Towards A Clear Understanding Of Rural Internet: What Statistical Measures Can Be Used To Assess, Compare And Forecast Internet Speeds For Rural Canadian Communities?

A Consulting Project for Math 6627 (1/3)

Benjamin Smith

26 February 2022

Contents

List of Figures

List of Tables

Introduction

(Quoted from the SSC website)

The Government of Canada has committed to helping 95% of Canadian households and businesses access high-speed internet at minimum speeds of 50 Mbps download and 10 Mbps upload (hereinafter referred to as the "Commitment") by 2026, and 100% by 2030. According to the CRTC, currently 45.6% of rural community households have access to the Commitment based on what's available to them via an Internet Service Provider (e.g. Shaw, Telus, etc.) in their region, rather than what a rural household actually realizes at home in terms of internet speeds.

For this case study, the SSC would like to understand the state of internet connectivity in both rural and underserved Canadian communities using consumer-provided data. The SSC claims that by using data directly from the consumer, it is possible to better understand connectivity in these communities as measured by the consumers in their own homes.

Specifically, the following is desired:

1. A statistical analysis of the current realized and forecasted internet speeds (upload and download) for rural and undeserved communities in terms of progress towards the Commitment;
2. A comparative analysis of rural and underserved communities in terms of progress towards the Commitment; and
3. The identification of statistically reliable methods to assess and compare rural and underserved communities' realized internet access.

The following analysis aims to address the above in a practical and concise manner.

The Data

The data was made available by the Statiscal Society of Canada with Ookla and Statistics Canada. One of the first things to check regarding the data is to see if any missing data is present in the dataset. Figure 1 shows that most of the missing data is related to population center information. Namely data on population center id, type and class (PCUID, PCTYPE, PCLASS). For this analysis, the dataset is used "as is". Reason why imputation was not applied to this dataset is discussed at the end of the report.

Analysis

Current Realized And Forecasted Internet Speeds - A Statistical Analysis

It is possible to describe the relationship between the current realized and forecasted internet speeds by use of a statistical model. For this analysis a mixed model structure is chosen with random effect being the individual tile measured. For choice of fixed effects in the model the use of directed acyclic graphs (DAGs) is employed for parameter selection. Figure 2 is the DAG which was used for model development. The fixed and random effects are thus listed in table 1 below:

Table 1: Fixed and Random Effects used to describe Upload/Download Speed

Fixed Effects	Random Effect
No. of devices	quadkey
Connection Type	
No. of Tests	
Year	
Quarter	
Province	

Fixed Effects	Random Effect
Population Center Class	
Year*Quarter	
Province*Population Center Class	

The model can be represented in the following form:

$$Y = X\beta + Zb$$

Where Y is the response variable (i.e. download/upload speed) X is the design matrix for the fixed effects, Z is the design matrix of the random effects and β and b are the fixed and random effects vectors.

Two separate models are constructed for download and upload speed and run in SAS. Figures 3 and 5 show that the fixed effects are all significant at the 0.001 level with the exception to the upload speed model which lists the p-values of the effects of the number of tests conducted and the population center class as 0.0011 and 0.0039 respectively. Figures 4 and 6 show that the random effects of the individual tile measured is significant at the 0.001 level.

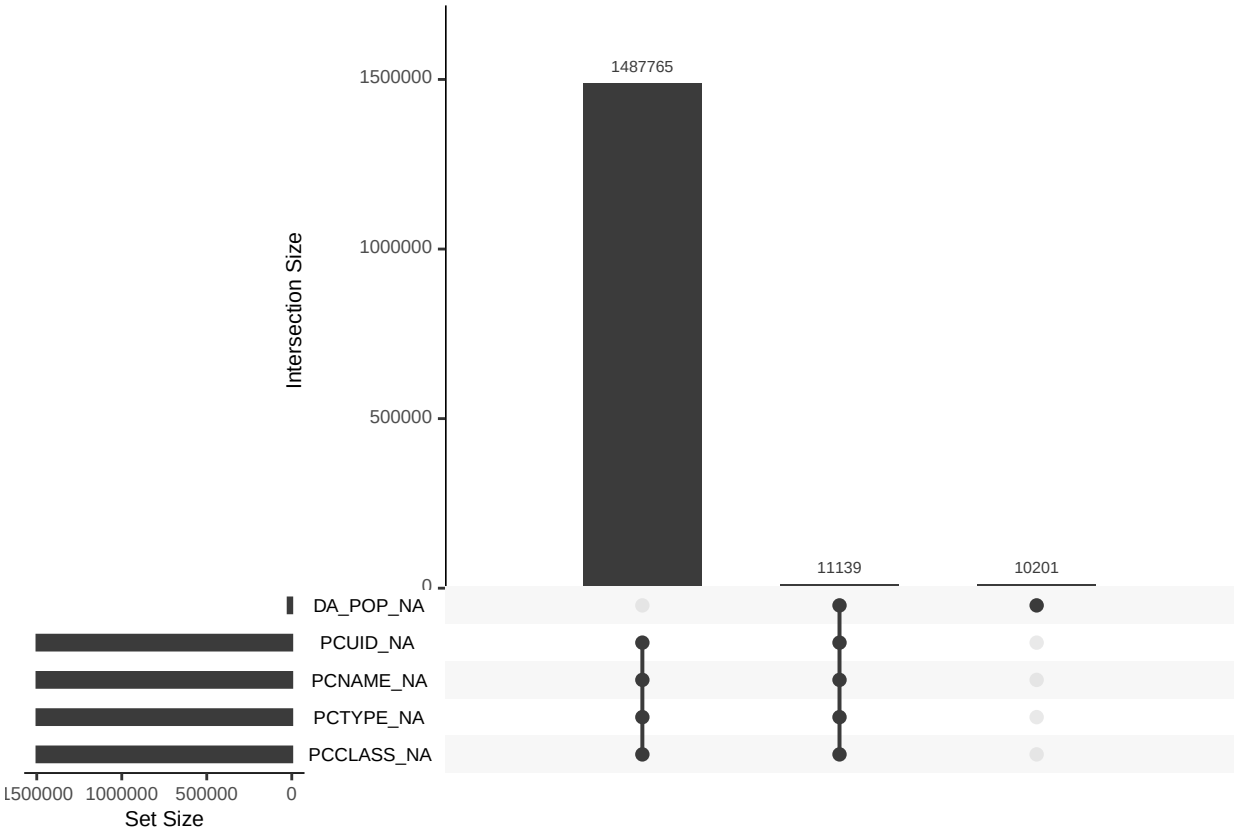

Figure 1: Missing data present in dataset provided

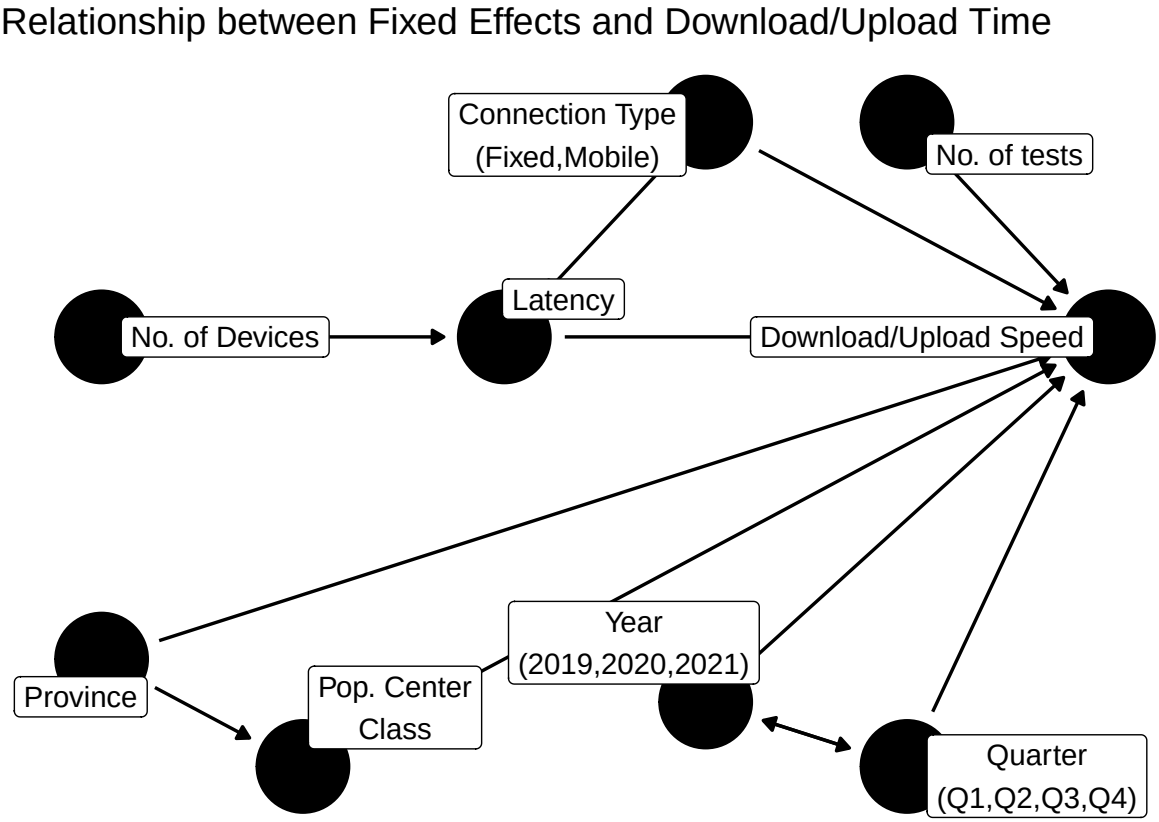

Figure 2: DAG representing the relationship between the fixed effects and upload/download time.

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
devices	1	13E5	1040.84	<.0001
conn_type	1	13E5	28970.4	<.0001
tests	1	13E5	455.98	<.0001
year	2	13E5	51456.5	<.0001
quarter	3	13E5	3231.77	<.0001
quarter*year	6	13E5	237.48	<.0001
PRNAME	12	13E5	24.94	<.0001
PCCLASS	2	13E5	112.37	<.0001
PRNAME*PCCLASS	15	13E5	75.64	<.0001

Figure 3: Type 3 tests of fixed effects for download speed model

Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
quadkey	9.277E8	64330039	14.42	<.0001
Residual	5.8519E9	7396119	791.21	<.0001

Figure 4: Covariance parameter estimates of the random effects in the download speed model

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
devices	1	13E5	993.01	<.0001
conn_type	1	13E5	107617	<.0001
tests	1	13E5	10.70	0.0011
year	2	13E5	23874.9	<.0001
quarter	3	13E5	2492.84	<.0001
quarter*year	6	13E5	100.49	<.0001
PRNAME	9	13E5	228.68	<.0001
PCCLASS	1	13E5	8.34	0.0039
PCCLASS*PRNAME	9	13E5	59.80	<.0001

Figure 5: Type 3 tests of fixed effects for upload speed model

Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
quadkey	4.6014E8	34046673	13.52	<.0001
Residual	1.7061E9	2156433	791.19	<.0001

Figure 6: Covariance parameter estimates of the random effects in the download speed model

Rural and underserved communities in terms of progress towards the Commitment

Figures 7 and 8 show that on average, most provinces are keeping to the Commitment above and beyond the requirements provided for all communities. The provinces which appear to be experiencing challenges with this are the Northwest Territories, Nunavut and Yukon- all of whom only have available data on small population centers[1].

While this does offer a "birds eye view" a more accurate portrayal is to look at the proportion of tiles in population centers which are meeting the agreement and which are not. Figures 9-14 outline such characteristics.

In terms of small population centers, all provinces appear to be making progress in the Commitment with the exception to Nunavut which appears to be struggling. For medium population centers, Manitoba appears to struggle with making any progress for upload time, but has improved in terms of download time.

For locations which were not associated with a population center, see "A note about missing data".

[1]According to Wikipedia there are only small population centers in the Canadian Territories as of 2016. Reference: https://en.wikipedia.org/wiki/List_of_population_centres_in_the_Canadian_Territories

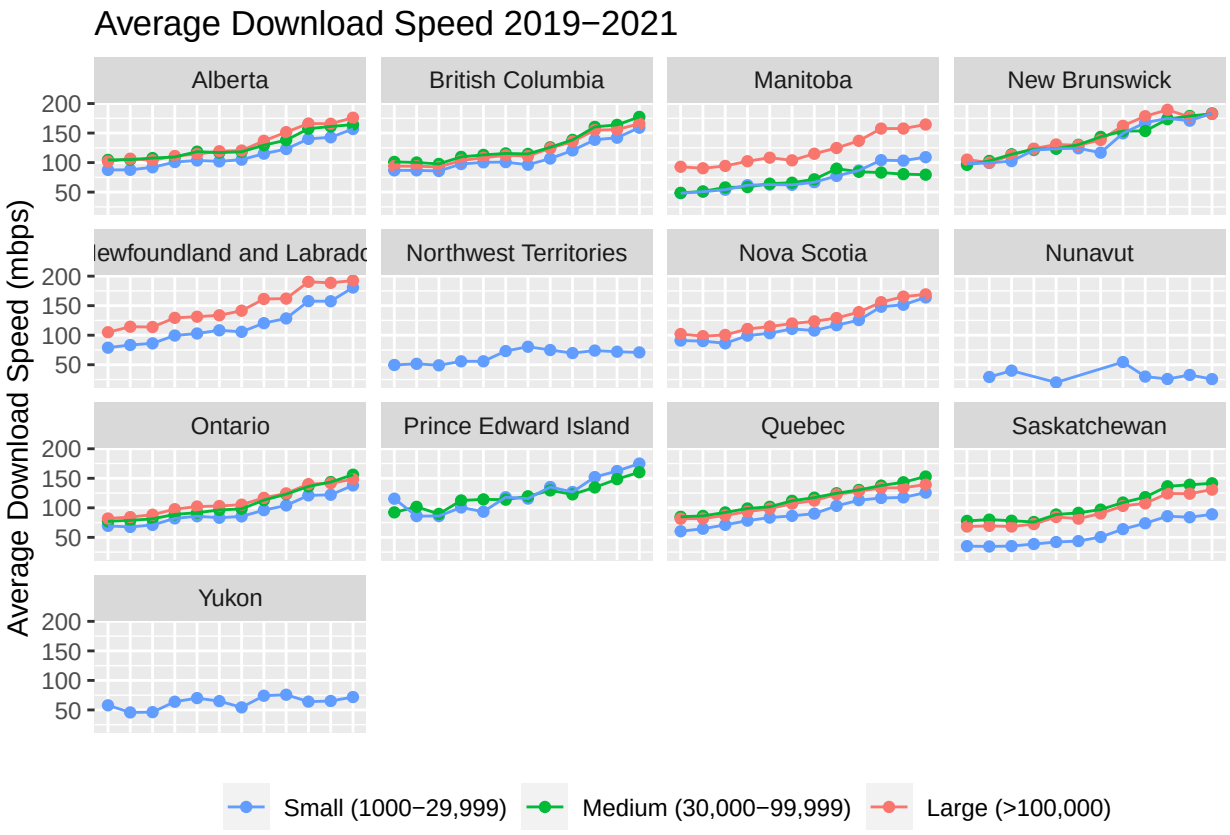

Figure 7: Average download speed accross provinces over time, by population center size

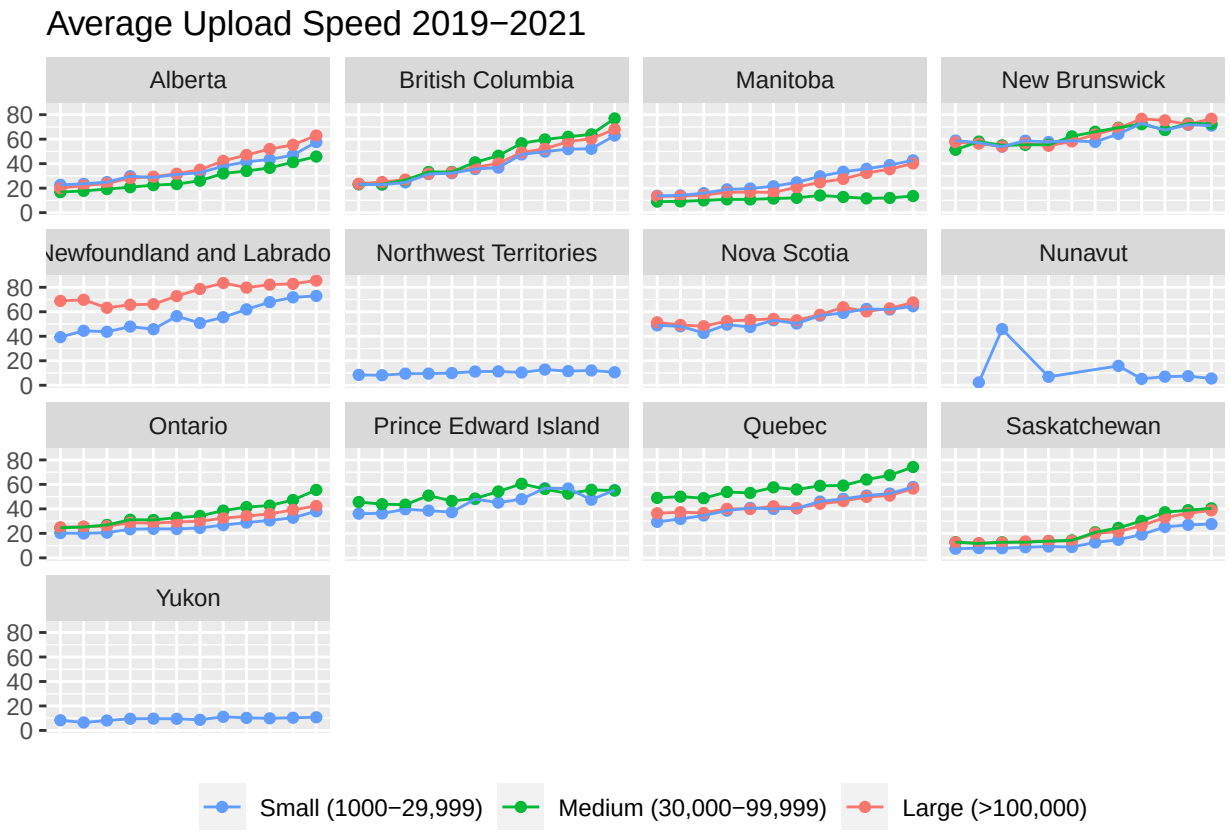

Figure 8: Average upload speed accross provinces over time, by population center size

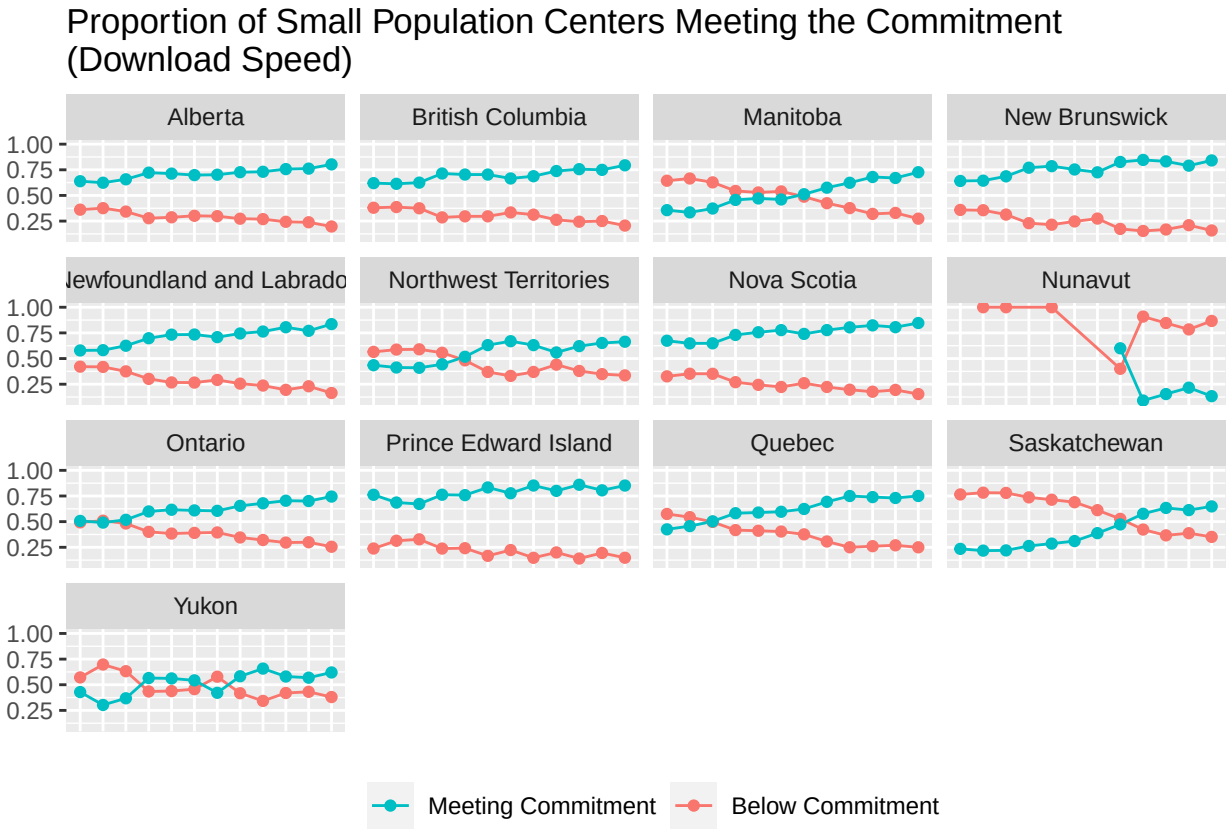

Figure 9: Proportion of small population centers accross Canada meeting the Commitment download speed over time

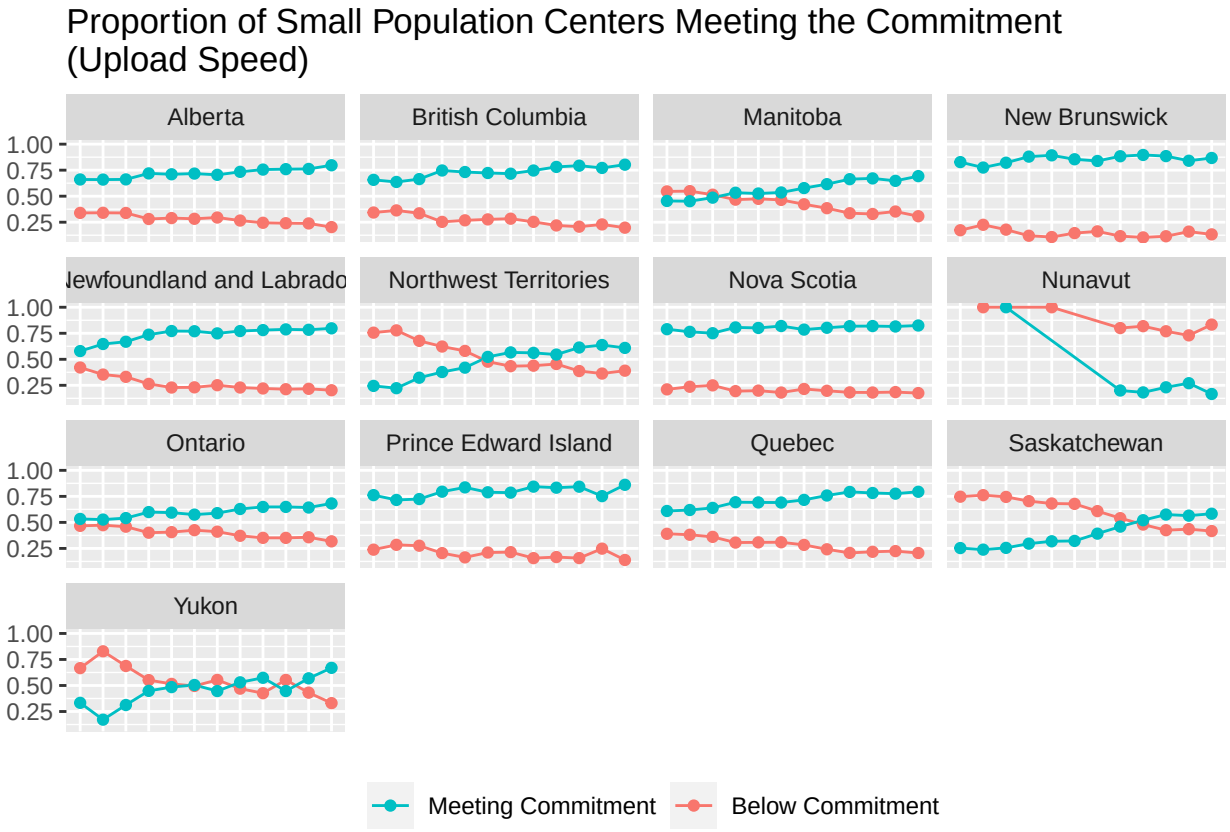

Figure 10: Proportion of small population centers accross Canada meeting the Commitment upload speed over time

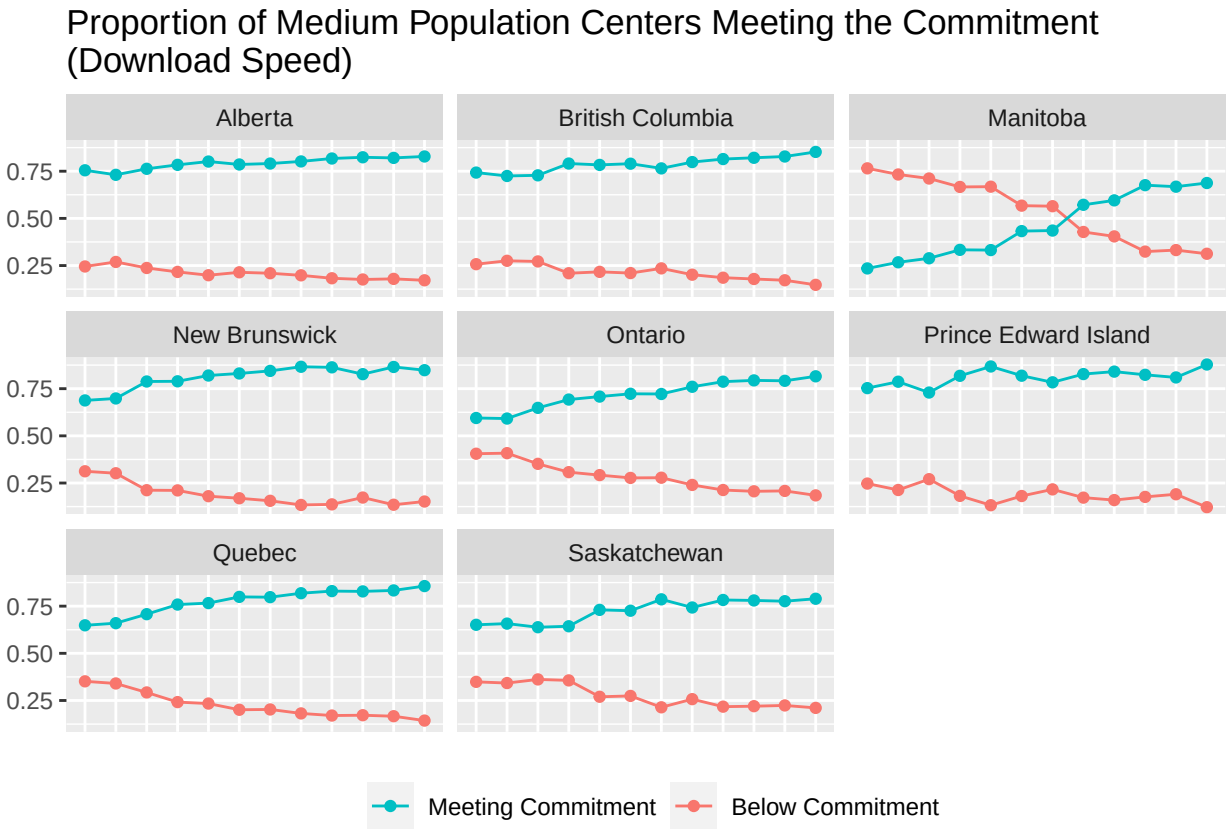

Figure 11: Proportion of medium population centers accross Canada meeting the Commitment download speed over time

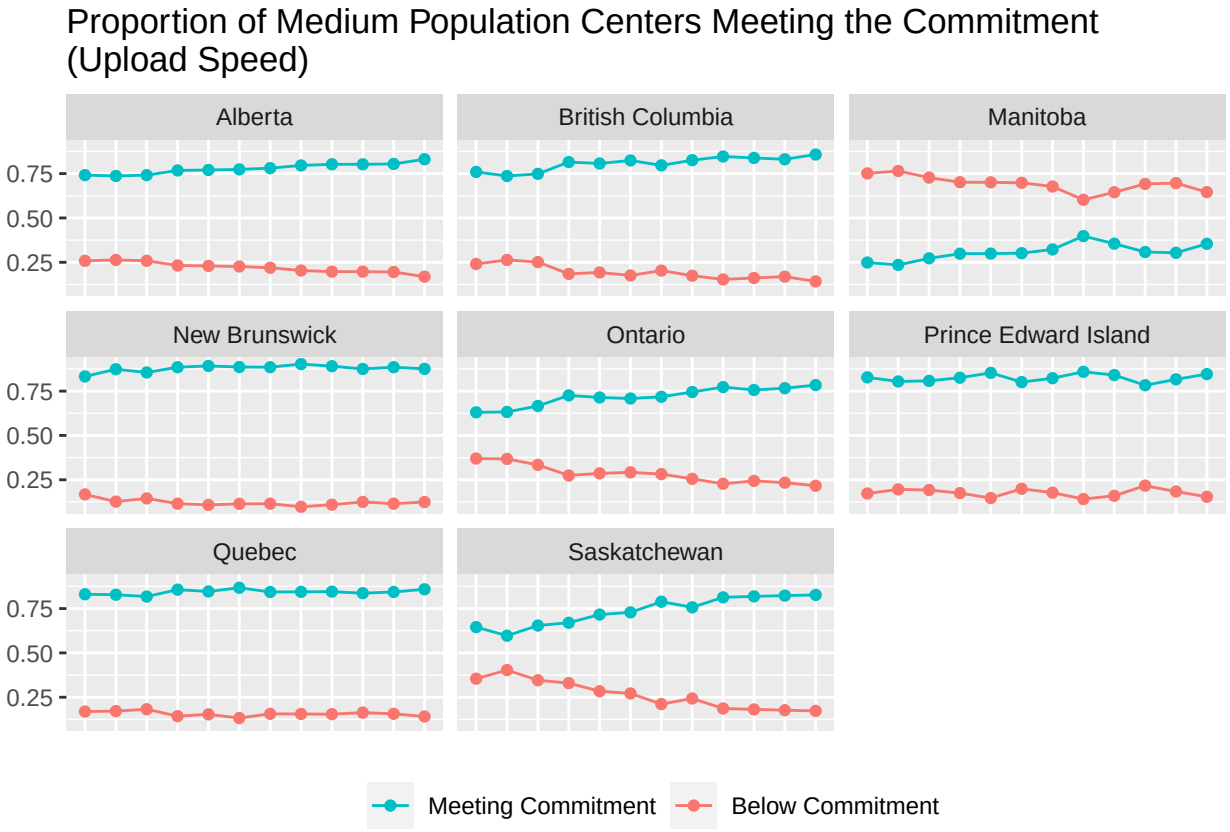

Figure 12: Proportion of medium population centers accross Canada meeting the Commitment upload speed over time

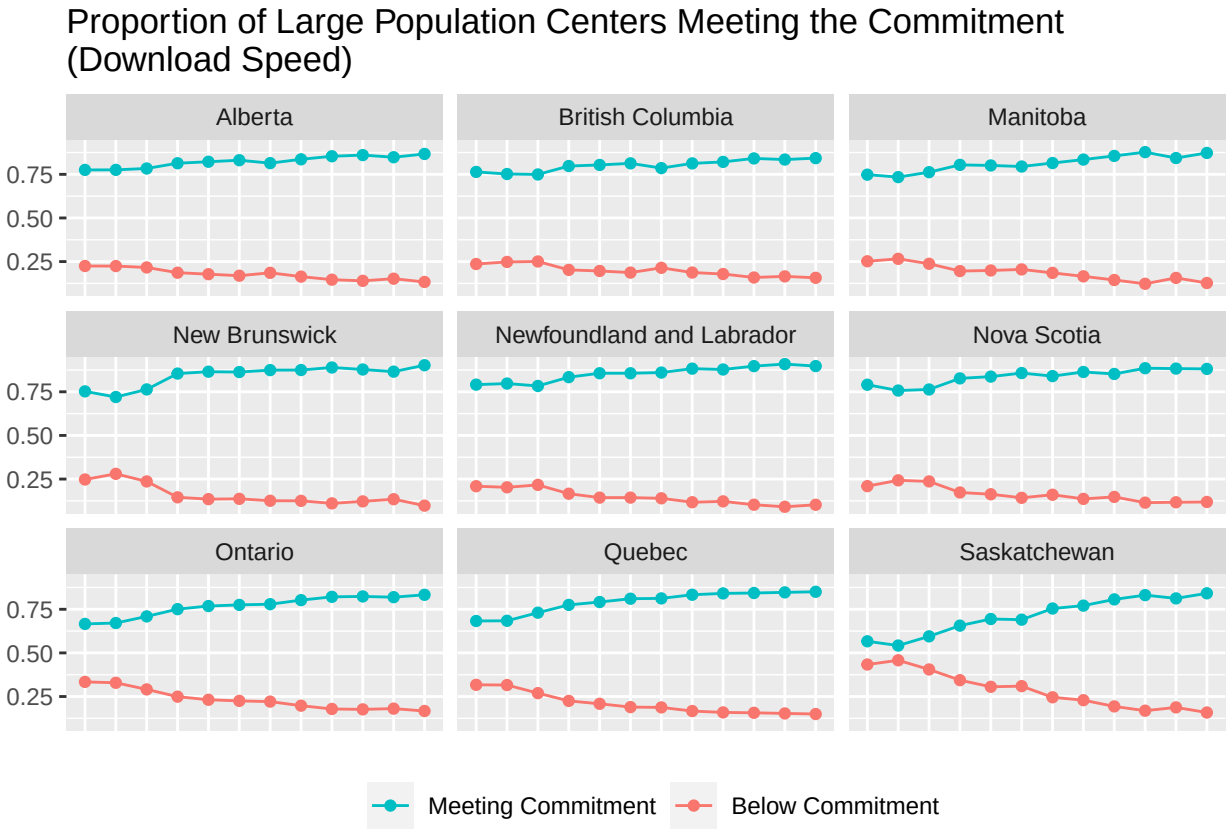

Figure 13: Proportion of large population centers accross Canada meeting the Commitment download speed over time

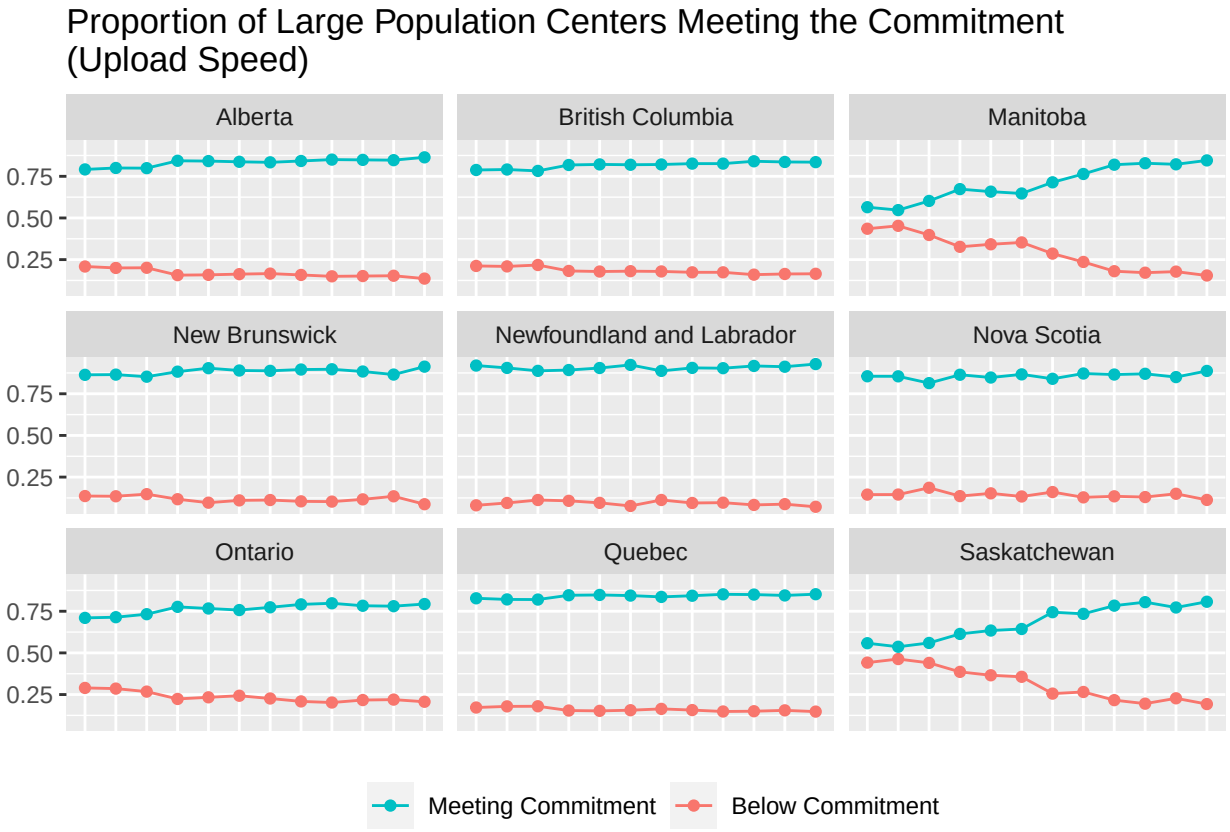

Figure 14: Proportion of medium population centers accross Canada meeting the Commitment upload speed over time

The identification of statistically reliable methods to assess and compare rural and underserved communities's realized internet access.

To identify statistically reliable methods with which to assess and compare rural and undeserved communities, LASSO regression with 10-fold cross-validation is applied to the data set where the variables of interest are download and upload speeds and the predictors are all the other variables with exception to the geometry, year and quarter of the observations as they are not meaningful in terms of statistically reliable methods moving forward. Additionally, superfluous variables such as individual classification id's (i.e. `PRUID`, `CDUID`, `PCUID`) have been removed. Similar to to the first part of the analysis, two separate models are created for download and upload speeds.

Figures 15 and 16 show the test MSE by λ value plotted. From the visuals it can be determined that $\lambda_d = \lambda_u = 35.96566$. Tables 2 and 3 show the respective sparse estimates. From the sparse estimates it is possible to determine which methods can be used to evaluate internet access. Specifically, it is found that location is largely important with looking at a location's statistical area of classification, the province it belongs to, its dissemination area and its census division. In addition to this, information such as its average latency, the number of tests conducted, the devices used and the connection type play prime importance. Surprisingly, information on the population center type does not prove to be significant.

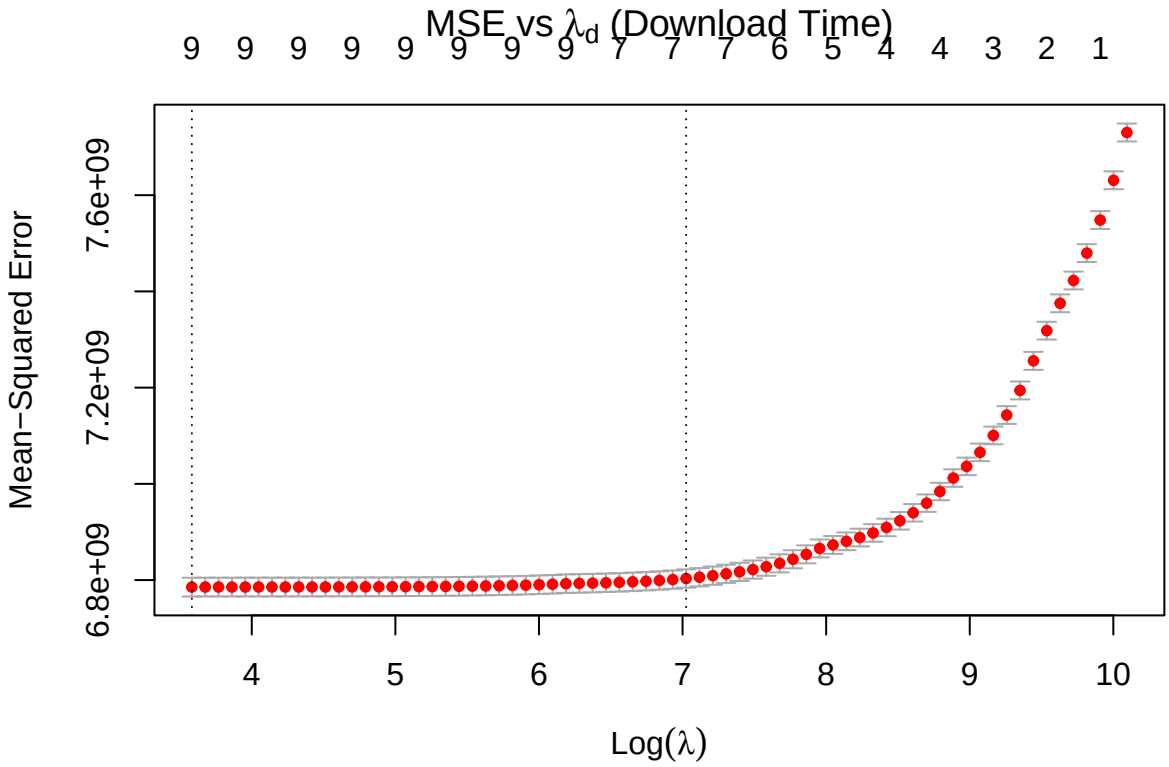

Figure 15: MSE vs λ_u (Download Time)

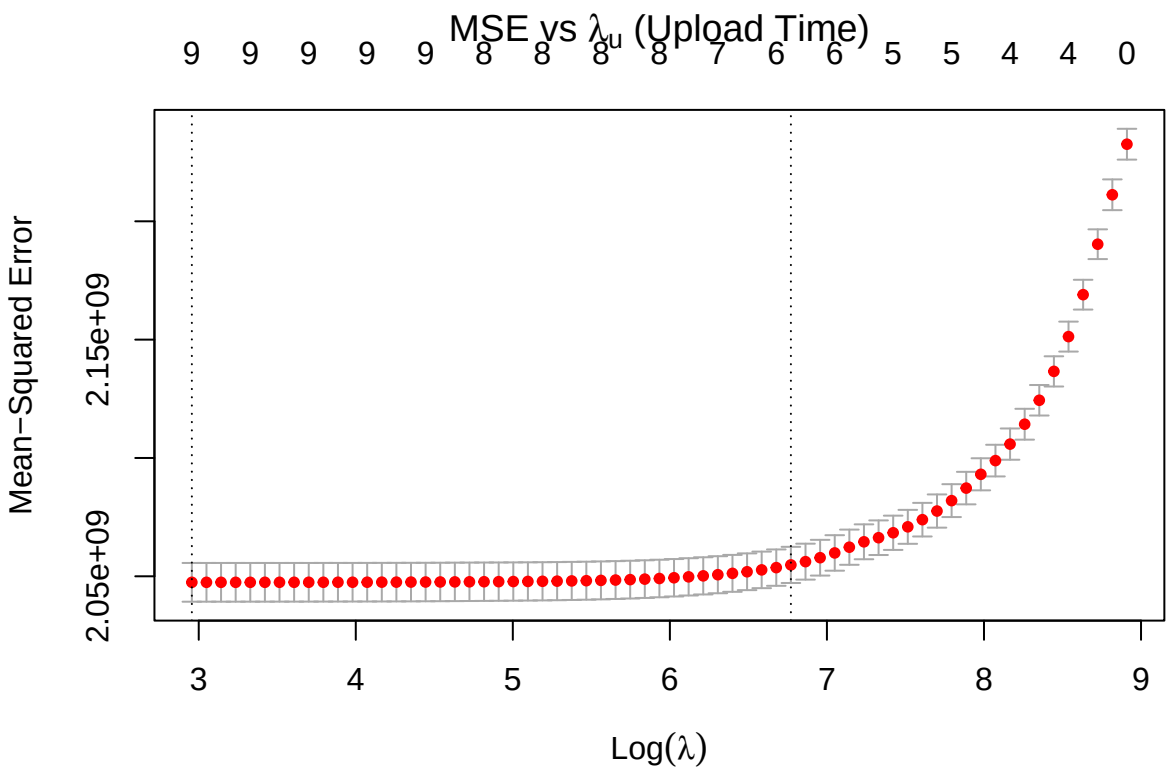

Figure 16: MSE vs λ_u (Upload Time)

Table 2: Sparse Estimates for Download Speed Predicton

	s0
(Intercept)	1.789319e+05
quadkey	-3.172130e-02
avg_lat_ms	-8.826692e+01
tests	8.155297e+00
devices	1.380494e+03
conn_type	-7.968878e+03
PRNAME	-3.125900e+03
CDNAME	3.101325e+01
DAUID	-8.939000e-04
SACTYPE	-1.122555e+04
DA_POP	0.000000e+00
PCNAME	0.000000e+00
PCTYPE	0.000000e+00
PCCLASS	0.000000e+00

Table 3: Sparse Estimates for Upload Speed Predicton

	s0
(Intercept)	7.561686e+04
quadkey	7.089300e-03
avg_lat_ms	-4.237735e+01
tests	-1.243000e+01
devices	5.798148e+02
conn_type	-1.851112e+04
PRNAME	-1.031682e+03
CDNAME	8.818899e+00
DAUID	-3.235000e-04
SACTYPE	-3.711885e+03
DA_POP	0.000000e+00
PCNAME	0.000000e+00
PCTYPE	0.000000e+00
PCCLASS	0.000000e+00

A note about missing data

After looking into the nature of the missing data, there is a reason for the missingness can be explained by the fact that classification in terms of population center does not make sense for it as the locations recorded with missing data on population center correspond to level 3 geographies. This can explains why the LASSO regression sparse extimates list anything related to population center is not significant while in terms of census division it is. Table 4 highlights this and it can be seen that the use of having a population center is not applicable as a statistical measure as opposed to census division.

With this discovery found later in the analysis, the topic of progress towards commitment was revisited. Figures 17 and 18 show that for level three geographies, all provinces are making some progress in terms of download speed with the exception of Nunavut and Yukon whom appear to struggle in meeting the commitment.

In terms of upload speed, Manitoba, Saskatchewan and Northwest Territories are the most noticeable in their progress. While Nunavut and Yukon appear to struggle. As far as other provinces are concerned progress is

either slow or not noticeable.

Table 4: Census divisions with missing data on population center

Census Division	Province
Region 1	Northwest Territories / Territoires du Nord-Ouest
Yukon	Yukon
Region 2	Northwest Territories / Territoires du Nord-Ouest
Region 3	Northwest Territories / Territoires du Nord-Ouest
Region 4	Northwest Territories / Territoires du Nord-Ouest
Region 6	Northwest Territories / Territoires du Nord-Ouest
Northern Rockies	British Columbia / Colombie-Britannique
Kitimat-Stikine	British Columbia / Colombie-Britannique
Peace River	British Columbia / Colombie-Britannique
Stikine	British Columbia / Colombie-Britannique
Division No. 17	Alberta
Region 5	Northwest Territories / Territoires du Nord-Ouest
Division No. 19	Alberta
Division No. 18	Saskatchewan
Division No. 16	Alberta
Division No. 12	Alberta
Division No. 22	Manitoba
Division No. 23	Manitoba
Skeena-Queen Charlotte	British Columbia / Colombie-Britannique
Bulkley-Nechako	British Columbia / Colombie-Britannique
Cariboo	British Columbia / Colombie-Britannique
Central Coast	British Columbia / Colombie-Britannique
Mount Waddington	British Columbia / Colombie-Britannique
Strathcona	British Columbia / Colombie-Britannique
Comox Valley	British Columbia / Colombie-Britannique
Powell River	British Columbia / Colombie-Britannique
Alberni-Clayoquot	British Columbia / Colombie-Britannique
Nanaimo	British Columbia / Colombie-Britannique
Sunshine Coast	British Columbia / Colombie-Britannique
Cowichan Valley	British Columbia / Colombie-Britannique
Fraser-Fort George	British Columbia / Colombie-Britannique
Division No. 18	Alberta
Division No. 13	Alberta
Division No. 14	Alberta
Division No. 15	Alberta
Division No. 11	Alberta
Division No. 9	Alberta
Division No. 10	Alberta
Division No. 8	Alberta
Division No. 7	Alberta
Squamish-Lillooet	British Columbia / Colombie-Britannique
Thompson-Nicola	British Columbia / Colombie-Britannique
Columbia-Shuswap	British Columbia / Colombie-Britannique
Fraser Valley	British Columbia / Colombie-Britannique
Greater Vancouver	British Columbia / Colombie-Britannique
Capital	British Columbia / Colombie-Britannique
Central Okanagan	British Columbia / Colombie-Britannique
North Okanagan	British Columbia / Colombie-Britannique
Okanagan-Similkameen	British Columbia / Colombie-Britannique

Census Division	Province
Kootenay Boundary	British Columbia / Colombie-Britannique
Central Kootenay	British Columbia / Colombie-Britannique
East Kootenay	British Columbia / Colombie-Britannique
Division No. 6	Alberta
Division No. 5	Alberta
Division No. 2	Alberta
Division No. 3	Alberta
Division No. 17	Saskatchewan
Division No. 13	Saskatchewan
Division No. 16	Saskatchewan
Division No. 12	Saskatchewan
Division No. 21	Manitoba
Division No. 15	Saskatchewan
Division No. 14	Saskatchewan
Division No. 4	Alberta
Division No. 8	Saskatchewan
Division No. 11	Saskatchewan
Division No. 7	Saskatchewan
Division No. 1	Alberta
Division No. 4	Saskatchewan
Division No. 3	Saskatchewan
Division No. 10	Saskatchewan
Division No. 6	Saskatchewan
Division No. 9	Saskatchewan
Division No. 20	Manitoba
Division No. 5	Saskatchewan
Division No. 16	Manitoba
Division No. 2	Saskatchewan
Division No. 15	Manitoba
Division No. 1	Saskatchewan
Division No. 6	Manitoba
Division No. 5	Manitoba
Division No. 19	Manitoba
Kenora	Ontario
Division No. 17	Manitoba
Division No. 18	Manitoba
Division No. 7	Manitoba
Division No. 8	Manitoba
Division No. 9	Manitoba
Division No. 4	Manitoba
Division No. 14	Manitoba
Division No. 13	Manitoba
Division No. 10	Manitoba
Division No. 11	Manitoba
Division No. 1	Manitoba
Division No. 12	Manitoba
Division No. 3	Manitoba
Division No. 2	Manitoba
Rainy River	Ontario
Thunder Bay	Ontario
Division No. 11	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Division No. 10	Newfoundland and Labrador / Terre-Neuve-et-Labrador

Census Division	Province
Nord-du-Québec	Quebec / Québec
Algoma	Ontario
Cochrane	Ontario
Abitibi-Ouest	Quebec / Québec
Sept-Rivières–Caniapiscau	Quebec / Québec
Le Saguenay-et-son-Fjord	Quebec / Québec
Maria-Chapdelaine	Quebec / Québec
Lac-Saint-Jean-Est	Quebec / Québec
Manicouagan	Quebec / Québec
Sudbury	Ontario
Timiskaming	Ontario
Rouyn-Noranda	Quebec / Québec
Témiscamingue	Quebec / Québec
Nipissing	Ontario
Manitoulin	Ontario
Bruce	Ontario
Greater Sudbury / Grand Sudbury	Ontario
Parry Sound	Ontario
Muskoka	Ontario
Haliburton	Ontario
Huron	Ontario
Lambton	Ontario
Middlesex	Ontario
Grey	Ontario
Simcoe	Ontario
Dufferin	Ontario
Kawartha Lakes	Ontario
York	Ontario
Durham	Ontario
Wellington	Ontario
Perth	Ontario
Waterloo	Ontario
Oxford	Ontario
Hamilton	Ontario
Brant	Ontario
Peel	Ontario
Halton	Ontario
Toronto	Ontario
Haldimand-Norfolk	Ontario
Niagara	Ontario
Essex	Ontario
Chatham-Kent	Ontario
Elgin	Ontario
Abitibi	Quebec / Québec
La Vallée-de-l'Or	Quebec / Québec
La Vallée-de-la-Gatineau	Quebec / Québec
La Tuque	Quebec / Québec
Le Domaine-du-Roy	Quebec / Québec
Antoine-Labelle	Quebec / Québec
Mékinac	Quebec / Québec
Renfrew	Ontario
Pontiac	Quebec / Québec

Census Division	Province
Hastings	Ontario
Les Collines-de-l'Outaouais	Quebec / Québec
Lennox and Addington	Ontario
Ottawa	Ontario
Lanark	Ontario
Les Laurentides	Quebec / Québec
Papineau	Quebec / Québec
Matawinie	Quebec / Québec
Maskinongé	Quebec / Québec
Les Pays-d'en-Haut	Quebec / Québec
D'Autray	Quebec / Québec
Joliette	Quebec / Québec
Argenteuil	Quebec / Québec
Prescott and Russell	Ontario
Gatineau	Quebec / Québec
Leeds and Grenville	Ontario
Stormont, Dundas and Glengarry	Ontario
La Rivière-du-Nord	Quebec / Québec
Montcalm	Quebec / Québec
Mirabel	Quebec / Québec
Thérèse-De Blainville	Quebec / Québec
Deux-Montagnes	Quebec / Québec
L'Assomption	Quebec / Québec
Pierre-De Saurel	Quebec / Québec
Marguerite-D'Youville	Quebec / Québec
Les Moulins	Quebec / Québec
Laval	Quebec / Québec
Montréal	Quebec / Québec
La Vallée-du-Richelieu	Quebec / Québec
Longueuil	Quebec / Québec
Les Maskoutains	Quebec / Québec
Vaudreuil-Soulanges	Quebec / Québec
Le Haut-Saint-Laurent	Quebec / Québec
Beauharnois-Salaberry	Quebec / Québec
Roussillon	Quebec / Québec
Rouville	Quebec / Québec
Le Haut-Richelieu	Quebec / Québec
Les Jardins-de-Napierville	Quebec / Québec
Charlevoix	Quebec / Québec
La Jacques-Cartier	Quebec / Québec
La Côte-de-Beaupré	Quebec / Québec
Charlevoix-Est	Quebec / Québec
L'Islet	Quebec / Québec
Montmagny	Quebec / Québec
La Haute-Côte-Nord	Quebec / Québec
Les Basques	Quebec / Québec
Rivière-du-Loup	Quebec / Québec
La Mitis	Quebec / Québec
Rimouski-Neigette	Quebec / Québec
Matane	Quebec / Québec
La Matapédia	Quebec / Québec
Kamouraska	Quebec / Québec

Census Division	Province
Témiscouata	Quebec / Québec
Madawaska	New Brunswick / Nouveau-Brunswick
Victoria	New Brunswick / Nouveau-Brunswick
Shawinigan	Quebec / Québec
Francheville	Quebec / Québec
Portneuf	Quebec / Québec
Lotbinière	Quebec / Québec
Bécancour	Quebec / Québec
Nicolet-Yamaska	Quebec / Québec
L'Érable	Quebec / Québec
Arthabaska	Quebec / Québec
Québec	Quebec / Québec
L'Île-d'Orléans	Quebec / Québec
Lévis	Quebec / Québec
Bellechasse	Quebec / Québec
La Nouvelle-Beauce	Quebec / Québec
Robert-Cliche	Quebec / Québec
Les Appalaches	Quebec / Québec
Les Etchemins	Quebec / Québec
Beauce-Sartigan	Quebec / Québec
Drummond	Quebec / Québec
Acton	Quebec / Québec
Les Sources	Quebec / Québec
Le Val-Saint-François	Quebec / Québec
La Haute-Yamaska	Quebec / Québec
Brome-Missisquoi	Quebec / Québec
Memphrémagog	Quebec / Québec
Sherbrooke	Quebec / Québec
Le Haut-Saint-François	Quebec / Québec
Coaticook	Quebec / Québec
Le Granit	Quebec / Québec
Carleton	New Brunswick / Nouveau-Brunswick
York	New Brunswick / Nouveau-Brunswick
Peterborough	Ontario
Northumberland	Ontario
Prince Edward	Ontario
Frontenac	Ontario
Minganie–Le Golfe-du-Saint-Laurent	Quebec / Québec
La Haute-Gaspésie	Quebec / Québec
La Côte-de-Gaspé	Quebec / Québec
Division No. 9	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Division No. 8	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Division No. 5	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Division No. 6	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Division No. 7	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Avignon	Quebec / Québec
Restigouche	New Brunswick / Nouveau-Brunswick
Bonaventure	Quebec / Québec
Le Rocher-Percé	Quebec / Québec
Gloucester	New Brunswick / Nouveau-Brunswick
Northumberland	New Brunswick / Nouveau-Brunswick
Les Îles-de-la-Madeleine	Quebec / Québec

Census Division	Province
Kent	New Brunswick / Nouveau-Brunswick
Queens	New Brunswick / Nouveau-Brunswick
Sunbury	New Brunswick / Nouveau-Brunswick
Westmorland	New Brunswick / Nouveau-Brunswick
Charlotte	New Brunswick / Nouveau-Brunswick
Kings	New Brunswick / Nouveau-Brunswick
Saint John	New Brunswick / Nouveau-Brunswick
Albert	New Brunswick / Nouveau-Brunswick
Cumberland	Nova Scotia / Nouvelle-Écosse
Kings	Nova Scotia / Nouvelle-Écosse
Prince	Prince Edward Island / Île-du-Prince-Édouard
Queens	Prince Edward Island / Île-du-Prince-Édouard
Kings	Prince Edward Island / Île-du-Prince-Édouard
Colchester	Nova Scotia / Nouvelle-Écosse
Hants	Nova Scotia / Nouvelle-Écosse
Pictou	Nova Scotia / Nouvelle-Écosse
Antigonish	Nova Scotia / Nouvelle-Écosse
Halifax	Nova Scotia / Nouvelle-Écosse
Guysborough	Nova Scotia / Nouvelle-Écosse
Division No. 4	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Division No. 3	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Victoria	Nova Scotia / Nouvelle-Écosse
Inverness	Nova Scotia / Nouvelle-Écosse
Cape Breton	Nova Scotia / Nouvelle-Écosse
Richmond	Nova Scotia / Nouvelle-Écosse
Digby	Nova Scotia / Nouvelle-Écosse
Annapolis	Nova Scotia / Nouvelle-Écosse
Lunenburg	Nova Scotia / Nouvelle-Écosse
Queens	Nova Scotia / Nouvelle-Écosse
Yarmouth	Nova Scotia / Nouvelle-Écosse
Shelburne	Nova Scotia / Nouvelle-Écosse
Division No. 2	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Division No. 1	Newfoundland and Labrador / Terre-Neuve-et-Labrador
Keewatin	Nunavut
Baffin	Nunavut
Kitikmeot	Nunavut

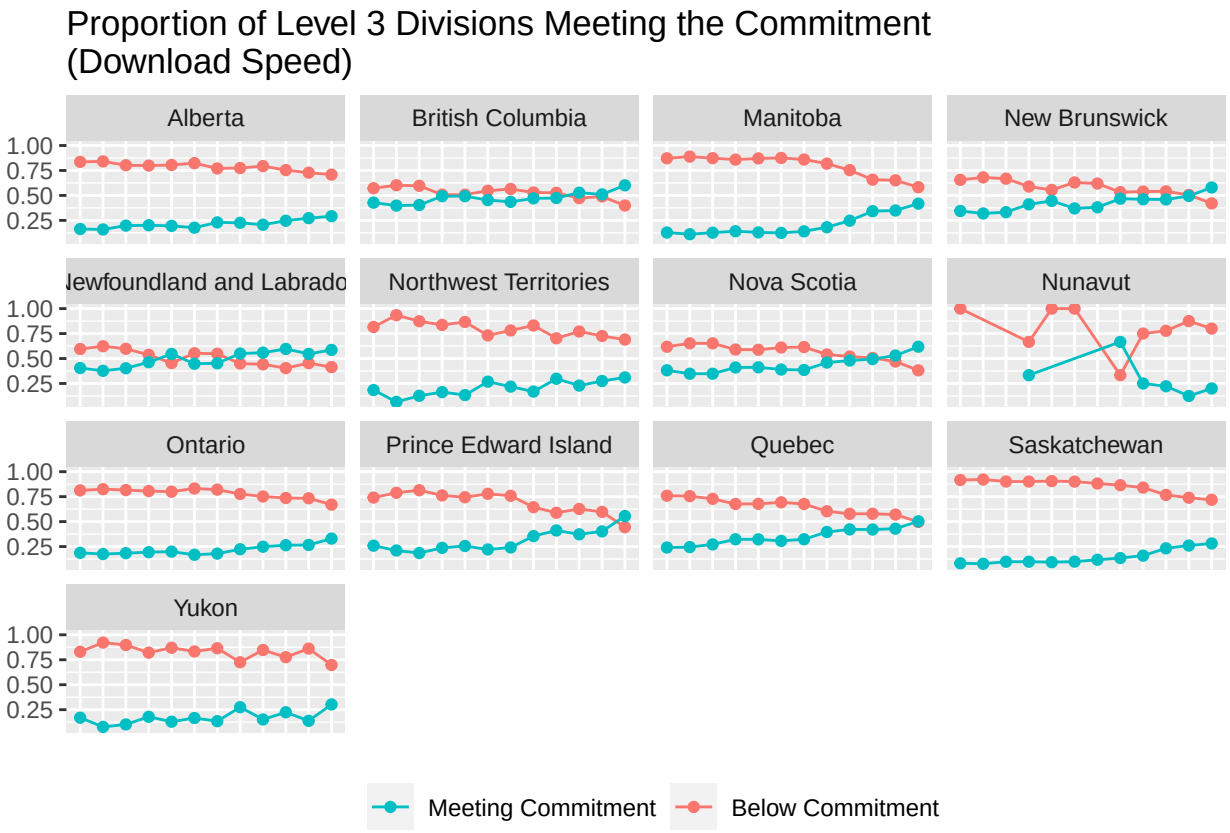

Figure 17: Proportion of level 3 divisions accross Canada meeting the commitment download speed over time

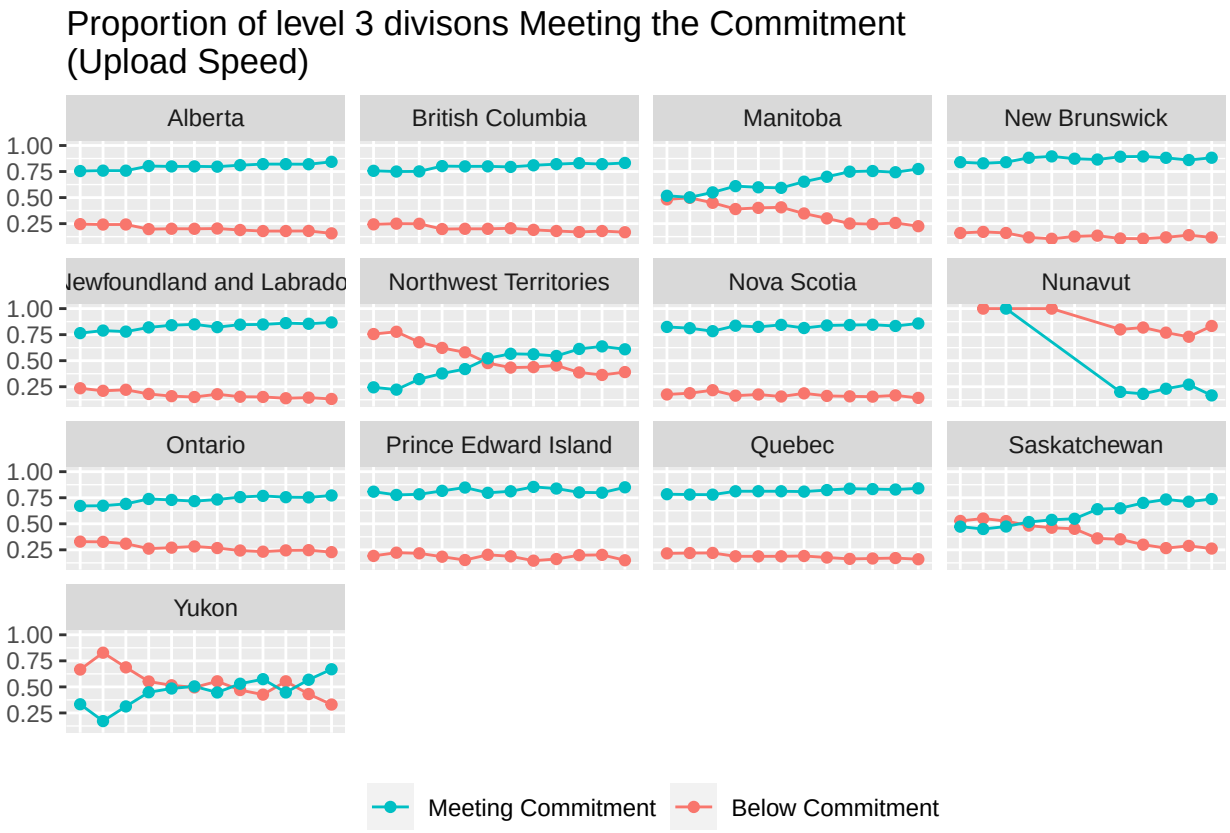

Figure 18: Proportion of level 3 divisions accross Canada meeting the commitment upload speed over time

Conclusion

Incorporating population center type was useful for descriptive purposes as shown in the data. However, in terms of it being a statistical measure, it is not adequate as shown in the sparse estimates. Missing data relating to population center is accredited to the locations measured with that field missing not being related to a particular population center. As such, when looking at location, attention should be paid to census division, statistical area of classification, dissemination area or the individual tile level when measuring internet speeds and their progress to the commitment.

In terms of measurement of other variables, measurements on latency, connection type, number of devices used and tests preformed are good benchmarks for accessing quality of upload and download speeds. This was possible to determine through LASSO regression and the sparse estimates produced.

References

1. Statistical Society of Canada, Towards A Clear Understanding Of Rural Internet: What Statistical Measures Can Be Used To Assess, Compare And Forecast Internet Speeds For Rural Canadian Communities?, https://ssc.ca/en/case-study/towards-a-clear-understanding-rural-internet-what-statistical-measures-can-be-used-assess
2. Ookla, https://www.ookla.com
3. Wikipedia,"List of population centres in the Canadian Territories", https://en.wikipedia.org/wiki/List_of_population_centres_in_the_Canadian_Territories

Code Appendix

SAS Code

```
/*Update File Path Accordingly*/
FILENAME REFFILE '.../ookla-canada-speed-tiles.csv';
PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=DT;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=DT;
RUN;

/*Download Time*/
PROC MIXED DATA=DT METHOD=REML COVTEST;
CLASS PRNAME PCCLASS quarter year conn_type quadkey;
MODEL avg_d_kbps= devices conn_type tests year quarter quarter*year PRNAME PCCLASS PRNAME*PCCLASS;
RANDOM quadkey/s;
RUN;


/*Upload Time*/
PROC MIXED DATA=DT METHOD=REML COVTEST;
CLASS PRNAME quarter year conn_type quadkey;
MODEL avg_u_kbps= devices conn_type tests year quarter quarter*year PRNAME PCCLASS PRNAME*PCCLASS;
RANDOM quadkey/s;
RUN;
```

R Code

```
library(mice)
library(tidyverse)
library(ggthemes)
library(ggspatial)
library(plotly)
library(rnaturalearth)
library(sf)
library(scales)
library(reshape2)
library(formatR)

# dt<-
# readr::read_csv('./ConsultingData/ookla-canada-speed-tiles.csv')
# Accomidating for mtor
setwd("/home/ben2908")
dt <- readr::read_csv("./ookla-canada-speed-tiles.csv")

# Visualizing missing data
naniar::gg_miss_upset(dt)

# Constructing DAG

library(ggdag)

dagify(download_time ~ latency, download_time ~ conn_type, download_time ~
    tests, download_time ~ year, download_time ~ quarter, quarter ~
    year, latency ~ conn_type, latency ~ no_of_devices, year ~
    quarter, download_time ~ pcclass, download_time ~ prname,
    pcclass ~ prname, labels = c(download_time = "Download/Upload Speed",
        latency = "Latency", conn_type = "Connection Type\n(Fixed,Mobile)",
        no_of_devices = "No. of Devices", tests = "No. of tests",
        year = "Year\n(2019,2020,2021)", quarter = "Quarter\n(Q1,Q2,Q3,Q4)",
        prname = "Province", pcclass = "Pop. Center\nClass")) %>%
    tidy_dagitty() %>%
    # prname (8,2) pcclass(8.5,1.5)
mutate(xend = c(10.5, 9, 10.5, 9, 10.5, 10.5, 8.5, 10.5, 9.5,
    10.5, 10.5, 10, NA), yend = c(0, 0, 0, 0, 0, 0, -2, 0, -1.7,
    0, 0, -2, NA), x = ifelse(name == "download_time", 10.5,
    ifelse(name == "latency", 9, ifelse(name == "conn_type",
        9.5, ifelse(name == "tests", 10, ifelse(name == "year",
            9.5, ifelse(name == "quarter", 10, ifelse(name ==
                "no_of_devices", 8, ifelse(name == "prname",
                8, 8.5)))))))), y = ifelse(name == "download_time",
    0, ifelse(name == "latency", 0, ifelse(name == "conn_type",
        1, ifelse(name == "tests", 1, ifelse(name == "year",
            -1.7, ifelse(name == "quarter", -2, ifelse(name ==
                "no_of_devices", 0, ifelse(name == "prname",
                -1.5, -2)))))))), effectType = ifelse(name %in%
    c("download_time", "latency", "conn_type", "no_of_devices",
        "tests", "year", "quarter"), "Fixed", "Random")) %>%
    ggdag(text = FALSE, use_labels = "label") + ggtitle("Relationship between Fixed
    ↪  Effects and Download/Upload Time") +
```

```
    theme_dag()

# Table 1
tibble(`Fixed Effects` = c("No. of devices", "Connection Type",
    "No. of Tests", "Year", "Quarter", "Province", "Population Center Class",
    "Year*Quarter", "Province*Population Center Class"), `Random Effect` = c("quadkey",
    rep("", 8))) %>%
    knitr::kable(caption = "Fixed and Random Effects used to describe Upload/Download
↪  Speed")


# Visuals

# Need to filter based on PCCLASS
dt %>%
    filter(!is.na(PCCLASS)) %>%
    group_by(PCCLASS, PRNAME, year, quarter) %>%
    summarize(average_d_time = mean(avg_d_kbps/1000)) %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER")))) %>%
    ggplot(mapping = aes(x = year_quarter, y = average_d_time,
        color = popcenter_class, group = popcenter_class)) +
    geom_point() + geom_line() + facet_wrap(~Province) + ggtitle("Average Download Speed
    ↪  2019-2021") +
    labs(y = "Average Download Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title.x = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")

# Need to filter based on PCCLASS
dt %>%
    filter(!is.na(PCCLASS)) %>%
    group_by(PCCLASS, PRNAME, year, quarter) %>%
    summarize(average_u_time = mean(avg_u_kbps/1000)) %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER")))) %>%
    ggplot(mapping = aes(x = year_quarter, y = average_u_time,
        color = popcenter_class, group = popcenter_class)) +
    geom_point() + geom_line() + facet_wrap(~Province) + ggtitle("Average Upload Speed
    ↪  2019-2021") +
    labs(y = "Average Upload Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")
```

```
# Small Provinces

smallProvs_Download <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER"))),
        d_status = ifelse((avg_d_kbps/1000) < 50, "Below Commitment",
            "Meeting Commitment"), u_status = ifelse((avg_u_kbps/1000) <
            10, "Below Commitment", "Meeting Commitment")) %>%
    filter(!is.na(PCCLASS), popcenter_class == "Small (1000-29,999)") %>%
    group_by(Province, year_quarter) %>%
    count(d_status)


smallProvs_Download <- do.call(rbind, by(smallProvs_Download,
    smallProvs_Download["Province"], tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

smallProvs_Download %>%
    ggplot(mapping = aes(x = year_quarter, y = prop, color = d_status,
        group = d_status)) + #geom_bar(stat='identity')+ group
        group = d_status)) + #geom_bar(stat='identity')+ =
        group = d_status)) + #geom_bar(stat='identity')+ d_status))
        group = d_status)) + #geom_bar(stat='identity')+ +
        group = d_status)) + #geom_bar(stat='identity')+ #geom_bar(stat='identity')+
geom_line() + geom_point() + facet_wrap(~Province) + ggtitle("Proportion of Small
↪  Population Centers Meeting the Commitment\n(Download Speed)") +
    labs(y = "Average Download Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")


smallProvs_Upload <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER"))),
        d_status = ifelse((avg_d_kbps/1000) < 50, "Below Commitment",
            "Meeting Commitment"), u_status = ifelse((avg_u_kbps/1000) <
            10, "Below Commitment", "Meeting Commitment")) %>%
    filter(!is.na(PCCLASS), popcenter_class == "Small (1000-29,999)") %>%
    group_by(Province, year_quarter) %>%
    count(u_status)
```

```
smallProvs_Upload <- do.call(rbind, by(smallProvs_Upload, smallProvs_Upload["Province"],
    tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

smallProvs_Upload %>%
    ggplot(mapping = aes(x = year_quarter, y = prop, color = u_status,
        group = u_status)) + #geom_bar(stat='identity')+ group
        group = u_status)) + #geom_bar(stat='identity')+ =
        group = u_status)) + #geom_bar(stat='identity')+ u_status))
        group = u_status)) + #geom_bar(stat='identity')+ +
        group = u_status)) + #geom_bar(stat='identity')+ #geom_bar(stat='identity')+
geom_line() + geom_point() + facet_wrap(~Province) + ggtitle("Proportion of Small
↪  Population Centers Meeting the Commitment\n(Upload Speed)") +
    labs(y = "Average Upload Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")

# Medium
medProvs_Download <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER"))),
        d_status = ifelse((avg_d_kbps/1000) < 50, "Below Commitment",
            "Meeting Commitment"), u_status = ifelse((avg_u_kbps/1000) <
            10, "Below Commitment", "Meeting Commitment")) %>%
    filter(!is.na(PCCLASS), popcenter_class == "Medium (30,000-99,999)") %>%
    group_by(Province, year_quarter) %>%
    count(d_status)


medProvs_Download <- do.call(rbind, by(medProvs_Download, medProvs_Download["Province"],
    tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

medProvs_Download %>%
    ggplot(mapping = aes(x = year_quarter, y = prop, color = d_status,
        group = d_status)) + #geom_bar(stat='identity')+ group
        group = d_status)) + #geom_bar(stat='identity')+ =
        group = d_status)) + #geom_bar(stat='identity')+ d_status))
        group = d_status)) + #geom_bar(stat='identity')+ +
        group = d_status)) + #geom_bar(stat='identity')+ #geom_bar(stat='identity')+
geom_line() + geom_point() + facet_wrap(~Province) + ggtitle("Proportion of Medium
↪  Population Centers Meeting the Commitment\n(Download Speed)") +
```

```
    labs(y = "Average Upload Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")

# Medium
medProvs_Upload <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER"))),
        d_status = ifelse((avg_d_kbps/1000) < 50, "Below Commitment",
            "Meeting Commitment"), u_status = ifelse((avg_u_kbps/1000) <
            10, "Below Commitment", "Meeting Commitment")) %>%
    filter(!is.na(PCCLASS), popcenter_class == "Medium (30,000-99,999)") %>%
    group_by(Province, year_quarter) %>%
    count(u_status)


medProvs_Upload <- do.call(rbind, by(medProvs_Upload, medProvs_Upload["Province"],
    tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

medProvs_Upload %>%
    ggplot(mapping = aes(x = year_quarter, y = prop, color = u_status,
        group = u_status)) + #geom_bar(stat='identity')+ group
        group = u_status)) + #geom_bar(stat='identity')+ =
        group = u_status)) + #geom_bar(stat='identity')+ u_status))
        group = u_status)) + #geom_bar(stat='identity')+ +
        group = u_status)) + #geom_bar(stat='identity')+ #geom_bar(stat='identity')+
geom_line() + geom_point() + facet_wrap(~Province) + ggtitle("Proportion of Medium
↪  Population Centers Meeting the Commitment\n(Upload Speed)") +
    labs(y = "Average Upload Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")


# Large
largeProvs_Download <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER"))),
        d_status = ifelse((avg_d_kbps/1000) < 50, "Below Commitment",
            "Meeting Commitment"), u_status = ifelse((avg_u_kbps/1000) <
```

```
            10, "Below Commitment", "Meeting Commitment")) %>%
    filter(!is.na(PCCLASS), popcenter_class == "Large (>100,000)") %>%
    group_by(Province, year_quarter) %>%
    count(d_status)


largeProvs_Download <- do.call(rbind, by(largeProvs_Download,
    largeProvs_Download["Province"], tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

largeProvs_Download %>%
    ggplot(mapping = aes(x = year_quarter, y = prop, color = d_status,
        group = d_status)) + #geom_bar(stat='identity')+ group
        group = d_status)) + #geom_bar(stat='identity')+ =
        group = d_status)) + #geom_bar(stat='identity')+ d_status))
        group = d_status)) + #geom_bar(stat='identity')+ +
        group = d_status)) + #geom_bar(stat='identity')+ #geom_bar(stat='identity')+
geom_line() + geom_point() + facet_wrap(~Province) + ggtitle("Proportion of Large
↪  Population Centers Meeting the Commitment\n(Download Speed)") +
    labs(y = "Average Download Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")


# Large
largeProvs_Upload <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER"))),
        d_status = ifelse((avg_d_kbps/1000) < 50, "Below Commitment",
            "Meeting Commitment"), u_status = ifelse((avg_u_kbps/1000) <
            10, "Below Commitment", "Meeting Commitment")) %>%
    filter(!is.na(PCCLASS), popcenter_class == "Large (>100,000)") %>%
    group_by(Province, year_quarter) %>%
    count(u_status)


largeProvs_Upload <- do.call(rbind, by(largeProvs_Upload, largeProvs_Upload["Province"],
    tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

# LASSO REGRESSION
```

```
library(glmnet)

y_download <- dt$avg_d_kbps
y_upload <- dt$avg_u_kbps
X <- data.matrix(dt[, -which(names(dt) %in% c("avg_d_kbps", "avg_u_kbps",
    "geometry", "year", "quarter", "PRUID", "CDUID", "PCUID"))])

# perform k-fold cross-validation to find optimal lambda
# value
cv_model_download <- cv.glmnet(X, y_download, alpha = 1)
cv_model_upload <- cv.glmnet(X, y_upload, alpha = 1)

# find optimal lambda value that minimizes test MSE
best_lambda_download <- cv_model_download$lambda.min
best_lambda_upload <- cv_model_upload$lambda.min

# produce plot of test MSE by lambda value
plot(cv_model_download, main = expression("MSE vs " * lambda[d] *
    " (Download Time)"))

plot(cv_model_upload, main = expression("MSE vs " * lambda[u] *
    " (Upload Time)"))


best_model_download <- glmnet(X, y_download, alpha = 1, lambda = best_lambda_download)
best_model_upload <- glmnet(X, y_upload, alpha = 1, lambda = best_lambda_upload)

as.matrix(coef(best_model_download), rownames) %>%
    knitr::kable(caption = "Sparse Estimates for Download Speed Predicton")


as.matrix(coef(best_model_upload), rownames) %>%
    knitr::kable(caption = "Sparse Estimates for Upload Speed Predicton")

# A note about missing data

temp <- dt %>%
    filter(is.na(PCNAME))

temp %>%
    transmute(`Census Division` = CDNAME, Province = PRNAME) %>%
    distinct() %>%
    knitr::kable(caption = "Census divisions with missing data")

rural_Download <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), popcenter_class = ifelse(PCCLASS ==
        2, "Small (1000-29,999)", ifelse(PCCLASS == 3, "Medium (30,000-99,999)",
        ifelse(PCCLASS == 4, "Large (>100,000)", "OTHER"))),
        d_status = ifelse((avg_d_kbps/1000) < 50, "Below Commitment",
            "Meeting Commitment"), u_status = ifelse((avg_u_kbps/1000) <
            10, "Below Commitment", "Meeting Commitment")) %>%
```

```
    filter(is.na(PCCLASS)) %>%
    group_by(Province, year_quarter) %>%
    count(d_status)


rural_Download <- do.call(rbind, by(rural_Download, rural_Download["Province"],
    tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

rural_Download %>%
    ggplot(mapping = aes(x = year_quarter, y = prop, color = d_status,
        group = d_status)) + #geom_bar(stat='identity')+ group
        group = d_status)) + #geom_bar(stat='identity')+ =
        group = d_status)) + #geom_bar(stat='identity')+ d_status))
        group = d_status)) + #geom_bar(stat='identity')+ +
        group = d_status)) + #geom_bar(stat='identity')+ #geom_bar(stat='identity')+
geom_line() + geom_point() + facet_wrap(~Province) + ggtitle("Proportion of Level 3
↪  Divisons Meeting the Commitment\n(Download Speed)") +
    labs(y = "Average Download Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")


ruralProvs_Upload <- dt %>%
    mutate(year_quarter = paste(year, quarter), Province = ifelse(grepl("\\/",
        PRNAME), PRNAME %>%
        str_extract(".*(?= \\/)"), PRNAME), d_status = ifelse((avg_d_kbps/1000) <
        50, "Below Commitment", "Meeting Commitment"), u_status =
        ↪  ifelse((avg_u_kbps/1000) <
        10, "Below Commitment", "Meeting Commitment")) %>%
    filter(!is.na(PCCLASS)) %>%
    group_by(Province, year_quarter) %>%
    count(u_status)


rural_Upload <- do.call(rbind, by(ruralProvs_Upload, ruralProvs_Upload["Province"],
    tibble) %>%
    lapply(function(x) do.call(rbind, by(x, x["year_quarter"],
        tibble) %>%
        lapply(function(y) y %>%
            mutate(prop = y$n/sum(y$n))))))

rural_Upload %>%
    ggplot(mapping = aes(x = year_quarter, y = prop, color = u_status,
        group = u_status)) + #geom_bar(stat='identity')+ group
        group = u_status)) + #geom_bar(stat='identity')+ =
        group = u_status)) + #geom_bar(stat='identity')+ u_status))
        group = u_status)) + #geom_bar(stat='identity')+ +
```

```
        group = u_status)) + #geom_bar(stat='identity')+ #geom_bar(stat='identity')+
geom_line() + geom_point() + facet_wrap(~Province) + ggtitle("Proportion of Level 3
↪  Divisons Meeting the Commitment\n(Upload Speed)") +
    labs(y = "Average Upload Speed (mbps)") + scale_color_discrete(guide =
    ↪  guide_legend(reverse = TRUE)) +
    theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        axis.title = element_blank(), legend.title = element_blank(),
        legend.position = "bottom")
```