自然资源学报, 2019, 34(6): 1331-1344
Journal of Natural Resources
http://www.jnr.ac.cn
DOI: 10.31497/zrzyxb.20190617

基于随机森林的耕地利用效率测度模型构建及其应用

陈丹玲[1],卢新海[2],匡 兵[2]

(1. 华中科技大学公共管理学院,武汉 430074;2. 华中师范大学公共管理学院,武汉 430079)

摘要: 构建合适的量化分析模型是科学把握耕地利用状况及利用效率的基础性工作,可为制定合理有效的耕地资源管控政策,实现耕地利用与生态环境的协调发展提供决策依据。为了更准确地反映耕地利用系统的复杂性、动态性及差异性等特征,鉴于随机森林的基本思想,运用随机抽样Bootstrap法在合理构建分类树的基础上,构造了耕地利用效率测度的RF模型,进而以中国粮食主产区172个城市为例训练该模型并将其运用至2003-2015年的耕地利用效率测度中,同时将BP神经网络和熵权法作为对比验证其一致性、代表性和优越性。结果表明:(1)耕地利用效率测度的RF模型不受量纲限制,运行所需参数少,运算过程简化,能够较为精确地模拟各评价指标间的复杂联系,科学量化各评价指标对耕地利用效率的贡献。(2)对同一空间单元的效率值而言,RF > BPNN > EW,RF与BPNN所得效率值的总体分布格局相似,且均与EW的测度结果存在较大差异。(3)从评价结果与现实的匹配度和精度表征参数来看,RF的测度结果与自然和社会经济发展等客观事实更相符,具有较高的适用性与可靠性。同时,与其余两种常用模型相比,RF能够降低计算复杂度,提高训练效率,其测度结果的相关系数R为0.8685,M_{RPD}为2.3533,且具有最小M_{MSE}0.0174和M_{MAE}0.0211,更适用于复杂非线性特征的耕地利用效率研究。

关键词: 耕地利用效率;随机森林;粮食主产区

科学测度耕地利用效率是考察耕地利用状态、测度耕地产出绩效和对耕地系统安全进行预警的核心实践内容[1]。随着粮食安全战略的深入推进，耕地资源对人类社会经济活动及农业可持续性的约束日益凸显，耕地利用粗放、生态环境恶化的空间格局初现端倪[2]。在这一现实背景下，耕地利用效率的理论研究与实践探索面临一系列挑战，厘清并客观把握耕地利用效率及现状成为经济学、土地科学及环境科学等学科研究的核心话题。

长期以来，国内外学者围绕耕地利用效率测度的原理和方法开展了大量的研究，其中常规方法以基于投入指标[3]、产出指标[4]、潜力指标[4]等的单指标测度法以及基于复合指标体系的熵权法（Entropy Weight，EW）[5]、AHP法[6]、PCA法[7]、模糊综合评价法[8]、DEA模型[2]等为代表。这类方法均具有数据兼容性强，模型简单且应用广泛等特点。AHP法的主观性较大，指标相对重要性难以权衡；EW、PCA法是客观赋权，不受主观判断影响，但未从指标的物理意义融入评价者的价值判断[9]；模糊综合评价法、DEA模型存在高维数据不易处理及稳健性弱等缺陷，难以深入挖掘指标数据间非线性和非正态

收稿日期：2018-11-12；**修订日期**：2019-03-18
基金项目：国家自然科学基金项目（71673096）；国家社会科学基金项目（16CGL054）
作者简介：陈丹玲（1993- ），女，江苏徐州人，博士研究生，研究方向为土地利用与管理。
E-mail: hustcdl93@163.com
通讯作者：卢新海（1965- ），男，湖北洪湖人，教授，博士生导师，研究方向为土地资源管理与粮食安全。
E-mail: xinhailu@163.com

的分布信息[10]。近年来，以机器学习为理论基础的人工智能算法开始运用于评价模型的构建中。其中，人工神经网络和支持向量机是该类方法中的典型。相比于常规赋权的综合评价方法，人工神经网络的突出特点是对大量高维度、非结构化、非线性和不确定性数据具有降维分解、并行处理和优化计算的能力。同时，它能够对指标间复杂的内在关联进行准确捕捉和评判，拟合并衡量评价对象与评价指标数据间的函数关系，在耕地利用效率评价中的应用也逐渐趋于成熟，但它在研究过程中也暴露出局部最优、网络结构难以确定、泛化误差大等问题[11]。与人工神经网络不同的是，支持向量机对非结构化数据具有鲜明的自适应特征，在评价对象上具有较强的泛化能力。同时，它能够将评价问题等价转化为凸优化问题，进而依据数据分布结构确定权重，避免了局部极值问题，这些都是其相较于人工神经网络取得的进步，但它在研究过程中也存在难以克服的缺陷，例如参数的经验选取、过学习以及对样本的划分状况直接影响其进一步应用[7]。可见，如何将多元指标数据有效统一到同一个评价单元仍是耕地利用效率测算研究的重点和难点问题。

随机森林（Random Forest，RF），是由Leo[12]提出的利用树形分类器（Classification and Regression Tree，CRAT）进行组合分类的机器学习算法，也称之为随机决策树。作为一种智能建模工具，RF不受量纲限制，具有极强的数据挖掘能力和较高的预测准确率，能够根据有限的训练样本以最优参数和最小误差实现较高的分类准确率并建立多个变量间的权重学习机制，进而解决复杂、非线性大系统内某一属性评价的“过拟合”问题[13]。此外，耕地利用效率是土地、社会经济及生态要素交互耦合的复杂巨系统，其测度体系具有复杂性、非结构性以及随机不确定性，需要更为稳健和灵活的测度方法来处理非线性关系、高阶相关性甚至是缺失值。同时，随着时空的推移，各指标对耕地利用系统的影响程度可能发生变化，初始权重不一定符合实际情况，进一步推动了测度模型向非参数化发展。而RF作为一种非参数树形模型，兼具以往复合指标体系测算方法的所有优点，且在多变量非线性关系及权重动态性的处理上具有更为优越的性能，可以防止由于训练样本存在噪声和数据缺失引起的精度降低，理论上能够成为耕地利用效率测度的理想工具。基于此，本文构建耕地利用效率测度的随机森林模型，以2003-2015年中国粮食主产区172个城市为研究样本，基于市域尺度测度耕地利用效率水平，并将目前测度耕地利用效率较为成熟的人工智能技术——BP神经网络模型（Back Propagation Neural Network，BPNN）和经典的EW进行比较，以此验证RF模型的可靠性和优越性，以期扩充耕地利用效率评价方法库，进而引起更多学者关注RF在解决耕地利用测度中的应用价值。

1 耕地利用效率测度的RF模型构造

1.1 基本原理

构建耕地利用效率测度的RF模型的核心是揭示各指标与测度结果间的对应规则，这个工作由分类树承担。分类树由根节点、子节点和叶节点三部分组成，其中，根节点代表指标变量的观测值，叶节点代表指标变量的线性分解，根节点到叶节点的路径对应分解规则，叶子就是没有下一分支的节点，分类树的生长过程代表样本集的训练过程。RF的基本原理[14]如图1所示，首先利用元分类器创造一片“森林”，通过自助法随机选择观

测数据生长成为分类树群，每棵树都会完整生长而不被修剪。在生成树的时候，通过自助法选择各指标变量的观测数据制造差异性、随机化的训练样本集，将训练数据集输入到分类树中形成分类器，采用随机森林法则去分裂每一个节点，把与评价对象具有非线性关系的各指标变量值分解为潜在的、存在线性关系的叶子，最多投票的结果返回作为指标分解结果；其次，对叶子的分布结构进行分析以提取并计算各分解指标变量对评价结果的贡献值（权重），将各分解指标权重按照分类树线性映射规则形成指标变量与权重间一一对应的关系集并输出RF模型，各叶子节点上指标变量的RF特征值与其对应权重值的综合加权的平均值即为所求结果。

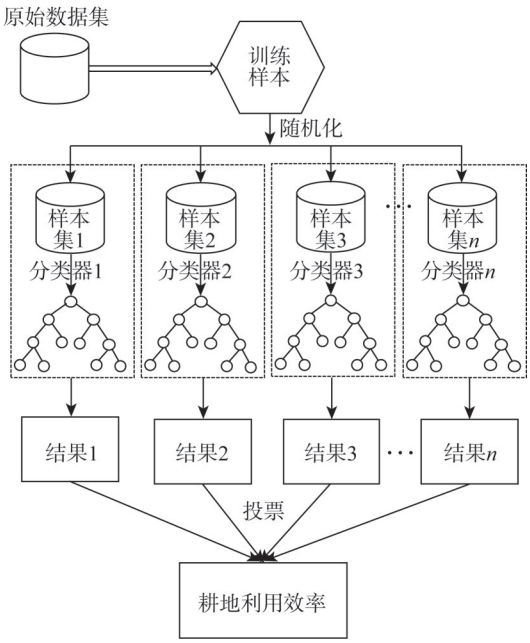


图1　RF的基本思路

Fig. 1　The basic principle of RF

1.2 构造过程

第一步，分类树构建

RF在具体实现时需要预先设置两个极为重要的参数，分类树的棵数（k）和每棵分类树构建时最优节点拆分次数（m），前者决定了随机森林的整体大小，后者决定了单棵分类树的状况。这两个参数必须优化以提高模型在数据处理中的精度。构造包含 k 棵耕地利用效率分类树模型的具体过程如下（图2）：（1）获取训练集。将耕地利用效率评价的原始训练样本集记为 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，采用Bootstrap抽样法从 T 中三分之二的样本形成一个随机向量序列 T_t，重复 k 次形成 k 个独立同分布的训练集 $\{T_k,t=1,2,\cdots,k\}$。（2）随机选取节点特征指标。对 k 个数据集分别建立不进行Pruned处理的深度最大的分类树，通过计算每个指标变量蕴含的 *Gini* 指数，节点 n 的 *Gini* 指数定义为：$Gini(n)=\sum_{i\neq j}p(w_i)p(w_j)=1-p^2(w_j)$，式中，$p(w_i)$ 是第 n 个节点上属于第 i 类样本个数占训练样本个数的频度。（3）确定分裂节点。对 *Gini* 指数最大的候选属性进行分裂，并重新计算 *Gini* 指数。重复分裂步骤直到 *Gini* 指数小于预定阈值，最终形成具有 k 棵数的决策“森林”。（4）递归分类。反馈 k 棵数的决策结果，按照票数最多原则确定指标变量的最佳线性分解模式。

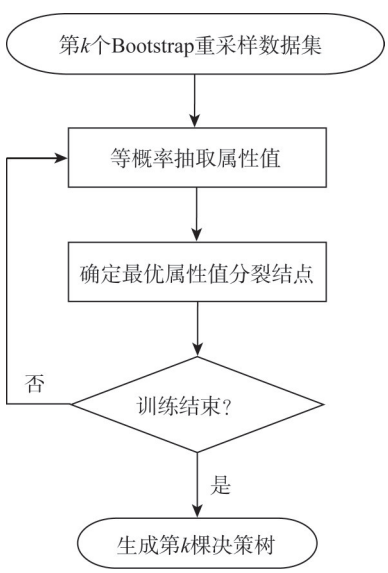


图2　分类树构建

Fig. 2　Construction of decision trees

第二步，RF权重计算

随机森林通过构造不同的训练集增加分类模型间的差异，从而提高组合分类模型的外推预测能力。在决策

树分类过程中只运用了原始数据集的部分数据量，剩余数据并没有发挥作用，为了避免算法过拟合，RF以这部分袋外数据（Out-of-bag，OOB）为基础，运用基尼值法计算各分解指标重要性，具体实现步骤主要包括：假设第i棵分类树原分割基尼指数为$Gini\ e_i$，对袋外数据第j个分解变量的属性值X进行随机序列变化，重新计算新的基尼指数为$Gini\ e_i^j$，则属性X在相应单棵分类树的权重表示为$Gini\ e_i$-$Gini\ e_i^j$，第j个分解指标的动态权重RFW为森林中所有分类树$Gini$指数的均值，计算公式如下：

$$\Delta_j = \left| \sum_{j=1}^{N} \left(Gini_{e_i} - Gini_{e_i^j} \right) \Big/ N \right|$$

$$RFW_j = \Delta_j \Big/ \sum_{j=1}^{N} \Delta_j$$

式中：Δ_j为$Gini$指数的减少值；N为指标个数；RFW_j为第j个分解指标的权重值，满足$\sum_{j=1}^{N} RFW_j = 1$。

第三步，加权组合

首先，输出各分解变量的RF权重，并按照分类树线性映射规则，生成各指标变量与响应变量（耕地利用效率）间的最优模拟关系；然后将待测样本集输入到各训练完毕的RF分类树中，各叶子节点上的特征值与其权重值的综合加权即为该棵树的耕地利用效率，最后，各棵分类树叶子节点上耕地利用效率的平均值即为所求的耕地利用效率值。

1.3 对比方法

基于多维度、多指标综合评价体系的耕地利用效率测度方法众多，但从方法性质和理论基础考虑，大致可归纳为常规赋权评价法和机器学习法，因此，为了更加全面地验证RF模型的应用效果和可靠性，本文从两类中各选取一种评价方法或模型作为对比。

第一，EW。从评价方法和过程的复杂性来看，相比于其他常规赋权评价法，EW的赋权最为基础和简单，且不需要多次循环一致性检验，评价结果相对稳定，因此本文选择EW作为常规赋权评价法的典型代表。其基本原理是通过考察数据变化的趋势和状态，根据评价指标提供有效信息量的多寡程度来确定其权数，进而将多维指标的个体指数直接相加得出综合评价数值的方法，由于该方法的应用较为成熟，其计算步骤参考文献 [5]，文中不多赘述。

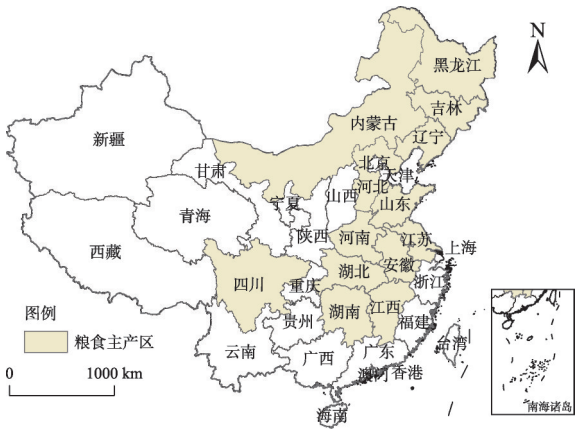
第二，BPNN。BPNN和支持向量机作为机器学习法的典型代表，则需要进行模型的复杂训练，但总体上两者皆比常规赋权评价法的稳定性强。然而，支持向量机虽具有公认的优越性，但参数选择是其本身难以解决的问题，这极大增加了支持向量机的应用难度，其深入研究也广受限制。相比之下，耕地利用效率评价的神经网络结构逐渐趋于成熟，且在实际应用中，大多数研究采用的是BPNN或其变形形式。因此本文选择BPNN作为机器学习法的经典代表。理论上来说，一个3层的BPNN能实现任意的连续映射，因此本文尝试利用Matlab 2015 a工具箱构建、训练耕地利用效率测度的3层BPNN结构。

2 研究方法与数据来源

2.1 研究区概况

粮食主产区是中国极其重要的粮食生产基地，对保障国家粮食安全、推动国民经济

发展及巩固农业基础等具有战略性和决定性作用。根据财政部2013年颁发的《关于改革和完善农业综合开发政策措施的意见》对中国的粮食主产区的分类和划分，确定了中国粮食主产区包括黑龙江、吉林、辽宁、内蒙古、河北、河南、山东、安徽、江苏、湖北、四川、湖南、江西等13个省（自治区），横跨东北、黄淮海和长江中下游3大地区（图3）。据统计数据显示，2015年，粮食主产区内耕地总面积约88.98万km^2，以约65.91%的耕地面积生产了全国77%以上的粮食。随着中国经济发展格局的调整和发展方式的转变，粮食主产区面临粮食增产潜能减弱与刚性需求增加的双重挑战，与此同时，耕地地力透支、土壤退化和面源污染加重等问题却进一步阻碍了粮食增产步伐[15]。因此，准确认识、把握进而提高该区域的耕地利用效率水平，对提高粮食生产能力和保障能力意义重大。本文以粮食主产区内的172个地级市为单位进行分析研究。



注：本图基于国家测绘地理系信息局标准地图服务网站下载审图号为GS(2016)2923标准地图制作，底图无修改。

图3　中国粮食主产区分布图

Fig. 3　Location of the major grain producing areas in China

2.2 指标体系构建及数据来源

第一，指标初选。既有文献中较为成熟的耕地利用效率测度思路主要有两种，第一是主要围绕“投入—产出”视角展开，重点体现合理增加投入并追求经济、社会、生态综合效益的最大化，尤其关注区域耕地利用系统在社会经济上的刻画[2,16]。第二是基于“压力—状态—响应”（Pressure-State-Response，PSR）视角，集中反映耕地利用系统的内部状态和区域人地关系[5-7,17]。两种测算思路下的耕地利用效率指标存在一定重叠，但后一研究视角下的指标体系具有更强的逻辑因果关系和更完整的内容体系。目前，PSR视角下的耕地利用综合评价体系已日臻成熟，对应于这一研究框架，耕地在利用过程中，“压力”表现为耕地利用主体的资本、劳动、科技等生产要素的投入数量及有效性，即投入水平；“状态”强调不同耕地质量及不同耕作制度下耕地利用的程度，即利用强度；“产出”表现为注重耕地利用模式优化背景下产出的提升，即产出效益。此外，在当前耕地数量持续减少的情况下，粮食安全保障尤为重要，因此，可持续发展能力也是考察耕地利用效率的关键。基于上述对耕地利用效率内涵的理解与界定并结合已有研究成果，初步建立由投入强度（压力系统）、利用程度（状态系统）、产出效益（响应系统）、可持续性4个子系统，19个指标构成的耕地利用效率测度指标体系（表1）。

第二，指标关联性分析。运用SPSS 22.0对属于同一个准则层的单项指标两两间进行Pearson相关分析，相关性显著的指标只保留一个，最终从18个指标中筛选出12个评价指标，构成同时适用于EW、BPNN和RF的中国粮食主产区耕地利用效率测度指标体系。

第三，指标体系效度检验。虽然上述构建指标框架（表1）的内容体系较为全面，但不能确定是否存在其他影响耕地利用效率的外部因素。因此，需要对这些指标进行残

表1 耕地利用效率测度指标体系

Table 1 Index system for cultivated land utilization efficiency

准则层	指标层	计算依据	说明
投入水平	地均农业机械投入/(kW·hm^{-2})	农业机械总动力/耕地面积	保留
	地均科技投入/(10^{4}元·hm^{-2})	技术投入总额/耕地面积	删除
	地均劳动力投入/(人·hm^{-2})	农业从业人数/耕地面积	保留
	地均化肥投入/(kg·hm^{-2})	化肥施用量/耕地面积	保留
利用强度	垦殖指数/%	耕地面积/土地总面积[9]	删除
	灌溉指数/%	耕地有效灌溉率	保留
	复种指数/%	农作物播种总面积/耕地面积[5]	保留
	稳产指数/%	旱涝收保面积/耕地面积[9]	保留
	农膜利用强度/(kg·hm^{-2})	农膜使用量/耕地面积	删除
产出效益	粮食单产/(10^{4}kg·hm^{-2})	粮食总产量/耕地面积	保留
	农民人均农业产值/(10^{4}元·人$^{-1}$)	农业产值/农业人口数	保留
	单位面积农业产值/(10^{4}元·hm^{-2})	农业产值/耕地面积	保留
	万元产值能耗/(kwh·元$^{-1}$)	农业用电量/农业总产值	删除
可持续性	人均耕地面积/(hm^2·人$^{-1}$)	耕地面积/总人口	保留
	非农指数/%	非农业人口总数/总人口	删除
	粮食安全系数/(kg·人$^{-1}$)	人均粮食占有量/400 kg	保留
	农业自然灾害成灾率/%	—	保留
	森林覆盖率/%	—	删除

差系数分析。首先，通过如下公式：$R^2=\sum_i^n\left(Y_{pi}-Y_m\right)^2/\sum_i^n\left(Y_i-Y_m\right)^2$ 计算得判定系数R^2=0.9912。式中，Y_{pi}表示各指标的实际值；Y_m表示各指标的平均值；n为城市个数，文中取172。其次，运用$e=\sqrt{1-R^2}$计算残差系数e=0.0938，虽然残余系数不等于0，但它足够小，这意味着除了本文所选用的12个指标之外，其他因素对耕地利用效率的影响小到可以忽略不计。因此，表1中的评价指标体系具有较强的可信性和代表性。

考虑到数据可得性和行政区划变动的原因，本文的研究期限为2003-2015年。基础数据来源于2004-2016年《中国农村统计年鉴》《中国城市统计年鉴》以及粮食主产区13个省、自治区及各地级市的官方统计资料。同时，样本中删除了济源市、神农架林区、天门市、潜江市、仙桃市、眉山市、资阳市、恩施土家族苗族自治州8市（区）。

2.3 RF算法实现及参数设定

第一，训练集生成及学习。训练样本是RF测度算法的核心部分，是建立权重分类树的过程，其目的是通过样本集指标与对应等级间的联系构建权重确定规则。基于表1，根据RF系统的稳健性测试结果，选择上述12项指标在2003-2015年的数据生成原始样本集，共2028组样本，随机选取60%，即1217组数据作为训练样本，25%的数据作为测试样本，剩余15%的数据作为检验样本。首先，运用Bootstrap从训练样本中随机、有放回地抽取（可抽取规则保证生成的权重模型具有可解释性）与样本数相同数目的样本，形成2028个不同的训练集（各训练集之间相互独立，数据可以有重复）；其次，调用randomForest（）命令，对指标属性特征进行多次采样学习（直到平均误差率几乎稳定不

变），依据各指标类别及数值属性的相对变化输出内在耕地利用效率等级划分规则，并用采到的样本、特征和等级做一棵决策树，重复步骤1、步骤2生成不同的权重决策树构成随机森林，进而在混淆矩阵（confusion matrix）监督下训练耕地利用效率测度的RF模型，并将其运用于2003-2015年耕地利用效率的测度中。训练结果表明，*P*值在5%的水平下显著为正，训练样本的R^2值在 [0.79, 0.87] 范围内波动，学习精度较高。此外，本文还将BPNN和EW中确定的指标权重作为对比，进一步检验所构建模型的数据挖掘优势。

第二，参数设置。RF分类筛选对样本数据的量纲和单位并不敏感，所以不需要对数据进行归一化处理。本文运用OOB无偏估计得到不同参数设置下RF模型的精度，并分别对训练样本与检验样本进行训练和测试，若满足训练与测试精度要求即可进行下一步计算，若不满足则需要设定参数或重新取样。操作命令为：library（randomForest）。从1一直到数据集的评价指标个数12，逐一尝试，通过对比模型总体误差来寻找最优分类节点*m*。表2为不同的*m*取值对应的模型误差值大小（表2）。为了保证随机森林求解结果的精确度，将节点分割数为7时误差最小，且经过多次调试和参数敏感性分析，当分类树的数量大于460后，OOB袋外误差趋于稳定，因此确定建模的分类树为460；对于BPNN而言，本文利用Matlab 2015 a的mapminmax（）函数将指标进行归一化处理，将输入的数值转换到 [0, 1] 之间，进而运用S型转移函数构造神经网络（12×12×1），权值由伪随机函数随机产生，主要参数设置如下：目标误差最小值为0.002，最大循环次数650，冲量系数为0.1，学习率为0.02。

表2　不同*m*取值的对应误差

Table 2　Errors corresponding to different m values

个数	1	2	3	4	5	6	7	8	9	10	11	12
误差值	0.3617	0.3299	0.2677	0.2173	0.2011	0.1944	0.1802	0.1814	0.1819	0.1817	0.1822	0.1830

第三，指标权重分析。当分类树设置为460时，RF在五重交叉训练过程中产生5个系列的Gini指数减少值，求取平均值后即得到各评价指标的平均RF权重（RFW），如图4所示，RF识别地均化肥投入、地均农业机械投入、粮食单产、农民人均农业产值是影响区耕地利用效率最重要的4个指标，四者总权重比例达40.29%。而人均耕地面积和粮食安全系数则被认为是最不重要的2个指标，两者总权重比例仅为9.97%。同时，求得BPNN中各指标权重（BPNNW）和EW中各指标权重（EWW）。其中，EW识别各指标的权重较为均衡，农民人均农业产值为最重要指标，稳产指数和农业自然灾害成灾率为最不重要指标，BPNNW识别农民人均农业产值为最重要指标，农业自然灾害成灾率为最不重要指标。可见，从各指标权重的数值还是从排名情况来看，EWW和RFW存在显著差异，而BPNNW和RFW间的差异较小。机械化的规模经营以及劳动力充足条件下的精耕细作是发挥耕地生产潜力的前提与保障，产出效益是耕地利用在经济层面上的直接体现，它们均对耕地利用效率具有重要影响。而人均耕地面积在时序上的变化以及空间分布结构没有与其余指标呈现一定规律，因而被RF识别为最不重要的影响因素。

第四，权重合理性验证。从方法性质和基本原理来看，EWW仅从客观数据变化趋势和状态的角度出发，揭示各指标系列独自的数据分布结构和排序特征，而RFW和BPNNW能够反映耕地利用效率与各指标属性间的内在联系，并将这些联系以*Gini*指数

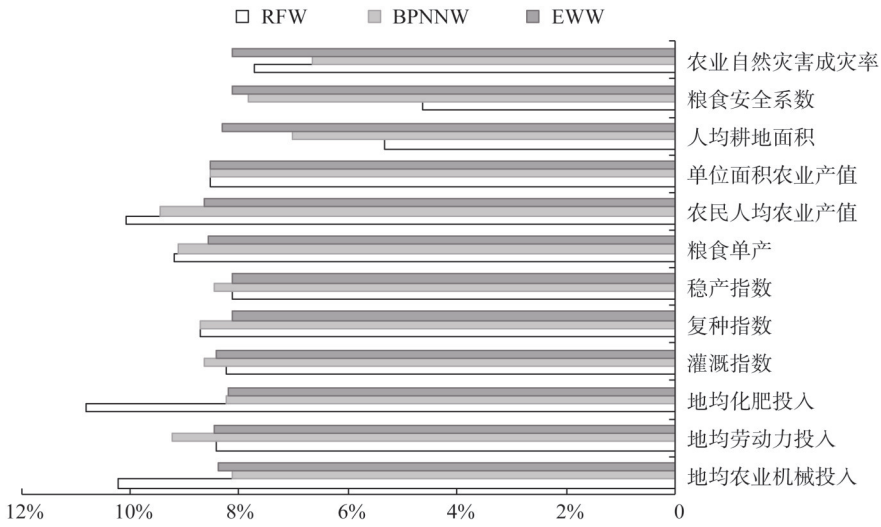


图4 三种评价方法的指标权重

Fig. 4 Values of each index of RFW, BPNNW and EW

减少值的形式表达出来，最终反映为指标权重间的差异，即两者的指标权重隐藏在分类树的内在演化规则或网络结构的映射关系之中，更符合耕地利用系统的非线性、复杂性和不确定性的特点，因此RF和BPNN权重对指标的解释力应该更强。同时，代码运行结果表明，RF进行数据挖掘的误差率集中在2.07%~7.35%，平均误差为3.19%，分类精度为89.23%，收敛时间为29.84秒，BPNN的误差率集中在5.19%~10.23%，平均误差为5.19%，分类精度为78.19%，收敛时间为43.27秒，对比之下，RF在权重确定的可解释性较强，运行精度较高，收敛速度更快，说明其权重设定对指标与测度对象间相互作用的拟合效果更好。此外，我们对权重进行十次十重交叉验证，检验BPNN和RF运行过程中权重的稳定性。由表3可知，RFW的稳定性优于BPNNW，泛化能力和稳定性较强，说明RFW更适用于指标具有时间序列特征的耕地利用效率测度。

表3 BPNNW与RFW稳定性检验

Table 3 Stability test of BPNNW and RFW

试验	BPNNW	RFW	试验	BPNNW	RFW
1	1.0000	1.0000	7	0.5000	0.9300
2	1.0000	1.0000	8	0.7500	0.7500
3	0.9300	1.0000	9	0.9300	1.0000
4	0.7500	1.0000	10	1.0000	0.7500
5	0.5000	1.0000	均值	0.8110	0.9430
6	0.7500	1.0000			

3 结果分析

3.1 耕地利用效率结果分析及比较

输入RFW值、BPW值和EWW值分别计算2003-2015年中国13个粮食主产区172个地市的耕地利用效率，以2003年、2009年和2015年为研究时点，并利用ArcGIS 10.2的

等间距断裂法将其空间可视化（依次划分为低值区、中值区、中高值区、高值区）。图5依次展示了基于RF、BPNN、EW的耕地利用效率测算结果。

（1）RF测算结果分析。时序变化上，中国粮食主产区耕地利用平均效率总体呈波动上升态势，从2003年的0.4793上升到2015年的0.7504，耕地利用效率依然较低且具有明显的阶段性特征，各省的效率指数在研究期内也都表现出不同幅度的增长趋势。其中，四川省的绝对增长量最大，由研究初期的0.6319增长至末期的0.8984，其次分别是黑龙江、江苏和吉林。分别由2003年的0.6220、0.5719、0.5633变化为2015年的0.8614、0.8374、0.8211。从各省历年耕地利用平均效率来看，黑龙江、四川、江苏、吉林的耕地利用平均效率最高，平均效率值分别为0.7552、0.7421、0.7262、0.7008，而内蒙古、湖南、河南、河北等地的耕地利用平均效率相对偏低，平均效率值分别为0.6210、0.6009、0.5851、0.5819。2009-2011年，粮食主产区内大部分省份的耕地利用平均效率在0.75以下，2012-2015年间，仅有四川、黑龙江的耕地利用平均效率始终保持在0.85以上。就不同城市而言，2003年、2009年和2015年粮食主产区172个地级市分别有32个、

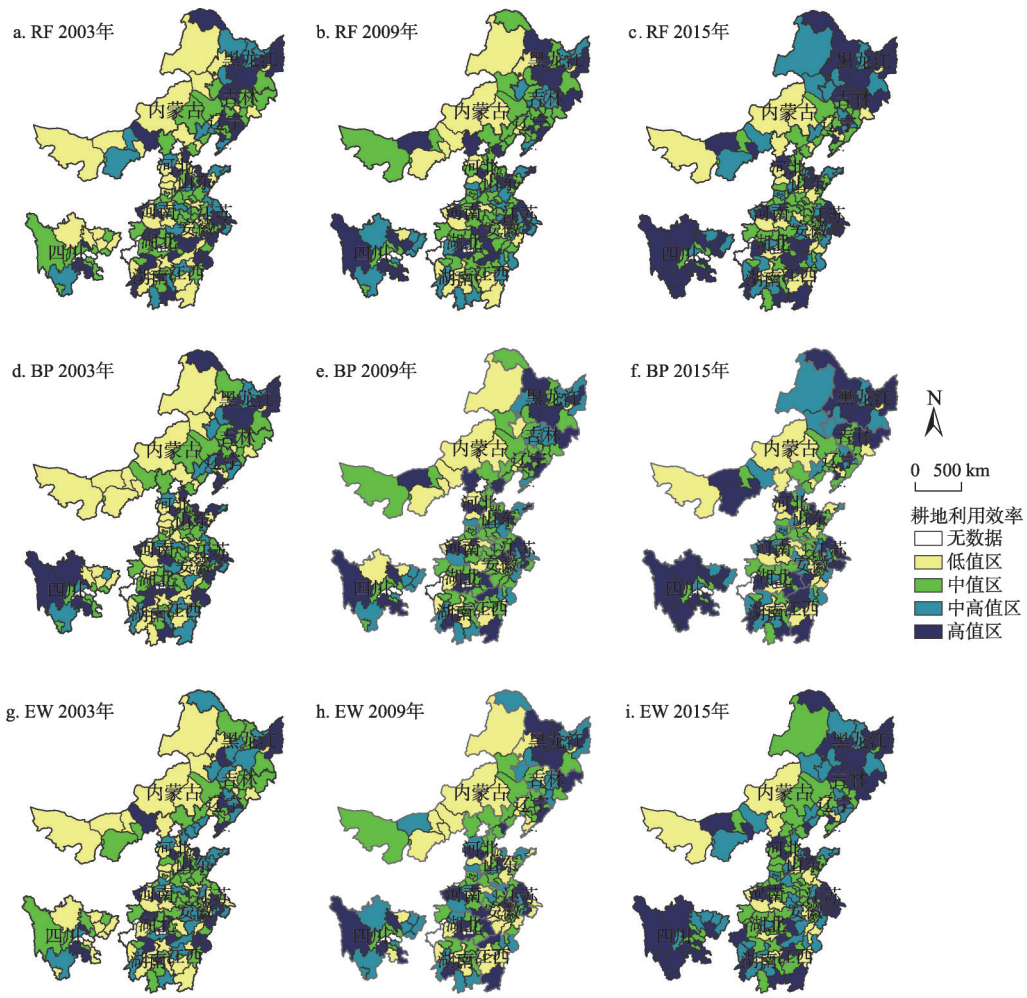


图5　2003-2015年耕地利用效率空间分布

Fig. 5　Spatial distribution of utilization benefit of cultivated land in 2003-2015

42个、59个市的耕地利用效率大于0.85，分别占总数的18.60%、24.42%、34.30%。此外，不同尺度下耕地利用效率还表现出显著的空间非均衡特征，整体呈现“南高北低”的分布格局。

（2）不同测度方法的结果对比。总体来看，就中国粮食主产区耕地利用年平均效率值而言，RF > BPNN > EW，分别从2003年的提高0.4793、0.4599、0.4037至2009年的0.6511、0.6453、0.6087，最终提高至2015年的0.7504、0.7322、0.7087。就省域耕地利用年平均效率而言，RF与BPNN的测算结果较为相近。但从局部区域来看，RF与BPNN测度值也存在显著的相对差异。如图5b与图5e，RF识别哈尔滨为耕地利用效率高值区，而BP则识别其为低值区。哈尔滨地区地势平坦、土地连片，适合大规模机械作业，具有天然的地理优势，同时人均耕地面积较多符合规模经营的理念，规模化经营耕地可以降低耕地生产成本，有利于采用更为先进的农业生产技术，提高耕地的单位产值，进而提高耕地利用效率，理论上应为耕地利用效率高值区，这一分析也得到众多文献[16,18]的实证支撑。而EW的测度结果与其余两种方法差异较大，各省按耕地利用水平由高到低的排序依次为黑龙江、吉林、四川、江苏、辽宁、河北、江西、湖北、河南、山东、湖南、安徽、内蒙古。

3.2 RF算法的合理性检验

衡量耕地利用效率测度方法的合理性，不能仅依靠以往文献与实际经验论述进行评判。因此本研究将对RF、BPNN和EW求解结果的一致性进行对比分析，并计算标准均方误差（M_{MSE}）、平均绝对误差（M_{MAE}）、百分比偏差率（M_{RPD}）和相关系数R等精度表征参数以揭示RF的优越性。

（1）一致性分析。BPNN和EW作为耕地利用效率测度中较为成熟和广泛应用的两种综合评价模型，具有其合理性和科学性。因此采用以下方法探讨RF、BPNN和EW测算结果的差异，验证RF测度的有效性：第一，将各分级图中高值区、中高值区、中值区、低值区依次标记为4，3，2，1；第二，利用ArcGIS 10.2中的栅格计算器对RF的耕地利用效率分级标记乘以10；第三，把BP神经网络的耕地利用效率分级图叠加到第二步的结果上，即可得到RF与BP神经网络耕地利用效率分级结果差异分布点图，同理可得RF与EW耕地利用效率分级结果差异分布点图（图6）。其中，差异标记的第一个序号为RF的分级结果，第二个为BP神经网络或EW的分级结果。如“21”表示“RF将该区域的耕地利用效率识别为中值区，而BP神经网络或EW则将该区域识别为低值区”，“12”则表示“RF将该区域的耕地利用效率识别为低值区，而BP神经网络或EW则将该区域识别为中值区”，其余依此类推。

从图6中可以看出，就RF与BP间的比较结果而言，RF与BP的误差点数从2003年的32个，减少到2009年的19个，最终下降至2015年的9个，相同点的个数占样本总数的比例从2003年的82.22%提高至2015年的95.00%。就RF与EW的比较结果而言，2003年、2009年和2015年误差点数分别为68、42、31，占样本总数的比例明显偏高（17.78%~37.22%），两种测度结果的误差主要集中于一级差异（“12”“21”“23”“32”“34”“43”），一级误差点数约占误差点总数的比例从研究初期的79.41%增加到研究末期的83.87%。上述结果表明，RF与BPNN的测度结果具有较高的一致性，且RF更适用于时间跨度较大的耕地利用效率测算中。

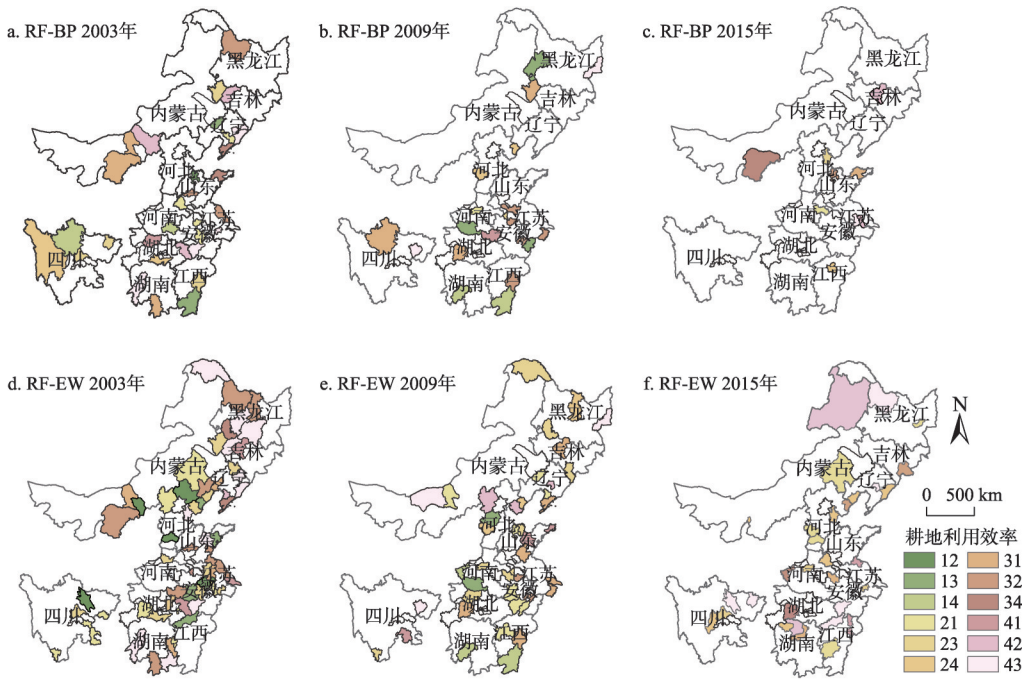


图6　RF、BP及EW测算结果的差异分布

Fig. 6　The difference distribution of the results of RF, BP and EW

在这些误差点中，EW识别为耕地利用效率高值区或中高值区而RF、BPNN识别为低值区的地区，部分为降水量少而不匀，耕地资源稀缺，经济发展水平不高的“化肥农业”城市，其耕地产出的提高主要依赖于地均化肥使用量的增加，且化肥投入在很大程度上掩盖了农业机械、水利等因素对耕地利用的正向影响，理论上耕地利用效率相对较低；EW识别为耕地利用效率低值区而RF、BPNN识别为高值区或中高值区的地区，部分为自然条件优良，经济相对发达的综合型粮食生产城市，这些城市从事经济作物以及瓜果蔬菜得到的经济产出远超于粮食，理论上具有较高的耕地利用效率，显然后两者的测算结果更具说服力，因此本文将进一步比较RF和BPNN的优越性。

（2）优越性检验。为了验证RF的优越性，选择检验集数据构建RF和BPNN模型对耕地利用效率进行测度，并与上文的测度结果进行比较。由表4可知，RF的相关系数R为0.8685，M_{RPD}为2.3533，均大于BPNN。F检验表明，唯有RF通过了5%的显著水平检验，表明RF在耕地利用效率中的评价能力更佳；RF具有最小M_{MSE}0.0174和M_{MAE} 0.0211，且与第一次测度值的拟合分布情况最佳。故RF的测度精度更高、稳定性更优。同时，与其他评价方法的平均值相比，RF的平均值更加接近第一次测度值的平均值。因

表4　测度结果误差分析

Table 4　Error analysis of measurement

检验	标准偏差	M_{MSE}	M_{MAE}	M_{RPD}	R	显著性
RF	0.1423	0.0174	0.1119	2.3533	0.8685	0.0000
BPNN	0.1629	0.0211	0.1644	1.6806	0.5722	0.1240

此，RF模型的测度结果泛化误差最小且精确度最高，能够代替BPNN进行耕地利用效率评价，具有代表性。

此外，虽然BPNN的标准偏差最小且最接近第一次测度值的标准偏差，但其偏差的位置及大小相差较大，究其原因，BPNN为“黑箱”测度模型，考虑了指标与耕地利用效率间的不确定拟合关系，但易陷入局部极小值，故其精确度最低。相比之下，RF测度结果的分布情况与训练函数的拟合度最高为91.25%。而EW为“白箱”评价方法，忽略了参数存在的不确定性以及其它影响不可测的不确定性影响，使得其标准偏差与其余两种方法的差距较大。因此，基于RF的耕地利用效率测度模型在全面考虑参数不确定性的条件下，评价结果更为准确，具有优越性。

4 结论与讨论

在耕地利用效率测度模型中，如何合理确定指标权重是效率评价的关键之一。基于RF在处理复杂性、动态性及不确定数据序列的天然优势，在分类树生成过程中引入随机因素，对样本进行Bootstrap采样得到训练集，对每个集样本用随机空间变量选取法建立权重决策树，以此构建耕地利用效率测度的RF模型。相比于常用方法（BPNN和EW）度量各指标变量对耕地利用效率的贡献时可能得到与实际情况不太符合的结果，RF能够更加有效地分解耕地利用效率与各指标间的内在联系并以线性关系表现出来，具有所需参数少、运算过程简化、解释性能佳等优点。

将该模型运用于中国粮食主产区172个城市的耕地利用效率测算中。在PSR的基础框架下，综合考虑农业可持续性、指标关联性及指标效度，从投入强度、利用强度、产出效益和可持续性4个方面，最终选取12个评价指标训练RF模型；使用构建好的RF模型对2003-2015年粮食主产区的耕地利用效率进行测度，并将BPNN和EW作为典型对比模型，从评价结果与现实的匹配度和M_{MSE}、M_{MAE}、M_{RPD}及相关系数R等精度表征参数两方面来验证RF的可行性和有效性。结果表明：对同一空间单元的效率测度值而言，RF > BPNN > EW，RF与BPNN所得效率值的总体分布格局相似，而EW的测度结果与其余两种方法差异较大。从现实情况来看，RF的测算结果与客观的自然条件和社会经济发展状况较为符合，测度结果较为客观。从模型精度来看，基于RF的耕地利用效率测度模型具有考虑参数不确定性影响的优势，其数据挖掘能力更强，评判精度更高，并且测度结果也更加符合耕地利用效率的空间分布和变化趋势，与其余方法的测度结果相比，具有一致性、代表性和优越性。

由于耕地利用系统的复杂性，借助人工智能技术的理论和方法，从RF算法的视角研究耕地利用效率是对其评价方法库的一个补充，但尚处于探索阶段。本文中构建的RF模型主要考虑了确定性因素对耕地利用效率的影响，只是一个初步的简化模型。事实上，自然环境和政策管理等因素的变化对耕地利用效率空间分布的影响也是非常显著的。此外，各指标权重是基于现实数据演变规则在训练过程中自动获取的，并不适用于未来耕地利用效率的测算，结合实际社会经济发展情况，进一步改进模型以提高评价性能将是后续研究的重点和难点。

参考文献(References):

[1] JIANG G ZHANG R, MA W, et al. Cultivated land productivity potential improvement in land consolidation schemes in Shenyang, China: Assessment and policy implications. Land Use Policy the International Journal Covering All Aspects of Land Use, 2017, 68: 80-88.

[2] 卢新海, 匡兵, 李菁. 碳排放约束下耕地利用效率的区域差异及其影响因素. 自然资源学报, 2018, 33(4): 657-668. [LU X H, KUANG B, LI J. Regional differences and its influencing factors of cultivated land use efficiency under carbon emission constraint. Journal of Natural Resources, 2018, 33(4): 657-668.]

[3] KLEIJN D, KOHLER F, BALDI A, et al. On the relationship between farmland biodiversity and land-use intensity in Europe. Proceedings Biological Sciences, 2009, 276(1658): 903-909.

[4] 龙禹桥, 吴文斌, 余强毅, 等. 耕地集约化利用研究进展评述. 自然资源学报, 2018, 33(2): 337-350. [LONG Y Q, WU W B, YU Q Y, et al. Recent study progresses in intensive use of cropland. Journal of Natural Resources, 2018, 33(2): 337-350.]

[5] 杜国明, 刘彦随. 黑龙江省耕地集约利用评价及分区研究. 资源科学, 2013, 35(3): 554-560. [DU G M, LIU Y S. Evaluating and zoning intensive utilization of cultivated land in Heilongjiang province. Resources Science, 2013, 35(3): 554-560.]

[6] 曹银贵, 周伟, 王静, 等. 基于主成分分析与层次分析的三峡库区耕地集约利用对比. 农业工程学报, 2010, 26(4): 291-296. [CAO Y G, ZHOU W, WANG J, et al. Comparative on regional cultivated land intensive use based on principal component analysis and analytic hierarchy process in Three Gorges Reservoir Area. Transactions of the CSAE, 2010, 26(4): 291-296.]

[7] WANG K, ZHANG P. The research on impact factors and characteristic of cultivated land resources use efficiency: Take Henan province, China as a case study. Ieri Procedia, 2013, 5(5): 2-9.

[8] 李强, 彭文英, 王建强, 等. 乡镇企业发达区耕地健康评价与驱动机理研究. 自然资源学报, 2015, 30(9): 1499-1510. [LI Q, PENG W Y, WANG J Q, et al. Health assessment and driving mechanism analysis of cultivated land in the township enterprises developed region. Journal of Natural Resources, 2015, 30(9): 1499-1510.]

[9] 石淑芹, 曹玉青, 吴文斌, 等. 耕地集约化评价指标体系与评价方法研究进展. 中国农业科学, 2017, 50(7): 1210-1222. [SHI S Q, CAO Y Q, WU W B, et al. Progresses in research of evaluation index system and its method on arable land Intensification: A review. Scientia Agricultura Sinica, 2017, 50(7): 1210-1222.]

[10] MENG X L, SHI F G. An extended data envelopment analysis for the decision-making. Journal of Inequalities & Applications, 2017, 2017(1): 240.

[11] 赖红松, 吴次芳. 基于粗糙集和支持向量机的标准农田地力等级评价. 自然资源学报, 2011, 26(12): 2141-2154. [LAI H S, WU C F. Productivity evaluation of standard cultivated land based on rough set and support vector machine. Journal of Natural Resources, 2011, 26(12): 2141-2154.]

[12] LEO B. Random forests. Machine Learning, 2001, 45(1): 5-32.

[13] 赖成光, 陈晓宏, 赵仕威, 等. 基于随机森林的洪灾风险评价模型及其应用. 水利学报, 2015, 46(1): 58-66. [LAI C G, CHEN X H, ZHAO S W, et al. A flood risk assessment model based on Random Forest and its application. Journal of Hydraulic Engineering, 2015, 46(1): 58-66.]

[14] LINDNER C, BROMILEY P A, IONITA M C, et al. Robust and accurate shape model matching using Random Forest Regression-Voting. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1862-1874.

[15] 刘影, 肖池伟, 李鹏, 等. 1978-2013年中国粮食主产区“粮—经”关系分析. 资源科学, 2015, 37(10): 1891-1901. [LIU Y, XIAO C W, LI P, et al. Relationship of grain output and economic development from 1978 to 2013 in the major grain producing area of China. Resources Science, 2015, 37(10): 1891-1901.]

[16] 张立新, 朱道林, 谢保鹏, 等. 中国粮食主产区耕地利用效率时空格局演变及影响因素: 基于180个地级市的实证研究. 资源科学, 2017, 39(4): 608-619. [ZHANG L X, ZHU D L, XIE B P, et al. Spatiotemporal pattern evolvement and driving factors of cultivated land utilization efficiency of the major grain producing area in China. Resources Science, 2017, 39(4): 608-619.]

[17] 倪超, 杨胜天, 罗娅, 等. 基于循环经济的黑龙江省耕地利用集约度时空差异. 地理研究, 2015, 34(2): 341-350. [NI C, YANG S T, LUO Y, et al. The spatial-temporal difference analysis of cultivated land use intensity in Heilongjiang

province based on circular economy. Geographical Research, 2015, 34(2): 341-350.]

[18] SONG X, ZHU O, LI Y, et al. Cultivated land use change in China, 1999-2007: Policy development perspectives. Journal of Geographical Sciences, 2012, 22(6): 1061-1078.

Measurement of cultivated land utilization efficiency: Construction and application of random forest

CHEN Dan-ling[1], LU Xin-hai[2], KUANG Bing[2]

(1. College of Public Administration, Huazhong University of Science and Technology, Wuhan 430074, China;
2. College of Public Administration, Central China Normal University, Wuhan 430079, China)

Abstract: Setting up a suitable quantitative analysis model is a basic work for scientific grasp of cultivated land utilization efficiency and its distribution pattern, and can provide reasonable decision- making basis for sustainable utilization of cultivated land then realizing the coordinated development of cultivated resources and environment. In order to effectively describe the complexity, dynamics and heterogeneity characteristics of cultivated land use system, a random forest (RF) model for measuring cultivated land utilization efficiency is constructed by applying random sampling Bootstrap to build a classification tree reasonably. Then by taking 172 cities in the major grain producing areas of China as an example, the RF model was trained to measure the cultivated land utilization efficiency in 2003-2015 compared with Back Propagation Neural Network and Entropy weight to verify the consistency, representative and superiority of RF. The results show that: (1) RF model has fewer parameters and simpler implementation. It can simulate the complex relations among the evaluation indexes, which makes it convenient to analyze the value of each index. (2) For efficiency measurement results of the same space unit, RF > BPNN > EW, the overall distribution pattern of the cultivated land utilization efficiency in RF and BPNN is similar while a great difference exists in EW. (3) Judged from the matching degree of evaluation results to reality and the accuracy parameters, the measurement results are reasonable and in accordance with the facts in RF, which reflected its high applicability and reliability. At the same time, compared with the other two commonly used models, RF can reduce the dimensions of input vectors and the computing complexity, then raise the training efficiency. The correlation coefficient R of RF is 0.8685, M_{RPD} is 2.3533, with the minimum M_{MSE} and M_{MAE} being 0.0174 and 0.0211, respectively, which is more suitable for the study of the cultivated land utilization efficiency with complex nonlinear characteristics, and this method has explored a new way for evaluating cultivated land utilization efficiency.

Keywords: cultivated land utilization efficiency; random forest; main grain producing areas