林业科学研究 2008, 21(6): 745～750
Forest Research

文章编号: 1001-1498(2008)06-0745-06

基于 k-NN和 Landsat数据的小面积统计单元森林蓄积估测方法

陈尔学[1], 李增元[1], 武红敢[1], 韩爱惠[2]

(1. 中国林业科学研究院资源信息研究所,国家林业局林业遥感与信息技术重点开放性实验室,北京 100091;
2. 国家林业局调查规划设计院,北京 100714)

摘要:基于吉林省一个试验区的森林资源一类清查固定样地数据、Landsat TM数据和土地利用数据,采用精度交叉评价方法研究了 k最近邻(k-NN)法用于小面积统计单元森林蓄积估计的有效性。结果表明: k-NN方法对样地覆盖区影像像元单位面积蓄积量的估测平均误差在 1.5 $m^3 \cdot hm^{-2}$ 之内,相对均方根误差(*RMSE*)低于传统的基于绿度指数的线性方程估测方法;采用 k-NN方法可以实现县市级统计单元的参数估计,估测效果优于只利用固定样地数据的传统成数估计方法。

关键词: k-NN方法;森林蓄积量; Landsat;森林资源调查
中图分类号: S757.2 **文献标识码**: A

Forest Volume Estimation Method for Small Areas Based on k-NN and Landsat Data

CHEN Er-xue[1], *LI Zeng-yuan*[1], *WU Hong-gan*[1], *HAN Ai-hui*[2]

(1. Research Institute of Forest Resource Information Techniques, CAF, Key Laboratory for Forest Remote Sensing and Information Techniques of State Forestry Administration, Beijing 100091, China; 2. Academy of Forest Inventory and Planning, State Forestry Administration, Beijing 100714, China)

Abstract: The effectiveness of k-Nearest Neighbour (k-NN) for forest parameters estimation of small area was evaluated using permanent forest plot data of national forest inventory (NFI), Landsat TM data and landuse map data in a test site located in Jilin Province. It was found that the bias of the mean volume per unit area estimated using k-NN was under 1.5 $m^3 \cdot hm^{-2}$, and the relative root mean square error (*RMSE*) was less than that of the conventional lineal regress method based on the relationship between Landsat ETM + greenness index and forest volume density; k-NN could be used to estimate forest parameters of small unit in the scale of counties or districts, whose performance could be better than that of traditional population based statistic method that only utilizes forest plot data.

Key words: k-NN; forest volume; landsat; forest resources inventory

我国森林资源清查是基于固定样地资料以省为统计单位进行的,难于对更小面积单元(如县/市、国有林场等)的森林调查因子进行有效估计,调查结果也难于生成具有较大比例尺的森林资源分布图。遥感影像在森林资源清查中的应用尚局限于获取遥感判读样地信息,并应用于多阶分层抽样统计。虽然在国内森林蓄积量、郁闭度等参数的遥感估测技术在一、二类调查中都有应用,但所采用的方法主要是多

收稿日期: 2007-08-03;修回日期: 2008-02-26
基金项目: 林业科技支撑计划专题"三北、长防及沿海防护林体系建设工程监测技术研究(2006BAD23B0504)";863课题"森林资源遥感监测定量化综合处理与业务运行系统"资助
作者简介: 陈尔学(1968—),山东成武人,博士,副研究员,长期从事林业遥感应用技术研究.

元统计回归方法[1-2]。为了克服最小二乘估计的缺陷,有人曾研究评价过基于岭估计和稳健估计的蓄积量估计方法[3]。除了这些参数统计估测方法外,还有一些非参数统计方法,但这类方法中,目前在国内还只有神经元网络方法用于蓄积估测的研究报道[3]。

k最近邻(k-NN)法也是一种非参数统计方法,不仅可用于分类识别,也可以用于森林参数的定量估计。Tomppo[4]将 k-NN应用于芬兰多源国家森林资源清查(MS-NFI),提出了基于 k-NN的 MS-NFI技术(oMS-NFI),得到了很好的估测结果。在此基础上又发展了改进定标的 MS-NFI法(cMS-NFI)[5-6]和分层的 MS-NFI法(sMS-NFI)[7]等。基于 k-NN的多源调查方法具有及时、经济和精确的特点,目前北欧一些国家多采用该技术估测小面积单元的类型面积和森林参数,如蓄积量、生物量等。芬兰、瑞典等北欧国家,已将这种综合利用样地调查数据和遥感数据的多源森林资源清查方法付诸实际应用[8-11],美国也进行了初步的应用试验[12]。但国内对该技术的研究和应用还基本上处于空白。本文将针对 oMS-NFI在我国森林资源清查小面积统计单元森林蓄积估计中的应用开展初步评价研究。

1 材料与方法

1.1 试验区

试验区位于吉林省中下部,左上角地理坐标为 44°06′24.71″N, 124°17′55.92″E,右下角为 42°15′55.70″N, 126°34′32.24″E,为一景 Landsat TM影像所覆盖,但不包括 TM影像左下角辽宁省的小部分区域(图 1-a)。该区域林区主要优势树种按所占比重由大到小依次是落叶松(*Larix* spp.)、栎类(*Quercus* spp.)、樟子松(*Pinus sylvestris* var. *mongolica* Litv.)。试验区属于农林交错区,林地覆盖面积仅占 32.95%,而农田占 57.51%。

1.2 森林资源清查样地数据

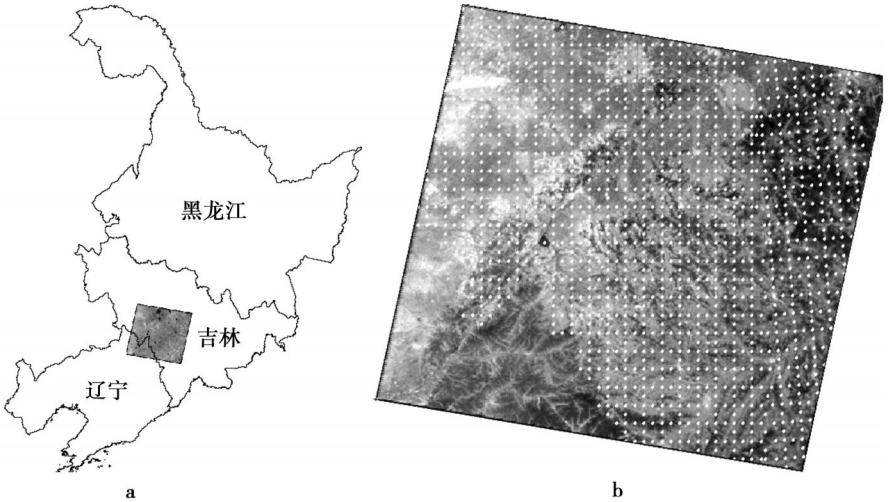
地面数据是我国第 6次全国森林资源清查固定样地,调查时间为 1999年。采用 4 km ×8 km间距系统布设正方形样地。每块样地的大小为 0.06 hm^2,按照资源清查样地调查规程实行外业调查,经整理后得到每块样地的属性:地类、优势树种类型、郁闭度、蓄积量、平均胸径、林龄等。覆盖试验区的样地总数为 1 276块(图 1-b)。

1.3 Landsat TM数据

Landsat TM数据获取时间为 1997年 6月 14日。以 1∶5万地形图为基准采集控制点,采用多项式法进行了几何精校正处理,几何校正误差在 2个 TM空间分辨率之内(小于 60 m)。几何校正后的影像像元大小为 30 m ×30 m。几何校正前曾按照国家林业局制定的《遥感影像处理技术规程》进行过简单的基于影像的辐射校正,但没有进行大气校正、太阳高度角校正处理。

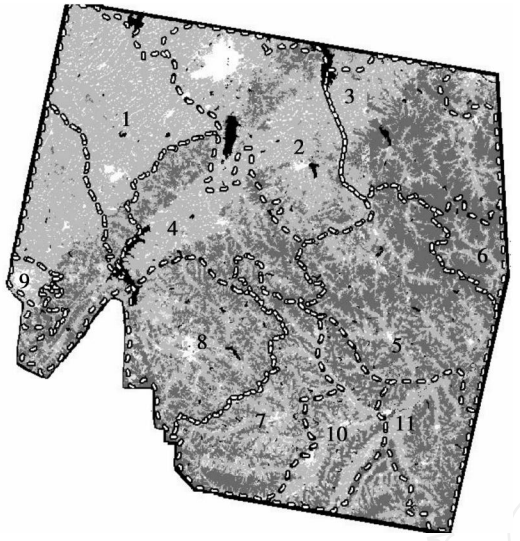
1.4 土地利用图

覆盖该试验区的数字化土地利用图,是在 1996年利用 Landsat TM影像通过计算机辅助分类和人工目视解译相结合生成的。本研究中将各地类归并为林地与非林地两大类,作为基于 k-NN估测小面积单元森林总蓄积量的输入数据(图 2)。



(a: 试验区为一景 Landsat TM影像所覆盖的吉林省的区域,主要在吉林省境内,小部分在辽宁境内; b: 在 TM影像上显示了所收集到的所有一类清查固定样地(白色小点), TM影像的右上角和左下角部分(位于辽宁省境内)没有固定样地分布)

图 1 试验区的位置及固定样地在 Lansat TM影像上的分布



(黑色为水体,深灰色为森林,浅灰色为农田,白色为城镇。每个单元用一个多边形表示,单元编号用黑色数字表示。)

图 2　试验区 11个小面积统计单元的边界,背景图为土地利用图

1.5　基于 k-NN的森林参数估测方法

遥感影像像元用 p表示,样地用 i表示,样地所在像元用 p_i表示。像元 p的影像光谱值用向量 X_p表示,本文直接利用 Landsat TM的 6个波段数据,因此 $X_p = [x_1, x_2, x_3, x_4, x_5, x_6]$。训练样地 i所在像元 p_i的光谱可用 X_{pi}表示。定义任一像元 p与 p_i光谱间欧氏距离为 $d_{pi,p}$,即

$$d_{pi,p} = |X_p - X_{pi}|^t \qquad (1)$$

其中 t是一指定常数,在本研究中 $t=2$。假设共有 N个样地,则可计算出像元 p与各样地对应影像像元光谱间的距离 N个,对 N个距离由小到大排序,排在最前面的 k个与 p具有最近距离的样地的集合可用$\{i_1(p), i_2(p), \cdots, i_k(p)\}$表示,则定义样地 i对像元 p的权重为

$$w_{i,p} = \begin{cases} \dfrac{1/d_{pi,p}^t}{\sum_{j\in\{i_1(p), i_2(p), \cdots, i_k(p)\}} 1/d_{pi,p}^t}, & \text{如果 } i\in\{i_1(p), i_2(p), \cdots, i_k(p)\} \\ 0, & \text{否则} \end{cases} \qquad (2)$$

则像元 p的观测变量 M的估计值

$$\hat{m}_p = \sum_{j\in\{i_1(p), i_2(p), \cdots, i_k(p)\}} w_{j,p} m_j \qquad (3)$$

其中,m_j是样地 j的观测变量 M的测量值;$\hat{m}_p$是观测变量 M在像元 p的估计值。例如若变量 M是单位面积蓄积量,则计算得到的就是该像元 p的单位面积蓄积量;若 M是样地的郁闭度,则估测结果就对应该像元 p的郁闭度,等等。但这并不是说 k-NN可用于估测固定样地的所有观测变量,只有那些和遥感数据光谱值有比较密切关系的样地调查因子才能被有效估测。

为了按统计单元(如行政区域)估计森林观测变量值,需要计算每个样地按单元 u统计的权重合计,$c_{i,u}$就表示样地 i对单元 u的权重,

$$c_{i,u} = \sum_{p\in u} w_{i,p} \qquad (4)$$

以 u为单元的观测变量 M的估计值 $\hat{m}_u$为

$$\hat{m}_u = \frac{\sum_{i\in Is} c_{i,u} m_i}{\sum_{i\in Is} c_{i,u}} \qquad (5)$$

这里,m_i表示样地 i的观测变量 M的测量值;Is是用于计算的层,如可以分别针叶林层、阔叶林层、落叶松层等估测 M的值。若 Is是针叶林层,则只有森林类型为针叶林的样地才参与计算,估计结果就是统计单元 u针叶林变量 M的估测值。上式分子乘以像元面积大小(30 m ×30 m)就得到单位 u内的 M变量的总量的估计(如总蓄积量)。分母乘以像元面积大小就是单元 u内层 Is的总面积估计。

这里介绍的是基于 k-NN进行样地观测因子(变量)的估测,其实 k-NN最早在遥感上的应用是地物的分类识别。因此 k-NN方法是一种可用于样地调查的多种因子(无论是类别还是连续变量)估测的非参数统计技术。当然在选择估测因子时,还要考虑这些因子和遥感影像光谱之间是否有一定的相关性,像样地记录的海拔、坡度、坡向等就不应该作为待估测因子,相反,若有这些因子的空间分布数据,可将他们作为用于计算权重的附加影像波段,可能还有助于估测精度的提高。这也正是 k-NN所具有的一大特点:按照统一的方法融合各种空间数据(土壤分布图、土壤湿度图、数字高程模型)到因变量的估测过程中。

若研究区域具备较为详细的地物类型分布图,可采用分层的 k-NN(如MF-NFI)估测方法。但本研究只采用了森林-非森林图层信息,只有属于森林图层的像元才进行蓄积量的估计,非森林图层的像元蓄积量估计值直接设为 0。参与 k-NN计算的固定样地只包括那些地类为林地、森林类型属于森林、样地总蓄积量不为零的样地,共有 307块。由于森林分布图不可避免有分类错误,样地、森林分布图和 TM影像间都也会存在一定的配准误差,因此这 307个样地还会有一些不落在森林图层内,但这些样地仍然用于 k-NN估测蓄积量的计算过程。

1.6 估测精度评价方法

由于本研究仅有 307个可用样地,对 k-NN估测精度的评价只能在像元尺度采用交叉评价方法进行。设样地总数为 N,每次从 N 个样地中不重复地抽出一个样地 i,利用剩余的 $N-1$ 个样地采用 k-NN估测样地 i所在像元的蓄积量值 Y_i,重复该过程共 N 次。设样地 i的蓄积量测量值为 X_i,N 次共得到 N 对 (X_i, Y_i), $i = 1, 2, \cdots, N$。则相对均方根误差 $RMSE$ 计算公式为:

$$RMSE' = \frac{\sqrt{\dfrac{\sum_{i=1}^{N}(X_i - Y_i)^2}{N}}}{\bar{X}} \times 100$$

其中 $\bar{X}$是 X的平均值。平均误差 $\bar{e}$为

$$\bar{e} = \frac{\sum_{i=1}^{N}(X_i - Y_i)}{N}$$

2 结果与讨论

2.1 单位面积蓄积估计精度评价

若将样地按优势树种或森林类型分层,就可以利用 k-NN估测不同树种或森林类型的单位面积蓄积量。表 1中“混合 ”一行结果是不按树种或森林类型分层,将所有的 307块样地混合在一起估测的森林单位面积蓄积量。其他几行是分别利用属于相应类型的样地得到的分类别估测结果。国家森林资源清查 (NFI)估计结果是直接对样地单位面积蓄积量观测结果取平均值得到。对 k-NN方法,这里试验了两种不同的 k取值:5和 10。当 k = 5时,单位面积蓄积量估计值误差最小为 -0.034 $m^3 \cdot hm^{-2}$,最大为 1.444 $m^3 \cdot hm^{-2}$;当 k = 10时,单位面积蓄积量估计值的误差最小值为 -0.814 $m^3 \cdot hm^{-2}$,最大为 1.333 $m^3 \cdot hm^{-2}$。这说明 k-NN方法对单位面积蓄积的估计误差是比较小的。无论是从 $RMSE$ 还是平均误差 $\bar{e}$上分析,k的大小对估计结果会有一定的影响。如,k = 10时,除了按照针叶林归类时的 $RMSE$ 增加外,其它几种方法的 $RMSE$ 都是降低的;从平均误差上对比,除了针叶林、阔叶林外,其它几组的平均误差也是降低的。因此在下面基于统计单元的定量参数估测中,将统一采用 k = 10的 k-NN方法。

表 1 像元级蓄积量估测精度交叉评价结果 (样地按树种组分层,精度检验也按树种组进行)

分层	NFI单位面积蓄积估计 / ($m^3 \cdot hm^{-2}$)	k = 5			k = 10		
		单位面积蓄积估测 / ($m^3 \cdot hm^{-2}$)	$RMSE$ '/%	平均误差 / ($m^3 \cdot hm^{-2}$)	单位面积蓄积估测值 / ($m^3 \cdot hm^{-2}$)	$RMSE$ '/%	平均误差 / ($m^3 \cdot hm^{-2}$)
混合	65.667	66.360	71.37	0.694	65.887 5	68.5	0.225
落叶松	58.539	59.528	71.32	0.990	59.475	67.9	0.934
栎树类	78.810	80.254	56.21	1.444	79.942 5	54.81	1.133
针叶林	53.738	53.715	73.74	-0.034	53.947 5	70.90	0.199
阔叶林	73.136	73.219	70.68	0.079	72.322 5	67.30	-0.814

作为与 k-NN方法的比较,本文还采用常规线性回归方法对森林平均蓄积进行了估计试验。森林蓄积量与 Landsat绿度指数的关系方程为 $y = a + b \times \ln(x)$, 其中 x是蓄积,y是 Landsat TM数据经缨帽变换得到的绿度指数。采用最小二乘法解算方程参数 a和 b,利用 $x = \exp[(y - a)/b]$,求算蓄积量 y的值。采用交叉评价方法,得到精度评价结果 (表 2)。虽然表 2平均误差的绝对值要普遍比表 1的相应值小,但表 2中的 $RMSE$ 都大于表 1中的相应值,说明 k-NN方法具有较低的相对均方根误差,比基于线性回归的估测方法精度高。

图 3是基于 k-NN方法,只采用具有林地属性的固定样地,每像元估测的单位面积蓄积量 ($m^3 \cdot hm^{-2}$)。黄色线条给出了县级行政单元的边界,不同的色阶代表了蓄积量的大小。

表 2 像元级蓄积量线性回归估测精度 (分树种类型单独计算的模型)

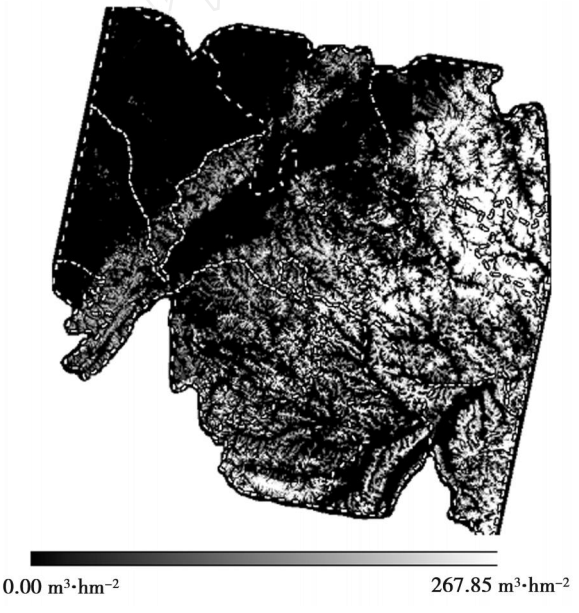
样地归并	NFI估计 / ($m^3 \cdot hm^{-2}$)	$x = \exp[(y - a)/b]$		
		平均值 / ($m^3 \cdot hm^{-2}$)	$RMSE$ '/%	平均误差 / ($m^3 \cdot hm^{-2}$)
混合	65.67	65.61	92.0	-0.056 25
落叶松	58.54	58.62	101.7	0.078 75
栎树类	78.81	78.45	83.9	-0.356 25
针叶林	53.75	53.76	105.0	0.030 00
阔叶林	73.14	72.96	91.9	-0.176 25

2.2 小面积统计单元蓄积估计

根据土地利用图将所有的地类划分为森林-非森林两个图层,提取落在森林图层的所有样地,假设有 N 个样地。基于这 N 个样地采用 k-NN方法(式 5),以县为统计单元估测森林总蓄积量;作为对比,还按照传统的成数估计方法计算了每个单元的森林总蓄积量(表 3)。

表 3　k-NN方法及常规统计方法对小面积统计单元平均蓄积和总蓄积的估计

单元编号	单元面积 / hm^2	单元内的森林样地数 / 单元内总样地数	k-NN估测结果		常规 NFI估测结果	
			平均蓄积 /($m^3 \cdot hm^{-2}$)	总蓄积 /(1 000 m^3)	平均蓄积 /($m^3 \cdot hm^{-2}$)	总蓄积 /(1 000 m^3)
1	280.867	7/135	36.51	255.494	68.016 7	990.559
2	346.811	7/163	47.11	2 008.648	54.881 4	1 985.086
3	279.901	56/137	71.87	10 012.310	87.410 4	10 000.810
4	452.789	45/211	44.33	5 079.207	47.954 4	4 630.783
5	348.923	51/158	67.15	12 922.300	69.903 3	7 873.002
6	64.495	19/31	91.10	4 018.883	103.691 0	4 098.822
7	230.505	34/110	51.42	4 583.385	60.490 7	4 309.781
8	259.435	31/124	47.63	4 413.927	59.676 3	3 870.530
9	32.011	2/16	41.25	346.640	45.108 3	180.495
10	155.383	12/73	51.79	2 649.446	32.098 6	819.876
11	140.555	24/64	60.82	4 140.450	54.694 4	2 882.839



(k = 10,利用了所有地类为林地的样地,多边形为县级行政单元)

图 3　基于 k-NN方法估测的像元尺度单位面积蓄积量分布图

图 2是试验区域的土地利用图,主要地类包括水体(黑色)、农田(浅灰色)、森林(深灰色)、城镇(白色)。从该图可以看出,单元 1只有很少的林地,林地面积大约和单元 9相当,单位面积蓄积不会相差特别大,单元 1的总蓄积量应该是所有单元中最低的,或者和单元 9不相上下。基于 k-NN计算得到的单元 1总蓄积量为 255 494.0 m^3,是 11个单元中最低的,和单元 9的估计值 346 640.0 m^3 相当;但根据常规统计方法得到的总蓄积量为 990 559.4 m^3,是单元 9估计值(180 495.2 m^3)的 5.5倍,这显然是不太合理的。统计单元 1、8、10的平均蓄积量与 NFI方法相比相对误差超过了 21%。统计单元 1、5、9、10和 11的总蓄积量与 NFI方法相比相对误差也超过了 21%。也就是说,单元 1和 10在平均蓄积和总蓄积上的相对误差都超过了 21%。假设单元 10与其左右相邻单元 7、11的平均蓄积量水平基本相同,则 k-NN 对单元 10 平均蓄积量的估计(51.79 $m^3 \cdot hm^{-2}$)更加合理。相比之下,NFI估计的平均蓄积仅为 32.098 6 $m^3 \cdot hm^{-2}$,和 7、11的平均蓄积量(60.490 7 $m^3 \cdot hm^{-2}$、54.694 4 $m^3 \cdot hm^{-2}$)相差较大。而且单元 10总蓄积量的 NFI估测结果为 819 875.6 m^3,也过于偏低。根据以上分析可以看出 k-NN对小面积单元的估测结果更加合理。但由于没有各小面积统计单元内的独立观测的加密样地数据,本文无法对小面积单元的估测精度给出客观的定量评价。Labrecque[8]利用加密样地,比较评价了几种小面积单元森林蓄积估计方法,其中就包括 k-NN和常规方法,得到的结论和这里观察结果是一致的:k-NN要优于常规的基于成数估计的方法。

常规的 NFI方法完全基于落在小面积统计单元内的固定样地,按照平均值法估测平均单位面积蓄积,按照成数比率法估算总蓄积,因此估测结果完全依赖一个统计单元内的样地。而通常情况下,县级行政单元或更小单元内固定样地的数量都是很有限的。本试验的统计单元属于县级水平,单元内的固

定样地数在 16～211块之间,其中属于林地的固定样地在 2～56块之间。这样少的样地必然会导致常规方法对小面积单元样地因子估测的较大误差。当然,我国森林资源清查样地的布设是为了省/区级尺度的调查而设的,不适用于县级以下的森林资源调查。但若有更多的空间信息数据,如调查区域的土地利用图、森林类型分布图、DEM、遥感数据等,就可以通过多源数据(空间数据加固地样地数据)的综合应用,实现小面积单元的参数估计。

3 结论

(1)根据森林资源一类清查样地和 Landsat数据,基于 k-NN方法可以在像元尺度给出整个样地覆盖区域某个森林观测因子平均值的估计,若对固定样地按类型分组,还可以分类型组(如针叶林、阔叶林)得到各观测因子平均值的估计。本试验表明,基于 k-NN的森林蓄积估测平均误差在 1.5 $m^3 \cdot hm^{-2}$ 之内,相对均方根误差(*RMSE*)低于传统的基于绿度指数的线性方程估测方法。

(2)基于森林资源一类清查固定样地、森林/非森林分布图和 Landsat数据,采用 k-NN方法可以实现小面积统计单元(县市级,单元总面积大小在 32～452.7 hm^2 之间)的森林总蓄积量的估测,估测精度优于只利用固定样地的传统成数估计方法。

参考文献:

[1] 赵宪文. 林业遥感定量估测[M]. 北京:中国林业出版社, 1997
[2] 游先祥. 森林资源调查、动态监测、信息管理系统的研究[M]. 北京:中国林业出版社, 1995
[3] 李崇贵. 用非线性理论研究以"3S"为基础的森林蓄积定量估测[D]. 北京:中国林业科学研究院, 2005
[4] Tomppo E. Satellite imagery-based national inventory of Finland[J]. International Archives of Photogrammetry and Remote Sensing, 1991, 28(7-1): 419 - 424
[5] Katila M, Tomppo E. Calibration of small-area estimates for map errors in multisource forest inventory[J]. Canadian Journal of Forest Research, 2000, 30: 1329 - 1339
[6] Katila M, Tomppo E. Selecting estimation parameters for Finnish Multisource National Forest Inventory[J]. Remote Sensing of Environment, 2001, 76: 16 - 32
[7] Katila M, Tomppo E. Stratification by ancillary data in the multisource forest inventories employing k-nearest-neighbour estimation[J]. Canadian Journal of Forest Research, 2002, 32: 1548 - 1561
[8] Labrecque S, Fournier R A, Luther J E, *et al*. A comparison of four methods to map biomass from Landsat-TM and inventory data in western Newfoundland[J]. Forest Ecology and Management, 2006, 226: 129 - 144
[9] Lappi J. Forest inventory of small areas combining the calibration estimator and a spatial model[J]. Canadian Journal of Forest Research, 2001, 31(9): 1551 - 1560
[10] Franco-Lopez H, Ek A R, Bauer M E. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method[J]. Remote Sensing of Environment, 2001, 77(3): 251 - 274
[11] Katila M. Empirical errors of small area estimates from the Multisource National Forest Inventory in eastern Finland[J]. Silva Fennica, 2006, 40(4): 729 - 742
[12] McRoberts R E, Nelson M D, Wendt D G. Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbours technique[J]. Remote Sensing of Environment, 2002, 82(2 - 3): 457 - 468