UNKNOWN TO KNOWN: PREDICTING TRUCK GPS COMMODITY USING MACHINE LEARNING

Mausam Duggal, WSP | Bryce Sharman, WSP | Rick Donnelly, WSP
Matthew Roorda, University of Toronto
Sundar Damodaran, MTO | Shan Sureshan, MTO

Keywords: Freight transportation, GPS, machine learning, data fusion

1.0 Introduction

To satisfy data needs for freight modelling and planning, the Ministry of Transportation Ontario (MTO) conducts the Commercial Vehicle Survey (CVS) approximately every five years. The last one, conducted in 2012, collected information on the route, commodity transported, and vehicle configuration for ~ 45,000 truck tours surveyed across 220 sites located in the Province. Since 2014, MTO has also purchased anonymized truck GPS data that show truck travel histories for trucks in Ontario. The GPS data is the most extensive depiction of truck movements on the Province's road network, although its market share is unknown and the vehicle characteristics and commodity carried are unobserved.

MTO and WSP have completed the development of a CVS-GPS processing package, whose objective is to fuse the information from the CVS, truck GPS and other data sources (such as truck counts) to better understand truck travel on Ontario roads. The steps of this package include:

1. Process raw GPS data into travel diary of stops, trips and tours.
2. Commodity prediction.
3. Routing onto assignable network.
4. Expansion of truck traffic to match available counts, and
5. Web-based visualization showing origin-destination totals and link-level truck volumes.

This paper focuses on step 2 of this package, namely commodity prediction of GPS-observed truck tours. The goal is to construct a succinct ML (machine learning) model using the CVS that can predict the commodity carried in the truck as a function of attribute variables such as geography, tour distance, nearby firms to tour stops. Once constructed, the ML model would be applied on the GPS data to predict a commodity for the entire truck tour.

2.0 Input data

MTO Commercial Vehicle Survey (CVS)
The 2012 CVS, intercepts trucks at data collection sites across the Province of Ontario highway network. The CVS, on average, achieves a 4% sample of commercial vehicles passing each data collection site, which represents 10% effective sampling rate after considering routing of the sampled vehicles that pass multiple station sites. Significant variation is seen across data collection sites. Each truck tour (consisting of the commodity origin, intermediate stops, commodity destination, truck configuration, tour length, etc.) is tagged with a single SCTG (Standard Classification of Transported Goods) commodity classification, which were aggregated to Canadian Freight Analysis Framework (CFAF) commodity groups for estimation.[i]

Pitney Bowes Firm Data
The 2012 Pitney Bowes firm data was obtained by the MTO during the development of the Province of Ontario's passenger and freight forecasting / travel demand model (TRESO). This dataset includes information on all the firms (390,162) in the Province, including its North American Industry Classification System (NAICS)[ii], size, and latitude and longitude.

Socioeconomic Data
The MTO has developed a detailed estimates and forecasts of GDP (gross domestic product) by NAICS2 industry classifications for each CSD in the province.

3.0 Methodology

Shipments made by large trucks generally occur between firms (although pickups/deliveries to households are possible). The CVS did not collect any information on shipping or receiving firms. However, this is a significant source of information for the learner and as such including firm information was deemed to be critical for model estimation. Attempts to match CVS shipments with individual firms were generally not successful due to data resolution, hence a simulated annealing approach was developed that calculated employment and probability scores by industry. The total industry score was summed over all stops in the tour. This procedure is summarized in Figure 1. Table 1 shows the final set of attributes used for training the ML model.

Two learners, Deep Neural Networks (DNN) and Gradient Boosting Machines (GBM), were initially tested but the GBM was selected given the number of categorical attributes. As per convention, the total commodity truck tours in the CVS were split into training and out-of-box (OOB) data using a stratified sampling technique on the CFAF groups. Given the relatively small dataset being used in the development of the ML, *k-fold cross validation*[iii] (5 folds) was used to overcome problems associated with overfitting and bias.

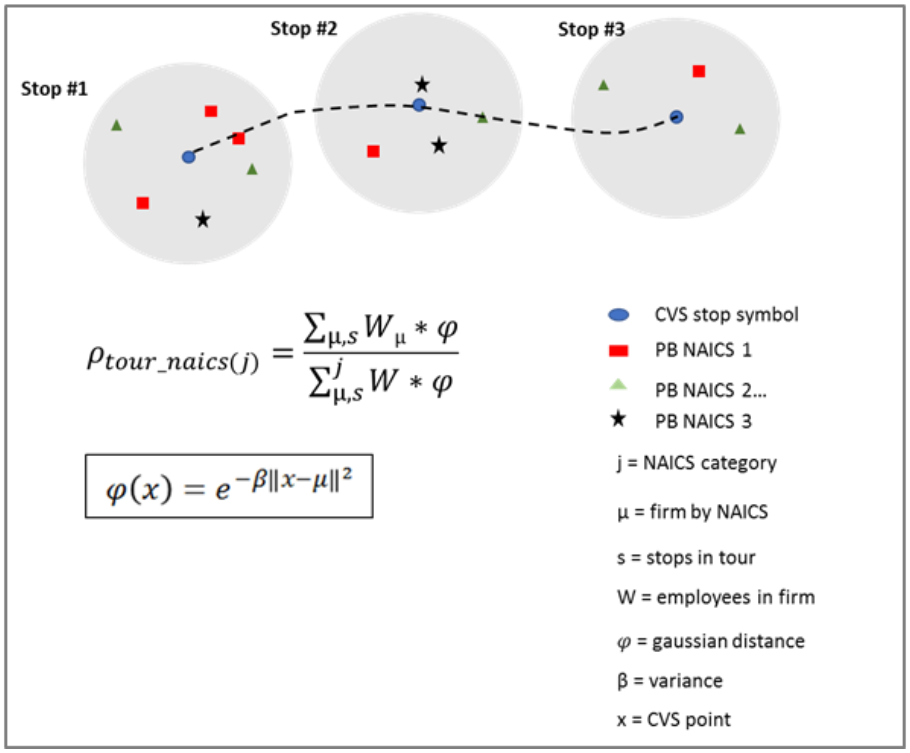
Simulated Annealing Framework

Initialize - Generate buffer around first CVS stop:
250m: urban area (GGH and Ottawa region);
500m: all others;
Starting variance: user defined (5th root of buffer: default);

While (no Pitney Bowes firm found in buffer):
Increase buffer size by 250m radius;
Compute Gaussian Distance from CVS point to each Pitney Bowes firm segmented by NAICS;
Compute weighted probability for every NAICS category in buffer;
Collect number of employees by each NAICS in buffer;

If (buffer radius reaches 4000m with no Pitney Bowes firm):
Exit
Return to next stop in tour

Gaussian Distance and Probability Calculations



$$\rho_{tour\_naics(j)} = \frac{\sum_{\mu,s} W_{\mu} * \varphi}{\sum_{\mu,s}^{j} W * \varphi}$$

$$\varphi(x) = e^{-\beta \|x-\mu\|^2}$$

CVS stop symbol
PB NAICS 1
PB NAICS 2...
PB NAICS 3
j = NAICS category
μ = firm by NAICS
s = stops in tour
W = employees in firm
φ = gaussian distance
β = variance
x = CVS point

Figure 1. Industry Probability Scoring

Table 1. Final Attribute Set in ML model

Attribute	Cols	Source	Description
Total tours in estimation data			31,491
CFAF group	1	CVS	11 groups
Tour length	1	CVS	km
Ln(tour length)	1	CVS	Natural log of tour length
Weekday flag	1	CVS	1 for weekday; 0 otherwise
First_csd	1	CVS	CSD of the first leg of the tour
Last_csd	1	CVS	CSD of the last leg of the tour
Prob(naics)	22	SA + PB	Weighted probability for each NAICS across the entire tour found in tour stop buffers
Emp(naics)	22	SA + PB	Number of employees for each NAICS across the entire tour found in tour stop buffers
First_gdp	1	SA + SED	GDP of the CSD that belongs to the first leg of the tour
Last_gdp	1	SA + SED	GDP of the CSD that belongs to the last leg of the tour
First_pop	1	SA + SED	Population of the CSD that belongs to the first leg of the tour
Last_pop	1	SA + SED	Population of the CSD that belongs to the last leg of the tour

CVS: commercial vehicle survey; SA: simulated annealing; SED: socioeconomic data; PB: Pitney Bowes

4.0 **Results**

The GBM's performance was evaluated on the OOB data. Two key metrics were chosen to evaluate the GBM with the following in mind: *Which CFAF groups are alike, causing the GBM to cross-classify tours between them? Is the GBM better than simply guessing randomly based on frequency of each class?*

Two metrics were chosen that would provide insight to the above questions: first is a confusion matrix (CM) while the second is the Cohen Kappa (CK) metric. In a CM, each row of the matrix represents instances in an actual class (observed) and each column represents the instances in a predicted class. In a perfect classifier, all non-diagonal entries will be zero. The CK is more robust than just a simple percent agreement as it considers the possibility of agreement occurring by chance. There is some ambiguity in interpreting the values of CK, but in general anything less than 0.2 is questionable and close to 1.0 is a perfect model. Figure 2 shows the CM obtained from the OOB data. Some observations are presented below:

- The overall accuracy of the model is at ~39% (1-0.6110). The *Accuracy by random chance* measure was 12.2%, meaning that the GBM is over three-times more accurate than random selection.
- The CK value achieved in the final model was 0.3.
- The MNRLS (minerals) category, has the highest precision and recall rates[iv].
- MISC (mixed freight) is often confused with the other CFAF groups. This is partially because MISC is the majority class in the training dataset and its non-homogenous nature.

▾ PREDICTION - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	AGRI	BMETL	FOOD	FRPAP	FUELS	MISC	MNRLS	OTHMF	PLCHM	TRANS	WASTE	Error	Rate	Recall
AGRI	130	51	54	16	1	82	13	29	26	9	7	0.6890	288 / 418	0.47
BMETL	17	288	59	17	7	123	22	77	47	47	20	0.6022	436 / 724	0.31
FOOD	23	64	280	28	0	162	5	57	52	22	6	0.5994	419 / 699	0.34
FRPAP	6	65	71	156	2	106	12	47	52	23	8	0.7153	392 / 548	0.46
FUELS	4	20	7	3	60	22	9	11	13	7	4	0.6250	100 / 160	0.66
MISC	28	114	122	32	8	668	9	95	61	65	9	0.4484	543 / 1,211	0.40
MNRLS	8	37	18	10	2	26	171	10	9	6	11	0.4448	137 / 308	0.63
OTHMF	21	96	71	23	4	164	10	163	61	55	6	0.7582	511 / 674	0.25
PLCHM	20	100	77	30	3	139	6	79	149	27	6	0.7657	487 / 636	0.29
TRANS	9	61	40	12	1	118	5	69	24	290	4	0.5419	343 / 633	0.52
WASTE	8	43	24	11	3	53	8	17	20	8	97	0.6678	195 / 292	0.54
Total	274	939	823	338	91	1663	270	654	514	559	178	0.6110	3,851 / 6,303	
Precision	0.31	0.40	0.40	0.28	0.38	0.55	0.56	0.24	0.23	0.46	0.33			

Figure 6. Confusion Matrix on OOB Data – GBM

5.0 Conclusions, Reflections and Future Directions

This paper presents a ML model that uses the CVS and available firm data to predict the commodity carried by trucks. Commodities were estimated using he 11 group CFAF classification system. This model has been designed to be applied to GPS-observed trucks. Model performance indicates a useful model, with an overall accuracy of 39% and a CK of 0.3.

An evaluation of CVS trucks that carried a GPS unit versus those that did not showed a nearly identical distribution of commodities. This supports our expectation that the GPS carrying trucks do not behave differently than those that did not and the learner could be safely applied on the GPS truck tour data.

The MTO is collecting the next round of CVS data. Assuming a similar effort such as the 2012 CVS, there is the potential to double the training and testing data. An evaluation of the GBM model's performance would also help focus survey efforts by geography, distance, and commodity stratifications. Thus, while the dataset used in this round of the GBM was relatively small, it is expected to help optimize future surveys and prove to be a valuable source of information in policy planning.

[i] https://www150.statcan.gc.ca/n1/pub/50-503-x/50-503-x2018001-eng.htm

[ii] https://www.statcan.gc.ca/eng/concepts/industry

[iii] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[iv] Precision is obtained by dividing the diagonal with the actual observations e.g. MISC = 0.55, which is obtained by 668/1211. Recall is obtained by dividing the diagonal with the predicted observations e.g. MISC = 0.4, which is obtained by 668/1663. Precision and Recall are bounded between 0 and 1 with a higher value reflecting a perfect classifier.