NC Broadband Survey Data README

September 2020 Edition

OVERVIEW

This document provides the starting point for understanding the documentation for the NC DIT Broadband Infrastructure Office 2020 Broadband surveys. The data consists of three specific surveys, with results merged into a single dataset.

The surveys are:

1. 2020 Farm Broadband Survey
2. 2020 General Broadbandy Survey
3. 2020 Phone-based Internet Reporting Tool

ACCESSING AND DOWNLOADING SURVEY DATA

Publicly available data are provided via these dashboards, and are updated regularly:

https://go.ncsu.edu/2020biopublicdash

https://go.ncsu.edu/2020biopublicmap

By viewing the interactive dashboard/map users may select specific data elements of interest. For example, a single county can be selected, visualized and the data associated with it then download.

See the document entitled ***Accessing Broadband Survey Public Data*** for specific details on downloading the data from the NC State University Tableau Server which hosts these dashboards.

For access to the dataset as a native MySQL table, please contact the NC DIT Broadband Infrastructure Office.

CODEBOOKS

The codebooks provide the content of the surveys as well as all the possible answers. Also included in the codebooks is the "flow logic" which shows which questions might be skipped based on a user's answers. For example, if a user indicated they do not have cellular service, they would not be presented with a question asking how satisfied they are with their cellular service.

There are three codebooks available, one for each of the surveys:

- ***NC_BIO_Farmer_Survey-2020-FullCodebook.pdf***
- ***NC_BIO_General_Survey-2020-FullCodebook.pdf***
- ***Broadband Survey Phone-Based Reporting Tool (Google Doc)***

It is important to understand the exact wording of each question and the options the user was presented when interpreting the data.

DATABASE FIELD DESCRIPTIONS

The data is provided to the public in the form of a single dataset, available in .csv or Excel format. Other formats are available upon request. Regardless of the format, the fields provided are documented in ***Broadband Survey Master Mapping - Field Descriptions-PUBLIC.pdf***

This document is key to understanding the names of the fields, their possible values, which of the three surveys provide input to the field and other details.

Fields that are only available from responses from the Phone-based Reporting Tool will begin with p_ or p_n_. The p_n_ indicates that they are "normalized" across media and languages.

Fields that are only available from response from the Farm Survey will begin with f_.

In addition, there are a number of fields that begin with c_, so called "calculated fields", which are discussed below.

CALCULATED FIELD DESCRIPTIONS

In addition to the raw data provided by users as they complete the survey, the dataset also includes a number of calculated fields, all of which being with c_.

For example, there is a field called c_k12_student_present which can either be True or NULL. This handy shortcut field allows SQL queries, or other data analysis tools to quickly filter out

records associated with users that have affirmed that they have a K-12 student present at their location.

The document entitled, ***Broadband Survey Calculated Fields (Google Doc)***, provides the exact definition as to how each of these fields is calculated from each of the three survey sources.

SPEEDTEST DETAILS

The speedtest is based on LibreSpeed, an open source internet connection testing tool. The tool uses approximately 10 simultaneous TCP connections to a server hosted in Research Triangle Park, NC at MCNC. That server has multiple 10-gig interfaces and is believed to be central located and peered with numerous top tier internet providers. The goal of placing the server here is to help reduce variability in testing as a result of packets traversing long distances to huge data centers in locations like Denver, San Jose or Ashburn.

Details about the speedtest internals as well as the source code are available at: https://github.com/librespeed

THIRD PARTY SOURCES

The dataset includes a number of fields that are acquired from third party sources.

Geocodes

Addresses are geocoded using the Google Geocode service: https://developers.google.com/maps/documentation/geocoding/overview

Note that the actual address is not available in the public dataset. Service providers with a bona fide need for street address detail may obtain that data from NC DIT Broadband Infrastructure Office. The public data does include an obfuscated lat/long coordinate for each address in the data set. Note that this is not a perfect process and Google sometimes returns results that are wildly incorrect.

Census Data

The census block number is obtained from the US Census system at: https://www.census.gov/data/developers/data-sets/Geocoding-services.html

Note that the actual (not the obfuscated) location is used as a parameter to this service. So the true census block of the user is always in the data set, even if the obfuscated lat/long is just outside of the block.

Service Provider Data

For users of the online surveys the service provider is determined based on the IP address of the client using IPinfo: https://ipinfo.io/developers

The BGP AS and service provider name are provided from this source. For a small number of users (less than half of a percent) we have found that their IP address is either not determined, or is somehow reported wrong. Those records will contain a NULL IP address and thus will not have a service provider associated with them.

It should also be noted that a small number of users completed the survey while connected to a VPN. These users will have the names of private companies or government organizations as the service provider, since their packets were routed through the VPN on their employer's AS.

The location of the service provider may or may not be near the user. For example, a number of satellite users have service provider locations on the west coast or central US. This is a result of which Earth ground station their IP address is associated with. Other technologies may also have their IP address assignment tied to a location that is seemingly nonsensical. However, in an attempt to provide as much data as possible, we have kept these elements in the data set and suggest more advanced users of the data dig deeper into understanding the derivation of these elements before drawing conclusions based on them.

Telephone Company Info

For users of the phone-based survey, the name of the telephone company associated with the caller as well as the type of phone (cellular, landline, voip) is determined from this service: https://www.twilio.com/docs/lookup/api

The Mobile Network Number is only available for cellular users.