小样本时间序列的数据处理

任劲涛，　朱家海，　邵玉梅

(空军工程大学工程学院，陕西西安 710038)

摘　要：针对工程实践中经常出现的测试数据样本容量小、测试时间间隔不等等问题，提出了一种基于插值法的小样本时序数据处理方法，基于此方法对某型启动箱的延时参数拟合了ARMA(n,m)模型，并进行了预测，结果表明：本方法对起动箱测试参数分析结论与实际情况吻合，效果良好

关键词：小样本；时间序列；数据处理；插值；建模

中图分类号：O24　　文献标识码：A　　文章编号：1009-3516(2005)03-0071-03

时间序列分析是一种行之有效的建模方法，要求观测数据样本不能少于50个，且测试时间等间隔。但在工程实践中，由于种种因素的制约，其观测数据样本往往会出现少于50个，且时间间隔不等的情况，因此研究小样本时序数据处理方法显得尤为重要。文献[1]利用支持向量机的方法对小样本数据进行建模预测，但如何合理地定义模型很困难，本文结合某型起动箱测试数据分析，给出一种简单实用的数据处理方法。

1　小样本测试数据的预处理

表1给出了某型起动箱10 s延时参数原始测量结果，其中2002.10表示2002年10月(其它同理)。从表1可以看出：① 测量参数只有16个，为小样本；② 缺少2003.1和2003.7的测量参数，说明测试间隔不等；③ 个别参数数据大小异常。若直接利用这些数据进行建模分析，会出现很大的误差，甚至出现错误的结果，因此在统计分析之前必须进行测试数据的预处理。首先，剔除明显有误差的值，同时为了保证数据的完整性和连续性，对剔除的数据进行插值；然后，为了保证测试数据的等间隔性，再次对原始数据进行插值补充。

表1　原始数据序列

序号	1	2	3	4	5	6	7	8
时间	2002.10	2002.11	2002.12	2003.2	2003.3	2003.4	2003.5	2003.6
x	10.030	10.045	10.035	10.055	10.045	10.040	10.050	10.030
序号	9	10	11	12	13	14	15	16
时间	2003.8	2003.9	2003.10	2003.11	2003.12	2004.1	2004.2	2004.3
x	10.050	10.055	9.9950	10.040	10.050	10.040	10.045	10.040

1.1　野点的剔除

数据处理前，需剔除原始数据中可能出现的野点。拉依达[2](PaЙta)准则是常见的野点剔除方法，拉依达准则判断方法如下：假定测试数据样本 x 服从正态分布，则有 $P(|x-\mu|>3\sigma)<0.003$。其中 μ 和 σ 分别为样本的数学期望和标准差。设测试数据为 $x_1,x_2,\cdots,x_n$，则均值 $\bar{x}=\dfrac{1}{N}\sum\limits_i x_i$，残差 $V_i=x_i-\bar{x}(i=1,2,\cdots,n)$，标准差 $\sigma=\sqrt{(n-1)^{-1}\sum\limits_i V_i^2}=\sqrt{(n-1)^{-1}[\sum\limits_i(x_i)^2-(-\sum\limits_i x_i)^2/n]}$。若某个测量值 x_d 的残差 V_d(1

收稿日期：2004-09-27
基金项目：军队科研基金资助项目
作者简介：任劲涛(1977-)，男，河南孟津人，硕士生，主要从事惯导与组合导航等研究.

$< d < n$) 满足 $|V_d| > 3\sigma$;则认为是异常值,应予以剔除。

对表1所示数据,应用拉依达准则,计算的均值 $\bar{x} = 10.0403$, $\sigma = 0.0143$,第11个数据对应的残差 V_{11} 为 $-0.045\ 312$,由于 $|V_{11}| = 0.045\ 312 > 3\sigma = 0.042\ 9$,根据拉依达准则,可判定 x_{11} 为野点,应予以剔除。

为维护数据的完整性,需对剔除的野点进行补值,根据拉格朗日插值原理,选取 x_9、x_{10}、x_{12}、x_{13} 四个点进行插值补充得 $x_{11} = 10.046\ 7$,考虑到测试电秒表的测试精度因素,取 $x_{11} = 10.045$。

1.2 测试数据等间隔处理

时间序列分析要求数据测试时间是等间隔的,由表1知,在时间点2003.1和2003.7上没有测试记录,利用拉格朗日插值多项式对其进行插值补充后,$x_4 = 10.045$, $x_{10} = 10.038\ 4$,同样考虑到测试电秒表的测试精度因素,取 $x_{10} = 10.040$,即得表2所示容量为18的基本数据序列,序号4、序号10为插值补充的两个数据。序号13为1.1中剔除野点后的补值。

表2 基本数据序列

序号	1	2	3	4	5	6	7	8	9
x	10.030	10.045	10.035	10.045	10.055	10.045	10.040	10.050	10.030
序号	10	11	12	13	14	15	16	17	18
x	10.040	10.050	10.055	10.045	10.040	10.050	10.040	10.045	10.040

2 构建建模所用的时间序列

2.1 三次样条插值函数

设函数 $y = f(x)$ 在区间 $[a,b]$ 上连续,且已知点: $a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$ 上的函数值为 $f(x_i) = y_i (i = 0,1,...,n)$,若函数 $S(x)$ 满足条件:① 在每个小区间 上是 x 的三次多项式;② 在整个区间 $[a,b]$ 上具有二阶连续导数;③$S(x_i) = y_i (i = 0,1,...,n)$。则称 $S(x)$ 为 $f(x)$ 的三次样条插值函数。

三次样条插值函数 $S(x)$ 的具体构造方法及具体公式见参考文献[4]。对三次样条插值函数来说,当插值节点逐渐加密时,可以证明:不但样条插值函数收敛于函数本身,而且其导数也收敛于函数的导数,基于此优点,下文用三次样条插值函数构建大样本时间序列。

2.2 插值构建时间序列

利用三次样条函数插值法,以原来的18个测试数据为插值基本点,在相邻的两个基本点之间等点数插入3个插值点,得到一个样本容量为69的插值时间序列 $\{Z_t\}$,基本点、插值点对应的曲线见图1,其中"*"为基本点,"+"为插值结果。

2.3 时间序列数据特性的检验

2.3.1 平稳性检验

利用时间序列建模,需保证所构建的时间序列是平稳的。本文采用逆序检验法进行检验。逆序检验法的基本原理是:若数据序列 $\{Z_t\}$ 平稳,则其分段子序列的均值或方差应无显著差异。其中均值平稳性检验的步骤如下:① 将 $\{Z_t: 1 \leqslant t \leqslant n\}$ 分成 k 段,每段 n/k 个数据,并相应计算各段均值,得到均值序列;② 计算逆序总数 A 及构造的统计量 $u = (A + 1/2 - k(k-1)/4)/\sqrt{k(2k^2 + 3k - 5)/72}$;③ 由给定置信水平,查标准正态分布得 $z_{\alpha/2}$,若 $|u| < z_{\alpha/2}$,则以置信度 $1-\alpha$ 认为该数据序列具有平稳性,否则认为是不平稳的。

对起动箱10 s延时参数序列 $\{Z_t\}$,计算得统计量 $\mu = 0.894\ 4$,当取显著性水平 $\alpha = 0.05$ 时,查标准正态分布得 $z_{\alpha/2} = 1.96$,因为 $|\mu| < 1.96$,故以置信度95%认为该数据序列是平稳的。

2.3.2 正态性检验

对时间序列 $\{Z_t\}$ 的正态性检验,一般采用偏峰态检验法,最基本的是检验 $\{Z_t\}$ 的偏态系数 ξ 与峰态系数 ν 是否满足正态随机变量的特性。ξ 与 ν 的定义为: $\xi = E[(x_t - \mu_x/\sigma_x)]^3$; $\nu = E[(x_t - \mu_x/\sigma_x)]^4$,理论上可以证明,若 Z_t 是正态随机变量,则有: $\xi = 0$; $\nu = 3$。因此对 $\{Z_t\}$ 计算 ξ 和 ν 的估计值: $\hat{\xi} = \frac{1}{N\sigma_Z^3}\sum_{i=1}^{N}(Z_t - \hat{\mu}_Z)^3$; $\hat{\nu} = \frac{1}{N\sigma_Z^4}\sum_{i=1}^{N}(Z_t - \hat{\mu}_Z)^3$,其中 $\hat{\mu}_Z$ 和 $\hat{\sigma}_Z$ 分别是 $\{Z_t\}$ 的均值和标准方差的估计值。当算得的 $\hat{\xi} \approx 0$; $\hat{\nu} \approx 3$ 时,则认为 $\{Z_t\}$ 是正态时序。

对起动箱 10 s 延时参数序列{Z_t},经计算得: $\hat{\xi} = -0.2215$, $\hat{v} = 2.5233$,所以可近似认为{Z_t}为正态时序。

经计算插值数据序列均值 $\overline{Z} = \frac{1}{69} \times \sum_{j=1}^{69} Z_j = 10.044$,对 $N = 69$ 个插值数据序列做零化处理,作变化 $W_t = Z_t - \overline{Z}$,即得到符合 ARMA 建模要求的 69 个新数据序列{W_t}。

3　建模预测

利用{W_t}拟合了 AR(3)模型,并进行了五步预测,预测曲线见图 2,预测误差曲线见图 3。从图 2、图 3 可以看到:在一定的误差范围内,预测曲线与实测曲线能够比较好地吻合。而且从预测的结果也能很好地判断该型起动箱 10 s 延时时间的变化趋势,从而实现对该型起动箱 10 s 延时时间的预测和控制。

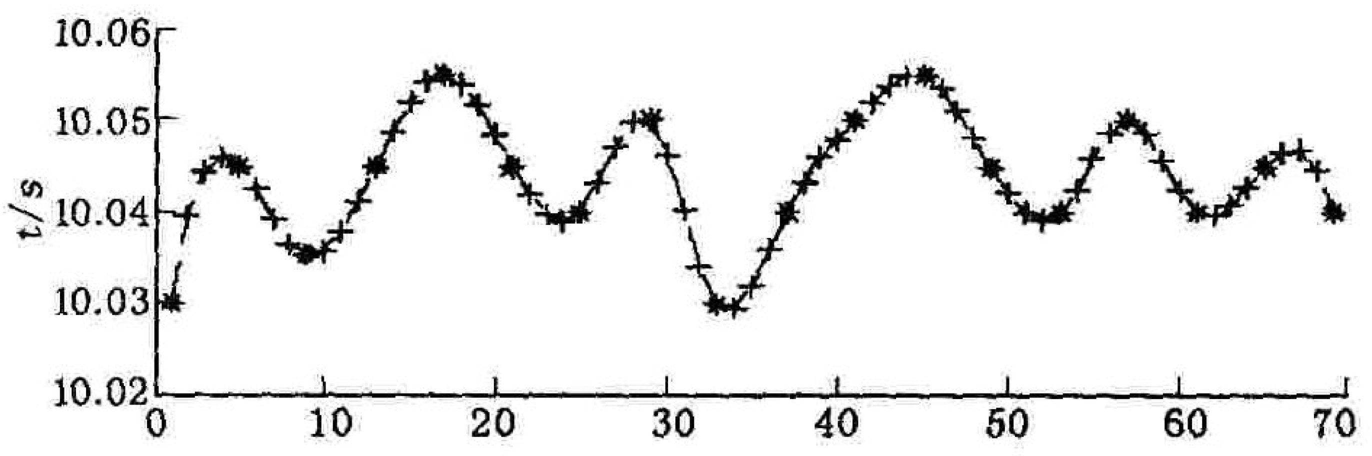

图 1　基本点、插值点关系图

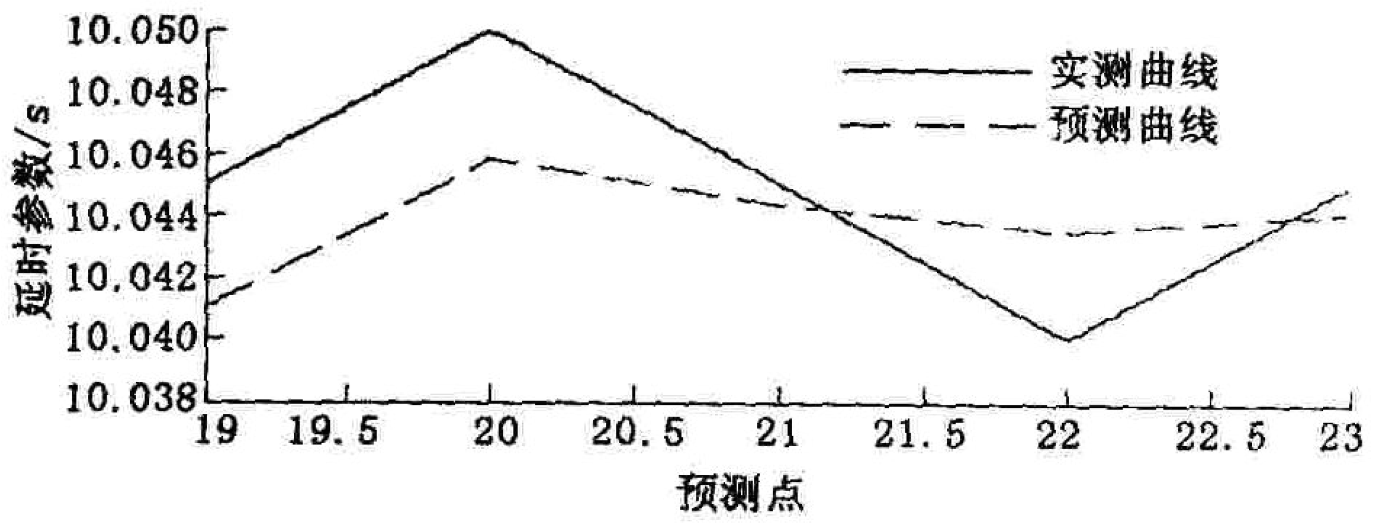

图 2　实测曲线及预测时间点比较

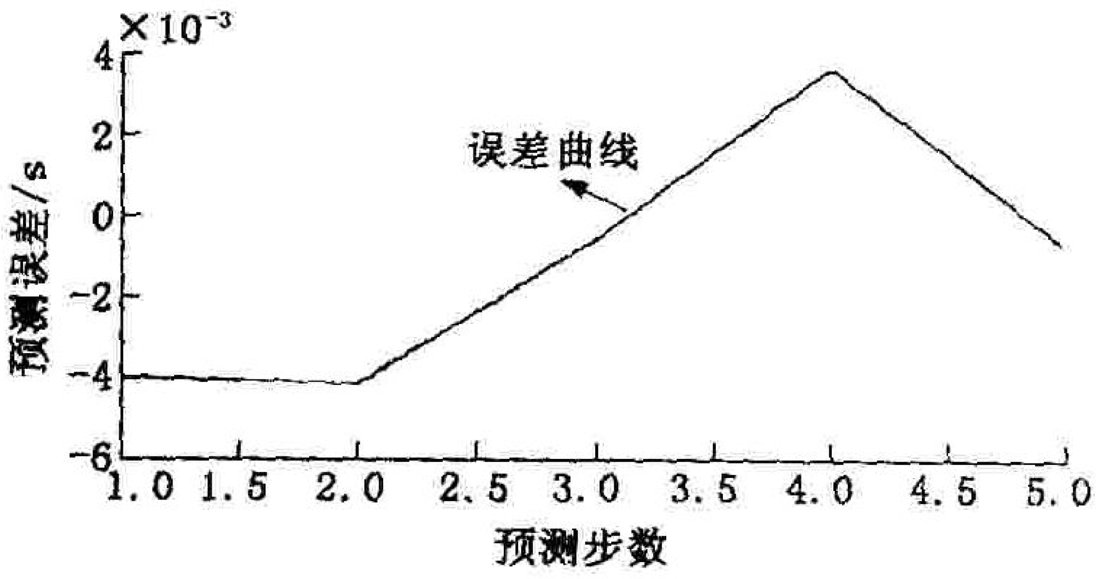

图 3　预测偏差曲线

4　结束语

本文针对某型起动箱延时参数的测试数据特点,提出了一种基于插值法的小样本时序数据处理方法,建模预测结果表明,本方法对起动箱测试参数分析结论与实际情况吻合,效果良好,对处理工程实践类似问题具有较大的推广应用价值。

参考文献:

[1]　周佃民. 支持向量机预测方法研究及其在电力市场预测中的应用[D]. 西安:西安交通大学,2003.

[2]　刘加丛,秦玉勋,刘占辰. 一种小子样试验数据分析方法[J]. 空军工程大学学报(自然科学版),2003,4(1):71-73.

[3]　刘进忙,冯有前,张晓刚. 基于最小二乘法 Lagrange 插值基函数的拟合推广[J]. 空军工程大学学报(自然科学版),2002,3(4):84-87.

[4]　李信真,车刚明,欧阳洁,等. 计算方法[M]. 西安:西北工业大学出版社,2000.

(编辑:姚树峰)

A Study of Data Processing of Small - Stylebook Time Series

REN Jin - tao, ZHU Jia - hal, SHAO Yumei

(The Engineering Institute, Air Force Engineering University, Xi'an, Shaanxi 710038, China)

Abstract: The problem that the testing sample capacity is small, and the testing intervals are not equal arises frequently in the project practice. In response to this kind of problems, a method of data processing of small - sample time series based on insert - data is put forward. Modeling ARMA (n, m) of delay - time of some type of starter combustor is made based on this method. A forecasting is carried out and a good result is obtained.

Key words: small - stylebook ; time series ; data processing; insert - data; modeling