文章编号: 1007-4619(2008) 05-0707-09

基于分区的局域神经网络时空建模方法研究

王海起[1, 2], 王劲峰[2]

(1. 中国石油大学 (华东)地球资源与信息学院, 山东 东营 257061;
2. 中国科学院 地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101)

摘 要: 区域数据表现为两种尺度的空间特性: 反映全局特征的空间依赖性和反映局域特征的空间波动性。空间波动性表现为空间数据在局部地区的聚集或高低交错现象。在研究区域数据时空预测性建模时, 从降低数据的空间波动和不平稳性对模型预测能力的影响角度出发, 提出了一种基于分区的局域神经网络时空非线性建模的思路。分区过程由基于空间邻接关系的 *K*-means聚类算法完成。不同的分区方案通过相关性、波动性、紧凑性等指标进行评价和优选。在确定最优分区方案的基础上, 对各子区分别采用两层前馈网络进行建模, 模型的输入不仅要考虑本区内单元的作用, 而且要考虑相邻子区的边界效应。各神经网络模型的时空预测能力通过平均相均差和动态相似率等指标进行衡量。最后, 通过对法国 94个县每周流感报告病例的时空建模分析表明, 与全局神经网络模型相比, 基于分区的局域神经网络模型具有更好的预测能力。

关键词: 格数据; 时空建模; 分区; *K*-means聚类; 神经网络; 边界效应
中图分类号: P208 **文献标识码:** A

1 引 言

在地理信息科学领域, Cressie将空间数据分为3种主要类型: 点模式数据 (point pattern data) 、地学统计数据 (geostatistics data) 和格数据 (lattice data)[1]。格数据, 也称为区域数据, 是指属性数据关联于固定多边形区域的数据类型, 其多边形区域既可以是规则的也可以是不规则的。区域数据分析侧重于对区域空间格局或趋势的探测、建模和解释; 区域数据时空分析研究在空间拓扑结构 (区域单元的空间排列、形状、大小等)保持不变的情况下, 属性数据随时间变化的时空过程和时空格局的探测、建模和预测。

区域时空数据具有空间和时间两方面的属性。从空间角度出发, 一方面, 正如 Tobler地理学第一定理阐述的观点, 空间对象呈现出相互依赖的空间格局, 并且这种相互依赖性随着空间对象之间距离的增加而减弱。另一方面, 假定空间结构在研究区域内具有平稳性是不现实的, 特别在空间单元数目较多的情况下, 它表现为在局部区域的高值 ("热点 ")和低值 ("冷点 ")聚集或异常, 从其潜在的空间运动过程来讲, 这种现象是由于局部地区空间过程的异质性和空间相互作用的程度不同造成的。

从时间角度出发, 正如单纯的时间序列分析, 估计 t时刻某一区域的属性值依赖于 t时刻之前同一区域的属性值, 对于一些在时间上具有马尔可夫性的过程, 如某些传染病的时空过程建模, 甚至仅需要考虑 $t-1$时刻的属性值。然而, 更重要的是, 如果忽视 t时刻之前其他空间关联区域的作用, 将严重影响时空建模的可信度和适用性。

对于区域数据的时空线性关系建模, 已经发展了一些时空线性回归模型[2]。例如, 时空自回归模型 STAR: $y_t = \rho \boldsymbol{W} y_{t-1} + \beta y_{t-1} + \varepsilon_t$, 通过引进空间权重矩阵 $\boldsymbol{W}$, 不仅考虑了 $t-1$时刻同一区域的属性值, 而且考虑了空间相邻区域属性值的作用。实际应用也表明, 对于区域数据时空线性建模, STAR模型的拟合与预测能力均优于单纯的时间序列模型

收稿日期: 2006-12-04; 修订日期: 2007-09-04
基金项目: 国家自然科学基金项目 (编号: 40471111)、国家 863计划项目 (编号: 2006AA12Z215)及中国石油大学 (华东)博士基金项目 (Y060124)共同资助。

作者简介: 王海起 (1972—), 男。讲师, 博士。主要研究方向为地学空间信息分析与智能计算。E-mail: wanghq@ lreis. ac. cn。

如 ARMA、ARMA 模型[3]。

实际上, 现实中的区域数据常常表现出非线性、复杂性等特点, 难以用简单的线性方程进行建模和逼近, 这时采用一些较为复杂的建模手段, 如神经网络模型, 也许能提高时空预测的结果。

神经网络 (artificial neural networks, 简称 NN) 模型应用于区域数据时空建模和预测有其自身的不足, 一方面, 大部分 NN 模型都是一种“黑箱”结构, 选择合适的网络结构和训练算法往往非常困难, 如果不考虑研究对象的先验知识, 应用 NN 进行建模常常导致错误的结果; 另一方面, 大部分神经网络算法的调整主要基于机器学习理论的角度, 很少基于研究问题的领域知识, 导致对其预测的结果往往难以进行解释。因此, 如果对于研究问题存在简单且有效的方法时, 没有必要使用如此复杂的模型; 然而, 当研究对象比较复杂且简单的方法不再适用时, 在牺牲模型易于解释的代价基础上, 为了得到更好的预测结果, 可以考虑使用 NN 模型[4]。

本文研究神经网络模型在区域数据时空分析中的应用。基于对空间格局局部不稳定的考虑, 首先采用基于空间邻接关系的改进 K-means聚类算法对研究区域的空间单元进行分区划分, 提出了利用全局和局部 Moran's I 统计量的关系对不同分区方案进行定量评价的方法; 在选择合适分区的基础上对各子区分别采用一个基于 BP 训练算法的多层前馈网络进行时空建模与预测。同时对整个研究区域建立一个全局的多层前馈网络模型。为了比较全局和局域 NN 模型效果, 采用法国 1990年第 1周至 1992年第 53周 3年共 157周 94个县流感报告病例数进行实例分析, 以第 t 周各区患病人数作为输入数据, 以第 $t+1$ 周各区患病人数作为预测数据。

2 区域数据的全局和局域 Moran's I 统计量

从空间数据探索性分析角度出发, 空间数据可认为由两部分组成[5]:

$$\text{Spatial data} = \text{spatial smooth} + \text{spatial rough} \quad (1)$$

空间趋势或平滑 (spatial smooth) 反映空间数据的全局或整体特征, 对于区域数据, 它与全局 (大尺度) 的空间自相关模式相关。空间正相关性表明研究区域内空间单元属性与其相邻空间单元具有相同的变化趋势 (同为高值或低值); 负相关性表明研究区域内空间单元属性与相邻空间单元具有相反的变化趋势; 无相关性则表明空间单元的属性值彼此相互独立, 在空间上随机分布。

空间波动 (spatial rough) 反映空间数据的局部特征。对于区域数据, 局域正相关性表现为单个空间单元与其相邻单元属性具有相同的趋势, 同为高值 (可称为“热点”) 或同为低值 (可称为“冷点”); 局域负相关性表现为单个空间单元与其相邻区域具有相反的趋势, 为高低或低高交错 (可称为“异常”)。

用于探测区域数据这两个不同尺度空间格局的全局和局域统计量包括 Moran's I, Getis' G 和 Geary's c 统计量等。

2.1 全局统计量 GlobalMoran's I

用于探测区域数据的全局空间自相关性, 其公式如下[6]:

$$I = \frac{n \cdot \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\left|\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right| \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (2)$$

式中, x_i 是空间单元 i 的属性值, $\bar{x}$ 是研究区域 n 个空间单元的属性平均值, w_{ij} 是空间权重矩阵 $\boldsymbol{W}$ 的元素, 定义了单元 i 与单元 j 的相关关系。

当 Moran's I 值为正数且显著时表明存在空间正相关性; 当 Moran's I 值为负数且显著时表明存在空间负相关性; 当 Moran's I 近似为零时表明为空间随机分布。

2.2 局域统计量 LocalMoran's I

Anselin将其称为 LISA, 即空间关联局域指标 (local indicator of spatial association), 对于空间单元 i 其公式为[7]:

$$I_i = \frac{n \cdot (x_i - \bar{x}) \sum_{j=1,j\neq i}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3)$$

当 I_i 为正数且显著时, 表明空间单元 i 与其相邻单元存在高值或低值的聚集现象; 当 I_i 为负数且显著时, 表明空间单元 i 与其相邻单元存在高低或低高交错现象; 否则表明存在局部的空间随机现象。

2.3 全局和局域 Moran's I 之间的关系

当空间权重矩阵 $\boldsymbol{W}$ 采用空间邻接形式, 即若单

元 i与单元 j具有共同边界则 $w_{ij}=1$, 否则 $w_{ij}=0$, 并且矩阵 W是行标准化形式 (每行元素之和为 1)时, 公式 (2)与公式 (3)之间的关系可表达为[7]:

$$I=\frac{1}{n}\sum_{i=1}^{n}I_i \tag{4}$$

公式 (4)表明, 对于一个研究区域, 局域 Moran's I的平均值即是全局 Moran's I值。因此, 当整个区域的空间过程较为平稳或空间波动不明显时, 可以预期局域 Moran's I值围绕全局 Moran's I值的波动较小; 反之, 当空间过程不平稳或空间波动较明显时, 局域 Moran's I与全局 Moran's I具有较大的差异。

因此, 可以用 Local Moran's I与其平均值 GlobalMoran's I的标准偏差作为度量一个区域波动程度 (或平稳性)的指标, 其公式如下:

$$\mathrm{Std}(I)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(I_i-I)^2} \tag{5}$$

在下一节中, Std(I)将作为不同分区方案是否稳定的一个评价指标。一个子区的 Std(I)值相对较小, 表明该子区的平稳性较好、波动较弱; 否则, 表明该子区的平稳性较差、波动较强。

3　分区标准

类似于地理学中的区划思想, 分区的目的是根据一组评价标准利用空间单元的属性数据对研究区域的单元进行划分, 使所有空间单元归到不同的子区中。对具有时空属性数据的区域单元进行分区时, 可以将同一属性在不同时期的观测值作为不同的属性来对待, 如: 某县报告的 12个月每月流感患病人数可以看作 12个不同的属性数据。

Cliff等人给出了一个最佳的 ("optimal")区划方案在一般情况下应满足的 3个标准: 简洁性 (simplicity)、均质性 (homogeneity)和空间紧凑性 (compactness)[8, 9]。根据局域神经网络建模的需要, 综合上述区划标准以及另外的两个附加标准来构建我们的分区评价指标。

(1) 简洁性 (simplicity)

对于局域时空建模, 需要对分区方案中的每个子区分别建立一个模型, 子区数目较少, 需要建立的模型及相应的计算量就较少, 对于 NN 模型来说用于模型学习时间的减少更为明显。因此, 当其他分区标准难以确定不同分区方案的优劣时, 分区数目较少的方案总是优于分区数目较多的方案。

(2) 邻接性 (contiguity)

邻接性意味着在分区时, 只有空间相邻单元才能归到同一子区中。邻接性的考虑将通过对 K-means聚类方法的改进而得以实现。

(3) 紧凑性 (compactness)

紧凑性关注于各子区的空间形状, 它保证在分区结果中总是优先考虑那些相邻空间单元距离较近的方案, 一些学者认为空间紧凑性与我们对社会和经济活动的"直观理解"是一致的[10]。对于一个子区, 通过计算该子区质心与子区包含的各空间单元质心的平均距离 (也称为离散度)作为衡量紧凑性的性能指标, 该公式如下:

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n}\sqrt{(x_i-x_g)^2+(y_i-y_g)^2} \tag{6}$$

式中, n是子区包含的空间单元数目; x_i, y_i是空间单元 i的质心坐标; x_g, y_g是该子区的质心坐标。如果离散度较小, 则表明该子区内的单元较为紧凑; 否则, 表明该子区较为松散。

(4) 相互依赖性 (interdependence)

相互依赖性保证一个子区内的空间单元之间具有关联性。对于基于分区的局域神经网络时空建模, 各空间单元 t时刻某个属性值的预测 (作为输出层节点)实际上是利用 t时刻之前各单元的属性值 (作为输入层节点)之间的相互作用 (通过隐含层)实现的, 因此, 分区方案必须保证子区内的空间单元存在实际的空间相互作用或依赖性。每个子区的空间依赖性可利用该子区的全局 Moran's I系数进行评估。

(5) 不平稳性 (instability)

虽然使每个子区具有完全平稳性是不可能的, 但是不平稳性越低意味着 NN 模型的预测效果越好[4]。各子区的平稳性由前述的 Std(I)指标衡量。

由上述分析可以看出, 在对不同的分区方案进行评价时, 各分区指标起到的作用是不同的。首先, 相互依赖性是必要条件, 不满足此条件的分区结果不能使用; 其次, 不平稳性是优先考虑的指标, 对于那些 Std(I)相差无几的方案, 可进一步利用离散度进行评价; 最后, 简洁性是可选标准, 在随后的应用实例中并没有使用。

需要指出的是, NN 模型研究的是区域时空数据, 而上述分区标准 (4)、(5)涉及的 Moran's I系数只是纯粹的空间相关性指标, 因此对时空数据采用空间 Moran's I指标衡量并不合适。然而, 从已有文献的检索中并没有找到合适的可用于区域时空

相关性计算和检验的全局和局域时空统计量, 这里, 采用一种折衷的方法, 对于研究的时空变量(如: 不同时间不同县的流感患病人数), 构造一个统计量, 使得对于每个空间单元, 该统计量是研究变量的不同时间观测值的函数(如: 不同时期流感病例的平均值、总和或最大值等)。

4 基于空间邻接关系的 K-means聚类方法

聚类方法是将研究对象按照其特征分组为多个类, 使每个类对象之间具有较高的相似性, 而不同类对象之间的差别较大。K-means 方法由 MacQueen于 1967年提出, 是目前应用最为广泛的一种聚类方法。

利用聚类方法对空间单元进行分区时, 要求同一子区的单元在空间上处于相邻的位置, 在地图上表现为彼此相连的状态。而传统的聚类方法仅利用空间单元的属性数据, 并没有考虑单元的空间邻接关系。有研究对空间单元聚类时, 将单元的空间坐标作为额外的属性变量加以考虑, 然而这种方法得到的同一子区的单元仍然可能出现在空间不相邻的位置[10]; 也有研究提出了新的空间单元分区方法[11, 12]。

本文利用 K-means聚类方法对空间单元进行分区, 在分区过程中将空间邻接关系作为约束条件加以考虑。在对每个空间单元进行类别归属判断时, 不仅要考虑单元与某类别中心的距离, 而且要考虑单元与该类别中空间单元的邻接关系; 只有当该类别与进行归属判断的空间单元之间存在邻接关系且距离最短时, 单元才可以归属于该类。这样, 对于最终的分区结果, 既保证了同一类单元的属性值差别较小、不同类之间属性值差别较大, 又保证了同一类的空间单元在空间上处于相邻的位置。

4.1 相关定义

首先对研究的区域时空对象作如下定义:

(1) 设研究区域 S 有 N 个空间单元 $S=\{s_1, s_2, \cdots, s_N\}$及邻接关系(neighbor relation)$R\subseteq S\times S$。空间单元 s_i 和 s_j 具有邻接关系当且仅当 $(s_i, s_j)\in R$, $i\neq j$。用空间邻接矩阵 $\boldsymbol{W}$ 表达邻接关系 R, $W(i, j)=W_{ij}=1$ 当且仅当 $(s_i, s_j)\in R$, 否则 $W(i, j)=W_{ij}=0$。

(2) 对每个单元 s_i, 设研究的单元时空属性变量为 $X_i\equiv X(s_i)=[x_{i1}, x_{i2}, \cdots, x_{iT}]$, T 是时间维的长度。

(3) 对每个单元 s_i, 构造一个统计量 $Q_i\equiv Q(s_i)=f(X_i)=f(x_{i1}, x_{i2}, \cdots, x_{iT})$, Q_i 是时空属性变量 X_i 的函数, 用于评价分区结果的空间 Moran's I 系数计算。

其次, 对于 K-means聚类算法作如下定义:

(1) 定义 $\{z_1, z_2, \cdots, z_K\}$为 K 个聚类中心, 每个聚类中心 $z_j=[z_{j1}, z_{j2}, \cdots, z_{jT}]$ $(j=1, 2, \cdots, K)$。

(2) 对每个聚类中心 z_j 定义一个集合 Z_j, 用于存放该类别中包含的空间单元, 初始化时集合 Z_j 为空。

(3) 定义 $N\times K$ 的二维距离矩阵 Dist, 用于存放每个空间单元与每个聚类中心的距离。同时定义矩阵 Dist的 $N\times K$ 辅助逻辑矩阵 DisMark, 用于标识在距离矩阵 Dist中搜索单元到聚类中心的最短距离时该距离是否参与搜索过程, 若矩阵 DisMark中某元素值为 True, 则矩阵 Dist中对应距离参与搜索, 否则不参与搜索。

4.2 算法流程

基于空间邻接关系的 K-means聚类算法详细流程请参考文献[13]。

利用该算法, 通过指定不同类别数 K, 可以得到不同 K 值的分区方案。对不同分区方案, 利用分区标准进行优选。

首先, 对指定类别数为 K 的分区方案的各个子区, 分别将各子区作为单独研究区域计算其统计量 Q 的全局 Moran's I 系数, 若存在没有空间相关性或相关性不显著的子区, 则类别数为 K 的分区方案将被淘汰; 其次, 对通过相关性检验的每个 K 类分区方案, 再分别以各子区作为单独研究区域计算各自的平稳性指标 Std(I)和离散度指标 $\bar{d}$, 将各子区指标的平均值作为每个 K 类分区方案的平稳性和紧凑型的指标结果; 最后, 从中选择平稳性最好(Std(I)值最小)、离散度最小的分区方案作为最终的分区结果。

5 神经网络建模及其边界效应

5.1 神经网络模型

在确定最终分区方案的基础上, 可以对各子区分别进行神经网络时空建模和预测。由于多层前

馈网络模型可以对任意的输入-输出映射进行建模并在实际应用特别是预测问题中得到了广泛的应用, 并且理论已经证明: 具有单隐层的前馈模型可以任意的精度逼近任意复杂的非线性函数, 因此, 采用两层前馈网络 (包括隐含层、输出层, 不包括输入层)进行建模。

对于区域单元时空预测性建模, 模型输出是 t 时刻各单元的预测值 X_t, 输入是 t时刻之前相关时段各空间单元的观测值, 神经网络建立如下的函数映射关系:

$$X_t = f(X_{t-1}, X_{t-2}, \cdots, X_{t-p}) \qquad (7)$$

因此, NN模型实际上是一个非线性的时空自回归模型。p 是时间滑动窗口步长, 用于决定建模时的相关时间滞后项, 例如, 对于 T 个时间的观测向量 $X_1, X_2, \cdots, X_T$, 每个 $X_t = [x_{t1}, x_{t2}, \cdots, x_{tN}]$表示空间 N 个单元的观测值, 进行一步预测 (1-step-ahead) 时, 第 1个输入-输出模式的输入是 $X_1, X_2, \cdots, X_p$; 预期输出是 X_{p+1}; 第 2个输入-输出模式的输入是 $X_2, X_3, \cdots, X_{p+1}$; 预期输出是$X_{p+2}, \cdots$, 最后, 第 $T-p$ 个输入-输出模式的输入是 $X_{t-p}, X_{t-p+1}, \cdots, X_{T-1}$、预期输出是 X_T。

目前, 滑动窗口步长 p 的确定并没有合适的方法, 有研究利用线性关系的时空自相关函数和时空偏自相关函数来确定时间阶数 p, 也有学者认为这种方法对于神经网络的非线性滞后并不合适[14]。实际使用时, 常采用多次试验 (try-and-error) 的方式。

NN模型的性能评价主要通过检验数据集衡量所建立的模型对于新输入的预测能力, 即泛化能力, 主要包括平均相均差 ARV 和动态相似率 DSR两个指标[15], 前者反映模型预测输出的准确程度, 后者反映模型预测的趋势与实际趋势的接近程度。

5.2　边界效应

采用分区的思路进行局域神经网络建模, 并不表明不同子区的空间单元之间没有关联性, 相反, 可能存在着其他形式的相关关系, 如经济、交通和人口等形式, 而这些形式的相关关系并不能被简单的空间邻接矩阵所表现和度量, 因此, 在建模时如果仅考虑子区内的单元对模型输出的影响, 而忽视子区周围单元的影响因素, 等于人为“割裂”了不同空间区域单元之间的相互联系和空间依赖关系, 这与地理学第一定理是相违背的, 模型的结果也是令人难以接受的。因此, 采用对各个子区分别进行 NN非线性建模时, 不仅要考虑子区内各单元的观测值对模型预测结果的作用, 而且应引入与其相邻的空间区域的作用因素, 即边界效应。

这里, 把与子区直接相邻 (边相邻或顶点相邻)的边界空间单元 t时刻之前的观测值也作为 NN模型的输入加以考虑, 这样, 局域 NN模型输出的各单元 t时刻预测结果不仅是子区内各单元 t 时刻之前观测值的函数, 而且是其周围边界单元 t 时刻之前观测值的函数。对于一步预测建模, 若设一个子区的单元数目为 n, 与其相邻的单元数目为 m, 时间滑动窗口步长为 p, 那么, 该子区的局域 NN模型的输入层节点个数为 $(n+m)\times p$, 输出层节点个数为 n。

6　应用实例

研究数据采用法国 94个县的每周流感报告病例[16], 时间为 1990年第 1周至 1992年第 53周共 157周, 图 1(a)为法国 94个县的数字编号。

以每周流感平均患病人数构造空间统计量 Q, 其空间分布见图 1(b)。空间邻接矩阵 $\boldsymbol{W}$ 采用边界直接相邻的一阶形式, 根据各县每周平均病例计算的 94个县全局 Moran's I= 0.1281, 假设检验表明流感病例具有显著的空间正相关 (图 2), 说明法国各县流感具有空间自相关性, 而且呈现出高发区与高发区相邻、低发区与低发区相邻的空间格局。

6.1　分区

以每周流感病例作为各县的属性数据, 各空间单元分别具有 157个属性数据, 以一阶邻接矩阵 $\boldsymbol{W}$ 作为约束条件, 对法国 94个县进行 K-means聚类分区。由于事先无法确定聚类的类别数 K, 依次取 K 值为 4—16之间的数值, 分别进行聚类计算, 通过不同分区方案的相关性检验, 最后具有显著空间相关性的类别数 K 分别为 8, 9, 10, 12, 14, 16, 分别计算这 6个不同分区方案的 Std(I)指标, 离散度指标 (表 1), 最终选择的最优类别数 K = 12。

分区数为 12的各子区空间分布见图 1(c), 从图 1(b)与图 1(c)的对比可以看出, 最终的分区方案也反映了流感病例的空间分布格局。表 2给出了

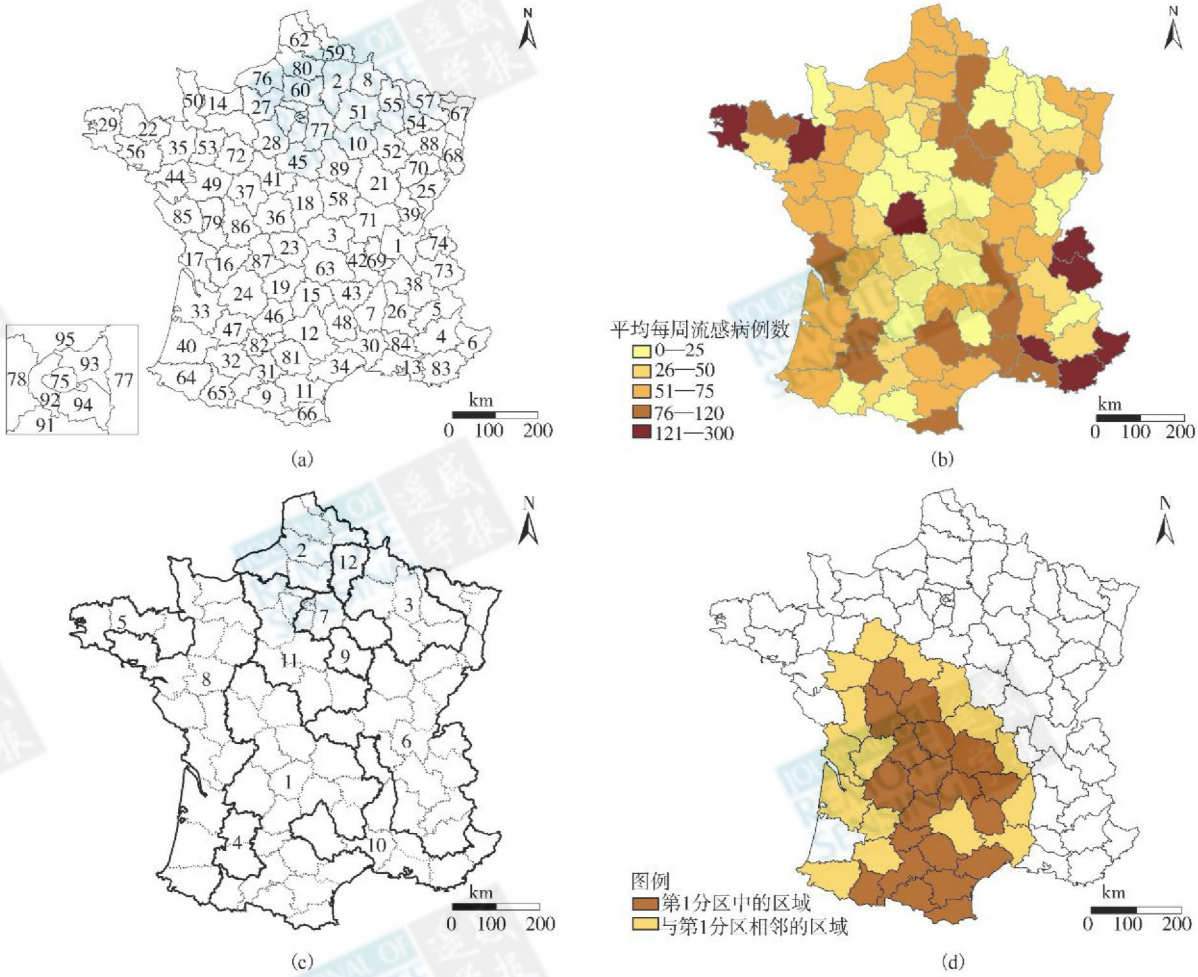

图 1 法国 94个县流感病例分区情况

(a) 94个县数字编号; (b) 94个县 1990年第 1周至 1992年第 53周平均每周流感报告病例分级图; (c) 类别数为 12的分区结果; (d)第 1分区及相邻的边界区域单元

Fig. 1 Partitioning for flu cases of 94 counties in France

(a) number IDs of 94 counties; (b) average weekly flu cases of 94 counties from the 1st week of 1990 to the 53th week of 1992; (c) the partition map of K = 12 for 94 counties; (d) the first subarea and its neighboring regions

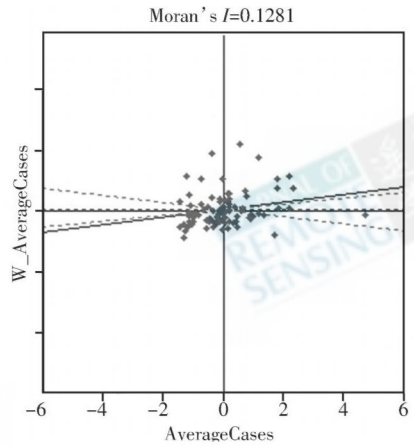

图 2 法国 94个县平均每周流感病例的全局 Moran's I散点图

Fig. 2 Global Moran's I of average weekly flu cases for 94 counties in France

分区数为 12的各子区包含的空间单元数目, 各自的空间相关性指标全局 Moran's I值和相应的 Z得分检验值, 平稳性指标 Std(I)值。

表 1 6个不同类别数 K 的分区方案评价指标结果

Table 1 The results of partition criteria for six partition schemes

类别数 K	平稳性指标 Std(I)	离散度指标 $\bar{d}$
8	0.6676	897.68
9	0.6614	612.34
10	0.6589	1434.10
12	0.5251	1030.27
14	0.5884	964.76
16	0.5345	664.39

表 2　分区数为 12的各子区相关指标结果

Table 2　Relevant statistic of each subarea in the partition scheme of K = 12

	空间单元数	全局 Moran's I	Z 得分检验	Std(I) 值
第 1子区	20	- 0.3392	- 1.9799	0.4227
第 2子区	5	0.7302	2.9224	0.4193
第 3子区	6	0.2141	2.0218	0.2792
第 4子区	2	- 1	- ∞	0
第 5子区	3	- 1.9190	- 4.0135	0.3422
第 6子区	18	0.2506	1.9725	0.6780
第 7子区	7	0.3948	2.7197	0.6516
第 8子区	15	0.3195	1.9934	0.4449
第 9子区	1			
第 10子区	8	0.5326	2.1490	0.6662
第 11子区	8	0.4070	2.2162	0.4026
第 12子区	1			

6.2　神经网络建模

对 12个子区分别建立一个神经网络模型, 每个局域 NN模型利用第 $t-1$ $t-2$ …, $t-p$ 周各县的流感病例, 预测第 t 周本子区各县的流感患病人数。由于流行性感冒的传染期约为 1周, 对于以周为时间单位的建模, 输入可以仅考虑第 $t-1$ 周的病例, 即 $p=1$。

因此, 各 NN模型的输出层节点数等于本子区包含的空间单元个数, 输入层节点数为本区单元个数与周围边界单元个数之和, 图 1(d)为第 1子区及其相邻的边界单元, 表 3为各子区 NN模型的输入层、输出层节点数。

为了比较局域 NN模型的效果, 同时对整个研究区域 94个县建立一个全局 NN模型, 输入数据为第 $t-1$ 周各县流感病例, 预期输出为第 t 周各县流感病例, 即输入与输出节点数均为 94。

表 3　各子区 NN模型的输入层和输出层节点数

Table 3　The number of input nodes and output nodes of each NN model for twelve subareas

	输入节点数	输出节点数
第 1子区	36	20
第 2子区	9	5
第 3子区	15	6
第 4子区	10	2
第 5子区	8	3
第 6子区	32	18
第 7子区	17	7
第 8子区	28	15
第 9子区	6	1
第 10子区	22	8
第 11子区	24	8
第 12子区	7	1

针对上述的局域和全局 NN模型, 将各子区 156个观测数据对 (X_{t-1}, X_t), 其中 X_{t-1}, X_t 分别是 $t-1$ 时刻, t 时刻空间单元的观测向量; 按 90%∶10%比例随机分为训练集 (train dataset)和检验集 (test dataset), 训练集为 140对样本, 检验集为 16对样本。采用 BP算法进行模型训练与调整, 再利用检验集对模型进行检验。最后由各局域和全局 NN模型得到的法国 94个县各县检验数据的平均相均差 ARV、动态相似率 DSR指标结果见图 3和图 4(横

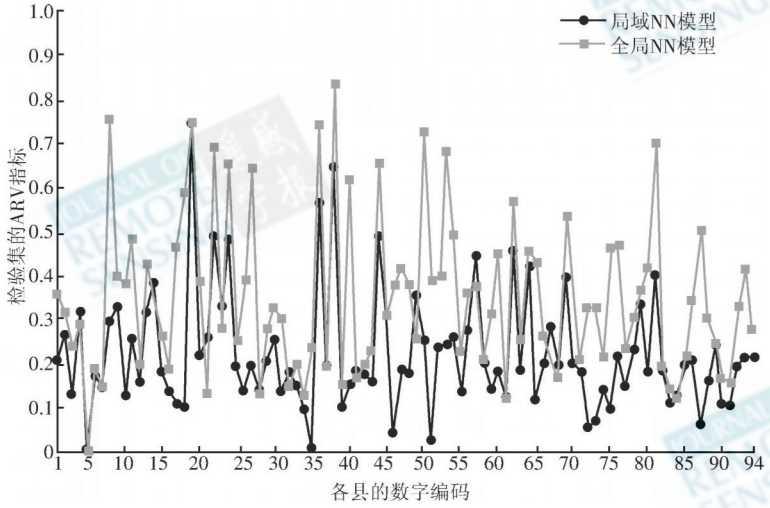

图 3　法国 94个县全局和局域 NN模型检验数据集的平均相均差对比图

Fig 3　Test dataset's ARV of local and global NN model for 94 counties in France

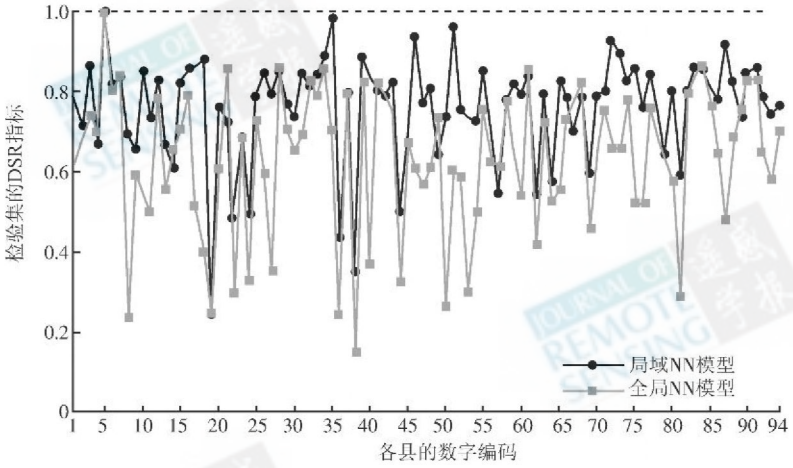

图 4 法国 94个县全局和局域 NN模型检验数据集的动态相似率对比图

Fig. 4 Test dataset's DSR of local and global NN model for 94 counties in France

坐标是各县的数字编号,与图 1(a)对应)。可以看出,基于分区的局域神经网络模型的预测能力明显优于全局 NN模型。

7 结论与讨论

针对 GIS格数据时空非线性建模,从降低数据的空间波动和不平稳性对模型预测能力的影响角度出发,提出了一种基于分区的局域神经网络建模的思路,分区的目的是使在全局尺度上表现为空间波动的局部区域,通过分区在较小尺度上表现为较强的空间相关性和较弱的空间波动性。

需要指出的是,在采用 *K*-means聚类算法进行分区时,初始聚类中心的选择对最终的分区结果具有重要的影响,对于空间聚类,随机选择初始聚类中心并不是一个合适的方法,进一步的研究应结合研究区域的空间格局,如:考虑局部的"热点"或"冷点"区域,使初始聚类中心的确定与空间格局建立联系。另外,对于分区标准,有必要进一步细化研究,对于不同类型的空间过程可能会有不同的评价标准,对于反映空间波动和不稳定性的指标需作更深入的分析,对于区域时空过程,研究相应的时空评价指标和检验方法更是势在必行。

参考文献 (References)

[1] Cressie A C. Statistics for Spatial Data[M]. New York: Wiley, 1991.

[2] Kamarianakis Y. Spatial-Time Series Modeling: A Review of the Proposed Methodologies[A]. Proceedings of the 8th AGILE Conference on GIScience[C]. Portugal, 2005.

[3] Han W G. Data-Driven and Model-Driven Spatio-Temporal Data Mining[D]. Institute of Geographic Sciences and Natural Resources Research, CAS: Graduate School of Chinese Academy of Sciences, 2005. [韩卫国. 数据驱动和模型驱动的时空数据挖掘[D]. 中国科学院地理科学与资源研究所: 中国科学院研究生院博士学位论文, 2005.]

[4] Gilardi N, Bengio S. Local Machine Learning Models for Spatial Data Analysis[J]. *Journal of Geographic Infomation and Decision Analysis*, 2000, **4**(1): 11—28.

[5] Haining R. Spatial Data Analysis: Theory and Practice[M]. London: Cambridge University Press, 2003.

[6] Anselin Luc. Spatial Econometrics: Methods and Models[M]. Dordrecht: Kluwer Academic, 1988.

[7] Anselin Luc. Local Indicators of Spatial Association-LISA[J]. *Geographical Analysis*, 1995, **27**(2): 93—115.

[8] Haining R, Wise S, Ma J. Designing and Implementing Software for Spatial Statistical Analysis in a GIS Environment[J]. *Journal of Geographical Systems*, 2000, **2**: 257—286.

[9] Cliff A D, Haggett P, Ord J K, *et al*. Elements of Spatial Structure: A Quantitative Approach[M]. London: Cambridge University Press, 1975.

[10] Wise S, Haining R, Ma J. Regionalization Tools for the Exploratory Spatial Analysis of Health Data[A]. Fisher M, Getis A. Recent Developments in Spatial Data Analysis: Spatial Statistics, Behavioral Modeling and Neuro-computing[C]. Berlin: Springer, 1997.

[11] Leung Y, Zhang J, Xu Z. Clustering by Scale-Space Filtering[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(12): 1396—1410.

[12] Luo J C, Zhou C H, Leung Yee, *et al*. Scale-Space Theory Based Regionalization for Spatial Cells[J]. *Acta Geographica Sinica*, 2002, **52**(2): 167—173. [骆剑承, 周成虎, 梁怡等. 多尺度空间单元区域划分方法[J]. 地理学报, 2002, **57**(2): 167—173.]

[13] Wang H Q, Wang J F. An Adapted *K*-means Algorithm Based on Spatial Contiguity Relations[J]. *Computer Engineering*, 2006, **32**(21): 50—51. [王海起, 王劲峰. 一种基于空间邻接关系的 *K*-means聚类改进算法. 计算机工程, 2006, **32**(21): 50—51.]

[14] Zhang G, Patuwo B E, Hu M Y. Forecasting with Artificial

Neural Networks: The State of the Art[J]. *International Journal of Forecasting*, 1998, **14**: 35—62.

[15] Wang J F. Structural Adaptive Modeling of Spatial Geo-information[J]. *Acta Geographica Sinica*, 1995, **50**(Suppl): 54—61. [王劲峰. 空间信息的结构自适应模型[J]. 地理学报, 1995, **50**(增刊): 54—61.]

[16] Data Source of France Flu: www.sph.umich.edu/geomed/data/france/. [法国流感数据来源: www.sph.umich.edu/geomed/data/france/.]

Local Neural Networks of Space-time Modeling Based on Partitioning for Lattice Data in GIS

WANG Hai-qi[1,2], WANG Jin-feng[2]

(1. *College of Geo-resources and Information, University of Petroleum (East China), Dongying Shandong* 257061, *China*;
2. *LREIS, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract This paper focuses on space-time nonlinear intelligent modeling for lattice data. Lattice data refers to attributes attached to fixed, regular or irregular, polygonal regions such as districts or census zones in two-dimensional space. Lattice data space-time analysis is aiming at detecting, modeling and predicting space-time patterns or trends of lattice attributes changed with time while spatial topological structures are simultaneously kept invariable. From the perspective of space, lattice objects have two different scale spatial properties influencing lattice data modeling: global dependence and local fluctuation. Global spatial dependence or autocorrelation quantifies the correlation of the same attribute at different spatial locations; and local spatial fluctuation or rough, coexisted with global dependence, is represented in the form of local spatial clustering of similar values or local spatial outliers. To consider simultaneously the effects of two properties above, local neural networks (NN) model is studied for space-time nonlinear autoregressive modeling. The main research contents include: (1) To reduce influence of spatial fluctuation on prediction accuracy of NN, all regions are partitioned into several subareas by an improved k-means algorithm. (2) Different partition schemes are evaluated and compared according to three essential criteria including dependence, continuity, fluctuation. Dependence means that an optimal partition must guarantee that there is real and significant spatial dependence among regions in a subarea because the results of output layer nodes in a NN model depending on the interactions of input layer nodes through hidden layers nodes. Spatial autocorrelation of a subarea can be measured by global Moran's I and its significance test can be done based on z-score of Moran's I. Continuity means that only neighboring regions can be grouped into a subarea, and this criterion is fused into the modified k-means algorithm. When the algorithm judges one region which subarea it belongs to, not only should the distance be considered to the centroid of a subarea but also the common borders between this region and the regions in a subarea. As to fluctuation, although it is impossible to make each subarea have complete spatial stability through partitioning, the less fluctuation means the better predicting results of NN model. For a subarea, standard deviation between local Moran's I of all regions in the subarea and global Moran's I of the subarea is regarded as an evaluation index to the fluctuation of the subarea. (3) Each multi-layer perceptrons (MLPs) network is used respectively in modeling and predicting for each subarea. The output nodes are the predicting values at time t of an attribute for all regions in a subarea. The input nodes are observations before time t of the same attribute of both regions in the subarea and regions neighboring to the subarea and the latter is called boundary effect. Finally, as a case study, all local models of all the subareas are trained, tested and compared with a single global MLPs network by modeling one-step-ahead prediction of an epidemic dataset which records weekly influenza cases of 94 departments in France from the first week of 1990 to the 53th of 1992. Two performance measures, including average relative variance (ARV) and dynamic similarity rate (DSR), indicate that local NN model based on partitioning has better predicting capability than global NN model. Several issues are still worth further study: (1) The initial subareas of partitioning are selected randomly in our research. In the further study, a reasonable approach should combine selection with spatial patterns, for instance considering the center of local cluster. (2) Partition criteria should be another issue and different types of spatial and space-time processes, such as rainfall, price waves, public data, etc, may have different objective criteria for choosing an optimal partition. (3) It may be more imperative to study feasible measures for quantifying global and local space-time dependence of lattice data and testing significance of this dependence.

Key words lattice data; space-time modeling; partitioning; *K*-means clustering; neural networks; boundary effect