同济大学
TONGJI UNIVERSITY

基于多源数据的管道安全性评估

学院：土木工程学院　　专业：建筑工程

答辩人：王子丰　　指导老师：李素贞

目 录

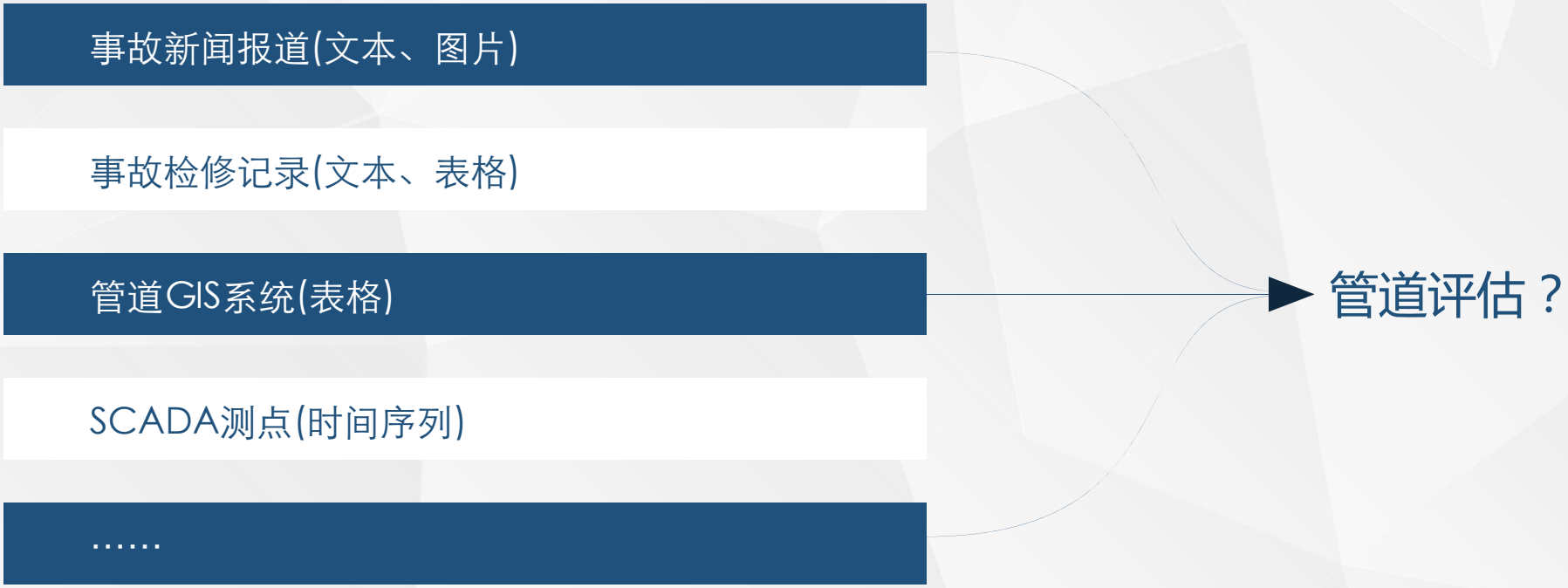
contents

01

PART ONE

背景简述

01/ 多源异构数据



01/ 本文主要工作

01

生命线管道事故数据系统

集成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

02

管道分段安全性评估

基于数据清洗、多种特征工程以及agglomerative clustering聚类算法的分段安全性评估模型

03

SCADA测点数据的实时异常检测模型

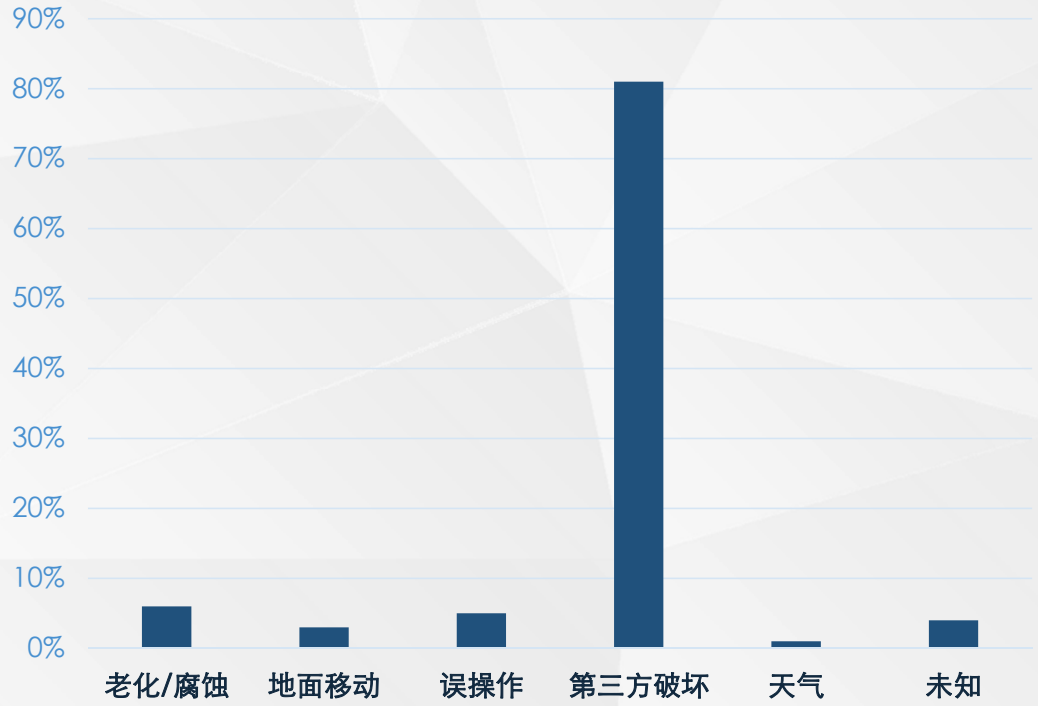
两层时序聚类，基于危险日内走势模版匹配的实时数据异常检测模型

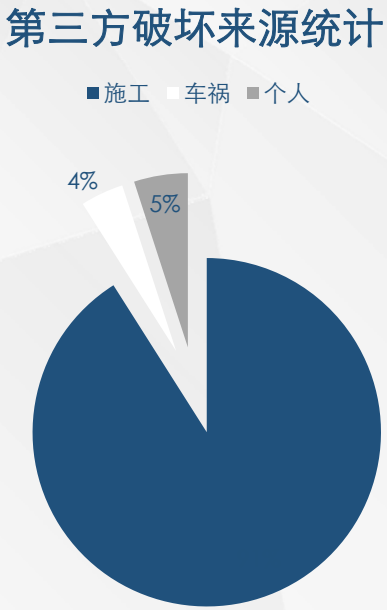
02

PART TWO

第一部分

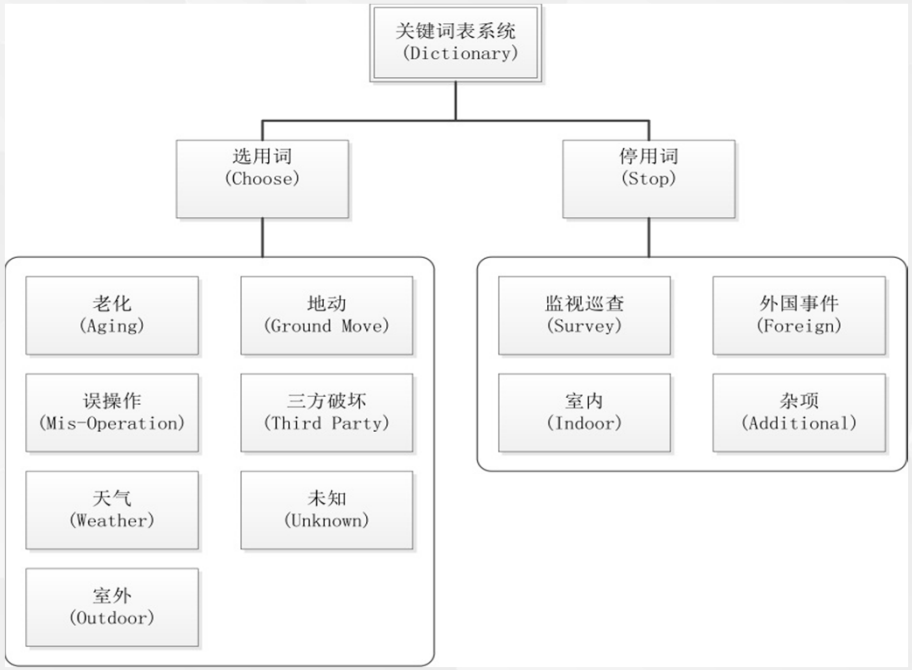
02 事故报道收集与统计(2010~2018)

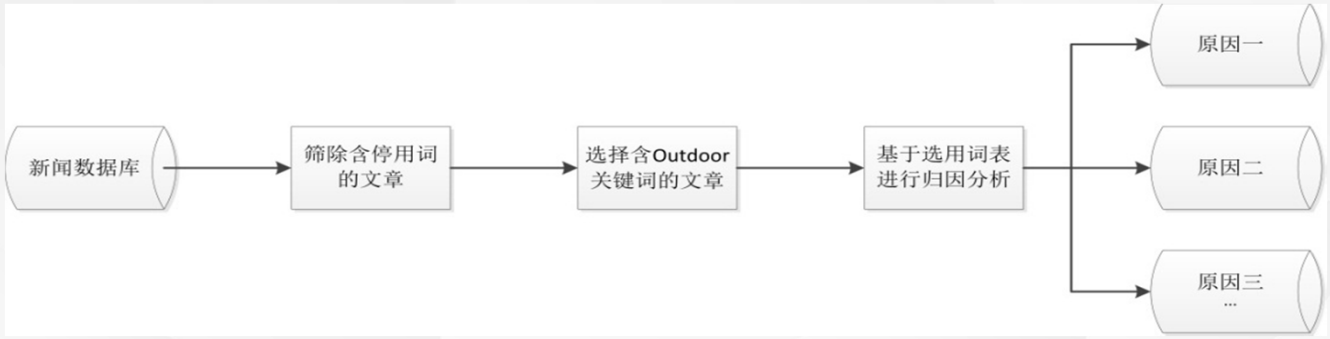




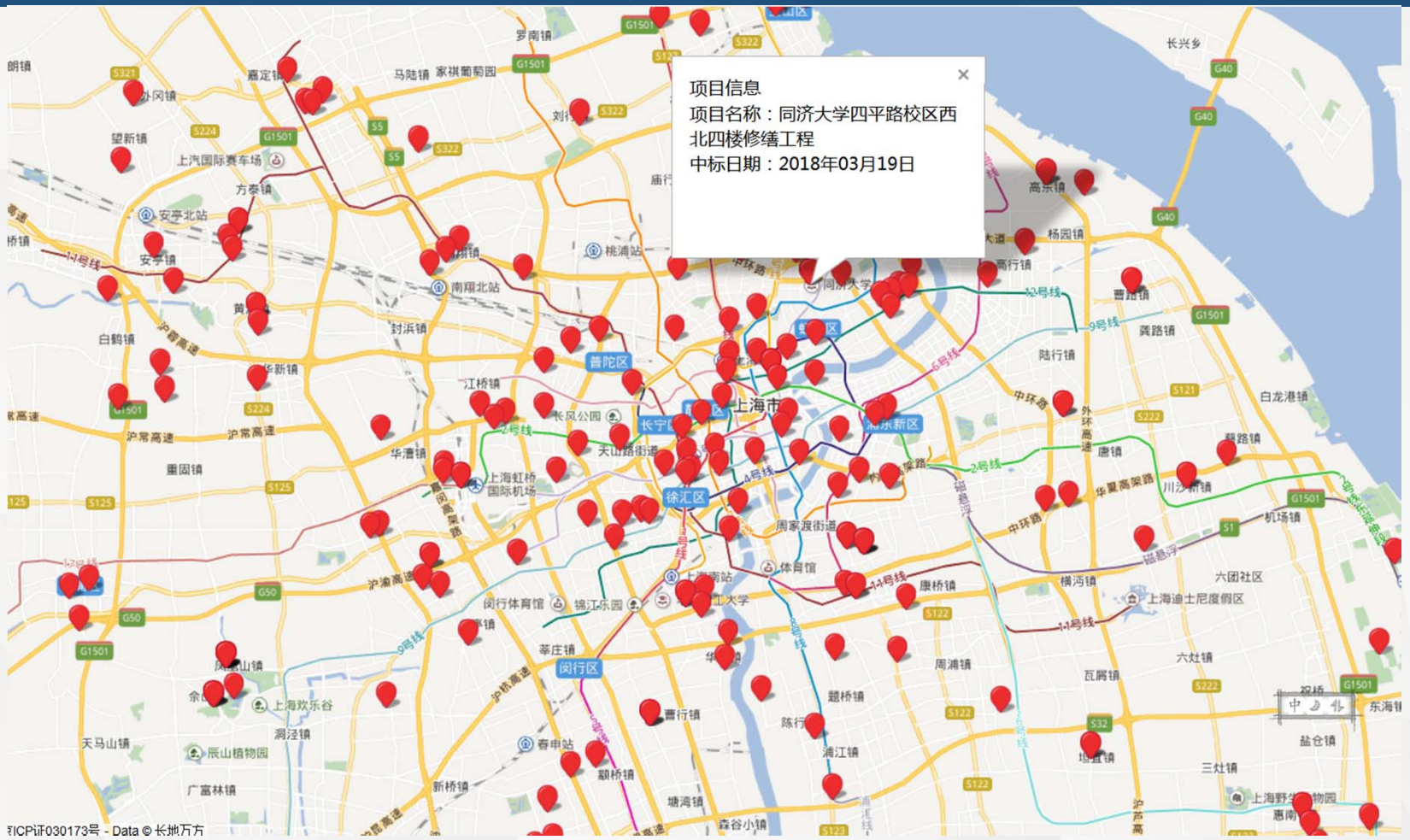
原因	老化/腐蚀	地面移动	误操作	第三方破坏	天气	未知	总计
数目	23	10	18	312	4	17	384
比例	6%	3%	5%	81%	1%	4%	100%

02 归因词表系统与词表系统



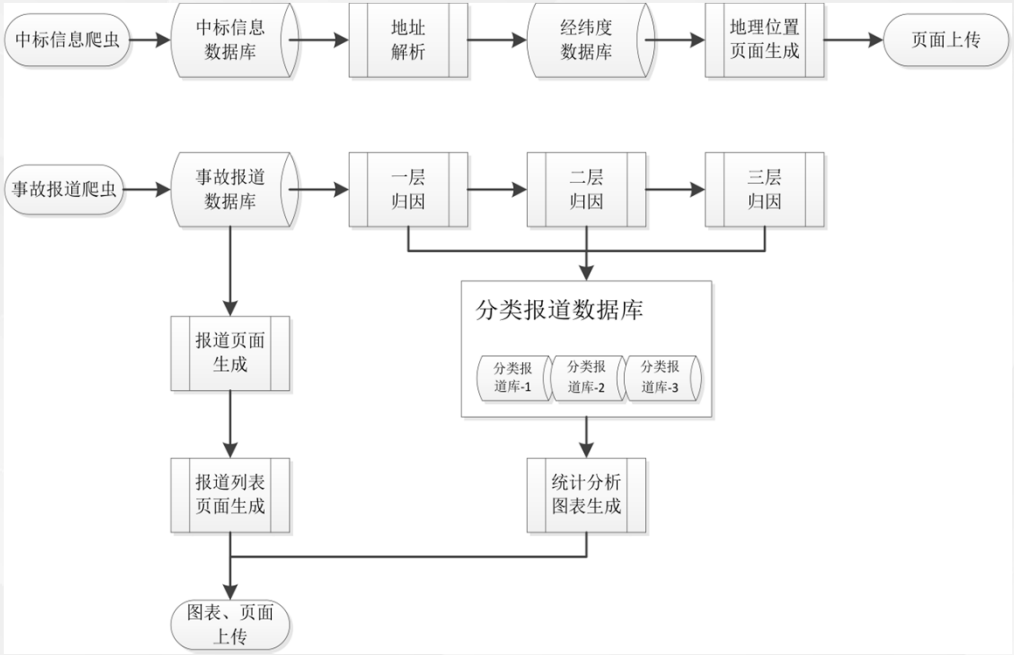


02/ 中标项目地理分布



tenders_shanghai.html

02 系统工作流程



JSON 文件 (1)
- 经纬度.json

Microsoft Excel 逗号分隔值文件 (2)
- news.csv
- tenders.csv

Python File (11)
- analysis.py
- analysis_3rdparty.py
- charts.py
- construction_sources.py
- genBaiduMap.py
- genHTML.py
- map.py
- post_article.py
- post_articlelist.py
- post_media.py
- pyh.py

QQBrowser HTML Document (1)
- Map.html

文本文档 (2)
- Map_module.txt
- Required Packages.txt

文件夹 (7)
- Classified
- Dictionary
- NewsHTML
- Pictures
- 博燃网
- 中标信息
- 中国燃气网

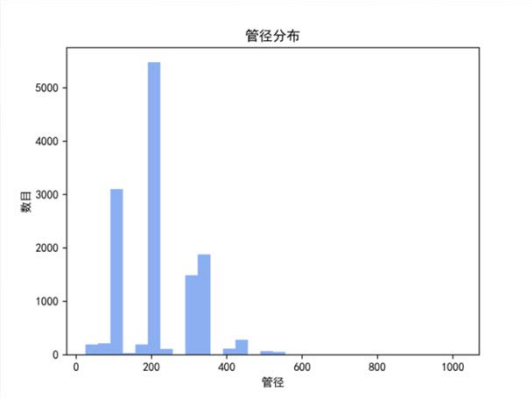
03

PART THREE

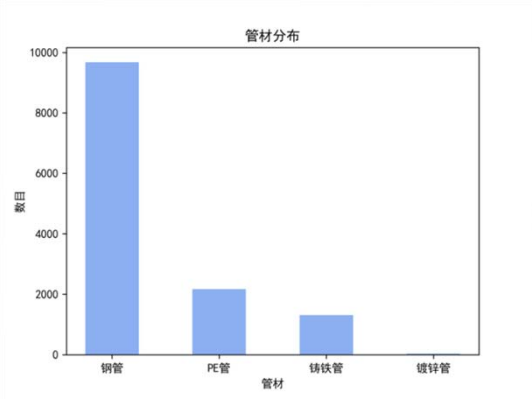
第二部分

03 数据清洗&特征工程

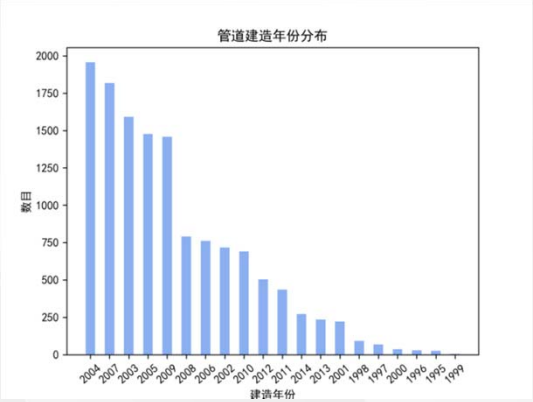
管径



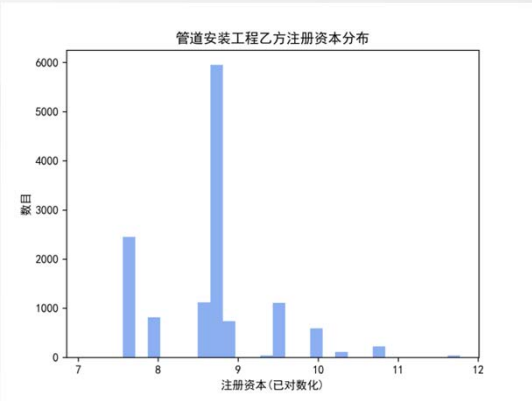
管材



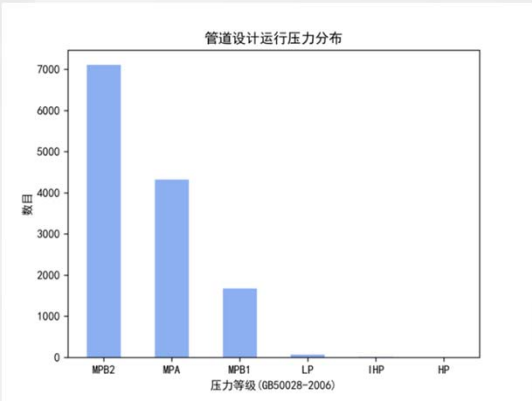
建造年份



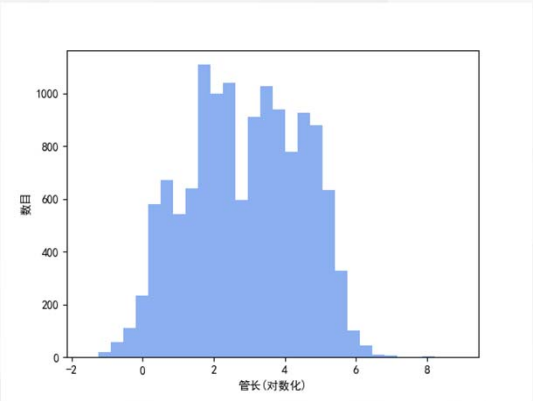
工程乙方注册资本



设计运行压力



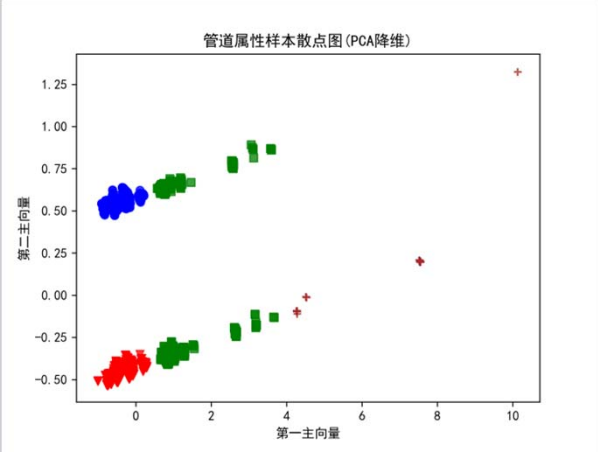
管长(对数化)

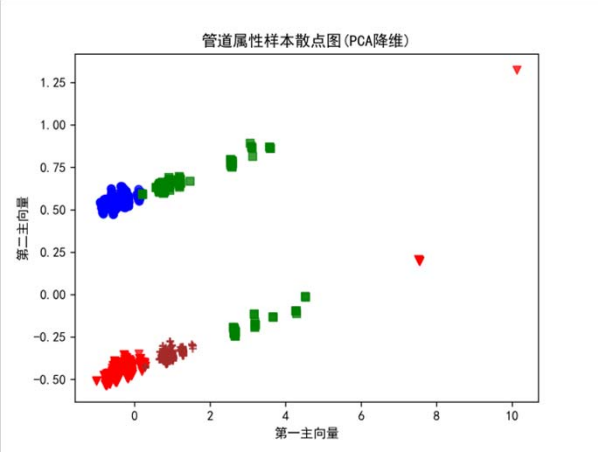


03 特征汇总

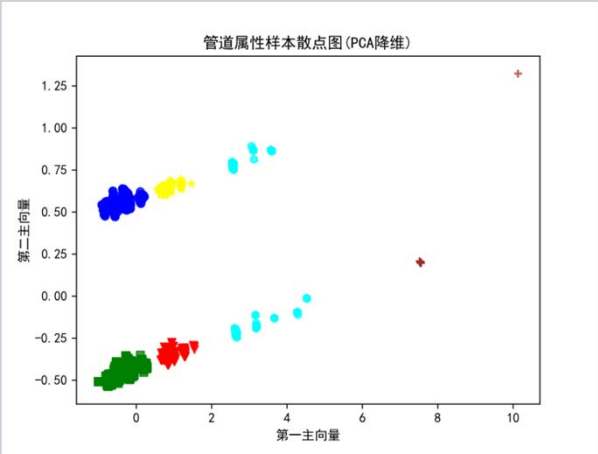
特征名称	类型	值域	单位	说明
WARN_TAPE	类别型	0,1	\	管道埋设处是否有警示标志,1为有,0为无
JJSIZE	数值型	[25,1020]	mm	管道管径大小
WEIGHTED_MATL	数值型	[30,60]	年	管道材料的设计使用寿命
WEIGHTED_DIAMETER	数值型	[0.092,0.540]	\	根据EGIG统计数据引入的管径的事故先验概率
DATE_BUILD_AGE	数值型	[1123,8341]	日	管道从安装完成起至2018年1月1日的天数
DATE_BUILD_SEASON	类别型	0,1,2,3	\	管道安装的季节，0为冬季，1为春季,2为秋季,3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

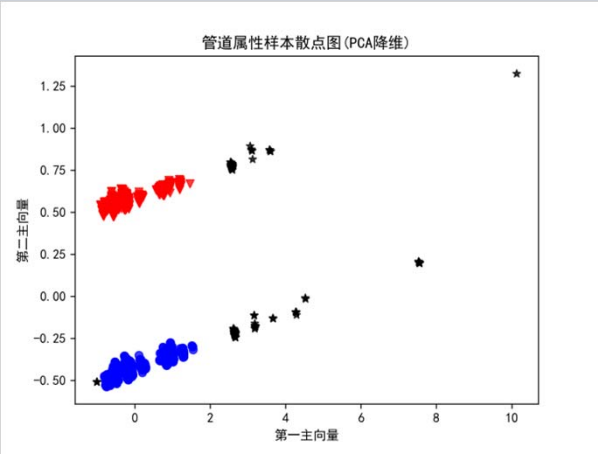
03 算法性能比较(一)

算法名称 & CH 指数	聚类结果散点图
K-means 参数： n_clusters = 4 CH index： 8171.02141577	

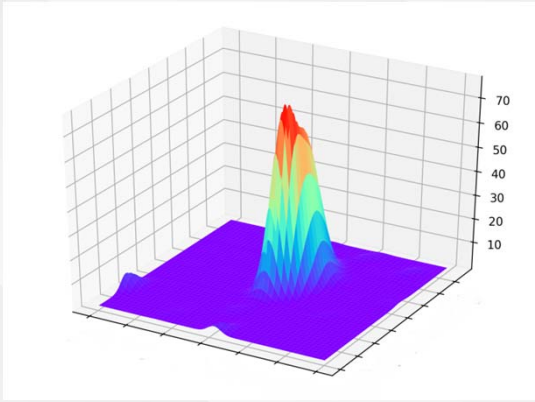
算法名称 & CH 指数	聚类结果散点图
Spectral Clustering 参数： n_clusters = 4 gamma = 0.5 affinity = “rbf” CH index： 5427.30246242	

03 算法性能比较(二)

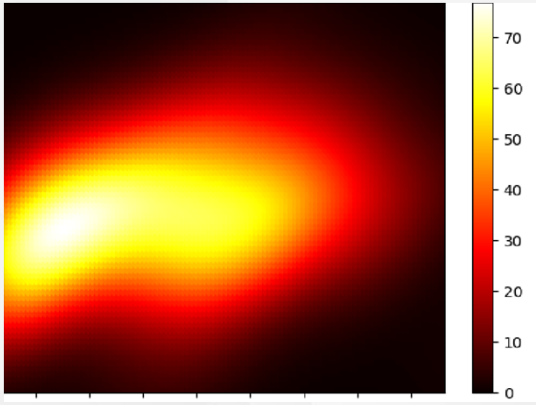
算法名称 & CH 指数	聚类结果散点图
Agglomerative Clustering 参数： n_clusters = 6 affinity = “euclidean” linkage = “ward” CH index: 8147.62242851	

算法名称 & CH 指数	聚类结果散点图
DBSCAN 参数： eps = 0.9 min_samples = 100 metric = “euclidean” CH index: 5165.90293589	

03 事故记录地理分布



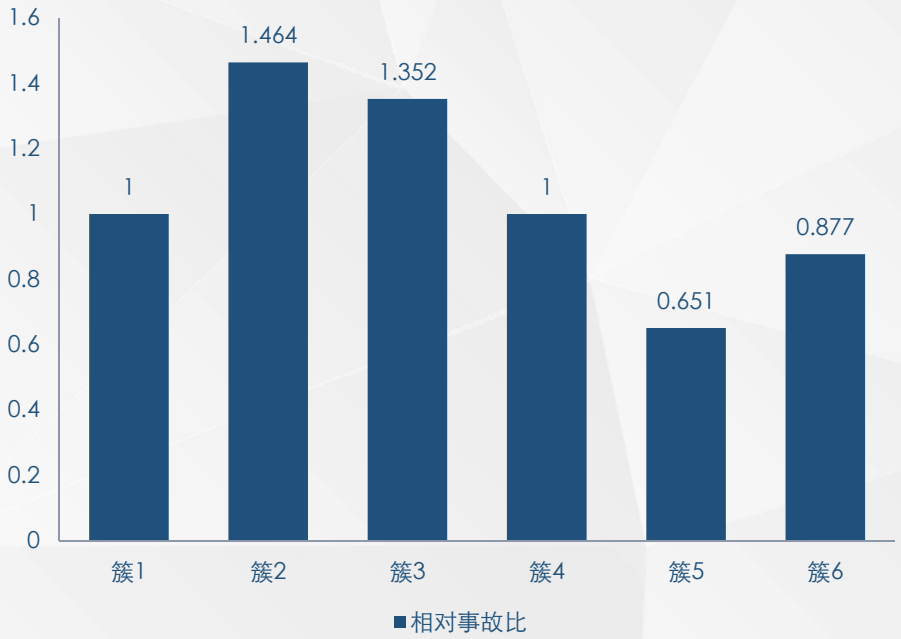
事故地理分布概率曲面



事故地理分布热图

failure_reports.html

03 聚类簇的风险定义



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)

卡方检验表

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

结论:
从危险到安全的排序为：2，3，1，4，6，5

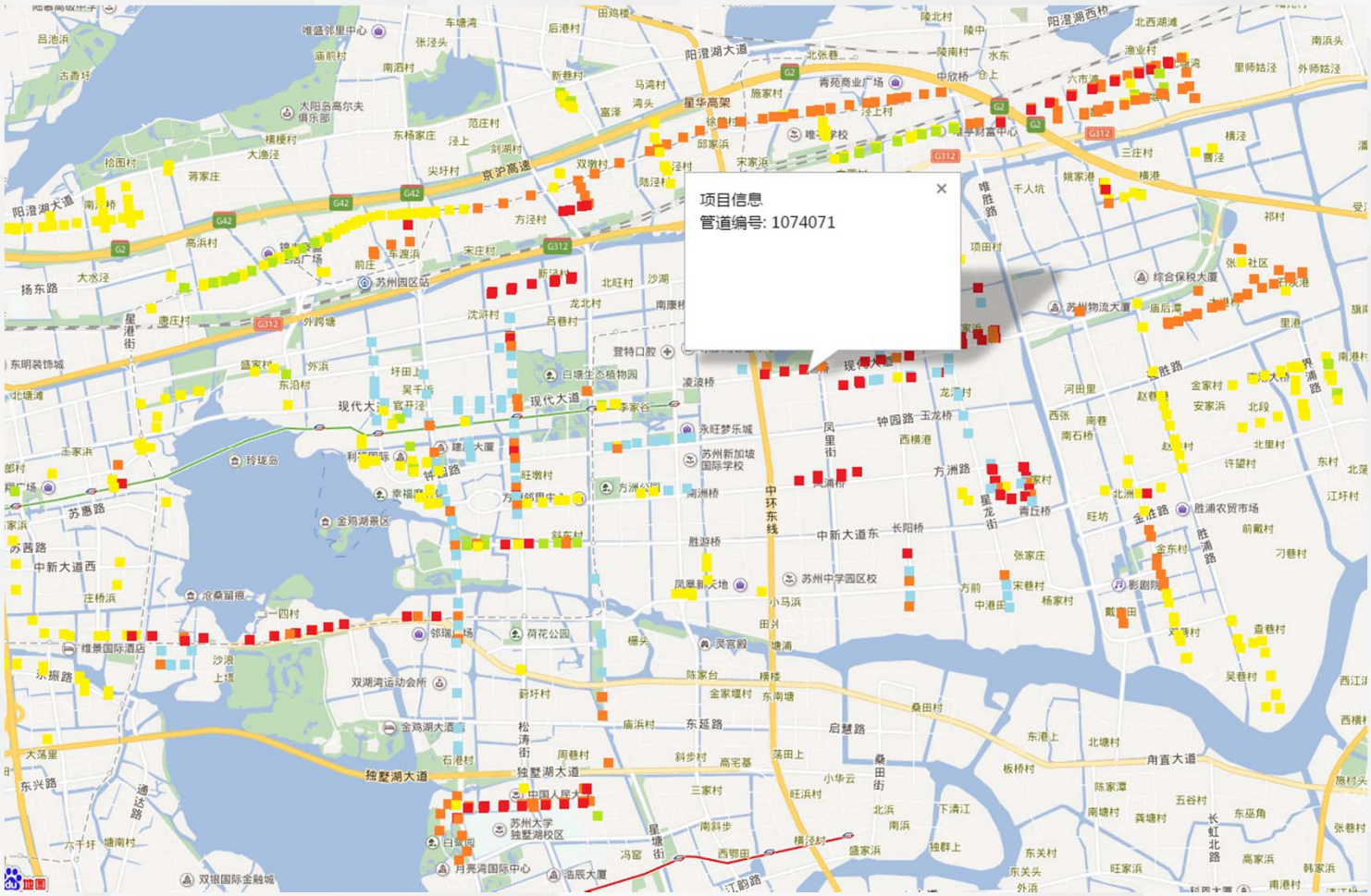
期望事故频数： $E(N_{F,i}) = \frac{n_i}{\sum n_k} \times N_F$

$N_{F,i}$ 第 i 类簇内的事故数

n_i 第 i 类簇内管道总数

N_F 事故记录总数

03 管道风险度的图形化展示



Map_Clustering.html

04

PART FORE

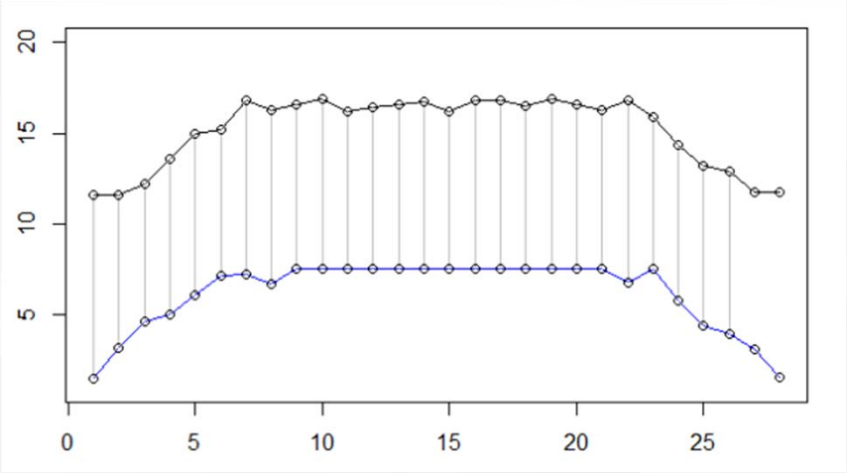
第三部分

04 SCADA实时测点数据异常检测模型

1. 利用SCADA测点历史数据进行时间序列聚类得到典型模版线
2. 用事故地测点序列匹配模版线，找到高风险模版线
3. 匹配实时序列与所有模版线，最匹配高风险模版线则说明出现异常

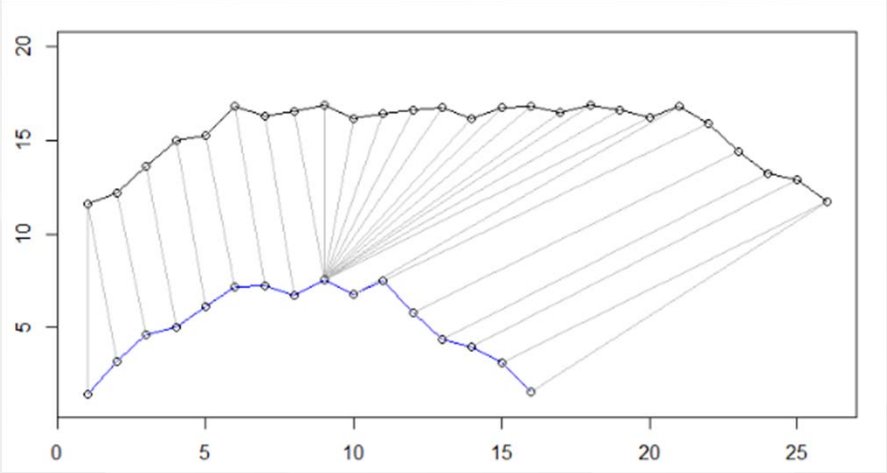
04 序列的相似度-DTW距离

欧式距离-
Euclidean Distance



V.S

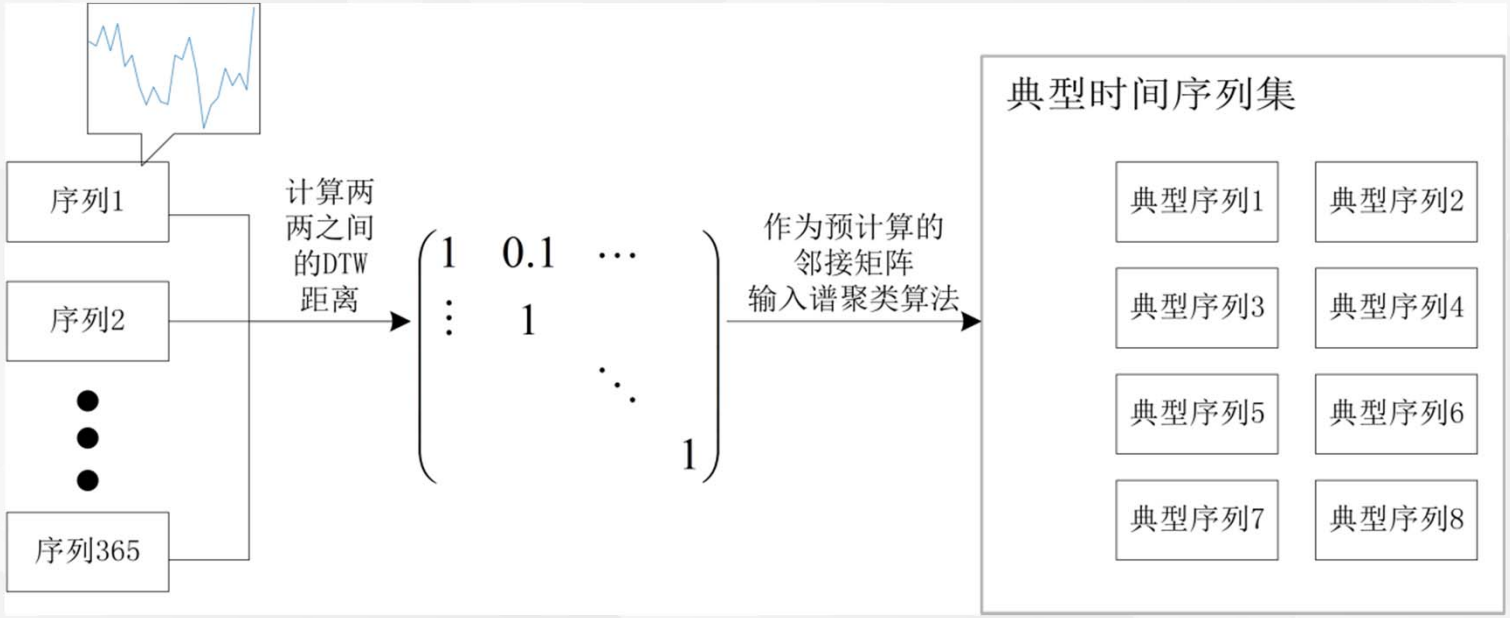
动态时间序列规整距离-
Dynamic Time-series Warping Distance



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

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

04 两层时间序列聚类



单层聚类:

$47 \times 365 = 17155$

$C_{17155}^{2} = 147138435$

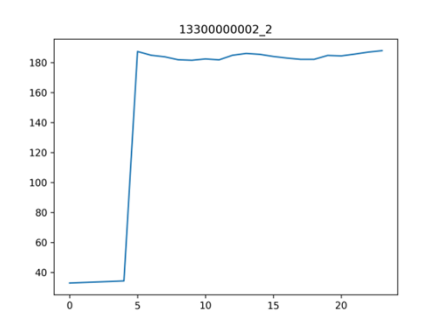
V.S

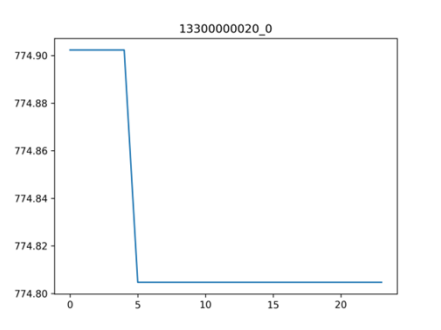
双层聚类:

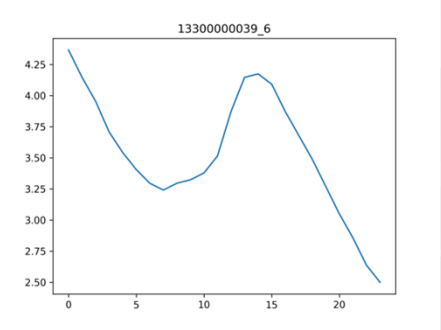
$47 \times 8 = 376$

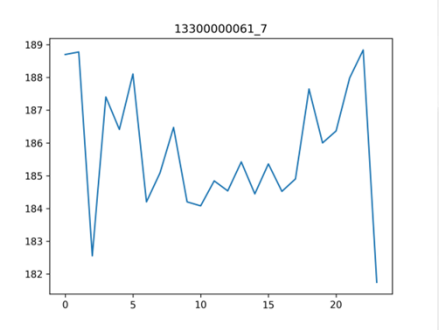
$47 \times C_{365}^{2} + C_{376}^{2} = 3192710$ $\approx 2\%$

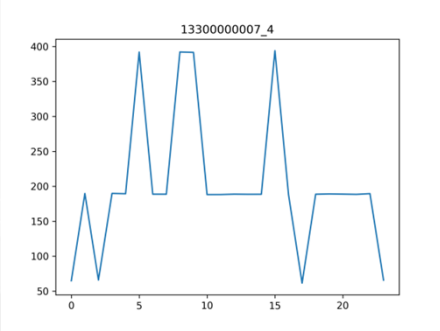
04 典型模版线

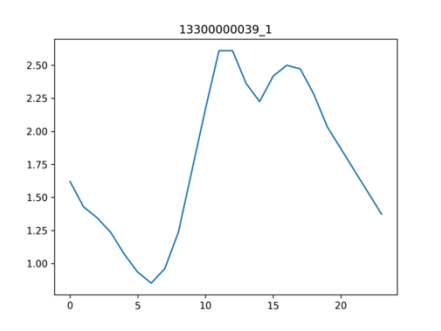

13300000002_2

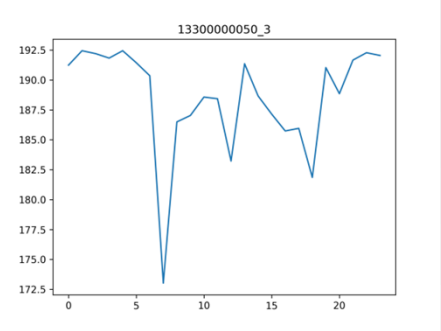

13300000020_0

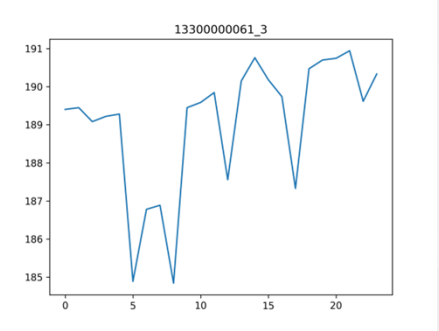

13300000039_6

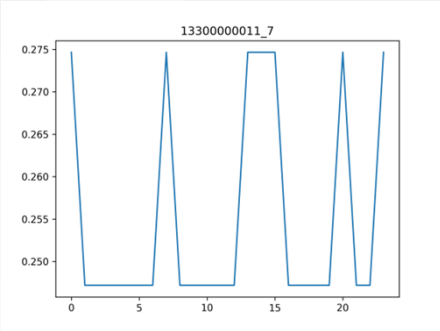

13300000061_7

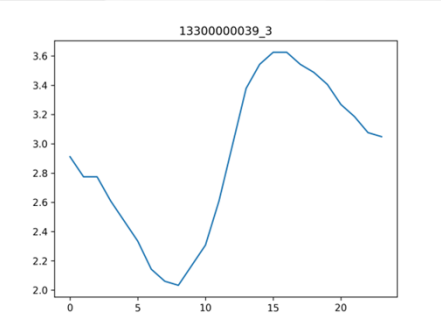

13300000007_4


13300000039_1

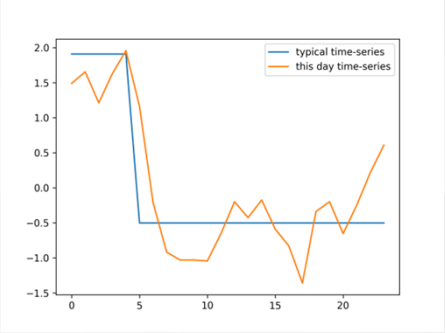

13300000050_3

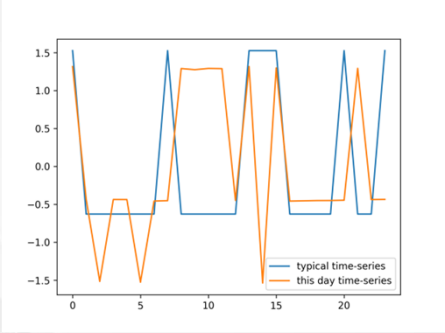

13300000061_3

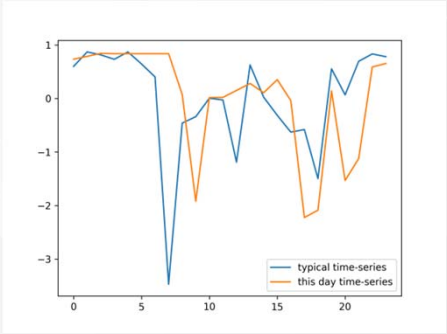

13300000011_7

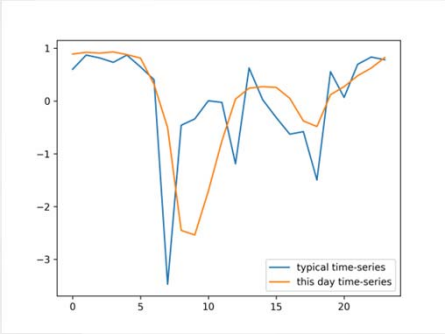

13300000039_3

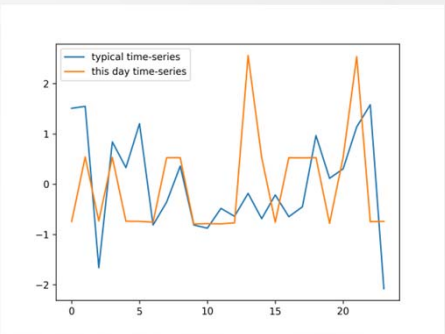
04 事故地序列匹配(2015)

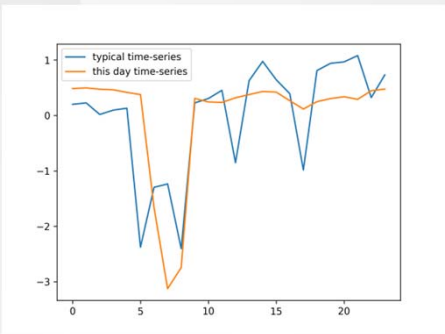


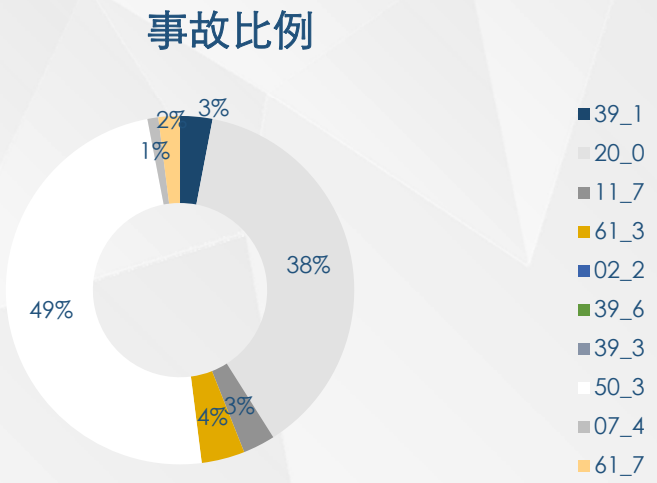








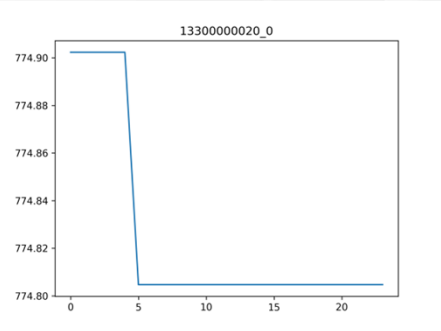




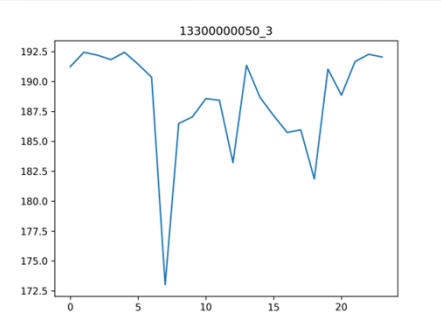
模版线编号	39_1	20_0
事故数量	5	68
事故比例	3%	38%
模版线编号	11_7	61_3
事故数量	5	7
事故比例	3%	4%
模版线编号	02_2	39_6
事故数量	0	0
事故比例	0%	0%
模版线编号	39_3	50_3
事故数量	0	87
事故比例	0%	49%
模版线编号	07_4	61_7
事故数量	2	3
事故比例	1%	2%

04 异常检测模型验证(2016)

事故地测点序列最匹配的两条模版(高风险)

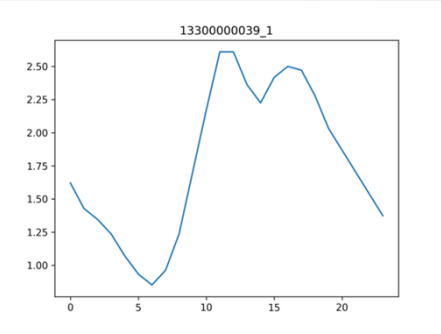


编号: 20_0

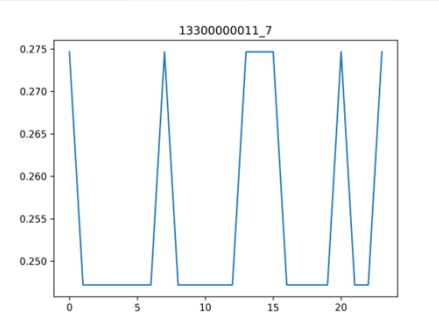


编号: 50_3

随机抽样测点序列最匹配的两条模版(正常)

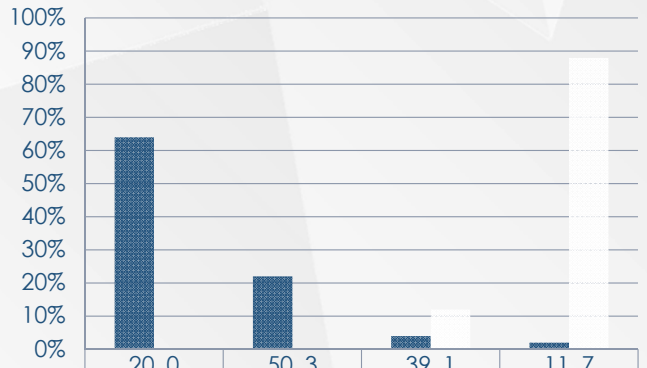


编号: 39_1



编号: 11_7

对比统计结果



	20_0	50_3	39_1	11_7
事故线样本集	64%	22%	4%	2%
随机抽样线样本集	0%	0%	12%	88%

05

PART FIVE

主要成果与展望

05 成果及创新点

01 生命线管道事故数据系统

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。完善的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。

02 管段聚类模型及安全性评估方法

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。

03 模式匹配的SCADA测点数据的实时异常检测模型

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。

06 研究方向展望

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。

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。

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。

THANKS

xxxxxx

王子丰

zifengwang2016@foxmail.com