

# 基于栅格面积成分数据的土地利用格局 解释模型稳健估计

邓祥征<sup>1</sup>, 林英志<sup>1</sup>, 战金艳<sup>2</sup>, 何书金<sup>1</sup>

(1. 中国科学院地理科学与资源研究所, 北京 100101; 2. 北京师范大学环境学院水环境模拟国家重点实验室, 北京 100875)

**摘 要:** 针对最小二乘估计不能应用于栅格尺度以面积成分表征的土地利用格局驱动机理分析的难题, 本文提出了一种利用偏最小二乘回归法稳健估计该类土地利用格局解释模型的方法。利用该方法能在解释变量间存在多重共线性的情况下, 获得基于栅格面积成分数据的土地利用格局解释模型的稳健估计。本文推导了应用偏最小二乘回归分析的数据处理和建模估计过程, 并运用该方法开展了针对黄淮海地区耕地、建设用地分布格局及其驱动因子的建模分析, 得到了拟合优度高的估计结果。研究表明, 偏最小二乘回归分析方法在开展栅格尺度以面积成分表征的土地利用格局驱动机理分析时具备高效与稳健的特征, 适宜在类似研究中推广应用。

**关 键 词:** 栅格成分数据; 偏最小二乘回归; 土地利用格局; 驱动力; 黄淮海地区

## 1 引言

随着 GIS 空间数据分析技术的不断发展与大量空间数据集的积累, 栅格数据逐渐成为地理学相关研究领域最主要、最基本的数据格式之一<sup>[1-3]</sup>。而基于栅格数据发展而来的栅格面积成分数据, 在土地利用格局演替、土地利用结构变化模拟等领域发挥着日益重要的作用。虽然广受地学分支学科及其研究领域的重视, 迄今为止, 栅格面积成分数据的建模及估计方法仍未取得实际性进展<sup>[4-7]</sup>。当前, 为了避免土地利用格局驱动机理分析中的共线性问题, 常用的土地利用格局驱动机理回归分析方法仍主要依靠通过增减自变量提升模型估计结果的准确性。利用栅格面积成分数据估计土地利用分布解释模型统计建模也面临着两个技术难题: 一是回归分析得到的栅格内各种土地利用类型面积之和常常不等于栅格总面积; 二是变量间的多重共线性问题导致模型估计结果存在偏误<sup>[8]</sup>。

针对上述问题, 本文提出并推导了栅格尺度以面积成分表征的土地利用格局解释模型的偏最小二乘回归分析方法, 并在黄淮海地区开展了案例研究, 建立起土地利用格局解释模型, 揭示了 2000 年

各驱动因子对黄淮海地区土地利用格局的驱动机理, 为土地利用格局演替、土地利用变化驱动机理分析等研究提供可资借鉴的方法。

本文第二部分介绍栅格成分数据的概念和基于栅格面积成分数据进行偏最小二乘回归分析的步骤。第三部分针对黄淮海地区开展案例研究, 验证栅格成分偏最小二乘回归分析方法的有效性与推广应用价值。最后一部分是结论与讨论。

## 2 模型与方法

### 2.1 栅格成分数据

成分数据的概念最早由 Ferrers 于 1866 年提出<sup>[9]</sup>, 经一个多世纪的发展, 现已成为统计分析中的一种重要数据类型。通常, 成分数据可用如下向量的形式表达:

$$S = \left\{ (s_1, s_2, \dots, s_m)^T \in R^m \mid \sum_{i=1}^m s_i = 1, 0 < s_i < 1 \right\} \quad (1)$$

$$s_i = S_i / \sum_{j=1}^m S_j \quad (2)$$

式中:  $m$  表示成分数据包含的分量个数;  $s_i$  表示成分数据的第  $i$  个分量;  $S_i$  表示第  $i$  个分量的原始观测值, 如耕地面积、建设用地面积等。

收稿日期: 2009-08; 修订日期: 2009-11.

基金项目: 国家自然科学基金项目(70873118、70821140353)、国家科技支撑项目(2006BAC08B03、2006BAC08B06、2008BAK47B02)。

作者简介: 邓祥征(1971-), 男, 博士, 副研究员, 主要从事区域环境变化、土地系统变化及效应研究。

E-mail: dengxz.ccap@igsnrr.ac.cn

栅格成分数据即是基于一定尺度栅格建立的成分数据。较一般意义上的成分数据,栅格成分数据不但受条件(3)的约束,还须满足条件(4)。

$$\sum_{i=1}^m s_i=1, 0<s_i<1 \tag{3}$$

$$\sum_{i=1}^m S_i=\Omega \tag{4}$$

式中: $\Omega$ 为常数,表示栅格面积。

采用栅格成分数据开展回归建模分析,必须首先解决如下三个主要问题。

首先,栅格成分数据各分量的取值在区间(0, 1)内部,但是现实情况中,样本的取值往往会出现一个或几个分量为 0 的情况,即存在分量  $s_i=0$  或 1 的情况。

其次,由于栅格成分数据各分量取值的局限性,因变量难以符合线性回归分析的正态分布假设,无法准确分析数据之间的相关性。

第三,无论自变量的取值如何,建立的回归模型必须保证约束条件(3)、(4)始终成立。

为了解决上述难题,以土地利用栅格面积成分数据为例,本文提出解决这三个问题的处理方案——非零化处理和对称 Logratio 变换。

### 2.1.1 非零化处理

现实中不存在一定空间尺度上某种土地利用类型面积绝对为零的情况<sup>[10]</sup>。鉴于此,当某种土地利用类型面积为零时,本文赋予其面积为一个微小的值。即对于样本向量

$$(s'_1, s'_2, \dots, s'_j)^T \in [0, 1]^p \tag{5}$$

式中: $p$ 表示土地利用类型的数量; $s'_j$ 表示第  $j$  种土地利用面积占栅格总面积的份额。当时  $s'_j=0$ ,我们将给它重新赋以一个微小的值  $s'_j=\varepsilon$ 。

进而,令  $y_j=s'_j/\sum_{i=1}^p s'_i$ ,则得到处理之后的土地利用栅格面积成分数据:

$$Y=\left\{(y_1, y_2, \dots, y_p)^T \in R^p \mid \sum_{j=1}^p y_j=1, 0<y_j<1\right\} \tag{6}$$

式中: $y_j$ 表示第  $j$  种土地利用面积占整个栅格面积的比例。

### 2.1.2 对称 Logratio 变换

对处理之后的栅格成分数据进行对称 Logratio 变换,建立成分数据的逻辑正态分布模型:

$$Z=(z_1, z_2, \dots, z_p)^T, z_j=\ln\left(y_j/\sqrt{\prod_{i=1}^p y_i}\right) j=1, 2, \dots, p \tag{7}$$

显然有  $z_j \in (-\infty, +\infty)$ 。令  $s_j=z_j-z_p, j=1, 2, \dots, p-1$ , 则可由反变换得:

$$y_j=\frac{e^{s_j}}{1+\sum_{i=1}^{p-1} e^{s_i}}, y_p=\frac{1}{1+\sum_{i=1}^{p-1} e^{s_i}}, j=1, 2, \dots, p-1 \tag{8}$$

这一变换不但解决了成分数据取值范围狭小与总和约束带来的问题,还将回归分析中的非线性问题线性化。另外,这一变换保持了原有成分数据的对称性,其建模分析结果可直接反映栅格成分数据的特征,从而使估计的结果易于解释性<sup>[11,12]</sup>。

## 2.2 偏最小二乘估计

变量之间的多重共线性问题是回归统计分析中的一个必须处理的问题,土地利用格局驱动机理分析也是如此。自 20 世纪 30 年代 Frisch 提出多重共线性问题以来<sup>[13]</sup>,对其的探究就没有间断过。多重共线性导致统计模型很难获得精确估计,甚至出现估计系数及其符号存在严重偏误的情况,并使得估计出的系数对样本数据的微小变化产生非常敏感的反应。1983 年, Wold 和 Albano 等人提出了偏最小二乘回归方法<sup>[14]</sup>。该方法基于因子分析,是一种将降维空间的每个元素组成的预测矩阵与被预测矩阵间的协方差最大化的降维技术。

假定自变量集为  $X=(X_1, \dots, X_q)^T$ , 因变量集为  $Z=(Z_1, \dots, Z_q)^T$ , Logratio 变换之后的因变量和自变量组的  $n$  次标准化观测数据阵分别为:

$$Z_0=\begin{bmatrix} Z_{11} & \cdots & Z_{1p} \\ Z_{21} & \cdots & Z_{2p} \\ \vdots & & \vdots \\ Z_{n1} & \cdots & Z_{np} \end{bmatrix}, X_0=\begin{bmatrix} x_{11} & \cdots & x_{1q} \\ x_{21} & \cdots & x_{2q} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nq} \end{bmatrix} \tag{9}$$

### 2.2.1 第一对成分

假设提取的第一对成分分别为  $U_1$  和  $V_1$ 。  $U_1$  是自变量集  $X=(X_1, \dots, X_q)^T$  的线性组合:

$$U_1=\omega_{11}X_1+\cdots+\omega_{1q}X_q=\omega_1^T X \tag{10}$$

$V_1$  是因变量集  $Z=(Z_1, \dots, Z_q)^T$  的线性组合:

$$V_1=v_{11}Z_1+\cdots+v_{1p}Z_p=v_1^T Z \tag{11}$$

由标准化观测数据阵  $Z_0$  和  $X_0$ , 可计算第一对成分的得分向量,分别记为  $u_1$  和  $v_1$ :

$$u_1 = X_0 \omega_1 = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \omega_{11} \\ \omega_{12} \\ \vdots \\ \omega_{1q} \end{bmatrix} = \begin{bmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{bmatrix} \quad (12)$$

$$v_1 = Z_0 v_1 = \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ z_{21} & \cdots & z_{2p} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{np} \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \\ \vdots \\ v_{1q} \end{bmatrix} = \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{n1} \end{bmatrix} \quad (13)$$

第一对成分  $U_1$  和  $V_1$  的协方差  $\text{Cov}(U_1, V_1)$  可用其得分向量  $u_1$  和  $v_1$  的内积来计算。故而,条件极值问题:

$$\begin{cases} \max\{(u_1, v_1)\} = \max\{\omega_1^T X_0^T Z_0 v_1\} \\ \omega_1^T \omega_1 = \|\omega_1\|^2 = 1, v_1^T v_1 = \|v_1\|^2 = 1 \end{cases} \quad (14)$$

可用以表达——求单位向量  $\omega_1$  和  $v_1$ , 使提取的第一对成分和满足如下要求:①各自尽可能多地提取所属变量组的信息;②二者之间相关程度达到最大。该极值通过计算  $q$  阶矩阵  $Q = X_0^T Z_0 Z_0^T X_0$  特征值与特征向量求解,  $Q$  最大特征值为  $(\omega_1^T X_0^T Z_0 v_1)^2$ , 对应的单位特征向量就是所求的解  $\omega_1$ , 而  $v_1$  由(15)计算得到。

$$v_1 = \frac{1}{\omega_1^T X_0^T Z_0 v_1} Z_0^T X_0 \omega_1 \quad (15)$$

### 2.2.2 第一对成分回归方程

建立  $Z_0, \dots, Z_p$  及  $X_1, \dots, X_q$  对  $U_1$  的回归方程。

$$\text{假定回归模型为:} \begin{cases} X_0 = u_1 \alpha_1^T + S_1 \\ Z_0 = u_1 \beta_1^T + T_1 \end{cases} \quad (16)$$

式中:  $u_1$  为  $U_1$  的  $n$  维得分向量,  $\alpha_1^T = (\alpha_{11}, \dots, \alpha_{1q})$  和  $\beta_1^T = (\beta_{11}, \dots, \beta_{1p})$  均为一个自变量对多个因变量的回归模型参数向量,  $S_1$  和  $T_1$  分别为  $n \times q$  和  $n \times p$  残差阵。易知, 回归系数向量  $\alpha_1$  和  $\beta_1$  的最小二乘估计

$$\text{为:} \begin{cases} \alpha_1^T = (u_1^T u_1)^{-1} u_1^T X_0 \\ \beta_1^T = (u_1^T u_1)^{-1} u_1^T Z_0 \end{cases} \quad (17)$$

### 2.2.3 最终回归方程

记  $\hat{X}_0 = u_1 \alpha_1^T, \hat{Z}_0 = u_1 \beta_1^T$ 。则残差阵  $S_1 = X_0 - \hat{X}_0, T_1 = Z_0 - \hat{Z}_0$ 。用残差阵  $S_1$  和  $T_1$  代替  $X_0$  和  $Z_0$ , 并重复 2.2.1 和 2.2.2 的运算, 即得第二对成分  $S_2$  和  $T_2$  的权重:

$$w_2 = (w_{21}, \dots, w_{2q})^T, v_2 = (v_{21}, \dots, v_{2p})^T \quad (18)$$

而  $v_2 = T_1 v_2$  和  $w_2 = S_1 w_2$  分别为第二对成分的得分向

$$\text{量。} \begin{cases} \alpha_2^T = (u_2^T u_2)^{-1} u_2^T S_1 \\ \beta_2^T = (u_2^T u_2)^{-1} u_2^T T_1 \end{cases} \quad (19)$$

为第二对成分的载荷量。此时有

$$\begin{cases} X_0 = u_1 \alpha_1^T + \dots + u_2 \alpha_2^T + S_2 \\ Z_0 = u_1 \beta_1^T + \dots + u_2 \beta_2^T + T_2 \end{cases} \quad (20)$$

设  $n \times q$  数据阵  $X_0$  的秩为  $r \leq \min(n-1, q)$ , 则存在  $r$  个成分  $u_1, u_2, \dots, u_r$ , 使得

$$\begin{cases} X_0 = u_1 \alpha_1^T + \dots + u_r \alpha_r^T + S_r \\ Z_0 = u_1 \beta_1^T + \dots + u_r \beta_r^T + T_r \end{cases} \quad (21)$$

设  $X_i^*$  ( $i=1, 2, \dots, q$ ),  $Z_j^*$  ( $j=1, 2, \dots, p$ ) 分别表示自变量和因变量的标准化变量, 将

$$U_k = \omega_{k1} X_1^* + \omega_{kq} X_q^*, \quad k=1, \dots, r \quad (22)$$

代入

$$Z_j^* = \beta_{j1} U_1 + \beta_{j2} U_2 + \dots + \beta_{jr} U_r \quad j=1, \dots, p \quad (23)$$

即得  $p$  个标准化因变量的偏最小二乘回归方程:

$$\hat{Z}_j^* = \alpha_{j1}^* X_1^* + \dots + \alpha_{jq}^* X_q^*, \quad j=1, \dots, p \quad (24)$$

然后再还原为原始变量的偏最小二乘回归方程:

$$\hat{Z}_j = \alpha_{j0} + \alpha_{j1} X_1 + \dots + \alpha_{jq} X_q, \quad j=1, \dots, p \quad (25)$$

容易验证, 上述偏最小二乘回归方程满足约束条件(3)和(4), 并对任意的  $i=0, 1, 2, \dots, q$ , 等式(26)

$$\text{恒成立。} \sum_{j=1}^p \alpha_{ji} = 0 \quad (26)$$

### 2.2.4 成分个数的判定

通常情况下, 偏最小二乘回归分析不需要选用所有的成分来建立回归方程, 而像主成分分析一样, 只选用前几个成分即可得到预测、解释能力较好的回归模型。常用的确定抽取成分个数的方法有“舍一交叉验证法”、“分批交叉验证法”、“分裂样本交叉验证法”和“随机样本交叉验证法”, 这些方法的差别在于交叉验证数据集的抽取方法不同<sup>[15]</sup>:

①舍一交叉验证法: 每次扣留第  $i$  ( $i=1, \dots, n$ ) 个观测作为检验数据, 用余下的  $n-1$  个观测按偏最小二乘回归方法建模;

②分批交叉验证方法: 每次扣留连续  $j$  个观测作为检验数据集,  $j=1$  时即为舍一交叉验证方法;

③分裂样本交叉验证方法: 扣留的用以作为检

验数据集的观测不必是连续的,而是按照一定的宽度抽取构成的;

④随机样本交叉验证方法:此方法中所扣留的用以作为检验数据集的观测是随机抽取的。

把交叉验证数据集  $I \subset \{1, \cdots, n\}$  代入抽取  $k$  个成分后拟合的回归方程,得到  $Z_j (j=1, \cdots, p)$  在这些观测点上的预测值  $\hat{z}_{ji}(k), i \in I$ 。对  $I$  重复以上验证,即得抽取  $k$  个成分时第  $j$  个因变量  $Z_j(j=1, \cdots, p)$  的预测参差平方和:

$$PRESS_j(k) = \sum_{I \subset \{1, \cdots, n\}} \sum_{i \in I} (z_{ij} - \hat{z}_{ji}(k))^2 \quad (27)$$

进而求得  $Z=(Z_1, \cdots, Z_p)^T$  的预测残差平方和:

$$PRESS(k) = \sum_{j=1}^p PRESS_j(k) \quad (28)$$

选取使  $Z$  的预测残差平方和达到最小的  $k$ ,即

为要抽取的成分的个数。

3 黄淮海地区案例分析

3.1 数据与处理

黄淮海地区土地利用数据来源于中国科学院资源环境科学数据中心。该土地利用数据包括了耕地、林地、草地、水域、建设用地和未利用地六大类土地利用类型的面积,大致反映了 2000 年黄淮海土地利用情况。本文将土地利用类型中的林地、草地、水域和未利用地面积合并为“其它用地面积”,与耕地面积、建设用地面积一起构成栅格成分数据(图 1)。此外,本文还综合考虑影响土地利用格局的自然环境条件和社会经济因素。选取的自变量分属气候、地形、土壤、人口、投资等多个层面(图 2)。上

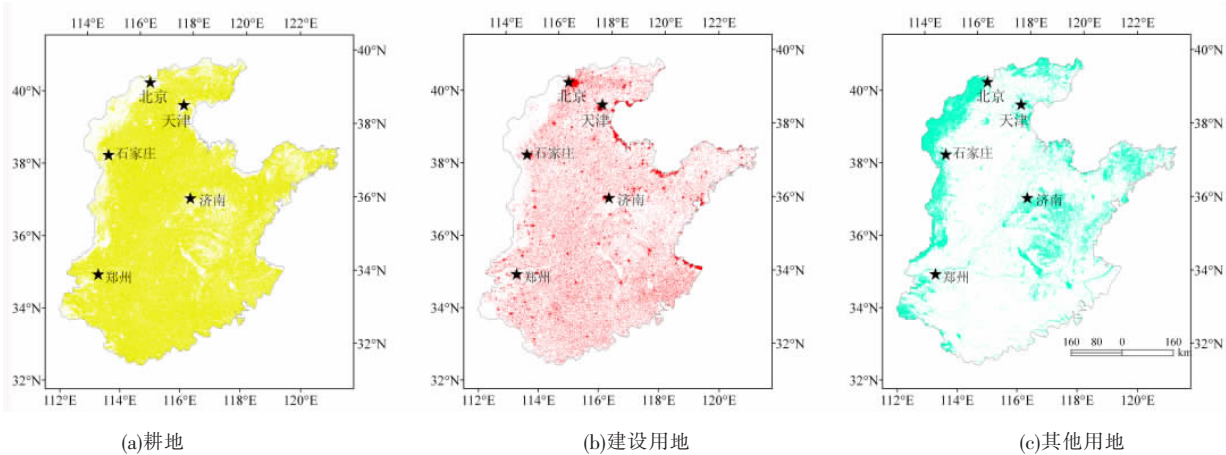


图 1 黄淮海地区 2000 年土地利用格局

Fig.1 Land use patterns of Huang-Huai-Hai Plain in 2000

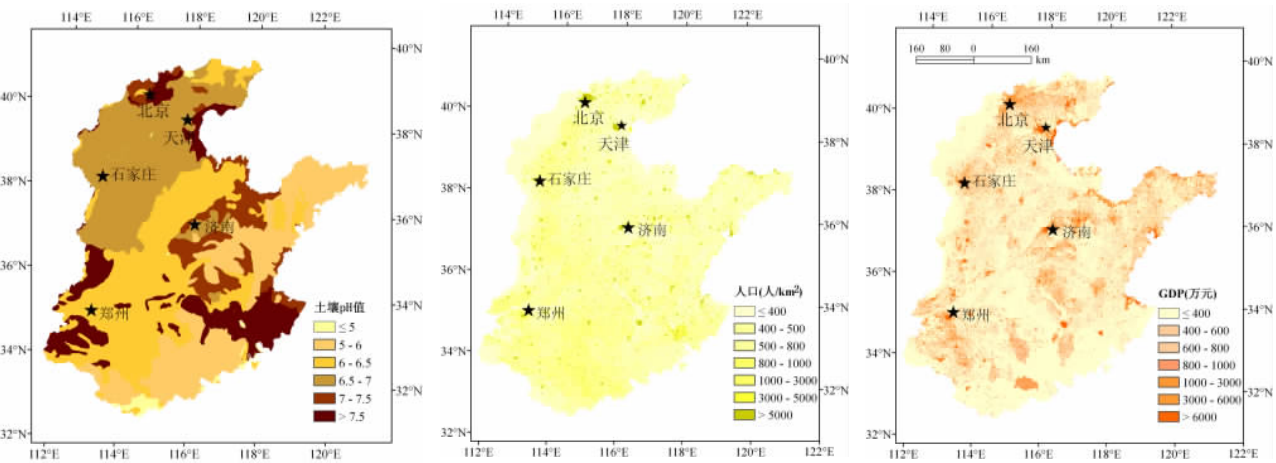


图 2 2000 年黄淮海地区土地利用格局解释模型部分自变量空间分异特征

Fig.2 Spatial heterogeneity of the explanatory variables for the determinant model of land use pattern for Huang-Huai-Hai Plain in 2000



述数据均被离散或集成到 1 km×1 km 栅格上。为避免空间自相关问题对估计结果的影响<sup>[6]</sup>,本文采用系统抽样法,每隔 50 个栅格抽取一个样本,共提取出观测值 183 个。具体变量及解释见表 1。

基于上述设定与变量处理,本文构建了黄淮海地区土地利用格局解释模型:

$$(cropld, bua, restld)^T + A_{3 \times 10}(soilpH, soilbulk, elevation, slope, temp, rainfall, urpopden, aglbr, aginvest, gdppc)^T + B_{3 \times 1}$$

(29)

式中:  $A_{3 \times 10}$  为待估系数矩阵;  $B_{3 \times 1}$  为待估截距项。

3.2 估计结果与分析

3.2.1 共线性检验

为规避自变量之间的多重共线性问题导致的估计结果偏误,在估计以栅格面积成分表征的土地利用格局解释模型之前,本文选用方差膨胀因子 VIF 测度了自变量之间的多重共线性(图 3)。

从共线性诊断结果来看,10 个变量中的高程、农村劳动力、年降水总量、坡度、年平均气温方差膨胀因子(VIF)超过了 2.0,说明多重共线性普遍存在。其中高程和年降水总量的 VIF 接近 5.0,年平均气温的 VIF 超过 6.0,说明自变量间的多重共线性较为严重。

3.2.2 普通最小二乘回归估计

在运用栅格成分数据的偏最小二乘回归分析之前,本文首先利用普通最小二乘回归分析方法估计了黄淮海地区耕地、建设用地格局驱动模型。

可以看出,运用普通最小二乘回归分析方法估计的黄淮海地区耕地、建设用地变化驱动机理模型获得的部分估计结果严重偏离了大家的一般认识。

表 1 黄淮海地区土地利用格局解释模型包含的变量

Tab.1 Variables included in the determinant model of land use pattern for Huang-Huai-Hai Plain

符号	含义	单位	均值	标准差
<i>cropld</i>	耕地面积成分对称 Logratio 变换后的值	—	6.15	4.70
<i>bua</i>	建设用地面积成分对称 Logratio 变换后的值	—	-9.76	5.63
<i>restld</i>	其他用地面积成分对称 Logratio 变换后的值	—	3.61	5.51
<i>soilpH</i>	土壤酸碱度	—	6.44	1.13
<i>soilbulk</i>	土壤容重	0.01 g/cm <sup>3</sup>	137.56	3.47
<i>elevation</i>	高程	m	112.36	187.73
<i>slope</i>	地形坡度	0.01 °	80.98	206.07
<i>temp</i>	年平均气温	0.01 °C	1629.01	178.24
<i>rainfall</i>	年降水总量	0.1 mm	7208.16	2472.28
<i>urpopden</i>	城镇人口密度	10 <sup>2</sup> 人/km <sup>2</sup>	65.07	93.16
<i>aglbr</i>	农村劳动力数量	人	101.6	62.4
<i>aginvest</i>	农业总投资	元	5360.49	3.34×10 <sup>4</sup>
<i>gdppc</i>	人均 GDP	元	6416.31	4189.79

例如,黄淮海地区普遍干旱,降水偏多有利于耕地保育,所以年降水总量对耕地面积的回归系数-0.013 显然不符合当地的实际情况。诸如此类的有违常理估计结果还包括土壤容重、农村劳动力、人均 GDP 等。另外,需要指出的是,上述回归系数显然不能满足等式(26),亦即回归结果不能满足各种土地利用面积之和等于栅格面积的收敛条件。总之,普通最小二乘回归分析的估计结果不能真正揭示黄淮海地区土地利用格局的驱动机理。

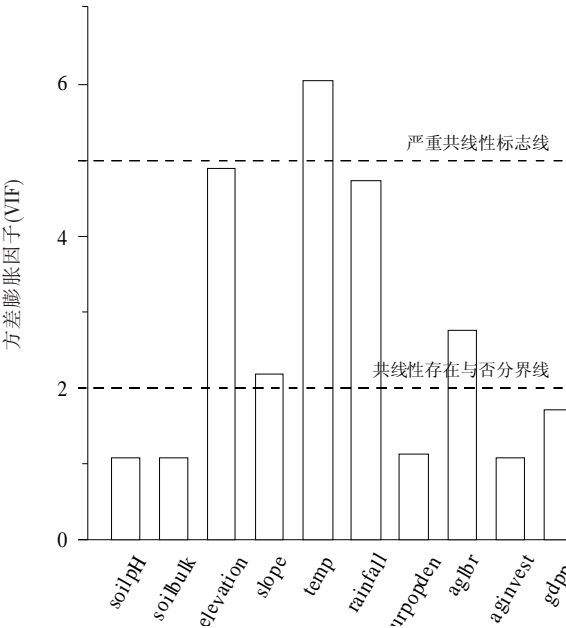


图 3 黄淮海地区土地利用格局解释模型自变量共线性诊断

Fig.3 Collinearity diagnostics for the explanatory variables included in the determinant model for land use pattern in Huang-Huai-Hai Plain

表 2 黄淮海地区土地利用格局解释模型普通最小二乘回归估计 \*

Tab.2 OLS estimates of the determinant model of land use pattern for Huang-Huai-Hai Plain

	<i>cropld</i>	<i>Bua</i>	<i>restld</i>
<i>soilpH</i>	0.063(0.09)	0.005(0.03)	-0.059(0.09)
<i>soilbulk</i>	0.017(0.10)	-0.077(1.06)	0.064(0.48)
<i>elevation</i>	-0.024(1.57)	-0.153(1.04)	0.177(2.20) <sup>***</sup>
<i>slope</i>	-0.161(2.81) <sup>****</sup>	-0.049(0.38)	0.188(3.16) <sup>****</sup>
<i>temp</i>	0.182(1.90) <sup>**</sup>	-0.097(1.00)	-0.056(1.45)
<i>rainfall</i>	-0.013(0.57)	-0.168(0.20)	0.183(0.47)
<i>urpopden</i>	0.172(0.50)	0.059(0.28)	-0.207(0.69)
<i>aglbr</i>	-0.004(0.39)	0.005(0.60)	-0.002(0.08)
<i>aginvest</i>	0.009(0.12)	0.107(0.36)	-0.117(0.32)
<i>gdppc</i>	-0.059(0.32)	0.014(0.17)	0.037(0.43)
截距项	0.072(0.96)	-0.005(0.21)	-0.056(0.90)
<i>R</i> <sup>2</sup>	0.429	0.047	0.465

\* 样本数 183。

3.2.3 偏最小二乘回归估计

鉴于最小二乘法估计的失效,本文利用随机样本交叉验证方法的偏最小二乘回归分析,随机抽取 15 个观测作为交叉验证数据集,最终结果显示抽取 2 个成分时所得的预测残差平方和的均方根达到最小,最小值为 1.011。计算得到各驱动因子对耕地、建设用地和其他用地的回归系数(表 3)。三个回归方程显著,决定系数 R2 分别达到 0.154、0.090 和 0.248。

偏最小二乘回归估计结果反映了各种因素对土地利用格局的影响。从估计结果来看,土壤酸碱度与耕地面积、建设用地面积均呈正相关关系。土壤容重与耕地面积、建设用地面积呈负相关关系,这与黄淮海地区土壤容重普遍偏高、土壤透气性较差的事实相符(表 3)。高程、地形坡度与耕地面积、建设用地面积呈负相关关系,符合坡度较大或海拔高度较高的地区较不适宜农业生产和设施建设的事实。年降水总量、年平均气温与耕地面积呈正相关关系,与建设用地面积呈负相关关系,符合黄淮海地区普遍干旱和农业生产力随气温有所上升的基本事实。城镇人口密度与耕地和建设用地面积均呈正相关关系。农村劳动力数量与耕地面积呈正相关,与建设用地面积呈负相关,说明黄淮海地区的用地矛盾主要是人口增长带来的建设用地紧张所致。农业总投资与耕地面积、建设用地面积间呈正相关关系,且主要表现为对建设用地的促进作用,说明黄淮海地区农业投资的去向主要用以改善了农业基础设施建设。人均 GDP 与建设用地呈正相关关系,基本符合黄淮海地区的实际情况,并反映出黄淮海地区城乡差距依然存在的事实。另外,上述回归系数显然满足等式(26)。

4 结论与讨论

通过开展数据的非零化处理和对称 Logratio 变换,本文构建了土地利用栅格面积成分数据,并采用偏最小二乘回归估计方法,解决了各种土地利用面积的回归结果不等于栅格面积之和的问题,避免了解释变量之间的共线性对回归结果的影响,得出了土地利用格局解释模型的稳健估计。研究结果表明,相对于普通最小二乘回归法,基于栅格成分数据的偏最小二乘回归法能够更好地处理土地利用格局解释模型的稳健性估计问题。虽然该方法会在

表 3 黄淮海地区土地利用格局驱动模型  
偏最小二乘回归的稳健估计 \*

Tab.3 Estimates of the partial least square regression for the determinant model of land use pattern in Huang-Huai-Hai Plain

	<i>cropld</i>	<i>bua</i>	<i>restld</i>
<i>soilpH</i>	0.053	0.232	-0.284
<i>soilbulk</i>	-0.017	-0.117	0.134
<i>elevation</i>	-0.003	-0.002	0.005
<i>slope</i>	-0.003	-0.002	0.005
<i>temp</i>	0.003	-0.001	-0.002
<i>rainfall</i>	0.001	-0.001	0.001
<i>urpopden</i>	0.002	0.002	-0.005
<i>aglbr</i>	0.001	-0.001	0.001
<i>aginvest</i>	0.001	0.001	-0.001
<i>gdppc</i>	-0.001	0.001	-0.001
截距项	4.226	7.812	-12.038
<i>R</i> <sup>2</sup>	0.154	0.090	0.248

\* 样本数 183。

一定程度上降低模型的决定系数,但从黄淮海地区案例研究的结果来看,估计出的系数仍足以满足预测、分析的要求,并稳健地揭示了各种驱动因子影响土地利用格局的机理。

面积成分数据与偏最小二乘回归方法在解释以栅格面积成分表征的土地利用格局驱动机理时十分有效,然而在推广应用时不能忽视其在模型设定中的各种假设。栅格中任何一种类型的土地利用面积均不为零的假设意味着这一估计方法更适用于在空间尺度较大的栅格上使用。尽管如此,该方法仍可以为获得稳健估计结果,从而较为准确地解释了各种驱动因子影响土地利用格局的驱动机理提供重要借鉴。可以预见,这一估计方法必将在以栅格面积成分表征的土地利用格局解释模型构建及估计中得到愈加广泛的应用。

参考文献

[1] 刘纪远, 庄大方, 张增祥, 等. 中国土地利用时空数据平台建设及其支持下的相关研究. 地球信息科学, 2002, (3): 3-7.

[2] 陈述彭, 周成虎, 陈秋晓. 网格地图的新一代. 测绘科学, 2004, 29(4): 1-4.

[3] 邓祥征. 土地系统动态模拟. 北京: 大地出版社, 2008.

[4] 李秀彬. 全球环境变化研究的核心领域: 土地利用/土地覆盖变化的国际研究动向. 地理学报, 1996, 51(6): 553-557.

[5] 陈百明. 试论中国土地利用和土地覆被变化及其人类驱动力研究. 自然资源, 1997, 19(2): 31-36.

[6] 蔡运龙. 土地利用土地覆被变化研究: 寻求新的综合途径. 地理研究, 2001, 20(6): 645-652.

[7] 鲁奇,战金艳,任国柱. 北京近百年城市用地变化与相关社会人文因素简论. 地理研究, 2001, 20(6): 688~696.

[8] 邓祥征. 土地用途转换分析. 北京: 大地出版社, 2008.

[9] Ferrers N M. An elementary treatise on trilinear coordinates. London: Macmillan, 1866.

[10] Bacon-Shone J. Modelling structural zeros in compositional data // Thió -Henestrosa S, Martín -Fernández J A. Proceedings of CODAWORK'03, the First Compositional Data Analysis Workshop. Spain: University of Girona, 2003, 15-17.

[11] Aitchison J. The statistical analysis of compositional data. London: Chapman and Hall, 1986.

[12] 张尧庭. 成分数据统计分析引论. 北京: 科学出版社, 2000.

[13] Frisch R. Statistical Confluence Analysis by Means of Complete Regression Systems, Economic Institute, Pub. No. 5, Oslo University, 1934.

[14] Wold S, Albano C, Dunn W J III, et al. Pattern recognition: finding and using patterns in multivariate data. In: Martens H and Russwurm H Jr (Eds.), Food Research and Data Analysis. London: Applied Science Pub. 1983, 147-188.

[15] 高惠璇. 应用多元统计分析. 北京大学出版社, 2005.

[16] Anselin L, Florax R, Rey S. Advances in spatial econometrics: Methodology, tools and applications. Berlin: Springer Verlag, 2004.

Robust Estimation for the Determinant Model of Land use Pattern  
Based on the Area Percentage Dataset of Land Uses

DENG Xiangzheng<sup>1</sup>, LIN Yingzhi<sup>1</sup>, ZHAN Jinyan<sup>2</sup>, HE Shujin<sup>1</sup>

(1. Institute of Geographic Sciences and National Resources Research, CAS, Beijing 100101, China;

2. State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing 100875, China)

**Abstract:** Given that the ordinary least square estimate is not adaptable to analyzing the driving mechanism of land use pattern represented by the area percentage data, this paper develops an estimation approach according to the partial least squares regression algorithm capable of analyzing the driving mechanism of land use pattern identified at the pixel specific resolution. The approach can be used to obtain the robust estimates for the determinant model of land use pattern and guarantee that the sum of predicted area of all kinds of land use categories equals the total land area of each grid pixel in addition to deducting the estimation bias resulting from the multi-collinearities between explanatory variables. This paper elaborately deduces the estimation algorithm for the explanatory model for the land use pattern according to the partial least-squares regression algorithm and based on the area percentage data. By using this kind of estimation approach, this paper analyzes the driving mechanism of land use pattern for cultivated land, built-up area and other kinds of land and obtains the estimates with high goodness of fit in Huang-Huai-Hai Plain of China. The research result shows that the partial least squares regression analysis can be used to estimate the driving mechanism of land use pattern represented by the area percentage data. The estimation approach proven applicable in this study is with unbiased, efficient and robust characteristics and worthy of being promoted to use in relevant case studies.

**Key words:** area percentage data; partial least squares regression analysis; land use pattern; driving mechanism; Huang-Huai-Hai Plain