

DOI:10.3876/j.issn.1000-1980.2021.01.006

基于矩阵分解的降水时空插值方法

陈 华¹,盛 晟¹,夏润亮²,马 瑞³,朱跃龙⁴,李 涛²

(1. 武汉大学水资源与水电工程科学国家重点实验室,湖北 武汉 430072;
2. 黄河水利委员会黄河水利科学研究院,河南 郑州 450003; 3. 长江空间信息技术工程有限公司,湖北 武汉 430010;
4. 河海大学计算机与信息学院,江苏 南京 210098)

摘要: 为了提高基于雨量站网观测值估算降水空间分布的精度,同时考虑降水的时序发展趋势和站点空间分布,将传统的插值方法(反距离权重法、普通克里金法)和 Funk 矩阵分解(Funk-SVD)模型结合,提出了一种基于矩阵分解的插值方法,并采用黄河流域小浪底到花园口区间 2009—2012 年多场降水的日观测数据进行了精度检验。结果表明,通过与 Funk-SVD 模型结合,反距离权重法和普通克里金法在平均绝对误差、百分误差等评价指标上的精度提升均超过 15%,改善效果在周围站点分布不均匀或降水量很大的情况下尤为明显,可以大幅度减少传统方法在降水插值过程中的估计误差,提升空间估算的精度。

关键词: 降水估算;矩阵分解;时空插值;反距离权重法;克里金法

中图分类号: TV125; P338 **文献标志码:** A **文章编号:** 1000-1980(2021)01-0035-07

Spatiotemporal interpolation method of rainfall based on matrix decomposition

CHEN Hua¹, SHENG Sheng¹, XIA Runliang², MA Rui³, ZHU Yuelong⁴, LI Tao²

(1. State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China;
2. Yellow River Institute of Hydraulic Research, YRCC, Zhengzhou 450003, China;
3. Changjiang Spatial Information Technology Engineering Corporation, Wuhan 430010, China;
4. College of Computer and Information, Hohai University, Nanjing 210098, China)

Abstract: To improve the estimation accuracy of rainfall spatial distribution based on the gauge network, an interpolation method is proposed based on the matrix factorization. By combining the traditional interpolation methods, including the inverse distance weighting (IDW) method and the ordinary Kriging (OK) method, and the Funk-SVD model, this method considered the temporal development of precipitation and the spatial distribution of gauges. The daily observation data of rainfall events from 2009 to 2012 between Xiaolangdi and Huayuankou were selected for the accuracy test. The results show that through combining with the Funk-SVD model, the estimation errors of the IDW and OK can be reduced over 15% in terms of MAE and PERC, and the improvement is especially obvious when the surrounding stations are unevenly distributed, or the rainfall is heavy. The proposed method can greatly reduce the estimation error of traditional methods in the precipitation interpolation process and help improve the accuracy of spatial estimation.

Key words: rainfall estimation; matrix decomposition; spatiotemporal interpolation; inverse distance weighting method; ordinary Kriging method

降水数据在国家防汛、抗旱等自然灾害应急决策管理中具有重要的作用,精确的雨量估算和空间分布确定能为自然灾害的应急决策提供科学准确的数据^[1],提升应急管理水平,社会效益显著。雨量站网是目前最主要的降水观测和采集系统,其主要特点是便捷、实时和准确;但由于站点是离散的,需要借助于空间插值方法来得到降水的空间连续分布。

现有的降水空间插值方法,不论是整体插值(如边界内差法、趋势面分析法)还是局部插值(如克里金

基金项目: 国家重点研发计划(2018YFC0407904)

作者简介: 陈华(1977—),男,教授,博士,主要从事水文模拟和水利信息化研究。E-mail:chua@whu.edu.cn

引用本文: 陈华,盛晟,夏润亮,等. 基于矩阵分解的降水时空插值方法[J]. 河海大学学报(自然科学版),2021,49(1):35-41.

CHEN Hua,SHENG Sheng,XIA Runliang, et al. Spatiotemporal interpolation method of rainfall based on matrix decomposition[J]. Journal of Hohai University(Natural Sciences),2021,49(1):35-41.

法^[2-3]、泰森多边形法^[4]、反距离权重法^[5]),都是基于插值点与雨量站的地理位置关系,利用插值的当前时刻降水信息进行插补。这些方法主要是基于“地理学第一定律”^[6]的基本假设:空间位置上越靠近的点,具有相似特征值的可能性越大;而距离越远的点具有相似特征值的可能性越小。这些方法都仅仅考虑了空间上的关系,没有考虑降水序列的过去临近信息对当前值的影响,无法将历史降水信息应用于当前时刻降水判断和未来趋势预测。也有不少学者开展了降水时空插值方法研究,比如 STARMA 和具有时间扩展模块的克里金插值方法^[7-9]。现有的结果都表明,考虑时间维度的插值方法精度都比只考虑空间维度的插值方法精度高,但由于算法过于复杂,并没有被广泛应用。

为了充分利用降水的时间序列和空间信息以提高降水空间分布及雨量估算的精度,本文引进了推荐系统中的矩阵分解方法,来实现降水数据的时空插值,并对降水空间插值结果进行了评价。

1 插值方法

1.1 Funk-SVD 模型

矩阵分解是解决协作过滤问题最常用的方法之一,其将用户的商品偏好程度看作用户-商品评分矩阵,并使用已知用户对商品的评分来预测用户在商品选择中的偏好并进行推荐^[10-11]。由于具有较高的预测准确度,矩阵分解已成为环境科学、生物医学等许多领域常用的插值方法^[12-14]。González-Macías 等^[15]使用正矩阵分解法来识别和确定海底沉积物中人为重金属的来源;Lee 等^[16]将非负矩阵分解法应用于基因数据的表达,量化了不同剂量的 pan-PPAR 兴奋剂在小鼠体内引起的分子变化。Funk-SVD 模型是 Funk^[17]提出的一种矩阵分解的变体模型,其基本思想是用户和商品可以分别由从用户-商品评分矩阵推断出的潜在特征向量来描述,由用户和项目特征之间的高度对应性形成预测和推荐^[18],这种模型也称作隐语义模型。

基于 Funk-SVD 模型的降水时空插值模型结构如图 1 所示。空间上不同点在多个时刻的降水数据可以看作是一个内在相关的矩阵 $P=(P_{ij})$,其中列代表时间,行代表空间。无降水记录的点在矩阵 P 中对应位置为空值,即为待插值点。为了得到该位置的值,可以根据已有的降水记录对矩阵 P 进行分解,通过分解后两个矩阵的乘积来补全原矩阵 P 中空缺位置的值。

具体来说,降水数据矩阵的时间和空间之间存在隐式关系,无法直接求解,所以需要潜在因子来建立间接关系。假设存在影响降水的 q 个潜在因子,从图 1 可以看出,该矩阵 P 可以分解为空间特征矩阵 X 和时间特征矩阵 Y ,分别包含 m 个 q 维行向量 $X_1, X_2, \cdots, X_m$ 和 n 个 q 维列向量 $Y_1, Y_2, \cdots, Y_n$,其中 $X_i(i=1,2,\cdots,m)$ 代表 i 站点 q 个潜在因子的影响程度,其包含的特征数据越相近,潜在因子对降水的影响效果越相近; $Y_j(j=1,2,\cdots,n)$ 描述 j 时刻潜在因子的关联关系,相关程度越高,对应的特征值越大。因此,可以基于潜在因子的影响程度和关联关系计算某时某地的降水量。也就是说, P_{ij} 可以由 X_i 和 Y_j 相乘得到,其计算公式如下:

$$P_{ij} = \sum_{u=1}^q X_{iu} Y_{uj}$$

(1)

1.2 插值方法计算流程

基于 Funk-SVD 模型的降水时空插值方法具体计算流程如图 2 所示。图 2 涉及的主要有 4 个计算步骤:

a. 对于 j 时刻在流域上的某点 I ,可以确定一个由 m 个站点和 n 个时刻组成的降水数据矩阵。所涉及的 m 个站点是在计算所有可能排列组合后由空间均匀度 L 选出的最优集合,空间均匀度是点集的空间关系度量,其值越大,点分布越均匀,计算公式为

$$L = \frac{4a}{\pi A}$$

(2)

式中: A ——涵盖所有站点的矩形网格面积; a ——独占圆面积的总和,某点的独占圆的定义为以该点为圆

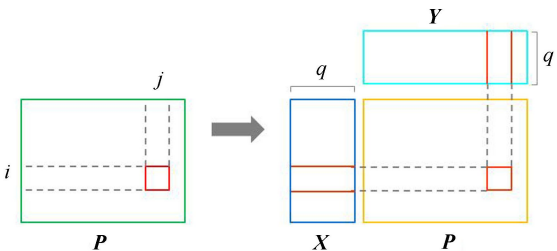


图 1 基于 Funk-SVD 模型的降水时空插值方法结构示意图

Fig.1 Structure diagram of spatiotemporal interpolation method of rainfall based on Funk-SVD model

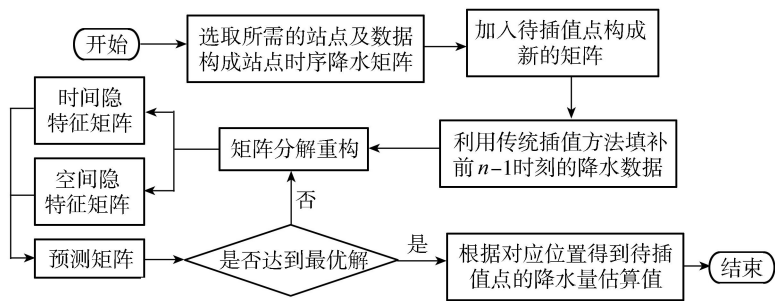


图 2 基于 Funk-SVD 模型的降水时空插值方法计算流程

Fig. 2 Calculation flowchart of spatiotemporal interpolation method of rainfall based on Funk-SVD model

心,与最近点距离的一半为半径的圆。考虑到矩阵分解的效率和精度,时刻数 n 的取值如下:

$$n = \begin{cases} j - t_0 + 1 & j - t_0 < N \\ N + 1 & j - t_0 \geq N \end{cases} \quad (3)$$

式中: t_0 ——降水事件的开始时刻; N ——前期影响时刻总数阈值。如果降水持续时间不长, n 个时刻就从降水开始时刻 t_0 开始取到插值时刻 j 。如果降水持续时间长,对于距开始时间超过 N 个时刻的插值时刻来说, n 个时刻就取 j 时刻以及前 N 个时刻。

b. 增加一行代表流域上一点 I ,得到一个新矩阵 R 。矩阵第 $m+1$ 行前 $n-1$ 个值为已知的 j 时刻前的降水,由传统插值方法计算得到。第 n 个值为空,代表待插值点。

c. 利用 Funk-SVD 模型将矩阵 R 分解成时间特征矩阵 X 和空间特征矩阵 Y 。Funk-SVD 模型通过最小化所有已知点的预测值平方误差来计算时间和空间的潜在特征。此外,为了防止出现过度拟合的情况,将正则化方法引入目标函数:

$$\min: S = \sum_{i=1}^{m+1} \sum_{j=1}^n E_{ij}^2 + \lambda \sum_{i=1}^{m+1} \sum_{u=1}^q |X_{iu}|^2 + \lambda \sum_{j=1}^n \sum_{u=1}^q |Y_{uj}|^2 \quad (4)$$

其中

$$E_{ij}^2 = (R_{ij} - R'_{ij})^2$$

式中: S ——目标函数; λ ——控制正则化程度的超参数; R'_{ij} 为重构矩阵 R' 的元素。为了最小化目标函数,使用随机梯度下降算法 (SGD) 来解决上述优化问题。 X_{iu} 和 Y_{uj} 会朝着梯度下降最快的方向减少,从而达到最优解。为了得到空间特征向量 X_i 的最优值,所有已知时刻的降水值 P_{ij} 都会被分解,即 X_i 中每个特征值和所有从历史降水信息中提取的时间特征向量 Y_j 都相关。随机梯度下降算法迭代公式为

$$X_{iu} = X_{iu} - \alpha \frac{\partial E_{ij}^2}{\partial X_{iu}} = X_{iu} + 2\alpha(E_{ij}Y_{uj} - \lambda X_{iu}) \quad (5)$$

$$Y_{uj} = Y_{uj} - \alpha \frac{\partial E_{ij}^2}{\partial Y_{uj}} = Y_{uj} + 2\alpha(E_{ij}X_{iu} - \lambda Y_{uj}) \quad (6)$$

式中: α ——机器学习的学习率。

d. 将优化得到的时空特征矩阵相乘,可以得到一个无限逼近于原矩阵的重构矩阵 R' ,且 R' 中每个位置都有值。原始矩阵 R 和重构矩阵 R' 中各元素为——对应的关系,重构矩阵 R' 第 $m+1$ 行第 n 列的值即为点 I 在 j 时刻的降水量。

1.3 评价指标

使用交叉验证的方法进行插值的精度验证,即每次将一个站点看作未知点,利用其余所有站点进行插值,将结果和观测值进行比较。采用均方根误差 (RMSE)、平均绝对误差 (MAE)、百分误差 (PERC) 和双样本 Kolmogorov-Smirnov 检验的概率值 (KS)^[19-20] 等 4 个评价指标来验证插值精度,采用反距离权重法和普通克里金法这两种广泛应用的空间插值方法与 Funk-SVD 模型结合进行插值。这两种方法都是基于站点的观测值和权重来得到待插值点的计算值,计算公式如下:

$$Z = \sum_{i=1}^m w_i Z_i \quad (7)$$

式中: Z ——待插值点的计算值; Z_i ——站点的观测值; w_i ——站点的权重,其中 IDW 和 OK 两种方法计算站

点权重的公式分别为

$$w_i = \frac{d_i^{-p}}{\sum_{i=1}^m d_i^{-p}}$$

(8)

$$\begin{cases} \sum_{i=1}^m \gamma(x_h, x_i) w_i + \mu = \gamma(x_h, x) \\ \sum_{i=1}^m w_i = 1 \end{cases}$$

(9)

式中: d_i ——站点到目标点的距离; p ——距离的幂; μ ——拉格朗日乘子; γ ——空间上两点间的半变异函数; x_i, x_h ——插值利用到的站点; x ——待插值点。

2 实例验证

2.1 研究区域及数据

选择黄河流域的小浪底到花园口区间为研究区域,该区域位于黄河下游(东经 $109^{\circ}42' \sim 113^{\circ}54'$ 、北纬 $33^{\circ}17' \sim 37^{\circ}30'$),是黄河流域主要暴雨来源区之一,暴雨来势急、汇流快,短时间内可能出现洪水陡涨局面,对黄河防汛安全威胁较大。区间内有伊洛河和沁河两条国家一级支流,是暴雨洪水多发地区。降水量年内分布不均匀,集中在6—9月。研究区有16个雨量站(图3),选取2009—2012年不同气象条件下8次降水事件的日降水数据来验证本文提出的基于 Funk-SVD 模型的降水时空插值方法(以下简称本文方法)。表1为8个降水事件的详细信息,每场降水的空间分布如图4所示。这些降水事件持续时间为6~10 d,分布在6—9月,面降水量跨度很大,从17~67 mm,具有较好的代表性。

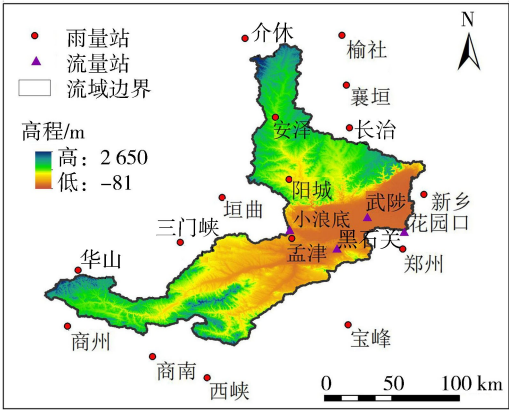


图3 黄河流域小浪底到花园口区间测站分布

Fig.3 Distribution of gauge stations between Xiaolangdi and Huayuankou

表1 降水事件信息

Table1 Information of selected rain events			
降水事件	时间	历时/d	面总降水量/mm
1	2009-07-05—2009-07-14	10	56.6
2	2009-09-03—2009-09-10	8	17.2
3	2010-06-29—2010-07-05	7	21.7
4	2010-08-08—2010-08-15	8	47.8
5	2011-07-01—2011-07-07	7	66.8
6	2011-08-15—2011-08-23	9	60.7
7	2012-07-21—2012-07-26	6	22.7
8	2012-08-17—2012-08-22	6	32.2

2.2 分解矩阵参数的影响

采用 Funk-SVD 模型利用矩阵进行插值时,需要 m 和 N 两个参数来确定矩阵的规模,其中 m 是插值所需的站点数,和矩阵的行数有关;结合 N 和插值时刻可以确定矩阵的列数。为了得到最佳的插值效果,对这两个参数选取了多种值进行演算和比较,结果见表2。可以看出,当 $m=7$ 时, L 达到最大值;当 $N=6$ 时, RSME 达到最小值,所以本研究中 m 取值为7, N 取值为6。

2.3 整体精度评价

使用4个不同的评价指标对各插值方法的计算结果进行了总体评价,结果如表3所示。从表3可以看出,4种插值方法的精度均在合理范围内,表明它们可以很好地应用于研究区域的降水插值计算。使用不同评价指标的评价结果是一致的,即通过与 Funk-SVD 结合, IDW 和 OK 精度均得到了明显的提升。在 RSME 上精度提升了9%左右,在其余3个指标上的精度提升均在15%以上。除去 KS,其他评价指标对 IDW 和 OK 的精度评价结果的差异很小,几乎可以忽略。

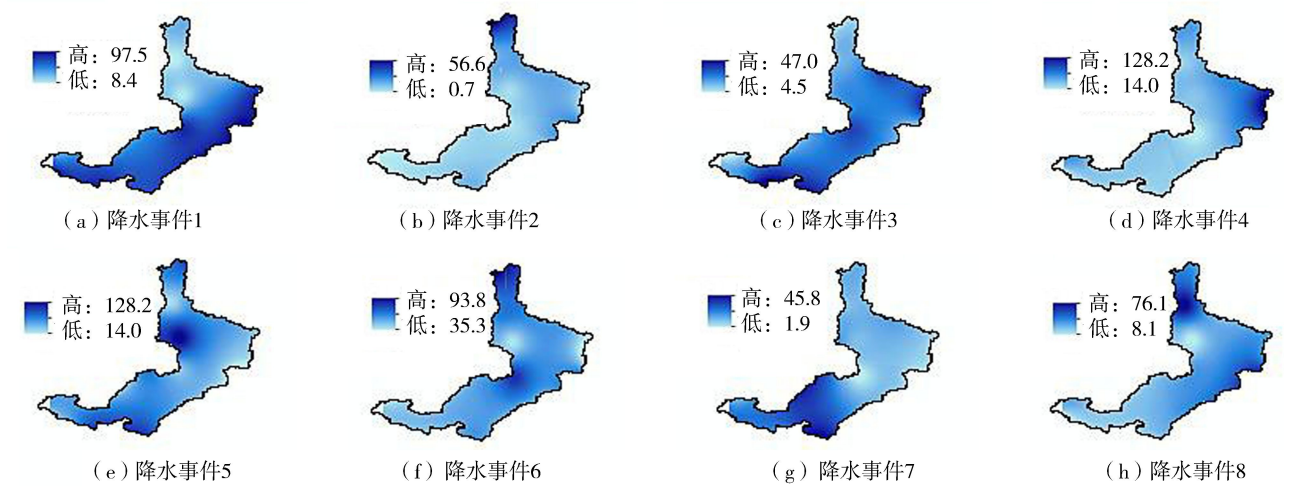


图 4 8 场降水事件的降水量空间分布 (单位:mm)

Fig.4 Spatial distribution of rainfall in eight selected events

表 2 不同 m 、 N 值的计算结果

Table 2 Calculation results of different m and N

m	N	L	RSME/mm
5	4	0.1527	5.354
6	5	0.1549	5.259
7	6	0.1649	5.081
8	7	0.1603	5.393
9	8	0.1601	5.421

表 3 8 场降水事件的综合评价结果

Table 3 Comprehensive evaluation results of 8 rainfall events

插值方法	RSME/mm	MAE/mm	PERC	KS
IDW	8.97	5.48	2.76	0.43
OK	8.98	5.47	2.75	0.39
Funk-SVD-IDW	8.20	4.64	1.93	0.32
Funk-SVD-OK	8.09	4.57	2.01	0.33

注:IDW、OK 分别为反距离权重方法和普通克里金方法,Funk-SVD-IDW、Funk-SVD-OK 分别为本文方法与前两种传统方法结合进行插值的方法。

图 5 为 4 种插值方法在每场降水事件交叉验证中的所有站点的误差均值,可以看出,对于每一场降水事件,通过与本文方法结合,IDW 和 OK 的精度在各项指标上均有所改善,这种改善在 PERC 和 KS 上更加明显。RSME 和 MAE 对不同事件的评价结果是类似的,这是因为它们的数值与降水量大小有非常大的关联,所以降水量越大,计算得到的误差就越大。而对个别事件,使用 PERC 和 KS 会给出不一样的评价结果。比如 5 号降水事件,它的面总降水量是所有事件中最大的,所以绝对误差也大,相应 RSME 和 MAE 也最大;而 PERC 的数值大小与相对误差的关系更大,受绝对误差影响较小,因此 PERC 反而小。

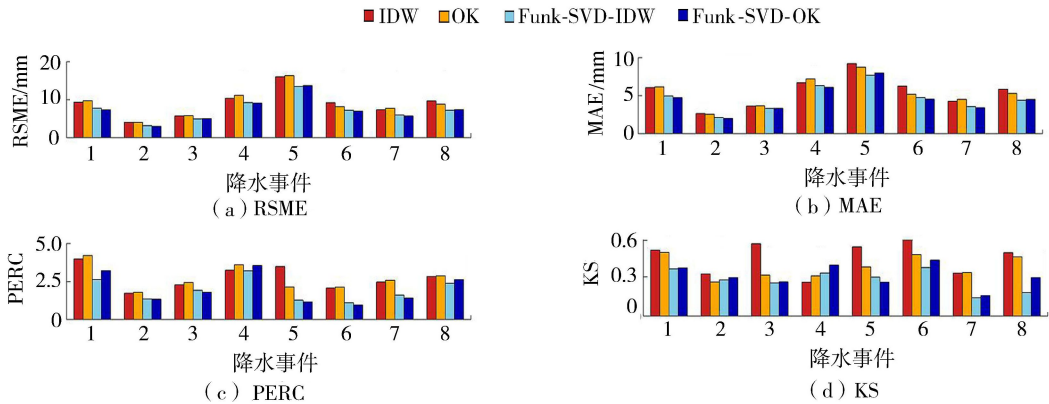


图 5 降水事件插值精度评价结果

Fig.5 Interpolation accuracy evaluation results of each rainfall event

2.4 代表性事件的站点误差比较

为了更好地评价本文方法的精度,选择 5 号大雨事件和 7 号小雨事件这两场典型的降水事件进行比较分析。图 6 和图 7 为使用不同插值方法得到的插值与误差结果。从图 6 和图 7 的(b)(c)和(e)(f)可以发现,无论是大雨还是小雨,降水量越大的点,往往插值产生的误差越大,并且大雨事件的插值误差会比小雨事

件的大;同时,周围站点分布不均的点也会产生相对大的误差。图6和图7的(d)(g)中红色(改善)的点比蓝色(未改善)的点多,说明无论大雨还是小雨,通过与Funk-SVD结合,可以明显提升传统方法插值的精度。

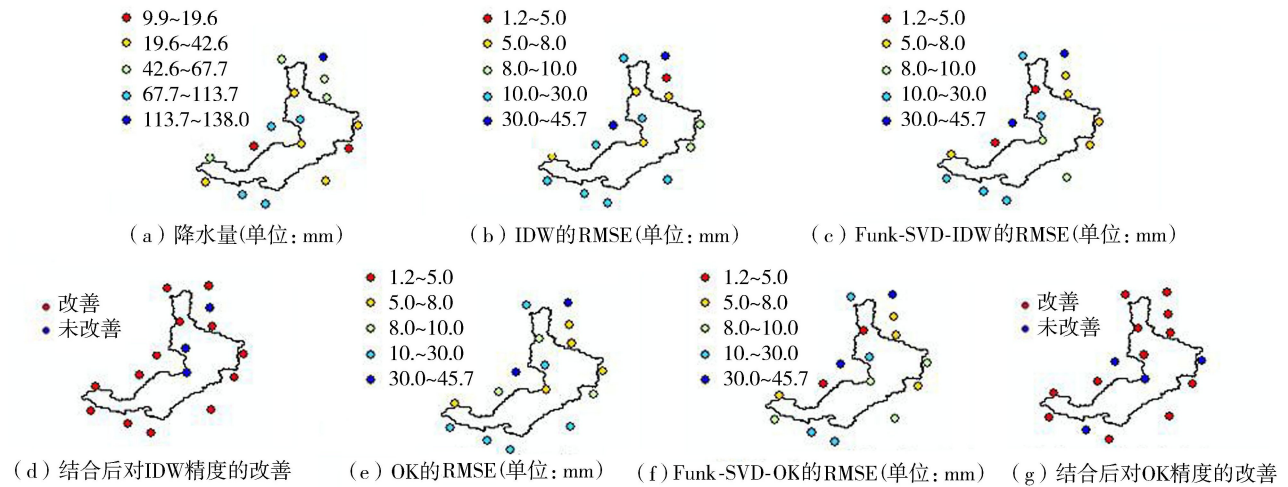


图6 5号降水事件插值结果

Fig.6 Interpolation results of No.5 rainfall event

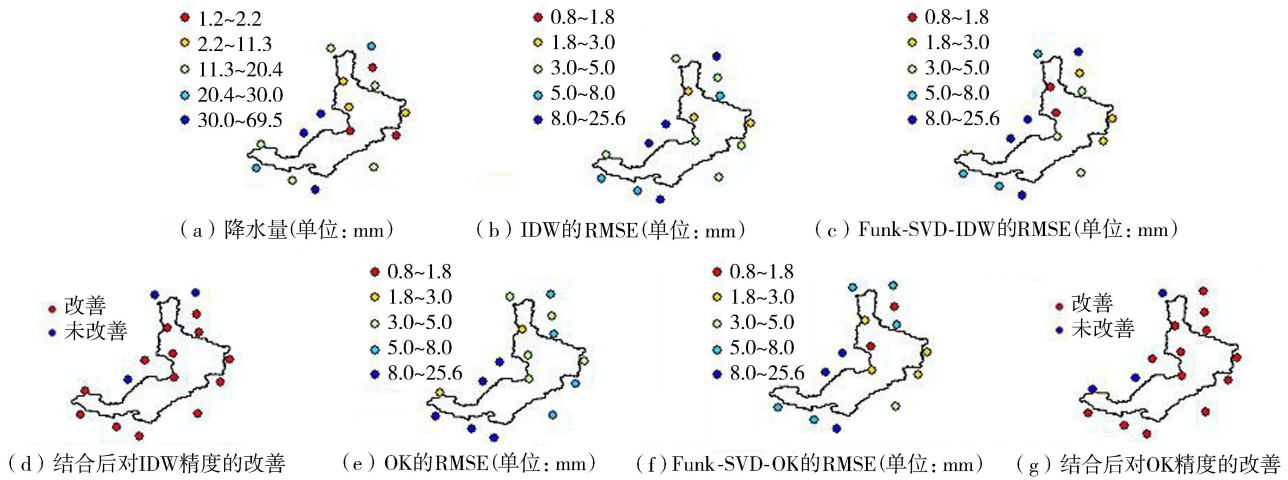


图7 7号降水事件插值结果

Fig.7 Interpolation results of No.7 rainfall event

3 结 语

本文基于站点的空间分布关系,利用当前时刻降水信息和历史降水信息,结合传统插值方法,提出了一种基于Funk-SVD模型的降水插值方法。采用黄河流域小浪底到花园口区间的8场降水事件中16个雨量站的日降水数据,通过交叉验证的方式对提出的方法进行精度检验,发现与Funk-SVD模型结合可以提升IDW和OK插值的精度,证实了本文方法的可行性。本文方法提高了降水空间估算的精度,为降水时空插值的相关研究提供了新的思路。

参考文献:

[1] 俞彦,张行南,张鹏,等. 基于SCS模型和新安江模型的雨量预警指标综合动态阈值对比[J]. 水资源保护,2020,36(3): 28-33. (YU Yan, ZHANG Xingnan, ZHANG Peng, et al. Comparison of comprehensive dynamic threshold of rainfall warning indicators based on SCS model and Xin'anjiang model[J]. Water Resources Protection,2020,36(3):28-33. (in Chinese))

[2] DELHOMME J P. Kriging in the hydrosiences[J]. Advances in Water Resources,1978,1(5):251-266.

[3] 范银贵. 空间插值方法在绘制降水量等值线中的应用[J]. 水利水电科技进展,2002,22(3):48-50. (FAN Yingui. Application of spatial interpolation method for drawing precipitation isolines[J]. Advances in Science and Technology of Water

- Resources,2002,22(3):48-50. (in Chinese)
- [4] THIESSEN A H. Precipitation averages for large areas[J]. Monthly Weather Review,1911,39(7):1082-1089.
- [5] SHEPARD D. A two-dimensional interpolation function for irregularly-spaced data [C]//Proceedings of the 1968 23rd ACM National Conference. New York:ACM,1968:517-524.
- [6] TOBLER W R. A computer movie simulating urban growth in the detroit region[J]. Economic Geography,1970,46: 234-240.
- [7] CLIFF A,ORD J K. Space-time modelling with an application to regional forecasting[J]. Transactions of the Institute of British Geographers,1975,64:119-128.
- [8] PFEIFER P E,DEUTRCH S J. A three-stage iterative procedure for space-time modeling phillip[J]. Technometrics,1980, 22 (1);35-47.
- [9] DALEZIOS N R, ADAMOWSKI K. Spatio-temporal precipitation modelling in rural watersheds [J]. Hydrological Sciences Journal,1995,40(5):553-568.
- [10] KUMAR B. A novel latent factor model for recommender system[J]. Journal of Information Systems and Technology Management, 2016,13(3):497-514.
- [11] TAKACS G, PILASZY I, NEMETH B, et al. Scalable collaborative filtering approaches for large recommender systems [J]. Journal of Machine Learning Research,2009,10:623-656.
- [12] XIE Y, BERKOWITZ C M. The use of positive matrix factorization with conditional probability functions in air quality studies:an application to hydrocarbon emissions in Houston, Texas [J]. Atmospheric Environment,2006,40(17):3070-3091.
- [13] XUE J L, ZHI Y Y, YANG L P, et al. Positive matrix factorization as source apportionment of soil lead and cadmium around a battery plant (Changxing County, China) [J]. Environmental Science and Pollution Research,2014,21(12):7698-7707.
- [14] ZHANG K, SHANG X, HERRMANN H, et al. Approaches for identifying PM2. 5 source types and source areas at a remote background site of South China in spring [J]. Science of The Total Environment,2019,691:1320-1327.
- [15] GONZÁLEZ-MACÍAS C, SÁNCHEZ-REYNA G, SALAZAR-CORIA L, et al. Application of the positive matrix factorization approach to identify heavy metal sources in sediments. a case study on the Mexican Pacific Coast [J]. Environmental Monitoring and Assessment,2014,186(1):307-324.
- [16] LEE C M, MUDALIAR M A V, HAGGART D R, et al. Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology [J] PLoS ONE,2012,7(12): e48238.
- [17] FUNK S. Netflix update: try this at home [EB/OL]. (2006-12-11) [2020-09-30]. <http://sifter.org/~simon/journal/20061211.html>.
- [18] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems [J]. Computer,2009,42(8): 30-37.
- [19] WASKO C, SHARMA A, RASMUSSEN P. Improved spatial prediction:a combinatorial approach [J]. Water Resources Research, 2013,49(7):3927-3935.
- [20] HUANG H, LIANG Z, LI B, et al. A new spatial precipitation interpolation method based on the information diffusion principle [J]. Stochastic Environmental Research,2019,33(3):765-777.

(收稿日期:2020 - 10 - 30 编辑:熊水斌)