Geographically Weighted Regression
the analysis of spatially varying relationships

A. Stewart Fotheringham
Chris Brunsdon
Martin Charlton
University of Newcastle, UK
Geographically Weighted Regression
Geographically Weighted Regression

the analysis of spatially varying relationships

A. Stewart Fotheringham
Chris Brunsdon
Martin Charlton

University of Newcastle, UK

JOHN WILEY & SONS, LTD
ASF:  To Barbara, Iain and Neill
CB:   To Francis and Anne
MEC:  To Ted and Avril
Contents

Acknowledgements xi

1 Local Statistics and Local Models for Spatial Data 1
  1.1 Introduction 1
  1.2 Local Aspatial Statistical Methods 3
  1.3 Local versus Global Spatial Statistics 6
  1.4 Spatial Non-stationarity 9
  1.5 Examples of Local Univariate Methods for Spatial Data Analysis 11
    1.5.1 Local Forms of Point Pattern Analysis 11
    1.5.2 Local Graphical Analysis 12
    1.5.3 Local Filters 13
    1.5.4 Local Measures of Spatial Dependency 14
  1.6 Examples of Local Multivariate Methods for Spatial Data Analysis 15
    1.6.1 The Spatial Expansion Method 16
    1.6.2 Spatially Adaptive Filtering 17
    1.6.3 Multilevel Modelling 18
    1.6.4 Random Coefficient Models 20
    1.6.5 Spatial Regression Models 21
  1.7 Examples of Local Methods for Spatial Flow Modelling 24
  1.8 Summary 25

2 Geographically Weighted Regression: The Basics 27
  2.1 Introduction 27
  2.2 An Empirical Example 27
    2.2.1 The Data 28
    2.2.2 A Global Regression Model 28
    2.2.3 Global Regression Results 34
  2.3 Borough-Specific Calibrations of the Global Model 38
  2.4 Moving Window Regression 42
  2.5 Geographically Weighted Regression with Fixed Spatial Kernels 44
  2.6 Geographically Weighted Regression with Adaptive Spatial Kernels 46
2.7 The Mechanics of GWR in More Detail
2.7.1 The Basic Methodology 52
2.7.2 Local Standard Errors 54
2.7.3 Choice of Spatial Weighting Function 56
2.7.4 Calibrating the Spatial Weighting Function 59
2.7.5 Bias-Variance Trade-Off 62
2.8 Testing for Spatial Non-stationarity 63
2.9 Summary 64

3 Extensions to the Basic GWR Model 65
3.1 Introduction 65
3.2 Mixed GWR Models 65
3.3 An Example 68
3.4 Outliers and Robust GWR 73
3.5 Spatially Heteroskedastic Models 80
3.6 Summary 82

4 Statistical Inference and Geographically Weighted Regression 83
4.1 Introduction 83
4.2 What is Meant by ‘Inference’ and How Does it Relate to GWR? 84
4.2.1 How Likely is it that Some Fact is True on the Basis of the Data? 85
4.2.2 Within What Interval Does Some Model Coefficient Lie? 85
4.2.3 Which One of a Series of Potential Mathematical Models is ‘Best’? 86
4.3 GWR as a Statistical Model 87
4.3.1 Local Likelihood 90
4.3.2 Using Classical Inference – Working with p-values 91
4.3.3 Testing Individual Parameter Stationarity 92
4.4 Confidence Intervals 94
4.5 An Alternative Approach Using the AIC 95
4.6 Two Examples 97
4.6.1 Basic Estimates 97
4.6.2 Estimates of Pointwise Standard Errors 99
4.6.3 Working with the AIC 99
4.7 Summary 102

5 GWR and Spatial Autocorrelation 103
5.1 Introduction 103
5.2 The Empirical Setting 104
5.3 Local Measures of Spatial Autocorrelation using GWR 104
5.4 Residuals in Global Regression Models and in GWR 112
5.5 Local Parameter Estimates from Autoregressive and Non-Autoregressive Models 117
5.6 Spatial Regression Models and GWR 121
5.6.1 Overview 121
5.6.2 Conditional Autoregressive (CA) Models 122
5.6.3 Simultaneous Autoregressive (SA) Models 122
5.6.4 GWR, Conditional Autoregressive Models and Simultaneous Autoregressive Models 123
5.7 Summary 124

6 Scale Issues and Geographically Weighted Regression 127
6.1 Introduction 127
6.2 Bandwidth and Scale: The Example of School Performance Analysis 130
   6.2.1 Introduction 130
   6.2.2 The School Performance Data 131
   6.2.3 Global Regression Results 133
   6.2.4 Local Regression Results 134
6.3 GWR and the MAUP 144
   6.3.1 Introduction 144
   6.3.2 An Experiment 147
6.4 Summary 153

7 Geographically Weighted Local Statistics 159
7.1 Introduction 159
7.2 Basic Ideas 161
7.3 A Single Continuous Variable 163
7.4 Two Continuous Variables 173
7.5 A Single Binary Variable 175
7.6 A Pair of Binary Variables 177
7.7 Towards More Robust Geographically Weighted Statistics 181
7.8 Summary 183

8 Extensions of Geographical Weighting 187
8.1 Introduction 187
8.2 Geographically Weighted Generalised Linear Models 188
   8.2.1 A Poisson GWGLM 190
   8.2.2 A Binomial GWGLM 193
8.3 Geographically Weighted Principal Components 196
   8.3.1 Local Multivariate Models 196
   8.3.2 Calibrating Local Multivariate Models 198
   8.3.3 Interpreting $\Sigma$ and $\rho$ 199
   8.3.4 An Example 200
8.4 Geographically Weighted Density Estimation 202
   8.4.1 Kernel Density Estimation 202
   8.4.2 Geographically Weighted Kernels 203
   8.4.3 An Example Using House Prices 203
8.5 Summary 205
9 Software for Geographically Weighted Regression

9.1 Introduction

9.2 Some Terminology

9.3 The Data File

9.4 What Do I Need to Specify?

9.5 Kernels

9.6 Choosing a Bandwidth

9.6.1 User-Supplied Bandwidth

9.6.2 Estimation by Cross-validation

9.6.3 Estimation by Minimising the AIC

9.6.4 The Golden Section Search

9.7 Significance Tests

9.8 Casewise Diagnostics for GWR

9.8.1 Standardised Residuals

9.8.2 Local $r$-square

9.8.3 Influence Statistics

9.9 A Worked Example

9.9.1 Running GWR 2.0 on a PC

9.9.2 The Outputs

9.9.3 Running GWR 2.0 under UNIX

9.10 Visualising the Output

9.10.1 Viewing the Results in ArcView

9.10.2 Point Symbols

9.10.3 Area Symbols

9.10.4 Contour Plots

9.10.5 Pseudo-3D Display

9.11 Summary

10 Epilogue

10.1 Overview

10.2 Summarising the Book

10.3 Empirical Applications of GWR

10.4 Software Development

10.4.1 Embedding GWR in Larger Packages

10.4.2 Software Extending the Basic GWR Idea

10.5 Cautionary Notes

10.5.1 Multiple Hypothesis Testing

10.5.2 Locally Varying Intercepts

10.5.3 Interpretation of Parameter Surfaces

10.6 Summary

Bibliography

Index
Acknowledgements

The UK maps in this book are based on copyright digital map data owned and supplied by Bartholomew Ltd and are used with permission. Some of the maps are also based on census data provided with the support of the ESRC and JISC and use boundary material which is the copyright of the Crown and the ED-LINE Consortium. The US census data and boundaries were obtained from CensusCD+Maps, a product of Geolytics Inc. The authors are grateful for the enlightened attitude of the US Government in making spatial data relatively freely available.

Throughout the book we make extensive use of house price data that has been supplied by the Nationwide Building Society to the University of Newcastle upon Tyne and we are extremely grateful for their generosity.

Dr Robin Flowerdew of the Department of Geography at the University of St Andrews supplied the school performance data used in Chapter 5 as part of a conference on local modelling with spatial data.

A number of people also deserve credit for assisting with various aspects of this book. Ann Rooke applied her usual superlative cartographic skills to some of the figures. Stamatis Kalogirou wrote the Visual Basic front end to the GWR software and Barbara Fotheringham did a very professional job of helping to proofread the manuscript.

We would also like to thank Sally Wilkinson, Lyn Roberts and Keily Larkins at John Wiley & Sons, Ltd for their encouragement, patience, assistance and good nature during the various evolutionary stages of this book. Further, we acknowledge a great debt to the reviewers of both the initial book proposal and an earlier version of the finished product for their strong support and useful insights. The book is far better for their comments.

Finally, we make the usual disclaimer that any errors remaining in the book are the sole responsibility of the authors – apologies for not catching them all!
This publication contains maps based on copyright digital map data owned and supplied by Bartholomew Ltd and is used with permission.
This applies to: Figures 2.1, 2.2, 2.4, 2.9, 2.12, 2.14–2.19, 5.1, 6.2–6.7, 7.1–7.6, 7.8–7.12, 8.1–8.3. Data taken from Bartholomews.
Figures 2.1, 2.2, 2.9, 2.12, 2.13–2.19, 5.1. Data taken from Bartholomews and UKBorders.
Figures 2.2, 5.1. Data taken from Bartholomews, UKBorders and Nationwide Building Society.
Maps are based on data provided with the support of the ESRC and JISC and use boundary material which is copyright of the Crown and the EDLINE consortium.
This applies to: Figures 2.1–2.3, 2.5–2.7, 2.9, 2.12, 2.14–2.19, 3.1, 3.3–3.5, 3.8, 3.9, 3.11, 3.12, 5.1–5.4, 5.6–5.10, 6.8–6.13.
Table 2.1 is calculated from data supplied by the Nationwide Building Society to the University of Newcastle upon Tyne.
1

Local Statistics and Local Models for Spatial Data

1.1 Introduction

Imagine reading a book on the climate of the United States which contained only data averaged across the whole country, such as mean annual rainfall, mean annual number of hours of sunshine, and so forth. Many would feel rather short-changed with such a lack of detail. We would suspect, quite rightly, that there is a great richness in the underlying data on which these averages have been calculated; we would probably want to see these data, preferably drawn on maps, in order to appreciate the spatial variations in climate that are hidden in the reported averages. Indeed, the averages we have been presented with may be practically useless in telling us anything about climate in any particular part of the United States. It is known, for instance, that parts of the north-western United States receive a great deal more precipitation than parts of the Southwest and that Florida receives more hours of sunshine in a year than New York. In fact, it might be the case that not a single weather station in the country has the characteristics depicted by the mean climatic statistics.

The average values in this scenario can be termed global observations: in the absence of any other information, they are assumed to represent the situation in every part of the study region. The individual data on which the averages are calculated can be termed local observations: they describe the situation at the local level.¹

¹ There is at least one other slightly different definition of ‘local’ and ‘global’ in the literature. Thioulouse et al. (1995) define a local statistic as one which is calculated on pairs of points or areas which are adjacent and a global statistic as one calculated over all possible pairs of points or areas. Their use of the term ‘local’, however, is not the same as used throughout this book because it still produces a global model; it merely separates the model applications into different spatial regimes.
Only if there is little or no variation in the local observations do the global observations provide any reliable information on the local areas within the study area. As the spatial variation of the local observations increases, the reliability of the global observation as representative of local conditions decreases.

While the above scenario might appear rather ludicrous (surely no one would publish a book containing average climatic data without describing at least some of the local data?), consider a second scenario which is much more plausible and indeed describes a methodology which is exceedingly common in spatial analysis. Suppose we had data on house prices and their determinants across the whole of England and that we wanted to model house price as a function of these determinants (such models are often referred to as hedonic price models and an example of the calibration of these models is provided in Chapter 2). Typically, we might run a regression of house prices on a set of structural attributes of each house, such as the age and floor area of the house; a set of neighbourhood attributes, such as crime rate or unemployment rate; and a set of locational attributes, such as distance to a major road or to a certain school. The output from this regression would be a set of parameter estimates, each estimate reflecting the relationship between house price and a particular attribute of the house. It would be quite usual to publish the results of such an analysis in the form of a table describing the parameter estimates for each attribute and commenting on their sign and magnitude, possibly in relation to some *a priori* set of hypotheses. In fact this is the standard approach of the vast majority of empirical analyses of spatial data.

However, the parameter estimates in this second scenario are *global statistics* and are possibly just as inadequate at representing local conditions as are the average climatic data described above. Each parameter estimate describes the *average* relationship between house price and a particular attribute of the house across the study region (in this case, the whole of England). This average relationship might not be representative of the situation in any particular part of England and may hide some very interesting and important local differences in the determinants of house prices. For example, suppose one of the determinants of house prices in our model is the age of the house and the global parameter estimate is close to zero. Superficially this would be interpreted as indicating that house prices are relatively independent of the age of the property. However, it might well be that there are contrasting relationships in different parts of the study area which tend to cancel each other out in the calculation of the global parameter estimate. For example, in rural parts of England, old houses might have character and appeal, thus generating higher prices than newer houses, *ceteris paribus*, whereas in urban areas, older houses, built to low standards to house workers in rapidly expanding cities at the middle of the nineteenth century, might be in poor condition and have substantially lower prices than newer houses. This local variation in the relationship between house price and age of the house would be completely lost if all that is reported is the global parameter estimate. It would be far more informative to produce a set of *local statistics*, in this case local parameter estimates, and to map these than simply to rely on the assumption that a single global estimate will be an accurate representation of all parts of the study area.

The only difference between the examples of the US climate and English house prices presented above is that the first describes the representation of spatial data,
whereas the second describes the representation of spatial relationships. It would seem that while we generally find it unhelpful to report solely global observations on spatial data, we are quite happy to accept global statements of spatial relationships. Indeed, as hinted at above, journals and textbooks in a variety of disciplines dealing with spatial data are filled with examples of global forms of spatial analysis. Local forms of spatial analysis or spatial models are very rare exceptions to the overwhelming tide of global forms of analysis that dominates the literature.

In this book, through a series of examples and discussions, we hope to convince the reader of the value of local forms of spatial analysis and spatial modelling, and in particular, the value of one form of local modelling which we term Geographically Weighted Regression (GWR). We hope to show that in many instances undertaking a global spatial analysis or calibrating a global spatial model can be as misleading as describing precipitation rates across the USA with a single value.

1.2 Local Aspatial Statistical Methods

Spatial data contain both attribute and locational information: aspatial data contain only attribute information. For instance, data on the manufacturing output of firms graphed against the number of their employees are aspatial, whereas the numbers of people suffering from a certain type of disease in different parts of a country are spatial. Unemployment rates measured for one location over different time periods are aspatial but unemployment rates at different locations are spatial and the spatial component of the data might be very useful in understanding why the rates vary. The difference between aspatial and spatial data is important because many statistical techniques developed for aspatial data are not valid for spatial data. The latter have unique properties and problems that necessitate a different set of statistical techniques and modelling approaches (for more on this, see Fotheringham et al. 2000, particularly Chapter 2). This is also true in local analysis.

There is a growing literature and an expanding array of techniques for examining local relationships in aspatial data. For example, there are techniques such as the use of spline functions (Wahba 1990; Friedman 1991; Green and Silverman 1994); LOWESS regression (Cleveland 1979); kernel regression (Cleveland and Devlin 1988; Wand and Jones 1995; Fan and Gijbels 1996; Thorsnes and McMillen 1998); and variable parameter models in the econometric literature (Maddala 1977; John- son and Kau 1980; Raj and Ullah 1981; Kmenta 1986; Casetti 1997) that are applicable to the local analysis of aspatial data. Good general discussions of local regression techniques for aspatial data are given by Hardle (1990), Barnett et al. (1991), Loader (1999) and Fox (2000a; 2000b).

The basic problem that local statistics attempt to solve is shown in Figure 1.1. Here there is a relationship between two aspatial variables, \( Y \) and \( X \), which needs to be determined from the observed data. A global linear regression model, for example, would produce a relationship such as that depicted by line \( A \); although the model gives a reasonable fit to the data, it clearly misses some important local variations in the relationship between \( Y \) and \( X \). Here, notice, ‘local’ means in terms of attribute space, in this case that of the \( X \) variable, rather than geographical
A local technique, such as a linear spline function, depicted by line $B$, would give a more accurate picture of the relationship between $Y$ and $X$. This would be obtained by essentially running four separate regressions over different ranges of the $X$ variable with the constraint that the end points of the local regression lines meet at what are known as ‘knots’. Finally, a very localised technique such as LOWESS regression would yield line $C$ where the relationship between $Y$ and $X$ is evaluated at a large number of points along the $X$ axis and the data points are weighted according to their ‘distance’ from each of these regression points. For example, suppose the regression point were at $x_1$. Then the data points for the regression of $Y$ on $X$ would be weighted according to their distance from the point $x_1$ with points closer to $x_1$ being weighted more heavily than points further away. This weighted regression yields a local estimate of the slope parameter for the relationship between $Y$ and $X$. The regression point is moved along the $X$ axis in small intervals until a line such as that in $C$ can be constructed from the set of local parameter estimates.

For something of a hybrid application of local modelling the reader is referred to McMillan (1996) in which land values in Chicago are regressed on distance to various features within the city. Although this is essentially an aspatial model because the local regressions are calibrated only in attribute space and not in geographical space, the use of distance as an independent variable does allow a spatial interpretation of the results to be made. As such, McMillan’s application can be thought of as ‘semi-spatial’.

Although a linear spline function is depicted in this example, cubic spline functions are often used in curve fitting exercises. The linear spline is shown here to distinguish it from the LOWESS fit.

The terms LOWESS and LOESS are used interchangeably in the literature; use is based on personal preference.
The difference between applying local techniques to aspatial data and to spatial data is that the relationship between $Y$ and $X$, as shown in Figure 1.1 might vary depending on the location at which the regression is undertaken. That is, instead of simply having the problem of fitting a non-linear function to a set of data, this non-linear function itself might vary over space as shown for two locations in Figure 1.2. Consequently, local statistical analyses for spatial data have to cope

*Figure 1.2 Local relationships in attribute space for two geographical locations*
with two potential types of local variation: the local relationship being measured in attribute space and the local relationship being measured in geographical space. Compounding the problem of measuring spatial variations in relationships is the fact that the relationships in geographical space can vary in two dimensions rather than just in one. That is, local variations in attribute space, such as those shown in Figure 1.1, take place along a line and the dependency between relationships is easier to establish than in the two-dimensional equivalent of geographical space.

Because local statistical techniques for aspatial data are already fairly well established and because such techniques do not always translate easily to spatial data, the remainder of this book concentrates almost exclusively on the local analysis of spatial data. Henceforth, any discussion of local analysis is assumed to refer to spatial data unless otherwise stated.

1.3 Local versus Global Spatial Statistics

Local statistics are treated here as spatial disaggregations of global statistics. For instance, the mean rainfall across the USA is a global statistic; the measured rainfall in each of the recording stations, i.e. the data from which the mean is calculated, represent the local statistics. A model calibrated with data equally weighted from across a study region is a global model that yields global parameter estimates. A model calibrated with spatially limited sets of data is a local model that yields local parameter estimates. Local and global statistics differ in several respects as shown in Table 1.1.

Global statistics are typically single-valued: examples include a mean value, a standard deviation and a measure of the spatial autocorrelation in a data set. Local statistics are multi-valued: different values of the statistic can occur in different locations within the study region. Each local statistic is a measure of the attribute or the relationship being examined in the vicinity of a location within the study

<table>
<thead>
<tr>
<th>Table 1.1 Differences between local and global statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global</strong></td>
</tr>
<tr>
<td>Summarise data for whole region</td>
</tr>
<tr>
<td>Single-valued statistic</td>
</tr>
<tr>
<td>Non-mappable</td>
</tr>
<tr>
<td>GIS – unfriendly</td>
</tr>
<tr>
<td>Aspatial or spatially limited</td>
</tr>
<tr>
<td>Emphasise similarities across space</td>
</tr>
<tr>
<td>Search for regularities or ‘laws’</td>
</tr>
<tr>
<td>Example:</td>
</tr>
<tr>
<td>Classic Regression</td>
</tr>
</tbody>
</table>
region: as this location changes, the local statistic can take on different values. Consequently, global statistics are unmappable or ‘GIS-unfriendly’, meaning they are not conducive to being analysed within a Geographic Information System (GIS) because they consist of a single value. Local statistics, on the other hand, can be mapped and further examined within a GIS. For instance, it is possible to produce a map of local parameter estimates showing how a relationship varies over space and then to investigate the spatial pattern of the local estimates to establish some understanding of possible causes of this pattern. Indeed, given that very large numbers of local parameter estimates can be produced, it is almost essential to map them in order to make some sense of the pattern they display. Local statistics are therefore spatial statistics whereas global statistics are aspatial or spatially limited.

By their nature, local statistics emphasise differences across space whereas global statistics emphasise similarities across space. Global statistics lead one into thinking that all parts of the study region can be accurately represented by a single value whereas local statistics can show the falsity of this assumption by depicting what is actually happening in different parts of the region. Consequently, local statistics are useful in searching for exceptions or what are known as local ‘hot spots’ in the data. This use places them in the realm of exploratory spatial data analysis where the emphasis is on developing hypotheses from the data, as opposed to the more traditional confirmatory types of analysis in which the data are used to test a priori hypotheses (Unwin and Unwin 1998; Fotheringham et al. 2000). It also suggests the techniques are not rooted fully in the positivist school of thought where the search for global models and ‘laws’ is important. However, this issue is not as clear-cut as it might seem because local statistics can also play an important role in confirmatory analyses as well as in building more accurate global models, a point expanded upon below.

The extent to which global estimates of relationships can present very misleading interpretations of local relationships is shown in Figure 1.3, a spatial example of Simpson’s Paradox (Simpson 1951). Simpson’s paradox refers to the reversal of results when groups of data are analysed separately and then combined. In the spatial example presented in Figure 1.3, data are plotted showing the relationship between the price of a house and the population density of the area in which the house is located. In Figure 1.3(a) data from more than one location are aggregated...
and the relationship, shown by the included linear regression line, is a positive one which suggests that house prices rise with increasing population density. However, in Figure 1.3(b) the data are separated by location and in both locations the relationship between house price and population density is a negative one. That is, for both individual locations, there is a negative relationship between house price and density but when the data from the two locations are aggregated, the relationship appears to be a positive one. Simpson’s Paradox highlights the dangers of analysing aggregate data sets. Whilst it is normally demonstrated in aspatial data sets where the aggregation is over population subgroups, the paradox applies equally to spatial data where the aggregation is over locations.
1.4 Spatial Non-stationarity

Social scientists have long been faced with a difficult question and a potential dilemma: are there any 'laws' that govern social processes, and if there are not, does a quantitative approach have any validity? The problem is more clearly seen as two sub-problems. The first is that models in social sciences are not perfectly accurate. There is always some degree of error (sometimes quite large) indicating that a model has not captured fully the process it is being used to examine. We continually strive to produce more accurate models but the goal of a perfect model is elusive. The second is that the results derived from one system can rarely, if ever, be replicated exactly in another. An explanatory variable might be highly relevant in one application but seemingly irrelevant in another; parameters describing the same relationship might be negative in some applications but positive in others; and the same model might replicate data accurately in one system but not in another. These issues set social science apart from other sciences where the goal of attaining a global statement of relationships is a more realistic one. Physical processes tend to be stationary whereas social processes are often not. For instance, in physics, the famous relationship relating energy and mass, \( E = mc^2 \), is held to be the same no matter where the measurement takes place: there is not a separate relationship depending on which country or city you are in.\(^8\) Social processes, on the other hand, appear to be non-stationary: the measurement of a relationship depends in part on where the measurement is taken. In the case of spatial processes, we refer to this as spatial non-stationarity. In essence, the process we are trying to investigate might not be constant over space. Clearly, any relationship that is not stationary over space will not be represented particularly well by a global statistic and, indeed, this global value may be very misleading locally. It is therefore useful to speculate on why relationships might vary over space; in the absence of a reason to suspect that they do vary, there is little or no need to develop local statistical methods.

There are several reasons why we might expect measurements of relationships to vary over space. An obvious one relates to sampling variation. Suppose we were to take spatial subsets of a data set and then calibrate a model separately with each of the subsets. We would not expect the parameter estimates obtained in such calibrations to be exactly the same: variations would exist because of the different samples of data used. This variation is relatively uninteresting in that it relates to a statistical artefact and not to any underlying spatial process, but it does need to be accounted for in order to identify more substantive causes of spatial non-stationarity.

A second possible cause of observed spatial non-stationarity in relationships is that, for whatever reasons, some relationships are intrinsically different across space. Perhaps, for example, there are spatial variations in people’s attitudes or preferences or there are different administrative, political or other contextual issues.

\(^8\) Even with this equation there is a controversy over whether the speed of light is actually a constant everywhere. However, the argument is only about extreme conditions not met in any practical circumstances and the argument has far from universal acceptance.
that produce different responses to the same stimuli over space. Contextual effects appear to be well documented, for example, in studies of voting behaviour as evidenced by, *inter alia*, Cox (1969); Agnew (1996) and Pattie and Johnston (2000). The idea that human behaviour can vary intrinsically over space is consistent with post-modernist beliefs on the importance of place and locality as frames for understanding such behaviour (Thrift 1983). Within this framework the identification of local variations in relationships would be a useful precursor to more intensive studies that might highlight why such differences occur.

A third possible cause of observed spatial non-stationarity is that the model from which the relationships are estimated is a gross misspecification of reality and that one or more relevant variables are either omitted from the model or are represented by an incorrect functional form. This view, more in line with the positivist school of thought and very much in line with that in econometrics, runs counter to that discussed above: it assumes a global statement of behaviour can be made but that the structure of the model is not sufficiently well formed to allow this global statement to be made. Within this framework, mapping local statistics is useful in order to understand more clearly the nature of the model misspecification. The spatial pattern of the measured relationship can provide a good clue as to what attribute(s) might have been omitted from the model and what might therefore be added to the global model to improve its accuracy. For example, if the local parameter estimates for a particular relationship tend to have different signs for rural and urban areas, this would suggest the addition of some variable denoting the ‘urban-ness’ or the ‘rural-ness’ of an area. In this sense, local analysis can be seen as a model-building procedure in which the ultimate goal is to produce a global model that exhibits no significant spatial non-stationarity. In such instances, the role of local modelling is essentially that of a diagnostic tool which is used to indicate a problem with the global model; only when there is no significant spatial variation in measured relationships can the global model be accepted.

Alternatively, it might not be possible to reduce or remove the misspecification problem with the global model by the addition of one or more variables: for example, it might be impossible to collect data on such variables. In such a case, local modelling then serves the purpose of allowing these otherwise omitted effects to be included in the model through locally varying parameter estimates.

The above discussion on the possible causes of spatial non-stationarity raises an interesting and, as yet unsolved puzzle in spatial analysis. If we do observe spatial variations in relationships, are they due simply to model misspecification or are they due to intrinsically different local spatial behaviour? In a nutshell, can all contextual effects be removed by a better specification of our models (Hauser 1970; Casetti 1997)? Is the role of place simply a surrogate for individual-level effects which we cannot recognise or measure? If the nature of the misspecification could be identified and corrected, would the local variations in relationships disappear? We can only speculate on whether, if one were to achieve such a state, all significant spatial variations in local relationships would be eliminated (see also Jones and Hanham 1995 for a useful discussion on this and the role of local analysis in both realist and positivist schools of thought). We can never be completely confident that our models are correct specifications of reality because of our lack of
theoretical understanding of the processes governing human spatial behaviour. In some ways, this is a chicken-and-egg dilemma. We can never completely test theories of spatial behaviour because of model misspecification, but model misspecification is the product of inadequate spatial theory.

However, the picture is not so bleak: in specific applications of any form of spatial model, we can ask whether the current form of the model we are using produces significant local variations in any of the relationships in which we are interested. If the answer is ‘yes’, then an examination of the nature of the spatial variation can suggest to us a more accurate model specification or the nature of some intrinsic variation in spatial behaviour. In either case, our knowledge of the system under investigation will be improved, in some cases dramatically.

Given the potential importance of local statistics and local models to the understanding of spatial processes, it is surprising that local forms of spatial analysis are not more frequently encountered. However, there have been some notable contributions to the literature on spatially varying parameter models that we now describe. These developments can be divided into three categories: those that are focussed on local statistics for univariate spatial data, including the analysis of point patterns; those that are focussed on more complex multivariate spatial data; and those that are focussed on spatial patterns of movement. We now describe some of the literature on local models and local statistics prior to a full description in Chapter 2 of one local modelling technique, Geographically Weighted Regression, that forms the focus of this book.

1.5 Examples of Local Univariate Methods for Spatial Data Analysis

Four types of local univariate analysis for spatial data can be identified. These are: local forms of point pattern analysis; local graphical analysis; local filters; and local measures of spatial dependency.

1.5.1 Local Forms of Point Pattern Analysis

Many data, such as the locations of various facilities, or the incidence of a particular disease, consist of a set of geocoded points that make up a spatial point pattern. The analysis of spatial point patterns has long been an important concern in geographical enquiry (inter alia, Getis and Boots 1978; Boots and Getis 1988). Traditionally, most methods of spatial point pattern analysis, such as quadrat analysis and neighbour statistics, have involved the calculation of a global statistic that describes aspects of the whole point pattern (inter alia Dacey 1960; King 1961; Tinkler 1971; Boots and Getis 1988). From this global analysis, a judgement would be reached as to whether the overall pattern of points was clustered, dispersed or random. Clearly, such analyses are potentially flawed because interesting spatial variations in the point pattern might be subsumed in the calculation of the average or global statistic. In many instances, particularly in the study of disease, such an approach would appear to be contrary to the purpose of the study, which is to
identify any interesting local clusters of disease incidence (see, for example, Lin and Zeng 1999). Typically, we are not interested in some general statistic referring to the whole point pattern: it is more useful to be able to identify particular parts of the study region in which there is a raised incidence of the disease. Consequently, there has been a growing interest in developing local forms of point pattern analysis.

One of the first of these was the Geographical Analysis Machine (GAM) developed by Openshaw et al. (1987) and updated by Fotheringham and Zhan (1996). As Fotheringham and Brunsdon (1999) note, the basic components of a GAM are:

1. a method for defining sub-regions of the data;
2. a means of describing the point pattern within each of these sub-regions;
3. a procedure for assessing the statistical significance of the observed point pattern within each sub-region considered independently of the rest of the data;
4. a procedure for displaying the sub-regions in which there are significant patterns as defined in 3.

The basic idea outlined in Fotheringham and Zhan (1996) demonstrates the emphasis of this type of technique on identifying interesting local parts of the data set rather than simply providing a global average statistic. Within the study region containing a spatial point pattern, random selection is made initially of a location, and then of a radius of a circle to be centred at that location. Within this random circle, the number of points is counted and this observed value compared with an expected value based on an assumption about the process generating the point pattern (usually that it is random). The population-at-risk within each circle is then used as a basis for generating an expected number of points which is compared to the observed number. The circle can then be drawn on a map if it contains a statistically interesting count (that is, a much higher or lower observed count of points than expected). The process is repeated many times so that a map is produced which contains a set of circles centred on parts of the region where interesting clusters of points appear to be located. The GAM and similar statistics are a subset of a much broader class of statistics known as ‘Scan Statistics’ of which there are several notable spatial applications, particularly in the identification of disease clusters (inter alia Kulldorf and Nagarwalla 1995; Hjalmars et al. 1996; Kulldorf 1997; Kulldorf et al. 1997; Gangnon and Clayton 2001).9

1.5.2 Local Graphical Analysis

One of the by-products of the enormous increases in computer power that have taken place is the rise of techniques for visualising data (Fotheringham, 1999a; Fotheringham et al. 2000, Chapter 4). Within spatial data analysis, exploratory

---

9 At the time of writing, software for calculating spatial, temporal and space-time scan statistics can be downloaded from http://dcp.nci.nih.gov/bb/SaTScan.html
graphical techniques which emphasise the local nature of relationships have become popular. For example, using software such as MANET (Unwin et al. 1996), or XLispstat (Tierney 1990; Brunsdon and Charlton 1996), it is possible to link maps of spatial data with other non-cartographical representations (such as scatterplots or dotplots). Selecting an object on one representation highlights the corresponding object on the other (an early example of this is Monmonier 1969). For example, if a scatterplot reveals a number of outlying observations, selecting these points will highlight their locations on a map. Similarly, selecting a set of points or zones on a map will highlight the corresponding points on a scatterplot. In this way, the spatial distribution of an attribute for a locally selected region can be compared to the distribution of the same attribute across the study area as a whole. Using techniques of this sort, combined with a degree of numerical pre-processing, it is possible to carry out a wide range of exploratory tasks on spatial data which are essentially local. For example, one can identify local clusters in data and investigate whether these are also associated with spatial clusters. Equally, one can also identify spatial outliers, cases that are locally unusual even if not atypical for the data set as a whole. More complex graphical techniques for depicting local relationships in univariate data sets include the spatially lagged scatterplot (Cressie 1984), the variogram cloud plot (Haslett et al. 1991) and the Moran scatterplot (Anselin 1996).

1.5.3 Local Filters

A number of techniques exist in image-processing that can be considered as ‘local’. The data for an image is usually presented as a regular array of intensity values each value referring to a single cell of known area (or a pixel). In order to determine which pixels are likely to represent edges in the image, a high-pass filter can be applied; this acts to increase high-intensity values, and decrease low ones. To remove isolated high values, a low-pass filter can be employed; its action is to make the values in nearby pixels more similar. Other filters may be applied to enhance the values of linear objects in the image; these are known as directional filters. Such filters are usually a square array of weights, often $3 \times 3$ pixels. The output from a filter is a weighted mean value of the pixel at its centre and its immediate neighbours; the filter is applied to each pixel in the input image to produce an output image. The reader is referred to Lilliesand and Kiefer (1995) for further information on the use of filters in image processing.

These filtering techniques have also been applied to raster GIS data (i.e., data stored as a regular lattice). Tomlin (1990) proposed a wide variety of functions that can be applied to local neighbourhoods in such data. Examples of these include the ‘focalmean’, the ‘focalmedian’ and the ‘focalvariety’. The focalmean function provides a weighted mean of the values in the raster which are immediate neighbours of the central one; in this way both high-pass and low-pass filters can be applied to raster GIS data. The focalmedian will return the median of the nine values in the surrounding $3 \times 3$ matrix (in some implementations the filter size can be varied). If the values in the raster are categorical (for example, they may represent land uses), then focalvariety will count the number of different values in the $3 \times 3$ matrix.
Some early examples of the use of filters for spatial analysis are contained in Schmid and MacCannell (1955) and Unwin (1981). More sophisticated examples are given by Cheng et al. (1996) who use variable window sizes and shapes for the local filtering of geochemical images. Rushton et al. (1995) apply a spatial filter to student enrolment projections. A similar technique, popular in fields such as geodesy and meteorology, is that of optimal interpolation in which data weighted by spatial proximity are used to estimate unknown values (Liu and Gauthier 1990; Daley 1991, Reynolds and Smith 1995). The technique is also known as objective analysis (Cressman 1959).

1.5.4 Local Measures of Spatial Dependency

Spatial dependency is the extent to which the value of an attribute in one location depends on the values of the attribute in nearby locations. Although statistics for measuring the degree of spatial dependency in a data set have been formulated for almost three decades (inter alia Cliff and Ord 1972; Haining 1979), until very recently these statistics were only applied globally. Typically a single statistical measure is calculated which describes an overall degree of spatial dependency across the whole data set. Recently, however, local statistics for this purpose have been developed by Getis and Ord (1992), Ord and Getis (1995; 2001), Anselin (1995; 1998) and Rogerson (1999). Getis and Ord (1992), for example, develop a global measure of spatial association inherent within a data set that measures the way in which values of an attribute are clustered in space. A local variation of this global statistic is then formulated to depict trends in the data around each point in space. There are two variants of this localised value depending on whether or not the calculation includes the point \( i \), around which the clustering is measured, although both are equivalent to spatially moving averages (Ord and Getis 2001). The local spatial association statistic allows that different trends in the distribution of one variable might exist over space. In some parts of the study area, for example, high values might be clustered; in other parts there might be a mix of high and low values. Such differences would not be apparent in the calculation of a single global statistic. In their empirical example, Getis and Ord (1992) find several significant local clusters of sudden infant death syndrome in North Carolina although the global statistic fails to identify any significant clustering.

Another local statistic for measuring spatial dependency is a local variant of the classic measure of spatial autocorrelation, Moran’s \( I \) (Anselin 1995). When spatial data are distributed so that high values are generally located in close proximity to other high values and low values are generally located near to other low values, the data are said to exhibit positive spatial autocorrelation. When the data are

---

10 At the time of writing, details of the application of spatial filters to health data, plus a downloadable copy of software for this purpose, DMAP, are provided by Rushton and his colleagues at http://www.uiowa.edu/%7Egeog/health/index11.html
distributed such that high and low values are generally located near each other, the data are said to exhibit negative spatial autocorrelation. However, it is possible that within the same data set, different degrees of spatial autocorrelation could be present; both positive and negative spatial autocorrelation could even exist within the same data set. Global measures of spatial autocorrelation would fail to pick up these different degrees of spatial dependency within the data. A global statistic might therefore misleadingly indicate that there is no spatial autocorrelation in a data set, when in fact there is strong positive autocorrelation in one part of the region and strong negative autocorrelation in another. The development of a localised version of spatial autocorrelation allows spatial variations in the spatial arrangement of data to be examined. Anselin (1995) presents an application of the localised Moran’s I statistic to the spatial distribution of conflict in Africa and Sokal et al. (1998) demonstrate its use on a set of simulated data sets. Other studies of local Moran’s I include those of Bao and Henry (1996), Tiefelsdorf and Boots (1997), and Tiefelsdorf (1998). Rosenberg (2000) provides a partially local measure of spatial autocorrelation through a directionally varying Moran’s I coefficient and Brunsdon et al. (1998) describe a different method of estimating local spatial autocorrelation through Geographically Weighted Regression.

Finally, Rogerson (1999) derives a local version of the chi-square goodness-of-fit test and applies this to the problem of identifying relevant spatial clustering. This local statistic is related to Oden’s (1995) modification of Moran’s I that accounts for spatial variations in population density and is a special case of a test suggested by Tango (1995). The local statistic incorporates a spatially weighted measure of the degree of dissimilarity across regions.

1.6 Examples of Local Multivariate Methods for Spatial Data Analysis

The local univariate statistical methods described above are of limited use in the large and complex spatial data sets that are increasingly available. There is a need to understand local variations in more complex multivariate relationships (see, for example, the attempts by Ver Hoef and Cressie, 1993 and Majure and Cressie, 1997 to extend some of the local visual techniques described above to the multivariate case). Consequently, several attempts have been made to produce localised versions of traditionally global multivariate techniques. Perhaps the greatest challenge, given its widespread use, has been to produce local versions of regression analysis. The subject matter of this book, Geographically Weighted Regression, is one response to this challenge but there have been others. Here we describe five of these: the spatial expansion method; spatially adaptive filtering; multilevel modelling; random coefficient models; and spatial regression models. We leave the description of GWR to Chapter 2. Each of the five techniques described below has limited application to the analysis of spatially non-stationary multivariate relationships for reasons we now explain.