

**CECS ×××:20××**

中国工程建设标准化协会标准

**城市供水监管中大数据应用技术指南**

**Guidelines for big data application technology in supervision of urban water supply system**

（征求意见稿）

**中国工程建设标准化协会标准**

城市供水监管中大数据应用技术指南

**Guidelines for big data application technology in supervision of urban water supply system**

**CECS ×××:20××**

主编单位：中国城市规划设计研究院

批准单位：中国工程建设标准化协会

施行日期：

# 前言

城市供水安全直接关系到广大人民群众的切身利益和身体健康，供水安全监管是我国城市供水行业普遍面临的压力。近年来，各地在相关政策法规的指导下开展了饮用水水量、水质、水压等监测能力的建设，具备了一定的供水安全监管能力。但一方面当前供水监管中存在实测指标不全面、监测频率不达标、风险预警不及时等问题，另一方面大中城市供水企业每日积累的大量供水数据所蕴含的巨大价值几乎尚未被挖掘。随着我国社会经济的快速发展，公众对供水安全的要求越来越高，行业主管部门面临着进一步提高供水监管能力的需求。因此，保障饮用水供水安全作为全面建成小康社会的一项重要任务，必须采取有效措施，推动先进的信息技术在供水行业的应用，引导城市供水监管向更科学、更高效、更精细的方向发展

为开拓大数据技术在城市供水监管领域的应用途径，完善供水安全监管技术工作体系，进一步提高城市供水的数字化监管水平，编制组编制了《城市供水监管中大数据应用技术指南》。

本指南编制依据国家和行业相关法律法规、标准规范，总结了近年来我国城市供水行业监管需求、信息化系统建设现状和大数据技术应用的实践经验，同时在编制过程中对主要问题开展了专题论证，对具体内容进行了反复讨论和修改。

本指南的主要内容包括：总则、供水大数据来源、平台架构与分析方法、水源和水厂大数据应用、供水管网运营大数据应用、公众反馈信息大数据应用。

本指南由中国工程建设标准化协会建标协会负责管理，由中国城市规划设计研究院负责具体技术内容的解释。请各单位在使用过程中，总结实践经验，提出意见和建议。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 主编单位： | 中国城市规划设计研究院 | | | |
| 参编单位： | （以下排名不分先后） | | | |
| 深圳水务（集团）有限公司 | | | |
| 济南水务集团有限公司 | | | |
| 北京首创股份有限公司 | | | |
| 山东省城市供排水水质监测中心 | | | |
| 中国航天科工三院三零四所 | | | |
| 主要起草人： | XXX |  |  |  |

目 录

[前言 I](#_Toc23522845)

[第一章 总 则 1](#_Toc23522846)

[第一节 编制目的 1](#_Toc23522847)

[第二节 适用范围 1](#_Toc23522848)

[第三节 指导思想 1](#_Toc23522849)

[第四节 供水大数据积累基本原则 1](#_Toc23522850)

[第五节 供水大数据应用基本原则 1](#_Toc23522851)

[第六节 术语与定义 2](#_Toc23522852)

[第二章 供水大数据来源 4](#_Toc23522860)

[第一节 内部来源 4](#_Toc23522861)

[第二节 外部来源 5](#_Toc23522864)

[第三章 平台架构与分析方法 9](#_Toc23522867)

[第一节 供水大数据分析平台架构 9](#_Toc23522868)

[第二节 数据挖掘方法 9](#_Toc23522869)

[第四章 水源和水厂大数据应用 11](#_Toc23522870)

[第一节 水源和水厂大数据应用目的 11](#_Toc23522871)

[第二节 水源和水厂大数据积累要求 11](#_Toc23522872)

[第三节 水源和水厂大数据典型应用场景 12](#_Toc23522876)

[第四节 水源和水厂大数据应用典型案例 15](#_Toc23522880)

[第五章 供水管网运营大数据应用 26](#_Toc23522884)

[第一节 管网运营大数据应用目的 26](#_Toc23522885)

[第二节 管网运营大数据积累要求 26](#_Toc23522886)

[第三节 管网运营大数据典型应用场景 27](#_Toc23522892)

[第四节 供水管网运营风险评估典型案例 29](#_Toc23522895)

[第六章 公众反馈信息大数据应用 33](#_Toc23522896)

[第一节 公众反馈信息大数据应用目的 33](#_Toc23522897)

[第二节 公众反馈信息大数据积累要求 33](#_Toc23522898)

[第三节 公众反馈信息大数据典型应用场景 34](#_Toc23522903)

[第四节 公众反馈信息大数据应用典型案例 36](#_Toc23522906)

[附录 42](#_Toc23522909)

# 总 则

## 编制目的

为贯彻执行《城市供水条例》，提高供水监管信息的价值挖掘效率，开拓供水大数据在水质风险识别、水量风险识别、设备安全动态监管、公众反馈问题处理等领域的辅助决策应用途径，进一步提高城市供水的数字化监管水平，根据城市水源、水厂、管网运营和客户服务等环节的供水安全监管需求，依据国家和行业相关法律法规和标准规范，编制本指南。

## 适用范围

本指南明确了供水业务中的大数据选择原则，以及常用应收集的数据类型及其应用条件。本指南适用于城镇供水主管部门、城镇供水单位及二次供水管理单位对供水业务中水质、管网运营、供水服务的风险预警和监管，城镇供水单位的工艺运行辅助决策也可参照本指南执行。

## 指导思想

本指南的指导思想是针对国内供水安全监管的实际需求，密切结合国情和供水行业技术现状，以供水行业中大数据技术应用行之有效的实践经验和科技综合成果为依据，按照协调性、系统性、实用性、先进性的原则，科学确定供水大数据积累途径、应用领域和应用技术路线要求。

## 供水大数据积累基本原则

涵盖供水服务从水源、水厂、管网到客户服务信息反馈全流程的大数据。

## 供水大数据应用基本原则

采用先进的大数据分析方法和模型开发技术，技术方法先进、适用、引领发展，应用途径符合法律和行政法规的规定，符合供水监管需求，契合供水监管更全面、更精细、更安全的发展方向。

## 术语与定义

### 供水监管

以提高监测能力、预报水平和减灾水平为目的，从水源到水龙头对供水系统的全面质量监管，包括：城市供水水质、水量和水压安全的监测、评估和控制；城市供水水质污染风险源的识别、监测与预警；现代化、智能化、定量化评估的供水系统管理辅助决策等。

### 城镇供水单位

从事城镇公共供水和自建设施供水专门服务的单位。

### 二次供水管理单位

受二次供水设施产权单位（人）委托，对二次供水设施进行日常维护和运行管理的单位。

### 供水大数据

供水活动从水源到龙头全过程所产生的具备体量巨大、来源多样、生成较快、且多变等特征的大量数据，包括水质检测、生产工艺、设备运营等环节的相关数据。

### 机器学习

通过利用数据或以往的经验训练出计算机模型的一种分析方法，模型可通过获取分析新数据而不断完善自身性能，可使用模型开展预测、决策等用途。

### 水质风险

相关水质指标超出标准限值，水质安全性受到威胁，存在损害人体健康的风险。

### 管网风险

指供水管网在特定的时间和供水条件要求下，不能够稳定、持续完成供水功能，引起不良后果的风险。

# 供水大数据来源

在开展供水大数据应用前，城镇供水相关单位应注意按时按质收集供水数据，保证数据的准确性、完善性和及时性，从而保障分析结果的科学性、有效性。随着信息化系统的不断完善、数据质量的不断提升，供水服务企业及相关单位应定期更新所需数据，不断提高数据质量，保证数据时效性。

## 内部来源

### 统计报表数据

城镇供水主管部门和城镇供水单位应定期收集各类统计报表数据。

一是根据相关标准和规范要求，由城镇供水单位、水质检测机构等单位按期检测的水质数据：按照《生活饮用水卫生标准》(GB5749-2006)、《城市供水水质标准》（CJT206-2006）、《城镇供水厂运行、维护及安全技术规程》等相关标准规范中要求的水质检测指标、检测方法和检测频率，对水源水、水厂各工艺段进出水、出厂水、管网水、龙头水等制水和输配水环节水质的检测数据。二是包括城镇供水单位的设备工况、材料库存、售水情况、用水户信息、管网信息、设备维护检修记录等相关生产数据。

### 实时在线传输数据

城镇供水单位应实时收集获取在制水和输配水现场，通过现场传感器或移动设备开展连续监测而创建或生成的远程数据。

一是包括在水源地自动在线监测采集的水质、水量、水位等实时数据，在水厂各工艺环节自动在线监测采集的水质、水位、流量等实时数据，在管网监测点自动在线监测采集的水质、水压、流量、噪声等实时运营数据。二是企业员工通过移动设备人为实时远传的地理位置、用户水量、事故特征、现场照片、视频等外勤作业信息数据。

## 外部来源

### 获取途径

除了部门内部数据，城镇供水主管部门和城镇供水单位还应加强与系统内其他供水主管部门和供水企业，以及环保、水利等相关部门的信息共享。

通过运用先进的网络爬虫等技术，可获取政府机构、企业等组织提供的与供水服务有关的免费开放数据，包括供水水质数据、供水事故信息、人口数据、建筑信息数据等。采用网络爬虫技术时，需明确供水服务相关数据获取需求，并以此设计完善系统总体架构，确认待获取信息量级以及待爬取页面类型，选择爬虫类型以及爬虫策略。对于爬取供水大数据，推荐采用分布式爬虫架构。一般情况下，开展网络爬虫的流程图如下：

明确信息

获取需求

设计爬虫策略

设计爬虫

总体架构

确认待获取信息

选择爬虫类型

爬取数据存储

爬虫结果展示

### 网络爬虫获取水质大数据典型案例

某供水主管部门拟运用分布式爬虫以及科学的数据分析模型，通过获取和分析多个城市的水质公开信息，实现水质大数据信息的收集和展示，提高供水业务监管能力和水平。

供水主管部门共收集了全国600多个城市关于水质公开的渠道共计2000余个，需要从此2000多个网址中爬取有关水质公开的有效信息。以此2000多个网址为起始地址向下探索，以此延伸开来的链接数据将是一个庞大的数量级，因此采用分布式的爬虫来完成所有的爬取任务。

分布式爬虫总体结构如图2-1所示：

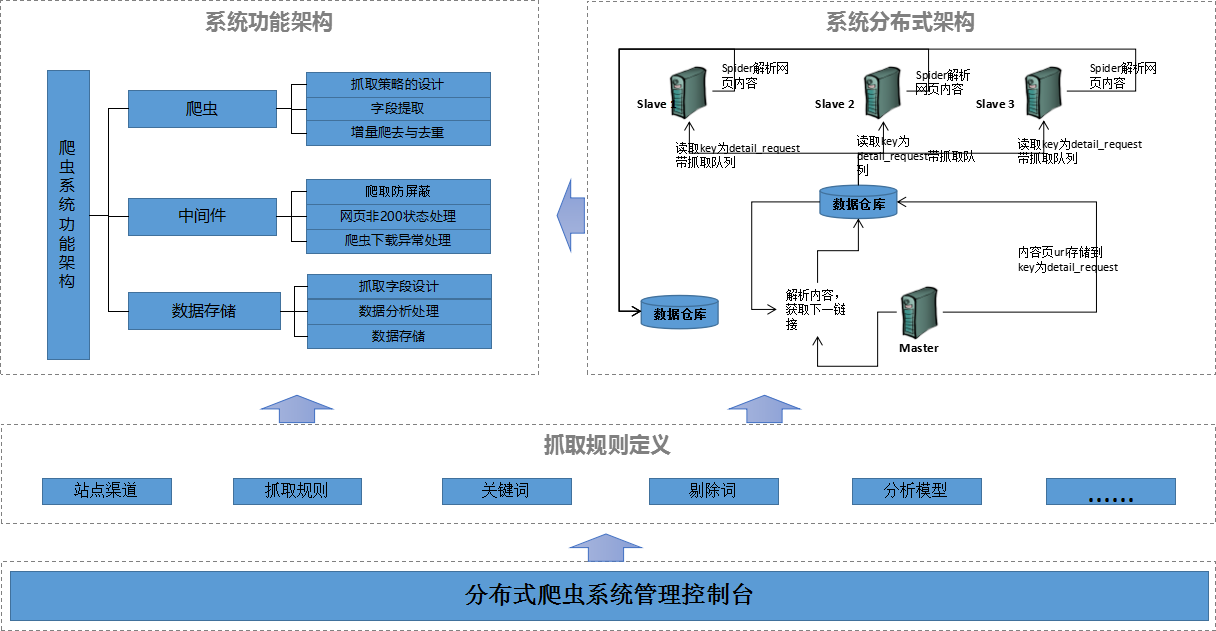


图2-1 分布式爬虫总体结构

对供水水质大数据的分布式爬虫总体结构包括分布式爬虫管理控制台，爬虫系统功能架构和系统分布式架构。分布式爬虫管理控制台是爬虫任务调度端；爬虫系统功能架构是爬虫任务执行端；系统分布式架构是数据服务端。

供水水质大数据的爬虫工作流程为：网络爬虫拥有一个控制中心，负责对整个爬虫系统进行管理和监控，主要包括控制用户交互、初始化爬虫器、确定主题、协调各模块工作、控制爬行过程等；然后，将初始的URL（Uniform Resource Locator，统一资源定位符，也被称为网页地址）集合传递给URL队列，页面爬行模块会从URL队列中读取第一批URL列表，然后根据这些URL地址从互联网中进行相应的页面爬取。爬取后，将爬取到的内容根据主题保存到数据库中，同时，在爬行过程中，会爬取到一些新的URL。此时，需要根据我们所定的主题使用链接过滤器过滤掉无效的链接，再将剩下的URL地址传递到队列中，供页面爬行模块使用。另一方面，将页面爬取模块爬取到的网页内容根据主题交由不同页面分析模块处理，并根据处理结果建立索引数据库，用户检索对应信息时，可以从数据库中直接查找对应的索引，并得到相应的结果。供水水质大数据爬虫工作流程图如图2-2所示。



图2-2 供水水质大数据爬虫工作流程图

爬虫过程中，需要对网页内容进行解析。由于网页内容解析需要解析来自不同网页的内容，目前尚无法实现通过通用的方法来解析网页内容，需通过解析网页文档Head（头节点）标签中的Title（标题）、HTML（超文本标记语言）文档Meta（元节点）标签中的Keywords（关键词）、Description（描述）来获取有效信息。同时为了更精准地捕获有效的信息，需要对爬取的网页进行多个层次的筛选工作。第一层筛选，在Redis（Remote Dictionary Server，远程字典服务）中维护了一个剔除词的集合，如果爬取的信息中包含剔除词则放弃解析该网页。第二层筛选，在Redis中维护了判断水质信息的关键词集合，根据这个集合来判断该网页是否是水质公开的网页。第三层筛选，为了在Redis中维护了四个分类的集合，依次来判断公开水质的所属类别，将水质公开信息按照水源水、出厂水、管网水、龙头水四类进行分类。在对爬虫抓取的内容进行解析和数据清洗后，再传送到数据库存储。

采用此技术爬虫获取了全国32个省份的水质信息公开数据，开发了信息平台对采用爬虫技术获取的供水水质大数据结果进行展示，如图2-3所示。该信息平台展示了全国各省市各类水质公开信息累计公开次数，可针对各城市的水源水、出厂水、管网水、二次供水情况等信息分类查询，并提供数据导出功能。

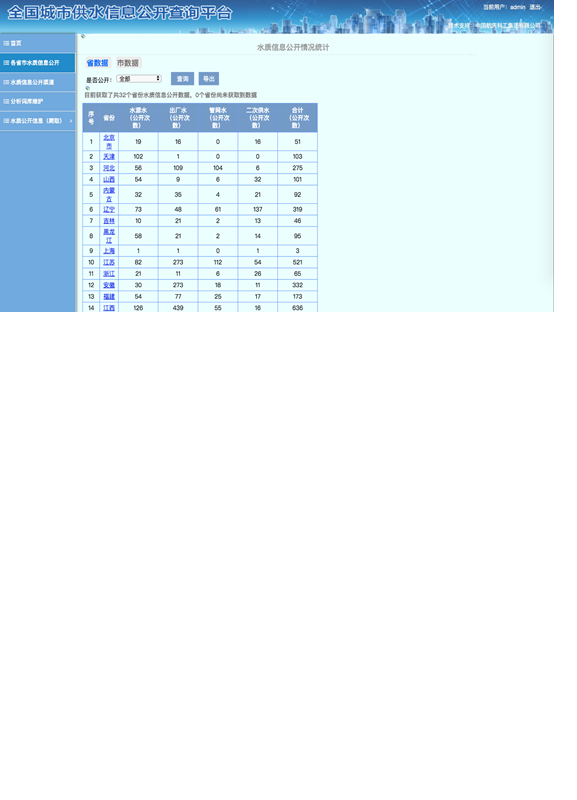


图2-3 供水水质大数据爬虫结果展示页面

# 平台架构与分析方法

## 供水大数据分析平台架构

开展供水大数据分析应用时，需根据应用、设计和保密等需求，构建供水大数据分析平台。供水大数据分析平台应由基础设施层、平台层、数据应用层（各类应用模型）和展示层等架构组成。

基础设施层主要包括存储设备、计算设备、网络设备、机房配套等相关软硬件基础设施。

平台层主要提供包括数据采集、传输、分布式存储、挖掘处理运算等功能，其中数据采集由物联感知信息采集系统、数据共享系统、网络爬虫系统等组成；数据传输是将物联感知、数据填报共享、网络爬虫等途径产生的数据，通过互联网、移动网络、卫星、超短波电台等渠道进行传输；数据存储是将各类数据信息进行存储与相应数据库中，实现对供水大数据的管理、调用、共享交换和更新，为应用层对数据的使用提供调取服务，对于结构化数据通过清洗、转换、集成处理后，加载到数据仓库中，对于非结构化数据，包括供水业务相关的图片、音视频、文档文件等，采用分布式存储的方式。

数据应用层包括服务于供水监管的各类业务系统，例如日常监管、实时监控、安全评估、监测预警、决策支持等业务系统，各业务系统通过清晰的工作流程、灵活的统计分析方法、统一的用户认证、标准化的海量数据，实现业务功能。

展示层将应用层的分析处理结果，通过文字、图表、报表等表现方式，对用户和公众实现页面可视化展现。

对供水大数据的采集、存储和应用，可参考《城市供水信息系统 基础信息加工处理技术指南》开展。

## 数据挖掘方法

供水大数据的数据特征挖掘是从海量的、关系模糊的供水数据中挖掘出潜在的数据规律，挖掘数据特征是开展供水大数据价值应用的重要基础性工作。机器学习是发掘、利用大数据价值的关键技术，机器学习按照大类可分为监督学习、无监督学习和半监督学习三大类。

监督学习利用已知类别的样本数据调整运算参数，使其达到所要求性能，发现和总结数据间的关联关系，并产生推断功能，从而可利用模型对于给定输入变量，对输出数据进行较准确的预测。监督学习主要包括分类和回归分析两种，目前，典型的有监督学习算法包括逻辑回归、决策树、贝叶斯判别、随机森林、支持向量机，以及目前在人工智能中被大量应用的神经网络等。

无监督学习的数据没有标注信息，其利用已有的样本数据，实现对样本聚类。无监督学习主要包括聚类分析、关联分析等。关联分析算法包括Apriori、FP-growth等，聚类分析算法包括K均值聚类、系谱聚类、密度聚类等。

半监督学习是监督学习与无监督学习相结合的一种学习方法。半监督学习同时使用大量的未标注数据，以及少量的标注数据。典型的算法包括自训练算法、半监督支持向量机、生成式方法等。

在供水大数据分析过程中，需要根据应用场景和目的选择合适的分析算法，不存在一个能够解决所有问题的分析算法，一种分析算法也并不能总是优于另外一种。供水大数据的数据挖掘方法选择需根据数据量大小、数据结构、数据特征、分析目的来综合比选。

# 水源和水厂大数据应用

## 水源和水厂大数据应用目的

对水源水质、水厂工艺等大数据进行存储、分析和价值挖掘，不仅可解决传统手段难以解决的海量数据问题，还有利于水质风险来源、时空变化规律、危害程度等特点的有效判定和识别，实现对水源和水厂水质风险的有效预警，并提出风险的最优处置方案；同时，通过对水源水质与水厂工艺大数据的联合建模分析，可有利于水厂根据不同的原水水质特点和风险对水处理工艺过程进行调整和优化，以节约水厂运行成本并保障水厂的出水水质。

## 水源和水厂大数据积累要求

城镇自来水供水单位应建立水文数据、水质数据和相应的水处理工艺运行参数数据库。

### 水质数据

城镇供水单位在日常运行中需按照《城镇供水厂运行、维护及安全技术规程》（CJJ58-2007）的要求开展原水、出厂水以及水厂各沉淀、过滤等工序的水质检测，原水以及出厂水日检、月检、年检的项目和频次，应符合《城市供水水质标准》（CJ-T206-2017）和《生活饮用水卫生标准》（GB5749-2006）的规定。此外，供水企业还需结合原水水质和净水工艺以及地方监管的要求确定是否需要增加相应的检测指标和检测频次。以湖泊、水库为水源的供水厂，应开展藻类相关指标的检测并在藻类爆发期增加检测的频次。

城镇供水单位应根据《城镇供水水质在线监测技术标准》（CJJ-T271-2017）对原水、工艺过程出水、出厂水及管网水安装必要的在线监测仪表，并同时开展班组级的手工检测。对实验室和在线监测的水质数据应及时记录并保存。

### 水文及气象数据

城镇供水单位需按周期建立水源地的水文资料台账，包含：流量、流速、水位、潮汐、天气、水温等，掌握气象因素（气温、湿度、日照、降雨等）与原水水质的关联关系以及由自然因素引起的水质突变。

### 工艺数据

城镇供水单位需按日建立工艺运行参数台账，包括：水处理工艺运行参数，例如处理水量、预处理药剂及其投加量、混凝剂及其投加量、沉淀池排泥周期、滤池运行周期及反冲洗参数等；各工序水质，例如沉淀后出水的pH值、浊度、耗氧量、余氯等，过滤后出水的pH值、浊度、余氯等。

## 水源和水厂大数据典型应用场景

### 水质相关分析

城镇供水主管部门及供水单位在水质日常监测、风险预警和管控过程中，通过对水源、水厂、管网和二次供水设施的水质指标及其环境类指标进行相关性分析，找出不同水质指标之间、水质指标与其它环境类指标的内在关联性，可实现水质风险预警，水质相关性分析具体包括以下几种情形：

1）同一供水环节不同水质指标之间的相关性分析：通过常规水质指标与特征水质指标的相关性分析，可揭示指标之间的内在联系，及时发现水质特征和规律，并可为回归分析提供数据参考。

2）不同供水环节同一水质指标之间的相关性分析：通过在供水全流程，对水源、水厂、管网和二次供水设施各环节的同一水质指标开展相关性分析，找出不同供水环节水质指标的关联性，可用于揭示不同环节的水质变化趋势，反馈指导水厂处理工艺调整，并有利于水质风险管控。

3）水质指标与环境类指标之间的相关性分析：通过水质指标与环境类指标的相关性分析，揭示其内在的关联性，可为水质问题溯源、回归分析建模因子选择等提供依据。

水质相关分析过程可采用的技术路线为：

供水水质

数据获取

数据预处理

数据分布

特征分析

选择分析方法/相关系数类别

相关系数计算

显著性检验

水质相关业务判断

### 水质风险预警

以历史水质数据、相关水文及环境类等数据为基础，城镇供水主管部门及供水单位可通过应用各类数据特征挖掘与分析技术和评价技术，构建水质时间序列、回归分析、评价和风险评估模型，以掌握水质指标的趋势性、周期性及其随机性变化规律，识别高风险水质指标和季节性易超标指标等，并对水质指标的未来数值和风险进行预测和预警。

对于已识别的具有显著变化趋势的水质指标、高风险水质指标等，在日常的水质检测或特定的季节时期，需要加密该项指标的检测，同时考虑加装该水质指标的在线监测仪表。同时也需要查找并明确引起该项水质指标超标的原因，并在今后的日常运行管理中进行相关参数的记录并重点关注。对于可预判的水质风险，可提前采取如调配水源、原位净化水体水质、调整水处理工艺等各种措施进行应对。

水质风险预警分析过程可采用的技术路线为：

预警相关指标

数据获取

数据预处理

数据分布

特征分析

模型选择

结果分析与

业务判断

水质风险预警场景/指标选择

### 水厂运行工艺调整辅助决策

城镇供水单位应建立起原水关键水质项目在工艺流程中水质变化情况及去除率情况，并对工艺运行参数如药耗、滤池反冲洗周期、排泥周期等同步记录，从而可基于不同进出水水质条件下的运行工况和水质预警结果反馈调整工艺方案。

水厂运行工艺的辅助决策方法主要可分为三步：水源地及水厂运行多年的常规数据获取、建立数据间的关联性、基于关联数据的模型构建。可采用的技术路线为：

数据预处理

建立数据关联性

辅助决策模型选择、构建及验证

工况输入试算

获取水源地数据

获取水厂数据

获得预警工况

出水水质达标

对应的临界参数

城镇供水单位进行辅助决策建模时，应基于关联大数据特征选择合适的模型，利用编程语言构建“获取信息-处理信息-反馈结果-调整权重”的数据处理流程，通过输入与输出值的对应关系，不断地降低误差梯度至具备拟合实际情况的能力，从而形成辅助决策模型。而后当面临药耗增加、水源地水质突变等相关工艺参数变化问题时，可将以上情况参数作为输入参数，利用辅助决策模型的数据识别和处理能力，输出相应的工艺运行结果，也即原水或工艺参数发生改变后的出水情况，从而避免了人为因素的主观判断性，同时也可预测出达标水质对应的水源地水质预警值及工艺药耗最小值。

## 水源和水厂大数据应用典型案例

### 原水水质指标间的相关性分析案例

叶绿素a能较好地反应水体的富营养化状况，黄河中下游地区XX市引黄水库为扩大筛选可预测水源水体富营养化的水质预警指标，整理了近5年监测的原水水质数据，包括溶解氧、总磷、总氮、氨氮、硝酸盐（以N计）、N/P、水温、pH、浑浊度、叶绿素a等10个水质指标，并对10个水质指标之间的相关性进行了分析。采用的分析步骤为：

1）整理相关水质数据并进行数据预处理。针对离群数据，核对数据的收集和录入过程，或者重复实验，校核数据的有效性，剔除无效或错误数据；

2）开展变量类型或正态性检验，并选择合适的相关系数分析公式；

3）计算各水质指标间的相关系数r，评估相关程度；

4）进行显著性检验，如果显著性系数p<a（a一般取0.05），认为水质指标间存在显著相关性；

5）对水质相关性分析结果进行解读，并给出相应的业务策略或建议。

整理的部分原始水质指标数据见表4-1。

**表4-1 XX市XX水库原水水质数据**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 水温(℃) | pH | 溶解氧（mg/L） | 氨氮（mg/L） | 总磷（mg/L） | 总氮（mg/L） | 硝酸盐（以N计）（mg/L） | 浑浊度（NTU） | 氮磷比 | 叶绿素a（μg/L） |
| 1 | 27.0 | 8.7 | 2.3 | 0.18 | 0.02 | 1.65 | 1.27 | 7.1 | 82.5 | 18.0 |
| 2 | 27.0 | 8.7 | 8.2 | 0.22 | 0.02 | 1.65 | 1.23 | 4.0 | 82.5 | 16.7 |
| 3 | 27.0 | 8.6 | 8.2 | 0.59 | 0.06 | 1.87 | 1.00 | 13.1 | 31.2 | 45.0 |
| 4 | 20.0 | 8.4 | 9.2 | 0.16 | 0.03 | 1.12 | 0.75 | 4.7 | 37.3 | 17.0 |
| 5 | 20.0 | 8.4 | 8.8 | 0.19 | 0.03 | 1.04 | 0.77 | 6.0 | 34.7 | 14.7 |
| 6 | 19.0 | 8.3 | 8.5 | 0.23 | 0.04 | 3.05 | 2.81 | 8.2 | 76.3 | 7.3 |
| 7 | 6.0 | 8.4 | 11.6 | 0.18 | 0.02 | 3.69 | 2.92 | 1.0 | 217.1 | 1.1 |
| 8 | 6.0 | 8.4 | 11.5 | 0.24 | 0.02 | 3.03 | 2.80 | 1.6 | 202.0 | 1.4 |
| 9 | 10.0 | 8.3 | 10.5 | 0.29 | 0.03 | 3.80 | 3.21 | 4.0 | 126.7 | 3.8 |
| 10 | 10.0 | 8.3 | 10.6 | 0.40 | 0.03 | 3.20 | 2.65 | 2.5 | 106.7 | 2.9 |
| 11 | 9.0 | 8.2 | 11.2 | 0.68 | 0.05 | 4.37 | 2.82 | 111.0 | 87.4 | 6.7 |
| 12 | 28.7 | 8.3 | 9.2 | 0.16 | 0.03 | 1.60 | 1.19 | 3.5 | 51.6 | 12.3 |
| 13 | 28.5 | 8.4 | 11.8 | 0.19 | 0.05 | 1.23 | 0.71 | 8.2 | 22.8 | 21.3 |
| 14 | 29.0 | 8.2 | 8.8 | 0.10 | 0.03 | 1.66 | 1.26 | 3.3 | 59.3 | 8.8 |
| 15 | 28.7 | 8.2 | 9.0 | 0.21 | 0.03 | 1.67 | 1.30 | 2.6 | 53.9 | 9.3 |
| 16 | 12.0 | 8.5 | 10.1 | 0.18 | 0.03 | 2.70 | 1.84 | 6.7 | 79.4 | 18.9 |
| 17 | 8.0 | 8.6 | 10.7 | 0.16 | 0.03 | 2.20 | 1.52 | 4.2 | 75.9 | 20.3 |
| 18 | 20.0 | 8.4 | 9.4 | 0.15 | 0.02 | 3.82 | 3.48 | 2.8 | 173.6 | 8.2 |
| 19 | 20.0 | 8.4 | 9.9 | 0.14 | 0.02 | 3.48 | 3.34 | 2.1 | 165.7 | 9.7 |
| 20 | 20.0 | 8.5 | 9.7 | 0.14 | 0.02 | 3.93 | 3.48 | 1.8 | 178.6 | 9.7 |
| … | … | … | … | … | … | … | … | … | … | … |
| N | 9.0 | 8.4 | 8.7 | 0.2 | 0.03 | 2.5 | 1.6 | 4.3 | 121.0 | 14.7 |

采用Kolmogorov-Smirnov（KS）检验查看样本数据是否符合正态分布。KS检验结果显示，样本数据服从正态分布，故采用皮尔森相关系数法进行相关性分析。皮尔森相关系数（P）及显著性系数（t）分析结果见表4-2。

**表4-2 相关性分析及显著性检验**

|  | | 水温 | pH | 溶解氧 | 氨氮 | 总磷 | 总氮 | 硝酸盐 | 浑浊度 | 氮磷比 | 叶绿素a |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 水温 | P | 1 | -.014 | -.622\*\* | -.060 | .083 | -.268\*\* | -.202\* | -.204 | .005 | .238\* |
| t |  | .907 | .000 | .543 | .415 | .007 | .038 | .155 | .959 | .014 |
| pH | P | -.014 | 1 | -.094 | .323\*\* | .240\*\* | -.268\*\* | -.439\*\* | -.042 | -.310\*\* | .518\*\* |
| t | .907 |  | .200 | .000 | .000 | .000 | .000 | .686 | .000 | .000 |
| 溶解氧 | P | -.622\*\* | -.094 | 1 | .002 | .119 | .073 | .094 | .186 | -.106 | -.050 |
| t | .000 | .200 |  | .978 | .078 | .285 | .184 | .100 | .118 | .457 |
| 氨氮 | P | -.060 | .323\*\* | .002 | 1 | .363\*\* | -.142\* | -.187\*\* | .462\*\* | -.219\*\* | .397\*\* |
| t | .543 | .000 | .978 |  | .000 | .021 | .003 | .000 | .000 | .000 |
| 总磷 | P | .083 | .240\*\* | .119 | .363\*\* | 1 | -.209\*\* | -.210\*\* | .306\*\* | -.613\*\* | .488\*\* |
| t | .415 | .000 | .078 | .000 |  | .001 | .001 | .002 | .000 | .000 |
| 总氮 | P | -.268\*\* | -.268\*\* | .073 | -.142\* | -.209\*\* | 1 | .947\*\* | .250\* | .523\*\* | -.550\*\* |
| t | .007 | .000 | .285 | .021 | .001 |  | .000 | .011 | .000 | .000 |
| 硝酸盐 | P | -.202\* | -.439\*\* | .094 | -.187\*\* | -.210\*\* | .947\*\* | 1 | .099 | .509\*\* | -.616\*\* |
| t | .038 | .000 | .184 | .003 | .001 | .000 |  | .316 | .000 | .000 |
| 浑浊度 | P | -.204 | -.042 | .186 | .462\*\* | .306\*\* | .250\* | .099 | 1 | -.070 | .028 |
| t | .155 | .686 | .100 | .000 | .002 | .011 | .316 |  | .488 | .767 |
| 氮磷比 | P | .005 | -.310\*\* | -.106 | -.219\*\* | -.613\*\* | .523\*\* | .509\*\* | -.070 | 1 | -.432\*\* |
| t | .959 | .000 | .118 | .000 | .000 | .000 | .000 | .488 |  | .000 |
| 叶绿素a | P | .238\* | .518\*\* | -.050 | .397\*\* | .488\*\* | -.550\*\* | -.616\*\* | .028 | -.432\*\* | 1 |
| t | .014 | .000 | .457 | .000 | .000 | .000 | .000 | .767 | .000 |  |
| \*\*. 在0.01 水平（双侧）上显著相关。 \*. 在 0.05 水平（双侧）上显著相关。 | | | | | | | | | | | |

由表4-2可知，XX市引黄水库原水中叶绿素a与硝酸盐、总氮、pH、总磷、氮磷比、氨氮浓度在0.01水平上显著相关，相关系数分别为-0.616、-0.550、0.518、0.488、-0.432以及0.397，其中，叶绿素a与pH、总磷、氨氮指标呈正相关，与硝酸盐、总氮、氮磷比指标呈负相关。分析原因可能为，该引黄水库中总氮含量较高，经常超出GB 3838-2002《地表水环境质量标准》中的Ⅴ类限值，且氨氮（氮营养盐的首先消耗）所占总氮比重较小，氮磷比较高，远远大于藻类生长所需的氮磷比，磷成为水中藻类增长最重要的限制因素。因此，初步筛选将硝酸盐、总氮、pH、总磷、氮磷比和氨氮指标纳入预测水源水体富营养化趋势的预警指标。

### 水质风险预警分析案例

XX市XX水库原水的高锰酸盐指数超标风险较高，为做好水质风险预警，根据2012年5月~2016年2月的高锰酸盐指数月检数据，预测未来6个月的高锰酸盐指数月度数据。采用的分析步骤为：

1）对水质风险预警应用场景进行分析并选择高锰酸盐指数作为预警分析指标；

2）对高锰酸盐指数数据进行整理和数据预处理；

3）根据应用场景，结合数据分布特征选择适宜的预警模型（回归、聚类、关联等），此处选择指数平滑模型；

4）对模型分析结果进行解读，并给出相应的业务策略或建议。

部分原始数据如表4-3所示。

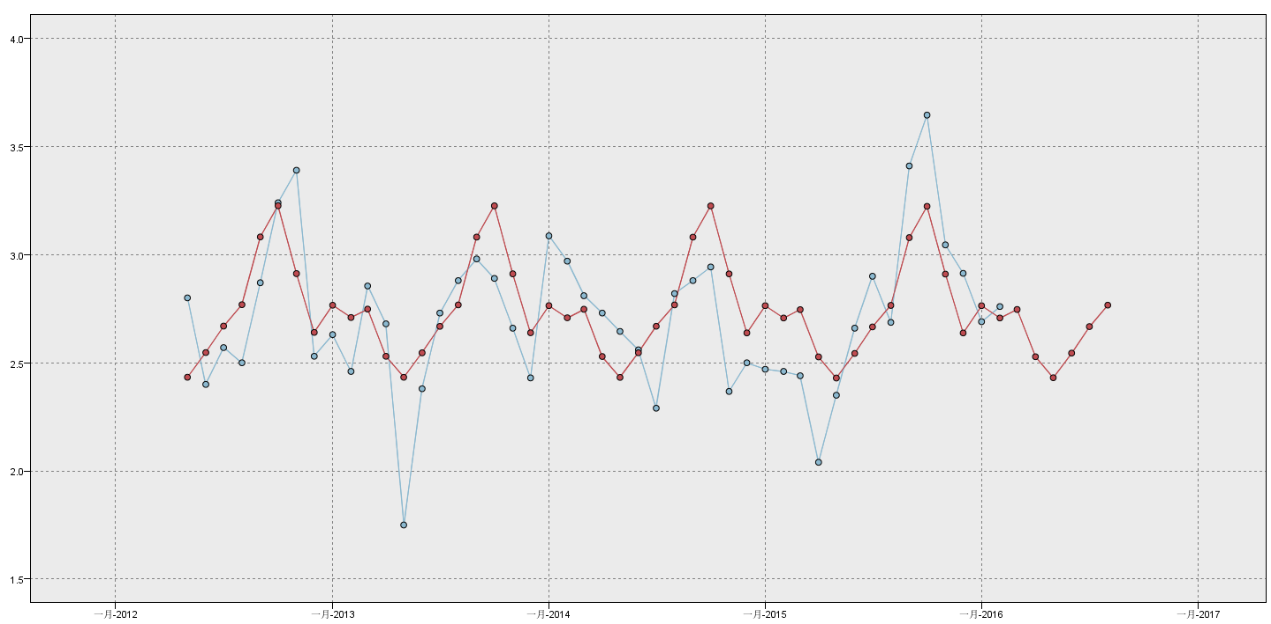
**表4-3 XX水库高锰酸盐指数浓度**

|  |  |  |
| --- | --- | --- |
| 序号 | 时间 | 实测值（mg/L） |
| 1 | 2012年5月 | 2.80 |
| 2 | 2012年6月 | 2.40 |
| 3 | 2012年7月 | 2.57 |
| 4 | 2012年8月 | 2.50 |
| 5 | 2012年9月 | 2.87 |
| 6 | 2012年10月 | 3.24 |
| 7 | 2012年11月 | 3.39 |
| 8 | 2012年12月 | 2.53 |
| 9 | 2013年1月 | 2.63 |
| 10 | 2013年2月 | 2.46 |
| 11 | 2013年3月 | 2.86 |
| 12 | 2013年4月 | 2.68 |
| 13 | 2013年5月 | 1.75 |
| 14 | 2013年6月 | 2.38 |
| 15 | 2013年7月 | 2.73 |
| 16 | 2013年8月 | 2.88 |
| 17 | 2013年9月 | 2.98 |
| 18 | 2013年10月 | 2.89 |
| 19 | 2013年11月 | 2.66 |
| 20 | 2013年12月 | 2.43 |
| … | … | … |

高锰酸盐指数随时间的变化趋势如图4-1所示，可见高锰酸盐指数的年际变化趋势性不明显，在年内存在简单的季节性波动：冬春季节（约从12月-次年5月）高锰酸盐指数浓度较低，夏秋季节（约从6月-11月）高锰酸盐指数浓度较高。根据数据波动特征，选择指数平滑法模型进行高锰酸盐指数浓度分析与预测。

**图4-1 高锰酸盐指数浓度的时间序列图**

应用指数平滑法对2016年3月至2016年8月的高锰酸盐指数进行预测，分析建模的46个月份数据预测误差（见图4-2），可见模型精度尚可。



预测值

实测值

**图4-2 高锰酸盐指数浓度预测曲线**

采用指数平滑模型预测2016年3月至2016年8月的高锰酸盐指数浓度分别为2.75、2.53、2.43、2.55、2.67和2.77 mg/L，预测未来6个月高锰酸盐指数暂不存在超标风险。

### 水厂运行工艺调整辅助决策案例

XX市XX水厂为了提前准备工艺预案，基于2014~2017年实测的进出水水质数据和记录的运行工况数据，建立“水源地水质/水量-水厂药耗-出水水质”在不同区间下一一对应的关联性，并采用神经网络模型预测出达标水质对应的水源地水质预警值或药耗最小值。采用的分析步骤为：

1）水源地及水厂运行多年的相关数据获取；

2）建立数据间的关联性；

3）基于关联数据构建辅助决策模型。

本模型以XX水厂为模拟对象，调取并整理2014~2017年输入值和输出值涉及的相关数据约1363组，由于数量关系仅列举部分数据组，见表4-4和4-5所示。

**表4-4 部分输入值的获取结果**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 日期 | 进水条件 | | 进水水质 | | | | | 进水量 | 药剂投加量 | |
| 温度 | pH | 浊度 | 色度 | CODMn | 氨氮 | 细菌总数 | PAC投加量 | 加氯量 |
| ℃ | - | NTU | 铂钴色度  单位 | mg/L | mg/L | CFU/mL | m3 | kg | kg |
| 2014/1/1 | 4 | 8.11 | 6.61 | 20 | 6.45 | 0.27 | 180 | 160090 | 5336 | 576 |
| 2014/1/2 | 4 | 8.16 | 6.15 | 25 | 6.65 | 0.17 | 100 | 156070 | 5202 | 587 |
| 2014/1/3 | 5 | 8.19 | 6.7 | 25 | 6.61 | 0.19 | 250 | 154410 | 5147 | 672 |
| 2014/1/4 | 5 | 8.16 | 8.56 | 20 | 6.41 | 0.23 | 480 | 153900 | 5130 | 683 |
| 2014/1/5 | 5 | 8.11 | 20.9 | 25 | 6.02 | 0.25 | 210 | 155670 | 5189 | 762 |
| 2014/1/6 | 5 | 8.12 | 31.2 | 25 | 6.18 | 0.25 | 100 | 149210 | 4974 | 782 |
| 2014/1/7 | 5 | 8.12 | 22.6 | 25 | 6.1 | 0.26 | 210 | 150360 | 5012 | 782 |
| 2014/1/8 | 5 | 8.19 | 25.1 | 20 | 6.06 | 0.28 | 87 | 154120 | 5137 | 586 |
| 2014/1/9 | 5 | 8.17 | 20.4 | 20 | 6.06 | 0.29 | 120 | 151750 | 5058 | 762 |
| 2016/7/10 | 25 | 7.91 | 3.01 | 18 | 5.04 | 0.03 | 385 | 292670 | 22087 | 1034 |
| 2016/7/11 | 25 | 7.78 | 24.8 | 30 | 5.04 | 0.15 | 255 | 319630 | 22268 | 999 |
| 2016/7/12 | 26 | 7.93 | 3.97 | 20 | 5.28 | 0.19 | 700 | 291590 | 23547 | 1057 |
| 2016/7/13 | 26 | 7.6 | 5.53 | 25 | 5.28 | 0.12 | 360 | 297530 | 23919 | 1084 |
| 2016/7/14 | 21 | 7.82 | 2.72 | 18 | 4.8 | 0.21 | 320 | 303040 | 23003 | 1042 |
| 2016/7/15 | 22 | 7.79 | 2.16 | 20 | 5.79 | 0.6 | 200 | 282680 | 21111 | 945 |
| 2016/7/16 | 22 | 7.89 | 2.62 | 20 | 5.92 | 0.88 | 600 | 291760 | 22580 | 966 |
| 2016/7/17 | 22 | 7.81 | 2.23 | 20 | 5.2 | 0.23 | 205 | 287960 | 22211 | 958 |
| 2016/7/18 | 25 | 7.79 | 39.7 | 18 | 6.8 | 0.7 | 295 | 287960 | 22999 | 1399 |
| 2016/7/19 | 24 | 7.89 | 16.4 | 25 | 6.84 | 0.8 | 215 | 288820 | 22856 | 1210 |
| …… | …… | …… | …… | …… | …… | …… | …… | …… | …… | …… |

**表4-5部分输出值的获取结果**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 日期 | 出水水质 | | | | 污染物去除率 | | | |
| 浑浊度 | 色度 | CODMn | 细菌总数 | 细菌总数 | 浑浊度 | 色度 | CODMn |
| NTU | 铂钴色度单位 | mg/L | CFU/mL | % | % | % | % |
| 2014/1/1 | 0.18 | ＜5 | 3.84 | 0 | 97.28 | 75.00 | 40.47 | 100.00 |
| 2014/1/2 | 0.27 | ＜5 | 3.88 | 0 | 95.61 | 80.00 | 41.65 | 100.00 |
| 2014/1/3 | 0.27 | ＜5 | 3.84 | 0 | 95.97 | 80.00 | 41.91 | 100.00 |
| 2014/1/4 | 0.26 | ＜5 | 3.72 | 0 | 96.96 | 75.00 | 41.97 | 100.00 |
| 2014/1/5 | 0.19 | ＜5 | 3.8 | 0 | 99.09 | 80.00 | 36.88 | 100.00 |
| 2014/1/6 | 0.26 | ＜5 | 3.64 | 0 | 99.17 | 80.00 | 41.10 | 100.00 |
| 2014/1/7 | 0.24 | ＜5 | 3.56 | 0 | 98.94 | 80.00 | 41.64 | 100.00 |
| 2014/1/8 | 0.26 | ＜5 | 3.6 | 0 | 98.96 | 75.00 | 40.59 | 100.00 |
| 2014/1/9 | 0.23 | ＜5 | 3.52 | 0 | 98.87 | 75.00 | 41.91 | 100.00 |
| 2016/7/10 | 0.22 | ＜5 | 2.72 | 0 | 92.69 | 72.22 | 46.03 | 100.00 |
| 2016/7/10 | 0.22 | ＜5 | 2.72 | 0 | 92.69 | 72.22 | 46.03 | 100.00 |
| 2016/7/11 | 0.21 | ＜5 | 2.68 | 0 | 99.15 | 83.33 | 46.83 | 100.00 |
| 2016/7/12 | 0.27 | ＜5 | 2.92 | 0 | 93.20 | 75.00 | 44.70 | 100.00 |
| 2016/7/13 | 0.27 | ＜5 | 2.68 | 0 | 95.12 | 80.00 | 49.24 | 100.00 |
| 2016/7/14 | 0.26 | ＜5 | 2.72 | 0 | 90.44 | 72.22 | 43.33 | 100.00 |
| 2016/7/15 | 0.26 | ＜5 | 2.89 | 0 | 87.96 | 75.00 | 50.09 | 100.00 |
| 2016/7/16 | 0.33 | ＜5 | 2.96 | 0 | 87.40 | 75.00 | 50.00 | 100.00 |
| 2016/7/17 | 0.29 | ＜5 | 2.72 | 0 | 87.00 | 75.00 | 47.69 | 100.00 |
| 2016/7/18 | 0.35 | ＜5 | 3.44 | 0 | 99.12 | 72.22 | 49.41 | 100.00 |
| 2016/7/19 | 0.43 | ＜5 | 3.24 | 0 | 97.38 | 80.00 | 52.63 | 100.00 |
| …… | …… | …… | …… | …… | …… | …… | …… | …… |

采用人工神经网络法进行建模。模型的主要框架结构由输入层神经元、输出层神经元和隐藏层神经元这三个基本处理单元组成。人工神经网络法最大的特点为具备自我学习能力，即每个函数都有一堆复杂参数，可以根据数据和已知结果，来进行参数的不断调整，最大程度提高正确率。人工神经网络的训练过程主要分为以下两个步骤：第一步骤为信号的前向传播，从输入层通过隐藏层，最终通过输出层得到结果；第二步骤为误差的反向传播，通过计算代价函数（通常为均方误差），由输出层向输入层，根据梯度下降法的基本原则，从后往前依次调节层间的连接权重和偏置。通过对以上两个步骤的反复迭代，神经网络不断更新和优化参数，最终将误差控制在较小的范围内，从而实现对输入相对精确的预测。

本神经网络模型以python为编码语言，以表4-4和4-5所示的大量输入、输出数据值为建模基础，构建形同人的神经元的“获取信息-处理信息-反馈结果-调整权重”的流程，通过输入与输出值的对应关系，不断地反向传播、降低梯度（将神经网络里的参数朝着误差函数斜率下降的方向进行调整）、最终趋于拟合的反馈系统，待训练到与真实输出的误差小于10-3时，可得到初步的模型，再利用一定数量的验证集（同样包含准确的输入、输出值），进行拟合度（R2）的回归性验证，若R2值接近于1，则证明神经网络在此类型的输入上已掌握充分，且具备较准确的反映能力。

模型的输入、输出值均以向量组的形式，形如（T1,T2,S1,S2,S3,S4,S5,L,P1,P2）和（Q1,Q2,Q3,Q4,D1,D2,D3,D4,），其中字母所代表的指标含义如表4-6所示。

**表4-6 关联性数据**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 功能 | 输入值 | | | | | | | | | | | | | | |
| 项目 | 进水条件 | | | 水源地水质 | | | | | | | | 进厂水量 | | 药品投加量 | |
| T1 | T2 | | S1 | | S2 | S3 | S4 | | S5 | | L | | P1 | P2 |
| 温度 | pH | | 浑浊度 | | 色度 | CODMn | 细菌  总数 | | 氨氮 | | 进水量 | | PAC  投加量 | 加氯量 |
| 功能 | 输出值 | | | | | | | | | | | | | | |
| 项目 | 出厂水质 | | | | | | | | 污染物去除率 | | | | | | |
| Q1 | | Q2 | | Q3 | | Q4 | | D1 | | D2 | | D3 | D4 | |
| 浑浊度 | | 色度 | | CODMn | | 细菌总数 | | 浑浊度 | | 色度 | | CODMn | 细菌总数 | |

输入值主要包括进水条件（温度、pH）、水源地水质（常规水质指标）、进水水量及药品投加量（PAC投加量、加氯量）等四方面内容，输出值主要包括经净水工艺处理后，出厂水质（常规水质指标）及所实现的污染物去除率。

本模型采集1363个样本用来进行模型的训练，按4:1分为训练集和验证集，验证集用来调整模型训练的超参数。在训练过程中，选用“adam”优化器算法，初始学习率（learning rate）为0.001，批尺寸（batch size）为16，训练时学习率（learning rates cheduler）的容忍次数（patience）为5，衰减率（decay ratio）为0.1，早期停止（early stopping）的容忍次数（patience）为10。

模型主要通过对各个卷积层的神经元数量和分布进行调整、处理器的优化选择、各其余层次的分布及参数调整，共计调整10次，每次均运算出20个误差值并取平均值，对比并进一步优化。模型优化过程结果见表4-7所示。

**表4-7 模型优化过程**

|  |  |  |
| --- | --- | --- |
| 优化选项 | 均方误差值 | 均方误差最终平均值（mean）/标准差（std） |
| 初始神经网络 | 0.00293458 0.00438405 0.00324121 0.0035162  0.00273117 0.00417119  0.00263683 0.00370243 0.00283205 0.00385743 0.00422358 0.0042471  0.00392071 0.00440823 0.00387573 0.00405216 0.00444746 0.0036238  0.00402404 0.00400557 | mean: 0.003741776265500966  std: 0.000564526269742618 |
| filter\_number（过滤器数量）从32改成16 | 0.00340019 0.00373671 0.00362741 0.00338937 0.00406527 0.0034718  0.00283635 0.003349  0.003416 0.00302272  0.00328775 0.00373456  0.00364304 0.00493004 0.00322945 0.00402437 0.00380561 0.00378313  0.00296916 0.00312062 | mean: 0.0035421265394161607  std: 0.00045773184663871087 |
| filter\_number从32改成8 | 0.00385583 0.00354566 0.00398654 0.00452405 0.00347082 0.00300605  0.0038133 0.00437118  0.0039284 0.00433513  0.00483271 0.00298115  0.00390666 0.00348626 0.00422772 0.00434381 0.00362064 0.00336281  0.00394402 0.00414961 | mean: 0.003884617220388238  std: 0.0004766241549775268 |
| filter\_number从32改成16，  优化器算法从adam改成sgd | 0.0088071 0.00635238  0.00542842 0.00662819 0.00633474 0.00602813  0.00547356 0.00585785 0.00509513 0.00860585 0.00506665 0.00568831  0.00565679 0.00553059 0.00646935 0.00675066 0.00609113 0.00607565  0.00539467 0.00539521 | mean: 0.006136516305617899  std: 0.000980010528316175 |
| filter\_number从32改成16，  kernel size（卷积核尺寸）从2改成3 | 0.0031447 0.00368669  0.00402096 0.00515812 0.00418097 0.00449132  0.00415816 0.00495875 0.00397378 0.00366731 0.00316181 0.00369592  0.00430191 0.00316735 0.00320676 0.00440672 0.00537768 0.00382584  0.00485529 0.0039899 | mean: 0.004071498106809119  std: 0.0006493458705114325 |
| filter\_number从32改成16，从6层卷积改成4层卷积（去掉最后两层卷积和池化层）  激活层从relu函数改成elu函数 | 0.00352578 0.00503989 0.00536297 0.0032764  0.00466299 0.00269701  0.00377756 0.00355275 0.00295632 0.00349901 0.00318469 0.00350986  0.0041002 0.00360342  0.00423021 0.00447223 0.00442995 0.0037775  0.00377594 0.0033936 | mean: 0.0038414157108122233  std: 0.0006675419101164789 |
| filter\_number从32改成16，  激活层从relu函数改成elu函数 | 0.00394467 0.00296702 0.00469443 0.00315081 0.00327057 0.00302116  0.00351136 0.00437996 0.00426543 0.00367095 0.00275784 0.00498831  0.00370524 0.00352701 0.00480724 0.00361135 0.00434492 0.00377661  0.00333595 0.00341832 | mean: 0.0037574578199157807  std: 0.0006216987108190994 |
| filter\_number从32改成16，  batch size（批尺寸）从16改成32 | 0.00346477 0.00357344 0.00486685 0.004031  0.00342979 0.00335211  0.00338035 0.00411582 0.00415355 0.00384353 0.00360677 0.00360305  0.00419632 0.00480416 0.00397911 0.00369778 0.00643534 0.00344511  0.00399647 0.00467953 | mean: 0.004032742386629532  std: 0.0007134235461666378 |
| filter\_number从32改成16，  batch size从16改成8 | 0.00334664 0.00428098 0.00367296 0.00463347 0.00358989 0.00286834  0.00333247 0.00317273 0.00460664 0.00355844 0.00381546 0.00437941  0.00353168 0.0055051  0.00360997 0.00320035 0.00431124 0.00343332  0.00437089 0.00447486 | mean: 0.003884742306431039  std: 0.0006373947191318135 |
| filter\_number从32改成16，  pooling size（池化尺寸）从2改成3 | 0.00341355 0.00417597 0.00421124 0.00545192 0.00398359 0.00315216  0.00379403 0.00435255  0.0038006 0.00545121  0.00479901 0.0052555  0.00324383 0.00436083 0.00367505 0.00419544 0.00388475 0.00290102  0.00328162 0.00426579 | mean: 0.00408248310124572  std: 0.0007173318515959882 |

利用五年的实测数据，建立起对应的神经网络模型，并进行了界面开发，开发的预测界面如图4-3所示。

****

**图4-3 水厂运行工艺调整辅助决策预测界面**

在具备拟合实际情况的模型基础上，通过改变不同的工况条件，可准确快捷地预测得到对应的出水水质和污染物去除率，同时可反推在出水水质达标要求下，进水条件或各工况工艺参数的预警值。

# 供水管网运营大数据应用

## 管网运营大数据应用目的

供水管网是城市不可或缺的基础设施，越来越多的城市供水管网破损事故、水质事件等对市民的生产生活用水安全带来了极大的不便甚至危险，因而有必要对供水管网的综合风险状态进行评估，这也是供水管网改扩建、优化、调度等工作的基础。供水管网发生破损事故或水质事件，通常是内外因综合作用的结果，影响因素众多且复杂，需要将各种因素进行梳理和提炼，形成管网综合风险评估方法，作为管网运营大数据分析路径进行推广。综合风险评估结果有利于城镇供水单位和主管部门开展科学决策。

## 管网运营大数据积累要求

城镇供水单位宜优先建立供水管网GIS系统（Geographic Information System，地理信息系统）、SCADA系统（Supervisory Control And Data Acquisition，数据采集与监视控制系统）等信息管理系统，对供水管网进行信息化管理，确保供水管网数据的准确性、全面性和时效性，并应实现管网信息化系统与综合风险评估系统的数据共享。供水管网综合风险评估系统应收集管网资产数据、环境数据、运营数据、水质数据及其它重要相关数据。

### 管网资产基础数据

管网资产基础数据即用于描述供水管网资产本身属性的数据，该数据宜从供水管网地理信息中获取，包括管材（闭路电视监控系统内部检测、超声波壁厚检测、管材实验室管材分析等数据）、管径、管长、建设时间、埋深、内外防腐情况、施工方式、接口方式、供应商、建设费用、修复日期、修复技术等。

### 环境数据

城镇供水单位应注意收集用于描述供水管网所处的内部环境及外部环境的数据，包括管道位置、水环境、土壤环境、道路负荷、气温气候变化、阴极保护情况、铁路与地铁的接近性等。

### 运营数据

城镇供水单位应注意收集供水管网运行状况的相关数据，包括水压检测数据、探漏数据、各区块漏损数据、运营费用、管网修复技术、更新改造技术、修复费用、更新改造费用等。

### 水质相关数据

水质相关数据包括在线水质监测、实验室检测、管材检测数据等，实验室与在线监测水质数据积累要求参见第四章第二节第1点。

### 管网破损历史数据

管网破损历史数据包括管道编号、事故日期、位置、破损类型、破损原因、破损维修方法、维修费用及故障后果（包括水量损失、水力重要性、用户影响情况、道路重要性、维修成本、水损指标、区域重要性等）等。

## 管网运营大数据典型应用场景

### 管网运营风险预警/更新改造辅助决策

通过供水管网运营风险评估，城镇供水主管部门可掌握供水管网破损风险、水质风险等预警情况，城镇供水单位可对供水管网运营维护与修复/更新改造的经济技术优化，实现以供水成本效益最优的方式，为用户提供满足或优于关键绩效目标的服务。通过开展供水管网运营风险预警，城镇供水单位可根据管道评估分析结果，制定管道修复/更新改造的优先级，同时可明确管网资产修复/更新改造的范围，并评估远期的供水服务水平及资产状况，并在满足此服务水平及资产状况下制定修复/更新改造所需的费用等。

管网运营风险评估及更新改造辅助决策的实施流程可概括为三大步骤：

1）风险概率预测。明确管网运营风险的具体评价对象，选择建模因子和适当的建模技术，对供水管网的运营风险评价指标发生概率进行定量预测。

2）风险评估。结合管道级别、道路等级、人口密度等因素，构建管道风险发生的影响程度定量判别指标体系和评判标准，评估管道运营风险，确定各管道的风险程度。

3）风险应对。基于风险评估的结果，制定相应的修复/更新改造计划，并跟踪评估实施效果。

可采用的技术路线图为：

数据获取

数据预处理

建立事故概率模型

评价模型预测效果

预测评估管道风险概率

制定修复/更新改造计划

风险评估指标体系

管道风险评估辅助决策支持系统

跟踪关键绩效指标，评估管道修复/更新实施效果

评估管道运营风险

在评估管网运营风险和辅助决策过程中，应根据自身数据质量和数据特征等情况，选择适合自身供水管网特点的模型对供水管网进行运营风险评估及后续更新改造辅助决策。

### 管网漏损控制

城镇供水单位在对供水管网现状运营大数据收集和整理的基础之上，建立相关管理系统和管理机制，可为供水管网漏损控制的评估和管理提供技术支撑。分区计量是管网漏损控制的科学有效的手段，通过建立全区域的计量管理体系，进行全网络水平衡分析，评估各区域管网漏损情况，有效的识别管网漏损的重点对象和范围，科学的指导开展管网漏损控制工作，减少资源浪费。

分区计量管理工作中，评估管网现状、收集管网信息需要建立起GIS系统管理基础管网资产数据；分区的划分依据和分区的管理论证需要依赖GIS系统中基础地图中反映的环境数据来开展；建立分区计量和对照体系并纳入管理，需要有管网水量、水压运营数据，以及SCADA系统或远程表务海量数据信息的支撑；漏损控制管理需要对漏损水量进行绩效核算，对水质、水压和高漏耗进行监督和报警，建立一套监督报警的分区计量管理系统。

利用供水大数据开展管网漏损控制可参照《城镇供水管网分区计量管理工作指南——供水管网漏损管控体系构建》要求，建立基于DMA（District Metering Area，独立计量区域）的分区计量管理系统，利用管网运营大数据，用科学高效的方法和精细化的管理来降低漏损，提升供水安全保障的能力。

## 供水管网运营风险评估典型案例

XX供水企业拟选择破损风险和水质风险作为供水管网运营风险的评估指标，建立供水管网运营风险评估模型，在此基础上对其服务范围内的供水管网运营风险进行评估，从而优化供水管网的运营维护与修复计划，实现以供水成本效益最优的方式，为当前及未来用户提供满足或优于关键绩效目标的服务。对管网运营概率评估采用的分析步骤为：

1）选择建模数据，进行数据预处理；

2）选择合适的建模方法，构建管网破损概率预测模型，验证模型准确性，预测管网破损风险；

3）选择合适的建模方法，构建管网水质风险预测模型，验证模型准确性，预测管网水质风险；

4）分别赋予管网破损风险和水质风险合适的权重系数，将两部分风险按照权重加和，得到管网运营综合风险指数，用于评估管网的运营风险。

选择参与建模的数据包括：环境数据（道路负荷）、管网数据（管材、管龄和管径）和水力模型数据（压力、流量和水龄）。将所需参与建模的数据分类收集并存储，部分数据如表5-1所示。

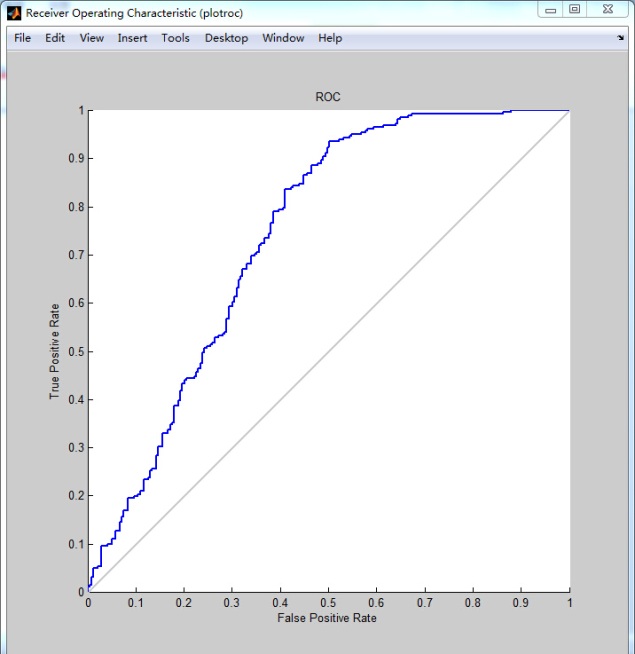
**表5-1 参与建模的部分基础数据整理示意表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 管径（mm） | 道路名称 | 管长（km） | 管道材质 | 建设年份 |
| 1 | 400 | 红岭南路 | 68.09 | CIP | 1983 |
| 2 | 200 | 滨河大道 | 5.16 | S | 1999 |
| 3 | 700 | 向园路 | 0.78 | S | 1994 |
| 4 | 600 | 园林路 | 122.44 | S | 1985 |
| 5 | 200 | 园林路 | 12.27 | S | 2003 |
| 6 | 600 | 中兴路 | 83.7 | RCP | 1985 |
| 7 | 400 | 东门北路 | 14.28 | CIP | 1995 |

在建立管网运营风险评估模型前，需对数据进行预处理，包括数据质量检查、分类和归一化检查。对于管网数据，由于同一条道路上多条管径、材质、铺设年份一致的管段会造成大量同质化严重的样本数据，因此需进行优化处理，合并同类数据。此外，GIS数据中的复杂边数据在水力模型中没办法直接运算，需要合并为简单边数据，便于关联水力模型成果。预处理（质量检查）后的部分数据示意如表5-2所示。

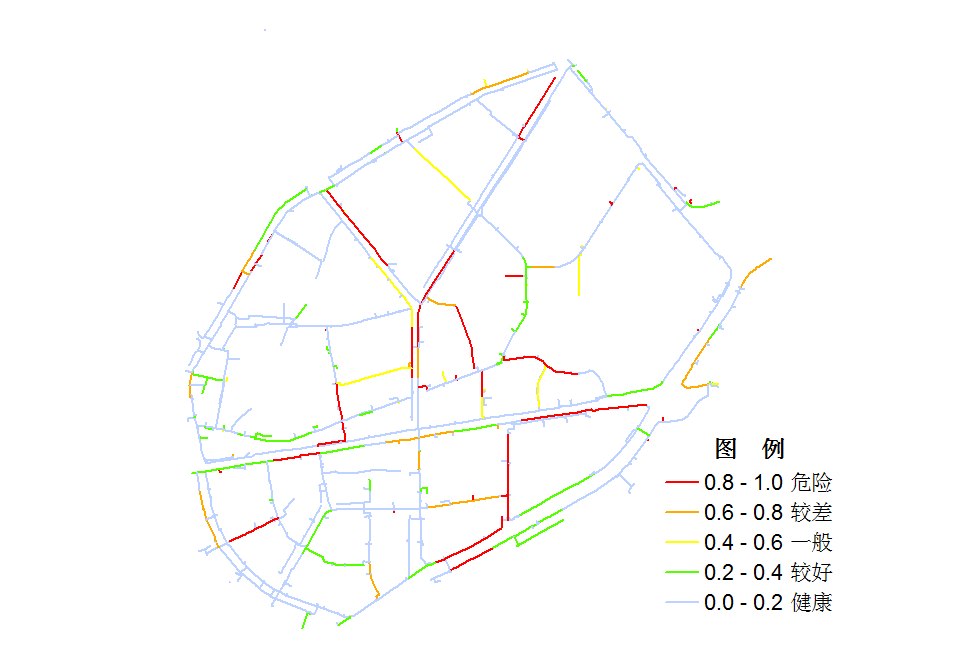
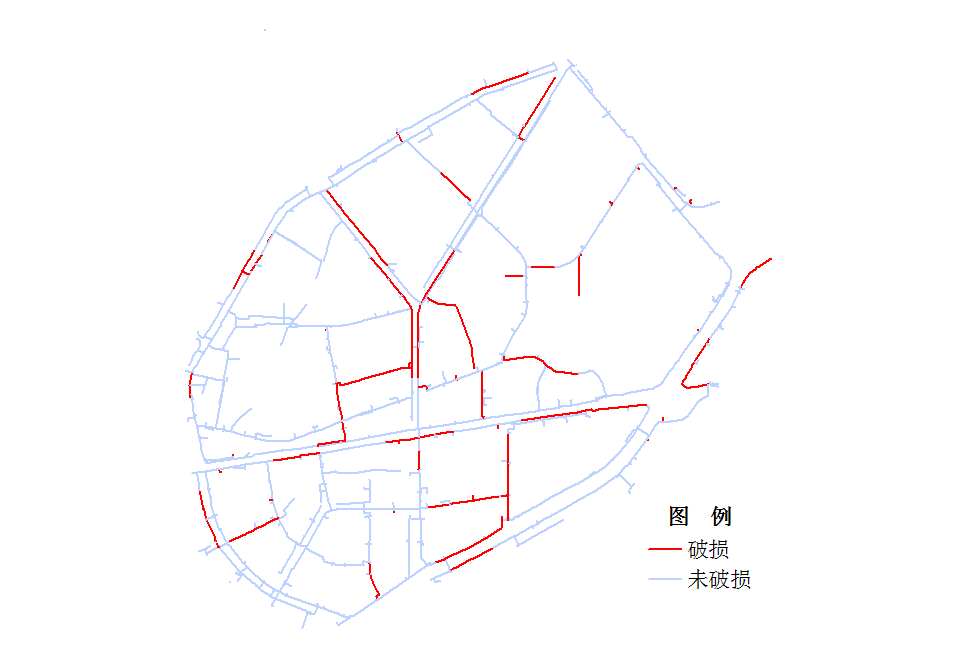
**表5-2 参与建模的部分预处理后数据整理示意表**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 管线编号 | 管径 | 管材 | 管龄 | 道路负荷 | 运行压力 | 杂散电流 | 是否发生破损 |
| 315711 | 400 | 2 | 9 | 4 | 34.07 | 1 | 1 |
| 106787 | 1000 | 5 | 14 | 2 | 42.78 | 0 | 1 |
| 489678 | 300 | 6 | 20 | 0 | 42.76 | 0 | 0 |
| 193536 | 250 | 4 | 4 | 3 | 37.14 | 0 | 0 |
| 102190 | 200 | 1 | 16 | 5 | 44.36 | 1 | 1 |
| 110772 | 800 | 5 | 32 | 0 | 41.75 | 0 | 1 |
| 309219 | 600 | 2 | 11 | 1 | 43.34 | 1 | 1 |
| 615496 | 200 | 6 | 5 | 0 | 29.66 | 0 | 0 |
| 507080 | 300 | 6 | 7 | 3 | 35.16 | 0 | 0 |
| 109813 | 800 | 5 | 17 | 0 | 41.98 | 0 | 0 |

数据预处理完成后，选择采用随机森林模型建立了管网破损风险预测模型，采用ROC曲线和AUC值对模型预测准确率进行校核。ROC曲线如图5-1所示。

**图5-1 预测模型的ROC曲线**

为验证模型准确性，对部分管道的实际破损情况进行了勘察，并利用建立的破损风险预测模型对管道破损风险进行预测，将破损风险模型预测结果与管网实际破损情况进行对比，如图5-2所示，可见基本上管网破损越严重，预测的破损风险越高，模型预测结果较准确。



管网实际破损情况

管网破损概率预测情况

**图5-2 供水管网破损概率预测结果与实际情况的对比图**

随后采用随机森林模型建立水质风险模型，在此过程中，将水质异常作为确定水质风险事件的指标，将管网水水质自身风险和爆管事件均考虑作为发生水质风险的外因。模型预测准确率校核过程与破损风险预测模型相同。

XX供水企业利用建立的破损风险预测模型和水质风险预测模型对更大范围的管道破损风险和水质风险进行了预测，按照各自0.5的权重系数对爆管风险预测模型与水质风险预测模型的结果加和作为综合运营风险评估分数值，以该数值的大小来量化评价管网运营综合风险，并作为制定管网更新改造计划的重要数据参考。

# 公众反馈信息大数据应用

## 公众反馈信息大数据应用目的

以城镇供水主管部门及供水单位对供水用户的客服经验为基础，以公众反馈的供水客服、网络舆情等数据作为数据应用的核心，结合与之相关的生产、营销等数据，运用大数据分析的思想和有效的数据分析手段，充分发掘围绕供水过程中的水质、水量、压力等服务痛点问题，发掘并解释问题成因，以及建立可能的解决路径，从而为掌握公众舆情、提高监管精度、提升供水服务质量提供相应的数据支撑。

## 公众反馈信息大数据积累要求

### 供水运营数据

城镇供水单位应积累包括当地供水管网地理信息数据、供水水力模型、厂站供水区域图例，以及供水过程中厂站、二次供水的监测数据。对水质数据积累要求参见本指南第四章第二节第1点。

对于厂站监测数据，还应包括但不限于进出厂水瞬时流量、进出厂水累计流量、进出厂水瞬时压力、进出厂水日均压力等。

对于二次供水加压站点的监测数据，还应包括但不限于进出站水瞬时流量、进出站水累计流量、进出站水瞬时压力、进出站水日均压力等。

### 供水服务业务数据

城镇供水单位应积累包括用户供水服务问题工单的信息来源（来电、市民服务热线、官网、网络社交平台等）、信息类别（咨询、问题、投诉、报装等）、产生时间、产生区划及地点、业务流转状态等业务数据。

### 营销业务数据

城镇供水单位应积累相关营销业务数据，主要包括用户开户基本信息、给水号、水表信息、用水状态（是否用水）等。

### 本地气象数据

城镇供水单位对气象数据的积累要求参见本指南第四章第二节第2点。

## 公众反馈信息大数据典型应用场景

### 公众反馈供水问题热词和热图解析

城镇供水主管部门及供水单位在供水服务过程中积累了大量的客服案例与经验，可以通过运用文本发掘的思想，以客户服务实际经验作为基础，对一段时间内的客户服务案例数据中所涉及到的用户信息、信息来源、信息类别、服务类型、受理时间、反应地点、服务内容等相关人工填报及自动采集数据，采用适当的数据挖掘技术，进行辨识、分类、总结出关键词，并以选取高频排序优先的关键词为原则构建出公众反馈问题热词库。根据确定分析的主题方向，以现有热词库为基础，对客服案例数据进行数据模糊匹配聚类，并对聚类结果进行文本可视化分析，构建出在时间、空间、内容等层面上的反馈问题热图，再结合与之相关的生产、营销等数据，选取适当的数据分析技术，构建出相关相异分析模型，继而可有针对性地开展问题溯源，从而起到精准掌握公众舆情、改进服务辅助决策、提高服务精度的作用。

供水客服信息热词分析过程可采用的技术路线为：

供水客服数据

数据获取

数据预处理

构建热词库

数据模糊匹配聚类

文本可视化分析

相关相异分析

结果分析判断

明确分析主题

分析影响因子

### 片区用户用水行为分析

城市用水量呈现一定的规律变化，主要体现在用户用水的周期性。以供水客服数据中的用户信息数据与管网末端小区二次供水数据为基础，结合其它相关数据，城镇供水主管部门及供水单位可以对用户所在小区二次供水泵房数据进行归类划分，量化泵房间的基本差异。运用时间序列分析思想，采用适当的数据挖掘技术，构建二次供水泵房指标变量时间序列，继而采用适当的数据分析技术，对不同序列之间的指标变量的趋势性、周期性及其随机性变化规律进行相关相异分析，识别出指标变量突出的相关相异点，并在此基础上总结得出该序列所代表小区用水行为变化趋势。对于已得出的用水行为变化趋势，根据不同小区之间用户平均年龄、小区位置、小区建造时间等因素，有针对性地定性溯源变化趋势产生原因，从而为提前做好客户服务预案和供水调度决策提供数据支撑。

供水用户用水行为特征分析过程可采用的技术路线为：

二次供水数据

数据获取

数据预处理

统计抽样

相关相异分析

结果分析判断

指标变量

供水客服数据

归类划分

分析影响因子

## 公众反馈信息大数据应用典型案例

### 公众反馈供水问题热词分析案例

XX市某供水企业在2016年10月至12月之间的客户服务月报中发现水质投诉类客服工单数量增加明显，为进一步完善客户服务知识库，以近三年（2015年至2018年）客户服务工单数据作为基础，结合与之相关的其它业务数据，溯源水质投诉类客服工单数量显著波动的原因。采用的分析步骤为：

1）以数据流条目的形式收集、获取客户服务工单数据；

2）对现有数据流条目进行数据预处理，筛除无效数据，筛选出包括用户信息、信息来源、信息类别、服务类型、受理时间、反应地点、服务内容等数据元素；

3）对预处理完毕的数据流条目，结合人工分析与机器学习的方法对高频词汇进行统计，筛选出关键词，再根据文本语义与组织结构，进行最小串分词，并以选取高频排序优先的关键词为原则构建出热词库；

4）围绕水质相关的热词，结合热词库，对现有数据流条目进行数据模糊匹配聚类；

5）对聚类于水质相关的数据流条目，分别以其中的受理时间、反应地点等数据元素进行数据文本的可视化分析；

6）根据可视化分析的数据分布特征，结合与水质相关的其它数据，选取适当的相关相异分析模型（回归、关联、信息熵及互信息等），此处选择回归分析，推断出显著的影响因子；

7）对分析结果进行判断，溯源形成的原因，给出客户服务的辅助决策建议。

现有公众反馈问题热词库的部分热词如表6-1所示。

**表6-1 工单类型的部分热词库**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 类型  热词 | 水质 | 施工扰民 | 水表故障 | 营销业务 | … | 管网及设施故障 |
|
| 1 | 水发黄 | 噪音 | 换表后漏水 | 抄收服务 | … | 管道损坏漏水 |
| 2 | 水发蓝 | 扬尘 | 换表维修后无水 | 水量的异议 | … | 闸门损坏漏水 |
| 3 | 水有虫 | 出行不便 | 水表自转 | 表制度的异议 | … | 消防损坏漏水 |
| 4 | 水垢多 | 施工垃圾 | 表蒙子漏水 | 违约金的异议 | … | 井盖损坏 |
| … | … | … | … | … | … | … |
| N | 水发白 | 无围挡 | 水表丢失 | 通知单及短信 | … | 管廊漏水 |

通过对2015年至2018年之间的与水质相关数据进行可视化分析发现，2016年10月至12月的水质投诉类工单确实显著增多，其中的高频热词为龙头水有异味（如图6-1所示）。

**图6-1 2016年度水质投诉及龙头水有异味热词工单逐月数量变化图**

将其中所有涉及龙头水有异味数据的发生地点进行数据抽取，发现所涉及地点围绕JS路周边沿线自西向东分布（如图6-2所示）。



**图6-2 龙头水有异味热词工单发生地点分布热图**

由于JS路主管线水源来自YQ水厂，在排除沿线加压站与管网可能污染水质的前提下，初步判定投诉问题与YQ水厂出厂水水质及其水源地YQ水库的水质相关。选择2016年度龙头水有异味数据数量与YQ水厂出厂水、YQ水库原水日检数据项、月检数据项作为数据分析的影响因子进行相关系数计算并作一元回归分析，发现YQ水库藻类物质含量与有异味数据条目数量存在显著相关性，相关系数为0.969。

通过对YQ水库的水源调查发现，2016年10月~12月，有南水北调的原水供给入YQ水库，供给入水量约为1000万立方米。南水北调原水的注入致使YQ水库水体富营养化程度增大，原水中藻类物质含量急剧增加，藻类物质分解产生土嗅素和2-甲基异莰醇，最终使水体异味增加，龙头水有异味工单数量也随之更加

经此分析，供水企业在外水注入YQ水库之际，增加了对藻类指标的检测，提前制定生产预案控制水质异味，减少了此类问题的投诉率。

### 供水用户用水行为特征分析案例

XX市某供水企业从2016年起陆续对供水区域内的小区二次供水泵房进行了改造、接收，并统一进行管理，为更好地利用现有二次供水设备优化管网末端的压力调度、改善客户服务质量，以用户信息与2018年的二次供水生产数据作为基础，结合与之相关的其它业务数据，分析不同小区之间的用水特征。采用的分析步骤为：

1）以数据流条目的形式收集、获取二次供水及客服用户信息数据；

2）对现有数据流条目进行数据预处理，筛除无效数据，筛选出包括客服用户信息、二次供水在线水质、水量、水压等数据元素；

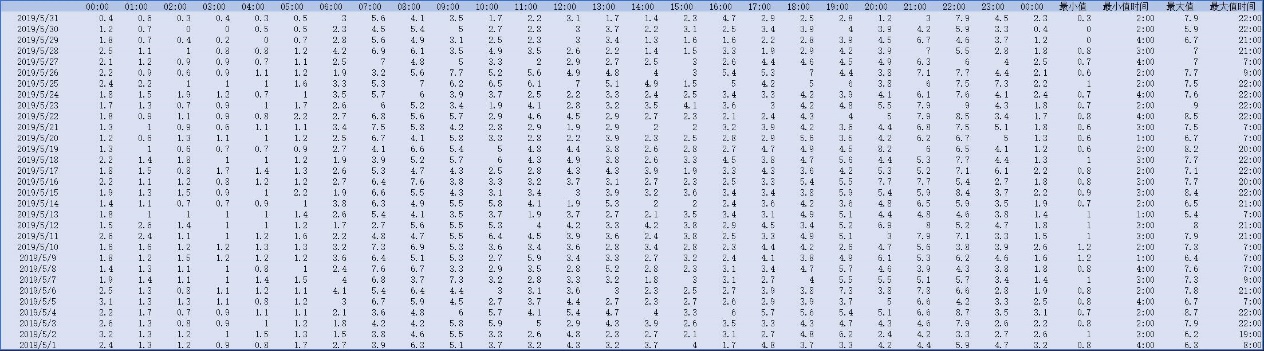
3）以数据元素中二次供水泵房位置信息与客服用户信息为基础，结合供水区域图示等数据对二次供水泵房数据进行分组，此处以相同供水水源、相同上级水厂与加压站作为前提对用户所在小区二次供水泵房数据进行分组；

4）对同组或不同组之间的小区二次供水泵房生产数据，运用时间序列的思想，按照某一抽样维度，选取某一指标变量数据进行抽样统计，此处选择按照通常作息时间划分为4组时段：0-5时（凌晨）、6-11时（上午）、12-17时（下午）和18-23时（晚上）。对上级加压站相同的同组两处不同建造年代小区二次供水泵房a、b的在线实时流量（即每小时累计流量）数据分别进行抽样统计，此处的统计方法是将每一时段内的最高与最低实时流量的所在时间进行抽样，即将每日的实时流量变化抽象为由4组、每组2个时间数据组成的一维数组，例如2018年1月1日泵房a实时流量变化数组：A0101={MAX1, MIN1, MAX2, MIN2, MAX3, MIN3, MAX4, MIN4}={0, 4, 8, 6, 17, 16, 22, 19}；

5）选取适当的相关相异分析模型，此处选择采用t分析，对分别泵房a、b数组内的差异显著性逐月分析，寻找出数组间的相似及相异规律；

6）对分析结果进行判断，溯源形成的原因，给出供水调度的辅助决策建议。

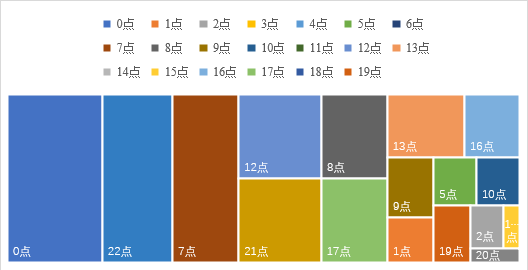
分时抽取的二次供水泵站a流量等数据信息如图6-3所示。

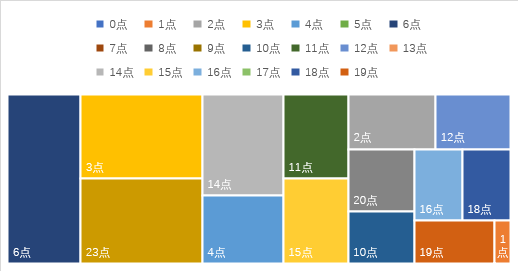


**图6-3 分时抽取的二次供水泵站流量等数据信息**

泵房a所在小区建成于2004年、泵房b所在小区建成于2011年。泵房a、b设备运行状态良好，在线仪表经过校准，并且上游供水水源地相同、水厂相同、上级加压站相同，在抽样过程中只单一抽取了最高、最低实时流量出现的时间，排除了过多其他因素对此次分析的影响。

通过对2018年泵房a、b抽样数组进行相似相异分析发现，泵房b在2018年2、7、8月的数组相异度明显，2018年3、4、5、6、9、11、12月相似度明显，2018年1、10月无明显相似相异倾向，而泵房a在2018年3、4、5、6、7、8、9、11、12月相似度明显，2018年1、2、10月无明显相似相异倾向。对泵房b相似度高的数组数据做数据累计排序，可观察到实时流量最高值在0，20，6时数量多，最低值在3、22、1时数量多。对泵房b相异度高的数组数据做数据累计排序，可观察到实时流量最高值在0，22，7时数量多（如图6-4所示），最低值在6，23，3时数量多（如图6-5所示）。对泵房a的数组数据做数据累计排序，可观察到实时流量最高值在0，18，7，9，10，15，16时数量多，最低值在5，23，2时数量多。



**图6-4 2018年2、7、8月泵房b实时流量最高值数据累计排序**

**图6-5 2018年2、7、8月泵房b实时流量最低值数据累计排序**

根据以上分析结果，结合小区业主用户平均年龄（泵房a所在小区平均48.5岁，泵房b所在小区平均38.3岁），可以推断出泵房b所在小区学前及义务教育阶段适龄人群较多，由于学生假期导致晨间与晚间用水高峰时间的波动。而泵房a所在小区，成年人尤其老年人较多，用水无明显高峰。

借此推断结论，可以对其它如用户信息准确度不高的新建小区、租住用户多的小区等进行以上分析，从而对小区用户用水特征进行画像，为诸如用水调度、生产预案等生产业务，应急供水、营业网点设立等服务业务提供数据辅助决策支撑。

# 附录

**编制依据**

1. GB/T 35295-2017 信息技术 大数据 术语
2. GB/T 35589-2017 信息技术 大数据 技术参考模型
3. GB/T 32063-2015 城镇供水服务
4. GB 5749-2006 生活饮用水卫生标准
5. GB/T 12504-12505-90 计算机软件质量保证及配置管理计划规范
6. GB 8566-88 计算机软件开发规范
7. CJJ 58-2007 城镇供水厂运行、维护及安全技术规程
8. CJ-T206-2017 城市供水水质标准
9. CJJT 271-2017 城镇供水水质在线监测技术标准
10. CJJ 61-2003 城市地下管线探测技术规程
11. CJJ /T 8-2011 城市测量规范