DOI:10.19361/j.er.2018.06.11

中国工业企业数据库的使用问题再探

陈 林*

摘要: 中国工业企业数据库是近年来国内外经济学界常用的企业级微观数据库,但从统计抽样调查的角度来看,该数据库存在着各种问题。为更好地了解该数据库可能出现的系统性误差与随机误差,本文使用 1996-2013 年数据对真实性、系统性误差及基层统计部门反映的各种数据问题,进行定量评估。结果发现:样本范围及统计口径的变动,缺失值较多、"回避规模以上"、"化整为零"等方面的数据问题,均会对数据产生一定的影响,并据此在行业代码整理、真实性检验、统计口径变更处理、统计偏误克服、通货膨胀平减等方面提出了相关建议。

关键词: 中国工业企业数据库;规模以上工业企业;工业统计报表;企业级微观数据;工业增加值

一、引言

"中国工业企业数据库"是近年来国内学术研究使用最多的数据库之一。即便该数据库存在着样本匹配、指标缺失等问题,且 2008 年后更新缓慢,但 2007 年至今还是出现了约 2 345 篇使用该数据库的期刊论文①,其中 CSSCI 期刊比例高达 78.4%。无论从量上还是从质上看,使用中国工业企业数据库的实证研究已成为国内的学术主流。

根据"谷歌学术数据库"和聂辉华等(2012)的不完全统计,截至 2016 年年底,已有超过 200 篇海外学术论文使用到中国工业企业数据库,所发表的期刊包括,综合性的《经济学季刊》(*Quarterly Journal of Economics*)、《美国经济评论》(*American Economic Review*)、《经济研究评论》(*Review of Economic Studies*)、《经济学杂志》(*Economic Journal*),专业学科性的《发展经济学杂志》(*Journal of Development Economics*)、《城市经济学杂志》(*Journal of Urban Economics*)等。上述英国、美国、荷兰的经济学期刊陆续刊登基于中国工业企业数据库写作的实证论文,除了论文本身优秀之外,理由可能还有两点:一是"中国问题"热;二是数据库质量高,能以之做出很好的微观计量论文。一个每年更新的基于官方全面调查的企业级微观数据库,在国外是较为罕见的。

* 陈林,北京师范大学经济与工商管理学院,邮政编码:100875,电子信箱:charlielinchen@qq.com。

本文获得国家自然科学基金面上项目"竞争政策与准入规制的协调机制研究"(项目号:71773039)的资助。作者感谢两位匿名审稿人和编辑部的宝贵修改建议,当然文责自负。

① 中国知网(CNKI)数据库内 2007-2016 年全文包含"工业企业数据库"或"规模以上工业企业数据"的期刊论文。

中国工业企业数据库是每年国家统计局及其下属机构的工业统计成果，是全国所有规模以上工业企业上报统计部门①的原始报表汇总。因此，该数据库可能出现一些系统误差与数据问题，导致学术界在使用时产生选择性偏误、关键指标缺失、统计口径变动等问题。

有鉴于此，本文在聂辉华等（2012）的基础上，进一步讨论了中国工业企业数据库的基本情况、整理办法、真实性检验以及未来的拓展方向，尤其针对新出现的2010-2013年工业企业数据，提出更有效的数据处理建议。本文还使用定量的方法对数据库自身存在的系统误差与数据问题进行评估，并提出初步的处理办法。对数据库的系统分析不仅有助于研究者了解该数据库的现存问题与使用注意事项，还能有助于拓展该数据库的适用领域，从而推进数量经济学的学科方法论发展。

二、中国工业企业数据库的行业代码整理

《国民经济行业分类》在2002年、2011年分别进行了修编，将1998-2002年、2003-2012年、2013年的中国工业企业数据库划分为三段截然不同的样本区间。

对于使用行业数据进行研究的学者而言，这个问题必须予以足够重视。这是因为，在同一个四位数代码行业中，企业生产的产品（指标“主要产品1”、“主要产品2”、“主要产品3”）已经是千差万别，而国有企业与民营企业之间的价值链位置（或者说产品附加值）也是各不相同，加上“相关市场”的界定千差万别，“同一行业”的界定尤需谨慎。如果使用三位数代码、二位数代码，对中国工业企业数据库的样本进行“同一行业”处理，这样的误差将更为严重。本文建议，研究具体的工业行业时，或者按照行业进行分类、分组研究时，学者应该使用更为精确的四位数行业代码，而非相对粗糙的三位数、二位数行业代码。

2003年前后，行业代码的变动主要有以下几种模式：（1）更改代码。比如镍钴矿采选业（0914）更改为镍钴矿采选（0913），水产罐头制造业（1433）更改为水产品罐头制造（1452），等等。（2）合并代码。比如锰矿采选业（0821）和铬矿采选业（0822）合并为其他黑色金属矿采选（0890），肉类罐头制造业（1431）和禽类罐头制造业（1432）合并为肉、禽类罐头制造（1451）。以上两种四位数行业代码的转换，使用计量、统计软件的编程可以很容易实现，而以下几种行业代码的转换则显得更为复杂。（3）分解代码。比如其他水产品加工业（1359）分解为鱼油提取及制品的制造（1364）和其他水产品加工（1369）；其他食品加工业（1390）分解为蔬菜、水果和坚果加工（1370）、其他未列明的农副食品加工（1399）。（4）分解后合并。比如煤炭开采业（0610）分解为烟煤和无烟煤的开采洗选（0610）、褐煤的开采洗选（0620）及其他煤炭采选（0690），其中第一部分与煤炭洗选业（旧的0620）合并成为新的0610；天然原油开采业（0710）分解为天然原油和天然气开采（0710）及与石油和天然气开采有关的服务活动（0790），其中第一部分与天然气开采业（0720）、油页岩开采业（0730）合并成为新的0710。以上两种行业代码的转换相对没有什么规律，因为，2003年后，企业工业统计人员在选择新的行业代码时，会根据企业自身经营而随机应变。而且，分解出来的新的细类行业与原来的行业大同小异，这增加了数据库处理上的难度。本文建议，使用计量、统计软件的循环语句，找出2003年后企业界定自身属于分解后的哪个行业，然后按照这个口径将1996-

①本文统称的“统计部门”主要指国家统计局、省市县统计局、乡镇统计站、各级统计调查队等官方统计机构。本文所使用的数据均来自于以上统计部门的官方统计。

2002 年该企业的四位数行业代码进行转换。

2013 年后,企业的四位数行业代码再次进行了大幅调整。最便捷的统一行业口径的处理方法是按照上述办法将 2013 年的数据转换为 2003-2012 年间的四位数行业代码。不过,这个办法并非一劳永逸,因为未来可能出现的 2014 年后的新数据,也是按照最新的四位数行业代码进行分类。因此,本文建议,使用最新的《国民经济行业分类(GB/T 4754-2011)》对 1996-2012 年间的中国工业企业数据库进行四位数行业代码的统一。在获得统一的四位数行业代码后,通过简单的编程,研究者就可以获得三位数行业代码及二位数行业代码,其中二位数行业代码即为每年《中国统计年鉴》等官方统计资料中的行业口径。

三、中国工业企业数据库的真实性检验

2013 年起,学界突然出现了 2010-2011 年的中国工业企业数据库,此数据库流传甚广,甚至有部分学术论文使用该数据库进行了实证研究,但事后证明该数据库是恶意编造出来的。对于每一位研究者,获得 2010 年之后的中国工业企业数据库是进行下一阶段实证研究的基础,自然迫不及待将该数据库投入研究。其理由如下:一是手头上的中国工业企业数据库可以更新了;二是 2008 年金融危机、"四万亿"财政政策冲击后的中国经济数据对于当前的中国问题研究至关重要。

笔者亦于 2013 年获得 2010-2011 年的虚假工业企业数据。经过大量数据整理、合并与实证分析,中国社会科学院经济研究所刘小玄研究员对该数据使用"与官方统计资料比对"的方法进行检验,发现了该数据的恶意编造迹象。随后,笔者使用下文提及的"资本结构比对"方法,进一步确认了该数据库的虚假。

有鉴于此,所有合并后的中国工业企业数据库都必须对数据的真实性进行一定程度的检测,尤其是 2010 年之后的样本。这是中国工业企业数据库构建工作中的必不可少环节,也是研究者获得该数据库之后的首要工作。

本部分将使用笔者手上的中国工业企业数据库(1998-2013)①,对该数据库进行真实性检验。

(一)与官方统计资料比对

根据工业统计报表制度,中国工业企业数据库与历年《中国统计年鉴》的工业部分和《中国工业统计年鉴》中的样本覆盖范围一致。因此,一旦研究者手上的历年中国工业企业数据库的行业或全国加总指标,与官方统计资料严重不符,即可基本断定该数据库存在一定真实性问题。为此,本文首先对比该数据库与历年《中国统计年鉴》《中国工业统计年鉴》《中国经济普查年鉴 2004》《中国经济普查年鉴 2008》中的规模以上工业企业个数,结果发现:除 2009 年的 13.1%、2010 年的 23.0%外,其余年份差距均小于 10%,详见表 1。

除了企业个数,数据库中的资产总计、流动资产、负债合计、主营业务收入、利润总额、工业总产值等指标,也可以加总为全国指标、二位数行业代码指标、分地区指标、分产权结构指标,将其与官方统计资料中的宏观指标进行比对。如果二者相差过大,即表明数据库的真实性存疑。但值得一提是,规模以上统计口径及行业代码的变动、经济普查年份等均会影响到

①1998-2007 年数据来自暨南大学,2008-2009 年、2011-2013 年数据来自中国社会科学院经济研究所,2010 年数据来自中国社会科学院工业经济研究所,部分指标来自中国社会科学院经济研究所提供的 2010 年全行业基础数据库。特此感谢刘小玄研究员、江飞涛副研究员、李晓萍副教授。

指标加总及上述比对法的有效性。

表1　　中国工业企业数据库(1998-2013)整体概况

年份	统计年鉴样本量（个）	样本量（个）	法人代码重复（个）	资本项不对应（个）	剔除缺失值后资本项不对应（个）
1998	165 080	164 726	0	1	1
1999	162 033	161 635	2	1	1
2000	162 885	162 488	4	0	0
2001	171 256	170 874	2	1	1
2002	181 557	181 213	2	1	0
2003	196 222	195 896	4	1	1
2004	276 474	278 724	8	3	3
2005	271 835	271 560	6	6	6
2006	301 961	301 686	8	3	3
2007	336 768	336 512	10	17	10
2008	426 113	411 311	0	1 569	11
2009	434 364	377 341	5 322	262 232	0
2010	452 872	348 536	4 124	346 886	346 878
2011	325 609	302 593	0	1 065	265
2012	343 769	311 314	0	1 986	108
2013	352 546	344 875	0	4 119	201

资料来源：作者整理。

(二)恶意编造的检验

根据戴天仕(2014)的发现，虚假的、恶意编造的2010-2011年中国工业企业数据库出现了以下几个特征：(1)使用真实的2008年之前的数据生成新的数据。比如，虚假数据库中的2010-2011年样本绝大部分可以在2007年之前的样本中找到法人代码、企业名称、资产合计、主营业务收入、实收资本以及企业成立月份完全一样的样本。(2)造假者的大意导致指标逻辑错误。2010-2011年的样本比2007年的样本在“成立年份”指标上分别增大了3年和4年，这是造假者错误理解“成立年份”的变量定义所造成的。(3)“整行数据重复”、“数据复制粘贴”也出现在虚假数据库之中。

在本文使用的数据库中，以上三种情况均没有出现。因此，其真实性得到一定保证。

(三)数据库匹配的效果

由于2009-2010年的中国工业企业数据库部分样本缺乏法人代码这个关键指标，该年份的数据库往往由其他学者通过匹配，填补法人代码获得。这样一来，评估同一年数据的“法人代码重复”现象，成为了检验该年数据库真实性、可信度的一个标准。本文使用的数据库显示，2009-2010年分别出现了5 322个、4 124个法人代码重复的样本，远高于其他年份。其原因正是，数据库生成过程进行了多个数据库的匹配工作，以致误差的出现。

对于2009-2010年出现较大面积“法人代码重复”现象的工业企业数据，本文建议：使用这两年数据作为企业级微观数据库构建企业面板数据模型时，需更加谨慎。一般而言，在企业“身份证”匹配过程中，行业代码、行政区划代码、产权结构等指标不会受到影响。因此，这两年数据可以加总为行业面板、地区面板、产权结构面板等进行面板数据回归。

(四)从资本结构出发的真实性检验

考虑到上文提及的造假者对“成立年份”具有明确经济学含义的指标理解偏差，而且结构性指标更难准确编造，本文使用一个与社会主义市场经济紧密相关的结构性指标——资本结构，对数据库的真实性进行检验(结果见表2)。

表 2　　中国工业企业数据库(1998-2013)的资本结构

年份	国有企业数量(个)	民营企业数量(个)	国有企业主营业务收入总额(亿元)	民营企业主营业务收入总额(亿元)	国有企业民营企业的总收入比例(%)	国有企业户均规模(千元)	民营企业户均规模(千元)
1998	63 171	20 417	32 322.1	4 265.1	7.6	51 166.0	20 890.0
1999	57 781	23 783	34 434.2	5 627.7	6.1	59 594.3	23 662.5
2000	51 146	34 714	39 622.3	9 232.7	4.3	77 469.0	26 596.6
2001	45 336	46 612	41 171.2	12 568.7	3.3	90 813.6	26 964.6
2002	40 240	60 010	44 544.4	17 679.0	2.5	110 696.7	29 460.0
2003	33 980	75 753	53 870.3	26 495.5	2.0	158 535.2	34 976.2
2004	35 176	123 037	66 798.6	41 114.1	1.6	189 898.1	33 416.0
2005	27 371	135 794	78 572.3	60 234.4	1.3	287 063.8	44 357.2
2006	47 618	200 287	113 192.3	112 141.6	1.0	237 709.1	55 990.4
2007	43 989	237 875	139 019.1	156 575.0	0.9	316 031.6	65 822.4
2008	38 336	322 300	155 019.5	263 734.3	0.6	404 370.5	81 828.8
2009	10 745	87 449	35 959.2	87 743.1	0.4	334 660.0	100 336.4
2010	52 650	233 791	291 998.3	237 205.6	1.2	554 602.6	101 460.5
2011	25 858	232 195	249 345.6	399 203.0	0.6	964 288.1	171 925.7
2012	25 237	237 979	255 677.2	437 615.3	0.6	1 013 105.0	183 888.2
2013	25 664	267 901	261 578.0	534 355.4	0.5	1 019 241.0	199 460.0

资料来源:作者整理。

本文首先以“实收资本总计”是否等于国有资本、集体资本、个人资本、法人资本、外商资本、港澳台资本之和,衡量数据造假的可能性。结果发现:2008 年之前,每年出现此问题的次数均小等于 10 次,可视为基层统计人员的随机填报失误;而 2008 年、2011-2013 年则在 1 065~4 119 次之间;2009-2010 年极为严重(详见表 1)。

究其原因,这并非是资本细项与“实收资本总计”不相等造成的,而是因为 2008-2013 年的资本结构细项指标出现了相对严重的缺失值现象。不考虑缺失值问题后,1998-2009 年出现上述问题的次数均小于 12 次;2011-2013 年则在 108~265 次之间;但 2010 年还是大面积出现此问题。由此看来,2008-2009 年、2011-2013 年部分企业没有上报资本构成细项,尤其 2009 年约有 69.5%的企业没有上报,影响了以上检验的结果。

2010 年数据的真实性存疑较大,为此,本文使用更为严密的方法对数据显示的资本结构进行分析。本文将国有资本占实收资本总计比例大于 50%的,或“控股情况”为“国有绝对控股”或“国有相对控股”的,或“法人资本”为零、同时国有资本为最大资本的,界定为国有企业;将个人资本占实收资本总计比例大于 50%的,或“控股情况”为“私人控股”的,或“法人资本”为零、同时个人资本为最大资本的,界定为民营企业。① 以之为标准计算历年全行业的国有企业、民营企业的数量、主营业务收入总额、户均规模。

从企业数量来看,民营企业数量在高速增长,国有企业数量在平稳下降,两者呈现反向发展态势。2010-2011 年间,由于“规模以上”统计口径的改变导致国有企业、民营企业数量大幅下滑。从企业产出来看,国有企业、民营企业的主营业务收入在 1998-2013 年间均进入高速增长通道,只是受 2011 年规模以上统计口径变动的影响,其增长曲线在 2010-2011 年间不平滑,二者的户均规模也出现了大幅上涨。从市场份额来看,2008 年前,国有企业与民

①在 1998-2013 年 4 321 284 个样本中,有 29 036 个样本因资本构成与控股情况矛盾,可同时界定为国有企业和民营企业。

营企业之间的市场份额比例一直在下降,其后基本维持稳定。无论是数量还是产出,均与近年来社会主义市场经济的发展态势基本相符。

然而,2009年国有企业、民营企业的数量、总产出及国有企业的户均规模均出现了大幅下滑,但本文认为其原因并非数据库本身出现的统计偏误或本文处理方法的问题,而是因为2009年约有69.5%企业没有上报资本细项。这样一来,本文界定国有企业、民营企业的方法会因大面积缺失值而失真,进而导致国有企业、民营企业的划分不准确(主要是低估数量)。这很可能是统计部门在公布信息时刻意删除所造成的,工业统计报表制度的运行其实并没有出现波动,这值得涉及企业产权结构的研究所注意。

2010年数据质量问题则较为严重。首先是,资本细项与实收资本合计不对应的样本量高达34.7万个(见表1)。其次是,国有企业数量出现了与前后年份均不对应的异常增长。最后是,国有企业与民营企业的主营业务收入之比,从前两年的0.4、0.6一下攀升到2010年的1.2,2011-2013年又回到0.5~0.6。上述奇异点的出现,表明2010年工业企业数据的资本结构是异常的,存在恶意编造的可能性。

综上所述,本文使用的中国工业企业数据库(1998-2013)基本真实,但2010年数据质量较差;以上方法可以作为检验研究者手头上中国工业企业数据库真实性的一种手段。本文不赞成研究者对部分期刊公开的数据采用"拿来主义",这是因为未对数据作处理,作者将无法保证其真实性。

四、系统误差与数据问题的评估

(一)样本范围突变及其误差评估

1."规模以上"统计口径的更改

1996年国家统计局根据《工业统计定期抽样调查试点方案》开始将"规模以上"作为一家工业企业是否进入全面调查的统计标准,从而导致1998年数据库样本容量激增。新进入工业统计的企业很可能不适应当时数十张统计报表的填报工作,因此,这些新加入企业的报表填报质量可能相对较差(文强,2012)。

本文使用"开工时间(年)"指标,对1996年前后样本的指标缺失值进行分界,结果发现:1996年(含)前开业的工业企业,在整个填报期间(1998-2013年)内,平均每家企业仅有22.3个缺失值;1996年后开业的工业企业,该指标上升到26.9个。考虑到巨大的样本量,这种差距还是显而易见的。

正如前文所述,2007-2008年和2010-2011年均是统计范围突变的时间节点,但是由于其前后年份的数据库来源渠道不统一,涵盖指标和质量均不一样,因此从缺失值的角度无法判断数据库质量的变化。为此,本文不考虑企业某些指标确实是零值,而以零值的出现作为基层统计人员填报报表质量的考量指标,即假设:统计人员遇到不懂的指标,往往以零值填报,从而损害数据质量。结果发现:剔除产值2 000万元以下(不含)国有企业的2008年数据,平均每个样本填报"零值"的次数从2007年的29.0次下降到12.5次;剔除产值500-2 000万元(不含)样本的2011年数据,平均每个样本填报"零值"的次数从2010年的16.4次下降到9.4次。虽然这种分析相对于缺失值而言较不严谨,但也可以作为数据库样本范围变动而导致质量变化的佐证。

对于今后的学术研究而言,本文建议:在使用中国工业企业数据库的特定指标之前,应对该指标的缺失值、零值问题进行简单的评估,以考察该指标的统计"信度"。

2.“规模以上”界定对数据库样本范围产生的影响

国家统计局企业调查总队在《工业统计定期抽样调查试点方案》中,对于为什么“规模以上”是总产值500万元,其实并未说明理由(申红,1999)。同理,2011年“规模以上”统计口径改为2 000万元以上,也是出于经济发展与通货膨胀的考虑,并非具有统计学意义的划分标准。这自然会对规模以下的工业抽样调查结果产生影响(雷平静,1997),但在大数定律下,对统计工作的影响不大。

而且“规模以上”分界线即具有自然实验中的随机分组效应,在一定程度上可以保证样本分布的随机性(尽管同样有“截断”问题)。工业统计中“规模以上”分界所具有的自然实验效应,与Meyer等(1995)的美国工伤补助标准提升改革研究类似。1980-1982年间美国两个州政府相继提高了高收入人群的工伤补助标准,该文假设州政府对高收入的界定标准(实验分组依据)实际上是由政府随机划分,并基于此进行自然实验研究。通过以上类比,不难看出,上市公司的界定是无法实现随机分组的,因为所有IPO都要经过相关部门的人为审核,数据库所带有的选择性偏误较为严重。更大的问题在于,上市公司数据库样本量相对较小,而且以大型企业为主,截断问题更严重,因此其是否能反映中国经济运行与发展的一般规律值得商榷。

不过,规模以上统计口径变动对经济学研究而言,却是值得考虑的问题。在500(2 000)万元产值左右波动的企业,有些年份进入中国工业企业数据库,有些年份则被剔除,直接导致了样本的选择性偏误(Selection Bias)。而且,这还会使近年兴起的“企业生存(存活)率”、基于企业进入和退出的生产函数半参数估计(如OP、LP、ACF等方法)等研究出现问题。因为,当企业消失于中国工业企业数据库中,很可能并非因倒闭而退出生产,而是没有参与“统计直报”而已。这个问题的危害会因为下文提及的企业统计人员填报问题而加剧。

3.“规模以上”统计口径改变的处理方法初探

该问题的严重性最早被聂辉华等(2012)所发现,但学界似乎至今仍未提出更进一步的处理方法。根据现有的学术文献,常用的处理方法为:在1998-2010年数据库中剔除2 000万以下的样本,便于和2011年后的“规模以上”统计口径统一。

本文认为:这样做看似合理,但实际上损害了数据库的信息。因为,1998年的2 000万元与2013年的2 000万元的产值规模其实量级相距甚远。即便要统一“规模以上”标准,也必须考虑通货膨胀、固定资产投资价格指数变动等因素。另一个要考虑的因素是,企业产值规模是当年价格的名义指标,而不同产业和不同地区的历年工业品出厂价格指数(PPI)和固定资产投资价格指数均是经年波动的。也就是说,在处理“规模以上”统计口径改变时,研究者应考虑到纵向(不同年份)与横向(不同产业与地区)的通货膨胀问题。

有鉴于此,本文建议,可以根据官方统计年鉴或统计公报中的历年PPI或固定资产投资价格指数对产值规模进行通货膨胀平减。不过,由于中国统计部门只公布二位数代码行业及各省、自治区、直辖市的省级PPI和固定资产投资价格指数,因此,同一个行业、同一地区内部只能使用同一通货膨胀率进行折算,这无疑影响了数据处理的精度。另一个可行的办法是,化繁为简——全国所有企业均使用当年全国的PPI或固定资产投资价格指数进行平减,货币型指标使用PPI,资产型指标使用固定资产投资价格指数。

(二)企业回避规模以上导致的误差评估

近年涉及进入和退出的半参数生产函数实证研究、生存率模型研究,都会将工业企业不再出现于中国工业企业数据库视为因倒闭而退出行业竞争的样本。但事实上,一家企业是

否退出市场,是否真正倒闭,只能通过其法律手续来确定。一个法人代码消失在中国工业企业数据库,很可能是因企业易主或其他理由而导致的法人代码变更,甚至可能是基层统计人员的填报失误。

更严重的是,企业回避规模以上统计而导致的问题,即"怕露富"(林文宏,2007)的企业家想方设法调低"主营业务收入",以避免进入《定期统计报表名录库》。因此,"消失于数据库的样本是退出企业"的先验假设,与当前国情较为不符。所谓"退出"其实可能并非代表着企业倒闭,相关学者应注意到此问题(聂辉华等,2012)。

为考察"回避规模以上"现象的客观性和普遍性,本文进行以下数据分析工作:(1)回避企业如果从头到尾一直隐瞒,那么,数据库无法体现这种回避现象,因为这些企业从来没有出现。(2)回避企业中途因隐瞒退出工业统计,其后一直没有回到中国工业企业数据库,那么,数据库也无法分辨其退出是因为倒闭,还是因为隐瞒收入或其他理由。(3)大型企业由于一直受到统计部门的重点监管,并受《统计法》规制,几乎无法隐瞒收入,因此不予考察。(4)这样一来,本文要寻找的是,同一个法人代码的中型企业(2011年前设为1 000万元及以下,2011年后设为4 000万元及以下)在特定年份试图隐瞒主营业务收入,从而使自己变为规模以下而没有进入当年中国工业企业数据库,但后续年份因为《统计法》、企业发展等原因,又重新进入工业统计范围。本文将这种存在隐瞒收入可能性的中型企业,称之为"中途退出企业",此类企业一般也会在生存(存活)率模型研究中剔除。(5)进入本次处理组的中途退出企业,有以下几种可能性:一是确实经营不善,某年的主营业务收入低于规模以上统计口径(500万元/2 000万元);二是企业或统计部门的基层统计人员误报;三是刻意隐瞒收入、回避规模以上的企业。

根据表3的检测结果,30 557家在2011年后曾经中途退出的企业,有17 018家是在2011年当年发生的。这很可能是因为,规模以上的标准由500万元攀升至2 000万元,造成上述企业达不到规模以上标准。从"营业收入/主营业务收入"指标的均值(表3最后一行)来看,这部分样本指标与432.1万个全样本指标是基本一致的,即这部分企业隐瞒收入的可能性较低。

表3　　中国工业企业的回避规模以上问题

样本范围	1999-2010年中途退出的企业			2011-2012年中途退出的企业			全部企业		没有经历中途退出的中型企业	
指标	样本量	主营业务收入(万元)	营业收入(万元)	样本量	主营业务收入(万元)	营业收入(万元)	主营业务收入(万元)	营业收入(万元)	主营业务收入(万元)	营业收入(万元)
个数/均值	52 273	659.1	1 833.2	30 557	2 863.2	2 903.8	18 145.2	19 303.9	651.9	941.3
比例	–	–	2.78	–	–	1.01	–	1.06	–	1.44

资料来源:作者整理。

对于学术研究而言,这从侧面佐证了:2011年之后,由于"规模以上"统计口径调整,企业隐瞒收入的难度激增,中国工业企业数据库的数据质量得到了提升。这也是因为,规模过小的企业生产经营缺乏稳定性,使规模临界点(500万元)附近的工业单位变动较大,给统计工作的连贯性和准确性造成负作用(张德宽等,2002)。

然而,1999-2010年间的中途退出企业却出现较大的异常——以均值看,"主营业务收入"在"营业收入"中的占比异常低。考虑到1 000万元规模以下的中型企业经营特性,这个占比还是远低于63.7万家没有经历中途退出的企业。这导致5.2万家中途退出企业的"营

业收入/主营业务收入”为2.8,是其他1 000万元规模以下企业的近两倍。

由于“主营业务收入”是进入工业统计的决定性变量,表3中途退出企业的“营业收入/主营业务收入”异常值,从统计数据上证明了:上文提及基层统计局公务员所反映的,规模不大的企业热衷于隐瞒收入(林文宏,2007),从而导致工业企业统计的系统性误差。

在统计合规性领域,中国只有一部《统计法》。企业通过一些会计手段,刻意调低营业收入中的主营业务收入,是游走在合法与违法边缘的。加上《统计法》的司法救济力度不大,企业即便被统计部门抓获篡改数据,由于统计部门的行政执法权限小,进入司法程序繁琐,造成1998-2010年间企业隐瞒收入现象频发。

对于今后的学术研究而言,本文建议:(1)在使用“主营业务收入”时,要考虑中型企业的隐瞒收入行为。(2)如果大量企业因隐瞒收入,而在1998-2010年间中国工业企业数据库中消失,那么,这些企业不应作为“退出”、“倒闭”的样本进行处理。这样的企业进入退出、生存率研究,似乎不符合中国国情(主要是工业统计报表制度),会导致一定的统计误差。(3)除了剔除“中途退出企业”,研究者在判断“退出企业”时,还需参考样本前一年的亏损额、利润总额等与倒闭息息相关的经济指标,甚至必要时进行人工处理。

(三)“化整为零”行为的误差评估与样本个体识别问题

本文使用同一个县区(“六位数行政区划代码”)内,同一年内,出现两家“法人代表”相同而法人代码不相同的企业,作为“化整为零”行为的证据。数据分析结果显示,共有106 352个样本①出现“化整为零”现象,即同一个法人代表在同一个县区内,同一年内,拥有两家或以上的工业企业,约占整体样本量的2.5%。

研究结果发现:正如基层统计工作人员反映的一样,工业统计中的化整为零现象较为严重。由于企业家的“跨界”经营、多样化经营、并购、设立分公司等投资战略均会导致上述数据现象发生,因此本文不能揣测这种“化整为零”的行为不是因生产经营的实际需要,而是因避税、“怕露富”等企业家行为所导致的。

然而可以肯定的是,“化整为零”行为会影响学者对于企业样本个体的识别精度。以学术界常引用的Brandt等(2012,2014)为例,该文将“邮政编码”、“四位数行业代码”、“主要产品”、“所在县名称”、“开工年份”相同的企业视作同一个企业,在基层统计“化整为零”现象的影响下,这种企业身份匹配方法可能导致“过度识别”问题。

首先,倘若企业家因各种原因而“化整为零”——在工业统计中申报两家企业,那么,Brandt等(2012)的识别法将会视这两家企业为同一家企业,因为二者的“邮政编码”、“四位数行业代码”、“主要产品”、“所在县名称”、“开工年份”是相同的。但是,将这两家企业的总产量、总资产等指标或合二为一,或合并加总,却均不能很好地识别这家企业。

其次,四位数行业代码与主要产品其实均是统计意义上的粗略划分,与实际的工业生产相距甚远。在实地调研过程中,本文发现:一旦部分工业园区开业,同一个“邮政编码”的地区就会出现很多“所在县名称”、“成立年份”一样的企业;而且在产业链与上下游分工中,不同的企业可能在同一最终产品生产的不同环节进行分工,因而它们的“四位数行业代码”和“主要产品”也是一样。其中,此问题在精细化工、石化及塑料制品等产业尤为严重。Brandt等(2012)将其识别为同一家,就会导致所谓的“过度识别”问题。

①含22个样本的法人代表为“厂长”,28个为“经理”,49个为“负责人”,诸如此类的个体会影响评估精确度。

“过度识别”问题较早由杨汝岱(2015)所发现,为此该文指出 Brandt 等(2012)的方法会使“本来不属于一家企业,但被他们匹配成一家企业了”,并提出以“行政区划代码”、“电话号码”、“成立年份”作为企业身份识别的办法。此方法可以在一定程度上缓解“过度识别”问题,值得推广。聂辉华等(2012)也曾提出,可以按“单位名称”、“电话号码”、“法人代表”分别进行三次分组,再考察同一名称组下的企业是否分属不同的法人代码组,然后进行“交叉匹配”或人工识别。

事实上,无论何种匹配方法,均不能完全克服“过度匹配”问题。正如基层统计人员发现“化整为零”现象普遍存在,那么即便“法人代表”、“电话号码”一样的两家企业,也可能是企业家在同一县区设立的两家分厂,而在避税、“怕露富”的心态下,才将其登记为两家企业。这样一来,将两家分厂合并为一家企业,也是一种“过度识别”。为此,本文认为:“法人代码”指标是中国工业企业数据库中样本企业的唯一“身份证”,企业身份的识别最好只依托法人代码。即企业样本的身份识别应“化繁为简”——只要法人代码不同,企业就是不同的两家企业。

其实学术界可能有所不知,全国工业统计“直报”及其互联网信息系统建立后,县、区一级统计部门的主要工作是:复核每家企业的年报、季报,不断给有疑问的企业打电话沟通指标错漏事宜。因此,法人代码的错漏是极为罕见的,况且“化整为零”现象已经被统计部门及其基层统计人员所关注。总之,本文认为,不同法人代码的企业,不论其最终控制人或者法人代表电话号码是否一致,均可视为不同的企业,而无需使用其他指标进行识别。如果法人代码缺失,综合使用聂辉华等(2012)、杨汝岱(2015)等的方法进行匹配,是可行的。

从某种意义上看,本文的企业身份识别使用的是“序贯识别法”,其实学界也可以使用王贵东(2017)的“交叉识别法”来鉴定“化整为零”问题,以处理共享法人代码或共享企业名称的“重复”样本。①

(四)指标缺失问题

根据本文使用的中国工业企业数据库,2008 年后很多关键指标,如工业增加值、累计折旧、当年折旧、工业中间投入等指标局部缺失。甚至部分研究者手上某些年份的数据连固定资产、从业人员平均人数这些生产投入要素的基本指标都没有。

具体问题应该具体分析,但亦有一个简单而又困难的方法——搜集新的数据库进行匹配与合并,其简单在于搜集到新数据库意味着新指标的加入,其困难在于数据库搜集的难度。例如,笔者搜集到国家统计局相关机构提供的部分年份“基础数据库”,该数据库不仅提供了全国所有纳入官方统计的农业、工业、服务业等全行业的企业层面指标,还有中国工业企业数据库没有公布或部分年份缺失的“年末从业人数”、“资产总计”等。经过平滑化、均值处理等方法,研究者就可以从这些新指标中,获得缺失指标的估计值。

在部分生产函数、成本函数的实证研究中,工业增加值、工业中间投入、当年折旧是必不可少的解释变量指标,一旦上述指标缺失将严重影响回归估计的时间跨度,甚至制约着实证结论的时效性。由于指标缺失,研究者一般使用估算法对一些指标进行补充,如刘小玄和李双杰(2008)的工业增加值补充法,陈林等(2015)、陈林和朱沛华(2017)的当年折旧估算法。这样做是可行的,是对数据库本身的一种补充。

本文建议,在增补或估算关键指标前,研究者应该适当地学习工业统计报表制度与历年

①在企业身份识别方面衷心感谢两位匿名审稿人详尽的宝贵建议。

会计准则,掌握每个指标在不同年份的定义与特性。如使用增补指标作为关键的被解释变量或解释变量时,则务必对数据进行更为细致和谨慎的分析与整理。

以“中间投入”指标估算为例。首先,在成本函数的经典研究(范建双、李忠富,2009,2010)中,一般可以使用存货作为“中间投入”的代理变量。因为,现行工业统计报表制度指出,存货“通常包括原材料、在产品、半成品、产成品、商品以及周转材料等”。但这或许是较为粗糙的估算方法。

其次,现行工业统计报表制度规定,主营业务成本(执行《企业会计准则》的企业则为营业成本)指的是,生产成本中存货已经销售的部分,即不包括存货中的“原材料”。其主要包括以下生产成本指标:直接材料消耗、直接人工、其他直接费用、制造费用,等等。由于其他直接费用与制造费用,没有在中国工业企业数据库中报告,只能对此忽略。那么,中间投入约等于存货中的“原材料”和主营业务成本中的“直接材料消耗”。综上所述,“中间投入”的估算值=“存货”-“存货中的产成品”+“主营业务成本”-“主营业务应付工资总额(或‘本年应付工资总额’)” -“主营业务应付福利费总额”。

最后,可以使用税务指标,如本年应交增值税、本年进项税额、本年销项税额等,结合产值指标,对原材料投入金额进行估算。具体方法有待后续的专题研究进行拓展,本文在此仅提出思路。

除了上述变量缺失外,像广告费、研发投入、女职工人数、无形资产等指标的缺失,也会影响该数据库在部分学科中的应用推广。但不同领域的论文写作过程中遇到的问题层出不穷,既然不能穷尽所有应对办法,本文关注更多的是,如何去推广中国工业统计报表制度及其数据库。只要理解好该制度及其指标算法,增补指标自然会得心应手。值得注意的是,无论使用何种估算方法,人为地、主观地增补指标,无疑会加剧计量模型的内生性等问题,因而应慎用。

(五)指标统计口径变化与经济普查导致的数据问题

每年年中,国家统计局会公开发布《工业统计 xxxx 年(当年)年报和 xxxx 年(下一年)定期报表填报说明》,对需要修订统计口径的指标和“五套表”的修订进行详细的说明。根据有据可查的留存文档,该制度最早可追溯到 1986 年。在此之前,国家统计局则是以内部公文形式,对工业统计报表每年需要修订的内容进行传达。

每年的指标统计口径修订会对数据库产生显著影响。以“补贴收入”指标为例,这个变量在 2000 年、2006 年均出现了重大的统计口径变化——自 2000 年起,该指标从原来的“国家补贴收入”改为“补贴收入”,指标定义变得更广;2007 年起,工业统计报表开始执行《会计准则(2006)》,“补贴收入”的计算产生了重大变动。2000 年的统计制度变动使补贴收入的定义更广,直接导致 2000 年全行业“补贴收入”均值从 1998 年和 1999 年的 16.6 万元和 16.8 万元,激增至 19.9 万元,并在 2000-2004 年间维持在 20 万元左右波动。2007 年的统计制度变动使无偿性成为“补贴收入”的必要条件,并计入“营业外收入”,而政府的资本性投入不再属于“补贴收入”。2007 年之后补贴收入采用收益法计提,资本性补贴不再属于“补贴收入”。具体而言,政府补贴一个企业技术改造项目,在原会计准则下完工后“补贴收入”立即进入所有者权益项目;但在新准则下,“补贴收入”会在当年进入“递延收益”,在技改设备使用寿命内按其折旧或摊销进度分配至“营业外收入”项目(郑仙萍、毛丽娟,2007)。

制度的变迁会明显改善获得资本性补贴的企业以后各期的总利润水平和净资产收益率,但也会大幅减少当年“补贴收入”的数值。从数据上看,2007 年的户均补贴收入水平从

2006年的28.3万元骤降至24.7万元,降幅高达12.7%,为历年最高水平。

近年来,关于政府补贴、产业政策的制度绩效重新引起学者们的热议。Aghion等(2015)正是使用中国工业企业数据库(1998-2007),围绕"补贴收入"指标构建了考察产业政策的关键变量,并得出了政府补贴、产业政策等政府干预市场的做法可以在竞争性行业获得成功的结论。但"补贴收入"指标在2000年和2007年的突变,是否意味着需要剔除1998年、1999年及2007年的样本进行稳健性检验?当一个关键指标出现两次巨大的统计口径变动,如果对工业统计报表制度不熟悉,对自身掌握的数据库不了解,上述稳健性检验也就较为必要了。

其他方面的统计口径变动也值得研究者留意,比如,2010年开始有资质的钢结构企业不再作为工业企业上报,2011年"利息支出"的定义按照《会计准则(2006)》进行修改,而之前还是沿用原来的会计准则。随着中国制造业的飞速发展与工业统计制度的变迁,工业增加值、工资、福利费等统计口径也经年变动。而在近年国家大力推行的混合所有制改革下,企业产权结构及其所有制属性的界定,也将是学术界、实务界关注的热点。因此,如何量化特定企业的所有制属性、"混合所有"后的产权结构,有待后续研究探索。

除了指标统计口径的变化外,2004年、2008年中国工业企业数据库很可能来自于2004年和2008年的经济普查,这也导致了数据库的指标均值波动。由于经济普查在统计制度上的差异,国家投入的人力物力也远高于工业统计年报,因此,2004年和2008年工业企业指标上报时可能更严谨。这在数据的变化上也有反映——工业企业各种产值、资产、财务指标,在这两年均出现了较大的跌幅,与中国1998-2013年间的经济上行周期严重不符。

因此,经济学研究在使用2004年和2008年中国工业企业数据库时,需要进行一些指标上的处理或更严谨的稳健性检验。

(六)其他问题

1. 名义指标与通货膨胀平减问题。对于工业产品价格与资产定价,不同产业和不同地区的价格指数是不同的,因此通货膨胀与价格指数应该在使用数据时有所体现。正如前文提及的,本文认为可以用以下三种方法进行价格指数平减:(1)分别使用官方统计资料中公布的二位数代码行业及各省、自治区、直辖市的省级PPI和固定资产投资价格指数,对货币型、资产型指标进行通货膨胀处理;(2)使用全国统一的PPI和固定资产投资价格指数进行通货膨胀处理;(3)部分年份的海关数据有出口商品单价,假定国际市场进行的是古诺(产量)竞争,那么,国内外市场的产品单价应该是一致的。因此,可以通过匹配中国工业企业数据库与海关数据库,获得四位数代码行业产品的平均价格。最后一种方法比较细致,值得推广,但近年海关公布的数据库不一定包括单价指标,且商品与行业的匹配工作量也较大。

2. 2008年前后数据合并可能带来的问题。正如前文所示,2009年后的数据准确度、指标健全程度均出现一定程度的下滑,尤以2010年数据为甚。由于本文无法确认其他研究者手上的数据库是否同样出现了上述问题,因而无法穷尽不同数据库的合并注意事项。但有一点是可以肯定的——合并数据库后,研究产业经济的应该考察行业均值的波动是否在合理范围,研究宏观经济与地区经济的应该考察区域均值波动是否平稳,研究微观企业的应该考察样本个体均值波动。总而言之,使用数据库前应检验合并历年数据可能导致的各种问题。

3. 关于剔除样本。剔除关键变量出现缺失值的样本无可厚非,但如果进一步剔除一些不符合会计原则的观测值(总资产小于流动资产,总资产小于固定资产净值,累计折旧小于当期折旧)以及过大或过小的异常值,则需更加谨慎。因为学者不可能获悉基础统计人员在

填报时所掌握的信息。正如前文所示,企业内部的统计口径变化也会造成部分指标不可比。退一步而言,总资产(流动资产或固定资产净值)、累计折旧(当期折旧)指标会不会因为“错报”而变小(大)?如果出现单一指标的“错报”,只要下一步的实证检验未使用到这个指标,那么,使用同一样本没有“错报”的指标进行回归,其信度也是能够得到保证的。同理,出现单一指标异常值的样本,也并非不能使用,除非发现该样本所有关键变量均为异常值。

总之,本文不建议对中国工业企业数据库进行过度的剔除处理,除非研究所必须,否则就会产生数据挖掘(Data-mining)影响回归结果的可能。

此外,本文使用的数据显示,“千元”错填为“元”和“万元”,年报错填为月报、季报,企业内部统计口径变更等误报现象也会出现。但上述三种误差都属于随机误差,对经济学研究的实证结果信度影响不大,所以,本文不对此进行深入分析。①

五、结论与展望

本文介绍了1998-2013年工业企业数据的整理、合并及真实性检验的方法,并对如何有效地整理、合并及使用中国工业企业数据库,尤其是针对新出现的2010-2013年数据,提出了具体的数据处理建议;最后,本文对样本范围及统计口径的变动、企业工业统计人员的填报等数据问题进行定量评估,结果发现部分问题客观存在,并提出对应的处理办法。

中国工业企业数据库在学术用途方面也可以作进一步的推广,这关键就在于合并其他数据库。根据现有的研究,中国海关数据库、历年城市统计年鉴、科技部门的专业数据库、国家知识产权局的专利数据库和《境外投资企业(机构)名录》,是当前与中国工业企业数据库合并的热门对象。其实在数据库合并方面,中国工业企业数据库还大有潜力。上市公司数据库、国家环保局的污染排放监控数据库、全国工商联的私营企业调查数据库等,均是可以考虑的合并对象。近期,2014年中国工业企业数据已经在学术界出现,最新的数据也应该进行相应的检查才能予以使用。

其实,每位学者都应掌握自身所使用数据的结构、特性及其问题,并有相对独立的处理方法,而不应未经思考,援引他人的数据处理方法,甚至直接使用其他学者整理过的数据。从某种意义来讲,这既是不熟悉数据结构的表现,也可以视作一种“取巧”。本文认为这种形式的数据传播,反而不利于学术探讨与智慧传递。因为,每一位初次接触中国工业企业数据库的学者都应该经历本文所探讨的“数据整理”过程,而不能简单地奉行“拿来主义”②。当然本文的数据处理方法仅为一家之言,本研究也仅为引玉之砖,希望能以之引起学界的进一步探讨。

参考文献:

1.陈林、汤秀梅、刘小玄,2015:《全要素生产率会影响成本函数估计吗》,《统计研究》第11期:26-35页。
2.陈林、朱沛华,2017:《一种新的考虑全要素生产率的成本函数估计法》,《数量经济技术经济研究》第5期:88-106页。

①本部分提及的所有数据问题,均为基层统计人员工作过程中发现与归纳的实际问题。因篇幅有限,正文没有把所有统计人员及其参考文献在每一处标注,具体见参考文献目录。

②甚至如本文其中一位匿名审稿人所言:国内大部分研究是直接利用多个学者累积整理的现有数据,但这存在一定风险,比如,如果中间的整理有误,那么后面再整理的人无论多么精确也无用了。而且,后面的学者根本就不知道其整理有误。

3.戴天仕,2014:《工业企业数据库的数据质量分析》,暨南大学工作论文。
4.范建双、李忠富,2009:《中国大型承包商规模经济和范围经济的实证研究》,《数量经济技术经济研究》第2期:47-59页。
5.范建双、李忠富,2010:《中国上市建筑企业规模经济和范围经济评价——一种随机边界成本函数方法》,《数理统计与管理》第5期:861-870页。
6.雷平静,1997:《<工业统计定期抽样调查试点方案>的设计》,《统计研究》第5期:69-74页。
7.林文宏,2007:《进一步做好规模上工业统计工作的思考》,《浙江统计》第7期:34-45页。
8.刘小玄、李双杰,2008:《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)》第3期:74-99页。
9.刘小玄、周晓艳,2011:《金融资源与实体经济之间配置关系的检验——兼论经济结构失衡的原因》,《金融研究》第2期:57-70页。
10.聂辉华、江艇、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:142-158页。
11.申红,1999:《对<工业统计定期抽样调查试点方案>的评价》,《统计研究》第S1期:26-27页。
12.王贵东,2017:《中国制造业企业的垄断行为:寻租型还是创新型》,《中国工业经济》第3期:83-100页。
13.王贵东、周京奎,2017:《中国制造业企业垄断势力测度——兼论市场边界》,《经济评论》第4期:30-44页。
14.文强,2012:《企业"一套表"报送中存在问题及对策》,《行政事业资产与财务》第8期:12页。
15.杨汝岱,2015:《中国制造业企业全要素生产率研究》,《经济研究》第2期:61-74页。
16.张德宽、刘绍辉、陆万明,2002:《山东省工业企业"一套表"实证研究》,《中国统计》第2期:9-11页。
17.郑仙萍、毛丽娟,2007:《<企业会计准则第16号——政府补助>解读》,《财务与金融》第5期:46-48页。
18.Aghion, Philippe, Jing Cai, Mathias Dewatripont, Luosha Du, Ann Harrison, and Patrick Legros. 2015. "Industrial Policy and Competition." *American Economic Journal: Macroeconomics* 7(4): 1-32.
19.Brandt, Loren, Johannes van Biesebroeck, and Yifan Zhang. 2012. "Creative Accounting or Creative Destruction? Firm-level Producvity Growth in Chinese Manufacturing." *Journal of Development Economics* 97(2): 339-351.
20.Brandt, Loren, Johannes van Biesebroeck, and Yifan Zhang.2014. "Challenges of Working with the Chinese NBS Firm-level Data." *China Economic Review* 30(3): 339-352.
21.Meyer, Bruce D., W. Kip Viscusi, and David L. Durbin.1995. "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment." *American Economic Review* 85(3): 322-340.

Re-exploring the Usage of China's Industrial Enterprise Database

Chen Lin
(Business School, Beijing Normal University)

Abstract: China's industrial enterprise database is one of the most commonly used micro-database at enterprise level both for domestic and foreign economic scholars in recent years. From the perspective of statistical sampling survey, however, it's more or less problematic. In order to better understand possible systematic errors and casual errors of this database, this paper uses the data from 1996 to 2013 to make quantitative assessments about its authenticity, systematic errors and various data problems reflected by the grassroots statistics department. The results show that the fluctuations of sample range and statistical caliber, a large number of missing values, and the problems related to enterprise statistics personnel' filling and reporting such as "the avoidance of above designated size" and "breaking up the whole into parts" will produce objective impacts on the data. Accordingly, this paper proposes corresponding solutions and suggestions.

Keywords: Chinese Industrial Enterprise Database, Industrial Enterprises above Designated Size, Industrial Statistical Reporting System, Micro Enterprise Data, Industrial Added Value
JEL Classification: C41, C42

(责任编辑:陈永清)