Route Clustering in Transportation with Geospatial Analysis and Machine Learning to Reduce CO_2 Emissions

Topic Areas: Last Mile, Environment, Database Analytics

By: Ade Barkah and Patrick Robert

Advisors: Dr. Josué C. Velázquez-Martínez and Dr. Karla M. Gámez-Pérez

Summary: This research examines carbon emissions and fuel efficiency characteristics of last-mile delivery vehicles for Coppel, a large Mexican retailer. Using GPS traces and applying machine learning algorithms, we segment routes into four different clusters based on geospatial and other factors (including road elevation, road gradients, average vehicle speed, length between delivery stops). We then rank vehicles according to their fuel performance in each cluster. Finally, we suggest a fleet composition that could minimize fuel consumption and CO_2 emissions for the company.

Before coming to MIT, Ade was a Sr. Solutions Architect at Capco Canada. He pursued a B.S. in Mathematics and Computer Science from the Colorado School of Mines. After MIT, Ade will be launching a technology startup.

Before coming to MIT, Patrick worked as a Supply Chain Manager at EY Advisory Services. He graduated with a Licentiate Degree in Industrial Engineering and an M.B.A. from the University of Costa Rica.

KEY INSIGHTS

1. Geospatial analysis using GPS traces from delivery vehicles enhances visibility into the drivers of fuel consumption.
2. Using machine learning to cluster delivery routes enables “apples-to-apples” comparisons and ranking of vehicle performance under differing road and traffic conditions.
3. Considering route clusters in vehicle-route assignments could minimize fuel consumption and CO_2 emissions.

Introduction

The main drivers of climate change are the greenhouse gas emissions (GHGs) from sources that are largely attributable to human activity. Moreover, from the sectors that are significant contributors to GHGs, the transportation sector is growing the fastest. Carbon emissions from this sector may double by 2050 due to the rate of adoption of vehicles in developing countries. Hence, it is important to reduce pollution from fuel-based delivery vehicles, especially with the expected growth in transportation requirements due to the rise of e-commerce.

In our research, we work with Coppel, a leading retail company in Mexico, to analyze their last-mile delivery fleet’s fuel efficiency and CO_2 emissions. This fleet distributes items from regional distribution centers to both retail stores and customers throughout Mexico. Moreover, Coppel’s fleet consists of a number of different vehicle models operating in varying road and traffic conditions. This diversity makes it difficult for the company to directly compare and study the fuel efficiency and CO_2 emissions of their vehicles.

Methodology

1. Approach

In the first phase, we use clustering to determine which combinations of road conditions most affect fuel consumption, which is our proxy for CO_2 emissions. Since we have a large number of factors we can cluster with, we iteratively perform *k*-means clustering using factors shown to be important in vehicle emission models, such road gradients and vehicle speed. We then quantitatively and qualitatively validate the resulting clusters and features.

In the second phase, we aggregate routes by similar weight utilization levels to account for the effect load on fuel consumption. Then, for each cluster, we rank truck types by their fuel consumption to see which perform best across clusters and utilization bins.

In the third phase, we validate our clusters against a field study conducted in Mexico to determine whether the model outputs are representative of the actual conditions observed by the study participants. Finally, we analyze the results and present our conclusions.

2. Data Model Design

The data sets from Coppel consist of structured data from different sources and systems of the company. Since the data is disparate in nature, there is a need to understand its meaning, relationships, and key process drivers. This step leads to the development of a valid data model.

3. Data Preparation

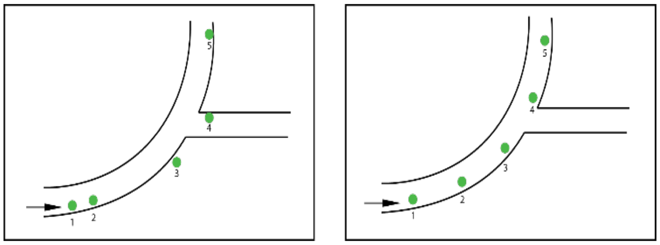
For the subset of Coppel vehicles with available GPS traces, we assemble and derive a number of data elements including roadway information (e.g. road elevation, gradient, segment length, etc.); fuel consumption; vehicle load data; vehicle characteristics (e.g. make and model, horsepower, torque, etc.); and the location (address) of each delivery point. We also augment and process this data with supplemental sources, such as the Google Maps set of Application Programming Interfaces (APIs).

4. Modeling Road Conditions

Our GPS data processing consists of 7 main steps:

i. *Initial quality analysis.* We analyze each of the 29,000+ GPS files to assess data breaks (gaps) in distance and in time, discarding records with breaks above pre-set thresholds from further processing.

ii. *Position normalization.* In addition to gaps, GPS data are subject to various accuracy limitations. We use the Google Maps "Snap to Roads" API to correct position data. Figure 1 (a) illustrates raw data from a delivery vehicle traveling along a certain road and Figure 1 (b) illustrates the normalized GPS data, with all points "snapped" to the correct road.



(a) (b)

Figure 1: Normalization of GPS traces

iii. *Elevation correction.* Due to satellite position geometry limitations, we cannot rely on the GPS altitude data for our calculations. At this step, we obtain corrected elevation data for each position using the Google Elevation API.

iv. *Segmentation.* Each GPS file data represents the entire route a particular truck takes throughout an entire day. Instead of considering average values for the entire day, we segment each route, based on stops the truck made during the day.

v. *Distance calculation.* From the normalized data we use the haversine formula (hav) to calculate the distance between each point, and sum the total distance traveled by the vehicle on that route for a particular day.

$$d = 2r \arcsin\left(\sqrt{hav(\varphi 2 - \varphi 1) + \cos(\varphi 1)\cos(\varphi 2)\, hav(\lambda 1, \lambda 2))}\right)$$

where φ1, φ2 are the latitudes at points 1, 2 and λ1, λ2 are the longitudes at points 1, 2.

vi. *Gradient calculation.* We calculate the road gradient at each GPS position. Additionally, we divide each segment into 100-meter sub-segments and estimate the average slope along the entire sub-segment. We use this information to generate a vertical profile of the segment.

vii. *Basic statistics.* We calculate basic statistics for various parameters (gradient, velocity, elevation).

We use these statistics as factors to the clustering process.

5. Fuel Emission Factors

Truck utilization (i.e., load) is another element that we consider in our analysis. Heavier vehicles use more fuel, all else being equal. To account for this effect, we bin our data into four groups of similar loads (i.e., low, medium, high, and overutilization bins).

Fuel consumption calculations enable the direct computation of emission factors for routes within the clusters. To estimate the emissions, we use the NTM methodology and use a factor of 2.615 kg of CO_2 emitted per liter of diesel fuel burned.

6. Field Study Validation

As part of this research, we conducted a 3-week validation field study in partnership with the Instituto Tecnológico y de Estudios Superiores de Monterrey. The objective of the study was to directly observe and capture different road and traffic conditions in diverse regions of Mexico. Each observation also captured metadata such as GPS position and event timestamp in a mobile app.

Captured events included observations such as climbing steep hills, experiencing heavy traffic, stopping or passing through a stoplight, among others. We use these observations to validate our cluster results.

Results

1. Cluster Analysis

We separately perform k-means clustering on each of the 4 utilization bin files. From the 24 parameters computed in the GPS processing steps, we choose 6 parameters to form our clusters with:

- Gradient variability (proxy for hilly conditions)
- Mean velocity
- Mean elevation
- Average segment length
- Percent of the route that's flat (road gradient is less than +/- 1%)
- Percent of the route that's steep (road gradient is 4% or greater)

For each utilization bin, we determine the optimal number of clusters by plotting the within-cluster sum of squared errors (SSEs). We pick k = 4 as the optimal number of clusters by using the "elbow method.

Setting k = 4, we examine the data to detect similarities between routes within the same cluster, and differences between routes in different clusters, as summarized in Table 1.

Parameter	Cluster A	Cluster B	Cluster C	Cluster D
Elevation	High	Low	Low	High
Topology	Hilly	Flat	Flat	Flat
Average Velocity	Low	Medium	High	Low
Segment length	Short	Medium	Long	Short

Table 1: Qualitative evaluation of the cluster centers

Some additional qualitative observations on the clusters may include:

- Cluster A primarily describes high altitude urban areas near Mexico City.
- Cluster B denotes small and medium-sized cities with low elevation.
- Cluster C is indicative of rural areas.
- Cluster D mainly describes outskirts areas of Mexico City.

2. Ranking of Vehicles

After assigning the different routes into clusters, we analyze the behavior of the emission factor across the clusters.

For the medium and low utilization scenarios, Cluster A is the cluster that shows the greatest impact on CO_2 emissions, being approximately 10% larger than the other clusters. In contrast, in the scenarios of high and overutilization we did not see any notable difference among the emission factors between clusters.

We observe in Figure 2 the combinatorial effect that the cluster, vehicle type, and computed utilization have on the emission factor. We notice how certain vehicle types have on average a higher emission factor across the clusters and rank them accordingly.

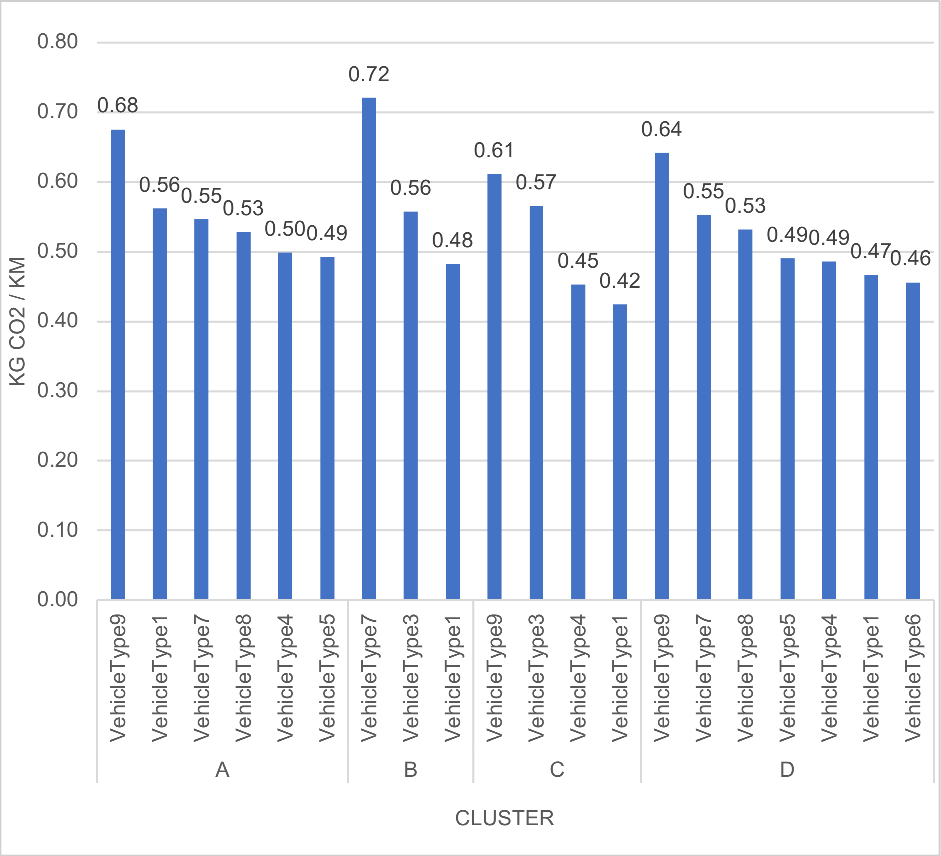


Figure 2: Average CO_2 emission factor per vehicle type for medium utilization. Values calculated for ranking purposes based on the number of routes per cluster by vehicle type.

For the Coppel fleet, vehicles that are more than 8 years old have the largest emission factor, on average.

Our analysis shows that the performance difference between the top and bottom performers within each cluster may be significant.

3. Potential CO_2 Reduction

If we consider a scenario where we exchange or substitute all other vehicle types by the best performing vehicle type in the cluster, we can reduce the average CO_2 emissions by 7.2%.

Conclusion

We applied our methodology to a fleet of delivery vehicles for a large Mexican retailer. We clustered road and traffic conditions based on over 29,000 GPS traces generated by a subset of the vehicles. We show that delivery routes can be meaningfully clustered based on factors such as road elevation, road gradients, average vehicle speed and the length between delivery stops. Furthermore, we found a cluster of routes associated with increased fuel consumption and ranked the most efficient vehicles for each cluster.

Our results support the notion that some vehicle types perform better in certain clusters, giving an opportunity to exchange vehicles between regions to optimally assign vehicle types to delivery areas. We estimate that using the best vehicle type in each cluster may yield up to a 7.2% reduction in fuel consumption and CO_2 emissions.