

An Evaluation of Non-parametric relative risk estimators for disease maps

Allan B Clark¹ and Andrew B Lawson¹

¹Spatial Epidemiology Group,
Department of Public Health,
University of South Carolina, Columbia, SC, USA
Phone: +1 803 777 4562 Fax: +1 803 777 2524
Email: allanbclark@sc.edu

**Presented at GeoHealth 2002
Victoria University of Wellington
December 3-5th 2002**

ABSTRACT

In geographical epidemiology it is often required to produce a map of the risk of disease over a study region, a disease map. This paper reviews a variety of approaches to produce disease maps when individual address locations are observed. These methods vary from kernel based smoothing approaches, e.g. Nadaraya-Watson, local linear and GAMs, to Bayesian partition models. The kernel based methods have the advantage of speed, but the partition model has the advantage of being able to adapt to local features (i.e. clustering) of the surface. Another advantage of the kernel based methodology is that the local linear model has a built in edge correction. A simulation study designed to assess the benefits of using an edge corrected estimator for relative risk estimation is described.

Keywords and phrases: disease mapping, generalised additive model, partition model, MCMC,

1.0 INTRODUCTION

There has been much interest in statistical models for the estimation of disease maps; see Lawson and Cressie (2000) for a review. In this paper we extend the work of Kelsall and Diggle (1998) and Ferreira et al (2002) who considered the production of a disease map when the data is available as the set of individual addresses of people with the disease within a study region. This data can be thought of as a single realisation of a spatial point process. This allows us to define the first order intensity of 'cases' at x as $\lambda_1(x)$. Of course, any analysis on this intensity would be meaningless unless we took account of the population at risk. In order to do this we define a 'control' disease that has similar characteristic to the disease under study. This defines another realisation of a point process with intensity $\lambda_0(x)$. These intensities play that exact same role as rates in standard cohort studies.

We shall use the following notation x_i is the location of the i th point; the first n_0 of these points are controls, the next n_1 are cases giving a total of n points; the binary labels y_i that will be attached to the points are such that $y_i=1$ if the i th point is a case, and $y_i=0$ if the i th point is a control. It is common to assume that the intensity of cases is linked to the intensity of controls via a multiplicative model, i.e.

$$\lambda_1(x) = \theta(x)\lambda_0(x),$$

where $\theta(x)$ is the relative risk function. The relative risk function can be expressed as the ratio:

$$\theta(x) = \frac{\lambda_1(x)}{\lambda_0(x)}.$$

Unfortunately, this parameter is difficult to estimate directly since, at each case location the observed relative risk is infinite, and at each control location the observation is zero. A simpler parameter to estimate is the conditional probability of a case, given an event at x , i.e.

$$p(x) = \frac{\lambda_1(x)}{\lambda_0(x) + \lambda_1(x)}. \quad (1)$$

Conditional on the observed events, $\{x_i\}$, the binary labels, $\{Y_i\}$, are distributed as mutually independent Bernoulli random variables, with probability $\Pr\{Y_i=1/X=x_i\}=p(x)$. Thus the estimation of $p(x)$ is a binary regression problem. Kelsall and Diggle (1998) demonstrated that $p(x)$ is directly related to $\theta(x)$.

In section 2 of the paper we give an introduction to the methodology that can be used to produce disease maps and locate any clusters that may exist. The Nadaraya-Watson regression estimator, used by Kelsall and Diggle (1998), has been shown to behave poorly near the boundary of the study window. Alternatives to the Nadaraya-Watson regression estimator have been examined, for example, Fan and Gijbels (1992) show that the local linear regression model has a built-in edge correction. Edge correction is an important problem in geographical epidemiology since it is not uncommon that a significant proportion of the population lie near edge of the study region. The purpose of this paper is to assess the benefits of having an edge-corrected estimator for relative risk estimation. This will be carried out via a simulation study, described in section 3. Finally, in section 4 we discuss future directions that could be investigated.

2.0 NON-PARAMETRIC MODELS

In the previous section, we described the problem of individual level disease mapping, and we showed that it could be viewed as a binary regression problem. Possible approaches to this are discussed in sections 2.1, 2.2 and 2.3. The simple estimator described in section 2.1 is used only to demonstrate the importance of treating the problem as a binary regression problem and should not be considered as an appropriate method. Kernel based regression estimators are described in section 2.2 they are relatively quick to estimate and are easily programmed in standard statistical software programs. The partition model, discussed in section 2.3, has recently grown in stature amongst the statistical community since they, unlike kernel based estimators, are not constrained by global parameterisations.

2.1 Plug-in estimator

A simple estimator of $p(x)$ is to estimate the intensities for controls and cases then substitute them into equation (1). The kernel estimators for the intensity of cases and controls are given below:

$$\hat{\lambda}_0(x; h) = \sum_{i=1}^{n_0} K_h(x - x_i),$$

$$\hat{\lambda}_1(x; h) = \sum_{i=n_0+1}^{n_0+n_1} K_h(x - x_i),$$

where $K_h(u)=h^{-2}K(h^{-1}u)$ and $K(\bullet)$ is a radially symmetric kernel function, we will assume a bivariate standard normal throughout this paper; and h is a smoothing parameter. The smoothing parameter is chosen to minimise some objective function, the objective we will use is the least squares cross-validation. The least squares cross-validation of a kernel estimator of the intensity $\lambda(x)$ is defined as:

$$LS(h) = \int_A \left[\hat{\lambda}(x; h) - \lambda(x; h) \right]^2 dx \quad (2)$$

For complex study regions this is difficult to compute, instead we use the computational form of this given by Bowman and Azzalini (1990, page 35). The computational form is an approximation since the integration is done over the whole plain rather than just the study region, however the effects should be minimal since the estimates decay rapidly to zero outside of the study region.

2.2 Kernel regression estimators

Overviews of kernel regression estimators can be found in a variety of textbooks (e.g. Bowman and Azzalini (1997), Hardle (1990) and Wand and Jones (1995)). These estimators can either be based without assuming a distribution of the binary labels (Y) or by assuming a Bernuolli distribution for the binary labels. We will use the term linear model for any model that results an estimator, of $p(x)$, which is a weighted average of the data points. Models that assume a distribution will be called generalised additive models (GAMs). The advantage of using GAMs is when covariates are available, since we can include these in the model specification.

2.2.1 Linear Model estimators

We shall consider the Nadaraya-Watson regression estimator and the local linear regression estimator, Bowman and Azzalini (1990, chapter 3). The Nadaraya-Watson regression estimator of $p(x)$ is given by

$$\hat{p}(x;h) = \frac{\sum_{i=1}^n K_h(x-x_i)y_i}{\sum_{i=1}^n K_h(x-x_i)}, \quad (3)$$

This has been proposed for binary regression by a number of authors, e.g. Copas (1983).

The local linear regression estimator of $p(x)$ is defined as the value of α for which

$$\phi(\alpha, \beta) = \sum_{i=1}^n [y_i - \alpha - \beta(x - x_i)]^2 K_h(x - x_i),$$

is minimised. The Nadaraya-Watson regression estimator can be derived in this way by omission of the β term. This is a weighted least squares problem, and the solution can easily be given in matrix notation, but it is not computationally convenient and closed form solutions are available.

Unlike most regression problems, in relative risk estimation the null hypothesis of constant risk is quite often true. This can be accommodated under the Nadaraya-Watson regression estimator with infinite smoothing parameter (i.e. $h=\infty$). However, this is not true for the local linear regression model. For the local linear regression model an infinite smoothing parameter corresponds to the ordinary least squares regression estimate.

The Nadaraya-Watson regression estimator compares poorly to the local linear regression estimator in terms of the asymptotic properties of the models (Fan and Gijbels, 1992). These disadvantages relate to having a large bias near the boundary of the study and having a variance that is dependent on the distribution of points. The disadvantages of the local linear regression estimator is that the resulting estimates of $p(x)$ are not constrained to lie within the interval $[0,1]$ and the variance is greater at the boundary than the corresponding Nadaraya-Watson estimator.

2.2.1.1 Choice of smoothing parameter

The choice of the smoothing parameter, h , is of prime importance in the application of kernel based models. Following Kelsall and Diggle (1998) we consider three methods to estimate the smoothing parameter. These are the likelihood cross-validation, least squares cross-validation and weighted least squares cross-validation. These essentially define three different functions that are minimised with respect to h , the resultant value of h is taken as its estimator. The three functions are:

$$CV_{lik}(h) = \left[\prod_{i=1}^n \hat{p}^{-i}(x_i;h)^{y_i} \{1 - \hat{p}^{-i}(x_i;h)\}^{1-y_i} \right]^{-1/n},$$

$$CV_{ls}(h) = \frac{1}{n} \sum_{i=1}^n \{y_i - \hat{p}^{-i}(x_i;h)\}^2,$$

$$CV_{wls} = \frac{1}{n} \sum_{i=1}^n \frac{\{y_i - \hat{p}^{-i}(x_i; h)\}^2}{\hat{p}^{-i}(x_i; h)\{1 - \hat{p}^{-i}(x_i; h)\}},$$

where $\hat{p}^{-i}(x_i; h)$ denotes the estimator computed without the i th data point, and h_0 denotes a preliminary choice of smoothing parameter. The weighted least squares cross-validation criteria requires a preliminary choice of smoothing parameter, since if this took an active role in the minimisation then the resulting value of h would be biased towards high values.

All three of the criteria can be used for the Nadaraya-Watson estimator, but the local linear estimator can result in values of $\hat{p}^{-i}(x_i; h)$ which can lie outwith $[0,1]$ and hence we can only use the least square cross-validation criteria for that model.

2.2.1.2 Effective number of parameters

In order to help compare models it is important to have a measure of the complexity of a model. One such measure is the number of parameters. Hall and Marron (1990) define the effective number of parameters for linear model estimators as

$$n_p = 2 \sum_{i=1}^n w_{ii} - \sum_{i=1}^n \sum_{j=1}^n w_{ij}^2,$$

where w_{ij} is the j th component in vector w_i which is used to estimate $p(x_i)$ via

$$\hat{p}(x_i; h) = w_i y.$$

The value of n_p is such that as $h \rightarrow \infty$ then $n_p \rightarrow 1$ for the Nadaraya-Watson regression estimator; and $n_p \rightarrow 3$ for the local linear regression estimator.

2.3 Generalised additive model estimators

Generalised additive models were popularised by Hastie and Tibshirani (1990). They consist of assuming a parametric model for the data, but the relationship between the response variable and the non-response variable is non-parametric. It is natural to assume a Bernoulli distribution for Y and a logit link function, i.e.

$$Y_i | x_i \sim \text{Bernoulli}(p(x_i)),$$

$$\text{logit}(p(x_i)) = \alpha + g(x_i),$$

The function $g(\bullet)$ can be estimated either by a weighted Nadaraya-Watson regression estimator or a local linear regression estimator.

The algorithm for fitting this model is the local scoring procedure used by Hastie and Tibshirani (1990):

1. Intialise $\hat{g}(x_i; h) = 0$ and $\hat{\alpha} = \text{logit}\{n_1 / (n_0 + n_1)\}$
2. Set $\hat{\eta}_i = \hat{g}(x_i; h)$ and $\hat{p}(x_i) = \exp \hat{\eta}_i / \{1 + \exp \hat{\eta}_i\}$
3. Construct the *adjusted dependent variable*

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{p}(x_i)}{\hat{p}(x_i)[1 - \hat{p}(x_i)]},$$

with weights $w_i = \hat{p}(x_i)\{1 - \hat{p}(x_i)\}$.

4. Fit a weighted linear model using either the Nadraya-Watson regression estimator

$$\hat{g}(x; h) = \frac{\sum_{i=1}^n w_i K_h(x - x_i) y_i}{\sum_{i=1}^n w_i K_h(x - x_i)},$$

or the local linear regression estimator.

5. Repeat these steps until the estimates don't change.

The smoothing parameter in the weighted linear model is estimated by minimising the weighted least squares cross-validation criteria:

$$LS(h) = \frac{1}{n} \sum_{i=1}^n w_i \{z_i - \hat{g}^{-i}(x_i; h)\}^2.$$

These models have the advantage, over the linear models, of being able to handle possible covariates.

Fan et al (1995) discuss the asymptotic properties of GAMs. They demonstrated that the GAM using a Nadaraya-Watson estimator has a large bias at, or near, the boundary, whereas if the GAM is fitted using the local linear model then the GAM has a built-in edge correction. The edge correction of the local linear model comes at the cost of increased variance of the estimate.

2.3.1 Effective number of parameters

The effective number of parameters of generalised additive models is given by

$$n_p = \text{trace}(2R - R^T A R A^{-1}),$$

where

$$A_{ij} = \begin{cases} [\hat{p}(x_i)\{1 - \hat{p}(x_i)\}]^{-1} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases},$$

and R is the projection matrix of Y onto $\hat{p}(x_i)$ see Hastie and Tibshirani (1990), page 157.

2.4 Partition models

In this section we consider a Bayesian partition modelling approach. Partition models have been used in geographical epidemiology by Knorr-Held and Raber (2000) for count data, and by Ferreira et al (2002) for individual level data. A recent review of partition models applied to spatial statistics is provided by Ferreira et al (2002) and for partition models in general the book by Denison et al (2002) is a valuable introductory guide. Partition models work by splitting the study region (A) into k disjoint regions (R_1, \dots, R_k) with a constant conditional probability of being a case, given an event occurs, in each region, say $\phi = (\phi_1, \dots, \phi_k)$ which are assumed to be exchangeable and to come from a common class of distribution f . In this paper we will only consider partitions that are formed by a Voronoi tessellation.

The model can be specified as

$$\begin{aligned} Y | x &\sim \text{Bernoulli}(p(x)), \\ p(x) &= \phi_j, \text{ if } x \in R_j, \\ f(c) &= \frac{1}{K} \frac{1}{|A|^k}, \\ f(\phi) &\propto 1, \end{aligned}$$

where $c=(c_1, \dots, c_k)$ is the vector of centres of the cells of the Voronoi tessellation; k is the number of cells; and K is the maximum number of cells allowed. The prior distribution of c is a distribution for both the location and number of cells. The model can be estimated using the Reversible jump MCMC algorithm of Ferreira et al (2002) and is not repeated here. Like the GAMs this model can be extended to deal with covariates, this is not considered here.

The conditional probability surface is estimated as the marginal posterior distribution of ϕ . That is we integrate the joint posterior distribution over the number and centres of cells. This results in a continuous relative risk surface and does not imply that, unlike the prior surface, the posterior surface has discontinuities, however if discontinuities do exist they can be present in the posterior surface.

3.0 SIMULATION STUDY

In this section we describe a simulation used to investigate the performance of the methods outlined in the previous section. The purpose of this simulation study is to examine how well the various model are at recovering the true relative risk surface. To this end, we will simulate 100 data sets from a variety of relative risk models. All of the models in the previous section will be fitted to the simulated data sets, and the bias and variance will be estimated via the average

It is necessary to automate the model fitting procedures. For the kernel based methods, we compute the cross-validation criterion for 101 values of h , 100 of which regularly in the interval $[0,2]$ and ∞ , and choose the value of h for which the criterion is minimised. For the GAM method, we followed the advice of Kelsall and Diggle (1998) and ran the algorithm for eight iterations, assessing h as above. For the partition model we ran the reversible jump MCMC algorithm starting from the solution $k=1$, $\phi=0.5$ and $c=(0.5,0.5)$ for a burn-in period of 30,000 iterations followed by using the next 20,000 iterations, with a thinning period of 10, for estimation purposes.

We consider the following methods for estimating $p(x)$:

1. Plug-in estimator;
2. Nadaraya-Watson regression estimator with least squares cross-validation;
3. Nadaraya-Watson regression estimator with likelihood cross-validation;
4. Nadaraya-Watson regression estimator with weighted least squares cross-validation;
5. Local linear regression estimator with least squares cross-validation;
6. GAM using a weighted Nadaraya-Watson estimator;
7. GAM using a weighted Local-linear estimator;
8. Partition model.

For this simulation study we consider a unit square on the range $[0,1] \times [0,1]$; fix the control density $f_0(x)$ to be uniform; and fix the number of cases and controls to be 100. We vary the relative risk function $\theta(x)$, which is scaled to integrate to unity over the study region, i.e.

$$\theta^*(x) = \frac{\theta(x)}{\int_A \theta(u) du}.$$

We generate a sample of controls from the density proportional to

$$f_0(x) \propto 1,$$

and the cases are simulated from the density proportional to

$$f_1(x) \propto \theta^*(x) f_0(x),$$

and obtain an estimate $\hat{p}(x)$ using each of the eight methods. This is compared to the true conditional probability

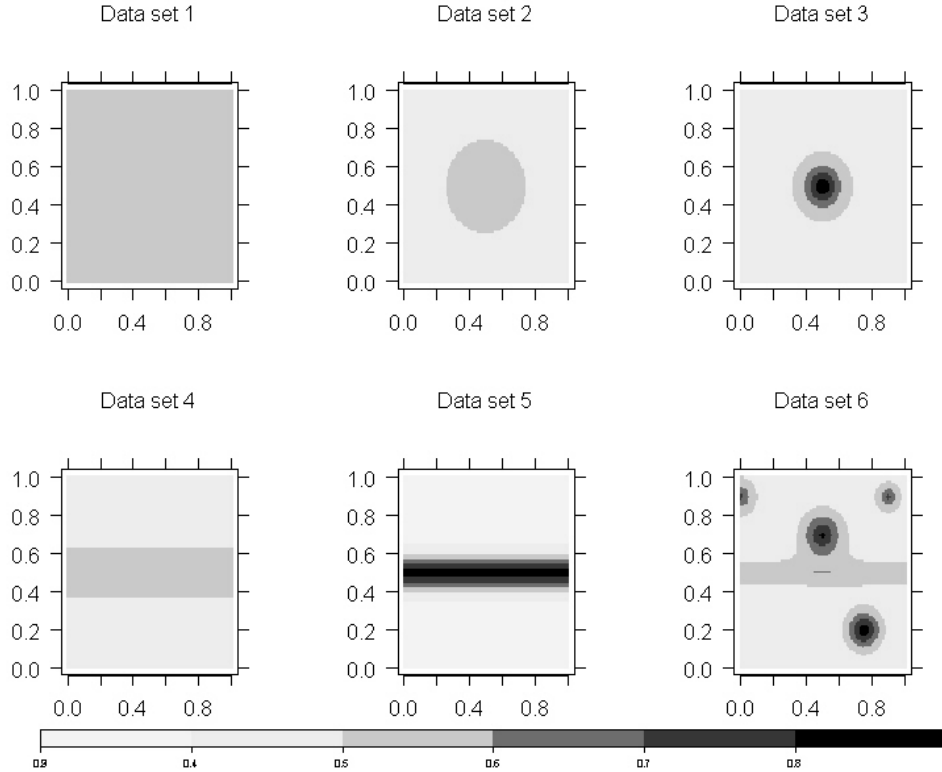
$$p(x) = \frac{\theta^*(x)}{1 + \theta^*(x)}.$$

The models that will be simulated from are:

Constant risk	$\theta(x) = 1$
Mild circular clustering	$\theta(x) = 1 + 0.1 \exp\{-16d(x, (0.5, 0.5))\}$
Strong circular clustering	$\theta(x) = 1 + 10 \exp\{-25d(x, (0.5, 0.5))\}$
Mild linear clustering	$\theta(x) = 1 + 0.1 \exp\{-16d_l(x, 0.5)\}$
Strong linear clustering	$\theta(x) = 1 + 10 \exp\{-25d_l(x, 0.5)\}$
Mixed clustering	$\theta(x) = 1 + 0.1 \exp\{-16d_l(x, 0.5)\}$ $+ 10 \exp\{-25d(x, (0.75, 0.2))\} + 6 \exp\{-18d(x, (0.5, 0.7))\}$ $+ 3 \exp\{-20d(x, (0.0, 0.9))\} + 3 \exp\{-25d(x, (0.9, 0.9))\}$

where $d(x, c)$ denotes the distance from x to the point c ; and $dl(x, 0.5)$ denotes the minimum distance from x to the line $y=0.5$. The first model gives constant probability across the region; the second and third models give circular clustering with probabilities in the ranges $[0.495, 0.519]$ and $[0.307, 0.830]$, respectively; the fourth and fifth models give linear clustering with probabilities in the ranges $[0.497, 0.519]$ and $[0.357, 0.845]$; the sixth model gives a mixture of linear and circular clustering with probabilities in the range $[0.419, 0.881]$ These models are displayed graphically in Figure 1.

Figure 1: Coloured contour plots of the conditional probability surfaces used to generate the data sets.



It is possible to simulate from a much larger set of clustering models, or any other type of model, however we feel that the current set is broad enough to give general results, and yet does not become cumbersome to fully examine the results.

3.1 Measures of comparison

In order to measure the difference between the true and estimate surfaces we can use either point wise or global goodness-of-fit (gof) measures. The point wise gof measures that we will consider are the bias and the variance of $\hat{p}(x)$ at x :

$$\text{bias}[\hat{p}(x)] = E\{\hat{p}(x)\} - p(x),$$

$$\text{var}[\hat{p}(x)] = E\{\hat{p}(x)^2\} - E^2\{\hat{p}(x)\}.$$

The bias is a measure of how far we expect the estimate to be from the true model. In practice we would require that the estimate never be too far away from the truth. The variance is a measure of how variable the estimate is, if the estimate is too variable then it will be not be much use beyond an exploratory tool.

The global gof measures that we will consider are the essentially aggregate measures of the point wise gof measures. Namely, we consider the average bias and variance over the study region.

3.2 Results

In order to fit the models to a large number of datasets it is necessary to automate the fitting procedures. For each relative risk surface we simulate 100 sets of data and use the average of the resulting estimated surfaces to give estimates of $\text{bias}[\hat{p}(x)]$ and $\text{var}[\hat{p}(x)]$.

3.2.1 Global gof results

In order to compute the global gof results it was decided to evaluate the bias and variance of the estimators on a 101 x 101 regularly spaced grid over the study region. The global gof results that we will present are the average of the local gof measures evaluated at these points. Table 1 gives the average value of the absolute bias at the grid points; similarly Table 2 gives the corresponding value for the variance. Table 3 gives the average value of the effective number of parameters; it is a measure of the model complexity. Unfortunately, it is not possible to estimate the effective number of parameters for the plug-in model.

Table 1: Average absolute bias of models fitted to the simulated data described in the text.

	Data Set					
	1	2	3	4	5	6
Plug-in	0.0107	0.0072	0.2798	0.0061	0.2855	0.0900
NW LS	0.0030	0.0065	0.0544	0.0059	0.0661	0.0510
NW WLS	0.0017	0.0042	0.1040	0.0042	0.1132	0.582
NW LIK	0.0026	0.0054	0.0599	0.0049	0.0729	0.0532
Local linear	0.0055	0.0074	0.0409	0.0072	0.0595	0.0482
GAM NW	0.0028	0.0053	0.0560	0.0049	0.0678	0.0527
GAM LL	0.0045	0.0051	0.0376	0.0055	0.0618	0.0516
Partition	0.0052	0.0068	0.1599	0.0067	0.1360	0.0594

Table 2: Average variance of models fitted to the simulated data described in the text.

	Data Set					
	1	2	3	4	5	6
Plug-in	0.0057	0.0078	0.0066	0.0072	0.0069	0.0086
NW LS	0.0021	0.0031	0.0109	0.0024	0.0126	0.0042
NW WLS	0.0005	0.0007	0.0085	0.0008	0.0058	0.0009
NW LIK	0.0017	0.0021	0.0091	0.0016	0.0099	0.0033
Local linear	0.0046	0.0053	0.0142	0.0047	0.0169	0.0068
GAM NW	0.0017	0.0021	0.0102	0.0017	0.0122	0.0036
GAM LL	0.0031	0.0034	0.0124	0.0030	0.0140	0.0042
Partition	0.0032	0.0039	0.0048	0.0037	0.0042	0.0036

Table 3: Average effective number of parameters of the models fitted to the simulated data described in the text.

	Data Set					
	1	2	3	4	5	6
NW LS	3.73	5.41	19.80	4.60	21.51	6.84
NW WLS	1.57	1.78	9.52	2.02	6.40	5.56
NW LIK	3.36	3.97	17.27	3.58	17.58	2.12
Local linear	5.60	5.43	17.94	5.24	19.39	7.56
GAM NW	3.31	4.04	19.67	3.71	21.45	6.02
GAM LL	5.16	5.36	16.40	5.20	16.63	6.13

From these results it appears that:

1. The plug-in estimator is not a suitable estimator for $p(x)$. This could have been expected since the smoothing parameters were chosen to optimally estimate each intensity rather than $p(x)$. We shall not consider the plug-in estimator any subsequent work.
2. Contrary to Kelsall and Diggle (1998) we find that the three methods for the estimation of the smoothing parameter in the Nadaraya-Watson estimator differ substantially. The weighted least squares cross-validation estimator results in unreasonably large values of the smoothing parameter, the least squares cross-validation estimator is more variable than the other two. For these reasons we consider that the likelihood cross-validation procedure should be used. We will only consider this procedure in subsequent work.
3. Local linear approach appears to have smaller bias, but larger variance, over the Nadaraya-Watson estimator when strong clustering is present in the data. This is true for both the linear models and the generalised linear models. However, all models do worse for strong clustering than for mild clustering.
4. The kernel smoothing models that take account of the distribution of the binary labels outperform the methods that do not.
5. The partition models, despite their better theoretical properties, do not perform better than the kernel based methods. This might be due to the partition models taking parameter uncertainty into account, whereas the kernel based methods are conditional upon the estimated value of the smoothing parameter.

In summary, the ‘best’ method appears to be the GAM using the local linear model regardless of whether or not covariates are to be included.

The local linear model was introduced as an edge-correcting estimator. Thus it is important to assess how it performs near the boundary of the study region. In order to do this we define a point to be an edge point if it lies within a distance of 0.1 of the boundary. This is contrary to the common approach, in kernel regression, which defines an edge point as any point for which the support of the kernel intersects the boundary. This is not a reasonable definition in our case since we have a number of model fits with an infinite smoothing parameter, and if the smoothing parameter were infinite then, by the traditional definition, every point would be an edge point.

Tables 4 and 5 are the same as Tables 1 and 2, but the average is taken only over the set of edge points. The general conclusions of the overall averages does not change, however, it is noticeable that in the

case of constant risk, the fit is generally worse in the edge region than overall. The comparison between the gof measures computed on the edge and overall is complicated by the presence of other clustering in data sets 2-6. It is also noticeable that, bar the plug-in model, the results are always more variable with the edge points than with all the points.

Table 4: Average absolute bias in the edge region of the models fitted to the simulated described in the text.

	Data Set					
	1	2	3	4	5	6
Plug-in	0.0090	0.0087	0.3815	0.0059	0.3184	0.1010
NW LS	0.0034	0.0065	0.0452	0.0056	0.0428	0.0514
NW WLS	0.0017	0.0039	0.1136	0.0036	0.1054	0.0582
NW LIK	0.0031	0.0055	0.0508	0.0043	0.0468	0.0534
Local linear	0.0067	0.0088	0.03088	0.0087	0.0411	0.0449
GAM NW	0.0035	0.0055	0.0495	0.0044	0.0462	0.0533
GAM LL	0.0059	0.0058	0.0214	0.0066	0.0408	0.0486
Partition	0.0057	0.0073	0.1897	0.0068	0.1389	0.0573

Table 5: Average variance in the edge region of the models fitted to the simulated data described in the text.

	Data Set					
	1	2	3	4	5	6
Plug-in	0.0057	0.0078	0.0064	0.0071	0.0068	0.0086
NW LS	0.0028	0.0045	0.0168	0.0032	0.0186	0.0060
NW WLS	0.0006	0.0009	0.0123	0.0011	0.0083	0.0013
NW LIK	0.0025	0.0030	0.0142	0.0023	0.0150	0.0047
Local linear	0.0083	0.0096	0.0281	0.0084	0.0337	0.0128
GAM NW	0.0023	0.0030	0.0155	0.0024	0.0178	0.0051
GAM LL	0.0056	0.0063	0.0226	0.0055	0.0262	0.0078
Partition	0.0057	0.0078	0.0064	0.0071	0.0068	0.0086

3.2.2 Pointwise gof results

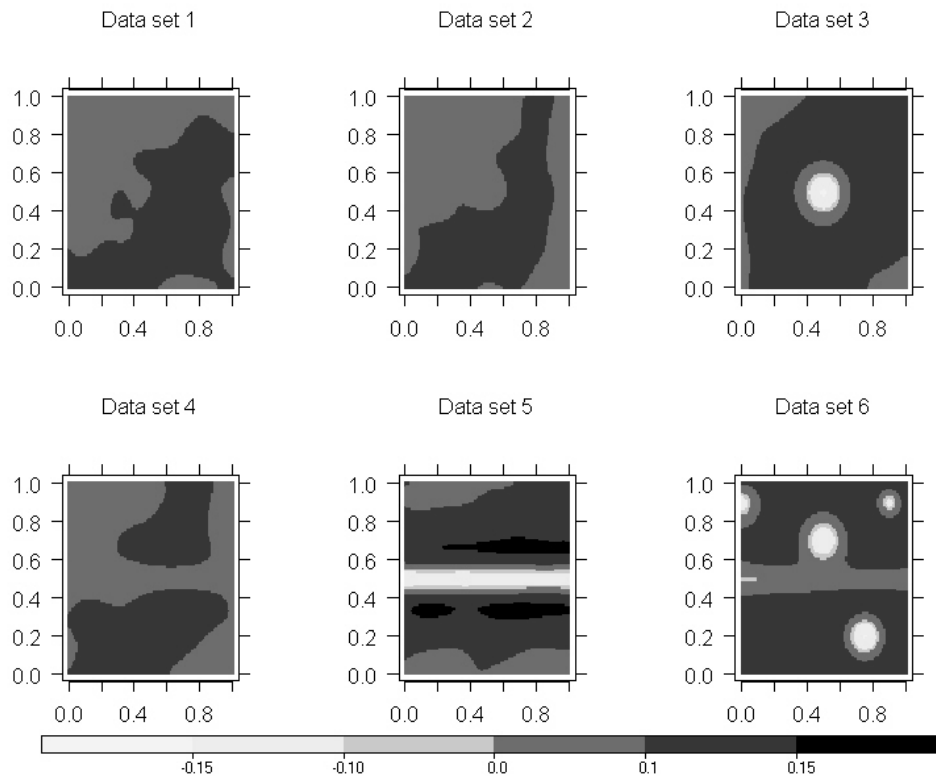
The examination of the spatial pattern of the bias and variance will reveal how the methods differ in the estimation of peaks in the true surface. A number of common features exist with all the estimation methods, namely:

1. they underestimate the height of the conditional probability surface near the peaks, and overestimate when a slope of the surface is quickly decreasing; and
2. as the distance from the edge decreases, the variability of the estimate increases.

From the global gof results the recommended method is the GAM with a local linear model. In this section we will discuss the pointwise gof results for this model only.

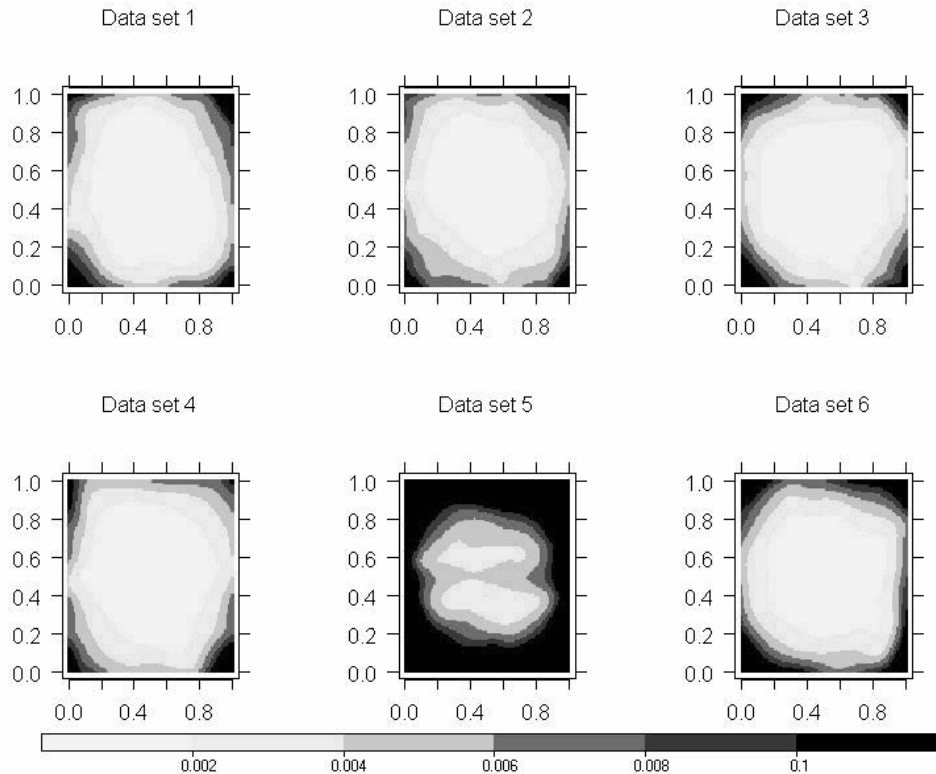
Figure 2 is a set of coloured contour plots of the bias of the fitted GAM with a local linear model for the simulated data sets. For data sets 1,2 and 4, no spatial pattern is immediately obvious, however, it does appear as though the model does underpredict at, or near, the cluster centres for data sets 2 and 4. For data sets 3 and 5 we see that the model severely underpredicts the conditional probability surface, and compensates by overpredicting near the boundary. For data set 6, the model underpredicts at all the cluster centres. Thus for any clustered, or peaked, surface the non-parametric models proposed here are not effective at estimating the true height of the surface for these peaks.

Figure 2: Coloured contour plots of the bias surfaces for the GAM with local linear model for the simulated data sets.



A coloured contour plot of the variability of the estimates is given in Figure 3. It appears as though the variance is stable away from the edge for data sets 1,2,4 and 6, near the edge the variance increases. For data sets 3 and 5 we find that the variance is small near the peaks of the conditional probability surface, but increases away from these peaks. This is due to the fact that most of the cases are near the cluster centres for these data sets.

Figure 3: Coloured contour plots of the variance surface for the GAM with local linear model for the simulated data sets.



4.0 CONCLUSIONS

We have demonstrated and compared a variety of non-parametric methodology for the estimation of the conditional probability surface, $p(x)$. The methods have been compared only in terms of goodness-of-fit. Other considerations may come in to the analysis of real data, for example, the availability of software. The time taken to compute the kernel based estimates is not great, however the partition model is quite cumbersome to handle and is perhaps hard to explain to non-statistical experts.

Of the methods considered it is our believe that the GAM using the local linear regression model should be used in any real analysis, since it can be generalised to deal with covariates and performs well in our simulation study. This importance of using this edge corrected estimator is seen best in Table 4 where the improvements are greater than in Table 1. It is interesting that while the method was introduced as an edge corrected estimator it has the advantage of being able to adapt to highly variable surfaces to a greater degree than the Nadaraya-Watson estimator. It is possible to further reduce this bias by using a higher-order polynomial, however the variability would be increased. Given that the variability of the local linear model is already high any further increase in this variability may render results of little use in applied analysis. Of course in any real analysis it is important to consider a range of smoothing parameters, and not just the optimum.

The models were observed to have problems in estimating the height of the conditional probability surface if the surface was peaked. A better fit might be obtained by one of the two possible extensions.

1. A variable bandwidth parameter may be introduced. The bandwidth can be made variable in a number of different ways. For example, Fan and Gijbels (1995) discuss having a different bandwidth in the different components of a partition of the study region.
2. A parametric model can be proposed for the clustering component of $p(x)$, e.g.

$$\text{logit}p(x) = \alpha + g(x) + \sum_{j=1}^{n_c} \alpha_j C(x - x_j; \kappa_j),$$

where $g(x)$ is a smooth function estimated by the local linear approach, n_c is the number of cluster centres, α_j is the height of the j th cluster centre, $C(\bullet)$ is a cluster distribution with dispersion parameter κ_j . This would be expected to improve the traditional clustering based model, since the non-parametric component will remove a large amount of noise. This model would fit into the GAM framework.

ACKNOWLEDGEMENTS

This work was made possible by the support of NIH grant number: 5R01CA092693-2.

REFERENCES

- Bowman, A and Azzalini, A (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press.
- Fan, J. and Gijbels, I (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, **20**, 2008-2036.
- Fan, J. and Gijbels, I. (1995). *Local polynomial Modelling and Its applications*. Chapman and Hall.
- Fan, J., Heckman, N. and Wand, M. (1995). Local polynomial kernel regression for generalised linear models and quasi-likelihood functions. *Journal of the American Statistical Association*. **90**, 141-150.
- Ferreira, J., Denison, D and Holmes, C. (2002). Partition modeling. In Lawson, A and Denison, D (Eds). *Spatial Cluster Modelling*, pp 125-146. London: Chapman and Hall.
- Hall, P and Marron, J. (1990) On variance estimation in nonparametric regression. *Biometrika*, **77**, 415-419.
- Hardle, W. (1990) *Applied Non-parametric regression*. Cambridge University Press.
- Hastie, T. and Tibshirani (1990). *Generalised additive models*. Chapman and Hall.
- Kelsall, J. and Diggle, P. (1998) Spatial variation in risk of disease. *Applied Statistics*, **47**, 559-573.
- Knorr-held, L. and Rabe, G. (2000) Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, **56**, 13-21.
- Lawson, A and Cressie, N. (2000) Spatial statistical methods for environmental epidemiology. In C. Rao and P. Sen (Eds). *Handbook of Statistics: Bio-Environmental and Public Health Statistics*, volume 18, pp 357-396. Elsevier.
- Wand, M and Jones, M (1995) *Kernel Smoothing*. Chapman and Hall.