Einbeck, Jochen; Tutz, Gerhard

Working Paper

Modelling beyond regression functions: an application of multimodal regression to speed-flow data

Discussion Paper, No. 395

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Einbeck, Jochen; Tutz, Gerhard (2004) : Modelling beyond regression functions: an application of multimodal regression to speed-flow data, Discussion Paper, No. 395, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1765

This Version is available at:
https://hdl.handle.net/10419/31026

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

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

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

8th December 2004

Abstract

An enormous amount of publications deals with smoothing in the sense of nonparametric regression. However, nearly all of the literature treats the case where predictors and response are related in the form of a function $y = m(x) + \text{noise}$. In many situations this simple functional model does not capture adequately the essential relation between predictor and response. We show by means of speed-flow diagrams, that a more general setting may be required, allowing for multifunctions instead of only functions. It turns out that in this case the conditional modes are more appropriate for the estimation of the underlying relation than the commonly used mean or the median. Estimation is achieved using a conditional mean-shift procedure, which is adapted to the present situation.

Key Words: Mean shift, Conditional density, Conditional mode, 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 Californian uninterrupted highway ("freeway") having 4 lanes, where only the lanes 2 and 3 are depicted here. The speed is measured in miles per hour, and the flow in vehicles per lane per hour. Each point is an average speed and hourly flow rate for data collected over a 30-seconds interval. The question is how can the shape of the data cloud be explained. For uncongested traffic, there is no significant

*einbeck@stat.uni-muenchen.de
†tutz@stat.uni-muenchen.de

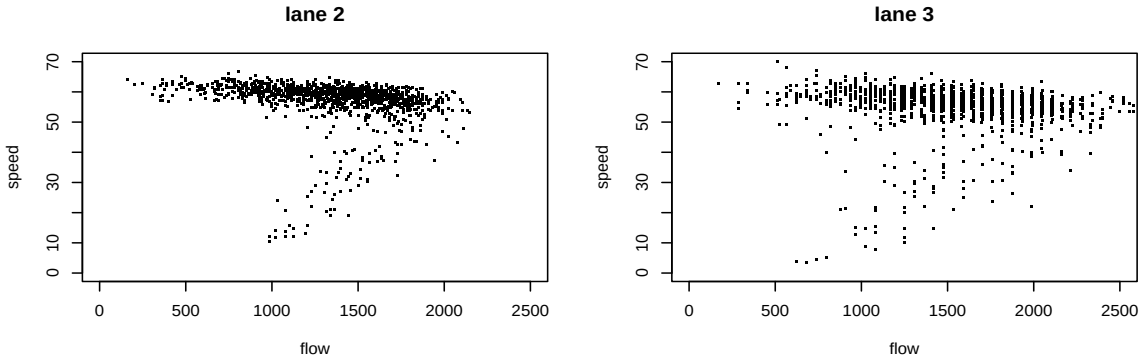

Figure 1: Speed-flow diagram for californian freeway.

association between traffic flow and speed - this is the big longish cluster at the top of the plots. When the traffic gets too dense, however, there might still be high traffic flow, but with a considerably diminished speed due to congestion. These are the less dense data points at the bottom of the figures. There has been considerable effort over the last 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). Recently, Kockelman (2001) applied mixture models of congested and uncongested conditions to flow-density relations.

Looking at Fig. 1, one notices immediately that the speed v cannot be described as a function $v(q)$ of the flow q. Thus, any attempt on modelling these data has been based on modelling the traffic flow as a function $q(v)$. 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 $v(q_0)$, 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-line highway. For a nice overview of available models see Li (2003). Since mathematical models generally involve functions, the question of how to model $v = v(q)$, as the usually applied type of plotting used in Fig. 1 would suggest, has never been considered. Let us assume that a nonparametric statistician has to find a solution for this problem, which in his terminology takes the form: Based on a value of traffic flow, give a prediction for speed. In practice, traffic speed prediction is

a very important issue, e.g. to construct and support Intelligent Transportation Systems (ITS), enabling drivers to obtain their expected arrival time. Recently, Huang & Ran (2003) worked on traffic speed prediction using neural networks.

In the following we present a general nonparametric approach to regression problems as that one described above, referring throughout to the application on speed-flow data.

2 Conditional Modes and Densities

The basic nonparametric regression problem is simply described from a statistical point of view. Assume a set of i.i.d. random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ sampled from a population $(X, Y) \in \mathbb{R}^2$ with joint density $f(x, y)$, where Y is a scalar response variable and X a 1-dimensional predictor. The objective is to find a smooth function $m : \mathbb{R} \longrightarrow \mathbb{R}$ relating X and Y in a proper way, which may be generally expressed as

$$m(x) = \Omega(Y|X = x). \tag{1}$$

The choice of the operator $\Omega(\cdot)$ is quite crucial. A common setting is $\Omega(\cdot) = E(\cdot)$, with expectation $E(\cdot)$. Early smoothing approaches, for instance Nadaraya (1964) or Reinsch (1967) in the context of localizing or penalizing, respectively, are nearly exclusively based on this setting, and due to its computational feasibility the expectation operator is undiminishedly popular. Recently, alternative choices of Ω have been considered. A natural alternative is

$$\Omega(\cdot) = \text{Med}(\cdot), \tag{2}$$

i.e. the regression function is the conditional median rather than the conditional mean. An early reference to local median smoothers is Härdle & Gasser (1984), see Fan, Hu & Truong (1994) for further references. Overviews on nonparametric regression methods have been given by Green & Silverman (1994), Simonoff (1996), Härdle & Schimek (1996), and Härdle, Müller, Sperlich & Werwatz (2004). Both above choices of $\Omega(\cdot)$ have in common that they are solutions of a specific minimization problem, namely

$$m_l(x) = \arg\min_a E(l(Y - a)|X = x), \tag{3}$$

where the loss function $l(\cdot)$ has the form $l(z) = z^2$ or $l(z) = |z|$, respectively (Fan, Hu & Truong, 1994). Applying these settings exemplarily on lane 2, one gets the results in Figure 2. Obviously, neither of the two curves is suitable, since they do not account for the flaked data points at the bottom, which obviously carry information and cannot be regarded as outliers. Another choice of Ω, that has received little attention, is the mode function

$$\Omega(\cdot) = \text{Mode}(\cdot). \tag{4}$$

A current reference on mode estimation is Bickel (2003). The mode also may be considered as a solution of a minimization problem of type (3). Let $l(\cdot) = -\delta(\cdot)$, where $\delta(\cdot)$ is the delta-function, i.e. $\delta(x) = 0$ for $x \neq 0$ and $\int \delta(x)\, dx = 1$. Then one obtains

$$\begin{aligned} m_l(x) &= \arg\min_a E(-\delta(Y-a)|X=x) = \\ &= \arg\min_a \left\{ -\int_{-\infty}^{\infty} \delta(y-a) f_{Y|X}(y|x)\, dy \right\} = \\ &= \arg\min_a \left\{ -f_{Y|X}(a|x) \right\} = \\ &= \arg\max_a f_{Y|X}(a|x) = \text{Mode}(Y|X=x). \end{aligned}$$

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 a. Apart from that, a conditional density function can have several conditional maxima on different levels, which may be interpreted as *local modes*, being defined by

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

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

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

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

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

and the counterpart to model (1) is given by the multifunction

$$M(x) = \{\Omega_{(j)}(Y|X=x) | 1 \leq j \leq k\}. \tag{5}$$

[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 about 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 $M(\cdot)$. Since conditional

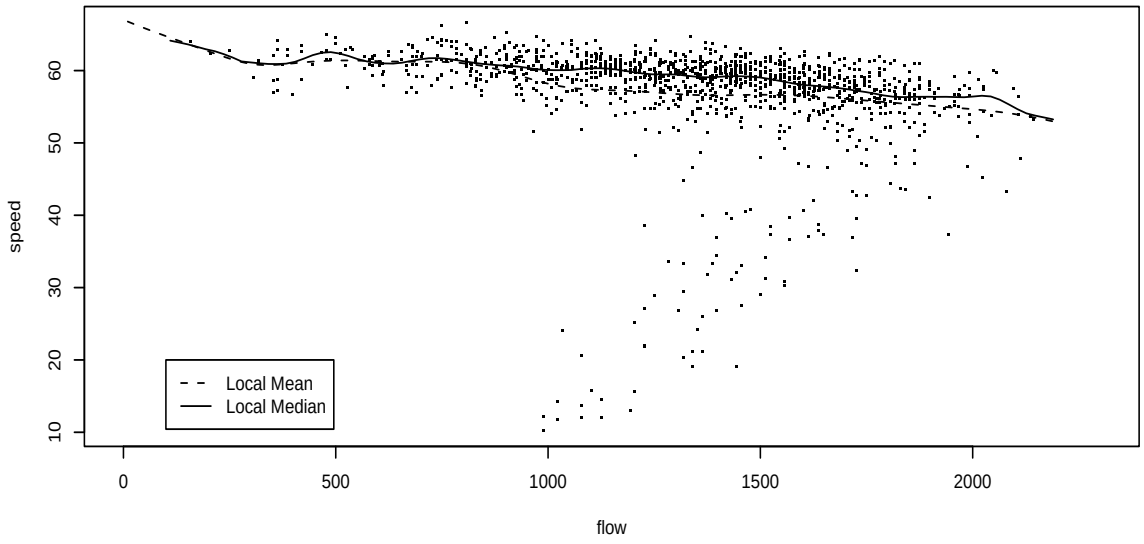

Figure 2: Speed-flow diagram for lane 2 with local (linear) mean and median smoother.

mode estimators are edge-preserving, they tend to have jumps. Thus, we only demand each branch to be smooth except for a finite set of points.

A multimodal distribution can be approximated by a finite mixture of unimodal distributions (McLachlan & Peel, 2000). Going one step further, one can construct finite mixtures of (generalized) linear regression models to account for multimodal response, as proposed by Wedel & Kamakura (1995). However, these approaches are (yet) restricted to parametric modelling. Our intention is to introduce a fully nonparametric method for multimodal regression. Therefore, we consider the conditional density estimator (Tutz, 1990, Fan, Yao & Tong, 1996, Hyndman & Yao, 2002, Fan & Yim, 2004), which in the case of univariate predictors has the form

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{f}(x)} = \frac{\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) K_2\left(\frac{Y_i - y}{h_2}\right)}{h_2 \sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right)}, \tag{6}$$

with kernels K_1, K_2 and bandwidths h_1, h_2. This enables a direct view of the conditional maxima of the distribution. Plotting the conditional densities at different values of flow yields Fig. 3 ($h_1 = 100, h_2 = 4, K_1, K_2$: Gaussian).

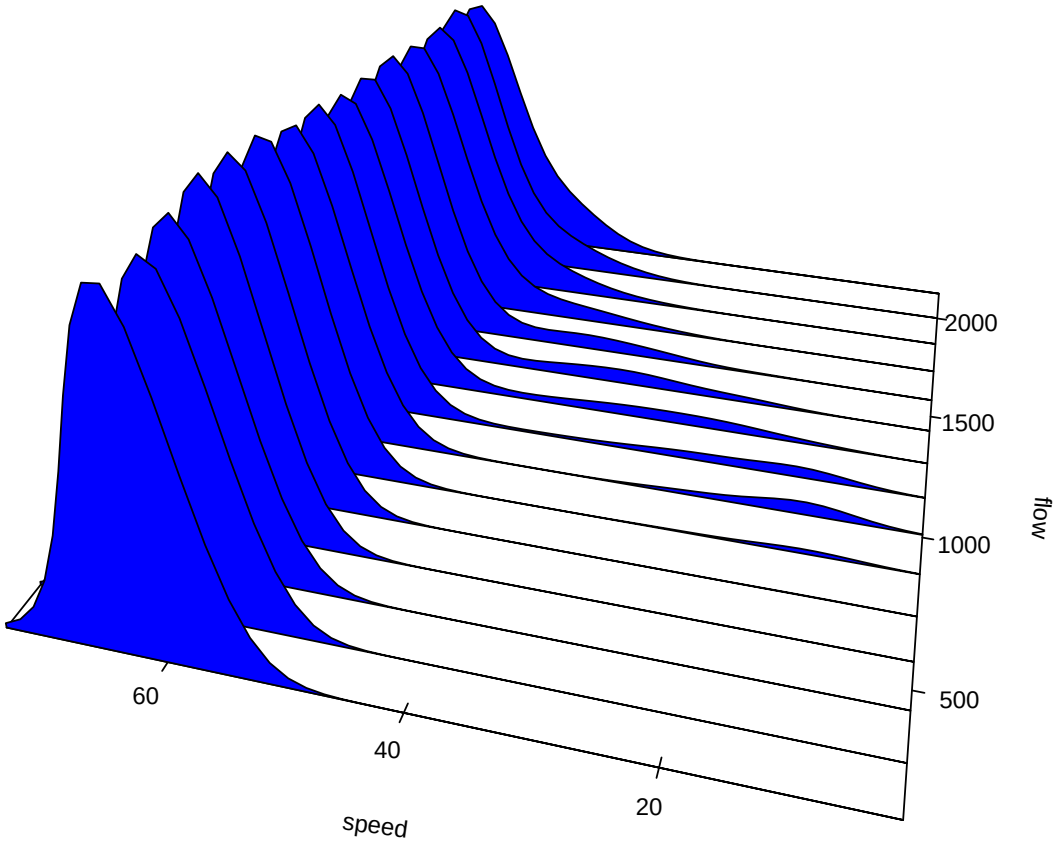

Figure 3: Conditional densities for speed-flow diagram (lane 2).

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 speed as follows: The expected speed will be *either* around 30 mph *or* around 60 mph, where *60 mph is more likely*. From this observation two problems arise which have to be clarified:

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

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.

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 *sample conditional mode*, is consistent and asymptotically normally distributed under suitable regularity conditions. However, it remains the problem of *how* to find the maxima of the conditional density estimates. Estimation of the maxima of a density function has been treated 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. Note that, in our setting, we do not only need the conditional global maximum, but, if possible, *all* conditional local maxima. There are two possible solutions at hand: grid search and the method of conditional mean shift.

3.1 Grid search

A grid search is undoubtedly an easily implemented and feasible tool to find the maximum of a (density) function. Searching for maxima on a grid is however computationally extremely demanding, especially when the space of predictors is multivariate. Moreover, a grid search algorithm finding *all local* conditional modes (rather than the global one) is not so easy to implement. Thus, we will not pursue this idea further in the following.

3.2 Conditional Mean shift

Maxima of the conditional density (6) have the property $\hat{f}'(y|x) = 0$, thus we turn our interest to the derivative of $\hat{f}(y|x)$. We 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 (6) 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 influences the definition of the weights $w_i(x)$.) By considering

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

one obtains for the mode estimator y_m

$$y_m = \frac{\sum_{i=1}^{n} w_i(x) k'\left(\left(\frac{Y_i - y_m}{h_2}\right)^2\right) Y_i}{\sum_{i=1}^{n} w_i(x) k'\left(\left(\frac{Y_i - y_m}{h_2}\right)^2\right)} \tag{7}$$

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

$$y_m = \frac{\sum_{i=1}^{n} w_i(x) G\left(\frac{Y_i - y_m}{h_2}\right) Y_i}{\sum_{i=1}^{n} w_i(x) G\left(\frac{Y_i - y_m}{h_2}\right)}. \tag{8}$$

In the case of conditional mode estimation, one has to set

$$w_i(x) = \frac{K_1\left(\frac{X_i - x}{h_1}\right)}{\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right)}, \tag{9}$$

yielding

$$y_m = \frac{\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) G\left(\frac{Y_i - y_m}{h_2}\right) Y_i}{\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h_1}\right) G\left(\frac{Y_i - y_m}{h_2}\right)} =: \mu(y_m). \tag{10}$$

This equation 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 idea of the mean shift procedure, which has been studied in the unconditional case (i.e. $w_i(x) \equiv 1$) by Comaniciu, Ramesh & Meer (2001), Comaniciu & Meer (2002) and Comaniciu (2003). For a given starting point y_0, Comaniciu & Meer (2002) showed that the sequence $(y_j)_{j=1,2,\ldots}$ defined by

$$y_{j+1} = \mu(y_j) \tag{11}$$

converges to a nearby mode y_m, which is a fix point of (11). (They give a proof in fact only for the unconditional case, but the extension to general weights is straightforward). To account for multimodal conditional distributions, one applies the mean shift procedure as follows: For a given x,

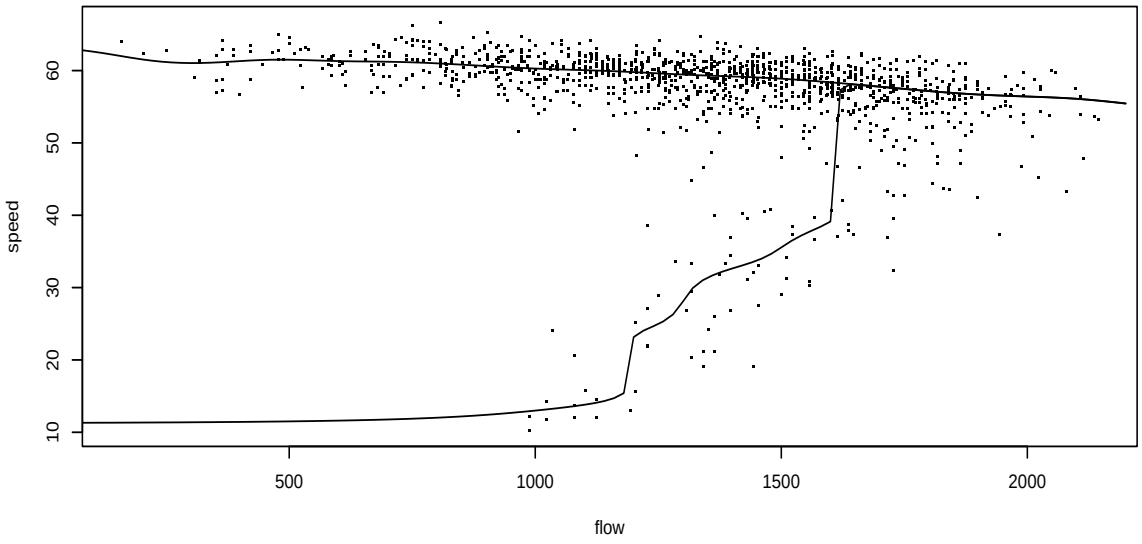

Figure 4: Multimodal regression for speed-flow data based on mean shift.

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

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

$$y_{j+1}^{(p)}(x) = \mu(y_j^{(p)}(x)) \tag{12}$$

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

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

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

The set $\hat{M}(x)$ 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 of branches. However, it is also possible when P is equal or smaller than the number of branches, as some modes may not have been found, while other modes are included several times in the multiset. To be certain that all modes are discovered, one has to install 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. When in doubt,

there is nothing wrong with starting more iterations than strictly necessary - in the worst case some modes are reached two or more times, but if the mean shift is iterated until convergence, all estimates belonging to the same mode are approximately equal.

In practice, around 30 iterations are enough to obtain convergence. Fig. 4 shows the results of a multimodal regression according to the presented algorithm. The conditional mean shift is calculated with Gaussian kernels K_1 and K_2 ($h_1 = 100, h_2 = 4$) and local constant weights as in (9). 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 one has some prior information about the shape of the underlying relation, it might be useful to work with variable starting points.

Remark 1.

The right side of (10) 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 procedure, which iterates (8) or (10) 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 relation of sigma filtering and mean shift is not surprising.

Remark 2.

We showed that setting $l(\cdot) = -\delta(\cdot)$ in the minimization problem (3) yields the conditional mode. In practice, the delta function has to be approximated. This is easily possible by means of the kernel function K_2, as

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

Applying the setting

$$l(\cdot) = -\frac{1}{h_2} K_2\left(\frac{\cdot}{h_2}\right) \equiv -\frac{c_k}{h_2} k\left(\left(\frac{\cdot}{h_2}\right)^2\right) \tag{13}$$

and minimizing (3) yields the equation

$$a = \frac{E\left(G\left(\frac{Y-a}{h_2}\right) Y | X = x\right)}{E\left(G\left(\frac{Y-a}{h_2}\right) | X = x\right)}. \tag{14}$$

Thus, the right side of (10) estimates the right side of (14). Comaniciu & Meer (2002) show that (13) corresponds (in the unconditional case) to *location M estimation*. Chu, Glad,

Godtliebsen & Marron (1998) make use of this relation by exploiting local M estimators for edge preserving smoothing and show improved performance of this estimator in comparison to the sigma filter. This is to be expected, since the sigma filter can be interpreted as a first step towards local M estimation.

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). One might wonder if the local constant conditional mean shift, which we employed, might be further improved to give a local linear mode estimator. This simply requires replacing the weights $w_i(x)$ in (8) by the corresponding weights for a local linear fit (Fan & Gijbels, 1996, p. 20, Fan & Yim, 2004), namely

$$w_i(x) = \frac{K_1\left(\frac{X_i - x}{h_1}\right)[S_{n,2} - (X_i - x)S_{n,1}]}{\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right)[S_{n,2} - (X_j - x)S_{n,1}]}, \tag{15}$$

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 (9) and local linear (15) settings is provided in Fig. 5. It is obvious 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 data structure is multimodal (Fig. 5 right), the local linear mode estimator behaves quite erratically and gives non-smooth results. Thus, for the rest 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 would formulate here

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

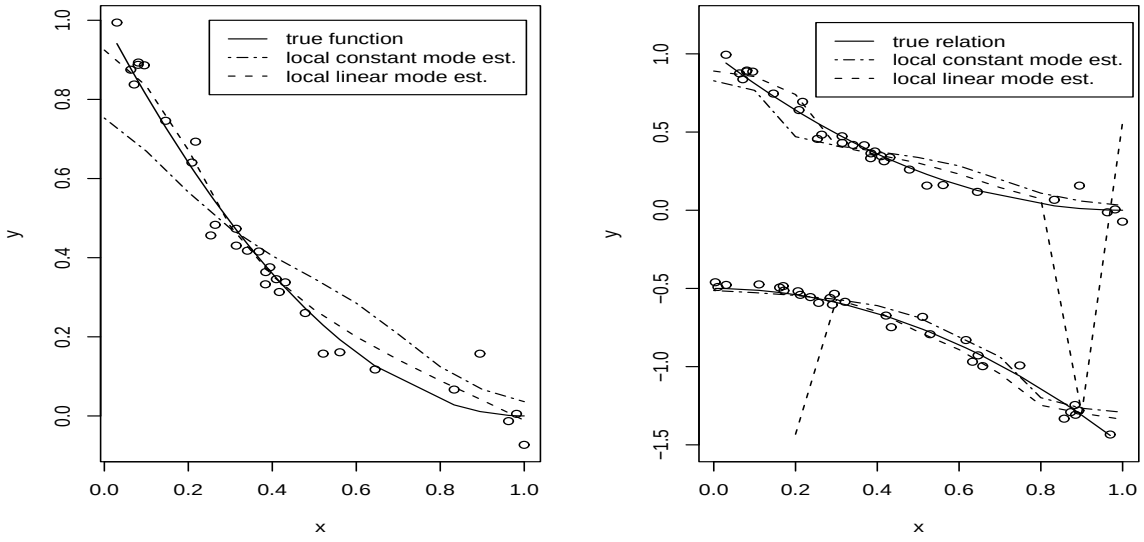

Figure 5: Comparison of a local constant and a local linear conditional mode estimator based on mean shift (left: $h_1 = 0.2, h_2 = 0.4$; right: $h_1 = h_2 = 0.2$)

To estimate these probabilities, one has to find the lows 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 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 will be more sophisticated tools that could be used to perform this 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. Thereby the choice of δ is not very crucial, because it is not a tuning parameter in this sense, but only influences the accuracy of the approximation. Fig. 7 shows the probabilities obtained in this manner for the data from lane 2. There is a point where the two branches merge and are no longer distinguishable. In this example, this point is achieved at a flow = 1620 vehicles/hour. At this point, the dashed line rises rapidly and catches up to the solid one. This is certainly an artifact and *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.

It is also helpful to store the positions of the minima found while calculating the integrals, and to look at the plots of these "minimum curves'. It turns out that these curves are also

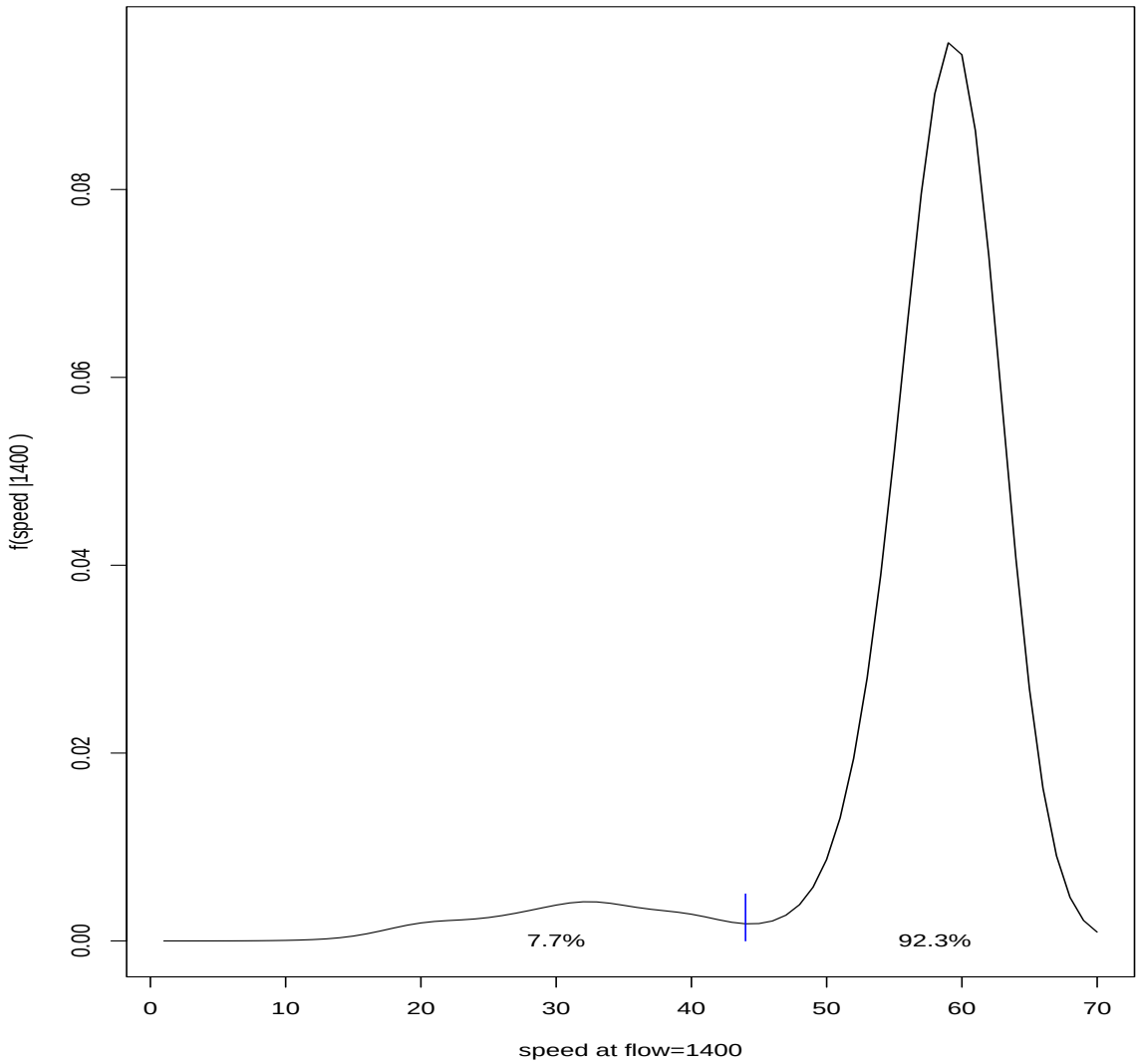

Figure 6: Estimated conditional density at a flow of 1400 vehicles/hour. The bottom of the valley at a speed of 43.00 mph is indicated by a vertical line.

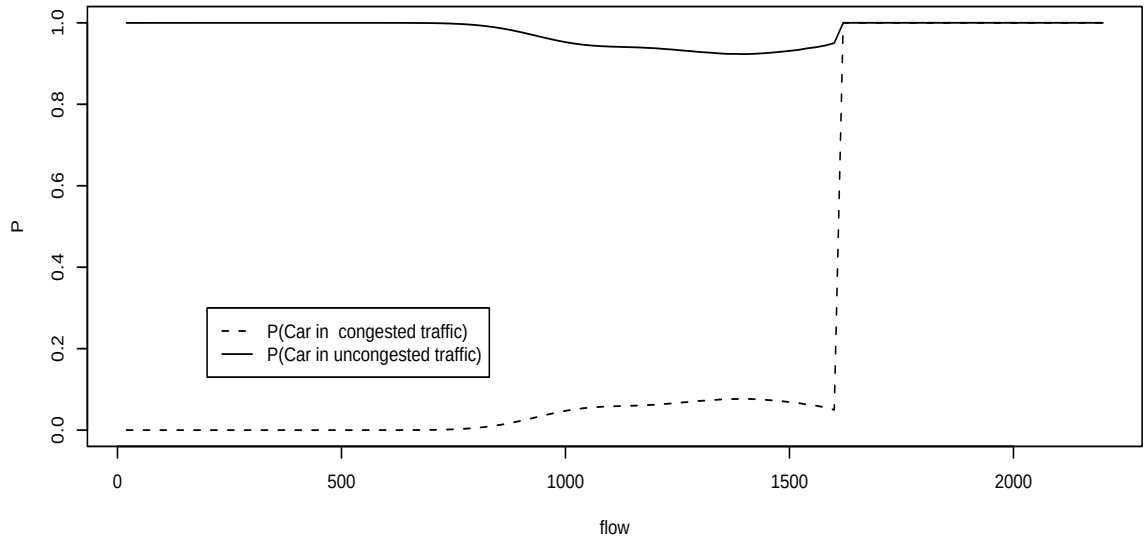

Figure 7: Probabilities of the branches of smooth multimodal regression curves of a speed-flow diagram

useful to classify the data into observations coming from congested or uncongested traffic. Fig. 8 shows both the minimum curves obtained by descending from the data cloud in the top (dashed line) as well as the corresponding curve obtained by descending from the cluster in the bottom. As can be seen from the figure, the two curves fall (certainly) together, and divide nicely the cars into those monitored in congested or uncongested situations, as long as a division is possible. The area in the right where no division is possible corresponds to the situation named "Queue discharge" by Hall, Hurdle & Banks (1992). It should be remarked that the minimum curve might also interesting from another point of view: it can be seen as a kind of "antiregression" curve, i.e. a curve which describes where the data are *not*. This curve might be useful in a wide range of other data situations, but this is not the topic considered here.

5 Bandwidth selection

One possible computational problem is that of bandwidth selection. However, fortunately there exist some results on how to select the bandwidth of local conditional density estimates, and we may apply these here. 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), however in the context of bandwidth selection for the conditional distribution function. Their idea was transferred to conditional density functions by Bashtannyk & Hyndman (2001). More recently, Hyndman & Yao (2002) (hereafter HY) developed another fast and simple rule

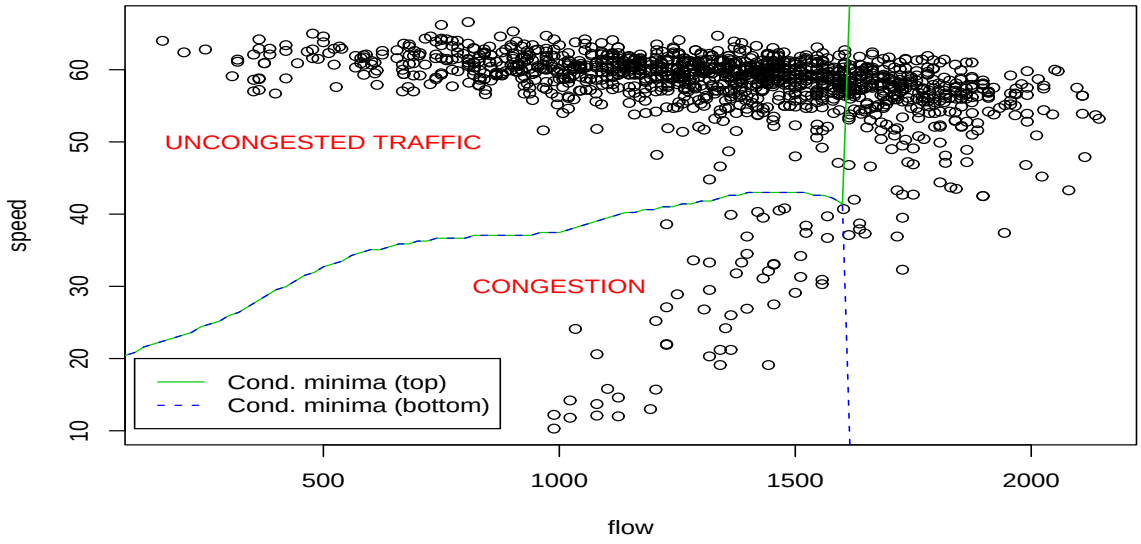

Figure 8: Minimum line as an instrument for classifying congested or uncongested traffic.

based on a combination of asymptotic properties of a local polynomial approximation of the conditional density and Silverman's normal reference rule (Silverman, 1986). HY developed an R package named `hdrcde` containing this bandwidth selection algorithm. Applying this rule to the two lanes in Fig. 1, we obtain for the x direction the bandwidths $h_{1,HY} = 133.480$ and $h_{1,HY} = 307.522$ for lanes 2 to 3, respectively. For the y direction, one gets $h_{2,HY} = 11.163$ and $h_{2,HY} = 11.100$, respectively. Looking at boxplots of high density regions (Fig. 9 top), one notices that the HY - bandwidths in y- direction are obviously oversmoothed, as also observed by Fan & Yim (2004). For the purpose of multimodal regression this is even more serious than for conditional density estimation, since one needs to distinguish the conditional modes clearly in order to identify them. From our experience with this and other data sets we suggest setting

$$h_{2,k} = \frac{h_{2,HY}}{1.5 \cdot k}, \tag{16}$$

where k is the number of branches which are supposed to be separated, and $h_{2,HY}$ is the bandwidth selected by the algorithm from HY. The bandwidth $h_{1,HY}$ can usually be left unchanged. Fig. 9 (middle) shows the boxplots of high density regions (HDR) for the corrected bandwidth $h_{2,2}$. Fig. 9 (bottom) shows the corresponding multimodal regression curves.

6 Discussion

We have demonstrated that, applying a simple conditional mean shift procedure, multimodal regression is easily tractable and yields reasonable results. The problem of finding

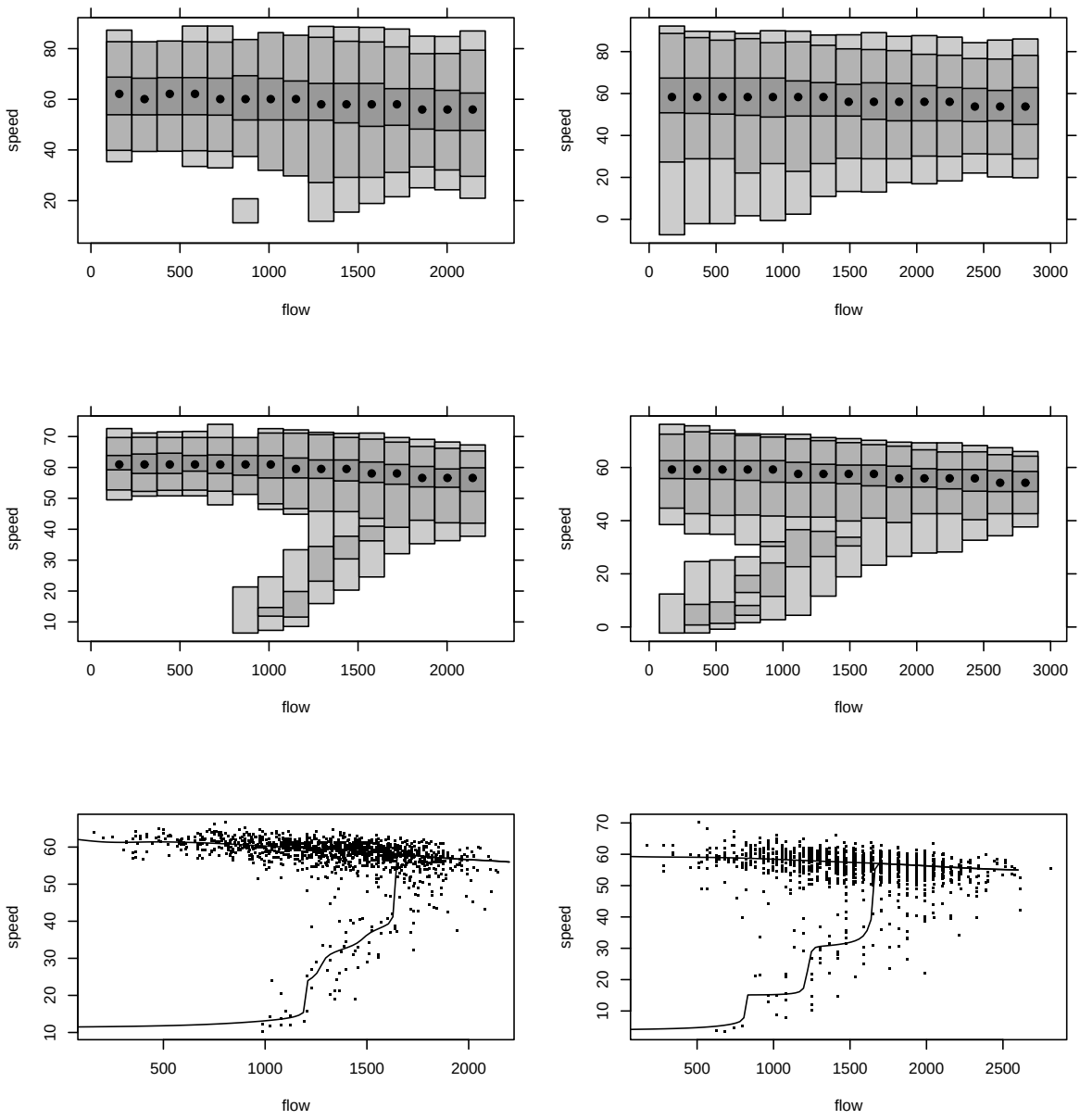

Figure 9: Top: HDR boxplots according to the HY rule; middle: HDR boxpolts according to the modified bandwidth selector (16); bottom: multimodal regression curves corresponding to the modified bandwidth selector, for lane 2 and 3.

the conditional modes is strongly related to the problem of conditional density estimation, which fortunately makes available some useful results concerning e.g. bandwidth selection. It should be noted that the mean shift procedure itself is extremely fast. Thus, calculating the modes via a mean shift procedure is much more faster than calculating all conditional densities and performing a grid search to find the maxima.

We applied the method to speed-flow relationships, and obtained smooth regression curves for speed given flow. Certainly the applicability of the method is not restricted to this kind of data. It can be used wherever multimodal conditional densities are to be expected. Another example for such a situation is the well-known Old Faithful geyser data (see Fig. 8. in Hyndman & Yao (2002)).

The proposed ideas leave plenty of room for further research and a number of open questions: Is it possible to construct a kind of anti-mean shift (i.e. an iterative algorithm converging to the valleys of a conditional density estimate) in order to obtain the "minimum curve"? How can this "antiregression curve" be exploited?

Finally, one word on theory. It is quite surprising that, though to our knowledge never applied for this purpose, the necessary theory related to nonparametric multimodal regression already exists to some extent. Asymptotic properties of estimators of the conditional mode have been derived by Berlinet, Gannoun and Matzner-Løber (1998, 2000). Asymptotic properties of the conditional density function have been analyzed in Fan, Yao & Tong (1996) and Fan & Yim (2004). The novel idea of the present paper is the use of a conditional mean shift procedure with the goal of multimodal regression.

Acknowledgements

We gratefully acknowledge support from Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 386: Statistical Analysis of Discrete Structures) in various aspects. The authors are grateful to John Hinde, National University of Ireland, Galway, for helpful suggestions.

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. (2000). Asymptotic normality of convergent estimates of conditional quantiles. *Statistics* **35**, 136–139.

Bickel, D. R. (2003). Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency. *J. Statist. Comput. Simul.* **73**, 899–912.

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

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.

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.

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 data-driven method to estimate conditional densities. *Biometrika* **??**, ??–??

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.

Greenshields, B. D. (1935). A study of traffic capacity. *Highway Reserach 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.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparemetric and semiparametric models*. New York: Springer Verlag.

Härdle, W. and Schimek, M. G. (1996). *Statistical Theory and Computational Aspects of Smoothing*. Heidelberg: Physica.

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

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*. www.topslab.wisc.edu/resources/ publications/ran/2003/ran_915.pdf.

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. 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.

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

Li, M. Z. F. (2003). Generic characterization of speed-flow relationships. under review.

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.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Nadaraya, E. A. (1964). On estimating regression. *Theory Prob. Appl.* **9**, 141–142.

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. Sta-*

tist. **33**, 1065–1076.

Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.* **10**, 177–183.

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

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

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer Verlag.

Tutz, G. (1990). Smoothed categorical regression based on direct kernel estimates. *Journal of Statistical Computation and Simulation* **36**, 139–156.

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