标题

摘要

一、问题重述

1.1 问题的背景

自来水公司对城市供水管网划分为多个独立计量区域(DMA)，在每个 DMA 区域入口处布设了流量仪表，以检测该区域的用水情况（流量和压力）。DMA 的建立能够主动确定区域的泄漏水平，同时通过监测 DMA 的流量，可以识别是否有新的漏点存在。由于管网泄漏是动态的，如果在泄露之初就得到控制，泄漏可以大幅减少；如果没有持续的泄漏控制，泄露会随着时间的延续而增大。因此，DMA 管理被视为在供水管网中减少和维持泄漏水平的有效方法。

现在 DMA 分区入口流量仪表测得瞬时流量中包含该 DMA 区域的各用户用水量之和以及管网漏失水量。同时可认为，当夜间(2:00~5:00)用水量较低时，最小夜间流量可以间接反映该区域的漏失水平。

1.2 问题的提出

基于以上背景，需解决如下问题:

1. 根据测定结果，辨识出该区域的典型用水模式(如居民小区、工厂、机关单位的用水量在时间上有显著不同的分布)；
2. 辨识该分区的漏水量模式；
3. 利用辨识出的典型用水模式，分析特定时刻的水量异常增加，用水高峰期不符合生活或声场规律节律变化的原因（如爆管、偷用生活用水做商业用途）；

二、问题假设

1. 假设数据来源真实可靠。
2. 问题 1 中认为夜间最小流量即为当天的漏失速率。
3. 问题 2 中认为夜间最小流量时刻无用户用水。
4. 问题 2 中认为附件所给压力即为管网的平均压力。

三、符号说明

符号	说明
i	第i天
j	每天第j个测量时间点
x_{ij}	第i天第j个点的数据，数据内容根据建模需要定义
k_{ij}	第i天第j个点的模式系数

四、问题分析

4.1 问题 1 的分析

问题一中需根据测定结果辨识出该区域的典型用水模式，附件中数据给出 2014/4/15 00:00:00 ~ 2014/6/12 22:00:00 每隔 15 分钟测定一次的瞬时流量数据序列，其时间跨度较小，则主要考虑分析其在一天中不同时间段，一周中不同天的变化趋势。由于数据较多，为找到典型数据，结合问题提示中给出实际自来水公司某供水片区的流量监测曲线得到各天的瞬时流量变化趋势相似，所考虑选择各天各时段的典型数据代表该时段，以绘制该区域典型用水模式曲线，查找资料获取各类典型用水的用水数据绘制模式曲线找到特征，与本问题中用水模式曲线进行比对，分析用水峰值和变化趋势，可得到该区域属于哪一用水模式。

4.2 问题 2 的分析

问题二需辨识该地区的漏水量模式，我们认为不同的漏水量模式即为，不同漏失速率登记，查阅资料发现各地区对漏失速率大小的等级划分有所不同，而本问题中所给数据并未指出数据来源地区。为得到更加科学且贴合实际的划分标准，我们希望利用数据本身的特征，通过聚类的办法得到根据漏失速率大小聚集而成的聚类簇，根据每个聚类簇代表的实际意义得到漏失水平的等级划分标准。由于问题一种使用每日的夜间最小流量当日的漏失速率，则漏失速率数率只有 58 个，对于聚类而言数量太少，最终得到的等级划分不合理。我们希望能够得到各天各时刻的漏失速率，将其作为聚类的数据集，从而得到更加准确的登记划分。

五、问题求解

5.1 问题 1 的求解

根据问题分析，考虑探究区域用户用水之和一周内各天各检测时间点在时间上的分布情况。

5.1.1 数据预处理

为找到典型的用户用水模式，首先需排除离群数据、缺失数据、异常数据对结果的影响。

5.1.1.1 离群值处理

为观测其数据总体情况，做数据可视化如下。

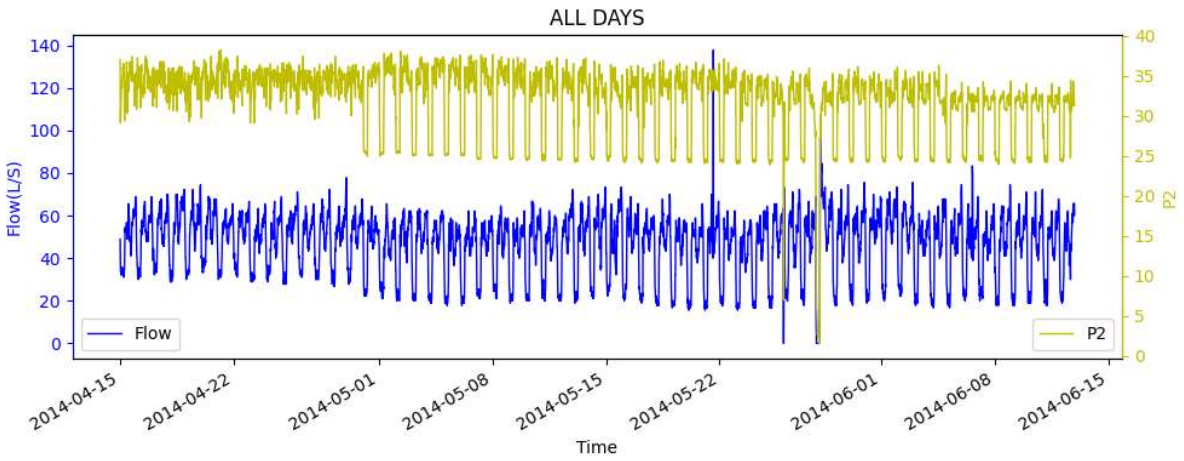

图 1 初始总体数据可视化

发现初始数据中存在部分数据非常大，属于离群数据，对其做下处理。

由于数据服从的分布未知，故使用箱线图识别异常值，更为客观，它是一种用作显示数据分散情况资料的统计图[1]。将该区域 2014/4/15 00:00:00 ~ 2014/6/12 22:00:00 的瞬时流量数据和压力数据按从小到大排列后位于中间的数据为中位数，下四分位数则是位于排列后序列的前 25%的末位数据上，上四分位数则是位于排列后序列的后 25%的首位数据上，如下所示。

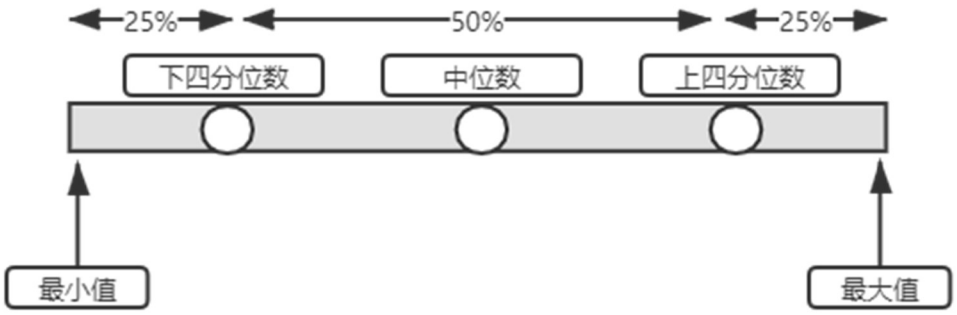

图 2 箱线图数据序列

反映数据的离散程度定义指标 IQR 为:

$$IQR = \text{上四分位数} - \text{下四分位数}$$

它是上四分位数和下四分位数的距离，反映了中间 50%数据的离散程度，其值越小,说明中间的数据越集中；数值越大，说明中间的数据越分散。

定义控制区间为:

$$[\text{上四分位数} - k \times IQR，\ \text{下四分位数} + k \times IQR]$$

当$k = 1.5$时，处于该区间内的数据为正常值，在区间外即为离群值。当$k = 3$时，处于该区间外的数据为极端离群值。

瞬时流量和压力数据序列使用 Matlab 利用箱线图识别异常值，得到结果，存于支撑材料表 1 中。

时间	流量
2014/5/21 15:00	137.78
2014/5/25 23:00	0
2014/5/25 23:15	0
2014/5/27 23:30	2.22
2014/5/27 23:45	0
2014/5/28 0:00	0
2014/5/28 4:30	3.33
2014/5/28 5:45	91.11

表 1 k=1.5 箱线法瞬时流量部分离群值

时间	压力
2014/5/25 22:45	13.41
2014/5/25 23:00	7.13
2014/5/25 23:15	6.97
2014/5/28 0:30	5.54
2014/5/28 0:45	4.36
2014/5/28 1:00	2.78
2014/5/28 1:15	2.7
2014/5/28 1:30	2.82

表 2 k=3 箱线法压力部分离群值

由于工作日、周末、节假日的用水模式存在不同，对离群值做如下修正:

1. 离群值为工作日，使用该工作日前一天的同一检测时间点的数据进行修正；
2. 离群值为周末，使用该工作日前一周周末同为周六或周日的同一检测时间点的数据进行修正；
3. 对离群值的日期分析后未发现离群值存在是节日的情况。

基于以上规则，对该区域进行数据修正。

5.1.1.2 缺失值处理

观察数据后发现，2014 年 6 月 12 日无 22:00 后的瞬时流量和压力数据，且该天为工作日，使用在 5.1.1.1 处理后的 2014 年 6 月 11 日 22:15 ~ 2014 年 6 月 12 日 22:23:45 的数据进行填充，更改后的数据存于支撑材料的表 2 中。

5.1.1.3 异常值处理

问题一中需要辨识区域的典型用水模式，根据题目可认为瞬时流量包含两部分水量：(1)用户用水量之和；(2)管网漏失的水量。而夜间(2:00~5:00)的最小流量可反映该区域的漏失水平。可得，

$$\text{用户瞬时流量之和} = \text{区域瞬时流量} - \text{管网漏失瞬时流量}$$

将 5.1.1.2 中处理后的区域瞬时流量序列使用 excel 求取每日夜间最小流量，并将各天各监测时间点的瞬时流量减去该天的夜间最小流量得到用户瞬时流量之和序列。

发现该序列中有部分值为负的情况，而用户瞬时流量不存在为负的可能。故认为这些值为异常值，且观测后发现异常瞬时流量都较为接近 0。在当日夜间最小流量固定

的前提下，更改该区域该时间点的瞬时流量为夜间最小流量，完成数据处理后的数据存于支撑材料表 2 中。

5.1.2 模式系数的计算

对上述处理后的数据进行如下编号。

i表示第i天，从所给材料中第 1 天(2014/4/15)起，按时间顺序对其编号知道最后一天(2014/6/12)为止，共 58 天，故i∈{1,2,3,…,58}。

j表示每天的第j个用户用水瞬时流量数据，从 00:00 起到 23:45 为止，每 15 分钟测量一次，共 96 次，故$j \in \{1,2,3,\dots,96\}$。

则x_{ij}第i天第j个用户用水瞬时流量数据。

模式系数是 24 小时内用水的变化因子[2]，其计算方法如下：

$$k_{ij} = \frac{x_{ij}}{\sum_{j=1}^{96} x_{ij}}，\ \forall i \in \{1,2,\dots,\} \tag{1.1}$$

其中k_{ij}为第i天第j个检测点的模式系数。

5.1.3 中位数法

由于每一区域内每周的用水模式并非完全相同，为找到典型数据考虑对一个时期内的模式系数数据进行横向处理，将所有数据按照其属于星期几分类。得到周一至周期的 7 个横向数据集合，对每一个数据集合内的不同日期的数据，针对某一时段，找出其中位数的模式系数。其意义在于使得结果免受过高或过低数据的影响破坏它的代表作用，并表达数据的集中趋势。

5.1.4 问题的求解

使用 Matlab 和 Excel 计算该区域每天各检查点的模式系数存放于补充材料表 3 中，将同为周一的相同时间点的数据进行排序，计算中位数，将其作为周 1 该时段的典型模式系数，依次计算周一各个时间点的中位数，同理计算周二到周日的典型模式系数，将结果存于补充材料表 4 中，并绘制曲线，通过分析曲线的变化规律得到每天各时段的用水模式，绘制典型用水模式趋势图进行分析。

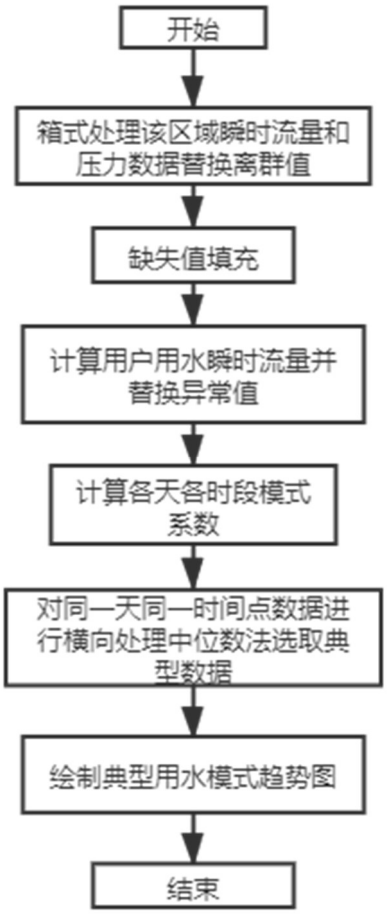

图 3 问题 1 算法流程图

5.1.5 用水模式分类

查阅资料[3]得到典型用水模式分类，并查找数据绘制其用水模式趋势图如下。

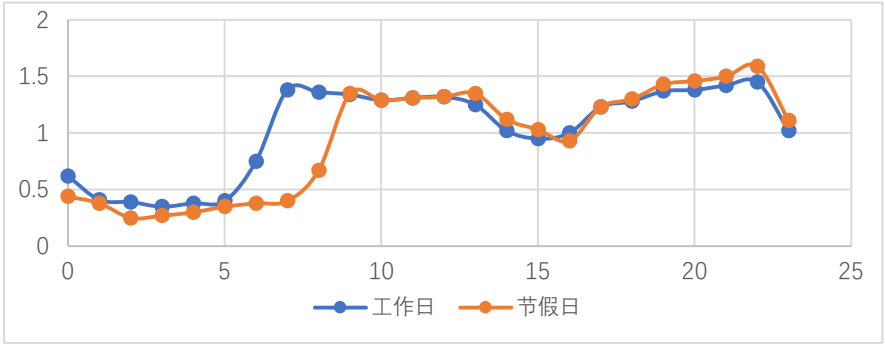

图 4 居民生活类

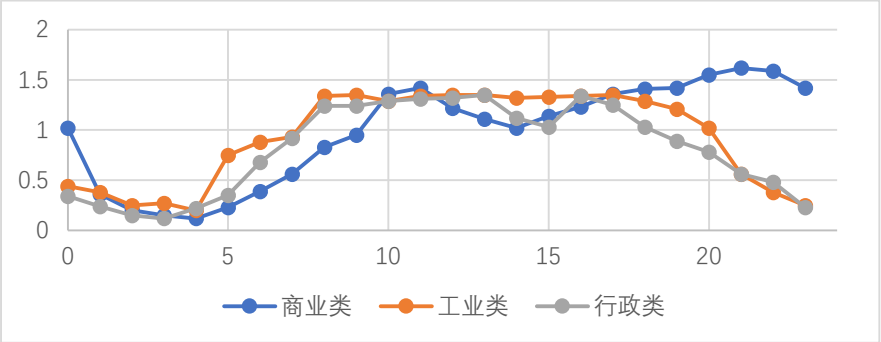

图 5 商业-工业-行政类

分析图片得到各典型用水模式的特征如下:

表 3 典型用水模式特征

典型用水模式	特征
居民生活	工作日与周末的用水量出现交错情况，用水高峰期早晨(7:30 左右)和夜晚(22:00 左右)在周末延后，0~8h 用水量周末少于工作日，9~23h 用水量周末多于工作日，节假日在 8~12h,13~18h 用水会多于工作日。
工业	未发现明显周末与工作日的区分，但从 6:00 点开始用水量出现回升，此后直到 20:00 都保持较高的用水水平，自此后开始下降。
机关单位	一周中各天的用水变化趋势相似，呈现 0:00 到 6:00 用水量较小且几乎不变，6:00 到 8:00 快速上升，8:00 到 12:00 用水缓慢上升，13 点后少量减少，但在 18:00 后呈下降趋势。
商业	从 7 点后逐渐上升但未出现向居民生活类一样的用水高峰期，从 12:00 左右开始下降从 16:00 左右再次回升一直保持较高用水水平知道 22:00 左右开始下降。

5.1.6 问题的结果及分析

得到典型用水模式趋势图如下。

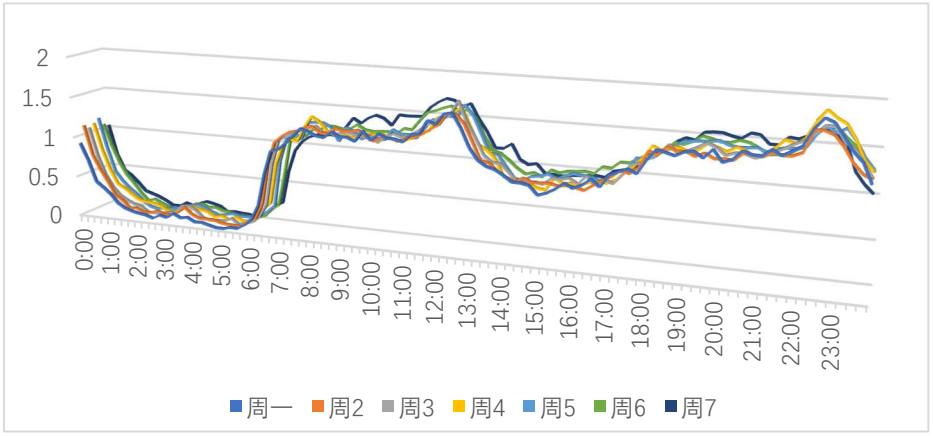

图 6 典型用水模式趋势图-各天

根据图 4 可得该区域的用户用水模式具有较强的日周期性，即一周之中各天用户

的用水都基本呈现出从 22:30 左右开始下降到 6:00 左右重新开始回升，6:00 到 8:00 快速上升，8:00 到 12:45 左右持续上升，12:45 到 15:00 左右下降，15:00 到 23 点持续上升的趋势；2:00 到 7:00 用水量几乎为 0：周末在 9:00 到 15:00，19:00 到 22:30 点的用水水平高于工作日，而用模式系数在 2:00 到 7：30，12:00 左右，18；30 左右，22:30 左右都存在较高重合性。

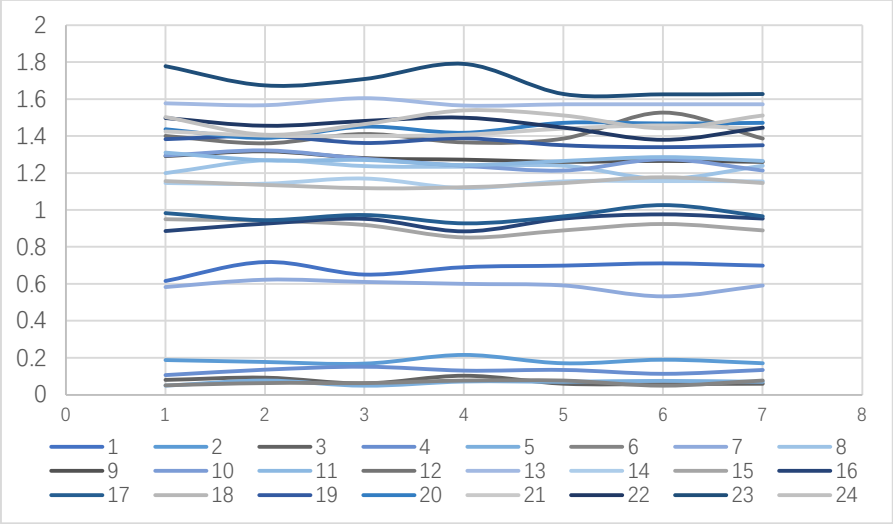

图 7 典型用水模式趋势图-各时段

根据图 5 分析该区域各时间点在一周中各天的分布情况，除 23：00 左右工作日的用水量高于周末；12 点用水周末多于工作日，17:00 点周末高于工作日其他时段外，一周中各天用水情况相差不大。

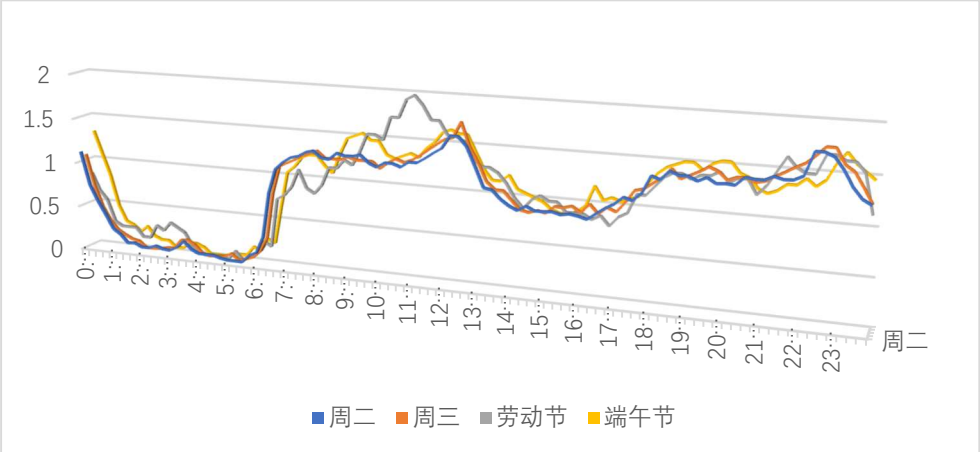

图 8 DMA1-节日工作日对比

分析可得 DMA1 区域节假日在 8:40 到 12:30，0:00 到 3:45 的用水水平高于工作日。

经过以上分析综合 5.1.5 该区域用水分布基本符合居民生活用水类特征，18~23h 用水再次上升，用水峰值(8:00,12:00,23:00)在周末延后，节假日在 8~12h 用水多于工作日，故可基本判别该区域用水模式属于居民生活用水模式。

5.2 问题 2 的求解

5.2.1 不同地区漏失水平划分

查阅文献[5]发现夜间最小漏水量可由如下部分组成。

表 4 夜间最小用水量组成

夜间最小用水量	用户夜间用水	
	漏损水量	表观漏失量
		物理漏失量

基于假设 3，认为夜间最小用水时刻无用户用水，且在问题一数据处理后认为不存在数值记录错误的情况，可认为夜间最小用水量即为物理漏失量。

问题分析中提到，我们认为该地区的漏失量模式，即为该地区的漏失速率等级。由于题目条件限制，且查阅资料[]后发现，各个地区对漏失等级的划分不同，如下所示。

表 5 物理漏失量分类评定

	技术指标	ILI	不同压力下的物理漏失量				
			10m	20m	30m	40m	50m
高等收入国家	A1	<1.5		<25	<40	<50	<60
	A2	1.5~2		25~50	40~75	50~100	60~125
	B	2~4		50~100	75~150	100~200	125~250
	C	4~8		100~200	150~300	200~400	250~500
	D	>8		>200	>300	>400	>500
低等或中等收入国家	A1	<2	<25	<50	<75	<100	<125
	A2	2~4	25~50	50~100	75~150	100~200	125~500
	B	4~8	50~100	100~200	150~300	200~400	250~500
	C	8~16	100~200	200~400	300~600	400~800	500~1000
	D	>16	>200	>400	>600	>800	>1000

且 ILI 的计算考虑了管线长度，供水时间、服务支管的数量等，而问题中并未给出该区域上述数据，故不能无法通过 ILI 作为漏失量分类评定的标准，只能使用漏失速率和压力作为漏失量分类评定的标准。

5.2.2 压力与漏损水量的关系

问题分析，漏失速率数据变化为一天一次，数据较少。且查阅资料[4]可知，管网运行压力与漏损水量呈指数关系，因而管网的平均压力越大，所造成的漏损水量就越大，并且发生漏损的概率就越大。管网各个节点压力 24 小时不断变化，所以管网任意时刻的漏失速率都为夜间最小流量。

将管网的漏点看作孔口，可得到水流速与压力的关系，表示如下。

$$v = Cd \times (2gP)^{0.5} \tag{2.1}$$

其中 v 为漏损速率，Cd为漏损系数，P为孔口处压力，g为中立加速度。

由于Cd并非常数，因此可以将漏失速率与压力的公式改写为如下形式。

$$\frac{v_1}{v_2} = \left(\frac{P_1}{P_2}\right)^n \tag{2.2}$$

其中P_1为t_1时刻的管网平均压力，P_2为t_2时刻的管网平均压力；v_1为t_1时刻的漏损速率，v_2为t_2时刻为漏损速率；n 为压力指数取值范围为 0.5~2.5，依赖于管道的管径，管材等具体情况而定，查找文献发现当对管道具体情况掌握较少是可取 n 为 1.18[5]。

由于问题仅给出一个压力数据序列，基于假设 4，认为这个压力数据能够代表管网中的平均压力。

5.2.3 任意时刻漏损速率的估算

根据夜间最小流量时刻(MNF)的漏损速率以及该点的压力变化，变化式(2.2)可求得管网中任意时刻的漏损速率如下。

$$v(t) = (\frac{P(t)}{P(t_{MNF})})^n \times v(t_{MNF}) \tag{2.3}$$

其中$v(t)$为t时刻的漏损速率，$P(t)$为t时刻的漏损速率，t_{MNF}为夜间最小时刻漏损速率。基于以上公式可得出各天各时刻的漏损速率。

5.2.4 KANN-DBSCAN 聚类

根据问题分析，本问题中所给数据并未指出数据来源地区。为得到更加科学且贴合实际的划分标准，我们希望利用数据本身的特征，通过聚类的办法得到根据漏失速率大小聚集而成的聚类簇，根据每个聚类簇代表的实际意义得到漏失水平的等级划分标准。

DBSCAN 是一种基于密度的聚类算法，它将将具有足够高密度的区域划分为簇，并可在存在噪声的空间数据库中发现任意形状的簇[6]。其存在两个重要参数：$MinPts$和Eps。前者为定义密度时的领域半径，后者为定义核心点时的阀值。

其核心思想可描述为:

从某个选定的核心点出发，不断向密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。

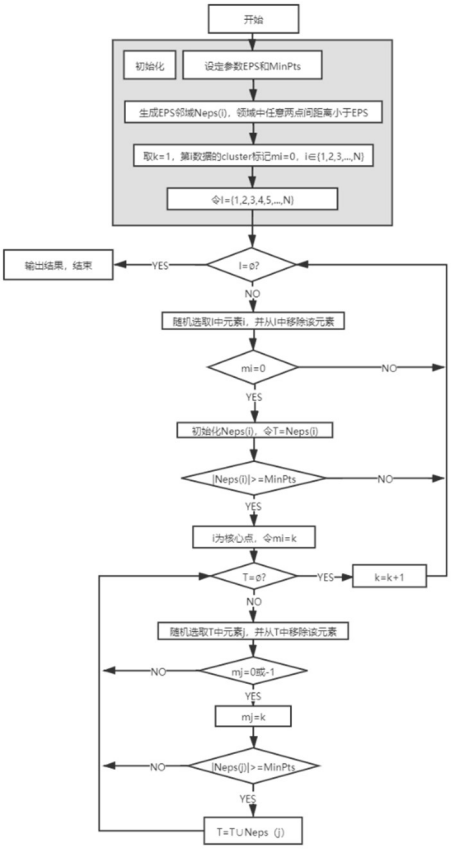

图 9 DBSCAN 算法流程图

由于 DBSCAN 算法需要Eps和$MinPts$两个参数来表示数据点分布的紧密程度，判断是否有足够的密度来划分簇，形成满足密度相连点的最大集合，故该算法对于Eps和$MinPts$两个参数的选取十分敏感，通过多次人工取值手工找到最佳取值不太可能。故查找文献[7]后找到了一种能够自适应确定 DBSCAN 算法参数的 KANN-DBSCAN 算法，他利用数据集自身的分布特征生成候选Eps和$MinPts$参数集合，并寻找数据集对应的最优 DBSCAN 算法参数。

定义密度阈值$Density$为以Eps为半径的圆内存在$MinPts$个数据点，公式为:

$$Density = \frac{Minpts}{\pi \cdot Eps^2} \tag{2.4}$$

5.2.4.1 生成Eps参数列表

采用$K-$平均最近邻算法和数学期望法生成Eps列表。具体过程如下:

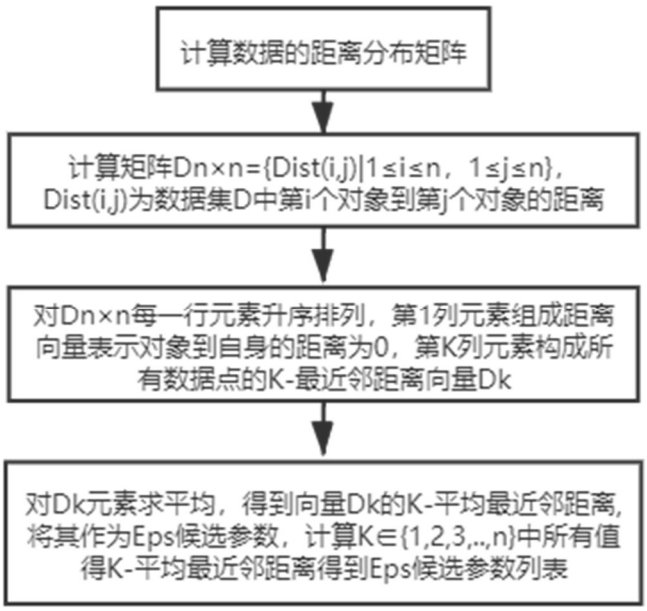

图 10 Eps参数列表生成

5.2.4.2 生成$MinPts$参数列表

采用数学期望法生成$MinPts$参数列表。对Eps参数列表，依次求出每个Eps参数对应的Eps领域对象数量的数学期望，作为数据集 D 的领域阈值$MinEps$参数。

$$MinPts = \frac{1}{n}\sum_{i=1}^{n} Neps(i) \tag{2.5}$$

其中$Neps(i)$为第i个对象的Eps领域对象数量，n为数据集D对象数量。

5.2.4.3 自适应确定最优参数

依次选用根据上述做法求出数据集 D 的Eps参数列表中元素作为候选Eps参数和有公式(2.5)得到$MinPts$参数，输入 DBSCAN 算法对数据进行聚类分析，分别得到不同 K 值下所生成的簇类。当生成的簇类连续三次相同时，认为聚类结果趋近于稳定，记簇

数 N 为最优簇数。

继续执行上述步骤直到簇数不为 N，并选用簇数为 N 时所对应的最大 K 值作为最优 K 值。最优 K 值对应的 K-平均最近邻距离作为最优Eps参数，最优 K 值对应的$MinPts$参数则为最优$MinPts$参数。

5.2.5 聚类结果及分析

根据以上算法，首先对 5665 条数据基于式(2.3)求解各时刻的漏失速率存与补充材料表 5 中，由于该漏失速率与其所属时刻的水压相关，即当漏失速率小时水压小，故仅对漏失速率聚类。使用 Python 基于 5.2.4 中的 KANN-DBSCAN 算法，在Eps为 0.53 时，得到最优簇数为 3，继续执行得到最优$MinPts$为 68，得到聚类结果如下。

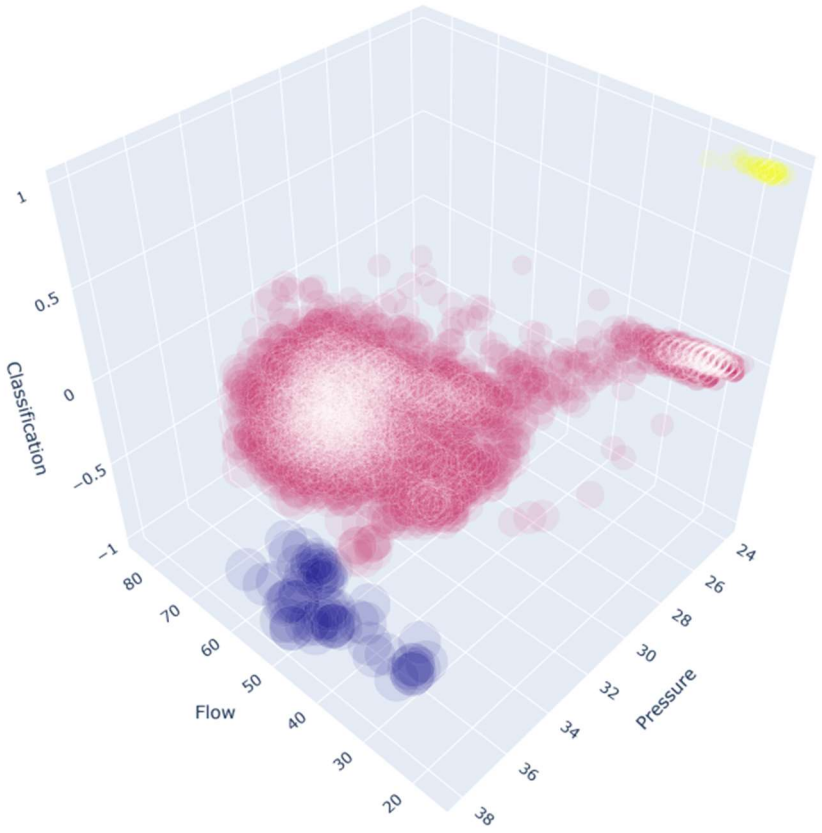

图 11 DBSCAN 聚类结果

分析各簇中具体数据的意义分析得到，簇 1 的夜间漏损速率较小，水压也小，定为等级 C。簇 2 的夜间漏损速率的范围较广，但与簇 1 相比较大，水压也相对较大，定为等级 B。簇 3 的夜间漏损速率较大，水压也大，在 DBSCAN 算法中其 cluster 值为-1，为噪声点，但结合其实际取值和本问对于聚类算法的应用认为这部分数据的是该区域夜间漏损速率的上限值，不做去除处理。

表 6 该 DMA 漏水量等级评定

等级	夜间漏损速率范围	数量	意义
A	*35.39552~38.61102*	64	大量漏水，水压大
B	*15.97309~35.29119*	5534	中量漏水，水压中等

C	*15.22290~15.88115*	69	少量漏水，水压小

六、模型评价

七、模型推广

八、参考文献

附录