·专题一:双清论坛“黄河流域生态保护与可持续发展”·

黄河流域人地系统研究的大数据支撑与方法探索

程昌秀[1,2*] 沈 石[1,2] 李强坤[3]

1. 北京师范大学 地表过程与资源生态国家重点实验室,北京 100875
2. 北京师范大学 地理科学学部,北京 100875
3. 黄河水利委员会 黄河水利科学研究院,郑州 450003

[摘 要] 从整体性出发,系统揭示黄河流域人地系统相互作用与耦合机制是实现黄河流域生态保护和高质量发展的关键。随着新时代的到来,地理学应该融合大数据范式,促进人地系统研究的认识革命,驱动人地耦合研究创新。论文系统分析了目前大数据+机器学习方法在人地系统模拟与预测中面临的两类挑战。一是在构建功能良好的地理机器学习模型时,面临地理标记样本贫乏、异质性、数据质量差、地理要素关系候选模式搜索空间巨大、伪相关等问题;二是面对迅速发展并成熟的地理机理模型以及地理复杂性解译需求,机器学习面临方法论困境。为实现黄河流域复杂人地系统的精准模拟与预测,本文提出将经典地理过程理论或建模方法与新的数据科学工具相结合,是大数据时代下人地系统研究的新范式和最佳路径,具体包括:以地理过程理解为基础的机器学习、机器学习与机理模型的耦合、机器学习与仿真的耦合和拓展远程耦合研究新方法。地理理论指导下的数据科学范式有望开启大数据在含有社会科学属性的黄河流域人地系统研究的潜力。

[关键词] 黄河流域;人地系统;大数据;机器学习;机理模型;地理学

1 黄河流域人地系统的耦合与复杂性特征

黄河作为“哺育中华民族、孕育中华文明”的母亲河,是人类活动最频繁、方式最复杂的地区之一。受自然生态本底影响,黄河流域各省区经济发展水平差异较大,上中游以农牧业生产模式为主,经济发展相对滞后,是巩固脱贫攻坚成果的重点区域;下游华北平原土壤肥沃、水热条件好、人口密集、经济发达,但面临水患频繁和水资源短缺的双重危机。近年来,气候变化悄然影响着黄河流域的人地关系,全球化和城市化加速了流域区域间自然与人文要素的联系。新时代黄河流域人地关系可能面临格局的重塑。

黄河流域是典型的人地耦合系统,自然要素变化与人类活动紧密关联,破解陆地表层多要素的耦合关系,是实现黄河流域生态保护和高质量发展的关键环节。黄河流域人地系统是多层嵌套的复杂、开放的人地耦合巨系统;各子系统内进行着地理要素间物质迁移、能量转换、信息传输的内循环;同时,更高层次的外循环推动了子系统间地理要素的关联,如图 1 所示。水土是黄河流域最核心的自然资源,是生物、粮食等衍生资源存在的基础。水是黄河流域人地关系的核心驱动要素,三水(大气水、地表水和地下水)平衡是流域不同尺度、不同区域需要共同关注的核心科学问题。不同地理环境下,水土资

程昌秀 北京师范大学地理科学部教授,国家自然科学基金委员会优秀青年科学基金项目获得者(2012)。主要从事地理数据组织管理与分析挖掘方法研究,重点应用和服务于自然灾害损失与风险评估、国土空间管理与优化等领域。主持国家自然科学基金、重点研发计划课题等多项国家级科研任务,发表科研论文百余篇,主持编制国家标准 2 部,获批国家发明专利 8 项。获茅以升科学技术奖:北京青年科技奖(2017);获高校 GIS 创新人物奖(2020)。

收稿日期:2021-02-17;修回日期:2021-04-22
* 通信作者,Email: chengcx@bnu.edu.cn
本文受到国家自然科学基金项目(42041007-03)的资助。

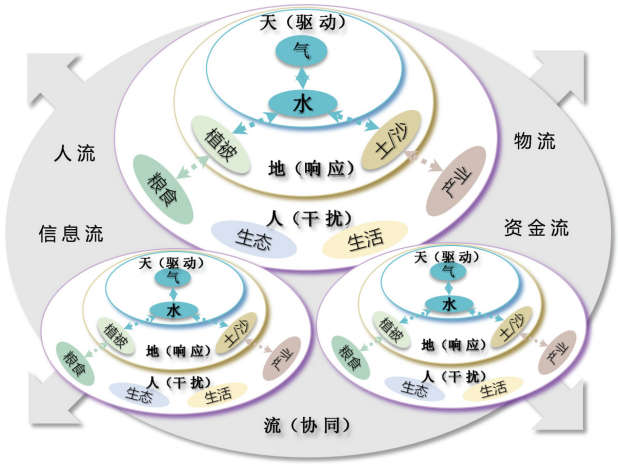

图 1 黄河流域人地系统-多层次嵌套、循环、开放

注：天—驱动系统、地—响应系统、人—干扰系统、流—协同系统

源可能表现出正效应（水源涵养），也可能表现出负效应（水土流失），因此水土资源适配是黄河流域人地协同、区域高质量发展的物质基础。在不同水土适配方案下，叠加上适宜的人类活动，则构成优化的“生态—生产—生活”配置空间，初步形成局地适宜的天地人和子系统。新时代不断增加的人流、物流、信息流、资金流加速了子系统间的自然与人文要素的相互作用与流动，局地人地系统作为子系统，将嵌套入到更大的系统中；局地的自然人文要素也将参与到更大尺度的循环中。

鉴于黄河流域人地系统表现出的多层、嵌套、复杂、开放等特征，需要建立系统思维、从“整体论”与“还原论”相结合的角度出发，运用大数据和机器学习等新时代科学研究范式，揭示黄河流域自然环境与人文社会要素、结构和功能的相互联系、级联特征[1]，探索远距离人地耦合关系[2]及其对生态环境、社会经济的影响；认识系统自组织、非线性、涌现、混沌等复杂特征[3]；实现对黄河流域复杂人地系统的准确模拟和科学预估；为提高黄河流域保护和治理的系统性、整体性和协同性提供科学基础。

2 大数据为人地系统研究提供机遇

地理科学经久不衰、蓬勃发展的根源在于它善于把握计量革命、地理信息系统、环境遥感等机遇促进自身变革。这些变革的表象在于方法或技术，但本质是影响深远的思想革命。随着新时代的到来，地理学应该融合大数据时代的科学范式[3,4]，促进人地系统研究的认识革命。

本文的“大数据”通常强调数据的 5V 特征及以机器学习为代表的大数据挖掘方法。“大数据”指存在非结构化、异构特征的大流量数据，因此具有体积（Volume）大、速度（Velocity）快和种类（Variety）多的 3V 特征；同时，大数据的噪声和不确定性使其增加了准确性（Veracity）低的第 4V 特征；而第 5V（Visualization）意味着需要以有效的可视化方式呈现大数据复杂而丰富的结构或信息[5]。机器学习是大数据分析常用的方法，涵盖了从简单的线性回归到复杂非线性的深度学习[6]，后续更关注深度学习的相关内容。

“大数据”为人地系统研究带来新思路。随着定位自动观测、对地观测和网络技术的迅速发展，地理数据的可持续获取能力大幅提升，数据量呈指数增长，加之公民科学观测的补充数据，构建了表征自然过程与人类活动的多类型、高密度、大范围的数据支撑体系，地理学研究从数据稀疏时代进入数据稠密时代。地理大数据为人地系统的过程与演化研究提供了坚实的数据基础，为人地系统的新内涵（例如层级、结构、物质/能量关系以及系统的复杂性等）提供丰富的原料。

“大数据＋机器学习”为人地系统研究创造新路径。近几十年来，在全球变化等重大环境问题和科学决策需求的推动下，地理系统模型发展迅速，但目前相关模型偏重自然过程，对人文社会过程以及系统的复杂性刻画不足，难以准确模拟和预测复杂的黄河流域人地系统[6]。大数据极大丰富了地理研究对人文社会要素的感知能力，一系列线性或非线性机器学习方法为人地系统的模拟和预测创造了新路径[3]。

3 大数据分析面临地理问题的挑战

大数据＋机器学习成功解决了商业领域的一些问题，例如，图书推荐、用户偏好等。大数据在商业领域的成功应用使人们猜测：大数据推理有望取代模型推演、计算机仿真等范式，成为科学的第四范式——数据密集型发现。事实上，目前以机器学习为代表的大数据分析方法，仅适用于弱人工智能范畴的“小问题”[7]，该文献中认为“小问题”具备如下三个必要条件：（1）对于感兴趣的问题，系统是可预测的；（2）为模型的初始训练提供了充足的样本数据；（3）有充足的新数据、且可根据观测结果定期评估，必要时调整模型。“小问题”并不意味着问题的重要性低（例如个性化医学）或复杂度小（例如语音识别），也不代表问题易于解决；仅代表它们具有目

标明确、范围狭窄、成功程度明确、预测出错的影响小、可反复预测、且可以根据新观测结果定期评价和改进预测模型的特点。这些特点与弱人工智能范畴的问题密切相关。

近年在大数据和高性能计算的支撑下，机器学习在解决地理学“小问题”方面取得了一些成就，例如地物识别与遥感图形分类[8]、城市内涝监测[9]、基于街景预测社区人口或环境[10, 11]等；但当遇到以人地关系为核心的地理问题时，由于多数问题不具备“小问题”的必要条件，以机器学习为代表的大数据方法面临挑战。为了构建功能良好的地理机器学习模型，在大数据收集、准备和存储、建模和计算等方面面临下述 1～3 的挑战；为了拓展超越“小问题”的科学研究范畴，大数据分析面临下述 4～5 的挑战。

挑战 1　地理现象或事件探测中标记样本的缺乏，在一定程度上造成机器学习的结果缺少精细的分类标准

人地系统中有些地理对象或事件（例如旋风、湍流、涡旋等）的探测非常重要，因为它们对系统中物质能量传输起着至关重要的作用。传统研究多采用人工解译的方法识别[12]。鉴于该类现象通常具有特定的时空模式，可用机器学习中模式挖掘技术实现相关对象或事件的自动探测。

然而，与社交网络的用户或零售商店的产品不同，多数地理对象或事件通常缺乏简明、清晰的定义，且通常在连续时空场中以无定形的边界出现，甚至有时还是高度动态的（例如洪水、森林火灾），这为标记训练样本带来困难。此外，人地系统关心的一些重大地理现象或事件（例如地震、台风等）属于很少发生的罕见现象或事件，使样本的标记和探测变得更加困难，主要包括：训练样本的积累时间长、分析的问题过于广泛、无法等待观测数据来测试预测结果、预测成功度指标不明确、难以进行预测后的调整等。地理现象或事件的上述特征严重制约了机器学习方法的预测性能。

挑战 2　地理现象的异质性和自相关性等特征，导致训练样本、分析方法等难以满足地理研究的需求，制约了地理变量数据推断的准确性

地理变量的数据推断，即利用已知观测数据或模拟变量数据，推断难以直接观测的地理变量的数值，是人地关系研究中的重要议题。准确推断地理变量为人地系统变化的预测和区域发展政策的制定提供辅助依据[13]。例如，森林覆盖率、植被健康、水质和地表水可用性、人类活动等地理变量在美丽中国和黄河流域高质量发展的研究中具有重要作用[14, 15]。

采用机器学习推断地理变量的挑战之一是：地理现象在时间、空间中呈现复杂的异质性。然而，导致这种异质性的地理过程或机制往往不为人所知，研究中通常简单归因于地形、土地覆盖、气候、季节、人类活动等因素的作用。异质性要求每个同质分区学习不同的模型，这进一步加剧了标记样本的贫乏。异质性、标记样本的缺乏、罕见地理现象或事件等综合效应，使得普通的机器学习算法难以达到良好的预测性能。

另一个挑战是：地理现象在时间、空间中的自相关性。这种自相关性可能增加虚假估计的风险。例如，云和气溶胶常常导致遥感数据的噪声和缺失值具有特殊结构特征（例如时空自相关），这与常见的随机椒盐噪声假设不同，因此常规的噪声和缺失值处理方法（例如马尔科夫随机场）不能直接应用于此类地理数据[7]。

挑战 3　大数据时代下地理要素关系候选模式的搜索空间巨大给数据挖掘和探测带来困难，自然界中常见的伪相关问题可能产生错误的因果关系或得到因果倒置的结论

以空间格局、时间过程以及要素相互作用为代表的人地关系研究，是人地系统的核心问题。例如，东太平洋海面温度的周期性变化（即厄尔尼诺—南方涛动）及其对洪水、干旱和森林火灾等若干陆表事件的影响。从地理数据中识别这种关系可以帮助我们捕获人地系统的重要信号，同时增进我们对人地系统过程的理解。

但是大数据时代地理数据的易得性和高空间分辨率，以及日益增加的人流、物流、信息流和资金流不断扩大着要素关联的影响区域。频繁、多样的人类活动也导致要素关联的复杂性持续增加。上述效应直接使得人地要素关联的候选模式的搜索空间呈非线性增长，巨大的搜索空间是机器学习在地理数据人地关系探测与挖掘中面临的重要挑战。以东太平洋气温上升对加利福尼亚州森林火灾影响的远程连接关系为例，即使采用 2.5°的粗空间分辨率数据，需要探测和挖掘的两地空间位置对也数以千计[16]。

此外，自然界常存在的“伪相关”现象使得重视相关关系的机器学习可能会错误解译因果无关或得到因果倒置的结论。因此，迫切需要引入和发展面向地理要素的因果归因方法，以更好解译人地关系相互作用的物理路径或机制。

挑战 4 如何与机理模型耦合实现复杂人地系统的精准模拟与预测，是机器学习面临的挑战

随着“人类世”的到来，人类活动日益成为人地系统演化的主导因素。地球外营力、地球内营力和人类活动的混合叠加影响，是人地系统耦合研究的重要内容。大数据在人文社会感知方面具有优势，而机器学习善于在因素众多、关系复杂、内部机制鲜为人知的数据中总结规律[17]，两者结合为人地耦合研究提供机遇。但是机器学习的“黑箱模型”难以显式刻画人地关系的内部机理以及物质流的传递规律，无法反映过程内部的运动机制，而且不能适应有不可测量的输入。在过去几十年，偏重自然过程的地理系统机理模型经历了从单要素到多要素、从静态到动态、从单点到区域和全球尺度模拟的发展历程，取得了较大成绩，但对人类活动和人类社会的刻画不足[6]。

大数据时代下，如何将善于刻画人类活动和社会经济的黑箱机器学习模型与善于刻画自然物质流传递机制的白箱机理模型耦合起来，系统揭示自然环境与人类社会之间的相互作用，实现复杂人地耦合系统的精准模拟与预测，是机器学习在人地系统研究中面临的挑战。此外，从系统科学观点来看，机理模型侧重于还原论，而机器学习侧重于整体论，机理模型与机器学习的有机结合也是系统科学方法论的根本。

挑战 5 如何充分融合大数据、复杂性研究方法是机器学习破解人地关系的重要挑战

人地系统作为动态、开放、复杂的巨系统，除在某些尺度上具备线性、均衡、简单的特征外，在更大范围表现出自组织、非线性、涌现、混沌等复杂性特征。复杂性对深刻认识人地系统关联，准确模拟预测人地系统演化有重要意义。复杂性研究通常可以发现复杂系统演化背后的一只“看不见的手”，例如，气雾栽培植物气生根形成是生物自组织性的表现，地震、滑坡等是在陆表系统达到某临界状态后产生的非线性变化，土壤湿度的长程记忆性[18]，城市的时空演化过程也存在自组织临界特征[19]，爱德华·洛伦兹发现的蝴蝶效应，股市上的羊群效应[20]等。人地系统中复杂性无处不在。英国著名物理学家霍金称“21 世纪将是复杂性科学的时代”。

大数据时代为认识人地系统复杂性研究提供数据基础[3]，复杂网络、统计物理为人地系统的复杂性提供方法基础。如何充分运用大数据及复杂性分析方法发现人地系统演化背后那只“看不见的手”，实现复杂人地系统的精准模拟与预测，是大数据时代人地系统研究的重大挑战和时代使命。

4 大数据时代下人地耦合系统的研究路径

不同尺度地理现象发生与演变规律的理解对确保预测和模型结果的可靠性至关重要。大数据有助于消除已有认知和资源的局限性。“地理过程与大数据机器学习相结合”是大数据时代下人地耦合系统的新范式和最佳路径[13, 21]，对于推动复杂人地系统的精确模拟及预测有重要作用。

4.1 路径 1：以地理过程理解为基础的机器学习

地理过程逻辑的准确理解是机器学习用于人地系统研究的核心，即充分利用地理现象或事件的先验知识和数据基础，标定高质量的训练数据，遴选合适的机器学习方法，构建功能良好的机器学习模型，结合地理理论（机理）知识解译机器学习的预测结果。

近十年来，针对挑战 1～3 提出了系列改进的机器学习算法。针对地理样本标记困难的问题（挑战 1），需要开发新的模式挖掘方法，以解释对象和事件的空间和时间属性；例如，采用空间一致性和时间持久性的方法处理无定形边界问题[22, 23]。为了解决异质性加剧样本稀缺的问题（挑战 2），将多任务框架中各种林地的森林覆盖度估算视作独立任务[24]，相似林地共享相关任务的机器学习模型，实现了异质性区域森林覆盖度的估算。为了解决气候数据的非平稳性（挑战 2），可用在线学习算法，结合多个气候模式输出结果生成更鲁棒的气象数据，从而更有效地捕捉气候数据的非平稳性[25]。针对地理大数据关系候选模式搜索空间大的问题（挑战 3），应发展探测地理数据相互作用新方法。例如，采用复杂网络的方法[26]，揭示 0.25°空间分辨率下的极端降水遥相关的全球模式；面向高维、稀疏数据提出三向聚类[27]，用于挖掘“空间-时间-属性-尺度”的联合作用模式。为了剔除伪相关，可以借鉴统计领域的系列因果归因分析方法。目前基于时间序列的格兰杰检验[28]、传递熵[29]、交叉收敛映射[30]等在地理科学领域已有应用。过去十年，VAR/LASSO 和 Pearl 多变量因果关系归因在生物学和医学方面取得了巨大突破，但地理学领域关注不足，此外近期提出的 PCMCI 方法值得关注[31]。这些因果归因分析方法在人地关系的探测与挖掘中具有巨大应用潜力。

针对训练样本少的问题，相关机理模型的输入和输出可作为标记样本参与机器学习模型的训练。

这不仅丰富了训练样本，更重要的是，机器学习模型可以学习到机理模型的相关知识，同时观测数据的加入还可纠正机理模型的偏差。例如，地表能量平衡约束下的机器学习模拟结果更符合蒸散发量的物理规律和实地观测[32]。

4.2　路径 2：机器学习与机理模型的耦合

地理理论与数据科学结合的另一种形式是机器学习与机理模型的耦合。机理模型属于理论驱动，机器学习属于数据驱动。事实上，两类方法相辅相成：机理模型基于事实原理可直接解释，并且提供超出观测条件之外的推断潜力；而机器学习在适应数据方面更加灵活，能够发现地理领域知识以外的一些模式。近数十年来，机理模型作为地理科学研究的基本方法之一，取得了十分显著的进展。但是，由于机理模型自身基础理论和框架的差异、模型参数偏差和观测数据误差等原因，不同机理模型的模拟结果差异较大，导致陆地表层系统中一些关键过程（如碳循环、水循环等）的模拟和预测还存在很大的不确定性[33]。此外，目前地理系统模型仍然偏重自然过程，对人类活动和社会经济过程刻画不足，难以准确模拟和预测复杂人地系统。机理模型和机器模型的耦合，是实现人地系统精确预测的关键途径之一。

Reichstein 等从系统建模角度总结了机理模型与机器学习从浅到深的五种潜在耦合模式[34]。

方式一：机器学习改进机理模型中的参数

参数估计对机理模型的模拟预测能力的提升至关重要，但有些参数难以从第一性原理中直接得到；可以考虑采用机器学习得到相关参数，即得到对观测数据或高分辨率模型数据的最佳描述。例如，陆地表面模型常需要将植被参数分配给不同的植物功能类型，然而根据第一性原理难以直接分配，机器学习可以从合适的统计协变量集中学习这些参数，从而使模型具有更好的动态性、依赖性和关联性。机器学习可以从根本上改善地球系统模式的参数优化方案[35]，改进陆表模型的不确定性估计[36, 37]。

方式二：机器学习取代半经验性的子模型

机理模型中可能存在一些半经验性的子模型（例如生物过程），其表现形式几乎没有理论基础。若有足够的观测数据，该子模型可以考虑用机器学习模型取代。替代后，模型同时兼具物理建模（理论基础、可解释性高）和机器学习（数据自适应性）的优点。例如，可以将成熟的植物水运移微分扩散（机理）方程与根据观测数据（机器）学习到的水分传导的生物调节模型耦合，该模型既符合物质能量运移规律，对生物过程也有良好的适应性。此外，一些大气科学家已开始试验相关方法，以规避大气对流物理参数化中长期存在的偏差[38, 39]。

方式三：机器学习分析机理模型与观测的不一致性，提高机理模型精度

假定没有观测误差，机理模型与观测值的偏差通常视为因认知不完善导致的模型错误。机器学习可以自动从观测数据中挖掘多种模式（甚至包括机理模型尚未显式表达的模式），分析机理模型与观测值的不一致性，有助于识别、可视化以及理解模型误差的模式，从而修正模型误差、改进机理模型（理论）。例如，Mcgovern 等人[40] 和 Vandal[41] 等人均用机器学习纠正了动态变量的模型偏差。与人为设计方法相比，这种方法可缩小模型误差范围，提升模型模拟的空间分辨率。

方式四：机器学习模型分解模型误差和不确定性

在多个子模型构成的复杂机理模型中，尤其是当子模型高度关联或耦合时，模型误差和不确定性来源的甄别是科学难题；借助机器学习可以对机理模型进行解耦，从而分离、解析模型误差和不确定性的来源与扩散方式。在离线模拟环境下，可用机器学习模型代替一个可能有偏的机理子模型，并将机器学习结果作为下一个子模型的输入驱动模型运行；通过对比分析两种不同输入数据情况下被驱动子模型的输出结果，有助于区分机理子模型的错误与其他耦合子模型的错误，从而简化模型参数校准或观测系统状态变量同化中的偏差和不确定性。例如，人工神经网络（ANN）能够很好地模拟欧洲中期气象中心数值天气预测模型的主要误差，同时借助 ANN 模型可以将弱约束模型的误差修正方法，扩展到整个大气柱的数值模型[42]。

方式五：机器学习代理机理模型，提升模型计算效率

对于计算复杂度大、代价高（算力高、耗时长）的机理模型，可考虑用机器学习作为机理模型的代理模型，提升模型的计算效率。经过训练的机器学习模型通常在不牺牲太多精度的情况下，可以实现快几个数量级的模拟[43]。高效的机器学习模型便于提高地理模型的灵敏度分析、模型参数校准以及估计置信区间推导等研究工作的效率。例如，用机器学习取代了计算代价高的辐射—植被—大气相互作用的辐射传输机理模型，提升遥感数据在陆表模型

中的解释和同化效率[44, 45]。此外，机器学习也可代理计算复杂度高的动态模型。例如，将机器学习用于气候动态建模[46]，探索机器学习在植被动态模拟中的应用[47]；用机器学习模型实现海气交换与海洋生物地球化学过程的耦合[48]；利用机器学习实现河口三角洲每日洪泛平原的高分辨率模拟[49]；在无需地下参数的情况下，借助机器学习实现地下水文演变的高精度模拟[50]，模拟结果与机理模型相当。

4.3 途径 3：机器学习与仿真的耦合

机器学习与仿真的耦合可以借鉴上节的系列技术方案，但本节侧重强调相关研究应该重点关注以耗散结构理论、协同理论、突变理论、超循环理论、分形理论、复杂自适应系统理论为基础，以机器学习与多主体仿真模拟为研究手段，探索人地系统的复杂性及其呈现的增益效应，实现复杂人地系统的精准模拟与预测。

在人地系统仿真的研究中，仿真模型与工具已经相对成熟，但研究通常很难准确地总结出系统中个体的行为规则或相关参数，致使仿真结果显得苍白[3]。大数据时代可以通过机器学习准确挖掘个体的行为规则或相关参数，尤其是复杂性特征，对提高人地系统模拟和推演的准确性意义重大。

此外鉴于机理模型的复杂性，仿真是很好的测试床，可以探测机器学习、深度学习系列方法的潜力，同时在训练条件范围之外进行外推。

4.4 途径 4：拓展远程耦合研究的新方法

交通和信息通讯技术的迅速发展，区域间人流、物流、信息流、资金流不断加强，增进了人地耦合系统之间社会经济要素与自然环境要素的远距离相互作用；日益加速的各种“流”作为一种重要的驱动力，深刻地重塑着开放系统中人地关系的格局。

新时代下，发展远程耦合框架下的人地关系网络系统理论创新，发展并完善远程耦合研究的方法和工具集，挖掘或探测人类与自然之间远程的时空滞后关系、因果关系、溢出效应以及解译相应的远程耦合的驱动机制，对于提升跨区域人地耦合的洞察力，推动地理学研究框架的适应性创新和变革有重大意义。

5 结 语

人地系统是综合理解陆地表层系统及其演变规律的基石。理解黄河流域人地耦合系统是实现黄河流域生态环境保护与高质量发展的关键。随着新时代的到来，地理学应该融合大数据的研究范式，促进人地系统研究的认识革命。当面向弱人工智能的大数据方法遇到非“小问题”范畴的人地关系问题时，面临两类挑战。一类是在构建功能良好的地理机器学习模型时，存在地理标记样本缺乏、异质性、数据质量差、地理要素关系候选模式搜索空间巨大、伪相关等问题（挑战 1～3）；另一类是面对迅速发展并成熟的地理系统机理模型以及地理复杂性需求时，机器学习存在方法论困境（挑战 4 和挑战 5）。

理论指导的数据科学是一个新兴的研究范式。将地理学经典理论或方法与新的数据科学工具相结合，是大数据时代下人地系统的新范式和最佳路径。为实现复杂人地系统的精准模拟与预测，针对挑战 1～3，论文提出以地理过程理解为基础的机器学习路径，并总结和展示了近十年相关研究的努力方向和成果；针对挑战 4，论文提出机器学习与机理模型耦合的路径，给出了 5 种可能的耦合方式；针对挑战 5，论文提出机器学习与仿真耦合的路径，重点强调应关注人地系统的复杂性及其呈现的增益效应。理论指导数据科学的范式有望开启地理大数据在含有社会科学属性的人地系统研究的潜力。

机器学习与机理模型的耦合是目前国内外前沿和热点。黄河流域人地系统是全球变化背景下最复杂、最脆弱的系统之一。黄河流域大数据平台的建设将为采用机器学习与机理模型耦合的方式开展人地系统的研究提供契机。论文总结的大数据相关支撑与方法，有助于开启地理大数据在黄河流域人地系统研究的潜力。

致谢 成文过程中得到宋长青教授的指导和帮助，在此表示感谢。

参 考 文 献

[1] Fu B, Wei Y. Editorial overview: keeping fit in the dynamics of coupled natural and human systems. Current Opinion in Environmental Sustainability, 2018, 33: 87091.

[2] 刘建国, Hull V, Batistella M, 等. 远程耦合世界的可持续性框架. 生态学报, 2016, 36(23): 7870—7885.

[3] 程昌秀, 史培军, 宋长青, 等. 地理大数据为地理复杂性研究提供新机遇. 地理学报, 2018, 73(8): 1397—1406.

[4] 熊巨华, 王佳, 张晴, 等. 地理科学的学科体系构建与内涵. 科学通报, 2021, 66(2): 153—161.

[5] Nativi S, Mazzetti P, Santoro M. Big data challenges in building the global earth observation system of systems. Environmental Modelling & Software, 2015, 68: 1—26.

[6] 彭书时, 朴世龙, 于家烁, 等. 地理系统模型研究进展. 地理科学进展, 2018, 37(1): 109—120.

[7] Knusel B, Zumwald M, Baumberger C, et al. Applying big data beyond small problems in climate research. Nature Climate Change, 2019, 9: 196—202.

[8] Yang L, Alan ME, Prasenjit M, et al. Visually-enabled active deep learning for (Geo) text and image classification: a review. ISPRS International Journal of Geo-Information, 2018, 7(2): 65.

[9] Jiang J, Qin C, Yu J. Obtaining urban waterlogging depths from video images using synthetic imagedata. Remote Sensing, 2020, 12(6): 1014.

[10] Gebru T, Krause J, Wang Y, et al. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(50): 13108—13113.

[11] 张丽英，裴韬，陈宜金，等. 基于街景图像的城市环境评价研究综述. 地球信息科学学报，2019，21(1)：46—58.

[12] Chelton DB, Schlax MG, Samelson RM, et al. Global observations of large oceanic eddies. Geophysical Research Letters, 2017, 34(15): L15606.

[13] Karpatne A, Ebert-Uphoff I, Ravela S, et al. Machine learning for the geosciences: challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(8): 1544—1554.

[14] 方创琳，王振波，刘海猛. 美丽中国建设的理论基础与评估方案探索. 地理学报，2019，74(4)：619—632.

[15] 崔盼盼，赵媛，夏四友，等. 黄河流域生态环境与高质量发展测度及时空耦合特征. 经济地理，2020,40(5)：19—57.

[16] Siegert F, Ruecker G, Hinrichs A, et al. Increased damage from fires in logged forests during droughts caused by El nino. Nature, 2001, 414(6862): 437—440.

[17] Zhang T, Shen S, Cheng C, et al. A topic model based framework for identifying the distribution of demand for relief supplies using social media data. International Journal of Geographical Information Science, 2021(10): 1—22.

[18] Shen S, Ye S, Cheng C, et al. Persistence and corresponding timescales of soil moisture dynamics during summer in the babao river basin, northwest China. Journal of Geophysical Research: Atmospheres, 2018, 123(17): 8936—8948.

[19] Li R, Dong L, Zhang J, et al. Simple spatial scaling rules behind complex cities. Nature Communications, 2017, 8(1): 1—7.

[20] Hou Y, Liu F, Gao J, et al. Characterizing complexity changes in Chinese stock markets by permutation entropy. Entropy, 2017, 19(10): 514.

[21] Karpatne A, Atluri G, Faghmous J, et al. Theory guided data science: a new paradigm for scientific discovery. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(10): 2318—2331.

[22] Deo RC, Şahin M. Application of the extreme learning machine algorithm for the prediction of monthly effective drought index in eastern Australia. Atmospheric Research, 2015, 153: 512—525.

[23] Tapia C, Abajo B, Feliu E, et al. Profiling urban vulnerabilities to climate change: an indicator-based vulnerability assessment for european cities. Ecological Indicators, 2017, 78: 142—155.

[24] Karpatne A, Khandelwal A, Boriah S, et al. Predictive learning in the presence of heterogeneity and limited training data. In Proceedings of the 2014 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2014: 253—261.

[25] Monteleoni C, Schmidt GA, Saroha S, et al. Tracking climate models. Statistical Analysis and Data Mining, 2011, 4(4): 372—392.

[26] Boers N, Goswami B, Rheinwalt A, et al. Complex networks reveal global pattern of extreme-rainfall teleconnections. Nature, 2019, 566: 373—377.

[27] 程昌秀，宋长青，吴晓静，等. 地理时空三向聚类分析方法的构建与实践. 地理学报，2020，75(5)：904—916.

[28] Granger CWJ. Testing for causality: a personal viewpoint. Journal of Economic Dynamics & Control, 1980, 2: 329—352.

[29] Schreiber T. Measuring information transfer. Physical Review Letters, 2000, 85(2): 461—464.

[30] Sugihara G, May R, Ye H, et al. Detecting causality in complex ecosystems. Science, 2012, 338(6106): 496.

[31] Runge J, Kretschmer M, Flaxman S, et al. Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 2019, 5(11): eaau4996.

[32] Zhao WL, Gentine P, Reichstein M, et al. Physics-constrained machine learning of evapotranspiration. Geophysical Research Letters, 2019, 46(24): 14496—14507.

[33] 方精云，朱江玲，王少鹏，等. 全球变暖、碳排放及不确定性. 中国科学：地球科学，2011，41(10)：1385—1395.

[34] Reichstein M, Camps-Valls G, Stevens B, et al. Deep learning and process understanding for data-driven earth system science. Nautre, 2019, 566: 159—204.

[35] Schneider T, Lan S, Stuart A, et al. Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations. Geophysical Research Letters, 2017, 44(24): 12396—12417.

[36] Chaney NW, Herman JD, Ek MB, et al. Deriving global parameter estimates for the Noah Land surface model using FLUXNET and machine learning. Journal of Geophysical Research Atmospheres, 2016, 121(22): 13218—13235.

[37] Sawada Y. Machine learning accelerates parameter optimization and uncertainty assessment of a land surface model. Journal of Geophysical Research: Atmospheres, 2020, 125(20): e2020JD032688.

[38] Gentine P, Pritchard M, Rasp S, et al. Could machine learning break the convection parameterization deadlock?. Geophysical Research Letters, 2018, 45(11): 5742—5751.

[39] Brenowitz ND, Bretherton CS. Prognostic validation of a neural network unified physics parameterization. Geophysical Research Letters, 2018, 45(12): 6289—6298.

[40] Mcgovern A, Elmore KL, Gagne D, et al. Using artificial intelligence to improve real-time decision-making for high-impact weather. Bulletin of the American Meteorological Society, 2017, 98(10): 2073—2090.

[41] Vandal T, Kodra E, Ganguly S, et al. Generating high resolution climate change projections through single image super-resolution: an abridged version. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), 2018.

[42] Bonavita M, Laloyaux P. Machine learning for model error inference and correction. Journal of Advances in Modeling Earth Systems, 2020, 12: e2020MS002232.
[43] Arcomano T, Szunyogh I, Pathak J, et al. A machine learning-based global atmospheric forecast model. Geophysical Research Letters, 2020, 47(9): e2020GL087776.
[44] Camps-Valls G, Martino L, Svendsen DH, et al. Physics-aware gaussian processes in remote sensing. Applied Soft Computing, 2018, 68: 69—82.
[45] Jochem V, Neus S, Juan R, et al. Emulation of leaf, canopy and atmosphere radiative transfer models for fast global sensitivity analysis. Remote Sensing, 2016, 8(8): 673.
[46] Castruccio S, Mcinerney DJ, Stein ML, et al. Statistical emulation of climate model projections based on precomputed GCM runs. Journal of Climate, 2014, 27(5): 1829—1844.
[47] Fer I, Kelly R, Moorcroft PR, et al. Linking big models to big data: efficient ecosystem model calibration through bayesian model emulation. Biogeosciences Discussions, 2018, 15(19): 5801—5830.
[48] Wang S, Kinnison D, Montzka SA, et al. Ocean biogeochemistry control on the marine emissions of brominated very short-lived ozone-depleting substances: a machine-learning approach. Journal of Geophysical Research: Atmospheres, 2019, 124(22): 12319—12339.
[49] Karimi SS, Saintilan N, Wen L, et al. Application of machine learning to model wetland inundation patterns across a large semiarid floodplain. Water Resources Research, 2019, 55(11): 8765—8778.
[50] Sahoo S, Russo TA, Elliott J, et al. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US. Water Resources Research, 2017, 53(5): 3878—3895.

Big Data Support and Method Exploration About Natural and Human Systems Research in the Yellow River Basin

Cheng Changxiu[1,2*] Shen Shi[1,2] Li Qiangkun[3]

1. *State Key Laboratory of Earth Surface Processes and Resource Ecology*, *Beijing Normal University*, *Beijing* 100875
2. *Faculty of Geographical Science*, *Beijing Normal University*, *Beijing* 100875
3. *Institute Yellow River of Hydraulic Research*, *Yellow River Conservancy Commission*, *Zhengzhou* 450003

Abstract The key science problem to realize ecological protection and high-quality development of the Yellow River basin is to discover the interaction and coupling of the nature-human system in the basin. With the advent of a new era, Geography should incorporate with a data-driven paradigm to promote research innovation and recognize revolution of the nature-human system. This paper analyzes two kinds of challenges when big data and machine learning faced to natural and human system simulation and prediction. One is that when building a well-functioning geo-machine learning model, it is faced with several problems, such as lack of geographical sample markers, heterogeneity, poor data quality, large search space for geo-feature relationship candidate patterns, and pseudo-correlation. The other is that machine learning is faced with the methodological dilemmas when machine learning face the rapidly developing and matured geo-system mechanic models and the interpretation of geographical complexity. This paper proposes that combining classical geographic process theory (modeling) with new big-data method maybe the new paradigm or the best path to realize the precise simulation and prediction of natural and human system. The pathes include the geography-theories-based machine learning, coupling of machine learning and mechanism models, coupling of machine learning and simulation, and new theory of telecoupled. The data science paradigm under the guidance of geographic theories is expected to open its potential on using geographic big data to study nature-human systems, with some social science property, in Yellow River basin.

Keywords Yellow River basin; natural and human systems; big data; machine learning; mechanism models; geography

(责任编辑 张 强)

* Corresponding Author, Email: chengcx@bnu.edu.cn