UTILIZING MACHINE LEARNING TO REDUCE THE PROCESSING TIME FOR GPS TRUCK BASED MAP-MATCHING

Kevin Gingerich, University of Windsor Cross-Border Institute
Hanna Maoh, University of Windsor Cross-Border Institute
Proceedings of the 52nd Annual Conference
Canadian Transportation Research Forum

Introduction

Map-matching is a process where point events, often derived from GPS pings generated by a moving vehicle, are converted into a line event that identifies the route of the vehicle. As evidenced by Figure 1, the actual route taken by a vehicle is inherently more complicated than a straight line connecting the two pings (points). A better approach is to identify the most probable route taken by the vehicle. Determining the route that is generated from map-matching provides a dual purpose: a more realistic depiction of the travel behavior of vehicles and a connection between the trip and road links that were utilized.

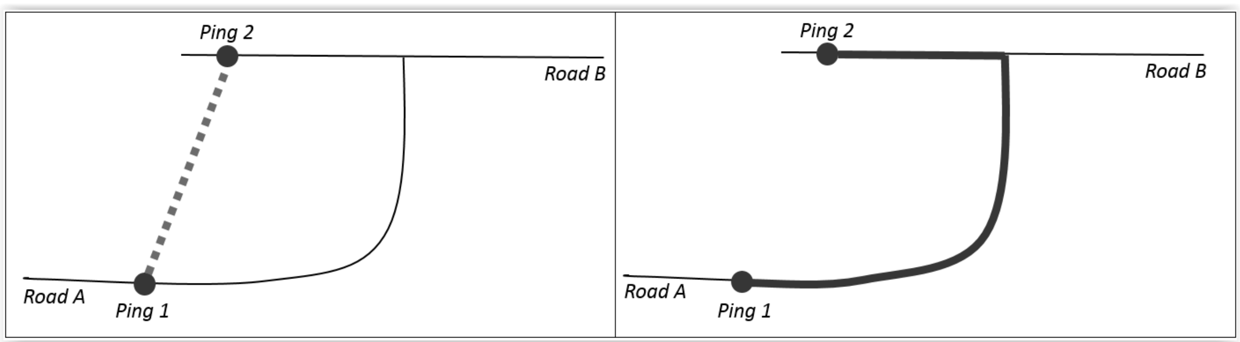

Figure 1. Connecting GPS pings with a straight line (left) and map-matched route (right)

Drivers often benefit from the map-matching process in real-time, as their GPS navigation device determines the current road that they are traversing and subsequently predicts the best route to their destination. For transportation analysts, off-line map-matching can be utilized to determine the routes travelled by vehicles in the past to create a rich dataset of vehicle movements. The goal of the map-matching is to estimate these routes, with the best information available, as continuous lines (i.e. no breaks where the vehicle jumps from one location to another) that preferably take into account the network connectivity of the road network. Processing a large dataset of trips can therefore become computationally intensive due to the quality of the data such as spatial location errors and uneven/large intervals between pings. The processing time required to map-match thousands or millions of individual routes can therefore become prohibitive.

This short paper presents a method designed to decrease the overall computational processing time of typical map-matching. This approach utilizes machine learning to identify the patterns in the volume of vehicles on each road link and apply this information to unprocessed data. The next section of the paper will describe the dataset of GPS pings used in this analysis. Next, the results of map-matching using traditional GIS-based techniques will be discussed. The resulting dataset will then be used as an input to train an initial artificial neural network (ANN) model that will eventually be capable of predicting the map-matching outcomes at a much faster speed.

Data

This analysis utilizes a subset of processed GPS data pertaining to Canadian owned truck movements occurring in the first week of March, 2016 (March 1 – March 5). The GPS dataset represents points (pings) that contain identifier fields for the truck and carrier, the latitude and longitude of the truck, and the corresponding timestamp. As such, the GPS pings can be sorted by time for each truck to observe the movements of the vehicle.

From the GPS data, trips are identified based on the processing approach discussed in Gingerich et al., (2016), resulting in 200,140 trips belonging to 477 carriers and 26,834 trucks travelling across Canada and the U.S. Since this paper focuses on non-local truck movements within Ontario, the processing identifies inter-regional movements where the truck travelled between census division zones. As a result, the trips are heavily focused on major roads such as the 400 series highways. The trips are originally associated with general information such as the starting location, ending location, industry, and border crossing information (if the trip was international). However, more detailed information on the route itself is not determined at this stage.

Map-matching

Detailed information on the trip routes traversed by trucks observed from the GPS data requires map-matching. This process converts the point based GPS pings to a line based route for each trip. Since this data is an ex-post analysis, offline-map matching is utilized. Quddus (2007) classifies map-matching algorithms into four categories: geometric, topological, probabilistic, and advanced.

Geometric map-matching is the simplest approach where the shape of the road network is the only feature utilized and the GPS point is assigned to the nearest road segment. Several issues arise from this approach. First, GPS pings are subject to spatial errors that can result from an insufficient number of connected satellites. Moreover, multipath errors are relatively commonplace in urban canyons where the signal can be reflected indirectly off other objects such as buildings. A previous analysis of this GPS data for March, 2013 found that the average lateral error (perpendicular to the road trajectory) was 15.4 m and 27.7 m at the Ambassador Bridge and a section of the Highway 401 corridor, respectively. The 95 percentile lateral error for these two locations was 89.6 m and 144.9 m. As a result of spatial errors, the GPS pings do not occur directly on the digitized road link and may end up closer to a different road segment.

In addition to spatial errors, the GPS data in this analysis exhibit fairly large distance intervals between successive pings where a 5 to 15 minute interval is typical but may also be exceeded. The relatively large gap between pings results in a set of road segments along the route that may not have any pings near them. As a result, simply matching pings to the nearest link will produce large gaps in the total route. To overcome this issue and provide a continuous route, the topological information of the road can be used to help determine the path that the vehicle likely utilized between the GPS pings. This information provides the connections between the road segments (i.e. which routes a vehicle can take at each node/junction).

The topological approach is therefore utilized here due to its ability to provide estimates of route choice between large gaps in GPS pings while still remaining relatively simple (compared to the probabilistic and advanced map-matching models). The software utilized for this task was created by Ron Dalumpines and Darren Scott at McMaster University. The custom application utilizes the network analyst extension of ESRI's ArcGIS software as the conduit for topological connections on the road network and the calculation of shortest path trip assignments. The application creates a feasible boundary corridor for potential routes based on the GPS pings of the trip. The route is then determined from a travel time based shortest path. More details on the map-matching application can be found in Dalumpines and Scott (2011).

The application designed by Dalumpines and Scott provides very good results after some calibration to account for longer intervals between pings in the GPS dataset used here. However, the map-matched results come at a cost of processing time with an average of one minute per trip in both our processing results and those stated by Dalumpines and Scott (2011). In comparison between the two, the former GPS dataset is composed of longer trips that would normally take longer to process, but benefits from more recent computer technology (Intel i7 3.4 GHz, 4 core processor and 32 GB of RAM). While a one minute processing time per trip is not a large concern for small datasets, the processing time adds up considerably when processing larger datasets. For the 200,140 identified trips, this would result in 139 days total. To reduce this time, the data is split into 1,000 trip sections with 6 instances running in parallel on a single computer (and multiple computers at the same time). Even so, the map-matching when taking into consideration pre-processing and post-processing of the data takes approximately 2 months for the one week dataset.

The results from the map-matching process provide individual routes pertaining to each trip. In addition, the individual trips can be aggregated to provide total volumes of trips on each road segment. The week 1 aggregate volumes are shown in Figure 2 with Southern Ontario focused in the center.

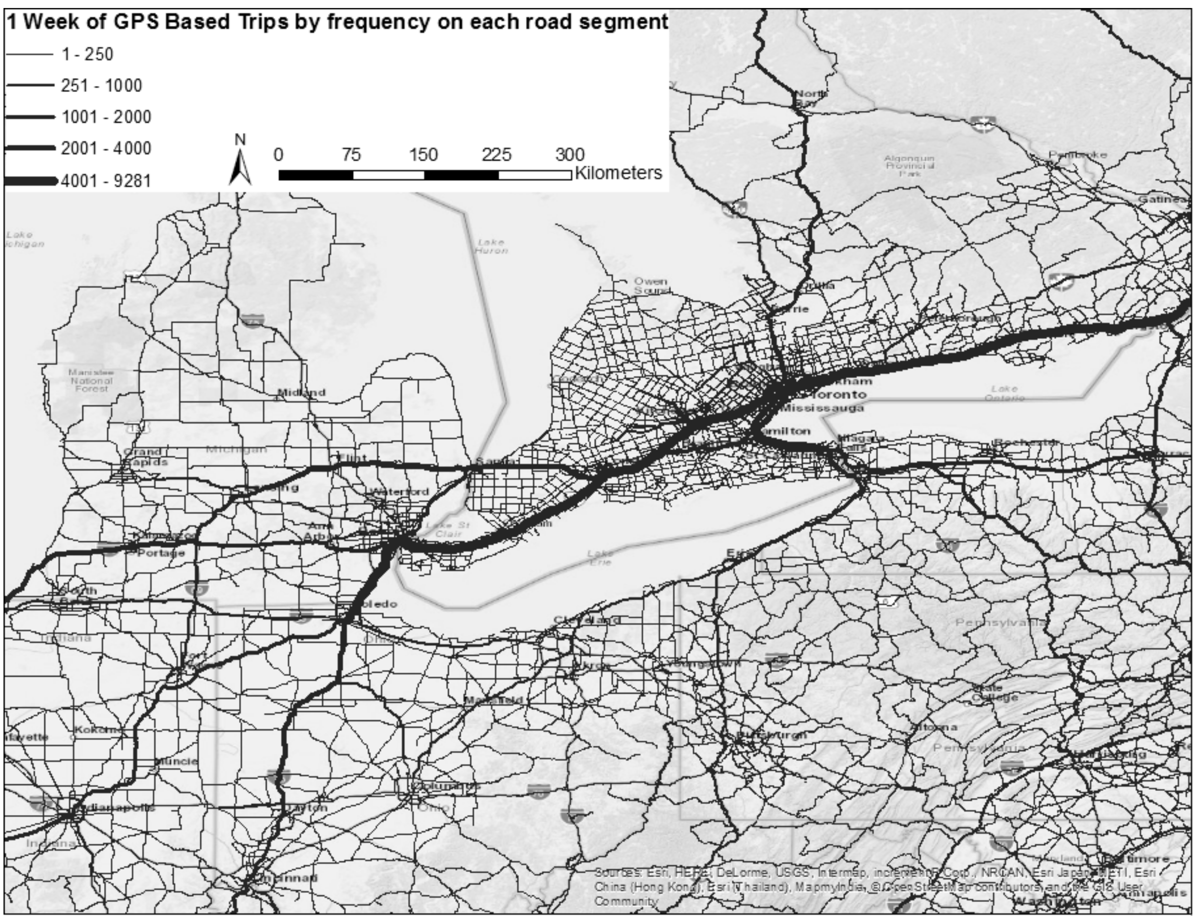

Figure 2. Trip volumes derived from map-matching

Artificial Neural Network and Initial Results

Due to the extensive processing time necessary to create the map-matched truck route, an alternative that can be performed on future datasets more efficiently is desirable. The objective is to predict the volume of truck traffic on road links using data available before the map-matching begins (bypassing the map-matching step altogether). An artificial neural network was selected as a feasible method in this case due to its predictive capabilities for large, complex datasets with substantial non-linear relationships. More detailed information on neural networks for predictions in transportation can be found in Moniruzzaman et al. (2016) that describes a short-term border crossing prediction model.

The ANN model requires dependent and independent variables similar to regression models. In this case, the dependent variable is the numeric volume of traffic on each road link (shown in Figure 2). The road network includes 84,423 segments that represent individual rows in the input table. The primary explanatory variable is the density of GPS pings that was created using a kernel density surface with a radius of 250 meters. The density was joined to the road links by averaging the values obtained from points created along the road segment at equal interval distances of 250 meters. The expectation is that a higher density of pings along a road segment will be a strong predictor of a larger traffic volume. Other explanatory variables available for the model include the standard deviation of the density on a road link, the category of road network (where 1 composes the larger highways, 2 represent smaller highways, and 4 represents arterial roads), road segment lengths, and indicator variables for one-way roads, highways, and ramps.

A linear regression was first utilized to confirm the expected relationships. The ANN model was then performed with a generalized regression neural net using Palisade Neural Tools software and default settings. A 20% hold-out dataset was utilized to test the performance of the model for prediction purposes. Unfortunately, the initial ANN model did not perform well, resulting in a prediction failure rate (based on a 30% tolerance) of 89%. On a positive note, the correlation between the observed and predicted volumes for the hold-out dataset is 75%. In addition, the relative weight of variables in the model confirmed that the GPS ping density was the most important variable (43%) followed by the density standard deviation (37%), segment length (11%), and class of road (6%).

The initial results and residual plots primarily demonstrated that numerous road links with a very low volume of traffic were predicted to have much higher traffic volumes. While an analysis of this result is still in progress, it may be occurring due to spillover effects from the GPS ping density surface on nearby (or perpendicular) road segments. Future work analyzing the results of the initial model and improving the prediction capabilities will be conducted to achieve suitable predictions.

References

Gingerich, K., Maoh, H., Anderson, W., 2016. Characterization of international origin-destination truck movements across major Ontario-Michigan border crossings, *Transportation Research Record*, vol. 2547, pp. 1-10.

Dalumpines, R., Scott, D., 2011. GIS-based map-matching: develompnet and demonstration of a postprocessing map-matching algorithm for transportation research, *Lecture Notes in Geoinformation and Cartography 1: Advancing Geoinformation Science for a Changing World*, vol. 1, pp. 101-120.

Moniruzzaman, M., Maoh, H., Anderson, W., 2016. Short-term prediction of border crossing time and traffic volume for commercial trucks: A case study for the Ambassador Bridge, *Transportation Research Part C*, vol. 63, pp. 182-194.

Quddus, M., Ochieng, W., Noland, R., 2007. Current map-matching algorithms for transport applications: State-of-the-art and future research directions, *Transportation Research Part C*, vol. 15, pp. 312-328.