**VOCS数据分析报告**

# 1 需求

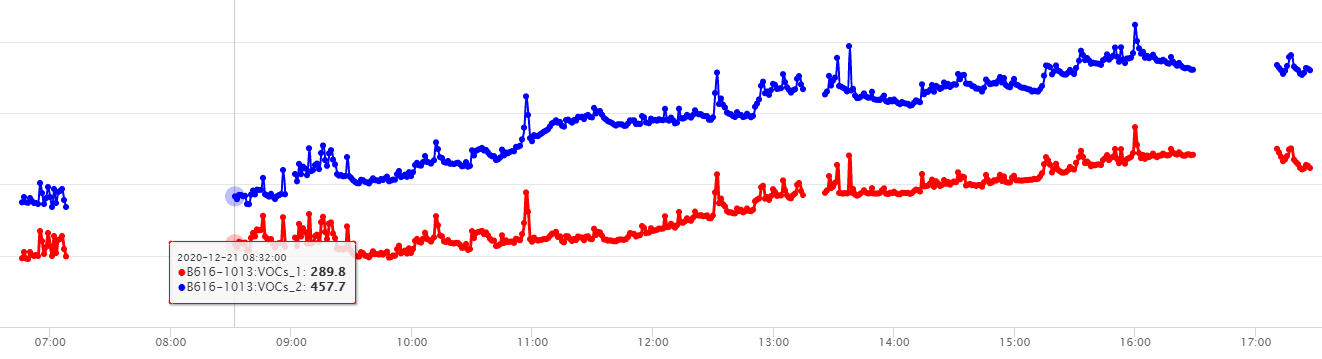
检测Vocs异常数据，解释脉冲点行为。

# 数据特征

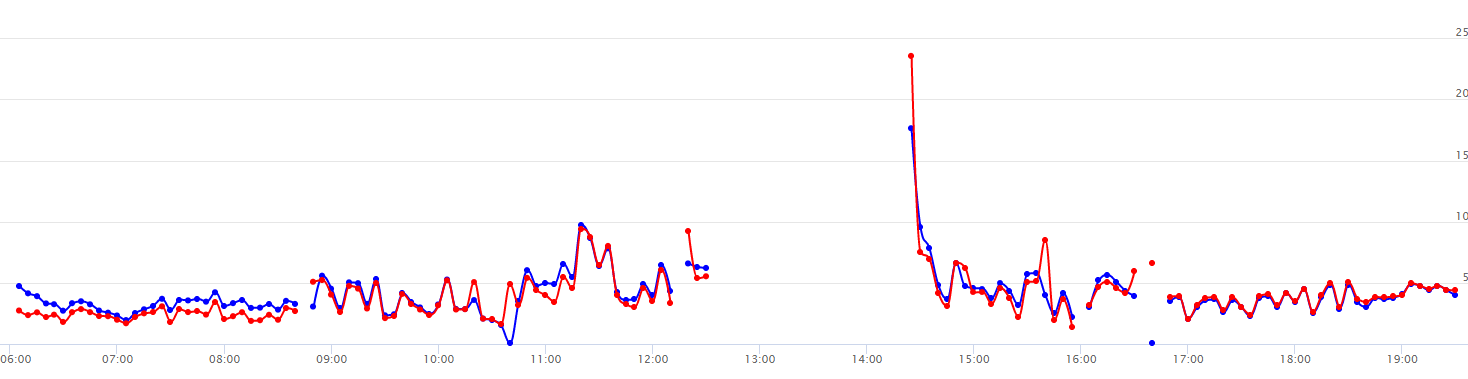
## 正常设备和异常设备的走势图

### 2.1.1正常设备

**B616-1013 12-21日数据**



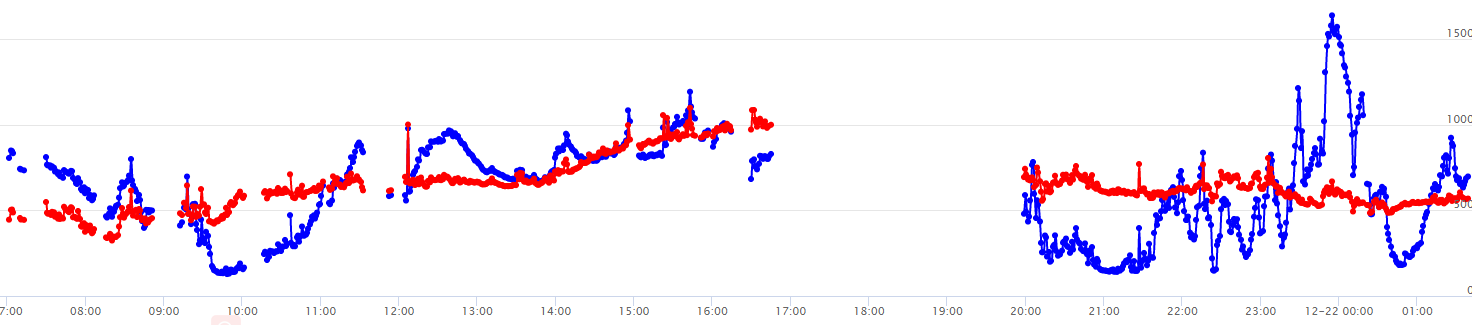
**B616-106D 12-21日数据**



双核设备走势趋同

### 2.1.2异常设备

**B616-1011 12-21日数据**



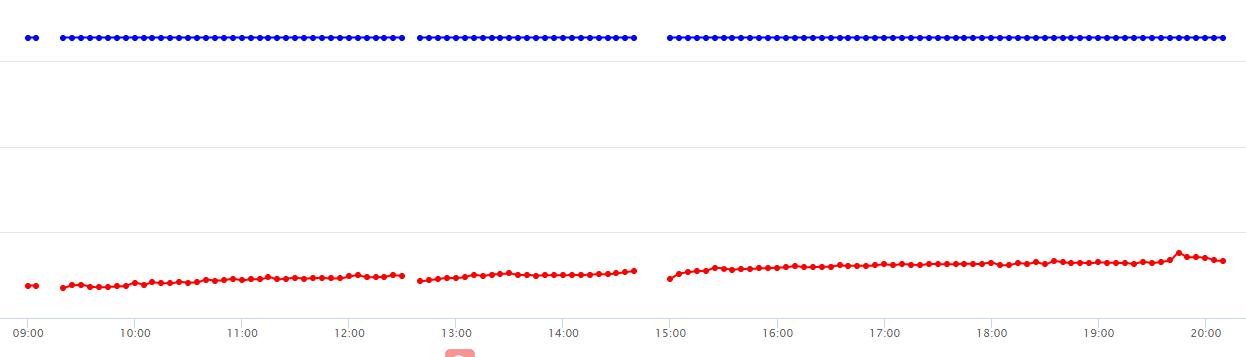
双核设备走势不同，交叉点很多

**B616-1020 12-21日数据**



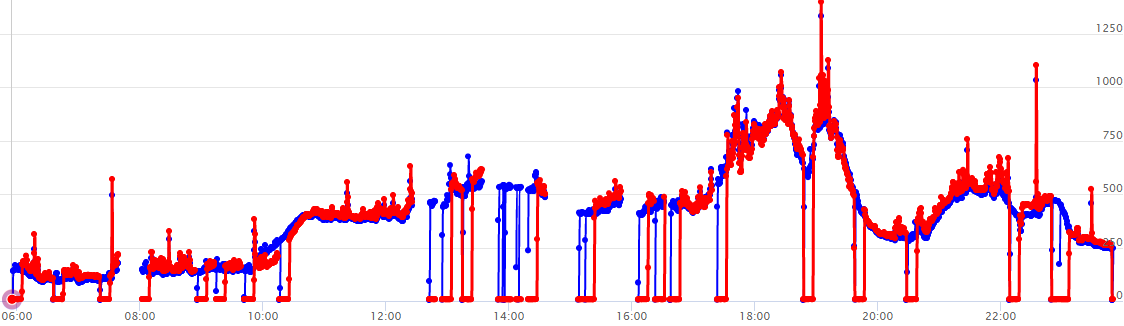
单核vocs\_2宕机，数据保持11不变

**B61A-10E7 12-21日数据**



单核vocs\_2宕机，数据保持0不变(WP页面中显示为6553.5，实际系统中不存在vocs\_2的数据)

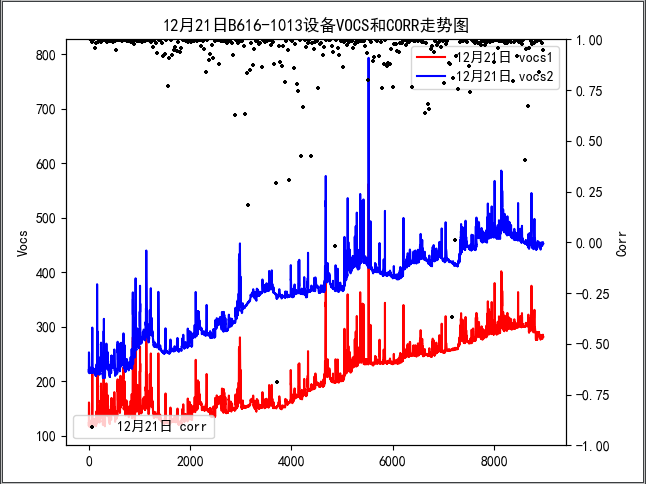
**B616-101D** 12-21日数据

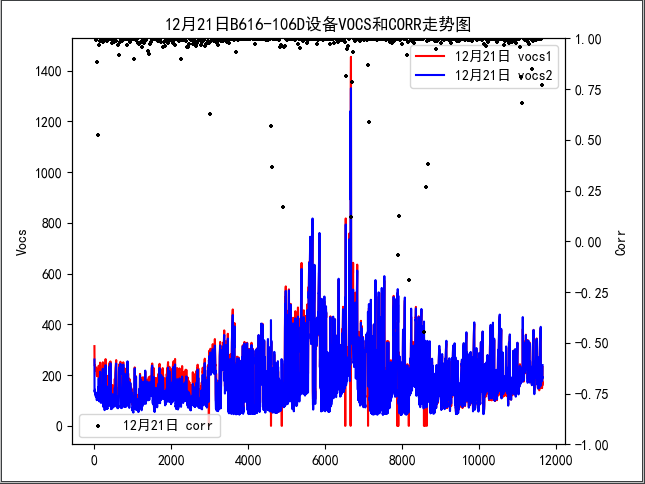


Vocs设备间歇性宕机

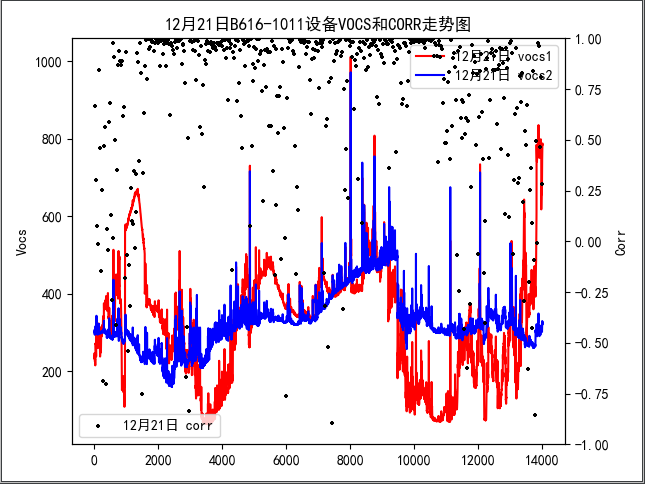
## 2.2 设备相关系数走势图（相邻30个数据为一组）

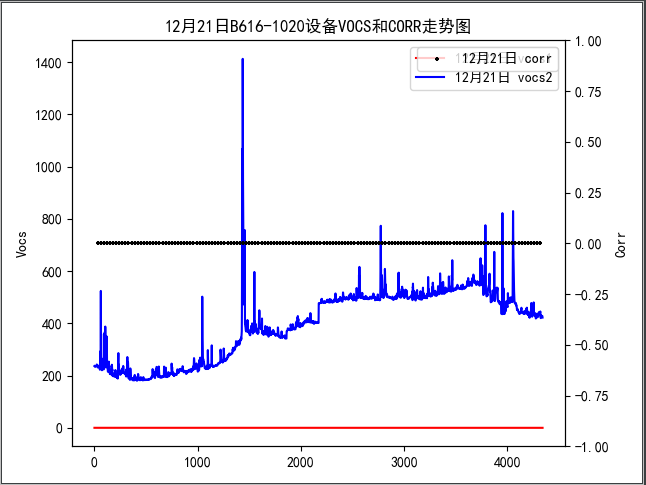
### 2.2.1正常设备（不考虑时间断层）

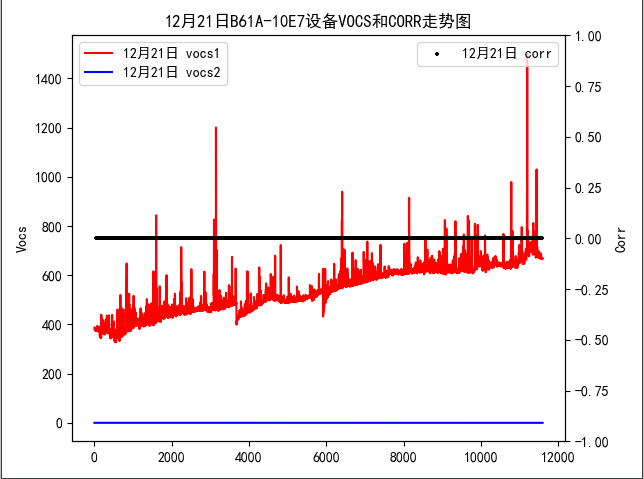


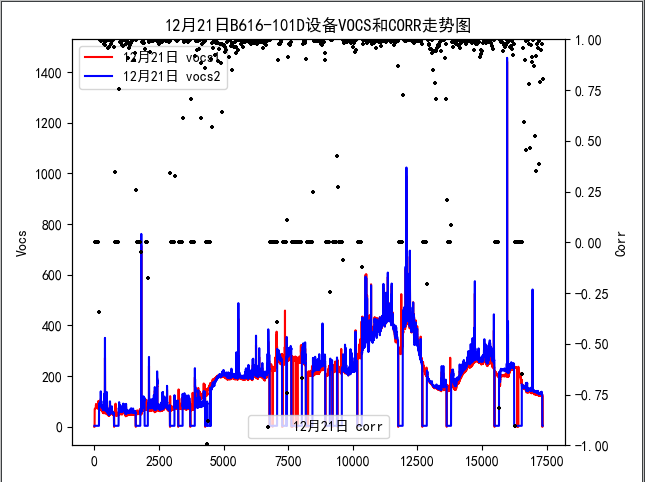


### 2.2.2异常设备（不考虑时间断层）



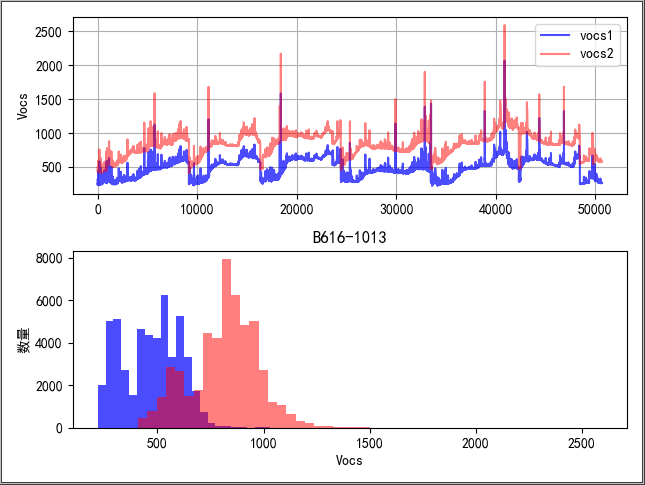


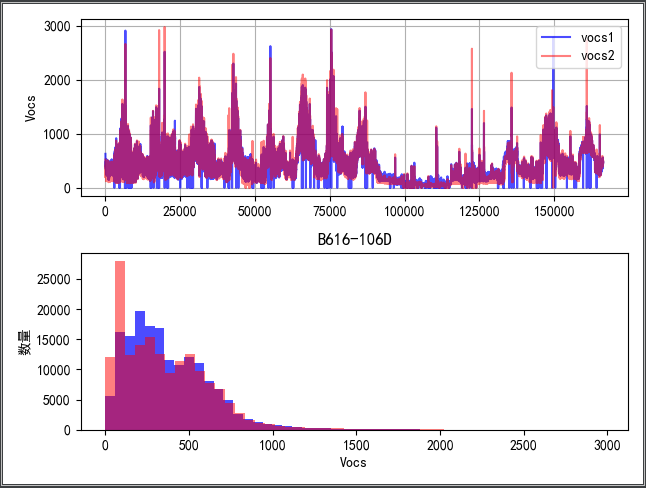




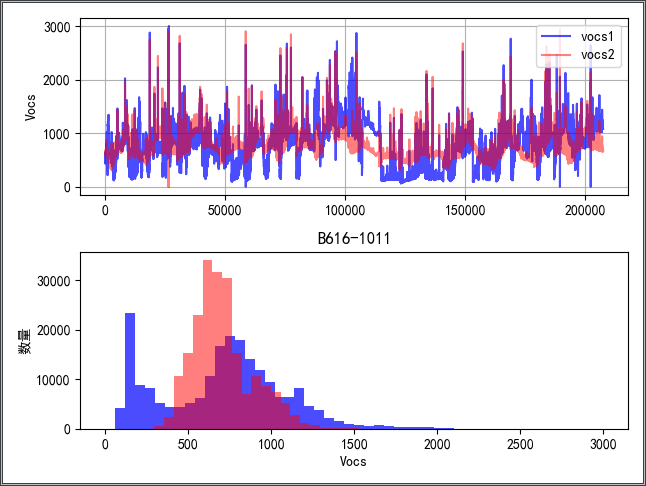
## 2.3 设备分布直方图

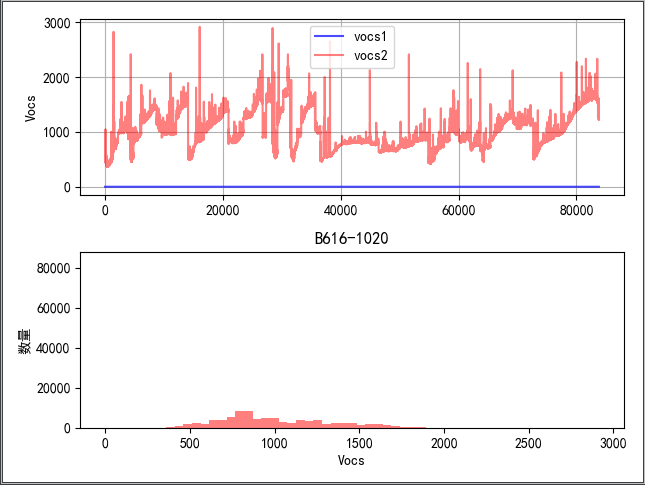
### 2.3.1正常设备（不考虑时间断层）

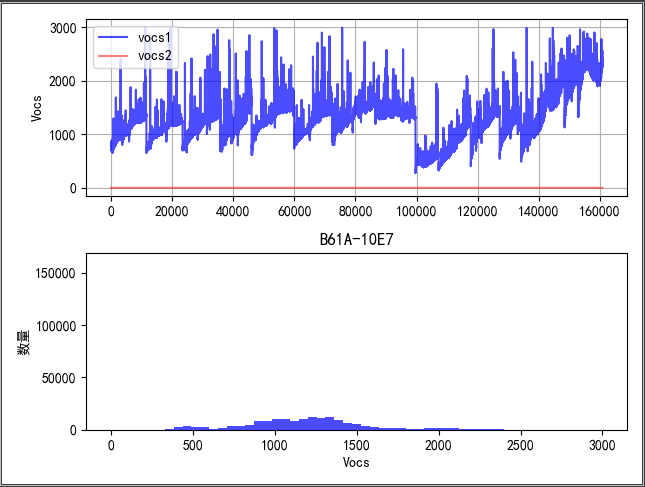


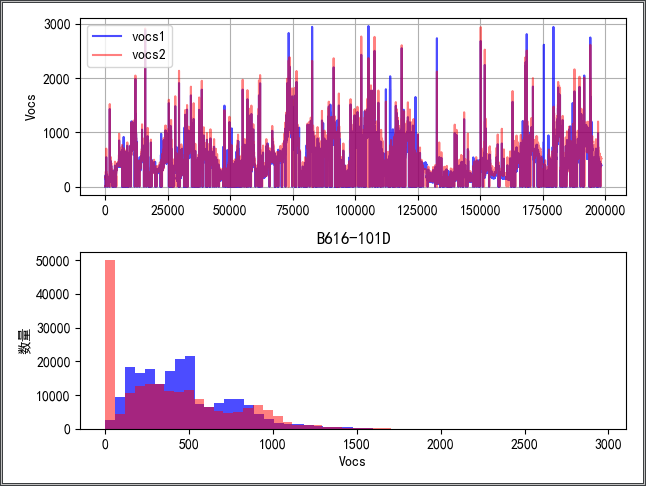


### 2.3.2异常设备（不考虑时间断层）









# 解决方案

检测异常设备，从而判断出其数据亦为异常，并分辨出正常设备；

设计模型检测出正常设备的正常数据，剔除异常数据以后，分析正常设备的数据统计指标，根据其统计指标分辨异常点。

注：

正常设备判断标准：

双核相关系数用corr表示

corr=0表示至少一核宕机；

0＜|corr|＜0.3表示低度相关

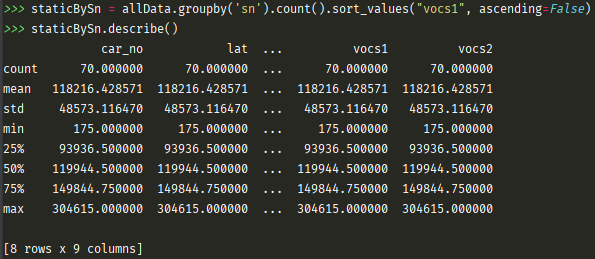
0.3≤|corr|＜0.5表示弱相关

0.5≤|corr|＜0.8表示中度相关

0.8≤|corr|＜0.95表示强相关

|corr≥0.95表示显著相关

由于本文目前只做正常设备判断，故不考虑时间因素，将所有数据连在一起计算相关度。（**此处做了一个假设**：假设设备每次开机都在该设备所属的可接受范围值内）



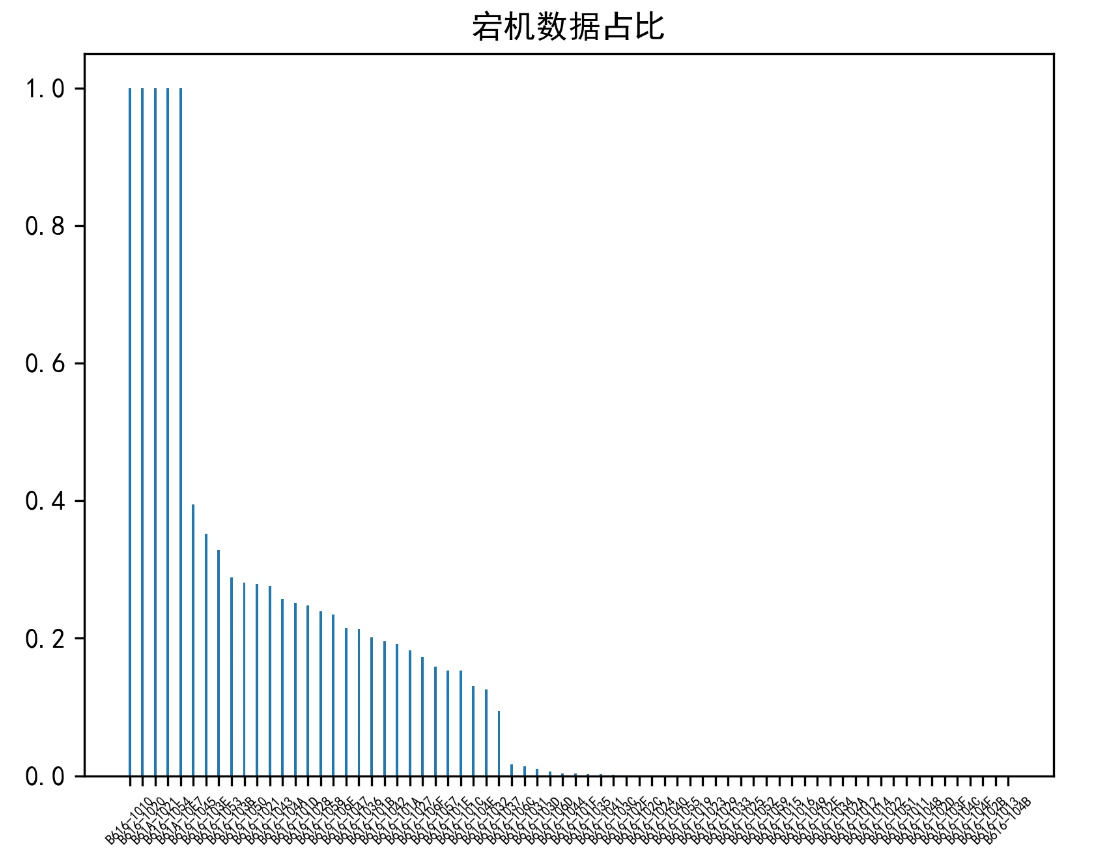
## 3.1 正常设备检测

### 3.1.1 离线计算

通过莱芜双核设备数据特征可以看出，虽然两个设备数据有截距差，但是正常设备的两份数据的走势是正常的、 趋同的，而异常设备则表现为3种特征：

1. 其中一个核宕机，数据表现为长时间Vocs数据为6.1、8.1、11；
2. 双核均长时间宕机；
3. 单核间歇性宕机或双核间歇性宕机；
4. 虽然双核都在运行，但数据走势不同。

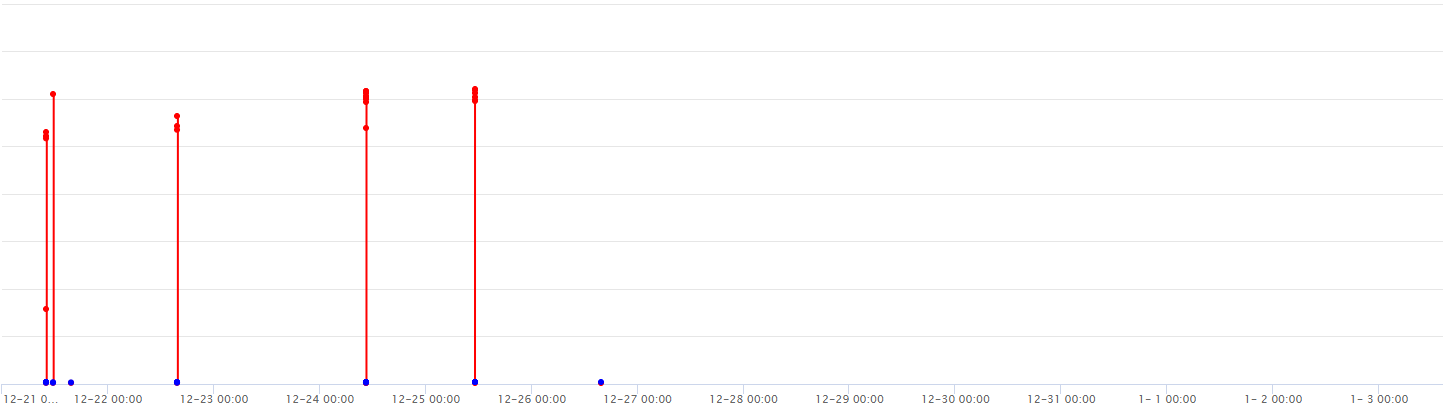
（一）剔除100%宕机设备数据



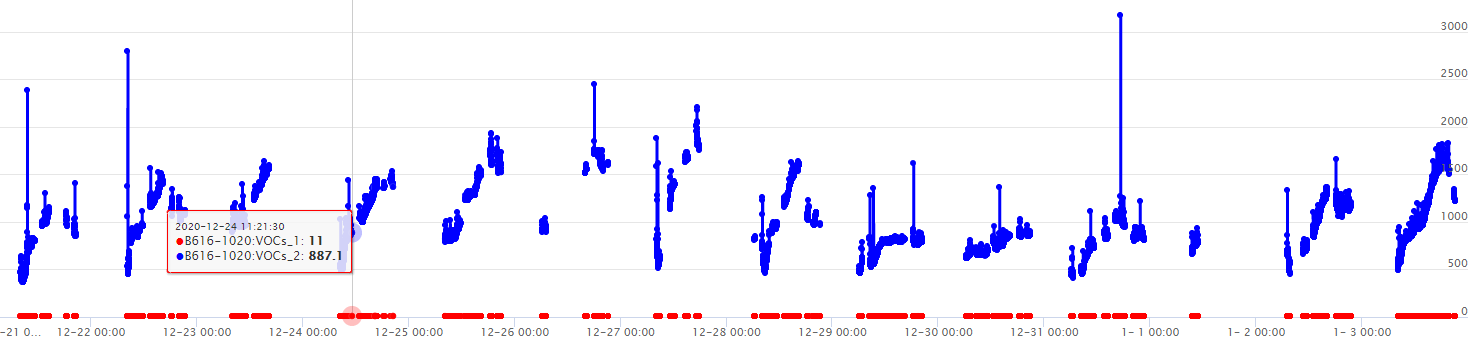
宕机数据计算逻辑：至少有一个核的数据≤11，则认为宕机。

通过宕机数据占比可以看出，有5台设备在2020-12-21至2021-01-03连续两周产生的数据中，宕机率为100%。分别是：

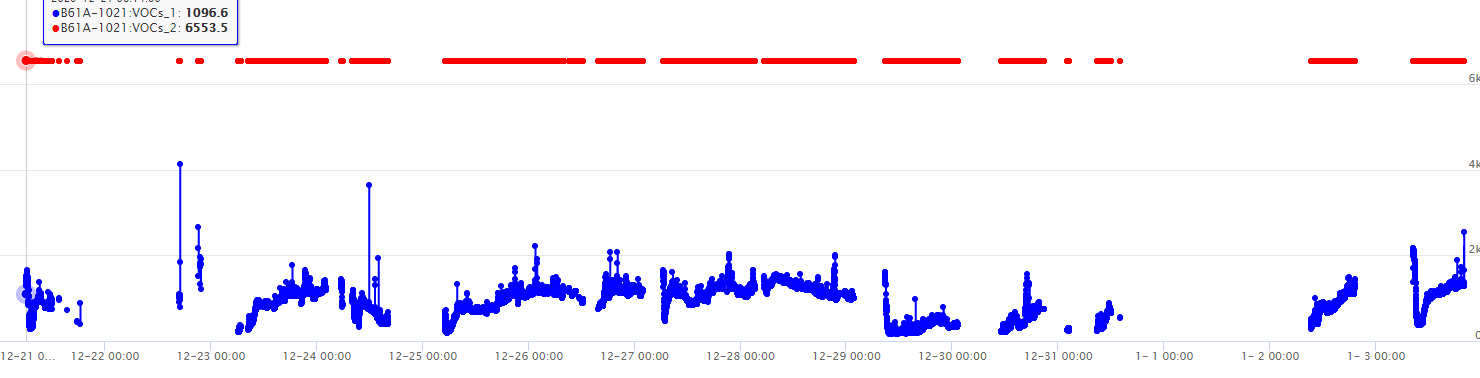
B616-1010



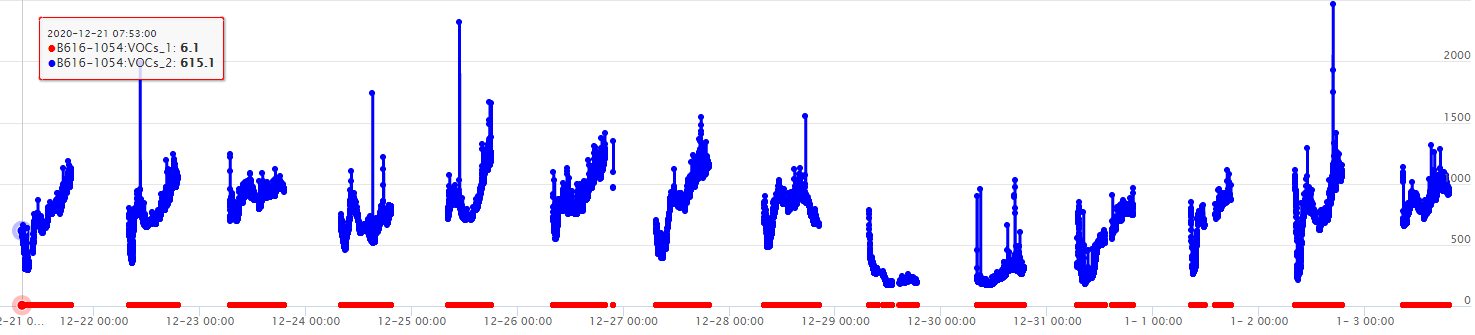
B616-1020



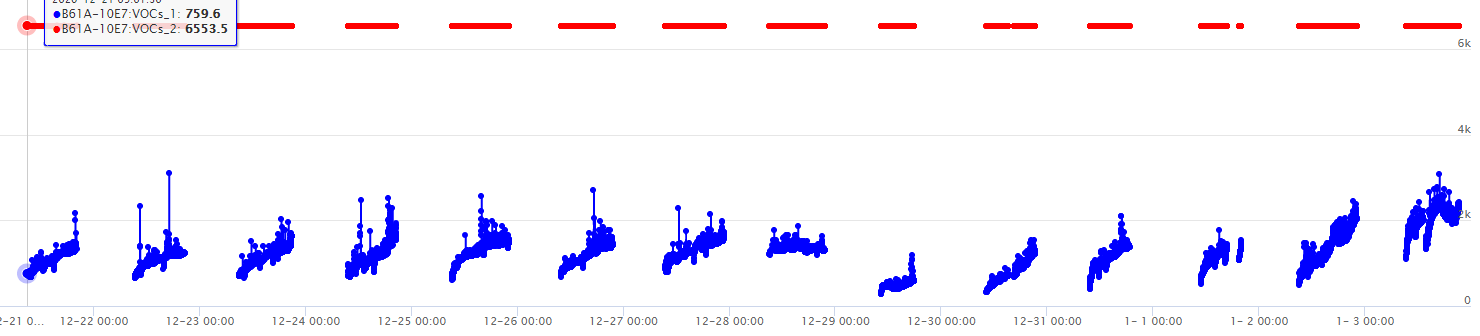
B61A-1021



B616-1054



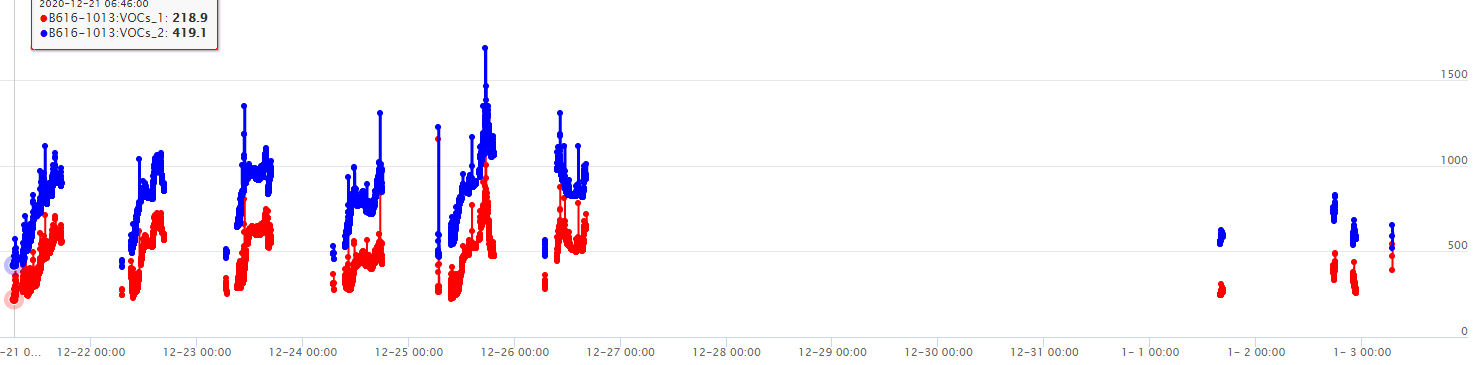
B61A-10E7



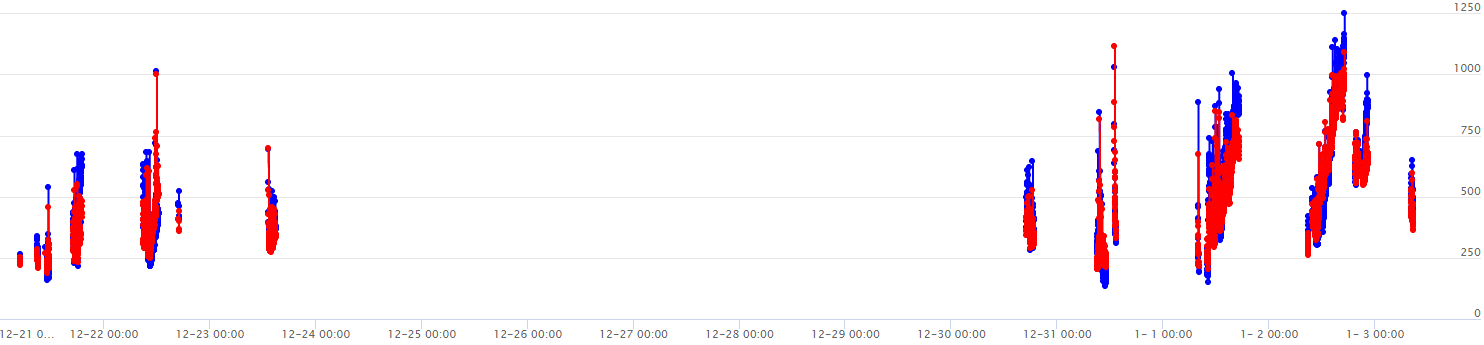
占70台设备的5/70=7%。

其余65台设备宕机率不超过40%。有38台设备宕机率不超过1%。有2台设备未宕机，分别是：

B616-1013



B616-104B



（二）通过相关系数corr判断设备的正常指数

针对最近一周的数据，对每30个数据计算两个设备的相关系数：

其中x, y分别是两个向量，本文中即vocs\_1和vocs\_2。相关系数corr取值范围是 （-1，1），正数表示x, y正相关，一个变大另一个也变大；负数表示负相关，一个变大，另一个变小；0表示零相关，即x和y正交。

通过上述描述可得，|corr|越接近1，表示双核的数据越同步，否则越不同步，即异常。

离线计算的主要目的是通过对历史数据的计算，找到一个判断是否正常设备的corr的阈值。这个阈值是需要不断迭代更新的，更新周期可以为一天。

### 3.1.2 实时计算

针对异常设备的特点，可以实现通过计数为30个的计数窗口（主要是针对数据不连续产生的情况）或者90秒的时间窗口（针对数据连续产生的情况）进行实时监测。算法采用计算两个设备数据的相关系数（相关系数能够很好的反应其走势趋同情况）是否超过阈值。

## 3.2 异常数据检测

# 4