors' calculation based on the TIC data.

As seen from the figure below, the kernel density function matches the Pareto function very well after the threshold of 7.7 million yuan, and the alpha parameter of the Pareto function is 1.506 at this time¹⁶. In other words, within the TIC data, only the samples after the threshold could satisfy the Pareto function, and the samples that are lower than the threshold might be underestimated or suffer from less sampling, so we have to simulate those samples using other methods, as discussed in the next section. Here, we drop a large number of the TIC samples if we use the threshold of 7.7 million yuan. Some might argue that using a lower threshold here might help to keep a much larger sample from the TIC. However, the Gini coefficient gap between the two methods increases once we use other higher or lower thresholds; in other words, the threshold of 7.7 million should be the unique threshold given the specific TIC sample. Moreover, on the basis of the given TIC data, we could make sure we are using good enough representatives for the top incomes. If we could find the threshold of the Pareto distribution, then we would be able to deduce

¹⁶ We have to say here, most of individual's whole income over this threshold belong to the samples in the Hurun Rich List.

the next top samples using the same Pareto parameter, if all of them share the same Pareto distribution. Therefore, we still think the threshold of 7.7 million yuan should be the unique and best point fitting the given TIC samples.

Figure 5.2 The gap between OLS method and Lorenz method



Source: authors' calculation based on the TIC data.

5.3 The Introduction of HS data

Apart from the income intervals that are higher than 7.7 million and satisfy the Pareto distribution, we are still worried that the threshold might move forward to a lower point as the income intervals for less than 7.7 million yuan are under-sampled according to the analysis above. Furthermore, we are also thinking that the first threshold for the overall income distribution could also be found in the household survey because the end of the income tail in the household survey is also high enough, while the maximum value of the household survey is 0.65 million yuan. In other words, some of the samples with high enough income in the household survey might satisfy the Pareto distribution but not the log normal distribution.

In practice, we use the CHIP 2013 as one choice of household survey data. This article focuses on the income distribution in 2016, so we convert the per capita disposable income by the increasing rate of for urban and rural residents, and the demographic and sociological characteristics remain unchanged. Generally, the CHIP represents approximately 1,293,409,000 individuals after weighting, which accounts for approximately 93.0% of the overall population in China 2016. The data for this study could reflect the overall map of the Chinese income distribution once we add the TIC data to CHIP, as the population in 2016 was 1.383 billion. Meanwhile, the mean value for income in CHIP is 2.80 ten thousand, while the minimum and maximum value are 0 and 0.65 million, respectively.

5.4 Linking the HS and TIC

In this section, we link the household survey data to the top incomes. Before the formal connection, we must highlight three aspects of the entire income distribution.

First, for the income interval from 0 to 0.65 million yuan, we think the household survey is good enough to ensure data quality. Then, we have to simulate the income intervals using the Pareto

function from 0.65 to 7.7 million. Afterwards, we use the actual collected top-income samples after 7.7 million from the TIC. In general, we can connect the four income intervals into the overall income distribution and organize it as the mixed dataset.

Second, we argue that the samples from 0 to 0.65 million yuan are the low- and medium-income groups, while the income intervals from 0.65 to 7.7 million yuan are the high-income groups. Finally, we think the income level that is higher than 7.7 million yuan is the real top income, as the top income is not captured by the ordinary household survey.

Third, according to our simulation based on the household survey and TIC data, we find that the low and medium-income groups satisfy the log normal distribution, while the high and top incomes satisfy the Pareto function according to the actual analysis. In other words, we should also keep in mind that the income distribution changes from log normal to Pareto distribution at the smallest threshold of 0.65 million yuan, and the Pareto parameter is the same from 0.65 million, though it was applied to different income intervals.

5.5 National representation of the mixed data

For the representativeness of the mixed data, we trust the data quality of the household survey to reflect the low- and medium-income groups. As CHIP is widely used by scholars, it represents approximately 1.29 billion individuals whose income ranges from 0 to 7.7 million yuan. Moreover, for the data quality of the high and top incomes we collected, although we do not have enough self-confidence to trust the second largest income samples, we have enough confidence about the largest high and top incomes not only during the data collection procedure but also from the data analysis. Furthermore, we can rely on the data for individuals whose income is higher than 7.7 million to estimate the parameter using the Pareto function. Afterwards, we can move the same parameter to the second largest top income from 0.65 to 7.7 million using a simulation approach based on the same Pareto distribution. Therefore, we could also trust the representativeness of the income interval from 0.65 to 7.7 million.

6 Data Quality Discussion

Generally, in those steps above, we assume that the household survey data are good enough to capture the medium and low-income population, except for missing the high and top incomes. Therefore, we use other sources to supplement the top incomes and then link it to the household survey data, in order to get an overall distribution of the income groups in China. However, there are still some points about the data quality that should be discussed.

6.1 Why TIC begins from 120 thousand

While we collect the TIC data at the beginning, we truncate the data from the 0.12 million yuan for the top income individuals. The starting point for individuals to declare personal income tax is 0.12 million yuan according to the China's State Administration of Taxation, so we think the 0.12 million should be a reasonable threshold for the top incomes. Meanwhile, the mean disposable income for the highest-earning 20% population in the year of 2016 is 0.06 million according to the NBS, and the mean value of the highest 10% and 5% group's income is 0.079 and 0.13 million, so we argue that the 0.12 million should be a reasonable starting point for the top incomes, as the highest 5% average income from the household survey is already higher than the 0.12 million. At last, we only keep the samples for the top incomes whose annual income is higher than 0.12 million

yuan, as it was widely regarded as the starting point of the high-income groups not only from the public but also from the ministry of tax system in China.

6.2 Household scale discussion in terms of the Hurun Rich List

The Hurun Rich List differs in ways of defining the household scale across years. Entries on the list can take the form of either individuals, couples, or households, which causes great difficulty for our subsequent analysis. To ensure comparability, we must estimate their household sizes and then calculate the wealth per capita for the richest households. Notice that we always view the wealth level on the lists as wealth belonging to the whole household, regardless of the way the entries are defined. This is because in China, household members are often engaged in family businesses or operations because this is traditional practice. We use information from various sources, such as the Lists from previous years, resumes, and other Internet sources to jointly determine the household sizes. If there is inconsistency among the sources, we use the minimal values. By splitting the household observations into individuals, the sample size is enlarged from 3,333 to 19,032. For the detailed information about collecting the household scale, we have many steps:

1) for the name appearing in the form of Individual in the List, it accounts for around 40% of all entries. We first link them with other individual persons who appeared together with them on the Lists in the years before, and treat also the latter as the members of the household (exceptions include who were dead or put into prison in 2016). Then we collect their firm's basic descriptions, homepages, advertisements etc. from the internet to make sure whether their immediate relatives (spouses, children, and parents) are also engaged in family businesses (The union of these two groups will be treated as household members).

2) for the name appearing in the form of Household or Family in the Hurun Rich List at 2016, it accounted for around 60% of all entries. We determine the family sizes as follows: if the entries are immediate relatives, we use again the procedure described above to identify all household members engaging in family businesses; if the entries are brothers or sisters alike, we identify the household members for them separately and add up to total household sizes; if the entries are not relatives, for example, business partners, or investors which are very few in our sample, we collect information on ownership structure of the firms, and we also consider individual persons who own at least 10% shares of the firms as "household members".

However, the procedure above would also introduce some new measurement errors about the richest groups. Nevertheless, the procedure could have the advantage of increasing sample size of the group of interest, which could in turn correct the biases introduced by the richest individuals that do not appear on the Hurun Rich List. In particular, there are large amounts of invisible top incomes in practice in China, and Hu Run also admits in the public that the Rich List could only guarantee 60 percent of the accuracy¹⁷. Therefore, we tend to believe that the procedure is also useful for us because of the increasing top income samples according to those methods.

6.3 Income definition for the TIC and HS

The HS dataset contains information on the wages, operative profits, property incomes and transfers, all of which are components of household disposable incomes. Following the NBS, we also calculate the imputed rents from the houses owned by the richest households and individuals, which can be expressed as $\text{Imputed Rent} = 0.02 * (\text{current market value of the housing} - \text{total value of the housing when purchased})$. Dividing the household disposable incomes by family size gives

¹⁷ <http://finance.sina.com.cn/chanjing/gsnews/2017-10-13/doc-ifymviyp0960018.shtml>

the per capita disposable incomes. Non-negative disposable incomes are excluded, though there are few in the sample.

As for the TIC dataset, because of the great difficulties involved in data collection, the data quality might not be as good as we expected. We describe all of the main components in detail below.

As the Hurun Rich List only reports the rank of the wealth, we deduce the incomes from wealth based on a 5% annual return rate to the wealth. Because the true return rate that applies to the richest group might be higher than this, the imputed incomes may be underestimated. Choosing 5% as the return rate involves certain arbitrariness, yet evidence from the Chinese private sector has lent some support for it. For instance, Li and Luo (2011) show that the average profit rates for so-called above-scaled private enterprises were approximately 5% in 2005. Considering that most of the firms run by the richest persons appearing on the Lists are leaders in their own industries, the return rates could exceed 5%.

The observations from the Signed Actors List are comprised of mostly famous actors and actresses. We estimate their incomes from playing in TV series and films, as well as giving concerts. Other income, such as income from advertisements, are not included.

In our dataset, for the owners of private enterprises, we adopt a similar procedure as we did for the Hurun Rich List to impute the incomes generated by wealth, and we do not include other income. In our dataset, for the senior executives or managers of the listed companies, we mainly consider their wages and salaries from these companies and exclude income from other sources.

For the rich writers in our datasets, we consider mostly their copyright royalties (including offline sales, online purchases, copyright sales and website payments,) and exclude income from other sources. For online celebrities, we use data mining methods to impute their gifts obtained from users into monetary terms and then attribute 30% of this income to these celebrities. Other income sources, such as endorsements, advertising and other offline activity income, are also excluded.

For professional e-sport players, we mainly consider their prizes from various competitions and advertisement incomes and do not include other types of income, such as income from investment. For athletes, we focus on their prizes from various competitions and advertisement incomes. Their investment income is excluded.

6.4 Weighting of the TIC

TIC data include top incomes from different data sources, and these weights are set according to several types of strong evidence provided by some industry associations or professional institutions.

Generally, the Hurun Rich List basically covers information on the richest people in China according to the specific organization, the Writer Rich list summarizes income information on China's top writers supported by the national writers association, and the athlete data are also released by authoritative organizations, all of which cover the data of the top earners in their fields, and these databases do not need to be weighted anymore.

For other categories of the databases that need to be weighted, we provide strong evidence to weight the data of private entrepreneurs and Online celebrities. Meanwhile, although the weight of the signed actors list is not supported by official or professional evidence, we also try to collect national information that can be accepted by different opinions. For the signed-up actors list and listed company executive data, we have tried to refer to the suggestions of experts and scholars or

used information from authoritative media. Although the weights may not be very accurate, the truth of the fact has been largely reflected.

In summary, it is undeniable that there may be some controversy about the specific weights used for each top income database, but we argue that the weight is much more accurate than before, and it is much closer to real facts. We do not make great demands to set the weight very precisely; what we think truly matters is to ensure the authenticity and reliability of the conclusion to the greatest extent.

6.5 Duplications of the seven top income databases

We obtained the TIC data by weighting the databases from seven different types of sources. However, one notable problem regarding the data is that there are duplications or overlaps from different sources. We conduct three activities to avoid overlaps to reduce their influence on the income distribution.

First, we are sure that there is a large difference between the occupations of individuals from different sources, so the possibility of sample repetition is very low. Second, we tried to avoid the same individual appearing in different lists during our data collection, which further reduces the possibility of overlaps. Finally, for those who may appear in different lists at the same time, although this is very rare, such as private entrepreneurs and individuals in the Hurun Rich list, we checked one by one through the actual names and manual searching from the Internet, which also avoids the duplication of samples. In summary, the data of seven categories of high- or top-income people will not be seriously affected by data duplication when they are merged together.

6.6 Overlaps of HS and TIC

Another issue related to the overlaps might also occur if we link the HS and TIC together. Although the HS is not able to collect the top incomes, the high incomes are very likely to be covered, which could also be found in the TIC data. In our CHIP data, the income level ranges from 0 to 0.65 million yuan, while the TIC also collects the income interval from 0.12 million to 2.15 billion, so the range from 0.12 million to 0.65 million is overlapped by the TIC and HS data¹⁸, and it is not realistic to check the duplication one by one.

Although the HS data cannot accurately contain top-income groups, we suppose that the HS data contain income data of stratum below the high-income groups. In other words, we trust that the representativeness of the low and medium samples is sufficient by using the household survey data, although it is impossible to capture the high or top incomes with the survey data. Therefore, we tend to use the household survey data for the income interval from 0 to 0.65 million yuan, and the range from 0.65 million to 7.7 million is simulated using the Pareto function, while the last income interval that includes more than 7.7 million are from the TIC data.

6.7 The meaning of collecting TIC

As the collected data's income interval from 0.65 to 7.7 million yuan is not fully used during the analysis above, someone might question the need to spend so much time collecting all of the top incomes starting with 0.65 million. Our argument relies on two aspects. First, on the basis of the overall collected samples from 0.12 million to 2.15 billion, we are able to simulate the distribution and finally find the threshold point for the Pareto distribution. In other words, we are not able to

¹⁸We think that the data from 0.12 to 7.7 million in TIC mainly plays the role of measuring the "7.7 million". The actual distribution of this interval is obtained by simulation, as shown in the following text. Therefore, even though there is repetition, the conclusion of this paper will not be affected.

find the best threshold if we do not have so many collected top incomes. Second, we can compare the distribution gap between the collected and simulated income distribution for the second largest top incomes if they are underestimated by the collected individuals. As seen from appendix 1, we find that a small representation of the collected samples as their density distribution is less than the simulated function; however, the gap is not as large as we expected from the figure. Therefore, we could have more confidence about the data quality for the collected top incomes.

7 Basic Results with the Mixed Data

It is worth noting that what we call “TIC” in this chapter includes data from 0.65 million to 2.15 billion yuan unless otherwise specified, while the number of observations after weighting in the TIC data is 2,247,357 after weighting. The ways for processing the data from 0.65 million to 7.7 million yuan can be classified as two methods: If we use the original value for this range, we call the TIC data the “original TIC after weighting from 0.65 million to 2.15 billion” (OTIC), when we use the simulated value with Pareto function for this range, we call the TIC the “simulated TIC after weighting from 0.65 million to 2.15 billion” (STIC). Meanwhile, if we connect the HS with the STIC data together, then we can obtain a whole map of the income distribution without any overlaps. This sample should also be a real and exact reflection of the Chinese income distribution, which we call mixed data (HS+STIC).

7.1 Comparison with NBS

Here, in this section, we try to compare the income level using different income sources. Taking the household survey as an example, the NBS’s average income per capita is the lowest, which is only 2.2 ten thousand yuan in 2016, while it is higher when using the CHIP household survey, as it reaches 2.8 ten thousand yuan. Meanwhile, if using aggregated data from other sources, such as the flows-of-funds table, then the average income per capita would reach as high as 3.1 ten thousand yuan. Moreover, it would even increase to 4.3 ten thousand yuan based on our mixed data in this paper, which demonstrates the apparent underestimation of the household income from different kinds of surveys. Lastly, based on the comparison of mixed data and GDP figures from NBS, we find that household income could account for 80 percent of the overall GDP, which is significantly higher than 62%, which is provided by the NBS.

Table 7.1 Different Income Sources Comparison

	NBS household survey	Flows-of-funds table from NBS	HS data in this paper	Mixed data in this paper	GDP from Statistical Yearbook
National average income per capita (Unit: ten thousand yuan)	2.1966	3.026	2.8034	4.3143	4.9992
National aggregate income (Unit: trillion)	30.120	41.488	32.890	54.879	68.551

Source: Authors.

7.2 Demographic characteristics using OTIC data

In the following section, we use the OTIC data to investigate the demographic characteristics of the top-income groups. The reason for this is that demographic characteristics such as gender, education level and age for those with an income between 0.65 and 7.7 million yuan cannot be

simulated in the STIC data¹⁹. Although the OTIC data lack the representativeness of the range from 0.65 million to 7.7 million, which might result in the existence of a new bias, it still reflects part of reality to a certain extent. Next, the primary target is the provincial distribution and the distribution in terms of gender, education and the year of birth.

First is provincial distribution, there are two types of definitions, one is the birth place and the other one is the workplace. In a sub-sample, such as the birth place or the native place of the Online celebrities, it is not easy to search by ourselves, but it is relatively easy to obtain their workplace. As for the Hurun rich list, the signed-up actors' data, rich writers and top athletes, the information of the workplace and birthplace are relatively easy to be obtained. And for the private entrepreneur owners and executives of listed companies, we can set the workplace based on the location of their corporate headquarters. To sum up, we chose the location of the workplace to identify the province. It is worth noting that a small sample failed to find the workplace, so the sample size after weighting decreased from 2,247,357 to 1,145,782 in the same year.

Table 7.2 population distribution for OTIC data

Province	Provincial OTIC population	Provincial OTIC population / Total OTIC population	Provincial population/ National population	Provincial OTIC population/ Provincial overall population
Guangdong	120422	10.51%	7.91%	0.13%
Beijing	98537	8.60%	1.58%	0.55%
Shanghai	92465	8.07%	1.76%	0.46%
Zhejiang	89944	7.85%	4.04%	0.20%
Jiangsu	85132	7.43%	5.82%	0.13%
Shandong	78028	6.81%	7.18%	0.10%
Fujian	72986	6.37%	2.80%	0.23%
Hunan	70809	6.18%	4.95%	0.13%
Laoning	63362	5.53%	3.20%	0.17%
Hubei	52477	4.58%	4.27%	0.11%
Tianjin	44801	3.91%	1.13%	0.35%
Shanxi	37925	3.31%	2.67%	0.13%
Helongjiang	31967	2.79%	2.80%	0.10%
Sichuan	24405	2.13%	5.98%	0.04%
Hebei	15124	1.32%	5.42%	0.02%
Chongqing	7791	0.68%	2.20%	0.03%
Inner Mongolia	6646	0.58%	1.83%	0.03%
Anhui	26238	2.29%	4.48%	0.05%
Hainan	24749	2.16%	0.66%	0.33%
Henan	23145	2.02%	6.91%	0.03%
Shanxi	21197	1.85%	2.77%	0.07%
Jilin	15124	1.32%	2.01%	0.07%
Jiangxi	11687	1.02%	3.33%	0.03%
Yunnan	7906	0.69%	3.46%	0.02%
Guangxi	6416	0.56%	3.50%	0.02%
Xinjiang	4124	0.36%	1.72%	0.02%
Guizhou	4011	0.35%	2.57%	0.01%
Tibet	2979	0.26%	0.24%	0.11%
Gansu	2292	0.20%	1.90%	0.01%
Ningxia	1604	0.17%	0.49%	0.03%
Qinghai	1489	0.14%	0.43%	0.03%

Source: authors' calculation based on the original TIC data.

Secondly is the gender distribution; it is worth noting that a small sample failed to find a workplace, so the sample size decreased from 2,247,357 to 981,762 after weighting. According to the

¹⁹ For the STIC, it is impossible to simulate the actual demographic characteristics for each individual, so in this section, we have to rely on the OTIC to describe the sample distribution of the top incomes.

comparison in Table 7.3, the proportion of males in the HS data (income below 0.65 million yuan) was only slightly higher than that of females, but the advantage of males was very obvious in the OTIC data, which also indicated apparent gender inequality in the top-income groups.

Table 7.3 Gender composition for the OTIC data and HS data

Gender	Specific gender OTIC population /overall OTIC population	Specific gender HS population /overall HS population
Male	78.64%	52.19%
Female	31.36%	47.81%

Source: authors' calculation based on the OTIC and CHIP data.

Thirdly, in terms of education attainment, it is worth noting that a small sample failed to find a workplace, so the sample size also decreased from 2,247,357 to 863,215 with weighting. Due to data limitations, we investigate only whether the individual obtains a college degree or above. According to the comparison in Table 7.4, only approximately 12 percent of the population in the HS data had a college degree and above, but that percentage almost completely reversed for the top incomes. This means that currently in China, people with relatively low educational backgrounds are significantly less likely to be in top-income groups, which also indicates the importance of education to accumulate human capital, especially to enter a higher social class.

Table 7.4 Education level composition for the OTIC data

Education	Specific education OTIC population /overall OTIC population	Specific education HS population /overall HS population
College and above	72.34%	12.43%
Below College	27.66%	87.57%

Source: authors' calculation based on the OTIC data.

Finally, in terms of birth year or age cohort, it is worth noting that a small sample failed to find a workplace, so the sample size decreased from 2,247,357 to 1,045,665 after weighting. For simplicity, the calculations are divided into different age groups. According to Table 7.4, most population proportions of the top-income groups in the OTIC data are mainly aged from 41 to 50 years old (33.09%), while the next is the group of those who are 51 to 60 years old (19.65%), 31 to 40 years old (18.62%) and 61 to 70 years old (15.05%). Young people dominate the top incomes according to Table 7.5, which is also in line with people's expectations. Compared with the conclusions using the OTIC data, the age of residents in the HS data was mainly in the following stages: under 20 years old (approximately 26.85%), 21 to 30 years old (approximately 19.36%), 41 to 50 years old (approximately 13.71%) and 51 to 60 years old (approximately 11.24%), as compared with OTIC data; as the age increased, the possibility of residents obtaining high income was significantly enhanced.

Table 7.5 Cohort composition for the OTIC data and HS data

Age group	OTIC data		HS data	
	Population within the Interval	Interval population/ overall OTIC population	Population within the Interval	Interval population/ overall HS population
70 years old	55629	5.32%	52585146	4.64%
61~70	157373	15.05%	90097394	7.95%
51~60	205473	19.65%	127382982	11.24%
41~50	346011	33.09%	155375506	13.71%
31~40	194703	18.62%	184161340	16.25%
21~30	86476	8.27%	219406987	19.36%
20 years old below	0	0.00%	304291198	26.85%
Overall	1045665	100.00%	1133300552	100%

Source: authors' calculation based on the TIC and CHIP data.

Table 7.6 shows the basic income distribution characteristics of the OTIC data. First, in terms of the mean value, the mean of the Hurun Rich List, private entrepreneur owners, signed actors and top athletes are much higher; all of them exceed 20 million yuan. The individuals on the Hurun Rich List have the highest income per capita.

Table 7.6 Income distribution for different databases within OTIC

	Signed actors	Hurun rich list	Top managers	Private entrepreneur	Online celebrities	Writer Rich list	Top Athletics
Mean	829.91	4564.29	128.56	450.45	139.27	1083.15	381.29
p90/p10	25.37	6.33	3.16	11.43	3.56	8.77	7.81
p90/p50	6.99	4.38	2.33	6.01	2.50	3.66	2.08
p10/p50	0.28	0.69	0.74	0.53	0.70	0.42	0.27
GINI	0.644	0.520	0.313	0.685	0.332	0.514	0.427

Note: after weighting.

Source: authors' calculation based on the OTIC data.

Second, for the ratio of p90 / p10, p90 / p50, or p10 / p50, there is a large difference between the private entrepreneur owners, the Hurun Rich List, and the rich writers, which demonstrates the considerable heterogeneity among the top incomes. Taking the Gini coefficient as an example, the indicator reaches as high as 0.685 among private business owners²⁰.

7.3 Income intervals for the mixed data using the STIC data

In the following section, we analyse the income interval distribution for the STIC data, while again limiting the range from 0.65 million to 2.15 billion yuan. The following is the share of top incomes for each income interval using the STIC data and the HS data.

²⁰ Lastly, the individuals on the Hurun list and the private entrepreneur owners of the OTIC accounted for the proportion of income, and the highest reached 86.58%. On the one hand, the private entrepreneurs tend to have higher incomes, which is because the number of private entrepreneur owners is larger. For the Hurun list, the income ratio is 12.43%. Even if the number is smaller, its large number of assets and income ensure that its income proportion is still very considerable, while the proportion of other people's income is very small, all less than 1%.

Table 7.7 The income intervals distribution using STIC + HS data

Income Interval (Unit: million)	Interval population	Interval population/overall STIC population	Interval population/overall Mixed data population	Accumulated Interval population	Accumulated Interval population/overall Mixed data population
<0	3705000	—	0.2842%	3705000	0.2841786%
0-0.12	1280461000	—	98.2131%	1284166000	98.4973140%
0.12-0.20	7631000	—	0.5853%	1291797000	99.0826223%
0.20-0.30	975000	—	0.0748%	1292772000	99.1574061%
0.30-0.40	351000	—	0.0269%	1293123000	99.1843283%
0.40-0.50	52000	—	0.0040%	1293175000	99.1883168%
0.50-0.65	234000	—	0.0179%	1293409000	99.2062649%
0.65-0.77	10098946	97.59%	77.460317%	1303507946	99.9808681%
0.77-1	49926	0.48%	0.382932%	1303557872	99.9846975%
1-2	119406	1.15%	0.915861%	1303677278	99.9938561%
2-3	34980	0.34%	0.268301%	1303712258	99.9965391%
3-5	11129	0.11%	0.085361%	1303723387	99.9973927%
5-10	15140	0.15%	0.116126%	1303738527	99.9985539%
10-20	18340	0.18%	0.140670%	1303756867	99.9999607%
20-50	399	0.00%	0.003060%	1303757266	99.9999913%
50-100	94	0.00%	0.000721%	1303757360	99.9999985%
>100	20	0.00%	0.000153%	1303757380	100.00%

Source: authors' calculation based on the STIC data.

Table 7.8 gives the minimum threshold for the different income quantiles, as well as the mean and the Gini coefficient of the group. Taking the top 0.8% as an example, 0.65 million yuan is the minimum threshold for entering China's top 0.1% income group; that is, the individual who is likely to enter into the top 0.1% income group earns at least 0.65 million yuan per year. Given that the maximum value of the HS data is 0.65 million, this means that the HS data basically covers the income of 99.2% of the residents in our country. Furthermore, the average income of China's top 0.1% top-income residents is 13.03 million, and the Gini coefficient of this group is 0.599. Meanwhile, the individuals whose income reaches 11.845 million, 2.568 million, 1.621 million, 1.023 million, 0.782 million, 0.649 million and 0.165 million can enter the top 0.01%, 0.1%, 0.2%, 0.4%, 0.6%, 0.8% and 1% of China's highest-income group, respectively. Furthermore, the threshold for entering the top 10% is 49.4 thousand.

Table 7.8 Income threshold for the top income groups using mixed data (HS+STIC)

Income level	Income threshold value (Unit: 10 thousand)	mean value within the Interval (Unit: 10 thousand)	Interval Gini index
Top0.01%	1184.45	3606.16	0.471
Top0.1%	256.82	1303.99	0.599
Top 0.2%	162.09	937.16	0.633
Top 0.4%	102.30	671.32	0.667
Top 0.6%	78.16	542.20	0.686
Top 0.8%	64.90	456.36	0.701
Top1%	16.45	236.97	0.784
Top2%	10.13	74.06	0.795
Top5%	6.74	29.01	0.698
Top10%	4.83	16.59	0.615
Top 20%	3.37	10.24	0.546
Top 30%	2.56	7.79	0.520
Top 40%	1.98	6.42	0.514
Top 50%	1.54	5.48	0.518
Top 60%	1.22	4.81	0.527
Top 70%	0.95	4.26	0.541
Top 80%	0.73	3.84	0.556
Top 90%	0.50	3.48	0.575

Source: authors' calculation based on the TIC and CHIP data.

7.4 Income share using STIC

From the following table 7.9, the STIC data are grouped by several income intervals, and note that these groups are not divided equally. Among them, the income interval ranging from 50 to 100 million accounts for as high as 2.57% of the total social wealth while the population rate only accounts for about 0.116‰. Meanwhile, in the range from 200 million to 500 million, 0.003‰ of the population share 0.28% of the overall social wealth. Last but not the least, the 0.212‰ of the population within the interval from 0.65 million to 7.7 million yuan also earn about 8.98% of the overall income in this country. Most important of all, only 0.0008‰ of the population's income is higher than 500 million yuan in this country, however their total income accounts for as high as 0.23% of the overall income in this country, which also shows the huge inequality level coming from the top incomes.

Table 7.9 Income intervals' share of total income in China using STIC

groups	Income intervals (ten thousand)	Population within intervals	intervals Population /national population within mixed data	Added intervals income /added national income within mixed data	Accumulated intervals income /added national income within mixed data
1	65-770	10098946	77.460317‰	8.98%	8.98%
2	770-1000	49925	0.382932‰	1.09%	10.08%
3	1000-2000	119406	0.915861‰	4.15%	14.22%
4	2000-3000	34980	0.268301‰	2.09%	16.31%
5	3000-5000	11129	0.085361‰	1.09%	17.40%
6	5000-10000	15140	0.116126‰	2.57%	19.97%
7	10000-20000	18340	0.140670‰	5.93%	25.90%
8	20000-50000	399	0.003060‰	0.28%	26.19%
9	50000-100000	94	0.000721‰	0.16%	26.35%
10	More than100000	20	0.000153‰	0.07%	26.43%

Source: authors' calculation based on the STIC data.

This part gives relevant information such as the mean value, minimum value, maximum value, proportion of total income and cumulative proportion after grouping ten from the statistics of STIC and HS data. From the table 7.10, first of all, the proportion of income from STIC data in ten-class subgroup increased steadily from 1.42% in group one to 74.96% in group ten, by contrast, different groups in the HS data increased their share of income by a larger margin, increasing from 0.63% in group one to 41.50% in group ten. Secondly, the cumulative percentage of income in the ten-class subgroup in STIC is higher than that in the HS data. Finally, the average, minimum and maximum values in the ten-class subgroup in the STIC data are extremely different in absolute values, but the ratio between the maximum and minimum values in each ten-class subgroup is relatively close.

Table 7.10 Income deciles description using STIC data and HS data(Unit: 10 ten thousand)

STIC data						HS data				
	mean	min	max	Overall Interval income/overall STIC income	Accumulated Interval income/overall STIC income	mean	min	max	Overall Interval income/overall HS income	Accumulated Interval income/overall HS income
1	67.25	64.90	69.50	1.42%	1.42%	0.31	-33.15	0.50	0.63%	0.63%
2	72.44	69.50	75.12	1.65%	3.08%	0.62	0.50	0.73	1.80%	2.43%
3	78.30	75.12	82.51	2.19%	5.26%	0.84	0.73	0.95	2.95%	5.38%
4	87.06	82.51	91.5	2.10%	7.36%	1.07	0.95	1.21	4.04%	9.42%
5	97.55	91.5	103.33	2.29%	9.65%	1.36	1.21	1.53	5.34%	14.76%
6	111.23	103.33	119.73	2.13%	11.78%	1.73	1.53	1.95	6.89%	21.66%
7	132.21	119.73	145.76	3.21%	14.99%	2.22	1.95	2.53	8.91%	30.57%
8	165.67	145.76	192.78	3.95%	18.94%	2.89	2.53	3.31	11.62%	42.19%
9	248.89	192.78	328.66	6.10%	25.04%	3.92	3.31	4.67	16.31%	58.50%
10	1620.48	328.66	215000	74.96%	100.00%	7.37	4.68	64.9	41.50%	100.00%

Note: after weighting.

Source: authors' calculation based on the STIC and HS data.

The following table gives information about the Mixed data which is grouped into ten deciles. It can be seen from the table that in the ten-class subgroup, the interval mean increases from 0.31 million yuan in group one to 15.74 million yuan in group ten; the difference between the minimum and maximum values in the top nine groups is not very large, regardless of the comparison in the absolute sense or in the relative sense, but there is a huge difference between the maximum value and the minimum value in group ten. This is also understandable because the sample size in the STIC data is small, even if group ten is even smaller.

For the proportion of income in the mixed data shared by different deciles, in the top six group, the income share does not exceed 10%, and the group seven slightly higher than 10%, and in the group ten its share reached 32.19% which reflects the huge difference in the income of residents in mixed data. As can be seen in the cumulative proportion of resident income, the cumulative proportion of the top five groups is slightly higher than 18%, and the cumulative proportion of the top eight groups is about 50%. The final two groups have a cumulative income of 50%, basically it's a universal consensus of the two-eight law.

Table 7.11 Income deciles description for Mixed data using STIC data (million)

	obs	mean	min	max	Overall Interval income/overall income in the mixed data	Accumulated Interval income/overall income in the mixed data
1	130286000	0.31	-0.33	0.5	1.37%	1.37%
2	130286000	0.62	0.5	0.73	2.77%	4.14%
3	130312000	0.84	0.73	0.95	3.78%	7.92%
4	130273000	1.08	0.95	1.22	4.84%	12.76%
5	130312000	1.38	1.22	1.54	6.15%	18.91%
6	130286000	1.75	1.54	1.98	7.80%	26.71%
7	130299000	2.24	1.98	2.56	10.03%	36.74%
8	130286000	2.93	2.56	3.37	13.13%	49.87%
9	130312000	4.01	3.37	4.83	17.94%	67.81%
10	131105380	15.74	4.83	215000	32.19%	100.00%

Note: after weighting.

Source: authors' calculation based on the STIC data.

7.5 Gini coefficient for the STIC and mixed data

Before we formally use the top income database to estimate the real overall income inequality, we can use the household data to conduct a test to help us to better understand the importance of the top incomes. Using the method suggested by Knight et al. (2018), who use only a household survey and expand the highest-income decile and then re-estimate the income inequality when expanding the highest household samples. Assume that we have a good household survey to capture enough about the high, medium and low-income population (except for the top incomes), so we would do an exercise to test the sensibility if expanding the highest-income samples from the household survey.

Using the HS data, we simulate the mean value and Gini coefficient of the income level before and after expanding the highest-income samples from the household survey. Table 7.12 shows the original sample without any sample expansion. The remaining rows represent the results of the 1%, 5% and 10% expansion for the top samples within the household survey. Obviously, the average income will increase with the expansion of the high-income sample. For the Gini coefficient, which we are more concerned about, in the CHIP sample in 2016, the Gini coefficient before expansion is approximately 0.464. Longitudinal comparisons show that when the high-income sample is expanded twice, the Gini coefficient increases with the proportion of the highest-income group, which is 0.485 (expanded 1%), 0.496 (expanded 5%) and 0.489 (expanded 10%). Furthermore, if all of the high-income samples are expanded three times, the Gini coefficient presents a trend of first an increase and then a decrease. The Gini coefficient dropped from 0.503 to 0.492 when the top 10% of the high-income sample was expanded three times.

Table 7.12 Impact of top incomes on the overall income distribution using HS data

	2 times expansion		3 times expansion	
	Mean value	Gini	Mean value	Gini
Original CHIP	2.976	0.464	2.23	0.464
Expansion of the highest 1%	3.239	0.485	3.333	0.503
Expansion of the highest 5%	3.516	0.496	3.629	0.513
Expansion of the highest 10%	2.976	0.489	3.938	0.492

Source: authors' calculation based on the CHIP data.

Apart from the simulation to test the importance of the highest-income groups, we also perform a simple calculation of the Gini coefficient. The Pareto distribution, based on the hybrid data, would be able to obtain the adjusted Gini coefficient, which includes the top-income population (Again, this is what we call the mixed data, the CHIP data and the TIC data, which include more than 0.65 million with weighting). The meanings of the adjusted Gini coefficient and related indicators are as follows: the specific calculation of the seven indicators is referred to in appendix 2 and is not described here.

$$G = p1^2 \times \frac{u1}{u} \times G1 + p2^2 \times \frac{u2}{u} \times G2 + p1 \times p2 \times \left| \frac{u2 - u1}{u} \right|$$

where, p1 represents the proportion of the total population whose income is less than 0.65 million in CHIP to the total population in the mixed data; u1 represents the average income if income is less than 0.65 million in the CHIP; and p2 represents the proportion of the population whose income is larger than 0.65 million to the total population in the mixed data. u2 represents income per capita if the income was higher than 0.65 million; u represents the average income in the mixed data; G1 represents the amount of the population whose income was less than 0.65 million; and G2 represents the Gini coefficient for those whose income was more than 0.65 million.

By substituting the above seven indicators, we can obtain the Gini coefficient, which is 0.646. Generally, the Gini coefficient of the household survey was 0.465 in 2016 according to the NBS. However, compared with the results in this paper, the Gini coefficient will increase to 0.646 after including the top incomes, or the Gini index would increase approximately 0.208 percentage points. Therefore, the considerable influence of the Gini coefficient because of the introduction of TIC data is very clear, and it also supports the great necessity to include the top incomes during income inequality research. Moreover, this is of great significance for the correct and comprehensive understanding of the reality of income distribution in China.

7.6 Income inequality using Mixed data

We calculate overall income inequality for the TIC in column one of Table 7.13. We also estimate the income inequality index using the HS data, STIC data and mixed data, and then, we calculate the results using the indicator of the weighting.

We estimate overall income inequality while including all of the top incomes from different sources. As seen in Table 7.12, income inequality using the STIC data is not as high, and the Gini coefficient is 0.497 for 2016. In comparison, income inequality using the HS data is only 0.464 in the same year. If we merge the household survey and top income data base together, the Gini coefficient of the entire income distribution reaches as high as 0.646.

This result shows that the HS seriously underestimates the income inequality of Chinese residents. Notably, the Gini coefficient for our HS data is close to 0.465, which was published by the NBS, with a difference of approximately 0.001. However, regardless of what it is, the HS data represent an undisputed reality about the disparity in the incomes of residents.

Then, for the mean value of the samples, the income per capita in HS data is approximately 28.0 thousand, while it is 1.932 million and 41.3 thousand for the STIC and the Mixed data, respectively. Compared with the HS data, income per capita for the mixed data actually increased by more than 47.5%, which also supports the necessity of including the TIC data together with the household survey data. Furthermore, if we use the index of p90/p10, we find that it jumps from 9.34 in HS to 4.73 if we add the database of the top incomes. Meanwhile, the mean value of the household

survey is 28.0 thousand, while it is 1931.6 thousand for the TIC data. After merging them together, the average income for the overall data should jump to 30.2 ten thousand.

Table 7.13 Income inequality

	HS data	STIC data	Mixed data (HS+STIC)
Mean(Unit: ten thousand)	2.80	193.16	4.13
P90/p10	9.34	4.73	10.25
P90/p50	3.06	3.18	3.18
P10/p50	0.32	0.67	0.35
GINI	0.464	0.497	0.646
Samples	1293409000	10348380	1303757380

Source: authors' calculation based on the CHIP and TIC data.

8 Conclusion

Because of the questionable representativeness of the top incomes in regular household surveys, the estimation of the top income distribution turns out to be challenging. The lack of consensus on empirical practices among researchers further highlights the difficulties. Keeping these shortcomings in mind, this paper sets out to review the definitions and the data sources that are available in the existing literature.

After comparing different estimation methods, this paper indicates that collecting income or wealth information of the top incomes from the public media and other available sources and then matching the mixed data sources together could be a new approach to address under-representativeness and under-reporting issues for household survey data.

As a new attempt to correct for the underestimated income inequality is practiced in this paper, we employ different data sources from the public to collect all types of top incomes in different industries and fields. On the basis of the considerable job of data collection, we clean and improve the data quality in different ways, and then, we merge the data into one whole dataset with interpolation and weighting. Finally, the Pareto function is used to estimate and test the sensitivity of the distribution parameter of the collected top income data. We then obtain the final top income dataset, which is called the "Top Incomes in China". It is the first dataset of top incomes in China.

The first finding of the paper is that the income level among the top incomes is extremely high, much higher than that obtained from the household survey data. This result implies that the current household surveys have missed capturing samples of the top incomes, leading to underestimating the quantity of the national wealth to a large extent. In particular, we reasonably expect that the extent of income underestimation is increasing with the trend of increased under-representativeness and under-reporting for the top incomes.

The second finding is that the omitted top incomes by household surveys are also unevenly distributed, as income inequality among the top incomes is large, while the Gini coefficient of the top incomes reaches as high as 0.497 for 2016. It is worth mentioning that the Gini coefficient for the household survey was only 0.464 in the same year.

The third finding comes from our observation of the regional, gender and educational distributions of the top incomes. We find that well-educated males are the main composition of the top incomes in China. Most of the top incomes worked in the coastal provinces in 2016, although some of them were born in the central or western part of China.

The fourth finding is the large impact of the high and top incomes on the overall income distribution in China. If we combine the top income data and household survey data together, the Gini coefficient of income inequality is approximately 0.646 for 2016, which is much higher than that estimated using only the household survey data.

Finally, we realize that the data included in the “Top Incomes in China” need to be improved. A panel of data covering the top incomes is also needed. These issues will be key aspects of our future studies on top incomes in China.

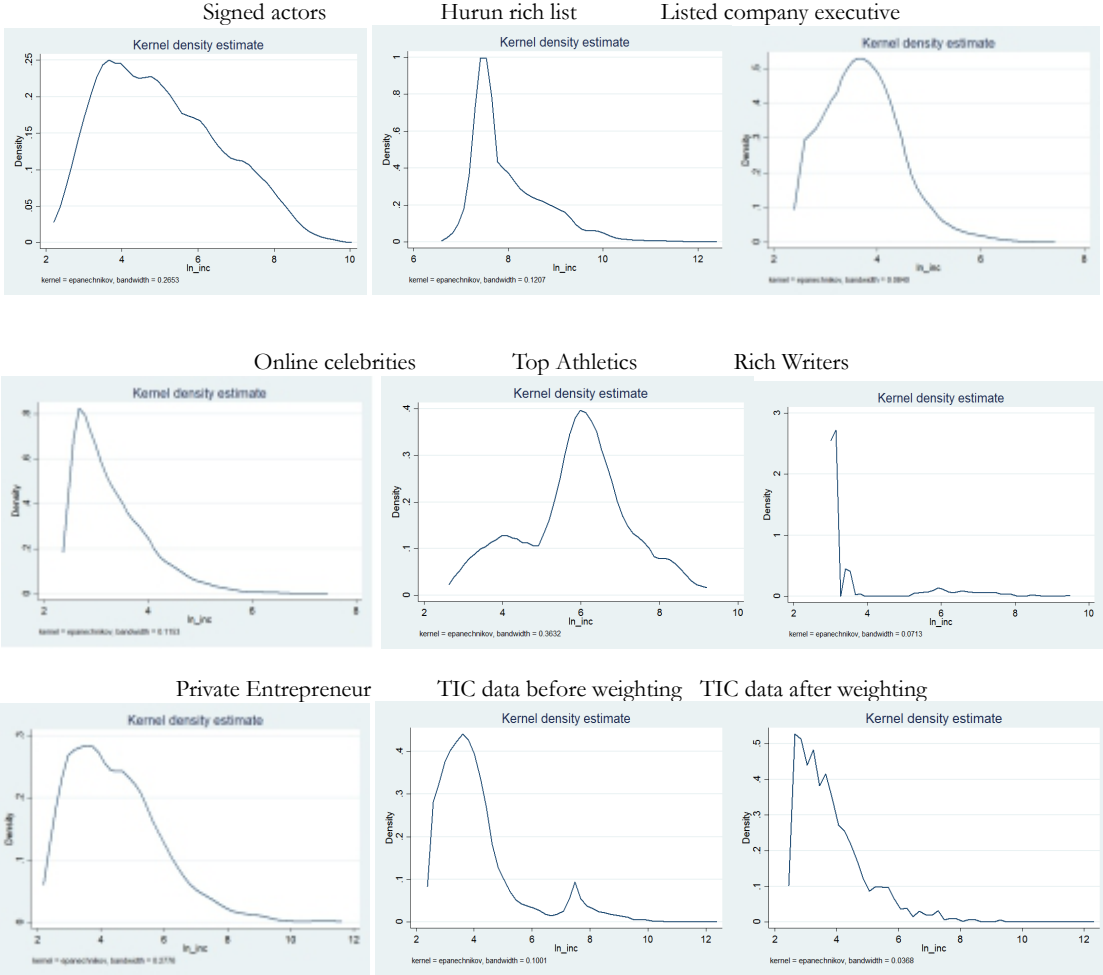
References:

- Alvaredo F., L. Chancel, T. Piketty, E. Saez, and G. Zucman (2016). "Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world". WID.world Working Paper 2016/02. World Inequality Database.
- Atkinson A.B. (2018). "Wealth and Inheritance in Britain from 1896 to the Present". . *The Journal of Economic Inequality*, 16(2): 137-169. <https://doi.org/10.1007/s10888-018-9382-1>.
- Brenner M. (2001). "Re-examining the Distribution of Wealth in Rural China." In C. Riskin, Z. Renwei, and S. Li. (eds), *China's Retreat from Equality: Income Distribution and Economic Transformation*. Armonk, NY:: M.E. Sharpe.
- Chen Y., Z. Huo, and J. Chen (2009). "Disaster Risk and Wealth Distribution of Chinese Urban Residents." *Economic Research Journal*, 11 (in Chinese).
- Credit Suisse Research Institute (2010). Global Wealth Data Book. www.credit.suisse.com.
- Davies J. et al. (2010). "The Level and Distribution of Global Household Wealth". *The Economic Journal*, 121, 223-254.
- Gan L. (2012). *Research Report of China Household Finance Survey*. Chengdu: Southwest University of Finance and Economics Press.
- Garbinti B., J. Goupille, and T. Piketty (2017). "Income Inequality in France, 1900-2014: Evidence from Distributional National Accounts (DINA)", WID.world Working Paper 2017/04. World Inequality Database.
- Knight J. (2014). "Inequality in China: An Overview". *World Bank Research Observer*, 29(1): 1–19.
- Knight J, S. Li, and H. Wan (2016). "The Increasing Inequality of Wealth in China, 2002-2013". Working Papers. Oxford: Oxford University.
- Li S., H. Sato and T. Sicular (2013). "Rising inequality in China: challenges to a harmonious society". Cambridge: Cambridge University Press.
- Li S., C. Luo (2011). "How Unequal Is China?" *Economic Research Journal*, 4 (in Chinese).
- Li S., and H. Wan (2015). "Evolution of wealth inequality in China". *China Economic Journal*, 2015,8(3): 264-287.
- Li S., and R. Zhao (2007). "Changes in the Distribution of Wealth in China 1995-2002." WIDER Working Paper. Helsinki: UNU-WIDER.
- Luo C., X. Yue, and S. Li (2011). "Question on Gray income estimation in China". *Comparative Studies*, vol 1.

- Meng X.(2007). "Wealth Accumulation & Distribution in Urban China". *Economic Development and Cultural Change*, 55, pp.761-791.
- Muellbauer J. (2007). "Housing and Personal Wealth in a Global Context". WIDER Working Paper 2007/27. Helsinki: UNU-WIDER.
- NBS (2013). The national Gini coefficient in China between 2003 and 2012 was released to the public. Beijing: National Bureau of Statistics of China. <http://news.sina.com.cn/c/2013-01-18/131726067826.shtml>.
- NBS (National Bureau of Statistics) (2011). *China Statistical Yearbook*. Beijing: China Statistics Press.
- Piketty T. (2014). *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.
- Piketty T., Y. Li, G. Zucman (2017). "Capital accumulation, private property, and inequality in China, 1978-2015". NBER working paper, 19 July 2017. Cambridge, MA: National Bureau of Economic Research.
- RCPE-CASS(Research Center for Private Entrepreneurs, Chinese Academy of Social Sciences)(2017). Chinese Private Enterprise Survey Report(2016). Beijing: China Association of Industry and Commerce Press.
- Torche F., and S. Spilerman (2008). "Household Wealth in Latin America". In J. Davies (ed.), *Personal Wealth from a Global Perspective*. Oxford: Oxford University Press.
- Wang H., K. Zhou (2006). "Does the income inequality is underestimated: based on the pareto distribution test". *Statistical Research*, 2006(04):8-15.
- Wang X. (2010). "Gray income and national income distribution". *Comparative Studies*, vol 48.
- Xie Y., and Y. Jin (2014). "Household Wealth." In *China Family Panel Studies 2014*". Peking University (in Chinese).
- Xie Y., Z. Xiang (2014). "Income Inequality in Today's China". *Proceedings of the National Academy of Sciences*, 19, 6928-6933.

Appendix 1 The kernel density function with TIC data

Figure appendix 1: kernel density function with TIC data



Source: authors' calculation based on the TIC data.

Appendix 2 The process of estimating the Gini index

Among them, $p_1 = 1293409000/1303757380$ represents the proportion of the population with its income less than 0.65 million in CHIPS to the total population in the mixed data;

$u_1 = 2.803431$ represents the average income if income is less than 0.65 million in the CHIPS;

$p_2 = 10348380/1303757380$ represents the proportion of the population whose income is larger than 1.87 million to the total population in the mixed data;

$u_2 = 193.158926$ represents income per capita among the groups whose income was greater than 0.65 million, we get it from the formula of pareto function $u_2 = X_0 \cdot \alpha / (\alpha - 1)$, among which the alpha is equal to 1.5060362;

The indicator of $u = (1293409000 \cdot u_1 + 10348380 \cdot u_2) / 1303757380 = 3.7490919$ actually represents the average income in the mixed data;

At last, the $G_1 = 0.46434$ represents the Gini coefficient of people with income less than 0.65 million, and we can calculate it based on the command of *ainequal* in the STATA; $G_2 = 0.497405$ represents the Gini coefficient of the intervals that large than 0.65 million.

Appendix 3 Robust check using different return rates of capital in Hurun list

Here we assume different return rates of the wealth along different wealth intervals for the Hurun rich list. Generally, we separate the Hurun list into five deciles, and assume the return rates increases from 5% to 9% with the increase of the wealth level. At this time, the threshold of the pareto function in the TIC data is about 7.8 million.

Moreover, we also estimate the Gini coefficient for the mixed data if using different return rates in the Hurun list. And we find that the new estimated Gini coefficient would be 0.681, which also don't change too much compared the 0.646 before. Therefore, using other return rate of the wealth would not change our basic results.

Table appendix 3a: Searching for different thresholds (Unit: 10 ten thousand)

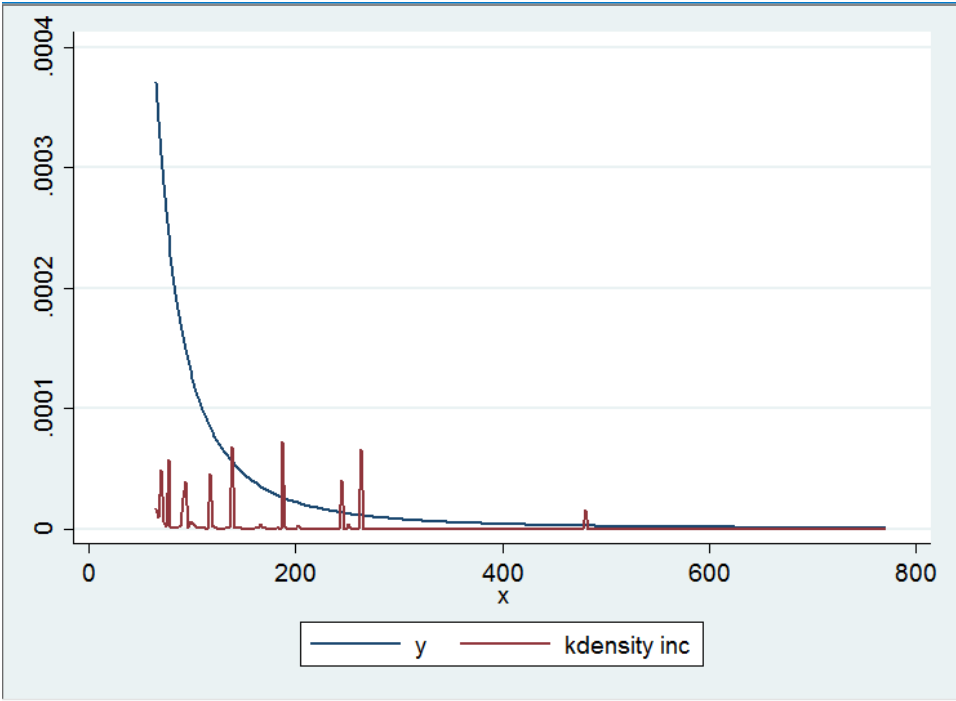
Iteration point	Actual Gini based on the OLS regression	Ideal Gini based on the Lortenz curve	GINI Gap between two approaches
700	0.760	0.498	0.26249
710	0.760	0.498	0.26238
720	0.760	0.498	0.26235
730	0.759	0.498	0.26223
740	0.759	0.498	0.26221
750	0.757	0.496	0.26187
760	0.757	0.495	0.26185
770	0.757	0.495	0.26185
780	0.756	0.495	0.26094
790	0.755	0.493	0.26227
800	0.755	0.493	0.26229

Source: authors' calculation based on the TIC data.

Appendix 4 Comparison between actual and simulated distribution for the overlapped income intervals

According to the figure below, we find that the actual density function is much higher than the simulated function for the income intervals from 0.65 to 7.7 million yuan. In other words, the actual samples don't satisfy the pareto distribution, which again demonstrates that the representativeness of the original TIC data for the second largest top incomes is not as good as expected. Therefore, it is necessary to simulate the samples for the income interval from 0.65 to 7.7 million.

Figure appendix 5: Comparison between actual and simulated distribution



Source: authors' calculation based on the TIC and CHIP data.