Driving Style Signatures: Who's Behind the Steering Wheel?

(Student: Vinh Luong - 442069)

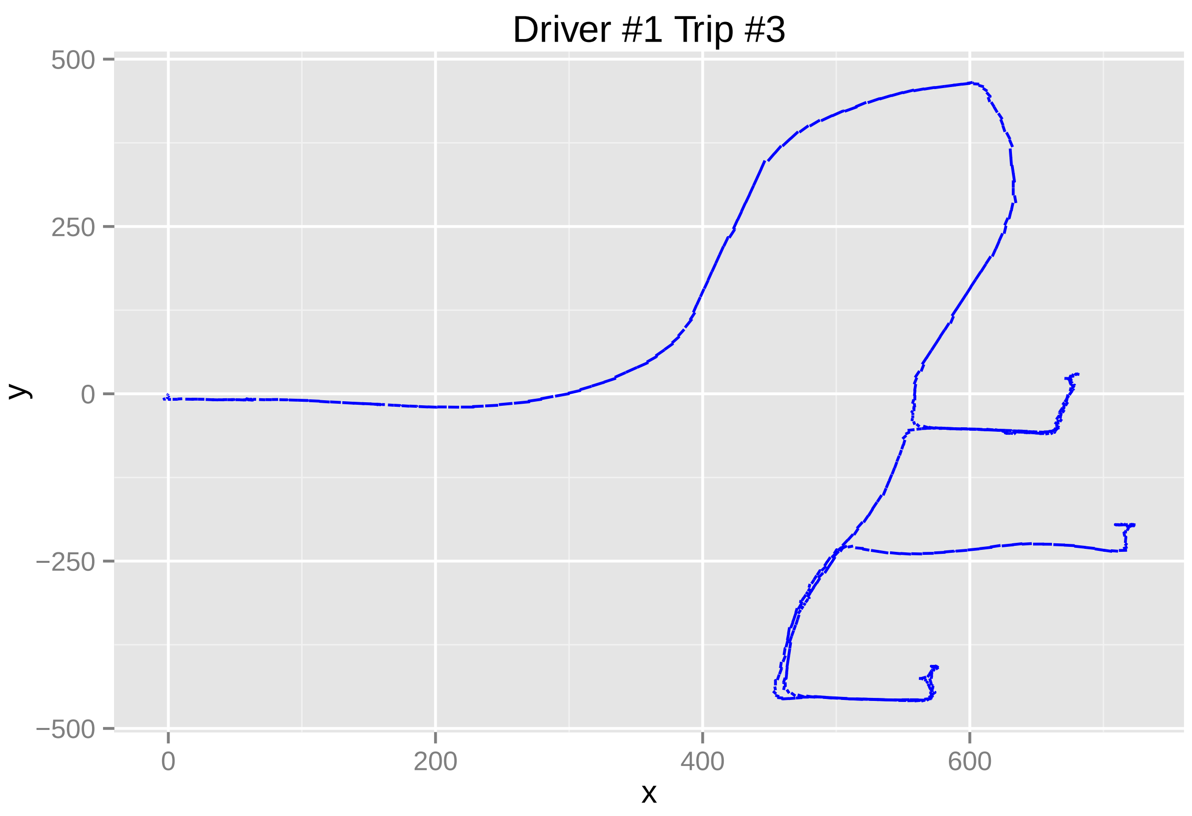
Executive Summary

This project attempts to develop a method to detect different people's own driving style "signatures" from a series of second-by-second GPS coordinate readings from their car's telematics devices.

1. Data and Data-Preprocessing

1.1. Data and Basic Higher-Order Features

We obtained second-by-second GPS (x_t, y_t) coordinate data from over half a million anonymized driving trips (200 trips by each of over 2,700 individual drivers) from French insurer AXA's Kaggle competition data set. A typical trip looks like the following:



For anonymization purposes, each trip's starting point was centered at (0, 0) and the subsequent coordinates were rotated by a random angle.

From the raw (x_t, y_t) data, we derived a number of basic higher-order features as follows:

x-velocity: $\Delta x_t = x_t - x_{t-1}$
y-velocity: $\Delta y_t = y_t - y_{t-1}$

x-acceleration: $\Delta\Delta x_t = \Delta x_t - \Delta x_{t-1}$
y-acceleration: $\Delta\Delta y_t = \Delta y_t - \Delta y_{t-1}$

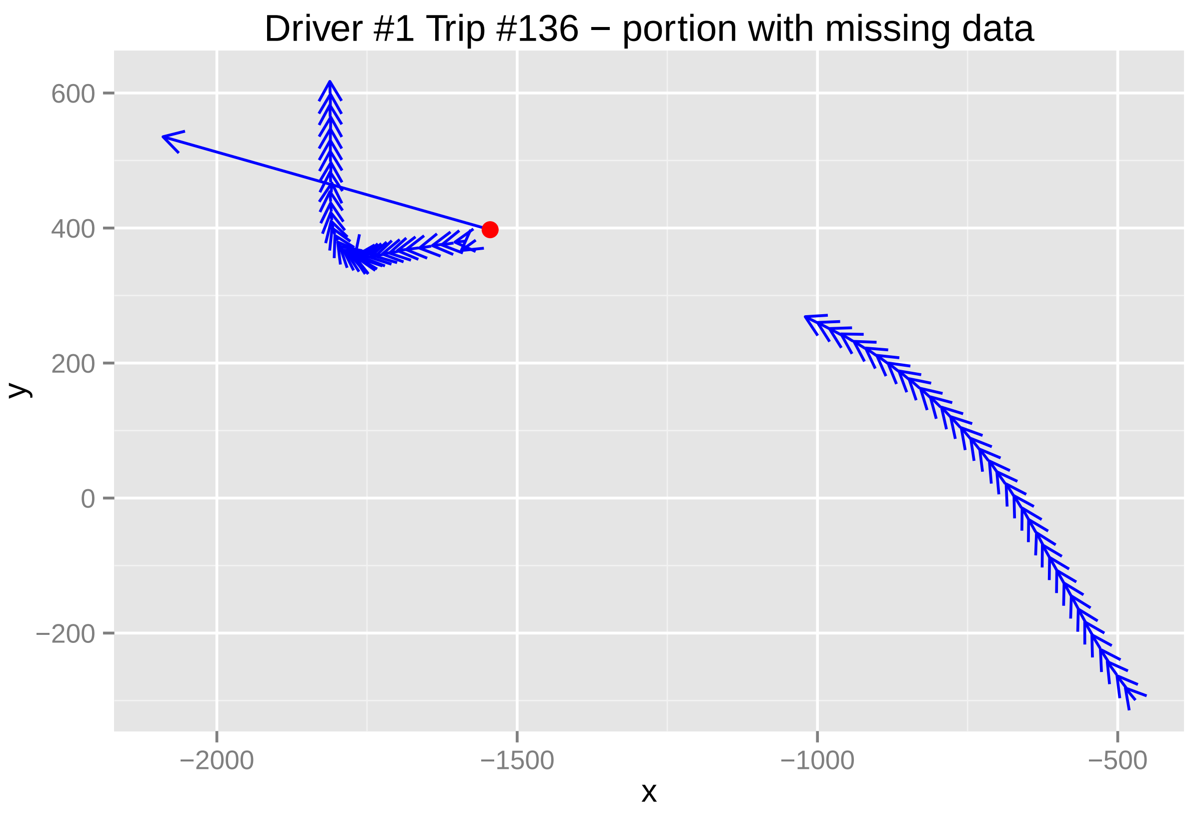
velocity magnitude: $v_t = \left\| \begin{array}{c} \Delta x_t \\ \Delta y_t \end{array} \right\|$

acceleration magnitude: $a_t = \frac{1}{v_t}\left\langle \left[\begin{array}{c} \Delta\Delta x_t \\ \Delta\Delta y_t \end{array} \right], \left[\begin{array}{c} \Delta x_t \\ \Delta y_t \end{array} \right] \right\rangle$
(i.e. acceleration in the direction of the velocity vector $\left[\begin{array}{c} \Delta x_t \\ \Delta y_t \end{array} \right]$)

angle: $\theta = \arctan(\Delta y_t, \Delta x_t)$
angular velocity: $\Delta\theta = \theta_t - \theta_{t-1}$
angular acceleration: $\Delta\Delta\theta = \Delta\theta_t - \Delta\theta_{t-1}$

1.2. Data Cleaning

Before we could proceed with analyzing this large data set, we had to attend to some data integrity issues. It turned out that due to lost communication signals and/or some extreme anonymization measures, many of the driving trip data sets were plagued with coordinate "jumps", i.e. i.e. missing chunks of GPS readings. One example is portrayed below:



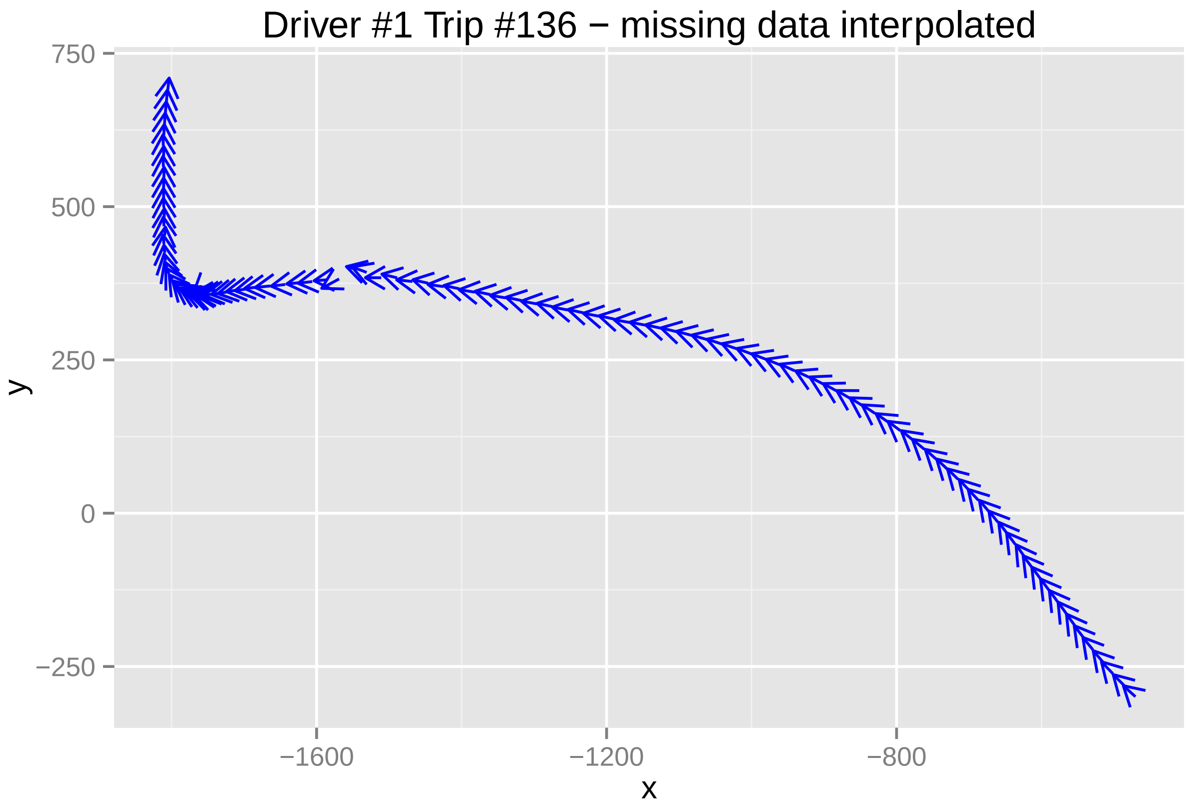
This problem was found present in over 25,000 driving trip data sets. Also, nearly 100% of the over 2,700 drivers have trips with missing data.

In the above depicted case, as well as in other missing data cases, the corrupt data portions (highlighted by red dots in the plots) manifest themselves quite apparently by an unreasonably large distance from (x_{t-1}, y_{t-1}) to (x_t, y_t), or equivalently an unreasonably high velocity v_t estimated from such consecutive pairs of coordinates. We hence devised a method to detect and interpolate the missing data, briefly as follows:

- detect data rows with derived velocity v_t over 50 meters per second;
- look at time windows of 3 seconds before and 3 seconds after each of such instance, and estimate the average velocities v_{before} and v_{after} and angular directions θ_{before} and θ_{after};
- estimate the length of a smooth circular arc spanning the locations $(x_{\text{before}}, y_{\text{before}})$ and $(x_{\text{after}}, y_{\text{after}})$ and with tangents at angles θ_{before} and θ_{after} at those points;
- estimate the number of seconds the vehicle needs to take to traverse such an arc at velocity $v_{\text{mean}} = \frac{1}{2}(v_{\text{before}} + v_{\text{after}})$;
- interpolate missing intermediate locations along the arc, with some minor technical adjustments to make the vehicle accelerate or decelerate evenly from v_{before} to v_{after}.

and interpolate cleaned the data by imposing a speed limit of 50 meters per second (over 100 miles per hour) and detect rows of data with estimated velocities over this limit.

With such a data interpolation method, the above case of Driver #1's Trip #136 was corrected to the following:



Overall, various data verification and cleaning steps took about 100 hours on a single computer running on seven cores.