空间软数据及其插值方法研究进展

罗 明[1,2],裴 韬[1,3]

(1. 中国科学院地理科学与资源研究所,北京 100101; 2. 中国科学院研究生院,北京 100049;
3. 华东师范大学地理信息科学教育部重点实验室,上海 200062)

摘 要:由于对地观测技术的迅速发展,空间数据的种类和数量增长迅猛,由空间数据反演得到的各种信息日趋膨胀,这些反演结果中的信息不少以软数据的形式出现。在实际应用中,这些软数据往往与空间插值的目标变量具有一定的相关性,甚至成为控制目标变量空间分布特征的重要因素。然而,由于这些数据通常表示为非数值形式,在计算和处理上存在着一定困难,以致被传统的插值方法所忽视,从而造成信息浪费。近来出现的空间软插值方法是一种利用空间软数据作为辅助信息并以改善插值效果的方法,能够较好的处理并利用软数据所隐含的信息,具有较好的应用发展前景。本文根据空间软数据的特点及其分类,系统综述了空间软插值方法及其应用领域。首先分析了空间数据软硬性质的根本区别,论述了软数据的分类和"硬化"方法,然后介绍空间插值模型中对空间软数据的集成方法和原理,最后对空间软插值方法及其应用研究领域进行了展望。

关 键 词:软数据;空间插值;空间信息统计;克里格;回归分析;贝叶斯最大熵

1 引 言

空间插值是利用采样点数据对研究区内的其他未知区域的特征数据进行推理和估计的方法[1-3],传统的空间插值方法主要用于空间定量属性(如土壤中的元素含量、某站点的降雨量等)为主的插值和模拟[4-7],其中的一些方法(如协同克里格等)也考虑了与目标变量相关的辅助变量的协同作用[8-10],但是却忽视了那些不具有定量数值而表现为类别的有用信息,例如,采样点的土地利用类型、岩石类型、某站点所属的气候类型、某城市所属的经济区带类型等等。这些数据往往与空间插值的目标变量之间存在着不同程度的相关性,并且隐藏着丰富的、不可忽视的有用信息,例如,土壤中某种元素的含量与土地利用类型密切相关,那么该元素的空间分布则很有可能受到土地利用类型的控制[11]。近年来,对地观测技术的迅速发展使得空间数据的种类和数量增长迅猛,所获的地表信息日趋膨胀,其中就包含了大量的这样的信息(如除土地利用类型信息之外的植被类型、地物类型、地貌信息等)。然而在实际应用中,由于这些数据不是直接表现为数值形式,在计算和处理上存在着一定的困难,因此未得到普遍而有效的利用,其中隐含的丰富信息也没有得到充分的发掘,从而造成了大量的信息浪费[12-14]。

近年来,一些学者开始注意到这些数据中所蕴藏的潜力,并尝试在实际应用中利用这些数据。有学者将其中这些分类的、不确定的、主观的或定性的数据称为软数据(Soft data)或者软信息(Soft information),并给出了对软数据的不同解释。Journel 指出,软数据是指这两种形式的数据:一种是不等式类型数据,如空间上某处的属性值大于或者小于某阈值等;另一种是定性数据,如关于断层、褶皱、矿化极限、连续性的地质解译等[15]。Hendriks 等将基于主观判断、定性的感官观察结果定义为软数据[16],而 Goovaerts 等学者则把土壤类型图、野外定性观测结果(如土壤颜色、气味等)看成是软数据[17-19]。此外,也有一些学者将带有不确定性的概率信息或者间隔数据当作软数据,如 Seibert 和 McDonnell 将软数据定义为不能直接以准确的数字形式来表达的定性知识[20],D'Or 和 Bogaert 把数据间隔、概率密度函数等当成是软数据[21],Serre 则直接将空间属性的

收稿日期:2009-03; 修订日期: 2009-05.
基金项目:国家自然科学基金项目(40601078);国家 863 项目(2006AA120106);华东师范大学地理信息科学教育部重点实验室开放研究基金资助项目。
作者简介:罗明(1986-),男,江西抚州人,硕士研究生,研究方向为空间信息统计。E-mail: luom@lreis.ac.cn
通讯作者:裴韬,peit@lreis.ac.cn

取值概率当作软数据[22],Douaik 等人把描述估计值的不确定性的间接数据认为是软数据 [23],Emery 认为由事件概率的不等式约束所构成的数据即为软数据等等[24]。上述学者虽然在其研究中都提到了软数据的概念,但仍没有对软数据进行系统的定义和分析,同时也未对数据的“软”与“硬”性质进行根本的区分。

通过对文献的总结并结合近年的实践,我们认为空间软数据是指描述对象表现为数值区间、类别等“集合”形式的空间数据,它与空间硬数据之间的本质区别在于:软数据所描述的是“集合”,而硬数据描述的是“具体的数值”。例如,某个气象站点所属的气候带类型是一种“集合”的表述,而站点处的温度、风速、降雨量是“具体的数值”;某点处所属的土壤类型所涵盖的也是“集合”,而土壤有机质含量是“具体的数值”。又如,遥感分类结果是一种软数据,其中某个(或某几个)波段的反射率范围代表林地,实际上,分类结果中的林地是由反射率在一定范围内的象素(即某个“集合”)所组成的。需要说明的是,这里的“集合”与误差“集合”不同,误差“集合”是指由测量值与误差所构成的区间,而软数据描述的“集合”则是指研究对象本身。

2 空间软数据

2.1 空间软数据的特点

与硬数据不同,空间软数据描述的对象是“集合”,它具有自身特殊的性质,主要体现在不等性、随机性、定性特征、综合性、复杂性等几个方面。

2.1.1 不等性

软数据不能像硬数据那样可以准确的表达为:空间上某点处的属性值“等于某个数值”,它一般描述成这样的形式:该点的属性属于某类别、大于或者小于某个阈值等。例如,我们一般说土壤类型图上的某一点是属于某种土壤类型[19,25],而不说它具体等于某个数值。

2.1.2 随机性

由于空间软数据反应的是不等性,描述的是类别或者区间范围,因此从一定意义上来说软数据都可以用“属于某类别”或者“落在某区间”的概率形式来表示。对于空间上的某点而言,空间软数据表示的实际上就是落在某个阈值区间内的概率、大于或者小于某个阈值的概率等信息。例如,在环境污染评价中,就常用某有害物质(如某类重金属等)的含量超过安全阈值的概率来进行评价[26]。

2.1.3 定性特征

与硬数据相比,软数据往往结合了更多的先验知识、专业认识等信息来对现象进行描述,因此非常适用于表达研究对象的定性信息。例如,遥感分类图实际上就是一种定性信息,我们关心的往往不是像素点的反射率值具体是多少,而是希望知道它属于是水体还是农田,是林地还是居民区等。

2.1.4 综合性

尽管软数据不能表示具体的数值,但是却包含了硬数据无法表达的综合性信息,而且在很多情况下, 用软数据描述地理对象往往比硬数据更加直观。例如,土壤类型图表示的不同土壤类型,实际上就隐含了该类型的土壤颜色、质地、颗粒度、潮湿度等综合信息。

2.1.5 复杂性

软数据描述的是“集合”而不是“具体的数值”,这就决定了对软数据与硬数据的处理需要使用不同的方法。对软数据的处理不能像对硬数据那样直接使用数值计算的方法进行处理[18],而一般都需要进行“硬化”处理,以便可以像硬数据一样方便的参与模型计算。因此,如何对软数据进行“硬化”是利用软数据首先要解决的问题,同时这也给软数据的利用带来一定的复杂性。

2.2 空间软数据的分类

空间软数据的分类可以有多种方式,常用的可以有两种。第一种可根据软数据所描述的“集合”边界的确定程度将其划分为确定型和模糊型两类:确定型软数据的“集合”边界相对确定、界限清楚,例如,砷元素的污染程度,由于砷元素污染的下限是明确的数值,因此,“某点砷污染”就表示该点的砷元素含量大于砷污染下限;模糊型软数据的“集合”边界则相对模糊,例如,某点的土壤类型 A,由于土壤类型之间的界限比较模糊,因此在土壤普查中可用模糊隶属度表示某点对土壤类型的归属性。

第二种分类方法可以按照表达软数据的变量的类型进行分类。通常可根据所采用的变量类型是否为类别变量或者序列变量等,可以将软数据分为类别型软数据、序列型软数据等类型。例如,遥感分类图、土壤类型图等以类别型变量表示的类别型软数据,以及如水质的重污染区、中度污染区、轻污染

区、无污染区等这些具有序列变化特征的序列型软数据。

2.3 软数据的“硬化”方法

由于当前大多数的地学环境分析模型都只能处理硬数据,而不能直接有效地处理和利用软数据[18],因此对软数据的利用首先要将其“硬化”为硬数据。对于不同类型的软数据,我们可以采用不同的“硬化”方法。本节分别对确定型软数据和模糊型软数据的“硬化”方法进行介绍。

2.3.1 确定型软数据的“硬化”方法

对确定型软数据来说,不同“集合”之间的界限是清楚的、确定的,它描述的实际上是某属性完全地等于或者属于某个“集合”。确定型软数据一般使用式(1)的指示函数形式来进行“硬化”:

$$i(u;\ S_k)=\begin{cases}1, \text{ if } \ z(u)\in S_k\\ 0, \text{ if } \ z(u)\notin S_k\end{cases} \tag{1}$$

式中:$z(u)$为位置 u 处的空间属性;S_k 是软数据描述的“集合”;$i(u;\ S_k)$表示“硬化后”的数据,当属性 $z(u)$属于“集合”S_k 时 $i(u;\ S_k)$为 1,不属于“集合”S_k 时为 0。

2.3.2 模糊型软数据的“硬化”方法

与确定型软数据不同,模糊型软数据的“集合”边界是模糊的、不确定的,它无法确定空间上某属性是否一定等于或者属于某个“集合”,而采用一种近似的或者不确定的形式来表达。糊型软数据的“硬化”常常使用概率或者模糊隶属度的形式:

$$i(u;\ S_k)=Prob[z(u)\in S_k] \tag{2}$$

式中:“硬化”后的数据 $i(u;\ S_k)$表示空间属性 $z(u)$在“集合”S_k 内出现的概率或者隶属于类别 S_k 的隶属度,为介于 0~1 之间的数。

3　空间软插值方法研究进展

空间软插值,即基于空间软数据的插值方法,是指利用空间软数据作为辅助信息对空间变量进行插值估计的方法,它是以改善插值效果为目的,将软数据有效地集成到空间插值模型中。与硬数据描述“具体的数值”不同,软数据描述的是“集合”,这使得空间软插值方法利用软数据作为辅助信息的方式与一般插值方法存在较大的区别,主要体现在对辅助数据的集成处理方式不同。一般的“硬”插值方法利用辅助数据一般通过两种形式:一种是利用辅助变量直接估计目标变量的空间变化趋势,如回归分析、回归克里格、漂移克里格等,另一种则是将辅助变量与目标变量同时集成到插值估计模型中,如协同克里格、同位克里格、指示克里格等。空间软插值对软数据的集成一般需要先对软数据进行“硬化”,然后再集成“硬化”后的数据。空间软插值对软数据的集成主要有三种方式:一是直接将“硬化”后的数据参与到插值估计模型;二是利用硬化后的数据产生条件概率,再将该条件概率融入插值模型中;三是将软数据作为分类或者分层的根据对采样点进行分类,并计算局部均值的估计值。

3.1 简单克里格

简单克里格方法(Simple kriging)是指在属性值的数学期望处处相等并且已知的情况下对待估点处的属性值进行估计的克里格方法。在此基础上,有学者提出数学期望已知并且在空间上是变化的简单克里格方法,并称之为变局部均值的简单克里格(Simple kriging with varying local means)[8,27],它的估计模型可以表示为:

$$z^*(u)=m^*(u)+\sum_{\alpha=1}^{n(u)}\lambda_a^{S_k}(u_\alpha)\cdot[z(u_\alpha)-m^*(u_\alpha)] \tag{3}$$

式中:$z^*(u)$和 $m^*(u)$分别为待估点 u 处的估计值和期望值,$z(u_\alpha)$为邻域观测点 u_α 处的已知值,$\lambda_a^{S_k}$为简单克里格权值,可由克里格方程组求解得到。

在简单克里格方法的计算中,对空间软数据的集成体现在局部变化均值由待估点所在软数据类别 S_k 内的条件均值 $m^*_{|S_k}$ 代替,即 $m^*(u)=m^*_{|S_k}$,条件均值 $m^*_{|S_k}$ 是指类别 S_k 内的所有采样点的观测值均值:

$$m^*_{|S_k}=\frac{1}{n_k}\sum_{a=1}^{n_k}i(u_a;\ S_k)\cdot z(u_a) \tag{4}$$

式中:n_k 是属于类别 S_k 内所有观测点个数,$i(u_a;S_k)$为“硬化”后的数据。

3.2 协同克里格

协同克里格可以同时利用多种辅助数据进行线性无偏估计,其中能够集成软数据的有普通协同克里格(Ordinary cokriging)和同位协同克里格(Colocated cokriging)。

3.2.1 普通协同克里格

普通协同克里格集成空间软数据有两种不同的方式,一种是直接将“硬化”后的数据集成克里格插值模型中[28]:

$$z^*(u)=\sum_{a=1}^{n(u)}\lambda_\alpha(u)\cdot z(u_a)+\sum_{K=1}^{K}\sum_{a'=1}^{n'(u)}\lambda_{\alpha'}(u;S_k)i(u_{a'};S_k) \quad (5)$$

式中：u、u_a、$u_{a'}$分别指待估点、主变量观测点、软数据观测点；$z^*(u)$为待估属性值；$\lambda_\alpha(u)$和$\lambda_{\alpha'}(u;S_k)$为克里格权值；$i(u_{a'};S_k)$为“硬化”后的数据。

第二种方式则是将式(8)中处理之后的条件概率$y(u;z_c)$参与到协同克里格模型中去[8]：

$$z^*(u;z_c)=m_z+\sum_{a=1}^{n(u)}\lambda_\alpha(u;z_c)\cdot[z(u_a)-m_z]+\lambda(u;z_c)\cdot[y(u;z_c)-m_y(z_c)] \quad (6)$$

式中：m_z和$m_y(z_c)$分别为目标变量和条件概率的均值。

3.2.2 同位协同克里格

同位协同克里格是普通协同克里格的一种简化形式，即如果辅助变量密集取样时，在计算过程中只使用与估计点同位的辅助变量[29]。同位协同克里格集成软数据的方式与协同克里格相似，也是直接将条件概率$y(u;z_c)$参与到估计模型中去[30,31]：

$$z^*(u;z_c)=\sum_{a=1}^{n(u)}\lambda(u_a;z_c)\cdot[z(u_a)-m_z(u)]+\lambda(u;z_c)\cdot[y(u_a)-m_y(z_c)]+m_z(u) \quad (7)$$

式(7)中各项意义与式(6)相同。

3.3 指示克里格

指示克里格方法 (Indicator kriging，IK) 由Journel 于 1983 年提出，它是一种非参数估计方法，一般用于估计目标变量落在某个阈值范围内(如污染物含量大于某个安全阈值等)的可能性大小[32]，或者估计未知点处的所属类别[33]。指示克里格方法中对软数据的集成利用方式是：

首先利用软数据产生条件概率，然后再将该概率值参与到克里格估计模型中去[34–37]。软数据条件概率一般是通过“硬化”后的数据(式(1)和式(2))和目标变量指示值(式(9))来产生：

$$y(u;z_c)=\mathrm{Prob}\{z(u)\leqslant z_c|s(u)=S_k]\}=\frac{\sum_{a=1}^{n}i(u_a;z_c)\cdot i(u_a;S_k)}{\sum_{a=1}^{n}i(u_a;S_k)}\in[0,1] \quad (8)$$

式中：z_c是对目标变量z进行指示变换时使用的阈值；S_k是软数据的“集合”，N为“集合”S_k内的所有观测点个数，$i(u_a;S_k)$和$i(u_a;z_c)$分别表示“硬化”后的数据和目标变量指示值(式(9))。

$$i(u;z_c)=\begin{cases}1, & \text{if } z(u_a)\leqslant z_c\\ 0, & \text{if } z(u_a)>z_c\end{cases} \quad (9)$$

指示克里格方法种类较多，其中能利用软数据作为辅助信息有具局部先验均值的简单指示克里格 (Simple indicator kriging with local prior means)、Markov–Bayes 算法(Markov–Bayes algorithm)等。

3.3.1 简单指示克里格

简单指示克里格对软数据的集成是将软数据条件概率$y(u;z_c)$作为待估点处的局部均值，再对目标变量指示值$i(u;z_c)$进行简单克里格估计[38–41]：

$$i^*(u;z_c)=y(u;z_c)+\sum_{a=1}^{n(u)}\lambda_\alpha(u;z_c)\cdot[i(u_a;z_c)-y(u;z_c)] \quad (10)$$

式中：$\lambda_\alpha(u;z_c)$为克里格权值，式中其他各项的意义则与式(3)相同。

3.3.2 Markov–Bayes 算法

由于协同克里格方法需要计算目标变量自协方差函数、软数据自协方差函数和二值之间的互协方差函数，数据计算量较大，一定程度上影响了该方法的应用。为了减少协方差函数的计算，Zhu 和 Journel 利用马尔可夫假设推导协方差函数的简化模型[42]。马尔可夫假设指：硬数据$i(u;z_c)$屏蔽了任何配位的软数据条件概率$y(u;z_c)$的影响，即：

$$\mathrm{Prob}\{z'(u')\leqslant z_c|i(u;z_c),y(u;z_c)\}=\mathrm{Prob}\{z'(u')\leqslant z_c|i(u;z_c)\} \quad (11)$$

在马尔可夫假设的基础上，可以建立软数据协方差$C_Y(h;z_c)$、互协方差$C_{IY}(h;z_c)$与目标变量指示值自协方差$C_I(h;z_c)$之间的联系：

$$C_Y(h;z_c)=\begin{cases}\left|m^{(1)}(z_c)-m^{(0)}(z_c)\right|\cdot C_I(h;z_c), & h=0\\ \left[m^{(1)}(z_c)-m^{(0)}(z_c)\right]^2\cdot C_I(h;z_c), & h>0\end{cases} \quad (12)$$

$$C_{IY}(h;z_c)=\left[m^{(1)}(z_c)-m^{(0)}(z_c)\right]\cdot C_I(h;z_c) \quad (13)$$

式中：$m^{(1)}(z_c)$和$m^{(0)}(z_c)$为分别表示$i(u;z_c)$为 1 和 0 时$y(u;z_c)$的条件期望，可以由采样点处$i(u;z_c)$分别为 1 和 0 时$y(u;z_c)$的算术平均值来估计。实际上，$m^{(1)}(z_c)$和$m^{(0)}(z_c)$的差值也反应了软数据条件概率$y(u;z_c)$区分$i(u;z_c)=0$和$i(u;z_c)=1$的能力，换句话说就是，它可以用来衡量软数据的精确度大小[8]。

3.4 指示协同克里格

指示协同克里格方法是将指示克里格和协同克里格相结合，利用辅助变量构建协同克里格模

型,对目标变量指示值进行空间预测。能够集成空间软数据的指示协同克里格方法主要有普通指示协同克里格(Ordinary indicator cokriging)和同位指示协同克里格(Collocated indicator cokriging)两种。

3.4.1 普通指示协同克里格

与简单指示克里格将软数据条件概率当成局部均值不同的是,指示协同克里格是把条件概率作为协同变量直接参与到克里格估计[8,43]:

$$i^*(u;z_c)=\sum_{a=1}^{n(u)}\lambda_\alpha(u;z_c)\cdot i(u_a;z_c)+\sum_{a'=1}^{n'(u')}\lambda_{\alpha'}(u';z_c)\cdot y(u'_{a'};z_c)] \quad (14)$$

式中:$\lambda_\alpha(u;z_c)$和 $\lambda_{\alpha'}(u';z_c)$为克里格权值,其他各项意义则与前述相同。

3.4.2 同位指示协同克里格

在同位指示协同克里格方法的计算中,对软数据的集成估计模型可以表示为:

$$i^*(u;z_c)=\sum_{a=1}^{n(u)}\lambda_\alpha(u;z_c)\cdot i(u_a;z_c)+\lambda_2(u;z_c)\cdot[y(u;z_c)-m_y(z_c)+m_z(u)] \quad (15)$$

式中:$m_y(z_c)$和 $m_z(u)$分别为条件概率与目标变量的均值,其他各项意义与前述相同。与普通指示协同克里格相比, 同位指示协同克里格的速度更快,而且对于克里格权值的求解不需要软数据的自变差函数或者自协方差函数[8]。

3.5 分层克里格

分层克里格(Stratified kriging)采用分类思想对具有不同类别的数据分别进行克里格估计,类别数据可以是土地利用图、土壤类型图和地质图等软数据[44-47],实际上,分层克里格对软数据的集成利用一般是将其作为分层或者分类的依据。

后来也有一些学者对分层克里格方法进行改进。例如,Lagacherie 和 Voltz 在进行分类后,根据地形和邻域点的所属类别,推导待估点属于不同土壤类别的条件概率,然后根据该条件概率,对各个土壤类别的属性平均值进行加权平均求和,从而获得待估点处的估计值 [48];Boucneau 和 Liu 等在分层克里格的基础上,根据类间变化(是否渐变或者突变)、制图精度以及类内空间变异性等 3 个特征,将土壤图的绘图界线分为几种不同的类型,并分别使用不同的克里格方法进行进行插值估计[49,50];此外,利用软数据对采样点进行分类或者分层之后,除了可以使用普通克里格方法进行估计外,一些学者也采用了其他的克里格方法, 例如, Hengl 等在分类后利用回归克里格方法对土壤有机质含量、表层 PH 值、表层厚度进行空间预测[51],Wu 等人则使用块段克里格方法对土壤中的金属含量进行预测[25]。

对于类别内具有足够多观测点的情况,分层克里格方法中的类别内变差函数可以直接根据实验变差函数进行拟合得到;而对于类别内观测点个数较少的情况,则可以将多个类别内实验变差函数进行合并之后再进行克里格估计[45]。

3.6 外部漂移克里格

外部漂移克里格 (Kriging with external drift)将估计值 $z(u)$分为漂移与残差两个部分[52,53]:

$$z(u)=m(u)+R(u) \quad (16)$$

式中: $m(u)$漂移部分一般为外部变量 $y(u)$(即辅助变量)的线性函数: $m^*(u)=a(u)+b(u)\cdot y(u)$,而残差部分 $R(u)$则通过克里格估计得到。根据该原理,Monestiez 等将外部漂移克里格方法应用到外部变量为软数据时的情况,此时对漂移的估计变为[54]:

$$m^*(u)=a(u)+i(u;S_k)\cdot m^*_{|S_k} \quad (17)$$

式中:$i(u;S_k)$、$m^*_{|S_k}$ 的意义则与式(4)相同,分别表示“硬化”后的数据和条件均值(即类别 S_k 内的所有采样点的观测值均值);漂移估计的线性系数 $a(u)$是通过求解克里格方程组得到。

外部漂移克里格集成空间软数据的方式是利用软数据类别内的条件均值及待估点与邻域采样之间的软数据类别是否一致来估计目标变量的漂移,其特点是对每一软数据类别内的采样点个数有一定要求,而且对软数据与目标变量之间的相关性要求较高[54]。

3.7 贝叶斯最大熵

贝叶斯最大熵 (Bayesian maximum entropy, BME)[55-57]作为一种非线性估计方法,可以将硬数据和软数据(如数值间隔、概率密度函数、物理定律等)集成到一起[19,23,58],对目标变量的后验条件概率进行预测估计。在 BME 方法中,对软数据的集成主要是利用软数据来求取条件概率密度函数,其估计模型可以表示为:

$$f_K(x_k|x_{hard},x_{soft})=\frac{\int f_G(x_{map})dx_{soft}}{\int f_G(x_{hard},x_{soft})dx_{soft}} \quad (18)$$

式中：$x_{map}=(x_k, x_{hard}, x_{soft})$，$x_k$、$x_{hard}$、$x_{soft}$分别表示待估点处预测值、采样处的硬数据和软数据；$f_G(x_{map})$为先验概率密度，是在满足信息熵 H 最大的条件下利用拉格朗日乘数法求出的；$f_G(x_{hard}, x_{soft})$是硬数据和软数据的联合概率密度函数，是一个归一化系数；$f_K(x_k|x_{hard}, x_{soft})$为待求的后验概率密度函数。信息熵 H 的定义是：

$$H=-\int \log[f_G(x_{map})]\cdot f_G(x_{map})dx_{map} \quad (19)$$

在 BME 方法的基础上，Bogaert 和 D'Or 将 BME 算法和蒙特卡罗过程结合(BME/MC)，利用少量采样的硬数据和详细的土壤图进行了土壤属性的空间预测[18,21]。

3.8 回归分析方法

回归分析是通过揭示呈因果关系的相关变量间(因变量和自变量)的联系形式，建立它们之间的回归方程，利用所建立的回归方程，由自变量来预测因变量。回归分析方法中对软数据的集成一般是直接利用软数据作为自变量来对目标变量(因变量)进行回归估计。例如，Ohlmacher 和 Davis 在对山体滑坡危险度预测的研究中，首先对采样点处是否滑坡进行逻辑变换，然后利用地质图、地形坡度作为回归因子，对山体滑坡危险度进行空间预测[59]：

$$P^*(u)=\frac{1}{1+\exp\left[-\left(\beta_0+\beta_1\cdot s(u)+\sum_{k=1}^{K}\beta_{k+1}\cdot i(u;S_k)\right)\right]} \quad (20)$$

式中：$P^*(u)$为滑坡危险度的估计值；$s(u)$是坡度；$i(u;S_k)$为地质图指示值，在 u 处属于地质类型 S_k 时为 1，反之为 0；$\beta_{0,1,2,\cdots,k+1}$为回归系数。除地质图之外，其他如土壤类型、土地利用图等软数据同样可以作为回归因子对目标变量进行回归估计[37,60,61]。

在回归分析的基础上，Brunsdon 等提出一种考虑空间自相关性的回归方法，并称之为地理加权回归方法[62]。该方法基于邻近样本点对待估点的影响大于距离较远的样本点的假设对目标变量进行估计：

$$z(u)=\beta_0(u)+\sum_{k=1}^{K}\beta_k(u)\cdot x_k(u)+\varepsilon(u) \quad (21)$$

式中：$x_k(u)$为回归因子；$\beta_0(u)$和 $\beta_k(u)$为回归系数；$\varepsilon(u)$是误差项。

与常规回归方法类似，地理加权回归也可以将软数据当成回归因子对目标变量进行预测。例如，Tu 和 Xia 利用地理加权回归方法，根据土地利用图，获得土地类型比例、建设用地比例，并结合人口密度来计算土地利用指示值，并以此作为回归因子，对水质指示剂做回归预测[63]。与常规回归方法不同的是，地理加权回归实际上是一种局部估计方法，因为它的回归系数在不同的位置上是变化的。

3.9 其他方法

除上述几种方法外，还有一些软方法是在常规插值方法(如反距离加权、最近邻域法等)的基础上进行改进，将软数据集成到空间插值模型中。例如，Voltz 和 Lagacherie 等利用软数据对采样点进行分类后，用每类内的采样点观测值均值重新给采样点赋值，然后使用反距离加权和最近邻域法进行插值估计[47]。

Kasimov 和 KoSheleva 等在反距离加权插值方法的基础，根据待估点与采样点的多维软属性是否一致来计算环境相似度，并使用该相似度来对反距离权值进行校正[64]：

$$z^*(u)=\frac{\sum_{a=1}^{n}z(u_\alpha)\cdot\left[\left(a_0+\sum_{k=1}^{K}a_k i_{ka}\right)\Phi(r_\alpha)\right]}{\sum_{a=1}^{n}\left[\left(a_0+\sum_{k=1}^{K}a_k i_{ka}\right)\Phi(r_\alpha)\right]} \quad (22)$$

式中：r_α为待估点 u 与采样点 u_α 之间的距离；$\Phi(r_\alpha)$为距离函数，可以是 $1/r_\alpha$ 或者 $1/r_\alpha^2$ 等；i_{ka}当待估点 u 与采样点 u_α 的软属性相同时为 1，反之为 0；a_0为背景系数；系数 a_k 可以在采样点处的估计误差之和为最小的条件下，利用最小二乘法求取。与反距离加权相比，该方法的优点在于它考虑了利用多维软数据来描述环境相似性，并对反距离权值进行校正，对相似度较大的采样点采用较大的权值；不足之处在于该方法还是存在反距离加权方法的“极值”问题，即插值结果容易受极大值或极小值的影响。

4 空间软数据及软插值方法应用展望

发展可利用空间软数据的插值方法有助于充分挖掘空间软数据所蕴藏的“潜能”，这些软插值方法将具有非常广阔的应用前景。可以预见，空间软数据的应用有可能将在数据多样性、分辨率等方面成为相关领域的关注热点；与此同时，数据的多样性、多尺度、数据量大等特点也给空间软插值方法带来了挑战。

4.1 空间软数据的应用展望

近年来，随着数据获取技术水平的提高和数据的不断积累并日益丰富，可获得的数据种类繁多、

形式多样。例如,我们既可以通过卫星遥感手段获得实时遥感数据,也可以通过研究古代文献和史料记载来获得对灾情信息的历史记录等 (如地震、洪涝灾情等),还可以通过野外观测(人工观测或传感器监测等)获得实验观测数据,而且在此基础上,我们还可以获得土壤图、地质图等次生信息。随着空间插值技术的发展和完善,这些易被忽视的、形式多样、信息量丰富的软数据无疑将会得到更加充分有效的利用。

由于对地观测技术的发展迅速,未来将会出现更多的高时间分辨率、空间分辨率、光谱分辨率的多尺度观测数据。不同尺度的数据具有各自的不同特点,如大尺度具有大范围长时间的宏观优势而小尺度数据具有小区域研究的微观优势等等。在未来的研究中, 我们可以根据实际情况和应用需求,将这些不同时空尺度的软数据结合起来,并充分发挥不同尺度数据的优势。例如,通过长时间尺度的史料记载和结合当代的观测数据等,可以对地理现象(如降雨、气温变化等)的变化规律进行长时间范围的分析和预测,在短期范围内则可以利用年、月等较小时间尺度的数据进行研究分析,在此基础上还可以利用实时监测方法进行每天、每小时等微观变化的研究等等。

4.2 空间软插值方法面临的主要问题和挑战

当前的空间软插值方法虽然能够一定程度地利用软数据所隐含的信息,但是往往都只能对一种软数据进行集成,而无法同时处理多种不同属性的软数据。同时,如何结合软数据的不同尺度的优势,利用多尺度数据也是空间软插值技术面临的主要问题。再者,对地观测技术的迅猛发展使得观测数据的数据量在急剧增加[65],可获得软数据的数据量在不断上升。海量的软数据一方面可以更真实的描述地理现象,但另一方面,如何处理、计算并充分利用这些海量的软数据给空间软插值技术带来了较大的困难和挑战。

5 结 语

空间软插值方法的发展是空间插值方法与 GIS 技术和遥感理论交叉的必然趋势,有关这方面的研究在国际上起步不久,但不难预见它将成为今后若干年内相关领域的研究热点。空间软插值方法的建立和完善不仅可以拓展空间信息统计与插值理论,而且也具有较为广泛的应用前景,它有望使一些原本只能用于定性解释的遥感解译结果、地质背景以及社会经济分区等软数据参与到定量的计算中,提高统计和插值的精度和准确度,并有可能改变现有的采样方式,减少一些不必要的样本采集,从而大大地节省采集观测的费用, 同时还可能在空间软、硬数据有机结合的基础上挖掘出有价值的知识。

参考文献

[1] Lam N S N. Spatial Interpolation Methods: A Review. American Cartographer, 1983, 10(2): 129–149.

[2] Myers D E. Spatial Interpolation: An Overview. 1st Conference of the Working–Group–on–Pedometrics of the International–Society–of–Soil–Science – Pedometrics–92: Developments in Spatial Statistics for Soil Science, Wageningen, Netherlands, 1992.

[3] Jeffrey S J, Carter J O, Moodie K B, et al. Using Spatial Interpolation to Construct a Comprehensive Archive of Australian Climate Data. Environmental Modelling & Software, 2001, 16(4): 309–330.

[4] 侯景儒, 肖斌, 赵鹏大. 地质统计学新进展. 地球科学进展, 2000, 15(3): 293–296.

[5] 柏延臣, 孙英君, 王劲峰. 地统计学方法进展研究. 地球科学进展, 2004, 19(2): 268–274.

[6] 刀谓, 郭怀成, 周丰. 地统计方法学研究进展. 地理研究, 2008, 27(5): 1191–1202.

[7] Dowd P A. A Review of Recent Developments in Geostatistics. Computers & Geosciences, 1991, 17 (10): 1481–1500.

[8] Goovaerts P. Geostatistics for Natural Resources Evaluation. New York: Oxford University Press, 1997.

[9] Juang K W, Lee D Y. A Comparison of Three Kriging Methods Using Auxiliary Variables in Heavy –Metal Contaminated Soils. Journal of Environmental Quality, 1998, 27(2): 355–363.

[10] 姜勇, 李琪, 张晓珂, 等. 利用辅助变量对污染土壤锌分布的克里格估值. 应用生态学报, 2006, 17(1): 97–101.

[11] McBratney A B, Santos M L M, Minasny B, et al. On Digital Soil Mapping. Geoderma, 2003, 117(1–2): 3–52.

[12] Chilès J P, Pierre D. Geostatistics: Modeling Spatial Uncertainty. New York: Wiley–Interscience, 1999.

[13] Myers D. E. Interpolation of Spatial Data: Some Theory for Kriging. International Journal of Geographical Information Science, 2002, 16(2): 205–207.

[14] Webster R, Oliver M A. Geostatistics for Environmental Scientists. New York: John Wiley, 2007.

[15] Journel A G. Constrained Interpolation and Qualitative Information: the Soft Kriging Approach. Mathematical Geology, 1986, 18(3): 269–286.

[16] Hendriks L A M, Leummens H, Stein A, et al. Use of Soft Data in a GIS to Improve Estimation of the Volume of Contaminated Soil. Water, Air and Soil Pollution, 1996, 101: 217–234.

[17] Goovaerts P. Geostatistics in Soil Science: State–of–the–Art and Perspectives. Geoderma, 1999, 89: 1–45.

[18] Bogaert P, D'Or D. Estimating Soil Properties from Thematic Soil Maps: The Bayesian Maximum Entropy Approach. Soil Science Society of America Journal, 2002, 66: 1492–1451.

[19] Brus D J, Bogart P, Heuvelink G B M, et al. Bayesian Maximum Entropy Prediction of Soil Categories Using a Traditional Soil Map as Soft Information. European Journal of Soil Science, 2008, 59: 166–177.

[20] Seibert J, McDonnell J J. On the Dialog between Experimentalist and Modeler in Catchment Hydrology: Use of Soft Data for Multicriteria Model Calibration. Water Resources Research, 2002, 38(11): 1241–1254.

[21] D'Or D, Bogaert P. Continuous –Valued Map Reconstruction with the Bayesian Maximum Entropy. Geoderma, 2003, 112: 169–178.

[22] Serre M L, Christakos G, Lee S J, et al. Soft Data Space/Time Mapping of Coarse Particulate Matter Annual Arithmetic Average over the U.S. 4th European Conference on Geostatistics for Environmental Appli–cations, Barcelona Spain, 2002.

[23] Douaik A, Van Meirvenne M, Toth T, et al. Soil Salinity Mapping Using Spatio–Temporal Kriging and Bayesian Maximum Entropy with Interval Soft Data. Geoderma 2005, 128(3–4): 234–248.

[24] Emery X. Simulation of Geological Domains Using the Plurigaussian Model: New Developments and Computer Programs. Computers & Geosciences, 2007, 33 (9): 1189–1201.

[25] Wu C F, Wu J P, Luo Y M, et al. Statistical and Geoestatistical Characterization of Heavy Metal Concen–trations in a Contaminated Area Taking into Account Soil Map Units. Geoderma, 2008, 144(1–2): 171–179.

[26] Tan M Z, Xu F M, Chen J, et al. Spatial Prediction of Heavy Metal Pollution for Soils in Peri–Urban Beijing, China Based on Fuzzy Set Theory. Pedosphere, 2006, 16 (5): 545–554.

[27] Goovaerts P. Geostatistical Approaches for Incorporating Elevation into the Spatial Interpolation of Rainfall. Journal of Hydrology, 2000, 228(1–2): 113–129.

[28] Zhu H, Journel A G. Indicator Conditioned Estimator. Transactions, Society for Mining, Metallurgy and Exploration, Inc., 1989, 286: 1880–1886.

[29] Almeida A S, Journel A G. Joint Simulation of Multiple–Variables with a Markov–Type Coregionalization Model. Mathematical Geology, 1994, 26(5): 565–588.

[30] W Xu, T Tran, M Srivastava R, et al. Integrating Seismic Data in Reservoir Modeling: The Collocated Cokriging Alternative. Society of Petroleum Engineers, 1992, 24742: 833–842.

[31] Goovaerts P. Ordinary Cokriging Revisited. Mathematical Geology, 1998, 30(1): 21–42.

[32] Journel A G. Nonparametric Estimation of Spatial Distributions. Mathematical Geology,1983,15(3): 445–468.

[33] Jerosch K, Schluter M, Pesch R. Spatial Analysis of Marine Categorical Information Using Indicator Kriging Applied to Georeferenced Video Mosaics of the Deep–Sea Hakon Mosby Mud Volcano. Ecological Informatics, 2006, 1(4): 391–406.

[34] Goovaerts P, Journel A G. Integrating Soil Map Information in Modelling the Spatial Variation of Continuous Soil Properties. European Journal of Soil Science 1995, 46(3): 397–414.

[35] Triantafilis J, Odeh I O A, Warr B, et al. Mapping of Salinity Risk in the Lower Namoi Valley Using Non–Linear Kriging Methods. Agricultural Water Management, 2004, 69(3): 203–229.

[36] Pardo–Iguzquiza E, Dowd P A. Multiple Indicator Cokriging with Application to Optimal Sampling for Environmental Monitoring. Computers & Geosciences, 2005, 31(1): 1–13.

[37] Lyon S W, Lembo A J, Walter M T, et al. Defining Probability of Saturation with Indicator Kriging on Hard and Soft Data. Advances in Water Resources, 2006, 29 (2): 181–193.

[38] Goovaerts P, Journel A G. Integrating Soil Map Information in Modeling the Spatial Variation of Continuous Soil Properties. European Journal of Soil Science, 1995, 46(3): 397–414.

[39] Brus D J, de Gruijter J J, Walvoort D J J, et al. Mapping the Probability of Exceeding Critical Thresholds for Cadmium Concentrations in Soils in the Netherlands. Journal of Environmental Quality, 2002,31(6):1875–1884.

[40] Park N W, Chi K H, Kwon B D, et al. Geostatistical Integration of Spectral and Spatial Information for Land–Cover Mapping Using Remote Sensing Data. Geosciences Journal, 2003, 7(4): 335–341.

[41] Ungaro F, Ragazzi F, Cappellin R, et al. Arsenic Concentration in the Soils of the Brenta Plain (Northern Italy):Mapping the Probability of Exceeding Contamination Thresholds. Journal of Geochemical Exploration, 2008, 96 (2–3): 117–131.

[42] Zhu H, Journel A G. Formatting and Integrating Soft Data –Stochastic Imaging Via the Markov–Bayes Algorithm. 4th International Geostatics Congress : Troia 92, Troy Portugal, 1992.

[43] Deutsch C V, Journel A G. Gslib: Geostatistical Software

Library and User's Guide. New York: Oxford University Press, 1998.

[44] Stein A, Hoogerwerf M, Bouma J, et al. Use of Map-Delineation to Improve Co-Kriging of Point Data on Moisture Deficits. Geoderma, 1988, 43: 311-325.

[45] Voltz M, Webster R. A Comparison of Kriging, Cubic-Splines and Classification for Predicting Soil Properties from Sample Information. Journal of Soil Science, 1990, 41(3): 473-490.

[46] Vanmeirvenne M, Scheldeman K, Baert G, et al. Quantification of Soil Textural Fractions of Bas-Zaire Using Soil Map Polygons and or Point Observations. Geoderma, 1994, 62: 69-82.

[47] Voltz M, Lagacherie P, Louchart X, et al. Predicting Soil Properties over a Region Using Sample Information from a Mapped Reference Area. European Journal of Soil Science, 1997, 48(1): 19-30.

[48] Lagacherie P, Voltz M. Predicting Soil Properties over a Region Using Sample Information from a Mapped Reference Area and Digital Elevation Data: A Conditional Probability Approach. Geoderma, 2000,97(3-4):187-208.

[49] Boucneau G, Van Meirvenne M, Thas O, et al. Integrating Properties of Soil Map Delineations into Ordinary Kriging. European Journal of Soil Science, 1998, 49: 213-229.

[50] Liu T L, Juang K W, Lee D Y, et al. Interpolating Soil Properties Using Kriging Combined with Categorical Information of Soil Maps. Soil Science Society of America Journal, 2006, 70(4): 1200-1209.

[51] Hengl T, Heuvelink G B M, Stein A, et al. A Generic Framework for Spatial Prediction of Soil Variables Based on Regression-Kriging. Geoderma, 2004, 120(1-2):75-93.

[52] Hudson G, Wackernagel H. Mapping Temperature Using Kriging with External Drift - Theory and an Example from Scotland. International Journal of Climatology, 1994, 14 (1): 77-91.

[53] Bourennane H, King D, Chery P, et al. Improving the Kriging of a Soil Variable Using Slope Gradient as External Drift. European Journal of Soil Science, 1996, 47 (4): 473-483.

[54] Monestiez P, Allard D, Navarro Sanchez I, et al. Kriging with Categorical External Drift: Use of Thematic Maps in Spatial Prediction and Application to Local Climate Interpolation for Agricultrure. geoENV98 - the Second European Conference on Geostatistics for Environmental Sciences, November 1998. Gómez-Hernández J, Soares AFroidevaux R. Valencia, Spain Springer: 163-174.

[55] Christakos G. A Bayesian Maximum Entropy View to the Spatial Estimation Problem. Mathematical Geology, 1990, 20: 763-787.

[56] Bogaert P. Spatial Prediction of Categorical Variables: The Bme Approach [C]. 4th European Conference on Geostatistics for Environmental Applications, Barcelona, Spain, 2002.

[57] Bogaert P. Spatial Prediction of Categorical Variables: The Bayesian Maximum Entropy Approach. Stochastic Environmental Research and Risk Assessment, 2002, 16: 425-448.

[58] Lee S J, Wentz E A. Applying Bayesian Maximum Entropy to Extrapolating Local-Scale Water Consumption in Maricopa County, Arizona. Water Resources Research, 2008, 44: W01401.

[59] Ohlmacher G C, Davis J C. Using Multiple Logistic Regression and Gis Technology to Predict Landslide Hazard in Northeast Kansas, USA. Engineering Geology, 2003, 69(3-4): 331-343.

[60] Giasson E, Clarke R T, Inda A V, et al. Digital Soil Mapping Using Multiple Logistic Regression on Terrain Parameters in Southern Brazil. Scientia Agricola, 2006, 63 (3): 262-268.

[61] Wang H B, Sassa K, Xu W Y. Assessment of Landslide Susceptibility Using Multivariate Logistic Regression: A Case Study in Southern Japan. Environmental & Engineering Geoscience, 2007, 13(2): 183-192.

[62] Brunsdon C, Fotheringham S, Charlton M, et al. Geographically Weighted Regression-Modelling Spatial Non-Stationarity. The Statistician, 1998, 47: 431-443.

[63] Tu J, Xia Z G. Examining Spatially Varying Relationships between Land Use and Water Quality Using Geographically Weighted Regression I: Model Design and Evaluation. Science of the Total Environment, 2008, 407 (1): 358-378.

[64] Kasimov N, KoSheleva N, Wagner V, et al. Modeling Geochemical Fields Based on Landscape-Guided Interpolation. Ecological Modelling, 2008, 212(1-2): 109-115.

[65] **周成虎, 骆剑承, 等. 高分辨率卫星遥感影像地学计算. 北京: 科学出版社**, 2009.

Review on Soft Spatial Data and its Spatial Interpolation Methods

LUO Ming[1,2], PEI Tao[1,3]

(1. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China;
3. Key Laboratory of Geographic Information Science, Ministry of Education, Shanghai 20062, China)

Abstract: In recent years, as the observation technologies develop rapidly, both type and number of spatial data is increasing, and information retrieved from spatial data expands increasingly, among which includes a large number of qualitative information, for instance, land-use type data, vegetation type data, topographic feature data, which some experts called soft information or soft data. These so-called soft data often have associations with the predicted target variable, even could become one of most important factors that influence the spatial distribution of target variable obviously in some cases, therefore, they can help improve prediction of target variable theoretically. However, in respect that non-numerical soft data can't be calculated directly and is neglected by traditional spatial interpolation methods, connotative useful information can not be utilized sufficiently and effectively, which results in a mass of wasted information. Lately, soft spatial interpolation technology was proposed, aimed to integrate soft spatial data as auxiliary or second information to help improve interpolation accuracy. According to the characteristics and categories of soft spatial data, this paper aimed to review on soft spatial interpolation methods and their applications. Firstly, we summarized some "harden" methods, hardening the soft spatial data to hard data. Then, we discussed several different type soft spatial interpolation methods afterward, such as simple kriging, cokriging, indicator kriging, ordinary kriging, stratified kriging, kriging with external drift regression, bayesian maximum entropy, inverse distance weighted. After that, prospects of application of soft data and soft spatial interpolations were proposed in the last part.

Key words: soft data; spatial interpolation; geostatistics; kriging; regression; bayesian maximum entropy