Transactions in GIS

Transactions in GIS, 2011, 15(3): 385–400

Research Article

Identifying Discontinuities in Trend Surfaces Using Bilateral Kernel Regression

Chris Brunsdon
Department of Geography
University of Leicester

Abstract
Following a brief review of the kernel regression approach to estimating surface models of the form $z = f(x,y) + \varepsilon$, this article will consider the situation where f is not a continuous surface function, and in particular where the discontinuities take the form of one-dimensional breaks in the surface, and are not specified *a priori*. This form of model is particularly useful when visualizing some social and economic data where very rapid changes in geographical characteristics may occur – such as crime rates or house prices. The article briefly reviews approaches to this problem and proposes a novel approach (Bilateral Kernel Regression) adapting an algorithm from the field field of image processing (Bilateral Filtering), giving example analyses of synthetic and real-world data. Techniques for enhancing the basic algorithm are also considered.

1 Introduction

There are a large number of techniques that are used to fit smooth surfaces to observational point data of the form (x_i,y_i,z_i), where x_i and y_i specify a location in space and z_i is some real-valued attribute of that space, for $i \in 1 \ldots n$ observations. For example x_i and y_i may be obtained from the postcode of a recorded house purchase i, with z_i being the purchase price of the house. Alternatively, (x_i,y_i) may be the centroid of a small areal unit, such as a UK Census Output Area, and z may be an attribute associated with that area such as a Townsend score for deprivation (Townsend et al. 1988). The aim of such an analysis is usually to obtain a visual image of geographical trends in the z value – for example to identify areas of very high house price or very low house price by overlaying such a visual representation semi-transparently over a map of the area under study, using

Address for correspondence: Chris Brunsdon, Department of Geography, University of Leicester, Leicester, LE1 7RH, UK. E-mail: cb179@le.ac.uk

doi: 10.1111/j.1467-9671.2011.01241.x

for example the Google Maps API, or materials provided by OS OpenData™ (or representing the trend as a three-dimensional surface). Broadly, the data model used for this kind of analysis takes the form

$$z_i = f(x_i, y_i) + \varepsilon_i \quad (1)$$

where f is a continuous (usually smooth) function, and ε_i is a random error term. Generally also $E(\varepsilon_i) = 0$ so that the random errors represent fluctuation centered around the trend described by f. However, in some situations, geographical changes in trends may not be smooth – for example when crossing the boundaries of 'golden postcodes' in the case of house prices – and in this case sudden jumps or breaks in the trend surface may be a more appropriate representation of the data model. Furthermore, since the aim of this modelling is to explore and identify potential discontinuities, one cannot assume that the location of such breaks are known *a priori*. To accommodate this it is necessary to relax the requirement of continuity for the trend surface function f and to require that any algorithm used to fit this kind of model will be able to identify locations for any potential discontinuities.

An alternative approach might be to aggregate data to a small set of areal units, and identify adjacent zones with notably different average z values. However, an issue here is that if the original data are in point form, it is difficult to account for distortions in pattern due to the modifiable areal unit problem (MAUP) (Openshaw, 1984) – hence a surface-based approach is advocated here. Another alternative might be to use interpolation methods that do not smooth at all – such as nearest neighbour methods or bilinear interpolation – these are different from smoothers in that they pass through every (x_i,y_i,z_i) triplet exactly, effectively assuming that $\varepsilon_i = 0$ for all i. Smith et al. (2005) demonstrate the use of these and other interpolators with urban LIDAR data, but in the context of social and economic data it is expected that there will be a relatively high degree of error, and the $\varepsilon_i = 0$ assumption is not realistic.

The article begins by providing a brief overview of trend surface modelling algorithms, then considers how these may be adapted to identify discontinuities – and how potential discontinuities may be validated. Examples are then given, together with a discussion of possible approaches to validation.

2 An Overview of Trend Surface Modelling

As noted earlier, a typical trend surface fitting problem has a data model of the form $z_i = f(x_i,y_i) + \varepsilon_i$, where no functional form for f is given. It is worth noting that the very general form of the model means that even for the case for f without discontinuities, typical approaches to estimation tend to be based on nonparametric methods, rather than perhaps more familiar approaches such as least squares or maximum likelihood estimates for a given functional form of f. Since f is allowed to take any functional form, given a finite data set a perfect fit can be achieved simply by specifying f that passes exactly through (x_i,y_i,z_i) for $i \in \{1 \ldots n\}$. This satisfies an unconstrained goodness-of-fit estimation rule, but is very likely to *overfit* the model, so that f describes the 'noise' component of the model in addition to the 'trend' component it was intended to identify. For example, when spatial trends in house price are being modelled, a number of non-spatial characteristics also influence the value of a house, such as its state of

decoration, or the urgency with which the owner wants to make a sale. In this case, overfitting a geographical trend might amount to producing a surface that responds to these kinds of fluctuations as they vary within streets or neighbourhood blocks. The phenomenon is of key importance in machine learning, where checks against overfitting are routinely carried out. Overfitting is also characterized by poor predictive ability of the fitted model. Although an overfitted model would provide a very good fit for the data points (x_i,y_i,z_i) for $i \in \{1 \ldots n\}$ it would not necessarily perform well for a new observation $(x_{n+1},y_{n+1},z_{n+1})$ – see for example Duda et al. (2000) or Hastie et al. (2001).

Since maximum likelihood approaches tend to lead directly to overfitting in this scenario, alternative approaches are used. One very common choice is the use of a technique known as *non-parametric regression* or *moving window smoothing*. Here, to estimate $f(x,y)$ at some new point (x,y) the estimator is:

$$\hat{f}(x, y) = \frac{\sum_i w_i z_i}{\sum_i w_i} \tag{2}$$

where

$$w_i = \exp\left[-\frac{(x_i - x)^2 + (y_i - y)^2}{2h^2}\right]. \tag{3}$$

This is a weighted mean of the observed z-values where the weights are a kernel centered on the location (x,y). The quantity h (the *bandwidth*) controls the degree of smoothing. Smaller values of h tend to undersmooth, possibly resulting in overfitting; but uneccesarily large values tend to oversmooth so that genuine features of the trend surface may be 'smoothed out'. Thus, h may be thought of as a tuning parameter – optimal values need to be discovered to allow the algorithm to perform as well as possible. Although non-parametric regression originates from the statistical literature, it has a great deal in common with machine learning approaches. In particular choosing parameters to avoid overfitting is a problem frequently encountered in the latter field.

An example of this approach is now given, using house price data from Leicester via Houseprices.co.uk[1] which has been postcoded, allowing location to be estimated by postcode centroids. The data relates to house sales in the post code areas LE1, LE2, LE3, LE4 and LE5, roughly corresponding to the area defined by Leicester City Council, see Figure 1. The prices are for terraced house sales during 2008, and there are 1123 sales in total.

Clearly there is some degree of spatial trend, with a number of clusters of high house prices (dark green) or low house prices at various locations on the map. It is also interesting to note that on the western side there is a cluster of relatively high prices just to the south of a cluster of low prices. It is also worth noting that changes in selling price do not necessarily correspond to potential physical dividing lines, such as main roads or areas without buildings.

To illustrate the issue of overfitting a kernel regression model is applied to this data, with (x_i,y_i) defined to be the centroid of OA i, and z_i the Townsend score for location i. Here $n = 1123$. The data is randomly partitioned into two subsets – a *training set* used to calibrate the model according to equations 2 and 3 (of size 561), and a *validation set* of size 562. The z-values for the latter are not used in the calibration, but are predicted

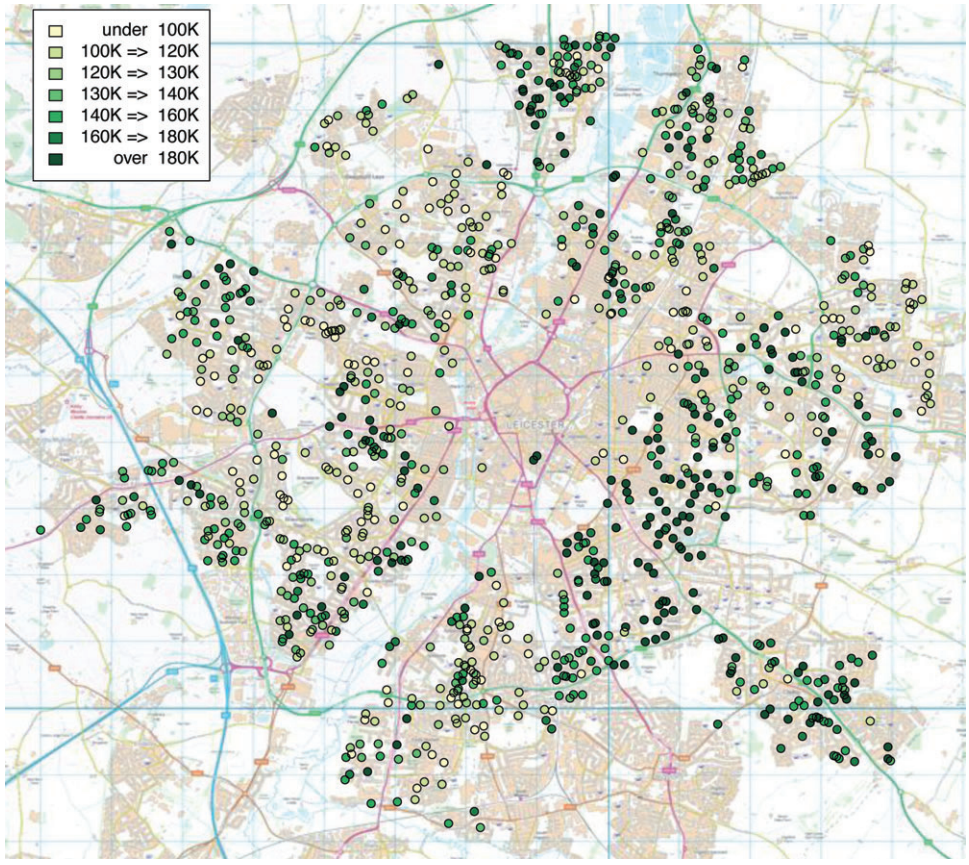

Figure 1 House price data for Leicester City and surrounding areas; background map is obtained from OS OpenData™ (Contains Ordnance Survey data © Crown copyright and database right 2010)

for each (x,y) pair. Next, the mean squared difference between the predicted z values and the true values is computed (referred to as the mean squared error or MSE). This is a measure of the predictive performance of the technique, and can be computed for a range of bandwidths. The results of doing this are shown in figure 2. The solid line shows the relationship between h and the MSE for the validation data set, while the dotted line shows the same relationship for the calibration set. As one might expect, for the calibration set the MSE approaches zero as the bandwidth decreases – as the regression surface becomes more flexible and closely matches the observed (x_i,y_i,z_i) points. However, for the validation set the gain in predictive ability is lost once the bandwidth drops below around 300m – suggesting that for bandwidths lower than this approximate value, overfitting occurs. Overfitting can also occur in a number of other surface modelling contexts – for example modelling f as a polynomial and including too many higher degree terms, under-penalising roughness in a spline-based approach (Wahba, 1990) or applying too much training using a specific data set for a machine learning algorithm.

In fact, using optimization software[2] the optimal MSE is found when $h = 390.8$ km. The surface using this bandwidth may then be fitted to a regular lattice of grids points

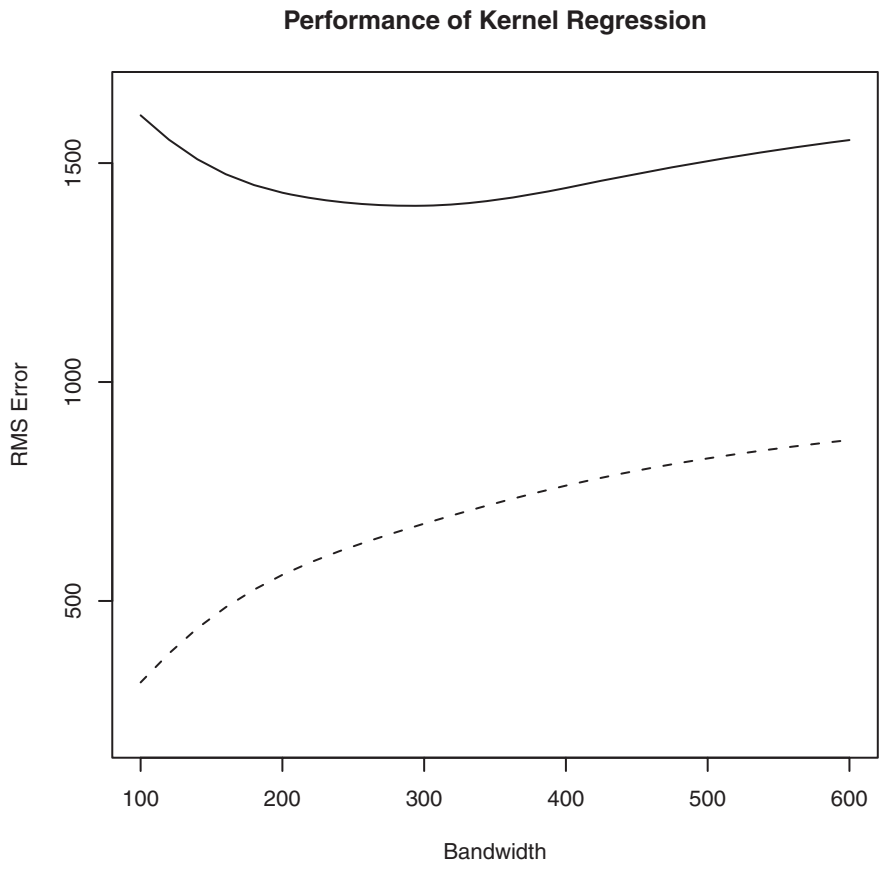

Figure 2 Performance of Kernel Regression against bandwidth for training data (solid line) and validation data (dotted line)

and visualised – as shown in Figure 3[3]. Note that even with cross-validation, the optimal bandwidth is relatively small, leading to a large number of peaks and troughs in the surface. Two observations may be made:

1. Leicester consists of a number of areas with few houses – commercial areas or green space for example, as well as residential areas. In the former, data is notably less densely distributed.
2. In the residential areas generally speaking there is a tendency for housing of similar price to cluster, but looking at both Figures 1 and 3 it is apparent that there are some rapid changes in levels of price as well as some geographical trends. To predict the rapid changes a reasonably small bandwidth is necessary.

The first of these observations may be addressed by applying an *adaptive* kernel regression technique. Here, the bandwidth is allowed to vary as a function of (x,y) so that in regions with more densely arranged data, a smaller bandwidth can be used, but in the sparser rural areas, larger bandwidths, and hence greater smoothing can apply. One simple way of achieving this is to replace Equation (3) in the standard kernel regression method with:

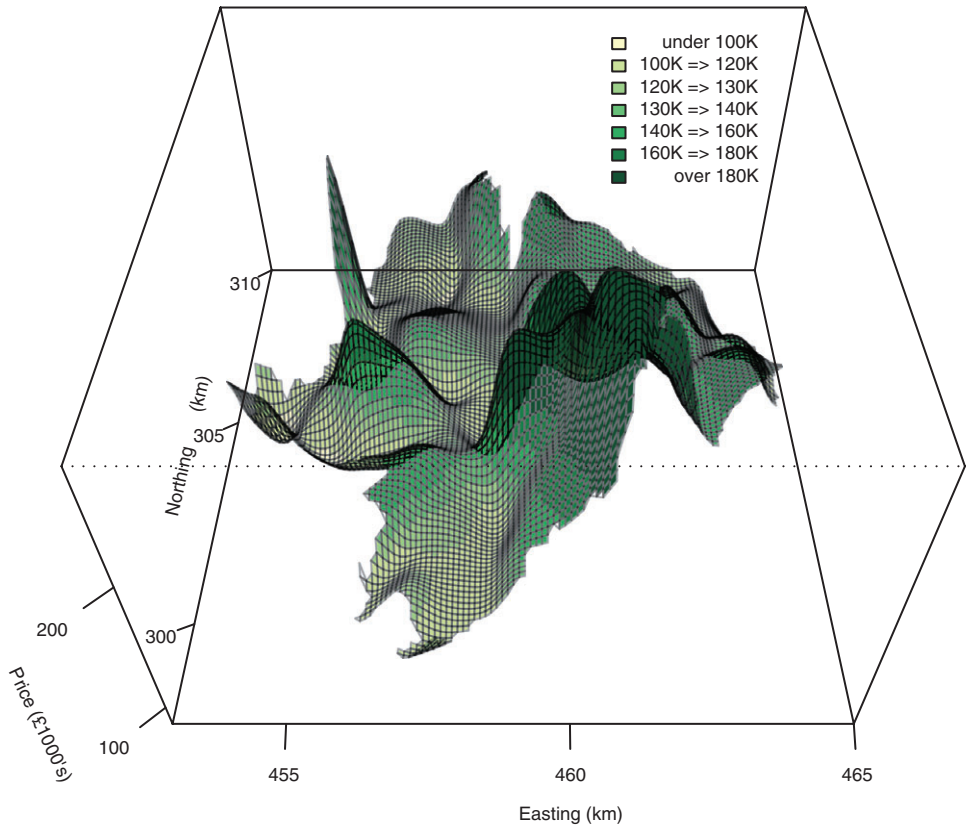

Figure 3 Fitted house price surface using kernel regression. Colour interpretation is as in Figure 1

$$w_i = \exp\left[-\frac{\mathrm{RANK}_{i\in 1\ldots n}\left[(x_i - x)^2 + (y_i - y)^2\right]}{2h^2}\right]. \tag{4}$$

where RANK[. . .] denotes that the quantity $(x_i - x)^2 + (y_i - y)^2$ for observation i is replaced with the rank of its value in the set of all n observations, where n is the rank of the largest observation and 1 is the rank of the smallest. Now h still corresponds to a bandwidth, but instead of having the dimension of a distance, it is now dimensionless. The weighting around a point (x,y) now depends on the rank of the distance from each (x_i,y_i) to (x,y) so that the kth nearest point always has the same weight regardless of the distance. Using the MSE minimizing approach applied previously, in this case the optimal h is 1.24 and the surface obtained is shown in Figure 4. While there are similar patterns in the residential areas, this has had the effect of greater smoothing in other areas. However, one issue with this approach is that the very low value of h here has led to 'stippled' patterns in the surface – effectively the weights are placed on only one or two nearest neighbours of each point, and this suggest that to some extent the 'noise' component ε_i is being included in the estimate of f. It seems some degree of undersmoothing is occurring.

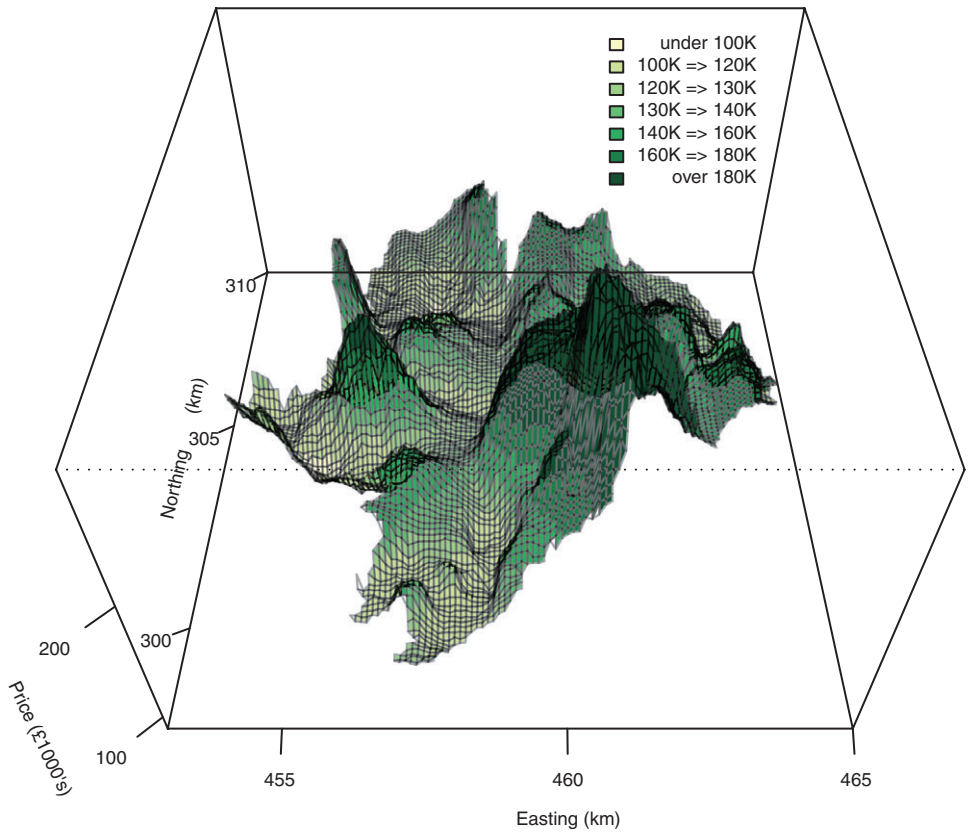

Figure 4 Fitted house price surface using adaptive kernel regression

Also, the second observation made above presents further issues that may not be easily addressed using the above approach. Essentially any regression surface produced using either standard or adaptive kernel regression will be smooth. To see this, note that Equations 3 and 5 are continuous and differentiable functions of x and y. Since in each equation $w_i > 0$ for all i, then Equation 2 is also a continuous differentiable function of x and y. Thus, regardless of the choice of h and the values of (x_i,y_i,z_i) the estimated trend function will inevitably be a continuous and differentiable function – that is, a *smooth* function. Thus, these approaches are not well equipped to model sudden changes in geographical trends.

Note that this is also the case for a number of other approaches. Another common approach is to approximate f using spline functions regularized by a roughness penalty (see for example Wahba 1990, Wood 2003 or Wood 2006). However, again the splines used are continuous, and if they are based on second or third order polynomials (as is often the practice) they will also be differentiable. A further alternative might be to consider geostatistical approaches – see for example Diggle and Ribeiro (2007) – but although such approaches can be very successful in general, in practice these also tend to smooth out discontinuities.

It is possible to adapt the spline approach to allow for 'barriers' – these are essentially boundaries supplied *a priori* over which no smoothing is applied. For example, ArcGIS Spatial Analyst provides such a function, where barriers are supplied by the user as either polygons or polylines, based on a method by Terzopoulos (1988). If there was prior evidence that, for example, a specific road, or river provided a physical barrier that was reflected in a discontinuity in the trend in house price, then this approach would be appropriate. However in Figure 1, it seems that although sometimes jumps in price do correspond to physical barriers, this is not always the case – there are jumps in other places, and some physical features over which no jump occurs. Thus, in the situation considered here, the location of discontinuities will not be assumed in advance, and the aim is to identify these purely on the basis of the observed (x_i,y_i,z_i) data. To address this issue, methods of smoothing that allow for, and detect, discontinuity will be considered. These will be discussed in the following sections.

3 Smoothing Allowing for Discontinuities – Bilateral Kernel Regression

In the previous section, approaches to smoothing that assumed $f(.,.)$ in Equation 1 were continuous functions. Here, methods of estimation when that assumption is relaxed will be considered. In general, the problem with the window regression approach in the presence of discontinuous trends is that when estimates are constructed, information from both sides of the discontinuity are used – so that trends 'bleed' over the discontinuity. This leads to two problems. Firstly, predicted z-values close to the discontinuity will be inaccurate, and secondly from a visual viewpoint, if the surface is plotted the discontinuity will not be evident.

A number of approaches to address this issue have been proposed – for example Crampton and Mason (2005) propose a method based on a modification of the support vector machine approach (Vapnik 1998). Bowman et al. (2006) propose a technique based on the comparison of several half-plane kernels – where at a given point (x,y) kernels are set to zero on a number of half-planes whose boundary passes through (x,y) – and differences are noted in the estimate of f for each kernel. Sufficient variation provides evidence that the point (x,y) lies on or near to a discontinuity. Both of these methods are relatively computationally intensive. Although this may yield specific advantages (for example the latter approach provides a means for statistically assessing the significance of the discontinuities), an alternative approach is set out here which, although cruder than these methods, is relatively straightforward to implement, requires only slightly more computational effort than ordinary kernel regression, and appears to work well in practice.

For this method, a modification is proposed to the moving window smoother than attempts to avoid smoothing over discontinuities. Assume that there are a number of locations with observed (x,y,z) values, and that we want to estimate $f(x,y)$ at these locations. Given the data model in Equation 1, this is equivalent to 'removing the noise' from the observed z-values. The general idea is that the smoother still takes the form of Equation 2, but that Equation 3 is replaced with:

$$w_i = I\{|z - z_i| < \delta\} \exp\left[-\frac{(x_i - x)^2 + (y_i - y)^2}{2h^2}\right] \tag{5}$$

where I {·} is equal to 1 if the expression in the braces is true, and equal to zero otherwise. The role of this multiplier is to remove the influence of any observations whose z_i value is very different from that at location (x,y,z). Exactly *how* different the observations have to be before this occurs is determined by a second tuning parameter δ.

However, there may be situations in which we have no information about z in a location (x,y), but still wish to estimate $f(x,y)$ at that location. For example, to visualise f as a trend surface, we may wish to obtain estimates on a regular grid. Using the method set out above, this cannot be done directly due to the lack of observed z-values at these locations. In a further modification to the method, suppose we have a new location (x^*,y^*,z^*) for which (x^*,y^*) are given, but z^* is unobserved. In this case, z^* is approximate by $\hat{z}^*$, where:

$$\hat{z}^* = z_i, \text{ where } i \text{ minimises} (x^* - x_i)^2 + (y^* - y_i)^2 \tag{6}$$

That is, z^* is approximated by the closest known observation to the point (x^*,y^*) – in practical terms the closest geographical point. This estimate is then 'plugged in' to the approach outlined above to obtain an estimate of $f(x,y)$ for any point (x,y) regardless of whether a z-variable exists there.

The technique described here has a number of similarities with an approach from the field of image processing, known as *Bilateral Filtering* introduced by Tomasi and Randuchi (1998). Here a similar approach is applied to pixel-based image data, the aim being to filter out noise in images, while preserving edges of objects. The two main differences here are that:

1. Data are irregular points, not z-values arranged in a rectangular grid and predictions are made at (x,y) locations where no z has been observed.
2. The function applied to the z-domain here is discontinuous, whilst in the original bilateral filtering, a smooth function was used. The reason a step function was used here is that empirical experimentation showed it to give better results.

To acknowledge the influence of Bilateral Filtering in the technique proposed here, it will be referred to as *Bilateral Kernel Regression* (BKR) for the remainder of this article. Some observations about this approach may be made:

1. If (x^*,y^*) coincides with a point (x_i,y_i) in the data set then $\hat{z}^* = z_i$ and the estimate obtained for $\hat{f}(x_i, y_i)$ will be identical to the one obtained from the basic bilateral filter approach.
2. If δ is larger than the maximum value of $|z_i - z_j|$ for any i and j then the estimate for $\hat{f}(x^*, y^*)$ will be identical to that obtained using kernel regression.
3. Any 'fault lines' in the fitted surface – over which no averaging occurs – will coincide with edges in a Voronoi tesselation (Okabe et al. 2000) of the locations (x_i,y_i), due to the way that $\hat{z}^*$ is estimated.

Whilst observations 1 and 2 may be thought of as advantageous, perhaps observation 3 is less so. Effectively 1 states that the method using the estimate $\hat{z}^*$ generalises the approach when the z-values are observed, in that it gives identical results if used in that situation, but extends the method to alternative situations. Similarly, 2 states that the method is a generalization of ordinary kernel regression, in that if δ is larger than any of the observed differences in z values, then the results are identical, but again, an alternative model can also be assessed by using smaller values of δ. However, observation 3

states that any observed faultlines will be any artefact of the x and y coordinates of the observed data. Possibly for large samples this will not present a problem, but care should be taken when considering small samples. It also suggests that in some situations where there is an idea of where potential faultlines might lie, and it is possible to exert some control of the locations where observations may be made, then experimental design should perhaps be mindful of observation 3. In particular, although the locations of observation points do not need to be arranged on a regular lattice, areas where they are more sparsely distributed will yield relatively crude estimates of faultlines. Thus if there is some idea of where a faultline might occur, priority should be given to sampling in this region. Also, if there is little prior knowledge of faultline locations, care should be taken to ensure that the point locations are of reasonably uniform density.

3.1 Demonstration of the Method with Synthetic Data

In this section the above method will be demonstrated for synthetic values of $f(x,y)$. In particular, two cases will be considered:

$$f_1(x, y) = \begin{cases} 1 & \text{if } y > 0.5 \\ 0 & \text{if } y \leq 0.5 \end{cases} \tag{7}$$

and

$$f_2(x, y) = 0.5 \tag{8}$$

The first of these contains a discontinuity along the line $y = 0.5$, the second has no discontinuities. In each case, a data set was created by generating (x_i,y_i)s drawn from the uniform distribution on $[0, 1] \times [0, 1]$ and then generating $z_i = f_m(x_i,y_i) + \varepsilon_i$ where $\varepsilon_i \sim N(0, 0.05^2)$ and m was equal to 1 or 2. In each case, data sets of size $n = 500$ were created, and randomly split into training and evaluation components of equal size. As before, tuning parameters were chosen to optimize the MSE for prediction on the validation data set. Since there are now two tuning parameters (h and δ) the problem is now one of multidimensional optimization. In practice, h was optimised using the univariate algorithm applied before, for a sequence of values of δ, and the best (h,δ) combination noted. For f_1 the optimal pairing was (0.096, 0.5) while for f_2 it was (35.0, 1.0). Applying the algorithm set out in Section 3 subject to this approach gave the reconstructions shown in Figure 5.

Generally the results are good – identifying a discontinuity in f_1 and equally importantly not finding one in f_2. The shape of the modelled discontinuity for f_1 is considerably more complex than the true shape (which is simply a straight line) – but this is a consequence of the 'voronoi discontinuity' property discussed earlier.

3.2 Demonstration of the Method with House Price Data

In the previous section, BKR was applied to synthetic data sets. In this section it will be applied to the house price data as used in Section 2. As in Section 3.1, the tuning parameters were chosen by stepping through values of δ, at each step finding the optimal bandwidth, noting the best performing pair. As in the previous examples the data were divided into training and evaluation sets – here the data were divided exactly as in

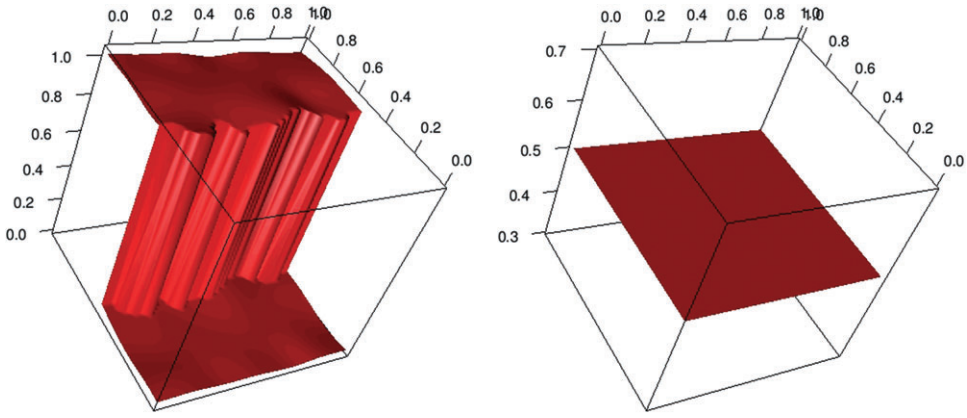

Figure 5 Reconstruction of synthetic functions f_1 (left) and f_2 (right) using BKR

Section 2, so that results may be compared directly. Here, the optimal values were (h,δ) = (495m, 152K pounds) and the MSE is 1325.5 pounds^2. The surface is shown in Figure 6.

This method has outperformed the optimal standard kernel regression approach, where the MSE was 1346.5 pounds^2.

4 Further Issues

In this section, two further issues associated with bilateral kernel regression will be considered. Both of these relate to how the method can be extended using bootstrap-based approaches.

4.1 Bagging

One issue noted earlier is the fact that any fault lines identified will lie along the edges of a voronoi tesselation of the (x_i,y_i) samples – and therefore observed geographic form in discontinuities will be an artefact of this. However, if we consider the observation locations as a random sample (in the house price example, they could be a random sample of houses which may be sold in a given year) then the estimated locations of the faultlines are themselves random (in the same way that, in classical statistical inference, although a population mean is a fixed quantity, a sample estimate is a random quantity), and the effect of faultlines on the estimated surfaces are also random. If the sample size is large enough, it is hoped that the degree of bias in estimating f is small, and that on average, local estimates $\hat{f}(x, y)$ for any location (x,y) will be reasonably close to the true $f(x,y)$. That is:

$$E(\hat{f}(x, y)) \approx f(x, y) \tag{9}$$

If this is a reasonable assumption, and it were possible to generate samples from the distribution of $\{(x_1,y_1,z_1), \ldots (x_n,y_n,z_n)\}$ and if we were to draw N such samples, labelled $j = 1..N$, and from each of these, derive a BKR estimate of f, labelled $\hat{f}_j$ then Equation 9 implies that:

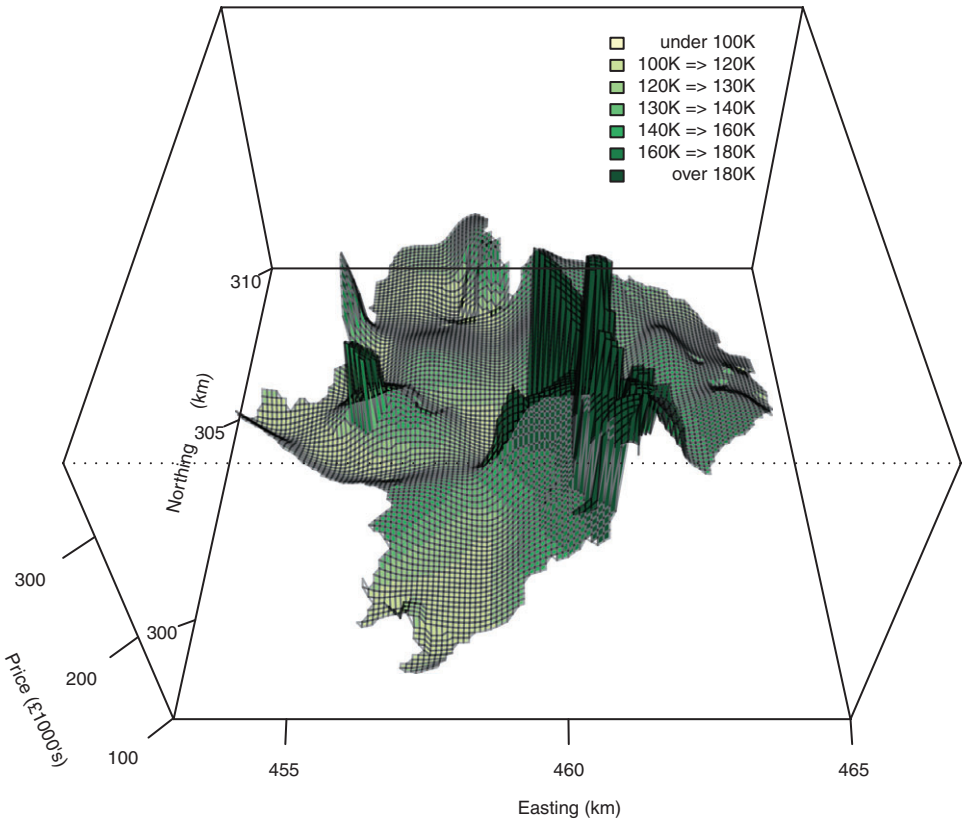

Figure 6 Fitted house price surface using bilateral kernel regression

$$E(\hat{f}(x, y)) \approx \frac{1}{N}\sum_{i}\hat{f}_i(x, y) \tag{10}$$

In reality, the distribution of $\{(x_1,y_1,z_1), \ldots (x_n,y_n,z_n)\}$ is not known, but it can be approximated using a *bootstrapping* approach (Efron 1979, 1982). That is, a sample of n draws *with replacement* of (x_i,y_i,z_i) triplets from the set $\{(x_1,y_1,z_1), \ldots (x_n,y_n,z_n)\}$ is used to approximate a draw from the underlying distribution – and produce a *pseudo-sample*. N of these pseudo-samples are then used in Equation 10 to produce an new approximation to $f(x,y)$. Using bootstrapping in this particular way was proposed originally by Breiman (1996) and is referred to as *bagging* – an abbreviation of **boot**strap **agg**regat**ing**. In several situations, the technique is able to reduce the MSE achieved with the base regression technique.

Here, the technique was applied to the house price data, with $N = 50^4$ yielding the result illustrated in Figure 7. Using the same validation and training data as before, the MSE achieved was 1323.9 pounds^2 which is a marginal improvement on the original bilateral kernel regression performance. Perhaps more notable, comparing this result to Figure 6, the effect of the discontinuities has been downplayed to an extent, although some are still apparent. This approach has perhaps avoided the qualitative problem of

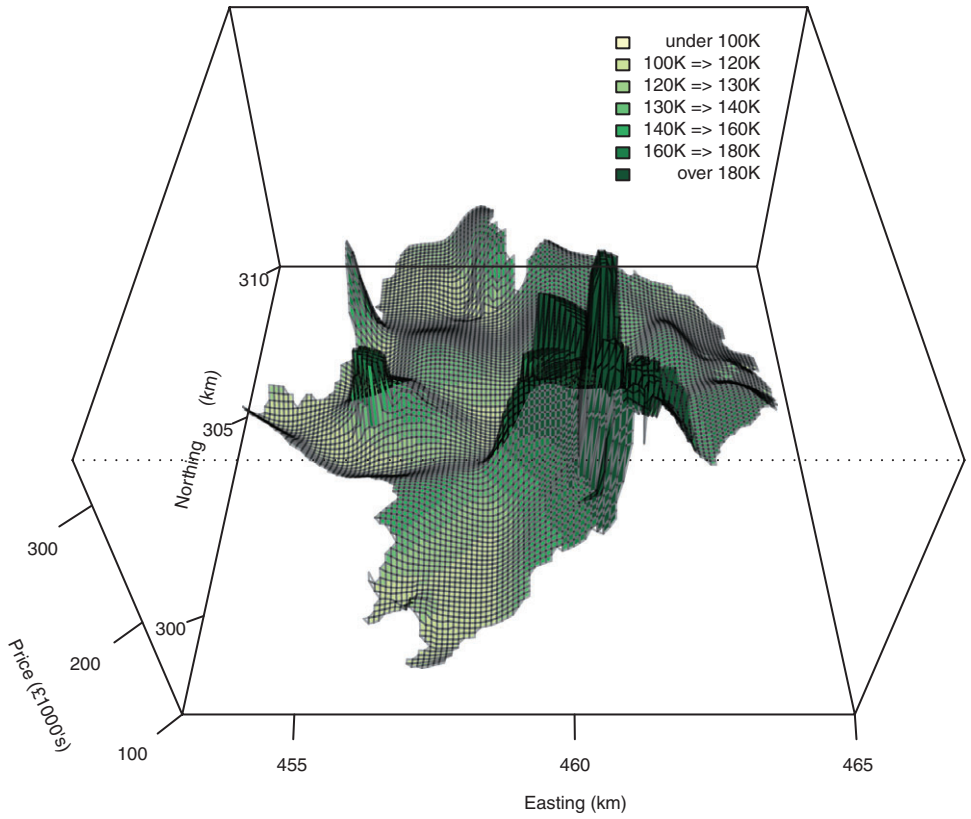

Figure 7 Fitted house price surface using bagged bilateral kernel regression

spurious 'spikes' appearing in the surface due to a very small number of outlying observations.

4.2 Standard Errors via the Bootstrap

Some of the computation needed for the bagging approach described in Section 4.1 may be put to further use. The original use of bootstrapping was as a method to estimate standard errors of statistics. The technique described in the previous section for generating pseudo-samples can be used to simulate the sampling properties of a given statistic. Suppose D is the real data set and there exists some statistic $s(D)$ – and that D_j^* for $j = 1 \ldots N$ are N pseudo samples generated as set out earlier, then $\{s(D_i^*)\}$ are a random sample of statistics whose distribution is an approximation of the true sampling distribution for s. Computing the standard deviation of these values is then a method to approximate the standard error of s.

Suppose s is the estimate $\hat{f}(x, y)$ for any given (x,y). It is then possible to estimate the standard error of f via the bootstrapping approach in this way. Since the bootstrap samples will be drawn in any case, if a bagged estimator is to be used, the same samples can be used to obtain estimates of standard error on a regression surface. The results of this approach applied to the house price data are shown in Figure 8. One

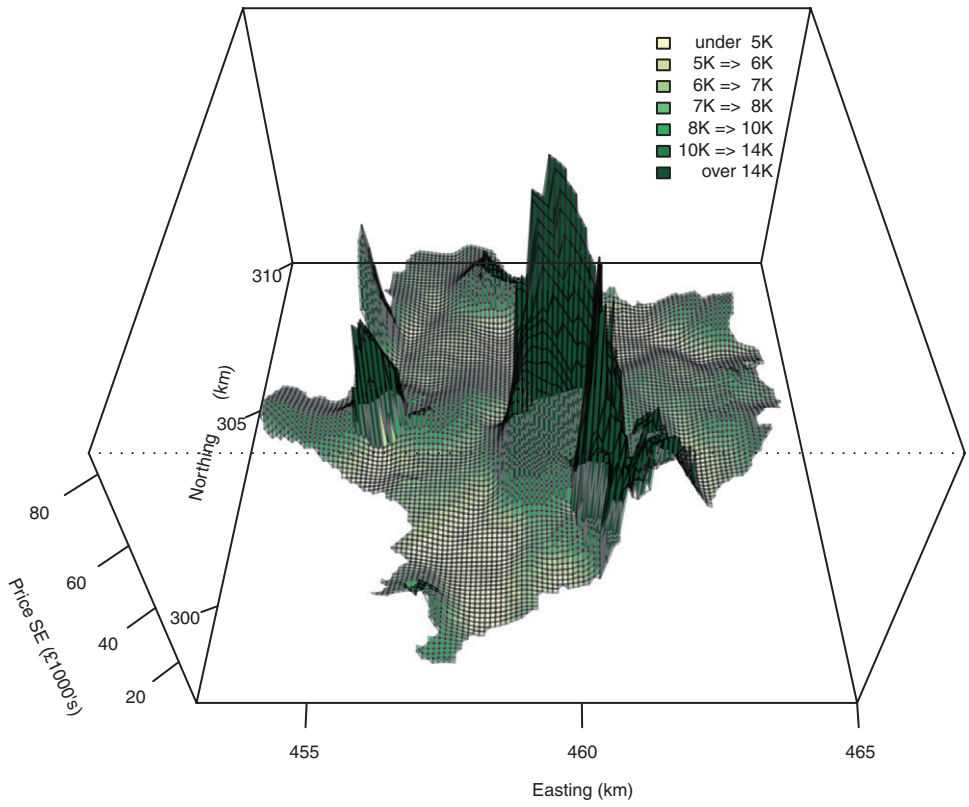

Figure 8 Standard errors of Bilateral Kernel regression from bootstrap estimates

interesting observation is that the regions nearest the 'cliff edges' show the highest standard error. This seems reasonable, as there is some uncertainty about the precise location of the discontinuity – and therefore for (x,y) locations near potential discontinuities there is uncertainty as to which side of the divide they are situated. This in turn implies a greater degree of uncertainty about $f(x,y)$, which is reflected in the higher standard error.

5 Conclusions

In this short article, a relatively simple method of surface fitting with the ability to detect discontinuities was introduced. The method can be implemented with a relatively low computational demand, as it is essentially a minor modification of kernel regression. The method can be augmented by using bagging, and standard errors can be estimated via the bootstrap. At this stage, the method is proposed on an *ad hoc* basis, with little theoretical background. For example, it is assumed that the effect of bias will be minimal in Equation 9, although this has not been formally demonstrated. This has been demonstrated for ordinary kernel regression (Staniswalis 1987a, b). At this stage, the empirical results suggest that techniques motivated by the assumption do give better results – and

therefore that the approach does have promise, but arguably a greater theoretical understanding could be achieved with further analytical work.

The ideas in this article could perhaps be extended to other surface-based methods. For example geographically weighted regression (GWR) (Brunsdon et al. 1996) with potential discontinuities might be a useful tool for exploring the possibility that the relationship between several variables undergoes a sudden change when some virtual or physical boundary is crossed. Similarly some other geographically weighted summary statistics (Brunsdon et al. 2002), such as localized standard deviations, could be adapted in a similar way.

Another area for further work is in the visualization of the fitted surfaces. Although surface visualization superimposed on maps is now commonplace, a characteristic that sets this approach apart is the appearance of 'cliff edges'. These are important in interpreting trends. Although perhaps these clefts in trend surfaces are a relatively abstract notion, they do draw attention to areas of notable geographical diversity, and highly contrasting circumstances, and perhaps more focussed study should be applied to the problem of visualizing them, and in particular visualizing uncertainty in their location. Such research would be valuable not only for the Bilateral Kernel regression approach proposed here, but also for other methods which detect discontinuities in trends.

Notes

1 See http://www.houseprices.co.uk/ for additional details.
2 Using the 'optimise' function in R.
3 Used the 'rgl' package in R for this figure.
4 It is generally suggested to use $N = 50$ or $N = 100$ in practice.

References

Bowman A, Pope A, and Ismail B 2006 Detecting discontinuities in nonparametric regression curves and surfaces. *Statistical Computing* 16: 377–90

Breiman L 1996 Bagging Predictors. *Machine Learning* 24: 123–40

Brunsdon C, Fotheringham A S, and Charlton M 1996 Geographically Weighted Regression: A method for exploring spatial nonstationarity. *Geographical Analysis* 28: 281–89

Brunsdon C, Fotheringham A S, and Charlton M 2002 Geographically weighted summary statistics: A framework for localised exploratory data analysis. *Computers, Environment and Urban Systems* 26: 501–24

Crampton A and Mason J 2005 Detecting and approximating fault lines from randomly scattered data. *Numerical Algorithms* 39: 115–30

Diggle P and Ribeiro P J 2007 *Model Based Geostatistics*. New York, Springer

Duda R O, Hart P E, and Stork D G 2000 *Pattern Classification* (Second Edition). New York, John Wiley and Sons

Efron B 1979 Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7: 1–26

Efron B 1982 *The Jacknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA, Society for Industrial and Applied Mathematics

Hastie T J, Tibshirani R J, and Friedman J 2001 *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, Springer

Okabe A, Boots B, and Sugihara K 2000 *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams* (Second Edition). New York, John Wiley and Sons

Openshaw S 1984 *CATMOG 38: The Modifiable Areal Unit Problem*. Norwich, Geo-Abstracts

Smith S L, Holland D, and Longley P A 2005 Quantifying interpolation errors in urban airborne laser scanning models. *Geographical Analysis* 37: 200–24

Staniswalis J G 1987a The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* 84: 276–83

Staniswalis J G 1987b *A Weighted Likelihood Formulation for Kernel Estimators of a Regression Function With Biomedical Applications*. Richmond, VA, Virginia Commonwealth University, Medical College of Virginia, Department of Biostatistics, Technical Report 5

Terzopoulos D 1988 The computation of visible-surface representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10: 417–38

Tomasi C and Randuchi R 1998 Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision*, Bombay, India: 839–46

Townsend P, Phillimore P, and Beattie A 1988 *Health and Deprivation: Inequality and the North*. London, Croom Helm

Vapnik V 1998 *Statistical Learning Theory*. New York, John Wiley and Sons

Wahba G 1990 *Spline Models for Observational Data*. Philadelphia, PA, SIAM

Wood S N 2003 Thin plate regression splines. *Journal of the Royal Statistical Society (Series B)* 65: 95–114

Wood S N 2006 *Generalized Additive Models: An Introduction wth R*. New York, Chapman and Hall