2021 Urban Rate Survey – Fixed Broadband Service Analysis

Introduction

Every year, the Wireline Competition Bureau (Bureau) conducts a survey of residential standalone Internet access service rates "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."[1] The Bureau adopted the general methodology for surveying terrestrial fixed broadband providers in 2013. The 2021 Urban Rate Survey (URS) for fixed broadband services follows the same methodology as the 2020 survey. This document shows how the fixed broadband reasonable comparability benchmark was calculated for 2021, including differences in the data received in the survey and changes from the analysis used for the 2020 data. As explained below, the 2021 reasonable comparability benchmarks calculated for fixed broadband services for the non-Alaskan portion of the United States are similar to the 2020 reasonable comparability benchmarks.[2]

Sample Design and Selection

The sampling unit for the 2021 fixed broadband survey was a (service provider, census tract) pair. The frame[3] for the 2021 URS was the set of sampling units of providers offering terrestrial fixed broadband service to residential customers in urban census tracts based on FCC Form 477 December 2019 data. The frame consisted of 155,805 sampling units, encompassing 1,430 service providers and 58,149 census tracts. The likelihood of a (service provider, census tract) pair receiving a survey is based on the number of potential subscribers for that provider in that census tract.

For each sampling unit other than the terrestrial fixed wireless providers, the number of potential subscribers is calculated as:

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

Provider Presence Ratio was calculated as the fraction of housing units in the census tract for which the provider reported service availability via Form 477.

We calculated the number of potential subscribers differently for the terrestrial fixed wireless providers because the number of potential customers for such services is limited by geographic and technological factors. Many terrestrial fixed wireless providers serve suburban areas that are of moderate population density. Also, the number of housing units in these areas is likely higher than fixed wireless providers have capacity to serve. Accordingly, for each sampling unit of these providers, the number of potential subscribers is calculated as:[4]

---

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

[2] The 2021 reasonable comparability benchmarks for fixed broadband services in Alaska are lower than the 2019 reasonable comparability benchmarks because the Alaska data is subject to high variability due to the small number of companies in the survey. Variation in standard deviation and average rate over time from 2018-2021 is likely due to an insufficient sample size. Commission staff is looking to address this issue.

[3] A frame is an inventory that lists all sampling units from which we select our samples.

[4] The number of potential subscribers for the terrestrial fixed wireless providers is calculated as 2 x number of residential subscribers, assuming that such providers could no more than double their number of existing residential customers within a few months.

Number of potential subscribers = 2 x (Number of residential subscribers in the sampling unit's census tract).

The number of potential subscribers was not allowed to exceed the number of households in the sampling unit's census tract.

The 2021 URS follows the stratification of the 2019 URS. The frame was divided into strata to account for the differing rate variability in each stratum.

The strata included in the 2021 URS are listed below. There are 27 strata: 13 strata for services with download bandwidth less than 500 Mbps, 12 strata for services with download bandwidth greater than or equal to 500 Mbps (high bandwidth strata),[5] and two Alaska strata.

- Service download bandwidth < 500 Mbps
    - AT&T (AT&T Services, Inc.)
    - CenturyLink (CenturyLink, Inc., CenturyLink Communications, LLC)
    - Charter (Charter Communications, Inc.)[6]
    - Comcast (Comcast Cable Communications, Inc.)
    - Cox (Cox Communications)
    - CSC Holdings (CSC Holdings LLC)
    - Frontier (Frontier Communications Corporation)
    - Verizon (Verizon New York Inc., Verizon Pennsylvania LLC, Verizon New Jersey Inc., Verizon California Inc., Verizon New England Inc., Verizon Virginia LLC, Verizon Maryland LLC, Verizon Florida LLC, Verizon Delaware LLC, GTE Southwest Incorporated dba Verizon Southwest, Verizon Washington, DC Inc.)
    - WideOpenWest (Knology, WideOpenWest, and Wiregrass Telcom)
    - Windstream (service providers identifying Windstream as their holding company)
    - Terrestrial fixed wireless providers
    - Major[7]
    - Minor
- Service download bandwidth ≥500 Mbps (high bandwidth strata)
    - AT&T
    - CenturyLink
    - Charter
    - Comcast
    - Cox
    - CSC Holdings

[5] There were 10 high bandwidth strata in last year's URS: AT&T, CenturyLink, Comcast, Cox, Verizon, WideOpenWest, Windstream, Terrestrial Fixed Wireless, Major, and Minor.

[6] Bright House Networks, LLC, Time Warner Cable Inc., and Charter Communications, Inc. have merged and operate as Charter Communications, Inc.

[7] The Major and Minor strata are divided based on the number of potential subscribers, the number of occupied housing units to which the provider offers service, and the Provider Presence Ratio. The algorithm used to divide the sampling units into the Major and Minor strata is "Partitioning Around Medoids." Partitioning Around Medoids is a type of cluster analysis to identify data clusters based on dissimilarities between clusters. Medoids are the medians for multi-dimensional data. *See* Kaufman, L. and Rousseeuw, P.J. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York; Park, H.S. and C.H. Jun. 2009. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications. 36(2):3336–3341.

    - Verizon
    - WideOpenWest
    - Windstream
    - Terrestrial Fixed Wireless
    - Major
    - Minor
- Alaska
    - Fixed wireline services
    - Terrestrial fixed wireless

The table below presents the sampling plan including the sample size for each stratum. Sampling units were selected randomly from each stratum, with unequal selection probability proportional to providers' number of potential subscribers in a given tract.[8] The sample sizes for each stratum are a reflection of the estimated number of potential subscribers in the stratum and the estimated variability of offered rates from last year's URS.

[8] The selection probability of a Probability Proportional to Size (PPS) sampling is a function of measure of "size." Measure of size is the number of potential subscribers from a provider in a given tract. The selection probability is higher for sampling units with higher number of potential subscribers.

	Frame				Sample			
Stratum	Units	Providers	Census Tracts	Offers	Units	Providers	Census Tracts	Offers
Overall	155805	1430	58149	199118081	500	184	496	739296
AT&T Services, Inc.	8534	9	8534	10924167	10	7	10	18607
CenturyLink	3771	1	3771	5211921	5	1	5	9312
Charter	29	1	29	40672	5	1	5	9860
Comcast Cable Communications, LLC	443	1	443	520392	5	1	5	7406
Cox	31	1	31	45530	5	1	5	7816
CSC Holdings LLC	490	1	490	480558	5	1	5	6201
Frontier	6775	1	6775	10180510	42	1	42	79934
Verizon	3328	8	3328	4877132	5	5	5	9758
WideOpenWest	44	4	43	60576	5	3	5	11069
Windstream	903	36	845	413225	5	5	5	4176
Terrestrial Fixed Wireless	17483	684	11528	954332	69	56	68	27422
Major	2363	205	2024	4090783	12	10	12	22588
Minor	6178	510	5480	1796671	24	17	24	15104
AT&T Services, Inc. (high bandwidth)	18891	9	18891	31045414	5	4	5	11210
CenturyLink (high bandwidth)	5411	1	5411	9309624	5	1	5	9556
Charter (high bandwidth)	21068	1	21068	32935061	5	1	5	8202
Comcast Cable Communications, LLC (high bandwidth)	26176	1	26176	39510832	111	1	111	203420
Cox (high bandwidth)	5241	1	5241	7646862	5	1	5	9911
CSC Holdings LLC (high bandwidth)	2362	1	2362	3590382	5	1	5	7234
Frontier (high bandwidth)	9751	8	9751	15251566	5	4	5	8662
Verizon (high bandwidth)	2107	2	2107	2720966	5	2	5	7507
WideOpenWest (high bandwidth)	745	28	722	1204998	5	5	5	8258
Windstream (high bandwidth)	635	25	618	82890	5	3	5	5347
Terrestrial Fixed Wireless (high bandwidth)	7234	278	6669	13252443	89	57	89	183340
Major (high bandwidth)	5578	410	5086	2650648	48	42	48	33215
Minor (high bandwidth)	175	4	87	317762	5	3	5	13015
Alaska	59	3	55	2164	5	2	5	1166
Alaska TFW	155805	1430	58149	199118081	500	184	496	739296

Survey Response

The table below presents the number of responses, the number of different service providers, and the number of different census tracts requested, received, and received with rates (service provided) in the 2021 URS for fixed broadband service.

Survey Status	Responses	Service Providers	Census Tracts
Requested	500	184	496
Received	480	171	476
Service Provided	476	167	472

The next table presents the number of responses, the number of different service providers, the number of different census tracts, and the number of rates for each technology among responses received with rates as of June 30, 2020 for the 2021 benchmark.

Technology	Responses	Service Providers	Census Tracts	Rates
Cable	257	59	256	1492
DSL	118	38	118	890
Fixed Wireless	75	54	74	330
FTTH[9]	108	64	108	499

A total of 3,211 rates were provided at a variety of service levels as of September 2020 for the 2021 benchmark. Several rates were excluded from the analysis due to business plans being reported rather than residential, resulting in a total of 3,203 rates available for the analysis. The table below presents the number of responses, the number of different service providers, the number of different census tracts, and the number of rates for each technology among responses received with rates available for the analysis.

Technology	Responses	Service Providers	Census Tracts	Rates
Cable	257	59	256	1492
DSL	118	38	118	890
Fixed Wireless	75	54	74	322
FTTH	108	64	108	499
All	475	166	471	3203

Monthly Rates and Rate Spreads

Monthly rates were treated as unique for a combination of census tract, FCC Registration Number (FRN), service name, technology, download bandwidth, upload bandwidth, and capacity allowance. The following average monthly rate was used if the service provider offered multiple rates in the census tract for each unique combination:

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

- Maximum Rate = Maximum Monthly Charge + Maximum Other Mandatory Charge + Maximum Surcharge
- Average Rate = (Minimum Rate + Maximum Rate)/2
- Rate Spread = Maximum Rate - Minimum Rate

The following average monthly rate was 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

Weights are required 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 2021 survey weight construction is consistent with the 2020 survey weight construction. Each rate was assigned a weight:

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.[9] 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.

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

Average Rate Model

The 2021 URS shows that broadband rate is nonlinear in proportion to download bandwidth and upload bandwidth (see Appendix A). To estimate an average rate for every possible bandwidth tier combination, we applied a weighted Generalized Boosted Model (GBM),[10] which is an algorithm allowing nonlinearity in our estimation,[11] to all terrestrial fixed broadband services with download bandwidths between 2 and 1000 Mbps, inclusive.[12]

This sub-sample of the data consisted of 3,203 rates from 475 responses encompassing 166 different providers in 471 different census tracts. The table below presents the number of responses, the number of different service providers, the number of different census tracts, and the number of rates for each technology used for constructing the average rate model.

Technology	Responses	Service Providers	Census Tracts	Rates
Cable	257	59	256	1492
DSL	118	38	118	890
FTTH	75	54	74	322
Fixed Wireless	108	64	108	499
All	475	166	471	3203

The rates in this sub-sample ranged from $14.99 to $599.95 with a weighted standard deviation ranging from $21.85 to $108.22. The rates vary widely across technologies. The following table shows the rate range, the weighted rate mean, the weighted rate standard deviation, and the weighted download bandwidth mean for different technologies in this sub-sample.

[10] The 2018 broadband average rate model was built with Generalized Additive Model (GAM), which is also a machine learning method that allows nonlinearity in estimation. However, GAM dramatically overfits the 2019, 2020, and 2021 URS data, which results in uncomfortable negative average rate estimates for download bandwidth, upload bandwidth, and capacity allowance combinations that do not have samples. This is one of the indications that the current URS sample size may need to be increased.

[11] 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.

[12] 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.

	min	max	Weighted rate mean	Weighted rate standard deviation	Weighted download bandwidth mean
Cable	14.95	599.95	85.78	31.02	371.02
DSL	14.99	199.95	61.36	21.85	36.34
Fixed wireless	25.00	639.95	104.55	108.22	48.55
FTTH	15.00	399.95	69.05	26.05	526.16

We undertook a weighted GBM[13] based on the following form:[14]

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

where D is download bandwidth in Mbps, U is upload bandwidth in Mbps, and A is the inverse of usage allowance in GB. ST includes 15 stratum groups: Alaska, Alaska TFW, AT&T Services, Inc. (all service download bandwidth), CenturyLink (all service download bandwidth), Charter (all service download bandwidth), Comcast Cable Communications, LLC (all service download bandwidth), Cox (all service download bandwidth), CSC Holdings LLC (all service download bandwidth), Frontier (all service download bandwidth), Major (all service download bandwidth), Minor (all service download bandwidth), Terrestrial Fixed Wireless (all service download bandwidth), Verizon (all service download bandwidth), WideOpenWest (all service download bandwidth), and Windstream (all service download bandwidth). The average monthly rate estimate is a function of D, U, A, and ST.

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

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

where n = 13, which represents 13 stratum groups in the continental[15] U.S. $E(Y| D, U, A, ST = ST_i)$ is the expected value conditioned on combinations of download bandwidth, upload bandwidth, and capacity allowance for a given stratum group. The γ_i is the proportion of total continental U.S. potential subscribers in a given stratum group. As of December 2019, the proportion of total continental U.S. potential subscribers in a given stratum group is listed in the table below.

Stratum Group	γ_i
AT&T Services, Inc.	21.08%
CenturyLink	7.29%
Charter	16.56%
Comcast Cable Communications, LLC	20.10%
Cox	3.86%
CSC Holdings LLC	2.04%

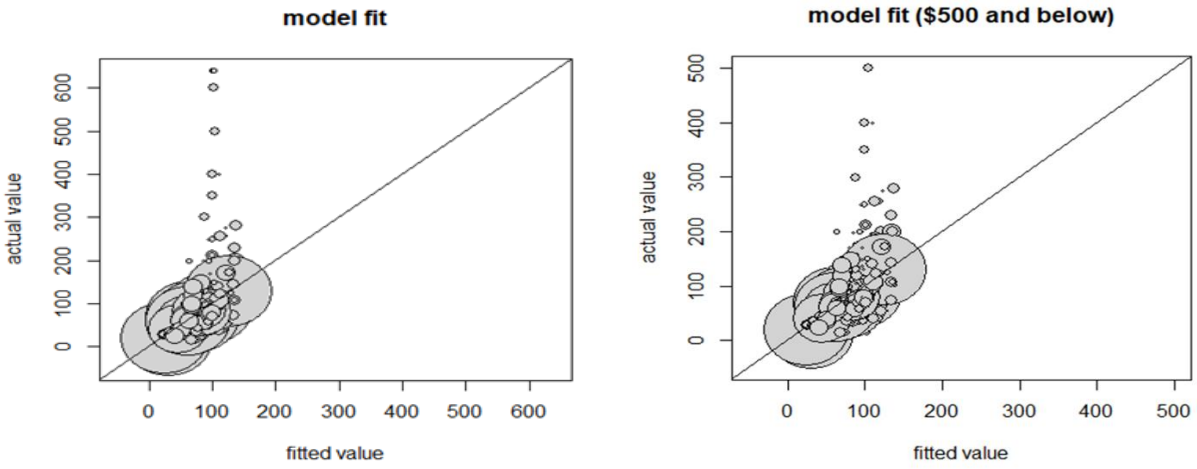
[13] The average rate model based on a weighted GBM for the 2021 URS allows nonlinearity in rate per download bandwidth and rate per upload bandwidth by stratum groups. For further information, see Appendix B.

[14] 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 best performance of root mean square errors. The optimal number of trees of our final model is 14,191 based on 10-fold cross-validation results.

[15] We use the term continental U.S. to mean all U.S. states and territories with the exception of Alaska, for which a separate benchmark is calculated.

Frontier	5.11%
Major	8.71%
Minor	2.23%
Terrestrial Fixed Wireless	0.52%
Verizon	10.11%
WideOpenWest	1.40%
Windstream	0.81%

The plot below shows 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.[16]

The 2021 URS broadband 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. For the 2021 URS, we calculate the RWMSR by Alaska and continental U.S. The table below shows the RWMSR by Alaska and continental U.S.

	RWMSR
Continental U.S.	17.35
Alaska	24.99

[16] 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.

The calculation of the U.S. reasonable comparability benchmark is the following:

U.S. reasonable comparability benchmark ($) =

U.S. Average Monthly Rate + 2 ($RWMSR_{ContinentalUS}$) =

U.S. Average Monthly Rate + 34.70

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:

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

where m = 2, which represents 2 stratum groups in Alaska (Alaska and Alaska TFW). $E(Y| D, U, A, ST = ST_j)$ is the expected value conditioned on combinations of download bandwidth, upload bandwidth, and capacity allowance for a given stratum group in Alaska. The γ_j is the proportion of total Alaska potential subscribers in a given stratum group. As of December 2020, the proportion of total Alaska potential subscribers in a given stratum group is listed in the table below.

Stratum Group	γ_j
Alaska	99.32%
Alaska TFW	0.68%

The AK reasonable comparability benchmark is the Alaska average monthly rate plus two RWMSR as the following:

AK reasonable comparability benchmark ($) =

AK Average Monthly Rate + 2 ($RWMSR_{Alaska}$) =

AK Average Monthly Rate + 49.98

Reasonable comparability benchmark results

The table directly below 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 bandwidths (greater than or equal to 4 Mbps) if needed and up to download bandwidths of 1,000 Mbps.

Download Bandwidth (Mbps)	Upload Bandwidth (Mbps)	Capacity Allowance (GB)	2021 U.S.	2021 AK
4	1	350	$70.69	$107.50
4	1	Unlimited	$75.72	$113.17
10	1	350	$79.51	$119.67
10	1	Unlimited	$85.11	$125.90
25	3	350	$80.97	$127.66
25	3	Unlimited	$86.72	$134.04
25	5	350	$89.13	$138.94
25	5	Unlimited	$94.89	$145.33
50	5	Unlimited	$102.04	$148.34
100	10	Unlimited	$106.20	$152.93
250	25	Unlimited	$125.78	$174.67
500	50	Unlimited	$131.51	$182.24
1000	100	Unlimited	$140.80	$191.20

Constraints

The reasonable comparability benchmark is the estimated average monthly rate plus twice the standard deviation of rates. While the estimated average monthly rate remains relatively unchanged over survey years, the reason for the increase in 2018 and the subsequent decrease in 2019 and 2020 is mainly because the standard deviation of rates is inconsistent. In other words, the survey data contains differing rate variation over survey years.

We have observed shifting in broadband service plans over time. A provider's rate for a given service plan may change along with shifting in the service plans that this provider currently offers to the new customers at a given location. For example, a provider may offer a 100/10 Mbps service plan to new customers in a city with the same rate as for a 25/5 Mbps service plan which was offered previously but is no longer available for new customers. The changes do not happen uniformly, either across providers or across geographic areas. Therefore, the national average monthly rate is not heavily influenced, but the rate variation for service plans may increase or decrease substantially over time.

Additionally, the sample size for a given service plan may change dramatically depending on what providers currently offer. The changes in sample size for a given service plan also increase the year-to-year changes in rate variation over survey years. The differing rate variation in data reflects quick and dynamic changes in the consumer market of fixed broadband services and inadequate sample size to capture all possible services that are offered in the market.

Our rate estimates are based on service plans with different combinations of download bandwidth, upload bandwidth, and monthly capacity allowance. The data collected present several difficulties to build such a model.

1. Not all bandwidth tiers have rate samples.
2. Some bandwidth tiers have very few samples.

3. Download bandwidth and upload bandwidth are correlated in practice. For example, it is common that a carrier provides a 10/1 Mbps service, but it is rare that a carrier would provide a 1000/1 Mbps service. Therefore, we do not have rate samples for services with a high download bandwidth and low upload bandwidth combination.

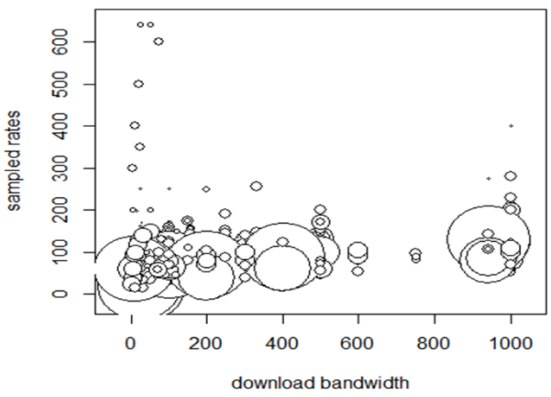
For these reasons, all models are subject to high risk of overfitting. Underlying data are essential for all model building and validation. In addition, we also have observed wider spread of service plan choices and geographic differentiations that have added and will continue to add more variation in our estimates over time.

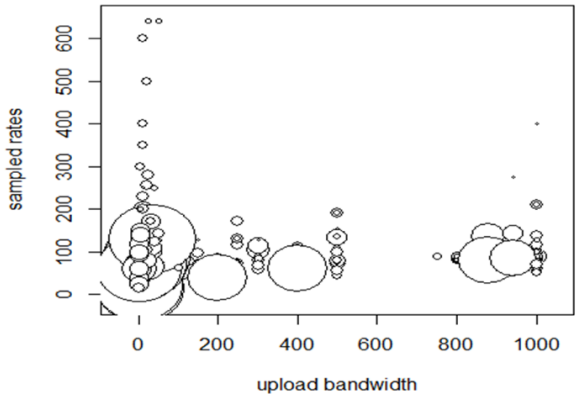
The total number of unique service plan choices on the market with different combinations of download bandwidth, upload bandwidth, and monthly capacity allowance is about 360 in both the 2018 and 2019 surveys, about 340 in 2020, and about 380 in 2021, though there are more high-bandwidth plans in the 2021 data than in previous years. Among the 380 combinations in 2021, 58 were high bandwidth (download bandwidth greater than or equal to 500 Mbps) service plan choices. In the 2018, 2019, and 2020 surveys, there were 30, 40, and 41 high bandwidth service plan choices respectively. The rest of the service plan choices remain relatively steady. In other words, the total number of unique service plans available nationwide is fairly steady, though with an increase in the number of high-bandwidth plans. With increasing changes in unique service plan choices on the market, the 2021 URS received about 617 responses of (service provider, census tract) pairs offering high bandwidth service plan choices, while the 2020 URS received about 417 responses offering high bandwidth service plan choices. In 2019, the URS received only about 407 responses offering high bandwidth service plans. This shows that more service providers provided different service plan choices to more customers, particularly for high-bandwidth services. Although more high bandwidth service plan choices have been offered over time, other service plan choices are still available for consumers. We suspect this trend will continue.

We have also observed that the rate varies based on geographic locations because services are not all equally deployed. Therefore, it is reasonable to see a 4/1 Mbps service that charges more than a 10/1 Mbps service at a different geographic location. Within most strata, it is more likely to see the rate as a systematic function of download bandwidth, upload bandwidth, and monthly capacity allowance. However, this relationship does not hold at a national level because of the geographic differentiations.

APPENDIX A

The 2021 URS modeled rates by download bandwidth and by upload bandwidth. Over this large range of bandwidths, the rates are not linear functions of download bandwidth and upload bandwidth. 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 illustrates nonlinearity and interactions well 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 outliners 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.