

# 水体富营养化驱动因子粗糙分析

吴义锋<sup>1</sup>, 吕锡武<sup>1</sup>, 薛联青<sup>2</sup>

(1. 东南大学环境工程系, 南京 210096; 2. 河海大学水资源环境学院, 南京 210098)

**摘要:** 基于水质污染因子及其污染特性存在不确定性的特点, 将粗糙集理论应用于水质污染因子及特性分析, 建立污染因子评价粗糙集数学模型。该模型仅依赖于数据本身提供的信息, 挖掘污染因子之间不确定的分类关系, 计算其对水体污染或水质富营养化的重要性即贡献率大小, 科学、快速和客观地揭示水体污染中起主导作用的污染来源, 为水体污染控制提供理论依据。应用数学模型分析巢湖 12 个监测点的水质数据, 确定巢湖主要污染因子及其导致湖泊富营养化的贡献率, 结果显示 TP、COD、TN 等是导致巢湖水体富营养化的重要因子, 今后应控制巢湖入湖水体的总磷及其有机污染物含量, 以改善巢湖水质富营养化状况。

**关键词:** 粗糙集; 不确定性信息; 污染因子; 模型; 水体富营养化

**中图分类号:** X824 **文献标识码:** A **文章编号:** 1671-1556(2005)04-0011-04

## Rough Analysis on Driving Factors of Eutrophic Water

WU Yi-feng<sup>1</sup>, LU Xi-wu<sup>1</sup>, XUE Lian-qing<sup>2</sup>

(1. Department of Environment, Southeast University, Nanjing 210096, China;

2. College of Water Resource and Environment, Hohai University, Nanjing 210098, China)

**Abstract:** Pollutant factors of water quality for rivers and lakes and their relationships are the key factors for water pollution control. Traditional theories based on ascertained maths are not appropriate to deal with them because of their unascertained information. The rough set theory as a new technically maths tool can analyse and mine unascertained information from water pollutant factors. And the maths model based on rough set theory is put forward in this paper to calculate the importance coefficients of pollutant factors and to discover hidden data from them. Then the maths model is used to analyse the water quality data of Chaohu Lake. The result shows that the pollutant factors such as TP and COD are the most important factors which describe correctly the contamination degree of Chaohu Lake. And it also points out that the future study of Chaohu Lake should be focused on reducing the concentration of TP, COD and TN so as to improve water quality of it.

**Key words:** rough set theory; uncertainty information; pollutant factors; model; water eutrophication

## 0 引言

作为一个开放的大系统, 水环境系统中的污染因子和污染物浓度等信息都存在着一定的不确定性。这些信息的不确定性有的是客观存在的, 有的是由资料不完全造成的, 也有的是因为人们对其内

在作用机理、变化规律认识不清引起的。忽视信息的不确定性, 就可能给问题的研究结果带来较大误差。就水质分析与评价而言, 不同污染因子对水体污染贡献率不同, 而且水质影响因素众多, 具有不精确、不确定和不完全性的特征, 因此在进行试验数据处理时, 必须运用不确定性信息处理从大量的、不完全的、有噪声的、模糊的和随机的数据中挖掘有用的

收稿日期: 2005-08-26

基金项目: 水文水资源及水利工程科学国家重点实验室对外开放科技创新基金项目(2003400519)资助。

作者简介: 吴义锋(1975—), 男, 博士研究生, 研究方向为富营养化河湖生态修复与生态水处理技术。

信息,简化和分析数据。水环境系统中信息的不确定性,可以借助粗糙集理论来解决。

粗糙集理论是一种刻画不完备、不确定性信息的数学工具,在机器学习、数据库知识发现、知识获取、决策支持系统、归纳推理等领域都有很大的利用空间。粗糙集理论主要是利用其下、上近似集对数据进行近似分类,不需要任何辅助信息,仅依赖于数据本身提供的信息,就能在保留关键信息的前提下,对数据进行化简,发现给定数据中隐含的信息,并求得知识的最小表达。目前,粗糙集理论已成功应用于机器学习、模式识别、逻辑控制、经济技术决策、专家系统中,而在环境科学与工程中的应用还不多见。

笔者基于粗糙集理论的思想,利用不确定性问题的近似分类方法,建立数学模型,分析巢湖各监测点的水质指标数据及其分类关系,确定巢湖水质富营养化主要污染因子及其贡献率。

## 1 粗糙集数学模型

### 1.1 基本思想<sup>[1-3]</sup>

设  $X \subseteq U, U = \phi, R$  是  $U$  上的等价关系,知识库可表示为  $K = (U, R)$ 。在  $K$  上,如果  $X$  是若干  $R$  的并集,则称  $X$  是  $R$  可定义的,否则  $X$  是不可定义的,不可定义集称为  $R$ -Rough 的。若存在一个等价关系  $R \in IND(U)$ ,其中  $IND(U)$  是  $U$  上给定的所有等价关系的交集,使得  $X \subseteq U$  是  $R$  一致的,则集合  $X$  称做  $U$  中一致集;若  $X \subseteq U$  对任意  $R \in IND(U)$  都是  $R$ -Rough 的,则  $X$  称为  $U$  上的粗糙集。

### 1.2 建立数学模型

粗糙集可用两个精确集来定义,即上近似集和下近似集。下近似集是对知识库  $R, U$  中所有一定能归入  $X$  的元素的集合,上近似集是对知识库  $R, U$  中可能归入  $X$  的元素的集合。上近似集是指一个集合不能利用有效的等价关系被恰当的分类,可通过另外的集合来表示这个集合的近似。设  $X \subseteq U$  是任意子集,  $R$  是  $U$  上的等价关系,有

$$R_*(X) = Y\{Y \in U/R : Y \subseteq X\}$$

$$R^*(X) = Y\{Y \in U/R : Y \cap X \neq \phi\}$$

分别称为  $X$  的  $R$  下近似集和  $X$  的  $R$  上近似集,  $Y$  是  $U$  上按等价关系  $R$  的等价类,上近似和下近似之差定义为  $X$  的等价关系  $R$  边界线集,为

$$BN_R(X) = R^*(X) - R_*(X)$$

它表示通过等价关系  $R$  既不能在  $X$  上被分类,也不能在  $X$  的补集上被分类的元素的集合。

通过已掌握的信息看论域  $U$  中的待研究子集

$X$ ,运用粗糙集原理只能观察到  $X$  的上近似集和下近似集,而不能观察到  $X$  的全部面貌。当且仅当  $BN_R(X) = \phi$ ,即  $R^*(X) = R_*(X)$  时,集合  $X$  是精确的,否则是粗糙的。

由于粗糙集存在边界线集,则有一些元素不能在全域的某个子集上被分类,也不能在其补集上被分类,而这些元素归于边界区域,因此,其大小就表示某子集  $X$  关于论域上的等价关系的近似精度,于是针对粗糙集中不确定性就引入了数值度量。

对于某水体水质监测数据,由粗糙集中不确定性的数值度量可定义水质参数的重要性程度,它可直接反映各水质指标对描述水体水质的重要性。方法如下:将水质监测数据视为论域  $U$ ,各水质指标视为等价关系  $R$ ,设  $X \subseteq U \cap X \neq \phi$ ,  $X$  的上、下近似集分别为  $R^*(X), R_*(X)$ ,且  $R_*(X) \subseteq R^*(X) \subseteq U$ ,则

$$\text{Card}(R_*(X)) \leq \text{Card}(R^*(X))$$

式中:Card 为集合中对象的基数。

所以,集合  $X$  在论域中的等价关系  $R$  对集合元素分类的有效分类程度即可定义水质因子对描述富营养化的重要性程度为  $\alpha_R(X)$ ,可用距离函数计算。

$$\begin{aligned} \alpha_R(X) &= 1 - d(R^*(X), R_*(X)) \\ &= 1 - \frac{\text{Card}(R_*(X) \setminus R^*(X)) + \text{Card}(R^*(X) \setminus R_*(X))}{\text{Card}(R_*(X) \cup R^*(X))} \\ &= \frac{\text{Card}(R_*(X) \cap R^*(X))}{\text{Card}(R_*(X) \cup R^*(X))} \\ &= \frac{\text{Card}(R_*(X))}{\text{Card}(R^*(X))} \end{aligned}$$

$\alpha_R(X)$  代表了等价关系  $R$  对集合元素分类的有效程度,可以衡量等价关系  $R$  的分类能力,数值越大,分类能力越强。对分析水体水质而言,获取多个水质指标的监测数据后,通过粗糙集理论进行数据约简,将无关或多余的信息丢掉,约简后的数据重新组合并产生新的决策规则,计算某水质子集的上、下近似集,运用上述模型计算某水质指标对水体划分的有效程度,继而得到水质污染及富营养化的贡献率,为水体水质污染控制提供决策依据。

## 2 巢湖水质污染因子分类应用

### 2.1 巢湖水质概况

巢湖位于安徽省中部,为我国五大淡水湖之一,水域面积 789 km<sup>2</sup>,具有防洪、灌溉、渔业等多种功能,是沿湖城镇生活和工业用水的主要水源。而 1980 年以后巢湖污染加剧,水质富营养化日益严重,根据巢湖水质富营养化状态评价标准,目前巢湖

已进入富营养—极富营养状态<sup>[4]</sup>。巢湖水质的严重恶化,降低了水体的使用功能,给沿湖乡镇的生产和生活带来了重大影响。笔者搜集了近期巢湖 12 个

监测点位关于 8 个水质指标的水质数据,计算和评价巢湖水质富营养化主要污染因子及其污染重要性程度。

表 1 巢湖湖区 12 个点位监测结果统计表(mg/L)

Table 1 Water quality statistics for 12 spots in Chaohu Lake(mg/L)

| 点位  | 湖区测点    | DO   | COD <sub>Mn</sub> | 非离子氨  | TP     | TN   | 透明度/cm | 叶绿素    | 水质类型 |
|-----|---------|------|-------------------|-------|--------|------|--------|--------|------|
| 010 | 南淝河入湖区  | 6.01 | 6.67              | 0.093 | 0.2028 | 3.00 | 39.4   | 24.36  | >V   |
| 011 | 十五里河入湖区 | 5.87 | 7.41              | 0.058 | 0.2051 | 2.58 | 33.5   | 26.54  | >V   |
| 012 | 巢湖塘西    | 6.19 | 5.93              | 0.026 | 0.1908 | 2.96 | 41.5   | 29.534 | IV   |
| 013 | 派河入湖区   | 6.12 | 6.73              | 0.065 | 0.1704 | 1.90 | 37.5   | 22.21  | IV   |
| 014 | 新河入湖区   | 6.47 | 5.85              | 0.026 | 0.1446 | 1.38 | 42.5   | 13.425 | IV   |
| 015 | 西半湖湖心   | 6.48 | 6.22              | 0.056 | 0.1819 | 1.99 | 42.5   | 17.487 | IV   |
| 016 | 巢湖坝口    | 6.94 | 4.28              | 0.011 | 0.2170 | 3.58 | 44.4   | 8.409  | >V   |
| 017 | 巢湖船厂    | 7.00 | 4.38              | 0.007 | 0.2210 | 3.63 | 49.2   | 6.533  | >V   |
| 018 | 中捍乡     | 7.08 | 5.24              | 0.010 | 0.2064 | 2.87 | 42.8   | 8.806  | >V   |
| 019 | 东半湖湖心   | 7.43 | 4.15              | 0.006 | 0.1752 | 3.45 | 44.8   | 8.287  | IV   |
| 020 | 忠庙      | 7.56 | 4.69              | 0.013 | 0.2067 | 3.88 | 41.7   | 11.674 | >V   |
| 021 | 兆河入湖区   | 7.37 | 4.10              | 0.008 | 0.1881 | 4.31 | 48.4   | 7.919  | IV   |

## 2.2 粗糙集数学处理

将 12 个监测点位水质监测数据集作为论域  $U$ , 7 个水质指标视作等价关系  $R$ , 将相同的或接近的归为一类, 建立一个粗糙集知识库  $K=(U, R)$ 。

根据各监测点水质状况, 将  $U$  论域划分  $D = \{D_1, D_2\}$ , 其中  $D_1 = \{010, 011, 016, 017, 018, 020\}$ ,  $D_2 = \{012, 013, 014, 015, 019, 020\}$ 。将 DO、COD<sub>Mn</sub>、非离子氨、TP、TN、透明度、叶绿素 7 个水质指标视作 7 个参数集, 即等价关系  $R =$

$\{R_1, R_2, \dots, R_7\}$ , 首先将监测数据离散化, 方法见参考文献[2], 将相同的或接近的归为一类, 对  $U$  进行划分, 首先计算 7 个水质指标对每一类水质划分的重要程度即贡献率  $a_{R_i}(D_j)$  ( $i=1, 2, \dots, 7; j=1, 2$ )。再将各水质指标对每一类水质的重要性程度  $a_{R_i}(D_j)$  按水质子集元素基数占总元素基数进行加权平均, 算得各水质指标对巢湖水质分类的综合重要性即水质富营养化水质数据知识库划分重要性程度  $\sum a_{R_i}(D)$ 。计算结果见表 2。

表 2 污染因子重要性程度及贡献率计算

Table 2 Calculation of the importance coefficients (rate of contribution) of pollutant factors

| 水质指标 $R_i$        | $R_i$ 的划分  | $a_{R_i}(D_1)$ | $a_{R_i}(D_2)$ | $\sum a_{R_i}(D)/\%$ |
|-------------------|--|----------------|----------------|----------------------|
| DO                | $\{\{010, 011, 012, 013\}, \{014, 015\}, \{016, 017, 018\}, \{019, 020, 021\}\}$ | 0.418          | 0.417          | 14.5                 |
| COD <sub>Mn</sub> | $\{\{010, 011, 013\}, \{012, 014, 015\}, \{016, 017, 019, 021\}, 018, 020\}$     | 0.417          | 0.667          | 23.2                 |
| 非离子氨              | $\{010, \{011, 013, 015\}, \{012, 014\}, \{016, 017, 018, 019, 020, 021\}\}$     | 0.250          | 0.250          | 8.7                  |
| TP                | $\{\{010, 011, 018, 020\}, 012, \{013, 015, 019, 021\}, 014, \{016, 017\}\}$     | 1.000          | 1.000          | 34.8                 |
| TN                | $\{\{010, 011, 012, 018\}, \{013, 015\}, 014, \{016, 017, 019, 020\}, 021\}$     | 0.333          | 0.250          | 10.1                 |
| 透明度               | $\{\{010, 013\}, 011, \{012, 014, 015, 016, 018, 019, 020\}, \{017, 021\}\}$     | 0.083          | 0.083          | 2.9                  |
| 叶绿素               | $\{\{010, 011, 012, 013\}, \{014, 020\}, 015, \{016, 018, 019, 021\}, 017\}$     | 0.167          | 0.167          | 5.8                  |

由表 2 可看出, 在反映巢湖水质的 7 个水质指标中, 总磷污染因子在巢湖水体富营养化中的贡献率最大, 其中污染因子总磷在水质富营养化的重要性程度最大, 为 34.8%; COD<sub>Mn</sub>、溶解氧、总氮、非离子氨、叶绿素、透明度在体现巢湖水质富营养化中的重要程度依次为 23.2%、14.5%、10.1%、8.7%、5.8%、2.9%。为了更直观地表现出各个污染因子对巢湖污染水质富营养化的贡献率, 将表 2  $\sum a_{R_i}(D)$

中的数据绘成图, 见图 1。

从图 1 可以看出, 导致巢湖水质富营养化的主要因子为总磷、有机污染物、溶解氧以及氨氮, 其中总磷和有机污染物在导致巢湖水质富营养化中重要性程度(即贡献率)占一半以上, 达 58.0%, 说明这两个主要的污染因子是导致湖泊富营养化的根源。因而, 水体中藻类等水生植物迅速生长, 首先引起水中的溶解氧含量下降, 继而导致水体中叶绿素含量

上升、透明度降低。相关文献研究也显示出入湖河流是巢湖营养物质的主要输入途径,资料表明巢湖水体中总磷浓度达 0.25 mg/L 以上,其中通过河道入湖的总磷占输入总磷的 95.23%<sup>[5,6]</sup>,这与本文应用粗糙集理论研究结果一致。因此控制巢湖水质富营养化及促进湖泊生态功能的恢复,应着眼于营养性污染物的控制,巢湖湖区必须在点源和面源上同时控制总磷、有机污染物的排放量。

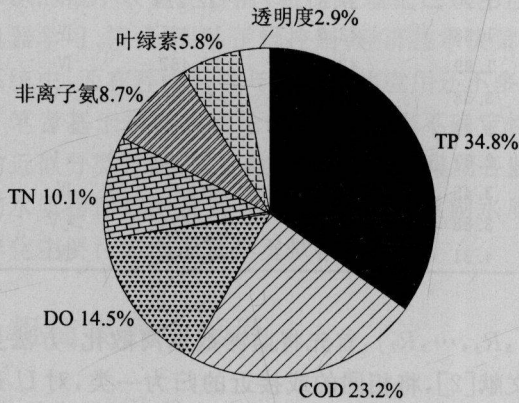


图1 各污染因子对巢湖水质富营养化的贡献率

Fig. 2 The rate of contribution of pollutant factors to eutrophic water in Chaohu Lake

### 3 结论

(1) 水体富营养化驱动因子数学模型优点显著,

适用性强,不需要任何辅助信息,仅依赖于数据本身提供的数据信息,能比较客观地反映水质污染程度,并确定各水质污染因子对水质污染的贡献率大小,在水污染控制及水环境管理领域具有广阔的运用前景。

(2) 实例研究表明,基于粗糙集理论的思想,利用不确定性问题的近似分类方法,通过分析巢湖各监测点的水质指标数据及其分类关系,确定巢湖主要污染因子及其导致湖泊富营养化的贡献率,理论上是可行的,计算结果是可靠的。

(3) 研究结果确定了巢湖水质富营养化的主要污染因子,指出今后巢湖水质富营养化的主要污染物控制方向,即主要控制总磷、有机物的污染浓度。

#### 参考文献:

- [1] Chang L. Y., G. Y. Wang, Y. Wu. An approach for attribute reduction and rule generation based on rough set theory[J]. *Chinese J. Software*, 1999, 10, 1: 206-1 211.
- [2] 刘清. Rough集及 Rough推理[M]. 北京:科学出版社, 2001. 11-38.
- [3] Lin T. Y. Granular fuzzy sets: A view from rough set and probability theories[J]. *International Journal of Fuzzy Systems*, 2001, 3(2): 373-381.
- [4] 郭培章, 宋群. 中外水体富营养化治理案例研究[M]. 北京:中国计划出版社, 2003.
- [5] 李如忠, 汪家权, 钱家忠. 巢湖流域非点源营养物控制对策研究[J]. *水土保持学报*, 2004, 18(1): 119-122.
- [6] 张之源, 王培华, 张崇岱. 巢湖营养化状况评价及水质恢复探讨[J]. *环境科学研究*, 1999, 12(5): 45-48.