CMPT 353: COVID-19 Risk Assessment

Motivation

In light of social lockdown measures for COVID-19, we observed that the dataset provided in the assignment document - which is polluted with an excessive amount of low interest data - created a unique opportunity to attempt risk assessment of the disease within a city. Most inquiry into COVID-19 risks focus on high-risk areas like public centres and institutions, which would normally overshadow the smaller surrounding amenities. In short: we understand the risks of a shopping mall, but know nothing about the risk of your average bus stop.

Data

The dataset used was the provided OSM (OpenStreetMap) data, which contained gps coordinates (lat, long), amenity type, a timestamp of the last edit, an optional name field, and secondary tagging information. These tags included items like washrooms and disabled access, as well as tags to elaborate on the amenity type (such as the type of fuel at a gas station, recycling availability at drop-off points, and hours of operation). The list of all available tags is in the OSM documentation, but a printSchema function was used to show only the tags present in the data set, since many nodes are incompletely tagged (or not at all). This data was flattened from the json format into a more manageable (albeit larger) csv for ease of use and organization, so individual tags could be indexed in bulk by comparing to a list.

We also used a data set for confirmed public covid exposures released by Vancouver Coastal Health. The data only includes cases within the Vancouver area, which we can assume to be roughly similar to the GPS bounds of the OSM data.

Because of the way we used the tagging info to create our baseline risk assessment, there is no need to remove outliers from the data set directly. We make the assumptions that an object tagged must exist, because it would not make sense for random GPS coordinates and labels to exist in the data set, and we have no way of reliably verifying if the data reflects a node being closed. The "outliers" in this case are nodes discarded in the initial risk assessment: these are either tagless or categorically uninteresting to us. Of the tags in SCHEMA.txt, only a handful were actually used for analysis, the rest indicating service capability rather than purpose (tags such as recycling availability, gas station fuel types, and whether or not the designated park bench has a backrest).

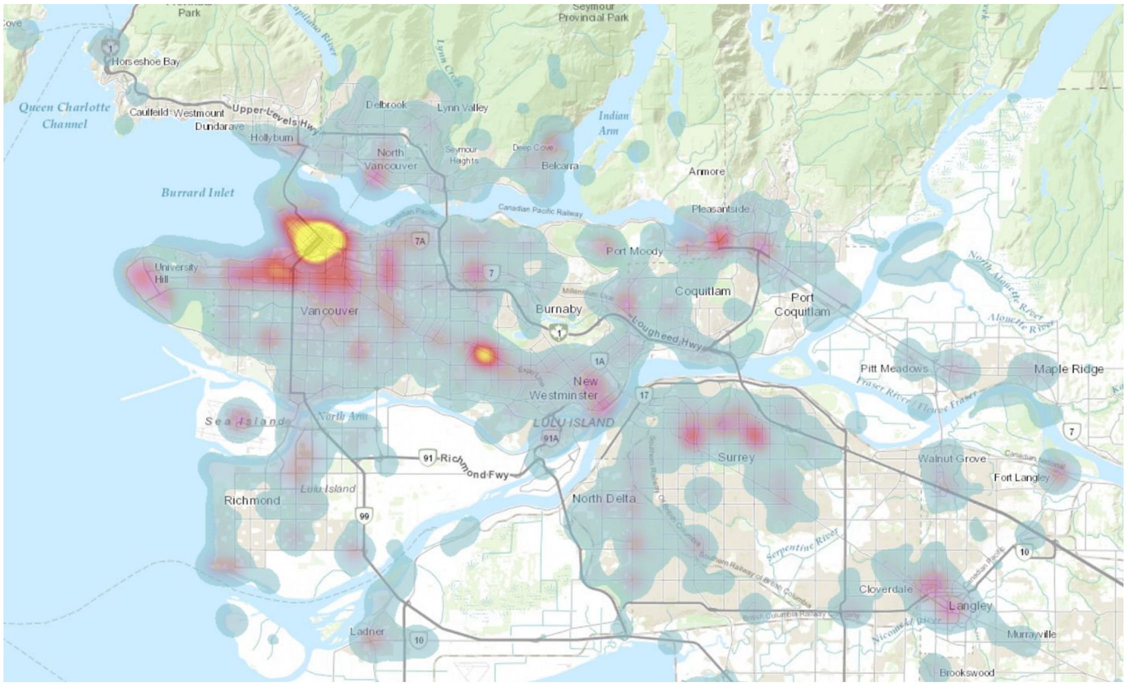
Analysis

The initial analysis assigns a risk factor to each node based on its tags, and is intended to be a primitive starting point. Relevant tags fall into five categories: food, medical facilities, transportation, gathering centres, and a miscellaneous "notable" category, allowing a node to fall into one, multiple, or no categories. The first four categories should make sense - diseases spread faster in confined, public spaces like transportation hubs, food vendors, and shopping centres. The second is more of a best guess based on skimming over the data: nodes with documentation tend to be more "notable". This includes branding info, wikidata entries, Wikipedia entries, and Yelp reviews. Weighing this category high in our risk assessment could pollute the data, but it seems too important to gloss over entirely.

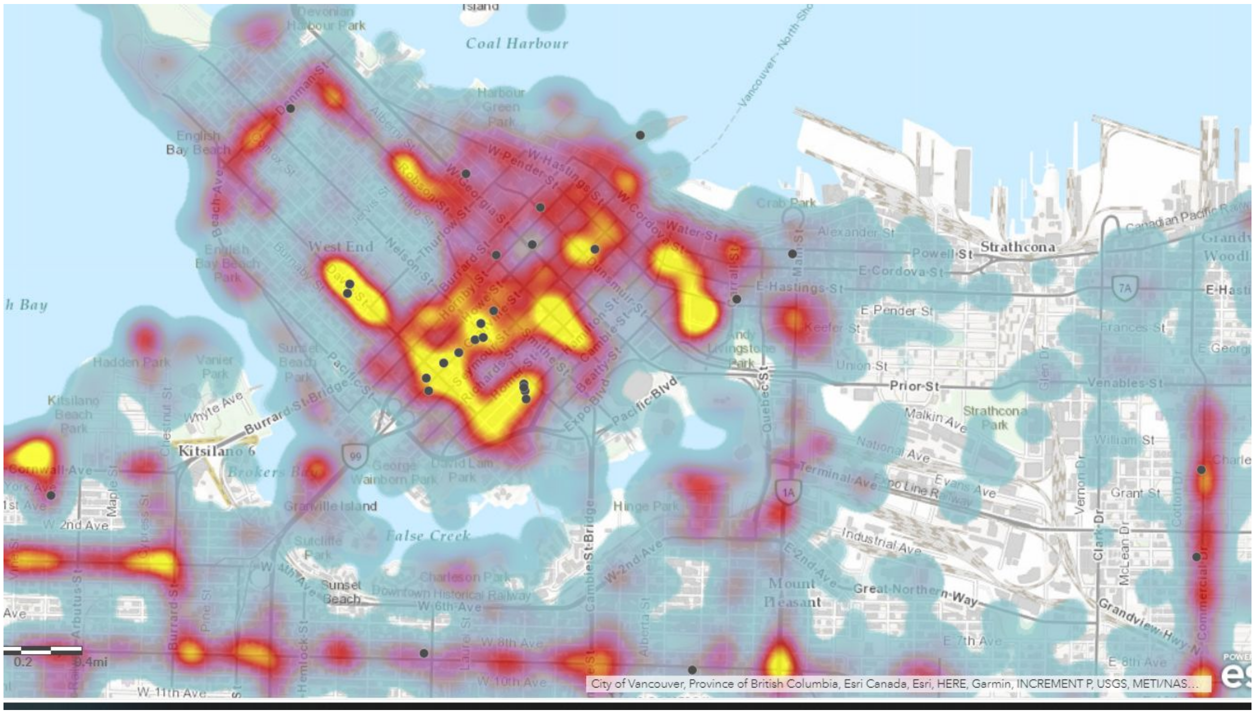
From here we refine our assessments by asking two questions: should proximity to a high-risk node make a node higher risk (yes), and would including this make for a better risk assessment (duh). We can use k-nearest-neighbours to factor in a node's surrounding risk factors, as well as their land distance (normalizing the data to [0, 1] for ease of model training).

APIs like ArcGIS also provide a better way to visualize this data over the region than something like Seaborn, and allows the higher risk regions to naturally combine rather than be represented in a grid. The initial heatmap is reassuring: our higher risk hotspots seem to overlap with what we would expect - and know from public lockdown measures - to be high risk. Higher density residential areas, commercial districts, and city centres all light up, with public transportation hubs being the brightest (most SkyTrain stations along the Expo line are clearly visible, for instance).

Even more encouraging is not just a population heatmap of the Vancouver area, as primitive data analysis tends to become. This means that our risk assessment is accounting for human activity over raw density, and the data is less likely to be polluted by small pockets of low-value nodes (bicycle parking, for example, would light up an area equivalent to multiple SkyTrain stations without the locality factor).



The below heatmap screenshot shows the IRL cases superimposed over the previous heatmap. Zoomed in to the downtown core to better see the details of the heatmap and which streets were more risky. As we can see, many of the covid exposures happened in the hottest area of the map, not surprising given the dense number of restaurants and people who live and use those areas. This further confirms the accuracy of our prediction model to predict possible covid exposure sites in the future.

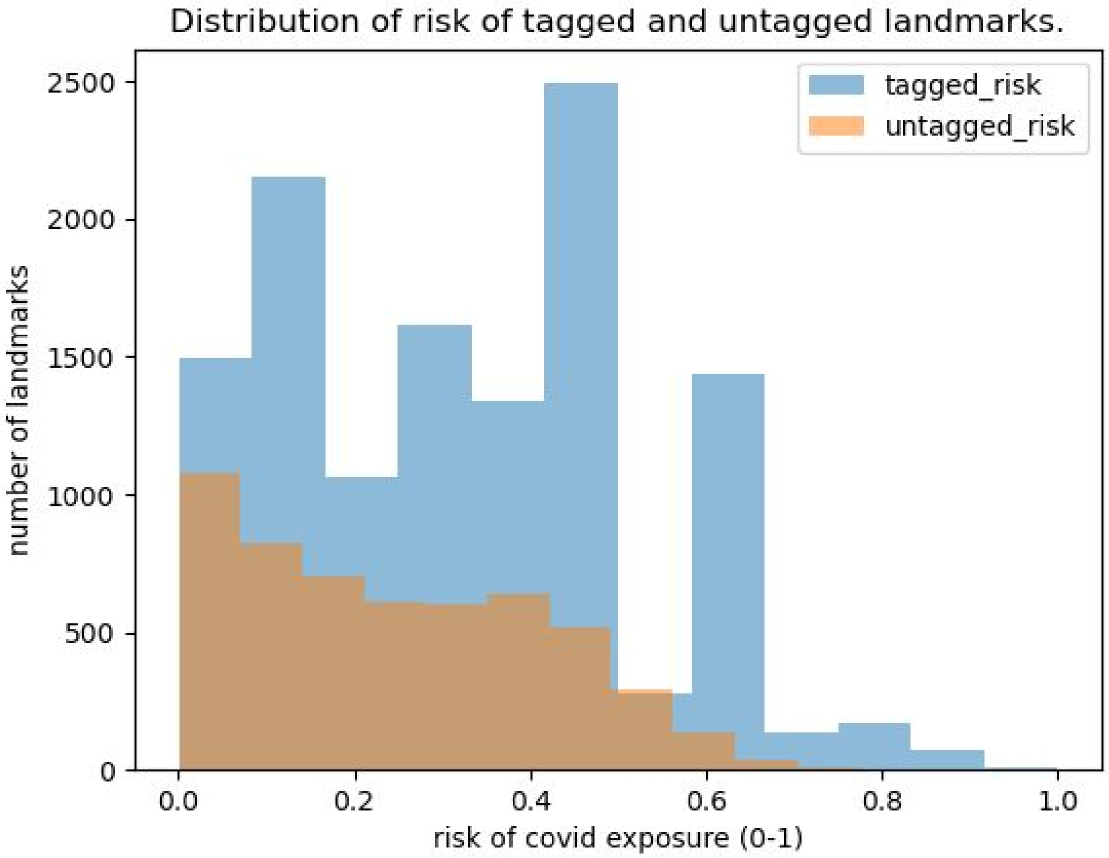


City of Vancouver, Province of British Columbia, Esri Canada, Esri, HERE, Garmin, INCREMENT P, USGS, METI/NAS...

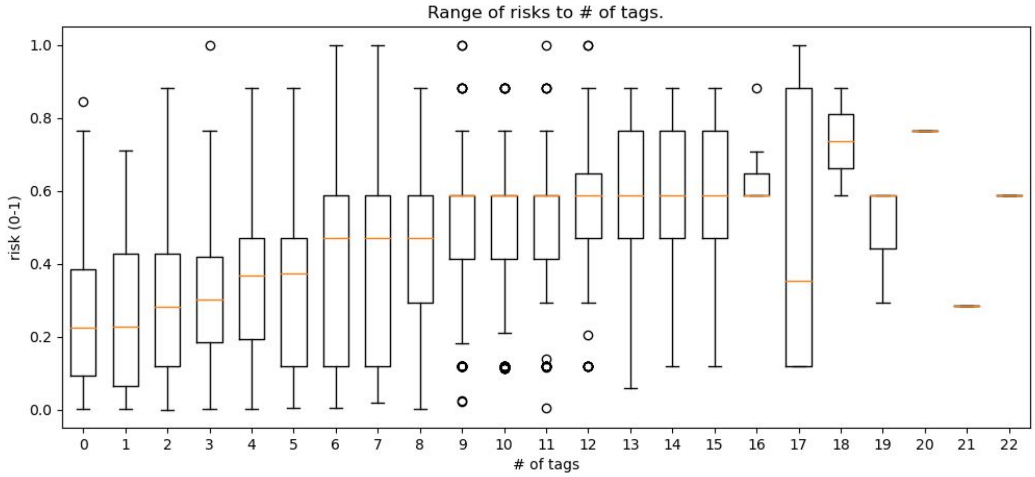
Results

It is difficult to make many conclusions from our analysis because of the assumptions made when processing, and the incompleteness and potential unreliability of the data. What we *can* do is compare our predictions to data we have access to: GPS locations of known COVID-19 cases. Our risk assessments are heavily right-skewed, so we take the squares of these values and compare their GPS data to known cases with a t-test, giving us p=0.123. LIkewise, a Mann-Whitney U test gives us a smaller (but still significant) p=0.079. With both p-values > 0.05 by a fair margin, we can say with some confidence that our prediction mode works.

As shown above, the distribution of untagged values which was generated from the tagged amenities have a similar distribution curve. However there is less noise in the untagged risks, as they derive their risk level based on their neighbours (as per k neighbours regression). However, the overall average of untagged areas still remains lower than those that are tagged. This is on purpose as many non risky landmarks such as benches, vending machines and fountains have few or no associated tags.



The below plot shows the overall distribution of risk for any given number of tags. Looking solely at the averages (the red lines), we can see a logarithmic curve that increases quickly before petering out due to lack of samples. This reflects how we decided to not use the number of tags (as many locations had an abundance of unrelated subtags) but instead focused on the primary tags and their overall associations. This meant that a gas station which may have tags for the fuel as well as the different fuel grades provided, is only counted once as a transportation tag. The wide variation of risk comes from some tags being unassociated with covid analysis all together, which can be seen in the data cleanup file.



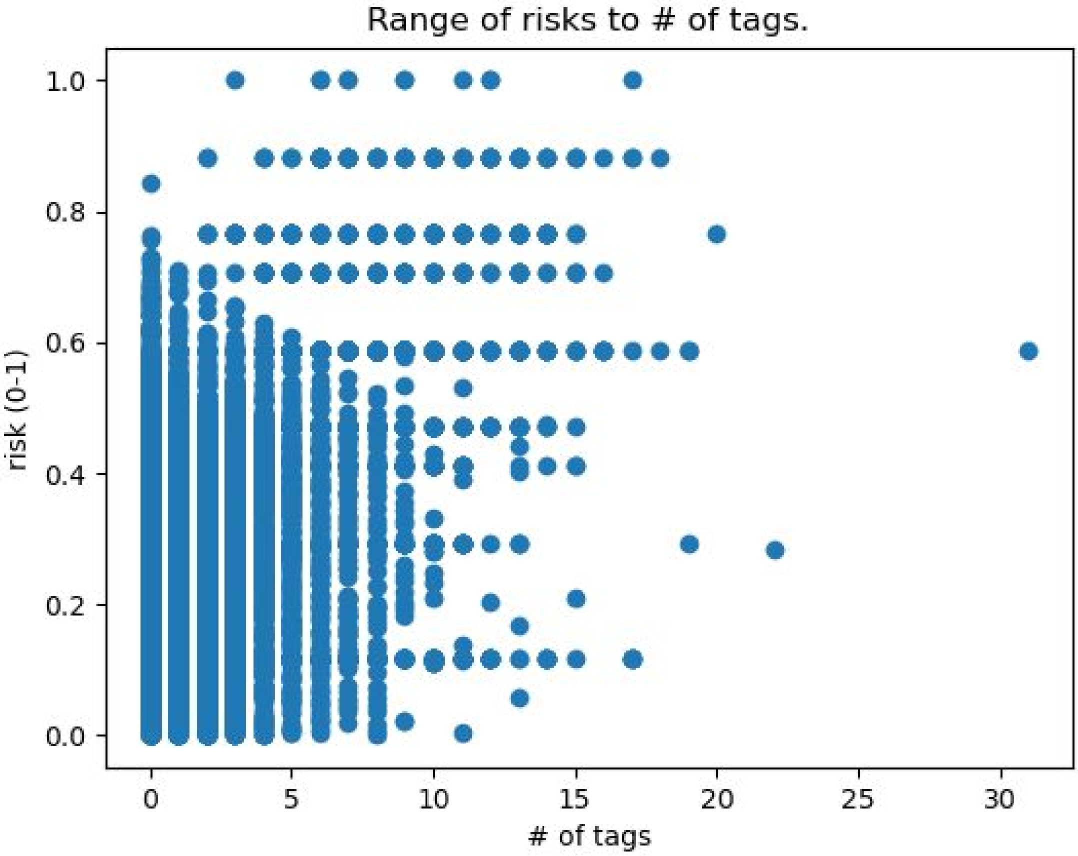
Limitations & Improvements

The obvious first choice for potential improvements is the human element in our data analysis: the manual classification of tagging information done at the beginning of our data cleaning is slow, prone to error, and may not be able to handle future data.

Our analysis also reveals that while tagging data was somewhat useful for determining the importance - and therefore risk - of a node, it was inefficient to use it for classification; the same information was reflected in the amenity type. If we were to repeat this kind of analysis with the same human categorization, the amenity type field would make this trivial.

Another oversight is the handling of "other" predictors of a node's human activity. The fifth tag category does a good enough job of this, but could easily incorporate unused fields like the timestamp - which indicates when a node was last updated. If a node being documented by Wikidata indicates its importance, that node being updated regularly by OSM might indicate the same.

Similarly, the number of tags may have been a factor in risk assessment. OSM allows access to tag heuristics, including which tags tend to coincide. This means tags like "fuel" and "fuel:octane89" can be counted as a single entity, and hierarchical tags like "medical:emergency" can default to the parent "medical". By using the number of *unique* tags in this way, we could have begun with a much better primitive risk assessment to train the KNN model.



Accomplishment Statements

Ryan

- Processed, evaluated, and formatted data, using efficient bulk procedures and categorical models to create a processing-friendly dataset.
- Applied statistical tests and modelling to identify relationships between tagging data and corresponding risk factors.
- Applied ML algorithm to dataset and compared alternative input values to expected real-world results to fine-tune model configuration.
- Filed and organized findings into cohesive project structure, and converted their findings to well-formatted report.
- Used ArcGIS to analyze and compare primitive assessment with machine-learning-produced models and produce sensible results.

Eric

- Evaluated data, analysing it and creating an algorithm to evaluate risk given location and tag information allowing us to quickly evaluate over 170,000 data points quickly and efficiently.
- Fixed data gaps using ML to account for locations with a lack of relevant data. Using kNeighbours regression to evaluate risk of locations based on their surroundings.
- Generated multiple graphs to show findings of the previously mentioned evaluation methods, and confirming that they would skew as expected.
- Created a new dataset of official real life covid public exposure events in the vancouver core which was used to evaluate the effectiveness of our risk model.
- Used statistical tests like the Mann-Whitney-U test and student T test to evaluate the kNeighbours model on accuracy of predicting high risk events vs. general risk areas.
- Used arcGIS to create a visual heatmap of vancouver and high to low risk covid exposure areas and further demonstrate its effectiveness by plotting the real life covid exposures reported by BCCDC.

Sources

2020 Vancouver Coastal Health. “Public Covid Exposures.” *COVID-19 Public Exposures*, 2020, www.vch.ca/covid-19/public-exposures.