

基于最小二乘谐波拟合的置信水平计算方法

马 民, 汪春峰

(河南师范大学 数学与信息科学学院, 河南 新乡 453007)

摘 要:结合 Lomb 最小二乘谐波拟合的方法和 Scargle 归一化功率谱的置信水平的解析公式, 给出了一种计算谐波拟合置信水平的方法. 该方法的优势在于: 直接计算拟合量的振幅, 并给出拟合振幅的置信水平, 物理意义明确; 能够对多个频率进行扫描和拟合并给出各频率对应振幅的置信水平; 通过对置信水平最大的频率进行拟合, 获得时间序列的周期性变化规律. 最后, 通过数值试验说明了本方法的优势和实用性.

关键词:最小二乘法; 谐波拟合; 置信水平

中图分类号:O241.6

文献标志码:A

分析时间序列的周期/频率(或者空间序列的波长/波数)在分析测量数据方面具有重要应用. 通过这一分析, 研究人员能够更好地解释测量数据, 从而揭示测量数据所蕴含的物理规律. 比如, 在研究大气参数(温度、风速、压强、密度)所表现出的周期性时空变化时, 如何准确获得时空序列的周期和波长及其对应的振幅在预测这些变化的周期和强度方面就非常重要^[1-2]. 深入了解这些周期性变化的规律有助于人们建立合理的模型来描述这些规律, 进而能够更为准确的了解大气运动^[3-4].

提取时间序列的周期和振幅的方法有傅里叶变换、小波分析等, 但这些方法仅适用于均匀采样的时间序列^[5-8]. 而在实际测量中, 受天气、地势和设备等因素的影响, 很容易遇到非均匀采样的情形. 对于非均匀采样数据可以通过插值得到均匀分布的数据, 然后用傅里叶变换或者小波分析来提取周期和振幅. 必须说明的是, 在将非均匀采样数据插值到均匀分布数据的过程中, 会因为插值而引入误差^[9-10]. 因此, 提取非均匀采样数据的周期和振幅并对其进行置信水平检验就成为一个重要研究问题.

Lomb-Scargle 周期谱能够给出非均匀采样时间序列的频率/周期的归一化功率谱及相应的置信水平^[11-12]. Scargle 引入归一化功率谱的重要作用是能够给出归一化功率谱置信水平的解析公式^[12], 然而, 归一化功率谱虽然在形式上具有振幅的量纲, 能够给出拟合量的振幅, 但不能给出拟合量的相位, Hocke^[13]给出了拟合相位的计算方法. 至此, 可以用 Lomb-Scargle 周期谱计算拟合量的振幅和相位, 但计算过程较为烦琐, 难以从拟合结果中直接分析物理规律.

值得注意的是, Scargle 虽然没有给出拟合振幅置信水平的方法, 但给出了归一化周期谱与最小二乘谐波拟合之间的等价关系. 本文将在此等价关系的基础上, 给出计算拟合振幅置信水平的方法.

1 谐波拟合与功率谱置信水平

1.1 最小二乘谐波拟合

考虑如下的非均匀采样时间序列

$$y_i = y(t_i), \quad i = 1, \dots, N. \quad (1)$$

希望分析此时间序列的周期/频率及其对应振幅的置信水平. Lomb^[1]曾给出用最小二乘谐波拟合计算非均匀采样时间序列的方法, 即用如下的函数来拟合采样点:

收稿日期:2015-11-25; 修回日期:2016-05-13.

基金项目:国家自然科学基金(41374158; 41574143); 河南省高等学校青年骨干教师资助计划项目(2014GGJS-047).

第1作者简介(通信作者): 马 民(1979-), 女, 河南社旗人, 河南师范大学助教, 研究方向为计算数学, E-mail: 013004@htu.cn.

$$y_{fit}(t) = A \cos \omega(t - \tau) + B \sin \omega(t - \tau), \quad (2)$$

使得

$$R(\omega) = \sum_{i=1}^N [y_{fit}(t) - y_i]^2 \quad (3)$$

达到最小,其中 $\omega = 2\pi f$ 是拟合函数的圆频率(angular frequency),时间延迟参数 τ 的定义为,

$$\tan(2\omega\tau) = \frac{\left(\sum_{i=1}^N \sin \omega t_i\right)}{\left(\sum_{i=1}^N \cos \omega t_i\right)}.$$

根据最小二乘法的标准推导过程,可以得到(2)式中 A 和 B 的表达式,

$$A = \frac{Y_C \cdot S_S - Y_S \cdot C_S}{C_C \cdot S_S - C_S^2}, B = \frac{Y_C \cdot C_C - Y_C \cdot C_S}{C_C \cdot S_S - C_S^2}, \quad (4)$$

其中,

$$C_C = \sum_{i=1}^N \cos^2 \omega(t_i - \tau), S_S = \sum_{i=1}^N \sin^2 \omega(t_i - \tau), C_S = \sum_{i=1}^N \cos \omega(t_i - \tau) \cdot \sin \omega(t_i - \tau),$$

$$Y_C = \sum_{i=1}^N y_i \cos \omega(t_i - \tau), Y_S = \sum_{i=1}^N y_i \sin \omega(t_i - \tau).$$

至此,可以通过(4)式计算非均匀采样时间序列(1)式在最小二乘意义下的振幅和相位,但无法计算振幅的置信水平. Lomb^[11]指出,可以通过大量(至少超过100次)的随机试验来计算置信水平.然而,大量的随机试验使得计算量迅速增大,降低了解决问题的效率.因此,有必要寻找更为简便的、计算效率更高的计算置信水平的方法.

1.2 功率谱的置信水平

Scargle^[12]通过引入一个新的功率谱公式,

$$P_y(\omega) = \frac{1}{2\sigma^2} \left\{ \frac{\left[\sum_{i=1}^N (y_i - \bar{y}) \cos \omega(t_i - \tau) \right]^2}{\sum_{i=1}^N \cos^2 \omega(t_i - \tau)} + \frac{\left[\sum_{i=1}^N (y_i - \bar{y}) \sin \omega(t_i - \tau) \right]^2}{\sum_{i=1}^N \sin^2 \omega(t_i - \tau)} \right\}, \quad (5)$$

其中,上式中用到了非均匀采样时间序列 $y_i, i = 1, \dots, N$ 的方差 σ 和均值 \bar{y} ,其定义分别为

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

文献[12]指出,公式(5)具有很好的统计行为,并与文献[11]给出的(3)式存在一定的等价关系.文献[12]根据公式(5)计算出的功率谱的拒绝水平($P_r, 1 - P_r$ 即为置信水平)的解析表达式如下,

$$P_r \{ P_y(\omega) > z \} = 1 - [1 - e^{-z}]^N. \quad (6)$$

上式用语言叙述为功率谱 $P_y(\omega)$ 大于某一数值 z 的概率表示为采样点数和数值 z 的函数.如果拒绝水平 P_r 较小,则置信水平较大.因此,数值 z 越小(即要求较小的功率谱 $P_y(\omega)$),则拒绝水平 P_r 较大,置信水平较小;反之, z 越大,置信水平越高,这就要求满足公式(6)的 $P_y(\omega)$ 越大.

2 谐波拟合振幅的谱置信水平

公式(6)给出了用公式(5)计算出的功率谱的置信水平,然而公式(5)虽然具有振幅的量纲,但不能直接与振幅联系在一起,因此,公式(6)不能给出通过(4)式计算出的振幅的置信水平.而振幅能够准确反映物理量的周期性变化强度,在数据分析中具有明确的物理意义.本部分借助于文献[12]给出的公式(5)与公式(3)之间的等价关系,给出计算振幅(由公式(4)给出)置信水平的解析表达式.把公式(4)中的 A 和 B 代入公式(2),可以使得 $R(\omega)$ 达到最小值,记为 $R_{\min}(\omega)$. Scargle^[12]引入

$$\Delta R(\omega) = \sum_{i=1}^N y_i^2 - R_{\min}(\omega), \quad (7)$$

通过选择合适的参数 τ , 使得 $C_s = 0$, 从而 $R(\omega)$ 可以简化为

$$\Delta R(\omega) = \frac{\left[\sum_{i=1}^N (y_i - \bar{y}) \cos \omega(t_i - \tau) \right]^2}{\sum_{i=1}^N \cos^2 \omega(t_i - \tau)} + \frac{\left[\sum_{i=1}^N (y_i - \bar{y}) \sin \omega(t_i - \tau) \right]^2}{\sum_{i=1}^N \sin^2 \omega(t_i - \tau)}.$$

把上式代入公式(5)中, 结合公式(7), 可以得到

$$P_y(\omega) = \frac{1}{2\sigma^2} \Delta R(\omega) = \frac{1}{2\sigma^2} \left[\sum_{i=1}^N y_i^2 - R_{\min}(\omega) \right]. \quad (8)$$

因此, 可以根据拟合振幅 A 和 B , 计算 $R_{\min}(\omega)$, 进而得到公式(8)的右端项, 这就可以借助于计算功率谱置信水平的公式(6)来计算振幅的置信水平. 此外, 如果在公式(6)的右端项中, 令 $z = P_y(\omega)$, 可以通过下式计算频率 ω 所对应振幅的置信水平 $\text{Sig}(\omega)$, 即公式(6)中的 $1 - P_r$,

$$\text{Sig}(\omega) = [1 - e^{-P_y(\omega)}]^N. \quad (9)$$

公式(9)给出了用单个频率 ω 拟合非均匀采样时间序列的振幅的置信水平. 需要注意的是, 公式(9)中虽然用到了 $P_y(\omega)$, 但在计算过程中并不需要通过公式(5)来计算 $P_y(\omega)$ 和相应的置信水平, 而是通过公式(8)的右端项来计算置信水平.

如果事先不知道时间序列的周期或者频率, 可以用公式(2)进行拟合, 再用公式(9)计算拟合结果的置信水平, 通过置信水平的极大值来判断此时间序列的周期或者频率(称为主要周期或者频率), 同时获取该周期或者频率对应的振幅, 这一过程可以称为周期或者频率扫描. 下面将通过数值试验来说明如何利用拟合振幅及其置信水平选择周期或者频率, 并给出利用主要周期或者频率的拟合效果.

3 数值试验

3.1 单频率试验

假设一物理量随时间的演化过程可以用如下的谐波函数描述,

$$y = \cos(4\pi ft) + R_n, \quad 0 \leq t \leq 1.$$

对此时间序列, 根据是否均匀采样和是否添加噪声 R_n 共4种方式获得物理量的采样值. 其中, 非均匀采样点用 $[0, 1]$ 之间具有均匀分布的随机数生成, 噪声用 $[-0.3, 0.3]$ 之间的均匀分布生成. 试验编号和试验条件如表1所示.

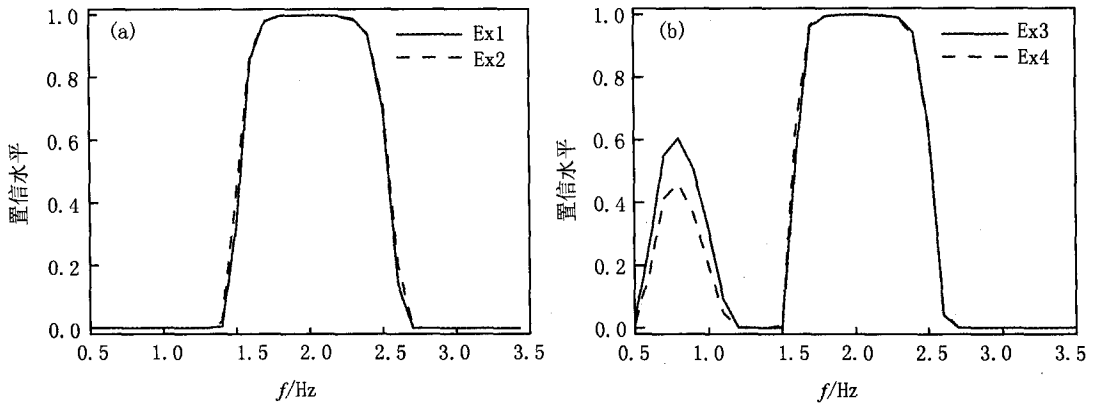
表1 试验编号和试验条件

试验编号	Ex1	Ex2	Ex3	Ex4
是否均匀采样	是	是	否	否
有无噪声	无	有	无	有

首先对4次试验的时间序列进行频率扫描, 以确定主要频率. 注意这里的频率扫描并非仅限于频率为整数的情形, 对频率不为整数的情况仍然可以扫描. 图1给出了扫描频率范围从0.5~3.5 Hz 每次增加0.1 Hz时, 置信水平随着扫描频率的变化情况. 从图1中可以看出, 不管对于均匀采样试验(Ex1和Ex2)还是非均匀采样试验(Ex3和Ex4), 置信水平在1.7~2.4 Hz 范围内较大(>0.95). 而对于非均匀采样试验, 置信水平在0.8 Hz处也有一个较小的极大值, 分别是0.6(Ex3)和0.45(Ex4). 通过与Ex1和Ex2比较, 这个较小的极值完全是由于非均匀采样引起的.

虽然图1中给出的置信水平在1.7~2.4 Hz 范围内都较大, 但其最大值均在2.0 Hz处. 下面用本文给出的方法给定频率 $f=2$ Hz 的情况计算拟合振幅及其相应置信水平.

表2中给出了这4次试验的拟合振幅以及相应的置信水平. 表2中给出的这4个试验的置信水平均超过0.995, 说明拟合结果是很可靠的. 在不加入噪声的情况下(Ex1和Ex3), 不管是否均匀采样, 拟合振幅的相对误差均在0.5%以内. 而加入噪声之后(Ex2和Ex4), 拟合振幅的相对误差变大.



注：图中实线和虚线分别表示无噪声和有噪声时的置信水平。
图1 对均匀采样(a)和非均匀采样(b)时间序列的置信水平随着扫描频率的变化情况

表 2 具有单频率采样数据拟合振幅以及相应的置信水平

试验编号	Ex1	Ex2	Ex3	Ex4
拟合振幅	0.9952	0.8583	0.9990	1.0125
置信水平	0.998 50	0.997 74	0.998 50	0.998 17

图 2 更加全面地展示了采样数据和拟合结果,其中图 2(a)和 2(b)分别给出了均匀采样及其拟合结果,图 2(c)和 2(d)分别给出了非均匀采样及其拟合结果.从图 2 中可以看出,不管数据采样是否均匀,最小二乘谐波拟合都可以给出很好的拟合结果,使得采样数据的物理意义更加明确.

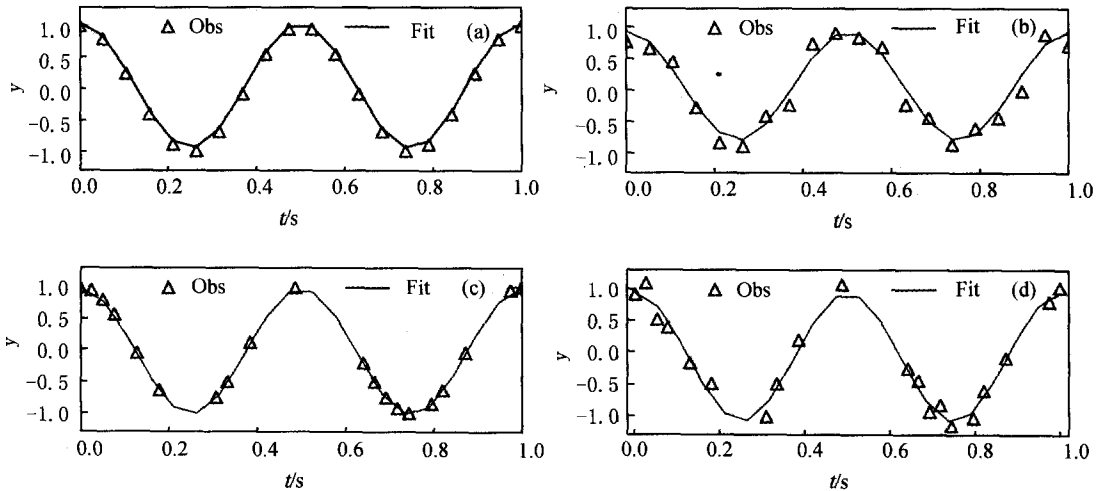


图2 试验Ex1 (a), Ex2 (b), Ex3 (c), Ex4 (d)的采样数据(Obs)和拟合曲线(Fit)

3.2 多频率试验

对于多频率试验,仍然按照单频率试验的过程,假设一物理量随时间的演化过程可以用如下的谐波函数描述,

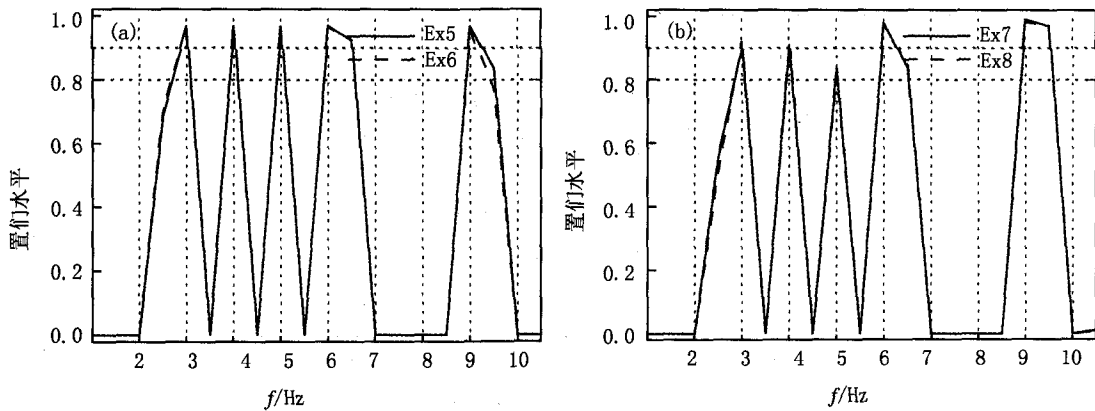
$$y = \sum_{i=1}^5 \cos(4\pi f_i t) + R_n, 0 \leq t \leq 1, f_i = 3, 4, 5, 6, 9.$$

这里的谐波函数包括5个主要频率,其中是 $f_i = 3, 4, 5, 6$ 是连续的,而 $f_i = 9$ 与其他几个频率是不连续的.对此时间序列,根据是否均匀采样和是否添加噪声 R_n 共4种方式获得物理量的采样值.其中,非均匀采样用 $[0, 1]$ 之间具有均匀分布的随机数生成,噪声用 $[-0.3, 0.3]$ 之间的均匀分布生成.试验编号和试验条件如表 3 所示.

表 3 试验编号和试验条件

试验编号	Ex5	Ex6	Ex7	Ex8
是否均匀采样	是	是	否	否
有无噪声	无	有	无	有

首先对 4 次试验的时间序列分别进行频率扫描,以确定主要频率.图 3 给出了扫描频率及其对应置信水平.从图 3 中可以看出,5 个主要频率(3, 4, 5, 6, 9)与理论值一致,均具有较大的置信水平,其中均匀采样试验的置信水平均超过 0.9(图 3(a)),非均匀采样试验的置信水平超过 0.8(图 3(b)).根据主要扫描得到的 5 个主要频率,对 4 种采样数据进行拟合的结果如图 4 所示,显示出最小二乘谐波拟合对多个频率拟合时也有较好的效果,更好的反映了采样数据的物理规律.



注:图中实线和虚线分别表示无噪声和有噪声时的置信水平,图中的两条点线表示的置信水平分别为0.8和0.9.

图3 对均匀采样(a)和非均匀采样(b)时间序列的置信水平随着扫描频率的变化情况

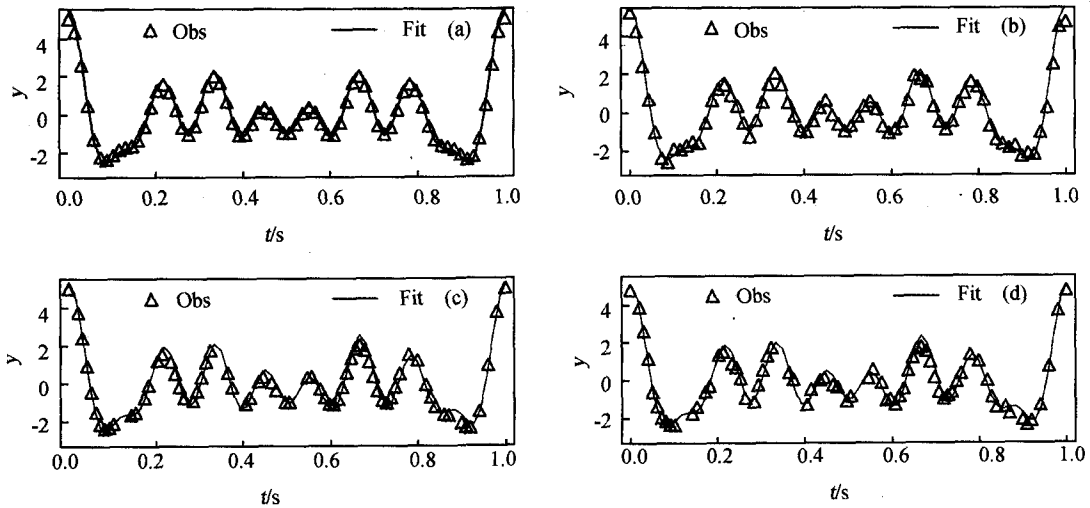


图4 试验Ex5 (a), Ex6 (b), Ex7 (c), Ex8 (d)的采样数据(Obs)和拟合曲线(Fit)

参 考 文 献

[1] Salby M L. Sampling theory for asynoptic satellite observations. Part 1: Space-time spectra, resolution, and aliasing[J]. J Atmos Sci, 1982,39:2577-2600.

[2] Wu D L, Hays P B, Skinner W R. A least squares method for spectral analysis of space-time series[J]. J Atmos Sci,1995,52(20): 3501-3511.

[3] Zhang X, Forbes J M, Hagan M E, et al. Monthly tidal temperatures 20–120 km from TIMED/SABER[J]. J Geophys Res,2006,111:

A10S08. doi:10.1029/2005JA011504.

- [4] Xu J, Smith A K, Liu H L, et al. Seasonal and quasibiennial variations in the migrating diurnal tide observed by Thermosphere, Ionosphere, Mesosphere, Energetics and Dynamics (TIMED)[J]. *J Geophys Res*, 2009, 114: D13107. doi:10.1029/2007JD011298.
- [5] 周永宏, 郑大伟. 小波变化及其在地球自转研究中的应用[J]. *计算物理*, 2000, 17(Z1): 209-214.
- [6] 范伟军, 周必方, 夏良正, 等. 基于最小二乘法准则的光学综合孔径观测目标傅里叶信息提取[J]. *光子学报*, 2004, 33(4): 472-475.
- [7] 王丹, 姜小光. 利用 NOAA 数据分析中国地区植被覆盖变化周期[J]. *中国图象图形学报*, 2006, 11(4): 516-520.
- [8] 徐华君, 柳林涛, 罗孝文. 标准 Morlet 小波变换检测地球自由振荡[J]. *测绘学报*, 2009, 38(1): 16-21.
- [9] 杜党波, 张伟, 胡昌华, 等. 含缺失数据的小波-卡尔曼滤波故障预测方法[J]. *自动化学报*, 2014, 40(10): 2115-2125.
- [10] 刘仁强, 易帆. 极区夏季中间层半日潮汐的 VHF 雷达观测[J]. *空间科学学报*, 2000, 20(3): 233-241.
- [11] Lomb N R. Least-square frequency analysis of unequally spaced data[J]. *Astrophys Space Sci*, 1976, 39: 447-462.
- [12] Scargle J D. Studies in astronomical time series analysis, II. Statistical aspects of spectral analysis of unevenly spaced data[J]. *Astrophys J*, 1982, 263: 835-853.
- [13] Hocke K. Phase estimation with the Lomb-Scargle periodogram method[J]. *Ann Geophys*, 1998, 16: 356-358.

A Method for the Confidence Level Calculation Based on the Least Square Harmonic Fitting

MA Min, WANG Chunfeng

(School of Mathematics and Information Science, Henan Normal University, Xixiang 453007, China)

Abstract: The least square harmonic fitting (LSHF) method proposed by Lomb has obvious physical meaning, the method of confidence level calculation proposed by Scargle has an explicit expression on the confidence of the normalized power spectrum. Based on the above two advantages, we propose a method for the confidence level calculations based on the LSHF. The advantages of our method are; the amplitude and its confidence level are calculated directly. And thus preserve the physical meaning the observation data; multiple frequencies can be scanned and fitted, their confidence levels can also be given simultaneously; the periodic nature of the time series can be explored by fitting the frequency with maximum confidence. Finally, numerical experiments are performed to illustrate the advantages and applications of our method.

Keywords: least squares method; harmonic fitting; confidence level