Path Estimation from GPS Tracks

Chris Brunsdon

Department of Geography
University of Leicester
University Road, Leicester LE1 7RU
Telephone: +44 116 252 3843
Fax: +44 116 252 3854
Email: cb179@le.ac.uk

1. Introduction

The widespread availability of hand-held GPS units has led to a proliferation in data on the tracks of individuals as they walk, drive or otherwise go about journeys. This data has been used in a number of ways - for example the OpenStreetMap project (The OpenStreetMap Foundation 2007). One characteristic of projects such as this is that there will often be several GPS tracks for the same stretch of road. In general, repeatedly measuring something and taking the average of measurements leads to a more accurate result. The question addressed here is "is it possible to 'average' GPS tracks and if so, does this lead to a better estimate of road location?".

2. Tracing Paths From GPS Data

In this section, the key technique for identifying paths from GPS data will be introduced. This approach is called *Principal Curve Analysis* (Hastie and Stuetzle 1989). Here, the basic principal curve algorithm will be used, as well as some proposed modifications to address specific issues relating to the estimation of cartographic data.

2.1 Description of the Data

The GPS data considered here is tracking data recorded in the GPX format. The track data coordinates were transformed from longitude and latitude to OS national grid coordinates [1] to allow comparison. Although the track data can be treated as line objects (with each line corresponding to a track), the technique outlined in the next section only requires the point information in each of the tracks. The points considered in this way will be referred to as a *point cloud*. The point cloud recorded by the author is shown in figure 1, and consists of 342 points.

   Using approximate bearings, and starting from the northernmost point, there is a short walk south-west (beside *Waterloo Road*), then a longer walk (south-east, along a footpath *New Walk*), then another long section (south west again, along *University Road*) and a final short walk south east into Leicester University's campus.

---

[1]Proj4 string: `+proj=tmerc +lat_0=49 +lon_0=-2 +k=0.999601 +x_0=400000 +y_0=-100000 +ellps=airy +towgs84=446.448,-125.157,542.060,0.1502,0.2470, 0.8421,-20.4894 +units=m +no_defs`

North –>

Scale: 200m

Figure 1: A point cloud showing recorded GPS tracks in Leicester.

2.2 Principal Curve Analysis

The idea here is to find a curve running through the 'middle' of the point cloud. One way of defining the 'middle curve' is to say that it is the curve minimising the total squared distances to each point in the point cloud. If we consider the point cloud to be a list of n coordinate pairs $\{\mathbf{p}_i = (x_i, y_i) : i = 1, n\}$, and our curve as a parametrised curve $\mathbf{f}(\lambda) = (f_x(\lambda), f_y(\lambda))$, then the distance between $\mathbf{p}_i$ and $\mathbf{f}$ is the closest point to $\mathbf{p}_i$ on $\mathbf{f}$:

$$\mathrm{D}(\mathbf{p}_i, \mathbf{f}) = \mathrm{argmin}_\lambda ||\mathbf{p}_i - \mathbf{f}(\lambda)|| \qquad (1)$$

and the 'middle' curve $\hat{\mathbf{f}}$ mimimises the expression $\sum_i \mathrm{D}^2(\mathbf{f}, \mathbf{p}_i)$. Curves satisfying these requirements are known as *principal curves*. It is noted that the parametrisation of $\mathbf{f}$ is not unique — $\mathbf{f}(\lambda)$ could be replaced by $\mathbf{f}(g(\lambda))$ where $g(.)$ is any monotone function. To resolve this ambiguity, it is specified that λ should be the distance travelled along the curve $\mathbf{f}$ — in this case, λ is therefore the distance travelled along the 'middle curve' of the GPS point cloud.

In order to estimate $\mathbf{f}$ (Hastie and Stuetzle 1989) attempt to reconstruct the curve at a number of discrete points $\{\mathbf{f}(\lambda_i) : i = 1...n\}$ where i corresponds to the index for the points in the point cloud. Given an initial guess at $\mathbf{f}$, the curve is reconstructed using the following method:

1. Find the nearest points to each $\mathbf{p}_i$ on the guessed curve, and compute the distance along the curve to each point. This provides a set of estimates for the λ_i's

2. The estimate of $\mathbf{f}$ is then updated by updating estimates for the two functions $f_x(\lambda)$, $f_y(\lambda)$ using a non-parametric smooth regression procedure (such as Cleveland (1979) or Green and Silverman (1994)) applied respectively to the (λ_i, x_i) and (λ_i, y_i) pairs.

3. Return to step 1 with the updated estimate of $\mathbf{f}$

The result of applying the algorithm to the point cloud data used here is shown in figure 2. The principal curve is shown in red - the black lines correspond to the distances of each individual point in the cloud to the line. Note a key difference between this technique and standard non-parametric regression — here the distances to be minimised can be in any direction, depending on the line joining a GPS point and the closest point to it on the principal curve, while in standard regression techniques, distances are always measured in the same direction — parallel to the y-axis

2.3 Adapting the Algorithm

Figure 2 demonstrates the general accuracy of the principal curve algorithm, but also highlights one of its pitfalls. At times GPS tracks can exhibit systematic errors - at the north end of New Walk, there is clearly a rogue track which veers noticeably from the true location - it may be seen as a 'dog leg' swinging away from New Walk, apparently crossing the railway line having passed through some buildings. This means that although in most places the curve provides a good estimate of the road or pathway, it swings out in locations near to the rogue track.

A way of overcoming this is to devise a robust variant on the principal curve algorithm - the approach is outlined below:

1. Fit a principal curve using the standard approach.

2. Note the distances from each point to the curve - call these $\{d_i\}$

3. Standardise these distances by dividing by their standard deviation - call these $\{d_i^*\}$.

4. Compute a set of weights as a monotone decreasing function of the d_i^*'s - typically let $w_i = 1$ if $d_i^* < 3$, $w_i = 4 - d_i^*$ if $3 < d_i^* < 4$ and $w_i = 0$ if $d_i^* > 4$.

North –>

Scale: 200m

Figure 2: Principal curve (red) fitted to GPS point cloud data.

North –>

Scale: 200m

Figure 3: Robust principal curve (green) fitted to GPS point cloud data.

5. Re-run the principal curve algorithm, but use weights $\{w_i\}$ in the nonparametric regression stage.

The result of applying this modified algorithm to the point cloud is shown in figure 3. From this, it is clear that the influence of the rogue track has been greatly reduced, and the estimated path now correctly follows the footbridge over the railway and main road.

3. Assessing the Quality of Principal Curves

As stated earlier, two ways of assessing the quality of the estimated curves are proposed - in terms of accuracy and precision. Accuracy — the ability of the curve to reproduce the 'ground truth' of path location — may be carried out visually using figures 2 and 3. Here, particularly with the robust modification, the results are encouraging.

Precision is essentially a measure of the reliability of the estimated paths, given that the logged GPS tracks are samples of locations on the actual paths containing some random error. Here, it is proposed to use a bootstrapping approach (Efron 1981; Effron 1982) to estimate confidence bands around the paths. Briefly, this method estimates the sampling distribution of an arbitary statistic s from a data set $\{X_i : i = 1...n\}$ by randomly sampling n items from the data set *with replacement* a number of times. Effectively we estimate the true distribution of the X_i's as a mass point distribution in which each individual value X_i has a probability of $\frac{1}{n}$ of occurring.

Here, s is not a number, but a line on a map. However, the bootstrap idea can still be applied. This is done here for both the standard and robust curves. The results are visualised by drawing each bootstrap sample of the principal curve on the same map, but using alpha blending (Porter and Duff 1984) when drawing the curves. In figure 4 the bootstrap paths for the standard method are shown in red, and those for the robust method are shown in blue. In general, the robust method has lower sampling variability.

4. Conclusions

A method has been proposed to reconstruct road or footpaths from GPS point cloud data, based on the idea of principal curves. The method has been assessed visually for both accuracy (by comparing the fitted paths against OS Landline data) and for precision (by considering bootstrap samples of the point cloud data). The results are encouraging — particularly for the robust curve estimation algorithm, suggesting that this may provide a viable method for automatic path detection from GPS data. This could be incorporated, for example, in the JOSM map data editing program (Scholz 2007) associated with the OpenStreetMap project.

One characteristic of the algorithm is that there are as many estimated λ_i values and $\mathbf{f}$ points as there are points in the point cloud - so that as the size of the point cloud increases, the principal curves contain an increasing number of points. For this reason, a final pass of the principal curve could be applied, to 'thin out' the number of points. This could be achieved, for example, with the Douglas-Peuker algorithm (Douglas and Peuker 1973). An example of this is shown in figure 5, with the thinned-out principal curve shown in red, and 100 bootstrap samples of this curve shown in blue.

A final issue to address is the comparison of the principal curve fitting algorithm used here with others, such as Einbeck, Tutz, and Evers (2005). For now, it seems that the algorithm in use at least satisfies the need to detect paths from GPS data, however further work may throw light on more efficient, precise or accurate approaches.
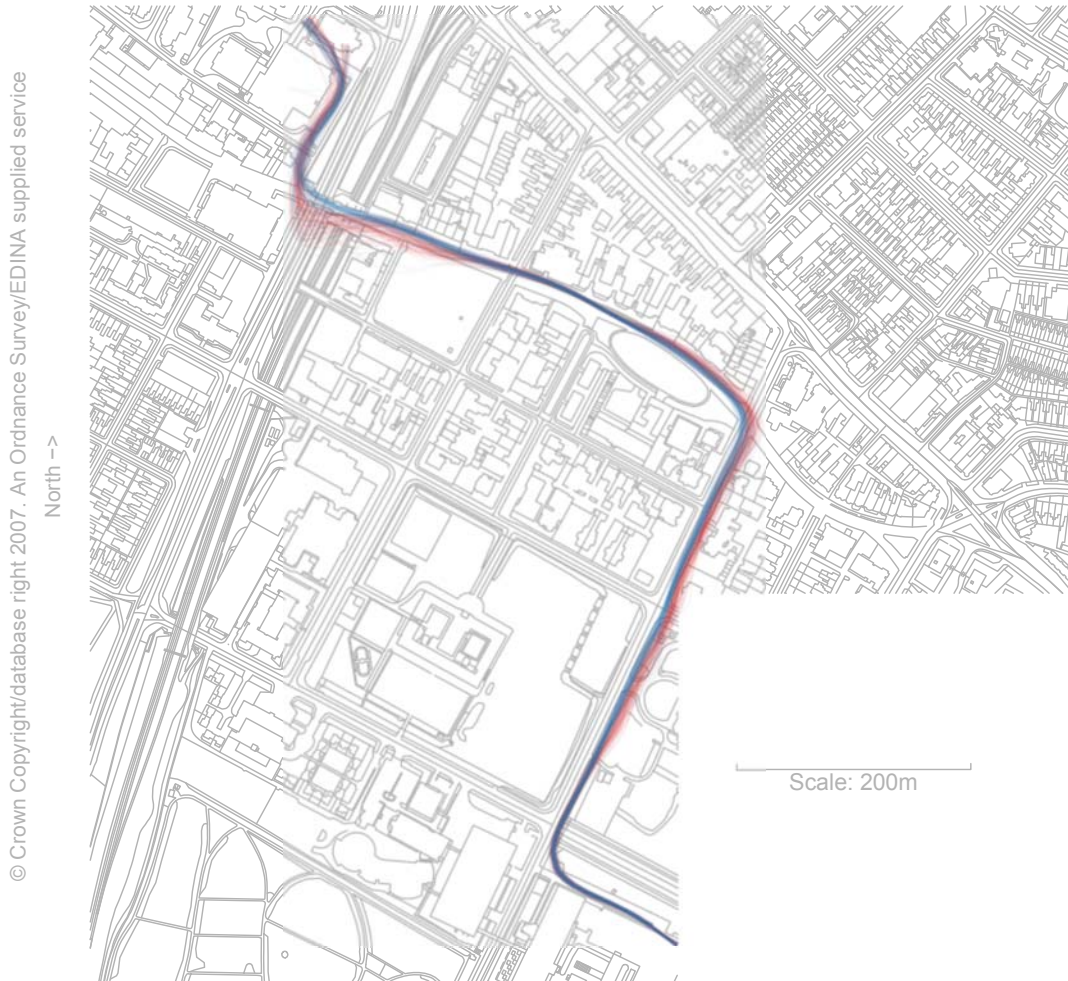
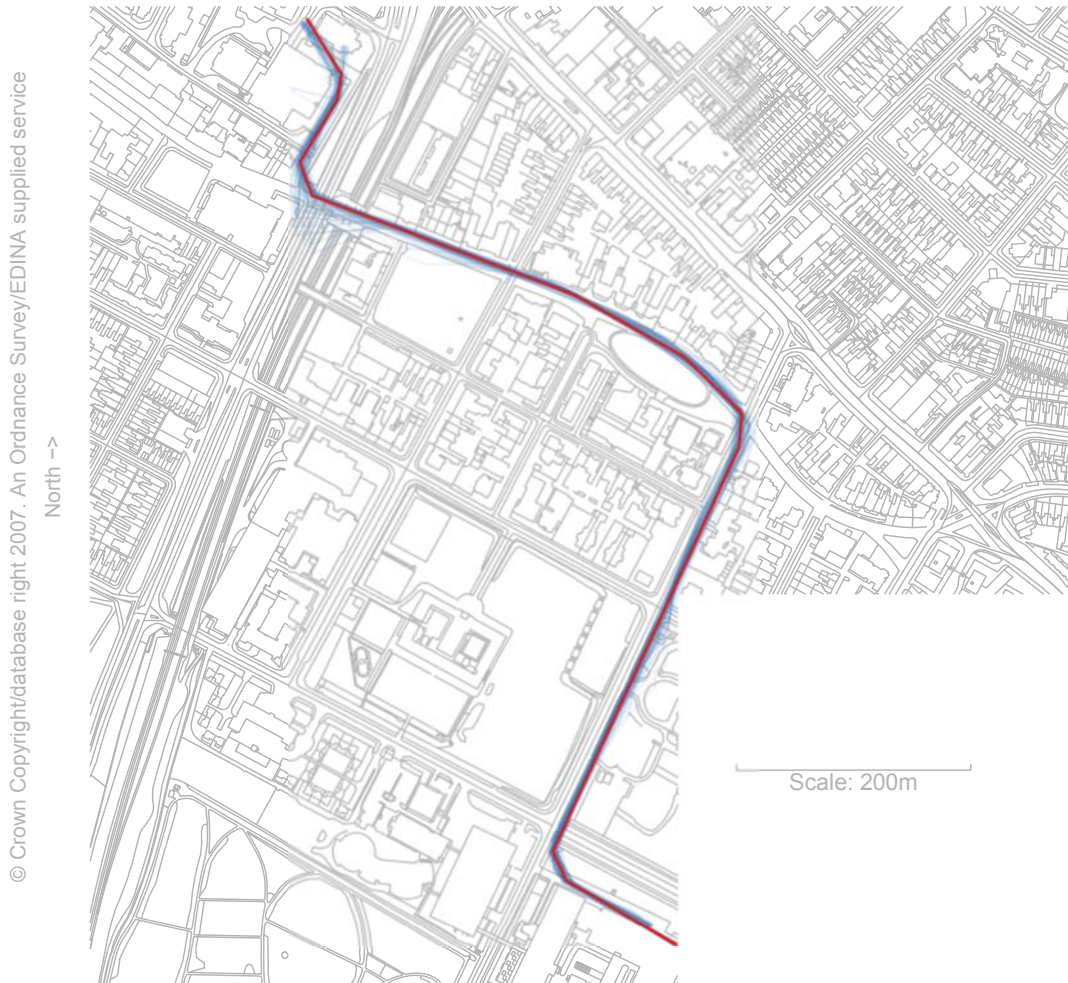Figure 4: Bootstrap sampling variability visualisations for standard (red) and robust (blue) principal curves.

Figure 5: A Generalized principal curve (red) and its bootstrap sampling variability estimate (blue).

5. References

Cleveland, W. S, 1979, Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

Douglas, D and Peuker, T, 1973, Algorithms for the reduction of the number of points required to represent a digitised line or its caricature. *The Canadian Cartographer 10*, 112–122.

Effron, B, 1982, *The Jacknife, the Bootstrap and Other Resampling Plans*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Efron, B, 1981, Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika 68*, 589–599.

Einbeck, J, Tutz, G, and Evers, L, 2005, Local principal curves. *Statistics and Computing 15*, 301–313.

Green, P and Silverman, B, 1994, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Hastie, T. J and Stuetzle, W, 1989, Principal curves. *Journal of the American Statistical Association 84*(406), 502–516.

Porter, T and Duff, T, 1984, Compositing digital images. *Computer Graphics 18*(3), 253–259.

Scholz, I, 2007, JOSM — OpenStreetMap. Web Site. `http://josm.eigenheimstrasse.de/` (Viewed April 3, 2007).

The OpenStreetMap Foundation, 2007, OpenStreetMap. Web Site. `http://www.openstreetmap.org` (Viewed April 3, 2007).