数据要素流通与收益分配机制研究：以风电场景融合气象数据为例

王衍之[1] 黄静思[1] 王剑晓[2] 高 锋[1] 宋 洁[1,2]
(1.北京大学工学院,北京 100871;
2.北京大学大数据分析与应用技术国家工程实验室,北京 100871)

摘要:在加快全国统一数据要素大市场建设背景下,如何设计符合我国国情的数据交易流通模式和与之相匹配的收益分配机制尚未得到充分研究。本文首先基于数据要素市场中普遍认可的三类主体设定,以气象数据供给为典型细分行业,以数据在电力预测中的应用作为典型场景,建立了数据要素交易流通模型。模型中嵌入了基于机器学习的风电预测模型,对多特征数据要素和数据服务的价值实现进行刻画。其次,在基于数据价值的收益分配机制的设计中,对比分析了平均法、留一法、沙普利值法、惩罚-沙普利值法四种收益分配机制对数据生产商(数商)的主体效益差异及市场影响。最后,研究表明:惩罚-沙普利值的收益分配策略能充分考虑市场中数据要素的差异化水平,同时能够识别并避免数据要素的复制导致的扰动。

关键词:数据要素流通;数据市场;收益分配;机器学习;沙普利值
DOI:10.14120/j.cnki.cn11-5057/f.2024.06.021

引 言

随着大数据技术的发展,数据的收集、存储和使用的方式发生了前所未有的变革。数据逐渐作为一种新型生产要素参与到各行各业的生产中,数字经济也成为推动国家发展、实现弯道超车的重要选择。我国拥有构建大规模数据要素市场的条件,也具备在数字经济时代获得战略先机的基础。据国际数据公司(IDC)测算,到 2025 年我国拥有的数据量将达到 48.6ZB,占全球 27.8%[1]。

在数据要素流通方面,国家数据局等相关部门联合发布的《“数据要素×”三年行动计划(2024—2026年)》中,明确提出了创新气象数据产品与服务的需求,以支持风能企业整合应用气象数据,从而实现成本降低和效率提升。此外,行动计划中还涉及在商贸流通、金融服务以及科技创新等多领域,通过跨行业跨主体间数据要素的流通与融合,依托大数据技术以进一步提高数据的内在价值,赋能经济发展。如今数据的流通和通过数据流通带来的产业赋能已经覆盖了能源、交通、农业、工业制造等各行各业,2022 年我国数字经济规模已经达到 50 万亿元[2]。贵州、北京、上海等省市相继成立了数据交易所,承担数据服务缔造价值的职能,推动数据正规化交易流通。这些数据交易所通过引入多类别数据主体,从数据要素生产、加工到销售环节各司其职,以实现高效化的价值实现。为了有效改善和监管数据要素价值实现流程中可能存在的数据隐私、数据定价、数据安全等问题,国务院相继印发《要素市场化配置综合改革试点总体方案》《关于构建更加完善的要素市场化配置体制机制的意见》等政策,完善数据要素市场建设、规范数据交易流通。在加强政策法规监管的同时,需要制定具体可持续的数据交易机制,通过有效的要素流通以及合理的收益分配,使数据交易成为衔接数据供给方与需求方的桥梁。理想的要素市场,一方面通过数据资源的整合、加工服务,实现数据资源向数据产品的转化;另一方面通过科学合理定价与收益分配,实现数据产品在供需双方间的有效匹配。因此,设计有效的数据要素流通机制和可持续的收益分配方案,是多数据要素主体参与数据市场建设面临的关键问题。

从数据服务的供给—加工—需求出发,本文建立了多主体数据要素流通模型,刻画了数据生产商、数据服务商和数据服务需求商三类主体的数据流通与交易过程。在数据要素-数据服务的过程中,数据生产商提供含有多种特征的数据要素,由数据服务商整合并建模形成更完善的数据服务,最终提交给数据服务需求商。收益分配过程中,数据服务需求商将一部分从数据服务获得的收益作为报酬支付给数据服务商;在扣除因服

收稿日期:2023-09-30
基金项目:国家资助博士后研究人员计划(GZC20230050);国家自然科学基金项目(72241420)。
作者简介:王衍之,北京大学工学院博士研究生;黄静思,北京大学工学院助理研究员,博士;王剑晓,北京大学大数据分析与应用技术国家工程实验室助理研究员,硕士生导师,博士;高锋,北京大学工学院助理研究员,博士;宋洁(通讯作者),北京大学工学院教授,博士生导师,博士。

务建模等成本后,数据服务商再将剩余收益分配给数商,最终形成数据交易的闭环。本研究通过选取平均法、留一法、沙普利值法及惩罚-沙普利值法作为收益分配策略,基于它们在预测性能提升中的贡献度评估及对市场公平性和数据复制性问题的考量,结合风电出力预测案例的实证分析,旨在得到一个既反映数据本质特性又确保市场公平的收益分配机制,为数据市场公平、高效的收益分配提供理论与实践的指导。通过对比分析,本文发现,惩罚-沙普利值法能够充分考虑市场中数据要素的差异化的同时,保障数据要素主体的经济效益,并在出现要素复制等扰乱市场行为时进行诊断和遏制。最后,从政策层面出发提出政策建议,数据要素市场的构建和发展,不仅需要制定有效的要素流通机制,还需制定合理的收益分配机制,才能够有效地维持数据市场的公平和秩序,促进市场良性发展。

综合考虑,通过引入可度量的价值评估方法,本文加深了对数据要素市场流通与收益分配机制设计的理解,在数字经济政策背景下实现了数据市场构建与收益机制的探析。此外,结合风电预测的案例分析,本文展示了研究方法在实际数据要素-数据服务应用中的有效性,并为数据市场的公平与有效性提供了实践层面的策略指导。

文献回顾

本文主要探讨了将气象特征数据用于电力预测的细分数据市场场景下的数据流通与收益分配机制,因此主要从数据流通机制研究和数据定价与收益分配机制研究两方面进行文献回顾。

在数据要素的价值链条中,数据流通与数据定价、收益分配是密不可分的。数据流通机制关注链条中参与主体的定位和功能划分,而科学合理的流通机制需要与之相匹配的数据定价与收益分配模式的支撑。基于学者们对于数据属性的探讨,数据流通可分为面对数据产品和数据资产时的流通机制研究[3],将数据产品定义为经过抓取、重新格式化、清洗、加密等处理后的数据集和数据集衍生的信息服务[4]。由此可见,数据产品与数据处理技术通常相伴而生。在上述定义下,数据要素交易的参与者主要包括三类市场主体——数据提供商、数据交易中介、数据买家[5],以及两类主要的交易模式:第一类为数据提供商主导的直接交易模式,即数据提供商直接向数据卖家出售相关数据产品或提供数据服务,并从中获得一定收益,交易实施的具体方式通常是由提供商直接向数据买家提供数据接口并授权访问,这类模式只涉及数据提供商、数据买家两类主体。第二类为基于数据平台的中介交易模式,即数据平台将数据提供商和数据买家聚集在一起,由其充当可信第三方中介机构对接数据供需双方,促成交易。充当第三方中介的机构可以是大数据交易所等专业数据交易平台,也可以是专业数据集成商。在这类模式中,数据提供商通过数据整合服务,将数据资源加工成可以交易的数据产品。数据交易中介不但承担了数据资产价值评估和交易撮合功能,还提供了数据采集、整理、聚合和分析的再加工服务以及安全性高、操作规范的交易场所,甚或发挥着争议仲裁等市场运营的职责,在数据要素市场的建设中发挥重要的作用[6]。从数据价值链主体出发,不盲目强调数据所有权,而是强调相关各方在数据价值创造过程中的主体地位和贡献,更容易实现合作和价值共创。本文所探讨的数据流通模式属于上述第二类。在这种依赖数据中介机构的集中式双边市场模式下,数据交易中介拥有较强的数据采集和数据分析能力,能提供整合程度和数据质量更高、种类更丰富的数据产品和服务[7],能够实现在不影响数据所有权的前提下交易数据使用权,从而在保护数据供应商权益的同时,最大限度地开发数据价值[8]。

由于数据资源具有可复制性、可共享性的特点,在数据流通与数据交易过程中,拥有数据资源的数商依赖于拥有大数据技术以及产品/服务包装能力的数据服务商提供的数据清洗、数据训练、隐私保护等服务[9]。因此,在我国数据市场的架构中,数商与数据服务商天然形成了合作博弈模式,与需要数据产品/服务来提升经济决策效益的数据服务需求商开展交易。虽然面向数据服务商的收益分配机制模型的相关研究较少,但是在一般商品交易的合作博弈研究中有较为丰富的理论成果。例如,沙普利值法(Shapley value,SV)是合作博弈论中经典的收益分配方法,沙普利值法通过遍历除了该元素对其以外的所有子集的边际效用做赋权平均,以得到其在整个博弈中的价值分配,其满足了现实收益分配场景中一系列公平性原则[10]。在数据使用场景中,每个数据提供者看作是合作博弈中的参与者,通过效用函数来表征任意数据子集的贡献[11]。在数商与数据服务商的合作模式下,除了合理的数据收益分配方案,数据服务商所选取的数据清洗、数据训练、隐私保护技术同样对于数据价值的实现发挥着决定性作用,模型准确性甚至成为价格的主要影响因素[12]。以阿里、腾讯为代表的科技公司是实践数据流通与数据赋能的先行者,通过收集大量的用户消费数据,以建立准确的机器学习模型[13-15]。其中,机器学习(machine learning)作为人工智能的一个重要分支[16],在工业界和学术界都取得了令人惊叹的快速发展。机器学习的相关技术方法具有如下的独特优势:从复杂高维数据环境中提取

有价值的信息,以更好地帮助变量测量、事件预测、因果推断、可解释模型构建[17-19]。在大数据环境下,通过机器学习对大量冗余的低质量数据开展数据治理,能够降低数值实验的计算成本与计算复杂度[20]。比如在风力预测场景中,通过筛选计算得到的高价值数据可以降低预测误差,进而实现系统的降本增效[21]。目前已经有许多国内外文献采用机器学习方法,研究资产定价[22]、私人信息[23]、信息和市场效率[24]等相关课题。在人工智能催生出更多数据要素使用逻辑的背景下,2022年末我国印发的《关于构建数据基础制度更好发挥数据要素作用的意见》聚焦于数据要素市场培育中的数据确权、流通交易、收益分配、安全治理的建设建议,国家寄希望于数据要素市场成为数据资源整合共享和开发利用的载体。在此背景下,王建东等[25]论述了推动数据要素市场化配置改革中的产权制度、供给制度、流通制度、分配制度和跨境制度五个方面的制度设计;杨铭鑫等[26]探讨了数据要素参与收益分配的三层次制度路径。

通过上述文献综述可以发现,截至目前,几乎没有研究从数据要素流通的角度研究潜在市场构建的可行性,以及基于数据要素特性设计合理的收益分配机制。针对在不同行业及特定数据要素应用场景中实施数据确权、定量化数据交易和收益分配策略,以既照顾到市场参与各方的利益,又推动细分行业数据要素市场向有序发展转型的问题,目前在理论与实践案例研究方面仍有广阔的探索空间。本研究旨在缩小这一研究空缺,通过具体分析细分气象行业数据要素在电力预测应用场景中的数据流通与收益分配机制,提出针对性的解决方案。本文设计了由“数据生产商”“数据服务商”和“数据服务需求商”为主体的数据流通机制模型。在场景应用方面,本文选择了结合气象数据的新能源预测作为案例进行分析,并关注于不同气象数据供应商与需求商之间的数据采购。风电作为最重要的可再生能源之一,具有很强的不确定性和波动性,而且受气象因素的影响尤为显著。实际上,对风电的高效预测可以帮助风能企业在即时市场中做出准确的交易决策,甚至减弱可再生能源波动对电力系统的不稳定性产生的影响。因此,风电预测体现了气象数据要素的价值,并有望将其推广到风能部门与气象部门在数据要素上的有机融合。本文以结合气象数据的风电数据预测为案例场景,通过机器学习方法建立数据价值评估模型,对比了平均分配法、留一分配法、沙普利值法和惩罚-沙普利值法四种收益分配方法的适用性,为数据要素市场流通分配机制的完善提供了参考与建议。

数据要素流通与收益分配机制设计

1.数据要素流通机制架构

为了便于提炼数据要素形成数据服务并最终实现价值增加的过程,本文将数据要素市场(以下简称“数据市场”)中的参与主体构成按照供给、加工、需求,总结为数据生产商、数据服务商(数据运营平台)、数据服务需求商三类。三类主体的定义介绍如下:

数据生产商(以下简称“数商”):通过收集等方式获得原始数据要素的机构,为数据服务的生产和加工提供了重要的要素来源。

数据服务商:以各类数据要素为基本生产资料,利用大数据技术[27]为数据应用提供增值服务的经济主体。数据服务商通常拥有核心技术或核心数据要素作为生产力,相比数商,其存在能够满足特定类型的数据服务需求的特殊性。数据服务商能够结合不同数商的原始生产要素,以实现多元数据主体的经济效益最大化,并根据需求交付数据服务。同时,数据服务商还兼具对数商提供数据要素层面经济补偿(收益分配)的职责。总结来说,数据服务商是数据市场运转的核心。

数据服务需求商:基于自身经济生产活动而产生数据服务需求的经济主体,通过委托数据服务商提供数据服务,并支付相应的服务费用。

其中,数据服务商作为要素流通的中间商,是联结数据供需双方的重要桥梁和纽带,它既是数据交易的组织者,也是交易活动的参与者,兼具市场监管主体和被监管对象双重角色,在数据市场和监管体系中占据着十分重要和特殊的地位[28]。数据要素通过在生产商和服务商之间流通,借助机器学习等大数据技术的加工,数据要素由资源转变为服务产品,进而与数据服务需求商交易以实现从生产到消费的流通。三主体通过数据流、价值流进行数据要素-数据服务的交易流通与收益分配,具体的交易流通过程如图1所示,包含如下5个步骤:

①“数据服务需求商”向“数据服务商”提出数据服务需求。

②“数据服务商”根据需求类型设计基于大数据技术的模型算法,并从相关的“数商”中收集不同类型的特征数据。

③“数据服务商”结合数据和模型进行计算和提取关键信息,向“数据服务需求商”提供数据服务。

④“数据服务需求商”根据服务质量以及之前签订的数据服务买卖协议,向“数据服务商”支付服务费用。

⑤“数据服务商”通过计算每一个“数商”所提供的数据要素资源对价值生成的贡献,将剩余收益合理分配给“数商”。

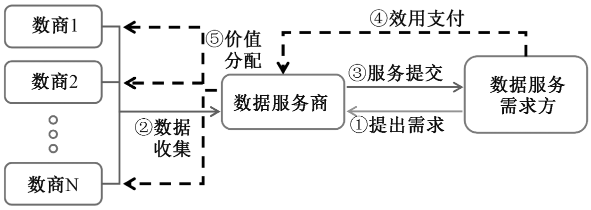

图 1　数据要素流通与收益分配流程

2.“要素流通-收益分配”模型构建

本部分通过对数据交易流通过程中各主体决策过程建模,深入解析数据要素通过交易流通创造的市场价值。以面向能源行业的数据服务为例,比如数据服务需求商需要获得未来某一地区风力发电总量的预测数据,以更准确地调整自身管理决策(调整地区火电调配策略等),数据服务内容则为对这一具有延伸经济价值的时间序列数据特征的估计,在模型中用时间序列 $Y \in \mathbb{R}^{T}$ 表示,其中 T 为特征数据的时间跨度。一般来说,关键特征数据的获取门槛相对较高,因此大多数据服务需求商无法通过自有渠道获得具有该特定特征的数据服务,而数据服务提供商具备这种数据服务的供应能力(比如电网企业的大数据部门)。然而,实际中时间序列往往具有显著的波动性和难以预测性,仅仅依靠对其内在惯性规律的掌握是不够准确的,需要结合其他类型的特征数据进行辅助分析以达到更高模型精度。因此,借助数据要素市场中众多数商提供的特征数据,数据服务商能够形成更精准更高质量的数据服务,更好地实现数据要素之间的有机结合和更高层次的价值实现。在本文的模型构建中,定义其他数商的数量为 N,其中每一个数商的数据集有 n_j 个特征,特征数据集为 $x_j \in \mathbb{R}^{n_j \times T}, j \in [1,2\cdots N]$。值得一提的是,以气象数据为例,这些数据一般可以在较短的时间范围内进行有效地评估,所以本文认为和待预测变量是同时间维度的。

定义满足数据服务的预测模型为 $F:\mathbb{R}^{(\sum_{j=1}^{N} n_j + m)T} \to \mathbb{R}^{T}$,那么,对 Y 的预测结果 $\hat{Y}$ 的表达式如下:

$$\hat{Y} = F(Y_{-mT}, x_1, x_2, \cdots, x_N) \tag{1}$$

其中,Y_{-mT} 表示使用 $m \times T$ 规模的历史时间序列主要特征信息作为输入,其中 $m \in \mathbb{R}^{+}, m \times T \in \mathbb{N}^{+}$。比如,当 $m = 1$ 时,则用过去 T 时间内的主要特征数据作为输入进行预测。x_i 为数商 i 提供的额外数据特征。对于服务需求商而言,预测结果的好坏直接反映了服务质量的高低,进而直接影响了数据服务指导决策而创造的经济价值规模。定义数据服务带来的效用函数为 $G:\mathbb{R}^{2T} \to \mathbb{R}$,预测结果 $\hat{Y}$ 带来的经济价值 K 为:

$$K = G(Y, \hat{Y}) \tag{2}$$

假设该价值函数对于数据服务商而言是已知的,所以数据服务商提升数据服务的目标和需求商带来经济效益的方向是一致的。基于此,本文定义数据服务商训练得到的预测模型(数据服务) F_{best} 的表达形式为:

$$F_{best} = \mathrm{argmax}_F G[Y, F(Y_{-mT}, x_1, x_2, \cdots, x_N)] \tag{3}$$

值得一提的是,为了书写方便和减少符号复杂度,在变量标书中,不对式(3)中训练数据集和式(1)中的预测数据集进行区分。式(3)表明,在历史数据集中训练得到的最优预测模型 F_{best},该模型在式(1)中将被应用未来 T 时间内的预测,结果为 $\hat{Y}$。

假设数据服务需求商也相应拿出一定比例的数据带来的经济效用增量用以支付服务费用 Q_F,定义为:

$$Q_F = b \times K = b \times \max_F G[Y, F(Y_{-mT}, x_1, x_2, \cdots, x_N)] \tag{4}$$

其中,$b \in (0,1]$,代表服务商的分成比例。数据服务需求商支付的服务费用由以下两部分构成:一部分与数据服务商自身的历史特征数据和模型技术相关,用 Q_S 表示,且 Q_S 表示为仅有数据服务商自身参与数据服务时的价值;另一部分与数商提供的其他特征数据对于预测效果提升的增益水平相关,用 Q_O 表示。Q_O 和 Q_S 分别通过式(5)和式(6)定义:

$$Q_O = Q_F - Q_S \tag{5}$$

$$Q_S = b \times \max_F G[Y, F(Y_{-mT})] \tag{6}$$

在式(6)中,$\max_F G[Y, F]$ 是一种简化写法,意味着在该预测模式下,利用历史数据训练得到的最优模型 F_{best} 之后,在待预测数据上的表现效果,展开为:

$$\max_F G[Y,F] = G[Y,\mathrm{argmax}_F G[Y,F]] \tag{7}$$

考虑到数据模型 F 的训练过程只使用历史数据,所以 $\max_F G[Y,F]$ 也可以理解为基于当前时间节点下,相对于未来价值最大化的服务效用。为避免混淆,之后类似写法意义相同。

3.数据服务商收益分配机制

目前我国数据要素市场尚处于起步阶段,如何设计合理的收益分配机制,吸引广大数商积极参与市场交易,打破数据孤岛,最大化发挥各方数据要素在经济生产中的作用,是数据要素市场需要解决的重要问题。考虑到数据服务商除关键要素外的其他技术成本不是本文考虑的重点,本文默认该成本可以被补贴,即通过数据要素流通融合带来的价值剩余全部分配给数商。在式(8)中,用 q_i 表示数商提供数据创造的价值剩余 q_i,$i \in [1,2,\cdots,N]$ 表示:

$$q_i = u_i \times Q_O \tag{8}$$

$$\text{s.t} \quad \sum_{i=1}^{N} u_i = 1 \tag{9}$$

$$u_i \geqslant 0 \tag{10}$$

其中,$u_i \in [0,1]$ 为每一个数商在价值分配中的占 Q_O 的配额。本文假设整个数据市场的价值分配是充分的,所以配额的总和为1。如何在兼顾效益与公平的前提下分配收益 Q_O,将是本文接下来需要深入探讨的问题。本文提出了4种可能的分配方式,如下:

(1)平等分配

由于每个数商的特征数据都以相同的格式整合到用于风电预测的机器学习模型中,因此,此机制将每个数商的贡献视为平等,具体情况如下:

$$u_i = 1/N \quad i \in [1,2,\cdots,N] \tag{11}$$

其中,N 为数商数量。

(2)留一分配

留一分配机制以最终的经济效用 K 作为基准,计算与去除了该数商提供的数据集后的模型的经济效用的差值,作为该数商在整个数据服务的边际贡献。在利用归一化函数 M 后,使其满足条件式(9)、式(10),并作为配额比例,以保证收益分配的非负性。

$X \coloneqq [x_1,x_2,\cdots,x_j,\cdots,x_N]$ 作为全数商提供的其他特征数据,$X_{-j} \coloneqq [x_1,x_2,\cdots,x_{j-1},x_{j+1},\cdots,x_N]$ 为去除特征 j 的数据集。边际效用下数商的价值分配 u_i 的具体表达式如下:

$$\tilde{u}_i = \max_F G[Y,F(Y_{-mT},X)] - \max_F G[Y,F(Y_{-mT},X_{-i})] \quad i \in [1,2,\cdots,N] \tag{12}$$

$$u_i = M(\tilde{u}_i) = \frac{\tilde{u}_i - \min \tilde{u}_i}{\sum_{i=1}^{N}(\tilde{u}_i - \min \tilde{u}_i)} \quad i \in [1,2,\cdots,N] \tag{13}$$

(3)沙普利值法分配

沙普利值(Shapley value)是劳埃德·沙普利为了解决博弈中公平分配合作收益而提出的概念。它是一种基于留一分配的衍生方法,通过遍历不包含该数商的所有数商组合,计算该数商边际效用的期望,以得到收益的分配结果。此处,忽略 b 的作用(或者认为 $b = 1$),需要分配的全部剩余价值总量为 $\max_F G[Y,F(Y_{-mT},X)] - \max_F G[Y,F(Y_{-mT})]$。本文将每一个数商提供的特征数据看作是参与博弈的玩家,将不同数据集之间的组合看作是玩家之间的合作,并将组合带来的预测服务的经济价值增量看作效用 v,表达式如下:

$$v(S) = \max_F G[Y,F(Y_{-mT},\{x_i \mid i \in S\})] - \max_F G[Y,F(Y_{-mT})] \tag{14}$$

其中,S为任意数商的组合,$v(S)$ 为该组合带来的收益增量。于是,基于沙普利值法的收益分配的具体表达式如下:

$$sv_i = \sum_{S \subseteq [N] \setminus \{i\}} \frac{|S|!\,(N-|S|-1)!}{N!}[v(S \cup \{i\}) - v(S)] \tag{15}$$

其中,$|S|$ 表示集合 S 中数商的数量,i 为数商编号,sv_i 为该数商的沙普利值。沙普利值通过对该数商所拥有的特征数据集于所有其他子集的边际效用加权求和。沙普利值在价值分配任务中具有良好的性质,具体如下:

A.累加性

通过对计算得到的每个数商的特征数据集的沙普利值进行累加,能够确保其总和相当于拥有全部数商的

完整特征数据集所带来的效用,即:

$$v([N]) = \max_F G[Y,F(Y_{-mT},X)] - \max_F G[Y,F(Y_{-mT})] = \sum_{i=1}^{N} sv_i \tag{16}$$

其中,$[N]$ 为数据全集 $\{1,2,\cdots,N\}$ 。所以,若沙普利值不存在小于 0 的值,则不需要进行标准化处理,即满足分配性质。

B.一致性

如果对于任意数商组合而言,数商 i 的数据集和数商 j 的数据集的作用存在一致的比较关系,其对应沙普利值 sv_i 和 sv_j 也存在一致的比较关系。

(4)惩罚-沙普利值法

通过对平等分配、留一分配以及沙普利值法分配机制的建模分析,发现沙普利值是根据每一个特征数据集对于其补集中的所有子集的边际效用的期望计算得到的。当两组特征数据完全相同,由于二者的特征补集也完全相同,则计算得到的沙普利值相同,最后对应的价值分配也会一致。对于绝大多数机器学习模型而言,额外相同的特征数据并不会影响回归结果,但在计算其沙普利值时,对于不包含重复数据的其他特征数据集合的边际效用却被重复计算了。在一定条件下,重复数据集中单个效用分配降低,但总效用会超过未复制的情况,这会引发潜在的扰乱市场秩序的复制行为。鉴于数据要素相较于其他生产要素具备无限可复制的特性,即复制数据要素的成本几乎可以忽略不计,数据生产者可能会通过复制自身数据来扰乱市场。具体来说,数据生产者可以复制自身数据(或者加入轻微扰动),并以此人为制造一个新的生产者加入市场,参与数据要素流通以及收益分配。在沙普利值法分配机制的框架下,由于待分配的收益总量是固定的,这种重复的效用计算可能会导致复制数据的生产者可以获得成倍的收益,其他数据生产者的效益受到侵占,从而影响数据市场的稳定性。所以,本文希望改进沙普利值法,使之能够识别数据要素的复制现象并对复制行为进行惩罚,从收益分配的角度遏制市场紊乱。

文中使用 psv_i 表示经过惩罚修饰的沙普利值法(penalty-modified Shapley value),在根据上文计算过程得到每一个数商的沙普利值 sv_i 后,遍历所有数据集并判断是否有重复集合,利用自然指数 e 进行调整,具体表达式如下:

$$psv_i = sv_i \times \exp\Big(-\lambda \sum_{j\in[N]\setminus\{i\}} \mathcal{S}_{\varepsilon}(x_i,x_j)\Big) \tag{17}$$

其中,λ 为惩罚参数,$\mathcal{S}_{\varepsilon}(x_i,x_j)$ 为相关性检测函数,可以通过定义 x_i、x_j 数据集之间的分布上的距离来进行复制的界定,ε 为距离参数。所以,如果两个数据集的距离小于 ε,则返回 1,即界定为该数据集与其他数据集存在复制或高度相似;否则返回 0,认为数据集相关性较低。所以,可以设置合适的距离参数 ε ,以惩罚当数据大部分重合时导致两个数商的数据集相关性较高,而引发的对市场的干扰。如果以皮尔逊(Pierson)相关系数的绝对值与 1 的差值作为时间序列数据集之间距离的定义,$\mathcal{S}_{\varepsilon}(x_i,x_j)$ 可以写成如下形式:

$$r_{i,j} = \frac{\sum_{t=Start}^{End}(x_{it}-\bar{x}_i)(x_{jt}-\bar{x}_j)}{\sqrt{\sum_{t=Start}^{End}(x_{it}-\bar{x}_i)^2}\sqrt{\sum_{t=Start}^{End}(x_{jt}-\bar{x}_j)^2}} \tag{18}$$

$$\mathcal{S}_{\varepsilon}(x_i,x_j) = \begin{cases} 1, & 1-|r_{i,j}| \le \varepsilon \\ 0, & 1-|r_{i,j}| > \varepsilon \end{cases} \tag{19}$$

其中,$Start$ 和 End 为用于检测相关性的时间序列历史样本的起始点和终至点。值得一提的是,惩罚-沙普利值不保证满足累加性,所以需要归一化处理,使其满足分配比例之和为 1。值得一提的是,本文为强调数据的无限可复制性,仅考虑数据完全复制情况。

实证结果及分析

1.数据服务场景和效用定义

在本文案例分析中,数据服务的核心内容为对可再生能源(风力发电)的每小时出力情况进行预测,数据服务的质量体现在对风力发电量的预测精度上,随着预测精度的提升,可为数据服务需求商带来更显著的经济收益。本文分析了 2018 年 3—7 月,共 150 天内四川省的风力发电每小时数据,总计 3600 个小时[29]。数据集按照 8∶2 的比例划分为训练集和测试集。通过对风力发电数据进行标准化处理,具体输出功率分布如图 2 所示,发现发电量的波动性十分显著:平均出力为 0.311,标准差为 0.274,显示出力在不同小时内的巨大差

异。最小出力几乎为零,而最大出力达到 1,这进一步凸显了风力发电的极端不稳定性。此外,中位数为 0.226,意味着超过一半的时间段内,发电出力低于平均水平;75%的数据不超过 0.453,表明较高出力水平出现的频率较低。统计表明风力发电量的显著波动性和高出力的低频出现,强调了进行高效和准确预测的迫切需求,这对于保证电力系统的稳定运作和提高风电资源的利用效率具有至关重要的作用。

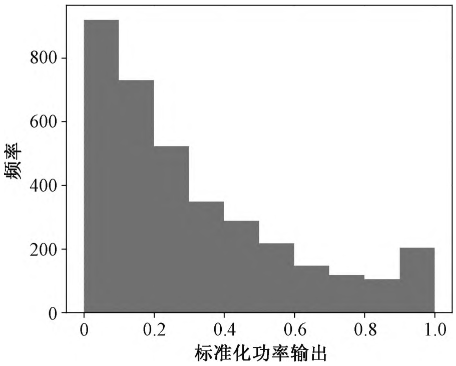

图 2　风电输出功率分布

基于已有的风电预测研究,本文以 $T=24$ h 为预测窗口大小,将式(6)中训练数据比例 m 取为 1,即利用过去 24 小时的历史风电数据预测未来 24 小时的风电出力[30]。将来源于美国 NASA 气象局官网的气象数据作为数商其他的特征数据输入预测模型。该数据集包含县域尺度的特征,本文对其进行聚类和平均处理,以实现对全省级别的气象特征的概括。

本文对所得到的小时级气象数据的特征根据类型进行分类,在数据市场中设计了三个代表性数商,分别是:

数商 1:提供离地面 2 米处的平均温度(temperature at 2 meters,T2M)、当天最高温度(T2M_MAX)和当天最低温度(T2M_MIN)的相关数据。

数商 2:提供离地面 2 米处的平均比湿度(specific humidity at 2 meters,QV2M)和相对湿度(relative humidity at 2 meters,RH2M)的相关数据。

数商 3:提供地标压力(surface pressure,PS)和降水量(precipitation,PRECTOTCORR)的相关数据。

数据服务商通过收集这三个数商的全部气象数据训练四川省风力发电的预测模型。在风力发电量预测模型的选择方面,本文基于统计学习理论考虑了多种实用模型适合于多特征风电预测[31],包括线性回归(linear regression,LR)、LASSO、K-近邻算法(k-nearest neighbor,KNN)、支持向量机(support vector machines,SVM)、随机森林(random forest,RF)、神经网络(neural network,NN)、梯度提升决策树(gradient boosted decision tree,GBDT)和 LightGBM。其中,线性回归是最基本的预测模型,用于建立一个或多个自变量与因变量之间线性关系的预测,其优势在于模型可解释性强,易于实现。在此基础上,LASSO 利用正则化以增强模型的预测准确性;K-近邻算法是一种基于距离的回归技术,不需要假设数据的分布,灵活性高;支持向量机在特征空间中寻找最优的分割边界,适用于高维数据的回归问题;随机森林是一个包含多个决策树的集成技术,其随机性带来了高准确性和对过拟合的控制;神经网络能够模拟非线性和复杂的关系,特别适合大规模数据处理,能够通过学习识别数据中的复杂模式和关系。此外,梯度提升决策树和 LightGBM(light gradient boosting machine)是一种集成回归树模型进行分析[32]。作为一种基于梯度提升决策树的改进模型,LightGBM 在训练时将整体预测任务拆分成多个子模型,即:

$$\hat{Y}_i=\sum_{k=1}^{K}f_k(\mathbf{x}_i) \tag{20}$$

其中,$\hat{Y}_i$ 表示预测目标的某一个分量,$\boldsymbol{x}_i$ 泛指预测 $\hat{Y}_i$ 所需的全部数据输入。每一个 f_k 均为一个子回归树模型,其训练的目标为:

$$\min_{f_k}\mathcal{L}^{(k)}(f_k)=\sum_{i=1}^{n}l(Y_i,\hat{Y}_i^{(k-1)}+f_k(\boldsymbol{x}_\mathrm{i}))+\Omega(f_k) \tag{21}$$

其中,$\hat{Y}_i^{(k-1)}$ 表示训练完前 $k-1$ 个子回归树模型之后对 Y_i 的预测结果,Ω 是对树模型的复杂性度量。此外,相比梯度提升树,LighGBM 在训练过程中提高了特征数据采样和排序的效率,在模型的训练速度和训练时内存消耗方面都有显著提升和优化。此外,LightGBM 已经被验证能有效捕捉时间序列数据中的内在规律,并能够在保证精度的同时处理多特征数据集[33]。

在经过风能与气象数据收集和预测模型的训练后,针对预测结果 $\hat{Y}_i$,本文使用平均绝对百分比误差(mean absolute percentage error,MAPE)作为预测精度的标准,表达式如下:

$$MAPE(Y,\hat{Y}) = \frac{100\%}{24}\sum_{t=1}^{T}\left|\frac{\hat{Y}_t - Y_t}{\max(Y_t,\eta)}\right| \quad (22)$$

其中, η 为小量,作为定义的误差下界,用于避免因为 Y_t 过小而导致微小的预测误差 $\hat{Y}_t - Y_t$ 引起百分比的剧烈变化。

从数据服务商的角度出发,本文利用自身的历史数据以及可获得的所有气象数据来训练不同的风电预测模型,以确定哪个模型可以实现数据要素的收益最大化。具体模型的性能如表1所示,从表1可以观察到,LightGBM 在所有模型中表现最出色。因此,数据服务商选择 LightGBM 作为基准预测模型,以确保数据要素的最大化收益。

表1　不同预测模型表现

模型编号	MAPE (%)
Linear Regression	78.7
LASSO	51.6
KNN	58.8
SVM	64.5
RF	51.7
NN	66.1
GBDT	53.8
LightGBM	45.4

研究表明,经济收益的上升与预测准确性的提高存在明显的对应关系[34]。本文将数据服务需求商的利润函数 G 与 $MAPE$ 近似为负线性关系,具体表达式为:

$$G(Y,\hat{Y}) = U - h \times MAPE(Y,\hat{Y}) \quad (23)$$

其中, U 为保证利润函数 G 正值的参数, h 为预测精度对需求方利润带来影响的比例系数。在本案例分析中,取 $\eta = 0.12$, $U = h = 100$。而之后的不同数商之间的收益分配,也均是基于此数值按比例划分。

2.风电预测场景下的收益分配

数据服务商汇总全部数商所提供的气象数据后,结合自有的历史风力发电数据训练预测模型,并将数据模型作为数据服务的核心技术。服务需求商利用数据模型获得未来24小时风力发电量的预测结果,基于预测开展精准决策,最终服务需求商将科学预测决策带来的部分经济收益作为报酬给予数据服务商。数据服务商在扣除自身风电数据管理、预测模型训练等成本费用后,将剩余的收益分配给数据提供商。数据服务商通过实际风力发电出力作为检验,事后分析利用所有不同气象特征数据组合带来的建模预测效果上的变化。图3展示了不同数商的气象特征数据组合建模的预测效用相较于仅使用数据服务商的风力发电历史数据(没有数商数据参与)的增益:实线圆形框对应仅采用单个数商气象特征;虚线方形框对应邻近直线所连接的两个数商气象数据;点线三角形框对应使用全部特征数据。可以发现,任意气象特征的组合使用都能带来风电预测精度的提升,说明了额外数据要素在风电预测场景下对价值提升的泛化性。同时,不同特征数据的组合带来的提升作用是不同的,例如组合"数商1+数商3"可以带来+7.324的效用增量,高于组合"数商1+数商2+数商3"的+6.420,此时数商2的特征数据对其他数据要素的效用带来了负增长;但数商2可以对数商3的效用产生促进作用,组合"数商2+数商3"的效用大于数商3本身的效用。由此可见,数据要素的单一累积并不一定会导致效用的单调上升,单一数据要素的边际收益会根据剩余子集不同而改变。因此,为了更好地适应数据要素的非均质性和场景依赖性,数据服务商亟须建立差异化的收益分配机制,以保证数据要素市场的公平和稳定。

更进一步,当式(4)中参数 $b = 1$,即数据服务商采用全部特征数据进行模型训练,且将由额外数商数据产生的收益全部用于数商的收益分配,数据服务商不收取任何额外费用。在平等分配、留一分配、沙普利值法、惩罚-沙普利值法四种收益分配机制下,不同数商分配的收益额度如表2所示。

平等分配使每一个数商的收益一致化(均为2.140),但不同特征数据带来的预测效用提升是不同层次的,所以该分配方式并不能凸显不同数据要素的特殊性。从数据要素市场机制的角度,平等分配不利于激励数商提供更高价值的数据以增强数据价值的发挥,抑制了市场活性。

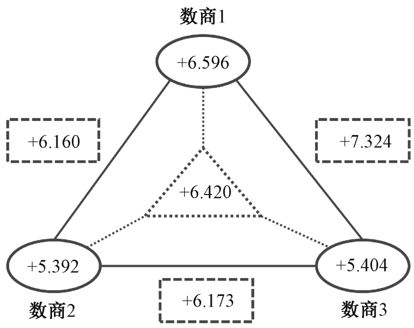

图3　不同数商数据组合带来的效用变化

相比之下,留一分配基于数据的后验边际效用判断数据要素的价值,考虑了数据要素的异质性,极大增强了收益分配差异性,使得数商1和数商3近似均分了全部收益(49.7%和50.3%),数商2没有被分配任何收益。这是因为数商2加入"数商1+数商3"组合后带来了效益的负向增长(-0.904),与此同时,虽然数商1和数商3加入剩余数商组合后带来的效益增益较小,但效果非常接近(+0.247和+0.260)。在留一分配模式下,后验边际效用表明"数商1+数商3"为最优的组合结果,从而否定了数商2参与数据要素市场的必要性。但从单一数商所提供的数据价值的视角出发,数商2的特征数据与数商3的特征数据的单独提升效果相差无几(+5.392和+5.404)。此外,如果数据要素市场中不存在数商1,那么数商2对数商3的提升效果是显著的。也可以从另一个视角解读导致以上效用结果的原因,原本数商1单独作用效果好于数商3,但是"数商2+数商3"组合拉低了数商1原本的高边际效用。那么,从这个角度分析,为了弥补数商1的损失,数商1应该获取更高收益,以补偿数商2和数商3对其在效用上的削弱作用,但这与数商1在实际留一分配中的结果不相符。综上所述,边际效用分配模式能够直接反映单个数商对其余数商之间的贡献,但是忽略了其在其他组合之间的促进/削弱关系,否定了对全集产生负作用的数据要素,忽视了单个数据要素在创造价值上的卓越表现。

相比前两种方法,沙普利值法的收益分配结果更为"折中":该方式肯定了数商参与市场的意义,即给予每一个数商一定的正收益;同时考虑了不同数商之间提供数据要素对数据服务效用的提升的多样性,即在最终的收益分配结果上保持区分度。数商2由于对组合"数商1+数商3"以及数商1存在负作用,其沙普利值最低。考虑到数商1、数商3对于它们的特征补集带来的边际效用提升作用类似,但是数商1单独的效用优于数商3,所以数商1的沙普利值更高。沙普利值法从多层次多角度综合考虑了数据供应商的边际贡献,在保证对每一个数商提供的数据价值进行充分计算的前提下,对不同数据质量进行区分,因而得到更为全面且公平的收益分配方案。

"惩罚-沙普利值法"重点关注了数商利用重复数据集套取收益扰乱市场的行为,表2中沙普利值法得到的收益分配方案与惩罚-沙普利值法相同,这是因为假设市场中的数商不存在复制数据的行为,其中相关性检验公式(19)中的距离参数 $\varepsilon = 1$。

表2　各数商的效用分配

数商编号	平等分配	留一分配	沙普利值法	惩罚-沙普利值法
1	2.140	3.192	2.729	2.729
2	2.140	0.000	1.551	1.551
3	2.140	3.228	2.140	2.140

考虑到数据要素具有"可复制"的特性,复制数据的特征信息与原数据完全重合,因此复制数据对于模型训练不具备提升效果。但是,在包含众多数商的市场环境下,复制数据会在不包含原数据的组合上产生与原数据相同的边际效益,进而损害沙普利值法的分配效果。在接下来的数值实验中,假设数商1通过直接复制的方式,佯装为数商4进入市场参与要素分配,企图获得更大的总收益(数商1+数商4)。在这一设定下,惩罚-沙普利值法的收益分配方案区别于沙普利值法,在数商4参与市场的情况下,具体收益分配效果如表3所示,在表3中,平均分配法和留一法分配的收益也区别于表2的分配结果。

由于在机器学习模型中,相同特征数据的增加不会对训练造成影响,所以数商所能够分配的效益总额与未复制情况一致。但数商4的加入使得数商之间的收益分配比例发生明显的变化。首先,平等分配由于不等式 $\frac{1}{n} < \frac{2}{n+1}$ 在数商数量 n 大于2时恒成立(此处为 $n=3$),所以数商1复制数据带来收益分配的增量为

1.070，接近未复制时的一半，同时，其他数商的收益均有所下降。对于留一分配而言，数商 1 和数商 4 相对于其余数商组合的边际效用为 0。但是由于此前分析的数商 2 对于组合“数商 1+数商 3”的负作用，经过配额比例标准化过后，数商 1 和数商 4 仍能得到正收益，收益总额为 3.904，大于原收益 3.192。所以，这两种分配方式在没有使数据服务总效用提升的前提下，都使数商 1 通过数据复制行为获得额外的收益，进而影响了其他数商的收益分配，干扰市场的有序进行。

这种现象也存在于沙普利值法的收益分配框架下。尽管数商 1 对于任意包含数商 4 的组合的边际增益为 0，但是其对于数商 2 和数商 3 的组合仍有边际效用。由于数商数量 n 的增加，沙普利值对边际效用的加和权重下降，进而使数商 1 的收益分配降低。数商 1 和数商 4 在数商组合的边际效用计算中对称，所以数商 4 的收益分配与前者一致。但发现，数商 1 通过复制使总收益增加了 38.8%，而相对应的，其他数商对于模型增益的相对大小虽然与未复制时保持一致，但收益的绝对大小进一步降低。这是因为，即使复制数据在包含原数据组合中的边际效用为 0，但复制数据与原数据的边际效用之和在其他组合中会存在重复计算的部分，可能导致复制数据的数商的总收益增加，同时削弱其他数商的收益。所以沙普利值法不能遏制数据复制现象的发生。

通过合理的分配策略重构抑制数商数据复制行为的发生，惩罚-沙普利值法通过在沙普利值法的基础上增加对数据相关性的指数级别的惩罚，使得存在复制现象的数商 1 的相对配额大幅下降，此处取 $\lambda = 1$。由前文可知，复制数据要素由于参与者数量的增加，会导致单个数商的沙普利值下降，但由于复制数商的收益为成倍增加，其总收益可能会提升。所以，对于数商 1 拥有的复制要素的总数量为 2，其沙普利值总收益相比未复制时的增益为 1.39 倍，但同时惩罚系数将这种提升效果抵消，甚至作为惩罚使之相比未复制时更小。在表 3 中可以看出，数商 1 在惩罚-沙普利值法的收益分配模式下，其总收益为 2.222，低于未复制时的收益，为原有收益的 81.4%。所以，惩罚-沙普利值能够根据数据要素进行相关性识别，并调整原沙普利值的效用分配方式，惩罚市场中的复制行为，进一步保障数据服务收益的有效分配，促进市场的稳定运行。

表 3　存在数据复制行为情况下的各数商的效用分配

数商编号	平等分配	留一分配	沙普利值法	惩罚-沙普利值法
1	1.605	1.952	1.894	1.111
2	1.605	0.000	0.926	1.477
3	1.605	2.516	1.706	2.721
4	1.605	1.952	1.894	1.111

结论及政策建议

本文基于我国数据要素市场化的改革方向，从数据要素的“供给—加工—需求”的全流程出发，设计了由“数据生产商(数商)”“数据服务商”和“数据服务需求商”为主体的数据流通机制用以体现数据要素价值实现中的具体流程，主要通过“要素流通”和“收益分配”两种关键市场机制解释据要素的经济链。以四川风力发电预测作为数据使用场景，本文通过机器学习预测模型，探究了不同气象特征数据作为额外数据集通过数据市场在预测模型中发挥的价值，展现了数据要素产生经济效用的方式。在此基础上，为了探讨数据服务商如何制订可持续收益分配方案，本文对比分析了四种收益分配方式。结果表明，平等分配没有考虑数据要素的质量差异；留一分配则过分强调数据要素边际价值对全局的作用，产生极端分配方案，不利于促进数据要素之间通过组合带来的价值提升。相比之下，沙普利值法通过遍历数据要素组合的计算方式，其分配结果兼顾了边际效益和公平性，不失为前两者的一种折中方案。但是，数据要素具备无限可复制的特性，在前三种收益分配方式下，数商能够接近无成本地通过复制数据参与市场交易，从而获得更高的收益，这将导致数据要素市场中同质数据的冗余，降低提供高品质数据的数商的市场竞争力，从而扰乱市场秩序。惩罚-沙普利值法通过对数据相关性的测度，能够很好地惩罚数据要素市场中的复制行为，极大地降低提供同质数据的数商的总收益，进而保障市场的有序运行。

在当前全国统一数据要素大市场建设的背景下，数据服务商作为数据流通和收益分配体系的核心，其培养和建设显得尤为重要。首先，应制定一套综合性评估标准和指标体系，全面衡量数据服务商的资质和专业能力，包括但不限于数据处理能力、模型的精确度、客户满意度以及安全合规性等方面。此外，建立数据服务商的信用评价体系也是必不可少，通过系统化收集和分析客户反馈、合作案例等信息，为数据持有者提供选择高质量数据服务商的依据。

鉴于数据要素可复制性的特点,对其进行有效的监管和监督显得尤为关键。这不仅包括开发和推广高效准确的数据异常检测工具,以便快速识别数据质量问题并应用到收益分配模型中,也涵盖了建立一套明确的异常数据处理指导方针和流程,例如数据清洗、错误修正、来源追溯等。与此同时,政府和行业监管机构应加强对数据质量的监控,定期进行数据质量审查,确保数据生产商遵循相应的数据质量标准,以完善市场流通和收益分配机制。此外,面对如多个数商平等地贡献数据而不存在集中式数据服务商的情况,则更加需要独立的中介机构进行监管,以确保收益分配机制的公平性和市场的有效运作。通过综合监管和技术手段,以实现对数据生产和流通的全面监控,从而提高整个数据市场的透明度和公正性。

最后,为了促进市场的健康发展,以收益分配为核心的激励机制的建立不容忽视。这包括制定与数据生产者和服务商的数据质量、服务水平及创新能力相匹配的绩效激励政策,以及对于那些提供高质量数据服务、积极采用先进技术和在数据治理方面做出显著贡献的企业和个人,提供税收优惠和财政补贴等激励措施。

综上所述,通过具体且有针对性的策略,可以有效地提升数据服务商的服务质量,加强数据治理,同时激发市场活力,促进数据要素大市场的健康发展。在未来的研究方向方面,可以纳入对数商参与市场策略的考虑,强调数商动态进入或者退出市场的可能性,增大建设数据要素市场的考虑范围,提升其稳定性和灵活性。此外,实践中沙普利值难以泛化到不同的数据集的问题,对动态数据市场的机制设计提出了挑战。针对这一点,可以考虑引入在线学习机制和迁移学习策略,通过持续调整分配方法及在不同分布数据集上迁移知识。这有助于增强分配方法的泛化能力,从而提高其在多变数据市场中的适用性和有效性。

参考文献:

[1] 蔡继明,刘媛,高宏,等. 数据要素参与价值创造的途径——基于广义价值论的一般均衡分析[J]. 管理世界, 2022,38(7):108-121
Cai J. M., Liu Y., Gao H., et al. The Approach of Data Factor Participating in Value Creation: A General Equilibrium Analysis Based on the General Theory of Value[J]. Journal of Management World, 2022,38(7):108-121

[2] Bertsimas D., Kallus N. From Predictive to Prescriptive Analytics[J]. Management Science, 2020,66(3):1025-1044

[3] 熊巧琴,汤珂. 数据要素的界权、交易和定价研究进展[J]. 经济学动态, 2021,(2):143-158
Xiong Q. Q., Tang K. Research Progress on the Right Delimitation, Exchange and Pricing of Data[J]. Economic Perspectives, 2021,(2):143-158

[4] Yu H., Zhang M. Data Pricing Strategy Based on Quality[J]. Computers & Industrial Engineering, 2017,112:1-10

[5] Spiekermann M. Data Marketplaces: Trends and Monetisation of Data Goods[J]. Intereconomics, 2019,54(4):208-216

[6] Muschalle A., Stahl F., Löser A., et al. Pricing Approaches for Data Markets[C]. Enabling Realtime Business Intelligence: 6th International Workshop, BIRTE 2012, Held at the 38th International Conference on Very Large Databases, 2012

[7] Kitchen R., Lauriaul T. P. Small Data in the Era of Big Data[J]. Geo Journal, 2015,80(4):463-475

[8] 欧阳日辉,杜青青. 数据要素定价机制研究进展[J]. 经济学动态, 2022,(2):124-141
Ouyang R. H., Du Q. Q. Research Progress on the Pricing Mechanisms of Data[J]. Economic Perspectives, 2022,(2):124-141

[9] 郭鑫鑫,王海燕. 大数据背景下基于数据众包的健康数据共享平台商业模式构建[J]. 管理评论, 2019,31(7):56-64
Guo X. X., Wang H. Y. Business Model of Health Data Sharing Platform Based on Data Crowdsourcing in the Context of Big Data [J]. Management Review, 2019,31(7):56-64

[10] Gul F. Bargaining Foundations of Shapley Value[J]. Econometrica: Journal of the Econometric Society, 1989:81-95

[11] Ghorbani A., Zou J. Data Shapley: Equitable Valuation of Data for Machine Learning[C]. International Conference on Machine Learning. PMLR, 2019

[12] Chen L., Koutris P., Kumar A. Towards Model-based Pricing for Machine Learning in a Data Marketplace[C]. Proceedings of the 2019 International Conference on Management of Data, 2019

[13] Bertsimas D., Kallus N. From Predictive to Prescriptive Analytics[J]. Management Science, 2020,66(3):1025-1044

[14] Jagabathula S., Subramanian L., Venkataraman A. A Conditional Gradient Approach for Non-parametric Estimation of Mixing Distributions[J]. Management Science, 2020,66(8):3635-3656

[15] Aggarwal V., Hwang E. H., Tan Y. Learning to Be Creative: A Mutually Exciting Spatiotemporal Point Process Model for Idea Generation in Open Innovation[J]. Information System Research, 2021,32(4):1214-1235

[16] Bell J. Machine Learning: Hands-on for Developers and Technical Professionals[M]. Indianapolis: John Wiley & Sons, 2020

[17] 洪永淼,汪寿阳. 大数据、机器学习与统计学:挑战与机遇[J]. 计量经济学报, 2021,1(1):17-35
Hong Y. M., Wang S. Y. Big Data, Machine Learning and Statistics: Challenges and Opportunities[J]. China Journal of Econometrics, 2021,1(1):17-35

[18] 洪永淼,汪寿阳. 大数据如何改变经济学研究范式?[J]. 管理世界, 2021,37(10):40-55
Hong Y. M., Wang S. Y. How Is Big Data Changing Economic Research Paradigms?[J]. Journal of Management World, 2021, 37(10):40-55

[19] 刘景江,郑畅然,洪永淼. 机器学习如何赋能管理学研究?——国内外前沿综述和未来展望[J]. 管理世界, 2023,39(9):191-216
Liu J. J., Zheng C. R., Hong Y. M. How can Machine Learning Empower Management Research? A Domestic-Foreign Frontier Review and Future Prospectsp[J]. Journal of Management World, 2023,39(9):191-216

[20] Kwon Y., Rivas M. A., Zou J. Efficient Computation and Analysis of Distributional Shapley Values[J]. arXiv Preprint arXiv: 2007.01357, 2020
[21] 赵越,徐博涵,王聪,等.基于风电场功率预测的数据价值研究[J].工程管理科技前沿, 2023,42(2):34-42
Zhao Y., Xu B. H., Wang C., et al. Research on Data Value Based on Wind Farm Power Prediction[J]. Frontiers of Science and Technology of Engineering Management, 2023,42(2):34-42
[22] Feng G., Giglio S., Xiu D. Taming the Factor Zoo: A Test of New Factors[J]. Journal of Finance, 2020,75(3):1327-1370
[23] Liu C. H., Nowak A. D., Smith P. S. Asymmetric or Incomplete Information about Asset Values?[J]. Review of Financial Studies, 2020,33(7):2898-2936
[24] 宫晓莉,熊熊,张维. 我国金融机构系统性风险度量与外溢效应研究[J]. 管理世界, 2020,36(8):65-83
Gong X. L., Xiong X., Zhang W. Research on Systemic Risk Measurement and Spillover Effect of Financial Institutions in China [J]. Journal of Management World, 2020,36(8):65-83
[25] 王建东,于施洋,黄倩倩. 数据要素基础理论与制度体系总体设计探究[J]. 电子政务, 2022,(2):2-11
Wang J. D., Yu S. Y., Huang Q. Q. Exploration of Basic Theory of Data Elements and Overall Design of Institutional System[J]. E-Government, 2022,(2):2-11
[26] 杨铭鑫,王建冬,窦悦. 数字经济背景下数据要素参与收入分配的制度进路研究[J]. 电子政务, 2022,(2):31-39
Yang M. X., Wang J. D., Dou Y. Research on the Institutional Progression of Data Factor Participation in Income Distribution in the Context of Digital Economy[J]. E-Government, 2022,(2):31-39
[27] 吴登生,冯钰瑶,张宏云,等. 数据治理现代化的关键支撑[J]. 国家治理, 2023,(13):34-38
Wu D. S., Feng Y. Y., Zhang H. Y., et al. Key Supports for Modernizing Data Governance[J]. Governance, 2023,(13):34-38
[28] 林镇阳,侯智军,赵蓉,等. 数据要素生态系统视角下数据运营平台的服务类型与监管体系构建[J]. 电子政务, 2022,(8):89-99
Lin Z. Y., Hou Z. J., Zhao R., et al. Service Types and Regulatory System Construction of Data Operation Platforms under the Perspective of Data Element Ecosystems[J]. E-Government, 2022,(8):89-99
[29] Lu X., McElroy M. B., Peng W., et al. Challenges Faced by China Compared with the US in Developing Wind Power[J]. Nature Energy, 2016,1(6):1-6
[30] Cassola F., Burlando M. Wind Speed and Wind Energy Forecast through Kalman Filtering of Numerical Weather Prediction Model Output[J]. Applied Energy, 2012,99:154-166
[31] James G., Witten D., Hastie T., et al. An Introduction to Statistical Learning[M]. New York: Springer, 2013
[32] Ke G., Meng Q., Finley T., et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017
[33] Sun X., Liu M., Sima Z. A Novel Cryptocurrency Price Trend Forecasting Model Based on LightGBM[J]. Finance Research Letters, 2020,32:101084
[34] Zhang J., Wang Y., Hug G. Cost-oriented Load Forecasting[J]. Electric Power Systems Research, 2022,205:107723

Research on the Circulation and Revenue Sharing Mechanisms of Data Elements: An Example of Integrating Meteorological Data in Wind Power Scenarios

Wang Yanzhi[1], *Huang Jingsi*[1], *Wang Jianxiao*[2], *Gao Feng*[1] *and Song Jie*[1,2]
(1.College of Engineering, Peking University, Beijing 100871;
2.National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871)

Abstract: In the context of accelerating the construction of a unified national data element market in China, the design of a data transaction and circulation model that suits the country's specific conditions, along with a corresponding revenue sharing mechanism, has not yet been fully explored. This paper starts by setting up a data element transaction model based on the commonly recognized three main parties in the data element market, using meteorological data supply as a typical subdivided industry and the application of data in power forecasting as a typical scenario. The model incorporates a wind power prediction model based on machine learning, depicting the value realization of multi-feature data elements and data services. Furthermore, in designing a revenue sharing mechanism based on data value, this study compares the differences in the main benefits to data producers (data vendors) and market impact among four revenue sharing methods: the average method, leave-one-out method, Shapley Value method, and Penalty-modified Shapley Value method. Lastly, the research demonstrates that the Penalty-modified Shapley Value revenue sharing strategy effectively considers the level of differentiation among data elements in the market, while also identifying and preventing disturbances caused by data element replication.

Key words: data element circulation, data factor market, revenue sharing, machine learning, Shapley value