# FCC Measuring Broadband America Validated September / October Data 2015 Cleansing

Info: http://www.fcc.gov/measuring-broadband-america/2014/validated-data-fixed-2014

This document outlines the data cleansing processes used to generate the ‘validated’ dataset from the ‘raw’ dataset. Please note that the ‘validated’ September / October 2015 data published on the FCC website has already had these operations performed upon it. The SQL scripts used to conduct these tasks are available in the file

<http://data.fcc.gov/download/measuring-broadband-america/2016/sql-cleanup-scripts-sept2015.tar.gz>

Remove all data outside of September 29th – 30th, October 2nd – 7th and October 10th – 31st.

The FCC reporting period ran between the following dates:

September 29th – 30th

October 2nd – 7th

October 10th – 31st

Data outside of this period was removed from the dataset.

Only participants who provided a minimum of 5 days of valid measurements and had valid data in each of the three time intervals were included in the September / October 2015 results.

Data was only charted when results from at least 25 separate Whiteboxes was available. Instances of 30 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 April 10th, we would remove data prior to April 10th because there would be a larger dataset 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.

Handle panelists that changed service tier intra-month

Some panelists changed service tiers mid-month (e.g. upgraded from a 768kbps plan to a 3Mbit/s plan). Where this occurred we removed the data for the tier that they spent the shortest period on. For example, if the panelist changed from tier A to tier B on April 10th, we would remove data prior to April 10th because there would be a larger dataset for their performance on tier B.

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:

1. 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)
2. If this difference is over 50%, we examined the downstream and upstream charts for this unit
3. Where an obvious step change is observed (e.g. from 768kbps to 3Mbit/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), curr\_webget (Web browsing) and curr\_netusage (Consumption) 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.

All tests with greater than 6 result intervals were removed. Speed tests should have no more than 6 result intervals (0-5s, 0-10s, 0-15s, 0-20s, 0-25s, 0-30s). A minority will have fewer than 6 intervals (e.g. if they exhausted the 100MB payload on each of the three connections in under 30 seconds), but this is acceptable.

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.

Consumption

All instances where the number of bytes consumed by the measurements exceeded the total bytes consumed were removed. The total bytes consumed should be greater than or equal to the number of bytes consumed by the measurements.