

## 基于多源数据的管道安全性评估

### 摘 要

在目前国内缺乏管道事故报道数据库的现状下, 本文构建了生命线管道事故数据系统, 其集成了城市燃气管道事故报道收集、事故归因统计分析以及图形化展示三大功能, 在该领域做出了有建设性的尝试; 在管道的分段安全性能评估问题上, 本文通过数据清洗、特征工程和算法处理三步骤, 在来自某工业园区 GIS 系统的管道属性数据集上比较了基于 K-means、层次聚类、DBSCAN 和谱聚类算法的管段聚类性能, 选取层次聚类作为最佳算法并实证检验了其能有效地区分不同风险状态的管段; 在如何利用 SCADA 系统实时监控管道运行状态的工程问题上, 本文提出了基于 DTW 距离度量的两层时序聚类模型, 从 2015 年内超过 16000 条序列中提取了 10 条典型序列, 并结合典型序列样本和事故记录报告构建了基于高风险日内走势识别的实时异常检测模型。实证检验该模型能成功识别 2016 年内事故地测点序列中的 86%, 证明了该模型的有效性和易用性。

**关键词:** 燃气管道, 安全评估, 爬虫, 无监督聚类, 异常检测

# **Risk Evaluation of Gas Pipeline based on Unstructured Multi-source Data**

## **ABSTRACT**

In the circumstance of the lack of domestic urban gas pipeline accident report database, this paper constructed a gas pipeline accident data system which consists of three main components including web crawler of accident reports, statistical analysis on failure inductions and results graphical display. It fills the void and is a constructive attempt.

Before applying clustering algorithms on pipeline datasets from the GIS, this paper conducts data cleaning and diverse feature engineering, and then compared the performance of K-means, Hierarchical Clustering, DBSCAN and Spectral Clustering on these datasets, selected Hierarchical Clustering as the best. The clustering result passed Chi-square test, being proved to be able to distinguish risk states of gas pipelines. This clustering model brings new insights in risk evaluation of gas pipelines.

This paper also built two-layered clustering model on time series, which were collected by the SCADA, and through which chose 10 typical series from more than 16,000 series in 2015. On the basis of series matching with the failure reports, 2 risky series were found and an anomaly detection model was created. At last, this paper tested this model on series in 2016 by which proved being capable of digging out 86% of risky series. That result certifies the validity, flexibility and implementation simplicity of the anomaly detection model based on risky series matching.

**Key words:** gas pipeline, risk evaluation, web crawler, unsupervised clustering, anomaly detection

## 目 录

1 引 言.....	1
1.1 问题的提出.....	1
1.2 课题背景.....	1
1.3 研究目的与意义.....	2
1.4 国内外研究文献综述.....	3
1.5 主要工作.....	6
2 生命线管道事故数据系统的设计与开发.....	7
2.1 国内外研究背景.....	7
2.2 事故报道收集的爬虫系统设计.....	9
2.3 事故原因的统计与分析.....	12
2.4 开工项目信息收集与可视化.....	14
2.4.1 图形化定向爬虫软件开发.....	14
2.4.2 项目地理信息的可视化.....	16
2.5 燃气管道事故信息化系统.....	17
3 生命线分段管道的风险评估模型.....	21
3.1 无监督聚类.....	21
3.2 特征工程.....	22
3.3 基于多种静态特征的管道聚类.....	23
3.3.1 数据清洗.....	23
3.3.2 特征工程.....	23
3.3.3 聚类算法简述.....	31
3.3.4 算法性能比较与评估.....	33
3.4 管道事故的地理分布.....	37
3.5 聚类簇的风险评估.....	38
4 SCADA 测点序列聚类及实时异常检测模型.....	41
4.1 时序集的交叉距离矩阵.....	41
4.2 序列的两层聚类模型.....	43
4.3 测点序列聚类簇的风险分析.....	47
4.4 测点实时数据的异常检测模型.....	49
5 总结与展望.....	53
5.1 本文所做工作的总结.....	53
5.2 下一步研究方向的展望.....	54
参考文献.....	55
谢 辞.....	57

## 1 引言

### 1.1 问题的提出

近年来，全球能源消费呈现低碳化趋势，天然气以其清洁、高效和高储量成为了支撑该趋势发展的重要支点，其发展速度显著高于煤炭和石油。根据英国石油公司 2014 年的报告《BP2035 世界能源展望》预测<sup>[1]</sup>，2012-2035 年全球天然气年均需求量增长速度为 1.9%，至 2035 年占全球能源消费比重将达到 26% 左右。欧洲燃气管网灾害数据协会（EGIG）的近期报告表明，截止到 2016 年欧盟成员国天然气管道总长已达 142794 km，其历年管道总长以及管道修建年份见图 1.1。根据我国的“十三五”规划报告，到 2020 年全国长输管网总规模将达 15 万千米左右，输气能力将达到 4800 亿立方米/年左右。随着生命线管道系统规模的飞速扩张，旧有管道与新管道同时存在，如何准确评估现有生命线系统中不同区段管道的安全性问题并构建精确的预警模型这一问题也越来越被学界与业界所重视。

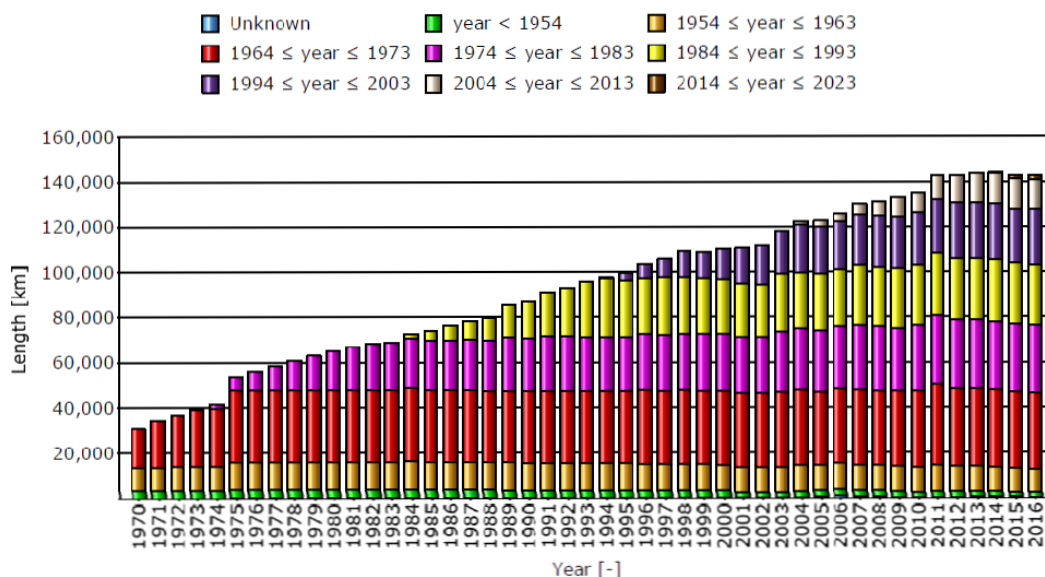


图 1.1 欧盟成员国历年管道总长以及管道修建年份图<sup>[18]</sup>

### 1.2 课题背景

城市燃气管道作为现代城市公共基础设施的一部分，已成为现代城市不可或缺的基础设施之一。为满足居民和企业不断增长的能源需求，近年来城市燃气管网规模正不断扩张。2004 年，中国油气管道总里程还不到 3 万公里，至 2016 年底已经达到总里程 11.64 万千米，其中天然气管道 6.8 万千米。在我国推进天然气管道的建设具有以下几大优势<sup>[9]</sup>：

- 中国天然气资源潜力大
- 中国天然气供应能力快速增长
- 中国天然气消费增长迅速
- 中国天然气管网基础设施建设处于快速发展期
- 中国天然气已形成完整产业体系

2015 年, 中国人均天然气消费量约 140 立方米, 天然气占一次能源消费总量的比重约 5.9%, 远低于世界平均水平 23.7%。随着国家绿色低碳能源战略的持续推进, 发展清洁能源将成为优化能源结构的重要途径。

随着中国城镇化建设的深入推进, 城镇化率稳步提升, 预计到 2020 年将达到 60%。未来城镇燃气发展方向包括三个方面: 一是稳步发展民用气, 提升居民气化水平, 城镇居民天然气化率 2020 年达到 50%~55%; 二是有序发展天然气采暖, 发展城市集中式采暖、燃气空调、分户式采暖; 三是推进重点地区天然气化, 加快煤改气进程。

然而, 人们在享受完整生命线系统带来的便利的同时也在面临其所带来的负面作用。天然气由甲烷、硫化氢与二氧化碳等成分组成, 具有密度低、易膨胀、易燃易爆等特性。因此, 天然气管道一旦发生泄漏, 极易波及整个城市生命线系统, 造成重大人员和财产损失, 并对环境产生威胁。城市用气的不断普及推进使事故发生频率居高不下, 据博燃网<sup>[8]</sup>报道, 2017 年我国共发生燃气爆炸事故 702 起, 造成 1100 余人受伤 126 人死亡。2017 年 7 月 2 日, 中石油天然气输气管道晴隆沙子段, 因持续强降雨引发边坡下陷侧滑, 挤断输气管道, 引起泄漏燃爆, 造成 8 人死亡, 35 人受伤, 其中重伤 8 人, 危重 4 人; 2017 年 7 月 4 日, 松原市宁江区建设街市人民医院后侧道路在维修排污管线过程中发生燃气管道泄漏, 燃气公司在抢修过程中发生爆炸, 致使临近的市人民医院也受到波及。造成 5 人死亡, 89 人受伤, 其中重伤 14 人。

目前城市燃气管道气源种类多, 设计标准不统一, 施工质量参差不齐, 老旧管道老化严重并与新修管道同时使用, 致使整个生命线系统可靠度下降。城区地下管道所处地带人员密集, 经济活动频繁, 燃气事故会造成重大的人员伤亡和财产损失。另外, 由于管网规模巨大, 传统人员巡视的方式因为效率低下越来越无法满足安全要求, 应运而生了针对现代化生命线管道运维管理系统开发的需求。现代生命线管道系统的运维管理将涉及多方面的数据, 包括人工检测形成的报告、记录、表格、日志、图片, 常规 SCADA 系统中压力、流量数据, 管网 GIS 的矢量数据、设施的 BIM 数据, 以及物联网架构下的安全监测数据等。如何综合运用结构化和非结构化的多源数据建模, 并对管网安全性进行实时准确监测便成为了中心议题。

## 1.3 研究目的与意义

风险(Risk)通常被定义为可能导致损失的事件的发生概率和潜在的损失规模<sup>[5]</sup>, 风险评估的目的主要是发现可能存在的事故隐患及原因, 分析事故的后果。为了及时发现事故隐患, 或是预测事故的发生概率, 我们需要一套进行风险评估的工具。

风险评估的方式可分为模型法和模拟仿真法, 模型法是对真实过程的简化, 用以方便人们对事件的理解。而仿真法则寻求尽可能真实地还原现实, 可能会导致模型的实用性和可解释性降低。虽然二者没有绝对的优劣之分, 但是模型法以其更好的可解释性以及高效性受到大部分现有的风险评估的研究的亲睐。一般的风险模型方法有以下几种:

### (一) 矩阵法(Matrix)

决策矩阵是最简单的评估方法，它将事故发生概率和后果按分级打分的形式综合起来，因此高发生概率以及高破坏性的事件会在结果中突出表示。该方法通过将事件的概率以及后果分开分析，给与评估人员简洁直观的评估结果。

## （二）概率法(Probabilistic)

概率评估模型(probabilistic risk assessment,PRA)，亦被称作 quantitative risk assessment(QRA)或者 numerical risk assessment(NRA)。PRA 是一种基于数学与统计学的模型，它依赖于对历史数据的统计建模分析，通过事件树或失效树(event-tree/fault-tree)将事故流程化并量化各个分支事件的概率分布，最终的事故概率由对所有单个事件逐层链式条件概率估计得到。

概率法能够得到事故失效的绝对风险评估(absolute risk assessment)，因此其预测结果可用于异类事故间的比较，比如比较火灾风险和水灾风险等。概率法能够尽量消除人的主观因素带来的偏差，得到更具综合性与科学性的评估结果，但它较其他方法对计算能力和数据量要求更加严苛，其应用一直受到一定限制。但随着近年来大数据与机器学习技术的兴起，PRA 将逐渐成为未来风险评估的热门方向。

## （三）指数法(Indexing)

指数法通过对系统一系列的重要状态因素进行打分来评估系统不同部分的相对风险水平(relative risk)，其不同因素的权重分别由统计方法和工程经验确定，反应了定性与定量方法的结合。指数法综合考虑了多种因素和多种模型，能提供对系统风险水平较为全面的视角，但它仍然存在主观性评分的影响，并且不能保证因素权重水平反映了真实的重要度。

综上，在拥有管网系统现代化细粒度的多源监控数据的条件下，通过建立风险模型评估管道风险水平并针对不同的管段建立预警机制，多管齐下致力于降低管网系统事故风险是本文研究所要探讨的中心问题，也是本研究的意义所在。

## 1.4 国内外研究文献综述

传统的管道安全维护是经验性的，主要根据安全生产管理条例进行反应式的维护，缺乏预见性防范安全事故发生的能力。20 世纪 70 年代，美国的 PRCI(Pipeline Research Committee International) 首先开始引入风险分析技术评价油气管道的风险性，其基于美国和欧洲的管道事故数据归纳总结出 22 种引起管道失效的因素；1985 年美国的 Battelle Columbus 研究院出版了《风险调查指南》，运用评分法评价管道风险；1992 年 W.Kent.Muhlbauer 编写的《管道风险管理手册》详细叙述了管道风险评估模型和各种评估方法，被各国接受作为开发风险评估软件的指导文献。在工程应用上，自 20 世纪 90 年代起，美国石油协会已出台了一系列风险评估的规范，包括 ASME B31.8S 和 API PR581 等。之后，北美以及欧洲发达国家均开始了管道风险管理技术的开发。英国 Advantica 公司经过统计分析大量管道资料以及灾害模拟试验，量化了天然气管道的危害因素以及事故后果，建立了专家分析软件 PIPESAFE；加拿大 C-FER 公司则开发了 PIRAMID 软件包用于分析管线失效概率、失效后果和总风险计

算。经验表明，管道风险分析的引入产生了巨大的经济效益和社会效益。例如，美国 Amoco 管道公司在 1987 年应用风险评价技术后，其管道和储罐泄露率显著下降，大幅提高了该公司的利润水平。

目前对生命线管道安全性的分析方法大体可以分为定性、半定量和定量两种。定性评估主要是针对管网事故特点和类别，借助指标体系进行风险评估；定量评估主要是针对管网自身特点，借助相关物理模型和计算方法，定量评估管网的风险。

## (1) 定性评估

目前国内外的定性评估方法主要分成两种：

**主观赋权评价法：**由经验丰富的专家根据自身经验和知识积累主观判断并得到相应比较权重。这类方法包括层次分析法、模糊综合评判法等；

**客观赋权评价法：**由安全管理人员根据各指标间的相关关系或变异系数来确定权重，例如事故树法、故障树法、灰色关联度法、TOPSIS 法、主成分分析法、数据包络分析法等。

## (2) 半定量评估

半定量法以风险的数量指标为基础，对识别到的事故首先为事故发生后果和事故频率各分配一个指标，然后对两个对应事故概率和严重程度的指标进行组合形成一个相对风险指标，其允许将风险量化为标准的等级。W.Kent Muhlbauer 提出了肯特指数法，综合了以定性法的以图表为基础的 HAZOP 模型和定量的事故概率分布，排除了不可预见的事故后果，使人们集中关注可能发生的事故后果上，提高了风险评估的实用性和准确性，被广泛接受为长输管道的风险评估方法<sup>[5]</sup>；

## (3) 定量评估

定量法是一种定量分析事故概率分布的统计建模方法。通过定量分析各种事件的概率分布，计算出导致安全性问题的最终发生概率和损失后果。目前，还没有广泛认可的管网风险整体评估模型，已有的研究主要集中在可能性分析、后果分析和失效传播分析等方面。

传统的研究方法大都集中在定性方法上，其中的代表有：Fred Henselwood<sup>[10]</sup>提出了一种基于矩阵法对灾害风险性的评估方法；王晓梅<sup>[2]</sup>采用模糊层次分析法对埋地管道失效各因素的权重进行了评估；郭峰<sup>[3]</sup>建立了城市地下管道失效事故树，采用布尔代数法确定了事故树的最小割集和最小径集；韩朱旻<sup>[4]</sup>采用灰色关联分析法，通过计算影响管道风险的各种评价指标经模糊综合评判处理后的评分序列与最优指标序列之间的关联度，对各级指标的权重进行了计算；韩朱旻<sup>[4]</sup>利用 Bathtub Curve 和可靠度、失效度、失效密度、修复率等进行了失效率的定量计算；Jaffee 等利用有限元建模对燃气泄露爆炸范围模拟仿真并基于此对燃气储运系统进行了爆炸后果评估<sup>[6]</sup>；黄超等建立了城市燃气管网失效传播模型，对事故在管网内部的传播过程进行了分析计算<sup>[7]</sup>。

基于目前生命线管道网络信息化的要求，未来的研究方向应逐渐加强定量法的研究，尤其是对事故可能性的定量分析。因为管道事故的发生往往在某一地点，这就必然涉及对管道分段分类的问题，传统基于指标法并人工分段的方法存在许多弊端，有很多研究已经针对定量方法分段展开：李大全<sup>[11]</sup>采用模糊聚类法选取管道属性对象构建模糊相似矩阵，通过传

递归包法进行管道动态分段，但该方法仍然存在人为确定参数的问题；张杰<sup>[13]</sup>采用主成份分析法(PCA)对管道多属性进行降维和提取，减弱了指标体系主观性过强的问题，再利用 K-Means 方法基于降维后的管道属性特征集对管段进行聚类分析，给出了不同管段的分类方法。但该种分段模式存在的问题，PCA<sup>[12]</sup>是一种线性降维方法，会导致降维过程忽略非线性关系，并且容易忽略贡献率小但对类别差异有重要影响的变量，另外 K-Means 聚类对初始化中心点位置可能比较敏感，导致分段结果具有一定的随机性。

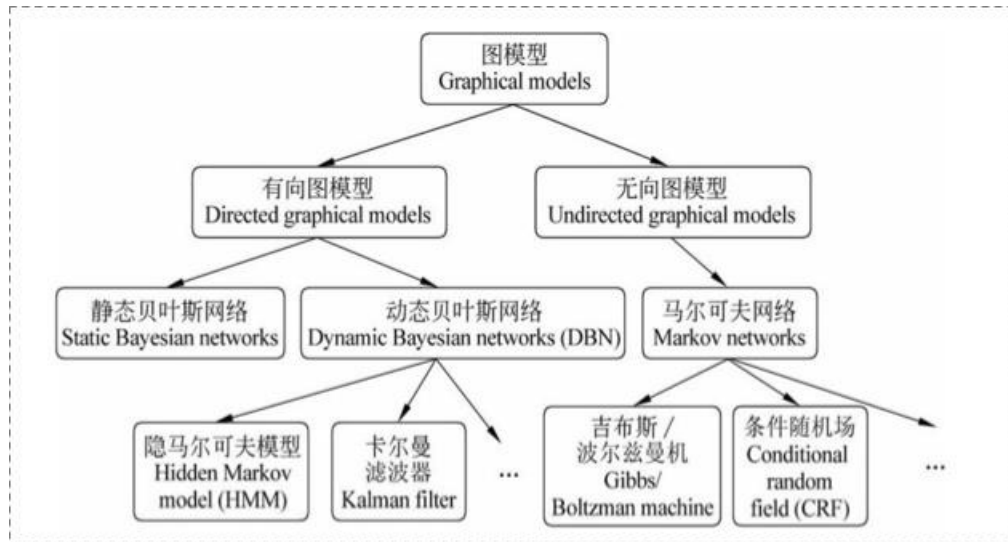


图 1.2 图模型(Graphical Models)的层次关系<sup>[19]</sup>

除了管道分段，定量分析发生可能性也是研究重点。Alireda Aljaroudi 等<sup>[16]</sup>针对管道时序数据，引入了电气工程中常用的信号分析方法，及时捕捉异常信号从而做到管道的安全监测；Jiansong Wu 等<sup>[14]</sup>提出一种基于静态贝叶斯网络的事件发生概率分析模型，该法首先采用 Dempster-Shafer 法对专家评分进行特征赋权，将事故分解为多级事件并以此构建贝叶斯网络，依据多级事件的条件概率传递对事故风险进行了建模分析。静态贝叶斯网络是图模型的一种，图模型的层次由图 1.2 所示。但是静态贝叶斯网无法进行增量学习，即其不能根据新输入的数据实时更新模型，因而限制了模型的效率提升，另外该模型仍然没有脱离半定量方法的范畴；王晓波等<sup>[15]</sup>进一步建立了事故树事件分割加动态贝叶斯网的管道事故快速辨识及溯源分析方法，克服了静态贝叶斯网络的缺陷。但是，贝叶斯网是一种有向图模型，其隐含了事件具有独立性和事件之间单向依赖的假设，这并不完全符合现实情况；Jingwei Qiu 等<sup>[17]</sup>结合了深度神经网络(DNN)和隐马尔科夫链式模型(HMM)构建了燃气管道压缩单元的预警监测模型，该模型的精妙之处在于利用 DNN 强大的特征提取能力，从原始数据中提取进一步输入 HMM 的特征，再利用 HMM 强大的序列标注能力预测系统的隐藏状态 (Hidden State)。但该模型只能利用时间序列型数据，对其它类型数据并没有做到完全利用。

综上所述，本文为综合之前研究方法的可取之处并克服其存在的种种问题，为消除主观性对管道分段的影响，充分利用多源异构数据，提出利用无监督方法进行管道分段的解决方



案；为了提高分段结果的准确性并针对性地建立异常检测模型，本文提取了时序数据特征进行了聚类，对预分类的管道温度、流量监测点建立异常检测模型，科学地完成管道安全性评估和安全监测的工作。

## 1.5 主要工作

通过对现有研究文献广泛阅读，结合现有条件和研究目标，本文拟从以下几个方面开展研究工作：

### (1) 生命线管道事故数据系统

通过广泛收集生命线管道安全事件的新闻报道、深度案例并进行文本信息挖掘，对管道事故的关键词进行统计分析，为进一步分析安全事故诱因奠定基础，并构建生命线管道事故数据系统。

### (2) 管段聚类模型及安全性评估方法

对目前现有的多源异构数据来源进行全面总结梳理，整理各数据的收集方法、数据结构、类别和统计指标等，建立基于多种静态属性的管段聚类模型及安全性评估方法。

### (3) 管道 SCADA 测点数据的时序聚类及异常检测模型

对 SCADA 测点序列数据进行时序聚类，挖掘其典型走势，寻找典型走势中的高风险模式，并基于该高风险模式建立模式匹配的异常检测模型。

### (4) 案例实证研究

选取某市工业园区的燃气输配系统为实例，结合多源异构的安全性评价模型进行管道的安全性评价案例研究；基于燃气公司的事故记录检验异常检测模型的精确性以及管道安全性评估模型的准确性。

## 2 生命线管道事故数据系统的设计与开发

### 2.1 国内外研究背景

国外已经有很多机构构建了完备的管道事故数据库，并定期发布相关事故的统计分析报告。比如美国的 PHMSA，欧盟的 EGIG，英国的 UKOPA 和加拿大的 EUB 等。其中，EGIG 数据库中共有发生于 1970 至 2016 年的 1310 条事故报告。其关于所收集事故的标准如下：

- (1) 事故必须导致异常的泄漏事件
- (2) 管道必须满足如下标准：
  - 金属制管道
  - 陆地上的管道，不含海底管道
  - 最大操作管压需超过 15 Barg
  - 处于燃气中心设施的围护设施之外
- (3) 发生于生产中的燃气事故、泵和压缩机事故均不被记录在案

基于以上标准，EGIG 所统计的事故发生频次趋势见图 2.1 所示。对 2007 到 2016 年所发生事故主要原因进行统计，结果见图 2.2 所示，腐蚀和外力破坏比例接近，但是前者造成的泄漏流量远小于后者<sup>[18]</sup>。

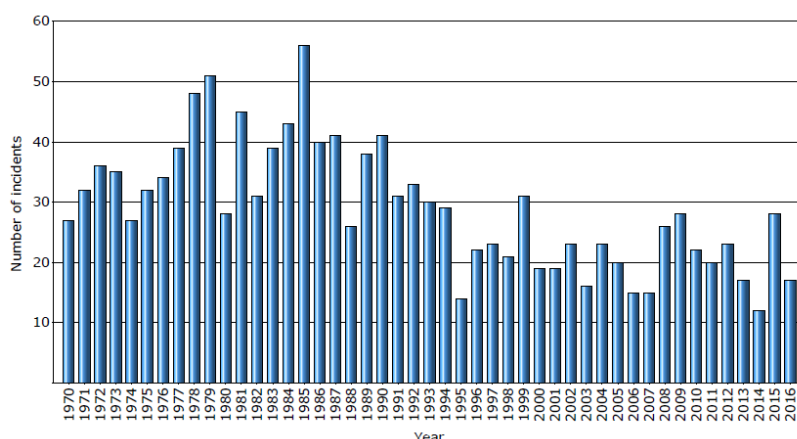


图 2.1 EGIG 统计燃气管道事故平均年发生频次（1970-2016）<sup>[18]</sup>

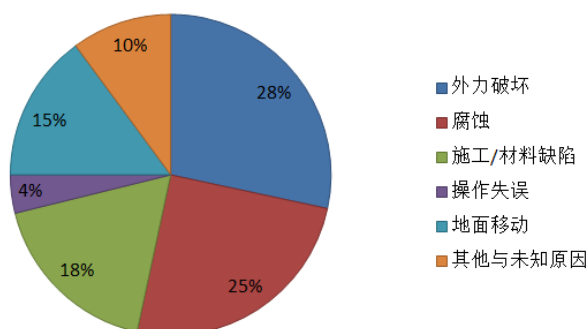


图 2.2 EGIG 所统计管道失效的原因（2007-2016）<sup>[18]</sup>

为更有针对性地设计管道安全监测系统，防患事故于未然，几大机构均对导致管道事故发生的原因进行了归类总结，戴联双等<sup>[20]</sup>对其收集并制作成的表格如表 2.1 所示。

表 2.1 外国不同组织机构对管道失效原因的分类<sup>[20]</sup>

组织机构名称	失效原因分类
PHMSA	挖掘损坏、腐蚀、误操作、材料/焊接/设备故障、自然力损坏、其他外力损坏、其他
EGIG	外部干扰、腐蚀、施工缺陷或材料本身缺陷、误操作、土体移动、其他或未知
UKOPA	外部干扰、内腐蚀、外腐蚀、土体移动、环焊缝缺陷、管道本体缺陷、纵焊缝缺陷、其他、未知
EUB	施工损坏、其他损坏、外腐蚀、内腐蚀、连接失效、超压、管道本体、阀门、焊接、其他

PRCI（国际管道研究协会）对输气管道失效原因进行总结，得出较为详细的分类清单，如表 2.2 所示。

表 2.2 PRCI 总结管道失效的基本因素<sup>[5]</sup>

一、发生变化且变化规律与时间关系密切的因素	1.外腐蚀	
	2.内腐蚀	
	3.应力腐蚀	
二、稳定存在的失效因素	4.与生产制造有关的缺陷	焊缝缺陷
		管材缺陷
	5.与焊接、装配有关的缺陷	管道周向焊接有缺陷
		管道焊接装配有缺陷
		起皱或翘曲
		螺栓掉落/管道破裂/接头失效
	6.与设备有关的失效	衬垫或密封圈失效
		控制/排放设备故障
		密封/捆扎失效
		装备混杂
三、发生变化但变化规律与时间无关的失效因素	7.第三方/机械破坏	由第一方，第二方或第三方引起的破坏（瞬间/随后失效）
		管道先前曾遭受破坏（延时失效模式）
		蓄意破坏
	8.误操作	错误的操作程序
	9.与自然和外力有关的失效因素	严寒
		闪电
		暴雨或洪水
		地层运动

中国官方的城市燃气网络故障报告数据库系统尚未建立。目前，各类城市燃气网络故障事件主要以新闻形式报道在各大新闻门户网站上，无法对其直接进行故障频率的具体原因分析和计算。戴联双等<sup>[20]</sup>收集了 2008 到 2012 年的 523 起城市燃气管道失效事件，并将事件原因总结成第三方破坏、操作失误、设备故障、腐蚀或老化、地面沉降、用户使用不当等因素，各个因素造成事故的占比见图 2.3 所示；张满可等<sup>[21]</sup>对我国的燃气事故进行了统计分析，将室外的燃气事故发生原因归结于野蛮施工、底面变形（重车碾压、地址松动等）、汽车撞断和地下管道腐蚀；刘爱华等<sup>[22]</sup>收集了 3927 起居民生活燃气事故报道，并将事故原因主要总结为了外力破坏、管道损坏（腐蚀老化、鼠咬、松脱）、阀门失效老化、燃具、人为破坏等，各个因素所占的比例见图 2.4 所示。

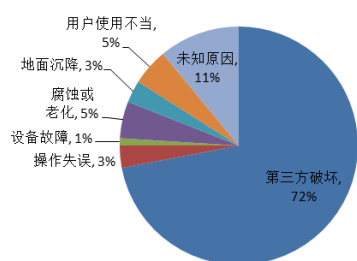


图 2.3 戴联双等<sup>[20]</sup>.2008-2012 国内城市燃气管网失效原因比例

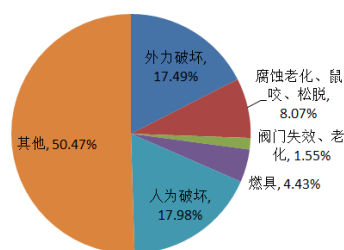


图 2.4 刘爱华等<sup>[22]</sup>.2012-2015 城市居民燃气事故原因比例

## 2.2 事故报道收集的爬虫系统设计

针对人工搜索整理新闻的低效高成本的问题，本文使用基于 Scrapy 框架的爬虫进行新闻爬取。Scrapy<sup>[23]</sup>是一个快速的高级 Web 爬虫框架，用于抓取网站并从其页面中提取结构化数据。它可用于从数据挖掘到监视和自动化测试等多个用途。

Scrapy 的核心组件有以下七个：

- Scrapy Engine: 负责组件之间数据的流转，当某个动作发生时触发事件
- Scheduler: 接受 requests，并把他们入队，便于后续的调度
- Downloader: 负责抓取网页，并传送给引擎，之后抓取结果将传送给 spider
- Spiders: 用户编写的可定制化部分，负责解析 response，产生 items 和 URL
- Item Pipeline: 负责处理 Item，典型的用途：清洗、验证、持久化
- Downloader Middlewares: 位于引擎和下载器之间的一个钩子，处理传送到下载器的 requests 和传送到引擎的 response
- Spider Middlewares: 位于引擎和抓取器之间的一个钩子，处理抓取器的输入和输出

在 Scrapy 中的数据流由执行引擎控制，其流程图见图 2.5 所示：

- (1) 引擎打开一个网站（Open a domain），找到处理该网站的 Spider、并向该 Spider 请求第一个要爬取的 URL(s)
- (2) 引擎从 Spider 中获取到第一个要爬取的 URL 并在调度器（scheduler）以 request 调度
- (3) 引擎向调度器请求下一个要爬取的 URL
- (4) 调度器返回下一个要爬取的 URL 给引擎，引擎将 URL 通过下载中间件（请求（request）方向）转发给下载器（Downloader）
- (5) 当页面下载完毕，下载器生成一个该页面的 response，并将其通过下载中间件（返回（request）方向）发送给引擎
- (6) 引擎从下载器中接收到 response 并通过 spider 中间件（输入方向）发送给 spider 处理
- (7) Spider 处理 response 并通过并返回爬取到的 item 及（跟进的）新的 request 给引擎
- (8) 引擎将（spider 返回的）爬取到的 item 给 item pipeline，将（spider 返回的）request 给调度器
- (9) （从第二步）重复知道调度器中没有更多的 request，引擎关闭该网站

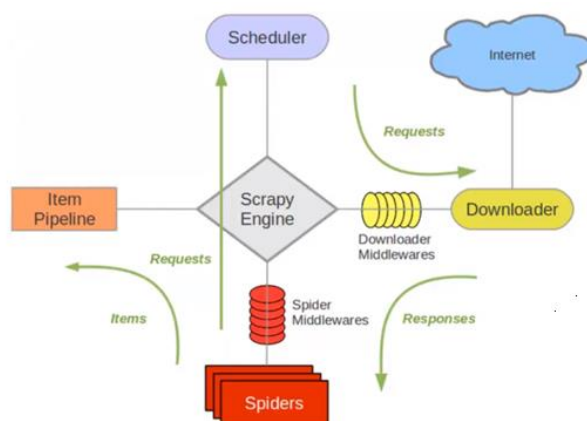


图 2.5 Scrapy 框架部件组成以及信息流传递示意图<sup>[23]</sup>

爬取的新闻内容来源于博燃网的安全速递频道以及中国燃气网的事故快报频道，从博燃网获得 4618 条新闻，从中国燃气网获得共 3854 条新闻，爬取了自 2010 年至 2018 年共 8473 条燃气事故的相关新闻。由于新闻内容繁杂、格式混乱，先通过关键词对新闻内容进行预筛选，建立停用词表并存入文本文档中，因此含有燃气器具、热水器、灶具、煤气器具、胶管等停用词汇的新闻被筛除。再建立好选用词表并存储，选择的关键词分别为管道、管段、燃气管道、天然气管道等。为了剔除国外的新闻建立了外国国家或地区名词表，但具体实施起来时间复杂度较高，可以酌情使用。

根据简单的停用词（未包含国家地区名词表）和选择词初步筛选后得到 988 条新闻，经抽样统计符合目标类型的新闻占比约为 77%，简单的关键词筛选模式已经能基本满足需求。

进一步分析被错误筛选的文章，主要为以下几类：

- 管道事故问责的制度法规介绍
- 居民或者商户室内发生的燃气事故
- 对管道及其附属设施进行安全检查的报道
- 境外相关的事故报告

综合已有的研究和报告对城市燃气管道失效原因的总结，结合上文对错误分类文章的整理分析，本文拟建立初步的关键词表系统对爬取到的新闻文本进行多层分类筛选。本文所构建的关键词表系统架构如图 2.6 所示。

关键词表总体分为选用词和停用词两部分，顾名思义含有选用词的文章则被保留，含停用词的文章则被剔除。选用词文件夹下共创建的七个词表基本上是根据事故原因划分的，包括老化(Aging)，地面移动(Ground Move)，误操作(Operation)，三方破坏(Third Party)，天气(Weather)，室外事故(Outdoor)和未知原因(Unknown)词表。通过比对文章内容和词表关键词便可以对文章进行归因分析，例如三方破坏词表中含有“挖断”、“挖破”、“挖穿”等词语，当文章中有形如“一辆挖掘机不慎将燃气井旁预留管道挖断”的语句时，该文章便被归因判别为第三方破坏的报道。特别的，室外事故词表并不是归因所用，而是在归因之前对文章进行初筛的词表，具体筛选层次系统接下来会具体介绍。

停用词表系统分为监视巡查(Survey)、外国事件(Foreign)、室内事故(Indoor)和杂项(Additional)四个文件，是基于以下逻辑设立了该四个词表：

- 燃气公司针对事故多发地段以及恶劣天气下进行隐患排查的报道，对事故报道的筛选具有极大的干扰，所以对含形如“摸底排查”、“及时制止”等词的报道进行剔除
- 是城市燃气管道事故但发生在境外的报道，不在本文研究范畴内，予以剔除
- 室内发生的居民燃气事故，往往是包含“软管”、“胶管”等迷惑性关键词的报道，需要建立相应词表予以处理
- 杂项是各种干扰项的集合，包括各种与本文研究目标不符的新闻报道关键词，由整理爬取到的上百条新闻正文归纳得来

构建好的词表系统在用于报道归因时并不是一股脑地一次性处理，而是根据由图 2.7 所示的筛选流程进行的。原始新闻数据库被导入后，先根据停用词表进行文章的第一步筛选工作，随后基于 Outdoor 词表对报道进行初筛，最后依次根据其他选用词表进行文章内容的归因判别处理。

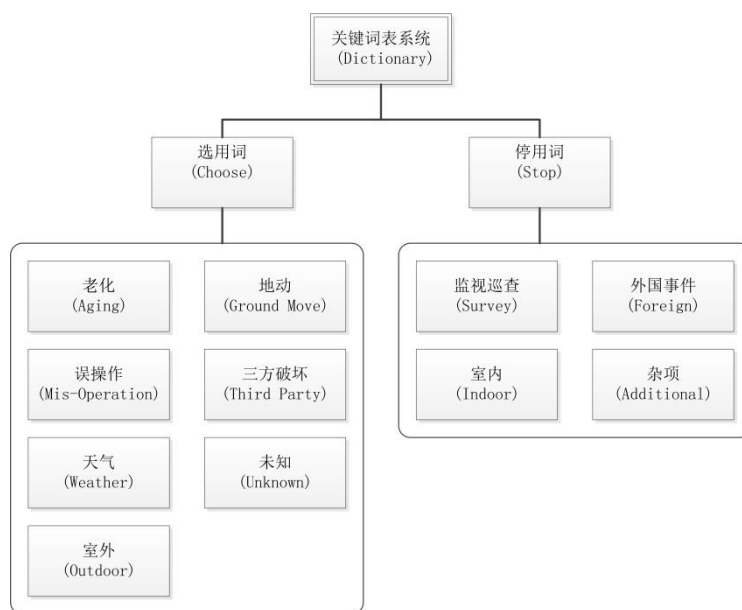


图 2.6 关键词表系统架构示意图

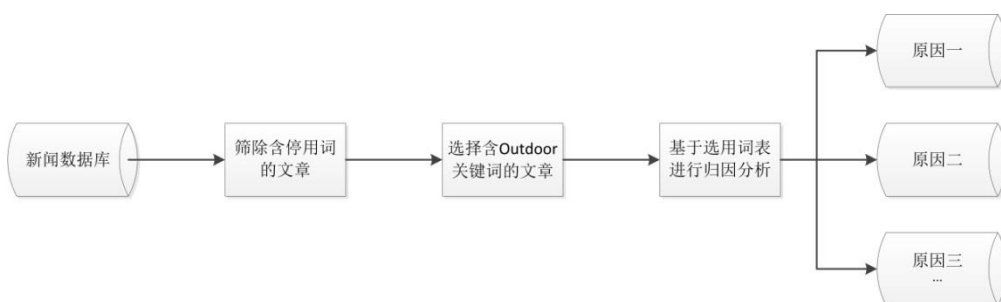


图 2.7 词表系统筛选流程

## 2.3 事故原因的统计与分析

对经过词表系统筛选过后的新闻报道进行分类归因，得到的统计结果如表 2.3 所示：

表 2.3 城市燃气管道事故报道(2010-2018)归因统计结果

原因	老化/腐蚀	地面移动	误操作	第三方破坏	天气	未知	总计
数目	23	10	18	312	4	17	384
比例	6%	3%	5%	81%	1%	4%	100%

据统计结果可见，有 81% 的事故报道指出第三方破坏是事故发生的主要因素。一方面，第三方破坏不在燃气公司的常规管理控制下，往往会造成较严重的居民恐慌、财产损失和人员伤亡后果，这使它们更被新闻媒体所亲睐；另一方面，我国的燃气公司与第三方的施工企业没有完善的交流协调机制，这让许多施工企业在埋地管道区域施工时缺少地下管道的布置

图纸参照，以至于摸黑施工、强行施工，最终导致事故频发。相对比而言，例如误操作、天气和可预测的地面移动等常规失效事故都在燃气公司的安全管理条例保护下被大部分提前排除或者及时抢险，在未造成事故后果之前即被妥善处理，这也是网上罕见此类事故报道的原因。但无论如何第三方破坏都是不可忽视的最重要因素，有必要对第三方破坏进行更深入的挖掘分析。

经过分析，第三方破坏的来源大致可分为个人、车祸和施工破坏三类。其中，个人因素主要指由个人疏忽或故意造成的管道损坏，比如行人乱扔烟头引燃管道、私自改造管道盗取燃气等事件。车祸指由于车辆失控撞断管道、车辆撞倒其他构筑物压断管道以及重型车辆碾压导致管道损坏等事故。以上三类来源各自占比见图 2.7 所示，可见施工破坏(91%)是第三方破坏的主要来源，个人(5%)和车祸(4%)共占不到 10%。针对施工破坏类型再进行细化分类，将施工破坏分成了堆载、挖掘机等若干类，统计结果如图 2.9 所示。结果表明，挖掘机是绝大部分管道施工破坏的来源，原因可能是挖掘机是建设工程地下施工的通用工程机械，被绝大部分施工工程所采用，挖掘机的作业方式也导致它更容易破坏地下的管道。另外可以看到除挖掘机外其他破坏类型占比很小，可能是因为本文所采用的词表系统对原始样本进行了有偏筛选，更容易挑选出挖掘机破坏的新闻，其他类型的新闻因为描述不清或者难以自动化识别的原因被挑选出的数量更少。

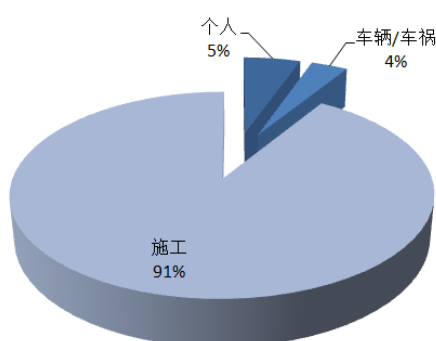


图 2.8 第三方破坏的来源统计



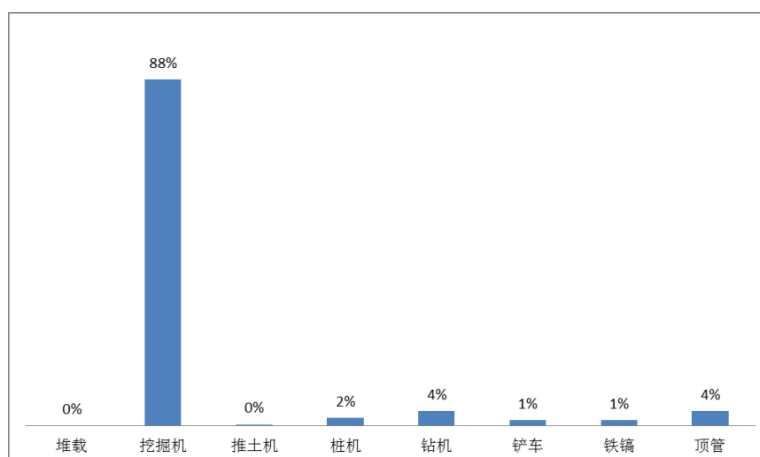


图 2.9 第三方施工破坏的主要类型

## 2.4 开工项目信息收集与可视化

### 2.4.1 图形化定向爬虫软件开发

前文的统计分析表明第三方破坏是国内城市燃气管网事故发生的主要诱因,而施工破坏又是第三方破坏最主要的来源。因此抓住第三方破坏的规律,提前预知第三方行为具有重要价值。对于燃气公司而言,提前获知工程开建的项目信息,继而及时与施工单位取得联系协调,并针对在建施工地区重点巡查,具有重要价值。

为了解决上述问题,本文拟开发对施工工程项目报道进行收集的爬虫,该爬虫应能定制化某日期间隔内上海地区全部的施工开建信息,从而为燃气公司对城市燃气管道的安全管理提供数据化支持。

本文所开发的爬虫软件是基于 Python 语言开发的,并将所有代码打包为 Windows 平台下可方便运行的 EXE 文件。

爬虫部分基于 Selenium 库编写,实现了自动化爬取并保存页面 HTML 源代码的功能。Selenium 是一个用于 Web 应用程序测试的工具,测试直接运行在浏览器中,就像真正的用户在操作一样。支持的浏览器包括 IE、Mozilla Firefox、Mozilla Suite、Chrome 等。

为了方便爬虫的使用和后续更新,本文基于 PyQt5 框架编写了爬虫的简易 UI 界面,并打包为 EXE 文件,方便了用户的无障碍使用。

安装好的爬虫软件文件根目录如图 2.10 所示:

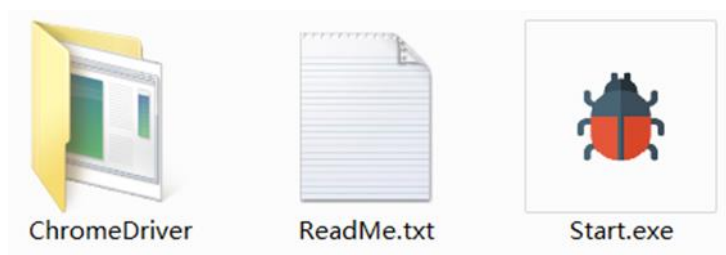


图 2.10 爬虫软件文件根目录文件

使用时只需单击 Start.exe 便可启动爬虫程序，启动后的爬虫 UI 界面如图 2.11 所示。分为起始、终止日期栏，URL 栏和开始运行按钮三个部分。用户在使用时，需要自定爬取起始和终止日期间隔内的建设工程中标公示报告。URL 框默认填充一般不需修改，如爬虫无法正常运行，应手动登录住建委网站找到中标公示对应界面，复制其链接并黏贴于 URL 框内即可。配置完毕后，单击开始运行，等待爬虫运行结束即可。

爬取完毕后，报道标题、日期和标类型均存储于根目录下名为 projects.csv 的表格中，表格内内容展示形如图 2.12 所示。

图 2.11 施工项目信息爬虫 UI 界面

中标日期	招标类型	项目名称			
2018年4月4日	设计招标	繁荣路（周达路~S3）新建工程			
2018年4月4日	施工招标	新港（敞开段）水环境综合整治工程			
2018年4月4日	监理招标	万洲厂区装修修缮工程			
2018年4月4日	勘察招标	长宁区程家桥街道357街坊新建工程			
2018年4月4日	监理招标	黄浦区劳动技术教育中心丽园路111号校区装修工程			
2018年4月4日	监理招标	特种光学薄膜研发平台			
2018年4月3日	设计招标	上海市宝山实验学校南校区抗震加固工程			
2018年4月3日	施工招标	金桥基地冲压辅助车间扩建项目			
2018年4月3日	施工招标	2018年新海小区住宅修缮工程			
2018年4月3日	工程总承包	松江南站大型居住社区C19-28-03地块经济适用房项目新建项目			
2018年4月3日	设计勘察	北沙港路（元电路—颛兴路）道路改建工程			
2018年4月3日	施工招标	永春东路（汽车洋桥—永春东路16号）积水点改造工程			
2018年4月3日	施工招标	新建梅园绿路（金玉路—沪杭铁路）道路工程			
2018年4月3日	工程总承包	松江南站大型居住社区C19-29-01地块经济适用房项目新建项目			
2018年4月3日	施工招标	翔黄路（沪杭铁路立交—虬江河桥）道路维修工程			
2018年4月3日	设计勘察	杨浦区益民小区二期西块新建项目			
2018年4月3日	监理招标	新建梅陇镇森安苑12号地块商业项目			
2018年4月3日	暂估价工程	徐汇滨江地区公共开放空间综合环境建设工程（一期）A配套建筑（幕墙工程）			
2018年4月3日	施工招标	绿华镇合作农场仓库、宿舍楼修缮工程			
2018年4月3日	施工招标	南桥新城18单元02-10地块（暂定名）项目			
2018年4月3日	施工招标	向化镇住宅小区二次供水设施改造			
2018年4月3日	施工招标	2018年华院小区住宅修缮工程			
2018年4月3日	暂估价工程	上海（金山）国际中小企业产业园（一期）消防控制系统专业分包工程			
2018年4月3日	设计施工	BAROMON淮海路商业体验中心装修工程			
2018年4月3日	监理招标	新建森安苑幼儿园			
2018年4月3日	施工招标	上海东源汇信股权投资基金管理有限公司办公室装修及搬迁工程项目			

图 2.12 爬取到的中标公示内容示例（节选）

## 2.4.2 项目地理信息的可视化

收集好的中标项目信息均存于名为 `projects.csv` 的文件中，考虑到对项目文字信息的人工再处理效率比较低，本文引入了百度地图 API 接口对中标项目进行地理位置解析，并通过 API 接口生成展示中标项目地理位置分布的 HTML 页面。可视化的项目地理信息系统能够更加直观的反映施工项目的地理位置，弥补部分施工工地未在燃气公司备案造成的管理缺失，方便管理者针对性地调整安排巡检工作。

本文通过编写 Python 脚本程序，从 `projects.csv` 文件中提取项目信息，调用百度地图 API，分条传输项目标题并传回解析好的结构化地址。针对存在部分标题未准确描述项目地址，或者百度地图引擎并未准确识别地点的情况，本文作了筛选机制，剔除了查找失败和结果不在上海市范围内的项目。对获得的结构化地址，本文编写了 HTML 页面并将其集成于同济大学智能燃气监测系统网页中，结果如图 2.13 所示。该页面不仅能展示项目的地理分布，点击单个标志点还能够显示该项目的标题和中标日期。

### Smart Gas

同济大学燃气管网安全监控平台

周界监测	安全评估	仿真评估	案例分析	青浦周界预警
------	------	------	------	--------

#### 案例分析



图 2.13 施工项目地理信息的可视化

## 2.5 燃气管道事故信息化系统

本章依据数据获取、数据处理分析和数据可视化展示的三步骤形式开展了研究与开发工作，对应到完成的具体工作主要有以下几点：

- 分别编写针对博燃网、中国燃气网以及上海市住建委中标公告信息网的爬虫脚本，并编写相应的爬虫 UI 界面
- 对爬虫爬取到的新闻内容进行事故原因分类，对中标信息内容进行地址解析并根据百度地图 API 编写可视化中标信息的地理分布示意图
- 编写将事故报道、事故报道可视化统计分析结果和中标信息地理分布示意图自动化生成和传输至 WordPress 页面的脚本，完善 HTML 页面展示效果

该系统所实现的业务流程图如图 2.14 所示，爬虫脚本运行分别从博燃网、中国燃气网和上海市住建委中标公告信息网爬取燃气事故报道和上海市新建工程项目中标公告。

对燃气事故报道采用词库系统对其进行三层归因分类，第一层初步筛除非目标事故报道以及对目标事故报道进行包括老化、地面移动等大类归因分析；第二层针对第三方破坏进行包括施工、个人等具体归因分析；第三层针对施工破坏进行包括钻机、挖掘机等具体归因分析。以上三层归因分析均会产生相应分类好的报道数据库文件，分别对这些分类库文件进行统计分析生成图表。

对中标信息的标题利用百度地图 API 进行地址解析并存储返回储存好的结构化经纬度数据，将经纬度数据嵌入预编写好的 HTML 模板中生成中标信息地理分布的可视化页面。脚本程序系统打包以上各种形式的原始数据和已处理数据，分别生成可视化的新闻报道数据库列表页、统计分析图表页和中标信息地理分布页并上传至 WordPress 服务端，在其前端用户即可查阅实时更新的动态页面信息。

整个系统下的文件和程序的文件目录图见图 2.15 所示，文件名称以及简略功能介绍见表 2.4、2.5 所示，(一)表主要是根目录下的脚本程序和数据文件，(二)表主要是根目录下数据和词表文件库。

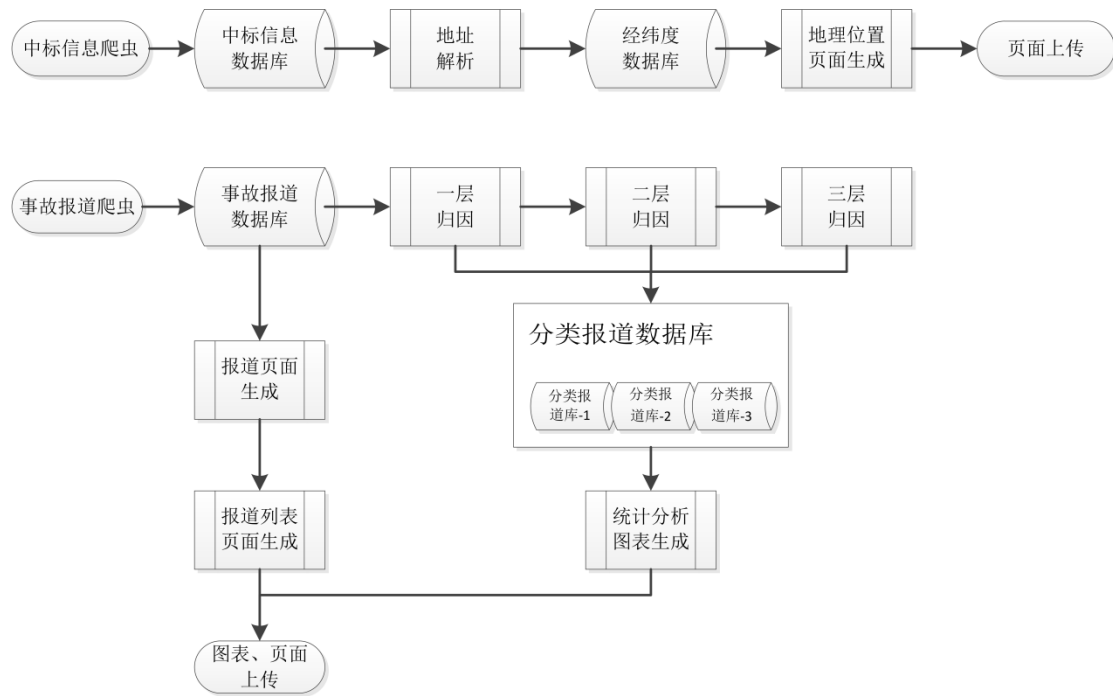


图 2.14 燃气管道事故信息化系统业务流程图

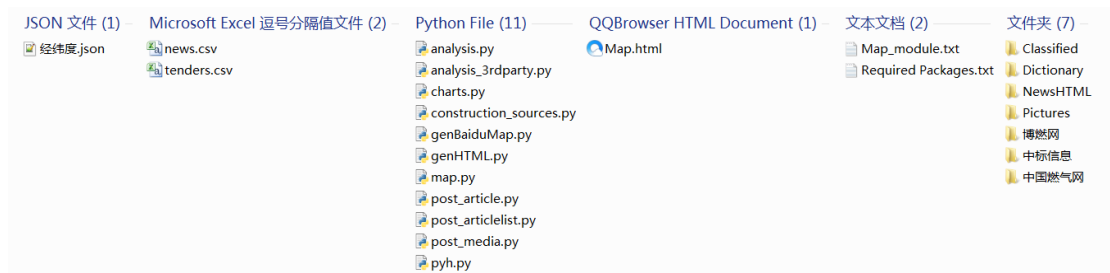


图 2.15 系统文件目录

表 2.4 系统文件名与简介(一)

NewsHTML	待上传的新闻报道页 HTML 代码(0.html/...)	中国燃气网	中国燃气网爬虫项目文件,Scrapy Project1
Pictures	待上传的统计分析页插图 (1.png/...)	博燃网	博燃网爬虫项目文件, Scrapy Project2
Map.html	待上传的中标信息可视化 HTML 代码	news.csv	Project1 和 Project2 爬取得到的新闻报道数据库
Map_module.txt	生成 Map.html 的模板	经纬度.json	待嵌入 Map.html 页中的中标项目地理经纬度数据库
analysis.py	输入 news.csv,利用词表系统按管道事故因素分类的新闻报道,存入对应的报道库中	charts.py	输入分类完成的报道库,输出归因好的统计数据饼图和表格
analysis_3rdparty.py	输入三方破坏报道库,利用词表系统按具体第三方破坏因素分类,存入对应报道库	construction_sources.py	输入施工破坏报道库,利用词表系统按具体施工破坏因素分类,存入对应报道库
genBaiduMap.py	输入 Map 模板文件和经纬度数据,生成待上传的 Map.html	post_article.py	输入生成好的新闻报道 HTML 文件,上传至 WordPress 服务器端
genHTML.py	输入 Classified 下的报道数据库,分别生成每篇报道的 HTML 页面文件	post_article_list.py	生成展示新闻报道列表页面的 HTML 文件,上传至 WordPress 服务器端
post_media.py	输入统计图表图片文件,上传至 WordPress 端	中标信息	爬取中标信息和生成经纬度 json 数据的项目文件夹
pyh.py	辅助生成 HTML 的工具库		

表 2.5 系统文件名与简介(二)

Classified	ThirdParty	construction_sources	报道库,具体施工行为因素
		car.csv	报道库,车祸因素
		construction.csv	报道库,施工因素
		pedestrian.csv	报道库,个人行为
	aging.csv	报道库,老化因素	
	ground_move.csv	报道库,地面移动	
	operation.csv	报道库,误操作	
	third_party.csv	报道库,三方破坏	
	weather.csv	报道库,天气异常	
	unknown.csv	报道库,未知	
Dictionary	Choose	aging.txt	选用词表,老化
		ground_move.txt	选用词表,地面移动
		operation.txt	选用词表,误操作
		outdoor.txt	选用词表,室外事故
		third_party.txt	选用词表,三方破坏
		unknown.txt	选用词表,未知事件
		weather.txt	选用词表,异常天气
	Stop	additional.txt	停用词表,杂项
		foreign.txt	停用词表,外国事故
		indoor.txt	停用词表,室内事故
		survey.txt	停用词表,安全检查
	ThirdParty	construction_sources	归因词表,具体施工行为因素
		car.txt	归因词表,车祸因素
		construction.txt	归因词表,施工因素
		pedestrian.txt	归因词表,个人行为因素

### 3 生命线分段管道的风险评估模型

城市燃气管道系统是由大量管段拼接而成的复杂系统，每条管道都可能有不同的管龄，管材，管厚等内部属性和周边环境等外部属性。鉴于管段之间存在显著的不同，差异导致管道系统的风险是空间不均分分布的，无论是采用半定量还是定量的评价方法，都需要考虑到管道之间存在的客观差异，对管道进行合理的分段<sup>[11]</sup>，并且有必要根据不同类型的管道分别构建评估或预测模型，有的放矢，能取得更加精确和具有普适性的管段安全检测模型。

管道系统的规模不断扩大，管道数据来源逐渐丰富和多元化，这对于现代管网安全性评估既是挑战又是机遇。一般来说，量化的风险评估模型可以利用的数据来源越丰富，数据量越大就可以得到越精确的结果，但综合有效地利用这些数据往往才是任务的关键。比如单看管段属性的每一个指标都可以分成许多类别，例如管龄可能就可以被分为 0-1 年组，1-2 年组等等组别，但是这样的处理会导致对管段的分类过于繁复，根本无法在工程上应用。所以采用一定的手段控制管段的分类，使其分类方式既可以表现每类管段的特点又不至于过于细分以至于不利于工程应用是亟待解决的现实问题。传统的多指标综合评价的基本思想是将包含各个侧面的多个单项指标组合起来形成一个综合指标，通过对每个管道的综合指标划分层次来进行管段分类，然而多指标的综合评价往往难以避免评价指标之间相关性和赋权的主观性问题<sup>[24]</sup>。张杰<sup>[13]</sup>在长输管道的分段问题中引入了无监督聚类的方法，即先通过对每个管段打分赋权取得其 10 个量化的风险因素指标，包括埋深、壁厚、人口密度、公众态度、上方活动、阴保电流、土壤腐蚀、杂散电流、敷设方式和土体类型，把对原始指标进行 PCA 降维后的指标进行层次聚类从而得到管道的分段结果。无监督聚类方法能够一定程度排除人为主观因素对管段分段的影响并且得到无法人工发掘的新的类别关系，是解决管道分段问题一条十分可行的思路。

本文获得的数据来源于某地燃气管道网络 GIS 系统以及当地某燃气公司 SCADA 系统，GIS 系统中存储了当地超过 13000 根现役燃气管道的多种属性，包括其长度、直径、材料、坐标、内压等，本章研究的目的在于充分利用 GIS 数据开展管段的聚类分析，挖掘不同聚类簇的风险等级关系，最终建立分段管道的安全评估模型。

#### 3.1 无监督聚类

无监督学习(Unsupervised Learning)是指训练样本的标记信息未知的统计学习方法，目标是通过通过对无标记训练样本的学习来揭示数据的内在性质及规律，为进一步的数据分析提供基础，聚类(Clustering)是此类学习任务中被研究和应用最广的一个方向<sup>[25]</sup>。

聚类方法的目标是把原始数据集划分为多个簇(Cluster)，每个簇类别所代表的分类概念是未知的，即每个簇所代表的类别是否符合工程或科学上的认知概念需要由使用者来把握，这也赋予了无监督聚类方法发现无法由人工发掘的复杂隐含关系的能力。

聚类的性能度量有两种<sup>[25]</sup>，一类是将聚类结果与某个参考模版(Referenced Model)进行比较，称为“外部指标(External Index)”；另一类是直接考察聚类结果而不利用任何参考模型，成为“内部指标(Internal Index)”。其中，外部指标包括 Jaccard 系数、FM 指数、Rand



指数等，内部指标包括 DB 指数、Dunn 指数等。无论哪种指标，都涉及到样本之间距离度量(Distance Measure)的问题，最常用的距离度量指标是闵可夫斯基距离(Minkowski Distance):

$$dist(x_i, x_j) = \left( \sum_{u=1}^n |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}} \quad (3.1)$$

当  $p$  取 1 和 2 时分别是闵可夫斯基距离的特例曼哈顿距离(Manhattan Distance)和欧式距离(Euclidean Distance)。

关于距离度量，有必要提到其直递性的概念，本章后面的研究工作基于此概念对模型进行了扩展。直递性的定义<sup>[25]</sup>为：

$$dist(x_i, x_j) \leq dist(x_i, x_k) + dist(x_k, x_j) \quad (3.2)$$

在某些任务不满足直递性的距离被称为“非度量距离”(non-metric distance)，举一个直观的说明，在某些任务中平面上两点之间的线段长度越短并不代表二者越接近，此时需要根据数据样本设计合适的距离计算式，或者把原数据点利用核函数(Kernel)投射到高维空间，再计算样本点之间的普通距离度量，后文会更具体地叙述这一点的应用过程。

聚类算法有很多且尚未成为完整的体系，周志华在其《机器学习》一书中列举了原型聚类(Prototype-based Clustering)、密度聚类(Density-based Clustering)和层次聚类(Hierarchical Clustering)三大类<sup>[25]</sup>，在工程上选择聚类算法要根据具体任务和数据确定，并没有通用的标准模式可以套用。

### 3.2 特征工程

在机器学习基础理论中有一个定理十分有趣，即“没有免费的午餐”定理(No Free Lunch Theorem, NFL<sup>[26]</sup>)。简单来说就是无论学习算法多么精妙复杂，其期望性能都与随机猜测算法相同。这并不意味着机器学习算法都是白折腾，而是指出了机器学习算法构建和应用的方向。周志华的《机器学习》一书对此问题有深刻的讨论：NFL 理论有一个重要的前提，所有的问题出现的机会相同或者所有问题同等重要，但是实际情况下我们时候只关注自己正在试图解决的问题(例如某个具体的应用任务)，希望为它找到一个解决方案，至于这个方案在别的问题甚至在相似问题上的性能并不重要<sup>[25]</sup>。

特征工程是机器学习在工程应用上的基石。特征工程(Feature Engineering)是指从原始数据样本中提取最能表达数据本质属性的特征以供算法使用，好的特征工程能尽可能地舍弃数据中的噪声、无用属性，保留显著特征和构建新的更具表达能力的特征。这里所说的特征，在机器学习理论里被成为假设(Hypothesis)，算法学习的过程就是一个在所有假设构成的空间里进行搜索的过程，假设的表示就是样本的特征的取值形式，当假设的表示确定以后，假设空间的规模和大小也同时被确定。打个比方，从多个方面描述一所大学的好坏可以有学生人数、教工人数、占地面积等属性，这些属性就是对描述一所大学这个问题的假设，当确定好了所用的所有假设，描述一所大学这个任务的假设空间也确定了。事实上，可以从无限

个角度观察一所大学，同时也可以得到无数种特征。从中有目的地选择特征和构建特征，就可以让学习算法在这一任务上具有优于随机猜测的性能，这是特征工程必要性的理论依据。

### 3.3 基于多种静态特征的管道聚类

基于上述讨论，本文在管道的聚类这一任务上确立了数据清洗，特征工程，算法聚类三步走的基本流程。上述环节每一个都不可或缺，都对最终取得鲁棒、符合工程经验的聚类结果具有重要意义，以下各流程均使用 Python 完成，Python 语言因其胶水语言的特性以及其强大丰富的第三方库，能被用于优雅方便地进行自动化的数据处理和分析工作。

#### 3.3.1 数据清洗

从 GIS 系统中取得的管道属性表 GASLINE.xls 由 OBJECTID、LENGTH 等几十列组成，共有 13000 余个样本。原始的数据表中存在很多问题，并不能直接用于算法的输入。首先是一些意义不明的列下基本都是空值，或者列下所有样本取值相同，处理方法是将列全部剔除；表的编号字段是表中所有样本的主键，对重复的主键样本需要去重处理；还有表中数据格式不符合算法输入的情况，需要对其格式进行处理，比如表的 SHAPE 列下数据是形如 SDE.ST\_GEOMETRY(4, 4, 65177.5273483712, 37037.6749225254, 65191.5687638808, 37097.2413789662, null, null, null, null, 0, 61.20134351600970, 300002)的一串字符串，其实际意义是管道的坐标、形状等信息，对其基于正则过滤进行分离处理可以得到管道的坐标、长度等新的符合输入格式的数据列；此外，机器学习算法不具有处理缺失值的能力，还需要对缺失的值进行填充或者舍弃的处理，这里本文对类别变量和数值变量采取的不同处理方式。对于类别变量，将其缺失值转为一类新的类别并填充；对于数值变量均采用所有非缺失值的中值填充，不采用均值填充是考虑到某些特征如管径取值是整数，均值填充会引入小数造成精度损失；针对某些列下数据格式错乱的情况，还要具体情况下具体处理，如归并同类、转为缺失类等。

#### 3.3.2 特征工程

特征工程需要对每一个特征设计针对性的处理模式，经过首轮的数据清洗，初步选择了 G3E\_FID, WARN\_TAPE, JJSIZE, MATL, PRESS\_D, PRESS\_O, DATE\_BUILD, CONTRACTOR, SUPERVISOR, SHAPE, LENGTH 等列。

特别的，对于类别特征而言需要进行额外的处理才能作为机器学习算法的输入。最常用简单的方法是直接对类别特征进行序数编码，如类 a 编码为 1，类 b 编码为 2，类 c 编码为 3...依次类推。但是，这种编码隐含了类别具有顺序性的假设，比如类 b 比类 c 更接近类 a，对于很多算法而言，这种编码还包含了类 a 和类 b 的“不同程度”与类 b 和类 c 相同的假设，因为二者的编码值都只相差了 1，很显然许多情况下这些假设并不成立，如果这样简单处理就会导致算法输出的偏差甚至错误。为了解决这个问题，许多类别编码方式被提出。一个方法是 one-hot 编码，即把类别样本转为哑变量(dummy variables)向量，列内每个样本被转为只含 0 和 1 的二值向量。该方法适用于内在不含顺序性且每种类别两两之间的距离相近的情

况，但是对于类别特别多的情况该方法会产生极高维度且及其稀疏(Sparse)的哑变量矩阵，造成对计算能力的爆炸性需求增长，同时使算法难以收敛。

本文综合考虑了以上两种编码方式的适用场景，针对性地对每一列类别数据分别编码，考虑到类别之间相似度的差异性，引入先验知识(priori knowledge)对不同类别赋予类的权值，用于聚类的每个数值特征都转为统一的越小越好的模式(smaller is better)，这样有助于聚类结果的可解释性，特征工程算法的细节在本节后部会详细介绍。

接下来将分别叙述数据集每列的具体含义，进行数据探索性分析并介绍对应的特征处理算法和处理流程。

### 1) WARN\_TAPE

该列为有无警示标志的类别变量，取值为：有、无和不适合，存在大量缺失值。这里对有标记为 1，其它均标记为 0。经处理后，共有 5595 个样本取值为 1，7590 个样本取值为 0。

### 2) JJSIZE

该列是管径数据，单位为 mm 毫米，无缺失值，数据分布如图 3.1 所示，管径集中在 200mm 附近，少量的管道超过 400mm 只有 3 根管径为 1020mm 的管道超过 600mm，有一部分小于 20mm 管径的管道，分布的正态性不显著。该列为数值特征且无缺失数据，为了防止 3 个离群值对于标准化的影响，1020mm 管径的数据先被归并到 600mm 组中，再采用 Z-score 标准化(zero-mean normalization)处理：

$$X = \frac{X_0 - \mu}{\sigma} \quad (3.3)$$

该方法将原始数据转化为标准差为 1，均值为 0 的正态分布，是把数据进行无量纲化处理的常用手段。不同的特征之间只有去量纲化才能够直接相互比较，使聚类结果更加稳健。

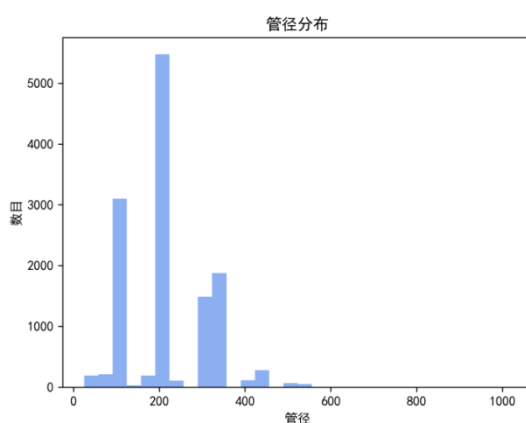


图 3.1 管径的样本分布

事实上，管径虽然是数值型特征，也可以被转换为类别特征。因为管道是批量生产和批量埋设的，每一批管道都具有统一型号，所以管道的管径实际上可以被编码为有限类别的一个类别特征。

### 3) MATL

该列是管道的管材，经统计管道的管材有钢管、铸铁管、PE 管、球墨铸铁管、PE 塑料管、灰口铸铁管、镀锌管和球墨管等，部分类别可能是由于人工录入过程失误产生的，如球墨管和球墨铸铁管，PE 管和 PE 塑料管等。为了减小复杂度需要归并部分类，如将球墨管类并入球墨铸铁管类，将含 PE 字段的管道全部合并为 PE 管，最终处理好的样本直方分布见图 3.2 所示。

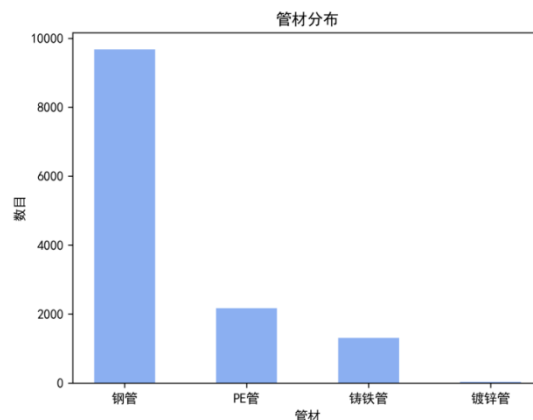


图 3.2 管材的样本分布

EGIG 每年的报告都对各种次级失效频率(Secondary Failure Frequencies)的统计分析，次级失效频率的计算是考虑不同参数(压力、直径、覆盖深度等)作为每个事故的原因的年发生频率<sup>[18]</sup>。EGIG 的报告涵盖从 1997 年到 2017 年近 20 年的统计数据，包括管径、覆土深度、管壁厚、管材、管龄等参数的次级失效频率，数据详实准确，样本容量大，为本文的研究提供了极有价值的先验知识。

对于上文所述将管径转为类别变量的概念，即可以采用管径的次级失效频率作为先验赋权，EGIG 关于管径的次级失效频率的部分统计结果如表 3.1 所示(原始数据以英寸为单位，已将其全部转为毫米)：

表 3.1 EGIG 次级失效频率-管道直径和泄漏程度统计表(1997-2016)<sup>[18]</sup>

管径(mm)	次级失效频率(1000km•year)				
/	未知	针眼&裂缝	孔洞	破裂	合计
(0,127]	0.011	0.333	0.122	0.074	0.54
[127,279.4)	0.011	0.138	0.08	0.027	0.256
[279.4,431.8)	0.007	0.055	0.04	0.017	0.119
[431.8,inf)	0.007	0.048	0.026	0.011	0.092

在处理管径这一特征时，先验知识实际上是对过去事件发生频率的总结，依据大数定理，当样本足够多时随机变量依概率收敛于常数，样本的频率均值收敛于其概率<sup>[27]</sup>。因此，可

以使用过去发生事件的频率均值作为其发生概率的估计。这里的先验知识实际上就是某类别的历史发生事故统计频率，将其作为数值特征分别给予不同类别一个权重，这个权重实际上就赋予了不同类别差异性的量化数值表示，解决了简单序数编码的性能问题。

直观上看管径越大，事故频率越低，本文将管径离散化归类到管径区间内形成四个管径类别，以其各自的合计列的值作为类权重，生成新特征并命名为 **WEIGHTED\_DIAMETER** 列，该列为数值型特征。

不同管材客观上具有不同的性能，使用简单的序数编码无法准确地量化它们之间性能的差异和潜在的性能优劣排序。而这里对于管材的先验知识无法引用 EGIG 的失效频率统计，原因是报告对于管材是基于美标 API SPEC 5L 标准分类的，包括 Grade B, X52, X60 等标准钢管，并不能在此直接套用。因此，本文拟用国标规定的不同管材的使用年限作为先验知识赋权。基于《城镇燃气设计规范》(GB50028-2006)规定的不同管材的燃气管道设计使用年限，本文分别对钢管、铸铁管、PE 管和镀锌管赋权 30, 60, 50 和 40，新特征命名为 **WEIGHTED\_MATL**。

#### 4) SHAPE

该列内值为字符串，实际上包含了每根管道在 GIS 系统中的坐标和形状数据，值得关注的是该坐标是地区局部坐标系中的坐标，单位为米，与经纬度并不对应。在数据清洗流程中已将其分离为两列坐标。坐标数据不能直接作为聚类算法的输入，需要结合坐标对应的实际地点提取相关信息。

本文依据由局部坐标系绘制的燃气管网示意图，在地图上选取了多个样本点及其坐标，同时找到样本点对应的经纬度，这样就找到了经纬度和局部坐标系的映射关系，根据该映射关系得到了表中所有管道的经纬度坐标。该坐标不直接用于聚类算法，而是用于事故维修记录表中事故地点对应管道的查找。

#### 5) DATE\_BUILD

**DATE\_BUILD** 是管道的建造日期，其原始格式是年/月/日，可被分解为年、月、日三列，仅取年和月两列，作其分布直方图如图 3.3、3.4 所示。直观可见绝大部分管道建成年份都在 2000 到 2010 年这十年间，只有少部分管道建成于 2000 以前。没有使用中且管龄超出最低设计使用年限 30 年的管道。从月份分布图中可见，大部分管道埋设于 7、8 和 9 月份之间，属于夏季。年份不适合直接作为聚类算法的输入，所以将其转换为距 2018 年 1 月 1 日的天数作为管龄属性。分离出的列为管龄和月份，列名分别为 **DATE\_BUILD\_AGE** 和 **DATE\_BUILD\_MONTH**。直观上看，管龄越小的管道安全性越高，符合 **smaller is better** 规则，但是建成月份显然不符合该规则。月份之间作为类别应当数值编码，为了降维将月份按季节分解为四个类别。从时间顺序看，类别应当按春夏秋冬排列，但从季节的气候来看，春和秋比春和夏似乎更加接近，考虑到这一点，本文对春夏秋冬分别编码为 1, 3, 2, 0。

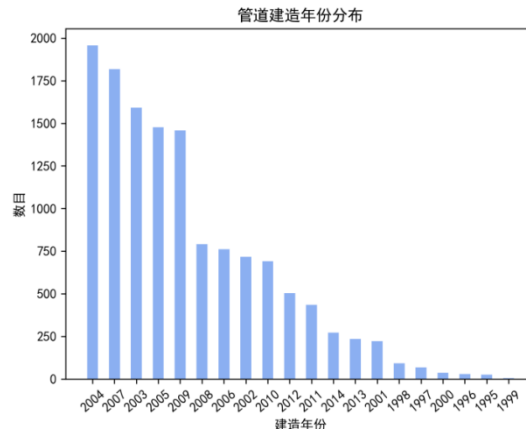


图 3.3 管道建造年代样本分布

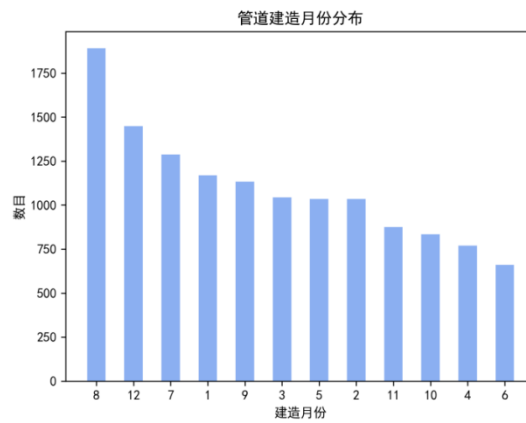


图 3.4 管道建造月份样本分布

#### 6) PRESS\_D & PRESS\_O

该两列分别为管道的设计压力和运行压力，其取值有 LP(低压)，MPB1(中压 B1)，MPB2(中压 B2)，MPA(中压 A)，IHP(次高压)和 HP(高压)，具体的分布见图 3.5 和 3.6 所示。

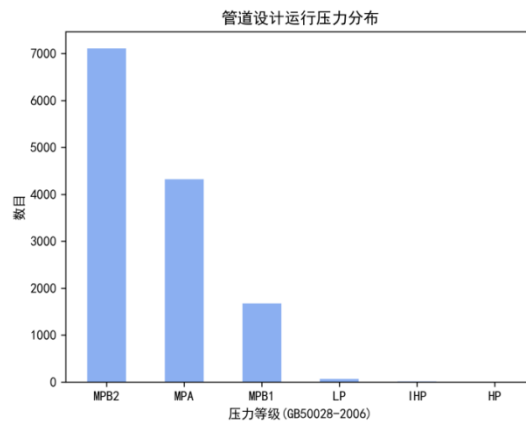


图 3.5 管道设计运行压力样本分布

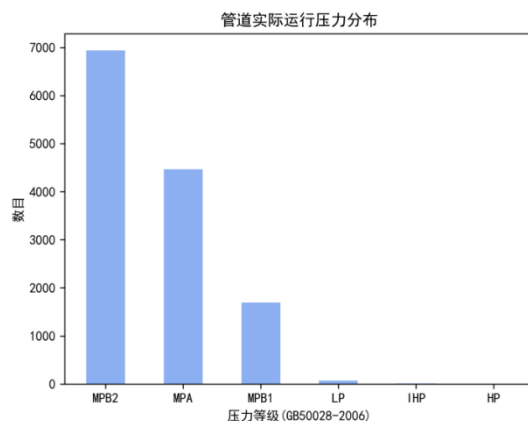


图 3.6 管道实际运行压力样本分布

对于管道运行压力特征，根据管道最大设计承压(内压)计算公式<sup>[28]</sup>：

$$S_1 = \frac{P \cdot D_1}{2([\sigma]E + P \cdot Y)} \quad (3.4)$$

其中  $P$  是设计压力， $S$  是管壁设计厚度，易得当设计内压增大，保证管道安全所需的设计壁厚越大。若将安全性作为因变量，则压力  $P$  符合前文预设的 **smaller is better** 规则，所以可将管道设计压力数值直接作为特征输入。城镇燃气设计规范(GB50028-2006)规定设计压力分级区间如表 3.2 所示；考虑规范的分级规定，本文对原始数据的压力类别赋权见表 3.3 所示：

表 3.2 城镇燃气管道设计压力(表压)分级<sup>[29]</sup>

压力等级	LP	MPB	MPA	IHPB	IHPA	HPB	HPA
区间	(,0.01]	(0.01,0.2]	(0.2,0.4]	(0.4,0.8]	(0.8,1.6]	(1.6,2.5]	(2.5,4.0]

表 3.3 管道设计内压特征赋权

压力类别	LP	MPB1	MPB2	MPA	IHP	HP
权值	0.01	0.07	0.2	0.4	0.8	2.5

由于原表中没有管道壁厚的数据，而最小剩余壁厚<sup>[30]</sup>又是管道抗腐蚀能力的一个重要指标，本文根据公式 3.4 对管道内压特征和管径特征基于公式 3.5 进行组合得到假设壁厚  $Hthickness$ ，式中下标 0 意为经 0.1 标准化的原始数据，该特征无物理意义的量纲，仅作为数值特征比较大小时用。为了使其符合 **smaller is better** 规则，对其取负数作为输入特征。

$$Hthickness = \frac{press_0 \cdot JJSIZE_0}{1 + press_0} \quad (3.5)$$

## 7) CONTRACTOR & SUPERVISOR

该两列分别为管道安装工程项目的乙方和监理方。该特征类别众多，其中涉及的乙方共有 28 家，涉及的监理方共 9 家。对其做简单序数编码不合适，因为公司之间的顺序不能直接确定；作哑变量矩阵也不合适，因为显然公司之间的差异性并不是两两相同的。考虑到上述两个因素，为了量化公司类别的差异性，在每家公司过去承揽工程项目的记录难以全部获得的情况下，本文拟采用每家公司的注册资本作为其权值，对公司描述不明确和缺失的全部采用中值填充。经处理后的乙方和监理方的注册资本(已对数化)分布见图 3.7、3.8，其中乙方主要在 6000~8000 万元之间，监理方则集中在 500 万元左右。一般来说，注册资本越多的公司实力越雄厚，所以该特征不符合 *smaller is better* 规则，故对其取倒数作为最终取值。

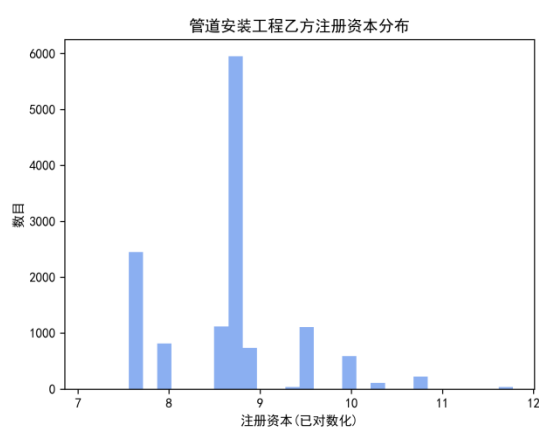


图 3.7 管道安装工程乙方注册资本直方图

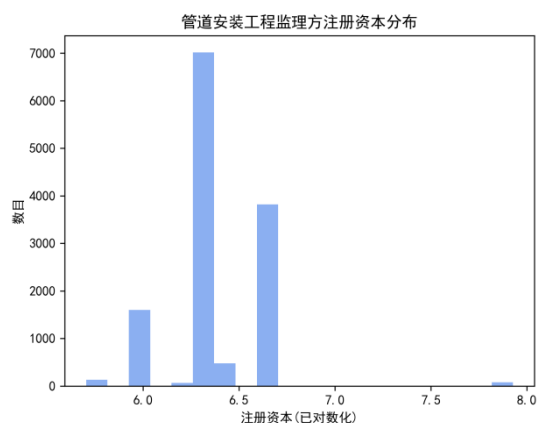


图 3.8 管道安装工程监理方注册资本直方图

## 8) LENGTH

该列为管道长度，单位为米。一般而言，管道越短则其发生事故的概率会较低，这是基于直觉的思考。事实上，EGIG 的统计报告中很多结果均以平均 1km 的事故发生率作为单位，其隐含了沿管道长度方向风险概率均匀分布的概念。因此，本文将管长视作符合 *smaller is better* 规则的特征，因其值域较宽，对其对数化后再做标准化处理。



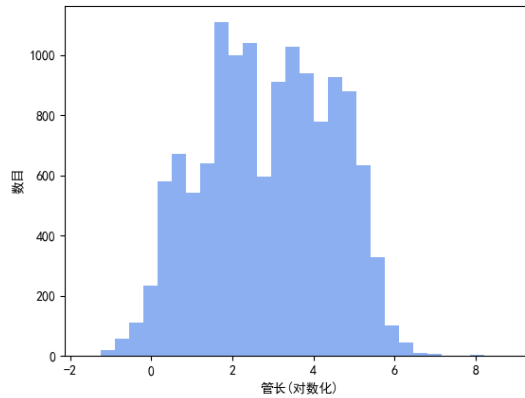


图 3.9 管道管长分布直方图

综合上文所述的特征工程，对最终选取用于聚类的特征名称和描述整理见下表：

表 3.4 聚类管道属性特征说明表

特征名称	类型	值域	单位	说明
WARN_TAPE	类别型	0,1	\	管道埋设处是否有警示标志,1 为有,0 为无
JJSIZE	数值型	[25,1020]	mm	管道管径大小
WEIGHTED_MATL	数值型	[30,60]	年	管道材料的设计使用寿命
WEIGHTED_DIAMETER	数值型	[0.092,0.540]	\	根据 EGIG 统计数据引入的管径的事先验概率
DATE_BUILD_AGE	数值型	[1123,8341]	日	管道从安装完成起至 2018 年 1 月 1 日的天数
DATE_BUILD_SEASON	类别型	0,1,2,3	\	管道安装的季节,0 为冬季,1 为春季,2 为秋季,3 为夏季
PRESS_O_WEIGHT	数值型	[0.01,2.5]	mPa	管道实际运行内压
PRESS_D_WEIGHT	数值型	[0.01,2.5]	mPa	管道设计使用内压
CONTRACTOR_WEIGHT	数值型	[1200,130000]	万元	管道安装施工乙方注册资本
SUPERVISOR_WEIGHT	数值型	[300,2773.9]	万元	管道安装施工监理方注册资本
HTHICKNESS	数值型	[0,1]	\	管道的假设壁厚,仿照规范公式计算得到
LENGTH	数值型	[0.2,7509]	米	管道长度

备注：特征输入聚类算法前都已进行标准化和 smaller better 处理，该表为了便于解释采用了处理前的特征属性

### 3.3.3 聚类算法简述

由于不同的聚类算法适合不同的业务背景,在无法预知适合本文数据的最好算法的情况下,本文在数据集上采用了多种聚类算法,以便对其进行相互比较。本文实验尝试的算法有 K-means、Spectral Clustering、Hierarchical Clustering 和 DBSCAN,上述四种算法分别是原型聚类、基于图论的聚类、层次聚类和密度聚类的经典,综合多种聚类算法有助于全面评估聚类结果的优劣。

#### 1) K-means

K-means 算法即“k 均值”算法,其针对聚类的簇划分  $C = \{C_1, C_2, \dots, C_k\}$  最小化平方误差和(式 3.6),

$$E = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|_2^2 \quad (3.6)$$

其中  $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$  是簇  $C_i$  的均值向量,实际上式 3.6 表示的是簇内样本围绕簇的均值向量的紧密程度,  $E$  值越小则簇内的样本相似度越高<sup>[25]</sup>。为了找到最优解需要考察样本集所有可能的簇划分,是一个 NP 难问题<sup>[31]</sup>,因此该算法通过贪心策略迭代以最小化式 3.6。

该算法一般基于欧式距离(Euclidean Distance)度量样本之间的相似度,所以很容易实现。但其计算量比较大,在高维大量数据下容易不能收敛;其对初始选取的中心点位置比较敏感,且在很多高维情况下欧式距离不适合作为距离度量,可能导致聚类结果比较差。

#### 2) Spectral Clustering

Spectral Clustering,一般译作“谱聚类”,是基于图论演化而来的一种聚类算法,它的主要想法是将所有数据视为空间中的点,点之间通过边连接。距离较远的两点之间的边权重低,反之则高。通过对图进行切割,使不同子图间的边权重低而子图内边权重高以达到聚类的目的。此处限于篇幅只给出其概要而不给出详细证明过程,详见[Ulrike, Luxburg., 2007]<sup>[32]</sup>。

采样点构成的网络图的方式主要有三种:  $\mathcal{E}$ -neighborhood、k-nearest neighborhood 和 fully connected,前两种可以构造出稀疏矩阵适合大型数据集,本文处理的数据样本只有一万多条,可以采用第三种形式,其计算样本  $i$  和  $j$  之间的相似性时一般使用高斯距离:

$$s_{i,j} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \quad (3.7)$$

因为高斯距离符合样本越相似则值越小的特点,所以每两条样本两两之间计算得到的相似性矩阵  $S$  可以直接作为邻接矩阵  $W$  (affinity matrix)。根据式 3.8、3.9 计算出度矩阵  $D$  (degree matrix) 和拉普拉斯矩阵  $L$  (Laplacians matrix):

$$D_{i,j} = \begin{cases} 1 & \text{if } i \neq j, \\ \sum_j w_{i,j} & \text{if } i = j \end{cases} \quad (3.8)$$

$$L = D - W \quad (3.9)$$

对连接图进行切割的目标是最小化各个子图连接边的和  $\min Cut(A_1 \dots A_k)$ ，一般采用 NCut 方式。通过构建指示矩阵  $H$ ，其中  $vol(A_j)$  是子图  $j$  中所有点的度（degree）之和：

$$h_{i,j} = \begin{cases} 0 & v_i \notin A_j \\ \frac{1}{\sqrt{vol(A_j)}} & v_i \in A_j \end{cases} \quad (3.10)$$

该方法将目标函数转换为式 3.11：

$$\arg \min_H tr(H^T L H) \text{ s.t. } H^T D H = I \quad (3.11)$$

通过构造  $H = D^{-\frac{1}{2}} F$ ，目标函数转为：

$$\arg \min_F tr(F^T D^{-\frac{1}{2}} L D^{-\frac{1}{2}} F), \text{ s.t. } F^T F = I \quad (3.12)$$

最小化目标函数问题被转为求  $D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$  的  $k$  个最小的特征值问题，对求得特征矩阵  $F$  进行原型聚类（比如 K-means）即可得到谱聚类的最终结果。

谱聚类相比 K-means 聚类而言处理高维稀疏数据的速度更快且效果更好，因其只需要数据之间的相似度矩阵且引入了降维。但其聚类结果比较依赖于相似矩阵的生成模式，不同的相似距离函数得到的聚类结果可能迥然不同。

### 3) Hierarchical Clustering

意为“层次聚类”，指通过采用“自下而上”聚合策略或“自上而下”的拆分策略方式形成树形的聚类结构，常用的层次聚类算法有 Agglomerative Clustering (AGNES)，它将每个样本看成一个初始簇，之后每迭代一步就将距离最近的两个簇合并知道达到预设的聚类簇个数<sup>[25]</sup>。度量聚类簇之间的距离最常用的是平均距离：

$$d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{z \in C_j} dist(x, z) \quad (3.13)$$

距离函数  $dist(\cdot, \cdot)$  的选取可参照 3.1 节的叙述，默认采用欧式距离。

### 4) DBSCAN

该算法为密度聚类算法（density-based clustering）中的一种，此类算法从样本密度的角度考察样本之间的可连接性，基于这种可连接性不断扩展簇，最终得到聚类结果。

DBSCAN 基于邻域参数  $(\epsilon, MinPts)$  描述样本之间的紧密性，具体概念见文献《机器学习》<sup>[25]</sup>，本文只简述保证研究逻辑通顺的必要概念。

给定数据集  $D = \{x_1, x_2, \dots, x_m\}$ ，该算法定义了以下概念：

- a)  $\epsilon$  领域：对  $x_j \in D$ ，其邻域包含样本集  $D$  中与该点距离不大于  $\epsilon$  的所有样本
- b) 核心对象： $\epsilon$  邻域的样本数量至少为  $MinPts$  的点
- c) 密度直达：在核心对象邻域中的所有点都称为由该点密度直达
- d) 密度可达：是密度直达直递性的表现

- e) 密度相连: 对  $x_i$  与  $x_j$ , 若存在  $x_k$  使  $x_i$  与  $x_j$  均由  $x_k$  密度可达, 则称点  $i$  与  $j$  密度相连

基于以上概念, DBSCAN 中定义的簇是满足以下两个性质的样本子集:

- a) 连接性 (connectivity):  $x_i \in C, x_j \in C \Rightarrow x_i$  与  $x_j$  密度相连  
b) 最大性 (maximality):  $x_i \in C, x_j$  由  $x_i$  密度可达  $\Rightarrow x_j \in C$

该算法先任选数据集中的一个核心对象为“种子”(seed), 以任意核心对象为出发点找出由其密度可达的样本生成簇, 直到所有核心对象均被访问过为止。该算法可以检测出异常值, 但是对参数设置比较敏感。

### 3.3.4 算法性能比较与评估

为了直观的观察样本的分布, 需要先把原始样本降维到 3 维或者 2 维形式然后作散点图。本文采用 PCA (主成份分析) 将原始样本降为 3 维, 前三个特征的单变量方差贡献率分别为 0.544、0.249 和 0.091, 累计贡献率超过 80%。作散点视图 3.10, 可见原始样本有比较明显的类别区分, 适合对其进行聚类分析。

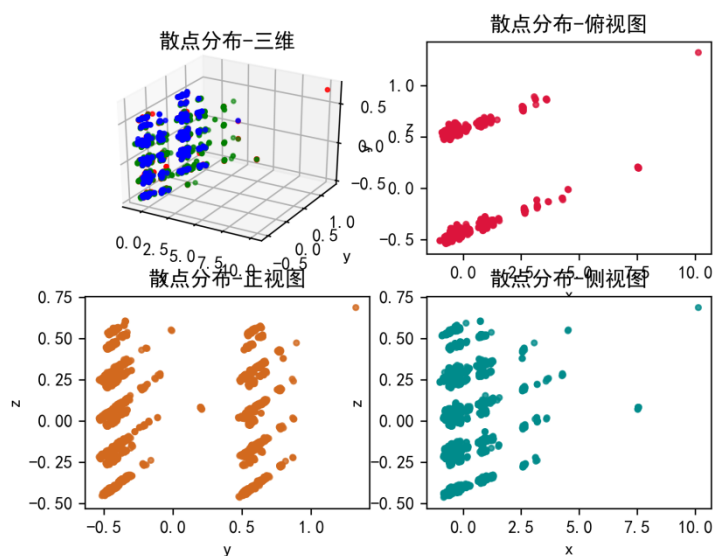


图 3.10 管道属性样本散点分布图

无监督聚类任务中数据都没有标签, 所以并不能对分类的正误直接评价。对聚类结果的评估已经有很多研究, 本文通过计算聚类簇的 Calinski-Harabaz Index 对聚类结果进行评估, 同时可视化每种聚类结果的分类散点图, 综合比较得到适合任务的最佳算法。

Calinski-Harabaz Index (CH index) 通过评估类的分离情况来决定聚类质量即类内越紧密, 类间距离越小则质量越高, 该方法计算迅速且易于互相比, 但该方法有时对密度聚类算法 (比如 DBSCAN) 给出不符合实际情况的高分, 因为该指数在凸簇 (convex cluster) 上会更大<sup>[33]</sup>, 需要对此给予关注。

表 3.5 K-means 聚类结果

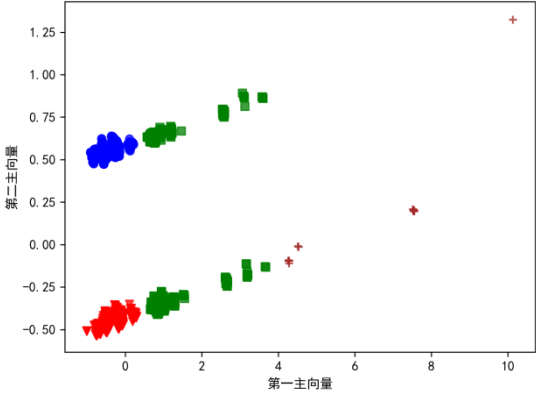
算法名称 & CH 指数	聚类结果散点图
<div>K-means</div> <div>参数:</div> <div>n_clusters = 4</div>	<div>管道属性样本散点图 (PCA降维)</div> 
CH index: 8171.02141577	

表 3.6 Spectral Clustering 聚类结果

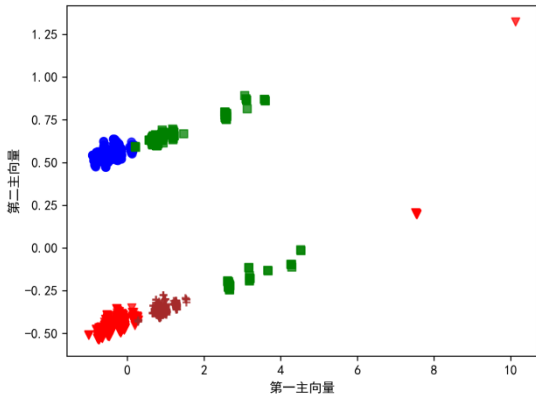
算法名称 & CH 指数	聚类结果散点图
<div>Spectral Clustering</div> <div>参数:</div> <div>n_clusters = 4</div> <div>gamma = 0.5</div> <div>affinity = "rbf"</div>	<div>管道属性样本散点图 (PCA降维)</div> 
CH index: 5427.30246242	

表 3.7 Hierarchical Clustering 聚类结果

算法名称 & CH 指数	聚类结果散点图
<p>Agglomerative Clustering</p> <p>参数:</p> <p>n_clusters = 6</p> <p>affinity = "euclidean"</p> <p>linkage = "ward"</p>	<p>管道属性样本散点图 (PCA降维)</p>
CH index: 8147.62242851	

表 3.8 密度聚类结果

算法名称 & CH 指数	聚类结果散点图
<p>DBSCAN</p> <p>参数:</p> <p>eps = 0.9</p> <p>min_samples = 100</p> <p>metric = "euclidean"</p>	<p>管道属性样本散点图 (PCA降维)</p>
CH index: 5165.90293589	<p>注: DBSCAN 不能指定簇的类别数,此处样本被分为了 2 类, 其中黑点是密度聚类算法分离出的噪声点</p>

表 3.9 聚类分簇的描述统计量

类别	统计量	长度(m)	管径(mm)	管龄(d)
1	mean	63.27	275.11	4778.53
	std	83.13	71.68	1052.47

续表 3.9

类别	统计量	长度(m)	管径(mm)	管龄(d)
2	mean	18.05	107.94	3310.61
	std	31.29	11.61	1117.32
3	mean	66.53	251.82	3853.43
	std	178.72	64.22	1121.97
4	mean	9.35	24.82	3682.22
	std	27.36	1.00	657.62
5	mean	23.29	107.92	4339.45
	std	42.44	6.69	1054.36
6	mean	39.59	55.04	3835.2
	std	76.42	4.54	1058.23

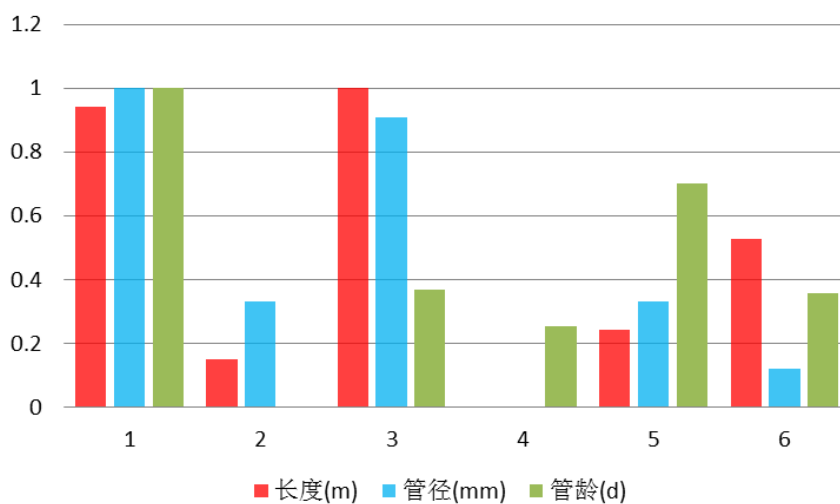


图 3.11 不同类别管道不同属性的均值大小(已缩放至 0,1 区间)



图 3.12 管段聚类结果的地理位置分布

总体来说,采用层次聚类的效果比较好,基本上能将几个簇干净地分隔开,对采用由层次聚类分离的几个类别进行描述性统计见表 3.9 所示,不同类的几个属性均值的直方分布见图 3.11 所示。为了更直观的展示管道分类效果,将聚类结果投射到了地理分布图 3.12 上。

### 3.4 管道事故的地理分布

本文从燃气公司得到 2015 至 2017 年内的管道异常检修记录共 738 条,每条记录大致记录了异常发生的地点描述、事故类型和检修经过。事实上,管段的 GIS 数据、测点 SCADA 数据以及事故发生记录来自三个不同的系统,因此无法直接对三方数据综合利用。本文对检修地点描述利用百度地图 API 解析得到其大致经纬度坐标,使得事故记录可以与管道和测点相对应,从而可以引入事故频率给与聚类簇分类风险度的定义。

根据已得的事实的经纬度坐标点画出事故发生的地理散点分布图 4.1,点越密集的地点则管道事故的发生频率高,且此处的管道或测点事故风险应处于一个较高的水平。为了直观地观察事故风险度的地理分布情况,本文绘制了事故发生的地理概率分布热图 4.2。该图由原始分布样本拟合而成的高斯核密度估计(Gaussian Kernel Density Estimation, Gaussian KDE)生成,可以很直观地描述管道的风险度地理分布。



图 3.13 事故地理分布散点图(2015-2017)

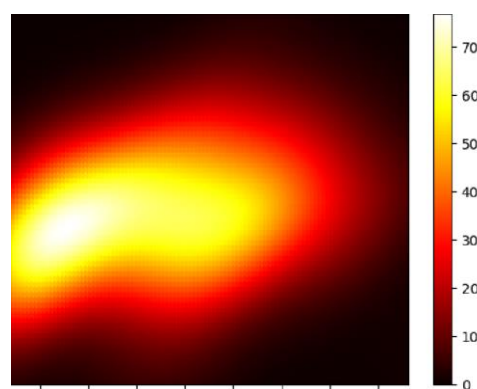


图 3.14 事故地理频率分布热图(2015-2017)

核密度估计(KDE)是一种给定样本集合求解随机变量的分布密度函数的非参数估计方法,非参数检验往往不假定总体的分布类型,直接对总体的分布的某种假设(如对称性、分位数大小等)作统计检验<sup>[38]</sup>。本文在此简述其概念,详见文献[38]或其他相关统计学文献。要观察任意一个有限的样本集合的分布,最直观的方式就是画出其直方分布图。假设该样本的真实分布概率密度函数为 $f(x)$ ,可以用直方图横轴 $x$ 处的邻域 $(x-h, x+h)$ 内的样本数估计该点处的概率密度函数,记为:

$$\hat{f}(x) = \frac{1}{2h} \lim_{h \rightarrow 0} \frac{N_{x_i \in [x-h, x+h]}}{N} \quad (3.14)$$



为了使估计的分布概率函数不对选择的直方个数敏感，分布函数可由一个核函数  $K(x)$  表达：

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x-x_i}{h}\right) \quad (3.15)$$

其中  $h$  是带宽，选用越大的带宽估计得到的密度函数会越平缓，核函数的选择多种多样，仅需要满足以下三个条件即可：

$$\begin{aligned} \int \kappa(v) dv &= 1 \\ \kappa(v) &= \kappa(-v) \\ \int v^2 \kappa(v) dv &< \inf \end{aligned} \quad (3.16)$$

本文选择基于高斯核的核密度估计，即采用标准正态分布密度函数作为核函数估计样本分布。此处对事故的空间分布为二维的密度估计，绘制的事事故地理分布的空间概率曲面如图 4.3 所示，可见事故的地理分布具有明显的聚集性，燃气公司如果加强高风险密度区的安全管理工作，将有效地控制燃气事故的发生频率。

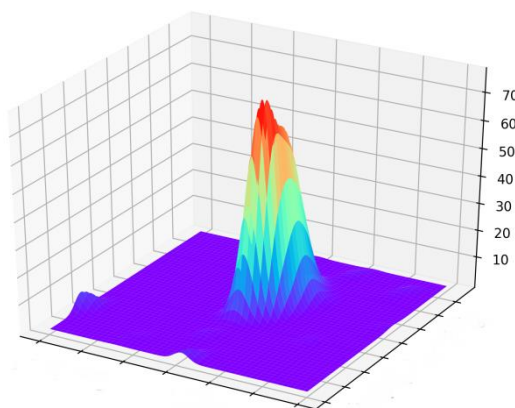


图 3.15 事故的空间概率密度曲面

### 3.5 聚类簇的风险评估

到此，本文已经得到了管道的分类标签、测点的时序典型走势线和事故记录发生地的经纬度坐标。可根据事故记录的经纬度坐标对应到相应的管段，比如当得到某事故地点坐标，分别计算其与各管段坐标的曼哈顿距离(Manhattan Distance)并找到距离最小的管段，则该管段可看作该事故发生的管道。由于管段数据可能存在缺失，部分管段可能未被收录，为了避免事故记录被张冠李戴的情况，本文设定可接受最小的距离为 0.001，超出该大小的事故记录将被忽略。经统计，各个管段分类簇的事故发生频数和事故样本比见下表，数值经过[0,1]缩放后的数量直方图见下图：

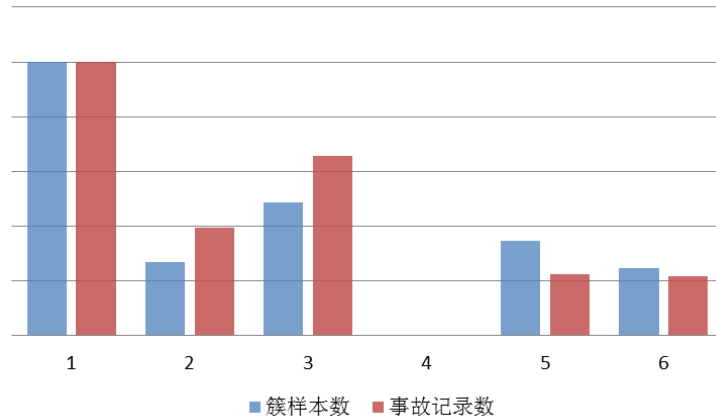


图 3.16 事故数与簇分类的条形分布

表 3.10 基于管段聚类簇分类的事故记录统计表

管段簇分类	簇样本数	事故记录数	相对事故比
1	5572	186	1
2	1517	73	1.464
3	2719	122	1.352
4	32	0	1
5	1953	42	0.651
6	1390	40	0.877

备注： 相对事故比=事故记录/簇样本(分子分母均已经 0,1 缩放)

若对原始样本集进行随机抽样将其分为 6 个大小不一的簇，则每个簇的期望事故发生频数应该与每个簇的样本数成正比，即第  $i$  个簇的期望事故频数为：

$$E(N_{F,i}) = \frac{n_i}{\sum n_k} \times N_F \quad (3.17)$$

其中  $N_{F,i}$  为第  $i$  类簇的事故总数， $n_i$  为第  $i$  类簇的管段样本总数， $N_F$  是所有事故记录数。为了检验聚类结果是否符合非随机抽样，换言之，簇的分类反映了不同管段集合风险的显著差异，本文对聚类结果进行了卡方检验(Chi-squared Test)，检验结果见表 3.11，检验的原假设(Null Hypothesis)和备选假设(Alternative Hypothesis)描述如下：

- 原假设  $H_0$ : 分类簇的实际事故频数同随机抽样的期望事故频数分布一致；
- 备选假设  $H_1$ : 分类簇的实际事故频数同随机抽样的期望事故频数分布不一致；

表 3. 11 基于簇分类的事故数量分布的卡方检验结果

簇分类	实际事故频数	期望事故频数	统计量
1	186	196.69	0.480222
2	73	53.28	7.300049
3	122	95.49	7.357215
4	0	1.12	1.123872
5	42	68.59	10.30884
6	40	48.82	1.592854
Chi-square			28.16305
自由度(df)			5
P<0.05			11.0705
P<0.01			15.08627

卡方检验结果表明，两个分布之间的  $\chi^2=28.16305$ ，而  $F_{0.05}(5)=11.0705$ ， $F_{0.01}(5)=15.08627$ ，因此至少有 99% 的把握拒绝原假设  $H_0$ ，即确定了管段分类结果反映了不同管段簇之间存在客观的风险度差异。

综合图的直观观察和表中的事故样本比，可以看到簇 5 和簇 6 的相对风险比较低，簇 2 和簇 3 的风险比很高。总的来说，分类的风险度从危险到安全的排序为：2, 3, 1, 4, 6, 5。可见前文所做的管段聚类工作成功地分离出了客观上具有不同风险等级的管段，并对工程上管段的风险度定义具有重要的参考价值。

## 4 SCADA 测点序列聚类及实时异常检测模型

SCADA(Supervisor Control And Data Acquisition)系统,即监测监控及数据采集系统。它可以实时采集现场数据,对工业现场进行本地或远程的自动控制,对工艺流程进行全面、实时的监视,并为生产、调度和管理提供必要的数据<sup>[34]</sup>。SCADA 系统在城市燃气管网上的应用在国内还处于推广阶段,天津燃气集团的工程实践表明,SCADA 系统能很好地完成日常的生产调度,保持管网输配的平衡,提高燃气生产和管理效率,降低成本,节约能源等目标,具有很高的实用性<sup>[35]</sup>。

本文所做的时序数据聚类的数据来源于 SCADA 的流量计量系统。SCADA 的流量计量系统被用来采集各个大用户的压力、温度、瞬时流量、累计流量等运行参数,将参数采集到终端,再通过专线上传至调度中心<sup>[35]</sup>。由于设立测点的成本较高,实际测点个数相对于管段数量要少得多,本文所有的测点个数不到 50 个,而管段个数有超过 13000 个,测点的分布极为稀疏,这就导致实际应用中基于测点时序数据对具体管段实时情况监控的困难。最自然的想法是增加测点,使单个测点覆盖的区域尽可能小从而得到尽可能精确的实时数据,然而限于实际成本,人们更多地期望通过充分利用现有的测点数据达到较为精确的监测结果。本章在此思路的基础上,通过对各个测点的一年内时序数据进行聚类分析,筛选出几类特征明显的日内时间序列,然后再针对各类管道内压的日内走势线(以下简称日线建立异常检测模型。

### 4.1 时序集交叉距离矩阵

本文使用了超过 40 个测点在 2015 年全年的流量计系统数据,每个测点每天的样本超过 2000 条且仪表接收测点数据的时间点不尽相同,因此本文对原数据采取了每小时定点抽样的处理,将每天的样本点固定在 24 个,减小了计算量并且取得了序列降噪的效果。某测点在若干天的日内内压数据的走势见下图:

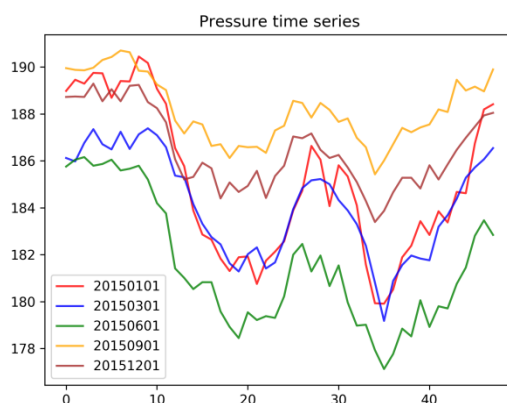


图 4.1 某测点的内压日内走势

很明显地观察到同一测点在一天内具有明显的周期性，在每天的午夜 0 点至 3 点和正午 13 点至 15 点时达到顶峰，在 9 点和 18 点左右是底部，且不同的日内序列存在纵轴上的绝对值差异，例如图中 9 月份比 6 月份平均高 4 个单位左右。

当计算不同的序列之间的相似度时，首先需要确定序列间的距离度量。给定两个等长的时间序列  $A = \{a_1, a_2, \dots, a_m\}$  和  $B = \{b_1, b_2, \dots, b_m\}$ ，最简单的度量是两序列之间的欧式距离 (Euclidean Distance):

$$dist(A, B) = \sum_{i=1}^m (a_i - b_i)^2 \quad (4.1)$$

该方法容易实现且时间复杂度低 ( $O(n)$ )，但无法度量不等长的序列且无法识别局部有拉伸和收缩的相似序列。

动态时间规整 (Dynamic Times series Warping, DTW) 是一种最初被用于语音识别的时间序列相似性度量，语音识别领域中由于不同人的语速不同，同一个单词的发音序列在时间上可能形态上非常不同。该算法基于动态规划 (Dynamic Programming) 的思想，将序列在时间轴上扭曲 (warping)，从而使非等长序列尽可能的对齐，再计算两个序列之间的距离<sup>[36]</sup>。简单来说，即扭曲后的两序列的点之间不再是一一对应地计算距离，而可能是一对多或多对一的关系，见示意图 3.14:

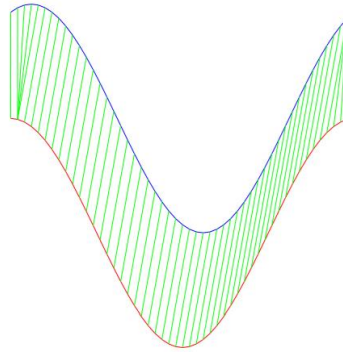


Figure 1: Two 1D sequences aligned with Dynamic Time Warping. Coordinates of the top and bottom sequences have been respectively computed by  $\cos(t)$  and  $\cos(t + \alpha)$ . For visualization purpose, the top sequence is drawn vertically shifted.

图 4.2 经过 warping 处理的两序列各点的对应关系示意图<sup>[36]</sup>

DTW 算法概念在文献 [Derivative Dynamic Time Warping, 2001]<sup>[37]</sup> 中被详细叙述，本文在此仅简述其基本概念。当有两个不等长时间序列  $Q = \{q_1, q_2, \dots, q_n\}$  和  $C = \{c_1, c_2, \dots, c_m\}$  时，对齐两个序列可以建立一个维度为  $n \times m$  的矩阵  $D$ ，其中  $D(i, j) = d(q_i, c_j)$ 。定义规整路径 (warping path) 为  $W$  :

$$W = w_1, w_2, \dots, w_k, \dots, w_K \quad \max(m, n) \leq K \leq m + n - 1 \quad (4.2)$$

$W$  是一段定义为两个序列之间的映射 (mapping)，其中  $w_k = (i, j)_k$ 。规整路径  $W$  必须满足以下三个约束条件:

a) 边界条件(Boundary conditions)

$w_1 = (1,1), w_K = (m,n)$ ，即保证路径从矩阵 $W$ 的左下角开始到右上角结束；

b) 连续性(Continuity)

如有  $w_k = (a,b)$ ，则其前一个对象  $w_{k-1} = (a',b')$  必须满足  $a-a' \leq 1, b-b' \leq 1$ ，即路径的前后两点必须相邻；

c) 单调性(Monotonicity)

如有  $w_k = (a,b)$ ，则其前一个对象  $w_{k-1} = (a',b')$  必须满足  $a-a' \geq 0, b-b' \geq 0$ ，即路径的前后两点必须在时间顺序上单调；

一个典型的 $W$ 路径如下图所示：

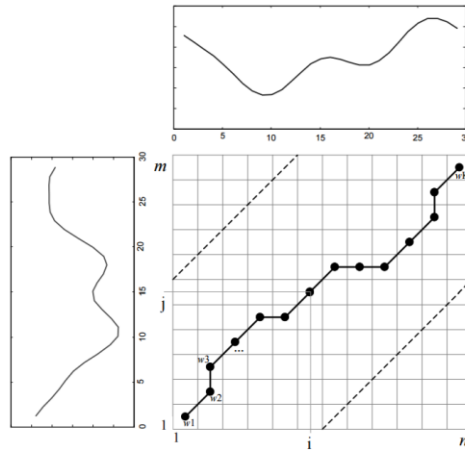


图 4.3 一个典型的规整路径样例<sup>[37]</sup>

定义累计距离 $\gamma$ ：

$$\gamma(i, j) = d(q_i, c_j) + \min\{\gamma(i-1, j-1), \gamma(i, j-1), \gamma(i-1, j)\} \quad (4.3)$$

这个累计距离可以通过动态规划(DP)方法计算得到，最终得到的 $\gamma(m,n)$ 即序列 $Q$ 和 $C$ 的DTW距离。该算法的时间复杂度为 $O(n^2)$ ，已有许多研究针对DTW计算的加速算法，包括FastDTW、SparseDTW等。

本文计算某测点一年内不同日之间的压力的走势序列的DTW距离，计算DTW之前已把所有序列标准化，所有计算好的距离可以整理成 $D$ ， $D$ 矩阵的维度为 $(365 \times 365)$ ，其中 $D(i, j)$ 表示第 $i$ 天序列和第 $j$ 天序列之间的距离，易得 $D$ 为对角线上全为0的对称矩阵。

## 4.2 序列的两层聚类模型

在得到序列之间的距离度量矩阵 $D$ 之后，即可进行序列的聚类分析工作。如果本文总共要对47个测点的2015年全年序列进行聚类分析，相当于总共有 $47 \times 365 = 17155$ 条日线。若对其计算每两条线互相之间的距离，相当于无放回的组合问题 $C_{17155}^2 = 147138435$ ，即需要计算147138435次DTW距离，时间上几乎不可接受。

针对这种情况,本文提出了分层聚类的思路,即先对特定测点的 365 条日线进行聚类提取出典型日线,再综合所有测点的典型日线进行聚类分析。比如,若第一层聚类每个测点共有 8 条典型日线,则第二层聚类需要针对  $47 \times 8 = 376$  条日线进行遍历计算,两层聚类共需计算 DTW 距离  $47 \times C_{365}^2 + C_{376}^2 = 3192710$  次,仅相当于第一种方法计算量的约 2%。除此之外,本文还对 DTW 的计算过程设计了多进程(Multiprocessing)并发处理程序,利用多核 CPU 的 4 个核心同时处理 4 个测点,最大限度地提高计算速度。本文的 PC 平台配置了 Intel i7-4700MQ CPU@ 2.40GHz 和 8.00GB RAM,数据处理过程时的终端运行效果见下图,时序聚类的流程示意图见图 4.5:

图 4.4 多进程并发的 DTW 计算终端运行处理过程

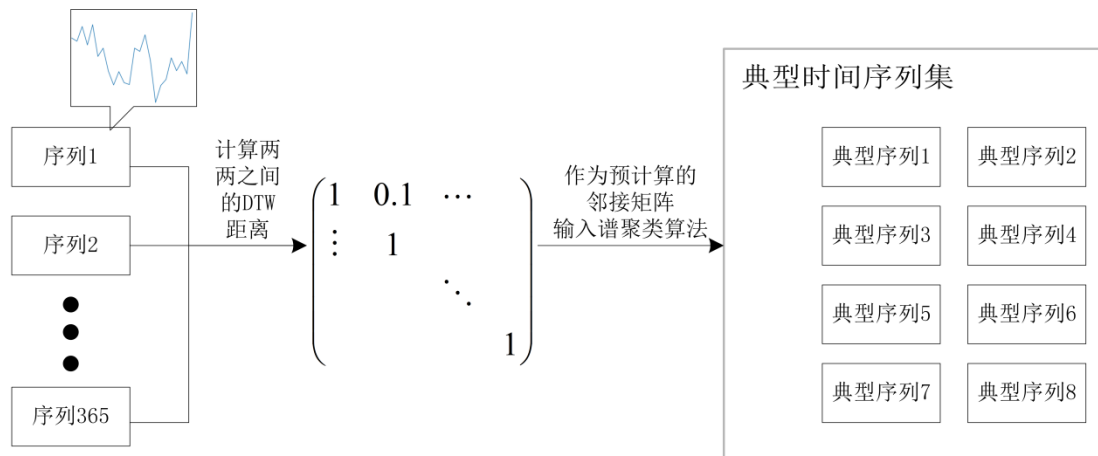


图 4.5 时序聚类的流程示意图

### 1) 第一层聚类

选取某测点计算好的交叉 DTW 距离矩阵  $D$ ，第一层聚类以  $D$  作为提前计算好的邻接矩阵(affinity matrix)输入进行谱聚类(Spectral Clustering)。这里的  $D$  在输入之前应该预处理，原因是邻接矩阵要求两样本越相似则边权重越大，这与  $D$  的两样本越相似距离越小相悖；另外，预处理标准化  $D$  内数值也能使聚类更易收敛。本文采用高斯核(Gaussian Kernel)对  $D$  预处理：

$$W = e^{\frac{-D^2}{2\sigma^2}} \quad (4.4)$$

经调试最终选择的  $\sigma$  为 10，预处理后的  $W$  矩阵对角线上的数值均为 1，其余数值均在 0,1 之间，符合谱聚类邻接矩阵的格式条件。

谱聚类将该测点的 365 条日线分为 8 类，取距离类内其他日线 DTW 距离之和最小的日线作为类的典型日线，该测点的 8 条典型日线走势如图 4.6 所示，可见结合交叉 DTW 矩阵和谱聚类算法的聚类模型成功找到了该测点一年内的典型日内管压力走势模式，可利用该 8 条典型日线作为接下来的第二层聚类的原始输入样本。

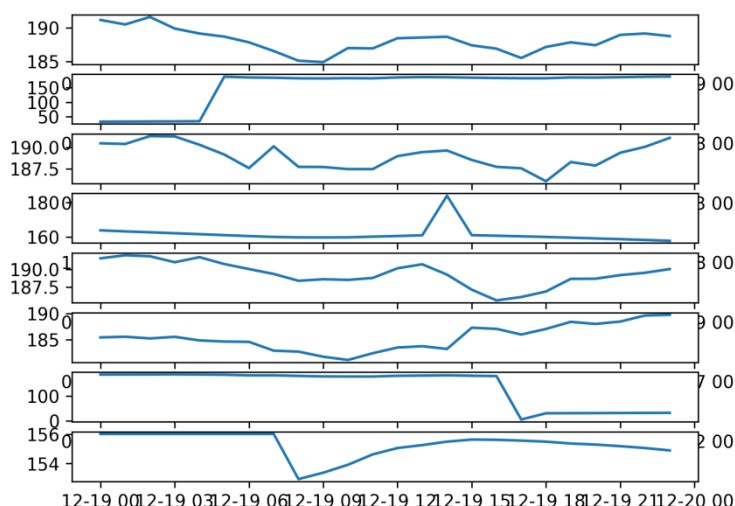


图 4.6 某测点的 8 条典型日内走势线

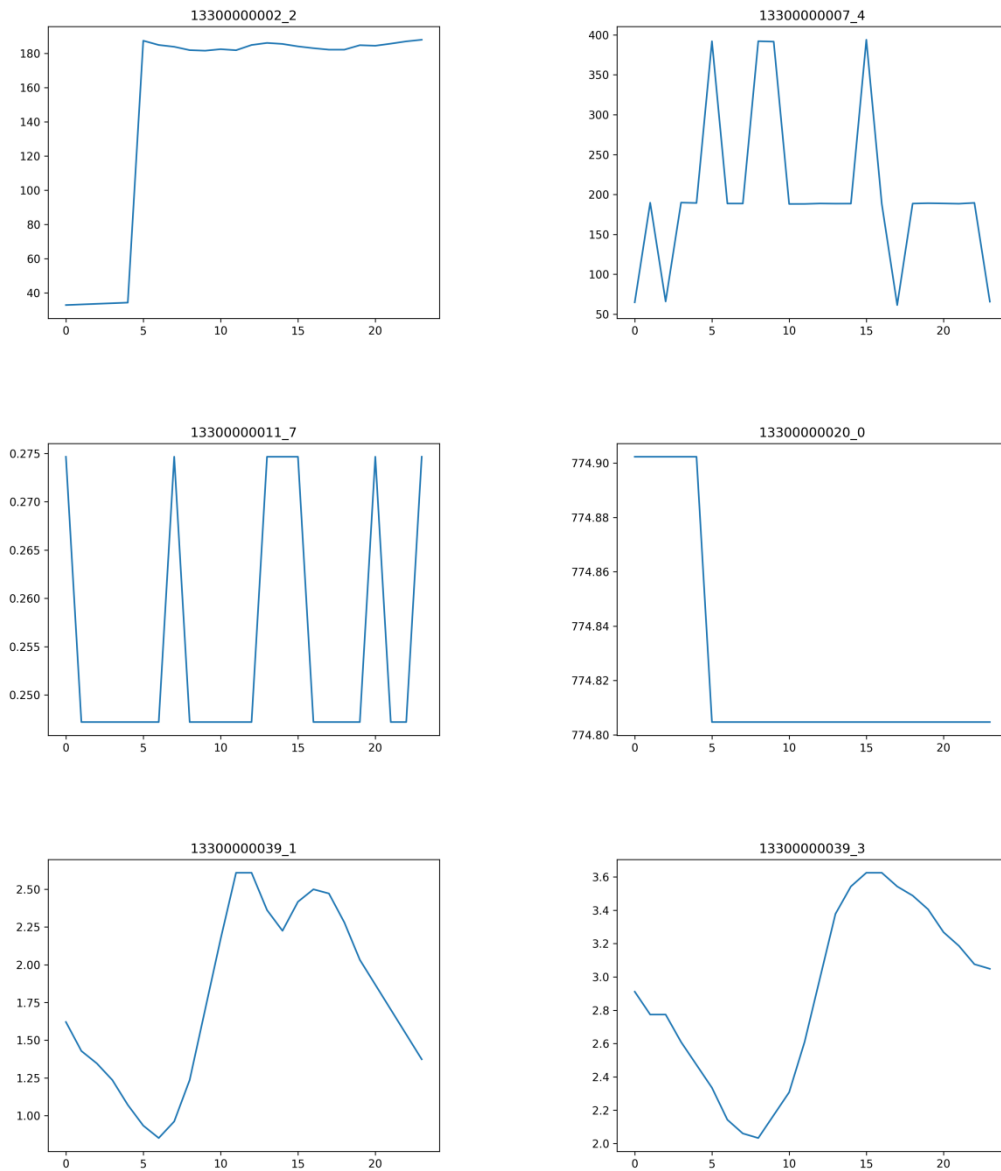
### 2) 第二层聚类

由第一层聚类共得到各测点的典型日线总共 360 条，接下来可以按照上述的构建交叉 DTW 距离矩阵  $D$ ，预处理得到邻接矩阵  $W$ ，运用与谱聚类算法得到日线的分类到最后取得若干类典型日线相同的流程，最终得到第二层聚类的结果。

本文获取了所有测点在 2015 全年的共 10 条代表日线，其走势各异，见图 4.7。通过两层的聚类，本文终于从浩如烟海般的 45 个测点共 16425 条日线，超过 10.0GB 的时序数据中获得 10 条最具代表性的趋势线，实现了海量时序的聚类。算力充足时，可以直接使用第一层聚类得到的 360 条线做走势匹配；而在算力比较有限时，可以仅就这 10 条典型日线进



行异常检测建模。考虑到成本因素，加两层聚类以最大限度的压缩模版数据是基于现实考虑的理性选择。



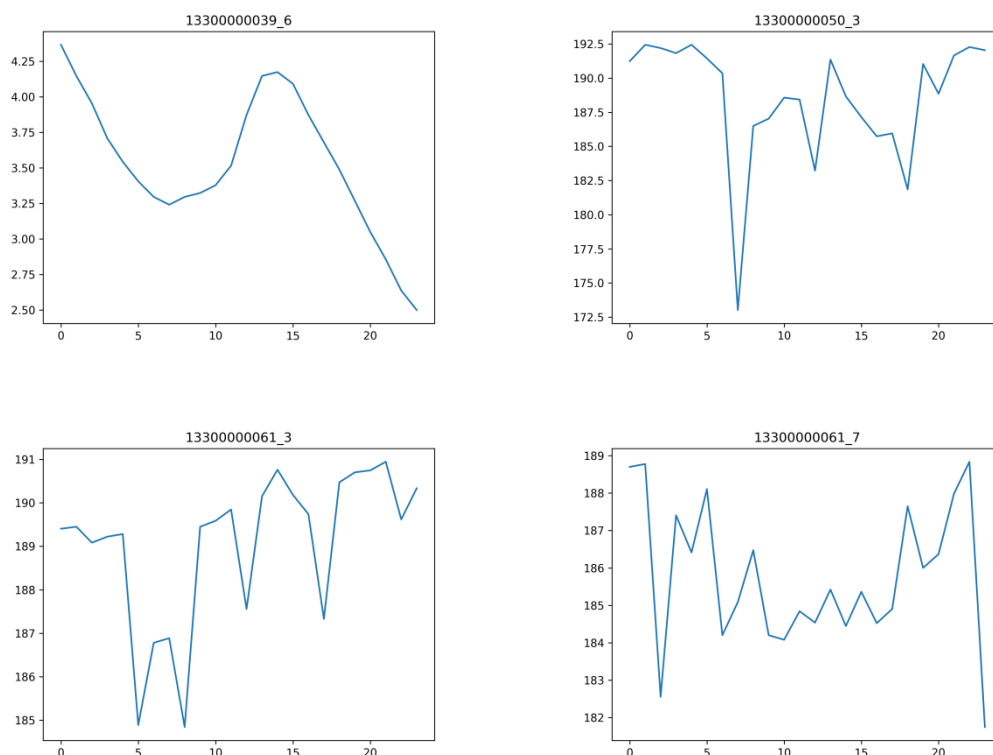


图 4.7 所有测点在 2015 年全年的 10 条典型日内内压趋势线

### 4.3 测点序列聚类簇的风险分析

前文建立了测点的两层时序聚类模型从而得到了 10 条代表性的内压日内趋势线。由于聚类算法是无监督的，即不存在客观的标签能对聚类结果进行评估，本文继续沿着聚类的思路拓展，引入管道的事故记录予以辅助建模，并指定簇类别在工程上的定义。本文希望能够通过发掘不同走势的日线所隐含的管道的运行状态找到高风险的日线模式，并在其后将其运用于时序实时的异常检测模型中。具体的，首先找到事故记录所在地附近最近测点的当天的日线模式，找到模版日线中与其走势最为接近的一条，则该类模版日线的风险度增加。

部分事故当天所在地附近测点内压走势与典型模版日线的匹配结果见下图 4.8 至 4.13 所示，很明显可见事故当日走势与模版日线匹配良好，但是不同事故当日测点日线形态各异，存在多个事故现场迥然不同的走势，其可能的原因有：

- 测点过于稀疏，部分事故地点距离测点距离过远而导致事故造成的内压波动并没有反映到测点的实时数据上；
- 事故程度极为微小，未造成明显的对管道运行的影响，管道仍然处于正常运行状态中；
- 测点仪表故障或传输故障导致部分数据缺失或者出现异常，未能记录实际的数据走势；
- 事故在发现之前已经出现多时，因此事故记录当日日线走势不一定能反映事故的影响

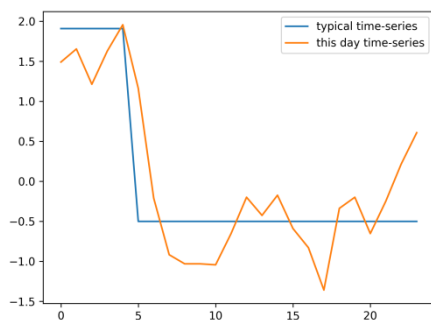


图 4.8 2015/1/4 事故所在地测点日线走势匹配

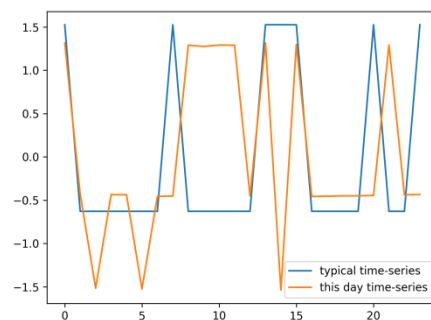


图 4.9 2015/2/4 事故所在地测点日线走势匹配

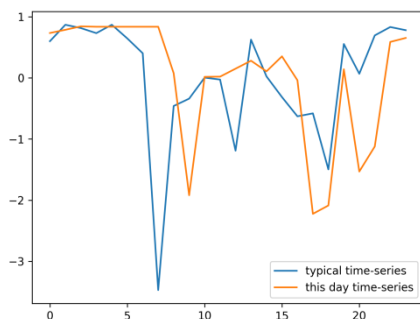


图 4.10 2015/1/28 事故所在地测点日线走势匹配

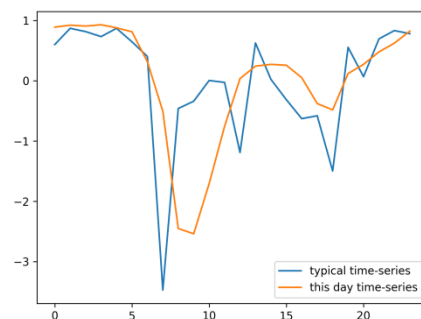


图 4.11 2015/2/2 事故所在地测点日线走势匹配

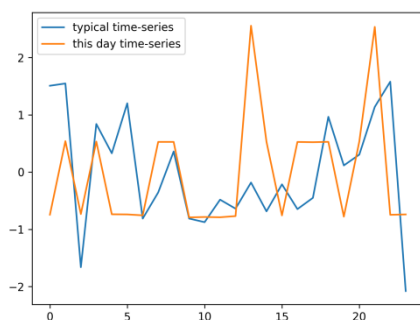


图 4.12 2015/3/1 事故所在地测点日线走势匹配

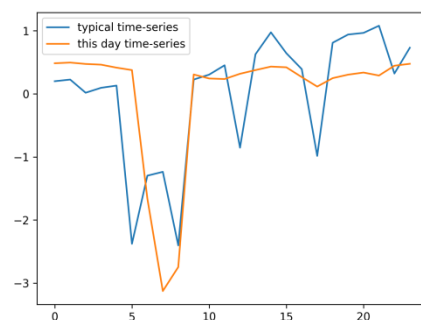
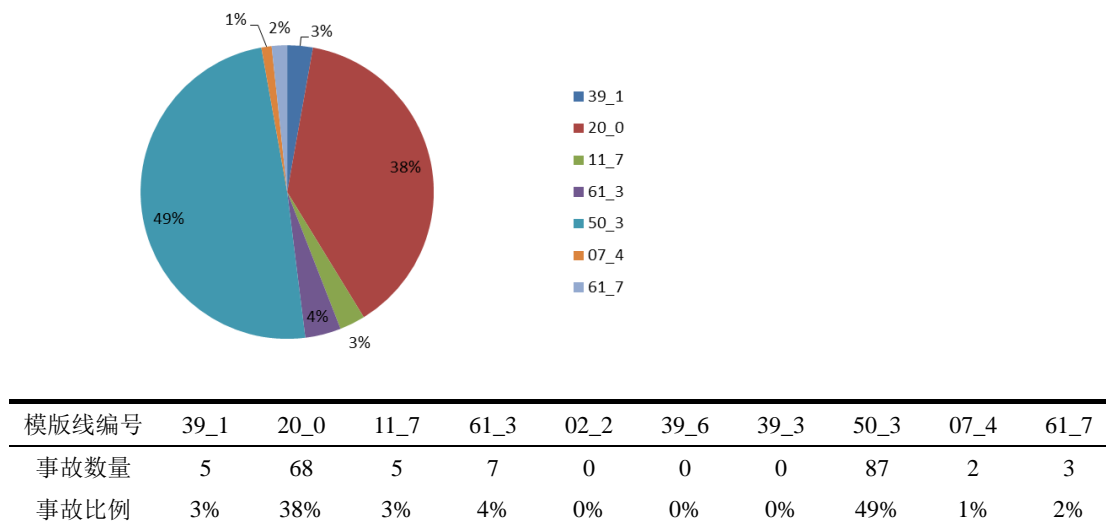


图 4.13 2015/6/23 事故所在地测点日线走势匹配

各模版日线其对应的最接近的事故个数统计结果见表 4.1 所示，模版线编号\*\*\_\*由测点序列号的后两位加上该测点典型日线类型组成。由统计结果可见，事故当日测点日线走势最多与编号为 13300000050\_3 与 13300000020\_0，其分别占比 49%与 38%。该两条线的走势见图 4.11，可以观察到这里两条线的走势都存在特定时点的突发大幅下降的现象，这很可能是由于管道受到外力变形或是出现了破损。

表 4.1 模版日线对应的事故记录数量统计结果(2015 年)



备注：所有模版编号前 9 位均相同，因此在表中仅取其后 4 位

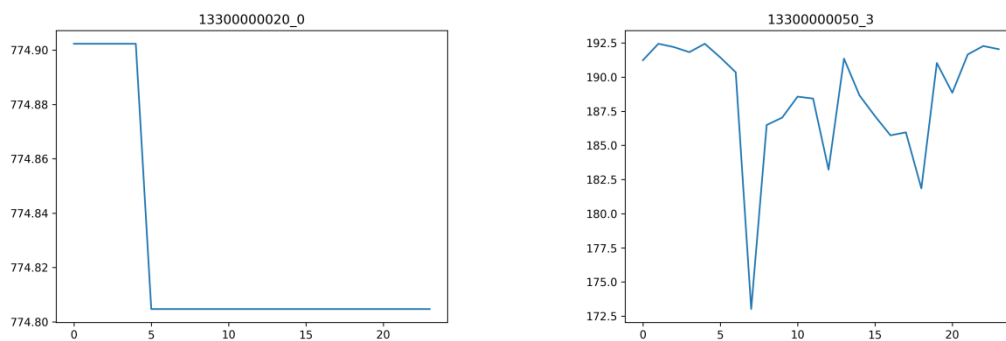


图 4.14 与所有事故发生地测点走势最接近的两条模版线

## 4.4 测点实时数据的异常检测模型

测点数据定时传输数据到达控制中心，工程上最关心实时数据所反映的管道的运行状态，此时可以通过异常检测模型对输入的测点数据进行分析处理并给出对运行状态的评估结果。

本章的前文已经对 2015 年全年共 177 条事故记录进行了模版走势匹配，结果表明匹配模版走势 20\_0 与 50\_3 的事故共有 155 条，占事故总数的 87%，因此有很大把握这两类模版线就是本文在寻找的高风险日线走势模式。为了验证这一想法，本文继续基于发生于 2016 年内的 197 条事故记录进行匹配，匹配结果见表 4.2。

表 4.2 模版日线对应的事故记录数量统计结果(2016 年)

模版线编号	39_1	20_0	11_7	61_3	02_2	39_6	39_3	50_3	07_4	61_7
匹配事故数量	8	127	4	0	0	2	3	43	0	10
比例	4%	64%	2%	0%	0%	1%	2%	22%	0%	5%

由统计结果可见，匹配 20\_0 线和 50\_3 的事故记录分别有 127 与 43 条，分别占比 64% 与 22%，总占比 86%，同 2015 年的该两类线的统计占比 87% 基本相当。为了从另一面验证该模型的准确性，本文在 2016 年共 217 天的序列中每天随机抽取一个测点的日线，并将其与模版线一一匹配，得到统计结果如表 4.3 所示，随机抽取的日线与 11\_7 最为匹配，占比为 88%，除此之外只与 39\_1 达成匹配，占比为 12%，而无日线与高风险模版线 20\_0 与 50\_3 匹配，这样就成功地验证了该两类线为高风险走势的观点。图 4.15 直观展示了事故线和随机抽样线分布的巨大差异，图 4.16 是与随机抽样的日线集合最为匹配的两条模版线。

表 4.3 随机抽样的日线与模版线的匹配结果统计(2016 年)

模版线编号	39_1	20_0	11_7	61_3	02_2	39_6	39_3	50_3	07_4	61_7
匹配日线数量	27	0	190	0	0	0	0	0	0	0
比例	12%	0%	88%	0%	0%	0%	0%	0%	0%	0%

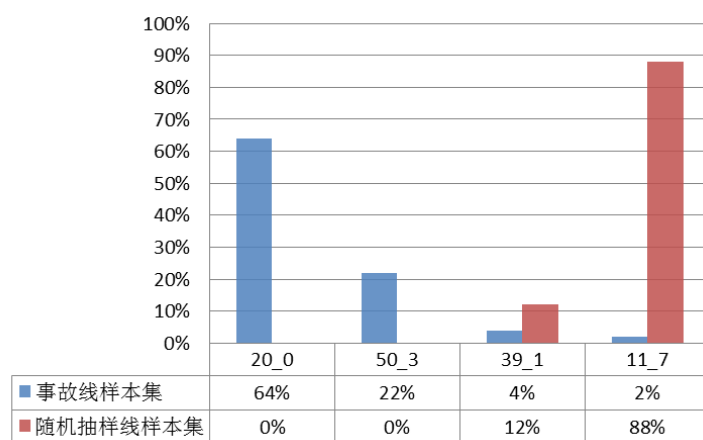


图 4.15 随机抽样日线集与事故日线集与模版匹配度分布差异示意图

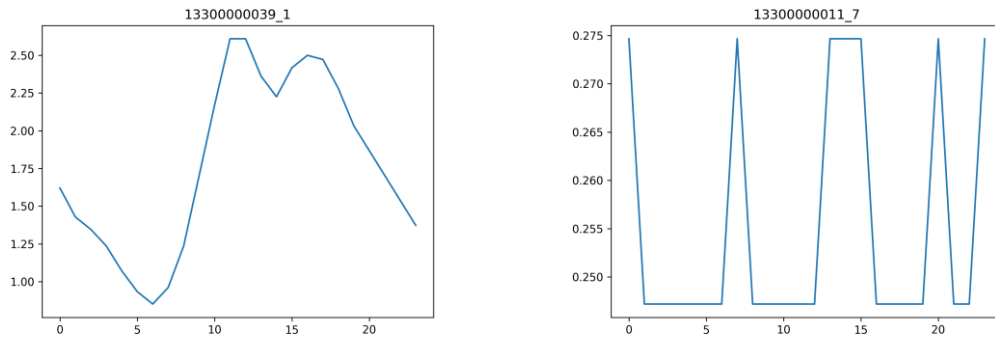


图 4.16 与随机抽样的日线集匹配度最高的两条模版线

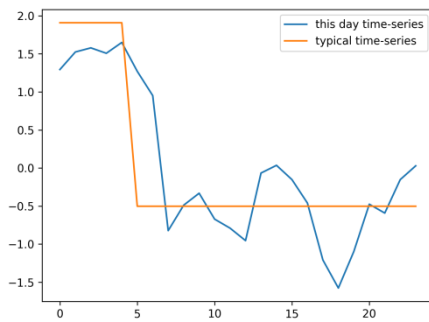


图 4.17 2016/3/7 事故所在地测点日线匹配结果

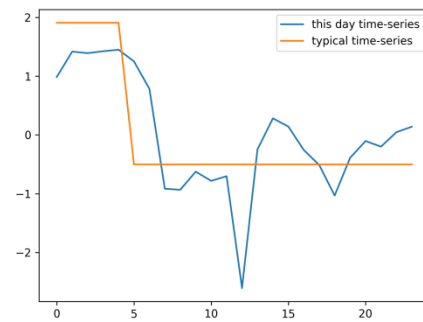


图 4.18 2016/3/4 事故所在地测点日线匹配结果

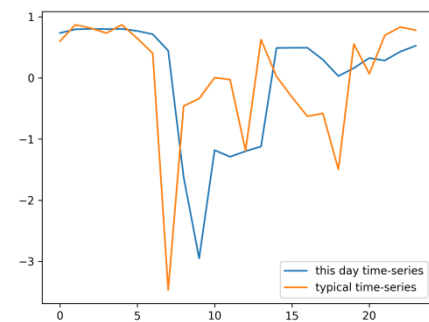


图 4.19 2016/2/2 事故所在地测点日线匹配结果

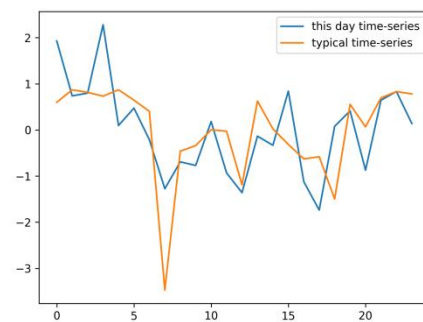


图 4.20 2016/3/27 事故所在地测点日线匹配结果

本文提出的异常检测模型即基于这一发现所建立，根据实时监控的测点内压走势，匹配 10 条典型日线，若实时数据接近高风险走势则表明测点附近管道很可能出现异常，需要派出人员检查和修理。在得到了典型日线模版和高风险日线模版后，将该模型应用于工程实际十分简便，实时计算量非常小，一台普通的 PC 即可完成该检测和计算任务。另外，模型采用的是基于对多个典型模版线的匹配度相互比较的匹配策略，不同于直接去匹配高风险线，该策略不用对匹配度设置阈值，这就使模型更加简单和稳健。比如说，如果直接对高风险线

进行匹配则必须要定义匹配度达到什么程度才能说明其足够匹配,而不同阈值的设定可能会导致迥然不同的模型评估结果,影响模型的可用性、可解释性和稳健性。

## 5 总结与展望

### 5.1 本文所做工作的总结

本文在城市燃气管道的安全问题上，沿着事故报道获取、事故归因分类、安全评估模型以及异常检测模型的逻辑线，展开了综合性研究工作。

本文的主要成果有：

#### (1) 构建了生命线管道事故数据系统

本文构建了生命线管道事故数据系统，该系统集成了城市燃气管道事故报道收集、事故归因统计分析以及图形化展示三大功能。系统程序包括针对博燃网和中国燃气网的燃气事故新闻收集爬虫、事故归因的统计分析和图形化展示脚本、以及针对地方施工招投标公告信息的爬虫和施工点地理分布的可视化脚本。其中，事故爬虫共爬取新闻 8400 余条，经编写的自动化脚本分析得到 380 余条处理好的新闻报道，将事故原因总结为老化腐蚀、地面移动、误操作、第三方破坏和天气五个因素，将第三方破坏再分解为个人、车祸和施工破坏三类，针对施工破坏因素细化为挖掘机、堆载等八个因素。

#### (2) 提出基于多种静态属性的管段聚类模型及安全性评估方法

本文从某地燃气管道网络的 GIS 系统获得超过 13000 根现役燃气管道的多种属性，从中选择了有无警示标志、管径、管材、长度、建成日期、设计内压、运行内压、安装工程甲方和乙方等属性作为原始样本集。针对所有类别属性和部分数值属性采用引入先验知识的方式将其转换为适合聚类算法的数值型数据，如根据规范的材料设计使用年限定义不同管材对应的权值；根据 EGIG 统计的 20 年事故概率离散化管径数据并赋予区间的权重；将建成日期分解为管龄和建成季节两个属性，季节按温度从低到高赋予权值；将安装工程的甲方和乙方公司的注册资本作为其权值等。在样本上尝试了 K-means、DBSCAN、层次聚类和谱聚类算法，比较分类结果并确定层次聚类作为最佳算法，管道被分为 6 类。

对聚类结果基于 2015 到 2017 年的管道事故记录进行实证检验，发现聚类结果确实反映了不同簇内的管道之间存在客观的风险度差异。综合分析不同簇内样本的多种统计值和事故样本比，确定类别风险度从危险到安全的排序为：2，3，1，4，6，5。

#### (3) 提出基于管道 SCADA 测点数据的时序聚类模型

本文提出一种两层聚类的架构用于寻找 SCADA 测点日内典型的走势形态。通过计算各个日线互相之间的 DTW 距离构建交叉距离矩阵，利用该矩阵构建连接图并基于谱聚类算法进行图分割得到若干典型的日线形态。第一层聚类是基于 45 个测点各自 365 条日线进行的，第二层聚类则基于第一层聚类结果，即每个测点的 8 条典型日线。将该聚类架构应用于 2015 年内超过 10GB 管道内压时间序列数据，共挖掘出 10 种形态各异的典型日线走势。

#### (4) 建立基于模式匹配的 SCADA 测点数据的实时异常检测模型

管道的危险运行状态会反映在测点的异常走势上，因此本文基于 2015 年内的 177 条事故记录当天的测点走势分别对模版线进行匹配，发现两类与事故地测点日线最匹配的模版线，占比达到事故记录数的 87%。针对这一发现建立异常检测模型，即对测点每日实时走势与



模版线分别计算匹配度，若其最匹配高风险日线则说明很可能存在异常。在 2016 年的共 197 条事故地日线进行模型实证检验，匹配高风险日线事故地日线共 170 条，占比达到 86%，另外用从 2016 年随机抽样取得的 217 条日线匹配模版线，发现无一条匹配高风险序列，这证明了该异常检测模型的有效性。

本文的创新之处有：

1. 进行了自动化燃气管道事故数据库构建的尝试，填补国内缺乏管道事故数据库的空白。完善事故数据库和事故归因统计报告将大大促进相关研究工作，提高我国燃气管道的精细化、规范化和现代化管理水平，降低事故发生率以及减小安全维护成本。
2. 引入聚类算法进行管段分类和安全状态评估，不同于以往基于项目打分的评估模式，该方法所使用的数据集均为管道的客观数值属性和先验概率赋权的类别属性，避免了人为主观误差对评估结果的干扰。
3. 采用双层聚类架构进行时序聚类，使计算量降低到单层聚类的 2%；使用 DTW 距离度量序列相似性，可以扩展到不等长和伸缩形状的序列比较上；采用模版匹配的方式寻找实时数据的异常走势计算量小、精度高，工程上应用起来十分方便。

## 5.2 下一步研究方向的展望

本文的研究尚存在一些不足之处，未来的研究可以基于以下几个方向进行改进：

1. 聚类所采用的原始属性数据并不丰富，如果能继续扩充可用的属性数据，如埋地管周围的土质类型、湿度、pH 值，管道埋深，管道的应力比等能反应管道风险度的数据，将提高聚类结果的精确性。
2. SCADA 测点比较稀疏，部分事故地点距离测点距离过远而导致事故造成的内压波动并没有反映到测点的实时数据上，密铺测点使每个测点的检测范围更加缩小将提高异常检测模型的精度和准确性。
3. 可以在管道分类评估中继续引入主动学习(Active Learning)，这是一种半监督方法(Semi-Supervised)，对聚类算法最不确定其分类的样本赋予标签，提高评估分类的准确性和精度。

## 参考文献

- [1] 英国石油公司(BP). BP2035 世界能源展望[EB/OL].  
[https://www.bp.com/content/dam/bp-country/zh\\_cn/Download\\_PDF/Report\\_BP2030EnergyOutlook/EO2035\\_Chinese\\_Version.pdf](https://www.bp.com/content/dam/bp-country/zh_cn/Download_PDF/Report_BP2030EnergyOutlook/EO2035_Chinese_Version.pdf), 2014-01-01
- [2] 王晓梅. 城市埋地燃气管道的风险评价[D]. 南京:南京工业大学, 2006.
- [3] 郭峰. 城市地下燃气管道风险评价研究[D]. 长沙:中南大学, 2011.
- [4] 韩朱旻. 城市燃气管网风险评估方法研究[D]. 北京:清华大学, 2010.
- [5] W.Kent Muhlbauer. Pipeline Risk Management Manual[M]. U.S.:Elsevier, 2004.
- [6] Jaffee A S;Jeff M J;Anthony S,et al. Fire and explosion as-sessment on oil and gas floating production storage offloading(FPSO):An effective screening and comparison tool[J]Process Safety and Environmental Protection, 2009,(2).
- [7] 黄超等. 城市燃气管网的故障传播模型[J]. 清华大学学报(自然科学版), 2008, 48(8): 1283~1286
- [8] 博燃网. 2017 年我国燃气爆炸事故分析报告[EB/OL].  
<http://www.gasshow.com/article/detail?id=964>, 2017-01-03
- [9] 国家能源局石油天然气总公司. 中国天然气发展报告(2016)[R]. 北京:国土资源部油气资源战略研究中心, 2016.
- [10] Fred Henselwood et al. A matrix-based risk assessment approach for addressing linear hazards such as pipelines[J]. Journal of Loss Prevention in the Process Industries, 2006, (19): 433-441
- [11] 李大全等. 模糊聚类法在油气管道风险评价管段划分中的应用[J]. 集输工程, 2012, 32(7): 63-67
- [12] D Xu et al. A Comprehensive Survey of Clustering Algorithms[J]. Annals of Data Science, 2015, 2(2): 165-193
- [13] 张杰. 基于主成分-聚类分析法的管道风险评价方法[J]. 油气储运, 2014, 33(2): 139-143
- [14] Jiansong Wu et al. Probabilistic analysis of natural gas pipeline network accident based on Bayesian network[J]. Journal of Loss Prevention in the Process Industries, 2017, (46): 126-136
- [15] 王晓波等. 基于事故树与贝叶斯网络的管道泄漏事故溯源方法[J]. 油气储运, 2017, 36(9): 1013-1018
- [16] Alireda Aljaroudi et al. Probability of Detection and False Detection for Subsea Leak Detection Systems: Model and Analysis[J]. TECHNICAL ARTICLE—PEER-REVIEWED, 2015, (15): 873-882
- [17] Jingwei Qiu et al. The early-warning model of equipment chain in gas pipeline based on DNN-HMM[J]. Journal of Natural Gas Science and Engineering, 2015, (27): 1710-1722

- 
- [18] EGIG. 10th EGIG report March 2018[R]. Europe:European Gas Pipeline Incident Data Group, 2018.
  - [19] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社, 2013.
  - [20] 戴联双等. 城市燃气管网失效事件统计分析[J]. 全国失效分析学术会议, 2013, (49): 547-550
  - [21] 张满可等. 2011—2014 年我国城市燃气事故统计分析[J]. 煤气与热力, 2016, 36(1): 41-46
  - [22] 刘爱华等. 城市燃气管道状况及燃气事故统计分析[J]. 煤气与热力, 2017, 36(10): 27-33
  - [23] Pablo Hoffman et al. Scrapy, a fast high-level web crawling & scraping framework for Python.[EB/OL]. <https://github.com/scrapy/scrapy>, 2018-03-22
  - [24] 姜扬. 聚类 and 主成分回归在经济指标数据中的应用研究[D]. 长春: 吉林大学, 2010.
  - [25] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016. 197-224
  - [26] Wolpert, D.H. and W.G. Macready. “No free lunch theorems for search.”[J] Technical Report SFI-TR-05-010, Santa Fe Institute, Santa Fe, NM. 1995
  - [27] 同济大学概率统计教研组. 概率统计[M]. 上海:同济大学出版社, 2009.
  - [28] 中华人民共和国原化学工业部, GB50316-2008, 工业金属管道设计规范[S]. 北京. 中国计划出版社
  - [29] 中华人民共和国住房和城乡建设部, GB50028-2006, 城镇燃气设计规范[S]. 北京. 中国建筑工业出版社
  - [30] 翁永基等. 腐蚀管道最小壁厚测量和安全评价方法[J]. 油气储运, 2003, 12(22): 40-43
  - [31] Aloise, D.,A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering[J]. Machine Learning, 2009, 75(2):245-248.
  - [32] Ulrike, Luxburg. A Tutorial on Spectral Clustering[J]. Statisitcs & Computing, 2007, 17(4): 395-416
  - [33] T. Calinski and J. Harabasz, A dendrite method for cluster analysis[J], Communications in Statistics -theory and Methods, 1974, 3(1), 1-27
  - [34] 项晓春, 刘广魁. SCADA 系统及应用[J]. 自动化技术与应用, 2000, 19(6): 19-22
  - [35] 白庆林等. 城市燃气 SCADA 及信息管理系统设计[J]. 自动化与仪表, 2009, 24(4): 47-50
  - [36] Petitjean, F.O et al. A global averaging method for dynamic time warping, with applications to clustering[J], Pattern Recognition, 2011, 44(3), 678
  - [37] Keogh E., Pazzani M.J. Derivative Dynamic Time Warping[C]. Proceedings of the 2001 SIAM International Conference on Data Mining, 2000,1-11
  - [38] 陈家鼎等. 数理统计学讲义[M]. 北京:北京大学出版社, 2014.

## 谢 辞

正文内容