

# 多层次格网模型的近邻指数聚类生态区划算法与实验 ——以新疆北部地区区划为例

袁焯城<sup>1,2</sup>, 周成虎<sup>1</sup>, 覃彪<sup>1</sup>, 欧阳<sup>1</sup>

(1. 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;

2. 中国科学院研究生院, 北京 100049)

**摘要:** 在综合考虑气候、植被、地貌等因素的基础上, 提出一种基于多层次格网模型的最近邻指数-模糊聚类生态区划算法 (Nearest Neighbor Index Fuzzy Clustering, NNI-FC)。该算法采用“自下而上”的方式, 首先, 利用离散格网单元之间的严格相似性形成区划的核心分区; 然后, 通过最近邻指数统计分析细碎分区的空间格局及其面积覆盖率, 再以模糊聚类方法将相似度最大的细碎区聚合归并, 即可得到相应的生态区划方案。数值实验证明了该算法可以很好地体现区域的分异特征, 并且具有较高的效率。

**关键词:** 生态区划; 多层次格网模型; 聚类; 最近邻指数; 空间分布

**DOI:** 10.3724/SP.J.1047.2011.00001

## 1 引言

生态区划是在对生态系统客观认识与充分研究的基础上, 利用生态学原理和方法揭示自然地域内在的相似性和差异性, 进行整合与分区, 划分生态环境的区域单元<sup>[1-4]</sup>。自1976年Bailey提出真正意义的生态区划方案以来, 生态地理区划已成为地理学界、生态学界所关注的焦点课题之一, 其成果在分析生态系统与环境变化问题的形成与机理方面发挥了重要作用。

经典的区划方法包括“自上而下”和“自下而上”两种。“自上而下”是指通过判别研究区内的差异性, 把整个研究区细分为小区的过程; “自下而上”则从基本地理单元开始, 通过归纳地理单元间的相似性, 依次往上合并为若干分区。

区划的技术手段主要有叠置法、主导标志法、主成分分析法、地理相关分析法、聚类分析法、多元线性判别法等<sup>[5-6]</sup>。叠置法和主导标志法属于定性分析方法, 存在主观性强、精确度不够等缺陷。而后两者则属于定量分析方法, 适用于基于行政统计单元的区域划分, 因为这类区划的基本单元个数较少(通常不超过一百个), 且无须制定新的边界线

(行政区本身已包含了边界线, 对其适当整合即可)。但定量分析法只能处理数值属性, 对于非量化的关键地理特征(如植被、地貌等)则无法恰当地表达或分析。

随着遥感技术、地理信息系统的发展和区划研究的进一步深入, 单纯的定量分析法或定性分析法已无法满足区划的系统性、合理性要求, 对需要考虑多种指标的生态区划尤其如此。另外, 对于给定一块自然地域, 按其内在的相似性与差异性从底层单元进行划分时(即“自下而上”方式), 定量分析法应用起来很繁琐且效率低下, 因为此时不仅要考虑区划的类型、数目, 还要考虑空间关系并确定区划边界线。针对这些问题, 本文提出一种基于多层次格网模型的最近邻指数-模糊聚类生态区划算法(NNI-FC), 将定性分析与定量分析的区划技术手段有机地结合起来。

## 2 多层次格网模型的近邻指数聚类生态区划算法

### 2.1 多层次格网模型

地理信息系统(GIS)的格网模型是空间信息的

收稿日期: 2009-03-30; 修回日期: 2010-11-03.

基金项目: 中科院院士咨询项目“新疆生态建设与可持续发展战略研究”(KZCX3-SW-347)。

作者简介: 袁焯城(1983-), 男, 浙江嵊州人, 博士研究生, 主要从事空间数据挖掘的研究。E-mail: yuanyzc@lreis.ac.cn

主要表达方式之一。它特别适合多源数据的融合分析。各种来源不同、格式相异的数据,如分辨率不同的航空影像、GPS 高程采样点矢量图层、面状的植被分布图等,可以通过重采样、插值等算法生成相应的格网图<sup>[7-8]</sup>。

多层次格网模型实际上就是通过对图层的叠加,给单个网格赋多重属性。另外,格网系统中,单元网格之间暗含位置关系,利用这点可以方便地实现各种空间分析。格网模型的用途相当广泛,在描述地球动力学现象、生态环境的地域综合评估、遥感像元的图像数据处理、描述高度离散的现象、面向空间分析的背景数据库等方面有特殊的适应性能<sup>[9]</sup>。本研究以多层次格网模型为基础,实现生态区域的自动划分。

2.2 最近邻指数-模糊聚类生态区划算法(NNI-FC)

基于格网模型的生态区划分,关键要解决两个问题:确定各格网单元所属的生态分区类型及生态区的边界。对于给定范围的自然区划,通常无法事先确定它所包含的分区类型及其数目。因此,我们利用非监督的模糊聚类来解决这个问题<sup>[10]</sup>。首先,根据需要对格网单元的多源属性数据进行适当的预处理;其次,利用类别属性组合和空间邻接的聚类方法,计算格网单元之间的相似度及其所属的类型,并根据相似度通过格网单元扩张的方式形成相应等级区划的核心区及初始边界;最后,由初始分区单元的最近邻指数分析结果,以模糊聚类方法整合相似度高的相邻小分区,形成最终的区域划分方案。

2.2.1 格网属性数据的预处理

格网单元的属性在数学上可分为两类:数值型和非数值型。数值型属性指的是有具体度量数据的属性,如海拔高程 DEM、干燥度指数等。非数值型属性是指一些分类别、分等级的文本数据,如各种植被类型、土地利用类型等。

在生态区划中,我们有时希望通过适当的方式将部分数值型属性转换成非数值型的,以便它们参与地理类型的组合。一般采用名称替代或以某种标准进行分等定级的方法进行转换。例如,年平均气温可分为温和、高寒等,年降水量与干燥度也可赋予相应的名称。表 1 是干湿状况的划分指标,它们在气候区的划分中有重要意义<sup>[3]</sup>。

表 1 干湿状况的划分指标

Tab. 1 Classification index of dry and wet condition

名称	干燥度
湿润	0.50~0.99
亚湿润	1.00~1.49
亚干旱	1.40~4.00; 1.50~5.00(在青藏高原)
干旱	≥4.00; ≥5.00(在青藏高原)

在生态区划中,主要考虑水分、热量、土壤、植被、地貌的空间分异特征、结构组合和区域分布、人类活动对生态系统的影响等。其中,土壤、植被、地貌为非数值型属性,它们是区划格局的核心和基础<sup>[10-11]</sup>。年平均气温、年降水量、干燥度等数值型属性可以通过恰当的转换,变成非数值型属性,这样就为区划的地理类型组合提供了便利。

2.2.2 核心区的形成

将所考察的自然地域格网化后,每个格网单元就是一个待分类对象,可用集合表示为  $X = \{x_1, x_2, \dots, x_n\}$ 。其中,  $n$  为单元总数,  $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$  表示第  $i$  个单元  $x_i$  的特征矢量,  $m$  为属性的个数,  $x_{ik} (k=1, 2, \dots, m)$  为  $x_i$  第  $k$  个属性值。令  $M = \{1, 2, \dots, m\}$  为指标集,  $M_1 = \{k | x_{ik} \text{ 是非数值型属性}, \forall 1 \leq i \leq n\}$ ,  $M_2 = \{k | x_{ik} \text{ 是数值型属性}, \forall 1 \leq i \leq n\}$ 。显然,  $M = M_1 \cup M_2$ ,  $|M| = |M_1| + |M_2|$ 。

可利用相似度的大小来判断两个格网单元  $x_i$ 、 $x_j$  是否同属一类别。相似度有多种定义方法,本文借鉴 ROCK (Robust Clustering using linKs) 算法<sup>[12]</sup>处理非数值型属性,用连接度来定义单元之间的相似度。格网单元  $x_i$ 、 $x_j$  之间连接度  $r_{ij}$  定义为:

$$r_{ij} = r(x_i, x_j) = \frac{\sum_{k \in M_1} P(x_{ik}, x_{jk})}{|M_1|} \tag{1}$$

其中,  $P(x_{ik}, x_{jk}) = \begin{cases} 1, & x_{ik} = x_{jk} \\ 0, & x_{ik} \neq x_{jk} \end{cases}$ ,  $|M_1|$  表示集合  $M_1$  元素的个数,  $r_{ij} \in [0, 1]$ 。两个单元的连接度越大,则它们的相似度越高。

由式(1)可知,格网单元  $x_i$  和  $x_j$  之间的相似度由它们属性值相同的个数所决定的。在初步聚类形成核心区时,只需考虑当前网格单元与其周围八个邻居单元的关系。若单元  $x_i$ 、 $x_j$  的相似度  $r_{ij} = 1$ ,即它们的非数值型属性完全相同,便将它们归为

一类;若  $r_{ij} < 1$ , 则暂不将  $x_i, x_j$  合并为一类。在图 1(a)中, 网格单元 S 属于的图中灰色斜线部分, 设该区为  $D_1$ , 下面以 S 为起点说明该核心区的形成过程:

(1) 计算 S 与它的 1 号邻居  $S_1$  的  $r$ , 若  $r=1$ , 则转到(2), 否则转到(3)。

(2) 将 S 和  $S_1$  归为同一类, 即将  $S_1$  所属区域赋为  $D_1$ 。然后以  $S_1$  为起点, 重复(1)。

(3) 计算 S 与它的 2 号邻居  $S_2$  的  $r$ , 若  $r=1$ , 则转到(2), 只是这里(2)中的  $S_1$  变成了  $S_2$ , 否则继续往下计算 S 和它的 3 号邻居的相似度  $r$ , 处理方式与前面的 1 号 2 号邻居一样。

遍历 S 的 8 个邻居网格单元, 便得到与 S 同类的可能连成片的所有网格单元。图 1(b)显示了该区各个网格遍历的顺序。

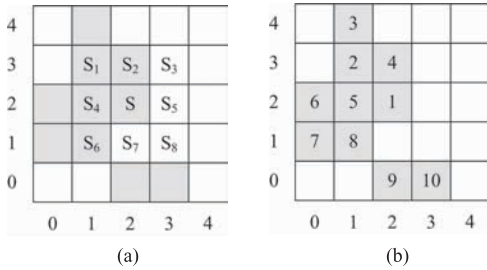


图 1 邻居单元遍历顺序示意图

Fig. 1 The ergodic sequence of Neighbor-elements

上述根据完全相似度进行划分的方式, 保证了每个分区空间上的连贯性, 又保证了属性的一致性, 且仅需扫描一次即可得到考察地域的一个初始划分, 效率较高。

一般地, 由初始划分可得到若干个分区。我们将那些包含的网格单元较多、总面积较大的代表性分区称为核心区; 将网格单元较少、总面积较小的零散分区称为细碎区。生态区划的基本格局就是以核心区为基础的。但是, 由于地理特征在空间分布上具有一定的连续性, 核心区与细碎区之间并不总是存在严格的界线。因此, 本文采用阈值(设为  $\theta$ )来界定它们: 对于任意一个初始分区, 如果其中的网格单元数不少于  $\theta$ , 则该分区是核心区; 否则, 该分区就是细碎区。

形成细碎区的原因主要有两个: 一是各叠加图层边界的不一致性所导致。例如, 土壤类型图和植被类型图叠加组合时, 由于数据来源的不同, 某种植被

类型的图斑和对应的土壤类型图斑未能完全叠合到一起, 它们的边界误差就表现为细碎狭长的多边形。这类小区域属于数据不匹配而造成的异常数据, 应作适当的校正。二是多重格网数据通过类型组合自然形成的, 但分区中网格单元数少于阈值  $\theta$ 。

### 2.2.3 细碎区的处理

初始划分中细碎区的界定与阈值  $\theta$  是直接相关的,  $\theta$  值一般随着格网尺度、区划级别的变化而变化。在给定格网尺度与区划级别的前提下, 细碎零散分区的统计与决策指导的意义均不大。因此, 应在充分考虑空间关系与相似度的基础上, 将它们适当地整合归并。

假设初始划分得到  $p$  个分区  $\{C_1, C_2, \dots, C_p\}$ ,  $p \leq n$  (其中, 包括  $q$  ( $q \leq p$ ) 个细碎区。参考式(1), 定义分区  $C_i, C_j$  的相似度为:

$$R(C_i, C_j) = \frac{|M_1| \cdot r_{ij}^a + |M_2| \cdot r_{ij}^b}{|M|} \quad (2)$$

其中,  $M, M_1, M_2$  的含义参见前一节, 且  $|M| = |M_1| + |M_2|$ ;  $r_{ij}^a$  为  $C_i$  与  $C_j$  的非数值属性之间相似度:

$$r_{ij}^a = \frac{\sum_{k \in M_1} P(C_{ik}, C_{jk})}{|M_1|},$$

$$\text{其中: } P(C_{ik}, C_{jk}) = \begin{cases} 1, & C_{ik} = C_{jk} \\ 0, & C_{ik} \neq C_{jk} \end{cases} \quad (3)$$

而  $r_{ij}^b$  为  $C_i$  与  $C_j$  的数值属性之间相似度:

$$r_{ij}^b = \frac{\sum_{k \in M_2} |C_{ik} - \bar{C}_i| |C_{jk} - \bar{C}_j|}{\sqrt{\sum_{k \in M_2} (C_{ik} - \bar{C}_i)^2 \cdot \sum_{k \in M_2} (C_{jk} - \bar{C}_j)^2}},$$

$$\text{其中: } \bar{C}_i = \frac{\sum_{k \in M_2} C_{ik}}{|M_2|}, \bar{C}_j = \frac{\sum_{k \in M_2} C_{jk}}{|M_2|} \quad (4)$$

由式(2)可知,  $0 \leq R(C_i, C_j) \leq 1$ 。  $R(C_i, C_j)$  越大, 则  $C_i$  和  $C_j$  的共性就越多, 相似度就越高。在处理细碎区时, 我们采用这种模糊聚类方法并结合最近邻指数 (Nearest Neighbor Index, NNI, 详见下文处理过程的第(4)步) 来整合相似度高的分区。

下面, 我们以地带性植被为划分主导因素为例, 说明细碎区的具体处理过程。首先, 我们把细碎区标记为  $T_i$  ( $1 \leq i \leq q$ ),  $\{T_1, T_2, \dots, T_q\} \subseteq \{C_1, C_2, \dots, C_p\}$ 。

(1) 把考察区以最小外接矩形为基准, 均分成四个小矩形  $R_a$  ( $a=1, 2, 3, 4$ ); 若某个细碎区  $T_i$  (1

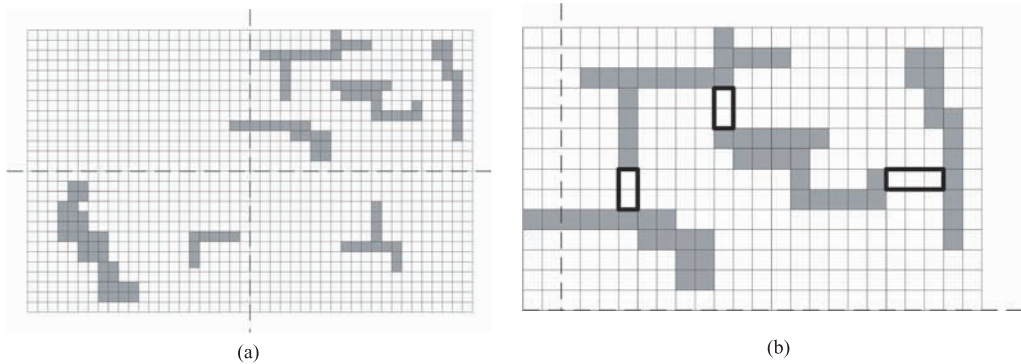


图2 区域分割及细碎区连接示意图(a)将区域四等分;(b)为(a)四等分之后的右上角部分,并用粗边框将细碎区连成一体

Fig. 2 The schematic diagrams of region division and fragmented-regions connection. (a)Dividing the region into quarters; (b)The right-top quarter of (a),connecting the fragmented regions by grids which have bold borders

$\leq i \leq q$ 被两个以上的小矩形共同占有,则按面积占优法来确定它属于哪个小矩形(如图2所示)。指定分区合并的面积比阈值  $\sigma (0 \leq \sigma \leq 1)$ ,并令  $a=1$ ,转入(2)。

(2)按地带性植被分别统计矩形  $R_a$  内各植被类型细碎区的数目  $N_a^b$  及面积总和  $S_a^b$ 。其中,  $b=1, 2, \dots, 8$ ,且依次对应水体、绿洲、沼泽、林地、草原、草甸、荒地、高山植被等八个植被类型。令  $b=1$ ,转入(3)。

(3)若  $N_a^b=0$ ,转 Step(9);若  $N_a^b=1$ ,转 Step(6);若  $N_a^b>1$ ,转入(4)。

(4)先计算  $N_a^b$  对应细碎区各自的最近邻细碎区,并从最小距离处将  $R_a$  内的细碎区连接为一体,记为  $R_{\text{link}}$ (如图2(b));其次计算多边形  $R_{\text{link}}$  的面积,记为  $S_{\text{link}}$ (它包括了细碎区面积及其连接处面积),显然,  $S_{\text{link}} \geq S_a^b$ ;最后计算  $R_{\text{link}}$  中细碎区的面积覆盖率  $S_{\text{ratio}} = \frac{S_a^b}{S_{\text{link}}}$  及最近邻指数  $NNI = 2\bar{d}\sqrt{\frac{N_a^b}{S_{\text{link}}}}$ 。其中,  $\bar{d}$  为  $N_a^b$  对应的细碎区最近邻之间的平均距离,而两个细碎区之间的距离定义为它们对应多边形之间的距离<sup>[13]</sup>。转入(5)。

(5)若面积覆盖率  $S_{\text{ratio}} \geq \sigma$  时,则将多边形  $R_{\text{link}}$  整体合成一个分区,其植被类型取为  $b$  值所对应的类型,转入(6);若  $S_{\text{ratio}} < \sigma$  且  $NNI < 1$  即细碎区趋于聚集分布,则参照(1)将  $R_{\text{link}}$  不断的四等分进行递归处理,直至它对应的某个最小外接矩形满足  $S_{\text{ratio}} \geq \sigma$ ,此时将该外接矩形当做一个分区,转入(6);若  $S_{\text{ratio}} < \sigma$  且  $NNI \geq 1$  即细碎区是随机乃至均

匀分布的<sup>[14-15]</sup>,转入(8)。

(6)若当前考察的分区包含的网格单元个数超过划分阈值  $\theta$ ,则它已是核心区,停止对它的处理,并转入(8);否则,转向(7)。

(7)若当前考察的分区与某矩形  $R_{a_0} (a_0 \neq a)$  中的某个同类细碎区距离很小,则将该分区并入  $R_{a_0}$  进行处理,转向(2);否则,若该分区与任意  $R_{a_0} (a_0 \neq a)$  中的任意细碎区距离都很大,则先利用式(2)计算该分区与所有的相邻核心区之间的相似度,再将它并入相似度最高的分区中。转入(9)。

(8)对于  $R_a$  中未处理的细碎区,若它与某矩形  $R_{a_0} (a_0 \neq a)$  中的某个同类细碎区距离很小,则将它并入  $R_{a_0}$  进行处理;否则,利用式(2)按模糊聚类方式将它归并到与它相邻且相似度最高的核心区。转入(9)。

(9)若  $b < 8$ ,令  $b=b+1$ ,转向(3);否则,转入(10)。

(10)若  $a < 4$ ,令  $a=a+1$ ,转向(2);否则,区域划分已完成,转入(11)。

(11)沿区划边界线做平滑处理,消除可能存在的尖锐狭长锯齿边界。输出当前的划分方案,即为最终划分结果。

应指出:一是本算法将所考察的自然地域逐步四等分处理(参见(5)),是一个递归算法。其需事先指定两个阈值:核心区与细碎区的划分阈值  $\theta$ 、分区合并的面积比阈值  $\sigma$ ;另在(5)中,细碎区的空间分布格局可采用  $\chi^2$  检验,此处略。二是地理要素在空间分布上有一定的连续性,因此,对(5)的处理

方式是合理可取的;同时将若干个细碎区合并为一个较大的分区时,其边界线可能会分割其他区域而产生一些新的细碎区,这些细碎区参照(7)处理;对

(11)可采用高斯滤波处理,具体处理步骤略。  
图 3 展示了细碎区整合归并前后的对比。图 4 给出了完整的近邻聚类区划算法处理流程。

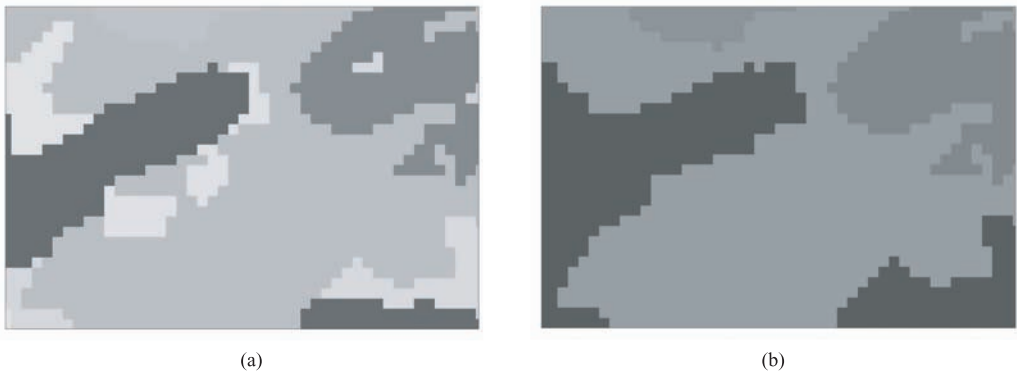


图 3 细碎区的整合:(a)为处理前(b)为处理后  
Fig. 3 Merging fragmented regions: (a) Before the merging; (b) After the merging

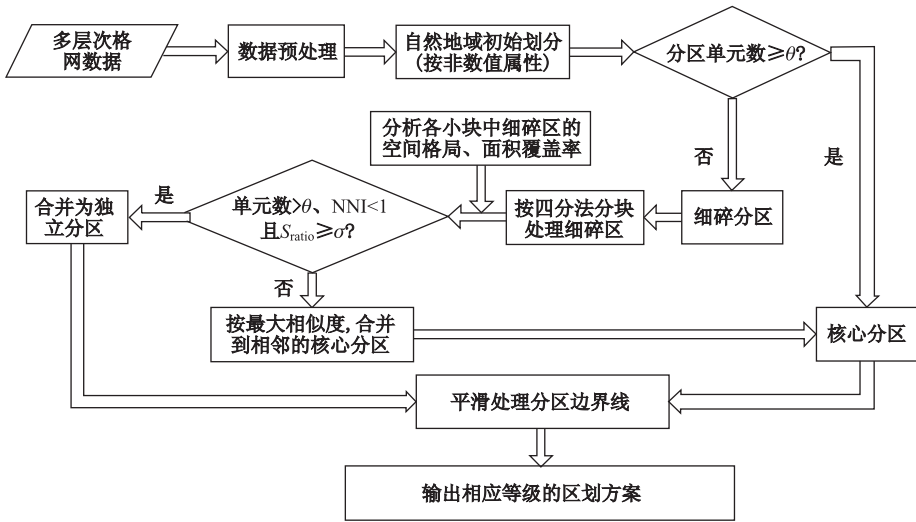


图 4 多源数据格网的近邻聚类区划算法模型流程图  
Fig. 4 Flow diagram of nearest neighbor index clustering algorithm based on the multi-layer grid model

### 3 新疆北部地区生态区划的实验

#### 3.1 实验数据

新疆北部包括新疆天山以北的地区,以及阿尔泰山和准格尔盆地。该地区属于温带大陆性气候,干旱少雨。该地区盆地年均气温 6~8℃,山区 2000m 以上为 0℃。山地年降水量 400~500mm,盆地边缘 150~200mm,盆地中心 100mm 左右。伊犁河和额尔齐斯河为该地区最大的河流水系,占整个地区径流量的 60%。从山麓到沙漠植被主要由

荒漠草原,荒漠草甸,河漫滩草甸,荒漠等组成,山地自上而下分布着高山垫状植物,高山蒿草荒原,山地云杉林,山地针茅草原,低山荒漠草原<sup>[16-17]</sup>。本区约占新疆总面积的 27%,却集中了全疆约 48%的人口,是整个新疆最富饶的地方。

生态系统的等级性是生态区划的基本原则之一,不同的等级有不同的指标体系<sup>[10-11]</sup>。我们对该地区的第三级生态区划分,所采用的指标包括地貌类型(高山、中山、低山、扇缘带、沙漠、湖盆、平原荒漠)、地带性植被(水体、绿洲、沼泽、林地、草原、草甸、荒地、高山植被)、年平均气温、年平均降



雨量和干燥度。其中,中地貌类型和地带性植被数据属于非数值型数据,分别根据中科院地理所地貌组绘制的 1:100 万全国地貌图和中科院植物所绘制的 1:100 万全国植被图,通过 ArcGIS 综合提取并矢栅转换得到。年平均气温、年平均降雨量和干燥度属于数值型数据,来源于中国农业科学院农业自然资源与农业区划研究所编制的生态环境背景层面数据,以 1km×1km 的 Arc/Info Grid 数据格式、以整数形式存储。

3.2 实验与结果

我们以 ESRI 的 ArcObject 类库为基础,Visual Studio .Net 2003 为开发工具自主开发出一套区划软件原型系统,实现了基本的图层加载、地图编辑、矢栅格式转换等功能,系统的核心是本文的多层格网近邻聚类区划算法。新疆北部地区生态区划实验在此基础上进行。

我们用 1km×1km 的网格对研究区进行格网化,以上一小节所述的地貌类型、地带性植被、年平均气温、年平均降雨量和干燥度数据作为实验参

数。

实验的第一步——核心区的形成处理完毕后,在阈值参数  $\theta=500$  的水平下,我们得到 2965 个分区,其中绝大多数是细碎区,但它们只占研究区总面积的 23.5%。表 2 是这些分区具体的统计信息。

表 2 核心区与细碎区的统计信息

Tab. 2 Statistic information of core regions and fragmented regions

分区类型	个数	占总分区个数的百分比(%)	网格数之和	占总网格数的百分比(%)
细碎区(网格数< $\theta$ )	2897	97.70	71622	23.50
核心区(网格数 $\geq\theta$ )	68	2.30	233073	76.50

注: $\theta=500$

实验的第二步——细碎区的处理,我们选定阈值参数  $\sigma=0.8$ ,最终聚类得到 68 个生态区,图 5 是原始区划结果栅格图,图 6 是图 5 矢量化后得到的结果矢量图,为了清晰起见,我们在图中标出了每个小区的编号。表 3 列出了其中 10 个典型分区的统计信息。

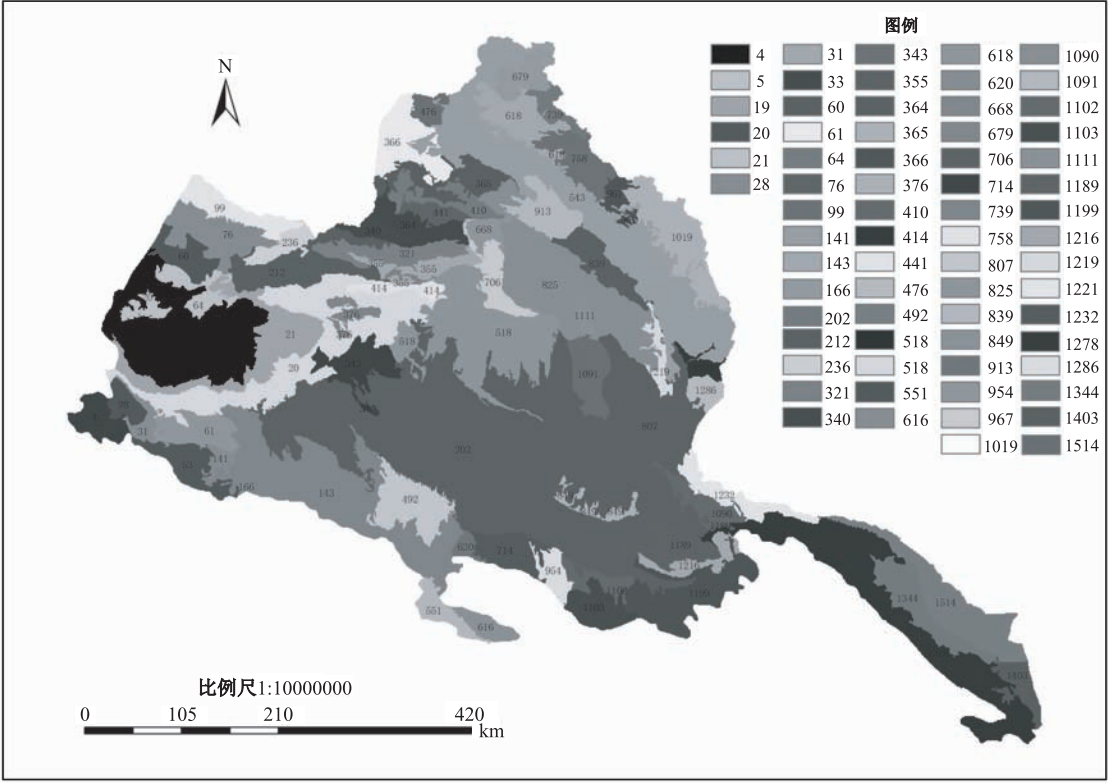


图 5 北疆生态区划分结果栅格图

Fig. 5 The raster map of ecological regionalization for North of Xinjiang

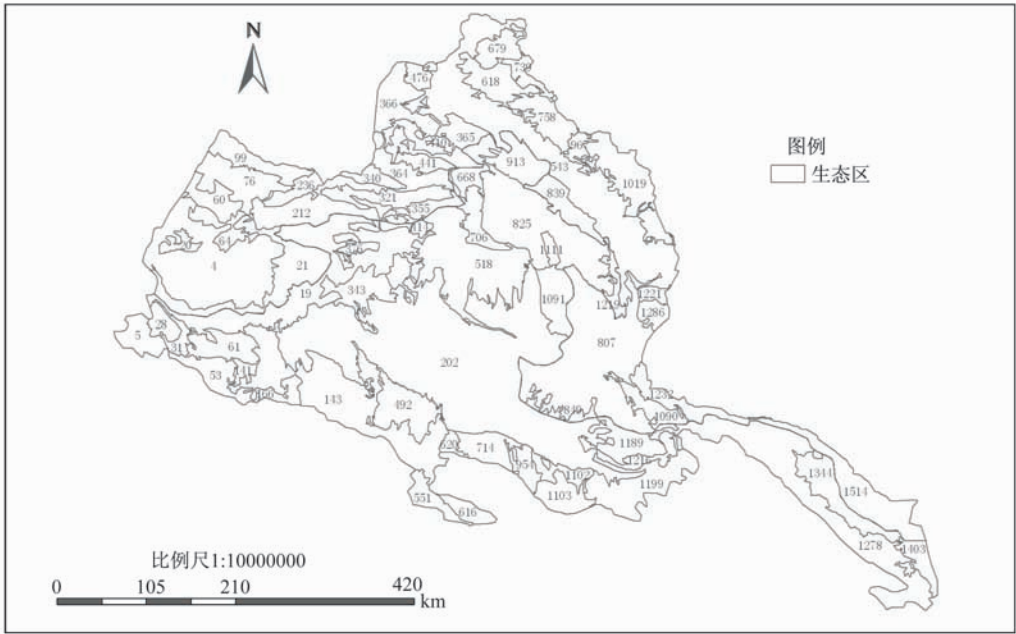


图 6 北疆生态区划分结果矢量图

Fig. 6 The vector map of ecological regionalization for North of Xinjiang

表 3 北疆地区第三级生态区划部分分区的指标参数统计

Tab. 3 Index parameter statistics of part of ecological regions for North of Xinjiang at the third ecological regionalization level

区域 编号	中地貌统计信息		地带性植被统计信息		年平均气温(℃)	年平均降雨量(mm)	干燥度
	中地貌类型	组成百分比	地带性植被类型	组成百分比			
20	平原荒漠	0.935	草甸	0.993	-0.8	253.1	6.138
	高山	0.065	高山植被	0.007			
64	扇缘带	0.987	绿洲	0.152	5.8	269.9	3.114
	平原荒漠	0.013	草原	0.009			
340	高山	0.988	荒地	0.839	-5.2	161.4	4.729
			林地	0.021			
			草甸	0.785			
			荒地	0.027			
849	扇缘带	0.945	高山植被	0.167	7.8	279.4	2.991
	沙漠	0.001	荒地	1			
	平原荒漠	0.054					
967	高山	0.102	林地	0.749	-3.2	164.9	8.102
	平原荒漠	0.898	草原	0.155			
28	湖盆	0.95	草甸	0.091	9.1	136.1	6.334
			高山植被	0.006			
			水体	0.85			
			草甸	0.15			
343	沙漠	0.005	水体	0.012	7.1	137.6	5.919
	湖盆	0.992	草甸	0.019			
1102	平原荒漠	0.003	荒地	0.97	5.6	332.8	3.982
	扇缘带	0.95	水体	0.004			
	平原荒漠	0.05	绿洲	0.013			
			草甸	0.85			
1103	扇缘带	0.773	荒地	0.133	4.2	328.5	3.942
			绿洲	0.767			
			草原	0.046			
1111	沙漠	1	荒地	0.188	2.7	158.3	4.83
			林地	0.041			
			荒地	0.959			

3.3 实验精度分析

生态区划与基于行政统计单元的区划(如主体功能区划)不同,后者只是基本行政区的合并,不需要考虑区域界线的重新划分,而生态区划主要是对生态区域分异规律的一种表达,各个分区之间的界线很难界定。到目前为止,还没有一个统一的标准来评判生态区划界线的"对错"。事实上,由于存在着大量的地貌、植被等过渡区,生态区之间并没有严格意义上的分界线。因此,我们只能从细碎区合并的角度,把区划结果的精度  $F$  定义为:某一区划指标的主要类型在该区域内所占的面积百分比,即:

$$F_{\mu} = \frac{S_{\mu}}{S_{total}} \tag{5}$$

其中  $\mu$  为区划指标,如中地貌类型,  $S_{\mu}$  为该指标在分区中的主要类型所占的面积,  $S_{total}$  为该区域的总面积。

一般而言,气候是生态系统的主要决定因素,而植被则是体现气候敏感性的最佳指标;地貌与地形由于对水热分子的分布起重要的作用,因而它们也是表征生态地理特征异同规律的重要指标<sup>[10-11]</sup>。我们通过计算各区域内的中地貌类型和地带性植被的  $F$  值,来粗略估计实验结果的精度。

表 4 列出了各生态区主要的参数指标,图 7、图 8 分别是区内主要中地貌类型、地带性植被所占的面积百分比。这两幅图的横标序号依次对应表 4 中各分区的序号。

从图 7、图 8 及表 4 可知,实验结果的中地貌类型精度最小值为  $F_{中地貌}^{最小值} = 0.638$ ,最大值为  $F_{中地貌}^{最大值} = 1$ ,平均值为  $F_{中地貌}^{平均值} = 0.881$ ,地带性植被的精度最小值为  $F_{地带性植被}^{最小值} = 0.402$ ,最大值为  $F_{地带性植被}^{最大值} = 1$ ,平均

值为  $F_{地带性植被}^{平均值} = 0.824$ 。结果表明,本方法的分类精度平均达到了 80% 以上。

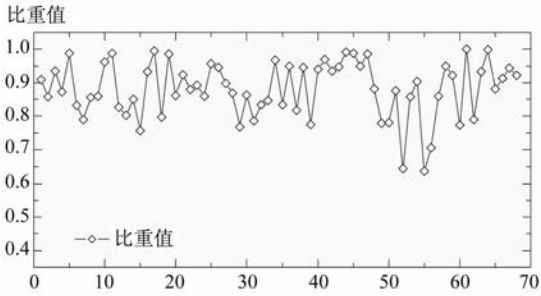


图 7 各分区主要中地貌类型所占面积比重  
Fig. 7 Area percents of main geo-morphological type in each region

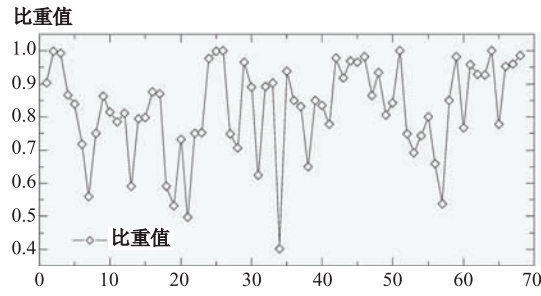


图 8 各分区主要地带性植被所占面积比重  
Fig. 8 Area percents of main zonal vegetation type in each region

另外,从表 4 可知,在同一分区的主要地貌与非主要地貌之间、主要植被类型与非主要植被类型之间都比较相似,并且区内的主要地貌类型与主要植被类型也一一对应。因此,本模型的实验结果很好地体现了分区内主要生态指标相一致性的原则,说明基于多层次格网模型的近邻指数聚类生态区划算法(NNI-FC)是有效可行的。

表 4 北疆地区第三级生态区划所有分区的主要指标参数统计  
Tab. 4 Main index parameter statistics of all ecological regions for North of Xinjiang at the third ecological regionalization level

序号	区域 编号	中地貌统计信息		地带性植被统计信息		年平均 气温(℃)	年平均 降雨量(mm)	干燥度
		主要中地貌类型	面积百分比	主要植被类型	面积百分比			
1	4	平原荒漠	0.908	草地	0.903	2.2	218.5	4.954
2	5	扇缘带	0.968	绿洲	0.402	7.6	153.7	5.432
3	19	扇缘带	0.834	荒地	0.938	6.9	148.4	5.225
4	20	平原荒漠	0.935	草甸	0.993	-0.8	253.1	6.138
5	21	平原荒漠	0.857	荒地	0.999	5.8	159.3	5.894
6	28	湖盆	0.95	水体	0.85	9.1	136.1	6.334
7	31	湖盆	0.818	荒地	0.832	8.5	132.4	6.394
8	53	扇缘带	0.946	荒地	0.65	6.7	176.5	5.272



(续表)								
序号	区域 编号	中地貌统计信息		地带性植被统计信息		年平均 气温(℃)	年平均 降雨量(mm)	干燥度
		主要中地貌类型	面积百分比	主要植被类型	面积百分比			
9	60	扇缘带	0.873	草甸	0.867	7.4	280.6	3.697
10	61	沙漠	0.775	荒地	0.849	7.4	146.1	5.968
11	64	扇缘带	0.987	荒地	0.839	5.8	269.9	3.114
12	76	扇缘带	0.832	绿洲	0.718	5.2	279.9	3.874
13	99	平原荒漠	0.79	草原	0.56	1	270.1	5.242
14	141	扇缘带	0.97	荒地	0.778	7.4	162.4	5.512
15	143	扇缘带	0.94	绿洲	0.835	6.9	188.4	4.915
16	166	扇缘带	0.947	荒地	0.917	6.8	175.9	5.297
17	202	沙漠	0.934	荒地	0.979	6.6	201.3	4.563
18	212	平原荒漠	0.856	草原	0.75	1.2	199.6	4.562
19	236	平原荒漠	0.86	草甸	0.863	−0.4	233.1	5.292
20	321	扇缘带	0.962	草原	0.815	1.9	149.9	4.162
21	340	高山	0.988	草甸	0.785	−5.2	161.4	4.729
22	343	湖盆	0.992	荒地	0.97	7.1	137.6	5.919
23	355	平原荒漠	0.826	草原	0.811	2.7	143.8	4.342
24	364	平原荒漠	0.803	草原	0.591	0.7	163.7	5.412
25	365	扇缘带	0.851	荒地	0.795	3.9	159.1	5.057
26	366	沙漠	0.757	荒地	0.798	5.1	173.9	4.563
27	376	平原荒漠	0.988	荒地	0.966	7.6	147.6	4.469
28	410	平原荒漠	0.933	荒地	0.875	3.4	160.8	4.936
29	414	平原荒漠	0.95	荒地	0.982	4.1	147.1	4.025
30	441	扇缘带	0.995	荒地	0.87	2.7	161.3	4.725
31	476	平原荒漠	0.985	荒地	0.531	2.1	175.7	5.179
32	492	扇缘带	0.985	荒地	0.865	6.8	186.5	4.781
33	518	平原荒漠	0.881	荒地	0.934	4.9	139.7	5.161
34	543	平原荒漠	0.798	草原	0.59	−1.2	169.6	6.118
35	551	扇缘带	0.778	荒地	0.806	4.7	307.5	2.518
36	616	扇缘带	0.78	荒地	0.843	3.8	167.2	4.616
37	618	平原荒漠	0.861	草甸	0.733	−3.3	159.7	9.361
38	620	扇缘带	0.876	荒地	1	7.2	237	3.464
39	668	湖盆	0.857	水体	0.691	4.2	130.1	5.816
40	679	高山	0.924	草甸	0.497	−7.4	159.4	9.282
41	706	湖盆	0.645	荒地	0.748	4.5	136.1	5.379
42	714	扇缘带	0.903	荒地	0.744	6.8	299.8	2.605
43	739	高山	0.88	草甸	0.75	−7.1	158.1	11.586
44	758	高山	0.893	草甸	0.752	−6.5	164.1	11.378
45	807	平原荒漠	0.859	荒地	0.977	4.7	218.5	3.3
46	825	平原荒漠	0.638	荒地	0.801	3.3	153	5.017
47	839	平原荒漠	0.956	荒地	0.998	2.3	169.1	4.929
48	849	扇缘带	0.945	荒地	1	7.8	279.4	2.991
49	913	平原荒漠	0.705	荒地	0.659	3.6	153.4	5.65
50	954	扇缘带	0.86	绿洲	0.538	5.9	236.4	4.363
51	967	平原荒漠	0.898	林地	0.749	−3.2	164.9	8.102
52	1019	高山	0.869	草甸	0.706	−7.8	172	6.649
53	1090	扇缘带	0.768	荒地	0.966	5.1	201.6	3.092
54	1091	平原荒漠	0.921	荒地	0.982	3.3	165.2	4.147
55	1102	扇缘带	0.95	草甸	0.85	5.6	332.8	3.982

(续表)

序号	区域 编号	中地貌统计信息		地带性植被统计信息		年平均 气温(℃)	年平均 降雨量(mm)	干燥度
		主要中地貌类型	面积百分比	主要植被类型	面积百分比			
56	1103	扇缘带	0.773	绿洲	0.767	4.2	328.5	3.942
57	1111	沙漠	1	荒地	0.959	2.7	158.3	4.83
58	1189	平原荒漠	0.79	荒地	0.929	7	259.7	3.262
59	1199	扇缘带	0.933	荒地	0.927	5.7	256	3.377
60	1216	沙漠	0.999	荒地	1	6.5	247.2	3.412
61	1219	扇缘带	0.881	荒地	0.7793	1.9	175.2	3.756
62	1221	平原荒漠	0.864	荒地	0.891	-0.5	174.5	4.289
63	1232	平原荒漠	0.786	草原	0.624	1.5	185	3.8
64	1278	扇缘带	0.912	荒地	0.952	9.1	84.9	12.082
65	1286	平原荒漠	0.833	荒漠	0.892	1.1	182.2	3.484
66	1344	平原荒漠	0.943	荒地	0.96	10.5	44.8	19.619
67	1403	平原荒漠	0.921	荒地	0.985	10.7	40.3	20.844
68	1514	平原荒漠	0.847	荒地	0.903	9.4	48	18.02

4 结语

本文提出的 NNI-FC 区划算法,是一种从最基本地理单元(即格网单元)进行划分的“自下而上”的递归区划算法。它可分为三个主要步骤:数据预处理、生成核心区、归并细碎区。NNI-FC 区划算法的时间复杂度可以用  $O(2n + 2nK \cdot \log_4(2n) + \log_2 n)$  来表示。若  $K$  远远小于  $n$ ,则算法的时间复杂度接近于线性;反之,若  $K$  接近于  $n$ ,则时间复杂度接近  $O(n^2)$ ,即核心区划分阈值  $\theta$  的大小直接影响算法的复杂度。

NNI-FC 需事先确定两个参数:核心区划分阈值  $\theta$ 、细碎区合并的面积比阈值  $\sigma$ 。 $\theta$  越大,则细碎区的个数越多,反之,若  $\theta$  过小,则可能导致部分核心区由于面积太小而不易辨识,且缺乏分析统计的意义;若  $\sigma$  过小,则可能将差异极大的多个小分区合并为一个大分区,这将与区划的分异原则相背。因此,阈值  $\theta$ 、 $\sigma$  的适当取值十分重要。为了兼顾区划结果的合理性与处理的高效, $\theta$  的取值应接近当前等级区划最小可辨识图斑面积对应的格网单元个数; $\sigma$  取值则应大一些(如:大于 0.7)。这样既可减少细碎区的个数,又能将性质最相近的小分区合理地归并为一个较大的分区;并且,当分区尺度变化时,可以很方便地将当前的等级区划方案扩展为多等级区划方案<sup>[15]</sup>。

综上,NNI-FC 算法的主要优点有:(1)充分考虑了数值属性与非数值属性的要素,并利用了“空间相邻、属性相似”的原则,因此,算法可以很好

地反映真实的地理空间信息;(2)通过设定适当的核心区划分阈值  $\theta$ ,即可提取某些具有特殊意义的但面积较小区域(如绿洲区);(3)对于任意给定的阈值  $\theta$ 、 $\sigma$ ,采用该算法的区划方案都是全局最优方案。

该算法的缺点主要有:(1)运算效率严重依赖于细碎区的边界特征,其边界越复杂,则算法的效率越低<sup>[16]</sup>;(2)阈值  $\theta$ 、 $\sigma$  的变化,会使算法的效率产生波动;(3)由于缺乏表征人类活动或对生态环境影响的恰当指标体系,算法中未考虑人类活动对生态区划的影响。

今后若能找到计算多边形距离的更好算法,则可以极大地降低 NNI-FC 算法的时间复杂度;若加入计算机可识别的表征人类活动影响生态环境的指标体系,则基于 NNI-FC 的生态区划将更趋合理。

参考文献:

[1] 刘国华,傅伯杰. 生态区划的原则及其特征[J]. 环境科学进展, 1998, 6(6):67-72.

[2] Omernik J M. Ecoregions: A Framework for Managing Ecosystems[J]. The George Wright Forum, 1995, 12(1): 35-50.

[3] 杨勤业,郑度,吴绍洪. 中国的生态地域系统研究[J]. 自然科学进展, 2002, 12 (3):287-291.

[4] 程叶青,张平宇. 生态地理区划研究进展[J]. 生态学报, 2006, 26(10): 3424-3433.

[5] 郑度,葛全胜,等. 中国区划工作的回顾与展望[J]. 地理研究, 2005, 24(13): 330-344.

[6] 吕晋,邬红娟,等. 主成分及聚类分析在水生态系统区划中的应用[J]. 武汉大学学报:理学版, 2005, 51(4): 461-466.

[7] 陈述彭,陈秋晓,周成虎. 网格地图与网格计算[J]. 测

- 绘科学, 2002, 27(4): 1-6.
- [8] 陈述彭, 周成虎, 陈秋晓. 格网地图的新一代[J]. 测绘科学, 2004, 29(4): 1-4.
- [9] Bailey R G. Delineation of Ecosystem Regions[J]. Environmental Management, 1983, 7(4): 365-373.
- [10] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. China Machine Press, 2006, 21-47.
- [11] Harding J S, Winterbourn M J. An Ecoregion Classification of the South Island, New Zealand[J]. J. of Environmental Management, 1997, 51(3): 275-287.
- [12] Guha S, Rastogi R, Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes[J]. Information Systems, 2000, 25(5): 345-366.
- [13] M. de Berg 等著, 邓俊辉译. 计算几何: 算法与应用[M]. 北京: 清华大学出版社, 2005, 71-87.
- [14] Bour O, Davy P. Clustering and Size Distribution of Fault Patterns[J]. Theory and Measurements. Geophysical Research Letters, 1999(29): 2001-2004.
- [15] Davis J H, Howe R W, Davis G J. A Multi-scale Spatial Analysis Method for Point Data[J]. Landscape Ecology, 2000(15): 99-114.
- [16] Nancy M A. Determining the Separation of Simple Polygons[J]. Journal of Computational Geometry & Applications, 1994, 4 (4): 457-474.
- [17] 新疆地理学会. 新疆地理手册[M]. 乌鲁木齐: 新疆人民出版社, 1993, 53-54.

## A Nearest Neighbor Index Clustering Algorithm for Ecological Regionalization Based on Multi-layer Grid Model

YUAN Yecheng<sup>1, 2</sup>, ZHOU Chenghu<sup>1</sup>, QIN Biao<sup>1</sup>, OU Yang<sup>1</sup>

(1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;

2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

**Abstract:** The article presented a Nearest Neighbor Index Fuzzy Clustering (NNI-FC) algorithm for ecological regionalization based on multi-layer grid model with consideration of spatial distribution of geographical factors such as climate, vegetation and topography. It's a "bottom-up" regionalization approach and solved the problem of how to determine the ecological regionalization's type and its boundary by calculating the Nearest Neighbor Index (NNI) and the similarity between grids. Numeric and non-numeric features were considered simultaneously in the NNI and similarity, so the algorithm integrated both qualitative and quantitative regionalization methods. The algorithm consisted of three consequence steps: data preprocessing, core region generation and fragmented region elimination. In the data preprocess, some numeric property values were transformed to non-numeric ones through classification of the combination of geographical factors. Next, core regions and fragmented regions were generated by ROCK algorithm, which clustering the adjacent discrete grids with the same property values. Then, on the basis of analyzing the fragmented regions' area coverage and its spatial distribution by using NNI, the algorithm divided the fragmented regions into small pieces and merged them into the core region which has the biggest similarity. Finally, an eco-regionalization scheme for the given natural section is formed. The experiment of ecological regionalization for North of Xinjiang shows that the algorithm achieves over 80% classification accuracy and can be very good at expressing the diversity of regional characteristics. Besides, different levels of eco-regionalization scheme can be obtained by adjusting the thresholds of the algorithm and its time complexity is between linear and quadratic ones depending on the thresholds.

**Key words:** ecological regionalization; multi-layer grid model; clustering; Nearest Neighbor Index; spatial distribution