第十六章　支持水务和污水处理系统脱碳的数据科学工具

Kathryn B. Newhart1*, Amanda S. Hering[2] and Tzahi Y. Cath[3]

[1]United States Military Academy, West Point, NY, USA
[2]Baylor University, Waco, TX, USA
[3]Colorado School of Mines, Golden, CO, USA
*Correspondence: kathryn.newhart@westpoint.edu

16.1 引言

数据科学工具可以利用历史和当前生成的数据来报告和影响水务和污水的分配和处理系统的监控和控制方式。尽管在工程环境系统的数据驱动建模方面有几十年的进展，但水务和污水处理设备仍在使用基本的监测、分析方法和控制模式。传统上，水务和污水处理模型源自对污染物去除现象的基本理解（例如重力分离和沉降、化学和微生物动力学）。由于全规模处理设备的规模和复杂性，少有模型能充分准确地进行过程监测和控制。相反，控制阈值（如单个过程变量的上限和下限）是根据历史表现和操作人员对特定系统的理解来确定正常的操作条件。这些数值是静态的，且包括一个很大的安全系数，以考虑所有可能的水质、环境和操作条件；最终大幅降低了系统的效率。替代这些静态、物理、数学模型的是经验、数据驱动模型。这些“智能”模型依赖于数据集中确定的变量之间的关系，而无需根据预先储存的知识确定这种关系。近年来，随着数据收集和存储的费用下降及数据处理速度呈指数增长，数据驱动建模（DDM）得到了发展。然而，水务和污水处理行业尚未完全实现这些技术进步。Manesis 等人（1998）认为限制污水处理行业采用 DDM 的原因如下：（1）智能控制领域不发达；（2）工程师对 DDM 不熟悉。尽管科学文献中人们对 DDM 越来越感兴趣，但由于上述的第二个原因，DDM 在水务和污水处理系统中的全规模应用仍然受限。本章的目的是让水处理工程师熟悉 DDM 方法，以实现水务和污水处理行业的脱碳目标。

DDM 包括统计方法和机器学习（ML）方法，虽然二者看起来很相似，但由于目的和要求的不同，不存在一种方法比另一种方法普遍“更好”。统计

模型本质上是概率模型，这意味着模型会自动测量不确定性。因此，当统计模型用于分析、总结和从数据中提炼结论时，会包含一个取决于数据噪声的误差范围。为了使这些统计模型有效，需要对数据中噪声的分布形状或变量之间函数关系的形式（例如，线性、指数、多项式）进行假设。另一方面，ML 模型非常灵活，可以对变量之间的非线性和复杂关系建模。它们不需要任何抽样分布或变量之间的关系形式的假设。然而，不确定性量化不能像 ML 模型那样容易得出，不仅涉及到大量内部参数的调整，而且往往需要非常大的样本量来拟合计算。这两种方法都可以用来实现相同的目标，并且不受特定的流程或系统的影响，但它们在哲学上有所不同，统计模型采用随机方法，而 ML 模型采用算法方法。Boulesteix 和 Schmid（2014）对统计模型和 ML 模型之间的区别进行了更深入的讨论。有些人可能会争辩说，与统计模型相比，ML 模型是“黑箱”；事实上，相对于基于物理学的模型，这两种类型的模型都是黑箱。模型中每个变量的重要性由统计模型自动提供，而在 ML 模型中需要多一些步骤才能得到，但是这两种模型都需要一些解释来理解每个变量对诱发因素响应的影响(Ljung，2010)。

DDM 脱碳的目标是最大限度地减少能源消耗和低效率，最大限度地回收资源和能源，最终减少直接和间接温室气体（GHG）排放。直接温室气体的排放源包括有机物的氧化、生物脱氮过程的副产品以及厌氧消化（AD）和燃烧产生的沼气。间接温室气体排放包括与电能消耗、反硝化作用的外部碳以及污泥处理和回收相关的排放(Flores-Alsina et al., 2011)。当前仿真研究已经建立了不同控制方案对水务和污水公用事业温室气体足迹影响的相关理论(Barbu et al., 2017)，但都是基于对运营成本和污水水质指标的定性假设。这是因为仿真研究只能近似但并不总是准确地表示全规模水处理厂 （WTP）和污水处理厂（WWTP）的变量之间的真实多元关系。例如，Oppong 等人（2013）比较了全规模 AD 和最流行的仿真模型（第二代基准仿真模型/BSM2）(Jeppsson et al., 2007)的输入和输出之间的 Pearson 相关系数。他们发现所有变量对在幅度和方向上存在很大差异。此外，全规模模型输出与 BSM2 输出不匹配，这种差异可能是由许多因素造成的，包括其他变量的影响和过程干扰，例如进水成分的变化、某些变量的抽样不频繁，以及全规模操作范围可能与仿真的范围不同。最终，该案例的仿真模型过于简单，无法捕捉到全规

模 AD 过程的真实行为。Dellana 和 West（2009）比较了统计模型和 ML 模型基于模拟的和真实的 WWTP 数据的预测性能，在哪个模型能“最优”预测出水氮和磷方面显示了相互矛盾的结果。。虽然统计模型对某些模拟仿真案例的预测误差较低，但 ML 模型对所有使用真实 WWTP 数据案例的预测误差最低。最终，以脱碳为目的的全规模 DDM 必须明确面向已知会影响能源消耗和温室气体排放的特征；然而，对于个别设备，实际影响可能难以推断。

本章介绍的工作旨在向读者介绍可在更大的脱碳战略中使用的 DDM 方法以及适当应用此方法的注意事项。本章分为五个部分，在 16.2 节中，介绍了数据准备、常见的 DDM 方法和用于比较模型性能的指标。在第 16.3 节中讨论了 WTP 和 WWTP 的单元工艺，其中第 16.2 节的方法可用于最大限度地脱碳。在第 16.4 节中，提出了全规模实施 DDM 的建议，第 16.5 节为结束语。

16.2 数据科学工具原理

DDM 主要分为统计学习和机器学习（人工智能的一个子集）；但对于脱碳系统来说，没有一种方法可以表现得“最优”。应根据应用（即系统）、消费者（即设备操作员、工程师、数据科学顾问）、可用数据的质量和数量以及分析的目标（即预测、预报、优化）等关键点选择采用统计、ML 还是混合统计-ML 方法最合适。由于统计模型和 ML 模型能够分别有效地捕获低维和高维关系，因此混合模型在预测和预报模型中有独特优势。用于预测的混合模型配置的例子有使用统计模型作为 ML 模型的输入(Newhart et al., 2020)，还有预测变量的统计模型与统计模型残差的 ML 模型的线性组合(Zheng &Zhong , 2011)。在本节中，我们将讨论开发智能水系统的重要组成部分：数据准备、特征选择的降维、重要过程控制变量的预测、机器学习模型的优化以及确定“最优”模型或方法的指标。开发 DDM 的过程如图 16.1 所示。

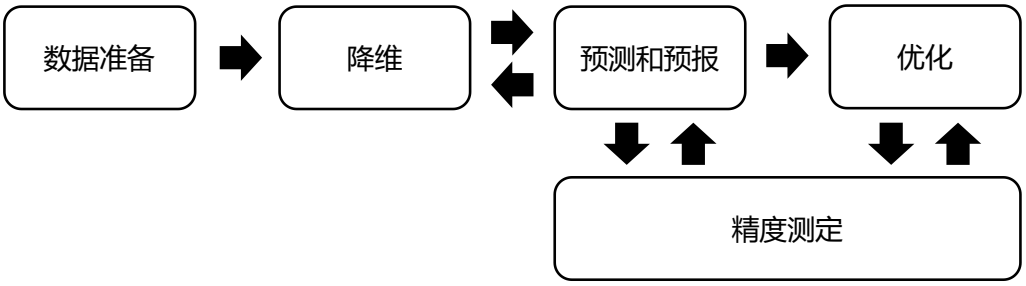

图16.1 DDM工作流程

16.2.1 数据准备

在收集用于 DDM 的数据时，需要考虑多种因素以确保数据能代表所模拟的过程：

（1）水质采样：由于成本、耗时或复杂的分析方法，或缺乏先进的传感技术，在线传感器或分析仪不适用于监测所有的水质变量。然而，监管报告（例如大肠杆菌）或性能评估（例如挥发性脂肪酸）需要其中的许多变量。抽样方法将确定分配给样本的时间戳，以及如何汇总在线数据令其能最好地代表采集样本数据时的环境和操作条件。用于实验室分析的典型水或固体基质样品的采集方式包括：

（a）抓取：分析结果仅代表采集样品时的条件。这些通常用于在环境条件下储存样品时随时间变化的水质变量（例如生物衰变）。

（b）时间复合：分析结果代表时间加权算术平均值，与流量无关。自动取样器以所需的时间频率抽取一定体积的样品。聚合样本被认为代表了一段时间内的情况。

（c）流量复合：分析结果代表事件加权算术平均值，它取决于流量。自动取样器抽取与水流量成比例的样品体积，高流量相对于低流量时采集的样本量更大。聚合样本是实际污染物负荷的最佳代表，因为它同时考虑了流量和时间。

（d）空间复合：当水质存在空间差异时（例如混合不均匀的矩形水箱），可以从不同位置采集并组合样本，聚合样本代表整个空间的平均状况。

（2）频率：WTP 和 WWTP 以不同的时间间隔收集数据，具体取决于测量设备（例如传感器、分析仪）和维护设备的人员的可用性，实验室设备和人员，以及特定变量的监管要求。当要汇总所有水质和运行变量时，必须考

虑范围广泛的插值间隔（例如秒、10 分钟、每天、每周 2-3 天）。在 DDM 中，具有不同频率的变量有两个主要影响：

（a）应该确定一个对于应用来说足够细粒度的间隔（例如，每天），但仍需足够大以避免不适当地计入那些不经常被收集的变量。聚合或插值方法必须是真实环境和操作条件的适当近似值，这些变量的频率分别比上面已确定的间隔更高或更低。聚合最常通过使用算术、时间加权或流量加权平均值来执行。插值可以通过线性方式进行，也可以将最后的测量值前移；但是应谨慎进行插值。例如，在实践中经常使用观测值之间的线性插值来“填充”缺失数据，但不一定能准确表示大多数水质变量随时间变化的情况(Newhart et al., 2021)。

（b）实时应用程序必须考虑瞬时数据的准确性和传感器的物理位置。许多在线传感器至少需要 5-20 分钟才能稳定下来并进行可靠的测量。因此，应根据关键预测变量实现平稳移动所需的平均时间来选择频率。此外，不同传感器测量之间的时间与流速（即水力停留时间）呈非线性关系，并且观测结果可能需要滞后才能准确反映对给定水质进水的处理性能。

（3）标准化：尽管某些 ML 模型有例外，但大多数 DDM 通常需要对数据进行标准化以使单个变量的变化与其量级无关(Maleki et al., 2018)。例如，对于某种成分，1 mg/L 可能属于浓度的较大变化，但对于第二种成分，这可能被认为是很小的变化。传统上，DDM 的数据使用以下任一方法进行标准化：（i）均值-中心（从每个值中减去均值并除以标准差）；或（ii）最大-最小法（从每个值中减去最小值并除以最大值减去最小值）。

（4）自相关：样本变量的许多水质测量值与先前测量值高度相关，也称为自相关。自相关和偏自相关函数图可以帮助确定在预测模型中应被视为预测变量的先前观测值（即滞后观测值）的数量(Maleki et al., 2018；Perendeci et al., 2009；Wu & Lo，2010)。

16.2.2 精度测量

为了衡量对于给定实际应用的预测方法的准确性，需要考虑多种指标。检验 DDM 误差的方法基本上有两种类型：训练和测试。训练数据用于拟合

模型，而测试数据不用于拟合模型，而是反映模型在实时或未知条件下的性能表现。文献中没有关于使用哪些特定指标的标准，但经常使用训练和测试误差的度量值来评估模型拟合和性能。因此，了解精度指标在不同应用中的优点或局限性非常重要。

可决系数（R^2）是环境工程中最著名的模型精度衡量指标。R^2 最常见的应用是训练值（y_i 或 $\mathbf{y}=(y_1,...,y_n)$）和模型预测值（$\hat{y}_i$）之间的比较，计算公式包含平方和(SST)和误差平方和（SSE）：

$$R^2 = 1 - \frac{SSE}{SST} \tag{16.1}$$

其中 $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$，$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$。当模型值与实际值完全匹配时，$R^2$=1。$\hat{y}_i$ 与 y_i 的线性回归图有助于了解多个模型的误差和 R^2 差异的分布。这些图可以回答诊断性问题，例如是否存在异常值、被高估或低估的特定观测组、或具有更大变化的 y_i 值范围。R^2 确实存在一些限制，例如对异常值敏感，不能很好地衡量差异的大小，且不适用于需要估计更多参数的更复杂的模型。因此，在比较两个模型的 R^2 值或使用不同的指标进行评估之前，了解模型及其预测值之间的潜在差异非常重要。

与测试数据相比，期望模型在训练数据上的 R^2 更高。但是，训练和测试 R^2 值的巨大差异可能表明模型对数据过度拟合。当模型过拟合时，模型中的参数多于捕获整体模式所需的参数。图 16.2c 是一个过拟合的例子，其中模型的参数太多，导致数据拟合错误。模型也可能欠拟合（图 16.2a），其中模型参数的数量不足以充分捕捉因变量的变化。鉴于过拟合模型的 R^2 将高于平衡模型，R^2 应始终与其他误差度量相辅相成。例如，模型拟合标准如 Akaike 信息标准（AIC, 方程（16.2））(Akaike，1974)和贝叶斯信息准则（BIC, 方程（16.3））(Schwarz, 1978)是平衡模型误差与模型中参数数量的指标，公式如下所示：

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2 \cdot p \tag{16.2}$$

$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \ln(n) \cdot p \tag{16.3}$$

其中 n 是观测数，p 是模型参数的数量，AIC 和 BIC 的区别在于对参数数量的适用性。在比较不同的模型时（例如给定模型类型的参数数量），具有最低误差和最少参数的模型将具有最低的 AIC 或 BIC。对同一模型进行比较时，AIC 会比 BIC 选择输入更多的模型。在这种情况下，建议选择 AIC 和 BIC 普遍偏爱的模型(Burnham & Anderson, 2004；Kuha, 2004；Vrieze, 2012)。例如，如果 AIC 在五个参数下达到最小，而 BIC 在三个参数下达到最小，那么具有四个参数的模型可能是最好的。

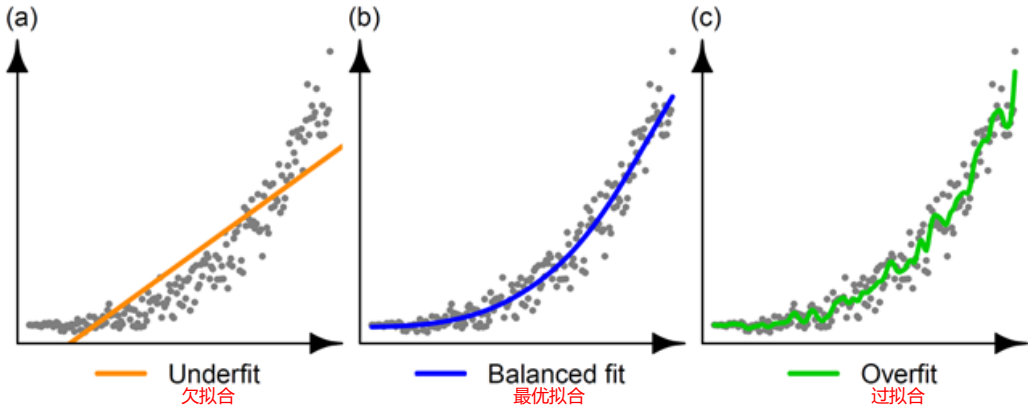

图16.2 欠拟合模型（橙色）、稳健/平衡模型（蓝色）、过拟合模型的示例

为了更好地理解模型误差的大小，可以使用未标准化（R^2）或惩罚性（AIC、BIC）的指标来衡量实际观测值和预测值之间的差异（或平方差异）。平均绝对误差（MAE，方程（16.4））、均方误差（MSE，方程（16.5））和均方根误差（RMSE，方程（16.6））是此类指标的代表：

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{16.4}$$

$$MSE = \frac{SSE}{n} \tag{16.5}$$

$$RMSE = \sqrt{\frac{SSE}{n}} \tag{16.6}$$

选择的单个指标取决于对大误差的期望敏感性。例如，MAE 取决于绝对误差，而不是 MSE 和 RMSE 的平方误差，因此它受实际值和模型值之间巨大差异的影响较小。另一个考虑因素是该指标是应用于训练数据还是测试数据。当 MAE、MSE 或 RMSE 用于单个出版物中的训练和测试数据时，一些作者将使用 MAPE、MSPE 和 RMSPE 来表征预测或测试指标。但是，也经

常将 AIC、BIC 和 R^2 用作训练指标，将 MAE、RMSE 和 RMSE 作为测试指标。

16.2.3 降维

在许多现实世界的 DDM 场景中，输入和输出变量之间的关系没有得到很好的理解或定义；因此，不相关的变量经常会无意中包含在 DDM 中。选择实现模型目标所必需且充分的变量子集称为特征选择(Kira&Rendell，1992)。彼此高度相关或只是噪声的输入变量会降低预测模型的有效性。首先，冗余信息将增加时间和计算要求，而不会显著提高预测准确性。其次，许多统计模型在存在多重共线性的情况下会变得数值不稳定，其中多个变量提供重叠信息。第三，模型的可解释性因额外的非必要输入变量而降低。最后，在检测故障时，在监测期间未发生实质性变化的噪声变量使得检测故障变得更加困难(Harrou et al., 2021)。

这里描述了几种通过特征选择来处理降维问题的方法。可以在构建模型之前使用统计降维方法，例如相关系数和主成分分析（PCA）。在建模步骤中经常使用逐步变量选择和套索建模方法来减少模型中的变量数量。这里还描述了一种用于 ML 模型中可比较的逐步选择变量方法。

16.2.3.1 相关统计

相关系数取值介于-1 和+1 之间。符号表示两个变量 X 和 Y 之间关系的方向，大小表示关系的强度。绝对值 1 表示两个变量完全相关，零表示它们完全不相关。此处仅介绍多种统计相关性指标中的 Pearson 指标和 Spearman 指标以说明基于幅度和基于秩的相关系数之间的差异，Helsel 和 Hirsch（2002）对水相关数据的相关系数进行了进一步讨论。

Pearson 相关系数（r，方程（16.7））是最流行的，它衡量一组 n 个独立观测值的线性关系的强度和方向$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$。定义如下：

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{16.7}$$

其中$\bar{x}$和$\bar{y}$分别是 x_i 和 y_i 值的样本均值。非线性相关的变量可能仍然相关，但 r 相对较低。Spearman 的秩相关系数（r_s，方程（16.8））是 r 的非参数变

体，能够衡量两个变量之间单调关系的强度和方向。例如，随着 X 的增加，Y 对所有 X 和 Y 都会增加，但不一定是线性的。Spearman 系数通过比较每对观测值的秩（即观测值从最小到最大排列时的位置）而不是值本身来实现这一点，如下所示：

$$r_s = 1 - \frac{6\sum (X_{irank} - y_{irank})^2}{n(n^2 - 1)} \tag{16.8}$$

其中$x_{i_{rank}}$是 x_i 的秩，$y_{i_{rank}}$同理。根据预期相关性的方向和幅度以及样本大小，应选择不同的相关系数。图 16.3 说明了 Spearman 对小样本量和非线性行为的敏感性不如 Pearson。但两者都无法量化在方向上既增加又减少的关系。需要注意的是，不同类型的相关系数值不能直接比较。例如，Pearson 的 r 值只能与其他 Pearson 的 r 值进行比较，以确定一对特征是否比另一对特征具有更高的线性相关性。

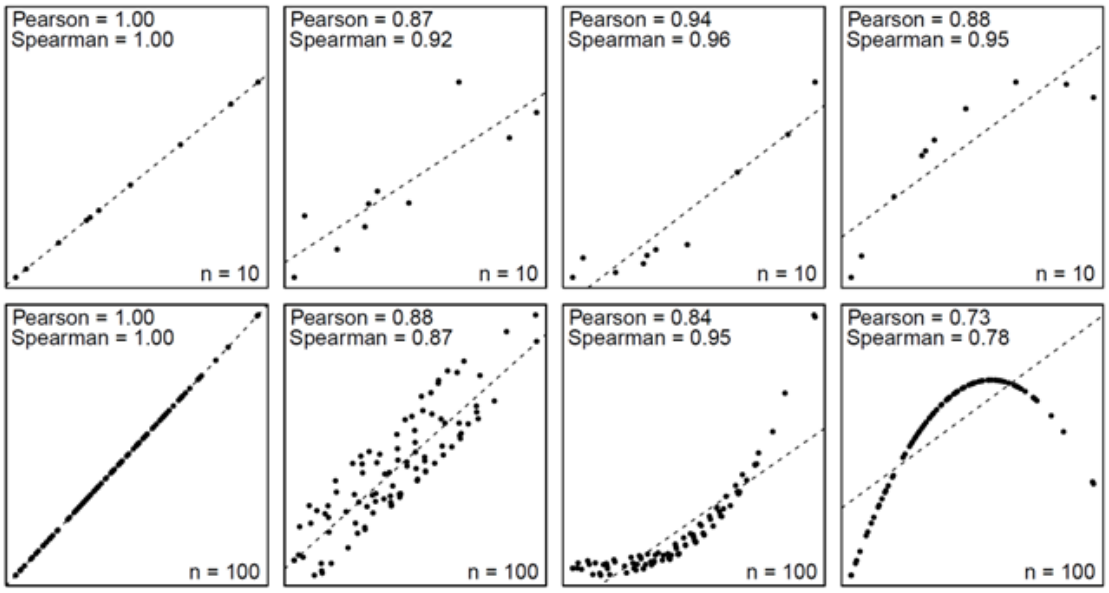

图16.3 不同样本大小的Pearson和Spearman示例，n =10（顶部）和n =100（底部），以及与拟合线性回归（虚线）相比的有噪声和无噪声数据（黑点）

16.2.4 主成分分析

PCA 是一种降维方法，它创建现有变量的线性组合以依次捕获数据中的最大变化(Jackson，1991；Wise 和 Gallagher，1996)。每个线性组合都是一个主成分（PC）并且与其他成分正交；因此，每个成分代表不同的变化来源或

变化方向，并且与其他成分线性无关。图 16.4 说明了如何使用 PCA 将三变量系统简化为两个独立的 PC，因为 PC1 和 PC2 所捕获的方差之和为 100%。

在一个假设的水处理系统中，y_1 可作为理想的藻类生长速率，y_2 为太阳辐射，y_3 为实际生长速率。尽管 y_1 和 y_2 是非线性函数，图 16.4b 显示了在第一个成分中它们的线性组合（y_3）如何被捕获。三变量系统中 PC1 未解释的其余变化被 PC2 捕获。

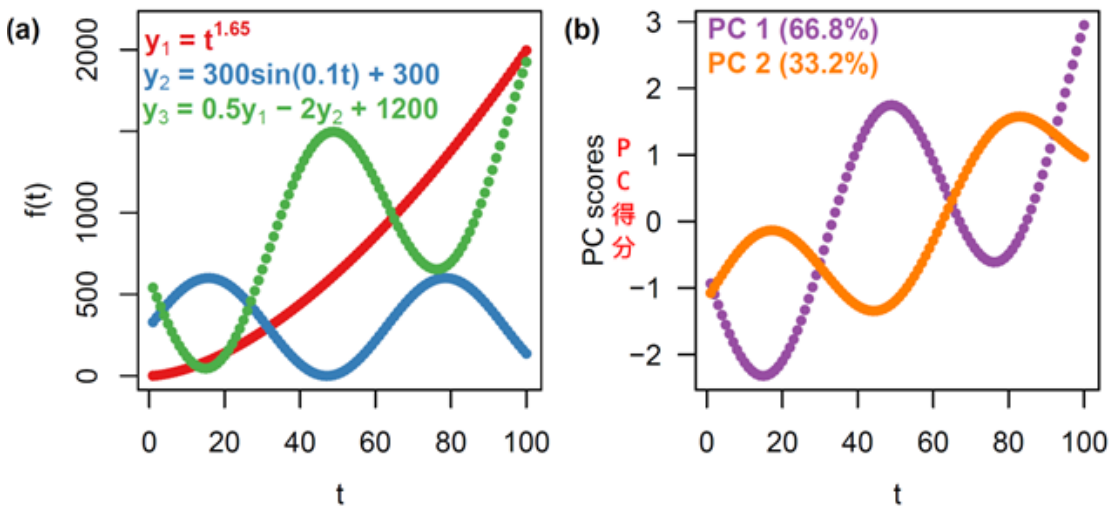

图 16.4（a）将两个非线性函数（y_1和y_2）和非线性函数的线性组合（y_3）绘制为t的函数；（b）PC1得分和PC2得分分别是（a）中的缩放观测值乘以PC1和PC2的变量载荷（即旋转矩阵）。括号内是每个PC捕获的总变化百分比

PCA 最流行的应用是多元回归模型的降维（第 16.2.4.1 节）(Wallace et al., 2016；Wang et al., 2017)和一般过程洞察(Corominas et al., 2018)。对于建模来说，具有最大载荷（即特定方向上的方差大小）的 PC 被保留用于模型构建。为模型构建保留的特定成分数量取决于 PCA 子空间描述的变化百分比，通常使用 90-99%范围内的值作为确定保留成分数量的阈值。

传统 PCA 以及用于水务和污水应用的大多数标准统计方法的一些限制因素是平稳性（恒定均值和方差）、线性和随时间独立性的假设。诸如滚动训练窗口、非线性降维方法和滞后观测等修正有助于接近 PCA 等方法所需的条件(Kazor et al., 2016；Odom et al., 2018)。Newhart 等人（2019）详细描述了这些为了适用于城市污水处理的 DDM 调整。

16.2.4.1 逐步变量选择

通常情况下，预测模型的拟合参数被用作相应变量重要性的指标；然而，

这种方法的适用性取决于 DDM 方法。例如，许多 ML 模型中的预测变量之间复杂的非线性关系在所有情况下都不容易通过单个预测变量的权重来概括。因此，如果目标是更好地理解过程而不是预测准确性，那么应该选择比 ML 方法更易解释的方法。如果预测准确性是主要驱动因素，并且使用的 ML 模型不能揭示机理信息，那么逐步变量选择是特征选择的一种方案，以降低维度并去除不相关的预测变量。

逐步变量选择方法，如前向和后向选择算法，使用信息标准（例如，RMSE、AIC、BIC）为给定模型分别迭代添加或删除预测变量，以确定哪些变量进入或退出模型。然而，逐步法可能会使参数估计出现偏差，即使变量的数量适中，对所有变量子集执行完整搜索在计算上也是不可行的。可以比较前向和后向的逐步选择以确保结果的相似性(John et al., 1994)。

16.2.4.2 套索算法

诸如多元回归之类的统计模型通常使用普通最小二乘法（OLS）进行拟合，它估计使得实际值和预测值之间的 SSE 最小化的参数。虽然 OLS 模型拟合方法会产生无偏估计值，但训练数据集中的测量误差会产生高方差，这使得模型难以泛化(Jameset al.，2013)。另一种引入少量偏差但降低变异性和模型复杂性的统计模型拟合方法是套索（lasso）算法（最小绝对值收敛和选择操作）(Tibshirani，1996)。套索算法同时执行变量选择和参数估计，而不是分两步执行这些任务。

套索算法将不重要的预测变量的系数缩小到零，从而仅选择那些具有非零系数的变量保留在模型中；然而，传统的套索算法并不总是选择正确的变量子集。例如，如果两个变量高度相关，则它们都将包含在最终的变量选择中。为了解决这个问题，Zou（2006）提出了自适应套索算法，它已被用于市政污水处理的序批式膜生物反应器中的故障检测(Newhart et al., 2020)。融合套索算法（fused lasso）是套索算法的另一种变体，它可以处理预计顺序时变系数相似的时间序列数据。在融合套索中，相邻系数之间的差异会受到惩罚，而不是系数本身(Hastie et al., 2015；Tibshiraniet et al., 2005)。Klanderman 等人（2020）举了一个在 WWTP 中用于故障分离的融合套索算法的例子。组套索（group lasso）是生物科学中一种有用的套索变体，可以识别出共同进入或

退出模型的变量组(Yuan & Lin, 2006)。Bai 等人（2019）的文章中提供了 WWTP 中的一个组套索示例。

16.2.4.3 沙普利加解释法

沙普利加解释法（SHapley Additive exPlanations，SHAP）类似于逐步变量选择，即将输入变量按顺序添加到模型中，但不同之处在于如何使用输出值来确定最佳子集。首先，根据平均训练数据假设一个基线期望值，并逐步添加输入变量以计算新的期望值。连续输入特征的期望值之间的差异表明了输入变量对输出值的影响大小和方向。但必须测试输入变量的所有可能排列组合，以考虑输入变量之间的交互效应。 沙普利值（Shapley value）是基于与基线相比的所有可能变量组合的输入变量的平均贡献(Shapley，1951)，是一种在 ML 中方法不可知的变量选择方法(Lundberg & Lee，2017)。

16.2.5 预测和预报

预测模型是过程的一种数学表示，使得一组预测变量可在给定的操作条件范围内估计响应变量。当响应变量是预测变量的未来值时，该模型称为预报（forecast）模型。预测模型和预报模型之间的主要区别在于它们的使用方式。预测（prediction）模型通常用于探索预测变量和响应变量之间的关系以及估计样本内的数值。相反，预报（forecasting）模型用于预测响应变量的未来值，并且还应考虑从一个观测值到下一个观测值的时间依赖性。预测和预报模型可以轻松地整合到 WTP 和 WWTP 的现有分布式控制系统（DCS）中，从而可以实时估算难以测量的变量以用于控制(Newhart et al., 2020)。在这种情况下，预测模型的使用通常被称为软传感器，其名称来源于“基于硬件”的传感器（包括传统的在线仪器）与“基于软件”的预测或预报模型之间的区别。

16.2.5.1 多元回归

最简单的建模方法是使用预测变量（X）的线性组合来计算响应变量（Y）。虽然预测变量可能会进行不同的转换（例如，对数正态、指数），但其模型做出了线性、无多重共线性和同方差的假设。模型误差通常被认为是正态分布的，均值为零和方差给定。以预测变量为条件，正态分布的假设也适用于 Y。使用散点图、直方图及拟合模型中预测变量和残差的相关系数验证数据是否

满足这些要求。鉴于水务和污水处理过程的复杂性，多元回归模型很少能提供最准确的预测，但一般来说它们是一个很好的起点。多元回归模型的一个好处是，当所有预测变量都标准化时，可以根据每个预测变量对响应变量的关系强度直接比较它们的估计系数。这些模型可以提供操作指导和对推动处理性能的现象的初步了解，而 ML 预测方法不会自动提供此信息。

通常情况下，线性关系适用于狭窄范围的操作条件，因此可能需要多个模型来估计一个更广泛、更实际的范围。包含不同预测变量范围的条件系数的多重线性模型是基于样条（spline-based）的模型。鉴于多元回归模型的线性形式的局限性，个别项也可以用非线性函数代替，例如多元自适应回归样条（MARS）(Friedman, 1991)中的基函数或广义加法模型（GAM）(Hastie & Tibshirani, 1999)中的多项式。

当随时间观测到的变量及其相应的自相关图具有明显的周期性变化模式时，可以使用具有正弦和余弦项（即傅里叶级数）的多元回归：

$$f(x) = \sum_{k=1}^{K} \alpha_k \cos(\frac{2\pi kX}{T}) + \sum_{k=1}^{K} \beta_k \sin(\frac{2\pi kx}{T}) + \varepsilon \qquad (16.9)$$

其中 K 是余弦和正弦对的数量；X 是某个时间段 T 中的时间；α_k 和 β_k 是估计的模型系数。例如，Newhart 等人（2020）使用正弦、余弦和过程变量的线性组合来模拟活性污泥系统中的氨浓度，正弦和余弦项可以捕捉氨浓度的日变化。在这种情况下，T 是 1440 分钟（相当于一个周期的长度为 1 天），X 是一天中的一分钟。可以使用变量选择方法来选择 K。

16.2.5.2 神经网络

神经网络是市政水务和污水处理中研究最广泛的 ML 预测方法之一(Khataee & Kasiri，2011)。神经网络是一种智能计算形式，它以模仿生物神经通路形成的方式将输入变量映射到输出变量（即分别为解释变量和因变量）(Beale& Jackson，1990；Bishop，1995；Kasabov，1996)。人工神经网络（ANN）是最简单的神经网络形式，包含三层计算节点（即神经元）：一个输入层、一个隐藏层和一个输出层。如果使用多个隐藏层，则该结构称为深度神经网络（DNN），可用于解决高度复杂的问题，但 DNN 需要大量数据和时间来训练(Schmidhuber，2015)。

当一个 ANN 被训练时，权重（w）和偏差（b）被调整以最小化实际输

出和预测输出之间的误差。前馈 ANN（一层的输出是下一层的输入）最流行的训练算法是反向传播。反向传播算法的工作原理是计算损失函数（也称为成本或目标函数）相对于每个权重的梯度，每次计算一层梯度，然后从最后一层向后迭代(Nielsen，2015)。可以在第 16.2.2 节中找到用作损失函数来比较不同模型结构（例如，隐藏层中的节点数、激活函数的类型）的误差度量。

每个节点都包括一个激活函数（步进、线性或非线性），该函数从前一层获取标准化的输入值，使用权重和偏差调整每个输入值。Sharma 等人（2020）对 ANN 中不同激活函数进行了总结。环境工程中最广泛使用的 ANN 激活函数是 sigmoid 函数，其中 x 是节点的输入向量；w 是节点的对应权重向量；b 是节点的偏差：

$$output = \sigma(\omega \cdot x + b) \tag{16.10}$$

其中 $\sigma(z) = (1/(1+e^{-z}))$。

在隐藏层使用 sigmoid 函数并在输出层使用线性函数的神经网络通常称为多层感知器（MLP）网络。另一个越来越受欢迎的 ANN 是径向基函数（RBF）神经网络。在 RBF 网络中，非线性径向距离函数被用于具有线性输出层的隐藏层。到目前为止讨论的所有 ANN 都假设训练和预测中使用的观测值是相互独立的，但是一种称为循环神经网络（RNN）的用于自相关数据的神经网络在环境数据设置中越来越受欢迎(Newhart et al., 2021)。在 RNN 中，节点的输出用作下一次观测的输入。RNN 的这种内部记忆功能允许以有序的顺序考虑观测结果。总之，通过在不同层中使用不同的激活函数定义了无数的神经网络配置。文献尚未建立用于水务和污水处理的“最优”神经网络；因此，在为特定应用开发预测模型时，尝试一系列选项非常重要。

16.2.5.3 模糊逻辑

模糊集理论（也称为“模糊逻辑”或 FL）通过分配部分隶属关系允许对没有明确界限的数据进行一般分类(Zadeh，1973)。简而言之，FL 允许将观测结果放在多个类别中（将某些类别指定为比其他类别更有可能）以解决不确定性。图 16.5 说明了用于将 FL 应用于加热系统的函数。输入数据（例如，传感器测量值、标签）通过应用隶属函数分配语言变量从而被“模糊化”，如图 16.5a 中分配温度级别的三角形隶属函数或图 16.5b 中分配加热器功率调整的

高斯隶属函数。隶属函数为每个语言标签分配 0 到 1 之间的多个值，其中 0 表示观测值不属于给定的模糊集，1 表示观测值完全属于模糊集。例如，如果图 16.5 中的温度介于两个值之间，则观测结果在两个标签中具有部分隶属关系，例如 0.7 寒冷（Cold）和 0.3 凉爽（Cold）；尽管所有语言变量的隶属度总和并不必须为 1。“如果-则”（“if-then”）规则会应用于每个模糊集（“推理”）。一个规则集可以确定“如果温度较为寒冷，则大幅增加加热器的功率”和“如果温度较为凉爽，则稍微增加加热器的功率”。在图 16.5b 中，推断值对应于 0.8 大幅增长和 0.2 轻微增长。可以通过使用推断值的加权平均值来计算中心值以产生一个单一数字输出（“去模糊化”），但也存在替代方法。在加热器的例子中，中心值方法使得加热器的功率调整在大幅和轻微之间。

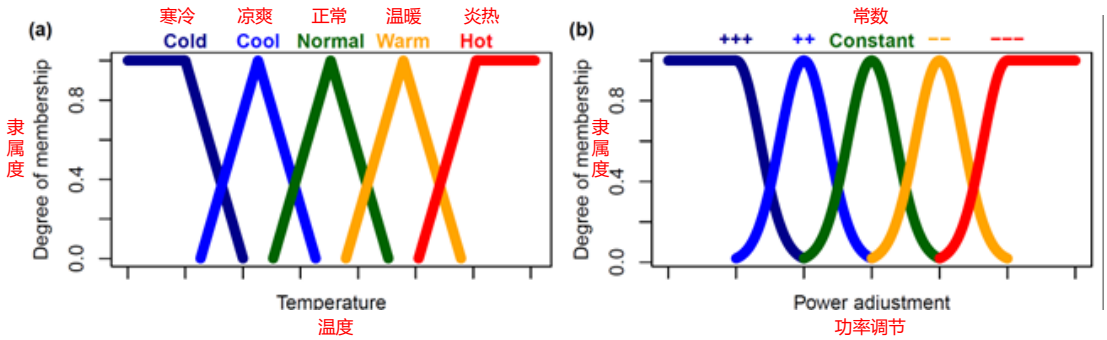

图16.5 (a)用于模糊化的三角隶属函数和(b)用于去模糊化的高斯隶属函数的示例。“++”或“--”表示功率略有变化，“+++”或“---”表示功率在相应方向上发生显着变化（增加或减少）

因为“if-then”规则是明确定义的，所以该方法被视为一个专家系统（expert system），而不是数据驱动系统。然而，模糊推理规则的权重也可以使用神经网络等 DDM 方法来识别更复杂的问题，但它们可能会失去“if-then”专家结构的真正可解释性(Hüllermeier，2015；Jang，1993)。数十年来，FL 控制器一直被提议用于具有时变和非线性系统的加工工业，包括水务和污水处理(Ferrer et al., 1998； Fiter et al., 2005)。

编写条件语句的两种最常见的 FL 方法是 Mamdani 和 Takagi-Sugeno（TS）。Mamdani 模糊规则遵循简单的“if-then”逻辑。在示例中，“如果酸流量低，则 pH 值高”，酸流量和 pH 值是语言变量，低和高是隶属函数的语言值。相比之下，TS 模糊规则使用类似的“if”逻辑和数学方程（如输入变量的常数、线性、非线性组合）(Takagi &Sugeno，1983)。例如，“如果酸流量低，则 pH

= k·flowz + c”，其中 k、z 和 c 是拟合的模型参数。

开发智能 FL 控制器的步骤如下(Manesis et al., 1998)：

（1）将变量分为控制变量和受控变量。受控变量量化了系统的特性（如性能、水质）。调节控制变量以将受控变量保持在其设定点。例如，再循环泵的流速是控制变量，而随再循环泵流速变化的悬浮固体浓度是受控变量。

（2）为每个控制变量建立一组语言描述符（例如，高、正常、低），使得工厂操作员可以理解这些语言描述符。粒度与描述符的数量直接相关，不过对于大多数控制应用程序来说，颗粒度在 3 到 5 之间较为合适。对于每个集合，确定一个隶属函数(Ross，2010)，但单个函数不如集合中语言描述符的数量重要(Sadollah，2018)。

（3）使用控制变量和受控变量的语言描述符定义“if-then”规则以形成知识库。“if-then”规则的形式取决于系统是 Mamdani 还是 TS。

（4）为控制变量和受控变量的隶属函数选择加权平均方法（例如，最大-最小、重心（COG））。

自适应神经模糊推理系统（ANFIS）是结合 ANN 和 TS FL 的优点的五层网络，它通过分别对 ANN 的输入和输出进行模糊化和去模糊化以提高噪声数据的预测精度(Abraham, 2005)。由于这种混合结构，ANFIS 被认为是一种通用估计器(Jang et al., 1997)。训练 ANN 的相同方法（如反向传播）可用于调整 FL 参数；但必须定义隶属函数本身。此外，与传统的 FL 模型相比，ANFIS 第一层中建立的 TS 规则不再具有可解释性的优势，必须使用替代变量重要性方法来理解输入-输出关系。第 16.3.2 和 16.3.4 节描述了 ANFIS 在水务和污水处理中的应用。

16.2.5.4 决策树

ML 中的“决策树”是一种基于一系列二元分类（例如 x > 1）的启发式建模技术，用于对变量进行分类或预测。有许多潜在的定量（回归）和定性（分类）问题可以用决策树来回答，如图 16.6 所示，在特定的水质条件下应添加哪种碱的问题。基于决策树的模型有许多优点，包括缺乏单个变量分布或解释变量和因变量之间的关系类型（如线性或非线性）的假设。通过分支可有效地创建多个模型，而不是使用一个全局模型来描述整个数据空间，并且可

以处理大量特殊情况。最后，基于决策树的模型对异常值干扰有抵抗力(Steinberg & Colla，1995)，但它们可能对分支分割的变化以及分割变量的指定很敏感。Sutton（2005）对分类树和回归树进行了很好的介绍和数学描述。

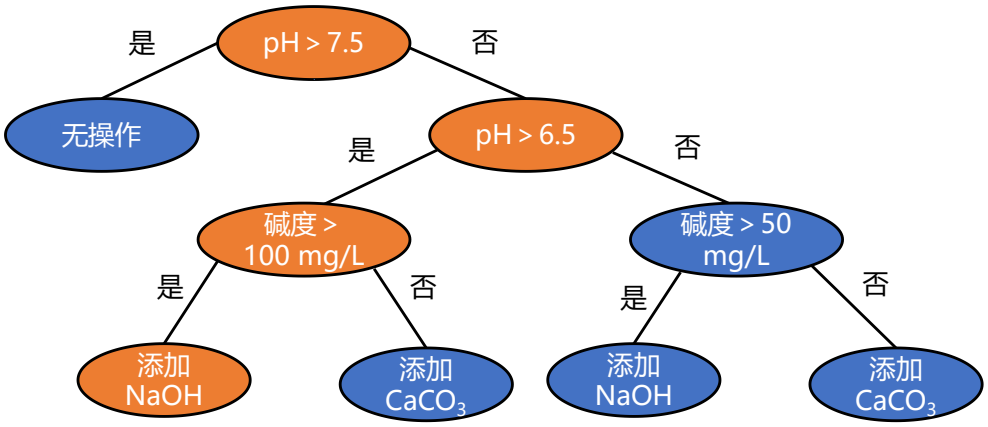

图16.6 根据水质特征确定添加哪种碱（NaOH 或CaOH）的决策树示例。每个节点（圆圈）代表一个二元分类器。橙色节点表示对于 pH值为7和碱度为150 mg/L的水达到“添加 NaOH”作用所采取的路径

最早和最普遍的决策树模型拟合技术是自举聚合（bootstrap aggregation），也称为 bagging(Breiman，1996)。在 bagging 中，通过对数据进行替换抽样来创建许多人工样本，这一过程称为自举法（bootstrapping）。随后每个自举样本都用于训练一个预测方法，每种方法都获得一个预测或分类，然后通过平均回归或投票分类将结果组合起来。boosting 是 bagging 的变体，其中加权平均值用于聚合回归模型的结果，且自举样本（boostrap sample）的重新抽样随着模型每次拟合迭代而变化。通过在后续训练步骤中包含更多错误预测的观测结果，可以创建能够处理特殊情况的模型。bagging 和 boosting 之后还开发了其他分类器算法，其中最流行的是 AdaBoost 算法，它使用一系列单一二元分类器的加权平均，这些分类器是通过赋予先前分类器的错误分类样本更高的权重来确定的(Freund & Schapire，1997)。

水务和污水处理中最流行的决策树方法是随机森林（random forests，RF）。随机森林是大量决策树的平均值，通过递归子集输入变量和训练观测值随机再抽样（即 bootstraping）创建(Breiman，2001)。拟合算法（特别是 boosting 或 AdaBoost）中每个二元节点的权重可用于确定模型中特定预测变量的重要性。但输入变量权重将根据节点的拆分方式而有所不同，因此可以作为变量

重要性的一个不一样的指标。最后，为了拟合 RF 模型，决策树的数量、最大深度、在节点处拆分的最小样本数以及评估最佳拆分的最大变量数是需要调整的重要超参数，这涉及计算负担和准确性之间的直接权衡。

16.2.6 优化

优化算法有两个主要的实际应用：优化预测模型（即通过调整内部模型参数或超参数来实现最低误差）(Le et al., 2019)或寻找现有预测模型的最佳输入集。本章重点讨论后一种优化应用，因为它与脱碳目标更相关。将数据驱动的预测模型与优化相结合，构成了模型预测控制。

元启发式算法是非精确框架，旨在搜索全局最优解空间，而不计算每个可能的解。鉴于很少有精确的水务和污水过程的力学模型能够准确地进行全面监测和控制(Newhart et al., 2019)，因此元启发式算法被用来确定第 16.2.4 节所述预测模型的最优解。在水务和污水分配和处理的文献中，将在本节中介绍最流行的三种算法是遗传算法、粒子群优化和模拟退火。对于此处未列出的其他精确和启发式方法，Beheshti 和 Shamsuddin（2013）提供了更完整的优化方法清单。

16.2.6.1 遗传算法

在预测模型中搜索最佳点的一个相对快速的方法是遗传算法（GA）(Reeves &Rowe, 2002)。第 16.3.1 节和第 16.3.4 节分别描述了遗传算法在泵调度和 AD 操作中的应用。基于达尔文进化论的思想，利用 GA 确定最优解的步骤为：

（1）初始种群：每个“个体”都由一组变量（“基因”）来描述，并显示出潜在的解。这些个体观察值可以随机生成以覆盖整个可能解的范围，或基于原始数据。

（2）适配度函数：根据适配度函数为个体分配一个“适配度分数”。该函数对各种标准下个体解的质量进行量化，如最小化能源消耗。

（3）选择：选择适配度得分最高的个体。

（4）交叉：最适个体对交换基因，由此产生的“后代”具有新的变量集并被添加到种群中。

（5） 突变：低比例的后代基因经历随机变化以维持种群内的多样性。

（6） 终止：为了保持一个种群规模恒定，最不适合的个体被剔除。适配度、选择、交叉和变异的顺序一直持续到种群收敛。

GA 的优势在于它能够处理解空间很大且解空间边界难以确定的问题。这是通过使用少量的分布在整个解空间的个体来实现的。虽然种群规模可以更大，但会显著增加计算时间。计算步骤的数量通常等于代数（第 2-4 步）乘以种群规模。由于个体的适配度用于确定最优解，而不是导数或梯度（如在传统优化中），所以 GA 倾向于识别解空间中的全局最优而不是局部最优(Kurek & Ostfeld，2013)。GA 的一个缺点是生成个体的过程可能会产生技术上不可行的解。在这种情况下，必须使用替代方法开发解空间的模型，明确定义合理边界，或使用离散个体而非连续个体(Sadatiyan Abkenar et al., 2015)。

16.2.6.2 粒子群优化

粒子群优化（PSO）是一种基于群体（例如鸟类、鱼类）的自然运动和信息的稳健随机优化程序(Eberhart & Kennedy, 1995)。与 GA 一样，PSO 是一种基于种群的搜索方法。一旦定义了模型空间，粒子群（即观测值）就会随机初始化为模型空间每个维度中的位置和速度向量。每个粒子将迭代地搜索模型空间中的最小值（即适配度值）。与自然界中的蜂群类似，粒子将使用来自其他粒子的局部最小值的信息来告知下一个搜索方向。然而，当初始化的粒子太少时，搜索可能会陷入局部最小值，这是 GA 更擅长避免的问题(Beheshti & Shamsuddin，2013)。当有大量粒子时，可以找到全局最小值，但计算量更大。当粒子数量合理时，PSO 可以成为 GA 的有效替代计算方案(Hassan et al.,2005)。

每个粒子包含三个向量：当前位置（x）、粒子目前遇到的最优解的位置（p）以及粒子行进的方向（即梯度）（v）。粒子将沿局部最佳（p，基于当前位置）和全局最佳（g，基于所有粒子的 p）组合的方向行进（方程（16.11））：

$$\begin{aligned} \upsilon_{i+1} &= W\upsilon_i + (c_1 r_1 (p - X_i)) + (c_2 r_2 (g - X_i)) \\ \upsilon_{\min} &\le \upsilon_{i+1} \le \upsilon_{\max} \end{aligned} \tag{16.11}$$

其中 W 是惯性权重；c_1 是局部最小值对速度矢量的影响（即自信心系数）； c_2 是全局最小值对速度向量的影响（群体置信度系数）；r_1 和 r_2 是随机生成的 0 到 1 之间的数字。

16.2.6.3 模拟退火

模拟退火（SA）是一种搜索技术，它基于一个常见的热力学原理，即处于平衡状态的原子集合在特定温度下的概率分布(方程 (16.12)，Metropolis et al., 1953)以确定模型空间的全局最优值(Kirkpatrick et al., 1983)：

$$e^{-\frac{\Delta D}{T}} > R(0,1) \tag{16.12}$$

其中 ΔD 是状态之间的距离变化，T 代表不同状态下解空间范围的合成温度，$R(0,1)$是 0 到 1 之间的随机数。概率方法对于避免陷入局部最小值非常重要，它可以通过探索合理的解空间以找到全局最小值。在 SA 的每一步中，邻近状态（s*）与当前状态（s）进行比较，并以概率方式决定将系统移动到状态 s*还是保持在状态 s。这些概率方式最终导致系统移动到低能量状态。通常情况下这个步骤会重复进行，直到系统达到足以满足应用的状态，或者直到给定的计算预算（例如，迭代次数）已用完。为了在策略上使得目标函数缓慢下降时实现全局最优，可以使用“退火计划”，当目标函数达到平稳状态时，T 会迭代减少。但如果状态之间的初始步长不够小，则无法保证会找到全局最小值。在实践中，这种粒度的计算要求通常超过了性能的改进(Trosset，2001)。在常规应用中，SA 会持续迭代，直到 300 次迭代中没有发现目标函数的变化为止(Prakash et al., 2008)。如果搜索空间一般是平滑的或存在多个局部最小值，则 SA 可能会提前终止或陷入局部最小值。在这些情况下，PSO 可能是更好的选择。第 16.3.3 节描述了 SA 如何用于污水处理厂的曝气控制。

16.3 数据科学在选择处理系统中的应用

在以下章节中，我们将举例说明 DDM 方法和框架的多样性，这些方法和框架用于实现 WTP 和 WWTP 中常见能源密集型工艺的类似目标（即能源优化）：泵送、化学添加（混凝）、曝气（硝化）和沼气生成（厌氧消化）。

16.3.1 泵送优化

水、污水和生物固体的泵送消耗占据公用事业能源需求和维护成本的很

大一部分(Shi，2011)；对于许多饮用水公用事业单位来说，这一比例高达90%(Cherchi et al., 2015)。在水处理分配系统中，泵送计划用于减少能源消耗；然而，由于分配系统是一个具有多个约束的高度非线性系统，因此很难用传统的建模方法确定最优解。尽管这些原则可以普遍适用于 WTP 和 WWTP 的大多数机械设备，但仍然存在对泵的现实约束，包括设计效率低下、最小和最大运行时间、每小时最大启动时间、最小休息时间、最小和最大流量、最大排放压力、最小和最大工厂生产率以及启动或关闭的超前-滞后顺序(Cherchi et al., 2015)。当前“无成本”能源优化的“最先进”控制方法包括基于储罐水位的开-关调度(Nybo et al., 2017)、允许调整单个泵速度的变频驱动器（VFD）、负载转移（即上游泵调度）和流程优化(Shankar et al., 2016)。下面讨论科学文献中的例子，并在表 16.1 中进行了总结。供水系统的商业优化软件、支持软件的基础设施（数字和物理）、劳动力和操作员培训通常有 2-5 年的投资回报期，能源成本降低 5-15%(Badruzzaman et al., 2014)。

表 16.1 用于泵优化的数据科学水应用示例

作者	目标	方法	配置	结果
Torregrossa 等人 (2017)	效率监控	FL	Mandami	泵能耗降低 18.5%
Sadatiyan Abkenar 等人（2015）	泵调度	GA	离散型	确定了尽量减少开关的最低能耗策略
Kebir 等人 (2014)	实时 VFD 调整	FL	Mandami	与开/关操作策略相比，节省 40%的能源（理论上）
Zhang 等人（2012）	泵调度	ANN	PSO	与开/关操作策略相比，可节省 8-24%的能源

Torregrossa 等人（2017）开发了一种 FL 泵的性能指标，用于监测效率并建议对流量条件进行预防性或即时维护。为此，基于提升的水量和消耗的能量计算效率指数，并使用滚动中位数来区分长期趋势和归因于条件变化的波动。连续多日的短期负波动表明需要维护。维护响应的即时性由权衡长期和短期效率的 FL 系统确定，并且假设维护能够使泵恢复到与新的、更高效的泵相比的基线效率，进而评估了维护与更换的经济后果。

在美国密歇根州门罗市的一个中等规模供水系统的水力模型中，Sadatiyan Abkenar 等人（2015）使用 GA 方法对两个泵的抽水计划进行了优

化，同时使能源消耗最小化，并包括对高压的额外惩罚。使用成对的开始和停止时间作为基因的连续方法产生了不可行的解（即冲突的 ON 或 OFF 时间）。为了缓解这种情况，在计算解的适用性之前，对产生不可行解的任何突变进行“修复”。另一种离散方法使用二进制 ON 或 OFF 指示符表示 1 小时间隔，其中每个间隔是一个基因，仅产生可行的解。

Kebir 等人（2014）对一个依靠顺序 ON/OFF 进水泵送策略的全规模污水处理厂（WWTP）进行了建模，其本质上是低效的，并提出了一种新的 FL 控制器，该控制器通过与上游水库平均高度的偏差来调整泵的 VFD；报告声称能耗减少 40%。Zhang 等人（2012）使用人工神经网络（ANN）开发了一个针对给定流量、并联运行的泵配置和上游水库水位的能源消耗模型。然后，他们使用 PSO 确定给定流量、所需水库水位和系统物理约束的最佳泵送计划；报告声称能耗减少 8-24%。

大型 WTP 或 WWTP 的一个重要考虑因素是能源成本，尤其是在能源成本全天变化或公用事业根据每月能源消耗峰值计费的情况下。科学文献中的作者在很大程度上忽略了随时间变化的能源成本。相反，能耗模型是基于 VFD 频率指标或单个设备的能源等级开发的。在大多数情况下，能耗最小化会使得成本最低；然而，预测模型可能需要包含描述真实成本的成本函数的变化，而不是使用消费作为代表指标。例如，在特定时间启动泵送可能不是最节能的行动，但如果即时需求（当必须降低储罐水位时）与增加的能源成本一致时，可能会降低一天中的泵送成本。

16.3.1.1 混凝

混凝是 WTP（在某些情况下是 WWTP）中添加化学物质（即混凝剂）以破坏胶体和悬浮颗粒物质稳定性的过程，使颗粒聚集（即絮凝物）并更容易通过重力去除较大的、带中性电荷的聚集体。用于混凝和絮凝的化学品的产生和运输可占 WTP 碳足迹的 5-20%(Biswas & Yek，2016)；因此，取决于处理设施的规模和初始水质，精密化学处理可以显著节省成本和减少碳排放。为了减少用于水处理的化学物质的量，必须设计剂量控制策略，以适应由于混合不良和水质变化导致的非理想的物理化学反应。然而，在全规模处理中很少出现这种情况。在 WTP 中，化学品计量主要是按流量进行的，通过调整

与水流量成比例的化学品计量泵的流量来维持每单位体积水中的化学品浓度。浓度设定点通常仅在发生重大水质变化或处理过程紊乱时才会调整，因为在实验室台架试验（即罐试验）中确定“理想”剂量既费时又费力，而且结果可能与全规模有很大差异。因此，使用数据驱动的化学剂量方法可以显著提高处理稳定性并减少处理设施的碳足迹。科学文献中的例子讨论如下。

应用人工神经网络（ANN）预测混凝剂剂量并不是一个新概念。Van Leeuwen 等人（1999）能够使用罐子历史测试数据和 ANN 预测给定水质的明矾剂量；尽管多元线性回归模型取得了类似的结果。十年后，Maier 等人（2009）基于与 Van Leeuwen 相同的数据，使用 DNN（两层人工神经网络）来预测处理后的水质（浊度、颜色、pH、UV-254、残留明矾）和最佳明矾剂量，能够将预测误差的标准偏差降低 37%。Zangooei 等人（2016）基于罐子历史测试数据，使用 pH、初始浊度、温度、混凝剂类型（如不同供应商的固体或液体聚氯化铝）和混凝剂浓度来预测浊度。具有两个隐藏层的 MLP 优于 RBF ANN 和 FL 回归模型，并且训练所需时间更少。同样，Wu 和 Lo（2008）发现，在有进水水质数据的情况下，ANN 在已处理水质预测方面优于 ANFIS 预测模型。在没有实时水质的情况下，ANFIS 模型能够根据历史趋势和当前给药剂量更准确地预测处理后的水质。当历史给药数据可用时，Wu 和 Lo（2010）发现加入前一个时间段的混凝剂剂量（DNN 模型的输出变量）可以减少测试误差。

Chen 和 Hou（2006）观察到，多元回归模型能够利用历史数据预测地表水的混凝剂剂量和 pH 调节剂量。然而，两个模型分别是针对低进水浊度和高进水浊度条件开发的。Chen 和 Hou 进一步使用 Mamdani FL 调整反馈控制参数，以最大限度地减少混凝剂剂量，同时实现污水浊度和 pH 目标。Bello 等人（2014）提出了一种线性化的 TS 模糊模型预测控制策略，通过保持处理水的表面电荷和 pH 值来提高混凝剂剂量控制的稳定性。根据可用数据和在线仪器的质量，FL 控制器可以提高传统级联控制的精度和稳定性。为了最大限度地减少混凝剂剂量，需要汇总进水水质、最终水质和混凝剂剂量，以训练一个处理后水质的预测模型。预测模型选项包括多元回归、ANN 或 DNN。为了使模型的预测能力得到全规模应用，可以通过多种方式将最优预测模型纳入控制策略。最基本的控制选项是标准级联控制，即当处理过的水质变量

（如浊度）超过阈值时，就增加混凝剂剂量。然而，这需要一个被充分理解的剂量-反应关系。在处理后的水质未达到目标时，增加剂量的严格规则可能会使混凝剂的浓度超过满足电中性的需要，从而导致污水浊度恶化，因为颗粒重新稳定在悬浮状态(Tchobanoglous et al., 2014)。拟议的 FL 控制器可以通过包含考虑水质恶化的规则集来防止这种过量，但这种方法需要更复杂的编程和调整结果语句参数的方法。调整可以在 WTP 现有的 DCS 结构内手动完成，但如果使用 ANFIS，则需要在外部完成。如果开发额外的 ANN 或 DNN 来确定特定污水水质所需的混凝剂剂量，同样会引起程序复杂性问题。在这种情况下，控制器必须在单独的服务器系统上运行，该系统可以为现有的级联控制策略提供输出。

16.3.2 硝化

生物养分去除（BNR）是 WWTP 中最昂贵、最易变且最难建模的过程；然而，世界上大多数现代设施都需要它来实现所需的氮和磷去除。大多数 WWTP 工艺共有的两个因素造成了建模和控制的困难：缺乏可靠的仪器和非理想的全规模工艺条件。由于仪器本身的生物膜生长和竞争性离子或固体干扰，处理过程中的微生物固液基质（即活性污泥或 AS）会干扰普通的在线仪器测量。利用光而不是离子传输的仪器，如溶解氧（DO）浓度，足以提供可靠的测量，且清洁和维护频率较低。DO 是在活性污泥系统中测量的关键水质参数，因为特定形式的氧气的可用性决定了微生物活性，从而决定了特定的污染物转化。以游离氧（O_2）形式提供的水氧会增加 DO 浓度，是好氧条件的指标。当氧气仅以硝酸盐（NO_3^-）的形式存在时，为缺氧条件的指标。当没有氧气可用时，为厌氧条件的指标。正是这些氧化条件的战略性交替将相关污染物（即碳、氮、磷，以及在较小程度上的硫）转化为气相或固相，从而降低了水中污染物的浓度。虽然 DO 传感器可以确保满足曝气条件，但测量本身是污染物转化完成的代表。例如，低浓度污水（如低浓度的有机物）将不需要那么多的氧气来实现处理目标；但在大多数系统中，无论需求如何，都会继续提供曝气以保持 DO 设定值。

序批式反应器（SBR）是一种常用的污水处理技术，其中一个具有生物

活性的完全混合的反应器经历一系列不同的操作条件以实现污染物的去除。SBRs 最常见的控制策略是在处理周期的每个阶段使用具有不同 DO 浓度设定值的时间序列。DO 设定值由操作员的经验和对每个阶段所需环境条件的一般知识确定，每个阶段都会激活一组独特的微生物。这种控制策略只需要一个在线仪器（DO 传感器），且可确保在稳定的进水条件下实现所需处理。但溶解氧是处理过程中去除的实际污染物的替代物，不能保证污水水质满足标准。从历史上看，这种不确定性已通过提高 DO 设定值来解决，以充分氧化化学污染物并确保微生物过程不受基质限制。这种方法增加了处理过程的能耗，占污水处理设施总能耗的 35-50%(Newhart et al., 2020)，是仅次于人工的第二大运营成本(Lindtner et al., 2008)。为了减少与 SBR 和其他二级生物处理系统（即传统和新型活性污泥配置）中的曝气相关的能耗，需要新的智能监测和控制策略。下面讨论了文献中的例子，并在表 16.2 中进行了总结。

表 16.2 用于曝气的数据科学水应用示例

作者	目的	方法	配置	结果
Traoré 等人（2005）	DO 控制	FL	Mamdani	在更广泛条件下提高稳定性
Ferrer 等人（1998）	DO 控制	FL	Mamdani	与 ON/OFF 相比，节能 40%
Fiter 等人（2005）	DO 控制	FL	Mamdani	与 ON/OFF 相比，节能 10%
Du 等人（2018）	DO 控制	NN	RBF	曝气能耗减少了 100kWh/天
Asadi 等人（2017）	DO 优化	SA	MRAS	气流减少 30%

Traoré 等人（2005）为处理城市污水的步进式中试规模的 SBR 提出了一种 FL 曝气控制策略。 FL DO 控制器的规则是根据测量的 DO 和循环阶段确定空气流量以维持 DO 设定值。与 ON/OFF DO 控制方法（当 DO 测量值超过设定值时，关闭空气；当 DO 测量值小于设定值时，打开空气）和传统的比例-积分-微分（PID）控制相比，模糊控制器能够在更广泛的环境条件下更精确地保持 DO 设定值。在模糊规则中加入 pH 值和摄氧率（OUR）可以缩短曝气周期时间并进一步降低能耗(Puig et al., 2006)。Ferrer 等人（1998）将类似的模糊 DO 控制器用于试验性 BARDENPHO 活性污泥系统；与 ON/OFF 方法相比，在提高精度方面显示出类似的结果，且节能高达 40%。Du 等人（2018）开发了一种 RBF NN 来调整级联控制参数以提高 DO 控制器性能，

包括显著降低变异性（干燥天气流量为 67%，潮湿天气流量为 59-93%）和略微减少曝气能耗（100kWh/d）。

调整单个 DO 控制器的另一种方法是在给定全系统模型下确定最佳控制策略。为了实现这一目标，Asadi 等人（2017）比较了底特律水务和污水处理行业二级曝气池中 MARS、ANN 和 RF 等的 DO。MARS 对 DO 和其他污水水质变量的预测优于 ANN 和 RF（基于 MAE 和 R^2 评价）。随后他们利用 MARS 预测模型比较了两组权重，一组强调最佳处理水质，另一组使用 SA 强调能源消耗。在对最佳水质进行优化时，他们表明可能在不影响处理后水质的情况下将空气流速降低 30%。然而，模型中没有考虑养分，这是大多数污水处理厂曝气需求和策略的一个重要驱动因素。此外，Asadi 等人（2017）认为，与统计模型相比，ML 模型需要对进水变量进行更频繁的采样。当对变量之间关系的形式所做的假设为真时，这种比较是成立的。一般来说，当数据较少时，统计模型做出的更简单的假设比 ML 模型能更准确地填补空白。相比之下，ML 模型至少需要训练数据中的条件实例，才能合理准确地预测测试数据中的类似条件。

16.3.3 厌氧消化

厌氧消化（AD）的主要功能是稳定固体和减少化学需氧量（COD）。但对脱碳至关重要的次要功能是产生能量。AD 产生的沼气中有 50%到 70%是甲烷(Holubar et al., 2003)，可用于现场产能或通过天然气管道进行净化、销售和分配。为了最大限度减少污水处理过程中的能耗，需要战略性地运行像 AD 这样的能量正收益工艺，以尽量减少过程干扰。例如，COD 负荷的大幅波动会导致中间化合物的积累，这些化合物对系统内的关键微生物群是有毒的。然而，AD 是最难建模、监控和控制的过程之一(Olsson，2006)，因此 AD 以高安全系数进行保守操作以确保稳定性。最终，这会导致大量处理过程效率低下，包括甲烷产量减少、由于运行中的 AD 反应器数量增加而导致的更高泵送成本，以及污水 COD 提高。由于控制变量和响应变量之间的复杂关系，且通过长保留时间进一步解耦，AD 工艺的 DDM 可以为更有效的操作提供见解。下面讨论科学文献中的例子，并在表 16.3 中进行了总结。

表 16.3 用于 AD 的数据科学水应用示例

作者	目的	方法	配置	结果
Akbaş 等人（2015）	预测、优化	ANN	PSO	甲烷百分比 +5%，沼气产量+64%
Holubar 等人（2003）	预测、优化	ANN	一次一个搜索	甲烷浓度 60-70%
Huang 等人（2016）	预测、优化	ANN	GA	沼气流量 R^2 = 0.91，MSE =2.0
Polit 等人（2002）	预测	FL	Mamdani	在不断变化的负荷下保持稳健
Turkdogan-Aydınol 和 Yetilmezsoy 等人（2010）	预测	FL	Mamdani	甲烷产量 R^2= 0.98, 沼气产量 R^2 = 0.98
Perendeci 等人（2009）	预测	ANFIS	滞后，相位	污水 COD R^2=0.89，RMSE=0.10

Turkdogan-Aydınol 和 Yetilmezsoy（2010）开发了一种多输入多输出（MIMO） FL 模型来预测沼气产量，该模型的性能优于多重非线性回归模型。Polit 等人（2002）使用具有模糊 pH 和温度系数调整的机械质量平衡模型来预测沼气产量，并且这种方法能够比单独的机械模型更好地跟踪负荷调整下的气体产量。 Holubar 等人（2003）使用 ANN 的分层系统来预测挥发性脂肪酸（VFA）的产量和 pH 值，然后预测在启动和稳定期间 AD 的沼气产量和成分。操作参数通过一次一个（one-at-a-time）搜索算法进行调整，以同时最大限度提高有机负荷率（OLR）和甲烷产量。

通过将 PSO 应用于沼气生产的 ANN 模型，Akbaş 等人（2015）确定了产生最高百分比的甲烷和沼气产量的操作条件。与历史数据的平均值相比，最佳条件使沼气中的甲烷含量提高了 5%，沼气产量增加了 64%。为了实现这一目标，污泥负荷率、温度、pH、总固体、总挥发性固体、VFA、碱度、固体停留时间（SRT）和 OLR 的日平均值被用作甲烷百分比和沼气产量预测模型的输入变量。使用提升树（boosting tree）算法(Breiman，1996)对输入变量进行降维，从而提高了预测性能。

FL 和 ANN 都可用于预测，虽然 ANN 已被证明更精确，但 FL 能够更好地处理输入和输出的变异性(Kambalimath &Deka，2020；Özcan et al., 2009)。有大量研究证明了 ANN 在预测沼气产量和 AD 性能方面的效率(Levstek & Lakota，2010)。因此，混合模糊模型（如 ANFIS）在解决 AD 系统方面出现

了热潮(Abrahart et al., 2008)。Perendeci 等人（2009）表明，ANFIS 模型能够预测季节性厌氧污水处理系统的污水 COD，通过添加输入变量，包括使用 10 天的 COD 历史数据来衡量系统是否处于启动或伪稳态条件的指标，从而提高性能。

与 WTP 中的混凝情况不同（第 16.3.2 节），WWTP 可以根据可用 AD 的数量和大小对有机负荷率、温度和 SRT 进行一些控制。为了充分利用 DDM 进行脱碳，预测模型应适用于沼气生产（包括数量和质量，例如特定的甲烷质量流速），并使用 PSO 或 GA 等优化方法来确定理想的操作条件。也可以针对无法快速测量但对理解性能至关重要的变量开发单独的模型，如 VFA。

16.4 全规模实施的建议

2020 年，水务和污水处理行业的领导者开会讨论数据驱动水系统的网络基础设施；在 DDM 从数据生成到使用、应用和展示的每个步骤中识别知识差距和有能力的人员(Ren et al., 2020)。虽然有一些工作组（如英国的非营利组织，智能水网论坛，Smart Water Networks Forum）和国际挑战（如由水研究基金会和水环境联合会共同发起的智能水系统挑战），但没有一个单一的联盟或文本涵盖与 DDM 相关的广泛主题。在没有一套全面的建议的情况下，目前需要各个公用事业单位来探索新的机会。大多数现代 WTP 和 WWTP 缺乏现代数字基础设施是实施 DDM 以实现脱碳的最大障碍。设施很少有适当的数据管理程序来广泛而有条不紊地组织数据库或一致的协议来清理固有的噪声数据。为了保持实时分析，还必须考虑一系列计划和实际影响。例如，大多数 SCADA 系统是为短期数据存储（最多三个月）和级联控制环路（主要是反馈）而设计的。集成 DDM 的方法包括：（1）必须升级 SCADA 系统以纳入历史数据和高级建模；（2）必须设计控制策略以允许 DDM 输出与现有数据框架进行信息传输。鉴于现有员工对基本控制结构的熟悉程度和可靠性，后一种选择是最实用且广泛的 DDM 实施策略。

在证明 DDM 系统的稳定性之后，可以添加复杂层级以提高处理过程效率。稳定性证明必须解决一些问题，如与传统控制策略相比的全流程变异性(Newhart et al., 2020)，以及在数据丢失或模型预测不可行的情况下的上限和

下限或其他应急措施。开发 DDM 控制的序列方案是与操作人员一起开发强大产品的机会，如果发生意外排放，操作人员可能会承担法律上的过失责任，从而要用他们对全规模处理过程动态的密切了解来平衡谨慎行事。当研究 DDM 作为水务和污水处理脱碳的潜在解决方案时，在确定 DDM 方法和集成策略之前，应明确讨论和定义以下因素：

- 涉及处理和能源绩效（如 kWh/MG、gCO_2/MG）的目标和关键绩效指标（KPI），包括工厂范围和特定处理过程；
- 限制性技术因素，包括操作限制、仪器、数据管理、控制系统结构和网络安全限制；
- 未知因素，例如对其他流程的影响或新技术的采用率。

一旦确定了上述项目限制因素，就开始正式选择 DDM 方法。一般步骤包括：

（1）识别可轻松与现有控制策略结合并提供实质性收益的预测变量（输入）和响应（输出）变量。

（2）开发实用和健全的数据混合协议，考虑现实世界的实现(即，合并变量具有不同采样频率时的观察结果)。这包括仔细考虑何时可以获得实验室数据。

（3）执行变量选择以最小化由不相关输入引入 DDM 的误差量。

（4）试验不同的建模框架，包括简单的和高级的，以评估所需响应的最佳候选者。如果优化是最终目标，按照预测模型开发进行实验以确定给定问题的“最佳”优化算法。

（5）通过提高和降低模型复杂性来调整预测和优化模型，以在测试数据上提供最准确的性能，此步骤不用于模型拟合。

（6）将模型编程到服务器上（使用 R 或 Python 等编程语言），该服务器可以将数据导出到特定公用事业公司使用的数据归档系统。数据归档系统经常可以访问 SCADA 系统而不会造成安全风险。

（7）随着时间的推移监控预测的稳定性，并在全面部署之前识别需要与新控制策略结合的不可预见的突发事件。

（8）为模型开发人员、控制专家和操作人员安排监测期，以同时观察全面实施情况。这些时间应跨越数周以彻底评估不同的环境条件，然后应逐渐

增加运行时间，直到操作人员对无监督操作满意。

（9）使用预先确定的 KPI 比较新 DDM 控制策略与原始策略的影响。如果新的 DDM 控制策略达到或超过原始策略的 KPI，则可以重复前面的步骤以纳入更多预测，或者通过消除控制环路的层级来更直接地依赖于预测。

16.5 结论

人们对将 DDM 集成到 WTP 和 WWTP 的关注度正在迅速增长，但公用事业公司在很大程度上被这项任务所压垮。同时开发良好的内部数据管理协议以及将 DDM 应对水务和污水处理的挑战可以显著降低清洁水的碳成本。鉴于目前水务处理每年消耗美国发电量的 3%，并且由于处理过程的需求和强度增加（即更高质量的污水），预计占比将增加到 6%(Chaudhry & Shrier，2010)，数据-驱动的处理过程优化是众所周知的降低碳和成本的“唾手可得的果实”。有大量科学文献将传统和新颖的 DDM 方法应用于工程环境系统，如 WTP 和 WWTP；然而，许多 ML 方法的启发式性质使得将任何一种方法视为特定处理过程应用的“最佳”方法是不可能的。使用已发表的文献作为指南，对单个公用事业的数据集进行实验才是真正的“最佳”框架。从根本上说，鉴于现有的机械和处理技术，对人员的投资和改进的操作策略将帮助 WTP 和 WWTP 以最小的环境影响实现其全部处理潜力。

致谢

这项工作得到了美国国家科学基金会 PFI:BIC 奖励编号：1632227 ；国家科学基金会工程研究中心计划合作协议 EEC-1028968（ReNUWIt）的支持。

参考文献

Abraham A. (2005). Adaptation of fuzzy inference system using neural learning. In: Fuzzy Systems Engineering, N. Nedjah and L. de Macedo Mourelle (eds), Springer, Berlin/Heidelberg, Germany, pp. 53–83.

Abrahart R. J., See L. M. and Solomatine D. P. (2008). Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications. Springer Science & Business Media, Berlin/Heidelberg, Germany.

Akaike H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723, https://doi.org/10.1109/TAC.1974.1100705

Akbaş H., Bilgen B. and Turhan A. M. (2015). An integrated prediction and optimization model of biogas production system at a wastewater treatment facility. Bioresource Technology, 196, 566–576, https://doi.org/10.1016/j.biortech.2015.08.017

Asadi A., Verma A., Yang K. and Mejabi B. (2017). Wastewater treatment aeration process optimization: a data mining approach. Journal of Environmental Management, 203, 630–639, https://doi.org/10.1016/j.jenvman.2016.07.047

Badruzzaman M., Cherchi C., Oppenheimer J., Gordon M., Bunn S. and Jacangelo J. G. (2014). Implementation of energy and water quality management systems modified with a GHG module. Proceedings of the Annual Conference & Exposition, Boston, MA, USA.

Bai H., Zhu R., An H., Zhou G., Huang H., Ren H. and Zhang Y. (2019). Influence of wastewater sludge properties on the performance of electro-osmosis dewatering. Environmental Technology, 40(21), 2853–2863, https://doi.org/10.1080/09593330.2018.1455744

Barbu M., Vilanova R., Meneses M. and Santin I. (2017). Global evaluation of wastewater treatment plants control strategies including CO_2 emissions. IFAC-PapersOnLine, 50(1), 12956–12961, https://doi.org/10.1016/j.ifacol.2017.08.1800

Beale R. and Jackson T. (1990). Neural Computing – An Introduction. Taylor & Francis, UK.

Beheshti Z. and Shamsuddin S. M. (2013). A review of population-based meta-heuristic algorithm. International Journal of Advances in Soft Computing and its Applications, 5, 1–35.

Bello O., Hamam Y. and Djouani K. (2014). Fuzzy dynamic modelling and predictive control of a coagulation chemical dosing unit for water treatment plants. Journal of Electrical Systems and Information Technology, 1(2), 129–143, https://doi.org/10.1016/j.jesit.2014.08.001

Bishop C. M. (1995). Neural Networks for Pattern Recognition. Oxford University Press, Inc., USA.

Biswas W. K. and Yek P. (2016). Improving the carbon footprint of water treatment with renewable energy: a Western Australian case study. Renewables: Wind, Water, and Solar, 3(1), 14, https://doi.org/10.1186/s40807-016-0036-2

Boulesteix A.-L. and Schmid M. (2014). Machine learning versus statistical modeling. Biometrical Journal, 56(4), 588–593, https://doi.org/10.1002/bimj.201300226

Breiman L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

Breiman L. (2001). Random forests. Machine Learning, 45(1), 5–32, https://doi.org/10.1023/A:1010933404324

Burnham K. P. and Anderson D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods and Research, 33(2), 261–304, https://doi.org/10.1177/0049124104268644

Chaudhry S. and Shrier C. (2010). Energy sustainability in the water sector: challenges and opportunities. Proceedings of Annual Conference and Exposition, 2010 Annual Conference and Exposition of the American Water Works Association, Chicago, IL.

Chen C.-L. and Hou P.-L. (2006). Fuzzy model identification and control system design for coagulation chemical dosing of potable water. Water Science and Technology: Water Supply, 6(3), 97–104, https://doi.org/10.2166/ws.2006.782

Cherchi C., Badruzzaman M., Oppenheimer J., Bros C. M. and Jacangelo J. G. (2015). Energy and water quality management systems for water utility's operations: a review. Journal of Environmental Management, 153, 108–120, https://doi.org/10.1016/j.jenvman.2015.01.051

Corominas L., Garrido-Baserba M., Villez K., Olsson G., Cortés U. and Poch M. (2018). Transforming data into knowledge for improved wastewater treatment operation: a critical review of techniques. Environmental Modelling and Software, 106, 89–103, https://doi.org/10.1016/j.envsoft.2017.11.023

Dellana S. A. and West D. (2009). Predictive modeling for wastewater applications: linear and nonlinear approaches. Environmental Modelling and Software, 24(1), 96–106,

https://doi.org/10.1016/j.envsoft.2008.06.002

Du X., Wang J., Jegatheesan V. and Shi G. (2018). Dissolved oxygen control in activated sludge process using a neural network-based adaptive PID algorithm. Applied Sciences, 8(2), 261, https://doi.org/10.3390/app8020261

Eberhart R. and Kennedy J. (1995). A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, IEEE, Nagoya, Japan, pp. 39–43.

Ferrer J., Rodrigo M. A., Seco A. and Penya-Roja J. M. (1998). Energy saving in the aeration process by fuzzy logic control. Water Science and Technology, 38(3), 209–217, https://doi.org/10.2166/wst.1998.0210

Fiter M., Güell D., Comas J., Colprim J., Poch M. and Rodríguez-Roda I. (2005). Energy saving in a wastewater treatment process: an application of fuzzy logic control. Environmental Technology, 26(11), 1263–1270, https://doi.org/10.1080/09593332608618596

Flores-Alsina X., Corominas L., Snip L. and Vanrolleghem P. A. (2011). Including greenhouse gas emissions during benchmarking of wastewater treatment plant control strategies. Water Research, 45(16), 4700–4710, https://doi.org/10.1016/j.watres.2011.04.040

Freund Y. and Schapire R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139, https://doi.org/10.1006/jcss.1997.1504

Friedman J. H. (1991). Multivariate adaptive regression splines. Annals of Statistics, 19(1), 1–67.

Hering A. S. (2021). Fault isolation, chapter 3. In: Statistical Process Monitoring Using Advanced Data-Driven and Deep Learning Approaches, F. Harrou, Y. Sun, A. S. Hering, M. Madakyaru and A. Dairi (eds), Elsevier, The Netherlands, pp. 71–117, https://doi.org/10.1016/B978-0-12-819365-5.00009-7.

Hassan R., Cohanim B., De Weck O. and Venter G. (2005). A comparison of particle swarm optimization and the genetic algorithm. 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Austin, Texas. American Institute of Aeronautics and Astronautics.

Hastie T. and Tibshirani R. (1999). Generalized Additive Models. Chapman & Hall/CRC, Boca Raton, FL, USA.

Hastie T., Tibshirani R. and Wainwright M. (2015). Statistical Learning with Sparsity: The Lasso and Generalizations. Routledge, United Kingdom.

Helsel D. and Hirsch R. (2002). Statistical methods in water resources. In: Techniques of Water-Resources

Investigations of the United States Geological Survey. Hydrologic Analysis and Interpretation, USGS, Reston, VA, pp. 209–218. Available at https://pubs.er.usgs.gov/publication/twri04A3

Holubar P., Zani L., Hager M., Fröschl W., Radak Z. and Braun R. (2003). Start-up and recovery of a biogas-reactor using a hierarchical neural network-based control tool. Journal of Chemical Technology and Biotechnology, 78(8), 847–854, https://doi.org/10.1002/jctb.854

Huang M., Han W., Wan J., Ma Y. and Chen X. (2016). Multi-objective optimisation for design and operation of anaerobic digestion using GA-ANN and NSGA-II. Journal of Chemical Technology and Biotechnology, 91(1), 226–233, https://doi.org/10.1002/jctb.4568

Hüllermeier E. (2015). From knowledge-based to data-driven fuzzy modeling. Informatik-Spektrum, 38(6), 500–509, https://doi.org/10.1007/s00287-015-0931-8

Jackson J. E. (1991). A User's Guide To Principal Components. John Wiley & Sons, Inc.

James G., Witten D., Hastie T. and Tibshirani R. (2013). An Introduction to Statistical Learning. Springer, New York, NY, USA.

Jang J.-S. (1993). ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on

Systems, Man, and Cybernetics, 23(3), 665–685, https://doi.org/10.1109/21.256541

Jang J.-S. R., Sun C.-T. and Mizutani E. (1997). Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Prentice Hall, Upper Saddle River, NJ, USA.

Jeppsson U., Pons M. N., Nopens I., Alex J., Copp J. B., Gernaey K. V., Rosen C., Steyer J. P. and Vanrolleghem P. A. (2007). Benchmark simulation model no 2: general protocol and exploratory case studies. Water Science Technology, 56(8), 67–78, https://doi.org/10.2166/wst.2007.604

John G. H., Kohavi R. and Pfleger K. (1994). Irrelevant features and the subset selection problem. In: Machine Learning Proceedings of the Eleventh International Conference, W. W. Cohen and H. Hirsh (eds), Morgan Kaufmann Publishers, San Francisco, CA, pp. 121–129.

Kambalimath S. and Deka P. C. (2020). A basic review of fuzzy logic applications in hydrology and water resources. Applied Water Science, 10(8), 191, https://doi.org/10.1007/s13201-020-01276-2

Kasabov N. K. (1996). Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. MIT Press, Cambridge, MA, USA.

Kazor K., Holloway R. W., Cath T. Y. and Hering A. S. (2016). Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility. Stochastic Environmental Research and Risk Assessment, 30(5), 1527–1544, https://doi.org/10.1007/s00477-016-1246-2

Kebir F. O., Demirci M., Karaaslan M., ünal E., Dincer F. and Arat H. T. (2014). Smart grid on energy efficiency application for wastewater treatment. Environmental Progress & Sustainable Energy, 33(2), 556–563, https://doi.org/10.1002/ep.11821

Khataee A. R. and Kasiri M. B. (2011). Modeling of biological water and wastewater treatment processes using artificial neural networks. CLEAN – Soil, Air, Water, 39(8), 742–749, https://doi.org/10.1002/clen.201000234

Kira K. and Rendell L. A. (1992). A practical approach to feature selection. In D. Sleeman and P. Edwards (eds), Machine learning proceedings 1992. Morgan Kaufmann, pp. 249–256. Available at https://doi.org/10.1016/B978-1-55860-247-2.50037-1

Kirkpatrick S., Gelatt C. D. and Vecchi M. P. (1983). Optimization by simulated annealing. Science, 220(4598), 671–680, https://doi.org/10.1126/science.220.4598.671

Klanderman M., Newhart K. B., Cath T. and Hering A. S. 2020 Case studies in real-time fault isolation in a decentralized wastewater treatment facility. Journal of Water Process Engineering. 38(2020) 101556, https://doi.org/10.1016/j.jwpe.2020.101556

Kuha J. (2004). AIC and BIC: comparisons of assumptions and performance. Sociological Methods & Research, 33(2), 188–229, https://doi.org/10.1177/0049124103262065

Kurek W. and Ostfeld A. (2013). Multi-objective optimization of water quality, pumps operation, and storage sizing of water distribution systems. Journal of Environmental Management, 115, 189–197, https://doi.org/10.1016/j.jenvman.2012.11.030

Le L. T., Nguyen H., Dou J. and Zhou J. (2019). A comparative study of PSO-ANN, GA-ANN, ICA-ANN, and ABCANN in estimating the heating load of buildings' energy efficiency for smart city planning. Applied Sciences, 9(13), 2630, https://doi.org/10.3390/app9132630

Levstek T. and Lakota M. (2010). The use of artificial neural networks for compounds prediction in biogas from anaerobic digestion – a review. Agricultura, 7, 15–22.

Lindtner S., Schaar H. and Kroiss H. (2008). Benchmarking of large municipal wastewater treatment plants treating over 100 000 PE in Austria. Water Science and Technology, 57(10), 1487–1493, https://doi.org/10.2166/wst.2008.214

Ljung L. (2010). Perspectives on system identification. Annual Reviews in Control, 34(1), 1–12, https://doi.org/10.1016/j.arcontrol.2009.12.001

Lundberg S. M. and Lee S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V.

Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds), Advances in neural information processing systems. Curran Associates, Inc., Vol. 30, pp. 4765–4774. Available at https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Maier R. M., Pepper I. L. and Gerba C. P. (2009) Environmental microbiology. Academic press. Cambridge, Massachusetts, vol. 397.

Maleki A., Nasseri S., Aminabad M. S. and Hadi M. (2018). Comparison of ARIMA and NNAR models for forecasting water treatment plant's influent characteristics. KSCE Journal of Civil Engineering, 22(9), 3233–3245, https://doi.org/10.1007/s12205-018-1195-z

Manesis S., Sapidis D. and King R. (1998). Intelligent control of wastewater treatment plants. Artificial Intelligence in Engineering, 12(3), 275–281, https://doi.org/10.1016/S0954-1810(97)10002-4

Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H. and Teller E. (1953). Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21(6), 1087–1092, https://doi.org/10.1063/1.1699114

Newhart K. B., Holloway R. W., Hering A. S. and Cath T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: a review. Water Research, 157, 498–513, https://doi.org/10.1016/j.watres.2019.03.030

Newhart K. B., Marks C. A., Rauch-Williams T., Cath T. Y. and Hering A. S. (2020). Hybrid statistical-machine learning ammonia forecasting in continuous activated sludge treatment for improved process control. Journal of Water Process Engineering, 37, 101389, https://doi.org/10.1016/j.jwpe.2020.101389

Newhart K. B., Goldman-Torres J. E., Freedman D. E., Wisdom K. B., Hering A. S. and Cath T. Y. (2021). Prediction of peracetic acid disinfection performance for secondary municipal wastewater treatment using artificial neural networks. ACS ES&T Water, 1(2), 328–338, https://doi.org/10.1021/acsestwater.0c00095

Nielsen M. A. (2015). Neural Networks and Deep Learning. Determination Press, San Francisco, CA, USA.

Nybo P. J., Kallesøe C. S. and Lauridsen K. G. (2017). Method for Operating A Wastewater Pumping Station. US Patent No. 9 719 241. US Patent and Trademark Office, Alexandria, VA, USA.

Odom G. J., Newhart K. B., Cath T. Y. and Hering A. S. (2018). Multistate multivariate statistical process control. Applied Stochastic Models in Business and Industry, 34(6), 880–892, https://doi.org/10.1002/asmb.2333

Olsson G. (2006). Instrumentation, control and automation in the water industry – state-of-the-art and new challenges. Water Science and Technology, 53(4–5), 1–16, https://doi.org/10.2166/wst.2006.097

Oppong G., Montague G. A., O'Brien M., McEwan M. and Martin E. B. (2013). Towards advanced control for anaerobic digesters: volatile solids inferential sensor. Water Practice and Technology, 8(1) 7–17, https://doi.

org/10.2166/wpt.2013.002

Özcan F., Atiş C. D., Karahan O., Uncuoğlu E. and Tanyildizi H. (2009). Comparison of artificial neural network and fuzzy logic models for prediction of long-term compressive strength of silica fume concrete. Advances in Engineering Software, 40(9), 856–863, https://doi.org/10.1016/j.advengsoft.2009.01.005

Perendeci A., Arslan S., Tanyola? A. and Çelebi S. S. (2009). Effects of phase vector and history extension on prediction power of adaptive-network based fuzzy inference system (ANFIS)

model for a real scale anaerobic wastewater treatment plant operating under unsteady state. Bioresource Technology, 100(20), 4579–4587, https://doi.org/10.1016/j.biortech.2009.04.049

Polit M., Estaben M. and Labat P. (2002). A fuzzy model for an anaerobic digester, comparison with experimental results. Engineering Applications of Artificial Intelligence, 15(5), 385–390, https://doi.org/10.1016/S0952-1976 (02)00091-X

Ross T. J. (2010). Properties of Membership Functions, Fuzzification, and Defuzzification. In Fuzzy Logic with Engineering Applications. John Wiley & Sons Ltd., pp. 89–116. https://doi.org/10.1002/9781119994374.ch4

Prakash A., Shukla N., Shankar R. and Tiwari M. K. (2008). Solving machine loading problem of FMS: an artificial intelligence (AI) based random search optimization approach. In: Handbook of Computational Intelligence in Manufacturing and Production Management, D. Laha and P. Mandal (eds), IGI Global,Hershey, Pennsylvania, USA, pp. 19–43.

Puig S., Corominas L., Traore A., Colomer J., Balaguer M. D. and Colprim J. (2006). An on-line optimisation of a SBR cycle for carbon and nitrogen removal based on on-line pH and OUR: the role of dissolved oxygen control. Water Science and Technology, 53(4–5), 171–178, https://doi.org/10.2166/wst.2006.121

Reeves C. and Rowe J. E. (2002). Genetic Algorithms: Principles and Perspectives: A Guide to GA Theory. Springer Science & Business Media. Boston, MA, USA.

Ren Z., Liner B., Ferguson C., Fisher A., Newhart K., Wang M. and Sharpless C. (2020). Report on NSF Mid-scale Research Infrastructure Workshop for Intelligent Water Systems. NSF Award #CBET 2035032 (online).

Sadatiyan Abkenar S. M., Stanley S. D., Miller C. J., Chase D. V. and McElmurry S. P. (2015). Evaluation of genetic algorithms using discrete and continuous methods for pump optimization of water distribution systems.

Sustainable Computing: Informatics and Systems, 8, 18–23, https://doi.org/10.1016/j.suscom.2014.09.003

Sadollah A. (2018). Which membership function is appropriate in fuzzy system? In: Fuzzy Logic Based in Optimization Methods and Control Systems and its Applications, A. Sadollah (ed.), InTech, London, UK, pp. 3–6.

Schmidhuber J. (2015). Deep learning in neural networks: an overview. Neural Networks, 61, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003

Schwarz G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464, https://doi.org/10.1214/aos/1176344136

Shankar V. K. A., Umashankar S., Paramasivam S. and Hanigovszki N. (2016). A comprehensive review on energy efficiency enhancement initiatives in centrifugal pumping system. Applied Energy, 181, 495–513, https://doi.org/10.1016/j.apenergy.2016.08.070

Shapley L. S. (1951). Notes on the N-Person Game–II: The Value of an N-Person Game. Rand Corporation, Santa Monica, California, USA.

Sharma S., Sharma S. and Athaiya A. (2020). Activation functions in neural networks. International Journal of Engineering Applied Sciences and Technology, 04(12), 310–316, https://doi.org/10.33564/IJEAST.2020.v04i12.054

Shi C. Y. (2011). Mass Flow and Energy Efficiency of Municipal Wastewater Treatment Plants. IWA Publishing, London, UK.

Steinberg D. and Colla P. (1995). CART: tree-structured non-parametric data analysis. San Diego, CA: Salford Systems.

Sutton C. D. (2005). Classification and regression trees, bagging, and boosting. In: Handbook of Statistics. Data Mining and Data Visualization, C. R. Rao, E. J. Wegman and J. L. Solka (eds), Elsevier, The Netherlands, pp. 303–329.

Takagi T. and Sugeno M. (1983). Derivation of fuzzy control rules from human operator's control actions. IFAC Proceedings Volumes, 16(13), 55–60, https://doi.org/10.1016/S1474-6670(17)62005-6

Tchobanoglous G., Stensel H. D., Tsuchihashi R., Burton F. L. and Abu-Orf M. (2014). Fundamentals of chemical coagulation. In: Wastewater Engineering: Treatment and Resource Recovery, G. Bowden, W. Pfrang and Metcalf & Eddy (eds), McGraw-Hill Education, New York, NY, USA, pp. 460–473.

Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

Tibshirani R., Saunders M., Rosset S., Zhu J. and Knight K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1), 91–108, https://doi.org/10.1111/j.1467-9868.2005.00490.x

Torregrossa D., Hansen J., Hernández-Sancho F., Cornelissen A., Schutz G. and Leopold U. (2017). A data-driven methodology to support pump performance analysis and energy efficiency optimization in waste water treatment plants. Applied Energy, 208, 1430–1440, https://doi.org/10.1016/j.apenergy.2017.09.012

Traoré A., Grieu S., Puig S., Corominas L., Thiery F., Polit M. and Colprim J. (2005). Fuzzy control of dissolved oxygen in a sequencing batch reactor pilot plant. Chemical Engineering Journal, 111(1), 13–19, https://doi.org/10.1016/j.cej.2005.05.004

Trosset M. W. (2001). What is simulated annealing? Optimization and Engineering, 2(2), 201–213, https://doi.org/10.1023/A:1013193211174

Turkdogan-Aydınol F. I. and Yetilmezsoy K. (2010). A fuzzy-logic-based model to predict biogas and methane production rates in a pilot-scale mesophilic UASB reactor treating molasses wastewater. Journal of Hazardous Materials, 182(1), 460–471, https://doi.org/10.1016/j.jhazmat.2010.06.054

Van Leeuwen J., Chow C. W. K., Bursill D. and Drikas M. (1999). Empirical mathematical models and artificial neural networks for the determination of alum doses for treatment of southern Australian surface waters. Aqua, 48(3), 115–127.

Vrieze S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological Methods, 17(2), 228–243, https://doi.org/10.1037/a0027127

Wallace J., Champagne P. and Hall G. (2016). Multivariate statistical analysis of water chemistry conditions in three wastewater stabilization ponds with algae blooms and pH fluctuations. Water Research, 96, 155–165, https://doi.org/10.1016/j.watres.2016.03.046

Wang X., Ratnaweera H., Holm J. A. and Olsbu V. (2017). Statistical monitoring and dynamic simulation of a wastewater treatment plant: a combined approach to achieve model predictive control. Journal of Environmental Management, 193, 1–7, https://doi.org/10.1016/j.jenvman.2017.01.079

Wise B. M. and Gallagher N. B. (1996). The process chemometrics approach to process monitoring and fault detection. Journal of Process Control, 6(6), 329–348, https://doi.org/10.1016/0959-1524(96)00009-1

Wu G.-D. and Lo S.-L. (2008). Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. Engineering Applications of Artificial Intelligence, 21(8), 1189–1195, https://doi.org/10.1016/j.engappai.2008.03.015

Wu G.-D. and Lo S.-L. (2010). Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial neural network. Expert Systems with Applications, 37(7), 4974–4983, https://doi.org/10.1016/j.eswa.2009.12.016

Yuan M. and Lin Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49–67, https://doi.org/10.1111/ j.1467-9868. 2005.00532.x

Zadeh L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man and Cybernetics, 1100, 38–45.

Zangooei H., Delnavaz M. and Asadollahfardi G. (2016). Prediction of coagulation and flocculation processes using ANN models and fuzzy regression. Water Science and Technology, 74(6), 1296–1311, https://doi.org/10.2166/wst.2016.315

Zhang Z., Zeng Y. and Kusiak A. (2012). Minimizing pump energy in a wastewater processing plant. Energy, 47(1), 505–514, https://doi.org/10.1016/j.energy.2012.08.048

Zheng F. and Zhong S. (2011). Time series forecasting using a hybrid RBF neural network and AR model based on binomial smoothing. World Academy of Science, Engineering and Technology, 75, 1471–1475.

Zou H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418–1429, https://doi.org/10.1198/016214506000000735