黄崇福，张馨文. 地理空间信息扩散技术实证研究——以四川省三台县洪水灾害为例[J]. 灾害学，2022，37(2)：89－101.［HUANG Chongfu and ZHANG Xinwen. Empirical Research on Geospatial Information Diffusion Technique —Taking Flood Disaster in Santai County, Sichuan Province, as an Example[J]. Journal of Catastrophology, 2022, 37(2): 89－101. doi: 10.3969/j.issn.1000－811X.2022.02.016.］

地理空间信息扩散技术实证研究*

——以四川省三台县洪水灾害为例

黄崇福[1,2]，张馨文[2]

（1. 北京师范大学 环境演变与自然灾害教育部重点实验室，北京 100875；
2. 北京师范大学 地理科学学部灾害风险科学研究院，北京 100875）

摘　要：插值，是推测地理空间中空白单元处地表现象的重要途径。协同克里金插值(CK)、地理加权回归(GWR)和回传神经网络(BP－ANN)等，在满足相应条件的情况下，都是很好的插值方法，但不具有普适性。在观测单元不多，数据离散性较大的情况下，信息扩散技术的插值，比这些模型的效果都好。该文以四川省三台县 2018 年和 2020 年发生的两次大洪水，采集的 25 个村的房屋损失、农业损失和庄稼被淹三类水灾灾情数据组成 6 个案例，以村庄与河流的距离、GDP 和坡度等为自变量，以灾情为因变量，实证了地理空间信息扩散技术用于插值的普适性。信息扩散的自学习离散回归模型(SLDR)，预测误差较小，且没有明显的预测误差小于基准误差的情况。CK 在所有案例中，均是预测误差小于基准误差，说明插值无效；GWR 在 5 个案例中也出现相同情况。虽然 BP－ANN 的基准误差很小，但预测误差却比基准误差高出近一个数量级，也远高于其他模型，表明能够高度拟合训练样本的回传神经网络模型，并不适用于复杂地表现象的插值。

关键词：空间插值；信息扩散；协同克里金；地理加权回归；神经网络；基准误差；预测误差；三台县
中图分类号：X43；X915.5；U44　**文献标志码：**A　**文章编号：**1000－811X(2022)02－0089－13
doi：10.3969/j.issn.1000－811X.2022.02.016

在世界格局发生急剧变化的今天，人们只有超越以往的经验化模式，才能更好地认识世界，包括形形色色的地表现象。自然灾害，是一种综合自然和人文特征的地表现象，用以往案例形成的成灾经验，很难正确认识环境和社会均发生了显著变化的自然灾害。

以重大自然灾害灾情快速评估为例，以往的经验化模式，是用历史灾害资料建立经验公式，一旦发生自然灾害，根据致灾因子强度和灾区的自然和社会数据，用这些公式对灾情进行快速评估。例如，一旦发生破坏性地震，可根据震级对灾情进行粗估[1]。这类远隔千山万水的快速评估，我们称之为“隔空判灾”[2]，缺点是精度较低，大多只能保证不出数量级错误(相差在 10 倍之内)，而且只能对县及县以上地理单元内的灾情进行评估[3]，很少细化到乡镇，更无法细化到村庄，评估结果支撑不了精准救灾。

现代信息技术的发展，为较高精度地快速评估灾情和救助需求，提供了一条新路径：由基层灾害信息员、卫星遥感和无人机等观测得到的局部数据，推测全灾区的情况。已观测地理单元是采点，空白单元是信息孤岛。外推的依据，是从观测得到的数据中总结出的因果关系。这类借助实时数字化技术进行的快速评估，我们称之为“采点外推”[4]，优点是自然和人文的变化已经在采点数据中体现，推测出的空白地理单元中的灾情精度较高，可细化到村庄，助力精准救灾。推测，是一种判断各种各样情况的行为，甚至于有纯主观性的层次分析法[5]，半主观的模糊综合评价[6]，常见的则是统计回归[7]。

当我们灾后第一时间采集灾区数据时，受信息员数量少、卫星扫描时段不凑巧、天气多云、投送无人机耗时长、灾区部分通讯中断等不利影响，2 h 内获得灾情数据，只能覆盖灾区的部分地

* 收稿日期：2021－11－17　　修回日期：2022－03－23
基金项目：国家重点研发计划项目(2017YFC1502902)；国家自然科学基金项目(41671502)
第一作者简介：黄崇福(1958－)，男，汉族，云南个旧人，博士，教授，博士生导师，主要从事自然灾害风险分析的研究.
E－mail：hchongfu@bnu.edu.cn

理单元，覆盖不了信息孤岛。只有推测出信息孤岛中的灾情，才能对整体灾情有较全面的认识，才能科学地制定出精准救灾方案。

在地理学中，有很多种数学插值法被用来研究推测问题。然而，除温度、降雨量等物理场外，多数地理特征值在空间上的分布，并不连续，数学插值的结果，误差较大。于是，统计回归方法(Statistical Regression Method)和人工神经网络(Artificial Neural Networks，ANN)等，被用来研究地表现象中的推测问题。但是，当观测数据提供的样本较小且样本点间有矛盾冲突时，这些方法的精度都不高。

理论和仿真实验证明，由于地理空间信息扩散技术，对被插值的参数，既没有连续性的要求，也没有与自变量间线性关系假设的约束，还具有优化处理小样本的功能，矛盾冲突也不影响总结学习因果关系的收敛性，所以能明显提高推测精度[8]。本文通过对四川省绵阳市三台县洪水灾害的实证研究，演示地理空间信息扩散技术在推测精度方面的优势。

1 地理学中的插值法

插值(Interpolation)是一个数学概念：给定函数$f(x)$在n个互异点的值$f(x_i)$，$i=1$，…，n，寻求函数$\varphi(x)$逼近$f(x)$，若要求$\varphi(x_i)$逼近$f(x_i)$，则称之为插值问题。$\varphi(x)$称为$f(x)$的插值函数，x_i称为插值节点。用$\varphi(x)$推测插值节点间任一点的函数值，称为插值；$\varphi(x_i)$对$f(x_i)$的逼近，称为似合。

直观地讲，用离散数据估算出其背后的函数在其它点处的近似值，就是数学插值。基本的数学假设是离散数据产生于一个连续函数。插值的目的是填充离散数据，形成较完整的函数图像。

物理空间中的插值，是人们试图对尚未实际测量场的连续场的值进行合理估计。空间插值用于地理学，则是指人们试图对尚未观测的地理单元的某个地理特征值进行合理估计。与数学插值较大的区别是，数学上的插值节点没有几何大小，而地理学中的插值节点是有几何大小的地理单元，从极细的栅格点，到行政单元，都是有几何大小的地理单元。

正如人们在地理制图中，利用有限个点处的取值，使用插值算法，计算丢失的信息，填充图像一样，人们在地表现象研究中，也利用有限个地理单元上的取值，使用插值算法，推测没有取值单元上的情况。所不同的是，制图插值涉及的单元通常很小且形状规则，能近似满足插值函数对连续性的要求；而地表现象中的单元，通常较大且不规则，其上的地理特征数据分布，并不连续。

人们在GIS中使用的插值技术，分确定性方法和地统计方法两种，例如，全局多项式、局部多项式、样条插值、反距离加权等，是确定性方法；而克里金法(Kriging)、地理加权回归(Geographically Weighted Regression，GWR)是地统计方法(Geostatistical Method)。即使是对连续表面的定量评估，这些插值方法的准确度也存在较大差异。研究表明，地统计方法优于确定性方法[9]。

确定性插值方法，又称“内插法”，也就是前述的数学插值。确定性是指，观测值只有测量误差，随机性可忽略不计。最简单的确定性插值方法，是求解由比例关系建立的方程，并由此衍生出多项式插值方法。为了让构造的函数既穿过观测点，函数图又形像，人们可将全部数据分割成若干部分，分段插值，再通过最高三阶的多项式，将插值用到的多个函数，尽可能平滑地连接起来。这些用到的函数，就是所谓的“样条”。

确定性插值方法中的反距离加权法，则是假定每个观测点都会存在局部影响，距离较近的事物更相似，因此对于被插值点，距离其越近的观测点影响越大。这种影响的大小，用权值来量化。通过加权求和，进行插值。权值计算方法不同，插值差异很大。最简单的取权值方法，是归一化距离倒数计算权值[10]；复杂一些的，则用到软化参数等[11]。

虽然确定性插值方法的精度不高，但由于简单、易操作，并能起到数据光滑作用，其在地理学中被广泛使用。

地统计方法，不仅仅是将空间坐标点和其地理特征值组成的空间分布，视为一个具有因果关系的随机场，而且认为空间中两个不同点处的取值具有相关性。借用随机过程理论，人们发展出了克里金插值，也译为克里格插值。如果仅仅考虑地统计方法中样本点的随机性，认为样本点具有空间独立性，则可用地理加权回归法来估计空间坐标点和其地理特征值间的因果关系。而通过随机样本训练的后传神经网络(Back Propagation Artificial Neural Network，BP－ANN)，也是一种统计关系。

为本文研究的方便，下面我们简述协同克里金(CollaborativeKriging，CK)插值、GWR和BP－ANN的基本原理和数学模型。为保持这三个模型的传统表述，在不引起混乱的情况下，各模型中的数学符号相对独立。也就是说，同一个符号，在不同模型意义可能不同。

1.1 协同克里金插值

克里金插值是依据协方差函数对随机场进行空间建模和插值的回归算法[12－13]。该方法20世纪60年代产生于地质学界，是一种地质统计学方法，后来被大量用于地理学中，才有了地统计方法的统称。

令集合X由一些空间点x组成。x的三个直角坐标通常记为x_u，x_v，x_w，即，$x=(x_u, x_v, x_w)$；空间点集合记为$\{x\}$，即，$X=\{x\}$。当一个空间变

量 Z 在 x 点的取值 $Z(x)$ 是一个随机数时，称 $\{Z(x) \mid x \in X\}$ 是一个随机场。克里金插值的思想，是将空间参数 x 视为随机过程中的时间参数 t，从而可以使用随机过程中的统计方法，实现插值。

当随机过程的统计特性不随时间的推移而变化时，称为平稳随机过程。具有相应性质的随机场，称为平稳随机场。在平稳随机场中，$Z(x)$ 的数学期望 $E[Z(x)]$ 与其位置 x 无关。此时，可推导出空间变量 Z 在某方向上相距 h 处增量 $Z(x)-Z(x+h)$ 的方差：

$$\gamma(x, h) = Var[Z(x) - Z(x+h)] = E[Z(x) - Z(x+h)]^2, \tag{1}$$

称 $\gamma(x, h)$ 为变异函数，如果与位置 x 无关，只与距离 h 有关，此称随机场是二阶平稳的，$\gamma(x, h)$ 可记为 $\gamma(h)$。

当估计值 $\hat{Z}(x)$ 和实测值 $Z(x)$ 使下式成立时，

$$E[\hat{Z}(x) - Z(x)] = 0,\ x \in X, \tag{2}$$

称 $\hat{Z}(x)$ 是统计学意义上的无偏估计。

设 x_0 是被插值点，$\hat{Z}(x_0)$ 是估计值，将空间点集合 $\{x\}$ 具体化为 $\{x_1, x_2, \cdots, x_n\}$ 时，记：

$$\gamma_{ij} = \gamma(x_i, x_j) = E[Z(x_i) - Z(x_j)]^2,\ i, j = 0, 1, 2, \cdots, n。 \tag{3}$$

使用拉格朗日乘数法[14]求解式(2)，得一个求权向量 $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ 和拉格朗日乘数 ϕ 的线性方程组：

$$\begin{cases} \sum_{j=1}^{n} \lambda_i \gamma_{ij} - \phi = \gamma_{i0}, i = 1,2,\cdots,n; \\ \sum_{i=1}^{n} \lambda_i = 1。 \end{cases} \tag{4}$$

于是，二阶平稳假设下的估计值是：

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)。 \tag{5}$$

式(4)－式(5)称为普通克里金法。外观上看，克里金插值公式(5)与反距离加权法[10]的插值公式完全一样，都是加权求和，但权向量的来源完全不同。克里金法建立在随机过程理论上，而反距离加权法并不考虑随机性。对随机场性质和点之间距离影响程度的理解不同，衍生出了大量的克里金方法。

如果我们不单单考虑空间位置 x 对随机场的影响，还考虑了其它因素 y 的影响，但将影响的主次分开，就拓展为特殊的多变量模型[15]，称为协同克里金插值法。例如，水灾灾情，不仅仅是地理空间上的随机场，而且灾情还与地理单元内的GDP、地理单元距河流的距离、地理单元的坡度等众多因素有关。在水灾随机场中，理论上空间位置是主因素，GDP、河流距离和坡度等是次因素，但由于它们紧密的相关性，这些次因素能充分地体现空间位置的随机场属性。

以一个次因素 y 为例，不失一般性，设随机场 $Z(x)$ 的空间点集合 $X = \{x_1, x_2, \cdots, x_n\}$，次因素随机场 $Y(x')$ 的空间点集合 $X' = \{x'_1, x'_2, \cdots, x'_m\}$。当 $Z(x)$ 与 $Y(x')$ 存在相关性(例如，水灾与坡度相关)，且均二阶平稳时，令：

$$\begin{cases} C\{Z(x_i)Z(x_j)\} = E[Z(x_i) - Z(x_j)]^2,\ i, j = 1, 2, \cdots n; \\ C\{Z(x_i)y_j\} = E[Z(x_i) - Y(x'_j)]^2,\ i = 1, 2, \cdots n; \\ \qquad\qquad\qquad\qquad\qquad\qquad j = 1, 2, \cdots, m; \\ C\{y_i y_j\} = E[Y(x'_i) - Y(x'_j)]^2,\ i, j = 1, 2, \cdots m。 \end{cases} \tag{6}$$

求解式(7)的线性方程组：

$$\begin{cases} \sum_{j=1}^{n} \alpha_j C\{Z(x_i)Z(x_j)\} + \sum_{j=1}^{m} \beta_j C\{Z(x_i)y_j\} + \mu_1 = \\ \quad C\{Z(x_i)Z(x_0)\}, i = 1,2,\cdots,n; \\ \sum_{j=1}^{n} \alpha_j C\{y_i Z(x_j)\} + \sum_{j=1}^{m} \beta_j C\{y_i y_j\} + \mu_2 = \\ \quad C\{y_i Z(x_0)\}, i = 1,2,\cdots m; \\ \sum_{j=1}^{n} \alpha_j = 1; \\ \sum_{j=1}^{m} \beta_j = 0。 \end{cases} \tag{7}$$

被插值点 x_0 的估计值：

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \alpha_i Z(x_i) + \sum_{j=1}^{m} \beta_i Y(x'_j)。 \tag{8}$$

式(7)－式(8)称为 CK 插值法。$m = n$，$x'_i = x_i$，$i = 1, 2, \cdots n$，是一个特例。

虽然克里金法在地学多领域中有大量成功应用的例子，协同克里金法显著改了精度，ArcMap 等软件系统提供了工具模块，但是，通常只在很局部的空间，随机场才满足二阶平稳条件，且一系列弱化条件的改进，也只在很特殊的情况下才有效。因此，协同克里金法并非普适性的空间插值方法。对于空间异质性的问题，地理加权回归更适于空间插值。

1.2　地理加权回归

在统计学中，“回归”是指依据观测样本，对两个或更多变量之间关系性质进行描述，并用于推测。任一回归方法用于地理学问题的研究，都称为地理回归。考虑到样本在各局部空间的统计性质常有明显差异，人们提出了地理加权回归，改进了全局回归。

设 x_1，x_2，$\cdots$，x_m 和 y 分别是空间位置 (u, v) 观测到的自变量和因变量。对于从 n 个空间位置 (u_i, v_i)，$i = 1, 2, \cdots, n$，得到的样本，我们记为：

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} = \{(x_{11}, x_{12}, \cdots, x_{1m}, y_1), (x_{21}, x_{22}, \cdots, x_{2m}, y_2), \cdots, (x_{n1}, x_{n2}, \cdots, x_{nm}, y_n)\}。 \tag{9}$$

当随机变量 $\boldsymbol{x}$ 的总体是一个 $m+1$ 维正态分布时，地理加权回归模型是[16]：

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{m} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i,\ i = 1,2,\cdots,n。 \tag{10}$$

式中：$\beta_0(u_i, v_i)$ 是空间位置 (u_i, v_i) 处的截距系数，$\beta_j(u_i, v_i)$ 是 (u_i, v_i) 处第 j 个自变量的局部回归系数，ε_i 为随机误差项。

y 服从正态分布时，随机误差 ε 的期望值为 0。此时，对插值点(u_0，v_0)，从自变量 x_{01}，x_{02}，…，x_{0m}，推测因变量，只须根据(u_0，v_0)与(u_i，v_i)的远近程度，定义(u_i，v_i)与(u_0，v_0)适当的空间权重 w_i，用它们和 $\boldsymbol{X}$，计算适用于 y_0 的局部系数列矩阵：

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_j \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_m \end{pmatrix}。 \tag{11}$$

则地理加权回归的推测值由式(12)给出：

$$y_0 = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{0j}。 \tag{12}$$

本文中，我们采用式(13)的自适应双平方(Adaptive bi－square)公式[17]来定义 w_i：

$$w_i = \begin{cases} (1 - \dfrac{d_{0i}^2}{\max\limits_{1 \leqslant k \leqslant n} \{d_{0k}\}})^2, & d_{0i} \leqslant \max\limits_{1 \leqslant k \leqslant n} \{d_{0k}\}; \\ 0, & d_{0i} > \max\limits_{1 \leqslant k \leqslant n} \{d_{0k}\}; \end{cases} i = 1, 2, \cdots, n。 \tag{13}$$

式中：d_{0i} 为(u_0，v_0)与(u_i，v_i)之间的欧氏距离。令：

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}; \tag{14}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1m} \\ 1 & w_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n2} & \cdots & x_{nm} \end{pmatrix}; \tag{15}$$

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}。 \tag{16}$$

则局部系数列矩阵的计算公式为：

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Y}。 \tag{17}$$

式中：$\boldsymbol{X}^{\mathrm{T}}$ 是 $\boldsymbol{X}$ 的转置矩阵，$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}$ 是($\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}$)的逆矩阵。式(12)，式(13)和式(17)构成了自适应双平方 GWR 插值法。

人们曾用 Logistic 回归和泊松回归等来探讨改进地理加权回归[18]，试图超越线性回归的限制，但不过是从正态分布假设变为另一种假设而已，并不具有普适性。

1.3 回传神经网络

人工神经元网络能以任意精度逼近任何一个连续函数[19]，为改进地理学中的插值提供了一条新途径。

神经元网络是一个从 p 维实数空间 $\boldsymbol{R}^p$ 到 q 维数实空间 $\boldsymbol{R}^q$ 的一个映射 f：$\boldsymbol{R}^p \rightarrow \boldsymbol{R}^q$，并且定义为 $y = f(x) = \varphi(\boldsymbol{W}x)$，此处 $x \in \boldsymbol{R}^p$ 是输入矢量，$y \in \boldsymbol{R}^q$ 是输出矢量。$\boldsymbol{W}$ 是一个 $p \times q$ 权值矩阵，且 φ 是一个非线性函数，常称为激励函数。典型的激励函数是 S 形函数：

$$\varphi(x) = \frac{1}{1 + e^{-\alpha x}}, \ \alpha > 0。 \tag{18}$$

映射 f 可以分解为多个映射；结果是一个多层网络：$\boldsymbol{R}^p \rightarrow \boldsymbol{R}^m \rightarrow \cdots \rightarrow \boldsymbol{R}^n \rightarrow \boldsymbol{R}^q$。

计算 $\boldsymbol{W}$ 的运算法则是训练算法。最常用的神经网络之一是 BP－ANN，算法的基本思想是，学习过程由信号的正向传播与误差的反向传播两个过程组成。正向传播时，输入样本从输入层传入，经连接各神经元的初始权值矩阵 $\boldsymbol{W}_0$ 处理后，传向输出层。若输出层的实际输出与期望的输出不符，则转入误差的反向传播阶段。误差反传是将输出误差以某种形式通过隐层向输入层逐层反传，并将误差分摊给各层的所有单元，从而获得各层单元的误差信号，此误差信号即作为修正 $\boldsymbol{W}$ 的依据。周而复始地修正 $\boldsymbol{W}$，直到网络输出的误差减少到可接受的程度，或进行到预先设定的学习次数为止。这种方法也称为自适应模式识别[20]。BP－ANN 可视为是用最小期望平方误差作为条件期望函数的一致性估计。

虽然 ANN 是一个黑箱，但对训练样本不需要任何假设，拟合函数时将空间位置作为输入的一部分即可，避免纠结空间相关性，在地理学中有较好的适应性。例如，可以较大程度地避免生态质量评价时人为主观假定对预测结果的影响[21]，构建植被指数对气候因子响应的复杂关系时拟合优度较高[22]，用于细颗粒物($PM_{2.5}$)的估算时短期预测结果更加稳定[23]。然而，ANN 却很难成为地理学中通用的插值法，因为宏观数据中有太多的随机、非随机因素干扰，并不存在一个可以逼近的，理论上的连续函数。大多数情况下，训练样本中存在明显冲突，调整权值矩阵 $\boldsymbol{W}$ 无可适从，训练进入死循环，导致训练后的 ANN 预测精度并不高[24]。

地理空间上的信息扩散模型，不须对观测样本作任何人为假设，并且能较好地处理样本点之间的冲突矛盾，较好地解决了地理学中常用插值法存在的问题，能有效提高插值精度。

2 地理空间信息扩散技术

在现实中，插值是因为缺失需要的信息。换句话说，只有插值节点间的空白处，才需插值，而插值节点上并不需要插值。拟合并不等于插值。人们对插值模型进行的精度检验，通常是对拟合度的检验。上述 CK、GWR 和 BP－ANN 的背后，都是将最小二乘法施于节点处，进行拟合。显然，拟合结果用于推测时，效果都会与拟合点处不同。为了使模型更具说服力，一些研究人员用“训练样本”用来训练模型，留出“验证样本”来展示预测准确性。由于“训练样本”和“验证样本”的选择不同，检验结果很不同，而研究者声称的“随机选择”，

很难查实，结果仍然可疑。地理空间信息扩散技术，将插值节点处的信息，向插值点外扩散，直接面对插值需求构造模型。

通信工程中的“信息”，是消除随机不确定性的东西，只有波形的形式因素，没有内容因素，也没有价值因素。现代人工智能理论中，将信息分为客体信息和感知信息。前者是指客体所呈现的关于其自身的“状态及其变化方式”；后者是指主体从客体信息所感知的客体状态及其变化方式的形式、内容和效用[25]。地理空间信息扩散技术中，插值节点处的观测值，是人们已经感知到的信息；模型试图推测的，是插值点外的客体信息。

信息扩散，是将观测点的感知信息，扩散到非观测点，力图对非观测点有所认识。信息扩散，是人类用有限的知识，认识无限世界的本能。信息扩散不同于联想，并不是由于某人或某种事物而想起其他相关的人或事物；信息扩散，也不是信息传播，并非是个人、组织和团体通过符号和媒介交流信息。近年来，许多文献将“信息传播”称为“信息扩散”，旨在借用大量的数学工具，但内涵并没有变化。

地理空间信息扩散技术来自于优化处理小样本的信息扩散原理[26]：当我们用一个不完备数据估计一个关系时，一定存在合理的扩散方式可以将一个没有几何大小的观测值变为一个集值（例如，模糊集），以填充由不完备性造成的部分缺陷从而改进非扩散估计。该原理不仅在概率空间中成立，而且在几何空间中也成立[27]。这就意味着，我们可以将信息扩散技术，拓展到在地理空间上去，以填补地理单元上的数据空白，使不完整的空间数据集，变为完整的数据集。

然而，概率空间中的信息扩散方法，并不能直接用于地理空间，而须借助在观测点和非观测点都有的同类背景数据作为桥梁[2]，才能将观测点的感知信息，扩散到非观测点。为此，我们先界定两个基本的概念：“空白单元”和“背景数据”。

2.1　空白单元和背景数据

定义 1：设 g 和 o 是研究区域 G 中的两个地理单元。如果在识别 G 上的地表现象 F 时，g 被观测并被赋值，而 o 没有，则对于识别 F 而言，称 g 是一个被观测单元，o 是一个空白单元。

例如，在洪水灾区，灾情是一种临时的地表现象，已经被调查过灾情并获得数据的地理单元，是被观测单元；没有被调查过灾情的地理单元，是空白单元。

对地理单元 g 的观测值（或向量）w_g 称为一个已观测数据；对空白单元 o 的相应值（或向量）w_o，称为一个待观测或待推测数据。设 b_{g1}，b_{g2}，…，b_{gt}和 b_{o1}，b_{o2}，…，b_{ot}分别是 g 和 o 的 t 个同类地理特征的属性值。记向量 $\boldsymbol{b}_g=(b_{g1}, b_{g2}, \cdots, b_{gt})$，$\boldsymbol{b}_0=(b_{o1}, b_{o2}, \cdots, b_{ot})$。

定义 2：设 g_1，g_2，…，g_n 是 n 个被观测单元，o 是一个空白单元，它们的属性值向量集合是 $\boldsymbol{B}=\{\boldsymbol{b}_{g_1}, \boldsymbol{b}_{g_2}, \cdots, \boldsymbol{b}_{g_n}, \boldsymbol{b}_o\}$。如果能用 $\boldsymbol{B}$ 依据观测数据 w_{g_1}，w_{g_2}，…，w_{g_n}推测 w_o，称 $\boldsymbol{B}$ 为背景数据集，简称背景数据。

例如，用“人口”、“人均 GDP”和“相对暴露度”等数据，依据被观测单元的灾情，推测空白单元的灾情时，“人口”、“人均 GDP”和“相对暴露度”等就是背景数据。此时，空间位置已经在计算“相对暴露度”时发挥过作用[28]。

任何能用背景数据，由多个被观测单元的观测值，推测空白单元上相应值的方法，都具有将被观测单元的信息扩散到空白单元的功能。例如，CK、GWR 和 BP－ANN 等插值方法，都具有某种信息扩散的功能，但并不明显，因为这些模型的控制规则，不是为了填补空白，而是为了最佳拟合。

设研究区域 G 由 $n-q$ 个被观测单元 g_1，g_2，…，g_{n-q}，和 q 个空白单元 g_{n-q+1}，…，g_n 组成，即，

$$\boldsymbol{G}=\{g_1, g_2, \cdots, g_{n-q}, g_{n-q+1}, \cdots, g_n\}。\tag{19}$$

设背景数据由 t 个地理特征的属性值组成。记地理单元 g_i 第 j 个特征的属性值为 b_{ij}，$i=1, 2, \cdots, n$；$j=1, 2, \cdots, t$。将 w_{g_i}简记为 w_i，$i=1, 2, \cdots, n$，于是，关于 $\boldsymbol{G}$ 上的信息可由表 1 示之。

表 1　研究区域 $\boldsymbol{G}$ 上的观测值和背景数据

地理单元	背景数据				观测值
g_1	b_{11}	b_{12}	…	b_{1t}	w_1
g_2	b_{21}	b_{22}	…	b_{2t}	w_2
…	…	…	…	…	…
g_{n-q}	$b_{n-q,1}$	$b_{n-q,2}$	…	$b_{n-q,t}$	w_{n-q}
g_{n-q+1}	$b_{n-q+1,1}$	$b_{n-q+1,2}$	…	$b_{n-q+1,t}$	Unknown
…	…	…	…	…	…
g_n	b_{n1}	b_{n2}	…	b_{nt}	Unknown

以背景数据 b_{ij}为桥梁，在地理空间 $\boldsymbol{G}$ 上进行信息扩散的方法，由构建因果关系矩阵和模糊近似推理两部分组成[2]。

2.2　用背景数据和已观测数据构建因果关系矩阵

令 $\tau=n-q$，$\lambda=t+1$，我们从表 1 中得到容量为 τ 的 λ 维样本 $\boldsymbol{X}$：

$$\boldsymbol{X}=\{(x_{i1}, x_{i2}, \cdots, x_{i\lambda-1}, x_{i\lambda}) \mid i=1, 2, \cdots, \tau\}。\tag{20}$$

式中：$x_{i1}=b_{i1}$，$x_{i2}=b_{i2}$，…，$x_{i\lambda-1}=b_{it}$，$x_{i\lambda}=w_i$，$i=1, 2, \cdots, \tau$。

设 U_j，$j=1, 2, \cdots, t$，是用于扩散背景数据中第 j 个地理特征属性值的监控空间，而 U_{t+1}是用于扩散已观测数据的监控空间。令 λ 维笛卡尔空间：

$$\boldsymbol{U}=U_1\times U_2\cdots\times U_\lambda。\tag{21}$$

式中：$U_j=\{u_{j1}, u_{j2}, \cdots, u_{jm_j}\}$，$j=1, 2, \cdots, \lambda$。理论上，对不同的分量 j，监控点的个数 m_j 可以不同，但由于监控点的密度达到一定程度后，用不同的 m_j 并不影响插值的精度，因此，我们取一个

m 作为所有分量监控空间中监控点的个数。

对于任意一个样本点，

$$\boldsymbol{x}_i = (x_{i1},\ x_{i2},\ \cdots,\ x_{i\lambda}) \in \boldsymbol{X}, \tag{22}$$

和任意一个监控点，

$$\boldsymbol{u} = (u_{1k_1},\ u_{2k_2},\ \cdots,\ u_{\lambda k_\lambda}) \in \boldsymbol{U},\ k_j \in \{1,\ 2,\ \cdots,\ m\},\ j = 1,\ 2,\ \cdots,\ \lambda。 \tag{23}$$

我们用式(24)的 λ 维初级扩散公式，将 $\boldsymbol{x}$ 的信息扩散到 $\boldsymbol{u}$。

$$\mu(\boldsymbol{x}_i, \boldsymbol{u}) = \prod_{j=1}^{\lambda} \exp[-\frac{(x_{ij} - u_{jk_j})^2}{2h_j^2}]。 \tag{24}$$

根据表 1 中的背景数据和已观测数据，分别用式(25)计算扩散系数 h_j，$j = 1,\ 2,\ \cdots,\ \lambda$。

$$h_j = \begin{cases} 0.814\ 6(b-a), & \tau = 5; \\ 0.569\ 0(b-a), & \tau = 6; \\ 0.456\ 0(b-a), & \tau = 7; \\ 0.386\ 0(b-a), & \tau = 8; \\ 0.336\ 2(b-a), & \tau = 9; \\ 0.298\ 6(b-a), & \tau = 10; \\ 2.685\ 1(b-a)/(\tau-1), & \tau \geqslant 11。 \end{cases} \tag{25}$$

式中：$b = \max\limits_{1 \leqslant i \leqslant \tau} \{x_{ij}\}$，$a = \min\limits_{1 \leqslant i \leqslant \tau} \{x_{ij}\}$。

令：

$$S_{k_1k_2\cdots k_\lambda} = \sum_{i=1}^{\tau} \prod_{j=1}^{\lambda} \exp[-\frac{(x_{ij} - u_{jk_j})^2}{2h_j^2}]。 \tag{26}$$

此数值表征了样本 $\boldsymbol{X}$ 在监控点 $\boldsymbol{u}$ 处的密集程度，可用于改进扩散模型，得到适应性扩散模型[29]：

$$Q_{k_1k_2\cdots k_\lambda} = (1 - \frac{S_{k_1k_2\cdots k_\lambda}}{\tau}) \sum_{i=1}^{\tau} \prod_{j=1}^{\lambda} \exp[-\frac{(x_{ij} - u_{jk_j})^2}{2h_j^2}]。 \tag{27}$$

于是，我们获得了一个 $U_1 \times U_2 \cdots \times U_\lambda$ 上的，关于 $\boldsymbol{X}$ 的信息矩阵：

$$\boldsymbol{Q} = \{Q_{k_1k_2\cdots k_{\lambda-1}k_\lambda}\}_{m \times m \times \cdots \times m}。 \tag{28}$$

$\forall j \in \{1,\ 2,\ \cdots,\ \lambda\}$，$k_j \in \{1,\ 2,\ \cdots,\ m\}$，令：

$$H_{kj} = \max_{\substack{1 \leqslant k_t \leqslant m \\ t \neq j}} \{Q_{k_1k_2\cdots k_\lambda}\}, \tag{29}$$

和

$$r^{(j)}_{k_1k_2\cdots k_\lambda} = \frac{Q_{k_1k_2\cdots k_\lambda}}{H_{k_j}}。 \tag{30}$$

此为针对分量 j 的归一化信息矩阵中的元素，此矩阵记为：

$$\boldsymbol{R}_j = \{r^{(j)}_{k_1k_2\cdots k_\lambda}\}_{\underbrace{m \times m \times \cdots \times m}_{\lambda 个}}。 \tag{31}$$

我们可由 X 构造出一个背景数据与观测数据之间的因果关系：

$$\boldsymbol{R} = \boldsymbol{R}_1 \wedge \boldsymbol{R}_2 \wedge \cdots \boldsymbol{R}_\lambda。 \tag{32}$$

此因果关系矩阵中的元素为：

$$r_{k_1k_2\cdots k_\lambda} = r^{(1)}_{k_1k_2\cdots k_\lambda} \wedge r^{(2)}_{k_1k_2\cdots k_\lambda} \wedge \cdots \wedge r^{(\lambda)}_{k_1k_2\cdots k_\lambda}。 \tag{33}$$

式(29)－式(32)的关系矩阵生成法，来自于模糊蕴含理论，适用于由小样本生成，离散性较大的原始信息矩阵 $\boldsymbol{Q}$(式(28))。如果样本较大，$\boldsymbol{Q}$ 的元素值呈现出一定的统计规律，可直接将 $\boldsymbol{R}_\lambda$ 作为关系矩阵使用。对所有归一化信息矩阵进行的取小运算，具有滤波的作用，也会丢失少量的统计信息。

2.3 用背景数据推测空白单元中未知数据

设 $\boldsymbol{b} = (b_1,\ b_2,\ \cdots,\ b_t)$ 为表 1 中一个空白单元的背景数据，$\lambda - 1$ 维笛卡尔空间 $U_1 \times U_2 \times \cdots U_{\lambda-1}$ 中的一个点记为 $\boldsymbol{u}_{\lambda-1} = (u_{1k_1},\ u_{2k_2},\ \cdots,\ u_{\lambda-1k_{\lambda-1}})$。用式(24)中用到过的扩散系数 h，由式(34)将 $\boldsymbol{b}$ 变为论域 $U_1 \times U_2 \cdots U_{\lambda-1}$ 上的一个模糊集，并用式(35)进行归一化处理。

$$\mu(\boldsymbol{b}, \boldsymbol{u}_{\lambda-1}) = \prod_{j=1}^{\lambda-1} \exp[-\frac{b_j - u_{jk_j}}{2h_j^2}]。 \tag{34}$$

$$\begin{cases} a_{k_1k_2\cdots k_{\lambda-1}} = \dfrac{q_{k_1k_2\cdots k_{\lambda-1}}}{s}; \\ s = \max\limits_{\substack{1 \leqslant k_j \leqslant m \\ 1 \leqslant j \leqslant \lambda-1}} \{q_{k_1k_2}\cdots k_{\lambda-1}\}; \\ q_{k_1k_2\cdots k_{\lambda-1}} = \prod\limits_{j=2}^{t} \exp[-\dfrac{(b_j - u_{jk_j})^2}{2h_j^2}]。 \end{cases} \tag{35}$$

此模糊集记 $\tilde{A}$ 为，即：

$$\tilde{A} = \mu_A(\boldsymbol{u}_{\lambda-1}) = \frac{a_{11\cdots1}}{u_{11\cdots1}} + \frac{a_{11\cdots2}}{u_{11\cdots2}} + \cdots + \frac{a_{k_1k_2\cdots k_{\lambda-1}}}{u_{k_1k_2\cdots k_{\lambda-1}}} + \cdots + \frac{a_{mm\cdots m}}{u_{mm\cdots m}}。 \tag{36}$$

式(36)为模糊集的扎德记法，并非分数求和。以 $\tilde{A}$ 为输入，使用近似推理公式(37)，由式(32)的因果关系矩阵，我们可以得到一个具有隶属函数 $\mu_C(u_{\lambda k_\lambda})$ 的模糊输出 $\tilde{C}$。

$$\mu_C(u_{\lambda k_\lambda}) = \max_{\substack{1 \leqslant k_j \leqslant m \\ 1 \leqslant j \leqslant \lambda-1}} \min\{a_{k_1k_2\cdots k_{\lambda-1}},\ r_{k_1k_2\cdots k_{\lambda-1}k_\lambda}\}。 \tag{37}$$

最后，使用式(38)的信息集中法[32]，我们获得了一个分明值 w：

$$w = \frac{\sum\limits_{k_\lambda=1}^{m} \mu_C(u_{\lambda k_\lambda}) \mu_C(u_{\lambda k_\lambda}) u_{\lambda k_\lambda}}{\sum\limits_{k_\lambda=1}^{m} \mu_C(u_{\lambda k_\lambda}) \mu_C(u_{\lambda k_\lambda})}。 \tag{38}$$

当 $\boldsymbol{R}_\lambda$ 可以作为关系矩阵使用时，用重心法[8]替代信息集中法，精度更高。

由公式(24)－式(38)组成的模型，称为自学习离散回归(Self－Learning Discrete Regression，SLDR)模型，是一种地理空间信息扩散技术。式(38)中的 w 是使用此技术，由空白单元的背景数据 $\boldsymbol{b}$ 和从被观测单元学习到的因果关系 $\boldsymbol{R}$，对空白单元的插值。

一个由“人口”、“人均 GDP”和“洪水相对暴露度”推测“洪水损失”的计算机仿真实验证明，在拟合精度上，SLDR 模型明显优于 GWR 和 BP－ANN，误差分别降低了 60% 和 33% 左右[8]。对空白地理单元“洪水损失”的推测，SLDR 和 BP－ANN 通过了平均基准误差小于平均预测误差的检验，证明了 SLDR 和 BP－ANN 插值的效性，而 GWR 无效[30]。此检验中，基准误差是指，给定样本除去一个测试点后模型的均方根误差；预测误差是指，测试点的实际值与估计值之间的误差。

样本中的每一个点均担任一次测试点任务，形成的平均误差用于检测插值的有效性，避免了使用主观“验证样本”存在的问题。

本文将以 2018 年和 2020 年发生在四川省三台县的两次大洪水的水灾灾情为例，实证研究地理空间信息扩散技术的可靠性，为从理论走向实践，进行必要的探索。

3　研究区概况

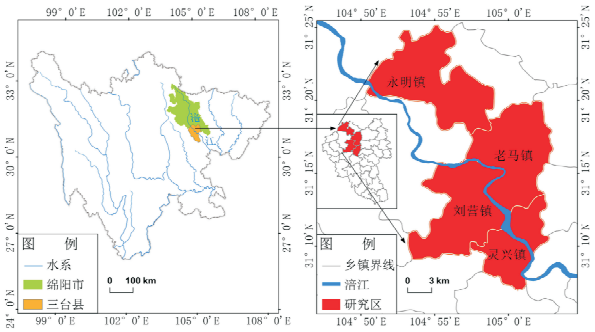
我国三分之二以上的国土面积受到洪涝灾害威胁，主要分布在长江、黄河、淮河、海河、珠江、松花江、辽河 7 大江河下游和东南沿海地区。这些大区域的水灾，相邻的较小地理单元上，灾情的同质性很高，只有进行大范围的调研，获得的数据才能支撑水灾插值模型的研究。为此，我们选用小范围内差异较大的四川省三台县涪江流域麦冬主产区作为实证研究区域。由于发生在当地的洪水具有一走一过的特点，涝灾并不明显，所以本文只研究洪水灾害的插值问题。

四川省绵阳市三台县位于四川盆地中部偏北，30°42′34″～31°26′35″N，104°43′04″～105°18′13″E；属于亚热带季风气候区，年平均降水量为 876.2 mm，降水在年内和年际变化大，年降水集中在夏秋两季，其中 6—9 月降水量占全年降水量的 72.4%；境内地质构造简单，全部由褶皱构造组成，无地质断层，海拔高度 307 m 至 672 m。属川中丘陵地区，地势北高南低。县境内大小江河溪流 46 条，均属于长江支流嘉陵江水系，其中涪江、凯江、梓江、郪江为四条大江。涪江由绵阳市涪城区丰谷镇进入三台县境内，经永明、芦溪、老马、刘营、里程、灵兴、新德、潼川、北坝出境至射洪县香山镇。

三台县历来易受洪水灾害影响。据历史资料记载，从唐贞观十八年（644）到民国三十八年的 1 300年中，三台发生严重的暴雨洪涝灾害计 38 次，其中有 19 次县城被淹。1949 年 10 月新中国成立后，共计出现洪水灾害 31 次，其中特大洪灾 6 次。截至 2018 年，近 30% 的涪江沿岸地段还暴露在无堤防状态下。

三台县幅员面积 2 659 km^2，丘陵面积占 94.39%，2021 年辖 31 个镇、2 个乡，总人口 139.12 万，其中农业人口 123 万。2020 年，生产总值 407.45 亿元，经济发展程度较高，是我国最大的生猪产地县。三台县享有“中国麦冬之乡”的美誉，麦冬产业带覆盖了芦溪镇、永明镇、老马镇、建设镇、刘营镇、灵兴镇、新德镇等乡镇，气候、湿度、土壤均适合麦冬生长，有 500 多年种植麦冬的历史，其“涪城麦冬”居全国麦冬之上品，目前三台全县常年种植麦冬面积达 3 333 hm^2，年均产量 1.2 万 t，占全国的 70% 以上，麦冬出口量占全国的 80% 以上。

2018 年 7 月和 2020 年 8 月，三台县发生了大洪水，涪江流域的永明镇、老马镇、刘营镇、灵兴镇受灾尤其严重，当地民众对灾情记忆犹新，为此，我们选择了这四个镇作为实证研究区域，其地理位置由图 1 所示。



(a) 四川省水系图　　(b) 涪江流域中的研究区

图 1　实证研究区域的地理位置

（基于自然资源部标准地图服务网站审图号为 GS(2019)1821 号的标准地图制作，底图无修改）

4　近年两次大洪水概况

4.1　2018 年“7·11”特大洪水

2018 年 7 月 9－11 日涪江流域上游县市区的强降雨和局地的大暴雨使得涪江、凯江、梓江、魏城河、郪江遭受了建国以来最大洪峰的洗劫。尤其是三台县涪江、凯江水位极速大幅上涨，流量均为历史最大峰值。

洪水期间，永明镇和花园镇（2019 年划归芦溪镇）等 40 个镇乡受灾，受灾人口达 25.1 万人，实施紧急转移安置 21 562 人，集中安置 1 612 人、分散安置 19 950 人，由于沿江的镇乡党政对于群众的疏散转移有效及时，无人员死亡。

涪江流域“7.11”特大洪水，导致三台县境内的道路、堤防、水库、渠系、电力、供水、通信、能源等基础设施毁损严重。洪水冲垮了 2 000 m 多的土堤造成决堤，有 50 km 多的基础设施需维修加固或重建，有 3 条县内公路中断，3 座大桥临时交通安全管制。全县水利工程 2 324 处受损，直接损失 4.2 亿元。江河干流堤防决口 9 处、损坏工程护岸 145 处，有 14 座水库管涌产生新的病险。2 条电力干线受损，导致 19 个镇乡突然停电。基础设施毁损 5 亿余元。其中，芦溪工业区殷家壕堤防、花园镇涪城村护堤、里程镇回龙村堤防和刘营镇下渡口堤防瞬间决堤导致洪水灾情最为惨重。芦溪工业区的大量厂房被淹，物料、机器、设备被洪水浸泡，损毁惨重，造成 24 户重点工业企业毁损、停产，直接工业损失达 6.3 亿元。

此次洪水共造成社会经济损失近 18 亿元，其中，农林水产受灾 19 442 hm^2，其中绝收 2 077 hm^2，农田（含鱼塘）毁损 198 hm^2，农业直接损失3.8 亿元。

4.2　2020 年“8·12”大洪水

2020 年 8 月 11—12 日，涪江流域普降大到暴

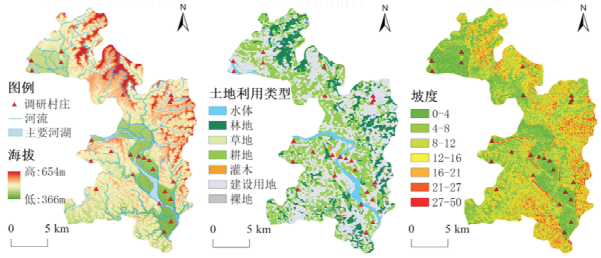
雨，上游的安州、北川、平武局地降下特大暴雨，洪灾压力加之疫情防控的重担，为三台县带来了70 年来最为严峻的大考。

三台县于 8 月 11 日启动并迅速提高至Ⅲ级防汛预警响应，16 日启动 II 级防汛应急响应。期间，涪江中下游超保证水位 1.8 m，13 000 m^3/s 的洪峰冲击导致明台库区尾段的新德镇马脊防洪堤基脚被洪水掏空 100 m 左右，出现了 560 m 迎水面“垮方险情”，省县部门紧急加固抢修，最终保证了洪峰顺利过境。

由于人员转移安置及时，抗洪抢险行动到位，全县此次洪灾并无人员伤亡情况。但极端降水重创了交通基础设施，造成三台县境内道路路基垮塌、山体塌方严重，出现 1 433 处灾毁险情，其中：国道 16 处、省道 199 处、县道 75 处、乡道 193 处、村道 950 处。中立路永明镇涪建村段受灾最为严重，车辆、群众出行受阻。因灾损毁道路于当月月底全部抢通。水路设施方面，共计受损 5 个渡口以及两岸码头。

5 灾区背景数据和灾情数据

为了研究三台县的洪水和地震灾害综合风险，2017 年北京师范大学与三台县合作建立了“安全科学与工程”教学实践基地，2018 年和 2020 年，分别对“7·11”特大洪水和“8·12”大洪水进行了一些调研。2021 年 6 月 17—20 日期间，本文作者前往研究区，对研究区的 25 个村庄进行了野外考察和入户调查，获得了第一手资料。根据调研村庄的海拔与水文特征(图 2a)、土地利用情况(图 2b)及坡度计算结果(图 2c)，经过整理和分析，我们得到了对地理空间信息扩散技术进行实证研究所需的背景数据(表 2)和灾情数据(表 3)。每一个地理单元 g 获得 3 个背景数据“与河流距离”、“GDP”和“坡度”，其中，“坡度”(b_{g3})，是用拟合曲面法[31]由 ArcGIS 平台计算而得。



(a) 海拔与水文特征 (b) 土地利用情况 (c) 坡度分析结果

图 2 调研区域水灾背景数据分析资料及调研村庄的地理位置

注：海拔来源于 ASTER GDEM V2 全球高程数据；水系数据由全国 1∶25 万地理信息数据库与县水利局河流水系平面图整理得到；土地利用资料来源于 Esri 提供的 2020 年 10 m 分辨率土地利用数据(https：//www. geoscene. cn/)。

表 2 研究区 25 个村庄的背景数据

乡镇	村名	编号	与河流距离/m	GDP/(万元/km^2)	坡度*/(°)
永明镇	长江村	1	159.15	299	3.09
	金翔村	2	57.61	301	4.07
	涪建村	3	405.89	336	11.98
	光辉场社区	4	38.36	338	7.07
	永和村	5	71.64	299	3.37
	团缘村	6	27.45	334	4.38
刘营镇	石鼓坝村	7	197.93	304	6.52
	大围坝村	8	333.33	337	3.51
	三道河村	9	251.91	252	8.02
	安宁村	10	520.75	337	6.19
	马家村	11	1 323.07	303	10.16
	龙沟村	12	141.37	311	6.63
	土地村	13	36.49	302	3.68
灵兴镇	木鱼村	14	179.22	307	2.73
	灵峰村	15	47.48	276	14.96
	花庙村	16	90.20	307	7.41
	石桥村	17	251.38	367	2.68
	争胜村	18	1 001.76	342	2.76
	灵兴村	19	449.09	348	4.24
老马镇	会龙村	20	146.24	305	3.88
	柳林子村	21	18.32	338	5.93
	木林村	22	247.61	302	4.65
	莲花村	23	247.99	338	8.49
	里程村	24	47.87	305	4.40
	回龙村	25	264.92	365	2.84

*300 m×300 m 范围内，以 30 m 分辨率像元高程值数据，用 ArcGIS 计算平均坡度，单位为角度。参见：https：//help. arcgis. com/zh－cn/arcgisdesktop/10.0/help/index. html#/na/009z000000vz000000。

表 3　两次大洪水的灾情数据

地理单元	2018 年“7・11”特大洪水			2020 年“8・12”大洪水		
编号	房屋损失/元	农业损失/元	庄稼被淹/hm^2	房屋损失/元	农业损失/元	庄稼被淹/ hm^2
1	0.00	52 500.00	0.13	0.00	22 000.00	0.06
2	4 000.00	20 000.00	0.09	0.00	0.00	0.00
3	1 666.67	4 333.33	0.23	10 000.00	14 500.00	0.52
4	911.11	444.44	0.06	2 855.56	444.44	0.02
5	11 100.00	9 583.33	0.23	29 000.00	11 012.50	0.31
6	6 683.33	4 816.67	0.19	47 600.00	14 218.75	0.16
7	63 333.33	119 666.67	0.24	0.00	0.00	0.00
8	110 000.00	105 000.00	0.30	10 000.00	160 000.00	0.33
9	0.00	1 000.00	0.27	0.00	1 300.00	0.08
10	0.00	27 500.00	0.10	0.00	16 500.00	0.05
11	0.00	3 666.67	0.16	0.00	4 333.33	0.29
12	200.00	3 000.00	0.29	233.33	3 000.00	0.09
13	0.00	0.00	0.03	1 000.00	3 000.00	0.20
14	0.00	14 000.00	0.09	0.00	7 666.67	0.10
15	750.00	2 500.00	0.13	750.00	1 600.00	0.17
16	0.00	875.00	0.18	375.00	1 750.00	0.07
17	0.00	2 600.00	0.00	0.00	6 800.00	0.09
18	0.00	1 233.33	0.08	0.00	866.67	0.05
19	27 500.00	24 250.00	0.05	0.00	2 500.00	0.02
20	4 000.00	2 033.33	0.14	4 600.00	0.00	0.09
21	250.00	4 650.00	0.15	2 750.00	5 025.00	0.18
22	0.00	750.00	0.27	0.00	1 716.67	0.13
23	0.00	0.00	0.00	0.00	375.00	0.03
24	6 666.67	45 555.56	0.26	2 888.89	34 888.89	0.13
25	61 666.67	126 666.67	0.43	833.33	69 000.00	0.40

6　用信息扩散技术推测灾情

我们以 2018 年“7・11”特大洪水中农业损失为例，演示如何用信息扩散技术构建因果关系矩阵，由背景数据推测灾情。然后，通过模型对全部 3 种灾情的预测误差分析，说明其插值是有效的。

6.1　构建因果关系矩阵

由表 2 中的背景数据和表 3 中的第 3 列数据，我们得到容量为 $\tau=25$，维度 $\lambda=4$ 的样本 $\boldsymbol{X}$：

$$\boldsymbol{X}=\{(x_{i1},\ x_{i2},\ x_{i3},\ x_{i4})\mid i=1,\ 2,\ \cdots,\ 25\}=\{(159.15,\ 299,\ 3.09,\ 52\ 500),\ (57.61,\ 301,\ 4.07,\ 20\ 000),\ \cdots,\ (264.92,\ 365,\ 2.84,\ 126\ 666.67)\}。\tag{39}$$

根据表 2 中河流距离、GDP、坡度和表 3 中农业损失的最大值和最小值，并依据样本容量大小，我们以等步长方式，各取 20 点构成它们的监控空间，即：

$$\begin{cases} U_1=\{u_{11},\ u_{12},\ \cdots,\ u_{1,20}\}=\{18.32,\ 86.99,\ \cdots,\ 1\ 323.07\};\\ U_2=\{u_{21},\ u_{22},\ \cdots,\ u_{2,20}\}=\{252,\ 258.05,\ \cdots,\ 367\};\\ U_3=\{u_{31},\ u_{32},\ \cdots,\ u_{3,20}\}=\{2.68,\ 3.33,\ \cdots,\ 14.96\};\\ U_4=\{u_{41},\ u_{42},\ \cdots,\ u_{4,20}\}=\{0,\ 6\ 666.67,\ \cdots,\ 126\ 666.65\}。\end{cases}\tag{40}$$

我们以式(39)中的第 1 个样本点

$$\boldsymbol{x}_1=\{x_{11},\ x_{12},\ x_{13},\ x_{14}\}=(159.15,\ 299,\ 3.09,\ 52\ 500)\tag{41}$$

为例，演示如何将其信息扩散给 4 维笛卡尔空间 $U_1\times U_2\times U_3\times U_4$ 中，与其距离较近的两个点：

$\boldsymbol{u}_{18407}=(u_{13},\ u_{27},\ u_{31},\ u_{47})=(155.66,\ 288.32,\ 2.68,\ 40\ 000)$；

和

$\boldsymbol{u}_{18408}=(u_{13},\ u_{27},\ u_{31},\ u_{48})=(155.66,\ 288.32,\ 2.68,\ 46\ 666.67)$。

笛卡尔空间点的编号，是按矩阵元素的序号排法所得。首先，由式(25)处理式(39)的样本数据，可得各分量的扩散系数 h_1，h_2，h_3，h_4 分别是 145.974，12.866，1.374 和 14 171.362。于是，

$$\begin{aligned}\mu(\boldsymbol{x}_1,\boldsymbol{u}_{18407})&=\prod_{j=1}^{4}\exp\left[-\frac{(x_{1j}-u_{jk_j})^2}{2h_j^2}\right]\\&=\exp\left[-\frac{(x_{11}-u_{13})^2}{2h_1^2}-\frac{(x_{12}-u_{27})^2}{2h_2^2}-\frac{(x_{13}-u_{31})^2}{2h_3^2}-\frac{(x_{14}-u_{47})^2}{2h_4^2}\right]\\&=1.000\times0.708\times0.956\times0.678\\&=0.459;\end{aligned}\tag{42}$$

$$\mu(\boldsymbol{x}_1,\boldsymbol{u}_{17408}) = \prod_{j=1}^{4}\exp[-\frac{(x_{1j}-u_{jk_j})^2}{2h_j^2}]$$
$$= \exp[-\frac{(x_{11}-u_{13})^2}{2h_1^2}-\frac{(x_{12}-u_{27})^2}{2h_2^2}-\frac{(x_{13}-u_{31})^2}{2h_3^2}-\frac{(x_{14}-u_{48})^2}{2h_4^2}]$$
$$= 1.000\times 0.708\times 0.956\times 0.919$$
$$= 0.622; \quad (43)$$

将式(39)中的所有25个样本点，在$U_1\times U_2\times U_3\times U_4$上完成初级扩散并累加后，我们得到一个初级信息分矩阵$\boldsymbol{S}=\{S_{k_1k_2\cdots k_4}\}_{20\times 20\times 20\times 20}$，例如$\boldsymbol{u}_{18407}$和$\boldsymbol{u}_{18408}$上获得的初级信息扩散总量分别是$S_{3,7,1,7}=0.841$，$S_{3,7,1,8}=0.868$。由式(37)进行适应性扩散，我们得到一个原始信息矩阵$\boldsymbol{Q}=\{Q_{k_1k_2\cdots k_4}\}_{20\times 20\times 20\times 20}$例如$\boldsymbol{u}_{18407}$和$\boldsymbol{u}_{18408}$上获得的适应性扩散总量分别是$Q_{3,7,1,7}=0.813$，$Q_{3,7,1,8}=0.838$。对$\boldsymbol{X}$的信息矩阵$\boldsymbol{Q}$，我们从第1分量到第4分量，分别进行归一化处理，得到相应的归一化信息矩阵。例如，对第4分量，即“农业损失”，$k_4=7$和$k_4=8$时，我们分别有：

$$H_7=\max_{\substack{1\leqslant k_t\leqslant 20\\1\leqslant t\leqslant 3}}\{Q_{k_1k_2k_37}\}=1.739,\ H_8=\max_{\substack{1\leqslant k_t\leqslant 20\\1\leqslant t\leqslant 3}}\{Q_{k_1k_2k_38}\}=1.637。\quad (44)$$

于是，

$$r^{(4)}_{3,7,1,7}=\frac{Q_{3,7,1,7}}{H_7}=\frac{0.813}{1.739}=0.467,$$
$$r^{(4)}_{3,7,1,8}=\frac{Q_{3,7,1,8}}{H_8}=\frac{0.838}{1.637}=0.512。\quad (45)$$

由式(32)对4个归一化信息矩阵进行“取小”运算，可得该地区此次洪水事件中，得到“农业损失”与“与河流距离”“GDP”、“坡度”因果关系的一个估计R。例如，在此因果型关系矩阵中我们有：

$$r_{3,7,1,7}=r^{(1)}_{3,7,1,7}\wedge r^{(2)}_{3,7,1,7}\wedge r^{(3)}_{3,7,1,7}\wedge r^{(4)}_{3,7,1,7}\wedge=0.236\wedge 0.384\wedge 0.285\wedge 0.467=0.236;\quad (46)$$

$$r_{3,7,1,8}=r^{(1)}_{3,7,1,8}\wedge r^{(2)}_{3,7,1,8}\wedge r^{(3)}_{3,7,1,8}\wedge r^{(4)}_{3,7,1,8}\wedge=0.244\wedge 0.396\wedge 0.294\wedge 0.512=0.244。\quad (47)$$

比较这两个元素的值可知，输入$\boldsymbol{u}=(u_{13},u_{27},u_{31})$时，“农业损失”是$u_{47}$(=40 000元)的可能性比$u_{48}$(=46 666.67元)的小。

6.2 用背景数据推测灾情

我们以背景数据$\boldsymbol{b}=(b_1,b_2,b_3)=(159.15,299,3.09)$为例，用6.1节中构建的因果关系矩阵，推测农业损失。选用的背景数据是式(41)中$\boldsymbol{x}_1$的前3个分量。推测的是长江村2018年“7·11”特大洪水中的农业损失。

首先，我们用6.1节中的扩散系数h_1，h_2，h_3，由式(42)将$\boldsymbol{b}$变为论域$U_1\times U_2\times U_3$上的一个模糊集$\tilde{B}$。

$$q_{ijk}=\exp[-\frac{(b_1-u_{1i})^2}{2h_1^2}-\frac{(b_2-u_{2j})^2}{2h_2^2}-\frac{(b_3-u_{3k})^2}{2h_3^2}];\quad (48)$$

$$\tilde{B}=\mu_B(\boldsymbol{u})=\frac{q_{111}}{u_{111}}+\frac{q_{112}}{u_{112}}+\cdots+\frac{q_{k_1k_2k_3}}{u_{k_1k_2k_3}}+\cdots+\frac{q_{20,20,20}}{u_{20,20,20}}。\quad (49)$$

式中：$u_{k_1k_2k_3}$是$U_1\times U_2\times U_3$中点$(u_{1k_1},u_{2k_2},u_{3k_3})$的简写。例如：

$$q_{371}=\exp[-\frac{(x_{11}-u_{13})^2}{2h_1^2}-\frac{(x_{12}-u_{27})^2}{2h_2^2}-\frac{(x_{23}-u_{31})^2}{2h_3^2}]$$
$$=1.000\times 0.708\times 0.956$$
$$=0.677。\quad (50)$$

用q的最大值0.979对$\tilde{B}$进行归一化处理，得模糊输入$\tilde{A}$：

$$\tilde{A}=\frac{a_{111}}{u_{111}}+\frac{a_{112}}{u_{112}}+\cdots+\frac{a_{k_1k_2k_3}}{u_{k_1k_2k_3}}+\cdots+\frac{a_{20,20,20}}{u_{20,20,20}}。\quad (51)$$

例如$a_{371}=0.677/0.979=0.692$。使用近似推理公式(37)，我们得到模糊输出：

$$\tilde{C}=\frac{c_{41}}{u_{41}}+\frac{c_{42}}{u_{42}}+\cdots+\frac{c_{4,20}}{u_{4,20}}$$
$$=\frac{0.885}{0}+\frac{0.952}{666.67}+\cdots+\frac{0.161}{12\,666.65}。\quad (52)$$

需注意上式并非分数求和，而是模糊集的扎德记法。该模糊输出表达的是：损失为0元，6 666.67元，…，126 666.65元的可能性分别是0.885，0.952，…，0.161。使用式(38)对此模糊集进行信息集中处理，我们得到由背景数据(159.15，299，3.09)推测的农业损失是：

$$w=\frac{\sum_{k=1}^{20}\mu_C^2(u_{4k})u_{4k}}{\sum_{k=1}^{20}\mu_C^2(u_{\lambda k_\lambda})}$$
$$=\frac{c_{41}^2u_{41}+c_{41}^2u_{41}+\cdots+c_{41}^2u_{41}}{c_{41}^2+c_{41}^2+\cdots+c_{41}^2}$$
$$=\frac{0.885^2\times 0+0.952^2\times 6\,666.67+\cdots+0.161^2\times 126\,666.65}{0.885^2+0.952^2+\cdots+0.161^2}$$
$$=22\,053.27(\text{元})。\quad (53)$$

6.3 预测误差分析

由背景数据(159.15，299，3.09)，用地理空间信息扩散技术的SLDR模型推测的，长江村2018年“7·11”特大洪水中的农业损失，是22 053.27元，与表2中观测值52 500有较大的出入，这是由于25个样本点的灾情标准差高达37 448.98所致。通常，我们用均方根误差，来比较两个模型拟合插值节点的误差。但拟合得很好的模型，不一定适合于节点以外的插值。只有节点以外的预测误差，才能体现插值精度[30]。

为了区分样本点中的自变量和因变量，我们将式(20)中的样本改写为

$$X=\{(x_{i1},x_{i2},\cdots,x_{it},y_i)\,|\,i=1,2,\cdots,\tau\}。\quad (54)$$

式中：$x_{it}=x_{i\lambda-1}$，$y_i=x_{i\lambda}$。对因变量y_i的估计值记为$\hat{y}_i$，均方根误差定义为：

$$\sigma=\sqrt{\frac{1}{\tau}\sum_{i=1}^{\tau}(y_i-\hat{y})^2}。\quad (55)$$

令：

$$\mathrm{E}=\{1,2,\cdots,\tau\}。\quad (56)$$

$\forall\eta\in E$，令：

$$\begin{cases} X_{L_\eta} = \{(x_{i1}, x_{i2}, \cdots, x_{it}, y_i) \mid i = 1, 2, \cdots, \eta - 1, \\ \qquad \eta + 1, \cdots, \tau\}; \\ X_\eta = \{(x_{\eta 1}, x_{\eta 2}, \cdots, x_{\eta t}, y_\eta)\}。 \end{cases} \tag{57}$$

称 X_{L_η} 为训练样本(有 $\tau - 1$ 个样本点)，称 X_η 为测试样本(只有一个样本点)。显然 $\boldsymbol{X} = X_{L_\eta} \cup X_\eta$。由 X_{L_η} 训练模型 f，其均方根误差称为 f 的一个基准误差，记为 σ_{L_η}；f 对 y_η 的预测误差，称为 f 的一个预测误差，记为 e_η，即

$$\begin{cases} \sigma_{L_\eta} = \sqrt{\dfrac{1}{\sigma} \sum\limits_{i \neq \eta} (y_i - \hat{y}_i)^2}; \\ e_\eta = |y_\eta - \hat{y}_\eta|。 \end{cases} \tag{58}$$

显然，对给定的样本 $\boldsymbol{X}$，σ_{L_η} 和 e_η 均受到 η 的影响。只有它们各自的平均值，即式(59)中的平均基准误差 $\overline{\sigma}_L$ 和平均预测误差 $\overline{e}$，才是较好的指标。在不引起混乱的情况下，下文中我们将 $\overline{\sigma}_L$ 和 $\overline{e}$ 简称为基准误差和预测误差

$$\begin{cases} \overline{\sigma}_L = \dfrac{1}{\tau} \sum\limits_{\eta=1}^{\tau} \sigma_{L_\eta}; \\ \overline{e} = \dfrac{1}{\tau} \sum\limits_{\eta=1}^{\tau} e_\eta。 \end{cases} \tag{59}$$

对表 2 和表 3 给出的数据，SLDR 模型的均方根误差、基准误差和预测误差见表 4。

根据文献[30]的研究，一个模型的预测是否有效，须通过两个准则来检验。

准则 I：平均基准误差必须小于平均预测误差，确保模型能从此样本中总结出规律。

准则 II：平均预测误差较小，确保模型的精度。

如果基准误差大于预测误差，就相当于说，你游历了欧洲而不是非洲，但是你对非洲的描述比欧洲更准确，这显然是荒谬的。如果基准误差明显大于预测误差，说明模型不符合逻辑，对给定样本的学习无效；或者说，使用的模型与给定的样本不匹配。表 4 中，3 个案例的基准误差小于预测误差，另 3 个案例的预测误差没有明显小于基准误差，说明 SLDR 用于学习相应 6 个样本是有效的，具有普适性。至于 SLDR 的预测误差是否较小，须同别的模型进行比较，才能显现出来。

7　与 CK、GWR 和 BP – ANN 模型的比较

分别用 CK、GWR 和 BP – ANN 模型对表 2 和表 3 给出的数据进行处理，所得结果列入表 5、表 6 和表 7。由表 4 比较这三个表可知，就本文的实例而言，只有 SLDR 模型能够通过预测有效性两个准则的检验。CK 模型的预测误差均明显小于基准误差，通过不了准则 I 的检验，插值无效。在 5 个案例中，GWR 模型的预测误差，均明显小于基准误差，也不合逻辑，插值无效。

表 4　自学习离散回归模型(SLDR)均方根误差 σ、基准误差 $\overline{\sigma}_L$ 和预测误差 $\overline{e}$

误差类型	2018 年“7·11”特大洪水			2020 年“8·12”大洪水		
	房屋损失	农业损失	庄稼被淹	房屋损失	农业损失	庄稼被淹
σ	24 271.70	32 816.91	0.091 903	10 054.90	29 911.46	0.092 738
$\overline{\sigma}_L$	24 408.62	32 929.26	0.091 841	10 086.38	30 094.87	0.093 329
$\overline{e}$	21 932.63	32 940.00	0.102 610	8 704.01	24 730.58	0.123 040

表 5　协同克里金模型(CK)均方根误差 σ、平均基准误差 $\overline{\sigma}_L$ 和平均预测误差 $\overline{e}$

误差类型	2018 年“7·11”特大洪水			2020 年“8·12”大洪水		
	房屋损失	农业损失	庄稼被淹	房屋损失	农业损失	庄稼被淹
σ	23 267.15	35 084.15	0.089 362	7 145.79	35 489.58	0.131 085
$\overline{\sigma}_L$	23 583.44	35 240.50	0.089 883	7 387.49	35 216.73	0.130 869
$\overline{e}$	15 547.34	24 115.46	0.064 794	4 210.72	20 083.27	0.106 381

表 6　地理加权回归模型(GWR)均方根误差 σ、平均基准误差 $\overline{\sigma}_L$ 和平均预测误差 $\overline{e}$

误差类型	2018 年“7·11”特大洪水			2020 年“8·12”大洪水		
	房屋损失	农业损失	庄稼被淹	房屋损失	农业损失	庄稼被淹
σ	25 090.09	35 653.23	0.103 160	10 314.18	31 182.00	0.123 738
$\overline{\sigma}_L$	24 967.76	35 511.85	0.102 688	10 247.70	30 908.53	0.123 073
$\overline{e}$	19 147.92	29 750.96	0.099 673	7 713.39	21 665.96	0.118 536

表 7　回传神经网络模型(BP – ANN)均方根误差 σ、平均基准误差 $\overline{\sigma}_L$ 和平均预测误差 $\overline{e}$

误差类型	2018 年“7·11”特大洪水			2020 年“8·12”大洪水		
	房屋损失	农业损失	庄稼被淹	房屋损失	农业损失	庄稼被淹
σ	4 165.71	6 678.87	0.020 714	2 615.23	7 622.64	0.032 035
$\overline{\sigma}_L$	4 464.21	12 645.68	0.061 728	2 923.63	7 708.43	0.051 851
$\overline{e}$	60 358.09	82 681.68	0.184 168	17 063.72	60 621.64	0.249 070

表7 中，BP – ANN 采用 3 × 9 × 1 拓扑结构，学习系数 0.9，惯性系数 0.7，系统误差 0.000 9。例如，用 2018 年“7·11”特大洪水时 25 个村庄的背景数据和农业损失组成的样本，训练出的神经网络，拟合的均方根误差是 6 678.87 元(见表 7 第 3 行第 3 列)。而从 25 个村庄中随机地取 24 个的数据组成样本训练网络时，一些样本的训练进入死循环，拟合的均方根误差较大；一些样本能顺利完成训练，均方根误差较小，平均基准误差为 12 645.68 元(见表 7 第 4 行第 3 列)。用 24 个村庄的数据训练出的网络对没有参加训练的村庄进行插值(预测)时，平均预测误差高达 82 681.68 元(见表 7 第 5 行第 3 列)。

对所有 6 个案例，BP – ANN 模型的基准误差均小于预测误差，通过了准则 I 的检验，但没有通过准则 Ⅱ 的检验，因为其预测误差远远大于 SLDR、CK 和 GWR 模型，精度太低，是一个无效的预测模型。这一现象说明，能够高度拟合训练样本的回传神经网络模型，并不适用于复杂地表现象的插值。

8 结论和讨论

由于成本和时效等诸多原因，用插值来完善地理空间数据，具有重要意义。在满足相应条件的情况下，许多插值模型都可以使用。但是，常用的插值模型，都不具有普适性。

虽然内插式的数学插值模型精度很高，但只适用于空间连续场；克里金法和地理加权回归等地统计方法考虑到了空间数据的随机性，但只适用于有大样本支撑的插值；回传神经网络模型能够高度拟合训练样本，但插值精度可能并不高。

模型的拟合精度高，并不等于插值精度也高。插值是因为缺失需要的信息，只有插值节点间的空白处，才需插值，节点上拟合并不是插值。插值是对空白处有关数值的预测。因此，一个模型是否可通过某样本的训练有效地进行插值，可通过两个准则来检验，一是平均基准误差必须小于平均预测误差，确保模型能从此样本中总结出规律；二是平均预测误差较小，确保模型的精度。

本文以 2018 年和 2020 年发生在四川省三台县的两次大洪水，造成 25 个村的房屋损失、农业损失和庄稼被淹等三类水灾灾情数据组成的 6 个案例，实证了地理空间信息扩散技术能通过两个准则的检验，是普适性插值模型。协同克里金模型在所有案例中，都没有通过准则 I 的检验，不合逻辑，说明插值无效；地理加权回归模型在 5 个案例中没有通过准则 I 的检验。虽然回传神经网络模型通过了准则 I 的检验，且基准误差很小，但预测误差却比基准误差高出近一个数量级，也比自学习离散回归模型、协同克里金模型和地理加权回归模型的预测误差都大很多，没有通过准则 Ⅱ 的检验。这说明，回传神经网络模型并不适用于复杂地表现象的插值。

信息扩散的自学习离散回归模型，是一种以离散数学表达的数学模型，能充分发挥现代计算机大容量存储、高速度运行的功能，具有某种人工智能的属性，如果能在扩散方式和近似推理等方面进一步完善，有望为地理空间插值提供一个重要的工具。

参考文献：

[1] 王晓青，丁香，王龙，等. 四川汶川 8 级大地震灾害损失快速评估研究[J]. 地震学报，2009，31(2)：205 – 211.

[2] HUANG C F. Geospatial information diffusion technology supporting by background data[J]. Journal of Risk Analysis and Crisis Response，2019，9(1)：2 – 10.

[3] 安基文，徐敬海，聂高众，等. 高精度承灾体数据支撑的地震灾情快速评估[J]. 地震地质，2015，37(4)：1225 – 1241.

[4] 黄崇福，田雯，王润东. 在救灾智联网中推测信息孤岛救助需求强度的空间信息扩散模型[J]. 自然灾害学报，2021，30(2)：1 – 13.

[5] CHEN J，LI W，CHEN W K，et al. Assessment of earthquake prevention and disaster reduction capability of county – level administrative units in Gansu Province[J]. Journal of Risk Analysis and Crisis Response，2018，8(3)：177 – 183.

[6] CHENG X F，SUN H H，YUAN Z，et al. Flood disaster risk assessment and spatial distribution characteristics along the Yangtze River in Anhui Province[J]. Journal of Risk Analysis and Crisis Response，2014，4(4)：238 – 242.

[7] DING Y，ZHANG M. Research on the development of county finance in Guizhou Province in the promotion of precise poverty alleviation[J]. Journal of Risk Analysis and Crisis Response，2018，8(1)：52 – 60.

[8] HUANG C F. Geospatial information diffusion based on self – learning discrete regression[J]. Journal of Environmental Informatics，2021，38(2)：93 – 105.

[9] ELDRANDALY K A，ABU – ZAID M S. Comparison of six GIS – based spatial interpolation methods for estimating air temperature in western Saudi Arabia[J]. Journal of Environmental Informatics，2011，18(1)：38 – 45.

[10] NISTOR M M，HARIANTO R，SATYANAGA A，et al. Investigation of groundwater table distribution using borehole piezometer data interpolation：Case study of Singapore[J]. Engineering Geology，2020，271：105590.

[11] SOUZA T J，MEDEIROS J，GONCALVES A C，et al. A method to identify an accidental control rod drop with inverse distance weighted interpolation[J]. Annals of Nuclear Energy，2020，145：107462.

[12] MATHERON G. Principles of geostatistics[J]. Economic geology，1963，58(8)：1246 – 1266.

[13] LE N D，ZIDEK J V. Statistical Analysis of Environmental Space – Time Processes[M]. New York：Springer，2006，p. 101 – 134.

[14] BERTSEKAS D P. Constrained Optimization and Lagrange Multiplier Methods[M]. New York：Academic，1982：231 – 256.

[15] BELKHIRI L，TIRI A，MOUNI L. Spatial distribution of the groundwater quality using kriging and Co – kriging interpolations[J]. Groundwater for Sustainable Development，2020(11)：100473.

[16] BRUNSDON C，FOTHERINGHAM A S，CHARLTON M E. Geographically weighted regression：a method for exploring spatial nonstationarity[J]. Geographical Analysis，1996，28(4)：281 – 298.

[17] NILSSON P. Natural amenities in urban space – A geographically weighted regression approach[J]. Landscape and Urban Planning，2014，121：45 – 54.

[18] FOTHERINGHAM A S，BRUNSDON C，CHARLTON M. Geographically Weighted Regression：The Analysis of Spatially Varying Relationships[M]. Chichester：John Wiley & Sons，2002：190 – 193.

[19] HORNIK K，STINCHCOMBE M，WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks，1989，2(5)：359 – 366.

[20] PAO Y H. Adaptive Pattern Recognition and Neural Networks

[M]. Reading MA: Addison - Wesley, 1989: 1 - 309.
[21] 刘焱序，王仰麟，彭建，等. 耦合恢复力的林区土地生态适宜性评价——以吉林省汪清县为例[J]. 地理学报，2015，70(3)：476 - 487.
[22] 唐见，曹慧群，陈进. 生态保护工程和气候变化对长江源区植被变化的影响量化[J]. 地理学报，2019，74(1)：76 - 86.
[23] 刘基伟，闵素芹，金梦迪. 基于分布式感知深度神经网络的高分辨率 $PM_{2.5}$ 值估算[J]. 地理学报，2021，76(1)：191 - 205.
[24] HUANG C F, LEUNG Y. Estimating the relationship between isoseismal area and earthquake magnitude by hybrid fuzzy - neural - network method[J]. Fuzzy Sets and Systems, 1999, 107(2): 131 - 146.
[25] 钟义信. 机制主义人工智能理论——一种通用的人工智能理论[J]. 智能系统学报，2018，13(1)：2 - 18.
[26] HUANG C F. Principle of information diffusion[J]. Fuzzy Sets and Systems, 1997, 91(1): 69 - 90.
[27] MAKO Z. Approximation with diffusion - neural - network[C]// G Horvath. Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence. Budapest: Budapest Tech - Hungarian Fuzzy Association, 2005: 589 - 600.
[28] 黄崇福. 自然灾害风险相对暴露度的研究[C]//黄崇福. 第六届中国西部风险分析与风险管理学术研讨会论文集. 巴黎：亚特兰蒂斯出版社，2019：1 - 5.
[29] 黄崇福. 一种评价台风风险模型可靠性的计算机仿真方法[J]. 自然灾害学报，2020，29(5)：24 - 35.
[30] HUANG C F. Two judging criteria to check validity of a model for filling gaps caused by incomplete geospatial data. Environmental Research[J], 2020, 186: 109401.
[31] 刘晓，赵荣，梁勇，等. 顾及地貌与 DEM 分辨率的坡度算法适应性研究[J]. 测绘科学，2017，42(3)：29 - 34.
[32] 王家鼎，黄崇福. 模糊信息处理中的信息扩散方法及其应用[J]. 西北大学学报(自然科学版)，1992，22(4)：383 - 392.

Empirical Research on Geospatial Information Diffusion Technique ——Taking Flood Disaster in Santai County, Sichuan Province, as an Example

HUANG Chongfu[1,2] and ZHANG Xinwen[2]

(1. *Key Laboratory of Environmental Change and Natural Disaster, Ministry of Education, Beijing Normal University, Beijing* 100875, *China*; 2. *Academy of Disaster Risk Sciences, Faculty of Geographical Science, Beijing Normal University, Beijing* 100875, *China*)

Abstract: Interpolation is an important approach to infer the earth surface phenomena where the geospatial data is incomplete. Under the corresponding condition, classical methods such as collaborative Kriging interpolation (CK), geographically weighted regression (GWR) and back propagation artificial neural network (BP - ANN) all have good performance. However, these methods are not applicable enough, especially in actual geographical survey, conditions of which are difficult to be satisfied. With fewer observation units and larger data dispersion, the best interpolation is based on the information diffusion technique, called Self - Learning Discrete Regression (SLDR). In June 2021, we visit and collect data from 25 villages in Santai County of Sichuan Province, tabulate the losses of each village during two flood events in 2018 and 2020. Choosing the distance from river, GDP and slope as the independent variable, the property damage, agricultural losses, crop inundated area separately as the dependent variable, 6 empirical cases prove the general applicability of the SLDR interpolation with two criteria based on the datum error and forecasting error. The forecasting error by SLDR is low, and there is no case showing the datum error higher than the forecasting error. In all cases of the CK, the forecasting error is less than the datum error, which illogically indicates that after learning, the prediction error is larger than the unlearned. The CK method is judged to be invalid, and the GWR model gets similar results in five cases. Although datum error of the BP - ANN is very small, its forecasting error is about an order of magnitude larger than the datum error, much larger than the other three models, which suggests that the BP - ANN, with a high ability to fit training samples, is not suitable for interpolating the complex surface phenomena on the earth.

Key words: spatial interpolation; information diffusion; collaborative kriging; geographically weighted regression; neural network; datum error; forecasting error; Santai County