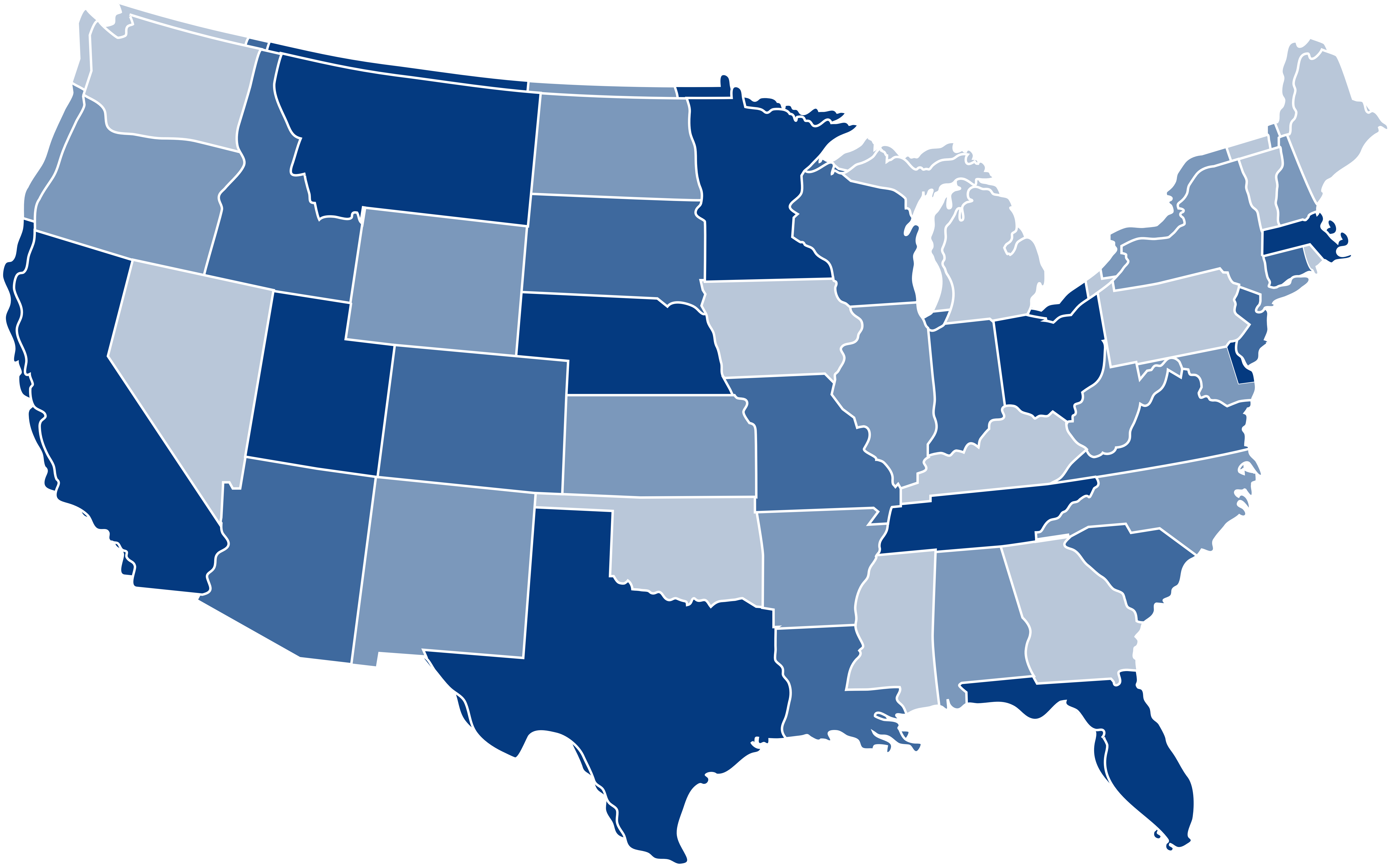

Measuring Broadband America

Validated Data Cleansing, Tenth Report

This document is based on September/October 2019 data. It outlines the data cleansing processes used to generate the 'validated' dataset from the 'raw' dataset. The 'validated' September/October 2019 data published by the FCC will have had these operations performed upon it.

Information

This document is based on September/October 2019 data. It outlines the data cleansing processes used to generate the 'validated' dataset from the 'raw' dataset. The 'validated' September/October 2019 data published by the FCC will have had these operations performed upon it. The SQL scripts used to conduct these tasks are available in the file:

Measuring Broadband America 10th Report Validated data can be found at:
https://files.samknows.com/~fcc/validated_data/sept2019/validated_data_sept2019.tar.gz

The SQL scripts used to conduct these tasks are available in the file:
http://Files.samknows.com/data/MBA%2010th%20Report/

Data Processing Flow Validation Operations

Remove all data prior to September 6th to October 3rd and from October 8th to October 9th

- The FCC reporting period ran between the following dates: September 6th – October 3rd and then October 8th to October 9th.
- Data for September prior to this period and for October after this period was removed from the dataset.
- Only participants who provided a minimum of 5 days of valid measurements during the testing period were included in the September / October 2019 results.
- Data was only charted when results from at least 45 separate Whiteboxes was available. Instances of 44 or fewer Whiteboxes were noted for possible future augmentation.

Handle panelists that changed ISP intra-month

Some panelists changed ISP mid-month. In situations where this occurred we removed the data for the ISP that they spent the shortest period on. For example, if the panelist changed from ISP A to ISP B on September 10th, we would remove data prior to this date because there would be a larger dataset (ie., September 12th to the 30th) for their performance on ISP B.

A daily log of the panelists' public addresses was used to determine when they changed ISP. This table records on a daily basis the owner of the netblock that the panelist's public facing IP address resides in. This allowed us to identify when people changed ISPs quickly and reliably.

We had two mechanisms for identifying panelists that changed service tiers:

Firstly, ISP-supplied panelist validation information informed us of which service tier the panelist was subscribed to. Some ISPs also provided the date at which they began this service.

However, in cases where the ISP was unable to validate the panelist in question or their validation was delayed, we used the following process:

- Find the difference between the average sustained throughput observed for the first three days in the reporting period from the average sustained throughput observed for the final three days in the reporting period (if the unit wasn't online at the start or end, then we take the first/final three days that they were actually online.
- If this difference is over 50%, we examined the downstream and upstream charts for this unit.
- Where an obvious step change is observed (e.g. from 25Mbit/s to 50Mbit/s), flag the data for the shorter period for removal.

General Data Cleansing

- Only the curr_httpgetmt (download throughput), curr_httppostmt (upstream throughput), curr_udplatency (UDP latency/loss) and curr_webget (web browsing) results were considered for the FCC analysis.
- All results from non-M-Lab and non-Level3 targets were removed from the curr_httpgetmt, curr_httppostmt and curr_udplatency tables.

Speed test (httpgetmt, httppostmt) cleansing

- All failed tests were removed. Failed speed tests were not considered in the analysis.

UDP latency/loss cleansing

- All test instances (one per hour, per unit) with less than 50 samples (out of a potential maximum of 600) were removed.
- Data was excluded where a unit's packet loss exceeded 10% within a single hour. Such a high level of loss would render a connection unusable and is considered an anomalous event.
- Data was excluded where a test node experienced more than 10% packet loss across all of the units testing against it within a single hour. This was intended to capture instances where the M-Lab or Level3 node was offline.

Web Browsing

- All test instances where the page load time exceeded 30 seconds were removed.
- All failed tests were removed. Failed results were not analyzed in this report.