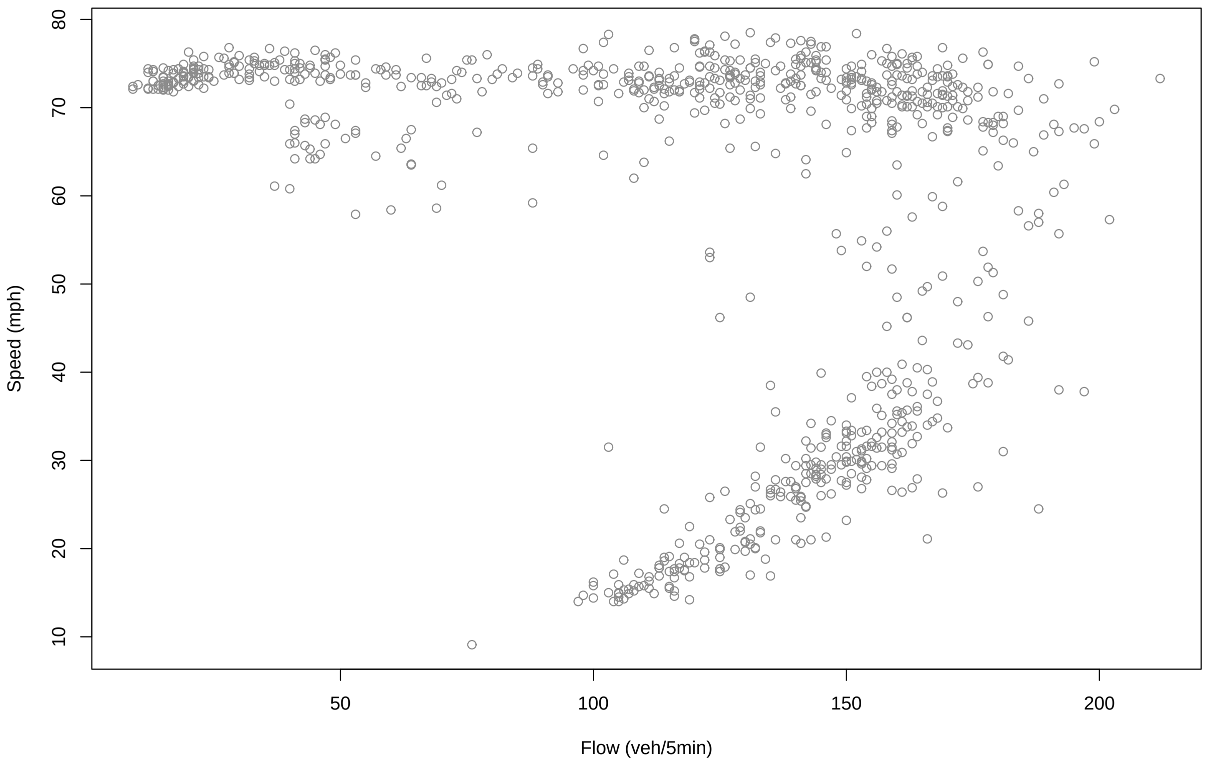
Analyzing traffic data with function-free smoothing methods

Approaches and challenges

Jochen Einbeck
jochen.einbeck@durham.ac.uk
Durham University

Cemapre, Lisboa, 20th of August 2007

Speed-Flow data



(data from: PemS)

- The plot shows average speed and flow aggregated over 5-minute-intervals on freeway I105-W in Los Angeles, California, collected from 10th July 2007 (00.00) to 13th July 2007 (23.59).
- The upper branch corresponds to uncongested traffic and the lower branch to congested traffic.

Collection of traffic data

- Data are collected through *loop detectors*, i.e. buried coils of wire, whose induction is altered when a vehicle drives over it.
- The flow is the number of vehicles that go over the loop per unit time (usually, 30 sec).
- The occupancy is the fraction of time that vehicles are over the detector. Occupancy is approx. proportional to density, which is the the fraction of the road covered by a vehicle.
- The speed can only be calculated through the time a vehicle of average length needs to pass the detector completely and is the less precise of all measurements.

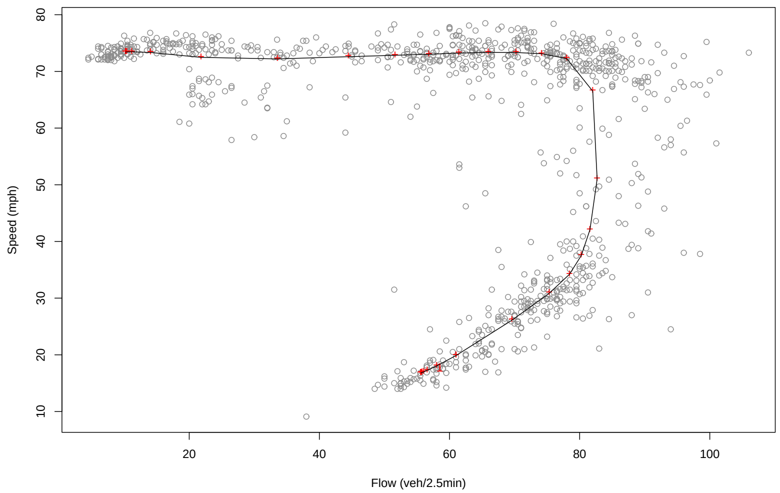
Speed-Flow data modelling

- Traffic speed prediction is of practical interest for topical issues as road pricing, journey time prediction, navigation systems etc.
- However, each value of traffic flow is associated to two different speeds. Hence, speed cannot be modelled as a *function* of flow.
- Hence, the literature has concentrated on descriptive analyses, and on finding mathematical models for *flow given speed*.
- Speed-flow data have scarcely been considered from a statistical point of view.
- They also require innovative approaches as
  - usual (parametric or nonparametric) regression models fail.
  - concepts on "switching regression" cannot be applied, as for a given point it is unknown to which regime it belongs.
  - speed-flow diagrams can differ strongly and there is no agreement on a suitable parametric model for the branches. A nonparametric modelling approach seems desirable.

Two modelling approaches

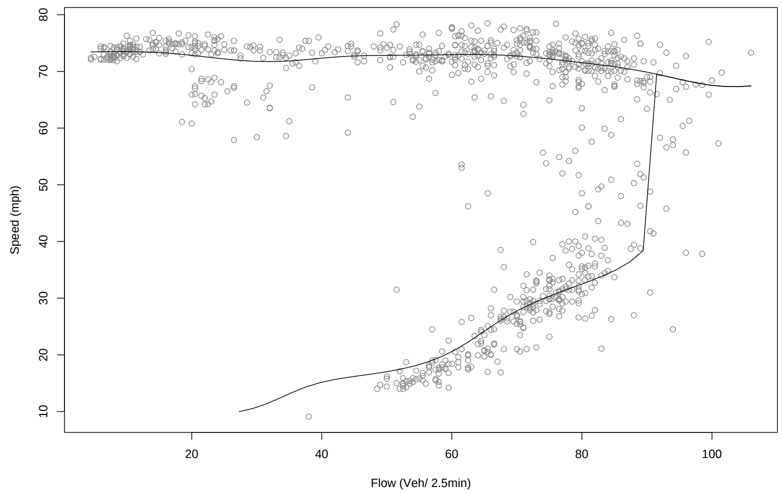
- Consider speed and flow as symmetric, and as a joint function $\begin{pmatrix} q(t) \\ v(t) \end{pmatrix}$ of some underlying parameter t.

  $\Longrightarrow$ Principal curves.



- Consider flow as the independent and speed as the dependent variable, i.e. $v = M(q)$ with some multifunction M.

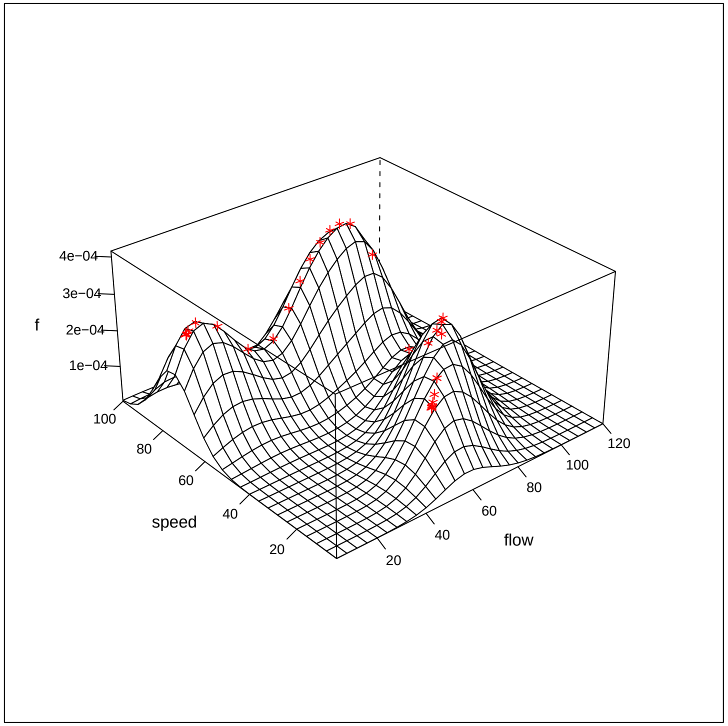
  $\Longrightarrow$ Multi-valued nonparametric regression.



Curve Fitting through Density Ridges

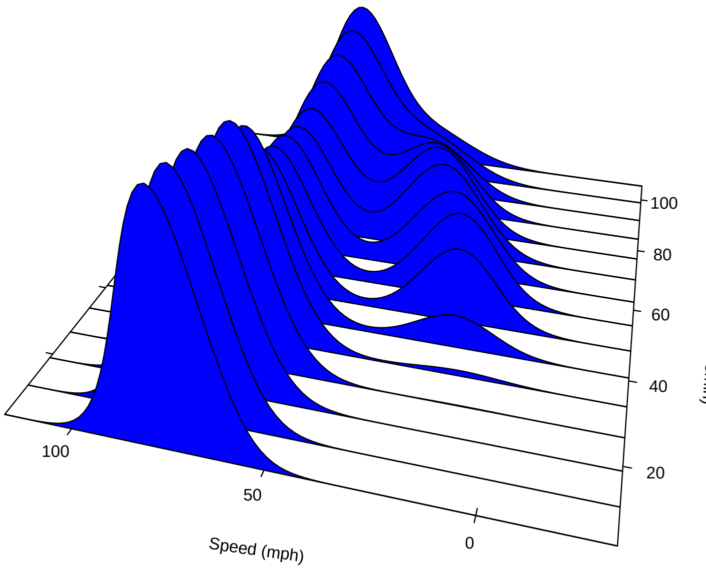
Principal curves

- Follow the ridge of a kernel density estimate $\hat{f}(q, v)$



Multi-valued regresion

- Follow the ridge made up by the conditional modes, i.e. the maxima of the estimated *conditional* densities $\hat{f}(v|q) = \frac{\hat{f}(q,v)}{\hat{f}(q)}$:



Ridge Estimation: Mean shift

- The mean shift $\mu(x)$ is the difference between a point x in $\mathbb{R}^d (d = 1, 2)$ and the local center of mass $m(x)$ of the points in its neighborhood, i.e.

$$\mu(x) = m(x) - x \equiv \frac{\sum_{i=1}^n K_h(X_i - x) X_i}{\sum_{i=1}^n K_h(X_i - x)} - x$$

(K_h: kernel weights with bandwidth h).

- Comaniciu & Meer (2002) showed that

(A) $\mu(x) \sim \nabla \hat{f}_h(x)$ where $\hat{f}_h(x)$ is a kernel density estimate.

(B) The sequence $m_{(0)} = x;\ m_{(k+1)} = m(m_{(k)})$ converges to a neighboring maximum of $\hat{f}_h$.

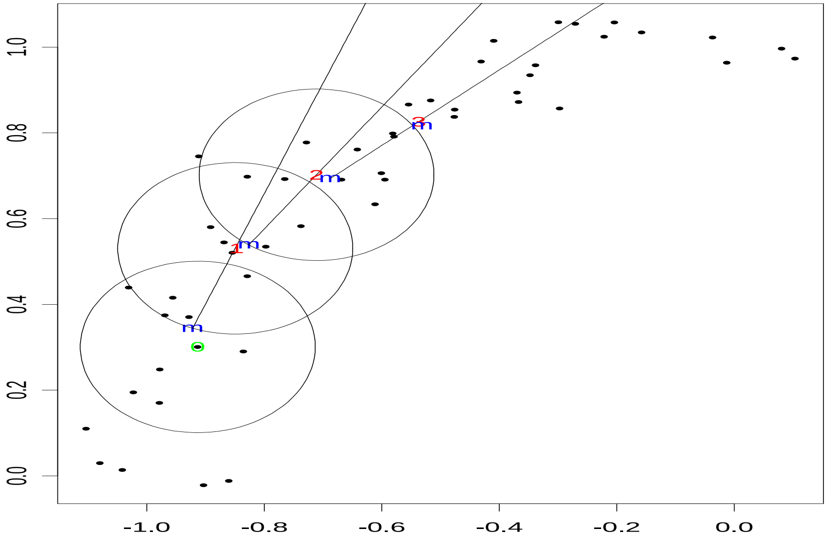
- Hence, iterating the local centre of mass $m(\cdot)$ leads us to the next available mode.

Curve Fitting Algorithms

Principal curves:

Starting from starting point (q_0, v_0), iterate between calculation of

1. a local centre of mass m
2. a localized first principal component (1,2,3,. . .)



Local principal curves, Statistics and Computing, Einbeck, Tutz & Evers (2005)

Multi-valued regression:

- For each q, define two starting points v_1 and v_2. Then, for both points, run the mean shift conditional on q until convergence to a conditional modes of $v|q$.
- One can show that this is equivalent to setting

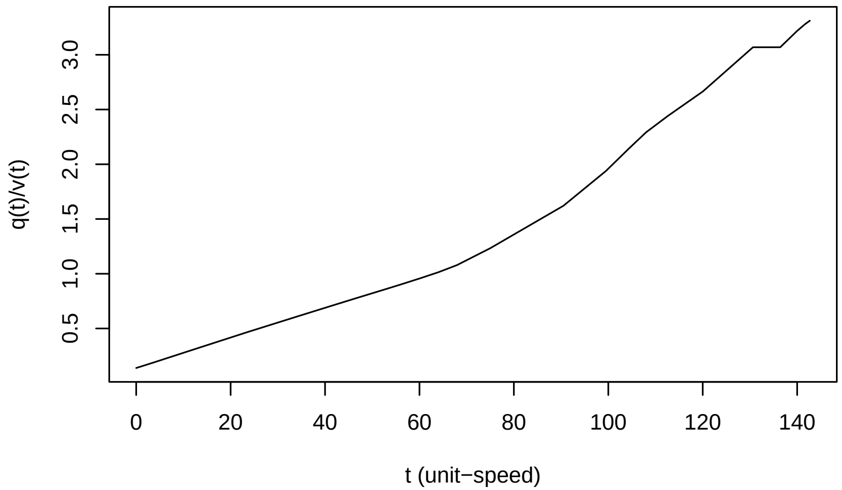
$$\frac{\partial \hat{f}(v|q)}{\partial v} = 0$$

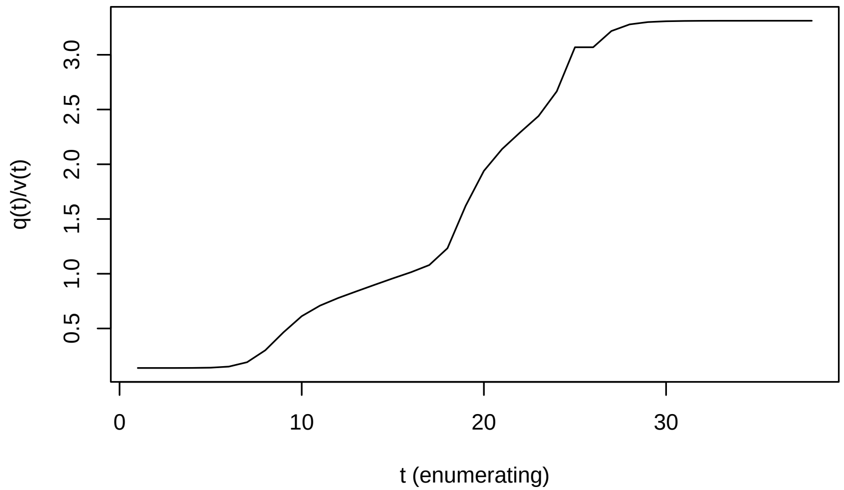
and solving w.r.t. v.

Multimodal regression, JRSSC, Einbeck & Tutz (2006)

What is the value of such curves?

- For principal curves we observe that, for any parameterization t, the traffic density $d(t) = q(t)/v(t)$ is a monotone function of the parameter t.

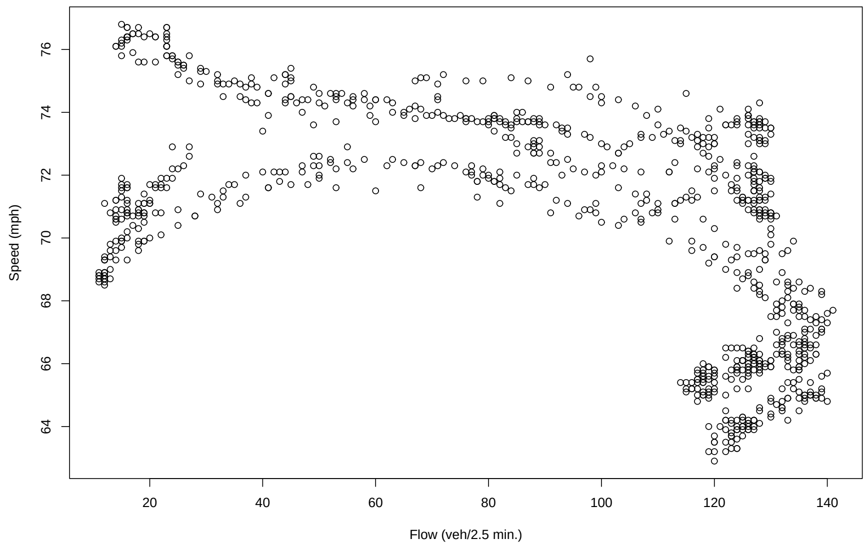




- This implies that, using such a calibration curve, one can use a principal curve to predict q and v simultaneously given d.
- However, one cannot use principal curves to predict v given q - here one needs the multi-valued regression approach.

Other variables involved?

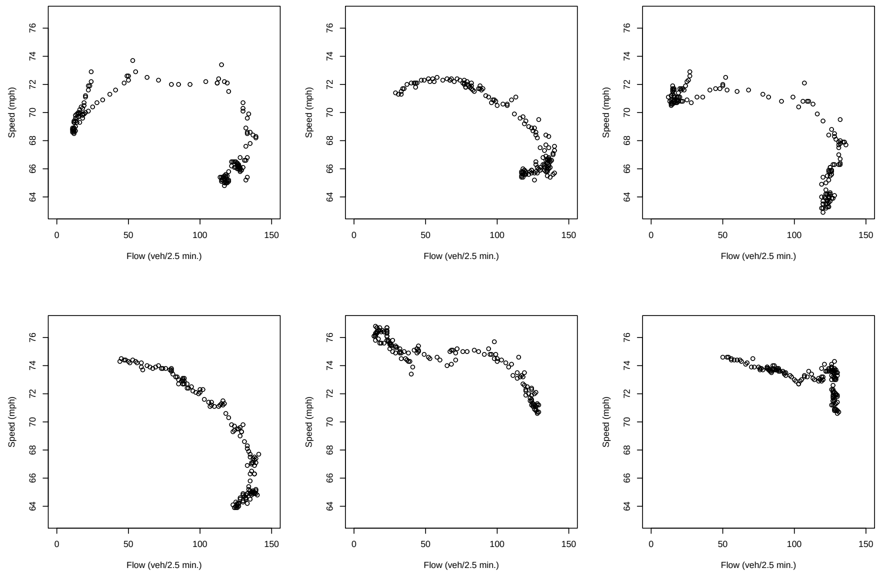
- Data from 12th (0.00) to 15th of July 2007 (23.59), Freeway SR22-E, California.



- Obviously, there are some further (observed or unobserved) variables involved, e.g. weather condition, road works, etc.
- Hence, both concepts have to be extended to allow for additional variables.

For the time being....

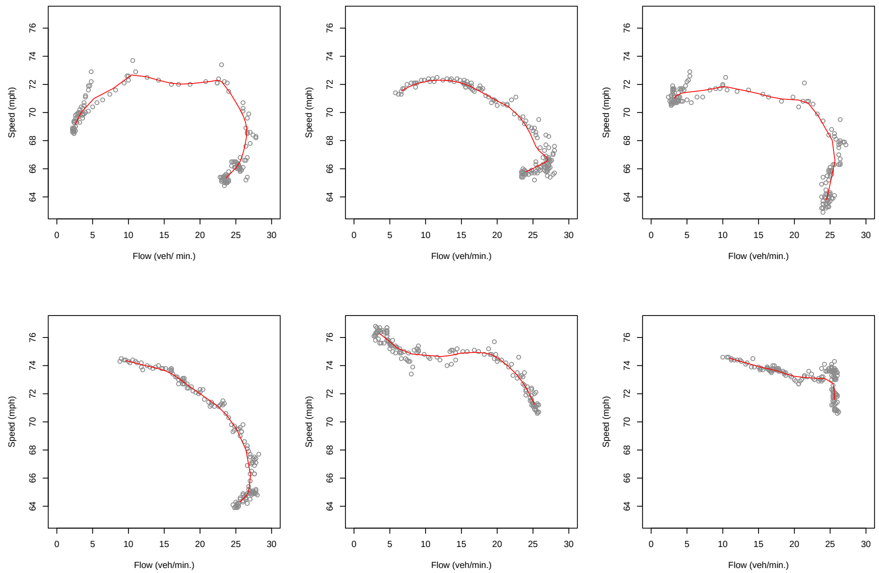
- As most variables will have an approximately constant value over a certain time span, they can be substituted by a time variable
- For example, divided into six 12-hour intervals, one obtains



- Principal curves or regression curves can then be fitted separately.

For the time being....

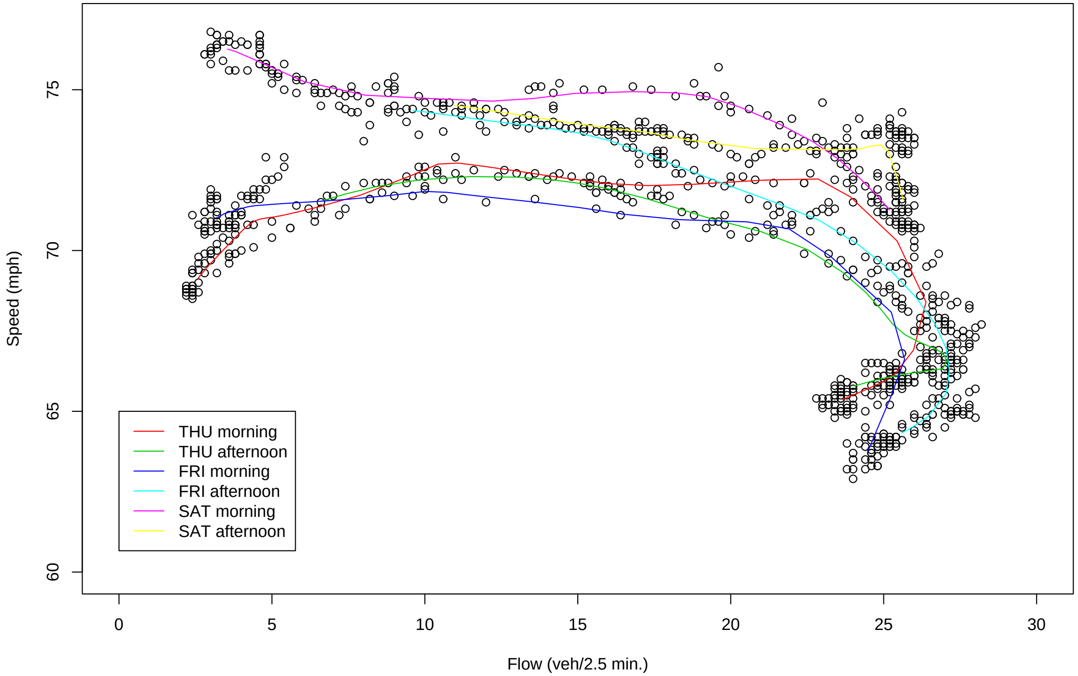
- As most variables will have an approximately constant value over a certain time span, they can be substituted by a time variable
- For example, divided into six 12-hour intervals, one obtains



- Principal curves or regression curves can then be fitted separately.

For the time being....

- As most variables will have an approximately constant value over a certain time span, they can be substituted by a time variable
- For example, devided into six 12-hour intervals, one obtains



Software and Literature

- Principal Curves
  - Hastie, T., & Stuetzle, W. (1989): Principal curves. *JASA* **84**, 502–516.
  - ⋮
  - Einbeck, J., Tutz. G. & Evers, L. (2005): Local principal curves. *Statistics and Computing* **15**, 301-303.
  - LPC Software at `http://www.maths.dur.ac.uk/~dma0je/lpc/lpc.htm`.
- Multi-valued nonparametric regression
  - Einbeck, J., & Tutz, G. (2006): Modelling beyond regression functions: an application of multimodal regression to speed-flow data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **55**, 461-475.
  - **R** function `modalreg` in **R** package **hdrcde** version 2.07 (Hyndman & Einbeck, 2007).