DOI:10.13203/j.whugis1993.03.015

第18卷第3期 武汉测绘科技大学学报 Vol. 18 No. 3
1993年9月 Journal of Wuhan Technical University of Surveying and Mapping Sept. 1993

人口函数的数据处理方法及其抗差估计的应用

黄幼才 黄 坚
(武汉测绘科技大学工程测量系,武汉,430070)

摘 要 为了提高人口数据分析和函数模型预测的可靠性,本文分别从纯化人口统计数据和合理选择拟合函数两个方面讨论了提高人口函数拟合质量的途经。在人口数据处理和函数拟合中引入了抗差估计理论,增强了人口函数的抗干扰的能力。对人口拟合函数作了一些特殊处理,提高了函数的拟合度。所得结果的精度和可靠性优于传统的人口数据处理方法。

关键词 人口控制论;人口发展方程;抗差估计;等价权
分类号 C921

0 引 言

定量人口研究主要有以下三方面的内容:1)人口状态分析,即人口按年龄分布状况,分析劳动人数,青少年人数与老年人数的比例等;2)人口短期与长期预测,分析在某种假定的情况下,人口状态可能产生的变化情况;3)人口稳定性分析及人口控制过程的优化。我们不仅要知道可能达到的人口状态,更重要的是希望能控制人口发展的进程,使其渐次达到理想人口状态,即人口总数适中的稳态人口结构。

由于各种原因,人口统计数据中不可避免地存在着粗差和异常值。人口函数模型是人口数据分析和预测的基础。建立一个符合客观规律的人口函数模型是保证人口数据分析质量之关键。人口统计数据中的粗差或异常值会严重地损害人口函数拟合的可靠性。此外,拟合函数选择不当也同样降低拟合精度。下面将从这两个观点出发,讨论如何建立一个符合客观情况的稳键人口函数模型。

本文根据"人口控制论"的方法对人口进行定量分析。人口控制论方法是从人口的出生,死亡和迁移三大方面来分析人口状态的变化的。它的基础是建立"人口发展方程"。利用人口发展方程可以进行人口状态分析和人口预测。人口控制论也是以人口发展方程为基础的。人口发展方程的数学表达式[1]如下:

$$\begin{cases} \dfrac{\partial p(a,t)}{\partial a}+\dfrac{\partial p(a,t)}{\partial t}=-\mu(a,t)p(a,t) \\ p(a,o)=p_0(a) \\ p(o,t)=\varphi(t)=\beta(t)\displaystyle\int_{a_2}^{a_1}k(a,t)h(a,t)p(u,t)da \end{cases} \tag{1}$$

收稿日期:1992-05-05.黄幼才,男,49岁,博士,教授,现从事抗差估计理论、信息系统和工程摄影测量数字处理的研究。

其中 $p(a,t)$ 为人口状态，即人口按年龄分布，a 表示年龄，t 表示年代；$\mu(a,t)$ 为死亡率函数；$k(a,t)$ 为妇女比例函数；$h(a,t)$ 为妇女生育模式。三者结合称人口函数。为了使该模型的人口预测和人口最优控制信息更为可靠，作者在人口数据处理中采用了抗差估计理论并对人口函数作了一些数学处理。下面详细叙述人口函数的数据处理方法及其结果。

1 妇女比例函数

妇女比例函数 $k(a,t)$ 为 t 年代年龄为 a 岁的妇女人数与同年代年龄为 a 岁的总人口数之比。

现有武汉市1981，1988，1989三年人口数据，现用这三年的数据对人口函数的处理加以说明。人口函数的数据处理目的是寻找体现发展趋势的稳定曲线。

图1是武汉市1988～1989三年妇女比例函数分布图。从图中可以看出妇女比例在[1,60]区间内基本上稳定在某一常数(48.25%)附近。区间[60,100]妇女比例随年龄逐渐增高。由于妇女比例函数仅用于计算出生率，即仅与生育区间的妇女有关，故取[1,60]区间上的数据进行处理。

以1988年数据为例，数据中有明显的异常点 A 和 B。考虑人口函数用于预测或最优人口控制，该函数应反映普遍规律，故应排除 A，B 两点对妇女比例函数的影响。

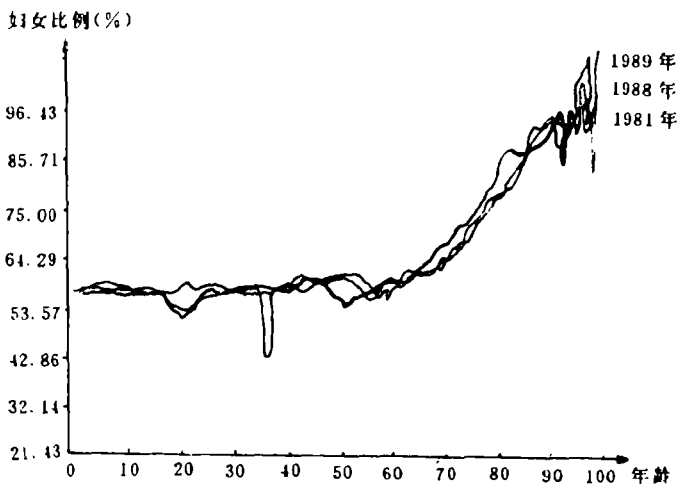

图1 三年妇女比例函数图

如果数据严格地服从某一标准分布，则传统的极大似然估计就是参数的最优估计，其目标函数为

$$\sum_{i=1}^{n}[-\ln f(x_i)] = \min \qquad (2)$$

其中 f 是随机变量 $(x_1,x_2,\cdots,x_n)$ 的概率密度，如果数据不严格服从某一标准分布或标准分布受到粗差或异常值的污染，通过式(2)不可能获得最优估值，甚至估值是完全错误，因为式(2)不具备抗差的能力。为了避免经典估计这个缺陷，用函数 ρ 代替上式中的 $\ln f$ 使其广义化，则有

$$\sum_{i=1}^{n}\rho(x_i,T_n) = \min \qquad (3)$$

于是估计量 T_n 满足方程

$$\sum_{i=1}^{n}\psi(x_i,T_n) = 0 \qquad (4)$$

这就是所谓的 M 估计。ρ 的选择应根据解决问题的性质而定。不同的 ρ 定义了不同的 M 估计，因而也就具备不同的抗差能力和效率。根据妇女比例函数的特点，我们选用了 Tukey 双权估计方法，因为该估计的 ψ 函数是平滑截尾函数，它能保持大部分正常观测值不变，完全排除粗。又因为是连续函数，函数计算不会出现急剧变化，这符合人口数据特点。Tukey 双权估计目标函数为：

$$\rho(u)=\begin{cases}\frac{1}{6}(1-(1-u^2)^2) & |u|\leqslant 1\\ \frac{1}{6} & |u|>1\end{cases} \tag{5}$$

相应的 ψ 函数为

$$\psi(u)=\begin{cases}u(1-u^2)^2 & |u|\leqslant 1\\ 0 & |u|>1\end{cases} \tag{6}$$

式中

$$u_i=\frac{x_i-T_n}{CS_n}$$

式中 x_i 为观测值，T_n 是由 n 个观测值求出的参数估值，S_n 为尺度参数。u_i 实际上是标准化后的余差，目的是减少由于观测值尺度不一致对平差估值不利的影响。

Tukey 双权估计的计算方法一般采用迭代趋近法。参数估值 T_n 的质量（抗差能力，优效性）取决于下面三个因子的选取：

1. 初始值 $T_n^{(0)}$ 的选取。为了保证迭代解正确地收敛于真值的附近，$T_n^{(0)}$ 应具有最强的抗差能力。这里选取 $T_n^{(0)}=\mathrm{med}\{x_i\}$，即取 x_i 的中位数作为迭代初始值，它的污染崩溃率为50%。考虑到妇女比例函数不受年龄影响，故数据容量 n 取60，即区间[1,60]内的数据。

2. S_n 的选取。用 S_n 使余差标准化的目的是使数据的尺度统一。因此，S_n 同样应具备很强的抗差能力。本文取中位绝对差 $S_n^{(0)}=MAD=\mathrm{med}\{|x_i-T_n^{(0)}|\}$，它和中位数具有同样的抗差能力。

3. 调制常数 C 的选取。C 值的选取关系到估值 T_n 的抗差能力和效率。C 值过小，排除的观测值多，虽然抗差能力增强，但失去的有效信息过多，效率下降。本项研究取 $c=9$。也就是说，如果数据是正态的话，大于6倍标准差的观测值看作是粗差（$9\times 0.745\approx 6$）。如果采用修正权法，则 Tukey 双权估计的迭代方程可写为初始值：

$$T_n^{(0)}=\mathrm{med}\{x_i\}\qquad S_n^{(0)}=\mathrm{med}\{|x_i-T_n^{(0)}|\}$$

迭代方程：

$$T_n^{(m+1)}=\frac{\sum\limits_{i\neq j}((1-(u_i^{(m)})^2)^2 x_{i(|u_i^{(m)}|\leqslant 1)}+O_{(|u_j^{(m)}|>1)})}{\sum\limits_{i\neq j}((1-(u_i^{(m)})^2)^2_{(|u_j^{(m)}|\leqslant 1)}+O_{(|u_j^{(m)}|>1)})}$$

式中 $u_i^{(m)}=\frac{x_i-T_n^{(m)}}{CSn^{(m)}}$。

数据处理结果如图2所示，实线代表 Tukey 双权估计，消除了异常值 A 和 B 对妇女比例函数不利的影响。虚线表示常规的最小二乘估计，异常点 A 和 B 对拟合直线有很强的吸引，使直线向下平移，妇女比例明显下降，不符合妇女比例函数的一般规律。

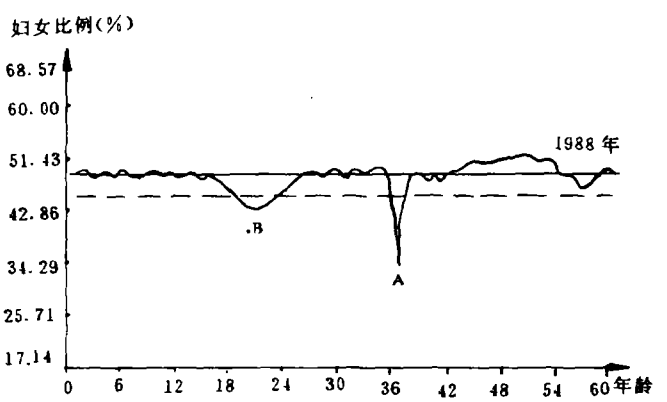

图 2　妇女比例 *Tukey* 双权拟合函数

2　死亡率函数

人口控制论中使用的前向死亡

率 $\mu(a,t)$。它是 t 年代死亡人数与 $t-1$ 年代总人口数之比。死亡率函数的特点是一盆形曲线，随着时间的推移，有些微下降的趋势。图3是武汉市1981，1988，1989三年死亡原始数据分布图。

因盆形曲线难以用某一曲线拟合，故采用分段拟合的办法。分1－2，2－3，3－4，4－5四段进行拟合。根据曲线走向，可选用下面几种拟合曲线。

1. $y=\beta_0 e^{\beta_1 x}$
2. $y=(\beta_0+\beta_1 x)^{-1}$
3. $y=\beta_0+\beta_1 e^{x}$
4. $y=\beta_0 x^{\beta_1}$
5. $y=x/(\beta_0+\beta_1 x)$
6. $y=\beta_0+\beta_1 \ln x$
7. $y=\beta_0+\beta_1 x^2$
8. $y=\beta_0 e^{\beta_1/x}$

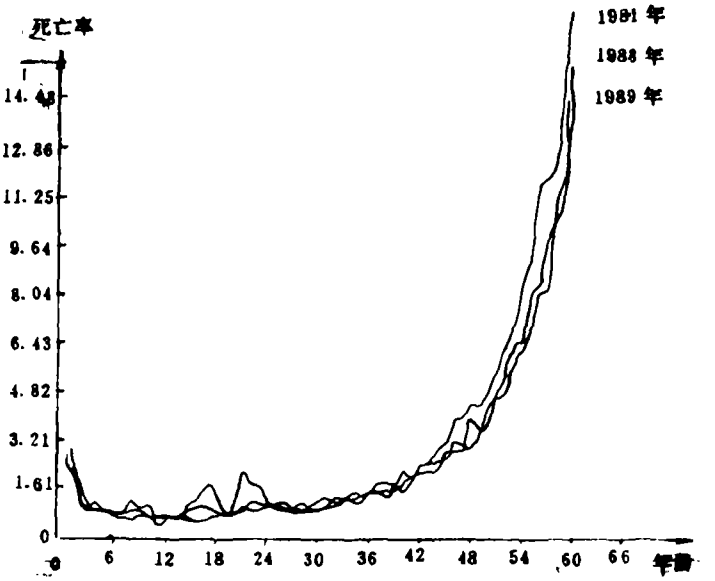

图3 武汉三年人口死亡率曲线

取1981，1988，1989三年数据对1－2，3－4，4－5三段分别用上述函数进行最小二乘似合（2－3段的趋势明显为线性函数）。拟合函数的选取依据下面两个统计量来确定，

$$\text{标准差：}S^2=\frac{\sum(y-\hat{y})^2}{n-2} \qquad \text{相关指数：}R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

式中 $\hat{y}$ 是 y 的最小二乘平差值，$\bar{y}$ 是 y 的平均值。S 越小，R^2 越接近于1，则说明拟合效果越好。通过三年数据计算，得出一致结果，均为

1－2段：$y=a+b\ln x$　　3－4段：$y=x/(ax+b)$　　4－5段：$y=ax^b$

因此，死亡率盆形曲线可用

$y=a+b\ln x$，$y=a+bx$，$y=x/(ax+b)$和 $y=ax^b$ 四段曲线来拟合。

用 S^2和 R^2来衡量拟合质量的前提是数据中没有异常值。为了提高拟合曲线抗干扰的能力，需要用抗差估计理论对最小二乘拟合抗差化。最小二乘回归可以通过权函数来实现抗差化。设抗差回归的目标函数为

$$\sum_{i=1}^{n}\rho(y_i-\sum x_{ij}\theta_j)=\min \tag{7}$$

或

$$\sum_{i=1}^{n}\varphi(y_i-\sum x_{ij}\theta_j)x_{ik}=0 \qquad (k=1,2,\cdots t)$$

式中的 t 为未知数的个数。由经典的平差理论可以导出等价权因子为

$$w(u)=\frac{\psi(u)}{u} \tag{8}$$

故 Tukey 双权估计的权因子为

$$w(u)=\frac{u(1-u^2)^2}{u}=(1-u^2)^2 \tag{9}$$

Tukey 双权估计的等价权可写为

$$\bar{p}_{ii}=\begin{cases}(1-u_i^2)^2 p_i & u_i\leqslant 1\\ 0 & u_i>1\end{cases}$$

式中 p_i 为观测值的权。

抗差化最小二乘拟合的计算步骤如下：

1. 计算被估参数 θ 的初始值 $\theta_j^{(0)}$；

a. 将与 θ_j 有关的观测值分成 m 组，每组包含的观测不完全相同。每组算出一个 $\theta_j^{(0)}$，共算得 m 个 $\theta_j^{(0)}$；

b. 将 m 个 $\theta_j^{(0)}$ 按大小排列成顺序统计量。得 $\theta_{j(1)}^{(0)},\theta_{j(2)}^{(0)},\cdots,\theta_{j(m)}^{(0)}$，取中位数得初始值 $\theta_j^{(0)}$；

2. 计算余差 $V^{(0)}=A\theta^{(0)}-L$ 和单位权中误差

$$\sigma_0^{(0)}=\frac{V^{(0)T}\bar{p}V^{(0)}}{n-t}$$

3. 求等价权 $\bar{p}=\mathrm{dia}(\bar{p}_{11},\bar{p}_{22},\cdots,\bar{p}_{nn})$；

4. 组成法方程式求得 θ 的第一次迭代解 $\theta^{(1)}$；

5. 用 $\theta^{(1)}$ 代替 $\theta^{(0)}$，重复步骤2—4直至 θ 的两次迭代值之差小于容许值为止。

现以死亡函数中1—2段拟合曲线为例来说明抗差最小二乘拟合的计算方法。设选取的拟合函数为

$$y=a+b\ln x \qquad 或 \qquad y=\begin{bmatrix}1 & \ln x\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}$$

其矩阵表达式为

$$Y=A\theta \qquad Y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix},\quad A=\begin{bmatrix}1 & \ln x_1\\ 1 & \ln x_2\\ \vdots & \vdots\\ 1 & \ln x_n\end{bmatrix},\quad \theta=\begin{bmatrix}a\\ b\end{bmatrix}$$

每次按顺序删除一个点，进行常规的最小二乘拟合求得 a_i,b_i。若有 n 个数据，则有 $n-1$ 组参数计算值。取 a_i,b_i 的中位数为迭代计算的初始值：$a^{(0)}=\mathrm{med}\{a_i\},b^{(0)}=\mathrm{med}\{b_i\}$。等价权的初始值取1，调制常数取9。具体计算步骤如下：

1. 计算初始值 $S_n^{(0)},u_i^{(0)},\bar{p}^{(0)}$

$$\bar{p}_{ii}^{(1)}=1,\qquad \bar{p}_{ij}^{(0)}=0$$

$$u_i^{(0)}=\frac{y_i-(a^{(0)}+b^{(0)}\ln x_i)}{9S_n^{(0)}}$$

$$S_n^{(0)}=\mathrm{med}\{y_i-(a^{(0)}+b^{(0)}\ln x_i)\}$$

2. 按最小二乘法求 θ 的值和余差

$$\hat{\theta}=(A^T\bar{p}A)^{-1}A^T\bar{p}L \qquad V=A\hat{\theta}-L$$

3. 由余差计算迭代权

$$u_i=\frac{V_i}{CS_n}$$

$$\bar{p}_{ii}=\begin{cases}(1-u_i^2)^2 P_i & u_i\leqslant 1\\ 0 & u_i>1\end{cases}$$

本例计算中取 $p_i \equiv 1$。

4. 用新的等权求得 θ 和 v 的新的计算值。如此迭代直至收敛到允许值为止。图4是用抗差最小二乘回归拟合方法所得的死亡率函数曲线。可以看出拟合的效果很好。该曲线已自动地排除了异常 值的干扰，如果人为地加入粗差也不会使这条曲线发生显著变化。图中虚线是常规的三段直线拟合法，显然传统的方法与实际情况有明显的差异。

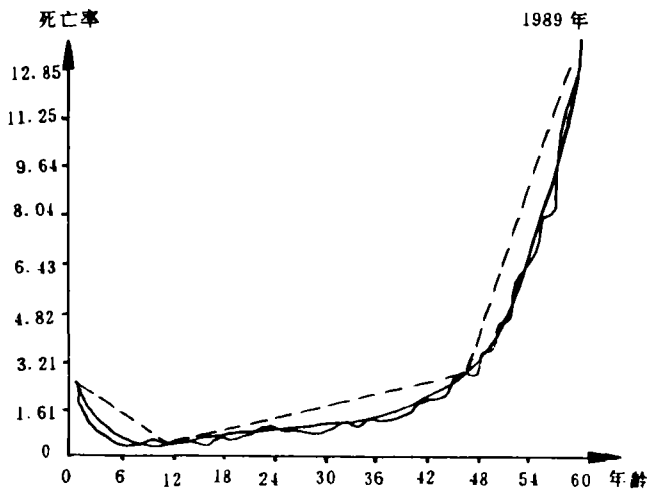

图4 死亡率函数拟合曲线

3 妇女生育模式函数

妇女生育模式函数 $h(a,t)$ 是计算出生率的一个重要函数。其含义可以理解为 t 年代年龄为 a 岁的妇女生育的概率。若 $h(a,t)=\max h$，则 a 岁为生育峰值年龄，或者说 t 年代生育妇女中，年龄为 a 岁的妇女最多。人口控制论中介绍，生育模式近于一 Gamma 分布函数曲线，用数学式子可表示为

$$h(a,t)=\begin{cases}\dfrac{(a-a_1)^{\sigma-1}e^{-\frac{a-a_1}{\theta}}}{\theta^{\sigma}\Gamma(\sigma)}, & a>a_1\\ 0, & a\leqslant a_1\end{cases}$$

式中 a_1 为最小生育年龄。

图5是1981年生育模式原始数据图。观其图形确实近于 Gamma 曲线，故可以直接用该曲线进行拟合。这类问题属于非线性曲线拟合，故首先要拟合函数线性化，其方法简述如下：设非线性模型一般表达式为

$$Y=f(x,\theta)+\xi$$

在 $\theta=\theta_0$ 处将 $f(\theta)$ 按泰勒级数展开，只取前二项得

$$f(\theta)=f(\theta_0)+J(\theta_0)(\theta-\theta_0)$$

式中

$$f(\theta)=(f_1(\theta),f_2(\theta),\cdots,f_n(\theta))$$

$J(\theta_0)$ 是 $n<p$ 阶雅可比矩阵，其表达式为

$$J(\theta_0)=\begin{bmatrix}\frac{\partial f_1(\theta)}{\partial\theta_1} & \frac{\partial f_1(\theta)}{\partial\theta_2} & \cdots & \frac{\partial f_1(\theta)}{\partial\theta_p}\\ \frac{\partial f_2(\theta)}{\partial\theta_1} & \frac{\partial f_2(\theta)}{\partial\theta_2} & \cdots & \frac{\partial f_2(\theta)}{\partial\theta_p}\\ \vdots & \vdots & \vdots & \vdots\\ \frac{\partial f_n(\theta)}{\partial\theta_1} & \frac{\partial f_n(\theta)}{\partial\theta_2} & \cdots & \frac{\partial f_n(\theta)}{\partial\theta_p}\end{bmatrix}$$

于是有

$$\theta=\theta_0+[J^T(\theta_0)J(\theta_0)]^{-1}J^T(\theta_0)[Y-f(\theta_0)]$$

故得逆推公式

$$\theta_{i+1}=\theta_i+[J^T(\theta_i)J(\theta_0)]^{-1}J^T(\theta_i)[Y-f(\theta_i)]$$

协方差矩阵为

$$\mathrm{cov}(\theta)=\sigma^2[J^T(\theta)J(\theta)]^{-1}$$

$$\sigma^2=\frac{1}{n-p}S(\theta)\qquad S=[Y-f(\theta)]^T[Y-f(\theta)]$$

利用武汉市1981年数据，自变量设为 a,θ。取 $a_1=21$，拟合结果如图5中虚线所示。

由拟合结果可以看出，拟合曲线与实际值附合较好，但拟合生育峰值比实际值偏低。这种差异在人口分布呈峰形时会有较大影响，故建议在拟合函数中加一乘变量，新的拟合方程为

$$h(a,t)=\begin{cases}\dfrac{k(a-a_1)^{a-1}e^{-\frac{a-a_1}{\theta}}}{\theta^a\Gamma(a)}, & a>a_1\\ 0, & a\leqslant a_1\end{cases}$$

此时回归变量为 k,a,θ 三个。同理可以拟合出曲线方程。

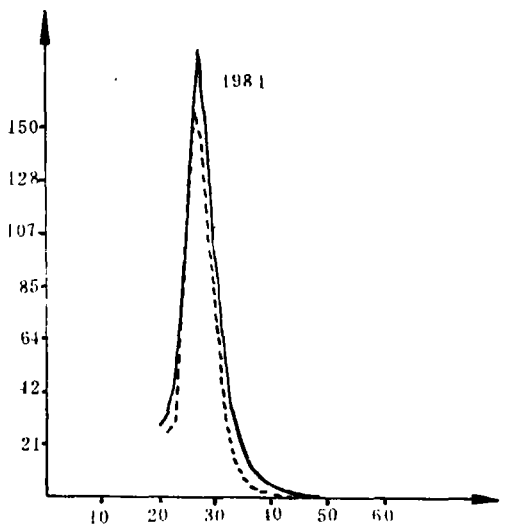

图5　妇女生育模式函数拟合图

4　结　语

人口函数在定量人口学中占有举足轻重的地位。人口函数的拟合质量关系到人口数据分析的可靠性。本文从拟合函数的选取和排除数据中异常值对拟合函数不利影响两个方面探讨了提高函数拟合质量的途径。实际上人口函数所呈现的状况是极为复杂的，需要通过大量的数据分析才能确定普遍规律。由于我国人口数据统计工作近几年才逐渐规范化，故本文所得的人口函数模型受到数据量不大的局限。本文侧重于人口函数处理的理论和方法的研究。

参　考　文　献

1　宋健等. 人口控制论. 北京：科学出版社，1985. 48

2　黄幼才. 数据探测与抗差估计. 北京：测绘出版社，1990.

3 方开泰. 实用回归分析. 北京:科学出版社,1988.

4 袁嘉新等. 系统论在区域规划中的应用. 社会科学文献出版社,1987.

5 黄幼才. Application of Robust Estimation in Close-Range Photogrammetry. Photogrammetric Engineering and Remote Sensing. 1987,53(2):171-175

6 Huber P J. Robust Statistics. John Wiley and Sons,1981.

7 Hample F R,et al. . Robust Statistics. John Wiley and Sons,1986.

Approaches for Population Data Processing and Application of Robust Estimation to Simulating the Population Models

Huang Youcai *Huang Jian*

(Dept. of Engineering Surveying, WTUSM Louyu Road 39, Wuhan,China,430070)

Abstract In order to improve the reliability of population data analysis and the prediction of population models,this paper presents various methods for enhancing the quality of simulation of the population functions based on cleaning statistic data and selecting suitable mathematical functions. The introduction of robust estimators to data processing and simulation could strengthen the resistance of the population models against outliers or gross errors. The accuracy of the simulation could be improved by choosing proper mathematical functions. The research in this article shows that the accuracy and reliability of population models derived from the suggested methods are better than classical approaches.

Key words theory of population control; equations of population development; Robust estimation; equivalent weight