**税调数据处理思路说明**

**一、数据清洗思路**

在对数据正式进行清洗以前，需要强调我们对原始数据进行了统一变量名和所属行业代码的预处理：

**首先是变量名统一**，由于原始数据中大多数数据名称为v1、v2…形式，难以辨识，同时不同年份数据之间的变量数量也不尽相同，因此需要根据数据说明表对各变量名称进行统一。具体结果请参见“**20230926-税调变量名称统一.xlsx**”文件，具体匹配程序请参见“**税调企业碳排放测算-step0 统一变量名称**”文件夹。

**注**：由于更新了2017-2020年的税调数据，所以请重点参考“**税调企业碳排放测算-step0 统一变量名称**”文件夹中的代码。

**其次是初步剔除，考虑到我们大多研究主要是围绕能源环境问题展开，所以**我们先依据原始数据中企业的能源消费量（煤、油、电、天然气和其他气体燃料）进行初步剔除，删除所有能源消费数据均未汇报的观测样本，以及能源数据中存在负数的观测样本。具体程序请参见“**税调企业碳排放测算-step1 初步剔除.do**”。

**再次是工商数据库匹配**，此举是为了更好地实现税调数据与其他数据库（专利、工企、土地出让、海关、污染排放…）的互联互通。我们希望将**统一社会信用代码和企业工商注册名称作为日后**与其他数据库进行匹配的唯二依据。

具体而言，我们先对税调数据库中“id”和“qymc（企业名称）”两个字段进行清理。其中，id主要由15位（6位区划代码+9位组织机构代码）及18位（统一社会信用代码，2位类型码+6位区划代码+9位组织机构代码+1位校验码）构成，我们在清洗过程中统一为11位代码（qyid，2位省级区划代码+9位组织机构代码），具体程序请参见“**税调企业碳排放测算-step1 字段生成.do**”。企业名称的清晰思路可以参见“**税调企业碳排放测算-step2.1 数据匹配.do**”“工商数据库预处理过程”（文件夹，原始数据链接**https://pan.baidu.com/s/1DqtLea7a1\_Pn6XTnJ6L0zg?pwd=webt**）。

**进一步地，**我们先根据qyid变量与工商数据库（库中包含18位统一社会信用代码，我们也将之清洗为11位代码）进行匹配，将匹配失败的观测样本用qymc变量再进行二次匹配。具体运行程序请参见“**税调企业碳排放测算-step2 数据匹配.do**”“**税调企业碳排放测算-step2 工商匹配.do**”。

**接着是所属行业代码及行政区划代码统一**，参考田彬彬等（2020）的做法，我们将税调原始数据中“industry（对应2007）”“hylb（2008-2011）”视为CIC2002版本，并统一转换为CIC2011版本（该版本于2012年正式使用）。具体转换代码请参见“**2011版统一为2002版stata代码do文件.do**”，实际在“税调企业碳排放测算-step3 折算碳排放因子\_行业和区域.do”中运行。

**注**：由于数据已经扩充至2020年，我们还需要将2017年版的行业代码统一至2011年版，具体数据请参见“Plus Step1 Data 2017版统一为2011版对照表.dta”。

此外，我们采用折算系数法对企业碳排放量进行测算，具体思路请见下文“二、碳排放核算方法”，具体程序请参见“**税调企业碳排放测算-step3 碳排放折算系数.py**”“**税调企业碳排放测算-step Plus 合并.do**”。

**注：**由于官方公布的电网平均碳排放因子数据时效性和颗粒度较差（仅有2010、2011、2012及2021年以后的相关数据），我参考了一篇2005-2021年省级电网平均碳排放因子的测算论文（请参见“**$2005—2021年区域和省级电网碳排放因子研究**”）。另外，由于作者不愿意分享具体的测算结果，我采用了邪道做法，用Origin软件提取了图中的具体数值，具体请参见“**$2005-2020年省级电网平均碳排放因子**”

**最后，**为了方便大家对税调企业进行更多的拓展性分析，我们首先将各个年份的数据进行了合并，具体参见“税调企业碳排放测算-step Plus 合并.do”。此外，我们还基于企业工商注册地址和企业名称，运用高德地图api对企业的所处省份、城市、区县和经纬度坐标进行解析。具体程序请参见“**税调企业-注册地址查询.py**”。同时，我们结合碳排放栅格数据，依次提取了税调企业周围1~5km范围内的二氧化碳排放量之和。具体程序可以参见“**提取点周围的栅格值.R**”（该方法也在后文有具体说明）。

**二、碳排放核算方法说明**

**（一）系数折算法（基础版，Cui2021）**

参考Cui等（2021）的做法，根据各类能源消费的碳排放因子，核算能源消费所产生的直接碳排放（主要来自燃料燃烧），具体的碳排放因子见表1，其中，电力能源消费引致的碳排放考虑了中国6大区域电网的碳排放系数差异。

另外，考虑Cui等（2021）忽视了天然气与其他燃气消费量数据在2014年之前不可得的问题，拟计算两类数据以分别用于基准回归分析与稳健性检验：一是按照Cui等（2021）的做法，在计算能源消费总量时，2014年之前不核算天然气与其他燃气消费引致的碳排放，而2014年之后则核算天然气与其他燃气消费引致的碳排放；二是仅使用电力、煤炭、成品油的消费量进行碳排放核算。另外，税调企业的电力、煤炭、成品油和天然气等各类能源消费的指标符号如表2所示。

**表1 各类能源消费的碳排放因子**

|  |  |  |
| --- | --- | --- |
| 能源种类 | 单位 | 折算系数 |
| 煤炭 | kgCO2/kg | 1.978 |
| 油 | kgCO2/kg | 3.065 |
| 天然气 | kgCO2/m3 | 1.809 |
| 电力消费量： |  |  |
| 北京、天津、河北、山东、陕西、内蒙古 | kgCO2/kWh | 0.8843 |
| 辽宁、吉林、黑龙江 | kgCO2/kWh | 0.7769 |
| 上海、江苏、浙江、安徽、福建 | kgCO2/kWh | 0.7035 |
| 河南、湖北、湖南、江西、重庆、四川 | kgCO2/kWh | 0.5257 |
| 陕西、甘肃、青海、宁夏、新疆 | kgCO2/kWh | 0.6671 |
| 广东、广西、云南、贵州、海南 | kgCO2/kWh | 0.5271 |

数据来源：Cui等（2021）

**（二）系数折算法（改进版）**

**Cui等（2021）未能有效考虑碳排放的行业异质性问题。**我们基于历年中国能源统计年鉴公布的“工业分行业终端能源消费量（标准量）-年份”和“全国能源平衡表（标准量）-年份”**分年度分行业折算碳排放因子（具体请见附件：碳排放因子折算）**。我们以制造业中纺织业为例，对应2位码行业，具体折算步骤如下：

首先，我们从中国能源统计年鉴中获取纺织业不同品类能源的终端消费量。其中，终端消费量是指行业直接消费的能源扣除用于加工转换二次能源的消费量和损失量以后的数量，直接消费的能源是指行业中用于燃料、动力和原材料的能源。**需要指出的是**，尽管能源统计年鉴给出了细致的多品类的能源消费情况，但与税调数据库的能源消费核算相比，**能源统计年鉴给出的能源消费量多出了用于原材料的能源消费量部分**。2020年纺织业终端能源消费量如表2所示，经过计算原煤的加权值为174.82/(174.82+3.93+3.82)，煤的碳排放因子为1.927（崔的结果为1.978）。同理，油的碳排放因子计算结果为3.0381（崔的结果为3.065）。

**表2 2020年纺织业终端能源消费情况**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 能源种类 | 能源消费量(吨标准煤) | 低位发热量(GJ/t) | 单位热值含碳量(tC/TJ) | 燃料碳氧化率 |
| **煤合计** | **207.58** |  |  |  |
| 原煤 | 174.82 | 20.908 | 26.37 | 0.94 |
| 洗精煤 | 3.93 | 26.344 | 25.4 | 0.9 |
| 焦炭 | 3.82 | 28.447 | 29.5 | 0.93 |
| **油合计** | **17.34** |  |  |  |
| 汽油 | 4.81 | 43.07 | 18.9 | 0.98 |
| 煤油 | 0.02 | 44.75 | 19.6 | 0.98 |
| 柴油 | 5.54 | 42.652 | 20.2 | 0.98 |
| 燃料油 | 1.67 | 41.816 | 21.1 | 0.98 |
| 润滑油 | 0.15 |  |  |  |
| 石油沥青 | 2.06 |  |  |  |
| **天然气** | **202.85** | **389.31** | 15.3 | 0.99 |

注：空白部分代表数据缺失。

进一步地，由于税调数据并未提供企业购入的电力消费数据，Cui等（2021）无法核算间接碳排放。而且，Cui等（2021）仅考虑了化石燃料燃烧引致的碳排放，但未能针对企业生产过程导致的碳排放量进行核算。该文为了排除这一误差，将可能产生大量工业过程排放行业（包括钢铁、化工、石化、水泥、石灰、玻璃制造等行业）的企业样本进行删除，并重新进行模型估计以开展稳健性检验。我们可在此基础上，考虑各行业生产流程差异，计算各行业平均碳排放强度（即碳排放/总产值），并结合企业总产值数据反推得到企业碳排放量。

另外，可考虑企业的股权穿透，在厘清企业上下游股权关联关系的基础上，再次核算企业碳排放以进行稳健性检验（**本次数据处理并未进行企业股权穿透，感兴趣的老师同学可以向我索要简易的股权穿透数据程序**）。事实上，某些企业虽然自身碳排放量较小，但其控股或入股的上下游关联企业可能排放了大量的二氧化碳。为排除具有股权关联的企业之间在能源消费方面的统计误差，可基于工商数据库识别税调数据库企业的股权关联企业，若税调企业与其股权关联企业同属一个行业（一般为四分位行业），则将该税调企业样本剔除（如果股权关联企业也在税调数据库中，也应一并剔除）。

**（三）夜间灯光数据反演法**

可结合夜间灯光数据反演的碳排放栅格数据，并使用企业地理信息，对企业碳排放量进行估算。具体而言，首先解析企业的精确经纬度坐标（至少到小数点后两位），再根据企业缴纳的城镇土地使用税估算企业用地面积，进而以其经纬度坐标为圆心提取附近x千米内所有栅格碳排放量的最大值或加和或中位数，最终得到企业的碳排放量。其中，x应取企业用地面积的开方（企业用地面积根据土地使用税估算得到），小于1则按1取值。此外，若在实际测算过程中城镇土地使用税数据缺失过多，则以1、2、…千米作为x的取值，再次获取栅格碳排放量的加和，并计算得到企业的碳排放量。

**三、跨数据库匹配思路**

考虑税调数据库中的企业标识码大部分为15位纳税人识别码（对于部分18位纳税人识别码的情况，可能因该企业办理过“三证合一”），而工业企业数据库、企业污染数据库、海关数据库、专利数据库、土地出让数据库等其他数据库的企业标识码为法人代码，如需完成上述数据库的嵌套使用，有两种匹配思路：其一，直接根据企业名称信息进行模糊匹配；其二，首先将税调企业与工商数据库进行匹配，以获取企业注册信息等企业基础信息，基于此可完成税调数据库与上述其他数据库之间的匹配。税调数据与工商数据匹配思路如图1所示。



**图1 税调数据与工商数据匹配思路**

参考文献

[1] 范子英，程可为，冯晨，2022：《用地价格管制与企业研发创新：来自群聚识别的证据》，《管理世界》第8期。

[2] 龙小宁，张靖，2021：《IPO与专利管理：基于中国企业的实证研究》，《经济研究》第8期。

[3] 陆菁，鄢云，黄先海，2022：《规模依赖型节能政策的碳泄漏效应研究》，《中国工业经济》第9期。

[4] 宋弘，封进，杨婉彧，2021：《社保缴费率下降对企业社保缴费与劳动力雇佣的影响》，《经济研究》第1期。

[5] 田彬彬，陶东杰，李文健，2020：《税收任务、策略性征管与企业实际税负》，《经济研究》第8期。

[6] Cui, J., C. Wang, J. Zhang, Y. Zheng, 2021, The Effectiveness of China's Regional Carbon Market Pilots in Reducing Firm Emissions, *Proceedings of the National Academy of Sciences of the United States of America*, 118(52), e2109912118.

[7] Fan, Z., Y. Liu, 2020, Tax Compliance and Investment Incentives: Firm Responses to Accelerated Depreciation in China, *Journal of Economic Behavior ＆ Organization*, 176, 1—17.