KU LEUVEN

CENTER FOR ECONOMIC STUDIES

DISCUSSION PAPER SERIES
DPS14.15

JUNE 2014

Challenges of working with the Chinese NBS firm-level data

Loren BRANDT, Johannes VAN BIESEBROECK and Yifan ZHANG

Econometrics

Faculty of Economics
And Business

Challenges of working with the Chinese NBS firm-level data[1]

Loren Brandt, Johannes Van Biesebroeck, and Yifan Zhang

April 22, 2014

Abstract

Over the reform period, industry has been the source of forty percent of GDP, and has contributed 90% of China's exports. Annual firm level surveys that begin in 1992, along with industry-wide census in 1995, 2004 and 2008 are rich sources of data on firms' actions in this important sector. It is well-known that working with Chinese data requires overcoming challenging measurement issues. Macroeconomic series are often suspected to suffer from political interference or reporting biases that stem from political incentives. Working with the firm-level data has its own challenges. Making sure that comparisons over time are consistent is perhaps the most difficult and pervasive issue. This is because of sampling as well as measurement issues for key variables, such as ownership type, real output, value-added, wages, and the capital stock. These problems are apparent, for example, in discrepancies between the evolution of aggregates from the firm-level data and aggregate statistics in the national income accounts. In this paper, we provide an introduction to these data sets. We discuss and illustrate several of the issues that make comparability over time difficult and we suggest solutions for many of them. The importance of a particular measurement issue often depends on the exact application. We illustrate this point by tracing the evolution of the relative productivity level of entrants and incumbents over time, trying to distinguish between changes in actual performance and changes driven by measurement problems. We conclude by identifying a few promising areas of future research and margins on which collaboration among users to improve these data might be beneficial.

[1] Corresponding author: Johannes Van Biesebroeck, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium, Email: jo.vanbiesebroeck@kuleuven.be. Loren Brandt (University of Toronto), Yifan Zhang (Lingnan University). We wish to acknowledge funding from a European ERC grant.

1. Motivation

The Chinese economy over the last decade is an amazing place to do economic research, especially studying the manufacturing sector. The pace of growth has been startling which accelerates the impact of any changes or trends. Partly this was due to the reforms and ensuing restructuring. Partly this was the result of China integrating rapidly into the world economy and becoming *the world's factory*.

Equally amazing is that coinciding with this boom of economic activity, the information from the Chinese firm-level industrial survey has become available to researchers. The official name of this dataset is "all state owned and all above-scale non-state owned industrial enterprise database". It contains information similarly as collected by other countries, but it has become more widely available without cumbersome access requirements. This source of information provides us with a unique window on the economic changes that have reshaped the Chinese manufacturing sector.

Dougherty et al. (2007) and Jefferson et al. (2008) were two of the earliest studies using these data. They analyzed topics of particular importance to China, namely, the emergence of the private sector and productivity convergence by ownership type and across space. Since then researchers have studied a variety of topics spanning almost all fields of economics. In macroeconomics, Hsieh and Klenow (2009) and Song et al. (2011) have used the data to study resource reallocation and aggregate TFP growth. In international economics, Park et al. (2010) study the impact of the Asian financial crises on Chinese manufacturing firms and Brandt et al. (2012) document large productivity effects associated with China's entry into the WTO. In industrial organization, Gao and Van Biesebroeck (2013) estimate the efficiency gains resulting from restructuring of the electricity sector. Aghion et al. (2012) evaluate the effectiveness of China's industrial policy more generally. Summarizing all the contributions would be far beyond the scope of this paper.

Chinese official statistics are often viewed with suspicion. It is often claimed, for example, that local officials have an incentive to overstate GDP growth in their jurisdiction to further their own careers. As a result, the sum of provincial GDP exceeds independent estimates of national GDP by China's National Bureau of Statistics (NBS) by a large margin. While we do not believe the firm-level data undergo similar manipulations, there are important data issues a researcher needs to address to obtain reliable information.

We set ourselves three tasks in this survey paper. First, we will describe the dataset in some detail and compare it to similar firm surveys from other countries. We discuss sampling and coverage, which has grown from 300,000 industrial firms in 1992, the first available survey year, to more than 10 million firms in the 2008 census. Finally, we describe the variables available.

Second, the availability of information on such a large panel of firms in such a rapidly evolving economic environment has enormous research appeal, but exploiting the panel and time dimensions of the dataset also entails big challenges. The following four issues have been particularly difficult in the Chinese case: (i) linking observations over time to identify firms, (ii) obtaining detailed price deflators, (iii) constructing the real capital stock, (iv) making sure that variables are defined consistently over time. On each of these dimensions, we will discuss what the nature of the problem is, how we have dealt with it in our own work, and in some cases what alternative solutions could be applied.

Third, we will illustrate the importance of dealing appropriately with measurement issues by illustrating the impact on one economic phenomenon. In particular, we calculate the gap in total factor productivity between new entrants and incumbent firms and trace the evolution of this gap over time. We show that the way some of the data challenges discussed above are handled can substantially change the economic findings.

We conclude the paper by highlighting a few areas of research on the Chinese economy that are still underdeveloped and where the NBS firm-level data could answer important questions.

2. The NBS survey of above-scale industrial firms

2.1 Coverage of the sample

The Chinese firm-level data is similar to the Longitudinal Research Database (LRD) maintained by the U.S. Bureau of the Census or to the widely used census data for Colombia and Chile. Bartelsman, Haltiwanger and Scarpetta (2009) document comparable data sources for 24 countries and use the information to compare patterns of firm dynamics internationally. Two of the most important differences between the Chinese data and that of many other countries pertain to sampling.

First, the unit of observation is a firm, defined as a legal unit (*faren danwei*). Large Chinese enterprises may have multiple subsidiaries. As long as these subsidiaries are legal units, they will enter the dataset as individual firms. Subsidiaries that are not legal units, so-called “industrial activity units (*chanye huodong danwei*), are excluded. According to *China Statistical Yearbook*, a legal unit needs to meet the following requirements: “(1) They are established legally, having their own names, organizations, location and able to take civil liability; (2) They possess and use their assets independently, assume liabilities and are entitled to sign contracts with other units; (3) They are financially independent and compile their own balance sheets.”[2]

In contrast, most other countries sample plants or establishments, i.e. physical entities operating at a particular address. The data sets of Bureau Van Dijk, which are compiled from company account filings and are available for most countries, are also at the firm level.[3] For many entities, the plant/firm definition coincides: for example, 93.5% of plants in the U.S.

[2] Source: Explanatory Notes on Main Statistical Indicators, Chapter 13, *China Statistical Yearbook* 2009.

[3] Its commercial database products are known as AMADEUS, which contains firm-level information for European countries and is most widely used by academic researchers, or ORBIS which covers data for all sizeable countries worldwide. Firm coverage varies by country.

Census belong to single-plant firms. In most years, Chinese firms are asked how many of their establishments are engaged in industrial activities. In 1998, 7.6% of firms report zero, 88.9% report a single production plant, 1.4% have two plants, and 2.0% more than two. In 2007, the share of single-plant firms increased to 96.6%. In the 2008 economic census, the NBS collected separate information for firms and their establishments.

A second unusual feature of the Chinese data set is the minimum threshold for inclusion in the sample. In the data sets for the Latin American countries mentioned, the sample consists of all active plants with at least 10 workers which are sampled every year. In the United States, all manufacturing establishments that have registered with the IRS to pay Social Security tax for their employees are sampled in census years (every five years). In the intervening years, the Annual Survey of Manufacturers includes all plants with at least 250 employees, all plants with annual shipments of at least 500 million USD (since 1984), and all other plants appearing in the preceding census as well as new entrants are sampled with a probability proportional to their size (and they remain in the sample until the next census year).

In China, we observe annual firm-level data for "above-scale" industrial firms, also called firms above designated size. This includes all state owned firms as well as non-state firms with sales exceeding 5 million RMB.[4] In 2011, the designated size increased from 5 million to 20 million RMB. In addition, in census years all industrial firms, irrespective of size, are sampled.[5] Panel (a) in Table 1 indicates the number of observations in each year. The above-scale sample grows from 165,118 observations in 1998 to 327,853 in 2008.

Three things are important to point out. First, it is impossible to know a firm's annual sales until some information has been collected on the firm. This is especially notable in 2004, a census year. The number of firms included in the above-scale sample jumped by 82,870 observations (net entry), or 42.2% of the number of firms active in the preceding year. This was several times higher than the usual rate of net entry, which averaged 4.9% over the four preceding years. Due to the rapid expansion of the economy, many firms had experienced strong sales growth and had already surpassed the sales threshold for some years without the NBS including them in the sample. For any analysis on entry behavior, it is important to take into account a firm's starting year, a variable in the survey, when identifying new entrants.[6]

Second, small firms can be included in the sample if they generate a very high level of revenue given their employment. This can be due to extremely high productivity or due to measurement error. Importantly, sample selection in the Chinese data set at the lower threshold

[4] At the exchange rate of 8.27 RMB per USD (in force between January 1997 and July 2005), this amount to 605,000 USD.

[5] Self-employed individuals are in principle excluded. Private firms with up to 8 employees could register this way and operate under a different legal system.

[6] 70.9% of new entrants in the 2004 above-scale sample reported a starting year before 2003, i.e. were at least 2 years old. The corresponding statistic in the previous year was 64.7% and in the following year only 51.4%.

is biased towards highly productive small firms. In contrast, most countries use a minimum employment threshold (e.g. at least 10 workers), which induces an opposite bias towards unproductive small firms.

Third, the inclusion threshold for non-state firms of having sales above 5 million RMB is not a hard rule. Firm that are in the sample one year, but whose sales dropped below the threshold the next year, are no longer required to report to the annual survey. However, many of these firms decided to continue their reporting and they are not automatically removed from the sample. In total, 5% of private or collectively-owned firms have sales below 5 million RMB (while there should be none) compared to 19% of state firms (where all small firms should be counted). These percentages are only slightly higher if we calculate them only for new entrants in the sample or only for firms that will exit the sample in the next year. It suggests that the sales threshold was not of prime importance. Moreover, the extremely high rate of growth in the Chinese manufacturing sector over this period makes sales reductions relatively rare anyway. We believe that churning in the sample because of the minimum size threshold is virtually nonexistent.

The NBS firm-level data available to us starts as early as 1992, a year when all firms with independent accounting were covered, about 350,000 firms. In 1998, the NBS changed the coverage of the annual survey to include all state owned firms and only the above-scale non-state firms. It also changed the firm identifiers as part of a wider overhaul of the statistical apparatus. As a result, firms can only be linked over time from 1998 onwards and most studies start their sample in that year. The survey has been conducted in all subsequent years, but to the best of our knowledge the latest year with reliable data available is 2008.[7] Therefore, we focus on the 1998-2008 sample period. The full sample covers all industrial firms, which is defined here to include mining, manufacturing, and public utilities. This corresponds to Chinese Industrial Classification (CIC) codes 0610-1210, 1311-4392, and 4411-4620. Many studies limit themselves to the manufacturing sector.

[Insert Table 1 approximately here]

In Table 1 we report the sum of several important variables in the firm-level data in panel (a) and the corresponding totals from the China Statistical Yearbook or the China Statistical Abstract in panel (b). In principle, the coverage should be identical and the discrepancies are indeed relatively small. For the majority of variables reported, the differences between the first two panels are less than 0.1%. For 1999, 2002, 2003, and 2007 all aggregates are identical, confirming that the firm-level data we use are also the basis for the numbers reported in the Chinese Statistical Yearbooks. For 1998 and 2000, only the employment aggregates are inexplicably lower in the sample, by 8.9% and 3.4%, while in 2006 only exporting is slightly

[7] The data file for 2009 misses important variables, such as revenue, wages, material input, and fixed assets. The data files for 2010 and 2011 that we obtained had incorrect information for employment.

higher (+1.5%). In 2001, the number of firms reported in the Yearbook suggests that we miss 1.3% of active firms and it is thus not surprising that the totals for all variables are slightly lower as well. Finally, in 2004 and 2005 there are small discrepancies, but without a pattern: some variables match, others do not; sometimes the sample aggregate is higher, other times it is higher in the Yearbook.

Overall, differences are minor, except for 2008, for which we miss a substantial fraction of the above-scale firms, approximately 30 percent. Some three-digit industries are missing entirely, but that does not explain the entire difference. The absence of these firms will lead to higher than usual apparent rates of exit in 2007. In principle, we can recover these firms from the 2008 census. Even though there is no explicit variable to identify above-scale firms exactly,[8] using the state registration type and the 5 million RMB sales threshold, the number of firms closely matches the number reported in the Yearbook. However, we have not completed the sample as the number of reported variables in the census is much more limited. The recovered observations would be useless for all the most basic analysis.

As mentioned, the NBS occasionally conducts a full industrial firm census which covers all active firms, irrespective of size or ownership. In 1995, 493,204 firms were surveyed, in 2004, more than 1.37 million firms and in 2008, 1.90 million firms. The statistics in panel (c) of Table 1 allow two additional comparisons. First, we extract from the different censuses all firms that are either SOEs or non-SOEs with sales larger than 5 million RMB, and compare them with our firm-level data set. The aggregates again correspond extremely well.

Second, we indicate the coverage of the above-scale sample relative to the full population of firms. In 1995, slightly less than one third of active firms was above-scale, but the large influx of relatively small non-state entrants reduced the coverage of the above-scale sample to only one fifth of the total by 2004. However, the excluded firms only account for a small fraction of economic activity. In 2004 they employed 28.8% of the industrial workforce, but only produced 9.3% of output and generated barely 2.5% of exports.[9]

2.2 Available variables

Compared to many other countries, the set of available variables in the Chinese data set is unusually extensive. Table 2 lists the most important variables. In the first panel are variables that are included throughout the entire sample period.[10] They include identifying information with detailed industry and geographic codes. Firm ownership can be identified using the official

[8] From our analysis of the 2004 data we know that inclusion in the firm-level annual sample does not follow the 5 million RMB threshold exactly.

[9] The original statistics reported in the 2005 Chinese Statistical Yearbook show smaller totals, for example output was only 18.72 trillion RMB. The higher figure reported in the 2006 Chinese Statistical Yearbook reflects the upward revision following the 2004 Census.

[10] A few exceptions with missing information in a single year are denoted by superscripts.

registration type or from the share in paid-up capital of different groups. Stock variables include various measures of assets, debt, inventory, and accounts receivable. Flow variables detail various dimensions of output, including export volumes, inputs, and taxes.

[Insert Table 2 approximately here]

In more recent years, several aspects of firms operations are broken down in much greater detail. Aspects of firm performance, on both the revenue and the cost side, are reported separately for the main line of business and limited to operational activities. Costs categories that were not available previously, and thus underestimated total costs, are now included. This includes various employment benefits beyond salaries, e.g. pension benefits. It also includes detailed cost categories such as advertising, transportation, and employee training, among others. Some other useful pieces of information, such as accounts payable, number of female employees, and cash flow variables are now reported as well.

Finally, in the bottom panel of Table 2 we list a number of variables that are interesting for specific research topics, but which are not systematically reported. These include output measured at constant prices, i.e. using a set of reference prices provided by the NBS, which we have used to construct detailed price deflators in some years. R&D expenditures are reported for six years and the number of computers in one. In the census year 2004, above-scale firms reported their employment broken down by education level and for various non-exhaustive worker categories, separately for total and for female employment.

3. Four measurement challenges

As mentioned, effectively exploiting the panel and time dimensions of the dataset raises several measurement challenges. For example, nearly 90% of firms report no R&D expenditures in any of the years between 2001 and 2007 (this variable is missing in 2004). The high percentage of zeroes could reflect reality or be due to reporting errors. Another example is the curious fact that seven percent of firms report “foreign or Hong Kong, Macau and Taiwan (HMT)” as ownership type, even though their paid-up capital owned by foreign or HMT investors is zero. Below we focus on four major issues and for each we address three key questions: What is the problem? How did we solve it in our own research? How good are the solutions and are there alternatives?

To foster collaboration in tackling these issues, we made available online files with concordance tables, supplementary information, and programs that were used in Brandt et al. (2012, 2013). We have now updated these files through 2008 and improved some of the programs, often in response to comments and feedback we received.[11]

[11] Data and program files with supplementary information are available online at http://www.econ.kuleuven.be/public/N07057/China/.

3.1 Matching of firms over time

The majority of firm linkages over time are made directly using the unique firm identifiers assigned by the NBS. The current system of IDs was implemented in 1998 and the same IDs are also used in the full census. Occasionally, however, a firm receives a new ID if it goes through a change in its legal registration, for example following a restructuring, merger or acquisition. Where possible, we have tracked firms when their boundaries or ownership structure changed, using information on their name, phone number, address, etc. This is especially important in the Chinese context as many incumbents were restructured or privatized. To investigate the effects of such restructurings we need to be able to link the old and new manifestations of the firm.[12]

The algorithm we used to establish firm linkages over time is available online. It consists of the following steps. First, we match firms in two consecutive years using the NBS ID.[13] Each year 10 to 30 observations have duplicate IDs, in which case we additionally use the firm name to find a match. When for a particular firm no observation with the same ID can be found in the next year, we rely on (combinations of) the firm name (in Chinese), the name of legal person representative (in Chinese), phone number, address, name of main product, geographic code, industry code, and the founding year to look for an alternative match.[14] Second, firms might disappear from the sample and re-enter later. Therefore, we subsequently try to match remaining observations in data files two years apart, again using both the IDs and other identifying information. We continue with this procedure across all years to establish an eleven-year long, unbalanced panel.

Table 3 shows how often we have been able to link observations in different years and the resulting firm histories. 7.2% of all entities are only observed in a single year. This amounts to 184,000 single-year firms, approximately 28% of the total number of firms we identify. At the other extreme, 11% of observations belong to firms that we observe in each of the eleven years. The table also shows that the number of firms active for 2 to 3, 4 to 5, or 6 to 10 years becomes successively smaller, but each group accounts for successively more observations. Approximately one half of observations belong to firms that are observed at least five years, which implies that they must have entered before the 2004 census year.

[Table 3 approximately here]

[12] At the same time, to make sure that such re-entry of a restructured firm with a new ID is not confused with *de novo* entry of a new firm, we always verify the reported startup year for entrants.

[13] In 1-2% of cases, there is a difference in case (upper or lower) for IDs that are otherwise unique and observed in two or more years. We uniformly changed all letters in the IDs to upper case.

[14] The geographic codes are updated yearly to reflect changes of administrative boundaries. These include the incorporation of adjacent rural areas as new districts into existing cities, spinning-off urban districts as independent cities, or promoting fast-growing cities, for example from county to prefecture level. The relevant codebooks are available on the NBS website. Baum-Snow et al. (2013) is one example of a research project where time-consistent city definitions are created, undoing the occasional administrative reclassification of well-defined geographic areas.

In the last two columns of Table 3 we show how often a link is established using information other than the NBS ID. On average, only 3.3% of the links use the additional information. This is misleading however, and at the firm-level more than 11% of the firms have at least one annual link established using information other than the firm ID. The probability of linking a firm using other information increases (slightly) in the number of years a firm is active—as establishing such links to bridge restructuring episodes is an important way to identify long-surviving firms. Because firms linked using this supplementary information can be followed over a longer period—almost 30% of firms active all eleven years experienced and ID change at least once—the total number observations that belong to a firm that experienced an ID switch stands at 15%.

The online appendix to Brandt et al. (2012) reports a few additional patterns about the linkages. As the sample period progresses, the number of links that can be established increases, i.e. the observed entry and exit rates go down. Moreover, the importance of other information than IDs to establish links becomes less important over time as the rate of firm restructuring slowed down. Not surprisingly, the probability a firm changes registration type (ownership) from one year to the next is more than twice as high when the match is made using other information than the ID.

We end up with an unbalanced panel of firms that increases in size from 163,295 firms in 1998 to 326,610 in 2007. Table 4 contains the total number of active firms in each year and breaks this down for each year between incumbents, entrants, and exiting firms. From 2001 to 2002, for example, the total number of firms increased from 167,259 to 179,811, or a net increase of 12,552. Gross entry was significantly higher, namely 30,733, but this was offset by the exit of 19,676 firms, or gross entry and exit flows of 17.1% and 11.8% of active firms.

[Table 4 approximately here]

The sharp increase in the number of firms between 2003 and 2004 that was mentioned already stands out. This increase in entry can be attributed to the 2004 Industrial Census, and the identification of many firms, mostly privately-owned, that should already have been covered in the sample in earlier years. The large number of entrants in 2008, in spite of our incomplete firm-level data file, can similarly be attributed to the 2008 Industrial Census.[15] As we observe the founding year for each firm, we can correctly identify a firm's actual entry year in the industry, as opposed to entry in the sample, such that sampling issues do not confound the true entry and exit patterns.

Important to note in this respect is that the average age of newly entered firms (in the sample) is declining over time. The median age for firms first showing up in the sample in 1999 or 2000 was four years. This declined to an average age of three years for new entrants in 2001 and 2002 and further to two years in 2003. The additional firms picked up in the 2004 census year

[15] The higher than usual fraction of exiting firms in 2007 is almost certainly due to the incomplete data file and it does not reflect an economic reality.

temporarily increased the median age of newly entered firms again, but the downward trend continues afterwards, with a median of two years in all subsequent years.

3.2 Price deflators

To make nominal variables comparable over time, we need a price deflator to express them in constant year prices. To construct an output deflator at the most detailed level possible, we use information from the 1998-2003 firm surveys. For those years, firms were asked to report the value of their output not only in nominal terms, but also in real prices using a set of "reference prices" provided by the NBS. The ratio of nominal to real output provides a firm-specific index of its price level in that year relative to the base year. The change in this index between two years measures the firm-specific price change, which we average to the four-digit industry level. We only use the information in the price changes, not the levels, as a change in the composition of active firms has noticeable effects on the average price level. We drop as outliers observations for which the price change differs by more than half of the standard deviation from the mean, or approximately 15-25% of observations. We then recalculate the weighted average price change for each sector, using current output weights.[16] Annual price changes are linked over time to construct an output deflator for each of the 458 four-digit industries. For the remainder of the sample period, 2004-2008, we use the two-digit (39 industries) ex-factory price index from the China Statistical Yearbook to extend the more detailed deflator.

To construct real value added—defined as output net of goods purchased for resale, indirect taxes, and material inputs—we need an input deflator for raw materials and intermediate inputs. This we construct using our output deflators and input shares calculated from the 2002 National Input-Output (IO) table. Most of the sectors defined in the IO table are less detailed than the industry definition used in the firm-level data and we have constructed a concordance table linking the IO sectors to the four-digit industries. We first calculate an aggregate output price index for each IO sector as an un-weighted average of underlying industry prices. We then obtain the input deflator for each IO sector by calculating a input-share weighted average of these output deflators.

The third deflator we need is for investment which we used in the construction of the real capital stock (below). Perkins and Rawski (2008) have constructed a chain-linked price deflator based on separate price indices for equipment-machinery and buildings-structures. The weights are the share of these items in fixed investment, as reported by China's National Bureau of Statistics.

[Figure 1 approximately here]

[16] This procedure to drop outliers generates deflators that are similar to those obtained using the median price change by sector.

Figure 1 shows the evolution over time of all three deflators. Output prices (solid lines) decreased in most industries between 1998 and 2002 following the Asian financial crisis and only began to rise almost universally after 2003. Over the full eleven-year period, the median sector experienced an increase in its output price of only 17.9% or 1.7% per year. The thinner lines show the 25th and 75th percentiles across industries, which illustrates that there is a large dispersion in the price evolution. Only one half of all industries experienced price inflation within an 8.7% to 45.3% band (cumulatively) over the sample period

Over the same period, input prices rose on average by 36.5%, with one half of the increase occurring over the last 2 years. This is twice as fast as output prices increases, but the difference across industries is more modest here—partially because of the higher aggregation at which we calculate input price growth. The input deflators are weighted averages of output deflators, with a much higher weight on industries producing raw materials and energy products than these industries represent in the overall economy. These products saw especially rapid price increases after 2003.

One potential source of measurement error is the input prices faced by export processors. These firms are allowed to import raw materials and intermediates duty-free. In the years preceding and following China's entry into the WTO, in 2001, import tariffs came down, also on intermediate goods and we expect this to be reflected in prices. In principle, intermediate goods prices for export processers should not have been affected, but we have no way to construct alternative indices for them. Input price inflation is likely to be biased upward for these firms, leading to an underestimate of real input use and an overestimate of value-added and productivity.

3.3 Firm-level real capital stock

A weakness of the Chinese data is the unusual way the capital stock is measured. In each year fixed assets are reported three ways: (i) "original fixed assets" is the sum of past investments at historical prices, (ii) "net" is original fixed assets less accumulated depreciation, (iii) "total" is net fixed assets with construction materials and ongoing construction added.

These book values sum nominal values for different years and should not be used directly. We make a number of assumptions to convert this information into a real value of the capital stock that is more comparable across time and across firms. Failure to do so is likely to introduce a systematic bias into the capital stock measure with respect to a firm's age.

Our procedure begins with estimating the real value of the capital stock in the first year that a firm appears in our data set. For simplicity of exposition, we assume this is 1998, the first year of our panel. In the absence of information on a firm's past investments and depreciation, we use information from the 1993 annual enterprise survey to construct estimates of the average rate of growth of the nominal capital stock between 1993 and 1998 at the two-digit industry level by province.[17] Combined with information on the age of each firm, these estimates are used to calculate the nominal capital stock in the firm's startup year. The real capital stock for that year is obtained by deflating with the investment deflator.

The nominal capital stock up through 1998 is then calculated by multiplying the firm's initial capital stock with the average sector-province growth rate for the number of years since the firm was established. Annual investment is the change in nominal capital stock between years plus depreciation, assumed to run at 9% annually. The real capital stock for 1998 is calculated using the perpetual inventory method, using the same depreciation rate and deflating annual investment. We continue this procedure for years after 1998, only using the observed change in the firm's nominal capital stock at original purchase prices as our estimate of nominal fixed investment.

In Table 5 we explore whether the capital stock thus obtained shows markedly different patterns from the net fixed asset value that is reported directly in the data set. In adjacent rows we compare for two groups of firms the capital-labor ratios constructed using either capital stock variable. The absolute values generated by our procedure tends to be higher, compare an average of 36.8 thousand RMB in real capital per worker in 1998 with 27.6 thousand RMB in net fixed assets in the same year. The average growth rate between 1998 and 2008 is very similar using both measures: 44% or 43% over ten years or 3.7% annually.

[Table 5 approximately here]

For other comparisons, however, the differences are larger. State owned firms appear to employ 19% (=1/0.81) more capital per worker using the book value, but 24% more using the real capital measure. Somewhat counterintuitively, small firms with fewer than 100 employees work with 9% more reported capital, but this gap disappears almost entirely using the calculated capital stock. Finally, and most strikingly, it is no surprise that incumbents have more capital per employee than new entrants, but while the difference is only 21% in the reported series, it is almost one half higher using the calculated real capital stock per worker.

In the right columns of Table 5, we make similar comparisons, but using the ratio of capital over value added. While the capital-labor ratio reflects only input substitution, the capital-value added ratio is also influenced by productivity differences. Lower numbers indicate a more efficient use of the capital stock, i.e. capital productivity. The most interesting discrepancy is in the comparison in the bottom two rows. While incumbents show a higher capital productivity using book values, entrants appear 10% more productive using the calculated capital stock. In two of the three other comparisons the capital productivity gap is merely accentuated when we use the new capital measure, while the change over time is more or less invariant, again, to the measure used.

[17] For firms entering after 1998 we use the nominal rate of growth in the capital stock from 1993 to their entry year.

3.4 Variables that changed definition or content over time

(a) Industry classification

Each firm is classified into an industry following the four-digit Chinese Industry Classification (CIC) system that resembles the older U.S. SIC system. The first Chinese industry classification was published in 1984 and it was revised in 1994, 2002 and 2011. Compared to previous versions, the 2003 CIC system (based on the 2002 revision) incorporates more detail for some industries, while other industries were merged. In several cases, the numeric code even changed without any change in coverage. To make the industry codes consistent across the entire period, we constructed a harmonized classification that groups some industries prior to or following the revision. The new classification covers fewer industries, a total of 458, and is made available online

In spite of this change, we still find that a large fraction of firms switch industries over time. As in the United States, a firm is classified into the sector of its main product by sales revenue. In the dynamic Chinese economy and with rapidly expanding export sales, it is natural to see more sectoral changes than in other countries. Statistics in Table 6 demonstrate that 18% of all firms change their four-digit sector affiliation at least once and this fraction is strongly increasing in the number of years we observe a firm in the sample. For firms observed at least six years, the likelihood they experience a sectoral switch surpasses 40%. Naturally, switches are more substantial and less common between industries that are defined at a higher level of aggregation. Even at the two-digit level, one tenth of all firms are observed to switch industries and more than one fifth of firms that are observed at least six years.

[Table 6 approximately here]

(b) Ownership

An important variable used in many studies is the firm's registered type (*qiye dengji zhuce leixing*). It distinguishes 23 exhaustive ownership types, which includes joint ventures between different types of owners. In Brandt et al. (2012) we classified each firm into one of five basic groups: state owned, hybrid or collective, private, and two types of foreign firms, those from Hong-Kong, Macau, and Taiwan and those from all other countries.[18] Foreign categories include both joint ventures and wholly-owned subsidiaries.

A major evolution in the Chinese economy is the growing importance of private firms. This is due both to new firms being predominantly private as well as firms changing ownership. One quarter of all firms move between one of the detailed categories and 17% even change between one of the five broadly defined groups. By far the most common switch is from the

[18] Hong Kong-based subsidiaries of foreign multinationals are included in the first foreign category as there is no straightforward way to distinguish them from other Hong Kong firms.

hybrid/collective category—which accounted for 38% of all firms in 1998, but for less than 6% of the total in 2008—to private.

An alternative approach to classify firms into different types is to use information on the registered capital. The share of each firm's "paid-up" capital is reported separately for six types of owners: state, collective, individual, legal person, Hong Kong-Macau-Taiwan, and foreign. One problem is that the "legal person" category can capture a wide range of possibilities—from investment stakes of state-controlled shareholding companies to private subsidiaries. In 1998 this category accounted for 18% of all paid-up capital, but it increased to 33% in 2008, the strongest increase of any of the six types. Gao and Van Biesebroeck (2013) show that the two alternative ways of identifying firms that were initially state owned—and thus directly exposed to the subsequent restructuring—has a pronounced impact on the estimated effect of reform in the electricity sector.

(c) Employment and wages

Two more variables that sometimes raise questions are employment and wage costs. When the economic reforms spread to the manufacturing SOEs in the urban sector in the mid-1990s, many workers were laid off and sent home, or *xiagang*. Initially, these workers remained on the payroll of their former employers, albeit at reduced wages and benefits, and were still included in NBS estimates of firm employment. By the late 1990s, the status of many of these individuals had been reclassified, which helps explain the sharp drop in employment in national figures in SOEs during this period. Some of the decline in total industrial employment we observe in our sample—from 58 million in 1999 to 53 million in 2001—probably still reflects these workers being dropped from the ranks of the employed in company records.

In contrast, there are two reasons to believe that firms may under-report the total number of workers. First, firms often pay taxes and fees to the labor department of the local government that are proportional to their total employment which provides an incentive for under-reporting. This problem is more serious in coastal regions where migrant workers account for a larger share of employment.

Second, it is common in China for firms may hire workers through a third party, called "labor dispatching" (*laodong paiqian*). The dispatched workers are employed by the labor dispatching companies and sent to industrial firms. While these workers are supposed to take only temporary jobs, many work for the firms for a long time. Firms "rent" the dispatched workers to increase flexibility and hold down labor costs. Because the dispatched workers are not official workers of the firms, the NBS survey does not include them. It was estimated that there were 37 million dispatched workers in 2011.[19] In some firms, the majority of the work force consists of the dispatched workers. The government revised labor contract law in 2012 to restrict this practice.

[19] Source: "Revision of Labor Contract Law," Xinhua News Agency, December 28, 2012.

One way this under-reporting shows up is through a lower than expected share of wages in value added. This averages only 32% across the firms in the sample, while in China's national accounts the wage share in GDP is almost 50%.[20] While the industrial sectors are expected to have lower than average labor-intensity, the observed average seems implausibly low.

Some aspects of worker compensation are also not reported throughout the entire period. Unemployment insurance and welfare expenditures are reported in every year, but pension contributions only since 2003, and housing subsidy since 2004. Together they only make up a small fraction of total worker compensation, on average 3.5% in 2007. The many reported zero values for these variables could indicate that total worker compensation, and thus the labor share in value added, is underreported.

Moreover, there is a further decrease in the wage share in value added over time, which is the result of three evolutions. First, average firm size is declining while the wage share is strongly increasing in firm size, as is the case in other countries. Second, the share of wages in state owned firms, which initially was much higher than average, converged over time towards the lower average recorded by private firms. And third, domestically owned private firms, which always had the lowest average wage share, increase in importance.

(d) Value added

Value added is not reported directly by firms. It is calculated by NBS using the expenditure approach:

$$\textit{value added} = \textit{output} \; - \; \textit{intermediate input} \; + \; \textit{value added tax payable.}^{21}$$

Holz (2008) shows that industrial value added is likely to be overestimated. Zhu and Qian (2012) investigate why the Chinese labor share for manufacturing firms is that much lower than in other countries and also argue that value added of the above-scale firms may be overestimated. An upward bias in measured value added will lead to an overestimation of TFP and if the bias changes over time, TFP growth estimates can also be affected.

In the first column of Table 7 we list industry GDP as reported in China Statistical Yearbook (2010), which is one component of GDP in the national accounts. In the second column we list aggregate value added summed over all above-scale industrial firms in our sample, which is almost identical to the reported numbers in the Statistical Yearbook (except for 2004). The two series converged markedly over the 1998-2007 period, but this implies that the latter series has grown much more rapidly. In 2008, the NBS stopped reporting value added in the firm-level data files, as well as in the Statistical Yearbook and the Census Yearbook. A possible reason for

[20] Hsieh and Klenow (2009) face the same discrepancy and they inflate wage payments by a constant factor for all firms to obtain a wage share in value added consistent with the national average.

[21] The last term equals zero if the value added tax payable is negative. In 2004, we use the same identity to construct output which was not reported directly in that year.

discontinuing this reporting is the growing inconsistency in the value-added estimates from the above-scale firm survey and China's national accounts, which relies on data from a separate firm-level survey.

Zhu and Qian (2012) investigate to what extent value added estimates differ if they are calculated using the (factor) income approach at the firm level. Unfortunately, as discussed previously, data on labor compensation are likely incomplete, which now introduces a downward bias in the value added estimates. Despite this shortcoming, we construct an alternative value added measure by adding four components: labor compensation, net indirect taxes (indirect taxes minus government subsidies), profit and depreciation.[22]

The last columns in Table 7 list both the alternative value added aggregates and the components. The new aggregate is lower than value added using the expenditure approach and the ratio of the expenditure to income value added tends to increase over time. It could be that the upward bias in the expenditure approach became more serious or that the underreporting of labor income in the income approach became more serious, or a bit of both.

[Table 7 approximately here]

(e) Exports

Prior to 2004, many private firms could only export through third parties, i.e. trade intermediaries. At the start of the sample, only very few firms, in particular state owned firms, multinationals, and very large exporters, had direct trading privileges. As a condition of China's WTO accession, the system was liberalized and from 2004 onwards every firm had in principle the right to trade directly—both to import or export.

When observations in the above-scale NBS firm-level data are linked to the Chinese information on trade transactions, the share of exports (and exporters) that can be matched is indeed increasing over time.[23] Nevertheless, a sizeable fraction of exports are by firms that cannot be found in the firm-level data. Imperfection in the matching process is one possible explanation, but there are likely to be many firms that continue to export indirectly even after 2004 as they had already developed an effective relationship with a trade intermediary. In addition, there are firms that report positive exports, but which cannot be found in the trade transaction data. It is highly likely that some indirect exporters report positive export levels, even though they only sell directly to domestic trading firms.

[22] Labor compensation consists of wage, unemployment insurance, welfare expenditures, pension contributions (after 2003) and housing subsidy (after 2004). Indirect taxes consist of three accounting items in our data: sales tax, value added tax and "other taxes under management expenses".

[23] We matched the two data sources by firm name, verifying several permutations of the generic characters in Chinese language firm names.

Statistics in the last two columns of Table 6 illustrate that 13.6% of firms are exporting continuously. For firms that are active over the entire sample period—which tend to be larger and predominantly state owned firms—this fraction is almost 20%. In addition, 9.6% of firms start exporting at some point. For firms operating throughout the sample period this even stands at 27.6%. By 2008, almost half of them are exporters. Among firms active between six and ten years, 36% are exporters at some point during the sample period, a much higher fraction than for other large economies.

4. Productivity gap between entrants and incumbents

The importance of the different measurement and sampling problems we discussed will naturally depend on the application. In this section we provide an illustrative example by documenting the evolution over time of the relative productivity gap between new entrants and incumbents.

We calculate total factor productivity levels using the index number method also used in Brandt et al. (2012). It is a Solow residual that uses the average of the firm-specific and the industry-average input shares in output as weight when subtracting input differences from the output difference. We enforce constant returns to scale and calculate the capital share as the sum of labor and material shares. Year by year, firm-level productivity is regressed on a set of provincial and ownership dummies as well as an entry dummy. This dummy takes the value of one if it is the first year we observe a firm and the firm’s reported startup year is at most one year prior to the current year.

If we deal with measurement issues the way we have described thus far, the benchmark estimates reveal that the relative performance of new entrants at the time of entry has declined noticeably over time. This pattern is illustrated by the solid black line in Figure 2 and the dashed band shows the 95% confidence interval. In 1999, the average entrant was slightly more productive than the average incumbent. The difference even was significant in a statistical sense, but at 2% the economic difference was negligible. By 2008, this difference had declined to a negative gap of approximately 9%, which is estimated very precisely. One way of explaining this evolution is by a changing selection process driven by falling entry barriers. Initially, only the most productive potential entrants were able to overcome entry restrictions. Over time, this constraint relaxed and the average entrant in China now underperforms the average incumbent at the time of its entry, just as in other countries.

[Figure 2 approximately here]

The other two lines in Figure 2 illustrate that the results would come out differently without some of the adjustments we have advocated. For example, the red dashed line shows the evolution of the productivity gap if the reported net value of fixed assets, directly from the company accounts, is used rather than the constructed real capital stock series. This adjustment has a minor impact on the productivity of entrants, but it lowers the capital stock for incumbents and hence raises their measured total factor productivity. This is especially true for incumbents that have been observed for many years in the sample, i.e. in later years. As a result, the pattern of a falling relative productivity level for new entrants becomes even more pronounced.

The adjustment goes in the opposite direction if we do not condition on the reported startup year and count all firms as entrants in the first year that we observe them in the sample. This erroneously classifies several older firms as entrants—the number of entrants almost doubles. On average, these older "entrants" have higher productivity levels than genuinely new firms. As a result, the pattern of a falling productivity level for entrants largely disappears. The green dashed line in Figure 2 moves somewhat erratically and the spike in 2004 is particularly worrying as we know that in that year many older firms entered the sample.

Yet another adjustment we could make is to ignore the linkages using other information than NBS firm identifiers. Firms whose ID changes following a restructuring would then be treated as exiting from the sample, while a new firm with a new ID enters the following year. Counting restructured firms as entrants again biases the average productivity of that group upwards. It would show up as a change in the same direction as the green line (not reported), but the difference with the black line is less pronounced.

Entry rates vary substantially by region and it is sometimes suggested that this is related to differences in the quality of local economic institutions or different degrees of opening-up. Figure 3 illustrates the relationship between the quality of local institutions and firm entry rates at the provincial level for 2004. The horizontal axis shows the marketization index constructed by Fan, Wang and Zhu (2010) which aggregates information on 19 indicators of institutional quality and 5 major areas of the market-oriented reforms. The vertical axis shows the number of newly entered firms as a fraction of the total number of incumbents using the 2004 firm census which has no minimum sales threshold for inclusion. There is a statistically significant positive relationship between the two variables, consistent with the hypothesis that institutional reform leads to higher rates of firm entry.

[Figure 3 approximately here]

5. Topics for future research

The NBS firm-level data has already been used extensively to study several aspects of firm dynamics in China, such as the process of entry, exit, and firm growth, both in terms of size and productivity. Several more studies evaluate the productivity gap and other dimensions of performance differences between state owned and private firms or between exporters and firms only selling domestically. Following Hsieh and Klenow (2009), factor market restrictions that lead to input misallocation and depress aggregate output have also received some attention. Researchers broadly in the fields of industrial organization, international trade, and macroeconomics have studied several more topics.

We believe, however, that the firm-level data is underexploited in several other fields. For example, in public economics the far-reaching reform of the tax system in 1995 has received a lot of research attention at the aggregate level, but it has not been studied yet at the micro level. In the firm-level data, we observe a wide range of taxes: income tax (levied on profits) and government subsidies, value added tax, indirect taxes on output and inputs, as well as several items of worker compensation that in other countries are covered by payroll taxes, e.g. pension contributions. An interesting feature is that the relative importance of the different taxes has changed over time. Other interesting topics would be to estimate the difference in the burden of taxation across firm types or the impact of tax exemptions that many multinational subsidiaries have received.

A second research area where this data could inform ongoing debates is the field of corporate finance. Several researchers have used company accounts for firms listed on the Shanghai or Shenzhen stock markets to study the performance of firms after they become listed or to study the relationship between financial performance and stock price evolution. Industrial groups that are listed are also included in this data set, in some cases both prior and following their stock market listing, and their economic performance can also be compared to non-listed firms. For state owned firms, the registration type variable indicates whether the firm has been reformed into a shareholding company or not, which is an interesting process in itself. The relationship between changes in the ownership structure, as evidenced by the share of the six types of investors in the paid-up capital, and firm performance also provides interesting information on the importance of governance.

Finally, some of the measurement issues we have discussed would benefit from additional scrutiny. The measurement of value added has generated questions that the statisticians at NBS also do not have good answers for. The extremely low value of labor income in value added is a related puzzle. It might reflect the reality in Chinese manufacturing or it might reflect remaining measurement problems. The development of detailed deflators covering the entire period, would be a third instance where most researchers using the firm-level data would benefit. In many applications an easy solution is to simply include industry-year interaction fixed effects in a regression, but sometimes this is not possible. Less than ideal double-deflation of output and inputs, which seem to be subject to different price evolutions, could also be a contributing factor to the declining wage share in income.

We do not want to suggest that these measurement problems are insurmountable or that they preclude rigorous analysis of economic phenomena. Making the firm-level annual surveys or censuses available and useful for economic research has taken many countries several decades. This is an ongoing process where many of the solutions can be and should be continuously improved. We believe that the best way to make progress is to tackle interesting economic questions and resolve the data problems that are most important for the question at hand. In this way, we can decentralize problem solving activities and benefit from each other's work. We hope that this paper provides one step in that direction.

References

Aghion, P., M. Dewatripont, L. Du, A. Harrison, and P. Legros (2012). “Industrial policy and competition.” NBER Working Paper No. 18048.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta (2009). “Measuring and analyzing cross-country differences in firm dynamics.” In T. Dunne, J. B. Jensen, and M. J. Roberts (eds.), *Producer Dynamics: New Evidence from Micro Data*, University of Chicago Press.

Baum-Snow, N., L. Brandt, J.V. Henderson, M.A. Turner, and Q. Zhang (2013). “Roads, railroads and decentralization of Chinese cities.” Working Paper.

Brandt, L., J. Van Biesebroeck, and Y. Zhang (2012). “Creative accounting or creative destruction? Firm-level productivity growth in Chinese manufacturing.” *Journal of Development Economics*, 97(2): 339-351.

Brandt, L., J. Van Biesebroeck, L. Wang, and Y. Zhang (2012). “The Impact of Entry into WTO on Chinese Enterprise Productivity,” CEPR Discussion Paper No. 9166.

Dougherty, S., R. Herd and P. He (2007). “Has a private sector emerged in China’s industry? Evidence from a quarter of a million Chinese firms.” *China Economic Review,* 18: 309-334.

Fan, G., X. Wang, and H. Zhu (2010). *China’s marketization index*, Economic Science Press, Beijing.

Gao, H. and J. Van Biesebroeck (2013). “Effects of deregulation and vertical unbundling on the performance of China's electricity generation sector.” *Journal of Industrial Economics,* forthcoming.

Holz, C. (2008). “How can a subset of industry produce more output than all of industry?" Working paper, Hong Kong University of Science and Technology.

Hsieh, C.-T. and P. J. Klenow (2009). “Misallocation and manufacturing TFP in China and India.” *Quarterly Journal of Economics,* 74(4): 1403-1448.

Jefferson, G. H., T.G. Rawski, and Y. Zhang (2008). “Productivity growth and convergence across China’s industrial economy.” *Journal of Chinese Economic and Business Studies,* 6(2): 121-140.

Park, A., D. Yang, X. Shi, and Y. Jiang (2010). “Exporting and firm performance: Chinese exporters and the Asian financial crisis.” *Review of Economics and Statistics,* 92(4): 833-842.

Perkins, D.H. and T.G. Rawski (2008). “Forecasting China's economic growth.” In: L. Brandt and T.G. Rawski (Eds.), *China's Great Economic Transformation*. Cambridge University Press, NY.

Song, Z., K. Storesletten, and F. Zilibotti (2011). “Growing like China.” *American Economic Review,* 101(1): 196-233.

Young, A. (2003). “Gold into base metals: Productivity growth in the People’s Republic of China during the reform period.” *Journal of Political Economy,* 111(6): 1220-1261.

Zhu, X. and Z. Qian (2012). “Misallocation or mismeasurement? Factor income shares and factor market distortions in China’s manufacturing industries.” Working paper, University of Toronto.

Table 1: Comparison of firm-level sample with China Statistical Yearbook and Census

(a) Firm-level data set

Year	Number of firms	Sales	Output	Value added	Employment	Net value of fixed assets (original value)	Export
1998	165,118	6.41	6.77	1.94	56.44	4.41	1.08
1999	162,033	6.99	7.27	2.16	58.05	4.73	1.16
2000	162,883	8.42	8.57	2.54	53.68	5.18	1.46
2001	169,030	9.24	9.41	2.79	52.97	5.45	1.61
2002	181,557	10.95	11.08	3.30	55.21	5.95	2.01
2003	196,222	14.32	14.23	4.20	57.49	6.61	2.69
2004	279,092	20.43	20.16*	6.62	66.27	7.97 (12.54)	4.05
2005	271,835	24.69	25.16	7.22	68.96	8.95	4.77
2006	301,961	31.36	31.66	9.11	73.58	10.58	6.05
2007	336,768	39.97	40.52	11.70	78.75	12.34	7.34
2008	327,853	39.04	38.77	11.82*	69.08	12.02	6.35

(b) China Statistical Yearbook (2009, Table 13-4): above-scale industrial firms

Year	Number of firms	Sales	Output	Value added	Employment	Net value of fixed assets	Export
1998	165,080	6.41	6.77	1.94	61.96	4.41	1.08
1999	162,033	6.99	7.27	2.16	58.05	4.73	1.15
2000	162,885	8.42	8.57	2.54	55.59	5.18	1.46
2001	171,256	9.37	9.54	2.83	54.41	5.54	1.62
2002	181,557	10.95	11.08	3.30	55.21	5.95	2.01
2003	196,222	14.32	14.23	4.20	57.49	6.61	2.69
2004	276,474	18.78	20.17	5.48	66.22	7.38	4.05
2005	271,835	24.46	25.16	7.21	67.85	8.81	4.79
2006	301,961	31.36	31.66	9.11	73.58	10.58	5.96
2007	336,768	39.97	40.52	11.70	78.75	12.34	7.34
2008	426,113		50.74		88.38	15.17	8.25

(c) Census Yearbooks

Year	Number of firms	Sales	Output	Employment	Original value of fixed assets	Export
Above Scale						
1995	162,114	4.85		66.82	4.17	
2004	279,040	20.42	20.17	66.22	12.58	4.05
2008	426,113		50.74	88.38	24.53	8.25
Below Scale						
1995	331,090	0.35		16.88	0.26	
2004	1,098,789	1.98	2.06	26.82	1.24	0.11
2008	1,477,267		3.66	31.69	2.20	
Fraction above scale						
1995	32.9%	93.3%		79.8%	94.1%	
2004	20.3%	91.2%	90.7%	71.2%	91.0%	97.5%
2008	22.4%		93.3%	73.6%	91.8%	

Notes: Above scale means SOEs and all other firms with sales above 5 million RMB. The export information in panel (b) is taken from China Statistical Abstract (2009, Table 10-2). All values denoted in trillions RMB and employment in millions of workers. Industrial sector covers mining, manufacturing, and utilities. Variables indicated by * are imputed.

Table 2: Variables reported in the annual NBS above-scale firm-level survey

(almost) Always reported (1998-2008)	
Identifying information:	ID, name, registration type, shareholding status, legal person name, industry code, geographic code, zip code, street, phone, start year, start month, "lishu" supervising body, industrial activity units[a]
Stocks:	Capital structure: owners' equity, paid-up capital -- split into six exhaustive categories: state, collective, foreign, HMT, individual, legal person Assets: total, current, average current, fixed assets (separately: total, net value, at original price), intangible[c], inventory, accounts receivable Debt: total, current, long-term
Flows:	Output: sales revenue, output value[b], of which industrial output[b] or new products[b], value added[c], exports, total profit, operation profit, production, top 3 products Input factors: employment, wages, materials and intermediate inputs, long-term investment, depreciation (current year & cumulative), property insurance, financial cost, interest expense, management cost Taxes: income tax payable, value added tax payable, sales tax, input tax, subsidy[bc]
More detailed breakdowns added since 2003 or 2004:	
- Sales, costs, and financial performance now reported separately for operational activities and for main line of business	
- Salary expenditures: pension, housing subsidy, welfare, labor and unemployment insurance	
- Additional cost categories: advertising, transportation, administration, staff training,..	
- Other: cash inflows and outflows, end-of-year employment, accounts payable, female workers	
Noteworthy variables:	
1998-2003:	Output at constant prices (in addition to usual current prices)
2001-2007:	R&D (not 2004)
Only 2004:	- Breakdown of employment (total and female) by: 5 education levels: postgraduate, university, college, high school, primary or less 3 technical titles (not exhaustive): senior, intermediate, junior 4 categories (not exhaustive): technician, senior tech., adv. engineer, mid-level worker - Number of computers, number of microcomputers - Special costs (unionization, trips, pollution), net cash flow of operational, financial, and investment activities - Parent firm, parent firm ID, registration ID

Notes: A number of variables that are reported only occasionally are omitted. [a] not in 2001, [b] not in 2004, [c] not in 2008

Table 3: Linking observations over time to identify firms

Years in the sample	Fraction of …		Fraction of links established not by ID	
	Observations	Firms	Observations	Firms[a]
1	7.2%	27.9%		
2-3	17.6%	27.6%	2.7%	3.8%
4-5	25.1%	21.3%	2.5%	8.0%
6-10	36.1%	18.3%	3.9%	21.1%
11	14.0%	4.9%	3.8%	29.8%
Total	2,539,519	660,510	3.3%	11.2%

Notes: [a] Fraction of firms with a year-on-year link established not by ID at least once over the observed period.

Table 4: Evolution of the panel over time

	Active	Entrants	Linked using		Exiting	Fraction of active firms			
	firms		NBS ID	other	(next year)	Entrant[a]	ID	Other	Exiting
				informatio			link	link	
1998	163,295				23,276				14.3%
1999	159,694	21,860	131,198	6,636	23,042	13.7%	82.2%	4.2%	14.4%
2000	160,956	24,183	132,949	3,824	32,256	15.0%	82.6%	2.4%	20.0%
2001	167,259	40,115	119,662	7,482	19,676	24.0%	71.5%	4.5%	11.8%
2002	179,811	30,733	144,238	4,840	21,621	17.1%	80.2%	2.7%	12.0%
2003	195,737	38,362	149,975	7,400	33,441	19.6%	76.6%	3.8%	17.1%
2004	278,183	115,667	146,848	15,668	39,518	41.6%	52.8%	5.6%	14.2%
2005	270,997	33,756	232,275	4,966	21,152	12.5%	85.7%	1.8%	7.8%
2006	301,165	48,849	248,245	4,071	23,663	16.2%	82.4%	1.4%	7.9%
2007	335,812	58,375	274,000	3,437	96,255	17.4%	81.6%	1.0%	28.7%
2008	326,610	85,315	236,919	4,376		26.1%	72.5%	1.3%	

Notes: [a] Firms that exit and later re-enter the sample (less than 1% of firms) are considered to be operating throughout. Entrant here means entering into the sample. To identify entrants into the industry, one should also incorporate the information about a firm's startup year.

Table 5: Comparison of capital stock estimates

	capital-labor ratio				capital-value added ratio			
	constructed capital	(diff.)	net fixed assets	(diff.)	constructed capital	(inverse diff.)	net fixed assets	(inverse diff.)
1998	36.8		27.6		1.71		1.26	
2008	52.9	(+44%)	39.6	(+43%)	0.52	(+229%)	0.38	(+232%)
SOE (1998)	41.4		30.4		3.26		2.33	
other (1998)	33.5	(-19%)	25.6	(-16%)	1.19	(+174%)	0.90	(+159%)
small (1998)	37.1		29.1		5.15		73.65	
large (1998)	36.6	(-1%)	26.8	(-8%)	4.57	(+13%)	69.87	(+5%)
entrant (2008)	39.0		34.3		0.48		0.41	
incumbent (2008)	58.3	(+49%)	41.3	(+21%)	0.53	(-10%)	0.38	(+8%)

Notes: Capital per worker is denoted in thousands of RMB (in 1998 prices). "Net fixed assets" is fixed assets at original purchase prices less depreciation, as reported in the data set, which we deflated using the investment deflator. Percentage changes on the capital-value added ratio are reported for the inverse (capital productivity).

Table 6: Fraction of firms that change status

the	Number	Fraction changing industry		Fraction changing "type"		Fraction exporting	
sample	of firms	4-digit	2-digit	23 types	5 types	Always	Starting
1	184,098					13.2%	
2-3	182,624	13.9%	7.6%	19.0%	11.8%	12.1%	5.3%
4-5	140,659	21.9%	12.1%	32.0%	21.2%	13.9%	13.1%
6-10	120,851	40.8%	22.8%	54.4%	39.0%	14.5%	21.9%
11	32,278	41.4%	22.6%	67.2%	49.9%	19.5%	27.6%
Total	660,510	18.0%	10.0%	25.3%	17.3%	13.6%	9.6%

Notes: "Industry" according to adjusted Chinese Industry Classification system that was made time-consistent; "type" according to 23 original entries in the NBS classification of registered enterprise type or 5 aggregate types (SOE, hybrid, private, HMT, foreign).

Table 7: Alternative approaches to calculate value added

	Total value added						
Year	Industry GDP (China Statistical Yearbook 2010)	Expenditure approach (firm-level data)	Income approach (firm-level data)	Labor Compensation	Net Indirect Taxes	Profit	Depreciation
1998	3402	1941	1446	591	392	146	317
1999	3586	2156	1624	603	429	230	363
2000	4003	2539	1984	648	496	439	400
2001	4358	2790	2118	668	535	469	445
2002	4743	3299	2424	749	603	579	494
2003	5495	4199	3090	921	750	834	586
2004	6521	6620[a]	4156	1299	915	1194	748
2005	7723	7218	5074	1533	1112	1479	950
2006	9131	9110	6384	1896	1407	1950	1131
2007	11053	11700	8254	2352	1824	2715	1363

Notes: [a] This was the only value where our firm-level aggregate differed noticeably from the reported value for above-scale firms in China's Statistical Yearbook (compare panels (a) and (b) in Table 1). All values reported in million RMB. Labor compensation includes wage, unemployment insurance, welfare expenditures, pension contributions (after 2003) and housing subsidy (after 2004). Net indirect taxes equal indirect taxes minus government subsidy. The indirect taxes covers three Chinese accounting items in our data: sales tax, value added tax and other taxes under management expenses.

Figure 1: Output, input, and investment deflators

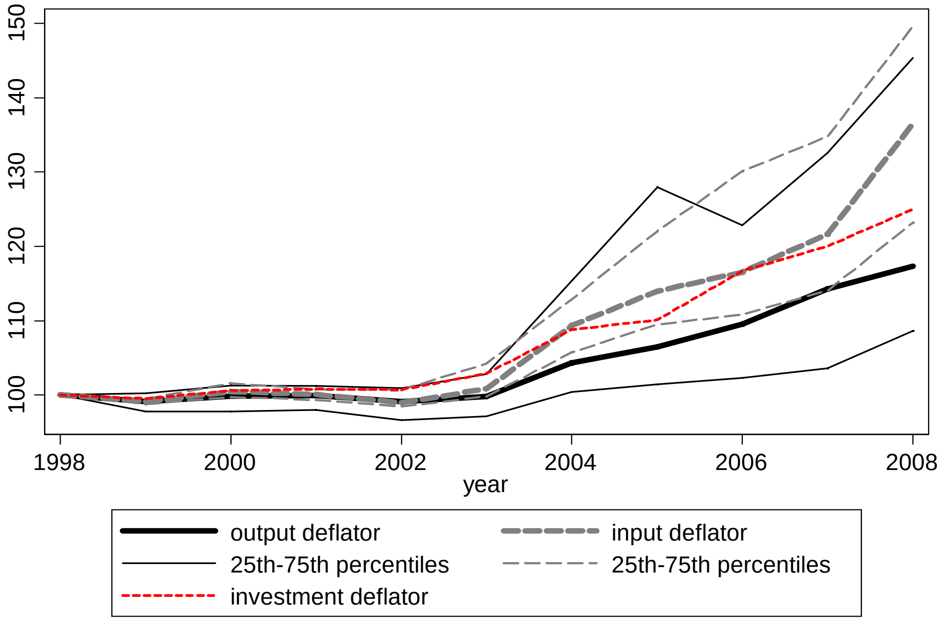


Figure 2: Productivity difference between entrants and incumbents

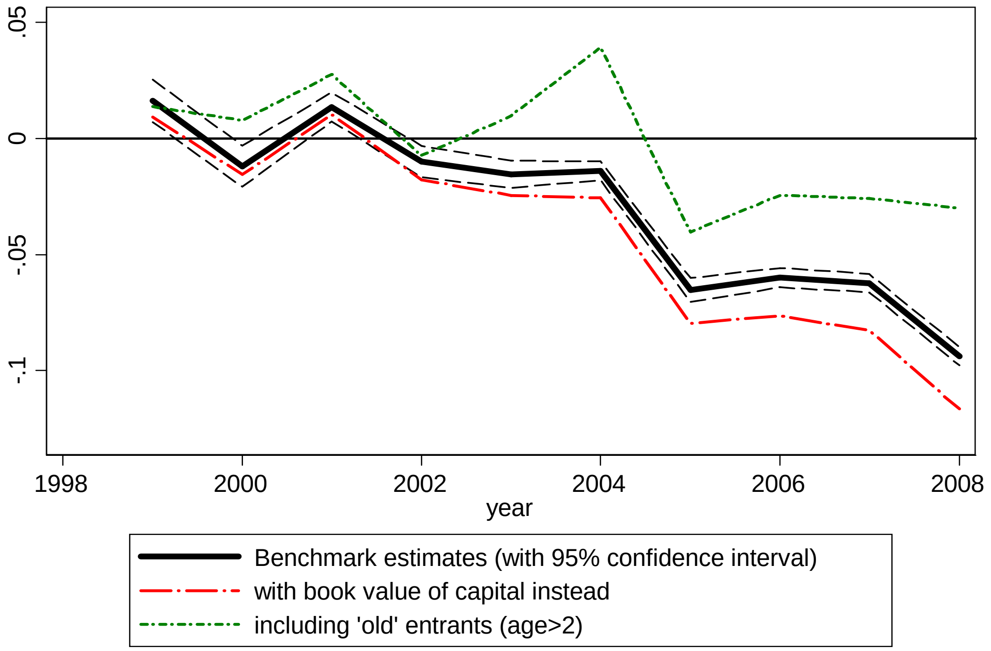
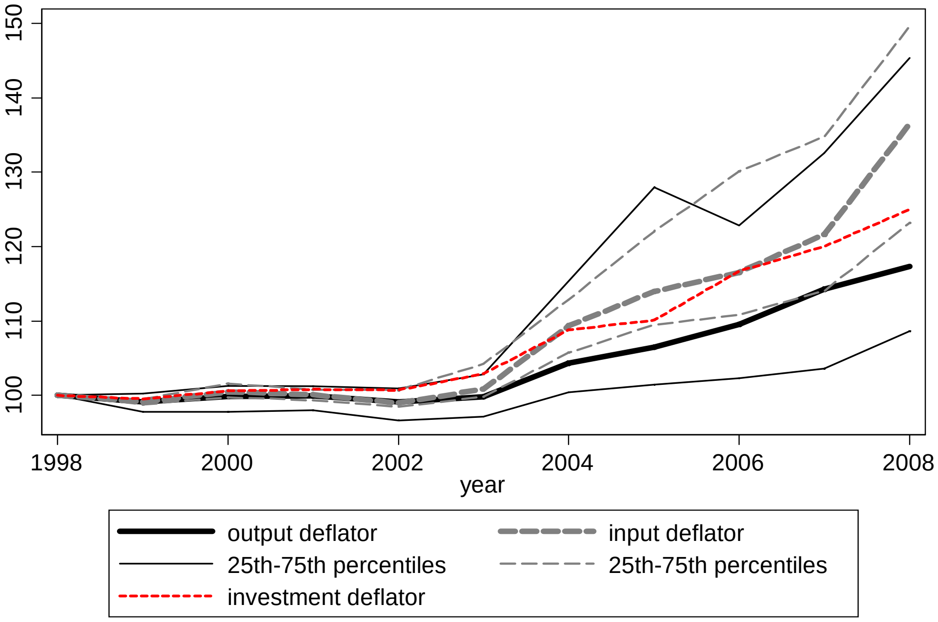
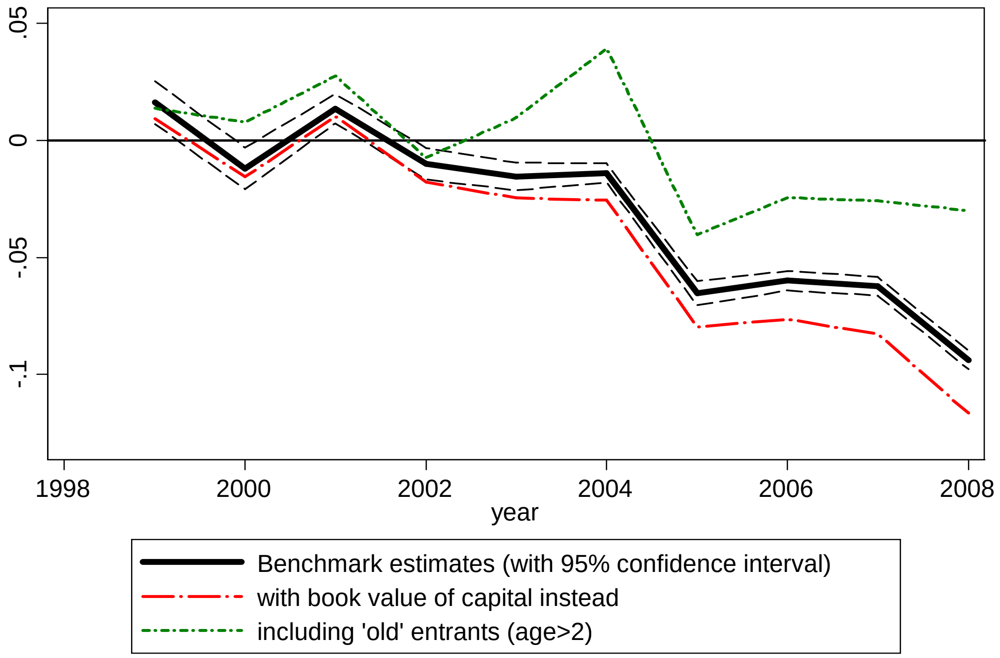


Figure 1: Output, input, and investment deflators



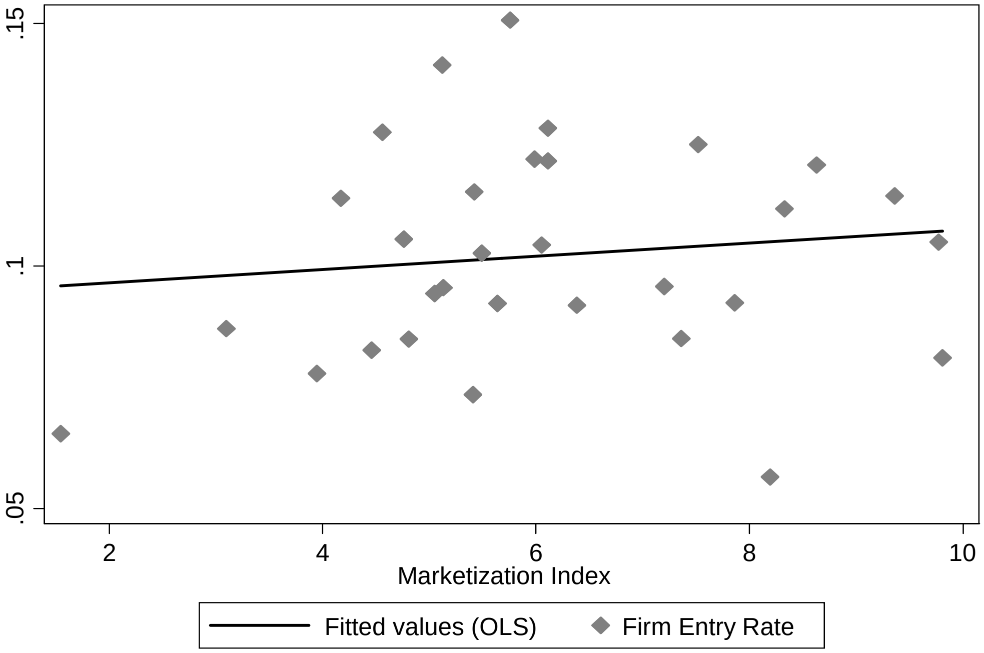
Notes: The output deflator is calculated from firm-level reports of nominal and real sales in the annual NBS above-scale industrial firm surveys (1998-2003). The series is extrapolated to 2008 using the 2-digit ex-factory price index from the China Statistical Yearbook. The input deflator is calculated by multiplying the output deflator (an industry-vector) with the 2002 National Input-Output table from China's NBS. The investment deflator is taken from Perkins and Rawski (2008).

Figure 2: Productivity difference between entrants and incumbents



Notes: Results are obtained from separate OLS regressions by year on the full sample of firms from the annual NBS above-scale industrial firm surveys (1998-2008). The dependent variable is total factor productivity and the statistics shown are regression coefficients and confidence intervals for the only explanatory variable of interest: an entry dummy for new firms.

Figure 3: Higher rates of firm entry in provinces that underwent broader institional reform



Notes: The Marketization Index measures the quality of local institutions and the extent of market-oriented reforms at the provincial level, taken from Fan, Wang and Zhu (2010). Firm entry rates are calculated from the 2004 firm census, which includes both above-scale and below-scale firms.