Statistical Analyses of Internet Speed and Quality in Canadian Communities

Hamid Hamidi*, Thet Htet Chan Nyein*, Yanzhao Qian*, Bahar Mousazadeh*

Objective

By using the consumer-provided data by Ookla, we are planning to study the progress towards the commitment for the rural and underserved areas by

- Using visualization and transformation to explore our data in a systematic way
- Implementing a reliable statistical model to forecast the progress towards the commitment in 2026 and 2030
- Performing comparative analysis of rural and underserved communities in terms of progress towards commitment

Data Structure

- Performed the statistical analysis based on Dissemination Areas [1] (DA) instead of tile-level since tiles are insufficient to represent the status of a geographical area
- An area is Rural:
  - If the DA population is not available, and if more than half of the tiles in a DA are missing the population centre.
  - If available, and the DA population < 1000, or population density for each DA < 400 people per km^2.
- An area is Underserved:
  - If it is in the 75th percentile of latency (ping) and the 25th percentile of average download and upload speed.
- Included the latency in the Underserved definition as high latency results in noticeable lag when performing online activities
- PCCLASS: Taking the highest PCCLASS among the tiles in each DA.
- Introduced the Time variable: Combined the year and quarter covariates and encoded them into a numeric variable; for instance, time=1 represents Q1 of 2019
- Target: A binary variable used to measure progress towards commitment
- Target is 1 if average download speed >= 50 Mbps and average upload speed >= 10 Mbps and 0 otherwise.

Exploratory Analysis

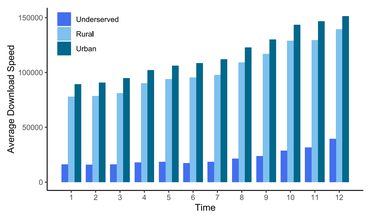

Fig. 1: Average Download Speed for Rural, Urban and Underserved Areas.

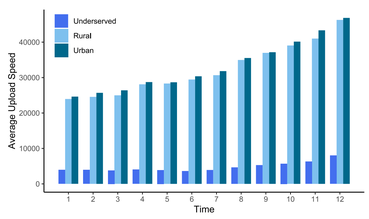

Fig. 2: Average Upload Speed for Rural, Urban and Underserved Areas.

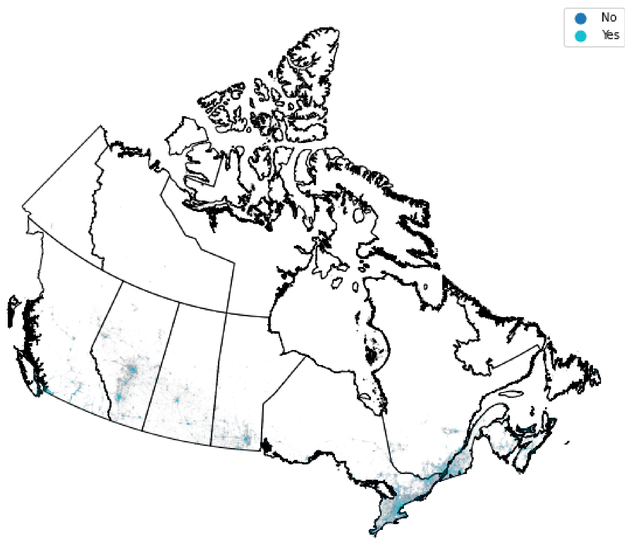

Fig. 3: Canada Map in Terms of Commitment for Fixed Connection Type in 2021.

- Fig 1 and 2. shows a comparison between the average download and upload speeds in different areas with respect to time. The download and upload speed in all areas are increasing.
- Fig 3. shows the map of Canada in terms of whether distinct areas have reached the commitment or not for fixed connection type by late 2021. The majority of the rural areas have not reached their commitment for fixed connection type.

Methods

Logistic Regression

- Split the dataset by connection type (Mobile, Fixed)
- Split the dataset to training set (time < 12) and test set (time = 12) which is ~9% of the data.
- Fitted three different logistic regression models; one for Fixed, one for Mobile, and for one the whole data set.
- Used predictors: number of tests, number of devices, DA population, time, province, Statistical Area Classification, Underserved or not, Rural or not, Population Centre Class
- Response: Target
- Included interaction terms [time × is_rural, time × is_underserved] to see the time effect for different categories
- The accuracy of the model for the whole dataset, the mobile dataset, and the fixed dataset are 0.8083, 0.6609, and 0.9156, respectively.

Results

Underserved communities form 12.09% of total Dissemination Areas by late 2021. Our logistic model for fixed connection predicts that both rural and underserved areas will achieve their commitment to high-speed internet by 2026. For mobile connection, the model predicts that 57.29% of rural areas will achieve the target; however, none of the underserved areas will have high-quality connectivity.

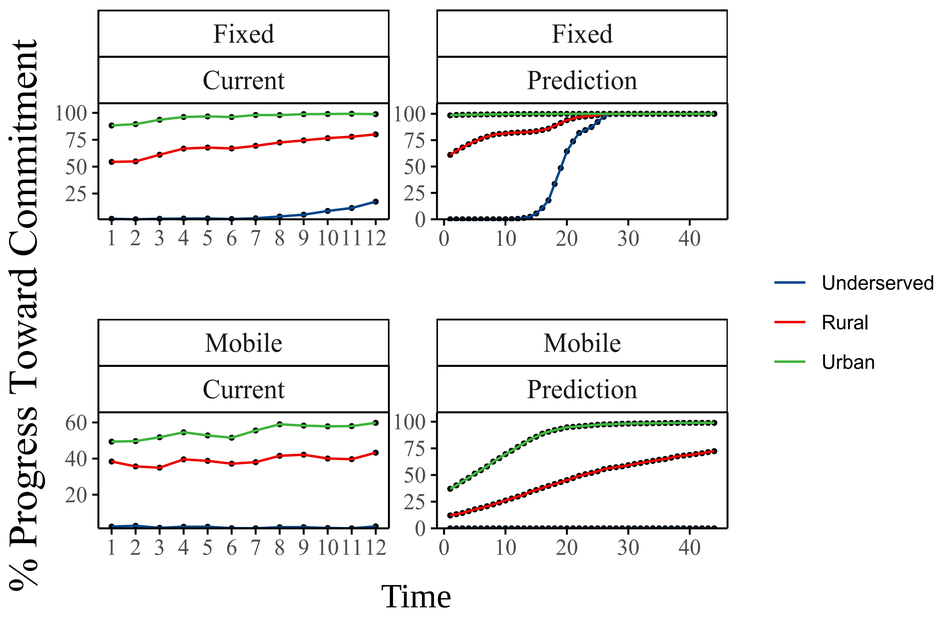

Fig. 4: Comparison for Current and Prediction Percentage of DA Reached Commitment for Mobile and Fixed Connections.

Table: Coefficients and Wald Statistic and P-value for Full Model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.25E-01	3.67E-02	-1.43E+01	1.49E-46
tests	-3.04E-04	1.98E-04	-1.54E+00	1.24E-01
devices	1.18E-01	1.31E-03	9.04E+01	0.00E+00
DA_POP	1.84E-04	6.88E-06	2.67E+01	4.88E-157
time	1.86E-01	3.64E-03	5.12E+01	0.00E+00
c(conn_type)mobile	-4.88E-01	1.37E-02	-3.56E+01	2.49E-277
c(is_underserved)True	-4.00E+00	6.00E-02	-6.67E+01	0.00E+00
c(is_rural)True	-4.40E-02	2.40E-02	-1.83E+00	6.70E-02
c(PCCLASS)	2.25E-01	2.87E-03	7.84E+01	0.00E+00
time:c(is_rural)True	4.27E-03	3.48E-03	1.23E+00	2.20E-01
time:c(is_underserved)True	6.42E-02	7.13E-03	9.01E+00	2.11E-19
time:c(conn_type)mobile	-1.60E-01	2.03E-03	-7.89E+01	0.00E+00

Goodness of Fit

- The model accuracy for the whole dataset, the mobile dataset, and the fixed dataset are 0.8083, 0.6609, and 0.9156, respectively.
- Fig.5 shows that the fixed model performs the best followed by the whole and mobile models which is supported by the AUC, 0.889, 0.78 and 0.943, respectively.
- Binned Residual Plot in Fig.6 shows that most residuals are located between the ±2 SE bands.
- The p-value of Residual Goodness of fit test is 1 for the three models.

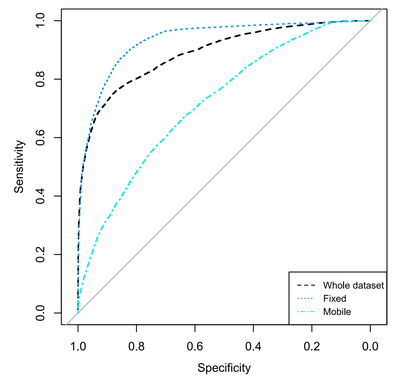

Fig. 5: ROC Plot for Three Models.

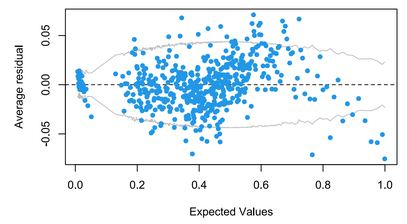

Fig. 6: Binned Residual Plot for Fixed Dataset.

Conclusion

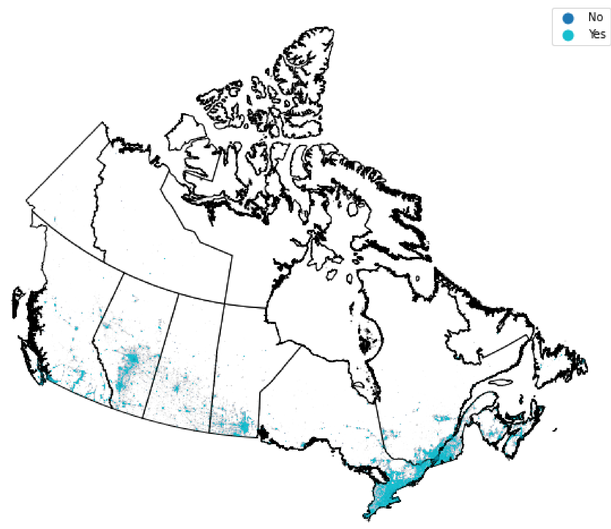

Fig. 7: Canada Map in Terms of Commitment for Fixed Connection Type in 2026.

Our analyses show steady development in internet speed in most areas of Canada for fixed and mobile connection types. However, underserved communities have large disparities in terms of internet access compared to rural and urban areas for both fixed and mobile connection types. Specifically, mobile connection with current trends would not make any significant progress toward commitment.

In this study, 40% of the mobile data is missing. Therefore, the fitted model has the lowest accuracy compared to other models for predicting the progress towards commitment. Additionally, the data is longitudinal data (multiple measurements) and should be analyzed with Generalized linear mixed effect models; however, fitting the GLEM models would need computational resources for big data sets which we did not have access to.

Acknowledgements

We thank Dr. Matthew Greenberg for his invaluable insights and discussions.

*Authors' affiliations: Department of Mathematics and Statistics, The University of Calgary.
*All authors have equal contribution.

[1] Dissemination areas are the smallest area for which public population statistics are distributed by Statistics Canada.