Functional Data Privacy Algorithms for User Based Insurance

Problem presented by

Steve Cowper

ControlF1

ESGI107 was jointly hosted by
The University of Manchester
Smith Institute for Industrial Mathematics and System Engineering

with additional financial support from
KTN Ltd
Natural Environment Research Council
Manchester Institute for Mathematical Sciences

Report author

Tamsin Spelman (University of Cambridge) and David Wood (University of Warwick)

Executive Summary

By monitoring each driver's driving characteristics individually, car insurance premiums can be set to directly reflect that driver's risk. Since such premiums tend to be cheaper, their uptake has increased in recent years, even outside the original market of young drivers. However, a lot of data (including GPS) is collected about each car journey in order to judge the driver's ability which raises privacy issues, with a particular concern being that every car journey can be reconstructed from that data. We looked at what to change (or remove) from the collected data so that driver's journeys couldn't be reconstructed, while retaining as much information about each driver as possible so the insurance company can still study a driver's charcateristics and potentially use the bulk data for testing new methods in future.

Control F1 does not want to use private key cryptography for customer relation reasons. We have deduced that GPS data would have to be deleted to retain privacy, however a quick experiment and a literature review suggests using heading and distance travelled data would still be sufficient to reconstruct journeys.

We considered deleting GPS data and time data and then randomising all the data points of a journey. This removes most of the information about the journey but the "estimated journey vector" constructed from the bearing and distance data will still be retained. The journey vector most accurately matched the actual GPS calculated journey vector for longer journeys and non-circular journeys. Also, particularly for longer journeys, it can be used to identify similar journeys e.g. someone's commute.

Cars slow down at junctions and traffic lights so data points are more commonly taken just before a turn. This effects the accuracy of the bearing and distance data. We studied the error in the "local estimated journey vector" caused by a junction. We suspect distance errors grow faster than bearing errors, which agrees with the analysed data.

Version 2.0
April 6, 2016
iv+10 pages

Contributors

Robert Whittaker (University of East Anglia)
Michal Kubiak (Polish Academy of Science)

Contents

1 Introduction

(1.1) The use of small devices placed in ones car to monitor driving are becoming more common, enabling insurance companies to price insurance based on driving ability rather than a more general system based on more generic information such as age, make of car etc. Since these boxes decrease premiums they are becoming more common, even outside the original market of young drivers. As such policies become more prevalent there will be an increasing number of privacy conscious people wanting these policies who are very concerned about the data that is being kept on them and what it could be used for. Such queries from customers have already started to emerge.

(1.2) A major concern is that the data collected on these devices is detailed enough that the driver's life could be tracked off their car journeys. These boxes typically create a data point every minute (or if there is a serious braking/accelerating event) and the data recorded at that moment include:
GPS Position
Time
Bearing
Distance travelled since last data point
Speed
Any sudden acceleration or de-acceleration events (categorised)

(1.3) One particular area of concern is the holding of GPS data, from which Journeys can easily be retraced. In America GPS data cannot be legally stored and such a change could potentially happen here.

(1.4) We are investigating the minimum amount of data to change (or remove) from the collected data so that driver's journeys can not be reconstructed. On the other hand we want to retain as much detail as possible so the insurance companies can still judge driving characteristics using their scoring algorithm and can also maintain a detailed historical database for testing against new algorithms.

(1.5) One simple mechanism to secure the data is public key cryptography, but Control F1 want to avoid this method due to problems explaining this to many customers. In addition, if a driver did want to reconstruct a journey (e.g. to fight a speeding ticket, prove to police their car was not where they believe it was) they would need to provide the private key, which many will have lost as it is difficult to stress that it must be kept safe. This then becomes a PR problem and a real one since such queries are already received.

(1.6) The problem was broken down into a number of different goals. We shouldn't be able to reconstruct a journey, its start and end points nor its journey vector from
(a) just one journey's information
(b) all journey data from the same driver.

A secondary goal is that
(c) we should not be able to identify the driver's ids from a given particular journey i.e. it should be difficult to recognise that two journeys are the same.

(1.7) (b) is a much harder problem than (a) since a driver's next journey will start where the previous one finished so theoretically this allows the number of data points about the drivers movements to tend to infinity. This also increases the chances of distinctive features of the journey to appear. For example, if the driver is doing 70mph, without speeding he must be on a motorway or dual carriageway which is a much smaller network than the entire UK road network. Then given the bearing data (and the fact these roads tend to wind less than other roads) that motorway/dual carriageway could probably be identified and the rest of the journey backward constructed from there.

(1.8) Question (c) is about relating journeys to each other. Even if you can not reconstruct the driver's journey, if you can identify a journey that is being done regularly (e.g. the commute) that also contains personal information. Additionally, if two journeys are identified as the same the data points could potentially be intertwined increasing the knowledge about the journey and hence increase the chance of identifying the route.

(1.9) In section 2 we consider reconstructing a journey without GPS data. In Section 3 we assume the data points at different points have been randomised. The "estimated journey vector" will still be retained so we consider how accurately it represents the actual "journey vector" and whether it can be used to relate similar journeys. Some of the errors arising from studying journeys using bearing, distance and speed data arises as cars slow down prior to turns so there is excessive sampling in these places which throws of the data's accuracy for calculating journeys. In section 4 we quantify the errors introduced by a junction.

2 Tracking a Journey without GPS Data

(2.1) To not be able to reconstruct journeys, it was clear that the GPS data would have to be removed since it is too specific. Was this sufficient to anonymise the data so journeys could no longer be reconstructed?

(2.2) To test this we naively tried reconstructing a journey experimentally. We assumed the car was travelling at a constant speed so data points were taken equal distance apart. Using a bit of string, marked at regular intervals, half the group laid out a route on a small map of Manchester. At each mark, a bearing was taken. This bearing data and the map was then passed to the other group who successfully managed to reconstruct the journey relatively quickly. The main points we noticed were:

1)For the human eye, drawing the instantaneous direction vector (bearing) at each point one after the other constructed a basic shape for the journey which could then be roughly identified on the road map.
2)Some features of the journey were easier to directly relate to roads on the map. For example, a section of two or three data points at a roughly constant bearing indicated a long straight section of road.
3)Parts of the journey where there was constant changes of direction (e.g. on small side roads) were harder to relate to specific roads. But when a major feature was identified you could work backwards to find where the car had been. This suggests less dense rural roads are easier to track a journey without GPS data. However even for an urban journey, although it might start on small back streets you tend to go onto the more major roads for the majority of the journey.

(2.3) This implied that journeys could be reconstructed using only bearing, distance and speed data, so a literature search was performed to back up this proposition. Chief among the papers found were methods of curve matching using algorithms to minimise *Frechet Distance* between a route on a map and routes garnered from sampled travel data. For two curves given as continuous maps $\pi_1 : [0, p_1] \to \mathrm{R}^d$ and $\pi_2 : [0, p_2] \to \mathrm{R}^d$, the Frechet distance is defined as

(2.4)
$$d_F(\pi_1, \pi_2) := \inf_{\substack{\alpha : [0,1] \to [0,p_1] \\ \beta : [0,1] \to [0,p_2]}} \max_{t \in [0,1]} ||\pi_1(\alpha(t)) - \pi_2(\beta(t))||_2$$

(2.5) Frechet distance is sometimes referred to as dog-leash distance, the minimum length of a leash required to connect a dog and its owner as they walk, without backtracking, along their respective curves from one endpoint to the other. Such matching algorithms are well established, and when tested on maps containing 430,000 vertices and 820,000 edges (Berlin, from Chen et. al. [2]) matches can be obtained with routes on standard desktop computers in average runtimes of minutes (also backed up by work of Wenk et. al. [5] using similar data from Athens which contained more noise due to extensive construction happening in the city at the time of the data being collected). For a good discussion of the algorithms see also Alt et. al. [1], and for one where GPS data is used (which would seem to be ripe for adapting for bearing data instead) see Rahmani et. al. [4]. This suggests that with more powerful computing equipment curve matching, even over larger areas, from journey vectors is either feasible currently, or will become much more so in future years.

3 Journey Vector

(3.1) From the previous section, it appears that even without GPS data the driver's

journey could still be reconstructed necessitating that the data must be further anonymised.

(3.2) It was then felt one of the other most informative pieces of information for journey reconstruction, which gives the insurance company relatively little information about the driver, was the time stamp. This information is mostly used to sort the data entries, to track the direction the car was travelling relative to the sun (as more accidents occur when driving into the sun) and determine how dark it was when driving (as more accidents happen at night and some premiums limit/restrict night time driving). However this useful data could be stored in extra, much less informative data entries such as: Darkness Level - Day,Transition,Night or Travel direction within 5 degrees of the sun - Yes/No. So if the time stamps were removed and replaced with these entries, then all of one driver's journey entries randomly rearranged (either amongst one journey or all journeys made by that driver) there is far less information to reconstruct journeys, while still retaining all necessary data for the insurance company.

(3.3) However what information about the journeys could you gather from these now randomised data entries? One crucial piece of information that is still retained is the "estimated journey vector" for the whole journey. At each time stamp using the bearing and distance data you can construct a vector of that journey segment. This is only an estimate since we assume the car has travelled at that bearing in a straight line the entire time since the previous time stamp. These vectors (at each time stamp) can be added (as vectors) to give an estimated journey vector for the full journey. Even if these time stamps are rearranged the estimated journey vector won't change (a standard property of vector addition).

(3.4) We want to know how much information this "estimated journey vector contains". Can it be used to identify similar journeys, or one done often e.g the commute? Is it an accurate method for identifying the direction and distance travelled? Does its accuracy vary with length of journey (as if you had all the data stamps for one driver you can theoretically construct an estimated journey vector as long as you want) or the frequency of time stamps?

3.1 Accuracy

(3.5) We looked at how accurate the estimated journey vector was at estimating the genuine journey vector. The genuine journey vector could be calculated using the GPS positions at the start and end of the journey. There are errors introduced by the GPS data but these are generally small and they only play a major role where the distance travelled between the start and end point are particularly small (e.g. a circular trip).

(3.6) Figures 1 and 2 consider the accuracy of the journey vector for data points taken every minute and every 2 minutes. The data we had was for journeys every 1 minute. Lower frequency data could be analysed by ignoring every 2nd (2nd and 3rd etc.) data point. However since most journeys were short, if we reduced the time stamps frequency significantly there was only two or three data points for a journey.

(3.7) Figures 1 and 2 show that the error between the actual and estimated journey vector was quite high. For the estimation of distance travelled, less than 50 percent of all journeys were within even 10 percent accuracy or the actual distance for both 1min or 2min time stamps. However for bearing a very high proportion (almost 40 percent) was within 10 degrees of accuracy for both 1min and 2min time steps.

(3.8) As expected when decreasing the number of time steps the accuracy decreased. However in both cases a significant minority of journeys had very large errors. One reason for this could be that they represent short journeys, and short journeys only have very few data points so any errors contribute to a proportionally larger percentage error with the journey vector. Comparing Figure fig:LongerJourneys to Figure 1 we can see evidence that indeed a much lower proportion of journeys have inaccurate journey vectors on these longer journeys. A much small minority still have large errors which we largely attribute to circular routes where you return to your starting point (e.g. dropped someone off at the train station and returning home without stopping the car). There journeys have very short journey vectors so small inaccuracies cause large errors.

(3.9) A small amount of data was available where time stamps were taken every second. This data was not from genuine customers using the boxes, but from tests of the overall system in a couple of cars. This data is shown in Figure 4 and, as expected, showed significantly more accuracy than the one or two minute data. However this most starkly illustrates that the "estimated journey vector" gives an accurate approximation of the bearing angle (to within 10 degrees) a much higher proportion of the the time than it gives an accurate estimation of the distance travelled in that direction.

3.2 Relating Two Journeys

(3.10) We have discussed in the previous section the accuracy of the journey vector. This considered the question of whether we could reconstruct information about a journey from one journey's data. However suppose there were many of these journeys with the time stamps randomised, relating to one driver. If similar journeys could be identified from the data then a commute (for example) could be identified and combining the data from all those journeys you think are the same would increase the likelihood of reconstructing it. Or similarly you might be able to identify one driver if you have a copy of

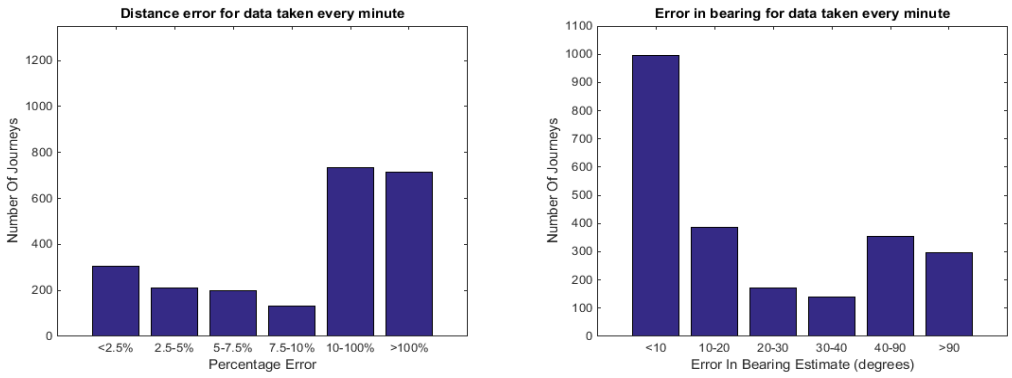


Figure 1: Accuracy of distance and bearing estimation of the journey vector for one minute time steps

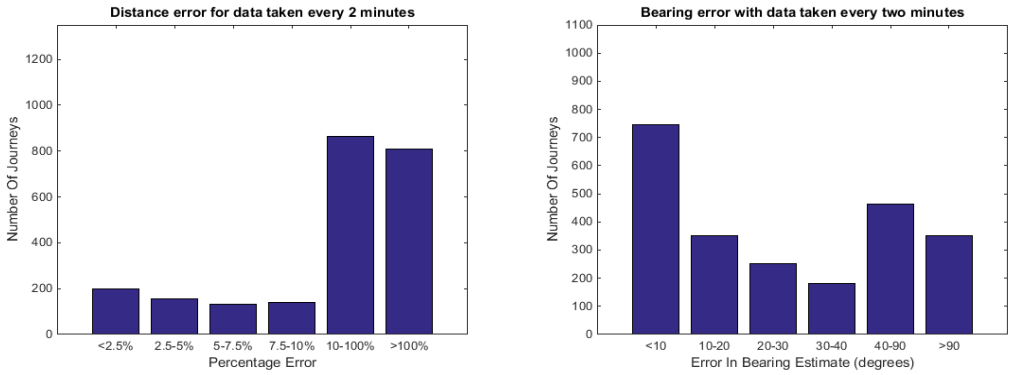


Figure 2: Accuracy of distance and bearing estimation of the journey vector for two minute time steps

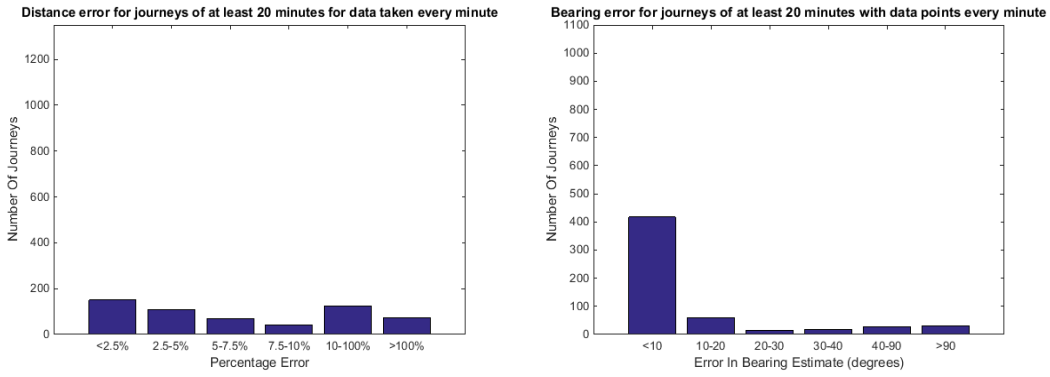


Figure 3: Accuracty of the distance and bearing data for journeys over 20 minutes long for time stamps taken at one minute time steps

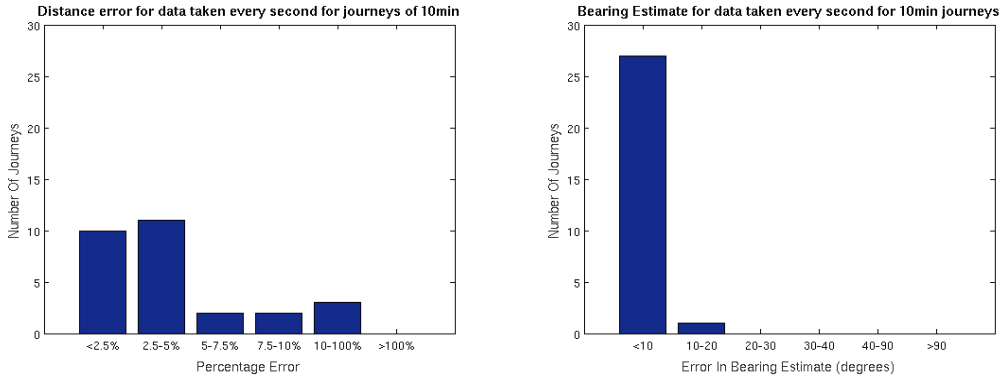


Figure 4: Accuracy of distance and bearing estimation of the journey vector for one second time steps only considering journeys over 10min in length

a journey made by one person, and it matches similar journeys but this one driver.

(3.11) We compared the journey vectors for every journey in our dataset. We defined two journeys to be the same if the length of their journey vectors was within 10 percent and the heading was within 10 degrees. We then compared these results to the GPS data to determine which journeys were actually the same and whether we had been able to identify them. Journeys under 100m in length were ignored for this test, as they are likely to be very inaccurate, and in general are not journeys we are interested in. The results are for all journeys over 100m in length is shown in Table 1 and for journeys over 10km in length in Table 2.

	Calculated Miss	Calculated Hit
True Miss	98,000	2,000
True Hit	3,000	2,000

Table 1: Vector pairs for journeys over 100m in length

	Calculated Miss	Calculated Hit
True Miss	2448	160
True Hit	200	1036

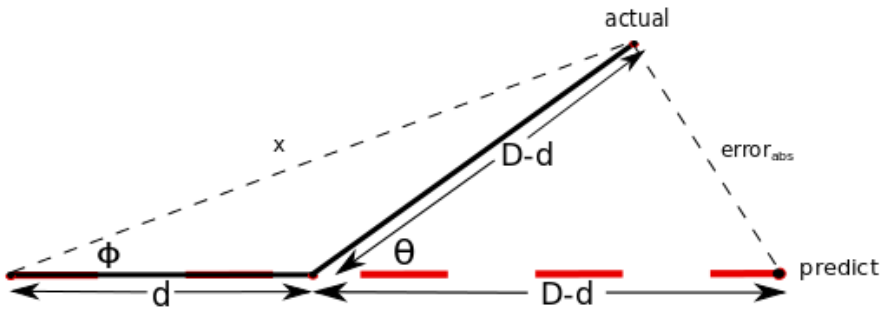
Table 2: Vector pairs for journeys over 10km in length

(3.12) From Table 1 we can see that the number of times we correctly predicted that two journeys were the same is of equal size to the number of times we predicted they were the same when the journeys were not. There were a lot of false hits here which suggest it couldn't accurately identify journeys which were the same. It also missed an even larger number that were the same which is didn't identify.

(3.13) The results were much better when only longer journeys were considered (over 10km) as seen in Table 2. Here 86 percent of the journeys we identified as the same were indeed the same journey which is far higher than before. However it still missed about 16 percent of journeys which were the same. However this is far better than before and suggests the journey vector can identify similar journeys when the journeys are long enough.

4 Junctions

(4.1) The greatest errors coming from journey vector amalgamations will be from junctions where the distance travelled after a junction will be assumed to be in the direction of the previous bearing before the junction (see figure).



(4.2) With the variables chosen some quick calculations suggest we can ignore the effects of decelerating before a junction and accelerating away so consider a car moving at constant speed before, at and after a single junction which makes an angle of θ with the incoming road (on which the latest bearing has been taken). A car travelling at an average 15 m/s (approximately 30 mph) will therefore travel 900m within a 60 second window between readings. In the general case we assume the car travels a distance D between readings, a distance d between the preceding data point and reaching the junction, and so a distance $D - d$ afterwards on the new bearing. If we let x denote the actual journey vector, then the true bearing between data points is denoted ϕ and the distance between the actual position and calculated position is $\text{error}_{\text{abs}}$ where:

(4.3)

$$\begin{aligned} \text{error}_{\text{abs}} &= s(D-d)\sin(\theta/2) \\ x^2 &= D^2 - 2d[(1-\cos\theta)(D-d)] \\ \cos\phi &= \frac{d+(D-d)\cos\theta}{[D^2-2d(1-\cos\theta)(D-d)]^{1/2}} \end{aligned}$$

(4.4) Clearly, largest errors occur when d is small compared to D (furthest distance travelled on "wrong" bearing) and as θ increases (which could easily include angles greater than 90 degrees at junctions that "double back". No further investigation was carried out during the meeting, but it may be useful to use these expressions to study how such errors accumulate over several time steps and multiple "realistic" junctions.

5 Conclusion

(5.1) Practical experimentation and a literature search makes us confident that given bearing and distance readings in the order that they are taken can be thought of as being sufficient to be able to reconstruct sufficiently long journeys using curve matching algorithms and suitably processed maps in acceptable time limits, and so if bearing and distance information is to be retained then it cannot be done so in this manner.

(5.2) If the GPS data is removed and the time stamps for each journey removed and the data entries randomised, then the "estimated journey vector" is still retained. This "estimated journey vector" can be an accurate estimate of the real journey vector but only if either data is taken very frequency (like in the every second data) or the journeys are long enough. The journey vector could be used to identify if two journeys are the same but due to the number of false positives, would only be accurate for longer journeys (greater than say 20km). As such it is limited in the cases where it could be used to reconstruct or identify journeys.

(5.3) Greatest errors in estimated journey vectors will most likely be introduced at junctions, in particular junctions making a large angle with the preceding road and readings that occur most recently to the arrival at a junction. Estimates on error have been derived which may be studied for the effect of cumulative errors arising from multiple junctions over a journey.

References

[1] *H.Alt et. al.* Matching planar maps *J. Algorithms* 49 (2003) 262-283

[2] *D. Chen et. al.* Approximate Map Matching with respect to the Frechet Distance *ALENEX* (2011) 75-83

[3] *S. Brakatsoulas et. al.* On Map-Matching Vehicle Tracking Data *Transport Res. Rec.* (2005) 853-864

[4] *M. Rahmani and H.N.Koutsopoulos* Path inference from sparse floating car data from urban networks *Transport Res. C* (2013) 41-54

[5] *C. Wenk et. al.* Addressing the Need in Map-Matching Speed: Localizing Global Curve-Matching Algorithms *SSDBM* (2006) 379-388