Area (2000) 32.1, 71–78

Spatial patterns in breast cancer incidence in north-west Lancashire

Janette E Rigby* and Anthony C Gatrell†

*School of Geographical Sciences, University of Bristol, University Road, Bristol BS8 1SS.
Email: jan.rigby@bristol.ac.uk
†Institute for Health Research, Lancaster University, Lancaster LA1 4YT.
Email: a.gatrell@lancs.ac.uk

Revised manuscript received 10 March 1999.

Summary *Breast cancer is a disease whose incidence is increasing in both developed and developing countries, but whose complex aetiology is not clearly understood. Recent research suggests that the environment may be an important factor, hence an investigation into spatial patterning of incidence could inform such research. We use data on incidence in north-west Lancashire and apply some techniques for exploratory spatial analysis, at a variety of spatial scales. Issues relating to the use of incidence data and the interpretation of results are discussed.*

Introduction

Breast cancer is a disease with high incidence rates in developed Western countries. In the United Kingdom, one woman in twelve will suffer from the disease at some stage in her life. The disease has a complex aetiology, and although some risk factors have been established, largely relating to age and family/reproductive histories, these factors explain less than half of the incidence of the disease (Madigan *et al* 1995). Considerable spatial variations in incidence and mortality are apparent (Swerdlow and dos Santos Silva 1993): for example, incidence rates are as high as 88·9 per 100 000 women in the United States, compared with 21·2 in China (Parkin *et al* 1992). In addition, where migrants from countries such as Japan, which have traditionally low incidence rates, settle in countries with much higher rates, the rates of the migrant groups move up towards those of the country of adoption (Buell 1973), suggesting that environmental influences may be important.

This paper examines the geographical incidence of breast cancer in north-west Lancashire in recent years, to explore whether the geography of the disease might inform the search for explanatory factors. The use of spatial analysis techniques with such data may provide guidance for subsequent investigation (for example, if incidence or mortality rates in certain areas are shown to be particularly high or low: Openshaw 1987; Haining 1998; Kulldorff 1998). The study area is that served by the North Lancashire Breast Screening Unit, which covers the districts of Lancaster and Morecambe, Blackpool, Wyre and Fylde and North Preston.

The data were investigated at a variety of spatial scales. An assessment of spatial patterning over the area as a whole is supported by a search for more localized spatial association, and subsequently for the presence of any localized clusters of the disease. Thus autocorrelation and local association statistics are used to detect any patterns in area data, while postcoded data are analysed as the outcome of a spatial point process, using a recent version of Openshaw's Geographical Analysis Machine (Openshaw and Turton 1998).

Data

Breast cancer data for the years 1982–92 were obtained from the North West Cancer Registry for investigation within an exploratory spatial data analysis framework. The data comprised 3694 cases of female breast cancer, giving a registry identification

ISSN 0004-0894

number, date of diagnosis, age at diagnosis, date of birth, residential postcode at time of diagnosis, occupation and, where applicable, date and cause of death. When the data were grouped by postcode, a few duplicate entries on the register were observed, and rationalized.

The introduction of the National Breast Screening Programme in 1988 resulted in an increase in recorded incidence in the 50–64 age group targeted for screening. Women are called for screening in order of their general practitioners. As these tend to serve small catchment areas, so 'clustering' of cases from 1988 onwards is inevitable. Hence the data analysis was restricted to the years 1982–87, and to women of 45–80 years of age at diagnosis. This gave a total of 1424 cases. The incidence curve for rates of breast cancer initially rises steeply, but then 'flattens out' from the menopausal age group of 45–55 years (Muir and Malhotra 1987). Incidence rates for this study were not further age-standardized, because of the small numbers of cases that would result in each age band.

As a compromise between the requirement for individual data and a need to protect confidentiality, the 'unit postcode' of the residence is used. This has the added advantage that automated procedures are available to convert postcodes to grid references that can be used for mapping the data, yet gives an acceptable approximation. The need for caution with respect to the accuracy of such representations is well established (Gatrell 1989; Gatrell *et al* 1991).

Approaches to spatial analysis

Traditionally, the protection of patient confidentiality has meant that medical data have not been available at an individual level, and have therefore been aggregated, requiring area-based analysis. Techniques for the analyses of such data by particular spatial units can therefore be applied; techniques with both implicit and explicit spatial considerations. However, there are several problems warranting serious consideration when investigating data grouped into areal units, particularly those relating to the modifiability of the areal units (Gehlke and Biehl 1934; Openshaw and Taylor 1979). Recent work is summarized by Green and Flowerdew (1996).

Goodchild notes that spatial autocorrelation,

> can be a descriptive index, measuring aspects of the way things are distributed in space, but at the same time it can be seen as a causal process, measuring the degree of influence exerted by something over its neighbours. (1987, 3)

We are using it in the first sense, as an exploratory tool for simultaneously examining locational and attribute information. Two common measures of spatial autocorrelation, popularized by Cliff and Ord (1973), are indices derived by Geary (1968) and Moran (1948); the latter is more commonly used. If positive spatial autocorrelation is found to be present, this indicates that spatial units (electoral wards) that are similar in location—near to one another—have similar attribute values. If these spatial objects have attribute levels that are more dissimilar than objects further apart, then there is said to be negative spatial autocorrelation. A zero result would indicate that the attribute values are independent of location. However, it should be noted that the scale of analysis can greatly affect the degree of spatial autocorrelation measured.

This is, however, a global or 'whole-map' statistic (Anselin 1996). It is possible that some more localized effects might be apparent, ie that the rates in some specific areas might be influenced by the rates in surrounding areas. Measures of such localized spatial association have been developed by Getis and Ord (1992) and Anselin (1995). The former suggest that:

> when used in conjunction with a statistic such as Moran's they deepen the knowledge of the processes that give rise to spatial association, in that they enable us to detect local 'pockets' of dependence that may not show up when using global statistics. (Getis and Ord 1992, 190)

Following Getis and Ord, there are two statistics that can be calculated, both of which were originally devised to measure the degree of association that results from the concentration of values included within a radius of distance d from the original point. The $Gi(d)$ statistic does not include the point under consideration; the $Gi^*(d)$ statistic does. The statistics measure the concentration (or the lack of concentration) of the sum of values associated with the variable in the region. $Gi^*(d)$ is a proportion of the sum of all those values within the specified distance of the point under consideration. Assuming that $Gi^*(d)$ is normally distributed, the Z value can be calculated and used to test significance. A large, positive Z value means that high values of the variable are within the distance d of the point,

whereas a large negative Z value indicates low values of the variable. However, because multiple comparisons are being made, and as the individual *Gi*(d)* will tend to be correlated, conventional tests for significance are problematic. One solution, though a very conservative one, is to scale the significance level (α) by the number of observations (Anselin 1995; Ord and Getis 1995).

Some unease with areal-based analyses has provided the impetus for obtaining data at an individual (point) level, and developing techniques for the analysis of patterns of these points. The development and application of the analysis of point locations of individuals within an epidemiological framework has been comprehensively reviewed by Gatrell *et al* (1996). The increasing use of computer systems for both the storage and manipulation of large datasets, and the development of software to assist their analysis, has provided the impetus for the further development of statistical methods. Although there exist statistical techniques to establish whether a group of points varies from a random positioning, these are of limited value when exploring human populations, which are not distributed homogeneously.

There are two fundamental questions that aspects of point pattern analysis can address. One seeks to establish whether there are specific clusters, ie raised incidence of events. Related to this are 'focused' tests of whether incidence is raised near a suspected source of pollution such as a waste incinerator or a chemical works. The second is to examine whether there is clustering over the region of interest. The nature of the information on the underlying 'population at risk' is important to the analysis: this may consist of a basic population count for an administrative area that is then attached to a somewhat arbitrary 'centroid' for that area, or may be in the form of residential addresses of individual 'controls' who are considered free of the disease under investigation.

This approach is also problematic. The point representing the location of an individual is generally taken to be his/her place of residence at the time of diagnosis. This may be helpful where, for example, the spread of an infectious disease such as influenza is under investigation, but can be misleading where an individual has moved residence within the time period between the onset of a disease and its diagnosis (Bentham 1988; Löytönen 1998). Some studies exclude cases based on length of residence, for example a study of breast cancer clustering in West Islip, Long Island New York excluded cases who had lived at their current addresses for less than 30 years (Timander and McLafferty 1998). However, this is a somewhat arbitrary limit, and therefore has a considerable effect on the completeness of the dataset.

One of the first inherently spatial approaches to clustering was devised by Openshaw *et al* (1987) and termed a Geographical Analysis Machine (GAM). This was primarily designed as a:

> descriptive technique designed for an exploratory purpose, that is, to identify areas of interest where further work will be necessary to either validate the findings or to test more specific hypotheses. (Openshaw *et al* 1987, 343)

The technique involved superimposing a lattice onto a study area that contained the location of cases, and information on the underlying population at risk. Circles were generated around the points of the lattice, and where an excess of cases occurred within the circle, its outline was drawn. By repeating this process across the study area, using circles of varying radii, visual inspection of the results could then identify areas of excess that warranted subsequent investigation.

A prominent application of GAM was an exploration of cases of childhood cancer in the North of England, which arose following concern over an apparent excess of cases of childhood leukaemia in the proximity of nuclear power installations. The excess was particularly apparent in terms of leukaemia rates in young children in the nearby coastal village of Seascale: for the period 1956–80, five cases were reported where the expected number was 0·45 (Urquhart *et al* 1984).

A comparative approach for evaluating the 'efficiency' of GAM and related techniques was initiated by Alexander and Boyle (1996). Simulated datasets were designed to represent a range of clustering scenarios, and the authors of a number of established clustering techniques were invited to apply their techniques to the datasets and provide accounts of how the analyses were performed, along with the conclusions reached. Alexander and Boyle subsequently commented on the findings. In addition to the correct identification of actual clusters, an important consideration is that 'false positives' should not emerge, ie a technique should not identify clusters that are not actually present in the data. As Alexander and Boyle note, 'the adverse social consequences of a false positive

cluster identification can be high, and must always be borne in mind' (1996, 155). GAM-K, a variant of GAM which creates a smoothed density surface of excess incidence for the significant circles (Openshaw and Turton 1998), performed well across the range of datasets, as did the technique of Besag and Newell (1991)—though the latter recorded more false positives.

Results

We first analyse the data at the scale of local authority wards. Incidence rates were calculated with the number of cases per ward as numerator, and the population 'at risk' (ie women in the same age range) as denominator. The latter was obtained from the 1991 Census data. Data from the 1981 Census might be considered to be better, but the advantage of 1991 was that digital boundary data at both ward and enumeration district level were readily available, and subsequent administrative boundary changes meant that data for 1981 would have been incompatible with 1991 boundaries. It was thought unlikely that major changes in the population at risk would have occurred over this time period.

Visual inspection of the incidence rates by ward (Figure 1) revealed little obvious geographic patterning, though a possible urban/rural variation was noted. However, formal tests of spatial autocorrelation have been carried out. The Moran *I* Statistic was estimated using the SpaceStat software package (Anselin 1992). The measure of spatial proximity applied was whether wards were adjacent, ie shared a common boundary. From this, SpaceStat generated a binary contiguity matrix for the wards. The result from this 'global' measure over the entire study area of 119 wards gave a value for Moran's *I* of 0·101, with a probability of 0·035, and hence there is some evidence of positive spatial autocorrelation, significant at the 95 per cent level.

To investigate more localized association, adjacency of the spatial units was again considered to be a better measure than distance, since a fixed distance (eg 30 kilometres) might prove quite adequate for rural areas, but not sufficiently sensitive for groups of urban areas. The *G* statistics could also be applied using the software SpaceStat, calculating the *Gi** statistic from the binary contiguity matrix developed for the spatial autocorrelation work, with the attributes being the incidence rates for the wards.

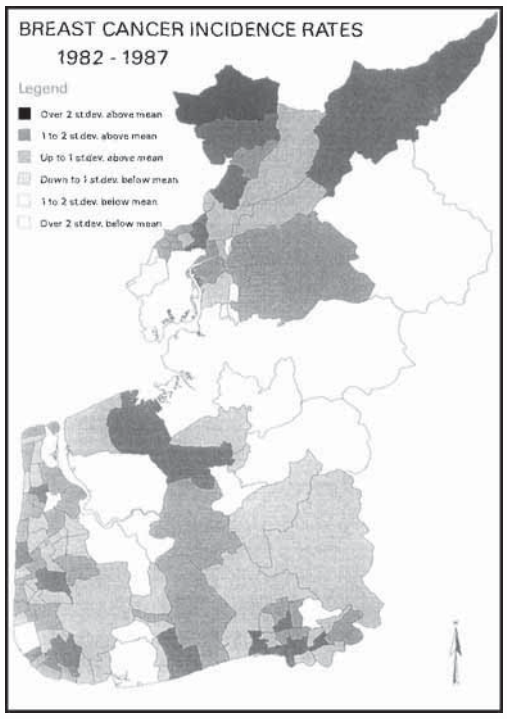

Figure 1 Incidence of breast cancer in the electoral wards of north-west Lancashire

The mapped results showing wards with significant local association statistics (Figure 2) are of some interest, although they should be regarded as indicative only, as the significance levels are unadjusted for multiple comparisons. The wards of 'high, local effect' to the north denote small areas of high incidence that tend to be surrounded by areas of similar value, while those of 'low, local effect' represent areas of low incidence adjacent to other low incidence wards. These suggest it might be appropriate to explore a more extensive study region, to see if these low rates to the east (and the high rates to the north) continue beyond the study region. The application of this technique can thus demonstrate that there is local spatial association for both high and low incidence rates of the disease, and this merits further investigation in terms of possible causal factors; had we been exploring mortality data there might have been implications in terms of access to health care.

It is important to note that, at ward level, data have been substantially aggregated over areas that might themselves contain considerable variations. Hence

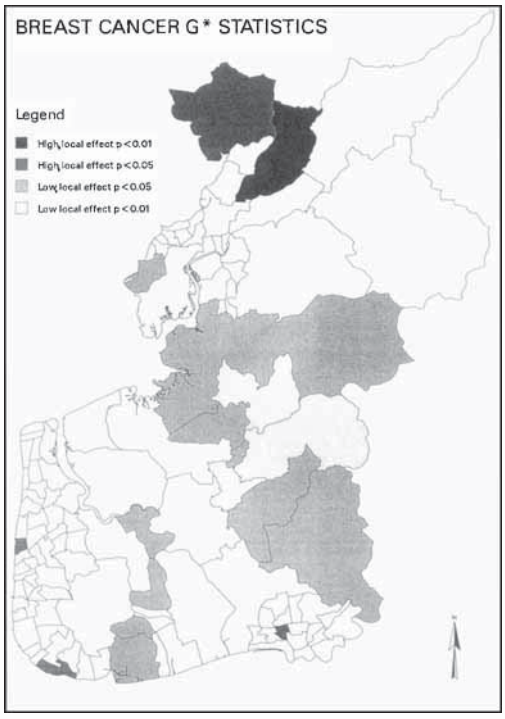

Figure 2 Significant local association measures of high and low breast cancer incidence in north-west Lancashire

more detailed inspection is warranted at the Enumeration District (ED) scale. Electoral wards are subdivided into enumeration districts, which comprise aggregations of around 130 households. It was anticipated that a reduction in the size of the area examined would lead to less variation within an ED than within a ward, ie an assumption that EDs are more homogeneous than wards. The raw data were therefore reaggregated to ED level, although in the study area there remained variability in the population density in the EDs from high-density housing in urban areas to hill-farming in the fells to the east.

One problem that can often occur when working with relatively small areas is that the number of cases for each area is itself small. It may therefore be that areas where rates are 'extreme' on the map are, in fact, those with the most extreme sampling error, based on small numerators and/or denominators (Kennedy-Kalaftis 1995). As statistical significance is directly related to sample size (Gardner 1989), areas with relatively large populations will reflect this.

A preferred approach is to map the data after adjusting for the reliability of the estimate as it changes across the map surface. For this, Bayesian statistical inference can be applied (Clayton and Kaldor 1987; Clayton and Bernardinelli 1992). This allows for the inclusion of prior information concerning the data, notably the distribution of relative risks between areas (Langford 1994). Empirical Bayes estimates were generated for the EDs using the MINITAB macro developed by Langford. As there were 1302 EDs in comparison with 119 wards, the study area was subdivided for visual inspection, with the EDs with highest and lowest values being emphasized. However, it can be seen from the example in Figure 3, which focuses on the north of the study area where the electoral ward rates were highest, that no consistent spatial patterning is apparent.

Individual (point) level

From the findings in Alexander and Boyle, the GAM-K approach appeared to offer high reliability in terms of both accurate cluster identification and low numbers of false positives. Openshaw and Turton (1998) have made the method available over the internet, so it was possible for us to access the software based at Leeds, supply it with data, process the data remotely, and receive the results back via the internet. Considerable efforts had been made by the Leeds group to make the interface as user-friendly as possible, and only minor modifications were required at the Leeds source to produce a successful run on the Cancer Registry data.

There tend to be two approaches in cluster analysis of this sort. One compares the distribution of the cases against a series of population controls; the alternative is to use an underlying population count for a small area. For the Cancer Registry dataset, individual controls were not available, which restricted the choice of method. The problem is not insurmountable, for example, Gatrell (1995) generated controls by stratified sampling from all possible residential postcodes for the study area. Whilst this has the advantage of producing appropriate postcodes for 'surrogate' controls, the data are obviously artificial, and hence problematic.

Underlying population counts were available for the registry data, although this again raises the modifiable areal unit problem. Population-at-risk counts for enumeration districts, and cases aggregated to ED level had already been established for the small area work, and these were used for the GAM-K analysis.

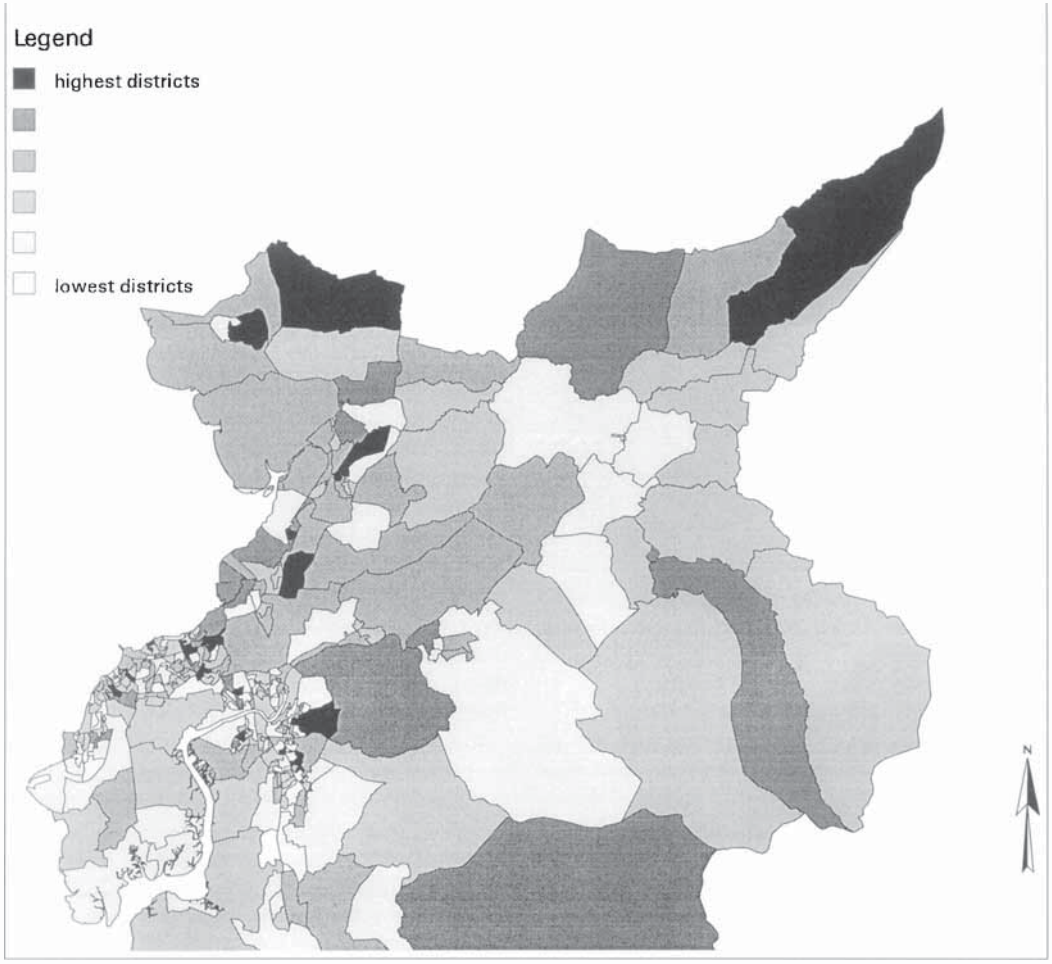

Figure 3 Empirical Bayes estimates for breast cancer incidence at Enumeration District level in north Lancashire

The results can be seen in Figure 4a. One drawback with this representation is that it is not visually apparent how 'strong' the clustering is; to assist with this, output can be viewed three-dimensionally (Figure 4b). Results indicate that the clusters on the Fylde coast (those with the highest peaks) are possibly the most interesting. The results of such an exploratory approach are designed for visual inspection to suggest small areas for further investigation; they are not intended as confirmatory. Where clusters are apparent, it is possible to investigate small areas in considerable detail. As the cancer data are postcoded and converted to grid references, these can be plotted as points on maps of, for example, the road network. Care must be taken to ensure that duplicated grid references are offset slightly so that each case is visible on the map. However, a major consideration is patient confidentiality, and the results are not presented here.

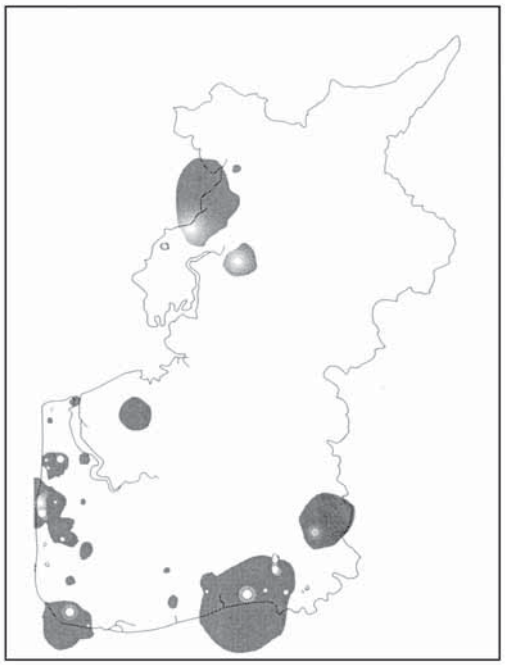

Figure 4a Clusters of high breast cancer incidence from a GAM-K search

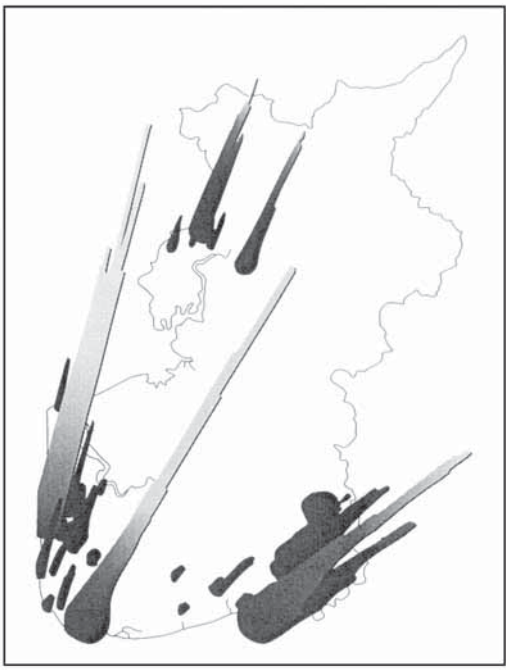

Figure 4b The relative intensity of the clusters indentified by the GAM-K search

Conclusion

This paper has demonstrated some techniques that can reasonably be applied to a dataset where there is an expectation that any spatial patterning that proves apparent might inform the aetiology of the disease. A major drawback is the lack of a single software package to perform all aspects of this analysis, and therefore work of this nature constantly necessitates the transferral of data in a variety of different formats. Whilst considerable progress has been made by, for example, Bailey and Gatrell (1995), software capable of comprehensive investigation of large-scale datasets, comprising several thousand records, remains on the agenda.

The results from this exploratory approach show no obvious, consistent spatial patterning over the study region. However, the local association measure has identified an area of low incidence in the south-east of the region, and the output from the cluster analysis has indicated a small number of areas for detailed study. It is important that the results of such analyses are carefully interpreted in the context of issues such as the modifiable areal unit problem; these issues are increasingly addressed by the geographic community, but have yet to permeate fully the range of other professionals who have an interest in patterns of disease.

Acknowledgements

This work forms part of a study funded by the Lancashire and Cumbria Foundation for Medical Research. We are grateful to the North West Cancer Registry for the incidence data on breast cancer, and to the ESRC Data Archive based with MIDAS, at the Manchester Computer Centre, for the boundary data. Special thanks to Adrian Maddocks and Chris Beacock for technical support, and to Robin Flowerdew and two anonymous referees for their helpful comments on an earlier version of this paper.

References

Alexander F E and Boyle P (1996) *Methods for investigating localized clustering of disease* IARC Scientific Publication No 135 (IARC, Lyon)

Anselin L (1992) *SpaceStat* (NCGIA, Santa Barbara, CA)

— (1995) 'Local indicators of spatial association' *Geographical Analysis* 27, 93–115

— (1996) 'The Moran scatterplot as an ESDA tool to assess local instability in spatial association' in Fischer M, Scholten H J and Unwin D (eds) *Spatial analytical perspectives on GIS* (Taylor and Francis, London), 111–25

Bailey T C and Gatrell A C (1995) *Interactive spatial data analysis* (Longman, Harlow)

Bentham G (1988) 'Migration and morbidity: implications for geographical studies of disease' *Social Science and Medicine* 26, 49–54

Besag J and Newell J (1991) 'The detection of clusters in rare diseases' *Journal of the Royal Statistical Society Series A* 154, 143–55

Buell P (1973) 'Changing incidence of breast cancer in Japanese-American women' *Journal of the National Cancer Institute* 51, 1479–83

Clayton D and Bernardinelli L (1992) 'Bayesian methods for mapping disease risk' in Elliott P, Cuzick J, English D and Stern R (eds) *Geographical and environmental epidemiology: methods for small area studies* (Oxford University Press, Oxford), 205–20

Clayton D and Kaldor J (1987) 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping' *Biometrics* 43, 671–81

Cliff A D and Ord J K (1973) *Spatial autocorrelation* (Pion, London)

Gardner M J (1989) 'Review of reported increases of childhood cancer rates in the vicinity of nuclear installations in the UK' *Journal of the Royal Statistical Society Series A* 152, 307–25

Gatrell A C (1989) 'On the spatial representation and accuracy of address-based data in the United Kingdom' *International Journal of GIS* 3, 335–48

— (1995) 'Spatial point process modelling of cancer data within a geographical information systems framework' in Cliff A D, Gould P R, Hoare A G and Thrift N J (eds) *Diffusing geography* (Blackwell, Oxford), 199–217

Gatrell A C, Bailey T C, Diggle P J and Rowlingson B S (1996) 'Spatial point pattern analysis and its application in geographical epidemiology' *Transactions of the Institute of British Geographers* 21, 256–74

Gatrell A C, Dunn C E and Boyle P J (1991) 'The relative utility of the Central Postcode Directory and Pinpoint Address Code in applications of geographical information systems' *Environment and Planning A* 23, 1447–58

Geary R C (1968) 'The contiguity ratio and statistical mapping' in Berry B J L and Marble D F (eds) *Spatial analysis: a reader in statistical geography* (Prentice Hall, Englewood Cliffs, CA), 461–78

Gehlke C E and Biehl K (1934) 'Certain effects of grouping upon the size of the correlation coefficient in census tract material' *Journal of the American Statistical Association Supplement* 29, 169–70

Getis A and Ord J K (1992) 'The analysis of spatial association by use of distance statistics' *Geographical Analysis* 24, 189–206

Goodchild M (1987) *Spatial autocorrelation* (GeoBooks, Norwich)

Green M and Flowerdew R (1996) 'New evidence on the modifiable areal unit problem' in Longley P and Batty M (eds) *Spatial analysis: modelling in a GIS environment* (GeoInformation International, Cambridge), 41–54

Haining R (1998) 'Spatial statistics and the analysis of health data' in Gatrell A C and Löytönen M (eds) *GIS and health* (Taylor and Francis, London), 29–47

Kennedy-Kalaftis S (1995) 'Reliability-adjusted disease maps' *Social Science and Medicine* 41, 1273–87

Kulldorff M (1998) 'Statistical methods for spatial epidemiology: tests for randomness' in Gatrell A C and Löytönen M (eds) *GIS and health* (Taylor and Francis, London), 49–62

Langford I H (1994) 'Using empirical Bayes estimates in the geographical analysis of disease risk' *Area* 26, 142–9

Löytönen M (1998) 'GIS, time geography and health' in Gatrell A C and Löytönen M (eds) *GIS and health* (Taylor and Francis, London), 97–110

Madigan M P, Ziegler R G, Benichou J, Byrne C and Hoover R N (1995) 'Proportion of breast cancer cases in the US explained by well-established risk factors' *Journal of the National Cancer Institute* 87, 1681–95

Moran P A P (1948) 'The interpretation of statistical maps' *Journal of the Royal Statistical Society Series B* 10, 243–51

Muir C S and Malhotra A (1987) 'Changing patterns of cancer incidence in five continents' in Kurihara M, Aoki K, Miller R W and Muir C S (eds) *Changing cancer patterns and topics in cancer epidemiology* (Plenum Press, New York), 3–23

Openshaw S, Charlton M, Wymer C and Craft A (1987) 'A mark 1 geographical analysis machine for the automated analysis of point data sets' *International Journal of GIS* 1, 335–58

Openshaw S and Taylor P J (1979) 'A million or so correlation coefficients: three experiments on the modifiable areal unit problem' in Wrigley N (ed) *Statistical applications in the social sciences* (Pion, London), 127–44

Openshaw S and Turton I (1998) 'A smart spatial pattern explorer for the geographical analysis of GIS data' (www.ccg.leeds.ac.uk/smart/intro.html) Accessed 10 May

Ord J K and Getis A (1995) 'Local spatial autocorrelation statistics: distributional issues and an application' *Geographical Analysis* 27, 286–306

Parkin D M, Muir C S, Whelan S L, Gao Y-T, Ferlay J and Powell J (eds) (1992) *Cancer incidence in five continents, volume VI* IARC Scientific Publication No 120 (IARC, Lyon)

Swerdlow A and dos Santos Silva I (1993) *Atlas of cancer incidence in England and Wales, 1968–85* (Oxford University Press, Oxford)

Timander L M and McLafferty S (1998) 'Breast cancer in West Islip, NY: a spatial clustering analysis with covariates' *Social Science and Medicine* 46, 1623–36

Urquhart J, Palmer M and Cutler J (1984) 'Cancer in Cumbria: the Windscale connection' *Lancet* i, 217–8