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

15 February 2022

Introduction

(Quoted from the SCC 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 SCC would like to understand the state of internet connectivity in both rural and underserved Canadian communities using consumer-provided data. The SCC 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's realized internet access.For this study in particular, the identification of reliable and reproducible statistical methods to understand connectivity of rural and underserved Canadian communities is critical.

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).

While it is possible to apply some treatment to the missing data, after removal, the remaining sample is 1,252,560 rows, which is still usable for this analysis. As such a filtered data set is used.

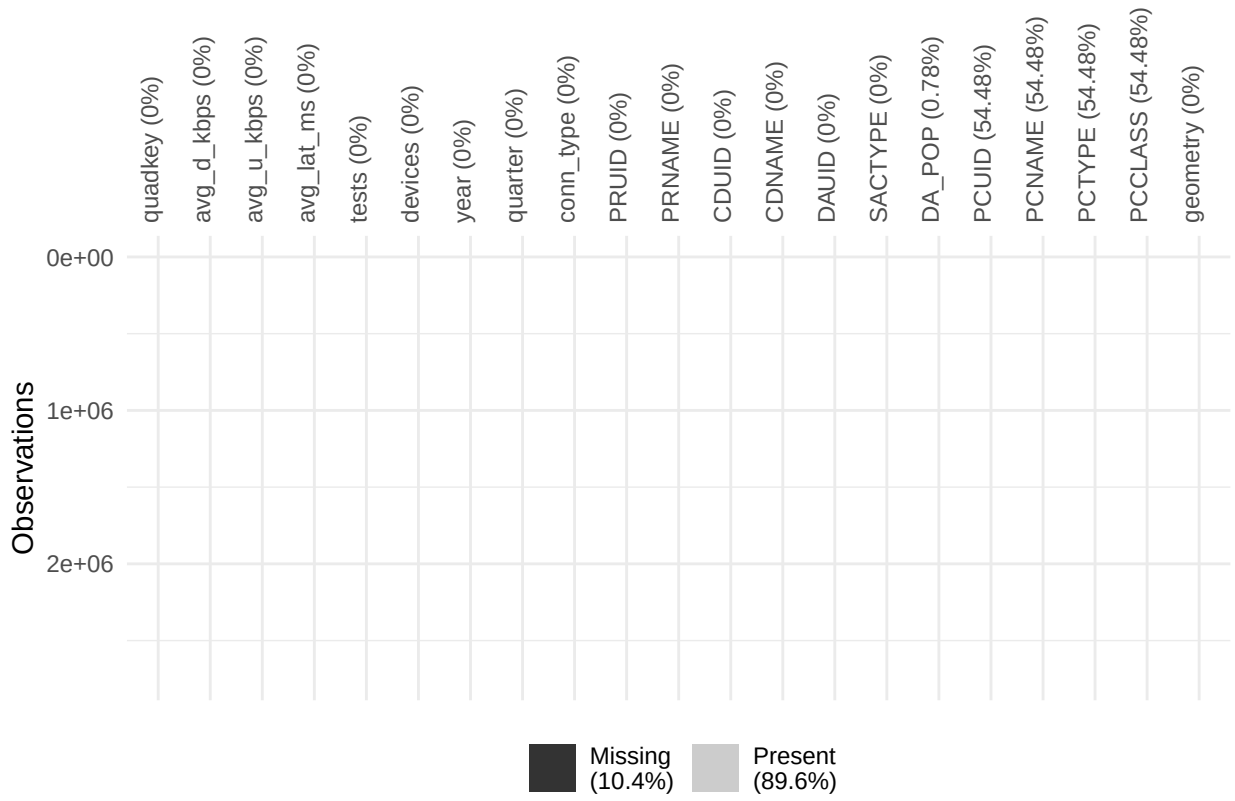

Figure 1: Missing data present in dataset provied

Analysis

Current Realized And Forecasted Internet Speeds (Upload And Download)- 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 effects being province, population class and their interaction. 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 this model development.

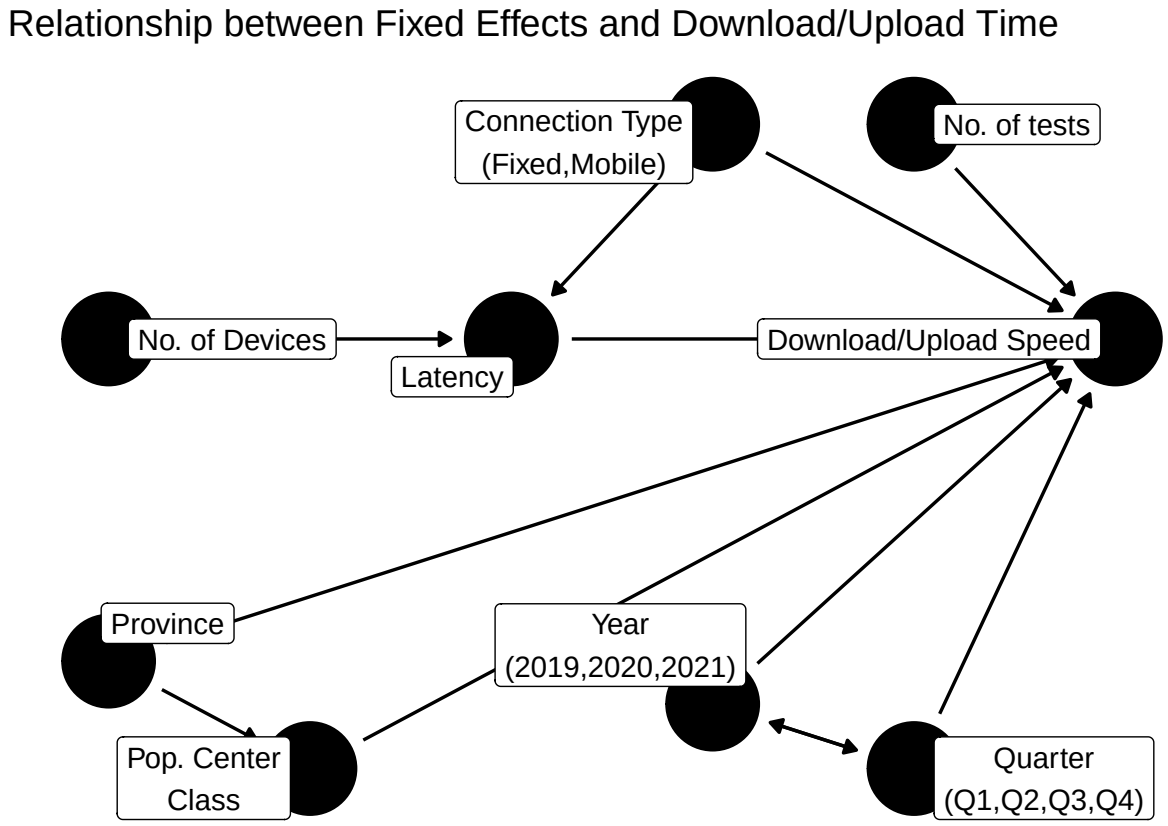

Figure 2: DAG representing the relationship between the fixed effects and upload/download time.

The fixed and random effects are 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	
Population Center Class	

Fixed Effects	Random Effect
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. 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.

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 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 4-9 outline such characteristics. In terms of small population centers, all provinces appear to be making progress in the Comittment 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.

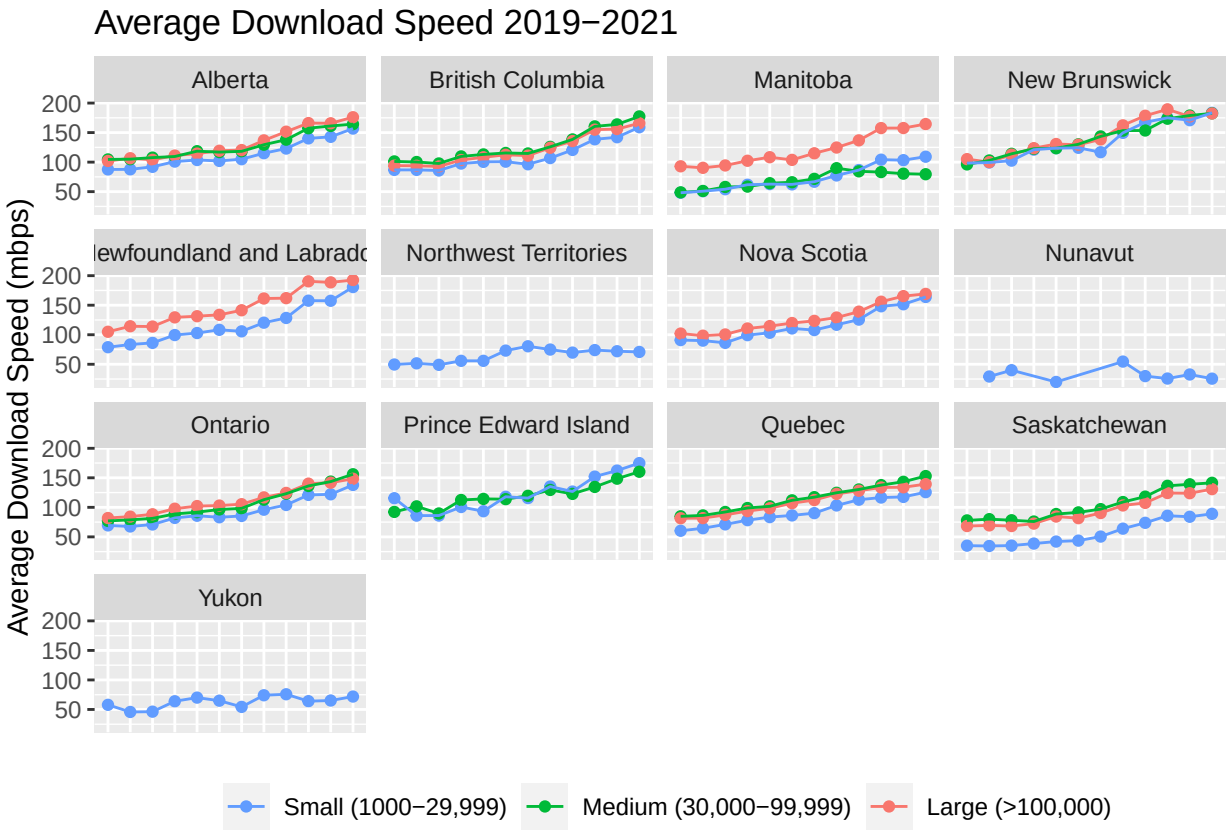

Figure 7: Average download speed accross provinces over time, by population center size

[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

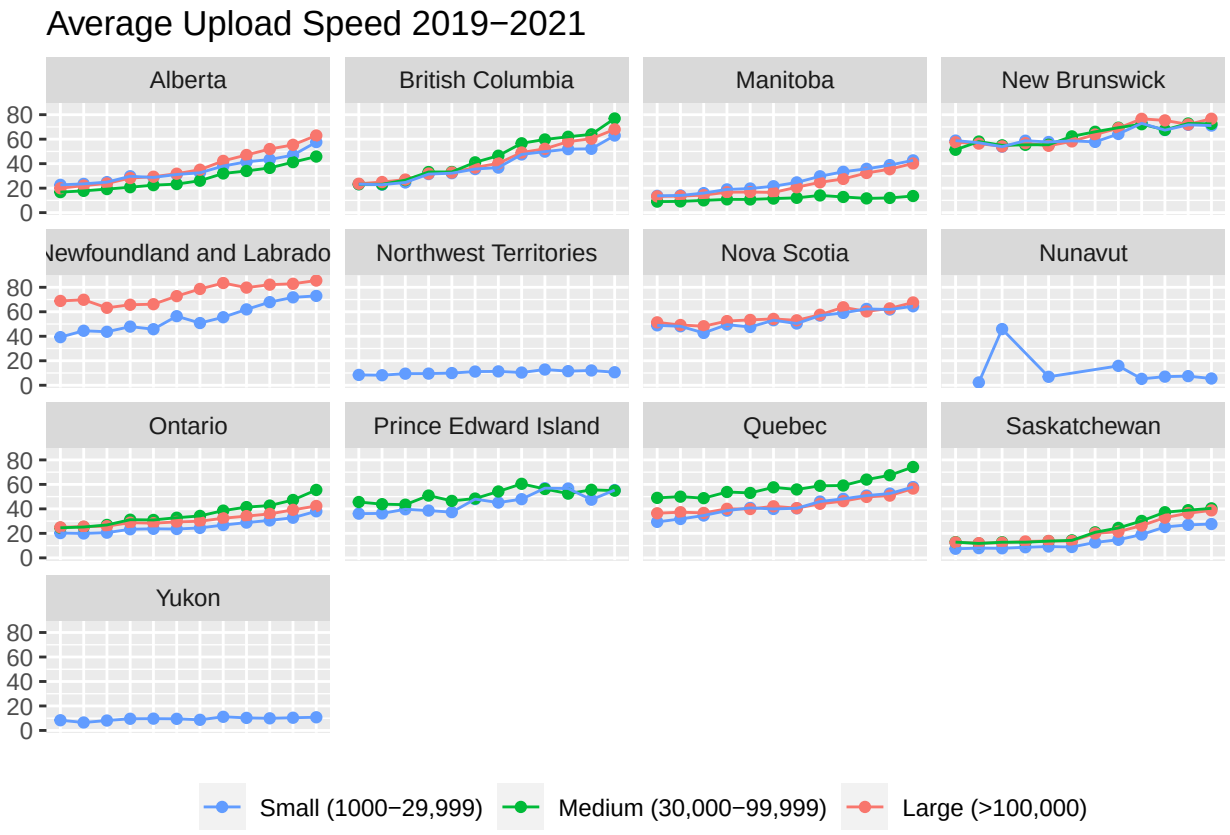

Figure 8: Average upload speed accross provinces over time, by population center size

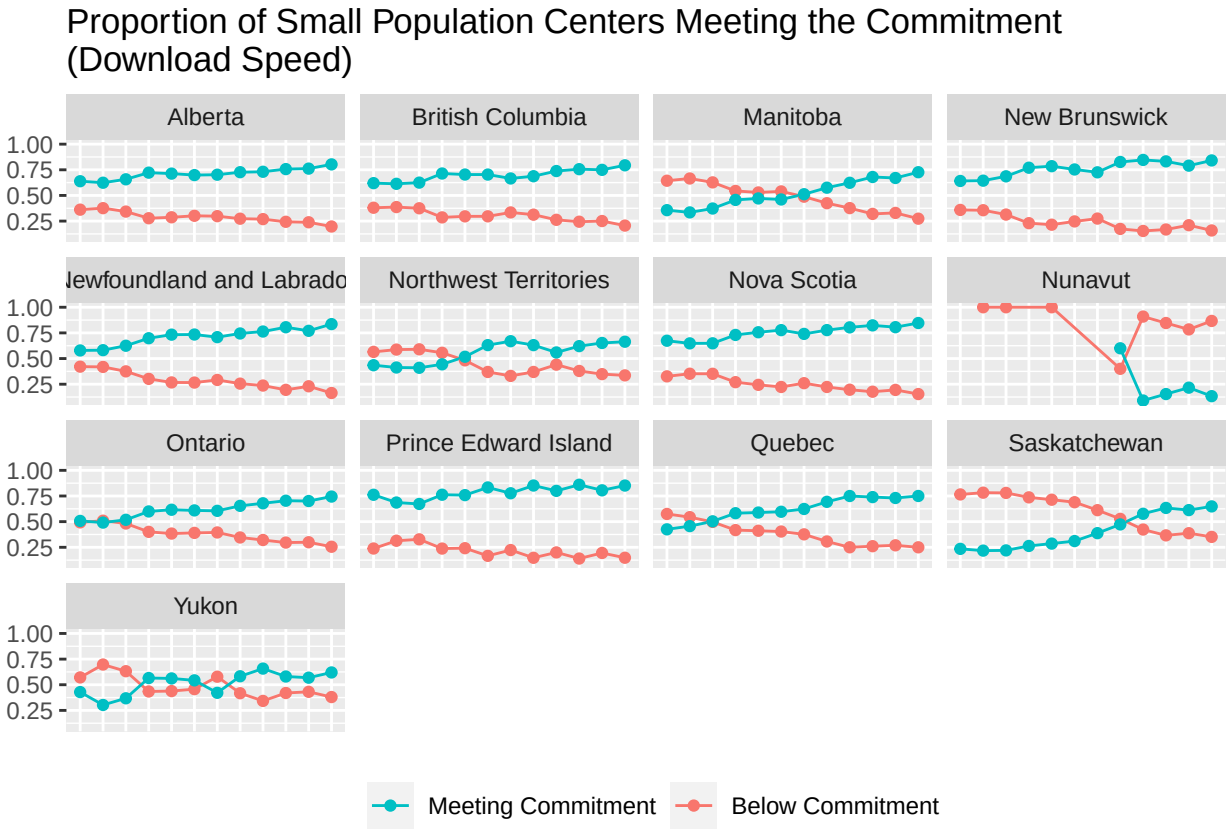

Figure 9: Proportion of small population centers accross canada meeting the commitment download speed over time

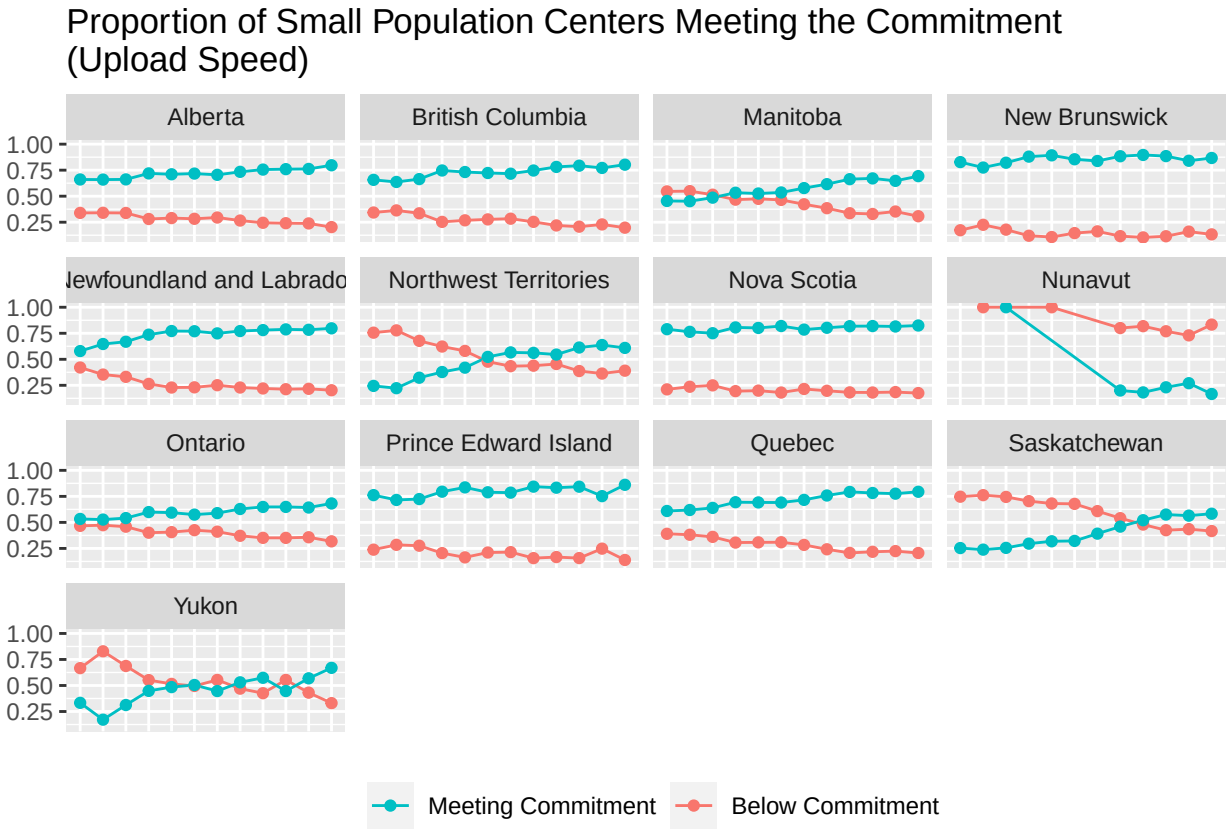

Figure 10: Proportion of small population centers accross canada meeting the commitment upload speed over time

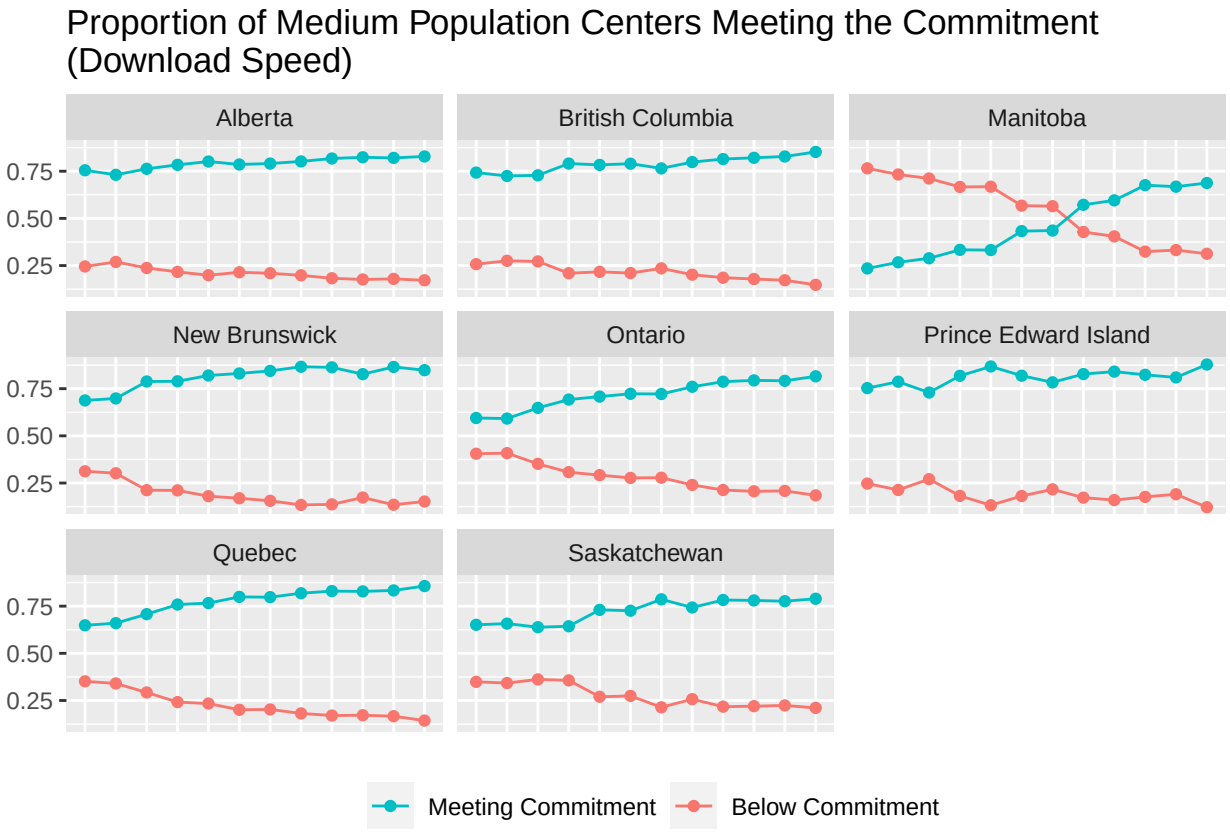

Figure 11: Proportion of medium population centers accross canada meeting the commitment download speed over time

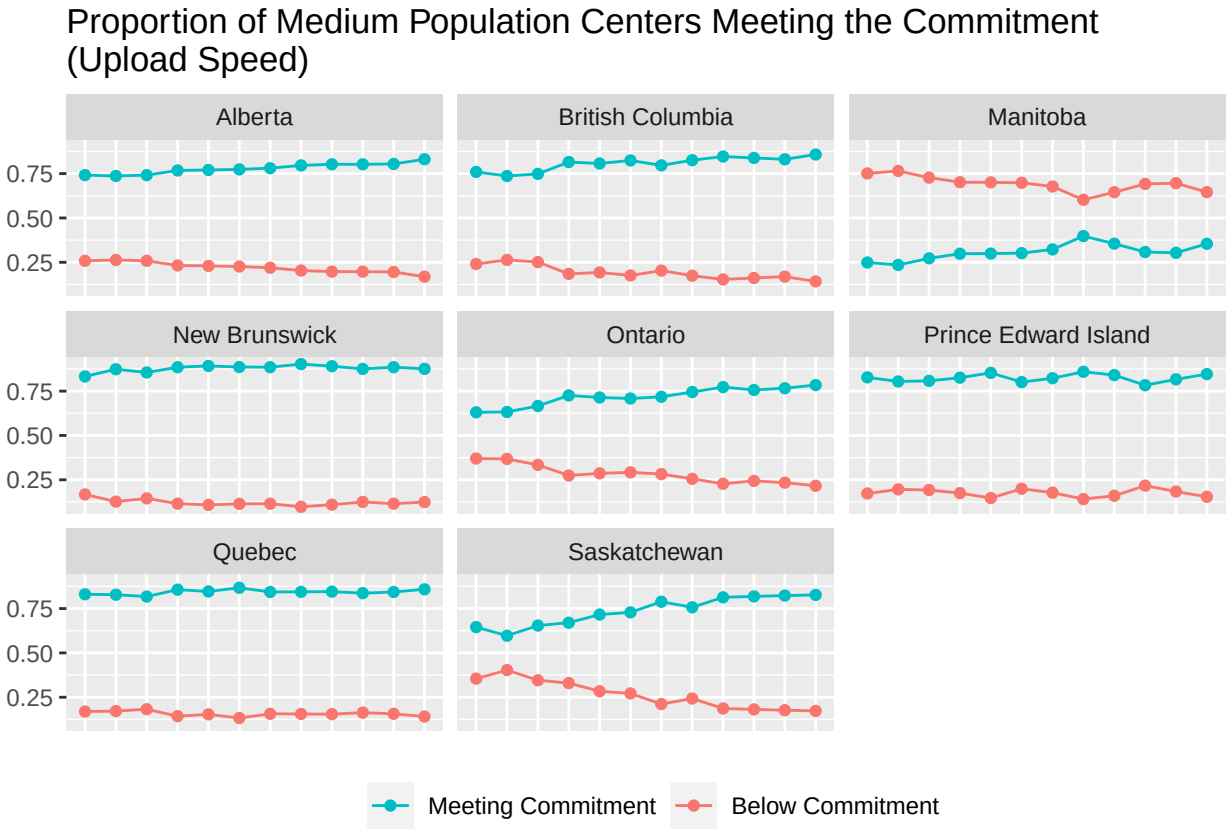

Figure 12: Proportion of medium population centers accross canada meeting the commitment upload speed over time

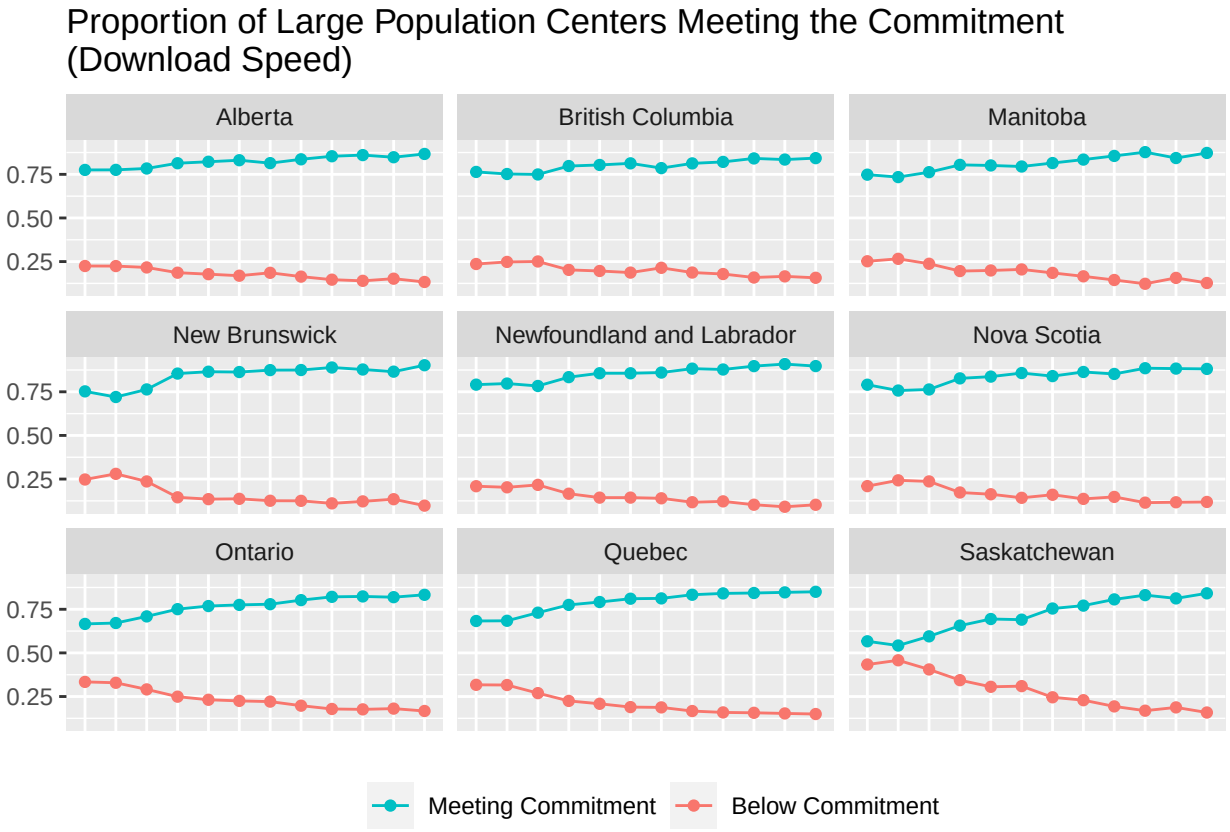

Figure 13: Proportion of large population centers accross canada meeting the commitment download speed over time

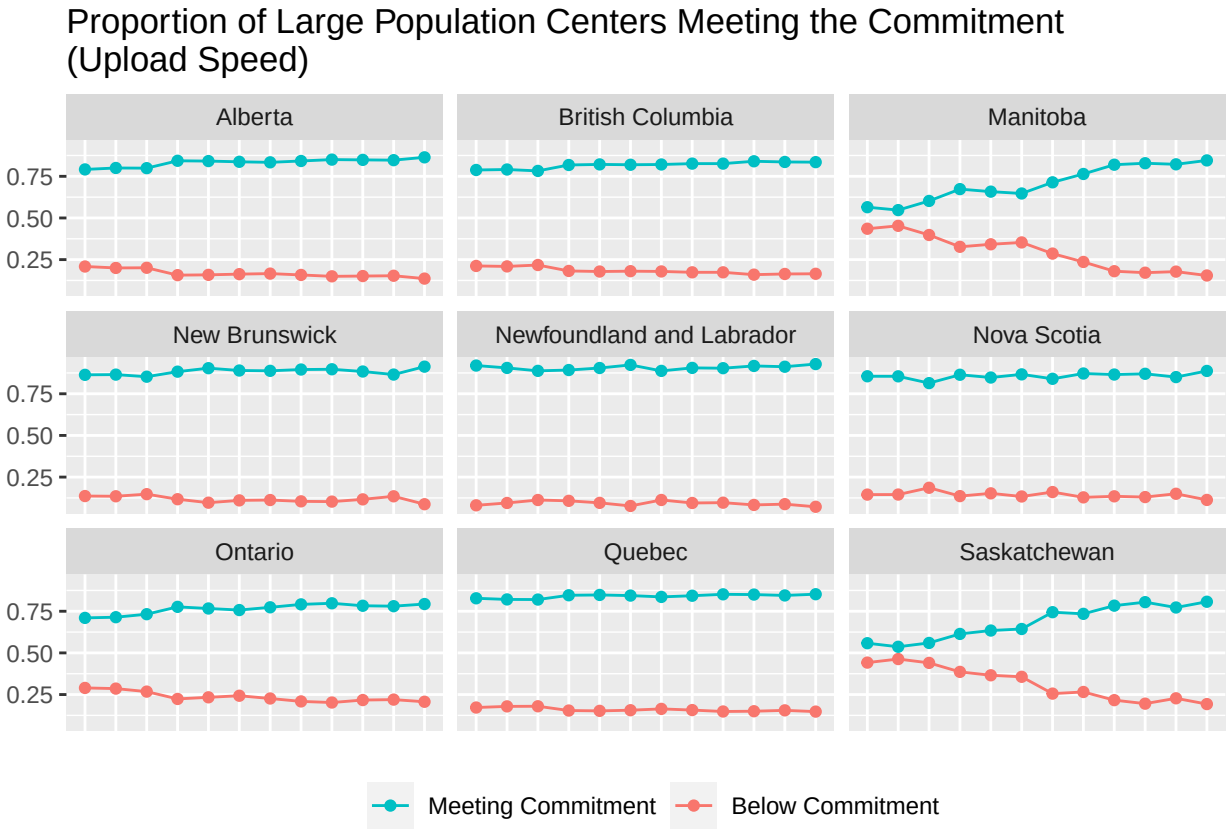

Figure 14: Proportion of medium population centers accross canada meeting the commitment upload speed over time

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.

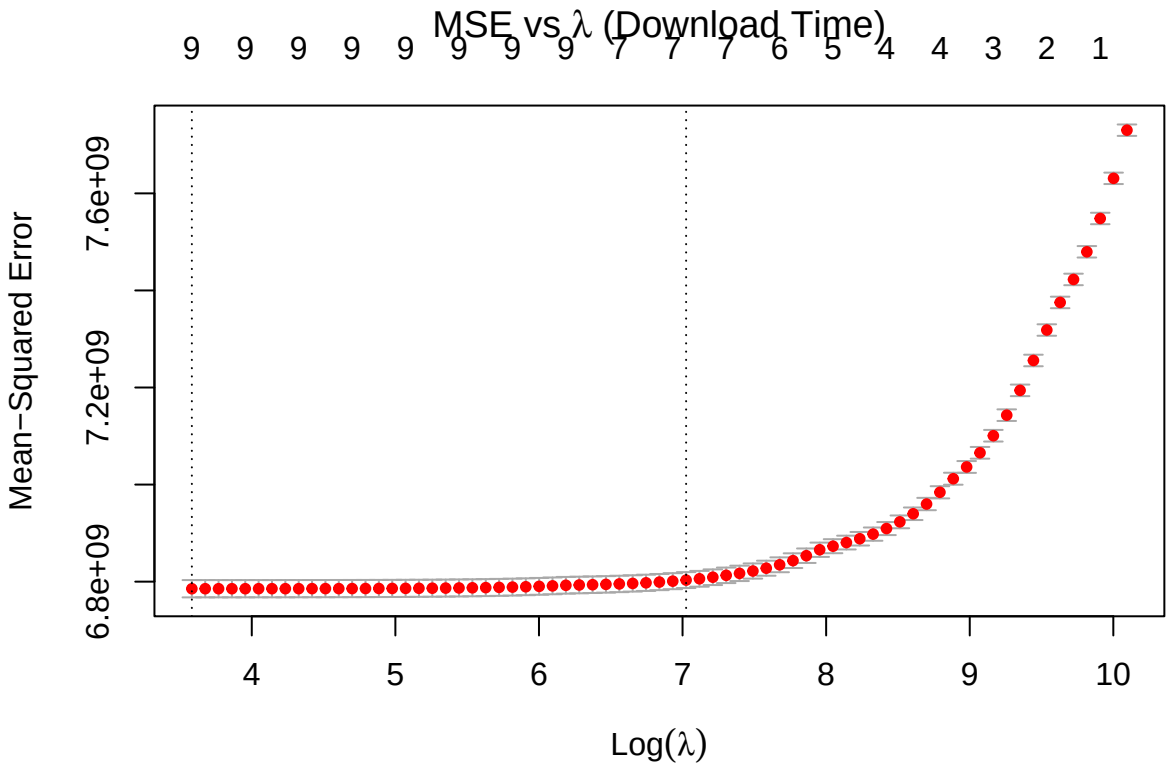

Figure 15: Download Time

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

	s0
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)	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

References

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*/
```

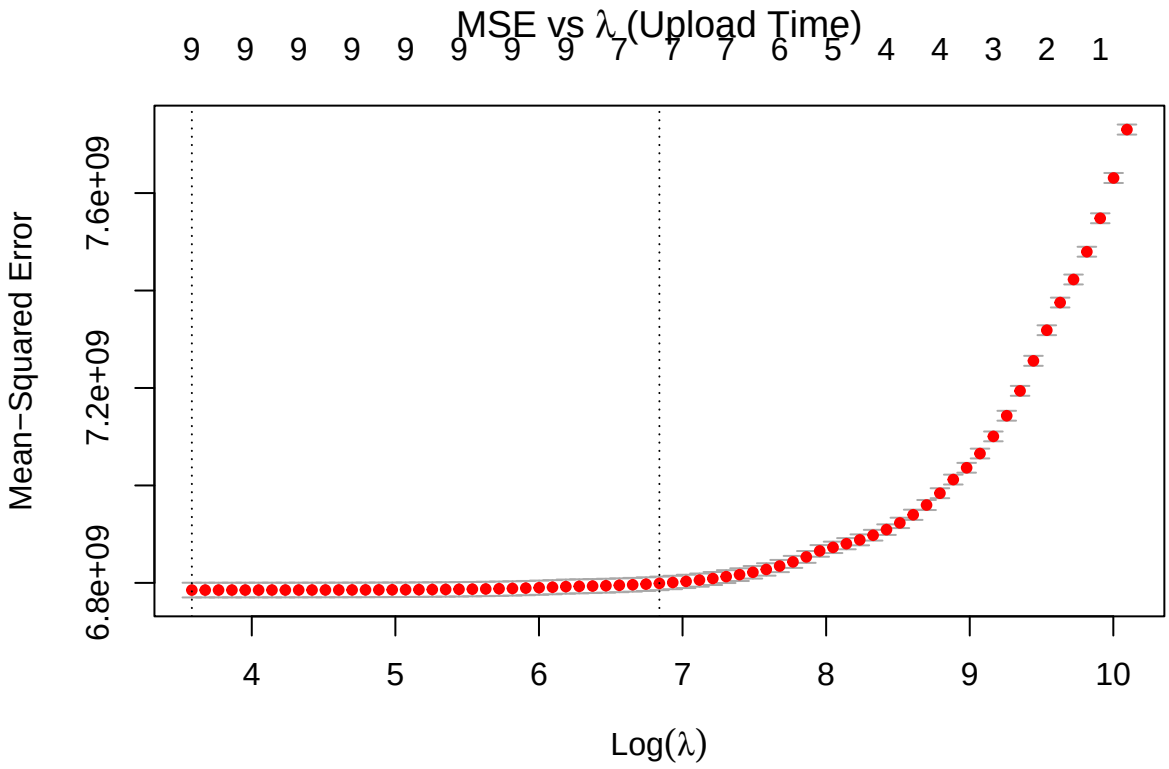

Figure 16: Upload

```
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;
```