

不同随机分布的相位随机化对实测 数据影响的分析研究*

马军海^① 陈予恕^② 刘曾荣^③

(1997 年 1 月 20 日收到, 1998 年 4 月 10 日收到修改稿)

摘 要

本文对服从不同分布的随机数相位充分随机化后对动力系统实测数据生成的替代数据的影响做了分析研究, 计算结果表明: 不同分布的随机数相位充分随机化后对临界值无质的影响。这对相位随机化方法的可行性, 实用性和广泛应用提供了理论依据, 并对它对于实际问题的应用提供了具体的实施方法。

关键词 实测混沌数据 替代数据 判值 相位随机化 随机时序

中图分类号 O175, O241

§ 1. 引 言

自八十年代以来, 国内外学者对时序问题已开展了大量的研究。由于在现实生活中存在着大量以随机为主的时序和以非线性混沌为主的时序, 或两者兼有之。所以如何区分一个现实的时序问题是随机的还是非线性混沌的, 或是以随机因素为主的还是以非线性混沌因素为主的就变得尤为重要。文献[1]给出了对现实的时序不同特性问题判别的基本方法即相位随机化方法。文献[2]给出了判值的计算方法。如何尽可能地减少原始数据中所含信息量在替代数据生成过程中的流失, 以及服从不同分布的随机数进行相位充分随机化是否会对判值产生影响, 是相位随机化方法能否如实地判定实测数据特性归属以及此方法能否被广泛地应用在各种各样时序问题的判定中的两个重要问题。第一个问题我们在文献[3]中已给出了相应结论, 第二个问题是本文研究的主要问题, 因为如果用不同分布的随机数进行相位随机化会对判值产生质的影响, 则会影响相位随机化方法在解决实际问题中的广泛应用。甚至人们会对此方法的切实可行性提出疑义。所以研究服从不同随机变量分布的随机数相位充分随机化对判值的影响程度如何, 便变得十分重要, 如影响程度较少或根本没有什么影响则对相位随机化方法的广泛应用提供了广泛的理论依据。本文在文献[3]的基础上, 对服从不同分布的随机数相位充分随机化后对动力系统实测数据生成的替代数据的影响做了分析研究, 计算结果表明:

* 国家自然科学基金资助项目(19672043)

① 东南大学管理学院系统所, 南京 210096

② 天津大学力学系, 天津 300072

③ 上海大学理学院数学系, 上海 201800

服从不同随机变量分布的随机数相位充分随机化后对临界值无质的影响。这对相位随机化方法的可行性、实用性和广泛应用提供了有力的理论依据,并对它对于实际问题的应用提供了具体的实施方法。

§ 2. 替代数据的生成和判据建立

令所得到的时间序列 $x_1, x_2, x_3, \dots, x_N, x_i$ 为第 T_i 时间采样所得,其中 $T_1, T_2, T_3, \dots, T_N$, 分别为:

$0, \Delta T, 2\Delta T, \dots, (N-1)\Delta T, \Delta T$ 为采样时间间隔。应用离散的付里叶变换,变换算子为 F , 得到:

$$X(f) = F\{x(t)\} = \sum_{n=0}^{N-1} x(t_n) e^{2\pi i f n \Delta t} \quad (2.1)$$

将式(2.1)可以进一步写为

$$x(t) = F^{-1}\{X(f)\} = A(f) e^{i\Phi(f)} \quad (2.2)$$

其中, $A(f)$ 为幅值, $\Phi(f)$ 为相位, f 分别取:

$f = -N\Delta f/2, \dots, -\Delta f, 0, \Delta f, \dots, N\Delta f/2; \Delta f = (1/N)\Delta T$ 。然后将 $\Phi(f)$ 随机地旋转一相位角 $\Psi(f)$, $\Psi(f)$ 是用计算机在 $[0, 2\pi]$ 内任意随机所生成的服从均匀分布和高斯分布的随机数。可以得到:

$$\tilde{X}(f) = A(f) e^{i[\Phi(f) + \Psi(f)]} = X(f) e^{i\Psi(f)} \quad (2.3)$$

再对(2.3)式进行付里叶逆变换可以得到:

$$\hat{x}(f) = F^{-1}\{\tilde{X}(f)\} = F^{-1}\{X(f) e^{i\Psi(f)}\} = \frac{1}{N} \sum_{n=0}^{N-1} X(n) \exp\left[\frac{i2n\pi k}{N}\right] \quad (2.4)$$

这样便得到替代数据 $\hat{x}_1, \hat{x}_2, \hat{x}_3, \dots, \hat{x}_N$, 我们在文献[3]中已提出了使替代数据虚部为零的依据和实现方法。

现以原始数据为例建立判定数据特性的方法:

$$\bar{V}_i = [x_i, x_{i+T}, x_{i+2T}, \dots, x_{i+(m-1)T}] \quad (2.5)$$

其中, $\bar{V}_i \in \mathcal{R}$ 。

为减少由于实际问题中所产生的数据间的自相关性,文献[4]已证明 $C(r, N, W)$ 应采用如下式子:

$$C(r, N, W) = \frac{2}{N^2} \sum_{n=W}^N \sum_{i=1}^{N-n} H[r - |\bar{V}_{i+n} - \bar{V}_i|] \quad (2.6)$$

这里, H 为 Heaviside 函数: $H(x) = \begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases}$

由于数据之间的自相关性和实际问题中的 N 取值的限制 Theiler^[2] 建议 $C(r, N, W)$ 应用(2.7)式:

$$C(r, N, W) = \frac{2}{(N-W+1)(N-W)} \sum_{n=W}^N \sum_{i=1}^{N-n} H[r - |\bar{V}_{i+n} - \bar{V}_i|] \quad (2.7)$$

这里要求 W 应满足条件: $W \geq \tau \left(\frac{2}{N}\right)^{2/m}$ (2.8)

τ 为时序之间的时间间隔, m 为嵌入维数,采用时序的关联维数为:

$$d = \lim_r \lim_N \frac{d[\log_2 \mathcal{C}(r, N, W)]/dr}{d[\log_2 \chi(r)]/dr} \quad (2.9)$$

其中, $d \leq m \leq 2d + 1$ 对原始数据取 $\langle S_{\text{orig}} \rangle = d$ 对替代数据取 $\langle S_{\text{surr}} \rangle = d_{\text{surr}}$ $\langle S \rangle$ 为 s 的统计平均。

$$\text{定义 } Z = | \langle S_{\text{surr}} \rangle - \langle S_{\text{orig}} \rangle | / \sigma_{\text{surr}} \quad (2.10)$$

σ_{surr} 为替代数据的均方差, Z 便为判据, 我们在文献[3] 中已指出: 临界值 $Z = 1.96$, 若 $Z \leq 1.96$ 则实测数据时序以 95% 的置信水平为随机因素为主的时序, 若 $Z \geq 1.96$ 则实测数据时序以 95% 的置信水平为非线性混沌因素为主的时序。

§ 3. 高斯随机数的生成

高斯随机数是一种有任意均值 μ 和任意方差 σ^2 的正态分布的随机数, 这种随机数在实际工程问题中最常见, 根据中心极限定理: 相互独立且服从同一分布的随机变量都和高斯分布之间存在非常密切的关系, 所以高斯分布具有广泛的代表性和实用性, 深入细致地研究服从正态分布的随机数相位随机化后对判值的影响是非常必要的。

根据概率论和数理统计理论, 随机变量 $X (X = [x_1, x_2, \dots]^T)$ 其均值和方差为:

$$\mu = E[X] = \int_{-\infty}^{\infty} xp(x) dx \quad (3.1)$$

$$\sigma^2 = \int_{-\infty}^{\infty} [x - E(x)]^2 p(x) dx \quad (3.2)$$

可算得区间 $[0, 1]$ 上服从均匀分布的随机变量 X 的均值和方差为:

$$\mu = 0.5, \quad \sigma^2 = 1/12 \quad (3.3)$$

根据中心极限定理可证明当 $n \rightarrow \infty$ 时 随机变量:

$$\left[\sum_{i=1}^n X_i - n\mu \right] \sqrt{\sqrt{n}\sigma} \quad (3.4)$$

服从标准正态分布 $N(0, 1)$, 实际问题中由于 n 不可能取 $+\infty$, 本文取 $n = 48$, 其近似程度已相当好了。由(3.4)式可推得:

$$Y = \mu + \sigma \left[0.5 \sum_{i=1}^{48} X_i - 12 \right] \quad (3.5)$$

近似服从 $N(\mu, \sigma^2)$

(3.5) 式便作为本文生成高斯随机数的方法。

§ 4. 计算结果

取如下四组数据对本文的方法进行验证:

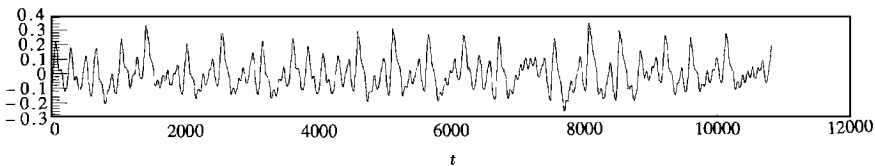


图 1 第四组数据的时间历程图

(1) 取 $[0, 2\pi]$ 内的随机数 7500 点(其相图如图 7)。

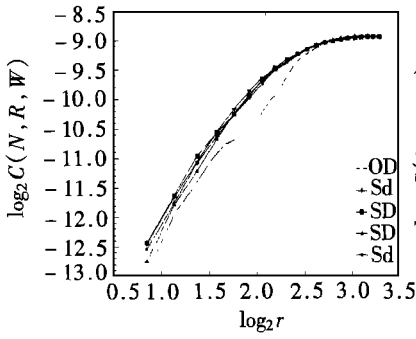


图2 第一组数据及其四组替代数据的 $\log_2 r \log_2 C(r, N, W)$ 图

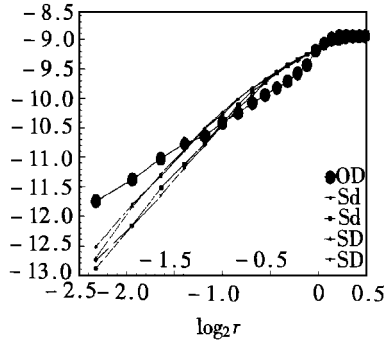


图3 第二组数据及其四组替代数据的 $\log_2 r \log_2 C(r, N, W)$ 图

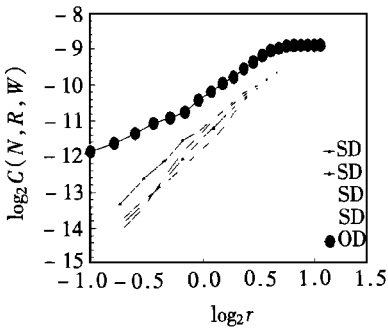


图4 第三组数据及其四组替代数据的 $\log_2 r \log_2 C(r, N, W)$ 图

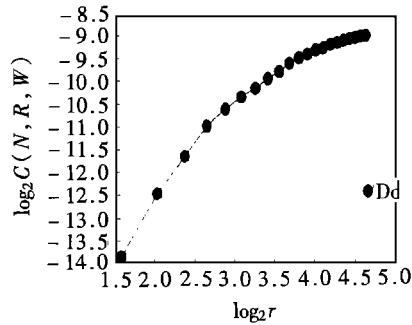


图5 第四组数据及其四组替代数据的 $\log_2 r \log_2 C(r, N, W)$ 图

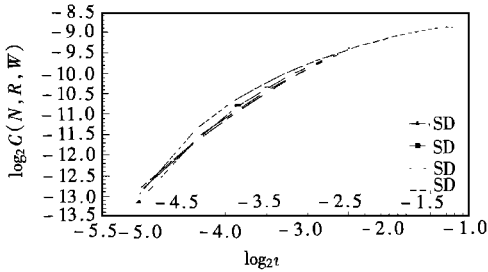


图6 第四组数据的四组替代数据的 $\log_2 r \log_2 C(r, N, W)$ 图

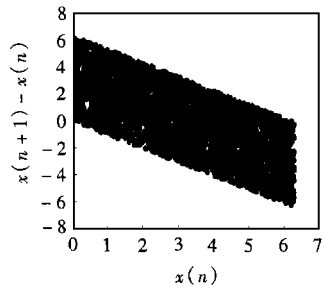


图7 第一组数据的相图

(2) 对 Logistic map $x_{n+1} = \lambda x_n(1 - x_n)$ 取 $\lambda = 4$ 标准混沌的情况, 前 1000 点作为暂态点去掉, 把后 7500 点作为(2) 组原始数据点。

分 (3) 取 Henon map $\begin{cases} x_{n+1} = 1 - \alpha x_n^2 + y_n \\ y_{n+1} = \beta x_n \end{cases}$ ($\alpha = 1.4, \beta = 0.3$) 标准混沌的情况 7500, 前 1000 点作为暂态点去掉, 把后 7500 点作为(3) 组原始数据点。

(4) 在天津大学非线性参数机械振动模型上得到 6225 点作为实验数据, 其时间历程图如图 1, 其相图如图 9。

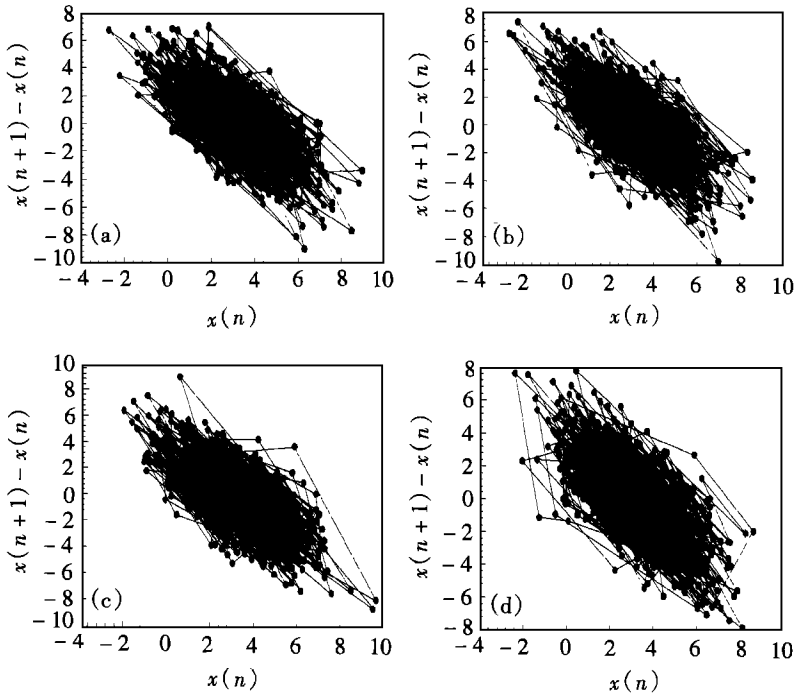


图 8 (a)、(b)、(c)、(d) 第一组数据的四组替代数据 a, b, c, d 的相图

表 1

	(1)组原始数据	(1)组替代数据 a	(1)组替代数据 b	(1)组替代数据 c	(1)组替代数据 d
均值 X	3.143593814	3.143593855	3.143593797	3.143593835	3.143593851
均方差	1.81342785275	1.813427871195	1.813427862013	1.813427861611	1.813427848001
$S_{\langle S_{orig} \rangle}$	1.920297179467				
$\langle S_{surr} \rangle$		2.090015691268	2.199898283823	2.317852380578	2.297979050142
	(2)组原数数据	(2)组替代数据 a	(2)组替代数据 b	(2)组替代数据 c	(2)组替代数据 d
均值 X	0.504697622	0.504697627	0.504697638	0.504697608	0.504697644
均方差	0.340114793301	0.340114776009	0.340114785928	0.340114780756	0.340114780652
$S_{\langle S_{orig} \rangle}$	1.129808967525				
$\langle S_{surr} \rangle$		1.855071391208	1.869004651628	1.876265034335	1.866391761901
	(3)组原数数据	(3)组替代数据 a	(3)组替代数据 b	(3)组替代数据 c	(3)组替代数据 d
均值 X	0.160699786848	0.160699800001	0.160699797002	0.1606992980	0.1606997890
均方差	0.545648587853	0.54564876051	0.545648599899	0.545648581472	0.545648583427
$S_{\langle S_{orig} \rangle}$	1.26999000001				
$\langle S_{surr} \rangle$		2.930198768601	3.090042856004	2.960000200001	3.001006022301
	(4)组原数数据	(4)组替代数据 a	(4)组替代数据 b	(4)组替代数据 c	(4)组替代数据 d
均值 X	0.004078231	0.004079012	0.004079144	0.004078875	0.004078998
均方差	0.063289616570	0.063289634001	0.063289596889	0.063289687043	0.063289696161
$S_{\langle S_{orig} \rangle}$	1.322466767501				
$\langle S_{surr} \rangle$		1.095651282801	1.110599347013	1.104820166101	1.105417170001

对上述四组数据的原始数据分别进行付里叶变换,在其逆付里叶变换的相位上分别加上在 $[0, 2\pi]$ 内产生的高斯分布的随机数,或加上用蒙特卡洛随机取余方法生成的均匀分布的随机数来生成替代数据。表中替代数据 a, b 由相位上加上不同的均匀分布的随机数所生成,其随机数种子分别取为 33.0 和 151.0,替代数据 c 由相位上加上 $\mu=0, \sigma=0$ 的高斯分布的随机数所生成,替代数据 d 由相位上加上 $\mu=0, \sigma=10.0$ 的高斯分布的随机数所生成。算上原始数据每组共有 5 组数据,其均值,均方差和算得的维

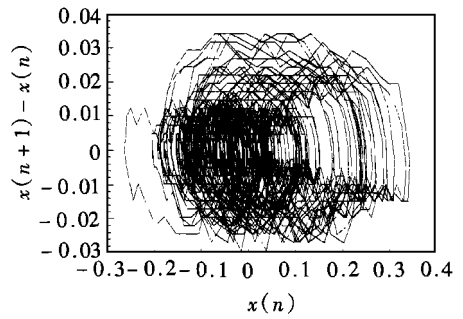


图9 第四组数据的相图

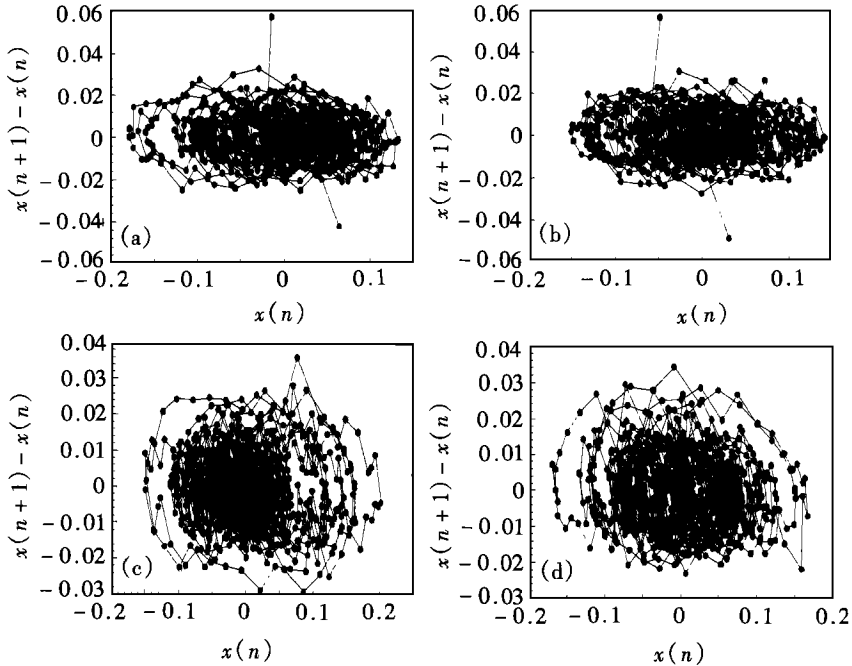


图10 (a)、(b)、(c)、(d) 第四组数据的四组替代数据 a, b, c, d 的相图

数均值如表 1:

从表 1 中的数据中可以看出:无论是相位上加上均匀分布的随机数还是加上高斯分布的随机数其原始数据与替代数据的均值和均方差基本上是相同的。但由于相位的充分随机化在频域上表现为数据相位的改变,在实域上则表现为数据的改变,从而通过(2.10)式来判断这种改变的激烈程度。

替代数据对应的判值 Z_a, Z_b, Z_c, Z_d 的值如下:

对第一组数据:

$$Z_a = 0.0935898, Z_b = 0.1541837,$$

$$Z_c = 0.21922857, Z_d = 0.208269$$

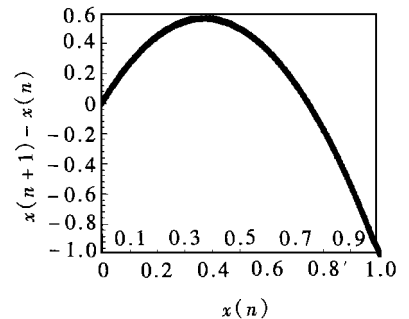


图11 第二组数据的相图

对第二组数据:

$$Z_a = 2.132403892, Z_b = 2.173371, Z_c = 2.194718, Z_d = 2.165688$$

对第三组数据:

$$Z_a = 3.042615, Z_b = 3.335558, Z_c = 3.0972319, Z_d = 3.172382$$

对第四组数据:

$$Z_a = 3.58376, Z_b = 3.347586, Z_c = 3.438895, Z_d = 3.4294618$$

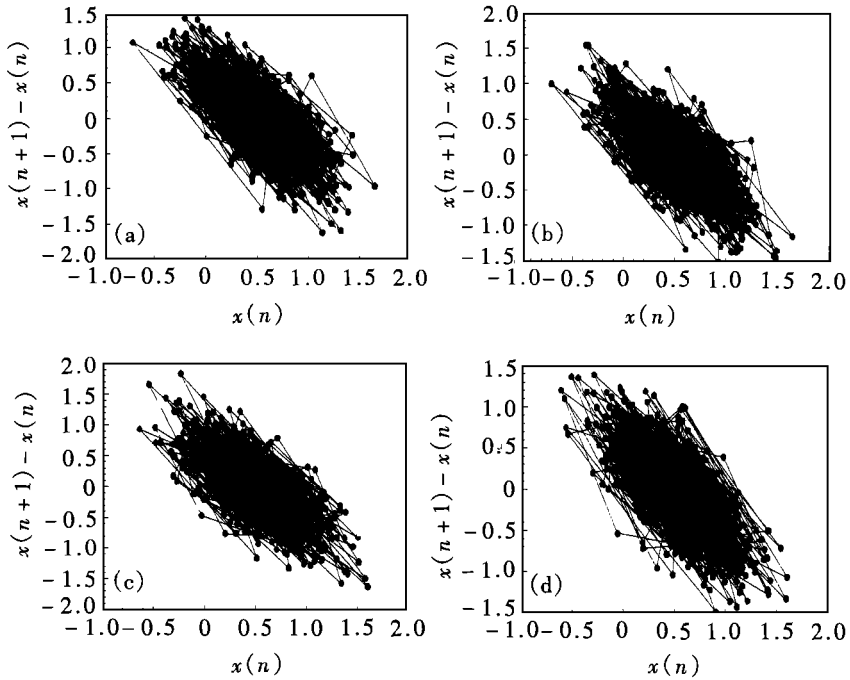


图 12 (a)、(b)、(c)、(d) 第二组数据的四组替代数据 a, b, c, d 的相图

由以上得算得的各组判值与临界值 1.96 作一比较, 可以得出这样的结论: (1) 组数据为随机数据, (2), (3) (4) 组数据为混沌数据。对 (1) (2) (3) 组数据其判定结论与所采用的数据类型完全一致。

图 2~ 图 6 为 (1)~ (4) 组数据的维数图。图 2 为 (1) 组数据原始数据和其对应的四组替代数据的 $\log r_log r(N, W, m)$ 图, 图中取 $N = 7500, W = 12, m = 4, \min r = 1.50, \max r = 11.50$ 。

图 3 为 (2) 组数据原始数据和其对应的四组替代数据的 $\log r_log r(N, W, m)$ 图, 图中取 $N = 7500, W = 12, m = 3, \min r = 0.18, \max r = 1.5$ 。

图 4 为 (3) 组数据原始数据和其对应的四组替代数据的 $\log r_log r(N, W, m)$ 图, 图中取 $N = 7500, W = 12, m = 3, \min r = 0.5, \max r = 2.85$ 。

图 5 为 (4) 组数据原始数据和其对应的四组替代数据的 $\log r_log r(N, W, m)$ 图, 图中取 $N = 6224, W = 12, m = 3, \min r = 2.85, \max r = 32.0$ 。

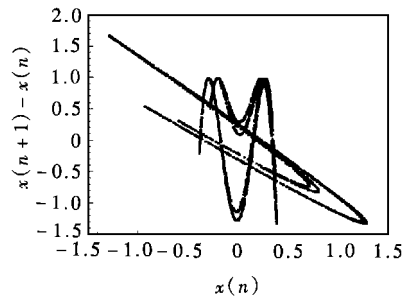


图 13 第三组数据的相图

图 6 为 (5 组数据原始数据和其对应的四组替代数据的 $\log r_{\log r}(N, W, m)$ 图, 图中取 $N = 6224, W = 12, m = 3, \min r = 0.022, \max r = 0.5$ 。

图 7 为 (1 组数据原始数据的相图。图 8(a), (b), (c) 和 (d) 为 (1 组数据原始数据的四组替代数据的相图。

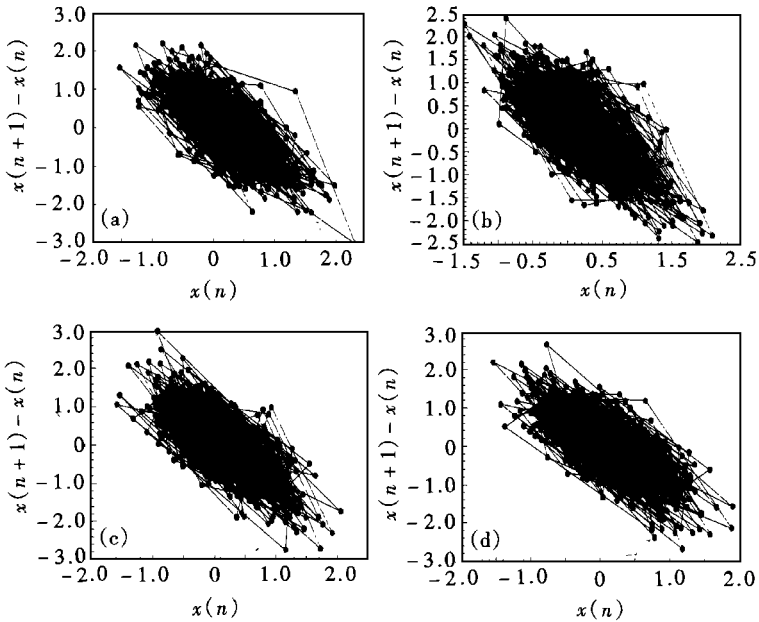


图 14 (a)、(b)、(c)、(d) 第三组数据的四组替代数据 a, b, c, d 的相图

图 9 为 (4 组数据原始数据的相图。图 10(a), (b), (c) 和 (d) 为 (4 组数据原始数据的四组替代数据的相图。

图 11 为 (2 组数据原始数据的相图。图 12(a), (b), (c) 和 (d) 为 (2 组数据原始数据的四组替代数据的相图。

图 13 为 (3 组数据原始数据的相图。图 14(a), (b), (c) 和 (d) 为 (3 组数据原始数据的四组替代数据的相图。

§ 5. 结 论

计算结果表明服从正态分布的随机数和服从均匀分布的随机数相位随机化后其相应的四组判值只有很少的差异, 从图 2~ 图 6 中可以看出四组数据中服从正态分布的随机数和服从均匀分布的随机数相位随机化后其相应的 $\log r_{\log r}(r, N, W)$ 图线相互交错在一起, 且其对应的替代数据的均值和均方差也基本相同, 这充分说明服从不同分布的随机数相位随机化后判值几乎无影响, 但必须保证随机数的生成是充分随机的。这一结论论证了相位随机化方法在实际工程问题中的可行性, 实用性和广泛应用前景, 并论证了本文的方法是正确的。

参 考 文 献

- 1 S. Rombouts and R. Keunen, Investigation of nonlinear structure in multichannel EEG, *Phys. Lett.*, **A202** (1995), 352—358•
- 2 James Theiler, Spurious dimension from correlation algorithms applied to limited time series data, *Phys. Rev.*, **A34** (1986), 2427—2432•
- 3 马军海、陈予恕、刘曾荣, 动力系统实测数据的非线性混沌特性的判定, *应用数学和力学*, **19**(6) (1998), 481—489•
- 4 Dean Prichard, The correlation dimension of differenced data, *Phys. Lett.*, **A191** (1994), 245—250•
- 5 Matthew B. Kennel, Method to distinguish possible chaos from colored noise and to determine embedding parameters, *Phys. Rev. Lett.*, **A46** (1992), 3111—3118•
- 6 P. E. Rapp and A. M. Albano, Filtered noise can mimic low-dimensional chaotic attractors, *Phys. Rev.*, **E47** (1993), 2289—2297•
- 7 Dean Prichard, Generating surrogate data for time series with several simultaneously measured variables, *Phys. Rev. Lett.*, **191** (1994), 230—245•
- 8 P. E. Rapp and A. M. Albano, Phase-randomized surrogates can produce spurious identifications of non-random structure, *Phys. Lett.*, **A192** (1994), 27—33•
- 9 Henry D. I. Abarbanel, Prediction in chaotic nonlinear systems methods for timeseries with broadband Fourier spectra, *Phys.*, **B5** (1991), 1347—1375•
- 10 M. Casdagli and Alistair Mees, Modeling chaotic motions of a string from experimental data. *Phys. Rev.*, **E54** (1992), 303—328•
- 11 P. E. Rapp and A. M. Albano, Predicting chaotic time series, *Phys. Rev.*, **E47** (1993), 2289—2297•
- 12 J. Luis Cabrera and F. Javier, Numerical analysis of transient behavior in the discrete random Logistic equation with delay, *Phys. Lett.*, **A197** (1995), 19—24•
- 13 Eric J. Kostelich, Problems in estimating dynamics from data, *Phys.*, **D58** (1992), 138—152•
- 14 S. J. Schiff and T. Chang, Information transport in temporal systems, *Phys. Rev. Lett.*, **A** (1992), 378—393•
- 15 Peter Grassberger, Finite sample corrections to entropy and dimension estimates, *Phys. Lett.*, **A125** (1988), 369—373•
- 16 James Theiler, Some comments on the correlation dimension of noise, *Phys. Lett.*, **A155** (1991), 480—493•
- 17 J. Timonen and H. Koskinen, An improved estimator of dimension and some comments on providing confidence intervals, *Geophys. Res. Lett.*, **20** (1993), 1527—1536•
- 18 D. Prichard and C. P. Price, Reconstructing attractors from scalar time series: a comparison of singular system and redundancy criteria, *Geophys. Res.*, **20** (1993), 2817—2825•

The Influence of the Different Distributed Phase_Randomized on the Experimental Data Obtained in Dynamic Analysis

Ma Junhai

(Institute of Systems Engineering, Southeast University, Nanjing 210096, P. R. China)

Chen Yushu

(Department of Mechanics, Tianjin University, Tianjin 300072, P. R. China)

Liu Zengrong

(Department of Mathematics, Shanghai University, Shanghai 201800, P. R. China)

Abstract

In this paper the influence of the differently distributed phase_randomized to the data obtained in dynamic analysis for critical value is studied. The calculation results validate that the sufficient phase_randomized of the different distributed random numbers are less influential on the critical value. This offers the theoretical foundation of the feasibility and practicality of the phase_randomized method.

Key words experimental data, surrogate data, critical value, phase_randomized, random timeseries, chaotic timeseries