文章编号:1671-8860(2007)02-0115-05　　**文献标志码**:A

混合地理加权回归模型算法研究

覃文忠[1]　王建梅[1]　刘妙龙[1]

(1　同济大学测量与国土信息工程系,上海市四平路 1239 号,200092)

摘　要:以迭代算法为基础,推导出混合地理加权回归模型的常系数(全局参数)和变系数(局域参数)的计算方法,并以上海市住宅小区楼盘销售平均价格为例进行验证。结果表明,混合地理加权回归模型的计算量略大于地理加权回归模型,但对样本数据的拟合更好,局域参数估计更稳健。

关键词:地理加权回归模型;混合地理加权回归模型;空间非平稳性;迭代算法;空间分析

中图法分类号:P208

在空间数据分析中,一般线性回归模型(ordinary linear regression,OLR)因其具有完备的理论体系和统计推断方法,常用来确定和分析变量之间的关系,有着非常广泛的应用。考虑到空间数据的空间非平稳性,OLR 模型的分析结果不能全面反映空间数据的真实特征,尤其是数据随空间区域的变化规律,为此有学者提出了空间变系数回归模型(spatially varying-coefficient regression model)[1],Fotheringham 等称之为地理加权回归模型(geographically weighted regression model,GWR)[2,3]。OLR 模型假设回归参数在空间上是不变的,而 GWR 模型允许回归参数随着地理空间的变化而变化。但在有些情况下,并不是所有参数都随地理空间变化而变化,有些参数在空间上是不变的,或者其变化非常小,可以忽略不计[4,5]。例如,预测城市房地产价格时,与房地产自身相关因素(如建筑结构、建筑质量等)及其所在区位相关因素(如交通状况、公共设施等)的影响力在空间上是变化的,而对房地产价格产生影响的社会经济因素(如失业水平等)在整个研究区域的影响力基本是一致的。因此,实际问题分析中应采用改进方案,即回归模型中的部分参数随空间位置改变而变化,其余参数为常数。这种新的回归模型,有的学者称之为半参数空间变系数回归模型[1],Brunsdon 等称为混合地理加权回归模型(mixed GWR model, MGWR)[6,7]。本文采用后一种名称。

地理加权回归模型中的参数皆为局域参数,而混合地理加权回归模型中的参数部分为全局参数,部分为局域参数,因此地理加权回归模型的参数估计方法不能直接用来估计混合地理加权回归模型的参数。对于如何解算混合地理加权回归模型问题,魏传华等提出了一种两步估计方法[1],并进行模拟试验。Brunsdon 等应用迭代算法得到常值系数的近似估计[6]。笔者以迭代算法为基础,进行严密的理论推导,得到 MGWR 模型的常系数和变系数估计的精确表达式,并给出了具体计算过程,最后以上海市住宅小区楼盘销售平均价格为例进行验证分析。

1　理论方法

1.1　地理加权回归模型

一般线性回归模型公式如下:

$$y_i = a_0 + \sum_{k=1}^{n} a_k x_{ik} + \theta_i \tag{1}$$

式中,$x_k(k=0,1,\cdots,n)$是独立变量;y_i是 x_k 的线性组合;$a_k(k=1,2,\cdots,n)$为参数;a_0为常数项;i为样本点($i=1,2,\cdots,m$);θ_i为符合正态分布的独立误差项($\theta_i \sim N(0,\sigma^2)$)。

以矩阵形式表示为:

$$\boldsymbol{Y} = \boldsymbol{XA} + \boldsymbol{\theta} \tag{2}$$

收稿日期:2006-10-23。

项目来源:国家自然科学基金资助项目(49971031);国家教育部长江学者奖励计划资助项目。

其中，$\boldsymbol{A}=[\,a_0 \quad a_1 \quad \cdots \quad a_n\,]^{\mathrm{T}}$，$\boldsymbol{X}=\begin{pmatrix}1 & x_{11} & \cdots & x_{1n}\\ \vdots & \vdots & & \vdots\\ 1 & x_{m1} & \cdots & x_{mn}\end{pmatrix}$，$\boldsymbol{\theta}=[\,\theta_1 \quad \cdots \quad \theta_m\,]^{\mathrm{T}}$，$\boldsymbol{Y}=[\,y_1 \quad \cdots \quad y_m\,]^{\mathrm{T}}$。采用最小二乘方法估计参数：

$$\boldsymbol{A}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

地理加权回归，也称为基础地理加权回归(basic GWR，BGWR)，是线性回归模型的扩展，允许参数在空间区域上变化，其公式为：

$$y_i=a_{i0}+\sum_{k=1}^{n}a_{ik}x_{ik}+\theta_i,i=1,2,\cdots,m \tag{4}$$

以矩阵形式表示为：

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{A}+\boldsymbol{\theta} \tag{5}$$

其中，$\boldsymbol{A}=\begin{pmatrix}a_{10} & a_{20} & \cdots & a_{m0}\\ a_{11} & a_{21} & \cdots & a_{m1}\\ \vdots & \vdots & \vdots & \vdots\\ a_{1n} & a_{2n} & \cdots & a_{mn}\end{pmatrix}$。采用加权最小二乘方法得到回归点 i 的参数估计为：

$$[\,\hat{a}_{i0} \quad \cdots \quad \hat{a}_{in}\,]^{\mathrm{T}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{Y},\quad i=1,2,\cdots,m \tag{6}$$

这里权重矩阵 $\boldsymbol{W}_i=\mathrm{diag}(w_{i1} \quad \cdots \quad w_{ij} \quad \cdots \quad w_{im})$；$j=1,2,\ldots,m$。权衡不同空间位置的观测值对于回归点 i 参数估计的影响程度，$\boldsymbol{W}_{ij}$ 是 i、j 两点之间距离 d_{ij} 的连续单调递减函数。将式(6)代入式(5)得到 $\boldsymbol{Y}$ 的估计值 $\hat{\boldsymbol{Y}}$ 为：

$$\hat{\boldsymbol{Y}}=\boldsymbol{S}\boldsymbol{Y} \tag{7}$$

其中，$\boldsymbol{S}=\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_i$ 为帽子矩阵[3]。

1.2 混合地理加权回归模型

混合地理加权回归模型是同时考虑常系数和变系数的回归模型，这里把常系数划为一组，记为 a_g，称为全局系数；把变系数也划为一组，记为 b_g，称为局域系数；对应的独立变量也分为两组分别记为 $\boldsymbol{X}_a$ 和 $\boldsymbol{X}_b$。于是 MGWR 模型可以写成如下形式：

$$y_i=\sum_{k=1}^{n_1}a_kx_{ik}(a)+\sum_{l=1}^{n_2}b_{il}x_{il}(b)+\theta_i \tag{8}$$

其中，$i=1,2,\cdots,m$；$n_1+n_2=n$；$\{x_{i1}(a) \quad \cdots \quad x_{in_1}(a)\}$ 为与 n_1 个全局系数 $\{a_1 \quad \cdots \quad a_{n_1}\}$ 相对应的独立变量；$\{x_{i1}(b) \quad \cdots \quad x_{in_2}(b)\}$ 为与 n_2 个局域系数 $\{b_{i1} \quad \cdots \quad b_{in_2}\}$ 相对应的独立变量；y_i 是独立变量的线性函数。

分析式(8)可以得到，如果去掉 a_g，MGWR 就变成 GWR，用加权最小二乘方法估计参数；如果去掉 b_g，MGWR 就变成 OLR，用最小二乘方法估计参数。

1.3 混合地理加权回归模型系数估计的算法

MGWR 模型系数估计迭代算法的基本思路是先假设 $\{a_1 \quad \cdots \quad a_{n_1}\}$ 为已知，用 BGWR 方法估计 $\{b_{i1} \quad \cdots \quad b_{in_2}\}$，然后将估计的系数代回方程(8)，用 OLR 方法估计 $\{a_1 \quad \cdots \quad a_{n_1}\}$；再将 $\{a_1 \quad \cdots \quad a_{n_1}\}$ 的估计值代回方程(8)，用 BGWR 方法估计 $\{b_{i1} \quad \cdots \quad b_{in_2}\}$。

为了方便推导，把 MGWR 模型改写成矩阵形式：

$$\boldsymbol{Y}=\boldsymbol{X}_a\boldsymbol{a}+\boldsymbol{M}+\boldsymbol{\theta} \tag{9}$$

其中，$\boldsymbol{Y}$ 为非独立变量向量；$\boldsymbol{X}_a$ 为 a_g 对应的独立变量矩阵；$\boldsymbol{a}$ 为 a_g 的系数向量；$\boldsymbol{\theta}$ 为误差向量；$\boldsymbol{M}$ 向量的第 i 个元素为 $\sum_{l=1}^{n_2}b_{il}x_{il}(b)$。假设 $\{a_1 \quad \cdots \quad a_{n_1}\}$ 已知，将 a_g 从 $\boldsymbol{Y}$ 中减掉，得：

$$(\boldsymbol{Y}-\boldsymbol{X}_a\boldsymbol{a})=\boldsymbol{M}+\boldsymbol{\theta} \tag{10}$$

由 §1.1 可知：

$$\hat{\boldsymbol{M}}=\boldsymbol{S}(\boldsymbol{Y}-\boldsymbol{X}_a\boldsymbol{a}) \tag{11}$$

这里 $\boldsymbol{S}=\boldsymbol{X}_b(\boldsymbol{X}_b^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{X}_b)^{-1}\boldsymbol{X}_b^{\mathrm{T}}\boldsymbol{W}_i$，$\boldsymbol{S}$ 为 b_g 的帽子矩阵。

将 $\boldsymbol{M}$ 向量的估计值 $\hat{\boldsymbol{M}}$ 代入方程(9)可得：

$$(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{Y}=(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}_a\boldsymbol{a}+\boldsymbol{\theta} \tag{12}$$

设 $\boldsymbol{Z}=(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{Y}$，$\boldsymbol{Q}=(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}_a$，则式(12)简化为：

$$\boldsymbol{Z}=\boldsymbol{Q}\boldsymbol{a}+\boldsymbol{\theta} \tag{13}$$

方程(13)为标准的 OLR 模型，用最小二乘法可以得到全局系数 $\boldsymbol{a}$ 的估计值 $\hat{\boldsymbol{a}}$：

$$\hat{\boldsymbol{a}}=(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q})^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Z}=(\boldsymbol{X}_a^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{X}_a)^{-1}\boldsymbol{X}_a^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{Y} \tag{14}$$

将式(14)代入式(10)，采用加权最小二乘方法，可以得到局域系数 $\boldsymbol{b}$ 的估计值 $\hat{\boldsymbol{b}}$ 为：

$$\hat{\boldsymbol{b}}_i=(\boldsymbol{X}_b^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{X}_b)^{-1}\boldsymbol{X}_b^{\mathrm{T}}\boldsymbol{W}_i(\boldsymbol{Y}-\boldsymbol{X}_a\hat{\boldsymbol{a}}) \tag{15}$$

计算过程如下。

1) 对 $\boldsymbol{X}_a$ 的每一行以 $\boldsymbol{X}_b$ 为独立变量用 BGWR 进行回归计算得到 $\hat{\boldsymbol{X}}_a$，并计算 $\boldsymbol{X}_a$ 的回归残差 $\boldsymbol{r}(\boldsymbol{X}_a)$；

2) 对 $\boldsymbol{Y}$ 以 $\boldsymbol{X}_b$ 为独立变量用 BGWR 进行回归计算出 $\hat{\boldsymbol{Y}}$，并计算 $\boldsymbol{Y}$ 的回归残差 $\boldsymbol{r}(\boldsymbol{Y})$；

3) 对 $\boldsymbol{r}(\boldsymbol{Y})$ 以 $\boldsymbol{r}(\boldsymbol{X}_a)$ 为独立变量，用 OLR 回归，得到全局系数 $\boldsymbol{a}$ 的估计值 $\hat{\boldsymbol{a}}$；

4) 对 $(\boldsymbol{Y}-\boldsymbol{X}_a\hat{\boldsymbol{a}})$ 以 $\boldsymbol{X}_b$ 为独立变量，用 BGWR 进行回归，得到局域系数 $\boldsymbol{b}$ 的估计值 $\hat{\boldsymbol{b}}$。

2 试验分析

根据上海市房地资源管理局提供的房地产销售价格以及楼盘相关信息等，选取市中心 83 个样本区域中 567 个住宅小区进行试验，所有小区开

发、销售时间在 2003 年 1 月到 2005 年 7 月之间,样本分布情况如图 1 所示。为了对上海市住宅楼盘销售平均价格(元/m^2)进行回归分析,经过比较研究,选取了 18 个典型的影响因子(变量)[8],分别为区域因素 10 个、个别因素 6 个、一般因素 2 个,见表 1。其中人口密度和失业率为社会经济等一般因素,对整个城市房地产市场的影响力是一致的,为 a_g 变量;其余因子皆随区域而异,为 b_g 变量。由于获取的样本存在时间差异,为了减少时间因素对回归结果的影响,在回归计算以前进行了数据归一化,根据上海市房地产评估中心提供的相关资料,把销售平均价统一归算到 2005 年 7 月。回归分析分别采用 BGWR 模型和 MGWR 模型进行拟合,其中试验中选取的权重函数为可变带宽的二次方核函数[3,9]。

两个模型的环境质量(qua-envi)变量的估计结果在空间上的分布如图 2 所示。通过对比分析可以发现,两个模型得出的环境质量变量的估计值的空间变化图景的整体趋势是非常相似的,影响峰值都沿着黄浦江与苏州河两岸,以及浦东世纪公园等大型绿地周围,与现实情况吻合良好,说明 MGWR 模型和 BGWR 模型都能较好地拟合样本数据。进一步仔细观察会发现两者还是存在一定的差异,以影响峰值周围区域为例,系数在图 2(a)中变化梯度较大,而在图 2(b)中要相对平滑,原因就在于代表住宅个性和区域特性的变量往往和一般性的社会经济因素相关联,把社会经济因素从 BGWR 模型中忽略掉,不可避免会导致住宅个性和区域特性变量的对应系数估计产生一定程度的区域波动,从而导致 BGWR 模型估计产生偏差。两个模型的其他变量的估计值在空间变化图景的整体趋势上也是相似的。

AIC(Akaike information criteria)信息规则是模型比较的常用方法,两个模型的 AIC 值相差

表 1　影响因子

Tab. 1　Independent Variables

变量名称	变量说明	BGWR	MGWR
容积率	小区容积率=总建筑面积/土地总面积	*	*
绿化率	小区绿化率=(土地面积-建筑占地面积)/土地面积	*	*
地铁站距离	小区到最近地铁站的距离,小于 1 km 为 1;其他为 0	*	*
道路宽度	小区前面道路的宽度,以 m 为单位	*	*
建筑类型	小区建筑类型,别墅为 2;6 层以上为 1;其他为 0	*	*
建筑质量	小区建筑质量好为 1;其他为 0	*	*
公园距离	小区到公园(s>50 000 m^2)的距离小于 1 km 为 1;其他为 0	*	*
占地面积	小区的宗地面积/m^2	*	*
临街距离	小区临街距离/m	*	*
公交站数量	小区周围公交站总数量	*	*
地铁站数量	样本区域内地铁站总数量	*	*
停车场数量	样本区域内社会停车场总数量	*	*
环境质量	样本区域综合环境质量,好为 1;其他为 0	*	*
公用设施数量	样本区域公用设施数量	*	*
道路覆盖率	样本区域道路覆盖率=机动车道道路总长度/样本区域面积	*	*
房屋出租情况	样本区域房屋出租超过 1/3 为 1;其他为 0	*	*
人口密度	城市人口密度 (100 人/km^2)	—	√
失业率	城市失业率,上海市政府 2004 年公布的数据	—	√

注:* 表示局域变量,√ 表示全局变量,— 表示未包含。

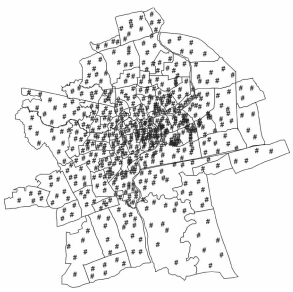

图 1　样本住宅小区示意图

Fig. 1　Distribution of Sample House-Blocks

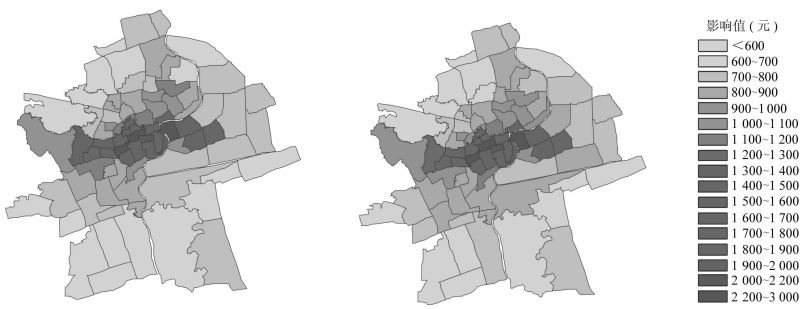

图 2　BGWR 中样本区域环境质量估计值分布图

Fig. 2　Map of Qua-envi Term in BGWR

3以上就认为有明显差异,模型的AIC值越小,模型越好。其计算公式如下:

$$\mathrm{AIC}=2n\ln(\hat{\sigma})+n\ln(2\pi)+n\left\{\frac{n+\mathrm{tr}(\boldsymbol{S})}{n-2-\mathrm{tr}(\boldsymbol{S})}\right\} \tag{16}$$

其中,$\boldsymbol{S}=\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_i$;$\hat{\sigma}=\sqrt{\frac{R_{SS}}{n-\mathrm{tr}(\boldsymbol{S})}}$,$R_{SS}=\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{y}$;$\mathrm{tr}(\boldsymbol{S})=\sum s_{ii}$ $(i=1,2,\cdots,n)$。

比较两模型的AIC值可以发现,2个模型对567个样本数据拟合的程度比较接近,其中BGWR为11 348,MGWR为11 342,也说明模型MGWR优于模型BGWR。

3 结 语

针对混合地理加权回归模型中既有全局参数,又有局域参数,不能直接应用地理加权回归模型方法进行参数估计的问题,本文以迭代算法为理论基础,推导出了混合地理加权回归模型中全局参数和局域参数估计的准确表达式,并给出了具体的计算方法,较好地解决了混合地理加权回归模型参数估计的问题。以上海市房地产住宅小区销售平均价格为例,验证了本文提出的参数估计方法的正确性和有效性,无论从可视化效果还是定量分析来看,MGWR模型拟合效果都要好于BGWR。当然由于MGWR模型在第一个回归点要运行(n_2+2)次BGWR模型,计算量偏大,但随着计算机性能的提高,计算时间差异将会越来越不明显。在本文的回归分析中,只考虑了空间因素,而城市空间分布和发展模式同时也受时间因素的影响,因此应进一步发展4维的时空MGWR模型。

参 考 文 献

[1] 魏传华,梅长林.半参数空间变系数回归模型的两步估计方法及其数值模拟[J].统计与信息论坛,2005,20(1):16-19

[2] Fotheringham A S, Charlton M, Brunsdon C. The Geography of Parameter Space: an Investigation into Spatial Nonstationarity[J]. International Journal of Geographical Information Systems, 1996,10:605-627

[3] Brunsdon C, Fotheringham A S, Charlton M. Spatial Nonstationarity and Autoregressive Models[J]. Environment and Planning A,1998,30(6):957-973

[4] Brunsdon C, Fotheringham A S, Charlton M. Geographically Weighted Regression: a Method for Exploring Spatial Nonstationarity[J]. Geographical Analysis,1996, 28(4):281-298

[5] 覃文忠,王建梅,刘妙龙.地理加权回归分析空间数据的空间非平稳性[J].辽宁师范大学学报(自然科学版),2005, 28(4):476-479

[6] Brunsdon C, Fotheringham A S, Charlton M E. Some Notes on Parametric Signficance Tests for Geographically Weighted Regression[J]. Journal of Regional Science, 1999,39(3):497-524

[7] Mei Changlin, He Shuyuan, Fang Kaitai. A Note on the Mixed Geographically Weighted Regression Model[J]. Journal of Regional Science, 2004, 44(1):143-158

[8] Gao Xiaolu, Asami Y. Influence of Spatial Features on Land and Housing Prices[J]. Tsinghua Science and Technology, 2005(3):344-353

[9] Huang Yefang, Leung Y. Analysing Regional Industrialisation in Jiangsu Province Using Geographically Weighted Regression[J]. Geography System, 2002(4):233-249

第一作者简介:覃文忠,副教授,博士生。现主要从事地理信息空间建模与空间分析研究。

E-mail:wenzhongq@mail.tongji.edu.cn

Algorithm for Mixed Geographically Weighted Regression Model

QIN Wenzhong[1] *WANG Jianmei*[1] *LIU Miaolong*[1]

(1 Department of Surveying and Geo-informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China)

Abstract: An iterative algorithm is developed to estimate global coefficients and local coefficients in MGWR. First independent variables are classified two groups, Group a_g in which variables are global associated with global coefficients and Group b_g in which variables are local associated with local coefficients. Second assuming that a_g is known, coefficients of b_g is

calibrated by using the basic GWR. Third ordinary linear regression (OLR) is used to estimated coefficients of a_g. Material formulations of two types of coefficients and computational progress are also produced, and further tested by using average prices of house blocks in Shanghai. The experiment proves that all formulations of coefficients are available, and comparison of the two models by Akaike information criteria value shows MGWR is more appropriate and stable for the local coefficients estimates than BGWR although it requires a greater computational effort.

Key words: geographically weighted regression; mixed geographically weighted regression; spatial nonstationarity; iterative algorithm; spatial analysis

About the first author: QIN Wenzhong, associate professor, Ph.D candidate. His research interests include spatial modeling and spatial analysis in GIS.
E-mail: wenzhongq@mail.tongji.edu.cn

(上接第 111 页)

An Elevation Model of InSAR and Its Accuracy Analysis

ZHANG Lei[1]　*WU Jicang*[1,2]　*CHEN Yanling*[3]

(1　Department of Surveying and Geo-informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China)
(2　Key Laboratory of Geospace Environment & Geodesy, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)
(3　Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, China)

Abstract: An improved elevation model of InSAR which expresses the direct relationship between height and phase has been developed based on the InSAR elevation model of Currie and Baker. The errors introduced by the parallel ray approximation applied in the improved model are analyzed, and the results verify that it can not be ignored for satellite radar system. The error propagation curves and the relation between elevation errors and baseline parameters are also given which can be used to rectify the approximation errors and reconstruct the high accurate digital elevation model (DEM). In addition, based on the improved model and error propagation law, an elevation error estimation formula is derived. It gives the definite effect of the baseline length and angle on the elevation accuracy.

Key words: InSAR; elevation model; baseline; parallel ray approximation; error propagation

About the first author: ZHANG Lei, postgraduate, interested in the theory of InSAR and its applications in surface displacement detection.
E-mail: zhanglen@gmail.com