

# Data Processing Methods

---

## Overview

In Round 8, the California Public Utilities Commission identified 225 potential broadband providers, 139 of whom did not submit data, and 86 who did. This represents an increase of 1 provider over the prior Round. Together, these 86 providers comprise over 99.9% of the total broadband connections in California, according to the June 2012 FCC Form 477 data.

In contrast to previous Rounds, we have eliminated zones from our submission areas where we could not validate the existence of broadband service availability prior to submitting data to the NTIA as well as the California state broadband availability map. This resulted in a 7% reduction overall in census blocks (wireline, mobile, and fixed wireless combined) compared to the December 2012 submission. The impact of the validation process was more significant on terrestrial fixed wireless providers, many of whom do not regularly file Form 477 data, if at all. Using the validation process outlined in the Appendix, any areas that we were unable to validate based on internal and third party data sources were removed from the provider's area. The validation process reduced the land area covered by fixed wireless providers, as measured in square miles, by 38% compared to the December 2012 data submission. In extreme cases where none of a provider's area could be validated, the provider's entire coverage area was removed. Out of the 32 providers of fixed wireless service, 10 were removed from the submission for this reason (only one wireline provider, Raw Bandwidth, was removed). A summary of the changes is provided in the Appendix.

## Data Collection

The California Public Utilities Commission (CPUC) sent out a Data Request to broadband providers to initiate the Round 8 data collection. Potential providers were strongly encouraged to submit broadband availability data. Providers who previously submitted data were also sent maps displaying their Round 7 coverage and validation results to guide their Round 8 submissions. Data submission instructions were posted online to assist providers along with template files, sample shapefiles and record formats on the CPUC Broadband Mapping Website at:

<http://www.cpuc.ca.gov/PUC/Telco/Information+for+providing+service/BroadBand+Mapping.htm>

The data submission instructions point each provider to the wireless and/or wireline datasets, which are separated into sections for those with GIS data (shapefiles or filegeodatabases) and those without GIS data (text or Excel files). For providers with GIS capabilities, statewide census block and TIGER/Line shapefiles were provided on the CPUC website. The square mileage of each block was calculated in advance in the sample census block shapefile. Using the shapefiles, providers were able to determine which blocks in their footprint were less than two square miles and which were two square miles or

greater and therefore needed to be represented using the road segment shapefile. For providers without GIS capabilities, Excel spreadsheets were provided incorporating record field formats adhering to the NOFA data submission requirements.

## Community Anchor Institutions (CAI)

CAI data is composed of the names and locations of schools, colleges, libraries, healthcare institutions and other community institutions, and the broadband technology and capacity of these institutions.

For each of these institutional categories, the facility data was sourced from the following locations:

- K-12 school data: <http://www.cde.ca.gov/ds/si/> and <http://nces.ed.gov/ccd/elsi/>
- College data: <http://nces.ed.gov/ipeds/datacenter/login.aspx>
- California health facility data: [http://oshpd.ca.gov/General\\_Info/Healthcare\\_Atlas.html](http://oshpd.ca.gov/General_Info/Healthcare_Atlas.html) and <http://www.caltelehealth.org/>
- Library data: <https://harvester.census.gov/imls/data/pls/index.asp>

Except for some of the health facility data, the above data sources do not have broadband technology information. We contacted the State Librarian to see if we could improve our reporting on library broadband capabilities. They are in the middle of conducting a survey of libraries throughout the state, and we hope to provide a more complete picture in the next round of data submission. We are also working with the California Department of Education, which is conducting a technology readiness assessment of all public K-12 schools prior to the rollout of computer-based standardized Common Core testing. We expect to have those survey results in October this year. Other areas we need to improve are medical and colleges. For medical, we are planning to contact California Telehealth Network, and for colleges, we are planning to contact the University of California and California State University systems to see if we can get more complete data from them.

Broadband technology and capacity for these facilities was captured using the following data sources:

- CPUC data from the California Teleconnect Fund (CTF) program with additional provider data from AT&T and TelPacific. The CTF program provides 50% discounts on telecommunications bills for qualifying schools, libraries, government-owned and operated hospitals and health clinics, and other community based organizations.
- Corporation for Education Networking Initiatives in California (CENIC) data. CENIC operates the K-12 High Speed Network (K12HSN) program which is funded by California Department of Education. K12HSN enables educators, students and staff across the state to have access to reliable high speed network to deliver high quality online resources to support teaching and learning and promote academic achievement.

Additionally, other nongovernment community organizations not listed above were sourced from the CTF data.

Finally, the CAI addresses were geo-coded to point locations and geo-matched to the Census Blocks 2010 shapefile to obtain the corresponding FULLFIPSID. Technology data was then associated with broadband technology and speed information.

Below is a summary by institution type of the broadband subscription data we have been able to collect for this round. There was an increase in the number of records for Community Based Organizations (CBOs) by roughly 1,000 compared to the previous round. This was due to the previous round’s data set being fairly old (mid-2012).

	All	Schools	Libraries	Medical	Colleges	CBOs
<b>Total Count</b>	26,018	12,999	1,137	5,338	768	5,776
<b>With Tech</b>	10,109	7,936	19	496	10	1,648
<b>Unknown</b>	15,909	5,063	1,118	4,842	758	4,128
<b>% with Tech</b>	38.9 %	61.1 %	1.7 %	9.3 %	1.3 %	28.5 %

## CPUC Initial Data Verification

Each data set submitted by broadband providers was reviewed against the GIS data model posted on the SBDD Network website and checked if mandatory fields were filled in, and if each field contained the appropriate range of values. Where possible, we made certain that appropriate field headers were used and that each field contained the correct data type. When data was found to be missing or incorrect, the provider was contacted and the issue was documented in the Changes and Corrections document.

## Geo-processing

After the initial CPUC review, data was transferred to the Geographical Information Center (GIC) at CSU Chico for geo-coding, geo-matching, propagation of wireless service by antenna, and validation of geographic data. In those cases where the CPUC received street address level data from broadband providers, such addresses were assigned a point location, (geo-coded) and then geo-matched to census blocks and street segments.

Wireless providers who were unable to submit a shapefile or geographic representation of their service area were asked for tower, antenna, and radio settings information. We used these and other parameters in EDX’s Signal software, version 11.0.1, to model the service area, and from that we created a shapefile. In cases where a fixed wireless provider offers service at different speed tier combinations, a separate propagation was run and a separate shapefile was created for each. All shapefiles were then overlaid and dissolved so that only the maximum advertised speed available for each area appeared. The EDX propagation used the Anderson-2D propagation model. Individual unit specifications were used to predict coverage based on frequency, transmit power, receiver sensitivity, antenna gain, and height. The propagation model took into account terrain based on two datasets, EDX universal .201, and SRTM

3-second .HGT format. The model also took into account land use/clutter using USGS 2006 30m data (.151 files).

## Validation Methods

Please refer to the Appendix for an explanation of the validation process employed for this data submission.

## CPUC Final Data Verification

The resulting datasets were delivered from Chico to the CPUC in the SBDD transfer model geodatabase for final review and verification. Data sets were checked again and reviewed for unexpected changes resulting from the geo-coding/geo-matching process. Geo-processed data was visually reviewed using ArcGIS to verify service area footprints, and the SBDD check submission Python script was run on each dataset to identify unexpected values.

## Deliverable Data

The final dataset is delivered to the NTIA/FCC in filegeodatabase format containing the following feature classes:

- BB\_ConnectionPoint\_MiddleMile – Point between the local “last mile” network and the middle mile network which goes on to connect to the internet backbone. This is a confidential dataset.
- BB\_Service\_CAInstitutions – Community Anchor Institutions: points geo-coded from address lists
- BB\_Service\_CensusBlock – Broadband availability polygons for areas less than 2 square miles
- BB\_Service\_Overview – Service overview by County including Subscriber Weighted Nominal Speed
- BB\_Service\_RoadSegment – Broadband availability line segments for areas 2 square miles and greater
- BB\_Service\_Wireless – Wireless service area polygons.

## Changes and Corrections Reporting

In reporting changes and corrections for this round of data collection, we included the following, where applicable:

- 1) Submitted new data, or no changes from previous round 7
- 2) New provider
- 3) Changes through company merger/acquisition
- 4) Changes to FRN number

- 5) Changes in speeds, middle mile, and/or spectrum
- 6) Changes in number of blocks and road segments for wireline
- 7) Changes in coverage area for mobile and fixed wireless

## Appendix: Validation Process

### Overview

One of the requirements of our NTIA Mapping Grant is that we validate the data submitted to us by broadband providers. This document explains the validation methodology we used. The following data sources were used to validate each provider’s data submission. Areas we were unable to validate do not mean there is no service there, or that service at a particular speed is not available, it simply means that we were unable to confirm the presence of service based on the data sources available to us.

Here is a summary of the changes between the December 2012 data submission and what is being submitted this round.

Measurement	December 2012 submission	June 2013 submission	% change
Census blocks (all technologies)	4,766,136 blocks	4,442,224 blocks	-7%
Census blocks (wireline)	1,178,045 blocks	1,105,446 blocks	-6%
Square miles (mobile)	357,204 sq. mi.	348,639 sq. mi.	-2%
Square miles (fixed wireless)	49,752 sq. mi.	30,792 sq. mi.	-38%

### Data Sources

The table below summarizes the validation method, data type, and to which type of broadband connection the validation method applies.

Method	Data Type	Fixed: Wireline	Fixed: Wireless	Mobile Wireless
<b>FCC Form 477</b>	Number of subscribers by speed tier combination by census tract	YES	YES	NO
<b>Broadband Scout (ID Insight)</b>	Online purchases sorted by provider and census block or street segment	YES	YES	YES
<b>TeleAtlas Wire Center (CalAtlas)</b>	Serving wire center locations of telephone companies	YES	NO	NO
<b>CPUC Mobile Field Tests</b>	Provider-specific, “In coverage” location results showing “No	NO	NO	YES

	Effective Service” (point data) from 3rd round mobile field testing			
<b>CalSPEED results</b>	Speed test results from LTE-capable devices and “No Effective Service” results from ANY device	NO	NO	<b>YES</b>
<b>Customer address service and speed information</b>	Provider-supplied list of customers showing their address and subscribed speeds, by provider (if available)	<b>YES</b>	<b>YES</b>	NO
<b>Tower data and/or EDX propagation image</b>	Coverage propagation of fixed wireless provider based on tower, radio, and antenna data submitted by the provider	No	<b>YES</b>	NO

### Explanation of Data Sources

- FCC Form 477** For fixed services, the FCC collects data from each broadband provider twice a year, including the number of broadband connections by technology type and speed tier combination for each census tract where the provider has customers. Mobile broadband service data only includes state-wide totals and is not useful for geographic validation. If a provider indicates it has broadband service in a particular census block but has not reported customers to the FCC for the census tract where that block resides, the Form 477 data cannot validate the actual presence of service. In the case of speed validation, if a provider has not reported any subscribers receiving the speeds which they advertise for any of the census blocks within the applicable census tract, then Form 477 cannot validate the speed for the entire census tract. However the existence of coverage in the entire census tract is still validated. As with any validation technique, there are inherent errors. For example, if Form 477 data shows that a particular provider has customers in a census tract and at the maximum advertised speeds submitted to us, we consider all blocks within that census tract validated for speed and/or availability for that provider. Because Form 477 data is only available at the census tract level, this validation tool tends to yield false positives and overstate the number of validated areas. Conversely, the most recent Form 477 data available to the CPUC is usually six months old or more. This time lag between when faster service is made available and when customers are shown for that service may yield false negatives and understate the number of areas where speed is validated.
- BroadBand Scout** is a third party, comprehensive dataset designed specifically to show the providers, connectivity, speed, and usage details of the national broadband landscape. ID Insight’s process analyzes hundreds of millions of Internet transactions that link a consumer's physical address to their Internet provider. BroadBand Scout data comes to us aggregated at the census block level.. The presence of an on-line purchase over a particular provider’s network validates service availability for that census block or street segment. The upstream and downstream throughput for an online purchase may be used to validate a provider’s advertised speed.

- **TeleAtlas Wire Center data** lists every Local Exchange Carrier (LEC) landline wire center in the United States. The term “wire center” refers to the location where the telephone company terminates its local lines; this is usually the same location as a central office, although a wire center might house multiple central offices. Buffers were created at 12,000 feet and 18,000 feet from provided Wire Center point datasets to cross reference ISP data submissions to the CPUC. The wire center boundary is a representation of the area served by all of the switching equipment housed at that physical location. When a provider indicates broadband availability in a particular census block, and that location is within the distance from the wire center to support a given speed, that census block is considered validated. If the location is farther from the central office than what can support the maximum advertised speed, we are unable to validate the service.
- **CPUC Mobile Field Tests** are conducted twice a year at 1,200 randomly selected points across the state and measure broadband performance for the four major mobile wireless operators: Verizon, AT&T, Sprint, and T-Mobile USA. The point data results from the April 2013 tests were compared against each operator’s advertised availability in the census block where the test was conducted. In census blocks where the test result for a particular operator was zero or “No Effective Service,” but the operator advertised coverage there, the coverage for that census block was considered un-validated.
- **CalSPEED Results** are crowd-sourced mobile test results from the CPUC’s Android mobile testing application. The CPUC launched CalSPEED on Google Play’s app store on April 5, 2013. The point data results through June 30, 2013 were compared against each operator’s advertised availability in the census block where the test was conducted. These results included operators beyond the four tested for the bi-annual mobile field testing. In census blocks where the test result for a particular operator was zero or “No Effective Service,” but the operator advertised coverage there, the coverage for that census block was considered un-validated.
- **Customer Address Service and Speed Information**  
In limited cases where we were unable to validate any areas of a provider’s availability (their entire footprint was a red zone), we requested customer address information to use as a validation data source. Census blocks where customers resided were considered validated.
- **Tower data and/or EDX propagation image**  
For fixed wireless providers, we used tower location and system parameter information, where available, to propagate a fixed wireless provider’s coverage area using EDX’s Signal software, version 11.0.1. The wireless propagation model is based on the Anderson-2D propagation model. System parameters included frequency, transmit power, receiver sensitivity, antenna gain, and height. EDX produced coverage patterns for each tower/sector combination taking into account terrain and land use/clutter that may hinder signal dispersion. For terrain, we used two data sets, EDX universal .201 and SRTM 3-second .HGT format. For land use/clutter, we used USGS 2006 30-meter .151 files. A separate propagation shapefile was created for each downstream and upstream speed tier combination, and all shapefiles were later overlaid and dissolved to where only the fastest advertised speed available was visible. In cases where a

provider submitted a shapefile or vector file of their coverage, the EDX propagation was used as one source for availability and speed validation.

## **Validation Procedures**

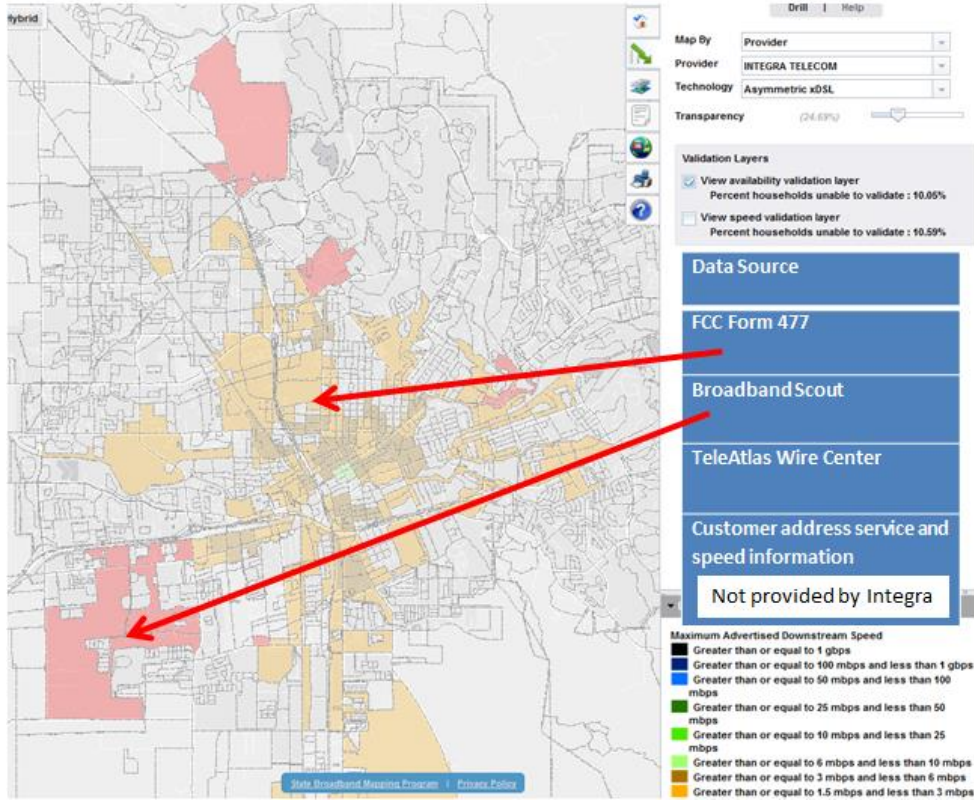
### **Wireline**

A spatial selection was performed on Census Block and Street Segment data, either submitted by the provider, or created from submitted address records through a geocoding/spatial selection process, to derive only those blocks or street segments which intersect polygons in a given validation layer. Counts are recorded as number of unique blocks or unique segments which share geographic area with any given validation layer, compared to the total number of unique blocks submitted by, or created for, a given provider. Percentages are recorded as percentage of the total number of unique blocks or street segments which share geographic area with any given validation layer, compared to the total number of unique blocks submitted by, or created for, a given provider.

Validation data sources: we used FCC Form 477 (December 2012 for California cable providers who submitted 477 data under the Digital Video and Competition Act, and June 2012 for all other providers), Broadband Scout's ID Insight, and TeleAtlas Wire Center data sets. In limited instances customer addresses were used to validate availability and/or speed.

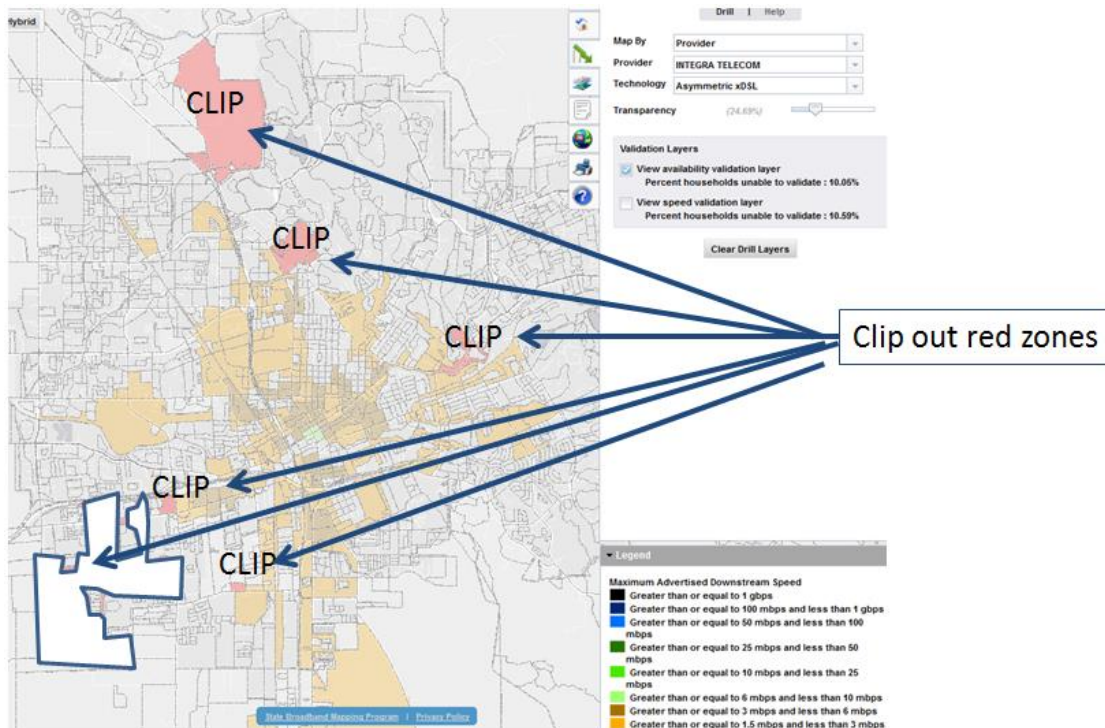
The image below is an example of how "red zones" were created for wireline providers. A red zone indicates a census block or census tract, depending on the third party data source, where we were unable to validate service availability based on subscriptions reported to the FCC or online transactions reported in Broadband Scout. In the example below, Integra, a provider of ADSL, DSL, and "other copper" broadband services, contains a number of red zones resulting from our validation. Conversely, areas shown in goldenrod were areas where we were able to validate service availability based on the presence of subscribers reported to the FCC or the presence of an online transaction from Broadband Scout. The goldenrod color represents Integra's maximum advertised downstream speed tier.





Following creation of the red zones from our validation process, the red zones were clipped from the provider's shapefile. The post-validation shapefile (goldenrod areas only in the example below) is what was submitted to the NTIA and shown on the California state broadband availability map for this round

of data collection.

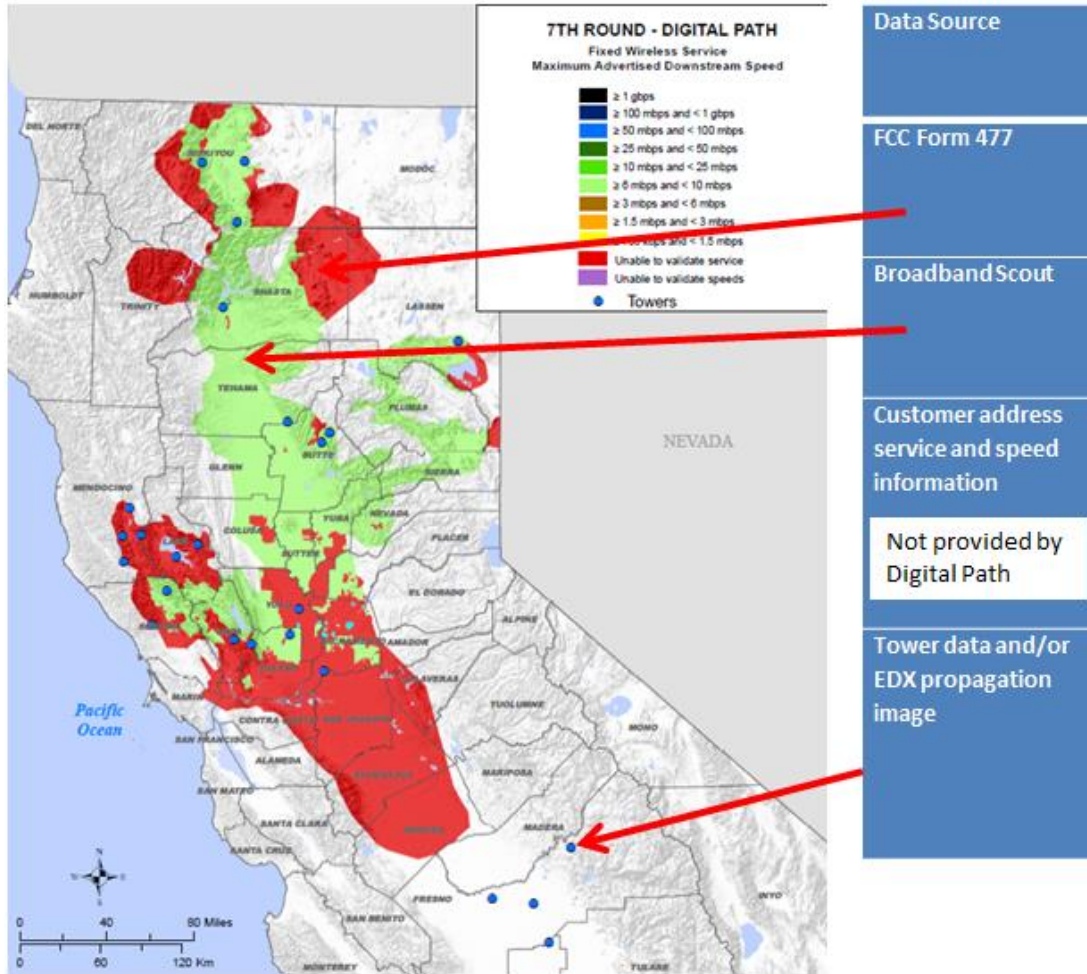


## Fixed Wireless

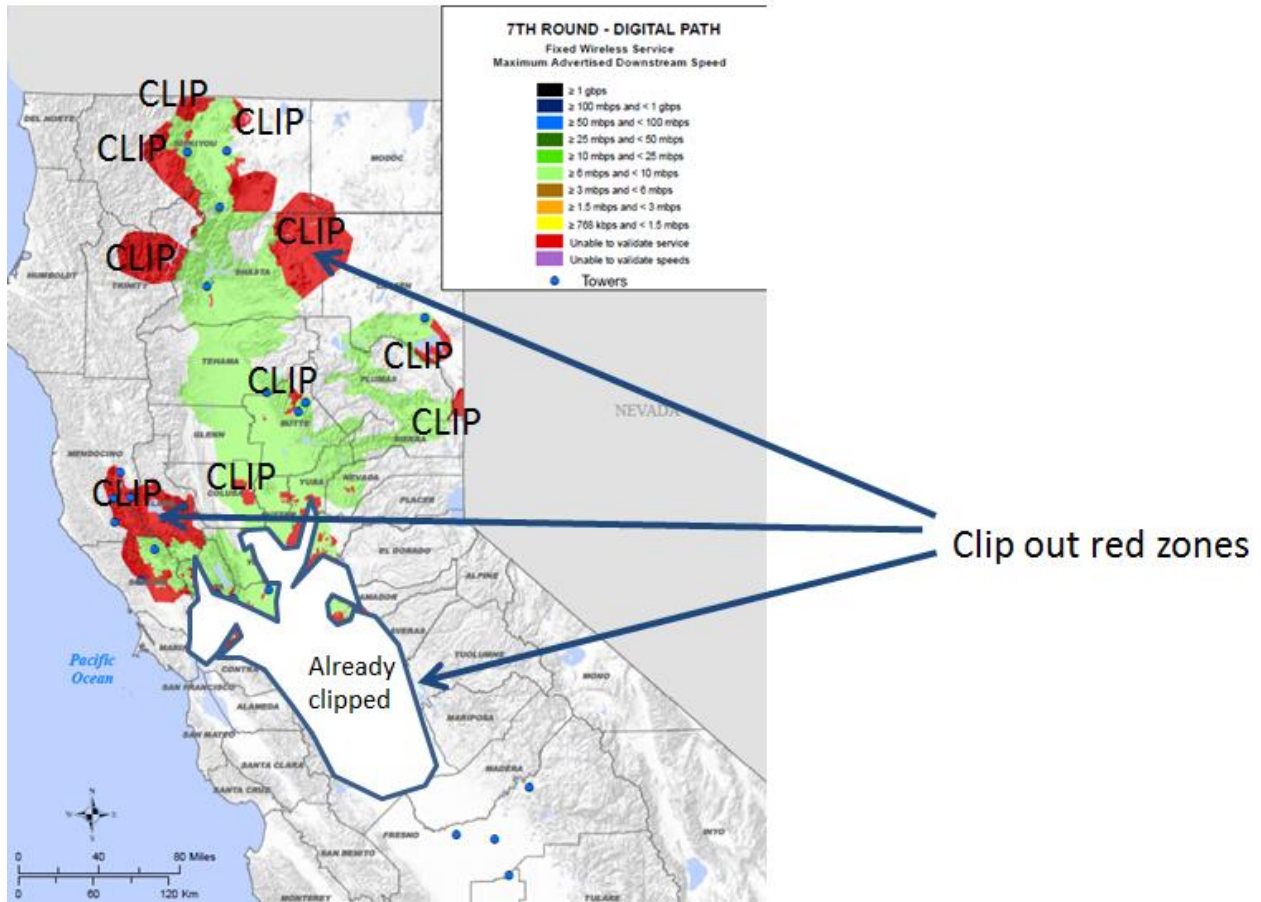
A spatial overlay was performed on Wireless Availability data, either submitted by the provider, or created from tower and antenna location information, to select only those polygons which intersect a given validation layer. Results are recorded as a percentage of the total geographic area of wireless coverage sharing geographic area with any given validation layer compared to the total coverage area submitted by, or created for, a given provider.

Validation data sources: we used FCC Form 477 (December 2012 for California cable providers who submitted 477 data under the Digital Video and Competition Act, and June 2012 for all other providers), Broadband Scout's ID Insight, and EDX propagation images. In limited instances customer addresses were used to validate availability and/or speed.

The image below is an example of how "red zones" were created for fixed wireless providers. A red zone indicates a provider's coverage area polygon where we were unable to validate service availability based on subscriptions reported the FCC, online transactions reported in Broadband Scout, and in limited instances, customer addresses or propagated tower data. In the example below, Digital Path provided both a shapefile of their service area and speeds as well as tower data. In comparing the two, we noticed large discrepancies between their claimed coverage versus their tower locations. We also noticed a large difference between areas we were able to validate, shown in green, versus areas we were unable to validate based on the third party data sources mentioned above.



Following creation of the red zones from our validation process, the red zones were clipped from the provider's shapefile. The post-validation shapefile (green areas only in the example below) is what was submitted to the NTIA and shown on the California state broadband availability map for this round of data collection.



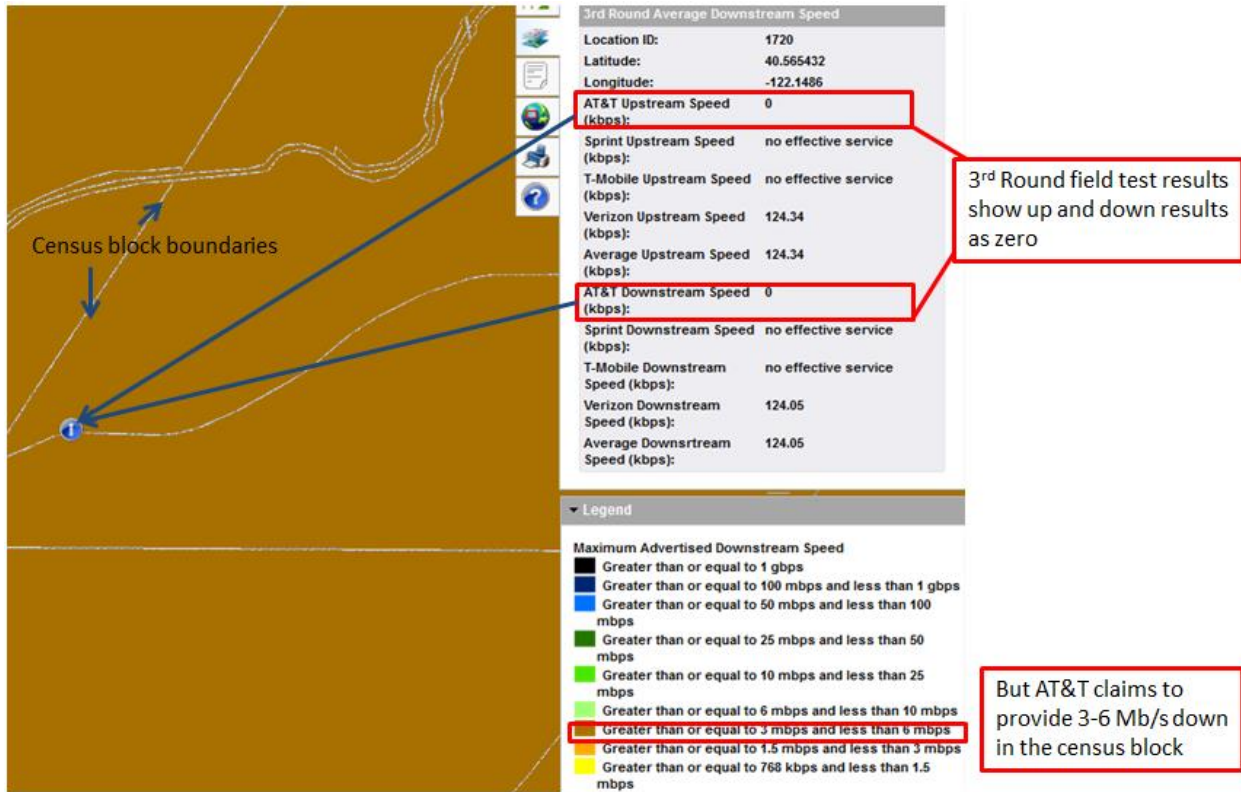
## Mobile

A spatial selection was performed on Wireless Availability data, either submitted by the provider, or created from tower and antenna location information, to select only those polygons which intersect a given validation layer. Results are recorded as a percentage of the total geographic area of wireless coverage sharing geographic area with any given validation layer compared to the total coverage area submitted by, or created for, a given provider.

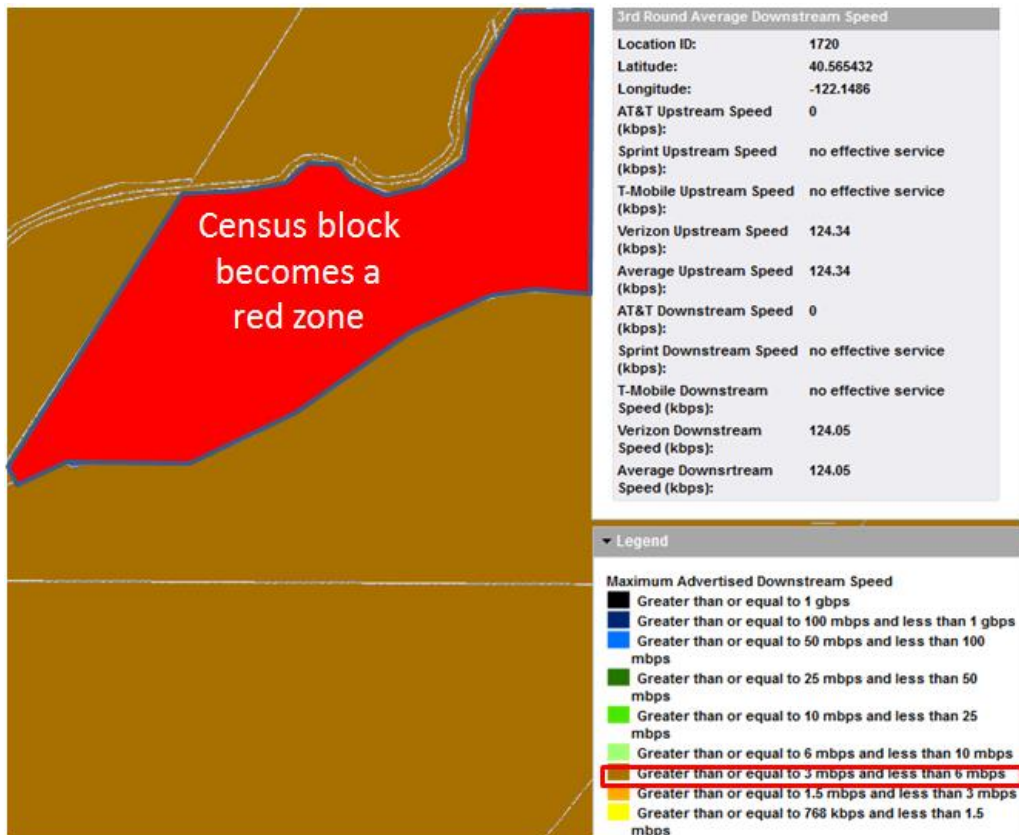
Validation data sources: we used Broadband Scout's ID Insight, CPUC's mobile field testing data from April 2013, and CalSPEED results through June 30, 2013. As stated earlier, FCC Form 477 for mobile broadband providers is aggregated the state level rather than census tract and is not useful for geographic validation. Due to the lack of geospatial data for validation, the default value for mobile provider coverage is validated area. Specific census blocks where a test result (mobile field test or CalSPEED) shows either "No Effective Service" or a zero value but falls in a provider's advertised coverage area become red zones.

The example below shows a mobile field test result of zero kilobits per second for AT&T Mobile at location #1720, which is within AT&T Mobile's advertised coverage area. In this example, the census block where the test location falls becomes a red zone.

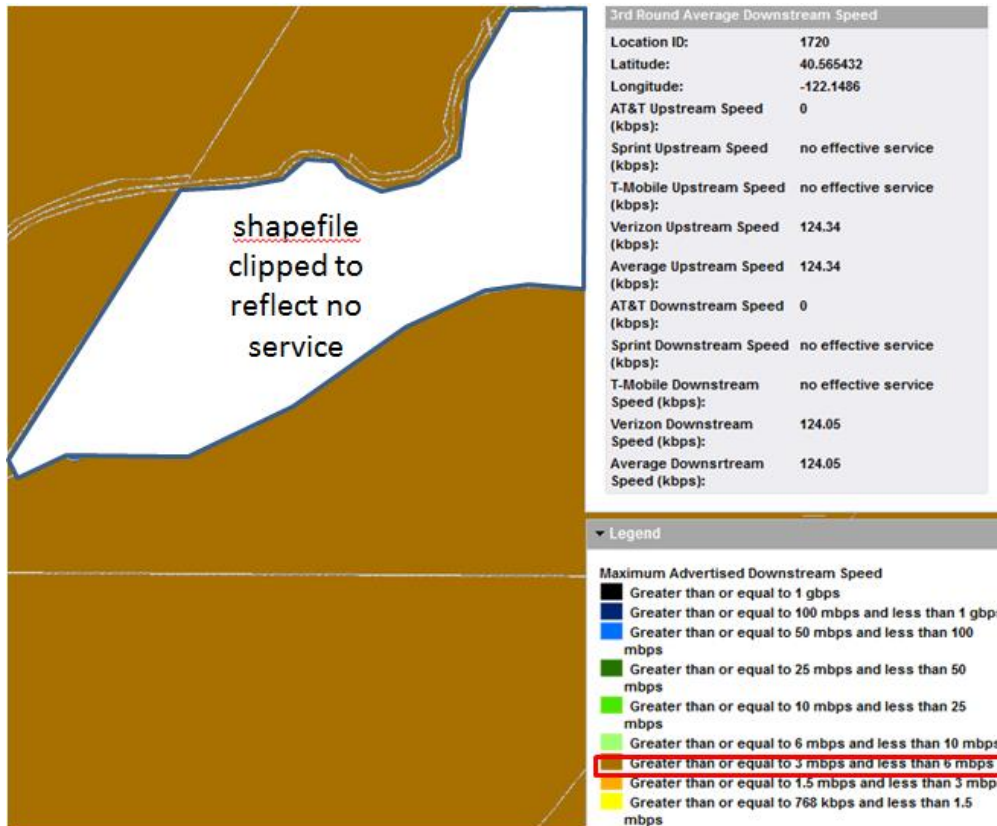




The census block where test result was zero becomes a red zone



Following creation of the red zones from our validation process, the red zones were clipped from the provider's shapefile. The post-validation shapefile (brown areas only in the example below) is what was submitted to the NTIA and shown on the California state broadband availability map for this round of data collection.



## Process Evolution

As we gather more validation data and add to our list of available third party data sources, we plan to improve the accuracy of validation. With the previous round of data collection, we sent a copy of the red zones to each provider and solicited feedback. Those that provided feedback often provided additional data such as customer addresses to help us better validate service availability and speeds. Many, however, did not provide feedback. It is our hope that this round of data collection will spur further engagement with the providers and help us improve validation for the next round.