Deliverable 2

This document summarizes the work done for Deliverable 2, for the BU Spark! project, Spring 2021, MAPC Broadband Digital Equity in MA.

Date: 03/04/21

Student Team

This project has two different teams, denoted MAPC team 1 and MAPC team 2. We represent team 2. There are five students, and one project manager for team 2:

- Adam Streich
- Jenny Li
- Nathan Lauer
- Yutong Shen
- Zhixing Zhao

The project manager is Kamran Arif.

Contact

- Ryan Kelly, RKelly@mapc.org Digital Services lead at the MAPC
- Matt Zagaja, mzagaja@mapc.org , Lead civic web developer at the MAPC

Organization

The Metropolitan Area Planning Council - MAPC

Purpose

The primary objective of the previous deliverable was simply to obtain the data; included within deliverable 1 were usable csv files of publicly available broadband data from both Ookla and MLAB for the 2020 calendar year. In this deliverable, we build upon the previous work, with two primary outcomes.

The first outcome is to begin in-depth analysis of these datasets, with a particular focus on the following points:

- Develop a statistical understanding of the datasets
- Label the data with Provider information, for both Ookla and MLAB
- Produce sub-datasets for each municipality, for both Ookla and MLAB
- Compute average broadband speed for each provider, per municipality.

The second outcome is to develop a clear and detailed plan for further development work in this project, by specifying the desired final analysis, obtaining the remaining data sets, and listing the steps required to complete this body of work.

The following sections will discuss each of the above primary outcomes of this deliverable in more detail.

Labeling Ookla Data

In the previous deliverable, the Ookla data was labeled with county information, as this information information is publicly available and easy to obtain. Unfortunately, labeling this data by county does not correlate well with the previous work done by MAPC; as much of MAPC's work is grouped by municipality -- a finer grain resolution than by county -- it was necessary to further granularize the data points in Ookla, by labeling each data point with a specific municipality.

Fortunately, MAPC already had a dataset with a geographically defined area for each polygon. That data can be found here: https://datacommon.mapc.org/browser/datasets/390. Using this dataset, we were able to label each row in the Ookla data with municipality information.

To be clear, previously a given data point in the Ookla data set was featured as such:

- Schema:
    - quadkey: a key that identifies the tile
    - avg_d_kbps: the average download speed in kilobits per second within the tile
    - avg_u_kbps: the average upload speed in kilobits per second within the tile
    - avg_lat_ms: the average latency of the tests in this tile.
    - tests: The number of tests that contributed to the other values in this tile.
    - devices: The number of unique devices that contributed to the data in this tile.
    - geometry: list of latitude/longitude pairs, that collectively form the polygonal shape of this tile.
    - STATEFP: state FIPS code. It's 25 for MA.
    - COUNTYFP: county FIPS code.
    - COUNTYNS: Another unique county identifier.
    - GEOID: unique ID for the geographic location of the county.
    - NAMELSAD: full name of the county
- Example data point: 0302332121321131,141598,56138,11,55,27,"POLYGON ((-71.1090087890625 42.3504251224346, -71.103515625 42.3504251224346, -71.103515625 42.3463653316019, -71.1090087890625 42.3463653316019, -71.1090087890625 42.3504251224346))",25,021,00606937,25021,Norfolk County

Note that this data point was labeled as having been obtained in Norfolk County, without any municipality information.

Now, the schema has changed slightly, as such:

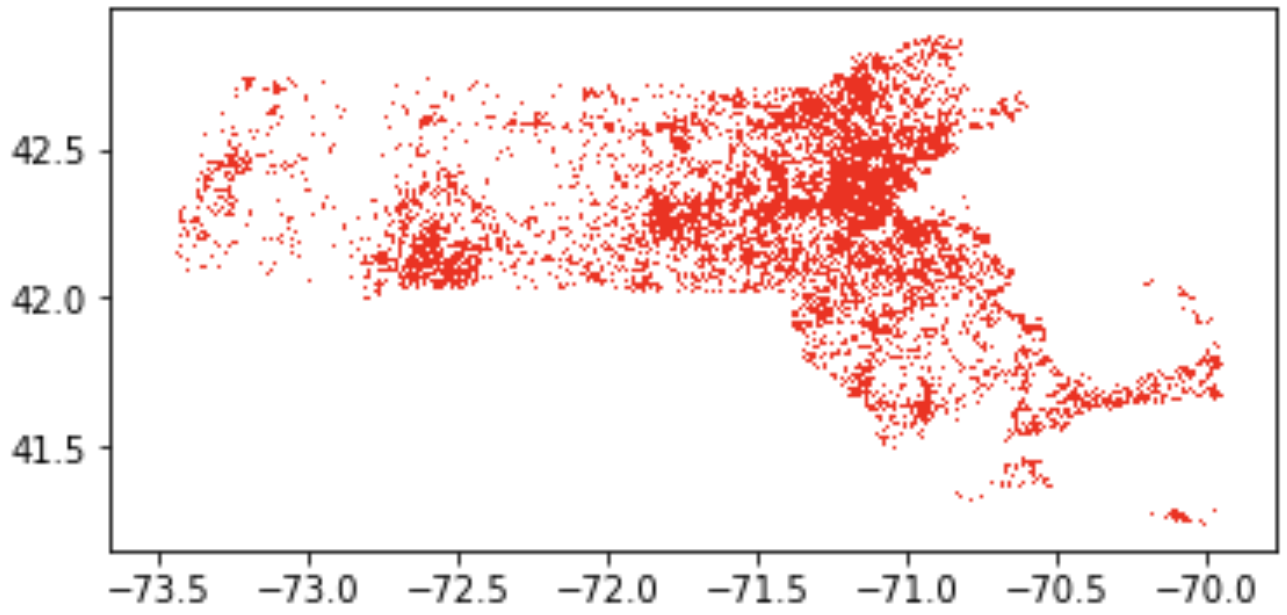
- Schema:

    - quadkey: a key that identifies the tile
    - avg_d_kbps: the average download speed in kilobits per second within the tile
    - avg_u_kbps: the average upload speed in kilobits per second within the tile
    - avg_lat_ms: the average latency of the tests in this tile.
    - tests: The number of tests that contributed to the other values in this tile.
    - devices: The number of unique devices that contributed to the data in this tile.
    - geometry: list of latitude/longitude pairs, that collectively form the polygonal shape of this tile.
    - index_right: joined column index number
    - objectid: identifier for object from joined table
    - muni_id: numeric identifier for the municipality.
    - municipal: name of the municipality
    - shape: list of latitude/longitude pairs, that collectively form the polygonal shape of this municipality.
- Example data point: 0302332123102031,240473,108651,9,5,3,"POLYGON ((-71.1749267578125 42.252917783302, -71.16943359375 42.252917783302, -71.16943359375 42.2488517007209, -71.1749267578125 42.2488517007209, -71.1749267578125 42.252917783302))",216,228,73,Dedham,8E080000660100000800100006A690000B621070001000000A5BAB4CEB2159EC09FB9A411E8F109A0EB0AE896CE019CF7D403C0F702A0BF08FC970488CC09ECD6068CE306ECE90888BF05D0A530ACF90BE4E96B8CA523FCBC4BA4EF1EE8B027A8DA19D0E313D0B024C094AD05F8B0C009FCB97EF4CAEC01F4820AE4F414F0C5

Note now that instead of being labeled with the County, this data point is now labeled with municipality.

Ookla Geographical Density

Since Ookla data is constructed with tiles labeled with geographical information, we were able to produce density maps showing where in the state the measurements were collected. Here is that map:



The vertical axis is latitude, and the horizontal axis is longitude.

From the scatter plot of the location of each data point, we can see that the data around Boston area and Springfield area are denser than average, and there are only a few data points in rural areas.

Labeling MLAB Data

Unlike the Ookla data, the MLAB data was already labeled with municipality, and thus the step of labeling each data point by municipality was unnecessary. The MLAB data was also labeled with "ASNumber," which refers to the number assigned to the Autonomous System which controlled the newtork via which each speed test was conducted. Unfortunately, the data did not come with the name of the organization that operates each Autonomous System, and therefore it became necessary to map each of these numbers to a well defined organization.

To do so, we used a publicly available listing of Autonomous Systems, found here: http://www.bgplookingglass.com/list-of-autonomous-system-numbers

With this information, we added a new column to the MLAB schema, containing the name of the Provider that runs each of the various Autonomous Systems. Unfortunately, this type of information is not obtainable with the Ookla data, and therefore this labeling was particularly important, in order to be able to run analyses on a per-provider level.

MLAB Descriptive Statistics

With the MLAB data labeled with providers, we computed a number of decsriptive statistics over the entirety of the dataset, to get a better understanding of the data contained within. In particular, we produced the following metrics:

Basic Statistics

- average MeanThroughputMbps: 43.7
- median MeanThroughputMbps: 12.6
- mode MeanThroughputMbps: 0 11.8
- Standard Deviation MeanThroughputMbps: 91.66

Fastest and Slowest Providers, with at least 50 measurements

Top 5 fastest providers on average

- TWRS-MA - Towerstream I, Inc.: 215.38 Mbps
- NMSU-AS - Checs-net: 206.89 Mbps
- HGE-NET - Holyoke Gas & Electric Department: 192.12 Mbps
- BIGLEAF - Bigleaf Networks LLC: 165.77 Mbps
- NORTHEASTERN-GW-AS - Northeastern University: 155.48 Mbps

Botton 5 slowest providers on average

- WB-DEN2 - Viasat Communications Inc.: 0.847753 Mbps
- OSRAM-SYLVANIA - OSRAM SYLVANIA INC: 1.562037 Mbps
- HNS-DIRECPC - Hughes Network Systems: 1.835326 Mbps
- SPCS - Sprint Personal Communications Systems: 2.376271 Mbps
- WIVALLEY - WiValley, Inc.: 2.506452 Mbps

Splitting MLAB Data into Sub-Datasets by Municipality

Since the desired granularity for the data is on the municipal level, we produced a series of csv files each limited to the MLAB data that was collected only within the relevant municipality. This would allow for easier analysis within each of the municipalities, since the overall data set is quite large. Thus, should some analysis focus on a particular municipality, or a short list of municipalities, these files would come in handy.

There are 101 municipalities within MAPC purview, and therefore we produced one output csv file for each of these. They each have the exact same schema as the overall MAPC data, but with data limited to the relevant municipality. For example, there is a single for Acton MA, with a total of 6645 data points contained.

Computing the Average Broadband Speed Per Provider Per Municipality

Here, we produced a csv file with the average broadband speed of each provider, in each municipality. For example, here is an excerpt from that file:

...

Abington,*BIGLEAF - Bigleaf Networks LLC,*1.9385489829867468,*44.9945*

Abington,*CELLCO - Cellco Partnership DBA Verizon Wireless,*14.1583484092067,*53.14081818181819*

Abington,*"COMCAST-7922 - Comcast Cable Communications, Inc.",*10.609942074630164,*25.27878453038673*

Abington,*LIGHTOWER Lightower Fiber Networks (LIGHT-141),*46.193328643945755,*22.193*

Abington,*"UUNET - MCI Communications Services, Inc. d/b/a Verizon Business",*175.27244063910456,*14.569966666666653*

...

The first column here is the name of the municipality; in this case, we are focused on Abington. The second column is the name of the Provider, or more specifically, the name of the organization that runs the Autonomous System via which a given test was conducted. The third column is the average MeanThroughputMbps, and the fourth column is the average MinRTT, which stands for Minimum Round Trip Time.

As can be seen here, Lightower had an average MeanThroughputMbps of 46.2 megabits-per-second, while Verizon was considerably faster, at 175.27 megabits-per-second.

This file is one of the primary outcome of this deliverable, and it contains average broadband speeds for all municipalities per provider.

Steps Towards Deliverable 3

TODO

Summary

TODO