引用格式：戴舒,付迎春,赵耀龙.基于Cubist模型树的城市不透水面百分比遥感估算模型[J].地球信息科学学报,2016,18(10):1399-1409. [Dai S, Fu Y C, Zhao Y L. 2016. The remote sensing model for estimating urban impervious surface percentage based on the cubist model tree. Journal of Geo-information Science, 18(10):1399-1409.] **DOI**:10.3724/SP.J.1047.2016.01399

基于Cubist模型树的城市不透水面百分比遥感估算模型

戴 舒,付迎春*,赵耀龙

华南师范大学地理科学学院 广东省智慧国土工程技术研究中心，广州 510631

The Remote Sensing Model for Estimating Urban Impervious Surface Percentage Based on the Cubist Model Tree

DAI Shu, FU Yingchun* and ZHAO Yaolong

School of Geography, South China Normal University, Guangdong Provincial Center for Smart Land Research, Guangzhou 510631, China

Abstract: As a typical land-cover type, impervious surface is a key indicator of urban environmental quality and urbanization scope. In comparison with the traditional remote sensing image processing methods, the assessment of impervious surface percentage (ISP) can offer the sub-pixel level exploration and acquire the fine-scale information. In this paper, the proposed method uses the Cubist model tree with both the high-resolution (Google Earth) and the medium-resolution (Landsat TM/ETM+) remote sensing data to establish an estimation model of impervious surface percentage (ISP). A base model (Base Cubist-ISP) is built integrating all the original bands from Landsat TM excluding the thermal infrared band. This paper tries to minimize the effects of noise by adopting the ensemble learning algorithm and by incorporating the median of each solar-reflective band within the adjacent temporal images. After that, the following variables are filtered to get the optimized results, including the TM thermal infrared band, the derived variables from the original bands such as Texture, and the tasseled cap transformation variables. Then the variables are simplified, and in that way, the optimized parameter of ensemble learning algorithm for Cubist tree and the well-chosen variables are used to establish an optimization estimation model (Optimal Cubist-ISP). The results of a case study for Haizhu district, which is located in Guangzhou city of Guangdong Province, show that the overall root mean square error between the estimated ISP value, which is based on the Optimal Cubist-ISP model, and the reference ISP value is 12.98%, with a determinant coefficient of 0.90. Moreover, this paper compares the Base Cubist-ISP model with the Optimal Cubist-ISP model. The accuracy of the Optimal Cubist-ISP model is better than the Base Cubist-ISP model, and the RMSE decreases by about 5.03%. It is illustrated that the Base Cubist-ISP model may over-estimate the pervious surface area and under-estimate the high density impervious surface area, which could be improved by the model optimization. In addition, the Optimal Cubist-ISP model can not only be able to well recognize the land types of soil and water, but also eliminate the influence of shadow on the high density building area to a certain extent. Thus, the proposed approach on impervious surface estimation based on the Cubist model tree as well as its optimization scheme can be applied for precisely obtaining the ISP in the urban areas.

Key words: impervious surface; Cubist model tree; ensemble learning algorithm; Optimal Cubist-ISP model

***Corresponding author:** FU yingchun, E-mail: fuyc@scnu.edu.cn

---

收稿日期：2015-09-14；**修回日期**：2015-12-20.

基金项目：国家自然科学基金项目(41101152)；“973”计划前期研究专项(2014CB460614)；广州市产学研协同创新重大专项民生科技项目(156100021)；广东省科技计划项目(2015A010103013)。

作者简介：戴 舒(1990-)，男，江西抚州人，硕士生，主要从事城市信息处理与定量遥感研究。E-mail: daishusysu@foxmail.com

***通讯作者**：付迎春(1976-)，女，云南武定人，博士，教授，主要从事定量遥感与地理过程模拟研究。E-mail: fuyc@scnu.edu.cn

摘要：不透水面是城市区域中一种典型的土地覆盖类型，是衡量城市环境质量和城市化水平的重要标志之一。与传统基于像元级的遥感研究方法相比，不透水面百分比（Impervious Surface Percent，ISP）的估算可以进入像元内部，获得更准确的城市信息。本文应用Cubist模型树，对Landsat TM的原始波段变量（除热红外波段），建立ISP估算的基础模型（Base Cubist-ISP）。通过基于模型树的集成学习优化算法和加入相邻时相影像的波段变量中值，以削弱噪声的影响。然后，优选热红外波段和各种衍生变量，并进行属性精简，继而应用集成学习算法得到的参数和精简后的变量建立ISP估算的优化模型（Optimal Cubist-ISP）。对广东省广州市海珠区的实验结果表明，Optimal Cubist-ISP模型估算不透水面的整体均方根误差（RMSE）为12.98%，决定系数（R^2）为0.90，精度明显优于Base Cubist-ISP模型，RMSE降低约5.03%，ISP在透水面区域被高估和高密度不透水面区域被低估的现象得到改善。本文提出的基于Cubist模型树建立ISP遥感估算的模型及优化方法可以适用于城市区ISP的提取。

关键词：不透水面；Cubist模型树；集成学习算法；Optimal Cubist-ISP模型

1 引言

不透水面是城市区域中一种典型的土地覆盖类型，指难以被水穿透的地表，主要由人造地物构成，如屋顶、停车场、广场、公路、街道和人行道等[1]。作为一种典型的人工地貌特征，不透水面是衡量城市环境质量和城市化水平的重要标志之一[2]。与传统的基于像元级的遥感研究方法相比，不透水面百分比（Impervious Surfaces Percent，ISP）的估算可以进入像元内部，获得更准确的城市信息。而土地利用类型为植被的像元中，同样有可能分布少量不透水面，只有在亚像元尺度下量化估算不透水面百分比，才能更准确地分析城市化与城市生态的影响及两者的联系[3]。

近年来，基于亚像元的不透水面百分比估算方法一般包括多元回归法[4-5]、线性光谱混合分析法[6-7]、人工神经网络法[8-9]、决策树法[10-11]。Weng[3]认为多元回归法最大的不足在于易受季节影响，落叶季节低估植被覆盖度，生长季节高估植被覆盖度，进而影响不透水面百分比的估算；而线性光谱混合分析法问题在于不透水面的光谱容易与其他某些地物混淆，使得不透水面呈小块分散分布的区域的估测结果往往被高估，而不透水面集中分布的区域被低估。人工神经网络法与上述2种方法不同，具有解决非线性问题的能力，得到的结果也更精确。然而，人工神经网络法也存在一定的局限性[12]：在选择拓扑结构时常缺乏充分的理论依据，以及网络连接权值的物理意义不明确，因此人们通常难以理解其推理过程，可信度较差；此外，人工神经网络算法将信息处理都归结为数值运算，对知识的表达、存储和推理是隐式的，因此，存在依赖学习样本数量和质量的好坏，以及在学习上收敛速度慢、网络记忆不够稳定等方面的不足。在决策树法中，较为广泛使用的是分类回归树（CART）算法。CART算法继承了一般决策树具备的所有优点，既可以用于分类研究，又能进行连续变量的预测和回归，且实现简单，运算效率高，成为研究不透水面的热点。Cubist模型树也是一种决策树，相对于CART算法，Cubist模型树与其最大的区别在于模型树的叶子节点上是一个线性回归模型，而回归树的叶子节点上只是一个具体的值。因此，Cubist模型树比回归树建模更灵活，且精度更高[13]。2001年，美国地质调查局EROS数据中心选择使用Cubist模型树估算不透水面百分比和植被覆盖度，并将获得的数据加入到国家土地覆盖数据库中。此后，Cubist模型树越来越多的结合Landsat数据用于反演不透水面百分比，高志宏等[14]以山东省泰安市为例，基于TM影像除了热红外波段外的6个波段反射率变量运用Cubist软件进行城市不透水面百分比（ISP）遥感估算，基于ISP制图结果对城市土地利用变化进行检测。除Landsat原始波段以外，相关研究引入了更多的变量组合建模，如缨帽变换的分量组合、归一化差值植被指数（NDVI）、归一化建筑指数（NDBI）和建筑指数（BU）[15-17]，并针对山谷区域不透水面百分比提取增加了坡度变量[16]。Walton[18]用Landsat ETM+数据的7个波段变量和缨帽变换三分量共10个变量分别基于Cubist模型树、随机森林法和支持向量回归3种方法估算城市的植被覆盖度和不透水面百分比，且估算结果表明Cubist模型树估算不透水面百分比的效果最好。

Cubist模型树节点上的回归方程建模比较灵活，可以针对不同区域构建合适的应用模型。但Cubist模型树也是一种学习算法，在对不透水面信息进行估算时，对数据噪声敏感。当训练样本存在大量噪声时，会降低Cubist模型树的学习能力，估算精度也会相应降低。因此，原始波段和变量的选择在模型树建模中显得非常重要。对于异质性很强的城市地表不透水面百分比估算的应用，增加有用的变量将有助于提高估算精度，而删除冗余变量

则可以减少数据容量并提高计算效率。但上述研究没有讨论遥感影像波段及相关变量的适用性及在Cubist模型树中变量优选的方法，故本文拟在应用Landsat提取城市不透水面百分比的研究案例中，从波段和变量优选角度探讨Cubist模型树建立和优化的方法，以提高运算效率和估算精度。

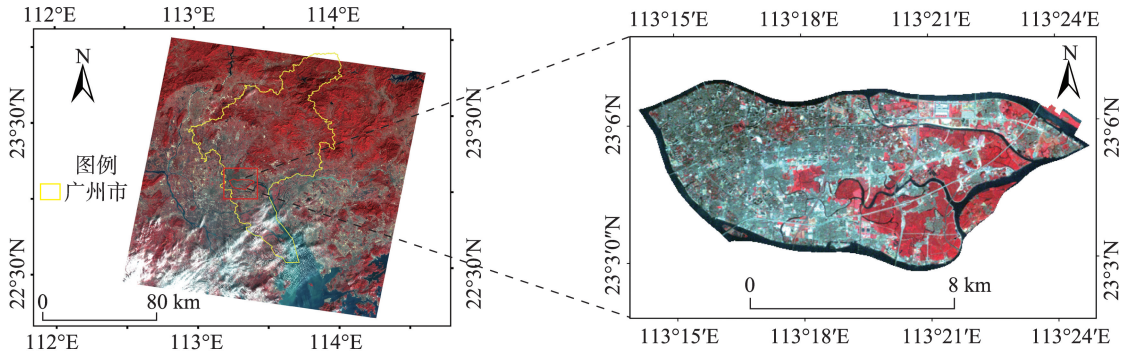
2　研究区与数据源

本文选取广东省广州市海珠区作为研究区。海珠区是广州市的老四区之一，位于广州市中部，珠江南面，由珠江水系广州河段前后航道所环绕，总面积92.11 km^2，是四面环水的天然良壤。研究区不透水面较集中，是城市化的典型区域，同时异质地表相嵌复杂多样，主要地物类型包括建筑物、道路、空地、湿地、林地、农田和河流等。

2.1　遥感影像数据

如图1(a)所示，研究区遥感影像目标数据选用2004年12月6日采集的Landsat TM图像，条带行列号为122/44，全景影像含有8%的云，黄色线所示区域为广州市，红色线所示区域即是本研究区。研究区(图1(b))图像质量较好，清晰无云。本文选取了研究区部分区域分辨率为1.1 m的Google Earth影像(图2)，采集时间为2005年1月6日，总面积10.8 km^2，用以获得ISP估算的训练样本和测试样本。对Google earth影像和TM影像进行了精确的几何配准，将Google earth影像配准到TM影像上，经投影和坐标转换后统一到UTM/WGS84投影坐标系。

同时，为了提高模型的估算精度，本文还选取了部分辅助数据，包括目标数据相邻的冬季多时相



(a) 研究区全景假色彩(TM4、3、2)合成影像图　　(b) 广州市海珠区假色彩(TM4、3、2)合成影像图

图1　研究区

Fig.1　The study area

图2　样本区Google Earth影像

Fig.2　Google Earth image of the sample area

Landsat影像和夏季多时相Landsat影像。其中，相邻的冬季多时相Landsat影像包括2004年11月20日的TM影像和2004年12月14日的ETM+影像；相邻的夏季多时相Landsat影像包括2004年6月13日、2004年6月29日、2004年10月19日和2005年7月18日这4个时相的TM影像。本文所有Landsat影像数据均来源于USGS(美国地质调查局)官方网站中的CDR(Climate Data Record)产品[19]。该产品为地表反射率产品，使用由美国航空航天局研发的陆地卫星生态系统干扰自适应处理系统(Landsat Ecosystem Disturbance Adaptive Processing System，LEDAPS)，经过几何校正、辐射定标、大气校正和Fmask云掩膜处理得到。LEDAPS大气校正基于6S辐射传输模型，获得有效的气溶胶分布数据，开展适用于TM/ETM+数据可见光、近红外及短波红外波段反射率的反演[20]。

2.2 ISP 训练数据和测试数据获取

应用面向对象分类法，将配准后的 Google Earth 影像分为不透水面（主要由建筑物、道路和水泥面空地等组成）、植被、耕地、裸地、水体和阴影。其中，为了防止样本区透水面所占比例过高，有2块植被区域（图2红线所示的区域）被排除在样本区之外。对分类后的结果进行排查、修改，以便精确提取最终的不透水面区域，并叠加到 TM 影像的像元网格，获得 30 m 尺度下 ISP 统计结果，即不透水面亚像元分布的参考值。考虑到很多阴影区的实际地物类型无法确定，含有阴影且无法确定实际地物类型的像元网格不参与 ISP 的统计。因此，在实验区随机选取 4000 个符合条件的样本点，其中 3200 个作为训练样本，剩下 800 个作为测试样本，训练样本和测试样本是相互独立的。

3 研究方法

本文基于 Cubist 模型树，通过基于模型树的集成学习算法[21]对 TM 影像原始波段变量（除热红外波段）进行降噪处理，增选有用变量并精简其相关属性对模型进行优化。在提供 Cubist 模型树优化评价方法的同时，建立城市不透水面百分比的估算模型。具体技术路线如图3所示。

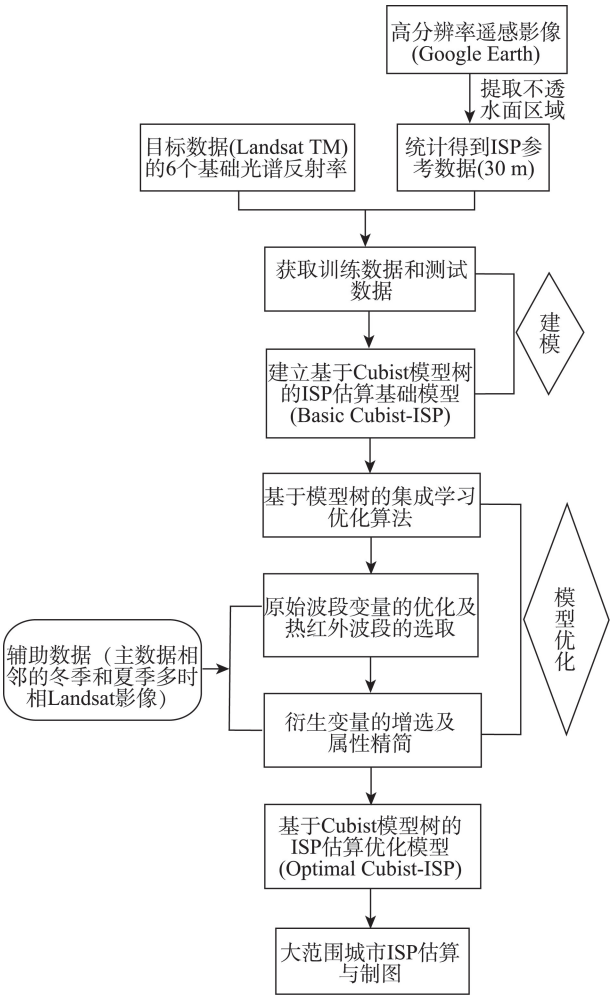


图3 技术路线图

Fig.3 Flow chart of the research

3.1 基于 Cubist 模型树的 ISP 估算模型

3.1.1 Cubist 模型树

Cubist 模型树是由 RuleQuest 公司开发的决策树算法，来源于 Quinlan 的 M5'模型树算法[13]。模型树是一种在叶子节点采用线性回归函数的决策树，表示一种分段式多元线性函数通过一系列的独立变量（称为属性）来预测一个变量的值。对给定的数据集，模型树将样本空间分为边缘相互平行的长方形区域（图4(a)），对每个分区确定一个相应的回归模型，使所建立的模型更为直观、清晰。如图4(b)所示，在模型树的每一层，选择最有识别力的属性作为子树的根节点，并以此将节点样本划分成若干子集。模型树持续划分，停止生长的条件有：(1)结点的样本数少于一定数量；(2)结点的样本目标属性标准差与总体样本目标属性标准差的比例小于某个限定值。模型树建立后，还需要对树进行剪枝，即是对某些子树进行归并后以叶子节点取代，以提高模型树的简洁性和效率。剪枝后，还需要使用平滑过程来补偿叶子节点的不连续性，平滑的方法需要考虑叶子结点的父结点，将子结点与父结点拟合方程合为一个新的线性方程。

模型树算法用一系列组合起来的分段线性模型，很好地解决了非线性问题。它与单纯的线性回归的区别在于，对输入空间的分割是由算法自动进行的，训练规则简单、有效，训练时间短，可以处理高维属性的问题。这种方法在估算不透水面上很好地结合了回归树和多元线性回归法，在预测连续值方面很成功。此外，该方法还根据样本信息将相关性不强的输入变量剔除，因此除了预测功能外，该方法还具有变量的重要性分析功能。本文采用R软件中的Cubist程序包。

3.1.2 基于 Cubist 模型树的 ISP 估算模型

3.1.2.1 ISP 估算基础模型（Basic Cubist-ISP）

首先，以目标数据除热红外波段以外的6个波

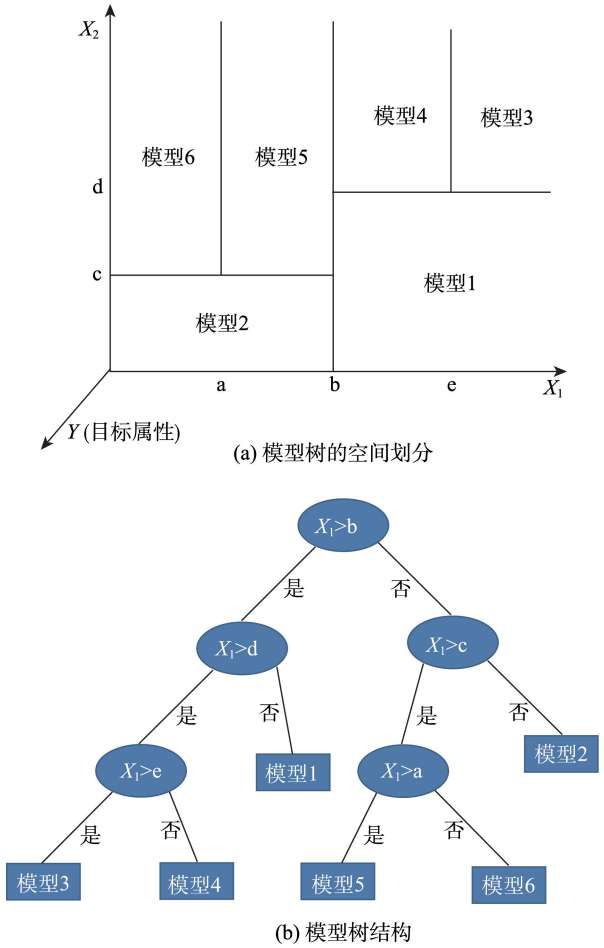


图4　Cubist模型树示意图

Fig.4　Illustrative diagrams of the Cubist model tree

段的光谱反射率作为估算模型的独立变量，将从高分辨率影像得到的30 m分辨率ISP参考数据作为目标变量。运用Cubist模型树对上述样本进行学习，建立对应于该时相遥感影像的ISP回归估算模型，并将此估算模型称为ISP估算基础模型，简称Basic Cubist-ISP。其中，模型树的结构可以表示成一系列if-then形式的决策规则：

Rule 1: [234 cases, mean 0.0072628, range 0 to 0.57196, est err 0.0114008]

if

BAND1 <= 0.092

BAND4 > 0.1659

then

ISP = − 0.0530795 + 2.7 BAND6 − 1.67 BAND5− 1.4 BAND3 + 1.6 BAND1+ 0.4 BAND2

Rule 2: [246 cases, mean 0.0208308, range 0 to 0.450773, est err 0.0301795]

if

BAND1 <= 0.1001

BAND4 > 0.2054

then

ISP = − 0.0091819 + 1.58 BAND6 − 1.6 BAND3 + 1.6 BAND2 − 0.4 BAND4- 0.29 BAND5

Rule 3: [152 cases, mean 0.0333124, range 0 to 0.810607, est err 0.0501680]

if

BAND6 <= 0.0588

then

ISP = − 0.1893327 + 0.84 BAND6 + 1.7 BAND2 - 0.6 BAND4 + 1.3 BAND1-0.22BAND5 + 0.2 BAND3

Rule ……

本文利用高分辨率影像提取的不透水面得到的ISP参考值来评估ISP估算模型的质量，即评价ISP参考值和ISP估算值的差异大小，所用的评价指标包括平均偏移误差（Mean-Bias-Error，MBE）、平均绝对误差（Mean-Absolute-Error，MAE）和均方根误差（Root-Mean-Square Error, RMSE）。各指标的计算公式如式（1）-（3）所示。

$$MBE=\frac{1}{N}\sum_{i=1}^{N}(\hat{f}_i-f_i) \tag{1}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|\hat{f}_i-f_i| \tag{2}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{f}_i-f_i)^2} \tag{3}$$

3.1.2.2　优化的ISP估算模型（Optimal Cubist-ISP）

（1）基于模型树的集成学习优化

Cubist模型选用了一种叫基于模型树的集成学习算法，该方法类似于boosting算法。当第一棵模型树遵循M5'模型树规则建立后，接下来的模型树是训练集结果的调整版本：如果模型高估某一目标值，那么下一个模型的响应为向下调整，以此类推。最终的估算值为每棵树的模型计算值的平均值。该优化方法可以降低模型树对数据噪声和训练样本误差的敏感性，提高估算精度。特别是在输入变量过多有大量属性冗余的情况下，基于模型树的集成学习算法可以提取有效信息，大大提高模型树的精度。而树的个数即参数committees的值需要设定，当committees的值足够大（≥10）时，估算结果及模型精度稳定，本文暂时设定committees为最大值100。表1为使用集成学习算法的优化方法前后的对比精度评价，从表中可知，使用集成学习算法的优化方法后，3个精度指标都有一定的提升。

表1 基于模型树的集成学习算法优化前后精度对比

Tab.1 The comparison of accuracy before and after the optimization through ensemble learning algorithm based on the model trees

committees	MBE/(%)	MAE/(%)	RMSE/(%)
1	−0.47	12.51	18.01
100	−0.25	12.19	17.50

(2)波段合成中值与热红外波段选择

Landsat影像经过辐射定标和大气校正后,由于阴影、云和雾霾等因素影响,仍然存在残余的噪声。为了减小噪声的影响,本文增加目标数据附近的冬季多时相Landsat影像。通常,多时相影像的合成取值可以最大限度地填补缺失数据和消除各种残余的噪声[22]。由于残余噪声的影响,影像数据中会出现过大或过小的异常值,显然,取多时相数据的均值是不合适的,而中值合成可以很好地解决这一问题[23]。利用目标数据和目标数据相邻的冬季影像共3个时相(2004年11月20日,2004年12月6日和2004年12月14日)的Landsat影像各波段反射率的中值代替原始波段反射率(表2实验1和2),RMSE降低了1.3%,模型估算精度明显提高。

Landsat TM影像数据包括6个波段反射率和1个热红外波段,相关研究在自变量中加了热红外波段[15-16,18]。根据表2的实验1和实验3对比证明,加入热红外波段数据可以提高估算精度,但提高得很少。研究表明,不透水面百分比与地表温度呈正相关关系[24]。热红外波段具有温度信息,有助于区分透水面与不透水面。本文尝试引入当年夏季(2004年6月29日)的TM影像的热红外波段,依据样本点的数据,把夏季和冬季的热红外波段分别与不透水面百分比做相关性分析。分析结果显示,夏季的相关系数为0.65,冬季的相关系数为0.41,表明夏季的热红外波段与不透水面百分比的相关度更高,能更好地区分透水面与不透水面。所以,本文用夏季的TM影像热红外波段代替冬季热红外波段。表2的实验1和实验4对比表明,加入夏季热红外波段可以有效地提高估算精度,RMSE也降低了约1.3%。遥感影像原始波段变量组合的实验表明,多时相中值和夏季热红外波段的加入有助于提高研究区不透水面百分比的估算精度。

表2 各种原始波段自变量的组合的MAE和RMSE

Tab.2 Estimation of MAE and RMSE using various combinations of the independent variables from the original bands

实验	MAE/(%)	RMSE/(%)	输入的自变量
1	12.19	17.50	b1-6
2	10.92	16.20	b1-6_med
3	12.26	17.45	b1-6,b7
4	11.26	16.21	b1-6,su_b7

注:b1-6为2005年12月6日的TM1~TM5和TM7;b1-6_med为目标数据附近的冬季多时相中值合成的TM1~TM5和TM7;b7为2005年12月6日的TM6波段;su_b7为夏季(2004年6月29日)TM影像的热红外波段

(3)衍生变量的增选及属性精简

在植被-不透水面-土壤模型中,不透水面主要与建筑用地有关,并区别于土壤和覆盖在上面的植被。由原始波段反射率变换得到的3类衍生变量,较好的突出了这3种地表覆盖组分的特征,分别是:①指数变量,即归一化差值植被指数(Normalized Difference Vegetation Index,NDVI)[25],归一化建筑指数(Normalized Difference Built-up Index,NDBI)[26]和归一化裸土指数(Normalized Difference Bareness Index,NDBaI)[27];②缨帽变换分量,能区分土壤地表平面与植被平面特征,包括绿度、亮度和湿度[28],跟植被和土壤有密切的关系;③纹理特征值,能重点突出地表成分空间结构关系。基于此,本文尝试加入指数变量、缨帽变换三分量和纹理特征值,实现多源信息辅助下ISP估算模型的建立。同时,对比各指标变量对ISP估算模型的影响(表3),得出以下2点结论:

①表3中实验1-3为研究NDVI相关变量的有效性。从实验1、2可以看出,加入NDVI后精度变化不明显,主要是因为NDVI容易受到物候的影响。因此,本文将单时相的NDVI换成目标数据相邻夏季多时相Landsat影像的NDVI最大合成值,即NDVI_max。从表3的实验1和实验3可知,加入NDVI_max后,精度有明显的提高。NDVI夏季最大值合成可以更好地区分植被和非植被区域,特别是某些只有在夏季覆盖植被的土壤,最大值合成便可以有效地区分土壤和不透水面。

②表3中实验4-9为测试每一个添加变量的有效性,设计了通过逐一去除变量测试变量影响的实验方案。其中,实验4为将所有考虑到的衍生变量均加入模型中,实验5-9分别为去除某一变量得到的结果,由表3可看出,去除了NDVI_max或Texture,精度均有明显的降低;而去除了TC、NDBI和NDBaI中的某一变量,精度差别不大。

表3　各种衍生自变量组合的精度评价指标MAE和RMSE

Tab.3　Estimation of MAE and RMSE using various combinations of the derived variables

实验	MAE/(%)	RMSE/(%)	输入的自变量(opt_b7;b1-6_med,su_b7)
1	10.41	15.39	opt_b7
2	10.34	15.42	opt_b7,NDVI
3	9.82	14.45	opt_b7,NDVI_max
4	8.77	12.67	opt_b7,NDVI_max,NDBI,NDBaI,TC,Texture
5	9.04	13.21	opt_b7,NDBI,NDBaI,TC,Texture
6	8.74	12.67	opt_b7,NDVI_max,NDBaI,TC,Texture
7	8.76	12.82	opt_b7,NDVI_max,NDBI,TC,Texture
8	8.84	12.72	opt_b7,NDVI_max,NDBI,NDBaI,Texture
9	9.38	13.83	opt_b7,NDVI_max,NDBI,NDBaI,TC

注：opt_b7为b1-6_med和su_b7合成的7个变量；NDVI：为目标数据的归一化植被指数；NDVI_max为2004年12月6日(TM)附近夏季NDVI最大合成值；NDBI为目标数据的归一化建筑指数；NDBaI为目标数据的归一化裸土指数；TC为缨帽变换三分量；Texture为TM图像主成分变换生成的第一主成分衍生出的8个纹理特征值，即均值(mean)、方差(var)、协同性(homog)、对比度(contr)、非相似度(diss)、熵(entropy)、角二阶矩(sec_mom)和相关度(correl)

因此，本文增加的衍生属性包括NDVI_max和Texture(8个纹理特征值)，既保留有用的变量、删除了冗余变量，又达到了预期的精度。

增选衍生变量后，还需要对现有的变量做属性精简。对6个波段中值进行最小噪声分离，最小噪声分离得到的前4个主分量波段涵盖了主要的信息，取前4个波段主分量，表示为MNF1-4_med。由表4可知，用MNF1-4_med代替b1-6_med，ISP的估算精度相差不大，所以，基于属性精简原则，本文用MNF1-4_med代替b1-6_med建模，并进行重要性分析。以Cubist模型树每个变量的重要度为指标，综合衡量以上增选后每个属性在各个Cubist模型树节点和多元线性回归模型中的贡献(表5)。在各变量的贡献中，最大是最小噪声分离变量的第二波段(79)，最小为纹理特征的熵变量(0)，本文通过实验选择重要度排在前面的10个属性，即MNF1-4_med、SU_TM6、NDVI_max和4个纹理特征值(均值、方差、对比度和非相似性)作为最终的自变量，建立估算不透水面百分比模型。上述可知，基于模型树的集成学习算法的参数committees暂时设为100，还需要设定优化参数。如图5所示，以10为间隔设置参数committees，选取优化的效果最佳，即RMSE最小时committees的值。

综上所述，本文以MNF1-4_med、SU_TM6、NDVI_max和4个纹理特征值(均值、方差、对比度和非相似性)作为自变量，设定参数committees=60，建立最终的估算不透水面百分比的优化模型，简称Optimal Cubist-ISP。

表4　属性精简前后精度评价

Tab.4　Evaluation of the accuracy before and after the simplification on attributes

实验	b1-6_med	MNF1-4_med	su_b7	NDVI_max	Texture	MAE(%)	RMSE(%)
1	√		√	√	√	8.84	12.86
2		√	√	√	√	8.94	13.09

注："√"表示被选择

表5　属性变量重要性分析

Tab.5　Importance analysis of the attribute variables

属性变量	MNF2_med	NDVI_max	MNF1_med	SU_TM6	var	MNF3_med	mean
重要度	79	64	59	55.5	47.5	47	46
属性变量	MNF4_med	diss	contr	homog	correl	sec_mom	entropy
重要度	38	30	28.5	19	6.5	1	0

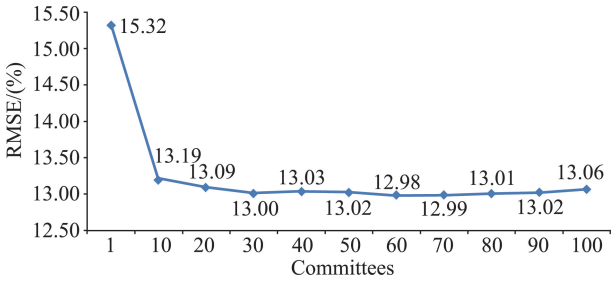


图5 参数Committees与RMSE关系图

Fig.5 The diagram showing the relationship between committees and RMSE

4 结果与讨论

4.1 ISP估算结果与Optimal Cubist-ISP模型评估

为了节省计算时间，本文基于改进的归一化差异水体指数（MNDWI）[29]，选用合适的阈值，既能够掩膜掉大部分水体，又避免误将城区的阴影掩膜。水体掩膜后，基于Optimal Cubist-ISP模型的海珠区的不透水面百分比估算结果如图6所示。统计可得，海珠区不透水面总面积39.92 km^2，占海珠区总面积43.34%。从图6可以看出，ISP估算结果图可以识别主要的公路，如广州大道南、广州环城高速和华南快速等。整体上，由于海珠区东部有大片的湿地和植被，东部的ISP平均水平要低于西部。

表6为模型的整体和分ISP高密度（70%~100%）、中密度（40%~70%）、低密度（10%~40%）和透水面（0%~10%）的精度评价，可以看出：(1)模型整体的RMSE为12.98%，MAE为8.88%，RMSE与MAE相差较大，主要是由于测试数据中异常值（离群值）的影响；(2)高密度不透水面和透水面精度较高，中、低密度不透水面精度较低；(3)高密度不透水面被低估，中、低密度不透水面和透水面被高估，其中

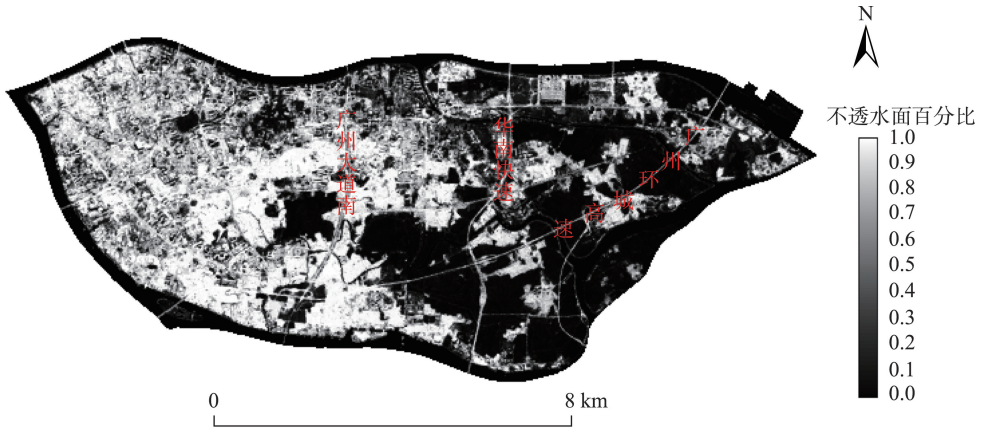


图6 海珠区ISP估算结果

Fig.6 Result of the ISP estimation over Haizhu district

表6 Optimal Cubist-ISP模型整体与分密度精度评价

Tab.6 Overall and hierarchy evaluation of the accuracy of Optimal Cubist-ISP model

评价指标	整体	高密度	中密度	低密度	透水面
MBE/(%)	−0.38	−5.76	1.84	6.42	5.91
MAE/(%)	8.88	7.46	13.24	14.83	6.31
RMSE/(%)	12.98	11.14	16.46	18.85	10.43

高密度不透水面和透水面的MBE的绝对值和MAE相差不大，表明大部分高密度不透水面都处于被低估的状态，而大部分透水面都处于被高估的状态。

对比表3中实验4和筛选有效的变量并进行属性精简后的建模时间，在参数committees相同（均设为100）的情况下，前者建模时间为38.70 s，后者的建模时间为21.80 s，在精度相差不大的情况下，建模效率大幅提升，提高了77.52%。

4.2 优化前后模型对比

如表7所示，相比基础模型（Base Cubist-ISP），优化后的模型MAE降低3.63%，RMSE降低5.03%。并且，RMSE在高、中、低密度和透水面均有降低，特别是透水面，降低7.06%，精度有大幅提升。

图7为基础模型（Base Cubist-ISP）与优化后的模型（Optimal Cubist-ISP）的ISP估算值与参考值散点图。对比散点图可以看出，优化后，异常点显著减少，特别是在透水面区域。R^2从0.80提高到0.90，斜率从0.79提高到0.87，ISP在透水面区域（图7中红圈所包含的点）被高估和高密度不透水面区域（图7中黄圈所包含的点）被低估均得到改善。

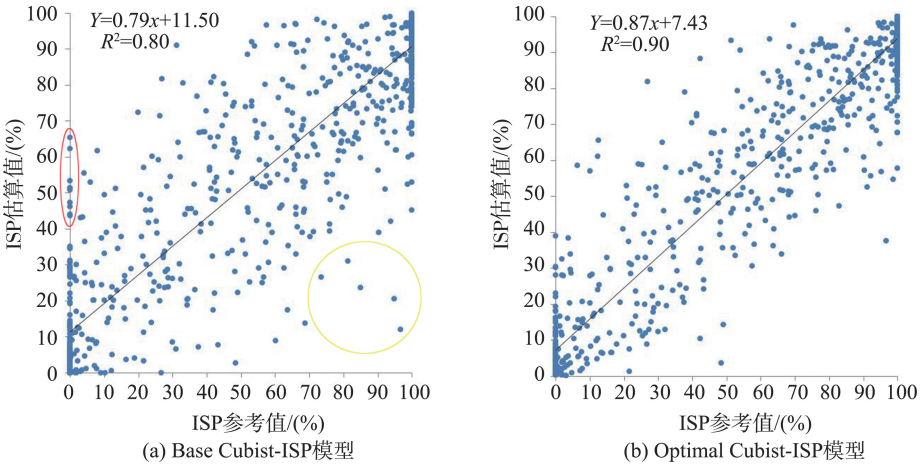
表7　模型优化前后整体与分密度评估精度对比

Tab.7　The comparison of overall and hierarchy evaluation accuracy before and after the model optimization

	MAE /(%)	RMSE/(%)				
		整体	高密度	中密度	低密度	透水面
Base Cubist-ISP	12.51	18.01	15.88	21.86	21.77	17.49
Optimal Cubist-ISP	8.88	12.98	11.14	16.46	18.85	10.43

如图8所示，选取了A、B、C、D、E、F 6个像元点。像元A、B、C分别为含有裸土的透水面像元、含有水的透水面像元和含有裸土和水的透水面像元，像元D、E、F分别为含有裸土的高密度不透水面像元、含有水的中密度不透水面像元和含有裸土和水的低密度不透水面像元。由于水体与城市里的阴影、透水的裸土或空地与不透水的人工地物存在光谱混淆，A、B、C、D、E、F的ISP值均被高估，优化后的模型能较好地识别裸土和水体，很好地改善了低值高估的现象（表8）。

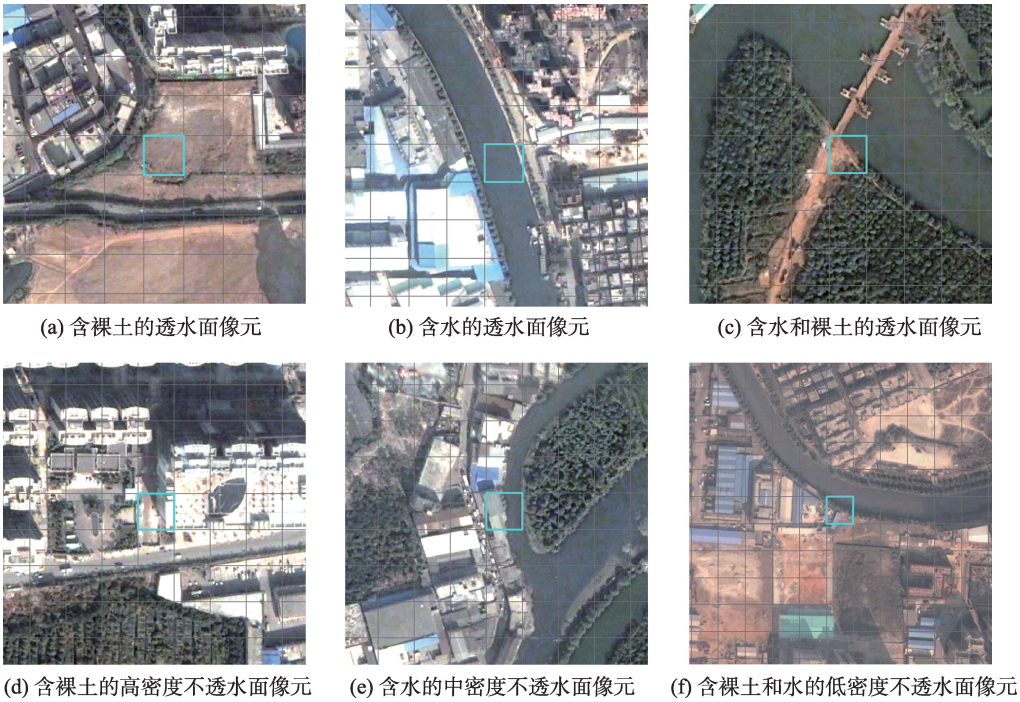
图9为优化前后海珠区ISP估算结果分级图。



(a) Base Cubist-ISP模型　　(b) Optimal Cubist-ISP模型

图7　ISP估算值与参考值散点图

Fig.7　Scatter plots between the estimated and reference ISP



(a) 含裸土的透水面像元　(b) 含水的透水面像元　(c) 含水和裸土的透水面像元

(d) 含裸土的高密度不透水面像元　(e) 含水的中密度不透水面像元　(f) 含裸土和水的低密度不透水面像元

图8　特殊像元点Google Earth影像图

Fig.8　Google Earth image of special pixels

表8 模型优化前后特殊像元点ISP估算值对比

Tab.8 The comparison of the ISP estimations of special pixels before and after the model optimization

像元点	ISP参考值/(%)	Base Cubist-ISP模型估算值/(%)	Optimal Cubist-ISP模型估算值/(%)
A	0.00	43.70	4.15
B	4.08	55.49	14.00
C	0.00	35.04	5.94
D	75.28	92.69	75.41
E	51.84	88.43	46.61
F	26.16	71.34	12.00

从图9可以看出,优化前本是大片高密度建筑区出现了很多破碎的高密度不透水面斑块,这主要是因为阴影的穿插使原本连续的高密度区域变得很破碎,而优化后在一定程度上消除了阴影对于高密度建筑区的影响。由优化前后的图中均可以看出,林地周围的植被较多,不透水面百分比降低,但却受不透水面地物的阻隔,斑块呈现小而破碎状。林地构成了大部分透水面,优化前后的ISP估算结果分级图中林地的分布差别不大。

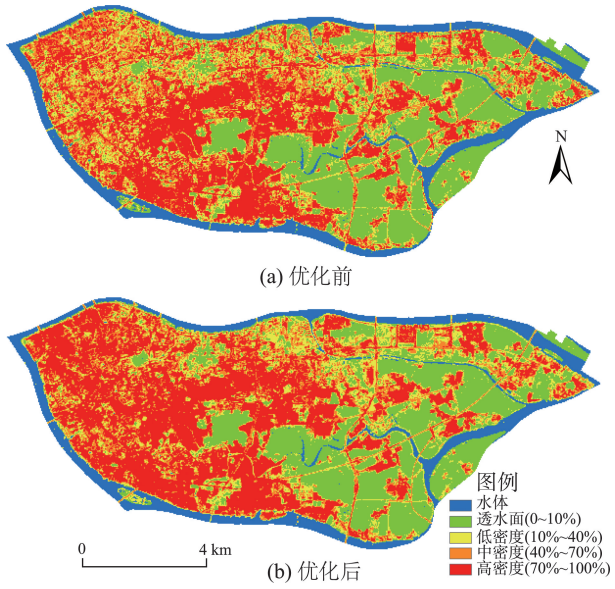


图9 模型优化前后海珠区ISP估算结果分级图

Fig.9 The classification map of the ISP estimation for Haizhu district before and after the model optimization

5 结论

本文基于Cubist模型树增加有用属性和删除冗余属性的重要性以及对数据噪声敏感这一现象,通过集成学习的优化算法、对TM原始波段变量通过辅助数据进行中值合成、加入有用的变量以及属性精简对基础模型进行优化,得到最终的自变量,建立优化的ISP估算模型(Optimal Cubist-ISP)。

(1)辅助的冬季数据可以一定程度上消除残余噪声,利用辅助的夏季数据可以得到夏季热红外波段和夏季NDVI最大值。增加2种数据变量均可以有效地提高模型的精度。

(2)Optimal Cubist-ISP模型整体的RMSE为12.98%,MAE为8.88%。其中,高密度不透水面和透水面精度较高,中、低密度不透水面精度较低;高密度不透水面被低估,中、低密度不透水面和透水面被高估。在Cubist模型树的集成学习算法的参数committees相同(均设为100)的情况下,筛选有效的变量并进行属性精简,建模效率大幅提升,提高了77.52%。

(3)基于模型树的集成学习的优化算法和中值合成等可以最大程度地消除噪声的影响,再通过加入有用的变量和属性精简后得到的Optimal Cubist-ISP模型精度明显优于基础模型(Base Cubist-ISP),均方根误差降低5.03%,R^2提高了0.10,异常点显著减少,ISP透水面区域被高估和高密度不透水面区域被低估的现象得到改善,适用于城市区ISP的提取。

(4)水体与城区里的阴影存在光谱混淆,透水的裸土或空地与不透水的人工地物(如停车场等)也存在光谱混淆,造成含有水体、裸土或空地的像元被高估。优化后的模型能较好地识别裸土和水体,很好地改善低值高估的现象。此外,优化后的模型可以一定程度上消除阴影对于高密度建筑区的影响,从而避免高密度建筑区斑块破碎化的现象。

参考文献(References):

[1] Yuan F, Bauer M E. Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery[J]. Remote Sensing of Environment, 2007,106(3):375-386.

[2] Arnold Jr C L, Gibbons C J. Impervious surface coverage: the emergence of a key environmental indicator[J]. Journal of the American Planning Association, 1996,62(2): 243-258.

[3] Weng Q. Remote sensing of impervious surfaces in the urban areas: requirements, methods, and trends[J]. Remote Sensing of Environment, 2012,117:34-49.

[4] Yang X J. Estimating landscape imperviousness index from satellite imagery[J]. IEEE Geoscience & Remote Sensing Letters, 2006,3(1):6-9.

[5] Bauer M E, Loffelholz B C, Wilson B. Estimating and mapping impervious surface area by regression analysis of Landsat imagery[A]. In: Weng Q. Remote sensing of impervious surfaces[M]. Boca Rotan, FL: CRC Press, 2008:39-58.

[6] Ridd M K. Exploring a V-I-S (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: comparative anatomy for cities[J]. International Journal of Remote Sensing, 1995,16(12):2165-2185.

[7] Wu C S, Murray A T. Estimating impervious surface distribution by spectral mixture analysis[J]. Remote Sensing of Environment, 2003,84(4):493-505.

[8] Flanagan M, Civco D L. Subpixel impervious surface mapping[C]. Proceedings of ASPRS Annual Convention, 2001.

[9] Hu X F, Weng Q. Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks[J]. Remote Sensing of Environment, 2009,113(10):2089-2102.

[10] Xian G, Crane M. Assessments of urban growth in the Tampa Bay watershed using remote sensing data[J]. Remote Sensing of Environment, 2005,97(2):203-215.

[11] 廖明生,江利明,林珲,等.基于CART集成学习的城市不透水层百分比遥感估算[J].武汉大学学报·信息科学版, 2007,32(12):1099-1102. [Liao M, Jiang L, Lin H, *et al*. Estimating urban impervious surface percent using boosting as a refinement of CART analysis[J]. Geomatics and Information Science of Wuhan University, 2007,32(12): 1099-1102.]

[12] 朱艾莉,吕成文.城市不透水面遥感提取方法研究进展[J].安徽师范大学学报:自然科学版,2010,33(5):485-489. [Zhu A, Lu C. Advances in the methods of extracting urban impervious surface based on remote sensing[J]. Journal of Anhui Normal University: Natural Science, 2010,33 (5):485-489.]

[13] Quinlan J R. Learning with continuous classes[C]. Proceedings of Australian Joint Conference on Artificial Intelligence World Scientific, 1992:343-348.

[14] 高志宏,张路,李新延,等.城市土地利用变化的不透水面覆盖度检测方法[J].遥感学报,2010,14(3):593-606. [Gao Z, Zhang L, Li X, *et al*. Detection and analysis of urban land use changes through multi-temporal impervious surface mapping[J]. Journal of Remote Sensing, 2010,14(3): 593-606.]

[15] Yang L, Huang C Q, Wylie B K, *et al*. An approach for mapping large-area impervious surfaces: synergistic use of Landsat-7 ETM+ and high spatial resolution imagery[J]. Canadian Journal of Remote Sensing, 2003,29(2):230-240.

[16] Xian G. Mapping impervious surfaces using classification and regression tree algorithm[A]. In: Weng Q. Remote sensing of impervious surfaces[M]. Boca Rotan, FL: CRC Press, 2008:39-58.

[17] Im J, Lu Z Y, Rhee J, *et al*. Impervious surface quantification using a synthesis of artificial immune networks and decision/regression trees from multi-sensor data[J]. Remote Sensing of Environment, 2012,117:102-113.

[18] Walton J T. Subpixel urban land cover estimation: comparing Cubist, random forests, and support vector regression[J]. Photogrammetric Engineering & Remote Sensing, 2008,74:1213-1222.

[19] USGS. Landsat higher level science data products[ER/OL]. http://landsat.usgs.gov/landsat_cdr_ecv.php, 2015-01-23/2015-11-06.

[20] Schmidt G, Jenkerson C, Masek J, *et al*. Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm description (No. 2013-1057)[R]. US Geological Survey, 2013.

[21] Quinlan J R. Combining instance-based and model-based learning[C]. Machine Learning Proceedings, 1993:236-243.

[22] Txomin H, Michael A W, Joanne C W, *et al*. An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites [J]. Remote Sensing of Environment, 2015,158:220-234.

[23] Neil F. Seasonal composite Landsat TM/ETM+ images using the medoid (a multi-dimensional median)[J]. Remote Sensing, 2013,5(12):6481-6500.

[24] Weng Q. A remote sensing-GIS evaluation of urban expansion and its impact on surface temperature in the Zhujiang Delta, China[J]. International Journal of Remote Sensing, 2001,22(10):1999-2014.

[25] Price J C. Using spatial context in satellite data to infer regional scale evapotranspiration[J]. IEEE Transactions on Geoscience & Remote Sensing, 1990,28(5):940-948.

[26] Zha Y, Gao J, Ni S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery[J]. International Journal of Remote Sensing, 2003, 24(3):583-594.

[27] Zhao H M, Chen X L. Use of normalized difference bareness index in quickly mapping bare areas from TM/ETM+ [J]. Geoscience & Remote Sensing Symposium Proceedings, 2005,3:1666-1668.

[28] Crist E P. A TM tasseled cap equivalent transformation for reflectance factor data[J]. Remote Sensing of Environment, 1985,17(85):301-306.

[29] 徐涵秋. 利用改进的归一化差异水体指数(MNDWI)提取水体信息的研究[J]. 遥感学报,2005,9(5):589-595. [Xu H Q. A study on iInformation extraction of water body with the modified normalized difference water Index (MNDWI) [J]. Journal of Remote Sensing,2005,9(5):589-595.]