**Fixed Broadband Monthly Data Usage Estimation Methodology for 2017 Usage Allowance**

**Introduction**

The objective of this analysis is to estimate the average data usage of a majority of fixed broadband customers[[1]](#footnote-1) using publicly available Measuring Broadband America (MBA) data.[[2]](#footnote-2) This analysis is based on MBA raw data files from July 2015 to September 2016 for NetUsage and the profile information files, which are the most recent data available.[[3]](#footnote-3)

The Commission describes the MBA program as “an ongoing nationwide performance study of broadband service in the United States that developed out of a recommendation by the National Broadband Plan to improve the availability of information for consumers about their broadband service. This program is built on principles of openness and transparency. The FCC has made available to stakeholders and the general public the open source software used on both its fixed and mobile applications, the data collected, and detailed information regarding the FCC's technical methodology for analyzing the collected data.”[[4]](#footnote-4) Although the raw data used in this analysis has not been audited,*[[5]](#footnote-5)* FCC staff believes that using the most recent data available provides the best basis for a data usage estimate.

**NetUsage Data**

The NetUsage metric is a passive test included in the MBA results. Under normal operations, there is one NetUsage report on total WAN usage and MBA test measurements usage transmitted per hour per individual panelist unit (whitebox) to the MBA data collection system. Both the download (rx) and upload (tx) bytes for the WAN and test usage are included.[[6]](#footnote-6)

The following are the fields in the MBA raw data file named curr\_netusage.csv that are publicly available at the MBA (Fixed) website:

|  |  |
| --- | --- |
| unit\_id  | Unique identifier for an individual unit  |
| dtime  | Time test finished in UTC  |
| wan\_rx\_bytes  | Total bytes received via the WAN interface on the unit (incl. Ethernet/IP headers)  |
| wan\_tx\_bytes  | Total bytes transmitted via the WAN interface on the unit (incl. Ethernet/IP headers)  |
| sk\_rx\_bytes  | Bytes received as a result of active performance measurements  |
| sk\_tx\_bytes  | Bytes transmitted as a result of active performance measurements  |
| location\_id  | Please ignore (this is an internal key mapping to unit profile data)  |

**Data for Analysis**

The monthly data usage for a reporting unit is the sum of total bytes of data transmitted and received via WAN interface, subtracting total bytes of data received and transmitted resulting from active MBA test measurements. Under normal operations, there is one NetUsage report transmitted per hour per individual panelist unit. These measurements were summarized by individual panelist (whitebox reporting unit) per calendar month. A copy of the data used is available on the FCC website in .csv format and the field labels are described in Appendix 1.

The MBA data consists of 34,594 records of reporting unit and month combinations. The table below presents the number of records by technology in each month.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|   | CABLE | DSL | FIBER | SATELLITE | UNKNOWN | WIRELESS |
| 7/1/2015 | 1835 | 614 | 206 | 21 | 174 | 0 |
| 8/1/2015 | 1802 | 594 | 200 | 21 | 170 | 0 |
| 9/1/2015 | 1735 | 581 | 194 | 21 | 165 | 0 |
| 10/1/2015 | 1700 | 572 | 192 | 18 | 159 | 0 |
| 11/1/2015 | 1648 | 560 | 188 | 20 | 154 | 0 |
| 12/1/2015 | 1614 | 555 | 189 | 19 | 145 | 0 |
| 1/1/2016 | 1582 | 539 | 185 | 18 | 140 | 1 |
| 2/1/2016 | 1541 | 525 | 187 | 18 | 137 | 1 |
| 3/1/2016 | 1507 | 510 | 184 | 17 | 137 | 1 |
| 4/1/2016 | 1484 | 504 | 185 | 16 | 138 | 1 |
| 5/1/2016 | 1463 | 505 | 180 | 18 | 135 | 1 |
| 6/1/2016 | 1427 | 494 | 180 | 17 | 128 | 1 |
| 7/1/2016 | 1420 | 490 | 174 | 16 | 133 | 1 |
| 8/1/2016 | 1388 | 478 | 172 | 15 | 133 | 1 |

Many MBA records are excluded from the analysis, resulting in a total of 12,827 monthly summary records available for the analysis. The reasons for exclusion include possible reporting error causing negative monthly data usage, incomplete data, and records out of scope for analysis. A record may be excluded from the analysis for more than one reason. There are 1,826 records with negative monthly data usage.[[7]](#footnote-7) Records in December 2015 (2,522) and June 2016 (2,247) are incomplete with missing days of information.[[8]](#footnote-8) Records with potential incomplete information (4,842) due to low number of reports are excluded from the analysis. The method to determine the number of reports cutoff is described in Appendix 2. Records without current validated profile information (1,433) and from ISPs not part of MBA program (2,766) are excluded. In addition, our analysis for purposes of developing a usage allowance for fixed broadband service only includes records with download speeds equal to or less than 25 Mbps from cable, DSL, and fiber services. Therefore, 15,937 records are out of scope for analysis based on download speed. The table below shows the number of qualifying monthly records by technology by month for analysis.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CABLE | DSL | FIBER |
| 7/1/2015 | 668 | 470 | 55 |
| 8/1/2015 | 659 | 461 | 56 |
| 9/1/2015 | 650 | 462 | 54 |
| 10/1/2015 | 636 | 454 | 54 |
| 11/1/2015 | 629 | 441 | 54 |
| 1/1/2016 | 592 | 419 | 49 |
| 2/1/2016 | 581 | 417 | 55 |
| 3/1/2016 | 567 | 404 | 56 |
| 4/1/2016 | 560 | 404 | 54 |
| 5/1/2016 | 555 | 385 | 53 |
| 7/1/2016 | 522 | 374 | 52 |
| 8/1/2016 | 520 | 356 | 49 |

**Results**

The table below shows the mean of monthly data usage for terrestrial fixed broadband by month and technology. We do not observe a time trend between July 2015 and August 2016.[[9]](#footnote-9)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | CABLE | DSL | FIBER | Overall Average |
| 7/1/2015 | 201.28 | 109.07 | 139.73 | 162.12 |
| 8/1/2015 | 202.68 | 110.75 | 126.37 | 163.01 |
| 9/1/2015 | 197.83 | 103.77 | 114.86 | 156.72 |
| 10/1/2015 | 197.84 | 110.21 | 136.35 | 160.16 |
| 11/1/2015 | 197.77 | 113.03 | 135.58 | 161.53 |
| 1/1/2016 | 226.79 | 122.12 | 150.46 | 181.88 |
| 2/1/2016 | 201.61 | 109.26 | 119.81 | 160.77 |
| 3/1/2016 | 205.63 | 114.60 | 115.63 | 164.91 |
| 4/1/2016 | 190.10 | 109.33 | 98.39 | 153.18 |
| 5/1/2016 | 198.78 | 113.88 | 109.87 | 161.12 |
| 7/1/2016 | 217.13 | 114.57 | 151.25 | 173.06 |
| 8/1/2016 | 215.78 | 124.69 | 145.31 | 176.99 |
| Overall Average | 204.11 | 112.63 | 128.25 | 164.32 |

The estimated average monthly data usage is 164GB with a 95% confidence interval of (160.75, 167.90).

**Limitations**

The data used for our analysis is subject to selection bias. The MBA reporting units are randomly sampled from a list of volunteers. The volunteers may not represent the population nationwide. In addition, the usage data is not currently collected from newly installed whiteboxes because either these whiteboxes do not support the data collection system, or these whiteboxes are not configured/set up for collecting usage data. A whitebox may be replaced because the whitebox malfunctioned, the reporting unit was a newly selected supplemental sample, or the service provider has alternated the services. The attrition rate of the legacy whiteboxes was an average of 2% per month between July 2015 and August 2016. Assuming the trend continues, the estimated number of whiteboxes in January 2017 would be about 34% less than it was in July 2015, and the estimated number of whiteboxes by the end of 2017 would be about 53% less than it was in July 2015. All of the above may introduce selection bias into the data.

**APPENDIX A**

**Labels for NetUsage-Analysis-File.csv**

|  |  |
| --- | --- |
| unit\_id | Unique identifier for an individual unit  |
| reports | Count of the individual reports submitted by a whitebox to the data collection system |
| data\_year | Calendar year of the record |
| data\_month | Calendar month of the record |
| jun\_exclusion | Flag to indicate the record is excluded because it was a June 2016 measurement |
| dec\_exclusion | Flag to indicate the record is excluded because it was a December 2015 measurement |
| download too high | Flag to indicate the record is excluded because it was a whitebox with a download speed tier > 25 Mbps |
| to\_few\_reports | Flag to indicate the record is excluded because it was a whitebox with too few hourly report in the month |
| negative\_number | Flag to indicate the record is excluded because it was a whitebox with negative result from (wan\_total\_GB - sk\_total\_GB) |
| use\_code | Summary flag that includes all other flags if they occurred for that whitebox, to allow visibility where multiple reasons excluded a whitebox |
| wan\_rx\_bytes | Total bytes received via the WAN interface on the unit (incl. Ethernet/IP headers)  |
| wan\_tx\_bytes | Total bytes transmitted via the WAN interface on the unit (incl. Ethernet/IP headers)  |
| sk\_rx\_bytes | Bytes received as a result of active performance measurements  |
| sk\_tx\_bytes | Bytes transmitted as a result of active performance measurements  |
| wan\_total\_GB | Calendar month total of all usage - Summation of (wan\_tx\_bytes + wan\_rx\_bytes) |
| sk\_total\_GB | Calendar month total of all usage - Summation of (sk\_tx\_bytes + sk\_rx\_bytes) |
| wan\_sk\_total\_GB | Calendar month total of consumer usage by taking out MBA test usage (wan\_total\_GB - sk\_total\_GB) |
| isp | Internet Service Provider network the the whitebox uses |
| download | Advertised download speed tier for the whitebox |
| upload | Advertised upload speed tier for the whitebox |
| technology | Last-mile technology for the whitebox |
| state | State where the whitebox is located |
| city\_nearest\_name | Nearest city to the whitebox |
| city\_nearest\_distance | Distance to the nearest city from the whitebox |
| mba\_isp | Flag to indicate if the ISP is an MBA ISP as identified in the 2016 MBA Report |

**APPENDIX B**

**Number of Reports Cutoff**

The data collection system is designed to collect and report data usage from each whitebox once per hour. There would normally be up to a total of 744 (24 x 31) reports from one whitebox in a 31-day month (720 reports in a 30-day month). In the case of skipped reports, the system is designed to capture the data in between reports in the next report. Therefore, the number of reports should not be correlated with the monthly data usage if the data collection system functions properly all times. However, our data suggests the opposite. The monthly data usage per number of reports does not appear to be random (see Figure 1, below). A possible reason would be that the reports from a given whitebox were sent, received, or processed incompletely for unknown reasons, or the whitebox was malfunctioning. Therefore, the usage data from the whitebox was only collected partially in a given month.

Figure 1.



To filter out possible partially collected usage data, we evaluated the randomness of monthly data usage with different levels of cutoffs, based on the deviance of a linear model that log monthly data usage is a function of log number of reports in a month.[[10]](#footnote-10) The randomness is approximated using the ratio of residual deviance and null deviance of the model, ranging from 0 to 1. When this ratio is closer to 1, the monthly data usage is closer to a random distribution over the number of reports. The optimal number of reports cutoff for each month is determined by maximizing monthly data usage randomness. Figure 2 shows the ratio of residual deviance and null deviance, as an approximation of randomness, by levels of cutoffs for multiple months. The blue lines graph where the optimal number of reports cutoff is for different months. Above this number of reports cutoff, the number of reports has minimum correlation with the monthly data usage. The distribution of the optimal cutoffs is fairly normal with a mean of 576.4. The number of reports cutoff for our analysis is then set to be 576. Records with a number of reports less than 576 are excluded from the analysis.

 Figure 2



1. We will use the following terms synonymously: consumer; customer; panelist; and whitebox. These terms represent the MBA consumer/customer premises data collection device known as a “whitebox.” [↑](#footnote-ref-1)
2. MBA website address at: https://www.fcc.gov/general/measuring-broadband-america [↑](#footnote-ref-2)
3. This document identifies the various details of each test unit, including Internet service provider (ISP), technology, service tier, and general location. Each unit represents one volunteer panelist. The unit ID’s were randomly generated, which serves to protect the anonymity of the volunteer panelists. We note that the Profile information is for the Sept/Oct 2015 time period. However, for the purpose of this analysis it provides sufficiently recent data to identify records as to their technology and download speed tier. [↑](#footnote-ref-3)
4. *See* FCC’s Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, Measuring Broadband America, Measuring Fixed Broadband Report -- 2016, <https://www.fcc.gov/general/measuring-broadband-america> [↑](#footnote-ref-4)
5. See FCC’s Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, Raw Data - Measuring Broadband America 2016, <https://www.fcc.gov/reports-research/reports/measuring-broadband-america/raw-data-measuring-broadband-america-2016#block-menu-block-4> [↑](#footnote-ref-5)
6. Total NetUsage numbers do include layer 2 and 3 protocol headers such as IP and Ethernet headers. [↑](#footnote-ref-6)
7. Occurs when the usage numbers reported for active testing exceed the WAN total usage. [↑](#footnote-ref-7)
8. Dec 2015 NetUsage file is missing dates 29 through 31, while Jun 2016 file is missing dates 24 through 30. Dates are inclusive. [↑](#footnote-ref-8)
9. There might be seasonality in monthly data usage. However, we would need more time points of data to examine the seasonality. [↑](#footnote-ref-9)
10. Deviance is a generalized quality-of-fit statistic of models. The null deviance shows how well the response is predicted by an intercept-only model (null model). The residual deviance shows how well the response is predicted by the model when the predictors are included. Small null deviance indicates the null model explains the data better than large null deviance, and likewise with residual deviance. [↑](#footnote-ref-10)