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:

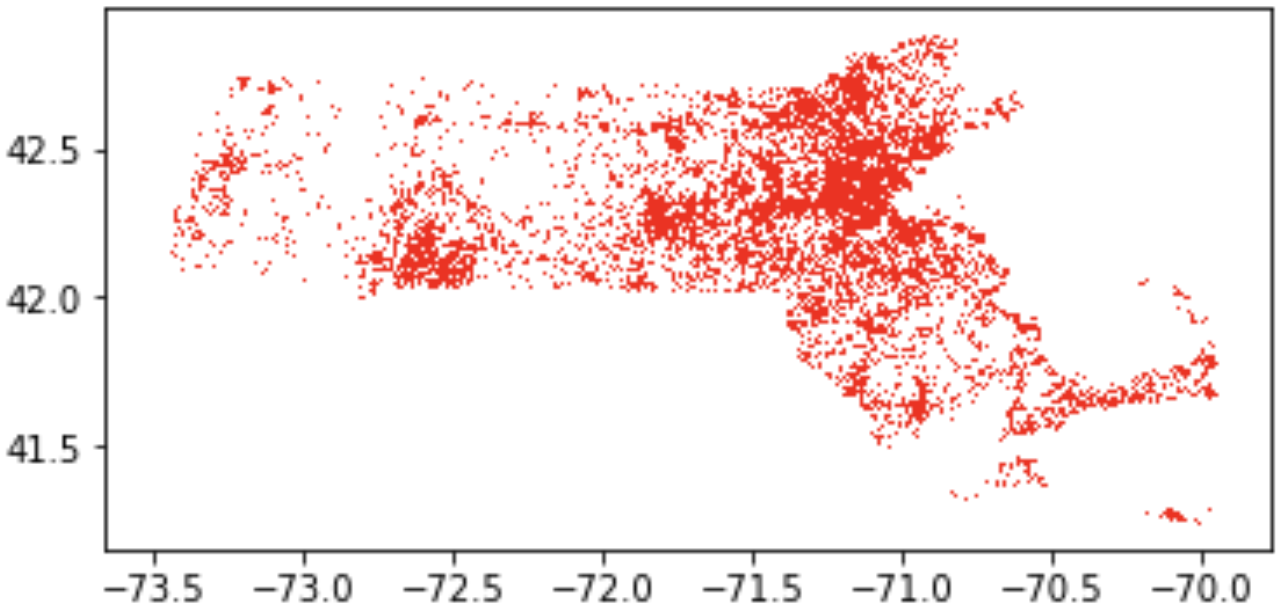
- Schema:

    - 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.

Ookla Basic Statistics

Top 25 Cities with Fastest Average Download Speed

City	Average Download Speed, Kbps
West Bridgewater	205943.28276
Montgomery	205546.3
Bridgewater	199698.68388
Gardner	197069.1943
East Bridgewater	191234.58481
Whitman	187003.88316
Ware	185472.76952

Erving	184774.20513
Everett	184351.24286
Dracut	183783.50973
Palmer	182919.57853
Ashby	181227.69079
Revere	180672.46409
Peabody	179239.06
Haverhill	178558.80971
Lowell	177470.17117
Norton	176841.28748
Chelsea	175945.49351
Fall River	175309.47971
Monson	174890.07237
Milton	174573.90667
Groveland	174141.37981
Westminster	173648.61179
Boston	173572.17082
Middleborough	172813.69739

Cities with the Most Average Number of Devices

City	Average Number of Devices
Somerville	24.791
Cambridge	21.39189
Brookline	18.10905
Boston	16.2856
Malden	15.30633
Everett	15.03333

Chelsea	14.59091
Watertown	13.70169
Medford	11.64208
Arlington	11.12895
Revere	10.8232
Quincy	10.56224
Belmont	10.22945
Newton	9.78877
Lawrence	9.05793
Lowell	7.88031
Provincetown	7.71429
Waltham	7.36449
Melrose	7.34375
Winthrop	7.25373
Lynn	7.14741
Winchester	7.04709
Swampscott	6.73604
Salem	6.61962
Worcester	6.52941

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

Top 5 Fastest Providers on Average, with at least 1,000 Measurements

Provider Name	Average Mbps
HGE-NET - Holyoke Gas & Electric Department	192.128319
UUNET - MCI Communications Services, Inc. d/b/a Verizon Business	134.443643
LIGHTOWER Lightower Fiber Networks (LIGHT-141)	82.131921
ASN-QWEST-US NOVARTIS-DMZ-US	49.514844
ALKERMES - ALKERMES INCORPORATED	23.363405

Botton 5 Slowest Providers on Average, with at least 1,000 Measurements

Provider Name	Average Mbps
SPCS - Sprint Personal Communications Systems	2.376271
CELLCO - Cellco Partnership DBA Verizon Wireless	5.019919
ASN-SHREWS - Shrewsbury Electric and Cable Operations	6.070376
T-MOBILE-AS21928 - T-Mobile USA, Inc.	8.656454
RR-NYSREGION-ASN-01 - Time Warner Cable Internet LLC	9.678461

Counts of Measurements per Provider, with at least 1,000 Measurements

ProviderName	Measurement Count
COMCAST-7922 - Comcast Cable Communications, Inc.	182906
UUNET - MCI Communications Services, Inc. d/b/a Verizon Business	89356
ALKERMES - ALKERMES INCORPORATED	68790
CHARTER-NET-HKY-NC - Charter Communications	21136
RCN-AS - RCN	15222
ASN-QWEST-US NOVARTIS-DMZ-US	12355
CELLCO - Cellco Partnership DBA Verizon Wireless	6878
T-MOBILE-AS21928 - T-Mobile USA, Inc.	6453
RR-NYSREGION-ASN-01 - Time Warner Cable Interne...	6096
SPCS - Sprint Personal Communications Systems	3502
ASN-SHREWS - Shrewsbury Electric and Cable Oper...	2579
LIGHTOWER Lightower Fiber Networks (LIGHT-141)	1369
HGE-NET - Holyoke Gas & Electric Department	1130

Top 25 Cities by Measurement Count

City	Measurement Count
Needham	34697
Boston	23807
Somerville	15488
Bedford	14515
Ashland	13688
Cambridge	11789
Devens	9150
Worcester	7709
Springfield	7146

Acton	6645
Dorchester	6479
Watertown	6094
Brighton	5325
Arlington	5212
Brookline	5135
Concord	5022
Newton Center	4791
Wellesley Hills	4639
Lexington	4532
Lowell	4415
Quincy	4402
Waltham	4274
Milton	4024
Framingham	3870
Andover	3795

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

In our most recent client meeting, which took place on Friday, March 19th, 2021, we established a clear goal for the next deliverable. This section describes what that goal is, and the steps we are planning to take to complete the requirement.

MAPC is particularly interested in a scatter plot which charts the average Mbps per city, against the median household income in that city. This chart should be indicative; is there a relationship between income and broadband speeds, and if so, what does that relationship look like? Are there patterns that emerge within the data, and are there identifiable aspects of the data that can be extracted for policy purposes?

One of the primary outcomes of the next deliverable will be to produce this graph. Notably, we will likely produce the following graphs specifically:

- A scatter plot of MLAB data, where each data point corresponds to a single municipality. As there are 101 municipalities within MAPC's purview, this chart will have 101 data points. Each data point is a tuple of average broadband speed in megabits-per-second, with the median household income in that municipality

- A similar scatter plot of Ookla download speed data
- A simialr scatter plot of Ookla upload speed data

We may possibly produce similar such charts for each provider, although it is not clear if this is currently something that MAPC desires. Further, each of these plots will be for the year 2020, and given time, will produce further plots for years before 2020 as well.

Data Sets

In addition to the Ookla and MLAB data sets, we will also need income data. We will obtain this information from the census, which can be found here: https://datacommon.mapc.org/browser/datasets/194

Procedural Steps

In order to produce this chart, we will first need to obtain the income data as listed in the previous section, and organize it appropriately. We will also likely need to do some cleaning and preprocessing of the data, in order to be able to easily work with it. Finally, once the data is cleaned and organized appropriately, producing these charts will be simple.

Summary

In this deliverable, we progressed with the state of both the Ookla data and the MLAB data. We labeled the Ookla data with municipality information, and we labeled the MLAB data with provider information. For each of the data sets, we ran a series of statistical analyses in order to better understand the data contained. Finally, for the MLAB data, we produced csv files for each municipality, and the average broadband speeds for each provider in each municipality.

Checklist

- Collect and pre-process a secondary batch of data
    - Data for this project was previously collected in deliverable 1. However, we have continued to refine and filter the initial data sets, and have the scripts in place to easily produce similar datasets for years prior to 2020.
- Refine the preliminary analysis of the data performed in PD1
    - As described, we have further processed each of the Ookla and MLAB datasets, labeled them appropriately, and produced sub-datasets on a per municipality level. We have also identified the data set that we will need to continue forwards with this work.
- Answer another key question
    - We have identified the average megabits per second broadband speed per municipality per provider in the MLAB data. We have also identified the municipalities with fastest average download speeds in the Ookla data.
- Refine project scope and list of limitations with data and potential risks of achieving project goal
    - We have clearly described the next desired outcome with MAPC for the next deliverable, and described the steps required to achieve this goal.

- Submit a PR with the above report and modifications to original proposal