doi: 10.19388/j.zgdzdc.2021.02.11
引用格式: 汤宇磊,吴杨杨,蒋兴征,等.面向自然资源信息提取的多源异构数据融合技术——以汉江流域NDVI数据为例[J].中国地质调查,2021,8(2): 74-82. (Tang Y L, Wu Y Y, Jiang X Z, et al. Multi-source heterogeneous data fusion technology for natural resource information extraction: A case study of NDVI data in Hanjiang Basin[J]. Geological Survey of China, 2021, 8(2): 74-82.)

面向自然资源信息提取的多源异构数据融合技术
——以汉江流域NDVI数据为例

汤宇磊[1,2], 吴杨杨[3], 蒋兴征[1], 冯 亮[1], 高 阳[4]
(1. 中国地质调查局地球物理调查中心,河北 廊坊 065000; 2. 自然资源要素耦合过程与效益重点实验室,北京 100055; 3. 四川大学建筑与环境学院,四川 成都 610065; 4. 中国农业大学土地科学与技术学院,北京 100083)

摘要: 高时空分辨率的自然资源指标数据对大尺度自然资源动态观测与趋势评估至关重要。大数据时代下的海量多源数据为数据高效融合利用提供了可能。以重构汉江流域归一化植被指数(Normalized Difference Vegetation Index, NDVI)数据为例,搭建了PostgreSQL自然资源时空大数据处理底层架构,集成了数据级融合法、特征级融合法和决策级融合法,基于机器学习算法构建了一套面向自然资源信息提取的多源异构数据智能融合技术,实现了多源数据的高效利用与特征空间优选。同时,重构了2000—2019年汉江流域NDVI 1 km逐年数据集,全面反映了汉江流域植被动态变化。研究结果可为地球科学时空大数据的高效提取与模拟分析提供科学参考,为定量核算林草资源禀赋规模、探究生态系统时空演变规律提供一种更精准、更便捷的技术手段。

关键词: 自然资源; 多源异构; 时空数据融合; 机器学习
中图分类号: P96; TP391; TP75 **文献标志码**: A **文章编号**: 2095-8706(2021)02-0074-09

0 引言

进入信息时代以来,人类对地球的观测与探测能力不断提升,获取的数据量成幂律增长,数据处理技术的不断丰富为数据融合利用提供了可能。各类自然资源时空属性信息充实于大量非关系型、非结构化和半结构化数据中,具有典型的多源、多维、多类、多尺度等特征[1]。已有研究表明,多传感器数据融合相较于单一来源数据在数据准确性和实际应用方面更具优势[2]。欧美等国基于不同卫星传感器,相继发布了各类归一化植被指数(Normalized Difference Vegetation Index, NDVI)遥感数据产品,在生态恢复工程评价[3]、林草资源监测[4]、生物多样性估算[5]、高分辨率森林覆盖分类[6]等诸多方面发挥了重要作用。但NDVI数据源的多源性同时也带来了植被评估的不确定性[7],一定程度上限制了遥感数据的价值挖掘及植被演变研究的延续性和准确性。不同流域之间植被种类与分布存在较大差异,NDVI反演参数差异较大,难以依据单一产品客观评估区域植被生长水平[8],迫切需要针对各类时空数据开展规则化重建、数学建模等工作,实现多源异构自然资源信息的融汇和海量观测数据的高效利用。汉江流域是我国南水北调工程的水源地,也是长江中游生态保护屏障区,国内外学者针对流域生态服务[9]、水文效应[10]、湿地变化[11]等开展了大量研究,但基于多源数据的资源-生态评估工作有待进一步深入。本文以汉江流域植被覆盖为研究案例,探索了一种基于数据规则化重构与机器学习算法的多源异构数据融合技术,有

收稿日期: 2021-02-26; **修订日期**: 2021-03-16。
基金项目: 中国地质调查局"汉江流域丹江口—钟祥段自然资源多要素综合调查(编号: DD20208024)"项目资助。
第一作者简介: 汤宇磊(1990—),男,工程师,主要从事环境大数据与机器学习应用研究。Email: tangyl_eco@gmail.com。

效融合了各类数据信息,获得了多年期高分辨率自然资源观测指标时空数据集,实现了地表植被演变的精准评估,进而定量核算了各类自然资源禀赋规模与时空演变规律,为区域长时间序列生态保护情况评估与社会经济发展策略回溯提供了数据支持,对我国自然资源调查和经济社会绿色健康发展具有现实意义[12]。

1　基于机器学习的多源异构数据融合技术

1.1　数据融合理论简介

数据融合指处理来自单一和多个来源的数据和信息关联的多层次过程,以实现重新定位,从而及时、完善地对其形势、风险及重要性进行评估[13],主要包括数据级融合、特征级融合和决策级融合 3 类。数据运营层主要针对数据读入、置信验证等方面进行原始数据融合; 数据仓库层主要针对权重筛查、时空定位、特征空间提取进行特征数据融合; 数据产品层主要针对全局模拟、需求导向等进行决策分析融合(图 1)。

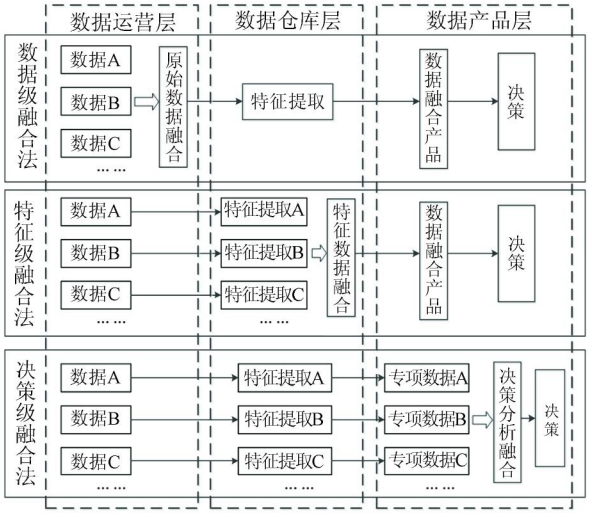

图 1　3 类数据融合方法技术路线

Fig. 1　Technical flow chart for 3 types of data fusion methods

1.2　数据规则化重构

数据规则化重构是数据融合的先决条件,也是数据管理的必要步骤。随着生态环境质量评估与自然资源存量调查的不断深入,数据源不断丰富,不同的变量数据在数据结构、格式、时空分辨率等方面均存在较大差异,需预先进行数据规则化重构。在数据建库过程中要兼顾服务器存储与计算效率,通常采用 PostgreSQL、MySQL、Oracle 等主流数据库软件平台搭建目标数据的底层架构,并通过搭建数据索引提高数据检索速度,建成融合研究前的环境基础数据库。这个环境基础数据库为自然资源变化区域的快速识别与精准定位提供了有效抓手。

1.3　机器学习建模评估

多源遥感 NDVI 在不同植被类型区域内的相关性不同,即在像元尺度上的相关性存在差异,难以依据线性关系进行有效拟合[14]。随机森林(Randon Forest,RF)是精细空间和时间分辨率下预测地面植被覆盖情况的有效工具,可以有效解决上述问题[15-16]。本文以 RF 为主体,辅以遗传算法进行因子权重与数据特征空间迭代筛查,实现机器学习数据融合。在模型训练过程中,导入训练数据集构建回归树。随机选择三分之一的预测变量用于构建每棵树[17]。首先,基于单个节点构建一个树; 然后,重复引导步骤,直到每个终端节点中只有一个数据条,从大量训练样本中提取特征,在回归树的每个节点处选择最佳分割,构建自变量与各协变量之间的相互关系,提取训练样本特征空间; 最后,建立指标因子预测子模型。植被变化不仅包括自然属性,还涵盖经济、社会、生态等多类人文属性。通过融合 3 类 NDVI 数据产品和 Landsat 部分解译数据,配合气象、地形、流域模式、人口密度等环境协变量对研究区域及时段进行模型预测。

2　材料与方法

2.1　研究区概况

汉江流域地处长江经济带中部,涵盖面积超过 15 万 km^2,位于我国南北气候过渡带,气候温和湿润(年均气温 14.1 ℃),水量较丰沛(年均降水量 972 mm),是我国重要的水源涵养地和长江中游生态保护屏障区。区域温带季风气候与平原地形特点赋予了流域良好的植被覆盖条件,流域天然植被主要为亚热带常绿阔叶林与常绿和落叶阔叶混交林。流域地势呈现西北高、东南低的特点,分别以干流丹江口和钟祥为节点,区分上、中、下游。上游高山耸立,峡谷多,植被景观丰富,丹江口水库是南水北调的中线水源区; 中、下游的江汉平原是我国中部地区重要的农作物产区[18],城市外延化进程

明显。区域工农业等社会、经济活动的不断加剧与人口的快速增长,造成流域生态功能弱化、自然资源减少等,这些问题值得关注。

2.2　基于机器学习的汉江流域多源 NDVI 数据重构

数据重构主要包括数据获取与清洗、特征工程建模、模型检验、产品输出等过程。本研究针对汉江流域上中下游植被的不同特点,结合区域林地、草地、湿地等主要土地利用类型,开展了基于机器学习的多源 NDVI 数据重构研究,通过交叉验证与真实值检验等方式评估了重构数据的准确性与精度。

2.2.1　数据获取与清洗

NDVI 数据来源于 MODIS(美国)、SPOT – VGT(法国等)、PROBA – V(欧洲)3 类卫星传感器,时间跨度分别为 2000 年 1 月至 2019 年 12 月,2000 年 1 月至 2014 年 5 月和 2013 年 10 月至 2019 年 12 月。MODIS 产品为 16 d 短期合成数据,一定程度上消除了大部分气象因素与云层的影响,但仍存在部分噪声干扰[19]。SPOT 产品对于常绿阔叶林和针叶林的指示准确,优于 MODIS[20],但受卫星寿命限制,已于 2014 年 5 月停止提供数据。PROBA – V 产品是一类植被专有观测传感器,具有与 SPOT – VGT 相似的光谱特征,旨在延续其地表植被观测任务,两者在整体上保持了观测一致性(均方根误差 RMSE 为 0.003),同时也存在某些未知的非系统差异[21]。

基础数据共涵盖 NDVI 数据、自然类环境协变量、社会经济协变量等 12 种不同数据来源(表 1)的 45 个数据信息。各类数据均进行了值域分布检查、异常值剔除、置信区间筛查,去除了部分不良噪音。根据不同数据源格式,基于 R、Python、SQL 等不同计算机编译语言,实现了数据批量导入[22]。

表 1　基础数据信息汇总

Tab.1　Data information summary of the basic database

类别	数据来源	数据格式	时间分辨率	空间分辨率
NDVI	Terra MODIS 遥感数据	hdf	每 16 d	250 m
	Aqua MODIS 遥感数据	hdf	每 16 d	250 m
	SPOT – VGT 遥感数据	hdf	每 10 d	1 km
	PROBA – V 遥感数据	hdf	每日	500 m
气温	中国气象数据网	txt	每日	0.5°
降水	中国气象数据网	txt	每日	0.5°
坡度	美国 SRTM 地形产品	tif	唯一	30 m
海拔	美国 SRTM 地形产品	tif	唯一	30 m
地形地貌类型	中国科学院资源环境科学数据中心[23]	hgt	唯一	1∶100 万
土地利用类型	欧洲航天局气候变化计划	GTiff	每年	300 m
人口密度	NASA 世界网格人口数据集	tif	每 5 a	1 km
国内生产总值	中国科学院资源环境科学数据中心[24]	ovr	每 5 a	1 km

2.2.2　基于机器学习的多源数据融合建模

本研究首先构建了汉江流域高分辨率空间网格(1 km × 1 km),获得基础网格单元 155 365 个。之后以盆地网格要素的单元格中心点为基准,将各类数据进行重采样处理,嵌套进入对应网格中。Landsat 辅助解译数据直接依据经纬度进行网格落定;NDVI 值(两组卫星数据插值后)、人口密度和国内生产总值(Gross Domestic Product, GDP)3 类数据的空间分辨率与基础网格一致,采用最近距离法进行重采样匹配;气象(差值后数据)、海拔、NDVI 值(年度最大值)和土地利用类型 4 类环境协变量数据的空间分辨率高于已有网格,采用嵌套与反距离权重插值相结合的方法,对源数据网格内多测量值进行加权和加和;PBLH 和排放清单数据的空间分辨率低于基础网格,采用反距离权重插值方法,基于源数据的多测量值的加权平均,进行网格值重采样。同时,为了保证数据的空间平滑性,对人口密度、海拔、NDVI 和土地利用类型 4 种数据均进行了二次空间卷积,卷积前后的两个变量均作为变量数据加入模型构建中,相关过程基于 PostGIS、Rstudio 等实现(图 2)。

经过梳理,20 a 的基础数据中,有效记录为 3 728.76 万条,每个数据集设立唯一的 DOI 编码,明确数据溯源,便于数据后期发布过程中的知识产权保护。数据均依据变量类别,通过数据时段和网格编号 ID 实现各类信息时空化识别与提取,为下一步数值建模提供支撑。模型训练样本为 2015—2019 年 Landsat 影像解译数据及部分实测值。模型添加了季节性变量,对变量取值空间进行了有效分隔。

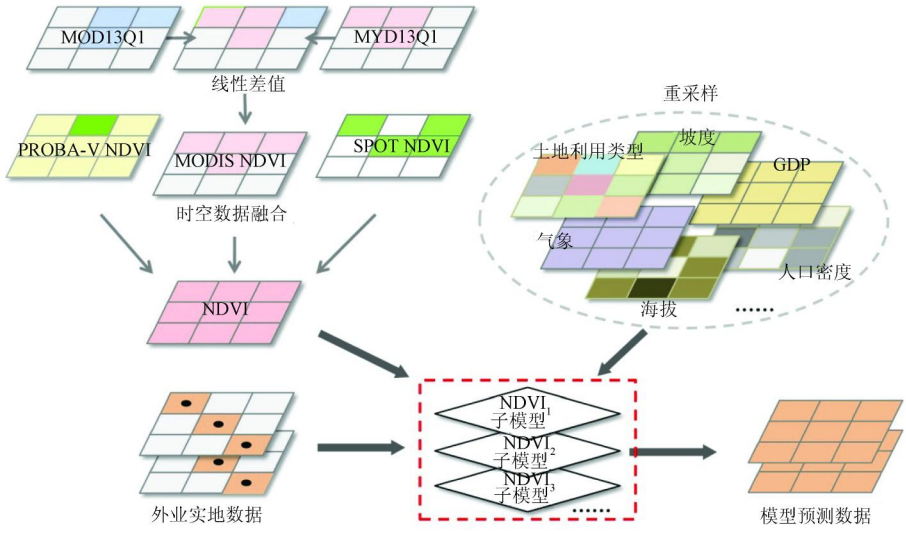

图 2　研究技术路线

Fig. 2　Technical flow chart of the research

通过量化各变量因子单一置换后的预测误差结果差异,筛查出每个变量的相对重要性[25]。基于袋外误差结果,剔除了各子模型中相关重要性低(<5‰)的自变量。依据多组模型超参数调整实验结果,各子模型中树的棵数设置为 500,最终预测结果取所有回归树结果的均值。在并行与并发运算支持[26]下,单次模型预测运行时间为 55 min,各子模型的模拟结果均达到近似最优的计算效率和预测性能。

2.2.3　模型准确性检验

k 折交叉验证是检验时空模型泛化能力的合理有效的方法,可以有效避免模型可能存在的过度拟合现象。将模型的训练数据根据数量大小,平均分为 k 份,每次使用其中的($k-1$)份数据进行模型训练,预测余下 1 组数据,最后将 k 次训练的结果全部合并,并与原始训练集数据进行比较,根据决定系数(R^2)、均方根误差(Root Mean Square Error, RMSE)等指标衡量模型的预测准确性。

3　结果与讨论

3.1　模型验证与评估

本文兼顾服务器计算效率,基于网格经纬度的分组方式将 32.7 万行训练数据进行 20 折交叉验证,得出决定系数 R^2 为 0.86,表明了模型在 NDVI 时空分布重构上的优越性(图 3)。同时,基于年份与月份进行交叉验证,R^2 分别为 0.77 和 0.82,基于流域上、中、下游分别建模验证,R^2 分别为 0.88、0.86 和 0.82,表明模型在时间外延与空间外延上均表现出较好的预测准确性。同时,根据流域 42 个实地林草样地调查结果比对,重构数据的植被覆盖准确度为 92.9%,高于单一数据源 MODIS(88.0%)、SPOT – VGT(83.3%)和 PROBA – V(76.1%),体现了基于机器学习的多源数据融合技术的优势。

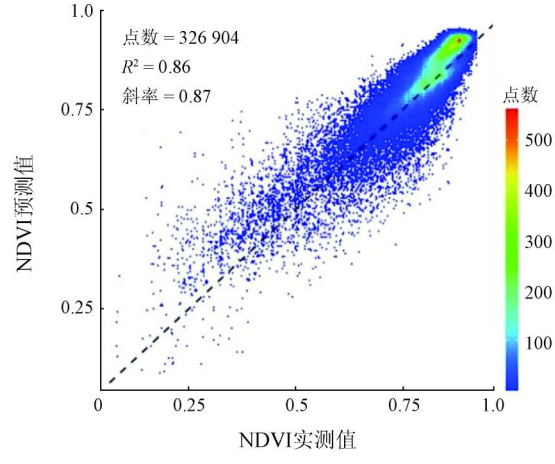

图 3　模型交叉验证结果

Fig. 3　Cross – validation results of the model

3.2　汉江流域 2000—2019 年植被 NDVI 时空变化

NDVI 值域高、低地区交错,受局部气候、地形、人文等因素分布差异影响,具有明显的空间异质性[27-28]。流域上游植被茂密,植被覆盖处于相对最高水平(NDVI >0.8),属亚热带山地湿润季风气候,降水与日照充足,气候温和,区域的水热条件非常适合植被的生长和更新[29],森林覆盖率高,汉中

市、安康市市区及周边地区是上游植被覆盖较低的区域；中、下游各城市及周边区域植被覆盖较低（$NDVI_{城区}$ =0.52 ±0.03），通过与县级及以上等级的居民点叠加分析，NDVI 低值区主要为城镇等人口聚集区，与 Landsat 影像解译结果一致，丹江口水库是汉江流域的重点水利工程，其改变了流域中、下游部分生态系统的原有面貌[30]。研究区植被覆盖水平相对较低（NDVI≤0.3）的区域主要分布于丹江口水库和武汉市、襄阳市、南阳市市区及其周边地区（图 4）。

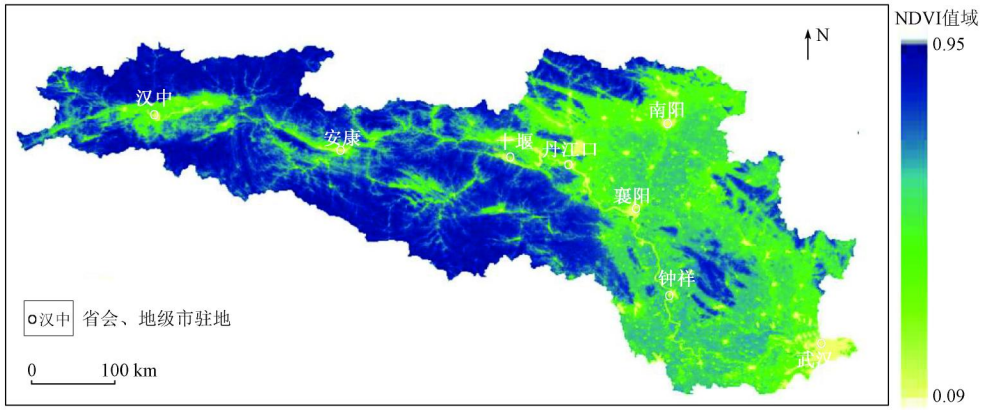

图 4　汉江流域 2000—2019 年 NDVI 年最大值空间分布

Fig. 4　Spatial distribution of annual NDVI maxima in Hanjiang Basin from 2000 to 2019

流域的植被覆盖率整体呈波动增加趋势，总增长率为 1.6%/10 a，中、上游增量较明显[31]（增长率分别为 2.2%/10 a 和 1.6%/10 a），下游植被覆盖率基本维持不变，一直处于波动阶段。流域植被改善面积达到 75.1%，其中 5.4% 面积的植被改善程度超过 10%，植被退化面积比例为 10.2%。植被覆盖变化分布存在地区差异，河流沿岸和人类活动密集区植被覆盖变化显著[32]（图 5）。计算结果表明，20 a 间流域植被覆盖上升区人口密度平均减少 0.3%，植被退化区人口密度平均增长 4.0%。植被覆盖上升区主要分布于汉江上游沿岸和流域东北部区域，丹江口水库周边与荆门市西部区域植被改善情况尤为明显，一定程度上表明国家水源保护地退耕还林、荒地造林、水土保持等政策的有效性，表明人类活动发挥了积极作用[33-34]。植被覆盖减少区则主要位于城市及周边区域，也是人类活动密集区。

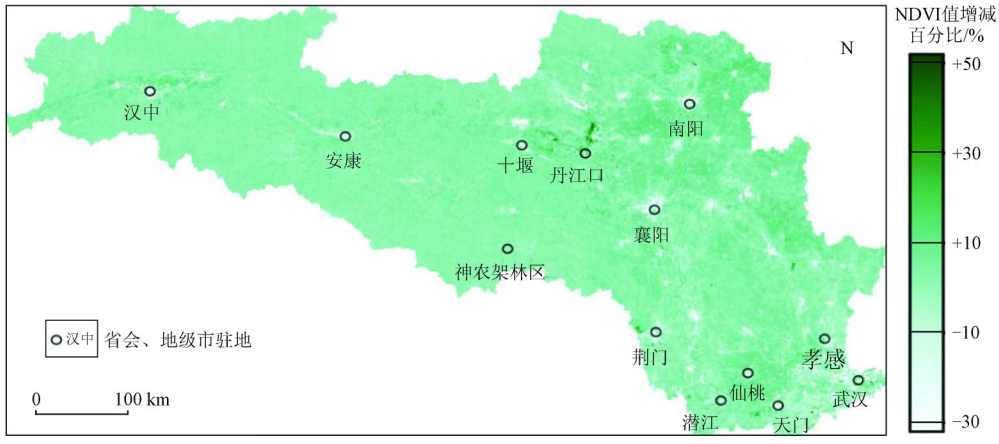

图 5　汉江流域 2000—2019 年 NDVI 空间变化趋势

Fig. 5　Spatial variation trend of NDVI in Hanjiang Basin from 2000 to 2019

3.3　汉江流域植被 NDVI 变化与人类活动间关联

流域的土地利用类型主要包括林地、园地、耕地、湿地/水体和城区，各类土地植被变化特征有所差异。本文将获得的 NDVI 数据集与流域两类土地利用类型数据相交叠加，得到流域各类土地 NDVI 时空变化序列（表 2），进而评估出区域自然资源赋存与生态环境质量情况。

由表 2 可知：汉江流域所属林地与园地主要位于

表 2　汉江流域 2000—2019 年不同土地利用类型下 NDVI 最大值汇总

Tab. 2　NDVI maxima of different land use types in Hanjiang Basin from 2000 to 2019

年份	林地	园地	耕地	湿地/水体	城区
2000	0.894	0.870	0.785	0.563	0.568
2001	0.899	0.877	0.792	0.563	0.573
2002	0.897	0.876	0.785	0.555	0.538
2003	0.897	0.876	0.785	0.568	0.533
2004	0.899	0.883	0.795	0.566	0.536
2005	0.898	0.881	0.801	0.575	0.531
2006	0.902	0.890	0.802	0.569	0.531
2007	0.902	0.891	0.801	0.577	0.523
2008	0.900	0.882	0.796	0.567	0.508
2009	0.902	0.886	0.802	0.571	0.508
2010	0.901	0.885	0.796	0.575	0.520
2011	0.898	0.884	0.791	0.570	0.498
2012	0.902	0.888	0.802	0.582	0.512
2013	0.906	0.898	0.805	0.587	0.521
2014	0.903	0.890	0.794	0.568	0.481
2015	0.912	0.904	0.817	0.594	0.535
2016	0.909	0.900	0.809	0.568	0.484
2017	0.915	0.905	0.808	0.582	0.496
2018	0.916	0.903	0.811	0.564	0.498
2019	0.911	0.896	0.796	0.566	0.483

上游地区,一直保持着整体较高的植被覆盖水平且稳中有升($NDVI_{林地}=0.903\pm0.006$,$NDVI_{园地}=0.888\pm0.010$),长期以来的森林抚育、封山育林等积极行为使森林生态系统保持了稳定向好的趋势[35-36];耕地主要位于中、下游的江汉平原,NDVI 维持稳定水平($NDVI_{耕地}=0.799\pm0.009$);湿地/水体主要分布于河流及周边区域,NDVI 水平中等($NDVI_{湿地/水体}=0.572\pm0.009$),变化不明显,丹江口水库大坝下游,即流域中、下游,湿地生态系统有所恢复;城区 NDVI 则下降较为突出,每 10 a 平均下降 4.7%,城市建设用地的不断扩张带来了植被的消极变化。上游森林资源与下游耕地资源均保持了相对稳定的水平,一定程度上体现了 20 a 间上游森林生态系统与中、下游耕地资源的相对稳定性[37]。但随着城镇化进程的不断推进,人类活动密集与城市向外扩张造成城区及周边区域植被覆盖显著减少,区域生态风险形势依然不容乐观。

基于研究区各网格单位计算流域多年期 NDVI 与人口密度 Spearman 秩相关系数,两者相关性空间分布具有明显的空间异质性(图 6)。NDVI 与区域人口密度正相关性区域占总面积的 28%,主要集中于河南省南阳市辖区,印证了该区域退耕还林工程成效明显[38];负相关性区域占总面积的 72%,主要分布于流域中游耕地区及人口密度较高的城市区域。两类截然不同的相关系数分布情况体现了人类活动对植被覆盖影响的不确定性和随机性,会受到国家政策和不同时期发展需求等多种因素的影响[39]。

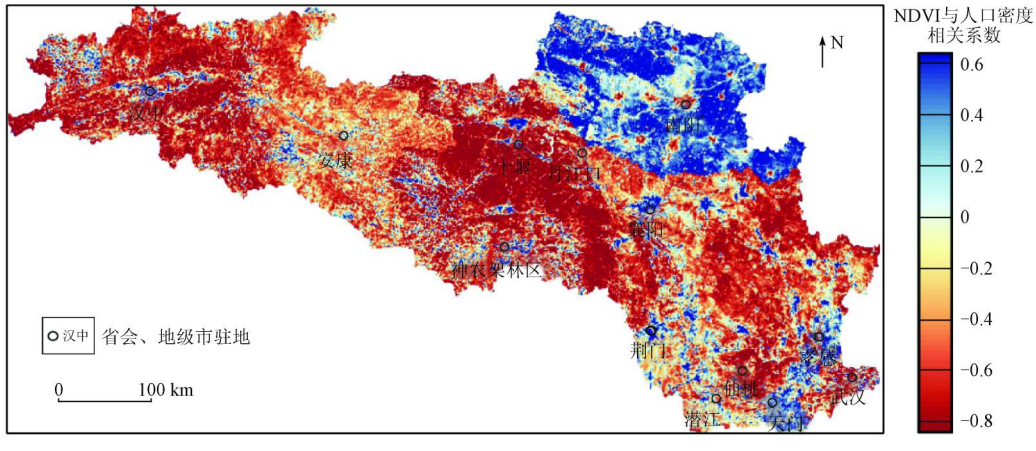

图 6　汉江流域 NDVI 与人口密度相关系数空间分布

Fig. 6　Spatial distribution of the correlation coefficient between NDVI and population density in Hanjiang Basin

3.4　不足与展望

本文主要针对植被每年的生长旺盛期进行逐年 NDVI 最大值模拟与分析,未进行植被生长季全周期的跟踪观测。未来可基于该融合技术方法,进一步提升数据的时空分辨率,模拟年内植被生长全过程,更精准地实现植被动态观测,更好地支撑自然资源管理与生态质量评估。

4　结论

本研究聚焦自然资源信息高效提取与利用,以

汉江流域 NDVI 数据为例,探索了一种多源异构数据融合技术,主要结论如下。

(1)基于机器学习的多源数据融合技术具有速度快、准确度高、经济高效等优势,本研究面向自然资源信息提取领域,形成了一个多源异构数据智能融合技术方法,可实现数据高效利用与特征空间快速优选。

(2)以汉江流域为例,基于随机森林算法,融合了 3 种异源 NDVI 数据产品,构建了 NDVI 回溯预测子模型,获得了 2000—2019 年汉江流域 NDVI 逐年时空分布数据集,模型交叉验证决定了系数 R^2 为 0.86,空间分辨率为 1 km。模型从多源数据中优化提取了数据特征空间,与原有单一数据产品相比,模拟结果更贴近实际,数据质量有所提升。

(3)汉江流域植被变化与区域人类活动密切相关,两者相关系数分布存在显著的空间异质性,正相关区主要为流域东北部区域,负相关区主要为流域中游耕地地区与城市周边区域。人类活动对植被的影响受国家政策、经济发展等多方面因素控制。

参考文献(References):

[1] 余辉,梁镇涛,鄢宇晨. 多来源多模态数据融合与集成研究进展[J]. 情报理论与实践,2020,43(11):169 - 178.
Yu H,Liang Z T,Yan Y C. Review on multi-source and multi-modal data fusion and integration[J]. Inf Stud:Theory Appl,2020,43(11):169 - 178.

[2] Zhu X L,Cai F Y,Tian J Q,et al. Spatiotemporal fusion of multisource remote sensing data:literature survey,taxonomy,principles,applications,and future directions[J]. Remote Sens,2018,10(4):527.

[3] 唐见,曹慧群,陈进. 生态保护工程和气候变化对长江源区植被变化的影响量化[J]. 地理学报,2019,74(1):76 - 86.
Tang J,Cao H Q,Chen J. Effects of ecological conservation projects and climate variations on vegetation changes in the source region of the Yangtze River[J]. Acta Geogr Sin,2019,74(1):76 - 86.

[4] 徐凯健,田庆久,徐念旭,等. 基于时序 NDVI 与光谱微分变换的森林优势树种识别[J]. 光谱学与光谱分析,2019,39(12):3794 - 3800.
Xu K J,Tian Q J,Xu N X,et al. Classifying forest dominant trees species based on high dimensional time - series NDVI data and differential transform methods[J]. Spectrosc Spect Anal,2019,39(12):3794 - 3800.

[5] Leveau L M,Isla F I,Bellocq M I. From town to town:predicting the taxonomic,functional and phylogenetic diversity of birds using NDVI[J]. Ecol Indicat,2020,119:106703.

[6] Zhang Y H,Ling F,Foody G M,et al. Mapping annual forest cover by fusing PALSAR/PALSAR - 2 and MODIS NDVI during 2007 - 2016[J]. Remote Sens Environ,2019,224:74 - 91.

[7] Cao R Y,Chen Y,Shen M G,et al. A simple method to improve the quality of NDVI time - series data by integrating spatiotemporal information with the Savitzky - Golay filter[J]. Remote Sens Environ,2018,217:244 - 257.

[8] Luo F L,Zhang L P,Du B,et al. Dimensionality reduction with enhanced hybrid - graph discriminant learning for hyperspectral image classification[J]. IEEE Trans Geosci Remote Sens,2020,58(8):5336 - 5353.

[9] 高艳丽,李红波,侯蕊. 汉江流域生态系统服务权衡与协同关系演变[J]. 长江流域资源与环境,2020,29(7):1619 - 1630.
Gao Y L,Li H B,Hou R. Evolution analysis on trade-offs and synergies of ecosystem services in Hanjiang River Basin[J]. Resour Environ Yangtze Basin,2020,29(7):1619 - 1630.

[10] 张翔,邓志民,李丹,等. 汉江流域土地利用/覆被变化的水文效应模拟研究[J]. 长江流域资源与环境,2014,23(10):1449 - 1455.
Zhang X,Deng Z M,Li D,et al. Simulation of hydrological response to land use/cover change in Hanjiang Basin[J]. Resour Environ Yangtze Basin,2014,23(10):1449 - 1455.

[11] Zhang Y Y,Ban X,Li E H,et al. Evaluating ecological health in the middle - lower reaches of the Hanjiang River with cascade reservoirs using the Planktonic Index of Biotic Integrity (P - IBI)[J]. Ecol Indicat,2020,114:106282.

[12] 葛良胜,夏锐. 自然资源综合调查业务体系框架[J]. 自然资源学报,2020,35(9):2254 - 2269.
Ge L S,Xia R. Research on comprehensive investigation work system of natural resources[J]. J Nat Resour,2020,35(9):2254 - 2269.

[13] White F E. Data Fusion Lexicon[M]. Washington:Joint Directors of Labs Washington DC,1991.

[14] Zhou J X,Chen J,Chen X H,et al. Sensitivity of six typical spatiotemporal fusion methods to different influential factors:a comparative study for a normalized difference vegetation index time series reconstruction[J]. Remote Sens Environ,2021,252:112130.

[15] Millard K,Richardson M. On the importance of training data sample selection in random forest image classification:a case study in peatland ecosystem mapping[J]. Remote Sens,2015,7(7):8489 - 8515.

[16] Belgiu M,Drăguţ L. Random forest in remote sensing:a review of applications and future directions[J]. ISPRS J Photogramm Remote Sens,2016,11(4):24 - 31.

[17] Breiman L. Random forests[J]. Machine Learning,2001,45(1):5 - 32.

[18] 李彩霞,邓帆,张佳华,等. 基于时序植被指数的湖北省物候空间特征分析[J]. 长江流域资源与环境,2019,28(7):1583 - 1589.
Li C X,Deng F,Zhang J H,et al. Phenological spatial characteris-

tics of Hubei Province based on time series vegetation index[J]. Resour Environ Yangtze Basin,2019,28(7):1583 – 1589.

[19] Guzmán Q J A, Sanchez – Azofeifa G A, Espírito – Santo M M. MODIS and PROBA – V NDVI products differ when compared with observations from phenological towers at four tropical dry forests in the Americas[J]. Remote Sens,2019,11(19):2316.

[20] Toté C, Swinnen E, Sterckx S, et al. Evaluation of the SPOT/VEGETATION collection 3 reprocessed dataset: surface reflectances and NDVI[J]. Remote Sens Environ,2017,201:219 – 233.

[21] Meroni M, Fasbender D, Balaghi R, et al. Evaluating NDVI data continuity between SPOT – VEGETATION and PROBA – V missions for operational yield forecasting in North African countries[J]. IEEE Trans Geosci Remote Sens, 2016, 54(2): 795 – 804.

[22] 汤宇磊,杨复沫,詹宇. 四川盆地 PM2.5 与 PM10 高分辨率时空分布及关联分析[J]. 中国环境科学,2019,39(12):4950 – 4958.
Tang Y L, Yang F M, Zhan Y. High resolution spatiotemporal distributionand correlation analysis of PM2.5 and PM10 concentrations in the Sichuan Basin[J]. China Environ Sci,2019,39(12): 4950 – 4958.

[23] 周成虎,程维明.《中华人民共和国地貌图集》的研究与编制[J]. 地理研究,2010,29(6):970 – 979.
Zhou C H, Cheng W M. Research and compilation of the Geomorphological Atlas of the People's Republic of China[J]. Geograph Res,2010,29(6):970 – 979.

[24] 徐新良. 中国 GDP 空间分布公里网格数据集[DB/OL]. 中国科学院资源环境科学数据中心,2017. http://www.resdc.cn/.
Xu X L. Spatial distribution of national GDP in 1km grid[DB/OL]. Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences,2017. http://www.resdc.cn/.

[25] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. J Chem Inf Comput Sci,2003,43(6):1947 – 1958.

[26] Wright M N, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C + + and R[J]. J Stat Softw,2017,77(1):1 – 17.

[27] Sun Q Y, Liu W W, Gao Y N, et al. Spatiotemporal variation and climate influence factors of vegetation ecological quality in the Sanjiangyuan National Park[J]. Sustainability, 2020, 12(16): 6634.

[28] Solangi G S, Siyal A A, Siyal P. Spatiotemporal dynamics of land surface temperature and its impact on the vegetation[J]. Civil Eng J,2019,5(8):1753 – 1763.

[29] 赵芳,张久阳,刘思远,等. 秦巴山地 NPP 及对气候变化响应的多维地带性与暖温带 – 亚热带界线[J]. 生态学报,2021,41(1):57 – 68.
Zhao F, Zhang J Y, Liu S Y, et al. Assessing the dividing line between warm temperate and subtropical zones based on the zonality discussion on multi – dimensional response of Net Primary Productivity to climate change in the Qinling – Daba Mountains[J]. Acta Ecol Sin,2021,41(1):57 – 68.

[30] Wang Y K, Wang D, Wu J C. Assessing the impact of Danjiangkou reservoir on ecohydrological conditions in Hanjiang river, China[J]. Ecol Eng,2015,81:41 – 52.

[31] 杨倩,刘登峰,孟宪萌,等. 汉江上游植被指数变化及其归因分析[J]. 南水北调与水利科技,2019,17(4):138 – 148.
Yang Q, Liu D F, Meng X M, et al. Vegetation index change in the upper reaches of Han River and its attribution analysis[J]. South North Water Transf Water Sci Technol,2019,17(4):138 – 148.

[32] 马梓策,于红博,曹聪明,等. 中国植被覆盖度时空特征及其影响因素分析[J]. 长江流域资源与环境,2020,29(6):1310 – 1321.
Ma Z C, Yu H B, Cao C M, et al. Spatiotemporal characteristics of fractional vegetation coverage and its influencing factors in China[J]. Resour Environ Yangtze Basin, 2020, 29(6): 1310 – 1321.

[33] 徐静文,肖飞,廖炜,等. 基于 MODIS NDVI 汉江中游植被时空变化及其地貌分异分析[J]. 长江流域资源与环境,2017,26(11):1895 – 1901.
Xu J W, Xiao F, Liao W, et al. Spatial – temporal changes of vegetation and its geomorphic differentiation in the middle reaches of the Hanjiang River based on MODIS NDVI data[J]. Resour Environ Yangtze Basin,2017,26(11):1895 – 1901.

[34] 刘海,黄跃飞,林苗,等. 基于 GIS 的汉江流域水土保持时空变化特征分析(2001—2017 年)[J]. 地域研究与开发,2019,38(3):154 – 159,164.
Liu H, Huang Y F, Lin M, et al. Analysis of temporal and spatial variation characteristics of soil and water conservation in Hanjiang River Basin based on GIS (2001 – 2017)[J]. Areal Res Dev, 2019,38(3):154 – 159,164.

[35] 王建邦,赵军,李传华,等. 2001—2015 年中国植被覆盖人为影响的时空格局[J]. 地理学报,2019,74(3):504 – 519.
Wang J B, Zhao J, Li C H, et al. The spatial – temporal patterns of the impact of human activities on vegetation coverage in China from 2001 to 2015[J]. Acta Geogr Sin,2019,74(3):504 – 519.

[36] 任正超,朱华忠,史华,等. 最后间冰期至未来 2070s 中国潜在自然植被时空分布格局及其对气候变化的响应[J]. 自然资源学报,2020,35(6):1484 – 1498.
Ren Z C, Zhu H Z, Shi H, et al. Spatio – temporal distribution pattern of potential natural vegetation and its response to climate change from Last Interglacial to future 2070s in China[J]. J Nat Resour,2020,35(6):1484 – 1498.

[37] 邓元杰,姚顺波,侯孟阳,等. 长江流域中上游植被 NDVI 时空变化及其地形分异效应[J]. 长江流域资源与环境,2020,29(1):66 – 78.
Deng Y J, Yao S B, Hou M Y, et al. Temporal and spatial variation of vegetation NDVI and its topographic differentiation effect in the middle and upper reaches of the Yangtze River Basin[J]. Resour Environ Yangtze Basin,2020,29(1):66 – 78.

[38] 刘兴,孙新杰. 南阳市退耕还林工程建设现状及展望[J]. 现代园艺,2017(17):183.

Liu X, Sun X J. The status quo and prospect of the program to return farmland to forests in Nanyang [J]. Xiandai Hortic, 2017 (17): 183.

[39] Liu Y, Li Y, Li S C, et al. Spatial and temporal patterns of global NDVI trends: correlations with climate and human factors [J]. Remote Sens, 2015, 7(10): 13233 - 13250.

Multi-source heterogeneous data fusion technology for natural resource information extraction: A case study of NDVI data in Hanjiang Basin

TANG Yulei[1,2], WU Yangyang[3], JIANG Xingzheng[1], FENG Liang[1], GAO Yang[4]

(1. *Center for Geophysical Survey, China Geology Survey, Hebei LangFang* 065000, *China*; 2. *Key Laboratory of coupling process and effect of natural resources elements, Beijing* 100055, *China*; 3. *College of Architecture and Environment, Sichuan University, Sichuan Chengdu* 610065, *China*; 4. *College of Land Science and Technology, China Agricultural University, Beijing* 100193, *China*)

Abstract: Natural resource indicator data with high spatio-temporal resolution are essential for large-scale natural resource dynamic observation and trend assessment. The large amount of multi-source data under big data era could provide the possibility for efficient utilization and fusion of data. Taking the Normalized Difference Vegetation Index (NDVI) in Hanjiang Basin as an example, the authors in this paper have built a spatio-temporal big data processing underlying architecture for natural resources based on PostgreSQL, and integrated three types of methods, including data-level fusion, feature-level fusion and decision-level fusion. Besides, the intelligent fusion system of multi-source heterogeneous data has been constructed based on the machine learning algorithms to achieve efficient utilization of multi-source data and feature spatial preference. Meanwhile, the year-by-year NDVI 1 km dataset of Hanjiang Basin from 2000 to 2019 has been reconstructed to comprehensively reflect the dynamic changes of vegetation in Hanjiang Basin. These results could provide some scientific reference for the efficient extraction and simulation analysis of spatio-temporal big data in earth sciences, and provide a more accurate and convenient technical means for quantitatively accounting the scale of forest and grassland resources endowment and exploring the spatio-temporal evolution of ecosystem.

Keywords: natural resources; multi-source heterogeneity; spatio-temporal data fusion; machine learning

(责任编辑: 刘丹)