Geographical Analysis ISSN 0016-7363

Geographically Weighted Discriminant Analysis

Chris Brunsdon,[1] Stewart Fotheringham,[2] Martin Charlton[2]

[1]Department of Geography, University of Leicester, Leicester, U.K. [2]National Centre for Geocomputation, National University of Ireland, Maynooth, U.K.

In this article, we propose a novel analysis technique for geographical data, Geographically Weighted Discriminant Analysis. This approach adapts the method of Geographically Weighted Regression (GWR), allowing the modeling and prediction of categorical response variables. As with GWR, the relationship between predictor and response variables may alter over space, and calibration is achieved using a moving kernel window approach. The methodology is outlined and is illustrated with an example analysis of voting patterns in the 2005 UK general election. The example shows that similar social conditions can lead to different voting outcomes in different parts of England and Wales. Also discussed are techniques for visualizing the results of the analysis and methods for choosing the extent of the moving kernel window.

Introduction

In this article, an extension to discriminant analysis is proposed, which allows the discrimination rule to vary over space. The term *Geographically Weighted Discriminant Analysis* (GWDA) is proposed for this new method. The motivation here is similar to that for Geographically Weighted Regression (GWR)—in some situations, relationships between variables are not universal, but dependent on location. A major distinction between the two techniques is that, while GWR attempts to predict a measurement or ratio scale variable y given a set of predictors $\boldsymbol{x} = \{x_1, \ldots, x_m\}$, GWDA analysis attempts to predict a categorical y variable. This is not the first article to suggest an application of geographical weighting to categorical data following the suggestions of Fotheringham, Brunsdon, and Charlton (2002). Atkinson et al. (2003) explore the relationships between riverbank erosion and various environmental variables using geographically weighted logistic regression, and Páez (2006) examines geographical variations in land use/transportation relationships using a geographically weighted probit model. There are similarities between discriminant analysis and logistic regression in that both are used to predict group membership from a set of predictor variables. The assumptions underlying each

Correspondence: Chris Brunsdon, Department of Geography, University of Leicester, Leicester, UK
e-mail: chrisbrunsdon@mac.com

technique are rather different. Logistic regression may be the method of choice when the dependent variable has two groups due to its more relaxed assumptions (Maddala 1983).

In common with existing discussions of discriminant analysis, it is helpful to regard the data used here as having been drawn from a number of distinct populations, one for each unique category of y. The task of GWDA is then to assess which population a given, unlabeled $\{x\}$ is likely to have come from. The difference between GWDA and standard discriminant analysis is that for GWDA this decision is made taking the geographical location of $\{x\}$ into account. The GWDA technique proposed here exploits the fact that linear and quadratic discriminant analyses (LDA and QDA) rely only on the mean vector and covariance matrix of $\{x\}$ for each population (the former assumes that the covariance is the same for all m populations). This being the case, the route to localizing LDA and QDA is via geographically weighted means and covariances, such as those proposed in Brunsdon, Fotheringham, and Charlton (2002) and Fotheringham, Brunsdon, and Charlton (2002). In the next section, LDA and QDA are reviewed. Following this, extending these techniques to forms of GWDA is considered. Finally, an empirical example using voting data in England and Wales is presented.

A review of discriminant analysis

Discriminant analysis is a technique used to identify which population a certain observation vector $\boldsymbol{x}$ belongs to, given a list of possible populations $\{1, \ldots, m\}$ and a training set of observations $\{x_{ij}\}$ where ij indicates the ith observation from population j. The simplest case occurs when $m = 2$ so that a binary classification has to be made. In this case, Fisher (1936) has used a decision theoretic approach to show that an optimal decision rule is to assign to population 1 if

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{C(1|2)p_2}{C(2|1)p_1} \tag{1}$$

and to assign to population 2 otherwise. Here, $f_j(\boldsymbol{x})$ is the probability density function for population j, p_j is the prior probability that an observation comes from population j, and $C(i|j)$ is the cost associated with wrongly classifying an observation in population i to population j. For this article, it will be assumed that all costs of wrong classification are the same.[1] In this case, (1) reduces to assigning $\boldsymbol{x}$ to population 1 if

$$p_1 f_1(\boldsymbol{x}) > p_2 f_2(\boldsymbol{x}) \tag{2}$$

and assigning to population 2 otherwise. The situation may be generalized (Rao 1948; Bryan 1951) to $m > 2$, by noting that equation (2) simply compares two scores of the form $p_j f_j(x)$, and extending this rule to state that x is assigned to population $k \in \{1..m\}$ where

$$k = \arg\max_{j \in \{1, \ldots, m\}} \{p_j f_j(\boldsymbol{x})\} \tag{3}$$

The interpretation of equation (3) is intuitive in terms of Bayes Theorem; $p_j f_j(\boldsymbol{x})$ is proportional to the posterior probability of an observation coming from population j once the value of $\boldsymbol{x}$ is known. Thus, assignment is to the population with the largest posterior probability. This is a general rule for assignment to populations, but it assumes that the prior probabilities $\{p_j\}$ and the probability density functions $\{f_j\}$ are known. In general, they are not, and they must be estimated from a set of training data. Much of the technique of discriminant analysis lies in this estimation task. The estimation of the p_j's is straightforward, if the training data are a random sample from the complete population obtained by merging each of the populations associated with individual y values. In this case, the estimate for p_j is just

$$p_j = \frac{n_j}{n_1 + n_2 + \ldots + n_m} \tag{4}$$

where n_j is the number of observations from population j.

If the training data set from each population is not generated in this way—for example, if they come from a case–control study where there are the same number of observations in each sample regardless of their relative abundance in reality—and there is no other information about the prior probabilities for each population, then a minimax argument shows that the optimal decision rule is found by setting each p_j to m^{-1}. In Bayesian terminology this is a noninformative prior.

Estimating f_j is less trivial. Two general possibilities are

1. Assume each f_j takes a known parametric form (such as multivariate Gaussian) and estimate the parameters from the data.
2. Use a nonparametric technique such as kernel density estimation to estimate f_j for each sample.

In this article, we concentrate on the first option and assume that the f_j Õs are multivariate Gaussian. This assumption gives rise to the standard techniques of LDA and QDA. The second option leads to a technique called *Kernel Discriminant Analysis* (KDA), which we will consider in the future.

In the multivariate Gaussian model, the decision rule (3) is

$$k = \arg\max_{j\in\{1,\ldots,m\}} \left[p_j \frac{1}{(2\pi|\mathbf{P}_j|)^{q/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_j)' \mathbf{X}_j^{-1} (\boldsymbol{x}-\mu_j)\right) \right] \tag{5}$$

where $\mathbf{P}_j$ is the variance–covariance matrix for population j, q is the number of predictor variables in $\boldsymbol{x}$, and μ_j is the mean value for population j. Taking logs, changing signs, and removing terms that are constant for all values of j, the rule may be written as

$$k = \arg\min_{j\in\{1,\ldots,m\}} \left[\frac{q}{2}\log\left|\mathbf{X}_j\right| + \frac{1}{2}(\boldsymbol{x}-\mu_j)' \mathbf{X}_j^{-1} (\boldsymbol{x}-\mu_j) - \log(p_j) \right] \tag{6}$$

This is the QDA decision rule. Each of the m score functions within the square brackets in equation (6) is quadratic in $\boldsymbol{x}$. If we make a further assumption that all populations have the same variance–covariance matrix, then the rule simplifies to

$$k = \arg\min_{j \in \{1, \ldots, m\}} \left[\frac{q}{2} \log \left| \Sigma \right| + \frac{1}{2} (\boldsymbol{x} - \mu_j)' \Sigma^{-1} (\boldsymbol{x} - \mu_j) - \log(p_j) \right] \quad (7)$$

Expanding the quadratic expression, and noting that the term quadratic in $\boldsymbol{x}$ appears identically in each score, the above rule simplifies to

$$k = \arg\min_{j \in \{1, \ldots, m\}} \left[-\boldsymbol{x}' \Sigma^{-1} \mu_j + \frac{1}{2} \mu_j' \Sigma^{-1} \mu_j - \log(p_j) \right] \quad (8)$$

which is linear in $\boldsymbol{x}$. This is the LDA decision rule. Note that maximum likelihood estimates may be used to estimate each μ_j and Σ_j, and that a pooled estimate for Σ in the case of LDA may be found by

$$\Sigma = \frac{n_1 \Sigma_1 + n_2 \Sigma_2 + \ldots + n_m \Sigma_m}{n_1 + n_2 + \ldots + n_m} \quad (9)$$

where $\Sigma_1, \ldots, \Sigma_m$ are maximum likelihood estimators of the respective variance–covariance matrices in each population.

The two rules are compared in Fig. 1. In this case, $\boldsymbol{x}$ is two-dimensional, $m = 3$, and the populations are clearly separated. Here, it is clear that LDA is just as good as QDA for separating the populations, although this is not often the case. In many real-world examples, there is a degree of overlap between the populations, so that neither LDA nor QDA can provide a perfect discrimination rule.

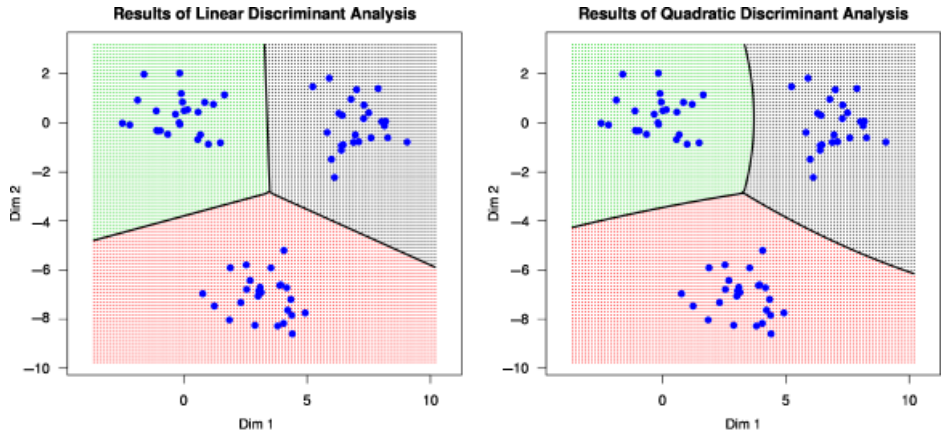

Figure 1. Discriminant analysis illustrated. LHS shows the result of LDA applied to the points; RHS shows the same for QDA. LDA, linear discriminant analysis; QSD, linear and quadratic discriminant analysis.

GWDA—a definition by extending discriminant analysis

Having outlined the ideas underlying QDA and LDA, this section discusses how these may be extended into geographically weighted methods. Essentially, the idea is that one now assumes that one or more of the quantities Σ_j, Σ, μ_j, and p_j are not fixed, but depend on spatial location $\boldsymbol{u}$. In effect, the decision rule now becomes localized, in the sense that any of the probabilities used to derive the decision rules are now conditional on $\boldsymbol{u}$. The dependency is modeled by assuming that the vector μ_j is a function of $\boldsymbol{u}$—so that each element of μ_j is a geographically weighted mean. Similarly, each element of Σ_j is a geographically weighted variance or covariance. These methods of estimation are in keeping with a local likelihood approach to estimating the population models of the form

$$f_p(\boldsymbol{x}|\boldsymbol{u}) = \frac{1}{(2\pi|\Sigma_j(\boldsymbol{u})|)^{q/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_j(\boldsymbol{u}))' \Sigma_j^{-1}(\boldsymbol{u})(\boldsymbol{x}-\mu_j(\boldsymbol{u}))\right) \tag{10}$$

Páez, Uchida, and Miyamoto (2002a, b) and Páez (2004) have made use of geographically weighted variances to provide an alternative interpretation of GWR.

Estimation

As suggested in the introduction, calibration of local discriminant functions is achieved through calibration of local coefficients. Under the Gaussian assumption, this amounts to the task of calibrating $\{\mu_1(\boldsymbol{u}), \ldots, \mu_m(\boldsymbol{u})\}$ and $\{\Sigma_1(\boldsymbol{u}), \ldots, \Sigma_m(\boldsymbol{u})\}$ in equation (10). This is essentially a nonparametric estimation task, because the objects to be estimated are functions of the vector $\boldsymbol{u}$. More specifically, $\{\mu_1(\boldsymbol{u}), \ldots, \mu_m(\boldsymbol{u})\}$ are *vector* functions of $\boldsymbol{u}$ while $\{\Sigma_1(\boldsymbol{u}), \ldots, \Sigma_m(\boldsymbol{u})\}$ are *matrix* functions of $\boldsymbol{u}$. As a shorthand, this collection of functions associated with the point $\boldsymbol{u}$ will henceforth be referred to as $C_j(\boldsymbol{u})$. One approach to estimating $C_j(\boldsymbol{u})$ at any point $\boldsymbol{u}$ is through *local likelihood* estimation. This is equivalent to calibrating a weighted *nonlocalized* model where the weights are obtained from a kernel function centered on $\boldsymbol{u}$. Weights are applied to log-likelihoods, and so for equation (10) $\hat{C}_j(\boldsymbol{u})$ is[2] chosen to minimize

$$L(C_j(\boldsymbol{u})|\boldsymbol{x}) = \sum_{i=1}^{n} w_i(\boldsymbol{u}) \left(-\log\left(|\Sigma_j(\boldsymbol{u})|\right) - \frac{1}{2}(\boldsymbol{x}_i-\mu_j)' \Sigma_j^{-1}(\boldsymbol{u})(\boldsymbol{x}_i-\mu_j)\right) \tag{11}$$

where $w_i(\boldsymbol{u})$ is the weight applied to observation i when calibration is taking place at the point $\boldsymbol{u}$. As stated earlier, this weight is determined by a kernel function, so that if $\boldsymbol{u}_i$ is the location associated with observation i, and $d_i(\boldsymbol{u}) = |\boldsymbol{u}-\boldsymbol{u}_i|$, then $w_i(\boldsymbol{u})$ is a decreasing function of $d_i(\boldsymbol{u})$, tending to 0 as $d_i(\boldsymbol{u})$ increases. One possibility is the bisquare kernel function

$$F_h(d) = \begin{cases} \left(1-\left(\frac{d}{h}\right)^2\right)^2 & \text{if } d < h \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where d is the distance, and h is the kernel *bandwidth*—a quantity controlling the smoothness of the nonparametric function estimation.

This approach leads to the following estimators for $C_j(\boldsymbol{u})$:

$$\hat{\mu}_j(\boldsymbol{u}) = \frac{\sum_{i=1}^{n} w_i(\boldsymbol{u})\boldsymbol{x}_i}{\sum_{i=1}^{n} w_i(\boldsymbol{u})} \tag{13}$$

and

$$\hat{\Sigma}_j(\boldsymbol{u}) = \frac{\sum_{i=1}^{n} w_i(\boldsymbol{u})(\boldsymbol{x}_i - \mu_j(\boldsymbol{u}))(\boldsymbol{x}_i - \mu_j(\boldsymbol{u}))'}{\sum_{i=1}^{n} w_i(\boldsymbol{u})} \tag{14}$$

The next step is to estimate $\{p_1, \ldots, p_m\}$. As before, this is straightforward, if the training data in the neighborhood of $\boldsymbol{u}$ are a random sample from the complete population obtained by merging each of the populations associated with individual y values. In this case, the p_j's are estimated as in equation (4). If we further assume that Σ is the same for each population, as in LDA, we may obtain an estimate for $\hat{\Sigma}$ using equation (9).

Note that in the above we are assuming that the proportions from different populations in the sample reflect the overall proportions. However, if this is not the case, then the assumption that each p_j is equal to $1/m$ is made. A final complication is that the proportions in the merged population associated with each of the y values may vary geographically in its composition. In this case, local likelihood estimates of p_j (written as $p_j(\boldsymbol{u})$) should be used. These are obtained by

$$p_j(\boldsymbol{u}) = \frac{\sum_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u})}{\sum_{j} \sum_{\boldsymbol{x}_i \in j} w_i(\boldsymbol{u})} \tag{15}$$

This choice between local and global estimation occurs in other aspects of the calibration, as discussed in the next subsection.

Choices

One key issue in the above approach is the decision as to what parts of the model should vary geographically. For any QDA analysis, we must estimate $\{\Sigma_1, \ldots, \Sigma_m\}$, $\{p_1, \ldots, p_m\}$ and $\{\mu_1, \ldots, \mu_m\}$. Similarly, for LDA we must estimate $\{\mu_1, \ldots, \mu_m\}$, $\{p_1, \ldots, p_m\}$ and Σ. We have a choice as to which of these we allow to vary geographically. In this study, we allow the mean, the variance–covariance matrix, and the prior probabilities to vary geographically although it would be possible to fix any combination of these to constant global values. However, the circumstances leading to such an action are not immediately obvious and the more logical approach would be the one followed here of allowing all three components to vary spatially.

These choices imply that there are a number of ways in which GWDA could be specified. It seems likely that different choices are likely to be better in different situations and that some methodology for selecting the best approach for a given data set should be proposed. In addition to these choices relating to the probability

model, there is also the matter of choice of bandwidth for the geographical weighting. Again, a selection methodology is required. A tentative approach proposed here is cross-validation. A comprehensive cross-validation would involve removing each observation from the training data, applying the GWDA to the rest of the data set, and then assigning the removed observation to a population using the discriminant rule thus derived. Applying this to every observation in the data set and noting the proportion of correct assignments gives a performance score for a given model working with a given bandwidth. This cross-validation score may then be used to identify the ''best'' combination of model and bandwidth for a given data set. An alternative, faster approach would be to randomly select a large proportion of the training data (say 90%), calibrate the GWDA using this, and then apply the cross-validation procedure to the remaining 10%.

A more sophisticated approach to cross-validation scoring may be to use *cross-validation likelihood*, rather than the proportion of correct assignments. Note that discriminant analysis assigns observations to populations on the basis of *posterior probabilities*. Applying Bayes Theorem, the probability that observation $\boldsymbol{x}$ is in population j is given by

$$\Pr(j|\boldsymbol{x},\boldsymbol{u}) = \frac{\Pr(\boldsymbol{x}|j,\boldsymbol{u})}{\Pr(\boldsymbol{x}|1,\boldsymbol{u}) + \ \ldots \ + \Pr(\boldsymbol{x}|m,\boldsymbol{u})} \tag{16}$$

For each observation in the held-back cross-validation data set, we know the actual value of j, and using a given model and bandwidth we can estimate the posterior probability that the observation is in its true population. Taking logs and adding up these quantities, a likelihood of the validation data set for a given choice of bandwidth and model is obtained:

$$\text{Score} = \sum_{\text{validation set}} \log(\Pr(j|\boldsymbol{x},\boldsymbol{u})) \tag{17}$$

This quantity can be used to select the best combination of model and bandwidth. This approach has one clear advantage over a straightforward score of classifications because it is a continuous function with respect to bandwidth—which is often advantageous in optimization algorithms, such as gradient descent (Fletcher and Reeves 1964) or the simplex method (Nelder and Mead 1965). The down side is that this score is computationally more expensive.

Spatially adaptive kernel bandwidths

As with GWR, one issue relating to the choice of h, the kernel bandwidth, is that one may expect the degree of spatial variability in the discriminant functions to vary geographically. For example, if the subjects of the analysis are people, then it is possible that in urban areas with denser populations, there may be more variety in social, economic, and cultural characteristics of places than there would be in equivalent-sized rural areas. If that were the case, one could reasonably expect geographical drift in GWDA models to be notable in much smaller geographical

windows. In such situations, a one-size-fits-all approach to bandwidth choice is unhelpful, because it assumes that the spatial scale of variation is the same everywhere. To overcome this problem, it is proposed to use a local bandwidth selection procedure, based on the lth nearest neighbor in the data set to the point $\boldsymbol{u}$ at the center of the kernel. To calibrate the model at $\boldsymbol{u}$, we simply choose the l nearest-neighbor distance from $\boldsymbol{u}$ as the bandwidth. In this way, in denser areas, where there are more sample observations, the bandwidth will be smaller, and the model calibration will take place on a smaller geographical scale. This is essentially the same approach as used in GWR (Fotheringham, Brunsdon, and Charlton 2002) and has been found to work well in practice.

At this stage, we are still left with the choice of an appropriate value for l. The approach to this is much the same as that described for bandwidth selection in ''Choices.'' Essentially, one of the cross-validation scoring methods outlined in that section is carried out over a series of possible l values, and the value with the best score is chosen. Unfortunately, because l is a discrete quantity, one cannot take advantage of some of the automatic optimization procedures suggested in the previous section, and so computation can take slightly longer. Once an optimal score has been found for the nearest-neighbor bandwidth method, this can be compared with that for the fixed bandwidth method, and the best performing of the two can be presented as the final GWDA model.

Visualization and interpretation

An alternative to mapping coefficients

A major difficulty when presenting the results of GWDA is that of visualization. For example, when expanding the expression in equation (7), then for each of the populations $1, \ldots, m$ there are $q+1$ coefficients. For a local QDA, there would be even more than this. Thus, in effect, for linear GWDA there is an m by $q+1$ matrix associated with each point in space. One possibility would be to plot $m(q+1)$ maps side by side (which is in keeping with Tufte's 1990 principle of ''small multiples'')—which is essentially mapping each of the coefficients in the way a typical GWR analysis would do. The main problem here is one of interpretability. For each variable, there is a separate coefficient for each population. For a given observation, one needs to compare these m variables to understand the relative likelihood that this observation will be assigned to a given group. This can be a large amount of information to take in, and there is often no obvious visual connection between the shading of a single map and likely group membership. It is for this reason that approaches to visualization other than the mapping of coefficients are considered here.

An alternative approach—and one that it is felt embodies the spirit of local modeling—is to consider the variation between a given fixed set of predictor variables $\{\boldsymbol{x}\}$ and the probability of membership of each of the populations $\Pr(j|\boldsymbol{x},\boldsymbol{u})$, as the location $\boldsymbol{u}$ varies. In a global model, $\Pr(j|\boldsymbol{x},\boldsymbol{u})$ would not depend on $\boldsymbol{u}$ and so

for a given $\boldsymbol{x}$ we would expect any map based on these probabilities to show no variations. The fact that local models do exhibit variation is their defining characteristic, and mapping this variation should provide an intuitive image of the nature of this variation for a given model. For example, if we were to choose $\boldsymbol{x}$ to be a global average or median of the sample data set, our map would show the geographical variation in the probable category to which a typical observation is likely to belong. Similarly, by choosing one of the variables of $\boldsymbol{x}$ to be above average (say the 80th percentile), we could assess changes in the assignment to category due to this increase in value across geographical space. In the example, we will consider how the level of deprivation affects voting behavior and show that higher levels of deprivation affect voting choice in different ways in different parts of England and Wales.

Mapping probabilities of membership

It has been argued that a helpful visualization strategy is to map variation in $\Pr(j|\boldsymbol{x},\boldsymbol{u})$ for a set of given $\boldsymbol{x}$ values. However, this in itself presents a minor challenge, as this is essentially a vector quantity because there are j potential values of k. Using the notational shorthand $p_j(\boldsymbol{u}) = \Pr(j|\boldsymbol{x},\boldsymbol{u})$ we need a way of mapping the vector $\boldsymbol{p}(\boldsymbol{u}) = (p_1(\boldsymbol{u}), p_2(\boldsymbol{u}), \ldots, p_k(\boldsymbol{u}))$. Here, two approaches will be suggested. First, assume that the map is divided into l zones $Z_1, \ldots, Z_l$ each associated with a representative point $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l\}$. These zones could be regular (such as rectangular pixels) or irregular (such as electoral wards). The representative point may be the centroid of the zone, but need not be, particularly in cases where the centroid of a zone is not within that zone. The approaches are:

1. ''Majority Vote Approach'': For each Z_i, find the largest element $p_j(\boldsymbol{u}_i)$ in $p(\boldsymbol{u}_i)$. Color Z_i with a color associated with j.
2. ''Mixture Approach'': Specify a color associated with each value of j using a color coordinate system (such as RGB or HSV). Use the elements of $p(\boldsymbol{u}_i)$ to obtain a weighted average of these colors in the color space. (This is possible because the elements of $\boldsymbol{p}(\boldsymbol{u}_i)$ sum to 1.) Color Z_i with this weighted average color.

The majority vote approach is less informative—for example, it is not possible to distinguish between probability vectors such as (0.34, 0.33, and 0.33) and (0.98, 0.01, and 0.01)—and thus one cannot tell how strong the degree of potential membership is to the most likely population. On the other hand, the mixture approach allows one to distinguish between more subtle changes in the profile, but requires that maps are produced in full color. Thus, it is not appropriate to display this kind of map in many refereed journals that can reproduce only monochrome images.

For the mixture approach, choice of color space is important. Colors for each of the populations should be clearly distinguishable—this should avoid confusion between high values of p_j for each possible value of j. Given that colors are chosen

by applying linear operations to the color coordinate space, it would also be helpful if the differences in *perception* of different colors correspond to distances between coordinates in the color space. This implies that, for example, if $k = 3$, the color used to represent (0.5, 0.5, and 0) should be equally distinguishable from the two colors used to represent the classes $j = 1$ and 2.

The issue of defining a color space whose distances corresponded to human color perception was addressed by the *Commission Internationale de l'Éclairage* (the CIE) in 1976. Two such spaces are the CIELAB and CIELUV specifications. Both are based on perceptual color spaces, with CIELUV generally preferred for work with additive color technologies, such as LCD displays and projection equipment, while CIELAB is more appropriate for subtractive color systems, such as the use of inks or dyes. Here, CIELUV will be considered, as a typical use of the mixture approach will be the creation of maps to be shown on LCD overhead projection equipment during conferences, lectures, or seminars.

The CIELUV space has three coordinates, L, u, and v, where $L \in [0, 100]$ is a degree of luminance, and $(u, v) \in [-100, 100] \times [-100, 100]$ correspond to balance between red/green and yellow/blue, respectively. Following the advice of Cleveland and McGill (1983), we note that areas on a graph with greater luminance tend to look larger, and to avoid any optical illusions that may draw unmerited attention to certain map regions purely on the grounds of color choice, we work with a fixed value of L. Thus, our color model resides entirely on the (u, v) plane. Next, it is important to note that not all (L, u, v) triplets correspond to a color—because the coordinate system is a nonlinear transform of the conventional RGB color cube it does not correspond to a regular cuboid itself, and so some points near the edges of the color space are undefined.

Thus, for the mixture approach to work, we must choose a set of colors for the k populations corresponding to k points in the (u, v) plane for a given L, and we must ensure that the convex hull of these points consists entirely of well-defined colors. A further condition is that the points should be as far apart as possible, to meet the requirement that colors for each of the populations should be clearly distinguishable. Finding such a set of points for any given L is an exercise in practical computational geometry.[3] For $k = 3$, a possible solution is illustrated in Fig. 2.[4] This uses a luminance value of 70—which has been found to work well in practice. For the three mixing colors, this scheme uses (70.0, −37.0, 63.6), (70.0, −37.0, −71.4), and (70.0, 80.0, −4.0).

An example using the 2005 UK election results for England and Wales

The results of the May 2005 UK General Election may be downloaded from the UK electoral commission.[5] As of May 6, 2005, the results for the 569 Parliamentary Constituencies contested in England and Wales are summarized in Table 1. The results in map form are shown in Fig. 3.

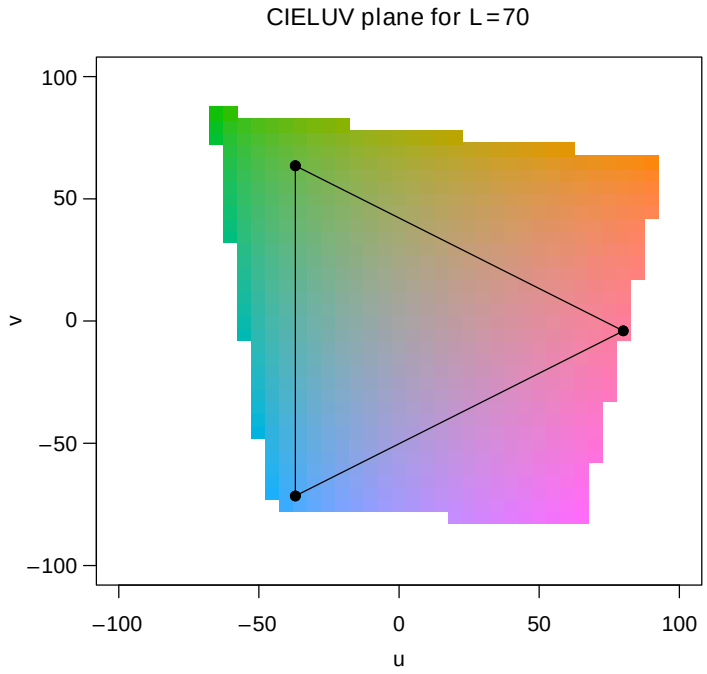

Figure 2. Using the CIELUV color space to visualize a three-way probability vector.

To obtain a greater understanding of the linkage of voting patterns with social and economic conditions in each of the constituencies, a number of socioeconomic variables were derived for each constituency. These are tabulated in Table 2.

These variables may be used in a straightforward (nongeographically weighted) discriminant analysis to predict the party elected in each constituency. This gives rise to the map in Fig. 4. One interesting feature of this map is that it predicts *no* seats for the Liberal Democrats or others. Clearly, this is at odds with the reality of voting patterns. Looking at Fig. 3, it seems that the Liberal Democrat/Other vote is stronger in certain regions than in others. In particular, North Wales and the West of England show a strong tendency to vote away from the Conservative/Labour axis.

However, because there are six predictor variables, an extra step of dimensional reduction using principal component analysis is carried out. This overcomes the difficulties associated with correlation among the predictor variables. In order to allow for differences in scale between the predictor variable, the principal com-

Table 1 May 2005 Election Results (England and Wales)

Party	No. of seats
Conservative	196
Labour	314
Liberal Democrat and Others	59

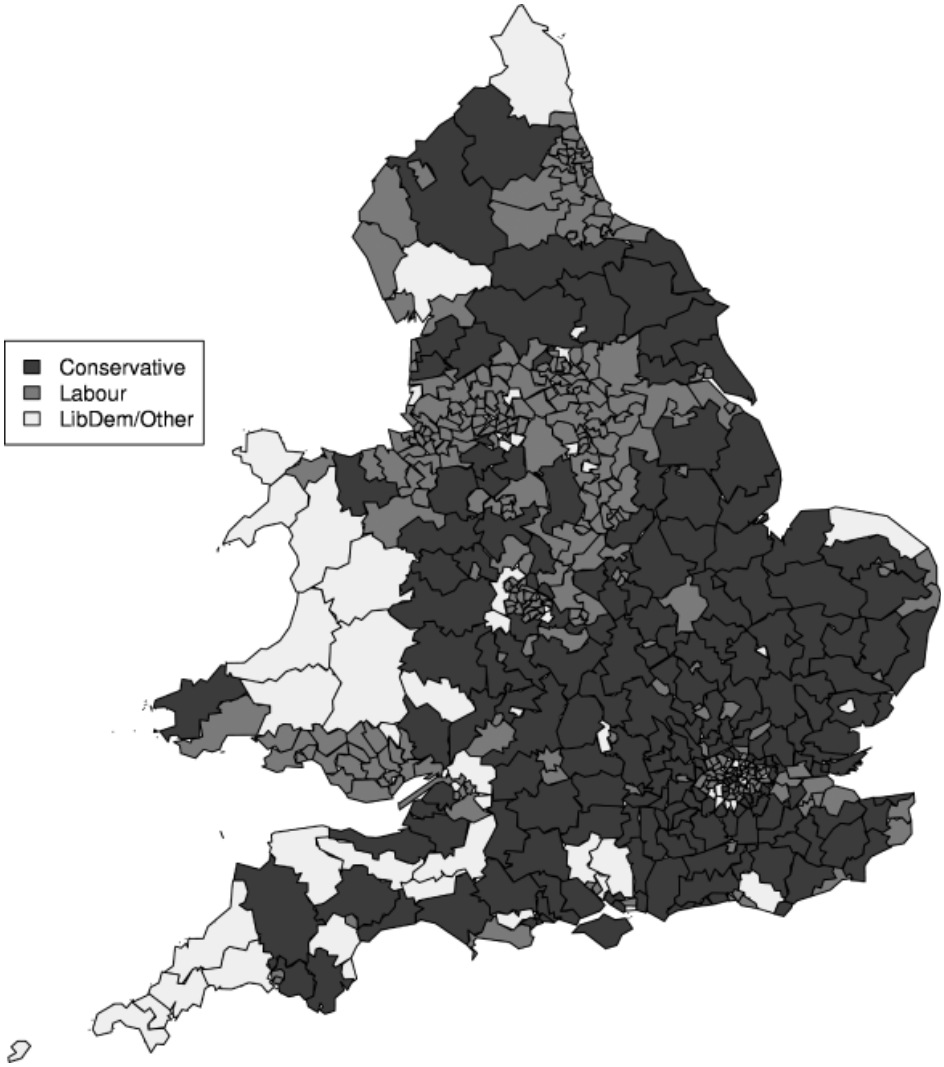

Figure 3. Election results for parliamentary constituencies in England and Wales.

Table 2 Constituency-Level Socioeconomic Variables Used in this Example

Variable name	Description
Unemp	Percentage of economically active males unemployed
NoQual	Percentage of adult population with no qualifications
OwnOcc	Percentage of owner occupied households
Pension	Percentage of pensioners in population
NonWhite	Percentage of nonwhite people in population
LoneParHH	Percentage of lone-parent households

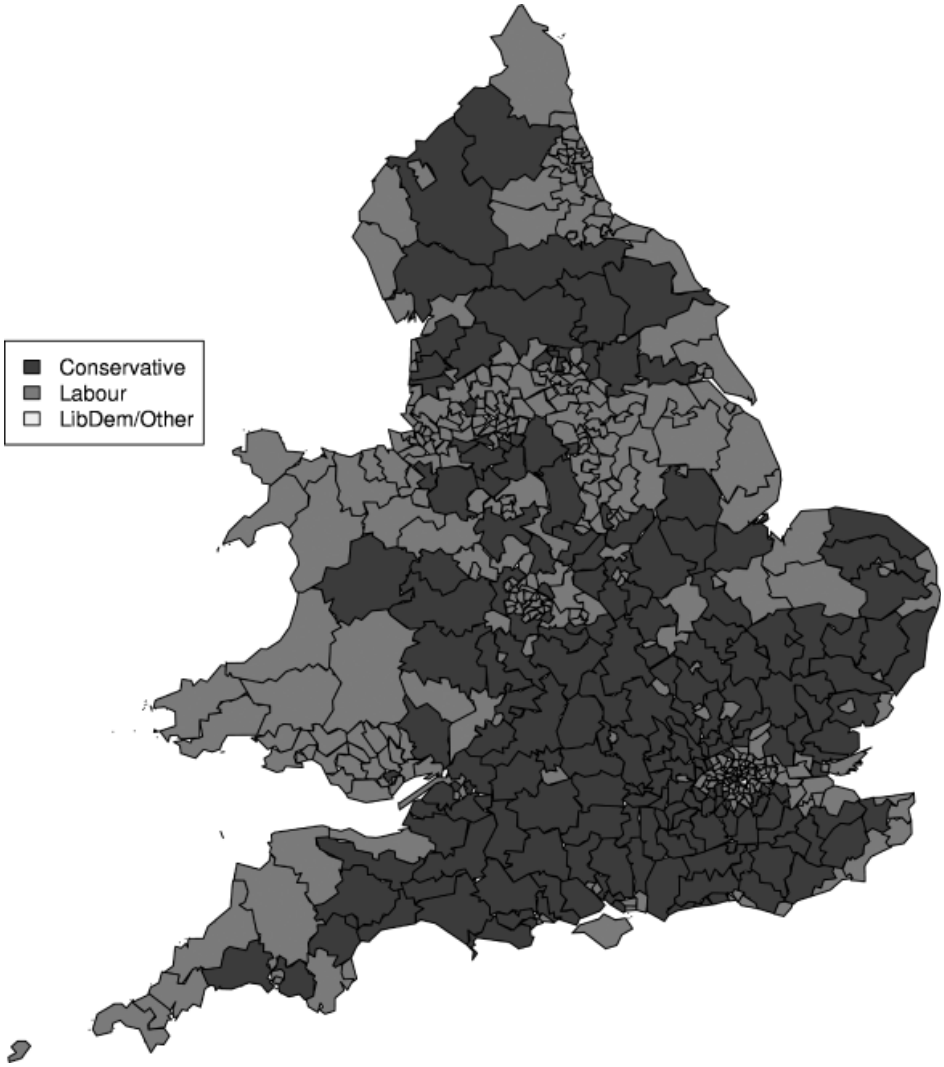

Figure 4. Predicted election results for England and Wales using global discriminant analysis.

ponent analysis is based on the correlation matrix, rather than the covariance matrix, of the variables in Table 2. The first two principal components are listed in Table 3. These could be interpreted as follows:

Table 3 Principal Component Analysis of Predictor Variables

Component	Unemp	NoQual	OwnOcc	Pension	NonWhite	LoneParHH
1	−0.135	0.027	0.090	0.111	−0.975	−0.096
2	0.716	0.471	−0.204	−0.021	−0.151	0.448

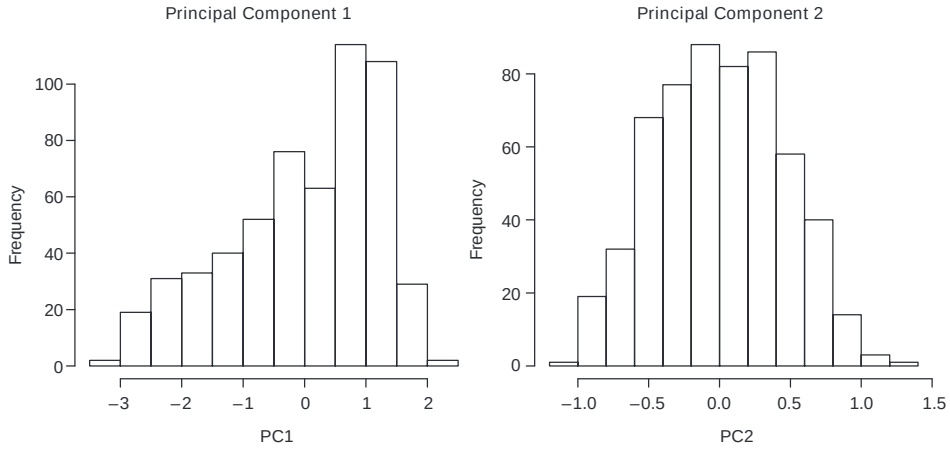

Figure 5. Histograms of principal components.

1. *Mainstream Britain*: high levels of home owner-occupiers, pensioners, low unemployment, lone-parent households, and nonwhites.
2. *Deprivation*: high unemployment, high levels of people without qualifications, and lone-parent households.

The histograms of the principal components are shown in Fig. 5. Although not perfectly normal in distribution (an assumption of LDA), they are reasonably close. Furthermore, for the geographically weighted approach, normality on a global scale is less important than the normality of components in the vicinity of any given local prediction point.

Next, both a fixed bandwidth and an adaptive GWDA were applied to this data. In each case, the optimal smoothing parameters were selected using cross-validation likelihood, as illustrated in Fig. 6 (fixed) and Fig. 7 (adaptive). From these it may be seen that the optimal fixed bandwidth is around 100 km, and the optimal adaptive bandwidth is at six nearest neighbors. It may also be seen that the optimal adaptive model has a better cross-validation likelihood than the fixed.

The results of predicting the party of the candidate returned by each constituency is shown for the fixed bandwidth GWDA in Fig. 8 and for the adaptive bandwidth GWDA in Fig. 9. In both cases, the models have to some extent captured the greater tendency to vote Liberal Democrat in the southwest of England and Plaid Cymru in Wales. This is also shown in the confusion matrices for the three methods (Tables 4–6), which confirm the superiority of GWDA over global LDA.

Next, the relationship between the predictor variables and the categorical outcomes will be considered. Mapping of the kind suggested in ''An alternative to mapping coefficients'' is carried out. Fixing the variables so that the first and second principal components take the values $\{-1, 0, 1\}$ in sequence gives the matrix of

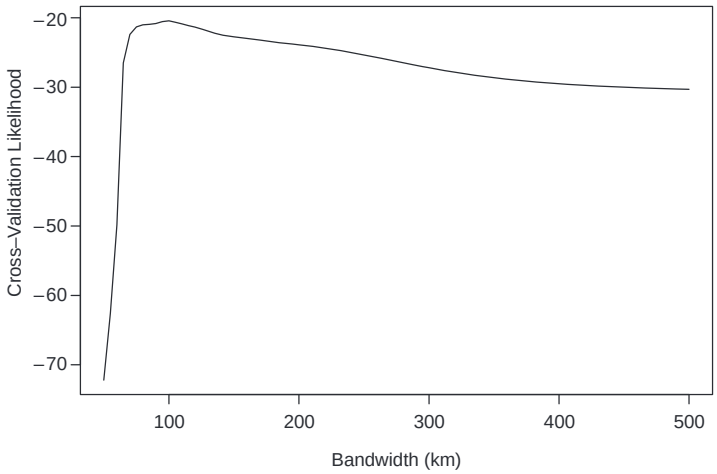

Figure 6. Fixed bandwidth versus cross-validation likelihood.

maps in Fig. 10.[6] These demonstrate the way in which the predicted parties elected alter as two key elements of the structure of the predictor variables alter. First, noting that the principal component scores are standardized, a value of 1 or -1 suggests one standard deviation above or below the mean value, while a score of 0 corresponds to the mean value itself for that particular component. From this, a number of features can be deduced. First, in all cases, increasing the ''middle Britain'' coefficient tends to increase the number of conservative MPs returned. When the second deprivation component is at the average level, increasing the middle-Britain component causes a conservative core in the south and east of England to grow until it covers most of the lower half of England. However, a localized effect is apparent, where in the southwestern part of England (Devon and Cornwall) the vote

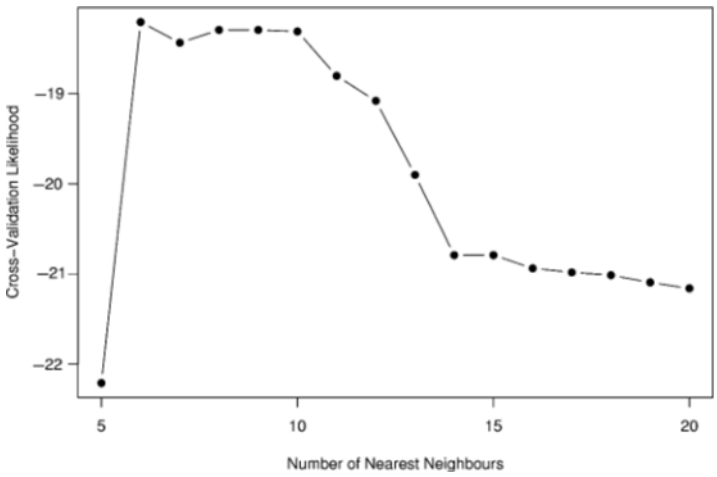

Figure 7. *l*th nearest-neighbor bandwidth versus cross-validation likelihood.

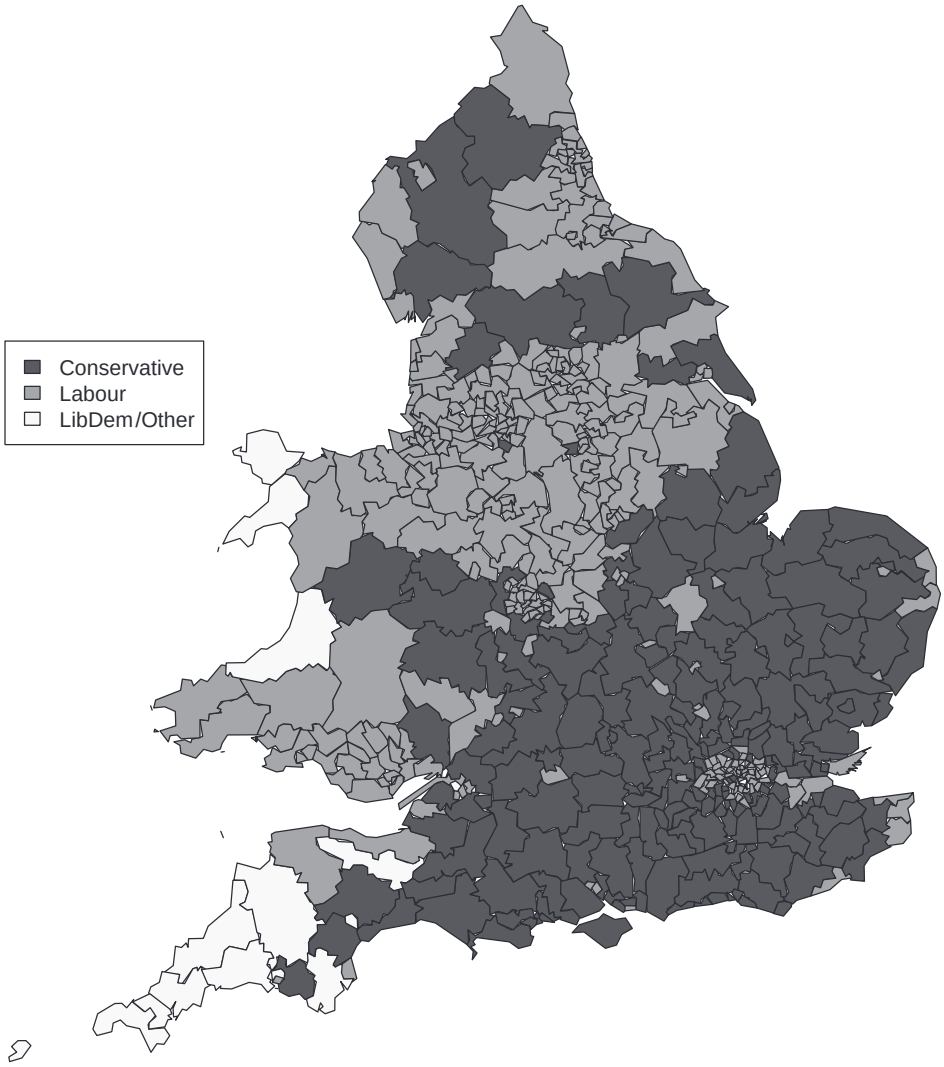

Figure 8. Predicted election results for England and Wales using fixed bandwidth GWDA. GWDA, geographically weighted discriminant analysis.

switches to liberal democrat rather than conservative. Another interesting effect is observed when fixing the middle-Britain component at the mean level and observing the effect of varying the deprivation component from -1 to 1. Again, this has the effect of varying the amount of cover gained by the conservatives, perhaps to a larger degree than varying the first component. However, the different voting pattern in the southwest of England, and also in Wales, is also apparent here.

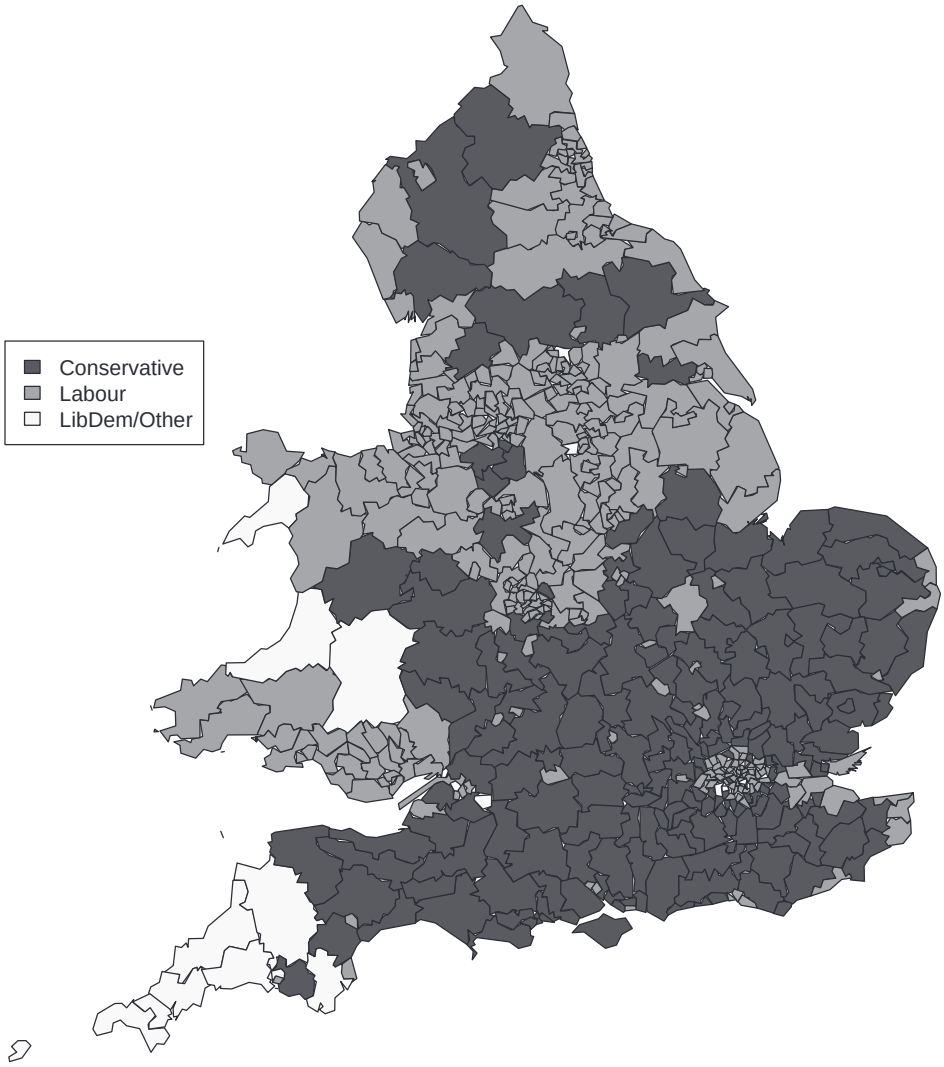

Figure 9. Predicted election results for England and Wales using adaptive bandwidth GWDA. GWDA, geographically weighted discriminant analysis.

Concluding discussion

In this article, the idea of local modeling has been extended to discriminant analysis, enabling the prediction of categorical data to be treated in a geographical weighted framework similar to GWR. Both adaptive and fixed kernel approaches have been considered, and a method for selecting the degree of smoothing to be applied has been outlined for these. There are still a number of methodological and theoretical issues to be considered, and these will now be outlined. First, the technique outlined here relies on the assumption that the predictor ($\boldsymbol{x}$) variables are

Table 4 Predictive Accuracy of the Adaptive Kernel GWDA Method

	Prediction		
	Conservative	Labour	Other
Result			
Conservative	160	33	3
Labour	16	297	1
Other	24	22	13

GWDA, Geographically Weighted Discriminant Analysis.

Table 5 Predictive Accuracy of the Fixed Kernel GWDA Method

	Prediction		
	Conservative	Labour	Other
Result			
Conservative	161	33	2
Labour	21	291	2
Other	25	24	10

GWDA, Geographically Weighted Discriminant Analysis.

Table 6 Predictive Accuracy of the Global LDA Method

	Prediction		
	Conservative	Labour	Other
Result			
Conservative	150	46	0
Labour	24	290	0
Other	30	29	0

LDA, linear discriminant analysis.

continuous and follow a Gaussian distribution. In many practical situations, this will not be the case. One potential approach here would be to consider KDA as discussed earlier—although there will be some methodological issues to be considered (see, e.g., Hastie, Tibshirani, and Friedman 2001).

A further issue is dealing with categorical *predictors*. Again, the current Gaussian approach is not well equipped to deal with these. One approach might be to consider all interactions between a categorical predictor and the distributions of continuous predictors—so that, for each level of the categorical variable, there is a different multivariate Gaussian distribution for the continuous predictors. This may work well for single categorical predictors, but there is perhaps a risk of multiplying potential parameterizations of distributions if there are a large number of discrete

Figure 10. Predicted election results for England and Wales using fixed bandwidth GWDA. GWDA, geographically weighted discriminant analysis.

predictors, and every possible combination of these corresponds to a different multivariate Gaussian distribution. A simpler approach might be to consider treating the categorical predictors as having independent effects with respect to the continuous predictors, so that

$$\Pr(j|\boldsymbol{x}, \boldsymbol{u}, z = m) \propto \Pr(j|z = m, \boldsymbol{u}) \Pr(j|\boldsymbol{x}, \boldsymbol{u}) \tag{18}$$

where m is one of the possible categorical values of the predictor variable z.

A further important issue is that of *robustness*. Currently, the mean and variance estimates at a given point $\boldsymbol{u}$ are based on weighted means, but it is worth noting that such statistics are vulnerable to contamination by outliers. An outlier in a particular locality could be particularly harmful, as it could have a great deal of leverage on local parameter estimates. Although the distortion would be restricted to the region of this rogue observation, the degree of distortion could be quite large if there were not many other normal observations nearby. This could clearly lead to misleading map patterns. For this reason, in the future we intend to investigate robust estimators of local mean and variance, such as those outlined in Rousseeuw and van Dreissen (1999) and Hubert and van Dreissen (2002), and adapt these for localized parameter estimation.

Notes

1 it is a conceptually simple but tedious task to modify results if this assumption does not hold.
2 The hat is intentional, denoting an *estimate* of $C_j(\boldsymbol{u})$.
3 That is trial and error plus a little guessing.
4 A color version of this diagram is available from homepage.mac.com/chrisbrunsdon/GWR_in_R/FileSharing5.html
5 http://www.electoralcommission.org.uk/elections/generalelection2005.cfm as of 12-12-2005.
6 A color version of these results using the CIELUV scheme is available from homepage.mac.com/chrisbrunsdon/GWR_in_R/FileSharing5.html

References

Atkinson, P. M., S. German, D. Sear, and M. Clark. (2003). ''Exploring the Relations Between Riverbank Erosion and Geomorphological Controls Using Geographically Weighted Logistic Regression.'' *Geographical Analysis* 35, 58–82.

Brunsdon, C., A. Fotheringham, and M. Charlton. (2002). ''Geographically Weighted Summary Statistics: A Framework for Localized Exploratory Data Analysis.'' *Computers Environment and Urban Systems* 26, 501–24.

Bryan, J. (1951). ''The Generalized Discriminant Function: Mathematical Foundations and Computational Routine.'' *Harvard Educational Review* 21, 90–95.

Cleveland, W., and R. McGill. (1983). ''A Color-Caused Optical Illusion on a Statistical Graph.'' *The American Statistician* 37, 101–05.

Fisher, R. (1936). ''The Use of Multiple Measurements in Taxonomical Problems.'' *Annals of Eugenics* 7, 179–88.

Fletcher, R., and C. Reeves. (1964). ''Function Minimisation by Conjugate Gradients.'' *Computer Journal* 7, 148–54.

Fotheringham, A., C. Brunsdon, and M. Charlton. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, UK: Wiley.

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

Hubert, M., and K. van Dreissen. (2002) ''Fast and Robust Discriminant Analysis.'' Technical Report, Department of Mathematics,Katholeike Universiteit Leuven.

Maddala, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

Nelder, J., and R. Mead. (1965). ''A Simplex Algorithm for Function Minimisation.'' *Computer Journal* 7, 308–13.

Páez, A., T. Uchida, and K. Miyamoto. (2002a). ''A General Framework for Estimation and Inference of Geographically Weighted Regression Models: 1. Location Specific Kernel Bandwidths and a Test for Locational Heterogeneity.'' *Environment and Planning A* 34, 733–54.

Páez, A., T. Uchida, and K. Miyamoto. (2002b). ''A General Framework for Estimation and Inference of Geographically Weighted Regression Models: 2. Spatial Association and Model Specification Tests.'' *Environment and Planning A* 34, 883–904.

Páez, A. (2004). ''Anisotropic Variance Functions in Geographically Weighted Regression Models.'' *Geographical Analysis* 36, 299–314.

Páez, A. (2006). ''Exploring Contextual Variations in Land Use and Transport Analysis Using a Probit Model with Geographical Weights.'' *Journal of Transport Geography* 14, 167–76.

Rao, C. (1948). ''The Utilization of Multiple Measurements in Problems of Biological Classification (with Discussion).'' *Journal of the Royal Statistical Society (B)* 10, 159–203.

Rousseeuw, P., and K. van Dreissen. (1999). ''A Fast Algorithm for the Minimum Covariance Determinant estimator.'' *Technometrics* 41, 212–23.

Tufte, E. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.