

基于 Logistic 回归的 CA 模型改进方法 以广州市为例

聂 婷, 肖荣波*, 王国恩, 刘云亚

(广州市城市规划勘测设计研究院, 广州 510060)

摘要: 基于 Logistic 回归的 CA 模型因其结构简单和数据要求相对较小的优势, 被广泛应用于城市模拟领域, 但数据的空间自相关性影响了模型机制挖掘与模拟精度。通过将影响城市发展演变的各种约束条件划分为强制和普通约束条件, 以及运用主成分分析降低普通约束条件的数据相关性, 构建了改进型 Logistic 回归 CA 模型, 并在 2000~2008 年广州市城市增长模拟研究中进行应用。结果表明: 与传统型 Logistic 回归 CA 模型相比, 改进型 Logistic 回归 CA 模型在模型拟合度和精度上均有 4% 左右的提高。其中约束条件划分对非城市像元模拟精度约有 6% 的提高, 对整体精度有 3% 的提高。更为重要的是, 降低数据相关性后, Logistic 回归 CA 模型对于城市扩展机制的解释更符合实际。本研究旨在寻求一种简单可行且易于构建的 CA 模型, 探求城市发展机理, 为城市规划管理提供更为准确的科学依据。

关键词: 逻辑回归; 元胞自动机; 空间自相关; 城市增长模拟

文章编号: 1000-0585(2010)10-1909-11

1 引言

作为微观动力学模型的代表, 元胞自动机 (Cellular Automata, 简称 CA) 广泛应用于城市模拟领域^[1], 并可与空间决策支持系统^[2,3]、生态学^[4]、情景规划^[5,6]及生命科学^[7]等技术或理论集成应用。CA 基本的组成要素包括元胞 (Cells)、元胞空间 (Lattices)、元胞状态 (States)、领域 (Neighbors) 及转化规则 (Rules), 最核心部分就是定义转换规则, 转换规则中的参数对模拟结果有着重要的影响^[8]。确定元胞转换规则的方法目前主要有基于 Logistic 回归、Markov 过程、多准则判断、主成分分析和神经网络等。其中较早的基于 Logistic 回归对土地利用动态变化的成功模拟由 Wu 等实现^[9], 研究显示基于土地转换概率的 Logistic 回归较多元回归分析有更强的能力解释城市增长。荷兰瓦赫宁根农业大学土壤科学与地质实验室的 Verburg 等人^[10]将该模型与土地利用非空间需求模块集成, 开发了 CLUE-S 模型, 实现了多种同时发生的土地利用方式变化的模拟。因其简单和数据要求相对较小, Logistic 回归 CA 模型被广泛于城市扩张机制挖掘和情景模拟应用中^[11~15], 然而模型中存在的自变量空间相关、约束条件差异化限定等问题多被忽视, 因此如何提高模型精度、更加准确地揭示城市扩张机制等需要有更为深入的研究。

针对传统型 Logistic 回归 CA 模型在数据处理方面存在的空间自相关问题, 本研究从

收稿日期: 2009-12-03; 修订日期: 2010-04-18

基金项目: 十一五 国家科技支撑计划 (2007BAC28B01)

作者简介: 聂婷 (1983-), 女, 江西宜春人, 硕士, 规划师。E-mail: nieting2007@gmail.com

* 通讯作者: 肖荣波 (1978-), 博士, 高级规划师。E-mail: ecxiaorb@163.com

两方面入手构建改进型 Logistic 回归 CA 模型: (1) 从约束条件对转换规则的影响出发, 将约束条件划分为强制约束条件和普通约束条件, 两种约束条件的数据处理方法有所差异; (2) 通过对普通约束条件进行主成分分析, 降低空间数据的相关性, 减少数据误差对模型模拟精度的干扰。并以广州市城市增长模拟为例, 从这两方面对传统型 Logistic 回归 CA 模型进行了改进和运用。

2 基于 Logistic 回归的 CA 模型改进

2.1 基于 Logistic 回归的转换规则定义

城市化过程就是非城市用地向城市用地转换, 如果给元胞赋状态集 $\{0$ (非城市化), 1 (城市化) $\}$, 只需要判断某个元胞的状态在 $[t, t+1]$ 时刻间是否从 0 变化到 1, 就可判断该元胞所代表的地块是否城市化, 这就是 Logistic 回归 CA 模型的最基本思想。假设一个区位的发展概率是一系列独立变量 (如离市中心的距离、离高速公路的距离、地形高程和坡度等) 所构成的函数, 但在实际应用中因变量是二项分类常量, 即将土地利用分为 0 (非城市化) 与 1 (城市化), 不满足正态分布的条件, 这时可用逻辑回归分析, 利用逻辑回归技术对 CA 的转换规则进行校正, 则 t 时刻元胞 (i, j) 的城市化最终发展概率为^[16]:

$$p_{ij}^t(S_{ij} = \text{urban}) = \frac{1}{1 + \exp(-z_{ij}^t)} \quad \text{con}(S_{ij}^t) \quad l_{ij}^t \quad (1)$$

式中, S_{ij} 是单元 (i, j) 的状态, 只有 0 (非城市化) 和 1 (城市化) 两种状态; z 是描述单元 (i, j) 开发特征的向量 $z = a + \sum_k b_k x_k$, a 是一个常量, b_k 是逻辑回归模型的系数, x_k 是一组约束条件变量; $\text{con}()$ 为约束条件, 其值 $[0, 1]$; l_{ij}^t 为领域影响, 表示 t 时刻元胞 (i, j) 的 3×3 窗口内的开发强度; ϵ 为随机变量, 其值大于 1。

抽样数目 (样本量) 的确定一般取决于被调查事物总体之间的差异程度和容许误差大小, 利用逻辑回归技术定义转换规则, 可按照总体数据 20% 的样本量来采集样本数据^[16]。获取到一定样本量的空间变量与土地利用变化的历史经验数据, 再利用逻辑回归得到合适的模拟权重参数。

2.2 模型改进

2.2.1 约束条件划分

广义的讲, 领域影响之外的转换规则可以统称为约束条件^[17]。对于影响城市发展演变的各种约束条件, 根据其对城市发展过程的作用效果, 分为强制约束条件和普通约束条件 2 大类。强制约束条件一般包括自然保护区、森林公园、湖泊、水库等因素, 在自然条件约束和政策调控下, 对城市发展起到绝对性的禁建作用, 其数值呈现二值离散化, 即只有 0 (不可发展) 和 1 (可发展) 两种; 普通约束条件一般包括到各等级城市中心的距离、交通可达性、坡度、高程等因素, 其限制作用体现在城市发展的经济成本上, 其数值是在某个值域范围内的连续分布。

由此可见, 强制约束条件是排除在转换规则之外的, 在符合法律法规的情况下, 其发展概率绝对是零。体现在 CA 模型的实现中, 主要有以下两点: (1) 在选取样本点获取转换规则时, 不允许在强制约束条件范围内选取样本点, 藉此消除数据采样中强制约束条件对模型精度的干扰; (2) 为避免强制约束条件范围内的元胞在模拟增长过程中受邻域影响城市化, 在城市化最终发展概率中 (公式 1), 强制约束条件内的约束条件 $\text{con}()$ 值均设为 0。

2 2 2 主成分分析降低数据相关性

在基于 Logistic 的 CA 模型进行城市增长模拟研究时, 获取转换规则时, 考虑了众多地形、区位和经济社会因素作为影响城市增长的普通约束条件, 但忽视了诸多因素的数据相关性对模型精度的干扰。因此采用 SPSS 统计学分析软件对参与获取转换规则的样本点数据进行数据相关性去除, 首先对各普通约束条件进行主成分分析, 其次用主成分因子代替普通约束条件原始数据进行二项逻辑回归, 最后将各普通约束条件回代入逻辑回归方程 (图 1)。

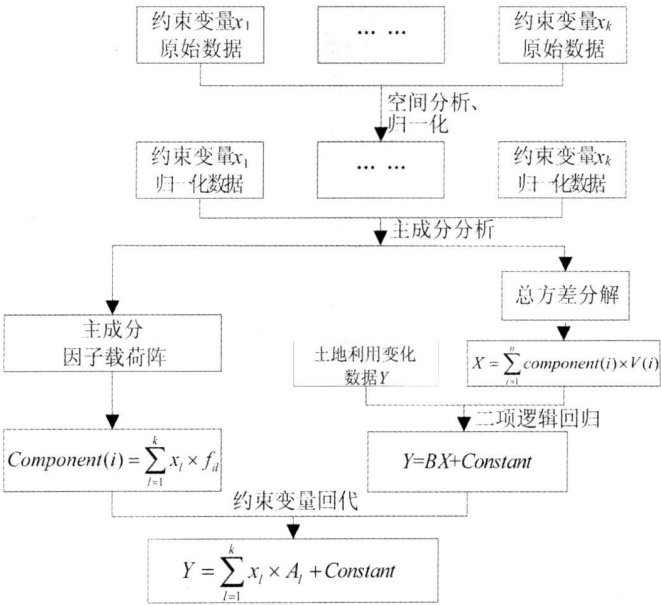


图 1 主成分分析降低数据相关性流程

Fig 1 Flow-chart of reducing data correlation in principal component analysis

2 3 模型模拟精度评价

将模拟结果像元与实际像元进行栅格叠加, 计算模型的模拟精度, 分为城市像元模拟精度、非城市像元模拟精度和总模拟精度, 计算公式如下:

$$= \frac{A}{B} \tag{2}$$

$$= \frac{C}{D} \tag{3}$$

$$= \frac{A + C}{B + D} \tag{4}$$

式中, A 为模拟正确的城市像元数, B 为实际城市像元数, C 为模拟正确的非城市像元数, D 为实际非城市像元数。

3 广州市实例应用

3 1 研究区概况

模拟区域为广州市全境, 跨度为 22 26 ~ 23 56 N、112 57 ~ 114 03 E, 包括中心城区的越秀、荔湾、海珠、天河、白云、黄埔, 近郊区的番禺、花都、南沙、萝岗和远郊区

的增城、从化 12 个区市级行政单元, 全市面积 7434.40 km^2 。2007 年全市常住人口 1004.58 万人, GDP 为 7171.28 亿元, 城镇建设用地 1600.38 km^2 。地势由东北向西南倾斜, 中部、南部是全市农业生产、城市生产、生活的主要集中区。

3.2 数据来源与处理

3.2.1 土地利用变化数据

模型中使用的土地利用/覆盖数据主要是来自于行列号 122/44 的 2 期 Landsat TM/ETM+ 数据 (2000 年 9 月 14 日和 2008 年 3 月 4 日), 像元大小为 $30\text{m} \times 30\text{m}$, 质量较好, 少云或无云, 覆盖广州全市。采用遥感影像处理软件 Erdas 9.0 对遥感影像进行解译, 得到各年份土地利用图。本次模拟研究城市增长过程, 仅涉及到非城镇建设用地向城镇建设用地的转化, 因此将土地利用类型合并为城镇建设用地和非城镇建设用地两类。依据高分辨率影像和相关历史经验数据, 对遥感分类结果进行精度检验, 2000 年遥感分类的 Kappa 系数为 0.9189, 2008 年遥感分类 Kappa 系数为 0.9017, 分类精度较高。根据分类结果, 2000 年建设用地面积为 1130 km^2 , 2008 年为 1595 km^2 , 文中所指现状建设用地均指遥感分类结果中的建设用地。

经统计, 2000~2008 年间, 强制约束条件内 1568.7 km^2 的非城市化用地中仅有 7.6 km^2 转化为城市化用地, 城市化比例仅为 0.5%, 远低于强制约束条件外 9.7% 的比例, 说明约束条件的划分非常有必要。对土地利用数据去除掉强制约束条件范围后, 通过 ArcGIS 的 Spatial Analysis 模块对相邻年份的土地利用数据进行栅格运算, 转化为相邻年份的土地利用变化数据, 栅格取值为 1 (表示两个年份间城市化) 或 0 (表示两个年份间未城市化)。

3.2.2 约束条件变量选取

根据广州实际情况, 强制约束条件含风景名胜区、森林公园、自然保护区和河流湖泊。普通约束条件变量的处理如下:

(1) 普通约束条件变量的选取。兼顾数据的可得性和操作性, 普通约束条件包括距市级中心距离 $DisCity$ 、距片区中心距离 $DisRegion$ 、距镇级中心距离 $DisTown$ 、距铁路距离 $DisRail$ 、距高速/快速路距离 $DisHighway$ 、距一般道路距离 $DisRoad$ 、坡度 $Slope$ 、高程 $Elevation$ 、常住人口密度 $PopDensity$ 、社会固定资产投资密度 $InvesDensity$ 和社会生产总值密度 $GDPDensity$ 等 11 个约束条件变量。

(2) 原始数据的选取 (图 2)。2000 年市级中心包括花园酒店、黄埔港、天河城; 片区中心包括北京路、中大商贸、上下九、花地湾、花都新华、番禺市桥、南沙中心区、增城荔城、从化街口; 镇级中心为广州市各镇。道路数据为 2000 年对应的路网数据。坡度和高程约束条件变量的原始数据为 DEM (数字高程模型) 数据。人口和经济约束条件变量的原始数据为 2000 年的社会经济统计数据。

(3) 数据处理方法。强制约束条件数值均设为 0。普通约束条件中, 距离变量通过对原始数据在 ArcGIS 中进行距离分析 (Distance) 获取, 坡度和高程变量通过对数字高程模型 (DEM) 进行三维分析 (3D Analysis) 获取, 密度变量通过对经济社会统计数据进行分析获取。由此将各类普通约束条件的原始数据转变为吸引力 (potential model) (表 1)。

(4) 归一化。按照 20% 的比例对土地利用变化和普通约束变量按照样本量进行描述统计 (表) 并归一化, 使其取值都位于 0~1 的范围。

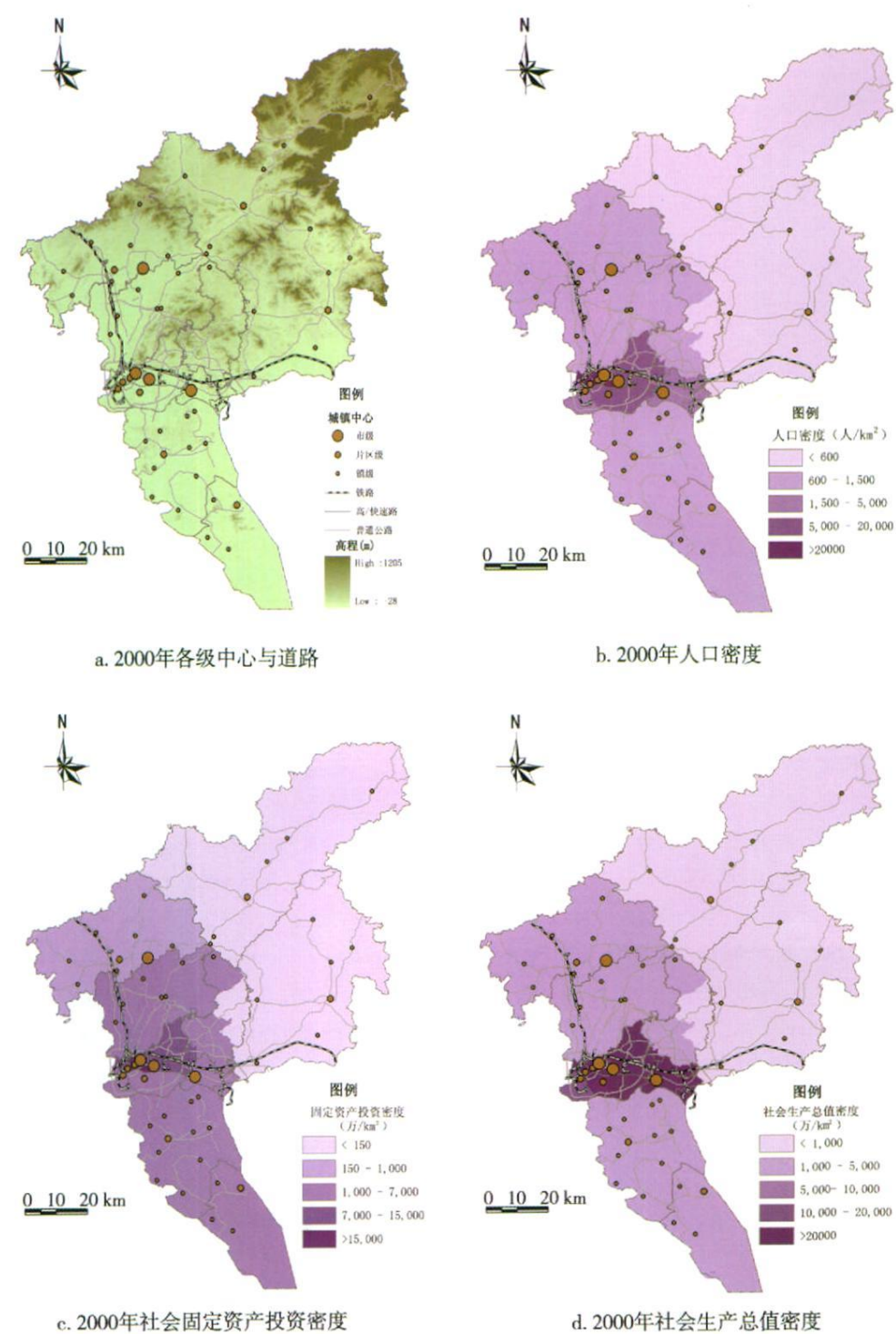


图 2 2000 年广州城市增长普通约束条件分布

Fig 2 Distribution of general constraint in Guangzhou in 2000

表 1 2000 年普通约束条件统计型描述表

Tab 1 Descriptive statistics of general constraints in 2000

约束条件变量	统计描述				相关系数									
	N	Mean	Std. Deviat ion	DisC ity	Dis Region	DisTow n	DisRail	Dis Highway	DisRoad	Slop e	Elevation	Pop Density	Inwes Density	GDP Density
DisCity	74014	44842	22817	1 00	0 49	0 24	0 89	0 93	0 34	0 39	0 62	- 0 31	- 0 28	- 0 32
DisRegion	74014	18136	10840	0 49	1 00	0 41	0 54	0 52	0 27	0 36	0 56	0 01	0 02	0 03
DisTown	74014	6909	3843	0 24	0 41	1 00	0 29	0 24	0 47	0 25	0 32	0 13	0 14	0 12
DisRail	74014	24253	20393	0 89	0 54	0 29	1 00	0 89	0 33	0 38	0 64	- 0 23	- 0 20	- 0 24
DisHighway	74014	23342	21853	0 93	0 52	0 24	0 89	1 00	0 19	0 45	0 68	- 0 24	- 0 23	- 0 26
DisRoad	74014	3421	3655	0 34	0 27	0 47	0 33	0 19	1 00	0 06	0 17	- 0 03	- 0 03	- 0 04
Slope	74014	6	8	0 39	0 36	0 25	0 38	0 45	0 06	1 00	0 70	- 0 10	- 0 09	- 0 10
Elevation	74014	76	127	0 62	0 56	0 32	0 64	0 68	0 17	0 70	1 00	- 0 13	- 0 11	- 0 14
PopDensity	74014	696	1228	- 0 31	0 01	0 13	- 0 23	- 0 24	- 0 03	- 0 10	- 0 13	1 00	0 90	0 89
InwesDensity	74014	871	2760	- 0 28	0 02	0 14	- 0 20	- 0 23	- 0 03	- 0 09	- 0 11	0 90	1 00	0 89
GDPDensity	74014	2472	6074	- 0 32	0 03	0 12	- 0 24	- 0 26	- 0 04	- 0 10	- 0 14	0 89	0 89	1 00
Valid N (listwise)	74014													

注：DisCity、DisRegion、DisTown、DisRail、DisHighway、DisRoad 描述统计单位为m；Slope 描述统计单位为；Elevation 描述统计单位为m；Pop Density 描述统计单位为人/km²；InwesDensity 描述统计单位为万元/ km²；GDPDensity 描述统计单位为万元/ km²。

3 3 模型构建

3 3 1 传统型 Logistic 回归 CA 模型构建

将 2000 年和 2008 年两个年份的归一化约束变量数据作为自变量，2000~ 2008 年的土地利用变化数据作为因变量，直接进行逻辑回归。各约束变量的回归系数在 0.05 的显著性水平上具有统计学意义。这说明 11 个特征变量确实对广州市城市化产生影响。各变量的系数中（表 2），InwesDensity、GDPDensity 等 2 个变量的系数为正值，说明该变量值越大，越有利于城市化；DisCity、DisTown、DisRail、DisRoad、Slope、Elevation、PopDensity、GDPDensity 等 9 个变量的系数均为负值，说明该变量值越小，城市化概率越大。回归方程分类预测总体拟合精度为 64.7%（分类截断值取 0.50），其中 38272 个城市化元胞中有 29916 个被正确预测，城市化预测拟合的正确率为 78.2%；35742 个非城市化元胞中有 17938 个被正确预测，非城市化预测拟合的正确率为 50.2%。

运用直接逻辑回归系数，按照公式（1）构建传统型 Logistic 回归 CA 模型。需要指出的是，城市增长过程中，必然要受到各种强制政策的影响。此处为了对比表现约束条件划分对模型精度的影响，将这种强制政策忽略，除河流外，con（）均取为 1。

3 3 2 约束条件划分 Logistic 回归 CA 模型构建

在传统逻辑回归模型基础上，仅做约束条件划分改进。将 2000 年和 2008 年两个年份的强制约束条件范围外的归一化约束变量数据作为自变量，土地利用变化数据作为因变量，直接进行逻辑回归。回归方程分类预测总体拟合精度为 69.3%（分类截断值取 0.50），其中 38588 个城市化元胞中有 30969 个被正确预测，城市化预测拟合的正确率为 80.0%；226749 个非城市化元胞中有 21581 个被正确预测，非城市化预测拟合的正确率

为 58.5%。相比传统型 Logistic 回归 CA 模型, 如果仅进行区分约束条件, 也可提高模型特别是非城市化像元精度。

运用逻辑回归系数, 按照公式(1), 将强约束条件内 $con()$ 值均设为 0, 构建 CA 模型。

3.3.3 改进型 Logistic 回归 CA 模型构建

对 2000 年和 2008 年两个年份的归一化约束变量数据进行主成分分析。将主成分因子作为自变量, 去除强约束条件范围后的 2000~2008 年的土地利用变化数据作为因变量, 进行逻辑回归。

(1) 主成分分析

在 SPSS 中对样本数据各约束条件进行主成分分析, 只有前三个因子的特征值大于 1, 所解释的方差占总方差的 77.37%。从主成分分析的因子载荷阵可以看出 (表 2), 第一主成分与 $DisRegion$ 、 $DisRoad$ 和 $DisHighway$ 这 3 个变量的相关较高, 说明第一主成分主要反映区位条件; 第二主成分与 $GDPDensity$ 、 $InvesDensity$ 和 $DisRegion$ 这 3 个变量的相关较高, 说明第二主成分主要反映经济社会条件; 第三主成分主要反映地形条件。

(2) 二项逻辑回归结果

将 11 个约束变量用 3 个主成分及其解释量归并为一个变量 $X = 0.41911 Component(1) + 0.24628 Component(2) + 0.1083 Component(3)$, 进行二项逻辑回归, 回归系数在 0.05 的显著性水平上具有统计学意义。回归方程分类预测总体拟合精度为 69.6% (分类截断值取 0.50), 其中 38272 个城市化元胞中有 30842 个被正确预测, 城市化预测拟合的正确率为 80.6%; 35742 个非城市化元胞中有 20673 个被正确预测, 非城市化预测拟合的正确率为 57.8%。

(3) 逻辑回归系数

将主成份因子用各约束条件变量的归一化数据回代, 则得到各类约束条件变量原始数据的逻辑回归系数 (表 3)。

运用改进逻辑回归系数, 按照公式 (1), 将强约束条件内 $con()$ 值均设为 0, 构建改进型 Logistic 回归 CA 模型。

3.4 模型比较

3.4.1 不同模型对城市扩展机制解释的比较

在 2000~2008 年广州市城市扩展过程的拟合中, 传统型、约束条件划分和改进型 Logistic 回归 CA 模型中各约束条件变量的逻辑回归系数与城市化的关系一致 (表 3), 其中反映区位、地形条件的变量及反映经济社会条件的 $PopDensity$ 变量呈负相关, 说明该变量值越小, 城市化概率越大; 反映经济社会条件的 $InvesDensity$ 和 $GDPDensity$ 变量呈正相关, 说明该变量值越大, 城市化概率越大。

表 2 主成分分析的因子载荷阵

Tab 2 Factor loading matrix of principal component analysis			
约束条件变量	主成分		
	1	2	3
$DisRegion$	0.903	0.051	-0.019
$DisTown$	0.628	0.399	0.031
$DisRail$	0.373	0.474	0.531
$DisHighway$	0.884	0.146	-0.024
$DisRoad$	0.896	0.113	-0.176
$Slope$	0.368	0.225	0.785
$Elevation$	0.586	0.198	-0.381
$PopDensity$	0.792	0.255	-0.286
$InvesDensity$	-0.440	0.842	-0.099
$GDPDensity$	-0.420	0.851	-0.109
$DisCity$	-0.449	0.837	-0.109

第三主成分与 $Slope$ 变量的相关较高, 说明

表 3 传统及改进 CA 模型的逻辑回归系数

Tab 3 Logistic regression coefficients of traditional and improved CA model			
普通约束条件变量	传统模型	约束条件划分模型	改进模型
	逻辑回归系数	逻辑回归系数	逻辑回归系数
<i>DisCity</i> (距市级中心距离)	- 0.0195*	- 0.0191*	- 0.0249*
<i>DisRegion</i> (距片区中心距离)	- 0.0352*	- 0.0356*	- 0.0491*
<i>DisTown</i> (距镇级中心距离)	- 0.0027	- 0.0034*	- 0.1254*
<i>DisRail</i> (距铁路距离)	- 0.0151*	- 0.0144*	- 0.0289*
<i>DisHighway</i> (距高速/快速路距离)	- 0.0171*	- 0.0162*	- 0.0257*
<i>DisRoad</i> (距一般道路距离)	- 0.0524*	- 0.0534*	- 0.1175*
<i>Slope</i> (坡度)	- 0.1066*	- 0.1058*	- 0.0453*
<i>Elevation</i> (高程)	- 0.0026*	- 0.0027*	- 0.0042*
<i>PopDensity</i> (常住人口密度)	- 0.0073*	- 0.0021*	- 0.0014*
<i>InvesDensity</i> (社会固定资产投资密度)	0.0072*	0.0040*	0.0011*
<i>GDPDensity</i> (社会生产总值密度)	0.0008	0.0005*	0.0001*
<i>Constant</i> (常量)	1.2774	1.2448	5.2488*

注：显著性处于 0.001 水平。

相对于传统和约束条件划分模型，改进型模型降低了数据的相关性，改变了约束条件变量的权重和相互间关系，使得对城市扩张机制的解释产生较大差异甚至是根本性的改变。从传统和约束条件划分模型可以看出，坡度（*Slope*）是影响广州城市扩张的首要因素，其次为到一般道路距离（*DisRoad*）。而在改进模型中，反映区位条件的变量权重均得到强化，反映地形和经济社会条件的权重得到降低，其中到镇级中心距离（*DisTown*）变量成为影响城市扩张最为重要的因素。可以理解为在 2000~ 2008 年期间，区位条件尤其是重点镇的发展带动成为影响广州城市扩张最为重要的因素。而在现实中，广州自 2003 年以来大力推动中心镇建设，村办、镇办企业多集中在镇区，诸多远离主城区的居住区也依托镇区良好的基础设施和公共服务设施资源进行建设；另外广州城市扩展主要集中在番禺、花都和增城南等平原地区，城市扩张空间腹地较大，实际受到地形条件约束较小。显然改进型模型对于城市扩张的解释更符合实际情况。

3.4.2 不同模型模拟精度比较

分别采用传统、约束条件划分与改进 Logistic 回归 CA 模型，以 2000 年为模拟基期，对 2008 年的广州市城市发展过程进行了模拟重建，并比较 2008 年实际，根据公式(2)~ (4) 分别计算城市化精度、非城市化精度和总精度（图 3、表 4）。基于传统 Logistic 回归 CA 模型的城市像元模拟精度为 73.53%，非城市像元模拟精度为

表 4 传统及改进 CA 模型的模拟精度

Tab 4 Simulation accuracy of traditional and improved CA model			
模拟精度 (%)	传统 Logistic	约束条件划分	改进 Logistic
	回归 CA 模型	Logistic 回归 CA 模型	回归 CA 模型
城市像元模拟精度	73.53	72.65	76.04
非城市像元模拟精度	83.91	89.46	90.67
总模拟精度	82.97	85.93	86.90
Kappa 系数	0.7639	0.8054	0.8161

传统 Logistic 回归 CA 模型的城市像元模拟精度为 73.53%，非城市像元模拟精度为

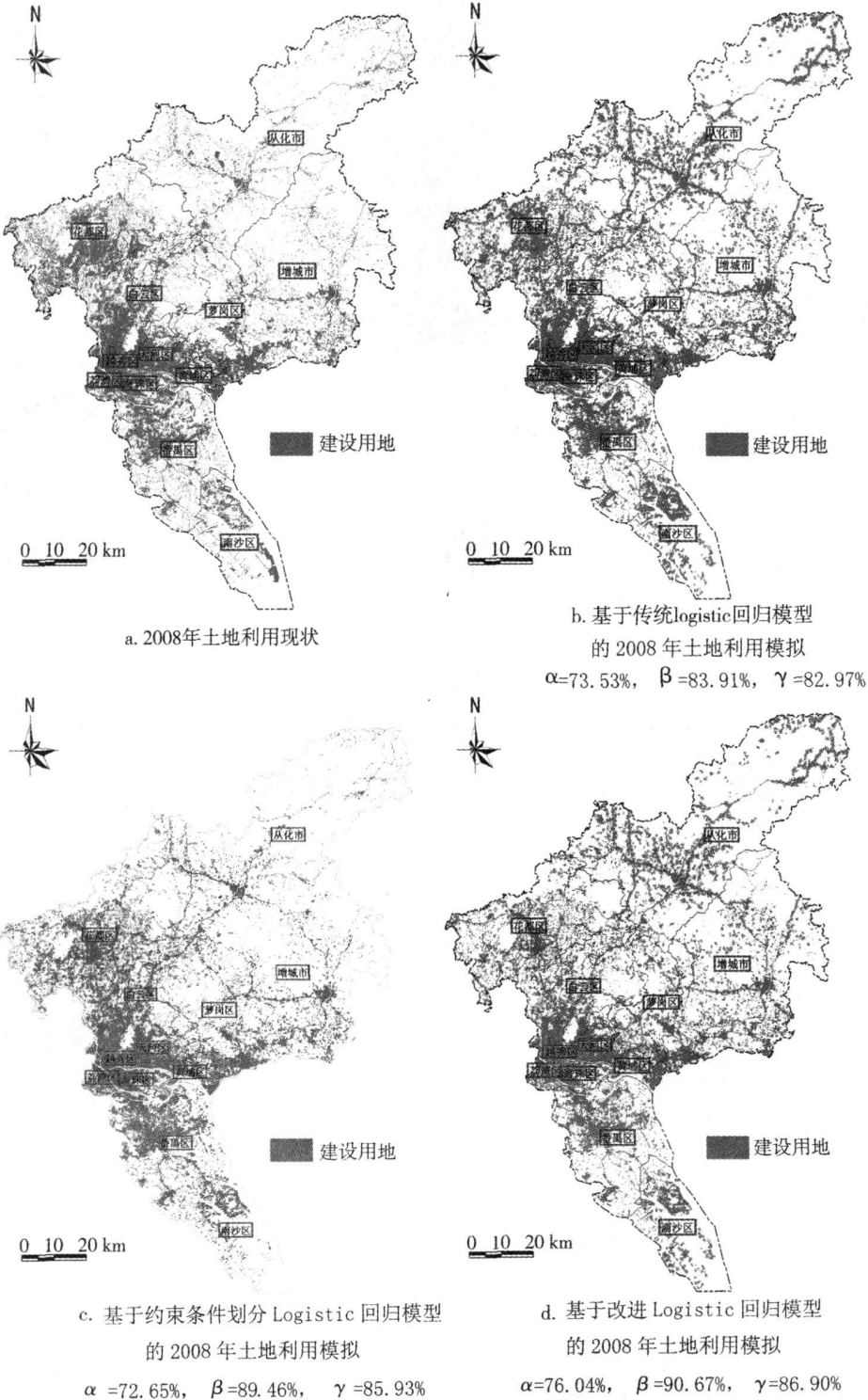


图3 2000~2008年广州市城市增长模拟

Fig 3 Urban growth simulation of Guangzhou during 2000~2008

83.91%, 总模拟精度为 82.97%, Kappa 系数为 0.7639。基于约束条件划分 Logistic 回归 CA 模型的城市像元模拟精度为 72.65%, 非城市像元模拟精度为 89.46%, 总模拟精度为 85.93%, Kappa 系数为 0.8054。基于改进 Logistic 回归 CA 模型的城市像元模拟精度为 76.04%, 非城市像元模拟精度为 90.67%, 总模拟精度为 86.90%, Kappa 系数为 0.8161。

从模拟精度看, 约束条件划分对模型模拟精度特别是非城市像元模拟精度的提高有重大贡献。改进型较约束条件划分 Logistic 回归 CA 模型的总体模拟精度提高不大, 但城市像元模拟精度的提高非常显著。

4 结论与讨论

通过主成分分析和划分约束条件类型对传统 Logistic 回归 CA 模型进行了改进, 结果表明改进型的模型能更加准确地揭示城市扩张机制, 为制定相关规划方案和政策提供更为科学地技术支撑。同时, 改进型的模型模拟精度达到 86.9%, 比传统模型高出 4%。

研究结果表明约束条件划分与相关性去除对模型改进的作用不相同。其中约束条件划分对于 CA 模型中非城市化像元的模拟精度有较大影响, 而数据的相关性对城市增长机制的解释有重要影响。在广州市应用实例中可以看出传统和约束条件划分模型揭示的城市扩张机制与实际情况有较大偏差, 如沿用传统和约束条件划分模型认为影响城市扩张的首要因素是坡度, 则很可能导致错误的规划决策因此, 对于目前诸多应用基于 Logistic 回归的 CA 模型进行城市模拟及辅助决策的研究, 数据相关性是一个不可忽视的问题。

本研究主要讨论逻辑回归 CA 模型的改进, 与已有的基于主成分分析的 CA 模型^[3]相比, 两者在消除变量的相关性上均有实现, 但根本区别在于转换规则获取方法不一样。基于主成分分析的 CA 模型定义了取值区间为[0, 1]的灰度值来反映元胞状态的连续变化, 一个元胞灰度值的增加值由两方面决定: 邻近元胞函数、与最适合于某种土地用途(例如城市用地)的理想点元胞的相似度。可见, 基于主成分分析的 CA 实现时需要确定理想点, 而理想点的界定比较模糊。相比而言, 基于逻辑回归的 CA 模型更简便和易操作。

土地利用变化是一个非常复杂的过程, 对其进行的建模和预测模拟, 往往受制于模型本身以及数据等问题。本文在传统的基于 Logistic 回归 CA 模型的基础上, 尝试性地进行了改进, 试图寻求一种相对简单而精度较高的模型, 以更为精确地探求城市发展过程中各种因素的影响机理, 对城市规划管理提供更为准确的科学依据。

参考文献:

- [1] Batty M, Xie Y From cells to cities Environment and Planning B: Planning and Design, 1994, 21: 531~ 548
- [2] White R, Stratman B, Engelen G Planning scenario visualization and assessment: a cellular automata based integrated spatial decision support system. In: Goodchild M F, Janelle D G, Shrore Z G Spatially Integrated Social Science London: Oxford University Press, 2004. 420~ 442
- [3] 黎夏, 叶嘉安 主成分分析与 Cellular Automata 在空间决策与城市模拟中的应用 中国科学(D 辑), 2001, 31(8): 683~ 690
- [4] 刘小平, 黎夏, 彭晓鹃 生态位元胞自动机在土地可持续规划模型中的应用 生态学报, 2007, 27(6): 2391~ 2402
- [5] 何春阳, 陈晋, 史培军, 等 大都市区城市扩展模型-以北京城市扩展模拟为例 地理学报, 2003, 58(2): 294~ 304
- [6] 吴晓青, 胡远满, 贺红土, 等 沈阳市城市扩张与土地利用变化多情景模拟 地理研究, 2009, 28(5): 1264~ 1275
- [7] 王行风, 汪云甲, 李永峰 基于生命周期理论的煤矿区土地利用演化模拟 地理研究, 2009, 28(2): 379~ 390
- [8] Wu F Calibration of stochastic cellular automata: The application to rural-urban land conversions International Journal of Geographical Information Science, 2002, 16(8): 795~ 818

- [9] Wu F, Yeh A G I Changing spatial distribution and determinants of land development in Chinese cities in the transition from a centrally planned economy to a socialist market economy: A case study of Guangzhou Urban Studies, 1997, 34(11): 1851~ 1879
- [10] Verburg P H, Soepboer W, Veldkamp A I Modeling the spatial dynamics of regional land use: The CLUE-S model Environmental Management, 2002, 30(3): 391~ 405
- [11] 段增强, Verburg P H, 张凤荣, 等 1 土地利用动态模拟模型的构建及其应用))) 以北京市海淀区为例 1 地理学报, 2004, 59(6): 1037~ 1047
- [12] 邓祥征, 刘纪远, 战金艳, 等 1 太仆寺旗土地利用变化时空格局的动态模拟 1 地理研究, 2004, 23(2): 147~ 156
- [13] 姜文亮, 张晓通, 李霖, 等 1 基于 GIS 和空间逻辑回归模型的城市空间扩展预测))) 以深圳市龙岗区为例 1 测绘科学, 2008, 33(4): 172~ 174
- [14] 龙瀛, 毛其智, 沈振江, 等 1 综合约束 CA 城市模型: 规划控制约束及城市增长模型 1 城市规划学刊, 2008, (6): 83~ 91
- [15] 龙瀛, 韩昊英, 毛其智 1 利用约束性 CA 制定城市增长边界 1 地理学报, 2009, 64(8): 999~ 1008
- [16] 黎夏, 叶嘉安, 刘小平, 等 1 地理模拟系统: 元胞自动机与多智能体 1 北京: 科学出版社, 2007. 51~ 53
- [17] White R, Engelen G I The use of constrained cellular automata for high-resolution modelling of urban land use dynamics 1 Environment and Planning B: Planning and Design, 1997, 24: 323~ 343

An improvement on CA model of logistic regression: A case study of Guangzhou

NIE Ting, XIAO Rong-bo, WANG Guo-en, LIU Yun-ya

(Guangzhou Urban Planning & Design Survey Research Institute, Guangzhou 510060, China)

Abstract: Due to its simple structure and less input data, CA model of logistic regression is widely applied in urban simulation. However, data dependency has some impact on the accuracy. Therefore, an in-depth research should be conducted to modify the traditional model.

This paper established an improved CA model of logistic regression in two major aspects. First, the urbanization factors were divided into forbidden constraint and general constraint. The input data were sampled only in general constraint, while the urbanization probability in forbidden constraint was set to be 0. Second, we reduced the data dependency of general constraint using principal component analysis in SPSS. In the case study of Guangzhou, the improved CA model was applied to simulate the urban growth from 2000 to 2008. Compared to the traditional CA model, the improved CA model made a 4% improvement both on model fitness and simulation accuracy, in which constraints division contributed a 3% improvement on overall simulation accuracy and a 6% improvement on non-urban simulation accuracy, while data dependency reduction gave a more reasonable explanation for urbanization mechanism. The study aimed to establish an improved CA model, which can mine a more reasonable urbanization mechanism, and provide more scientific support for urban planning and land management.

Key words: logistic regression; cellular automata; data dependency; urban simulation