第9卷 第3期
1997年9月

湖 泊 科 学
JOURNAL OF LAKE SCIENCES

Vol. 9, No. 3
Sep., 1997

261－268

X832
P343.3

主成分监督分类及其在水质特征遥感图像识别中的应用*

余丰宁 蔡启铭

（中国科学院南京地理与湖泊研究所，南京 210008）

提要 建立了一种水域水质状况图像识别的主成分监督分类方法，首先通过TM水域图像数据的主成分分析，将原有各波段图像的显著且独立的信息集中在数目尽可能少的合成图像中；再依据不同类型水体的光谱特性，分析各主成分图像的构成及其环境生态学含义，由此对整个研究区域内存在的不同标志类型及其分布特征有所了解；在此基础上，选定训练样本集，从而根据具有清楚的环境生态意义的标志类型，应用监督法得到较好的识别分类结果。分析表明，这一方法采用主成分分析确定标志类型，无需大量的现场调查，因而具有非监督聚类成本低的优点，分类结果则优于非监督法，且各类型的生态意义明显，分布特征与环境因子相互吻合，是水域水质环境图像识别的有效而实用的方法。

关键词 主成分分析 遥感识别 水质特征 太湖

化学成分，水质环境 湖水

水体的悬浮质、营养盐、藻类浮游植物等物理、化学和生物要素场在中尺度内的特征和动态对于水体生态系统的研究至关重要。由于水的流动变化，水域环境生态调查周期很短，高频次、大面积的水域布点观测受到人力、物力和风浪等环境条件的约束，困难很大；而利用卫星遥感信息，通过统计分析、图像识别等手段可建立起与地面要素场及其所反映的生态环境特征的相互联系，具有高效率、高频次、高分辨率和低成本的优势。

通常的遥感图像识别聚类方法（监督法和非监督法）[1]在水质分析中的应用受到很大限制，因为监督法必须先有研究区域内已知类型的训练样本集，即对识别分类的各标志类型及其分布特征要事先了解，但水质状况及其分布变化的时间尺度很短，训练样本集的获取极为困难。而非监督法虽然依据同类物体具有相同（似）光谱特性，按各样本点在变量空间分布的相似性得到不同类型的点群，无需样本集的先验知识，但各点群所代表的地物特征还需经地面实况调查或光谱特征比较加以确定，而且一般说来，该方法的精度较低。用于水质分类时，不同点群的分布往往离散度很高，常常会出现在一个区域内各类型点状混合分布的结果，而水体的流动变化通常是渐变的，这使得实况与分类结果不能很好地对应，且分类的生态学意义不清晰。

本文建立了一种水域水质状况图像识别的主成分监督分类方法。通过水域图像数据的主成分分析，依据不同类型水体的光学特性，分析各主成分图像的构成及其生态学含义，由此对整个研究区域内存在的不同标志类型及其分布特征有所了解；在此基础上，选定训练样本集，

* 国家自然科学基金（39500027）和江苏省社会发展研究基金（BS95035）资助项目。
收稿日期：1996－01－16；收到修改稿日期：1996－04－03。
作者简介：余丰宁，女，1961年生，副研究员。1988年中国科学院大气物理所硕士毕业生，现主要从事物理湖泊学和数值模型研究。

从而根据具有清晰的生态意义的标志类型，应用监督法得到较好的识别分类结果.

1 不同类型水体的光学特征

水体对入射光有很好的吸收性，同时，水中所含杂质成分及其浓度或数量对水体光学特性有很大的影响. 这些杂质的光学特性与水体光学特性叠加，即组合成不同类型水体的反射光谱特征[2]：a）清洁水体：以天池中部泥沙含量小于 10mg/L、基本无污染的水体为例，其反射率很小，约 1.0% -2.0%，且基本不随波长变化. b）含泥沙水体：以黄河泥沙含量为 781.0mg/L、有机污染很小的水体为例，当波长在 0.55－0.85μm 之间，其反射较大、达 10%以上. 3）含藻类水体：以太湖有密集藻类的水面观测为例，约在 0.55μm 处有绿色反射峰，0.7－0.9μm 为强烈的近红外反射峰，在此双峰之间 0.65－0.7μm 处是吸收峰[3]. 可见反射率随波长变化非常明显、水体、泥沙和浮游植物光学特性之间的这些明显差异、是利用遥感光谱信息进行水质判别的依据所在.

2 水域遥感图像的主成分构成及其环境生态意义

本研究选用了美国陆地卫星 TM 数据产品为信息源[4]，波段设有 7 个通道、其中 TM2、TM3、TM4、TM7 等与水体中泥沙、藻类反射光谱特征有很好的对应性：TM2、TM4 分别对应藻类的绿反射峰和近红外强反射，TM3 对应泥沙的反射峰和藻类的红外波吸收峰，TM7 对应藻类在中红外区的吸收带. 因此，TM 具有适宜水体泥沙、藻类等水质状况调查的光谱波段设置，能够捕捉和反映一定的水质特征. 对于这样多维且量大的信息源，将 7 个通道图像数据所反映的不同测面和片断组合起来，才能有效并充分提取地物的特征信息. 若将 7 个波段的图像数据作为 7 个变量，通过主成分分析[5]，可消除原有图像（变量）间的相关性，将原有数据的显著且独立的信息集中在数目尽可能少的合成图像中. 本文用 1991 年 7 月 23 日、1992 年 7 月 25 日、1994 年 6 月 29 日三个时相太湖北部湖区（含梅梁湖、贡湖、竺山湖和五里湖）的 TM 遥感图像数据分别进行主成分变换、从而提取太湖北部夏季的水体环境特征.

2.1 波段选择

首先 TM6 为热红外信息、在水面上差异极小，故可舍去. 其次，从不同波段数据的相关分析中发现湖区 TM1、TM2 和 TM3 之间有很好的相关性，两两相关系数均在 0.99 以上、而且图像特征一致. 分析其原因，作者认为由于太湖水浅风浪大，水体中泥沙含量相对较高，而泥沙在可见光区的反射较显著，因此 TM1、TM2、和 TM3 都主要反映泥沙分布的信息、以至于浮游植物和藻类密集区在 TM2 波段应有的绿色反射峰也被掩盖了；另一方面，浮游植物绿色反衬峰很窄，而 TM2 波段通道相对较宽，也使得该波段对藻类的可分辨性下降. 故此认为 TM3 可以表达 TM1 和 TM2 的信息，在本文中再舍去 TM1 和 TM2. 最后选定以 TM3、TM4、TM5 和 TM7 为自变量进行主分量变换.

2.2 水体识别

由于水体强烈吸收红外辐射，红外波段的水体辐射明显地低于其它地物，而理想的识别水体的波长为 1.5－1.8μm 之间，因此用 TM5 波段（1.55－1.75μm）进行水体识别. 在 TM5 频率分布直方图上可发现在水体与非水体之间有一过渡区，为了排除有水生植物等覆盖的岸边浅水水体，将水体判别阈值定在过渡区靠近水体的临界点，再将图像中遥感值大于阈值的像元

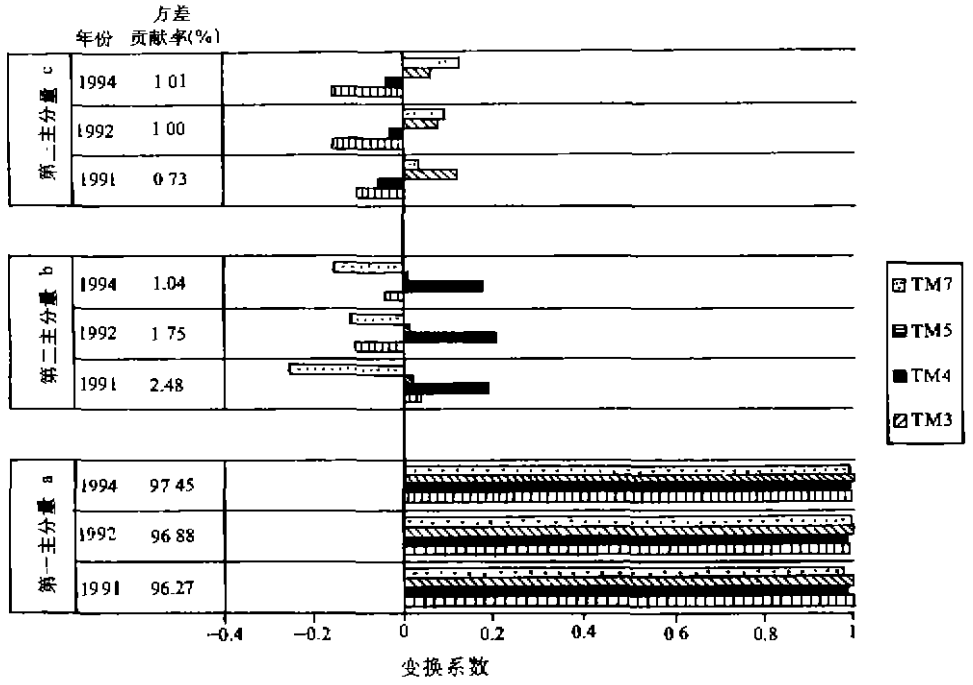

图 1　前三个主分量的变换系数及方差贡献率

Fig. 1　Conversion coefficients and SD contribution of the first three principal components

置零，由此将水域从图像中分离，即得到水域的遥感图像。

2.3　各主分量的构成及含义

对上述 4 个波段的水域图像进行主成分变换，得到的前三个主分量的变换系数（特征向量组成）见图 1，分析其构成可见：第一主分量的方差贡献率均在 95%以上，说明水体的主要差异来自这一主分量，它近似为原来 4 个波段遥感值的和，比照清洁水体和含泥沙水体的光学特征，不难理解这一主分量代表了水中悬浮质的情况，即水体的清洁程度。对于清洁水体，各波段的反射值均很小，第一主分量取低值，而混浊的水体中泥沙等悬浮质将产生散射，第一主分量取高值（图 2a）；第二主分量的方差贡献率为 1%－3%，其构成为 TM4 与 TM7（或 TM3）之差，参照绿色植物和藻类的反射光谱，可知这一主分量的值可表示水体中含藻类情况及叶绿素含量（图 2b）；第三主分量的方差贡献率为 1%左右，其构成为中红外（TM5 或 TM7）与可见光波段（TM3）的遥感值之差。观测发现绿色植物的红外反射峰在植物发黄腐败时，其位相向长波方向移动，由此初步推断，这一主分量可能表示了腐质植物在水中的分布情况（图 2c），但这一推测还有待进一步确认。

由此可见，从主成分变换特征图像中可提取水域悬浮质（或透明度）、藻类等分布和含量的特征信息。

3　主成分监督分类法

鉴于非监督法在水域分类判别中的效果不好，而监督法受到训练样本集选取困难的限制，

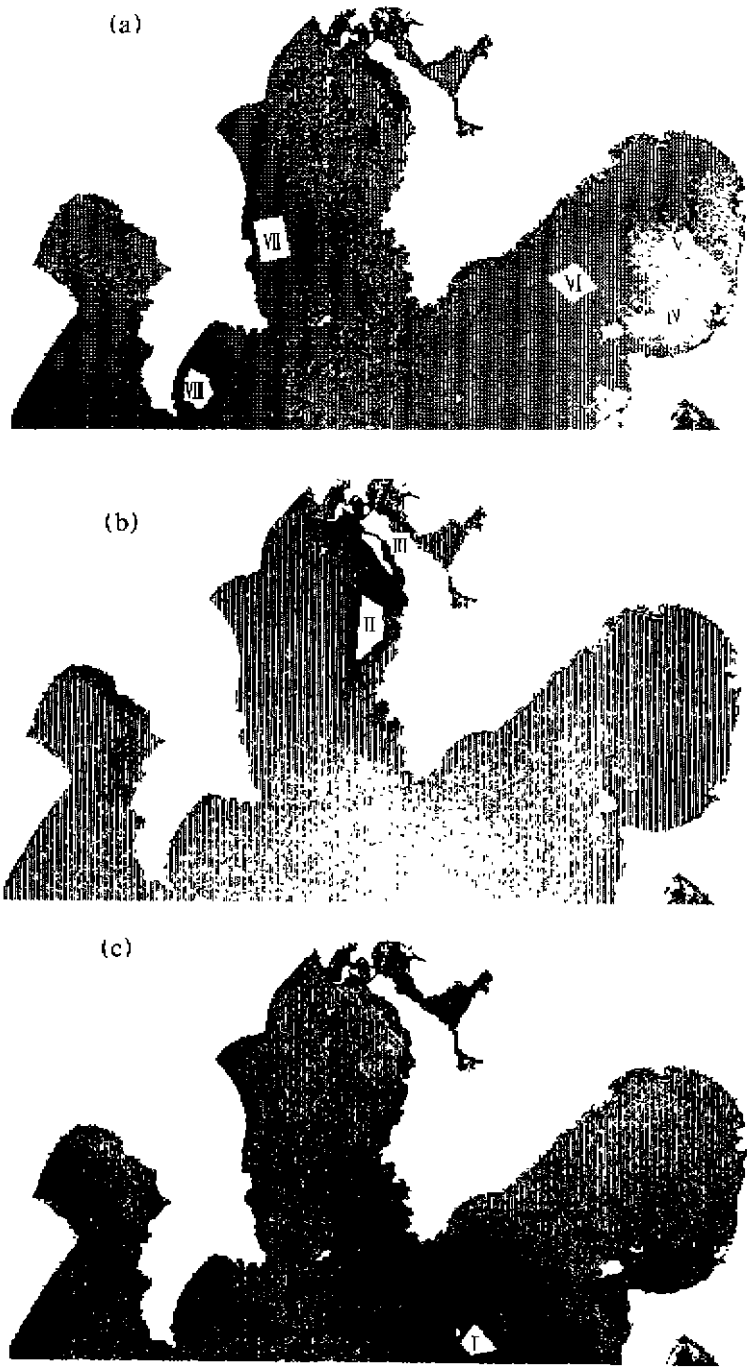

图 2 太湖北部遥感图像的主分量合成图和选定的标志类型区 I-VII(1992-07-25)
a. 第一主分量图； b. 第二主分量图； c. 第三主分量图

Fig. 2 The image of the principal components and selected sample areas in the northern part of Taihu Lake on July 25, 1992

很难在水质特征分类中应用，本文提出的主成分监督分类法，通过水域图像数据的主成分分析，解决监督法所需的训练样本集问题，从而根据具有清楚的生态意义的标志类型，应用监督分类法得到较好的分类效果。其工作程序如下：

(1) TM 遥感数据的主成分分析，用 TM3、TM4、TM5 和 TM7 四个波段的图像数据为自变量进行主成分变换。

(2) 观察分析各主分量及其合成图像的特征，判断前三个主分量的波段组成，并对照地物光谱特征，分析其代表的生态学意义。

(3) 用第三主分量确定一个标志类区。在第三主分量合成图像的最高值区域中选一个标志类区 1，在夏季，这一类区可能代表含较多腐质植物的水体情况(图 2c)。

(4) 用第二主分量确定两个标志类区。在第二主分量合成图像的次高和最高区中各选一个标志类区Ⅰ、Ⅱ，这两个类区在夏季通常分别代表含较多、很多藻类的水体类型(图 2b)。

(5) 由第一主分量确定五个标志类区。将第一主分量图像数据分成五个等级区，在各级区域中分别选一个标志类区Ⅳ－Ⅷ，这五个类区分别代表含不同程度的固体悬浮质(或水体透明度)类型(图 2a)。由此，在整个研究区域中根据主成分分析结果选定约 8 个标志类区作为训练样本集，每个选定的类区唯一代表一个标志类型，即训练样本集内没有公共区(图 3)，以整圆代表所研究的水域，被分割成的五个扇形区即表示根据第一主分量划分的五个等级区，△区、□区和○区域分别表示在第三、第二、第一主分量图像中选定的标志类型区。

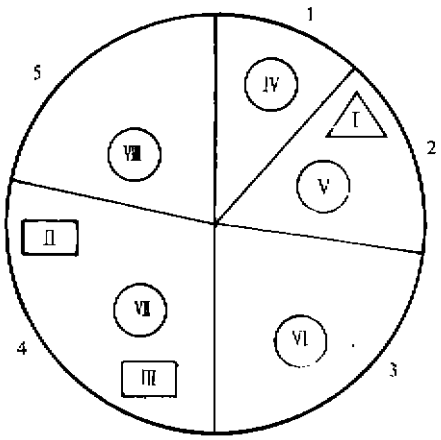

图 3　标志类区的选取

Fig. 3　Schematic diagram for selecting training sample based on PCA

(6) 针对上述训练集，计算确定每一标志类型的聚类"重心"，再依据统计上的最大相似或最小距离原则，将研究区域内的每个像元划分到相应的标志类区中。

(7) 对每个像元分类后，重新计算各类的"重心"。

(8) 考察每个像元，如果把该像元从其原来所属的类移动到另一类能够减少样本对于各自"重心"的总离差，就将其移到另一类。

(9) 重复(7)、(8)直至再没有像元点被移动到另一类中，即得最终的 8 个聚类，它们将整个研究水域分成不同的水质特征类型。

4　主成分监督法评价

此方法具有非监督法分类的优点，即不需要所研究水域的实况调查，而是通过主成分分析，直接从遥感图像中提取水质特征状况的信息，从而解决了确定标志类型样本集的困难，在此基础上应用监督法分类，故又具有聚类的生态含义清楚、水域划分有连续性和精度较高的优点。其分类效果可以从以下三个方面来考察。

4.1　统计指标

对于各种不同方法的聚类效果，可用不同类型的"重心"(变量均值)之间的距离和每一类型内部各点之间的平均距离比较不同类型的分离程度和同一类型中各点的聚集程度。因此我

们可以选定这样一个判断分类效果的统计指标：

$$K = \text{类间平均离差} / \text{类内平均离差和}$$

这一指标，在形式上与方差分析中所用的齐性检验相似，其含义是分类后不同类型之间的差异越大，而同一类型区域内各像元之间的差异越小，则分类效果越好，K 值就越大。据此，对 1991－07－23、1992－07－25、1994－06－29 三个时相的遥感数据分别用假彩色合成簇分聚类和主成分监督分类，然后计算其 K 值。假彩色合成得到的非监督簇分聚类的 K 值分别为 0.3756、1.2569 和 1.0332，主成分监督法分类的 K 值分别为 0.4184、1.5910 和 1.3724。可见后者的分类效果较好。

4.2 不同类型水体的生态学意义

主成分监督法的标志类型是依据主成分分析的结果确定的，而根据夏季水体遥感数据的各主分量组成和不同类型水体光谱特征的对比，可赋予主分量一定的生态意义。由此选定的标志类型，在聚类后仍具有明显的生态学意义。图 4 为采用这一方法分类得到的不同类型水体遥感光谱特征，其中清洁Ⅰ－Ⅲ三个类型的各波段反射均较弱，说明水体较清洁。“清洁Ⅰ”的 TM3 较低，表明泥沙很少。“清洁(Ⅱ)”含有少量悬浮质和一定量的腐质植物；“一般”的光谱与平均情况相似，水体主要含一些泥沙；“多藻”的光谱 TM4－TM3 较高，说明水体含有较多藻类。“水华”的特点很突出，TM4 很高且 TM4－TM3 也很高，可以判断为有藻类堆集形成的严重水华区；多泥沙Ⅰ、Ⅱ类型最突出的特点是 TM3 为高值，其它各波段的值也有相应提高，基本配置与“一般”和平均情况相似，这是含泥沙较高的水体。其中“多泥沙Ⅱ”更严重一些。由此可见，不同类型水体的光谱特征差异明显并对应有清楚的生态意义。

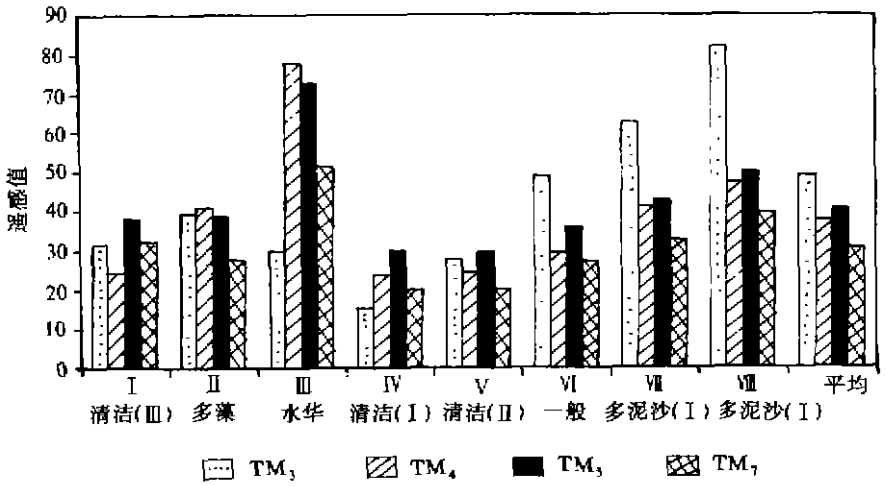

图 4 对 1992－07－25 太湖北部水域采用主成分监督法分类得到的不同类型水体的遥感光谱特征

Fig. 4 Spectral features of different water types divided with PC-supervised method in the northern part of Taihu Lake on July 25, 1992

4.3 与环境条件的配合

仍以 1992 年 7 月 25 日太湖北部为例。根据气象观测，该区域 7 月上旬出梅，25 日前一周内气温 28－36 C，为晴热高温天气，盛行东南风，但 24 日为西到西北风，25 日凌晨转为西南风。该水域遥感分类显示(图 5)：

(1) 水质从东南(贡湖东南部)向西北递次下降，在西南部即竺山湖南部及焦山至拖山一

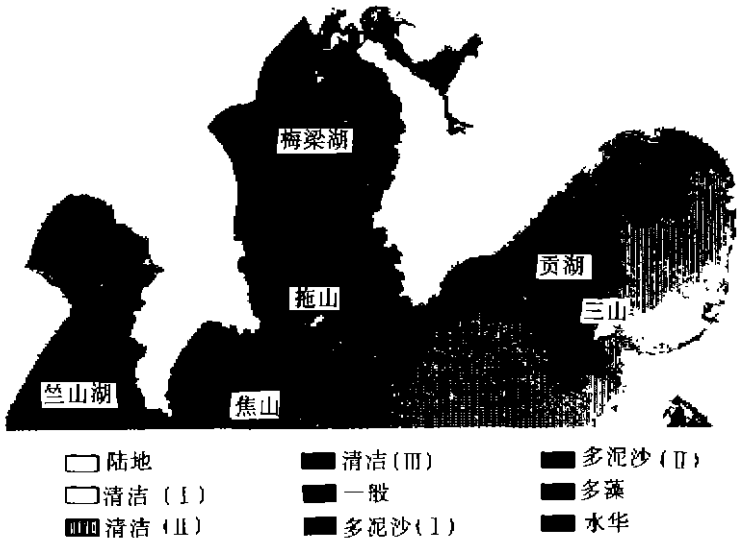

图5　1992－07－25太湖北部水域水质状况分类

Fig. 5　The classification image of water quality in the northern part of Taihu L. on July 25, 1992

带悬浮泥沙很严重，梅梁湖西部及贡湖西北沿岸也有明显的悬浮泥沙，这说明受盛行东南风影响，在下风湖岸区风浪大，扰动强，同时还有风生流输送的悬浮质，这些以泥沙为主的悬浮质受短期风向变化的影响较小，故主要分布在盛行东南风的下风湖岸区.

(2) 由于长时期的高温晴热天气，湖中藻类大量繁殖，并受盛行东南风影响滞留或漂移到位于太湖北部的梅梁湖和竺山湖，图中可见大片的藻类密集区，分布在梅梁湖东部和竺山湖北部的湾内，其中梅梁湖东北沿岸还有一带状水华覆盖区. 由于藻类漂浮于水面易随风漂移，故会受到短时风向变化的影响. 24－25日的偏西风，使藻类大多聚集在湖湾东部.

(3) 贡湖南部是夏季东南季风的迎风区，又由于湖底有水生植被[6]，水体清澈，悬浮泥沙很少，但有少量藻类和腐质植物.

可见，分类结果与环境要素吻合较好.

5　结语

(1) 一般说来，监督聚类的性能优于非监督聚类，然而其代价昂贵，对于水域而言，则无法预知标志类型的训练样本集，使监督法的应用受到更多的限制. 本文提出的主成分监督法，从解决监督法所需的训练样本集入手，采用主成分分析确定标志类型，突破了这一限制. 由于无需大量的现场调查，因而既具有非监督聚类成本低的优点，又得到了较好的分类效果. 是水域水环境图像识别的有效而实用的方法.

(2) 本文所分析的是太湖北部夏季的情况，对于不同的湖泊或不同的季节，各主分量的组成及其所代表的水质特征将有一定差别，这需要对比光谱特征进行分析和确定.

(3) 此方法和分类效果在一定程度上会受到人为因素的影响. 在对主分量合成图进行分析并选取标志类型时，要选择每一合成图像中最显著的特征区域，并且训练样本集中的像元组

成要尽可能"单纯",减少杂点.因此,关于在主分量合成图上选取训练样本集的原则和定量化标准,还需更进一步的研究和试验.

参 考 文 献

1 Cormack RM. A review of classification. *J R Statist Soc A*,1971,13(4)

2 中国科学院空间科学技术中心编.中国地球资源光谱信息资料汇编.北京:能源出版社 1987

3 李旭文等.太湖藻类的卫星遥感监测.湖泊科学,1995,7(1):65—68

4 阎守邕等编译.地球资源技术卫星.北京:科学出版社,1980

5 (英)M·肯德尔.多元分析.北京:科学出版社,1983

6 孙顺才等.太湖.北京:海洋出版社,1993

PRINCIPAL-COMPONENT-SUPERVISED CLASSIFICATION AND ITS APPLICATION TO IMAGE RECOGNITION OF WATER QUALITY

She Fengning Cai Qiming

(*Nanjing Institute of Geography & Limnology, Chinese Academy of Sciences, Nanjing 210008*)

Abstract

A new classification method of remote sensing image recognition, called Principal Component-Supervised Classification, is presented. Firstly, by means of principal component analysis, the component images are uncorrelated with each other and explain progressively less of the variance found in the original Landsat Thematic Mapper (TM) data in water area. After analyzing the composition of each component image and its eco-environmental implication according to spectrum features of different water types, the existing water types and their distribution features in the water area are known. Then, the training samples are selected based on the sample water types in PCA images and the classification image is produced following one of the decision rules and programs of supervised methods. This PC-Supervised method, selecting training samples based on the result image of PCA without large-area investigation on the ground or water surface, has thd advantages of unsupervised classification and a partition resolution higher than that of cluster analysis. Furthermore, its distinguishing result, applied to water quality recognition in the northern part of Taihu Lake, shows that the presented water types and their distributions are concordant with the conditions of lake body and environmental factors. So, it is indicated that PC-Supervised classification is an effective and practical method for dynamic analysis of water quality using remote sensing information.

Key Words Principal component analysis, image recognition, water quality, Taihu Lake