Proposal for FSDN standardization of waveform quality metrics

Knowledge of quality of seismic waveform data and related metadata is essential for any scientific analysis and interpretation of the data. Automated processes to calculate data quality parameters are required (a) to handle huge amounts of data, (b) to enable data centers to automatically monitor changes or variations in data quality over different time scales and (c) to provide services to the research community to search for and harvest the best quality data based on these parameters.

Currently, parallel developments are on-going (e.g. IRIS DMC, ORFEUS EIDA) to calculate data quality parameters and use these in different services. This document proposes the standardization of a number of basic metrics of which most are common in both systems.

Quality parameters are calculated for a time windows time series. In the following a time series is considered to belong to a data stream uniquely identified by a SEED network code, stations code, channel code and location code.

Metrics based on sample values

The following metrics are calculated from all sample values within time window [t_0, t_1]. Samples at $t=t_0$ are included, samples at $t=t_1$ are excluded. Default time window length is 24 hours, starting at 00:00:00.0 UTC.

- **sample_mean**

Average value of all samples (x_1, .., x_N).

$$mean = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- **sample_max**

Maximum value of all samples (x_1, .., x_N).

- **sample_min**

Minimum value of all samples (x_1, .., x_N).

- **sample_rms**

The root mean square of all samples (x_1, .., x_N).

$$rms = \sqrt{\sum_{i=1}^{N} \frac{x_i^{\ 2}}{N}}$$

- **sample_stdev**

The standard deviation of all samples (x_1, .., x_N).

$$stdev = \sqrt{\sum_{i=1}^{N} \frac{(x_i - mean)^2}{N}}$$

Metrics based on mini-SEED data records

The following metrics are extracted from headers in mini-SEED records fitting time window **[t_0, t_1].** Default time window length is 24 hours, starting at 00:00:00.0 UTC.

Data records having start time $\mathbf{T_s}$ and end time $\mathbf{T_e}$ for which $\mathbf{t_o \le T_e \le t_1}$ are included.

Data records having start time $\mathbf{T_s}$ and end time $\mathbf{T_e}$ for which $\mathbf{t_o \le T_s \le t_1}$ are included.

These metrics follow from flags that are specifically defined in SEED. To distinguish these from metrics that are calculated from the sample values (time series) the names are pre-pended by 'ms_'.

- **ms_amplifier_saturation**

Number of records fitting time window [t_0, t_1] in which bit 0 in the Data Quality Flags byte is set to '1'.

- **ms_digitizer_clipping**

Number of records fitting time window [t_0, t_1] in which bit 1 in the Data Quality Flags byte is set to '1'.

- **ms_spikes**

Number of records fitting time window [t_0, t_1] in which bit 2 in the Data Quality Flags byte is set to '1'.

- **ms_glitches**

Number of records fitting time window [t_0, t_1] in which bit 3 in the Data Quality Flags byte is set to '1'.

- **ms_missing_padded_data**

Number of records fitting time window [t_0, t_1] in which bit 4 in the Data Quality Flags byte is set to '1'.

- **ms_telemetry_sync_error**

Number of records fitting time window [t_0, t_1] in which bit 5 in the Data Quality Flags byte is set to '1'.

- **ms_digital_filter_charging**

Number of records fitting time window [t_0, t_1] in which bit 6 in the Data Quality Flags byte is set to '1'.

- **ms_suspect_time_tag**

Number of records fitting time window [t_0, t_1] in which bit 7 in the Data Quality Flags byte is set to '1'.

- **ms_calibration_signal**

Number of records fitting time window [t_0, t_1] in which bit 0 in the Activity Flags byte is set to '1'.

- **ms_timing_correction**

Number of records fitting time window [t_0, t_1] in which bit 1 in the Activity Flags byte is set to '1'.

- **ms_event_begin**

Number of records fitting time window [t_0, t_1] in which bit 2 in the Activity Flags byte is set to '1'.

- **ms_event_end**

Number of records fitting time window [t_0, t_1] in which bit 3 in the Activity Flags byte is set to '1'.

- **ms_event_in_progress**

Number of records fitting time window [t_0, t_1] in which bit 6 in the Activity Flags byte is set to '1'.

- **ms_clock_locked**

Number of records fitting time window [t_0, t_1] in which bit 5 in the I/O & Clock Flags byte is set to '1'.

- **ms_timing_quality**

Average of the timing quality percentage value stored in miniSEED blockettes 1001. Value is NULL if not present in the data records.

- **ms_timing_quality_max**

Maximum of the timing quality percentage value stored in miniSEED blockettes 1001. Value is NULL if not present in the data records.

- **ms_timing_quality_min**

Minimum of the timing quality percentage value stored in miniSEED blockettes 1001. Value is NULL if not present in the data records.

Metrics based on gaps and overlaps

The following metrics are calculated within time window [t_0, t_1]. Samples at $t=t_0$ are included, samples at $t=t_1$ are excluded. Default time window length is 24 hours, starting at 00:00:00.0 UTC

Definitions: Continuous time series

A continuous (discrete) time series is defined as a time series in which the time interval between two adjacent samples (a) is constant (Δt) or (b) does not differ from this constant by more than a certain time tolerance (**ε**).

$\Delta t - \varepsilon <= T_{i+1} - T_i <= \Delta t + \varepsilon$.

i is the sample index number in the time series, T_i is the time corresponding to sample *i*, Δt the (constant) sample rate (in s), ε is the time tolerance is s. Default tolerance is 0.

Gap

A gap in the time series is defined as the time interval (in seconds) between two adjacent, continuous time series for which the time difference between the start of the second time series ($T_{s,2}$) and the end of the first time series ($T_{e,1}$) exceeds the sample rate interval (Δt) by more than the time tolerance ε: $T_{s,2} - T_{e,1} > \Delta t + \varepsilon$

gap = $T_{s,2} - T_{e,1}$

Within a time window (t_0, t_1) a time series may have a start time (T_s) later than t_0 or a stop time (T_e) before t_1. These start and end gaps are defined as:

- when $T_s - t_0 > \Delta t + \varepsilon$ gap: $T_s - t_0$
- when $t_1 - T_e > \Delta t + \varepsilon$ gap: $t_1 - T_e$

Here, T_s and T_e are the times of the first sample and last sample in the time window (t_0, t_1).

Overlap

An overlap in the time series is defined as the time interval (in seconds) between two adjacent continuous time series for which the time difference between the end of the first time series ($T_{e,1}$) and the start of the second time series ($T_{s,2}$) exceeds the sample rate interval (Δt) by more than the time tolerance ε: $T_{e,1} - T_{s,2} > \Delta t + \varepsilon$

overlap = $T_{e,1}$ - $T_{s,2}$

Within a time window (t_0, t_1) an overlap can be intersected:

- when $T_{e,1}$ - t_0 > $\Delta t + \varepsilon$ overlap: $T_{e,1}$ - t_0
- when t_1 - $T_{s.2}$ > $\Delta t + \varepsilon$ overlap: t_1 - $T_{s,2}$

- **num_gaps**

The number of gaps in time interval [t0, t1].

- **sum_gaps**

The sum of all gaps in time interval [t0, t1].

- **num_overlaps**

The number of overlaps in time interval [t0, t1].

- **sum_overlaps**

The sum of all overlaps in time interval [t0, t1].

- **max_gap**

Largest gap in seconds.

- **max_overlap**

Largest overlap in deconds.

- **percent_availability:**

Data availability is the percentage of data available in the time window (t_0, t_1). It is calculated as length of the time window (t_0, t_1) minus the sum of all gaps in this time window, relative to the length of the time window t_1-t_0.

$$availability = 100 \times \frac{(t_1 - t_0) - sum\_gaps}{t_1 - t_0}$$