BEAD/CPF
Perspective
Ookla: using Ookla crowd-sourced data to visualize fixed broadband performance
Introduction

As individual states and territories prepare to allocate federal funding for broadband infrastructure projects a subtle problem remains: how does one know with a high degree of certainty and objectivity which areas have the greatest need?

The most common approach is to look to federal data sources for insights. The FCC’s Form 477 data reports the offered speeds of internet service providers on a block-by-block basis. It has been criticized as being insufficiently granular and in many cases overstating real-world performance. A new source of optimism is the Broadband DATA Act. It replaces the census block paradigm with a “fabric” of locations. The Act also proscribes a challenge process. In theory it should produce a more granular and representative view of broadband availability. The Broadband Equity, Access, and Deployment Program (BEAD) Notice of Funding Opportunity (NOFO) indicates that the coverage calculated using the fabric will form the basis for identifying unserved or underserved areas for the purpose of allocating BEAD funds.

There are a wide variety of other metrics used in describing the lack of broadband access. These include measures of social welfare and survey-based (check the box on a survey) indications of broadband access and affordability. These metrics, while relevant to a broad discussion of the digital divide, reflect much more inclusive and abstract constructs than the simple question of last mile access.

What if the Broadband DATA Act falls short? What if communities are still highly skeptical (as any critical thinker might be) and say “Show me”. Decision-makers are then faced with the challenge of measuring and reporting actual results.

A number of states and not-for-profit organizations and for-profit organizations collect broadband performance data. In some cases these include geolocated results. In other cases they do not. Collecting data to assess broadband availability in small geographic areas is particularly challenging. One must collect a large number of samples to produce a high resolution “image”. One must also locate – accurately – each result. A number of US states have attempted to do this and many have fallen short because of the difficulty of collecting a sufficient number of responses.

One of the success stories is the State of North Carolina. They collected 111,000 samples from locations across the state. This represents a sample for every 36 households. It required a grass roots effort by local political leaders encouraging their constituents to take the test. The speed test results were coupled with survey data and precise street addresses. The effort is described in a recent whitepaper, State of the States: broadband mapping initiatives at the state level.

What if the leader of a state broadband office wakes up one morning and realizes, “We really need measured data and we need it now”? The only choice is to find someone who has already collected the data. While open source software that enables data collection is available the process of collecting a sufficient number of results may take months or years.

When one types “speed test” into Google one comes up with a site operated by Ookla. It turns out that Ookla has collected 45 billion speed test results from users around the world, over 5 tests for every man, woman, and child on the planet (including in the denominator those who have no internet access and therefore couldn’t run a speed test).
Figure 1: Pueblo County, CO with Block Group boundaries, 38,360 Ookla results (red points), and FCC 447 blocks with qualifying coverage (green).
Figure 2: Downtown Pueblo with FCC 447 blocks with FCC-reported qualifying coverage (>=100 Mbps, >=20 Mbps) in green.
Brand Names, Trademarks, and Disclaimer

Anyone using Ookla data will encounter a number of trademarks. Ookla has six registered trademarks and twelve product trademarks described in its press kit. This paper will discuss the Ookla Broadband Performance Dataset (a specific data set) which is part of a broad set of fixed and mobile data product offerings.

Ookla has a branded name for its speed test. It called it a “Speedtest”, followed by a registered trademark symbol (®). Ookla similarly refers to a collection of tests as “Speedtest® results”.

We use industry-standard language (e.g. “speed test”) in this whitepaper, since much of the discussion applies to data sources other than Ookla and because this whitepaper is the second in a series on crowd-sourced data. We recently published a whitepaper entitled State of the States: broadband mapping initiatives at the state level, which shows some similar visualizations.

The required data set can be downloaded in CSV form from the Ookla’s proprietary portal, with an appropriate subscription. Ookla’s portal also visualizes some data. We have not included any data visualizations from the portal in this whitepaper.

When we present a figure in this whitepaper that uses Ookla data we include an overlay identifying it as Ookla data. In most of our other whitepapers detailed contextual information is included, but in the text of the whitepaper, not as a label overlaid on the graphic.

The vast majority of images include data derived from other high quality sources, not just Ookla. These other sources are described in detail in the Introduction (a 42-page presentation) to the Infrastructure Essentials BEAD Toolkit, the tool used to generate each image. Every visualization includes a selection of data inputs (which layers to display) and, often, analyses of the source data that result in the specific output. In the results presented in this whitepaper Ookla provided one of several data sources but did not participate in the analysis associated with any specific visualization, apart from providing guidance on data definitions.

BroadbandToolkit.com has not independently assessed the quality / accuracy of Ookla speed test results or conducted a head-to-head comparison with any other speed test technology or data source. We are writing about Ookla primarily because it is a widely available data source for an entity (a state, a county, an ISP) to accurately assess broadband offerings in a specific geographic area.

The whitepaper describes a series of methodologies for filtering the source data set then visualizing the result. The end visualizations are similar to those one might achieve using precisely geo-coded data from other data sources. The process of cleaning the data is specific to the Ookla Broadband Performance Dataset - hence this whitepaper.

Local Perspective

In 1982 Tip O’Neill famously said “All politics is local”. One can say with similar confidence “Broadband coverage is local”. A typical household doesn’t care whether the national average broadband speed, as reported by the FCC, is going up or down. They do care about the connectivity available in their neighborhood.

This section provides a quick overview of Pueblo County, Colorado, a randomly selected geographic area that we are using to show how Ookla speed test data can be interpreted and visualized with other
Figure 3: Downtown Pueblo w. Ookla speed test results colored to reflect qualifying (green), underserved (salmon), and unserved (mustard) areas.

Qualified (>=100/20 Mbps)
Underserved (>=25/3 Mbps)
Unserved (neither of the above)

Based on BroadbandToolkit.com analysis of Ookla speed test data for 8/1/2021 to 7/31/2022.
Figure 4: Downtown Pueblo with Census Blocks categorized as Qualifying (green), Underserved (salmon) or Unserved (mustard) based on the best result. Based on BroadbandToolkit.com analysis of Ookla speed test data for 8/1/2021 to 7/31/2022.
Figure 5: Downtown Pueblo, Percentage of Test Results Categorized as Qualifying (>=100 Mbps down, >=20 Mbps up, <100 msec) by Block Group

Based on BroadbandToolkit.com analysis of Ookla speed test data for 8/1/2021 to 7/31/2022.
Figure 6: Downtown Pueblo, Block Groups Showing Median Download and Upload Speeds in Mbps, and Color-Coded to Reflect % Qualifying.

Based on BroadbandToolkit.com analysis of Ookla speed test data for 8/1/2021 to 7/31/2022.
Figure 7: Pueblo County, Tracts, Showing Median Download and Upload Speeds in Mbps, and Color-Coded to Reflect % Qualifying.
Figure 1 shows Pueblo County. Block group boundaries appear in magenta. Elevation data is enabled in the form of hillshade shadowing to show steep changes in elevation. The county of 167,412 (2020 population) has robust town centers (e.g. Pueblo and Blende), several clusters of population outside of the town centers, and significant expanses of rugged and largely unpopulated terrain. FCC Form 477 data suggest that a large portion of the population enjoys qualified broadband (>=100 Mbps down, >=20 Mbps up, and <100 msec latency). These areas, represented at a block level, are shown in green. One will discover shortly that measured data provides a much more nuanced perspective. Finally, the image shows 38,360 red dots, each representing a speed test result collected by Ookla from August 1st, 2021 to July 31st, 2022. The Ookla Broadband Performance Dataset for the specified period started with 91,616 results. After a robust filtering process (discussed in detail in a few pages) 38,360 remained.

Figure 2 shows the same information as Figure 1, except that it zooms in to the city center of Pueblo.

Figure 3 characterizes each test result as either qualifying (>=100 Mbps down, >=20 Mbps up, and <100 msec latency), underserved (not qualifying but >=25 Mbps down, >=3 Mbps up, and <100 msec latency), or unserved (not qualifying or underserved). These thresholds reflect the current BEAD definitions. In Figure 3 the layer reflecting FCC Form 477 data has been removed.

Figure 4 takes the individual test results and uses them to attribute a quality of service to the census block. Every block with speed test data is tagged as either qualifying (green), underserved (salmon), or unserved (mustard). Blocks without data (where no tests were run) are not colored. This color scheme approximates the methodology historically used by regulators. Each census block is characterized by its most favorable data point. In this case those data points are actual measurements, as opposed to advertised speeds. Measured data, as seen in Figure 4, presents a much more nuanced picture than the nearly ubiquitous green of Figure 2, based on 477 data. In blocks with multiple speed test results the most favorable result is used.

Figure 5 shows the results within each block group on an aggregate basis. The coloring of the block (red to green) reflects the percentage of test results that are qualifying.

Figure 6 is similar to Figure 5 in that its coloring reflects the percentage of tests representing qualifying coverage. Its labels show the median download and upload speeds in Mbps by block group.

Figure 7 aggregates data to the tract level (appropriate for a more zoomed out view) and displays the same data as Figure 6 in a county-wide view. Figure 7 shows a much more nuanced view using speed test data than Figure 1.

Ookla Data Sources

Ookla makes data available in many different forms. They publish two data sets for not-for-profits, one showing fixed broadband results, and the other showing mobile broadband results. The data is visualized for the entire world via scalable square pixels that become as small as 610.8 x 610.8 meters (data, description) at the maximum zoom point.

The FCC also cites Ookla data in discussions of broadband performance, as in the Broadband Performance Report.
In 2021 the NTIA released a data set called “Indicators of Need”, with tract and county level results. It includes download and upload speeds from Ookla and M-Lab and more abstract metrics from Microsoft. These data sets and others are visualized in the Infrastructure Essentials BEAD Toolkit.

The best data source for assessing fixed broadband networks is a commercial product: the Ookla Broadband Performance Dataset. It includes over 40 fields and is laser-focused on fixed broadband networks.

The Ookla Broadband Performance Dataset has several important attributes:

- **Density.** A one-year sample in the US, in the data sets we have reviewed, approaches 1 test per household. A US city with 1 million households approaches 1,000,000 crowd-sourced results over a 1-year period.
- **Precise Locations.** A large portion of the tests include precise locations. We’ll discuss in a moment how such data is collected.
- **Download speed, Upload speed, Latency, and Jitter** are among the standard metrics.
- **Broadband Providers.** A user can filter for specific broadband providers.
- **A Rich Set of Parameters.** The data set provides enough specialized fields to enable the user to remove test results that might not be representative.

The following pages describe a process of preparing the data, ensuring that each sample has a precise location stamp, then filtering it to extract the most definitive possible results for assessing the quality of the fixed broadband access, then visualizing it within the context of other geospatial data.

### Preparing the Data

The Ookla portal has a “Data Extracts” section that allows the user to download monthly files of data. To visualize the data, as described in this whitepaper, the user needs to do several things:

1. Download the data as a series of monthly CSVs.
2. Concatenate the data – preferably over 12 consecutive months - into a single data file.
3. Extract specific fields.
4. Apply a series of filters that remove unusable data points (discussed below).
5. Do a set of statistical calculations, mapping point results to geospatial polygons (block, block group, and tract).

Version 1.2 of the Infrastructure Essentials BEAD Toolkit will include a set of visualizations – similar to those shown in this whitepaper. BroadbandToolkit.com will also offer data conversion as a service, for those wishing to simply purchase CSV data from Ookla then see it integrated into the Toolkit (a simple turnkey approach).

Every state broadband office and every serious investor or broadband provider should consider using speed test data to compliment the other layers of information they must consider in choosing where to invest.

Here is quick overview of the process for Ookla data:

1. The downloaded data must be integrated into a single database. For many locations (a typical city or county) the
data set is small enough to include in a spreadsheet. For larger areas where the number of data points exceeds 1 million (e.g. a large county or a state) a more sophisticated database may be required.

2. Apply a series of quality control filters (extremely important, described below).

3. Discard those data points that fail the quality control process.

The "discarded" data points may still be very useful for other applications (calculating larger area averages, observing changes in performance over time, etc.). They are simply removed from the data set used for the precise block-level calculations required by the visualizations of this whitepaper.

Extracting Location Information

In the downloaded data each row represents a speed test. Tests are taken on laptops, desktops, and mobile devices. Desktops and laptops may be connected via an Ethernet cable or via Wi-Fi. Mobile devices are almost always connected via Wi-Fi. If a mobile device is connected directly to a mobile network using LTE or another macrocellular radio access network technology it will not be included in Ookla's Broadband Performance Dataset.

When someone runs a speed test Ookla does not ask them for their address. Everything happens automatically in the background. The location will be determined by a GPS chip in the computer or device (if there is one and if it is enabled). Alternatively, if GPS location data is not available, Ookla will calculate an approximate location based on the IP address. This concept of “location based on IP” is widely used in the industry, but is not very accurate. It will often miss the actual address by several miles. If the goal is to estimate the data speeds across a large area, such as a state, using IP-based location may be used. If the goal is to map speed test data to a census block such an approach is grossly inadequate.

The final data set visualized in this report, after all the filters have been applied, excludes any test that does not include a GPS-based location. It also excludes, based on a “location accuracy filter”, any test with a self-reported accuracy of greater than 100 meters. This estimated accuracy is most likely calculated by the device operating system. Both iOS and Android have location APIs.

Someone is on a mountain top with a high performance GPS unit that integrates over a long period of time can theoretically locate himself with sub-meter accuracy. Someone connected to Wi-Fi associated with a fixed broadband connection is most likely inside a building, not on a mountain top, and most likely using a smartphone, not an expensive high precision surveying instrument. This means that the device will see fewer satellites and that the signals it receives will have bounced off of multiple reflective surfaces in the home or office, in what is called “scatter”. In an in-building environment there are several tens of meters of inaccuracy that are unavoidable, due to the physics of the radio environment. When one opens the location app on a typical smartphone the “dot” representing the user often moves, even if the phone is still. Sometimes it moves outside of the walls of the building, suggesting that the user might be falling through space. This, of course, is not the case. The device is simply struggling to calculate its location.

Ookla reports GPS locations to three decimal places, for several possible reasons. First, there is a scientific limitation to the accuracy of the reported result. Second, there may be a desire to provide some privacy to the user by locating the result to a very small area.
but not necessarily to a specific address. Three decimal places translate into 111 meters of uncertainty at the equator. At higher latitudes (such as the continental United States) the uncertainty diminishes.

In our filtering process we reject data points with a reported uncertainty greater than 100 meters, on the belief that the GPS chip is likely to be struggling (because it sees too few satellites, perhaps because it is having difficult with reflections, or for some other reason) if the reported uncertainty is greater than this.

What remains, after these filters are applied are the following use cases:

- Laptops and desktops with GPS chips that are connected via Ethernet
- Laptops and desktops with GPS chips that are connect via Wi-Fi
- Mobile devices with GPS chips that are connected via Wi-Fi

In reality, there are relatively few laptops and desktops that have GPS chips. This means that the vast majority of Ookla tests with usable GPS locations come from mobile devices (smart phones and tablets) connected via Wi-Fi.

In contrast, the data collected by the State of North Carolina, which located based on an address provided by the user, probably has a much more representative mix of devices (a lot more laptops and desktops) and a much greater portion connected via Ethernet.

The fundamental trade-off in capturing speed test data is the choice between making it easy for the user by automating everything (the Ookla approach) and accepting a sample of geolocated results that largely excludes laptops and desktops or asking the consumer to enter an address (the North Carolina approach) and living with fewer responses.

Ookla data points that lack precise location data can still be very useful. They can be used to estimate speeds over a large area. They can also be used to estimate the average speed associated with a particular broadband provider if knowing the exact location of the test is not required.

Finally, when a device is connected via Wi-Fi it is testing not only the ISP connection to the home or business, but also the quality of the Wi-Fi indoors. A consumer with an excellent ISP connection may have done a poor job at setting up his Wi-Fi network and therefore may record a poor result.

**Additional Filters**

We apply several other filters to ensure an accurate representation of fixed broadband performance.

- The connection type is limited to either “Wi-Fi” or “Ethernet”. Connection types that represent macrocellular technologies include a long list of cellular technologies. These samples should be excluded by Ookla from the Broadband Performance Dataset. In case any slip through we will filter them out.

- Ookla also has a quality metric. If a data point is deemed problematic by Ookla (perhaps because it has been run repeatedly or for some other reason) then it is excluded from the Speedtest Intelligence Portal. We recognize this quality flag and remove tests that are identified as problematic.
There is a use case that falls in a gray area. If someone has a mobile hotspot (either a standalone hotspot device or a smart phone with a built-in hotspot feature) and has another device connected via Wi-Fi or Ethernet (standalone hotspots sometimes have integrated Ethernet ports) should we include those results? Our response in this study has been “no” although others may choose differently. The argument for a “no” response is that the user is consuming data over a mobile device and is choosing portability over performance. This use case is important but it falls in the category of “mobile broadband”.

Finally, there are an increasing number of mobile providers that also offer fixed broadband service. These providers:
- File with the FCC in specific locations as “fixed” providers.
- Deliver service to a specific physical address (as opposed to a mobile phone or hotspot that can be anywhere).
- Provide CPE that plugs into the wall and looks like a cable or DSL router.
- Provide “flat” pricing plans similar to any cable, fiber, or DSL provider, and unlike typical mobile plans.

We have included these providers in results in this report because, from a consumer perspective, they are competing directly with fixed broadband providers.

We have also included satellite providers in the results. To be categorized as “qualifying” or “underserved” under BEAD a provider must deliver service with less than 100 msec of latency. LEO and hybrid LEO/GEO systems are likely to meet this requirement. A GEO-only system will not meet this requirement because of the time it takes for the signal to reach geosynchronous orbit and return (twice). It will record a result with a high latency. The result will ultimately be categorized as “unserved” due to its latency.

Viewing the Data through the Lens of BEAD

In recent federal grant programs census blocks have been characterized as either “qualifying” or not. In many cases this designation reflects the claim of one or more broadband providers that they offer service at one or more locations within the block at advertised speeds that meet or exceed a threshold.

Under BEAD the thresholds are:
- Qualifying: >=100 Mbps down, >=20 Mbps up, <100 milliseconds of latency
- Underserved: not qualifying, but >=25 Mbps down, >=3 Mbps up, <100 milliseconds of latency
- Unserved: everything else

Definitions of “qualified”, “underserved”, and “unserved” are provided in the Broadband Equity Access and Deployment Program (BEAD) Notice of Funding Opportunity (NOFO) which is available from the NTIA as a 98-page PDF. Page 17/98 defines unserved as:

The term "unserved location" means a broadband-serviceable location that the Broadband DATA Maps show as (a) having no access to broadband service, or (b) lacking access to Reliable Broadband Service offered with - (i) a speed of not less than 25 Mbps for downloads; and (ii) a speed of not less than 3 Mbps for uploads; and (iii) latency less than or equal to 100 milliseconds.

The average man or woman on the street might object to this definition on the premise that “Slightly less than 25 Mbps is good
enough for me” or on the premise that it is important to distinguish between someone who has no connectivity at all and someone who has modest but usable service. Similarly, state regulators might conclude that getting most of the population access to 25/3 Mbps is far more important than achieving 100/20 Mbps. Such officials might resist the more ambitious definitions of BEAD.

Since the federal government is providing $65 billion to states and territories under IIJA and since the BEAD portion (focused on infrastructure deployment) is tied to the BEAD NOFO definitions many readers will end up adopting the vocabulary of BEAD, even if it is not entirely satisfactory to them.

The Infrastructure Essentials BEAD Toolkit provides the user with a wide range of choices on how to visualize data, including an ability to customize the output. This whitepaper is designed to offer ideas on how to visualize data, rather than being prescriptive.

In the future, under the Broadband DATA Act, decisions will reflect performance across a fabric of locations, rather than across census blocks. BEAD grants must therefore wait for results using the fabric.

The methodology in this whitepaper and in the Infrastructure Essentials BEAD Toolkit is to attribute performance at a block level, until the FCC, under the Broadband DATA Act, presents the public with data using a different framework.

Historically, if a block is “covered” in one location it is considered covered, period. This is a reasonable approximation in many urban locations where a block is a physical city block. It is less a reasonable assumption in remote areas where a census block may be very large.

In processing speedtest results we tag and color code each result as either “qualifying”, “underserved” or “unserved” based on the BEAD definition.

We attribute to each census block the status of “qualifying”, “underserved” or “unserved” based on the best single test within the block. If a block has no test then it remains uncharacterized.

Choosing the best result within a block is consistent with the historic FCC methodology. It also mitigates the issue described above with poor reported results as a result of poorly configured Wi-Fi networks. If four homes within a neighborhood have a similar broadband connection and all report results then hopefully one of the four will have a thoughtfully configured Wi-Fi network and will report results reflective of the speed of the ISP connection.

Having multiple simultaneous users within a household impacts the reported results, since various family members are competing with the speed test application for bandwidth. Choosing a “cheap” broadband plan also tends to portray the ISP in a poor light. The broadband provider may offer very robust service but a particular consumer might say “I don’t need it” or “I can’t afford it” or “basic service is good enough for me”. Choosing the best of several tests within a block at least mitigates these concerns.

In larger areas of aggregation (block groups, tracts, counties, and other polygons) the following metrics can be calculated based on a statistical characterization of individual speed tests:

- Percentage of qualifying test results within each polygon (block group / tract / etc.)
- Median download speed within each polygon
- Median upload speed within each polygon
• Median latency within each polygon

Since these statistical calculations examine many test results within a larger polygon one is less likely to see polygons that cannot be characterized as a result of a lack of data.

**Background Information**

BroadbandToolkit.com (a business of Signals Analytics, LLC) asked Ookla to provide data for this whitepaper for a small geographic area within the United States, preferably one that includes both suburban and rural geographies. Ookla chose Pueblo County, Colorado. The visuals are based on a 12-month sample that includes 91,616 Speedtest results (before filtering) run from August 1st, 2021 to July 31st, 2022. Both the trademarks and data of Ookla are used with permission.

In Pueblo County three “traditionally mobile” operators have filed as fixed operators under Form 477. These include T-Mobile, Verizon, and Viaero Wireless. In many cases we can distinguish between fixed and mobile (e.g. hotspot) speedtests. When in doubt we have included these tests in an effort to recognize the emerging category of “historically mobile” firms competing in the fixed broadband space.

Adding even more complexity: many ISPs are still using IP v4. Just as telephone company once ran out of telephone numbers ISPs are now running out of IP addresses (at least IP v4 addresses). The result is that blocks of IP v4 address are often traded between ISPs and the same block may be used by a particular ISP for both mobile and fixed applications. Such practices complicate the task of data analysts to distinguish between ISPs (who is providing the service) and between service offerings (fixed vs. mobile or consumer vs. business).

In this report 38,360 tests (41.9%) out of 91,616 passed the rigorous “filter” criteria. The most common reason for not using a test was the lack of a precise GPS measurement (44%).

Ookla has provided a robust data set to BroadbandToolkit.com and has explained field definitions and the data collection process.

**Conclusions**

Robust speed test data offers a compelling tool for identifying unserved and underserved areas. In some cases it will confirm ISP data submissions to the FCC. In other case it may challenge potentially overstated numbers. Either way, it represents an independent data-driven view.

The data is likely to be current, showing new competitors, improved levels of performance, and expanded service areas relative to the most recently released set of FCC Form 477 data.

State broadband offices, investors, and grant applicants using the Ookla Broadband Performance Dataset will soon be able to integrate that data with a multitude of other data sets using the Broadband Essentials BEAD Toolkit.

Such an integrated view will allow broadband offices and infrastructure providers to identify precise areas where ISP-provided data may be overstated (triggering a range of possible challenges) or where socioeconomic needs are acute. States will be able to identify those precise unserved or underserved areas where infrastructure investment is likely have a very significant impact.
Resources

Valuable resources include:

White papers:  
https://broadbandtoolkit.com/pages/whitepapers

Product Documentation:  

Need help understanding the concepts and/or the available tools and data sets? If so, send us an email or give us a call (415-346-5393). We’ll be happy to step through the available resources and – if desired – demonstrate various tools via a Zoom / conference call.

Disclaimer. This whitepaper reflects one possible view of the subject. Readers are encouraged to carefully read all original data sources, to run the numbers themselves, to discuss these concepts, methodologies, and interpretations with their colleagues and with other subject matter experts, and to consult competent legal counsel for any issues involving an interpretation of the law.

Publication Date: November 17th, 2022