

基于数据预处理的平面不确定度评定技术研究

王乔义, 黄美发, 李向前, 鲍家定

(桂林电子科技大学 机电工程学院, 广西 桂林 541004)

摘 要: 针对目前应用三坐标测量机进行相应测量以及不确定度评定时, 其内部软件少有错误数据剔除功能致使测量数据中的粗大误差对评定结果产生一定影响的问题, 结合概率论与数理统计理论, 讨论了粗大误差的剔除技术. 分析了平面最小二乘法拟合的不确定度, 研究了结合基于数据剔除方法和最小二乘法的平面不确定度评定方法以及具体实现手段, 以平板为例对上述方法进行了实例验证.

关键词: 平面度; 不确定度; 数据剔除; 最小二乘法

中图分类号: TB9

文献标识码: A

文章编号: 1000-7180(2010)09-0147-04

Study on Technique in Flatness Uncertainty Based on Data Preprocessing

WANG Qiao-yi, HUANG Mei-fa, LI Xiang-qian, BAO Jia-ding

(School of Electromechanical Engineering, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: At present, there is a lack of method in the internal software of coordinate measuring machine (CMM) to deal with the error data, as a result, the evaluation is influenced. To solve the problem above, methods for excluding error data are discussed in this article based on probability theory and mathematical statistics. Then the least square method for calculating plane uncertainty is analyzed and details in data processing based on the least-squares fitting methods are described. At last an example is given to verify the program. The results show that the schemes discussed in this paper are feasible and have certain application prospects in CMM.

Key words: flatness; uncertainty; data elimination; least-squares method

1 引言

在计算机处理大量数据时, 数据的可靠性、有效性一直是困扰应用者的瓶颈问题. 作为新一代 GPS 标准体系中的一项重要内容, 数据处理在不确定度的计算过程中起着至关重要的作用. 为了保证测量误差以及不确定度的正确性, 使用计算机实现实验数据的快速预处理, 剔除粗大误差及错误数据是获得有效测量结果的技术手段, 也是实现新一代 GPS 标准体系下不确定度管理的必然要求^[1-2].

在对工件进行检测的指标中, 平面度一直倍受关注, 评判方法也多种多样, 较为常用的有最小区域

法、最小二乘法等. 最小二乘法只要求使误差平方和达到最小, 在各方程的误差之间建立一种平衡即可, 从而防止了某一极端误差取得支配地位. 更为重要的是借助较小的运算量就能揭示系统的更接近真实的状态, 达到寻优的目的^[2].

文中利用最小二乘法进行平面度不确定度评定时对测量数据也进行相应的剔除操作. 主要研究罗曼诺夫斯基判据和格罗布斯准则的具体应用.

2 数据异常值的判别与剔除

2.1 数据异常判别与剔除的原则

根据测量数据的特点一般可分为时变数据(动

态测量数据)和非时变数据(即被测物理量是常数的多次测量数据)两种. 对时变数据而言, 不易采用反复测试的方式以检验测试结果的正确性. 而非时变数据可以进行反复测量, 并且在剔除粗差后还可以降低外界干扰因素对测试结果的影响. 目前在加工制造过程中很少能够在线检测零件质量并做出相应的加工修正, 所以当前的质量检测尚属于事后检测, 满足非时变数据的特点^[3].

在静态数据情况下^[1,4], 判别粗大误差数据的准则有 3 δ 准则、肖维涅准则、罗曼诺夫斯基准则和格罗布斯准则等. 其中 3 δ 准则以测量次数充分大为前提, 而实际测量时次数皆较少, 因此不宜采用. 肖维涅准则虽然对测量次数的要求不高, 但不能保证足够的精度而不宜用于三坐标测量数据的剔除. 而罗曼诺夫斯基准则、格罗布斯准则对测量次数要求较低且实际应用表明以上两种方法可以达到相当高的精度.

由于测量数据大部分情况下是服从某种分布的, 所以对测量数据中异常值的剔除可依据残差的统计判别准则进行. 根据概率论与数理统计的思想, 剔除可疑数据的基本思想是: 首先根据实际情况规定一个显著性水平 α , 而后确定一个置信限, 凡超过这个限度的误差, 就认为它超出随机误差的范围是粗大误差而予以剔除. 假定直接测量列为 $X_1, X_2, X_3, \dots, X_n$, 那么该测量列的标准差可表示为 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n v_i^2}$, 其中 n 为测量点个数, v_i 为残差. 根据数理统计的相关知识, 合理数据的残差的绝对值不超出标准偏差 s 的某一倍数 k , 所以当 $|v_i| > ks$ 时, 剔除该数据. 为了避免漏删和误删, 在每次剔除时只剔除其中残差最大的可疑值, 而对其余数据再做新的剔除处理, 直至无可疑值为止. 在上述判别式中, 由于剔除的侧重点和稳健性等因素考虑不同, k 可有不同的取值和相应的判别准则. 出于对新一代 GPS 标准体系的考虑, 文中将工件的功能要求作为给出 k 值的重要依据之一, 判别式可统一改写为:

$$|v_i|_{\max} > s \cdot k \quad (1)$$

当测量数据满足该式时, 说明该数据是错误数据, 应对该数据予以剔除.

(1) 格罗布斯准则

设独立重复测量的一个正态分布样本为 $x_1, x_2, \dots, x_n$, 对其中的数据 x_d 构造统计量 $g = \frac{x_d - \bar{x}}{s}$. 选定显著性水平 α , 通常取 0.05 或 0.01, 求

得按式(2)定义的临界值 $G(\alpha, n)$

$$P(|g| \geq G(\alpha, n)) = \alpha \quad (2)$$

因此得到如下判别准则: 若 $|x_d - \bar{x}| \geq G(\alpha, n)s$, 则数据 x_d 含有粗大误差, 应予剔除; 否则保留. 式中 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 为均值, $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ 为标准差.

按照 Grubbs 导出的理论分布和临界值 $G(\alpha, n)$ 的分布规律^[5] 可以得到如下关系式:

$$n < 30, k = C_n = \ln(n - 2.65)/2.31 + 1.305 + \rho \quad (3)$$

$$n > 30, k = C_n = \ln(n - 3)/2.301 + 1.36 - n/550 + \rho \quad (4)$$

式中 ρ 为调节系数, 可以根据新一代 GPS 标准体系有关功能要求和测量认证相统一的具体要求确定. 如果机床导轨面在给出平面度要求的基础上还要保证导轨正常工作 10 年, 那么平面度只能要求是上偏差以满足我们的使用年限. 对导轨材质进行分析, 应用材料力学相关知识分析导轨在正常使用条件下的磨损情况, 在此基础上给出合理的 ρ 值. 此时剔除异常值的判别式变为 $|v_i/s|_{\max} > C_n$, C_n 为临界系数.

(2) 罗曼洛夫斯基准则(t 分布检验法)^[3]

与格罗布斯准则相同, 该方法也是针对正态分布的情况. 假定 X 服从 $N(0, 1)$, Y 服从 $\chi^2(n)$ 并且 X, Y 独立, 则随机变量 $t = \frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布. 此时的判据为: 对于 x_i 中的异常值 x_m , 其余 $n-1$ 个数的平均值为 $\bar{x}_{n-1}$, 标准差为 $s_{n-1} = \sqrt{\frac{\sum_{x_i \neq x_m} (x_i - \bar{x}_{n-1})^2}{(n-2)}}$, x_m 相对于 $\bar{x}_{n-1}$ 的残差为 $x_m - \bar{x}_{n-1}$. 若 x_m 满足不等式 $\left| \frac{x_m - \bar{x}_{n-1}}{s_{n-1}} \right| > k(\alpha, n)$, 则剔除 x_m , 式中系数 $k(\alpha, n)$ 即对应的 k 值可依据显著性水平 α 的取值不同, 由式(5)、式(6) 计算:

$$\alpha = 0.01, k(\alpha, n) = \sqrt{\frac{n}{n-1}} (2.57 + \frac{5.44}{n-3.26}) + \rho \quad (5)$$

$$\alpha = 0.05, k(\alpha, n) = \sqrt{\frac{n}{n-1}} (1.958 + \frac{2.481}{n-2.943}) + \rho \quad (6)$$

式中 ρ 为调节系数, 确定方法与上同, 均是根据功能

要求来确定。

2.2 平面度最小二乘拟合的不确定度计算

假定采用不确定度评定的白箱模型对平面度不确定度进行评定,测量结果 Y 作为各次测量变量 x_i 的函数,即:

$$Y = G(X_1, X_2, \dots, X_i, \dots, X_{p+r}) \quad (7)$$

p 表示彼此不相关($\rho = 0$)的变量个数, r 表示具有强相关性($\rho = 1$ 或 -1)的变量个数。那么,测量结果的不确定度可以表示为:

$$u_c = \sqrt{\sum_{j=1}^r \left(\frac{\partial Y}{\partial X_j} \times u_{X_j} \right)^2 + \sum_{i=1}^p \left(\frac{\partial Y}{\partial X_i} \times u_{X_i} \right)^2} \quad (8)$$

为了构造式(8)中的相关系数,假定最小二乘拟合平面为:

$$z = ax + by + c \quad (9)$$

由点到平面的距离公式可知,点 (x_i, y_i, z_i) 到该拟合平面的距离为:

$$d_i = \frac{z_i - ax_i - by_i - c}{\sqrt{1 + a^2 + b^2}} \quad (10)$$

假定 (x_1, y_1, z_1) 和 (x_2, y_2, z_2) 是到最小二乘平面距离最大和最小的两个点,最小二乘检验结果为:

$$\delta = \frac{(z_1 - z_2) - a(x_1 - x_2) - b(y_1 - y_2)}{\sqrt{1 + a^2 + b^2}} \quad (11)$$

下面推导出每一元素的传递系数:

$$\frac{\partial \delta}{\partial a} = \frac{-x_1 + x_2}{\sqrt{1 + a^2 + b^2}} - \frac{a(z_1 - z_2 - a(x_1 - x_2) - b(y_1 - y_2))}{(1 + a^2 + b^2)^{\frac{3}{2}}} \quad (12)$$

$$\frac{\partial \delta}{\partial b} = \frac{-y_1 + y_2}{\sqrt{1 + a^2 + b^2}} - \frac{b(z_1 - z_2 - a(x_1 - x_2) - b(y_1 - y_2))}{(1 + a^2 + b^2)^{\frac{3}{2}}} \quad (13)$$

$$\frac{\partial \delta}{\partial c} = 0 \quad (14)$$

$$\frac{\partial \delta}{\partial x_1} = \frac{-a}{\sqrt{1 + a^2 + b^2}}; \frac{\partial \delta}{\partial x_2} = \frac{a}{\sqrt{1 + a^2 + b^2}} \quad (15)$$

$$\frac{\partial \delta}{\partial y_1} = \frac{-b}{\sqrt{1 + a^2 + b^2}}; \frac{\partial \delta}{\partial y_2} = \frac{b}{\sqrt{1 + a^2 + b^2}} \quad (16)$$

$$\frac{\partial \delta}{\partial z_1} = \frac{1}{\sqrt{1 + a^2 + b^2}}; \frac{\partial \delta}{\partial z_2} = \frac{-1}{\sqrt{1 + a^2 + b^2}} \quad (17)$$

在函数 $\delta = G(x_1, y_1, z_1, x_2, y_2, z_2, a, b, c)$ 中,由于 $x_1, y_1, z_1, x_2, y_2, z_2$ 的坐标值的测量结果是相互独立的,它们是彼此不相关的量,所以一般情况下认为只有变量 a 和 b 相关。同时对参数 c 求导后为 0,根据 ISO/TS 14253-2 的规定,按照保守评估,凡是具有相关性的量其相关系数都按 $\rho = 1$ 或 -1 处理,本例中, a 和 b 取为正相关,那么 $\rho = 1$,可以得到:

$$u_\delta^2 = \left(\frac{\partial \delta}{\partial x_1} u_{x_1} \right)^2 + \left(\frac{\partial \delta}{\partial x_2} u_{x_2} \right)^2 + \left(\frac{\partial \delta}{\partial y_1} u_{y_1} \right)^2 + \left(\frac{\partial \delta}{\partial y_2} u_{y_2} \right)^2 + \left(\frac{\partial \delta}{\partial z_1} u_{z_1} \right)^2 + \left(\frac{\partial \delta}{\partial z_2} u_{z_2} \right)^2 + \left(\frac{\partial \delta}{\partial a} u_a \right)^2 + \left(\frac{\partial \delta}{\partial b} u_b \right)^2 + 2 \frac{\partial \delta}{\partial a} \frac{\partial \delta}{\partial b} \rho_{ab} u_{ab} \quad (18)$$

根据数理统计随机变量相关性的理论知识,在求解式(18)中的 u_a, u_b, u_{ab} 时,将 a, b, c 看作随机向量 r ,在对数据进行处理时可以多次测量或将测量数据分为多组,从而求得随机向量 r 的期望、二阶中心矩以及协方差。在此将测量数据分为三组,那么通过这三组数据就可以得到该平面三组相关参数值: $u_{a1}, u_{b1}, u_{ab1}; u_{a2}, u_{b2}, u_{ab2}; u_{a3}, u_{b3}, u_{ab3}$ 。根据数理统计中 n 维随机变量协方差矩阵的概念,对上述三组平面参数值进行类似处理。根据三组平面参数值得到随机向量 r 的协方差矩阵:

$$\text{Cov}(r) = \begin{bmatrix} u_a^2 & u_{ab} & u_{ac} \\ u_{ba} & u_b^2 & u_{bc} \\ u_{ca} & u_{cb} & u_c^2 \end{bmatrix} \quad (19)$$

从而得到 u_a, u_b 和 u_{ab} ,进而得到平面度最小二乘拟合的认证结果 δ 的不确定度。

2.3 异常数据判别与剔除的实现流程

直接测量值 $x[i]$ 可通过手动或读取数据文件输入,为了便于处理,将输入的数据存储为数据文件。由于采用 MATLAB 作为实现相应算法的软件环境,故可以方便地将数据文件存储为 .txt 格式。由图 1 知在数据处理过程中,测量数据并没有直接送入剔除单元,而是送到平面度最小二乘处理单元。这是因为三坐标测量机测得的数据具有 (x, y, z) 三个坐标的形式,而在数据剔除过程中无法对三个坐标分别处理,而只能是当做一个整体来处理,而且不符合平面度不确定度评定的要求。故首先将坐标值转换为距离,即将三维问题转换为一维问题^[2-3]。由于 (x, y, z) 服从正态分布,并且由公式 $d_i = \frac{z_i - ax_i - by_i - c}{\sqrt{1 + a^2 + b^2}}$ 可知距离值 d 是关于 (x, y, z)

的线性函数. 根据概率论与数理统计, 在转化过程中其正态分布的性质并未发生改变, 这就保证了在对距离值进行相应判别与诊断过程中数据建模的正确性和鲁棒性. 从而可以对转换后的数据按照上述讨论的方法进行处理.

借助计算机的强大处理功能, 在剔除过程中主要采用逐次执行方式, 每次只剔除一个异常值, 尽量避免仅作一次性可疑值剔除造成的数据漏删或误删, 更合理和更大限度地保留有效数据. 当距离值 $d[i]$ 满足剔除准则时, 记录该值及其相关信息, 其

中包括坏值的警示与位置序号等. 从而可以由距离值的相应信息追踪到坐标值的信息, 完成数据剔除任务. 对余留值再次检测时的数据更新采用自动方式快捷完成, 也可通过手动方式依次执行. 为了便于理解和及时修正剔除策略, 可以对每一个值进行手工剔除以便达到人机交互的目的, 同时可以观察到每个异常值以及正常值对检测结果的影响情况. 剔除结束后, 对有效数据以文件形式进行存储做好备份. 程序流程如图 1 所示.

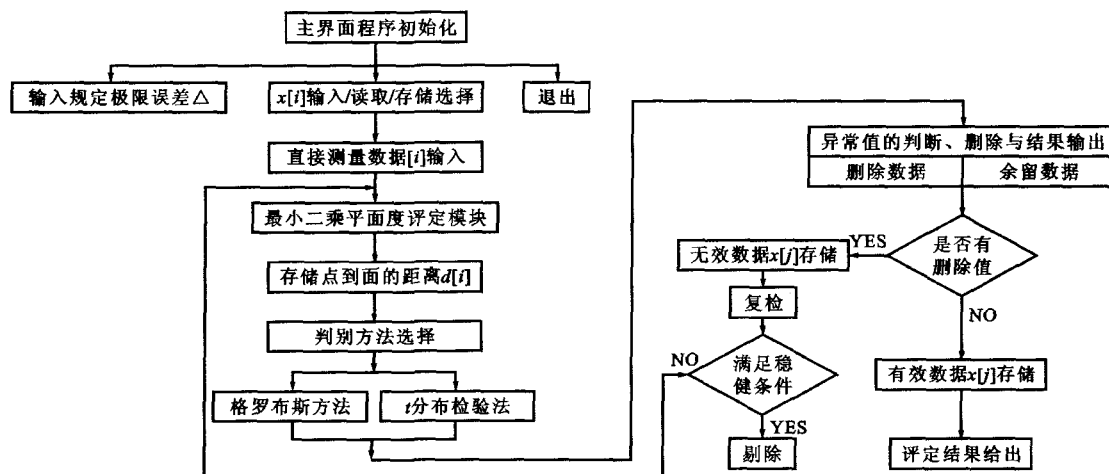


图 1 系统流程图

3 结束语

文中讨论了数据剔除方法以及结合新一代 GPS 标准体系的判别及其表达式, 为在新一代 GPS 标准下进行应用提供了借鉴. 结合最小二乘方法, 应用实例对该方案进行了验证. 验证过程中结合三坐标测量机的相关特点对数学模型的转换以及新模型的建立进行了详细的论述. 实验结果表明该方案具有一定的可行性和应用前景.

参考文献:

- [1] 王中宇, 刘智敏, 夏新涛, 等. 测量误差与不确定度评定[M]. 北京: 科学出版社, 2008.
- [2] 蒋向前. 新一代 GPS 标准理论与应用[M]. 北京: 高等教

育出版社, 2007.

- [3] 苏变萍, 侯筱婷. 文本数据的非线性降维方法研究[J]. 微电子学与计算机, 2007, 24(10): 106-108.
- [4] 余剑. 高精度智能测量系统中粗大误差的处理技术[J]. 测试技术学报, 2003, 17(3): 258-261.
- [5] 曹菊英, 赵跃龙, 王明亮. 异常证据及其检测算法研究[J]. 微电子学与计算机, 2009, 26(4): 111-114.

作者简介:

王乔义 男, (1982-), 硕士研究生. 研究方向为新一代精度理论体系中的不确定度.

黄美发 男, (1962-), 教授. 研究方向为计算机辅助公差设计、GPS 标准体系、制造过程建模.

李向前 男, (1982-), 硕士研究生. 研究方向为 GPS 标准体系建模.

(上接第 146 页)

- [5] 杜景林, 陈力军, 谢立. 无线传感器网络与互联网集成体系结构[J]. 微电子学与计算机, 2008, 25(10): 203-209.

作者简介:

陈真 男, (1985-), 硕士研究生. 研究方向为通信系统与

通信网络.

汪小燕 女, (1963-), 教授. 研究方向为通信系统与通信网络.

王钰 女, (1982-), 硕士研究生. 研究方向为通信系统与通信网络.