In the near future, vehicles in the US will be equipped with short range telematics capability. The promise is to improve safety and efficiency for vehicular traffic. The data set in this challenge was obtained during a pilot study conducted at the University of Michigan. The data were emitted by appropriately instrumented vehicles to a number of transcievers at curves, parking lots and intersections. Essentially, as a vehicle comes within a certain distance of a transciever, a handshake commences, a temporary ID is established for the vehicle and data are transmitted nominally at 10Hz until the vehicle is out of range. These data are meant to illustrate what is called “vehicle to infrastructure” or V2I technology. The promise of this technology is that the data can be used for optimizing traffic signal timing, warning drivers about approaching a curve or intersection too fast, etc. Below is a table of the data fields. Many more are allowed in the protocol, but we limit to a few that describe vehicle dynamics.

|  |  |  |
| --- | --- | --- |
| **Field** | **Description** | **Unit** |
| GenTimeSec | Number of Seconds since Jan 1, 2004 | Seconds |
| Latitude | GPS Latitude Position | deg |
| Longitude | GPS Longitude Position | deg |
| Elevation | GPS Elevation | m |
| Speed | GPS Estimated Speed | m/s |
| Heading | GPS Heading | deg |
| Ax | Estimated Longitudinal acceleration: acceleration in a straight line | m/s2 |
| Ay | Estimated Lateral acceleration | m/s2 |
| Az | Estimated Vertical acceleration | m/s2 |
| Yawrate | Estimated Yaw rate: vehicle's angular velocity around its vertical axis | deg/s |
| ID | Mysterious string that contains id's of vehicles and transponders | Character |

As one would expect, in the real world, the data are fraught with quality issues. The data selected here are multivariate time series lasting at least 5 seconds. Some time series are quite a bit longer. There are dropouts (not every transmitted sample is received), some are “stuck at zero” or are contaminated with bad “reads”. Nevertheless, a simple “pairs” plot shows some interesting behavior:



The challenge is to develop a smoothing algorithm to estimate the true values of all ten variables such that one could infer precisely the dynamics of the vehicle from historical data. The laws of physics dictates that a vehicle should have continous first and second derivatives. Vehicles do come to a stop by breaking and accelerate in some of the time series, but they don’t crash (at least in this data set). All of the measurements here theoretically vary smoothly, but are obviously measured unevenly and noisily.

All the vehicles traveled on roads (some are in parking lots), so mapping the data will give a reasonable ground truth about where the vehicle traveled. Road data are not part of the data set however and cannot be used by your algorithm, which will be tested on data you have not seen. But that might help you figure out what happened during gaps in the data and help you determine whether you reconstructed the paths correctly.

Here is a handy R-function that implements a distance calculation for two points on a great circle, returned in meters using the Spherical Law of Cosines:

LatLon2Dist = function(Lon1, Lat1, Lon2, Lat2){

point1.lat.rad = Lat1\*pi/180

point1.lon.rad = Lon1\*pi/180

point2.lat.rad = Lat2\*pi/180

point2.lon.rad = Lon2\*pi/180

d = acos(sin(point1.lat.rad)\*sin(point2.lat.rad)+cos(point1.lat.rad)\*cos(point2.lat.rad)\*cos( point2.lon.rad-point1.lon.rad))\*6371000

return(d)

}

It allows you to calculate the distance in meters between two points expressed in degrees of longitude and latitude. You might find it useful to measure the distance between your estimates and the data in units that are compatible with the other fields.

Your task is to produce an algorithm that reconstructs curves for each of the ten variables at 10Hz sampling rate. For example, if the input time series is 20 seconds long, your output should be a ten-component time series of length 200 with equally spaced samples (every tenth of a second). You must supply the source for our execution and testing on a hold -out data set. You must also produce a write-up of the details and scientific justification for how it works.

There are 50 time series in the training set and 25 time series in our test set. Smoothing evaluation will be done by computing a root mean squared error for each component where the errors have been normalized to be unitless. RMSE’s will be summed for a total score on the error. We will also measure the smoothness of each curve. Each curve will also be inspected to assess behavior on missing data and any unrealisting behavior will be flagged as a penalty.

The final score will be based on 1) accuracy in tracking the data; 2) structural plausibility (smoothness and behavior on missing data); and 3) scientific justification in the write up, e.g. mathematical rigor and generality.