浙 江 农 林 大 学 学 报，2019，**36**(3)：515–523
Journal of Zhejiang A&F University
doi:10.11833/j.issn.2095-0756.2019.03.012

基于优化 *k*-NN 模型的高山松地上生物量遥感估测

谢福明，字　李，舒清态

（西南林业大学 林学院，云南 昆明 650224）

摘要：针对传统 *k*-最近邻法(*k*-nearest neighbor, *k*-NN)在搜索最近邻单元时赋予特征变量相等的权重，缺少对特征变量加权优化等不足问题，在云南省香格里拉市，以高山松 *Pinus densata* 为研究对象，基于 49 块实测标准地，116 株高山松样木和 Landsat 8/OLI 影像，在前期进行基于遗传算法(genetic algorithm, GA)优化的 *k*-NN 模型实现的基础上，对 *k*-NN 的 3 个参数(*k*，*t* 和 *d*)进行反复测试优化组合，在像元尺度上对研究区高山松地上生物量进行遥感估算。结果表明：基于遗传算法优化的 *k*-NN 模型精度优于传统的 *k*-NN 模型，优化前均方根误差为 30.0 t·hm^{-2}，偏差为–0.418 t·hm^{-2}，相对标准误差百分比(R_{MSE})为 54.8%；优化后均方根误差为 24.0 t·hm^{-2}，偏差为–0.123 t·hm^{-2}，R_{MSE} 为 43.7%。基于优化 *k*-NN 模型的研究区高山松地上生物量总储量估测结果为 0.89×10^7 t。图 7 表 6 参 20

关键词：森林测计学；*k*-NN 模型；遗传算法；Landsat 8/OLI；地上生物量；高山松
中图分类号：S758.5　　**文献标志码**：A　　**文章编号**：2095-0756(2019)03-0515-09

Optimizing the *k*-nearest neighbors technique for estimating *Pinus densata* aboveground biomass based on remote sensing

XIE Fuming, ZI Li, SHU Qingtai

(College of Forestry, Southwest Forestry University, Kunming 650224, Yunnan, China)

Abstract: For the traditional *k*-nearest neighbor (*k*-NN), there are insufficient problems that give the weight of the feature variables equally when searching the nearest neighbor population units and a lack of weight vectors for the feature variables. In this study, Shangri-la City, Yunnan Province, was selected as the research area, and *Pinus densata* was taken as the research object. Based on 49 field data plots, 116 *P. densata* data samples, and Landsat 8/Operational Land Imager (OLI) imaging, a genetic algorithm was used to optimize the *k*-nearest neighbor model in the early stages, and the aboveground biomass of *P. densata* in the study area was estimated at the pixel scale after the *k*-NN three parameters (*k*, *t*, and *d*) were repeatedly tested and optimized. Results showed that accuracy of the *k*-NN model optimized by a genetic algorithm was better than the traditional *k*-NN model. Before optimization, the root mean square error was 30.0 t·hm^{-2}, deviation was –0.418 t·hm^{-2}, and R_{MSE} was 54.8%; after optimization, the root mean square error was 24.0 t·hm^{-2}, deviation was –0.123 t·hm^{-2}, and R_{MSE} was 43.7%. Finally, the estimated total aboveground biomass of *P. densata* in the study area was 0.89×10^7 t based on the optimized *k*-NN model. [Ch, 7 fig. 6 tab. 20 ref.]

Key words: forest mensuration; *k*-NN model; genetic algorithm; Landsat 8/OLI; aboveground biomass; *Pinus densata*

大气中温室气体浓度上升引起的全球气候变化，导致极端气候事件频发，严重威胁着人类生存与社会经济的可持续发展，成为各国政府和科学家关注的重大环境问题。在应对全球气候变化背景下，森林

收稿日期：2018-05-23；修回日期：2018-09-10
基金项目：国家林业公益性行业科研专项(201404309)；国家自然科学基金资助项目(31460194，31060114)；云南唐守正院士工作站资助项目
作者简介：谢福明，从事地理信息系统与遥感应用研究。E-mail: geoxfming@qq.com。通信作者：舒清态，副教授，博士，从事“3S”技术及森林景观经营研究。E-mail: shuqt@163.com

碳汇的相关研究成为科学界关注的热点[1-3]。生物量是森林生态系统碳汇潜力评估的重要基础，如何快速、准确地获取森林生物量信息，在 20 世纪 90 年代就成了森林生态系统与全球气候变化研究的关键[4]。准确评估森林碳储量的时空变化，不仅可以为森林资源的经营管理和林业可持续发展提供的科学依据，而且对碳循环及碳汇研究具有重要的意义。随着遥感技术的不断发展，利用数学模型结合实测样地数据进行生物量的大尺度快速估测变得有效可行。*k*-最近邻法(*k*-nearest neighbor, *k*-NN)作为一种非参数方法，已被广泛用于多源林业调查和森林参数估计的反演。1990 年，TOMPPO[5]首次将 *k*-NN 技术应用于芬兰森林资源监测中并取得了较好的效果。MCROBERTS[6]记录了该技术在国际范围内被广泛用于林业应用领域，包括森林调查空间插值预测、数据库监测、反演制图、小区域估测和统计推理。从数据层面上来讲，*k*-NN 与 Landsat 影像，机载激光扫面数据和 MODIS 数据联合使用估测评价森林属性的研究较多，并且将机载激光扫描指标等主动遥感变量与光学遥感、大尺度森林变量等参数结合使用有助于提高 *k*-NN 模型的预测精度[7]。国外研究者在遗传算法的优化下，利用 *k*-NN 和机载激光扫描数据对森林资源调查、森林参数估测与评价等方面取得了较好的研究成果[8-10]。KATILA 等[11]和 TOMPPO 等[12]运用数字地图进行数据分层和使用遗传算法对特征变量进行加权来作为一种提高预测精度的手段后，该方法得到了加强。利用遗传算法对卫星影像数据特征变量加权优化将会提高估测精度，并且将优化好的模型应用于单一森林属性变量(如某个树种)比同时应用于多变量的精度会提高许多[13]。然而，国内的研究学者缺少对 *k*-NN 模型算法进行优化改良的研究，仅局限于将传统的 *k*-NN 运用于不同的森林参数估计。如陈尔学等[14]运用 Landsat 数据和传统的 *k*-NN 法对小面积统计单元森林蓄积量估测，其结果表明采用 *k*-NN 法对县市级统计单元森林参数的估测效果明显优于只利用固定样地数据的传统参数估测方法。郭颖[15]利用 *k*-NN 非参数回归模型对甘肃省西水林场的森林地上生物量进行估测，并用随机森林算法(RF)进行特征选择后估测精度得以提升，优化后的算法在处理错误样本时具有良好的容错能力。本研究使用遗传算法对 *k*-NN 模型进行优化，使模型预测结果的偏差、均方根误差等最小化，以期提高模型的估测精度，实现对研究区高山松 *Pinus densata* 地上生物量储量估计与空间反演制图。

1 研究区概况

研究区位于滇西北迪庆藏族自治州香格里拉市境内(26°52′11.44″~28°50′59.57″N，99°23′6.08″~100°18′29.15″E)(图 1)。研究区地势高耸，热量不足，气温偏低，海拔为 1 503~5 545 m，多年平均气温为 5.5 ℃，历年平均降水量为 618.4 mm，平均降雪日为 35.7 d，年日照率为 40%~50%，属山地寒温带季风气候。境内密集的金沙江水系支流、冰雪融水和高原湖泊等水资源以及以棕壤、红壤为主的森林土壤类型孕育了丰富的植物资源。森林植被面积大，覆盖率高，南北差异分布明显，主要分布有 10 种植被类型，常见的树种有云杉 *Picea asperata*，冷杉 *Abies fabri*，高山松，云南松 *Pinus yunnanensis* 和高山栎 *Quercus semicarpifolia* 等。其中，高山松适应性广，更新能力强，是喜光、耐旱、耐瘠薄的优势树种。一般分布于云杉、冷杉林下限，海拔为 2 800~3 500 m，林分外貌整齐，成片分布，以同龄单层林常见，占全市乔木林面积的 22.7%。

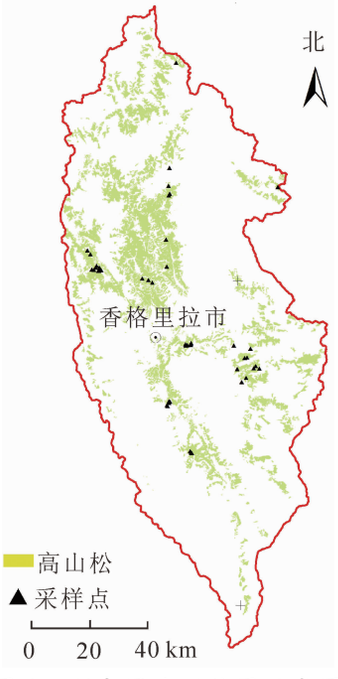

图 1 研究区地理位置示意图

Figure 1 Location of the study area

2 数据与方法

2.1 遥感数据及信息提取

从地理空间数据云(http://www.gscloud.cn/)获取 Landsat 8/OLI 影像 3 景覆盖整个研究区：2015 年 11 月 9 日(2 景)，轨道号分别为 132/040 和 132/041；2015 年 12 月 20 日(1 景)，轨道号为 131/041(图 1)。并采用软件 ENVI 5.3 对卫星影像进行辐射定标、大气校正(FLAASH)和几何精校正等预处理后提取单波段、多波段组合、主成分变换、缨帽变换、植被指数、纹理和地形特征（由 DEM 提取)等共计 123 个因子，作为建模因子备选参数(表 1)。

表 1　遥感因子一览表

Table 1　A list of factors derived from remote sensing

变量	数量	公式及说明
ρ_{Bi}	6	Landsat 8/OLI 数据第 i 波段原始发生率 ρ_{Bi}（i=2, 3, 4, 5, 6, 7）
V_{IS234}	1	$V_{IS234}=\sum_{i=2}^{4}\rho_{Bi}$
A_{lbedo}	1	$A_{lbedo}=\sum_{i=2}^{7}\rho_{i}$
P_{CAj}, I_{CAj}, M_{NFj}	9	分别为主成分分析、独立主成分分析、MNF 变换的第 j 成分（j=1, 2, 3）
T_{CB}, T_{CG}, T_{CW}	3	分别为缨穗变换的亮度、绿度、湿度分量
D_{VI}	1	差值植被指数：$D_{VI}=\rho_{NIR}-\rho_{R}$，$\rho_{NIR}$，$\rho_{R}$ 分别为近红外波段、红波段的反射率
N_{DVI}	1	归一化植被指数：$N_{DVI}=(\rho_{NIR}-\rho_{R})/(\rho_{NIR}+\rho_{R})$
E_{VI}	1	增强植被指数：$E_{VI}=2.5\left[\frac{(\rho_{NIR}-\rho_{R})}{(\rho_{NIR}+6.0\rho_{R}-7.5\rho_{BLUE}+1)}\right]$，$\rho_{BLUE}$ 为蓝波段的反射率
R_{VI}	1	比值植被指数：$R_{VI}=(\rho_{NIR}/\rho_{R})$
S_{AVI}	1	土壤调节植被指数：$S_{AVI}=\frac{(1+L)(\rho_{NIR}-\rho_{R})}{(\rho_{NIR}+\rho_{R}+L)}$，$L$ 为土壤调节系数，因研究区植被覆盖率大，本研究取 0.25
$B_{i\_N\_T}$	96	纹理特征，即第 i 波段 $N\times N$ 窗口下的纹理滤波 T。i=2, 3, 4, 5；N=3, 5, 9；T 为纹理滤波，依次分为：均值 ME，方差 VA，协同性 HO，对比度 CO，相异性 DI，信息熵 EN，二阶矩 SM，相关性 CR
$E_{levation}$	1	海拔
S_{lope}	1	DEM 派生的坡度因子

2.2　地面实测数据及处理

地面实测数据 49 块标准地和 116 株高山松样木数据（表 2）：实测标准地数据于 2014 年 10–11 月，在云南省香格里拉市境内的高山松分布范围内采集，在高山松分布范围布设了 49 个大小为 30 m × 30 m 的样地，记录了树高、胸径、样地差分 GPS 定位坐标和海拔等。其中：林分地上生物量依据式（1）进行计算。

$$W=0.095\ 5(D_{BH}^{2}H)^{0.832\ 9}。\tag{1}$$

高山松样木数据记录了不同龄组下（包括幼龄林、中龄林、近熟林、成熟林、过熟林）116 株高山松胸径（D_{BH}）和树高（H），并测定了树干、树皮、树叶、树枝、树冠生物量，用于拟合高山松地上生物量计算模型。本研究中的地上生物量由树干、树枝和树叶 3 个部分的生物量构成，生物量调查参照胥辉等[16]生物量测定方法。

表 2　生物量实测数据基本信息表

Table 2　Basic information of biomass measured data

变量	样木数据（N=116）			标准数据（N=49）	
	树高/m	胸径/cm	单株地上生物量/kg	标准树高/m	标准胸径/cm
均值	15.061	24.094	276.381	9.275	15.295
最大值	33.00	76.00	2 058.50	14.77	23.10
最小值	4.20	5.60	4.03	5.61	8.62
标准差	6.480	14.082	370.847	2.092	3.373

首先，采用随机抽样法将 116 株样木数据分成 2 个部分：2/3 样本作为建模样本建立生物量估算模型，1/3 作为检验样本对模型进行检验。其次，采用相对生长模型（非线性模型），运用最小二乘法对高山松单木地上生物量（W）模型进行拟合，结果见式（1），拟合决定系数 R^2 为 0.980 7，均方根误差 R_{MSE} 等于 46.73 kg，模型的验证结果如图 2 所示，检验决定系数 R^2 等于 0.995 7。

2.3　基于传统和优化 k-NN 模型的生物量估测

2.3.1　*传统k-最近邻法（k-NN）*　在 k-NN 的专业术语中，将待测变量及其特征变量的观测值样本指定为参考集，将待测变量的预测集指定为目标集，特征变量定义的空间成为特征空间。对于诸如生物量或蓄积量等连续性变量 M 在像元 p 上的预测值 m_p 的计算方法如下：

$$m_p=\sum_{i=1}^{k} w_{ip} m_i 。 \tag{2}$$

式(2)中：m_i 为变量 M 参考样地点 i 上的实测值；k 为计算预测值 m_p 时考虑的近邻个数；w_{ip} 为像元权重值，其计算如下：

$$w_{ip}=\begin{cases} d_{\mathrm{pwp}}^{-t} \Big/ \sum_{j=i_1(p)}^{i_k(p)} d_{p_j p}^{-t} & ，当且仅当 i\in\{i_1(p),\cdots,i_k(p)\} \\ 0，其他情况 \end{cases} 。 \tag{3}$$

式(3)中：i 是参考集样本；p 是目标集像元；p_j 是与参考集样本 j 对应的样本；d_{pwp}^{-t}为距离分解因子；k, t 为常量，一般通过实验反复测试选取最佳值；$\{i_1(p),\cdots,i_k(p)\}$是与待测像元 p 在特征空间上最相似的 k 个参考集样本。特征变量空间相似度由度量 $d_{pi,p}$，其计算方法如下：

$$d_{p_i,p}=\sqrt{\sum_{l=1}^{n_f}\omega_{lf}(f_{l,p_j}-f_{l,p})} 。 \tag{4}$$

式(4)中：f_{l,p_i}和$f_{l,p}$ 分别为参考集和目标集样本对应的遥感影像光谱波段及其派生因子等特征变量；n_f 为特征变量个数；p 为目标集像元；p_i 为参考集样本 i 对应的像元；ω_{lf} 为赋予特征空间中第 l 个特征变量的权值。

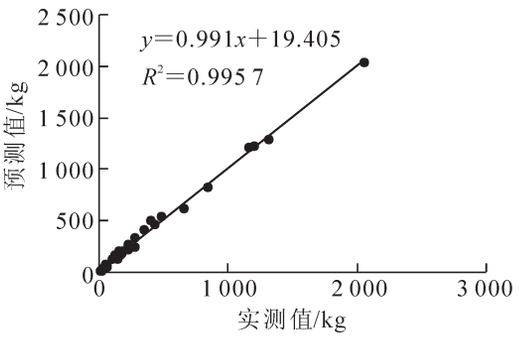

图 2 高山松单木地上生物量模型验证

Figure 2 Validation of *Pinus densata* aboveground biomass model

2.3.2 优化 k-最近邻法(ik-NN) ik-NN 与 k-NN 在方法原理上是一样的，改进之处在于前者使用遗传算法赋予了特征空间里的所有变量一个评价其重要性指标的权重向量，即式(4)中的 ω_{lf}；而后者则赋予所有变量相同的权值。优化的非单位矩阵 ω_{lf} 降低了不相关因子对因变量的影响，间接的起到了因子筛选的作用。

遗传算法优化 ω_{lf} 过程：(1)初始化。便于描述，将初始化权重向量群体比作染色体群体，权重向量的元素个体比作基因。随机生成大小为[n_{pop}, n_{f}]的数组作为初始化群体，运用二进制(0/1)对基因进行编码，并计算每一个染色体的适应度 γ，其计算公式见式(5)，用于对初始染色体及子代染色体选择的评价指标。(2)选择。采用随机遍历采样，根据自定义选择概率 p_{s} 将已有的优良染色体复制后添入新染色体群体中，删除劣质染色体；染色体是否被选择的依据是其适应度的大小，适应度大者被复制，小者被淘汰，确保新群体中的基因总数和初始群体相同。(3)交叉。利用交叉算子对染色体的基因编码进行重组，发生的概率为 p_{c}，通过交叉操作可以得到新一代染色体，子代的染色体组合了父辈的特性。交叉是遗传算法中最主要的操作，体现了信息交换的思想。(4)变异。变异首先在染色体群体中随机选择 1 个个体，对于选中的个体以突变概率 p_{m} 随机地改变其基因的编码。同生物界一样，遗传算法中变异发生的概率很低，通常取值很小。

2.4 模型的精度评价方法

留一法交叉验证，即对于 N 个样本，每次从 N 个样本中抽出 1 个样本作为测试集，利用剩余的 N−1 个样本作为参考集，重复 N 次循环，直至结束。本研究将 N 个样本的模型预测值$\hat{y}_i$=(i=1, …, N)与对应样本的实测值(y_i)进行统计分析，利用均方根误差$\hat{\sigma}$[式(6)]和偏差$\hat{e}$[式(7)]及相对标准误差百分比 R_{MSE}[式(8)]来检验模型的精度。

$$\gamma(\omega,\hat{\sigma},\hat{e})=\sum_{j=1}^{n_f}\hat{\sigma}_j(\omega)+\left|\sum_{j=1}^{n_f}\hat{e}_j(\omega)\right| ； \tag{5}$$

$$\hat{\sigma}=\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i-y_i)^2}{N}} ； \tag{6}$$

$$\hat{e}=\frac{\sum_{i=1}^{N}(\hat{y}_i-y_i)}{N} ； \tag{7}$$

$$R_{MSE}=\frac{\hat{\sigma}}{\bar{y}}\times 100\%。\tag{8}$$

式(5)~(8)中：γ 为遗传算法适应度；ω 为赋予特征变量的权值；n_f 为特征变量个数；y_i 和$\hat{y}_i$ 分别为第 i 个样本的实测值与模型预测值；$\bar{\hat{y}}$为模型预测值的平均值。

3　结果与分析

3.1　建模特征变量的筛选

筛选特征变量的目的在于：①降低特征空间的维数提高算法的运行速率，保证研究的可行性；②排除不相干变量、选择相关性显著的特征变量来提高模型的精度。在 SPSS 软件中分析特征变量与生物量之间的相关性显著水平，综合考虑特征空间的维度和模型精度后，从 123 个特征变量中选取 16 个与生物量极显著相关的特征变量作为建模变量。表 3 是将特征变量分为原始、显著相关和极显著相关 3 个等级后逐一评价的结果，客观地反映了不同特征变量等级下的模型精度。

表 3　不同特征变量等级下的模型精度对比

Table 3　Comparison of model accuracy under different level feature variables

特征变量等级	数量	$\hat{\sigma}$/(t·hm^{-2})	$\hat{e}$/(t·hm^{-2})	R_{MSE}/%
原始	123	33.96	0.03	61.6
显著相关	35	33.34	–2.70	63.6
极显著相关	16	29.95	–0.42	54.8
显著或极显著相关	51	34.52	0.01	62.6

3.2　模型参数优化配置

3.2.1　*k-NN 模型参数优化配置*　k-NN 模型需要确定 3 个重要的参数：评估特征变量空间相似度的距离参数$d_{p_n p}$；计算待测像元 p 的预测值时考虑的在特征空间上最相似的参考集样本个数 k 及其加权方案 w_{ip}。依据 CHIRICI 等[17]和谢福明等[18]的研究，在 k-NN 模型距离参数指标度量方式中，最常用的是欧氏距离(70%)，其次是马氏距离(3.5%)，以及典型相关分析度量(1.9%)，故本研究选取欧氏距离作为特征变量空间相似度的评价标准。k 的选择通常介于 1~10，本研究结合相应的实验数据选取的 k=5，如图 3A：$k\leqslant 5$ 时，模型精度随 k 的增大而提高，并在 k=5 时达最佳精度；$k>5$ 时，模型的精度逐渐降低。t 即距离分解因子，在模型中的应用见式(3)，t 通常取 0~2 内的值，其对模型精度的影响较小(图 3B)，本研究 t 取 2。

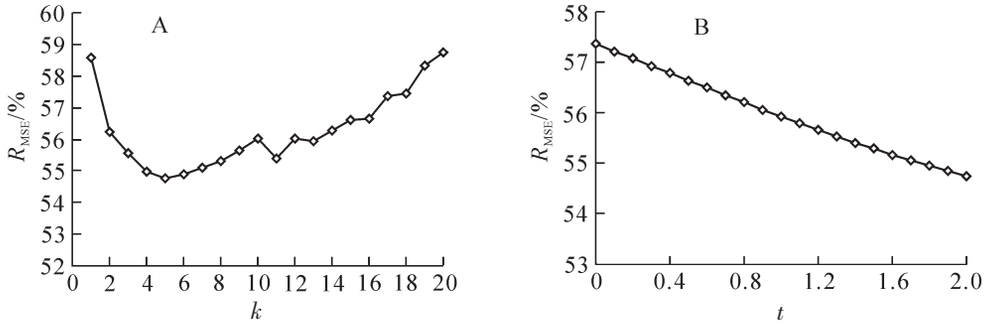

图 3　k-NN 模型精度随 k 和 t 的变化曲线

Figure 3　Change curve of model accuracy with the value of k and t

3.2.2　*遗传算法参数说明*　遗传算法最终的目的是为每一个特征变量计算出权重，并将其运用于 k-NN 模型来提高生物量的预测精度。算法中的主要函数调用于 Sheffield 遗传算法工具箱，其中的参数值在实验中反复测试、调试后确定。其中，图 4 表明了适应度 $\gamma(\omega, \hat{\sigma}, \hat{e})$随着遗传迭代次数(10，30 或 50)的增加呈缓慢下降，当迭代次数大于 50 时，适应度随迭代次数的变化比较平稳，并趋向于稳定。表 4 记录了算法的最佳初始参数值和调用的主要算子。

表 4 遗传算法有效参数值与主要算子汇总

Table 4 Parameters and main functions of genetic algorithm

自定义有效参数值	主要算子(算法调用于 Sheffield 遗传算法工具箱)
初始化染色体群体个数 n_{pop}：50	crtbp.m，创建任意离散随机种群
遗传迭代次数 n_{gen}：30~80	bs2rv.m，二进制串到实值的转换
染色体选择操作概率 p_s：0.95	ranking.m，基于排序的适应度分配
染色体基因交叉操作概率 p_c：0.7	sus.m，随机遍历采样选择方式
染色体变异操作概率 P_m：0.01	xovsp.m，单点交叉；mut.m，离散变异
优化权重上限值：0.5	reins.m，一致随机和基于适应度的重插入

3.3 模型效果分析

k-NN 模型及其优化算法在 MATLAB 环境下调试、运行，算法给予每个特征变量初始化的权重值均相等(第 0 代)，表 5 为第 50 代优化的特征变量权重值(算法的参数设置同 2.2 所述)，表 5 数据为标准化后的数值，其和为 1。

本研究的主要目标是通过优化方法降低像元尺度下模型的估测误差，提高对高山松地上生物量的估测精准度。表 6 和图 5 表明：(1)基于传统 *k*-NN 模型的样本生物量预测结果为 16.2~92.6 t·hm^{-2}，平均值为 54.7 t·hm^{-2}，模型均方根误差为 30.0 t·hm^{-2}，偏差为–0.418 t·hm^{-2}，R_{MSE} 为 54.8%(图 5A)；(2)遗传算法优化后的 *ik*-NN 模型精度得到了提升，均方根误差为 24.0 t·hm^{-2}，偏差为–0.123 t·hm^{-2}，R_{MSE} 为 43.7%(图 5B)。与传统 *k*-NN 模型相比，*ik*-NN 模型的精度均方根误差值降低了约 6.0 t·hm^{-2}，偏差下降比例达 75.6%，模型精度 R_{MSE} 提高了 11.1%；(3)*ik*-NN 模型的样本估计值为 23.3~95.2 t·hm^{-2}，在均值上与实测值比较相近，约 55.0 t·hm^{-2}。但对于高生物量或低生物量区域的估测残差仍较大，均出现高值低估，低值高估的现象。

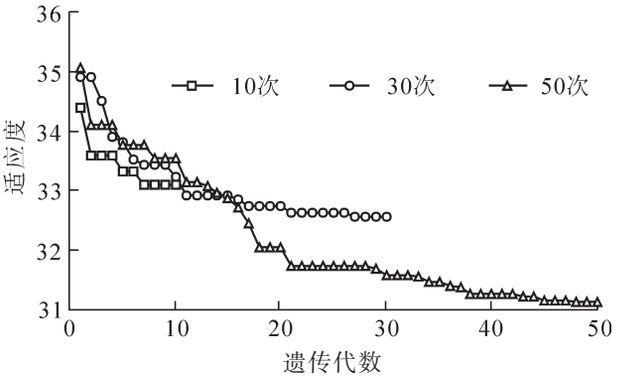

图 4 遗传算法优化中适应度值随遗传代数的降低曲线

Figure 4 Reduction of fitness value curve with the number of generations in optimization of genetic algorithm

3.4 生物量估计与反演

像元尺度下的定量反演是一项极其密集的任务，需要逐一计算研究区内的每一个像元，对计算机内存需求大，故本研究把研究区分成多块区域后再逐一估测反演。图 6 为 *k*-NN 和 *ik*-NN 2 个模型局部反

表 5 第 50 代优化的特征变量权重值(遗传代数为 50，上限值为 0.5)

Table 5 Values of the elements of the weight vector for feature variables for the 50th optimization (with upper bounds 0.5 and 50 generations)

项目	B_2	$B_{2\_3\_ME}$	$B_{2\_3\_HO}$	$B_{2\_3\_DI}$	$B_{3\_3\_HO}$	$B_{3\_3\_DI}$	$B_{3\_3\_EN}$	$B_{3\_3\_SM}$
权重	2.10×10^{-3}	2.50×10^{-2}	7.53×10^{-2}	1.41×10^{-1}	1.14×10^{-1}	1.24×10^{-1}	1.16×10^{-1}	6.12×10^{-2}
项目	$B_{4\_3\_ME}$	$B_{2\_5\_ME}$	$B_{3\_5\_ME}$	$B_{3\_5\_EN}$	$B_{3\_5\_SM}$	$B_{4\_5\_ME}$	$B_{2\_9\_ME}$	$B_{3\_9\_ME}$
权重	2.42×10^{-2}	2.88×10^{-2}	2.29×10^{-2}	2.75×10^{-2}	9.52×10^{-2}	2.59×10^{-2}	4.04×10^{-2}	7.50×10^{-2}

说明：$B_{i\_N\_T}$ 为纹理特征，即第 i 波段 $N\times N$ 窗口下的纹理滤波 T。纹理滤波依次分为：均值 ME，方差 VA，协同性 HO，对比度 CO，相异性 DI，信息熵 EN，二阶矩 SM，相关性 CR。如 $B_{2\_3\_ME}$，即第 2 波段 3×3 窗口下的均值(ME)纹理滤波，依次类推

表 6 高山松地上生物量实测值与模型预测值统计结果

Table 6 Statistics of observations and model predictions of aboveground biomass of *Pinus densata*

变量	生物量/(t·hm^{-2})			
	最小值	最大值	均值	标准差
样地实测	10.2	141.2	55.1	34.9
k-NN 预测	16.2	92.6	54.7	18.9
ik-NN 预测	23.3	95.2	55.0	20.1

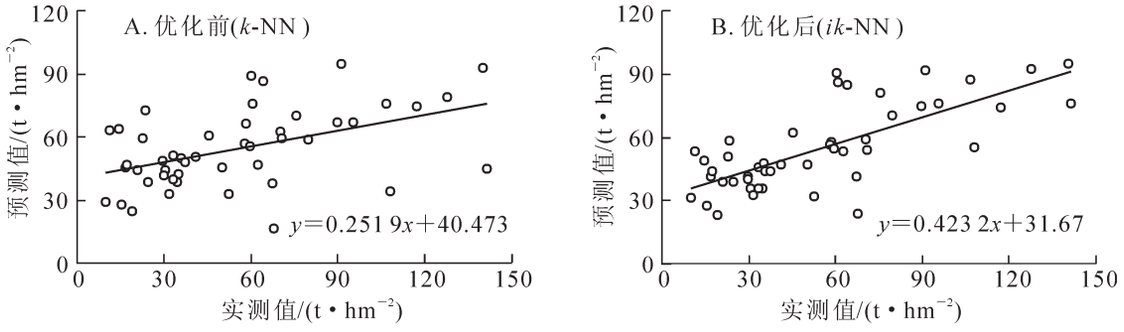

图 5　模型优化前后生物量的估测精度对比

Figure 5　Comparison of estimation accuracy of aboveground biomass of *Pinus densata* between *k*-NN and *ik*-NN model

演结果：*k*-NN 模型的预测为 20.0~97.5 t·hm^{-2}，平均值为 49.5 t·hm^{-2}，标准差为 13.1 t·hm^{-2}(图 6A)；*ik*-NN 模型的预测值则为 18.4~113.7 t·hm^{-2}，平均值为 49.3 t·hm^{-2}，标准差为 13.5 t·hm^{-2}(图 6B)。模型的反演结果中离散分布的像元较少，近邻相关性好，更好地体现了变量的区域相关性。依据森林资源二类调查统计数据，研究区高山松分布区面积为 1.74×10^5 hm^2，其地上总生物量估测结果为 0.89×10^7 t。图 7 为 *ik*-NN 模型的高山松地上生物量空间分布等级图，生物量等级在 16.8~108.9 t·hm^{-2} 之内，主要分布在 45~75 t·hm^{-2}。

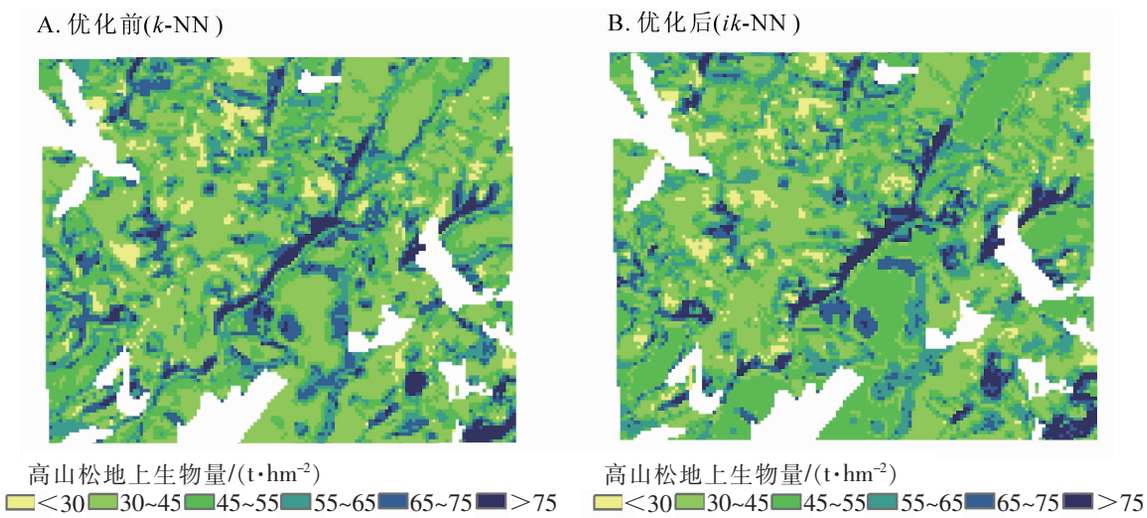

图 6　像元尺度下的*k*-NN/*ik*-NN 模型局部反演对比

Figure 6　Comparison of local inversion of *k*-NN/*ik*-NN model on pixel scale

4　结论与讨论

本研究使用遗传算法实现对 *k*-NN 模型中的特征变量赋予相应的权重值后，构建加权欧氏距离，结合卫星数据和地面实测样地数据建立了优化的 *k*-NN 估测回归模型，估算出香格里拉高山松地上生物量储量，反演出地上生物量分布等级图。结果显示：*k*-NN 算法参数 *k* 和 *t* 分别取值为 5 和 2 时，模型的预测效果最佳；基于遗传算法优化的 *ik*-NN 模型预测精度优于传统的 *k*-NN 模型，均方根误差为 24.0 t·hm^{-2}，偏差为−0.123 t·hm^{-2}，R_{MSE} 为 43.7%。研究区像素级水平下高山松地上生物量的预测值为 16.8~108.9 t·hm^{-2}，总估计值为 0.89×10^7 t。

CHIRICI 等[17]研究显示：使用卫星光谱数据作为特征变量时，需要大量的样本来获取较小的相对标准误差百分比，这与本研究结果相符合。本研究 *k*-NN 模型的参考样本偏少，且参考样本在空间分布上相对集中(图 1)，所以生物量的预测结果残差较大，出现高值低估，低值高估的现象；造成这一现象的另一个原因是 *k*-NN 法本身存在的缺陷，即只能局限于实测值范围内对未知单元进行估测，预测值不会超出实测值的范围，模型算法中 *k* 个参考样本间的加权求和降低了估计值的方差，从而产生了更大的估

测误差。但 *k*-NN 在大尺度区域上的森林资源监测中有很大的潜力，不仅适用于森林参数的估测反演，还适用于森林调查空间插值预测、数据库监测、小区域估测和统计推理等研究[19-20]，并且从以下方面做出突破可以有效提升其预测能力，为生活生产实践提供更好的技术借鉴：①*k*-NN 在搜索最近邻个体时应限制搜寻的范围，如限制一个搜寻半径或在指定的图斑区域，而不是全局搜索，充分利用区域化变量的特性来提高模型的估测精度。②利用地物光谱的差异性，结合星载、机载高光谱数据和地面实测高光谱数据或者其他能够区分地物的单波段，利用最近邻法或其他机器学习算法实现对地物的精细识别，提高区域尺度上的地物分类精度，进而提高对其生理生化参数定量估测的准确性。

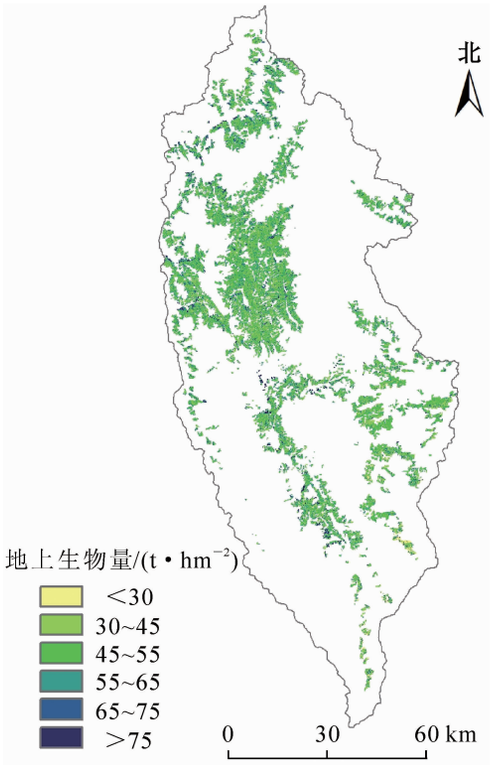

图 7 像元尺度下香格里拉市高山松地上生物量反演结果示意图

Figure 7 Spatial distribution of *Pinus densata* aboveground biomass in Shangri-la at the pixel scale

5 参考文献

[1] 王效科，冯宗炜. 中国森林生态系统中植物固定大气碳的潜力[J]. 生态学杂志，2000，**19**(4)：72 – 74.
WANG Xiaoke, FENG Zongwei. The potential to sequester atmospheric carbon through forest ecosystems in China[J]. *Chin J Ecol*, 2000, **19**(4): 72 – 74.

[2] 胡会峰，刘国华. 中国天然林保护工程的固碳能力估算[J]. 生态学报，2006，**26**(1)：291 – 296.
HU Huifeng, LIU Guohua. Carbon sequestration of China's National Natural Forest Protection Project [J]. *Acta Ecol Sin*, **26**(1): 291 – 296.

[3] 胡会峰，刘国华. 森林管理在全球 CO_2 减排中的作用[J]. 应用生态学报，2006，**17**(4)：709 – 714.
HU Huifeng, LIU Guohua. Roles of forest management in global carbon dioxide mitigation [J]. *Chin J Appl Ecol*, 2006, **17**(4): 709 – 714.

[4] 汤旭光，刘殿伟，王宗明，等. 森林地上生物量遥感估算研究进展[J]. 生态学杂志，2012，**31**(5)：1311 – 1318.
TANG Xuguang, LIU Dianwei, WANG Zongming, *et al*. Estimation of forest aboveground biomass based on remote sensing data: a review [J]. *Chin J Ecol*, 2012, **31**(5): 1311 – 1318.

[5] TOMPPO E. Satellite imagery-based national inventory of Finland [J]. *Int Arch Photogramm Remote Sensing*, 1991, **28**(7/1): 419 – 424.

[6] MCROBERTS R E. Estimating forest attribute parameters for small areas using nearest neighbors techniques [J]. *For Ecol Manage*, 2012, **272**(3): 3 – 12.

[7] MCROBERTS R E, NÆSSET E, GOBAKKEN T. Optimizing the *k*-Nearest Neighbors technique for estimating forest aboveground biomass using airborne laser scanning data [J]. *Remote Sensing Environ*, 2015, **163**: 13 – 22.

[8] MURA M, MCROBERTS R E, CHIRICI G, *et al*. Statistical inference for forest structural diversity indices using airborne laser scanning data and the *k*-Nearest Neighbors technique [J]. *Remote Sensing Environ*, 2016, **186**: 678 – 686.

[9] MCROBERTS R E, DOMKE G M, CHEN Q, *et al*. Using genetic algorithms to optimize *k*-Nearest Neighbors configu-

rations for use with airborne laser scanning data [J]. *Remote Sensing Environ*, 2016, **184**: 387 – 395.

[10] MCROBERTS R E, CHEN Q, WALTERS B F. Multivariate inference for forest inventories using auxiliary airborne laser scanning data [J]. *For Ecol Manage*, 2017, **401**: 295 – 303.

[11] KATILA M, TOMPPO E. Stratification by ancillary data in multisource forest inventories employing k-nearest neighbor estimation [J]. *Can J For Res*, 2002, **32**(9): 1548 – 1561.

[12] TOMPPO E, HALME M. Using coarse scale forest variables as ancillary information and weighting of variables in k-NN estimation: a genetic algorithm approach [J]. *Remote Sensing Environ*, 2004, **92**(1): 1 – 20.

[13] TOMPPO E, GAGLIANO C, NATALE F D, *et al*. Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery [J]. *Remote Sensing Environ*, 2009, **113**(3): 500 – 517.

[14] 陈尔学，李增元，武红敢，等. 基于 k-NN 和 Landsat 数据的小面积统计单元森林蓄积量估测方法[J]. 林业科学研究，2008，**21**(6)：745 – 750.
CHEN Erxue, LI Zengyuan, WU Honggan, *et al*. Forest volume estimation method for small areas based on k-NN and Landsat data [J]. *For Res*, 2008, **21**(6): 745 – 750.

[15] 郭颖. 森林地上生物量的非参数遥感估测方法优化[D]. 北京：中国林业科学研究院，2011.
GUO Ying. *Optimum Non-Parametric Method for Forest Aboveground Biomass Estimation based on Remote Sensing Data* [D]. Beijing: Chinese Academy of Forestry, 2011.

[16] 胥辉，张会儒. 林木生物量模型研究[M]. 昆明：云南科技出版社，2002.

[17] CHIRICI G, MURA M, MCINEMEY D, *et al*. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data [J]. *Remote Sensing Environ*, 2016, **176**(2): 282 – 294.

[18] 谢福明，舒清态，字李，等. 基于 k-NN 非参数模型的高山松生物量遥感估测研究[J]. 江西农业大学学报，2018，**40**(4)：743 – 750.
XIE Fuming, SHU Qingtai, ZI Li, *et al*. Remote sensing estimation of *Pinus densata* aboveground biomass based on k-NN nonparametric model [J]. *Acta Agric Univ Jiangxi*, 2018, **40**(4): 743 – 750.

[19] BEAUDOIN A, BERNIER P Y, GUINDON L, *et al*. Mapping attributes of Canada's forests at moderate resolution through k-NN and MODIS imagery [J]. *Can J For Res*, 2014, **44**(5): 521 – 532.

[20] MCROBERTS R E. Estimating forest attribute parameters for small areas using nearest neighbors techniques [J]. *For Ecol Manage*, 2012, **272**(3): 3 – 12.