Bridging the Digital Divide in the US

Augusto Espin[*]

October 10, 2022

Abstract

The internet plays a vital role in everyday life across the world. The US, however, has seen a slowdown in household broadband adoption since 2010, creating a gap between connected and unconnected households usually referred to as the "digital divide." While prior studies have documented how the digital divide is related to income, demographics, and geographic location, this paper takes a different approach and focuses on the mechanisms that could help bridge this gap. To this end, we use a two-stage approach. First, we construct a comprehensive and detailed dataset on household internet usage and prices to estimate broadband demand. Second, we employ the estimated income-dependent demand elasticities to assess multiple counterfactuals aimed at evaluating a number of public policy initiatives, including those recently approved in the Biden Infrastructure Act. We contrast the effectiveness of the policies on three metrics: a) policy costs, b) reduction of the digital divide, and c) and consumer surplus increase. We find that affordability policies (i.e., subsidies) can have a larger impact on decreasing the gap and on increasing consumer surplus vis-à-vis infrastructure deployment policies (i.e., increased coverage or bandwidth).

1 Introduction

The internet has become a vital part of everyday life. Americans rely on this network for accessing a variety of products, services, and government benefits. The key societal role of high-speed (also known as "broadband") internet was most recently evidenced during the COVID-19 pandemic, when it allowed the country to keep its economy running and helped

[*]PhD candidate, University of Massachusetts Amherst, aespin@umass.edu

support many daily activities. As a result of its pivotal role in society, governments around the world are constantly pursuing and adopting policies aimed at enabling universal access to broadband internet.

Although broadband internet access via mobile technology is technically feasible, broadband access policy has focused on fixed broadband, which offers users larger capacity and more affordable access. Fixed technologies can not only accommodate faster speeds but also allow more users (e.g., all household members) to access content more economically. That is, unlimited internet access at a single flat rate is commonly offered by fixed broadband companies but not by mobile operators. Thus, as in policy discussions, the analysis in this paper focuses on broadband internet service provided via fixed technologies.

Since 2010, the rate of household broadband adoption has slowed (see Figure 1). While this slowdown is to be expected as product adoption gets closer to full coverage, the sizable fraction of unconnected households has received attention from governments and international organizations around the world. The term "digital divide" reflects the fact that the unconnected are at a disadvantage in not being able to access the ever-growing universe of information and services (and consequent opportunities).

Multiple studies report factors that may prevent households from adopting broadband internet; unsurprisingly, these factors include income, educational attainment, race, and location. In this paper, we offer a significant advance in the analysis and explore mechanisms and policies that can help bridge the digital divide. Specifically, we first estimate a model for household broadband demand and then use the estimated structural parameters to simulate the outcomes of various recently proposed policies directed at bridging the digital divide. We evaluate a) policy costs, b) reduction of the digital divide, and c) increases in consumer surplus.

To estimate demand, we use publicly available data to assemble a novel dataset that contains detailed information on coverage, prices, internet speed, and usage at a very granular level: 63,900 tracts covering almost 90% of the US population.[1] One hurdle in the construction of the dataset is that publicly available operators' prices are only available for a subset of tracts. To circumvent this issue, we rely on machine learning algorithms that assign prices to tracts based on similarities in demographics and technology (e.g., fiber) in nearby tracts where prices are available.[2] Finally, we enrich the dataset with demographic information from the US Census Bureau.

[1]Census tracts are small, relatively permanent statistical subdivisions of a county. More detail is available in `https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf`.

[2]See Section 3.6 and Appendix A for details.

Our demand model follows the discrete choice literature. Given the nature of our data (see Section 3), households face three choices: high-speed internet, low-speed internet, and no internet (outside option). A key outcome of our demand estimation is price sensitivity, which is modeled as a function of income. The overarching idea of our counterfactuals is to simulate how consumers would react if market conditions were altered by two types of policies: a) government subsidies targeted to lower income households (i.e., a targeted price drop) and b) government-incentivized network deployment (i.e., greater infrastructure coverage).

The Biden Infrastructure Act (BIA) budgeted \$14.2 billion (22% of the total \$65 billion BIA budget) for direct subsidies and \$42.25 billion (65% of \$65 billion) in infrastructure deployment. Our results show that direct subsidies could increase household connectivity by 4 percentage points and increase consumer surplus by \$260 million.[3] Conversely, policies intended to increase coverage through infrastructure deployment could result in an increase of less than 1% connected households and almost negligible impacts on connectivity and consumer surplus.

We carry out additional counterfactuals to better understand the costs and benefits of closing the digital divide. One scenario estimates the required price drop in each tract with an average income of less than \$75,000 so that all households in that tract would enroll in a high-speed broadband plan; we then compute the costs and benefits of this price drop. While we find that this strategy could boost fixed broadband connectivity by 13% and consumer surplus by \$1.3 billion, it would require a budget 2.7 times larger than that allocated for income-targeted subsidies in the BIA.

Another counterfactual is aimed at quantifying the consumer surplus that would be gained if the (minimum) speed of broadband plans increased from the 10 megabits per second (Mbps) download threshold to a more stringent 25 Mbps (the current FCC threshold for high-speed internet). While we cannot quantify the policy cost of this counterfactual, we find that consumer surplus would increase by \$201 million (or \$2.6 per household/year).

The paper is organized as follows: In Section 2, we provide important concepts and background descriptive information regarding the digital divide. Section 3 describes the datasets used and the required transformations to compute the demand estimation as well as summary statistics. Section 4 describes the demand model and identification. Section 5 presents the results of the demand estimation, including interaction with income. Section 6 presents the counterfactuals performed, which include the evaluation of the BIA policy and a more aggressive proposal to close the digital divide. Finally, Section 7 concludes.

[3] As we later explain, these estimations are done with data that precedes the BIA Act and are, hence, an approximation.

2 The Digital Divide

In this section, we first explain the notion of the digital divide as it pertains to the US and its importance for policy. We then describe the divide and some of the factors associated with it. The stylized facts that are presented provide the background and motivation for our modeling and counterfactual choices.

The term "digital divide" was introduced in the mid-1990s to name the gap separating people with and without access to information and telecommunication technologies. One of the most important indicators used to understand this divide has been internet access. Demand for telecommunication systems has been extensively studied in economics; a takeaway from this literature is the distinction between demand for access[4] and demand for use (i.e., conditional on access).[5]

As is usual in the use of new technologies, adoption follows an S-shaped curve, with an initial period of slow growth until a critical mass is reached, followed by a period of rapid growth that eventually levels off. Figure 1 shows the adoption curve of fixed broadband internet from 2005 to 2020. The deployment of fixed broadband technologies started around 2000. By 2005, 40% of US households had already adopted broadband services; 10 years later, 60% of households had subscribed. However, in the last decade we notice a slowdown in adoption, with an increase of less than 20%. While this flattening is not unexpected, it is significantly more pronounced than that observed in Europe: Although both Europe and the US exhibited similar broadband adoption rates in 2005, by 2020 Europe has pulled ahead of the US by approximately 10%.

Most countries have adopted policies aimed at expanding telecommunication access to all households. This policy can be traced back to the concept of universal service used by Bell System to transform the industry to a regulated monopoly (see Mueller, 1997). As a consequence, many governments have established some form of digital agenda for closing the digital divide. Feijóo et al. (2018) explore the high-speed broadband situation in the European Union and estimate that the Digital Agenda for Europe (DAE) would require an investment of €137.5 billion to resolve the digital divide.

[4] What the Federal Communication Commission (FCC) refers to as availability. See https://us-fcc.app.box.com/v/bdc-availability-spec.

[5] Some examples include Rohlfs (1974), who developed demand modeling that simultaneously allowed for access and usage components. Empirical studies, such as Train et al. (1987), analyze users' preferences for usage charges as well as substitutability across different types of services.

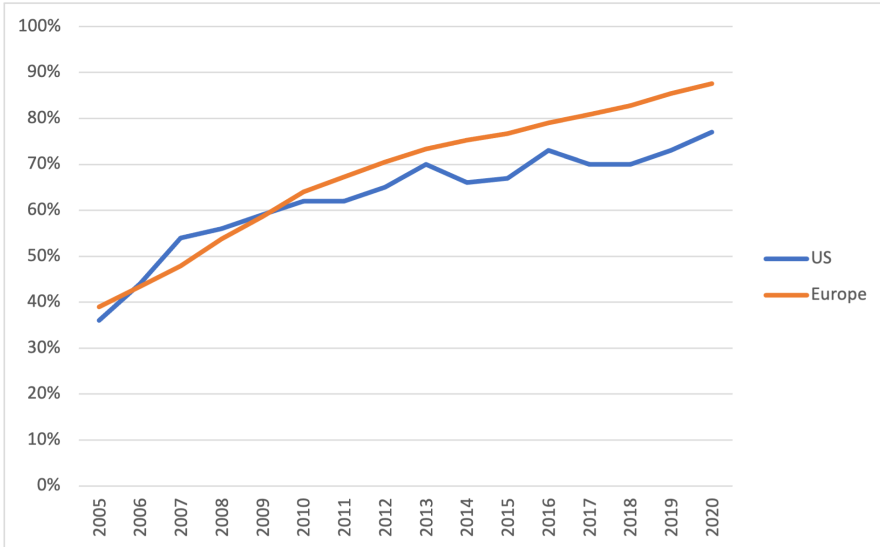

Figure 1: Household broadband connections in the US and Europe

Source: Pew Research Center and International Telecommunications Union (ITU).
(https://www.pewresearch.org/internet/fact-sheet/internet-broadband/)
(https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx)

In the US, several programs at both the federal and state level address the digital divide.[6] For example, the FCC's Connect America Fund provides funding for broadband operators to defray the cost of operating in high-cost areas across the US and supports smaller cooperatives and independent companies.

More recently, the Emergency Broadband Benefit was established during the COVID-19 pandemic to help low-income households connect to the internet. According to the FCC, 9 million people have benefited from this program.[7] Finally, in 2022, the Biden Infrastructure Act of 2021 (BIA) included an entire chapter on broadband infrastructure, which gave rise to the Digital Adoption Act.[8]

The BIA includes $65 billion to improve US broadband access. The Department of Commerce supervises the allotment of $42.25 billion for infrastructure development (ID) in unserved and underserved areas, while the FCC established the Affordable Connectivity Program (ACP) with an assigned budget of $14.2 billion. An additional $8.35 billion was

[6]More generally, the Department of Commerce is responsible for achieving "digital equity" and "digital inclusion" and building capacity for broadband adoption among US residents. Digital equity is defined as a condition in which individuals and communities have the information technology capacity that is needed for full participation in the society and economy of the United States; the term digital inclusion encompasses the activities that are necessary to ensure that all individuals in the United States have access to, and the use of, affordable information and communication technologies

[7]Source: `https://docs.fcc.gov/public/attachments/DOC-378908A1.pdf`

[8]H.R.3684 - Infrastructure Investment and Jobs Act.
`https://www.congress.gov/bill/117th-congress/house-bill/3684/text`

assigned to other programs, including digital readiness, rural deployment in tribal lands, telehealth, and distance learning. The two main objectives of the BIA are thus to a) improve network availability (through the ID initiative) and tackle affordability problems (through the ACP). As we explain in Section 6, our policy evaluation scenarios are motivated by the government interventions contemplated in the BIA.

Clearly, one important enabler in the use of fixed broadband internet is the availability of the service to possible subscribers. According to the FCC (2020), between 2016 and 2018 the number of Americans without a terrestrial broadband (defined by a 25/3 Mbps threshold)[9] service provider in their area has declined by 30%. By 2018, 97.4% of the US population could subscribe to a provider offering speeds of least 10/1 Mbps. Infrastructure deployment appears to have kept a steady pace: $80 billion were invested in network infrastructure in 2018 alone. These figures suggest that the digital divide observed in Figure 1 is not driven by lack of infrastructure (i.e., availability) but by lack of adoption.

A central aspect of an internet connection is its speed (download/upload). Since policy, as well as our work, considers "fast" (i.e., broadband) internet to be the focus of bridging the digital divide, it is important to determine an appropriate threshold at which a connection should be deemed to be broadband. The FCC uses a threshold of 25/3 Mbps to separate broadband and non-broadband links.[10] However, this definition does not appear to be universally accepted. To determine eligibility for the Connect America Fund (CAF),[11] a program designed to subsidize broadband service to high-cost and rural areas, the FCC set a threshold of 10/1 Mbps. In our analysis below, we use the 10/1 Mbps standard as the threshold for determining whether a plan offers broadband; to understand the importance of this discrepancy in definition, some of our counterfactuals study policy scenarios in which broadband definition is raised to the more stringent 25/3 Mbps standard.

Fixed (terrestrial) broadband uses several technologies; the most commonly used in the US are cable modem, fiber optic, digital subscriber lines (DSL), and fixed wireless access

[9]This notation represents a link with download speeds of 25 Mbps and upload speeds of 3 Mbps. In general, most internet connections offered to households are asymmetric due to the fact that households consume services from the internet, requiring much slower upload speeds than download speeds to keep the connections working properly, as explained in Andreica and Tapus (2010). One important aspect to consider, as discussed by Mangla et al. (2022), is that the FCC coverage datasets are constructed with data reported by providers and may be inconsistent with reality, especially in rural areas.

[10]Federal Communication Commission. Inquiry Concerning the Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion, and Possible Steps to Accelerate Such Deployment Pursuant to Section 706 of the Telecommunications Act of 1996, as Amended by the Broadband Data Improvement Act. GN Docket No. 14-126

[11]Petition of USTelecom for Forbearance Pursuant to 47 U.S.C. § 160(c) from Obsolete ILEC Regulatory Obligations that Inhibit Deployment of Next-Generation Networks. WC Docket No. 14-192

(FWA). In general, fiber optic technology provides the highest possible bandwidth in both directions. Nevertheless, cable modem with the DOCSIS 3.1 standard, although asymmetric, can provide download bandwidths similar to those available using current fiber optic technology. Cable modem is the most popular technology in the US and cable operators respond to fiber providers adjusting their investment strategy (improving their infrastructure to support fiber-like speeds to respond to competition only where they feel threatened) to successfully compete with fiber optic operators (Skiti, 2020). Many authors, such as Cardona et al. (2007) and Dutz et al. (2009), estimate fixed broadband demand at the technology level (e.g., whether consumers choose cable modem vs. DSL); however, we believe that when choosing a broadband provider, subscribers are more likely to be concerned with the speed and quality of their connections than with the underlying technology (Bauer, Steven et al., 2010). Further, there is a general perception that faster broadband speed is more beneficial for the population because it allows subscribers to access richer content on the internet. The demand modeling that we adopt in this paper is consistent with these observations: We estimate a discrete choice model in which consumers decide to subscribe to either a high-speed (i.e., broadband) or a low-speed (i.e., non-broadband) connection.

Year	**Unserved (%)**	**Underserved (%)**
2016	1.77	8.94
2017	1.54	8.58
2018	1.36	8.11

Table 1: Household broadband internet availability

To better understand the extent to which fixed broadband internet is available in the US, we use our constructed dataset (see Section 3) to compute two availability measures. First, we compute the percentage of the population that is not able to connect to *any* fixed internet provider; the FCC refers to this population as unserved. The second measure corresponds to the percentage of the population for whom the only choice is to subscribe to a low-speed (less than 10/1 Mbps) internet provider; the FCC refers to this population as underserved. Table 1 reports these figures for the three years of data available in our study; although small, the evolution of these two measures indicates an improvement in availability over time.[12]

Figure 2 breaks down the two availability measures by state. There is a wide variation between states; in the worst cases, almost 30% of households are underserved. As before, availability has increased over time. Positive and significant correlation between the two measures (0.668 for 2016, 0.629 for 2017, and 0.585 for 2018) suggests that, unsurprisingly,

[12]These figures are similar to those reported in FCC (2020).

they move in tandem.

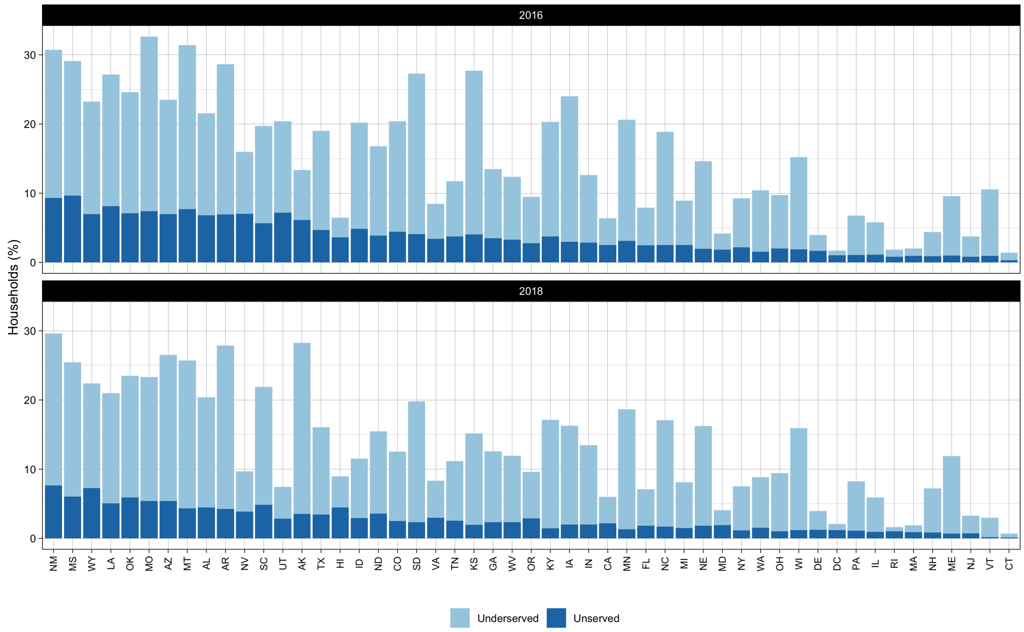

Figure 2: Availability of internet services by state

Another important aspect for understanding the availability issue is its relationship with income. Table 2 reports the percentage of unserved and underserved households, broken into four income brackets, for the years 2016 and 2018.[13] For the lowest incomes, we observe that the percentage of both unserved and underserved households is higher and has higher dispersion than for higher incomes. At the same time, we see an increase in availability from 2016 to 2018 in all cases. These patterns are consistent with Goldfarb and Prince (2008), who find a correlation of usage with income, education, and the opportunity cost of leisure time.[14]

Another dimension of the digital divide is the difference in availability between urban and rural areas.[15] Table 3 reports the percentage of unserved and underserved households separately for urban and rural locations (and separately for 2016 and 2018). Three patterns emerge. First, as before, availability is improving over time, with the urban-rural difference in the unserved measure declining from 5.1% to 3.3%. Second, the availability problem is substantially larger in rural areas than in urban areas (95% confidence intervals of t-test in

[13] As we explain in Section 3, our data is at the tract level. Thus, the figures in Table 2 report the average (and SD) of the availability measure (e.g., percentage unserved) across tracts.

[14] Prieger (2003), however, using observations at zip code level, conclude that there is little evidence of unequal availability across income levels or areas of varying ethnicity concentration.

[15] The US Census Bureau's classification of rural comprises all territory, population, and housing units located outside of urban areas and urban clusters (i.e., blocks with a population density of at least 1,000 people per square mile). See `https://www2.census.gov/geo/pdfs/reference/GARM/Ch12GARM.pdf`

	y<50[1]	**50<y<100**[1]	**100<y<150**[1]	**y>150**[1]
2016				
Unserved (%)	5.08 (11.87)	3.86 (9.25)	1.30 (4.39)	0.86 (3.12)
Underserved (%)	16.29 (24.20)	16.13 (23.21)	8.27 (17.65)	8.50 (19.83)
2018				
Unserved (%)	3.29 (8.18)	2.85 (6.85)	1.24 (3.87)	0.81 (2.49)
Underserved (%)	13.29 (20.76)	14.92 (21.63)	6.27 (14.18)	4.67 (12.97)

[1] Income (y) in thousands of dollars per year. Mean (SD)

Table 2: Household internet availability by income

means in parentheses); in particular, rural areas register 19 more percentage points in the underserved measure; more importantly, this gap has not changed over time. Third, the figures suggest that the infrastructure impediment to broadband access is not the complete lack of infrastructure (i.e., unserved percentages are relatively low) but the lack of sufficiently fast infrastructure (i.e., the "underserved" percentages are of an order of magnitude higher).[16]

The stark difference in urban and rural availability is not surprising given the substantially larger cost of network deployment in rural areas; the main reason for this difference is that the sparseness of household locations demands more infrastructure on a per-subscriber basis. To illustrate this cost differential, Vergara et al. (2010) develop a cost model for network roll-out in different settings and show that, for the same take-up rate, deploying network in rural settings can be 6 to 8 times more expensive than in urban areas. To economize on the cost of deploying in rural settings, many operators have opted to use wireless technologies, which are more cost effective but often have bandwidth limitations. In cases of very isolated populations, wireless access networks could be the only economically feasible solution. For instance, Chiha et al. (2020) propose the use of satellite technology and 4G networks to close the digital divide in Europe.

As we have seen so far, broadband availability has increased; in 2018, almost 92% of households had the option of at least one provider offering fixed broadband internet. However, only 73% of Americans subscribed to a fixed broadband provider in 2018 (FCC, 2020). The substantial gap between availability and adoption raises questions about what prevents households from subscribing to fixed broadband. Prior research provides some answers.

Unsurprisingly, affordability appears to be a consistent factor. Higher income households are, conditional on availability, more likely to adopt fixed broadband (e.g., Goldfarb and

[16] While the availability issue is, on average, less pronounced in urban areas, the issue can arise in certain urban locations. For instance, Reddick et al. (2020) study the digital divide in the San Antonio, Texas, area and found an important (un)availability issue in intra-city locations, especially in low-income areas.

	Rural[1]	Urban[1]	Difference[2]	95% CI[2,3]
2016				
Underserved (%)	23.84 (25.97)	5.34 (13.70)	19	(18.9, 19.3)
Unserved (%)	6.16 (12.10)	1.04 (3.41)	5.1	(4.9, 5.3)
2018				
Underserved (%)	22.11 (23.85)	3.56 (9.88)	19	(18.9, 19.2)
Unserved (%)	4.21 (8.61)	0.92 (2.78)	3.3	(3.2, 3.4)

[1] Mean (SD)

[2] Welch Two Sample t-test

[3] CI = Confidence Interval

Table 3: Availability of internet connections in rural and urban households

Prince, 2008; Silva et al., 2018). In addition, adoption is greater among households with higher educational attainment (e.g., Goldfarb and Prince, 2008; Silva et al., 2018). Further, there is some evidence that ethnicity may play a role, with Hispanics and Black households registering lower adoption rates even after controlling for income and education (Prieger and Hu, 2008). Competition, which induces lower prices, can also increase adoption (Prieger and Hu, 2008; Wilson, 2016).[17] It is worth noting, however, that at the rural level availability ends up being the most important factor in increasing adoption rates (Silva et al., 2018).

Year	Unconnected (%)	No high-speed (%)
2016	13.2	39.3
2017	12.5	35.6
2018	11.0	31.8

Table 4: Households without internet connection

Table 4 reports adoption rates from the data used in this paper, including (separately for each year) the percentage of households that have not adopted internet or broadband services. The number of unconnected households (i.e., households that have broadband service available in their area and choose to not subscribe) is around 4 times the number of unserved (reported in Table 1). More strikingly, the percentage of households with no high-speed connections (although the technology is available to them) is around 12 times larger than that of unserved households and more than 4 times larger than underserved households (reported in Table 1).[18] These patterns further confirm that adoption is a significantly more important issue than availability.

[17] Other literature has also explored behavioral factors, such as motivation, as drivers of adoption (e.g., Drouard, 2011).

[18] The correlation between unconnected and unserved households is still positive but much lower than that between unserved and underserved, with 0.360 for 2016, 0.336 for 2017 and 0.287 for 2018.

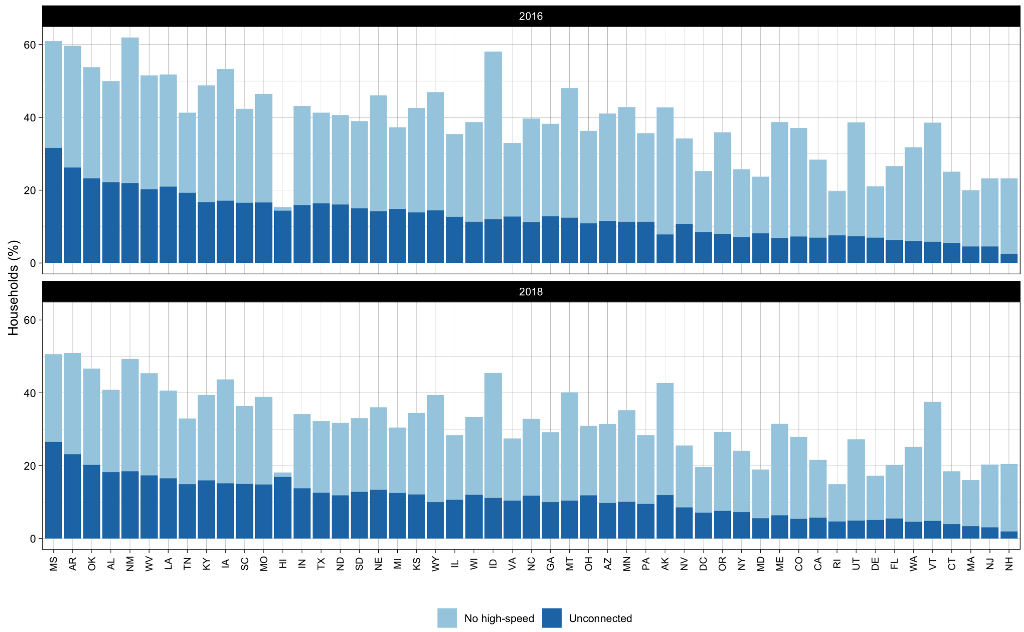

Figure 3: Households without internet connection

We explore heterogeneity in adoption across states in Figure 3 and across income in Table 5. As with availability, adoption varies widely across states.[19] As expected, the number of unconnected households is higher than that of unserved and underserved households for all states. Results in Table 5 are in line with much of the literature, which finds that adoption is strongly income-related. This can be attributed to an affordability issue, but, as is well-known, income is positively correlated with other demographics (e.g., education).[20]

	y <**50**[1]	**50**< y <**100**[1]	**100**< y <**150**[1]	y >**150**[1]
2016				
Unconnected (%)	28.26 (16.42)	11.88 (14.88)	1.53 (6.15)	0.73 (4.44)
No high-speed (%)	53.92 (19.71)	39.53 (23.86)	19.32 (15.74)	14.75 (11.28)
2018				
Unconnected (%)	25.84 (15.23)	11.51 (14.45)	1.74 (7.06)	0.75 (4.51)
No high-speed (%)	45.68 (18.75)	33.21 (21.78)	17.00 (15.17)	13.27 (10.55)

[1] Income (y) in thousands. Mean (SD)

Table 5: Internet adoption by income

[19] There is a high correlation across the two adoption variables (0.747 for 2016, 0.765 for 2017, and 0.768 for 2018).

[20] We would expect adoption in rural areas to be lower than in urban areas because availability is more restricted.

3 Data

The main purpose of this paper is to estimate a discrete choice demand model for fixed broadband. Because regulations for fixed internet provision are very flexible, there are many fixed internet providers across the US (in many cases, local governments participate in the market). As a result, there is wide variation across the country in terms of the number of providers available in any given area; further, most providers are not present in broad regional areas but provide service locally (e.g., county, state).

Further, because of the fixed nature of the service, consumers are constrained to choose from providers available in their local area. Thus, a sensible approach (for demand estimation purposes) is to use very narrow market definitions. For a given household, this approach would result in a choice set that is more realistic than if one were to use broad market definitions (i.e., the choice set would reflect those broadband providers that are truly available in their neighborhood/county). Thus, the approach in this paper is to model demand at the narrowest possible geographical level, the tract level. This approach, however, presents several hurdles as the comprehensive data necessary to estimate demand for broadband internet at this level of granularity are not readily available.

The two main components required for demand estimation are number of subscribers (or market shares) at the market (in our case, tract) level and prices. While market share data is available at the tract level, the price data is only available at the state level (via a large and representative survey of providers' internet plans). To deal with this mismatch and reliably assign prices at the tract level, we rely on a) detailed coverage data from the FCC and b) machine learning methods. We next provide a road map of the procedure, some details of each dataset, and data assembly details.

3.1 Data Road Map

The first step in the procedure is ensure that each internet provider (and data plan) in our database has a price assigned to it. To increase reliability, we carry out this matching procedure at the narrowest geographic unit possible, which in our case is a block.[21] For this, we rely on the FCC's detailed block-level coverage information. This information contains the identity of providers available to households in a given block as well as maximum advertised

[21] Census blocks are the smallest geographic area for which the Bureau of the Census collects and tabulates decennial census data. More information at https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf.

download and upload bandwidths and the technology used. This data, unfortunately, does not provide plans or pricing information.[22] We note that blocks are much narrower geographic units than tracts (our level of analysis for demand estimation); we explain in Section 3.6.2 how we deal with this mismatch.

To assign prices to each provider in a block, we use a national survey of prices at the provider-plan level carried out by the FCC. Because plans (and their corresponding pricing information) comes from a representative state-level sample, it is not always feasible to match a price from the survey to a provider (and plan) in each of the 11.16 million blocks in the FCC's coverage dataset. To circumvent this hurdle, we rely on machine learning techniques. The idea, which we describe in detail in Appendix A, is to assign prices to a provider in the coverage dataset by searching for the most similar provider and plan in the same (or its hierarchical geographic area) in the survey data.[23] To maximize reliability, we probe our matching procedure by carrying out an out-of-sample prediction exercise (see Appendix A). Further, we report sensitivity results that restrict demand (and counterfactual) estimation to the subsample of data for which price assignment is direct (i.e., the sample of providers and plans that can be directly matched to survey price data).

The assembled block-provider-plan price dataset is then matched to usage (i.e., market share) data. Usage data, also available from the FCC, is recorded at the tract level and details the percentage of households in a given tract that subscribe to each internet provider. An important limitation of the usage data is that it is not available at provider level but at internet-speed level. Specifically, the database reports the percentage of households that have a) subscribed to a high-speed internet provider, b) subscribed to a low-speed internet provider, and c) have not subscribed to either. Our demand model is thus based on these three discrete choices. Since the three mutually exclusive sets are defined as a function of internet speed, one can think of our demand model as one of vertical differentiation.

The last step involves matching the assembled price dataset with the usage data. Since the price dataset was assembled at the block level, it needs to be aggregated so that it can be matched to the usage dataset. The aggregation is done in two dimensions: a) up to the tract level and b) up to the speed (instead of provider/plan) level. Section 3.6 describes this aggregation.

Finally, we complement the dataset by matching it with tract-level demographics from

[22]Nor does it provide the number of households subscribed to each provider. We deal with assigning usage (market share) information in the last step of the data construction, explained in Section 3.6.2.

[23]Providers' characteristics (e.g., technology used and the maximum advertised download and upload bandwidths) are part of the detailed coverage dataset; see A.

the US Census Bureau (see Section 3.5). The resulting panel dataset contains approximately 64,000 tracts per year. To our knowledge, the assembled dataset provides the most comprehensive information for broadband demand. Note that we are not able to capture approximately 15% of tracts. There are several reasons for these missing geographic units, including incomplete information, unreliable/unfeasible price assignment, and issues derived from privacy concerns that limit what could be published. However, the tracts that we are able to include in our data cover 87% of US households.

3.2 Coverage Data

The FCC provides highly detailed coverage datasets on an annual basis.[24] In this paper, we limit our analysis to 2016 to 2018 because the usage dataset (explained below) is not available beyond 2018. The dataset records the characteristics of each provider's internet offer in a given block. A record includes information of a provider's technology offered,[25] as well as their maximum advertised download and upload speeds.[26] The dataset also includes the provider's name and its parent company (if any). Our period of study includes up to (approximately) 75 million records per year, 7,802 internet providers, and 2,227 parent companies.

3.3 Usage Data

Usage data are obtained from the FCC's Form 477 Census Tract Data on Internet Access Services datasets.[27] This form provides information on the number of households (out of 1,000) using a fixed internet connection. The data reports connections for two speed levels: slow (over 200 Kbps in at least one direction) and fast (at least 10/1 Mbps). At the time of analysis, data was available only up to 2018.[28] The data does not provide an exact number of connected households but instead reports the fraction of households belonging to one of six mutually exclusive (and ascending) bins, as shown in Table 6. Our discrete choice model uses the midpoint of each bin as the dependent variable. For example, an entry (tract) with a code 2 for high-speed connections would be assigned a market share of 0.3. We also carry out sensitivity analyses for alternative assignments for the dependent variable (e.g., assigning

[24]Data available at `https://broadbandmap.fcc.gov/#/data-download`. The FCC offers a tool to visualize the latest available broadband coverage at `https://broadbandmap.fcc.gov/#/`.

[25]xDSL, cable, fiber, fixed wireless access (FWA), power line, or satellite.

[26]We exclude satellite providers due to our interest in the terrestrial broadband internet market.

[27]Available at `https://www.fcc.gov/form-477-census-tract-data-internet-access-services`.

[28]The FCC offers maps with data for each year. 2018 data is available at `https://www.fcc.gov/reports-research/maps/tract-level-residential-fixed-connections-dec-2018/`.

a 0.2 or a 0.4 for code 2. See Appendix C).

Code	Connections
0	0
1	$0 < x \leq 200$
2	$200 < x \leq 400$
3	$400 < x \leq 600$
4	$600 < x \leq 800$
5	$800 < x \leq 1000$

Table 6: Codes used in the FCC internet access dataset

3.4 Price Data

The FCC publishes price data through the Urban Rate Survey Data & Resources, which is produced through a representative collection (survey) of prices offered by fixed broadband providers in urban tracts. The purpose of this survey is to produce a reasonable broadband benchmark for every service tier 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."[29] Given its intended purpose, we posit that this data can be used to generate a good proxy for rural providers, especially for lower tier connections.[30] In 2018, the survey used around 500 sampling units to produce a representative state-level sample.[31] Each record includes the name of the provider, the state where the sample was taken, the technology used, the offered download and upload speeds, the number of gigabytes allowed in the plan, whether a data cap is included, and its price.[32] Survey weights reflect how widely available each entry (plan) in the dataset is in a particular state.

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

[30]Lower tier connections are those offered at the cheapest rate in the area, usually the lowest quality link that can be purchased from a provider.

[31]Detailed information on this survey can be found at https://www.fcc.gov/file/22209/download.

[32]Most fixed internet access plans offer unlimited download data (or at least a very high limit), but many providers limit the amount of data that a user can access through her connection. These kinds of plans are used when there are technical limitations on the available bandwidth, as is the case for technologies that use a shared bandwidth such as satellite or FWA.

3.5 Other Datasets

We also use several tract-level datasets from the US Census Bureau,[33] including basic demographics, housing estimates, and ACS estimates on internet subscription and computer ownership.[34] These variables were used either directly in the model (e.g., income), to create relevant variables (e.g., population density to identify rural and urban tracts) or to check the consistency of the internet usage derived from the FCC.

3.6 Data Assembly Details

As stated previously, our datasets are not all at the same geographical level, nor can they be directly linked. Our discrete choice model requires us to construct shares for each of the three choices (high-speed, low-speed, and no internet) in each market (i.e., tract) as well as a set of product characteristics, including price. We used the following steps to assemble the dataset for the estimation:

3.6.1 Price Assignment

This step includes two procedures: direct assignment and indirect assignment. Direct assignment occurs when the pricing dataset contains plan information for a provider in the coverage dataset. Indirect assignment occurs when a provider in the coverage dataset includes no information in the pricing dataset.

Direct assignment is not always automatic as the pricing dataset often registers multiple plans (i.e., speed-price combinations) for a provider, while the coverage dataset registers the provider's available technology (e.g., fiber) and maximum advertised speeds in a block. The first step in direct assignment is to filter the plans in the pricing data so that they fall within the maximum advertised speed parameters in the coverage data. Once this set is identified for a provider in a block, price assignment is carried out by selecting the cheapest plan (and its corresponding characteristics, such as speed) in the identified set. The logic behind choosing the cheapest plan is that it provides the most affordable connectivity at any location and is usually correlated with the minimum bandwidths that a subscriber offers. Direct assignment allows us to match approximately 17% of cases in the coverage dataset.

For indirect assignment, we rely on a machine learning algorithm.[35] The logic behind the

[33]Data accessed at https://data.census.gov/cedsci/.
[34]American Community Survey (ACS), https://www.census.gov/programs-surveys/acs.
[35]See Appendix A for details on the algorithm used.

algorithm is to produce a "predicted" price (as well as plan characteristics) for a provider in the coverage dataset. The overarching idea of the algorithm is to use the pricing data to create a cluster of pricing plans (for each available technology) in each US division.[36] The features used to clear these clusters are a) survey weights (provided in the FCC's pricing sample) and b) prices.

An ensemble classifier learns parameters from the data obtained through direct assignment and predicts the most likely weight (which measures how widely is the plan available) in the survey. The weight is used to find the most likely cluster, given the technology and division, from which to choose the plan. Then, a plan is randomly sampled from such cluster and its parameters (price and speed) are assigned to the provider. The clusters are constructed with the cheapest plans available in the division. The root mean square error computed with the matched data for high-speed plans is \$8, which shows that our current assigned prices deviate on average by $\pm$\$4 from the matched prices. For low-speed plans, this measure is \$4 and the total computed error is on the order of 10% compared with matched prices.

3.6.2 Data Aggregation

We then aggregate the assembled pricing dataset to match the usage dataset. First, we group block-level data at the tract level. This is straightforward: We identify all blocks that belong to a tract and then group all observations (providers and corresponding price-speed information) registered in that tract. Note that there could be multiple entries for a given provider in a tract because a given provider in the coverage dataset could provide slightly different types of services (e.g., different technologies or speeds) across blocks in the same tract.

The assembled coverage and price data was then aggregated up to the speed level (i.e., speeds above 10/1Mbps as high-speed connections and everything else as low-speed connections). That is, we aggregate all high-speed (and low-speed) plans (across providers) within a tract to generate one weighted average price (and weighted average speed). We use the proportion of households in the tract to whom a plan was available as the weight.

[36] US Census Bureau aggregates states in divisions and then in regions. See `https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf`.

3.7 Summary Statistics

The dataset used for estimation has an entry for each option (high or low speed) in each geographic market (i.e., tract). As is common with discrete choice datasets, in some cases (1% of the markets) only one of the two options is available. There are 405,665 observations for the three years of data (2016 to 2018). The dataset comprises over 63,000 tracts, which represent more than 85% of all tracts in the US.[37] Further, these tracts encompass 90% of the US population. Table 7 summarizes the available observations and the corresponding number of tracts per year.

In addition to market shares, the dataset includes download and upload bandwidths (Mbps) and price for each of the two options as well as information on the number of internet providers in the tract, percentage of served households (i.e., households that could subscribe to at least one provider), percentage of connected households, and the take-out rate (ratio of connected households to served households).

Year	Observations	Tracts
2016	125,960	63,905
2017	124,504	63,095
2018	126,520	63,935

Table 7: Observations and tracts considered per year

Table 8 reports the summary statistics for these variables as they pertain to low-speed connections. As can be seen, download and upload speeds remain very similar over the period, while the average price per month increased from \$36.34 to \$40.64. The average number of providers remains around five, and there is a slight increase in the number of served households. However, the percentage of connected households and the take-out decrease over time.

Table 9 reports summary statistics for high-speed connections. In this case, average download speeds increase from 20 Mbps in 2016 to 24.6 Mbps in 2018, while upload speed remains fairly constant. The average price decreases from \$56.7 in 2016 to \$53.3 in 2018. It is interesting to mention that in some tracts, high-speed service is offered at relatively high prices. The average number of providers increases over time; a similar trend is observed for the percentage of households served and the percentage of connected households. The take-out is significantly higher than that registered by low-speed connections and registers

[37] The number of tracts per year in the dataset depends on whether the price assignment procedure was feasible for a tract as well as the availability of coverage or income data.

	2016 (N = 63,872)	2017 (N = 63,062)	2018 (N = 63,909)
Download (Mbps)			
Mean (SD)	2.26 (1.12)	1.84 (1.12)	2.38 (1.32)
(IQR)	(1.50, 3.00)	(1.00, 2.33)	(1.40, 3.00)
Range	0.25 - 8.00	0.38 - 7.00	0.50 - 8.00
Upload (Mbps)			
Mean (SD)	0.52 (0.25)	0.57 (0.28)	0.65 (0.23)
(IQR)	(0.28, 0.75)	(0.37, 0.77)	(0.48, 0.77)
Range	0.13 - 3.00	0.06 - 5.00	0.25 - 5.00
Price (USD)			
Mean (SD)	36.34 (10.64)	38.29 (9.84)	40.64 (9.34)
(IQR)	(31.46, 46.33)	(33.34, 46.08)	(35.42, 48.27)
Range	14.99 - 82.84	19.70 - 222.96	14.99 - 79.99
Number of providers			
Mean (SD)	5.87 (3.46)	4.85 (3.28)	5.12 (3.85)
(IQR)	(3.00, 8.00)	(3.00, 6.00)	(2.00, 7.00)
Range	1.00 - 77.00	1.00 - 42.00	1.00 - 40.00
Served households (%)			
Mean (SD)	48.16 (32.94)	49.53 (32.35)	50.14 (33.01)
(IQR)	(18.02, 78.01)	(20.38, 78.84)	(19.68, 80.95)
Range	0.00 - 100.00	0.00 - 100.00	0.00 - 100.00
Connected households (%)			
Mean (SD)	18.29 (13.02)	16.67 (11.77)	14.57 (10.20)
(IQR)	(10.00, 30.00)	(10.00, 30.00)	(10.00, 18.39)
Range	0.00 - 90.00	0.00 - 90.00	0.00 - 90.00
Take-out (%)			
Mean (SD)	55.41 (33.71)	50.55 (34.04)	47.01 (34.34)
(IQR)	(27.59, 100.00)	(19.37, 95.72)	(15.74, 84.64)
Range	10.00 - 100.00	10.00 - 100.00	10.00 - 100.00

Table 8: Low-speed connection summary statistics

a substantial increase over the period, from 68.2% to 75.7%. These patterns suggest that low-speed connections are being substituted by high-speed connections.

	2016 (N = 62,088)	**2017** (N = 61,442)	**2018** (N = 62,611)
Download (Mbps)			
Mean (SD)	20.0 (41.3)	22.6 (44.1)	24.6 (33.2)
(IQR)	(11.3, 16.0)	(11.0, 18.3)	(12.9, 24.0)
Range	10.0 - 1,000.0	10.0 - 1,000.0	10.0 - 1,000.0
Upload (Mbps)			
Mean (SD)	8.3 (41.4)	10.3 (44.2)	8.4 (28.8)
(IQR)	(1.0, 2.7)	(1.5, 4.4)	(1.7, 5.4)
Range	1.0 - 1,000.0	1.0 - 1,000.0	1.0 - 1,000.0
Price (USD)			
Mean (SD)	56.7 (15.5)	55.7 (13.6)	53.3 (14.0)
(IQR)	(49.3, 59.5)	(46.2, 62.4)	(45.0, 59.3)
Range	30.0 - 484.5	15.0 - 199.9	15.0 - 259.0
Number of providers			
Mean (SD)	6.5 (4.0)	7.0 (4.3)	7.5 (4.9)
(IQR)	(4.0, 8.0)	(4.0, 8.0)	(4.0, 9.0)
Range	1.0 - 108.0	1.0 - 77.0	1.0 - 61.0
Served households (%)			
Mean (SD)	90.1 (21.3)	90.6 (20.0)	91.2 (18.9)
(IQR)	(94.4, 100.0)	(94.3, 100.0)	(95.0, 100.0)
Range	0.0 - 100.0	0.0 - 100.0	0.0 - 100.0
Connected households (%)			
Mean (SD)	61.0 (24.4)	64.6 (23.4)	68.6 (22.0)
(IQR)	(50.0, 87.1)	(50.0, 90.0)	(50.0, 90.0)
Range	0.0 - 90.0	0.0 - 90.0	0.0 - 90.0
Take-out (%)			
Mean (SD)	68.2 (22.2)	71.6 (20.9)	75.7 (19.0)
(IQR)	(50.1, 90.0)	(55.0, 90.0)	(70.0, 90.1)
Range	10.0 - 100.0	10.0 - 100.0	10.0 - 100.0

Table 9: High-speed connection summary statistics

Table 10 reports the most relevant demographics at the tract level. Here, the average number of households per tract is around 1,700, while the mean population per tract is around 4,500. The urban/rural distribution (location) remains stable in the period, with around 37% of households in rural areas and 63 % in urban areas. The average income per tract increased from \$75,900 in 2016 to \$82,100 in 2018. As explained later, we use location (rural vs. urban, computed using population data and density) and income (to model heterogeneity of price sensitivity) in the demand model.

Table 10 also includes four variables that measure tract-level intensity of computer and internet usage. These variables, obtained from the US Census Bureau (and collected via survey instruments), are not used in the estimation but are reported here for reference purposes. We can see that the percentage of households owning a computer decreases from 75.9% in 2016 to 73.6% in 2018, while smartphone ownership increases from 68.6% to 76.5%. Broadband connections and no internet per household are relatively consistent with the statistics previously shown; the consistency in broadband connections between the US Census Bureau survey data and those generated using providers' data from the FCC (See Table 9 "connected households," compared to Table 10 "has broadband") serve as one validity check for the data we employ in our estimation purposes.[38]

Finally, we report mean prices of each of the two types of connections (low and high speed) by state. Figure 4 shows a high variation of prices across states for both types of connections; Oregon and Alaska show the highest prices, while Vermont, Connecticut and Hawaii register the lowest prices. Low-speed prices have remained stable of the period (and in some case have increased). On the other hand, the price of high-speed connections appears to have decreased in most states. As a result, the average price of high-speed connections in 2018 is much closer to that of low-speed links than what is observed in 2016.

Figure 4: Internet prices by state

[38]One drawback of survey data is that respondents may not know or understand what is defined as a broadband connection. In this sense, FCC data (which we use for our estimation) is more precise as is defined in terms of specific download/upload bandwidths, as it is reported directly by broadband providers.

	2016 (N = 63,905)	2017 (N = 63,095)	2018 (N = 63,935)
Households			
Mean (SD)	1,684.2 (754.3)	1,695.5 (754.7)	1,714.0 (773.1)
(IQR)	(1,158.0, 2,100.0)	(1,166.0, 2,115.0)	(1,175.0, 2,139.0)
Range	6.0 - 17,829.0	0.0 - 15,141.0	2.0 - 18,506.0
Location			
rural	24,012 (37.6%)	23,624 (37.4%)	23,535 (36.8%)
urban	39,893 (62.4%)	39,471 (62.6%)	40,400 (63.2%)
Population			
Mean (SD)	4,454.2 (2,156.4)	4,465.4 (2,202.1)	4,482.0 (2,268.2)
(IQR)	(2,983.0, 5,540.0)	(2,975.5, 5,555.0)	(2,963.0, 5,569.0)
Range	24.0 - 61,133.0	19.0 - 65,528.0	17.0 - 70,271.0
Mean income (1000s)			
Mean (SD)	75.9 (39.0)	78.8 (40.2)	82.1 (42.2)
(IQR)	(51.1, 89.4)	(53.3, 93.1)	(55.3, 96.9)
Range	6.6 - 506.7	6.0 - 539.7	7.4 - 589.8
Owns a computer (%)			
Mean (SD)	74.9 (15.4)	74.1 (15.9)	73.6 (16.1)
(IQR)	(65.4, 86.4)	(64.2, 85.9)	(63.6, 85.5)
Range	0.0 - 100.0	0.0 - 100.0	0.0 - 100.0
Owns a smartphone (%)			
Mean (SD)	68.6 (13.5)	72.8 (13.2)	76.5 (12.8)
(IQR)	(59.8, 77.8)	(64.3, 81.7)	(68.6, 85.0)
Range	0.0 - 100.0	0.0 - 100.0	0.0 - 100.0
Has broadband (%)			
Mean (SD)	63.6 (18.6)	64.0 (18.6)	64.9 (18.5)
(IQR)	(51.0, 77.7)	(51.5, 78.1)	(52.7, 78.6)
Range	0.0 - 100.0	0.0 - 100.0	0.0 - 100.0
No internet (%)			
Mean (SD)	22.5 (12.8)	20.1 (12.0)	18.0 (11.2)
(IQR)	(12.6, 30.2)	(11.1, 27.0)	(9.6, 24.1)
Range	0.0 - 100.0	0.0 - 100.0	0.0 - 100.0

Table 10: Demographic summary statistics

Before proceeding to our model, we summarize some patterns regarding the digital divide in the US that emerge from the data presented thus far. First, the lack of service availability does not affect a large number of households: Approximately 92% of the population can feasibly subscribe to a high-speed provider. Conversely, the main factor that appears to be driving the divide is the lack of adoption, which, in turn, is highly correlated with income. While we cannot draw conclusions about the role that prices might have played, it is interesting to note that average prices for high-speed connections have not changed significantly over the period (and they have even increased for low-speed connections). Despite the absence of substantial price decreases over time, the average percentage of households connected using high-speed links has increased by more than 7%.

4 Model and Identification

Our demand model follows the discrete choice modeling framework introduced by Berry (1994). The indirect utility is defined as

$$U_{ijt} = x'_{jt}\beta - \alpha \cdot \log(y_t - p_{jt}) + \xi_{jt} + \epsilon_{ijt} \tag{1}$$

where i denotes a household and j the available choices: high-speed internet, low-speed internet, or the no-purchase (outside) option.[39] The subscript t denotes a market, which is defined as a tract-year pair. Price is denoted p_{jt} and income income y_t. We note that, given the nature or our data, income only varies by market (i.e., we use the average household income reported by ACS in that tract). Our specification allows for household's demand (price sensitivity) to depend on income. This is not only a realistic assumption but also a key aspect of our model results and counterfactuals. The term ξ_{jt} captures the product-market unobservables that can potentially be correlated with price (we later discuss endogeneity issues), and ϵ_{ijt} is the usual idiosyncratic Type I extreme-value-distributed term.

We define δ_{jt} in equation 2 to obtain choice-market specific probabilities $s_{jt}(x, \beta, \alpha, \xi)$, as shown in equation 3 (Train, 2009).

$$\delta_{jt} = x'_{jt}\beta - \alpha \cdot \log(y_t - p_{jt}) + \xi_{jt} \tag{2}$$

$$s_{jt}(x, \beta, \alpha, \xi) = \frac{\exp(\delta_{jt})}{\sum_{j=1}^{J} \exp(\delta_{jt})} \tag{3}$$

[39]As stated before, a connection is defined as high speed if it registers a speed of at least 10/1 Mbps. All other connections are cataloged as low speed.

Following Berry (1994), we assume that, for aggregated data, $s_{jt}(x, \beta, \alpha, \xi) = S_{jt}$, where S_{jt} is the observed market share of a given type of service in each market. The outside option of any household is not connecting to the internet. Therefore, considering that this option does not provide utility, we can obtain equation 4, which can be estimated using standard linear methods.

$$\log(S_{jt}) - \log(S_{0t}) = x'_{jt}\beta - \alpha \cdot \log(y_t - p_{jt}) + \xi_{jt} \tag{4}$$

We deal with the endogeneity of prices in two ways. First, we add a rich set of fixed effects; specifically, we control for market unobservables by adding tract-specific fixed effects. Second, and more directly, we address price endogeneity by using a variety of instruments, which we explain in Subsection 4.1.

For product characteristics, x'_{jt}, we include an indicator variable for the type of connection (high or low speed) and the (weighted average of) download speed.[40] Specifically, we add an indicator variable for low-speed connections; this variable serves as a control for quality (as measured by speed) and its coefficient, which is expected to be negative, quantifies the average (dis)utility from a low-speed connection (relative to high-speed connections).

Besides price, download speed is the most important characteristic for an internet connection. To account for the fact that utility from a speedier connection may exhibit decreasing marginal utility (i.e., after a certain bandwidth, households may not perceive a meaningful difference in the quality of the service received), we include a quadratic term for download bandwidth.[41]

Finally, we add an urban location indicator, which picks up the difference in utility that urban households receive from having an internet connection compared to that received by rural households.

Own-elasticities are computed using equation 5, while cross-elasticities are computed using equation 6. While the usual limitation of the logit model is the independence from irrelevant alternatives, this is not an issue in our case given that we only model two alterna-

[40]As explained in Section 3.6, a provider offer is weighted by the number of households that have such bandwidth available.

[41]Our data include other characteristics such as upload bandwidth or usage caps (applied by some providers). However, we do not include these variables in our chosen specification because they do not produce economically meaningful results. This is not surprising given that these technical parameters are usually not usually well understood by a household (nor do they necessarily harm the quality of service).

tives:

$$\varepsilon_{ii} = -\frac{\alpha}{y_t - p_{jt}} \cdot p_{jt} \cdot (1 - s_{jt}) \tag{5}$$

$$\varepsilon_{ii} = -\frac{\alpha}{y_t - p_{jt}} \cdot p_{jt} \cdot s_{jt} \tag{6}$$

Finally, we are interested in computing the consumer surplus at an aggregated level. Following Train (2009), we can compute the expected consumer surplus tract for a typical household under the same alternatives of internet service. Therefore, multiplying this expected surplus by the number of households served in the tract, hhs_t, we can estimate the tract-level aggregated consumer surplus, CS_t, as shown in equation 7.

$$CS_t = \left[\frac{y_t - p_{jt}}{\alpha_{jt}}\right] \cdot \log\left(\sum_{j=1}^{J} e^{x'_{jt}\beta - \alpha \cdot \log(y_t - p_{jt})}\right) \cdot hhs_t \tag{7}$$

4.1 Instrumental variables

We construct three instrumental variables (IV) using principles from both BLP and Hausman instruments (see Berry and Haile, 2015, for details). First, we construct a Hausman-type instrument by computing the average price in neighboring tracts. To reduce the possibility that common demand shocks exist across the instrumented area and the areas used for instruments, we exclude immediately adjacent neighbors and use second-order neighbors for the calculation.[42]

Our other two instruments are also computed using information from tracts other than the one being instrumented. The difference with our first IV is that instead of price we use a) a product characteristic (as in BLP) and b) the number of providers (for high speed as well as for low speed).[43] The product characteristic that we use is the advertised maximum speed.[44] As with price, we compute the IVs using the average across neighboring tracts.

The logic behind all instruments is that since providers typically serve larger areas than a single tract, they will face supply conditions (e.g., infrastructure deployment, advertising costs, etc.) that are common across multiple tracts. These IVs would then be correlated with price but likely uncorrelated with demand conditions that are specific to the tract being

[42]This is a procedure similar to that used by Wilson (2016).

[43]Our instrument based on number of providers in neighboring tracts is similar to that used by Wilson (2016).

[44]As opposed to BLP instruments, however, we do not use rival characteristics but own characteristics.

instrumented. The inclusion of tract fixed effects and the exclusion of adjacent tracts in the calculation of IVs increases the validity of our instruments. Fixed effects control for time-invariant tract unobservables, whereas excluding adjacent neighbors reduces the possibility of common shocks between instrumented market and markets used to construct instruments.

5 Results

Table 11 reports the results of OLS and 2SLS estimation. All coefficients are estimated to be significant at conventional levels (all p-values are below 0.1%) and have the expected sign. Results and diagnostic tests from first-stage regression results confirm the strength and validity of the instruments (see Appendix B). Perhaps more importantly, we note the dramatic increase in the estimated price coefficient when instruments are used, a change that is theoretically predicted to occur when endogeneity bias exists and proper instruments are being employed.

Table 12 reports demand price elasticities for each year. As expected, own-price elasticities are negative and differ between low- and high-speed connections. Demand for high-speed internet becomes more price inelastic over time, while the opposite occurs for low-speed internet. In addition, demand for high-speed connections is less price elastic than that for low-speed internet.

Our results are largely consistent with those reported in earlier work (Dutz et al., 2009; Cardona et al., 2007).[45] Substitution across speeds is asymmetrical: For a given price decrease, consumers are more likely to switch away from low-speed service than from high-speed service (a result that is also consistent with previous literature reporting on internet demand). Table 12 also reports the corresponding consumer surplus for each year; in line with other research, there is an increasing value as more subscribers connect to the internet. For reference purposes, the annual consumer surplus from broadband internet in the US is of similar order of magnitude as the funds that the Biden Infrastructure Plan has set aside for internet infrastructure (see Section 2).

Our model allows price sensitivity to vary by income; further, since income varies across regions, we can compute location-specific elasticities. Figure 5 depicts a box plot of elastici-

[45] Although this earlier literature estimated elasticities for different technologies (i.e., dial-up vs. cable modem), we can make some comparisons. Slower technologies (such as dial-up) would be somewhat comparable to our low-speed category, whereas faster technologies (cable modem) would be similar to our high-speed definition. Our low-speed (high-speed) price elasticities are consistent with dial-up (cable modem) price elasticities reported in earlier work.

	Dependent variable:	
	$\log(S_{jt}/S_{0t})$	
	OLS	**2SLS**
Type:low-speed	-1.591^{***}	-1.662^{***}
	(0.006)	(0.020)
Loc:urban	-0.335^{***}	-0.935^{***}
	(0.069)	(0.214)
Download_bw	$8.645\text{e}{-4}^{***}$	$2.861\text{e}{-3}^{***}$
	(2.105e−4)	(6.526e−4)
Download_bw^2	$-1.295e-6^{***}$	$4.518\text{e}{-3}^{***}$
	(3.903e−7)	(1.204e−6)
log(income - price)	0.626^{***}	
	(0.064)	
log(income - price)		101.679^{***}
		(6.832)
Observations	376,984	376,954
R^2	0.857	−0.314
Adjusted R^2	0.825	−0.610
Residual Std. Error	1.768 (df = 307574)	5.362 (df = 307544)
Note 1:		*p<0.1; **p<0.05; ***p<0.01
Note 2:		Regressions include tract and year fixed effects

Table 11: Demand model estimation

	2016	**2017**	**2018**
Elasticities			
high-speed own	-0.372	-0.337	-0.262
low-speed own	-0.463	-0.556	-0.588
high-speed cross	0.108	0.117	0.105
low-speed cross	0.407	0.445	0.407
Consumer surplus (billion USD)			
Internet access	41.88	43.83	49.85

Table 12: Elasticities and consumer surplus by year

ties for high-speed connections (cross-price elasticities from low- to high-speed links).[46] Lower income tracts show greater price sensitivity for both own and cross-price elasticities. Cross-price elasticities imply a similar inference: As high-speed links decrease price, low-income households are more willing to switch to high-speed connections (vis-à-vis high-income households).

At the same time, we can observe the variation from 2016 to 2018: Own-elasticity for high-speed links decreases (in absolute value) over the period while cross-elasticity (from low to high speeds) increases. Figure 6 depicts a box plot of elasticities for the different divisions of the country. In 2016, East-South-Central had the highest median own-price elasticity for high-speed connections, while the Pacific division had the lowest median own-price elasticity. On the other hand, if we look at cross-price elasticities in 2016, households in East-South-Central are more willing to switch to high-speed connections if their high-speed links drop their prices. For 2018, we see generally lower own-price elasticities for all divisions in the country. New England shows the lowest value and West-South-Central the highest value. In the case of cross-price elasticities, New England again shows the lowest value; therefore, households in that area using low-speed links are less willing to switch to high-speed connections than those in any other area in the country.

6 Counterfactuals

In this section, we use the estimated demand parameters to understand the impact of a number of policies to close the digital divide in the US. The policies can be grouped in two categories: affordability and availability. Affordability scenarios evaluate the impact of price reductions (e.g., a direct subsidy), whereas availability scenarios focus on how the digital divide would decrease if broader and better (speedier) infrastructure were to be made available.

Before providing more details on the policies we consider, it is important to note some important simplifications and assumptions that we make. First, the evaluations do not consider the bureaucratic and operational costs that fielding a policy usually has nor the time required for it (i.e., we assume that take-up increases immediately after the policy). Despite this limitation, the exercise is still useful for contrasting the upper-bound gains (i.e.,

[46] The lower and upper edges of the box represent the first and third quartiles of the distribution, and the median is marked with the middle line inside the box. The end point of the horizontal lines represent the location of $Q1$ and $Q3$ multiplied by 1.5; dots are outliers. For readability, the box plot for certain income levels is cut short.

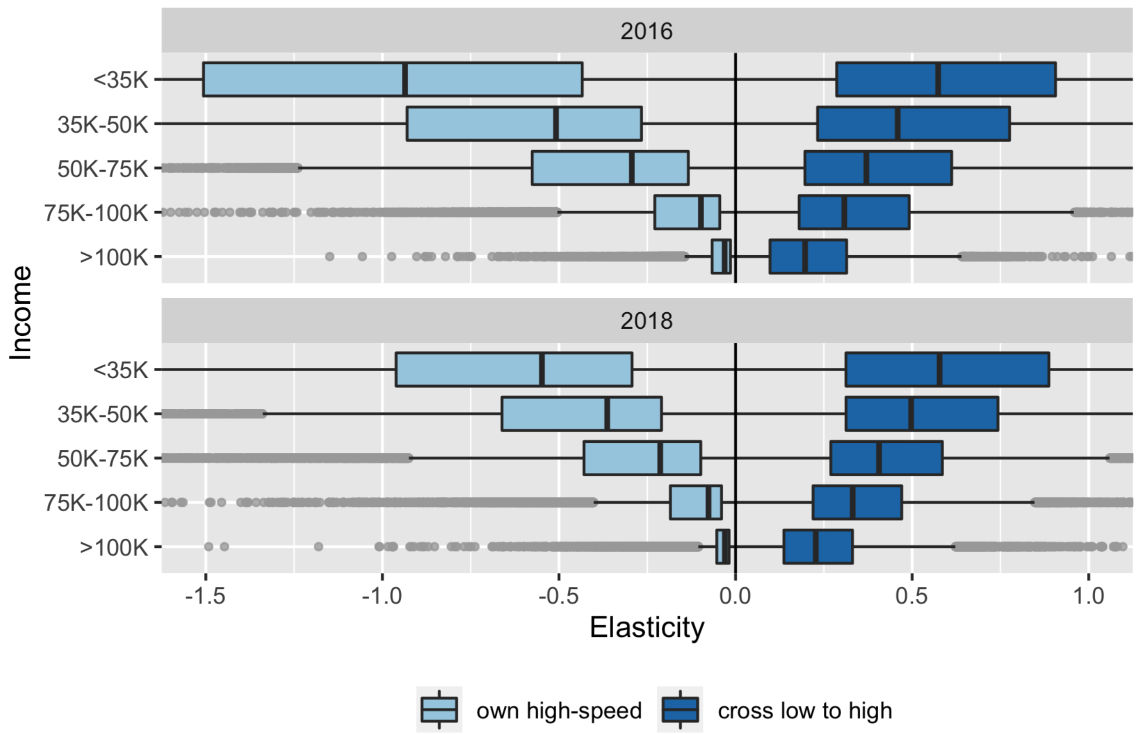

Figure 5: Elasticity variation by income

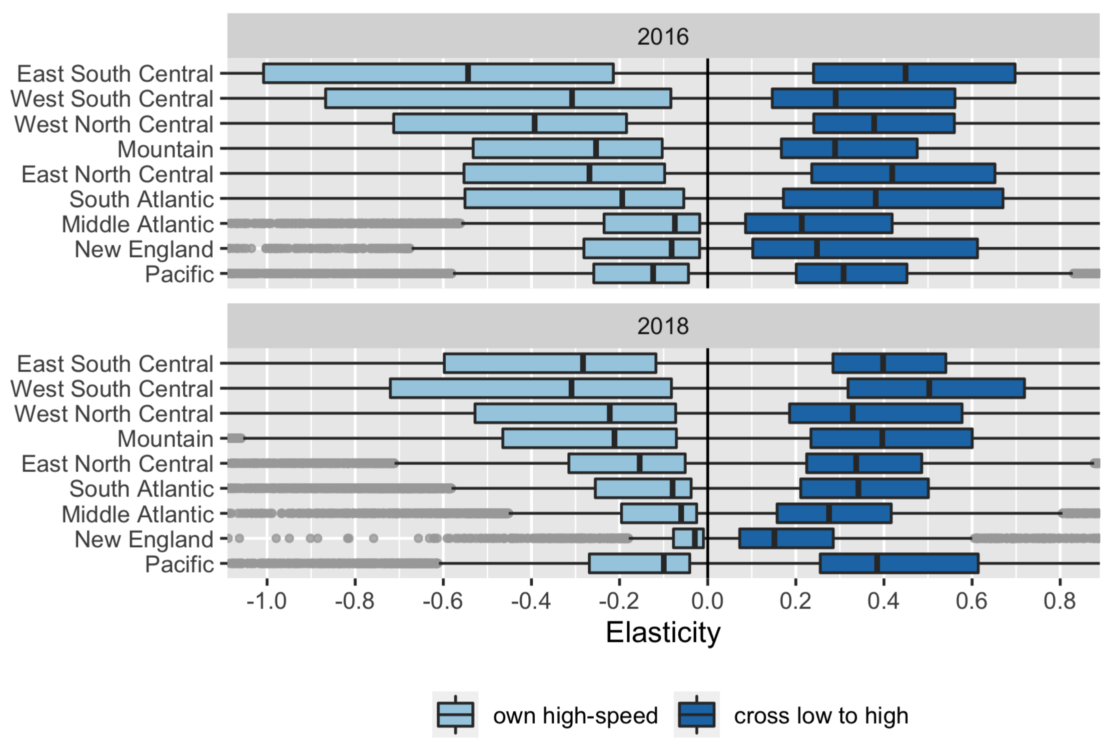

Figure 6: Elasticity variation by geography

reduction in digital divide and gains in consumer surplus) that may be feasible for each policy.

Second, we assume that competition remains unaltered after the intervention. The reason for this assumption is that the nature of our data (available only at the speed level but not at the firm level) does not allow us to model the supply side. This assumption can have important implications for our results. While we cannot predict policies' supply-side effects, a sensible prediction is that government intervention may produce supply reactions that further boost consumer well-being. For example, firms may react to the policies by modifying offering plans that are appealing to lower income population. Alternatively, consumer subsidies effectively expand the market size which, in turn, would accommodate more firms (Bresnahan and Reiss, 1991). To the extent that this conjecture is true, our results might be conservative relative to an evaluation that considered supply-side reactions.

In a first group of counterfactuals, we analyze affordability and availability scenarios that resemble those stipulated in the broadband section of the Biden Infrastructure Act (BIA) of 2021. The BIA provides two main types of support to close the digital divide: a direct subsidy to low-income families and support (e.g., grants) for infrastructure upgrade and deployment in underserved and unserved areas.

The affordability section of the BIA provides support (in the form of a subsidy provided to specific low-income households) through a bill for the Affordability Connectivity Program (ACP) run by the FCC. The counterfactual simulates the demand reaction that subsidized internet would have on tracts that meet the low-income criterion. The simulated increased demand is then used to compute the resulting decrease in the digital divide as well as gains in consumer welfare. The availability section of the BIA provides substantial resources (e.g., grants, loans) for network deployment and upgrades to underserved areas. Our counterfactual computes the additional consumers in low-income areas (as stipulated in the ACP) who would take up broadband internet if it were rolled out in areas with no current coverage. As with affordability, we then use the estimates to calculate the decrease in the digital divide and the consumer surplus gains. To include the portion of the BIA stimulus aimed at network upgrade, we consider an additional availability counterfactual that increases the internet speed.

In a second group of counterfactuals, we consider more ambitious policies that aim to close (or eliminate) the digital divide in a broader set of tracts (those with average income below $75,000, which is a less stringent threshold than that stipulated in the ACP). The availability scenario in this set of counterfactuals simulates the price drop that would be required in each tract (with average household income below $75,000) to ensure 100% take-up (of the existing infrastructure). The availability counterfactuals are similar to those in the BIA counterfactuals, with the difference that deployment is considered using the broader income criterion just described.

6.1 Counterfactual 1: The Biden Infrastructure Act

As mentioned previously, the broadband division of the BIA comprises $65 billion to improve broadband access, of which $14.2 billion are to be invested in the Affordable Connectivity Program (ACP) and $42.25 billion in broadband deployment programs. We first focus on the ACP (i.e., affordability) and, using our computed elasticities, assess the potential for that program to close the digital divide.

The second set of counterfactuals, motivated by broadband deployment programs stipulated in the BIA, quantifies the impact of deploying additional broadband infrastructure (both to provide more coverage and to increase download speeds). We explain the mechanics of these exercises next.

6.1.1 BIA: Subsidy (affordability)

The ACP considers a subsidy of $30 per month to all eligible households in the US. Eligibility is given by the formula $a+(n-1)b$, where n is the number of household members and a and b are the parameters shown in Table 13. Thus, households in the continental US have slightly different conditions than those in Hawaii and Alaska, although the eligibility structure is kept consistent for all cases. The program considers additional criteria (e.g., participating in other government assistance programs or living in tribal lands).[47] An eligible household needs to submit an application online or by mail and then contact the appropriate participating provider before the discount is applied to their bill.

Region	**Base income (b) (USD)**	**Additional member (a) (USD)**
48 states	27180	9440
Alaska	33980	11800
Hawaii	31260	10860

Table 13: Affordable Connectivity Program eligibility parameters

To implement this counterfactual, we start by finding eligible households using the pre-

[47]Detailed conditions can be found at `https://www.fcc.gov/acp`. We do not (cannot) account for these additional criteria in our counterfactuals given the lack of data.

viously explained criteria and assuming that all of those eligible simultaneously receive the allocated subsidy at no transactional cost. We assume that since most of other government assistance programs follow similar low-income eligibility process, we will capture a great majority of eligible households. We note, however, that we compute eligibility not per household but per tract; therefore all households in a tract are assumed to be eligible. Then, using the demand parameters, we compute additional households in that tract that would take up broadband internet as dictated by our demand estimates.

Table 14 reports the effects after applying the ACP policy as if it were applied entirely in the given year over its current status (i.e., the baseline for each calculation is the status quo in that year). The results show a 6.44 percentage points (ppt) increase in high-speed internet adoption for 2016, a 5.16 ppt increase for 2017, and a 3.96 ppt increase for 2018.Due to substitution effects, low-speed connection adoption exhibits a 2.71 ppt decrease in 2016, a 5.16 ppt decrease in 2017, and a 3.96 ppt decrease in 2018.

To calculate the cost of the policy, we assume that all households that are eligible—even those that are already connected—will receive the benefit. As a result, the cost of additional connections is much larger than $360/year/household: $2,256 in 2016, reaching $2,680 in 2018. The overall cost of the policy shows that because fewer households are connected to high-speed internet in earlier years, the cost is higher in 2016 ($9.2 billion) than in later years ($6.79 billion in 2018). Therefore, the BIA provides for around two years of subsidy (assuming instant take-up and no transactional costs). Finally, we compute the change in consumer surplus if this policy were applied; the maximum value is reached in 2016, with an additional $330 billion, decreasing to $260 billion in 2018. Although the additional consumer surplus resulting from greater internet adoption is well below the amount required to support the ACP policy, we note that we are not quantifying other benefits that result from greater adoption (e.g., online services, remote working, or accessing telemedicine services).

Figure 7 presents the changes in high-speed broadband adoption (measured by percentage of connected households) with respect to the baseline for each state for 2016 and 2018. At both, the start and the end of our period of study, the states that benefit most from this policy are New Mexico, Mississippi, and Arkansas. Since availability increases with time, we can see that the overall effect of the policy on adoption is smaller later in the sample. Regardless, the policy helps bring greater adoption in lower income states, thereby providing less unequal adoption across states.

	2016	2017	2018
High-speed adoption			
Baseline (%)	61.57	64.99	69.15
With ACP (%)	68.02	70.15	73.11
Low-speed adoption			
Baseline (%)	17.86	16.38	14.28
With ACP (%)	15.14	14.20	12.63
Cost of policy			
ACP subsidy (billion USD)	9.06	7.82	6.79
Per connection (1000 USD)	2.26	2.45	2.68
Consumer surplus			
Baseline (billion USD)	41.88	43.83	49.85
With ACP (billion USD)	42.21	44.13	50.11
Additional surplus (MUSD)	329.63	301.03	260.26

Table 14: Effects of the Affordable Connectivity Program

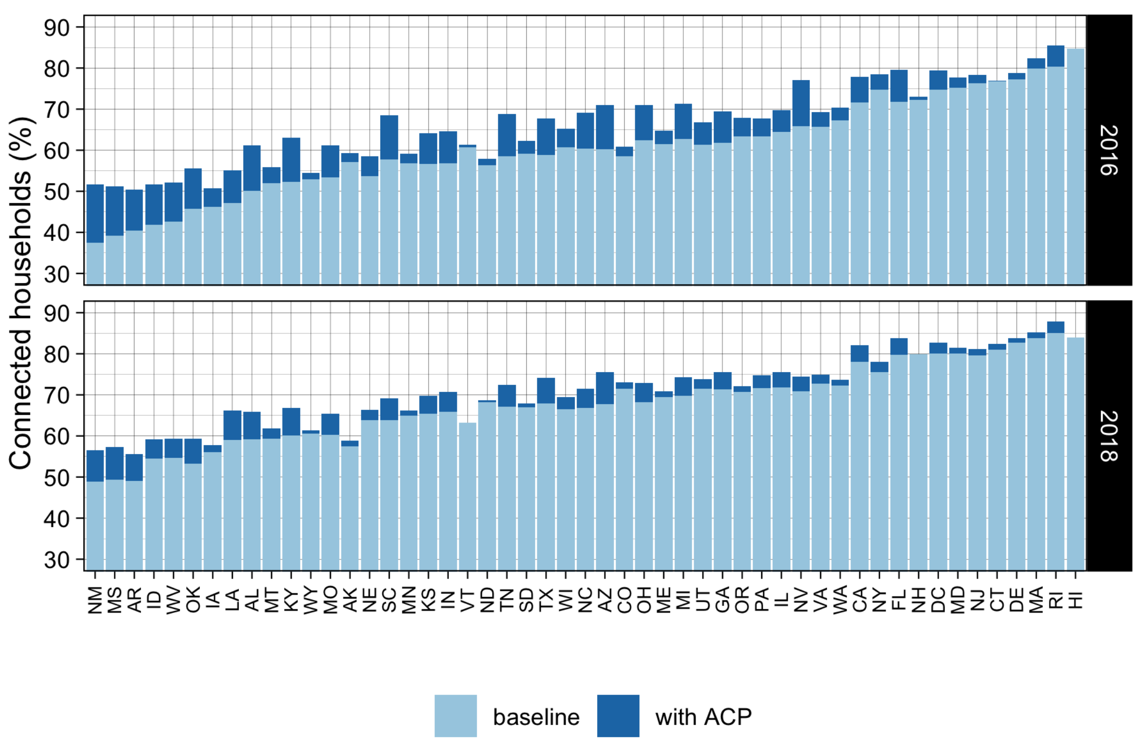

Figure 7: Adoption per state with the Affordable Connectivity Program

6.1.2 BIA: Infrastructure Deployment (availability)

We look at two possible improvements in availability. First, we improve coverage in tracts that are eligible for the ACP policy; second, we improve the minimum bandwidth available in ACP-eligible tracts. For bandwidth improvement, we increase the speed of high-speed connections that are below 10 Mbps used for demand estimation to the more stringent 25 Mbps. In both cases, we report the incremental effects of the policy that result after the affordability piece of the policy (i.e., subsidy) is implemented.

Both cases are directly related to network rollout policies, similar to the broadband deployment portion of the BIA. To estimate the effects of increasing availability in eligible sites where it is less than 100%, we simulate the demand reaction (take-up) that would result in those ACP tracts if they were instantly covered with high-speed networks.

As before, we are not considering technical issues related to network deployment nor construction time. However, we do estimate network deployment CAPEX for the first scenario (increase in coverage).[48]

Increasing availability through greater coverage decreases the digital divide by increasing the number of households that can access the choice set for that tract; at the same time, increasing availability results in an increase in consumer surplus.

Increasing bandwidth, however, only affects consumer surplus. As a consequence, for this second availability counterfactual, we only report increases in CS. For the second case, we cannot estimate additional CAPEX requirements as we do not possess information on the associated costs. However, we can evaluate the consumer surplus effect of increasing the minimum download bandwidth to the current 25 Mbps minimum considered by the FCC.

The effects of improving availability in tracts eligible for the ACP improves adoption slightly: by around 1.29% in 2016, 1.01% in 2017, and 0.76% in 2018 (see Table 15). The cost of the policy increases slightly as well by around $560 million in 2016 and $330 million in 2018; the increase in consumer surplus is also almost negligible. However, estimated CAPEX ranges from $3.49 billion in 2016 to $1.89 billion in 2018, with an average CAPEX per connection of $2,200, which is within the range of known values.

Finally, in Table 16 we show the effects in consumer surplus of increasing the minimum bandwidth available to 25 Mbps in all tracts that are eligible for the ACP. As can be seen,

[48]CAPEX means capital expenditure. It is the required investment to deploy additional network. We use an average of two methodologies proposed by Cartesian. The lower end is based on auction-based data and the higher end based on estimates to build FTTH. See https://www.cartesian.com/addressing-gaps-in-broadband-infrastructure-availability-and-service-adoption/.

	2016	2017	2018
High-speed adoption			
Baseline (%)	61.57	64.99	69.15
Additional coverage only (%)	61.91	65.29	69.41
ACP (%)	68.02	70.15	73.11
Additional coverage with ACP (%)	69.31	71.16	73.87
Cost of policy			
ACP with additional coverage (billion USD)	9.62	8.25	7.12
Per connection (1000 USD)	1.78	1.93	2.11
Consumer surplus			
Baseline (billion USD)	41.88	43.83	49.85
Additional coverage only (billion USD)	41.89	43.83	49.85
ACP (billion USD)	42.21	44.13	50.11
Additional coverage and ACP (billion USD)	42.25	44.15	50.11
Surplus from additional coverage and ACP (MUSD)	40.64	17.58	0
Estimated CAPEX for additional coverage			
Total (billion USD)	3.49	2.66	1.89
CAPEX/connection (1000 USD)	2.24	2.23	2.08

Table 15: Additional coverage effects on the BIA Affordable Connectivity Program (ACP)

	2016	2017	2018
Consumer surplus			
ACP (billion USD)	42.21	44.13	50.11
Additional bandwidth (billion USD)	42.42	44.32	50.31
Surplus from additional bandwidth (MUSD)	218.90	199.51	201.39

Table 16: Increased minimum bandwidth effects on the BIA Affordable Connectivity Program (ACP)

the effects in surplus are greater than those resulting from greater coverage and of a similar order of magnitude as those obtained from the affordability scenario (Table 14).

6.2 Counterfactual 2: Broader Policies to Close the Digital Divide

To better understand the costs and benefits of closing the digital divide, in this section we examine similar but more ambitious counterfactuals than those in the BIA-type scenarios.

6.2.1 Subsidy (affordability)

We find the price that would allow all households with an income of less than $75,000 to afford a high-speed internet subscription. Then, we compute the total cost for the industry that such price drop will imply and the number of households that will switch from low-speed links due to the substitution effect. Additionally, we compute the consumer surplus generated from the policy.

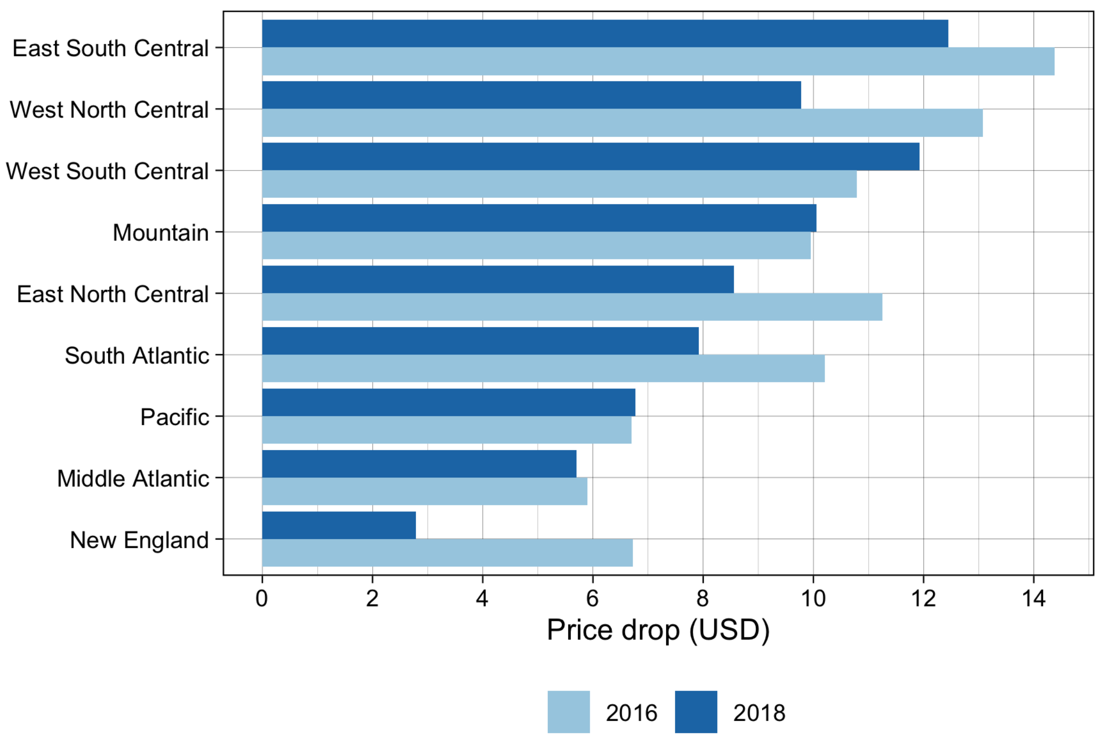

Figure 8: Price drop required to close the digital divide

In Figure 8, we show the average price drop required per division to connect households with an annual income below $75,000 for 2016 and 2018. East-South-Central states require the highest price drop, around \$12 for 2018, while New England the lowest, around \$3 for 2018. The average required price drop is \$9.69 for 2016, \$9.59 for 2017, and \$8.4 for 2018.

	2016	**2017**	**2018**
High-speed adoption			
Baseline (%)	61.57	64.99	69.15
With subsidy policy (%)	79.64	80.83	82.33
Low-speed adoption			
Baseline (%)	17.86	16.38	14.28
With subsidy policy (%)	7.74	7.41	6.91
Cost of policy			
Total (billion USD)	20.58	20.55	18.29
Per connection (1000 USD)	1.06	1.21	1.27
Consumer surplus			
With subsidy policy (billion USD)	43.40	45.36	51.23
Additional surplus (MUSD)	1519.24	1536.04	1378.65

Table 17: Effects of the policy of subsidy for closing the digital divide

Table 17 reports the results of applying this policy. The baseline is the actual situation for a given year and the parameters shown are the results of applying the policy at a given year. Thus, if the policy were applied in 2018, where we already computed that 69.15% of households are using high-speed internet connections, the proposed policy would increase the percentage of connected households to 82.33%. Since we assume that low-speed providers will not change prices, many actual low-speed subscribers will move to high-speed internet, bringing the number of low-speed subscribers from 14.28% to 6.91% in 2018. The cost of the policy for 2018 will be \$18.29 billion, where an additional subscriber connected will cost \$1,266 to the industry. This policy can increase consumer surplus by \$1.47 billion on average with respect to the baseline.

6.2.2 Infrastructure Deployment (availability)

Table 18 reports the results of a simulated increase in availability in all tracts under the \$75,000 level. We obtain an important increase of around 5.5% in adoption for all years. However, the cost of the subsidy increases by around \$3 billion on average, while the surplus generated by additional connected households increases by only \$577 million on average.

	2016	2017	2018
High-speed adoption			
Baseline (%)	61.57	64.99	69.15
Additional coverage only (%)	63.41	66.86	71.07
Subsidy policy only (%)	79.64	80.83	82.33
Additional coverage with subsidy (%)	85.74	86.36	87.22
Cost of policy			
Subsidy with additional coverage (billion USD)	23.94	23.60	20.97
Per connection (1000 USD)	0.92	1.03	1.06
Consumer surplus			
Baseline (billion USD)	41.88	43.83	49.85
Additional coverage only (%)	42.39	44.21	50.17
Subsidy policy only (billion USD)	43.40	45.36	51.23
Additional coverage with subsidy (billion USD)	44.09	45.92	51.71
Surplus from additional coverage and subsidy (MUSD)	694.58	559.49	478.30
Estimated CAPEX for additional coverage			
Total (billion USD)	19.55	17.47	15.56
CAPEX/connection (1000 USD)	2.66	2.68	2.65

Table 18: Additional coverage effects on the subsidy policy for closing the digital divide

	2016	2017	2018
Consumer surplus			
With policy (billion USD)	43.40	45.36	51.23
Additional bandwidth (billion USD)	43.62	45.56	51.43
Surplus from additional bandwidth (MUSD)	224.36	204.56	205.40

Table 19: Minimum bandwidth of 25 Mbps on the policy for closing the digital divide

Finally, in Table 19, we assume that the minimum bandwidth offered to high-speed subscribers will be increased to a minimum of 25 Mbps and compute the change in consumer surplus. The results are similar to those obtained previously, with an average increase in consumer surplus of $211.44 million.

6.3 Infrastructure Deployment: Additional Considerations

To understand the impact of developing infrastructure, we carry out two different counterfactuals. In the first exercise, network is deployed to cover all areas (not only ACP areas) where it is not yet available. The second counterfactual increases the minimum bandwidth across all households (not only ACP areas) connected to the high-speed broadband service. Figure 9 displays the effects of increasing coverage to 100% of households. The baseline is the current coverage in each year; incremental changes in coverage are plotted for each year. As we can see, results do not vary much across years. The change in consumer surplus is not linear and slightly upward growing on coverage. For instance, an additional 0.5% increase in availability would generate an increase in consumer surplus of approximately $200 million; a 1% coverage increase could boost consumer surplus by about $524 million. However, given that the current coverage is already close to 100%, there is not much room to increase surplus; the estimated CAPEX required for the additional 1% would be around $11 billion.

Figure 10 shows the effects of increasing the minimum bandwidth available across all households that currently have high-speed connections. Since our estimation requires the assumption that households are connected with the lowest tier connection available in the area,[49] the calculation provides an upper bound for possible gain in consumer surplus. After the minimum bandwidth offered exceeds 200 Mbps, the gain in consumer surplus flattens quickly. In the more linear area of the curve, below 200 Mbps, a 10 Mbps increase in the minimum offered bandwidth is associated with a $100 million increase in surplus. Intuitively, this result is consistent with the fact that a subscriber's benefit decreases over a determined download bandwidth since no real gain in usability can be perceived.

7 Conclusions

Many studies have identified factors associated with broadband adoption (e.g., income, educational attainment, age, race, geographic location). At the same time, many government

[49] The lowest priced plan offered in the area, which almost always provides the lowest download bandwidth for the offered technology (e.g., cable, DSL, fiber optic).

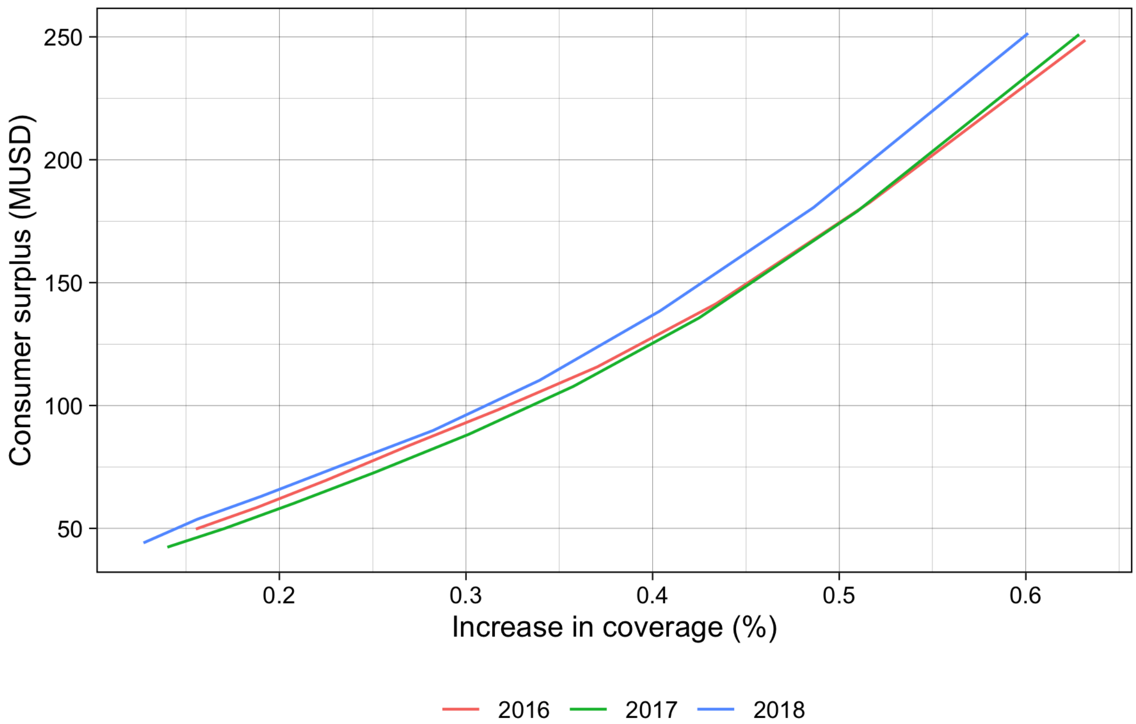

Figure 9: Change in consumer surplus due to additional availability

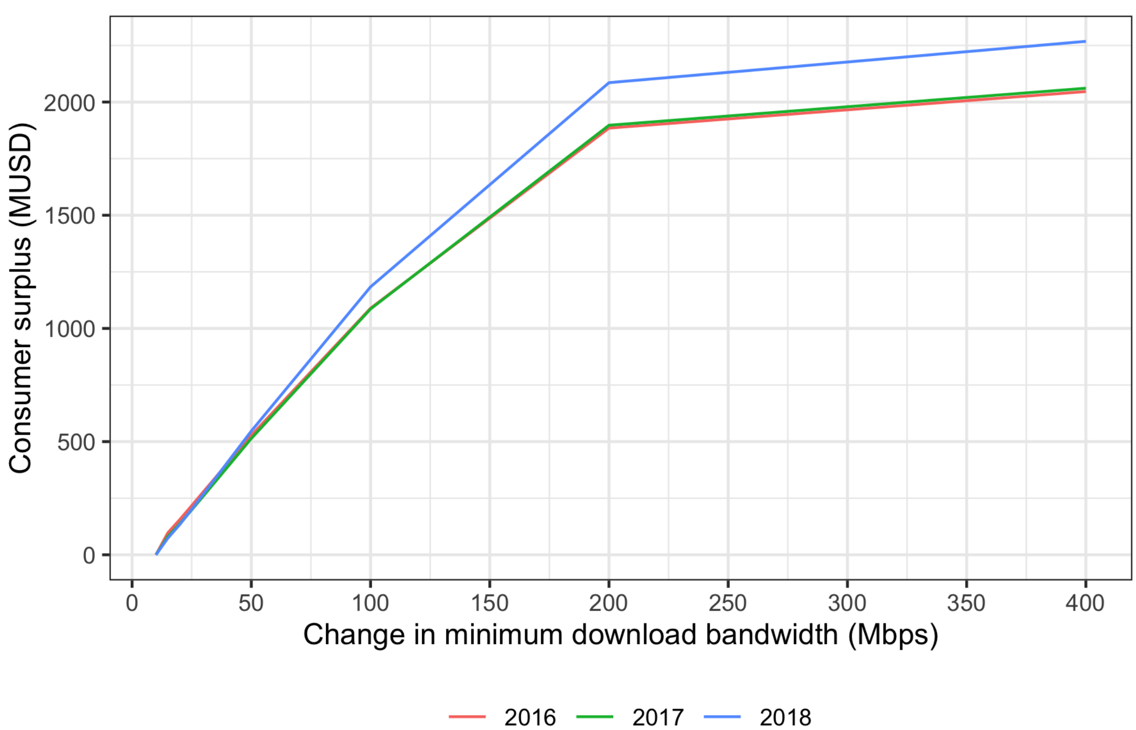

Figure 10: Change in consumer surplus due to increase in minimum bandwidth offered

efforts have focused on developing infrastructure and increasing the number of broadband providers as strategies for closing the digital divide. Recently, the BIA allocated around 65% ($42.2 billion) of its $65 billion budget to deploy broadband infrastructure; only around 22% ($14.2 billion) was assigned to overcome affordability issues. However, as we show in this study, for 2018, only 8.11% of US households were not covered by (at least) one broadband provider, while 32% of households were not connected to high-speed internet.

We use granular (tract-level) data from multiple public data sources to estimate broadband demand across most of the US. We find that price elasticity is highly correlated with income. An implication of this finding is that any policy aimed at lowering prices in lower income areas (where adoption is particularly deficient) will result in significant increases in adoption. For instance, we show that a price drop in 2018 of around $8.4 in entry level broadband plans in tracts with a mean income of less than $75,000 per year could close the digital divide by around 12 percentage points.

We use our estimated parameters to evaluate possible effects of the BIA. Our findings show that if the ACP had been applied in 2018, broadband adoption could have increased by around 4% at an estimated cost of $6.79 billion in subsidies, generating an additional $260 million in consumer surplus. We also simulate an increase in network availability in areas eligible for the ACP and find a marginal increase in adoption of 0.76% at an estimated additional cost of $7.12 billion, with negligible increase in consumer surplus. The main takeaway from this evaluation is that affordability efforts generate more impact in both adoption and consumer surplus. This contrasts with the current BIA budget allocation, which assigns almost 65% of the total to solve infrastructure rather than affordability issues.

In addition to BIA policies, we evaluate more aggressive consumer subsidy and infrastructure deployment policies. As with our BIA policy evaluation, results indicate that addressing affordability via subsidies will do more to close the digital divide and will provide higher consumer surplus than just developing infrastructure. For example, increasing coverage by 1% could potentially generate around $524 million of additional surplus but have a very limited effect on adoption. Similar results are found if the minimum available download bandwidth is increased: Each 10 Mbps increase in minimum download speed nationwide could generate around $100 million in consumer surplus, an effect that becomes negligible beyond a speed of 200 Mbps. In these more aggressive scenarios, affordability policies could reduce the digital divide by 13.1 percentage points (ppt) at a cost of $18.3 billion in annual subsidy ($1.38 billion/ppt), whereas infrastructure-only policies (which do not solve the affordability issue) would close the digital divide by only 1.92 ppt at a cost of $15.56 billion ($8.1 billion/ppt).

Although one could argue that the infrastructure cost is a multiyear investment, maintenance and operating costs are required yearly to maintain infrastructure, and the limited impact in closing the digital divide still holds. An alternative to direct subsidies to consumers could be subsidizing providers' operational costs in low-income areas, which could also reduce prices and improve affordability relative to a policy that focuses solely on infrastructure deployment.

Finally, network availability and adoption rates vary greatly by state. In our study, we quantify the importance of income as one driver in these disparities. An implication of this heterogeneity is that effective policies should consider these differences. Aside from allocating greater resources to lower income areas, infrastructure deployment policies in rural areas should consider that—in many cases—it can be too costly to find terrestrial solutions; in such cases, recent innovative satellite services could provide a solution to the availability problem while avoiding unnecessary infrastructure roll-out costs.

References

Andreica, M. I. and N. Tapus (2010, January). Efficient Upload Bandwidth Estimation and Communication Resource Allocation Techniques. Number: arXiv:1001.1451 arXiv:1001.1451 [cs].

Bauer, Steven, Clark, David, and Lehr, William (2010, August). Understanding Broadband Speed Measurements. In *TPRC 2010*.

Berry, S. and P. Haile (2015, August). Identification in Differentiated Products Markets. Technical Report w21500, National Bureau of Economic Research, Cambridge, MA.

Berry, S. T. (1994). Estimating Discrete-Choice Models of Product Differentiation. *The RAND Journal of Economics 25*(2), 242.

Bresnahan, T. F. and P. C. Reiss (1991, October). Entry and Competition in Concentrated Markets. *Journal of Political Economy 99*(5), 977–1009.

Cardona, M., A. Schwarz, B. B. Yurtoglu, and C. Zulehner (2007). Demand Estimation and Market Definition for Broadband Internet Services. *SSRN Electronic Journal*.

Chiha, A., M. Van der Wee, D. Colle, and S. Verbrugge (2020, April). Techno-economic viability of integrating satellite communication in 4G networks to bridge the broadband digital divide. *Telecommunications Policy 44*(3), 101874.

Drouard, J. (2011, March). Costs or gross benefits? – What mainly drives cross-sectional variance in Internet adoption. *Information Economics and Policy 23*(1), 127–140.

Dutz, M., J. Orszag, and W. Robert (2009, July). The Substantial Consumer Benefits of Broadband Connectivity for U.S. Households. Technical report, Internet Innovation Alliance.

FCC (2020, April). 2020 Broadband Deployment report. Technical report, Federal Communication Commision. GN Docket No. 19-285.

Feijóo, C., S. Ramos, C. Armuña, A. Arenal, and J.-L. Gómez-Barroso (2018, October). A study on the deployment of high-speed broadband networks in NUTS3 regions within the framework of digital agenda for Europe. *Telecommunications Policy 42*(9), 682–699.

Goldfarb, A. and J. Prince (2008, March). Internet adoption and usage patterns are different: Implications for the digital divide. *Information Economics and Policy 20*(1), 2–15.

Mangla, T., E. Showalter, V. Adarsh, K. Jones, M. Vigil-Hayes, E. Belding, and E. Zegura (2022, March). A tale of three datasets: characterizing mobile broadband access in the U.S. *Communications of the ACM 65*(3), 67–74.

Mueller, M. (1997). *Universal service: competition, interconnection, and monopoly in the making of the american.* Place of publication not identified: Aei Press. OCLC: 228288985.

Prieger, J. E. (2003, April). The Supply Side of the Digital Divide: Is There Equal Availability in the Broadband Internet Access Market? *Economic Inquiry 41*(2), 346–363.

Prieger, J. E. and W.-M. Hu (2008, June). The broadband digital divide and the nexus of race, competition, and quality. *Information Economics and Policy 20*(2), 150–167.

Reddick, C. G., R. Enriquez, R. J. Harris, and B. Sharma (2020, November). Determinants of broadband access and affordability: An analysis of a community survey on the digital divide. *Cities 106*, 102904.

Rohlfs, J. (1974). A Theory of Interdependent Demand for a Communications Service. *The Bell Journal of Economics and Management Science 5*(1), 16.

Silva, S., N. Badasyan, and M. Busby (2018, June). Diversity and digital divide: Using the National Broadband Map to identify the non-adopters of broadband. *Telecommunications Policy 42*(5), 361–373.

Skiti, T. (2020, June). Strategic technology adoption and entry deterrence in broadband. *Industrial and Corporate Change 29*(3), 713–729.

Train, K. (2009). *Discrete choice methods with simulation* (Second edition ed.). Cambridge New York Melbourne Madrid Cape Town Singapore São Paulo Delhi Mexico City: Cambridge University Press.

Train, K. E., D. L. McFadden, and M. Ben-Akiva (1987). The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices. *The RAND Journal of Economics 18*(1), 109.

Vergara, A., A. Moral, and J. Perez (2010, September). COSTA: A model to analyze next generation broadband access platform competition. In *2010 14th International Telecommunications Network Strategy and Planning Symposium (NETWORKS)*, Warsaw, Poland, pp. 1–6. IEEE.

Wilson, K. (2016). Does Public Competition Crowd Out Private Investment? Evidence from Municipal Provision of Internet Access. *SSRN Electronic Journal*.

A Machine learning algorithm for price assignment

As explained in Section 2, we aggregated detailed, block-level datasets to the tract level, preserving their structure; therefore, we have detailed information for each provider operating in each tract, including the number of households where service is available, the advertised download and upload speeds, and the technology used. On the other hand, the price dataset has a large sample of data plans offered by providers that can be matched with providers at the state level. Many plans could be offered by a provider under the maximum advertised download and upload speeds. Under the assumption that the lowest priced plan for each provider in each tract furnishes the maximum number of subscribers in that tract, we can reduce the number of possible plans that needs to be matched. Using this approach, it is possible to find plans by matching providers by state between the survey and detailed dataset that use the same technology and provide download and upload speeds under the provider's maximum advertised parameters. Using this procedure, we can match 17% of the providers in the detailed coverage dataset with an entry level plan.

To perform the price assignment for the remaining providers in the detailed coverage dataset, we develop a machine learning algorithm that creates clusters from the survey data using two features: price and weight (this parameter in the survey dataset quantifies the number of subscribers to whom a plan is being offered while considering the size of the sample and other technical factors).[50] We computed clusters for each division in the US and each access technology available (e.g. DSL, cable, fiber optic).

We use a higher hierarchical geographical level because most national providers operate subsidiaries at that level. Therefore, similar plans will be deployed in all states and regions within the subsidiary because a consistent offer of plans in a similar market is a common industry practice.[51] For instance, if we look to a national provider (e.g., Verizon), a subsidiary company (Verizon New England) operates in most regions of New England (some areas may be excluded due to the lack of availability) with a consistent market offer. It is reasonable to assume that smaller local providers need to offer competitive plans in the same areas. Therefore, by building clusters for each geographic division and available technology, we have a larger basket of plans likely to be offered in the area.

On the other hand, we use the technical parameters (e.g., advertised download and upload speed and access technology) and demographic parameters (e.g., households served, total

[50] https://www.fcc.gov/file/22209/download.

[51] There is a trend in the industry to offer consistent plans over large geographical areas. See https://www.cnet.com/home/internet/best-no-contract-internet-plans/.

number of households, plan weight) from the portion of successfully matched providers with entry level plans to determine the most likely weight (we know the weight of matched plans) for a given cluster (which is already segregated by geography and technology). Then, once the algorithm is trained, we can use that predictor to find, within the clusters segregated by geography and technology, the most likely weight. With that parameter, we can find the closest cluster and sample one of the available plans to assign it to that provider. The algorithm used to predict the weight from matched data is the histogram gradient boosting classifier,[52] an ensemble of decision trees that are added sequentially to correct prediction errors.

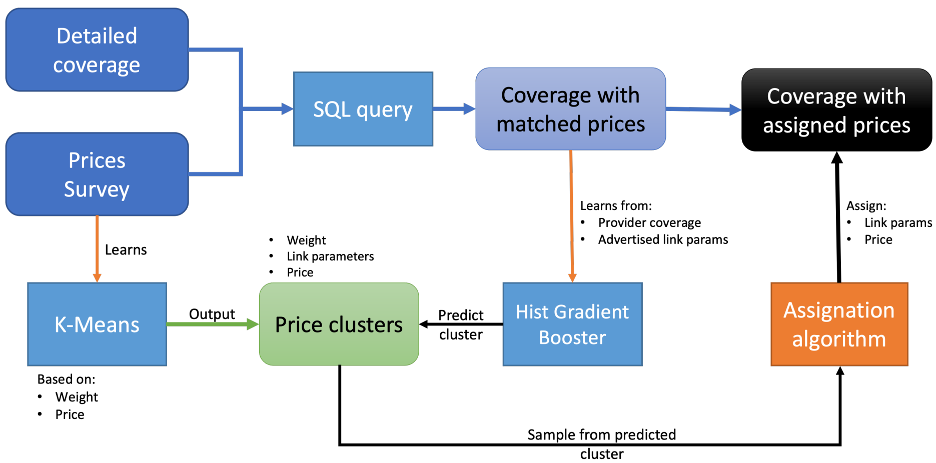

Figure A1: Machine learning algorithm for price assignment

Figure A1 depicts the whole process. Initially, plans are matched to coverage using an SQL query, while the K-means algorithm is used to produce clusters for each division and technology.[53] The histogram gradient booster classifier (from the sklearn implementation) is trained with the matched plan data. There may be cases in which there are no plans in the division for a given technology or in which the plan download and upload bandwidths lie outside the advertised limits, in which case no match is found. Finally, the algorithm produces a dataset with matched and predicted prices and other technical parameters at the tract level but with block-level detail. This data is later aggregated at the tract level to create the estimation dataset.

Table A1 evaluates the performance of the algorithm by comparing the matched price with the predicted price for the available subset of data. We use standard machine learning

[52]For an explanation of the algorithm, see `https://www.analyticsvidhya.com/blog/2022/01/histogram-boosting-gradient-classifier/`.

[53]See https://www.analyticsvidhya.com/blog/2021/11/understanding-k-means-clustering-in-machine-learningwith-examples/ for an explanation of this algorithm.

	Predicted price[1] (USD)	Matched price[1] (USD)	Overall Error (%)
2016			
High-speed	60.18	60.60	13.65
Low-speed	40.71	43.63	11.03
2017			
High-speed	60.85	61.95	11.55
Low-speed	40.22	40.25	6.37
2018			
High-speed	52.33	53.54	15.17
Low-speed	43.34	45.83	11.80

[1] Mean value

Table A1: Overall performance

procedures, leaving 20% of the matched data for testing and the remaining 80% for training in order to search for the hyper-parameters of the classifier. We evaluated the parameters using the testing portion of the data. The evaluation shown in Table A1 reflects an overall evaluation of predictions over matched and aggregated data at the tract level, similar to the one used in the demand estimation. The prices shown in Table A1 are mean values for predicted and matched prices; both are quite similar. As we can see, overall errors, computed using all the matched data, vary by year and are in the 6% to 15% range. The consistency in errors give us confidence in the inference estimation that we later perform using this data. For this estimation, we used matched data, where available, and predicted otherwise.

B First-stage regression

As can be seen from Table A2, most regressors in the first stage have p-values well below 5% level. (The only exception is Download_bw^2, which has a p-value close to 5%). Importantly, the F-statistic for excluded instruments is highly significant. The instruments used are as follows: $instr1$ is computed using the advertised download bandwidth, $instr4$ is computed using the number of providers, and $instr5$ is the price. For a detailed discussion on how these instruments are calculated, see Section 4.1. Other instruments that were computed but did not provide significant results were advertised upload bandwidth and usage allowance (cap defined in plans).

	Price			
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
Type: low-speed	0.002562	0.0003374	7.6	3.162e-14
Loc: urban	0.005899	0.001966	3	0.002702
Download_bw	-1.362e-05	6.007e-06	-2.3	0.02342
Download_bw^2	2.132e-08	1.11e-08	1.9	0.05479
instr1	6.872e-06	7.774e-07	8.8	9.588e-19
instr4	-0.0001184	4.425e-05	-2.7	0.00745
instr5	-6.766e-05	7.884e-06	-8.6	9.385e-18

Multiple R^2(full model): 0.9897. Adjusted R-squared: 0.9873
Multiple R^2(proj model): 0.0009709. Adjusted R-squared: -0.2245
F-statistic(full model): 424.5 on 69411 and 307542 DF, p-value: $< 2.2e-16$
F-statistic(proj model): 42.7 on 7 and 307542 DF, p-value: $< 2.2e-16$
F-statistic(excl instr.): 81.89 on 3 and 307542 DF, p-value: $< 2.2e-16$

Table A2: First-stage regression

Table A3 reports weak instrument diagnostic tests and confirms that the chosen instruments are reliable. The first two tests confirm that the instruments are correlated and exogenous, while the last test shows that we do not have an over-identification problem.

Test	df1	df2	statistic	p-value
Weak instruments	3	376944	2537.28	<2e-16 ***
Wu-Hausman	1	376945	2893.64	<2e-16 ***
Sargan	2	NA	24.72	4.28e-06 ***
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table A3: Weak instruments tests

C Robustness check

	Dependent variable:		
	$\log(S_{jt}/S_{0t})$		
	Matched	**Lower limit**	**Upper limit**
	(1)	(2)	(3)
Type:low-speed	-1.861^{***}	-1.072^{***}	-2.440^{***}
	(0.023)	(0.006)	(0.007)
Loc:urban	-0.325	-0.343^{***}	-0.309^{***}
	(0.296)	(0.070)	(0.076)
Download_bw	$1.872\text{e}{-3}^{**}$	$7.890\text{e}{-4}^{***}$	$1.301\text{e}{-3}^{***}$
	($8.760\text{e}{-4}$)	($2.119\text{e}{-4}$)	($2.314\text{e}{-4}$)
Download_bw^2	$-5.539\text{e}{-6}^{***}$	$-1.141\text{e}{-6}^{***}$	$-2.097\text{e}{-6}^{***}$
	($1.232\text{e}{-6}$)	($3.929\text{e}{-7}$)	($4.292\text{e}{-7}$)
log(income - price)	0.739^{***}	0.617^{***}	0.705^{***}
	(0.254)	(0.064)	(0.070)
Observations	81,534	376,982	376,984
R^2	0.890	0.858	0.831
Adjusted R^2	0.809	0.826	0.793
Residual Std. Error	1.990 (df = 46958)	1.741 (df = 307572)	1.926 (df = 307574)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A4: Robustness check using OLS

Here, we provide estimations in 3 different settings. First we only use the data that was possible to match directly from price surveys with coverage data, which accounts for 17% of the providers in the dataset. We apply the same estimation method to such data and the result is shown in the column "Matched". The other 2 settings are using the borders of the bins defined in the usage datasets, which could be seen as lower and upper bounds for the estimated parameters. Two tables are produced, Table A4 that shows results using OLS, and Table A5, which shows the results using instruments and 2SLS. In general, results have the expected signs and their magnitudes are in-line with the origin of the underlying data.

	Dependent variable:		
		$\log(S_{jt}/S_{0t})$	
	Matched	**Lower limit**	**Upper limit**
	(1)	(2)	(3)
Type:low-speed	-2.038^{***}	-1.142^{***}	-2.739^{***}
	(0.066)	(0.019)	(0.032)
Loc:urban	-2.909^{***}	-0.929^{***}	-1.488^{***}
	(0.860)	(0.210)	(0.406)
Download_bw	$7.216\text{e}{-3}^{**}$	$2.743\text{e}{-3}^{***}$	$5.231\text{e}{-3}^{***}$
	(2.433e−3)	(6.406e−4)	(1.235e−3)
Download_bw^2	$-1.015\text{e}{-5}^{**}$	$-4.295\text{e}{-6}^{***}$	$-8.440\text{e}{-6}^{***}$
	(3.325e−6)	(1.182e−6)	(2.279e−6)
log(income - price)	135.621^{***}	99.498^{***}	199.556^{***}
	(18.369)	(6.706)	(12.931)
Observations	81,491	376,952	376,954
R^2	0.226	−0.245	−3.595
Adjusted R^2	-0.344	−0.526	−4.632
Residual Std. Error	5.273 (df = 46917)	5.263 (df = 307542)	10.149 (df = 307544)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table A5: Robustness check using 2SLS (instruments)