**VOCS数据分析报告**

# 1 需求

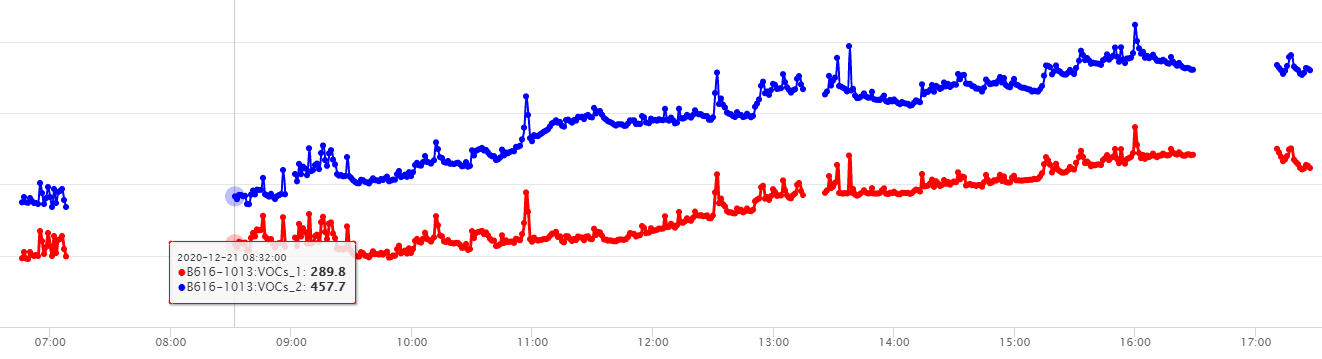
检测Vocs异常数据， 发现真实污染数值。

# 数据特征

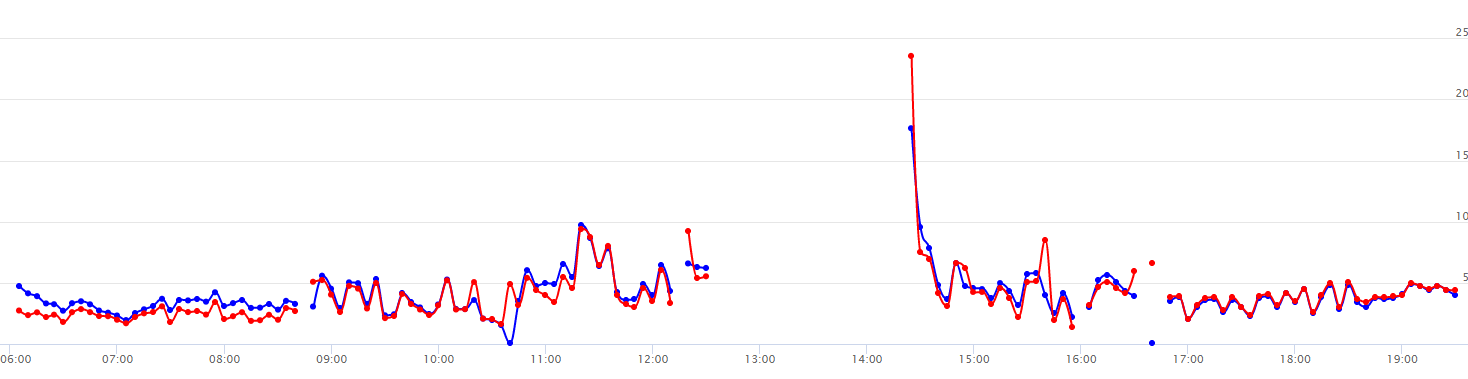
## 设备时序图

### 2.1.1正常设备

**B616-1013 12-21日数据**



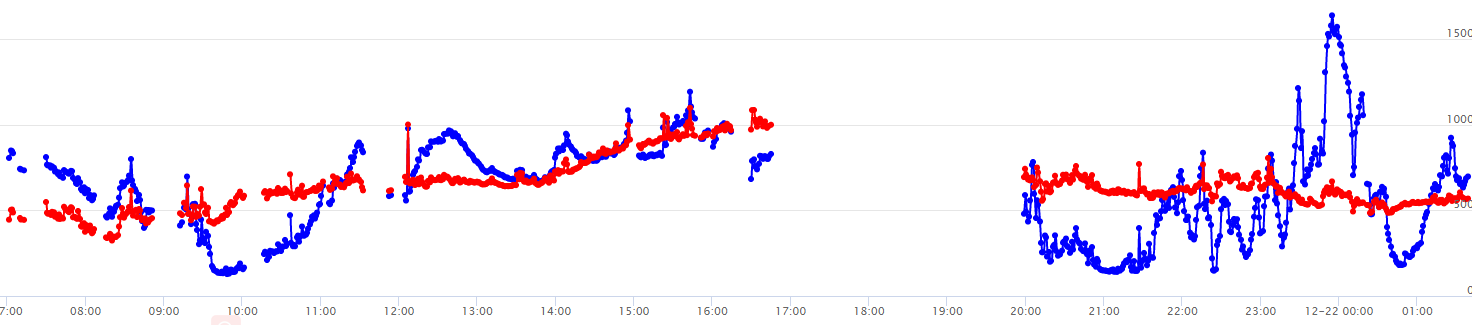
**B616-106D 12-21日数据**



双核设备走势趋同

### 2.1.2异常设备

**B616-1011 12-21日数据**



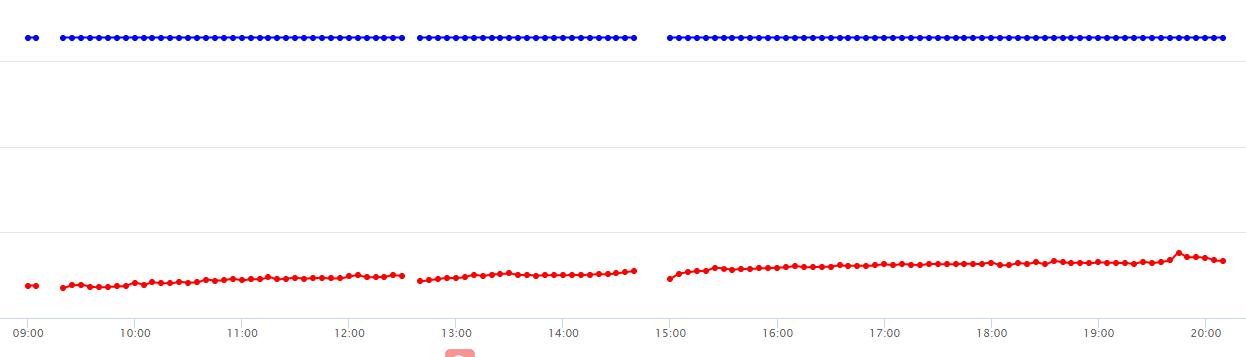
双核设备走势不同，交叉点很多

**B616-1020 12-21日数据**



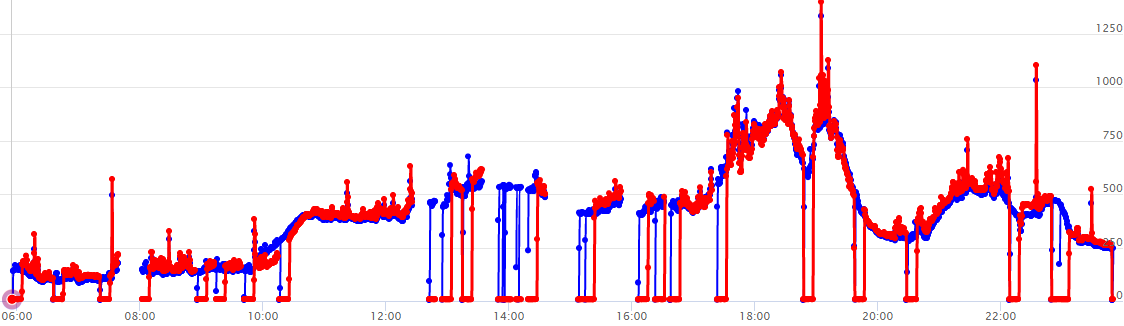
单核vocs\_2宕机，数据保持11不变

**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)

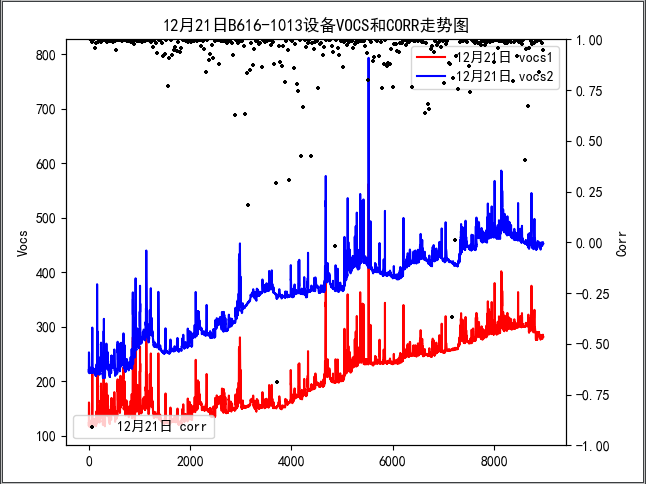
**B616-101D** 12-21日数据

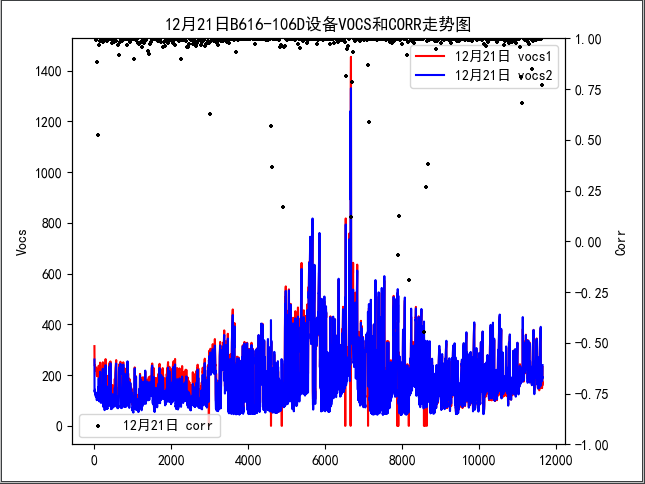


Vocs设备间歇性宕机

## 2.2 设备相关系数走势图（相邻30个数据为一组）

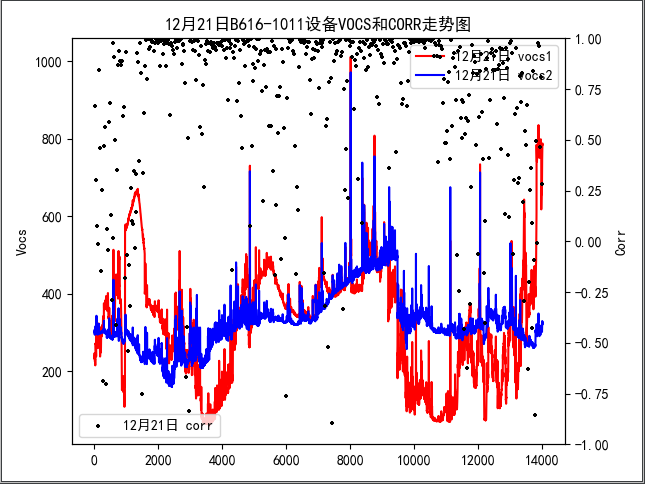
### 2.2.1正常设备（不考虑时间断层）

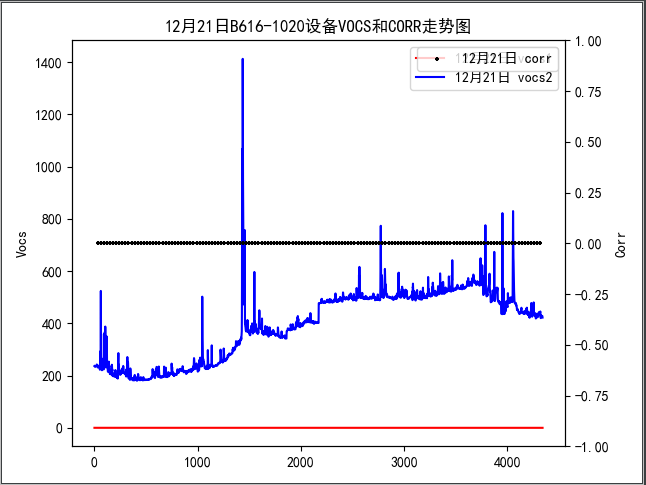


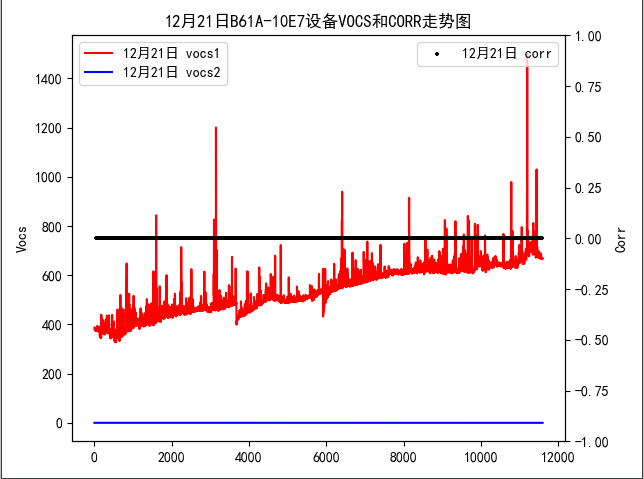


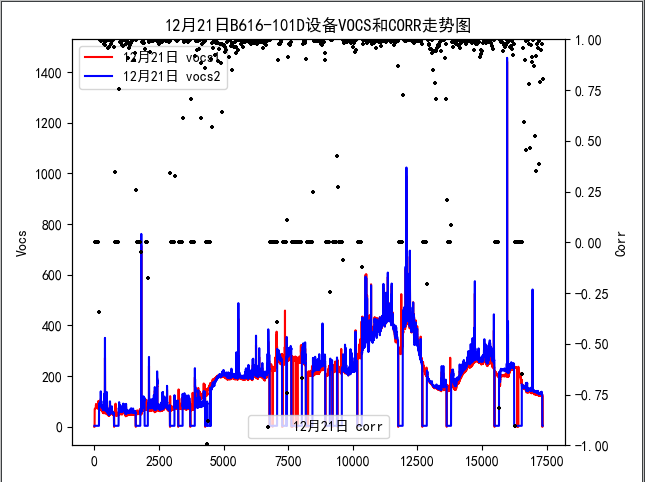
可以看出, 正常设备的双核相关系数大部分在1附近, 表示显著正相关.

### 2.2.2异常设备（不考虑时间断层）





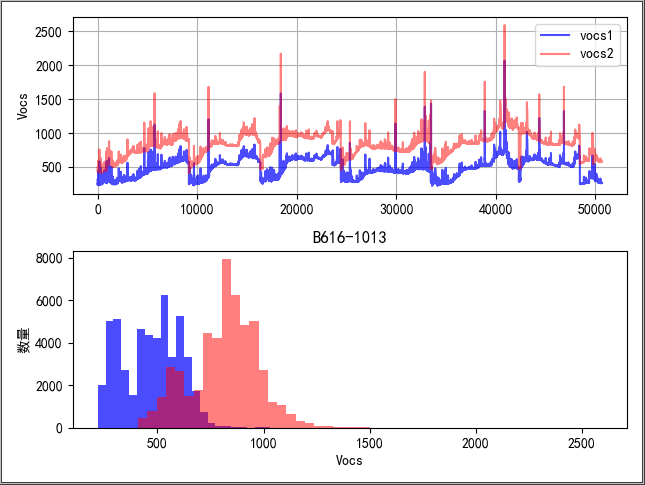


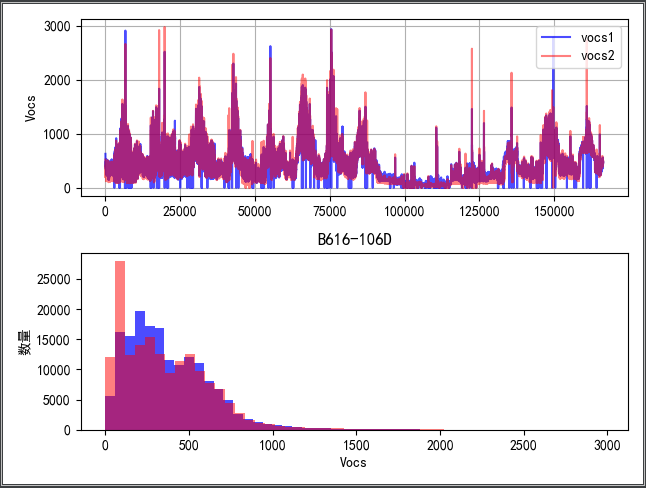


异常设备相关系数比较分散, 而且双核中存在宕机的机器时, 相关系数表现为0.

## 2.3 设备分布直方图

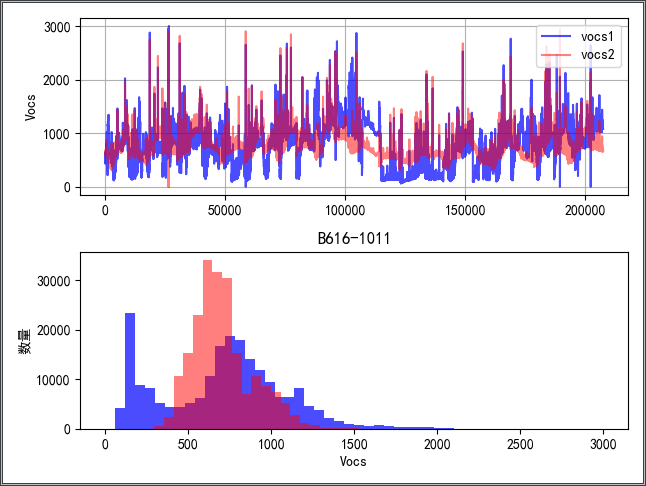
### 2.3.1正常设备（不考虑时间断层）

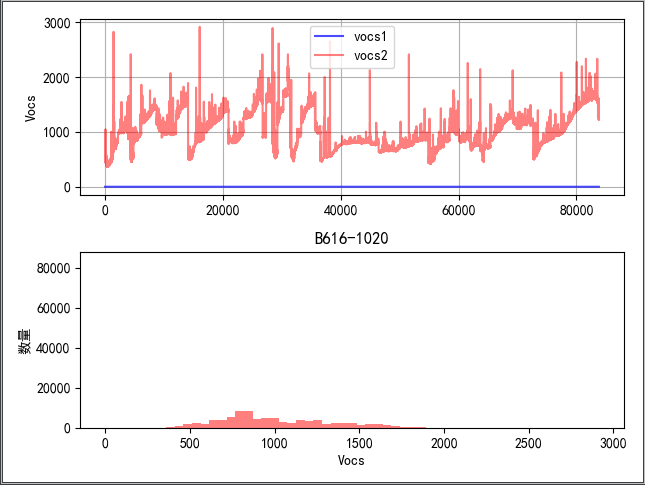


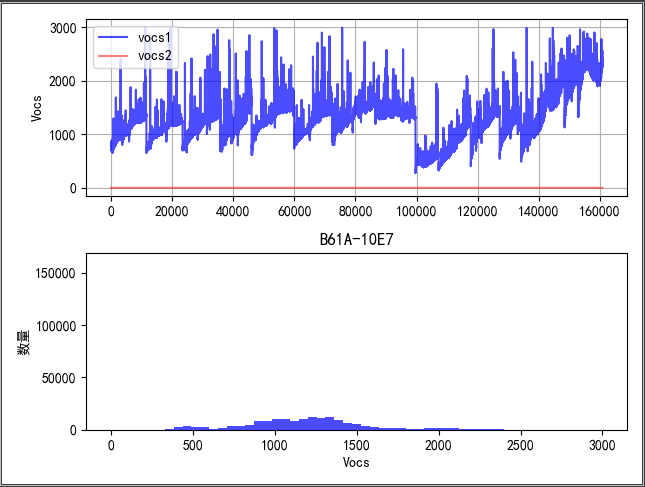


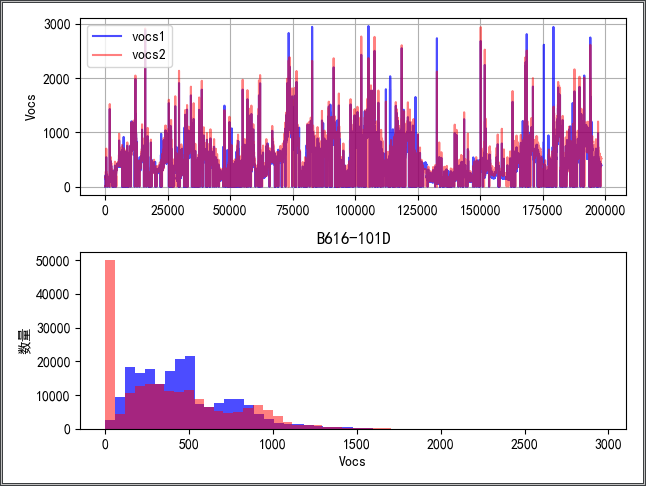
正常设备的分布形状基本相同, 只是存在均值的差别, 表现为平移.

### 2.3.2异常设备（不考虑时间断层）









异常设备的分布图明显可以看出, 双核数据分布情况不同.

# 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：**

相关系数判断标准：

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

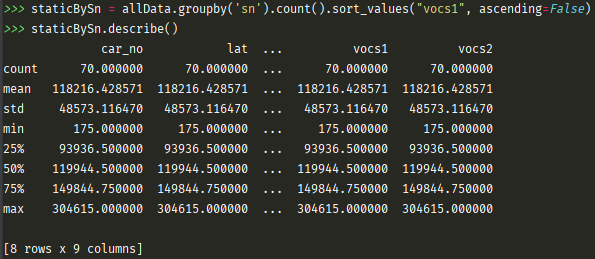
0.5≤|corr|＜0.8表示中度相关

0.8≤|corr|＜0.95表示强相关

|corr≥0.95表示显著相关

本文目前不考虑时间因素，将所有数据连在一起计算相关系数。

（**此处做了一个假设**：假设每个设备都有各自的统计属性，设备每次开机都在该设备所属的统计属性范围值内波动。）



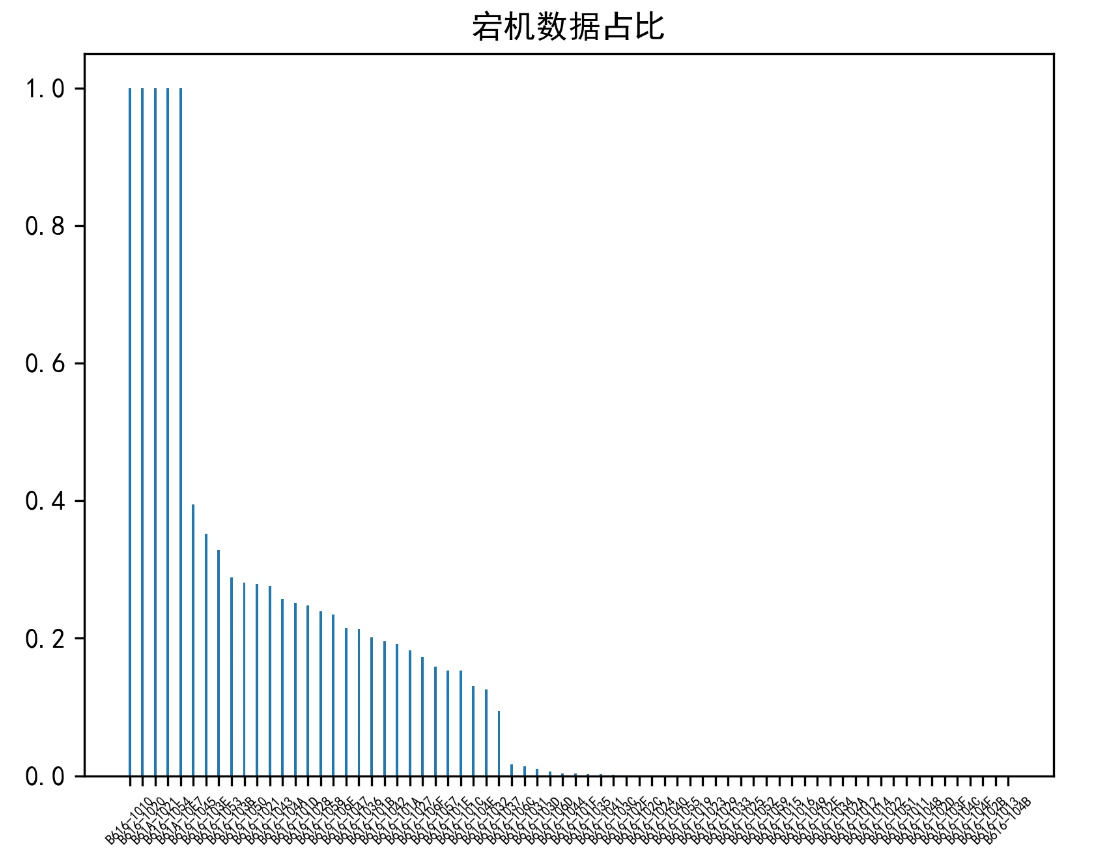
## 3.1 异常设备和数据检测

### 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。

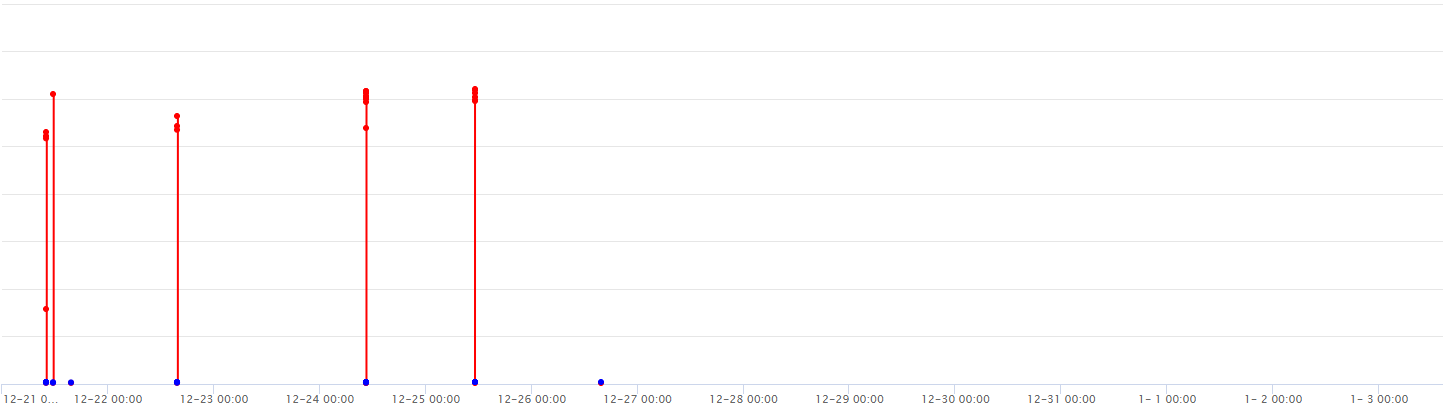
#### （一）剔除100%宕机设备数据



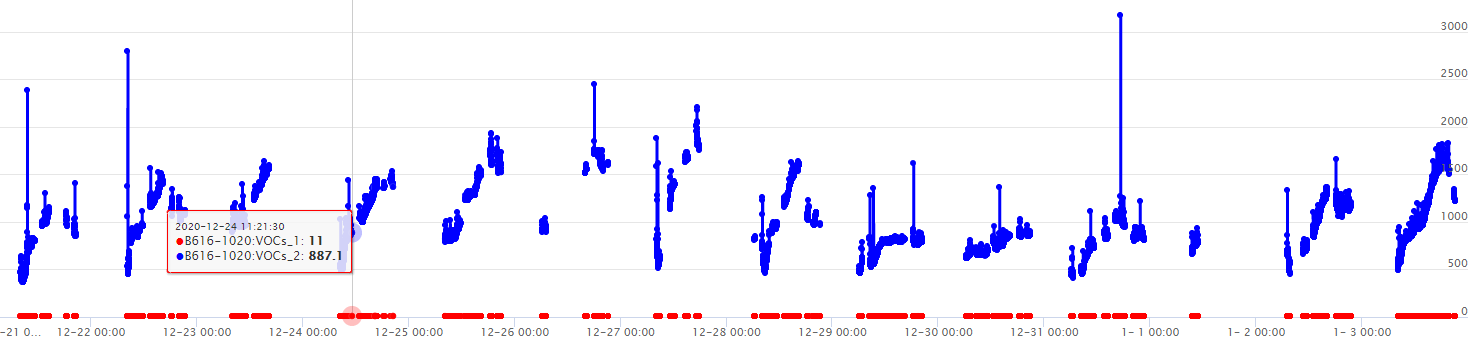
宕机数据计算逻辑：至少有一个核的数据≤11，则认为宕机。

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：

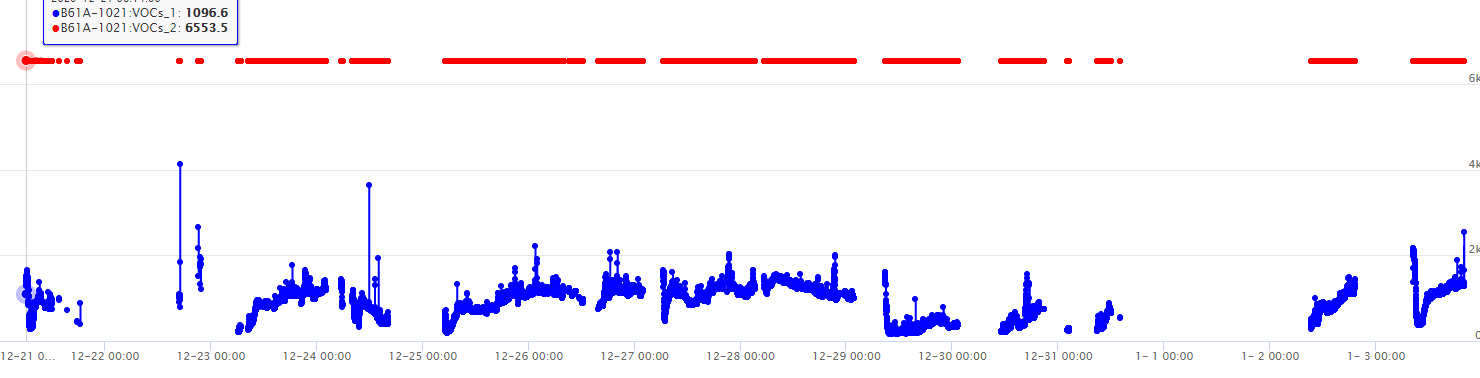
B616-1010



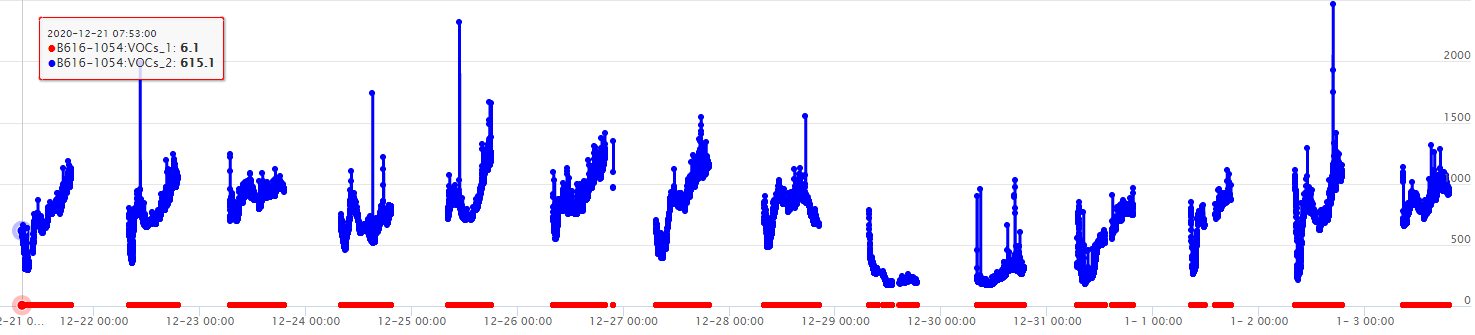
B616-1020



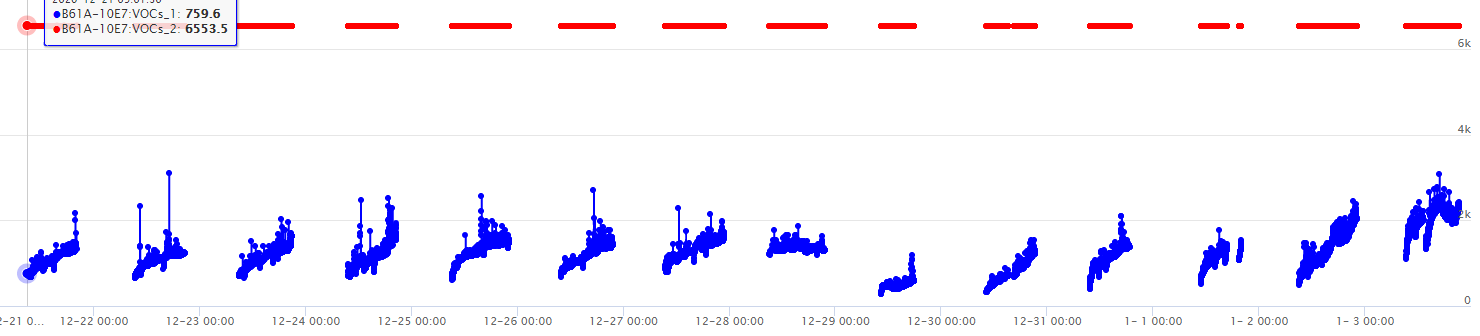
B61A-1021



B616-1054



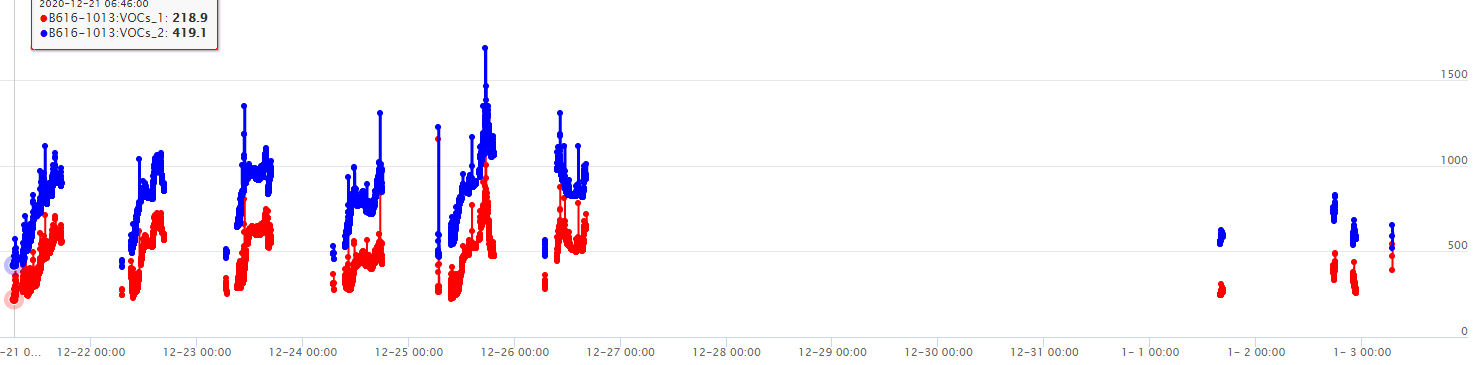
B61A-10E7



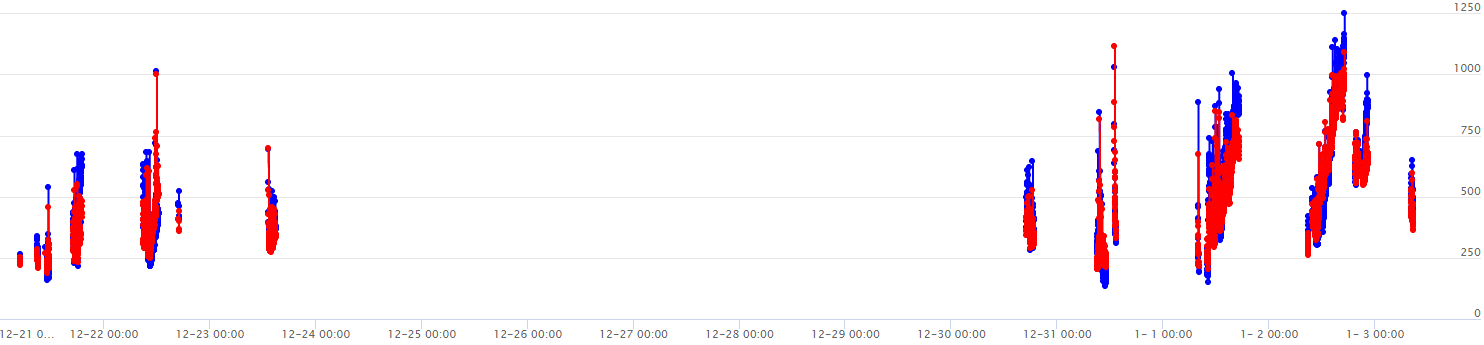
占70台设备的5/70=7%。

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



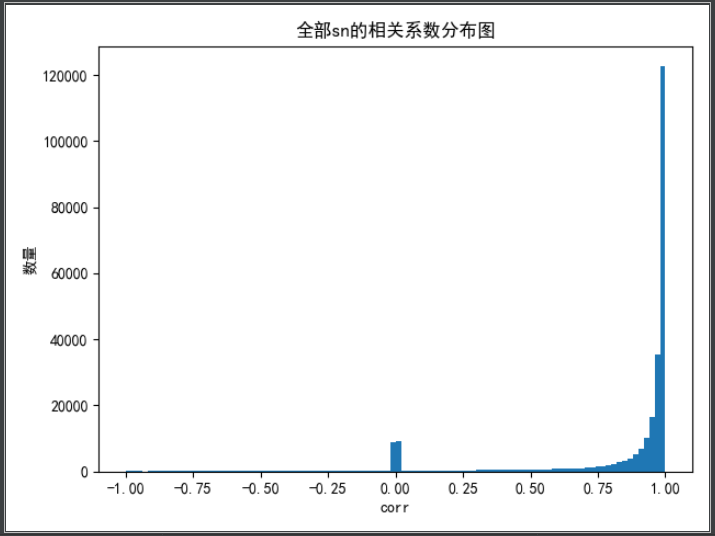
#### （二）通过相关系数corr判断设备的是否正常的标准定义

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：**

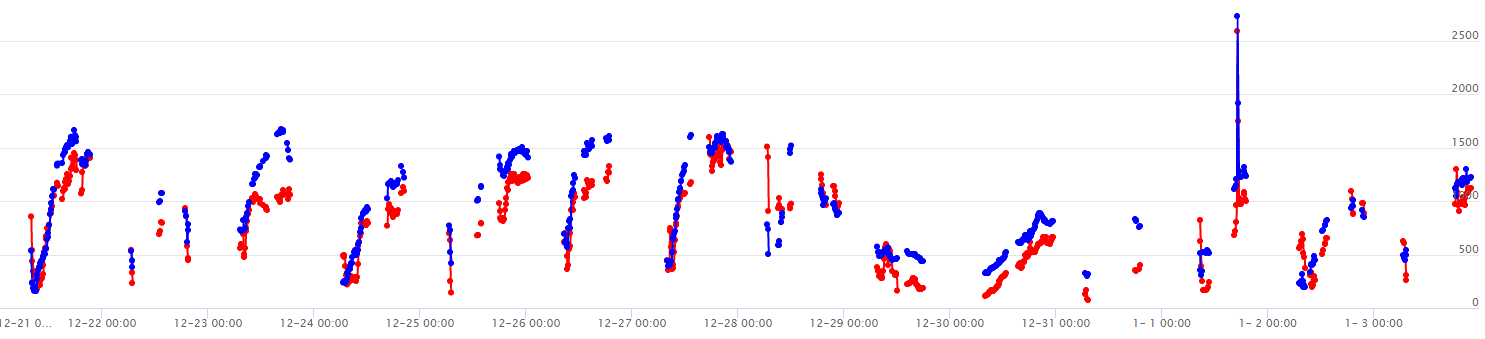
corr≥0.8比例：表示corr≥0.8的数据占总数据的比例

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）

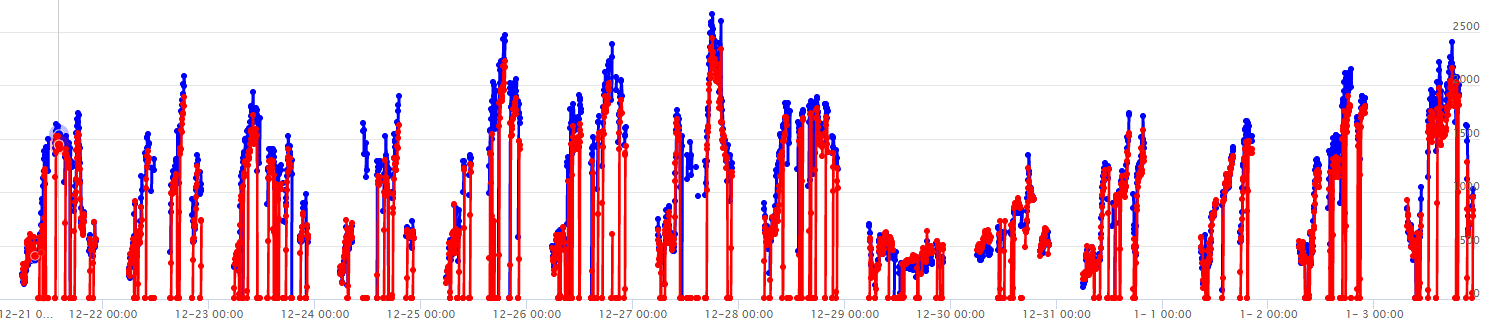
corr均值：表示所有corr的总和/corr的数量

corr数量：表示每30个相邻数据计算得出的相关系数

B616-1059（表中39位）



B616-1057（表中40位）



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%**

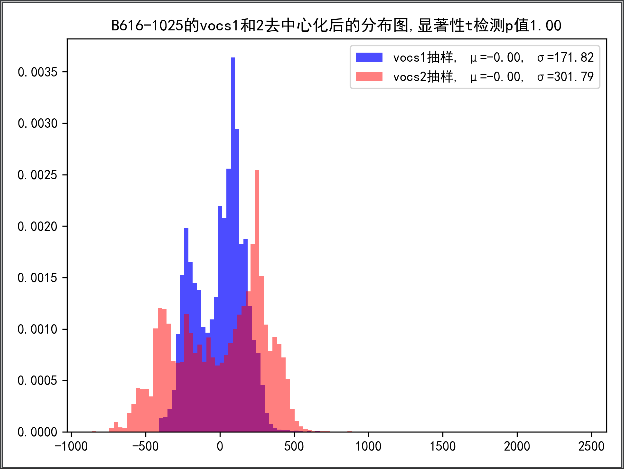
**corr≥0.8比例 ≥ 80%**

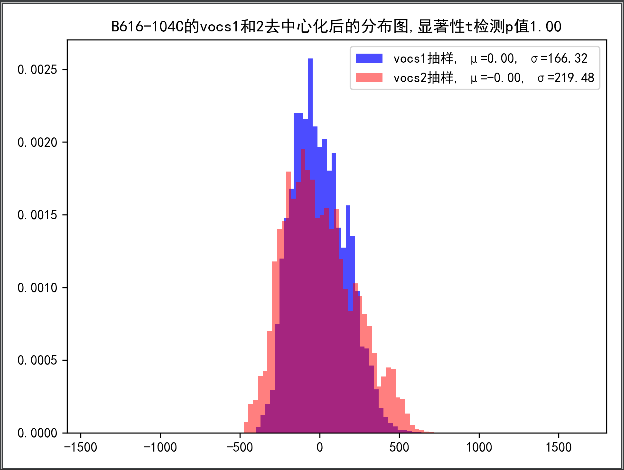
**corr=0比例 ＜ 0.1%**

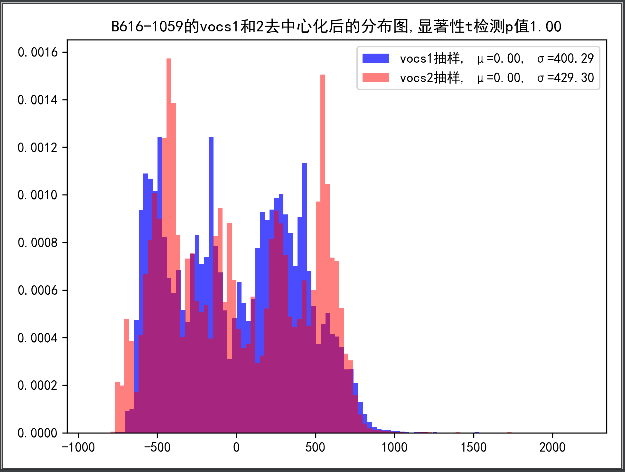
**正常设备的标准定义为：**

同时满足以上3个条件, 可以认为设备正常。否则，存在异常。

#### （三）抽样分析正常设备的统计指标



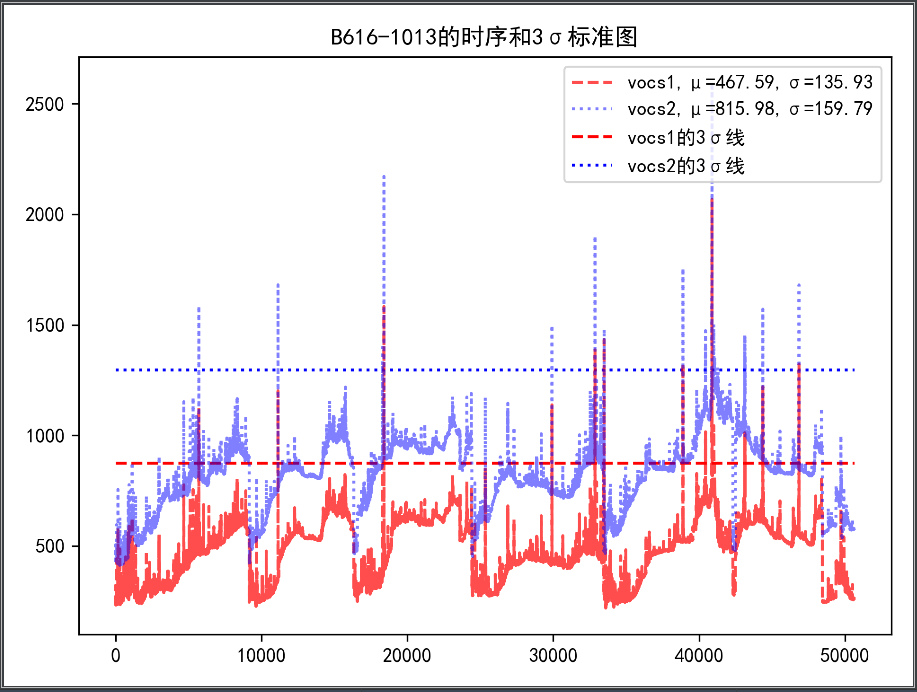


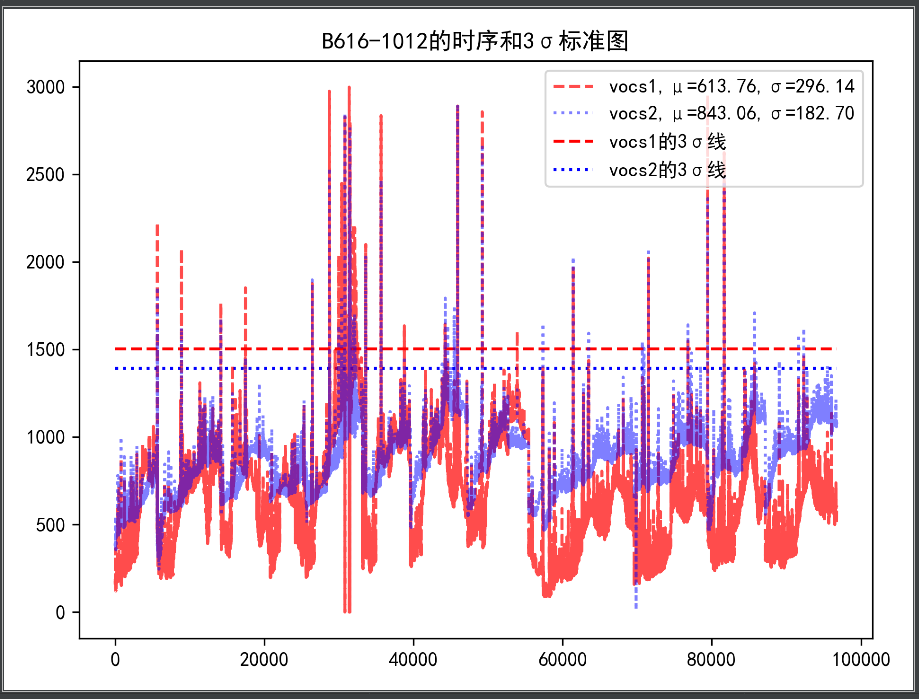


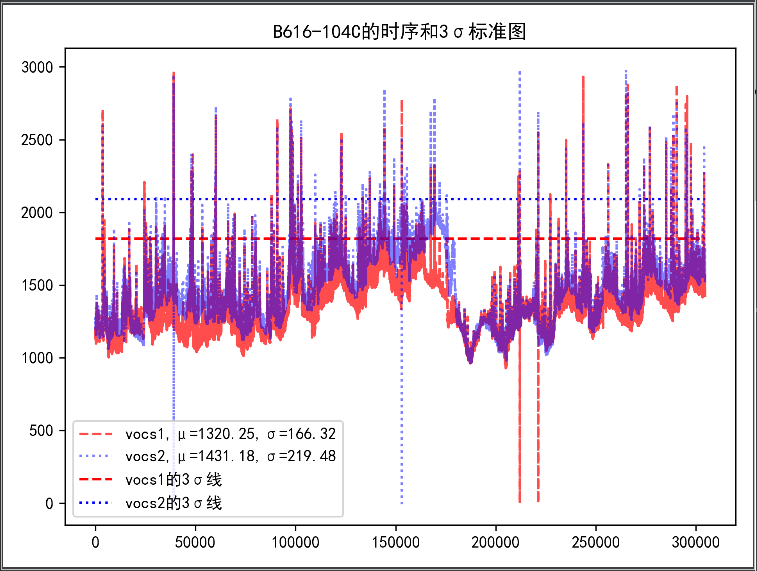
由以上抽检的3台设备的数据分布图可以看出，数据呈现类似正态分布。

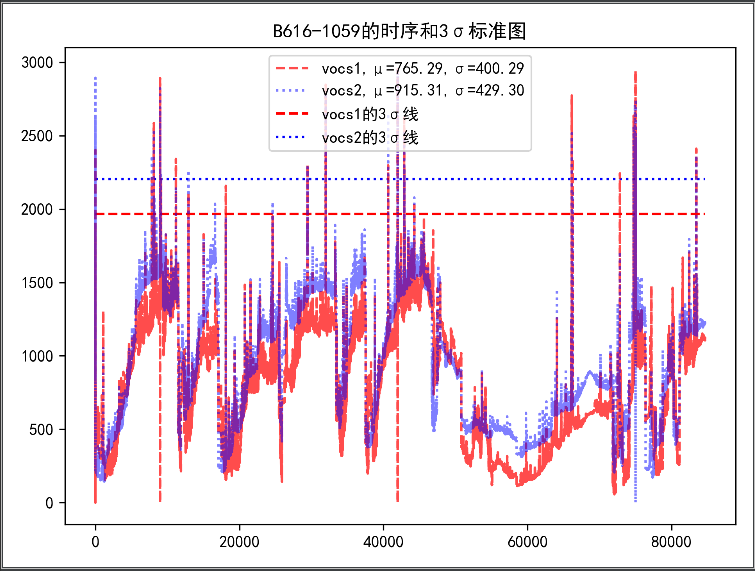
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sn | vocs1均值 | vocs2均值 | vocs1标准差 | vocs2标准差 | 3σ以外数据占比 |
| B616-1012 | 613.76 | 843.06 | 296.14 | 182.70 | 0.47% |
| B616-1013 | 467.59 | 815.98 | 135.93 | 159.79 | 0.30% |
| B616-1014 | 1021.39 | 1082.85 | 392.89 | 300.05 | 0.10% |
| B616-1015 | 887.90 | 1096.06 | 271.25 | 404.59 | 0.09% |
| B616-1016 | 682.17 | 609.00 | 334.69 | 147.56 | 1.16% |
| B616-1019 | 1227.04 | 768.91 | 320.39 | 267.27 | 0.71% |
| B616-1022 | 999.83 | 695.48 | 209.32 | 151.04 | 0.18% |
| B616-1023 | 1328.24 | 703.95 | 381.34 | 267.08 | 0.13% |
| B616-1024 | 990.40 | 1372.62 | 363.34 | 513.27 | 0.10% |
| B616-1025 | 522.25 | 862.49 | 171.82 | 301.79 | 0.23% |
| B616-1029 | 1585.59 | 1171.70 | 432.80 | 339.19 | 0.04% |
| B616-102A | 701.13 | 558.44 | 305.29 | 311.97 | 0.19% |
| B616-102B | 1047.23 | 950.38 | 353.43 | 347.46 | 0.04% |
| B616-102C | 707.01 | 795.00 | 524.20 | 351.50 | 0.27% |
| B616-102D | 872.33 | 1066.14 | 340.32 | 310.21 | 0.07% |
| B616-102E | 756.16 | 899.12 | 184.44 | 317.64 | 0.15% |
| B616-102F | 1585.43 | 1384.52 | 586.21 | 457.39 | 0.04% |
| B616-1031 | 279.07 | 302.97 | 155.66 | 183.87 | 0.34% |
| B616-1033 | 1260.19 | 1196.06 | 406.44 | 393.99 | 0.05% |
| B616-1034 | 1043.51 | 1100.94 | 332.85 | 298.06 | 0.08% |
| B616-103D | 1024.23 | 1257.60 | 288.15 | 264.93 | 0.42% |
| B616-103F | 687.88 | 849.17 | 306.31 | 251.71 | 0.15% |
| B616-1040 | 484.71 | 756.62 | 352.27 | 399.53 | 0.12% |
| B616-1041 | 1147.12 | 364.40 | 454.49 | 169.69 | 0.29% |
| B616-1044 | 641.80 | 633.90 | 205.77 | 256.73 | 0.16% |
| B616-1048 | 932.34 | 1094.90 | 398.36 | 377.59 | 0.05% |
| B616-1049 | 584.00 | 908.33 | 291.00 | 386.92 | 0.06% |
| B616-104B | 523.03 | 488.97 | 230.01 | 194.93 | 0.16% |
| B616-104C | 1320.25 | 1431.18 | 166.32 | 219.48 | 0.24% |
| B616-104F | 516.25 | 847.29 | 189.04 | 188.29 | 0.87% |
| B616-1051 | 553.65 | 942.99 | 217.31 | 407.04 | 0.16% |
| B616-1052 | 1016.63 | 660.67 | 450.25 | 352.47 | 0.10% |
| B616-1055 | 764.99 | 857.36 | 287.00 | 289.60 | 0.11% |
| B616-1059 | 765.29 | 915.31 | 400.29 | 429.30 | 0.08% |
| B616-106C | 398.72 | 387.28 | 246.82 | 253.40 | 0.08% |
| B616-106D | 376.07 | 358.14 | 253.09 | 272.90 | 1.36% |

可以看出这36个设备的均值和标准差都不相同，但3σ以外数据占比几乎都低于0.03%。









以上是3σ标准线, 所处数据中的位置, 可以看出它所标记出的异常值.

#### （四）持久化离线数据

将各设备的μ和3σ数据持久化存储到数据库，并实现离线日更新。为数据校准和实时异常值判定提供数据基础。

### 3.1.2 实时计算

根据离线计算的判断规则，在不影响实时性的前提下，最大化的记录历史数据。通过记录的历史数据，根据判断标准，判断设备是否正常。

实时计算场景下，针对正常设备的异常数据检测，采用2步：

1. 计算当前时间窗口数据相关系数是否≥0.8；
2. 满足1的条件下，判断当前双核数据是否超出μ+3σ（读取3.1.1中存储的数据）。如果未超出，则数据正常；否则，**数据异常**。

代码逻辑还未写。。。

# 综述

本文做了一个假设：假设每个设备都有各自的统计属性，设备每次开机都在该设备所属的统计属性范围值内波动。

本文论证了正常设备的离线判断标准（3条）, 以及正常设备的异常值检测标准（3σ）。

但对于实时计算依然存在问题需要论证：

1. 离线计算中，正常设备的判断标准是具有时间滞后性的，需要论证新的观点；
2. 异常数据的检测依赖离线计算的σ值，实时情况是否延续历史表现有待验证。

本文只阐述了一种异常检测方案，即3σ，与之同效的还有z-score方法（将数据先标准化，然后根据正态分布判断异常值）。但两者都需要大量的数据样本做支撑。

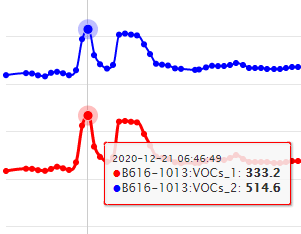
为了更好的满足实时计算条件，需要论证其他的方法，比如，LOF、PCA、孤立森林、k距离局部异常检测等异常值检测方法。这类方法的特征是根据少量样本判断异常值。但他们本身也存在缺陷，此处还需要论证，以求更好的效果。

# 系统问题

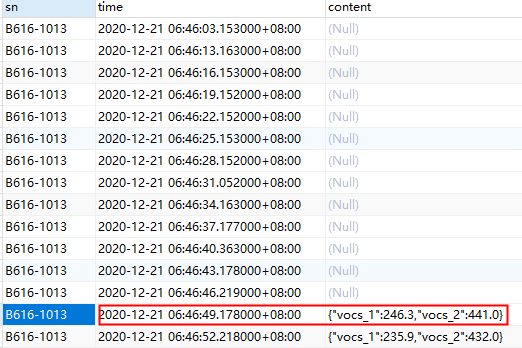
192.168.1.173系统中目前发现2个问题：

1. vocs有些为0的数据，显示为6553.5

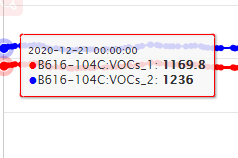
2. sqlite数据库文件中的数据与<http://192.168.1.173/wp/#/>所显示数据有出入，例如：



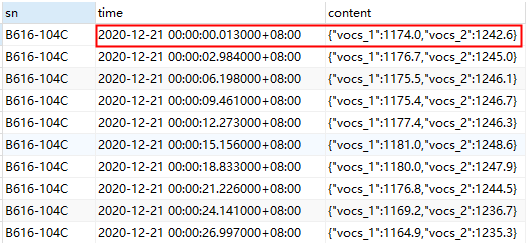
图中可以看出B616-1013设备在06:46:49秒的数据是vocs1：333.2，vocs2：514.6，而且之前也是有数据的。



而数据库中，49秒处的数据不同，而且之前是没有数据的。



B616-104C的21日数据，第一条是1169.8， 1236；



与表中是有出入的。

类似的还有很多。