A Review and Reflection on the Use and Abuse of Chinese Industrial Enterprises Database

Nie, Huihua and Jiang, Ting and Yang, Rudai

School of Economics, Renmin University of China, School of Economics, Renmin University of China, Academy for Consumption Studies, Xiangtan University

1 November 2012

A Review and Reflection on the Use and Abuse of Chinese Industrial Enterprises Database*

Huihua NIE

(School of Economics, Renmin University of China, Beijing 100872; niehh@ruc.edu.cn)

Ting JIANG

(School of Economics, Renmin University of China, Beijing 100872; ting.jiang@gmail.com)

Rudai YANG

(Academy for Consumption Studies, Xiangtan University, Hunan 411105; rudaiyang@gmail.com)

Abstract: Empirical research has called for intensive use of disaggregated data. One of the most heavily used data set in studying the corporate behavior and performance in China is Chinese Industrial Enterprises Database. As we will show, this data set suffers from data matching problems as well as measurement errors, unrealistic outliers and definition ambiguities etc., all of which practically lead to research results thereupon that are at best questionable. In this article, we briefly summarize the data set, selectively review its uses in previous studies, address some critical issues regarding its usage, and propose some remedies and recommendations.

Keywords: Firm-level data, Industrial enterprises, Microeconometrics, Manufacturing sector, Productivity

JEL Classification: C33 D24 L22 L60

1 Introduction

Data is the cell of empirical research, thus the quality of data can directly decide the vitality of it. In the past ten years or more, international economics field has focused more and more on researches using longitudinal micro-level data. Compared to macroeconomic data or industrial data, micro-level data such as firm-level data and individual data has some significant advantages. First, longitudinal micro-level data contains more information, such as the ownership, scale and export of firms. This information is essential for analyzing firms' behavior. Second, longitudinal micro-level data contains both time dimension and individual dimension, which would help solve the individual heterogeneity problem and thus can ensure the consistency of the estimates. Third, longitudinal level-data increases the number of samples, making the estimates more efficient. For research fields including industrial organization theory, firm theory, corporate finance, international trade, income distribution, labor supply, etc., empirical researches mainly use microeconomic data.

With the introduction of microeconometrics and the access to domestic and

* This is a translation version of a Chinese article: 聂辉华、江艇和杨汝岱，2012，《中国工业企业数据库的使用现状和潜在问题》，《世界经济》，第 5 期，第 142-158 页. Citation: Nie, Huihua, Ting Jiang, and Rudai Yang, "A Review and Reflection on the Use and Abuse of Chinese Industrial Enterprises Database" (in Chinese), *World Economy*, 2012, no.5, pp. 142-158.

overseas micro databases, economists in China have paid more and more attention to the development and use of micro-level data and have achieved many research results based on it. Some databases about China are even being used by scholars from all over the world. On one hand, this phenomenon shows that Chinese problems have received more and more attention from international economics field. On the other hand, it indicates that the quality of these Chinese databases have gained more and more recognition. Especially, many scholars from home or abroad have used Chinese industrial enterprises database and published their research results on some famous international or Chinese journals including *American Economics Review*(e.g. Song, et al. 2011), *Quarterly Journal of Economics* (e.g. Hsieh and Klenow, 2009) and *Economic Research*. As a database constructed by National Bureau of Statistics of China, its advantages are large sample, many indices, and large time span. However, it is not published by any academic institutions, so in many aspects it can hardly meet the strict requirement of academic research. Its disadvantages include sample match problems, missing indices, unrealistic outliers, measurement errors and definition ambiguities. We can infer that if some researchers have not noticed these defects and taken effective methods to release or eliminate them, then these defects would negatively affect their empirical researches, or even lead to improper results. Therefore, we believe it is necessary to discuss the background and current use of Chinese industrial enterprises database detailedly and strictly, point out its problems and make our best to provide solutions. We hope that this paper can help potential researchers use this database more accurately as well as show them the current analyses and future directions, thus can improve researches in relevant fields. Of course, as users of this database, we cannot guarantee that we have a full picture about it. And some tendencies may be unavoidable in our analysis.

2 Basic Information of Chinese industrial enterprises database

We start by a brief introduction of the database. Chinese industrial enterprises database is constructed by National Bureau of Statistics of China. Its data mainly comes from the annual reports or quarterly reports that the sample enterprises submitted to the local Bureau of Statistics. The full name of this database is “all state-owned and above-scale non-state-owned industrial enterprises database”. Its statistical unit is business entity. Here, “industrial” includes “mining industry”, “manufacturing industry”, “production and supply industry of electricity, gas and water” in “industrial classification for national economic activities”, and “manufacturing industry” takes up 90% of the enterprises. Here, “above-scale” requires the main business income (that is, sale) of an enterprise to be no less than 5 million RMB, and this standard was revised to 20 million RMB in 2011. Database based on the above classification was collected first in 1998, but the database used by most scholars only involves between 1999 and 2007. Because most samples in this database belong to manufacturing industry, the classification is consistent with other countries, and some variables (for example, capital, R&D investment, and export value) are easier to measure, users of this database always only take samples of

manufacturing firms in it. Manufacturing industry includes 30 categories (two-digit industry) ranging from "agricultural product and byproduct processing", "food production to craftwork and other manufacturing" and" waste resources and materials recovering", whose codes are 13-43 (without 38) in "codes of industrial classification for national economic activities" (GB/T4754-2002). To keep the completeness of the samples, and making our analysis comparable with current studies, we will use all state-owned and above-scale non-state-owned manufacturing enterprises samples within 1999-2007as the main samples in our analysis of the database.

Chinese industrial enterprises database (1999-2007) contains over 2 million observations, and the per-year numbers of observations increase from 160 thousands in 1999 to 330 thousands in 2007. [1]During these 9 years, there are about 550 million enterprises (including public enterprises) appearing in the database. Obviously, it is a huge non-balanced panel data set. Because of closure, structure reform, restructuring, etc., only 46 thousand enterprises (take up 8% of the total number of sample enterprises) appears in every period continuously. The samples of this database account for most of manufacturing enterprises in China. According to the comparable annual report of the first national economic census in 2004, the total sales of manufacturing enterprises are 21844.281 billion RMB. And the total sales of all sample enterprises in Chinese industrial enterprises database that year are about 19560 billion RMB, accounting for about 89.5%[2]of the national sales. Now, Chinese industrial enterprises database is the largest database except the economic census database. Table 1 summarizes the enterprises number and the ratio change of state-owned, collectively-owned, private, foreign-invested (including Hong Kong, Macao, and Taiwan) enterprises during 1999-2007. We can see that the ratios of state-owned and collectively-owned enterprises has decreased significantly, dropping from 2/3 in 1999 to less than 1/10 in 2007. But the ratio of private enterprises has increased rapidly from less than 20% to more 70%. The table shows the tremendous change of the Chinese economic structure.

Table1 The type, number and ratio of Chinese manufacturing enterprises

year	State-owned	ratio%	Collectively-owned	ratio%	private	ratio%	foreign	ratio%	total
1999	52817	32.86	53507	33.29	27757	17.27	26652	16.58	160733
2000	44665	27.66	49383	30.58	39192	24.27	28240	17.49	161480
2001	36781	21.67	42528	25.06	59208	34.89	31178	18.37	169695
2002	31570	17.55	38237	21.25	75884	42.18	34208	19.02	179899
2003	25157	12.93	32334	16.62	98698	50.74	38318	19.70	194507
2004	27403	9.89	26896	9.70	165864	59.85	56976	20.56	277139
2005	18520	6.86	23875	8.84	171603	63.53	56112	20.77	270110
2006	16209	5.40	20983	6.99	202417	67.43	60585	20.18	300194

[1] Different researchers may get this database from different resources, but these versions have only a few differences.

[2] Sales of manufacturing enterprises comes from the website of National Bureau of Statistics "Report of the First National Economic Census (No.2)"; sales of manufacturing enterprises in Chinese industrial enterprises database comes from the authors' computation.

2007	11724	3.50	19355	5.78	236823	70.68	67174	20.05	335076

Source: computed by the authors according to the database

Actually, Chinese industrial enterprises database is also the most comprehensive database of enterprises. This database contains two kinds of information about enterprises: one is the basic information of the enterprises, and the other is the financial data of them. The basic information includes: code of certificate, name of the enterprise, legal representative, phone number, postal code, address, industry, registration type (ownership), administrative subordination, opening year, number of staff and workers, etc. The financial data includes: current assets, account receivable, long-term investment, fixed assets, accumulated depreciation, intangible assets, current liabilities, long-term liabilities, paid-up capital, prime operating revenue, operating costs, operating expenses, administrative expenses, finance charge, operating profits, profits, advertising fees, research and development expense, total salaries, total employee benefits, value added tax, industrial intermediate inputs, total industrial output value, value of export delivery, etc. There are about 130 indices in total. Especially, because 2004 is the first year of economic census, data of that year in the database also includes the numbers of male and female staff and workers of different education backgrounds (graduate, undergraduate, junior college, technical secondary school, high school, middle school and under) and different job titles (technical titles, technician and etc.). Moreover, data of 2004 also includes whether a enterprise joined Worker's Union, the number of people joining Worker's Union and some other information which is not included in data of other years.

Undoubtedly, Chinese industrial enterprises database has some obvious advantages. First, it has a very large sample size, covering all state-owned industrial enterprises and above-scale non-state-owned industrial enterprises. The total amount of observations during the nine years is over 2 million. From 2006, the number of sample enterprises per year is over 300 thousands. No other enterprises database can be compared to it besides the census database. From the view of statistics and econometrics, the advantage of large sample is a lower bias and a higher efficiency of the estimates. Second, the database has a lot of indices, including the basic information and financial data of the enterprises, which can comprehensively reflect the enterprises' behaviors such as market entry, investment, loans, advertisements, research and development, export, etc. as well as the enterprises' short-term and long-term performance. In addition, the aggregated data of the enterprises can reflect the market structure of the industry or of the district that the enterprises belong to. From the view of industrial organization theory, researchers can do almost all kinds of research once acquiring data about market structure, enterprise behavior and performance. Researchers of corporate finance, firm theory, international trade, industrial agglomeration and other relative fields can also make use of this database, including doing interdisciplinary research. If combining this database with other databases, researchers will find more research perspectives. The more the indices, the more the explanatory variables and controlling variables that researchers can use in constructing econometric formulas. Thus the omitting variables problem can be

mitigated. Third, the database has a long time series. It was collected first in 1998, and is updated to 2008 by now, which means it contains 11 years. Thus it is feasible for researchers to use dynamic panel data methods, which may help find out the effect of historical factors and study the transaction of enterprises or industries from a dynamic view.

Comparatively, some other prevalent enterprise databases such as Wind financial database, Sinofin Economic and Financial Database、GTA Listed Company Database use listed companies as samples and have more comprehensive, accurate and frequently indices. For instance, these listed company databases usually contains major shareholders' holdings, the personal characteristics and position change of board members and senior executives, thus can be used to study corporate governance structures. Moreover, listed company databases contain not only industrial listed companies, but also financial and service industrial listed companies, which are not included in Chinese Industrial Enterprises Database. Furthermore, some specific census also results in enterprises database. For example, World Bank and National Bureau of Statistics held a census of more than 1200 enterprises in 12 provinces in China in 2006, involving social responsibility, inner management, quality management, labor management, environment management, market competition, technique development, and some other aspect of the enterprises. From 1991 to 2006, United Front Work Department and All-China Federation of Industry and Commerce successively held sampling surveys of private-owned enterprises, involving their basic information, management system, entrepreneur backgrounds and labor relations.[3]

3 The Current applications of the Database

Because of the special advantage of Chinese Industrial Enterprises Database, a lot of home and abroad economists use it to write and publish papers every year during the past years, with themes covering industrial organization, firm theory, corporate finance, transaction economics, international trade, labor economics, regional economics and so on. Following we will give brief introductions of the applications of Chinese Industrial Enterprises Database on these economic branches. On one hand, we hope it can be helpful for interested researchers to know about what researchers in different fields have done with it and what can be done with it in the future; on the other hand, we hope it can be helpful for interested researchers to know about how current researchers did their study. Of course, because of limited article length and our energy, we can hardly cover all literatures using this database. Thus we concentrate on those major journals and widespread English literatures.

3.1 Productivity

Among all relevant literatures using this database, productivity is the most popular theme. Productivity is the most important efficiency measure, just as what

[3] About other enterprises databases, interested readers can visit the website of Service Center for Chinese Studies ,The Chinese university of Hong Kong.

Krugman said (Krugman, 1997):"Productivity isn't everything, but in the long run it is almost everything". Moreover, for computing enterprises' productivities, Chinese Industrial Enterprises Database provides with special advantages that aggregated data does not. Using sales or economic value added (as Y), fixed assets (as K) and employee number (as L), and deflating them with price index, we can compute labor productivity and total factors productivity (TFP) for every enterprise. Because that labor productivity cannot reflect the efficiency of capital, most literatures use TFP as the measure of productivity. And because these industrial enterprises are more comparable with international industrial classification, existing literatures always use industrial enterprises as samples when computing TFP. When computing TFP, some researchers use traditional Solow residual method, such as Xie et al (2008)、Hsieh and Klenow (2009); some researchers use main-stream OP method (Olley and Pakes, 1996), such as Zhang et al (2009)， Yu (2010), Nie and Jia (2011), Yang and Xiong (2011), Brandt et al (2012); some researchers use LP method (Lecinsohn and Petrin, 2003), such as Zhou et al (2007); some researchers use SFA (Stochastic Frontier Approach), such as Liu and Li (2008).[4]

3.2 International Trade

International trade is highly relevant to productivity. More specifically, it is the relation between enterprise's export and its productivity that matters. According to the famous firm heterogeneity hypothesis (Melitz, 2003), enterprises with high productivity tend to export, which means productivity and export have a positive correlation. Chinese Industrial Enterprises Database includes enterprises' export delivery value, but we cannot differentiate general trade enterprises and process trade ones. Using Chinese Industrial Enterprises Database, some researchers have tested whether it is true for the hypothesis in China. Zhang et al (2009) uses industrial enterprises data samples from 1998 to 2007 and finds that export will help enterprises improve TFP, that is, there exists the learning effect of export. But Li (2010) uses samples from 1998 to 2007 and finds that the average TFP or labor productivity of export enterprises is lower than domestic enterprises, which he called the paradox of productivity. Moreover, Zhao et al (2011) finds that labor productivity is negatively correlated with export choice, but TFP is sometimes positively correlated with export choice. It seems to indicate that researches using this database don't show strong evidences supporting the firm heterogeneity hypothesis. However, Lu (2010) gives a theoretical explanation for this. There are also some other researchers doing relative studies using Chinese Industrial Enterprises Database. For example, Yu (2010) finds that trade liberalization(tariff cut) will increase the TFP of export enterprises; Bao et al (2011) finds that the exports of enterprises haven't significantly improved their employees' income; Yang and Zheng (2011) finds that the vertical specialization of industries have different impacts of employees' salaries.

3.3 Foreign Direct Investment

It is ten years since China joined WTO. What role has foreign direct investment

[4] Nie and Jia (2011) compare the advantages and disadvantages of several TFP methods.

(FDI) played in China's economic development? Qi et al (2008) uses industrial enterprises data from 1998 to 2001 to study the spillover effect that foreign enterprises have on the TFP of domestic enterprises, and finds that the spillover effect is not significant within a single industry but positive among industries and among districts. Luo et al (2008) uses industrial enterprises data of 2000 and 2002 and finds that foreign enterprises have significant positive spillover effect on domestic enterprises of the same industry or the same district. What's interesting, Lu (2008) uses industrial enterprises data from 1998 to 2005 and finds that the spillover effect foreign enterprises have on domestic enterprises decrease with geographical distance. It is positive within a city and positive within the whole country and it is negative for the state-owned enterprises and positive for the private-owned enterprises. Du et al (2011) finds that the spillover effect that foreign enterprises have on domestic enterprises is formed by the industrial relations forward and backward, and horizontal industrial relations didn't produce significant spillover effect. What's more, foreign enterprises from Hong Kong, Macao, Taiwan and enterprises from foreign countries have different effect on domestic enterprises. Xu and Sheng (2011) has a similar conclusion. Sheng et al (2011) finds that FDI increases the value of domestic enterprises exports by industrial relations backward and increases the export willingness of domestic enterprises by the demonstration effect within the industry. Chen et al (2011) finds that foreign enterprises have an obvious wage premium and a inhibiting effect on domestic enterprises wages, thus aggravate the wage inequity among enterprises.

3.4 R&D

Technological Innovation is one of the main sources of enterprises' productivity. Thus the research and development (R&D) behaviors of enterprises are always been focused. Literature about R&D can be divided into two categories: The first category studies the determinants of R&D or enterprises innovation to test Schumpeter hypothesis; the second category studies the effect that R&D have on enterprises' performance. Nie et al (2008) uses industrial enterprises data from 2001 to 2005 and finds that the R&D intensity (a measure of innovation) have an inverted-U relation with the scale and market competition of enterprises. Moreover, although state-owned enterprises have higher R&D intensity than private-owned enterprises, but their efficiency is lower. Hu et al (2009) finds that FDI and the restricting have positive effect on R&D intensity of enterprises. Chen and Zhu (2011) uses industrial enterprises data from 2005 to 2006 and separates industries into ones with high administrative barrier to entry and ones with low administrative barrier to entry according to the proportion of state-owned economy of an industry. They find, within high administrative-barrier-to-entry industries, innovation and market structure have an inverted-U relation and the Schumpeter hypothesis is true, but within low administrative-barrier-to-entry industries, things are opposite. Chesbrough and Liang (2007) uses semiconductor industry as an example and finds that market orientation will affect the payoffs of R&D. That is, the global market orientated enterprises can get higher R&D payoffs than domestic market orientated enterprises. Dai and Yu

(2012) finds that R&D before export can improve the productivity after export.

3.5 Privatization

One of the main achievements of China's state-owned-enterprises reform is that a huge amount of state-owned enterprises make a transformation from a totally state-owned enterprise to a state holding enterprise or private enterprise. This is obvious from the change of paid-in capital composition of state-owned industrial enterprises. Tong (2009) uses industrial enterprises data from 1998 to 2003 and finds that the aggravation of market competition, the increase of FDI concentration, and the hardening of budget constraint are the main reasons for the privatization of state-owned enterprises. And state-owned enterprises with better performance are more likely to be privatized. Bai et al (2009) studies the effect of the privatization of state-owned enterprises and finds it increases sales and labor productivity mainly through diminishing administration cost. Dougherty et al (2007) finds privatization improves enterprises' productivity through increasing their profitability and specialization level

3.6 Corporate Finance

Many researchers use Chinese Industrial Enterprises Database to study the investment, financing and tax avoidance behaviors of enterprises because it contains abundant financial indexes. Cai and Liu (2009) raises an interesting question: will competition lead to tax avoidance? They measure the level of tax avoidance by comparing the profit that an enterprise reports and the profit calculating using accounting rules. Using industrial enterprises data from 2000 to 2005, they find competition will aggravate tax avoidance. Cull et al (2009) holds that there exists a substitution relation between bank loan and trade credit in China. Badly-performing state-owned enterprises will reallocate bank loan to business customers through trade credit, and well-performing private enterprises are more likely to develop trade credit than badly-performing enterprises. Yu and Pan (2010) use industrial enterprises data from 2004 to 2007 and finds that enterprises, especially private enterprises, will use trade credit as a method of product market competition. This supports the competition hypothesis of trade credit. Guariglia et al (2011) finds that the internal financing (cash flow over total assets) of private enterprises is an important constraint condition of their growth, but state-owned enterprises are free of this kind of constraints.

3.7 Industrial Agglomeration

Using enterprise level data, we can get the aggregated data of industry level or of district level. This will reflect industrial agglomeration of China. Using industrial enterprises data from 1998 to 2005, Lu and Tao (2009) studies the determinant of industrial enterprises agglomeration (measured by EG index) of China. They find that local protectionism (the proportion of employees employed by state-owned enterprises) is the main determinant of impeding industrial agglomeration. Some other researchers study the impact that industrial agglomeration has on enterprises. Li et al (2011) finds that industrial agglomeration has a positive effect on the scale of

enterprises. Lin et al (2011) finds that there is an inverted-U relation between industrial agglomeration and productivity. Yang and He (2011) finds that trade will affect the geographical agglomeration of export enterprises through information and division of labor.

3.8 Micro-effect of Macro-policy

Researchers can also use Chinese Industrial Enterprises Database to study the effect that macro-policy has on the behaviors and performance of micro-enterprises, and thus provide micro-foundations for macro-policy analysis by empirical researches. Nie et al (2009) uses difference in difference model (DID) and finds that the added-value tax transformation policy in 2004 significantly improves enterprises' fix assets investment and labor productivity, but decreases the number of their employees. Yuan and Zuo (2011) uses DID model to study how the "county-power enlarging" policy of Zhejiang Province from 2003 to 2005 affects enterprises' growth and finds that the policy has improved the sales growth and assets growth of county enterprises. Peng and Lian (2010) uses industrial enterprises data from 2000 to 2007 and finds that macro policies increasing interest rate in the short run will lead to an increase of enterprises' production cost, thus leading to inflation. This is what we called monetary cost. Song et al (2011) infers that discriminative financial policies will lead to a rapid growth of saving rate efficient enterprises and thus lead to a huge amount of foreign exchange reserve and trade deficit. They use industrial enterprises data from 1998 to 2007 to test it.

3.9 Others

Other empirical researches with Chinese Industrial Enterprises Database mainly focus on employment. Fang et al (2010) uses industrial enterprises data from 1999 to 2005 to compare micro employment elasticity of enterprises of different ownerships by GMM method (generalized method of moment) and finds that the employment elasticity of state-owned enterprises is lower than that of private enterprises. Zhang et al (2010) uses industrial enterprises data from 1998 to 2006 to get the aggregated data of infrastructure, output and employment of every province, and calculates the elasticity that infrastructure has on output and employment. Dong and Xu (2009) discusses how the labor flow between public sectors and private sectors contributes to China's economic growth.

4 Potential Problems of Chinese Industrial Enterprises Database

Chinese Industrial Enterprises Database provides indispensable materials for micro-econometrics researches. However, this database is not perfect, but problematic. As users, we finds the database has many problems including sample match problems, missing indices, unrealistic outliers, sample selection, and measurement errors. Ignoring these problems, the results of empirical researches may not be robust, even may be wrong. Existing literatures find some of these problems and provide some

solutions. Now, on the basis of existing literatures and with our own experiences, we will sum up the potential problems of the database and make our best to give advises on solving them.

4.1 Sample Match Problem

For multiple-year data, the first step of reorganizing data is to construct a two-dimensional panel data using enterprise ID and year. This easy step is always very difficult for Chinese Industrial Enterprises Database, because in this database it is very hard to find a unique feature as ID to identify every sample enterprise. Usual solution is using some basic information such as code of certificate, name of the enterprise, legal representative, address, postal code, phone number, industry code, name of main product, opening year to identify whether different samples come from the same enterprise. However, because the basis information is not required to fit a certain format, without effective intelligent approximate string matching methods, accurate matching can hardly realize. Among the basic information, code of certificate and name of the enterprise are more accurate, thus can be the main basis of matching. For example, Brandt et al (2012) firstly identify the same enterprise by the same code of certificate, then identify it by the same name, finally referring to other basic information. This sequential identifying method assumes that the code of certificate is more accurate than the name of a enterprise. That is, samples with the same code of certificate will definitely be identified as the same enterprise, but samples identified as the same enterprise may have different code of certificate. In this database, there exist cases that the same enterprise changes it code of certificate (for example, after transformation or recombination) as well as cases that different enterprises share the same certificate of code (maybe because of statistical errors). The same problem exists for the name of the enterprise. Many enterprises changed their names when transforming, recombining or expanding. For example, there are many enterprises in China whose names are at first “XX Factory”, then changed to “XX Limited Liability Company”, and finally “XX limited company”. Sometimes, geographical locations in enterprises’ names will be different. For example, an enterprise name changes from “XX city mechanical and electrical factory“ to “XX mechanical and electrical factory”. Accurate matching by names of enterprises will improperly identify too many enterprises.

Our suggestion is: divide all the enterprises in groups twice respectively by code of certificate and by name, and then check whether enterprises belonging to the same name group belong to the same code group. If yes, put all these enterprises into the same group (do the same steps for every name group and keep regrouping, we can call this cross-matching); If there are not any observations of the same year in the new group, than identify this group as the same enterprise; If there are some observations of the same year in the new group, then use manual identification. There may be several cases in the manual identification phase, so we need to consider data features and basic information to make comprehensive judgment. For example, samples in a group may belong to a same enterprise, but there are two observations for some year. We only need to keep one of the two observations because the other one may be

totally the same or may lack some important index. Samples in a group may belong to different enterprises, but this may be caused by the mistake when enterprises report their code of certificate. Then we need to consider the order of magnitudes of some important indexes such as name, legal representative, address, code of industry, sales, and registered capital and make sure which samples belong to the same enterprise. We find that, after using cross-matching method, about 10% of the observations (about 200 thousands) belong to the same-name-different-code case or the same-code-different-name case. Obviously, ignoring matching problem will seriously affect the authenticity and veracity of samples.

Besides the problems of matching enterprises, there are also problems of matching industries. Users must notice that National Bureau of Statistics of China uses two different standards of grouping industries before and after 2002. It uses GB/T 4754—1994 before 2002 (including 2002), and GB/T 4754—2002 after 2002. These two standards are the same on 2-digit industries, a little different on 3-digit industries, and very different on 4-digit industries. Yang and Zheng (2011) and Yang and Xiong (2011) transfer 4-digit industries of 1994GB to 3-digit industries of 2002GB. It is a possible solution.[5]

4.2 Missing Indices

Chinese Industrial Enterprises Database changes its objects and caliber of statistics every year, leading to some important indices missing at some years. First, Chinese Industrial Enterprises Database of some source mixes the economic census data of 2004with data of other years without matching with data of other years. This leads to the missing of some important indices of the data of 2004, such as total industrial output value, industrial added value, export value, and R&D cost. At the same time, compared with data of 2004, data of other years lacks indices about union, education background and technical titles of male and female employees. Therefore, users should check the differences between data of 2004 and of other years before analyzing it.

Second, there are also some differences between the data before and after 2003. For example, data before 2001 doesn't include R&D cost. Chinese Industrial Enterprises Database from 1999 to 2003 of some source lacks industrial added value and account receivable, but has net account receivable. According to accounting rules, net account receivable equals account receivable minus the ending balance of bad debt reserves. Thus account receivable and net account receivable cannot be simply compared. For years lacking industrial added value, users can estimate industrial added value according to accounting rules: industrial added value = total industrial output value - intermediate industrial input + added-value tax. For years lacking total industrial output value (such as 2004), the estimating equation is: industrial added value = product sales – opening stock + closing stock – intermediate industrial input

[5] Recently, many researches combine Chinese Industrial Enterprises Database with China Customs Database. Researches about these themes should match codes of industrial classification for national economic activities with codes of Harmonized System of customs. For details, please refer to Yang (2008).

+added-value tax. For example, Liu and Li (2008) uses this estimating method. We use the first equation to estimate the industrial added value from 2005 to 2007, and finds that on average the estimate values are a little smaller than the report values (see Table 2). Users should notice this difference when calculating productivity by industrial added value.

Table 2: Report value and estimate value of industrial added value

	2005	2006	2007
Report value	26229.8	29875.86	34447.48
Estimate value	26206	29849.54	34401.34
Number of observations	270110	300194	335076

Note: unit of money is thousand yuan; the report value and the estimate value are average values. Outliers are not eliminated.

We need to especially point out that, although Chinese Industrial Enterprises Database includes the export value of enterprises, we can only know whether an enterprise exports. We cannot differentiate general trade and process trade. For labor intensive process trade enterprises, their labor productivity and TFP may be lower than general trade enterprises. Moreover, process trade is a special phenomenon of developing countries. If we don't make a distinction between process trade enterprises and general trade enterprises while simply compare the productivity of all export enterprises with the productivity of domestic enterprises, it is no wonder getting the result that the average productivity of export enterprises is lower than that of domestic enterprises.[6] In fact, Dai et al (2011) matches Chinese Industrial Enterprises Database with Custom Database. After eliminating process trade enterprises, they find that the paradox of productivity is not true.[7]

4.3 Unrealistic Outliers

Although Chinese Industrial Enterprises Database contains over 130 indices, there are many unrealistic outliers of them. Unrealistic outliers make many observations invalid, thus must be eliminated before any regression. We notice that, the eliminating method that Cai and Liu (2009) uses is comparatively comprehensive and is widely borrowed by other researchers. Firstly, they eliminate observations that lack important indices (for example, total assets, number of employees, total industrial output value, net value of fixed assets, and sales); Secondly, they eliminate observations that don't satisfy "above-scale" standard, that is, the net value of fixed

[6] In this sense, we believe that Chinese Industrial Enterprises Database is not suitable for testing firm Heterogeneity hypothesis of international trade.

[7] Some researchers have matched Chinese Industrial Enterprises Database and Custom Database. Because their original data and methods are different, their matching results are quite different. Yu (2010) matches 30% export enterprises of Custom Database according to information such as postal code and telephone number. Tang (2011) matches 70% export enterprises of Chinese Industrial Enterprises Database. Based on information such as name, postal code, telephone number, Yang and He (2011) uses searching and matching in key words database methods and matches 60% of export value of Custom Database.

assets is less than 10 million RMB, or the sales is less than 10 million RMB, or the number of employees is less than 30[8]; Thirdly, they eliminate observations that obviously don't satisfy accounting rules, including ones with total assets less than current assets, total assets less than net value of fixed assets, or accumulated depreciation less than current depreciation; Fourthly, they eliminate the extreme value of important indices (top and bottom 0.5%).

Using 1999-2007 as an example, we analyze the unrealistic outliers problem. Firstly, we find that, we eliminated over 5900 observations lacking sales, numbers of employees, total assets or net value of fixed assets. That is about 0.3% of all 2048833 observations. Secondly, according to Xie et al (2008), we eliminate over 28000 observation with number of employees less than 8 (we don't believe this kind of enterprises have reliable accounting system), accounting for 1% of all observations. Thirdly, we eliminate over 200 observations with total assets less than current assets, total assets less than net value of fixed assets, or accumulated depreciation less than current depreciation. Finally, we eliminate over 176500 observations with sales less than 5 million RMB, accounting for 9% of all. After four steps above, we have eliminated about 200 thousand observations, accounting for 10% of all. Although we have done these four steps, we still find many observations abnormal. For example, if we treat observations with profit rate higher than 99% or lower than 0.1% as outliers according to Bai et al (2009), there are about 430 thousands outliers, accounting for about 23% of all. Alternatively, we can use a looser standard and find about 11 thousands outliers with paid up capital equals to or less than 0, accounting for 6% of all. Therefore, even after eliminating these outliers, users still have to eliminate outliers of regression's important variables or parameters. We put all the abnormal situations before eliminating any important indices in Table 3.

Table 3 Abnormal situations of indices

Eliminating standards	Number of outliers	Total observations	Percentage of outliers
lack sales, number of employees, total assets, or value of fixed assets	5912	2048833	0.29
Number of employees less than 8	28307	2048833	1.38
total assets less than current assets, total assets less than net value of fixed assets, or accumulated depreciation less than current depreciation	2832	2048833	0.14
Sales less than 5 millions	201540	2048833	9.84
profit rate higher than 99% or lower than 0.1%	570157	2048833	27.83
Paid up capital equals to or less than 0	33026	2048833	1.61

4.4 Measurement Errors

[8] Their "above-scale" standard is different from the official standard (sales more than 5 million RMB).

When National Bureau of Statistics collects data of industrial enterprises, they don't give enterprises a single form to fill in. Instead, enterprises are asked to report data by annual report or other periodic report. Then National Bureau of Statistics gathers all the data. In fact, enterprises have to report at least four forms: combined annual report, combined periodic report, annual report, periodic report. This means, because of different statistical time and caliber, enterprises may report different value for the same indices at different time. Moreover, many small-scale enterprises don't have reliable accounting system. They may also conceal or falsely report some indices for tax avoidance. These factors will all lead to measurement errors.

Use R&D cost as an example. Between 2001 and 2007 (except for the census year 2004), there are over 1.2 million observations with R&D cost equals 0, accounting for 89% of all the 1.4 million observations. There are three situations in which enterprises report R&D cost 0: (1) An enterprise doesn't pay for R&D, thus the cost is 0; (2) An enterprise doesn't know how much it pay for R&D, thus they report 0 randomly; (3) An enterprise doesn't report its R&D cost, thus the statistician fill 0 in it. The first and second cases are more likely to happen for small and middle-sized enterprises. Therefore, we eliminate small and middle-sized enterprises with sales less than 300 million and eliminate all the export enterprises. We find there are also over 20 thousand enterprises with R&D cost equals to 0, accounting for over 70% of the 28 thousand observations left. This remind all the user of two points: First, if most of small and middle-sized enterprises don't have R&D cost and it is truth, then when analyzing the determinant of enterprises' R&D (or innovation) cost, it is better to use Tobit truncation model (for example, Nie et al, 2008) because it is more likely to get consistent estimates than OLS; Second, if we can't distinguish the second case and the third case, then the accuracy and authenticity of R&D cost is questionable. It may be improper to use Chinese Industrial Enterprises Database to study the determinants or performance of R&D.

Other indices with obvious measurement errors are profit and added value. Enterprises tend to underreport or misreport profit and the added value when the regulation is poor, because the tax is positively correlated with these two indices that an enterprise reports. In fact, Cai and Liu (2009) calculates enterprises' profit according to accounting rules (profit = total industrial output value– intermediate input – financial cost – wages - current depreciation - added-value tax) and finds that the average of profit rate between 2000-2005 is 0.1431. But the average of profit rate that enterprises report is 0.0515. The latter is over 2/3 less than the former. Moreover, according to our estimate of industrial added value between 2005 and 2007 above, it is obvious that the estimate is smaller than the report value.

Another problem is fake indices, which doesn't belong to classic measurement error problem but still relates to it. According to registration type, there are about 1/5 observations (about 400 thousand) belonging to foreign enterprises (including foreign enterprises from Hong Kong, Macao, Taiwan and enterprises from foreign countries). This proportion is much higher than our intuition. What is known to all is that foreign enterprises can get many kinds of preferential tax policies. After deeper analysis, we find that although registered as foreign enterprises, 6% of them have 0 paid up capital

from Hong Kong, Macao, Taiwan and foreign countries, with about half having explicit foreign identification on their registration number (for example, “No. XX of 企合津总字). There are two possibilities. First, these enterprises used to be foreign enterprises, but they didn’t change their registration type after changing their paid up capital. Second, these enterprises mistakenly report their registration type. For the 94% of foreign enterprises left, we cannot confirm their real type and cannot eliminate fake foreign enterprises as well.

4.5 Sample selection

For sample selection, a serious problem of Chinese Industrial Enterprises Database is that it contains all the state-owned industrial enterprises, but only contains above-scale non-state-owned industrial enterprises. Therefore, when users are to compare the behaviors or performance of state-owned enterprises and non-state-owned enterprises, it is better to eliminate all the below-scale state-owned enterprises. When users study industrial agglomeration, the agglomeration level of non-state-owned enterprises may be underestimated (Lu and Tao, 2009). Samples of above-scale enterprises are not random. Among the over 2 million observation between 1999 and 2007, only 8% of enterprises exist every year and only 22% of enterprises exist for the last three years. If an enterprise doesn’t exist for some year, it may be because this enterprise’s sales is less than 5 million RMB that year, or it may be because it is bankrupt, is reconstruct, changes it name, or being missed. In this sense, we can hardly define enterprises’ entry and quit. Therefore researchers should try their best to solve or mitigate this problem when using this database to analyze enterprise dynamics. Moreover, this database contains detailed information about where an enterprise is, so users can know whether it is in a special economic zone or in an economic development zone. Because enterprises in a special economic zone or an economic development zone have special features, distinguishing them from general enterprises can help mitigate sample selection problem when comparing productivity, industrial agglomeration, profit rate, financial cost, etc. Another troublesome problem is that National Bureau of Statistics use enterprise as statistical caliber, not enterprise group or factory. Thus many enterprises belonging to a same group may be recognized as different enterprises, while the differences among many factories may be covered with a single enterprise.

4.6 Definition Ambiguities

Ownership cannot be ignored when analyzing Chinese enterprise. We notice that existing literature generally use two methods to identify enterprises’ ownership: registration type and paid up capital. These two methods are actually different. The former represents the type that an enterprise registers at the Industrial and Commercial Bureau, and the latter represents an enterprise’s real shareholding type. We may define enterprises with registration types “state-owned”, “state-owned joint”, “state-collective joint” and “wholly state-owned” as type I state-owned enterprises, while define enterprises with state-owned capital over 50% of paid up capital as type II state-owned enterprises. In Chinese Industrial Enterprises Database of 1999-2007,

after eliminating observations with unrealistic paid up capital, there are 245376 type I observations and 252629 type II observations. The latter is about 3% more than the former, and their overlapping observations account for about 84% of type I state-owned enterprises. It means although registered as state-owned enterprises, at least 15% of them are not real state-owned any more. Because shareholding type can better represent enterprises' ownership, thus we recommend users use paid up capital ratio to define enterprises' ownership.[9] The same problem happens for foreign enterprises too. According to law, foreign capital ratio should be more than 50% of foreign enterprises. Although about 1/5 observations are registered as foreign enterprises, 10% of them have Hong Kong, Macao and Taiwan or foreign capital ratio less than25%. Some researchers use whether foreign paid up capital ratio exceeds 25% as the method to identify foreign enterprises (for example, Lu, 2008).

Besides ownership, another variable hard to define is "capital". Theoretically, capital is the sum of the stock of fixed asset and investment flow. Most literatures define capital as the original cost or net value of fixed assets, and then calculate investment $I_{it}=K_{it}-(1-\delta)K_{it-1}$ by perpetual inventory method, where I represents investment, K represents current capital stock, δ represents depreciation rate. This methods means the investment of the first period will be missed. Depreciation rate is always chosen to be 5%, 10%, or 15%. Moreover, users should also use different price indexed to depreciate output value, capital, investment, intermediate input, and some other indices. What we should remind users of is that different capital definition methods, depreciation rates, and price indexes will lead to different results. Brandt et al (2012) provides detailed interpretation and operating procedure.

5 Conclusions

Because of the wide use of Chinese Industrial Enterprises Database by domestic and overseas researchers, we have introduced its basic information in this article, including types of sample enterprises and main indices. We have reviewed its application on 9 topics such as productivity, international trade, FDI, R&D, privatization, corporate finance, etc. This will help potential and current researchers know what they can do and what else they can do with the database. Chinese Industrial Enterprises Database itself has many problems, including sample match problems, missing indices, unrealistic outliers, measurement errors, sample selection and definition ambiguities and so on. Ignoring them will impact basic results of econometric analysis. Therefore we have summarized them problems based on current literatures and tried our best to give advices on solving them. Although this article is written for Chinese Industrial Enterprises Database, we believe some problems may happen for other enterprises databases (for example, Large and Middle-sized Industrial Enterprises Database of 1995-2001), so these problems are worth studying by researchers on other occasions. Because of limited article length and our

[9] Every enterprise reports it total paid up capital and state-owned, collective, corporate, private, Hong Kong, Macao, and Taiwan, and foreign paid up capital. Thus we can also identify other ownership types by the ratio of paid up capital.

perspectives, we don't focus a lot on empirical researches of other topics except those above. We haven't talked about empirical researches driven by the combination of Chinese Industrial Enterprises Database and other databases.

References

Bai, Chong-En; Lu, Jiangyong and Tao, Zhigang "How does privatization work in China?" *Journal of Comparative Economics*, 2009, 37: 453–470

Brandt, Loren; Biesebroeck, Johannes Van and Zhang, Yifan, "Creative Accounting or Creative Destruction? Firm-level Productivity Growth in Chinese Manufacturing", *Journal of Development Economics*, 2012, 97(2): 339-351

Cai, Hongbin and Liu, Qiao, "Competition and Corporate Tax Avoidance: Evidence from Chinese Industrial Firms," *Economic Journal*, 2009, 119: 764-795

Chen, Zhihong, Ge, Ying and Lai, Huiwen, "Foreign Direct Investment and Wage Inequality: Evidence from China", *World Development*, 2011, 39(8): 1322–1332

Chesbrough, Henry and Liang, Feng, "Return to R&D Investment and Spillovers in the Chinese Semiconductor Industry: A Tale of Two Segments", working paper, 2007

Cull, Robert; Xu, Colin, Lixin and Zhu, Tian, "Formal finance and trade credit during China's transition", *Journal of Financial Intermediation*, 2009, 18: 173–192

Dai, Mi; Maitra, Madhura and Yu, Miaojie, "Unexceptional Exporter Performance in China: The Role of Processing Trade",Peking University, working paper, 2011

Dong, Xiao-yuan and Xu, Colin Lixin, "Labor restructuring in China: Toward a functioning labor market", *Journal of Comparative Economics*, 2009, 37: 287–305

Dougherty, Sean; Herd, Richard and He, Ping, "Has a Private Sector Emerged in China's Industry? Evidence from a Quarter of a Million Chinese Firms", *China Economic Review*, 2007, 18: 309–334

Du, Luosha; Harrison, Ann and Jefferson, Gary, "Testing for horizontal and vertical foreign investment spillovers in China, 1998–2007", *Journal of Asian Economics*, 2011, forthcoming

Guariglia, Alessandra; Liu, Xiaoxuan and Song, Lina, "Internal Finance and Growth: Microeconometric Evidence on Chinese Firms", *Journal of Development Economics*, 2011, 96: 79–94

Hsieh, Chang-Tai and Klenow, Peter, "Misallocation and Manufacturing TFP in China and India", *Quarterly Journal of Economics*, 2009, 124(4):1403-1448

Hu, Albert Guangzhou and Jefferson, Gary, "A Great Wall of Patents: What is behind China's Eecent Patent Explosion?" *Journal of Development Economics*, 2009, 90: 57–68

Jefferson, Gary and Su, Jian "Privatization and Restructuring in China: Evidence from Shareholding Ownership, 1995–2001", *Journal of Comparative Economics*, 2006, 34: 146–166

Krugman, Paul, *The Age of Diminished Expectations*, the 3rd edition, MIT Press, 1997

Levinsohn, J., and Petrin, A., "Estimating Production Functions Using Inputs to Control for Unobservables", *Review of Economic Studies*, 2003, 70(2): 317–342

Li, Dongya; Lu, Yi and Wu, Mingqin, "Industrial Agglomeration and Firm Size: Evidence from China", *Regional Science and Urban Economics*, 2011, 42: 135–143

Lin, Hui-Lin; Li, Hsiao-Yun and Yang, Chih-Hai, "Agglomeration and Productivity: Firm-level Evidence from China's Textile Industry", *China Economic Review*, 2011, 22: 313–329

Lu, Dan, "Exceptional Exporter Performance? Evidence from Chinese Manufacturing Firms", University of Chicago, working paper, 2010

Lu, Jiangyong and Tao, Zhigang, “Trends and Determinants of China’s Industrial Agglomeration”, *Journal of Urban Economics*, 2009, 65: 167–180

Lu, Jiangyong; Tao, Zhigang and Yang, Zhi, "The Costs and Benefits of Government Control: Evidence from China's Collectively-owned Enterprises", *China Economic Review*, 2010, 21: 282–292

Melitz, Marc J., “The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity”, *Econometrica*, 2003, 71(6): 1695-1725

Olley, Steven and Pakes, Ariel, “The Dynamics of Productivity in the Telecommunications Equipment Industry”, *Econometrica*, 1996, 64(6): 1263-1297

Sheng, Yu; Chen, Chunlai and Findlay, Christopher, "Impact of FDI on Domestic Firms’ Exports in China", University of Adelaide School of Economics, Research Paper No. 2011-15, 2011

Song, Zheng; Storesletten, Kjetil and Zilibotti, Fabrizio, "Growing Like China", *American Economic Review*, 2011, 101: 202–241

Tang, Hei-Wai, "Factor Intensity, Product Switching, and Productivity: Evidence from Chinese Exporters", Tufts University, mimeo, 2011

Tong, Sarah Y., 2009, "Why Privatize or Why not? Empirical Evidence from China's SOEs Reform", *China Economic Review*, 20: 402–413

Xu, Xinpeng and Sheng, Yu, "Productivity Spillovers from Foreign Direct Investment: Firm-Level Evidence from China", *World Development*, 2011, forthcoming

Yang, Rudai and He, Canfei, "Intermediaries and Export Agglomeration", Peking University, mimeo, 2011

Yu, Miaojie, “Processing Trade, Firm’s Productivity, and Tariff Reductions: Evidence from Chinese Products”, Peking University, CCER working paper, 2011

包群、邵敏、侯维忠，《出口改善了员工收入吗？》，《经济研究》，2011年，第9期

陈林、朱卫，《创新、市场结构与行政进入壁垒——基于中国工业企业数据的熊彼特假说实证检验》，《经济学季刊》，2011年，第10卷，第2期

戴觅、余淼杰，《企业出口前研发投入、出口及生产率进步》，北京大学，工作论文，2012年

方明月、聂辉华、江艇、谭松涛，《中国工业企业就业弹性估计》，《世界经济》，2010年，第8期

李春顶，《中国出口企业是否存在“生产率悖论”：基于中国制造业企业数据的检验》，《世界经济》，2010年，第7期

刘小玄、李双杰，《制造业企业相对效率的度量和比较及其外生决定因素(2000—2004)》，《经济学》（季刊），2008年，第3期

路江涌，《外商直接投资对内资企业效率的影响和渠道》，《经济研究》，2008，第6期

罗雨泽、朱善利、陈玉宇、罗来军，《外商直接投资的空间外溢效应：对中国区域企业生产率影响的经验检验》，《经济学季刊》，2008年，第7卷，第2期

聂辉华、方明月、李涛，《增值税转型对企业行为和绩效的影响——以东北地区为例》，《管理世界》，2009年，第5期

聂辉华、贾瑞雪，《中国制造业企业生产率与资源误置》，《世界经济》，2011年，第7期

聂辉华、谭松涛、王宇锋，《创新、企业规模和市场竞争——基于中国企业层面面板数据的证据》，《世界经济》，2008年，第7期

彭方平、连玉君，《我国货币政策的成本效应——来自公司层面的经验证据》，《管理世界》，2010年，第12期

亓朋、许和连、艾洪山，《外商直接投资企业对内资企业的溢出效应：对中国制造业企业的实证研究》，《管理世界》，2008年，第4期

谢千里、罗斯基、张轶凡，《中国工业生产率的增长与收敛》，《经济学季刊》，2008年，第4卷，第3期

杨汝岱，《中国工业制成品出口增长影响因素研究》，《世界经济》，2008年，第8期。

杨汝岱、熊瑞祥，《干中学与中国工业企业出口生产率》，工作论文，2011年

杨汝岱、郑辛迎，《垂直专业化对员工工资的差异化影响》，工作论文，2011年

余淼杰，《中国的贸易自由化与制造业企业生产率》，《经济研究》，2010年，第12期

余明桂、潘红波，《金融发展、商业信用与产品市场竞争》，《管理世界》，2010年，第8期

袁渊、左翔，《"扩权强县"与经济增长：规模以上工业企业的微观证据》，《世界经济》，2011年，第3期

张光南、李小瑛、陈广汉，《中国基础设施的就业、产出和投资效应——基于1998-2006年省际工业企业面板数据研究》，《管理世界》，2010年，第4期

张杰、李勇、刘志彪，《出口促进中国企业生产率提高吗？——来自中国本土制造业企业的经验证据：1999－2003》，《管理世界》，2009年，第12期

赵伟、赵金亮、韩媛媛，《异质性、沉没成本与中国企业出口决定：来自中国微观企业的经验证据》，《世界经济》，2011年，第4期

周黎安、张维迎、顾全林、汪淼军，《企业生产率的代际效应和年龄效应》，《经济学季刊》，2007年，第6卷，第4期

Translated by Yuxiao Zhang