基于回归模型与优化算法的样本检验方法探讨

王 平, 唐 旭, 祝国瑞
(武汉大学 资源与环境科学学院, 武汉 430079)

摘要: 阐述了样本地价与土地级别具有相关性的空间特性, 分析了基准地价评估中传统样本检验方法存在的不足, 提出了基于回归分析模型的样本检验方法, 在此基础上设计了局部搜索算法和模拟退火算法以实现样本的合理检验, 并通过实例进行了验证. 文中阐述的样本检验理论与方法对于科学、合理地评估基准地价, 完善地价评估理论具有重要意义.

关键词: 回归分析模型; 模拟退火算法; 基准地价模型
中图分类号: P285.2 文献标识码: A 文章编号: 1007-855X(2004)03-0017-04

Research on the Approach to the Sample Test Based on Regression Model and Optimization Algorithm

WANG Ping, TANG Xu, ZHU Guo-rui
(School of Resource and Environment Science, Wuhan University, Wuhan 430079, China)

Abstract: The characteristics of spatial distribution of sample land-value and land grade are expounded, and the deficiency of the traditional sample test in evaluating urban land benchmark price is analyzed. The new methods of the sample test based on land-value regression model is brought forward, in which local search algorithm and simulation annealing algorithm are given to realize the reasonable test, and some examples are illustrated to prove the methods compared with the traditional sampling inspection. The above mentioned theory and the methods have a great significance in evaluating the land value in a scientific and reasonable way. Finally, some reasonable advice to improve sampling is put forward.

Key words: regression analysis model; simulation annealing algorithm; land-value model

0 引 言

基准地价是国家宏观控制土地市场价格的基础, 是国家征收土地使用税、评估标定地价的依据. 在城镇基准地价评估中, 地价样本则是基准地价的测算依据. 通常, 在同一土地级别或同一均值地域内的同行业中, 由于某些特殊因素影响造成一些样本地价明显高于或低于其它宗地地价, 不能反映当地的实际地价水平; 或者由于同一宗地调查样本数目过多、不同区域交易发生的密度相差较大或外业采点的人为倾向, 会出现地价样本价位以及价格在空间分布上不合理、或者点位的空间分布不均匀等问题. 因此, 样本数据的分析、检验、剔除, 去伪存真、去粗取精, 直接影响着基准地价评定结果. 所以, 研究一种合理、科学的地价样本检验方法对基准地价的合理评估具有十分重要的意义.

1 样本检验与基准地价模型

1.1 样本的空间特征

按照马克思地租地价理论和区位理论, 城镇土地因其相对位置不同, 在空间上表现出不同的使用价值、不同的经济效益和不同的地租地价. 不同级别的土地, 其土地价格也应不同, 级别越高地价也相应越

收稿日期: 2003-12-26.

第一作者简介: 王平(1977~), 男, 在读硕士研究生. 主要研究方向: 土地信息系统. **E-mail:** fly_sky_wp@163.com

高. 所以, 市场交易样本的地价并非完全随机的统计集合, 它与样本在城镇所处土地区位存在着规律性的关系.

1.2 样本子集与基准地价模型

在基准地价评估中, 首先要对外业采集的样本总体进行检验, 剔除异常样本数据, 选取有效的样本子集(有效样本子集的选取过程也就是样本的检验剔除过程). 然后, 根据选取的有效样本子集, 利用最小二乘法建立基准地价评估模型. 因此, 基准地价模型反映着有效样本的地价与土地级别的二维关系. 一个有效样本子集对应着一个基准地价模型, 两者之间是一一映射关系的, 其关系如图1所示. 因此, 基准地价评估模型是否合理也就是有效样本选取合理与否的判定标准.

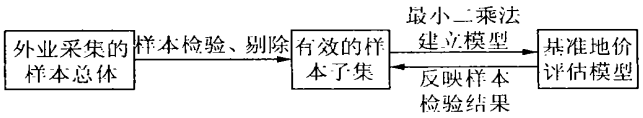

图1 样本子集与基准地价模型关系

2 基于数理统计的传统样本检验方法的不足

按照《城镇土地估价规程》规定, 样本检验应以土地级别或均质地域为单位, 分土地利用类型进行抽样样本的总体和方差检验. 通常, 当样本总体呈正态分布时采用 t 检验法, 样本总体呈非正态分布时采用均值- 方差法.

t 检验法的基本原理是将一个按样本值大小顺序排列的样本集合数列中的最大值样本和最小值样本均视为可疑异常样本, 通过设置的拒绝域来不断的进行检验、剔除, 直到样本集合数列两端的样本均不为异常值为止. 置信度为 $1-\alpha$ 的拒绝域可以表示为:

$$|t| = \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| \geq t_{\alpha/2}(n-1)$$

图2表示某一土地级别的样本检验结果, 其中 μ_1, μ_2 分别表示拒绝域上下限临界状态所对应的样本地价值.

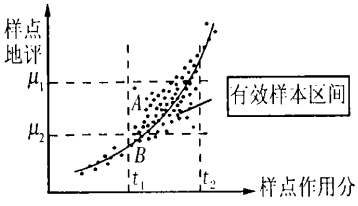

图2 正态分布样本的 t 检验法

均值 —方差法主要通过数量级比较, 剔除偏离地价均值过大的样点, 剔除标准一般取 $\bar{X} \pm 2S$, 其中 $\bar{X}$ 为同级别内同行业的地价均值, S 为标准差. 图3表示某一土地级别的样本检验结果, 其中 $\mu_0 = \bar{X} - 2S$, $\mu_2 = \bar{X} + 2S$.

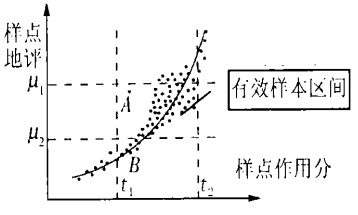

图3 偏态分布样本的均值 - 方差法

从上面的图2和图3中可以看出: 样点 A 作用分较低, 但对应的地价却很高, 应属于异常样点; 样点 B 作用分较低, 但对应的地价却也低, 其作用分与其对应的地价关系合理, 故属于合理样点. 但在传统检验方法中样点 A 没有被剔除, 样点 B 反而被剔除了, 故传统的样本检验方法存在一定的不合理性.

传统的检验方法忽视了样本地价作为空间数据的特点. 数理统计中的样本变量一般具有独立性, 而地价样本并非独立的、完全随机的统计集合, 它与样本在城镇所处的土地区位存在着规律性的关系(一般情况下级别越高的土地其地价也相应会很高). 因此, 在进行样本检验时, 应把地价和土地级别联系起来. 否则在进行样本检验剔除时, 会出现正常样本被剔除或异常样本保留的现象, 从而影响到基准地价评估结果的准确性.

3 基于回归模型的地价样本检验理论

基于回归模型的样本检验剔除方法是建立在回归分析理论基础上的, 它是通过回归方程建立样本地价与因素作用分之间的关系, 即建立地价模型曲线方程 $F(x)$. 基准地价模型方程 $F(x)$ 可以是线性方程 $y = \alpha + \beta x$ 或指数方程 $y = \alpha(1+\beta)^x$, 指数方程通常可以转化为线性方程求解, 回归方程中的表示样本作用分值, y 表示样本地价.

考虑到地价样本的空间特征, 如图4(假设地价模型为指数模型) 所示, 可以将样本检验剔除与回归模

型建立结合, 通过建立的基准地价模型, 确定有效样本的地价偏差范围, 分析离散样点相对于模型曲线偏离程度, 就可以判定该样本是否是合格样点. 这样在进行样本检验剔除时, 就可以充分考虑样本地价和因素作用分值的关系, 从整体上保持地价样本分布的空间相关性和分布连续性.

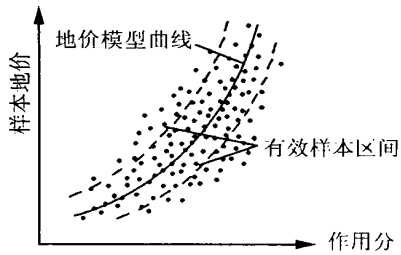

图 4　基于地价模型的样本检验

4 智能优化算法在样本检验中的应用

4.1 基于局部搜索算法的样本检验方法实现

局部搜索算法主要是基于贪婪思想利用邻域函数来进行搜索的, 它通常可描述为: 从一个初始解出发, 利用邻域函数持续的在当前解的邻域重搜索比它好的解, 若能够找到如此的解, 就以之成为新的当前解, 然后重复上述过程, 否则结束搜索过程, 并以当前解作为最终解. 可见, 局部搜索算法尽管具有通用易实现的特点, 但搜索性能完全依赖于邻域函数和初始解. 邻域函数设计不当或初值选取不合适, 则算法最终的性能将会很差[1].

因此, 在利用局部搜索算法进行样本检验剔除的过程中, 如何设计邻域函数和选取初始解是解决问题关键所在. 下面就这几个关键问题进行说明:

1) 初始解的选取. 首先对样本总体进行粗差剔除, 即剔除那些诸如空间点位错误、土地收益为负数或土地价格低于成本等错误样点, 然后在剩下的有效样本总体 S 中来确定初始样本子集, 从而确定初始解. 初始解的选取通常有两种方法: 一种是直接以有效样本总体 S 为初始解, 即算法开始时选取所有的样本; 另一种方法与前者恰恰相反的, 它考虑到了样本的总体分布趋势, 先求取各级别内样本地价均值和样本作用分均值, 以这些各级别的均值样点集合为初始解. 这里我们采用后者来确定初始解.

2) 邻域函数的设计. 根据初始解(初始有效样本子集) 建立的地价回归方程为 $\hat{y} = \hat{\alpha} + \hat{\beta}x$, 对有效样本总体 S 中的每个样点, 求出对应的样本地价与其估计量的偏差: $d_i = y_i - \hat{\alpha} - \hat{\beta}x_i$. 由此得到地价偏差集合 $U = \{d_1, d_2, \Lambda, d_n\}$, 记 σ 为集合 U 的方差.

则邻域函数可以设定为:

$$\begin{cases} f_1(x) = \hat{y} + 2\sigma \\ f_2(x) = \hat{y} - 2\sigma \end{cases}$$

邻域函数曲线与回归模型曲线关系如图 5 所示.

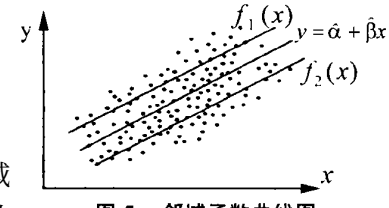

图 5　邻域函数曲线图

3) 算法的实现过程. 根据初始解建立地价模型, 利用设计的邻域函数, 选取落在地价模型曲线的上下临界区间内的样点, 作为新的有效样本子集 S' (新解). 根据样本子集 S' 可以建立新的地价模型, 再对有效样本总体进行检验, 又可选取新的样本子集, 再产生新解. 这个过程重复进行, 当样本子集 S' 的有效样本数目及其对应的地价模型的复相关系数、回归平方和、偏差平方和等指标达到设定值时, 算法结束.

4.2 基于模拟退火算法的样本检验方法实现

模拟退火算法(Simulated Annealing, 简称 SA) 是一种智能优化算法, 常用于解决大规模组合优化问题. 它是基于Monte Carlo 迭代求解策略的一种随机寻优算法, 其出发点是模拟固体退火过程中, 其自由度随温度的降低而不断减小, 能量最终达到极小值– 固体达到基态, 从而寻找目标函数的极小值. SA 算法在解空间中进行随机搜索时, 通过应用 Metropolis 抽样策略, 能使算法避免陷入局部最小, 从而最终得到问题的全局最优解.

基于上述模拟退火算法的基本思想, 在地价样本检验的实现过程中几个关键问题的设计如下:

1) 目标函数与初始解的确定. 已知地价模型曲线方程 $F(x)$ 为线性回归方程(指数方程可转化为线性方程求解), 记 $\theta = (\hat{\alpha}, \hat{\beta})$ 为回归方程系数的估计量, R^2 为线性回归方程的复相关系数, 则目标函数可以由地价模型回归显著性检验中的相关参数来表示: $C(\theta) = a(1 - R^2)$, 其中为 a 常数, $R^2 = S_{回}/S_{总} = \sum(\hat{y} - \bar{y})^2 / \sum(y_i - \bar{y})^2$, 而初始解则可以选定为经过粗差剔除后的有效样本集合.

2) 初温 t_0 和温度更新函数的确定. 初始温度 t_0 可以依据函数 $t_0 = -\Delta_{max}/\ln p_r$ 来确定, 其中 Δ_{max} 为随机产生一组状态中, 两两状态间的最大目标值差的绝对值, p_r 为初始接受概率(理论上应接近于1). 此处, 由目标函数 $C(\theta)$ 可求得 $|\Delta_{max}| = a$, 为提高算法优化效率, p_r 可定为1/3, 表示应用Metropolis准则接受较差解的状态转移概率阈值为1/3.

温度更新函数可选用目前最常用的指数退温函数 $t_{k+1} = \lambda t_k$, 其中 λ 为退温速率. 文中采用线性下降方法进行降温处理, 这里取 $\lambda = 0.9$.

3) 目标函数状态值的改变. 将有效样本总体以样本作用分值的高低进行排序, 并以相等的分值间隔进行分组, 在各组内随机选取一定比例(保证样本子集反应总体分布)样点, 组成一个新的样本子集, 该样本子集即为产生的新解. 新的样本子集对应新的地价模型, 所以, 产生新解后, 回归方程 $F(x)$ 复相关系数改变, 目标函数状态值发生改变.

4) 循环终止准则的确定. 若在某一在温度 t_k 下, 样本集合(回归方程)连续 N_1 次保持不变, 则退出内循环; 若在连续 N_2 次退温过程中, 所选取的样本集合(回归方程)保持不变, 或者目标函数 $C(\theta_i) < \varepsilon$, 两者满足其一则退出外循环, 即算法结束.

算法的流程图如图6所示.

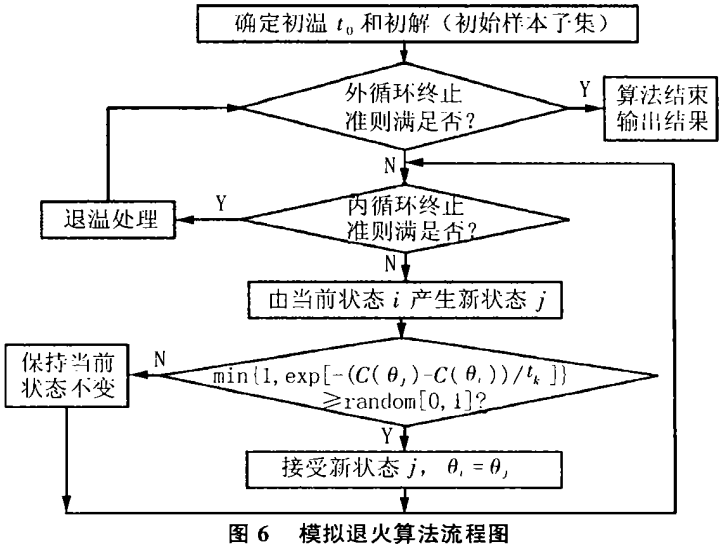

图6 模拟退火算法流程图

5 实例应用与分析

现有某市商业用地样本726个, 土地级别为4级, 样本具体情况见表1.

表1 某市商业用地样本数据情况表

土地	样点	样本均价	样本地价范围/元·m^{-2}	级别作用分区间	级别平均分
一级	130	5 577.58	(167 ~ 20 717)	(62.16 ~ 80.85)	69.89
二级	362	4 068.78	(150.84 ~ 45 815)	(47.09 ~ 62.16)	54.67
三级	200	2 632.10	(43.111 ~ 12 943)	(29.08 ~ 47.09)	39.38
四级	34	1 500.22	(4.14 ~ 4 636)	(5.01 ~ 29.08)	24.41

据前所述, 分别采用传统样本检验方法、基于局部搜索算法的样本检验方法和基于模拟退火算法的样本检验方法, 对样本数据进行检验剔除, 并求取该商业用地基准地价, 所得的结果见表2.

表2 样点剔除及基准地价评估结果

检验方法	剔除点数	各级别地价 一级	二级	三级	四级	地价和作用分值模型	复相关系数 R^2
传统方法	95	4 921	2 873	1 673	985	$Y = 415.524^{*}(1+0.036)^{x}$	0.461
局部搜索算法	141	4 713	2 633	1 467	827	$Y = 325.096^{*}(1+0.039)^{x}$	0.508
模拟退火算法	118	5 361	2 743	1 399	724	$Y = 247.278^{*}(1+0.045)^{x}$	0.601

通过对传统的样本检验方法和基于回归模型与优化算法的样本检验方法的实例比较可以得出以下结论:

1) 样本的地价不是简单的、完全的随机样本, 传统的检验方法只是从数学的角度对样本进行检验剔除, 从统计方法、检验结果都不是十分合理.

2) 基于回归模型与优化算法的检验方法从地价和土地级别及因素作用分的关系考虑, 将那些满足于数理统计, 不满足地价规律(模型)的样点剔除掉, 样本更合理.

(下转第28页)

翡翠的历史.但我们利用拉曼光谱测定出石斧的成分不是硬玉,而是蓝晶石.

2) 所谓"清朝孝庄皇后墓中的翡翠"只不过是现代商人为追逐暴利,而通过高科技手段作假注入环氧树脂的翡翠B货,根本不可能是清朝时期的翡翠制品.

3) 通过对上述文物的鉴别,我们发现拉曼光谱技术非常适合于易损和不允许取样的珍贵艺术品的无损分析,可以为文物的真伪鉴别提供依据.同时可以预计,拉曼光谱分析技术必将在我国珍贵文物的研究中发挥重要的作用.

致谢 感谢云南省腾冲县文物局提供的石斧样品,同时感谢昆明理工大学材料学院光电新材料研究所张鹏翔教授的大力支持和帮助.

参考文献:

[1] 高大伦. 玉器鉴赏[M]. 桂林: 漓江出版社, 1995. 1~12.
[2] 王寒竹. 玉石鉴赏与商贸指南[M]. 武汉: 中国地质大学出版社, 1997. 24~25.
[3] 马宝忠. 云南珠宝王国[M]. 昆明: 云南科技出版社, 1999. 5~6.
[4] Lowell I. Mccann, K. Trentelman, T. Possley and B. Golding. Corrosion of Ancient Chinese Bronze Trees Studied by Raman Microscopy[J]. Journal of Raman Spectroscopy, 1999, 30(2): 121~132.
[5] Edwards H G M, Farwell D W, Heron C P, Croft H, David A R. Cat's Eyes in a New Light: Fourier Transform Raman Spectroscopic and Gas Chromatographic-Mass Spectrometric Study of Egyptian Mummies[J]. Journal of Raman Spectroscopy, 1999, 30(3): 139~146.
[6] Philippe Colomban, Francoise Treppoz. Identification and Differentiation of Ancient and Modern European Porcelains by Raman Macro- and Micro- Spectroscopy[J]. Journal of Raman Spectroscopy, 2001, 32(2): 93~102.
[7] Peter Vandenabeele, Francis Verpoort, and Luc Moens. Non-destructive Analysis of Painting Using Fourier Transform Raman Spectroscopy with Fibre Optics[J]. Journal of Raman Spectroscopy, 2001, 32(4): 263~269.
[8] 左健, 许存义. 古壁画、陶彩颜料的拉曼光谱分析[J]. 光散射学报, 1999, 11(3): 215~219.
[9] 祖恩东, 陈大鹏. 翡翠B货的拉曼光谱鉴别[J]. 光谱学与光谱分析, 2003, 23(1): 64~66.

(上接第20页)

3) 从地价模型的复相关系数可以看出,利用基于优化算法的样本检验剔除结果进行基准地价评估,其评估结果比利用传统的样本检验剔除结果进行的基准地价评估结果更合理.

4) 对于局部搜索算法和模拟退火算法,从样本的检验剔除结果和基准地价评估结果来看,后者优于前者;从算法的时间效率来看,前者优于后者.

6 小结

地价样本不是简单的、完全的随机样本,它往往和所处的土地级别有关.文中提出的基于回归模型与优化算法的的样本检验方法,充分的考虑了样本的空间分布特性,其检验、剔除结果比传统的样本检验方法更加科学、合理,在城镇基准评估工作中具有重要意义.

参考文献:

[1] 王凌. 智能优化算法及其应用[M]. 北京: 清华大学出版社, 2001. 17~29.
[2] 国土资源部土地估价师资格考试委员会. 土地估价理论与方法[M]. 北京: 地质出版社, 2000. 12~24.
[3] Damodar N. Gujarati. 计量经济学[M]. 北京: 中国人民大学出版社, 2000. 67~128.
[4] 唐旭. 基于基准地价评估模型的修正体系编制方法研究[J]. 中国土地科学, 2002. 34~38.
[5] 陈幼松, 杨位钦. 实用数理统计方法及应用详解[M]. 北京: 北京科学技术出版, 1998. 83~134.
[6] 王新生, 姜友华等. 模拟退火算法及其在非线性地学模型参数估计中的应用[J]. 华中师范大学学报(自然科学版), 2001, (3): 103~106.
[7] 祝国瑞. 模拟退火算法在动态建立基准地价模型中的应用[J]. 武汉大学学报(信息科学版), 2003, (10): 593~595.