Data Quality Assurance at the IRIS DMC: Expanding and Improving the MUSTANG System

Gillian Sharer[1], Mary Templeton[1], Laura Keyson[1], and Jerry Carter[1]

[1]IRIS

November 21, 2022

Abstract

The IRIS Data Management Center (DMC) maintains a large assemblage of pre-computed and dynamically generated quality assurance metrics for seismic data by means of its MUSTANG system. Freely accessible through web services at http://service.iris.edu/mustang, this collection of measurements includes basic statistics, data latency, data availability, miniSEED flag counts, Power Spectral Densities (PSD), Probability Density Functions (PDF), and more. The metrics produced are suitable for a wide range of uses such as data selection for research projects, identification of data and metadata problems, assessment of station health and upgrades, and characterization of environmental noise, among others. Currently, MUSTANG measurements for seismic channels span our entire primary data repository from the year 1972 to the present. We are expanding MUSTANG's utility to the broader geosciences community by calculating measurements on our active source, PH5 data repository and by working towards including other types of data such as infrasound. We will also present improvements to metric visualization and quality assessment tools like the MUSTANG Databrowser (basic metric plots and boxplots), MUSTANGular (map-based metric plots), and ISPAQ (a stand-alone Python utility for calculating metrics for miniSEED data stored locally or at any FDSN data center).

SAGE Seismological Facility for the Advancement of Geoscience

Data Quality Assurance at the IRIS DMC: Expanding and Improving the MUSTANG System

Operated by IRIS

Gillian Sharer, **Mary E. Templeton***, Laura Keyson, and Jerry Carter
Incorporated Research Institutions for Seismology (IRIS): Data Services

S21H-0631

As part of the IRIS Data Service's commitment to characterizing and optimizing the quality of data within its archive, we provide over 40 automated data quality metrics to our user community through MUSTANG web services services.iris.edu/mustang/). Metric measurements include amplitude statistics, data latency, completeness, power spectral density-based noise characterization and more. We have recently expanded MUSTANG and related visualization tools to include

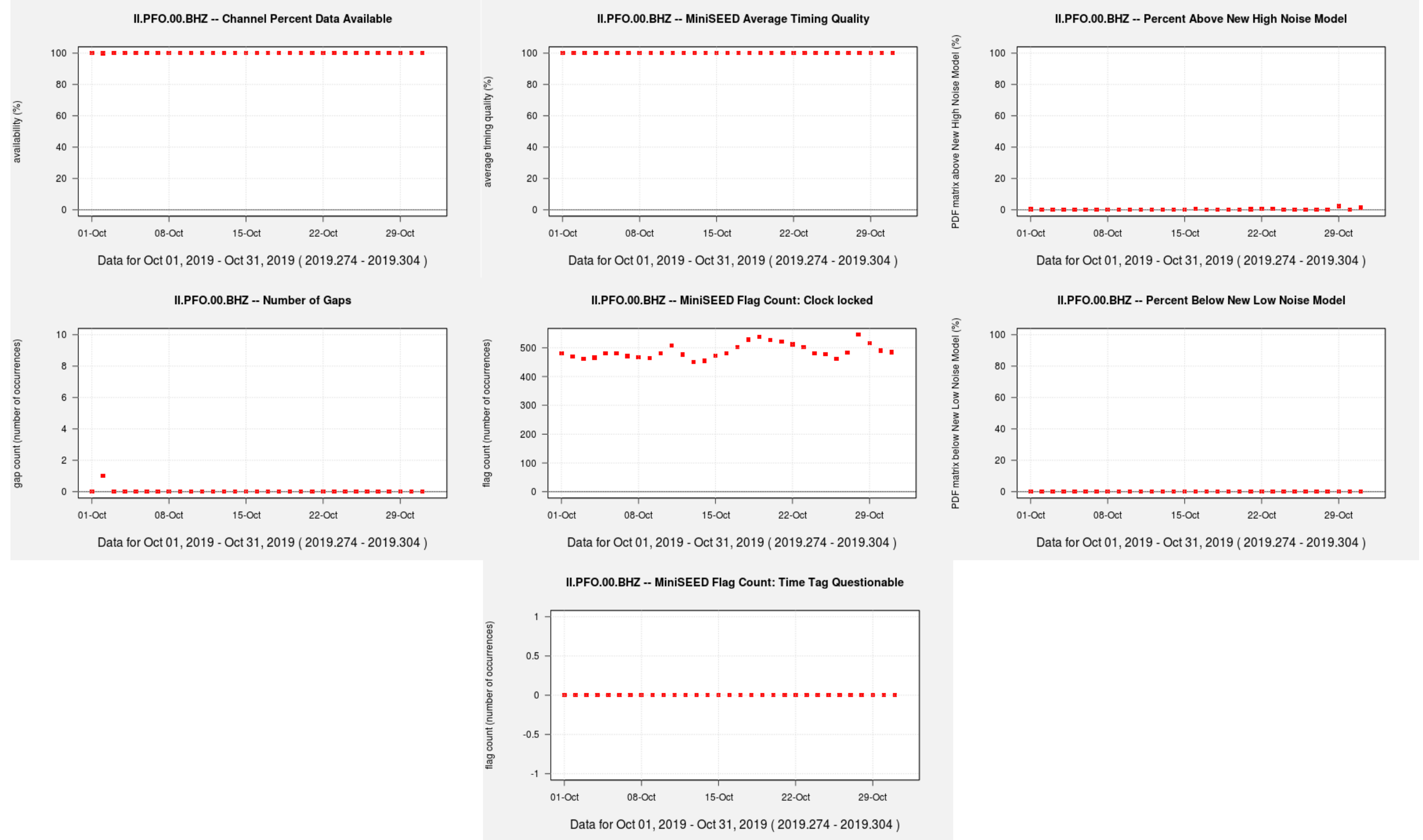
- Gap duration plotting in MUSTANG Databrowser
- QuARG – a utility in development for analyzing, tracking and reporting network data quality issues (available in 2020)
- PDF daily mode spectrograms viewable in an array of color palettes
- Measurements for data stored in PH5 format – in progress
- Improvements to ISPAQ – a Python utility for generating metrics for local data or any data available through FDSN web services.

Here we showcase how to use these enhancements to answer common data quality questions.

Is this channel healthy?

Channels are likely to be healthy if they

- are complete and continuous (*percent_availability*=100; *num_gaps*=0),
- have good timing (*clock_locked*>0; *timing_quality*>>0%; *suspect_time_tag*=0),
- record seismic energy (*pct_below_nlnm*<20), and
- have low noise levels (*pct_above_nhnm*<20).



Plots from **MUSTANG Databrowser**, a web-based visualization client (ds.iris.edu/mustang/databrowser/), show that but for one gap, II.PFO.00.BHZ fit these criteria during October 2019. We plan to incorporate these plots into a single view in 2020.

Tip

The **RRDS** (Research-Ready Datasets: service.iris.edu/irisws/rrds/1/) web service allows you to retrieve only data complying with metric constraints that you choose. To retrieve gap-free data having amplitudes consistent with the presence of seismic signal:

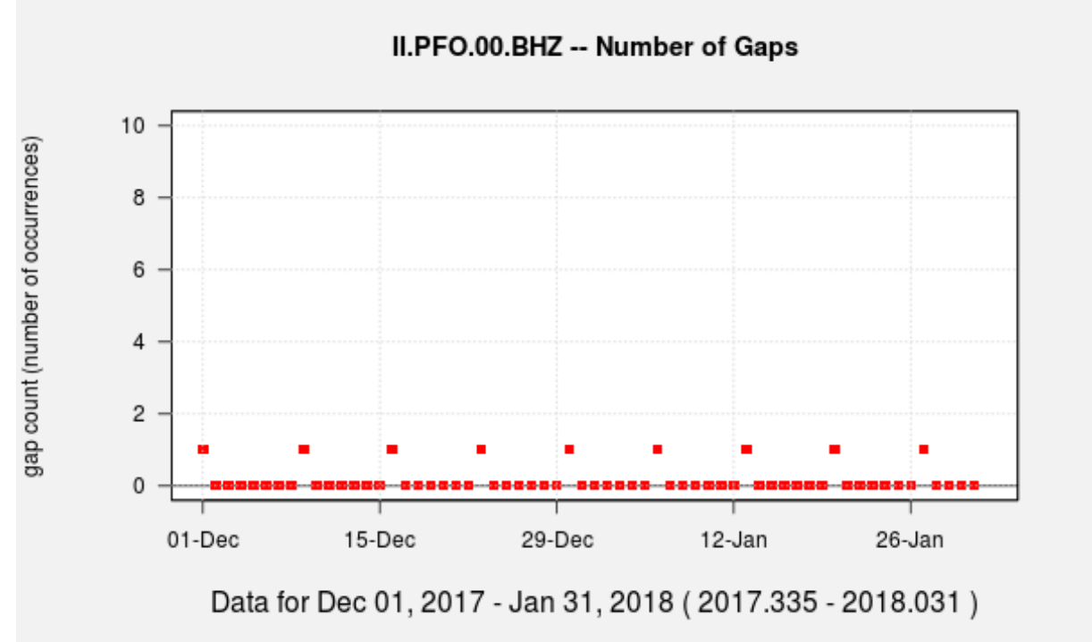
- build an **rrds** query using the URL builder found at the link above
- append metric constraints: &*num_gaps_eq=0&pct_below_nlnm_lt=20*
- send this query by web browser or **curl** to receive a list of suitable channels
- send this list using **FetchData** to receive your data. (**FetchData** is available at seiscode.iris.washington.edu/projects/ws-fetch-scripts/.)

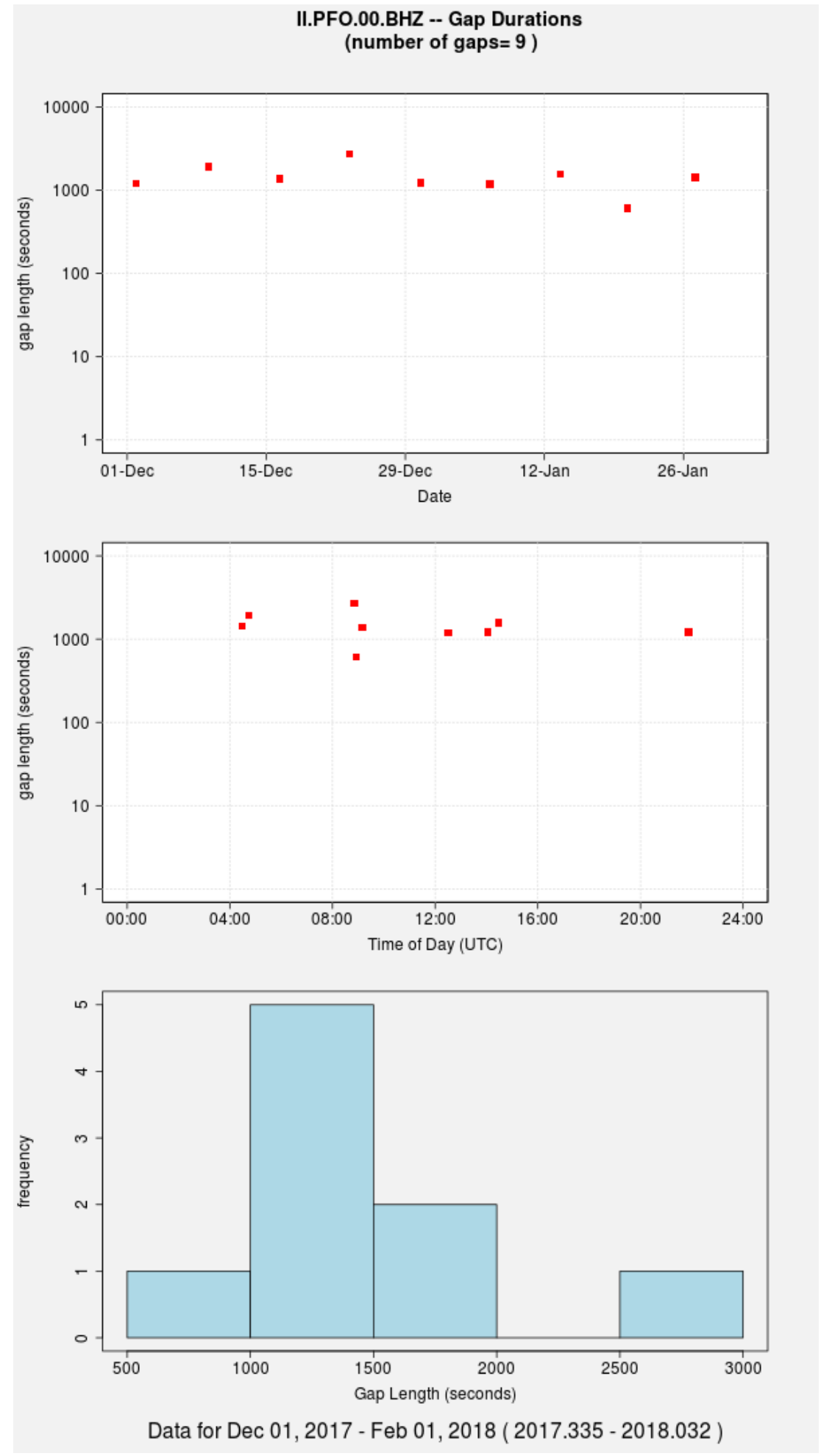
My station has gaps – what can I learn about this?

In late 2017 and early 2018, station II.PFO experienced a weekly data gap during which both dataloggers at the site rebooted.

Plotting gap duration by date, time of day and frequency of occurrence using **MUSTANG Databrowser** shows that gaps with similar durations had occurred each Saturday night.

Discussion revealed that a computer security contractor was probing the PFO dataloggers, causing them to reboot. An additional firewall solved the problem.





For Updates on these Projects

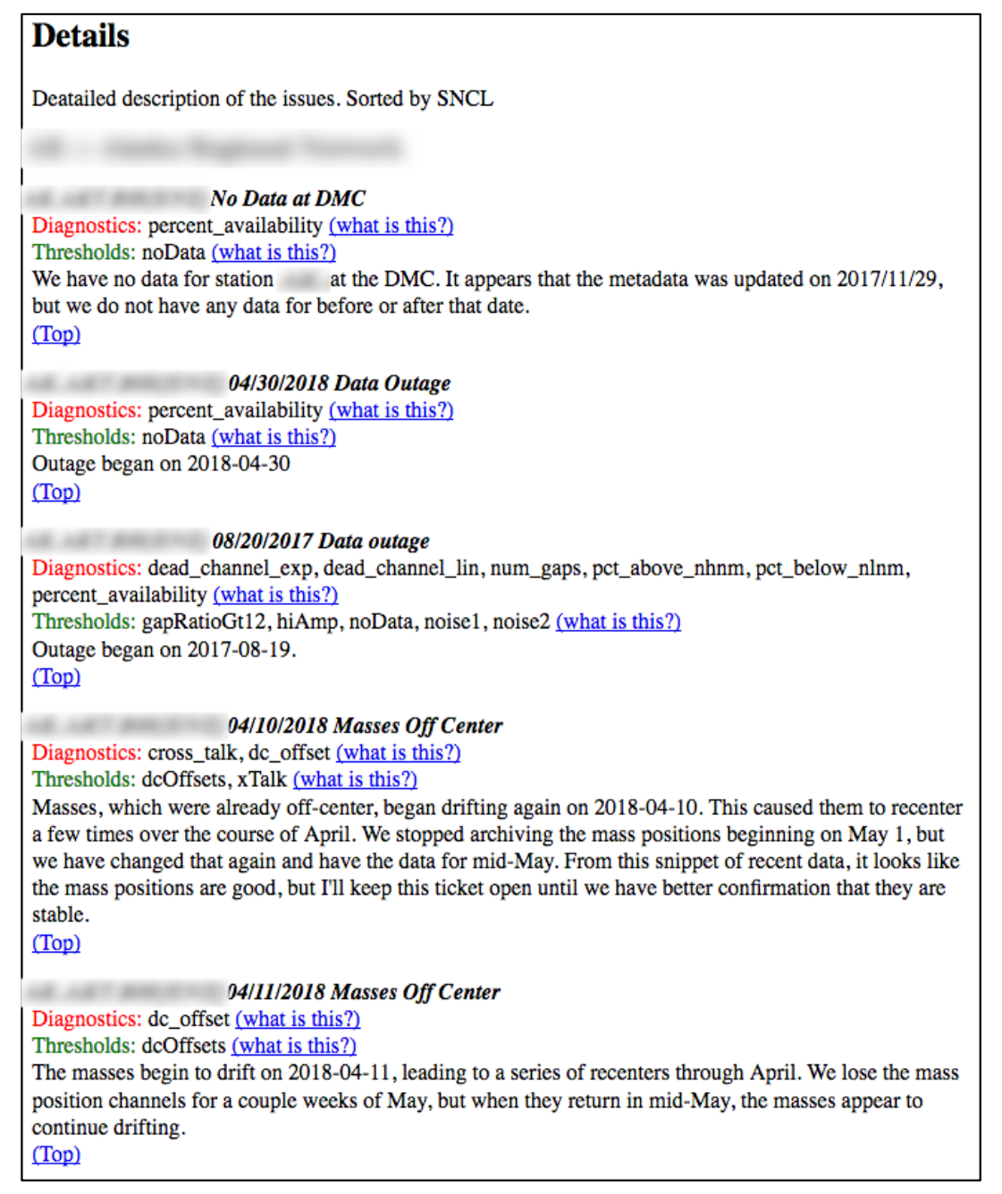
see **News** at ds.iris.edu/ds/nodes/dmc/quality-assurance/.

Which stations in my network have quality issues?

We are developing a new utility called **QuARG** (the Quality Assurance Report Generator) that queries MUSTANG metrics and guides the user through the steps of discovering, examining, tracking and reporting quality issues. It is particularly helpful for examining potential issues by providing an easy way to utilize several quality assurance tools that IRIS provides.

The end result is a **Quality Assurance Report** about problematic data in an easy-to-read format.

Once released, QuARG will be available at github.com/iris-edu.



Details

Deatailed description of the issues. Sorted by SNCL

[redacted]

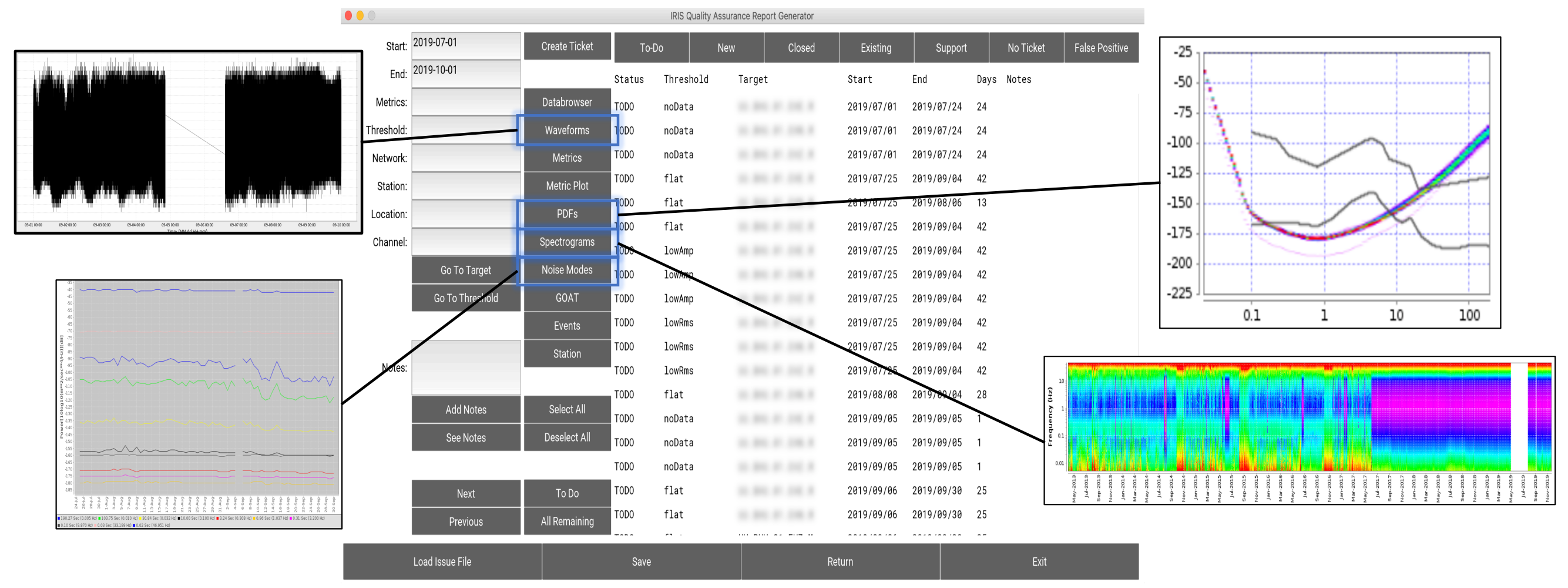
[redacted] *No Data at DMC*
Diagnostics: percent_availability (what is this?)
Thresholds: noData (what is this?)
We have no data for station [redacted] at the DMC. It appears that the metadata was updated on 2017/11/29, but we do not have any data for before or after that date.
(Top)

[redacted] *04/30/2018 Data Outage*
Diagnostics: percent_availability (what is this?)
Thresholds: noData (what is this?)
Outage began on 2018-04-30
(Top)

[redacted] *08/20/2017 Data outage*
Diagnostics: dead_channel_exp, dead_channel_lin, num_gaps, pct_above_nhnm, pct_below_nlnm, percent_availability (what is this?)
Thresholds: gapRatioGt12, hiAmp, noData, noise1, noise2 (what is this?)
Outage began on 2017-08-19.
(Top)

[redacted] *04/10/2018 Masses Off Center*
Diagnostics: cross_talk, dc_offset (what is this?)
Thresholds: dcOffsets, xTalk (what is this?)
Masses, which were already off-center, began drifting again on 2018-04-10. This caused them to recenter a few times over the course of April. We stopped archiving the mass positions beginning on May 1, but we have changed that again and have the data for mid-May. From this snippet of recent data, it looks like the mass positions are good, but I'll keep this ticket open until we have better confirmation that they are stable.
(Top)

[redacted] *04/11/2018 Masses Off Center*
Diagnostics: dc_offset (what is this?)
Thresholds: dcOffsets (what is this?)
The masses begin to drift on 2018-04-11, leading to a series of recenters through April. We lose the mass position channels for a couple weeks of May, but when they return in mid-May, the masses appear to continue drifting.
(Top)



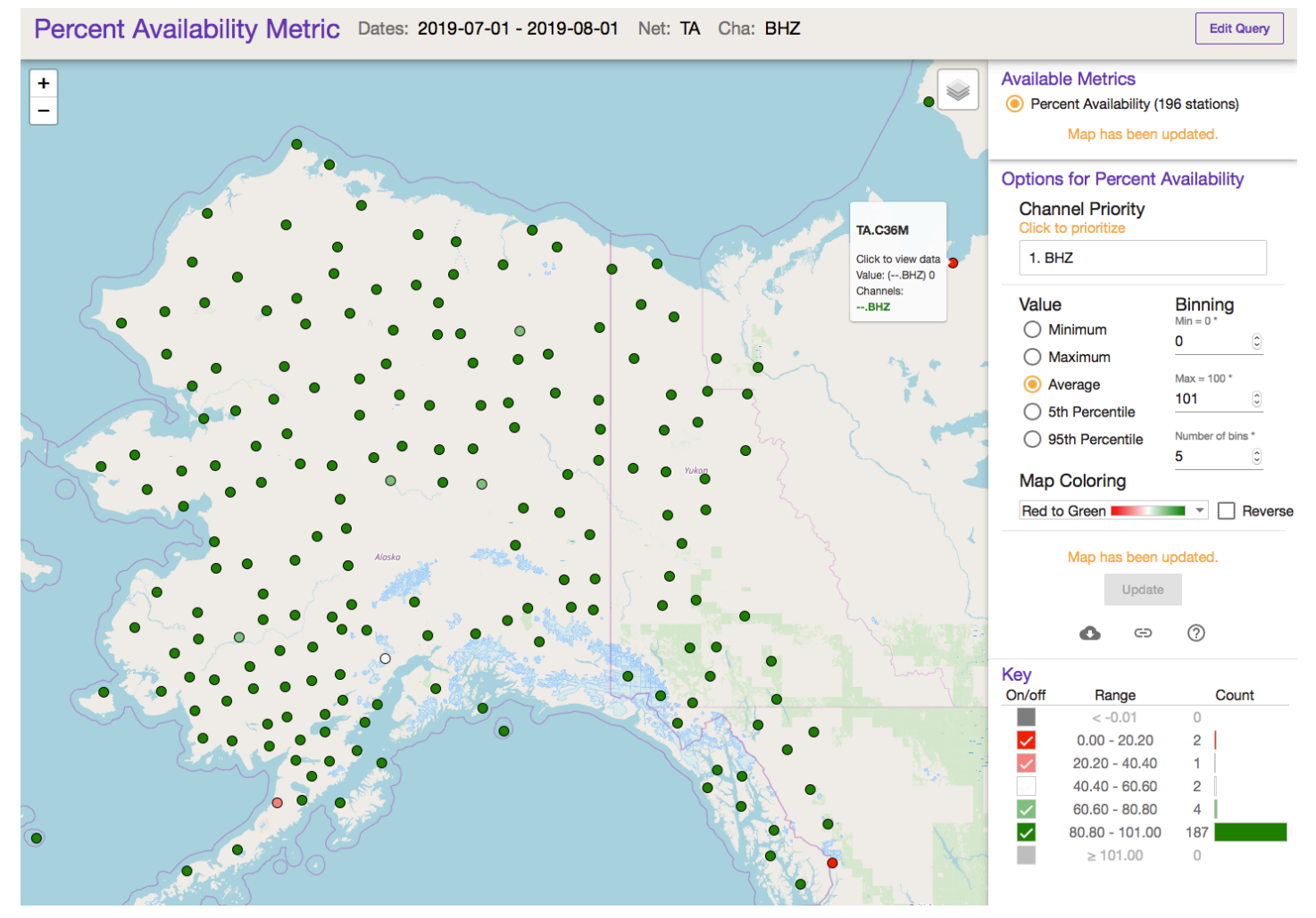
What other tools are available to quickly identify stations in my network that have anomalous metric values?

IRIS has two additional visualization tools that can summarize MUSTANG metrics over a range of time: MUSTANGular and LASSO.

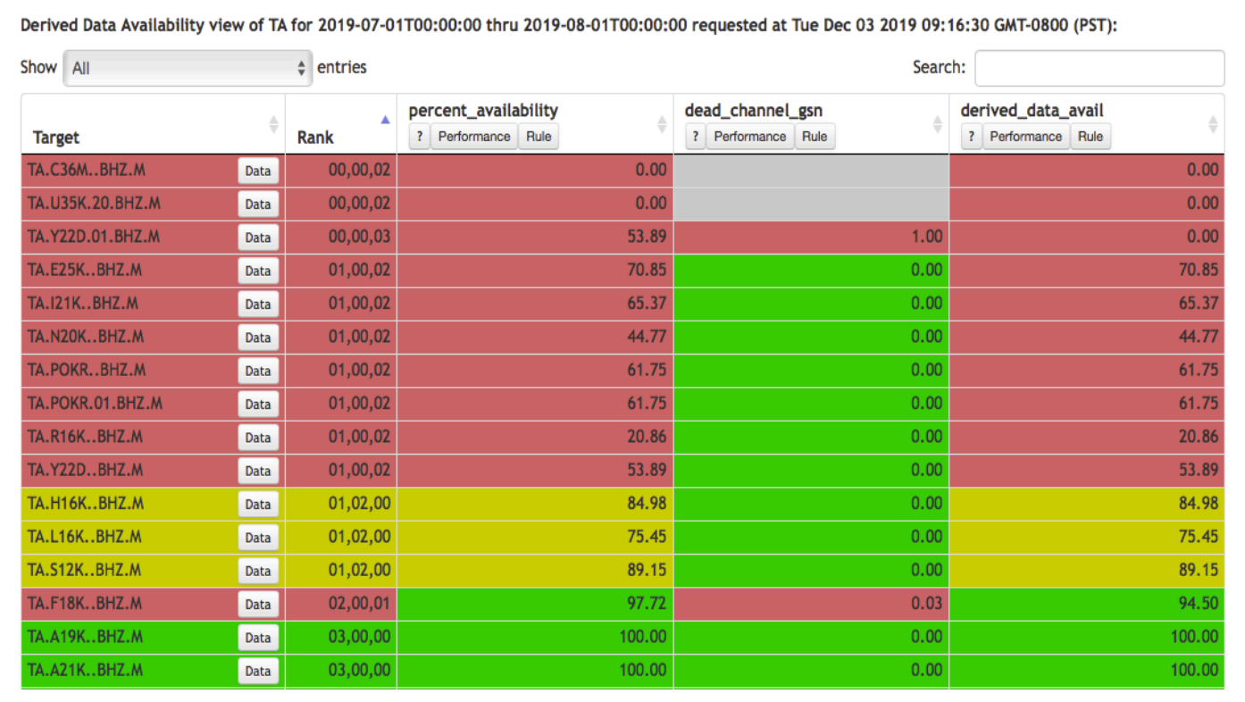
MUSTANGular displays a color-coded summary of station metric values on a map so that you can quickly find problem stations. This example identifies TA and AK network stations that have good to poor average data availability in July 2019.

Coming in 2020:

- combine all channel-locations from a station into a single value
- prioritize location code
- improvements to binning



ds.iris.edu/mustang/mustangular



Derived Data Availability view of TA for 2019-07-01T00:00:00 thru 2019-08-01T00:00:00 requested at Tue Dec 03 2019 09:16:30 GMT-0800 (PST):

Show All entries Search:

Target		Rank	percent_availability	dead_channel_gsn	derived_data_avail
TA.C36M..BHZ.M	Data	00,00,02	0.00		0.00
TA.U35K.20.BHZ.M	Data	00,00,02	0.00		0.00
TA.Y22D.01.BHZ.M	Data	00,00,03	53.89	1.00	0.00
TA.E25K..BHZ.M	Data	01,00,02	70.85	0.00	70.85
TA.I21K..BHZ.M	Data	01,00,02	65.37	0.00	65.37
TA.N20K..BHZ.M	Data	01,00,02	44.77	0.00	44.77
TA.POKR..BHZ.M	Data	01,00,02	61.75	0.00	61.75
TA.POKR.01.BHZ.M	Data	01,00,02	61.75	0.00	61.75
TA.R16K..BHZ.M	Data	01,00,02	20.86	0.00	20.86
TA.Y22D..BHZ.M	Data	01,00,02	53.89	0.00	53.89
TA.H16K..BHZ.M	Data	01,02,00	84.98	0.00	84.98
TA.L16K..BHZ.M	Data	01,02,00	75.45	0.00	75.45
TA.S12K..BHZ.M	Data	01,02,00	89.15	0.00	89.15
TA.F18K..BHZ.M	Data	02,00,01	97.72	0.03	94.50
TA.A19K..BHZ.M	Data	03,00,00	100.00	0.00	100.00
TA.A21K..BHZ.M	Data	03,00,00	100.00	0.00	100.00

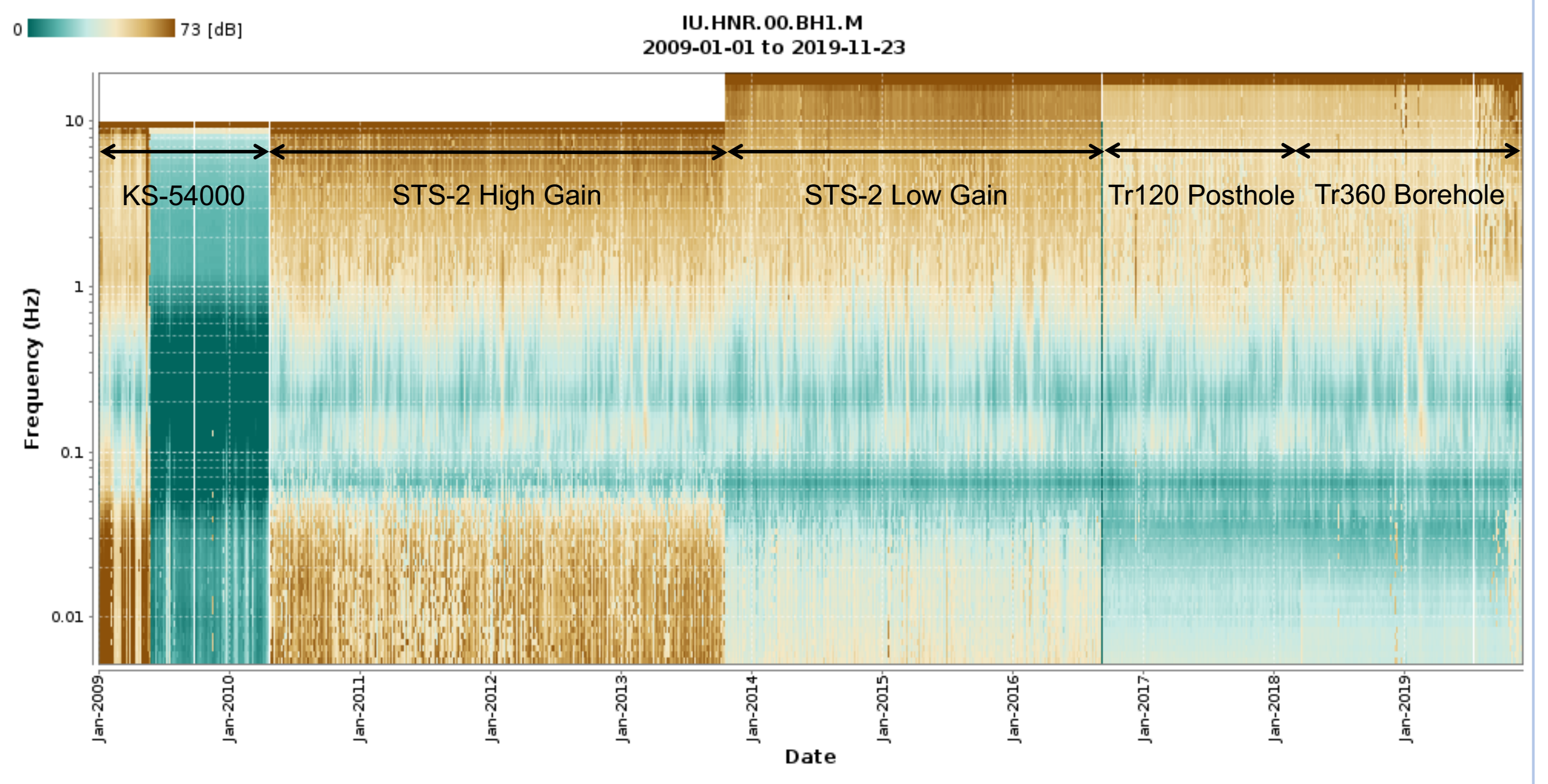
lasso.iris.edu

LASSO (Latest Assessment of Seismic Station Observations) ranks and color-codes stations based on metric values. This example combines the average *percent_availability* metric with *dead_channel_gsn* (a Boolean indicator of likely digitizer noise) to create a combined data availability value, then ranks stations from worst to best for the TA network in July 2019 (not all stations are shown).

How did hardware upgrades affect my station?

Noise spectrograms (service.iris.edu/mustang/noise-spectrogram/1/) of daily Probability Density Function (PDF) modes can elucidate small Power Spectral Density (PSD) changes over time.

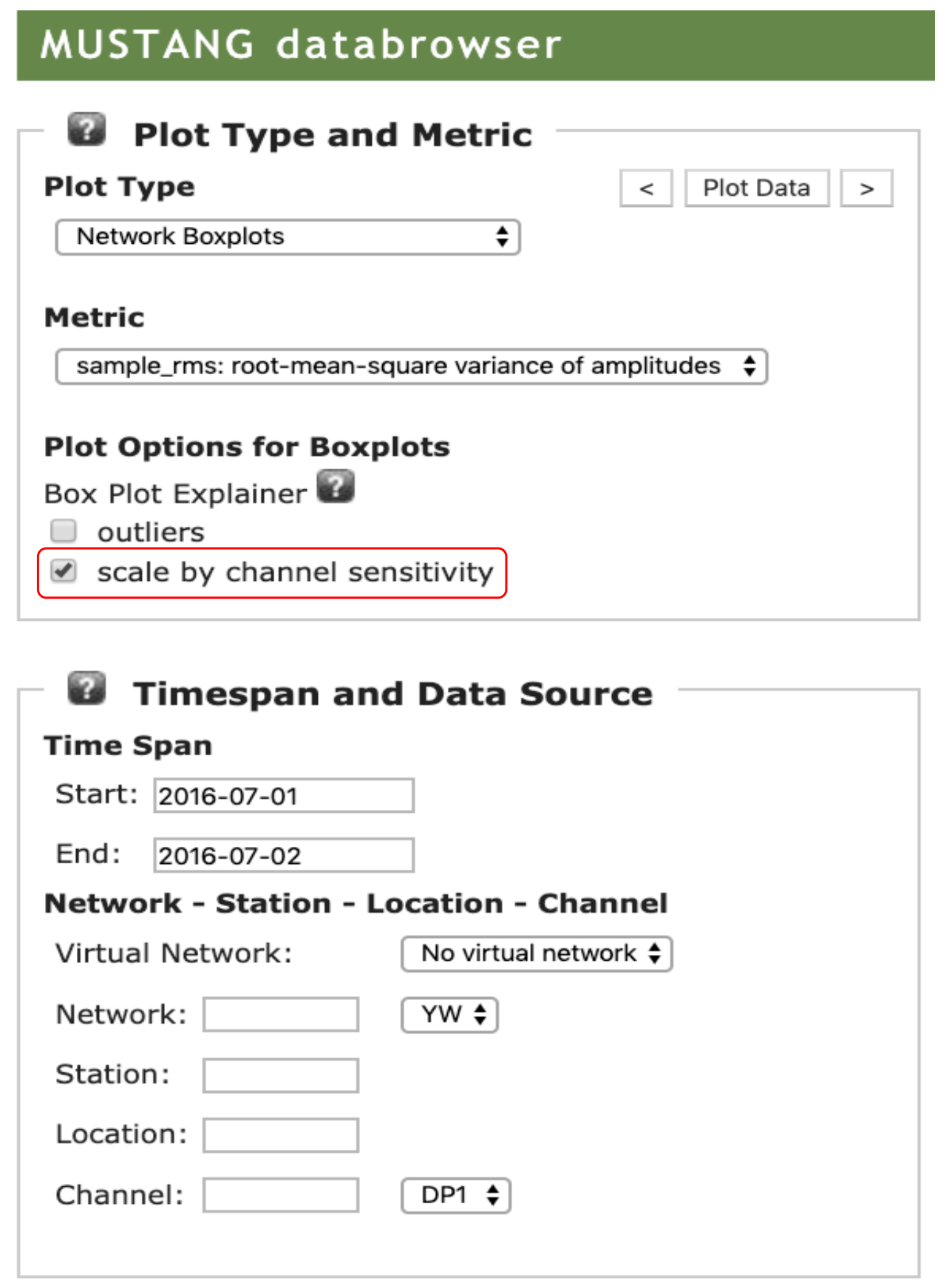
The spectrogram below shows that after the failure of the KS-54000 seismometer in 2010, low-frequency noise levels improved with each of three sensor upgrades. Noise levels for the final Trillium 360 borehole sensor appear comparable to the Trillium 120 posthole that it replaced.



Spectrograms can be displayed in a variety of color palettes that are compatible with greyscale or color printing, and/or are colorblind-friendly.

Which PH5 traces should I omit from my shot gathers?

We now calculate metrics for PH5 data and our next **MUSTANG Databrowser** release will display them. Plotting amplitude metrics as Network Boxplots will make culling geophones with anomalous amplitudes easier. High-amplitude outliers for the metric *sample_rms* appear at the top of the box plots. This metric is in units of Counts, so we recommend that you "scale by channel sensitivity" when comparing heterogeneous instruments.



MUSTANG databrowser

Plot Type and Metric

Plot Type: Network Boxplots

Metric: sample_rms: root-mean-square variance of amplitudes

Plot Options for Boxplots

Box Plot Explainer

□ outliers

☑ scale by channel sensitivity

Timespan and Data Source

Time Span

Start: 2016-07-01

End: 2016-07-02

Network - Station - Location - Channel

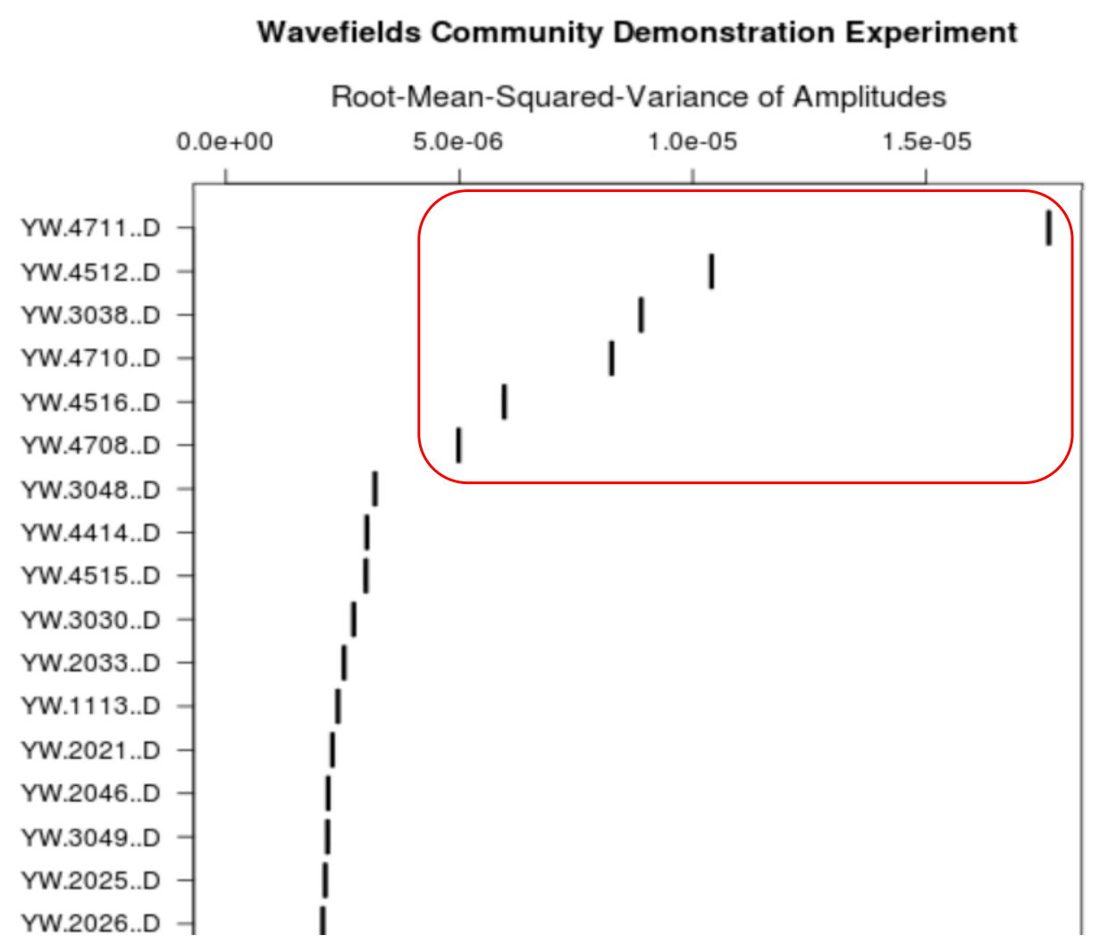
Virtual Network: No virtual network

Network: YW

Station:

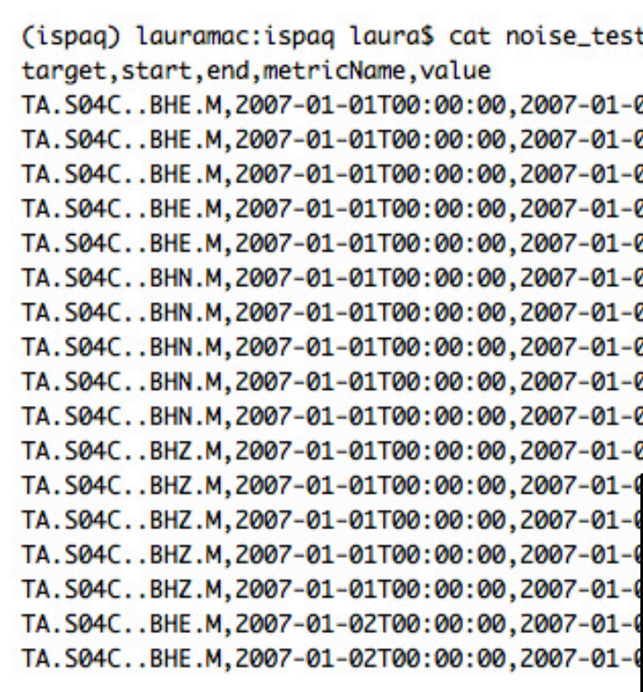
Location:

Channel: DP1

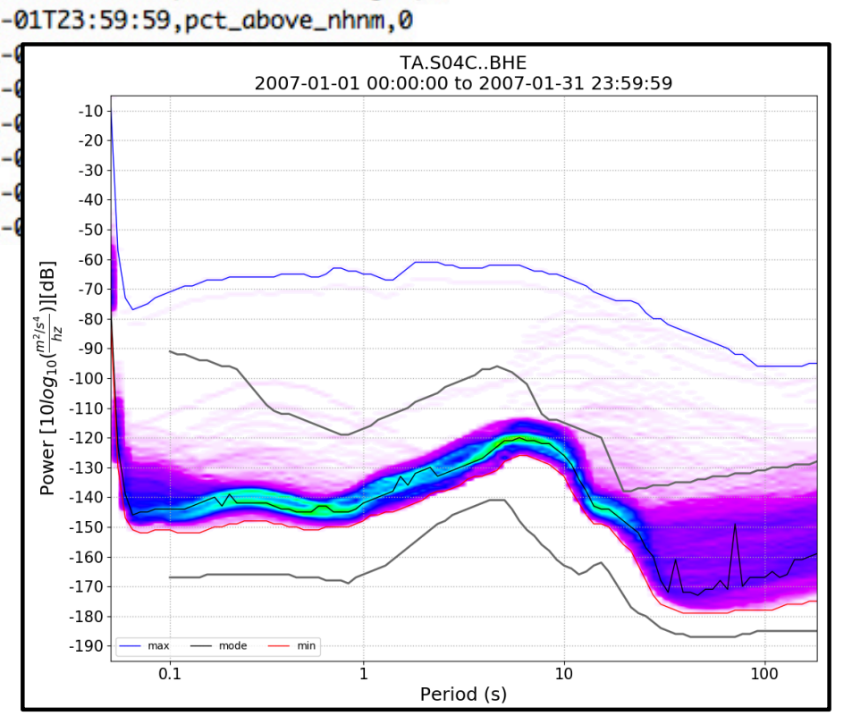


How can I generate QA metrics for data not housed at IRIS?

ISPAQ (IRIS System for Portable Assessment of Quality) is a command-line utility that allows users to generate MUSTANG metrics on local seismic data, or data from any of the Federated Data Centers. While ISPAQ is wrapped in Python code, at its core is the exact same R code that the DMC uses in-house to calculate metrics. ISPAQ stores metrics as text csv or PNG files. Improvements over the past year include the ability to create PDFs for arbitrary time periods.



```
(ispaq) lauramac:ispaq laura$ cat noise_test_test_local_2007-01-01_2007-01-03_PSDMetrics.csv
target,start,end,metricName,value
TA.S04C..BHE.M,2007-01-01T00:00:00,2007-01-01T23:59:59,pct_above_nhnm,0
TA.S04C..BHE.M,2007-01-01T00:00:00,2007-01-01T23:59:59,pct_below_nlnm,0
TA.S04C..BHE.M,2007-01-01T00:00:00,2007-01-01T23:59:59,dead_channel_exp,0.490119
TA.S04C..BHE.M,2007-01-01T00:00:00,2007-01-01T23:59:59,dead_channel_lin,10.2708
TA.S04C..BHE.M,2007-01-01T00:00:00,2007-01-01T23:59:59,dead_channel_gsn,0
TA.S04C..BHN.M,2007-01-01T00:00:00,2007-01-01T23:59:59,pct_above_nhnm,0.338491
TA.S04C..BHN.M,2007-01-01T00:00:00,2007-01-01T23:59:59,pct_below_nlnm,0
TA.S04C..BHN.M,2007-01-01T00:00:00,2007-01-01T23:59:59,dead_channel_exp,0.479492
TA.S04C..BHN.M,2007-01-01T00:00:00,2007-01-01T23:59:59,dead_channel_lin,9.95372
TA.S04C..BHN.M,2007-01-01T00:00:00,2007-01-01T23:59:59,dead_channel_gsn,0
TA.S04C..BHZ.M,2007-01-01T00:00:00,2007-01-01T23:59:59,pct_above_nhnm,0
TA.S04C..BHZ.M,2007-01-01T00:00:00,2007-01-0
TA.S04C..BHZ.M,2007-01-01T00:00:00,2007-01-0
TA.S04C..BHZ.M,2007-01-01T00:00:00,2007-01-0
TA.S04C..BHZ.M,2007-01-01T00:00:00,2007-01-0
TA.S04C..BHE.M,2007-01-02T00:00:00,2007-01-0
TA.S04C..BHE.M,2007-01-02T00:00:00,2007-01-0
```



ISPAQ is available from Github: github.com/iris-edu/ispaq.git.

Acknowledgements

Many thanks to Pete Davis, director of Project IDA for providing the II.PFO station gap example. LASSO is funded through IRIS Instrument Services. IRIS Data Services are funded through the National Science Foundation Cooperative Agreement EAR-1851048 (SAGE).