

主成分分析与 Cellular Automata 在空间决策 与城市模拟中的应用*

黎 夏

(广州地理研究所, 广州 510070)

叶嘉安

(香港大学城市规划及环境管理研究中心)

摘要 探讨了多准则判断技术(MCE)在空间分析中所碰到的变量相关的问题. 变量之间的相关使得使用的权重不恰当, 导致分析结果不合理. 以城市发展的空间模拟为例, 指出了在一般的空间分析中所使用的因子有很大的相关性. 提出了利用主成分分析与“理想点”的方法来解决这一问题. 结合单元自动演化(Cellular Automata)和地理信息系统(GIS), 将该方法应用在城市形态的优化模拟中.

关键词 主成分分析 Cellular Automata 地理信息系统 城市模拟

Cellular Automata(单元自动演化模型, 简称 CA)首先由 von Neumann 和 Ulam 在 1948 年提出, 随后被广泛应用于探讨和模拟各类生物与物理系统的演化过程, 例如气体的变化, 生物的繁殖及演化. 在 CA 模拟中, 利用简单的规则可以形成复杂的系统, 这与生物学中的 DNA 所起的作用十分相似. CA 模型的优点已引起不同领域学者的注意, 它已被应用在物理、化学、生物、生态及城市研究中^[1, 2].

由于 CA 具有强大的空间运算能力, 它很适合于土地利用变化及城市形态演变等复杂系统的模拟. 近年来, CA 越来越多地被应用于城市系统的模拟中, 取得了许多有意义的研究成果^[2-7]. CA 模型可以方便地用于检验各种假设, 如城市的分形结构及其演化等. 将 CA 与地理信息系统(GIS)相结合, 可以大大提高它们的模拟能力. 这也有助于将一系列约束性条件引入 CA 模型来产生可持续的城市发展形态^[5, 8]. 另外, 城市的空间模拟涉及复杂的因素, 需要考虑多种空间变量, 已有研究将 CA 与 MCE 结合来进行城市形态的模拟^[4].

在空间分析及规划决策的制定过程中, 往往涉及到多种复杂和冲突的因素, 需要采用多准则判断技术(MCE)^[9]. 多准则判断是在 70 年代发展起来的, 现已被广泛地用于复杂的多因素分析及规划决策过程中^[10]. 但是, MCE 所处理的准则必须是相互独立的, 准则之间的相关性会导致决策的失误^[11].

本研究将探讨在城市模拟中的各种空间变量的相关性. 提出了使用主成分分析方法(PCA)来消除这种相关性, 使得 CA 的模拟更具有合理性. PCA 方法已经被广泛用于空间数据的处理及压缩中, 有十分显著的效果. 该原理是通过正交旋转变换, 产生新的变量集, 消除原数据中的冗余度. 将 PCA 引进 CA 模拟中, 将能明显改善 CA 的模拟效果, 消除一般 MCE 的不合理性, 目前在国际上还没有开展有关研究. 所提出的方法不仅仅局限于 CA 模拟的应用,

也可以使用于其他空间决策及评价过程中。

1 主成分分析与 CA 空间模拟

当准则较多时, 确定各个准则的权重将很困难. 而且, 当准则之间有较大的相关性时, 所选取的权重也会不准确. 引进 PCA 方法可以有效地解决这个问题. PCA 方法是通过正交旋转变换的方法来消除原数据中的相关性或冗余度. 其正交旋转的公式如下^[12]:

$$pc_{ij} = \sum_{k=1}^n X_{ik} E_{kj}, \quad (1)$$

这里 pc_{ij} 是对应于象元 i 的第 j 个主成分; X_{ik} 是对应于象元 i 的第 k 个准则; E_{kj} 是对应第 k 行第 j 列的特征向量矩阵.

特征向量和特征值可以由下列方程来求解:

$$E \text{Cov} E^T = V. \quad (2)$$

这里 Cov 是协方差阵; V 是以特征值为对角值的矩阵; E 是特征向量矩阵; T 为转置.

利用 PCA 可以生成一系列独立不相关的新变量(主成分). 将新变量代替原变量用于 CA 模拟中, 摆脱 MCE 权重不合理性的弊端, 并能方便地使用更广泛的空间变量, 以改善模型的精度. 以下介绍如何将 PCA 引进 CA 模型中的方法. 一般的 CA 模型可以表达归纳为如下的形式^[8]:

$$S^{t+1} = f(S^t, N). \quad (3)$$

这里 S 是状态, f 是邻近函数, N 是邻近范围, t 是迭代运算时间. CA 模型的特点是 $t+1$ 时的状态取决于 t 时邻近范围内的状态.

CA 模型的状态一般是离散的. 在城市模拟中, 不同的状态用来反映不同的土地利用类型. 由此可以利用 CA 来模拟土地利用变化和城市的发展过程. 一般的 CA 模型是用一个二进位的数来表达不同状态转换的过程: 1 为转换, 0 为不转换. 该方法有一定的局限性, 不能反映状态转换的连续过程. 我们在这里使用‘灰度’(G)来表达状态的连续变化过程. 当‘灰度’值从 0 逐渐变为 1 时, 表示该单元最终完成状态的转变. 例如, 由农业用地转换为城市用地. 该 CA 模型的迭代公式如下:

$$G_i^{t+1} = G_i^t + \Delta G_i^t, \quad G_i \in (0, 1). \quad (4)$$

‘灰度’的增加值由两个方面决定: 邻近函数和相似度. 邻近函数反映了周围象元状态对中心象元状态转换的影响. 例如, 在时间 t 时邻近范围已经转变为城市用地的象元越多, 其中心象元在时间 $t+1$ 时转变为城市用地的概率越高, 它们呈比例关系^[13]. 除此之外, 我们还提出相似度来度量中心象元与‘理想点’在各项属性方面的差异, 由此来确定‘灰度’的增加值. ‘理想点’是指最适合于某一土地利用用途(例如城市用地)的点, 它能获得最大理论收益值. 例如, 根据土地适宜性评价, 可以获得每个象元对不同准则关于城市发展适宜性的一系列值^[14]. ‘理想点’就是具有所有准则最大值的点. 在 CA 模拟中, 某个象元其属性越接近‘理想点’, 其‘灰度’的增加速度越快.

图 1 显示了由最大准则值来确定‘理想点’的方法. ‘理想点’可以表达如下:

$$\xi = (S_1^{\max}, S_2^{\max}, S_3^{\max}, \dots, S_j^{\max}, \dots, S_K^{\max}), \quad (5)$$

这里 S_j^{\max} 是第 j 个准则的最大值.

事实上, 这个‘理想点’是一个虚拟点. 可以通过主成分变换求出其变换后的对应主成分值. 通过‘理想点’的方法, 可以把一系列环境和可持续发展要素引进 CA 模型中, 以形成合理的城市形态. 可持续发展的‘理想点’应该保证能获得这些准则的最大值. 由于这些准则往往是相关的, 需要消除它们的相关性. 通过主成分变换, 把反映各种经济、环境和资源等要素的空间变量, 来作为象元的属性. 由此可以计算某个象元与‘理想点’的相似度. 其公式如下:

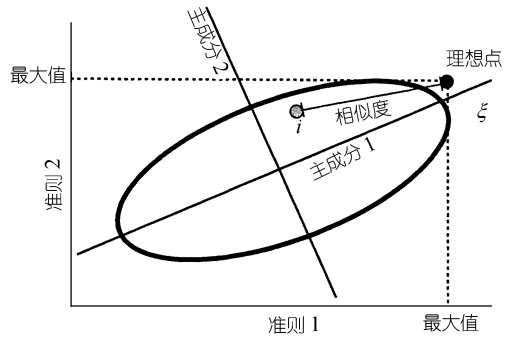


图 1 主成分变换与‘理想点’

$$d_{i\xi} = \sqrt{\sum_j^m w_j^2 (pc_{ij} - pc_j^0)^2} . \tag{6}$$

这里 $d_{i\xi}$ 是象元 i 与‘理想点’ ξ 之间的相似度; pc_{ij} 是象元 i 的第 j 个主成分的值; pc_j^0 是‘理想点’的第 j 个主成分值; w_j 是第 j 个主成分在计算相似度时的权重.

可将相似度进行标准化, 使得其值在 0~1 之间. 标准化的相似度为:

$$SIM = 1 - \frac{d_{i\xi}}{d_{i\xi}^{\max}} , \tag{7}$$

这里 $d_{i\xi}^{\max}$ 是 $d_{i\xi}$ 的最大值.

由此, ‘灰度’的增加值应该与邻近函数和标准化的相似度成比例关系. 有

$$\Delta G_i^t = f_i(q^t, N) \times SIM^t = \frac{q^t}{\pi l^2} \times \left(1 - \frac{d_{i\xi}^t}{d_{i\xi}^{\max}} \right)^k . \tag{8}$$

这里 q^t 为在时间 t 时邻近范围内已经转变为城市用地的象元数; l 为邻近范围的半径; k 为非线性指数变换的参数.

通过非线性变换及参数 k 的选择, 可以有效地产生不同的模拟形态^[4,5,8]. 可以把随机变量引进 CA 模型中, 使得模拟效果更接近现实^[3]. 随机变量项可以由下式表达:

$$RA = 1 + (-\ln \gamma)^\alpha , \tag{9}$$

其中 γ 为值在 {0, 1} 范围内的随机数; α 为控制随机变量影响大小的参数.

公式(8)引进随机变量项后, 变为:

$$\Delta G_i^t = RA \times \frac{q^t}{\pi l^2} \times \left(1 - \frac{d_{i\xi}^t}{d_{i\xi}^{\max}} \right)^k = [1 + (-\ln \gamma)^\alpha] \times \frac{q^t}{\pi l^2} \times \left(1 - \frac{d_{i\xi}^t}{d_{i\xi}^{\max}} \right)^k . \tag{10}$$

该公式决定象元状态的转变. 在 CA 的每次迭代运算过程中, 如果某一象元的‘灰度’变为 1 时, 该象元就转变为城市用地. 提高 CA 模型的不迭代运算, 可以模拟出城市这个复杂系统的演变及其优化形态.

2 应用及分析

将上述模型应用在深圳和东莞市的城市发展的空间模拟中. 首先需要对影响城市发展的空间因素进行度量. 这些空间变量的影响往往由一系列距离函数来度量. 例如, 转变为居民点的概率取决与一系列基础设施, 包括学校、医院、交通和公园等. 可以用距离的梯度函数来表达这些因素对城市发展的影响^[15]. 这些影响可以是正面或负面的. 例如, 越接近交通要道、城市中心, 城市发展所得到的‘收益’越大; 反之, 越接近一些环境和生态脆弱或保护地区, 城市发展所得到的‘收益’就越小(或‘代价’越高). ‘收益’的具体大小可以由距离的梯度函数来表达, 并作为决策判断的准则. 利用遥感和 GIS 可以方便地获得这些空间判断准则. 本研究所采用的正面空间判断因子有:

a) 离市中心的距离; b) 离镇中心的距离; c) 离铁路的距离; d) 离高速公路的距离; e) 离公路的距离; f) 离河流的距离.

对于这些正面影响, 城市发展所获得的‘收益’, 即进行空间决策判断的准则, 可以由下面梯度函数来表达:

$$X_j = e^{-\beta_j dist_j} \quad (11)$$

这里 X_j 是变量 j 的准则值; $dist_j$ 是离开目标物的距离; β_j 为衰减系数. 准则值大小在 0~1 范围内.

所采用的负面空间变量有:

a) 离耕地的距离; b) 离果园的距离; c) 离菜地的距离; d) 离鱼塘的距离; e) 离水库的距离; f) 离森林的距离; g) 离湿地的距离.

离这些目标物的距离越近, 给资源和环境所带来的干扰和破坏就越大, 城市发展带来的‘收益’就越小. ‘收益’的梯度函数可以由下式来表达:

$$X_j = 1 - e^{-\beta_j dist_j} \quad (12)$$

这些空间判断准则是进行一般 GIS 选址以及城市模拟的主要依据. 这些空间变量往往是相关的, 直接利用多准则判断存在着一定的问题. 而且, 有时空间判断所使用的准则达几十及上百个之多^[16], 更无法给出合适的权重值. 为了使模拟结果更加合理性, 我们将 PCA 方法引进 CA 模拟中, 消除原数据中的冗余度.

表 1 是深圳和东莞市各个空间变量的主成分分析的结果. 由表 1 可以看到, 前 5 个主成分已经包含了原始 13 个变量中高达 90% 以上的信息(深圳 93.9%, 东莞 92%). 即使前 3 个主成分也已经包含了原始数据中 80% 以上等信息(深圳 88.8%, 东莞 81.4%). 这说明原始数据中的冗余度还是相当高的, 变量之间有很高的相关性, 不能直接使用多准则判断的方法.

表 2 仅显示了东莞每个主成分包含原来的 13 个变量的信息承载量(loading)情况. 主成分所对应的系数值越大, 包含原变量的成分越高. 由此可以具体分析各个主成分的组成. 例如, 主成分 I 集中反映了原变量中农业和生态的有关成分, 主要包含了菜地、鱼塘、湿地和果园等信息; 主成分 II 则基本反映了交通的信息, 主要包含了公路、高速公路等信息; 主成分 III 主要反映了离市中心和镇中心距离的空间信息. 可以看到, PCA 变换可以有效地将相似(相关)的空间变量放到相同的主成分上. 因此, 基于 PCA 的模拟比基于 MCE 的模拟能获得更合理的结果. 因为当准则之间相关时, MCE 所给出的权重就不准确, 会导致重复使用某些准则.

表 1 主成分及信息量

主成分	深圳		东莞	
	特征值	信息量/%	特征值	信息量/%
I	90.4	64.1	62.9	44.4
II	25.9	18.4	38.9	27.5
III	8.8	6.2	13.5	9.5
IV	3.7	2.6	8.5	6.0
V	3.6	2.5	6.5	4.6
VI	3.1	2.2	3.2	2.3
VII	1.8	1.3	2.6	1.9
VIII	1.2	0.9	1.9	1.4
IX	1.0	0.7	1.7	1.2
X	0.5	0.4	0.9	0.7
XI	0.5	0.4	0.5	0.3
XII	0.4	0.3	0.3	0.2
XIII	0.1	0.1	0.1	0.1

表 2 各个主成分的信息承载量

距离变量	主成分												
	I 农业和生态	II 交通	III 城市中心	IV 河流	V 高速公路	VI 耕地	VII	VIII	IX	X	XI	XII	XIII
市中心	-0.10	0.07	0.47	-0.50	0.02	0.04	-0.03	-0.07	0.06	0.07	-0.07	-0.17	-0.69
镇中心	-0.15	0.05	0.45	-0.52	-0.06	-0.05	0.01	-0.11	-0.04	-0.03	0.06	0.15	0.67
铁路	0.16	0.17	-0.07	-0.15	-0.72	0.15	0.27	0.53	-0.04	0.04	0.00	-0.13	0.01
高速公路	-0.26	0.62	-0.11	-0.09	0.51	0.03	0.09	0.50	-0.10	-0.05	0.03	0.05	0.01
公路	-0.07	0.64	-0.34	-0.08	-0.29	-0.07	-0.10	-0.59	0.07	0.06	-0.02	-0.01	-0.02
河流	-0.43	0.21	0.54	0.63	-0.24	0.11	-0.05	0.00	0.07	-0.06	-0.02	0.01	0.00
耕地	0.18	0.06	0.06	0.03	0.11	0.74	0.05	-0.22	-0.58	0.05	-0.01	-0.07	0.03
果园	0.23	0.10	0.15	0.11	0.18	0.00	0.85	-0.22	0.30	0.03	0.06	-0.01	0.02
菜地	0.49	0.20	0.17	0.05	0.07	0.01	-0.18	0.06	0.16	-0.32	-0.71	0.07	0.08
鱼塘	0.48	0.19	0.17	0.05	0.03	0.10	-0.31	0.05	0.25	-0.25	0.68	0.03	-0.03
水库	0.21	0.09	0.14	0.08	-0.10	-0.34	0.08	0.01	-0.45	0.14	0.06	0.71	-0.21
森林	0.16	0.09	0.15	0.10	0.02	-0.52	0.06	-0.05	-0.49	-0.17	0.07	-0.62	0.06
湿地	0.25	0.12	0.15	0.11	0.11	-0.06	-0.19	0.09	0.12	0.87	-0.06	-0.17	0.14

我们提出了使用‘理想点’的方法将反映经济、环境和资源的要素放进 CA 模型中. 为了消除数据的冗余度, 各个要素最后用主成分来表达. 城市用地的‘理想点’应具有最大的准则值, 其主成分变换前的坐标值为(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1). 为了取得最大的经济及环境保护的‘收益’, 应该首先选取最接近该‘理想点’的象元进行城市发展, 从而获得城市形态的 CA 优化模拟结果.

由于大部分信息已经包含在前面几个主成分上, 这里可以仅用前 6 个主成分来计算相似度. 主成分变换后的‘理想点’坐标为:(1.2, 2.6, 1.9, -0.2, -0.4, 0.1). 在多准则判断中, 需要对不同的因子根据其重要性给予一定的权重. 对于不同的目标和假设, 可以有不同的权重组合, 从而得到不同的结果. 例如, 在城市规划中, 可以利用不同的权重组合来获得不同的规划

方案. 但当因子多达几十及上百时, 就很难给出合适的权重. 利用主成分分析, 可以有效地将多因子压缩到少数的几个分量上, 对它们给出权重就非常容易了.

本文利用前 6 个主成分来计算相似度. 由表 1 可以具体分析每个主成分所包含的要素的构成, 从而可以根据不同的规划目标给出权重. 权重值在 0~1 范围内, 其赋值一般是根据专家经验来决定. 在实际应用中, 可以通过多个专家的平均打分来决定. 越大的权重值表示该主成分的重要性越高. 例如, 假如需要严格保护生态和农业时, 就应该给予第一主成分较大的权重值. 表 3 中给出了 5 种不同的权重组合, 分别对应了 5 种不同的城市发展方案. 在实际应用中, 可以根据具体需要来修改这些权重或给出其他权重组合, 这是十分方便的.

表 3 不同权重组合及发展方案

主成分	各主成分权重				
	方案 1 以城市中心为主 (市中心和镇中心)	方案 2 以交通为主 (高速 公路, 公路, 河流)	方案 3 保护耕地	方案 4 保护生态与农业 (蔬菜, 鱼 塘, 水库, 湿地和森林)	方案 5 经济与保护 环境
I 农业和生态	0.25	0.25	0.25	1.00	1.00
II 交通	0.25	1.00	0.25	0.25	1.00
III 城市中心	1.00	0.25	0.25	0.25	1.00
IV 河流	0.25	0.25	0.25	0.25	0.50
V 高速公路	0.25	0.25	0.25	0.25	0.50
VI 耕地	0.25	0.25	1.00	0.25	1.00

权重赋值: 非常重要 - 1.00; 很重要 - 0.75; 重要 - 0.50; 不太重要 - 0.25; 不重要 - 0

对应于这些不同组合, 可以很方便地由 CA 自动产生不同的规划方案. 图 2 是深圳以交通为主(高速公路和公路)的 1988~1997 的城市发展模拟方案. 该方案给主要包含了交通信息的第二主成分赋予了较大的权重, 使得城市主要在交通方便的地方快速发展; 图 3(a)是东莞以市中心和镇中心为主的发展方案(方案 1). 在该方案中, 主要包含市中心和镇中心信息的第三主成分被赋予较大的权重, 使得城市主要围绕现有的市中心和镇中心发展; 由于第六主成分主要包含了耕地信息, 对该主成分赋予较大的权重, 就可以形成保护耕地的发展方案. 图 3(b)是

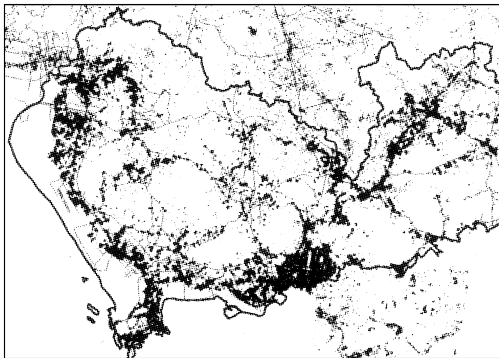


图 2 深圳市沿交通干线为主的城市发展的 CA 模拟

由此方案所得的东莞市模拟结果. 如图 3(b)所示, 该方案可以有效地保护位于西北部东江三角洲平原上的农田; 由于第一主成分主要包含农业和生态信息, 如蔬菜、鱼塘、水库、湿地和森林. 对该主成分赋予较大的权重, 形成了能有效地保护农业生态环境的发展方案. 由方案 4 所得到的东莞市的模拟结果见图 3(c); 在大多数情况下, 经济发展与环境保护是有一定的冲突的. 在实际应用中, 只顾及经济利益或只考虑环境保护都是行不通的, 必须兼顾不同方面的考虑. 方案 5 通过给各种因子一定的权重, 反映了各方面的需要, 图 3(d)是由此方案所获得的模拟结果.

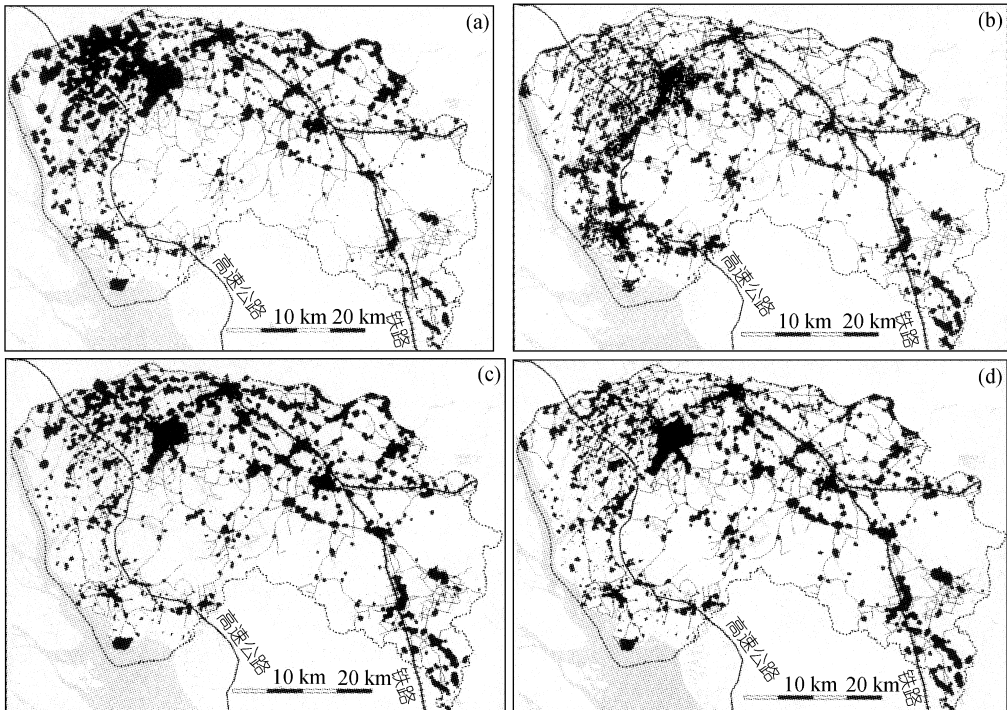


图 3 东莞市的城市发展 CA 模拟

(a) 基于市和镇中心为主; (b) 强调保护耕地; (c) 保护农业和生态; (d) 基于经济和环境综合因素

3 结论

在大多数的空间决策过程中, 往往涉及到许多空间要素, 要使用多准则判断技术来进行评价分析. 这些空间要素可以由一般的地理信息系统(GIS)分析功能来获取. 研究表明, 空间要素之间存在着很大的相关性, 直接使用多准则判断并不恰当. 要素之间的相关性会使得在多准则判断中重复使用相关的因子, 不符合多准则判断的原则. 本研究提出了利用主成分分析及‘理想点’的方法来消除这种相关性, 使得多准则判断更具有合理性. 将 PCA 方法引进 CA 模型进行空间决策模拟, 能十分方便地产生不同的城市发展方案, 为城市发展规划提供依据, 可以作为城市规划工作者有用的工具.

本文利用了一系列空间变量进行空间决策和城市发展的 CA 模拟. 在实际应用中, 还需考虑到不同的开发强度对城市形态的影响. 如何将开发强度引进 CA 模拟中, 目前有关研究不多, 还需进一步探讨.

参 考 文 献

- 1 Binder P. Evidence of lagrangian tails in a lattice gas. In: Manneville P, Boccara N, Vichniac G Y, eds. Cellular Automata and Modeling of Complex Physical Systems. Berlin : Springer-Verlag, 1989. 155~160
- 2 Batty M, Xie Y. From cells to cities. Environment and Planning B: Planning and Design, 1994, 21: 531~548
- 3 White R, Engelen G, Uijee I. The use of constrained cellular automata for high-resolution modelling of urban land-use

- dynamics. *Environment and Planning B*, 1997, 24: 323~343
- 4 Wu F, Webster C J. Simulation of land development through the integration of cellular automata and multicriteria evaluation. *Environment and Planning B*, 1998, 25: 103~126
 - 5 黎 夏, 叶嘉安. 约束性单元自动演化 CA 模型及可持续城市发展形态的模拟. *地理学报*, 1999, 54(4): 289~298
 - 6 黎 夏, 叶嘉安. 单元自动演化模型自动生成农田保护区的研究. *中国环境科学*, 2000, 20(4): 318~322
 - 7 周成虎, 孙战利, 谢一春. *地理元胞自动机研究*. 北京: 科学出版社, 1999
 - 8 Li X, Yeh A G O, Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science*, 2000, 14(2): 131~152
 - 9 Carver S J. Integrating multi-criteria evaluation with geographical information systems. *International Journal of Geographical Information Systems*, 1991, 5(3): 321~339
 - 10 Nijkamp P, van Delft A. *Multi-criteria analysis and regional decision-making*. Netherlands: H E Stenfert Kroese BV, 1977
 - 11 Malczewski J. *GIS and multicriteria decision analysis*. New York: John Wiley & Sons Inc, 1999
 - 12 Gonzalez R C, Wintz P. *Digital image processing*. Reading, Massachusetts: Addison-Wesley Publishing Company, 1977
 - 13 Batty M. Cellular automata and urban form: a primer. *Journal of the American Planning Association*, 1997, 63(2): 266~274
 - 14 Yeh A G O, Li X. Sustainable land development model for rapid growth areas using GIS. *International Journal of Geographical Information Science*, 1998, 12(2): 169~189
 - 15 Batty M, Xie Y C, Sun Z L. Modeling urban dynamics through GIS-based cellular automata. *Computer, Environment and Urban Systems*, 1999, 23: 205~233
 - 16 Bauer V, Wegener M. A Community information feedback system with multiattribute utilities. In: Bell D E, Keeney R L, Raiffa H, EDS. *Conflicting Objectives in Decisions*. West Sussex, England: Wiley, 1977. 323~357