Truck Parking Inventory and Utilization Appendixes

WA 3 Task 2.3/2.4/2.7

Final 1: February 28, 2020
Appendix B: Procedures for Preparing Location Based Data for Analysis

The goal of this task was to process raw GPS point data provided by the American Transportation Research Institute (ATRI) into usable stop information. Knowing the location of a truck at regular points throughout the day gives a great deal of information about its travel (including stops). For example, a GPS unit may return a report at a series of mileposts on a highway, then multiple reports are recorded at the same location over an hour or two, then the points begin moving down the highway again. The location of the stop, the length of stop, and the relative direction of travel before and after can all be discerned from this information.

However, the wide range of possible activities and situations that influence where a truck is and what it is doing at any given moment can make interpreting each set of points complex. For example, traffic may cause multiple points to be in the same place though the truck has not made a stop. When stops are short or movements are very slow, the GPS traces of stops vs. non-stops start to become more similar. The process of discerning these must keep the true stops while filtering out everything else. It must also maintain consistency in its definition of stops and non-stops across all trucks.

Another challenge is handling the massive amount of data involved when looking at a state-wide, near-population-scale data set of trucks. Algorithms which are computationally intensive or poorly programmed can become prohibitively time-consuming.

To tackle these issues, the team developed a series of sophisticated heuristics to iteratively check every data point and analyze patterns of time, distance, and location. The team also used a combination of Spark, Python, and R to efficiently process the data through a series of steps.

1.1 Raw Data

The raw truck GPS data from four periods during different seasons of the year (shown in Exhibit 1) was purchased from ATRI for this truck parking analysis. These periods included at least two of each day of the week, accounting for weekday variations. Seasonal effects were captured by analyzing different months of the year.

Exhibit 1: ATRI Raw Data Periods

<table>
<thead>
<tr>
<th></th>
<th>February</th>
<th>May</th>
<th>July</th>
<th>October</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Date</td>
<td>02/01/2018</td>
<td>05/01/2018</td>
<td>07/16/2018</td>
<td>10/16/2018</td>
</tr>
<tr>
<td>End Date</td>
<td>02/15/2018</td>
<td>05/15/2018</td>
<td>07/30/2018</td>
<td>10/31/2018</td>
</tr>
<tr>
<td>Number of Truck (Thousands)</td>
<td>168</td>
<td>161</td>
<td>229</td>
<td>165</td>
</tr>
<tr>
<td>Number of Points (Millions)</td>
<td>385</td>
<td>407</td>
<td>415</td>
<td>380</td>
</tr>
</tbody>
</table>

Trucks appear in the data any period in which they were inside the state of Texas. For each period, ATRI provided GPS traces for every truck that operated in Texas at any time. If any of these trucks recorded points in other states, these were included, but without an exact location. The devices recorded data at regular intervals as frequent as 60 seconds apart, but could be as long as 15 minutes. Typically no data was recorded when a truck was shut off – at which time the truck would not move either. As it would report the same location before and after shut down, our process marks the whole period as part of a stop.

Exhibit 2 shows the points from a single example truck, reported over a period of 15 days in October when it was moving or stopped. This truck was clearly in many places around Dallas, Laredo, and especially Houston. The truck may have stopped, stopped multiple times, or passed through any of these. The time stamp (not shown) indicates that the first point is recorded on the state border by Texarkana on the 3rd day of the period, when the truck entered Texas. It proceeded to the Dallas area and onto Laredo, though it does not appear to have crossed the border. For the rest of the study period was within the state. At one time the truck is turned off and stored for a weekend at a facility in Houston. Mapping all 1,950 points illustrates the amount of information as well as the need for the process described here. In this form, there is little useful information about truck parking.
Exhibit 2: Example of a Single Truck's Points

This study is focused on identifying locations where trucks park rather than fully understanding travel patterns of individual vehicles. In order to determine this in a state where many facilities are located close to roadways and other businesses, a high level of accuracy and precision is necessary. These two terms are often used interchangeably, but they are important concepts when using location data of any type. Exhibit 3 provides an illustration of the definition applied for this study.

For truck parking, the points must be precise enough to distinguish between those inside a lot and those in another nearby facility. This is especially true when identifying overflow from legal parking areas into illegal spaces. Imprecise data, like in target c) below, would produce a scattershot of stops in and around a lot. Rules which draw a border tightly around the space would erroneously miss many points. Points from trucks not in the facility of interest may land within the facility. Inaccurate reporting, as in b) and d), could bias the data by putting shifting points in a particular direction, resulting in similar issues to imprecision.

Exhibit 3: Diagram of Accuracy and Precision

```
| a) Accurate, Precise | b) Inaccurate, Precise | c) Accurate, Imprecise | d) Inaccurate, Imprecise |
```

Source: Cambridge Systematics (2019).

Raw GPS data from ATRI is fortunately accurate and precise. It is, in fact, so precise that none of the points in Exhibit 4 fall more than a few feet from pavement. This map contains records of 847 stops over the data collection period. While there are too many to clearly see each point even with a sample, they form an outline of the places where trucks go.

Exhibit 4: Example of Raw Data Points at Iraan Rest Area


Given there are so many points and they are not far to the side of the road, it is safe to assume that their place along the length of the roadway or ramp is very close also. This
allows us to differentiate trucks on the ramp versus those in the rest stop. This drastically increases the amount of information about individual facilities and how they are used.

1.2 Methodology

1.2.1 Rules and Heuristics

A stop was defined most generally as an occurrence in which a truck was in one place for at least 15 minutes. This amount of time is long enough to discount stops due to traffic congestion or construction but short enough to capture drivers stopping to obtain food or meet the 30 minute break requirement. Trucks stopped at a location may move short distances within the facility or report slightly different locations due to satellite positions, so the stop ended only when it traveled a certain distance based on the moving average of the pings of the stop. Pings which were identified as waypoints – reports from a truck moving on the highway – were marked and filtered out. This was based on the speed and the frequency of reporting. Unreasonable speeds indicated that points were given and erroneous location or time. They were compared to the adjacent points and points most likely to be incorrect were filtered out.

Even with some allowance for movement within a facility, two stops would occasionally be produced just far enough apart to technically be marked as separate. Consecutive “stops” meeting very specific distance and time criteria were allowed one more chance to merge at the end.

At the conclusion of stop-making, categories were imposed:

- Overnight – Stop which starts before 3 am and ends after 3 am, with minimum duration of 4 hours (and max 48 hours).
- Long-haul – A stop between 2 trips of at least 3 hours travel time each.
- Staging – A stop with duration between 1 and 4 hours, following a trip of at least 1 hour and followed by a trip of less than 2 hours.
- Local – A stop of less than 2 hour stop duration and between 2 trips of up to 1 hour travel time each.

1.2.2 Data Handling and Pre-Processing

The datasets from the 4 periods (see Exhibit 1) were received and handled separately, but using the same process. The programming language Spark (implemented in Python) was used first as it is designed specifically to handle large datasets. Spark read and combined all the data – inside and outside of the state of Texas – for each period. After filtering out unused out of state points, assigning new truck ids, sorting and doing some other formatting changes, data were exported into a series of smaller files, each containing mutually exclusive sets of trucks. The smaller files could be more easily managed in R and run independently.
The next step occurred in R and prepared the data for the main processing steps. Some minor cleaning was done to preclude unnecessary or erroneous calculations. Additionally, initial calculations were done and the format of the data object set to save time in the next steps. Lastly, the (only slightly processed) data were summarized to show basic characteristics or unusual trends.

### 1.2.3 Analysis

Setting up the processing steps required understanding the initial state of the data and the new state after each step. This generally consisted of checking overall changes like statistics describing what was dropped or kept. Maps of truck points showed how a random sample or a specifically targeted set of individual trucks’ points were treated. Diagnostic runs allow for close examination of what was passing through each filter to become a “stop” by:

- Comparing results vs raw data;
- Comparing results vs previous step’s results;
- Analysis of trip-stop patterns; and
- Checking of unusual rates of stopping, lengths of time, speeds, etc.

The result was a series of filters that dropped records, grouping algorithms that clustered multiple records together, and calculations which analyzed their status, along with the carefully honed parameters which controlled them. Each step was closely monitored to address the following questions:

- Was anything filtered out or changed that should have remained?
- Was anything retained that should have been dropped or changed?
- If yes for 1 or 2, is this the step to take care of it or is there a later step that will?

When the answer for question 3 was “yes”, changes were made and tested until a desired result was reached. Not every pause in a truck’s movement is a stop and not every stop consists of many reports at exactly the same location. Consequently, the heuristic approach requires repeatedly taking different perspectives of each trace while carrying along the information inferred in previous steps. As the process progressed, the number of data points decreased and the changes became more limited and precise. Early steps allowed a lot of non-“stops” through to avoid taking out the real ones. For many trucks the later stages had no impact on their attributes because their stops were already identified.

### 1.3 Special Cases and Errors

#### 1.3.1 Trucks Traveling Outside of Texas

Many trucks which recorded travel in Texas also visited other states. In these cases, the information from the time and location (zone) where it entered or left the other state was
kept along with its stop within Texas. That is, the data used followed the truck as it traveled through Texas and when it left the state kept a single record stating the first zone in Arkansas (or LA, NM, OK). Conversely, a truck that was elsewhere and entered Texas has a single record stating the New Mexico (or LA, AR, OK) zone it had reported prior to the point inside Texas.

Data on these trips made it clear that there were no stops or trips inside Texas before the entry point, after the exit point, or between exit and entry points. This produced trip-stop patterns such as: Truck left location A in Plano at 2 PM to location B in Mesquite, TX where it stopped for 45 minutes at 2:30. It then drove for at least 2 hours and 30 minutes, reporting a location in or around Shreveport, LA at 5:45. The program effectively interpreted this as: I do not know what it was doing and it is not relevant to this study. Data on travel outside Texas was dropped as being outside the scope of this study.

An issue occurred in the data that disrupted this process. Some trucks that almost certainly left the state were not reporting, leaving the appearance that they had either inexplicably stopped reporting near the border only to show up again at another border crossing, or more problematically that they were stopped on the border. The latter could be misinterpreted as a stop if the following criteria were met: 1) the truck entered and exited on the same road, 2) the pre-exit point and post-entrance points – temporally consecutive in the database – are close together, and 3) no outside locations were reported.

This data error was corrected in post-processing after analysis of the problem. Fortunately the points had several characteristics which could be used for identification. First, they were predominately very long “stops”. Second, they were in the right-of-way. Third, most occurred within a mile or even a few hundred feet of the border. Fourth, they consisted of very few pings – often only 2. (Stops, especially those in which the truck was not turned off, could record many pings in the stop location.) “Stops” within a mile of the border, in the right-of-way, were dropped unless the duration was less than an hour. To catch those reported further into the state, a 10 mile buffer was used and stops within this buffer were dropped only if they had a duration over 10.5 hours (to account for trucks stopping to fulfil a 10-hour break).

This process is shown in Exhibit 5 (before filtering) and Exhibit 6 (after filtering). Truck parking in the lot near the border remain as do many on other roads. A few erroneous points may remain or genuine stops disappear, but overall this was effective.
1.3.2 Blackouts

One additional special case occurred when an unexplained disruption to location information occurred. Often, gaps in data started and ended in the same location suggesting that the truck was simply turned off. A small number of trucks experienced a disruption which started after a report in one place and ended up with another in a clearly different place. These were noted and not considered stops. The very limited set of trucks which had several of these were considered to be malfunctioning and dropped from the analysis.

1.4 Results

The output from this process is a set of all stops by all trucks in each period. Exhibit 7 shows a summary of points, stops, and trucks. ATRI trucks made almost 175,000 stops of at least 15 minutes every day. The number of trucks given is the sum of unique trucks in each period – a truck that appears in the February and July data will be counted twice in this total.
The average number of stops per truck is around 16, but this includes trucks which only spent a short time in Texas as well as those which traveled the whole 15 days without leaving. The 25% of trucks with the most stops in the first half of February averaged almost 50 stops and a few at the very top. A small percentage (~2%) but still significant number of frequently stopping trucks made at least 100 stops in the February period.

**Exhibit 7: Summary of Stop Processing Results**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Average Per Truck*</th>
<th>Average Per Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Data Points</td>
<td>1,587,668,148</td>
<td>2,195</td>
<td>26,027,347</td>
</tr>
<tr>
<td>Processed Stops</td>
<td>10,590,349</td>
<td>15.9</td>
<td>173,612</td>
</tr>
</tbody>
</table>


*Trucks are counted once for each period in which they appear.

A key element to understanding truck parking is the duration of the stop and the travel pattern. In Exhibit 8, the stops are broken out by their length. Nearly 60% of the stops were less than an hour. The next largest segment were trucks stopping between 1 and 4 hours. Since this data includes trucks stopped for any purpose at any location in the dataset, the high number of stops of short magnitude is not surprising. The proportions of stops by length did not vary from one season to the next.

**Exhibit 8: Stops by Duration**

![chart showing stop duration]


### 1.4.1 Examples of Individual Trucks

Below are some examples of the truck data before and after being processed into stops. The stops are the critical input to the truck parking analysis using ATRI data described in the associated technical memos for this project. Exhibit 9 shows a truck that started traveling
from the Texarkana border through Dallas, San Antonio, Austin, Nuevo Laredo then back to Dallas, although only a handful of (in-state) stops were made until it reached the greater Houston region.

**Exhibit 9: Example Truck 1 – Truck Stop Locations Before and After Processing**

![Map showing truck stop locations](image)


The inset view in Exhibit 10 shows the activity even further. Both Dallas stops occurred at the same private facility, one of which was over 12 hours. In the Houston region there were many stops less than 1 hour in which it went to locations outside the beltway. All of its extended and overnight truck parking occurred at a facility owned by a freight company rather than public or private truck parking locations.
Exhibit 10: Example Truck 1 – Inset of Raw (Red) and Stop Data (Blue)

The truck mapped in Exhibit 11 did more intercity travel, returning multiple times to Dallas. As a result it had fewer stops. It made several trips between Dallas and the Abilene area, and a few elsewhere. It parked several times at privately owned truck parking properties like an 11 hour stop at a Love’s West of Abilene. The recording frequency varied significantly, resulting in the sparse trace on I-10. On the highway the frequency was often 1 point every 15 minutes, and while moving it never fell below this.

Exhibit 11: Example Truck 2 – Truck Stop Locations Before and After Stop Processing

The last example, Exhibit 12, shows another truck which makes a few unique trips in and out of Texas. It also sat for several days at a shipping facility outside of Brownsville. While it made several long stops, these were not at an authorized public or private truck parking facilities. This truck is somewhat unusual because its combination of travel pattern and consistent reporting, even when stopped, produced only 17 stops from almost 5,000 points.

**Exhibit 12: Example Truck 3 – Truck Stop Locations Before and After Stop Processing**