Master of Data Science - University of British Columbia Okanagan

Minutes for May 6, 2020 - 11:00 AM - 12:00 PM PST via Microsoft Teams Meeting

Present: Bruno St-Aubin (Statistics Canada), Marian Radulescu (Statistics Canada), Sofia Bahmutsky, Ngan Lyle, Kaitlyn Hobbs (Minutes), Shreeram Murali

Agenda

Discussion Points

1 — Potential statistical analysis using PCA or FA.

2 — Working on a server and reproducible methods.

Meeting Notes

Reviewed Data and Notes Received from Bruno —

- HR level COVID data (feature server link) can run queries on database in backend. Shows all health regions in Canada as polygons with cases reported. Data is updated in real time. Health authorities data includes the provincial polygons.
- COVID-19 dataset is compiled from a Professor in Toronto.
- ODHF includes geocoding, it does not have geometries while ODB does. As a result, centroids in ODB will not match on latitude and longitude in ODHF (e.g. if a hospital has multiple buildings). Could create a buffer between ODB layer, which would represent geometries for health facilities. If we are showing on a map, we don't need to merge, we could layer ODHF on top of ODB. **"Record Linkage"** is terminology used for a join.
- Open Database of Addresses may be available to us at it's current stage of development if we choose to complete the ODB.
- Dissemination blocks adjust in size according to density (large density shrinks dissemination block size and vice versa).
- Potential to expand beyond long-term facilities in ODHF (ie. ambulatory services) and include proximity data.
- Publishing a report is dependent on our results and requires stringent assessment from Statistics Canada.
- Aggregating proximity data for a health region may provide better comparisons between regions.

Incorporating PCA — Classification of long-term care facilities using PCA. Cross reference with cases in long-term facilities. Potential variables include: *Density of neighbourhood, ratio of staff to patients, dissemination blocks, population density per building, proximity data.* The result would be an inferential model. Case data would be a proportion of outbreak and normalized by testing count.

Factor Analysis — To determine latent variables not observed.

Questions and Answers —

Question from Ngan: Has anyone tried to link attributes to ODB?

Answer from Statistics Canada: Yes, there's a project under way. Open Database of Addresses is coming as a basis for linkage of all databases together. It will evade troubles with geometry issues but will not be ready in time for this project.

Question from Ngan: Using a server or whatever methods is standardized for you for reproducibility.

Answer from Statistics Canada: Generally, code is stored on Github. Servers may require more effort and technical skills. If databases are modified then load as a temporary CSV until Statistics Canada can integrate it. For long-term large data, using arcGIS or Mapbox, a server will be required. You would need someone familiar with coding networking or servers.

Mapbox infrustructure is used as a 3rd party server and can optimize data in a web application using tile layers of JSON data on varying zoom levels (exponential division, allowing for granular data). Running a query on rendered data only. Search bars rely on servers. You can create an account and receive a token. 50 000 hits per month are free.

Question from Sofia: Is there a preference for coding in R or python?

Answer from Statistics Canada: Whatever you're more proficient in. There is open source QGis software to visualize spatial data files. GeoPandas can be unintuitive but is commonly used. Sas is also used by StatCan but has a required cost.

Final Thoughts

1. Use GC collab email for further communications.

Next Meeting: Wednesday, May 13, at 11:00 AM PST with Bruno and Marian via Microsoft Teams Meeting