DOI: 10.20103/j.stxb.202312112704

周婷,徐含乐,徐奇刚,朱本挥,陆亚刚.基于随机森林分类模型填补森林资源连续清查缺失因子.生态学报,2024,44(18):8269-8282.

Zhou T, Xu H L, Xu Q G, Zhu B H, Lu Y G.Filling missing factors for China′s forest resources inventory based on random forest classification models.Acta Ecologica Sinica,2024,44(18):8269-8282.

基于随机森林分类模型填补森林资源连续清查缺失因子

周 婷[1],徐含乐[1],徐奇刚[2],朱本挥[1],陆亚刚[2,*]

1 浙江大学环境与资源学院, 杭州 310058

2 国家林业和草原局华东调查规划院, 杭州 310019

摘要:森林资源连续清查是评估林业生态建设成效和制定发展战略的重要依据。然而,由于监测体系和不同时期需求的升级,清查数据存在调查因子不连贯的问题。基于华东地区六省一市第六期至第九期清查数据,根据五个林分结构缺失因子和不同因子之间的相关性选取特征因子,采用随机森林分类模型填补缺失因子并分析特征因子的重要性。结果显示:(1)缺失因子和当期因子之间的相关系数普遍高于后期因子,其中植被类型、树种结构和其他相关因子的平均相关系数为 0.868 和 0.733,显著高于其他三个缺失因子;(2)所有随机森林分类模型的准确度均达到 0.770 以上,并且在省级和县级尺度都具有出色的外部有效性,其中相关性系数高的缺失因子对应的模型准确度也相对更高;(3)特征因子重要性结果与相关性分析的结果基本吻合,显示特征因子组合中当期因子的占比高有助于提高模型填补性能,此外缺失因子本身对应的后一期数值对提高模型填补性能的贡献较大。研究可用于完善森林资源动态监测数据库,为科学评估我国生态保护建设成效以及完善森林分类经营管理制度提供支撑。在未来的研究中,基于国家森林高质量发展的基本策略,通过严密的实验设计评估我国森林分类经营制度的建设成效,对于完善我国森林生态效益补偿制度,建立稳定、健康、优质、高效的森林生态系统具有至关重要的作用。

关键词:缺失因子;森林资源清查;生态保护;机器学习;森林分类经营

Filling missing factors for China′s forest resources inventory based on random forest classification models

ZHOU Ting[1], XU Hanle[1], XU Qigang[2], ZHU Benhui[1], LU Yagang[2,*]

1 *College of Environmental and Resource Sciences, Zhejiang University, Hangzhou* 310058, *China*

2 *East China Academy of Inventory and Planning of National Forestry and Grassland Administration, Hangzhou* 310019, *China*

Abstract: Forest resource inventory is an important scientific basis for comprehensively understanding the effectiveness of China′s forestry ecological construction and formulating forestry sustainable development strategies. However, due to updates in forestry monitoring systems and changing requirements for ecological environment construction over time, there are some missing factors existing in forest resource inventory. This study is based on the data from the sixth (1999—2003) to ninth (2014—2018) periods of forest resource continuous inventory in East China. First, we selected feature factors based on the results of correlation analysis of five forest stand structure factors including vegetation type, tree species composition, forest community structure, regeneration level, and naturalness. Then, we used random forest classification models to fill in the missing factors and assessed the accuracies of models, as well as examined the external validity of models. Finally, we analyzed the importance of feature factors. Our results show that: (1) the correlation coefficients between missing factors

基金项目:国家重点研发计划(2022YFE0112700);国家留学基金委国家建设高水平大学公派研究生项目(202306320489)

收稿日期:2023-12-11; **网络出版日期**:2024-07-12

* 通讯作者 Corresponding author.E-mail: hdyluyagang@ qq.com

and factors from the same period are overall higher than those with factors from later periods. Among them, the average correlation coefficients between vegetation type and tree species composition and other related factors are 0.868 and 0.733, significantly higher than the other three missing factors; (2) overall, the random forest classification models achieve an accuracy of 0.770 or higher in all five missing factors, demonstrating outstanding external validity at both provincial and county scales. Moreover, models corresponding to factors with high correlation coefficients exhibit relatively higher accuracy (the accuracies of vegetation type and tree species composition are 0.900 and 0.841, respectively); and (3) the results of feature factor importance align closely with the findings of the correlation analysis, indicating that a higher proportion of the same-period factors within the feature factor combination contributes to enhancing the model imputation performance. Additionally, the missing factor itself exhibits a significant contribution to improving the model imputation performance in relation to the subsequent-period values. The findings of the study can be utilized to refine the dynamic monitoring database of forest resources, thereby providing support for a scientific evaluation of the effectiveness of China's forestry protection programs, and the enhancement of China's forest classification and management system. In future research, based on the fundamental strategy of promoting the high-quality development of national forests, rigorous experimental designs should be employed to evaluate the effectiveness of China's forest classification and management systems. This effort is crucial for refining our country's compensation mechanism for forestry ecological construction and establishing a stable, healthy, high-quality, and efficient forest ecosystem.

Key Words: missing factors; forest resources inventory; ecological protection; machine learning; forest classification management

森林资源定期清查是全面了解我国森林资源动态变化的基础,也是全面把握我国林业生态建设成效和制定林业发展战略的重要科学依据[1—2]。2021 年,为统筹推进山水林田湖草沙一体化保护和修复,我国森林资源清查从单一的森林调查监测转向每年开展森林、草原、湿地、沙化、石漠化土地综合监测。而在此之前,森林资源清查(下面连续清查简称连清)每五年开展一次,从 1973 年开始至 2018 年已完成九次连清工作,第一期(1973—1976)至第六期(1999—2003)的森林资源连清监测内容较为单一,调查因子主要反映森林面积及蓄积量[2—4]。随着林业向以生态建设为主转化,我国森林资源连清体系不断优化改进,从第七期(2004—2008)开始增加了林业管理及森林生态功能等调查因子以体现我国对森林生态效益研究的重视[4—5]。新增的调查因子可用于更加全面科学地评估我国森林资源动态变化,以及评估我国从 20 世纪 90 年代开始实施的一系列森林生态保护工程及政策绩效并揭示其变化规律和原因[3]。然而,我国大部分林业生态保护工程的起始年份都在 2000 年之前[6—9],对应时期的森林资源连清数据库存在部分调查因子不连贯的情况,也因此给后续分析(如评估政策生态效益和揭示政策作用机制)带来困难[10]。采用科学合理的方式填补森林资源连清第六期缺失的调查因子对于全面评估我国森林资源动态变化,科学评估我国生态保护建设成效以及完善我国森林分类经营管理制度具有重要的意义。

森林资源调查数据不连贯是一个全球性的问题,其本质是数据处理问题,目前国外在缺失数据的填补方法方面已经有大量研究并且取得了较多成果[11—12]。20 世纪 90 年代,Reams 等人便用移动平均法(Moving Average)和加权移动平均法(Updated Moving Average)用于补充美国南部森林资源连清缺失数据[13]。90 年代后,随着计算机水平的快速发展,Rubin 和 Lipsitz 等学者优化了早期基于极大似然原理的迭代逼近算法[14],并将贝叶斯算法应用在缺失数据的填补中[15]。但是贝叶斯算法不适用于缺失数据非随机且高维度的情况,Astebro 和 Chen 在应用该方法填补缺失的分类数据时也发现该方法的填充效果不佳[16]。21 世纪初,回归的思想被应用于填补森林连清缺失数据。回归既可以用于连续型缺失数据的填补,也可以用于离散或分类型缺失数据的填补,但是回归模型填补效果的好坏依赖于数据的分布,在数据分布不明确及拟合维度较多的情况下回归模型的填补效果较差[12,17]。另外,最近邻(K-Nearest Neighbor)填补法由于其方法的灵活性(可以设置

不同的权重及匹配规则)在过去几十年中被大量应用于森林资源连清缺失数据的填补[11—12,18—19]。近些年,随着机器学习方法的盛行和集成学习的发展,基于监督学习(Supervised Learning)的随机森林(Random Forest)模型在填补缺失数据中得到广泛的使用。Tang 和 Ishwaran[20]对比了包括随机森林、邻近插补、即时插补、利用多元无监督和监督分类的插值等不同的缺失数据处理方法,最终发现随机森林方式性能最佳且最稳健,并且在高缺失的情况下依然能保持良好的性能。

国内这方面的研究还处于起步阶段,针对填补缺失数据方法的应用案例较少,尤其是针对森林资源缺失数据的实证案例。金勇进和庞新生等是国内较早研究缺失数据填补方法的学者,他们在国外学者 Rubin 等提出的多重缺失数据填补法的基础上探究了缺失数据填补过程中的有效信息[21—22],并且通过实证分析比较了单一插补法与多重插补法的性能差异,对不同缺失数据的处理方式做出了调整[23]。后来乔珠峰[24]、梁怡[25]、胡玄子[26]和靳国栋[27]等人又分别对朴素贝叶斯分类法、均值填补法、回归分析法和插值填补法填补缺失数据的效果进行了研究。刘菲在 2019 年基于湖南省郴州市 2014 年森林资源连清数据,对比了不同方法对缺失因子林木胸径的填补效果,研究结果表明随机森林算法的综合性能最优[28]。目前国内在缺失数据处理方法上取得了一定的进步,但是将这些方法应用到实际的案例解决实际需求方面的尝试较少,尤其是大尺度的应用。

填补我国森林资源调查缺失数据本质上虽然是一个缺失数据处理问题,但森林资源调查有其自身的特点和侧重点,其缺失数据的填补是一个需要结合其自身特性的缺失数据处理问题。根据国内外填补缺失数据不同方法性能的对比结果,结合我国森林资源第六期连清数据缺失的实际情况,并且考虑到数据的可得性,本研究选取我国亚热带落叶阔叶和常绿阔叶气候带华东片区为研究区,基于第六、七、八、九次森林连清数据,采用了随机森林分类模型填补该地区第七次森林资源连清中相应的缺失因子并验证其模型的准确度,将填补性能优秀的模型应用到填补第六次森林资源连清中相应的缺失因子,为全面评估自 90 年代实施森林生态保护工程以来,我国森林生态保护建设成效提供技术支持。此外,我国森林资源连清以省为单位开展统计分析,在数据采集上以县为单位自下而上进行汇总。虽然本研究范围集中在亚热带森林生态系统,但是考虑到各个省份之间自然条件和森林资源的差异,本研究同时在省级和县级尺度验证随机森林模型填补缺失因子的性能。本研究具体的研究目标如下:(1)分析森林资源连清缺失因子和所有调查因子及气象因子的相关性和显著性;(2)验证随机森林分类模型在填补缺失因子上的性能;(3)分析不同特征因子对提高缺失因子填补性能的重要性。

1　研究区概况

本研究选取中国华东地区包含上海市、江苏省、浙江省、安徽省、福建省、江西省和河南省六省一市为研究区域(图 1)。该地区位于中国东南沿海,主要的植被类型为亚热带常绿阔叶林(以人工林为主),年均气温及降水分别为 17℃和 1363mm,地区海拔高度的范围为-68—2368m。2004—2020 年间,华东地区是中国人口密度(483 人/km^2)最高和人均 GDP(46297 元)最高的地区[29],提高该地区森林质量和效益以适应该地区对良好生态系统服务功能的需求是我国森林分类经营的整体思路[30],也是新时期我国森林高质量提升的主要任务和宏观策略[31]。

2　数据来源与研究方法

2.1　数据来源及预处理

本研究采用华东地区第六次(1999—2003)、第七次(2004—2008)、第八次(2009—2013)和第九次(2014—2018)四期森林资源连清样地数据。本研究使用的森林资源连清数据在调查时间上存在细微差异:上海、浙江、安徽始于 1999 年,江苏始于 2000 年,福建和河南始于 2003 年,江西始于 2001 年。尽管起始时间略有不同,然而每个省份相邻调查阶段之间的调查时间间隔都是五年。由于模型的训练和测试均基于相邻两

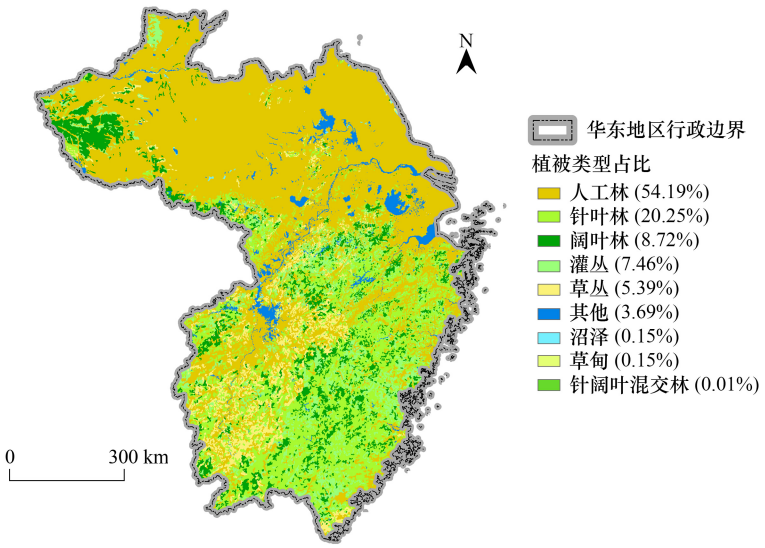

图 1 研究区地理位置及植被类型图

Fig.1 Geographical location and vegetation type of the study area

个调查阶段的数据,因此在模型分析中,可以视每一个样本为相对独立。这种一致的调查时间间隔有助于维持样本的一致性,从而确保模型分析结果的可靠性。根据国家森林资源连续清查技术规程[32],提取每一期一级地类为林地(包括乔木林地、竹林地、疏林地、灌木林地、未成林造林地、苗圃地、迹地和宜林地八个二级地类)的样地,得到每期有效样地数量为 12996 个。林分结构因子直接影响着森林生态系统的稳定性和功能性,并且是揭示政策对森林生态系统的作用机制的关键因子。所以,本研究根据第六次(1999—2003)未调查,第七次(2004—2008)、第八次(2009—2013)和第九次(2014—2018)三期均完整调查的因子中选择林分结构缺失因子作为填补对象,共计 5 个(表 1)。

表 1 华东森林森林资源连续清查缺失因子

Table 1 Category and definition of missing factors of the National Forest Inventory in Eastern China

因子类别 Factor category	因子名称 Factor name	数据类型 Data types	定义 Definition
林分结构因子 Stand structure factor	植被类型	名义分类变量	样地植被所属的植被型,主要依据《中国植被》分类系统,按面积优势法确定,将植被分为自然植被和栽培植被两大类别,其中:自然植被分 9 个植被型组,31 个植被型;栽培植被分 3 个植被型组,11 个植被型。
	树种结构	名义分类变量	反映乔木林分的针阔叶树种组成,共分 7 等级,分别为针叶纯林、阔叶纯林、针叶相对纯林、阔叶相对纯林、针叶混交林、针阔混交林和阔叶混交林。
	森林群落	有序分类变量	乔木林的群落结构划分为 3 个等级,分别为完整结构、较完整结构和简单结构。
	更新等级	有序分类变量	根据幼苗各高度级的天然更新株数确定。因子共 3 个等级,分别为良好、中等和不良。
	自然度	有序分类变量	按照现实森林类型与地带性原始顶极森林类型的差异程度,或次生森林类型位于演替中的阶段确定。因子共 5 个等级,分别为 I、II、III、IV 和 V。

2.2 研究方法

本研究的技术路线图如图 2 所示。首先将缺失因子根据数据类型进行分类,利用统计检验方法对缺失因子和所有调查因子进行相关性分析和显著性检验。第二步,采用不同的方式选择特征因子构建训练集,应用于随机森林分类模型并对比不同模型的性能。第三步,选择准确度(Accuracy)作为精度验证指标检验随机森林分类模型的性能[33],计算公式如(1)所示。然后,选择模型性能最佳的最优特征因子组合来确定最终随机森林决策树的数量。将以上得到的最优特征因子和决策树数量的组合作为最终随机森林分类模型的参数,检

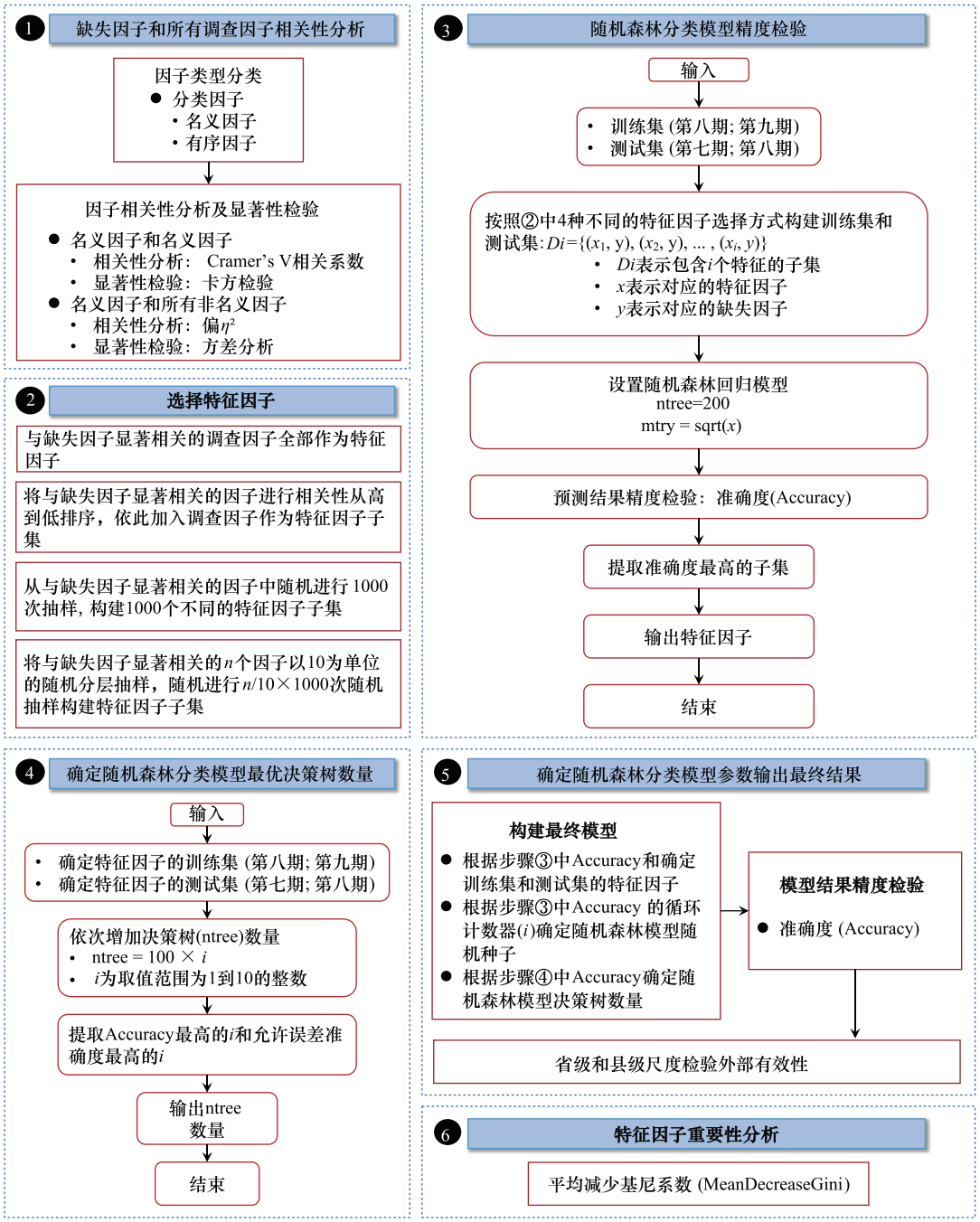

图 2　研究技术路线图

Fig.2　Roadmap of the methodology

mtry:每个节点随机抽样的变量数 Number of variables randomly sampled at each split;sqrt:平方根 Square root

验华东地区模型填补的准确度,并且计算基于华东地区数据训练得到的模型在单个省份上的准确度,然后再将模型应用到省级和县级尺度检验外部有效性。最后,根据最优随机森林分类模型输出每个缺失因子对应特征因子重要性的分析结果。本研究所有的数据分析过程都在 R 语言 4.2 版本中进行。

$$\text{Accuracy}=\frac{p_c}{p_t} \tag{1}$$

式中,p_c是模型正确预测的数量(即一个样本预测得到的类别与该样本的真实类别完全一致),p_t是所有样本

的总数量。该数值的取值范围为 0—1,数值越高代表模型填补的性能越好。

2.2.1 相关性分析及特征因子的选取

在大规模数据分析问题中,选择合适的方法消除无关的变量可以强化模型的可解释性并且避免过度拟合从而提高模型的准确性[34]。由于本次用于填补缺失因子的数据库因子众多,通过使用独立编码将多分类名义变量转化为二分类名义变量后,除缺失因子之外的调查因子达到 451 个。为了提高模型的准确性,我们根据数据类型将所有因子进行分类,再选取和缺失因子显著相关的调查因子用于后续随机森林分类模型的构建。我们此次分析的因子都是类别因子,首先将分类型因子细分为名义因子和有序因子。然后,根据因子的不同特性使用不同方法分析每两类因子组间的相关性分析并检验其相关系数的显著性。其中,对于分类型因子中的名义因子,采用克拉默(Cramer's V)相关系数来计算两个因子之间的相关性[35],然后用卡方检验(Chi-squared test)检验其相关性是否显著,使用 R 开源程序软件包"vcd"[36]实现,相关系数的计算公式如(2)所示。

$$V=\sqrt{\frac{x^2}{n\times\min\ (k-1,r-1)}} \tag{2}$$

式中,x^2是卡方检验的卡方值,n 是观测总样本数,k 是列数,r 是行数。克拉默相关系数是一种用于衡量分类变量之间关联程度的统计量。该数值的取值范围为[0, 1],0 表示没有关联,而 1 表示完全关联。对于名义分类因子和其他所有因子,我们采用偏 η^2(Partial Eta-Squared)来计算他们之间的相关性[37],然后用方差分析(ANOVA)检验数值型因子和分类型因子之间不同组别是否存在显著差异,该分析使用 R 语言内置函数"eta_squared"和"aov"实现,偏 η^2 相关系数的计算公式如(3)所示。

$$\eta^2=\frac{SS_{\text{effect}}}{SS_{\text{effect}}+SS_{\text{error}}} \tag{3}$$

式中,SS_{effect}是因变量(因素)的效应平方和,表示所有因变量均值与总均值之间的差异的平方和的加总,具体计算公式如(4)所示。SS_{error}是误差平方和,具体计算公式如(5)所示。

$$SS_{\text{effect}}=\sum_{i=1}^{k} n_i\ (\bar{Y}_i-\bar{Y})^2 \tag{4}$$

式中,k 是因素的个数;n_i是第 i 个因素水平的观测值数量;$\bar{Y}_i$是第 i 个因素水平的观测值均值;$\bar{Y}$ 是总体观测值的均值。

$$SS_{\text{error}}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\ (Y_{ij}-\bar{Y}_i)^2 \tag{5}$$

式中,n_i是第 i 个因素水平的观测值数量;Y_{ij}是第 i 个因素水平的第 j 个观测值;$\bar{Y}_i$是第 i 个因素水平的观测值均值。

在完成缺失因子和所有调查因子之间的相关性和显著性分析后,选择合适的方法进行特征因子的选择可以有效减少模型预测的维度,进一步提高模型预测的性能[38—39]。在结合传统线性统计中向前选择、向后除去和逐步式选择等方式的基础上,加入不同模式的随机抽样,制定出以下四种不同的特征选择方式,每种选择方式都是基于对前一种选择方式的进一步补充和完善:第一种是放入和缺失因子显著且相关的所有因子作为训练集,这种方式的优点是比较全面地囊括了所有信息,缺点是信息维度高和信息冗余。为了解决以上问题,综合分析了当期缺失因子之间的相关性及前后两期因子之间的相关性并且对五期相关系数取平均,按照相关系数进行降序排列,依此加入因子构建训练集。该方法存在的问题是如果因子之间相关性结果和随机森林分类特征因子重要性判定结果存在较大差异,那么该选择方式得到的结果将不是最优。用第三种对所有显著相关因子随机抽样 1000 次构建训练集的方式解决该问题。但是随机抽样仍然会存在组件差异,即相关因子的不同区间不一定会产生相同的训练集数量,因此我们又采用了第四种随机分层抽样的方式来保证对于每一层(本研究中设定每 10 个因子为一层,每层随机选取 100 次训练集)得到相同数量的训练集组合。

2.2.2 确定最优随机森林分类模型参数及检验模型外部有效性

选择合适的特征因子和决策树对于提高随机森林分类模型的预测性能至关重要[39—41],本文根据准确度

来确定最终的特征因子数据集,根据袋外误差(Out-of-bag Error)来确定决策树的数量。准确度作为最常见的分类模型性能指标之一,衡量了模型正确预测的样本数量占总样本数量的比例,能够直观地表示模型的总体精度。根据准确度得到最优的特征因子组合后,设定随机森林分类模型决策树的数量为 100 到 1000 之间的不同组合,运行由最优特征因子组合的随机森林分类模型,根据袋外误差最小值确定随机森林分类模型决策树数量。

同时在省级和县级尺度验证随机森林分类模型填补缺失因子的外部有效性,对比不同尺度和地区模型精度验证的差异。由于森林资源连清每一期的清查因子都基本保持一致,故此次外部有效性的检验不考虑替换特征变量的情况。在省级尺度上,首先将华东地区得到的最优模型的特征因子集合应用到单个省份,保持其余参数一致运行模型得到每个省份每个缺失因子的填补准确度。然后,分别对比单个省份和华东地区的模型填补的准确度,单个省份和华东分省统计的模型填补准确度(即根据华东地区整体模型得到所有样本的准确度后再根据省份进行分区统计其准确度)。在县级尺度上,第一采用不重复随机抽样的方式,对于每一个缺失因子,分别随机抽取华东地区不少于 15 个县(市、区)进行缺失因子填补和精度检验。其次保存县级尺度所有模型的精度检验结果,计算模型在不同县的填补准确度的离散系数(标准差/均值),对比不同地理单位的模型精度验证结果。将这两个尺度的分析结果用于检验本研究构建的随机森林分类模型的外部有效性。

2.2.3　特征因子重要性分析

最后,将以上得到的最优特征因子和决策树数量的组合用于最终随机森林分类模型参数的设定,填补缺失因子并验证模型结果的精度,最终输出每个缺失因子对应特征因子重要性的分析结果。本研究基于平均基尼减少量来衡量特征因子重要性[42],该评价指标通过计算特征因子在随机森林分类模型中用于划分节点时减少的基尼系数的平均值来评估特征的重要性。特征因子重要性大小反映该特征对于模型性能的影响程度,数值越大表示该特征对于缺失因子的解释能力越强。本研究特征因子重要性分析使用 R 语言 4.2 版本的公开程序软件包“randomForest”[42]中的“importance”函数实现。

3　研究结果

3.1　因子间相关性排序及不同模型精度对比

缺失因子和不同因子之间的相关性分析结果显示缺失因子和当期因子之间的相关系数普遍高于后期因子(图 3)。本文最终结果呈现了与缺失因子相关性排名前十的显著($P < 0.05$)相关因子。整体而言,当期因子占林分结构排名前十的显著因子的 64.0%,后期因子占比为 36.0%,除此之外,每个缺失因子对应的相关系数排第一的因子都是当期因子,相关系数的均值为 0.750,显著高于后期因子相关系数的均值 0.561。不同缺失因子之间的相关性系数均值差异显著,具体表现为植被类型(0.868)> 树种结构(0.733)> 更新等级(0.498)> 自然度(0.435)> 森林群落(0.376)。由此可见,植被类型因子的平均相关系数遥遥领先于其他 4 个缺失因子,并且其对应的最高相关系数也显著高于其他类别。树种结构因子的平均相关系数最低。

分层随机抽样的选择方式得到的随机森林分类模型的精度验证结果最优(图 4)。整体而言,分层随机抽样的选择方式下,有 80.0%的模型的准确度均高于前三种选择方式。随机抽样和分层抽样得到的模型精度验证结果十分接近,第一种全部加入所有因子构建模型的方式和后三种得到的结果相对而言后者的精度验证结果会更优。总的来说,随机森林模型在对林分结构因子的预测上性能良好且不同模型之间精度验证结果的一致性较高。

3.2　最优模型精度及外部有效性验证

本文构建的随机森林分类模型性能整体表现出色(图 5)。所有林分结构缺失因子的模型填补结果的准确度均可达到 0.770 及以上。在不同因子之间,植被类型、更新等级和树种结构的模型填补准确度较高,其中前两个因子的准确度达到 0.900 以上。分省统计的结果进一步显示出随机森林模型填补结果的高度一致性和稳健性。除了自然度因子波动较大之外,其余 4 个因子中 89.3%的省份模型填补的准确度与模型整体准确

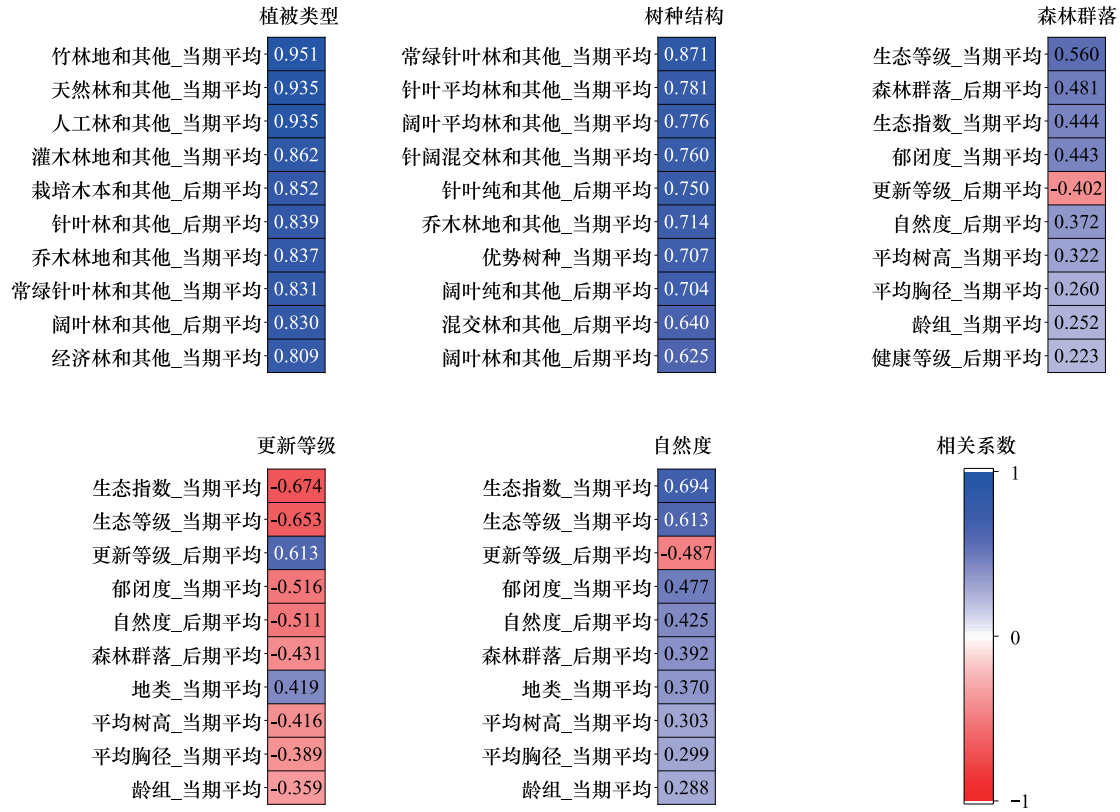

图 3 缺失因子及对应特征因子相关性系数排序

Fig.3 Correlation coefficient ranking of missing factors and corresponding feature factors

由于缺失因子当期没有调查数据,所以和缺失因子名称一致的相关因子的相关系数计算采用该因子对应的后一期值;该图所有相关系数的显著性 $P<0.05$

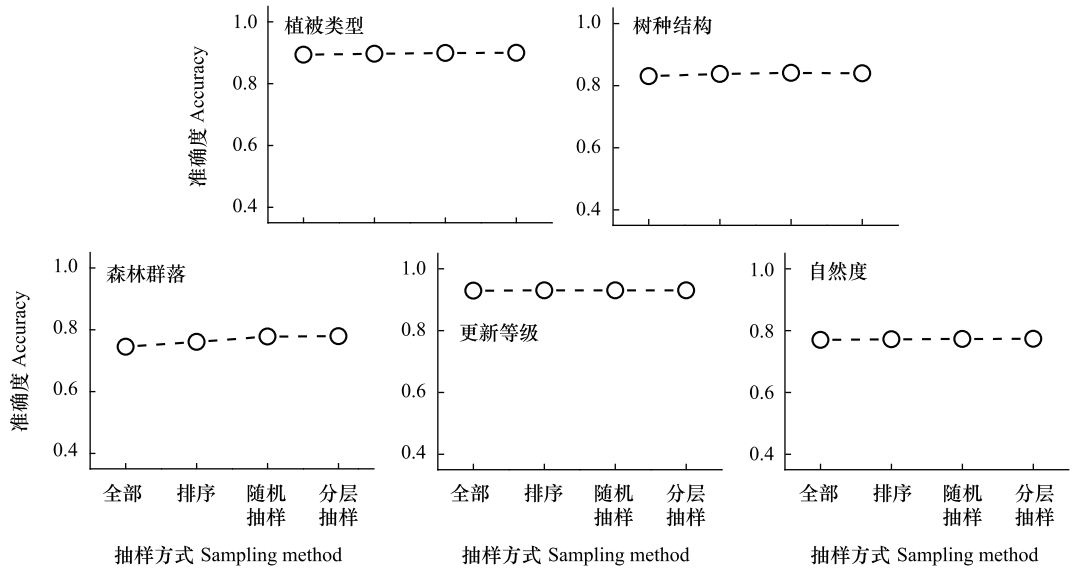

图 4 基于不同抽样方式的随机森林分类模型填补性能对比

Fig.4 Performance comparison of random forest model imputation based on different sampling selection methods

度之间的差异均在 10.0%以内,说明模型填补性能的内部一致性较高。整体模型准确度较高的因子其分省统计的准确度具有较高的稳健性,尤其是植被类型因子,其各个省份的平均差异百分比仅为 4.8%。此外,准确度高于模型整体准确度的占比为 60.0%。对于自然度因子而言,分省统计准确度波动最大的省份是福建省,其模型填补的准确度仅为 0.402,由于福建省样本个数占华东地区总体的 28.3%,该省份较低的准确度也在一定程度上影响了该因子的模型在华东地区的整体准确度。

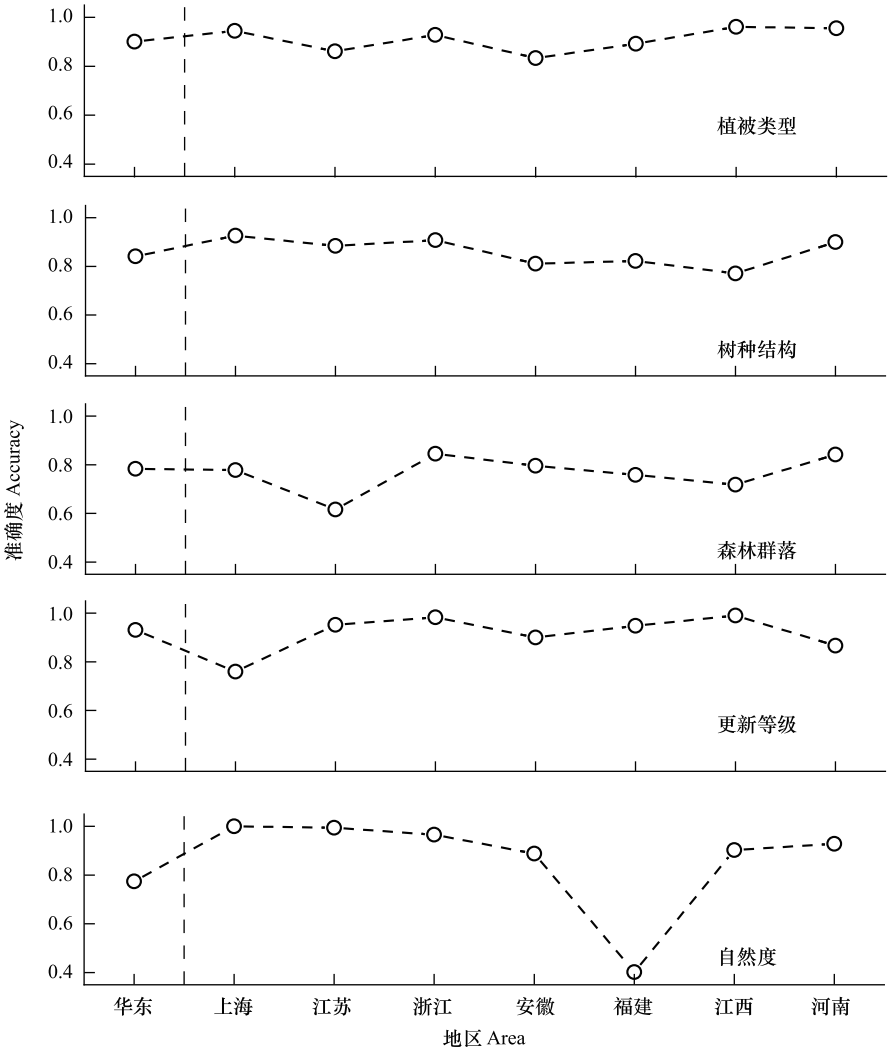

图 5　随机森林分类模型填补精度验证

Fig.5　Imputation accuracy validation of the random forest classification model

随机森林分类模型在省级和县级尺度的交叉检验结果说明我们构建的模型外部有效性高,并且模型填补性能稳健。每个缺失因子在省级层面的模型填补准确度及所有省份准确度的平均结果如图 6 所示。首先,根据省级层面模型填补结果的准确度计算得到的华东地区的准确度均值和华东地区整体模型得到的准确度基本保持一致,差异均在 5.0%以下。并且,省级层面 91.4%的单个模型填补结果的准确度和华东分省统计的准确度的差异均小于 10.0%,80.0%的单个模型填补结果的准确度和华东分省统计的准确度的差异均小于 5.0%。此外,省级层面所有缺失因子的地区整体均值均略微低于华东地区整体模型的准确度,具体表现为植被类型(-1.0%),树种结构(-2.5%),森林群落(-3.9%),更新等级(-0.4%),自然度(-1.0%)。在自然度因子的填补性能评估中,模型在福建省的填补准确度依旧最低,为 0.402,与华东分省统计的结果一致。森林群落和树种结构因子在上海市的填补准确度较低,并且与华东分省统计之间的差异较大。

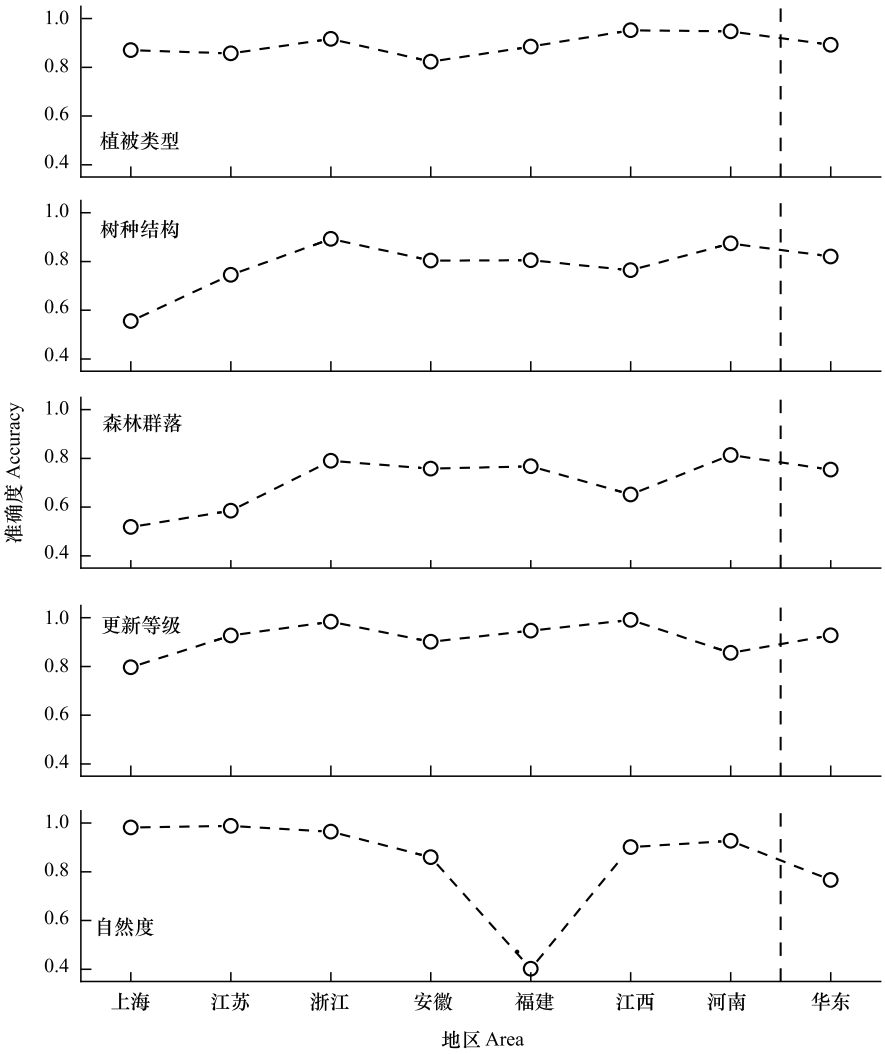

图 6 省级尺度随机森林分类模型填补精度验证

Fig.6 Imputation accuracy validation of the random forest classification model at provincial level

县级尺度 73.3%的模型的准确度均大于 0.750,并且在华东地区填补结果准确度较高的模型在县级尺度上也会得到相对一致的结果(图 7)。所有缺失因子在县级尺度的模型准确度和华东地区和根据省级层面单个模型计算得到的华东地区均值准确度之间的差异均小于 10%。在县级外部有效性检验中模型填补准确度达到 0.900 以上的因子为更新等级,其平均准确度为 0.911±0.047,该因子在华东地区和省级层面的填补性能准确度也均在 0.900 以上。同时,80.0%的缺失因子模型在不同县(随机抽样>15)的填补准确度的离散系数均小于 0.2,表明模型填补准确度在不同县之间的差异较小。省级和县级两个地理尺度的结果均一致说明本研究所采用的模型得到的缺失因子填补性能稳健,并且有较高的外部有效性,可以有效用于填补其他地区的森林连清缺失数据。

3.3 特征因子重要性

缺失因子本身对应的后一期数值和特征因子组合中当期因子的占比高有助于提高模型填补性能。特征因子重要性排序的结果如图 8 所示,针对每个缺失因子,取其对应平均减少基尼系数排名前 10 的特征因子展示重要性。对于大部分的缺失因子而言,缺失因子的后一期数值对提高模型性能的贡献最大。除此之外,生态等级和生态指数这两个因子对提高所有缺失因子模型的填补性能贡献都一致较高。此外,优势树种、林地起源和郁闭度对提高林分结构因子模型性能的贡献度也较高。通过对比不同缺失因子对应特征因子的调查

时间,进一步发现当期特征因子会比后期特征因子更能提高模型填补性能。结合图 4 中模型填补准确度均大于 0.900 的两个缺失因子(更新等级和植被类型)进行对比分析,发现对填补这两个缺失因子重要性排在前几位的特征因子中当期因子占绝大部分,并且整体而言特征因子组合中当期因子的比例均高于后期因子。

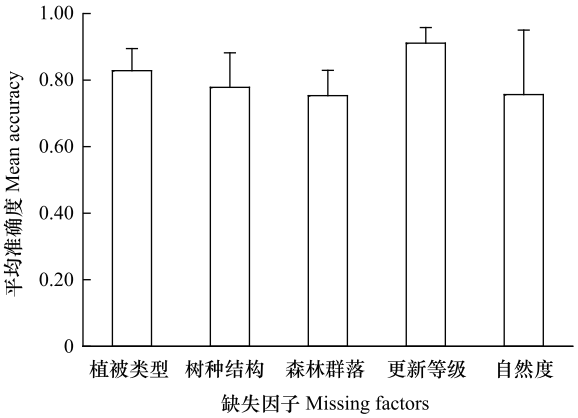

图 7　随机森林分类模型外部有效性验证

Fig.7　External validity assessment of random forest classification models

图中的误差棒为标准差(Standard deviation)

4　讨论

本研究利用随机森林分类模型填补森林资源连清数据库中 5 个林分结构缺失因子。整体而言,随机森林分类模型在填补缺失因子上性能表现良好,所有缺失因子的模型填补准确度均能达到 0.770 以上,此外针对部分指标,模型填补结果的准确度可以达到 0.900 以上,并且在省级和县级外部有效性的检验上展示出良好的泛化能力。不过,的模型在个别因子的分省测试上表现不佳以及个别因子在省级尺度测试和华东分省统计上差异较大。以下将对个别因子填补性能不佳的原因及解决对策,如何将本研究结果用于科学评估我国森林生态保护建设成效及未来如何完善我国森林分类经营管理制度三方面进行讨论和展望。

4.1　因子模型性能不佳的原因及解决对策

福建省的缺失因子自然度在省级尺度模型和华东分省统计下均性能不佳,以及上海市的森林群落和树种结构两个因子的省级尺度模型和华东分省统计的差异较大,其主要原因是模型所能捕捉到的特征信息相对有限,再加上训练样本不足和数据结构的不平衡,共同导致了个别模型在泛化能力上的限制。影响随机森林模型填补性能的关键因素有特征的选择、样本质量、数据的分布和模型的参数选择[43]。在设计模型初期,本研究充分考虑了关键因素,采用了多种规则来选取特征变量,以确保模型在给定数据条件下达到最佳性能。然而,由于基于经验数据的分析,样本量、数据分布以及可用有效特征信息都是固定的,这导致了不同缺失因子之间模型信息捕获的差异。填补五个缺失因子(更新等级、植被类型、树种结构、森林群落和自然度)对应的特征变量集数量分别是 115、62、59、53 和 26。模型填补准确度的排序与特征变量数量呈正相关。特征数量的差异部分解释了为何自然度、森林群落和树种结构的模型填补准确度相对较低。另外,上海市在分省建模时的样本数量较少,仅为 54 个,进一步降低了模型填补性能。除了特征因子数量的影响外,数据不平衡进一步降低了福建省自然度因子的模型填补准确度。以华东整体模型为例,福建省训练集和测试集之间不同类别的百分比差异最大为 66.5%,比其余 6 个省市的百分比差异高出 2.9 倍。在条件允许的情况下,使用更广泛范围的训练样本来建立综合模型,可以有效提升单独建模的填补准确度(以本研究省级尺度模型和对华东分省统计模型的准确度对比为基础)。此外,在时间维度增加福建和上海的特征数量也将有助于提高模型的填补准确度。

4.2　模型结果的应用场景及展望

本研究除了可以为全面评估我国森林资源动态变化提供数据支撑外,也可以用于科学评估我国森林生态保护建设成效。自 20 世纪 90 年代开始,我国陆续推行了一系列林业保护工程,其中天然林保护工程和退耕还林工程因其巨额的资金和人力投入、长时间和大尺度的实施,在全球范围内引起了广泛关注[44]。目前已有大量研究定量评估了这两个工程的生态和社会经济效应[45]。但是,大部分的论文在生态效应的评估上都是使用土地覆被指标[6,46],少部分研究涉及到政策对生态系统服务的影响[47],鲜有研究能够全面地揭示宏观政策对生态服务变化的因果机制。开展政策对生态系统服务的因果机制研究有两部分的难点,第一是基础数据

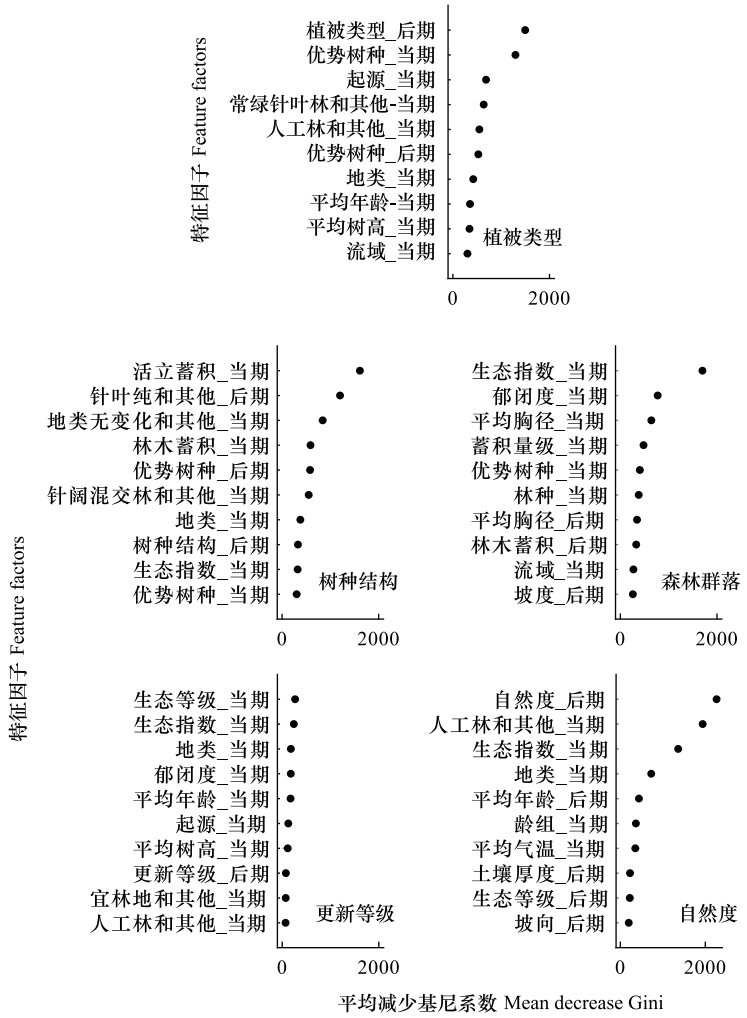

图 8 随机森林分类模型特征因子重要性

Fig.8 Feature attributes importance of random forest classification models

的可得性,第二是严密的实验设计[10]。本研究填补的缺失因子中的植被类型和树种结构既是开展以上研究必要的基础数据,也是在构建严密实验设计时建立可靠的反事实评估基准线的必要数据。

科学评估我国森林生态保护建设成效是完善我国森林分类管理制度的重要环节,也是贯彻落实森林可持续经营理念的关键步骤。自 20 世纪 60 年代颁布《森林保护条例》至今,我国的林业发展顺应国家经济社会发展大势,经历了以木材生产为主向以生态建设为主的历史性转变,如今走向了高质量发展之路。2019 年新修订的《森林法》强调充分发挥森林的多种功能,明确商品林和公益林的不同经营管护制度,进一步强调公益林的生态主导功能和提高公益林的经营质量的重要性。同时强调完善森林生态效益补偿制度为国家林业建设发展提供制度保障。在未来的研究中,本研究将基于完善后的森林资源清查数据库,积极整合不同部门的多种数据源,立足于国家森林高质量发展的基本策略,以严密的实验设计评估我国森林分类经营制度的建设成效,评估成果将用于完善我国森林生态效益补偿制度,以期为建立稳定健康优质高效的森林生态系统提供技术支撑。

参考文献(References):

[1] 自然资源部办公厅. 国家林业和草原局办公室关于统筹推进 2021 年度全国森林资源调查监测和林草生态综合监测评价工作的通知. (2021-11-18)[2023-8-17]. https://www.gov.cn/zhengce/zhengceku/2021-11/18/content_5651615.htm.

[2] Zeng W S, Tomppo E, Healey S P, Gadow K V. The national forest inventory in China: history-results-international context. Forest Ecosystems, 2015, 2(1): 23.
[3] 国家林业局. 中国森林资源及其生他功能四十年监测与评估. (2018-06-22)[2023-8-17]. https://cfern.org/portal/article/index/id/12654.html.
[4] 陈娟. 山东省森林资源清查体系发展的思考. 林业科技情报, 2023, 55(1): 137-140.
[5] 李忠平. 森林资源连续清查体系优化问题的思考. 林业建设, 2014(6): 1-3.
[6] Zhou T, Shen W W, Qiu X, Chang H, Yang H B, Yang W. Impact evaluation of a payments for ecosystem services program on vegetation quantity and quality restoration in Inner Mongolia. Journal of Environmental Management, 2022, 303: 114113.
[7] Yang W, Liu W, Viña A, Luo J Y, He G M, Ouyang Z Y, Zhang H M, Liu J G. Performance and prospects of payments for ecosystem services programs: evidence from China. Journal of Environmental Management, 2013, 127: 86-95.
[8] Ouyang Z Y, Zheng H, Xiao Y, Polasky S, Liu J G, Xu W H, Wang Q, Zhang L, Xiao Y, Rao E M, Jiang L, Lu F, Wang X K, Yang G B, Gong S H, Wu B F, Zeng Y, Yang W, Daily G C. Improvements in ecosystem services from investments in natural capital. Science, 2016, 352(6292): 1455-1459.
[9] 李文华, 李芬, 李世东, 刘某承. 森林生态效益补偿机制与政策研究. 生态经济, 2007(11): 151-153, 159.
[10] 杨武, 陆巧玲, 周婷. 生态保护项目绩效评估的技术方法体系. 生态学报, 2020, 40(5): 1779-1788.
[11] Barrett T, Maltamo M. Missing data in forest ecology and management: advances in quantitative methods. Forest Ecology and Management, 2012, 272: 1-2.
[12] Eskelson B N I, Temesgen H, Lemay V, Barrett T M, Crookston N L, Hudak A T. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research, 2009, 24(3): 235-246.
[13] Reams G A, Roesch F A, Cost N D. Annual forest inventory-Cornerstone of sustainability in the South. Journal of Forestry, 1999, 97(12): 21-26.
[14] Rubin D B. Multiple imputation after 18+ years. Journal of the American Statistical Association, 1996, 91(434): 473-489.
[15] Lipsitz S R, Zhao L P, Molenberghs G. A semiparametric method of multiple imputation. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1998, 60(1): 127-144.
[16] Chen G Y, Åstebro T. How to deal with missing categorical data: test of a simple Bayesian method. Organizational Research Methods, 2003, 6(3): 309-327.
[17] Little R J A, Rubin D B. Statistical Analysis with Missing Data. New Jersey: Wiley, 2002.
[18] LeMay V, Temesgen H. Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Science, 2005, 51(2): 109-119.
[19] Stage A R, Crookston N L. Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. Forest Science, 2007, 53(1): 62-72.
[20] Tang F, Ishwaran H. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2017, 10(6): 363-377.
[21] 金勇进. 处理缺失数据中辅助信息的利用. 统计研究, 1998, 15(1):43-45.
[22] 庞新生. 缺失数据处理中相关问题的探讨. 统计与信息论坛, 2004, 19(5): 29-32.
[23] 金勇进, 朱琳. 不同差补方法的比较. 数理统计与管理, 2000, 19(4): 50-54.
[24] 乔珠峰, 田凤占, 黄厚宽, 陈景年. 缺失数据处理方法的比较研究. Agent 理论与应用学术会议, 2006.
[25] 梁怡. 缺失数据的插补调整方法. 西安文理学院学报: 自然科学版, 2009, 12(1): 74-76.
[26] 胡玄子, 陈小雪, 钱叶亮, 姜正龙, 赵彤洲. 数据处理中缺失数据填充方法的研究. 湖北工业大学学报, 2013, 28(5): 82-84.
[27] 靳国栋, 刘衍聪, 牛文杰. 距离加权反比插值法和克里金插值法的比较. 长春工业大学学报: 自然科学版, 2003, 24(3): 53-57.
[28] 刘菲, 李明阳, 刘雅楠, 江一帆, 王子. 森林资源抽样调查缺失数据填充方法. 林业资源管理, 2018(6): 130-137.
[29] 国家统计局. 中国统计年鉴 2004—2020. 北京: 中国统计出版社, 2004—2020.
[30] 周生贤. “东扩、西治、南用、北休”——相持阶段林业发展区域战略方针. 人民论坛, 2005(10): 25-26.
[31] 郝天象, 王兵, 牛香, 刘世荣, 于贵瑞. 全面提升我国森林生态系统质量和稳定性的实践与思考. 陆地生态系统与保护学报, 2022, 2(5): 13-31.
[32] 国家林业和草原局. 国家森林资源连续清查技术规程. 北京: 国家市场监督管理总局, 国家标准化管理委员会, 2020.
[33] Fox E W, Hill R A, Leibowitz S G, Olsen A R, Thornbrugh D J, Weber M H. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environmental Monitoring and Assessment, 2017, 189(7): 316.
[34] Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Statistics and Computing, 2017, 27(3): 659-678.
[35] Akoglu H. User's guide to correlation coefficients. Turkish Journal of Emergency Medicine, 2018, 18(3): 91-93.

[36] Meyer D, Zeileis A, Hornik K, Friendly M. vcd: Visualizing Categorical Data. R package version 1.4-12, 2023.

[37] Richardson J T E. Eta squared and partial eta squared as measures of effect size in educational research. Educational Research Review, 2011, 6(2): 135-147.

[38] Speiser J L. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. Journal of Biomedical Informatics, 2021, 117: 103763.

[39] Chen R C, Dewi C, Huang S W, Caraka R E. Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 2020, 7(1): 52.

[40] Boehmke B, Greenwell B. Hands-On Machine Learning with R. New York: Chapman and Hall/CRC, 2019.

[41] Karabadji N E I, Korba A A, Assi A, Seridi H, Aridhi S, Dhifli W. Accuracy and diversity-aware multi-objective approach for random forest construction. Expert Systems with Applications, 2023, 225: 120138.

[42] Liaw A, Wiener M. Classification and regression by random forest. R News, 2002, 2(3): 18-22.

[43] Breiman L. Random forests. Machine Language, 2001, 45(1): 5-32.

[44] Liu J G, Li S X, Ouyang Z Y, Tam C, Chen X D. Ecological and socioeconomic effects of China's policies for ecosystem services. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(28): 9477-9482.

[45] Ma Z H, Xia C Q, Cao S X. Cost-benefit analysis of China's natural forest conservation program. Journal for Nature Conservation, 2020, 55: 125818.

[46] Zhao A, Zhang A, Liu J, Feng L, Zhao Y. Assessing the effects of drought and "Grain for Green" Program on vegetation dynamics in China's Loess Plateau from 2000 to 2014. Catena, 2019, 175: 446-455.

[47] Huang L S, Wang B, Niu X, Gao P, Song Q F. Changes in ecosystem services and an analysis of driving factors for China's Natural Forest Conservation Program. Ecology and Evolution, 2019, 9(7): 3700-3716.