Geographically Weighted Local Statistics Applied to Binary Data

Chris Brunsdon, Stewart Fotheringham, and Martin Charlton

Department of Geography
University of Newcastle upon Tyne, NE1 7RU, United Kingdom
{chris.brunsdon,stewart.fotheringham,martin.charlton}@ncl.ac.uk

Abstract. This paper considers the application of geographically weighting to summary statistics for binary data. We argue that geographical smoothing techniques that are applied to descriptive statistics for ratio and interval scale data may also be applied to descriptive statistics for binary categorical data. Here we outline how this may be done, focussing attention on the odds ratio statistic used for summarising the linkage between a pair of binary variables. An example of this is applied to data relating to house sales, based on over 30,000 houses in the United Kingdom. The method is used to demonstrate that time trends in the building of detached houses vary throughout the country.

1 Introduction

Previous work by the authors has developed the method of *Geographically Weighted Regression* (GWR) [1]. In this technique, the geographical stability of coefficients in regression models can be modelled, by locally calibrating regression models using a moving window or moving kernel technique. However, this approach need not be confined to regression models. Before more advanced statistical analysis takes place, it is generally good practice to carry out some initial exploratory data analysis (EDA), and to compute some descriptive statistics for the data set under consideration. As well as giving an overview of typical values and levels of variation for variables in the data set, EDA can help to identify outliers, detect general trends in the data, and identify potential problems that may occur in any modelling or more advanced statistical analysis that may subsequently take place. We argue here that "geographical weighting" as used in GWR is an approach that may also be applied to a broad range of statistical methods, including the computation of descriptive statistics.

In addition to the graphical methods for EDA such as those cited in [2], summary statistics are also a useful tool. Typical summary statistics include the mean and standard deviation of continuous variables, frequency tables (or proportion tables) for discrete variables, and correlation coefficients between pairs of continuous variables. We have argued elsewhere that these summary statistics are good candidates for "geographical weighting" [3]. In this paper, we consider another summary statistic—the *odds ratio*—which measures the dependency between pairs of binary variables. We argue that a geographically weighted version of this statistic may also be a useful exploratory tool.

M.J. Egenhofer and D.M. Mark (Eds.): GIScience 2002, LNCS 2478, pp. 38–50, 2002.

As an example, we consider a sample of 34,110 cases from the Nationwide UK property price data set[1]. This dataset contains details of properties (houses and flats) sold in England and Wales in 1992 where mortgages were granted by the Nationwide Building Society. A number of variables were recorded, but here we focus attention on just two: the variable DETACHED denotes whether a given property is detached or not, and the variable POST74 denotes whether a property was built after 1974. Both are binary variables—they can only take the values "Yes" or "No"—which may be denoted respectively by the numbers 0 and 1. The relationship between the two variables gives information about changes in trends in building over the last quarter century or so. Here, we investigate geographical trends in this relationship by developing a geographically weighted version of the global odds ratio statistic.

2 The Data Set

The data comprise anonymised records for property sales where the sale was completed between January and December 1992 inclusive. As well as the selling price of the property, details of the building include its type (detached, semi-detached house/bungalow, purpose-built flats, flat conversion), the number of bedrooms, number of bathrooms, nature of vehicle storage, details of central heating and floor area. Information is recorded for 34,110 houses and flats. The locational information is in the form of a postcode that can be matched to a grid reference using the UK Central Postcode Directory.

3 Geographically Weighted Summary Statistics for Binary Data

A useful basic summary statistic here is the proportion of "Yes" responses in the data set for each of the two binary variables DETACHED and POST74. Viewed as a global statistic, there are 6,667 "Yes" responses and 27,443 "No" responses for the variable DETACHED, so the proportion of detached properties in the data set as a whole is around 0.24. This provides some useful "overview" information—around one property in four is detached in England and Wales viewed as a whole. However this information, although useful, is rather general from a geographical viewpoint. Anyone who has travelled within the UK will be aware that it is a diverse place, and that the nature of its housing stock varies from locality to locality. For example, some more affluent areas may consist almost entirely of detached housing, but other equally affluent places, such as the London Docklands area, which has been dramatically redeveloped in the last decade, are dominated by luxury flats. As a (rather obvious) rule, there are more detached properties in sparsely populated areas. It would perhaps be more useful to divide the UK into a number of sub-regions (Census districts for example), and to tabulate or map the proportion of detached properties in each of these. Although this approach would provide a more helpful summary than the single figure of 0.24 given earlier, it relies on the assumption that the choice of sub-regions reflects the spatial patterns in the housing stock. If a "cluster" of detached housing straddles

[1] These data were kindly provided by the Nationwide Building Society, for which we are extremely grateful.

two sub-regions without dominating either of them, then it may fail to show up in the above analysis. This is an example of a phenomenon first noted by Gehlke and Biehl [4] termed the *Modifiable Areal Unit Problem* (MAUP) by Openshaw [5].

An alternative approach is outlined here based on the notion of geographical weighting. The underlying principle is very similar to that used in GWR—a kernel is placed around an arbitrary point *(u,v)* and a weighted proportion *p* of detached properties for all of the data located inside this window is computed:

$$p(u,v) = \frac{\sum x_i w_i(u,v)}{\sum w_i(u,v)} \tag{1}$$

where the term x_i denotes a binary indicator variable taking the value 1 if property *i* is detached, 0 if it is not and $w_i(u,v)$ is the weight assigned to property *i* according to some kernel function depending on the location of interest *(u,v)*. A typical function might be the bisquare kernel:

$$w_i(u,v) = \begin{cases} \left(1 - \left(\frac{d_i(u,v)}{h}\right)^2\right)^2 & \text{if } d_i(u,v) \le h \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

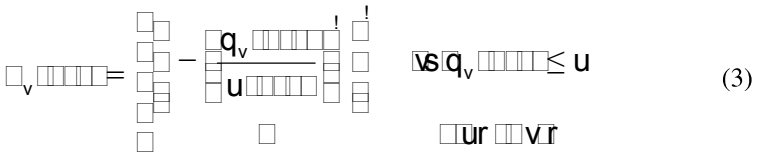
where $d_i(u,v)$ is the distance between the location of property i and the arbitrary point *(u,v)*. The constant *h* is termed the *bandwidth* and has the physical units of length. As the value of *h* increases, the bigger the circular window around *(u,v)* used to compute the local proportion of detached properties.

Note that if all of the w_i's were set to 1, p_i would simply be the number of detached properties divided by the total number of properties—that is, the global proportion. More generally, if $0 \le w_i \le 1$ then the weighting scheme can be thought of as turning some properties into "fractional" observations. For example, if $w_i(u,v) = 0.5$ we can think of property *i* as *half* an observation at least in the context of equations (1) and (2). For typical kernels, $w_i(u,v)$ decreases with distance from *(u,v)* so that properties further from the centre of the kernel have less influence on the local statistic. For the kernel specified in equation (2) properties further than a distance *h* from *(u,v)* have no influence at all.

One important issue here is the choice of the parameter *h*. When *h* is very large—for example notably larger than the distance between the furthest pair of x_i points in the data set, then $w_i(u,v)$ is very nearly equal to one for all values of *i*. In this case, the geographically weighted proportion estimate is virtually constant, and equal to the global proportion, regardless of the values of *u* and *v*. When *h* is very small, say in the order of magnitude of the distances between neighbouring x_i's, then $w_i(u,v)$ is zero for many of the *i*'s and therefore the geographically weighted proportion is computed on the basis of a very small number of x_i's in the vicinity of (u,v). Hence the value changes very rapidly as (u,v) scans over the geographical study area. If the localised statistic is depicted as a surface (which it may be since it is a function of *u* and *v*) then very small values of *h* tend to result in very spiky surfaces.

The above discussion suggests when *h* is either too big or too small, there are problems with the resulting surface, but it is hard to choose an *exact* value. In geo-

graphically weighted regression, the aim is to predict a dependent variable in terms of several independent variables and one of any number of indicators of quality of prediction might be used to "automatically" select h. However, the intention with geographically weighted summary statistics is more based on the desire to explore rather than model data, so perhaps there is an argument for experimentation over a reasonable range of values of h here. One further issue is whether the same value of h should be used for all values of u and v. For example, one might expect more detailed spatial variability in the age and type of housing in an urban area, where properties are packed closely together, than in an area outside of the city where properties are more sparsely positioned. In this case, one may wish to use smaller values of h in areas where the data points are more densely packed. To allow for this, equation (2) could be modified as:



(3)

The difference between this and equation (2) is that h is now expressed as a function of u and v, allowing the bandwidth to change geographically. To take into account the aim of reducing h in areas where the data is most dense, a k-nearest neighbour distance is a good choice for $h(u,v)$—that is, h is defined to be the distance between (u,v) and the kth nearest point to (u,v) . A typical choice for k might be 10% of the size of the data set, n.

By allowing this window-based statistic to be computed at closely spaced regular grid points, a surface or pixel image of the spatial variation in the proportion of detached properties can be constructed. If the spacing of the grid is sufficiently close, then any "cluster" of detached properties should become apparent—at least one of the grid points should be reasonably close the centre of the cluster. This should address the "straddling" problem discussed above. Note that the above approach is a specific example of a much more general group of methods. Here we regard the proportion of detached properties as a summary statistic and we suggest exploring variation over geographical space with a *geographically weighted summary statistic*. As an example, a map of showing the variation of the DETACHED variable for our data set is given in Fig. 1. Here the bandwidth h was selected according to the kth nearest neighbour method discussed above—with k being 5% of the sample size. There is a distinct east/west trend—the local proportion of detached properties in our sample tends to be highest along the east coast of the UK. The highest proportions of geographically weighted proportions of detached houses would appear to be in the rural lowland parts of eastern Britain. On the other hand, the "low spots" seem to pick out "Metroland" in London and Surrey/Sussex and Kent. Large parts of East Anglia, Lincolnshire, and eastern Yorkshire and Northumberland are in the high areas, as well as a band to the north of London including areas such as Stevenage, Chelmsford and Luton. The area covering London itself, perhaps not surprisingly, has a lower proportion of detached properties.

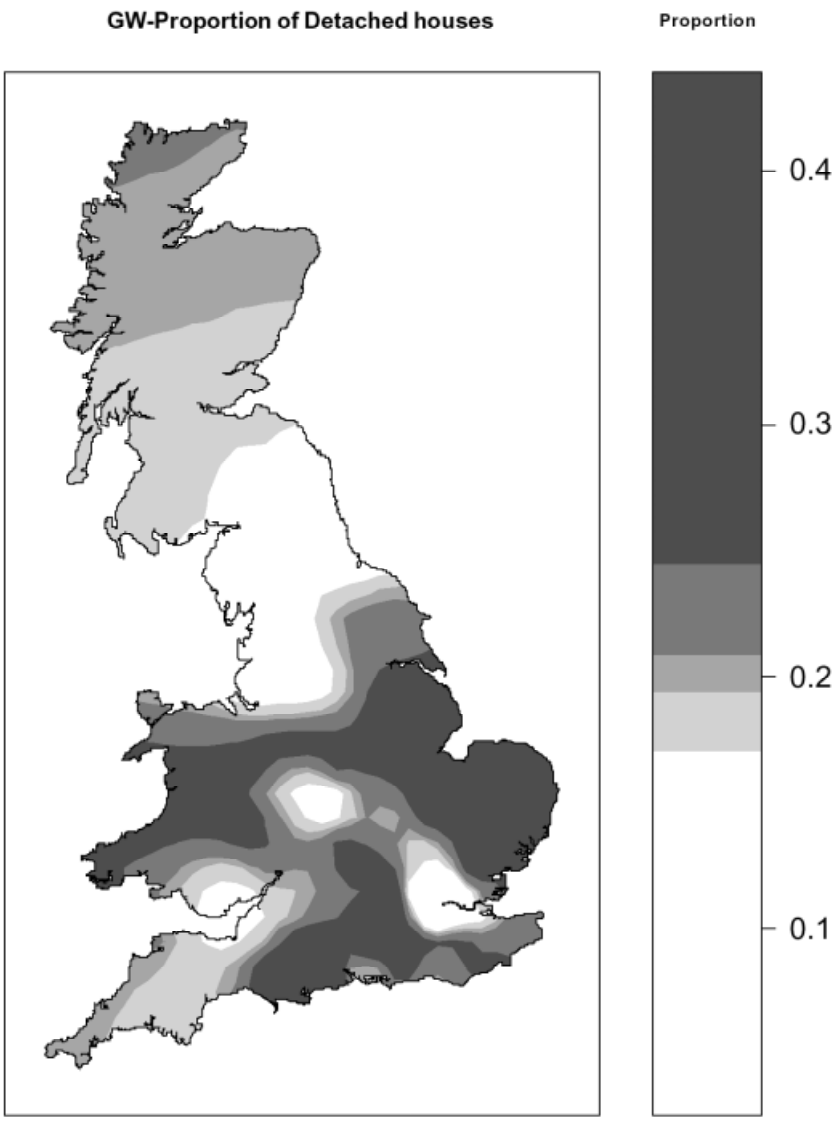


Fig. 1. Geographically Weighted Proportion of Detached Houses

4 A Pair of Binary Variables

Now, we consider the bivariate case where there are two binary variables. In this instance, we can summarise the global relationship between the variables using a 2x2 contingency table. For example, in addition to the DETACHED variable used in the last section, we will consider the variable POST74 defined in the previous section. To address geographical variation in the relationship between a pair of binary variables, we propose using a geographically weighted variation of the *odds ratio*. In this section, we define the odds ratio, and the *log odds ratio* before considering how this statistic can be geographically weighted.

The relationship between DETACHED and POST74 is summarised in table 1.

Table 1. Contingency table for the binary variables DETACHED and POST74

	DETACHED	
POST74	True	False
True	2,200	4,455
False	4,467	22,988

From this table, it appears that for the UK as a whole, newer (post-1974) properties are more likely to be detached. This could be confirmed using a χ^2 test. Here, χ^2 =958.9 with 1 d.f., which is extremely significant—the 95% point is 3.54. The χ^2 test is therefore a single number summary that measures the association between the two binary variables. However, for 2x2 tables another, perhaps more intuitive, measure exists. This is the so-called *odds ratio*. According to the data, the odds of a property being detached if it were built during or after 1975 are 2,200:4,455. The same odds for a property built before 1975 are 4,467:22,988. We can express these odds as real numbers, by carrying out the implicit divisions: 2,200:4,455 = 2,200/4,455 = 0.494 and 4,467:22,988 = 4,467/22,988 = 0.194. We can now divide one set of odds by the other: 0.494/0.194 = 2.54. This tells us that the odds of a property being detached if it were built during or after 1975 are about two and a half times what they are if the property were built before then. The quantity derived in this way is termed the *odds ratio*. In general, if a 2x2 contingency table is denoted as

a	b
c	d

(4)

then the odds ratio is defined as

$$a/c \div b/d \quad \text{or} \quad \frac{ad}{bc}. \tag{5}$$

It is possible to obtain an approximate confidence interval for the odds ratio. If we assume that *a, b, c* and *d* are independently Poisson distributed (as we would when applying a generalised linear model, for example), then the variances of *a, b, c* and *d* are approximately *a, b, c* and *d* also—assuming that the actual values *a*, *b*, *c* and *d* are reasonable estimates for their expected values. Now consider taking the natural logarithm of the odds ratio. This quantity is equal to

$$\log_e(a) + \log_e(d) - \log_e(b) - \log_e(c). \quad (6)$$

Now, it can be shown that approximately

$$\text{var}(\log_e(x)) = \{E(x)\}^{-1}\text{var}(x) \quad (7)$$

so that

$$\text{var}(\text{log odds ratio}) = a^{-1} + b^{-1} + c^{-1} + d^{-1}. \quad (8)$$

The log odds ratio may be reasonably modelled by a Normal distribution—since the log transform reduces the skewness of the distribution of the untransformed odds ratio. Note that while the odds ratio may take values between zero and infinity, the log odds ratio may take values anywhere on the real line. Confidence intervals for the log odds ratio may be computed on the basis of (8). From this it is the possible to compute confidence intervals for the untransformed odds ratio by taking natural antilogarithms.

For our data, the odds ratio is

$$\frac{!\,!\,\square\square\square!\,!\,(\,'\,'}{\#\#\$\$\square\square\#\#\%\&} = !\,\square\$\# \quad (9)$$

giving a log odds ratio of 0.932. Assuming approximate normality of the log odds ratio as suggested above, this gives an approximate 95% CI of (0.87,0.99) for the log odds ratio, and therefore a 95% CI of (2.39, 2.69) for the untransformed odds ratio.

The odds ratio is quite commonly used in epidemiological applications, but is perhaps less well known in other areas. For 2x2 tables it has the advantage that it is more "intuitive" than the χ^2 statistic—statements such as "The odds that a post-1974 property is detached are about two and a half times as high as those for a pre-1975 property" are easily interpreted, whereas the interpretation of a χ^2 statistic of 15.2 is less immediately linked to the "real world." The confidence intervals also have a natural interpretation—for the 95% CIs given above, we can say that subject to sampling variability, the odds that a post-1974 is detached is between around 2.4 to 2.7 times those for a pre 1974 property. Another advantage is that the odds ratio gives an indication of the "direction" of the relationship between the variables—in this case whether a property being built during or after 1975 makes is more or less likely to be detached. The standard χ^2 only measures the level of association, not the direction. This latter problem is addressed using the signed-χ^2 measure [6] although this approach still suffers from the lack of directness of interpretation in comparison to the odds ratio.

However, it would be unreasonable not to also point out an obvious advantage of the χ^2 statistic. Such a statistic may be applied to tables other than 2x2, and so may be used as a geographically weighted measure of association for categorical variables having more than two values. In this paper we consider the odds ratio approach mainly because of its intuitive appeal for binary variables. Rogerson [7] has already considered the idea of localised χ^2 -statistics in depth. Also, we feel that the odds ratio is a more appropriate statistic in the special case of 2x2 tables: the aim of here is more to consider localised *summary* statistics than inferential statistics, and since the odds ratio is an intuitively interpretable quantity, the odds ratio best satisfies this requirement.

To define a GW-odds ratio, we first consider the idea of a localised 2x2 contingency table. For each binary variable pair (x_i,y_i) we can define four indicator variables, say A_i, B_i, C_i, D_i, one for each of the four possible states that (x_i,y_i) can take, as set out in table 2.

Table 2. Values of indicator variables A_i, B_i, C_i and D_i in relation to (x_i,y_i)

x_i	y_i	A_i	B_i	C_i	D_i
0	0	1	0	0	0
1	0	0	1	0	0
0	1	0	0	1	0
1	1	0	0	0	1

Then it may be checked that $a = \Sigma A_i$, $b = \Sigma B_i$, $c = \Sigma C_i$ and $d = \Sigma D_i$. Suppose now we wish to consider the odds ratio locally—that is, in the vicinity of some point (u,v). We may then redefine *a, b, c* and *d* in terms of the *weighted* sums of A_i, B_i, C_i and D_i.

As before, the weighting function will be some form of distance-based kernel—for example

$$h\square\square\square\square\square\square = \sum_{v} 6_{v}\square_{v}\square\square\square\square\square\square \qquad (10)$$

where $w_i(u,v)$ is defined as in equation (2). Here we write *a* as a function of *u* and *v* to demonstrate that as the point (u,v) varies, the value of *a* will change. We can construct similar expressions to (10) for $b(u,v)$, $c(u,v)$ and $d(u,v)$. These may then be combined to produce a local odds ratio around the point (u,v) giving

$$PS\square\square\square\square\square\square = \frac{h\square\square\square\square\square q\square\square\square\square\square}{i\ \square\square\square\square\square p\square\square\square\square\square}. \qquad (11)$$

Using equation (11) we may therefore define and map a geographically weighted odds ratio, in much the same way as we defined a geographically weighted proportion earlier. We may also consider confidence intervals based on equation (8), by substituting *a, b, c* and *d* with $a(u,v)$, $b(u,v)$, $c(u,v)$ and $d(u,v)$ respectively. However, some caution should perhaps be exercised here. We state this for a number of reasons:

1. One of the assumptions made to obtain the approximate variance of the log-odds ratio was that *a, b, c* and *d* were Poisson random variables, but $a(u,v)$, $b(u,v)$, $c(u,v)$ and $d(u,v)$ are not even integers. However, since $a(u,v)$ … $d(u,v)$ are proportional to density estimates, the "variability band" approach of Bowman and Azzalini [8] suggests that this is a reasonable assumption.
2. If we do obtain CIs for OR(*u, v*) it is important to remember that these are just pointwise intervals. This implies that although these are the correct intervals for any given point (u,v) we cannot construct "confidence surfaces" from the upper and lower bounds such that there is a 95% probability of the true surface being sandwiched between our upper and lower bounding surfaces. This is mainly because the CIs for different points are not independent of one another.

With these provisos, it is reasonable to consider mapping some quantity related to the upper and lower confidence intervals—but mainly to give an informal idea of the

uncertainty of our geographically weighted log odds ratio, rather than to provide the basis for any kind of formal testing.

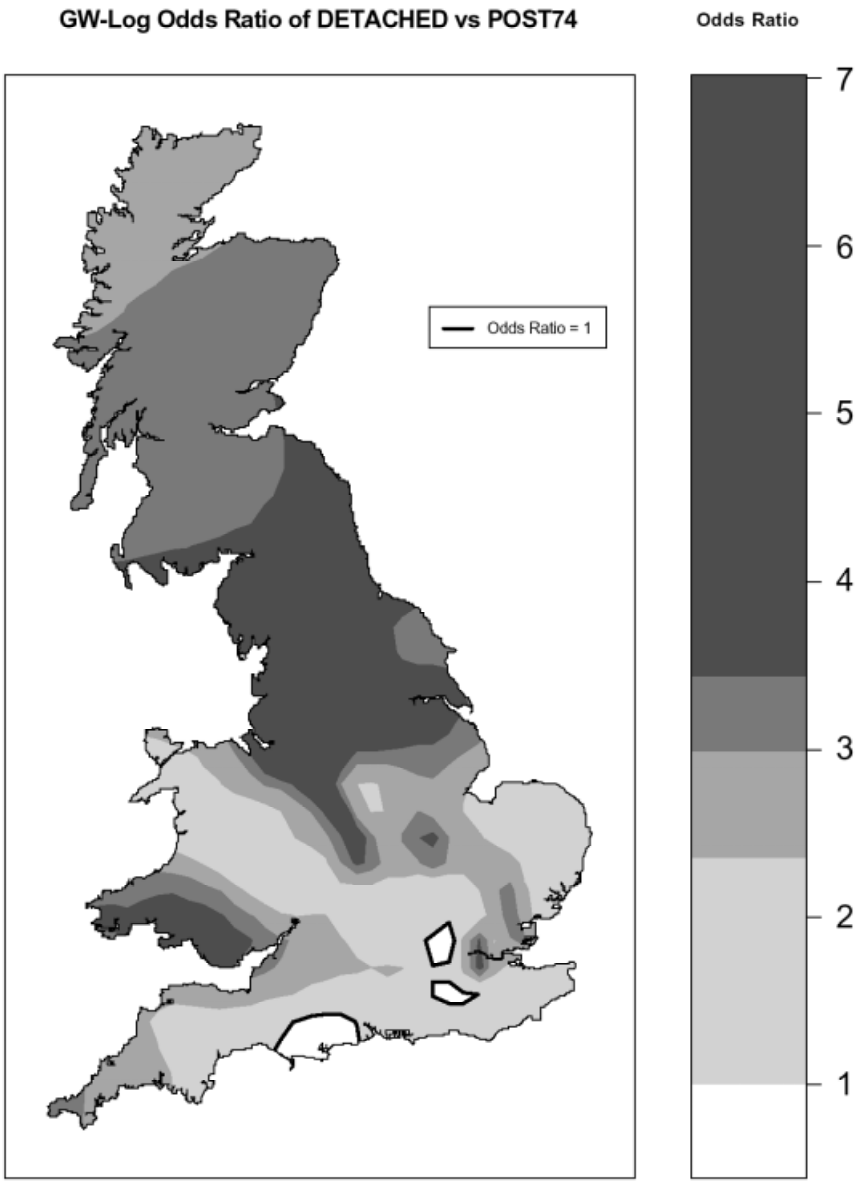


Fig. 2. Geographically Weighted Odds Ratio for DETACHED vs POST74

We demonstrate the technique in Fig. 2 by mapping the geographically weighted odds ratio for the variables DETACHED and POST74 introduced earlier in the paper. We choose the bandwidth locally according to the same method we used in the earlier analysis of proportions. For this shaded map, areas in white corresponded to places where the geographically weighted odds ratio is below 1. In these places, properties built post 1974 are *less* likely to be detached than those built in 1974 or before then. This suggests that in these areas, there is has been a reduction in the proportion of detached properties being built in recent years. This occurs in a two zones to the north and south of London as well as an area to the South-West of the other two areas, incorporating part of the South Coast. For everywhere else in the country, there has been a greater tendency to build detached properties in recent years—with this effect being more marked as one heads north. However, it is possible that some of the effects in Scotland could be a result of the relative lack of data from that part of the country.

Finally, in Fig. 3 we consider the variability of these odds ratio estimates, based on the technique for estimating standard errors described above. Showing the estimated odds ratio surface sandwiched between the upper and lower confidence limit surfaces would prove a difficult visualisation problem for the medium of monochrome print, so we do not consider this directly. Instead, we focus on the uncertainty of the location of the bounding lines between regions where the geographically weighted odds ratio is above unity, and those where it is below unity. Fig. 3 shows the contour line where the odds ratio is equal to unity (as thin black lines) and the contour line where the odds ratio minus one standard error is equal to unity (thicker black lines). The second set of lines suggest that rather than two distinct zones to the north and to the south of London having odds ratios greater than one, there is now a single zone covering Greater London and some surrounding area. In some circumstances, it would also be helpful to consider the unity contour line for the odds ratio *plus* one standard deviation, but in this case, it is the latter quantity is greater than one throughout the UK and so such a contour line does not exist.

Fig. 4 is similar to Fig. 3 except that here the thicker black line shows the unity contour for the odds ratio minus two standard deviations. In this instance, all three distinct zones in Fig. 2 are now part of one larger zone. A reasonable way of interpreting Fig. 3 and Fig. 4 is to note that allowing for sampling errors in the data, one may be fairly clear that are areas to the north and south of London where detached houses are less likely to be built after 1974—but there is also a possibility that these are part of one larger zone which also incorporates part of London itself. It also seems likely that there is a region south-west of London incorporating Portsmouth, Southampton and Brighton also having this characteristic. Fig. 4 suggests a possibility that all three zones may be part of a larger zone—and that one or two other similar zones may exist. The general north/south trend observed in Fig. 2 appears to a reflect current concern in UK housing—there are currently many news items discussing the increasing demand for housing in southern parts of England—and it seems reasonable that this demand would result in more compact housing (i.e., non-detached properties) being built.

Fig. 3. Comparison of contour lines for Odds Ratio and Odds Ratio minus Odds Ratio Standard Error

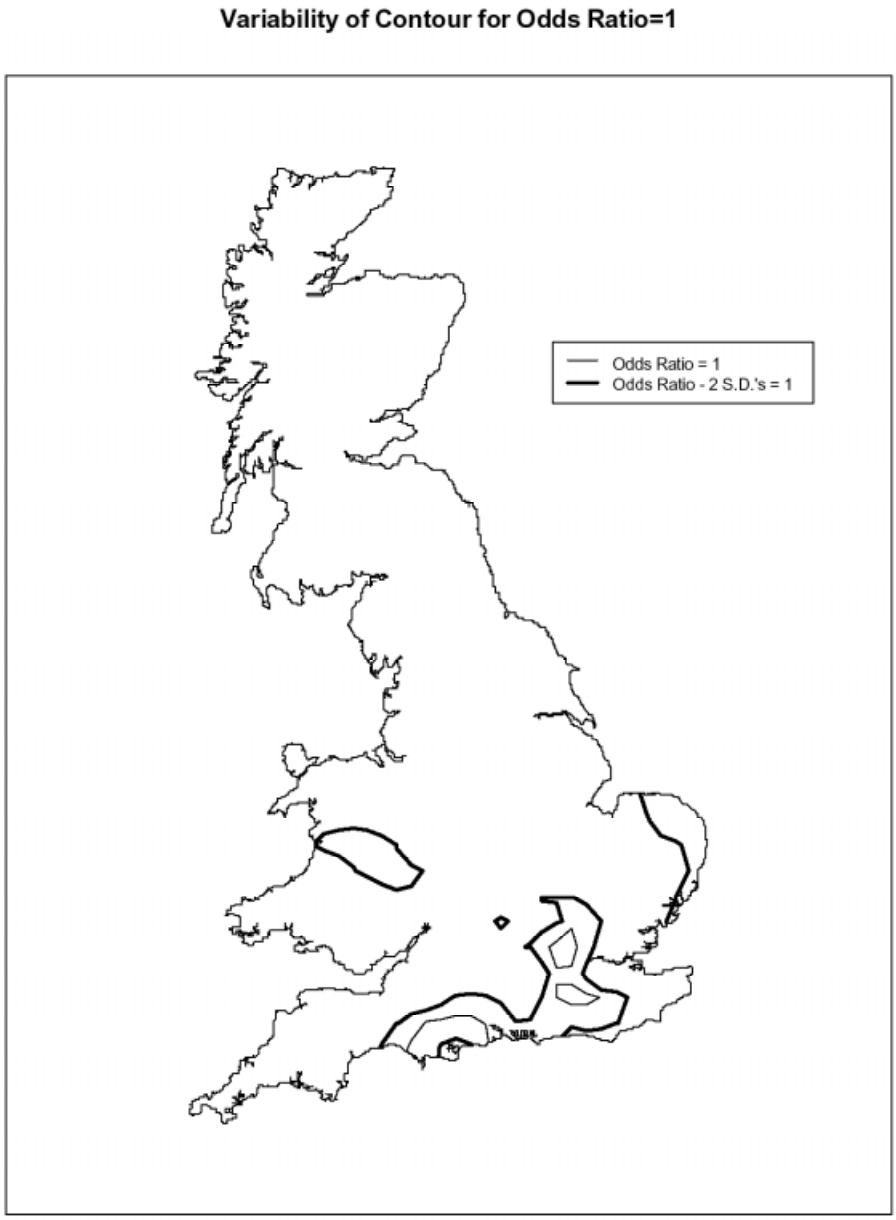


Fig. 4. Comparison of contour lines for Odds Ratio and Odds Ratio minus twice the Odds Ratio Standard Error

5 Conclusions

In this paper two geographically weighted summary statistics have been discussed. The first of these, the geographically weighted proportion, is essentially a variant on a moving window mean smooth, whereas the second, the geographically weighted odds ratio, is quite different from many moving window approaches considered before. An important feature of both of these methods is that they focus on binary variables rather than ratio or interval scale variables, whereas most moving window or kernel-based approaches (with the notable exception of [7]) tend to focus on the latter. This may become useful in a number of areas of geographical research where local statistical approaches have not often been applied—for example the results from post-coded questionnaires in which several yes/no questions are asked may be analysed using techniques such as these. One such area is epidemiological research—an area of study that already makes much use of the global odds ratio. For example, one could consider spatial variation in gender differences in the risk of contracting certain types of illness, and how these change geographically.

The general intention here is to expand the kinds of data to which geographically weighted methods may be applied. This paper outlines how the relationship between a pair of binary variables may alter over space. Some important future challenges include considering other combinations of variable types, for example the spatial changes in the relationship between one binary and one continuous variable. It is hoped that future research will address issues such as this.

References

1. Brunsdon, C., Fotheringham, A.S., and Charlton, M. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity, *Geographical Analysis* **28** (1996) 281-298
2. Fotheringham, A.S., Brunsdon, C. and Charlton, M. *Quantitative Geography: Perspectives on Spatial Data Analysis*, Sage, Thousand Oaks CA (2000)
3. Brunsdon, C., Fotheringham, A.S., and Charlton, M. Geographically Weighted Summary Statistics—A Framework for Localised Exploratory Data Analysis, *Computers and Urban Environmental Systems,* (Forthcoming)
4. Gehlke, C. and Biehl, K. Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material*, Journal of the American Statistical Association*, **29** (1934) 169-170
5. Openshaw, S. *The Modifiable Areal Unit Problem,* CATMOG 38, Geo-Abstracts, Norwich (1984)
6. Visvalingham M. The Signed Chi-Square Measure for Mapping, *Cartography Journal* **15** (1978) 93-98
7. Rogerson, P. The Detection of Clusters Using a Spatial Version of the Chi-Square Goodness-of-Fit Test , *Geographical Analysis* **31** (1999) 351-357
8. Bowman, A. and Azzalini, A. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations, Oxford University Press (1997)