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

(Student: Vinh Luong - 442069)

Introduction

The field of automotive insurance has a number of important questions regarding individuals' driving behaviors:

i. What data features are needed to characterize a person's driving habits? - such features can be used to appropriately price accident risk as well as cross-sell related insurance products;

ii. In the case of an accident claim, given such data features, how well can we tell if the insured person - and not another person - is really behind the steering wheel when the incident occurs?

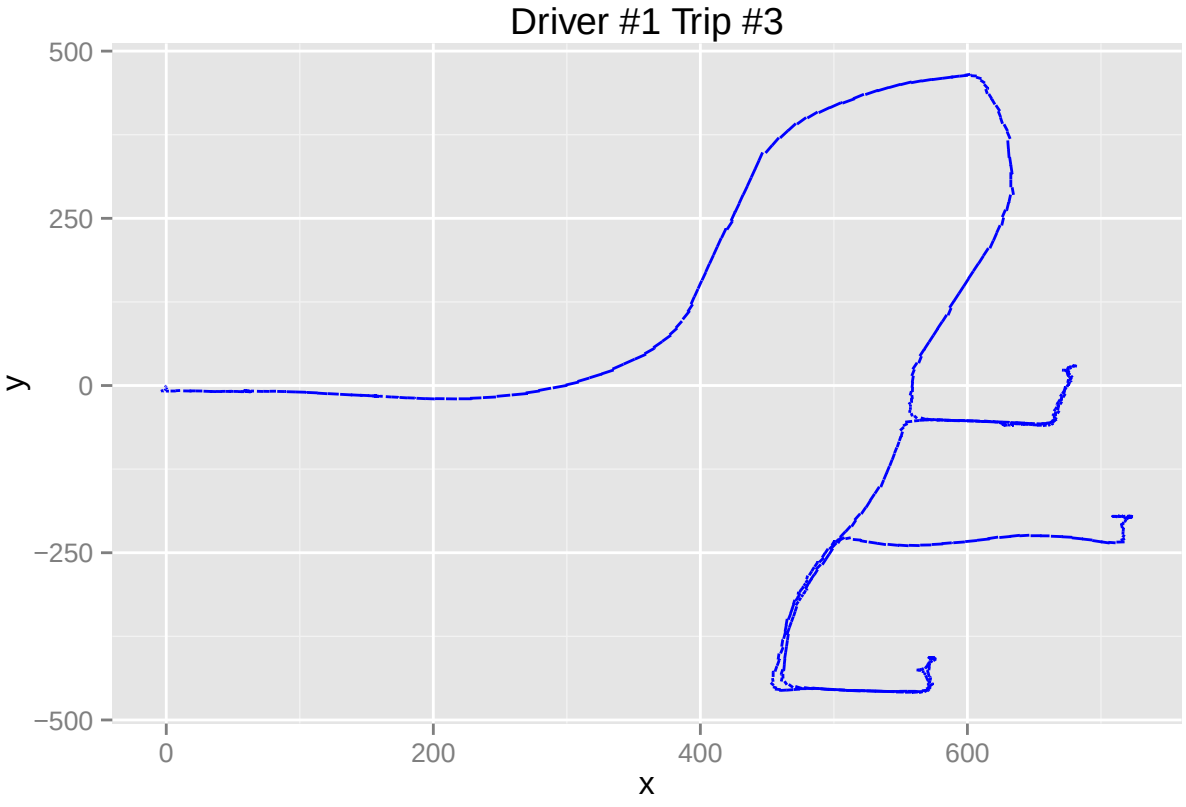
The advent of vehicle-mounted telematics devices has provided rich new data sources to address these issues. In this project, we attempt to develop a method to detect different people's own driving style "signatures" from a series of second-by-second GPS coordinate readings from their cars' telematics. We will show that by using just simple features such as velocity, acceleration, angular velocity and angular acceleration, we could correctly identify with [90%] accuracy whether the insured driver is driving his/her car.

1. Data and Data-Preprocessing

1.1. Raw Data and Processed 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. This large dataset occupies nearly 6 GB of storage space when unpacked.

One trip is depicted below:



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

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

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

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

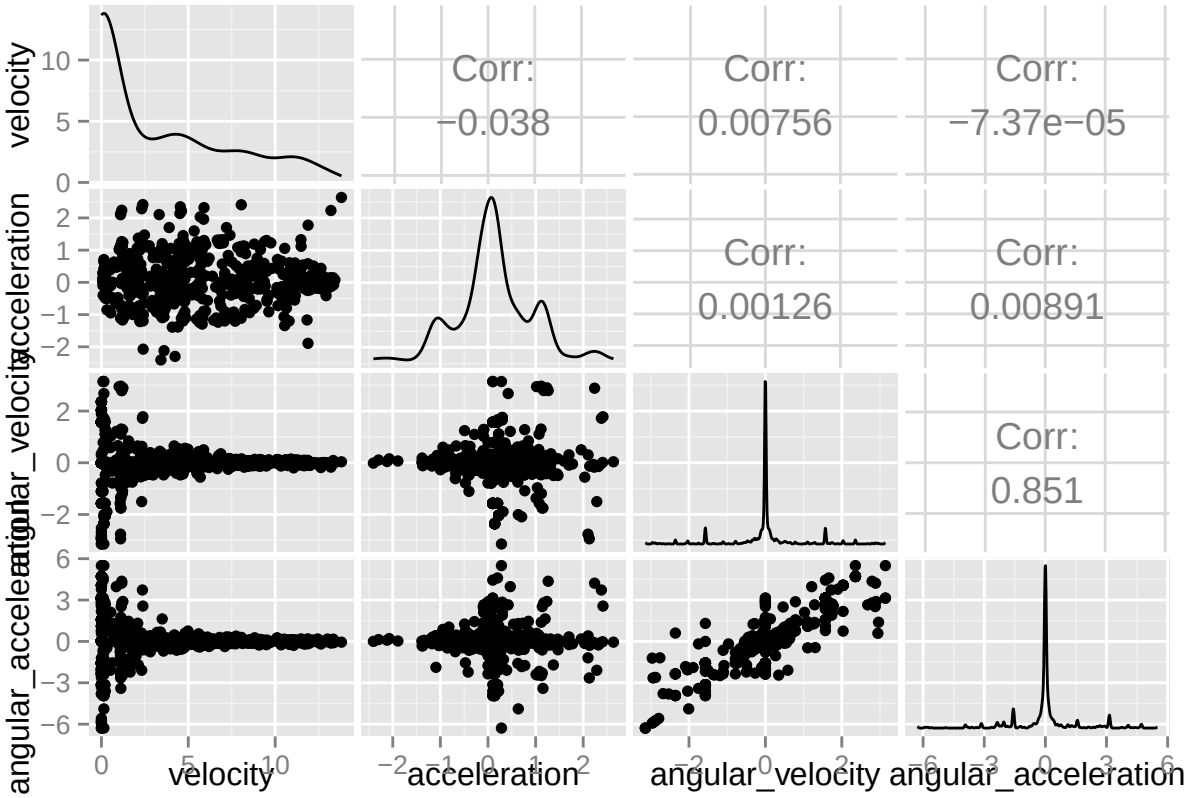
$$\textbf{absolute velocity magnitude: } v_t = \left\| \begin{array}{c} \Delta x_t \\ \Delta y_t \end{array} \right\|$$

$$\textbf{signed acceleration magnitude: } a_t = \frac{1}{v_t} \left\langle \left[\begin{array}{c} \Delta\Delta x_t \\ \Delta\Delta y_t \end{array} \right], \left[\begin{array}{c} \Delta x_t \\ \Delta y_t \end{array} \right] \right\rangle$$

(i.e. acceleration in the direction of the velocity vector $\left[\begin{array}{c} \Delta x_t \\ \Delta y_t \end{array} \right]$)

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

These features are measured for every second of each driving trip. Among them, we will focus on several features measuring **rates of change**, namely **velocity**, **acceleration**, **(signed and absolute) angular velocity** and **(signed and absolute) angular acceleration**. For the above depicted trip, the distributions and correlations among these variables are as follows:



1.2. Data Cleaning

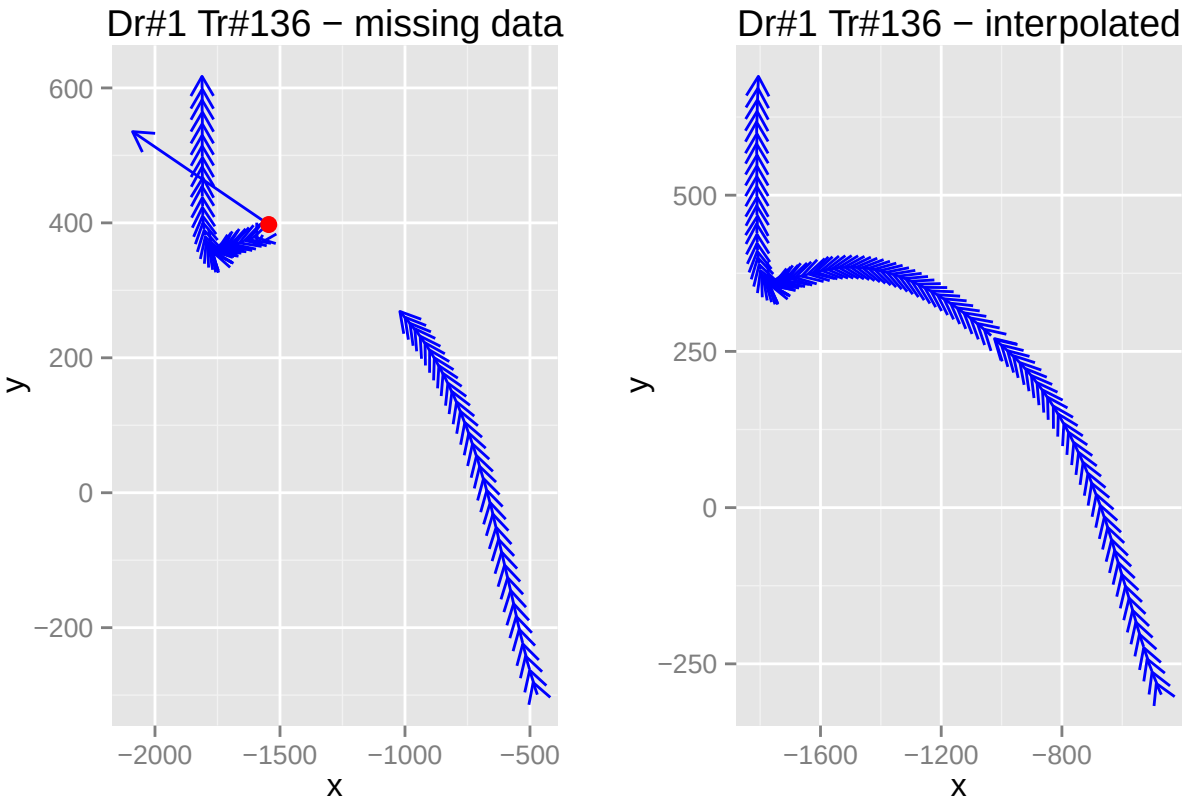
Before we could proceed with analyzing this large data set, we had to attend to some data integrity issues. It turns out that due to lost communication signals and/or some extreme anonymization measures, numerous raw driving trip data sets are plagued with coordinate "jumps", i.e. missing chunks of GPS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

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

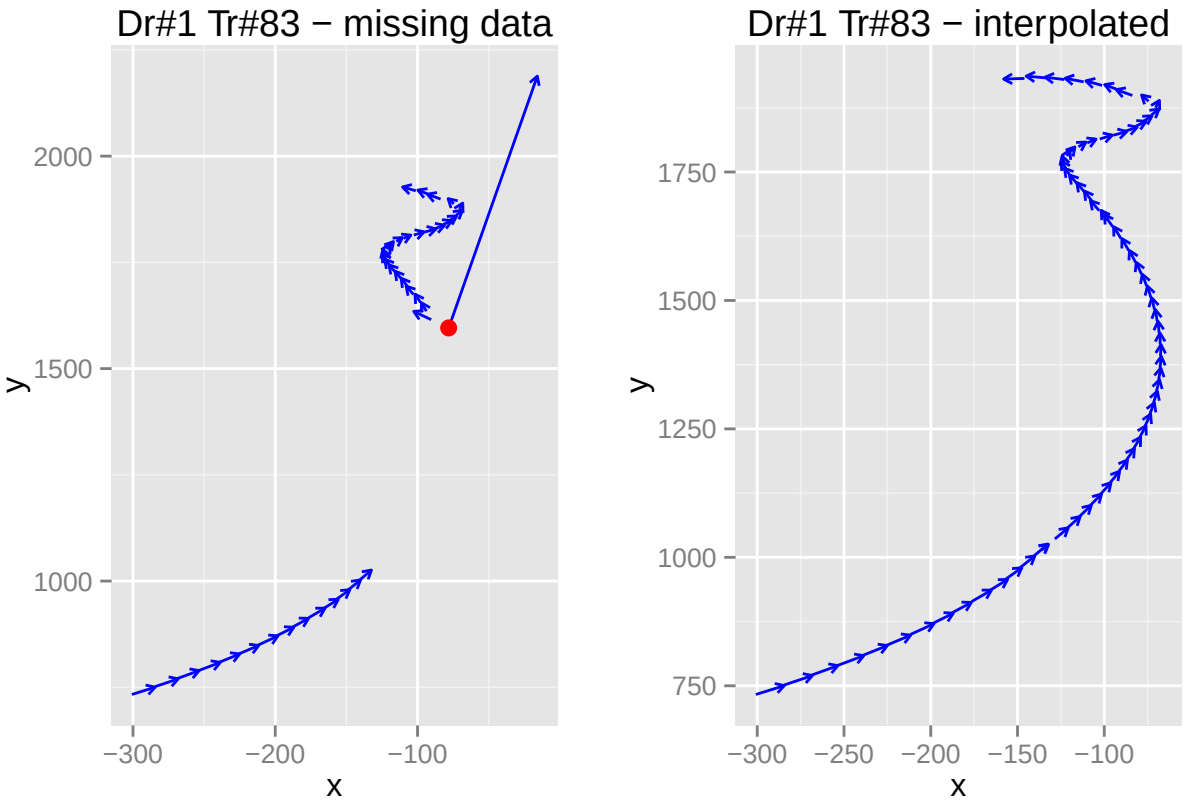
In the above depicted case, as well as in other missing data cases, the corrupt data portions (highlighted by red dots in the plots) manifest themselves quite apparently by an unreasonably large distance from (x_{t-1}, y_{t-1}) to (x_t, y_t), or equivalently an unreasonably high velocity v_t estimated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

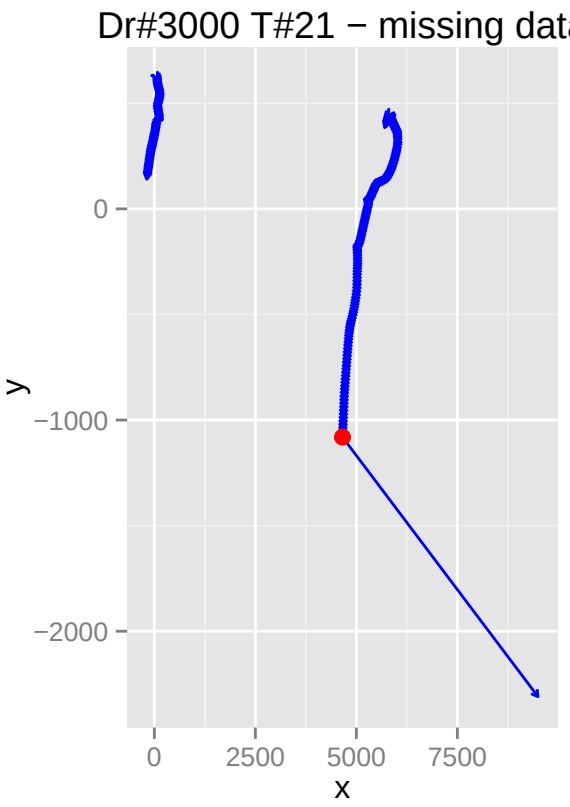
- detect data rows with derived velocity v_t over 50 meters per second;
- look at time windows of 3 seconds before and 3 seconds after each of such instance, and estimate the average velocities v_{before} and v_{after} and angular directions θ_{before} and θ_{after};
- by certain polygonal approximations, estimate the length of one or several smooth parabolic arcs 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; (*it will become apparent in certain visualizations below why parabolic curves are more natural than straight lines or circular curves*)
- estimate the number of seconds the vehicle needs to take to traverse such parabolic arc(s) at velocity $v_{\text{mean}} = \frac{1}{2}(v_{\text{before}} + v_{\text{after}})$;
- interpolate missing intermediate locations along the parabolic arc(s), with certain technical adjustments to make the vehicle accelerate or decelerate evenly from v_{before} to v_{after}.

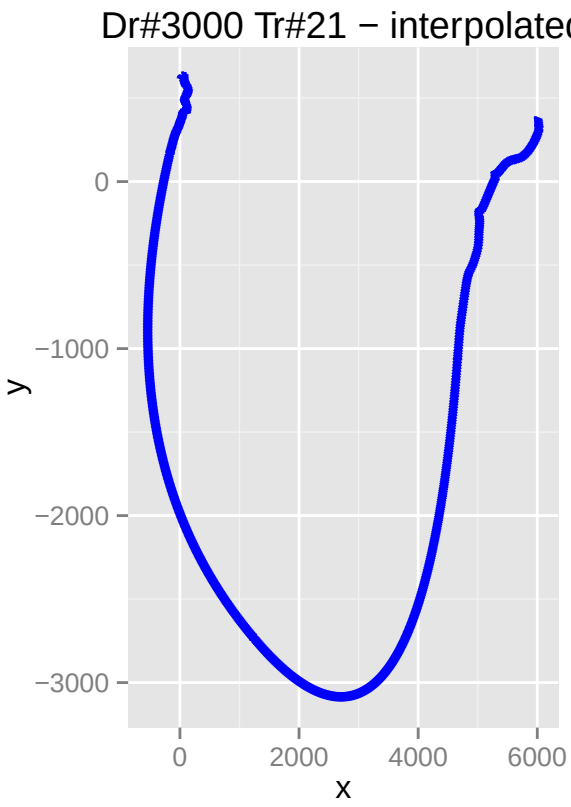
With such a data interpolation method, the above case of Driver #1's Trip #136 could be corrected to the following:

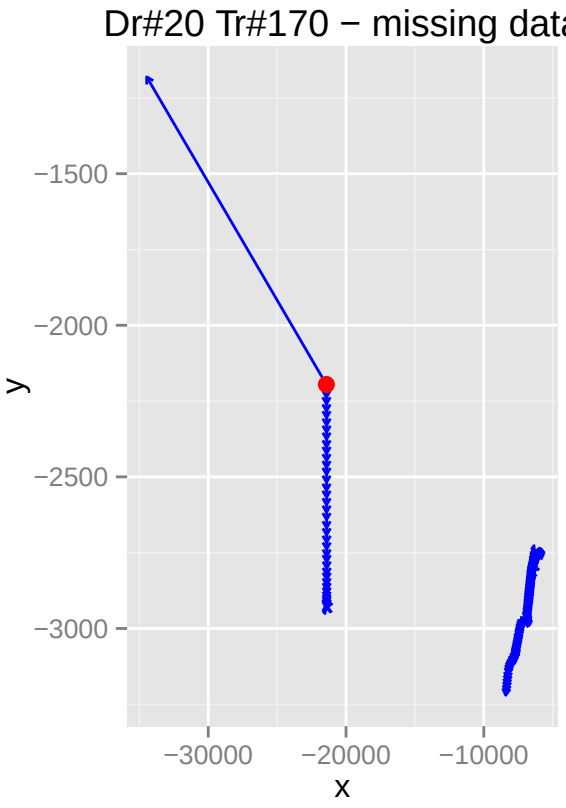


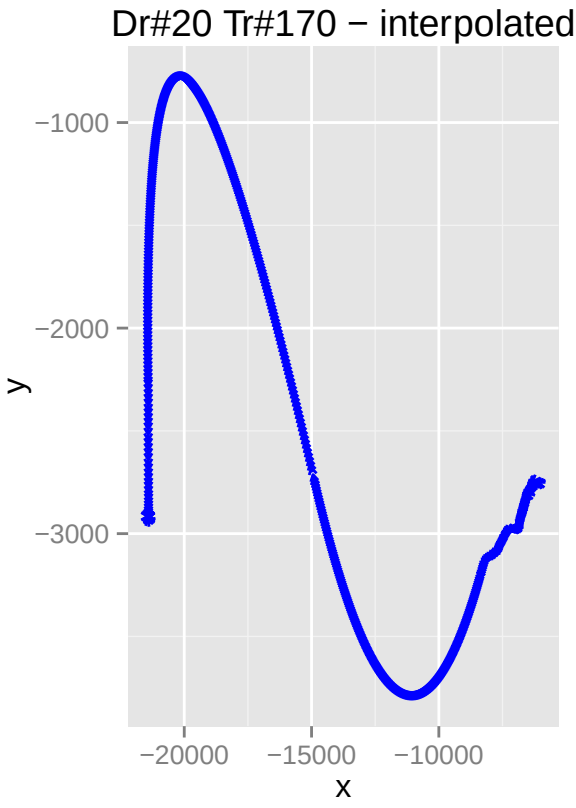
Below are several other examples demonstrating the efficacy of this method in recovering smooth, realistic-looking paths to replace missing data (notice how parabolic-curve approximation works really well, while using straight lines or circular arcs would have created much less believable trajectories):



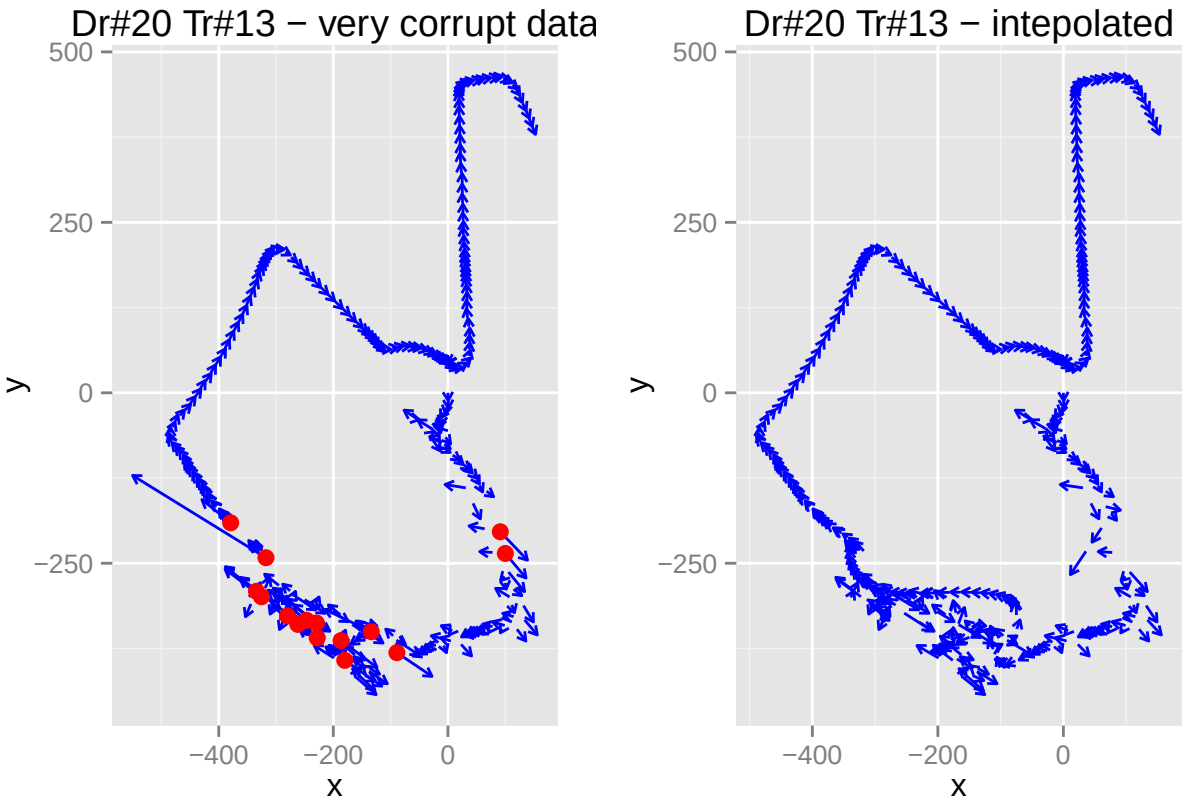








However, there are also many cases with data so corrupt that they cannot be reliably recovered:



We hence decided to limit recovery of missing data to cases with three or fewer missing sections, and discard the more seriously impaired cases. Overall, we recovered missing data for about 18,000 out of the 25,000 affected driving trip data sets.

In terms of time cost, our various data verification and cleaning steps took us about 100 hours on a single computer running on seven cores.

2. Driver Identification as Classification Problem

Following the above data verification and cleaning procedures, the question for us to address now is that, with such a database of labeled personal driving trip data (raw GPS plus derived features), whether we can effectively distinguish among different drivers' different driving styles.

2.1. Problem Framing

For each individual Driver D, we have got a collection of labeled driving trips representing his/her typical driving habits. Because we have over 2,700 such drivers in the database, for each Driver D we have also got an abundance of labeled driving trips that are ***not*** by Driver D.

With such labeled data and a "one-vs.-all" approach, we can train a discriminative classification model to distinguish the driving style of Driver D versus those of other drivers. At test time, the trained discriminative model will be given an unlabeled driving trip data set comprising GPS coordinate readings and the related derived higher-order features, and the model will be evaluated according to how well it classifies whether or not the trip is by Driver D.