应用回归树分析双季稻区水稻土地力特征*

王莉莎[1,2] 李 勇[2**] 沈健林[2] 刘新亮[2] 付晓青[2] 石 辉[1] 黄铁平[3]
([1]西安建筑科技大学, 西安 710055; [2]中国科学院亚热带农业生态研究所, 长沙 410125; [3]湖南省土壤肥料工作站, 长沙 410005)

摘 要 地力评价对推进农业结构调整、改良土壤和科学施肥有重要的指导作用。基于长沙县2007—2010年耕地地力调查与评价工作,本研究选取影响地力的立地条件、土壤理化性质、农田基本设施建设等方面共22项指标,以实际产量与潜在产量比作为地力高低的替代指标,采用回归树统计分析方法,对影响双季稻区水稻土地力的因素进行了分析和建模。结果表明:海拔、地形部位、田面坡度、排水能力、输水方式、土壤母质、土壤质地、耕层厚度、土壤有机质含量和有效氮含量为影响长沙县双季稻水稻土地力的主要因子,并且可用这10个因子对地力特征进行较好的模拟。回归树方法对分析大容量、多变量数据间的复杂关系具有很好的效果,可推广用于中国其他地区的地力评价工作。

关键词 双季稻区; 回归树; 地力评价; 水稻产量; 模拟
中图分类号 S158 **文献标识码** A **文章编号** 1000-4890(2013)1-0232-09

Using regression tree to analyze the fertility characteristics of paddy soil in double-rice cropping region. WANG Li-sha[1,2], LI Yong[2**], SHEN Jian-lin[2], LIU Xin-liang[2], FU Xiao-qing[2], SHI Hui[1], HUANG Tie-ping[3] ([1] *Xi'an University of Architecture and Technology*, *Xi'an* 710055, *China*; [2] *Institute of Subtropical Agriculture*, *Chinese Academy of Sciences*, *Changsha* 410125, *China*; [3] *Soil and Fertilizer Station of Hunan Province*, *Changsha* 410005, *China*). *Chinese Journal of Ecology*, 2013, **32**(1): 232-240.

Abstract: Soil fertility evaluation plays an important guidance role in promoting agricultural reconstruction, improving soil productivity, and recommending proper fertilization. Based on the investigation and evaluation of arable land fertility in Changsha County of Hunan Province, South-central China in 2007-2010, this paper analyzed and modeled the determining factors of paddy soil fertility in the double-rice cropping region of Changsha by using the statistical analysis method of classification-regression tree (CRT). A total of 22 independent variables associated with paddy soil fertility, such as site condition, soil physical and chemical properties, and farmland basic facilities were selected, and the ratio of surveyed yield to potential yield for eliminating the rice variety impact was treated as the dependent variable for representing the paddy soil fertility level. Our results suggested that elevation, topographic position, field surface slope, drainage capacity, water transport mode, soil parent material, soil texture, soil plow layer depth, soil organic mater content, and soil available nitrogen concentration were the main factors affecting the fertility characteristics of paddy soil in the double-rice cropping region of Changsha, and these 10 factors could be used to construct a regression tree model to well predict the paddy soil fertility level. Considering its good performance in analyzing complicated relationships of numerous variables and in handling large volume of data set, the CRT method was recommended to be used in the arable land fertility evaluation in other places of China.

Key words: double-rice cropping region; regression tree; soil fertility evaluation; rice yield; modeling.

* 国家重点基础研究发展计划项目(2011CB100506)资助。
** 通讯作者 E-mail: yli@isa.ac.cn
收稿日期: 2012-06-17 接受日期: 2012-10-17

20 世纪 80 年代以来,中国耕地数量急剧减少,人多地少的矛盾日益尖锐。全国人大农业与农村委员会公布,2011 年中国耕地面积约为 1217333 km^2,比 1997 年的 1299333 km^2 减少 82000 km^2(陈磊,2011)。湖南的耕地变化趋势与全国类似,8 年时间共计净减耕地 1362 km^2。耕地的减少对粮食生产提出了巨大的挑战。水稻是中国第二大作物,尽管其面积仅占中国作物总面积的 29.1%,但产量却占全国粮食总产的 43.7%。鉴于中国处于世界上最大的水稻生产国地位,水稻产量的变化不仅对中国,对亚洲乃至世界的粮食安全都将产生很大影响(姚凤梅等,2007)。通过分析水稻土地力影响因素,可为稻田改良和提高水稻产量提供有重要价值的科学依据。湖南省长沙县的平原和低谷地区自古以来一直适宜水稻种植,该县水稻常年种植面积在 800 km^2 左右,年总产量 54×10^4 t,是一个以双季稻为主的粮食主产区和商品粮县。以该县的水稻土地力特征为研究对象可以找出一些基本规律,对于更好地了解中国南方水稻土区域的一般现状,提高区内土壤肥力,发展粮食生产具有典型而深远的意义。

以往关于水稻产量与土地地力特征的研究中,多采用不同地力影响因素与水稻产量之间的相关性分析,相关系数越高,则代表该地力影响因素对水稻产量的影响越强。这种方法在判断一系列地力影响因素对水稻产量的独立作用时具有一定的优势。不过,通常我们的研究目的是要在一系列繁杂的因素中筛选出起主要作用的因素,以此来因地制宜地开展土壤改良和地力培肥,相关分析法在这种情况下就显得稍有逊色。多元线性回归分析虽然可以实现上述目标,但在众多预测变量中,往往又会出现变量之间的共线性,这种分析方法无法处理变量之间的共线性,对多预测变量间复杂的交互作用分析困难,因此使回归系数的正确估计受到影响(武艳华等,2008)。近几年,数据挖掘技术引起了信息产业界和整个社会的广泛关注,在这个信息化的社会,随着数据量不断增加,迫切需要将这些庞大的数据提取和转化为有用的信息和知识。获取的信息和知识能够用于各种各样的应用中,包括市场分析、欺诈检测、图像识别、风险预测和科学研究等。回归算法是数据挖掘领域中在解决预测问题所必须的核心工具,我们可以在数据挖掘工具中利用一些常用的回归算法来建立相应的预测模型,以解决实际中的回归问题(蔡晖,2011)。回归树建模方法在数据统计方面具有很好的灵活性,因此它在医疗判断、气象预测、物流管理、数据挖掘、投资风险分析等行业领域均已经有成功的应用(张立彬等,2002),但在土壤学中的应用还不为多见,本文通过对研究区水稻产量与各环境因子之间建立回归树模型实践,使其统计方法能广泛地应用于土壤科学和农业生产之中。

1 材料与方法

1.1 数据来源

为了对长沙县域范围内的耕地地力进行评价,湖南省土肥站于 2007—2010 年组织开展了全县耕地地力调查,共采集大田土壤样品 6072 个,采样点覆盖全县 20 个乡镇和 58 个土种,并对立地条件、土壤理化性质、灌排条件、农田基本设施建设和水稻产量等 38 项指标进行了分析测定和归类。各土种面积差异很大,布点时,以大面积土种为重点,同时兼顾小面积土种。以样点总数为基础,面积>66.67 km^2 的土种布点 50 个以上,33.33 到 66.67 km^2 的布点 30 到 50 个,13.33 到 33.33 km^2 的布点 15 到 30 个,0.67 到 13.33 km^2 的布点 8 到 15 个,小于 0.67 km^2 的布点 1 到 10 个。由于受到采样点数限制,土壤类型多而分布差异大,在采样点的选择时:以土种为基础,兼顾山区、丘陵等地貌类型,以乡镇为布点平台,以种植制度为布点框架,以第 2 次土壤普查的典型剖面为基点。

本研究以 2007—2010 年长沙县耕地地力调查与评价资料为依托,提取双季稻田土壤采样点(共 4742 个)数据,并基于以往对土壤地力特征的研究,选取 38 项指标中对地力影响较强的 22 个环境因子来进行回归树分析。在构建模型的过程中,经多次试验表明,这 22 个环境因子在本研究中影响显著且基本可以解释模型。22 个环境因子中,有 14 个为分类型变量,另外 8 个为数值型变量,按照《全国耕地地力调查与质量评价技术规程》所确定的耕地地力评价技术流程的要求,对它们的说明分别参见表 1 和表 2。

1.2 研究方法

作物产量是土壤地力的综合反映,可以作为衡量地力高低的指标。由于各采样点水稻品种不一致,从而使得非地力属性的品种因子可能会对以产量为指标的地力分析产生干扰。为消除品种的影响,本研究以实际产量与该品种在研究区域内的潜在产量比值(简称产量比)作为衡量地力高低的指

表1 分类型环境因子的说明

Table 1 Classification of category environmental factors

环境因子	英文说明	分类说明
地貌类型	LANDFORM	1 中山,2 丘陵,3 平原
地形部位	TOPO_POSITION	1 低山丘陵坡地,2 低山缓坡地,3 丘陵坡地中上部,4 丘陵坡地下部,5 丘陵缓坡地,6 丘陵低谷地,7 山垄上部,8 冲垄中下部,9 冲垄下部,10 岗地,11 河流阶地,12 河网平原
坡向	ASPECT	1 东,2 东南,3 南,4 西南,5 西,6 西北,7 北,8 东北,9 平地
农田基础设施	INFRASTRUCTURE	1 完全配套,2 配套,3 基本配套,4 不配套,5 无设施
排水能力	DRAINAGE	1 强,2 较强,3 中,4 较弱,5 弱
灌溉能力	IRRIGATION	1 保灌,2 能灌,3 可灌,4 无灌
水源条件	WATER	1 河流,2 湖泊,3 井水,4 水库,5 塘堰,6 无
输水方式	WATER_TRANSPORT	1 提水防渗渠,2 提水土渠,3 自流防渗渠,4 自流土渠,5 无
灌溉方式	IRRIGATION_METHOD	1 沟灌,2 管灌,3 漫(畦)灌,4 无
土壤亚类	SOIL_SUBORDER	1 典型红壤,2 漂洗水稻土,3 潜育水稻土,4 石灰性紫色土,5 淹育水稻土,6 潴育水稻土
土壤母质	PARENT _ MATERIAL	1 非石灰性壤质冲积物,2 非石灰性砂质冲积物,3 非石灰性粘质冲积物,4 基性结晶岩类残积物,5 砾质洪积物(砾石占剖面30%以上),6 砾质坡积物(砾石占剖面30%以上),7 泥岩类残积物,8 石灰性泥岩类残积物(钙质、紫色泥、页岩),9 石灰性壤质冲积物,10 石灰性砂岩类残积物(钙质、紫色砂岩砾岩),11 石灰性砂质冲积物,12 石灰性粘质滨海沉积物,13 石灰性粘质冲积物,14 酸性结晶岩类残积物,15 酸性泥岩类残积物(红色泥页岩),16 酸性砂岩残积物(红色砂砾岩),17 土质洪积物(砾石占剖面1% ~ 30%),18 土质坡积物(砾石占剖面1% ~ 30%),19 中性砾质洪积物,20 中性砾质坡积物,21 中性泥岩类残积物(紫色泥页岩),22 中性砂岩残积物(紫色砂砾岩),23 中性土质洪积物,24 中性土质坡积物
土体构型	PROFILE	1 A-B-C,2 A-P-C,3 A-P-E-C,4 A-P-G-C,5 A-Pg-Wg-C,6 A-P-W1-W2-C,7 A-P-W-C,8 A-P-W-G
土壤质地	TEXTURE	1 砂土及壤质砂土,2 砂质壤土,3 壤土,4 粉砂质粘壤土,5 粘壤土,6 粘土
土壤结构	STRUCTURE	1 块状,2 粒状,3 团块状,4 团粒状,5 微团粒,6 无

表2 数值型环境因子的数据分布

Table 2 Statistical distribution of numeric environmental factors

环境因子	英文说明	最大值	最小值	均值
海拔(m)	ELEVATION	240.0	18.0	54.0
坡度(°)	SLOPE	60.0	0.0	5.5
地下水水位(cm)	GW_DEPTH	470.0	5.0	75.7
耕层厚度(cm)	PLOW_LAYER_DEPTH	30.0	2.0	15.7
土壤有机质($g \cdot kg^{-1}$)	SOM	81.5	2.0	34.1
土壤有效氮($mg \cdot kg^{-1}$)	AV_N	400.0	2.0	167.2
土壤速效磷($mg \cdot kg^{-1}$)	AV_P	258.2	0.9	22.1
土壤速效钾($mg \cdot kg^{-1}$)	AV_K	1749.0	28.0	298.4

标。在此基础上,以产量比作为响应变量,以22个环境因子作为预测变量,采用回归树统计方法对研究区双季稻田地力特征进行了相关分析。以下是该统计方法的简要步骤,应用R统计软件及其相关回归树分析包tree和rpart完成。

1.2.1 变量重要性评分 建立随机森林的基本思想是,通过自助法(bootstrap)重采样技术,不断生成训练样本和测试样本,由训练样本生成多个分类树组成随机森林,测试数据的分类结果按分类树投票多少形成的分数而定。随机森林有两个重要参数,一是树节点预选的变量个数 m_{try};二是随机森林中树的个数 n_{tree}。本研究设定 $m_{try}=\sqrt{m_{all}}$ (m_{all}为所有变量的个数),$n_{tree}=500$。算法通过R语言编制的程序实现。

设原始样本含量为 N,表示各基因的变量为 x_1,$x_2,\cdots,x_m$。应用bootstrap法有放回地随机抽取 b 个新的自助样本,并由此形成 b 个分类树,每次未被抽到的样本则组成 b 个袋外数据(out-of-bag,OOB)。袋外数据作为测试样本可以用来评估各个变量在分类中的重要性,实现过程如下:

(1)用自助样本形成每一个树分类器,同时对相应的OOB进行分类,得到 b 个自助样本中OOB中每一个样品的投票分数,记为 $vote_1, vote_2, \cdots, vote_b$。

(2)将变量 x_i 的数值在 b 个OOB样本中的顺序随机改变,形成新的OOB测试样本,然后用已建立的随机森林对新的OOB进行分类,根据判别正确的样品数得到每一个样本的投票分数,所得结果可用矩阵表示。

(3)用 $vote_1, vote_2, \cdots, vote_b$ 与矩阵对应的向量相减,求和平均后得变量 x_i 的重要性评分。

1.2.2 变量选择 变量重要性评分是针对特定样本数据做出的,一般情况下评分排在前面位置的变量相对重要一些,但在数据中含有大量对分类不起作用变量的情况下,其评分是不准确的。研究人员

在前期的模拟试验中注意到,在变量数目很多的情况下,评分排在前面的变量不一定对分类有贡献,但位于后面的变量总是些不起作用的变量。为此,我们给出一种新的算法,其基本思想是,根据一定的准则逐步从数据中删除“不起作用”的变量,自动建立“最优”随机森林判别模型。具体算法如下:

(1)确定迭代过程中的 ROC 曲线下面积 AUC 的容许误差△值和变量选择的容许误差 R_{min},本研究设置△=0.005,$R_{min}=2$。

(2)采用留一方法,将样本量为 N 的样本分为两部分,其中 N-1 个样品作为训练部分,另外一个样品作为测试样品,该过程用于每个样品,建立 N 个随机森林,并对相应的每一个测试样品进行判别,然后用 ROC 曲线下面积对其进行评定,记为 AUC_0;同时作赋值 $k=0$,forward=true。

(3)对当前数据中的所有变量的重要性评分进行排序。

(4)如果 forward=true,采用向前折半算法,保留排在前面位置的变量,去除其余变量,得到经变量筛选后的新数据集;如果 forward= false,采用向回折半算法,保留排在前面位置的变量,去除其余变量,得到经变量筛选后的新数据集。应用随机森林算法对新数据集采用留一法进行分类,并计算 AUC_{k+1}。

(5)判断折半的方向:如果 $AUC_k-AUC_{k+1}\leqslant\triangle$,则 forward=true,如果 $AUC_k-AUC_{k+1}>\triangle$,则 forward=false。

(6)判断是否停止迭代:如果折半后变量的数目小于 R_{min},则停止变量筛选过程,建立“最优”随机森林判别模型;否则 $k=k+1$,返回第(3)步进入循环。

1.2.3 树的生成及分叉原则 回归树的基本原理是根据响应变量,利用循环二分形成二叉树形式的决策树结构,将由预测变量定义的空间划分为尽可能同质的类别(郑海峰,2009)。这类模型的拟合采用二元递归分解法,即连续地把一个数据集分解到不断增加的同质子集直至无法再继续分解(Clark & Pregibon,1992)。从概念层面可以将二元递归分解法描述为一种减少“杂质”的量的过程(Breiman *et al.*,1984)。对于回归问题,这种“杂质”是用某个特定节点 i 的离差平方和 D_i 来测量的。在初始时刻,所有的观测值被分配给同一个“节点”(根节点)。每次分解把观测值分成两个子节点(左边和右边),分解后的离差平方和为:

$$D_{i,child}=D_{i,L}+D_{i,R}$$

在给定节点,能够最大化离差平方和的减少量 $\triangle D=D_i-D_{i,child}$ 的分解方式被最终选中。特别地,模型按顺序用于每个预测变量,然后将响应变量分为两组。分解是基于现有的预测变量。如果预测变量 x_j 是连续或者顺序分类的变量,分解则根据预测变量 x_j 小于或者大于某个特定值把响应变量划分为两组。如果 x_j 具有 n_j 个唯一值,那么就有 n_{j-1} 种分解响应变量数据的可能方法。模型将对所有可能的分解方法进行尝试,计算每次分解造成的离差平方和减少量。如果预测变量 x_j 是字符型变量,模型将尝试各种可能的二分方式并记录每次分解造成的离差平方和减少量。在对所有预测变量做完计算后,具有能够使得离差平方和减少量最大化的预测变量就被选中为最佳预测变量(钱松,2011)。一般而言,位于第一级节点处(分类树最上端分枝)的分类参数对因变量变异的贡献程度最大,即对因变量变异起主导作用的因子(White *et al.*,2005)。这种方法被称为贪婪算法,算法会为当前节点选出最佳分解而不考虑整个树的变现。在每次分解后,原始数据集被划分成两个子集,对这两个子集重复相同的过程,就“种”出了一棵树。

1.2.4 树的修剪 分类回归树的优化算法主要有预剪枝和后剪枝两类。预剪枝技术主要是按照某种标准限制树的充分生长,控制树的规模,它不能把握树的全局信息,存在一定的盲目性,难以确定被剪掉的结点的子结点是否具有存在价值,可能使决策树过早停止生长,从而不能得到最优的分类与回归树,因此这一方法未获得广泛应用。后剪枝技术则是待分类与回归树充分生长后,通过某种标准剪掉某些枝干。后剪枝法是基于分类与回归树的全局信息,通常优于预剪枝法,因此得到了较为广泛的应用(王代星等,2010)。

目前主要的后剪枝方法有如下 4 种(王琴竹,2011):悲观错误剪枝(pessimistic error pruning,PEP)、最小错误剪枝(minimum error pruning,MEP)、代价复杂度剪枝(cost-complexity pruning,CCP)和基于错误的剪枝(error-based pruning,EBP)。本文采用第 3 种方法进行剪枝处理。

在讨论代价复杂度剪枝算法之前,先给出文章中出现的一些术语和符号。

T_{max}:由决策树生成算法生成的未剪枝的完全决策树;

T_t:以内部节点 t 为根的一棵子树;

$n(t)$:到达节点 t 的所有样本数目;

$n_i(t)$:到达节点 t 且属于第 i 类的样本数目;

$e(t)$:到达节点 t 但不属于节点 t 所标识的类别的样本数目;

$r(t)$:错分样本率,其值为 $e(t)/n(t)$。

CCP 剪枝算法为子树 T_t 定义了代价(cost)和复杂度(complexity),以及一个可由用户设置的衡量代价与复杂度之间关系的参数 α。其中,代价指在剪枝过程中因子树 T_t 被叶节点替代而增加的错分样本,复杂度表示剪枝后子树 T_t 减少的叶节点数,α 则表示剪枝后树的复杂度降低程度与代价间的关系,定义为 :$\alpha=\frac{R(t)-R(T_t)}{|N_t|-1}$。其中 $|N_t|$ 表示子树 T_t 中的叶结点数;$R(t)$ 表示子树 T_t 被剪枝并代以叶节点 t 后产生的错分样本数,计算公式为 $R(t)=r(t)\times p(t)$,$r(t)$ 为节点 t 的错分样本数,$p(t)$ 为落入节点 t 的所有样本;$R(T_t)$ 表示子树 T_t 没有被剪枝时的错分样本数,计算公式为 $R(T_t)=\sum R(i)$,i 为子树 T_t 的叶节点。

CCP 剪枝的结果是得到一棵错分样本数和树的规模相折中的回归树,其剪枝的过程分为两步:

(1)对于完全回归树 T_{max} 的每个非叶节点计算 a 值;循环,剪掉具有最小 a 值的子树,直到剩下根节点;在该步可得到一系列剪枝树 $\{T_0,T_1,T_2,\cdots,T_m\}$,其中 T_0 为 T_{max},T_{i+1} 为对 T_i 进行剪枝得到的结果。

(2)用新的剪枝集对步骤(1)中的 T_i 进行评估并找出其中的最佳剪枝树作为结果(范洁和杨岳湘,2005)。

1.2.5 树型的确定 Breiman 等(1984)建议使用 1 倍标准误(SE)规则,即尺寸小一些且与产生最小的交叉检验误差的树的差异的一倍标准误之内的树。一个模型预测的准确性是用交叉检验误差(E)和交叉检验标准差(SE)来度量。要估计交叉检验误差,可以通过将原始数据集随机划分为 10 个(默认值)子集合的模拟过程来完成。将 1 个子集合放在一边,把它作为检验数据集,而其他 9 个子集合用来拟合模型。放在一边的检验数据集合则用来评估模型。重复 10 次这个过程,每次采用不同的检验集合。E 是 10 个误差之和,而 SE 则是 10 个误差的标准误(钱松,2011)。记 E_i 为剪枝树 T_i 交叉检验误差,E_{min} 为 $\min\{E_i\}$,SE_i 为 E_i 的标准误。所得的最佳剪枝树 T_{best} 是满足条件 $E_{best}\leqslant E_{min}+SE_{min}$ 且包含的节点数最少的那棵剪枝树 T_i。

2 结果与分析

2.1 影响因子重要性分析及筛选

对长沙县 22 个影响产量的环境因子的重要性分析见表 3,表中各个指标值的大小代表其重要的程度。地貌类型(LANDFORM)的重要性相比较其他预测变量明显偏低,回归树作为一种有效的变量选择模型将它剔除出了模型的拟合。

2.2 双季稻田地力特征的回归树分析

构建一棵包含许多层及节点的回归树,如果不加以修剪,这棵树必然可以完全解释响应变量。模型复杂性参数(complexity parameter, cp)可以限制模型的复杂性,而模型的复杂性与树模型的大小直接相关,即树枝越多,模型越复杂。cp 值越小,模型越复杂。由于数据量过于庞大,应用预剪枝方法在本文研究中设定 cp 值为 0.0015,采用二元递归分解法初步生成回归树。这棵树经过 69 次分解具有 88 个终端节点。由于这样的回归树过于庞大,并且对于非重要预测变量而言,它们的存在对模型解释结果并没有很大的作用,所以对回归树进行后修剪就显得十分重要。

根据代价复杂度剪枝法的第一步骤,首先得出一系列由交叉检验分析得出的剪枝树,接着就可对模型预测能力和交叉检验误差分别与树的大小相关

表 3 在回归树分析中表征水稻土地力特征的各环境因子的重要性评估

Table 3 Evaluation of the importance of environmental factors related to paddy soil fertility based on regression tree analysis

指标	回归树分析重要性	指标	回归树分析重要性
ELEVATION	2.32	DRAINAGE	0.67
ASPECT	1.72	TEXTURE	0.57
AV_N	1.69	INFRASTRUCTURE	0.56
SOM	1.58	WATER	0.49
AV_K	1.56	STRUCTURE	0.45
AV_P	1.42	IRRIGATION	0.43
TOPO_POSITION	1.42	PROFILE	0.40
PARENT_MATERIAL	1.28	WATER_TRANSPORT	0.40
PLOW_LAYER_DEPTH	1.19	IRRIGATION_METHOD	0.27
SLOPE	1.15	SOIL_SUBORDER	0.17
GW_DEPTH	0.90	LANDFORM	0.05

作图(图1)。左图实线表示非交叉检验的情况下随着树结点的增多,拟合度增加,虚线表示交叉检验情况下随着树结点的增多,拟合度增加至最大值就趋于平缓。右图中竖直线段是 E_{min} 加上/减去 1 倍 SE_{min},水平实线是最小的 E_{min} 加上 1 倍 SE_{min}。

根据交叉检验方法以及 1 倍标准误规则,得出 $E_{min}=0.8781$ 和 $SE_{min}=0.0184$,即 E_{best} 的取值范围是[0.8781,0.8965]。E_{best} 对应的复杂性参数 cp 的取值范围是从 0.0033 到 0.0058。当 cp = 0.0055 时,修剪后得出了具有 13 个分叉的回归树。在最终的模型中,并没有包括所有的候选预测变量,仅用到了海拔、地形部位、田面坡度、排水能力、输水方式、土壤母质、土壤质地、耕层厚度、土壤有机质、有效氮来拟合处理。经过预剪枝构建的回归树模型能够解释产量比总方差的 30%,最终优化后的模型能够解释产量比总方差的 16%。

作为一种有效的变量选择模型,回归树方法最终构建的树模型不仅筛选出了对水稻产量影响显著的上述 10 个环境因子,并且给出了它们是怎样影响产量的变化。根据以上结果修剪后的优化回归树如图 2 所示。

修剪以后的回归树具有 13 次分解或者说 14 个终端结点。每个节点顶部的文字给出的是将父节点分解成两个子节点的变量和标准。根节点(在最顶

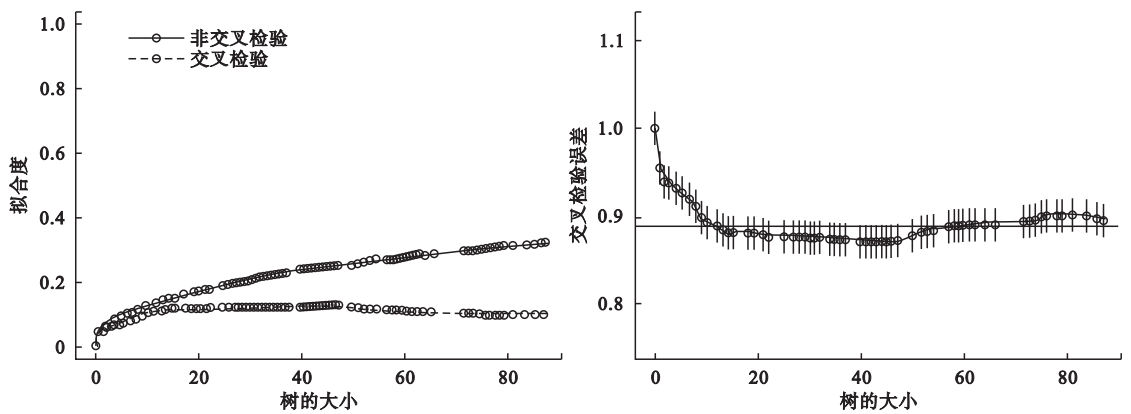


图 1 回归树交叉检验分析

Fig. 1 Result of regression tree with cross-validation analysis

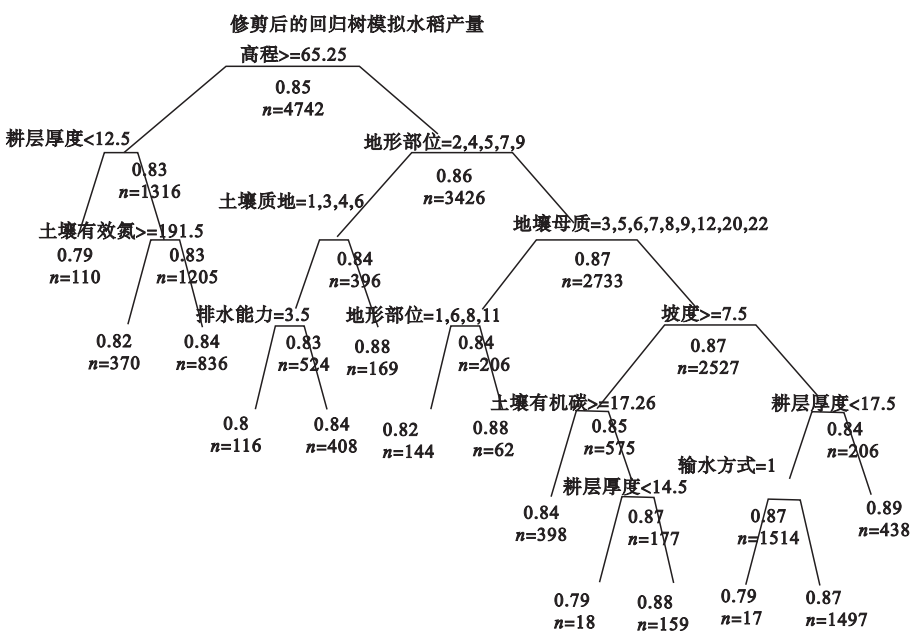


图 2 修剪以后的回归树

Fig. 2 Pruned regression tree

端)代表的是整个数据集。构成回归树的结点分别为海拔(ELEVATION)、地形部位(TOPO_POSITION)、田面坡度(SLOPE)、排水能力(DRAINAGE)、输水方式(WATER_TRANSPORT)、土壤母质(PARENT_MATERIAL)、土壤质地(TEXTURE)、耕层厚度(PLOW_LAYER_DEPTH)、土壤有机碳(SOC)、有效氮(AV_N)。在此以根结点为例说明数据型结点的意义如下:未分解时的树具有样本数为4742,实际水稻产量与潜在水稻产量比值的平均值为0.85,经过一次分解后样本变成1316与3426。当海拔高度高于65.25 m时选择左分支,在1316个样本中,实际水稻产量与潜在水稻产量比值的平均值为0.83。相反,右分支的实际水稻产量与潜在水稻产量比值的平均值为0.86。以地形部位为例说明字符型结点的意义是:有3426个样本的海拔低于65.25 m,当地形部位为低山缓坡地、丘陵坡地下部、丘陵缓坡地、山垄上部、冲垄下部这样的类型时,其中的693个样本符合条件,实际水稻产量与潜在水稻产量比值的平均值为0.84,其余样本的实际水稻产量与潜在水稻产量比值的平均值为0.87。拟合出的模型可以被看作是预测实际水稻产量与潜在水稻产量比值的一组规则。即对任一给定的观测值,可以利用分解标准提出一系列问题,而对这些问题的回答就可以从根节点走到某个终端节点,终端节点的实际水稻产量与潜在水稻产量比值的均值就是模型的估计值。

从图2可知,海拔是回归树第一层节点上的分类变量,说明海拔对地力变异起主导作用。耕层厚度和地形部位均为第二层节点上的分类变量,且二者在回归树中作为分类变量分别出现3次和2次,说明耕层厚度和地形部位对水稻田地力也有重要影响。

2.3 回归树预测能力

优化回归树具有14个叶子结点,如图3所示,样本空间最终划分为7个子样本,代表产量比有0.79、0.80、0.82、0.84、0.87、0.88和0.89这7个级别。箱图给出了7个级别的叶子节点各自的预测能力。以比率0.79为例,它包括了样本数为110、17和18这3个子样本,将其合并为样本数为145的子样本,这些样本的产量比如第一个箱图所示。由图3可知,在7个子样本中,预测得到的产量比中值与实际样品的产量比平均值十分接近,说明回归树方法对双季稻田的地力特征具有很好的模型效果。

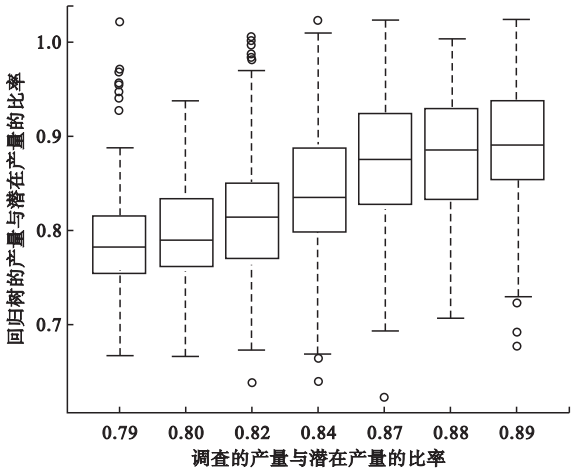


图3 优化后回归树预测能力的箱图

Fig. 3 Boxplot of predictive ability by pruned regression tree

3 讨 论

3.1 采用回归树法分析地力特征

本研究表明,耕地的地形,如海拔、地形部位、母质等是调控稻田地力的上层因子,而一些受耕作措施影响的因子如耕层深度、土壤有机碳含量、排灌方式等却处于影响稻田地力的下层,这反映出我国南方丘陵地区耕地地形如海拔、地形部位等可能通过影响农民的水肥投入而影响地力。比如,位于高海拔和丘陵坡地的稻田,可能对施用粪肥等农家肥造成困难,同时水源也将受到限制,从而不利于稻田地力的提高。海拔和地形部位作为上层因子,重要性在其他地区也有所体现。在东南亚大部分旱稻田生长在缓坡区域,区域中由于几米高度差别的不同地形系列的土壤性状相异。不同高程会导致土壤性质和水文条件的不同,进一步可能会影响农作物的产量表现(Boling & Tuong,2008)。在台湾地区,农民常常在高程只有几米差异的坡地上种植作物,虽然这种海拔差异只有几米,但是这也会造成土壤性质和水文条件的不同(Hseu & Chen,2001;Tsubo *et al.*,2006)。有研究表明,地形部位较高的区域水稻产量低于地形部位较低的区域(Boling & Tuong,2008)。这与本文的研究结果相同。然而,耕地地形特征在中国大陆地区以往的地力评价中,通常受到的重视程度不够。如在对重庆梁平县的耕地地力评价研究中,海拔和地貌类型对地力的影响程度仅排位于倒数第一和第二位(刘永文等,2009)。在对湖北省鄂州市、江陵县、荆州区的地力评价中,地形

在对地力权重的影响中也同样排位于倒数第一或第二的位置(鲁明星,2007)。本研究表明,立地条件虽然不直接影响作物产量而影响地力,但却会通过影响土壤水肥投入、农田基础设施建设的难易程度(或成本)来间接地对地力产生重要影响。可见,地形在地力评价中应该受到高度重视。同时,回归树通过层层分级,可找出在特定的环境因子条件下,影响地力的关键因子,从而为地力提升调控提供了理论依据。

值得注意的是,本研究,耕层厚度作为回归树的分类变量出现过 3 次,要高于其他任何一个因子出现的次数。具体表现为,当耕层厚度小于某一数值时,产量比会出现显著下降,成为影响地力的一个很重要因素。这说明耕层厚度变浅在研究区的水稻生产中带有一定的普遍性,而当地日益普遍使用的机耕状况可能是导致耕层变浅的重要原因。以往的研究表明,耕层厚度影响着土壤水分、养分的总存储量和根系分布的空间范围,是决定农作物高产稳产的重要因素之一。通常土壤疏松、深厚,作物根系分布也深,且能吸收较多的水分和养分,作物生长良好,抗逆性强。所以,作物产量和生物量会随着耕层的增加而提高,随着表层的移失而降低(李浩和高鹏,2010)。近年来,各地农业机械化程度不断提高,越来越多的耕地由于采用小型或轻型机械操作的方式,耕层厚度逐渐变浅的现象也屡见不鲜。鉴于耕层厚度对稻田地力具有重要的影响,开发出适宜深耕的稻田耕作机械,防止土壤耕厚度变浅和板结,将对稳定和提高水稻产量有着不可轻视的良好效果。

关于长沙县水稻田的管理,应当因地制宜,采取相应的耕地地力定向培育与土壤改良措施。对涝渍潜育型低产田主要是加强排水设施建设,通过排除田间渍涝,降低地下水位,使土壤逐步脱潜。对干旱灌溉型低产田主要是解决干旱缺水问题。对水源困难,全部依靠降水解决灌溉的望天田,应主动调整种植结构,通过改制改种,变高耗水作物为节水抗旱作物,变被动抗旱为主动抗旱。对障碍层次型低产田主要是消除障碍层次,如针对砂层,主要措施是培土,增施有机肥和科学施用化肥,合理轮作,改善灌溉条件,增强抗旱能力;对重砂、重粘改良培肥型低产田,主要是针对质地粘重或砂重的问题,广辟有机肥源,增施有机肥,改革耕作制度,实行水旱轮作,逐步加深耕作层,测土配方施肥,协调土壤养分,有条件的地方通地客土改良土壤质地。对坡地梯改型低产田主要是从根本上改变顺坡种植的习惯,变坡地为梯田,防止水土流失。

3.2 回归树分析地力特征的特点

回归树分析是一种简单实用的数据挖掘技术,通过递归地划分自变量空间和用验证数据进行剪枝,构建二叉树,从而达到预测的目的(陈辉林和夏道勋,2011)。它是一种新型的多因素分析方法,其结果不受变量间共线性的影响,可用于表征多种预测变量对响应变量变异的相对贡献(刘洋等,2010)。此方法具有以下特点:能对不同类型(如数值型、分类型等)的响应变量进行分析;能进行交互式的解释、描述和预测;对存在相互作用的复杂结果可以用简单的图表来解释;通过交叉检验来选择模型;能处理缺失数据(De'ath & Fabricius,2000)。与单纯的回归相比,分类与回归树也省去了选择变量的麻烦,因为在构建树的过程中它会自动选择最优划分变量(谢益辉,2007)。由回归树方法构建的预测树很简单,很多情况下比常用的统计方法构建的代数学预测准则更准确。事实上,数据越复杂,变量越多,回归树法相比其他方法的优越性就越显著(张立彬等,2002;谢益辉,2007)。

一个地区的地力往往受多个因子(包括地形地貌、灌溉、土壤培肥等)的调控,且因子之间存在着复杂的相互关联,这就需要一种合适的统计方法来剖析它,以便有效地服务于地力提升。以往的统计方法,如多元线性回归、主成分分析等,能找出影响地力的主要因子,但却不能找出具体条件下影响地力的某一关键因素,不利于地力提升工作的开展。回归树方法通过层层分解,可以用简单的图表对复杂的相互关系进行解释,能够对特定条件下制约地力的关键因子进行解析,从而有利于针对性的开展地力提升工作。本文的研究正是基于大量的数据信息,因此将回归树方法应用于中国广大水稻田区的地力评价工作,将能对促进中国耕地地力提升起到很好的指导作用。

本研究中,修剪后的回归树模型的拟合程度和预测还不是很理想。由于采样点多和分布广的限制条件,致使在信息采集过程中可能有不准确或不统一的现象发生,这可能是影响准确度的因素之一。在以后的调查中应尽量保持一致和准确,减少误差。当然回归树也并非最完美的统计方法,但是,随着时间推移,各种方法的交叉运用也使它的缺陷在逐渐改善,例如预测的准确度问题,可以通过 Bagging

(Breiman *et al.*,1984)等方法来提高。

致谢 感谢中国科学院亚热带农业生态研究所王毅老师和邹刚华博士对本论文提出宝贵意见。

参考文献

蔡 晖. 2011. 基于树模型回归算法在预测问题中的研究. 软件导刊, **10**(10): 58-60.

陈 磊. 2011. 中国18亿亩耕地红线遭严峻挑战. [EB/OL]. [2012-6-10]. http://house.ifeng.com/news/detail_2011_02/25/4849426_0.shtml

陈辉林, 夏道勋. 2011. 基于CART决策树数据挖掘算法的应用研究. 煤炭技术, **30**(10): 164-165.

范 洁, 杨岳湘. 2005. 决策树后剪枝算法的研究. 湖南广播电视大学学报, (1): 54-56.

李 浩, 高 鹏. 2010. 耕层深度与土壤改良对作物产量及土壤特性的影响. 水土保持应用技术, (6): 1-3.

刘 洋, 吕一河, 郑海峰, 等. 2010. 用回归树模型分析陕北黄土丘陵沟壑区气候因子对NDVI变异的影响. 应用生态学报, **21**(5): 1153-1158.

刘永文, 樊 燕, 刘洪斌. 2009. 丘陵山地耕地地力评价研究. 中国农学通报, **25**(18): 420-425.

鲁明星. 2007. 湖北省县域耕地地力评价及其应用研究. 湖北: 华中农业大学.

钱 松. 2011. 环境与生态统计: R语言的应用. 北京: 高等教育出版社.

王代星, 张小平, 王翰虎. 2010. 基于决策树结构特性的后剪枝技术研究. 电脑与信息技术, **18**(4): 1-4.

王琴竹. 2011. 决策树算法中过度拟合的研究与解决方案. 运城学院学报, **29**(2): 53-57.

武艳华, 史宝林, 葛丽平. 2008. 分类与回归树分析方法及其在医学研究中的应用. 河北北方学院学报, **25**(6): 72-73.

谢益辉. 2007. 基于R软rpart包的分类与回归树应用. 统计与信息论坛, **22**(5): 67-70.

姚凤梅, 张佳华, 孙白妮, 等. 2007. 气候变化对中国南方稻区水稻产量影响的模拟和分析. 气候与环境研究, **12**(5): 659-666.

张立彬, 张其前, 胥 芳, 等. 2002. 基于回归树(CART)方法的统计解析模型的应用与研究. 浙江工业大学学报, **30**(4): 315-318.

郑海峰. 2009. 典型黑土区农田生态系统生产力的多尺度变异性研究 ——以黑龙江省海伦市大豆系统为例(博士学位论文). 北京:中国科学院研究生院.

Boling AA, Tuong TP. 2008. The effect of toposequence position on soil properties, hydrology, and yield of rainfed lowland rice in Southeast Asia. *Field Crops Research*, **106**: 22-33.

Breiman L, Friedman JH, Olshen R, *et al*. 1984. Classification and Regression Trees. Wadsworth New York: Chapman and Hall.

Clark LA, Pregibon D. 1992. Statistical Models in S. Monterey: Chapman and Hall/CRC.

De'ath G, Fabricius KE. 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, **81**: 3178-3192.

Hseu ZY, Chen ZS. 2001. Quantifying soil hydromorphology of a rice-growing Ultisol toposequence in Taiwan. *Soil Science Society of America Journal*, **65**: 270-278.

Tsubo M, Basnayake J. 2006. Toposequential effects on water balance and productivity in rainfed lowland rice ecosystem in Southern Laos. *Field Crops Research*, **97**: 209-220.

White AB, Kumar P, Tcheng D. 2005. A data mining approach for understanding topographic control on climate-Induced inter-annual vegetation variability over the United States. *Remote Sensing of Environment*, **98**: 1-20.

作者简介 王莉莎,女,1988年生,硕士研究生,主要研究方向为区域环境模拟的研究。E-mail: 80237215@qq.com

责任编辑 李凤芹