Title
Entropy-based China income distributions and inequality measures

Journal
China Economic Journal, 12(3)

ISSN
1753-8963

Authors
Fu, Qiuzi
Villas-Boas, Sofia B
Judge, George

Publication Date
2019-09-02

DOI
10.1080/17538963.2019.1570620

Peer reviewed

China Economic Journal

ISSN: 1753-8963 (Print) 1753-8971 (Online) Journal homepage: https://www.tandfonline.com/loi/rcej20

Entropy-based China income distributions and inequality measures

Qiuzi Fu, Sofia B. Villas-Boas & George Judge

To cite this article: Qiuzi Fu, Sofia B. Villas-Boas & George Judge (2019): Entropy-based China income distributions and inequality measures, China Economic Journal, DOI: 10.1080/17538963.2019.1570620

To link to this article: https://doi.org/10.1080/17538963.2019.1570620

Published online: 05 Feb 2019.

CHINA ECONOMIC JOURNAL
https://doi.org/10.1080/17538963.2019.1570620

Entropy-based China income distributions and inequality measures

Qiuzi Fu[a], Sofia B. Villas-Boas[b] and George Judge[c]

[a]International Department, the People's Bank of China, Beijing, China; [b]Agricultural and Resource Economics, University of California, Berkeley, CA, USA; [c]Graduate School and Giannini Foundation, University of California Berkeley, Berkeley, CA, USA

ABSTRACT
We use information theoretic information recovery methods, on a 2005 sample of household income data from the Chinese InterCensus, to estimate the income distribution for China and each of its 31 provinces and to obtain corresponding measures of income inequality. Using entropy divergence methods, we seek a probability density function solution that is as close to a uniform probability distribution of income (with the least inequality), as the data will permit. These entropy measures of income inequality reflect how the allocation and distribution systems are performing, and we show the advantages of investigating province variation in income inequality using entropy measures rather than Gini coefficients. Finally, we use a sample of data from the China Family Panel Study to recover an estimate of the 2010 and the 2016 to investigate possible directions of inequality changes using these different additional data sources, given that the 2015 Inter-Census is not yet available.

ARTICLE HISTORY
Received 7 June 2018
Accepted 14 January 2019

KEYWORDS
Income probability distribution function; micro-income data; information theoretic methods; cressie-read divergence; entropy maximization; Pareto's law

JEL
D31; E21; C1; C10; C24

1. Introduction

In this paper, information theoretic entropy-based divergence methods are used to estimate and investigate the heterogeneity of income probability density functions (PDFs) across Chinese provinces using micro-household income data from a sample of the Chinese 2005 Inter-Census. Given the estimated PDF's we recover the corresponding implied entropy measures of income inequality. In the entropy criterion-measure, we seek a probability density function solution that is as close to a uniform PDF-distribution of income (an equal distribution with the least inequality), as the micro-sample data will permit.

In economic behavioral systems, markets provide a basis for processing information and determining the value of the components in the income portfolio. At the economic unit-country level, the income probability density function-distribution contains information on how the market is functioning, how the allocation and distribution system is performing, and in terms of dynamics, how the economic system has changed over time. Like the behavioral system in other developed countries, the distribution of China's income or wealth that results

CONTACT Sofia B. Villas-Boas sberto@berkeley.edu Agricultural and Resource Economics, University of California, Berkeley, CA 94720

is based on a very complex set of market interactions, governmental policies and interventions, and changes that are institutional in nature.

As a basis for estimating, ordering, and determining the informational content of countrywide income PDFs, likelihood is the common loss function used in econometric information recovery. However, the optimality of a given likelihood method is fragile and unstable with respect to estimation and inference under model uncertainty. In addition, the precise functional representation of the data sampling process cannot usually be justified from economic-behavioral theory. Given this situation, a natural solution is to use estimation and inference methods that are designed to deal with systems that are fundamentally ill-posed and stochastic in nature where uncertainty and random behavior are basic to information recovery. Therefore, instead of traditional information recovery methods, we use informational theoretic entropy based econometric methods to analyze the 2005 sample of Chinese micro-income data. In this entropy approach to information recovery, moment constraints provide a basis for representing our knowledge of the microeconomic behavioral system in terms of a probability density function-distribution. Within this framework, breaking up the data by provinces, we obtain provincial level PDF's and a unique province-based entropy measure of income inequality results and illustrate its advantage relative to traditional measures such as the Gini coefficient. Finally, by comparing the 2005 income distribution with 2010 and 2016 based income distributions from an additional income survey data source of a smaller sample size, given that the inter-census from 2015 is not available, we also investigate the way income inequality has evolved in China.

The study of income distribution recovery has a long history. In the search for a universal regularity in income and wealth, Pareto (1896) originally suggested an exponential-power law income distribution to describe the allocation of wealth among individuals, and to demonstrate that a larger portion of the wealth of any society is owned by a small percentage of the people. We contribute to this literature by presenting a uniform entropy based behavior-related method that simultaneously recovers country-based income PDFs under uncertainty and the corresponding income equality-inequality measure, from samples of micro-income data in China. In a related paper, Villas-Boas, Fu, and Judge (2017) focus on European repeated cross-sections of micro-income data over several years, by recovering country-based probability density-income distribution functions and measuring the nature of income inequality across European countries and over time. We feel this information recovery approach provides a measure of disorder and thus offers a unique way to analyze Chinese micro-income data. From an empirical standpoint, the availability of Chinese data series that consists of millions of household-province level observations allows us to pursue some of these questions and to use entropy information recovery methods to recover income probability distributions for China and each of its provinces and obtain measures of income inequality.

There is a rich literature relative to studies of China's income distribution, inequality and wealth. Many of the studies note that although China is second in the world in terms of GDP, serious problems exist in terms of income inequality. In a recent article, Han, Zhao, and Zhang (2015), discuss China's income inequality in a global context and discusses inequality issues between China and European countries. In the following, we note a few studies in this rich literature. Xie and Xiang (2014), note that China's income inequality has reached the high levels of a Gini coefficient of 0.55. They note

that China's income inequality is due to regional disparities and the rural-urban gap. Xie and Yongai (2015), use data from the China Family Panel Study to investigate the level and distribution of household wealth in China and find a Gini coefficient of 0.73 in 2012. They note the urban-rural divide in the wealth distribution. Zhou and Song (2016), note rapid growth in China has been accompanied with the rapid increase in income inequality. They highlight the high return to capital relative to wage incomes. Knight (2017), notes that inequality in China is evolving in different directions. He notes that it appears that income inequality is now falling, whereas wealth inequality continues s to rise. His analyses are based on surveys conducted in 2002, 2007 and 2013.

The paper proceeds as follows. In section 2 we argue that the probability distribution of income is given by the possible ways in which a collection of non-interacting micro-incomes may occupy a set of discrete income states. Consistent with the micro-sample data generation process, we suggest an information theoretic entropy-based method (see Judge and Mittelhammer 2012a, 2012b) to recover the unknown probability density-distribution functions. The micro-level Chinese income data at the Country and province level is described in Section 3. In Section 4, we indicate how economic behavior and the nature of income inequality is captured by the entropy-based distribution and inequality measures. In the empirical information recovery process of Section 5, we recognize the connection between adaptive intelligent behavior and causal entropy maximization in a self-organized open economic behavioral system, and use the Read and Cressie (1988) family of divergent measures to determine the nature of the distribution of the Chinese micro-income data and corresponding inequality measures for the year 2005. In section 6 we recover income PDFs and measures of inequality for each of the 31 provinces and show the advantages of investigating province level variation in income inequality using entropy measures rather than traditional measures like the Gini coefficient. In section 7 we use 2010 and 2016 micro-data sample to recover an estimate of the income PDF for China and a measure of inequality over time. In section 8 we comment on the income distribution-inequality results and look to an analysis of the forthcoming 2015 sample of micro-income data.

2. An entropy-information recovery framework

As we focus on the recovery of the underlying (PDFs) from micro-income data, we recognize and emphasize the behavior related nature of the observations. This means we recognize that like prices, incomes do not behave, but that people behave. Thus, a country's income distribution is one way to exhibit and summarize economic behavior and its allocation and distribution performance. In recovering information regarding the unknown parameters of a micro-based behavioral income distribution system, we follow Wissner-Gross and Freer (2013) and recognize the connection between adaptive intelligent behavior, causal entropy maximization and self-organized equilibrium-seeking behavior in an open dynamic economic system. Under this optimizing criterion, each microstate can be seen as a causal consequence of the macro-income state to which it belongs. This connection between causal adaptive behavior and entropy maximization, which is based on a causal generalization of entropic forces, suggests that economic social systems do not evolve in a deterministic or a random way,

but tend to adapt behavior in line with an optimizing principle. In the sections ahead, we exhibit new ways to think about income information recovery and the causal adaptive behavior of large complex microeconomic systems, and the use of entropy as the system's status inequality measure.

2.1 *Problem formulation and solution*

In this paper, we use a particular form of information theoretic methods on the problem of establishing a databased link to recover income probability density functions-distributions from samples of micro-income data. In recovering the income probability density function-distribution from a sample of N positive real China-based income numbers, we assume the income probability to be represented by partitions-histograms that span the income sample space and count the relative frequencies of the incomes within each subinterval. Samples from these partitions yield histogram outcomes of the discrete random income variable d_j, for $j = 1, 2, \ldots, n$, and under repeated observation, one of n histograms-micro-configurations associated with the macro-state income is observed with probability p_j. Further, suppose after a large number of trials, we have first-moment sample information in the form of the mean value of a country's income:

$$\sum_{j=1}^{n} d_j p_j = \bar{d}. \tag{2.1}$$

Given this first-moment sample information and the inverse problem of identifying an income distribution from the sample income data, we seek the best predictions of the unknown probabilities $p_1, p_2, \ldots, p_n$. It is readily apparent that there is one data point $\bar{d}$, and n unknown p_i. From an information recovery standpoint, there is an infinite number of possible discrete probability distributions with $\bar{d} \in [1, n]$. Based only on the information $\sum_{j=1}^{n} d_j p_j = \bar{d}$, $\sum_{j=1}^{n} p_j = 1$, and $0 \leq p_j \leq 1$, the problem cannot be solved for a unique solution. Thus, a function must be inferred from insufficient information when only a feasible set of solutions is specified. In such a situation, it is useful to have an approach that allows the investigator to use sample-based information recovery methods without having to choose a parametric family of probability densities on which to base the income function. If we replace the expectation value of the macro-variable income by its most likely value, this is equivalent to maximizing entropy with respect to the macro-state. This is the problem to which we now turn.

2.2. *The information theoretic family*

In this section, we specify information theoretic entropy-based methods to provide a natural basis for establishing a causal influence-econometric-inferential link to the data and solving the resulting ill-posed stochastic inverse problem. In this ill-posed inverse problem context, we face a system which may have more unknowns than data points, or in which the solutions depend discontinuously upon the initial data. This type of uncertainty, regarding the economic-econometric model, the associated

estimating equations and the data sampling-probability distribution function, create unsolved problems with respect to information recovery. As noted earlier, although the likelihood is a common loss function used in fitting econometric models, the optimality of a given likelihood method is fragile inference-wise under model uncertainty. In addition, the precise functional representation of the data sampling process cannot be justified using economic-behavioral theory. Given this situation, a natural solution is to use estimation and inference methods that are designed to deal with systems that are fundamentally ill-posed and stochastic in nature and uncertainty and random behavior are basic to information recovery. To identify estimation and inference measures that represent a way to link the model of the process to a family of possible likelihood functions associated with the income data, we use the Cressie and Read (1984) and Read and Cressie (1988) single parameter CR family of entropic function-power divergence measures given by:

$$I(\mathbf{p}, \mathbf{q}, \gamma) = \frac{1}{\gamma(\gamma+1)} \sum_{i=1}^{n} p_i \left[\left(\frac{p_i}{q_i} \right)^{\gamma} - 1 \right] \tag{2.2}$$

In (2.2), γ is a parameter that indexes members of the CR-entropy family of divergence measures-distributions, $p_i{}'^{s}$ represent the subject probabilities and $q_i's$ are usually interpreted as uniform reference probabilities but they may also represent any prior pre-data information. Being probabilities, the usual probability distribution characteristics of $p_i, q_i \in [0,1] \forall i$, $\sum_{i=1}^{n} p_i = 1$, and $\sum_{i=1}^{n} q_i = 1$ are assumed to hold. In (2.2), as γvaries, the resulting CR-entropy statistical family of estimators that minimize power divergence, exhibit qualitatively different sampling behaviors that includes Shannon's entropy, the Kullback-Leibler measure and in general a range of independent (additive) and correlated systems (see Gorban, Gorban, and Judge (2010), Judge and Mittelhammer (2012a) and (Judge and Mittelhammer 2012b)). In identifying the probability space, the CR family of power divergences is defined through a class of additive convex functions and the CR power divergence measure leads to *a broad family of likelihood functions* and test statistics. The CR measure exhibits proper convexity in **p**, for all values of γ and **q**, and embodies the required probability system characteristics of additivity and invariance with respect to monotonic transformations of the divergence measures following Gorban, Gorban, and Judge (2010).

In the context of extremum metrics, the general CR family of power divergence statistics represents a flexible family of pseudo-distance measures from which to recover the joint distribution probabilities and encompasses a wide array of empirical goodness-of-fit and information recovery criteria. As γ varies, power law Pareto behavior is efficiently described and the resulting estimators that minimize power divergence exhibit qualitatively different sampling behaviors. To place the CR family of power divergence statistics in an entropy perspective, we note, following Gorban, Gorban, and Judge (2010), that there are corresponding families of entropy functions–divergence measures. Over defined ranges of the divergence measures, the CR and entropy families are equivalent.

If in the CR family of entropy functions, in the limit as $\gamma \to 0$, the solution of the first-order condition leads to Shannon maximum entropy and the logistic expression for the conditional probabilities. Alternatively, if in the family of CR entropy

functional in the limit $\gamma \rightarrow -1$, the solution of the first-order condition leads to the maximum empirical likelihood distribution for the conditional probabilities. These alternative entropy functions permit us to use the information content of the structure of the income-probability density function-distribution as a measure of inequality. The income statistical system may be characterized by a macro-state, for which many potentially compatible micro-configurations exist, that are compatible with it. To obtain the PFD of the income distribution we use the principle of maximizing the CR-entropy functional, subject to the constraints (2.1) to identify the most likely distribution function-histograms for a given economic statistical system. In this context, recovering the income distribution from a sample of positive real numbers through the use of the CR-entropy criterion (2.2), suggests we seek a solution to the following extremum problem:

$$\hat{\boldsymbol{p}} = \arg\min_p \left[I(\boldsymbol{p}, \boldsymbol{q}, \gamma) \mid \sum_{j=1}^{n} p_j d_j = \bar{d}, \sum_{j=1}^{n} p_j = 1, p_j \geq 0 \right]. \tag{2.3}$$

Solving this optimization problem provides a solution to the income probability distribution function and to the entropy inequality measure. In general, the solution to this extremum problem does not have a closed-form expression and the optimal values of the unknown network parameters must be numerically determined.

3. The Chinese micro-household income data

The data used in this paper originate from the China's inter-census Survey from 2005. The inter-census survey represents 1% of the population and is conducted every 10 years for years ending in 5. The sample contains over one million observations on personal characteristics and income data among the population of current residence. A full population census happens every 10 years at the year ended by 0 and an inter-census population survey, also called 1% population sample survey, is conducted between two full censuses at the year ended by 5. Except for the 1% coverage of population, the inter-census survey is almost the same as the full census in terms of organization, procedures, and questionnaire design. The survey covers all the 2861 counties of China and is representative of the 333 prefectures using a three-stage cluster sampling method probability proportional to estimated size (PPES) where the minimal sampling unit is an enumeration district (equivalent to a resident group of village committee in rural areas and equivalent to a neighborhood committee in urban areas). The variable we extract form the survey for our analysis consists of recent monthly income for sampled individuals. Income data are obtained from questions about last-week employment, including work time, industries, occupation, recent monthly income, and employment status. Information from all 31 provinces is included in Table 1.

China's inter-census Survey includes sample in all 31 provinces. The number of observations is in proportion to the population of each province at that time, and ranges from 9,881 in Tibet to 191,729 in Guangdong. Mean monthly income varies greatly among provinces as well. People in Beijing had the highest average monthly income. At the same time, there are six provinces with mean income lower than 400.

Table 1. Information on the data from the 31 provinces.

Province	Obs	Mean income	Std.Dev.	Min	Max
Anhui	44,490	474.45	486.98	1	20,000
Beijing	21,324	1605.31	1554.07	20	50,000
Chongqing	32,512	460.34	527.30	1	20,000
Fujian	27,353	775.01	755.16	1	22,000
Gansu	51,540	375.04	421.75	7	10,000
Guangdong	191,729	889.86	1136.74	5	99,000
Guangxi	37,079	403.86	436.60	5	30,000
Guizhou	34,330	375.89	420.89	5	10,000
Hainan	17,935	504.66	521.51	10	10,000
Hebei	47,149	489.37	474.29	7	10,000
Heilongjiang	31,322	577.16	461.47	16	10,000
Henan	51,461	359.96	355.73	5	10,000
Hubei	49,312	484.21	451.90	10	10,000
Hunan	50,412	523.36	517.92	5	30,000
Inner Mongolia	26,820	681.81	655.78	6	20,000
Jiangsu	49,692	805.85	918.24	10	90,000
Jiangxi	31,771	509.33	453.64	1	10,000
Jilin	34,157	562.31	437.86	7	10,000
Liaoning	34,599	606.66	643.72	10	30,000
Ningxia	13,406	540.06	551.15	10	18,000
Qinghai	17,095	444.26	530.59	10	12,000
Shandong	74,477	556.22	560.05	1	30,000
Shanghai	40,175	1559.23	1536.27	20	50,000
Shannxi	61,601	408.49	433.04	2	10,000
Shanxi	55,664	576.75	543.95	2	12,500
Sichuan	54,105	398.51	443.74	5	20,000
Tianjin	38,446	1049.26	701.91	2	10,000
Tibet	9881	308.15	462.06	2	10,000
Xinjiang	22,606	541.75	554.39	5	10,000
Yunnan	92,993	330.10	535.23	4	80,000
Zhejiang	36,819	1038.26	1072.06	1	50,000

4. An information theoretic data income analysis

We focus on the analysis of samples of Chinese micro-household income data for 2005. In the analysis of the income data, we make use of the information theoretic methods of Section 2 as a basis for summarizing the income data in the form of a probability density function, and use entropy as an income distribution measure of inequality. We use income levels-histograms to span the micro-sample spaces to investigate the patterns of income distributions over the provinces, and make comparisons in terms of income inequality based on entropy measures.

4.1. *Maximum empirical exponential likelihood (MEEL) formulation*

Two information-theoretic variants of the CR-entropy $I(p, q, \gamma)$ discrepancy-distance measure are prominent in the literature. The choice of γ is concerned with a measure of uncertainty about the realization of the micro-sample data. Out of all the distributions consistent within the constraint set (2.1), we choose in the limit $\gamma \to 0$, which leads to the distribution that can be achieved in the greatest number of distinctive ways (Judge and Mittelhammer 2012a). Thus, in the limit $\gamma \to 0$, the Maximum Exponential Empirical Likelihood (MEEL), is the most likely distribution to be observed from

a statistical or combinatorial point of view. It is also the most appropriate measure of effective support size.

Consistent with the discussion of identifying the micro-configurations compatible with the income macro-state, for discussion purposes assume that K income histogram levels are chosen to span the range of the micro-income data space. Consequently, in analyzing the Inter-Census sample of income data we use the CR-MEEL in the limit criterion, $\gamma \to 0$, a uniform reference distribution q $(q_j = \frac{1}{n}, \forall j)$, K income-histogram levels, and first-moment information $\sum_{j=1}^{K} d_j p_j = \bar{d}$, as a basis for recovering discrete income probability density function-distributions. Under this specification, when in the limit $\gamma \to 0$, the CR $I(p, q, \gamma)$ converges to an estimation criterion equivalent to the maximum exponential empirical likelihood (MEEL) metric $H(p) = -\sum_{j=1}^{K} p_j \ln(p_j)$. Our extremum problem likelihood-entropy function may then be formulated as:

$$\max_{p}\left[-\sum_{j=1}^{K} p_j \ln p_j \middle| \sum_{j=1}^{K} p_j d_j = \bar{d}, \sum_{j=1}^{K} p_j = 1, p>0\right]. \tag{4.1}$$

The corresponding Lagrange function-extremum problem is:

$$\begin{aligned} L(\boldsymbol{p}, \eta, \lambda) \equiv & -\sum_{j=1}^{K} p_j \ln(p_j) + \lambda\left(\bar{d} - \sum_{j=1}^{K} p_j d_j\right) \\ & + \eta\left(1 - \sum_{j=1}^{K} p_j\right). \end{aligned} \tag{4.2}$$

Solving the first-order conditions yields the exponential result:

$$\hat{p}_j = \frac{exp\ (-d_j \hat{\lambda})}{\sum_{j=1}^{K}\ exp\ (-d_j \hat{\lambda})} \tag{4.3}$$

for the jth income outcome and the mean-related income distribution. As the mean of income varies over a range of micro-data sets, a family of exponential distributions results. In Equation (4.3), $\hat{p}_j$ is a function of $\hat{\lambda}$, the Lagrange multiplier for constraint (4.3). This information may be used as a basis for modifying the distribution of income probabilities.

The CR-MEEL-entropy criterion provides an empirical representation of the joint income probability distribution function, where the p_j are chosen to assign the maximum joint probability among all of the possible probability assignments. Using the CR ($\gamma \to 0$) entropy functional and the mean of the Chinese country's (or the mean for each province) income data, we recover the resulting probability density income distribution.

4.2. *Entropy measure of income inequality*

In order to provide the information that is needed to compare the income distributions, we make use of the entropy measure $E = \gamma \to 0 = \sum_{j=1}^{K} p_j \ln p_j$. In the entropy

criterion-measure, we seek a probability density function solution that is as close to a uniform PDF-distribution of income (an equal distribution with the least inequality), as the micro-sample data will permit.

5. 2005 income PDF for china and income inequality measure

To develop income probability density functions that attempt to capture a statistical measure for China, we use the 2005 data sample, and in line with the information recovery method discussed section 4, we divide the income data for each country into 20 histograms-categories. Using the information recovery methods, we obtain an estimate of the distribution of income in China, given the entropy criterion-measure, that is as close to a uniform PDF-distribution of income (an equal distribution with the least inequality), as the micro-sample data will permit. The resulting PDF is depicted in Figure 1.

Using the information recovery methods discussed in Sections 2 and 4, the distribution of income in China is of an exponential form, and has an entropy inequality measure value $H(p) = -\sum_{j=1}^{K} p_j \ln(p_j)$ of 2.058, indicating a moderate level of income inequality. As an indication of how the Chinese income allocation and distribution system was performing, we note that the first 10 income histograms of the PDF contained 95.3% of the samples of data. In terms of income inequality, the resulting. To understand an entropy inequality measure for China, it is useful to compare this measure to similar inequality measures obtained from micro-survey data for other countries. For instance, in Europe using samples of data from 2008 to 2013, the entropy measure of inequality ranges between 1 for Portugal and 2.5 for Norway (Villas-Boas, Fu, and Judge 2017). Thus, China's entropy inequality measure reflects less income inequality than Portugal but more inequality than Norway and Northern European countries. In particular, it is important to note that the entropy inequality measure H(p) is directly related to the entropy-based PDF.

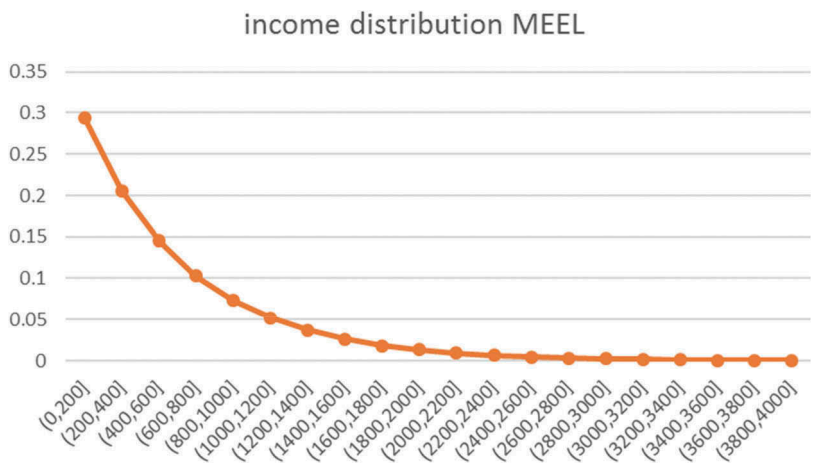

Figure 1. Information theoretic-entropy 2005 China income PDF.

6. Province income PDF and entropy income inequality measures

In this section, we identify residence location for each respondent into each of the 31 provinces in the 2005 data sample, and develop income probability density functions and a statistical measure of income inequality for each Province. For this analysis, we divide the income data for each Province into 10 income histograms-categories and the income PDF-distributions for the 31 Provinces are given in Figure 2.

In Figure 2 all of the estimated PDFs for the different provinces exhibit an exponential shape and differ among the different provinces. The corresponding entropy inequality measure for the 31 Provinces are given in Figure 3 and are arranged in increasing order of the entropy measure (that is in decreasing measure of income inequality).

6.1. *Comparison of Gini coefficients as a basis to study income inequality*

Given that the estimated MEEL income distributions were different as shown in Figure 2, then the corresponding province level entropy measures are different and thus have different implications for discussing income inequality at the province level. This is not guaranteed to be the case if we were to use common measures of income inequality such as the Gini coefficient. construct for the Gini Coefficient measure. In fact, we will show next that even though provinces have different income distributions we would have the same values for the Gini coefficients and then same implications and interpretations in terms of income inequality.

Given the estimated MEEL income distributions by the province in Figure 2 we obtain for each province the Lorenz curve and the resulting Gini coefficients by province as well. We compute the Gini measure based on the MEEL estimated probability income distributions.

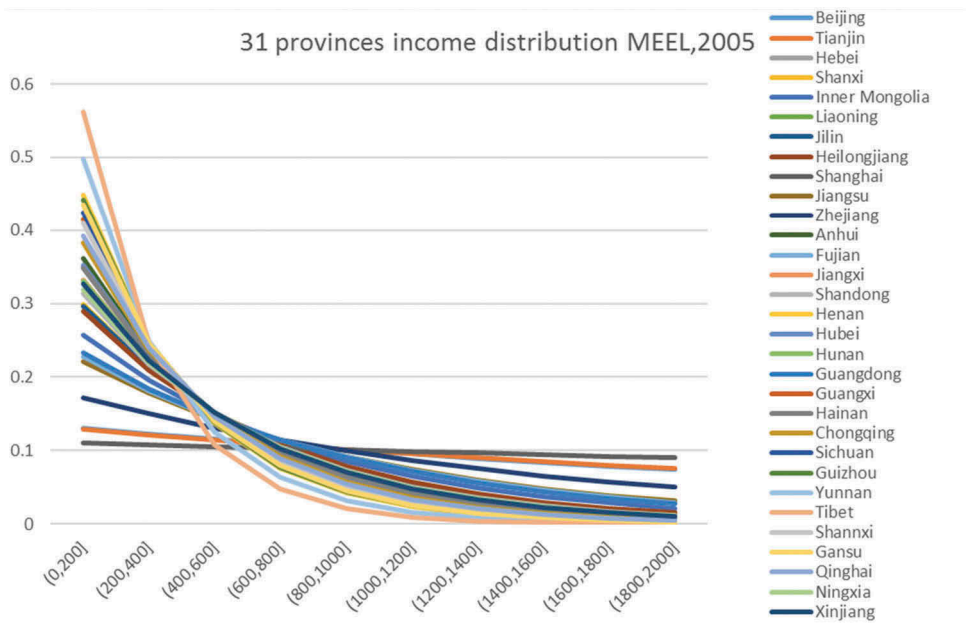

Figure 2. Income distributions for the 31 provinces.

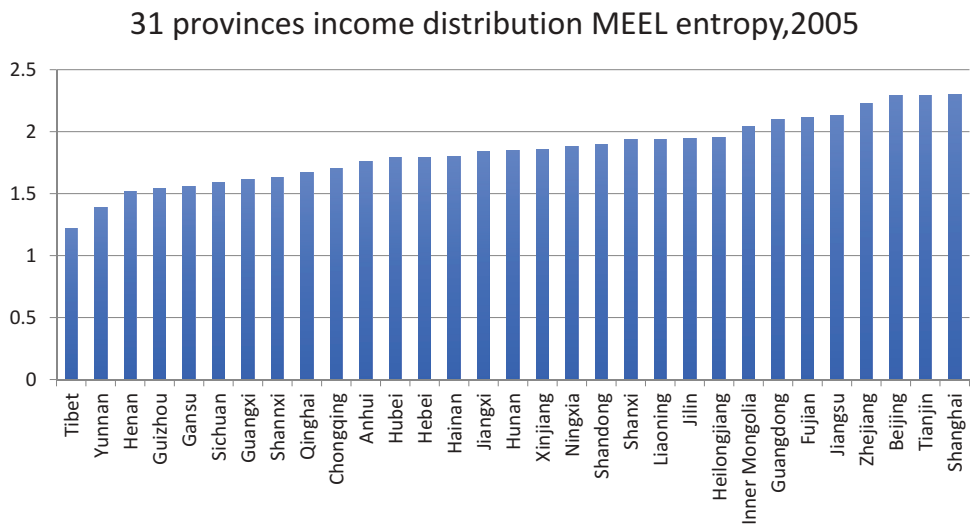

Figure 3 Entropy measures for the 31 provinces.

Then, we compare our province level entropy measure of inequality with the Gini coefficient measures as in Khan and Riskin (2005) using our MEEL pdf estimates. Ranking the 31 provinces by lowest to highest entropy measure as in Figure 3, that is in terms of highest (on the left of the Figure) to lowest (in the right of the Figure) income inequality as measured by the increasing entropy measure, we show in Figure 4 the estimated Gini Coefficients by province.

First, we see that there is less province level variation in Gini measures than in the estimated entropy measures. Second, the ranking of provinces differs for three provinces, namely, Beijing, Tianjin, and Shanghai. Those are the provinces that according to the highest measures of entropy correspond to the lowest income inequality provinces among the 31 provinces in China. However, based on the Gini measure the three are not the ones with the highest Gini coefficients (see also Figure A.1 in Appendix for the overlapped measures by provinces).

The different insights from Gini relative to our approach are quite understandable given the Gini coefficient's main weakness as a measure of income distribution. The main weakness is that it is incapable of differentiating different kinds of inequalities. Lorenz curves may intersect, reflecting differing patterns of income distribution, but nevertheless resulting in very similar Gini coefficient values which is exactly what happens for the provinces of Beijing, Shanghai, and Tianjin. For example, Beijing has a very different income distribution than the province of Heilongjiang, but both provinces have the same Gini coefficient. Similarly, Shanghai and Shandong have the same Gini coefficient but their income distributions could not be more different as we will show in the next Figures of the estimated MEEL income distributions by provinces. This troubling property of the Lorenz framework complicates comparisons of Gini coefficient values and may confound rankings of regions in terms of income inequality based on Gini coefficients.

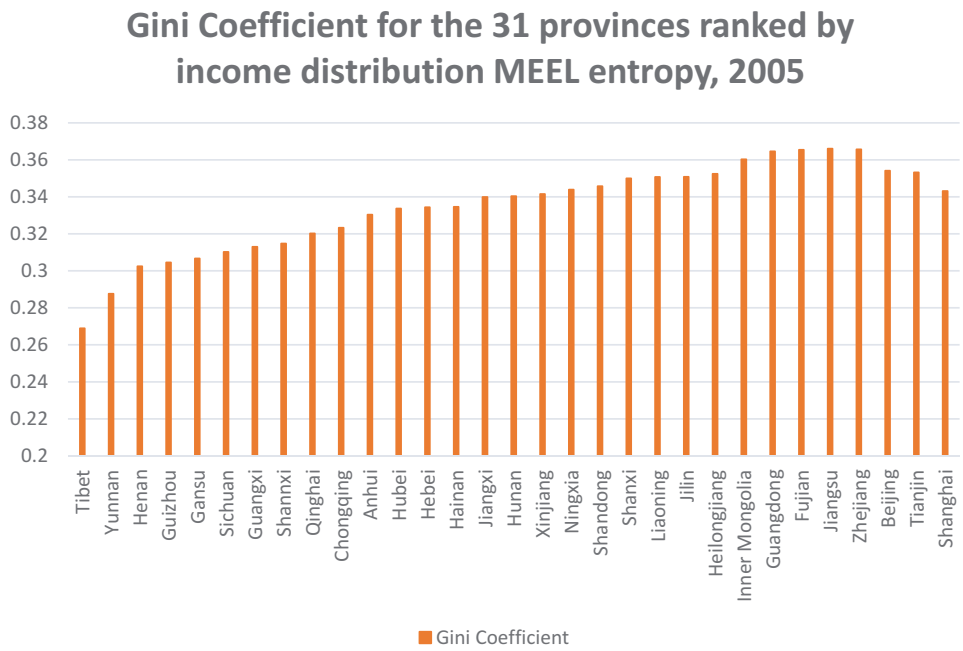

Figure 4 Province level Gini coefficients.

6.2. *Entropy measures as a basis to study income inequality*

In the entropy criterion PDF estimation, we seek a probability density function solution that is as close to a uniform PDF-distribution of income (an equal distribution with the least inequality), as the micro-sample data will permit. Hence, the higher the value of entropy for a given Pdf, the lower income inequality. To investigate the heterogeneity in the PDFs formally we compute the implied entropy inequality measure for each Province. The highest value of the entropy inequality measure is 2.3 for Shanghai, and the lowest measure of income inequality is 1.3 for Tibet. Shanghai, Tianjin, and Beijing are major urban Provinces and all of them are located in the east coastal area of China. These three provinces exhibit the highest entropy values of 2.3 and thus have the lowest levels of income inequality. As we move inland and Westward, the entropy measures are lower and the lowest entropy measure-income inequality Provinces, such as Tibet, Yunnan, and Henan, are depicted in Figure 5 with smaller circles.

To investigate the patterns of income inequality more carefully, we break up the Provinces into groups. When we do this, we find that the resulting Province level PDF's fall broadly into three main types of income probability density distributions and these presented in Figures 6–8.

7. Suggestive evidence-income inequality over the years

In order to investigate patterns of inequality over time, we investigate micro-income data for another year, given that changes within China are affecting the level of economic activity

Figure 5. Highest entropy and lowest entropy provinces in China.

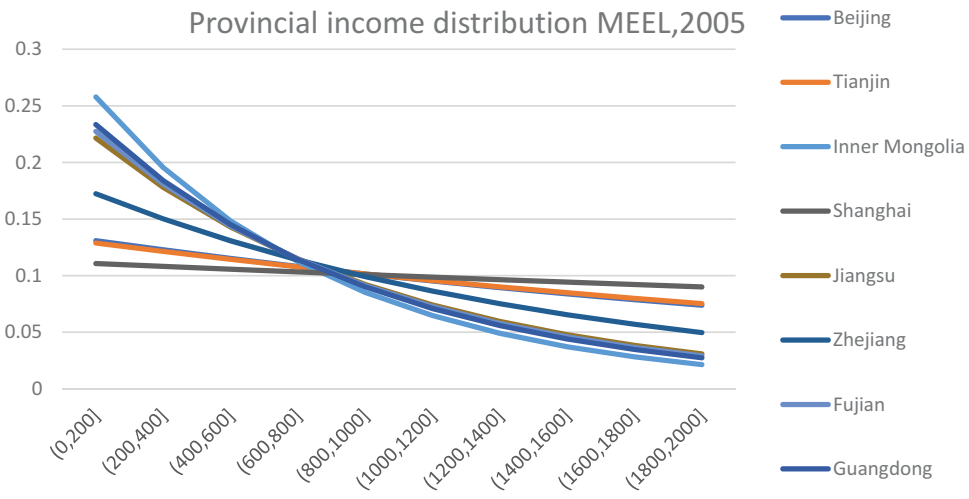

Figure 6. Income PDFs for eight high entropy-low inequality measure provinces.

throughout the world. We make use of the China Family Panel studies (CFPS) for 2010 and also for 2016. The CFPS is a nationally representative, annual longitudinal survey of Chinese communities, families, and individuals launched in 2010 by the Institute of Social Science Survey (ISSS) of Peking University, China. The CFPS is designed to collect individual, family, and community-level longitudinal data in contemporary China (see Knight 2017). For the analysis to follow, we get micro-information from the family roster part that provides monthly salary per capita for each of 13,000 households.

Making use of this sample information and using the information recovery methods of sections 2 and 4, in Figure 9 we combine the 2010 and the 2016 income distributions using the smaller CFPS sample with the inter-Census based 2005 income distribution

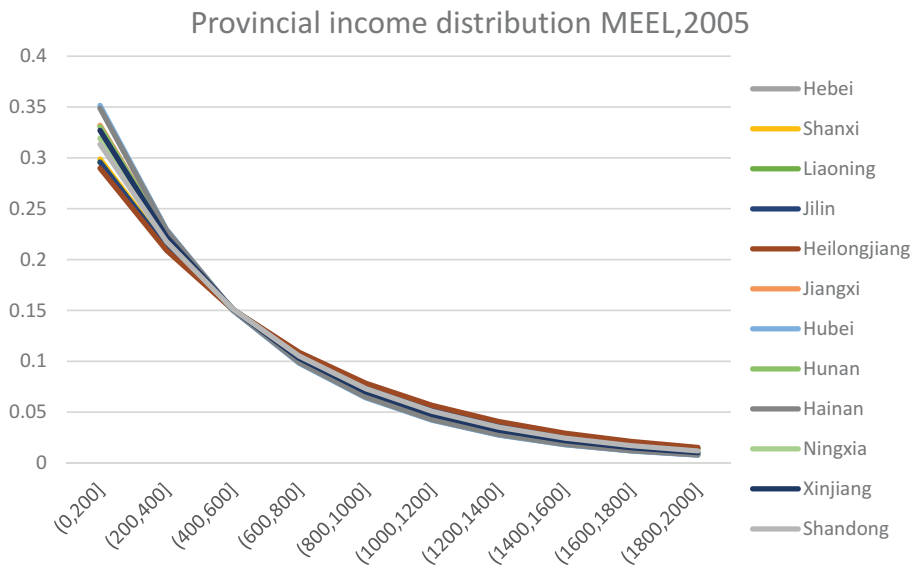

Figure 7. Income PDF's of provinces with middle entropy–inequality measures.

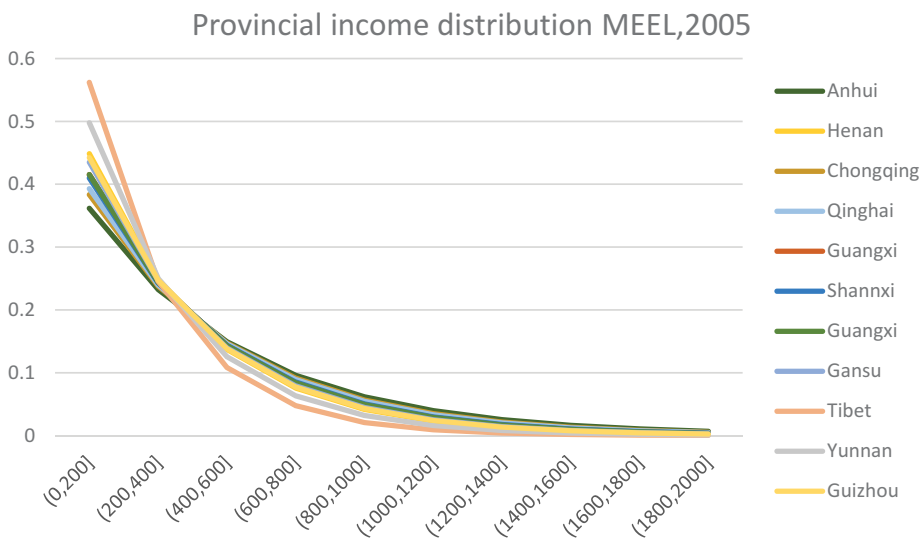

Figure 8. Income PDFs with low entropy measure-high inequality provinces.

developed in Section 6. Once again, we see an exponential shaped estimate for all the estimated distributions. The comparison between 2005, 2010, and 2016 is meant to be suggestive of the direction of change and, when the 2015 census data are available, in the future work we hope to do a new complete analysis.

First, the difference in the distributions and the entropy measure of inequality is quite significant. For the 2016 sample of income, the entropy measure of income inequality is 2.06. Relative to the 2005 inequality measure of 1.35, this indicates a major change toward income equality and more in line with the inequality measures

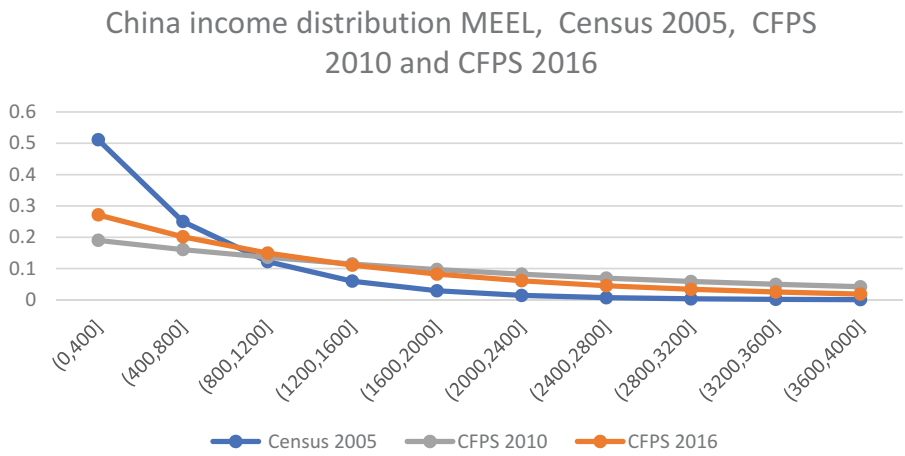

Figure 9. A comparison of the 2005, 2010, and 2016 income distributions.
Note: Entropy Inequality Measure for 2005: 1.35, CFPS 2010: 2.26, CFPS 2016: 2.06.

of Northern European countries. Interestingly for 2010, the measure of inequality is even higher than in 2016, with a value of 2.26. Considering that CFPS only contains more than 10 thousand households from 25 provinces in China and that the size of the 2005 inter-Census data is larger, and contains all the 31 provinces, this comparison is to be seen with caution. When the 2015 China micro-income data becomes available we will be able to identify the validity of the 2016 income distribution estimate and the inequality measure.

8. Concluding remarks

In this paper, we have presented a methodological basis for recovering country-based income probability density functions from samples of Chinese household income survey data. In terms of information recovery, the Cressie-Read family of entropy-based information divergence measures has been used to provide a flexible family of functions to recover the unknown country-based probability density functions-income distributions and yield an integrated entropy measure of income inequality that has less problems than traditional measures such as the Gini coefficient. The estimated income PDFs provide a macro-basis for indicating how the pricing mechanism is performing, how the allocation and distribution system is functioning and how the income distribution may have changed over time. We feel that this paper provides a methodological, empirical and dynamic basis for analyzing the income results from China's 2015 micro-sample of income data.

Acknowledgments

We thank the editor Yan Shen and two anonymous referees for their suggestions. Special thanks are due Maoyong Fan who helped us obtain access to the sample from the 2005 China's Inter-Census income survey and thank Peter Berck for comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Cressie, N., and T. Read. 1984. "Multinomial Goodness of Fit Tests." *Journal of Royal Statistical Society of Series B* 46: 440–464.

Gorban, A., P. Gorban, and G. Judge. 2010. "Entropy: The Markov Ordering Approach." *Entropy* 12: 1145–1193. doi:10.3390/e12051145.

Han, J., Q. Zhao, and M. Zhang. 2015. "China's Income Inequality in a Global Context." *Perpectives in Science* 7: 24–29. doi:10.1016/j.pisc.2015.11.006.

Judge, G. G., and R. C. Mittelhammer. 2012a. *An Information Theoretic Approach To Econometrics*. New York: Cambridge University Press.

Judge, G. G., and R. C. Mittelhammer. 2012b. "Implications of the Cressie-Read Family of Additive Divergences for Information Recovery." *Entropy* 14: 2427–2438. doi:10.3390/e14122427.

Khan, A., and C. Riskin. 2005. "China's Household Income and Its Distribution, 1995 and 2002." *The China Quarterly* 182: 356–384. doi:10.1017/S0305741005000238.

Knight, J. 2017. "China's Evolving Inequality." *Journal of Chinese Economic and Business Studies* 15: 307–323. doi:10.1080/14765284.2017.1317171.

Pareto, V. 1896. *Cours d'Economique Politique*. Geneve: Library Droz. Reprinted 1964.

Read, T.R.C., and N.A.C. Cressie. 1988. *Goodness of Fit Statistics for Discrete Multivariate Data*. New York: Springer Verlag.

Villas-Boas, S. B., Q. Fu, and G. Judge, 2017. "Entropy Based European Income Distributions and Inequality Measures," CUDARE working paper, U. C. Berkeley.

Wissner-Gross, A., and C. E. Freer. 2013. "Causal Entropic Forces." *Physical Review Letters* 110: 168702-1-5. doi:10.1103/PhysRevLett.110.168702.

Xie, Y., and Z. Xiang. 2014. "'Income Inequality in Today's." *China.' Proceedings of the National Academy of Sciences* 111 (19): 6928–6933. doi:10.1073/pnas.1403158111.

Xie, Y., and J. Yongai. 2015. "Household Wealth in China." *China Sociological Revue* 47 (3): 203–229. doi:10.1080/21620555.2015.1032158.

Zhou, Y., and L. Song. 2016. "Income Inequality in China." *Causes and Policy Responses, China Economic Journal* 9: 186–208.

Appendix

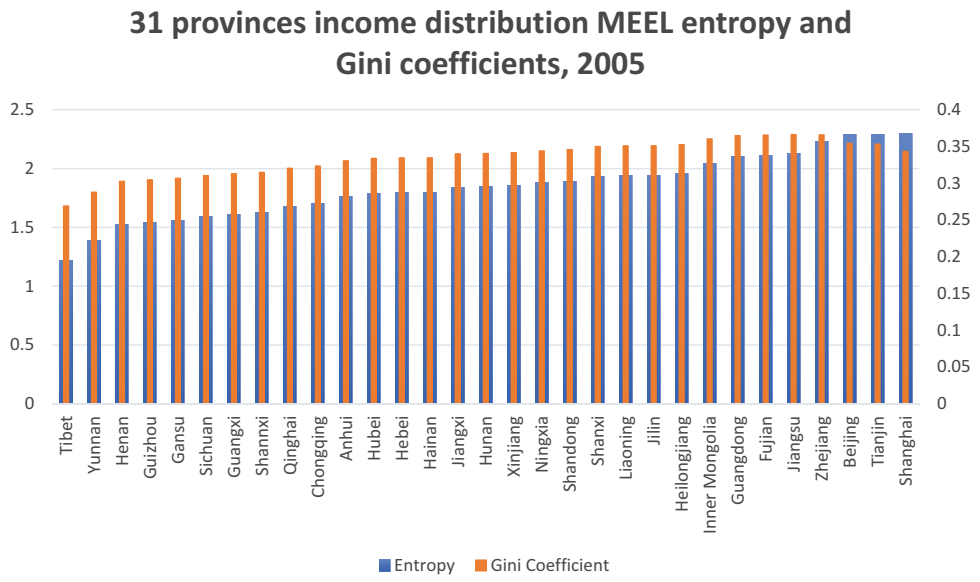

Figure A1. Province level entropy measures of inequality and Gini coefficients.

If income distributions are different then the entropy measure is different. This is not a guarantee for the Gini Coefficient measure as we can see in Figure A1. It depicts in blue bars the estimated entropy measures for the 31 provinces, in increasing entropy value (that is in decreasing measure of income inequality) jointly with the Gini coefficients for the same 31 provinces using the 2005 MEEL based estimated income distributions for each province in orange vertical lines. We see that ranking of provinces differs for three provinces, namely, Beijing, Tianjin, and Shanghai. Those are the provinces that according to the highest measures of entropy correspond to the lowest income inequality provinces among the 31 provinces in China. However, based on the Gini measure the three are not the ones with the highest Gini coefficients.