

文章编号: 1005-0523(2006)02-0067-04

基于马尔可夫模型的地理信息系统研究

刘艳丽, 刘觉夫, 谢 昕

(华东交通大学 信息工程学院, 江西 南昌 330013)

摘要:以“江西国土资源数字化研究”为项目背景,阐述了马尔可夫模型与地理信息系统结合的方法和意义,重点论述了基于数值区间划分的马尔可夫数值预测模型的理论基础和实现方案,理论论证和比较实验得出了马尔可夫预测模型在大量历史数据进行转移概率统计的情况下对于波动较大的数据有较好的预测效果.

关键词:地理信息系统;马尔可夫模型;数值域划分

中图分类号:TP391.41

文献标识码:A

1 引言

本课题来源于“江西国土资源数字化研究”这个项目.国土资源包含内容复杂多样,各种数据庞杂.以前,这些数据是以报表的形式汇总分析,由于办公自动化水平低下,大量工作主要靠人工完成,工作量巨大.随着地理信息系统(GIS)^{[1][2]}充分利用,国土资源研究管理逐步走向现代化和科学化.但是,在实际应用中,地理信息系统经常会出现数据分析不够充分,建立的数学模型不尽合理等现象.这是因为地理信息数据库基础之上的,地理信息往往是海量数据库,其信息动辄达到GB甚至TB级.这些地理数据的取得一般都需耗费大量的人工、时间和资金.当面对海量地理信息数据库和大量的应用需求的时候,GIS和地理学中数学方法合理的结合^{[3][4][5]},构建更为有效安全的数据分析和决策支持平台.同样,数学方法只有程序化并和GIS平台紧密结合才能发挥其数据分析处理的优势.

另一方面,地理信息系统需要用数学方法实现对地理数据的综合分析和处理以满足不同应用的需要,包括运用数学方法为其建造空间分析模型,如空间统计分析模型、叠加分析模型、缓冲分析模

型等,以及其他的分析模型,如综合评价模型,预测模型等,就连GIS中的一些基本技术如数据格式的转换都需要借助有关的数学方法来实现.因此本文利用马尔可夫模型对土地利用变化进行模拟,观测其未来的演变趋势,以促进研究国土资源的合理永续的利用.

2 土地利用变化的马尔可夫过程模拟

马尔可夫过程是指每次状态的转移都只与前一刻的状态有关,而与过去的状态无关,或者说是无后效性的状态转移过程.这样的过程适合于研究土地利用的动态变化.其原因是在一定条件下,土地利用的动态演变具有马尔可夫过程的性质:(1)在一定区域内,不同土地利用类型之间具有相互可转化性;(2)土地利用类型之间的相互转化过程包含着较多尚难用函数关系准确描述的事件.

要运用马尔可夫过程,须首先确定土地利用类型之间相互转化的初始转移概率矩阵P,其数学表达式为:

收稿日期:2005-12-12

基金项目:江西省自然科学基金资助(0311046)

作者简介:刘艳丽(1979-),女,湖北汉川,助教,研究方向为GIS及图像处理.

$$P = (P_{ij}) = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1n} \\ P_{21} & P_{22} & \dots & P_{2n} \\ \dots & \dots & \dots & \dots \\ P_{n1} & P_{n2} & \dots & P_{nn} \end{pmatrix} \quad \text{其中 } P_{ij} = \frac{C_{i-j}}{LU_i} \quad (1)$$

式中 n 为研究区域的土地利用类型数目, 为初始到末期由类型 i 转化为类型 j 的概率, 表示由研究区域中在研究期内第类土地利用类型转化为第类的面积, LU_i 为研究期初第类土地利用类型面积, P_{ij} 它就满足以下条件:

$$\begin{cases} 0 \leq P_{ij} \leq 1 & (i, j = 1, 2, 3, \dots, n) \\ \sum_{i=1}^n P_{ij} = 1 & (j = 1, 2, 3, \dots, n) \end{cases} \quad (2)$$

根据马尔可夫过程的无后效性及 Bayes 条件概率公式, 有:

$$P(n) = P(n-1)P_{ij} \quad \text{或} \quad P(n) = P(0)P^n$$

式中 $P(n)$ 为经过 $n-1$ 次状态转移到达 n 次转移后的状态概率向量, $P(n-1)$ 为经过 $n-2$ 次状态转移到达 $n-1$ 次转移后的状态概率向量, P 或 P_{ij} 为转移概率矩阵, $P(0)$ 为它的初始状态概率向量.

3 马尔可夫预测算法

一般来说, 预测在数据分析中是比较难掌握的. 首先, 它是对将来未知事件的数值估计, 其实, 目前的预测算法一般对于被预测数据有较高要求, 主要是原始数据的波动性不能太大, 但是, 有时正是对不平稳的数据序列进行预测才有实用价值, 这就是预测算法的难点所在. 马尔可夫预测算法主要由三个部分组成, 即数据序列预处理模块、状态转移概率统计模块、预测模块.

3.1 预处理模块

主要完成原始数据序列数值域的等值划分工作, 这个过程必须在原始数据序列对应于划分的子区间映射而成的状态序列保持马氏性的前提下进行才有意义. 预处理模块得出可以划分的最大子区间数, 这个参数不仅将决定马尔可夫预测模型的最高数值精度, 而且还与预测模型的预测精度紧密关联, 一般来说该参数越大则数值精度越高; 要保证预测精度则该参数也不能太小. 但是, 这个参数越也将极大影响预测算法的运行效率, 参数越大矩阵规模越大, 当然, 能提供较大子区间参数的数据序

列一般也会有相当的规模, 这就意味着预处理的时间越长. 所以需要数据序列进行马氏性检验, 具体的方法^[6]如下: 设数据序列 ζ 为 $\zeta_1, \zeta_2, \dots, \zeta_p$ 将其数值域划分为 $[x_0, x_1), [x_1, x_2), \dots, [x_{n-1}, x_n]$, 设状态空间 $E = \{E_1, E_2, \dots, E_n\}, E_1 \overset{\Delta}{=} [x_1, x_2), \dots,$

$E_n \overset{\Delta}{=} [x_{n-1}, x_n)$, 设 $n_i, 1 \leq i \leq p$ 为数据序列 ζ 属于状态 $E_j, 1 \leq j \leq n$ 的数据个数, $n_{ij}, 1 \leq i \leq n, 1 \leq j \leq n$ 表示 $\zeta_1, \zeta_2, \dots, \zeta_p$ 经状态 $E_i, 1 \leq i \leq n$ 的数据个数, $n_{ij}, 1 \leq i \leq n, 1 \leq j \leq n$ 表示 $\zeta_1, \zeta_2, \dots, \zeta_p$ 经状态 $E_i, 1 \leq i \leq n$ 一步转移至 $E_j, 1 \leq j \leq n$ 的频数. 令

$$Q_{ij} = \frac{n_{ij}}{n_j} \quad 1 \leq i \leq n, 1 \leq j \leq n \quad (3)$$

$$Q_{0j} = \frac{\sum_{i=1}^n n_{ij}}{\sum_{i=1}^n n_i} \quad 1 \leq j \leq n \quad (4)$$

$$\lambda^2 = 2 \sum_{i=1}^n \sum_{j=1}^n n_{ij} |\log \frac{Q_{ij}}{Q_{0j}}| \quad (5)$$

当数据序列具有一定规模时, 统计量 λ^2 服从自由度为 $(n-1)^2$ 的 λ^2 分布^[20], 在置信度为 α 的情况下, 如果 $\lambda^2 \geq \lambda_{\alpha}^2(n-1)^2$, 则认为该序列为马尔可夫链. 实验中, 置信度 α 为 0.001 以保证检测合格的序列具有较好的马氏性.

3.2 状态转移概率统计模块

马氏链状态划分算法可以确定某一数据序列在保持马氏性的前提下其数值域最多可以等分为多少子区间, 从而确定数据片的大小. 这个算法可以用来对原始序列进行预处理把其构造为马尔可夫数值预测可以处理的序列, 从而使马尔可夫数值预测可以在安全的状态下预测任意的数据序列. 所以, 可以认为在保持马氏性前提下, 任何数据序列的数值域都可以进行一定程度的等值划分, 并且相应地具有稳定的状态转移矩阵(至少是单位阵^[3]).

假设 ζ 具有稳定的状态转移矩阵, 所以预测的真实值 x 处于某一个数值子区间 $[x_{i-1}, x_i]$ 的真实概率为常数 p_{ki} , 其中, k 指被预测数据序列 ζ 在上一个状态处于子区间 $[x_{k-1}, x_k]$, 由于处于不完备统计的情况下, 状态转移矩阵中的相应概率为 p_{ki} , 假设在子区间内 ζ 服从等概率分布, 于是, 预测值为:

$$x = \sum_{i=1}^n p_{ki} (x_{i-1} + x_i) / 2 \quad (6)$$

则加权平均误差为

$$\begin{aligned} AEP &= \frac{1}{S} \sum_{i=1}^n p_{ki} \int_{x_{i-1}}^{x_i} (x - x) dx \\ &= \frac{1}{S} \sum_{i=1}^n p_{ki} \int_{x_{i-1}}^{x_i} [x - \sum_{j=1}^n p_{kj} (x_{j-1} + x_j) / 2] dx \end{aligned}$$

(数值域为等值划分, $x_i = x_0 + is$ (s 为子区间长度, 或称为数据片长度))

$$\begin{aligned}
 &= \frac{1}{s} \sum_{i=1}^n p_{ki} \int_{x_{i-1}}^{x_i} [x - \frac{1}{2} \sum_{j=1}^n p_{kj} (2x_0 + (2j-1)s)] \\
 d_x \\
 &= \frac{1}{s} \sum_{i=1}^n p_{ki} [\frac{1}{2} (x_i^2 - x_{i-1}^2) - (x_i - x_{i-1}) \sum_{j=1}^n p_{kj} (x_0 + \frac{2j-1}{2}s)] \\
 &= \frac{1}{s} \sum_{i=1}^n p_{ki} [s(x_0 + \frac{2i-1}{2}s) - s \sum_{j=1}^n p_{kj} (x_0 + \frac{2j-1}{2}s)] \\
 &= \sum_{i=1}^n p_{ki} [x_0(1 - \sum_{j=1}^n p_{kj}) + (\frac{2i-1}{2})s - \sum_{j=1}^n p_{kj} \frac{2j-1}{2}s] = s \\
 &[\sum_{i=1}^n ip_{ki} - \sum_{i=1}^n p_{ki} (\sum_{j=1}^n jp_{kj})] \text{ (由状态转移矩阵的性质得} \\
 &\sum_{i=1}^n p_{ki} = 1, \sum_{i=1}^n p_{ki} = 1; \sum_{i=1}^n ip_{ki} = C, C \text{ 为常数; } \sum_{i=1}^n ip_{ki} = C, \\
 &C \text{ 为常数)} \\
 &= s [\sum_{i=1}^n i(p_{ki} - p_{ki})] \tag{7}
 \end{aligned}$$

设 $d = (p_{ki} - p_{ki})$, 从上面的推导结果可以发现, 加权平均平均仍然取决于两个因素: 一个是数据片 s ; 另一个是真实状态转移概率和统计转移概率的偏差 d . 当数据序列统计概率接近真实概率时, 如果上一时刻状态已知, 则下一时刻状态的转移将服从稳定概率分布, 那么, 可以近似地认为该情况服从稳定连续概率分布.

状态转移概率统计模块主要对预处理模块得出的状态序列进行状态转移的频度统计, 其统计的目的在于得出状态转移概率矩阵用于后面的预测运算. 统计是按状态序列的顺序进行的, 从 p 个状态中提取出 $p-1$ 次状态转移信息, 并按起始状态和目的状态的序号在转移状态概率矩阵的相应位置进行概率统计.

3.3 预测模块

预测模块主要有两个部分, 概率预测部分和数值预测部分.

概率预测部分这一部分是基于条件概率的相关理论和马尔可夫链的性质, 它的准确程度只取决于状态转移概率统计的准确程度. 如果数值域被划分为 n 个子区间, 则状态转移概率矩阵 A 将会是 $n \times n$ 规模的矩阵相乘, 其时间复杂度显然为 $O(mn^3)$, 若 m 的值比较大, 则运算时间会大大增加. 但是, 由于参与连乘的矩阵都是完全相同的, 于是可以简化这一矩阵连乘过程. 先假设 m 为偶数, 采用二分法进行分割^[7] 如图所示:

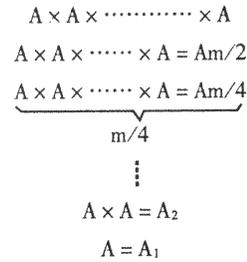


图 1 矩阵链的二分过程示意图

图 1 中, 当 m 为偶数时矩阵连乘次数为 $\log_2 m$ 次, 如果为奇数, m 减 1 成为偶数, 二分过程进行 $\log_2(m-1)$ 次, 矩阵连乘的次数为 $\log_2(m-1) + 1$ 次. 所以, 得到 m 步状态转移概率矩阵的时间复杂度为 $O(n^3 \log_2 m)$. 应用 strassen 矩阵相乘法^[8], 还可以将其复杂度达到 $O(n^{\log_2 7} \log_2 m)$.

数值预测部分, 是从概率预测到数值预测的重要一步, 若有数据序列 ζ 为 $\zeta_1, \zeta_2, \dots, \zeta_p$ 将其数值域划分为 $[x_0, x_1), [x_1, x_2), \dots, [x_{n-1}, x_n]$, 设状态空间 $E = \{E_1, E_2, \dots, E_n\}$, $E_1 \xrightarrow{\Delta} [x_0, x_1), E_2 \xrightarrow{\Delta} [x_1, x_2), \dots, E_n \xrightarrow{\Delta} [x_{n-1}, x_n)$ 如果数据序列 ζ 对应于该状态转移矩阵, 从而预测的真实值 x 处于某一个数值子区间 $[x_{k-1}, x_k)$, 由于处于不完备统计的情况下, 状态转移矩阵中的相应概率为 p_{ki} , 它为 p_{ki} 的估计值, 则可以根据公式: $x = \sum_{i=1}^n p_{ki} (x_{i-1} + x_i)/2$ 来计算预测值, n 表示原始序列数值域划分的子区间数.

4 总结

马尔可夫作为重要的数学工具已经应用在很多领域, 应用 GIS 支持下的马尔可夫模型从空间上进行利用格局变化的模拟是适宜的, 是一种量化的实用方法. 为土地利用的规划、管理及决策提供了良好的研究手段. 转移概率的确定是马尔可夫过程模拟的重要环节, 因此用于确定转移概率的基础数据就准确可靠. 同时, 对于受人为干扰较大的地类, 可根据实际情况对其转移概率做一定调整, 使模拟结果更加合理.

参考文献:

[1] 孙九林. 广泛开展资源信息科学研究和应用[J]. 自然资源学报, 2000, 15 (1): 11- 16.
 [2] Lam bin E F, Baulies X, Bockstaeln N, et al. Land use and

land cover change (LUCC) implementation strategy [R]. IGBP Report No. 48 and HDP Report No. 10. Stockholm: IGBP, 1999.

- [3] 徐建华. 现代地理学中的数学方法(第2版)[M]. 北京: 高等教育出版社, 2002.
- [4] Longley P A, Brooks S M, McDonnell R, Macmillan W. Geocomputation: a primer. John Wiley, 1998.
- [5] 常欣, 程序, 邱化蛟, 等. 计算机模拟模型在黄土丘

陵区土地可持续利用中的实证研究[J]. 农业工程学报, 2003, 19(4): 295—298.

- [6] 岳朝龙, 等. 股票价格的灰色—马尔可夫预测[J]. 系统工程, 1999(6): 55—59.
- [7] 邹海明, 余祥宣, 崔国华. 计算机算法基础(第2版)[M]. 武汉: 华中科技大学出版社, 2000.
- [8] 王晓东. 算法设计与分析(第1版)[M]. 北京: 清华大学出版社, 2003.

Research of Geographical Information System Based on Mathematical Model

LIU Yan-li, LIU Jue-fu, XIE Xin

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: The main contents of this paper are derived from the project “Jiangxi Country Resource Digitized Research”. This paper expatiates the method and signification of combining Markov model with geographical information system, and discourses upon the theory basis of Markov data estimation model based on the mechanism of numerical division and realizing plan. It is proved by theoretical reasoning that Markov data estimation model has comparatively precise result on estimating data with acute fluctuation especially.

Key words: geographical information system; Markov model; numerical division

(上接第33页)

5 结束语

工程质量监督机构通过检测报告监管系统, 可以在检测实验室的市场竞争中起到强有力的监督管理作用, 规范整个检测市场. 检测实验室通过使用检测报告监管系统, 大大提高实验室的管理能力和检测水平, 提高工作效率, 全面提升实验室质量

管理水平, 有效增强其参与市场竞争的能力.

参考文献:

- [1] 检测和校准实验室能力的通用要求. ISO/IEC17025:1999 (idt GB/T15481—2000)
- [2] 建设部. 2003—2008年全国建筑业信息化发展规划纲要(建质[2003]217号文)
- [3] 建设部. 建设事业信息化“十五”计划(建科[2003]239号文)

The Building and Application of System of Supervision and Management for Test Report in Test Lab for Engineering Quality

YU Jian-guo, WAN Zhi-feng

(Total Department of Jiangxi Construction Engineering Quality Supervision, Nanchang 330077, China)

Abstract: This paper introduces the concept of system of supervision and management for test report in construction test lab, and probes into its development situation. then lays emphasis on its basic designing thoughts, systematic structure, and technology to carry out the system.

Key words: lab; test report; system of supervision and management