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.