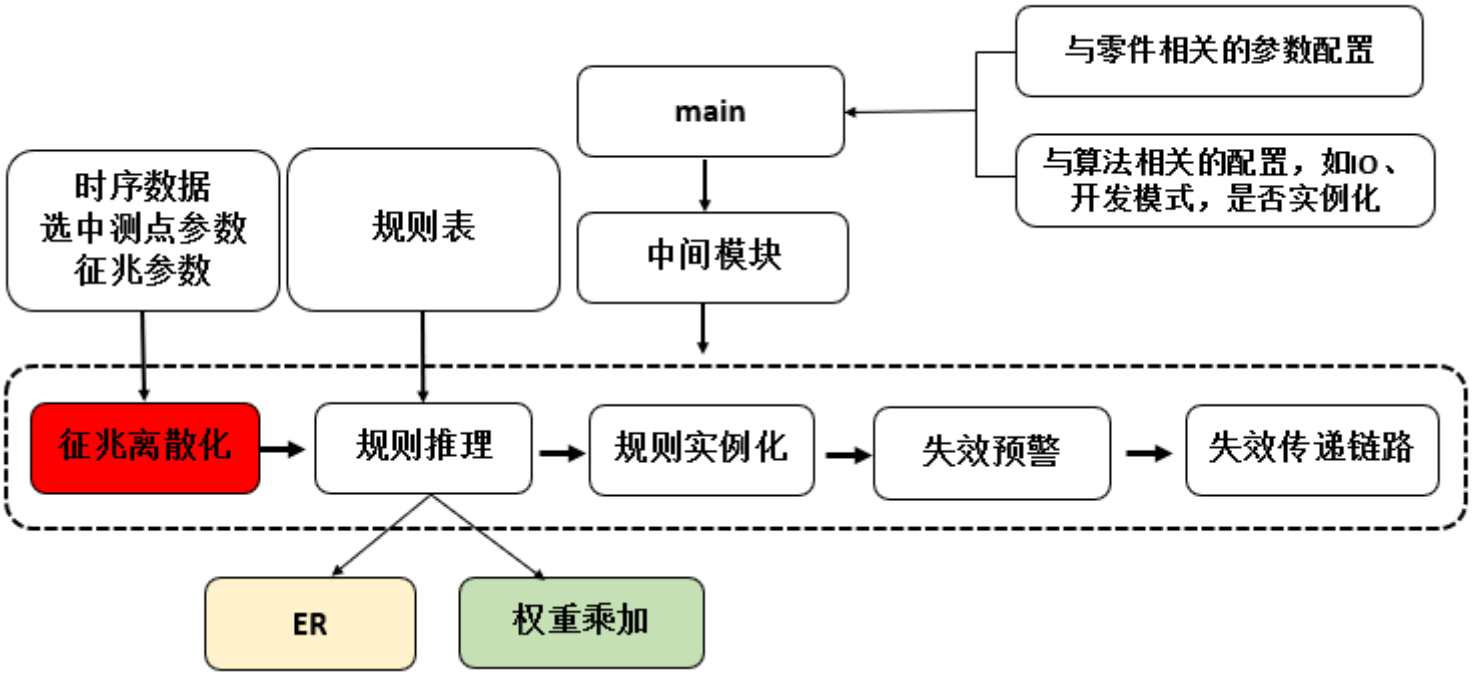
时序数据处理

田宝林

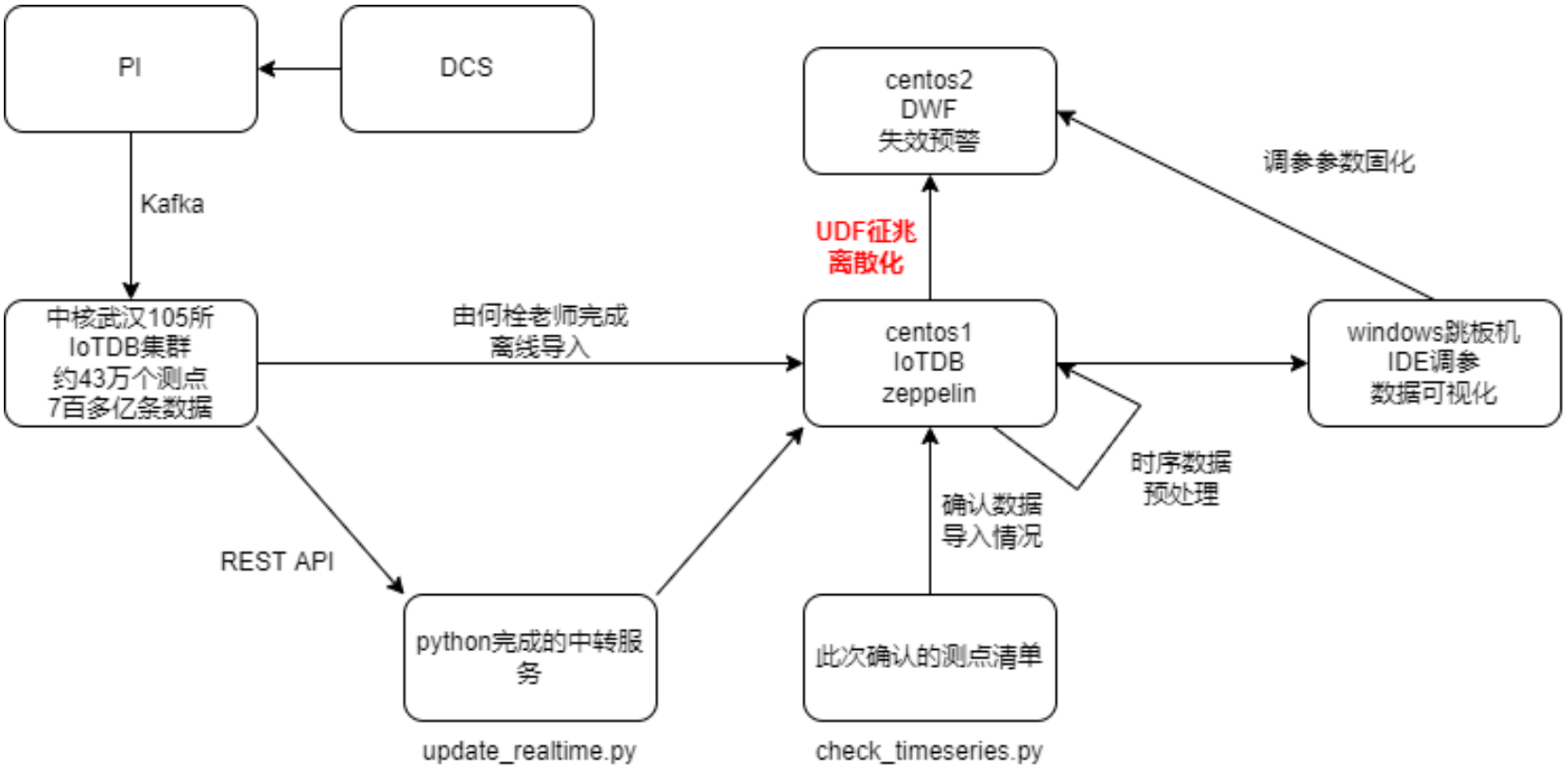
CONTENTS

目录

0. 征兆判断所属流程


与零件相关的参数配置
main
与算法相关的配置，如IO、开发模式，是否实例化
时序数据
选中测点参数
征兆参数
规则表
中间模块
征兆离散化
规则推理
规则实例化
失效预警
失效传递链路
ER
权重乘加

0. 整体架构


PI
DCS
Kafka
中核武汉105所
IoTDB集群
约43万个测点
7百多亿条数据
由何栓老师完成
离线导入
centos2
DWF
失效预警
调参参数固化
UDF征兆
离散化
centos1
IoTDB
zeppelin
windows跳板机
IDE调参
数据可视化
时序数据
预处理
确认数据
导入情况
REST API
python完成的中转服务
update_realtime.py
此次确认的测点清单
check_timeseries.py

1. 原始时序数据查看

- zeppelin/Grafana/python脚本

```
# 全量数据的查看
select re_sample(QF_01_1RCP604MP_AVALUE, ‘every’=’60.0m’) from root.CNNP.QF.01;


select re_sample(QF_01_1RCP604MP_AVALUE, ‘every’=’60.0m’) from root.CNNP.QF.01
where time > 2019-04-01 00:00:00 and time > 2019-06-30 00:00:00 order by time desc
limit 1000;
```

1. 原始时序数据查看

- TimeSeries Analysis 📈 A Complete Guide 📚

https://www.kaggle.com/andreshg/timeseries-analysis-a-complete-guide

1. 通过python pandas 查看时序数据是否有NaN，采样的频率等。

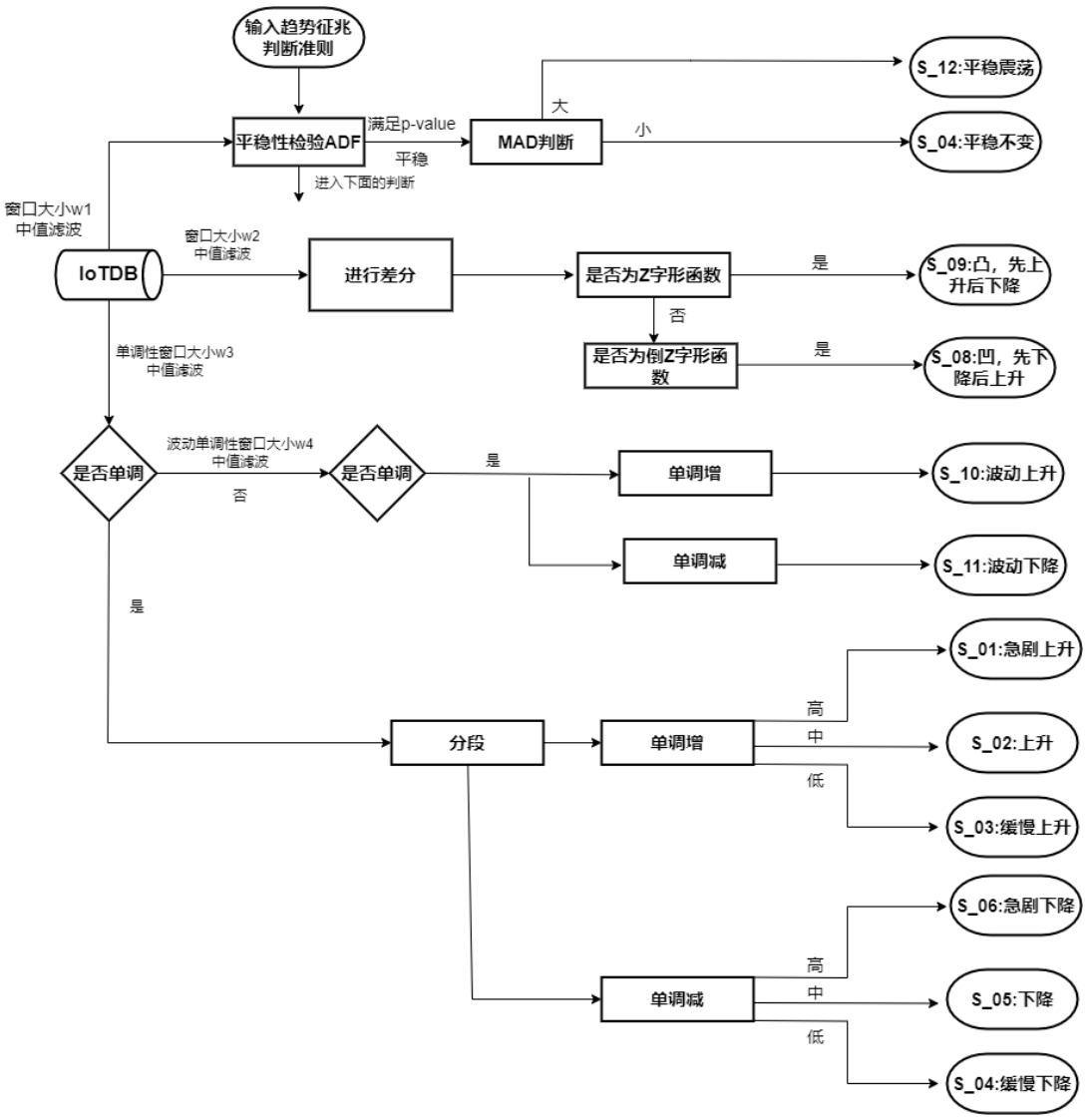
2. 时序数据预处理方法

- 处理缺失时序数据：
  - 一段时间没有变化的时序数据，为了减少存储量而未存入时序数据库。
  - NaN：previous, mean, linear等方法

2. 时序数据预处理方法

- 平滑/重采样：
  - 上采样，下采样：目前建议先上采样数据对齐，然后进行下采样
  - 如何使用合适的采样方法，使得重要的突变能保留，而不重要的噪音能够滤去
  - 目前全部使用中值滤波

3.1 征兆


输入趋势征兆判断准则
平稳性检验ADF
满足p-value
平稳
进入下面的判断
MAD判断
大
小
S_12:平稳震荡
S_04:平稳不变
窗口大小w1
中值滤波
IoTDB
窗口大小w2
中值滤波
进行差分
是否为Z字形函数
是
否
S_09:凸，先上升后下降
是否为倒Z字形函数
是
S_08:凹，先下降后上升
单调性窗口大小w3
中值滤波
是否单调
波动单调性窗口大小w4
中值滤波
是否单调
是
否
单调增
S_10:波动上升
单调减
S_11:波动下降
是
分段
单调增
高
中
低
S_01:急剧上升
S_02:上升
S_03:缓慢上升
单调减
高
中
低
S_06:急剧下降
S_05:下降
S_04:缓慢下降

3.1 征兆判断流程

序号	变量名	中文含义	类型	初始值
1	time_point	用于判断征兆的时间点	String	2019-06-30 11:00:00
2	time_type	时间类型: past, now, future	String	past
3	resample_method	重采样使用的方法：average, mad	String	mad / average
4	resample_fre	从IoTDB中读取时序数据时的采样频率（时间单位：min，可以取小数）	Double	
5	trend_range_day	判断趋势征兆时，所用的时间长度（时间单位：天，可以取小时）		
6	threshold_range_day	判断阈值征兆时，所用的时间长度（时间单位：天，可以取小时），不重采样，防止减少信息	Double	0.01
7	monotonicity_window	进一步进行重采样的滑窗的大小（单位：个）	Int	
8	vibrate_window	判断波动上升窗口的大小，一般比monotonicity_window大（单位：个）	Int	

3.1 征兆判断流程

序号	变量名	中文含义	类型	设定值
9	ADF_pvalue	用于趋势征兆判断中的ADF平稳性检验。当ADF平稳性检验的P-value大于该值时，时序数据非平稳；小于该值时，时序数据平稳	float	0.05
10	z_window	判断是否稳定不变时用到的标准差下限。当小于标准差下限时，判断为平稳不变（单位：个）	float	
11	segment_method	分段的方法：jenks, 极大极小值	String	jenks
12	slope_method	变化率的计算方法：slope, ratio	String	Slope
13	classification_number	分段的段数，一般设置大一些没有问题	Int	4

3.1 征兆判断流程

序号	征兆变量	变量含义	变量类型	默认值
1	S04_std	判断测点平稳不变MAD的阈值	float	0.05
2	S12_std	判断测点平稳震荡MAD的阈值	float	根据数据设置1e9
3	S01_rise_range	单调急剧上升：斜率	float	1e9
4	S02_rise_range	单调上升：斜率	float	1e9
5	S03_rise_range	单调缓慢上升：斜率	float	1e9
6	S05_drop_range	单调缓慢下降：斜率	float	1e9
7	S06_drop_range	单调下降：斜率	float	1e9
8	S07_drop_range	单调急剧下降	float	1e9

3.2 代码框架

- code: 包含

1. 针对一个测点的征兆判断代码；
2. 征兆判逻辑

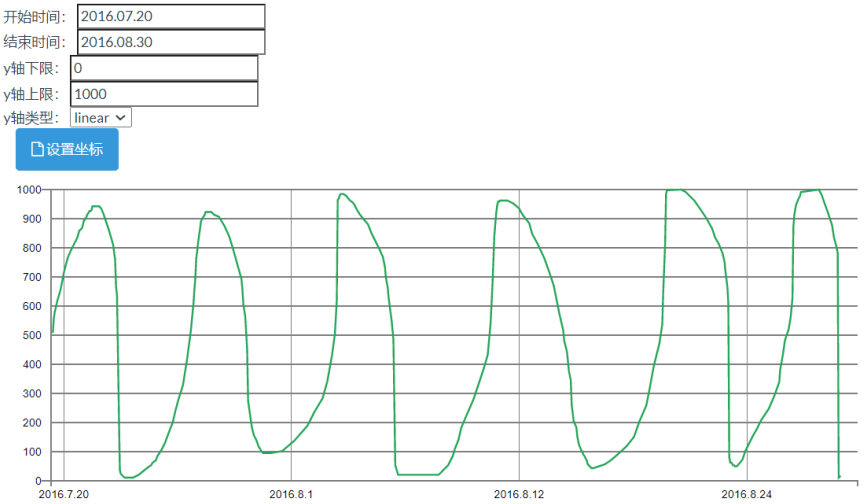
- config: 针对时间序列，趋势征兆，阈值征兆的配置
- images: 存储中间滤波分段后的时间序列展示，方便参数的调试
- timeseries_generation: 时序数据生成，征兆测试

3.2 代码框架

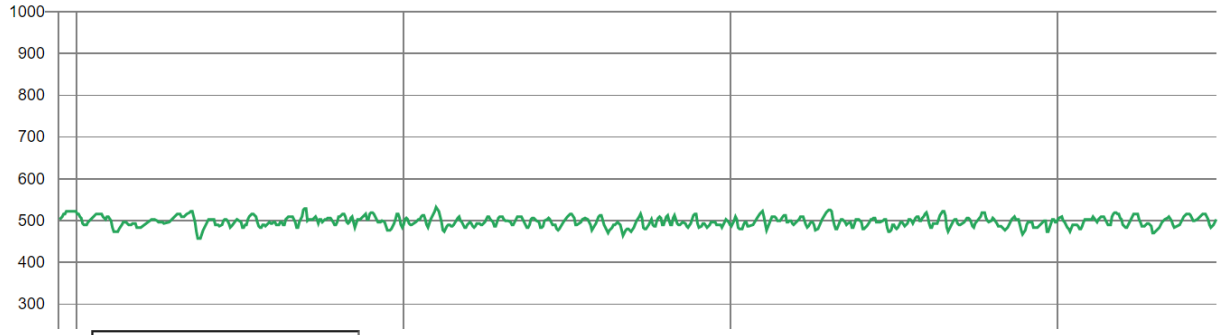
- code: 包含

1. 针对一个测点的征兆判断代码；
2. 征兆判逻辑

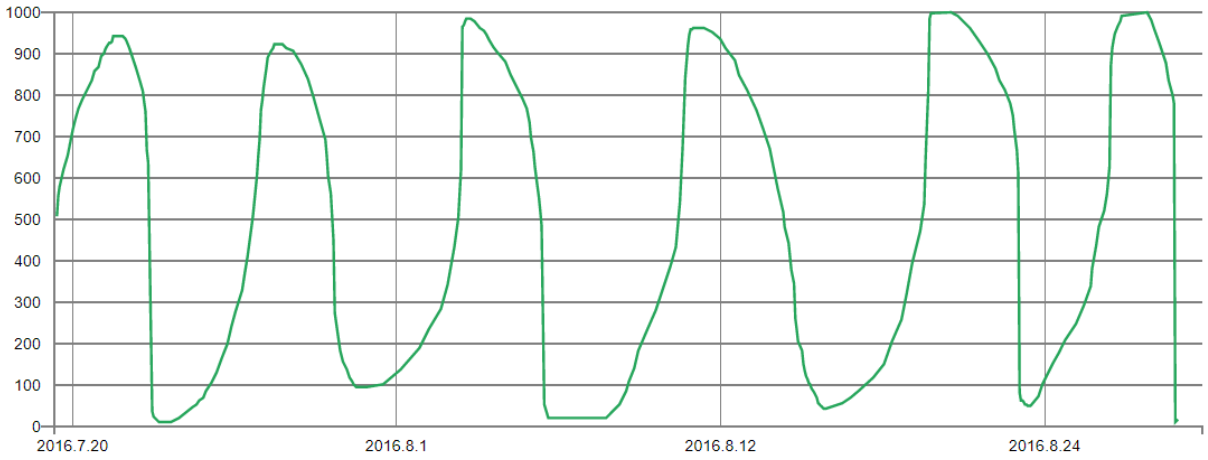
- config: 针对时间序列，趋势征兆，阈值征兆的配置
- images: 存储中间滤波分段后的时间序列展示，方便参数的调试
- timeseries_generation: 时序数据生成，征兆测试



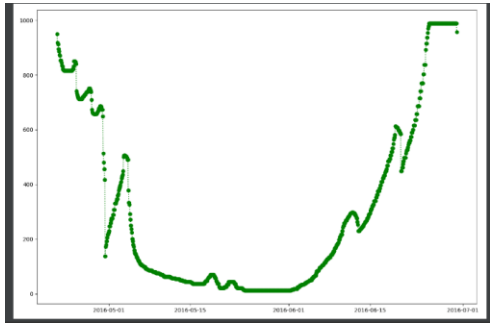
3.3 各种征兆测试：平稳震荡、平稳不



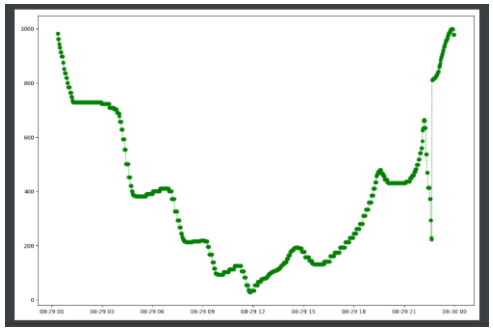
开始时间：2016.07.20
结束时间：2016.08.30
y轴下限：0
y轴上限：1000
y轴类型：linear
设置坐标



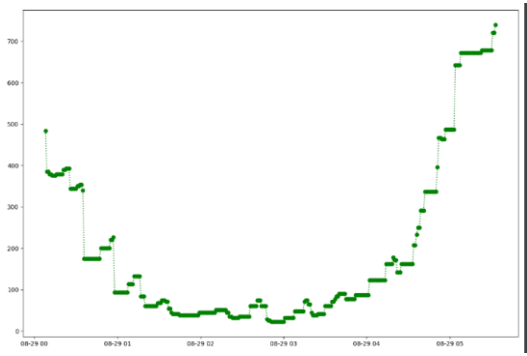
3.3 各种征兆测试：凹凸性



Length = 2months

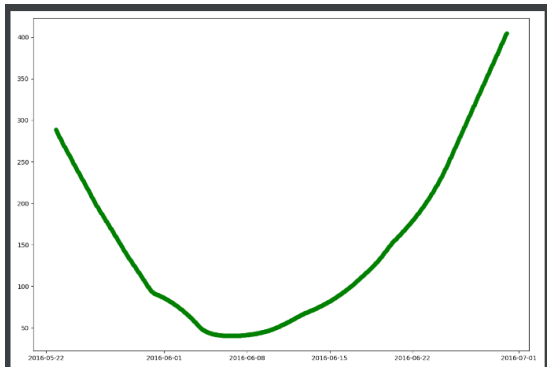


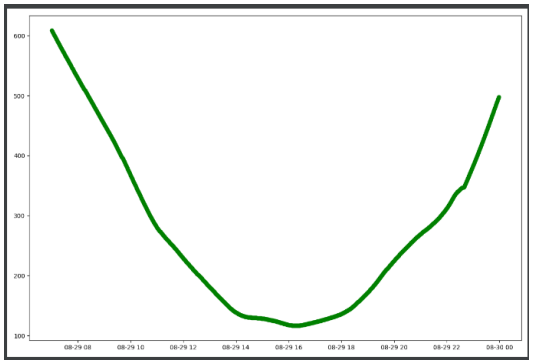
Length = 1day

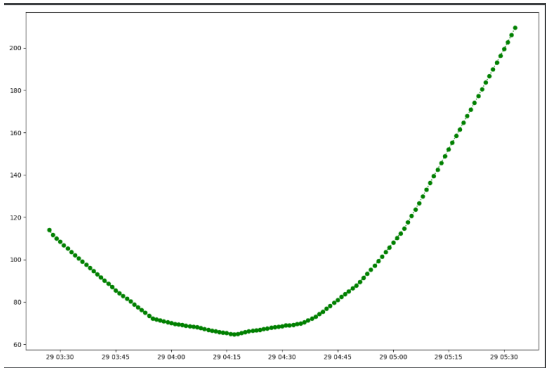


Length = 6hours

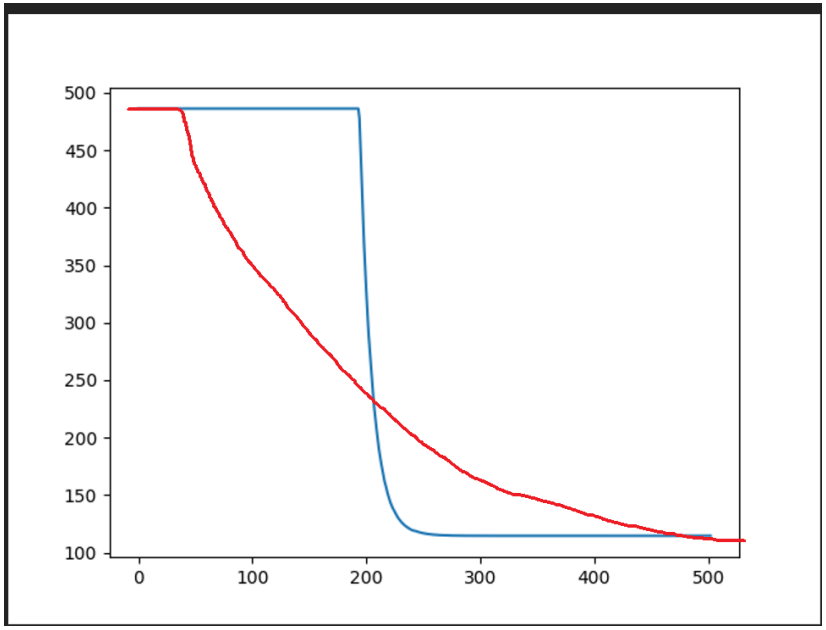
滤波之后

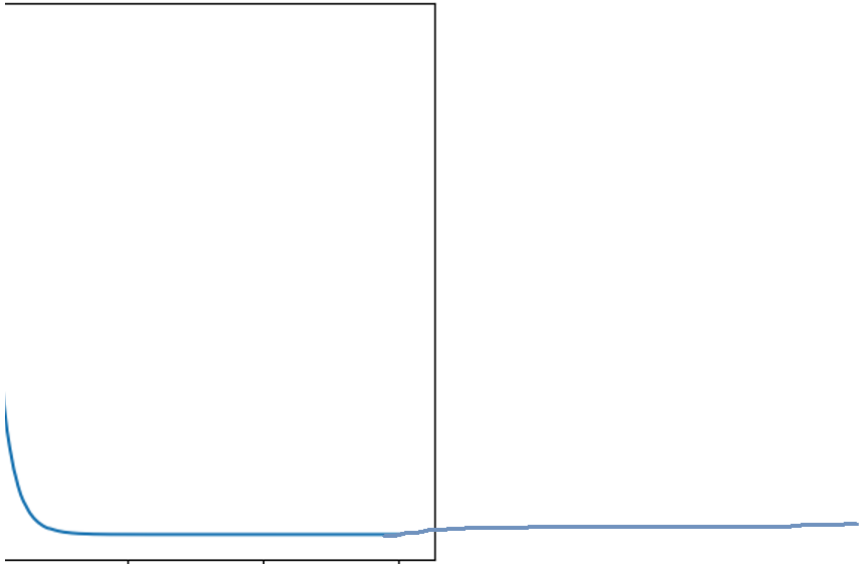




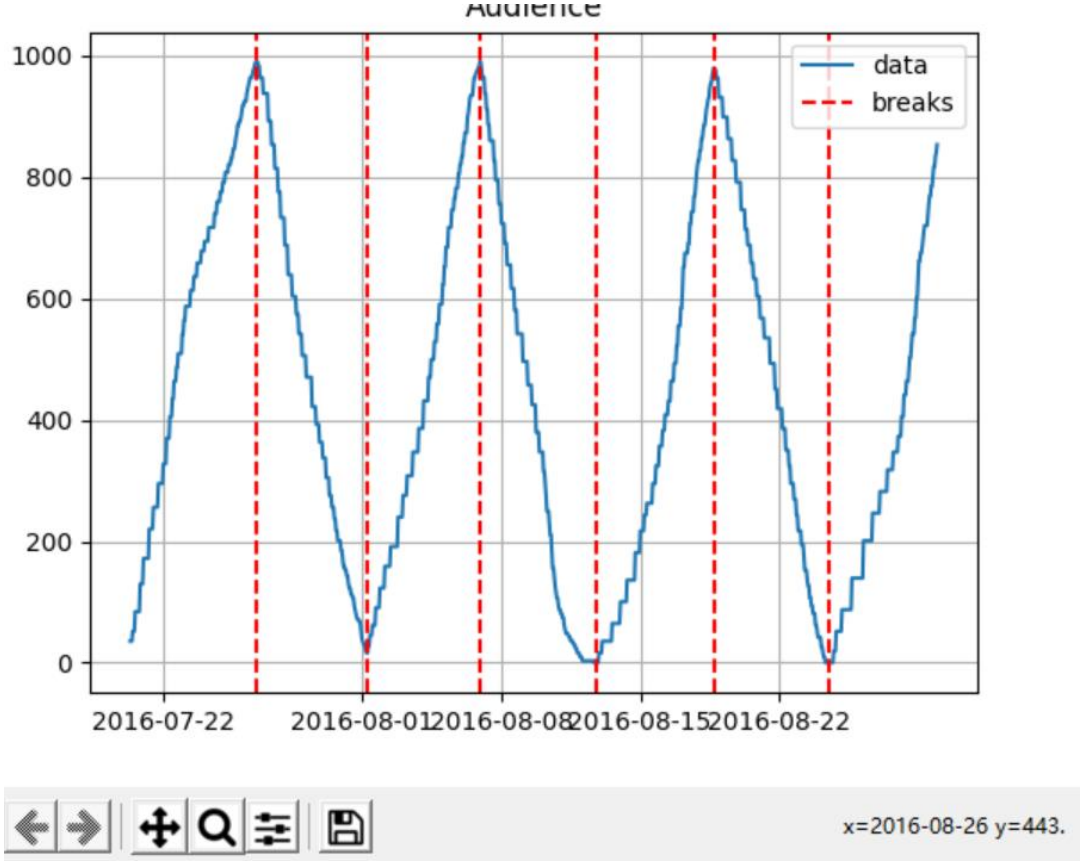


3.3 各种征兆测试：单调性-分段



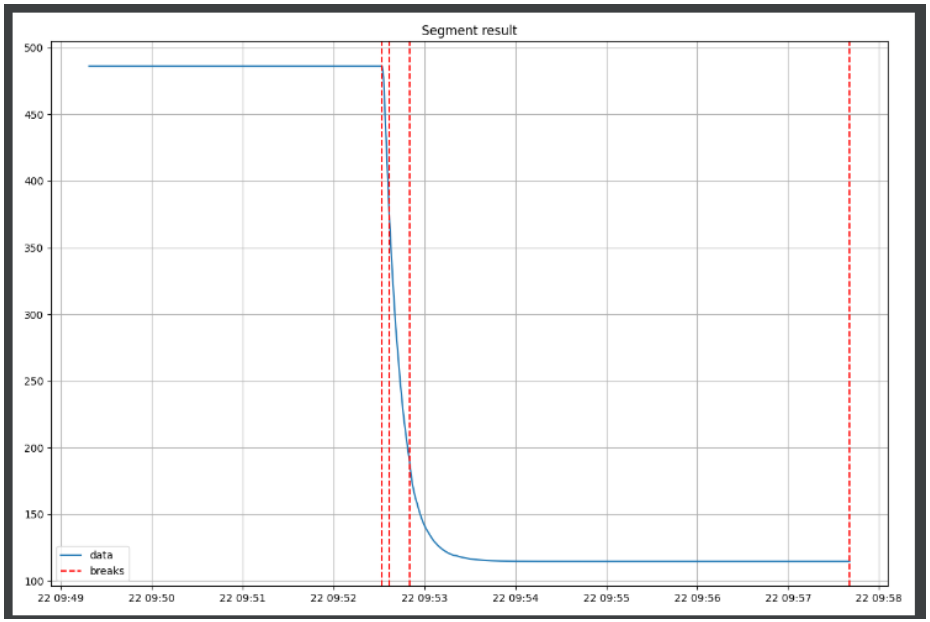


3.3 各种征兆测试：单调性-极值点



3.3 各种征兆测试：单调性-jenks

- 核心思想：
  - 同一组内的数据方差应该尽可能的小
  - 不同组的数据之间差距尽可能的大
- 分段段数设置

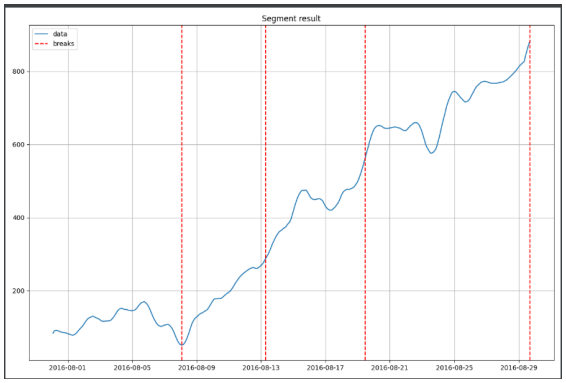


3.3 各种征兆测试：单调性

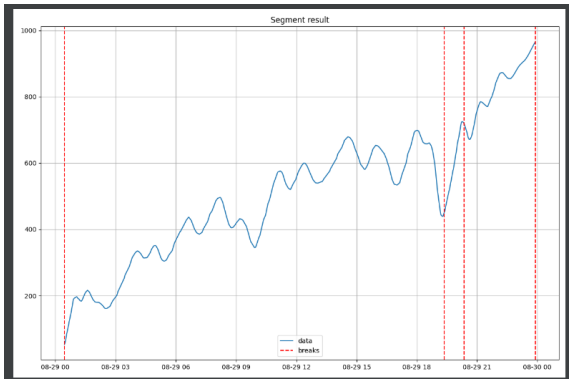
纵看对窗口不敏感

横看对时间不敏感

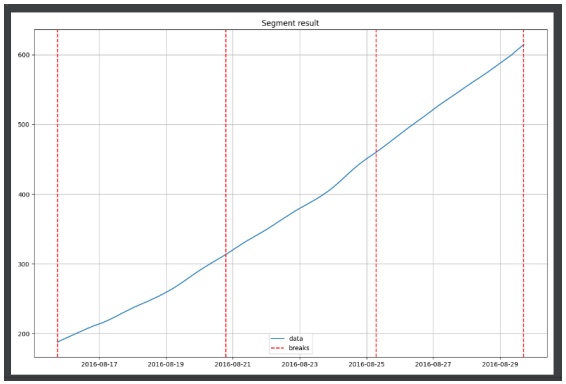
滤波之后

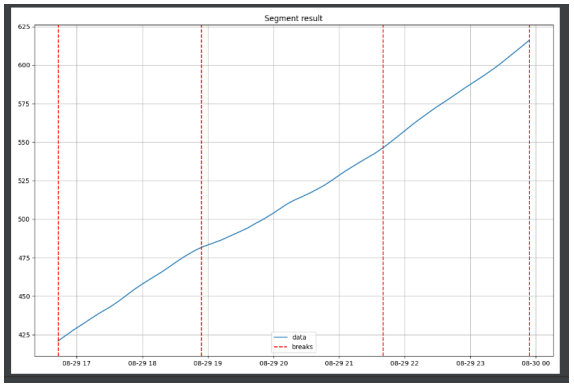


Length = 1month



Length = 1day





3.3 各种征兆测试：波动单调性

- 唯一的区别：比单调上升滤波的窗口更大

3.4 征兆回测

- 正确的时间段能否被报出来。
- 误报率是多少，能否降低误报率，合理的误报。

3.5 时序数据replay

- 可以加速
- 查看具体实时的系统是如何运行和正确的进行预警

4.1 经典时序数据研究方法：AR, MA, ARMA

- Auto Regression
- Moving Average
- x(i)为时刻i时序序列的值
- ε_t 时刻t的冲击信号

$$
\begin{aligned}
AR(1) &: \quad x_t = \phi x_{t-1} + \varepsilon_t \\
AR(p) &: \quad x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t \\
MA(1) &: \quad x_t = \varepsilon_t + \theta \varepsilon_{t-1} \\
MA(q) &: \quad x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \\
ARMA(p,q) &: \quad x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t \\
&\quad + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}
\end{aligned}
$$

4.1 经典时序数据研究方法：AR, MA, ARMA

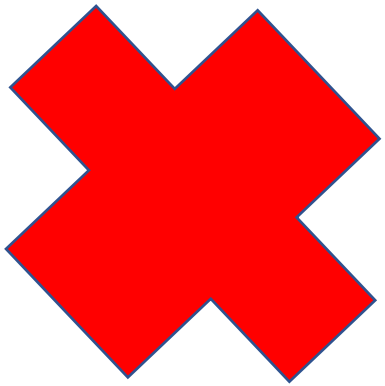
- 平稳性检验，非平稳时序数据通过差分转化为平稳时间序列，或者通过非线性拟合
- 参数估计：AIC, BIC指标

4.2 时序选择

- 时序数据相关性计算，PCA进行相关时序数据删除
- 由于目前测点较少，重要测点较少，没有进行特征的选取流程

4.3 监督学习

- 对时序数据打标记，通过训练，验证，测试流程预测是否发生故障。



4.4 无监督学习的故障诊断

- 3 sigma原则
- 孤立森林：将少量迭代就能进行划分的点判定为离群点

4.5 不考虑时序数据相关性的预测

- 选取多个时间序列构建训练矩阵
- 回归问题：Logistic Regression

4.6 考虑时序数据相关性的预测

- LSTM

4.6 时序数据预测准确评估指标

y_i为真实值，$\hat{y}_i$为预测值

1. Mean Squared Error:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

2. Root Mean Square Error:

$$RMSE = \sqrt{MSE}$$

3. RMSLE(Root Mean Square Logarithmic Error):

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|log(y_i + 1) - log(\hat{y}_i + 1)|}$$

这种方法适用于数值序列出现长尾分布的情况。

4. RMSPE(Root Mean Square Percentage Error):

$$RMSPE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{y_i - \hat{y}_i}{y_i})^2}$$