doi: 10.6046/zrzyyg.2021214
引用格式: 孙肖,彭军还,赵锋,等. 基于空间统计学的高光谱遥感影像主成分选择方法[J]. 自然资源遥感,2022,34(2):37-46. (Sun X, Peng J H, Zhao F, et al. Principal component selection method for hyperspectral remote sensing images based on spatial statistics[J]. Remote Sensing for Natural Resources, 2022, 34(2): 37-46.)

基于空间统计学的高光谱遥感影像主成分选择方法

孙 肖[1], 彭军还[2], 赵 锋[3], 王晓阳[1], 吕 洁[2], 张登峰[4]
(1. 中国地质调查局廊坊自然资源综合调查中心,廊坊 065000; 2. 中国地质大学(北京)土地科学技术学院,北京 100083; 3. 中国地质调查局乌鲁木齐自然资源综合调查中心,乌鲁木齐 830057; 4. 中国地质调查局西安矿产资源调查中心,西安 710100)

摘要: 主成分分析是一种广泛使用的高光谱遥感影像降维方法,在面向任务的工作中,基于累计方差贡献率的主成分选择方法效果并不理想。针对主成分分析变换后主成分选择的问题,提出基于空间统计学的主成分选择方法。计算各主成分的半变异函数参数变程、拱高、基台值,综合变程和拱高/基台值实现主成分的选择。变程的大小用以判断每一个主成分空间相关性的范围,拱高/基台值的大小用以判断每一个主成分空间相关性的强弱。仿真实验证明了变程和拱高/基台值可以有效表达高光谱遥感影像空间相关性的范围和强弱。在真实高光谱遥感影像实验的基础上,从主观和客观 2 个方面来综合确定主成分选择的经验阈值,即变程为 2.5、拱高/基台值为 0.2。从基于支持向量机算法的分类结果来看,和传统方法相比,利用变程和拱高/基台值可以筛选出图像质量较好的主成分,不仅能够达到降维的目的,同时能够保证足够高的分类精度。

关键词: 高光谱; 主成分分析; 空间统计学; 半变异函数; 支持向量机
中图法分类号: P 962 **文献标志码**: A **文章编号**: 2097-034X(2022)02-0037-10

0 引言

在面向任务的高光谱遥感影像数据分析中,由于高光谱遥感影像波段数相对比较多,庞大的数据量给后续的分析处理带来了极大的挑战。在高光谱分类中,把分类精度随着波段数量的增加先升高后降低的现象叫作“Hughes”现象[1]。Chang[2]研究发现最高有 94% 的波段是可以舍弃的,而且不会影响分类的精度。因此,在高光谱遥感影像的研究中,一般会首先进行降维处理,主成分分析(principal component analysis, PCA)就是一种常用的线性降维方法[3]。

PCA 将数据的方差作为线性变换的标准,因此,一般按照变换后数据的累积方差进行主成分的选择。降维主要的目的就是降低数据维数的同时,尽可能保留信息,而选择方差较大的主成分必然会带来信息的损失。目前没有一种有效的方法来决定该选择哪个主成分。

Jolliffe[4]通过大量实验研究将 PCA 变换后主成分选择的经验阈值定为累计方差贡献率大于 0.85,但是该阈值在高光谱遥感的研究中具有局限性。PCA 在高光谱遥感领域的应用主要包含两大类:一类是应用于解混、变化检测、数据压缩、目标探测、去噪等的研究中,根据研究的目的和内容不同,一般选择某一特定主成分或某几个主成分进行研究[5-10];另一类是应用于分类研究,Chang 等[11]认为利用累计方差贡献率大于 0.99 的主成分进行分类研究效果比较好。Li 等[12]认为累计方差贡献率大于 0.9 就能保证分类精度大于 0.85。臧卓等[13]对高光谱遥感影像降维后的主成分进行分类测试,发现累计方差贡献率与分类精度没有必然联系,而主成分的个数对分类结果的影响较为明显,认为保留前 15~20 个主成分较为合适。黄鸿等[14-16]认为可以给定一定数量的主成分进行分类。臧卓等[17-19]逐次增加主成分数量进行分类,根据分类精度确定合适的主成分个数,该方法虽然能保证分类精度最高,但是效率较慢。Mather 等[20]指出不能仅依靠特征值对应的主成分来做图像分类,还应考虑图像的实际视觉效果。Rodarmel 等[21]分别采用编号 1—5、1—

收稿日期: 2021-07-14; **修订日期**: 2021-12-08
基金项目: 中国地质调查局项目“京津冀廊坊地区生态修复支撑调查”(编号: DD20208073)资助。
第一作者: 孙 肖(1988-),男,硕士,助理工程师,主要从事高光谱遥感解译研究。Email: sunxiao@mail.cgs.gov.cn。
通信作者: 彭军还(1964-),男,博士,教授,主要从事测绘理论研究。Email: pengjunhuan@163.com。

10、1—25、1—50 的主成分分段计算了分类的精度，认为可以用 5% ~ 10% 的主成分个数进行分类。Ibarrola - Ulzurrun 等[22]将常用的主成分选择方法分为 4 类(基于特征值、纹理特征、类别变换和感兴趣区分离)，分别利用前 2,5,10,15,20 主成分对这 4 类方法的分类精度进行了对比研究，认为特征值不是最适合的主成分选择方法，类别变换和感兴趣区分离需要人为的确定感兴趣区，因此纹理特征是比较适合的主成分选择方法，典型的纹理特征指标即信息熵。

以上主成分选择方法虽然取得了一定的效果，但是仍存在依据不充分、效率较低、主观性较强的问题。而且，PCA 变换结果不随噪声排列，方差也不能判断噪声的大小。以上方法会导致部分图像质量较好的主成分被舍去，而部分图像质量较差的主成分参与分类的现象。从实际应用来看，编号较大的一些主成分对分类结果也有一定的影响[23]。

目前，定量的 PCA 变换后主成分选择方法的研究还比较少。Zheng 等[24]提出 GA - Fisher 算法，能有效地增加编号较大的有效主成分。Zhang 等[25]利用免疫克隆选择算法对主成分进行二次处理，提出 ICSA - PCA 算法，在一定程度上解决了以上存在的问题。本文从空间统计学的角度，利用高光谱遥感影像的空间相关性，提出了一种定量的主成分选择方法。

1 研究方法

空间统计学是建立在相邻地理单元存在某种联系的基本假设之上的统计学，将统计学和现代图形计算技术结合起来，用直观的方法展现空间数据中所隐含的空间分布、空间模式以及空间相互作用等特征[26]。地统计学是空间统计学的重要组成部分，而半变异函数理论是地统计学处理空间数据的方法，是探索展布于空间并呈现出一定的随机性和结构性的自然现象的重要技术和方法，被用于描述空间相关性[27]。数据空间分布的相关性越大，即空间上聚集分布的现象越明显。若所测值不表现出任何空间依赖关系，那么，这一变量表现出空间不相关性或空间随机性。

变程、基台值、块金值是常用的表达空间自相关性的半变异函数基本参数。变程的大小反映了区域化变量影响范围的大小，或者说反映该变量的自相关尺度。在变程距离之内，空间上越靠近在一起的点之间的相关性越大，相隔距离大于变程的点之间没有自相关性。基台值与块金值之差表示由于采样数据中存在空间自相关性引起的方差变化范围，反映了数据的随机性，称为拱高。拱高和基台值的比值可以反映数据空间相关性的强弱[28-29]。利用空间数据的以上特性可以很好地研究主成分选择的问题。

本文利用变程、拱高/基台值两个半变异函数参数进行 PCA 后主成分的选择，结合以上原理，基于空间统计学的高光谱遥感影像主成分选择流程见图 1，主要过程如下：

1)对高光谱遥感影像数据集进行 PCA 变换，获取主成分。在去除各主成分的趋势项影响后对各主成分数据进行正态化转换。

2)选定理论半变异函数模型对计算出的各主成分的实验半变异函数进行拟合，从而获取各半变异函数参数变程、基台值、块金值，由此计算出用于主成分选择的变程、拱高/基台值两个参数。

3)联合变程、拱高/基台值进行主成分选择，结合高光谱遥感影像实验结果，从主观和客观两个方面来综合确定主成分选择的经验阈值。

4)利用支持向量机(support vector machine，SVM)算法进行分类，通过分类的 Kappa 系数、制图精度、选择的波段数等评价主成分选择方法的效果[30-31]。与 3 种传统选择主成分的方法进行比较，评价本文提出的方法的好坏。

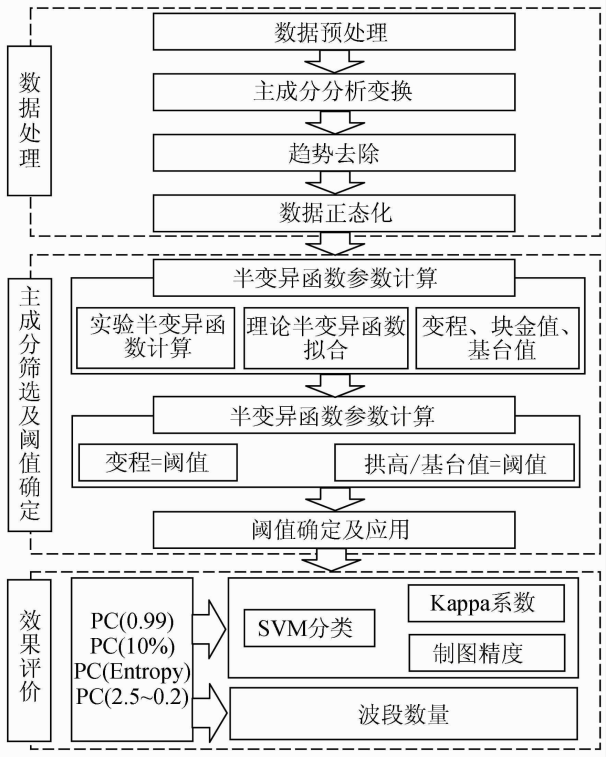


图 1 算法流程图

Fig. 1 Flowchart of algorithm

2 实验及其结果分析

2.1 仿真实验

为便于研究，仿真数据大小设计为 72 × 72。从

美国地质调查局网站提供的地物波谱库里随机选择 4 类地物波谱，按照规则格网设计仿真图像（格网大小 $W=1,3,\cdots,23$）。为更加接近真实数据情况，添加信噪比分别为 10,20,30,45 的零均值高斯噪声（图 2）。

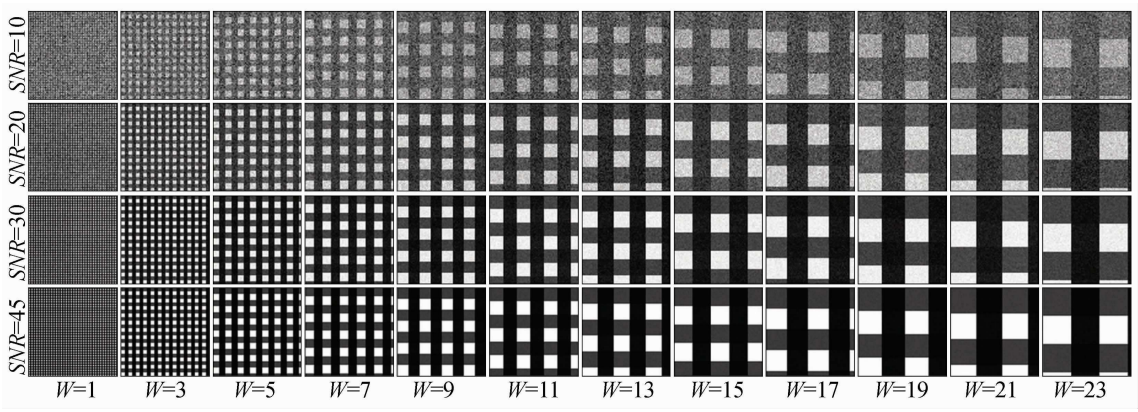


图 2　仿真图像（不同栅格大小 $W=1,3,\cdots,23$；信噪比 $SNR=10,20,30,45$）

Fig. 2　Simulation images (different grid sizes $W=1,3,\cdots,23$; $SNR=10,20,30,45$)

如图 3、图 4 所示，当 $W=1$ 时，即相邻的像元均为不同地物，图像以随机性为主，实验半变异函数主要表现为块金值。当 $W=3,5,\cdots,23$ 时，图像的变程计算结果与设计的网格大小基本一致，反映了图像空间相关性的范围大小。拱高/基台值计算结果受噪声影响比较明显，可以作为利用变程选择主成分的辅助参数。仿真实验从数据的空间相关性和随机性两个方面验证了利用半变异函数参数拱高/基台值、变程进行 PCA 后主成分选择的有效性。

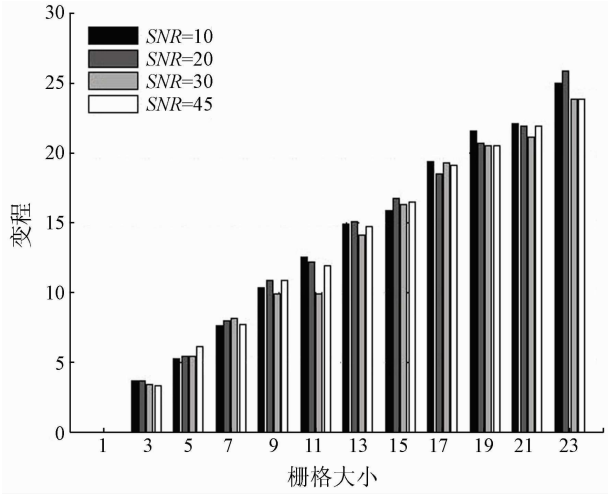


图 3　不同栅格大小仿真图像计算的变程结果

Fig. 3　Results of the range of the simulation image

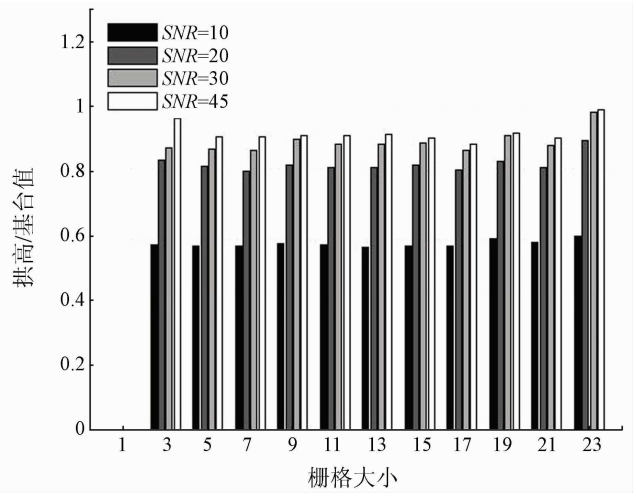


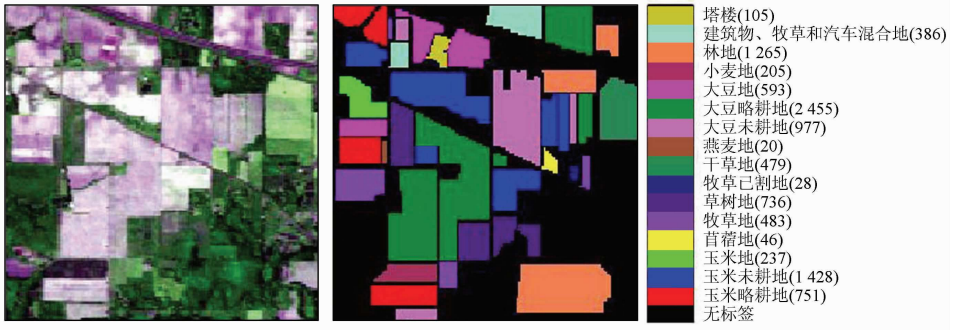
图 4　不同栅格大小仿真图像计算的拱高/基台值结果

Fig. 4　Results of the partial sill/sill of the simulation image

2.2　实际数据

2.2.1　实验数据集

Indian Pines 高光谱数据集是 1992 年由 AVIRIS 传感器获取的印第安纳州西北部农业区的高光谱遥感影像数据的一部分，图像大小为 145 × 145 像素，包含 16 类地物（图 5）。AVIRIS 数据光谱范围为 0.4 ~ 2.45 μm，共 224 个波段，空间分辨率 20 m。

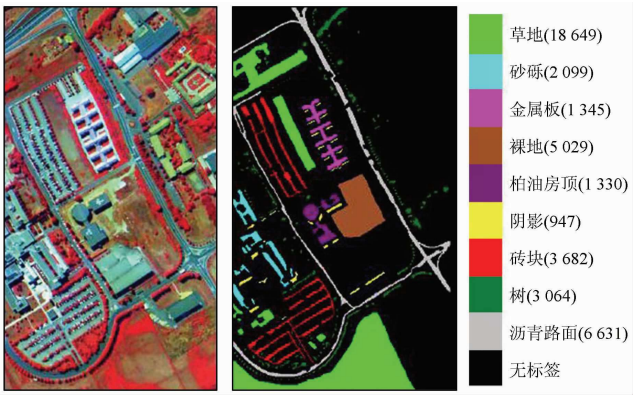


（a）影像　　（b）真实标签

图 5　Indian Pines 数据集

Fig. 5　Indian Pines data set

ROSIS 传感器于 2003 年在意大利的北部 Pavia 大学获取了 2 幅高光谱影像，University 高光谱数据是该数据集的其中之一（图 6）。图像大小为 610 × 340 像素，空间分辨率 1.3 m，包含 9 类地物。ROSIS 传感器共 103 个波段，光谱范围为 0.43 ~ 0.86 μm。

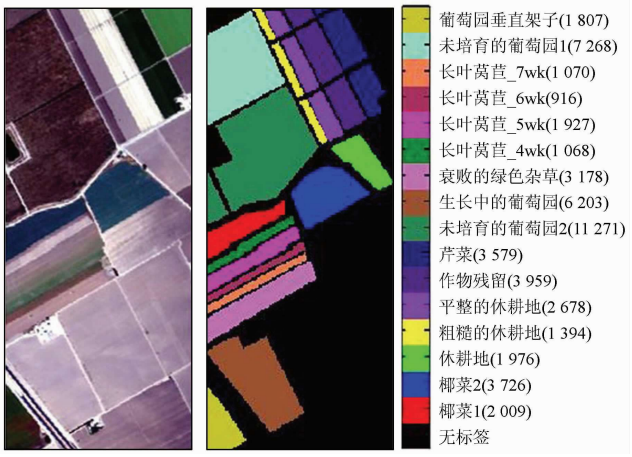


（a）影像 （b）真实标签

图 6 Pavia U 数据集

Fig. 6 Pavia university data set

Salinas 高光谱数据集由 AVIRIS 传感器获取的美国加利福尼亚州萨利纳斯山谷区域。图像大小为 512 × 217 像素，空间分辨率 3.7 m，包含 224 个波段，该数据集地物类别包含 16 类（图 7）。



（a）影像 （b）真实标签

图 7 Salinas 数据集

Fig. 7 Salinas data set

研究中使用的 Indian Pines，Pavia University（简写为 Pavia U）和 Salinas3 种高光谱数据集获取网站网址如下：http：//www. ehu. eus/ccwintco/index. php? title = Hyperspectral_Remote_Sensing_Scenes。

2.2.2 数据处理及结果

数据处理主要包含无信息波段的剔除、趋势项去除、数据正态化、实验半变异函数计算、理论半变异函数拟合等过程。

由于受水汽影响，Indian Pines 数据集部分波段的成像效果比较差，本数据集中剔除的波段为 104 ~ 108，150 ~ 163，220。同样，Salina 数据集剔除 108 ~ 112，154 ~ 167，224 波段。通过二阶数据漂移估计，消除趋势项影响。空间统计学中一般都假设数据是服从正态分布，本文在正态检验的基础上，采用常态得分变换（normal score transform，NST）方法进行数据正态化的转换，该方法相比传统方法不受数据负值的影响[32]。

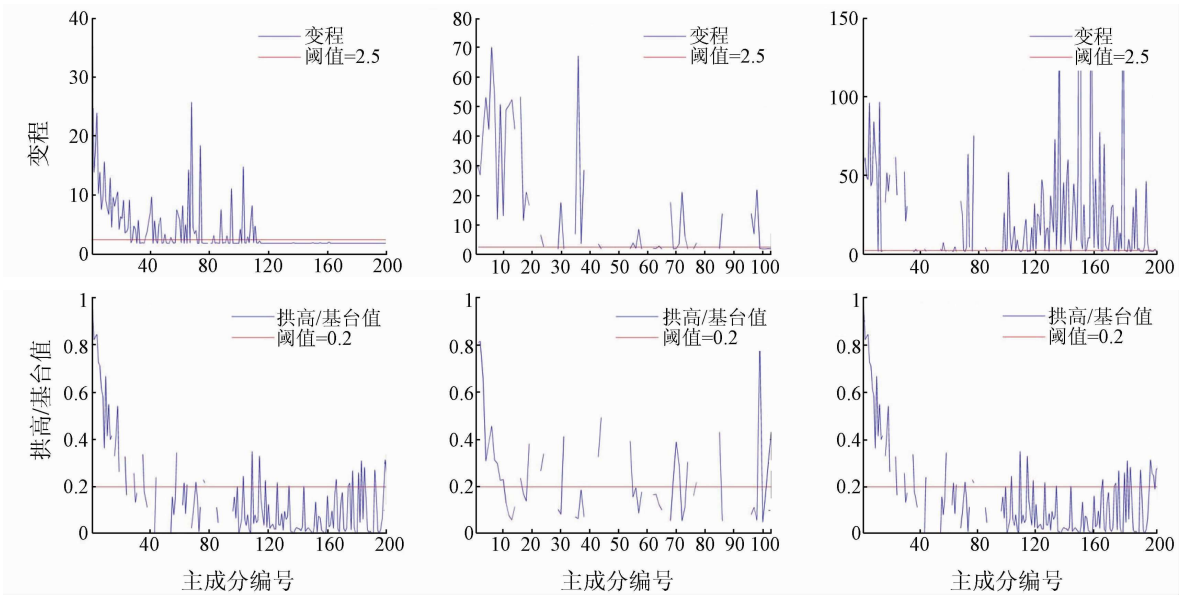
对于高光谱遥感影像的实验半变异函数的计算，一般是分别计算图像的 0°，45°，90° 和 135° 方向的实验半变异函数曲线，研究其曲线变化特征，通过套和获取最终的实验半变异函数。选择 5 种常用的模型——球状模型、指数模型、高斯模型、线性有基台值模型和普通线性模型模型，采用线性规划法进行拟合，最终，选择交叉验证结果最佳的一个作为最终的理论模型[33-34]。一般采用交叉验证后判定系数 R^2 较大，或者残差标准差较小的理论模型作为最终结果。实验中采用判定系数 R^2 和残差标准差的比值作为理论模型的选择标准。半变异函数参数计算结果如图 8 所示，为便于图面表达，图示中舍弃了拱高/基台值小于 0.05 的无意义主成分。

2.2.3 主成分选择方法

从变程、拱高/基台值的计算结果来看（图 8），高光谱遥感影像 PCA 变换后编号较大的主成分主要表现为随机噪声，结果主要体现为块金值，该特征与仿真实验比较一致，主成分选择中舍弃该类主成分。单独利用变程或者拱高/基台值也可以进行主成分的筛选，但是对于编号较大的主成分计算出的无意义结果不能很好的判断。同时，变程和拱高/基台值的结果具有明显的互补性，对于一些无意义结果，同时通过 2 组参数可以有效的进行剔除。因此，本文提出综合利用以上 2 组参数进行主成分选择的思路。

2.2.4 阈值确定

利用变程、拱高/基台值选择主成分关键的问题就是阈值的确定。通过仿真实验可以知道，图像最小的空间相关性范围大小为 2，即相邻像元是相关的，也就是变程为 2。为便于研究，将变程增加到 2.5 作为对比实验。拱高/基台值体现了图像的随机性，一般认为当该值小于 0.2 ~ 0.25 时，数据表现为强的随机性。为了便于研究，分别采用拱高/基台值为 0.2 和 0.25 进行对比研究。在此基础上，分别测试了变程 = 2、拱高/基台值 = 0.2（表示为 PC（2 ~ 0.2））；变程 = 2、拱高/基台值 = 0.25（表示为 PC（2 ~ 0.25））；变程 = 2.5、拱高/基台值 = 0.2（表示为 PC（2.5 ~ 0.2））；变程 =



(a) Indian Pines 数据集参数　　(b) Pavia U 数据集参数　　(c) Salinas 数据集参数

图 8　真实数据半变异函数参数计算结果图

Fig. 8　The results of the semi – variogram parameters of real data

2.5、拱高/基台值 =0.25(表示为 PC(2.5 ~0.25))的主成分选择效果。

为了说明利用几种阈值进行主成分选择的效果,从主观和客观 2 个方面进行评价。主观评价方法即观察利用几种阈值选择出的主成分的图像质量。以 Indian Pines 数据集为例,图 9 为该数据集 PCA 变换后各主成分的缩略图,表 1 为不同阈值筛选的主成分。图 9 中排列顺序为从左至右,从上至下,主成分编号依次增大,表示为 PC1, PC2, …, PC200。实验中发现,无论哪种阈值组合都可以剔除图面质量较差的 PC9。当拱高/基台值固定时,增大变程会剔除更多的主成分。从表 1 结果来看,当拱高/基台值 = 0.25 时,PC103 被剔除,同时,当变程由 2 增加到 2.5 会导致 PC108, PC109 被剔除。从图 9 来看,PC103, PC108, PC109 主成分的图像细节仍然比较清楚。当变程固定时,减小拱高/基台值会增加更多主要表现为随机性的主成分。从表 1 结

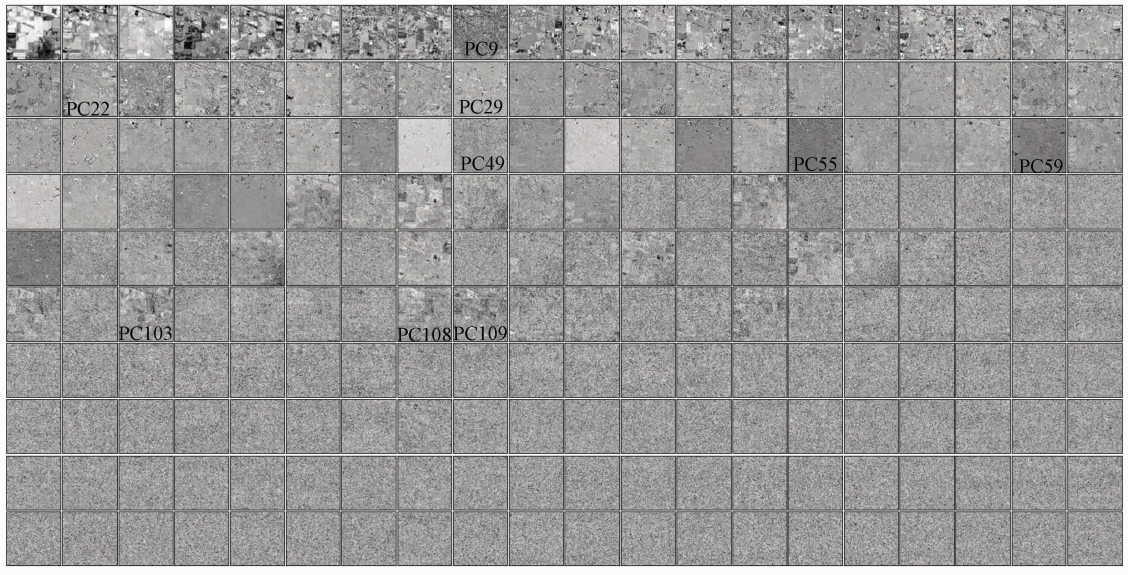


图 9　Indian Pines 数据集 PCA 后各主成分缩略图

Fig. 9　Thumbnails of each principal component after the PCA transformation of the Indian Pines data set

表 1　Indian Pines 数据集不同阈值筛选的主成分

Tab. 1　Principal component selected by different thresholds in Indian Pines data set

不同方法	筛选的主成分编号
PC(2 ~0.25)	1 ~8,10 ~21,23 ~25,28,38,66,68,88,92,101,108,109
PC(2 ~0.2)	1 ~8,10 ~25,28,29,38,55,59,66,88,92,101,103,108,109
PC(2.5 ~0.25)	1 ~8,10 ~21,23 ~25,28,38,66,68,88,92,101
PC(2.5 ~0.2)	1 ~8,10 ~25,28,29,38,59,66,68,88,92,101,103

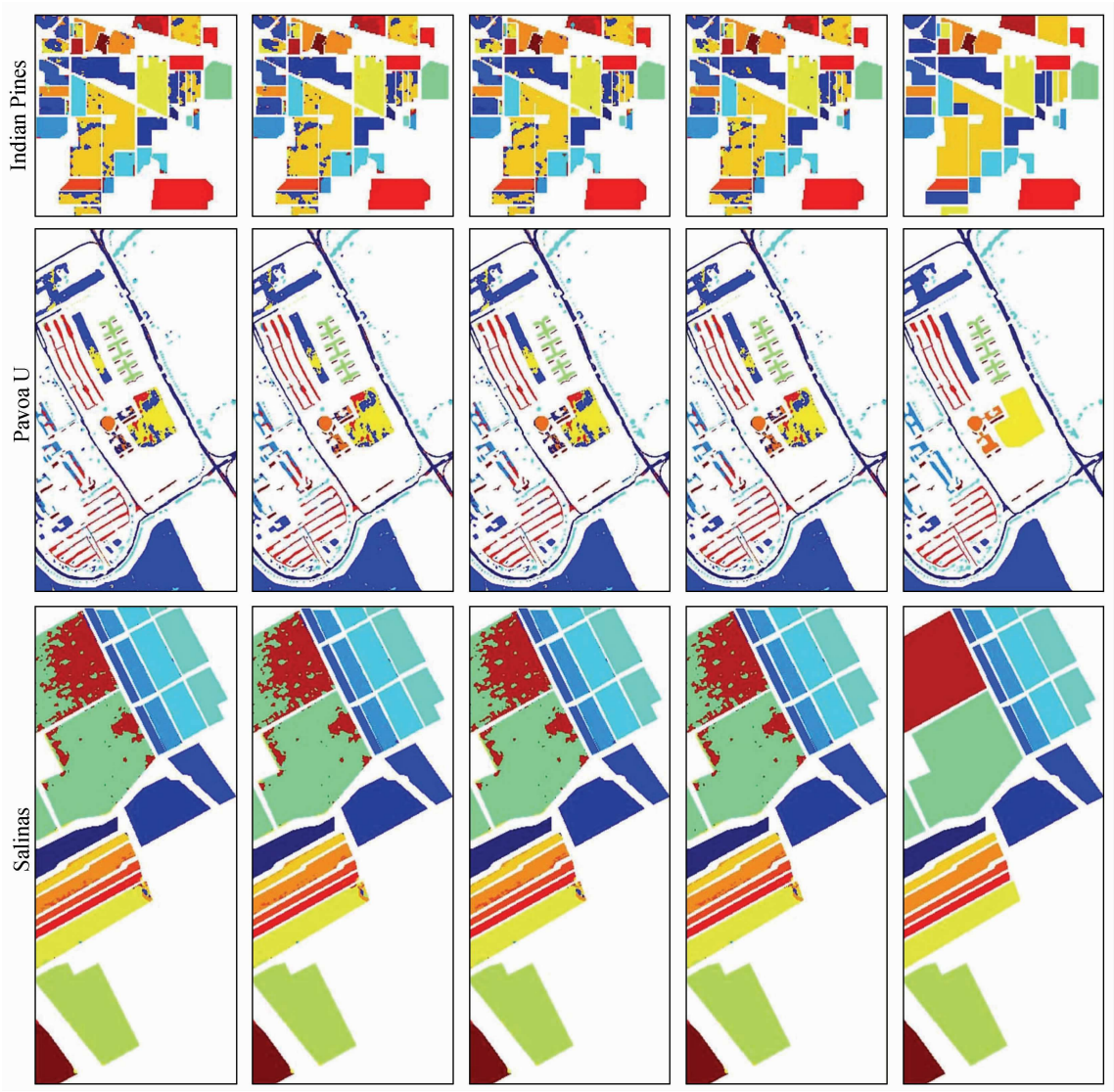
果来看，当变程 =2 时，拱高/基台值由 0.25 减小到 0.2 会将 PC22，PC29，PC55，PC59，PC103 筛选进来。从图 9 来看，PC22，PC29，PC59，PC103 图像细节仍然比较清楚，但是 PC55 图像质量较差。因此还不能完全说明拱高/基台值的阈值确定为哪个比较合适。

客观评价方法是计算所选择的主成分分类的 Kappa 系数进行对比评价。本文利用常用的 SVM 方法对几种阈值的筛选结果进行分类，分类过程通过 ENVI 软件实现，利用 RBF 核，设置 *gamma* = 0.1，*penalty* = 100。利用几组阈值选择的主成分进行分类，从表 2 的分类精度结果来看均能得到较高的分类精度，从图 10 的分类效果来看结果差别不明显。综合考虑主观评价结果，利用 PC(2 ~ 0.2) 筛选出的主成分较多，利用 PC(2.5 ~ 0.25) 筛选出的主成分较少，PC(2.5 ~ 0.2) 总体分类精度较高，因此，最终选择变程 =2.5，拱高/基台值 =0.2 作为主成分选择的阈值。

表 2 不同阈值的 Kappa 系数

Tab. 2 Kappa coefficient of different thresholds

数据集	PC (2 ~ 0.25)	PC (2 ~ 0.2)	PC (2.5 ~ 0.25)	PC (2.5 ~ 0.2)
Indian Pines	0.912	0.912	0.905	0.906
Pavia U	0.856	0.881	0.858	0.883
Salinas	0.953	0.953	0.954	0.953



(a) PC(2 ~ 0.25) 分类结果　(b) PC(2 ~ 0.2) 分类结果　(c) PC(2.5 ~ 0.25) 分类结果　(d) PC(2.5 ~ 0.2) 分类结果　(e) 真实标签

图 10 Indian Pines，Pavia U 和 Salinas 数据集分类结果图

Fig. 10 Classification results of Indian Pines, Pavia U and Salinas data sets

2.2.5 效果评价

为了进一步验证提出的方法的有效性，采用了 2 种使用较为广泛的传统方法和一种创新方法进行对比研究。第一种是利用累积方差贡献率进行主成分筛选，实验中阈值定为 0.99（表示为 PC(0.99)）[11]；第二种是 Rodarmel 等[21]提出的可以用 5% ~10% 的主成分个数进行分类研究，实验中采用分类效果较好的 10%（表示为 PC(10%)）；

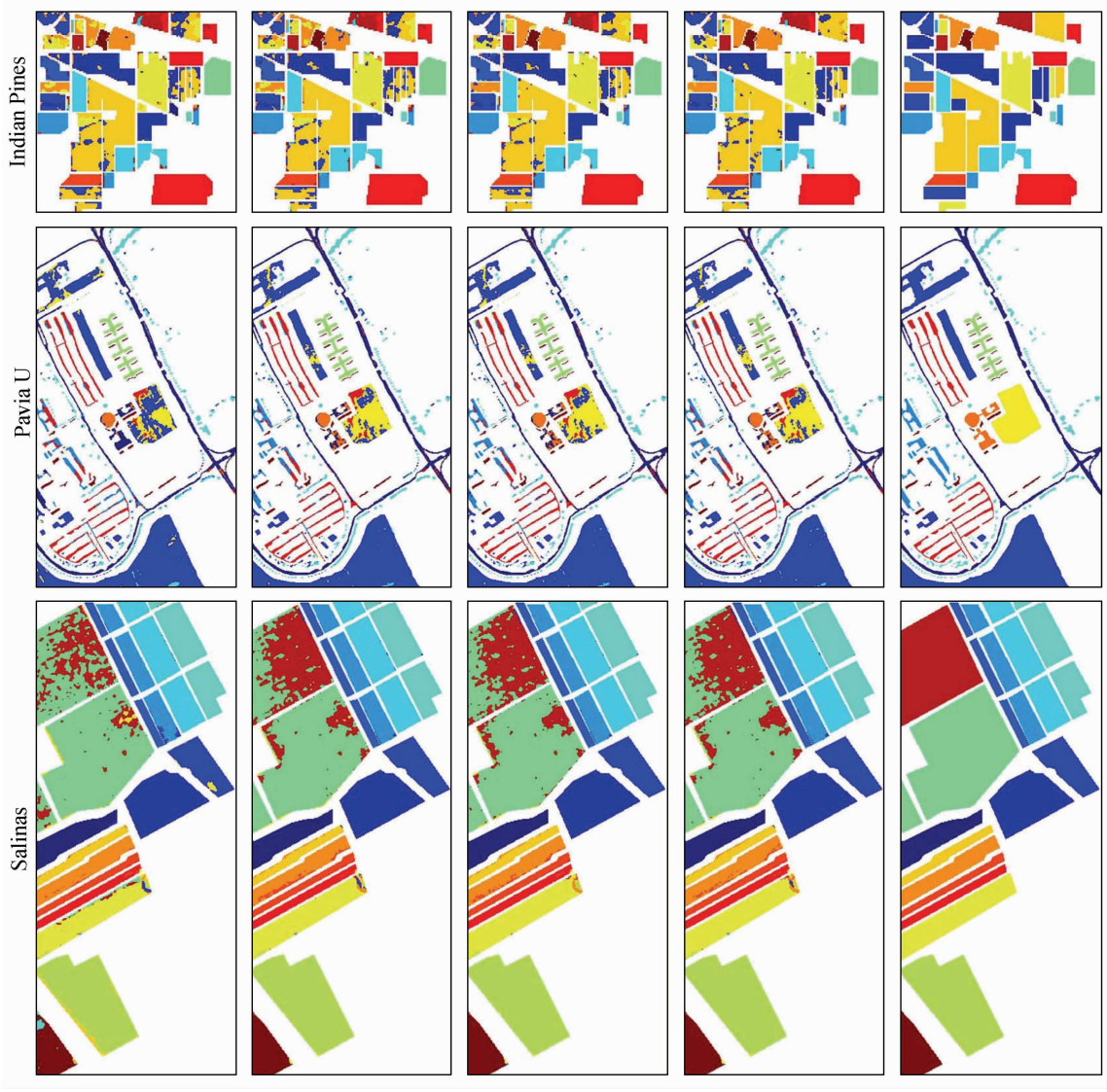
第三种是 Ibarrola – Ulzurrun 等[22]创新提出的反映纹理特征指标即信息熵，选择大于信息熵标准差的主成分进行分类（表示为 PC（Entropy））。

各主成分选择方法分类精度统计结果见表 3，分类结果图见图 11。对于 Indian Pines 和 Salinas 数据集，利用 PC（2.5 ~ 0.2）分类的结果优于其他方法。对于 Pavia U 数据集，利用 PC（2.5 ~ 0.2）分类的结果明显优于 PC（0.99），且与 PC（10%）、PC（Entropy）2 种方法的分类精度差别不大。

表 3　不同方法 Kappa 系数

Tab. 3　Kappa coefficient of different methods

数据集	PC（0.99）	PC（10%）	PC（Entropy）	PC（2.5 ~ 0.2）
Indian Pines	0.864	0.853	0.899	0.906
Pavia U	0.718	0.884	0.884	0.883
Salinas	0.910	0.952	0.951	0.953



（a）PC（0.99）分类结果　（b）PC（10%）分类结果　（c）PC（Entropy）分类结果　（d）PC（2.5 ~ 0.2）分类结果　（e）真实标签

图 11　Indian Pines，Pavia U 和 Salinas 数据集分类结果图

Fig. 11　Classification results of Indian Pines，Pavia U and Salinas data sets

表 4 为 Indian Pines 数据集各地类的制图精度统计结果，数据为百分比。从表 4 可知，本文方法对于林地、大豆略耕地、燕麦地、牧草已割地、牧草地、玉米地、玉米未耕地和玉米略耕地的分类精度优于其他方法。同时，本文方法筛选出的主成分对于数量较少的地物较为敏感，塔楼、小麦地、燕麦地、牧草已割地、玉米地、苜蓿地、玉米地等分类效果明显优于其他方法。表 5 为 Pavia U 数据集各地类的制图精度统计结果，数据为百分比。从表 5 可知，本文方法对于裸地、柏油房顶、树的分类精度优于其他方法。同时，本文方法筛选出的主成分对于数量较少的树较为敏感，分类效果优于其他方法。表 6 为 Salinas 数据集各地类的制图精度统计结果，数据为百分比。从表 6 可知，该数据集各地类总体分类精度比较高，本文方法对于未培育的葡萄园、长叶莴苣的分类精度优于其他方法。同时，本文方法筛选出的

表 4 Indian Pines 数据集制图精度

Tab. 4 Prod. Acc of Indian Pines data set

地物种类	PC (0.99)	PC (10%)	PC (Entropy)	PC (2.5 ~ 0.2)
塔楼	100	100	100	100
建筑物	74.02	75.2	81.45	79.53
林地	98.89	98.89	99.43	99.51
小麦地	100	100	100	100
大豆地	84.46	79.77	98.4	97.95
大豆略耕地	86.78	85.56	90.82	92.7
大豆未耕地	89.4	89.18	88.26	86.98
燕麦地	77.78	77.78	60	88.89
干草地	100	100	100	100
牧草已割地	81.82	90.91	94.12	100
草树地	100	99.77	100	100
牧草地	100	99.24	100	100
苜蓿地	85.71	80	86.84	85.71
玉米地	95.95	93.92	96.95	97.97
玉米未耕地	73.32	73.08	80.39	82.93
玉米略耕地	62.14	57.96	65.44	65.54

表 5 Pavia U 数据集制图精度

Tab. 5 Prod. Acc of Pavia U data set

地物种类	PC (0.99)	PC (10%)	PC (Entropy)	PC (2.5 ~ 0.2)
草地	95.3	94.72	94.72	93.54
砂砾	42.87	72.63	72.63	72.00
金属板	100	99.91	99.91	99.82
裸地	31.51	79.64	79.64	81.02
柏油房顶	37.84	67.86	67.86	74.26
阴影	100	100	100	99.83
砖块	88.78	94.31	94.31	93.75
树	94.58	95.18	95.18	95.41
沥青路面	94.45	97.07	97.07	96.44

表 6 Salinas 数据集制图精度

Tab. 6 Prod. Acc of Salinas data set

地物种类	PC (0.99)	PC (10%)	PC (Entropy)	PC (2.5 ~ 0.2)
葡萄园垂直架子	98.23	99.92	99.84	99.83
未培育的葡萄园 1	46.83	74.3	72.72	75.26
长叶莴苣_7wk	98.66	99.81	99.04	99.81
长叶莴苣_6wk	98.95	99.48	99.12	99.48
长叶莴苣_5wk	100	100	99.89	100
长叶莴苣_4wk	100	100	100	100
衰败的绿色杂草	99.35	100	100	100
生长中的葡萄园	100	100	100	100
未培育的葡萄园 2	94.21	94.62	94.62	94.77
芹菜	100	100	100	100
作物残留	100	100	100	100
平整的休耕地	99.83	100	100	100
粗糙的休耕地	100	100	100	100
休耕地	100	100	100	100
榔菜 2	99.95	100	100	100
榔菜 1	99.49	99.91	100	100

主成分对于类别较少的长叶莴苣_6wk 地物较为敏感,分类效果优于其他方法。表 7 为不同方法所选择的主成分个数。从表 7 可知,利用累计方差贡献率大于 0.99 选择的主成分个数因数据不同差别比较大,直接影响分类效果,不能作为一种适用于所有遥感影像数据的方法。利用 10% 的主成分个数的主成分开展分类效果可以,但无法解释其物理意义,且结果受波段总数影响较大,结果具有随机性。利用信息熵选择的主成分个数因数据不同差别比较大,当地物易分类时,所选择的主成分个数过多。利用信息熵进行主成分选择会选择出无信息的个别主成分,例如 Indian Pines 数据集的第 49 主成分,从图 9 来看,该主成分无明显的图像信息。本文方法受数据影响较小,均能筛选出数量适中的主成分。本文方法不仅可以剔除一些编号虽然较小,但是图像质量比较差的主成分,而且可以将编号较大,但是图像质量较好的图像参与运算。

表 7 不同数据集筛选出的主成分信息

Tab. 7 Principal component information selected by different data sets

数据集	不同方法	筛选的主成分编号	个数
Indian Pines	PC(0.99)	1 ~ 25	25
	PC(10%)	1 ~ 20	20
	PC(Entropy)	1 ~ 46,49	47
	PC(2.5 ~ 0.2)	1 ~ 8,10 ~ 25,28,29,38,59,66,68,88,92,101,103	34
Pavia U	PC(0.99)	1 ~ 4	4
	PC(10%)	1 – 10	10
	PC(Entropy)	1 – 10	10
	PC(2.5 ~ 0.2)	1 ~ 10,16,19,23,24,43,71	16
Salinas	PC(0.99)	1 ~ 3	3
	PC(10%)	1 ~ 20	20
	PC(Entropy)	1 ~ 100	100
	PC(2.5 ~ 0.2)	1 ~ 10,12,15 ~ 23,25,28,29,34,42,61,66,75,77,118,126,134	32

3 结论

本文提出一种基于空间统计学的 PCA 变换后主成分选择的新方法,利用半变异函数参数变程、拱高/基台值的特性进行 PCA 变换后主成分的选择,取得了理想的效果,得出如下结论:

1)仿真实验证明了变程、拱高/基台值可以有效表达高光谱遥感影像空间相关性的范围和强弱。

2)变程、拱高/基台值的结果具有明显的互补性,联合两组参数可以有效剔除无意义主成分。

3)对比变程 2 ~ 2.5 和拱高/基台值 0.2 ~ 0.25 的不同参数组合的分类结果,变程 2.5、拱高/基台值 0.2 的参数组合可以更加有效地筛选主成分。

4)和传统方法相比,本文提出的方法可以剔除主成分编号较小,但是图像质量较差的主成分,同时,筛选出主成分编号较大,但是图像质量较好的主

成分。

5)在基于分类的研究中,利用变程 2.5、拱高/基台值 0.2 进行 PCA 变换后主成分选择,不仅能够达到降维的目的,同时能够保证足够高的分类精度。和传统方法相比,本文方法可以更好地识别数量比较少的地类。

由于半变异函数参数的计算也是一个研究比较多的问题,确定更加准确的半变异函数参数会对主成分的选择产生一定的影响。除此之外,本文方法仅对 PCA 变换后的主成分选择进行了探讨,同时可以推广到最大噪声分数变换、独立成分分析等降维方法中去。

参考文献(References):

[1] Hughes G. On the mean accuracy of statistical pattern recognizers [J]. IEEE Transactions on Information Theory, 1968, 14(1): 55 – 63.

[2] Chang C – I. Hyperspectral Data Exploitation: Theory and Applications[M]. John Wiley & Sons Inc, Hoboken, New Jersey, 2007, 205.

[3] 宋海峰,陈广胜,杨巍巍. 基于 PCA 的高光谱遥感图像分类[J]. 测绘工程,2017,26(12):17 – 20,26.
Song H F, Chen G S, Yang W W. Principal component analysis for hyper spectral image classification[J]. Engineering of Surveying and Mapping, 2017, 26(12): 17 – 20, 26.

[4] Jolliffe I T. Principal Component Analysis[M]. Wiley, 2nd edition, 2002.

[5] Pu R, Gong P, Tian Y, et al. Invasive species change detection using artificial neural networks and CASI hyperspectral imagery [J]. Environmental Monitoring and Assessment, 2008, 140(1 – 3): 15 – 32.

[6] Chen G, Qian S. Denoising and dimensionality reduction of hyperspectral imagery using wavelet packets, neighbour shrinking and principal component analysis[J]. International Journal of Remote Sensing, 2009, 30(18): 4889 – 4895.

[7] Canty M J, Nielsen A A. Linear and kernel methods for multivariate change detection[J]. Computers & Geosciences, 2012, 38(1): 107 – 114.

[8] Foca G, Ferrari C, Ulrici A, et al. The potential of spectral and hyperspectral – imaging techniques for bacterial detection in food: A case study on lactic acid bacteria[J]. Talanta, 2016, 2016(153): 111 – 119.

[9] Li B, Hou B, Zhang D, et al. Pears characteristics (soluble solids content and firmness prediction, varieties) testing methods based on visible – near infrared hyperspectral imaging[J]. Optik International Journal for Light and Electron Optics, 2016, 127(5): 2624 – 2630.

[10] 韩彦岭,崔鹏霞,杨树瑚,等. 基于残差网络特征融合的高光谱图像分类[J]. 国土资源遥感,2021,33(2):11 – 19. doi:10.6046/gtzyyg.2020209.
Han Y L, Gui P X, Yang S H, et al. Classification of hyperspectral image based on feature fusion of residual network[J]. Remote Sensing for Land and Resources, 2021, 33(2): 11 – 19. doi: 10.6046/gtzyyg.2020209.

[11] Chang Y, Wang Y, Liu T, et al. Fault diagnosis of a mine hoist using PCA and SVM techniques[J]. Journal of China University of Mining and Technology, 2008, 18(3): 327 – 331.

[12] Li C, Liu L, Lei Y, et al. Clustering for HSI hyperspectral image with weighted PCA and ICA[J]. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2017, 32(5): 3729 – 3737.

[13] 臧 卓,林 辉,杨敏华. 利用 PCA 算法进行乔木树种高光谱数据降维与分类[J]. 测绘科学,2014,188(2):146 – 149.
Zang Z, Lin H, Yang M H. Dimension reduction and classification of hyperspectral data of tree species using PCA algorithm[J]. Science of Surveying and Mapping, 2014, 188(2): 146 – 149.

[14] 黄 鸿,曲焕鹏. 基于半监督稀疏鉴别嵌入的高光谱遥感影像分类[J]. 光学精密工程,2014,22(2):434 – 442.
Huang H, Qu H P. Hyperspectral remote sensing image classification based on SSDE[J]. Optics and Precision Engineering, 2014, 22(2): 434 – 442.

[15] Lee C, Youn S, Jeong T, et al. Hybrid compression of hyperspectral images based on PCA with pre – encoding discriminant information [J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(7): 1491 – 1495.

[16] 袁 林,胡少兴,张爱武,等. 基于深度学习的高光谱图像分类方法[J]. 人工智能与机器人研究,2017,6(1):31 – 39.
Yuan L, Hu S X, Zhang A W, et al. A classification method for hyperspectral imagery based on deep learning[J]. Artificial Intelligence and Robotics Research, 2017, 6(1): 31 – 39.

[17] 臧 卓,林 辉,杨敏华. ICA 与 PCA 在高光谱数据降维分类中的对比研究[J]. 中南林业科技大学学报,2011,31(11):18 – 22.
Zang Z, Lin H, Yang M H. Comparative study on descending dimension classification of hyperspectral data between ICA algorithm and PCA algorithm[J]. Journal of Central South University of Forestry & Technology, 2011, 31(11): 18 – 22.

[18] 叶 珍,何明一. PCA 与移动窗小波变换的高光谱决策融合分类[J]. 中国图象图形学报,2015,20(1):132 – 139.
Ye Z, He M Y. PCA and windowed wavelet transform for hyperspectral decision fusion classification[J]. Journal of Image and Graphics, 2015, 20(1): 132 – 139.

[19] Abdolmaleki M, Fathianpour N, Tabaei M. Evaluating the performance of the wavelet transform in extracting spectral alteration features from hyperspectral images[J]. International Journal of Remote Sensing, 2018, 2018: 1 – 19.

[20] Mather P, Koch M. Computer processing of remotely – sensed images[M]. John Wiley & Sons, Ltd, 2011, 265.

[21] Rodarmel C, Shan J. Principal component analysis for hyperspectral image classification[J]. Surveying and Land Information Systems, 2002, 62(2): 115 – 122.

[22] Ibarrola – Ulzurrun E, Marcello J, Gonzalo – Martin C. Assessment of component selection strategies in hyperspectral imagery[J]. Entropy, 2017, 19: 1 – 17.

[23] Stephan K, Hibbitts C, Hoffmann H, et al. Reduction of instrument – dependent noise in hyperspectral image data using the principal component analysis: Applications to Galileo NIMS data[J]. Planetary and Space Science, 2008, 56(3 – 4): 406 – 419.

[24] Zheng W, Lai J, Yuen P. GA – Fisher: A new LDA – based face

recognition algorithm with selection of principal components[J]. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), 2005, 35(5): 1065 - 1078.

[25] Zhang X, Li R, Jiao L. Feature extraction combining PCA and immune clonal selection for hyperspectral remote sensing image classification[C]. International Conference on Artificial Intelligence & Computational Intelligence, IEEE, 2009.

[26] Tobler W. Cellular geography[M]. Springer Netherlands, 1979, 379 - 386.

[27] 戴平生,陈建宝. 空间统计学研究应用综述[C]//烟台:国际应用统计学术研讨会,2008.
Dai P S, Chen J B. Review of spatial statistics research applications [C]//Yantai: International Symposium on Applied Statistics, 2008.

[28] Liu Q, Sun J, Chen Y, et al. Spatial variability of soil heavy metals at different sampling scales[J]. Chinese Journal of Soil Science, 2009, 40(6): 1406 - 1410.

[29] 刘昌振,舒 红,张 志,等. 基于多尺度分割的高分遥感图像变异函数纹理提取和分类[J]. 国土资源遥感,2015,27(4):47 - 53. doi:10.6046/gtzyyg.2015.04.08.
Liu C Z, Shu H, Zhang Z, et al. Variogram texture extraction and classification of high resolution remote sensing images based on multi - resolution segmentation [J]. Remote Sensing for Land and Resources, 2015, 27(4): 47 - 53. doi: 10.6046/gtzyyg.2015.04.08.

[30] 张 亮. 基于 PCA 和 SVM 的高光谱遥感图像分类研究[J]. 光学技术,2008,34(s1):184 - 187.
Zhang L. Study on the hyperspectral remote sensed image classify based on PCA and SVM[J]. Optical Technique, 2008, 34(s1): 184 - 187.

[31] 王 华,李卫卫,李志刚,等. 基于多尺度超像素的高光谱图像分类研究[J]. 自然资源遥感,2021,33(3):63 - 71. doi:10.6046/zrzyyg.2020344.
Wang H, Li W W, Li Z G, et al. Hyperspectral image classification based on multiscale superpixels[J]. Remote Sensing for Natural Resources2021, 33(3): 63 - 71. doi: 10.6046/zrzyyg.2020344.

[32] Goovaerts P. Geostatistics for natural resources evaluation[M]. NewYork City: Oxford University Press, 1997, 266.

[33] Cressie N. Fitting variogram models by weighted least squares[J]. Mathematical Geology, 1985, 17(5): 563 - 586.

[34] 矫希国,刘 超. 变差函数的参数模拟[J]. 物探化探计算技术,1996,18(2):157 - 161.
Jiao X G, Liu C. Estimation of variation parameter[J]. Computing Techniques for Geophysical and Geochemical Exploration, 1996, 18(2): 157 - 161.

Principal component selection method for hyperspectral remote sensing images based on spatial statistics

SUN Xiao[1], PENG Junhuan[2], ZHAO Feng[3], WANG Xiaoyang[1], LYU Jie[2], ZHANG Dengfeng[4]

(1. *Langfang Natural Resources Comprehensive Survey Center, China Geological Survey, Langfang* 065000, *China*; 2. *School of Land Science and Technology, China University of Geosciences (Beijing), Beijing* 100083, *China*; 3. *Urumqi Natural Resources Comprehensive Survey Center, China Geological Survey, Urumqi* 830057, *China*; 4. *Xi'an Center of Mineral Resources Survey, China Geological Survey, Xi'an* 710100, *China*)

Abstract: The principal component analysis is a widely used method for dimensionality reduction of hyperspectral remote sensing images. In task - oriented work, the principal component selection method based on cumulative variance contribution rate is not ideal. To address the problem of principal component selection after principal component analysis transformation, a method of principal component selection based on spatial statistics is proposed. The selection of principal components is performed by calculating the values of the semi - variogram parameter range and partial sill/sill of each principal component. The magnitude of a range is used to judge the range of spatial correlation of each principal component, and the partial sill/sill is used to judge the strength of spatial correlation of each principal component. The simulation proves that the variable range and partial sill/sill can effectively express the range and strength of spatial correlation of hyperspectral remote sensing images. Based on the experiment of real hyperspectral remote sensing images, the empirical threshold of principal component selection is determined from subjective and objective aspects, that is, the range is 2.5, and the partial sill/sill is 0.2. According to the classification results based on the support vector machine algorithm, compared with traditional methods, the principal components with better image quality can be screened by using variable range and partial sill/sill, which can not only achieve the purpose of dimensionality reduction, but also ensure high classification accuracy.

Keywords: hyperspectral; principal component analysis; spatial statistics; semi - variogram; support vector machine

(责任编辑: 李 瑜)