同济大学
TONGJI UNIVERSITY

基于多源数据的地下管道安全性评价

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？

事故新闻报道(文本、图片)

事故检修记录(文本、图片、表格)

管道地理信息系统(管径、管材、长度……)

管道监测点(压力、流量、)

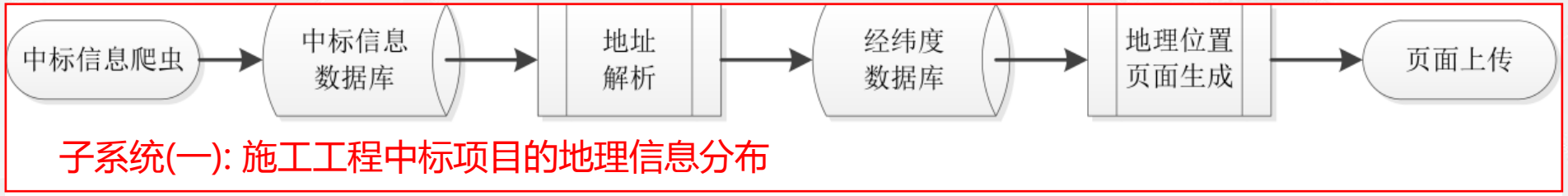
……

02

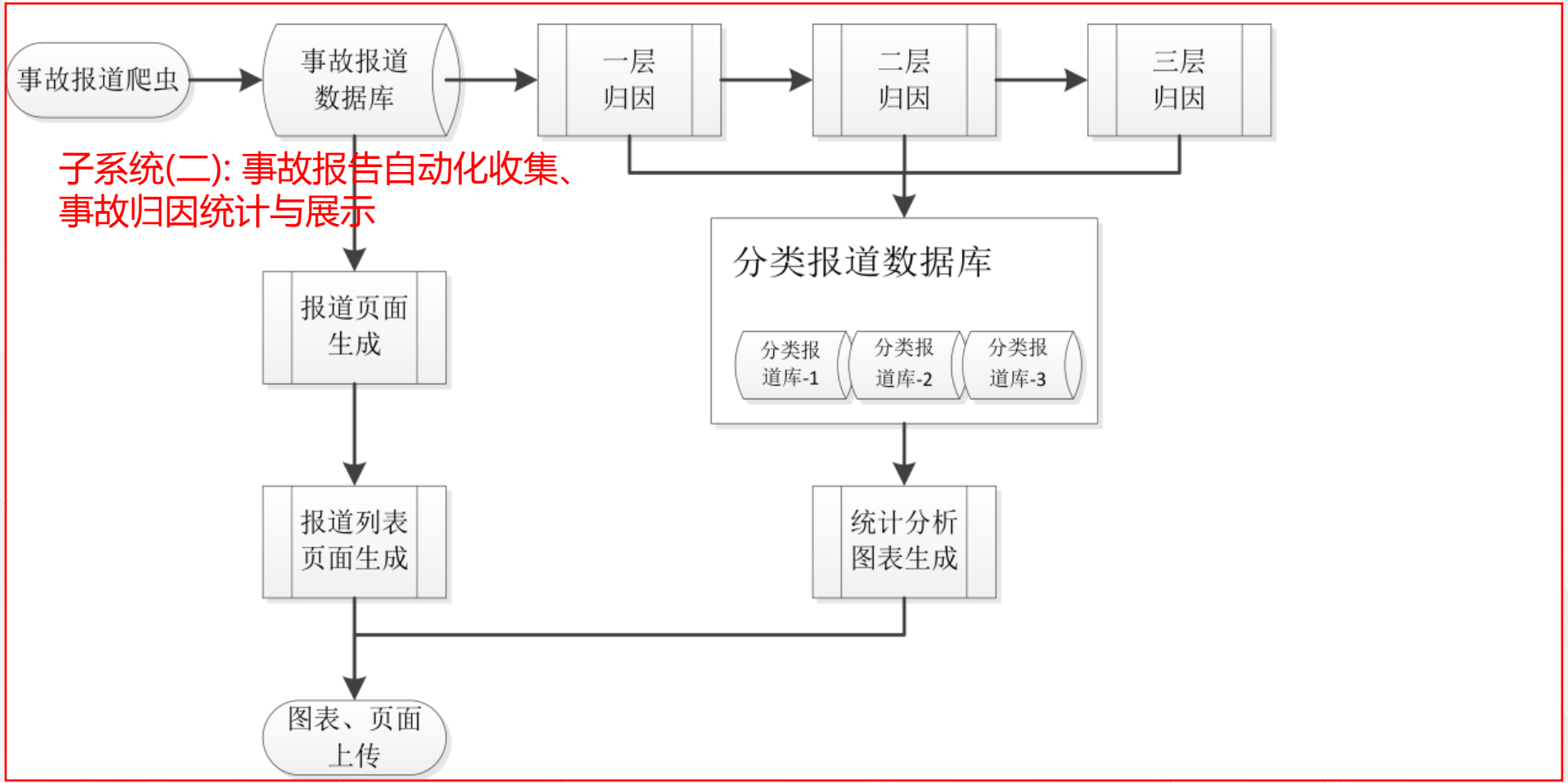
PART TWO

管道事故数据系统

02 系统的组成部分



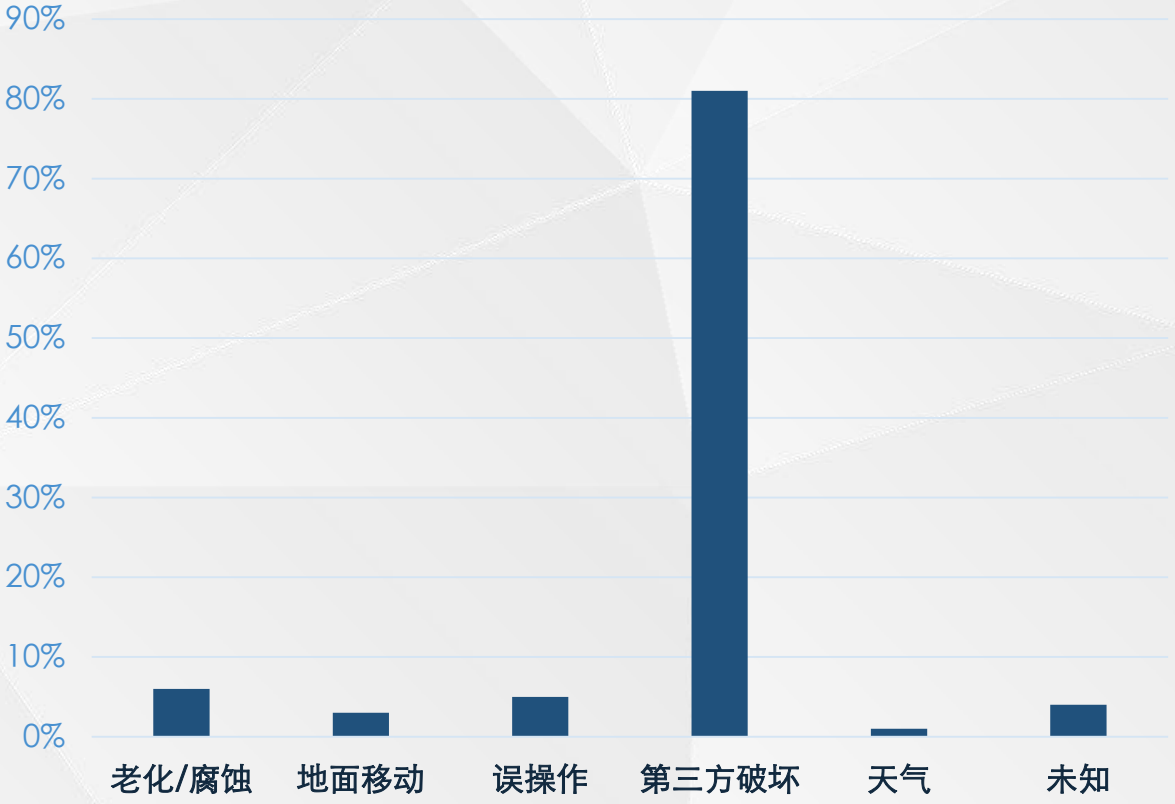
子系统(一): 施工工程中标项目的地理信息分布

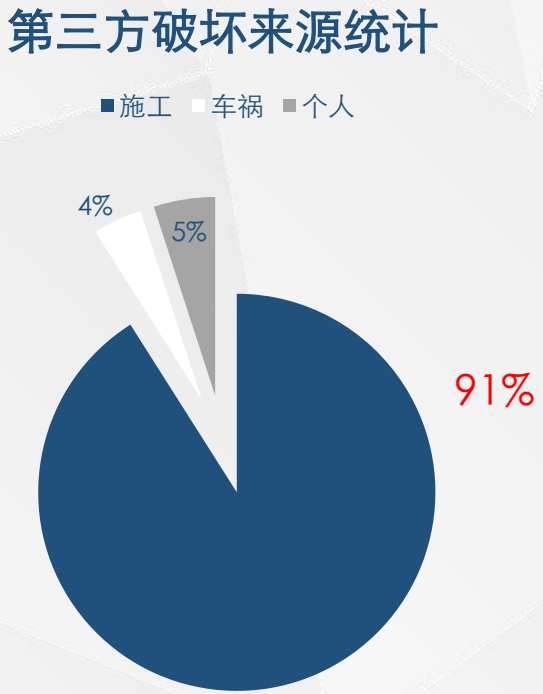


子系统(二): 事故报告自动化收集、
事故归因统计与展示

02 事故报道收集与统计(2010~2018)

从8000余条新闻中自动化筛选出384条新闻
第三方破坏占总数的81%
施工占第三方破坏的91%





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 中标项目地理分布

帮助燃气公司与施工单位提前沟通，加强监管以降低施工造成的事故发生率。

03

PART THREE

管道分段安全性评估

02 安全性评估的建模流程

01

原始数据(管径、管材、长度……)的预处理，数据来自苏州工业园区的GIS系统

02

特征构建(对原始数据的二次处理，使其更适合作为算法的输入)

03

多种算法性能的比较(比较了k-均值、谱聚类、密度聚类和层次聚类算法的性能)

04

定义管道的风险等级(统计不同管道类别的事故发生频数)

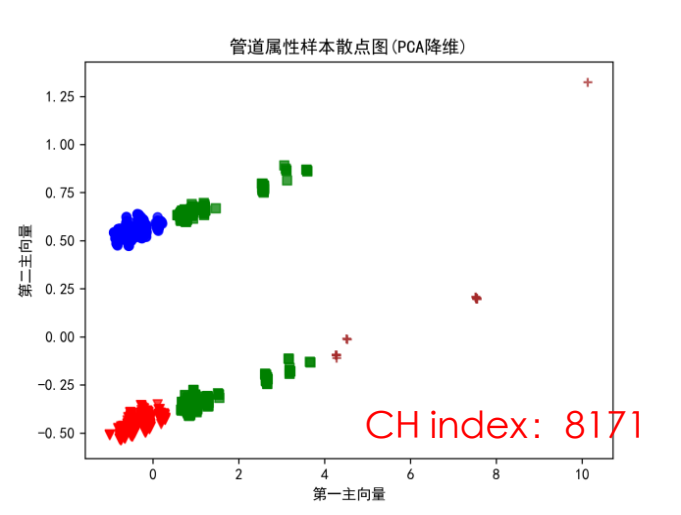
03 特征汇总

特征名称	变量名称	类型	值域	单位	说明
警示标志	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
管径	JJSIZE	数值型	[25,1020]	mm	管道管径大小
管材(设计寿命加权)	WEIGHTED_MATL	数值型	[30,60]	年	管道材料的设计使用寿命
管径(事故概率加权)	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
管龄	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
建造季节	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
实际内压	PRESS_O_WEIGHT	数值型	[0.01,2.5]	mPa	管道实际运行内压
设计内压	PRESS_D_WEIGHT	数值型	[0.01,2.5]	mPa	管道设计使用内压
工程乙方规模	CONTRACTOR_WEIGHT	数值型	[1200,130000]	万元	管道安装施工乙方注册资本
工程甲方规模	SUPERVISOR_WEIGHT	数值型	[300,2773.9]	万元	管道安装施工监理方注册资本
壁厚(假设)	HTHICKNESS	数值型	[0,1]	\	管道的假设壁厚,仿照规范公式计算得到
长度	LENGTH	数值型	[0.2,7509]	米	管道长度

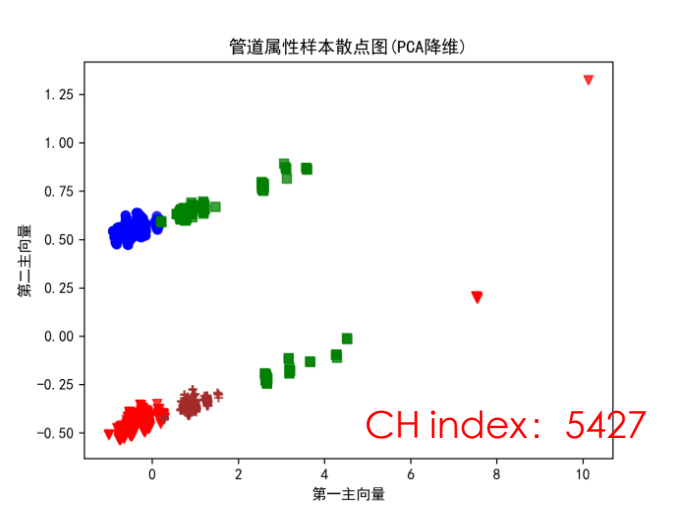
经过第一步预处理与第二步特征构建，得到这些用于输入聚类算法的数据集

03 聚类算法性能比较

K-means



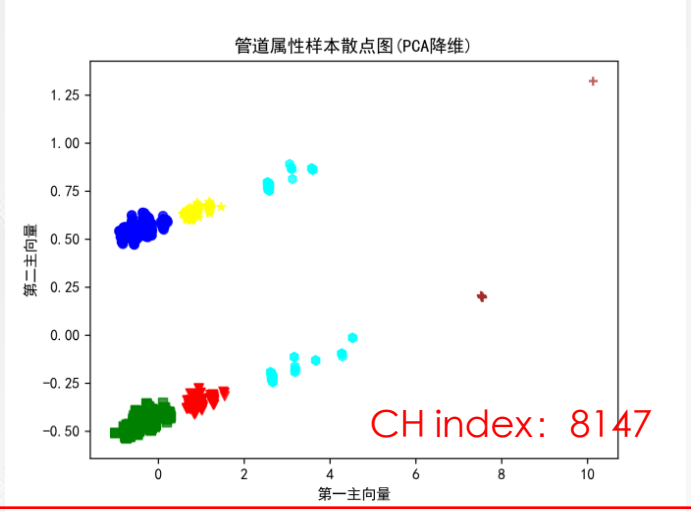
谱聚类



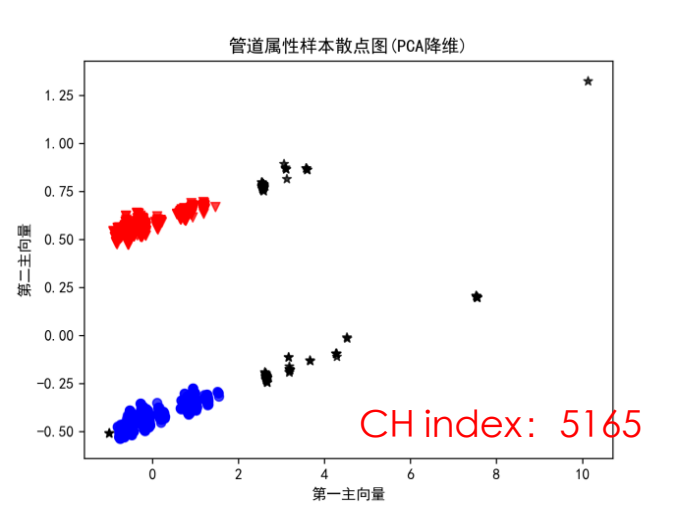
不同颜色代表不同的管道类别

CH index是度量聚类好坏的标准，一般越大越好

层次聚类 最佳



密度聚类



层次聚类将管道分为了6类，此时尚不清楚各类管道的风险等级顺序，需要引入事故记录来定义。

02 事故记录(2014~17)的分布频数

管道分类	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

期望事故频数：

假设事故在所有管道上均匀分布，
则每类管道上所发生的事故频数应与该类管道的总数成正比。

检验结果显示至少99%的把握认为事故在不同类的管道上发生频率不同，即管道分类确实体现了其风险度的差异。

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 类别的风险等级定义

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)

定义相对事故比：

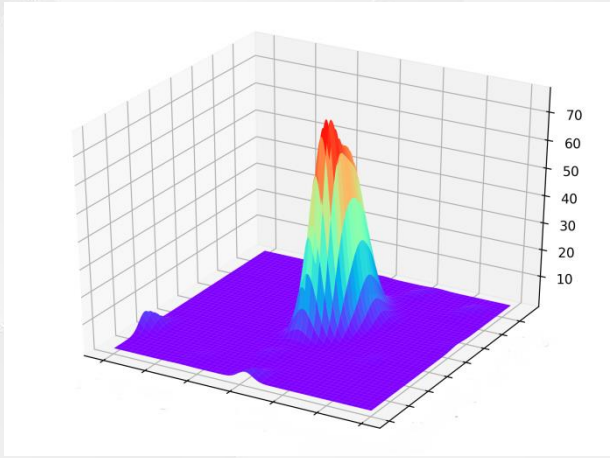
在某类管道上发生的事故频数÷该类管道的数量

相对事故比越高，则说明该类管道风险越高。

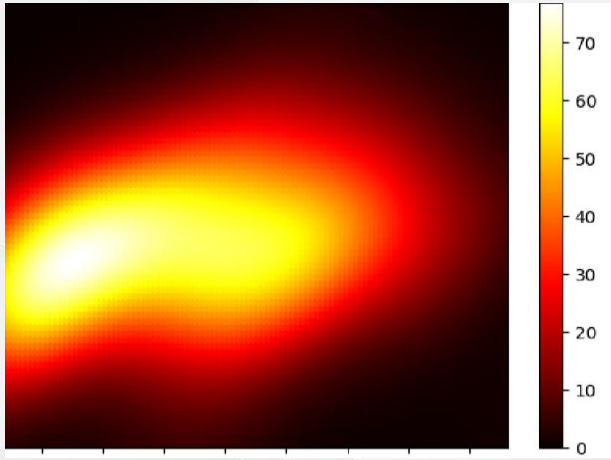
结论：

从危险到安全的排序为：2，3，1，4，6，5

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

03 管道风险度的图形化展示

红色：高风险
橙色：较高风险
黄色：中风险
蓝色：较低风险
绿色：低风险

管道共有13,138条

1,517条被评为高风险

1,953条被评为低风险

04

PART FORE

压力的异常检测模型

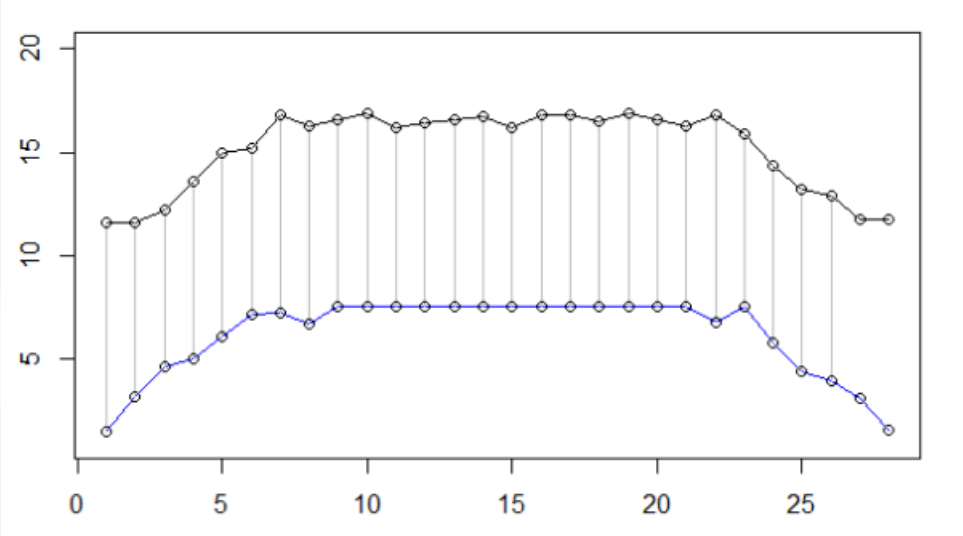
04 建模流程

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. 实证检验，确定高风险模版线确实表示不正常的管道运行状态

处理流程：

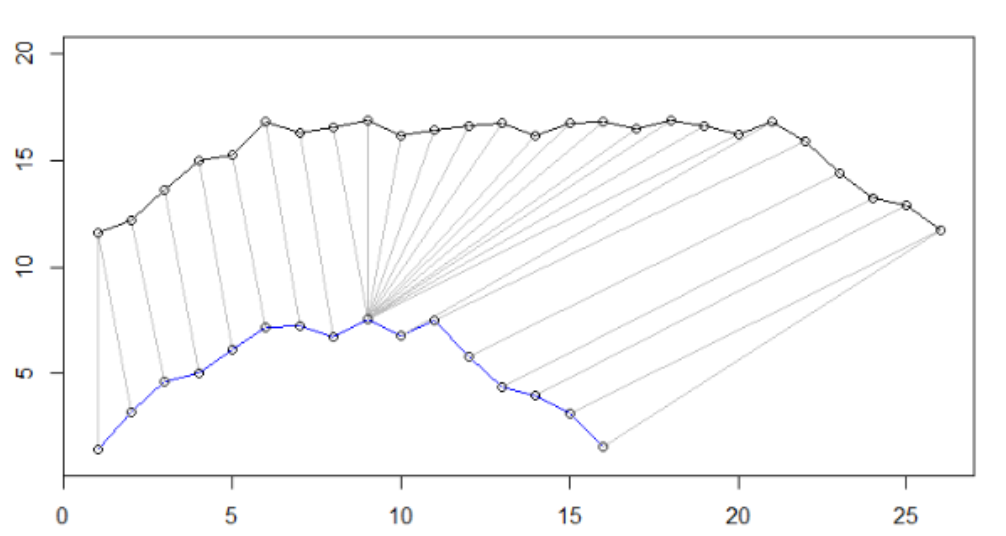
04 序列的相似度-DTW距离

欧式距离-
Euclidean Distance



V.S

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

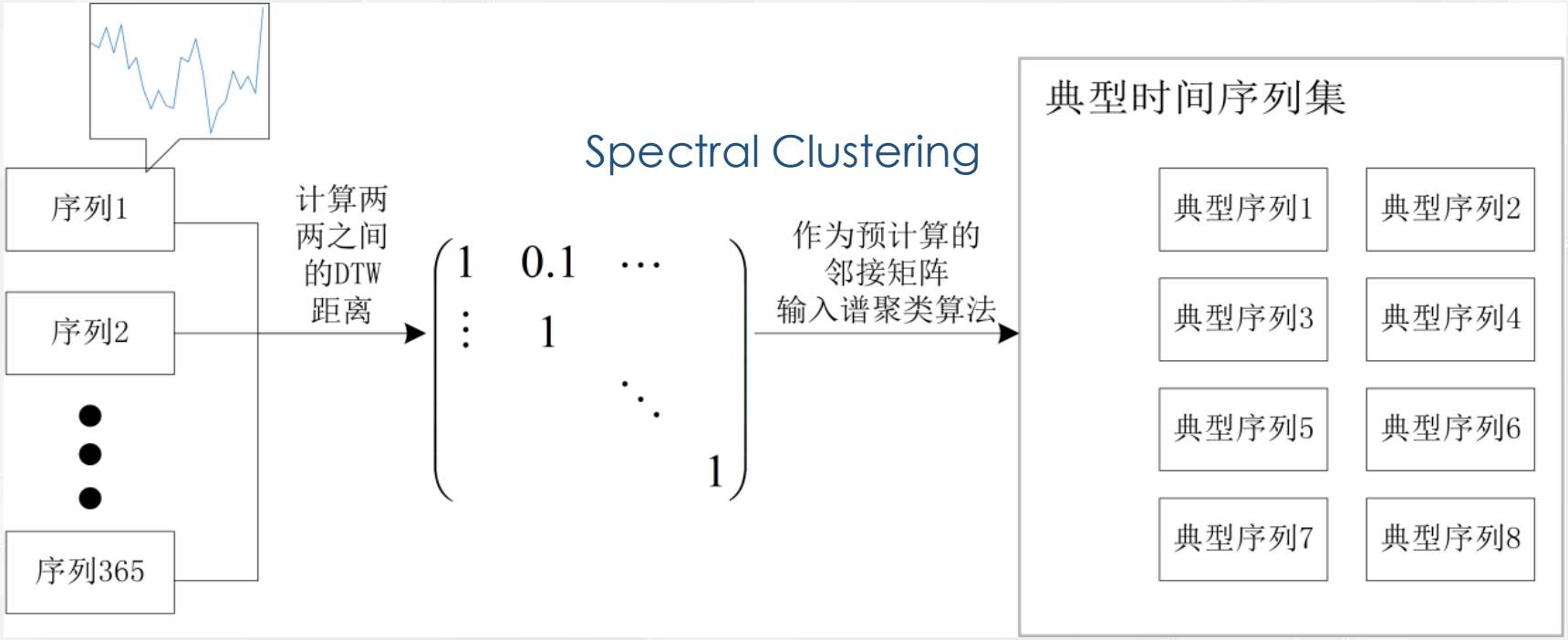


$$dist(A,B)=\sum_{i=1}^{m}(a_i-b_i)^2$$

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

DTW能用于不等长的序列之间的比较，更优

04 两层的聚类？



单层聚类:

$$47 \times 365 = 17155$$

$$C_{17155}^{2} = 147,138,435$$

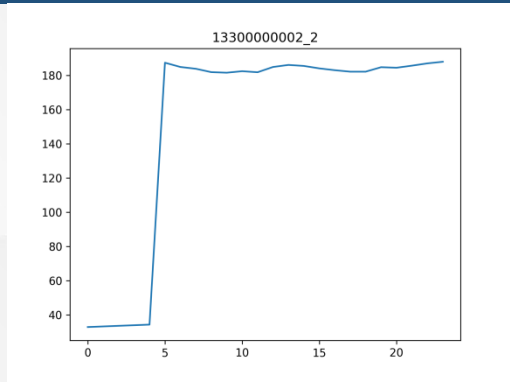
V.S

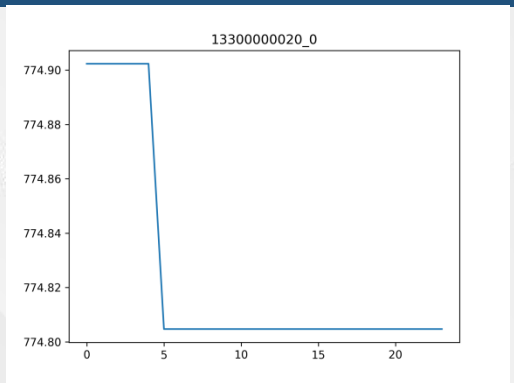
双层聚类: 两层聚类的计算量仅为单层的2%
大大节省了计算时间

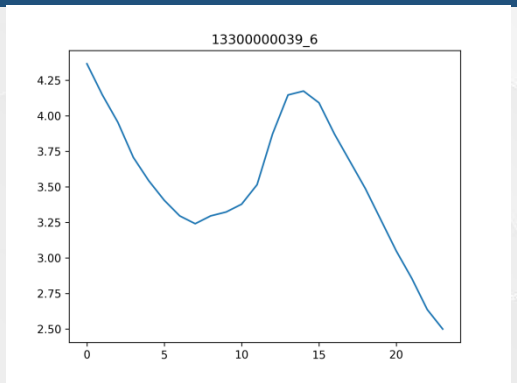
$$47 \times 8 = 376$$

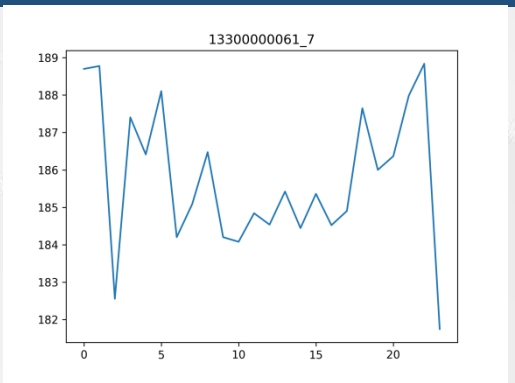
$$47 \times C_{365}^{2} + C_{376}^{2} = 3,192,710$$

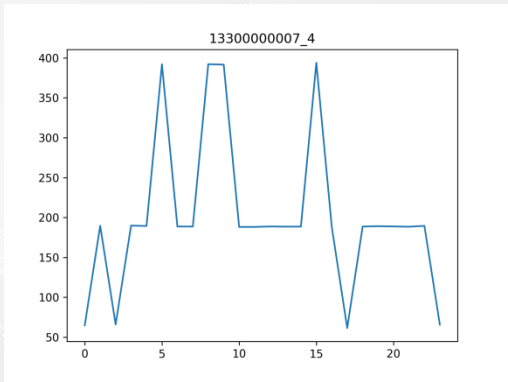
04 典型模版线

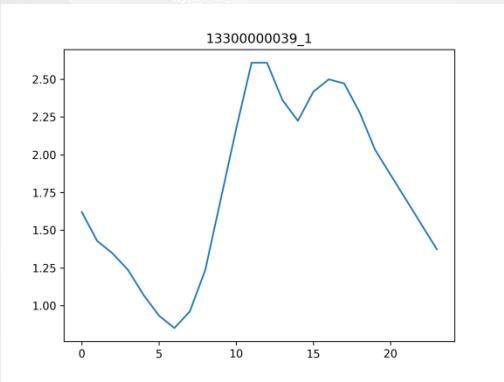

13300000002_2

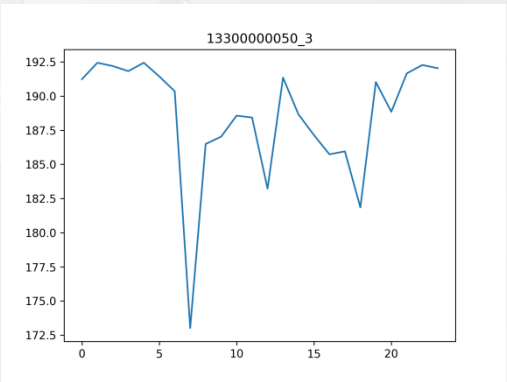

13300000020_0

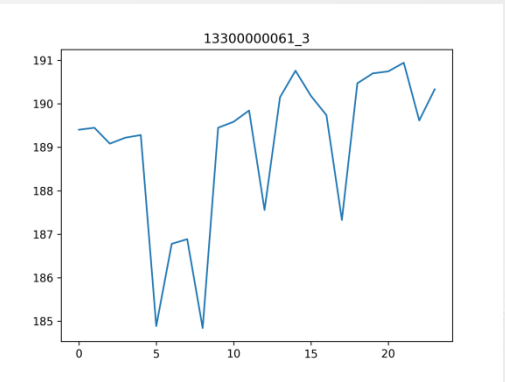

13300000039_6

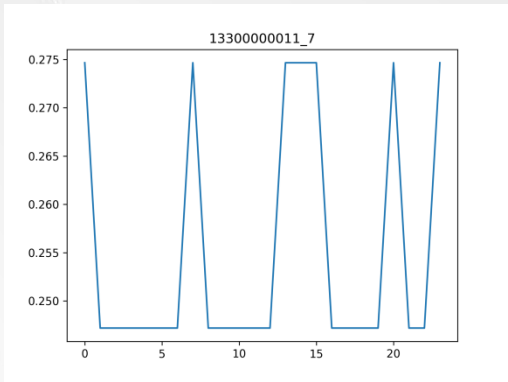

13300000061_7

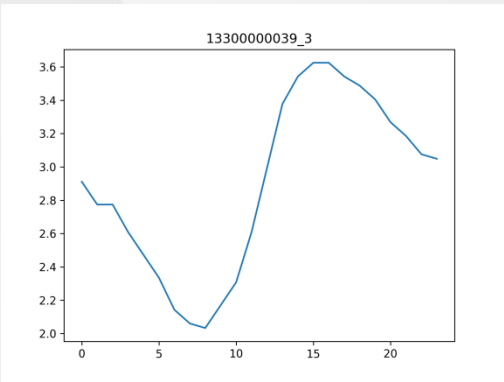

13300000007_4


13300000039_1


13300000050_3


13300000061_3

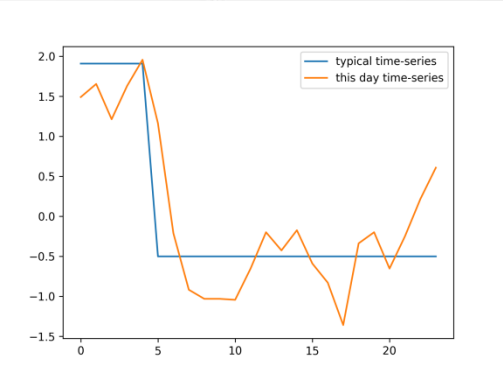

13300000011_7

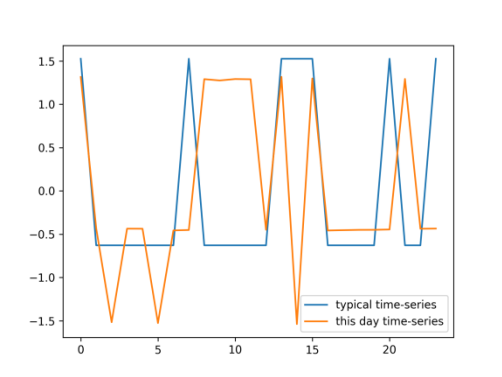

13300000039_3

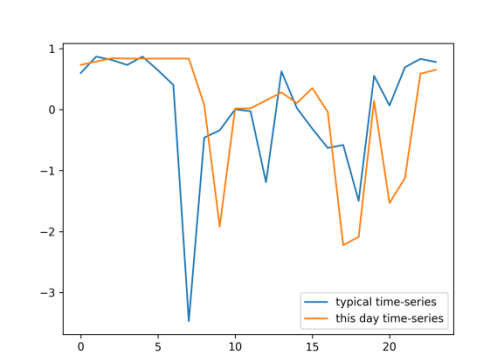
04 事故地序列匹配(2015)

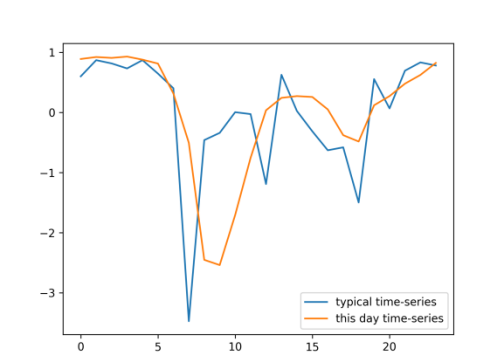
找到事故发生地当日测点序列，计算与其最匹配的模版线，
统计每个模版线的事故频数，
发现20_0和50_3共占有87%。

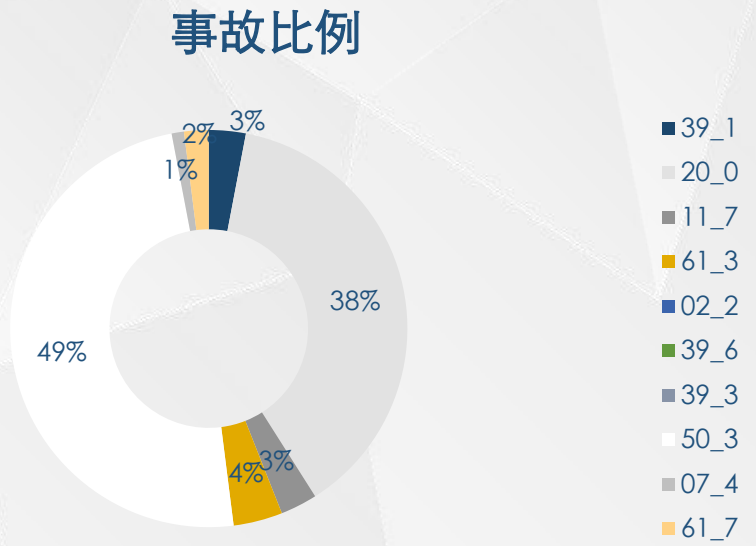
认为20_0和50_3是两条高风险模版线。







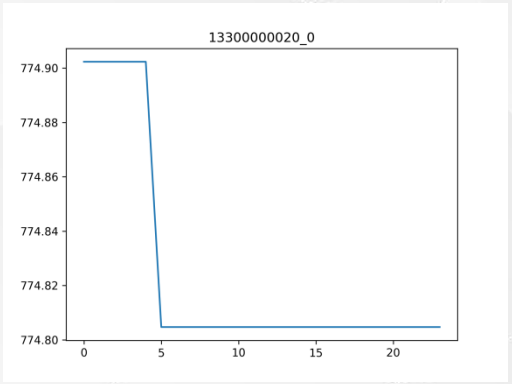




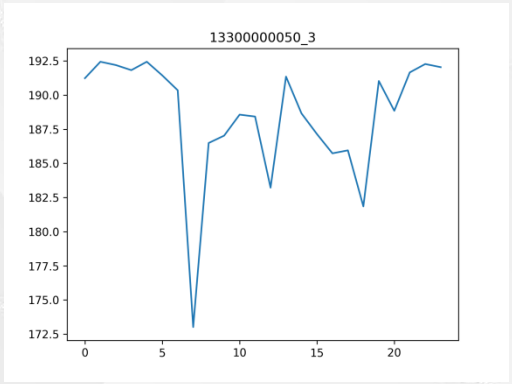
模版线编号	39_1	20_0
事故数量	5	68
事故比例	3%	38%
模版线编号	11_7	61_3
事故数量	5	7
事故比例	3%	4%
模版线编号	02_2	39_6
事故数量	0	0
事故比例	0%	0%
模版线编号	39_3	50_3
事故数量	0	87
事故比例	0%	49%
模版线编号	07_4	61_7
事故数量	2	3
事故比例	1%	2%

04 异常检测模型验证(2016)

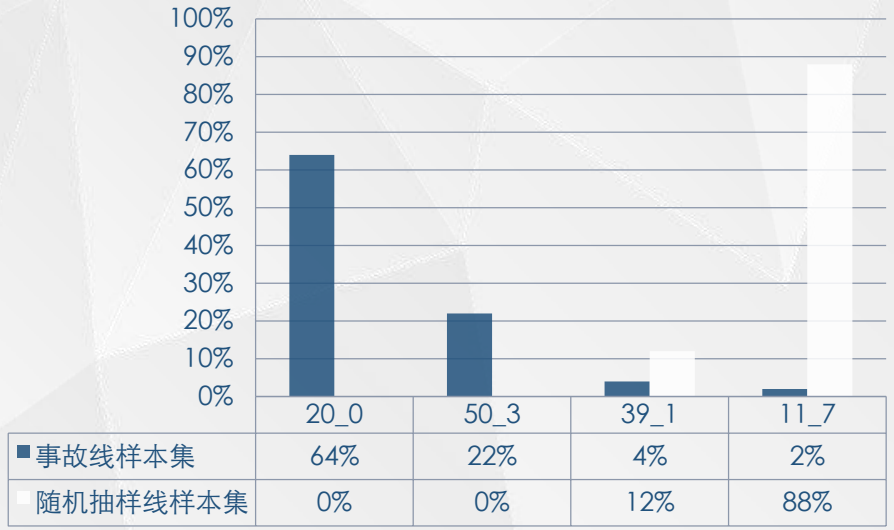
事故地测点序列最匹配的两条模版(高风险)



编号: 20_0

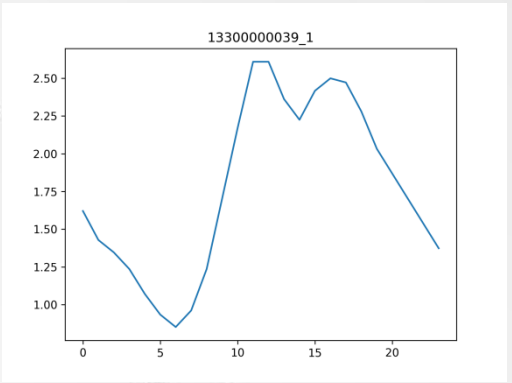


编号: 50_3

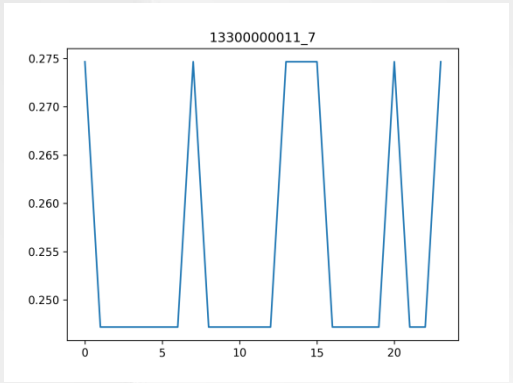


	20_0	50_3	39_1	11_7
事故线样本集	64%	22%	4%	2%
随机抽样线样本集	0%	0%	12%	88%

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

精确率(Precision)=100%
召回率(Recall)=86.3%

即模型的误报率为0%，漏报率13.7%。

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