

# 基于随机森林的元胞自动机城市扩展模拟 ——以佛山市为例

陈 凯<sup>1,2</sup>, 刘 凯<sup>1,2\*</sup>, 柳 林<sup>1,2</sup>, 朱远辉<sup>1,2</sup>

(1. 中山大学地理科学与规划学院, 综合地理信息研究中心, 广州 510275;  
2. 广东省城市化与地理环境空间模拟重点实验室, 广州 510275)

**摘 要:**本文提出一种基于随机森林的元胞自动机城市扩展(RF-CA)模型。通过在多个决策树的生成过程中分别对训练样本集和分裂节点的候选空间变量引入随机因素,提取城市扩展元胞自动机的转换规则。该模型便于并行构建,能在运算量没有显著增加的前提下提高预测的精度,对城市扩展中存在的随机因素有较强的容忍度。RF-CA模型可进行袋外误差估计,以快速获取模型参数;也可度量空间变量重要性,解释各空间变量在城市扩展中的作用。将该模型应用于佛山市1988-2012年的城市扩展模拟中,结果表明,与常用的逻辑回归模型相比,RF-CA模型进行模拟和预测分别能够提高1.7%和2.6%的精度,非常适用于复杂非线性特征的城市系统演变模型与扩展研究;通过对影响佛山市城市扩展的空间变量进行重要性度量,发现对佛山城市扩张模拟研究而言,距国道的距离与距城市中心的距离具有最重要的作用。

**关键词:**随机森林;元胞自动机;城市扩展;佛山

## 1 引言

元胞自动机(Cellular Automata, CA)由数学家Stanislaw M.Ulam与Von Neumann于1948年提出,最初用于模拟生命系统所特有的自复制现象,是一个描述自然界复杂现象的简化数学模型(段晓东等, 2012)。自1970年Tobler首次将CA应用于城市扩展模拟以来,已经有众多学者开展有关城市CA的研究,CA在模拟城市土地利用与土地覆盖变化方面的潜力受到持续关注(廖江福等, 2014)。例如,综合系统动力学模型和元胞自动机模型对中国北方农牧交错地区进行模拟(何春阳等, 2005),探索城市形态与能源消费的关系(Chen et al, 2013),对城市增长情景进行模拟(张鸿辉等, 2008),基于粮食安全约束和经济发展区域差异进行土地资源优化配置

(柯新利等, 2014),以及进行规划方案的模拟与评价(龙瀛, 2011)等。这些研究表明,CA能够很好地模拟城市用地的扩展,并可将模拟结果应用于生态、经济、农业、规划等众多领域。

CA模型的核心问题是定义元胞的转换规则(冯永玖等, 2011)。在每次循环迭代运算中,转换规则是由元胞的当前状态及其邻域状态确定下一时刻该元胞状态的动力学函数。CA模型是否成功,很大程度上在于转换规则的设计是否合理,能否真实地反映事物间发生变化的内在本质。为使CA能够模拟出城市扩展现象,众多学者从不同角度提出了确定转换规则的方法。一方面,以Wu为代表的学者开创性地提出使用空间统计方法获取CA模型的转换规则,例如采用逻辑回归模型确定CA的转换规则(Wu, 2002),这类方法运算方便、易于理解、

收稿日期:2014-12;修订日期:2015-04。

基金项目:国家高技术研究发展计划(863)项目(2012AA121402);国家自然科学基金项目(41001291);中央高校基本科研业务费专项资金项目(13lgpy61)。

作者简介:陈凯(1988-),男,湖南湘潭人,硕士研究生,主要从事土地利用模拟与GIS应用研究,E-mail: cksysu@foxmail.com。

通讯作者:刘凯(1979-),男,黑龙江伊春人,博士,副教授,主要从事环境遥感与GIS应用、湿地遥感研究,  
E-mail: liuk6@mail.sysu.edu.cn。

引用格式:陈凯, 刘凯, 柳林, 等. 2015. 基于随机森林的元胞自动机城市扩展模拟: 以佛山市为例[J]. 地理科学进展, 34(8): 937-946. [Chen K, Liu K, Liu L, et al. 2015. Urban expansion simulation by random-forest-based cellular automata: a case study of Foshan City[J]. Progress in Geography, 34(8): 937-946.]. DOI: 10.18306/dlkxjz.2015.08.001

简单实用,被众多学者用来获取CA的转换规则,获得了广泛的应用。但是逻辑回归方法要求空间变量间线性无关,而影响城市扩展的空间变量之间往往存在相关性,比如城市中心附近的元胞往往也邻近交通要道,传统的空间统计方法难以消除变量多重共线带来的不利影响,其模拟结果较难反映城市的真实形态(冯永玖等, 2010)。另一方面,以黎夏等为代表的学者对复杂的非线性城市系统引入人工智能和机器学习方法提取CA模型的转换规则,例如Li等(2001, 2002)提出使用神经网络训练的方法自动获取转换规则,Liu等(2008)提出使用蚁群智能算法模仿蚂蚁寻找食物的方式来构造转换规则,冯永玖等(2010)提出基于核主成分分析,通过核函数映射来消除空间变量的相关性,以反映城市发展的非线性过程。这些智能化建模方法能够较好地解决复杂城市系统的非线性分类问题,取得了不错的模拟结果,但仍存在一定的局限性,如收敛速度慢、对影响城市扩展的空间变量的作用不清晰等问题。本文探索了在保证精度较高的基础上,采用随机森林这样一种具有计算复杂度适中、有较好解释性、非线性等特点的方法来提取CA模型转换规则的参数,不仅能够通过模拟较好地反映城市的真实形态,并且可以度量空间变量在城市扩展中的重要性。

随机森林(Random Forest, RF)是由美国科学院院士Breiman(2001a)提出的一种利用多个决策树进行预测的组合算法。城市CA模型具有大量的数据需要处理,而构成随机森林的决策树不需要进行剪枝,可以在运算量没有显著增加的前提下提高预测的精度。由于在决策树生成过程中引入了随机性,不易出现过拟合现象,能够对城市扩展中存在的随机因素有较好的容忍度,是一种自然的非线性建模工具(Breiman, 2001b; 方匡南等, 2011)。随机森林还能够根据各变量对预测的贡献程度,对其进行重要性度量,从而解释各空间变量在城市扩展中的作用。由此可见,运用随机森林算法提取CA转换规则对城市扩展进行模拟,不仅能够获得较好的精度,还有助于我们了解城市演变的机制。

本文提出利用随机森林算法来提取城市扩展CA模型的转换规则,模拟佛山市1988-2000年的城市用地增长,并且根据其发展趋势以12年为一个阶段预测佛山市2012年和2024年的城市发展状况。

## 2 随机森林及RF-CA模型

### 2.1 随机森林的基本原理

随机森林是一种组合算法,它以决策树(Decision Tree)作为基分类器(Basic Classifier)。决策树是一种由节点和有向边组成的树状预测模型,在树的结构中包含有根节点、分支节点和叶子节点,其算法有ID3, C4.5, CART等多种形式,这些算法均采用自上而下的贪婪算法,每个内部节点选择分类效果最好的属性来分裂节点。Breiman等(1984)提出的随机森林使用的是决策树中的分类回归树(Classification and Regression Tree, CART)。

随机森林由一系列决策树模型 $\{h(\mathbf{X}, \Theta_k), k=1, 2, \dots\}$ 组成,其中 $h(\cdot)$ 表示决策树模型, $\mathbf{X}$ 是输入向量,参数集 $\{\Theta_k\}$ 是独立同分布的随机向量, $\Theta_k$ 决定单棵树的生长过程;采用简单多数投票法(针对分类)或单棵树输出结果的简单平均(针对回归)得到RF的最终输出。预测变量可以是数值型变量,也可以是类别型变量,无需对变量进行转换。具体过程如图1所示。

自随机森林算法提出以来,机器学习(Genuer et al, 2010)、生物医学(Kandaswamy et al, 2011)、生态(Peters et al, 2007)、遥感(Rodriguez-Galiano et al, 2012)等领域的学者进行了大量的理论和实证研究,证明了RF具有很高的预测准确率,对样本中存在的异常值和噪声具有很好的容忍度,并且不易出现过拟合现象。

随机森林分类算法的具体步骤(Peters et al, 2007)为:

- (1) 应用Bootstrap方法从原始训练集 $\mathbf{X}$ 中有放回地随机抽取一个自助样本集 $\mathbf{X}_i$ 。
- (2) 对于每个自助样本集 $\mathbf{X}_i$ ,用如下过程生成一棵不剪枝的决策树:设共有 $M$ 个原始变量,给定一个正整数 $mtry$ ,满足 $mtry \leq M$ 。在每个内部节点,

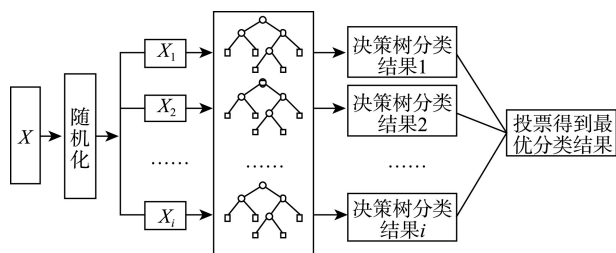


图1 随机森林算法示意图(方匡南等, 2011)

Fig. 1 Schematic diagram for random forest(Fang et al, 2011)

从  $M$  个原始变量中随机抽出  $mtry$  个预测变量作为该分裂节点的候选变量, 在  $mtry$  个候选变量中选出最好的分裂方式对该节点进行分裂。在生成整个森林的过程中,  $mtry$  不变。

(3) 重复(1)、(2), 直到生成  $ntree$  棵决策树( $ntree$  足够大)。

(4) 对未知类别的数据进行预测时, 输出的类别标签由  $ntree$  棵树的多数投票决定, 即:

$$H(x) = \arg \max_Y \left( \sum_{k=1}^{ntree} I(h(X, \Theta_k) = Y) \right) \quad (1)$$

式中:  $H(x)$  表示组合分类模型,  $h(\cdot)$  表示决策树模型,  $Y$  表示输出变量(或称为目标变量),  $I(\cdot)$  为示性函数,  $\arg \max_Y$  表示使得  $\sum_{k=1}^{ntree} I(h(X, \Theta_k) = Y)$  取最大值时  $Y$  的取值。

对于训练集中每一个变量  $x_i$ , 组成  $ntree$  棵决策树中的每一棵树都会对其标定一个唯一的类别。因此, 每一个变量  $x_i$  将会被分类  $ntree$  次, 可将某个类别  $c_j$  得到的投票数量  $N_{c_j}$  占有所有投票数量  $ntree$  的比例解释为: 该变量的类别是  $c_j$  的概率(Peters et al, 2007), 即:

$$P(c_j) = \frac{N_{c_j}}{ntree} \quad (2)$$

式中:  $P(c_j)$  为变量  $x_i$  被分到类别  $c_j$  的概率,  $N_{c_j}$  为变量  $x_i$  被投票为类别  $c_j$  的数量,  $ntree$  为随机森林中决策树的总个数。

随机森林在 Bagging 的基础上引入随机选择属性, 更大程度降低了树之间的相关性, 同时建立的单棵不剪枝的决策树能够保证较低的偏差, 从而保证了随机森林的分类性能(王云飞等, 2013)。由于在决策树的构建过程中不需要进行剪枝处理, 因此相对其他机器学习模型, 随机森林的生成速度并不会显著增加。在生成随机森林过程中, 决策树的数量  $ntree$  以及用于分裂节点的候选变量的个数  $mtry$  是由用户自定义的参数。

由于自助样本集  $X_i$  是使用 Bootstrap 方法有放回地随机抽取得到的, 因此原始训练集  $X$  中大约有 36.8% 的样本不会出现在自助样本集  $X_i$  中(方匡南等, 2011), 这些数据就称为这个自助样本集的袋外(Out-of-Bag, OOB)数据。这部分数据可以求取 OOB 误差以估计随机森林的性能, 也可以用来评价变量的重要性以进行变量选取。

OOB 误差可通过下述方法得到:

(1) 在每个 Bootstrap 自助样本集构建好决策树

后, 用其预测该自助样本集的 OOB 数据。

(2) 对原始训练集中的每个样本, 合计所有上述 OOB 预测的结果, 计算得到的错分比例即为 OOB 误差。

Breiman(2001a)通过实验证明, 上述的 OOB 估计是无偏估计, 因此使用 OOB 估计可以取得和  $N$  折交叉验证同样的效果。用 OOB 数据衡量变量重要性的常用方法有两种, 分别为平均精度减少衡量和平均基尼减少衡量(Hastie et al, 2008), 本文采用 Breiman(2001a)使用的平均精度减少衡量。平均精度减少衡量把一个变量的取值变为随机数, 在其他变量不变的情况下, 通过分析随机森林预测准确性的降低程度得到, 该值越大表示该变量越重要。

## 2.2 基于随机森林的元胞自动机模型

### 2.2.1 利用随机森林获取元胞开发适宜性

在 CA 城市扩展模拟中, 使用开发适宜性  $P_{ij}$  来衡量各空间变量影响下的元胞转变为城市用地的可能性。本文使用随机森林方法获取元胞自动机的开发适宜性  $P_{ij}$ , 构建随机森林-元胞自动机模型(RF-CA 模型)。随机森林在分类时根据多个决策树的投票结果来决定预测类别, 由于在训练生成随机森林的过程中, 对原始训练集  $X$  和原始空间变量都引入了随机性, 因此很多情况下各棵决策树的分类结果并不完全一致, 对于同一元胞, 可能有部分决策树投向转变为城市用地, 另一部分投向不转变为城市用地。本文中, 一个元胞的开发适宜性是随机森林中所有决策树的有关该元胞发展为城市用地的平均预测概率, 计算方法如下:

$$P_{ij} = \frac{N_{ij}}{ntree} \quad (3)$$

式中:  $P_{ij}$  是  $ij$  位置元胞的转变城市用地的开发适宜性,  $N_{ij}$  为在所有决策树中, 将该元胞分类为发展成城市用地的决策树数量,  $ntree$  为随机森林中决策树的总个数, 本文通过使 OOB 误差最小化来确定该值的大小。

### 2.2.2 城市扩展 CA 模型

CA 模型中转换规则的提取是整个模型的核心任务, 模型的转换规则由 4 个部分组成: 开发适宜性部分、邻域函数部分、随机因子部分和约束条件部分。

(1) 开发适宜性部分。在城市 CA 模型中, 非城市用地转换为城市用地都是基于元胞的开发适宜性得到的。开发适宜性是衡量非城市用地元胞在多个空间变量, 包括区位因素、交通条件、自然环



境、社会经济因素等的共同作用下发展为城市用地元胞的适宜程度。目前开发适宜性的获取主要通过统计学习与数据挖掘方法获取,本文的开发适宜性通过随机森林方法获得。

(2) 邻域函数部分。CA模型中元胞之间的相互作用是局部的,体现在一个元胞下一个时刻的状态由其周围的邻域共同决定。邻域函数的计算如下式(张亦汉等, 2013):

$$\Omega_{ij}^t = \frac{\sum_{3 \times 3} \text{con}(S_{ij}^t = \text{urban})}{3 \times 3 - 1} \quad (4)$$

式中:  $\Omega_{ij}^t$  表示  $t$  时刻  $ij$  位置元胞的  $3 \times 3$  邻域作用值;  $\text{con}(\cdot)$  为条件函数;  $S_{ij}^t$  为该元胞的当前状态,如果元胞为城市元胞,则值为1,否则为0。

(3) 随机因子部分。城市的扩展除了受到纳入开发适宜性的各个确定性因素影响之外,还会受到政策调整、经济环境、自然灾害等随机因素的作用,为了体现这些不确定的随机因素,本模型引入随机因子(Li et al, 2001):

$$RA = 1 + (-\ln \gamma)^\alpha \quad (5)$$

式中:  $RA$  为随机因子部分,  $\gamma$  为值在  $(0,1)$  范围内的随机数,  $\alpha$  为控制随机变量影响大小的参数(取值范围为1~10的整数)。

(4) 约束条件部分。城市扩展模拟必须考虑客观的空间约束条件,如水体、山地、公园、优质农田等限制发展单元。元胞发展为城市用地的空间约束条件可由下式表达:

$$\text{con}(S_{ij}^t) = \begin{cases} 0 & \dots \text{该元胞禁止发展为城市用地} \\ 1 & \dots \text{该元胞可以发展为城市用地} \end{cases} \quad (6)$$

式中:  $\text{con}(S_{ij}^t)$  判断  $t$  时刻  $ij$  位置元胞是否可发展为城市用地,  $S_{ij}^t$  为该元胞的当前状态值,由用户预先确定的不可发展图层的属性决定元胞是否限制发展,在限制性开发区域  $S_{ij}^t$  值为0。

综合考虑以上4个部分,在CA模型中,  $t+1$  时刻  $ij$  位置元胞转变为城市用地的城市发展概率  $P_{ij}^{t+1}$  为:

$$P_{ij}^{t+1} = RA \times P_{ij}^t \times \text{con}(S_{ij}^t) \times \Omega_{ij}^t = (1 + (-\ln \gamma)^\alpha) \times P_{ij}^t \times \text{con}(S_{ij}^t) \times \Omega_{ij}^t \quad (7)$$

元胞是否发生土地利用类型转变则由下面的条件决定:

$$\begin{cases} P_{ij}^t \geq P_{\text{threshold}} \dots \text{转变为城市用地} \\ P_{ij}^t < P_{\text{threshold}} \dots \text{不转变为城市用地} \end{cases} \quad (8)$$

式中:  $P_{\text{threshold}}$  为土地利用转变的阈值,由用户根据其

不同取值时的模拟精度确定其大小。式(8)表示,如果  $P_{ij}^t$  的值大于或等于用户自定义的土地利用转变阈值  $P_{\text{threshold}}$ ,且  $ij$  位置元胞的土地利用类型为非城市用地,则该元胞的土地利用类型转变为城市用地,否则不发生转变。

运用随机森林确定CA的转换规则,模拟城市扩展的流程如图2所示。

### 3 模型应用及结果

#### 3.1 研究区概况

本文选取珠江三角洲的广东省佛山市作为研究区,模拟1988-2012年的城市扩张,以检验模型的效果,并根据其发展趋势预测2024年的城市扩展状况。佛山市位于广东省中南部,与广州市共同组成繁荣的广佛都市圈,是珠江三角洲的核心城市之一。过去30年是佛山市的城市快速扩张时期,土地利用不断发生变化,城市建成区范围持续扩大,以这样一个快速城市化区域作为RF-CA模型的试验区,能够较好地检验该模型的有效性。

#### 3.2 空间变量及数据获取

城市CA模拟假设城市发展概率受到一系列空间距离变量、邻近范围已城市化元胞数、元胞本身的状态等因素的影响(杨青生, 2008)。结合研究区的特点,本文选取以下空间变量,这些变量及获取方法见表1。

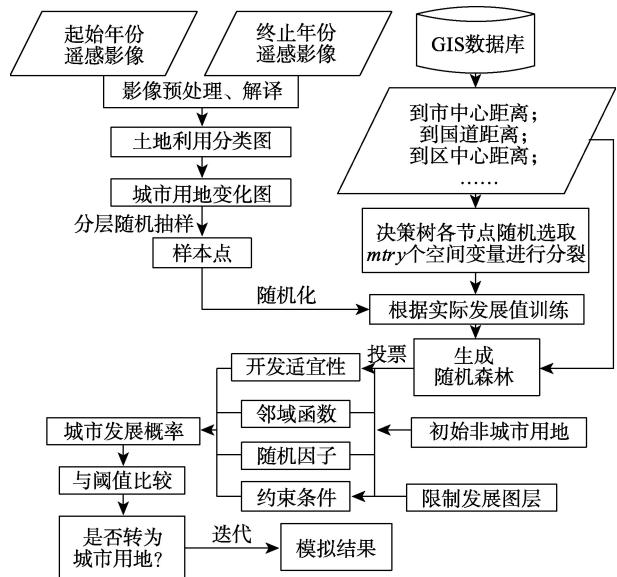


图2 基于随机森林CA模型的城市扩展模拟

Fig.2 Random forest-based cellular automata for simulating urban expansion

本文利用佛山市1988年Landsat 5 TM影像,2000年Landsat 7 ETM+影像和2012年HJ-1A影像作为数据源,使用ENVI 5.1对各期影像进行几何校正、影像拼接与裁剪等预处理,通过eCognition 9.0对预处理好的影像进行多尺度分割、目视解译,得到各年份30 m分辨率的佛山市土地利用分类图,通过有经验解译人员的交叉验证与历史地图对比,解译的精度高于90%。考虑到在对遥感影像的目视解译中难以区分城市绿地、园地、林地,因此本文的城市用地特指建设用地。各空间距离变量可使用ArcGIS 10.1中的Distance功能进行计算,模拟时需要动态计算中心元胞邻域的变化,ArcGIS 10.1的Focal函数可以很容易地计算中心元胞邻域中已城市化的元胞数目(刘小平等,2007)。

3.3 模型训练与参数校准

为了构建RF-CA模型,首先需要用历史数据对随机森林模型进行训练。本文选取1988、2000年两期土地利用数据计算得到的城市用地转化图作为模型的训练数据源,将该时期转变为城市用地的元胞编码为1,没有发生转变的其他元胞编码为0,作为该时期的城市发展值。运用随机分层抽样的方法,从转变为城市用地的元胞和可以转变为城市用地而尚未转变的元胞中分别抽取20000个样本点,获取这些样本点的空间坐标,运用ArcGIS 10.1的Sample功能读取这些样本点对应的城市发展和空间变量值,得到原始训练集X。

本模型使用Python为编程语言,结合国际上流行的机器学习开源工具包Scikit-learn,使用原始训练集X训练生成随机森林模型。

在对随机森林进行训练的过程中,涉及到2个用户自定义的关键参数:*ntree*和*mtry*。其中*ntree*为决策树的数量,即使用Bootstrap重抽样的次数;*mtry*为预测变量的数量,即决策树分裂节点的候选空间变量个数。在随机森林分类过程中,由于选取样本的方法含有放回随机抽样,因此,建立回归树时约有36.8%的样本数据不会被选中,而作为检验样本出现,起到样本内部交叉验证的作用。本文通过OOB无偏估计,得到随机森林在不同参数设置情况下的精度,以进行参数设置。

随机森林将多棵决策树集成在一起,在每棵决策树的生成过程中,对每个节点都会从M个原始变量中随机抽取*mtry*个预测变量,从这*mtry*个预测变量中选择最具分类能力的变量进行节点分裂。参数*mtry*不同,最后预测精度也会随之发生改变。经过测试发现决策树数量越大,预测精度相应也会趋于增大。为了确定参数*mtry*,在决策树的数量较大的情况下(*ntree*=1000),测试参数*mtry*取不同数值时的精度。如图3所示,随着预测变量数的增加,分类精度总体呈现先增加后减少的趋势。当预测变量个数*mtry*=1时分类精度为89.48%,随着*mtry*的增加,精度不断提升,当*mtry*=4时,精度达到最高,为89.93%,因此在本文中*mtry*的参数设置为4。

随机森林对每个Bootstrap样本分别进行决策树建模,组合多棵决策树的预测,最后通过投票得到最终的预测结果,决策树的数量设置也会对预测的精度造成影响。如图4所示,当预测变量的数量*mtry*=4时,通过比较不同数量的决策树下的预测精

表1 空间变量及获取方法		
Tab.1 Spatial variables and acquisition methods		
变量类型	变量	获取方法
因变量	是否转变为城市元胞	遥感分类
空间距离变量	距市中心的距离( $x_1$ )	利用ArcGIS的Distance获取
	距区中心的距离( $x_2$ )	
	距镇中心的距离( $x_3$ )	
	距高速公路的距离( $x_4$ )	
	距省道的距离( $x_5$ )	
	距县道的距离( $x_6$ )	
自然因素变量	高程栅格图( $x_7$ )	
	坡度栅格图( $x_8$ )	
限制因素变量	不可发展区域图( $x_9$ )	
局部变量	3×3邻域已城市化元胞数( $x_{10}$ )	用ArcGIS的Focal函数

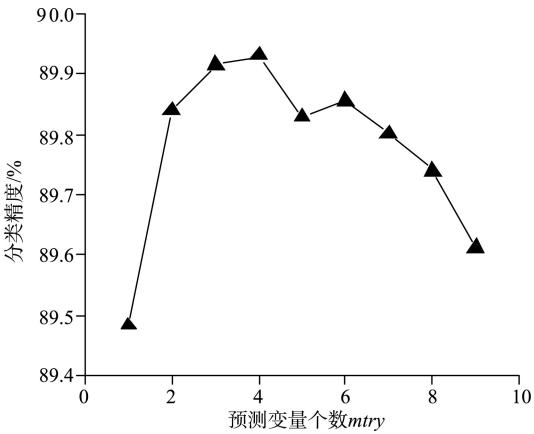


图3 分类精度与预测变量个数之间的关系  
(树的数量为1000)

Fig.3 Relationship between classification accuracy and the number of predictive variables with 1000 trees

度变化情况,可知随着树的数量增加,预测精度不断增加。当树的棵数  $ntree=1$  时,精度仅有 63.7%;  $ntree=2$  时,精度迅速提升到 70.7%;  $ntree=1000$  时,精度达到 89.90%,之后精度仍然会有小幅增加并不断趋近于 90%,但精度增加的幅度越来越小。综合考虑计算机运算性能以及精度要求,本文选取  $ntree=1000$  作为树的数量的设置参数。

在对随机森林进行训练后,可以使用袋外数据在生成好的随机森林中进行变量重要性度量。如图 5 所示,可以看到距国道和城市中心的距离两个因素对非城市用地转变为城市用地最为重要,其中距国道的距离对预测精度的影响最大,这是因为国道相对其他道路来说通达性好,位置优越,进入车道的路口较多;其次是距城市中心的距离,这是由于城市中心市政设施良好,政府机关较多,卫生教育机构齐全,购物方便,吸引较多住户和商户,从而

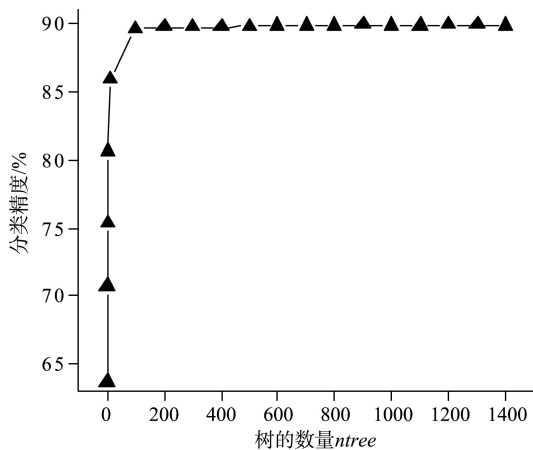


图 4 分类精度与树的数量之间的关系(预测变量个数为 4)

Fig. 4 Relationship between classification accuracy and the number of trees with 4 predictive variables

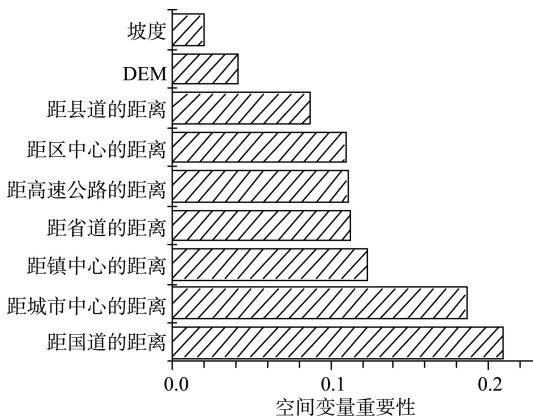


图 5 空间变量重要性度量

Fig. 5 Variable importance plots for classification random forest

推动接近城市中心的非城市用地转变为城市用地。距镇中心距离、距省道距离、距高速公路距离、距区中心距离、距县道距离等变量也具有较重要影响。逻辑回归模型获取的参数也能够用于评价各空间变量对城市扩展的贡献大小(杨青生等, 2007)。但是,由于逻辑回归模型要求空间变量间线性无关,同时空间变量间往往难以满足线性无关的条件,因此可能会得到与实际情况不太符合的结果,例如本文使用逻辑回归 CA 模型模拟佛山市城市扩展的过程中,得到距区中心的距离与城市扩展概率呈正相关的结果。相比较而言,随机森林模型是一种自然的非线性建模工具,因此能够更加有效地评价空间变量的重要性。

根据随机森林的算法流程,可知随机森林需要构建大量的决策树。传统的随机森林算法通常采用串行的构建方法,即一个决策树构建好之后,才开始构建第二个决策树,依此类推。串行构建决策树比较消耗时间,特别是当随机森林中需要构建的决策树足够多时。由于随机森林算法中用于构建每棵决策树的自助样本集  $X$  都是独立通过 Bagging 抽取的,同时每棵决策树的生长也是相互独立的,它们都是用自己的随机特征子空间进行分裂,因此随机森林很适于并行化构建(蔡林霖, 2013)。为提高训练的速度,充分利用多核 CPU 的运算性能,本文在构建随机森林模型的决策树时,分发给 CPU 多颗核心并行独立运行。

### 3.4 城市扩张动态模拟与预测

使用 RF-CA 模型进行城市扩展模拟时,首先使用生成好的随机森林计算每个元胞在空间变量作用下的发展适宜性,在此基础上计算受邻域元胞、随机变量、约束条件共同影响下初始状态非城市元胞的城市发展概率。模拟阶段的初始状态从 2000 年遥感影像分类图中获取。模拟过程中,邻近范围已转变为城市用地的元胞数在每次迭代过程中动态计算。

约束条件的设置可从初始的土地利用类型、地形图、政府规划等资料中获取,例如河流、水库等水体以及城市绿地的发展概率非常小,禁止发展为城市用地,约束值可设为 0;政府规划的开发区发展概率很大,相应约束值可设为 1。

在模拟的过程之中,需要设定城市发展的阈值  $P_{\text{threshold}}$ 。阈值  $P_{\text{threshold}}$  设置过大,则模拟得到的城市形态可能会过于集中;反之可能会过于分散。因此阈值的设置过大或过小都会影响到模拟的精度和效



果。本文通过1988-2000年提取的转换规则,验证不同阈值对应的1988-2012年模拟精度大小,确定模型的城市发展阈值 $P_{\text{threshold}}=0.8$ ,如表2所示。

城市系统是一个复杂的系统,可能会受到政策、经济等各方面的外界影响,因此应该适当引入随机性,在计算转变概率时,随机因子参数 $\alpha$ 取较小值2,以便尽量展现RF-CA模型在外界随机因素影响较小时的模拟结果。

根据随机森林所挖掘到的转换规则,模拟了2000-2012年期间佛山市城市空间演变情况。城市空间演变的模拟是通过Python语言,使用ArcGIS的ArcPy工具包以及机器学习开源工具包Scikit-learn实现的。

3.5 精度检验与对比分析

使用CA模型模拟城市扩展时,需要使用实际数据检查其有效性。一种简便的检验方法是用模拟结果与对应年度的实际城市用地进行对比(刘小平等,2007)。图6是使用随机森林CA模型得到的佛山市模拟结果和实际情况的对比图,其中图6a为1988年的初始状态,图6b、图6c分别为2000年的模拟城市扩展状况和实际城市扩展状况,图6d、图6e分别为2012年的模拟城市扩展状况和实际城市扩展状况,图6f则为根据其发展趋势以2012年的城市用地为起始年份预测佛山市2024年的城市发展状况。从2000年和2012年的对比图可以发现,模拟结果中城市用地的整体空间分布与真实城市用地相当接近,而预测得到的2024年城市发展状况表明,未来佛山市的城市用地格局将更加紧凑,在佛山各区中,南海区、顺德区将会有更多的非城市用地转变为城市用地。

为了定量评价模拟结果,本文采用全样本方式对随机森林模型的模拟结果进行精度评价,并与逻辑回归模型模拟结果比较,进行精度检验。全样本方式将整个佛山市全部城市用地与非城市用地元胞进行逐点对比,与部分已发表论文使用抽样的方法进行精度评价不同(使用抽样的方法会随着抽样样本的不同得到不同的精度)。逐点对比评价将从遥感图像获取的实际城市用地与对应时段的模拟城市用地进行叠加分析,得到模拟精度评价的混淆矩阵。将遥感图像分类得到的实际城市用地与随机森林模型、逻辑回归模型模拟的结果进行比较,得到的混淆矩阵如表3所示。以1988年为起始年份,由1988-2000年的实际用地总量为控制,经过迭代运算,得到的2000年模拟结果表明:随机森林模

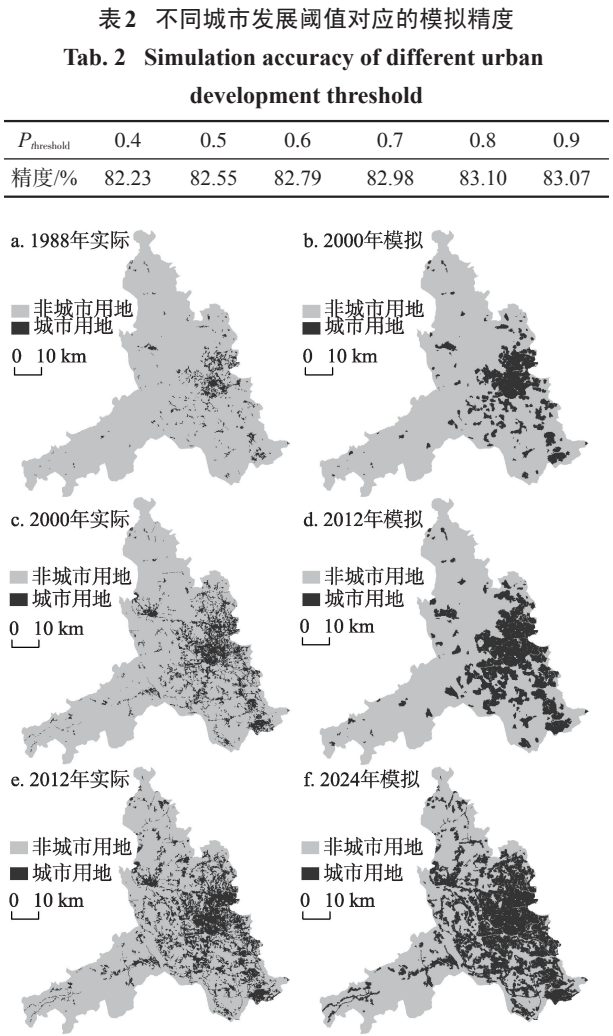


图6 佛山市城市用地模拟结果与实际情况对比图

Fig. 6 Simulation results of urban and non-urban land use from 1988 to 2024 based on the RF-CA model

型的总精度为89.7%,Kappa系数为0.580;逻辑回归模型的总精度为88.0%,Kappa系数为0.508。模拟结果显示,基于随机森林的CA模型比基于逻辑回归的CA模型有更好的精度,特别是Kappa系数方面的优势更加明显,说明随机森林CA模型的模拟结果与真实城市的一致性更好。CA城市扩展模型可以用于预测,本文根据1988-2000年提取转换规则,以1988年为起始年份,用于预测2000-2012年的城市扩展,将预测得到的结果与2012年实际城市用地进行对比,随机森林模型的总精度为83.1%,Kappa系数为0.531;而逻辑回归模型的总精度为80.5%,Kappa系数为0.458。可见,随机森林模型相对于逻辑回归模型具有更好的可靠性,更加适用于城市扩展的模拟与预测。值得指出的是,总精度包括遥感分类图非城市用地对应的不转变精度和城

表3 随机森林模型与逻辑回归模型模拟的混淆矩阵

Tab. 3 Confusion matrix of the RF-CA model and logistic CA model simulation results

年份	类型	随机森林模型			逻辑回归模型			
		非城市	城市	精度/%	非城市	城市	精度/%	
1988–2000 年	遥感分类图	非城市	3444759	220694	94.0	3406151	259302	92.9
		城市	220008	383507	63.6	253128	350387	58.1
	总精度			89.7			88.0	
	Kappa	0.580			0.508			
2000–2012 年	遥感分类图	非城市	2901659	360751	88.9	2844209	418201	87.2
		城市	360647	645911	64.2	416346	590212	58.6
	总精度			83.1			80.5	
	Kappa	0.531			0.458			

市用地对应的转变精度,其中不转变精度由于受佛山市未转变为城市用地面积所占比例较大的影响(表3),导致随机森林模型与逻辑回归模型的不转变精度相差不大,但随机森林模型的转变精度分别在1988-2000年、2000-2012年较逻辑回归模型提高5.5%、5.6%,这表明随机森林CA模型相对逻辑回归模型来说,具有一定的优势。

4 结论与讨论

在城市CA模型中,如何合理有效地确定转换规则是CA模拟的关键。本文运用随机森林方法,在决策树的生成过程中引入随机因素,对训练集进行Bootstrap重采样得到多个样本,对每个Bootstrap样本用随机空间变量选取法建立决策树,然后组合多棵决策树的预测,通过投票得出城市发展适宜性,以此构建CA模型的转换规则。由于在决策树生成过程中引入了随机性,随机森林算法是一种自然的非线性建模工具,不易出现过拟合现象,能够对城市扩展中存在的随机因素有很好的容忍度。随机森林方法还能够根据各变量对预测的贡献程度,对各变量进行重要性度量,从而可解释各空间变量在城市扩展中的作用。相比逻辑回归评价各空间变量对城市扩展的贡献时可能得到与实际情况不太符合的结果(杨青生等, 2007),随机森林方法能够更加有效地评价空间变量的重要性。对随机森林方法的训练时间进行检验,表明通过并行构建方法训练能够大幅减少训练所需的时间。由此可见,基于随机森林的CA模型具有精度高、运算性能快、解释性好等优点。

将该模型应用于广东省佛山市,以不同年份的

卫星遥感影像得到的城市转变状况为因变量,各空间变量为自变量,使用随机森林方法提取CA转换规则。在训练过程中得到的空间变量重要性度量表明,距国道的距离和距城市中心的距离是影响佛山市城市扩展最重要的两个空间变量。使用构建好的随机森林模型模拟佛山市1988-2000年和2000-2012年两个时段的城市用地增长,利用目视对比和逐点对比方法对模型精度进行评价。研究结果表明,基于随机森林的CA模型相对于逻辑回归模型,在模拟佛山市2000年以及预测2012年的城市发展格局时,分别能够提高1.7%和2.6%的模拟精度,更加适用于具有复杂非线性特征的城市系统演变模拟。

城市是一个非常复杂的非线性系统,城市用地的扩张不仅受距城市中心的距离、地形条件等确定性因素的作用,还会受到政策、经济环境等反映城市系统不确定性的随机因素的影响,在本文中主要考虑了确定性因素的作用,只是对现实世界的一个初步的简化模型。事实上,政策和经济环境等因素的变化对城市扩展的总量和分布的影响也是很明显的。此外,由于本文所提出的模型是基于城市扩展的历史演变数据提取CA的转化规则,而未来的城市扩展特征可能与过去存在一定的差别,因此,在应用该模型进行城市扩展的模拟与预测时应结合当地的实际情况进行细致的分析和判断。

参考文献(References)

蔡林霖. 2013. 随机森林的模型选择及其并行化方法[D]. 哈尔滨: 哈尔滨工业大学. [Cai L L. 2013. Model selection of random forest and its parallelization[D]. Harbin, China: Harbin Institute of Technology.]

段晓东, 王存睿, 刘向东. 2012. 元胞自动机理论研究及其



- 仿真应用[M]. 北京: 科学出版社. [Duan X D, Wang C R, Liu X D. 2012. Cellular automata theory research and simulation applications[M]. Beijing, China: Science Press.]
- 方匡南, 吴见彬, 朱建平, 等. 2011. 随机森林方法研究综述[J]. 统计与信息论坛, 26(3): 32-38. [Fang K N, Wu J B, Zhu J P, et al. 2011. A review of technologies on random forests[J]. Statistics & Information Forum, 26(3): 32-38.]
- 冯永玖, 刘妙龙, 童小华, 等. 2010. 基于核主成分元胞模型的城市演化重建与预测[J]. 地理学报, 65(6): 665-675. [Feng Y J, Liu M L, Tong X H, et al. 2010. Kernel principal components analysis based cellular model for restructuring and predicting urban evolution[J]. Acta Geographica Sinica, 65(6): 665-675.]
- 冯永玖, 刘艳, 韩震. 2011. 不同样本方案下遗传元胞自动机的土地利用模拟及景观评价[J]. 应用生态学报, 22(4): 957-963. [Feng Y J, Liu Y, Han Z. 2011. Land use simulation and landscape assessment by using genetic algorithm based on cellular automata under different sampling schemes[J]. Chinese Journal of Applied Ecology, 22(4): 957-963.]
- 何春阳, 史培军, 陈晋, 等. 2005. 基于系统动力学模型和元胞自动机模型的土地利用情景模型研究[J]. 中国科学: 地球科学, 35(5): 464-473. [He C Y, Shi P J, Chen J, et al. 2005. Developing land use scenario dynamics model by the integration of system dynamics model and cellular automata model[J]. Science in China: Earth Sciences, 48(11): 1979-1989.]
- 柯新利, 孟芬, 马才学. 2014. 基于粮食安全与经济发展区域差异的土地资源优化配置: 以武汉城市圈为例[J]. 资源科学, 36(8): 1572-1578. [Ke X L, Meng F, Ma C X. 2014. Optimizing land resource allocation based on food security and regional difference in economic development: a case study in Wuhan metropolitan[J]. Resources Science, 36(8): 1572-1578.]
- 廖江福, 唐立娜, 王翠平, 等. 2014. 城市元胞自动机扩展邻域效应的测量与校准研究[J]. 地理科学进展, 33(12): 1624-1633. [Liao J F, Tang L N, Wang C P, et al. 2014. Measuring and calibrating extended neighborhood effect of urban cellular automata model based on particle swarm optimization[J]. Progress in Geography, 33(12): 1624-1633.]
- 刘小平, 黎夏, 叶嘉安, 等. 2007. 利用蚁群智能挖掘地理元胞自动机的转换规则[J]. 中国科学: 地球科学, 37(6): 824-834. [Liu X P, Li X, Yeh A G O, et al. 2007. Discovery of transition rules for geographical cellular automata by using ant colony optimization[J]. Science in China: Earth Sciences, 50(10): 1578-1588.]
- 龙瀛. 2011. 面向空间规划的微观模拟: 数据、模拟与评价[D]. 北京: 清华大学. [Long Y. 2011. Urban microsimulation for spatial plan: data, modelling, and evaluation[D]. Beijing, China: Tsinghua University.]
- 王云飞, 庞勇, 舒清态. 2013. 基于随机森林算法的橡胶林地上生物量遥感反演研究: 以景洪市为例[J]. 西南林业大学学报, 33(6): 38-45. [Wang Y F, Pang Y, Shu Q T. 2013. Counter-estimation on aboveground biomass of Hevea brasiliensis plantation by remote sensing with random forest algorithm: a case study of Jinghong[J]. Journal of Southwest Forestry University, 33(6): 38-45.]
- 杨青生. 2008. 地理元胞自动机及空间动态转换规则的获取[J]. 中山大学学报: 自然科学版, 47(4): 122-127. [Yang Q S. 2008. Dynamic transition rules for geographical cellular automata[J]. Acta Scientiarum Naturalium Universitatis Sunyatseni, 47(4): 122-127.]
- 杨青生, 黎夏. 2007. 基于遗传算法自动获取CA模型的参数: 以东莞市城市发展模拟为例[J]. 地理研究, 26(2): 229-237. [Yang Q S, Li X. 2007. Calibrating urban cellular automata using genetic algorithms[J]. Geographical Research, 26(2): 229-237.]
- 张鸿辉, 尹长林, 曾永年, 等. 2008. 基于SLEUTH模型的城市增长模拟研究: 以长沙市为例[J]. 遥感技术与应用, 23(6): 618-623. [Zhang H H, Yin C L, Zeng Y N, et al. 2008. Study on urban growth simulation based on SLEUTH model: Changsha City as an example[J]. Remote Sensing Technology and Application, 23(6): 618-623.]
- 张亦汉, 黎夏, 刘小平, 等. 2013. 耦合遥感观测和元胞自动机的城市扩张模拟[J]. 遥感学报, 17(4): 872-886. [Zhang Y H, Li X, Liu X P, et al. 2013. Urban expansion simulation by coupling remote sensing observations and cellular automata[J]. Journal of Remote Sensing, 17(4): 872-886.]
- Breiman L. 2001a. Random forests[J]. Machine Learning, 45(1): 5-32.
- Breiman L. 2001b. Statistical modeling: the two cultures[J]. Statistical Science, 16(3): 199-231.
- Breiman L, Friedman J H, Stone C J, et al. 1984. Classification and regression trees[M]. Boca Raton, FL: CRC press.
- Chen Y M, Li X, Wang S J, et al. 2013. Simulating urban form and energy consumption in the Pearl River Delta under different development strategies[J]. Annals of the Association of American Geographers, 103(6): 1567-1585.
- Genuer R, Poggi J -M, Tuleau-Malot C. 2010. Variable selection using random forests[J]. Pattern Recognition Letters, 31(14): 2225-2236.
- Hastie T, Tibshirani R, Friedman J H. 2008. The elements of statistical learning: data mining, inference, and prediction

- (2nd ed.)[M]. New York, NY: Springer.
- Kandaswamy K K, Chou K C, Martinetz T, et al. 2011. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties[J]. *Journal of Theoretical Biology*, 270(1): 56-62.
- Li X, Yeh A G O. 2001. Calibration of cellular automata by using neural networks for the simulation of complex urban systems[J]. *Environment and Planning A*, 33(8): 1445-1462.
- Li X, Yeh A G O. 2002. Neural-network-based cellular automata for simulating multiple land use changes using GIS[J]. *International Journal of Geographical Information Science*, 16(4): 323-343.
- Liu X P, Li X, Liu L, et al. 2008. A bottom-up approach to discover transition rules of cellular automata using ant intelligence[J]. *International Journal of Geographical Information Science*, 22(11-12): 1247-1269.
- Peters J, De Baets B, Verhoest N E C, et al. 2007. Random forests as a tool for ecohydrological distribution modelling[J]. *Ecological Modelling*, 207(2-4): 304-318.
- Rodriguez-Galiano V F, Ghimire B, Rogan J, et al. 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67: 93-104.
- Wu F L. 2002. Calibration of stochastic cellular automata: the application to rural-urban land conversions[J]. *International Journal of Geographical Information Science*, 16(8): 795-818.

## Urban expansion simulation by random-forest-based cellular automata: a case study of Foshan City

CHEN Kai<sup>1,2</sup>, LIU Kai<sup>1,2\*</sup>, LIU Lin<sup>1,2</sup>, ZHU Yuanhui<sup>1,2</sup>

(1. Center of Integrated Geographic Information Analysis, School of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China; 2. Guangdong Key Laboratory for Urbanization and Geo-simulation, Guangzhou 510275, China)

**Abstract:** Cellular automata (CA) has been frequently used to investigate the logical nature of self-reproducible systems and simulate the evolution of complex geographical phenomena such as urban expansion. The core of cellular automata is to define transition rules. Traditionally, approaches for defining the transition rules of cellular automata had difficulty to balance the interpretability, accuracy, and convenience. This article presents a new cellular automata model for simulating urban expansion based on random forest algorithm. The proposed model extracts CA transition rules of urban expansion by introducing random factors in training samples and candidate spatial variables that split nodes during the multiple decision trees building process. One significant advantage of our approach is that it can be easily adopted for parallel implementation and has high prediction accuracy and tolerance to random factors in urban expansion. Another strength of the proposed approach is that it can estimate out-of-bag errors to obtain model parameters quickly and measure the importance of spatial variables and explain the contribution of each variable in urban expansion. The model was applied to simulate urban expansion in Foshan City, Guangdong Province. We used the urban land change of 1988 and 2000 as the dependent variable and the spatial variables as the independent variables to construct the CA model based on random forests, then simulate and predict urban expansion of 2000 and 2012. The results show that random forest model can improve the simulation and prediction accuracies by 1.7% and 2.6%, respectively, when compared to the logistic regression model commonly used in CA simulation. This suggests that random forest model is superior for modeling complex nonlinear urban evolution. Urban expansion of Foshan City in 2024 was also predicted according to its urban development trend. Through measuring the importance of some spatial variables that affect urban expansion, we found that distance to national roads and to the city center are the two most important spatial variables for urban expansion simulation in Foshan City.

**Key words:** random forest; cellular automata; urban expansion; Foshan City