文章编号:1672 - 3317(2019)02 - 0084 - 09

点源时间序列数据缺失值的估值不确定性分析
——以小流域气象和水文数据为例

石 锦[1,2],周脚根[2],王 辉[1*],甘 蕾[2,3],沈健林[2],李 希[2],李裕元[2],吴金水[2]
(1.湖南农业大学 工学院,长沙 410128;2.中国科学院 亚热带农业生态研究所 亚热带农业生态过程重点实验室,长沙 410125;3.湖南师范大学 资源与环境科学学院,长沙 410081)

摘 要:【目的】对普遍存在的时间序列缺失值进行有效估值,进而改善时间序列数据的质量。【方法】以亚热带典型小流域长期定位观测的气象(最低气温、最高气温、太阳辐射)及水文(降水量、地表径流量)数据为样本并利用计算机模拟的方法,比较了线性内插法、*K*-最近邻插值法、多项式插值法、样条插值法和核密度估值法5种估值方法的性能差异,分析了不同取样时间步长(日和月)及不同数据缺失量(1%、5%、10%、15%、20%)条件下对缺失值进行估值的不确定性。均方根误差(*RMSE*)、绝对值平均误差(*MAE*)和Pearson相关系数(*r*)3个交叉验证指标用于评估5种估值方法的性能优劣。【结果】①5种方法估值性能较好,Pearson相关系数为0.62~0.99(P<0.05),且以核密度估值法和多项式插值法的估值性能为优;②数据缺失量和取样步长增加降低了5种估值方法的估值精度;③数据集的变异系数(*CV*)与估值评估指标(*RMSE*、*MAE*及*r*)显著相关(P<0.05)。【结论】核密度估值法和多项式插值法的估值结果相对更可靠,变异系数是影响估值不确定性的重要因素。

关 键 词:缺失值;估值方法;变异系数;不确定性;时间序列

中图分类号:O171 **文献标志码:**A doi:10.13522/j.cnki.ggps.2017.0421

石锦,周脚根,王辉,等. 点源时间序列数据缺失值的估值不确定性分析——以小流域气象和水文数据为例[J]. 灌溉排水学报,2019,38(2):84-92.

0 引 言

时间序列数据是生态环境、水文及气象等研究领域必不可少的基础数据。环境参数需要进行长期定位监测,但是由于仪器设备故障、恶劣环境或者人为操作失误等不同原因,采集的时间序列数据难免会出现数据缺失的问题[1],从而影响观测数据质量。有效估算时间序列数据的缺失值或者未观测值,可以完善时间序列数据的质量,提升数据使用效率,是空间分析与统计领域的研究热点之一[2]。时间序列的估值问题,目前主要涉及3个方面研究内容:①面源尺度上对未观测位点环境参数属性值的估算;②点源尺度上对观测参数缺失值的估算;③观测参数估值的不确定性。

由于人力和物力的有限性,面源尺度上环境参数通常只是部署一定量具有代表性的点源观测单元,再通过这些点源观测数据实现观测数据的面源拓展。简单而言,这是一个用一定量点源观测数据估算面源上未观测单元参数值的过程。空间插值方法或统计插值方法常用于解决该问题。例如,国外有学者利用协同克里格法[3]有效地实现了降雨的空间插值且估值效果优于普通克里格法;Murphy等[4]利用反距加权和克里格法,有效完成了水温及含盐量的水质监测数据空间插值,整体上克里格方法的表现通常优于反距离加权。Chen等[5]发现普通克里格法对我国东部地区降雨量的估值效果较优;有研究发现克里格估值对河套地区地下水的空间变异性研究效果较优[6-7]。

收稿日期:2017-07-11
基金项目:水利部公益性行业科研专项经费项目(201501055);国家科技支撑计划项目(2014BAD14B02);河南省重大科技专项项目(161100310600)
作者简介:石锦(1993-),男。硕士研究生,主要从事水文生态与环境方面的研究。E-mail: 570187642@qq.com
通信作者:王辉(1973-),男。教授,主要从事水文环境生态方面的研究。 E-mail: wanghuisb@126.com

点源尺度上时间序列缺失值的估值，主要是对一定观测时间段内缺失的观测数据进行有效的补充插值。一些研究直接将缺失数据的样本剔除，也有研究采用均值替换所有缺失值的单一插补法[8]。缺失值删除和单一插补法操作简单，但会导致潜在信息丢失，局限性大。鉴于点源时间序列实为二维数据集，实际研究中通常用线性内插法[9]、*K*-最近邻插值法[10]以及多项式插值法[11]等二维曲线拟合数学方法插补缺失数据。戴新刚等[12]、姜晓剑等[13]采用最近邻法、反距离加权法及样条法等对全国范围内气温缺失值进行估算。

分析观测参数估算结果不确定性来源，有利于提高估值可靠性和精度。观测参数估值的不确定性，通常源于样本数据量变化、样本或环境协变量数据的空间或时间尺度变化以及所用估值模型的参数变化。目前，有关观测参数估值不确定性研究主要集中在面源尺度上。例如评估不同取样尺度[14]、取样量[15]及估值方法[16-17]下对土壤属性估值不确定性，比较不同估值方法下对降雨[18-19]以及对高程数据插值的不确定性[20-21]。

总体上，当前国内外对点源时间序列数据缺失值的估值问题研究，通常集中于某一估值方法对特定类型的数据集缺失值的估值分析，缺乏不同估值方法对缺失值估值结果的性能差异比较，也少有分析影响点源时序数据缺失值的估值不确定性。为此，选用线性内插法、*K*-最近邻插值法、多项式插值法、样条插值法和核密度估值法等5种估值方法，以湖南金井小流域气象站点数据（最低气温、最气高温、太阳辐射）及水文数据（降雨量、地表径流流量）为例，研究不同取样时间步长（日和月）及不同数据缺失量（1%、5%、10%、15%、20%）条件下5种估值方法对缺失值进行估值的不确定性。

1 材料与方法

1.1 数据来源

本研究数据来源于中国科学院亚热带农业生态研究所长沙农业环境观测研究站，试验站地处湖南省长沙县金井镇，所在水系属于湘江一级支流捞刀河的上游，即金井河小流域。该小流域地理坐标为27°55'—28°40' N、112°56'—113°30'E，属亚热带湿润性季风气候，为典型亚热带红壤丘陵地貌，年平均降水量为1 200～1 500 mm[22-23]。

文中采用的气象数据为2010—2012年的日最高气温、日最低气温数据以及日降雨量、日太阳辐射量数据，水文数据为2010—2012年金井河小流域出水口的日地表径流量数据。各气象因子数据由小型气象站（Intelimet Advantage，Dynamax Inc.，美国）观测获得。径流流量数据采用Simpson’s Parabolic Rule方法，用螺旋杯式流速仪实测而得。该系统每10 min自动采集并记录流量数据，据此计算流域研究时段内的日径流量。

1.2 估值方法

1.2.1 线性内插法

线性内插法（LIM）利用时间与观测值之间的等比关系近似求解时间序列的缺失值。给定时间序列集t，已知t_i、t_k时刻对应的观测值分别为$Y(t_i)$、$Y(t_k)$，t_j时刻数据样点值$Y(t_j)$缺失，其中$i<j<k$，$Y(t_j)$计算式为：

$$Y\left(t_j\right)=Y\left(t_i\right)+\left[Y\left(t_k\right)-Y\left(t_i\right)\right]/\left(t_k-t_i\right)\times\left(t_j-t_i\right)。\tag{1}$$

由式（1）可知，若数据缺失位点处于时序序列的端点，即$j=i$或$j=k$，LIM方法将无法运行。

1.2.2 *K*-最近邻插值法

K-最近邻插值法（KNNM）的核心思想是：搜索与待估算点最邻近的k个观测点样本，用这些样本点观测值的加权和赋予待估值点。样点之间的邻近关系计算式为：

$$D_i=\frac{1}{\left|t_j-t_i\right|}。\tag{2}$$

给定缺失值$Y(t_j)$和与t_j临近的k个临近点集，$Y(t_j)$计算式为：

$$Y\left(t_j\right)=\sum_{i=1}^{k}\frac{D_iY\left(t_i\right)}{\sum D_i}。\tag{3}$$

1.2.3 核密度估值法

核密度估值法（KDEM）[24]是一种从数据样本本身出发研究数据分布特征的密度函数近似估值算法，不需要有关数据分布的先验知识。对给定缺失值$Y(t_j)$，$Y(t_j)$计算式为：

$$Y\left(t_j\right)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t_i-t_j}{h}\right)，\tag{4}$$

式中：$K(t)$为核函数；h为核函数的带宽；n为参与估值的观测值数目。本研究中，核函数$K(t)$采用高斯核函数；该核函数是1个权函数，离缺失点t_j越近的点对函数值的影响越大，其权值也越大。核函数带宽h统一为缺失点t_j到其他观测点的距离集的中段值，参数n统一为15。

1.2.4 多项式插值法和样条插值法

多项式插值法（PIM）是用多项式对一列数据进行线性拟合再对给定待估值点进行估值的过程。给定时间序列数据集$Y=\{Y(t_1),Y(t_2),\cdots,Y(t_n)\}$和待估值点$Y(t_j)$，首先用多项式函数$f(t)=a_0+a_1t+a_2t^2+\ldots+a_nt^n$对时间序列数据集$Y$进行线性拟合，以求解最优的参数$\beta=(\beta_0,\beta_1,\beta_2,\cdots,\beta_n)$。本研究用最小二乘法求解最优参数$\beta$。

样条插值法（SIM）是一种特殊的分段3次多项式插值法。相对普通多项式插值，通常样条插值方法对数据集的拟合更平滑，输出的插值误差更小。给定n+1个不同的观测时刻t_i，并满足$t_0<t_1<\cdots<t_{n-1}<t_n$以及$n$+1个观测值$Y(t_i)$，样条插值实质上就是构建一个$n$阶样条函数（式（5））逼近观测数据集。

$$Y(t)=\begin{cases}Y_0(t) & t\in(t_0,\ t_1)\\ Y_1(t) & t\in(t_1,\ t_2)\\ \cdots & \\ Y_{n-1}(t) & t\in(t_{n-1},\ t_n)\end{cases}。\tag{5}$$

1.3 缺失值设置及模型校验

时间序列数据集按取样时间步长分为日和月，即：日最高气温、日最低气温、日太阳辐射量、日降雨量、日地表径流量、月最高气温、月最低气温、月太阳辐射量、月降雨量以及月地表径流量。这10个实例数据集的数据点分布见图1所示。

通常时间序列数据集中数据缺失位点以及数据缺失量是随机和不确定的。为有效地评估LIM、KNNM、SIM、PIM及KDEM这5种方法对缺失值的估值不确定性，对上述每个实例数据集设置了5个样本抽取处理（1%、5%、10%、15%、20%）。样本抽取时，采用随机模式进行；样本抽取后每个实例数据的剩余样本则作为训练样本数据集，用于筛选估值方法的最佳模型参数。参数优化后的所有估值方法用于对1%、5%、10%、15%、20%缺失样本进行估值。

均方根误差（*RMSE*）、平均绝对误差（*MAE*）和Pearson相关系数（r）3个评估指标用于评估估值结果的优劣。*RMSE*和*MAE*反映了模型的估值与真实观测的接近程度；*RMSE*和*MAE*的值越小，模型的估值越接近真实值。Pearson相关系数反映了模型的估值与真实值的线性相关性。*RMSE*、*MAE*以及r计算式分别为：

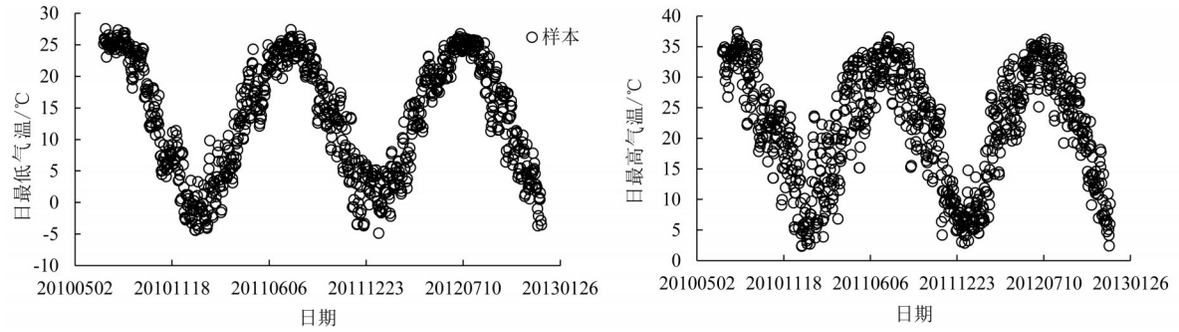
$$RMSE=\left\{\frac{1}{n}\sum_{i=1}^{n}\left[Y(t_i)-\hat{Y}(t_i)\right]^2\right\}^{\frac{1}{2}},\tag{6}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|Y(t_i)-\hat{Y}(t_i)\right|,\tag{7}$$

$$r=\frac{\sum Y(t_i)\hat{Y}(t_i)-\dfrac{\sum Y(t_i)\hat{Y}(t_i)}{n}}{\sqrt{\left(\sum\left(Y(t_i)\right)^2-\dfrac{\left(\sum\hat{Y}(t_i)\right)^2}{n}\right)\left(\sum\left(\hat{Y}(t_i)\right)^2-\dfrac{\left(\sum\hat{Y}(t_i)\right)^2}{n}\right)}},\tag{8}$$

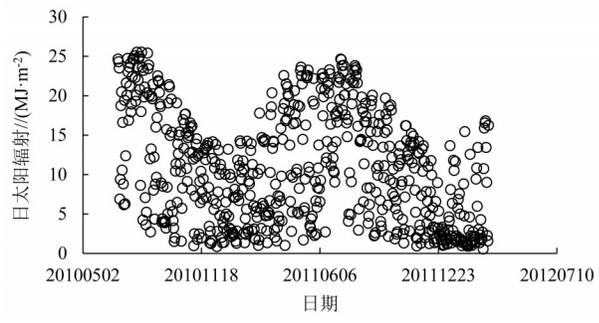
式中：$Y(t_i)$为真实值；$\hat{Y}(t_i)$为模型的估值；n为参与估值计算的观测样本数。

为减少单次抽样估值结果的随机误差，采用100次随机抽样估值输出的评估指标（*RMSE*、*MAE*和r）的均值，评估上述估值方法的性能差异。涉及的LIM、KNNM、SIM、PIM及KDEM的代码实现及模型运行均在Matlab2011b软件平台完成。其中，LIM、KNNM、SIM及PIM 4个插值方法通过调用Matlab2011b软件的内置包运行，KDEM为自主编码实现。以月为取样时间步长（地表径流量、最低气温、最高气温、太阳辐射及降雨量）的样本量偏小（<30个）。按照1%、5%、10%、15%的抽离比例，抽取的样本个数不足4个，故本研究舍弃了上述各数据集的1%、5%、10%及15%的抽样处理。

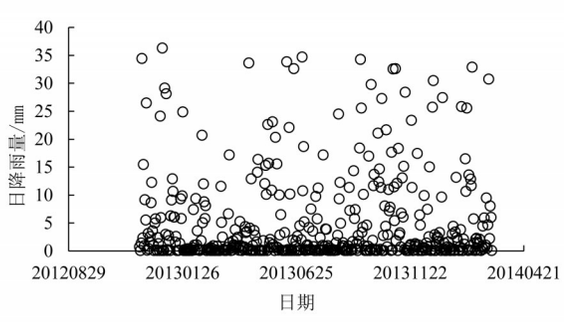


(a)日最低气温

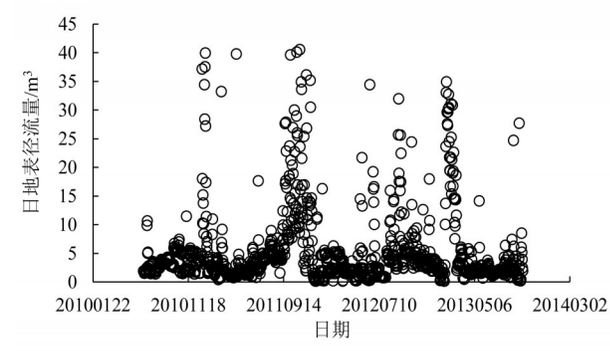
(b)日最高气温



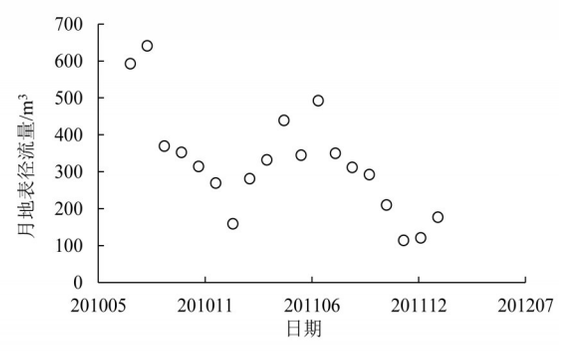
(c)日太阳辐射



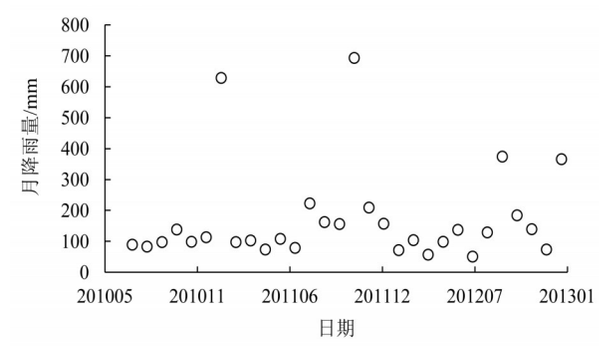
(d)日降雨量



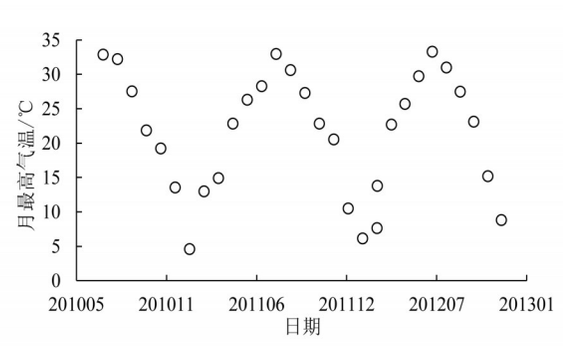
(e)日地表径流量



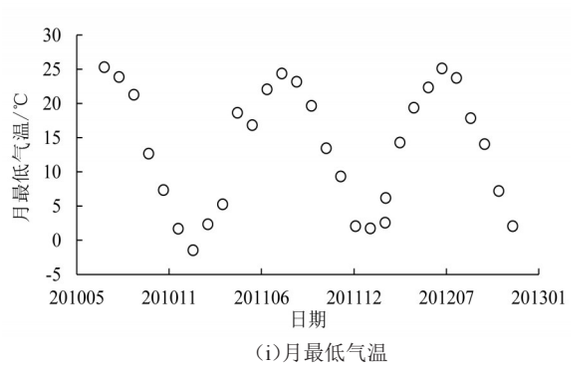
(f)月地表径流量



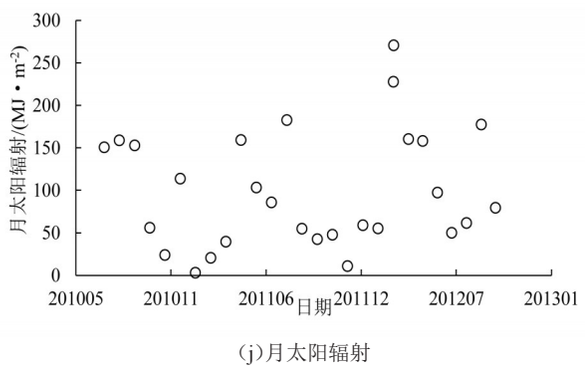
(g)月降雨量



(h)月最高气温



(i)月最低气温



(j)月太阳辐射

图1　10组点源时间序列数据集

2 结果与分析

2.1 金井小流域气象和水文水质数据集的统计特征

研究区日最高气温、日最低气温、日太阳辐射、日降雨量、日地表径流量、月最高气温、月最低气温、月太阳辐射、月降雨量、月地表径流量、小流域水文和气象数据的统计特征见表1。日最高气温和日最低气温变程分别为2.31~37.39、-0.49~27.47 ℃，平均值分别为21.98、13.36 ℃。日太阳辐射值变程为0.49~25.46 MJ/m^2，平均值为10.60 MJ/m^2；日降雨量和日地表径流量数据的变程分别为0.01~36.24 mm、0.07~41.93 m^3，平均值分别为4.70 mm、5.07 m^3；月最高气温和月最低气温变程分别为4.55~33.22、-1.52~25.25 ℃，平均值分别为21.42、13.75 ℃；月辐射值变程为112.94~591.39 MJ/m^2，平均值为319.82 MJ/m^2；月地表径流量和月降雨量数据的变程分别为49.56~692.44 m^3、2.58~270.27 mm，平均值分别为173.61 m^3、100.10 mm。上述各观测数据的变异系数差异较大。日最高气温、月最高气温数据与月太阳辐射数据的*CV*值较小（小于50%），属于弱变异水平。日最低气温、月最低气温、日太阳辐射及月降雨量数据的*CV*值居中分别为67.93%、64.84%、67.14%、69.75%，属于中度变异水平。日降雨量和日地表径流量、月地表径流量数据*CV*值最大，分别为130.7%、162.6%和92.81%，属于强变异水平。

表1 气象和水文数据集的统计特征

指标	变程	均值	变异系数/%
日最低气温/℃	-4.9～27.47	13.36	67.93
日最高气温/℃	2.31～37.39	21.86	42.91
日地表径流量/m³	0.07～41.93	5.07	129.96
日降雨量/mm	0.01～36.24	4.70	166.06
日太阳辐射量/(MJ·m^{-2})	0.49～25.46	10.60	67.14
月最低气温/℃	-1.52～25.25	13.75	64.84
月最高气温/℃	4.55~33.22	21.42	40.90
月地表径流量/m³	49.56～692.44	173.61	92.81
月降雨量/mm	2.58～270.27	100.10	69.75
月太阳辐射量/(MJ·m^{-2})	112.94～591.39	319.82	41.23

表2 不同抽离比例下日数据的交叉校验结果（100次随机重复）

数据集	估值方法	1%			5%			10%			15%			20%		
		RMSE	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*
日流量/m³	LIM	7.66	3.09	0.83	11.41	3.34	0.76	11.99	3.32	0.72	12.40	3.39	0.71	12.94	3.59	0.71
	KNNM	10.74	4.38	0.80	12.69	3.80	0.67	14.58	4.07	0.62	14.14	3.95	0.63	15.27	4.07	0.62
	SIM	7.56	3.12	0.82	1.92	1.10	0.97	3.43	1.48	0.89	3.76	1.70	0.85	5.09	2.37	0.71
	PIM	8.01	3.07	0.79	1.58	0.87	0.98	2.70	1.21	0.92	3.79	1.53	0.84	4.52	1.91	0.76
	KDEM	8.23	3.14	0.77	1.93	1.04	0.97	2.82	1.49	0.90	3.99	1.65	0.82	4.55	2.03	0.75
日降雨量/mm	LIM	14.18	9.62	0.26	15.71	9.14	0.14	17.16	9.65	0.14	17.17	9.26	0.14	16.56	9.19	0.13
	KNNM	14.92	9.62	0.22	18.63	10.01	0.13	19.20	10.25	0.09	19.68	10.04	0.08	19.48	10.28	0.08
	SIM	17.29	13.31	0.24	15.91	10.60	0.09	13.23	8.71	0.01	9.96	6.98	0.32	11.25	7.86	0.20
	PIM	11.29	9.16	0.32	13.47	8.34	0.12	10.00	6.65	0.04	7.97	5.47	0.30	8.77	5.85	0.22
	KDEM	5.15	4.93	0.61	12.35	8.02	0.15	8.89	6.03	0.08	7.77	3.35	0.26	8.37	5.61	0.25
日最低气温/℃	LIM	1.45	1.11	0.99	1.53	1.14	0.99	1.65	1.23	0.98	1.68	1.24	0.98	1.71	1.26	0.98
	KNNM	2.11	1.57	0.97	2.24	1.58	0.97	2.24	1.58	0.97	2.25	1.59	0.97	2.29	1.61	0.97
	SIM	1.38	1.04	0.99	1.57	1.24	0.99	1.44	1.16	0.99	1.69	1.25	0.98	2.02	1.47	0.97
	PIM	0.77	0.65	1.00	1.53	1.16	0.99	1.37	1.05	0.99	1.52	1.08	0.99	1.88	1.33	0.98
	KDEM	0.65	0.58	1.00	1.74	1.19	0.98	1.41	1.07	0.99	1.59	1.12	0.98	1.93	1.39	0.98
日最高气温/℃	LIM	2.42	1.81	0.96	2.50	1.85	0.96	2.57	1.88	0.96	2.61	1.92	0.96	2.71	1.99	0.96
	KNNM	3.16	2.45	0.94	3.35	2.48	0.94	3.39	2.52	0.94	3.37	2.50	0.94	3.44	2.56	0.94
	SIM	1.49	1.26	0.99	3.07	2.33	0.95	2.54	1.88	0.97	3.17	2.25	0.95	2.92	2.22	0.95
	PIM	1.69	1.23	0.98	2.90	2.16	0.95	2.28	1.96	0.97	2.79	2.10	0.96	2.82	2.15	0.95
	KDEM	1.69	1.16	0.98	2.84	2.14	0.95	2.28	1.71	0.97	2.97	2.19	0.95	3.02	2.25	0.95
日辐射量/(MJ·m^{-2})	LIM	4.04	3.24	0.76	4.36	3.28	0.80	4.30	3.21	0.81	4.61	3.43	0.78	4.49	3.36	0.79
	KNNM	4.67	3.70	0.71	5.45	4.00	0.71	5.51	3.99	0.70	5.54	4.04	0.71	5.52	4.00	0.71
	SIM	4.04	3.18	0.76	5.47	3.99	0.62	4.86	3.58	0.79	6.51	4.51	0.65	5.61	4.22	0.70
	PIM	4.53	3.51	0.79	4.34	3.12	0.72	3.81	2.77	0.85	4.75	3.30	0.77	4.71	3.69	0.76
	KDEM	4.16	3.37	0.79	3.61	2.94	0.80	4.09	2.90	0.82	4.29	3.18	0.81	4.68	3.76	0.76

2.2 数据缺失量对估值精度的影响

气象和水文数据缺失值的估值结果表明，5种估值方法的估值性能具有较大差异（结果见表2、表3）。对于日最高气温、月最高气温、日最低气温、月最低气温及日太阳辐射和月太阳辐射数据，LIM、PIM、KDEM、KNNM及SIM 5种方法皆表现较佳，输出的预测值和实测值的相关性显著（P<0.05）；上述数据使用KDEM及PIM的*RMSE*为0.65~97.92，而LIM、KNNM及SIM的*RMSE*为1.38~151.95。综合考虑*RMSE*、*MAE*和*r*发现，KDEM和PIM方法的估值准确性最佳，LIM方法的性能居中，KNNM和SIM这3种方法表现最差。对日降雨量、月降雨量和日径流通量与月径流通量数据，LIM、PIM、KDEM、KNNM及SIM这5种方法表现都不佳，输出的*RMSE*和*MAE*偏大，且预测值与实测值的相关性不显著（P>0.05），上述数据使用KDEM及PIM的

*RMSE*为1.58~64.69，而LIM、KNNM及SIM的*RMSE*为1.92~283.13。

表3 15%和20%抽离比例下月数据的交叉校验结果（100次随机重复）

数据集	估值方法	15%			20%		
		RMSE	*MAE*	*r*	*RMSE*	*MAE*	*r*
月最低气温/℃	LIM	2.50	2.05	0.95	2.61	2.06	0.94
	KNNM	5.07	4.45	0.77	5.25	4.57	0.79
	SIM	1.41	1.31	0.99	2.40	2.15	0.97
	PIM	1.35	1.26	0.99	1.66	1.40	0.99
	KDEM	1.96	1.77	0.97	1.57	1.38	0.99
月最高气温/℃	LIM	2.85	2.38	0.96	2.76	2.27	0.96
	KNNM	5.27	4.78	0.79	5.33	4.80	0.81
	SIM	3.59	2.70	0.98	1.62	1.16	0.99
	PIM	3.45	3.21	0.99	2.24	1.95	0.99
	KDEM	3.25	2.77	0.97	2.72	2.34	0.99
月辐射量/（MJ·m^{-2}）	LIM				77.01	60.93	0.86
	KNNM				107.83	90.36	0.76
	SIM				151.95	110.57	0.71
	PIM				97.92	74.29	0.88
	KDEM				81.75	65.14	0.88
月流量/m^3	LIM	233.89	166.63	0.45	233.12	160.97	0.50
	KNNM	283.13	190.98	0.42	280.29	182.40	0.35
	SIM	167.96	123.10	0.44	185.51	143.44	0.52
	PIM	107.48	67.82	0.49	142.88	117.59	0.53
	KDEM	89.10	59.86	0.52	128.85	103.53	0.52
月降雨量/mm	LIM				86.81	70.57	0.51
	KNNM				105.48	83.26	0.45
	SIM				75.49	65.57	0.50
	PIM				66.63	58.04	0.53
	KDEM				64.69	54.97	0.54

在综合分析了10组数据的均方根误差（*RMSE*）、平均绝对误差（*MAE*）和Pearson相关系数（*r*）后，可以明显看出由于数据缺失量的不同，各方法的性能差异明显。总体上，随着数据量的增加，均方根误差（*RMSE*）和平均绝对误差（*MAE*）增加而Pearson相关系数（*r*）降低，说明随着数据缺失量的增大，5种估值方法的估值精度降低。对于同一组数据，KDEM优于PIM，LIM、SIM和KNND较差，单从Pearson相关系数（*r*）来看，日太阳辐射、日地表径流量、日最低气温、日最高气温、月最低气温、月最高气温和月太阳辐射数据，输出的预测值和实测值的相关性显著（$P<0.05$），而日降雨量、月降雨量和月地表径流量数据，其预测值与实测值的相关性不显著（$P>0.05$）；然而在预测值与实测值的相关性不显著的数据集中（例如日降雨量和月降雨量数据），KDEM相对于其他方法体现出其优异的性能。

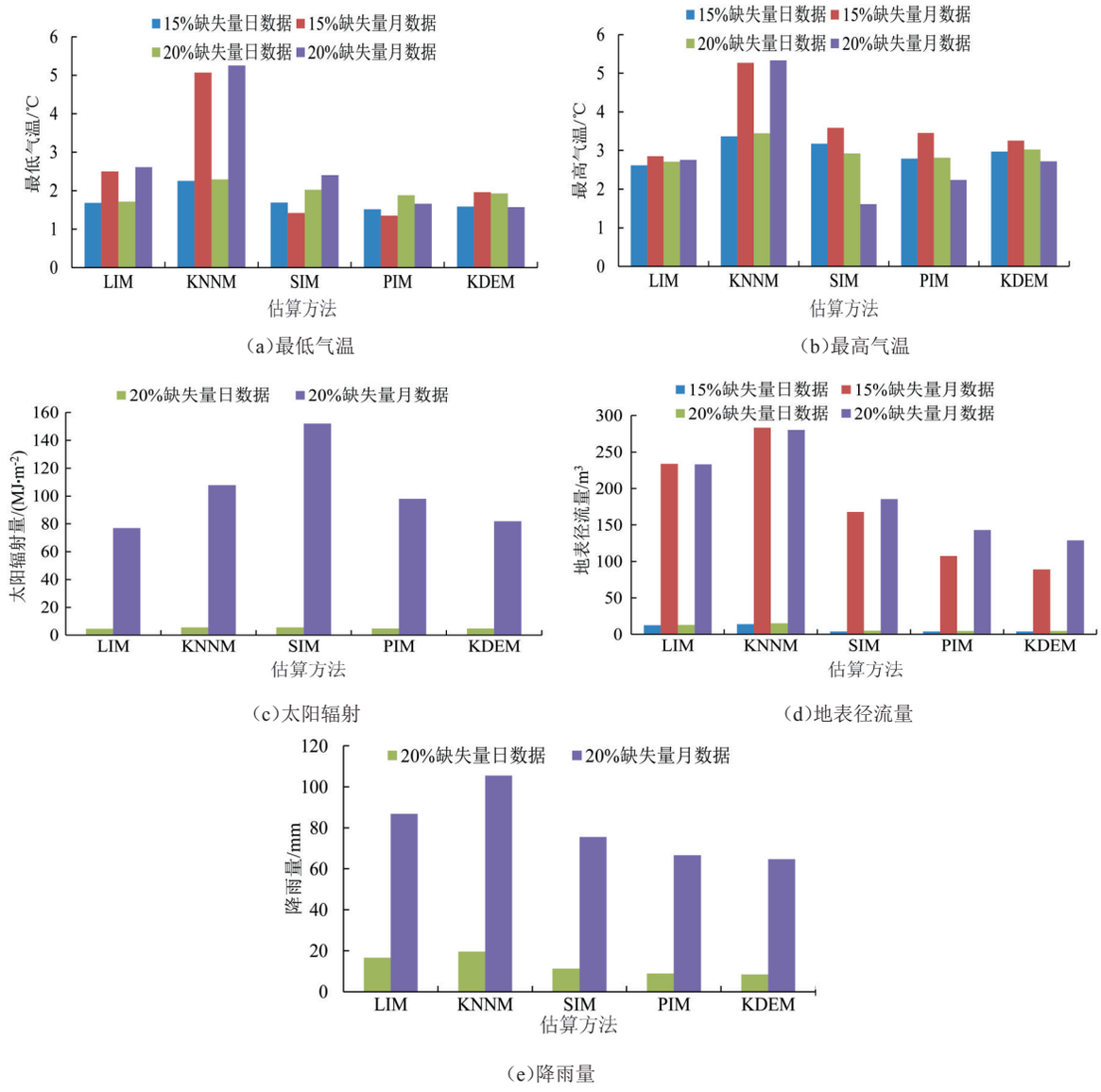
综上所述，随着数据缺失量的增加而导致估值精度的降低，且KDEM及PIM估值性能较优于LIM、KNNM与SIM。

2.3 时间尺度对估值精度的影响

本研究分析了最低气温、最高气温、太阳辐射量、地表径流量以及降雨量5类观测样本的取样时间步长变化（日和月）对估值方法性能的影响。对上述5类观测样本，相同抽样水平下取样时间步长由日变月时，各估值方法的性能下降明显（结果见图2）。

对于地表径流量，日步长数据的*RMSE*为3.45~15.27，而月步长数据的*RMSE*为89.09~283.13；其中，KDEM与PIM的*RMSE*（3.45~142.88）低于其他3类方法的（3.76~283.13）。对于最低气温，日步长数据的*RMSE*为1.52~2.45，而月步长数据的*RMSE*为1.57~5.25；其中KDEM与PIM的*RMSE*为1.35~1.96，低于其他3类数据的*RMSE*（1.41~5.25）。对于最高气温，日步长数据的*RMSE*为2.61~3.37，而月步长数据的*RMSE*分别为2.85~5.27；其中，KDEM与PIM的*RMSE*为2.61~3.45，低于其他数据的*RMSE*（2.24~5.27）。对于降雨量，日步长数据的*RMSE*为8.37~19.48，而月步长数据的*RMSE*为64.69~105.48，其中KDEM与PIM的*RMSE*为8.37~66.63，低于其他数据的*RMSE*（11.25~105.48）。对于太阳辐射量，日步长数据的*RMSE*为4.49~5.61，月步长数据的*RMSE*为77.00~151.95，其中KDEM与PIM的*RMSE*为4.49~97.92，低于其他数据的*RMSE*（5.61~151.95）。

整体上，日步长数据的估值精度高于月步长数据；相对LIM、KNNM及SIM方法，KDEM和PIM方法的估值性能对取样步长变化的敏感性较弱。

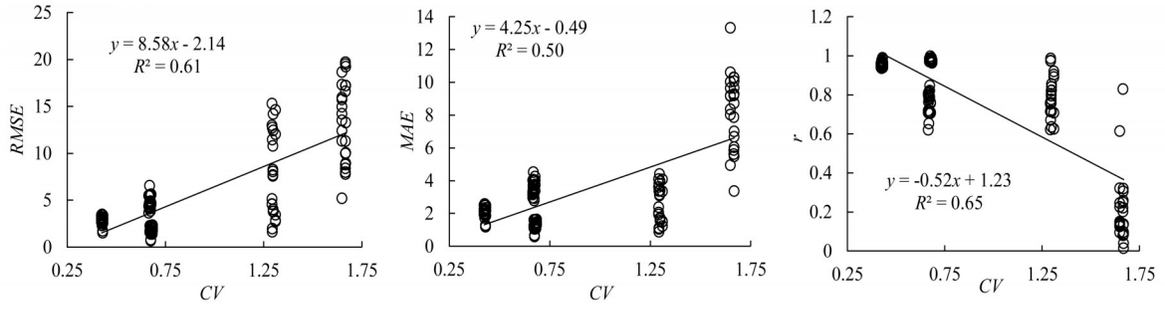


（a）最低气温　（b）最高气温

（c）太阳辐射　（d）地表径流量

（e）降雨量

图2 不同时间尺度下估值方法的*RMSE*



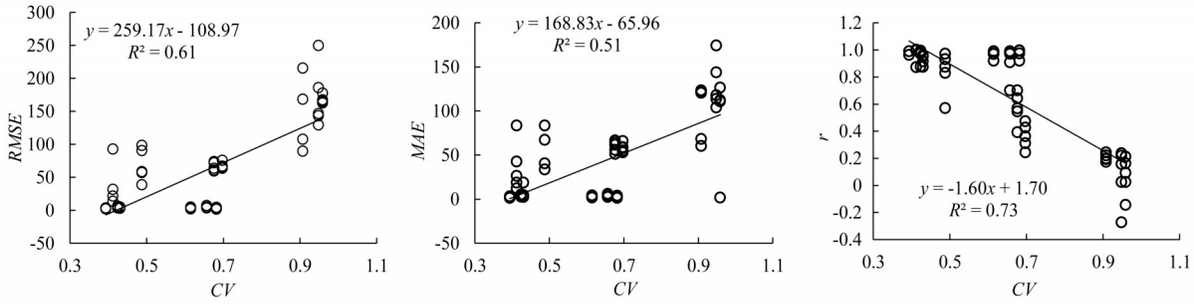
（a）日数据集样本与*RMSE*的相关性图　（b）日数据集样本与*MAE*的相关性图　（c）日数据集样本与*r*的相关性图

图3 各数据集变异系数与性能指标的相关性（日数据）

2.4 变异系数对估值结果的影响

为验证气象和水文数据的离散性对估值结果的影响，本研究对比了10组数据（日最高气温、日最低气温、日辐射量、日降雨量、日径流通量、月最高气温、月最低气温、月辐射量、月降雨量、月径流通量）的变异系数与5种估值方法插补后的均方根误差（*RMSE*）、平均绝对值误差（*MAE*）、Pearson相关系数（*r*）3个交叉验证性能指标（图3、图4）之间的相关关系。由图3、图4可知，变异系数（*CV*）与Pearson相关系数（*r*）相关性程度最高（R^2=0.65~0.73），呈现负相关关系，变异系数值（*CV*）与均方根误差（*RMSE*）存在一定的相关性（R^2=0.61），而变异系数（*CV*）与平均绝对误差（*MAE*）相关性较弱（R^2=0.50~0.51），均呈正相关关系；*RMSE*和*MAE*

均随着*CV*增大而增大，但*r*随着*CV*增大而减小，说明变异系数越大，数据集越不稳定。与2.3综合分析得出，5种估值方法的估值精度随数据变异性的增加而降低。



（a）月数据集样本与*RMSE*的相关性　（b）月数据集样本与*MAE*的相关性　（c）月数据集样本与*r*的相关性

图4 各数据集变异系数与性能指标的相关性（月数据）

3 结 论

兹以点源时间序列数据（最低气温、最高气温、太阳辐射量、降雨量及地表径流量数据）为例，比较了线性内插法、*K*-最近邻插值法、多项式插值法、样条插值法和核密度估值法5种估值方法的性能差异及其主要影响因素。总体上，5种估值方法随着数据缺失量的增加而估值精度下降。其中，核密度估值法和多项式插值法的估值性能优于线性内插法、*K*-最近邻插值法和样条插值法，预测误差小且预测值与实测值的相关性显著。

由日尺度变月尺度时，5种估值方法的精度呈现下降趋势，但核密度估值法和多项式插值法的估值性能还是优于其他3种插值方法。特别以流量数据为例，在缺失率为15%及20%时，核密度估值法体现出其稳定性。一些文献研究也证实了核密度估值法对点源时间序列数据缺失值的估值性能较优。例如，鲁帆等[25]在估算丹江口水库不同分期降雨径流的概率计算与丰枯风险分析时，表明了核密度估值法性能的优越性。

不同时间尺度下估值方法的性能差异表明：对日步长数据的估值精度高于月步长数据，但估值精度的差异程度与数据类型有关。总体上，当时间步长由日变月时，地表径流量、降雨量和太阳辐射量3类数据的*RMSE*增幅程度显著高于气象数据（最低和最高气温）。这与地表径流量、降雨和太阳辐射的月步长数据*CV*（41.23%～92.81%）较大而月气温数据的*CV*较小（40.9%～64.84%）有关。

当数据集离散度相对小时，5种估值方法对数据集缺失值的估值性能较优；当数据集离散度较大时，5种估值方法的估值性能皆显著下降。数据集的变异系数*CV*与评估指标*RMSE*、*MAE*保持显著的线性正相关（$P<0.05$），而与Pearson相关系数*r*显著负相关（$P<0.05$），这充分证实了数据集的离散度是影响估值不确定性的重要因素。该结论也与文献研究结果相吻合，例如，赵彦锋等[14]发现有机质数据变异系数小于10%时对数据集估值结果的准确性最高；Yozgatligil等[26]也证实土耳其降水、温度数据集*CV*值越小，对缺失值估值结果越可靠。

参考文献：

[1] MEHMED K. Data mining: concepts, models, methods, and algorithms[M]. American: Hoboken, 2011.

[2] 关宏强, 蔡福, 王阳，等. 短时间序列气温要素空间插值方法精度的比较研究[J]. 气象与环境学报, 2007, 23(5):13-16 .

[3] HENGL T, HEUVELINK G B M, ROSSITER D G. About regression-kriging: From equations to case studies[J]. Computers and Geosciences, 2007, 33 (10):1 301-1 315 .

[4] MURPHY R R, CURRIERO F C, BALL W P. Comparison of spatial interpolation methods for water quality evaluation in the Chesapeake Bay[J]. Journal of Environmental Engineering, 2010, 136 (2): 160-171 .

[5] CHEN D, OU T, GONG L, et al. Spatial interpolation of daily precipitation in China: 1951—2005[J]. Advances in Atmospheric Sciences, 2010, 27 (6): 1 221-1 232 .

[6] 王卫光, 薛绪掌, 耿伟. 河套灌区地下水位的空间变异性及其克里金估值[J]. 灌溉排水学报, 2007, 26 (1): 18-21.

[7] 汪昌树, 杨鹏年, 于宴民, 等. 焉耆盆地绿洲区地下水硝态氮污染空间变异性研究及成因分析[J]. 灌溉排水学报, 2016, 35 (4): 65-70.

[8] 鲍晓蕾, 高辉, 胡良平. 多种填补方法在纵向缺失数据中的比较研究[J]. 中国卫生统计, 2016, 33 (1): 45-48.

[9] 李新, 程国栋, 卢玲. 空间内插方法比较[J]. 地球科学进展, 2000, 15 (3): 260-265.

[10] 张晓琴, 王敏. 基于主成分分析的成分数据缺失值插补法[J]. 应用概率统计, 2016, 32 (1): 101-110.

[11] 陈林. 基于GIS的流域水文数据的时空分析一以格兰德河流域径流数据为例[D]. 青岛: 山东科技大学, 2010.

[12] 戴新刚, 陈洪武. 对气象场进行地质统计插值研究[J]. 地球物理学报, 2004, 47 (6): 983- 990.

[13] 姜晓剑, 刘小军, 黄芬，等. 逐日气象要素空间插值方法的比较[J]. 应用生态学报, 2010, 21 (3): 624-630.
[14] 赵彦锋, 陈杰, 齐力，等. 不同采样尺度下土壤图和Kriging法的空间估值精度比较：以砂姜黑土典型地区的研究为例[J]. 土壤通报, 2011, 42 (4): 872-878.
[15] 张贝尔，黄标，赵永存，等. 采样数量与空间估值方法对华北平原典型土壤质量评估空间预测精度的影响[J]. 土壤, 2013, 45 (3): 540-547.
[16] 赵永存，黄标，孙维侠，等. 张家港土壤表层铜含量空间预测的不确定性评价研究[J]. 土壤学报，2007, 44 (6): 974-981.
[17] NATHAN P O, ALEX B M, BUDIMAN M. Digital soil property mapping and uncertainty estimation using soil class probability rasters[J]. Geoderma, 2015, 237/238: 190-198.
[18] 朱会义, 贾绍凤. 降雨信息空间插值的不确信性分析[J]. 地理科学进展, 2004, 23 (2): 34-42.
[19] COULIBALY P, EVORA N D. Comparison of neural network methods for infilling missing daily weather records[J]. Journal of Hydrology, 2007, 341: 27-41.
[20] JOHN B L. Sensitivity of channel mapping techniques to uncertainty in digital elevation data[J]. International Journal of Geographical Information Sciences, 2006, 20 (6): 669-692.
[21] 赵明伟，汤国安，田剑. AMMI模型的DEM内插方法不确定性研究[J]. 地球信息科学学报, 2012, 14 (1):62-66.
[22] 孟岑, 李裕元, 吴金水，等. 亚热带典型小流域总氮最大日负荷(TMDL)及影响因子研究：以金井河流域为例[J]. 环境科学学报, 2016, 36 (2):700-709.
[23] 刘梦霞, 周脚根, 黄新，等. 亚热带小流域COD负荷及影响因子分析[J]. 农业现代化研究, 2017, 38 (1):168-175.
[24] 王国荣, 俞耀明, 徐兆亮，等译. 数值分析(原书第三版)[M]. 北京: 机械工业出版社, 2005.
[25] 鲁帆, 朱奎, 宋昕熠，等. 基于核密度估计和Copula函数的降水径流丰枯组合概率研究[J]. 中国水利水电科学研究院学报, 2016, 14 (4): 297-303.
[26] YOZGATLIGIL, ASLAN, IYIGUN, et al. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data[J]. Theoretical and Applied Climatology, 2013, 112 (1/2): 143-167.

Analyzing the Uncertainty Induced by Methods Used to Calculate the Missing Data in Time Series: A Case Study Based on Meteorological and Hydrological Data in Small Watershed

SHI Jin[1,2], ZHOU Jiaogen[2], WANG Hui[1*], GAN Lei[2,3],
SHEN Jianlin[2], LI Xi[2], LI Yuyuan[2],WU Jinshui[2]

（1. College of Engineering, Hunan Agricultural University, Changsha 410128, China; 2. Key Laboratory of Agro-ecological Processes in Subtropical Region, Institute of Subtropical Agriculture, Chinese Academy of Sciences, Changsha 410125, China; 3. College of Resources and Environmental Sciences, Hunan Normal University, Changsha 410081, China）

Abstract:【Objective】Incomplete data is common in meteorological and hydrological analysis and this paper analyzed uncertainty caused by estimating such missing date using different interpolation methods.【Method】We take meteorological data, including minimum temperature, maximum temperature, solar radiation; and hydrological data, including rainfall and stream flow, collected from a long-term field experiment in a typical small watershed in a subtropical zone as examples. We developed a computer model to simulate them. The difference between the simulated results using five interpolation methods: the linear interpolation method (LIM), the K-Nearest neighbor interpolation method (KNNM), the polynomial interpolation method (PIM), the spline interpolation method (SIM) and kernel density estimation method (KDEM), was compared. We then analyzed the uncertainty resulted from sampling frequency (daily and monthly) and data missing degree (1%, 5%, 10%, 15%, 20%). Root mean square error (*RMSE*), absolute mean error (*MAE*) and the Pearson correlation coefficient (*r*) were used as criterion to evaluate the five methods.【Result】① All five methods worked well in estimating the missing meteorological data with r varying from 0.62 to 0.99 ($P<0.05$). In general, the KDEM and PIM were more accurate than other three methods. ② Accuracy of all five methods deteriorated when the sampling time frequency changed from daily to monthly and data missing degree increased. ③The coefficient of variance (*CV*) of the data sets was significantly correlated with the valuation indexes (*RMSE*, *MAE* and *r*) ($P<0.05$).【Conclusion】The KDEM and PIM are relatively more reliable, and the coefficient of variance (*CV*) of data sets is critical to the accuracy of all five interpolation methods.

Key words: missing data; interpolation methods; coefficient of variance; uncertainty; time series

责任编辑：白芳芳