Latent Gaussian Models for Spatio-Temporal Internet Connectivity Data

Christian Mitrache, Philipp Schröppel

Department of Statistics and Actuarial Science, University of Waterloo

Current State and Future Directions

Assessing the State of the Commitment

- To estimate the share of the population which lives in locations where the commitment is met, internet-speed data (tiles) has to be combined with census data.
- Tiles are associated to the dissemination area with the largest overlap (see left figure below).
- The population per tile is estimated by dividing the population of a dissemination area evenly between the tiles which are associated to it.
- Based on the population per tile and the measured internet speeds it is possible to estimate the share of the population with access to high-speed internet.

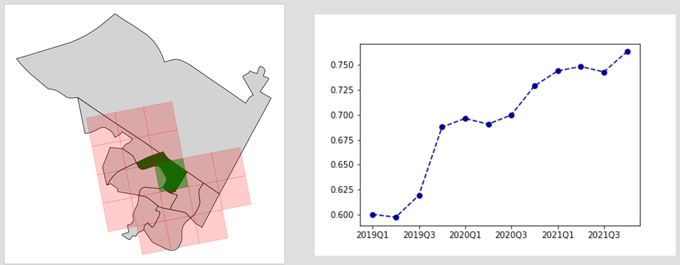

Fig. 1: Left: Dissemination areas (gray) overlaid with speed tiles (red). The green cells visualize the mapping of tiles to dissemination areas by largest overlap.

Right: Estimated share of Canadian population with access to high-speed internet (as defined in the commitment).

Going Forward: Expanding Connectivity

- Our estimate indicates that given the historical trend it is unlikely that the intermediate goal for 2025 will be reached.
- We propose a metric called action score to recommend locations for future investment in connectivity related infrastructure

$$\text{a(T)} = \text{(1-s)} * \log\left(\frac{population(T)\ devices(T)\ tests(T)}{\text{1+ } dist2closestPC(T)/1000}\right),\ T\ tile$$

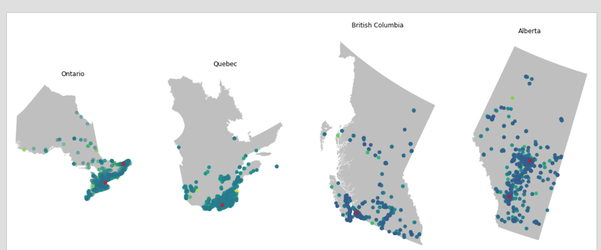

Fig. 2: Locations of top 1000 action scores with marked positions of major cities / IXPs (Toronto, Ottawa, Montreal, Vancouver, Calgary, Edmonton).

Feature Engineering

Features Considered:

1. $s_u = \min(1, \text{avg}_{u\_kbps}/10000)$, $s_d = \min(1, \text{avg}_{d\_kbps}/50000)$, $s = \min(s_u, s_d)$
2. $Y = \lfloor s \rfloor$: (boolean representing if tile has access to commitment)**
3. x_1 : Distance to Closest Internet Exchange Point (IXP) **
4. x_2 : Closest Internet Exchange Point (IXP)
5. x_3 : Distance to Closest Population Center
6. x_5 : Median Household Income Per Census Region
7. x_6 : Dissemination Area Population

** indicates our belief in the significance of the new feature. (After modelling and exploratory analysis)

Aggregation for Computational Feasibility

To lower the computational requirements, we considered the aggregation of relevant model features on the level of dissemination areas.

Dessem_id	Province	x	y	Commitment Counts	closest_city_dist	closest_PC_dist	Population	number_of_tiles	conn_type
<int>	<chr>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
10010165	Newfoundland and Labrador / Terre-Neuve-et-Labrador	8978078	2146503	3	807758.0	3269.055	505	4	fixed
10010168	Newfoundland and Labrador / Terre-Neuve-et-Labrador	8978528	2146948	1	808389.3	3713.691	520	1	fixed
10010170	Newfoundland and Labrador / Terre-Neuve-et-Labrador	8978289	2147287	1	808429.8	3511.169	400	1	fixed

Fig. 3: Distances were averaged, population and counts were summed.

Conclusions

- Based on our action score we conclude that there are a large number of under-served communities close to population centers with a high potential to improve connectivity quickly and cost-efficiently.
- There are significantly different spatial dependence structures among the different provinces.
- Given the historical trends, it is unlikely that government of Canada will meet their commitments.

References

- https://www.cira.ca/community-investment-program/canadas-internet-infrastructure-internet-exchange-points-ixps
- https://www.paulamoraga.com/book-geospatial/sec-geostatisticaldataexamplespatial.html

Modelling The Proportion of Households Where the Commitment is Met

Computational Considerations

We acknowledge that there is a small increasing time dependency. However, due to computational limitations, this is taken into account by only considering the counts for each tile in the previous quarter.

Model Definition:

Let $\mathbf{z_i}$ be the centroid of the tiles for dissemination area i. For a random variable (X_j) where j indicates a tile, define the aggregated random variable on dissemination area i as X'_i. Denote $Y'_{i\,(-1)}$ as the counts in dissemination area i for the previous quarter. Finally, for the model definition, we use the notation defined in the feature engineering section.
Then the model is defined as:

$$Y'_i | P(\mathbf{z'_i}, x'_{6,i}, x'_{1,i}, Y'_{i\,(-1)}) \sim Binomial(N_i, P(\mathbf{z_i}, x'_{6,i}, x'_{1,i}, Y'_{i\,(-1)}))$$
$$logit(P(\mathbf{z_i}, x'_{6,i}, x'_{1,i}, Y'_{i\,(-1)})) = \beta_0 + \beta_1 x'_{1,i} + \beta_6 x'_{6,i} + \beta_{(-1)} Y'_{i\,(-1)} + S(\mathbf{z_i})$$

Where $S(z_i)$ is a spatial random effect with Matern Covariance:

$$cov(S(\mathbf{z_i}), S(\mathbf{z_j})) = \sigma^2(||\mathbf{z_i} - \mathbf{z_j}||K_1(||\mathbf{z_i} - \mathbf{z_j}||))$$

and K_1 is the modified Bessel function of the second kind with order 1.

As part of the model validation, we train this model on the 2021 Q3 data, in order to predict the 2021 Q4.

Performance of Model in Different Provinces

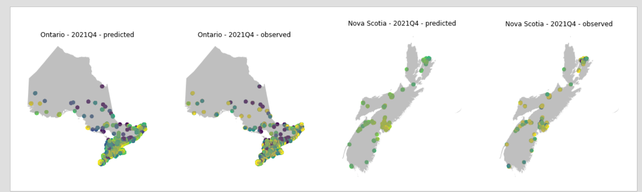

Fig. 4: Plotting the predicted proportions for 2021 Q4 against the true observed proportions.

- Different regions have different spatial effect structures. However, spatial covariance isn't strong.
- Model estimates closely resemble true observed proportions.
- Time series component appears to be more significant then spatial component.
- High internet speed areas contain small pockets of low internet speeds.