

基于 K 均值聚类 EEMD 的 CORS 高程时间序列信号分析方法

张恒璟^{1,2} 齐 昕^{1,2} 文汉江²

1 辽宁工程技术大学测绘与地理科学学院,阜新市玉龙路88号,123000

2 中国测绘科学研究院,北京市莲花池西路28号,100830

摘 要: 针对传统 EEMD 进行信号分解时信噪比低和部分模态混叠的问题,提出基于 K 均值聚类的 CORS 高程时间序列改进分析方法。通过添加正负白噪声的 EEMD 提高信号分解信噪比,基于 K 均值聚类方法对 EEMD 迭代过程中分解的各个 IMF 分量进行聚类分析。实验结果表明,该方法提高信噪比 3% 以上,基于正交指数的分解精度提高 26% 以上,聚类结果能够解决 IMF 中近似的 0.5 a、1 a、2 a 周期信号的模态混叠问题。

关键词: 高程时间序列;模态混叠; K 均值聚类;信噪比;正交指数

中图分类号: P228

文献标识码: A

整体经验模式分解(EEMD)是基于经验模式分解(EMD)存在模态混叠现象进行改进的方法^[1],虽解决了 EMD 的一些缺陷,但仍存在分解后信噪比降低、模态混叠等问题^[2-4]。

本文提出基于 K 均值聚类 EEMD 的 CORS 高程时间序列信号分析方法,添加正负白噪声提高信噪比,引入统计学中的 K 均值(K -means)对 CORS 高程时间序列进行聚类分析^[5-6],并用 2 个 CORS 站连续多年的高程时间序列验证该方法的适用性。

1 改进的 EEMD 算法

1.1 基于正负白噪声的 EEMD 算法

1.1.1 EEMD 算法

EEMD 的基本思路是在分解前将原始序列信号加入高斯白噪声生成新的待分解信号,对多个本征模态函数(IMF)分量进行平均处理,进而保留具有物理意义的 IMF 分量,同时消除加入的噪声。对 CORS 高程序列添加白噪声:

$$X_i(t) = X(t) + w_i(t) \quad (1)$$

式中,下标 i 为第 i 次 EMD 分解。该过程的循环次数可以取 100,噪声与信号的标准差之比可以取 0.1、0.2、0.4 等^[1,4]。经过 EMD 生成一系列 IMF 分量,由于每次随机生成的白噪声都是不同

的,可以通过重复分解过程得到信号多次分解的分量,将对应相同分解尺度的分量取平均值,得到最终信号分解的分量。采用式(2)判定产生 IMF 分量的个数:

$$n_{\max} = \text{fix}(\log_2 N) - 2 \quad (2)$$

式中, N 为 CORS 高程时间序列的长度。

1.1.2 添加正负白噪声

白噪声的加入会影响原始时间序列的纯净程度,污染原始信号。分解后的 IMF 分量中存在消除不完全的噪声,使时间序列的信噪比降低,导致信号的重构误差较大。因此采用在加入白噪声的过程中随机添加正、负白噪声^[2]:

$$X_i(t) = X(t) + (-1)^q d_0 w_i(t) \quad (3)$$

式中, $X(t)$ 为原始 CORS 高程时间序列, d_0 为白噪声的幅值标准差。当 q 为奇数时,加入负白噪声;当 q 为偶数时,加入正白噪声,保证所加入的正白噪声总和与负白噪声总和相等。

1.2 K 均值聚类算法

K 均值聚类分析的 EEMD 算法采用划分聚类方法中的 K 均值算法,将每次 EEMD 循环产生的 IMF 分量进行聚类分析,保留聚类个数多的一类,取该类中 IMF 分量的平均值作为分解产生的 IMF。 K 均值聚类算法首先要确定常数 K 为最终分类的数量。令 K 从 2 到某个固定值,在每个 K

收稿日期:2018-11-09

项目来源:高分对地观测系统重大专项(42-Y20A09-9001-17/18);辽宁省教育厅高等学校基本科研项目(LJ2017QL008);辽宁工程技术大学博士启动基金(13-1120)。

第一作者简介:张恒璟,博士,副教授,主要研究方向为空间大地测量数据处理与 GNSS 高程非线性运动,E-mail:sun_winter2009@163.com。

通讯作者:齐昕,硕士生,主要研究方向为空间大地测量数据处理,E-mail:lntuqixin@163.com。

值上运行数次 K 聚类,避免局部最优解,计算 K 的轮廓系数,最后选取轮廓系数最大的值对应的 K 作为最终分类个数^[7]。轮廓系数计算方法为:

$$S(i) = \frac{o(i) - p(i)}{\max[p(i), o(i)]} \quad (4)$$

式中, $o(i)$ 为 i 向量到所有它属类中其他点的距离, $p(i)$ 为 i 向量到所有非本身所在类的点的平均距离。

常数 K 确定后,选取初始点作为质心,分成几类则选择几个初始点。通过计算每个样本与质心的距离来判定样本之间的相似程度,将样本点归到最相似的类中,完成一次聚类。继续计算新生成的各个类的质心,重复计算距离直到质心不再改变,最终确定按所需分成的各类中的样本和各个类的质心。算法本质就是不断更新质心向量,寻找目标函数最小化的过程,通常采用式(5)计算最小方差函数^[8]:

$$E(a_1, \dots, a_k) = \frac{1}{n} \sum_{i=1}^k \sum_{m \in a_i} \|m_j - a_i\|^2 \quad (5)$$

式中, m 为每一个元素, a_i 为第 i 个聚类中心, k 为聚类中心数目。

本文采用欧氏距离作为判断2个IMF分量

相似程度大小的依据,距离越小则相似度越高,距离越大则相似度越低。对于一个 N 维数组,数组 $A(A_1, A_2, A_3, \dots, A_n)$ 与数组 $B(B_1, B_2, B_3, \dots, B_n)$ 之间的欧氏距离为:

$$D_{(A,B)} = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \dots + (A_n - B_n)^2} \quad (6)$$

针对 EEMD 算法 100 次迭代计算产生的 IMF 分量矩阵进行 K 均值聚类分析,选取 IMF 聚类中成员多的那类作为 IMF 的合集。聚类过程如下:1)对 IMF_1 进行聚类,选取第 1 个和第 50 个矩阵作为起始聚类中心,计算剩余矩阵分别到这 2 个聚类中心的欧氏距离;2)根据距离的大小将各个矩阵中的 IMF_1 聚类到距离最近的一类中,取各类的平均值作为新的聚类中心,再次计算各矩阵到聚类中心的距离;3)不断重复这个过程,直到聚类中心不再发生改变;4)选取 IMF_1 聚类完成后对象多的一类,继续将 IMF 各个其他分量和最后的残余项按照上述过程进行 K 均值聚类,得到每个 IMF 分量和残余项的聚类结果;5)对聚类结果分别取平均值,即为 K 均值 EEMD 分解的各个 IMF 分量和残余项(图 1)。

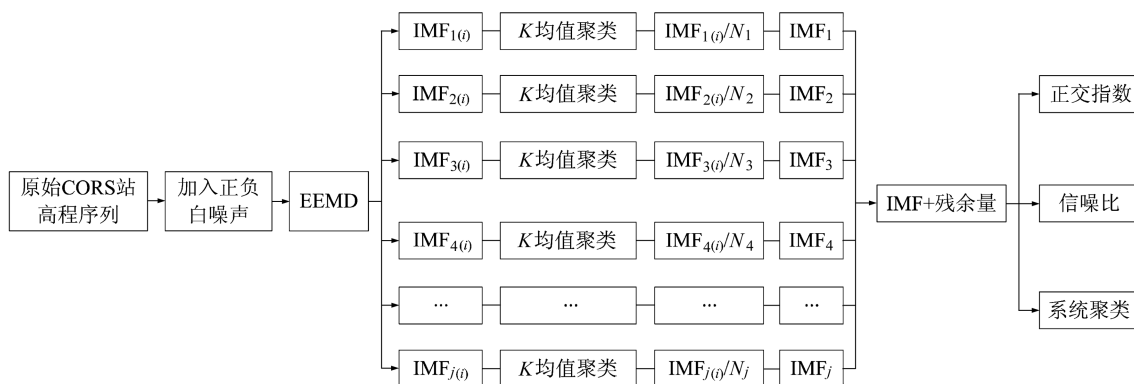


图 1 K 均值聚类 EEMD 流程

Fig. 1 K -means clustering EEMD flowchart

1.3 正交指数与信噪比指标

正交指数 (IO) 指标是衡量模式分解结果精度的标准^[9],正交指数越小分解得到的精度越高。在 CORS 高程时间序列分解的结果中,正交指数可以用来衡量任意 2 个 IMF 分量:

$$IO_{ij} = \sum_t \frac{C_i(t)C_j(t)}{C_i^2(t) + C_j^2(t)} \quad (7)$$

式中, IO_{ij} 为第 i 个和第 j 个 IMF 分量之间所求得的正交指数, C_i 为对应的 IMF 分量。

CORS 高程时间序列信号的信噪比 (SNR) 计算的是 EEMD 加入白噪声分解前后的去噪效果,计算公式为:

$$SNR = 10 \log_{10} \left(\frac{E[y(n)^2]}{E\{[y(n) - \hat{y}(n)]^2\}} \right) \quad (8)$$

式中, $y(n)$ 为没有加入白噪声的原始高程序列, $\hat{y}(n)$ 为加入白噪声 EEMD 分解后得到的所有分量 (IMF+残余项) 重构的信号。

1.4 方差贡献率

方差贡献率是各个 IMF 分量的方差与分解得到的 IMF 分量方差之和的百分比,IMF 分量方差贡献率的大小反映该频率的分解信号在整个序列信号运动能量中的贡献大小^[10],统计任一 IMF 分量的方差公式为:

$$\rho_x^2 = E(x^2) - E^2(x) \quad (9)$$

2 实验分析

2.1 EEMD 和 K 均值 EEMD 分解实验

利用 BJFS 站和国外 BOGO 站近 20 a 的高程时间序列进行实验。实验数据来自 SOPAC 网站,已进行去常数处理,并利用 3 倍中误差法剔除粗差。图 2、图 3 分别为 BJFS 站和 BOGO 站的原始高程时间序列,2 个站分别进行 EEMD 和 K 均值聚类 EEMD(K -EEMD)实验。

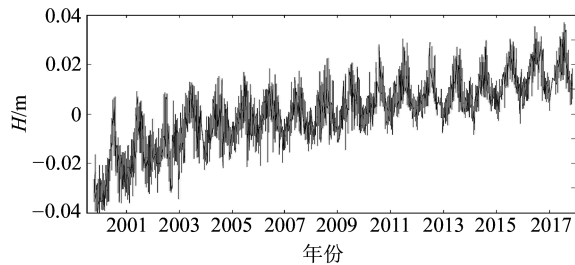


图 2 BJFS 站原始高程序列
Fig. 2 Height time series of BJFS station

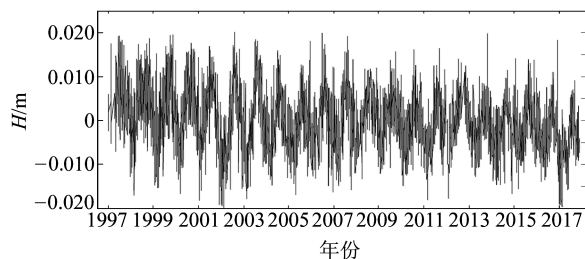


图 3 BOGO 站原始高程序列
Fig. 3 Height time series of BOGO station

根据 Wu 等^[1]的建议,EEMD 迭代次数设置为 100,噪声与信号标准差之比设置为 0.1。李鹏^[5]指出, K 均值聚类分析的迭代终止条件一般是聚类中心点收敛或达到最大的迭代次数。本文实验发现,2 个 CORS 站的聚类迭代次数均为 12 时聚类中心不再发生改变,即完成聚类过程。

图 4、图 5 分别是 BJFS 站和 BOGO 站基于 2 种分解方法得到的 IMF 分量,图中仅给出中低频

的 IMF 分量。各个 IMF 分量的方差贡献率如表 1 (单位%)所示。

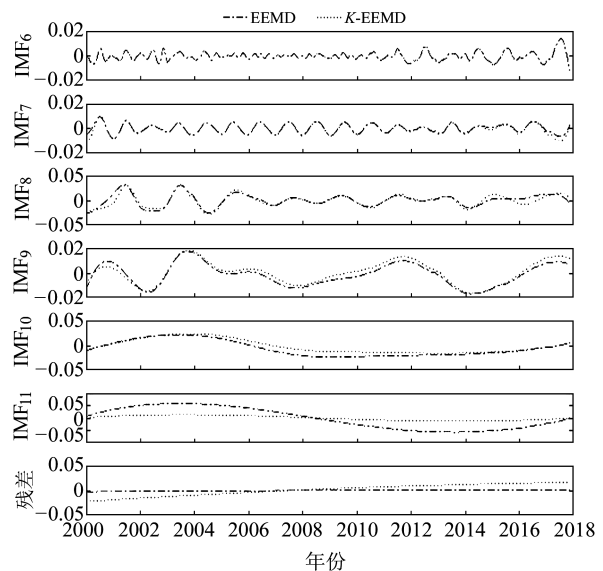


图 4 BJFS 站 IMF₆~IMF₁₁ 分量
Fig. 4 IMF₆-IMF₁₁ components of BJFS station

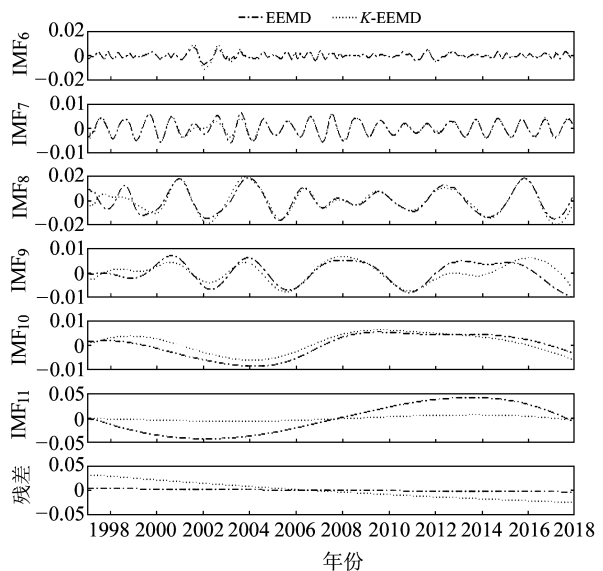


图 5 BOGO 站 IMF₆~IMF₁₁ 分量
Fig. 5 IMF₆-IMF₁₁ components of BOGO station

表 1 2 个测站各 IMF 分量的方差贡献率

Tab. 1 The variance contribution rate of each IMF

站名	IMF ₁	IMF ₂	IMF ₃	IMF ₄	IMF ₅	IMF ₆	IMF ₇	IMF ₈	IMF ₉	IMF ₁₀	IMF ₁₁
BJFS	19.88	7.67	6.82	5.43	3.59	21.19	30.59	3.27	1.67	0.81	0.01
BOGO	18.52	10.90	10.04	8.23	5.52	20.12	22.86	2.85	0.46	0.46	0.01

BJFS 站 IMF₁~IMF₃ 为明显高频分量,主要是信号中的有色噪声;IMF₄~IMF₅ 的方差贡献率分别为 5.43% 和 3.59%,在序列中占比较小;IMF₆ 与 IMF₇ 的方差贡献率分别为 21.19%、30.59%,在序列中占比最大,为序列主要周期项。从图 4 可以看出,IMF₆ 为近似 0.5 a 周期项,IMF₇ 为近似 1 a 周期项,IMF₈ 为近似 2 a 周期项,IMF₉~IMF₁₁ 周期明显变长,趋势项单调性明

显,符合经验模式分解的要求。

BOGO 站分解后与 BJFS 站类似,IMF₆ 与 IMF₇ 的方差贡献率分别为 20.12%、22.86%,分别为近似 0.5 a 和 1 a 的主要周期贡献项;IMF₈ 方差贡献率 2.85%,为近似 2 a 周期项。除 IMF₁~IMF₃ 为高频分量外,其余各分量方差贡献率大致相同。

2.2 精度对比

对 BJFS 站和 BOGO 站分解后产生的 IMF 分量进行正交指数检验。由于 IMF₆ 与 IMF₇ 是主要的周期贡献量,对这 2 个分量进行正交指数计算,结果见表 2。

表 2 2 种分解方法的正交指数与信噪比
Tab. 2 IO_{67} and SNR of EEMD and K-EEMD

	正交指数		信噪比	
	EEMS	K-EEMD	EEMD	K-EEMD
BJFS	0.055 3	0.040 7	5.592	5.919
BOGO	0.069 0	0.049 0	5.080	5.220

可以看出,K-EEMD 算法计算的 BJFS 站和 BOGO 站 IMF₆ 与 IMF₇ 之间的正交指数均小于 EEMD 算法,正交指数分别减小 26% 和 29%。K-EEMD 算法的信噪比明显大于 EEMD 算法,分别提高 6% 和 3%。

2.3 IMF 系统聚类分析

判断 IMF 分量之间是否存在相似性,可利用聚类分析画出系统聚类图,若出现模态混叠现象则存在相似性。

图 6、图 7 分别是 BJFS 站 EEMD 方法与 K-EEMD 方法的 IMF 分量系统聚类图。当欧氏距离为 25 时,EEMD 方法出现模态混叠,IMF₆ 和 IMF₇ 被分成一类,具有相似性;当欧氏距离为 10 时,EEMD 方法的 IMF₄ 与 IMF₅ 同样被分成了一类。K-EEMD 方法的 11 个 IMF 分量在同等欧氏距离条件下没有出现模态混叠现象,但当欧氏距离小于 10 时,高频的 IMF₄ 与 IMF₅ 混叠,可能是聚类分析过程中起始质心点的选择与欧氏距离选择过小引起。

图 8、图 9 分别是 BOGO 站 EEMD 与 K-EEMD 的 IMF 分量系统聚类图。当欧氏距离在 30 以上时,EEMD 方法的 IMF₆ 与 IMF₇ 分量发生模态混叠,而 K-EEMD 方法的各个 IMF 分量在与 EEMD 同等欧氏距离条件下,没有出现模态混叠。

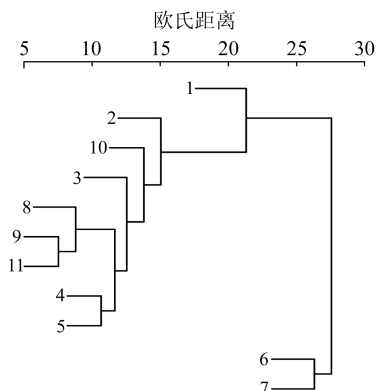


图 6 EEMD 方法 BJFS 站 IMF 聚类图
Fig. 6 IMF clustering by EEMD of BJFS station

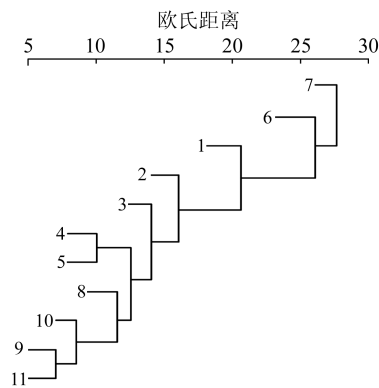


图 7 K-EEMD 方法 BJFS 站 IMF 分量聚类图
Fig. 7 IMF clustering by K-EEMD of BJFS station

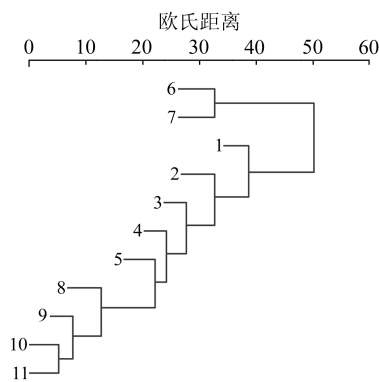


图 8 EEMD 方法 BOGO 站 IMF 聚类图
Fig. 8 IMF clustering by EEMD of BOGO station

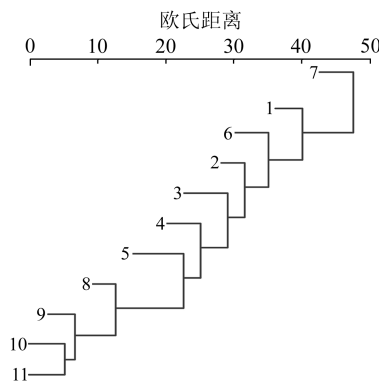


图 9 K-EEMD 方法 BOGO 站 IMF 聚类图
Fig. 9 IMF clustering by K-EEMD of BOGO station

叠,不存在相似性。

2 个 CORS 站的实验结果表明,EEMD 分解后 IMF₆ 与 IMF₇ 之间存在模态混叠现象,而 K 均值聚类 EEMD 在同等条件下不存在模态混叠现象。CORS 高程时间序列的 IMF 分量包含高频噪声项、周期项和低频趋势项。中低频 IMF 分量对高程时间序列信号的周期运动给出清晰的解释,包含明显的季节性、1 a 和 2 a 周期变化、长周期变化。已有研究表明,IMF 周期信号通过 Hilbert 变换后的频率并不是一个常数,而是瞬时频率,说明 IMF 的周期随时间而变化,这个问题需要进一步研究。

3 结 语

利用 2 个 CORS 站近 20 a 的高程时间序列,采用 EEMD 和 K 均值聚类 EEMD 方法进行 CORS 高程时间序列分解,获取频率从高到低的 IMF 分量,识别出近似 0.5 a、1 a、2 a 周期项,有效减少了 IMF 分量中近似 1 a 与 0.5 a 周期信号间存在的模态混叠现象,将信噪比提高 3% 以上,分解精度提高 26% 以上。对于分解结果中部分分量之间存在的模态混叠现象,是否是聚类分析过程中起始聚类点和聚类终止条件的选择问题,仍需进一步研究。

参考文献

- [1] Wu Z H, Huang N E. Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method[J]. *Advances in Adaptive Data Analysis*, 2009, 1(1): 1-41
- [2] 徐健,周志祥,唐亮,等. 基于总体平均经验模态分解算法的自适应改进[J]. *振动与冲击*, 2017, 36(11): 215-223 (Xu Jian, Zhou Zhixiang, Tang Liang, et al. Adaptive Improvement of EEMD Algorithm[J]. *Journal of Vibration and Shock*, 2017, 36(11): 215-223)
- [3] 郑近德,程军圣,杨宇. 改进的 EEMD 算法及其应用研究[J]. *振动与冲击*, 2013, 32(21): 21-26 (Zheng Jinde, Cheng Junsheng, Yang Yu. Modified EEMD Algorithm and Its Applications[J]. *Journal of Vibration and Shock*, 2013, 32(21): 21-26)
- [4] 施闯,牛玉娇,魏娜,等. HHT-EEMD 用于 IGS 站高程时间序列分析[J]. *大地测量与地球动力学*, 2018, 38(7): 661-667 (Shi Chuang, Niu Yujiao, Wei Na, et al. Application of the HHT-EEMD Approach in Analysis of GPS Height Time Series[J]. *Journal of Geodesy and Geodynamics*, 2018, 38(7): 661-667)
- [5] 李鹏. 基于层次 K 均值的聚类算法的研究[D]. 哈尔滨: 哈尔滨工程大学, 2015 (Li Peng. The Study and Development of Hierarchical K -Means Based Clustering Algorithm[D]. Harbin: Harbin Engineering University, 2015)
- [6] Chung K L, Lin J S. Faster and More Robust Point Symmetry-Based K -Means Algorithm[J]. *Pattern Recognition*, 2007, 40(2): 410-422
- [7] Rousseeuw P J. Silhouettes: A graphical Aid to the Interpretation and Validation of Cluster Analysis[J]. *Journal of Computational and Applied Mathematics*, 1987, 20: 53-65
- [8] 唐东明. 聚类分析及其应用研究[D]. 成都: 电子科技大学, 2010 (Tang Dongming. Study on Clustering Algorithm and Its Applications[D]. Chengdu: University of Electronic Science and Technology, 2010)
- [9] Huang N E, Shen Z, Long S R, et al. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis[J]. *The Royal Society*, 1998, 454(1 971): 903-995
- [10] 张恒璟,程鹏飞. 基于经验模式分解的 CORS 站高程时间序列分析[J]. *大地测量与地球动力学*, 2012, 32(3): 129-134 (Zhang Hengjing, Cheng Pengfei. Analysis on Time Series of Two CORS Stations' Height Based on EMD[J]. *Journal of Geodesy and Geodynamics*, 2012, 32(3): 129-134)

Analytical Method of CORS Height Time Series Based on K -Means Clustering EEMD

ZHANG Hengjing^{1,2} QI Xin^{1,2} WEN Hanjiang²

¹ School of Geomatics, Liaoning Technical University, 88 Yulong Road, Fuxin 123000, China

² Chinese Academy of Surveying and Mapping, 28 West-Lianhuachi Road, Beijing 100830, China

Abstract: We advance an analytical method of CORS height time series based on K -means clustering EEMD. Aiming at the problem of low signal-to-noise ratio and partial mode mixing in signal decomposition of traditional EEMD, we add positive and negative white noise EEMD to improve signal decomposition SNR. Based on the K -means clustering method, we cluster the various IMF components decomposed in the EEMD iterative process. The experience results show that the method improves the SNR by more than 3%, and the decomposition accuracy based on index of orthogonality increases by more than 26%. The clustering results can solve the problem of mode mixing of the approximate year, half-year and two years periodic signals in the IMF.

Key words: height time series; mode mixing; K -means clustering; SNR; index of orthogonality

Foundation support: The Major Project of High Resolution Earth Observation System, No. 42-Y20A09-9001-17/18; Fundamental Research Funds for Universities of Education Department of Liaoning Province, No. LJ2017QL008; PhD Start-Up Fund of Liaoning Technical University, No. 13-1120.

About the first author: ZHANG Hengjing, PhD, associate professor, majors in spatial geodetic data processing and nonlinear motion of GNSS height time series, E-mail: sun_winter2009@163.com.

Corresponding author: QI Xin, postgraduate, majors in spatial geodetic data processing, E-mail: lntuqixin@163.com.