Journal of Risk Analysis and Crisis Response
Vol. 9(1), March (2019), pp. 2–10
DOI: https://doi.org/10.2991/jracr.b.190328.001; eISSN: 2210-8505, ISSN: 2210-8491
https://www.atlantis-press.com/journals/jracr

Geospatial Information Diffusion Technology Supporting by Background Data

Chongfu Huang[1,2,3]

[1]*Key Laboratory of Environmental Change and Natural Disaster, Ministry of Education, Beijing Normal University, Beijing 100875, China*
[2]*State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China*
[3]*Faculty of Geographical Science, Academy of Disaster Reduction and Emergency Management, Beijing Normal University, Beijing 100875, China*

Received December 20, 2018
Accepted January 10, 2019

Abstract

In this paper, we express the initial concept of geospatial information diffusion supporting by background data, which plays a role as a bridge to diffuse the information carried by the observations, obtained from observed units, to gap units. The self-learning discrete regression, based on the multivariate normal diffusion, is suggested to supplement incomplete geospatial data to be complete. The suggested method has obvious advantages over the geographic weighted regression and the artificial neural network for inferring the observations in gap units

Keywords: geographic unit, background data, information diffusion, normal diffusion, self-learning discrete regression

借助背景数据的地理空间信息扩散技术*

黄崇福 [1,2,3]

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，中国
2. 北京师范大学，地表过程与资源生态国家重点实验室，北京 100875，中国
3. 北京师范大学，地理科学学部，减灾与应急管理研究院，北京 100875，中国

摘要：本文提出了以背景数据为桥梁，将已观测数据携带的信息，扩散到空白单元的信息扩散基本思想，并给出了基于多元正态扩散的自学习离散回归模型，从而实现了将地理空间上的不完整数据补充为完整数据。本文的方法，在空白地理单元数据的推测方面，比地理加权回归和人工神经网络，有明显的优势。

关键词：地理单元；背景数据；信息扩散；正态扩散；自学习离散回归

1. 引言

在全球环境变化驱动下，人们对陆地表层系统中各种现象的研究，须使用更精细的观测数据测度更具体的研究对象[1]。然而，由于资源或时间的制约，为研究某一问题，人们获得的数据，常常不完整。例如，2008 年汶川大地震[2]发生一天后，救援人员对通信中断的受灾地区，仍然不知其灾情如何。灾情信息的不完整，严重制约着灾害救助的时效性与精准性。为解决此问题，人们统计回归历史灾害数据，得出经验公

*本研究由国家重点研发计划（编号：2017YFC1502902）和国家自然科学基金项目（编号：41671502）资助。

式，一旦发生破坏性地震，就可根据震级对灾情进行粗估，也称为快速评估[3]。这种经验性的“隔空判灾”，能将灾情的估计精度控制在数量级误差之内，就已经很不错。精细一点，人们可借助地理信息系统（GIS）中使用的数学插值模型[4]，依据在一些点上采集到的灾情，推测其它点上的灾情。问题是，数学插值要求数据具有连续性，而除了温度场以外，大多数地表数据都不满足连续性这一条件。因此，数学插值的结果，往往不尽人意。

由于概率空间中的信息扩散方法[5]，能在小样本条件下显著提高概率分布的估计精度，文[6]尝试着将此方法发展到地理空间上，将不完整数据补充为完整数据。本文继续此项工作，完善重要概念的定义，规范化如何借助背景数据进行地理空间信息扩散。

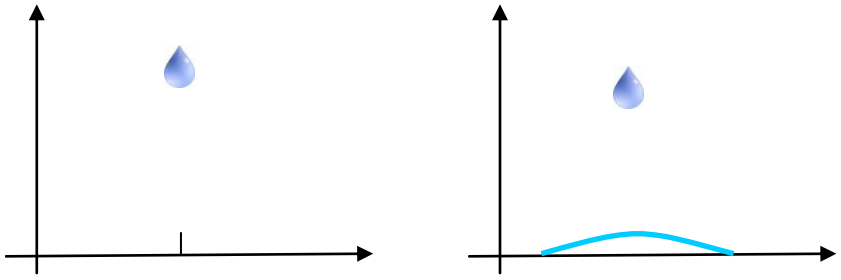
2. 概率空间中的信息扩散方法

让我们用一个简单的例子来说明信息扩散的初衷。当我们对所研究系统进行一次观测时，假定我们得到了一个数据。我们可以将此数据看作是天上掉下来的一滴水(图 1(*a*))，这滴水落在地面的什么位置，有一定的随机性。只有统计较多的水滴，我们才可能知道一段时间内落下水滴在地面的分布。但如果统计水滴落下后的影响面积(图 1(*b*))，统计较少的水滴，就能知道分布。

这种集值化后能提高概率分布估计精度的现象，被称为信息扩散原理[7]：设 X 是从概率密度函数为 $p(x)$ 的总体中随机抽取的一个样本，$\hat{p}(x)$ 是用 X 对 $p(x)$ 的一个非扩散估计。如果 X 不完备，则一定存在一个基于 X 的扩散估计 $\tilde{p}(x)$，使得 $\tilde{p}(x)$ 比 $\hat{p}(x)$ 更精确。

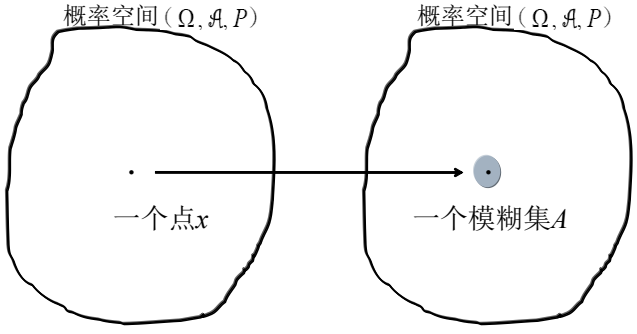
估计概率密度函数时，将 X 中各样本点 x 携带的信息在观测空间中的扩散，其实是一种在概率空间中的扩散。本文将传统的信息扩散方法，称为“概率空间中的信息扩散方法”。最常用的有线性信息分配法和正态信息扩散法[8]。用正态扩散函数 μ 将样本点 x 的信息在概率空间中进行扩散，如图 2 所示。

目前，概率空间中信息扩散理论的基础比较稳固，应用涉及面较广，尤其在风险分析方面。例如，信息扩散方法被用于分析太湖蓝藻暴发风险，为政府决策提供了重要依据[9]；用于研究美国东海岸飓风风险，计算出保守风险值和冒险风险值[10]；用于研究中国最近 20 年的水灾数据，计算出不同可能性的多值风险



(*a*) 观测值是 x　　(*b*) x 被变成一个模糊集 A

图 1. 信息扩散类似于水滴影响面积。统计少量水滴的影响面积，就能估计水滴的分布。



$$A:\mu(x,u)=\frac{1}{h\sqrt{2\pi}}\exp[-\frac{(x-u)^2}{2h^2}],x\in X,u\in\Omega$$

图 2. 用正态扩散函数将样本点信息在概率空间中进行扩散

[11]；用于研究中国北方草原火灾风险，为畜牧业生产制定补偿计划提供了依据[12]；用于研究洞庭湖区洪水、干旱、虫害和鼠害，绘制出粮食生产的自然灾害风险图[13]；用正态扩散处理我国水利部数据库中的水灾数据，生成可变模糊集，拟合出了水灾风险曲线[14]；用于北塞浦路斯旅游保险气候指数计算，绘制了风险图[15]；将信息扩散模型嵌入土壤流失方程，评估环渤海地区不同降雨情况下的土壤侵蚀风险[16]；用于分析液化天然气站过去 12 个月的运行数据，能及时发现潜在的风险[17]；用于研究草原生物灾害，绘制出我国北方 10 省区的风险图[18]；用于研究农业保险业务数据，计算不同损失的可能性，为政府财政支持农险提供依据[19]；用于广东省北江、西江和绥江三江汇流区的洪水风险评估，能根据洪峰水位推断水灾面积的几种可能性[20]；用于广州南沙区黄阁镇和南沙镇的区域环境风险评估，帮助当地政府优化工业区布局，建立风险防范管理程序[21]。

文[22]用解析几何学的方式证明，信息扩散原理，不仅在概率空间中成立，而且在几何空间中也成立（见图 3）。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。

3. 地理单元上的不完整数据

当我们对一个研究区域上的某种地表现象展开研究时，该区域中各地理单元上的地理位置、人口等信息很容易得到，但有些信息却不易得到。例如，当我们对某县的自然灾害风险进行研究时，一些乡镇的历史灾害资料就难以收集到。又例如，当重大自然灾害发生后，灾害管理部门很难在 1 小内得到灾区所有乡镇的灾情信息。依据不完整灾情制定的应急救灾方案，直接影响灾害救助效果。

设 G 是一个研究区域，其上的某一现象 F，例如地震风险，是被研究的对象。不失一般性，假定 G 由 n 个地理单元 $g_1, g_2, ..., g_n$ 组成，即，

$$G=\{ g_1, g_2, ..., g_n \} \tag{1}$$

更进一步地，我们假设，通过观测 G 上所有的地理单元与 F 有关的数据（或向量），可以识别出现象 F。对地理单元 g_i 进行观测得到的数据，记为 w_i。

例如，当我们研究由 129 个县级地理单位组成的云南省“地震风险”（Earthquake risk）现象时，该被研究的现象可写为 $F=E_{\text{risk}}$。研究区域 G 是云南省（Yunnan），记为：

$$G_{\text{Yunnan}}=\{g_1, g_2, ..., g_{129}\}$$

式中，g_1=昆明市五华区，g_2=昆明市盘龙区，...,g_{129}=临沧市双江拉祜族佤族布朗族傣族自治县。

为了识别 G_{Yunnan} 上的 E_{risk}，我们必须(观测)知道每个单元的“地震危险性[23]”和“地震易损性[24]”。例如，对玉溪市通海县，即地理单元 g_{27}，我们必须知道下面向量中每个分量的具体数值：

w_{27}=（通海县地震危险性，通海县地震易损性）

如果我们能获得研究区域 G 上识别 F 所需的所有地理单元上的值，则称该研究区域上的数据对于识别 F 是完整的，否则称 W 为不完整。据此，本文给出“数据不完整”的形式化定义如下：

定义 1：设 G 由地理单元 $g_1, g_2, ..., g_n$ 组成，对 g_i 的观测记为 w_i, i=1, 2, …, n。假定 G 上的地表现象 F 可用这些地理单元上的观测数据集合

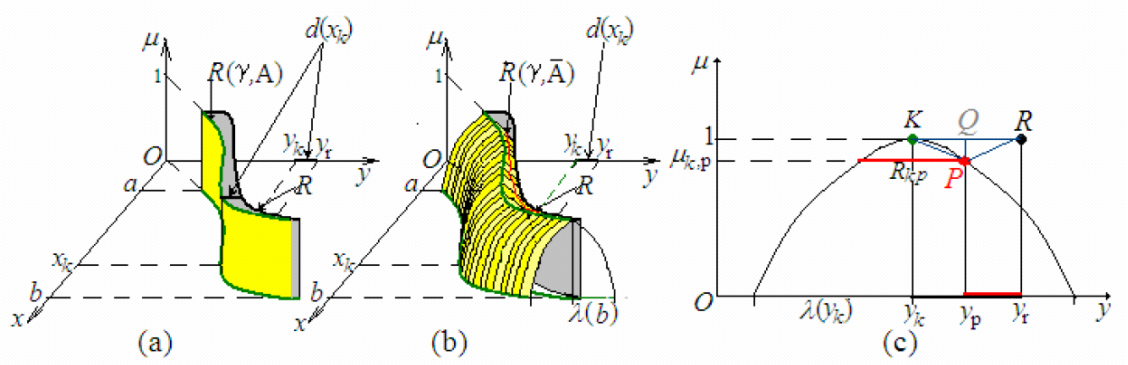


图 3. 信息扩散原理的解析几何表述。(a) 用算子γ和数据集 A 估计 x 和 y 间的关系 R（真实关系，假定为右侧曲线），得一个估计 $R(\gamma, A)$ （假定为左侧曲线）的所有元素（$R(\gamma, A)$中的点）的隶属度值是 1。（b）将 A 中的所有数据用信息扩散函数变为模糊集，生成数据集 $\overline{A}$ 。用算子γ和数据集 $\overline{A}$ ，得对 R 的另一个估计 $R(\gamma, \overline{A})$，其所有元素都是拟三角形模糊数。（c）$R(\gamma, \overline{A})$在 Oy 的投影。R_{kp} 表示 $R(\gamma, \overline{A})$的水平集$\mu_{k,p}$。一定存在 R_{kp} 上的一个点 $P(x_p, y_p)$，使 R 和 P 间的距离小于 R 和 K 间的距离。

$$W=\{w_1, w_2, \ldots, w_n\}, \quad (2)$$

来识别。当 W 的所有元素均被赋值时，称 W 是在 G 上识别 F 的完整数据，否则称为**不完整数据**。

例如，图 4 的研究区域是一次大洪水造成的灾区，一些地理单元上的死亡人数、重伤人数和灾民人数没有统计上来，灾区的灾情数据不完整。此时，我们难以识别出研究区域上的地表现象“灾情”。

显然，数据完整与否，与所研究的地表现象和区域大小有关。越是复杂或精细的地表现象，越是容易出现数据不完整；研究区域涉及的地理单元越多，要获得完整的数据越是困难。

4. 地理空间上的信息扩散

当研究区域上的数据不完整时，有两条途径或许能补齐数据，使之完整。一条途径是追加调研；另一条途径是用数学模型进行插值。由于时间或成本的制约，追加调研的途径常常走不通；由于有一定跨度的相邻地理单元上的大多数地理特征的属性值并不连续，插值法生成的数据，并无价值。

人们自然想到，是否可以用概率空间中的信息扩散方法，将地理空间上已经得到的数据，扩散到缺少数据的地理单元上，得到虽然与客观数据相比会有所偏差，但却完整且有价值的数据。

显然，概率空间中的信息扩散方法，并不能直接用于地理空间上。事实上，概率空间中的信息扩散，是从样本 X 到到以样本空间 Ω 为论域的模糊幂集 $\mathcal{P}(\Omega)$ 上的一个数学映射 μ：

$$\begin{aligned} \mu:\ & X \to \mathcal{P}(\Omega) \\ & x \mapsto \mu_x(u), u \in \Omega \end{aligned} \quad (3)$$

式中的 μ 称为信息扩散函数。式(4)是人们常用的正态信息扩散函数，将样本点 x 携带的量值为 1 的信息，由扩散系数 h 控制，按正态分布的形式，在样本空间 Ω 中进行扩散。

$$\mu(x,u)=\frac{1}{h\sqrt{2\pi}}\exp[-\frac{(x-u)^2}{2h^2}],\ x \in X, u \in \Omega \quad (4)$$

只有当 X 是小样本（少于 30 个样本点），不足以用其估计其所来自总体的分布时，为提高估计精度，对 X 进行扩散处理，才有意义。样本空间 Ω 中的点 u 从样本点 x 能扩散得到多少信息量，由 u 与 x 之间的距离决定。

式(3)中，被扩散的 x，本身就是概率空间中的一个点，所以称为概率空间“中”的信息扩散。当我们将式(2)中对地理单元 g 的观测值（或向量）w 视为一个随机抽样时，当然也可以在 w 的概率空间中进行信息扩散，但当 w 不是经纬度时，w 就不是地理空间上的点，w 携带信息就不是在地理空间上被扩散。

我们需要通过某种媒介的帮助，才能将 w 携带的信息，在地理空间上进行信息扩散。为此，我们界定三个基本的概念：“空白单元”、“媒介”和“背景数据”。

定义 2：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**。

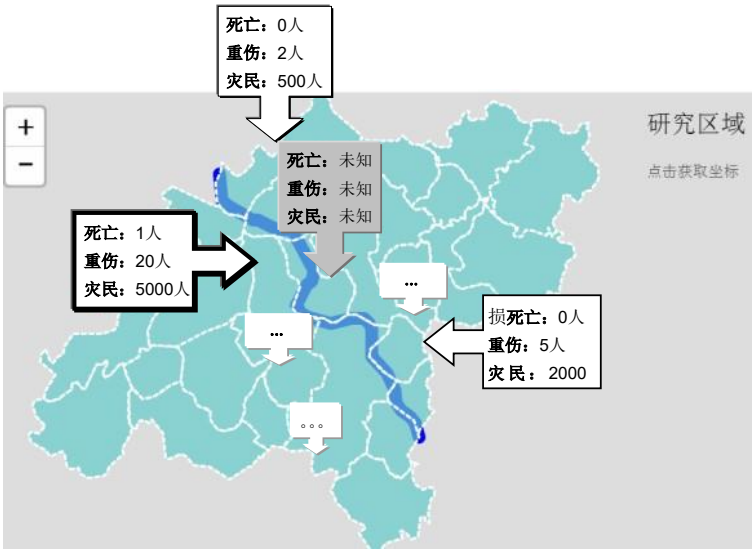


图 4. 研究区域上的数据不完整，难以对总体“灾情”做出判断。

例如，在洪水灾区，对识别总体灾情而言，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 称为一个**已观测数据**；对空白单元 o 的相应值，称为一个**待观测数据**。

定义 3：设 o 是一个空白单元。如果能用数据 ω，依据一些已观测数据，对 o 赋值，则称 ω 是对空白单元进行估值的**媒介**。

例如，根据 Tobler 的第一地理定律，“每一事物都与其他事物有关。与近处事物的关系，比与遥远的事物关系密切”[25]，GIS 分析师经常使用反距离加权（IDW）插值来填补数据空白[26]。在 IDW 法中，x_o 为点 o 处的估计值，x_i 为已知点 i 处的 x 值，d_i 为点 i 和点 o 之间的距离。借助于一系列距离 d_1，d_2，...，d_n，使用一系列已知值 x_1，x_2，...，x_n 估计 x_o。IDW 法中的 d_1，d_2，...，d_n 就是媒介。

什么样的媒介，能帮助我们实现信息在地理空间上的扩散呢，我们自然想到了地理特征的属性值。

在认识论中，事物本身具备的性质，称为特性，通常是内在的，独有的，例如，熔点是晶体的特性。人们根据某种事物所共有的特性抽象出的某一概念，称为特征，一般从表面上就看得出来，例如，生命是生物的特征。

地球自然创造的地理特征，称为自然地理特征，例如，气候、地形、土壤、水文、日照时间、无霜期等。人类社会创造的地理特征，称为人文地理特征，例如，人口、交通、经济、宗教、国家政策等。

量化描述地理特征的数值，称为地理特征的属性值，例如，100 万是“人口”特征的一个属性值。所要讨论属性值涉及的范围，叫做该属性值的论域。下面，我们用地理特征的属性值来定义“背景数据”。

定义 4：设 $g_1, g_2, ..., g_n$ 是 n 个被观测单元，o 是一个空白单元，记

$$G=\{ g_1, g_2, ..., g_n, o\}. \tag{5}$$

设 z 是多个特征的属性值向量，

$$Z_G = \{z_{g_1}, z_{g_2}, \cdots, z_{g_n}, z_o\} \tag{6}$$

是地理特征属性向量的集合。如果 Z_G 中的数据是媒介，称 Z_G 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，我们能对空白单元的灾情，进行估值。此时，“人口”、“人均 GDP”和“相对暴露度”等就是背景数据。

依据定义 4 可知，一个属性值的数据是否为另一个属性值的背景数据，由两个因素决定，一是能否发挥制约作用；二是能否容易获得。当我们研究一个地区的洪水灾情时，各地理单元上的水灾程度，受地理单元的位置、人口和经济发展程度制约，而且这些信息很容易得到。地理位置，常常决定了地理单元暴露于洪水的程度。河流流经区域较大的地理单元，其暴露度较高。“相对暴露度”是由地理位置转化而来的，具有地理位置的信息。

至此，我们可以提出借助背景数据在地理空间上进行信息扩散的形式化定义如下：

定义 5：设 W 是用于识别区域 G 上的现象 F 的不完整数据集，Z_G 是背景数据。如果模型 γ 能用 Z_G 使 W 变成完整数据，则称 γ 用 Z_G 对 W 的信息在 G 上进行了扩散。

显然，任何内插算法都不能在地理空间上进行信息扩散，因为背景数据在内插法中发挥不了作用。Brunsdon 等人提出的地理加权回归方法(GWR)[27]，本质上是一个地理空间上的信息扩散模型。

GWR 是一种空间预测模型[28]，基本表述式为：

$$y(u,v) = b_0(u,v) + b_1(u,v)x + \varepsilon(u,v) \tag{7}$$

式中，y 表示服从高斯分布的因变量，x 是自变量，u 和 v 是数据的坐标，b_0 是截距系数，b_1 是斜率系数，ε 是随机误差项。多元 GWR 用式(8)表之。.

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{8}$$

式中，y_i 表示在位置 i 处的因变量，$\beta_0(u_i, v_i)$ 是在 i 处的截距系数，x_{ik} 是在位置 i 处的第 k 个自变量的值，$\beta_k(u_i, v_i)$ 是第 k 个自变量的局部回归系数。在 GWR 模型中，(u_i, v_i) 表示笛卡尔点坐标，ε_i 为随机误差项。

事实上，GWR 模型仍是一个线性回归模型，通常用最小二乘法来估计模型中的系数。GWR 当然也可拓展为广义的线性模型，例如 Logistic 回归和 Poisson 回归[29]。如果我们清楚地知道什么类型的函数可以表达 y 和 x 之间的关系，无论它是多么的高度非线性或非单调，我们都可以用此函数来改造 GWR 模型，达到最佳的回归效果。然而，问题在于特别难以确定哪种非线性函数适合于表达特定的地球表面现象。逃避这一问题的最常见做法，是想当然地

人为假设出某种函数；略为装饰一下的做法是追加统计检验。

为解决不依靠假设函数而识别非线性关系的问题，人们提出了人工神经网络（ANN）模型。由于训练后的 ANN，能够担当定义 5 中模型 γ 的角色，ANN 可以作为一个地理空间上的信息扩散模型使用。理论上，只要有足够多的隐含层神经元，ANN 能以任意精度逼近任何一个连续函数[30]。现实中，由于 ANN 是在计算机上学习训练样本，因此任意精度逼近的事实并不会出现[31]。更致命的是，如果训练样本由于随机干扰等而出现矛盾冲突，则 ANN 不收敛[31]。

为了填补地理空间上的数据空白，用 GWR 模型或 ANN 模型，在背景数据的帮助下，我们都可以将已观测数据的信息，扩散到空白单元中去，将地理空间上的不完整数据，变为完整数据。但是，当遇到非线性问题或矛盾样本时，由这两种模型推测出的空白单元中的待观测数据，可能与客观数据相差过大，失去使用价值。为解决这一问题，我们建议用基于信息扩散理论的自学习离散回归(SLDR)模型[7]来实现地理空间上的信息扩散。

5. 自学习离散回归模型

不失一般性，假设研究区域 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，

$$G=\{ g_1, g_2, ..., g_{n-q}, g_{n-q+1}, ..., g_n\} \tag{9}$$

更进一步，假设 t 个地理特征的属性值是背景数据，地理单位 g_i 第 j 个特征的属性值是 z_{ij}。于是，关于 G 上的信息可由如表 1 示之。

表 1. 研究区域 G 上的观测值和背景数据

地理单元	背景数据	观测值
g_1	z_{11} z_{12} ... z_{1t}	w_1
g_2	z_{21} z_{22} ... z_{2t}	w_2
...	...	...
g_{n-q}	z_{n-q1} z_{n-q2} ... z_{n-qt}	w_{n-q}
g_{n-q+1}	$z_{n-q+1,1}$ $z_{n-q+1,2}$ ... $z_{n-q+1,t}$	Unknown
...	...	...
g_n	z_{n1} z_{n2} ... z_{nt}	Unknown

统计回归的经验告诉我们，除非用于回归的样本，其容量 n-q≥30×(t+1)，或者我们知道背景数据和观测值的因果关系函数类型（例如，线性函数或指数函数），否则 GWR 模型填补出来的数据不可信；训练 ANN 的经验告诉我们，除非用于训练的样本，即表 1 中非 Unknown 行构成的样本，内中没有矛盾冲突，否则 ANN 模型在不能收敛情况下填补出来的数据也不可信。

SLDR 模型，则能解决 GWR 和 ANN 面临的上述问题。该模型由“构建关系矩阵”和“用背景数据推测待观测数据”两部分构成。

5.1. *用背景数据和已观测数据构建关系矩阵*

设 x 是定义在集合Ω上的一个变量。如果我们在讨论 x 时所关注的是Ω上的一个子集 U，则称 U 是 x 的论域。当 U 用于接收观测值所扩散的信息时，称 U 是监控空间。

设U_j，j=1,2,...,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视空间。令λ=t+1，则背景数据监控空间和已观测数据监视空间构成了一个λ维监控空间：

$$U_1 \times U_2 \times \cdots \times U_\lambda. \tag{10}$$

式中，$U_j = \{u_{j1}, u_{j2}, \cdots, u_{jm_j}\}, j = 1,2,\cdots,\lambda.$

令τ=n-q。从表 1 中，我们得到容量为τ的λ维样本 X：

$$X = \{(x_{i1}, x_{i2}, \cdots, x_{i\lambda-1}, x_{i\lambda}) \mid i = 1,2,\cdots,\tau\} \tag{11}$$

式中，

$$x_{i1} = z_{i1},\ x_{i2} = z_{i2}, \cdots,\ x_{i\lambda-1} = z_{it},\ x_{i\lambda} = w_i,\ i = 1,2,\cdots,\tau.$$

对于λ维样本点，

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{i\lambda}) \in X, \tag{12}$$

和λ维监控点，

$$\boldsymbol{u} = (u_{1k_1}, u_{2k_2}, \cdots, u_{\lambda k_\lambda}) \in U_1 \times U_2 \times \cdots \times U_\lambda, \tag{13}$$

（此处 $k_j \in \{1,2,\cdots,m_j\}, j = 1,2,\cdots,\lambda$），我们用式(14)的$\lambda$维正态扩散公式，将 $\boldsymbol{x}$ 的信息扩散到 $\boldsymbol{u}$.

$$\mu(\boldsymbol{x}_i, \boldsymbol{u}) = \prod_{j=1}^{\lambda} \exp[-\frac{(x_{ij} - u_{jk_j})^2}{2h_j^2}]. \tag{14}$$

式中的扩散系数 h_j, j=1,2,..., λ根据表 1 中的背景数据和已观测数据，分别用式(15)进行计算。

$$h_j=\begin{cases}0.8146(b-a), & \tau=5;\\ 0.5690(b-a), & \tau=6;\\ 0.4560(b-a), & \tau=7;\\ 0.3860(b-a), & \tau=8;\\ 0.3362(b-a), & \tau=9;\\ 0.2986(b-a), & \tau=10;\\ 2.6851(b-a)/(\tau-1), & \tau\geq 11.\end{cases} \tag{15}$$

式中，$b=\max_{1\leq i\leq\tau}\{x_{ij}\}$，$a=\min_{1\leq i\leq\tau}\{x_{ij}\}$.

令

$$Q_{k_1k_2\cdots k_\lambda}=\sum_{i=1}^{\tau}\prod_{j=1}^{\lambda}\exp[-\frac{(x_{ij}-u_{jk_j})^2}{2h_j^2}]. \tag{16}$$

我们获得了一个 $U_1\times U_2\times\cdots\times U_\lambda$ 上的，关于 X 的信息矩阵，如式(17)所示：

$$\boldsymbol{Q}=\{Q_{k_1k_2\cdots k_{\lambda-1}k_\lambda}\}_{m_1\times m_2\times\cdots\times m_{\lambda-1}\times m_\lambda} \tag{17}$$

$\forall k_\lambda\in\{1,2,\cdots,m_\lambda\}$，令

$$S_{k_\lambda}=\max_{\substack{1\leq k_j\leq m_j\\ 1\leq j\leq m_{\lambda-1}}}\{Q_{k_1k_2\cdots k_{\lambda-1}k_\lambda}\}, \tag{18}$$

和

$$r_{k_1k_2\cdots k_{\lambda-1}k_\lambda}=\frac{Q_{k_1k_2\cdots k_{\lambda-1}k_\lambda}}{S_{k_\lambda}}. \tag{19}$$

我们可构造出表 1 中背景数据和观察之间的关系矩阵，记为：

$$R=\{r_{k_1k_2\cdots k_{\lambda-1}k_\lambda}\}_{m_1\times m_2\times\cdots\times m_{\lambda-1}\times m_\lambda} \tag{20}$$

5.2.*用背景数据推测待观测数据*

设 $\boldsymbol{z}=(z_1,z_2,\cdots,z_t)$ 为表 1 中空白单元的背景数据，且

$$\boldsymbol{u}_{\lambda-1}=(u_{1k_1},u_{2k_2},\cdots,u_{\lambda-1k_{\lambda-1}})\in U_1\times U_2\times\cdots\times U_{\lambda-1}.$$

我们可以用式(21)的λ-1 维正态扩散公式将此 $\boldsymbol{z}$ 变为论域 $U_1\times U_2\times\cdots\times U_{\lambda-1}$ 上的一个模糊集，并取其隶属度最大值用式(22)进行归一化。

$$\mu(\boldsymbol{z},\boldsymbol{u}_{\lambda-1})=\prod_{j=1}^{\lambda-1}\exp[-\frac{(z_j-u_{jk_j})^2}{2h_j^2}]. \tag{21}$$

$$\begin{cases}a_{k_1k_2\cdots k_{\lambda-1}}=\dfrac{q_{k_1k_2\cdots k_{\lambda-1}}}{s},\\ s=\max\limits_{\substack{1\leq k_j\leq m_j\\ 1\leq j\leq\lambda-1}}\{q_{k_1k_2\cdots k_{\lambda-1}}\},\\ q_{k_1k_2\cdots k_{\lambda-1}}=\prod\limits_{j=1}^{\lambda-1}\exp[-\dfrac{(z_j-u_{jk_j})^2}{2h_j^2}].\end{cases} \tag{22}$$

此模糊集记为 $\widetilde{A}$，其各隶属度值是：

$$a_{k_1k_2\cdots k_{\lambda-1}},k_j=1,2,\cdots,m_j,\ j=1,2,\cdots,\lambda-1.$$

对于模糊输入 $\widetilde{A}$，使用近似推理公式（23），我们可以得到一个具有隶属函数 $\mu_B(u_{\lambda k_\lambda})$ 的模糊输出 $\widetilde{B}$。

$$\mu_B(u_{\lambda k_\lambda})=\max_{\substack{1\leq k_j\leq m_j\\ 1\leq j\leq\lambda-1}}\min\{a_{k_1k_2\cdots k_{\lambda-1}},r_{k_1k_2\cdots k_{\lambda-1}k_\lambda}\}. \tag{23}$$

最后，使用式（24）的重心方法，我们获得了一个分明值 w：

$$w=\frac{\sum_{k_\lambda=1}^{m_\lambda}\mu_B(u_{\lambda k_\lambda})u_{\lambda k_\lambda}}{\sum_{k_\lambda=1}^{m_\lambda}\mu_B(u_{\lambda k_\lambda})}. \tag{24}$$

由上述公式(14)-(24)组成的模型，被称为自学习离散回归（SLDR）模型。

一个关于洪水灾区中空白单元损失数据推测的研究表明，相比地理加权回归法，本文提出的自学习离散回归信息扩散法，能减少大约 60％的误差。

6. 结论与讨论

面对地理学中的数据不完整问题，人们对插值和地理加权回归（GWR）等方法，进行了大量卓有成效的研究，解决了一系列现实问题。但当地理单元较大时，大多数地理特征的数属性值在相邻单元上并不连续，插值法对变量连续性的基本要求得不到满足，插值得到的数据，没有意义。而当所研究的地理学现象过于复杂时，任何假设出来的，描述此现象的函数类型，都无法证实。基于不能证实的假设函数进行的地理加权回归，统计结果的可靠性，疑问重重。

由于将观测值变模糊集的信息扩散方法，能部分弥补小样本缺陷，提高系统识别精度，似乎我们可以将信息扩散方法用于相邻地理单元上的信息扩散，由几个相邻地理单元上的已知数据，推导出空白单元上的数据，使研究区域的不完整数据，变成完整数据。遗憾的是，这条路不通，因为传统的信息扩散模型，

是在所处理观测值的论域上，将观测值变模糊集。当观测值是随机样本点时，此论域，是随机事基本空间。因此，传统的信息扩散，是在概率空间中进行，而不是在地理空间上进行。

在传统信息扩散理论框架之内，本文提出了，以背景数据为桥梁，将已观测数据携带的信息，扩散到空白单元的方法，称为借助背景数据的地理空间信息扩散技术，并给出了自学习离散回归（SLDR）模型。

从原理上讲，SLDR 与 GWR 及人工神经网络（ANN）一样，都属于统计回归方法，但 SLDR 不仅能统计回归出非线性关系，而且容忍观测数据间存在矛盾，而不会出现 ANN 的收敛问题。

案例研究表明，SLDR 的估计比地理加权回归法减少了大约 60％误差，优势明显。

概率空间中的信息扩散模型，自然要比地理空间上的信息扩散模型简单，因为前者是无约束扩散，后者是有约束。随机样本在扩散模型中不受样本以外的因素影响，地理数据在扩散模型中则受到背景数据的约束。少数几个背景数据支持下的信息扩散，其结果与大量背景数据时的肯定不同；制约程度较高的背景数据支持下的信息扩散，其修补得到的完整数据的质量，肯定高于制约程度较低的质量。

地理空间上的信息扩散的研究，为信息扩散理论和方法的发展，提供了巨大的空间。

背景数据支持下的，地理空间上的信息扩散，本质上是某种“联想”。人类通过联想对事物进行判断，成功与否，既与经验和知识有关，也与联想者的智能水平有关。地理空间上信息扩散中用到的背景数据和给定的不完整数据，相当于联想者的经验和知识；扩散模型的好坏，相当于联想者智能水平的高低。

在人工智能席卷全球的今天，地理空间上信息扩散理论和方法的研究，必定能在智慧地学的建设中发挥重要作用。

参考文献

[1] 傅伯杰. 新时代自然地理学发展的思考. 地理科学进展 , 2018, 37 (1) :1-7.
[2] 张勇, 许力生, 陈运泰. 2008 年汶川大地震震源机制的时空变化. 2009, 52 (2): 379-389.
[3] 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. 地震学报, 2009,31(2):205-211.
[4] Eldrandaly K A，Abu-Zaid M S. (2011). Comparison of six GIS-based spatial interpolation methods for estimating air temperature in western Saudi Arabia. Journal of Environmental Informatics, 2011, 18(1), 38–45.
[5] Huang C F. Principle of information diffusion. Fuzzy Sets and Systems, 1997, 91(1): 69-90.
[6] 黄崇福. 地理空间上的信息扩散及其在风险分析中的应用. 一带一路背景下的风险分析和危机反应---中国灾害防御协会风险分析专业委员会第八届年会论文集(西安，2018 年 10 月 20-21 日), 1-7.
[7] Huang C F, Shi Y. Towards Efficient Fuzzy Information Processing: Using the Principle of Information Diffusion, Heidelberg, Germany: Physica- Verlag (Springer), 2002.
[8] Huang C H, Zong T, Chen Z F. Four models to calculate a fuzzy probability distribution with a small sample. International Journal of Information Technology & Decision Making, 2007, 6(4): 611–623
[9] Chen Q, Rui H, Li W, Zhang Y. Analysis of algal bloom risk with uncertainties in lakes by integrating self-organizing map and fuzzy information theory. Science of the Total Environment, 2014, 482–483(1): 318–324.
[10] Feng L, Luo G, Application of possibility-probability distribution in risk analysis of landfall hurricane—A case study along the east coast of the United States. Applied Soft Computing, 2011, 11 (8): 4563–4568.
[11] Zou Q, Zhou J, Zhou C, Guo J, Deng W, Yang M, Liao L. Fuzzy risk analysis of flood disasters based on diffused-interior-outer-set model. Expert Systems with Applications, 2012, 39 (6): 6213–6220.
[12] Liu X, Zhang J, Cai W, Tong Z. Information diffusion-based spatio-temporal risk analysis of grassland fire disaster in northern China. Knowledge-Based Systems, 2010, 23 (1): 53–60.
[13] Zhong L, Liu L, Liu Y. Natural disaster risk assessment of grain production in Dongting Lake Area, China. Agriculture and Agricultural Science Procedia, 2010, 1(1): 24-32.
[14] Li Q, Zhou J, Liu D, Jiang X. Research on flood risk analysis and evaluation method based on variable fuzzy sets and information diffusion. Safety Science, 2012, 50(1): 1275-1283.
[15]Olya H G T, Alipour H. Risk assessment of precipitation and the tourism climate index. Tourism Management, 2015, 50: 73-80.
[16] Xu L, Xu X, Meng X. Risk assessment of soil erosion in different rainfall scenarios by RUSLE model coupled with information diffusion model: A case study of Bohai Rim, China. Catena, 2012, 100(2): 74-82.
[17]Chu Y Y, Dong W L, Li Y, Liang D. Risk prediction model of LNG terminal station based on information diffusion theory. Procedia Engineering, 2013, 52: 60-66.
[18] Hao L, Yang L, Gao J M. The application of information diffusion technique in probabilistic analysis to grassland biological disasters risk. Ecological Modelling, 2014, 272: 264-270.
[19]Xing L, Lu K. The importance of public-private partnerships in agricultural insurance in China: based on analysis for Beijing. Agriculture and Agricultural Science Procedia, 2010, 1(1): 241-250.
[20] Zou Q, Zhou J, Zhou C, Song L, Guo J, Liu Y. The practical research on flood risk analysis based on IIOSM and fuzzy α-cut technique. Applied Mathematical Modelling, 2012, 36(7): 3271-3282.

[21] Xu L, Liu G. The study of a method of regional environmental risk assessment. Journal of Environmental Management, 2009, 90(11): 3290-3296.
[22] Mako Z. Approximation with diffusion-neural-network. Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence (November 18-19, 2005, Budapest), 2005, pp. 589-600.
[23] Yan J P, Li S S, Bai J, Liu X Y. The spatial symmetry axis of earthquake hazard in China. Journal of Risk Analysis and Crisis Response, 2013, 3(1): 59-64.
[24] Cartier S, Vallette C, Mediene H. Urban seismomorphoses seismic vulnerabilities, an embarrassing legacy. Journal of Risk Analysis and Crisis Response, 2012, 2(2): 96-106.
[25] Tobler W R. A computer movie simulating urban growth in the Detroit region. Economic Geography, 1970, 46:234-240.
[26] Eldrandaly K A, Abu-Zaid M S. Comparison of six GIS-based spatial interpolation methods for estimating air temperature in western Saudi Arabia. Journal of Environmental Informatics, 2011, 18(1): 38–45.
[27] Brunsdon C, Fotheringham A S, Charlton M E. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical Analysis, 1996, 28(4): 281-298.
[28] Lieske D J, Bender D J. A robust test of spatial predictive models: geographic cross-validation. Journal of Environmental Informatics, 2011, 17(2): 91-101.
[29] Fotheringham A S, Brunsdon C, Charlton M. Geographically Weighted Regression: the Analysis of Spatially Varying Relationships, Chichester, USA: John Wiley & Sons, 2002.
[30] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks, 1989, 2 (5): 359–366.
[31] Wray J, Green, G G R. Neural networks, approximation theory, and finite precision computation. Neural Networks, 1995. 8 (1): 31–37.
[32] Huang C F, Moraga C. A diffusion-neural-network for learning from small samples. International Journal of Approximate Reasoning, 2004, 35: 37–161.