AUTOMATIC TRAFFIC INCIDENT DETECTION

Author: Andres Namm

Supervisor: Dr Reynold Cheng

Introduction:

The goal of this research is to apply and develop a novel automatic incident detection System and test out the current state of the art method based on Topic Modeling [1] in Hong Kong using Probe Car GPS observations. This method is further developed to integrate speedpanel data to the training and testing dataset. In addition a well documented framework is developed for the preprocessing of data (Steps 1-6) Accurate incident detection is important mostly for 2 reasons: Routing and Acknowledging.

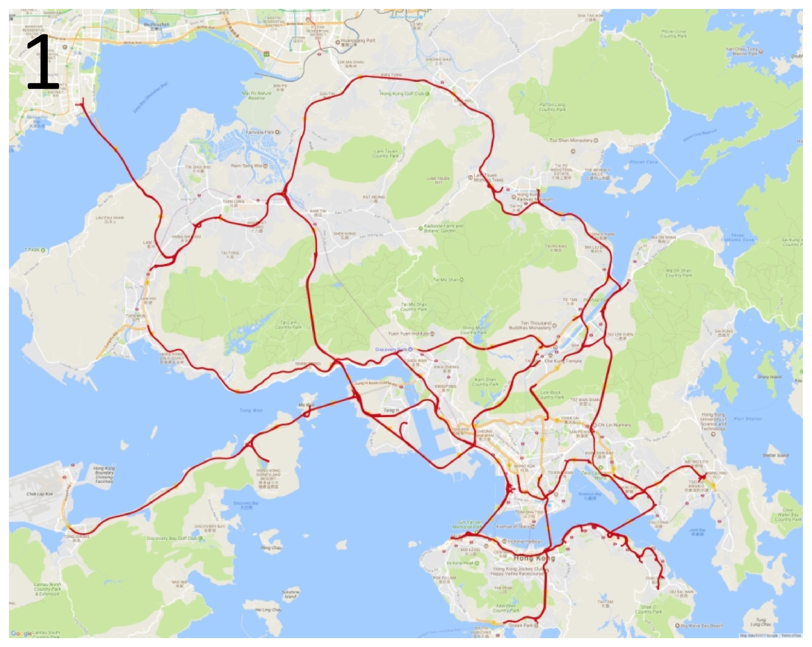
Step 1: Choosing the right Roads for Analysis

First step in this project was to choose the right map format to match GIS data – GPS observations, accident data and speedpanel data. The choice was made towards Open Street Map (OSM) format because it's open source and multiple tools have been built for parsing and analyzing it. The cornerstone of OSM are nodes and ways. Nodes carry the geographic coordinates in the OSM data model. A way only gets geometry via the nodes that are members of it. After choosing the map format, Hong Kong City map data was downloaded from Mapzen.com. Because it contained a lot of other map features like parks, buildings and geometry, filtering was necessary to keep the dataset only for roads.
Less important roads were eliminated from analysis as these roads had too few observations.

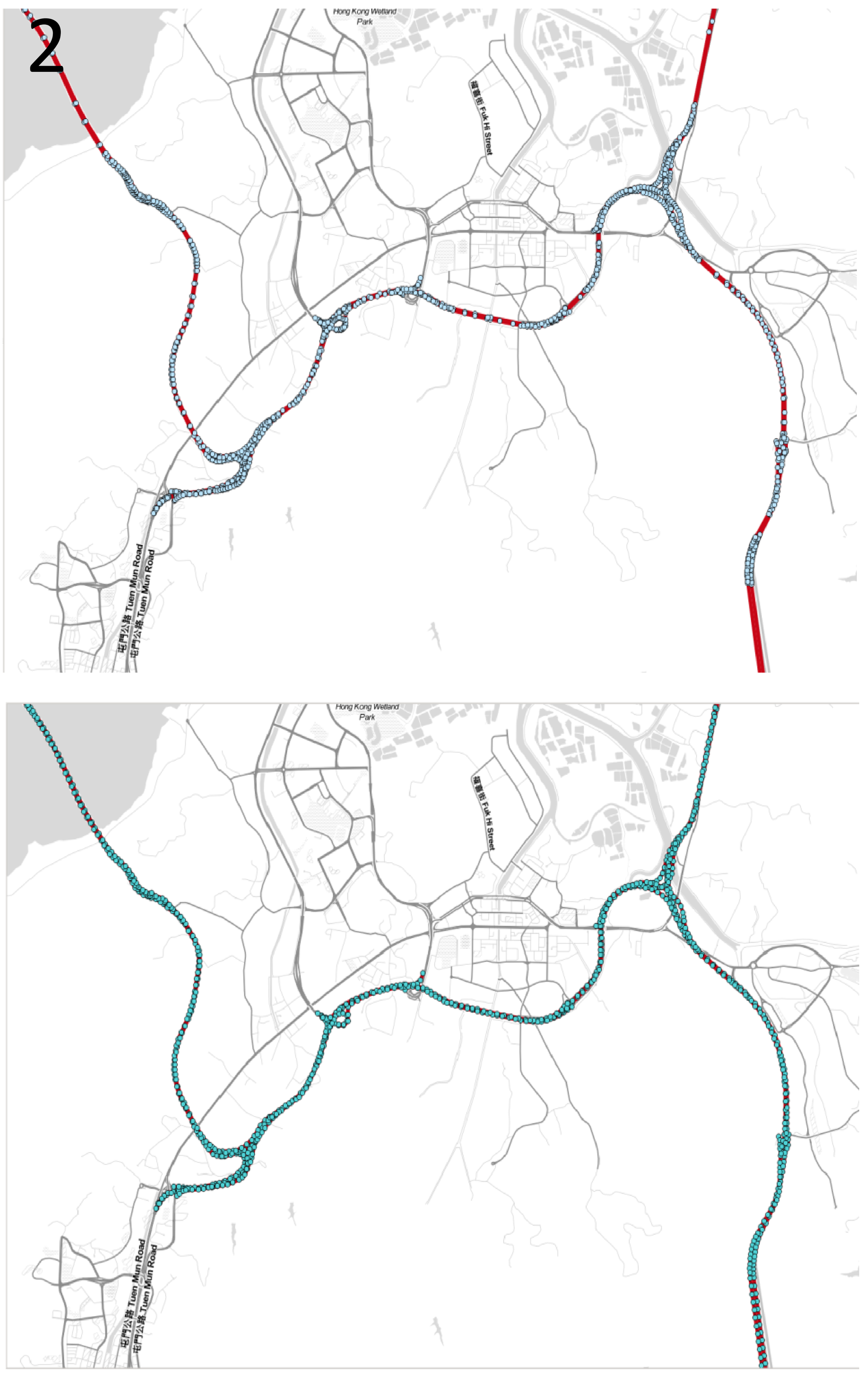
Step 2: Segmenting Roads

This step takes ways and nodes and modifies them so that in each way the distance between sequential nodes is inside a predefined interval (50m-100m). The ways both in the input and the output have only one direction. In principle, this step takes the map with Ways built from nodes(shown as points) like shown in Figure upper part and transforms these nodes so the nodes will be distributed like shown in Figure 2 bottom part. In later phases, the road straights between the nodes are going to to be the structural background for map-matching, model building and incident detection.

1



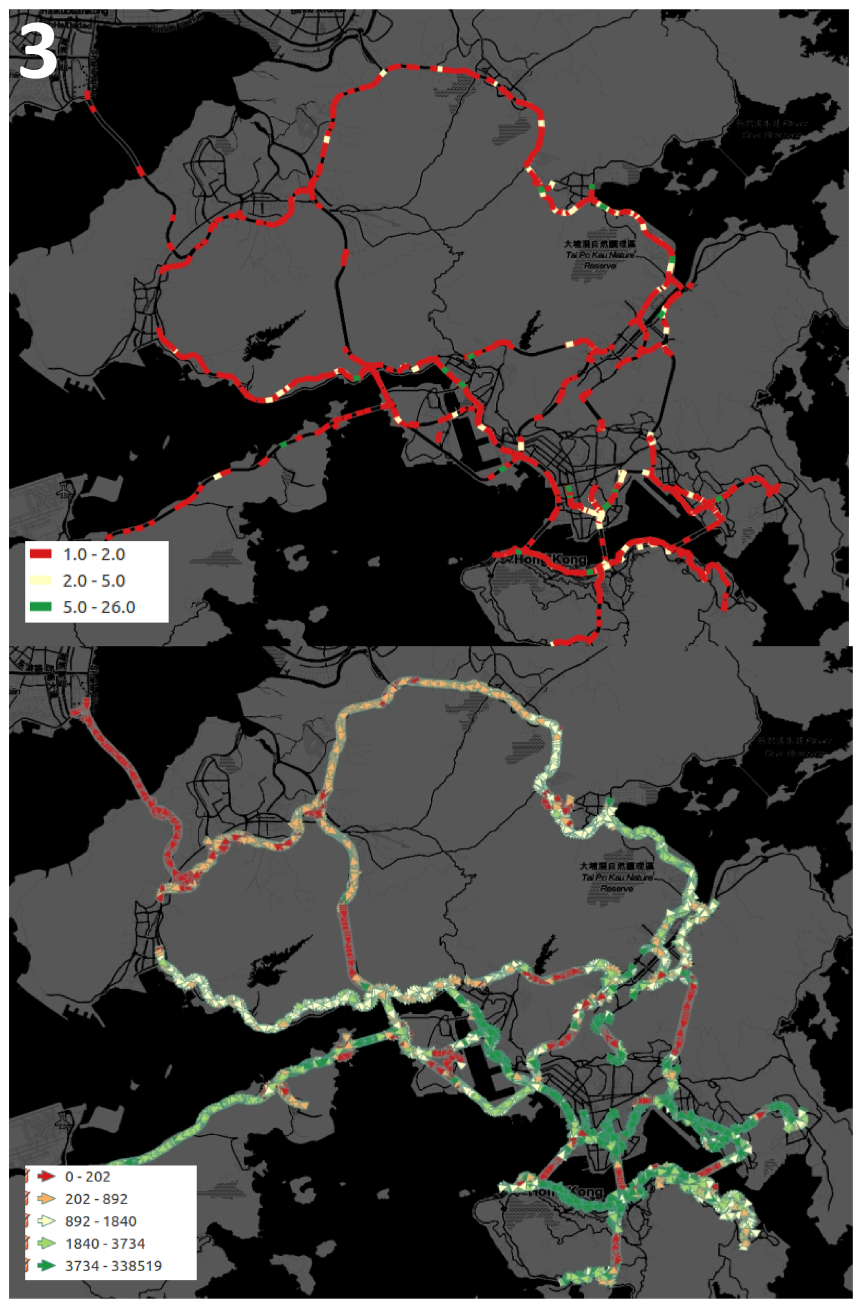
2



Step 3: Map Matching

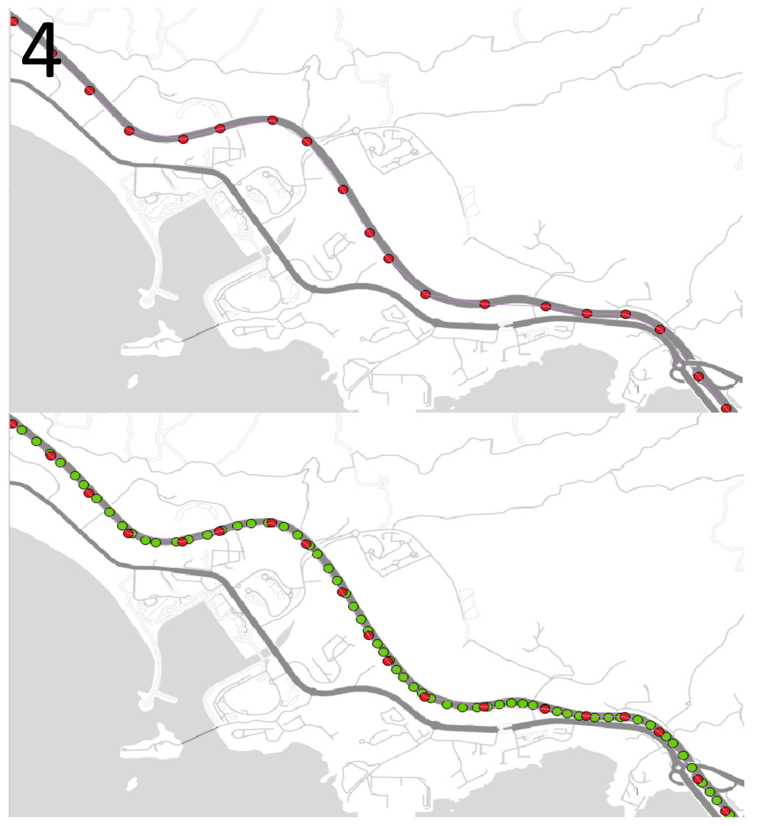
This part reads in the GPS records collected from Hong Kong Taxis in year 2010 and accidents information and matches them to specific road segments. This is a very important and very complicated part in the whole Incident Detection project because slight inaccuracies in matching the GPS data and Accident data can have negative results in the overall performance of the incident detection model. To perform mapmatching the segmented map is read in and Segment objects with list datastructures for Accidents and Records are created. Each generated segment will be assigned a direction. After the generation of segments a grid with 100 m x 100 m squares will be generated over the whole Hong Kong area. Each generated is assigned into 1 to many grid squares. After creation of the grid, the map is ready for matching GPS records and accidents to it. For each record that is matched, first its location inside the gird map is calculated. Afterwards the record is compared with all the segments inside the grid elements using direction and distance. The matching of accidents is similar however without direction comparison

3



Accidents Density (UP) GPS records density (BOTTOM)

4



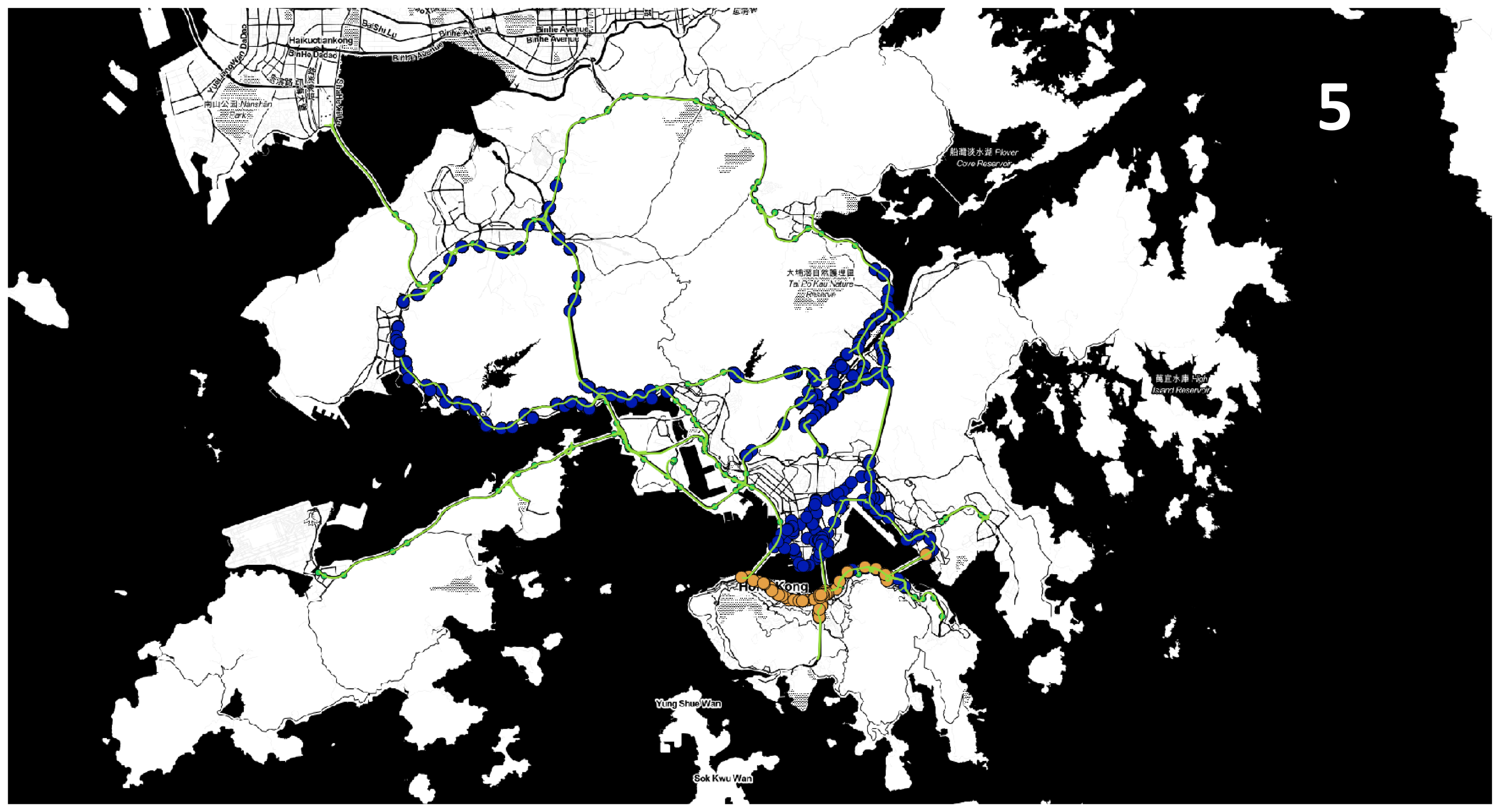
Step 4: Trajectory Generation & Interpolation

On faster roads, the taxis traveling speed is so high that GPS readings which take place in 30 second time intervals cannot create an observation in each sequential segment . On the other hand, on slower routes with heavy congestion, taxis can generate multiple observations inside 1 segment. In this phase, trajectory generation is done for each device. If inside a trajectory, there are some segments skipped as shown in the upper part of Figure 4, interpolated records are going to be created between the sequential segments. At the same time, 1 trajectory can only contribute an observation to 1 segment.

Step 5: Integrating Speedpanel (SP) Data

One of the key contributions of this Dissertation was to integrate speedpanel data with GPS data, The speedpanel data was obtained from the Hong Kong Government. It consists of average speed observations in 2 minute time intervals. As the taxi dataset was available for year 2010, only speedpanels that were installed in year 2010 have been integrated as a data source to this project. The speedpanels that are available before year 2013 are shown in color beige. The speedpanels that are available after year 2013 are shown in color blue and beige. The integration of speedpanel data has so far beend rather straigthforwards. First speedpanel pairs were transformed into ways in the OSM format. From ways they were turned into segments. After that, the segments generated in based on Hong Kong map in the segmentation step nr 2 were matched to the speedpanel based segments. For this, exactly the same function was used as in mapmatching and when assessing closeness, it took into account the absolute distance between segments start and end points and the segments directions. In the end 62 speedpanels were matched with 262 segments. After automatic map matching, manual filtering was done to eliminate false matches.

5



6

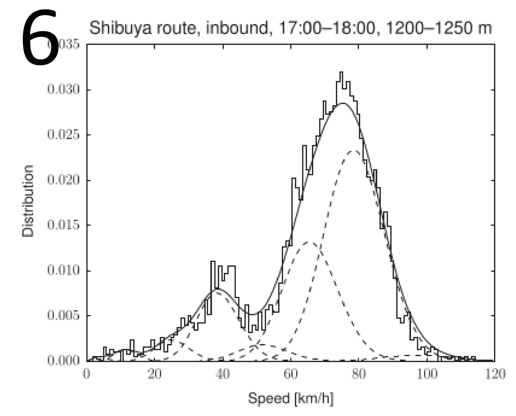


Fig. 5. Histogram of the speed of probe cars in a segment and the estimated Poisson mixture.

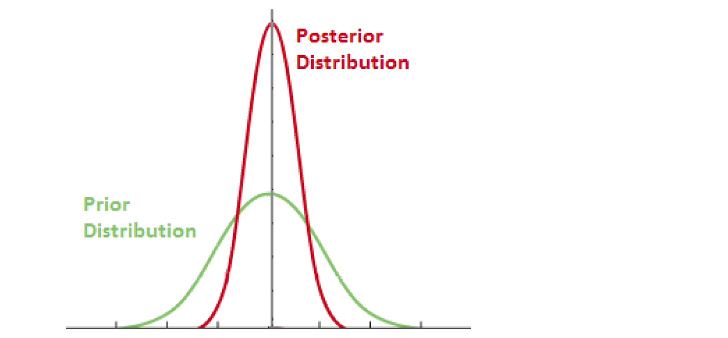
Step 6: Training The Model

The speed observations in each segment are assumed to be genertated by a mixture model of K Poisson distributions. The model trains:

1. K global Poisson distribution means
2. For each road segment multinomial distribution parametes over the K global distributions

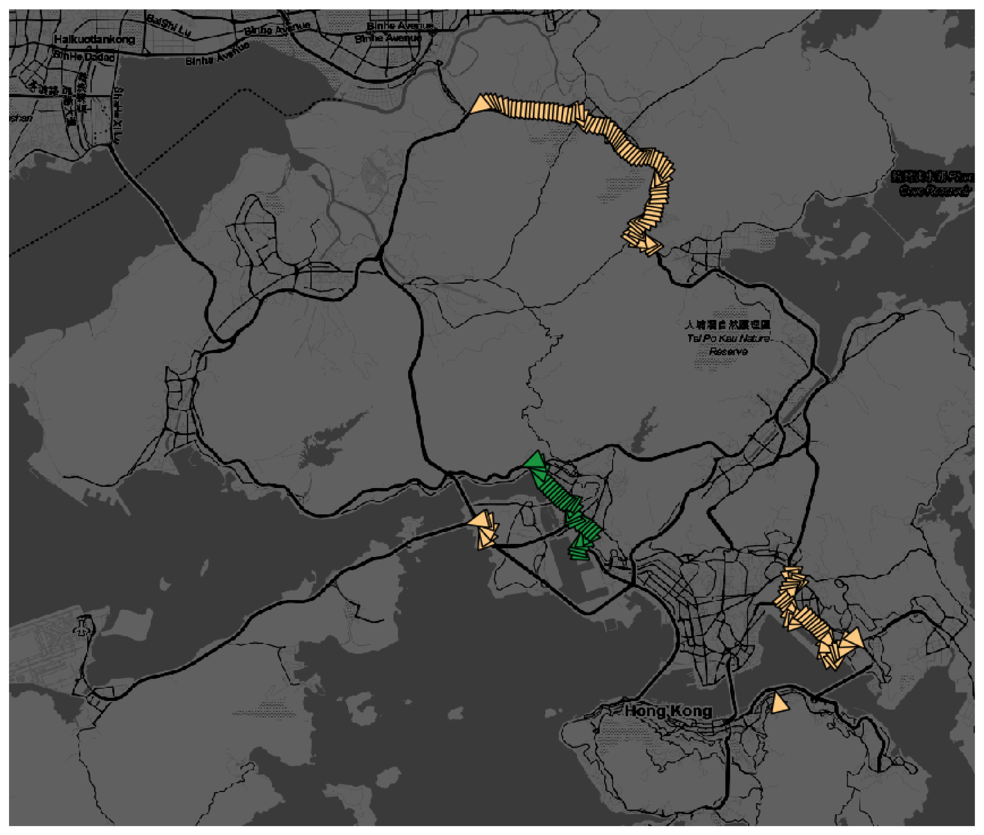
To train the model, EM algorithm was used. The basic training can be compared to soft clustering.

7



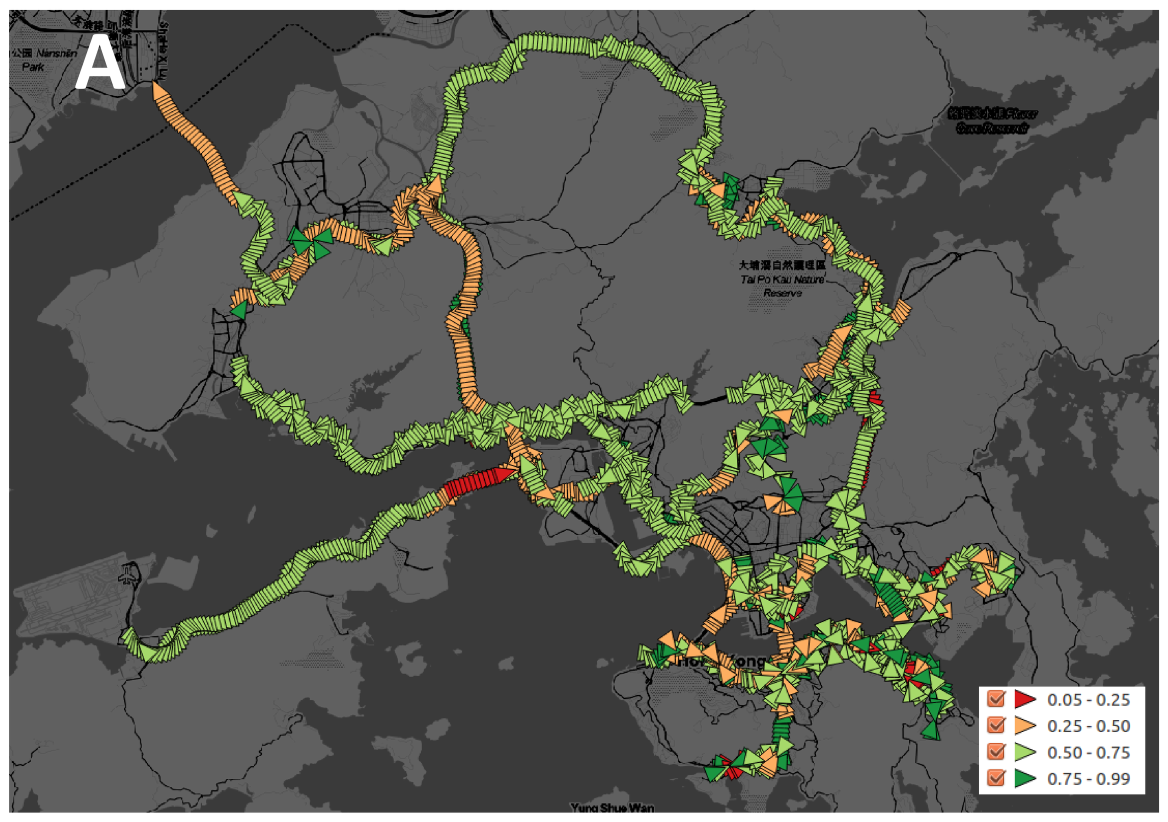
Step 7: Incident Detection

After the parameters have been inferred globally and for each of the segments new observations will be compared with the historical data based model. To do that Bayesian inference is used again inside a specific time window to build a new distribution for the specific local segments using again the global traffic state parameters. Then the new posterior distribution is compared with the old distribution using weighted KL divergence.



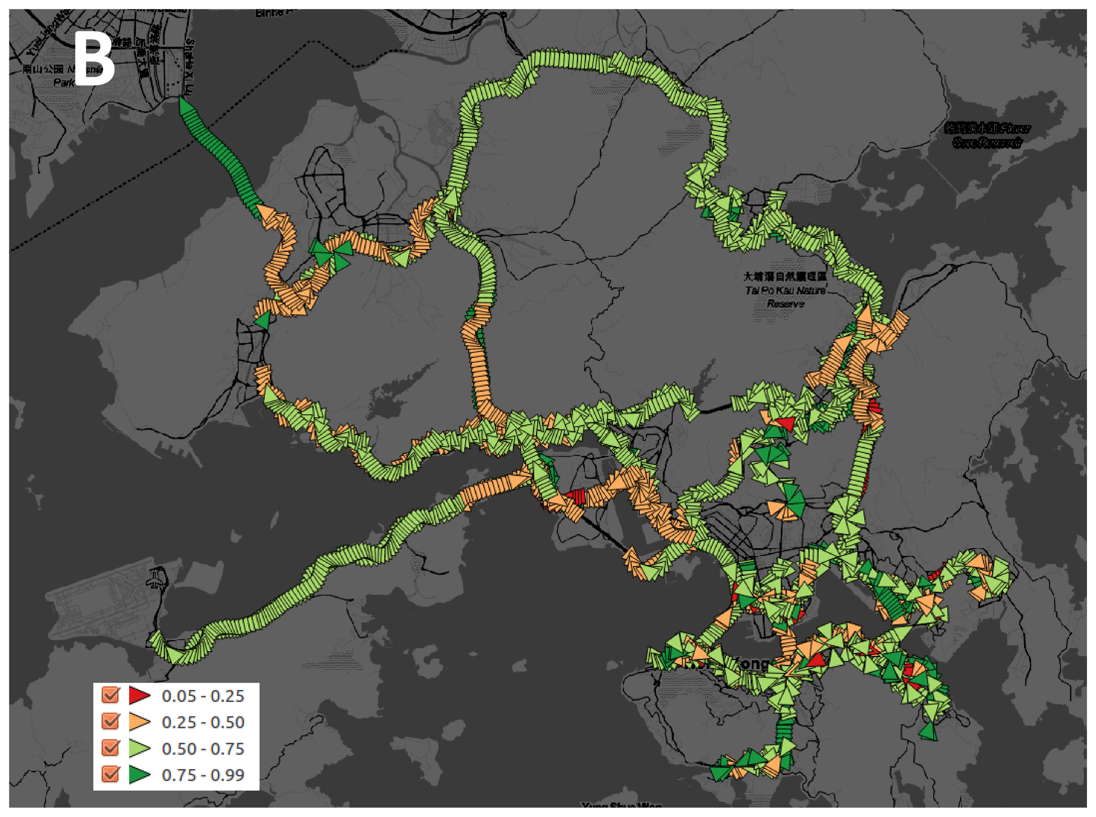
Initial project results (GPS)

A

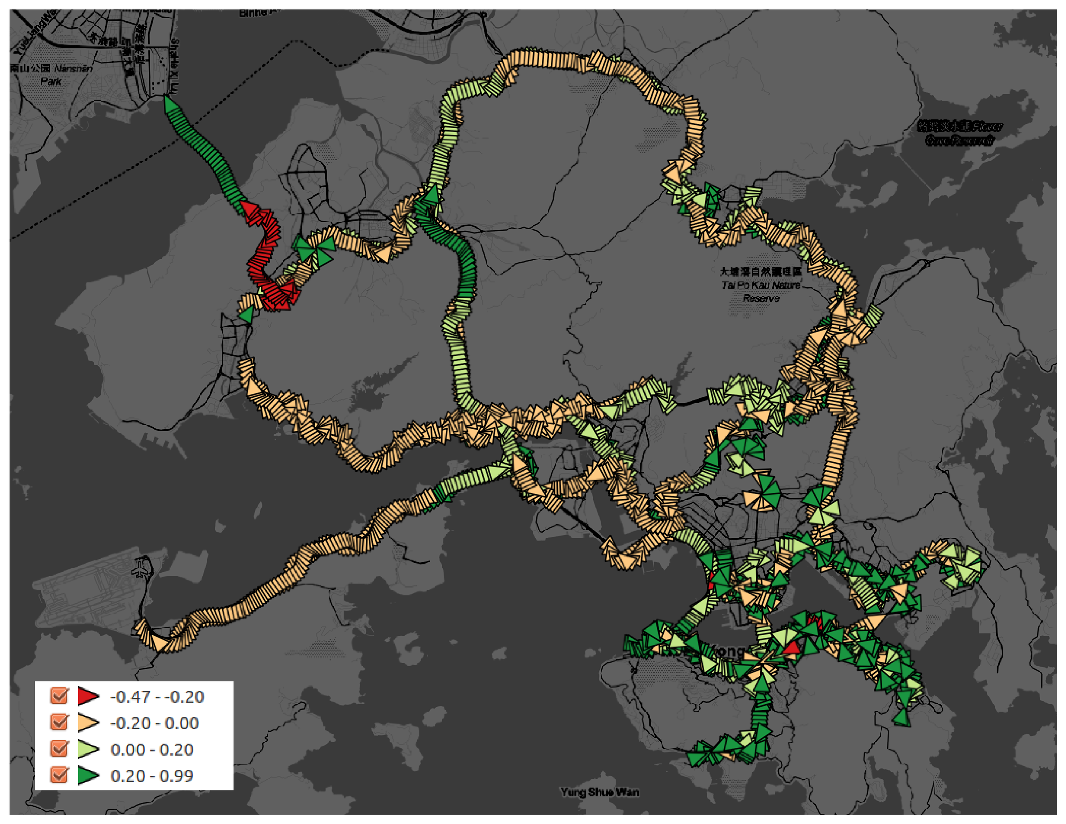


Redesigned Project results (GPS)

B



Redesigned Project results (GPS+SP)



B results – A results

Experiments

To evaluate the performance of the algorithm, a 4-fold cross-validation experiment was designed.The dataset was partitioned into 53 weeks, 13 of which was considered as testing data. Detection Rate (DR) and False Alarm Rate(FAR) was measured and plotted the receiver-operating curve (ROC) to review the result for each road. We also calculated the area under the ROC curve (AUC) for each road. The results) for all experiments for different folds varied a lot within a segment. This was due to different distributions of traffic accidents in each fold. To solve this issue, the AUC average was calculated using weights which were determined by the number of accidents in a specific fold. For testing, each observation generated by GPS or speedpanel data was labeled as belonging to an accident or not.

Experiment using only GPS data

The average AUC result distribution received using only GPS data for the older project can be seen on the leftmost figure. After redesigning the project, incident detection in Hong Kong improved a lot. When not dismissing outliers, the AUC for 441 ways is 0.5714. The distribution of AUC values on roads over Hong Kong can be seen on Figure A.

Experiment using GPS and Speedpanel data

Adding speedpanel data resulted in a slightly higher overall result of 0.2%. The ways which had both speedpanels and GPS data average was 0.6. Overally however its clear that addition of speedpanels improved the result in the city area as can be seen on the leftmost Figure comparing the results.

Analysis

The focus to the development of data preprocessing pipeline payed off significantly. Alltogether 3 mistakes were detected in the project under development and fixing them improved the accuracy and the results drastically . Based on the observations of the all of experiments (1-4) it seems the overall amount of data as an input to the model plays a strong role in producing the results. The worst result came from using only the speedpanel dataset which due to lack of interpolation could produce the least amount of data to segments. The second worst results were received with the dataset using only GPS observations assigned to segments longer than 1.5 km and belonging in trajectories longer than 500 m. This filtering done in experiment 3 due to filtering produced less data than the methodology in experiment 1 . Using GPS data with interpolation without the filtering produced the second best result and adding speedpanel data to it produced the best result.
The averages of the best result in Hong Kong are still 0.2 less compared to the AUC results received in [1] This based on the experiments (1-4) conducted on Hong Kong dataset likely results from roughly 6 times smaller dataset. The results are however not that much worse which the first experiments lead to believe.Comparing these results should also not be taken too seriously as Et. al Kinoshita article considered roadworks as traffic incidents as well. Due to inability to get access the roadworks database in Hong Kong, this paper used traffic accidents recorded by Hong Kong Police for testing the model. This likely introduced additional false results.

References

1. A. Kinoshita, A. Takasu, and J. Adachi. Real-time traffic incident detection using a probabilistic topic model, Information Systems, 54:169--188, 2015
2. OSM wiki - http://wiki.openstreetmap.org/wiki/Key:highway
3. QGIS – software for GIS data analysing and visualization