**VOCS数据分析报告**

# 1 需求

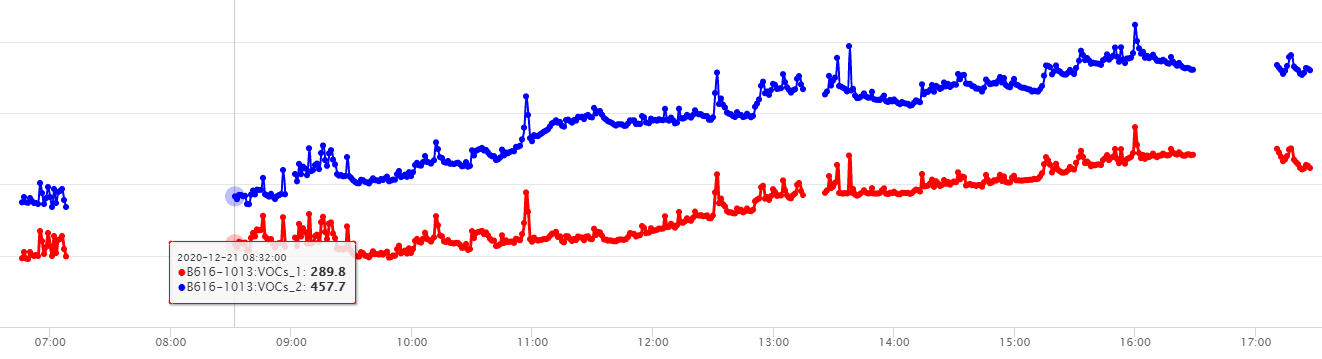
检测Vocs异常数据，解释脉冲点行为。

# 数据特征

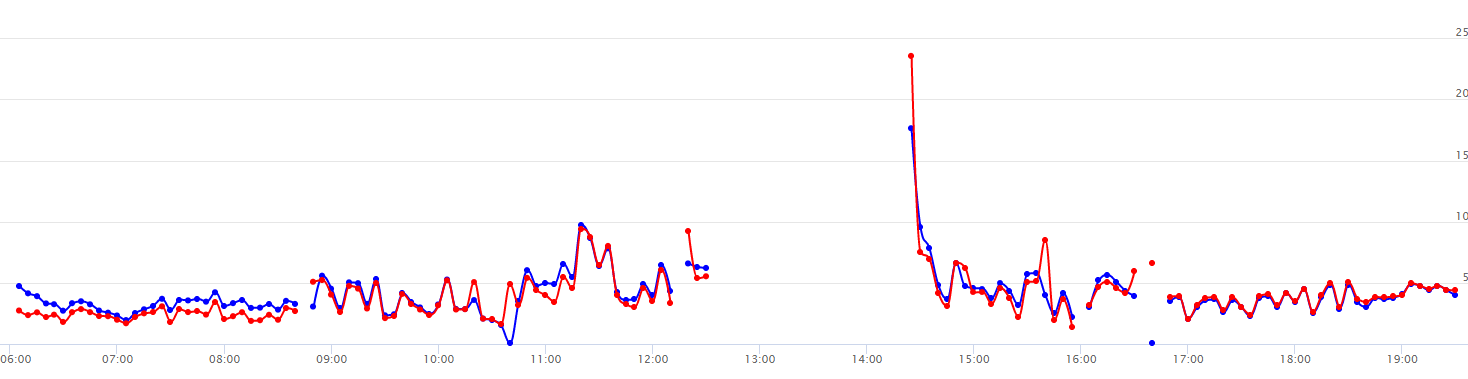
## 正常设备和异常设备的走势图

### 2.1.1正常设备

**B616-1013 12-21日数据**



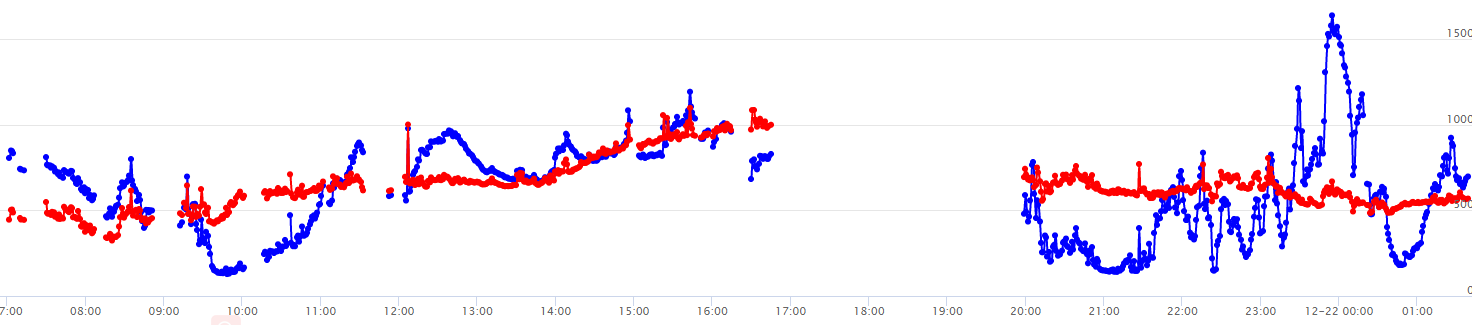
**B616-106D 12-21日数据**



双核设备走势趋同

### 2.1.2异常设备

**B616-1011 12-21日数据**



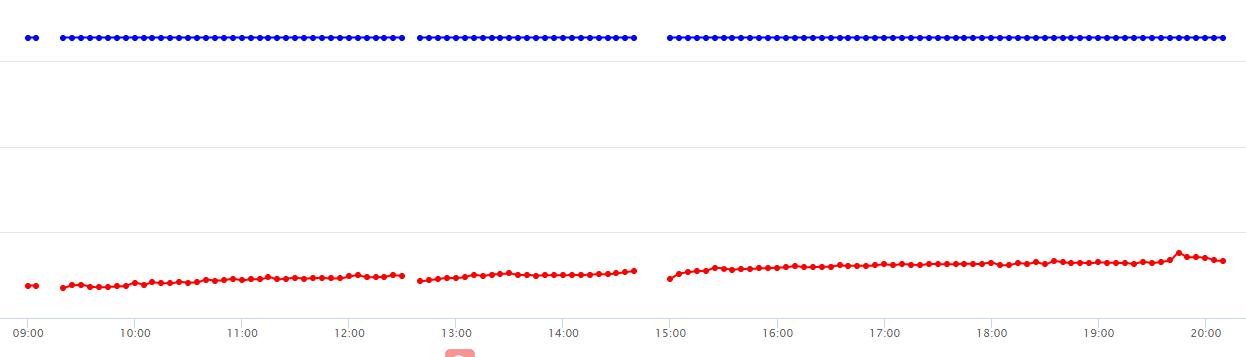
双核设备走势不同，交叉点很多

**B616-1020 12-21日数据**



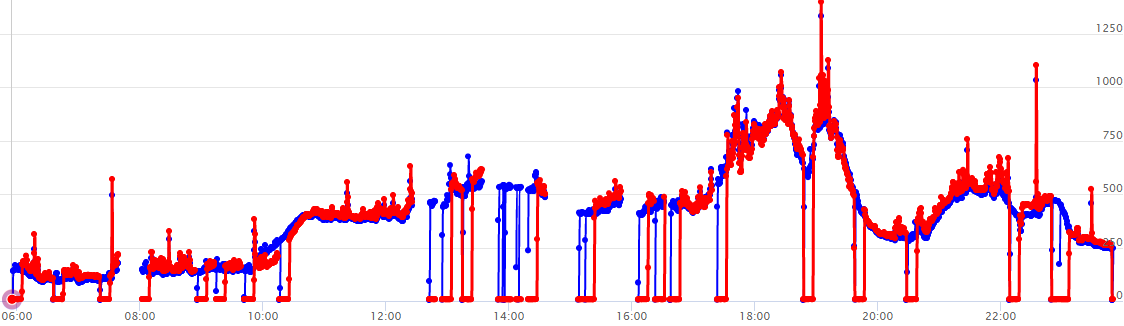
单核vocs\_2宕机，数据保持11不变

**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)

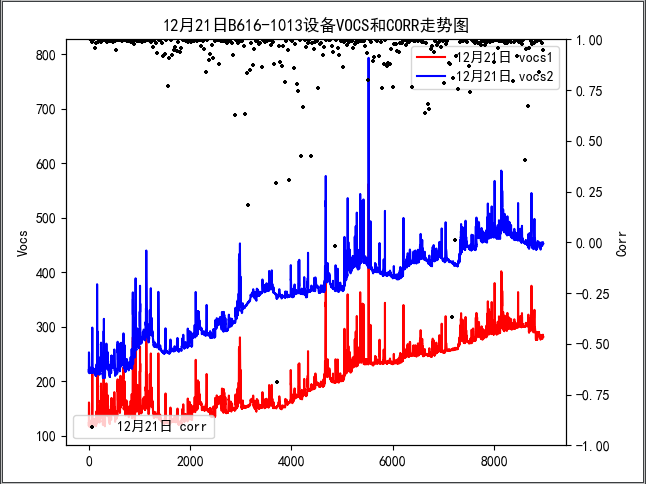
**B616-101D** 12-21日数据

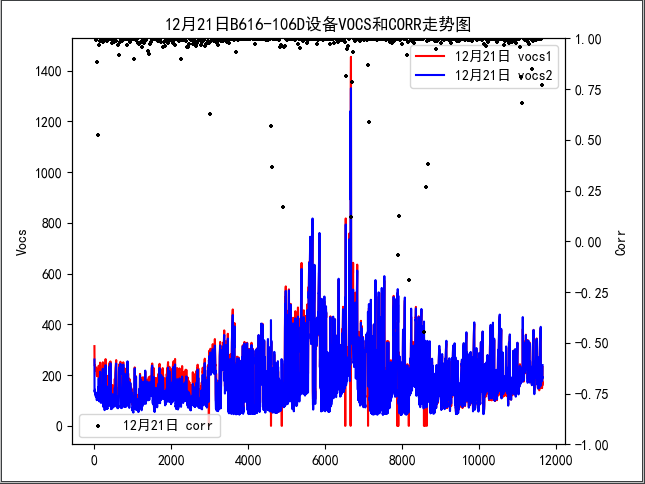


Vocs设备间歇性宕机

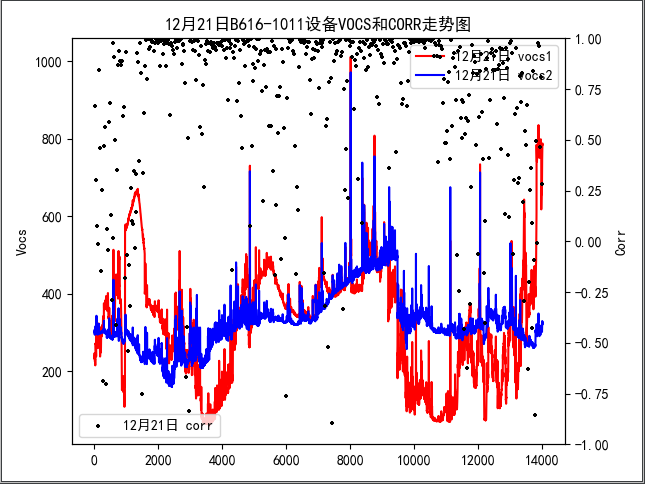
## 2.2 设备相关系数走势图（相邻30个数据为一组）

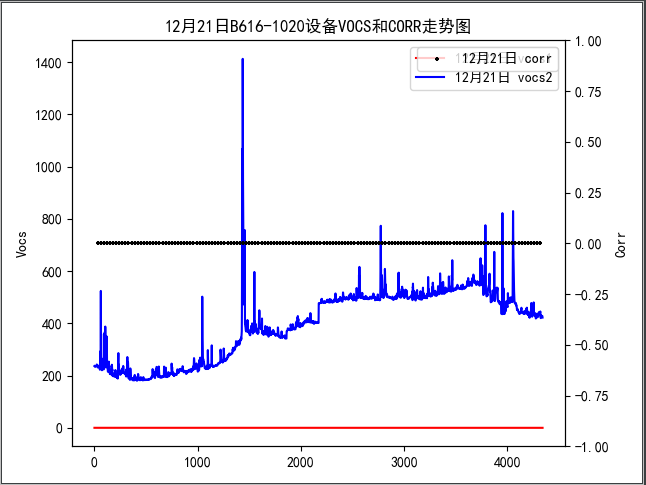
### 2.2.1正常设备（不考虑时间断层）

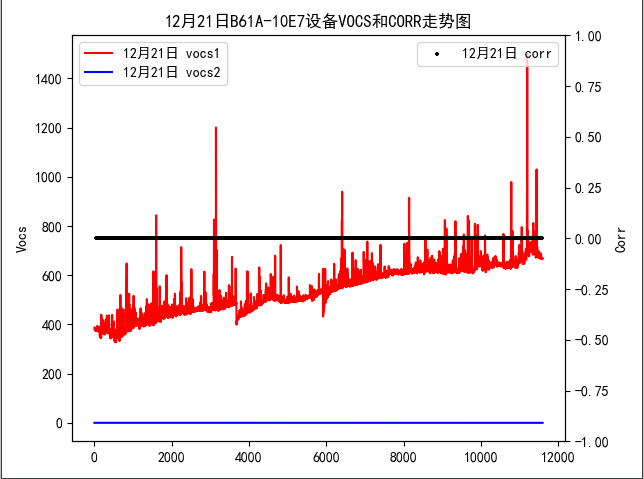


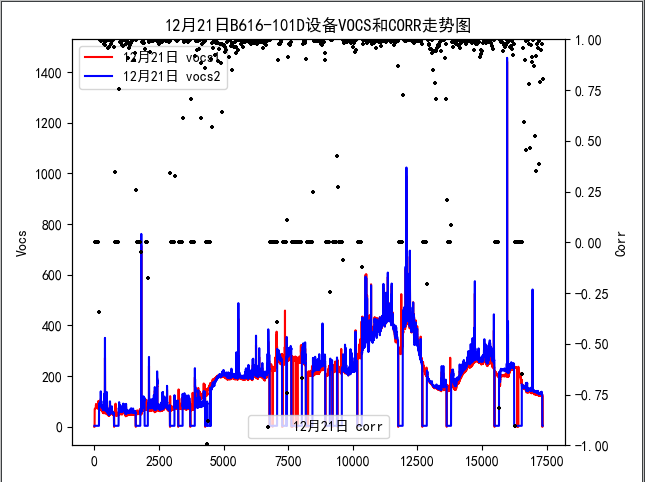


### 2.2.2异常设备（不考虑时间断层）



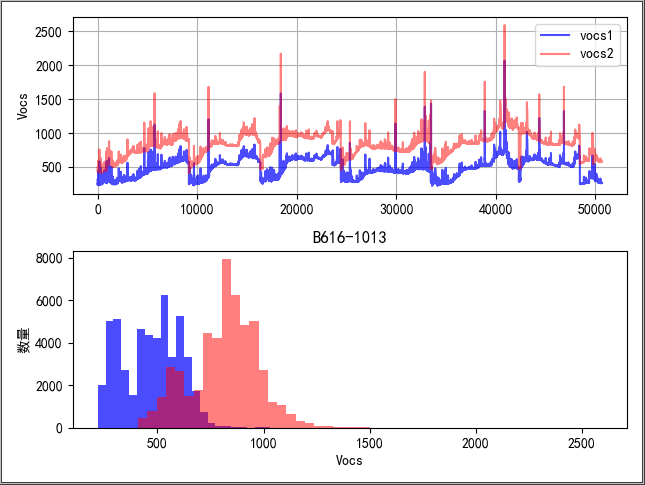


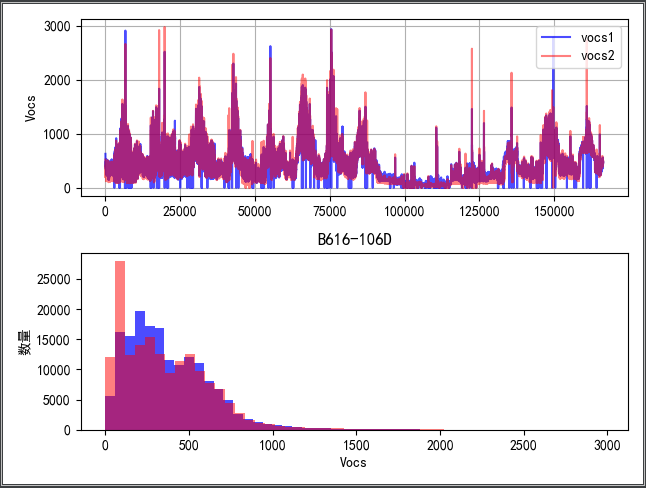




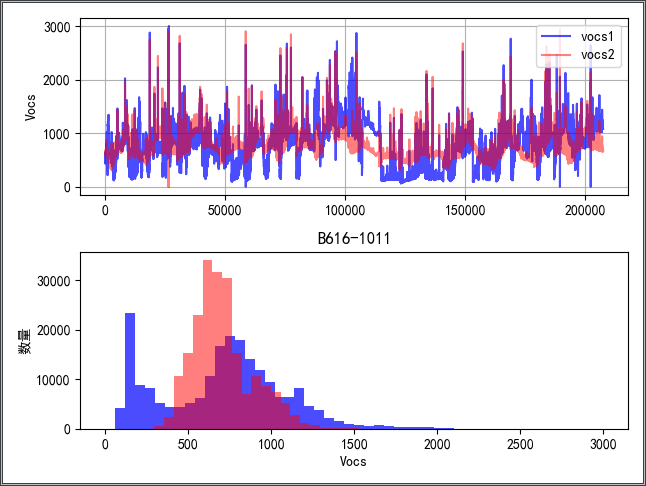
## 2.3 设备分布直方图

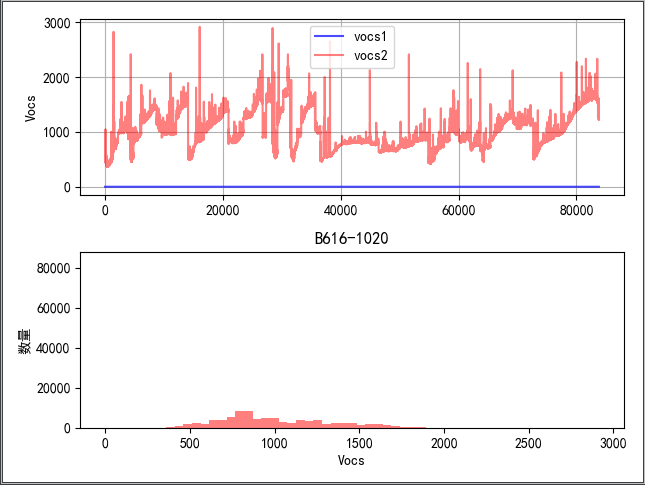
### 2.3.1正常设备（不考虑时间断层）

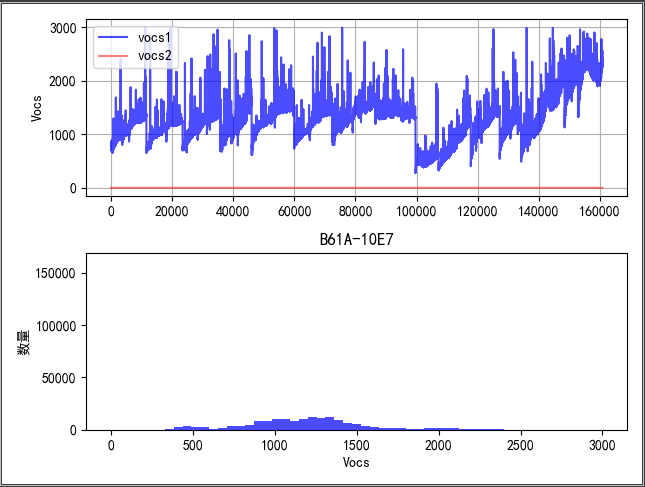


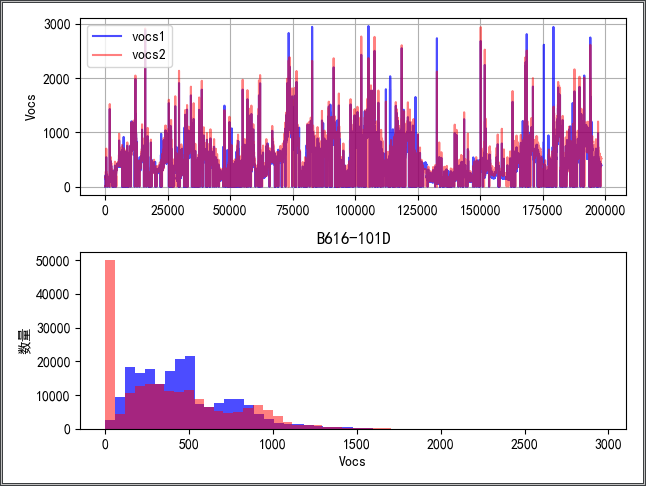


### 2.3.2异常设备（不考虑时间断层）









# 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

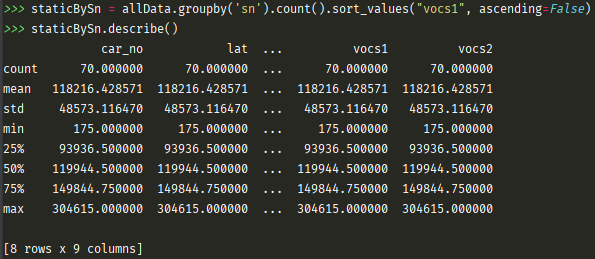
0.3≤|corr|＜0.5表示弱相关

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）



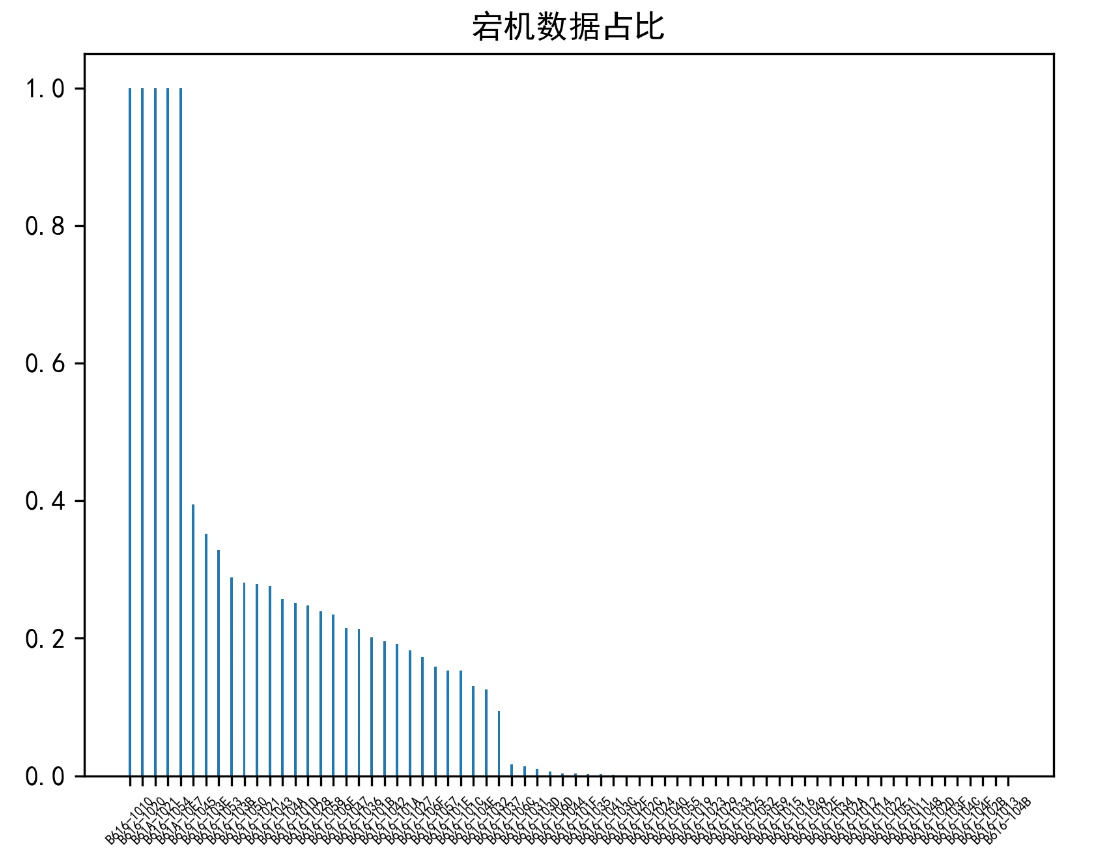
## 3.1 正常设备检测

### 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。

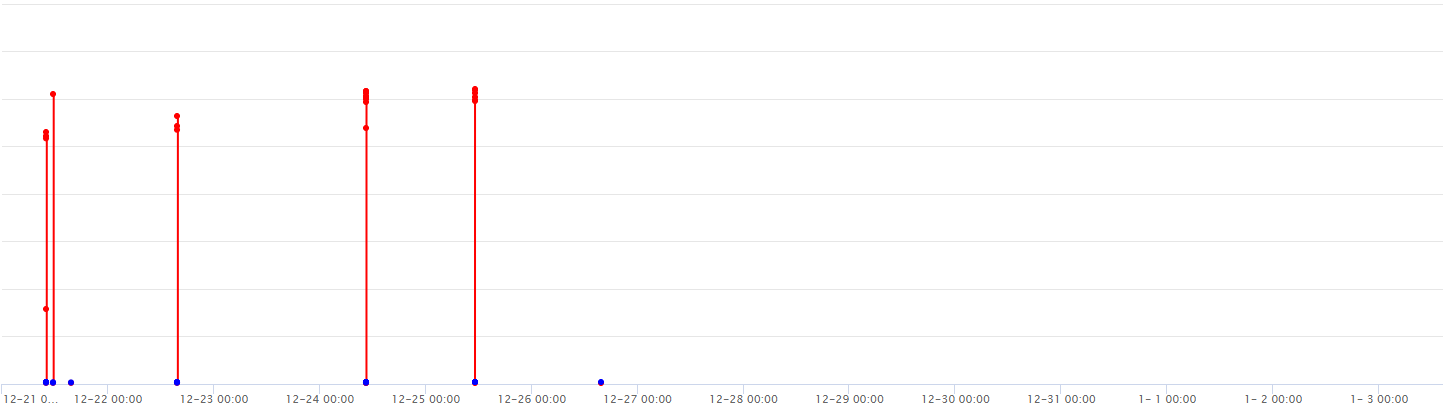
（一）剔除100%宕机设备数据



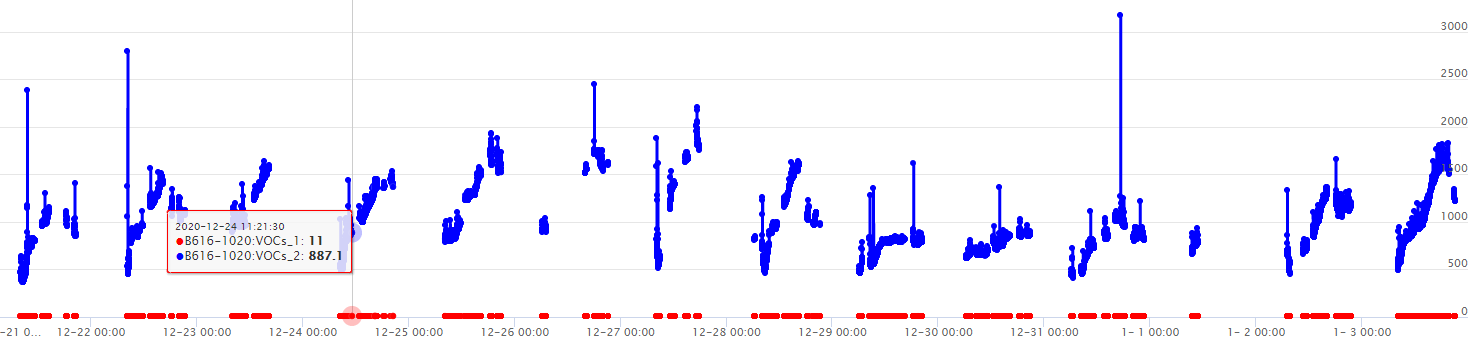
宕机数据计算逻辑：至少有一个核的数据≤11，则认为宕机。

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：

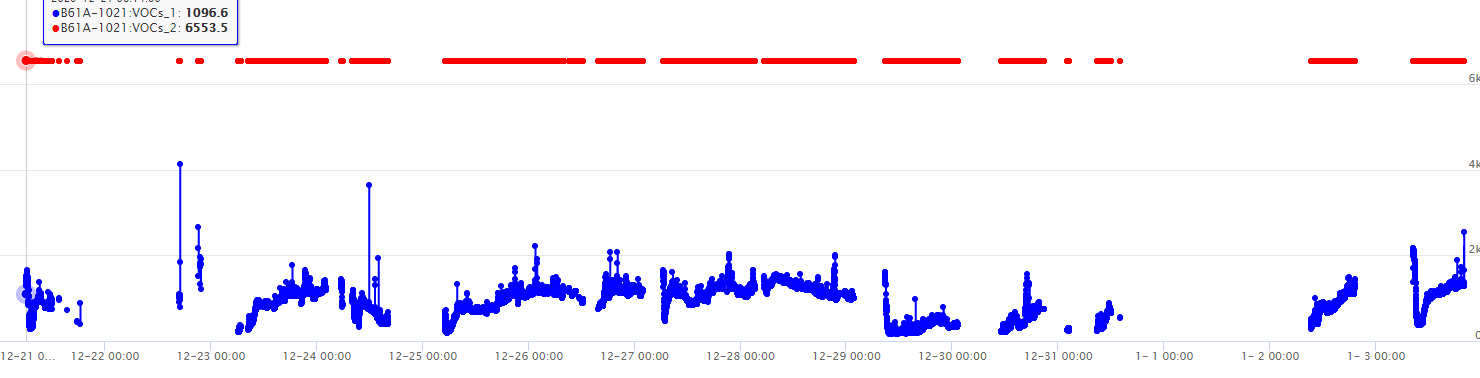
B616-1010



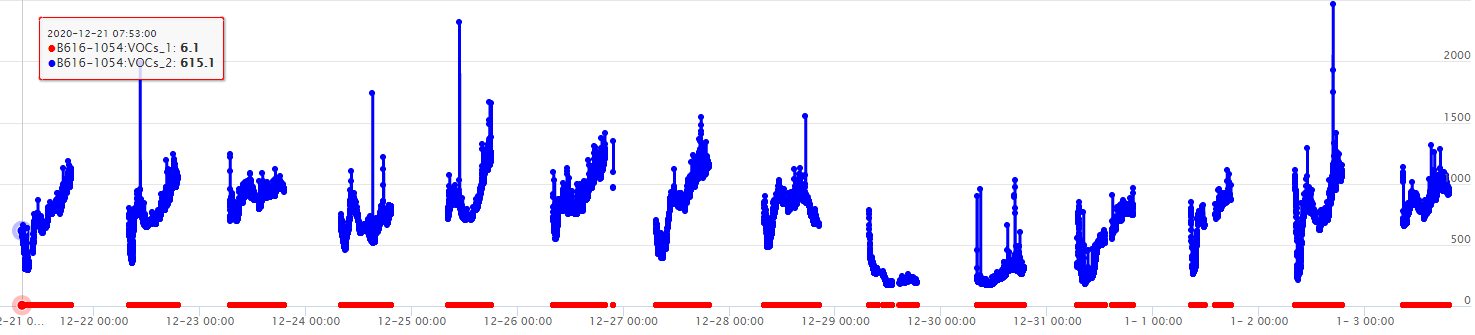
B616-1020



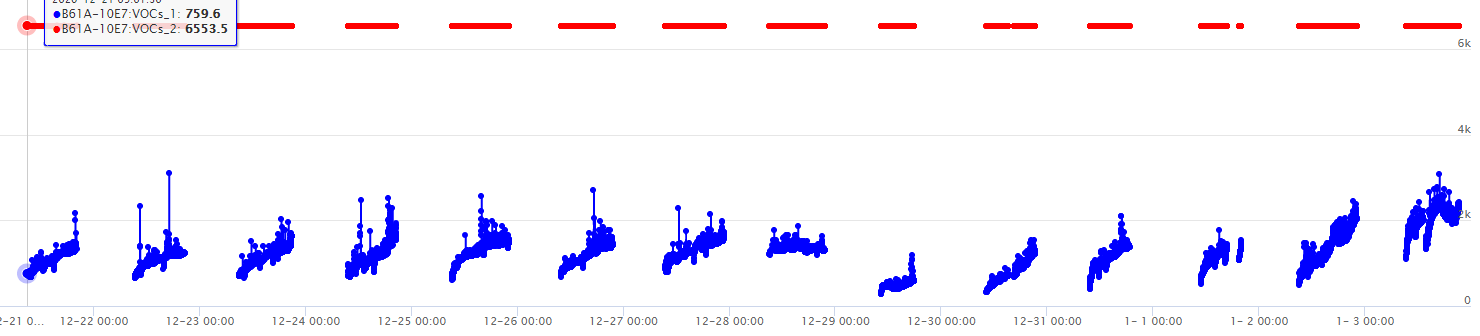
B61A-1021



B616-1054



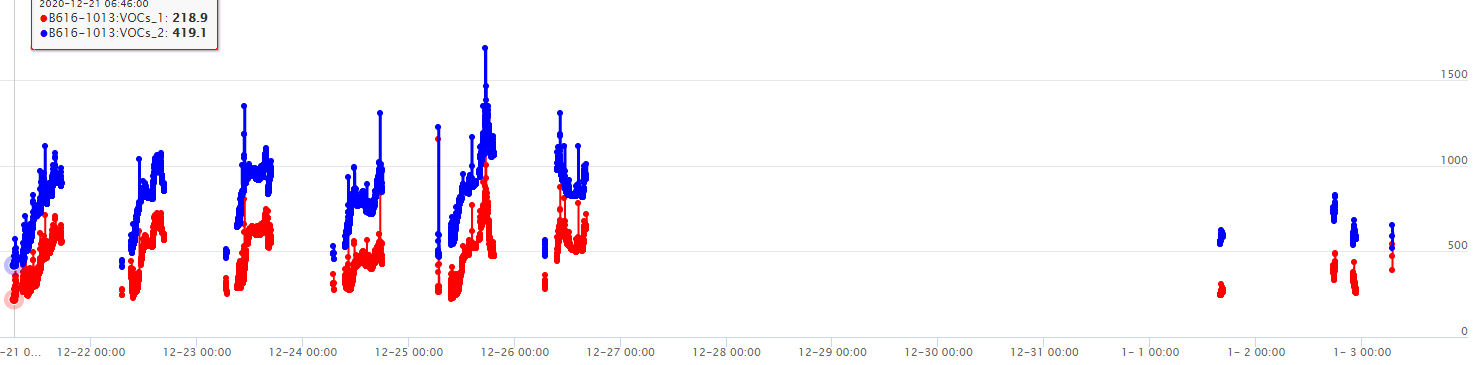
B61A-10E7



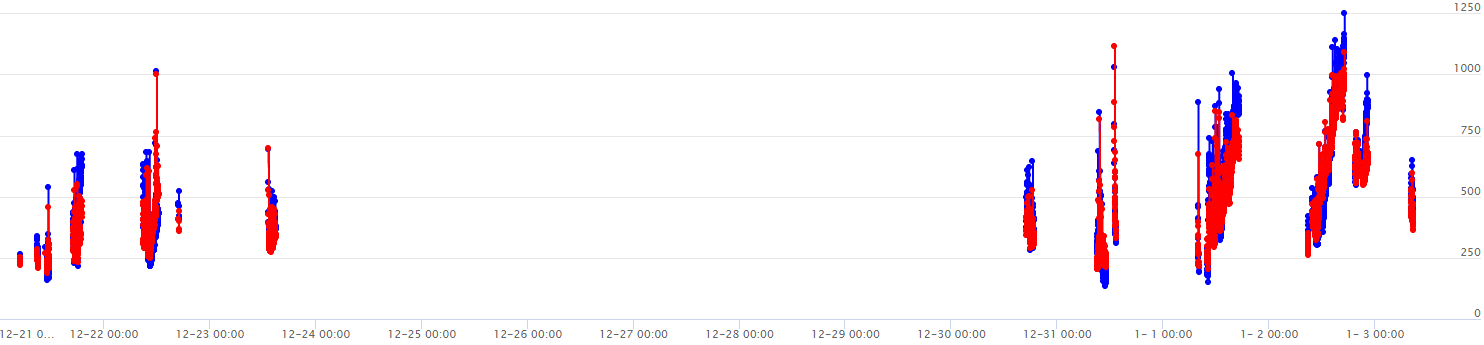
占70台设备的5/70=7%。

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



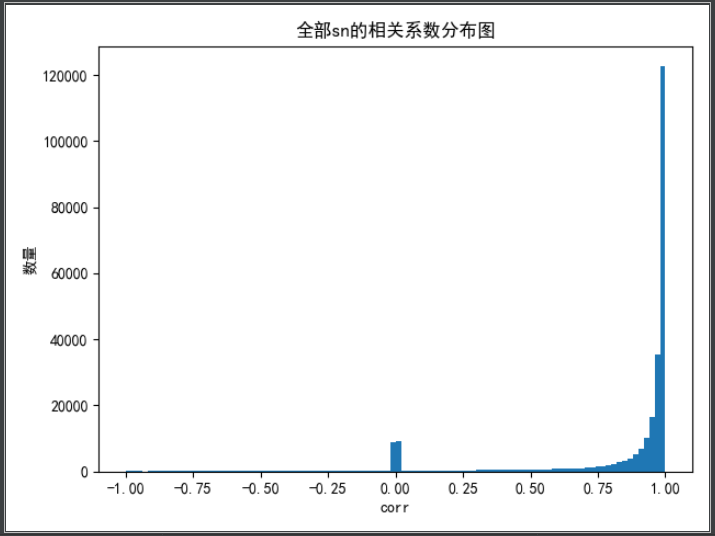
（二）通过相关系数corr判断设备的正常指数

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。



下表为12-21至01-03的相关系数统计结果：

其中，corr的是在不考虑数据存在时间断层的情况下，根据相邻30条（90秒）的双核数据计算得出。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | sn | corr≥0.8比例 | corr=0比例 | corr均值 | corr数量 |
| 1 | B616-104B | 98.34% | 0.00% | 97.54% | 1023 |
| 2 | B616-103D | 98.13% | 0.00% | 97.28% | 160 |
| 3 | B616-1022 | 96.17% | 0.00% | 95.81% | 4522 |
| 4 | B616-1025 | 94.60% | 0.00% | 95.15% | 3168 |
| 5 | B616-102B | 94.78% | 0.00% | 94.99% | 3948 |
| 6 | B616-1031 | 95.11% | 0.00% | 94.94% | 3638 |
| 7 | B616-1024 | 94.74% | 0.00% | 94.64% | 3367 |
| 8 | B616-102A | 93.03% | 0.00% | 94.14% | 3739 |
| 9 | B616-106D | 94.64% | 0.00% | 93.62% | 5545 |
| 10 | B616-1044 | 92.78% | 0.00% | 93.52% | 5234 |
| 11 | B616-1033 | 92.42% | 0.00% | 93.47% | 2586 |
| 12 | B616-1055 | 92.45% | 0.00% | 93.45% | 5124 |
| 13 | B616-1029 | 93.14% | 0.00% | 93.36% | 5041 |
| 14 | B616-1023 | 92.06% | 0.00% | 92.94% | 4748 |
| 15 | B616-1041 | 91.42% | 0.00% | 92.29% | 3148 |
| 16 | B616-1049 | 91.25% | 0.00% | 92.20% | 2801 |
| 17 | B616-1048 | 90.98% | 0.00% | 91.95% | 3571 |
| 18 | B616-1034 | 89.89% | 0.00% | 91.57% | 4965 |
| 19 | B616-103F | 88.57% | 0.00% | 90.97% | 5075 |
| 20 | B616-104C | 88.68% | 0.00% | 90.46% | 10153 |
| 21 | B616-102E | 88.03% | 0.00% | 90.19% | 4368 |
| 22 | B616-102F | 88.92% | 0.00% | 90.11% | 4099 |
| 23 | B616-102C | 87.47% | 0.00% | 89.62% | 4094 |
| 24 | B616-1013 | 84.83% | 0.00% | 89.36% | 1688 |
| 25 | B616-1012 | 87.59% | 0.00% | 89.10% | 3224 |
| 26 | B616-102D | 86.19% | 0.00% | 88.99% | 4344 |
| 27 | B616-1052 | 85.97% | 0.00% | 88.83% | 2887 |
| 28 | B616-104F | 86.51% | 0.00% | 88.20% | 3313 |
| 29 | B616-1051 | 83.88% | 0.00% | 87.68% | 3928 |
| 30 | B616-1035 | 87.12% | 0.03% | 87.62% | 3913 |
| 31 | B616-1014 | 84.82% | 0.00% | 86.89% | 4228 |
| 32 | B616-1040 | 82.47% | 0.00% | 86.23% | 4222 |
| 33 | B616-1037 | 86.09% | 0.08% | 86.09% | 5004 |
| 34 | B616-1019 | 83.76% | 0.00% | 85.64% | 2636 |
| 35 | B616-106C | 83.79% | 0.00% | 85.60% | 3195 |
| 36 | B616-1016 | 82.01% | 0.00% | 84.94% | 5419 |
| 37 | B616-101F | 79.78% | 0.00% | 84.37% | 3457 |
| 38 | B616-1015 | 81.85% | 0.00% | 84.15% | 4722 |
| 39 | B616-1059 | 80.04% | 0.00% | 82.93% | 2821 |
| 40 | B616-1057 | 80.36% | 0.16% | 79.93% | 5061 |
| 41 | B616-1032 | 76.82% | 0.17% | 79.20% | 5452 |
| 42 | B616-101C | 78.08% | 0.17% | 78.65% | 4771 |
| 43 | B616-101E | 76.71% | 0.31% | 78.31% | 3868 |
| 44 | B616-1027 | 77.03% | 0.23% | 77.02% | 3012 |
| 45 | B616-101B | 74.49% | 0.33% | 75.05% | 5137 |
| 46 | B616-101A | 72.87% | 0.24% | 73.91% | 5363 |
| 47 | B616-1042 | 73.37% | 0.42% | 73.65% | 2122 |
| 48 | B616-1047 | 71.33% | 0.14% | 72.47% | 3666 |
| 49 | B616-1058 | 71.22% | 0.36% | 71.88% | 5587 |
| 50 | B616-1036 | 71.11% | 0.23% | 71.83% | 6109 |
| 51 | B616-104E | 74.53% | 0.26% | 70.62% | 3125 |
| 52 | B616-1011 | 63.06% | 0.00% | 70.39% | 6914 |
| 53 | B616-106F | 67.24% | 0.30% | 70.26% | 4048 |
| 54 | B616-101D | 69.46% | 0.36% | 69.84% | 6611 |
| 55 | B616-1028 | 66.60% | 0.35% | 68.46% | 4848 |
| 56 | B616-1021 | 68.49% | 0.41% | 68.28% | 3703 |
| 57 | B616-103B | 66.99% | 0.40% | 67.07% | 3254 |
| 58 | B616-1043 | 65.48% | 0.37% | 66.78% | 5133 |
| 59 | B616-104A | 64.68% | 0.18% | 66.29% | 1628 |
| 60 | B616-1050 | 63.72% | 0.26% | 65.34% | 3062 |
| 61 | B616-106E | 59.27% | 0.16% | 63.02% | 3032 |
| 62 | B616-103C | 48.18% | 0.00% | 62.99% | 4159 |
| 63 | B616-1053 | 58.81% | 0.54% | 60.86% | 4101 |
| 64 | B616-103E | 58.86% | 0.28% | 59.52% | 722 |
| 65 | B616-1045 | 53.69% | 0.25% | 56.36% | 393 |

说明：

corr≥0.8比例：表示corr≥0.8的数据占总数据的比例

corr=0比例：表示corr=0的数据占总数据的比例（这种情况很大程度说明存在宕机，小概率说明双核数据无关）

corr均值：表示所有corr的总和/corr的数量

corr数量：表示每30个相邻数据计算得出的相关系数

B616-1059

B616-1057的图形

通过上表和图，得出结论，在最近2周的历史数据尺度下：

corr均值 ≥ 80%

corr≥0.8比例 ≥ 80%

corr=0比例 ＜ 0.1%

满足以上3个条件, 可以认为设备正常。否则，存在异常。

### 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