doi:10.3969/j.issn.1671-637X.2013.06.016

基于双向检验的异常数据剔除与修复方法

姜大治， 韩先平
（中国人民解放军92941部队，辽宁 葫芦岛 125000）

摘 要：针对靶场复杂的试验环境及测量数据的特点，深入分析了异常数据产生的原因，提出了一种利用多项式拟合对测量结果数据进行双向检验，剔除异常数据的方法，并在检验结果满足修复条件时，对剔除数据进行加权修复。实际数据测试表明：本方法能够有效克服数据突变和段落性阶跃等干扰因素的影响，具有较好的模型稳定性和较高的异常数据剔除率，满足试验任务实际需求，具有较高的工程应用价值。

关键词：数据预处理；异常数据剔除；数据修复
中图分类号：V27；TN9 **文献标志码**：A **文章编号**：1671-637X(2013)06-0070-05

Abnormal Data Eliminating and Repairing Method Based on Two-Sided Test

JIANG Dazhi, HAN Xianping
(No.92941 Unit of PLA, Huludao 125000, China)

Abstract: In view of the complex test environment and characteristics of measurement data, through a deep analysis of the causes of abnormal data, a method for eliminating abnormal data based on two-sided test of measurement result data by use of polynomial fitting was presented. When the test result met the repairing condition, weighted repairing was conducted for the eliminated data. Real data test showed that this method can effectively overcome the influence of data mutation and paragraph step, has better model stability and higher abnormal data eliminating rate, which can meet the actual requirement of experimental task and is of high practical value in engineering.

Key words: data preprocessing; abnormal data eliminating; data repairing

0 引言

在靶场试验中，目标轨迹由光学测量、雷达测量和卫星测量等装备完成，它们共同构成了天地一体化的立体测量系统。整套系统连续实时地提供目标高精度的位置信息和速度信息，是增强靶场试验外弹道测量能力的有效手段[1-3]。特别是随着卫星导航与测量技术的迅猛发展，其在目标轨迹测量中的应用也不断深入。但由于靶场试验环境复杂，参试装备较多，各种干扰因素对数据影响较大，导致外弹道信息中含有大量异常数据。数据预处理作为数据处理的重要组成部分，其效果如何直接决定了数据处理结果的质量和精度[4-5]。异常数据剔除与修复是数据预处理的重点和难点，经过对以往大量实测数据的分析和总结，针对GPS测量数据具有可正、逆双向解算，以及光测图像逐画幅判读的特点，研制了利用变阶多项式拟合外推模型对数据进行双向检验，综合两个方向的检验结果完成异常数据剔除的方法，并在条件满足时完成数据修复。实际数据测试表明，本方法能够有效剔除独立及成片异常数据，修复精度满足任务需求。

收稿日期：2012-06-25

作者简介：姜大治(1977—)，男，辽宁葫芦岛人，硕士生，工程师，研究方向为GPS数据处理，外测数据建模与处理。

1 异常数据来源分析

在靶场试验中，由于受陆海复杂环境影响，多体制、多台套装备共同使用，以及目标高动态飞行等因素的共同制约，导致测量结果中的异常数据来源非常复杂，对不同体制测量装备的异常数据来源分析如下。

1.1 光学测量

光测装备是靶场测控系统的基石。CCD技术的出现，使光测装备的捕获、跟踪和识别能力有了质的飞跃。靶场光测装备的典型代表是光电经纬仪，主要由光学系统、测角系统、跟踪伺服系统和记录系统等部分

组成。

由于光电经纬仪是一个复杂而精密的测量系统，各环节的制造和安装缺陷以及环境和使用上的波动都会带来各种各样的误差因素。主要包括垂直轴倾斜误差、水平轴倾斜误差、照准轴误差，零位差、定向差、光学系统几何畸变误差、蒙气差等静态测量误差源和电视动态测量误差、仪器跟踪运动误差、CCD 脱靶量输出滞后误差、大气抖动误差、视轴晃动误差等动态测量误差。光电经纬仪系统误差的绝大部分可进行调整或修正，而且随机误差较小。异常数据主要来源于装备故障或操作不当；此外，当图像质量较差或目标距离装备较远时，图像判读误差较大，是异常数据的一个重要来源。

1.2　雷达测量

雷测装备作为集中了现代电子科技成就的高科技系统，是靶场测控系统的支柱。靶场雷测装备较多，测量体制包括连续波、单脉冲和相控阵等，而且作为测控装备，其波束宽度较窄。

产生雷达测量误差的原因很多，性质也不同。距离系统误差主要影响因素包括：用真空光速代替空气中的速度，忽略传播损耗和信号折叠，测量量化误差，信号处理时间延迟和信号处理方法误差，检测和点迹提取方法误差等。方位和俯仰系统误差主要影响因素包括：测量量化误差，信号处理方法误差，检测和点迹提取方法误差，天线稳定转台不完全水平、机械轴与正北及机械轴与电轴的不完全平行误差，方位编码器自身误差，波束扫描在高仰角和低仰角不完全误差。雷测装备的大部分系统误差同样可进行调整或修正，但随机误差较大。雷达测量异常数据来源较为复杂，而且常连续成批出现。各种干扰，地杂波，海杂波，目标过捷径和目标闪烁等因素都可产生异常数据。

1.3　卫星测量

卫星测量在靶场应用较多的是比较成熟的 GPS 测量系统。由于 GPS 卫星运行在 2000 km 的高空，而其发射功率不可能很大，因此 GPS 接收机在地面上接收到的信号功率很弱，容易受到来自近地空间和地面的各种电磁干扰。干扰导致 GPS 接收机的信噪比降低，使得接收机跟踪环路噪声增加，从而增加了测量误差[6-8]。GPS 定位误差按误差性质可以分为系统误差和偶然误差。其中，系统误差无论大小还是对定位结果的危害都比偶然误差大得多，但大多数系统误差有规可循，可通过各种方法消除。GPS 卫星导航定位误差主要分成 3 类。

1）GPS 信号自身误差，主要包括卫星钟差和卫星的轨道偏差。当 GPS 卫星钟差通过钟差模型改正后，表现在距离偏差约为 1 m[9]。GPS 信号自身误差主要表现为系统性，除非卫星信号异常，一般不会在定位结果中产生明显的异常数据，可通过差分处理消除[10]。

2）GPS 信号传播误差，主要包括对流层时延、电离层时延、相对论效应、地球自转、多路径及信号干扰影响等。由于靶场外场试验参试装备较多，且功率较大，试验环境复杂，难以满足 GPS 接收机需要的电磁环境和开阔空间，微弱的 GPS 信号极易受到干扰，而且 GPS 天线周围经常有各型装备及设施，多路径影响复杂。信号干扰及多路径是异常数据最主要的来源[11-13]。

3）GPS 接收机及天线所产生的信号测量误差，主要包括观测噪声、内时延、天线相位中心误差等。试验时，GPS 接收机都装配于载体中，由于载体启动加速度较大，运动速度较快，且部分载体具有高动态或运动姿态变化较大的特点，可能导致观测噪声较大，严重时可造成信号失锁，是异常数据另一个主要来源。

2　异常数据剔除与修复

2.1　方法模型

设目标参数的离散观测量为 $y_i(i=1,2,\cdots,n)$，对应的时间序列为 $t_i(i=1,2,\cdots,n)$。为进行异常值识别，需保证前 m 项数据正常。当采样频率足够高时，相邻采样值之间差距很小，因此可利用三阶差分模型进行连续 4 点可用性判断。

设连续 4 点观测量为 $y_i, y_{i+1}, y_{i+2}, y_{i+3}$，三阶差分公式为

$$\Delta_i = y_i - 3y_{i+1} + 3y_{i+2} - y_{i+3} \tag{1}$$

如果 $|\Delta_i| < \delta$，则认为选取信息中无异常数据，可作为起始拟合数据。否则选取 $y_{i+1}, y_{i+2}, y_{i+3}, y_{i+4}$ 继续检验，直到满足条件。δ 为判断门限，当测量数据质量较好时，δ 取元素测量精度 σ 的 3 ~ 5 倍，当测量数据质量较差，丢点较多时，可适当放宽门限。当测量精度超差或精度未知时，可利用样条平滑方法对参数残差 $\hat{\sigma}$ 进行估计。公式为

$$\hat{\sigma} = \sqrt{\frac{n\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 - \left(\sum_{i=1}^{n}(y_i - \tilde{y}_i)\right)^2}{n(n-1)}} \tag{2}$$

式中：n 为目标参数测量数量；$\tilde{y}_i$ 为目标参数平滑值。

由于三阶差分只能判断连续 4 点中是否有异常数据，因此要利用滑窗连续检验，直到连续正常数据数量满足拟合递推点数为止。

设正向拟合点数为 m_f，拟合阶数为 l_f，正向递推公式为

$$\begin{cases} [a_{f0} \quad a_{f1} \quad \cdots \quad a_{fl}]^{\mathrm{T}} = (\boldsymbol{B}_f^{\mathrm{T}}\boldsymbol{B}_f)^{-1}\boldsymbol{B}_f^{\mathrm{T}}\boldsymbol{U}_f, \ i=k+1,k+2,\cdots, \ m_f>2 \\ \hat{y}_{fi} = a_{f0} + a_{f1}t_i + \cdots + a_{fl}t_i^l, \ i=k+1,k+2,\cdots, \ m_f>2 \\ \hat{y}_{fi} = \dfrac{y_k - y_{k-1}}{t_k - t_{k-1}}t_i + \dfrac{t_k y_{k-1} - y_k t_{k-1}}{t_k - t_{k-1}}, \ i=k+1,k+2,\cdots, \ m_f=2 \end{cases} \tag{3}$$

式中：

$$\boldsymbol{B}_f=\begin{bmatrix}1 & t_{k-m_f+1} & \cdots & t_{k-m_f+1}^{l_f}\\ 1 & t_{k-m_f+2} & \cdots & t_{k-m_f+2}^{l_f}\\ \vdots & \vdots & & \vdots\\ 1 & t_k & \cdots & t_k^{l_f}\end{bmatrix};\boldsymbol{U}_f=\begin{bmatrix}y_{k-m_f+1}\\ y_{k-m_f+2}\\ \vdots\\ y_k\end{bmatrix}。$$

当拟合数据点数为2时，拟合方程衰变成线性。除非目标发生异动，目标轨迹方程在短时间内一般不会超过三阶，因此首先计算三阶递推，然后逐次降阶，直至衰减到线性。记结果为$\hat{y}_{fi}^{(3)}$、$\hat{y}_{fi}^{(2)}$、$\hat{y}_{fi}^{(1)}$，仍设δ为判断门限，判别下式：if $|\hat{y}_{fi}^{(j)}-y_i|<\delta, j=1,2,3$。只要有任一阶递推结果符合条件，则认为$y_i$正常，否则认为$y_i$数据可能异常，标记$y_i$，此点数据不可再用于数据递推。继续判断下一点，当再次有数据被判断正常时，记录此刻拟合外推结果与测量结果最接近时的拟合点数$m_{fi(\min)}$和拟合阶数$l_{fi(\min)}$，并重新构造外推模型。

设逆向拟合点数为m_b，拟合阶数为l_b，逆向递推公式为

$$\begin{cases}[a_{b0}\quad a_{b1}\quad\cdots\quad a_{bl}]^{\mathrm{T}}=(\boldsymbol{B}_b^{\mathrm{T}}\boldsymbol{B}_b)^{-1}\boldsymbol{B}_b^{\mathrm{T}}\boldsymbol{U}_b,\ i=k'-1,k'-2,\cdots,\ m_b>2\\ \hat{y}_{bi}=a_{b0}+a_{b1}t_i+\cdots+a_{bl}t_i^l,\ i=k'-1,k'-2,\cdots,\ m_f>2\\ \hat{y}_{bi}=\dfrac{y_{k'}-y_{k'+1}}{t_{k'}-t_{k'+1}}t_i+\dfrac{t_{k'}y_{k'+1}-y_{k'}t_{k'+1}}{t_{k'}-t_{k'+1}},\ i=k'-1,k'-2,\cdots,\ m_b=2\end{cases}\tag{4}$$

式中：

$$\boldsymbol{B}_b=\begin{bmatrix}1 & t_{k'} & \cdots & t_{k'}^{l_b}\\ 1 & t_{k'+1} & \cdots & t_{k'+1}^{l_b}\\ \vdots & \vdots & & \vdots\\ 1 & t_{k'+m_b-1} & \cdots & t_{k'+m-1}^{l_b}\end{bmatrix};\boldsymbol{U}_b=\begin{bmatrix}y_{k'}\\ y_{k'+1}\\ \vdots\\ y_{k'+m_b-1}\end{bmatrix}。$$

逆向递推和判别原理与正向同理，不再赘述。设$m_{bi(\min)}$和$l_{bi(\min)}$为逆向拟合外推检验为异常后，再次有数据被判断正常时，逆向拟合外推结果与测量结果最接近时的拟合点数和拟合阶数。

当某点数据正向与逆向都判定异常时，则说明此点数据异常。当只有一个方向判断异常，另一方向判断正常时，说明数据位置可能发生阶跃等目标运动轨迹异常现象，因此本方法亦可作为判断目标轨迹是否存在异常的一种手段。

当双向判断数据都为异常，且$m=m_{fi(\min)}=m_{bi(\min)}$，$l=l_{fi(\min)}=l_{bi(\min)}$时，数据满足修复条件（如果$m=2$，则只允许中间存在一个异常数据且数据无丢失，当连续异常数据多于两个或同时存在数据丢失现象时，无法保证拟合模型符合数据规律，不能进行数据修复）。数据修复公式为

$$\begin{cases}[a_{r0}\quad a_{r1}\quad\cdots\quad a_{rl}]^{\mathrm{T}}=(\boldsymbol{B}_r^{\mathrm{T}}\boldsymbol{B}_r)^{-1}\boldsymbol{B}_r^{\mathrm{T}}\boldsymbol{U}_r,\\ \qquad i=k-m+1,k-m+2,\cdots,k,k',\cdots,k'+m-1\\ \hat{y}_{ri}=a_{r0}+a_{r1}t_i+\cdots+a_{rl}t_i^l,i=k+1,k+2,\cdots,k'-1\end{cases}\tag{5}$$

式中：

$$\boldsymbol{B}_r=\begin{bmatrix}1 & t_{k-m+1} & \cdots & t_{k-m+1}^{l}\\ 1 & t_{k-m+2} & \cdots & t_{k-m+2}^{l}\\ \vdots & \vdots & & \vdots\\ 1 & t_k & \cdots & t_k^{l}\\ 1 & t_{k'} & \cdots & t_{k'}^{l}\\ 1 & t_{k'+1} & \cdots & t_{k'+1}^{l}\\ \vdots & \vdots & & \vdots\\ 1 & t_{k'+m-1} & \cdots & t_{k'+m-1}^{l}\end{bmatrix};\boldsymbol{U}_r=\begin{bmatrix}y_{k-m+1}\\ y_{k-m+2}\\ \vdots\\ y_k\\ y_{k'}\\ y_{k'+1}\\ \vdots\\ y_{k'+m-1}\end{bmatrix}。$$

在整个剔除与修复过程中应注意控制参数的设定，当连续判断异常数据数量或连续数据丢失数量超过最大允许值时，模型失效，应从第一个被判断为异常的数据或连续丢失段落后第一个有效数据开始重新进行判断。

设允许最多连续异常判断次数为$n_{r\max}$，允许最大连续丢失数量为$n_{l\max}$。$n_{r\max}$和$n_{l\max}$的值依据工程背景进行确定，主要考虑参数采样频率和目标运动状态等因素。当采样频率较高，目标动态变化较平稳时，可适当增大$n_{r\max}$与$n_{l\max}$值，反之应相应减小$n_{r\max}$与$n_{l\max}$值。

2.2 算法流程

1）依据数据质量设定异常数据判断门限、允许最多连续异常判断次数和允许最大连续丢失数量。

2）利用式(1)对初始数据进行可用性判断，直到连续满足条件数据数量符合要求。

3）利用式(3)构建正向拟合外推模型，进行外推判断。

4）当某点数据判断为异常时，进行降阶判断，直至衰减到两点线性外推，如果各判断结果都为异常，则认为此点数据可能异常，进行标记，否则认为数据正常。

5）当数据可能异常时，继续向后判断，当再次出现正常数据时，记录外推结果与测量结果最接近时的拟合点数和拟合阶数。将此点数据加入数据序列，并舍去第一点数据，重构拟合外推模型。

6）当被连续判断为异常的数据数量或连续丢失数据数量超过允许门限时，表明当前递推模型失效，进行分段处理，在各数据段内分别完成数据检测。

7）参照正向递推判断方法，利用式(4)完成逆向递推判断。

8）双向检测都为异常表明结果数据异常，只有单向检测异常，说明目标轨迹在此处可能存在异常，需进一步判断。

9）对满足修复条件的异常数据利用式(5)进行修复。

程序流程见图 1。

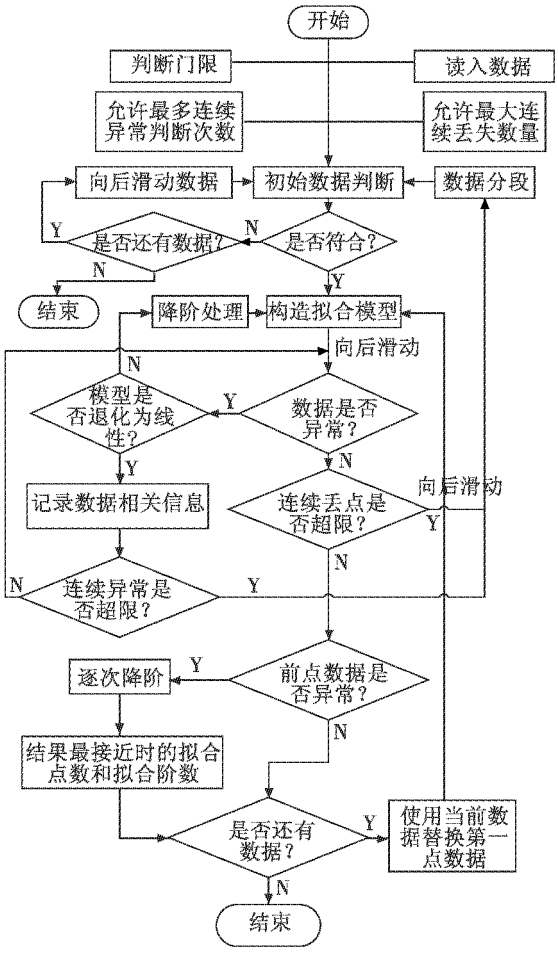

图 1　程序流程图

Fig. 1　Program flow chart

3　实例分析

以某民用目标 GPS 实测数据作为验证用例，GPS 天线安装于目标前端顶部，目标飞行滚动角度小于 15°，运动初始段目标动态变化较大，接收机短暂失锁。接收机天线安装于表层，没有多路径抑制功能，存在干扰、多路径和遮挡影响[14-15]。收星数 9 ~ 11 颗，稳定跟踪 6 颗以上，多数时间 PDOP 小于 2.3，处理模式为伪距差分定位，多普勒测速，差分距离 15 ~ 80 km。以目标在地心坐标系下的 X 方向速度参数为例，共记录 19285 个数据，数据记录间隔为 0.1 s；丢失 192 个数据，数据丢失率为 0.96%；数据中有 135 个异常数据，异常数据比例为 0.68%；异常数据剔除前残差估计值为 2.91 m/s。异常数据剔除后残差估计值为 1.08 m/s，由于残差估计值较大，说明数据受随机误差影响较大，因此，判断门限设为 9.0 m/s($3\hat{\sigma}$)。允许最多连续判断 5 次数据异常(0.5 s)，允许最多连续丢失 5 个数据(0.5 s)。数据统计情况如表 1 所示，异常数据剔除效果曲线如图 2 ~ 图 5 所示。

表 1　异常数据剔除情况汇总

Table 1　Eliminating abnormal data status summary

	方法	实判点数	漏判点数	误判点数	漏判比例/%	误判比例/%
线性模型	正向检验	383	3	251	0.74	12.10
	逆向检验	406	5	276	1.23	13.30
	双向检验	178	9	52	2.21	2.51
二阶模型	正向检验	179	35	79	8.58	3.81
	逆向检验	151	49	65	12.01	3.13
	双向检验	89	53	7	12.99	0.34
三阶模型	正向检验	784	57	706	13.97	34.02
	逆向检验	513	72	450	17.65	21.69
	双向检验	37	107	9	26.23	0.43
变阶模型	正向检验	189	4	58	0.98	2.80
	逆向检验	216	5	86	1.23	4.14
	双向检验	162	9	36	2.21	1.73

表 1 中：漏判比例表示采用某方法完成剔除计算时漏判点数占全部方法漏判点数之比；误判比例表示采用某方法完成剔除计算时误判点数占全部方法误判点数之比。

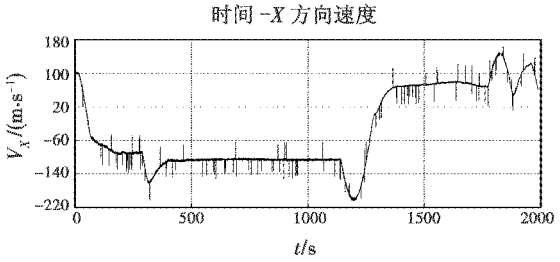

图 2　异常值剔除前参数曲线图

Fig. 2　Parameter curve before abnormal data eliminating

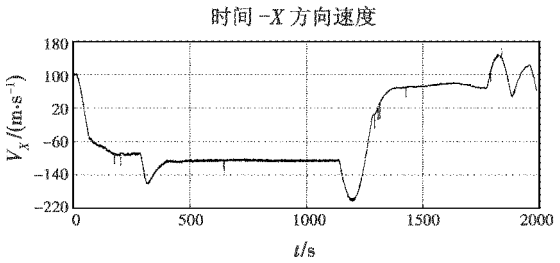

图 3　异常值剔除后参数曲线图

Fig. 3　Parameter curve after abnormal data eliminating

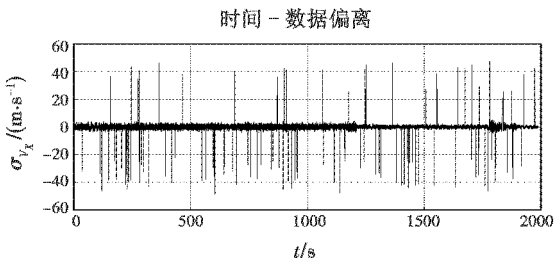

图 4　异常值剔除前残差估计曲线图

Fig. 4　Residuals estimated curve before abnormal data eliminating

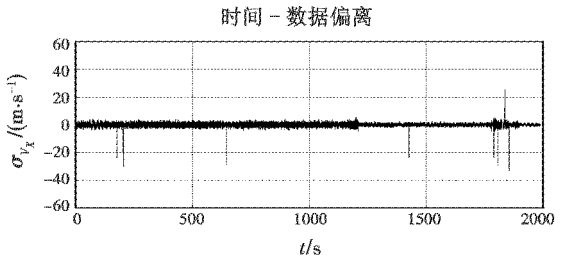

图5 异常值剔除后残差估计曲线图

Fig. 5 Residuals estimated curve after abnormal data eliminating

由数据可知,由于干扰、多径、遮挡等因素的影响,数据随机性偏差较大,数据在1239.5 s连续丢失2.1 s,并出现大约2 m/s的阶跃。在剔除效果上,二阶模型可靠性优于线性模型,但漏判现象较多,三阶模型效果最差,说明当采样频率足够高时,采用低阶模型对数据进行拟合外推是合理的。对各阶模型,双向检验结果均优于单一方向检验结果。通过统计结果可知,双向变阶模型综合效果优于各阶模型,而且不受数据频繁丢失和数据阶跃的影响,表现出较强的工程稳定性。由于数据随机误差较大,造成部分异常数据无法剔除。放宽门限值虽然可减少漏判,但误判数量增加,反之,严格门限值虽然可减少误判,但漏判数量增加。如何降低数据随机误差对模型的影响,是本方法需完善的一个方向。

4 结束语

异常数据的剔除和修复是外测数据处理的重点和难点。异常数据判断的关键不是数据偏离程度的大小,而是数据是否符合目标的运动规律。当目标动态变化较大时,数据变化同样剧烈,所以是否符合目标运动规律才是最有效、最可靠的判断依据。本文提出的双向变阶异常数据检验方法充分考虑了目标短时间内的运动特性,有效克服了数据丢失和数据阶跃等因素对模型的影响,通过综合考虑双向检验结果,大大降低了数据误判率。实际数据测试结果证明了方法的有效性和稳定性。

参考文献

[1] 夏南银. 航天测控系统[M]. 北京:国防工业出版社,2002:235-237.

[2] RIZOS C. Network RTK research and implementation: A geodetic perspective[J]. Journal of Global Positioning Systems, 2002, 1(2):144-150.

[3] 刘利生. 外测数据事后处理[M]. 北京:国防工业出版社,2000:6-10.

[4] 刘利生,吴斌,杨萍. 航天器精确定轨与自校准技术[M]. 北京:国防工业出版社,2005:24-33.

[5] EULER H J, SEEGER S, ZELZER O, et al. Improvement of positioning performance using standardized network RTK messages[C]//Proceedings of the National Technical Meeting of the Institute of Navigation, ION NTM 2004, San Diego, USA, 2004:453-461.

[6] 侯者非,王学东,陈国军. GPS干扰与抗干扰技术研究[J]. 现代电子技术,2004(23):99-101.

[7] 叶宝盛. GNSS抗干扰接收机技术研究[D]. 西安:电子科技大学,2010:12-15.

[8] WARD P W. GPS receiver RF interference monitoring mitigation and analysis techniques[J]. Navigation, 1995, 41(4):367-391.

[9] 朱祥维,肖华,雍少为,等. 卫星钟差预报的Kalman算法及其性能分析[J]. 宇航学报,2008,29(3):966-970.

[10] 刘磊,盛峥,王迎强,等. 利用广播星历计算GPS卫星位置及误差分析[J]. 解放军理工大学学报:自然科学版,2006,7(6):592-596.

[11] 刘基余. GPS卫星导航定位原理与方法[M]. 北京:科学出版社,2003:12-17.

[12] 王惠南. GPS导航原理与应用[M]. 北京:科学出版社,2003:134-139.

[13] 李天文. GPS原理及应用[M]. 北京:科学出版社,2010:120-127.

[14] 张敏虎,任章,华春红. 超级紧组合中弱GPS信号跟踪算法[J]. 电光与控制,2010,17(8):33-36.

[15] 吕艳梅,李小民,孙江生. 高动态环境的GPS信号接收及其算法研究[J]. 电光与控制,2006,13(4):24-27.