UNITED NATIONS
UNIVERSITY

UNU-WIDER
World Institute for Development
Economics Research

Research Paper No. 2006/134

Estimating Income Inequality in China Using Grouped Data and the Generalized Beta Distribution

Duangkamon Chotikapanich,[1]
D.S. Prasada Rao,[2] and Kam Ki Tang[2]

November 2006

Abstract

There are two main types of data sources of income distributions in China: household survey data and grouped data. Household survey data are typically available for isolated years and individual provinces. In comparison, aggregate or grouped data are typically available more frequently and usually have national coverage. In principle, grouped data allow investigation of the change of inequality over longer, continuous periods of time, and the identification of patterns of inequality across broader regions. Nevertheless, a major limitation of grouped data is that only mean (average) income and income shares of quintile or decile groups of the population are reported. Directly using grouped data reported in this format is equivalent to assuming that all individuals in a quintile or …/

Keywords: Gini coefficient, generalized beta distribution, urban and rural inequality
JEL classification: D31, C13, C16

[1]Department of Econometrics and Business Statistics, Monash University, Melbourne; [2]School of Economics, University of Queensland, Brisbane

decile group have the same income. This potentially distorts the estimate of inequality within each region. The aim of this paper is to apply an improved econometric method designed to use grouped data to study income inequality in China. A generalized beta distribution is employed to model income inequality in China at various levels and periods of time. The generalized beta distribution is more general and flexible than the lognormal distribution that has been used in past research, and also relaxes the assumption of a uniform distribution of income within quintile and decile groups of populations. The paper studies the nature and extent of inequality in rural and urban China over the period 1978 to 2002. Income inequality in the whole of China is then modelled using a mixture of province-specific distributions. The estimated results are used to study the trends in national inequality, and to discuss the empirical findings in the light of economic reforms, regional policies, and globalization of the Chinese economy.

Acknowledgements

1 Introduction

As China accounts for about a quarter of the world population, changes in income and income inequality in China have important implications to global income inequality. In fact, Milanovic (2002) shows that rising income gap between rural China (together with rural India and Bangladesh) and several large and rich OECD countries, plus the gap between urban China and rural China (and rural India), were the main reasons why the world Gini coefficient increased from 62.8 in 1988 to 66.0 in 1993. This means that any advancement in the measurement of income inequality within China is not only important for understanding the economic development and well-being of the people inside the 'middle kingdom', but also important in the global context.

Thereby, it is not surprising that studies on income inequality in China are abundant.[1] This strand of research is predominately empirical and, therefore, data and methodology concerns are inevitably central issues. There are two main types of data sources of income distributions in China: household survey data and grouped data. Although household surveys have been conducted on an annual basis in most provinces since the mid 1980s, the release of unit record data from those surveys is sporadic. For instance, Meng (2004) has access to the unit record data of three urban household surveys conducted in 1988, 1995 and 1999, with only five provinces in common across all three years; Gustafsson and Li (2002) obtained the unit record data of two rural household surveys conducted in 1988 and 1995, with 18 provinces surveyed in both years. In comparison, aggregate or grouped data are typically available more frequently and usually have a national coverage. In principle, grouped data allow investigation of the change of inequality over longer and continuous periods of time, or identification of patterns of inequality across broader regions. Notwithstanding, a major limitation of grouped data is that only mean (average) income and income share of quintile or decile groups of the population are reported at the provincial level. Directly using grouped data reported in these formats is equivalent to assuming that all individuals in a quintile or decile group having the same income. This potentially underestimates inequality within each province. The distortion could be even more severe when one aggregates provincial data into regional data in order to investigate regional disparities.

The main objective of the paper is to examine levels and trends in inequality in China using income distribution data available in grouped form, with the help of recently developed econometric methods that can fit fairly flexible income distributions to grouped data. Two types of income distributions belonging to the class of generalized beta distributions are employed to model income inequality in rural and urban China and also to model inequality at the provincial level. The generalized-beta distribution is

---

[1] Most recent contributions include Knight and Song (2003); Meng (2004); Zhang and Kambur (2001); Gustafsson and Li (2002); Tsui (1998); Wan et al. (2004); Wei (1999).

more general and flexible than the lognormal distribution used in past research, and at the same time it relaxes the assumption of an equal distribution of income within quintile or decile groups of the population. The method was recently applied to study income distributions of eight East Asian countries in 1988 and 1993, and proved to be very useful in handling grouped data in these cases (Chotikapanich et al. 2006). The econometric methodology employed here can be used regardless of whether the income distribution data are in the form of size classes or in the form of income classes. In addition to modelling income distributions for different regions (rural and urban) or for different provinces, the methodology can also be used to generate income distributions for China as a whole, thus making it possible to study income inequality at the national level.

The rest of the paper is organized as follows. Section 2 explains the data sources and describes the income distribution data. Section 3 briefly describes the econometric methodology. Section 4 presents the estimation results and inequality estimates for rural and urban China, and Section 5 provides some concluding remarks.

2   Data

The current study makes use of published income distribution data on China from various public sources. An attempt is made to use the most recent data so that the results from the current study can be linked to some of the previous studies in assessing the impact of globalization on inequality in China. It is frequently found that distributional data are not available in the most ideal *unit record* format with survey data at the household level. Most of the published income distribution data are available in an aggregated form where households are grouped into size classes or into income classes. Some of the principal sources of data and a few salient features of the data are discussed here.

Income data are sourced from various issues of the China Statistical Yearbook. In cases where figures for the same year are different in different Yearbooks, the figure from the latest Yearbook is used. Population data are sourced from China Statistical Yearbook 2004 and China Population Statistics Yearbook 2000.

2.1 Data for the study of urban and rural inequality

The population is classified as either urban or rural, where urban includes individuals residing in cities and townships. The manner in which income distribution is reported in the China Statistical Yearbook for rural and urban residences is very different. For urban areas, the mean per capita annual income[2] or disposable income for different

---

[2] Admittedly, per capita annual disposable income has limitations as a welfare measure. For instance, Chen and Ravallion (1996) pointed out that consumption is more closely related to living standards than disposable income, especially in rural areas, and can be more accurately measured. However, since most previous work in inequality in China is based on income, the use of income allows us to compare our

percentiles of households (size classes) is reported. For instance, in 2002, the 10 per cent of households with the highest income earned about RMB20,200 per capita per year, the next highest 10 per cent households earned about RMB12,600, and so forth.[3] Figure 1 shows the income distributions of urban households in 1987, 1995, and 2002. The level of income for each percentile is expressed as a ratio of the national average. It therefore eliminates the effects of general inflation on the absolute income measures. The figure shows clear divergence across different urban household income groups over the past one and a half decades. For instance, in 1987, the highest 10 per cent group earned about twice of the lowest 10 per cent group did, the ratio increased to close to three times in 1995, and over six times in 2002.

Figure 1: Urban household income distribution, selective years
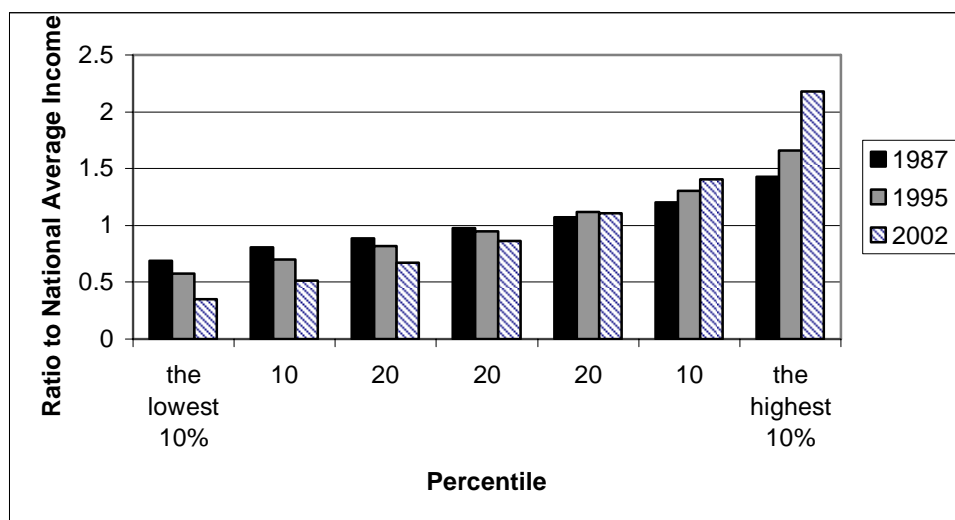
In estimating inequality in urban China, data on the distribution of households are converted into the distribution of persons by using data on household size within each income class. In general it was found that average household size decreased as household income increased. Therefore inequality estimates based on household income distribution data are likely to understate the overall inequality.

In the case of rural areas, the percentages of households grouped by per capita annual net income are reported. For instance, in 2003, 0.49 per cent of the rural households earned RMB100 or below, 0.18 per cent earned RMB100 to RMB200, and so forth. The range of income has changed over the last decade, reflecting the fact that there are substantial increases in nominal household income. However, even after controlling for inflation, the rise in household income is still significant. Figure 2 shows the income

results with the existing literature. More importantly, in the *China Statistical Yearbook*, which is the most widely available data source, there is only information on the percentage of households grouped by per capita income, but not by per capita expenditure.

[3] Before 1989, the income distribution of urban areas was reported in two forms, one was in a format similar to the rural data used in this study.

distributions of rural households in 1980, 1990, and 2000, all measured in 1980 price level.[4] For instance, in 1980, only 13 per cent of rural households earned more than RMB300; in 1990, 56 per cent of rural households passed the RMB300 income level, and the figure increased further to 83 per cent in 2000.

Figure 2: Rural household income distributions, selective years
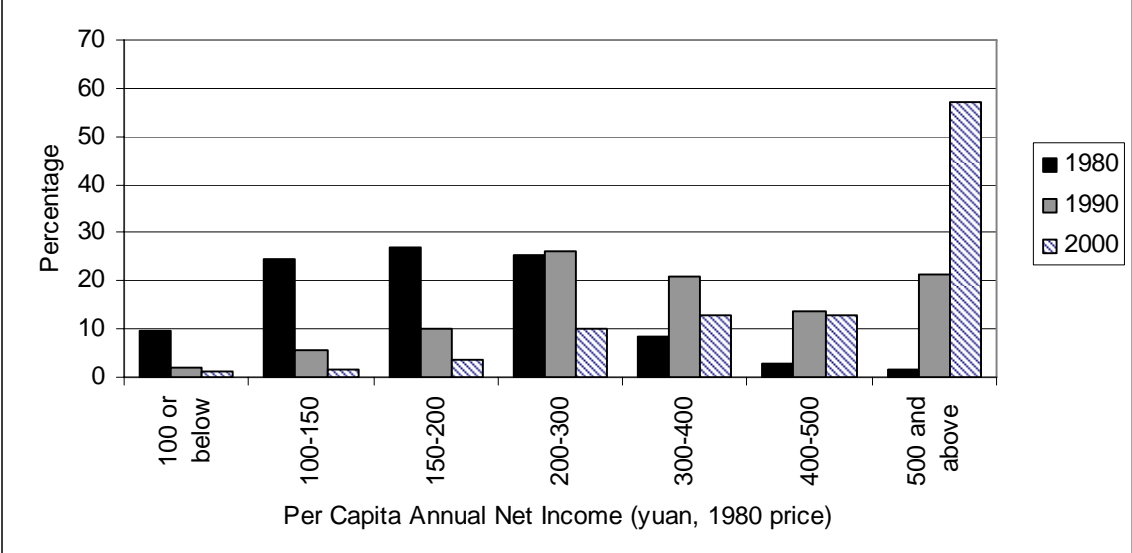
In summary, for urban areas the percentage of households for each income group is fixed but the mean income for each group varies; for rural areas the income range is fixed but the percentage of rural households that fall within each income range varies over time. The difference in the way the data are reported call for different estimation procedures.

In addition to the income distribution data, data on the size of rural and urban populations have been utilised in combining rural and urban income distributions into a single distribution for the whole country. The actual methodology used in deriving the combine distribution is fully explained in the next section.

2.2 Adjustment for price level differences using spatial deflators

As the current study makes use of income data from rural and urban China, it is necessary to make appropriate adjustments for price level differences across rural and urban areas and for price movements over time before combining the corresponding income distributions. For instance, Chen and Ravallion (1996) show that proper accounting for changes in price levels is crucial in estimating inequality in rural China. In this study we construct the necessary price index numbers using a data set recently developed by Brandt and Holz (2006) (hereafter BH). We follow a simple two-step procedure. First, we derive a price index for rural china with urban China as the base for

---

[4] The rural CPI index was drawn from China Data Online.

the year 1990 using BH data.[5] For purposes of adjusting the index for other years, we use the urban and rural CPI data to extrapolate the 1990 rural-urban price index to cover other years included in our analysis.

3 Methodology

Given the nature of income distribution data available in an aggregated form, income shares for different size classes, or population shares in different income classes, we make use of recently developed techniques to fit flexible income distributions to limited aggregated data.[6] In this section we briefly outline the methodology, which essentially involves three stages.

At the first stage, we fit a selected statistical distribution to the income distribution of each population subgroup (rural or urban). We consider two special cases of the generalized beta distribution, namely, the beta-2 distribution and the Weibull distribution. Both of them have been found to adequately describe income distributions. For the urban data where they are in the form of population shares and mean incomes we utilise a recently developed econometric methodology based on the generalized method of moments (GMM) to estimate the parameters of the beta-2 distribution. For the rural data where the available information is in the form of population shares for each income interval we use the simple maximum likelihood approach to the estimation of the Weibull distribution.

In the second stage, we derive national income distributions by combining distributions for urban and rural population subgroups. This approach is based on the notion that the overall distribution is essentially a population-share weighted mixture of income distributions for population subgroups.

Finally, income distributions derived at the national level are used in studying the levels and trends in inequality using Lorenz curves and the estimated Gini coefficients.

---

[5] The price index is essentially of a 'fixed-basket' type which does not allow for substitution. Notwithstanding this limitation this is the only spatial price index available.

[6] In this paper the main focus is on fitting income distributions instead of fitting Lorenz curves to grouped data. Fitting Lorenz curves with grouped data could be accomplished using POVCAL package available on the World Bank website. In fact POVCAL allows generalized-beta specification for the Lorenz curve. If the primary aim is to simply obtain a measure of inequality like the Gini coefficient, it is sufficient to estimate a Lorenz curve using a package like POVCAL. However, our emphasis is on income distributions as we aim to combine distributions (rural and urban combine to describe distribution for the whole of China). It is not possible to derive the properties of the income distribution underlying a Lorenz curve. For example, it is difficult to derive the density function of the income distribution associated with the Lorenz curves that could be fitted using POVCAL. It is clear that information on income distributions is more useful and density and distribution functions provide a rich tapestry of detail of the underlying distribution that cannot be studied using Lorenz curves.

3.1 Modelling income distributions for population subgroups

We assume that income distributions can be modelled using either beta-2 distribution with three parameters or the Weibull distribution with two parameters. The choice of the distribution and the method of estimation depend on the data available. Both distributions are special cases of the generalized beta distribution. The generalized beta distribution is a flexible distribution and it has been shown to provide a good fit to a variety of empirical income distributions—see McDonald (1984); McDonald and Ransom (1979). Bandourian et al. (2002) compared the performance across countries and over time of different functional forms for income distributions that belong to the family of the generalized beta distribution. They have shown that the Weibull distribution is the best-fitting two parameter special case of the generalized beta distribution.

The beta-2 distribution

A number of generalizations of the normalized beta distribution have been used to fit income distributions (see, for example, McDonald 1984; and Kleiber and Kotz 2003: Ch6). The one that we adopt here is sometimes known as the beta-2 distribution and has the probability density function (pdf)

$$f(y) = \frac{y^{p-1}}{b^p B(p,q)\left(1 + \frac{y}{b}\right)^{p+q}} \qquad\qquad y > 0 \qquad\qquad (1)$$

where $b > 0$, $p > 0$ and $q > 0$ are parameters and

$$B(p,q) \;=\; \frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)} \;=\; \int_0^1 t^{p-1}(1-t)^{q-1}\,dt$$

For the mode of $f(y)$ to be nonzero $p > 1$ is required; for the mean to exist $q > 1$ is required. Its corresponding cumulative distribution function (cdf) is given by

$$F(y) \;=\; \frac{1}{B(p,q)} \int_0^{[y/(b+y)]} t^{p-1}(1-t)^{q-1}\,dt \;=\; B_{y/(b+y)}(p,\;q) \qquad (2)$$

The function $B_t(p,q)$ is the cdf for the normalized beta distribution defined on the (0,1) interval. It is a convenient representation because it is commonly included as a readily-computed function in statistical software. If *T* is a standard beta random variable defined on the interval (0, 1), then the relationship between *T* and *Y* is

$$T = \frac{Y}{b+Y} \qquad\qquad\qquad Y = \frac{bT}{1-T}$$

The mean, mode and variance of *Y* are given by

$$\mu = \frac{bp}{q-1} \qquad\qquad\qquad m = \frac{(p-1)b}{q+1}$$

$$\sigma^2 = \mu\left[\frac{b(p+1)}{q-2} - \mu\right] = \frac{b^2 p(p+q-1)}{(q-1)^2(q-2)} \tag{3}$$

The estimation procedure requires starting values for b, p and q. It is often easier to suggest reasonable starting values for μ, m and σ^2. In this case corresponding values for b, p and q can be found from the relationship between the parameters of the distribution and the standard measures like the mean, mode and the variance

$$b = \frac{\mu^2(\mu-m) - (3m-\mu)\sigma^2}{\sigma^2 - \mu^2 + \mu m}$$

$$p = \frac{\mu}{b}\left(\frac{2m+b}{\mu-m}\right) \qquad\qquad q = \frac{\mu+m+b}{\mu-m} \tag{4}$$

For future reference we note that the Gini coefficient (see McDonald 1984) is given by

$$G = \frac{2B(2p, 2q-1)}{pB^2(p,q)} \tag{5}$$

If the parameters b, p and q are known then the distribution is completely known and the Gini coefficient can be computed. So it is sufficient for us to have an estimation procedure to estimate these three parameters.

The Weibull income distribution

The characteristics of the Weibull distribution are as follows. The cumulative distribution function (cdf) is given by

$$F(y; \alpha, \beta) = 1 - e^{-(y/\beta)^{\alpha}} \qquad\qquad \alpha > 1, \beta > 0 \tag{6}$$

It has a probability density function (pdf)

$$f(y; \alpha, \beta) = \frac{\alpha\, y^{\alpha-1} e^{-(y/\beta)^{\alpha}}}{\beta^{\alpha}} \tag{7}$$

The mean is

$$\mu = \beta\,\Gamma\left(1 + \frac{1}{\alpha}\right) \tag{8}$$

The Gini coefficient (see McDonald 1984) is defined as

$$\gamma = 1 - \frac{1}{2^{1/\alpha}}$$

Estimation of income distributions for urban China

For urban China, the data are in the form of mean incomes for different percentiles of households, we fit a beta-2 distribution using the method suggested by Chotikapanich et al. (2006). Suppose we have N income classes $(a_0, a_1), (a_1, a_2), \ldots, (a_{N-1}, a_N)$, with $a_0 = 0$ and $a_N = \infty$. Let the mean class incomes for each of the N classes be given by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N$; and let the population proportions for each class be given by $c_1, c_2, \ldots, c_N$.

Given available data on $\bar{y}_i$ and c_i, but not on a_i, our problem is to estimate the parameters of a beta-2 distribution, along with the unknown class limits $a_1, a_2, \ldots, a_{N-1}$.

The approach we use is to fit a beta distribution to the data such that the sample moments $\bar{y}_i$ and c_i are 'close' to their population counterparts. This approach is equivalent to fitting a distribution such that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{2N}$ are 'close to zero' where

$$c_i = \int_{a_{i-1}}^{a_i} f(y)dy + \varepsilon_i \qquad\qquad i = 1, 2, \ldots, N \qquad\qquad (9)$$

and

$$\bar{y}_i = \frac{\displaystyle\int_{a_{i-1}}^{a_i} yf(y)dy}{\displaystyle\int_{a_{i-1}}^{a_i} f(y)dy} + \varepsilon_{N+i} \qquad\qquad i = 1, 2, \ldots, N \qquad\qquad (10)$$

Chotikapanich et al. (2006) show how to find estimates of the parameters, b, p, q and the class limits $a_1, a_2, \ldots, a_{N-1}$ that minimize the weighted sum of squares function

$$\sum_{i=1}^{N}\left[\left(\frac{\varepsilon_i}{c_i}\right)^2 + \left(\frac{\varepsilon_{N+i}}{\bar{y}_i}\right)^2\right] \qquad\qquad (11)$$

This can be achieved by recognizing that equations (9) and (10) can be rewritten in terms of the beta distribution function as follows.

$$c_i = B_{a_i/(b+a_i)}(p,q) - B_{a_{i-1}/(b+a_{i-1})}(p,q) + \varepsilon_i \qquad\qquad (12a)$$

and

$$\bar{y}_i = \frac{bp}{q-1}\left(\frac{B_{a_i/(b+a_i)}(p+1,q-1) - B_{a_{i-1}/(b+a_{i-1})}(p+1,q-1)}{B_{a_i/(b+a_i)}(p,q) - B_{a_{i-1}/(b+a_{i-1})}(p,q)}\right) + \varepsilon_{N+i} \qquad\qquad (12b)$$

where $B_{a_0/(b+a_0)}(p,q) = 0$ and $B_{a_N/(b+a_N)}(p,q) = 1$.

Chotikapanich et al. (2006) provide details of the estimation procedure which involves minimization of equation (11) with respect to the unknown parameters. The estimation can be done using the non-linear least squares options available in standard econometric package like EViews.[7] Starting values for the non-linear optimization problem are derived using descriptive statistics from the sample relating to the population moments described in equation (3).

In Appendix Table A1 we present the observed income shares of different size classes and the corresponding fitted income shares computed using estimates of parameters of the fitted beta-2 distribution. The fitted shares are quite close to the observed shares. It is shown in Chotikpanich et al. (2006) that the beta-2 distribution used here fits the data

---

[7] The code used in estimating parameters of the beta-2 distribution using EViews is available with the others if some readers are interested in using this approach.

better than the lognormal distribution which is routinely fitted to income distribution data. The superior performance can be attributed to the flexible nature of the beta-2 distribution as well as the efficiency associated with the generalized method of moments estimators of parameters. As the basic data, which is in the form sample moments like the observed shares for size classes, is derived from household expenditure surveys with very large samples, the parameters of the beta-2 distribution possess useful asymptotic properties.

Estimation of income distributions for rural China

For rural China, the data are in the form of the proportion of individuals belonging to a given income class. Availability of grouped data in this form means that the likelihood function is in the form of a multinomial distribution with probabilities prescribed by the cumulative distribution function of the assumed income distribution. For an income distribution, the likelihood function is the pdf for a potential sample of numbers of income units in each of the groups, m_i. It is given by the multinomial distribution

$$f(\theta) = M! \prod_{i=1}^{N} \frac{\left[F(a_i; \theta) - F(a_{i-1}; \theta) \right]^{m_i}}{m_i!} \tag{13}$$

where $F_i(.)$ is the cumulative distribution function (cdf) for an income distribution, θ is a vector containing the distributional parameters and N is the number of income classes. The a_i and a_{i-1} are the upper and lower limits of the ith income class, M is the total number of observations.

Taking the log of Equation (13) and dividing through by M it can be shown that:

$$\frac{1}{M} \log f(\theta) = \frac{1}{M} \log \left(\frac{M!}{m_1! m_2! ... m_N!} \right) + \sum_{i=1}^{N} \frac{m_i}{M} \log \left[F(a_i; \theta) - F(a_{i-1}; \theta) \right] \tag{14}$$

Since M and m_i are constant the maximum likelihood estimate for θ can be found by maximizing the log-likelihood function $\sum_{i=1}^{N} \frac{m_i}{M} \log \left[F(a_i; \theta) - F(a_{i-1}; \theta) \right]$.

We initially estimated the income distribution assuming that $F(.)$ follows the beta-2 distribution. We found that the estimates were quite unstable. We also found that the correlation between the parameters b and q is very high. This suggests that a two-parameter distribution may be a good representation of the income distribution. We thus choose to use the Weibull distribution for the case of rural data.

3.2 Method for combining income distributions for different population subgroups

The methods described in Section 3.1 can be used to fit income distributions to a population subgroup, such as a particular rural or urban region. After estimating the rural and urban income distributions we are in a position to combine them to form a

national income distribution. Given rural and urban income distributions each with an income pdf[8] $f_k(y)$, $k = 1, 2$, and population proportions λ_k, the pdf for the income distribution for the whole China is given by the mixture

$$f(y) = \sum_{k=1}^{2} \lambda_k f_k(y) \tag{15}$$

The national cumulative distribution function is given by the same weighted average of the country-specific cdf's

$$F(y) = \sum_{k=1}^{2} \lambda_k F_k(y) \tag{16}$$

where $F_k(y)$ is defined as $B_{y/(y+b_k)}(p_k, q_k)$ for the case of the beta-2 distribution or $1 - e^{-(y/\beta_k)^{\alpha_k}}$ for the case of the Weibull distribution.

National mean income is given by

$$\mu = \sum_{k=1}^{2} \lambda_k \mu_k \tag{17}$$

where μ_k is defined as $b_k p_k / (q_k - 1)$ or $\beta_k \Gamma(1 + (1/\alpha_k))$ for the case of beta-2 and Weibull distributions, respectively.

The national cumulative income shares are given by

$$\eta(y) = \frac{1}{\mu} \int_0^y z f(z) \, dz$$

$$= \frac{1}{\mu} \sum_{k=1}^{2} \lambda_k \int_0^y z f_k(z) \, dz \tag{18}$$

where $\int_0^y z f_k(z) \, dz$ can be shown to be equal to $\mu_k B_{y/(y+b_k)}(p_k + 1, q_k - 1)$ (see, Kleiber and Kotz, 2003: 192) or $\mu_k \Gamma_{(y/\beta_k)^{a_k}}(1 + 1/a_k)$ for the case of beta-2 and Weibull distributions, respectively. (For the case of the Weibull distribution, see our Appendix for a proof.)

A national cumulative distribution function can be graphed by using equation (16) to compute $F(y)$ for a grid of values of y. A national Lorenz curve, relating income shares to population shares, can be graphed by using equations (16) and (18) to compute $F(y)$ and $\eta(y)$ for a grid of values of y.

The Gini coefficient for China is calculated using η_i and F_i that are obtained from $\eta(y)$ and $F(y)$ for a grid of values of y. The expression is:

---

[8] We may use the Weibull distribution or any other income distribution in place of the beta distribution. The only difference is in the actual implementation of the formulae in equations (11) to (14).

$$Gini = \sum_{i=1}^{N} \eta_{i+1} F_i - \sum_{i=1}^{N} \eta_i F_{i+1} \tag{19}$$

See Kakwani and Podder (1973: 288) for this expression.

The Theil index[9] is calculated as

$$Theil = \sum_{i=1}^{N} s_i \left(\log(s_i) - \log(p_i) \right) \tag{20}$$

where s_i is the income share and it is calculated as $s_i = \eta_i - \eta_{i-1}$. The variable p_i is the population share and is calculated as $p_i = F_i - F_{i-1}$.

In fact, once the distributions at the regional and national level are derived, the fitted income distributions offer the possibility of undertaking an in-depth analysis of income inequality.

4   Empirical Results

Given the massive amount of results emanating from the econometric estimations, only key results are presented below. Detailed results are available from the authors upon request. We present the fitted distributions and estimated inequalities computed using rural/urban and provincial data in separate sections. We present the results for urban and rural China, followed by results for the whole country by combining the results from rural and urban regions.

4.1 Income distributions and inequality in urban China

Data for the study of urban inequality is in the form of size classes showing mean income of the population (and households) belonging to the poorest 10 per cent, the next 10 per cent, middle income groups, and the richest 20 and the top 10 per cent of the population in urban areas. Income distributions have been fitted using the beta-2 form of the generalized beta distribution which involves three parameters *b*, *p*, *q*. Though the data are available for all the years of 1985 and 1987 to 2003, we present the estimated coefficients for selected years.[10] For purposes of examining trends in inequality, we also obtain the 1981 data for urban area. This data set is different from the rest of urban data and it is in the form of population share in each income interval. The data for 1981 are in the same format as those for rural areas. For this case we estimate the income

---

[9] For purposes of illustration we opted to use Theil's T-measure of inequality. The Theil index is included here as an additional measure to check the robustness of the trends in inequality computed using the Gini-coefficient.

[10] It would normally be useful to compute standard errors of the estimated parameters and the Gini and Theil coefficients. Given that estimates of *b*, *p* and *q* are derived using a generalized method of moments type estimator and the fact that the sample size on which the moment conditions are based are quite large (usually the sample sizes are in excess of 30,000 households representing more than 75,000 individuals if the average household size is taken to be around 2.5). The asymptotic standard errors of the estimators are likely to be very small given such large sample sizes.

distribution using the maximum likelihood estimation assuming that the distribution follows the Weibull distribution.

Table 1: Estimates of parameters of urban income distributions

beta2	b	p	q	Gini	Theil
2003	2661.8385	6.9390	5.5226	0.3293	0.1550
2000	2652.8252	10.1003	10.1152	0.2528	0.1041
1995	733.6429	27.7792	10.2085	0.2114	0.0727
1991	345.6764	61.8043	14.0357	0.1703	0.0462
1985	679.7618	33.9795	17.9029	0.1665	0.0436
Weibull	β	α		Gini	
1981	88.0058	3.7979		0.1668	0.0427

Source: See text.

The parameters of the distribution show accelerated shifts in the income distribution indicated by the changes in the estimated coefficients. Specifically, half of the total increase in the Gini coefficient between 1985 and 2003 took place in the last four years only.[11]

In Chotikapanich et al. (2006), the beta-2 distribution fits the observed data very closely. In Appendix Table A1, we present the observed and expected income shares for different size classes for urban China for all the years considered in the study. It can be seen that the beta-2 distribution also fits the observed data very well for our case.[12] The density function for urban China in each year can be graphed by using Equation (2) and the estimated parameters b, p and q to compute $f(y)$ for a grid of values of y. Shifts in the income distribution are shown in Figure 3. It is evident from Figure 3 that the location of the distribution has shifted rightward indicating an increase in the mean of the distribution during the study period. These results are consistent with the rapid growth in real per capita income over the reform period.

The shifts in per capita income evidenced in Figure 3 are accompanied by increases in inequality in income as measured by the Gini coefficient, which has nearly doubled

---

[11] Following the suggestion of an anonymous referee we computed approximate standard errors for the Gini coefficients reported in Table 1 using the formula given on page 117 of Cowell (2000). For a large sample size (say 30,000 sample households) underlying the sample moments used in this study, the standard error of the Gini coefficient for the year 2003 is found to be 0.0011 which is small compared to the Gini coefficient value of 0.3293. We note, however, that the formula in Cowell (2000) provides an approximate standard for the Gini coefficient when the distribution is symmetric (our distributions are skewed here). We believe that the true standard errors associated with these Gini coefficients (which are non-linear functions of the parameters of the beta-2 distribution) are likely to be very small.

[12] Chotikapanich et al. (2006) provide a comparison of goodness-of-fit of beta-2 and lognormal distributions. Both distributions are fitted using the same econometric methodology outlined in this paper. Their results clearly show that the beta-2 distribution provides a better fit than the lognormal distribution in all the countries studied.

from 0.1665 in 1985 to 0.3293 in 2003.[13] We also note that our estimates of the Gini coefficient exhibit a trend similar to that reported by Bramall (2001), although inequality is somewhat lower for each of the years. Similarly, the Theil index has increased fourfold from 0.0436 in 1985 to 0.1550 in 2003. However, the most important result to report here is the rapidly increasing urban inequality accompanying rapid increases in per capita income, and the acceleration of urban inequality in recent years. Another point worth noting is the gradual change in the shape and location of the income distributions for urban China over the study period. This implies that the use of scalar measures of inequality or reliance on Lorenz curves for the analysis of income distribution is likely to mask potentially useful information on income distributions.

Figure 3: Shifts in income distribution in urban China, 1985 to 2003
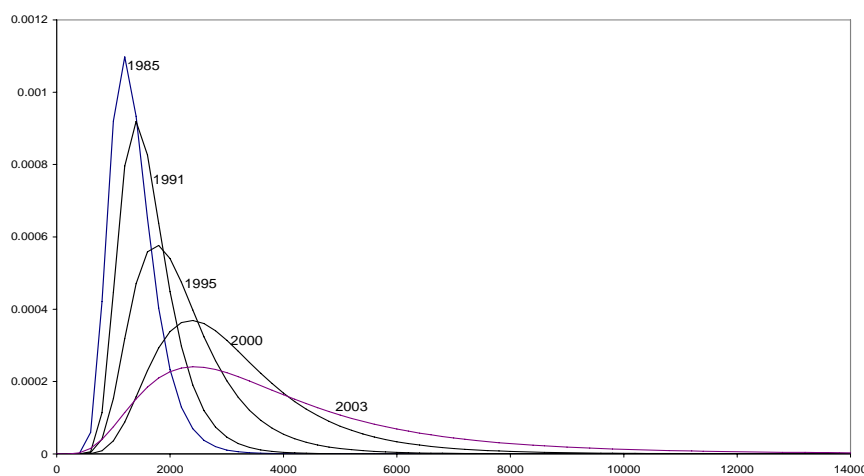
4.2 Income distributions and inequality in rural China

The income distribution data available for our study from the rural region of China is quite different from that available for urban China. Data for rural China is in the form of population shares in different income groups. We also found some internal inconsistencies in the data.[14] As a starting point we tried to estimate the beta-2 distribution using a maximum likelihood estimation. We experienced some serious convergence problems and also found that the estimates were fairly unstable. Further estimates of b and q were highly correlated, indicating that a two-parameter model may be adequate. Based on the excellent performance of the Weibull distribution reported by Bandourian et al. (2002), we decided to use it and the results are reported in Table 2.

---

[13] The value of the Gini coefficient is calculated using the expression for Gini coefficient associated with a beta-2 distribution (given in equation 5) and the estimates of parameters presented in Table 1.

[14] For some years, we had population shares adding up to more than 1. In such cases, in the absence of additional information, we scaled the population shares to sum to unity.

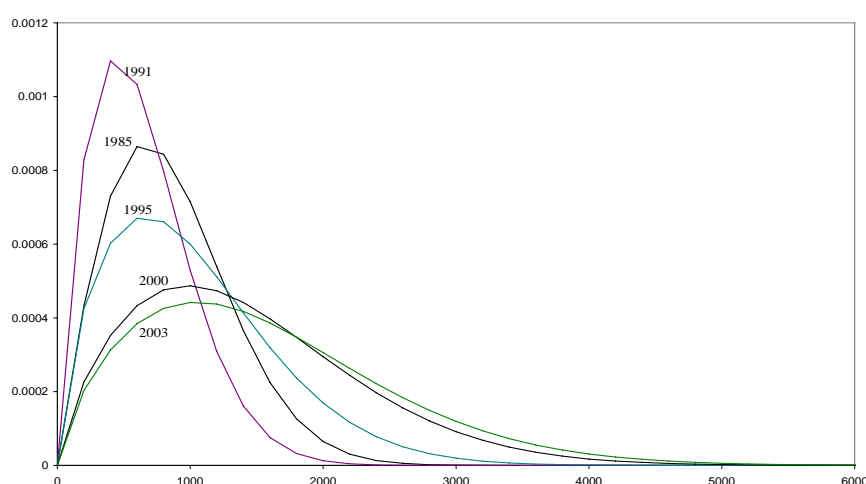Table 2: Estimates of parameters of income distributions in rural China

Weibull	β	α	Gini	Theil
2003	1767.6530	1.7031	0.3344	0.1805
2000	1617.8088	1.7298	0.3302	0.1757
1995	1150.0166	1.6720	0.3394	0.1846
1991	1068.4383	1.7655	0.3247	0.1652
1985	966.0514	1.9479	0.2994	0.1421
1981	675.3239	2.3949	0.2513	0.0969

Source: See text.

Before discussing the results, we briefly comment on the goodness-of-fit of the Weibull distribution. As the data we have here consists of population shares in different income classes, we examine the goodness-of-fit using the observed and estimated population shares presented in the Appendix Table A2. It is clear from Table A2 that the fits are not as good as that of the beta-2 distribution.[15] We therefore urge caution from the reader in using these results.

An interesting aspect of the result is that most of the increase in the Gini coefficient between 1978 and 2003 took place in the 1980s, which was the early period of reform in China that granted peasants with rights to land and financial assets, and to transferring goods and labour (Yang et al. 1992). We also point out that the inequality level in rural China was relatively large in 1985 with a Gini coefficient of 0.2995 compared to an urban Gini coefficient of 0.1665. One principal reason for such a large Gini coefficient is inequality between rural regions in different provinces. Furthermore, inequality in rural China in recent years, as indicated by both the Gini coefficient and the Theil index, seems to have stabilized, in great contrast to the acceleration in urban areas.

Figure 4: Shifts in income distribution in rural China, 1985 to 2003

---

15 Efforts are continuing to solve the problem of highly correlated parameter estimates when the beta-2 distribution is used.

Figure 4 shows the profiles of income distribution in rural China since 1985. While there is a general shift towards the right, the shifts are not as spectacular as in the case of urban China. The mode has shifted from around RMB700 to RMB1200 over the study period.

A comparison of the results in Tables 2 and 3 show that inequality in rural China has been higher than in urban China. However, inequality in urban China has increased so rapidly in recent years that a real possibility is that it will soon bypass that of rural China. The Gini coefficients in Table 2 are fairly similar to those reported by Bramall (2001) which are 0.24, 0.26, 0.31 0.34 and 0.35 respectively for the years 1981, 1985, 1991, 1995 and 1999. Our estimates are also very similar in magnitude to the results reported by Tang (1994) as cited in Tsui (1998).[16] Therefore, despite the less than satisfactory fit of the Weibull distribution, the overall inequality estimates and the trends implicit are consistent with those reported in recent research.

Figure 5: Income distribution in rural and urban China, 1985 and 2003

1985   2003
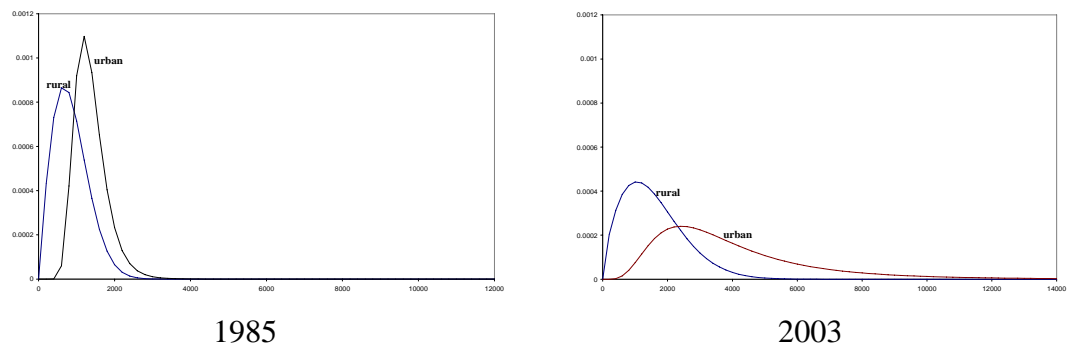
In Figure 5, we present the fitted income distributions for rural and urban China for the years 1985 and 2003. The rural incomes have been expressed in urban prices using appropriate price deflators. The scales of the two graphs are kept the same for the purpose of observing the movement over time. The figure clearly demonstrates the widening income gap between the two regions. In 1985 the distribution for urban China appears to be a lateral shift of that for rural China. However, in 2003 the two distributions are quite dissimilar, indicating that the distribution of incomes in urban China has become quite different from that of rural China. An explanation for this may be found in the asymmetric influence of globalization and economic restructuring on urban and rural regions (see Meng 2004; Kanbur and Zhang 2005).
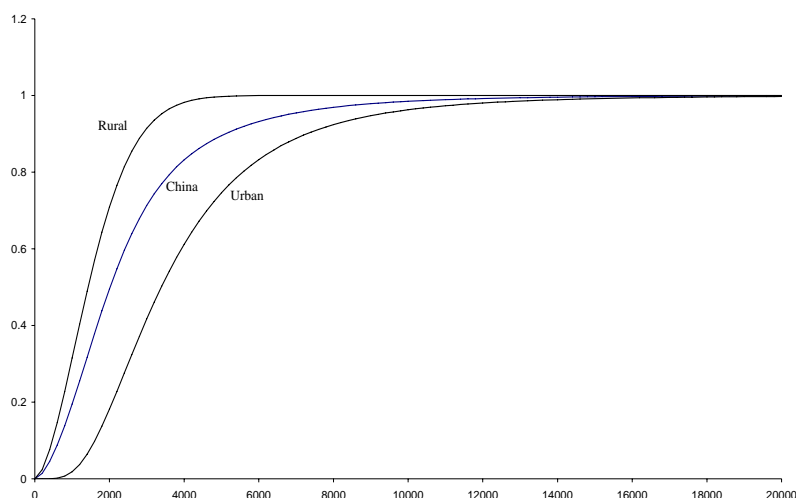
4.3 Income distributions and inequality in the whole of China

In this study, we examine inequality in China using income distributions fitted separately for the rural and urban regions of China. These regional distributions are then combined using population shares as weights and the methodology described in

---

[16] Tang (1994), a paper in Chinese, was cited in Tsui (1998).

Section 3.2. In Figure 6, we present the cumulative distribution functions for rural and urban China and for the country as a whole.

Figure 6: Cumulative Distributions of Income, rural, urban and whole China, 2003

Though we have not attempted the task in this study, it should be possible to model the income distribution data generated for the whole of China as a mixture of region-specific distributions. Table 3 shows the Gini measure of inequality for the two regions and for the country as a whole, derived using the methodology in Section 3.2. Results for China are presented only for the years 1985 to 2003, the only years for which we are able to find the population weights and consistent data.

Table 3: Inequality in rural and urban, 1985-2003

	Gini coefficient			Theil		
	Urban	Rural	China	Urban	Rural	China
2003	0.3293	0.3344	0.4018	0.1550	0.1805	0.2697
2000	0.2528	0.3302	0.3471	0.1041	0.1757	0.2001
1995	0.2114	0.3394	0.3506	0.0727	0.1846	0.2027
1991	0.1703	0.3247	0.3061	0.0462	0.1652	0.1548
1985	0.1665	0.2994	0.2827	0.0436	0.1421	0.1322

Source: See text.

It is very clear from Table 3 that inequality has been steadily increasing in both rural and urban China using both the Gini coefficient and Theil index. A more significant aspect of Table 3 is the significant increases in inequality in the whole of China from 1991 with the Gini coefficient and Theil index well above the levels for rural and urban areas. This trend indicates a widening gap between the rural and urban regions, and not just widening inequality within these regions.[17]

---

[17] We have not attempted any decomposition analysis of inequality in the whole of China. This has been the subject matter a number of studies including Bhalla et al. (2003), Gustafsson and Shi (2002) and Wan et al. (2004). However, as the Theil index is additively decomposable it is possible to measure the contribution of between region (between rural and urban) inequality to over inequality in China. Simple calculations suggest that the contribution of 'between-region inequality' has gone up from around 7 per

Results reported in Table 3 are broadly consistent with estimates of inequality presented in studies by other authors. For example, Kanbur and Zhang (2005) show a Gini coefficient of 0.303 for China for the year 1999. Also, there are many attempts in the past (e.g. Wan et al. 2004) to identify the causes of this increase in inequality.

5   Conclusions

China has experienced a series of structural changes since the late 1970s. The structural changes, induced by economic reform and China's increasing integration with the global economy, are not uniform in both time and spatial dimensions. The early phase of economic reform was concerned with rural areas and product markets. In the 1990s, reforms in urban labour markets and state owned enterprises increased and, more importantly, China's globalization process accelerated, especially in the coastal region. This non-uniform economic restructuring process is observed to be accompanied by rising inequality between the rural and urban areas, and between different regions.

The present study makes use of a recently developed method to estimate income distributions with grouped data (Chotikapanich et al. 2006). The latest available income distribution data are used to examine the levels and trends in income inequality for rural and urban regions in China. We employ the beta-2 and Weibull distributions from the class of generalized beta distributions to study regional and provincial income distributions. We find that the beta-2 distribution fits particularly well to urban whereas we had difficulty in fitting the beta-2 distribution to data on rural China. We instead employed the Weibull distribution to rural data and found that the fits are good but there is further scope to improve on the fits obtained.

Given that the main objective of the study was to examine the feasibility of using a new econometric method to fit income distributions to grouped data, we have not been able to explore a number of interesting and potentially important aspects of inequality in China. For instance, we have not conducted any decomposition analysis to assess the exact nature of the widening gap between rural and urban regions or to examine the contribution of *within-* and *between*-region inequalities to overall inequality in China.
In conclusion, the results reported in this study do clearly demonstrate the feasibility of fitting fairly flexible income distributions to grouped income distribution data, thereby providing a technique which can we be used to answer these and other important inequality questions in the near future.

---

cent to about 37 per cent over the period 1985-2003. This widening gap between rural and urban China is consistent with results reported in recent studies on regional disparities in China.

References

Bandourian, R., J.B. McDonald and R.S. Turley (2002). 'A Comparison of Parametric Models of Income Distribution Across Countries and Over Time', mimeo, Brigham Young University: Provo UT.

Bhalla, A.S., S. Yao, and Z. Zhang (2003). 'Causes of Inequalities in China, 1952 to 1999', *Journal of International Development* 15: 939-55.

Bramall, C. (2001). 'The Quality of China's Household Income Surveys', *The China Quarterly* 167: 689-705.

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

Chen, S., and M. Ravallion (1996). 'Data in Transition: Assessing Rural Living Standards in Southern China', *China Economic Review* 7: 23-56.

Chotikapanich, D., W.E. Griffiths, and D.S.P. Rao (forthcoming 2006). 'Estimating and Combining National Income Distributions using Limited Data', *Journal of Business and Economic Statistics*.

Cowell, F.A. (2000). *Measuring Inequality*, 3rd edn, Oxford University Press: Oxford.

Gustafsson, B., and S. Li (2002). 'Income Inequality Within and Across Counties in Rural China 1988 and 1995', *Journal of Development Economics*, 69: 179-204.

Johnson, N.L., and S. Kotz (1972). 'Power Transformations of Gamma Variables', *Biometrika* 59: 226-29.

Kakwani, N.C., and N. Podder (1973). 'On the Estimation of Lorenz Curves from Grouped Observations', *International Economic Review* 14: 278-92.

Kanbur, R., and X. Zhang (2005). 'Fifty Years of Regional Inequality in China: A Journey through Revolution, Reform and Openness', *Review of Development Economics* 9: 87-106.

Kleiber, C., and S. Kotz (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*, John Wiley and Sons: New York.

Knight, J., and L. Song (2003). 'Increasing Urban Wage Inequality in China', *Economics of Transition* 4: 597-619.

McDonald, J.B. (1984). 'Some Generalized Functions of the Size Distribution of Income', *Econometrica* 52: 647-63.

McDonald, J.B. and M.R. Ransom (1979). 'Functional Forms, Estimation Techniques and the Distribution of Income', *Econometrica* 6: 1513-25.

Meng, X. (2004). 'Economic Restrictions and Income Inequality in Urban China', *Review of Income and Wealth* 56(1): 357-379.

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

Tang, W.S. (1994). Urban Land Development under Socialism: China between 1949 and 1977, *International Journal of Urban and Regional Research* 18: 392-415.

Tsui, K.Y. (1998). 'Factor Decomposition of Chinese Rural Income Inequality: New methodology, Empirical Findings and Policy Implications', *Journal of Comparative Economics* 26: 502-28.

Wan, G., M. Lu, and Z. Chen (2004). 'Globalization and Regional Income Inequality: Evidence from within China', *WIDER Discussion Paper 2004/10,* UNU-WIDER: Helsinki.

Wei, Y.D. (1999). 'Regional Inequality in China', *Progress in Human Geography*, 23(1): 49-59.

Yang, X., J. Wang and I. Wills (1992) 'Economic Growth, Commercialization, and Institutional Changes in Rural China', *China Economic Review* 3(1): 1-37.

Appendix

Table A1: Observed and estimated income shares

Urban 1985			Urban1991	
observed	estimated		observed	estimated
0.0657	0.0664		0.0656	0.0664
0.0795	0.0790		0.0787	0.0785
0.1758	0.1760		0.1751	0.1746
0.1944	0.1955		0.1932	0.1943
0.2162	0.2194		0.2159	0.2190
0.1212	0.1226		0.1207	0.1232
0.1472	0.1410		0.1507	0.1439

Urban1995			Urban 2000	
observed	estimated		observed	estimated
0.0571	0.0575		0.0483	0.0487
0.0710	0.0714		0.0631	0.0640
0.1647	0.1648		0.1547	0.1564
0.1894	0.1904		0.1888	0.1909
0.2207	0.2232		0.2266	0.2307
0.1284	0.1305		0.1380	0.1419
0.1687	0.1622		0.1806	0.1671

Urban 2003	
observed	estimated
0.0351	0.0356
0.0523	0.0521
0.1353	0.1344
0.1775	0.1772
0.2257	0.2296
0.1448	0.1494
0.2293	0.2215

Note: From the beta-2 Distribution fitted using Chotikapanich et al. (2006) method.

Source: See text.

Table A2: Observed and estimated population shares

1981			1991			2003		
class	observed	estimated	class	observed	estimated	class	observed	estimated
0-100	4.7	9.0	0-100	0.4	2.1	0-100	0.5	2.14
100-150	14.9	13.0	100-150	0.77	2.2	100-200	0.2	4.95
150-200	23.0	17.0	150-200	1.56	2.8	200-300	0.3	6.88
200-300	34.8	33.9	200-300	6.64	6.9	300-400	0.5	8.15
300-400	14.4	19.7	300-400	11.08	8.2	400-500	0.8	8.86
400-500	5.0	6.3	400-500	13.35	8.9	500-600	1.2	9.06
500-	3.2	1.1	500-600	12.56	9.1	600-800	3.3	17.22
			600-800	21.21	17.2	800-1000	4.9	14.39
			800-1000	14.19	14.4	1000-1200	5.5	10.79
			1000-1500	15.32	20.8	1200-1300	3.0	4.08
			1500-2000	4.82	6.2	1300-1500	6.4	5.91
			2000-	4.06	1.4	1500-1700	6.5	3.58
						1700-2000	9.4	2.63
						2000-2500	13.8	1.20
						2500-3000	10.8	0.16
						3000-3500	8.0	0.01
						3500-4000	5.84	0.00
						4000-4500	4.2	0.00
						4500-5000	3.16	0.00
						5000-	11.85	0.00

Note: The Weibull Distribution fitted using the Maximum Likelihood Method

Source: See text.

Appendix to show that $\int_0^y z\,f_k(z)\,dz = \mu_k \Gamma_{(y/\beta_k)^{a_k}}(1+1/a_k)$ for the case of Weibull distribution.

$$f(y) = \frac{a}{\beta^a} y^{a-1} e^{-(y/\beta)^a}$$

$$y\,f(y) = a\,(y/\beta)^a\, e^{-(y/\beta)^a}$$

$$\int_0^z y\,f(y)\,dy = \int_0^z a\,(y/\beta)^a\, e^{-(y/\beta)^a}\,dy$$

Let $\qquad u = (z/\beta)^a$

and $\qquad w = (y/\beta)^a \quad$ or $\quad y = \beta w^{1/a}$

$$dy = \frac{\beta}{a} w^{(1/a)-1} dw$$

$$\int_0^z y\,f(y)\,dy = \int_0^z a\,(y/\beta)^a\, e^{-(y/\beta)^a}\,dy$$

$$= \int_0^u a\,w\,e^{-w}\beta\frac{1}{a} w^{(1/a)-1} dw$$

$$= \int_0^u w^{1/a}\, e^{-w}\beta\,dw$$

$$= \beta \int_0^u w^{1/a}\, e^{-w}\,dw$$

$$= \beta \int_0^u w^{\alpha^*-1}\, e^{-w}\,dw$$

$$= \beta \times \Gamma\left(1+\frac{1}{a}\right) \times \frac{1}{\Gamma\left(1+\frac{1}{a}\right)} \times \int_0^u w^{\alpha^*-1}\, e^{-w}\,dw$$

$$= \mu\,\Gamma_u\left(1+\frac{1}{a}\right)$$

where $1/a = \alpha^* - 1$ or $\alpha^* = (1/a)+1$