数字化水位观测数据缺失填补研究

杨盛梅

西藏自治区水文水资源勘测局昌都水文分局　西藏　昌都　850000

[摘　要]由于水位观测数据的时间尺度连续性遭到破坏，导致对缺失数据填补的准确性较低，为解决这一问题，本文对数字化水位观测数据缺失填补进行研究。以观测到的水位数据序列的局部最大值和最小值作为包络线的上、下限，按照时间尺度对数据进行经验模态分解，通过对最邻近算法的k值进行优化计算，水位分解结果的填补值的最佳时间尺度，并将最小误差k值对应的数据作为缺失填补值。测试结果表明，设计方法的填补数据与实际数据之间的误差始终小于5.00mm。

[关键词]水位观测数据；缺失填补；时间尺度；经验模态分解；最邻近算法；k值优化

【DOI】10.12252/j.issn.2096-6288.2021.12.335

0 引言

受环境以及相关因素的影响，在对其水源的发展情况进行分析和预测的过程中，对于观测数据的依赖性较强。也正因如此，准确、有效的水位数据成为相关工作开展的重要基础。但在实际的观测阶段，受干扰因素的影响，水位观测结果中难以避免地会存在部分缺失，为了降低该部分数据对后续研究的影响，实施合理有效的缺失值修复处理成为水位数据预处理中极为重要的内容之一。传统模式下，对于水位观测数据处理的重视程度相对较低，当观测数据中出现中断、缺失时，大多直接放弃该部分数据序列，或者不顾及缺失数据，直接进行使用。这样的处理方式不仅在一定程度上降低了对水位实际发展情况的观测效果，也在数据分析阶段表现出了更高的局限性。在此基础上，为解决这些问题，本文提出一种数字化水位观测数据缺失填补方法，秉持着避免数据资源浪费的基本原则，在充分挖掘残缺样本中的蕴含信息的基础上，最大限度填补缺数的数据信息，并通过对比测试的方式分析验证了设计方法有效性。

1 数字化水位观测数据缺失填补

1.1数字化水位数据经验模态分解

数字化水位观测数据之间存在一定的关联性，为了实现对缺失数据的准确填补，本文以观测到的数字化水位数据序列的局部最大值和最小值作为包络线的上、下限，利用极值点计算出上包络线和下包络线的均值分别为

$$\max(x)=\frac{\bar{x}+a}{2} \quad (1)$$

$$\min(x)=\frac{\bar{x}-a}{2} \quad (2)$$

其中，$\max(x)$和$\min(x)$分别表示上包络线和下包络线的均值，a表示水位数据极值点的参数值，$\bar{x}$表示整体数据的均值。

本文采用经验模态分解（Empirical Mode Decomposition，经验模态分解）这一方法对观测到的数据序列进行处理。首先，以数字化水位数据的时间尺度为基础，将其自适应分解为多个IMF（Intrinsic Mode Function，固有模态函数）和一个残余量RES，其分解方式可以表示为

$$D(x)=nsim(t\to T)+\lambda R_x$$
$$s.t.x\geq\min(x),x\leq\max(x)$$

其中，$D(x)$表示缺失数据序列，n表示分解固有模态函数数量，t表示在时间尺度上数据的局部特征信息，$sim(*)$表示自适应参数，T表示单位时间尺度值，λ表示数据的关联系数，R_x表示分解后的残余量。

通过这样的方式得到具有不同时间尺度参量的IMF分量，且在n个分量中对应数据出现的频率逐渐降低，序列的平稳性也逐渐提高。此时水位数据极值点及过零点数目是一致的，当数据规模较大时，其数据最大差异值不会超过1。

通过这样的方式，以实现对数字化水位信息的经验模态分解处理。

1.2缺失数据填补

在上述基础上，本文采用最邻近算法实施对缺失数据的修复。考虑到k值选取直接影响到填补数据准确性，因此本文对k值进行寻优处理，上文中已经以时间制度为基础对数据进行分析，因此本文同样以单位时间尺度对此时的k值进行计量。假设k值为m个单位时间尺度，且有$m=1,2,\ldots\ i$，其中，i表示时间尺度参数的最大值。此时k对应的填补数据误差可以表示为

$$\Delta=\frac{k(x_t-x_{t+m})}{m} \quad (3)$$

其中，Δ表示填补数据误差，x_t和x_{t+m}分别表示在t时间尺度和$t+m$时间尺度上的水位数据。

经过时间尺度排序和k值设定后，对时间尺度最近的k个数据进行赋权处理，考虑到待填补数据最终数值是由时间尺度最近的　个参数值决定，因此本文对其的计算方式为

表 1 地下水位数据信息 /m

观测点	数据编号									
	1	2	3	4	5	6	7	8	9	10
D1	-12.44	-12.36	-12.30	-12.26	-12.30	-12.26*	-12.24	-12.20	12.72	-12.29
D2	-13.56	-13.44*	-13.20*	-13.02	-13.10	-13.00	-12.84	-12.56	-12.30	-12.12
D3	-10.66	-10.56	-10.40	-10.50	-10.42*	-10.21	-1.006	-9.85	-9.52	-9.42
D4	-9.96	-9.75	-9.82	-9.72*	-9.66*	-9.53*	-9.36	-9.22	-9.10	-9.00

$$x' = w\frac{\sum_{i=1}^{k} x_i}{k} \quad (4)$$

其中，x''表示填补的数据值，w表示对k个邻近时间尺度范围内数位数据的赋权结果。需要注意的是，当k值的取值结果大于经验模态分解得到的固有模态函数数量时，需要对单位时间尺度大小进行调整，以确保k值的取值结果能够覆盖所有数据。

按照式（5）所示的方式，逐个计算不同k值下的填补数据误差，计算得到的结果中，误差结果会呈现出“U”的发展趋势，本文取波谷处的k值对应的数据结果作为最终补充的参数，实现对缺失数据的填补。

通过这样的方式，实现对水位观测缺失数据的填补。

2 测试分析

在上述基础上，分别采用文献[1]和文献[2]提出的数据填补方法作为对照组，采用本文设计的方法作为测试组，以实际水位数据为基础，开展了测试。

2.1测试数据准备

本文以某区域的实际地下水位信息为测试数据，随机选取了1条地下水，对其水位数据中的4个观测点（D1-D4），以其中连续10个观测数据为基础，构成待修复数据集，其中，完整的数据信息如表1所示。

在此基础上，分别对数据中的部分数据进行敲除处理，考虑到实际观测数据的缺失规模存在一定的差异，因此本文在数据敲除阶段的敲除对象分别为D1-6、D2-2、D2-3、D3-5、D4-4、D4-5、D4-6（表1中“*”标记数据）。分别采用三种方法对敲除数据进行填补，并对比其与实际数值之间的差异。

表 2 缺失数据填补效果对比表 /mm

缺失信息	填补误差		
	文献[1]方法	文献[2]方法	本文方法
D1-6	6.67	5.44	1.24
D2-2	11.55	10.63	3.55
D2-3	12.02	10.24	3.40
D3-5	5.92	5.31	1.16
D4-4	16.82	12.86	4.66
D4-5	17.11	12.91	4.62
D4-6	16.90	13.03	4.53

2.2测试结果

采用三种方法对填补的缺失数据与实际数据之间的差异如表2所示。

由表2可以看出，对比三种方法，其中文献[1]方法对于单一缺失数据的填补结果与实际结果的误差相对较小，但是当连续缺失数据的规模增加时，其误差明显升高，最大误差达到了17.11mm（D4-5），文献[2]方法整体表现出与文献[1]方法相同的趋势特征，但是其误差范围明显较小，最大误差为13.03mm（D4-6），相比之下，本文方法对缺失数据的填补效果明显优于2种对比方法，与实际数据的误差始终稳定在5.00mm以内，其中，最大误差仅为4.66mm（D4-4）。测试结果表明本文提出的数据缺失填补方法能够实现对水位观测数据的准确修复，最大限度还原数据的真实情况。

3 结束语

水位观测数据结果是相关地质研究工作开展的基础，对观测数据中的缺失数据进行有效修复作为数据预处理阶段的重要环节之一，其关系到后续相关研究的可靠性。为此，本文对数字化水位观测数据缺失填补进行研究，以期为实现对缺失数据的准确填补，提高观测数据的利用价值，达到实际地质研究工作良好开展的目标提供一定帮助。

参考文献

[1]鹤壁台静水位观测现状及存在问题和建议[J].张达.华北自然资源.2021（04）

[2]黑龙江干流畅流期水位观测防浪方案的应用[J].王化鑫，常春超，曲志强. 东北水利水电.2017（02）

[3]水位观测的不确定度计算[J].马勇，徐长春，柴国武，柴颖. 河南水利与南水北调.2012（23）

作者简介:

杨盛梅（1993年11月），女，汉族，籍贯（四川省凉山州）。助理工程师，大学本科，研究方向。