Modelling beyond Regression Functions: an Application of Multimodal Regression to Speed-Flow Data

Jochen Einbeck*
National University of Ireland
Department of Mathematics
Galway, Ireland

Gerhard Tutz†
Ludwig Maximilians Universität
Institut für Statistik
80799 München, Germany

Abstract

For speed-flow data, which are intensively discussed in transportation science, common nonparametric regression models of the type $y = m(x) + \text{noise}$ turn out to be inadequate since simple functional models are unable to capture the essential relationship between predictor and response. Instead a more general setting is required, allowing for multifunctions rather than functions. The proposed tool is conditional modes estimation which, in the form of local modes, yields several branches that correspond to the local modes. A simple algorithm for computing the branches is derived. This is based on a conditional mean-shift algorithm and is shown to work well in the considered application.

Key Words: Conditional density, multi-valued regression, smoothing, speed-flow curves.

1 Introduction

Speed-flow diagrams have been widely used and discussed in traffic engineering. Fig. 1 shows two speed-flow diagrams for a 4-lane Californian uninterrupted highway ("freeway"); only lanes 2 and 3 are shown here (data from University of California, Berkeley). Speed is measured in miles per hour, and traffic flow in vehicles per lane per hour. Each plotted point is an average speed and hourly flow rate for data collected over a 30-second interval. The question of interest is how can the shape of the data cloud be explained. For uncongested traffic, there is no significant association between traffic flow and speed - this is the large cluster at the top of the plots. However, when the traffic gets too dense, there may

*jochen.einbeck@nuigalway.ie
†tutz@stat.uni-muenchen.de

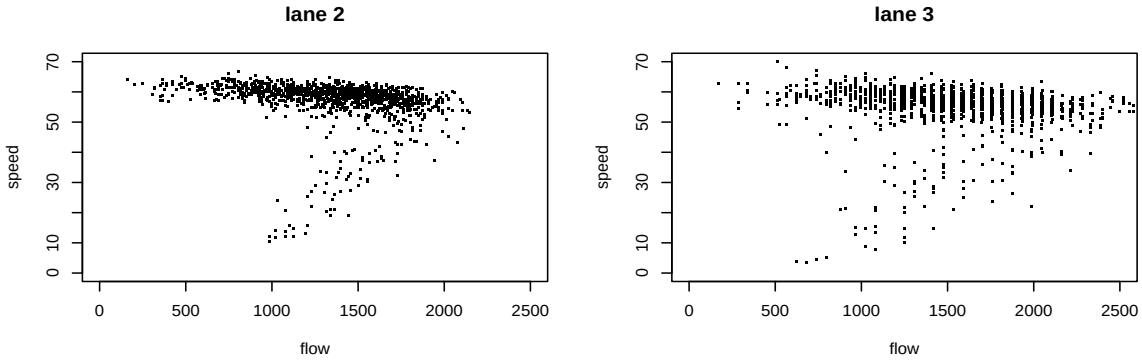

Figure 1: Speed-flow diagram for Californian freeway.

still be high traffic flow, but with a considerably diminished speed due to congestion. The less dense cloud of data points at the bottom of the figures corresponds to this situation.

Looking at Fig. 1, it is clear that the speed v cannot be adequately described as a function $v(q)$ of the flow q. Thus, any attempts at modelling data of this form have been based on modelling the traffic flow as a function $q(v)$ of speed. For example, a Greenshields-type model (Greenshields, 1935) as given in the Highway Capacity Manual 2000 (HCM, 2000) has the form

$$q = q_0 \left[\frac{v_f - v}{v_f - v_0} \right]^{1/\beta}.$$

In this equation, q_0 is the maximum flow, v_0 is the speed $v(q_0)$ at maximum flow, and v_f is the free-flow speed, assuming that the vehicle is alone on the highway. The constant β is specific for the type of the highway, e.g. $\beta = 1.31$ for a multi-lane highway. For an overview of available models see Li (2005). Direct modelling $v = v(q)$, the natural quantity of interest, has hardly been considered, as mathematical modelling is generally based on functional forms. It is obvious that for this kind of data standard regression models are inappropriate. Regression modelling that allows prediction of speed given traffic flow has to account for the specific variability of the data which is characterized by the two branches.

There has been considerable effort over recent decades to understand data of this type. In the eighties most studies concentrated on just reporting the graphical relationship between flow and speed (see Hall & Hall (1990) and Hall, Hurdle & Banks (1992) for an overview on this literature), while in the last decade the research interest focussed on finding mathematical models for the data as in Daganzo (1995), Del Castillo & Benitez (1995), or more recently Li & Zhang (2001), to name a few. However, there are few instances of using statistical tools to analyze speed-flow diagrams. An early approach in this direction was given by Drake, Schoefer & May (1967). Kockelman (2001) applied mixture models of congested and uncongested conditions to flow-density relations.

Traffic speed prediction is becoming an increasingly important issue, e.g. to construct

and support Intelligent Transportation Systems (ITS), enabling drivers to obtain their expected arrival time (Xiao, Sun & Ran, 2003). Huang & Ran (2003) noted that traffic flow may be considered as one of the most important variables in this model, though an accurate model for the speed will also involve other variables, e.g. weather conditions. However, citing the latter authors, 'it is very difficult to use traffic flow directly ... because each traffic flow might be corresponding to two different traffic speeds', and for a given value of flow, 'it is *impossible* to decide whether it is congested or not' (If it was possible, one could simply apply methods for nonparametric estimation of branching curves, see Silverman & Wood (1987)). The common – suboptimal – way out is to use other variables instead of flow, e.g. time. As a first step towards a more adequate solution of this problem, we present in this paper an approach to modelling speed given flow directly which may also be used for general non-standard regression problems with several branches of data. Our approach is based on a new and simple method to find the conditional modes of a multivariate distribution.

2 Conditional Modes and Densities

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote an iid. sample from a population $(X, Y) \in \mathbb{R}^2$ with joint density $f(x, y)$, where Y is a scalar response variable. A common way of deriving regression functions is by considering the minimization problem

$$m(x) = \arg\min_a E(l(Y - a)|X = x), \tag{1}$$

where $l(\cdot)$ denotes a loss function. The quadratic loss function $l(z) = z^2$ yields the most widely used regression function $m(x) = E(Y|X = x)$, whereas $l(z) = |z| + (2p - 1)z$ yields regression quantiles (see Koenker (2005) for an overview), and in the special case $p = 0.5$ the conditional median. An early reference to local median smoothers is Härdle & Gasser (1984). If the loss function is taken as $l(\cdot) = -\delta(\cdot)$, where $\delta(\cdot)$ is the delta-function, i.e. $\delta(x) = 0$ for $x \neq 0$ and $\int \delta(x)\, dx = 1$, one obtains the conditional mode $\text{Mode}(Y|X = x)$. As observed by Berlinet, Gannoun & Matzner-Løber (1998, 2000) in the context of nonparametric regression, and Hyndman (1995) and Matzner-Løber, Gannoun & Gooijer (1998) in the context of nonparametric forecasting, the conditional mode has clear advantages when the conditional distributions (the forecast densities, resp.) are multimodal. To refer to the resulting curve Scott (1992) uses the term modal regression curve or trace.

The mode differs from the mean and the median in one important aspect. While the conditional mean and median always represent a single value, the conditional mode is not

necessarily unique, as the maximum density $f_{Y|X}(a|x)$ might be achieved for more than one value. Moreover, a conditional density function can have several conditional maxima at different levels, which may be interpreted as *local modes*, being defined by

$$\text{local Mode}(Y|X=x) = \arg\max_{a\in U} f_{Y|X}(a|x) \tag{2}$$

where U (in the unidimensional case) is a closed interval and the maximum is taken over the interior of the interval. It is the multiplicity of local modes which makes them attractive for the analysis of data such as speed-flow diagrams. When the conditional distribution of the data is multimodal, then the data cannot be described properly by a function. The underlying structure represents a relation rather than a function. It is assumed that the underlying relation $R \subset \mathbb{R}^2$ decomposes into several branches, which are defined by the operators

$$M_{(j)}(\cdot) = j^{\text{th}}\,\text{local Mode}(\cdot),$$

where $j = 1, \ldots, p$ is a suitable enumeration of the branches (e.g. from bottom to top). The underlying relation has the form

$$R = \{(x, M_{(j)}(Y|X=x)); x \in \mathbb{R}, j = 1, \ldots, p\},$$

which may also be described by the multifunction

$$M_f(x) = \{M_{(j)}(Y|X=x) | 1 \le j \le p\}. \tag{3}$$

[A mapping $M : A \longrightarrow B$ is said to be a multifunction if $M(x) \subset B$ for all $x \in A$. For details on multifunctions, see standard books on set-valued analysis, e.g. Aubin & Frankowska (1990).]

As for the function $m(\cdot)$ in (1), which is usually assumed to be "smooth", i.e. at least once continuously differentiable, one has to impose some smoothness on $M_f(\cdot)$. Since conditional mode estimators are edge-preserving, they tend to have jumps. So, we only require each branch to be smooth except at a finite set of points.

Equation (2) provides some motivation for estimating the conditional mode(s) as the local maxima of the conditional density estimator, which in the case of univariate predictors takes the form

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{f}(x)} = \frac{\sum_{i=1}^n K_1\left(\frac{X_i - x}{h_1}\right) K_2\left(\frac{Y_i - y}{h_2}\right)}{h_2 \sum_{i=1}^n K_1\left(\frac{X_i - x}{h_1}\right)}, \tag{4}$$

with kernels K_1, K_2 and bandwidths h_1, h_2 (Samanta & Thavaneswaran, 1990, Hyndman, Bashtannyk & Grunwald, 1996). Extensions to local linear estimators have been considered by Fan, Yao & Tong (1996), Hyndman & Yao (2002), and Fan & Yim (2004). Plotting the conditional densities at different values of the flow x, using the graphical tools

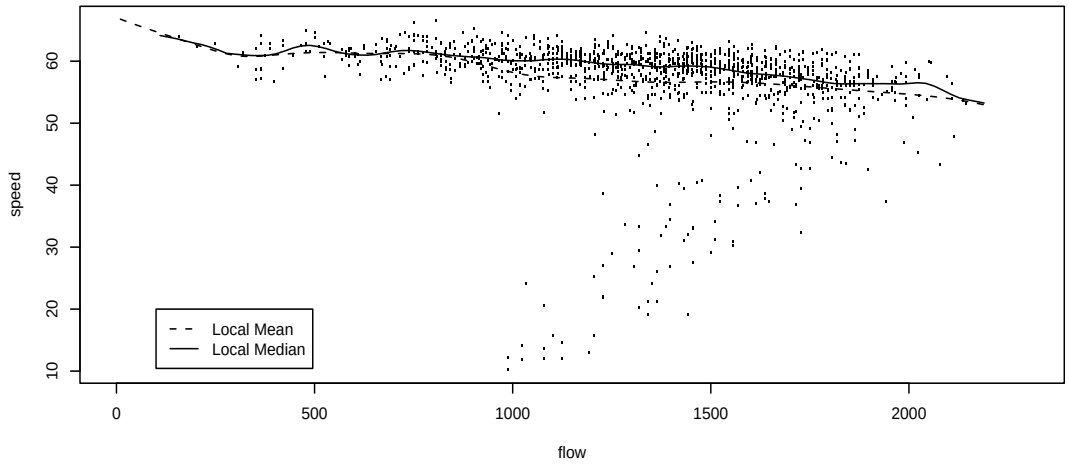

Figure 2: Speed-flow diagram for lane 2 with local (linear) mean and median smoother.

developed by Hyndman, Bashtannyk & Grunwald (1996), gives Fig. 3 ($h_1 = 100, h_2 = 4, K_1, K_2$: Gaussian) which provides a direct view of the conditional maxima of the distribution.

Obviously there is a wide range of flow values reaching from about 1000 to 1600 vehicles/hour where the conditional distribution of speed given flow is multimodal. Considering exemplarily the conditional distribution of speed at flow=1400, one can predict the speed as follows: the expected speed will be *either* around 30 mph *or* around 60 mph, where *60 mph is more likely*. From this we see that there are two problems which need to be addressed:

a) How can the conditional modes be estimated? This is the topic of the next section.

b) How does one quantify that one estimated mode is more likely than another one? This issue will be treated in Section 4.

3 Estimating the conditional modes

Estimation of the maxima of a density function is quite an old problem and has been considered by a large number of authors, starting with early publications from Parzen (1962) and Nadaraya (1965). However, there has been comparatively little work in the literature on the estimation of the maxima of a *conditional* density function. In the previous section we suggested using the value(s) maximizing the conditional kernel density estimate as estimator(s) for the conditional mode(s). Samanta & Thavaneswaran (1990) and Berlinet, Gannoun & Matzner-Løber (1998) show that this estimator, also called the *sample conditional mode*, is consistent and asymptotically normally distributed under

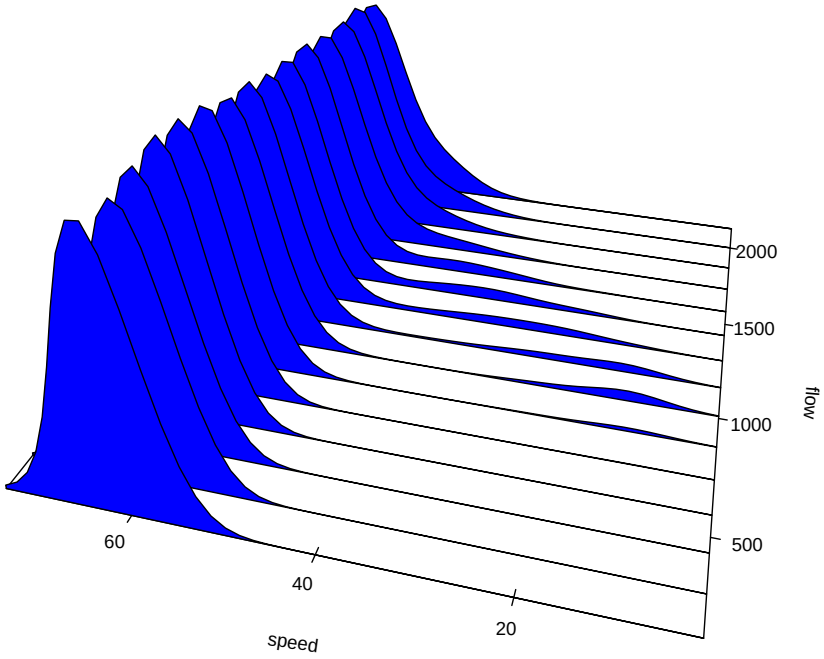

Figure 3: Conditional densities for speed-flow diagram (lane 2).

suitable regularity conditions. A separate issue is *how* to find the maxima of the conditional density estimates.

Mehra, Ramakrishnaiah & Sashikala (2000) apply smoothed rank nearest neighbor estimators, but constrain themselves to the case of a unique conditional mode. Alternatively, a simple grid search based on the conditional density estimates (4) could be applied. However, searching for maxima on a grid is computationally extremely demanding, especially when the space of predictors is multivariate, and implementation is not straightforward if *all* conditional local maxima have to be found, rather than just the global one. Hence, more sophisticated methods are needed. Scott (1992) uses averaged shifted histograms. Carreira-Perpiñan (2000) uses an EM algorithm to model X and Y jointly as a Gaussian mixture and then proposes a gradient ascent algorithm for mode finding starting from the locations of the centroids. Based on the observation that maxima of the conditional density (4) have the property $\hat{f}'(y|x) = 0$, we propose a simpler alternative concept. Let us assume that K_2 belongs to a special class of radially symmetric kernel functions satisfying

$$K_2(\cdot) = c_k k((\cdot)^2),$$

with c_k being a strictly positive constant. The function $k(\cdot)$ is called the *profile* of K_2. We work with a slightly more general setting than in equation (4) and analyze the conditional density estimator

$$\hat{f}(y|x) = \frac{c_k}{h_2} \sum_{i=1}^{n} w_i(x) k\left(\left(\frac{Y_i - y}{h_2} \right)^2 \right),$$

where $w_i(x)$ is some weight function, usually a kernel function, not depending on y. (The extension to multivariate predictors is straightforward and only affects the definition of the weights $w_i(x)$.) From the estimation equation

$$\frac{\partial \hat{f}(y|x)}{\partial y} = \frac{2c_k}{h_2^3} \sum_{i=1}^{n} w_i(x) k' \left(\left(\frac{Y_i - y}{h_2} \right)^2 \right) (y - Y_i) = 0$$

the mode estimator y_m is given by

$$y_m = \frac{\sum_{i=1}^{n} w_i(x) k' \left(\left(\frac{Y_i - y_m}{h_2} \right)^2 \right) Y_i}{\sum_{i=1}^{n} w_i(x) k' \left(\left(\frac{Y_i - y_m}{h_2} \right)^2 \right)}. \tag{5}$$

Note that the dependence of $y_m \equiv y_m(x)$ on x is suppressed for notational ease. Let

$$g(\cdot) = -k'(\cdot)$$

and consider g as a kernel profile belonging to a kernel function $G(\cdot) = c_g g((\cdot)^2)$. When K_2 is the Gaussian kernel, then G is Gaussian as well. The kernel K_2 has been named the *shadow* of G by Cheng (1995). By use of G one obtains the equation

$$y_m = \frac{\sum_{i=1}^{n} w_i(x) G \left(\frac{Y_i - y_m}{h_2} \right) Y_i}{\sum_{i=1}^{n} w_i(x) G \left(\frac{Y_i - y_m}{h_2} \right)}. \tag{6}$$

In the case of conditional mode estimation, one has the weights

$$w_i(x) = \frac{K_1 \left(\frac{X_i - x}{h_1} \right)}{\sum_{j=1}^{n} K_1 \left(\frac{X_j - x}{h_1} \right)}. \tag{7}$$

The resulting equation $y_m = \mu(y_m)$, with

$$\mu(y_m) = \frac{\sum_{i=1}^{n} K_1 \left(\frac{X_i - x}{h_1} \right) G \left(\frac{Y_i - y_m}{h_2} \right) Y_i}{\sum_{i=1}^{n} K_1 \left(\frac{X_i - x}{h_1} \right) G \left(\frac{Y_i - y_m}{h_2} \right)}, \tag{8}$$

cannot be solved analytically, but the solution y_m can be obtained iteratively by calculating a series of local means. An important tool is the so-called *mean shift*

$$\mu(y) - y,$$

which for a mode y_m takes the value zero. This is the basic idea of the mean shift algorithm, which has been recently studied in the unconditional case (i.e. $w_i(x) \equiv 1$) by Comaniciu,

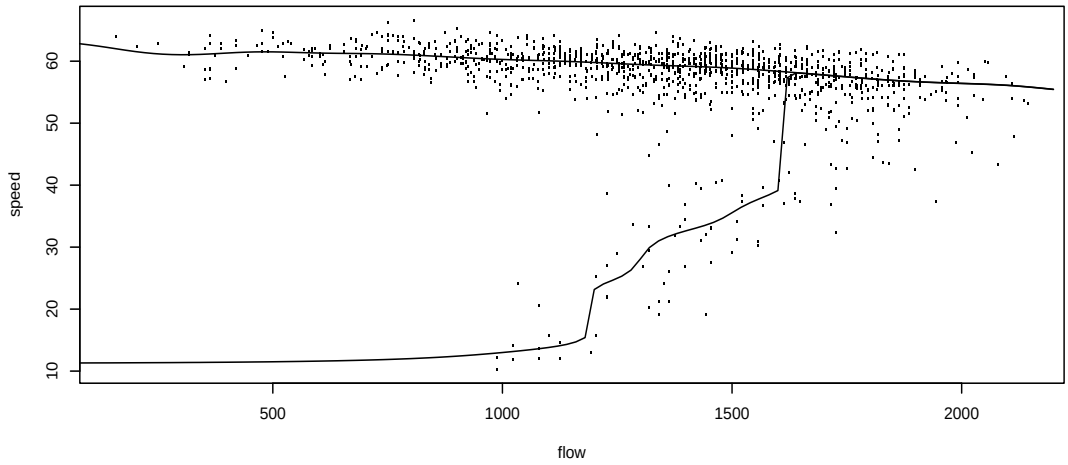

Figure 4: Multimodal regression for speed-flow data based on mean shift.

Ramesh & Meer (2001), Comaniciu & Meer (2002) and Comaniciu (2003). For a given starting point y_0, Cheng (1995) showed that the sequence $(y_\ell)_{\ell=1,2,\ldots}$ defined by

$$y_{\ell+1} = \mu(y_\ell) \tag{9}$$

converges to a nearby mode y_m, which is a fixed point of (9). To account for multimodal conditional distributions, one applies the mean shift procedure as follows: For a given x,

1) Choose a set of starting points $y_0^{(1)}(x) < \ldots < y_0^{(P)}(x)$.

2) For $j = 1, \ldots, P$:
   Set $\ell = 0$. Iterate

$$y_{\ell+1}^{(j)}(x) = \mu(y_\ell^{(j)}(x)) \tag{10}$$

   until convergence is reached, resulting in estimates $\hat{y}_m^{(1)}(x), \ldots, \hat{y}_m^{(P)}(x)$.

3) The estimator for $M_f(x)$ is the random set

$$\hat{M}_f(x) = \{\hat{y}_m^{(1)}(x), \ldots, \hat{y}_m^{(P)}(x)\}.$$

The set $\hat{M}_f(x)$ is ordered, i.e. $\hat{y}_m^{(1)}(x) \leq \ldots \leq \hat{y}_m^{(P)}(x)$. This follows immediately from the properties of the mean shift, as the series of local means converges to a nearby conditional mode (see Comaniciu & Meer, 2002, Theorem 1). This ordering makes it easy to identify the branches. Note that $\hat{M}(x)$ may actually be a multiset, because some modes might have been reached more often than once. This will certainly occur when P exceeds the number p of branches. However, it is also possible when P is equal to or smaller than the number of branches, as some modes may not have been found, while other modes may be included several times in the multiset. To be certain that all modes are discovered, one has

to use a sufficiently large number of starting points. Each point gives an iteration process, which will find a conditional mode within its *basin of attraction*. If one may assume that the data are bimodal (as in the speed/flow example), it is sufficient to start one mean shift procedure from the bottom and one from the top of the distribution of the data. Then each one will find its corresponding mode automatically. For more than two modes, there are a variety of methods to detect the number of modes of a univariate distribution (Silverman, 1981, Muller & Sawitzki, 1991) or to test for multimodality (Fisher, Mammen & Marron, 2002). Recently, Davies & Kovac (2004) presented another tool for controlling the modality based on the taut string method. However, most of these methods are either not directly applicable to the case of a *conditional* distribution, or they suffer from some drawbacks (see Hartigan, 2000). If there is uncertainty about the number of branches, we recommend simply using more starting points than is strictly necessary. This will certainly mean that some modes are reached two or more times, but if the mean shift is iterated until convergence, all estimates belonging to the same mode will be approximately equal.

Fig. 4 shows the results of a multimodal regression using the above algorithm. Note that the fitted curve corresponds nicely to the schematic 'generalized speed-flow relationship' depicted in Fig. 2 in Hall, Hurdle & Banks (1992). The conditional mean shift is calculated with Gaussian kernels K_1 and K_2 ($h_1 = 100, h_2 = 4$) and weights as in (7). The starting points are constant w.r.t. x, i.e. $y_0^{(1)}(x) \equiv y_0^{(1)} = \min\{Y_1, \ldots, Y_n\}$ and $y_0^{(2)}(x) \equiv y_0^{(2)} = \max\{Y_1, \ldots, Y_n\}$. In other situations, when there is some prior information about the shape of the underlying relation, it might be useful to work with variable starting points. Though the mean shift algorithm has been accused of being quite slow (Comaniciu & Meer, 2002), we almost always observed convergence after about 30 iterations. Of course, this is still much faster than calculating all conditional densities and performing a grid search to find the maxima.

Remark 1.

The right side of (8) is already well-known: This is exactly the formula for the sigma filter, firstly applied in Lee (1983) for digital image smoothing. However, in contrast to the mean shift algorithm, which iterates (6) or (8) until the mode is found, the sigma filter only runs the first loop of this iteration. Consequently, the sigma filter can be seen as a one-step approximation to the conditional mode. An important property of the sigma filter is that it is edge-preserving (Chu, Glad, Godtliebsen & Marron, 1998). The sigma filter exploits the fact that the conditional mode has better edge-preserving properties than the conditional mean, and therefore the close relationship of sigma filtering and the mean shift is not surprising.

Remark 2.

The use of $l(\cdot) = -\delta(\cdot)$ in the minimization problem (1) yields the conditional mode. In

practice, the delta function has to be approximated. This is possible by means of the kernel function K_2, since

$$\delta(\cdot) = \lim_{h_2 \longrightarrow 0} \frac{1}{h_2} K_2 \left(\frac{\cdot}{h_2} \right).$$

Applying this for a fixed h_2 gives

$$l(\cdot) = -\frac{1}{h_2} K_2 \left(\frac{\cdot}{h_2} \right) \equiv -\frac{c_k}{h_2} k \left(\left(\frac{\cdot}{h_2} \right)^2 \right) \tag{11}$$

and minimizing (1) yields the equation

$$a = \frac{E\left(G\left(\frac{Y-a}{h_2} \right) Y | X = x \right)}{E\left(G\left(\frac{Y-a}{h_2} \right) | X = x \right)}. \tag{12}$$

Thus, the right-hand side of (8) estimates the right-hand side of (12). Comaniciu & Meer (2002) show that (11) corresponds (in the unconditional case) to *location M-estimation*. Chu, Glad, Godtliebsen & Marron (1998) make use of this relationship by exploiting local M-estimators for edge preserving smoothing and show improved performance of this estimator compared to the sigma filter.

Remark 3.

It is well known that local linear smoothers perform distinctly better than local constant smoothers, as pointed out by Fan (1992) and Hastie & Loader (1993). Fan, Hu & Truong (1994) showed that the nice properties of local linear estimators carry over to any smoothing problem of type (1) based on a convex loss function $l(\cdot)$ with a unique minimizer at 0. Although the negative δ-function obviously does not share this property, one might still wonder if the local constant conditional mean shift, which we used, might be further improved to give a local linear mode estimator. This simply requires replacing the weights $w_i(x)$ in (6) by the corresponding weights for a local linear fit (Fan & Gijbels, 1996, p. 20, Fan & Yim, 2004), namely

$$w_i(x) = \frac{K_1\left(\frac{X_i - x}{h_1} \right) [S_{n,2} - (X_i - x) S_{n,1}]}{\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1} \right) [S_{n,2} - (X_j - x) S_{n,1}]}, \tag{13}$$

where $S_{n,\ell} = \sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1} \right) (X_i - x)^\ell$. An example demonstrating the difference between the local constant (7) and local linear (13) settings is provided in Fig. 5. It is clear that the disadvantages of local constant mean estimators carry over to local constant mode estimators. In particular, they are heavily biased at the boundary and for clustered designs (Fig. 5 left). Although the local linear mode estimator corrects the deficiencies of the local constant mode estimator in this example, it can be recommended only for the case of functional dependence, i.e. where the mode is unique. In cases where the

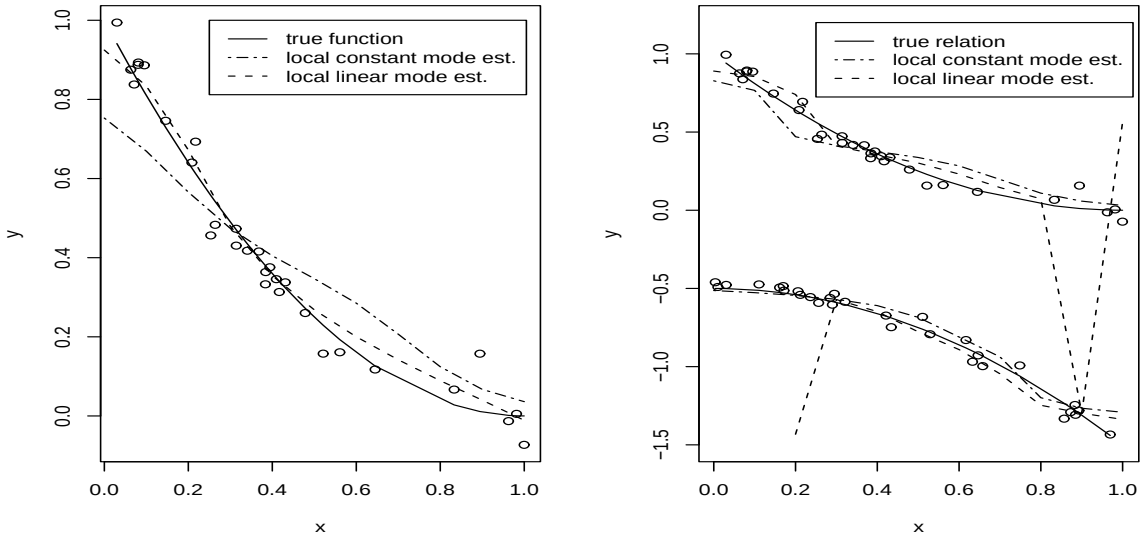

Figure 5: Comparison of a local constant and a local linear conditional mode estimator based on mean shift (left: $h_1 = 0.2, h_2 = 0.4$; right: $h_1 = h_2 = 0.2$)

data structure is multimodal (Fig. 5 right), the local linear mode estimator behaves quite erratically and gives non-smooth results. Thus, for the remainder of this paper we will restrict attention to local constant mode estimators.

4 Assigning Probabilities

A crucial point is the evaluation of the relevance of a conditional mode. Intuitively, the probability mass inside the basin of attraction of a conditional mode (in other words: the probability mass between the neighboring valleys surrounding the mode) is a useful measure for the relevance of a mode. Fig. 6 illustrates this concept for the speed-flow data given a flow of 1400 vehicles/hour. The area between the left border and the valley contains an estimated probability of 0.077, and the second mode corresponds to the probability 0.923. Thus, one obtains

$$\hat{M}(1400) = \begin{cases} 32.65 & \text{with est. prob.} \quad 0.077 \\ 59.18 & \text{with est. prob.} \quad 0.923. \end{cases}$$

To estimate these probabilities, it is necessary to find the low points of the valleys and to integrate over the estimated conditional densities between them. Without too much effort one can do the search for the minimum and the integration simultaneously. For a given (local) mode y_m at x, one descends from the (local) maximum $f(y_m|x)$ in small steps of length δ, say, to the right (steps $k = 0, 1, 2, \ldots$) as well as to the left (steps $k = -1, -2, \ldots$), and augments the integral in each step by $\delta \cdot f(y_m + k\delta|x)$ until the

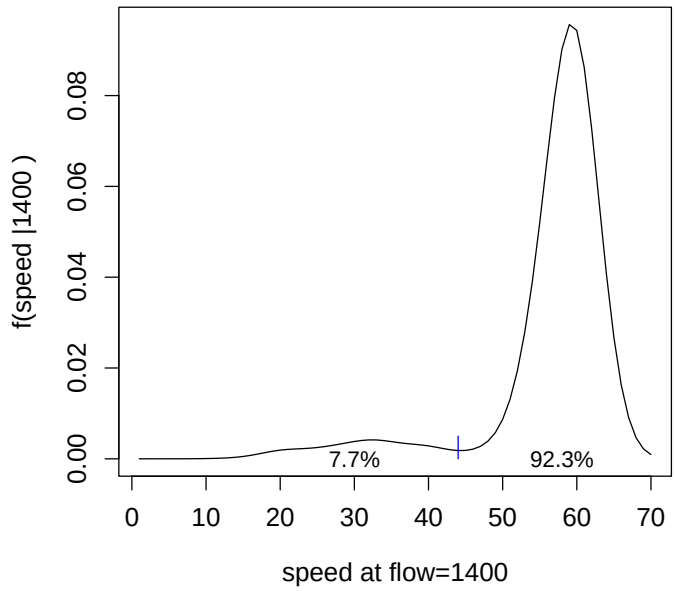

Figure 6: Estimated conditional density at a flow of 1400 vehicles/hour. The bottom of the valley at a speed of 43.00 mph is indicated by a vertical line.

minimum is reached, i.e the sequence $(f(y_m + k\delta|x))_k$ stops to fall. Note that the number of steps until the next minimum to the left and to the right do not necessarily need to be the same. Undoubtedly, there are more sophisticated tools that could be used to perform integration and the search for the minima. However, although being approximated by a step function, this integral is usually surprisingly accurate, since the approximation errors on the left and on the right side of the maximum tend to cancel out. So the choice of δ is not too crucial, it only influences the accuracy of the approximation. Fig. 7 shows the probabilities obtained for the data from lane 2. At a flow = 1620 vehicles/hour the two branches merge and are no longer distinguishable. At this point, the dashed line rises rapidly and catches up with the solid one. This is certainly *not* a sign for a suddenly rising probability of congested traffic. Beyond this point, there is no longer any dip to separate the components, although the data may still be seen as a - not clearly separated - mixture of the congested and uncongested regime.

For a given value x, storing the positions of the minima found while calculating the above integrals gives a vector of *conditional antimodes*. An antimode can be seen as an *'antiprediction - a value that is specifically identified as not likely to be seen'* (Cobb, 1998). Conditional antimodes have previously received attention particularly in catastrophe theory (Cobb & Zacks, 1985). In the present context, it is a straightforward step to connect the conditional antimodes in the x-direction in order to obtain a smooth curve, which then can be interpreted as a nonparametric *antiregression* or *antiprediction curve*, i.e. a curve describing where the data are *not* to be expected. In the case of speed-flow data, one ob-

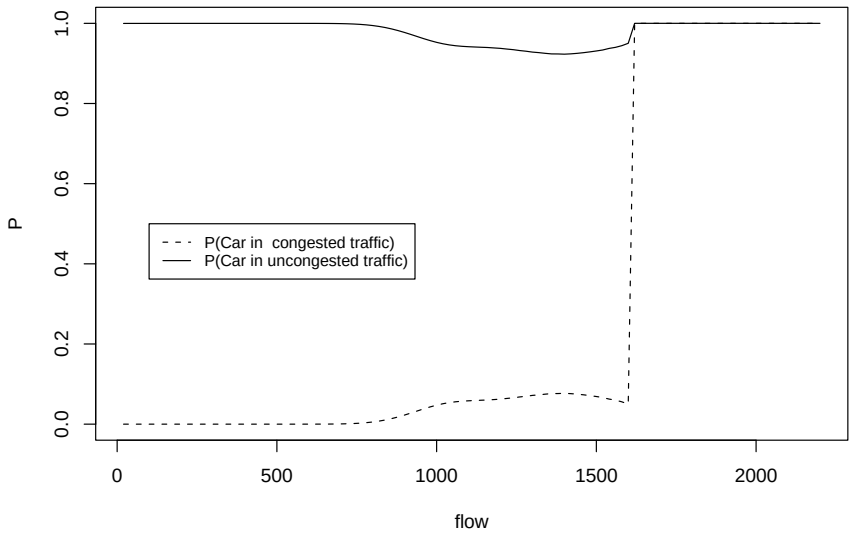

Figure 7: Probabilities of the branches of smooth multimodal regression curves of a speed-flow diagram

serves that this curve is also useful to classify the data into observations coming from the congested or uncongested regime. Fig. 8 shows both of the antiregression curves obtained by descending from the data cloud at the top (dashed line) as well as the corresponding curve obtained by descending from the conditional modes of the cluster at the bottom. As can be seen from the figure, the two curves fall (certainly) together, and nicely divide the cars into those monitored in congested or uncongested situations, as long as a division is possible. The concept of an antiregression curve deserves further attention and might be useful in a wide range of other data situations, but this is not the topic considered here.

5 Bandwidth selection

Since the problem of conditional mode estimation is strongly related to that of conditional density estimation, the bandwidth selectors developed in the latter context may be applied. The first bandwidth selection rule, developed by Fan, Yao & Tong (1996), is based on the RSC criterion (Fan & Gijbels, 1995). A further rule using bootstrapping was suggested by Hall, Wolff & Yao (1999) in the context of bandwidth selection for the conditional distribution function. Their idea was transferred to conditional density functions by Bashtannyk & Hyndman (2001) (hereafter BH). BH developed a variety of bandwidth selection rules, the computationally fastest of which is the normal reference rule. Applying this rule to the two lanes in Fig. 1, we obtain for the $x-$direction the bandwidths $h_1 = 133.48$ and $h_1 = 283.19$ for lanes 2 to 3, respectively. For the $y-$direction, one gets $h_2 = 11.16$ and $h_2 = 10.56$, respectively. Looking at boxplots of high density regions (Fig. 9 top), one

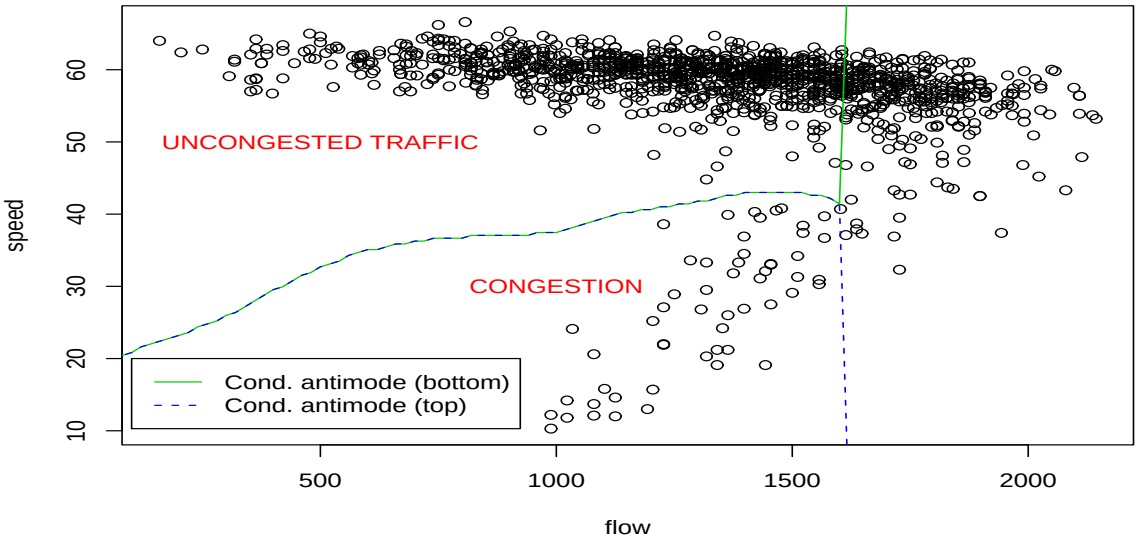

Figure 8: Antiregresssion curve as an instrument for classifying congested or uncongested traffic.

notices that these bandwidths are obviously oversmoothing, as also observed by BH and Fan & Yim (2004). This problem was already noted by Silverman (1986) in the context of univariate density estimation, who suggested employing a correction factor of 0.85 to the bandwidth selected by the normal reference rule. BH solved this problem in a less ad hoc manner and implemented a 'hybrid' bandwidth selection method. This uses the normal reference rule to select h_2 initially, then uses a regression-based method to find h_1 given h_2, and finally a bootstrap method to update h_2 for fixed h_1.

This 'hybrid' bandwidth selection rule (corresponding to the setting `method=2` and `deg=0` in function `cde.bandwidths` in the R package `hdrcde`) gives the bandwidths $h_1 = 108.08$ and $h_2 = 3.10$ for lane 2 and $h_1 = 199.56$, $h_2 = 3.55$ for lane 3. One sees from Fig. 9 that these bandwidths give reasonable results. As we obtained similarly pleasing results with this bandwidth selection tool for other data sets, we recommend this rule for multimodal regression. One drawback is that the computation of the hybrid bandwidths can be quite slow. In the special case of a data set with bimodal response, a helpful starting point for further fine tuning may be obtained by calculating the normal reference bandwidths, leaving the (less crucial) bandwidth h_1 as it is, and scaling the bandwidth h_2 by the factor 1/3. In our example, this gives nearly exactly the hybrid bandwidths for h_2.

6 Discussion

We have introduced a simple method to find the conditional modes of a multivariate density based on a conditional version of the mean shift procedure, and demonstrated by means of speed-flow data that this gives a fully nonparametric multi-valued regression

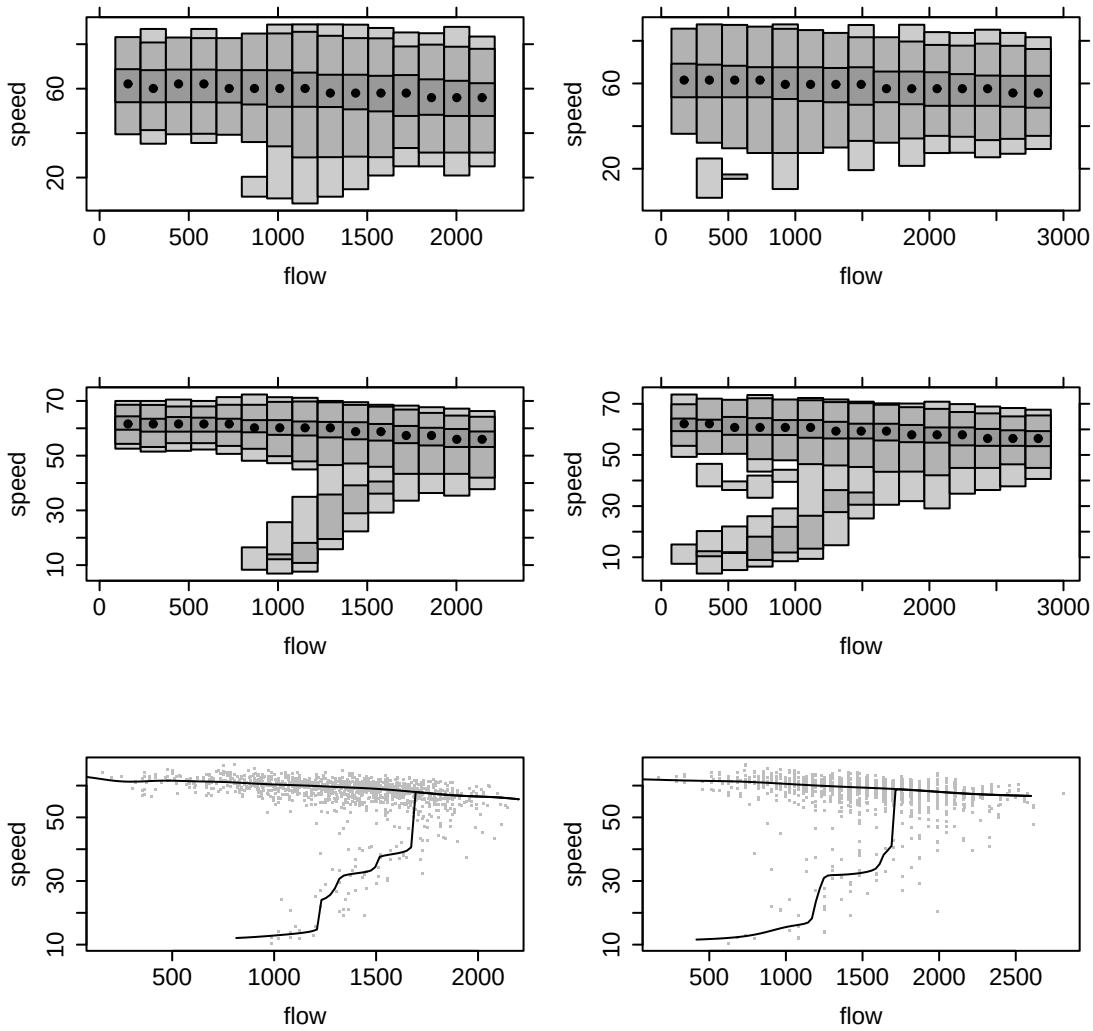

Figure 9: Top: HDR boxplots according to normal reference rule; middle: HDR boxpolts according to hybrid rule; bottom: multimodal regression curves corresponding to the hybrid bandwidth selector, for lane 2 and 3.

tool. As the problem of finding the conditional modes is highly related to the problem of conditional density estimation, we used some current research results from the latter area, which turned out to be helpful for the bandwidth selection problem.

Problems with multi-valued output have received attention in various research areas from both statistical and non-statistical perspectives. Beyond the various ideas arising from maximizing the conditional density listed in Section 3, a number of topics are worthy of mention: (1) Approaches based on mixtures of regression models, e.g. mixtures of generalized linear models (Wedel & Kamakura, 1995). A simple tool to fit models of this type is the R package `flexmix` developed by Leisch (2004). (2) 'Multiple model regression estimation' by Cherkassy & Ma (2005): this fits parametric regression models consisting of several regimes, where no prior assignment of points to regimes is required. The tool used for estimation is the Support Vector Machine (SVM); see also Suykens, Gestel, Brabanter, Moor & Vandewalle (2002) for related contributions from the machine learning community. (3) Switching regression (Quandt, 1972). (4) Bayesian approaches (Hurn, Justel & Robert, 2003). (5) Branching curves or *multicurves* (Silverman & Wood, 1987). This is a fully nonparametric approach based on smoothing splines, but every point is a priori assigned to a certain branch. (6) Principal curves (Hastie & Stuetzle, 1989) have been successfully applied on speed-flow data by Chen, Zhang, Tang & Wang (2004).

The dividing lines between the concepts (1) to (4) are rather thin, and there are many overlaps. For instance, (4) can be seen as a Bayesian version of (3); (3) can be seen as a special case of (2); and (2) itself could be reformulated to fit the framework of mixture modelling as well, so that approaches (1) to (4) are actually all based on mixture modelling. The essential difference to our work is that all these approaches, at least to the best of our knowledge, only allow fitting of mixtures of *parametric* regression models. The nonparametric alternatives (5) and (6) cannot be applied to the data situations considered in this paper, because in the former case we do not have prior knowledge about which point belongs to which branch, and in the latter case principal curves are not suitable for prediction.

Acknowledgements

We gratefully acknowledge support from Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 386: Statistical Analysis of Discrete Structures) in various aspects. This work was partly supported by Science Foundation Ireland Basic Research Grant 04/BR/M0051. The authors are grateful to Rob Hyndman, Monash University, Australia, for helpful hints concerning his bandwidth selection tools, and John Hinde, National University of Ireland, Galway, for suggestions and discussions.

References

Aubin, J.-P. and Frankowska, H. (1990). *Set-Valued Analysis*. Birkhäuser.

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comp. Stat. Data Analysis* **36**, 279–298.

Berlinet, A., Gannoun, A., and Matzner-Løber, E. (1998). Normalité aymptotique d'estimateurs convergents du mode conditionnel. *Canadian Journal of Statistics* **26**, 365–380.

Berlinet, A., Gannoun, A., and Matzner-Løber, E. (2001). Asymptotic normality of convergent estimates of conditional quantiles. *Statistics* **35**, 136–139.

Carreira-Perpiñan, M. A. (2000). Reconstruction of sequential data with probabilistic models and continuity constraints. In S. A. Solla, T. K. Leen, & K. R. Müller (Eds.), *Advances in Neural Information Processing Systems*, pp. 414–420. Cambridge, MA.: MIT Press.

Chen, D., Zhang, J., Tang, S., and Wang, J. (2004). Freeway traffic stream modeling based on principal curves and its analysis. *IEEE Trans. Intell. Transp. Systems* **5**, 246–258.

Cheng, Y. (1995). Mean shift, mode seeking and clustering. *IEEE Trans. Pattern Anal. Machine Intell.* **17**, 790–799.

Cherkassy, V. and Ma, Y. (2005). Multiple model regression estimation. *IEEE Transactions on Neural Networks* **16**, 785–797.

Chu, C. K., Glad, I. K., Godtliebsen, F., and Marron, J. (1998). Edge-preserving smoothers for image processing (with discussion). *J. Amer. Statist. Assoc.* **93**, 526–541.

Cobb, L. (1998). An introduction to cusp surface analysis. *http://www.aetheling.com/models/cusp/Intro.htm* .

Cobb, L. and Zacks, S. (1985). Applications of catastrophe theory for statistical modelling in the biosciences. *J. Amer. Statist. Assoc.* **80**, 793–802.

Comaniciu, D. (2003). An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Anal. Machine Intell.* **25**, 281–288.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.* **24**, 603–619.

Comaniciu, D., Ramesh, V., and Meer, P. (2001). The variable bandwidth mean shift and data-driven scale seclection. In *Proceedings 8th International Conference on Computer Vision*, Vancouver, BC, Canada, pp. 438–445.

Daganzo, C. F. (1995). Requiem for second-order approximation of traffic flow. *Transportation Research* **29B**, 277–286.

Davies, P. L. and Kovac, A. (2004). Densities, spectral densities and modality. *Ann. Statist.* **32**, 1093–1136.

Del Castillo, J. M. and Benitez, F. (1995). On the functional form of the speed-density relationships. *Transportation Research* **29B**, 373–406.

Drake, L. S., Schoefer, J. L., and May, A. D. (1967). A statistical analysis of speed-density hypotheses. *Highway Research Record* **154**, 53–87.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998–1004.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society, Series B* **57**, 371–395.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

Fan, J., Hu, T.-C., and Truong, Y. K. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics* **21**, 433–446.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* **91**, 819–834.

Fisher, N. I., Mammen, E., and Marron, J. S. (2002). Testing for multimodality. *Comp. Stat. and Data Analysis* **18**, 499–512.

Greenshields, B. D. (1935). A study of traffic capacity. *Highway Research Board Proc.* **14**, 448–477.

Hall, F. L. and Hall, L. M. (1990). Capacity and speed-flow analysis of the Queen Elisabeth way in Ontario. *Transportation Research Record* **1287**, 108–119.

Hall, F. L., Hurdle, V. F., and Banks, J. M. (1992). Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relations on freeways. *Transportation Research Record* **1365**, 12–17.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods of estimating a conditional distribution function. *Journal of the American Statistical Association* **94**, 154–163.

Härdle, W. and Gasser, T. (1984). Robust nonparametric function fitting. *Journal of the Royal Statistical Society, Series B* **46**, 42–51.

Hartigan, J. A. (2000). Testing for antimodes. In W. Gaul, O. Opitz, & M. Schader (Eds.), *Data Analysis, Scientific Modeling and Practical Applications*. New York: Springer.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry. *Statistical Science* **8**, 120–129.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84**, 502–516.

HCM (2000). Highway Capacity Manual 2000, Transportation Research Board.

Huang, S. H. and Ran, B. (2003). An application of neural network of traffic speed

prediction under adverse weather condition. In *TRB 2003 Annual Meeting CD-ROM. www.topslab.wisc.edu/publications/ran/2003/ran_2003_0915.pdf.*

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* **12**, 55–79.

Hyndman, R. J. (1995). Highest-density forecast regions for non-linear and non-normal time series models. *Journal of Forecasting* **14**, 431–441.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics* **5**, 315–336.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimations and symmetry tests for conditional density functions. *Nonparametric Statistics* **14**, 259–278.

Kockelman, K. M. (2001). Modeling traffic's flow-density relation: Accomodation of multiple flow regimes and traveler types. *Transportation* **24**, 363–374.

Koenker, R. (2005). *Quantile Regression*. Cambridge, UK.: Econometric Society Monograph Series, Cambridge University Press.

Lee, J. S. (1983). Digital image smoothing and the sigma filter. *Computer Vision, Graphics and Image Processing* **24**, 255–269.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* **11**, 1–18.

Li, M. Z. F. (2005). Generic characterization of equilibrium speed-flow relationships. Unpublished Working Paper, Nanyang Business School, Singapore.

Li, T. and Zhang, H. M. (2001). The mathematical theory of an enhanced nonequilibrium traffic flow model. *Journal of Networks and Spatial Economy* **1**, 167–177.

Matzner-Løber, E., Gannoun, A., and Gooijer, J. G. D. (1998). Nonparametric forecasting: A comparison of three kernel-based methods. *Comm. Statist. - Theory Meth.* **27**, 1593–1617.

Mehra, K. L., Ramakrishnaiah, Y. S., and Sashikala, P. (2000). Laws of iterated logarithm and related asymptotics for estimators of conditional density and mode. *Ann. Inst. Statist. Math.* **52**, 630–645.

Muller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746.

Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Applied Probabilty* **10**, 186–190.

Parzen, E. (1962). On estimation of a probability function and mode. *Ann. Math. Statist.* **33**, 1065–1076.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *J. Amer. Statist. Assoc.* **67**, 306–310.

Samanta, M. and Thavaneswaran, A. (1990). Non-parametric estimation of the conditional mode. *Commun. Statist. - Theory Meth.* **19**, 4515–4524.

Scott, D. W. (1992). *Multivariate Density Estimation*. New York: Wiley.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B* **43**, 1–21.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Silverman, B. W. and Wood, J. T. (1987). The nonparametric estimation of branching curves. *J. Amer. Statist. Assoc.* **82**, 551–558.

Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. Singapore: World Scientific.

Wedel, M. and Kamakura, W. A. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification* **12**, 21–55.

Xiao, H., Sun, H., and Ran, B. (2003). Fuzzy neural network traffic prediction framework with wavelet decomposition. In *TRB 2003 Annual Meeting CD-ROM. www.topslab.wisc.edu/publications/ran/2003/ran_*2003_1899*.pdf*.