文章编号：0253-374X(2010)04-0608-05 DOI:10.3969/j.issn.0253-374x.2010.04.026

基于偏最小二乘地理元胞模型的城市生长模拟

冯永玖[1,2]，童小华[1]，刘妙龙[1]

(1.同济大学 测量与国土信息工程系，上海 200092；2.上海海洋大学 海洋科学学院，上海 201306)

摘要：提出了一种基于偏最小二乘回归(PLS)方法的地理元胞(cellular automata，CA)模型 PLS-CA，并用来模拟城市生长和扩展.CA 模型的定义涉及存在严重相关性的众多空间变量，而传统的多准则判别技术(MCE)和主成分分析(PCA)不能够彻底地解决变量相关性问题.利用偏最小二乘回归从空间变量中提取线性无关的主成分，从而获取地理元胞自动机(CA)的转换规则，在地理信息系统(GIS)环境下建立 PLS-CA 模型，可以优化城市生长和扩展的模拟.利用提出的 PLS-CA 模型，模拟了上海市嘉定区 1989 年与 2006 年城市生长和扩展情况.

关键词：城市生长模拟；元胞自动机；偏最小二乘回归；地理信息系统

中图分类号：TP 79 **文献标识码**：A

Modelling Urban Growth Based on Geographical Cellular Automata with Partial Least Squares Regression

FENG Yongjiu[1,2], *TONG Xiaohua*[1], *LIU Miaolong*[1]

(1. Department of Surveying and Geo-informatics, Tongji University, Shanghai 200092, China; 2. College of Marine Sciences, Shanghai Ocean University, Shanghai 201306, China)

Abstract: Based on partial least squares regression, a novel geographical cellular automata model (PLS-CA) is proposed for simulating urban growth and expansion. In definition of cellular automata (CA) transition rules, numerous highly correlated independent spatial variables are utilized for obtaining more actual simulation results. Conventional methods, such as multi-criteria evaluation(MCE) and principal component analysis (PCA), have difficulties in removing the harmful effects of correlation. Using partial least squares regression (PLS) integrated with CA and geographical information system (GIS), a new CA model is created for optimizing the simulation of urban growth and expansion. The PLS-CA model has been successfully applied to simulating urban growth of Jiading district, Shanghai from 1989 to 2006. And the simulation results show that the accuracy of PLS-CA is higher than that of conventional CA models.

Key words: urban growth simulation; cellular automata; partial least squares regression; geographical information system

元胞自动机(cellular automata，CA)由美国数学家 Ulam 于 20 世纪 40 年代提出，是一种时间和空间都离散的动力学系统[1]，用于模拟和分析集合空间内的各种现象[2].近年来，元胞自动机越来越多地应用于城市模拟领域，新的 CA 模型不断被提出，如 DEUM 模型[3]、SLEUTH 模型[4]、MCE 模型[5]、主成分分析模型[6]、模糊逻辑模型[7]、神经网络模型[8]、蚁群智能模型[9]等.这些地理元胞模型对于探索城市生长机理、评价城市土地利用规划方案以及进行城市宏观控制具有理论价值和实际意义.

地理 CA 模型的核心是确定各种影响因素在转换规则中的作用.影响城市生长的因素非常多，并且这些因素通常存在严重的相关性，如何消除相关性并确定模型参数，是 CA 转换规则的首要问题.Wu 等用多准则判断技术(MCE)和 Logistic 回归来解决空间变量的冲突[5]，但是当变量相当复杂且存在严重的相关性时，这些方法并不适合[8].Li 等提出使用主成分分析方法(PCA)来消除空间变量的多重相关性[6]，但是 PCA 方法解析出来的主成分并没有和因变量形成关联，因此无法肯定这些成分对因变量具有最佳解释[10].

收稿日期：2008-12-24

基金项目：国家自然科学基金资助项目(40771174)；教育部新世纪优秀人才计划资助项目(NCET-06-0381)；教育部博士点基金资助项目(20070247046)；上海高校选拔培养优秀青年教师科研专项基金资助项目(ssc09018)

作者简介：冯永玖(1981—)，男，工学博士，主要研究方向为遥感与 GIS、地学信息模型.E-mail：yjfeng@shou.edu.cn

童小华(1971—)，男，教授，博士生导师，工学博士，主要研究方向为遥感与空间数据处理.E-mail：xhtong@tongji.edu.cn

刘妙龙(1944—)，男，教授，博士生导师，主要研究方向为 GIS 理论、方法与应用.E-mail：liuml@tongji.edu.cn

近年来，偏最小二乘回归（partial least squares regression，PLS）在消除变量相关性研究中应用较多. PLS 不仅可以从多重共线的变量中提取线性无关的主成分，而且所提取的主成分能够很好地解释因变量，因此非常适合于获取地理 CA 转换规则. 笔者利用 PLS 获取 CA 模型参数和转换规则，从而建立 PLS-CA 模型，并以上海市嘉定区 1989～2006 年城市生长模拟为例，验证了提出的 PLS-CA 模型.

1　偏最小二乘地理元胞模型（PLS-CA）

1.1　PLS-CA 模型

通常，一般的 CA 模型可以表达归纳为[5]

$$P_{t,c} = \frac{1}{1+\exp(-z_{ij})}\mathrm{con}(S_{t,ij}=S)\cdot \frac{\sum_{3\times 3}\mathrm{con}(S_{ij}=U)}{3\times 3-1}(1+(-\ln\gamma)^{\beta}) \tag{1}$$

式中：$P_{t,c}$为城市生长的联合概率；$\frac{1}{1+\exp(-z_{ij})}$是元胞的局部概率，$z_{ij}=\alpha_0+\alpha_1 x_1+\cdots+\alpha_i x_i+\cdots+\alpha_p x_p$（$i=1,2,\cdots,p$），$\alpha_i$ 是空间变量的权重参数，x_i 是影响城市生长的空间变量；$\mathrm{con}(S_{t,ij}=S)$是元胞状态转换的限制函数，可取值为零或 1，$S_{t,ij}$为第 i 行第 j 列的元胞；S 代表土地元胞适合于开发为城市；$\frac{\sum_{3\times 3}\mathrm{con}(S_{ij}=U)}{3\times 3-1}$表示元胞的 Moore 邻域；$1+(-\ln\gamma)^{\beta}$ 是 CA 模型的随机因子，γ 为值在（0，1）范围内的随机数，β 为控制随机变量影响大小的参数，取值范围为 0～10 之间的整数. 将该联合概率与元胞转换阈值比较，可以判断非城市元胞是否可以向城市元胞转变.

现有的 CA 模型中，式（1）中 z_{ij}的权重参数通常是利用 MCE，Logistic 或 PCA 等方法确定，但是这几种方法无法消除变量相关性的影响，以至确定各个因素的权重 α_i 将很困难[5-6,10]. PCA 虽然在一定程度上能够消除数据冗余、提取能够解释自变量的主成分，但是这些主成分没有与因变量形成关联，因此很难保证 PCA 所提取的主成分能够很好地解释因变量. 研究表明，引入 PLS 可以解决此问题. PLS 所分解出来的主成分不仅能表达影响因素，而且能更好地解释因变量（城市生长的转换概率）. 将 PLS 提取出来的主成分用于 CA 模拟中，可以摆脱 MCE 权重不合理性、纠正 PCA 主成分对因变量的解释乏力的弊端[10]，从而可以更广泛地使用空间变量，达到改善模拟效果的目的.

设 $\boldsymbol{Y}$ 为单一因变量集合（在此为转换概率），各种空间变量的集合为 $\boldsymbol{X}=[x_1,x_2,\cdots,x_i,\cdots,x_n]$. 记 F_0 是因变量 y 的标准化变量，有

$$F_{0i} = \frac{y_i-\overline{y}}{s_y} \quad i=1,2,\cdots,n \tag{2}$$

式中：$\overline{y}$ 是 y 的均值；s_y 是 y 的标准差. 而记 E_0 为自变量集合 X 的标准差.

按照偏最小二乘回归方法及其分析步骤，可知变量空间的第 h 个主成分有

$$\begin{cases} w_h = \dfrac{\boldsymbol{E}_{h-1}^{\mathrm{T}}F_0}{\|\boldsymbol{E}_{h-1}^{\mathrm{T}}F_0\|} \\ t_h = \boldsymbol{E}_{h-1}^{\mathrm{T}}w_h \\ p_h = \dfrac{\boldsymbol{E}_{h-1}^{\mathrm{T}}t_h}{\|t^h\|^2} \\ \boldsymbol{E}_h = \boldsymbol{E}_{h-1}-t_h\boldsymbol{p}_h^{\mathrm{T}} \end{cases} \tag{3}$$

式中：w_h 为 PLS 的轴；t_h 为成分；p_h 为回归系数；$\boldsymbol{E}_h$ 为残差向量；这时得到 m 个成分 $t_1,\cdots,t_m$，求 F_0 在 $t_1,\cdots,t_m$ 上的回归 $\hat{F}_0$，得到

$$\hat{F}_0 = r_1t_1+\cdots+r_mt_m \tag{4}$$

且存在 $t_h=E_{h-1}w_h=E_0w_h^*$，所以 $\hat{F}_0$ 可以写成 E_0 的形式，即

$$\hat{F}_0 = r_1\boldsymbol{E}_0w_1^*+\cdots+r_m\boldsymbol{E}_0w_m^* = \boldsymbol{E}_0\Big[\sum_{h=1}^{m}r_hw_h^*\Big] \tag{5}$$

式中：r_m 为 $\hat{F}_0$ 的回归系数；$w_h^*=\prod_{j=1}^{h-1}(\boldsymbol{I}-w_j\boldsymbol{p}_j^{\mathrm{T}})\cdot w_h$，$\boldsymbol{I}$ 为单位矩阵.

也可以写成回归方程

$$\hat{y} = \alpha_0+\alpha_1x_1+\cdots+\alpha_px_p \tag{6}$$

则，利用 Logistic 回归可以得到[5]

$$P_1 = 1/(1+\exp(-\hat{y})) = 1/(1+\exp(-(\alpha_0+\alpha_1x_1+\cdots+\alpha_px_p))) \tag{7}$$

式中：P_1 是区域空间变量作用下的元胞转换概率. 将 P_1 替换式（1）的$\frac{1}{1+\exp(-z_{ij})}$，则可得到 PLS-CA 模型的城市生长联合概率 $P_{t,c}$.

$$P_{t,c} = 1/(1+\exp(-(\alpha_0+\alpha_0x_1+\cdots+\alpha_px_p)))\cdot \mathrm{con}(S_{t,ij}=S)\frac{\sum_{3\times 3}\mathrm{con}(S_{ij}=U)}{3\times 3-1}(1+(-\ln\gamma)^{\beta}) \tag{8}$$

则,非城市元胞是否转换为城市元胞的最终判断条件为

$$S_{t+1,ij} = \begin{cases} 1 & P_{t,c} > P_{threshold} \\ 0 & P_{t,c} \leqslant P_{threshold} \end{cases} \quad (9)$$

式中:1 代表城市;0 代表非城市.从式(9)可知,当 $P_{t,c}$ 大于阈值 $P_{threshod}$ 时,该元胞转化成为城市用地,当小于该阈值时,元胞保持原有状态不变.

1.2 PLS-CA 模型模拟步骤

本模型在 VS.NET 环境下利用 ArcGIS Engine 组件编写而成.该框架可以集成其他已有的 CA 模型,同时在此框架下可以开发更多新的 CA 模型.

利用 PLS-CA 模型进行城市生长模拟的步骤为:①利用模拟框架随机产生 10% 的样本点,通过 ArcGIS Spatial Analyst 分层获取所需空间变量;②通过 PLS 对空间变量进行分析,提取相应主成分,利用符合条件的主成分预测因变量值;③对因变量进行 Logistic 变换,得到局部概率值 P_1;④加入城市或地理发展的其他因素,如邻居、限制条件、随机因子等,结合前一项转化为联合发展概率 P_c^t;⑤进入 SimUrban 框架进行模拟,评价模拟结果,输出符合要求的结果.

2 模型应用及结果分析

2.1 研究区域及 PLS 数据处理

选择城市化进程较快的上海近郊区嘉定区作为模型的试验区域.嘉定区位于上海市西北部,与江苏省昆山市毗连,属于上海城市的近郊区,近 10 余年正经历快速城市化过程.

影响城市生长的空间因素较多,且互相之间存在严重的相关性.用偏最小二乘法回归方法提取变量主成分,能最好地解释因变量,这对于提高模型的性能非常重要.

首先,从遥感影像和 GIS 数据提取 CA 模型所需的相应空间变量.遥感影像为 1989 年和 2006 年 2 个时相的影像,通过 2 个年份的变化对比来获取其转化的概率,而通过 1989 年遥感影像获取空间变量.获取空间变量如下:因变量(转换概率)y;自变量 5 个,分别为:到市中心的距离 D_{uc}、到镇中心的距离 D_{tc}、到主要道路的距离 D_{mr}、到耕地的距离 D_{pl}、到菜园的距离 D_{ky}.此外,模拟中动态获取 Moore 3×3 邻域的元胞数量和发展限制条件函数的值.

在模拟之前,通过模拟框架获取 1 000 个样本点.对这一系列变量值形成的变量矩阵作偏最小二乘回归,提取能最好地解释因变量的各个成分.利用主成分分析和偏最小二乘回归,对嘉定区随机选择的样本点进行分析,所获取的主成分分别如表 1 和表 2.

对于 PCA,第 1 主成分主要表示以道路为主,体现道路对城市生长的"推动力"作用;第 2 主成分主要表示以市中心为主,体现市中心对城市生长的"引力"作用;第 3 主成分主要表示非市中心的信息;第 4 主成分则主要表示耕地和园地的信息;第 5 主成分则可以忽略不计.从表 1 可知,主成分分析所获得的成分中,前 4 个成分累计贡献率已达 99.32%,因此只须利用前 4 个成分进行 CA 转换规则的获取.

表 1 利用主成分分析提取的空间变量主成分

Tab.1 Extract principal components of spatial variables using principal component analysis

成分	特征值及贡献率			距离变量				
	特征值	贡献率/%	累计贡献率/%	D_{uc}	D_{tc}	D_{mr}	D_{pl}	D_{ky}
1	3.187	63.737	63.737	0.460 0	0.533 0	0.977 0	0.189 3	−0.387 2
2	0.992	19.847	83.584	0.750 0	0.425 0	0.482 0	−0.239 0	−0.232 0
3	0.503	10.067	93.651	−0.187 0	0.234 0	0.216 3	0.115 5	0.116 4
4	0.283	5.670	99.320	−0.152 0	0.042 0	0.044 5	−0.019 8	−0.010 1
5	0.034	0.680	100.000	0.003 0	0.003 0	−0.002 0	−0.013 0	−0.002 8

表 2 利用偏最小二乘回归提取的空间变量主成分

Tab.2 Extract principal components of spatial variables using partial least squares regression

成分	正交有效性			距离变量				
	R	Q_{fn}^2	临界值	D_{uc}	D_{tc}	D_{mr}	D_{pl}	D_{ky}
1	0.836 1	0.824 0	0.097 5	0.747 8	0.247 0	0.953 2	−0.122 4	−0.092 3
2	0.338 9	0.113 0	0.097 5	0.157 2	0.784 5	0.658 3	−0.106 0	0.044 0
3	0.194 6	−0.007 7	0.097 5	−0.119 6	−0.240 6	0.278 8	−0.379 2	−0.093 0

利用 PLS 对嘉定区样本点进行分析，提取了 3 个主成分：第 1 主成分主要表示以道路为主、市中心和镇中心为辅的信息；第 2 主成分主要表示以镇中心为主、市中心和道路为辅的信息；而第 3 个主成分由于 $Q_h^2 \leqslant 0.097\ 5$，因此不再符合正交有效性的要求，所以只需要利用前面 2 个主成分对因变量进行解释即可，这 2 个主成分的累计贡献率为 93.70%，达到了 PCA 方法前 3 个主成分对因变量的解释能力.

为使获取的参数具有明确的地理意义，在标准化模拟样本数据时，将空间数据进行了反变换，即到地理实体（如市中心、道路）等的距离越小，所得到的计算距离越大（即 D_{uc}，D_{mr} 越大），这样得到的 CA 模型参数就显得比较容易理解，即空间正距离的 CA 参数为正值，而空间负距离的 CA 参数为负值（如表 3）.

表 3　PLS-CA 模型与 PCA－CA 模型参数

Tab.3　CA Parameters obtained by PLS-CA and PCA-CA

CA 模型	D_{uc}	D_{tc}	D_{mr}	D_{pl}	D_{ky}
PLS-CA	0.678 5	0.472 4	1.020 1	－0.138 3	－0.062 3
PCA-CA	0.404 5	0.442 5	0.731 4	0.084 1	－0.277 0

从表 3 可知，利用 PCA-CA 模拟城市生长的 CA 参数 D_{uc}，D_{tc} 和 D_{uc} 在数值上较为均衡，说明市中心、镇中心和道路对城市生长的贡献是近似等同的，但是这与实际情况并不相符. 从嘉定区城市实际发展和模型模拟的情况来看，市中心对城市生长的贡献非常大，且城市从市中心发散状地沿主要道路延伸，因此说明主要道路对城市生长的作用也是非常明显的. 相比之下，镇中心对城市生长的作用并不明显，这可以从模拟结果中得到验证. 利用 PLS-CA 优化后的参数，拉大了市中心和道路的参数与镇中心参数的数值差距，使得 CA 参数更加合理.

利用 PLS 对上海市嘉定区 1989 年及 2006 年城市变化分析可知，对该区域城市生长贡献较大的参数依次为到道路的距离（1.020 1）、到市中心的距离（0.678 5）以及到镇中心的距离（0.472 4），而到农田和菜园的距离参数均为负值. 通过反变换得到的 CA 参数其地理意义显得非常明显.

2.2　模拟研究

利用 PLS-CA 模型，模拟了嘉定区 1989 年、2006 年的城市生长空间模式. 为评定模拟结果，将 1989 年和 2006 年的遥感影像分类结果作为基准，见图 1，“模拟”即 PLS-CA 模拟结果，“实际”则为遥感影像分类结果. 同时，比较了 PCA-CA 和 PLS-CA 模型应用在嘉定区城市生长模拟中的精度. 模拟过程中将城市土地利用分为 3 类：城市、非城市和水体. 城市水体独立设类，视为城市生长的限制因素，即遇到水体元胞时其状态保持不变，实际赋予元胞限制值为零.

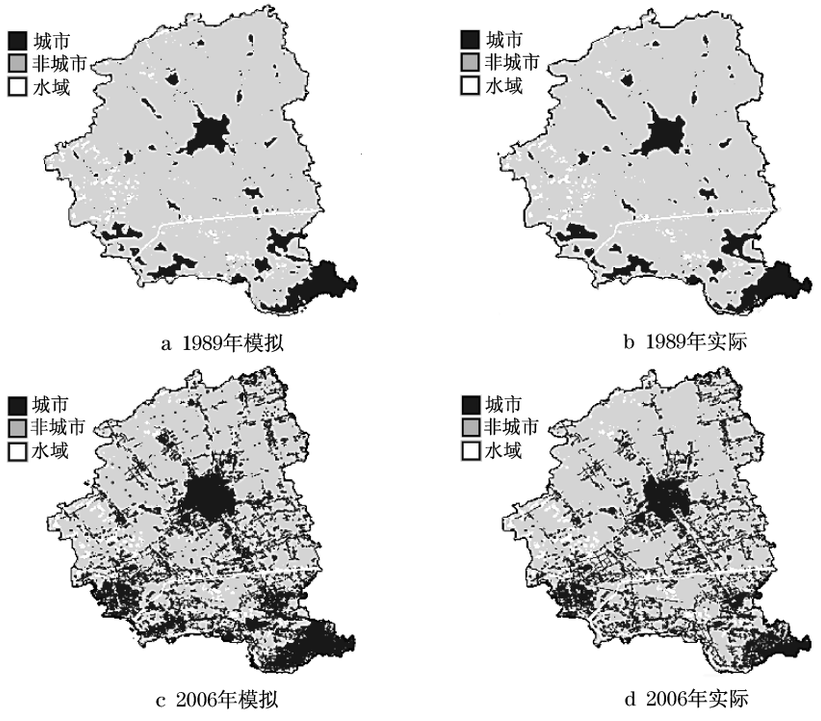

图 1　利用 PLS-CA 模拟上海市嘉定区 1989 年与 2006 年城市生长

Fig.1　Simulating urban growth of Jiading district with PLS-CA in 1989 and 2006

为了检验PLS-CA模型的模拟精度,利用2006年的模拟情况,分别计算了PLS-CA和PCA-CA模拟与基准遥感分类影像的逐点对比精度(如表4).表4表明,非城市元胞类的模拟精度高于城市元胞类的模拟精度.对实际城市生长中土地元胞未转化的情况,PLS-CA模型模拟精度为84.61%,比PCA-CA模型(82.04%)高近2.4%;而对于实际发生了转变的土地元胞,PLS-CA模型(75.49%)比PCA-CA模型(70.23%)高5.2%左右;总体精度上PLS-CA模型(80.37%)比PCA-CA模型(76.55%)高近3.8%.另外,通过Kappa系数来检验模拟结果的一致性,计算结果表明PLS-CA模型模拟结果的Kappa系数为60.36%,比PCA-CA模型的Kappa系数(52.58%)高约7.8%.可见,利用PLS对CA参数进行优化,明显提升了模拟精度和整体一致性,使得模拟结果也更加合理和接近于城市实际发展情况.

表4 PLS-CA与PCA-CA城市生长模拟的混淆矩阵

Tab.4 Confusion matrixes of simulation results with PLS-CA and PCA-CA model

实际分类		模拟结果/元胞		精度/%	总精度/%	Kappa系数/%
		非城市	城市			
PLS-CA	非城市	24 509	4 458	84.61	80.37	60.36
	城市	6 171	19 005	75.49		
PCA-CA	非城市	23 765	5 202	82.04	76.55	52.58
	城市	7 495	17 681	70.23		

3 结语

城市是一个异常复杂的巨系统,影响城市生长的空间因素众多,且因素(变量)之间通常存在严重的相关性.PLS不仅能够方便地提取一组相关变量中独立的空间变量,且能最大限度地解释因变量.结合GIS,PLS和元胞自动机,提出了PLS-CA模型,用于模拟城市生长及形态的扩展.

将该模型应用于上海市嘉定区,利用不同年份的TM遥感影像和GIS数据作为主要的空间数据,利用PLS获取CA模型的转换规则以及参数设置,模拟了该区1989年及2006年的城市生长变化情况.分析表明,模拟结果与该区的城市生长情况非常相符.与PCA-CA的模拟情况作相应比较,结果表明PLS-CA比PCA-CA更接近城市实际发展的空间格局.

针对嘉定区计算得到的模型参数虽然并不能够完全适合于其他区域,但是PLS-CA模型框架却具有较好的普适性.当需要研究新的区域时,根据其初始空间变量,就能够通过PLS回归计算得到相应的CA模型参数,从而可以方便地模拟新的区域.

参考文献:

[1] John von neumann. Theory of self-reproducing automata[M]. Champaign:University of Illinois Press,1966.

[2] 黎夏,叶嘉安,刘小平,等.地理模拟系统——元胞自动机与多智能体[M].北京:科学出版社,2007.
LI Xia,YEH Anthony Garon,LIU Xiaoping,et al. Geographical modelling systems—cellular automata and multi-agent[M]. Beijing:Science Press,2007.

[3] Batty M,XIE Y,SUN Z. Modeling urban dynamics through GIS-based cellular automata[J]. Computers,Environment and Urban System,1999(23):205.

[4] Clarke K C,Gaydos L J,Hoppen S. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area[J]. Environment and Planning B,1997(24):247.

[5] WU Fulong. Calibration of stochastic cellular automata: the application to rural-urban land conversions[J]. International Journal of Geographical Information Science,2002,16(8):795.

[6] LI Xia, YEH A G O. Urban simulation using principal components analysis and cellular automata for land-use planning [J]. Photogrammetric Engineering & Remote Sensing,2002,68(4):341.

[7] LIU Yan. Modelling urban development with geographical information systems and cellular automata[M]. New York:CRC Press,2008.

[8] Almeida C M,Gleriani J M,Castejon E F,et al. Using neural networks and cellular automata for modelling intra-urban land-use dynamics [J]. International Journal of Geographical Information Science,2008,22(9):943.

[9] LIU Xiaoping,LI Xia,LIU Lin,et al. A bottom-up approach to discover transition rules of cellular automata using ant intelligence [J]. International Journal of Geographical Information Science,2008,22(11):1247.

[10] 王惠文,吴载斌,孟洁.偏最小二乘回归的线性与非线性方法[M].北京:国防工业出版社,2006.
WANG Huiwen,WU Zaibin,MENG Jie. Partial least squares regression-linear and nonlinear methods[M]. Beijing:National Defence Industry Press,2006.