

文章编号:1007-6735(2013)04-0307-06

基于消除趋势波动分析的上海世博会场馆排队分析

郭 伟¹, 顾基发², 徐山鹰², 张 毅¹, 李 力¹

(1.清华大学 自动化系, 北京 100084; 2.中国科学院 数学与系统科学研究院, 北京 100190)

摘要: 2010 年,上海市举办了第 41 届世界博览会,愈 7 300 万的参观人次也创下了历届之最.如此巨大的参观人次势必造成大量场馆特别是热门场馆的排队和等待现象.研究排队和等待现象的时间关联性,有助于引导参观人群合理选择场馆、提高参观效率,同时避免过度拥挤酿成事故.为此,引入消除趋势波动分析法,利用各场馆日平均排队长度和日平均等待时间数据,分析其长时程关联性.测试表明,大部分场馆的排队和等待现象呈现长时程正相关性,部分热门场馆呈现长时程反相关性.

关键词: 上海世博会; 消除趋势波动分析法; 长时程关联性
中图分类号: O 226 **文献标志码:** A

Analysis on Queuing Behaviors in Shanghai World Expo
Based on Detrended Fluctuation Analysis

GUO Wei¹, GU Ji-fa¹, XU Shan-ying¹, ZHANG Yi¹, LI Li¹

(1. Department of Automation, Tsinghua University, Beijing 100084, China;
2. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In 2010, the 41st World Expo was held in Shanghai. More than 73 million visitors visited the Expo and created a new record. Such a huge number of visitors cause queuing and waiting phenomena around popular pavilions. Noticing that the time correlation of queuing length and waiting time helps visitors better choose their desired pavilions, a detrended fluctuation analysis (DFA) method to analyze the potential long-range dependence of queuing behaviors. The results show that the queuing length time series of most pavilions yield the long-range correlations; while some of popular pavilions yield the long-range anti-correlation.

Key words: Shanghai World Expo; detrended fluctuation analysis; long-range correlation

2010 年 5 月 1 日到 10 月 31 日,上海市举办了第 41 届世界博览会,共有 240 个国家和地区组织参与,总投资约 450 亿元人民币,建造了占地面积约为 5.28 km² 的世博园区.在世博会举办的 184 天中,总共迎来了 7 308.44 万人参观,也创下了历届世博会之最^[1].如此庞大数量的参观人数,势必造成参观各场馆时的排队现象.根据统计,沙特馆、石油馆、中国馆等热门场馆的平均等待时间达到 5 h 之久,节假日

收稿日期: 2013-07-10
基金项目: 国家重点基础研究发展计划(973)资助项目(2010CB731400)
第一作者: 郭 伟(1988-),男,博士研究生.研究方向:智能交通. E-mail: guo-w11@mails. tsinghua. edu. cn
通讯作者: 顾基发(1935-),男,研究员.研究方向:运筹学、系统科学. E-mail: ifgu@amss. ac. cn

日可能达到 6~7 h. 因此, 研究世博会各场馆排队长度和排队等待时间的长时程相关性就显得尤为重要^[2].

分析采集到的排队长度和排队等待时间序列的一个重要难点在于其非平稳性. 为此, 引入消除趋势波动分析法 (detrended fluctuation analysis, DFA), 将各个场馆每日的平均排队长度、平均等待时间以及每日入园总人数作为非稳定时间序列, 通过研究其长时程相关性, 分析不同特性的场馆 (热门场馆、一般场馆和冷门场馆) 每日排队和等待状况, 希望对大型活动参观时人群的路径选择提供参考.

DFA 方法^[3]是一种研究时间序列长期相关性的常用方法, 在非平稳时间序列的分形和多重分形特性、长时程关联性识别等方面已成为一项广泛使用的技术^[4-5], 并在 DNA 序列^[3,6]、心律动态^[7,8]、神经元尖峰^[9]、人体步态^[10]、长期天气记录^[11]、云结构^[12]、地质学^[13]、人类学^[14]、经济学^[15]中得到了成功应用. 相比于传统的功率谱和关联性分析方法, DFA 的最大优势在于可以消除因为时间序列不稳定而造成的伪相关性的干扰.

DFA 方法不仅可以用于研究非平稳时间序列的长时程幂律相关性, 也有助于识别同一系统在不同尺度行为下的不同状态. 例如, 在心率动态检测中, 心跳间隔对应的不同幂律指数可以用来区分健康状态和患病状态^[16-18]. 在交通系统中, 出行者的路径选择相关性一直是该领域研究的十分重要的问题, Wu 等^[19]引入 DFA 方法对城市轨道交通系统中乘客的路径选择行为进行分析, 从路径交通流量演化过程入手, 探究在路网到达均衡状态之前的路径选择行为的相关性.

本文对 DFA 方法进行数学描述, 指出不同取值区间的赫斯特指数 H 所代表的不同含义. 以石油馆平均排队长度和平均等待时间的序列为例, 逐步计算配置序列、均方根值、局部趋势、波动函数, 并最终求取每个场馆的 H ; 然后把所有场馆的 H 进行聚类分析, 挖掘热门场馆和冷门场馆 H 的内在联系, 并解释了序列中零值对 H 的影响, 最终给出了相应的结论.

1 消除趋势波动分析法

DFA 方法通常用以下 5 个步骤进行描述, 首先假定一个长度为 N 的时间序列 x_k .

步骤 1 定义配置序列 Y

$$Y(i) \equiv \sum_{k=1}^i [x_k - \langle x \rangle], \quad i = 1, \dots, N \quad (1)$$

其中 $\langle x \rangle$ 为序列 x_k 的平均值.

步骤 2 将配置序列 Y 划分为等长度为 l 的 N_l 非重叠数据段, 其中 $N_l \equiv \text{int}(N/l)$. 通常情况下, 数据序列长度 N 并不是数据段长度 l 的整数倍, 因配置序列会有少量数据被遗漏, 为了避免该情况, 再从尾至首重复步骤 2, 总共获取 $2N_l$ 个数据段.

步骤 3 利用最小二乘拟合方法计算出步骤 2 中获取的 $2N_l$ 个数据段的局部趋势 (local trend). 然后计算每一个数据段的方差.

$$M^2(l, v) \equiv \frac{1}{l} \sum_{i=1}^l \{Y[(v-l)l+i] - y_v(i)\}^2 \quad (2)$$

其中, $v = 1, \dots, N_l$. 且

$$M^2(l, v) \equiv \frac{1}{l} \sum_{i=1}^l \{Y[N - (v - N_l)l + i] - y_v(i)\}^2 \quad (3)$$

其中, $v = N_l + 1, \dots, 2N_l$. 这里, $y_v(i)$ 是数据段 v 的拟合多项式. 除了线性拟合之外, 还可使用二阶、三阶甚至高阶拟合.

步骤 4 计算所有数据段的 q 阶波动函数.

$$M(q, l) \equiv \left\{ \frac{1}{2N_l} \sum_{v=1}^{2N_l} [M^2(l, v)]^{q/2} \right\}^{1/q} \quad (4)$$

其中, q 可以去任何非零实数. 若 $q = 2$, 则为标准 DFA 算法.

步骤 5 计算波动函数的广义赫斯特指数值.

$$F_q(l) \sim l^{h(q)} \quad (5)$$

其中当 $q = 2$ 时, $H = h(2)$ 为赫斯特指数 (Hurst exponent). 同时把 $h(q)$ 称为广义赫斯特指数 (generalized Hurst exponent).

赫斯特指数 H 是用来衡量时间序列长时程记忆性的重要指标. 表 1 给出了不同 H 值所代表的不同含义^[20].

表 1 不同取值区间的赫斯特指数含义表
Tab. 1 Illustration of Hurst exponent among different intervals

H 值	代表含义	中文释义
$0 < H < 0.5$	long-range anti-correlation	长时程反相关性
$H = 0.5$	white noise	白噪声
$0.5 < H < 1.0$	long-range correlation	长时程正相关性
$H = 1.0$	pink noise	粉红噪声
$1.0 < H < 1.5$	random walks	随机游走
$H = 1.5$	brown noise	棕褐噪声

2 试验与讨论

按照描述的 DFA 方法,将每个场馆的日平均等待时间、日平均排队长度两个非稳定的时间序列应用 DFA 方法,并将相关结果绘制成图.本试验所使用的数据集时间跨度为 2010 年 8 月 1 日至 2010 年 10 月 31 日共 92 天.涉及到总共 134 个场馆的日平均排队长度(单位:人)、日平均等待时间(单位:min).

2.1 DFA 方法的直接应用

一般而言,可以将按照描述的 DFA 方法中的 5 个步骤,依次得到配置序列 Y 、均方根值(root mean square, RMS)、局部趋势(local trend)、波动函数(fluctuation function)以及赫斯特指数(Hurst exponent).

以石油馆的排队长度和等待时间数据为例,绘制每一步骤得到的结果.

步骤 1 根据时间序列 X 求取配置序列 Y ,见图 1.

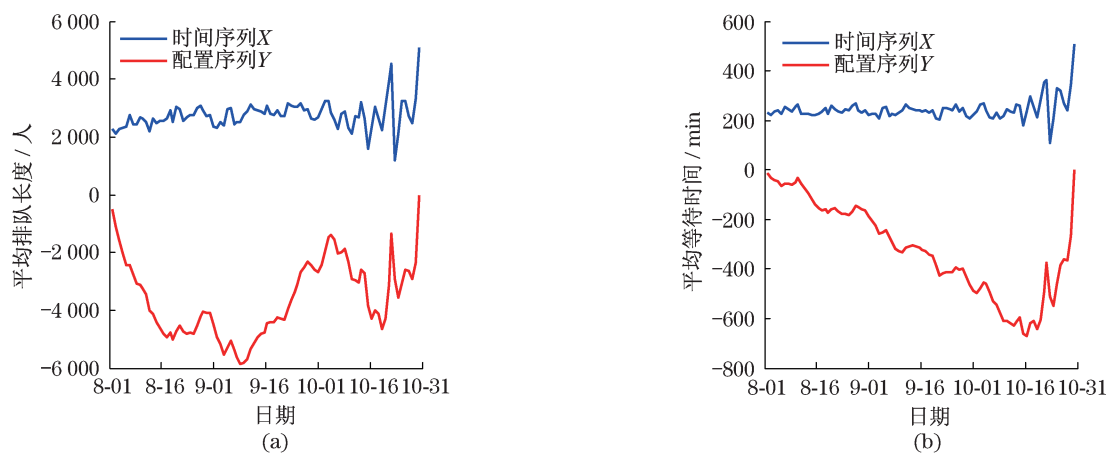


图 1 石油馆排队长度和等待时间的配置序列 Y

Fig.1 Profile Y for queuing length and waiting time about the Oil Exhibition Hall

步骤 2 计算时间序列 X 的均方根值,见图 2, 为下一步的局部趋势拟合做准备.

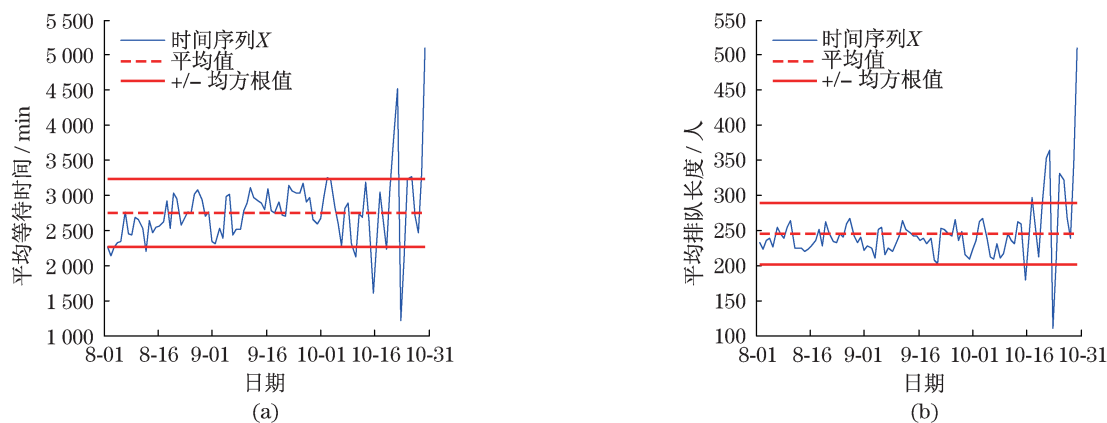


图 2 石油馆排队长度和等待时间序列的均方根值

Fig.2 RMS values for queuing length and waiting time about the Oil Exhibition Hall

步骤 3 采用线性、二阶和三阶拟合计算各数据段的局部趋势,见下页图 3.

步骤 4 计算不同数据段长度 l 情况下的局部波动,见下页图 4.

步骤 5 通过得到的局部波动函数,计算其广义赫斯特指数,见下页图 5.

通过以上步骤,得到了每一个场馆平均排队长度和平均等待时间的广义赫斯特指数,并可根据表 1 得出每一个场馆的长时程关联性.

2.2 去除异常零值的 DFA 方法应用

但是如果时间序列中出现大量异常零值,则会对 Hurst 指数的计算产生偏差.通过石油馆中的异

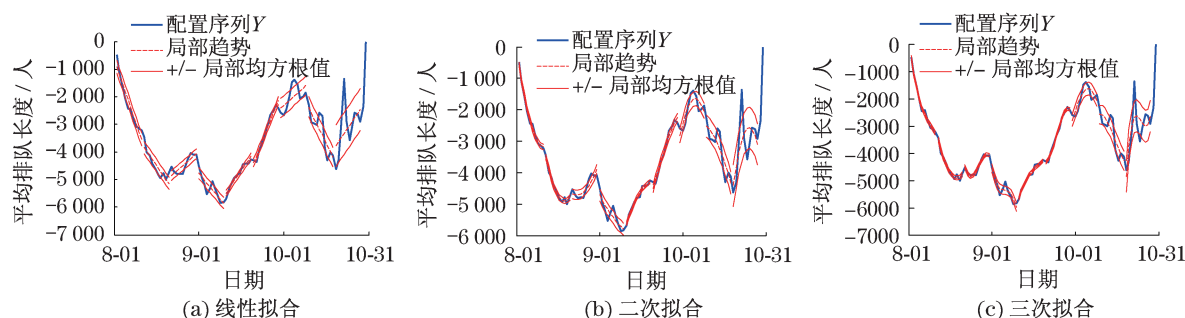


图3 石油馆排队长度序列的局部趋势示意图

Fig.3 Computation of local trend based on queuing length about the Oil Exhibition Hall

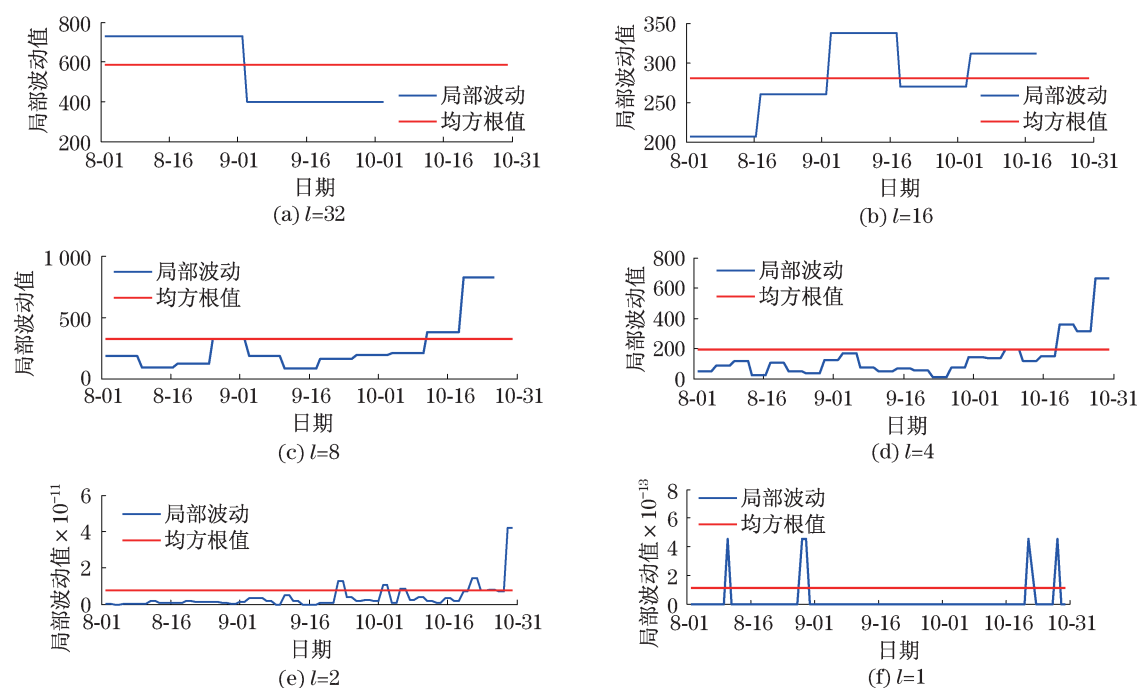


图4 石油馆不同数据段长度下的局部波动示意图

Fig.4 Local fluctuations for segments with multiple sizes about the Oil Exhibition Hall

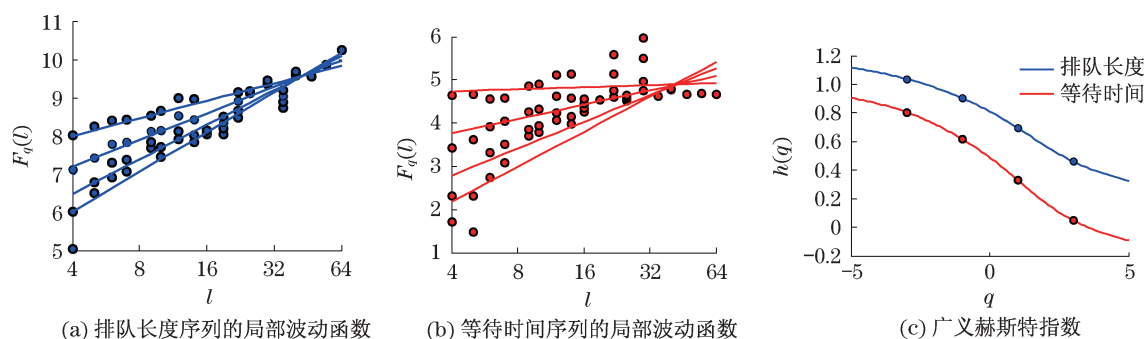


图5 石油馆的局部波动函数和广义赫斯特指数

Fig.5 Local fluctuation function and generalized Hurst exponent for the Oil Exhibition Hall

常零值数据来说明以上结论. 根据上述步骤算出的结果, 沙特馆和石油馆都是呈长时程反相关性, 图6给出了沙特馆和石油馆时间序列的对比示意图. 从图中可以看出, 沙特馆的平均等待时间为锯齿状, 呈

现明显的长时程反相关性. 而石油馆第44, 78, 80, 82天出现异常零值或者异常峰值, 有可能是检测设备失灵或者异常情况闭馆造成的.

将这些零值做最简单的线性插值, 令 $x_a =$

$(x_{a-1} + x_{a+1})/2$, 其中 a 为代表出现异常的日期. 将石油馆平均排队长度去除零值之后重新计算 Hurst 指数, 得到 H 值由原来的 0.34 修正为 0.57, 其长时程关联性也由反相关性修正为正相关性. 可见, 零值的存在会使得计算出的 Hurst 指数存在偏差.

通常情况下, 冷门场馆存在大量零值, 即在大部分时间里, 参观冷门场馆是不需要排队和等待的. 用大量存在零值的时间序列计算出的 Hurst 指数势必存在较大误差. 因此在校正之后, 这部分存在零值的冷门场馆在分析长时程关联性的时候不予考虑.

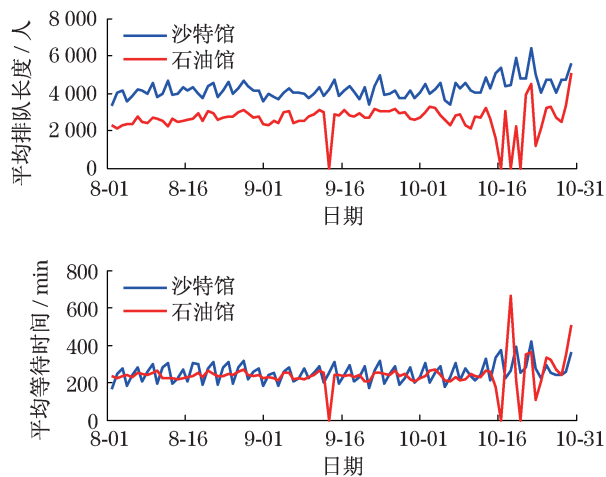


图6 沙特馆和石油馆的时间序列对比示意图

Fig.6 Comparison between the time series of the Saudi Arabia and Oil Exhibition Hall

2.3 去除异常零值后的 DFA 结果分析

根据之前的计算, 每个场馆的排队长度和等待时间序列都得到一个 H 值. 不妨令第 i 个场馆的 H 值为 $H_i = (H_{\text{queue}, i}, H_{\text{wait}, i})$, 其中 $i = 1, 2, \dots, L$, L 为总场馆数. 将所有场馆的 H 值绘制出来, 如图 7 所示.

下面按照聚类的结果对每一类 H 值的特性进行分析:

a. 青色点: $H_{\text{queue}, i} \in (0.5, 1)$, $H_{\text{wait}, i} \in (0.5, 1)$. 绝大部分场馆的平均排队长度和平均等待时间呈现长时程正相关性(long-range correlation).

b. 红色点: $H_{\text{queue}, i} \in (0, 0.5)$, $H_{\text{wait}, i} \in (0, 0.5)$. 这部分场馆的平均排队长度和平均等待时间呈现短时程反相关性(long-range anti-correlation). 主要有以下 12 个场馆: 沙特馆、荷兰馆、城市人馆、石油馆、中南美洲联合馆、城市人馆、古巴馆、加共体联合馆、立陶宛馆、尼日利亚馆、亚洲联合馆 2、宁波

案例馆. 其中沙特馆和石油馆是非常热门的场馆, 沙特馆平均排队长度为 4 237 人, 平均等待时间为 254 min; 石油馆的平均排队长度是 2 639 人, 平均等待时间为 240 min. 剩余场馆均为时间序列中存在大量零值的冷门场馆.

c. 蓝色点: $H_{\text{queue}, i} \in (0, 0.5)$, $H_{\text{wait}, i} \in (0.5, 1)$. 这部分场馆主要有: 土耳其馆、瑞典馆、芬兰馆、民营企业联合馆和中国企业联合馆. 通过绘制其排队长度和等待时间的时间序列, 发现这些时间序列的相关性并不明显, 其 H 值可近似认为是 0.5, 即呈现高斯白噪声特性.

d. 绿色点: $H_{\text{queue}, i} \in (0.5, 1)$, $H_{\text{wait}, i} \in (0, 0.5)$. 这部分场馆为日本馆、韩国馆、伊朗馆、台湾馆、朝鲜馆和万科馆. 其中日本馆和韩国馆为热门场馆. 日本馆的平均排队长度 3 510 人, 平均等待时间为 237 min; 韩国馆平均排队长度 2 444 人, 平均等待时间 173 min. 朝鲜馆为存在大量零值的冷门场馆.

e. 黄色点: $H_{\text{queue}, i} \in (1, 1.5)$, $H_{\text{wait}, i} \in (1, 1.5)$. 这部分的场馆主要有以下 5 个: 分别为澳门馆、欧洲联合馆 1、欧洲联合馆 2、生命阳光馆、世博会博物馆. 其中欧洲联合馆 1 和欧洲联合馆 2、生命阳光馆为存在大量零值的非常冷门的场馆.

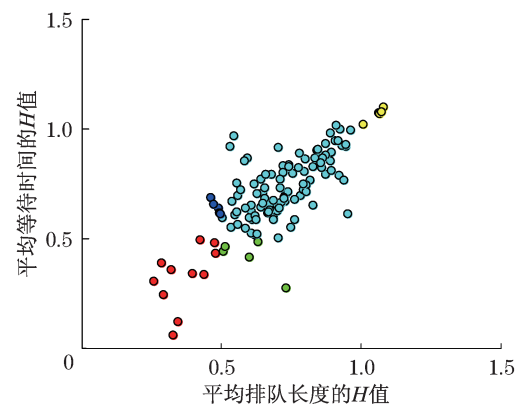


图7 所有场馆的 H 值聚类示意图

Fig.7 Cluster analysis of Hurst exponent for all exhibition halls

3 结 论

本文通过将 DFA 方法引入上海世博会场馆的排队和等待数据的测试和分析中, 通过计算不同场馆之间的 Hurst 指数, 分析各场馆的长时程相关性, 得到以下结论:

a. 时间序列中零值的存在会对长时程关联性产生影响,需要进行合理的补偿之后再进行计算.

b. 大部分场馆呈现长时程正相关性.

c. 部分非常热门的场馆呈现长时程反相关性,代表场馆为沙特馆、日本馆和韩国馆,其具体原因尚有待考察.

参考文献:

- [1] 2010 上海世博会概况 [EB/OL]. (2010-12-18) [2013-07-01]. <http://www.expo2010.cn/>.
- [2] 顾基发,徐山鹰,房勇,等. 世博会排队集群行为研究[J]. 上海理工大学学报,2011,33(4):312-320.
- [3] Peng C K, Buldyrev S V, Havlin S, et al. Mosaic organization of DNA nucleotides[J]. Physical Review E,1994,49(2):1685-1689.
- [4] Taqqu M S, Teverovsky V, Willinger W. Estimators for long-range dependence: an empirical study [J]. Fractals,1995,3(4):785-798.
- [5] Kantelhardt J W, Zschiegner S A, Koscielny-Bunde E, et al. Multifractal detrended fluctuation analysis of nonstationary time series [J]. Physica A: Statistical Mechanics and Its Applications, 2002, 316 (1): 87-114.
- [6] Buldyrev S V, Goldberger A L, Havlin S, et al. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis [J]. Physical Review E, 1995, 51(5): 5084-5091.
- [7] Viswanathan G M, Buldyrev S V, Havlin S, et al. Optimizing the success of random searches [J]. Nature, 1999, 401(6756): 911-914.
- [8] Bunde A, Havlin S, Kantelhardt J W, et al. Correlated and uncorrelated regions in heart-rate fluctuations during sleep [J]. Physical Review Letters, 2000, 85 (17): 3736.
- [9] Blesić S, Milošević S, Stratimirović D, et al. Detrended fluctuation analysis of time series of a firing fusimotor neuron [J]. Physica A: Statistical Mechanics and Its Applications, 1999, 268(3): 275-282.
- [10] Hausdorff J M, Edelberg H K, Mitchell S L, et al. Increased gait unsteadiness in community-dwelling elderly fallers [J]. Archives of Physical Medicine and Rehabilitation, 1997, 78(3): 278-283.
- [11] Koscielny-Bunde E, Bunde A, Havlin S, et al. Indication of a universal persistence law governing atmospheric variability [J]. Physical Review Letters, 1998, 81(3): 729-732.
- [12] Ivanova K, Ausloos M, Clothiaux E E, et al. Break-up of stratus cloud structure predicted from non-Brownian motion liquid water and brightness temperature fluctuations [J]. EPL, 2000, 52(1): 40-46.
- [13] Malamud B D, Turcotte D L. Self-affine time series: measures of weak and strong persistence [J]. Journal of Statistical Planning and Inference, 1999, 80(1): 173-196.
- [14] Alados C L, Huffman M A. Fractal long-range correlations in behavioural sequences of wild chimpanzees: a non-invasive analytical tool for the evaluation of health [J]. Ethology, 2000, 106(2): 105-116.
- [15] Liu Y, Gopikrishnan P, Stanley H E. Statistical properties of the volatility of price fluctuations [J]. Physical Review E, 1999, 60(2): 1390-1400.
- [16] Ashkenazy Y. The use of generalized information dimension in measuring fractal dimension of time series [J]. Physica A: Statistical Mechanics and Its Applications, 1999, 271(3): 427-447.
- [17] Peng C K, Havlin S, Stanley H E, et al. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series [J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 1995, 5(1): 82-87.
- [18] Havlin S, Buldyrev S V, Goldberger A L, et al. Fractals in biology and medicine [J]. Chaos, Solitons & Fractals, 1995, 6: 171-201.
- [19] Wu J, Sun H, Wang D Z W, et al. Bounded-rationality based day-to-day evolution model for travel behavior analysis of urban railway network [J]. Transportation Research Part C: Emerging Technologies, 2013, 31: 73-82.
- [20] Ihlen E A F. Introduction to multifractal detrended fluctuation analysis in Matlab [J]. Frontiers in Physiology, 2012, 3: 141.

(编辑:董伟)