**2022 Urban Rate Survey – Fixed Broadband Service**

## Introduction

Every year, the Wireline Competition Bureau (Bureau) and the Office of Economics and Analytics (OEA) (together, Bureau/OEA) conduct the fixed broadband Urban Rate Survey (broadband URS). The broadband URS collects data on rates for standalone Internet access service charged by a representative sample of fixed broadband providers in urban Census tracts[1] in the United States.

The main purpose of the broadband URS is to produce reasonable broadband comparability benchmarks for every possible service tier (i.e., a service plan with specified minimum download speed, minimum upload speed, and monthly capacity allowance). These benchmarks serve as rate caps to "help ensure that universal service support recipients offering [fixed voice and] broadband services do so at reasonably comparable rates to those in urban areas."[2]

To calculate these benchmarks, the Bureau/OEA have, over the years, used essentially the same sample design, data collection mode, edit checks, and estimation methodology originally adopted in 2013. In particular, the same fixed sample size of approximately 500 sampling units had been selected and the same benchmark definition had been used every year.

To account for the fact that the Bureau/OEA must calculate benchmarks for a much wider range of speeds than when the broadband URS was first initiated, for the 2022 broadband URS, the Bureau/OEA have increased the sample size to about 2,000 sampling units. Correspondingly, we have slightly modified the sample design and estimation methodology to capture the variation in broadband service rates across the United States more adequately, and thereby improve the quality of the benchmark estimates.

This methodology report follows the format of previous years' reports and describes how the Bureau/OEA calculated the fixed broadband reasonable comparability benchmarks for 2022.

## Sample Design

### *Primary sampling unit and sampling frame*

The 2022 broadband URS retains the same definition of primary sampling unit (PSU) as had been used in past survey cycles. That is, a PSU is a pair consisting of a broadband service provider and an urban Census tract where the provider offers at least one terrestrial fixed broadband service tier to residential customers therein. In rare cases where this pair is distinguishable based on the provider's designation as both an incumbent local exchange carrier (ILEC) and a non-ILEC in the Census tract, or the availability of both fixed wired and wireless service options, the PSU definition accommodates these distinctions.

As in previous years, the Bureau/OEA developed the sampling frame[3] for the 2022 broadband URS based on data from the FCC Form 477, as of December of the year prior to data collection. The 2022 broadband URS frame consists of 190,536 PSUs, encompassing 1,600 service providers and 58,224 Census tracts.

---

[1] Census tracts with at least one populated block located within an urban area or urban cluster that is also located within a county designated as a metropolitan statistical area.

[2] *Connect America Fund*, WC Docket No. 10-90, Order, 28 FCC Rcd 4242 (WCB/WTB 2013).

[3] A sampling frame, or frame for short, is a list of all primary sampling units.

***Stratification***

The broadband URS uses a stratified sample design. Stratification is the division of a heterogenous population (represented by the sampling frame), into subpopulations called strata (singular: stratum), each of which is internally homogenous with respect to the population characteristic(s) of interest. When properly implemented, this commonly used sample design element can produce gains in precision in the estimates of characteristics of the whole population.[4]

For the past several years, the Bureau/OEA has stratified the broadband URS sampling frame based on combinations of the following five major factors:

1. Continental United States[5] versus Alaska;
2. Provider's affiliated holding company;
3. Low- versus high bandwidth, where the distinction lies in whether the provider's service tier(s) in the Census tract has minimum download speed lower than, or at least equal to, 500 Megabits per second (Mbps);
4. Providers of terrestrial fixed wireless (TFW) versus non-TFW (i.e., wired) service; and
5. "Major" versus "Minor" providers, where the distinction is algorithmically determined by a clustering method that captures dissimilarities in the number of occupied housing units to which the providers offer service.

Because of the significant increase in sample size, the Bureau/OEA has stratified the frame slightly differently for the 2022 broadband URS. Specifically, the Bureau/OEA has implemented the following modifications:

- More granular division of the range of download speeds based on the following cut points, in Mbps: 2, 36, 75, 115, 155, 250, 300, 500, 750, and 1,000

  This modification is consistent with the design proposal that the Bureau/OEA submitted to the Office of Management and Budget in 2019 in support of its application to increase the broadband URS sample size from the usual 500 to up to 2,000. The goal of this design modification is "to reduce the number of survey responses requested … while still capturing the [variation in] offered rates in accordance with their estimated effect on the reasonable comparability benchmark."[6] This granular division does not apply to holding company-based strata, which retained the low- versus high bandwidth distinction.

  If not enough sampling units in the frame are available to construct separate strata based on this more granular scheme, the Bureau/OEA collapsed the strata appropriately.

- Removal of the "Major" versus "Minor" strata

  Bureau/OEA staff's analysis of historical broadband URS sampling frame data suggests that carving out separate strata based on the number of occupied housing units introduces design redundancy while contributing little, if anything, to achieve the goals of stratification. This is because the sample design uses this number to assign a measure of size to each sampling unit

[4] William G. Cochran, Sampling Techniques ch. 5 (3rd ed. 1977).

[5] All 50 U.S. states except Alaska, together with the District of Columbia and Puerto Rico.

[6] Supporting Statement, 2020 Urban Rate Study Statistical Methodology Part B, OMB Control Number 3060-1192 (Dec. 2019) *available at* https://omb.report/icr/201911-3060-013/doc/97437001.

before selecting the sample (as described in more detail below). As a result of this finding, the Bureau/OEA decided not to define strata based on the number of occupied housing units.

These modifications resulted in the formation of a total of 45 strata for the 2022 broadband URS. The table on the next page shows summary information on these strata.

By comparison, the 2021 broadband URS had 27 strata.

***Sample Allocation***

Consistent with sampling procedures used in prior survey cycles, the Bureau/OEA carefully mapped the 45 strata for the 2022 broadband URS onto the 27 strata from the 2021 cycle. This mapping is required to generate stratum-level estimates of the variance of fixed broadband service rates. These variance estimates, in turn, are essential ingredients for calculating how many sampling units to select from each of the 45 strata; i.e., for allocating the fixed sample size of up to 2,000 sampling units among these strata.

For the broadband URS, the Bureau/OEA uses optimal allocation[7] to produce sample counts that vary according to the estimated rate variance: the higher the estimated variance in a stratum, the more sampling units are allocated for selection from that stratum. The table on the next page includes information on the sample allocation for the 2022 broadband URS.

***Measure of Size and Sample Selection***

The broadband URS implements probability sampling, which means that every sampling unit has *some* chance of being selected in the sample. However, it does not use equal probability sampling, where every sampling unit has an *equal* chance of selection. Instead, the broadband URS sample design calculates a measure of size (MOS) for every sampling unit in the frame, and selects the sample independently within each stratum based on this MOS. Thus, for example, if sampling unit A has a MOS that is twice that of sampling unit B, then A is twice as likely to be selected in the sample compared to B. This type of unequal probability selection is called probability proportional to size (PPS) sampling.[8]

Following historical procedures, the Bureau/OEA calculated the MOS for the 2022 broadband URS sampling units (which, as described above, are pairs of provider-Census tract) by estimating the provider's number of potential subscribers in the Census tract.

As before, the Bureau/OEA calculated this MOS differently for TFW and non-TFW providers. This is because the number of potential customers of TFW services is limited by geographic and technological factors. Many TFW providers serve suburban areas that are of moderate population density. Also, the number of housing units in these areas is likely higher than TFW providers have capacity to serve.

Accordingly, for TFW providers, the Bureau/OEA calculated the number of potential subscribers as follows:

*Number of potential subscribers = 2 x (Number of residential subscribers in the sampling unit's Census tract)* [9]

For non-TFW providers, the Bureau/OEA calculated the number of potential subscribers as follows:

---

[7] Cochran, *supra* note 4 at 96-99.

[8] *Id*. at 251.

[9] The factor of 2 appearing in this calculation is based on the assumption that TFW providers could no more than double their number of existing residential customers within a few months.

*Number of potential subscribers = Provider Presence Ratio x (Number of households in the sampling unit's Census tract),*

where *Provider Presence Ratio* is the fraction of housing units in the Census tract for which the provider reported service availability in its December 2020 Form 477 filing.

In either case, the number of potential subscribers may not exceed the number of households in the sampling unit's Census tract.

After completing the stratification, sample allocation, and measure of size calculation steps, the Bureau/OEA selected the final sample using the SAS SURVEYSELECT Procedure with systematic PPS selection option.[10] The procedure selected a total of 1,995 sampling units.

The table below summarizes the sample design for the 2022 broadband URS.

Main strata groups with one or two strata are holding company-based strata that retained the low- versus high bandwidth distinction from previous survey cycles. Frontier does not have a high bandwidth stratum. América Móvil and Radiate are new holding company-based strata.

The Non-TFW main strata group consists of ten strata defined by the previously identified cut points that divide the range of download speeds. In the TFW main strata group, the Bureau/OEA collapsed three of the strata into one, due to a small number of sampling units. In the Alaska main strata group, there is even more collapsing of strata.

| Main Strata Group | Number of Strata | Frame | | | | Sample | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Sampling Units | Providers | Census Tracts | Number of Potential Subscribers | Sampling Units | Providers | Census Tracts | Number of Potential Subscribers |
| AT&T | 2 | 28,712 | 9 | 27,571 | 41,945,830 | 40 | 9 | 40 | 74,004 |
| Lumen[11] | 2 | 9,189 | 1 | 9,171 | 14,524,644 | 38 | 1 | 38 | 67,811 |
| Charter | 2 | 21,148 | 1 | 21,148 | 33,557,727 | 33 | 1 | 33 | 62,603 |
| Comcast | 2 | 25,739 | 1 | 25,739 | 40,748,545 | 80 | 1 | 80 | 155,265 |
| Cox | 2 | 5,283 | 1 | 5,283 | 7,94,5228 | 13 | 1 | 13 | 28,228 |
| Altice[12] | 2 | 3,902 | 2 | 3,902 | 5,700,903 | 20 | 2 | 20 | 40,427 |
| Frontier | 1 | 6,283 | 1 | 6,264 | 9,344,119 | 19 | 1 | 19 | 35,647 |
| Verizon | 2 | 13,790 | 9 | 13,776 | 20,193,202 | 65 | 7 | 65 | 115,807 |
| WideOpenWest | 2 | 2,227 | 6 | 2,225 | 2,783,380 | 13 | 4 | 13 | 21,917 |
| Windstream | 2 | 1,513 | 35 | 1,384 | 1,625,533 | 20 | 18 | 20 | 30,517 |
| América Móvil | 2 | 825 | 1 | 825 | 1,235,052 | 75 | 1 | 75 | 131,490 |
| Radiate | 2 | 2,452 | 8 | 2,452 | 3,109,345 | 48 | 8 | 48 | 86,432 |
| Non-TFW | 10 | 18,477 | 932 | 14,669 | 18,433,982 | 1,144 | 263 | 1,112 | 1,938,025 |
| TFW | 7 | 50,753 | 811 | 36,202 | 1,439,008 | 328 | 175 | 317 | 145,026 |
| Alaska | 5 | 243 | 7 | 87 | 323,975 | 59 | 4 | 47 | 115,007 |
| **Overall** | **45** | **190,536** | **1,600** | **58,224** | **202,910,473** | **1,995** | **485** | **1,906** | **3,048,206** |

[10] The SURVEYSELECT Procedure, SAS User's Guide, https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/statug/statug_surveyselect_toc.htm.

[11] Formerly CenturyLink.

[12] Formerly CSC Holdings.

**Survey Response**

This section describes how the sample of 1,995 sampling units responded to the 2022 broadband URS.

This year, there are four ineligible sampling units because the selected provider did not offer, or stopped offering, fixed broadband service in the selected Census tract. Of the remaining 1,991, 158 did not respond to the survey. Thus, the overall response rate is 92% ((1,991 - 158) ÷ 1,991, times 100%).

The table below shows the number of responses, the number of different service providers, and the number of different Census tracts requested and received at the close of data collection for the 2022 broadband URS.

| Survey Status | Responses | Service Providers | Census Tracts |
|---|---|---|---|
| Requested | 1,995 | 485 | 1,906 |
| Received | 1,833 | 432 | 1,757 |

After conducting edit checks, including the removal of submitted rate data for business instead of residential plans, a total of 1,829 responses had useable unique monthly rates. Monthly rates were treated as unique for a combination of Census tract, FCC Registration Number (FRN), service name, technology, download speed, upload speed, and capacity allowance. A total of 10,149 unique monthly rates were used to estimate the 2022 broadband comparability benchmarks.

The next table shows summary information on how these 10,149 unique monthly rates distribute by technology.

| Technology | Responses | Service Providers | Census Tracts | Rates |
|---|---|---|---|---|
| Cable | 634 | 100 | 631 | 3,238 |
| DSL | 632 | 94 | 626 | 3,130 |
| Fixed Wireless | 337 | 169 | 324 | 1,405 |
| FTTH | 649 | 207 | 641 | 2,376 |
| **Total** | **1,829** | **431** | **1,753** | **10,149** |

**Monthly Rates and Rate Spreads**

The main analysis variable for the broadband URS is the monthly rate which broadband providers charge their customers in urban Census tracts. It is common, however, for providers to offer multiple service tiers in the same Census tract at different monthly rates. For this reason, the survey asks for the minimum and maximum of these rates and calculates an "average" monthly rate based on these two extreme values. Specifically, the following equations are used to calculate this average monthly rate, if the service provider offered multiple rates in the Census tract:

- Minimum Rate = Minimum Monthly Charge + Minimum Other Mandatory Charge + Minimum Surcharge
- Maximum Rate = Maximum Monthly Charge + Maximum Other Mandatory Charge + Maximum Surcharge
- Rate Spread = Maximum Rate - Minimum Rate

- Average Rate = (Minimum Rate + Maximum Rate)/2

The following equations were used if the service provider did not offer multiple rates in the Census tract:

- Average Rate = Minimum Monthly Charge + Minimum Other Mandatory Charge + Minimum Surcharge
- Rate Spread = 0

**Weights**

The broadband URS uses weights to ensure the contributions of each response properly represent the offers that consumers possibly receive nationwide. Weights are also used to ensure that a service provider's rates do not exert extra influence on the estimate only because the provider offers different services using multiple technologies.

The 2022 broadband URS weight construction is consistent with the method used in previous years. That is, each rate was assigned a weight based on the following equation:

*Weight = Sampling Weight x Nonresponse Weight x Same Rate Weight x Service Level Weight x Number of Potential Subscribers*

*Sampling Weight* is the inverse of the selection probability for each sample unit. The selection probability is determined by the total number of units in each stratum, the sample size in each stratum, and the units' number of potential subscribers described in the sample selection section earlier. Each sample is assigned a sampling weight to reflect its selection probability.

*Nonresponse Weight* is assigned to each stratum in order to compensate for unit nonresponse in each stratum. It is the total number of potential subscribers sampled over the total number of potential subscribers in the sampled Census tracts of a given provider who has provided rate responses in each stratum.

*Same Rate Weight* is assigned to the respondents who provided i) multiple service levels or ii) equal service levels via different technologies for the same rate in the same Census tract.[13] In such cases, the rate was assigned a Same Rate Weight equal to 1/R, where R is the number of rate responses provided by a service provider at the same rate in the Census tract.

*Service Level Weight* is assigned to the respondents who provided multiple rates for the same service level offered via different technologies and/or service names. Each rate was assigned a Service Level Weight equal to 1/L, where L is the number of responses with different rates provided by a service provider for the same service plan (same download bandwidth, upload bandwidth, and monthly capacity allowance) in the Census tract.

*Number of Potential Subscribers* is the estimated number of potential customers to whom the providers advertise their service.

The final weight is the product of Sampling Weight, Nonresponse Weight, Same Rate Weight, Service Level Weight, and the Number of Potential Subscribers.

[13] Such a situation could arise when a provider uses different technologies to provide similar services to customers in different parts of a Census tract.

**Average Rate Model**

The 2022 URS shows that broadband rate is nonlinear in proportion to download speed and upload speed (see Appendix A). To estimate an average rate for every possible combination, we applied a weighted Generalized Boosted Model (GBM),[14] which is an algorithm allowing nonlinearity in our estimation,[15] to all terrestrial fixed broadband services with download speeds between 2 Mbps and 1 Gigabit per second (Gbps), inclusive.[16]

The 10,149 rates used in the final analysis ranged from $9.95 to $1,289.00, with a weighted standard deviation ranging from $21.73 to $96.15. The rates vary widely across technologies. The following table shows the range, weighted mean, and weighted standard deviation of these rates, as well as the weighted mean download speed for different technologies.

| Technology | Min | Max | Rate, weighted mean | Rate, weighted standard deviation | Download speed, weighted mean |
|---|---|---|---|---|---|
| Cable | 9.95 | 500.00 | 82.64 | 35.32 | 409.97 |
| DSL | 13.99 | 190.00 | 60.69 | 21.73 | 42.53 |
| Fixed Wireless | 10.00 | 1,289.00 | 92.12 | 96.15 | 51.63 |
| FTTH | 14.99 | 549.99 | 77.17 | 29.99 | 615.56 |
| **All** | **9.95** | **1,289.00** | **78.68** | **34.51** | **412.32** |

We undertook a weighted GBM[17] based on the following form:[18]

$$\text{Average Monthly Rate (\$)} = Y = f(D\ , U\ , A, ST)$$

---

[14] See Appendix B.

[15] Ideally, we would calculate directly the weighted means and the weighted standard deviations of rates for all services. However, our samples do not cover all possible combinations of services provided to consumers nationwide. Therefore, we use a statistical model to estimate rates for all possible services.

[16] The 2018 broadband average rate model was the first year to include data with download bandwidths between 2 and 1000 Mbps. The 2017 broadband linear regression only models average rate between 2 and 50 Mbps.

[17] The average rate model based on a weighted GBM for the 2022 URS allows nonlinearity in rate per download bandwidth and rate per upload bandwidth by stratum groups. For further information, see Appendix B.

[18] We used the R package "gbm: Generalized Boosted Regression Models" to perform model fitting. We used random 50% of data as training set and 50% of data as validation set for each regression tree phase. Multiple GBM models were constructed and compared. Our final model was selected based on root mean square errors. The optimal number of trees of our final model is 17,095 based on the out-of-bag error statistic, which is a method of measuring the prediction error of boosted decision trees.

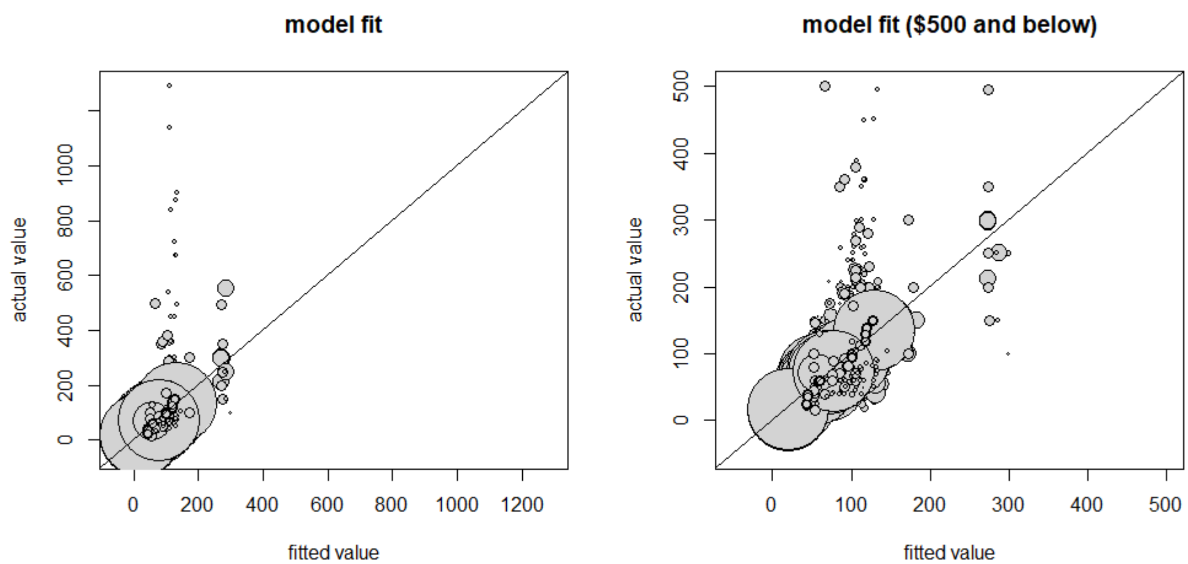
where *D* is download speed in Mbps, *U* is upload speed in Mbps, and *A* is the inverse of monthly capacity allowance in GB. *ST* includes the 15 main strata groups as shown in the sample design summary table on page 4. The average monthly rate estimate is a function of D, U, A, and ST.

We estimated the U.S. average monthly rate as:

$$\text{U.S. Average Monthly Rate (\$)} = \sum_{i=1}^{n} \gamma_i \, \mathrm{E}(\mathrm{Y} \mid \mathrm{D}, \mathrm{U}, \mathrm{A}, \mathrm{ST} = \mathrm{ST}_i)$$

where $n = 14$, which represents 14 main strata groups in the continental U.S. $E(Y| D, U, A, ST = ST_i)$ is the expected value conditioned on combinations of download speed, upload speed, and capacity allowance for a given main strata group. The $\gamma_i$ is the proportion of total continental U.S. potential subscribers in a given main strata group.

The plots below show how the model fits the raw data. The closer the dots are to the 45-degree line, the better the fit. The size of the circles represents the weights of the sample rates.




**U.S. reasonable comparability benchmark**

Under the methodology previously adopted by the Bureau, the reasonable comparability benchmark is the estimated average monthly rate plus twice the standard deviation of rates for terrestrial fixed broadband service plans with download bandwidths of 10 Mbps or greater, upload bandwidths of 1 Mbps or greater, and meeting or exceeding the minimum monthly usage allowance. The root weighted mean squared residual (RWMSR) is an estimate of the standard deviation of rates for service plans meeting the reasonable comparability benchmark criteria.[19]

The 2022 broadband URS average rate model approximates rate per download bandwidth and upload bandwidth closely. Therefore, the RWMSR of rates does not show a trend by download bandwidth and upload bandwidth. As before, we calculated the RWMSR values separately for the continental U.S. and Alaska. The table below shows the final RWMSR values.

[19] RWMSR is the square root of the weighted average of the square of residuals (observed rate minus average rate as defined by the Average Monthly Rate equation) plus the square of the spreads divided by 12.

| | RWMSR |
|---|---|
| Continental U.S. | 16.37 |
| Alaska | 19.76 |

Following the definition adopted by the Bureau, the U.S. reasonable comparability benchmark is calculated as follows:

$$\text{U.S. reasonable comparability benchmark (\$)} = \text{U.S. Average Monthly Rate} + 2\,(\text{RWMSR}_{\text{ContinentalUS}})$$

$$= \text{U.S. Average Monthly Rate} + 32.75$$

The U.S. average monthly rate estimator is described in the previous section.

**Alaska reasonable comparability benchmark**

For the Alaska reasonable comparability benchmark, the average monthly rate model is defined as follows:

$$\text{AK Average Monthly Rate (\$)} = \sum_{j=1}^{m} \gamma_j \, \mathrm{E}(\mathrm{Y} \mid \mathrm{D}, \mathrm{U}, \mathrm{A}, \mathrm{ST} = \mathrm{ST}_j)$$

where $m = 5$, which represents the 5 strata in Alaska. $E(Y| D, U, A, ST = ST_j)$ is the expected value conditioned on combinations of download speed, upload speed, and monthly capacity allowance for a given stratum in Alaska. The $\gamma_j$ is the proportion of total Alaska potential subscribers in a given Alaska stratum.

The AK reasonable comparability benchmark is the Alaska average monthly rate plus two times its RWMSR:

$$\text{AK reasonable comparability benchmark (\$)} = \text{AK Average Monthly Rate} + 2\,(\text{RWMSR}_{\text{Alaska}})$$

$$= \text{AK Average Monthly Rate} + 39.51$$

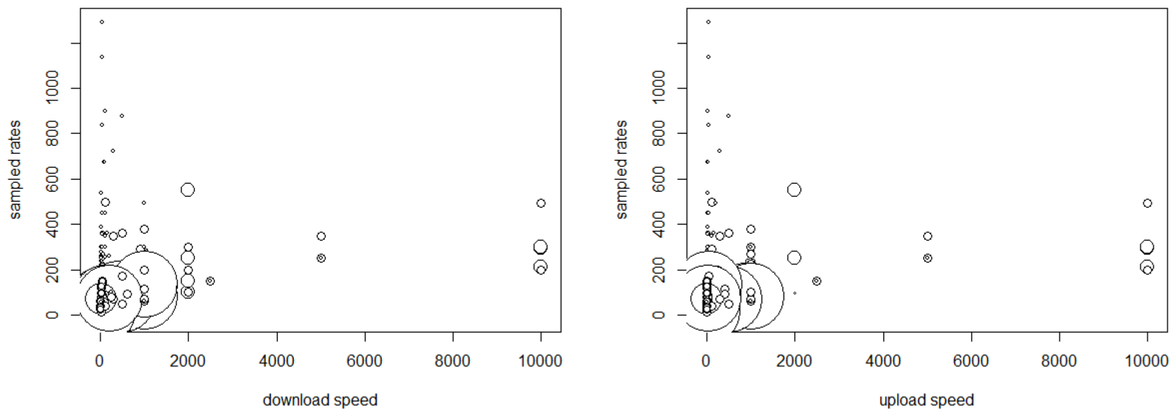
**Reasonable comparability benchmark results**

The table on the next page provides examples of reasonable comparability benchmarks (rounded up to the nearest cent) for several service plan levels. The estimates are available for a reasonable comparability benchmark for lower download speeds (greater than or equal to 4 Mbps) if needed and up to download speeds of 1,000 Mbps. Upload speed may not exceed download speed.

| Download Bandwidth (Mbps) | Upload Bandwidth (Mbps) | Capacity Allowance (GB/mo) | 2022 U.S. Benchmark | 2022 AK Benchmark |
|---|---|---|---|---|
| 4 | 1 | 500 | $65.37 | $87.03 |
| 4 | 1 | Unlimited | $70.21 | $96.56 |
| 10 | 1 | 500 | $73.12 | $115.17 |
| 10 | 1 | Unlimited | $77.97 | $124.70 |
| 25 | 3 | 500 | $71.09 | $121.63 |
| 25 | 3 | Unlimited | $75.93 | $131.16 |
| 50 | 5 | 500 | $77.96 | $126.85 |
| 50 | 5 | Unlimited | $82.96 | $136.53 |
| 100 | 20 | 500 | $100.74 | $147.18 |
| 100 | 20 | Unlimited | $105.67 | $156.79 |
| 1000 | 500 | 500 | $131.33 | $180.64 |
| 1000 | 500 | Unlimited | $134.62 | $188.61 |
| 25 | 5 | 500 | $71.07 | $121.63 |
| 25 | 5 | Unlimited | $75.91 | $131.16 |
| 100 | 10 | Unlimited | $105.21 | $156.34 |
| 250 | 25 | Unlimited | $120.12 | $175.59 |
| 500 | 50 | Unlimited | $124.20 | $181.21 |
| 1000 | 100 | Unlimited | $134.65 | $188.67 |

## APPENDIX A

The 2022 URS modeled rates by download speed and by upload speed. Over this large range, the rates are not linear functions of either quantity. The size of the circles in the plots below represents the weights of the sample rates. Sampled rates represent common services provided to the customers and do not include all possible combinations of download bandwidth, upload bandwidth, and monthly capacity allowance.



## APPENDIX B

A Generalized Boosted Model (GBM) is a machine learning algorithm that combines regression trees and gradient boosting techniques. The GBM framework does not assume a specific pattern between the independent variables and the dependent variable. It allows nonlinearity and interactions without the need to define complex mathematical equations.

The algorithm first selects a portion of data to “train” a regression tree model (regression tree phase). The regression tree model used in GBM is usually a stump-only model or with only very few branches. Then, it uses the unselected data to “validate” the model and output a user defined performance statistic or loss function (validation phase). The algorithm repeats the same procedure on the residuals from the previous modeling phases until the performance gain stabilizes or loss function optimizes (gradient boosting phase). The outputs of a GBM are model fits from a series of regression tree models. Therefore, conventional coefficients are not applicable. Independent variable collinearity and data outliers have very little impact on the model fit because only the most influential variables are selected during each regression tree phase (only one most influential variable is selected if fitting a stump-only model). The interactions are naturally embedded in the structure of a series of regression tree models. Overfitting is safeguarded by inserting a cross-validation technique. Therefore, the GBM algorithm is considered to have high predictive accuracy. However, its predictive performance is weakened when the relationship between an independent variable and the dependent variable is very linear. More information about GBM can be found in the following references:

Y. Freund and R.E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 55(1):119-139.

G. Ridgeway. 1999. The state of boosting. Computing Science and Statistics. 31:172-181.

J.H. Friedman, T. Hastie, and R. Tibshirani. 2000. Additive Logistic Regression: a Statistical View of Boosting. Annals of Statistics. 28(2):337-374.

J.H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics. 29(5):1189-1232.

J.H. Friedman. 2002. Stochastic Gradient Boosting. Computational Statistics and Data Analysis. 38(4):367-378.