DOI:10.19511/j.cnki.jee.2020.04.009

环境大数据应用的最新进展与趋势

钱浩祺*

摘要:社会科学研究中更好地利用环境数据开展定量研究,是帮助我国提升环境治理能力的重要一环。本文基于大数据5V模型提出了环境大数据体系的概念,并基于该数据体系构建了针对环境大数据应用的分析框架,从宏观环境统计数据、微观环境数据、环境监测数据、卫星数据以及异构数据五个类别分析环境数据在我国社会科学研究中的应用现状与不足。研究发现,在环境大数据体系中,不同类别的数据各有其优势与劣势,相比传统环境统计数据,新形式的环境数据虽然在时间频度和数据粒度上得到了巨大的提升,但是其提供的环境信息种类较少,且数据质量参差不齐,目前主要适用于相对有限的环境问题研究,但是在未来有着较大的提升潜力。为了进一步拓展和深化环境大数据的应用,则需要从提升数据质量、引入新研究方法以及加强协同合作三个方面来进行改进。

关键词:环境大数据;微观环境数据;监测数据;卫星数据;异构数据

一、引言

大数据及其相关技术在过去十余年中飞速发展,对当今社会运行与经济生活产生了巨大影响,人们的生活方式、企业生产与管理、市场商业模式以及政府决策过程都或多或少发生了转变。社会科学研究同样如此,当研究者们面临数量更大、来源更广的数据时,会主动去挖掘这些数据中所蕴含的新信息,也会竭尽所能探索全新的研究方法来收集并处理海量数据。我国生态环境部(原环境保护部)于2016年颁布了《生态环境大数据建设总体方案》,进一步加

*钱浩祺,复旦大学全球公共政策研究院,复旦–LSE全球公共政策研究中心,复旦大学能源经济与战略研究中心,上海市大数据社会应用研究会,邮政编码:200433,电子信箱:qianhaoqi@fudan.edu.cn。

本文是国家重点研发计划资助课题“应对气候变化科学数据与知识集成共享平台建设”(2018YFC1509007)、国家自然科学基金青年科学基金项目“碳排放峰值约束下的中国绿色电力转型研究——基于电力大数据与中国多区域CGE模型”(71703027)、国家杰出青年科学基金项目“能源环境经济与政策分析”(71925010)、国家社会科学基金重大项目“基于大数据的宏观经济现时预测理论与方法研究”(15ZDB148)的阶段性成果。感谢匿名审稿人提出的宝贵意见;文责自负。

大在环境大数据方面的投入，实现全面提高生态环境保护治理能力。在该方案所提出的“一个机制、两套体系、三个平台”的生态环境大数据总体架构中，目前阶段主要以环保大数据的基础设施平台与数据资源平台建设为主，对于大数据应用平台的建设，还远远没有达到充分挖掘环境大数据潜力的阶段。并且从实践的角度看，目前在各地区具体实施过程中还存在着建设进展不一、机构设置不完善、数据共享与公开较为有限以及大数据应用不足等问题（赵海凤等，2018；张毅等，2019）。虽然环境大数据的建设和应用还处于起步阶段，存在着不少的困难亟需克服，但是其在环境管理、经济管理和社会管理方面已经开始发挥重要作用。与此同时，社会科学研究作为帮助我们认识环境污染问题、评价环境政策效果以及优化环境经济与社会管理的重要基础支撑，随着环境大数据的出现与兴起，正逐步开启一个全新的“第四研究范式”纪元（米加宁等，2018），这对于环境交叉研究而言，既存在着无限的潜力，同时也充满了各式挑战。

环境大数据为社会科学研究带来的好处是多方面的。第一，相比于利用传统宏观数据进行趋势分析，研究者们能够从更微观的污染物排放主体以及高时空分辨率视角去开展研究，从而挖掘出原本可能因宏观加总而被掩盖的个体异质性和时空异质性等特征。第二，利用环境大数据能够开展更加精确的环境经济核算工作，为不同政策、不同项目提供更加精细的成本收益分析，从而实现资源有效分配。第三，随着社会科学其他领域不断引入大数据，深入应用环境大数据能够使环境交叉研究更好地与之相匹配，提供更全面的研究视角。

而环境大数据所面临的挑战则主要集中在数据和研究方法方面。首先，环境问题本质是从自然科学问题所引申出来，因此对于社会科学研究者而言就存在天然的研究壁垒。第一个壁垒在于数据本身，由于环境研究使用的数据来源多样、数据特征差异较大、数据管理与发布主体复杂，因此社会科学研究者在获取数据、识别数据和校验数据的过程中需要跨越极大的门槛（赵苗苗等，2017）。第二个壁垒在于对环境专业知识的理解与应用，不同污染物的环境暴露与健康危害机制存在差异，即不同污染物在影响的空间范围、时间尺度以及影响对象方面是不一样的，需要专业理论与知识的支撑，而这些差异又将影响如何去发现科学问题以及如何应用正确的分析方法。其次，环境大数据对于研究所需要的技术和方法也提出了全新的挑战。由于新数据在特征和体量方面完全不同于传统社会科学研究所涉及的数据，需要适当地采用新技术来分析不同的环境问题，例如机器学习、复杂网络分析以及自然语义处理等方法，都有可能在关键问题上提供全新的研究视角和解决思路。因而，涉及环境大数据的研究将不断促使社会科学研究者了解和掌握新的工具方法，并与其他专业领域的研究者开展深入合作。

但是，环境大数据为社会科学研究带来的综合影响毫无疑问是正面的。从整体而言，虽然目前环境大数据在社会科学研究中的应用依然处于起步阶段，但是已经产生了一系列极具

影响力的研究,为政府决策者制定有效的环境保护措施以及社会大众了解环境污染产生的深层原因提供了重要的实证经验依据。目前的研究主要集中于利用环境大数据构建新指标、分析企业环境效率、评估政府环境治理和政策效果以及估算环境健康效应等方面。而从专业领域看,目前的探索性研究和应用则主要集中在应用经济学与公共管理这两个社会科学二级学科,鉴于此,本文将主要聚焦于这两个领域的现有文献来对环境大数据的应用进展进行分析与评估。

本文的结构安排如下:第二部分对环境大数据的概念进行界定,并提出本文的分析框架;第三部分分析环境大数据的相关分析技术和方法在现有研究中的应用现状,找出目前社会科学研究中应用环境大数据的局限与困难;第四部分对未来进一步应用环境大数据的潜在突破方向进行分析并提出展望,最后对全文进行总结。通过本文的研究和分析,期望有助于国内环境交叉研究领域的研究者进一步拓展和深化环境大数据的应用。

二、环境大数据体系

根据业界与学界公认的大数据5V模型,大数据拥有五大特征维度:数量(Volume)、种类(Variety)、时效(Velocity)、价值(Value)和准确性(Veracity),目前社会科学研究中所使用的新型环境数据虽然在数量维度上无法完全满足大数据的严格定义,但是根据学界所形成的共识,新型环境数据实质上已经具备了大数据的基本特征,即相对于传统数据而言具有更大的数据规模与复杂性,具有5V模型所描述的特征(赵苗苗等,2017;赵海凤等,2018);通过大量数据的整合与分析,能够发现新知识、创造新价值(米加宁等,2018)。

首先,从数量维度看,虽然现有研究中所使用的环境大数据还不足以称之为真正意义上的"大"数据,其容量往往最多达到MB或GB级别,与大数据业界通常所宣称的TB级别甚至是PB级别相去甚远①,但是相比于过去的研究中仅使用数百或数千个宏观观测值的做法,其数量级已经呈现了几何级数的增长、数据的复杂度也不断增加。数据容量与复杂度的提升,也促使研究中不断引入新的分析技术来有效处理和分析数据,此时,大数据的其余四个特征维度所扮演的角色就愈发重要。鉴于此,在讨论环境大数据在社会科学研究中的应用时,必须关注到其他四个特征维度,而非仅仅关注数据的体量是否足够大。

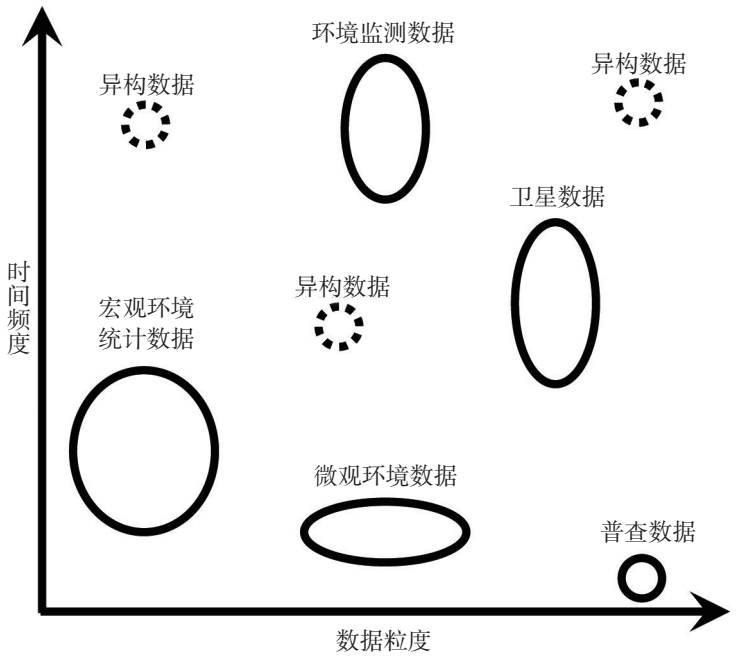
其次,考虑其他四个特征维度,环境大数据已经具备大数据的基本内涵。从种类维度看,环境数据的形态日趋多样,从最初以统计数字与监测数字为主,到现在开始使用卫星遥感数据、位置数据以及文本数据等,已经初步具备大数据的多源异构特征。从时效维度看,部分环

①MB、GB、TB和PB都是存储单位,此处用来表征数据在计算机中的容量大小。存储容量的基本单位是字节(Byte),1MB=1048576B,1GB=1024MB,1TB=1024GB,1PB=1024TB。

境数据已经能够做到“准实时”级别，例如部分在线污染监测系统能够做到每小时向公众及时公布周边环境污染及重点企业排污信息，部分卫星遥感数据也能在短短几天内公布经过校准计算的数据，这对于及时进行环境监管与污染预报起到了重要的作用。从价值维度看，环境数据的价值密度随着其容量的不断提升而下降，单个个体或时间观测点对研究分析所起到的贡献不断下降，研究者更关注整个时间周期内、整个区域内的环境污染变化及其对社会经济的交互影响。最后从准确性维度看，传统的环境统计数据质量的高低完全取决于环境统计制度的完善与否，相比于我国正在不断完善和规范中的环境统计制度而言，新的环境监测系统、卫星遥感数据等能够在统计数据之外，提供相对客观和准确的数据。

最后，与目前大数据在特定商业或专业场景中的应用有所不同，环境数据并非如图像或文本大数据经过简单数据标注后，就能够直接用于复杂的深度学习模型进行训练。若要在社会科学研究领域中充分发挥环境大数据的科研价值，需要上述五个数据特征维度同时发挥作用，即不同类型的环境大数据之间存在着千丝万缕的内在关联。此外，这五个数据特征之间也并非完全互斥，例如数据来源增加、数据频度提升必然会带来数据体量的增加，而不同的数据来源能够互相进行交叉验证或稳健性检验，以提升数据的准确性与可靠性。

根据以上分析，考虑到环境数据之间的内在关联性，本文将社会科学中所应用的环境大数据定义为环境大数据体系。而为了更好地对环境大数据应用情况进行分析，本文依据大数据5V模型将环境大数据体系从不同维度进行分类。首先，从数据粒度（对应数量Volume）和



注：不同类型数据的圆圈跨度表明在不同维度上的常见形式。

图1　社会科学研究中的环境大数据体系

时间频度(对应时效Velocity)两个维度将环境数据划分为不同的数据类别(对应种类Variety);其次,数据价值(Value)和数据准确性(Veracity)则由具体的环境数据应用问题所体现。上述整个环境大数据体系及其分类由图1所示。在环境大数据体系中,环境数据主要分为宏观统计数据、微观数据库(含普查数据)、监测数据、卫星数据以及异构数据五类。其中,宏观统计数据主要由不同地区与行业的加总数据构成,其时间频度和数据粒度均为最低;微观数据库则由企业与个人观测值组成,在数据粒度上得到了提升;监测数据与卫星数据则均有着较高的时间频度和数据粒度;异构数据则由除上述之外的各类新型数据构成,在时间频度和数据粒度上具有更多的可能性。本文对环境大数据应用的分析也将基于上述针对环境大数据的五个分类展开,并分别从以下四个方面进行着重分析:现有研究中采用了哪些环境大数据、数据特征是什么?现有研究利用这些数据研究了什么样的科学问题?现有研究中主要使用了什么样的分析方法来处理环境大数据?现有研究中存在哪些不足与可改进的空间?

三、环境大数据应用现状

根据本文所界定的环境大数据体系和对应的分析框架,接下来将分别从宏观统计数据、微观数据库、监测数据、卫星数据以及异构数据这五种数据对环境大数据的应用现状进行评述。值得注意的是,在目前我国社会科学研究领域,环境污染的研究对象主要集中于空气污染和水污染这两个方面,而土壤污染与固体废弃物污染等其他类别的污染物虽然也是重要的环境污染物,但是受限于数据可得性等原因,涉及到这些污染物的相关环境问题研究还没有在社会科学研究领域中得到重点关注。鉴于此,本文对环境大数据应用的讨论,也将集中在大气污染和水污染这两类环境污染物之上展开。

(一)宏观环境统计数据

我国从1980年起才由国务院环保办与国家统计局联合建立环境保护统计制度,起步相对较晚,且环境统计制度一直处于持续的完善过程中,这使得环境统计数据所能提供的信息相对较为有限(彭立颖、贾金虎,2008)。“十二五”和“十三五”时期,环境统计制度中依然存在着诸如统计执行困难、资金紧缺、调查员素质不高以及企业瞒报等困难,使得环境统计数据质量存在一定的问题(高峰,2014;黄璇,2018)。但是,作为社会科学研究中最重要的数据来源之一,环境统计数据为研究者和决策者从宏观角度初步了解中国的环境演化及其主要驱动因素提供了巨大的帮助。

环境统计数据主要来自于《中国环境年鉴》《中国环境统计年鉴》和《中国环境统计年报》。本文对中文权威期刊中引用了上述三本统计资料数据的论文进行统计,来分析利用环境统计数据进行研究的总体发展趋势。选取的期刊来自于《中国人文社会科学期刊AMI综合

评价报告(2018年)》中被评选为顶级与权威级别的期刊,学科范围则包含管理学、环境科学、经济学、统计学和综合类,共计21本期刊。统计的截止时间为2019年12月,共查询获取符合引用数据要求的文献共计665篇[①]。

统计结果显示,国内针对环境问题的量化研究主要从2005年左右开始,随后进入快速发展期,图2显示,该上升趋势一直持续到近几年才逐渐减缓,2015—2019年的年均发文量已达到81篇。在三本统计资料中,《中国环境年鉴》主要以文本形式记录大事件为主,也同时提供了结构化的统计数据,是早年环境统计数据的主要来源,此后其提供结构化统计数据的功能逐渐被《中国环境统计年鉴》所替代,后者目前已经成为最主要的研究数据来源。此外,《中国环境年报》与《中国环境统计年鉴》类似,主要以提供结构化的环境统计数据为主,两者在数据内容上存在着一定的重复,前者在部分废气、废水等重点研究关注的指标上比后者提供了略微详细的分类数据。总体而言,宏观环境统计数据具有类别最全、细分指标最全的特征,其中具体类别包括大气环境、水环境、土壤环境、海洋环境、固体废弃物、林业环境以及自然灾害等,因此能够满足绝大多数宏观环境研究的需要,而且其另外一个重要的作用在于能够为微观研究提供基本的整体趋势判断。

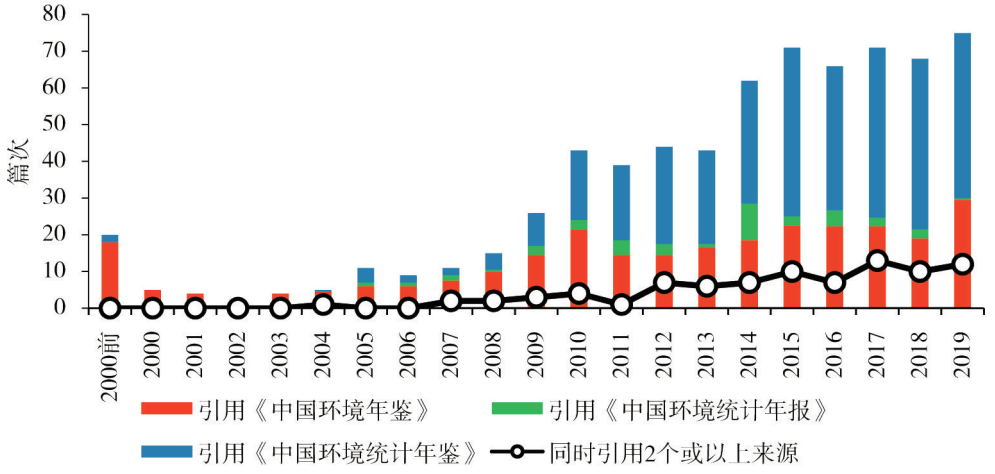


图2　引用三大环境统计资料的论文数量发展趋势

但是与此同时,我们必须注意到,环境统计数据实质上存在着非常明显的缺陷。首先,作为国内最主要的官方环境统计数据来源,三本统计资料中均没有给出既分省又分行业的环境

①本文选取的21本期刊主要包括:《中国社会科学》《经济研究》《管理世界》《经济学(季刊)》《世界经济》《金融研究》《中国工业经济》《财贸经济》《经济学动态》《南开管理评论》《数量经济技术经济研究》《统计研究》《中国人口·资源与环境》《财政研究》《中国农村经济》《经济管理》和《中国软科学》等17本综合性或专业领域期刊,并且根据本文研究的主题,另外选取了高校学报中的四本期刊:《北京大学学报(哲学社会科学版)》《清华大学学报(哲学社会科学版)》《中国人民大学学报》和《北京师范大学学报(社会科学版)》。

数据,使得现有研究仅能够分析不同省的总体环境污染情况,或者仅能够从全国层面对不同行业环境污染情况进行分析[①]。考虑到其他经济社会数据如产出水平、资本投入、劳动投入、能源投入以及研发活动等主要指标的分省分行业层面数据较为容易获得,若要结合社会经济数据深入分析环境问题,分省分行业的环境数据缺失是一个重要阻碍。由于现有研究在这一方面存在着较大的空白,在未来的研究中,对分省分行业的环境数据进行估算将是一项极其重要的工作。此外,地区层面污染物加总数据与行业层面污染物加总数据之间存在一个差值,即两个维度的污染物加总数据是不同的,这点在研究中需要格外注意。产生这个差异的原因在于,在计算各个地区的污染总量时,会将该地区中占比15%的非重点调查企业的数据进行估算后加入,因而导致两个维度的加总数据产生差异(李斌,2010)。

其次,环境统计数据的公开程度在近年呈现大幅度下降趋势。三大统计资料之一的《中国环境统计年报》自2016年起不再出版,因此该统计资料能够提供的环境数据年份仅仅截止到2015年。而作为另一主要数据来源的《中国环境统计年鉴》,其公开的环境统计数据自2016年起出现了大幅简化[②]。以大气污染数据为例,2016年的统计数据仅仅给出了工业废气、二氧化硫、氮氧化物、烟(粉)尘等指标的全国加总排放数据,分省市的主要污染物排放数据则需要查询各省市的《统计年鉴》进行获取。虽然地级市层级的排放数据可以通过查询《中国城市统计年鉴》获取,但该年鉴仅包含我国地级及以上城市,缺少地区、盟和民族自治州等行政区域的数据,且可获取的数据中也不再包含污染物产生量这类指标,也难以对环境污染的末端治理效果进行评价。而有关全国分行业的污染物排放数据,目前也不再有明确的数据获取途径。

总体而言,宏观环境统计数据的一个较明显的问题在于指标层级分类相对较粗且关键统计指标的年际连续性较差,部分关键环境污染物指标质量在"十三五"期间出现了大幅滑坡。如何克服数据上的困难来对我国近年来的环境治理成效进行宏观评价,是当前学界所面临的最大难点。

(二)微观环境数据

微观环境数据指以企业、家庭(个人)或站点为基础的个体数据,虽然这类数据所包含的观测值往往只是全体样本的抽样子集,但是相对于宏观统计数据而言,其数据粒度更细、更易于考察微观个体之间的行为差异以及对政策的响应区别,而且经过科学抽样设计和调研问卷设计所形成的微观数据库在一定程度上能够揭示出全体样本的特征。近年来,微观数据库在社会科学领域的应用不断增加,这些数据库主要分为家庭和个人追踪调查微观数据库(甘犁、冯帅章,2019),以及企业追踪调查微观数据库等(聂辉华等,2012)。目前,最主要的环境微观

①从本文对665篇论文进行统计的结果显示,凡是涉及到省市研究或行业研究,均存在这种现象,即囿于数据限制而无法开展分省市分行业的细致研究。

②由于年鉴的命名规则,这里所指即从《中国环境统计年鉴2017》起。

数据包括由生态环境部(原环境保护部)管理的环境统计数据库以及每十年进行一次的全国污染源普查数据库。

环境统计数据库是我国官方统计资料《中国环境统计年鉴》等的基础数据来源,目前学界通过公开渠道所获取的《中国工业企业环境统计数据库》(下称《工企环统数据库》),是环境统计数据库的工业子数据库,包含了全部工业源的调查样本及部分主要污染物统计指标[①]。《工企环统数据库》中所涵盖的重点调查单位,全部为重点污染物(原则上为当年实行总量控制的污染物)排放量占地区排放总量85%以上的工业企业,由县级环保部门要求其自主填表上报,并进行不定期检查以确保数据质量,目前可获得的数据库涵盖年份为1998—2014年,因而这一数据库也被视作当前最全面、可靠的长时间序列微观环境数据库。该数据库提供了包含企业水资源使用量、废气污染、废水污染、化学需氧量、氨氮、二氧化硫、氮氧化物、烟尘和粉尘等一系列目前学界所重点关注的环境污染物排放与处理指标。

本文将《工企环统数据库》与官方环境统计年鉴的数据进行对比,从表1可以看到,该数据库的企业观测样本数量与官方统计年鉴的数据总体上保持一致,除了1998年、1999年以及2002年的总观测值比官方公布数据分别少了约24.6%、8.0%和3.7%之外,其余年份的观测值总量与官方公布数据的误差范围基本上在3%之内。除了部分省市在个别年份存在样本遗漏外,2005年之后数据误差主要来自于火电企业的口径差异[②]。由于"十一五"期间环境统计制度发生变化,火电企业独立于《工企环统数据库》进行统计,因此2006年至2010年的数据库中并不包含火电企业的污染数据。值得注意的是,"十二五"期间环境统计制度再次发生了变化,最大的改变则是企业的统计口径根据第一次全国污染源普查结果进行了修正,因此"十二五"期间数据库所包含的企业数量大幅上升,2011年的企业数相对于2010年增加了约35.7%。鉴于此数据库是宏观环境统计的基础,因而工业部门的宏观环境统计数据也在2011年发生了巨大改变。此外,"十二五"期间环境统计报表制度中的产排污系数也根据第一次全国污染源普查结果进行了调整,对于2010年前后都存续的企业,其部分污染物指标的产生量和排放量数据会在2011年产生一个明显的系统性变化,这一变化也体现在了宏观环境污染物加总数据中。因此,无论是使用该微观环境数据库进行微观研究还是使用宏观环境数据进行宏观研究时,这是必须注意到的一个问题,以避免对相关数据产生误读。此外,鉴于火力发电对我国环境质量的重要影响,2006—2010年期间火电企业环境污染数据的长时间缺失无疑会对研究产生重大影响,如果未来没有出现更好的数据来源对其进行数据补充,意味着研究者需要利用多源数据对该时间段的缺失数据进行插值填补,而部分省份在个别年份的缺失数据,在实际应用的过程中也同样需要进行类似处理。

①此处的公开渠道获取指的是各科研院校与研究机构可以通过专业数据商进行数据采购而获得数据。

②这里的火电企业指的是火力发电企业,即四位数国民经济行业代码为4411的企业。

表1　　环境微观数据库企业数情况　　(单位:个)

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
北京市	0	1015	1517	1133	1004	4	838	794	730	850	787	808	711	914	896	932	1534
天津市	0	0	1228	1839	1596	1617	1615	1372	1765	1610	1767	1871	1940	1988	2021	2657	3315
河北省	2818	2563	2671	3011	2724	2634	2865	2844	2871	4958	5549	5677	5802	10524	10074	10488	10856
山西省	2891	2769	2759	3395	3482	3375	3189	3116	3304	3817	3918	3661	3942	6464	6367	6258	6509
内蒙古自治区	1163	1134	1022	927	903	962	1192	1266	1282	1823	2211	2137	2259	3187	3025	3110	3724
辽宁省	6	2590	2508	2810	2595	2597	2495	2988	3976	4768	4912	4885	4559	6586	6316	6305	8366
吉林省	1062	1058	1066	989	897	881	844	808	749	845	941	963	967	1489	1466	1483	1540
黑龙江省	108	1839	1770	1544	1596	1619	1506	1414	1314	1417	1500	1537	1480	2083	1927	1884	1914
上海市	2215	2052	2000	1938	1808	1732	1597	1575	1763	1721	1802	1749	1850	2283	2158	2089	2228
江苏省	5557	6235	6369	4898	5068	5604	5518	5800	6081	7674	7896	7905	7887	11291	11107	10743	10731
浙江省	4660	4751	5433	5537	5634	5664	5749	5848	6199	9999	10119	10889	10767	13931	13541	13211	12342
安徽省	1978	1776	1981	1728	1691	1714	1644	1586	1788	2717	2944	3150	4112	8552	8403	8365	8402
福建省	3503	3694	3737	3213	3043	3184	3098	3135	3108	6196	6153	6091	6053	5802	5740	5755	5767
江西省	0	0	1107	942	992	1042	1116	1160	1285	2453	2861	2970	3308	5197	5118	5115	5682
山东省	1754	5446	5425	5246	5107	5186	4994	5038	5014	5838	5569	6330	6052	8008	7791	7708	7911
河南省	2510	3416	3611	4108	2866	3158	4047	3402	3737	4644	4101	4280	4333	6678	6550	6494	6967
湖北省	2302	2046	2187	2222	2236	2589	2283	2312	2194	2388	2505	2518	2613	3911	3699	3590	3912
湖南省	3526	3021	2845	2839	2823	3156	3139	3015	2716	3310	3670	3852	3906	5008	4800	4668	4831
广东省	5507	5451	6500	7037	7068	7017	6724	6376	6702	12948	12974	11785	11939	15907	14890	14959	15145
广西壮族自治区	1716	1769	1885	1886	1687	1734	1719	1730	1785	4333	4795	4631	4430	3565	3517	3532	3464
海南省	287	298	282	294	293	283	295	282	242	282	338	357	327	483	460	458	496
重庆市	1241	1336	1240	1449	1432	1476	1367	1388	1916	2453	2758	2875	2458	3212	3107	3145	3677
四川省	3268	3458	3463	3465	3638	3965	4020	4136	5370	6217	5904	5516	6571	7720	7548	7510	7750
贵州省	2273	2021	2148	2567	2636	2590	2949	2990	2832	2942	3504	3375	3450	4131	3801	3452	3618
云南省	1121	1308	1357	1399	1402	1497	1549	1557	1771	1959	2018	2031	2037	4112	4105	4143	4252
西藏自治区	0	0	0	0	0	0	0	0	0	0	0	0	0	93	86	95	98
陕西省	1825	1848	1800	1803	1874	1892	1832	1760	1683	2797	2702	2603	2580	4070	3760	3650	3763
甘肃省	1276	1190	1198	1115	1077	1082	1058	1027	1062	1162	1297	1350	1443	2473	2290	2341	2532
青海省	243	214	210	170	169	202	239	244	252	342	520	559	566	597	611	603	691
宁夏回族自治区	224	229	212	228	270	270	301	332	363	411	529	619	650	938	884	819	842
新疆维吾尔自治区	821	755	691	455	580	597	675	798	863	1184	2054	2175	2132	1830	1938	2095	1771
未知区域	0	0	0	0	0	0	0	0	1784	0	0	0	0	0	0	0	0
数据库火电企业数	640	848	910	1048	1055	1122	1205	1402	138	74	45	37	31	1828	1820	1853	1908
年鉴火电企业数	na	na	na	1033	1077	1158	1196	1403	1571	1715	1742	1715	1642	1828	1824	1853	1908
数据库企业总数	55855	65282	70222	70187	68191	69323	70457	70093	76501	104058	108598	109149	111124	153027	147996	147657	154630
年鉴企业总数	74097	70978	70944	71377	70797	69665	70462	70514	76185	106457	110373	110905	112799	153027	147996	147657	154633

注:na表示该数据不可得。未知区域表示数据库中无法通过行政区划代码对企业的所在省市进行分类。

目前，学界已经利用《工企环统数据库》对中国的企业环境表现和相关环境政策进行深入分析（Wu et al.，2017；Wang et al.，2018；Zhang et al.，2018）。但是，根据前文所述，《工企环统数据库》的统计口径与各类官方统计年鉴中的经济数据统计口径存在区别，例如统计年鉴中对工业的统计主要以是否是规模以上企业进行划分，而非按照污染物排放规模进行划分，因此该数据库并不适合对数据加总后与其他宏观经济数据进行合并分析。因此，《工企环统数据库》最大的一个优势在于可以与经济学研究中应用最广泛的微观数据库《中国工业企业数据库》进行样本匹配，从而结合企业的经济财务数据将研究问题进行外延拓展（Liu et al.，2017；陈钊、陈乔伊，2019；王班班等，2020）。本文通过将两个数据库中的企业名称与组织机构代码进行匹配，发现1998年至2013年的总体匹配率约为46%，即《工企环统数据库》中将近一半的观测值能够通过数据匹配对其经济财务数据进行补充，每一年的匹配数量与匹配率则由图3所示。相信未来随着学界对《工企环统数据库》不断深入研究，能够从中挖掘出更多的知识与内涵。

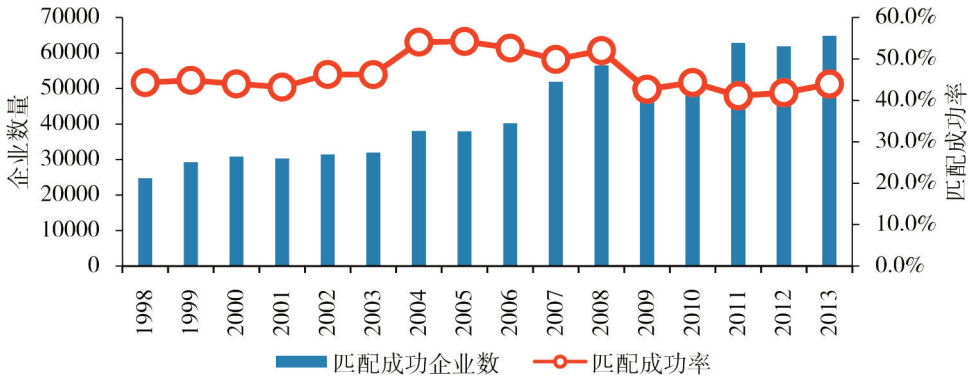


图3　微观数据库匹配结果

相比于《工企环统数据库》，全国污染源普查数据库更加能够反映全国总体环境污染物排放的全貌，能够对环境统计的数据进行全面的偏误校准（何能勇、姜平，2010；李斌，2010）。第一次全国污染源普查结果显示，截止到2007年12月31日，全国普查对象总计592.6万个，其中工业源对象达到了157.6万个[①]，远远超过了2007年环境统计中的重点调查单位数（约10.6万个），由此可以看到，污染物普查能够为我国整体环境污染情况给出最为全面的描述。虽然如此，环境统计数据能够凭借具有时间序列数据的优势，为研究者提供微观企业污染行为的时间动态变化，为决策者更好地理解环境政策的微观影响机制提供实证依据。

如上所述，污染物普查数据最大的一个局限在于其时间维度上的信息缺失较多，我国于

①国家统计局：《第一次全国污染源普查公报》，2010年2月6日，http://www.stats.gov.cn/tjsj/tjgb/qttjgb/qgqt-tjgb/201002/t20100211_30641.html。

2020年完成了第二次全国污染源普查工作,距离第一次普查工作时隔十年之后再次向全社会通报了最新的全国环境污染全貌。根据《第二次全国污染源普查公报》,截至2017年12月31日,全国普查对象总计358.32万个,比第一次普查有所下降,但是工业源对象上升至247.74万个,十年期间增加了约90万个。从评估环境政策效果的角度看,2007—2017年间我国出台了包含"大气十条""水十条"和"土十条"等一系列环境政策①,因此两次污染源普查之间存在的十年跨度使得很难从普查数据上去甄别不同环境政策的具体贡献,故目前基于全国污染源普查数据的工作仍然集中在对高空间分辨率排放清单的估算上(谈佳妮等,2014;叶贤满等,2015)。因此,借助多源数据以及机器学习模型来对两次污染物普查年份之间的环境污染数据进行插补,将为理解我国这十年间各项环境政策和各项环境治理的效果提供巨大的帮助。

(三)环境监测数据

环境监测技术和通讯技术的发展,促使环境监测数据对传统环境统计数据起到了极大的补充作用。环境监测数据的特征主要有三个方向:第一,环境监测可以对企业污染物排放口进行直接监测,也可以在关键的空间位置如人群密集区、河流交汇处等对环境进行直接监测;第二,与环境统计记录污染物排放总量不同,环境监测直接监测污染物的实时排放浓度,经过一定的计算可以换算出污染物排放总量;第三,环境监测在技术条件允许的情况下(如满足一定的传感器敏感性、数据存储与网络传输条件),可以以较高的时间频率采集环境污染数据,经过进一步加工处理后形成可供下一步分析的数据产品。目前我国已经初步建成了一个包括空气环境监测、水环境监测、噪音监测和土壤监测的全国环境监测网络,以此为基础产生了海量的环境监测数据②。由于数据公开可得性的问题,环境空气国控监测站点数据和地表水水质国控监测站点数据是目前学界使用最多的两类监测数据来源,图4显示了我国2007—2017年间这两类国控监测站点数量及其占全部监测站点比重的变化趋势③。在环境空气监测数据方面,我国生态环境部(原环境保护部)于2012年颁布了新的《环境空气质量标准》,从2012年开始分四个步骤逐渐从重点区域扩展到全国来实施新标准④。在空气质量新标准的监测实施方案中,规定了向社会公开六种污染物指标(SO_2、NO_2、PM_{10}、$PM_{2.5}$、O_3、CO)的实时小时

①"大气十条"指《大气污染防治行动计划》,"水十条"指《水污染防治行动计划》,"土十条"指《土壤污染防治行动计划》。

②《"十一五"期间产生监测数据1亿多个,出具监测报告8000多份》,《中国环境报》2011年8月18日。

③国控监测站点数量和全部监测站点数量均来自于历年《中国环境年鉴》,由于《中国环境年鉴2019》(2018年数据)不再公布具体的监测站点数量,因此图中数据截至2017年。但是根据《中国环境年鉴2019》,2018年大气环境国控监测站点数量依然为1436个,与2017年持平。而根据《关于印发<"十三五"国家地表水环境质量监测网设置方案>的通知》(环监测〔2016〕30号)推断,地表水水质国控监测站点数量在"十三五"期间会维持在2767个。

④环境保护部:《关于实施<环境空气质量标准>(GB3095-2012)的通知》,环发〔2012〕11号,http://www.mee.gov.cn/gkml/hbb/bwj/201203/t20120302_224147.htm。

浓度值和日均浓度值以及AQI指数，2015年起全国全部338个地级以上城市的1436个国控监测站点均公布上述数据。利用国控监测站点数据，研究者们对大气环境政策进行了更加细致的评估（罗知、李浩然，2018），也结合其他社会经济数据研究了空气污染对劳动率、精神状态和支付意愿的影响（Zhang et al.，2017；Chang et al.，2019；Ito & Zhang，2020）。此外，基于国控站点数据的高空间颗粒度与高时间频度特性，不少研究也将其运用于对交通限行、轨道交通建设等政策效果的评估中（曹静等，2014；Viard & Fu，2015；梁若冰、席鹏辉，2016），也进一步研究了空气污染对公众健康的影响（Zhong et al.，2017）。

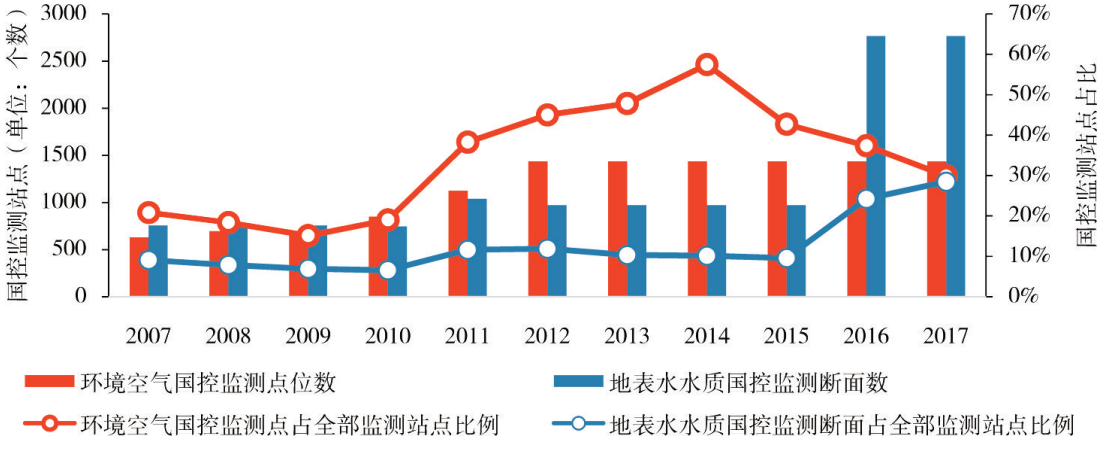
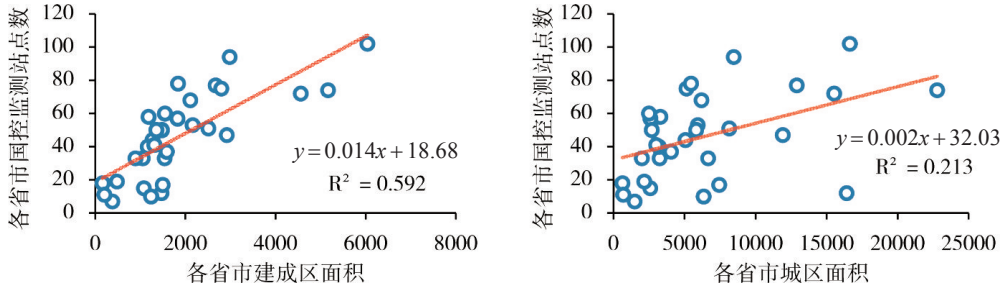


图4　我国环境空气与地表水水质国控监测站点信息

总之，环境空气国控监测站点因其数据公开度高、完整性好、可获取性强，且对各地区的空气污染情况具有较好的样本代表性（见图5），已经在当前的研究中得到了广泛的应用。以2020年暴发的全球COVID-19疫情为例，我国的大气监测数据为分析疫情防控效果以及深度理解我国大气污染物成因等提供了实时且完整的数据支撑（He et al.，2020；Huang et al.，2020）。



注：数据来源于《中国环境年鉴2018》《中国统计年鉴2019》；单位为个，平方公里。

图5　各省市建成区面积（左）、城区面积（右）与国控监测站点数量关系

在地表水水质监测数据方面，虽然全国已经完成了2767个国控监测断面的建设，但是由

于中国环境监测总站的国家地表水水质自动监测实时数据发布系统尚处于调试阶段[①],因而目前只能通过该系统查询到少数断面的实时数据,这些数据每四小时更新一次。与此同时,各省市的生态环境厅(局)网站也没有全部公开地表水水质实时监测数据平台,因此目前学界还没有办法从公开渠道获取全部地表水水质国控监测断面的实时监测数据。若把时间频率降低到周度或月度,生态环境部数据中心则提供了截至2018年底的全国主要流域重点断面水质自动监测周报,其中涵盖了148个重点断面的污染物浓度监测数据,包含PH值、溶解氧、五日生化需氧量、氨氮等10种主要监测指标[②]。此外,各省市的生态环境厅(局)网站也提供了地表水水质月报报告,但由于报告格式不统一,且大多数报告仅报告断面的水质分类结果而非监测浓度数据,因而在实际研究应用时也存在着较大的难度。目前断面样本覆盖最全的数据来自于《中国环境年鉴》,历年的年鉴中报告了2004年至2010年各水系国控监测断面的年均污染浓度数据,虽然数据时间序列较短且为年度均值,但是通过对这些国控监测断面进行地理信息编码,能够得到不同断面在空间上的上下游位置关系,因而可以利用这些数据开展有关水环境边界污染效应的研究(Kahn et al.,2015;李静等,2015;沈坤荣、周力,2020)。结合这些数据,也有研究者对分权治理成效、河长制成效、产业政策效果以及企业生产率等问题进行了研究(王兵、聂欣,2016;蔡嘉瑶、张建华,2018;沈坤荣、金刚,2018;金刚、沈坤荣,2019)。

除了环境监测数据外,还有一类监测数据同样重要,那就是企业排污口的直接污染物排放监测数据。我国最早在《关于加强和改进环境统计工作的意见》(环发〔2005〕100号)中提出要建设重点污染源的自动在线监测系统,并于2007年正式制定了国家重点监控企业名单,要求企业安装自动监控系统并与各级环保部门监控系统联网,并由监测部门至少每月对国家重点监控企业进行一次监督性监测,与自动监测数据比对。该措施在保持原有环境地方分权自治的基础上,通过在线监测系统加强了中央直接监管的力度,极大地提升了企业环境统计数据的质量(Zhang et al.,2018)。截至2012年,环境保护部调查了全国13352家重点监控企业,其中10285家企业开展了自行监测(李莉娜等,2014)。

起初这些企业的自行监测数据并没有办法为社会公众所知晓,但是2014年1月1日起执行的《国家重点监控企业自行监测及信息公开办法》(环发〔2013〕81号)则规定了企业自行监测数据必须向社会公众公开,由此使得学界也能够通过公开渠道获取这些企业污染排放监测数据。更进一步,我国于2016年正式全面实行污染物排放许可证措施,《控制污染物排放许可制实施方案》(国办发〔2016〕81号)中明确规定了获取排污许可证的条件之一就是要企业开

①可以通过中国环境监测总站(http://www.cnemc.cn/)"实时数据"功能中的"水质自动监测实时数据"模块进入查询。

②生态环境部数据中心的在线访问地址为:http://datacenter.mee.gov.cn/。

展自行监测，并且要编制排污许可证执行报告。至此，学界可以通过各省市的公开企业自行监测平台获取企业的高频度污染排放数据①，也能够通过生态环境部的全国排污许可证管理信息平台中各企业的排污许可证执行报告来获取企业的月度排污总量数据。目前基于企业监测数据，研究者们开始研究这种新形式的环境执法监督对企业环境绩效与创新等方面的影响（沈洪涛、周艳坤，2017）。鉴于火电行业污染排放是空气质量的重要影响因素，在线监测数据对于研究火电行业的环境表现和环境政策效果尤为重要，可以用来重点考察如超低排放限值、环保电价、排污费差异以及可再生能源上网等政策的具体实效等问题（Karplus et al.，2018；马北玲等，2019；Tang et al.，2019）。

通过以上分析可以发现，现阶段环境监测数据还远未发挥出其巨大的数据价值，一方面是因为我国的整体环境监测网络依然处于建设的过程中，不少环境监测领域尚未实现监测全覆盖，另一方面则是因为现有的监测设施没有全部开启，或者即使进行了监测活动但相应的监测数据也尚未完全得到公开。此外，监测数据本身可能也存在一定的作假因素（Ghanem & Zhang，2014），增加了学者在使用和分析数据时的障碍。但是随着国家不断加大对环境监测的资金、人力与制度保障的投入，如逐步公开环境监测数据、上收国控站点监事权、颁布《环境监测数据弄虚作假行为处理办法》（环发〔2015〕175号）等一系列措施，环境监测数据未来将在社会科学研究的环境大数据体系中扮演愈来愈重要的角色。

值得关注的是，上述所有监测数据基本都是固定点（监测站、监测断面、企业排污口）监测数据，随着近年来技术的快速发展，移动监测技术在国内外正逐步得以测试应用。移动监测技术通过在城市运营车辆以及公共市政车辆上安装空气质量监测设备，利用其轨迹随机、深入居民区与园区的特性，对城市范围内固定监测点所无法覆盖的空间进行高时空分辨率的监测，并且监测所得数据与固定监测点数据高度一致，其可靠性较好（Wu et al.，2020）。我国部分城市已经开展类似移动监测活动，社会公众已经能够通过各类移动应用软件查询实时监测结果（司书春等，2020）。这类数据目前主要被用于对高时空分辨率空气污染排放数据进行估算（Messier et al.，2018），对于社会科学研究中所关心的诸如环境政策评估、人群健康效应以及经济社会影响等问题所展开的研究较少。随着覆盖区域的不断增加以及数据可获得性的不断提升，移动监测数据在将来拥有较大的应用前景。

（四）卫星数据

在近十年的时间里，社会科学研究中使用卫星数据的情况变得逐渐普遍，最耳熟能详的例子便是使用夜间灯光数据来对经济活动进行分析（Henderson et al.，2012；徐康宁等，

①目前也有一些第三方机构会对官方公开的企业环境数据进行汇总，例如公众环境研究中心（IPE）在其网站（http://www.ipe.org.cn/）汇聚了海量全国各级环保部门官方发布的数据，它在2014年6月推出的可以查看重点污染源实时排放信息的应用“蔚蓝地图”为公众所知晓。

2015)。环境数据也不例外,由于卫星是在高空利用其所搭载的传感器对地面进行监测,且能够在较短时间内对整个地球表面进行扫描,因而卫星数据能够比地面监测数据提供覆盖范围更广泛的环境数据。目前最成熟的环境类卫星数据产品主要以监测大气气溶胶为主,如TERRA卫星、AQUA卫星和风云三号卫星等,也有部分卫星对除大气气溶胶之外的包括O_3、SO_2以及NO_2等在内的痕量气体进行监测,如METOP卫星、TOMS卫星和AURA卫星等,以及对绿地、水体等地表环境相关数据进行监测,如环境一号卫星、高分五号卫星等。国际上利用卫星来对全球大气进行监测的科学研究活动已经开展了40余年,我国在相关领域的大气监测活动也已开展了十余年,而在《生态环境监测网络建设方案》(国办发〔2015〕56号)中,更是明确提出要建立天地一体化的生态遥感监测系统。2017年我国生态环境部试点"千里眼计划",对重点区域实行网格化监管①,及时发现环境问题并进行针对性的解决,试点取得较好的成效,于2018年全面展开"千里眼计划"并逐步扩大实施范围②。

在社会科学研究中使用卫星数据的最大障碍在于存在大气科学知识以及反演算法等专业壁垒,因而社会科学研究者几乎不会使用原始的卫星数据作为研究所需的环境指标数据。以$PM_{2.5}$数据为例,由于环境卫星并不直接监测$PM_{2.5}$浓度,通常是将卫星遥感监测得到的气溶胶光学厚度(Aerosol optical depth,AOD)数据,结合边界层高度和相对湿度等因素,将其与地面监测$PM_{2.5}$浓度数据相结合来构建基于统计回归或机器学习的反演模型,最终得到高时空分辨率的$PM_{2.5}$反演数据(吴健生、王茜,2017)。SO_2和NO_2等痕量气体的高时空分辨率浓度数据也是通过类似的反演算法计算得到(Zhan et al.,2018;Zhang et al.,2018)。因此,为了保证数据的可用性、可比性和可靠性,社会科学研究中通常会使用经过权威机构专业处理的二次加工数据来开展进一步的研究(Donaldson & Storeygard,2016)。

从具体的数据应用角度而言,哥伦比亚大学社会经济数据和应用中心所提供的卫星遥感二次加工数据是目前被学界所广泛采用的卫星数据产品之一③。在该中心提供的"卫星遥感环境指标"数据库中,共包括了全球$PM_{2.5}$网格浓度、全球NO_2网格浓度、全球火点排放和全球热岛指数等8个与环境相关的数据集,其中三个长时间序列数据集的具体情况见表2。

①网格化监管,即对京津冀及周边地区"2+26"城市(以下简称"2+26"城市)全行政区域按照3千米×3千米划分网格,利用卫星遥感技术,筛选出$PM_{2.5}$年均浓度较高的3600个网格作为热点网格,进行重点监管。

②生态环境部:《生态环境部启动"千里眼计划"全面开展热点网格监管工作》,http://www.mee.gov.cn/gkml/sthjbgw/qt/201808/t20180826_454253.htm,2018年8月26日。

③哥伦比亚大学社会经济数据和应用中心网址为:https://sedac.ciesin.columbia.edu/,数据免费向社会公众开放。

表2　　卫星遥感环境指标数据库主要数据集情况

数据集名称	主要环境指标	空间分辨率	数据时间跨度
全球$PM_{2.5}$网格数据	$PM_{2.5}$年度平均浓度	0.01度×0.01度（约1公里×1公里）	1998—2016年
全球NO_2网格数据	NO_2年度平均浓度	0.1度×0.1度（约10公里×10公里）	1996—2012年
全球火点排放指数	火烧面积、碳排放量	0.25度×0.25度（约25公里×25公里）	1997—2015年

在该卫星遥感环境指标数据库中，全球$PM_{2.5}$网格数据在针对我国的空气污染问题中最受关注。如前文所述，我国大气环境监测网络于2012年起才逐步具有监测$PM_{2.5}$浓度的能力，对于2012年前很长一段时间均没有相关的数据记录，全球$PM_{2.5}$网格数据提供了自1998年起的高空间分辨率$PM_{2.5}$浓度数据，极大填补了我国在这块领域的数据空白。在社会研究层面，该数据最大的优势在于可以通过利用ArcGIS等地理信息系统软件对全球$PM_{2.5}$网格数据的栅格数据进行不同区域层级（省级、地市州级、区县市级或乡镇级）的加总，并与现有的社会经济统计数据相结合构建面板数据。该数据目前已经被我国学者广泛应用于空气污染与社会经济活动相关的研究中（邵帅等，2016；严雅雪、李锴，2016；黄寿峰，2017），也被用来研究空气污染对公众健康影响的问题中（Feng et al.，2019）。除了由哥伦比亚大学社会经济数据和应用中心提供的$PM_{2.5}$浓度数据外，现有研究中也会采用其他研究团队估算或由作者自行估算的$PM_{2.5}$浓度数据对我国空气污染问题展开研究（He et al.，2016；陈诗一、陈登科，2018；Zhao et al.，2018）。

相比于$PM_{2.5}$需要进行复杂的反演计算，NO_2数据可获得性相对较高，但是卫星遥感环境指标数据库所覆盖的NO_2数据时间范围较早，2012年之后的浓度数据暂时无法获得，因此不少研究会直接使用现成的NO_2柱总量数据产品展开研究（Cui et al.，2019）①。NO_2相关的卫星数据产品在COVID-19疫情期间发挥了巨大的作用，尤其是在分析全球疫情应对下的空气质量时，提供了高质量的国际可比数据（Le et al.，2020；Liu et al.，2020）。

除了研究卫星提供的各类浓度数据，另一类较为独特的卫星数据——卫星监测火点数据，在近年来引起了研究者们的注意。除了自然火灾外，由人为活动所带来的农业秸秆焚烧等行为不仅会带来意外火灾、影响交通和破坏生态环境，更会导致严重的空气污染的，进而损害公众健康（冒海燕，2014）。卫星遥感环境指标数据库中的火点排放指数实际来自于第四版全球火点排放数据库（Global Fire Emissions Database v4，GFEDV4），除此之外，中国科学院遥感与数字地球研究所的SatSee-Fire卫星看火项目提供了火点可视化服务，并且提供了可供下载的基于多种卫星遥感的火点数据产品②。目前，卫星火点数据正成为一个重要的数据源，被

①柱总量即卫星进行扫描探测时，探测到的垂直大气柱空间内的污染物总量。

②项目主页地址为：http://satsee.radi.ac.cn:8080/index.html。

应用于对我国空气污染问题和公众健康问题的研究(罗知、李浩然,2018;He et al.,2020a)。

除了上述卫星数据之外,SO_2柱总量与地表浓度数据也是经常被学者用来研究我国大气污染问题的卫星数据之一(Karplus et al.,2018;Le et al.,2020),因其与NO_2数据较为类似,本文不再进行过多的展开论述。结合前文所述,可以看到卫星遥感数据在监测与研究我国环境污染问题时具有非常大的潜力,但是由于跨学科专业壁垒的存在,在社会科学研究中进一步推广使用卫星数据存在的主要困难,依然在于数据的可获得性与可用性。目前社会科学研究中采用的卫星数据主要来自于国外权威机构所提供的数据产品,随着我国卫星技术提升以及天地一体化生态遥感监测系统不断完善,希望我国相关部门与科研机构能够在保证数据安全的前提下向公众开放更多自主开发的成熟可靠的卫星数据产品,以更好地提升卫星环境数据在交叉学科中的应用。而从研究的科学问题角度来看,卫星数据监测与收集的是全球范围的数据,数据具有覆盖完整、标准统一等特性,因此其非常适合用来对全球环境问题进行研究,我国的研究者在未来应该更多利用这类数据研究全球环境问题,在提升自身的环境治理能力的同时,增加我国在全球环境研究领域的话语权。

(五)异构数据

异构环境数据是目前存在形式最多样、给研究者想象空间最大的一类环境数据。它包含文本数据和位置数据等不同的形式,该数据最大的特点是大多数以非结构化的形式存在于生活的各方面,在面对特定的环境问题研究需求时,都有可能经过一定的数据处理后被用于研究。随着大数据概念与相关分析技术在各个领域的渗透,研究者们一方面试图利用新颖的数据来估算更精确的污染物排放清单,另一方面也在不断挖掘和探索不同数据所蕴含的价值,以帮助我们更好地理解环境治理效果、公众健康影响以及公众行为响应等问题。

在构建污染物排放清单方面,全国污染源普查与环境统计数据可以用来构建相对全面完整的固定源污染物排放清单,而车载监测设备可以获取细分区域内空气中污染物的浓度数据。但是对于诸如小汽车、卡车、轮船与飞机等移动源交通工具,它们本身所实时排放的污染物数据会因交通工具型号、移动速度以及燃料品种等因素的不同而存在极大的异质性。针对这一情况,目前北斗导航卫星、GPS导航卫星等全球导航系统提供的位置数据,为采用"自下而上"方法来估算高时空分辨率的交通部门污染排放提供了可能。例如,利用全球导航系统获取汽车或卡车的运行轨迹与速度,可以估算道路上汽车运行的高时空分辨率污染物排放数据(Deng et al.,2020)。在船舶污染物方面,目前使用最广泛的是利用船舶自动识别系统(Automatic Identification System, AIS)所提供的各类船舶高频位置数据,结合每艘船舶的船型、发动机和燃料等信息来对SO_2、NO_x甚至是CO_2等污染物排放量进行估算(Smith et al.,2014)①,目

①AIS数据是国际海事组织(International Maritime Organization, IMO)定期出版的温室气体报告中,使用自下而上方法对船舶污染物排放进行估算的最主要数据基础。

前有大量研究利用AIS数据估算了我国沿海或内陆航线船舶污染物排放，并对航运部门相关污染物控制政策进行分析（Liu et al.，2016；Zhang et al.，2019）。

大数据在交通部门的应用绝不仅限于完善污染物排放清单的估算，通过对各类新型商业模式和行为习惯变化进行分析，能够评估其所带来的环境效益。例如研究者对共享单车和滴滴出行等运营数据进行分析，发现近几年逐渐流行的共享经济能够带来可观的环境效益（Yu et al.，2017；Zhang & Mi，2018；许宪春等，2019；Chen et al.，2020），为决策者在制定与这类新型经济现象相关的政策提供有力的实证证据。

除了对交通部门分析时应用地理位置大数据，文本数据是另一类已经被广泛应用于社会科学研究的大数据类型（沈艳等，2019；姚加权等，2020），其中与环境关联性较高的有环境政策文本、环境行政处罚信息、环境影响评价报告等。政策文本分析目前主要采用词频分析与文本编码的方法来分析环境治理体系的变迁（吴芸、赵新峰，2018；许阳，2018），这一类分析以趋势分析为主，鲜见使用统计方法进行量化研究的例子。环境行政处罚是反映企业环境绩效以及环境监管力度的重要风向标，监管部门对每个企业开出的处罚书中，内容虽然按照结构化数据形式展列，但是其中具体的内容，如处罚事由、处罚依据等都需要另行处理，该数据对微观企业环境数据起到了较好的补充作用，也填补了政府监管与企业响应之间联动的数据空白（沈坤荣等，2017；宋平凡、祁毓，2017）。环境影响评价报告是另一个能够充分反映企业项目建设中所涉及的环境污染问题的数据，根据《建设项目环境影响评价政府信息公开指南（试行）》（环办〔2013〕103号）的要求，我国自2014年1月1日起正式要求环保部门需要主动公开环境影响评价报告书、表的全本内容，应该说环境影响评价报告为公众和科研人员了解企业项目建设的环境影响提供了非常丰富的信息。但是由于报告公示分散在各个地区环保部门的网站，且报告本身的内容是高度非格式化的文字、表格与图片，因而需要花费大量精力进行数据收集与格式化处理，目前这方面的数据还没有被学界所充分利用。

公众参与是当今环境治理体系中非常重要的一个环节，除了政府部门制定环境政策并参与监管外，新闻媒体、社交网络等不同的社会公众参与形式，能够起到唤醒民众环保意识、监督突发环境事件以及提升环境信息透明度等作用。2015年9月1日由环境保护部颁发的《环境保护公众参与办法》（部令第35号）正式实施，在前期《中华人民共和国环境保护法》的内容规定上，对公众参与环境保护的原则、方式和义务等做出了明确规定。公众参与环境保护的过程会产生大量的数据，如媒体报道数据、互联网搜索数据以及社交平台文本数据等，这些数据都已经或多或少地被学者用于研究我国环境治理相关的问题。首先，媒体报道的监督对于环境信息的及时公开以及提升公众环保意识起到了重要的作用，2015年纪录片《穹顶之下》的推出，更是激发了公众对雾霾污染问题前所未有的关注度（Tu et al.，

2020),而在一系列研究中,环境相关的关键词互联网搜索指数也被作为公众对环境质量需求与关注的代理指标(李欣等,2017;李子豪,2017),同时媒体报道或环境类非营利性组织(Environmental Non-governmental Organizations, ENGOs)等参与监督也会对企业与城市环境治理产生正面效果(王云等,2017)。而2014年出现的覆盖全国范围的"环境污染随手拍"活动,更是带来了公众直接参与环境监督的热潮,一款名为《环保随手拍》的手机应用程序,能够让实名用户直接通过拍摄、描述等方式对环境污染现象进行上传举报,经过审核后由环境管理执法部门进行处理①。这项活动随后在新浪微博平台大规模展开,公众可以通过直接将举报信息告知各地区官方环境管理部门的形式参与环境监督。相比于使用专用手机应用程序,通过社交网络等渠道的公众参与过程更加透明,数据也更容易被研究者所获取。

除了上述数据之外,目前学者们也采用互联网大数据,从"以人为本"的角度出发研究环境污染所造成的影响。举例而言,以人们在社交媒体上自主发布的文本内容作为数据源,使用自然语言处理(Natural Language Processing,NLP)方法对其进行情绪分析,所得到的结果被用来研究环境污染对人群情绪的影响,进而分析环境污染的健康福利效应(Zheng et al.,2019;Wang et al.,2020)。研究者也采用在线零售数据来研究我国公众与人群个体对环境污染的响应行为,例如口罩销售与空气净化器销售等预防性商品销售情况与空气污染之间的关联性等(Zhang & Mu,2018;Ito & Zhang,2020)。

综上所述,异构数据目前已经逐渐成为研究中数据来源的一部分,甚至核心数据来源。与自然科学研究不同,社会科学研究者们正努力地收集和加工社会经济运行所产生的各类数据,提取其中所蕴含的社会经济内涵,并将其应用于我国环境问题的分析。随着互联网技术、大数据技术等发展,未来这种数据还有更加广阔的空间等待着研究者去挖掘和探索。

四、进一步讨论及未来展望

通过前文对五类环境数据的基本特征及其目前在文献中的应用现状进行分析,可以得到由表3所示的相关特征总结。此外,由前文论述可以看到,在目前的社会科学研究中,环境数据已经不再局限于环境统计数据的范畴,越来越多的数据从微观个体、高时间频率以及高空间分辨率的维度,促使社会科学领域的环境研究进行着内延式和外延式的快速发展。但是为了更好地应用这些环境大数据,本文认为仍需要从提升数据质量、引入新研究方法和加强协同合作这三个方面来进一步地提升和加强。

①生态环境部:《绿侠"环保随手拍App上线 公众监督进入移动互联时代》,http://www.mee.gov.cn/ywdt/hjnews/201406/t20140623_277270.shtml,2014年6月23日。

表3　　不同类型环境数据的特征总结

数据类别	时间频度	数据粒度	数据来源	优势	劣势
宏观环境统计数据	低	低	官方渠道	数据结构统一，指标完整且时间跨度较长，能够进行总体趋势分析。	缺少分省分行业数据，数据公开与连续性质量不断下降。
微观环境数据	低	高	官方渠道数据采购	数据结构统一，时间跨度较长，能够从污染物排放主体视角进行微观行为分析。	指标相对较少，需要与其他经济微观数据库进行匹配；数据的时效性较差。
环境监测数据	普通	高	官方渠道自行收集	数据结构统一，时间频度较高，以浓度数据为主，适用于结合地理信息开展环境监测与控制措施效果分析。	指标较为有限，目前数据集中于空气污染监测；企业自行监测数据涵盖面较为有限，但在不断增加；移动源监测没有统一的数据来源。
卫星数据	较低/普通	高	专业机构自行计算	空间覆盖范围较广，时间跨度较长，主要以浓度数据为主。涵盖无法通过常规手段进行监测的异常环境事件。适用于全球环境研究。	需要对原始卫星数据进行专业处理，专业壁垒较高，需要依赖跨学科团队合作来获取可靠的数据。
异构数据	普通/较高	普通/较高	自行收集	数据形式多样，样本代表性较高，且时效性较好。适用于对新事物、新形势以及对个体行为与政策响应模式进行分析。	数据结构不统一，需要进行结构化处理；由于数据提供方包含各类社会主体，需要对数据准确性进行额外判别。

（一）提升数据质量

为了保证研究结论的可靠性，环境大数据体系的快速发展不能为了“数量”而放弃了“质量”，数据是否可获取、是否有代表性、是否有延续性是广大研究者复现（Replicate）并检验论文结论可靠性和普适性的重要基础，从这个角度看，目前环境大数据体系的整体质量还有非常大的提升空间。

第一，宏观环境统计数据作为体现我国环境污染整体趋势的权威数据，其数据公布的连续性和稳定性应该在官方年鉴中得到基本的保证。当缺乏连续可靠的宏观环境统计数据时，势必极大地限制研究者们开展社会科学研究，减弱了其对于优化提升我国环境管理的支撑作用。对微观环境统计数据而言，虽然获取壁垒较高，但是由于各地监管水平与数据统计的能力存在差异，再加上环境统计体制在不断完善的过程中所遗留的历史问题，数据中实质上存在着不少的“数据噪音”，在使用前必须进行仔细的清洗与整理（王班班等，2020）。

第二，虽然涉及主要环境污染物的卫星数据的公开程度较高，但由于这些数据需要经过前期复杂的专业技术处理，数据质量会因为不同机构或团队的处理方法存在差异而受到影响。除了技术原因导致的数据质量波动，不同数据源代表的数据含义也会存在差异，例如监测数据与卫星数据可能会针对同一个污染物研究对象得到不同的结果（Karplus et al.，2018），

因而在数据可得性条件允许的情况下,尽可能使用不同的数据源(如采用不同卫星的观测数据或以监测数据为辅等)来进行稳健性检验,以降低数据质量波动带来的影响,提升研究结论的可靠性。

第三,异构数据的数据质量在当下则面临更大的挑战。第一个挑战来源于数据采集过程,对于文本数据或互联网数据这类数据,需要研究者自行采集或委托相应的专业技术团队进行采集[①],数据采集的过程中会经常遇到数据来源渠道不稳定、历史数据流失等情况,导致数据的完整性和代表性都受到影响。第二个挑战来自于数据处理与分析过程,目前,在对文本数据和互联网数据进行异常值处理、特征信息抽取以及核心指标生成的过程中,由于缺乏统一的处理标准或是缺乏理论基础而不得不进行简化处理,因而所得到的关键数据与分析结果的可靠性与可复现性较低。为了克服这方面的不足,建议研究者在采用此类数据时,除了对数据质量与可靠性进行详尽的论述之外,可以将相关数据及其详细介绍发表在专业的数据类学术期刊,通过同行评审来进一步提升研究数据的可靠性[②],而且通过公开数据的形式也有利于增加研究的透明度和提升结论的可信度。

(二)引入新研究方法

大数据时代的到来,为社会科学研究带来的最大变化便是促进了"计算社会科学"的兴起(Lazer et al.,2009),这一领域与量化社会科学研究有所不同,更强调针对大数据的特性使用新的方法和范式来开展研究。而从我国社会科学研究的演变历程来看,也仅仅是在近二十年的过程中经历了从定性研究向定量研究的逐步转变,涉及环境相关问题的研究也不例外。由于过往的数据以宏观环境统计数据为主,因此定量研究中也主要采用传统时间序列分析、空间计量分析、效率分析以及构建指标体系等分析方法。即使是利用环境大数据所开展的研究,也主要集中于利用传统定量方法进行实证分析,并没有在方法论方面产生具有影响力的突破。

第一个潜在的提升方向,可以利用复杂网络分析方法来从环境大数据中发现新知识。计算社会科学重点关注的领域之一,是探索社会经济生活中网络特性所蕴含的新知识(Lazer et al.,2009),环境问题无论从自然科学角度还是从社会科学角度,均具有鲜明的复杂网络特性。从自然科学角度来讲,环境污染具有高度的时空关联性,例如空气污染物扩散会受到地形、风速和风力等自然因素的影响,水污染物会因河流流向与流速不同而呈现不同的扩散模

①如社交网络网站的千万级用户数据、电子商务网站的亿级商品数据等,而且采集过程中会遇到各类"反爬虫"的技术封锁,因此在没有专业团队的技术支撑下,数据采集任务是对社会科学研究者的极大考验,很多时候几乎是"不可能完成的任务"。

②代表性的专业数据期刊包括由自然出版集团(NPG)发行的开放获取期刊 *Scientific Data* 等,更多有关专业数据期刊的信息可以参考刘晶晶和顾立平(2015)。

式，因此不同地区之间会随着污染物的动态扩散而具备网络关联特征。而从社会科学角度来看，污染物作为社会经济活动的负产出，会受到不同污染物排放主体之间如产业链关系、供需关系等因素影响，而环境规制又涉及到不同层级主管部门以及同一层级不同主管部门之间的协同，上述现象都体现了环境问题存在着极为复杂的社会网络关系（Bodin，2017）。借助于环境大数据的日渐丰富，虽然研究者们已经认识到环境问题中的复杂网络特性的重要性，并正逐渐开展相应的研究，但是采用的方法依然集中于传统的定量分析方法，对整个环境复杂网络特性探索和知识发现的程度还远远不够。

第二个潜在的提升方向，则是可以更好地利用机器学习方法的精准预测特性，将其运用于社会科学研究。机器学习作为大数据分析方法中的核心方法之一，其本质目标是提升对研究事物与现象的预测能力。以目前在环境政策分析中最为学界所接受的因果效应分析方法为例，环境大数据的出现使得数据本身能够较好地满足因果效应分析对数据颗粒度的要求，如匹配法、双重差分法和断点回归法等对数据的数量规模与颗粒度均有较高要求，利用机器学习则可以提升因果效应分析中的分组匹配效果与反事实结果预测效果，从而估算出更加精准的环境政策效应（Athey & Imbens，2017；Lewbel，2019）。其次，机器学习可以在构建新指标等方面起到重要作用，例如可以利用NLP等技术对未知的文本内容进行情绪分类预测，构建与环境相关的情绪指数（Zheng et al.，2019），也能够从众多环境规制文本信息中提取有效信息，用来测度环境规制程度等。第三，缺失信息插补也是一类预测任务，虽然环境类数据在个体、时间和空间维度的规模不断提升，但是由于技术因素或人为因素，不可避免的会带来信息缺失，使用机器学习技术对关联信息中的数据模式进行训练，进而对缺失信息进行插补，能够帮助研究者们获得质量更高的数据，增加研究结论的可靠性。

（三）加强协同合作

为了在社会科学中更好地应用大数据，不同数据拥有主体之间以及不同专业研究机构之间是否能够形成高效可靠的协同合作是一个至关重要的影响因素。在计算社会科学兴起的近十年中，良性的协同合作并未取得令人乐观的改善，不管是私人公司数据还是政务数据，学界能够与之共享的渠道非常有限，此外，大学和研究机构中的跨学科合作也受到不同的制度约束而进展缓慢（Lazer et al.，2020）。

在环境大数据的研究应用中，上述现象也同样存在。首先，从环境大数据的数据合作角度而言，目前存在着非常巨大的障碍。本文前述的环境大数据中，仅仅是部分数据目前可以通过公开渠道获得，而对于一些诸如微观统计数据以及互联网数据等，需要通过数据采购或项目合作等渠道获取，而本文所提到的这些数据，还仅仅是环境大数据体系中的一小部分，其余数据的获取门槛则更高。例如，目前智能家居设备逐渐开始普及，家用智能空气质量监测及空气净化设备的渗透率逐年提升，若要研究微观个体对环境污染的响应行为或者是支付意

愿等问题,这类智能家居是采集数据与使用数据的最佳选择,但是与拥有这类数据的公司进行合作往往会因为联系渠道、商业秘密或者是个人隐私等问题而难以实现。再如,对于研究环境污染对人体健康的影响这一问题,国家卫生部门的个人健康数据是最佳的选择,但是对绝大多数的研究者而言,该类政务数据是几乎无法获得的。在这种情况下,研究者们往往选择退而求其次,利用城市级或省级的宏观数据进行研究,这样势必会得到不太精准且忽略人群异质特征的结果。因此,不管是公司还是政府部门,应该致力探索不同的数据协作与共享模式。我国目前在这方面已经开始了相关尝试,以国家统计局与清华大学共建的"国家统计局–清华大学数据开发中心"为例,其为国内学校和科研机构提供了使用官方微观数据库进行研究的机会。而以上海和浙江为代表的一批省市,于近年先后建立了政务数据和公共数据的数据开放平台,为全社会有数据需求的个人或机构提供了申请使用数据的机会。学界也应该在呼吁建立更多数据开放与合作平台的同时,好好利用机会,做出更多有价值的研究。

其次,若想要真正实现环境大数据在社会科学研究中得以充分利用,就必须依赖不同专业学科的通力合作。就以数据或分析技术本身而言,想要获得高质量的环境大数据,离不开专业数据团队进行数据采集与清洗,而类似卫星数据则需要具有高度专业知识的团队来进行数据的深加工。此外,仅仅依赖于第三方提供数据也不利于数据的及时更新,有时也会对数据本身缺乏深入的理解。若以研究问题而言,环境问题同样会涉及到不同的专业学科,如污染物的扩散特征、产生原因、二次污染物反应机制等都会对社会科学研究者开展具体的研究产生阻碍,因此这就要求研究团队必须具有一定的环境科学与环境管理知识。例如若要在环境污染研究的基础上,进一步考虑环境污染的人群健康效应,则必须需要专业的公共卫生和医学相关知识来进行支撑。因此,鼓励研究者开展跨学科合作研究,无疑能够更好地发挥环境大数据所带来的数据红利。

五、结论

社会科学研究中更好地利用环境数据开展定量研究,是帮助我国提升环境治理能力中的重要一环。本文基于大数据5V特征对环境大数据体系进行了概念上的界定,构建了针对环境大数据应用的分析框架,并按照时间频度和数据粒度两个维度将环境数据分为五类:宏观环境统计数据、微观环境数据、监测数据、卫星数据以及异构数据。本文发现,宏观环境统计数据提供了最完整的数据指标,且数据质量在我国的环境统计体系下能够得到较高的保证,但是进入"十三五"时期后,部分关键指标的连续性呈现不升反降的趋势,这点对于学术研究而言尤为不利。微观环境统计数据给研究者提供了从微观排放主体视角开展研究的机会,但是其数据获取门槛较高、开放的数据指标也较为有限,且数据时效性较差,较为适合利用历史数据对环境治理机制进行研究。环境监测数据与卫星数据是目前研究应用中较为新颖的环

境数据类型，虽然在时间频度和数据粒度上得到了巨大的提升，但是其提供的环境信息种类非常有限，目前主要集中在对空气污染与水污染的环境问题研究。最后是异构数据，其数据形式与获取途径多样，得益于此，研究者们所开展的研究内容也逐渐丰富，从辅助完善污染物排放清单核算、改进成本收益分析，到公众参与环境治理和公众对环境污染的响应等，异构数据为社会科学研究提供了极佳的数据支撑，也是未来社会科学研究中应用环境大数据的一个重要发展趋势。

结合全文的分析，并针对环境大数据的特征及其应用现状，本文从三个方面进一步讨论在未来研究中更好利用环境大数据的潜在途径。首先，为了得到更加全面可靠的研究结论，必须全方位地提升环境数据质量，强化数据的样本代表性与连续性，并且利用大数据的特性对数据进行交叉检验，同时可以考虑将一部分重要且具有典型性的数据发表于专业数据期刊。其次，针对环境污染问题，可以结合计算社会科学新研究范式中所强调的复杂网络分析方法，对环境问题展开深入分析，此外也可以利用机器学习的预测能力来对环境管理以及环境政策效果进行更加精准的评估。最后，为了增强数据可得性，建议建立适当的数据共享机制来打通政府或企业与学界的合作渠道。此外，大学与研究机构也应该加强跨学科的合作研究来充分挖掘环境大数据带来的红利。综上所述，环境大数据在社会科学研究中已经初露锋芒，并在未来具有巨大的潜力和发展空间，能够在我国构建现代化环境治理体系的进程中大有可为。

参考文献：

[1] 蔡嘉瑶，张建华. 财政分权与环境治理——基于“省直管县”财政改革的准自然实验研究[J]. 经济学动态，2018，(1)：53–68.

[2] 曹静，王鑫，钟笑寒. 限行政策是否改善了北京市的空气质量?[J]. 经济学(季刊)，2014，(3)：1091–1126.

[3] 陈诗一，陈登科. 雾霾污染、政府治理与经济高质量发展[J]. 经济研究，2018，(2)：20–34.

[4] 陈钊，陈乔伊. 中国企业能源利用效率：异质性、影响因素及政策含义[J]. 中国工业经济，2019，(12)：78–95.

[5] 甘犁，冯帅章. 以微观数据库建设助推中国经济学发展——第二届微观经济数据与经济学理论创新论坛综述[J]. 经济研究，2019，(4)：204–208.

[6] 高峰. “十二五”环境统计工作中存在的问题及建议[J]. 环境保护与循环经济，2014，(9)：69–71.

[7] 何能勇，姜平. 环境统计与污染源普查数据比对研究——以贵州省COD排放总量为例[J]. 贵州师范大学学报(自然科学版)，2010，(2)：44–48.

[8] 黄寿峰. 财政分权对中国雾霾影响的研究[J]. 世界经济，2017，(2)：127–152.

[9] 黄璇. “十三五”环境统计工作存在的问题及建议[J]. 环境与发展，2018，(7)：213–214.

[10] 金刚，沈坤荣. 地方官员晋升激励与河长制演进：基于官员年龄的视角[J]. 财贸经济，2019，(4)：20–34.

[11] 李斌. 辽阳市污染源普查与环境统计数据对比分析[J]. 科技信息，2010，(7)：357–358.

[12] 李静,杨娜,陶璐. 跨境河流污染的“边界效应”与减排政策效果研究——基于重点断面水质监测周数据的检验[J]. 中国工业经济,2015,(3):31–43.

[13] 李莉娜,唐桂刚,万婷婷,陈敏敏,景立新. 我国企业排污状况自行监测的现状、问题及对策[J]. 环境工程,2014,(5):86–89+94.

[14] 李欣,杨朝远,曹建华. 网络舆论有助于缓解雾霾污染吗?——兼论雾霾污染的空间溢出效应[J]. 经济学动态,2017,(6):45–57.

[15] 李子豪. 公众参与对地方政府环境治理的影响——2003–2013年省际数据的实证分析[J]. 中国行政管理,2017,(8):102–108.

[16] 梁若冰,席鹏辉. 轨道交通对空气污染的异质性影响——基于RDID方法的经验研究[J]. 中国工业经济,2016,(3):83–98.

[17] 刘晶晶,顾立平. 数据期刊的政策调研与分析——以Scientific Data为例[J]. 中国科技期刊研究,2015,(4):331–339.

[18] 罗知,李浩然. “大气十条”政策的实施对空气质量的影响[J]. 中国工业经济,2018,(9):136–154.

[19] 马北玲,吕欣,陈星,陈晓红. 火电厂大气排放监测大数据分析及政策影响研究[J]. 中国人口·资源与环境,2019,(7):73–79.

[20] 冒海燕,黄润州,杨蕊,周定国. 秸秆焚烧:秋收季节“狼烟”又起[J]. 生态经济,2014,(11):6–9.

[21] 米加宁,章昌平,李大宇,林涛. 第四研究范式:大数据驱动的社会科学研究转型[J]. 学海,2018(2):11–27.

[22] 聂辉华,江艇,杨汝岱. 中国工业企业数据库的使用现状和潜在问题[J]. 世界经济,2012,(5):142–158.

[23] 彭立颖,贾金虎. 中国环境统计历史与展望[J]. 环境保护,2008,(4):52–55.

[24] 邵帅,李欣,曹建华,杨莉莉. 中国雾霾污染治理的经济政策选择——基于空间溢出效应的视角[J]. 经济研究,2016,(9):73–88.

[25] 沈洪涛,周艳坤. 环境执法监督与企业环境绩效:来自环保约谈的准自然实验证据[J]. 南开管理评论,2017,(6):73–82.

[26] 沈坤荣,金刚. 中国地方政府环境治理的政策效应——基于“河长制”演进的研究[J]. 中国社会科学,2018,(5):92–115+206.

[27] 沈坤荣,金刚,方娴. 环境规制引起了污染就近转移吗?[J]. 经济研究,2017,(5):44–59.

[28] 沈坤荣,周力. 地方政府竞争、垂直型环境规制与污染回流效应[J]. 经济研究,2020,(3):35–49.

[29] 沈艳,陈赟,黄卓. 文本大数据分析在经济学和金融学中的应用:一个文献综述[J]. 经济学(季刊),2019,(4):1153–1186.

[30] 司书春,许宏宇,张子珺,高健,秦孝良. 城市出租车走航大气监测系统研究——以山东省济南市为例[J]. 环境保护,2020,(7):54–57.

[31] 宋平凡,祁毓. 企业捐赠对环境处罚的影响研究——来自工业类上市公司的证据[J]. 环境经济研究,2017,(4):93–106.

[32] 谈佳妮,余琦,马蔚纯,马剑丽,程杰,张艳. 小尺度精细化大气污染源排放清单的建立——以上海宝山区为例[J]. 环境科学学报,2014,(5):1099–1108.

[33] 王班班,莫琼辉,钱浩祺. 地方环境政策创新的扩散模式与实施效果——基于河长制政策扩散的微观实证[J]. 中国工业经济,2020,(8):99–117.

[34] 王兵,聂欣. 产业集聚与环境治理:助力还是阻力——来自开发区设立准自然实验的证据[J]. 中国工业经济,2016,(12):75–89.

[35] 王云,李延喜,马壮,宋金波. 媒体关注、环境规制与企业环保投资[J]. 南开管理评论,2017,(6):

83–94.

[36] 吴健生，王茜. 基于AOD数据反演地面$PM_{2.5}$浓度研究进展[J]. 环境科学与技术，2017，(8)：68–76.

[37] 吴芸，赵新峰. 京津冀区域大气污染治理政策工具变迁研究——基于2004–2017年政策文本数据[J]. 中国行政管理，2018，(10)：78–85.

[38] 徐康宁，陈丰龙，刘修岩. 中国经济增长的真实性：基于全球夜间灯光数据的检验[J]. 经济研究，2015，(9)：17–29+57.

[39] 许宪春，任雪，常子豪. 大数据与绿色发展[J]. 中国工业经济，2019，(4)：5–22.

[40] 许阳. 中国海洋环境治理政策的概览、变迁及演进趋势——基于1982—2015年161项政策文本的实证研究[J]. 中国人口·资源与环境，2018，(1)：165–176.

[41] 严雅雪，李锴. 中国城市化对$PM_{2.5}$浓度影响的门槛效应研究[J]. 环境经济研究，2016，(2)：93–106.

[42] 姚加权，张锟澎，罗平. 金融学文本大数据挖掘方法与研究进展[J]. 经济学动态，2020，(4)：143–158.

[43] 叶贤满，徐昶，洪盛茂，焦荔，沈建东，张天，何曦. 杭州市大气污染物排放清单及特征[J]. 中国环境监测，2015，(2)：5–11.

[44] 张毅，贺桂珍，吕永龙，马艳飞，宋帅. 我国生态环境大数据建设方案实施及其公开效果评估[J]. 生态学报，2019，(4)：1290–1299.

[45] 赵海凤，李仁强，赵芬，刘丽香，赵苗苗，徐明. 生态环境大数据发展现状与趋势[J]. 生态科学，2018，(1)：211–218.

[46] 赵苗苗，赵师成，张丽云，赵芬，邵蕊，刘丽香，赵海凤，徐明. 大数据在生态环境领域的应用进展与展望[J]. 应用生态学报，2017，(5)：1727–1734.

[47] Athey, S. and G. W. Imbens. The State of Applied Econometrics: Causality and Policy Evaluation[J]. The Journal of Economic Perspectives，2017，31(2)：3–32.

[48] Bodin, Ö. Collaborative Environmental Governance: Achieving Collective Action in Social–ecological Systems[J]. Science，2017，357(6352)：eaan1114.

[49] Chang, T. , J. Graff Zivin, T. Gross, and M. Neidell. The Effect of Pollution on Worker Productivity: Evidence from Call Center Workers in China[J]. American Economic Journal: Applied Economics，2019，11(1)：151–172.

[50] Chen, F. , Z. Yin, Y. Ye, and D. Sun. Taxi Hailing Choice Behavior and Economic Benefit Analysis of Emission Reduction based on Multi–mode Travel Big Data[J]. Transport Policy，2020，97：73–84.

[51] Cui, Y. , W. Zhang, H. Bao, C. Wang, W. Cai, J. Yu, and D. G. Streets. Spatiotemporal Dynamics of Nitrogen Dioxide Pollution and Urban Development: Satellite Observations over China, 2005—2016[J]. Resources, Conservation and Recycling，2019，142：59–68.

[52] Deng, F. , Z. Lv, L. Qi, X. Wang, M. Shi, and H. Liu. A Big Data Approach to Improving the Vehicle Emission Inventory in China[J]. Nature Communications，2020，11(1)：2801.

[53] Donaldson, D. and A. Storeygard. The View from Above: Applications of Satellite Data in Economics[J]. The Journal of Economic Perspectives，2016，30(4)：171–198.

[54] Feng, Y. , J. Cheng, J. Shen, and H. Sun. Spatial Effects of Air Pollution on Public Health in China[J]. Environmental and Resource Economics，2019，73(1)：229–250.

[55] Ghanem, D. and J. Zhang. 'Effortless Perfection:' Do Chinese Cities Manipulate Air Pollution Data?[J]. Journal of Environmental Economics and Management，2014，68(2)：203–225.

[56] He, G. , T. Liu, and M. Zhou. Straw Burning, $PM_{2.5}$, and Death: Evidence from China[J]. Journal of Development Economics，2020a，145：102468.

[57] He, G. , Y. Pan, and T. Tanaka. The Short–term Impacts of COVID–19 Lockdown on Urban Air Pollution in China[R]. 2020b.

[58] He, Q. , M. Zhang, and B. Huang. Spatio-temporal Variation and Impact Factors Analysis of Satellite-based Aerosol Optical Depth over China from 2002 to 2015[J]. Atmospheric Environment, 2016, 129: 79-90.

[59] Henderson, J. V. , A. Storeygard, and D. N. Weil. Measuring Economic Growth from Outer Space[J]. American Economic Review, 2012, 102(2): 994-1028.

[60] Huang, X. , A. Ding, J. Gao, B. Zheng, D. Zhou, X. Qi, R. Tang, J. Wang, C. Ren, W. Nie, X. Chi, Z. Xu, L. Chen, Y. Li, F. Che, N. Pang, H. Wang, D. Tong, W. Qin, W. Liu, Q. Fu, B. Liu, F. Chai, S. J. Davis, Q. Zhang, and K. He. Enhanced secondary pollution offset reduction of primary emissions during COVID-19 lockdown in China[R]. 2020.

[61] Ito, K. , and S. Zhang. Willingness to Pay for Clean Air: Evidence from Air Purifier Markets in China[J]. Journal of Political Economy, 2020, 128(5): 1627-1672.

[62] Kahn, M. E. , P. Li, and D. Zhao. Water Pollution Progress at Borders: The Role of Changes in China's Political Promotion Incentives[J]. American Economic Journal: Economic Policy, 2015, 7(4): 223-242.

[63] Karplus, V. J. , S. Zhang, and D. Almond. Quantifying Coal Power Plant Responses to Tighter SO_2 Emissions Standards in China[J]. Proceedings of the National Academy of Sciences, 2018, 115(27): 7004-7009.

[64] Lazer, D. , A. Pentland, L. Adamic, S. Aral, A. L. Barab ά si, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. Computational Social Science[J]. Science, 2009, 323(5915): 721-723.

[65] Lazer, D. M. J. , A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, A. Nelson, M. J. Salganik, M. Strohmaier, A. Vespignani, and C. Wagner. Computational Social Science: Obstacles and Opportunities[J]. Science, 2020, 369(6507): 1060-1062.

[66] Le, T. , Y. Wang, L. Liu, J. Yang, Y. L. Yung, G. Li, and J. H. Seinfeld. Unexpected Air Pollution with Marked Emission Reductions During the COVID-19 Outbreak in China[J]. Science, 2020, 369(6504): 702-706.

[67] Lewbel, A. The Identification Zoo: Meanings of Identification in Econometrics[J]. Journal of Economic Literature, 2019, 57(4): 835-903.

[68] Liu, F. , A. Page, S. A. Strode, Y. Yoshida, S. Choi, B. Zheng, L. N. Lamsal, C. Li, N. A. Krotkov, H. Eskes, R. van der A, P. Veefkind, P. F. Levelt, O. P. Hauser, and J. Joiner. Abrupt Decline in Tropospheric Nitrogen Dioxide over China after the Outbreak of COVID-19[J]. Science Advances, 2020, 6(28): eabc2992.

[69] Liu, H. , M. Fu, X. Jin, Y. Shang, D. Shindell, G. Faluvegi, C. Shindell, C. , and K. He. Health and Climate Impacts of Ocean-going Vessels in East Asia[J]. Nature Climate Change, 2016, 6(11): 1037-1041.

[70] Liu, M. , R. Shadbegian, and B. Zhang. Does Environmental Regulation affect Labor Demand in China? Evidence from the Textile Printing and Dyeing Industry[J]. Journal of Environmental Economics and Management, 2017, 86: 277-294.

[71] Messier, K. P. , S. E. Chambliss, S. Gani, R. Alvarez, M. Brauer, J. J. Choi, S. P. Hamburg, J. Kerckhoffs, B. LaFranchi, M. M. Lunden, J. D. Marshall, C. J. Portier, A. Roy, A. A. Szpiro, R. C. H. Vermeulen, and J. S. Apte. Mapping Air Pollution with Google Street View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression[J]. Environmental Science & Technology, 2018, 52(21): 12563-12572.

[72] Smith, T. W. P. , J. P. Jalkanen, B. A. Anderson, J. J. Corbett, J. Faber, S. Hanayama, E. O'keeffe, S. Parker, L. Johanasson, L. Aldous, and C. Raucci, C. Third IMO GHG Study 2014[R]. 2014.

[73] Tang, L. , J. Qu, Z. Mi, X. Bo, X. Chang, L. D. Anadon, S. Wang, X. Xue, S. Li, X. Wang, and X. Zhao. Substantial Emission Reductions from Chinese Power Plants after the Introduction of Ultra-low Emissions Standards[J]. Nature Energy, 2019, 4(11): 929-938.

[74] Viard, V. B. and S. Fu. The Effect of Beijing's Driving Restrictions on Pollution and Economic Activity[J].

Journal of Public Economics, 2015, 125: 98–115.

[75] Tu, M., B. Zhang, J. Xu, and F. Lu. Mass Media, Information and Demand for Environmental Quality: Evidence from the 'Under the Dome.' [J]. Journal of Development Economics, 2020, 143: 102402.

[76] Wang, C., J. Wu, and B. Zhang. Environmental Regulation, Emissions and Productivity: Evidence from Chinese COD-emitting Manufacturers[J]. Journal of Environmental Economics and Management, 2018, 92: 54–73.

[77] Wang, J., N. Obradovich, and S. Zheng. A 43-Million-Person Investigation into Weather and Expressed Sentiment in a Changing Climate[J]. One Earth, 2020, 2(6): 568–577.

[78] Wu, H., H. Guo, and B. Zhang. Westward Movement of New Polluting Firms in China: Pollution Reduction Mandates and Location Choice[J]. Journal of Comparative Economics, 2017, 45(1): 119–138.

[79] Wu, Y., Y. Wang, L. Wang, G. Song, J. Gao, and L. Yu. Application of a Taxi-based Mobile Atmospheric Monitoring System in Cangzhou, China[J]. Transportation Research Part D: Transport and Environment, 2020, 86: 102449.

[80] Yu, B., Y. Ma, M. Xue, B. Tang, B. Wang, J. Yan, and Y. M. Wei. Environmental Benefits from Ridesharing: A Case of Beijing[J]. Applied Energy, 2017, 191: 141–152.

[81] Zhan, Y., Y. Luo, X. Deng, K. Zhang, M. Zhang, M. L. Grieneisen, and B. Di. Satellite-Based Estimates of Daily NO_2 Exposure in China Using Hybrid Random Forest and Spatiotemporal Kriging Model[J]. Environmental Science & Technology, 2018, 52(7): 4180–4189.

[82] Zhang, B., X. Chen, and H. Guo. Does Central Supervision Enhance Local Environmental Enforcement? Quasi-Experimental Evidence from China[J]. Journal of Public Economics, 2018, 164: 70–90.

[83] Zhang, J., and Q. Mu. Air Pollution and Defensive Expenditures: Evidence from Particulate-filtering Facemasks[J]. Journal of Environmental Economics and Management, 2018, 92: 517–536.

[84] Zhang, X., X. Zhang and X. Chen. Happiness in the Air: How Does a Dirty Sky Affect Mental Health and Subjective Well-being?[J]. Journal of Environmental Economics and Management, 2017, 85: 81–94.

[85] Zhang, X., Y. Zhang, Y. Liu, J. Zhao, Y. Zhou, X. Wang, X. Yang, Z. Zou, C. Zhang, Q. Fu, J. Xu, W. Gao, N. Li, and J. Chen. Changes in the SO_2 Level and $PM_{2.5}$ Components in Shanghai Driven by Implementing the Ship Emission Control Policy[J]. Environmental Science & Technology, 2019, 53(19): 11580–11587.

[86] Zhang, Y. and Z. Mi. Environmental Benefits of Bike Sharing: A Big Data-based Analysis[J]. Applied Energy, 2018, 220: 296–301.

[87] Zhang, Z., J. Wang, J. E. Hart, F. Laden, C. Zhao, T. Li, P. Zheng, D. Li, Z. Ye, and K. Chen. National Scale Spatiotemporal Land-use Regression Model for $PM_{2.5}$, PM_{10} and NO_2 Concentration in China[J]. Atmospheric Environment, 2018, 192: 48–54.

[88] Zhao, B., H. Zheng, S. Wang, K. R. Smith, X. Lu, K. Aunan, Y. Gu, Y. Wang, D. Ding, J. Xing, X. Fu, X. Yang, K. N. Liou, and J. Hao. Change in Household Fuels Dominates the Decrease in $PM_{2.5}$Exposure and Premature Mortality in China in 2005–2015[J]. Proceedings of the National Academy of Sciences, 2018, 115(49): 12401–12406.

[89] Zheng, S., J. Wang, C. Sun, X. Zhang, and M. E. Kahn. Air Pollution Lowers Chinese Urbanites' Expressed Happiness on Social Media[J]. Nature Human Behaviour, 2019, 3: 237–243.

[90] Zhong, N., J. Cao, and Y. Wang. Traffic Congestion, Ambient Air Pollution, and Health: Evidence from Driving Restrictions in Beijing[J]. Journal of the Association of Environmental and Resource Economists, 2017, 4(3): 821–856.

Recent Development and Research Prospect of Application of Environmental Big Data

Qian Haoqi[a,b,c,d]

(a:Institute for Global Public Policy, Fudan University;

b:LSE-Fudan Research Centre for Global Public Policy;

c:Center for Energy Economics and Strategy Studies, Fudan University;

d:Shanghai Association for Social Application of Big Data)

Abstract: Conducting quantitative researches by using high quality environmental big data in social science fields is one of the key aspects in enhancing China's environmental governance capacity. This paper defines the environmental big data system based on 5V model of big data, and establishes an analytical framework for analyzing the application of environmental big data. The analyses consist of reviewing researches in social science which involve five different environmental data types such as macro-level environmental statistics, micro-level environmental data, environmental monitoring data, satellite data and miscellaneous data. This paper finds each type of environmental data has its own advantages and disadvantages. When compared to the traditional statistical data, although environment data in new forms typically have higher time and spatial resolutions, they provide less information on pollution types and typically have low data qualities. As a result, the latter ones are currently applied to study some specific environmental problems, but have great potentials in the future. In order to expand and deepen the applications of environmental big data, more efforts should be made in three aspects such as improving data qualities, adopting new analytical techniques and enhancing collaborations.

Keywords: Environmental Big Data; Micro-level Data; Monitoring Data; Satellite Data; Miscellaneous Data

JEL Classification:C81, C82, Q50, Q56

(责任编辑:卢玲)