# Building a Health and Environment Geographical Information System: An Evaluation

## Looking at Childhood Cancer in Northern England

Anna. E. Cross BA (Hons)

Thesis submitted for the degree of
Doctorate of Philosophy in the
Department of Geography,
University of Newcastle upon Tyne

September 1991

# ABSTRACT

The aim of this research was to evaluate a relatively young technology, Geographical Information Systems (GIS), in a specific applications environment. The application adopted was that of searching for environmental causes of childhood cancer, in particular that of Acute Lymphoblastic Leukaemia (ALL), in Northern England. It is also relevant in terms of the WHO's intention to develop a Health and Environment GIS, and therefore the research aims to satisfy their recommendations for pilot studies.

The subject matter of this thesis therefore covers two very high profile topics, which it is believed will mutually benefit from the research carried out. Firstly, very little is known about the aetiology of ALL, and thus any new methodology which is introduced to help analyse sensitive issues of causation is welcomed not only by those in the medical field but also the public. The application was made possible with the provision of detailed cancer data for Northern England and a weak but interesting hypothesis that environmental factors may be an attributable mechanism for causation. Key questions which are asked include; Where are incidences of ALL located? Why are they there? Is there a cluster? and What could be the cause?

Secondly a Geographical Information System, in this case the proprietary software package ARC/INFO, was considered an excellent medium for tackling this spatial epidemiological problem. Especially with its capability to store large volumes of diverse data, and its inherent flexibility to deal with spatial information pertaining to health and environmental factors. More importantly the application itself offered a means of evaluating the implementation of a GIS. Establishing the advantages and pitfalls which accompany all stages of 'The GIS Process' and an invaluable documentation of the experiences acquired as an initiator, developer and implementor of this new technology.

In addition, this research offers fresh ideas and techniques for improving those areas of the technology which appear to be lacking in these early phases of its development. The problems of spatial analysis in GIS and the provision of useful tools such as 'pattern spotters', 'relationship seekers' and 'error handlers' are discussed as alternative techniques. To ensure an exciting future for GIS technology in application environments the latter and other key areas of research which should be persued are highlighted in this thesis.

# ACKNOWLEDGEMENTS

There are a number of people who in their own small or huge way have contributed to the completion of this thesis. Not everybody will appear in this short section but whoever they are, they are warmly thanked by default.

Firstly, this thesis would not have been possible without the contribution of Dr Louise Parker and Dr Alan Craft of the Department of Child Health, Newcastle upon Tyne. They trusted me with very sensitive data and were willing to let me, armed with a GIS have a go at tackling a particularly interesting but complex problem of childhood cancer causation. In addition, my supervisor Professor Stan Openshaw played a major part in guiding me along in these last four years. His enthusiasm helped when GIS failed to meet my expectations and it was nice to think someone had faith in my ability to sustain a thesis.

In turn the working environment at Newcastle provided me with an excellent team of consultants, for whom I must owe a few gallons of beer in lieu of payment. They include Dr Christopher Brunsdon who was my statistical advisor and who patiently listened to my problems sometimes two or three times over, but always seemed to fathom out my dilemma and solve it. Daniel Dorling and his Archimedes helped to produce some key illustrations in this thesis. Martin Charlton, who will probably hope that he has heard the last of 'Martinnnnn, why has...?' -- Ha well he hasn't! helped to decipher some of ARC/INFOs unfriendly error messages and a little 'Word' and 'FORTRAN' also went along way. Special thanks goes to Jacqueline Nicol who added the Scottish touch to these 'words of wisdom', in the invaluable, but less than glamorous task of proof reading.

A number of other key members of staff include Dr Steve Carver for letting me 'acquire' some of his environmental databases, Zhilin Li for producing some mathematical equations for the text and David Waugh who helped me out of some FORTRAN snags whilst the others deserted me to go and enjoy themselves in Stockholm. Thanks also goes to Ann Rooke who helped to reproduce a number of diagrams for Chapter 5, and Betty Robson and other members of the CURDS General Office who succeeded in deciphering my writing in order to type up Appendices and the Bibliography.

Finally it is noted that a thesis is not centred purely around the academic arena. I have made a number of friends in my four years at Newcastle, all of which have helped me to carry out this research. Thanks too, to Mum and Dad who didn't always understand what I was going through but frustratingly had faith in my completion whatever the circumstances, xxx. Also to Steve Burn who was very understanding throughout the hectic few weeks that preceeded the submission date. Finally it is noted that the completion of this thesis should, I hope, hear the last of that most irritating of questions 'Hows it going then?!!!'

# CONTENTS

# LIST OF FIGURES

# List of Tables

# SECTION ONE:
# RESEARCH OBJECTIVES AND THE FEASIBILITY STAGE IN IMPLEMENTING A GIS

# CHAPTER 1

# A PILOT STUDY FOR BUILDING A HEALTH AND ENVIRONMENT GIS IN NORTHERN ENGLAND

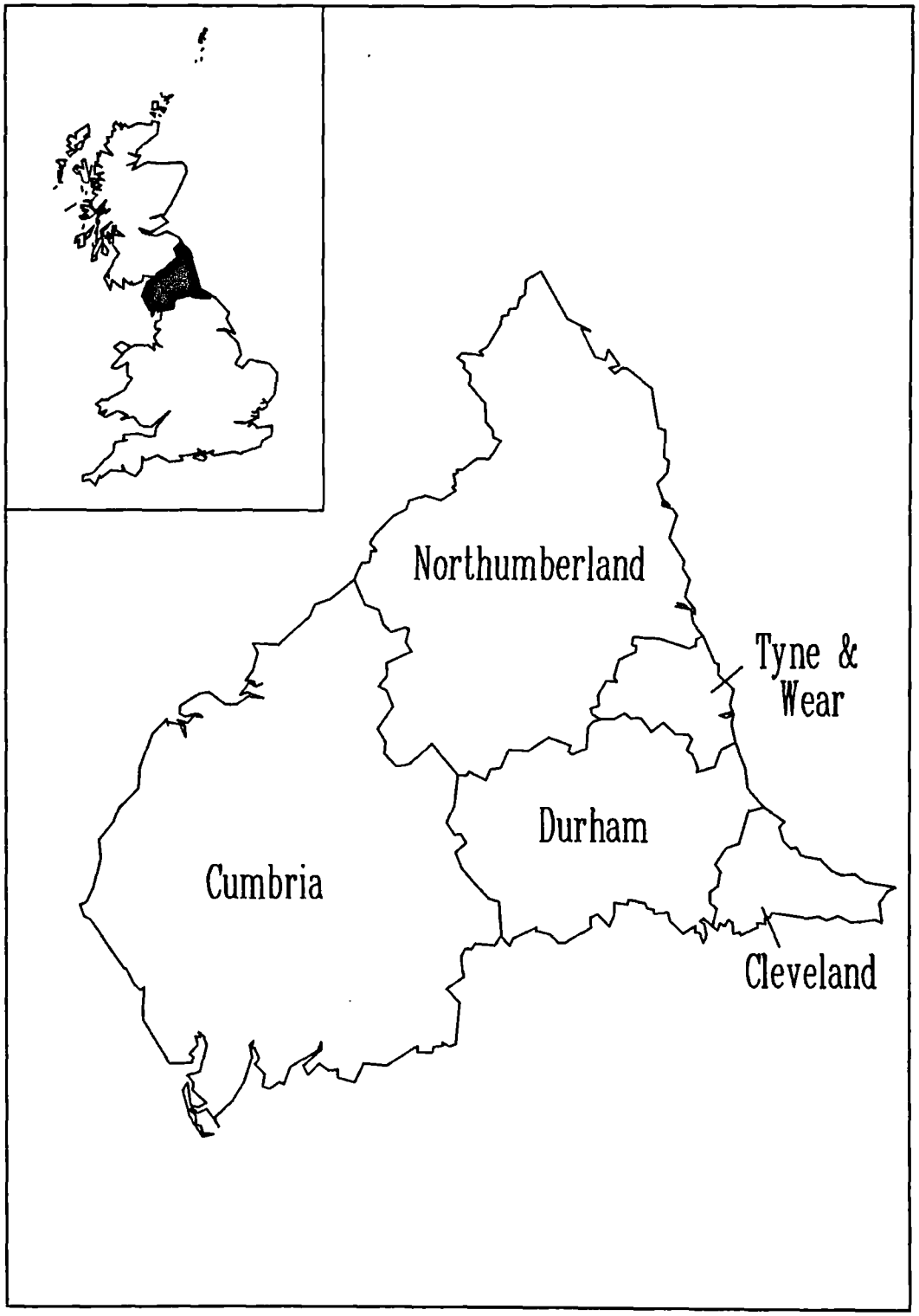
## 1.1 Research Objectives

This research has a dual objective. The first is to evaluate a relatively young technology, Geographical Information Systems (GIS), in a specific applications environment. The second aim is to use this alternative methodology for analysing the possible environmental causes of childhood cancer, in particular Acute Lymphoblastic Leukaemia (ALL) in Northern England, the extent of this study area is shown in Figure 1.1. At present these constitute two high profile topics, and this research will demonstrate how they can mutually benefit from adopting this twofold approach.

The characteristics, distribution and causes of Acute Lymphoblastic Leukaemia have received considerable attention in recent years from a wide range of disciplines, including geography and statistics, which will be discussed further in section 1.3. The main reason for this interest is the lack of knowledge concerning the aetiology of such childhood cancers, ie. the identification of a causal mechanism. Thus any technique and/or technology which can further research in this area of spatial epidemiology is welcomed, not only by those in the medical field but also the general public.

The application was made possible by the provision of a comprehensive and detailed Children's Cancer Registry, produced for the Northern Region at the Department of Child Health, Newcastle upon Tyne. Acute Lymphoblastic Leukaemia is one of the main cancers to be found in the Registry, it is also the one which has been the subject of various forms of contentious analysis. In general, these have tended to emphasise the possible localised nature of incidences, which in turn has lead to a number of weak but interesting hypotheses linking Acute Lymphoblastic Leukaemia to environmental factors. It is these two aspects which are focused upon in more detail in this research.

Figure 1.1: The study area



The distribution of health, or more particularly ill-health, and geographical entities representing elements of the environment both possess a spatial component. Thus with the availability of spatially referenced health data and a spatial epidemiological problem, the new technology of Geographical Information Systems seemed an ideal medium to offer a fresh approach to the complex problem of establishing the possible causes of Acute Lymphoblastic Leukaemia.

There are many definitions of what a Geographical Information System is, in general though it is referred to as;

> 'a system for capturing, storing, checking, integrating, manipulating, analysing and displaying data which are spatially referenced to the Earth' (pp 132, DoE, 1987)

This research will encapsulate every aspect of this statement and will use the spatial epidemiological application as a vehicle for evaluating the advantages, potential and limitations of this new technology in practice. As a result this thesis will form one of the first complete and fully documented examples of a GIS in the application world. Its structure therefore is one of a pilot study for the development and implementation of a Health and Environment Geographical Information System (HEGIS) for Northern England. This approach is not only intentional but appropriate in the light of recent developments in Geographical Information Systems. Both national and international bodies have called for detailed reviews of working systems in order to evaluate their progress and potential. For instance, The Chorley Report (1987) which looked into the handling of spatial data in the United Kingdom (UK) recommended;

> '..that user groups, data and system suppliers, and academics should actively disseminate information, via appropriate media, on the applications and benefits of Geographic Information Systems..' (pp98)

In addition initiatives have been established at the European scale to develop a more comprehensive approach to a Health and Environment Geographical Information System (WHO, 1988). Although on a much larger scale than this research the aims are essentially the same, Appendix A provides a brief history of the European objectives and the timetable for implementation of a European HEGIS. However,

amongst the milestones set at the First European Conference on the Environment and Health (Frankfurt, 1989) it was recognised that;

> 'Pilot studies or demonstration projects are likely to be extremely useful in ensuring the success of HEGIS. Such studies or projects would demonstrate the potential of HEGIS and would allow potential users critically to evaluate opportunities for their further application' (pp14, on Target 19)

Even though this statement was made some two years after this research had commenced and was only formalised when this thesis was being completed, it still serves to aptly summarise the purpose of this study. However after four years experience in developing a prototype HEGIS a number of key factors can be identified which may effect the success of this European approach. In particular the implications of health related systems for the whole of Europe may be far more problematic than that of creating a comprehensive set of environmental databases, which was the responsibility of the Coordination of Environmental Information (CORINE, CEC, 1985) since 1985. This will also be discussed in Chapter 10. In addition the timetable for the implementation of the HEGIS was set to start in 1988 and culminate in a full programme for operation by the end of 1992. Whilst the experience of this research suggests that the framework for a HEGIS can be established in that time. In order to meet the goals of such a large project the programme will require a much longer time period and sustained input from all Member States if it is to succeed.

It is noted that the aims of this research and that of a European HEGIS are similar but that the actual organisation of such a pilot study compared to that of a large scale approach is very different, some of the key factors are summarised in Appendix B. These include the fact that this pilot study will benefit from the researcher having control over each stage of the GIS implementation. The complete responsibility for the project and a clear perspective of application needs should allow a relatively logical route into the building of a GIS for spatial epidemiology. On the other hand an isolated research does not have the advantages of information sharing, and the benefits of a wide range of expertise which could stimulate a wider scope of ideas and further research development. However the aspects that this research does raise are generically applicable and will prove important feedback for both small scale and large scale projects.

This introduction has formalised the general aspects of the research. The remaining sections in this chapter will provide further background to the study. In particular section 1.2 will emphasise the significance of a geographer's contribution to the field of spatial epidemiology. Whilst section 1.3 discusses some of the previous research and public demands to establish new techniques which may provide evidence for the causes of Acute Lymphoblastic Leukaemia. These provide further justification for the adoption of a Geographical Information Systems approach.

## 1.2 Spatial Epidemiology

The intellectual research relationship between geography and medicine is not new, indeed Gould (1985) describes it as the 'old partnership', albeit a slow and not always fruitful one. In the past geographers have tended to only get involved in spatial epidemiology when a disease is at the forefront of investigation, and has had a strong environmental component postulated for its aetiology. This explains to some extent its present popularity, especially as more and more diseases and health related problems are now being ascribed to one or more environmental cause (both physical and social).

The actual term 'spatial epidemiology' serves to combine the study of the determinants and prevalence of disease in a geographical context. It can be tackled in three ways; (a) Descriptive studies, which may involve the observation of disease patterns and their development over time. (b) Analytical studies, which investigate the hypotheses or general ideas suggested by descriptive studies and, (c) Experimental or intervention studies concerned with measuring the effects of potentially harmful environmental influences on the population, or concerned with introducing preventative and/or ameliorative services under controlled conditions (Alderson, 1983).

To some extent the descriptive studies, strategy (a), have already been exploited with respect to childhood leukaemia, and the results have led to the development of several atlases, including Gardner (1983), Kemp (1985) and Alexander (1990). However it also provides the stimuli for an analytical approach (b), which aims to discover possible causative effects, which is the purpose of GIS in this study. It should be noted though that whilst GIS technology may be innovative, its ability to produce

maps for epidemiological research is not new to this field. The use of maps as an aid to the human mind to assimilate and understand the spatial dynamics of disease is in fact well established. The 'Golden Age' of medical geography was said to be between 1835 and 1855, when the much quoted medical scientist Dr John Snow demonstrated that there was a link between the geographical distribution of cholera deaths and contaminated water supplies (1854). This occurred at a time when little or nothing was known about the aetiology of cholera. The power of inference from such map-based analysis was summarised by Petermann (1852) in that;

> 'while symbols of the masses of statistical data....present a uniform appearance, the same data embodied in a map, will convey at once, the relative bearing and proportion of the single data, together with their position, extent and distance'

From these results the final strategy in spatial epidemiology, the experimental approach (c) can be tackled with intervention and the introduction of preventive measures as and when is necessary.

Two aspects have recently strengthened the 'partnership' between geography and medicine. These include the increased availability of data at a fine geographical resolution, discussed further in Chapter 3, and considerable technological developments, described in Chapter 2. Since prior to the 1960's, and the onset of large computational facilities geography was simply too ill-equipped to undertake studies which depended upon 'quantitative and comparative' geographical analysis at scales larger than a single region. As this is no longer a major problem, the question has now become one of 'what can be achieved with the new and ever increasing range of analytical techniques in order to accomplish sensible interpretation and analysis?'

Techniques which have already been employed to study the possible environmental correlates of Acute Lymphoblastic Leukaemia are outlined in the next section. These also provide justification for the development of a GIS platform to tackle this complicated spatial epidemiological problem.

## 1.3 Research Background

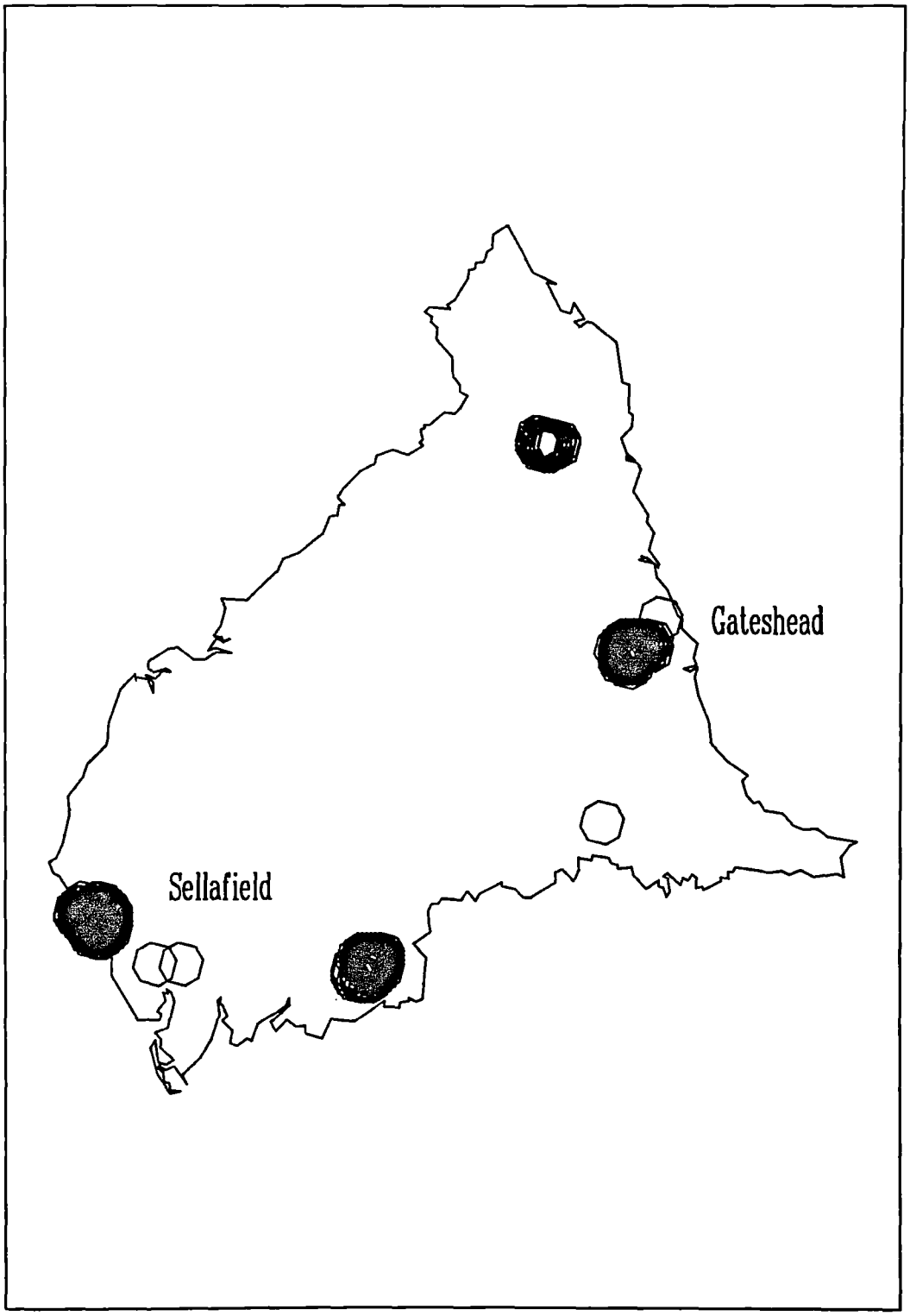
This section takes three areas that have served to highlight key aspects concerning the distribution and understanding of incidences of Acute Lymphoblastic Leukaemia. The first looks at the possible clustering effect of the cases of childhood cancer, employing what has been termed a Geographical Analysis Machine (Openshaw, 1986). Secondly, reference is made to previous research that explored the possible environmental causes of Acute Lymphoblastic Leukaemia cases in the county of Tyne and Wear. Finally, the increasing awareness of the general public on health related issues is noted because they too have served to pressurise change and developments in the search for disease causation.

### 1.3.1 Geographical Analysis Machine

The results from the locational analysis of Acute Lymphoblastic Leukaemia, using a Geographical Analysis Machine served to promote this GIS approach. Figure 1.2 demonstrates some of the early results from GAM, whereby the circles represent an increase in the number of childhood cancer cases in certain areas which were greater than would be expected for the region as a whole. The interesting aspect of these results was the additional cluster picked out in Gateshead (Tyne and Wear), especially in comparison to previous analysis (Black, 1984) which tended to focus on the incidences around Sellafield. This implied that radiation may not be the only cause of Acute Lymphoblastic Leukaemia. It also supported a suspicion by the staff at the Department of Child Health in Newcastle, that a disproportionately large percentage of their cases did come from a particular housing estate, the Leam Lane Estate, in Gateshead. At about the same time Kinlen (1988) was also arguing for non-radiation causes for those cases around Sellafield, discussed in Chapter 3. Thus the result of GAM was to reinforce the need for a more general look at geographical features in the region. In other words the clusters were being put forward as a possible reflection of a single or combination of environmental factors, which could be tackled by a GIS given the relevant databases.

The specific methodology involved in GAM will be expanded upon in Chapter 8, which is concerned with the need for point pattern analysis techniques as part of 'The GIS Process'. However the main aim of GAM was to alleviate previous research shortcomings, by; (i) avoiding bias in the results which may occur due to the

Figure 1.2: Early GAM results involving the cluster analysis of incidences of Acute Lymphoblastic Leukaemia



conscious or accidental selection of time periods, categories of disease groups, age groups etc, ii) preventing the invalidation of hypotheses by prior knowledge of the data, iii) alleviating some of the problems which surround the determination of the significance of results, iv) the measure of associated errors in analysis and finally, v) reducing the heavy reliance on more favourable causes of cancer which can lead to the neglect of other potential environmental factors (Openshaw et al, 1987). This was in direct response to the nature of the Acute Lymphoblastic Leukaemia debate over causation and the problems that behold the use of geographical data in any complex form of spatial analysis, again this is an issue which will be discussed in Chapter 8. Therefore the question this research will explore, is to what extent GIS can help to explain the geographical differences highlighted by GAM.

### 1.3.2 A Previous Tyne and Wear Study

A thesis carried out between 1984-1987, entitled 'Environmental Correlates of Childhood Leukaemia in Tyne and Wear' (Raybould, 1988), involved a superficially similar methodology to that expressed in this particular study. It too attempted to answer some of the various theories that relate to whether a cancer is the result of; the environment of the patient, an infectious agent (which is probably viral), or a genetic predisposition. His approach tended to be extremely isolated and restricted mainly due to the lack of analytical methods available (including GIS) and ultimately the time and man-power that could be sustained by a single thesis. It did not identify any specific cause for childhood leukaemia, but each chapter was considered to provide food for thought with rather weak yet tantalisingly associations being found. Research findings included the inability to establish overhead power lines as a possible source of childhood leukaemia, despite the beliefs of local paediatric oncologists who considered this a likely cause. In addition some weak associations between drinking water quality and heavy metal soil pollution were suggested. Raybould's thesis also served to produce a number of spatial databases specific to Tyne and Wear which could be incorporated into a GIS. As well as detailed information from field studies that could ideally be used to calculate possible areas of impact of certain pollution sources, discussed further in Chapter 4 where the environmental databases employed in this research will be reviewed.

Raybould concluded that new clustering techniques needed to be explored to avoid the speculation that surrounded those observed to date, by Knox (1964) and Kellett (1937). The GAM (Openshaw et al 1987) has gone some way in answering this call. The other issue raised was to look to a far wider range of environmental topics. Since this would require a more flexible framework for analysis it would appear that a GIS approach is at present the most apt technology available to fit this criteria.

### 1.3.3 Media attention

The public have become increasingly aware of, and interested in the environment and its possible impact upon health. They are constantly being informed by environmental organisations such as Greenpeace, Friends of the Earth and the media in general. The late 1980's in particular proved to be an extremely enlightening and informative period due to the accidents of Chernobyl and Seveso.

The nuclear industry is one particular area which has come under extreme criticism. The first and perhaps well known television coverage was 'Windscale, the Nuclear Laundry' filmed by Yorkshire TV. It generated the desirable level of concern, and led to an enquiry group in 1983/4, chaired by Sir Douglas Black. The 1990's will see a continuation of this concern, as the governments, especially in the more developed nations start to strengthen their policies to alleviate and prevent hazardous elements effecting the environment. The latter has also been stimulated by the increased availability of disease data and a higher level of research activity in health related areas, including the establishment of a Small Health Statistics Unit (SAHSU) in the UK since 1987.

However low-level radiation discharges from nuclear reprocessing plants are no longer solely blamed for health problems. It has been brought to light that other environmental factors should also be investigated, such as rivers which should meet quality control standards, especially with certain chemicals namely aluminium now being related to the brain wasting disease, Alzheimers (Edwardson, 1988). In addition other source outlets of pollution are being met with concern, including toxic waste dumps, coke works, chemical plants, and mines, as reflected by the media on a frequent basis. The following headlines therefore are just a taste of the media attention given to the impacts of the environment on health.

'Presence of metal in water linked to Alzheimer's disease' (The Guardian 30/1/89)

'Toxic Waste plant inquiry' (The Guardian 7/3/89)

'Ban all CFCs immediately, says the Prince of Wales" (The Journal 7/3/89)

'Tapwater tests over cancer link to children' (Today 1/1/90)

As Macquil (1988), has pointed out diseases, especially those affecting children that can be attributed to man-made intervention, will always (and quite rightly so) attract considerable media, political and public concern and sympathy. However whilst the latter headlines provide reasons for a GIS method of analysis into health and environmentally related issues, the following provides a warning that we need to get it right! Findings must be 'correct' from a scientific point and not as happens, all too often, unduly distorted for 'sensational headlines'.

'Cancer link to ketchup' (Today 14/2/90)

'No babies advice at Sellafield' (The Guardian 22/2/90)

In these cases public concern is being heightened because of an insufficient information or medical knowledge to explain the causes of such malignancies. The need therefore is to improve this situation with calls on any research, however vague its contribution, in order to achieve this goal.

## 1.4 Can GIS do it?

This chapter has provided evidence to substantiate the contribution that geography can make to spatial epidemiology and in particular the search for environmental causes of Acute Lymphoblastic Leukaemia. In addition it demonstrates how previous research and a general desire for more knowledge in this area demands that any relevant technology which is introduced should attempt to tackle this problem. In both instances it is suggested that GIS's ability to store a variety of spatially referenced

data in a single coherent framework and a 'toolbox' of cartographical and analytical facilities may be the answer.

If the GIS software manufacturers' descriptions and the associated 'hype' is to be believed then GIS will succeed as a spatial epidemiological tool. The essence of this thesis therefore is to evaluate to what extent these claims are true and if the objectives to find evidence of causes of childhood cancer can be achieved. However, even if the results from this study prove to be negative, it is still comforting to know that something is being done and the best available technology is being employed.

It should be stressed though that given the nature and quality of the data available, to be discussed in Chapters 3 and 4, and the limited medical knowledge, the geographer with or without a GIS will never be qualified to offer anything more than weak evidence for a medical hypothesis. Consequently, the culmination of any geographically based argument is to recognise possible causative factors at a preliminary level ie 'x' plus a little 'y' could cause 'z', this is known as aetiology (McGlashan, 1972). This can be useful in the management, and highlight of, possible social and/or environmental risks which was recognised by the European HEGIS initiatives, see Appendix A.

The role of the geographer and this GIS approach to health issues is aptly summarised by Alderson (1983), who suggests that by providing additional knowledge we can;

> ' learn more about the causes of leukaemia in the hope that it will be possible to remove these causes, or at least reduce the population exposure to them.'

## 1.5 Contents of the Thesis

As mentioned in section 1.1 this thesis is designed as a small scale pilot study for developing a health application, with particular emphasis on GIS evaluation. It is broken down into three basic sections. The first section provides an overall discussion of the research, thus this chapter emphasised the factors surrounding the spatial epidemiological application to be employed and the need for new techniques to tackle this. From this the use of GIS technology has been suggested as an alternative approach and thus Chapter 2 will complete the overall introduction by concentrating on the general aspects and technicalities of the GIS environment.

This will be followed by Chapter 3 and 4 which describe and assess the data employed. Chapter 3 specifically reviews the medical data and associated hypotheses that surround the distribution and causation of Acute Lymphoblastic Leukaemia. Chapter 4 provides an overview of the environmental databases employed, their respective sources and links to child health. The links described are not always directly related to Acute Lymphoblastic Leukaemia because the idea of using a GIS framework is to attempt to overcome some of the more restrictive univariate approaches which have tended to dominate the environmental debate in the past. Thus any aspect of the environment which can be represented will be included allowing both vague and favoured hypotheses to be tested.

The second section constitutes 'The GIS Process'. Chapter 5 chronicles data capture processes and manipulation techniques. Chapter 6 outlines the procedures which determine GIS analysis. This is followed by Chapter 7, which summarises the results of the GIS analysis stage, but also assesses the effectiveness of GIS in terms of tackling a spatial epidemiological problem.

The third section of this thesis raises some of the pertinent questions which epidemiologists are asking but GIS failed to answer. Many of the problems discussed in this research are attributed to a lack of spatial analysis functionality. Chapters 8 and 9 therefore explore two key areas which are considered important to GIS; including 'pattern spotting' and 'relationship seekers'. Chapter 10 also provides a brief evaluation of some of the problems which are likely to be experienced and may prove critical in the development of GIS. Chapter 11 summarises the research with some concluding remarks on the use of GIS as a spatial epidemiological tool and the possible way forward for this technology.

# CHAPTER 2

# AN OVERVIEW OF GIS

A brief reflection of the development of Geographical Information Systems (GIS) will be included at this point as a means of emphasising the infancy of this technology and the importance of this study for providing one of the first complete documentations of GIS as a spatial epidemiological tool.

Existing paper-based maps are in a sense an early version of a GIS serving to collect and store spatially referenced data from a variety of sources. Thus it may be argued that the concept of GIS is as old as cartography. The sophisticated version of GIS referred to in this research though was first introduced in the 1950's, when the necessary platform for development was created. This coincided with the arrival of digital computers and a general improvement in the power of computers. Whilst at the same time authors such as Tobler (1959) began advocating the basic automated concept of GIS. Tobler made specific reference to the fact that there would be

> '..some day stock decks of punched cards containing basic geographic information...a separate deck for each category of geographic information, such as coastlines, state boundaries, cities, contours, railroads and population..'

and he stated that this in turn would increase the flexibility of data handling and storage. This essentially describes the model of present GIS software packages.

The Canada Geographic Information System set up in 1964 was the first major operational GIS and, unlike the New York Land Use and National Resources Information System (1967), is still used today. The UK response to the potential of GIS for application work however was not fully recognised until twenty years later, with the development of the Regional Research Laboratories (Economic and Social Research Council, 1987). This served to provide academia with access to suitable software packages which had previously been restricted by the cost of technology. For example, ARC/INFO (Doric, Environmental Systems Research Institute) was not acquired by the University of Newcastle upon Tyne until 1987 coinciding with the onset of this research application. Thus this researcher had the ominous task of

starting at the very bottom of the GIS 'learning curve'. The advantage of this is that a true overall evaluation can be made, and hopefully documenting procedures and experiences that accompany the building of a GIS will benefit others contemplating a similar route.

Since the mid 1980's the situation in the UK has radically changed, with virtually every geography department now owning at least one GIS. It should be noted at this point however, that despite the wording Geographical Information Systems and the use of this technology by a geographer for a geographical approach to spatial epidemiology it does not make this technology the sole prerogative of the geography discipline. The expansion in GIS is being parallelled in the private end-user market too, with over 200 software manufacturers and some 1000 plus users, including Local Government, Health and Police authorities. In each case they have adopted the technology in the belief that it will answer most of their spatial data handling problems. This research will examine to what extent 'state-of-the-art' GISs can satisfy specific end-user needs. Thus this chapter, and those to follow, are designed to provide an overview of GIS and a continuous assessment of the advantages and pitfalls that can be encountered at each stage of the development of a GIS application.

## 2.1 The GIS Environment

This section will focus on the hardware and software actually made available for this study. Although the data model itself and the requirements for any particular application are basically generic. They include;

### 2.1.1 Software

This involved the use of ARC/INFO, the major proprietary GIS in most British Universities. The dominance of ARC/INFO in academia is the result of enormous discounts on copies of the software negotiated by the Committee of Higher Education Software Team (CHEST). This situation leads to a number of advantages;

i) ARC/INFO will run on a variety of platforms, notably the DEC VAX using VMS and Sun SPARC station 2, which are both used in this application. Other platforms

include Prime computers using the PR1MOS operating system, and IBM Personal Computers compatible with MS DOS.

ii) With a large number of universities using the same system Britain to some extent can avoid the North American GIS experience, where a vast range of software systems running on independent and incompatible machines have been adopted. This is perhaps an important factor for the European approach to consider, even though a number of data conversion programs are now available to reduce such inconsistency problems.

iii) ARC/INFO is one of the world leaders in GIS. An important consideration therefore is that leading software manufacturers are more likely to keep their technology up-to-date, incorporating new modules and analytical techniques whenever possible.

The ARC/INFO software package is a cartographic system built around a hybrid data model which organises geographic data via a relational and topological model. It has two main components to facilitate the efficient handling of data these are;

**ARC** which is the main program environment containing all the commands to initiate any of the subsystems that manipulate the spatial data and,
**INFO** which is the relational database manager storing the tabular/attribute data that accompanies geographic features.

Locational information is stored in ARC/INFO as a series of cartesian coordinates (x- and y-), allowing an exact representation of geographical features. This type of storage defines ARC/INFO as a vector based system. As a result the software contains a number of sophisticated techniques for manipulating vector data, but the detail involved to store point and areal features does render simple overlay and buffer operations to be considerably slow due to the complex mathematical calculations required to create new coverages in this way. This is compensated to some extent though by the preciseness of the resultant coverages.

### 2.1.2 Hardware

The ARC/INFO software was run on a Tektronix 4207 colour graphics terminal, connected to a DEC MICROVAX II, and in the final year of this research a Sun SPARC station 2 was made available which served to speed up the analysis procedure and extend the available GIS tools which could be supported by this improved hardware platform. The majority of data conversion and manipulation required to produce ARC/INFO compatible datasets was executed on the Newcastle University mainframe system (Amdahl 5860). Other peripheral hardware employed (although not essential to a GIS environment) included a Calcomp 91480 (A0) digitiser, used for data capture purposes. The maps presented in this thesis utilised the plot output interface of ARC/INFO to send maps to PostScript devices, such as the IBM 4216 Personal Pageprinter (laser printer) and the Calcomp 5912, 8 colour thermal wax plotter.

## 2.2 Main stages of a GIS

Whilst the terminology and software design may vary from system to system, all will contain four basic components, the contents of which are summarised in Figure 2.1.

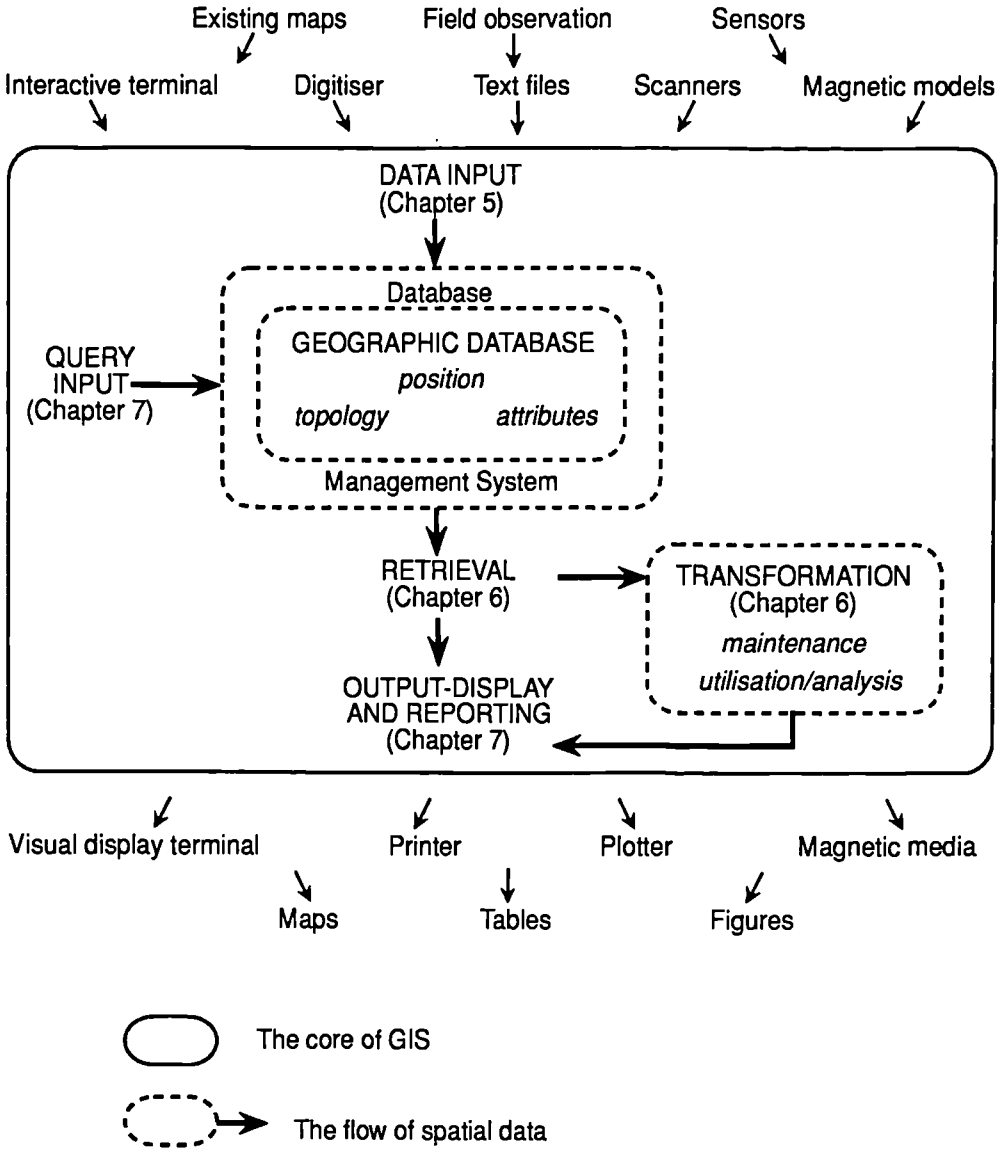
*Stage I*

This includes a data input subsystem which collects and/or processes spatial data derived from existing maps, documents, remote sensors etc. This refers to the conversion of all databases, described in Chapters 3 and 4, into a GIS compatible format. This is possible by either the direct manual keyboard entry of points, for example the records for childhood cancers diagnosed, or the conversion of maps, using ADS the ARC Digitising System, which offers full digitising capabilities with an easy-to-learn menu interface. This is explained in greater detail in Chapter 5.

*Stage II*

This includes a data storage and retrieval subsystem which organises the spatial data to allow quick and easy retrieval for subsequent analysis, and permits rapid and accurate updates of databases. In the case of a Cancer Registry therefore new records can be added to the health database as soon as they have been verified. This stage

**Figure 2.1** A schematic view of the GIS environment




Source: The integration of several diagrams from Burrough (1986)

may also include the use of an editing subsystem, ARCEDIT, a unique graphics and database editor that enables interactive editing of incomplete data (Chapter 5).

*Stage III*

This is a data manipulation and analysis subsystem which performs a variety of tasks. These include; changing the format of data by employing user defined aggregation rules, for instance aggregating individual cancer cases to County level. As well as the ability to produce estimates for population thresholds or constraints on hazardous levels of pollution which can then be employed in models for monitoring the environment. Thus the manipulation of data can range from simple reselections of certain aspects of a database to carry out more detailed analysis, such as a breakdown of Acute Lymphoblastic Leukaemia into age/sex categories. Alternatively these procedures can involve more complicated transformations of data into new coverages, using modules such as TIN (Triangular Irregular Network) which will convert points (irregularly spaced) to a three dimensional surface. In this research this technique was particularly useful in managing rainfall and air pollution datasets, highlighted in Chapter 6.

An important point to make here is the limitation of GIS as an analysis tool. The reasons for this and methods for alleviating the problems that exist will be demonstrated in Chapters 8 and 9. However the main problem is identified as one of conflicting terminology whereby the GIS software manufacturer's definition of 'spatial analysis' does not match the end-users interpretation of 'spatial analysis'. As a result Chapters 6 and 7 will determine the potential of GIS in terms of analysis and how its existing tools can, or cannot, tackle the complex spatial epidemiological problems of disease causation.

*Stage IV*

This is a data reporting subsystem which is capable of displaying all or part of the original database. This involves the creation of map displays via computer cartography, reports and basic statistics. ARCPLOT provides an interactive cartographic and mapping subsystem.

In addition, GIS command languages can be coupled with extensive macro and menu building tools. This allows the production of 'tailor-made' interfaces which suffice to bring GIS into the realms of a 'user-friendly' system. This is particularly important within the epidemiological field as it allows medical practitioners to query data and benefit from the facilities it provides without knowing the commands and technicalities surrounding GIS. Customisation issues are discussed in Chapter 10.

This chapter has put GIS technology into context. In order to even contemplate exploiting its potential though, the relevant spatial databases need to be established. Chapters 3 and 4 therefore aim to provide a rationale for the datasets included in this research and highlight the kind of theories to be tested using GIS technology.

# CHAPTER 3

# CHILDHOOD CANCER: EPIDEMIOLOGY AND DATA SOURCES

This chapter details the characteristics of the cancer database used in this research. It also substantiates the relevancy of the environmental and socioeconomic elements hypothesised to be important causal factors.

## 3.1 The Childhood Cancer Database

The relevant details for developing a cancer database were gathered from the Northern Region Children's Cancer Registry (1968-) which is based in the Department of Child Health, at the Royal Victoria Infirmary Hospital, Newcastle. The registry records the diagnoses of all types of malignancies in children up to and including the age of 15. It covers the whole of the Northern Region, which for this research includes the counties of Cleveland, Durham, Tyne and Wear, Northumberland and Cumbria. The Registry is believed to be approximately 98 percent comprehensive with updates completed by active liaison with the treatment centres in the area and reinforced by double checks based on referrals to the 'Hospital Action Record Sheets' and death certificates. Additional quality is maintained by frequent cross checks made against the statutory register of cases aged 0-25 years which is held at the Newcastle General Hospital. Thus it does not rely on passive notification which could potentially lead to undesirable lag periods in data capture or missing data. The Registry contains information on; the type of cancer diagnosed, the date and place of diagnosis, the date and place of birth and the sex of the person. Appendix C provides a complete list of the types of cancers that are registered in the Cancer Registry. For the purpose of this research however attention is given to one particular type of childhood cancer, that of Acute Lymphoblastic Leukaemia, accounting for 30 percent of all diagnoses in the Registry (in all subsequent chapters this cancer will be referred to as ALL for short).

Leukaemia is the most common of childhood cancers. The term leukaemia is used to describe a group of cancers that produce an excess of white cells caused by the proliferation of malignant clones in the tissue from which white cells are normally

produced. Different types of leukaemia correspond to the different types of cell from which the cancers arise and from the severity of the malignant degeneration. The ultimate effect of the imbalance of blood cells means that the patients blood can no longer function normally, increasing his/her vulnerability to infection.

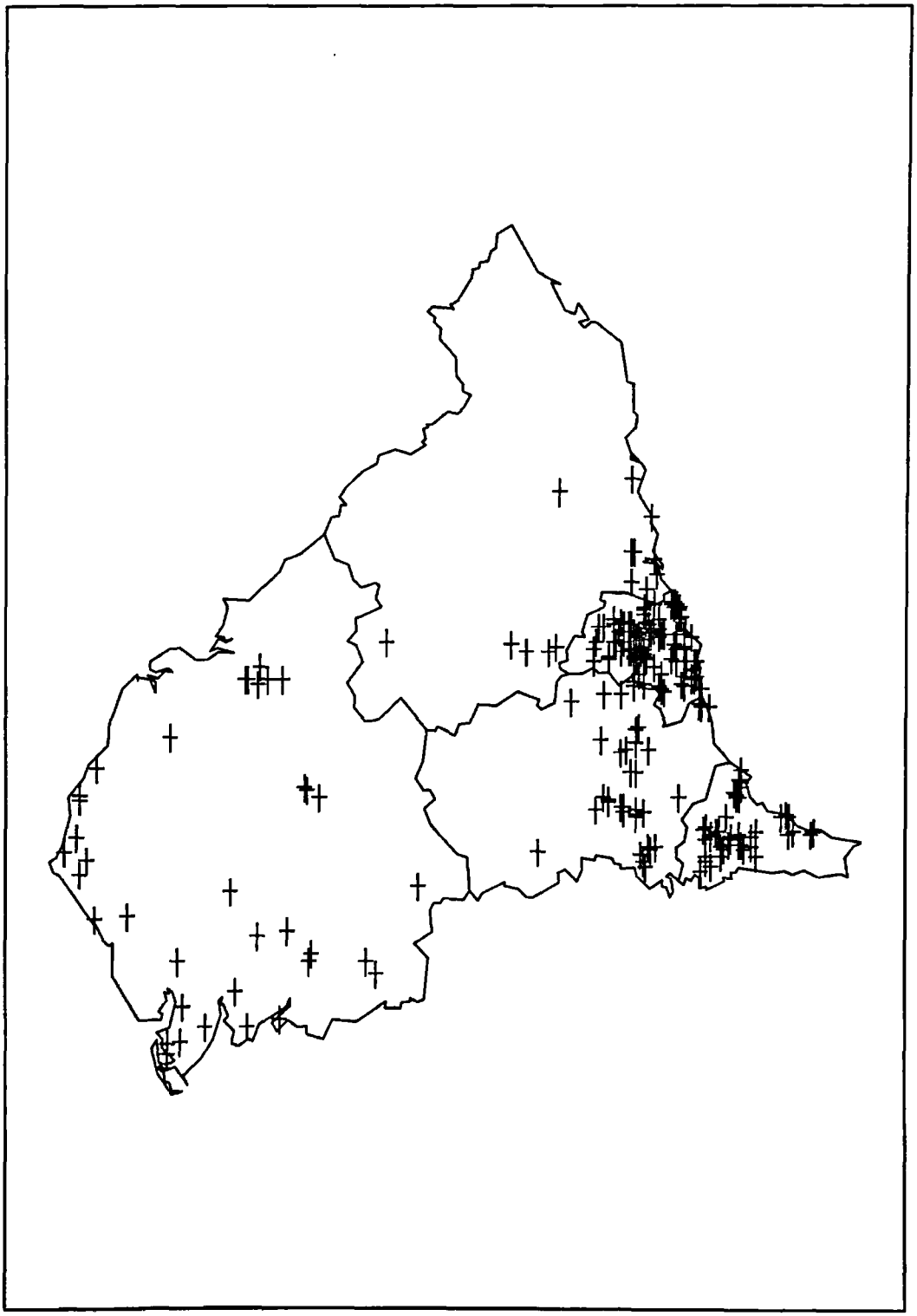
The distinction between childhood and adult leukaemia is not absolute, but the separation is useful for this research because Acute Lymphoblastic Leukaemia is more common in children and it differs from the common leukaemias that effect adult life. This in turn may reflect different causes. The upper limit to childhood is commonly taken as 15 years of age, this is also sufficiently close to the age at which the incidence of leukaemia is at a minimum. This does not however discount some adult leukaemias occurring in under 15 year olds, or some childhood leukaemias continuing into adulthood. For this research though, children between the ages of 0 to 15 years of age have been taken as an acceptable working limit for analysing incidences of Acute Lymphoblastic Leukaemia. The study region and the distribution of this childhood cancer are shown in Figure 3.1.

Whilst the main interest is upon the distribution of Acute Lymphoblastic Leukaemia and the possible environmental correlates it is useful to have some means of comparing results from subsequent analysis. This is achieved in this chapter by using three other cancer categories from the Registry, including Non-Hodgkin's Lymphoma, Bone and Brain Tumours, outlined in Table 3.1.

**Table 3.1**: Cancer categories to be referred to in this research

| |
|---|
| A Acute Lymphoblastic Leukaemia (ALL) |
| B Non-Hodgkin's lymphoma (NHL) |
| C Bone Tumours |
| D Brain Tumours |

Figure 3.1: The distribution of Acute Lymphoblastic Leukaemia in Northern England



Since the cancer database is the crux of this GIS application it also played an important part in determining key operational decisions concerning database design. For instance, the geographical resolution for all the environmental datasets employed. This was set at 100 metres because the spatial reference for all the cancer cases was based on the addresses of patients at the time of diagnosis, to this a postcode was attached which can be converted to a 100 metre grid-reference, as described in Chapter 5.

Secondly the research restricts its analysis to those patients which were diagnosed with Acute Lymphoblastic leukaemia between the years of 1976 and 1986, leading to a sample size of 225 cases. This is despite the fact that the Cancer Registry includes all diagnoses from December 1968, and was recently extended from 1986 to cover all cases up to and including 1989. The reason for this particular time period is based on the need to calculate rates of cancer incidences, requiring a reliable population base compatible with the cancer data available. In the UK though a comprehensive population count is only taken every 10 years, therefore an accurate count for the period can only be gathered from the 1971 and/or 1981 census. However the administrative units employed to store key population information for these two census, ie enumeration districts, are incompatible. It was decided therefore, to concentrate on a time period that could best be explained by one census count alone. This would reduce any inaccuracies which may have occurred in attempting to combine the two censuses through a fuzzy matching process. Thus the variables for the 1981 census were selected as the population base for those cancer cases occurring between 1976 to 1986.

This section has provided a summary of the criteria which determined the development of the main cancer database. Considerable effort was made to ensure the accuracy of data and this was given a figure of 98 percent, but it must be stressed that this is based upon the known problems of the data recorded. There are however a number of errors which may reduce this figure further many of which will be expanded upon in Chapter 10. However, the next section will highlight some of the additional factors which may have indirectly effected the quality of data recorded in the Cancer Registry. This in turn will emphasise the importance of fully documenting key characteristics of any database in order to render future interpretations of data and analysis successful, particularly at this stage of the building of a GIS.

## 3.2 Completeness and Accuracy of data collected

Missing data may include late registrations, which artificially reduce the number of cancers available for analysis. More seriously, some cases may be eliminated completely because the child died before the malignancy could be diagnosed, or moved out of the region with the disease and was diagnosed elsewhere. Conversely, duplications of registrations can artificially inflate the figure produced. This is most likely to occur when patients cross over regional boundaries, ie the patient lives in one Local Health Authority area but is treated in another. Another scenario is where the patient may have consecutively developed a number of malignancies all of which have been recorded individually, resulting in multiple entries in the Registry.

In addition the various categories and absolute numbers of incidences may be distorted by changes in the cancer type codes or even differences in diagnosis. For instance, cancer may occur in an individual but not be diagnosed, or may be mis-diagnosed as a non-malignant condition (ie leukaemia diagnosed as an infection). Similarly a diagnosis of cancer may be made where no malignancy is in fact present. The total extent of these types of error within the Registry are not calculable, but are probably only a matter of a few percent (Swerdlow, 1989). The latter suggests that the data may never be 100 percent accurate, or in this case the 98 percent stated, and this should be borne in mind throughout the stages of the development of a GIS.

## 3.3 Descriptive Epidemiology of Acute Lymphoblastic Leukaemia

In the absence of any known aetiology the data contained in the Cancer Registry can be used to carry out 'descriptive epidemiology'. This acts as a fall back mechanism with the hope that it may generate some clues as to where to start looking for possible causes of certain diseases. Doll (1989) suggested five areas for speculative investigation, including the variation of cancers according to; age, sex, place of ethnicity, socioeconomic factors and time. Thus with all the relevant information stored in a comprehensive cancer database GIS can be employed to manipulate and summarise data to satisfy similar descriptive epidemiological analysis on Acute

Lymphoblastic Leukaemia, which in the following sections will be compared to the other three categories mentioned in Table 3.1.

### 3.3.1 Variation with Age

A characteristic of all childhood cancers is the rapid increase in numbers after birth, reaching a peak followed by a decline. Figure 3.2 summarises this point with a barchart representing the distribution of Acute Lymphoblastic Leukaemias diagnosed according to three main age groups, 0-4, 5-9 and 10-15 years of age. The age at which childhood cancers peak though varies from one type of cancer to another, as demonstrated by the comparison of cases in each age group for the four cancer categories, see Table 3.2.

**Table 3.2:** Proportion of cases in each age group according to the cancer categories specified in Table 3.1

| | | Cancer Category | | | |
|---|---|---|---|---|---|
| | | ALL | NHL | Bone | Brain |
| Age Group | 0- 4 years | 70% | 37% | 0% | 57% |
| | 5- 9 years | 27% | 48% | 58% | 36% |
| | 10-15 years | 3% | 15% | 42% | 7 % |

Detailed haematological and immunological examinations show that the sharp peak in incidences of leukaemia between 1-4 years of age is entirely due to Acute Lymphoblastic Leukaemia. This can be seen in Figure 3.3 which breaks down the incidences of Acute Lymphoblastic Leukaemia into yearly intervals showing a peak at 4 years.

**Figure 3.2:** Distribution of ALL according to 3 standard age groups

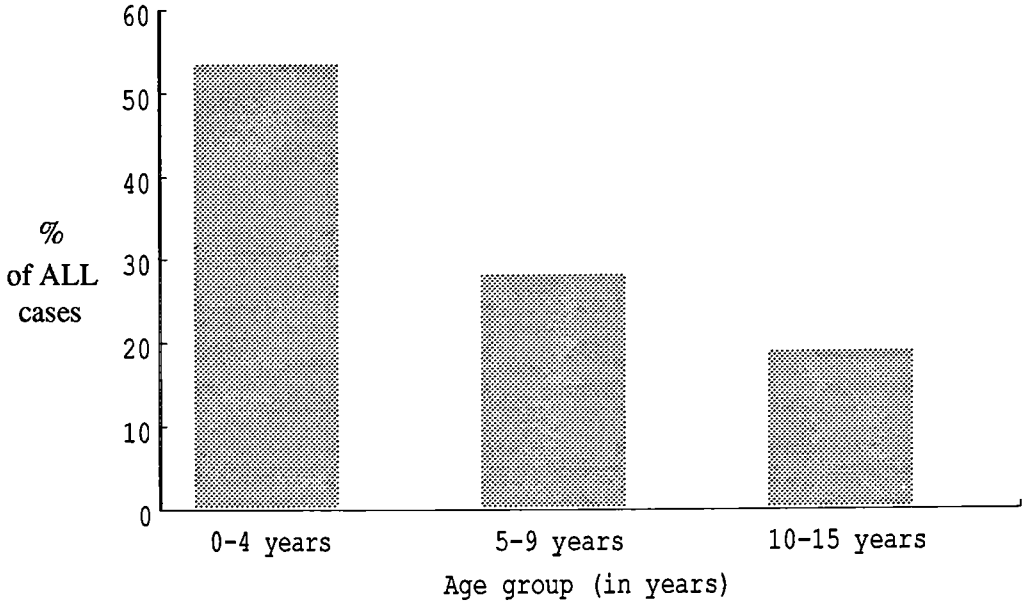
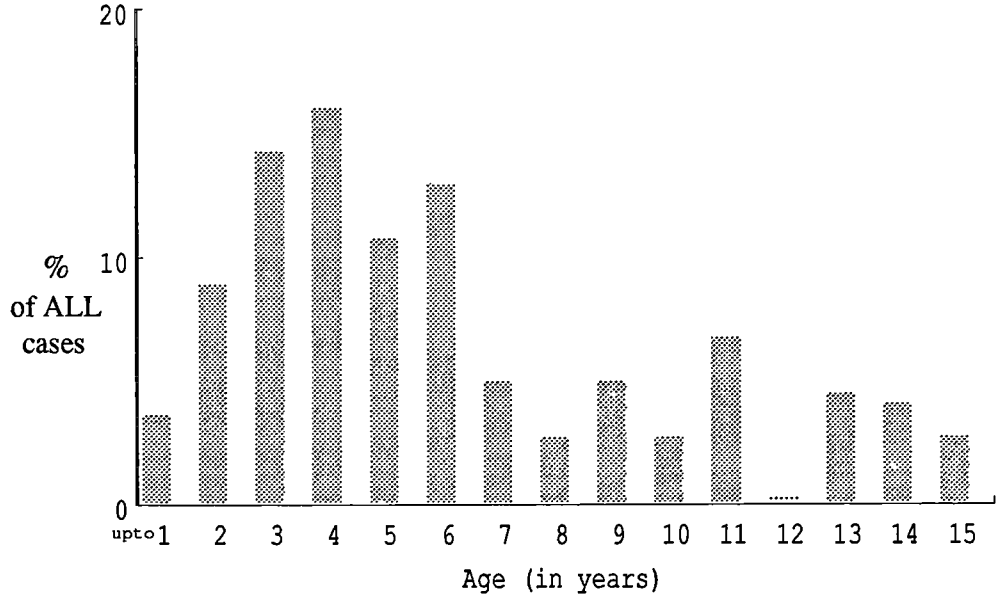



**Figure 3.3:** A more detailed view of the age distribution of ALL



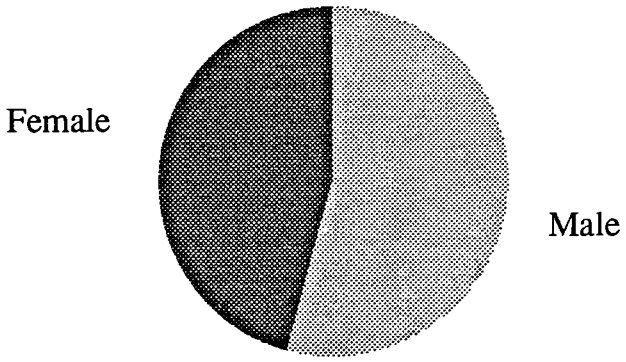
### 3.3.2 Variation with Sex

Childhood cancers tend to be slightly more common in boys than in girls, with a sex ratio consistently at about 1.2 to 1.0 (Doll, 1989). This is illustrated in Figure 3.4, and Table 3.3 provides a list of sex ratios for the four different cancer categories.

**Table 3.3:** Sex ratios for each of the cancer categories specified in Table 3.1

| | | Cancer Category | | | |
|---|---|---|---|---|---|
| | | **ALL** | **NHL** | **Bone** | **Brain** |
| **Sex** | **Female** | 1.0 | 1.0 | 1.0 | 1.0 |
| | **Male** | 1.3 | 1.5 | 1.4 | 1.0 |

**Figure 3.4:** An illustration of the ALL sex ratio




One possible explanation concerns the average birth rate of male infants, because on average they tend to be 4 percent heavier than females, therefore there is a relatively larger number of cells at risk which may account for an increased cancer rate in males.

### 3.3.3 Variation with Place and Ethnicity

This is not easily measured where the area concerned is fairly small or homogeneous. However previous research has found that the number of incidences of childhood cancers seldom vary across different communities, and childhood leukaemia is no exception. Any variations which may be observed are considered dubious, because of the error and differences in diagnosis procedures (Parkin et al, 1988). This is not a problem in this research given the high quality of data available and the fact that only 1.9 percent of the total population in the Northern Region is of ethnic origin anyway, which would render any significant analysis in this area difficult.

### 3.3.4 Variation with Socioeconomic factors

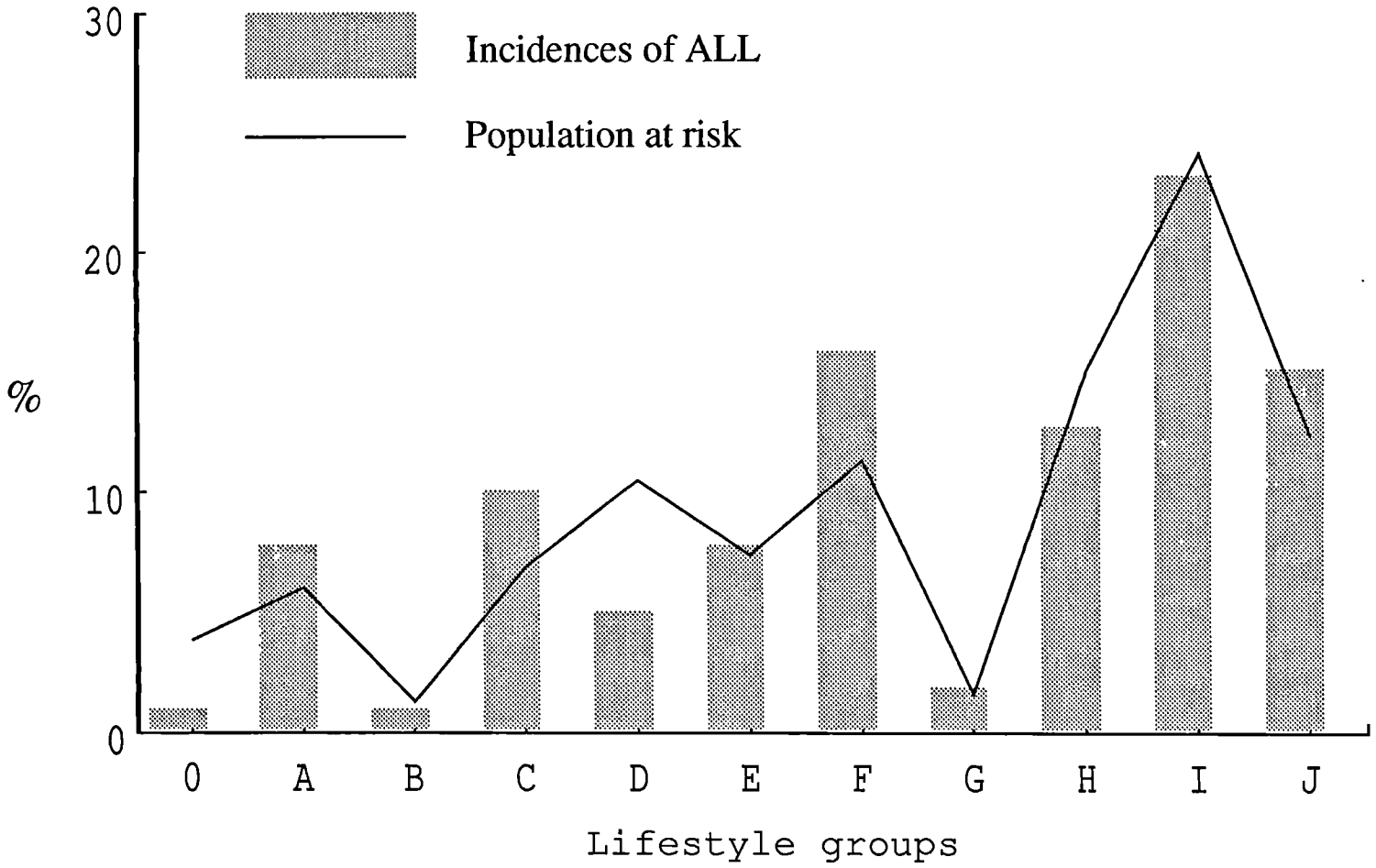
Differences have been found in Acute Lymphoblastic Leukaemia between localities with different proportions of the population in different socioeconomic classes (Greaves, 1986). Since the incidences of Acute Lymphoblastic Leukaemia have a postcode reference it is a relatively simple task to tag socioeconomic information to these incidences, the method of which is provided in chapter 4 which highlights the development and manipulation of complementary datasets such as these. Table 3.4 summarises the type of categories which are used to represent socioeconomic class, whilst Figure 3.5 provides a graph of the distribution of Acute Lymphoblastic Leukaemia in the Northern Region. This is based on the proportion of cases found in each 'lifestyle' group with the population at risk superimposed to demonstrate how the percentage of cases differ from that which may be expected.

**Table 3.4:** Socioeconomic variables based on 'lifestyle' groups

| Lifestyle | Description | % Hholds in UK |
|---|---|---|
| A | Affluent Minority | 8.3 |
| B | Metro Singles | 5.1 |
| C | Young Married Suburbia | 8.6 |
| D | Country and Retiring Suburbia | 9.0 |
| E | Old Suburbia | 8.8 |
| F | Aspiring Blue and White Collars | 14.3 |
| G | Multi-ethnic Areas | 7.4 |
| H | Fading Industrial | 12.2 |
| I | Council Tenants | 16.0 |
| J | The Under-privileged | 8.7 |

Source: Super Profiles: The new generation in market segmentation (CDMS Marketing Services, 1987)

**Figure 3.5:** Breakdown of ALL according to socioeconomic 'lifestyle' groups



It can be seen that for a majority of the 'lifestyle' groups the proportion of cases in each group is comparable to that which would be expected from the general socioeconomic profile for the region as a whole, suggesting that the likelihood of a child developing Acute Lymphoblastic Leukaemia is equally probable whatever their socioeconomic background. Only three categories; affluent minority, young married suburbia and aspiring blue/white collar show a marked increased proportion of cases compared to the population distribution for those groups (A, C and F), but this may be explained by the fact that these sections of the population are more likely to have children falling into the critical age bracket of 0-4 years. The only group which shows a negative distinction, ie there are less cases of Acute Lymphoblastic Leukaemia compared to the percentage of the population at risk, is that of D but since this is categorised as 'country and retiring suburbia' it is perhaps not that unusual.

### 3.3.5 Variation with Time

Variation with time is the final descriptive variable that Doll(1989) assigns to the distribution characteristics of childhood leukaemia with particular reference to mortality rates. However it would be expected that an overall reduction in mortality rates would occur given improved diagnostic techniques leading to cases being treated earlier, and the fact that there has been an improvement in treatment of Acute Lymphoblastic Leukaemia, although it can be superseded by other malignancies! Thus with this cancer any investigation into the possible variations in time are best looked at in terms of the numbers being diagnosed. Figure 3.6 shows that this is fairly consistent with slightly fewer numbers in 1979 and 1980, but in general for the decade 1976-86 about 10 percent a year were diagnosed. This temporal element is also exploited in Figure 3.7 which maps the incidences according to four three yearly time periods and this serves to demonstrate that in spatial terms the distribution seems to be consistent over time too.

This descriptive approach is only one aspect of spatial epidemiology and it is obvious that it does not generate enough knowledge to pin-point the aetiology of Acute Lymphoblastic Leukaemia, and that a more rigourous form of analysis is necessary to deduce possible causation. Thus the figures and tables produced in this section are analytically limited and in terms of GIS do not really stretch its capabilities very far. Even so a number of hypotheses have been formulated with respect to the causation of Acute Lymphoblastic Leukaemia and the next section will briefly outline some of

**Figure 3.6:**
The variation in ALL diagnosis, over the study period 1976-86

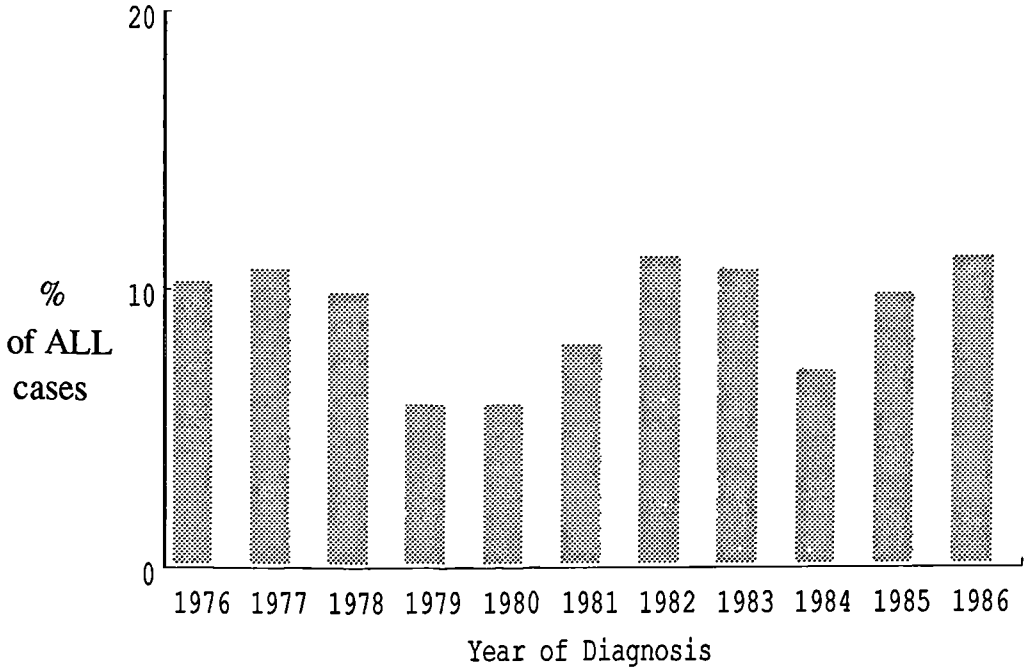
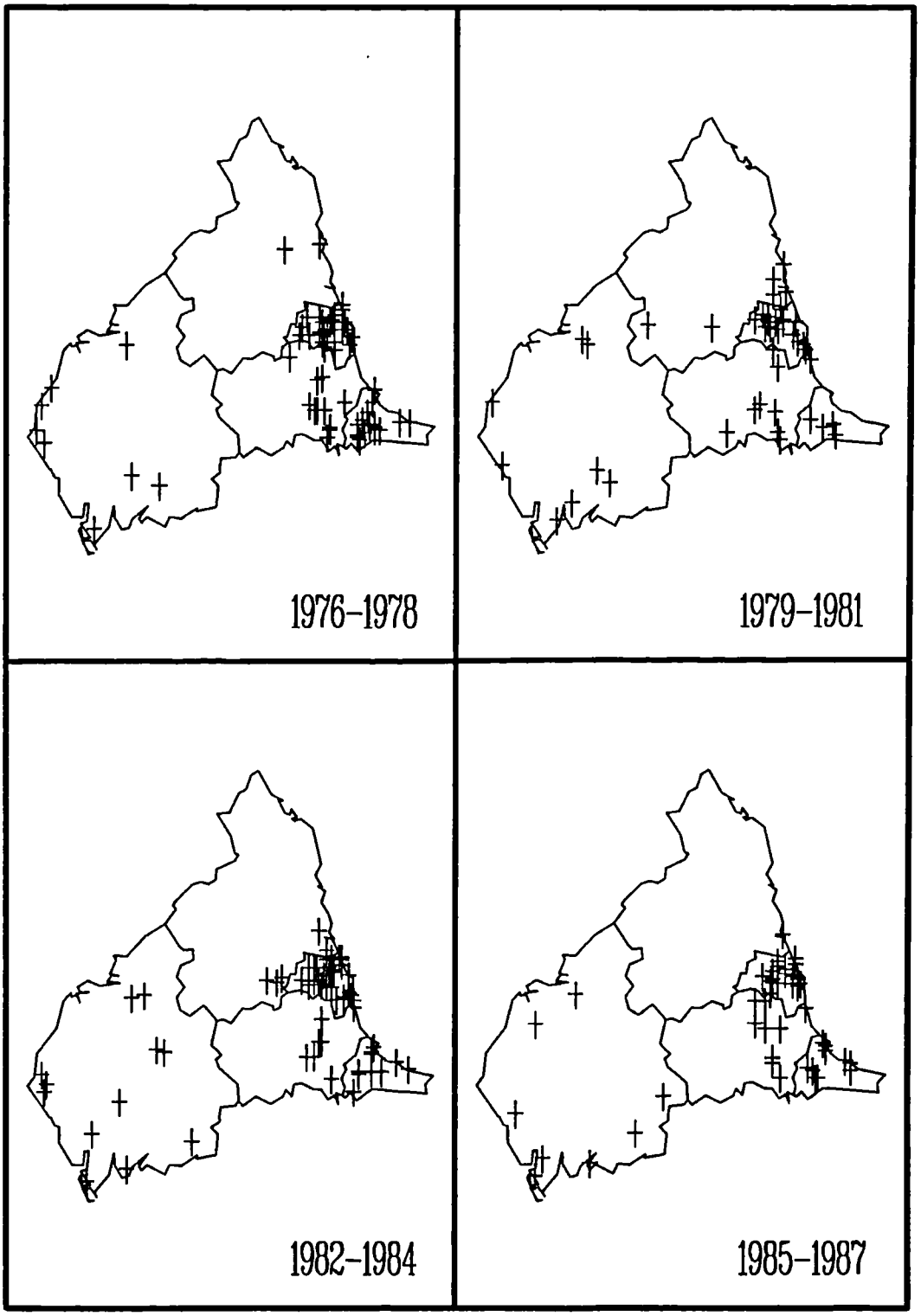


Figure 3.7: A temporal view of the distribution of ALL using GIS to select on date of diagnosis



the most common theories to date. Emphasising in turn the dominance of an environmental explanation, lending further credence to the importance of a Health and Environment Geographical Information System such as this to help to extend this particular area of research.

## 3.4 Existing Hypotheses

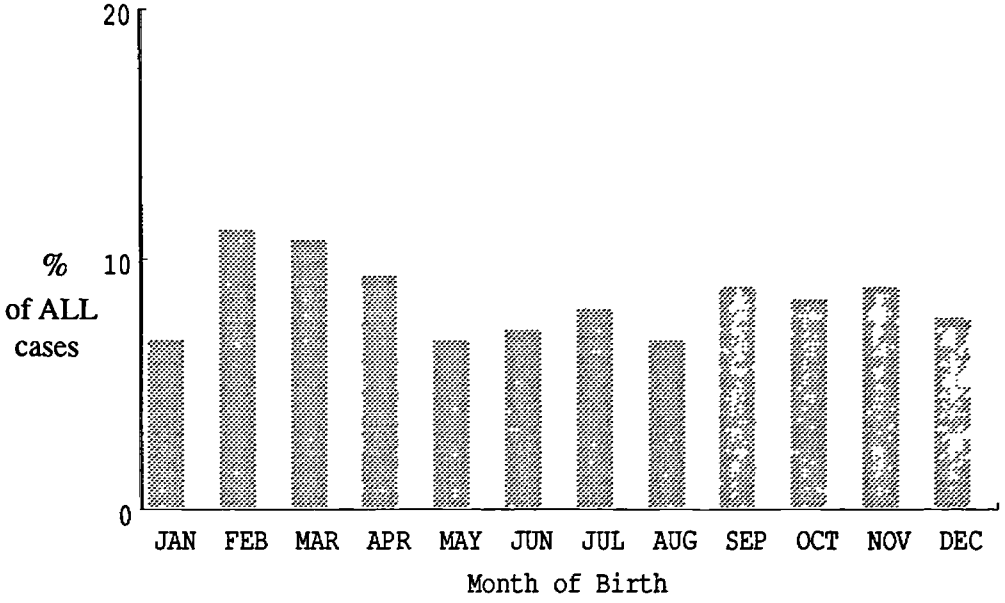
These cover a variety of interesting ideas ranging from occupational risk, urban/rural differentiation, space-time clustering and seasonal variation. This section is summarised according to the main researchers in this field and the theories that they have put forward for the causation of childhood cancers, and in particular that of Acute Lymphoblastic Leukaemia.

### 3.4.1 Knox - Space-time clustering

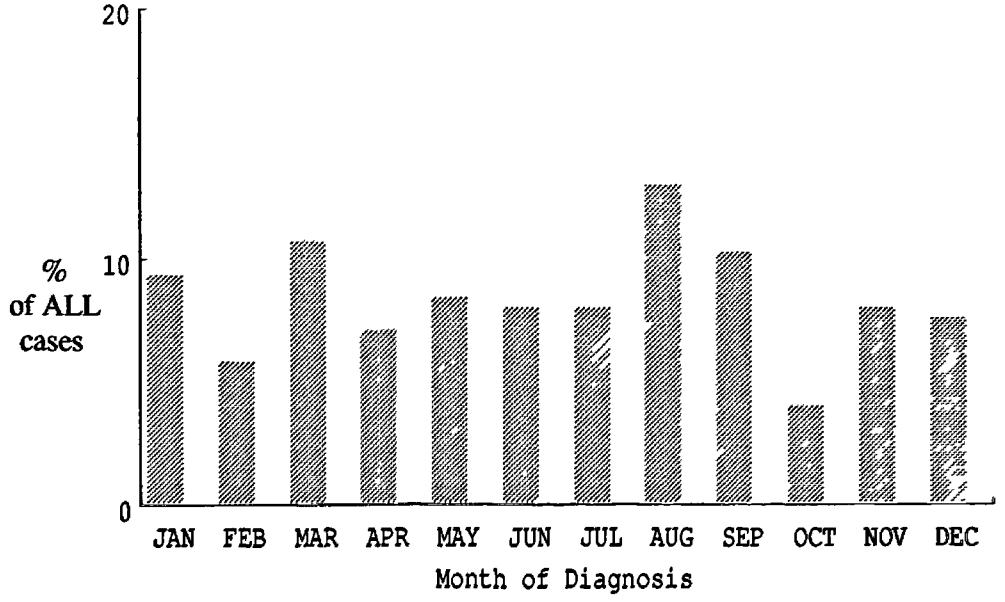
Knox (1964) carried out his investigations on incidences of childhood leukaemia in Northern England from 1951 to 1960, noting a possible clustering effect in both space and time, which had also been observed by Kellett (1937). He stated that there was evidence of (i) seasonal variation with a summer peak particularly for cases of Acute Lymphoblastic Leukaemia. Figure 3.8 summarises the data available for this research according to both the month of diagnosis and that of the month of birth to explore the latter idea, but in both cases Knox's idea of seasonal clustering does not appear to be evident. He also suggested (ii) a high risk in children living in larger towns, this can be tested in Chapter 7 which will compare incidences according to landuse type. Additional aspects which Knox noted from his results was that there was nothing to suggest that leukaemia was transmitted from one case to another, only that two cases may sometimes have a common source. This could be interpreted as a toxic rather than an infective agency spread through atmospheric pollution, through contamination of food and/or water supplies (Knox, 1964). Moreover, the risk concerning these hypothetical factors might also be seasonal with a vague link to the environment. The flexibility of GIS to store many features of the environment, described in Chapter 4, should allow such vague links to be explored.

**Figure 3.8:** Is there a seasonal variation?
Distribution of ALL according to;

(a) Month of Birth




(b) Month of Diagnosis



The idea of space-time distributions of childhood leukaemia is an on-going debate, with not only research by Knox (1964) in the '50's, but Van Steessel-Moll et al (1983) working on data from the 70's in The Netherlands and Openshaw et al (1987,88) in the '80's for Northern England. The Netherlands study (Van Steessel-Moll, 1983) actually proved a good example of conflicting theories, because it used similar techniques to Knox but found no seasonal variation in either months of births or the months of diagnosis. Although in this study it was suggested that long latency periods between the onset of malignancies and diagnosis may mask any clustering effect. At the same time though they did suggest that environmental agents within a limited geographical area may be a likely cause.

### 3.4.2 Greaves - Socioeconomic factors

Greaves (1986) attempted to look at several alternative causal mechanisms of Acute Lymphoblastic Leukaemia, including the idea of 'spontaneous mutation' (Greaves and Chan, 1986). This section though focuses upon his theory which concerned the increased incidence rates of cases associated with certain socioeconomic conditions. Greaves found that incidences of Acute Lymphoblastic Leukaemia tended to be more prevalent amongst children living within areas of high socioeconomic status. A similar analogy was drawn by McWhirter (1982) who compared African countries in the 1980's to Europe and the USA earlier this century. These factors though need to be set against a background of improving general medical care especially during neonatal periods and early infancy. Consequently any variation in socioeconomic characteristics may simply be a reflection of the degree of medical care that the child receives.

### 3.4.3 Kinlen - An Infectious cause

Kinlen (1988) suggests that Acute Lymphoblastic Leukaemia is the result of an unreasonable response to an infective agent due to a large number of people coming together to live in one place, but originating from different socioeconomic and behavioural backgrounds. His research findings showed a significant increase in childhood leukaemias in one particular rural district of Scotland, that of Glenrothes which was a New Town created in the 1950's. Since it was the only area at a distance

from a conurbation (without a nuclear installation close by) to receive such an influx of people, the causal mechanism for the observed childhood cancer cases was attributed to the mixing of a variety of different communities leaving them more susceptible to an unreasonable response to infection.

These New Towns also attract higher proportions of young people resulting in an unusually high density of children which ultimately increases the number of susceptible individuals. In the study area for this research there are the New Towns of Newton Aycliffe and Peterlee which have developed under similar circumstances. It should be noted though that this theory does not apply to those new towns which result from the over spill of large conurbations, such as London, because the people involved in this case exhibit similar habits and characteristics of residency.

### 3.4.4 Gardner - Occupational exposure

A final hypothesis relates to the occupational exposure to radiation of fathers with children diagnosed with Acute Lymphoblastic Leukaemia. Gardner (1990) suggested that the fathers occupation lead to an increased risk of their offspring developing such childhood cancers. If this is the case, then the risk of radiation to children around a nuclear power station, for example Sellafield in North West England is indirectly caused by the paternal employment and contact with radiation. This hypothesis though has insufficient evidence to prove causation but is important enough to warrant further investigation.

The immediate conclusions were genetic, hereditary causes, but it is possible that the most heavily exposed workers had inadvertently brought radioactive material home on contaminated clothes. This would serve to alter the micro-environment of the child by increasing the amount of their received radiation dosage. This angle of interpretation again brings the possible causal mechanism back to one of the environment, but in this case it is one which cannot be easily incorporated into a GIS framework because medical records are not detailed enough to provide histories for the child and his/her guardians. Gardner's (1990) study covered a small sample population and thus the conclusions should be noted with caution. It may be easier to look for geographical surrogates and detect areal anomalies in disease incidences rather than prove causation by identifying key mechanisms. Thus in a GIS approach

to this problem distance from Sellafield may be a reasonable geographical proxy for studying the effect of radiation.

Despite these hypotheses, substantial advances in the treatment of Acute Lymphoblastic Leukaemia and the understanding of its biology, the aetiology of this disease remains enigmatic (Greaves, 1986). This section has demonstrated that hypotheses are put forward, only to be proved by one author and then refuted by another, whilst others remain unsubstantiated. Gaining additional knowledge into the basic pathophysiology of the disease is therefore crucial, Linet (1982) suggested that this may be achieved by;

> 'identification of micro-environmental factors that may play a key role in disease onset, remission and relapse'

## 3.5 Causation: Looking to the environment

Even though Knox's theories are now some thirty years old, they still hold considerable weight in this research area. It is also interesting to note that whatever the stance the authors adopted, they all accepted that there may be a link with geographical and/or environmental factors. Thus with new GIS technology a more flexible exploratory approach into the possible environmental causes of Acute Lymphoblastic Leukaemia can be achieved. Providing an excellent medium to view the problem with an open mind, whilst accepting at the same time that the impact of the environment upon human health may not be simple or direct but more on the lines of a compounding or cumulative effect. Or as Easterly (1981) suggested the environment may establish an

> '..initiator promoter concept of cancer causation'

This assumes that the possibility of a malignancy developing does so or not depending on the presence of certain environmental factors, which then may act as a 'trigger' mechanism. A hypothesis free GIS framework therefore will allow ideas such as this to be investigated, as well as being able to explore several environmental databases. Thus from the datasets described in Chapter 4 favoured theories such as the proximity to nuclear installations can be tested or even vague ideas concerning the impact of landuse types. Both cases though may serve to stimulate new insights into

the distribution of Acute Lymphoblastic Leukaemia and its aetiology, from which the epidemiologist may develop new theories and areas of research.

# CHAPTER 4

# AN OVERVIEW OF ENVIRONMENTAL AND SOCIO-ECONOMIC DATABASES RELEVANT TO A HEGIS

## 4.1 Macro versus Micro Environments

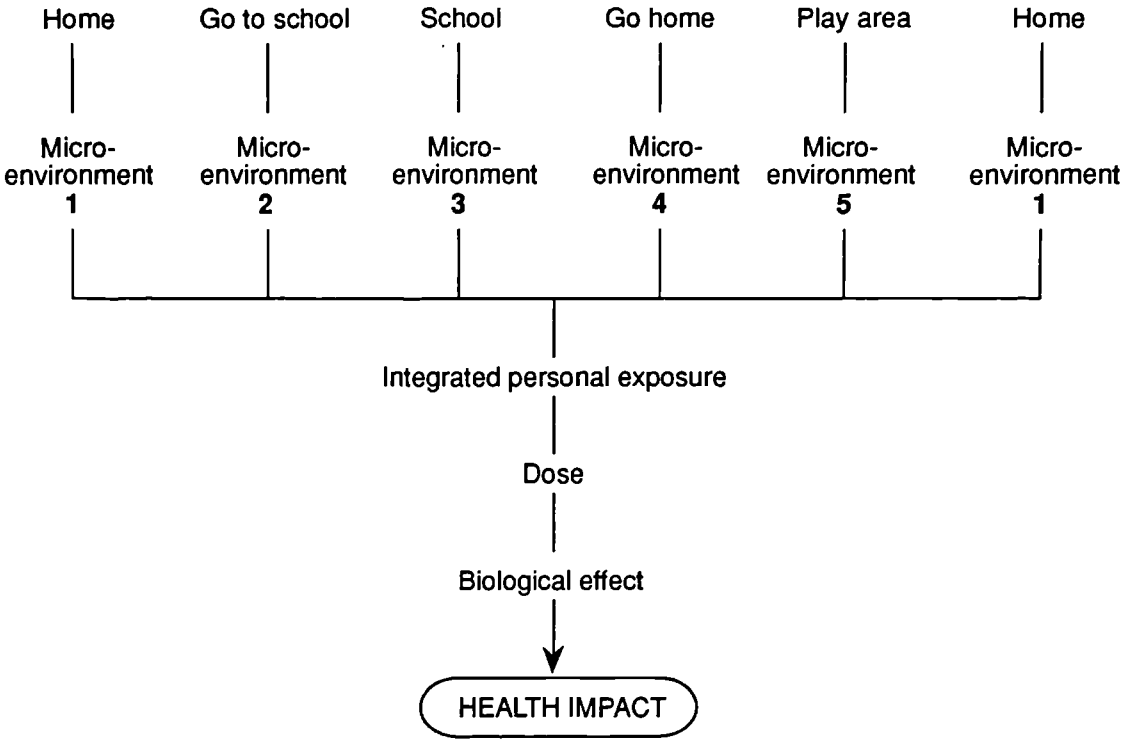
This chapter considers the sources of data and substantive evidence necessary to investigate the possible relationships which may exist between exposure to environmental contaminates and childhood cancer. This is not exhaustive since both the medical data and GIS technique dictate the use of a 'macro or general environment' focusing on the living space of the individual.

Causation may be determined by the 'micro- or personal environment' of an individual though including social and/or working spheres. A 'micro environment' therefore is more likely to vary over time and according to location ie. in the street, at home, or at school. In addition the concentration of any pollutant that an individual may receive may be determined by aspects of dilution, transport and physical/chemical processes, some individuals being more susceptible than others. Figure 4.1 illustrates schematically some of the variations that a child's micro-environment may be subject to. It is estimated though, that approximately 70% of an individuals time is spent at home and another 20% or more is spent elsewhere indoors (work, school, other buildings etc), see Lebret (1990). This coupled with the assumption that children are less mobile than adults and therefore more likely to remain in a reasonably homogeneous location, ie. at home, may suggest that the limitations of the medical data do not affect the test for an environmental health relationship of Acute Lymphoblastic Leukaemia (ALL) at a macro geographic environmental scale. A European model looking at public health in general should not, if possible, ignore the impact of other micro- environments.

## 4.2 Environmental Database Design

What constitutes an environmental database of relevance to a disease analyst? How do you justify the time and the effort required to include it in a GIS application?

**Figure 4.1** An illustration of the different micro-environments which can effect a child




Adapted from Lebret (1990)

Where is it possible to acquire the necessary information? What resolution should be considered, and how do you ensure the relevancy and accuracy of the data? These issues are fundamental to successful database development, and represent a sample of the questions needing to be considered at this stage. In some cases, these answers will be obvious, as they will be dictated by the availability of data, others will require some subjective decisions if data acquired are to reflect the needs of the hypotheses to be tested.

Many of the databases in this study have been selected because previous research findings have demonstrated an association (however weak) with the spatial distribution of ALL patients. Others have been included because of a general link with child health. It is thought that a more flexible GIS approach will allow all possible environmental aspects to be investigated which in turn may discover alternative aetiological factors that were ignored in more restricted studies.

It should be noted however that not all relevant elements of the environment can be directly measured or have specific spatial references. As a result geographical surrogates must be used to represent these missing features. The mapping of lead pollution is an example of this, whereby car exhaust fumes are a main source of lead pollution forming an integral part of everyone's micro-environment. However it does not warrant systematic monitoring by the government or health related agencies. In a GIS context the absence of any direct data measurements can be overcome to some extent by adopting the road network as a proxy for lead pollution corridors, the details of which will be provided in section 4.3.2.2.

An important advantage of the GIS approach over previous analysis procedures concerning environmental health issues is its inherent multivariate analysis capabilities. GIS enables the exploitation of databases in such a way as to test the combination or accumulation of causal factors upon individual childhood cancer cases. For instance, taking radiation as a health risk, there are obvious man-made pollution sources including power stations, waste dumps and special radioactive sites. Another source of pollution is background 'ionising' radiation produced by certain rock types, which up until recently was considered to have only a minor impact (NRPB, 1987). The implication of various sources of radiation are very important. Background radiation may be ignored as a significant risk factor when observed

independently, but the radiation dosage received by an individual may become more dangerous after exposure to both sources, man-made or otherwise.

The following section outlines all the datasets employed in this research, starting with what is perceived to be the most important risk factor; radiation, and then following with man-made, natural and socioeconomic factors. This section will serve to reflect the oncologist's most favoured environmental theories, such as nuclear power stations, incinerators etc, and then look at those factors which are not so strongly linked to health but have sufficient evidence to warrant inclusion, for example landuse, estuaries, socioeconomic status etc. A summary of these databases can be found in Table 4.1 which highlights the databases acquired for this HEGIS, references to chapter sections and the sources of key information. Table 4.5, found at the end of this chapter compliments Table 4.1 by providing important details regarding factors which are needed for successful design and implementation of GIS databases. These include the type of spatial referencing involved, the resolution of the source data, and the geographical extent of the resultant coverages.

## 4.3 Database: Justification and Sources

The perspective adopted for all the databases described in this research is that of a 'geographer'. In other words the spatial distribution of environmental factors is foremost rather than the exact concentration, dosage and pathways of the pollutants concerned.

### 4.3.1 Radiation

The first database discussed is that concerning the impact of radiation. Not simply because it is the leading contender in terms of environmental causation of childhood cancer, but also the fact that as a database it encompasses most of the issues raised so far, ie. (i) It does not have a singular source, there are man-made and natural emitters of harmful radioactive material. (ii) The research to-date is contradictory although more substantial than other causal factors investigated. (iii) It demonstrates the importance of both direct and surrogate characteristics of database design when building up a GIS application.

**Table 4.1:** Environmental Databases: Quick Reference

| COVERAGE | CHAPTER REFERENCE | ORIGINAL SOURCE | SOURCE YEAR |
|---|---|---|---|
| 1) Acute Lymphoblastic Leukaemia | 3 | Childrens Malignancy Disease Registry | 1968-1990 |
| 2) Power Stations | 4.3.1.1 | Central Electricity Generating Board | 1990 |
| 3) Special Rad. Sites | 4.3.1.1 | Department of the Environment | 1988 |
| 4) Solid Geology | 4.3.1.2 | British Geological Survey | 1937/46/52 |
| 5) Background Radiation | 4.3.1.2 | National Radiological Protection Board | 1989 |
| 6) Overhead Power Lines | 4.3.2.1 | North Eastern Electricity Board | 1979 |
| 7) Substations | 4.3.2.1 | North Eastern Electricity Board | 1979 |
| 8) Road Network | 4.3.2.2 | Bartholomews | 1989 |
| 9) Railways | 4.3.2.2 | Bartholomews | 1989 |
| 10) Incinerators/Waste | 4.3.2.3 | Aspinwall and Co. | 1987 |
| 11) Waste Disposal Sites | 4.3.2.4 | Aspinwall and Co | 1987 |
| 12) Mine and Quarries | 4.3.2.4 | Mines and Quarries (HMSO) | 1988 |
| 13) Smoke concentrations | 4.3.2.5 | Warren Springs Laboratory | 1988 |
| 14) Sulphur Dioxide concn | 4.3.2.5 | Warren Springs Laboratory | 1988 |
| 15) Landuse | 4.3.3.1 | Agriculteral Census | 1989 |
| 16) Bracken | 4.3.3.1 | Lunn | 1976 |
| 17) Estuaries | 4.3.3.2a | Ordnance Survey | 1977 |
| 18) Stream Network | 4.3.3.2b | Bartholomews | 1989 |
| 19) Rainfall | 4.3.3.2c | Meteorological Office | 1989 |
| 20) Population counts | 4.3.4.1 | Government Census (OPCS) | 1981 |
| 21) Ward boundaries | 4.3.4.1 | ESRC Data Archive | 1981 |
| 22) County boundaries | 4.3.4.1 | Ordnance Survey | 1977 |
| 23) Social Class | 4.3.4.2 | CDMS Marketing Services | 1987 |

**NOTE:** CDMS - Credit and Data Marketing Services Ltd
OPCS - Office of Population Census and Surveys
HMSO - Her Majesty's Stationary Office
ESRC - Economic and Social Research Council

#### *4.3.1.1 Man-made sources of radiation*

Man-made sources of radiation fall into two categories. The first has a global effect which includes fall out from nuclear explosions such as Hiroshima (1945) and more recently Chernobyl, USSR (1986). These expose a high proportion of hapless people to large doses of radiation over a fairly short period of time. The increased risks of developing cancer can be best understood from the study of survivors of these incidences. For example, out of 54 000 people close enough to receive excess doses in the explosions of Hiroshima and Nagasaki, 320 died of cancer before 1982. Leukaemia was the first cancer to appear, with excesses of the malignancy continuing for up to 20 years after the event (Cooke et al 1986). The latent period for other cancers is much longer and therefore the full risk involved is still uncertain.

The global effects occur in isolated incidences but are devastating events in terms of the impact upon man, the second category of localised sources which includes sources of nuclear power and similar outlets is considered to be far more important. The research into the effects of very low dosage levels though is on-going and highly contentious. Publications such as the Black Report (1984) investigated the incidences of childhood cancer in Seascale which were quoted to be five times higher than the regional average. However, the result of this study, and follow up reports, was to state that the proximity to a nuclear installations could not be proved as the cause of childhood cancer. In addition the government maintain that radioactive by-products and effluent from nuclear power stations are not hazardous to public health, stating that exposure levels to the surrounding population are well below the acceptable International Commission on Radiological Protection recommendations, set at an average of 1mSv per annum.

In spite of these claims the proximity to nuclear installations remains a key factor for concern and even the UKs' National Radiological Protection Board (NRPB, 1987) have intervened in a bid to dismiss such fears. Between 1985 and 1988, the NRPB published results from a series of studies into the effects of radiation and the estimated radiation doses from discharges at Sellafield (Stather 1986) and Dounreay (Hill 1986). Both studies indicated that it was most unlikely that radiation doses released into the environment and taken up by the individual through the atmosphere, sea, freshwater, diet or a combination of these would have contributed to the increase

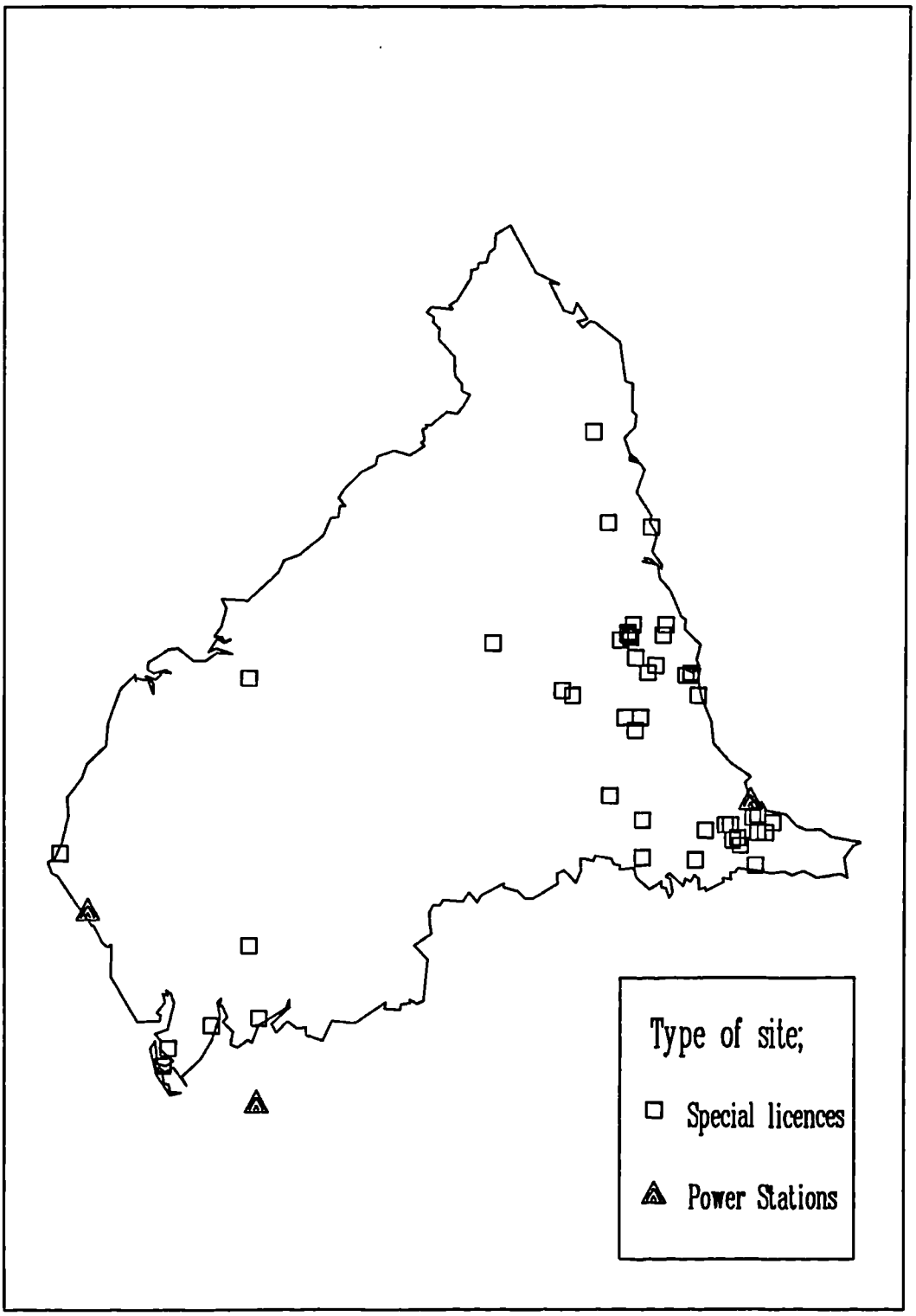
in leukaemia incidences recorded in the local communities of Seascale and Thurso respectively. Even in the absence of any obvious causative agent it is still disturbing to find two areas of increased incidences near sites of the only two reprocessing sites of spent fuel in Britain (Darby 1987).

The latter therefore demands further investigation into doses of radiation which are being received by the general public and the associated risks of cancer. In addition, it may be far more realistic to have a safe dosage level which takes into account the vulnerability of children to hazardous elements as opposed to the current dosage level for adults. Another angle of approach may be to look at the possible combination effect of several hazardous sources operating in the same geographical region, each of which would serve to 'top up' an individuals received dosage of radiation. This theory of causation would render one to one searches into the relationship between ALL and radiation sources insufficient, and thus other sources of man-made radiation should also be included as separate coverages.

Other outlets that deal with radioactive substances must conform to the controls set out by the government on the handling, storage and disposal of radioactive waste. Thus these can easily be traced by their special site licences, which under the Radioactive Substances Act 1960 (RSA60, HMIP, 1988)) must be recorded. The required spatial referencing for a GIS can be derived from the addresses of the establishments. In addition, disposal sites that deal with any radioactive effluent should be included, particularly since recent statistics show that the total hazardous waste in England designated as 'hazardous waste' has risen from some 1.8 to 2.2 million tonnes between 1988 and 1990. Spatial references for these disposal sites can be found in the directory published by Aspinwall and Co. (1987), which contains a general description of the type of waste that a particular site handles and a postcode. Figure 4.2 shows the distribution of these point sources of radiation.

Additional characteristics of man-made sources of radiation should also be noted even though they cannot be incorporated into a GIS database. These include occupational exposure to radiation in industry and medicine, such as x-rays and radiotherapy treatment. All of these form an integral part of society and serve to expose individuals to yet more low doses of radiation.

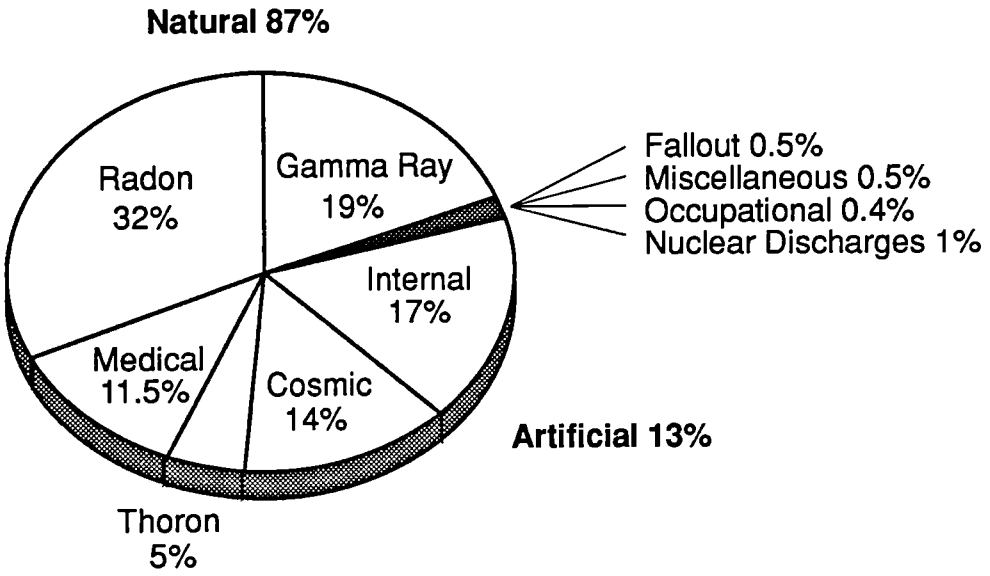
Figure 4.2: Man-made sources of radiation



*4.3.1.2 Natural Sources of radiation*

The irony surrounding radiation exposure though is that in comparison to man-made sources, the effects of naturally occurring radiation upon the population are far more dominant, as reflected in Figure 4.3. This, coupled with the increased incidences of ALL in the Gateshead area where there is a notable absence of man-made radioactive sources, demands that other environmental factors should be reviewed. The first of these being the impact of naturally occurring radiation in the environment.

**Figure 4.3** A pie chart showing the difference in percentages, between manmade and natural sources of radiation




Source: Radiation Risks in Perspective (CEGB, 1988)

The natural sources of radiation come in several forms, all of which can vary greatly over space. The main source is that of; **Terrestrial radiation** which is usually connected with radon daughters and involves the naturally radioactive gas Radon 222 produced by the decay of trace quantities of uranium. This is transferred through the pores of rocks and depending on the local geology may eventually seep into houses

situated on the surface and accumulate indoors (Bradley 1987). Consequently, since soil and rock compositions vary greatly over space, it is not surprising to find that the absorbed dosage of gamma rays from this source varies from one area to another.

Results from recent studies (NRPB, 1987) suggest that the highest radon concentrations are generally associated with intrusive igneous rocks, especially granites found in South East England. Mean concentrations in dwellings built on granite in Cornwall and Devon proved to be more than ten times that of the national average. High radon concentrations however, are not exclusively associated with igneous rocks, many different sedimentary formations (with the exception of clays) were also associated with higher concentrations (NRPB, 1987)

Until recently the 'safe action' level was considered to be a dose of 20 mSvy-1 or 400 Bequerels. This has subsequently been reduced to only 200 Bequerels (NRPB, 1990), bringing other areas, such as Northamptonshire into consideration. Table 4.2 demonstrates that even with the reduction in levels the Northern Region still falls well below the 'supposed' hazardous radon daughter levels.

**Table 4.2** Radon Daughter concentrations

| County | No Sampled | Housing stock | Mean Concn Bqm-3 | Approx No dwelling ge 400 |
|---|---|---|---|---|
| **Cornwall** | 19 | 164k | 110 | 8000 |
| **Durham** | 25 | 228k | 24 | 20 |
| **Cumbria** | 11 | 182k | 21 | 0 |
| **Cleveland** | 23 | 212k | 20 | 0 |
| **Lancaster** | 67 | 520k | 15 | 0 |
| **Northumberland** | 12 | 114k | 14 | 0 |
| **Tyne & Wear** | 58 | 430k | 12 | 0 |

Extracted from the Mean indoor activity concentrations of radon gas in areas of the UK from the National Survey Bqm-3 in NRPB(1986)

In addition these terrestrial radiation levels can be modified by 'man' through the building of houses made out of local stone. This can have the effect of either shielding

the inhabitants from natural radiation sources, or conversely exposing them to increased doses because the radiation becomes trapped inside the building.

Recent research has linked radon daughters to an increased risk in cancer (Henshaw, 1990), however there are very few detailed and specific measurements to reinforce these studies and the statistics and methodologies actually employed are questionable. Despite this, radon daughters remain a key environmental data source, and two coverages can be used as influential proxies. The first coverage is solid geology. This provides information on the rock distribution of the region. From this, areas of particular interest can be identified, such as areas of acidic igneous rocks with high silicon content, which are perceived to be very important in terms of radioactive emitters, see Figure 4.4.

This geological database was created by digitising a number of 1:253440 hard copy maps, compiled by the British Geological Survey (see Chapter 6 for more details on the digitising process). The use of several map sheets surveyed by different people at different times and the aggregation of small areas of lithology into one geological type, ie. limestone, may lead to some data quality problems (discussed in Chapter 10). These problems should not be ignored but it is considered that any relationship between ALL and this database should be suitably reliable so long as special consideration is given to those cases which are found at the edge of lithological boundaries.

The second source for background radiation comes from that of outdoor gamma ray readings, which can be used to complement the geological coverage. These were made available in the form of point source measurements supplied by the National Radiological Protection Board (1989). Measurements were taken at every 10 kmsq based on the Ordnance Survey National Grid, and at 1m above the ground. Any important localised variations therefore could not be observed and in order to produce a more applicable coverage for this research these points were converted into a surface to represent background radiation throughout the Northern Region, Figure 4.5.

Other natural sources, noted in the pie chart in this section include **cosmic radiation**, which reaches the planet from the depths of space. However, the majority of this energy is removed by the upper atmosphere, and a final source is **internal radiation.** This constitutes approximately one fifth of all natural radiation dosage arising from

Figure 4.4: The geological coverage for Northern England highlighting igneous rocks

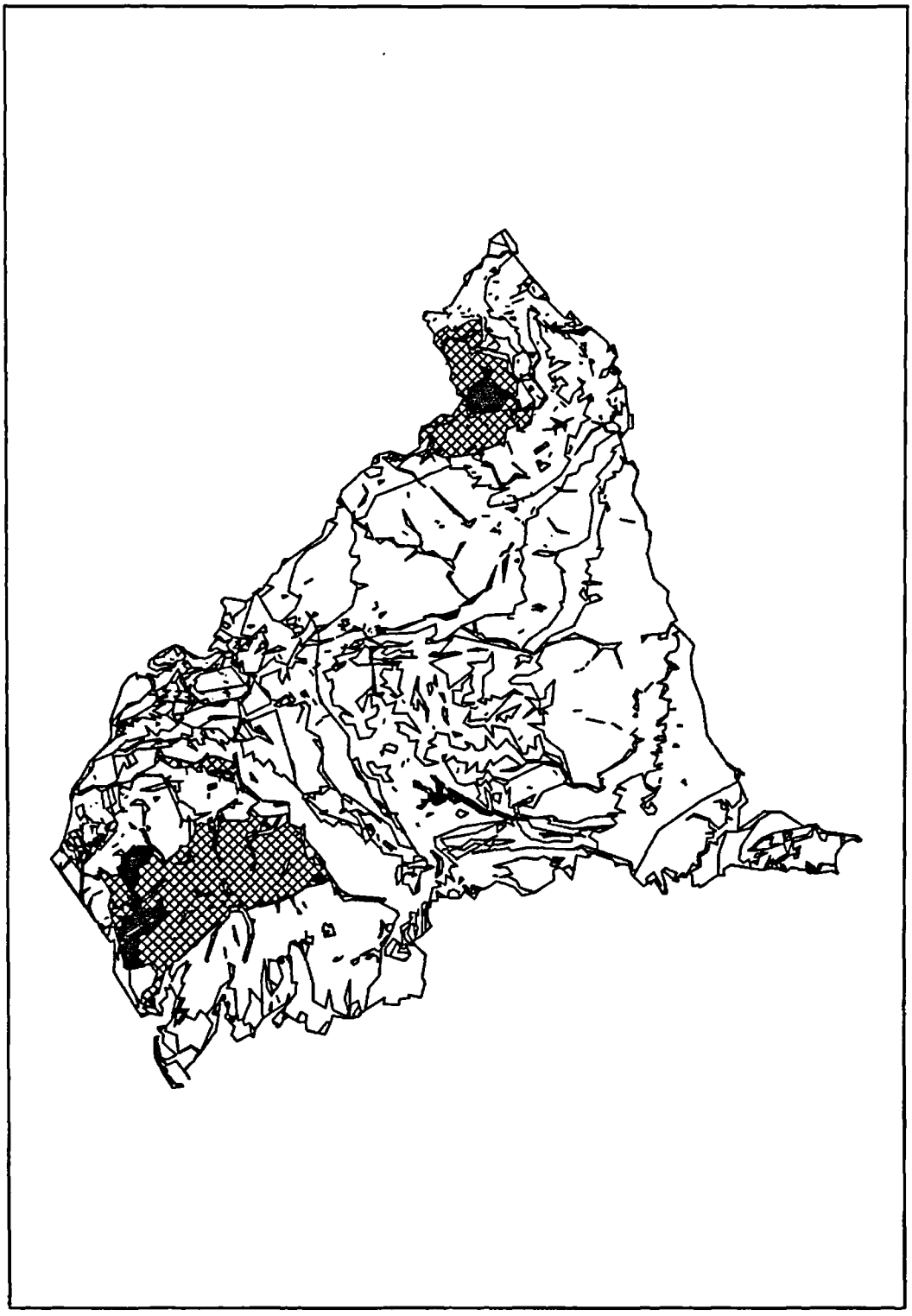
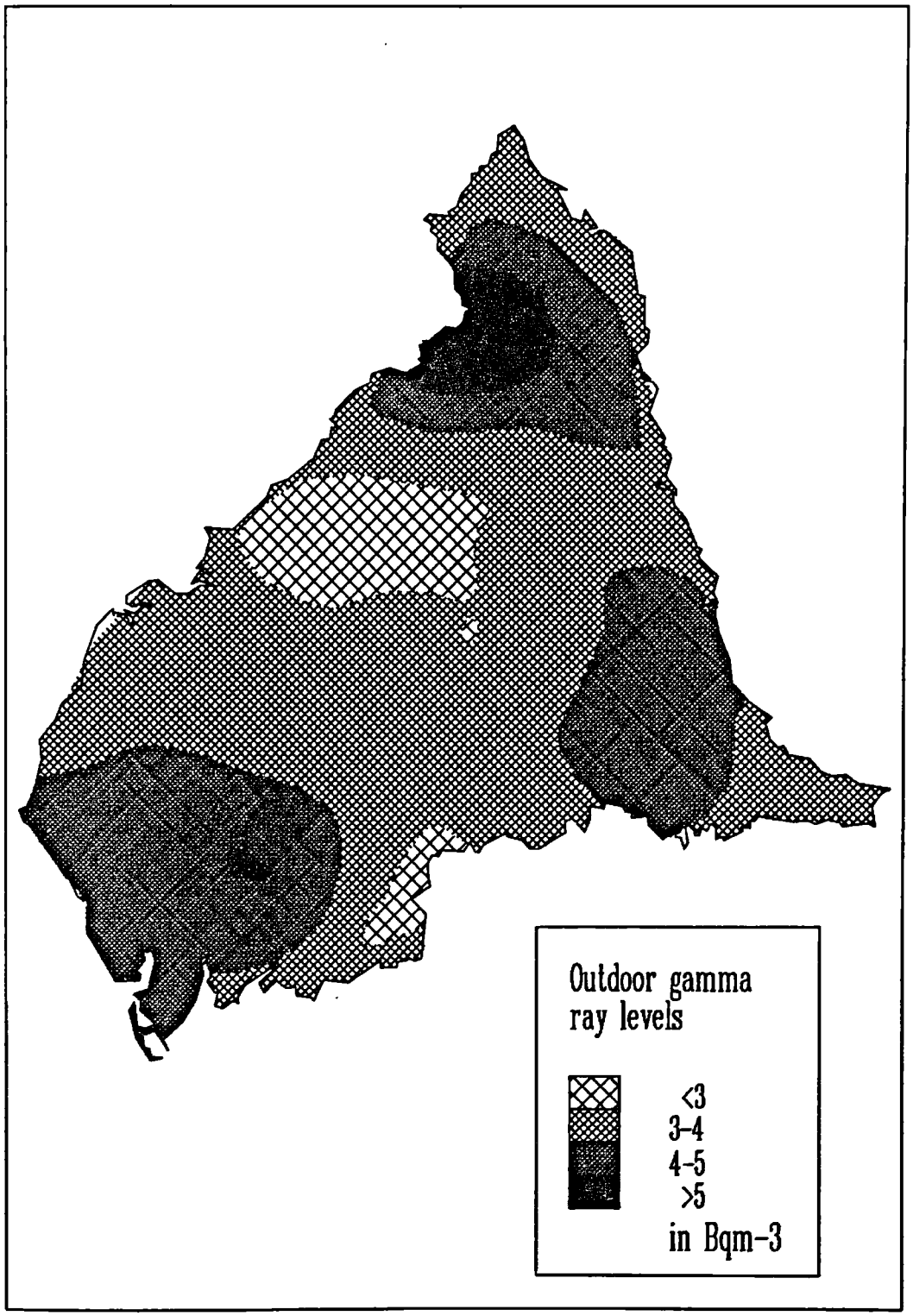


Figure 4.5: Background radiation levels in Northern England



radio-isotopes already present in living tissue (Rumsey, 1973). However, unlike terrestrial radiation these two factors are not easily recorded within a GIS framework.

### 4.3.2 Man-made environmental sources

The importance of man-made sources are investigated in this section. Most of the research on ALL to date focuses on links to these types of environmental factors.

#### *4.3.2.1 Electromagnetic fields*

In terms of electricity generation hazardous sources include electricity transmission lines and points which network the country. These have been placed high on the list of 'suspected' causes of ALL due to the impact of electromagnetic fields associated with overhead power lines and substations, because whenever electricity is used electric and magnetic fields are produced. These fields are undetectable by human senses but can be measured with special equipment. Both electric and magnetic fields are strongest immediately beneath a power line and diminish rapidly with distance. For the highest voltage lines, (400 kilovolts (kv)), the fields merge into normal background levels at a distance from two to three hundred metres, whilst local low voltage electricity distribution lines have a range of ten to twenty metres. However, these fields are not confined to power line supplies, trace doses can also be received from everyday objects used in the household, such as washing machines, electric blankets, and television sets. There is an argument therefore for a cumulative effect of electromagnetic doses. In addition, there are the doses from neighbourhood electricity substations, the fields from which are said to fall to background levels within five metres but this may vary through time and have a directional component. It is an interesting fact that many of these sources of pollution tend to be located in areas dominated by children, for example near schools and play grounds, because they were perceived to be completely harmless.

A possible leukaemagenic effect of high voltage equipment was raised some time ago by Wertheimer and Leeper (1979) following a study in Denver City, Colorado. The investigation was a case-control study of children who had died from childhood cancer. The actual exposure of the children to electromagnetic fields was inferred from the type and proximity of neighbouring overhead electrical distribution lines.

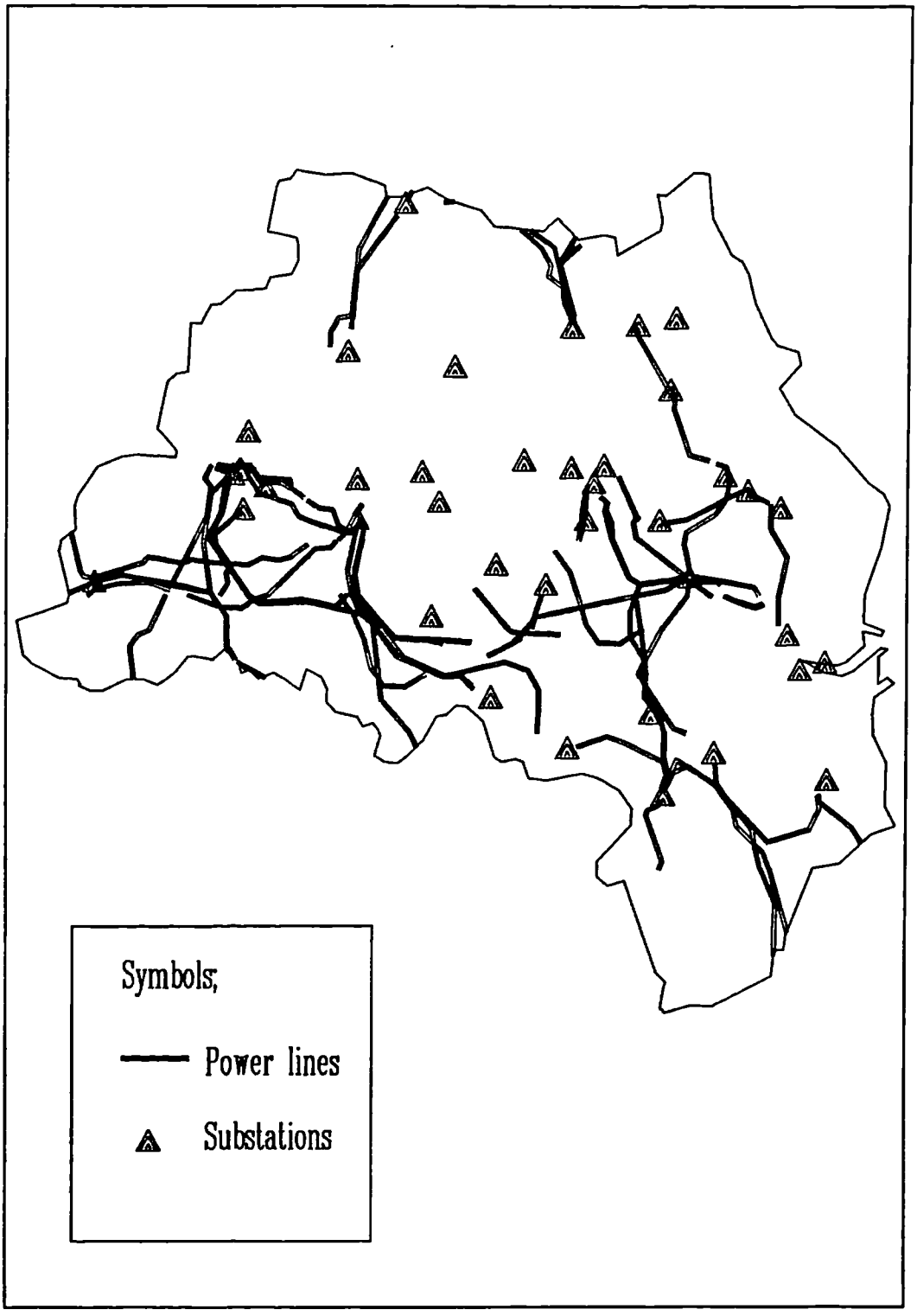
There were similar findings in Stockholm (Tomenius, 1986) however, Fulton et al (1980) carrying out a study in Rhode Island found no such evidence. Again the source of causation proves to be very contentious.

Despite the speculation which surrounds the ALL causation debate, the Central Electricity Generating Board (CEGB, 1988) suggested that there is a danger of drawing premature and wrong conclusions from statistical associations such as those mentioned in the latter studies (Wertheimer et al 1979). They did however support four studies designed to investigate the effects of high electric and magnetic fields on surrounding populations. These looked at the impact upon staff, patients and other people in close contact with electric and magnetic fields. The most recent, by the Leeds Leukaemia Research Fund Project, studied childhood cancer cases found within the proximity to overhead power lines in the Yorkshire Health region, over the period 1970 to 1979. The preliminary results though showed no significant findings (Myers et al, 1990).

In order to test the effect of these fields data were acquired from a case study (Raybould 1989), and for the purpose of this exercise are limited to the county of Tyne and Wear. The location of substations were derived from the North Eastern Electricity Board (NEEB) in the form of hard copy maps, at two scales 1:1250 for the extreme urban areas of Newcastle upon Tyne city centre and 1:10560 for the remainder of the county. The resultant coverage includes substations and overhead power lines of voltages greater than or equal to 20kv, shown in Figure 4.6. This is because; (i) these are the most likely to have an impact upon health and (ii) despite increased prevalence in urban areas the lower voltage power lines were poorly represented on the documentation made available. The linear power lines and substation centroids were both digitised to a resolution of one metre. Although the accuracy of these coverages will also be subject to human and hardware problems, as well as a lack of knowledge as to the criteria under which the maps were originally compiled by NEEB.

An important assumption made in the acquisition of these data, and for that matter many of the other datasets compiled for this research, is that the features recorded were presumed to have been present throughout the study period, 1976 to 1986. In cases where longer time periods are involved and/or the data will be used in major policy issues, such as the European HEGIS, attention to temporal detail

Figure 4.6: The distribution of overhead power lines and substations in Tyne and Wear



would be obligatory. In addition information on the exact dose rate and measurements of fields would be advantageous, although this would involve more time and resources. At this point it becomes necessary to weigh up the benefits which can be accrued from the inclusion of this extra information against the extra resources that would be required.

The latter provides a clear indication of how man's activities in todays society can result in the production of dangerous elements that ultimately effect human health. However the generation of electricity and its perceived risks is simply one example. The following will serve to highlight other major sources of hazardous material which can be linked detrimentally to child health, if not to ALL directly.

#### *4.3.2.2 Road Network*

The establishment of this database for a HEGIS application provided a classic example of how features such as the road infra-structure could be employed as a surrogate for other unmeasured aspects of the environment, namely that of the atmospheric pollution associated with motor vehicles. This research particularly focused on the generation of two chemicals which are associated with car exhaust fumes and considered extremely hazardous to public health, lead and benzene.

Although lead is not directly linked to the causation of ALL evidence shows that severe damage can result from the intake of lead long before any obvious symptoms of poisoning start to show. Children in particular are more vulnerable to this type of poisoning because their brain and nervous systems are still developing. They also absorb lead approximately five times quicker than adults, and once there it tends to be very difficult to eliminate, thus creating a cumulative effect (Rutter, 1986).

The actual measurements of lead concentration in the atmosphere are limited despite the setting up of a Working Party in 1974 to review the situation. This Party specifically stated that monitoring of the atmosphere should take into account lead concentrations over a range of urban and rural locations. However, the result was the setting up of just 21 sites throughout Great Britain, of which only three can be found in the Northern Region (McInnes, 1986). The extent of, or lack of, information available is illustrated in Tables 4.3a and 4.3b.

**Table 4.3a** Monitoring Lead, Site Details

| Site Name | Location | Type of Site | Grid Reference | Start Date |
|---|---|---|---|---|
| North Tyneside | Holy Cross School | Urban | 4309 5677 | June 1984 |
| Newcastle | Gosforth High School | Urban | 4247 5688 | July 1984 |
| Windermere | Wragmines Lancs | Rural | 3362 4974 | 1972 |

**Table 4.3b** Annual Statistics for Lead concentrations ($ngm^{-3}$) in Northern England 1976-1985

| Site | '76 | '77 | '78 | '79 | '80 | '81 | '82 | '83 | '84 | '85 |
|---|---|---|---|---|---|---|---|---|---|---|
| Windermere | 75 | 47 | 49 | 40 | 46 | 40 | 46 | 45 | 48 | 35 |
| N. Tyneside | | | Missing | | | | | | | 290 |
| Newcastle | | | Missing | | | | | | | 180 |

Extracted from Tables in the Department of Trade and Industry Report (McInnes, 1987)

The second of the chemicals noted, benzene, has been recognised as an important leukaemogen. Recent studies demonstrated that workers occupationally exposed to benzene showed a five fold increase in mortality risk for all leukaemias, and a ten fold increase for myeloid leukaemias (Brandt 1978). Although workers in direct contact with petrol are more susceptible, even a trace amount of benzene ingested by children should not be ignored. As with lead there is no means of directly representing the impact of benzene upon the child's micro-environment, therefore the regional road network was adopted as a surrogate for the emissions of lead and benzene into the surrounding environment. This dataset, based on the digitised version of the

Bartholomews motorist maps (1990), at a scale of 1:253440, was made available to the University through the academic Committee of Higher Education Software Team (CHEST). It included a network of up-to-date digitised maps at a resolution of one metre and contained information on the status of the roads (A, B, M and so on).

The determination of the overall area of impact of car exhaust fumes upon the surrounding environment was problematic. Since traditional models available for representing the distribution of atmospheric chemicals and heavy metal deposition tend to be extremely specific and complex, none were considered suitable for defining areas of effect from an effluent source. In addition a review of these models in urban areas suggested that they were inadequate for this type of study and were found to have 'very limited general applicability' (Zib, 1977). Thus the areas over which lead and benzene chemicals were presumed to have an effect were based on subjective decisions and previous field studies carried out in Tyne and Wear (Raybould, 1989).

The field studies found that the maximum width around a road, over which lead particles were deposited, was 250 metres and that this was constant irrespective of the load of the road (Raybould, 1989). The actual concentrations within that 250m corridor would be higher with increased traffic density. For this research therefore corridors were calculated at 250 metres plus or minus 100 metres. In addition, greater attention was given to those areas which overlapped and junctions where the traffic produced increased amounts of heavy metals, through low gear work, acceleration/braking and stagnation of vehicles. Figure 4.7 shows a sample of the road network for the Northern Region. Chapter 6 will outline in more detail the necessary GIS manipulation techniques needed to isolate areas of increased pollution.

From the Bartholomews dataset another linear transport feature could be acquired, that of the distribution of railways in the study area, shown in Figure 4.8. This has not been looked at in terms of ALL or child health in general. However, who is to say that factors such as the type of load carried by trains and/or other characteristics such as granite chippings or effluent associated with the siting of this form of transport, do not have some influence upon the surrounding environment. This coverage therefore provides the best example of GISs ability to store all types of spatial data available, as well as to test for the most unlikely of proxies for environmental factors and their possible relationship with ALL.

Figure 4.7: A sample of the road network for the study area. A proxy for lead pollution

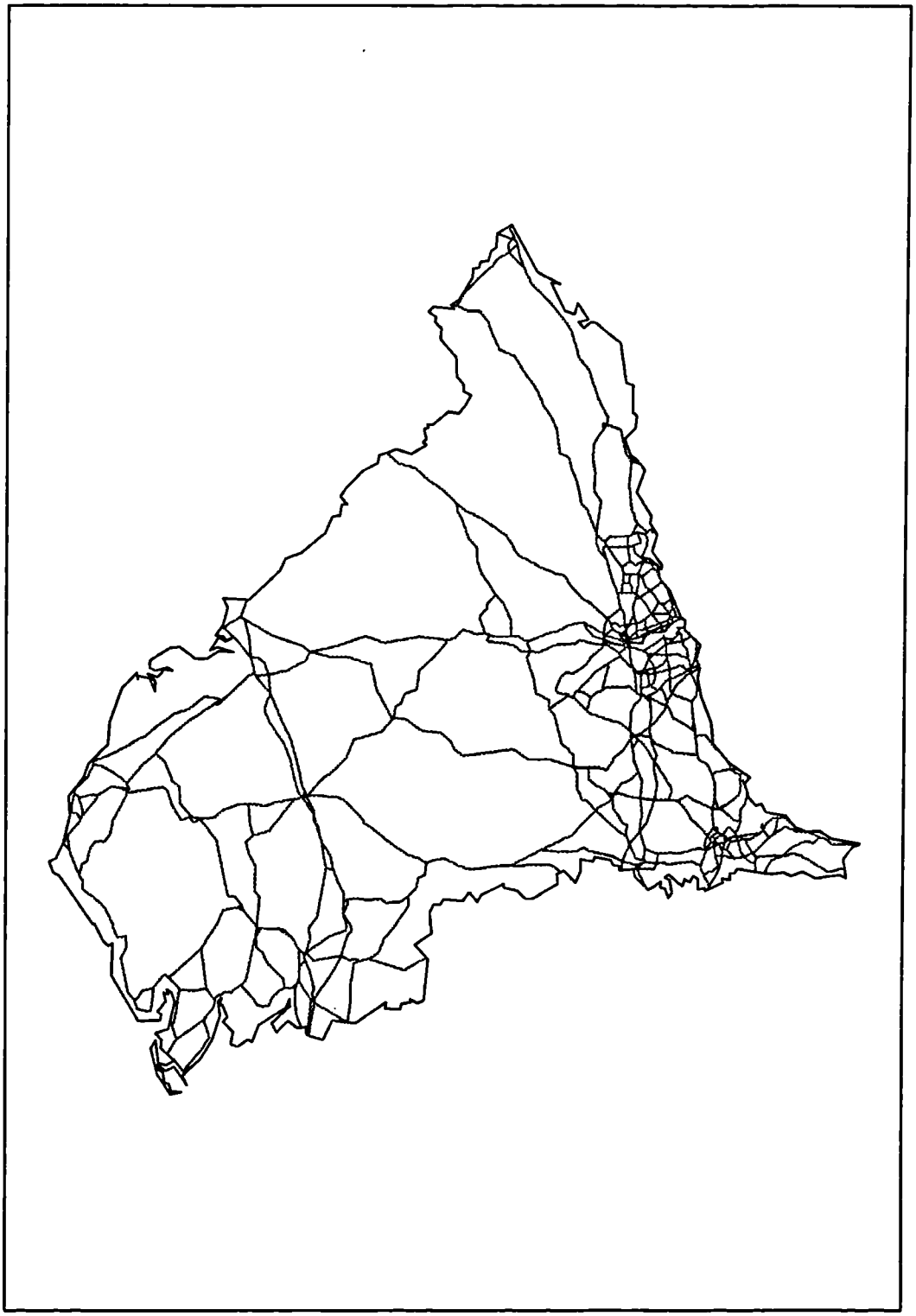
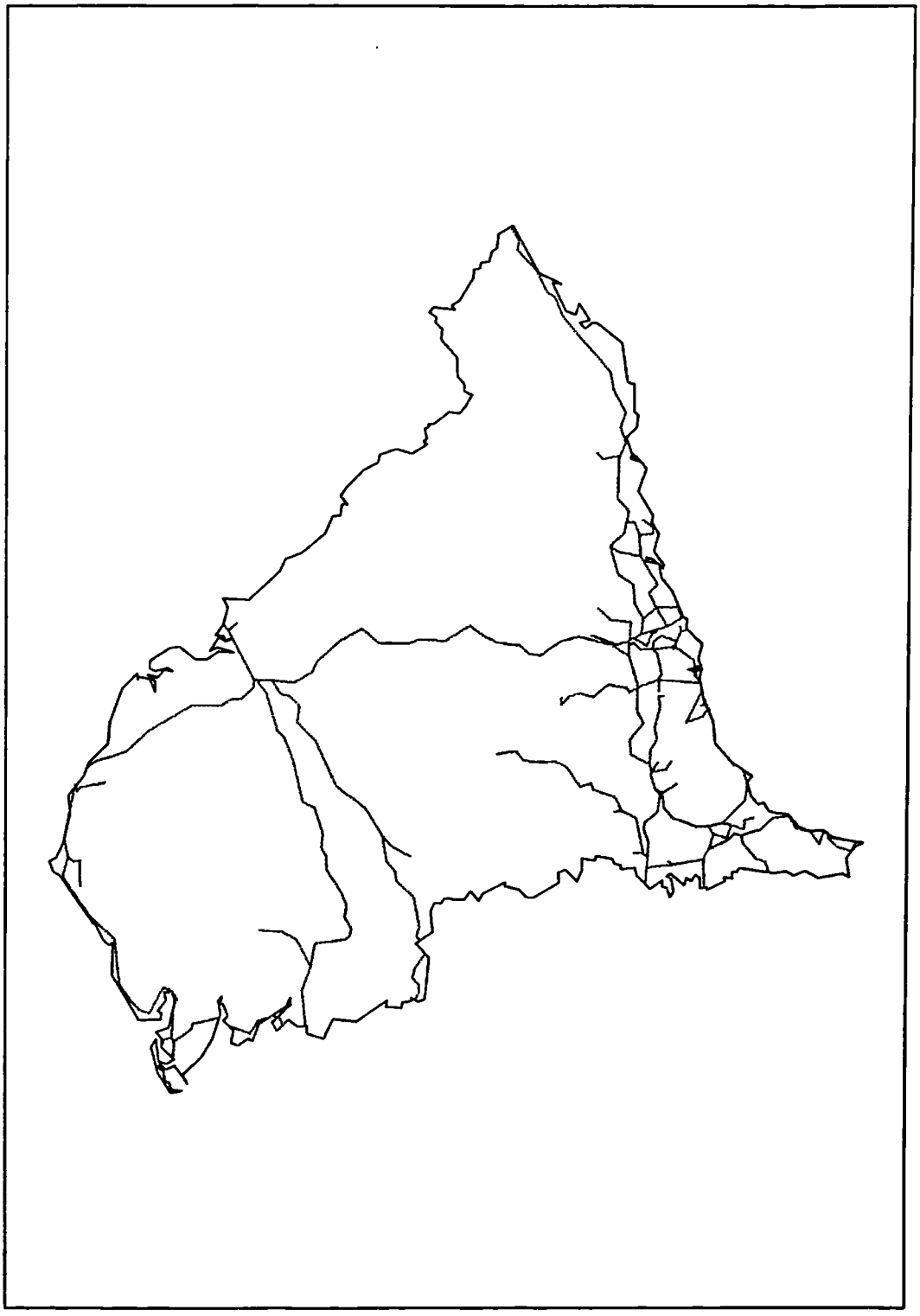


Figure 4.8: The railway network



#### *4.3.2.3 Incinerators*

Atmospheric pollution is not simply confined to the linear corridors of the road or rail network. There are a number of other point sources which are of equal importance These include; waste disposal sites, incinerators and manufacturing installations dealing with volatile chemicals. These form key areas for concern, the incidences of ALL in the Gateshead area have been attributed by the public as a result of the inadequate working practices of the Wrekenton incinerator found downwind from the observed cluster.

The most documented link to health and incinerators comes from the review of the practices of the ReChem plants in Bonnybridge and Pontypool. The Bonnybridge plant was closed in October 1984 due to local allegations that 'dioxins' and polychlorinated biphenyls (PCB's) emissions were caused by inadequate incineration and this was linked with various adverse human and veterinary health problems that had occurred in the vicinity of the plant. The Lenihan Report (1985) was unable to confirm these effects, although it was heavily criticised for being biased (The Independent, 1989).

Large municipal incinerators therefore are a major source of localised and widespread air pollution. They often burn a wide range of materials, many unknown, using fairly primitive technology. The content of the exhaust gases is not monitored, thus the actual chemical content of effluent is a matter of theoretical guess-work. Essentially waste in the UK is disposed of using the 'best practicable means', in terms of cost and environmental considerations. What is meant by 'best practicable means' and how much emphasis is put on cost as opposed to the environment is not outlined by the government and is left as a matter for local negotiation. This will change in 1995/6 when the European Regulations come into force.

At present two incinerators within the study region are being reviewed. The Portrac incinerator, Stockton, will be closed down for modifications because it is producing over ten times the hydrochloric acid levels allowed and considerable amounts of heavy metals and sulphur dioxide (Northern Life, 18/6/91). The incinerator located in North Shields also requires up-dating in order to reduce air pollution in the

surrounding environment and the incinerator in South Shields has already been demolished (1991).

A coverage to represent incineration sites was derived from the postcode of addresses, converted to a 100m grid reference (Aspinwall and Co, 1987). However, as with the road versus atmospheric pollution problem the total area of impact of these sites upon the surrounding population is subject to theoretical estimates. For this research, several areas of effect were created using GIS. This involved producing circular zones of impact at distances of 1km, 2km and 5km respectively. Further modifications could be made on these zones if any significant results were obtained and this would probably take the form of searching for a directional component.

#### *4.3.2.4 Other important sites*

There are other ways in which waste can be disposed of besides incineration, and these include dumping of material at designated sites on land and at sea. Initially, these may appear to be very specific in their location and as such have little impact upon the surrounding environment, but one cannot ignore possible indirect effects. For example, the decomposition of material from these sites can produce harmful gases which may spread into the surrounding environment and subsequently build up in homes. Alternatively, poor management of landfill sites may lead to water percolating through the waste, absorbing harmful chemicals which are ultimately carried into water supplies and passed on to the individual. There is no statistical evidence to insinuate a link between these types of locations and possible health risks. Thus the relationship to health has tended to stem more from public speculation, as reflected in major media headlines which appear whenever there is a change in or commencement of waste dumping and/or mining activities;

'Uranium sludge' (5/11/81) at High Urpeth, Urpeth Grange Tip

'Dubious burning of pellets' (7/3/86) Byker Heating Scheme

'Dangerous levels of radiation, from mining carried out 20 years ago' (6/10/88) Garmondsway near Bishop Middleham

This concern extends to mining sites and chemical works. These are automatically targeted to be potentially hazardous because of the nature of the materials being processed, for example volatile chemicals, petroleum based solvents, benzene etc.

The sources of these databases were acquired from directories and secondary documents, including 'A Digest of Authorised Waste Treatment and Disposal Sites in Great Britain' produced by Aspinwall and Co. (1987). This contains information on waste disposal sites compiled from over 8000 statutory permits, site licences and resolutions required under the Control of Pollution Act 1974. This directory also includes details on the type of waste handled by the site ranging from household waste to difficult/hazardous material. All the sites which are registered in this directory are shown in Figure 4.9a, but using the INFO subsystem of GIS these can be manipulated and reselected according to the type of waste disposal ie. incineration, thus Figure 4.9b shows the location of incinerators.

The location of mines can be found through the 'Mines and Quarries Directory' (HMSO, 1988), which enables spatial referencing by postcoded address. This directory is based on 2700 locations, described in terms of name, location, ownership, base geology and the commodity produced, and is assembled from the records of the British Geological Survey, local authority, industrial sources, the Health and Safety Executive and the Inland Revenue Valuation Office (Minerals), these sites are shown in Figure 4.10.

It should also be considered that local pollution might be caused by dumping activity that occurred in the nineteenth century and the first half of the twentieth century. The directories and registries available cannot provide comprehensive or detailed records for all the historic mining sites or dumping grounds, in many cases though those that did exist will probably have been built upon with no historical or visible trace belt anyway. Yet these sites may under certain circumstances still prove to be very hazardous.

The inclusion of mines and waste disposal sites as separate databases offers a clear example of how GIS can be used to follow up the vaguest of hunches, even those expressed by the general public and media. Its importance may only prove to be one of making people 'feel' that something is being done to satisfy their curiosity.

Figure 4.9a: The distribution of all waste disposal sites in the study area

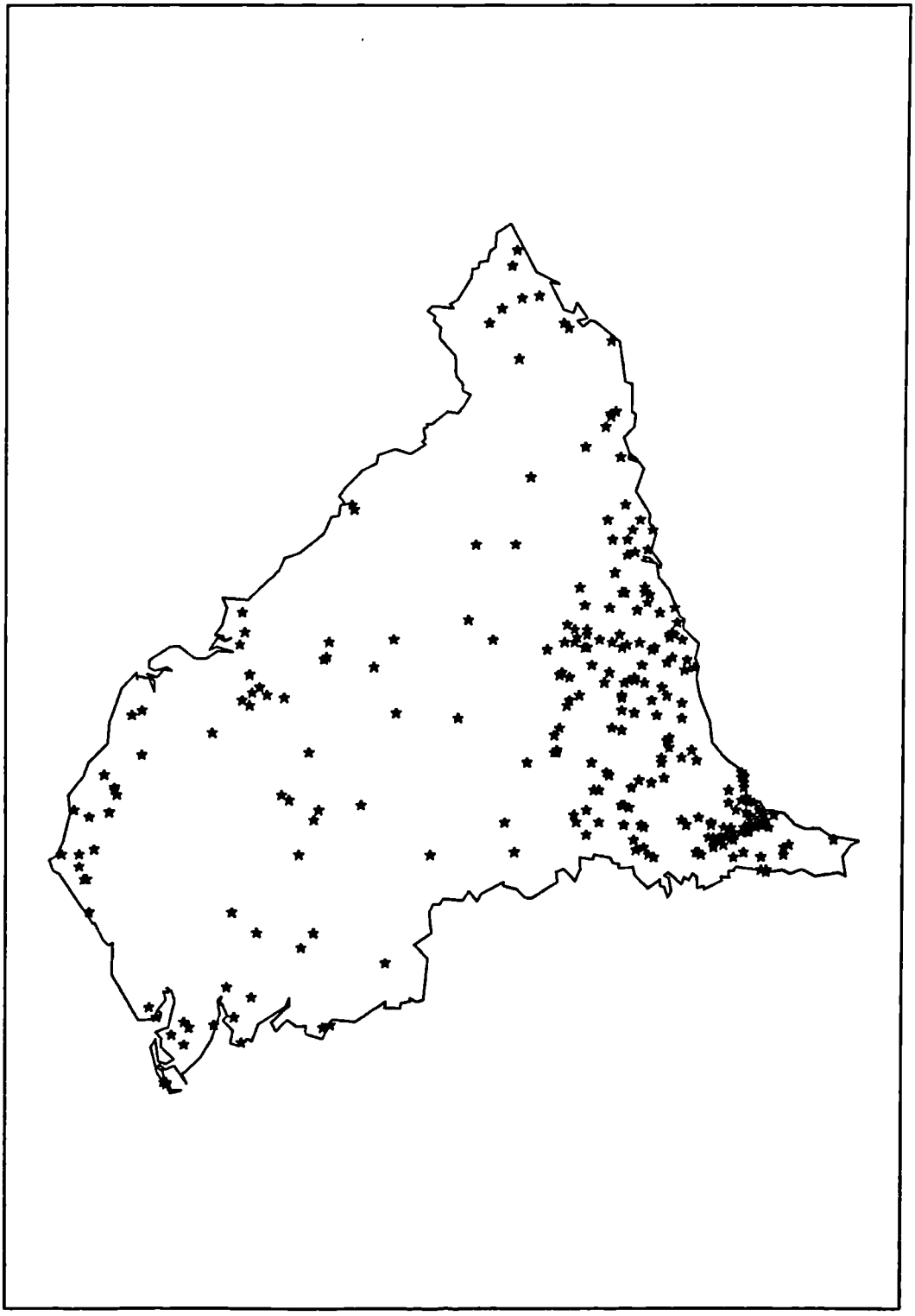


Figure 4.9b: Incineration sites in Northern England

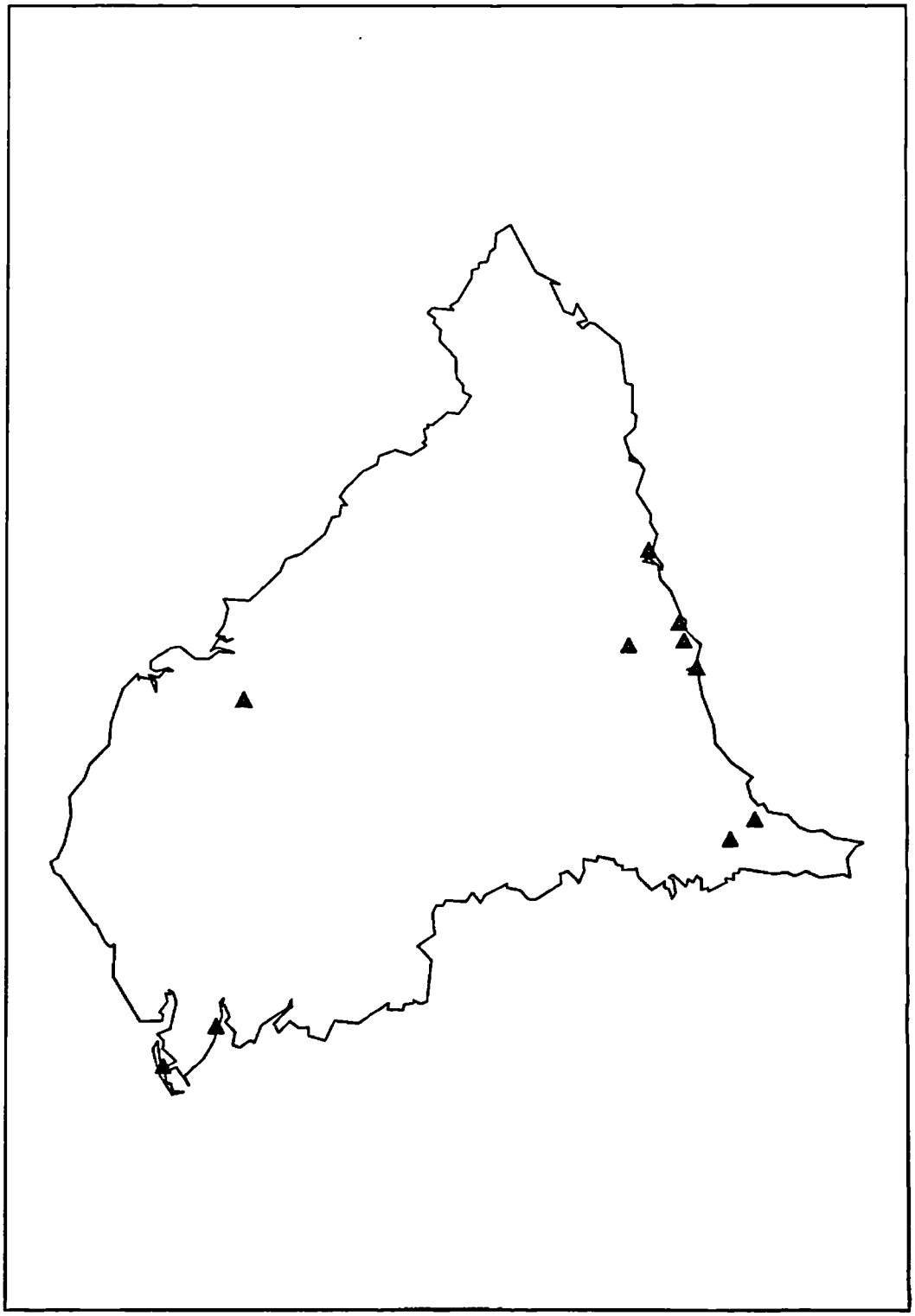
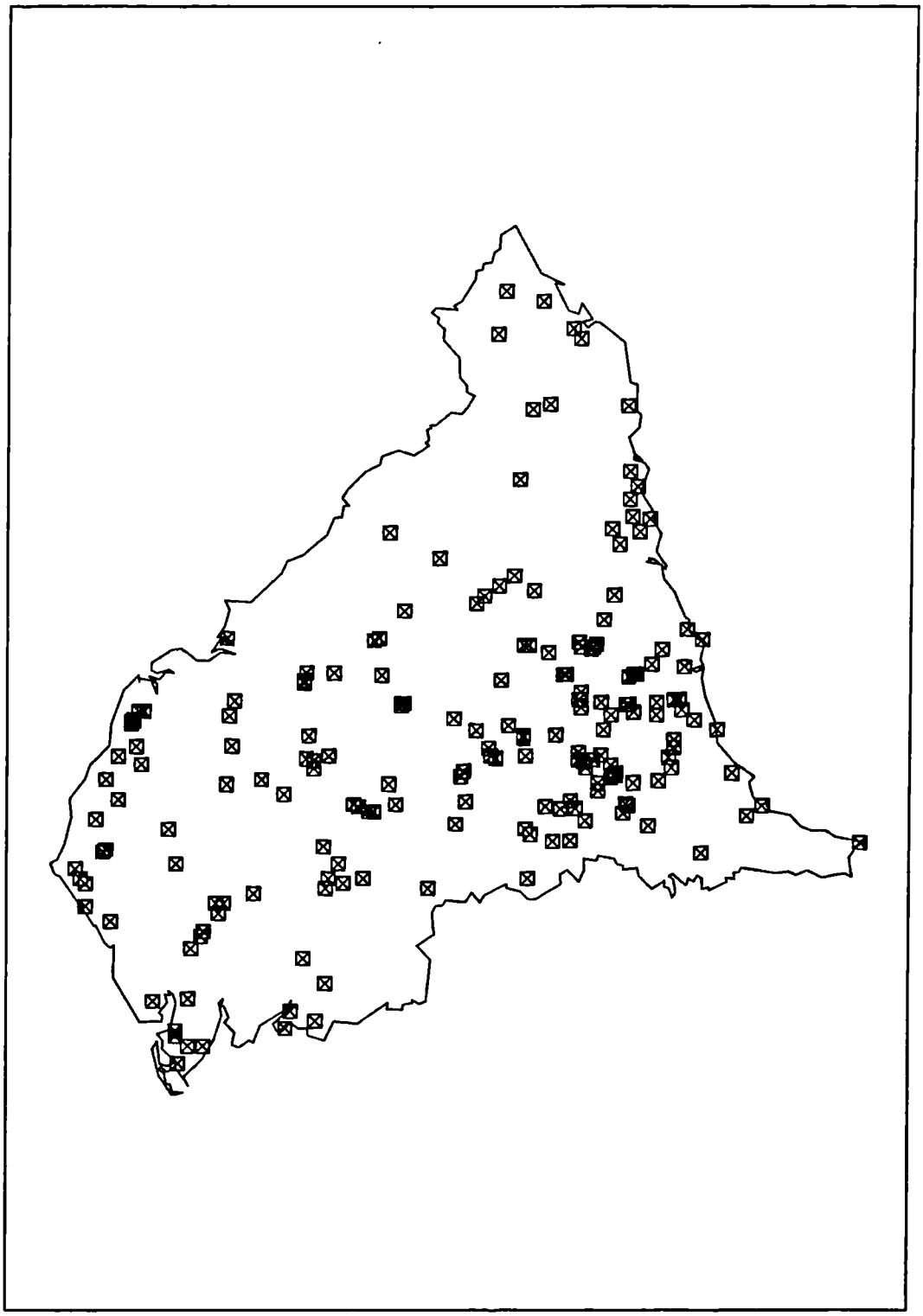


Figure 4.10: The distribution of operational mining sites



Alternatively such speculations may prove to have some significance, you do not know until you look. This role of GIS has yet to be appreciated.

#### *4.3.2.5 Smoke and Sulphur Dioxide concentrations*

It is apparent that man's impact upon any one micro-environment is quite considerable and diverse, even when the elements to be discussed have been limited to those which are GIS compatible. To complete the databases concerning atmospheric pollution coverages representing smoke and sulphur dioxide concentrations were also created. The emissions of these pollutants are one aspect of air pollution which the government does continuously monitor. In the past, these pollutants have been responsible for causing the deaths of thousands of people, the classic example being the London smog of December 1952. However, with the introduction of smokeless zones and tighter legislation on industrial output the levels have been drastically reduced. These controls have not eliminated the impact of smoke and sulphur dioxide completely, and rather like radon daughters, little is known about the effect of smaller doses over prolonged periods.

The relevant information for this type of database was provided by the Warren Springs Laboratory, and covers all the sites found in Northern England that were in operation during the period 1976 to 1986. For both pollutants monthly concentrations were given in micro grams per cubic metre and this was accompanied by additional data on the status of the station ie. urban, rural, industrial etc. Unfortunately, these came in computer readable format without a spatial reference thus in order to make these data GIS compatible a spatial reference had to be obtained from a separate paper-based directory 'The Investigation of Air Pollution, Directory Part 1' (1985) which provided a link in the form of a common station code. It is trivial problems such as these which makes the data capture stage of any GIS application time consuming, and similar problems will be highlighted further in Chapter 5.

In order for this air pollution database to be remotely acceptable for integration into this research application it was subjected to several manipulation techniques. It is considered however that the eventual coverage is a good proxy for smoke and sulphur dioxide concentrations in the region.

The first decision involved the simplifying of the data. This was accomplished by averaging all the monthly concentrations for each recording station throughout the period 1976 to 1986. This resulted in a dataset which contained one set of x- and y-coordinates with a single smoke and sulphur dioxide concentration value. This was considered necessary on two accounts; (a) because not all the sites were continuously in operation throughout this period and; (b) because the vast amount of temporal data was not considered practical at this stage of the development of a HEGIS. Since GISs flexibility does not extend to providing the necessary tools to fully utilise such temporally based data. This limitation will be discussed further in Chapter 7.

Another advantage of simplifying the point dataset was that these new points could now be converted into an air pollution surface using Triangular Irregular Network algorithms available in ARC/INFO (see Chapter 5 for more details). This serves to provide a more realistic representation of the environmental factor. The latter also lead to the refinement of the area of analysis to only cover the county of Tyne and Wear, illustrated in Figure 4.11. This was because the surface for this area was considered to be more accurate due to it containing a greater density of recording stations and some degree of stability in measurements over time.

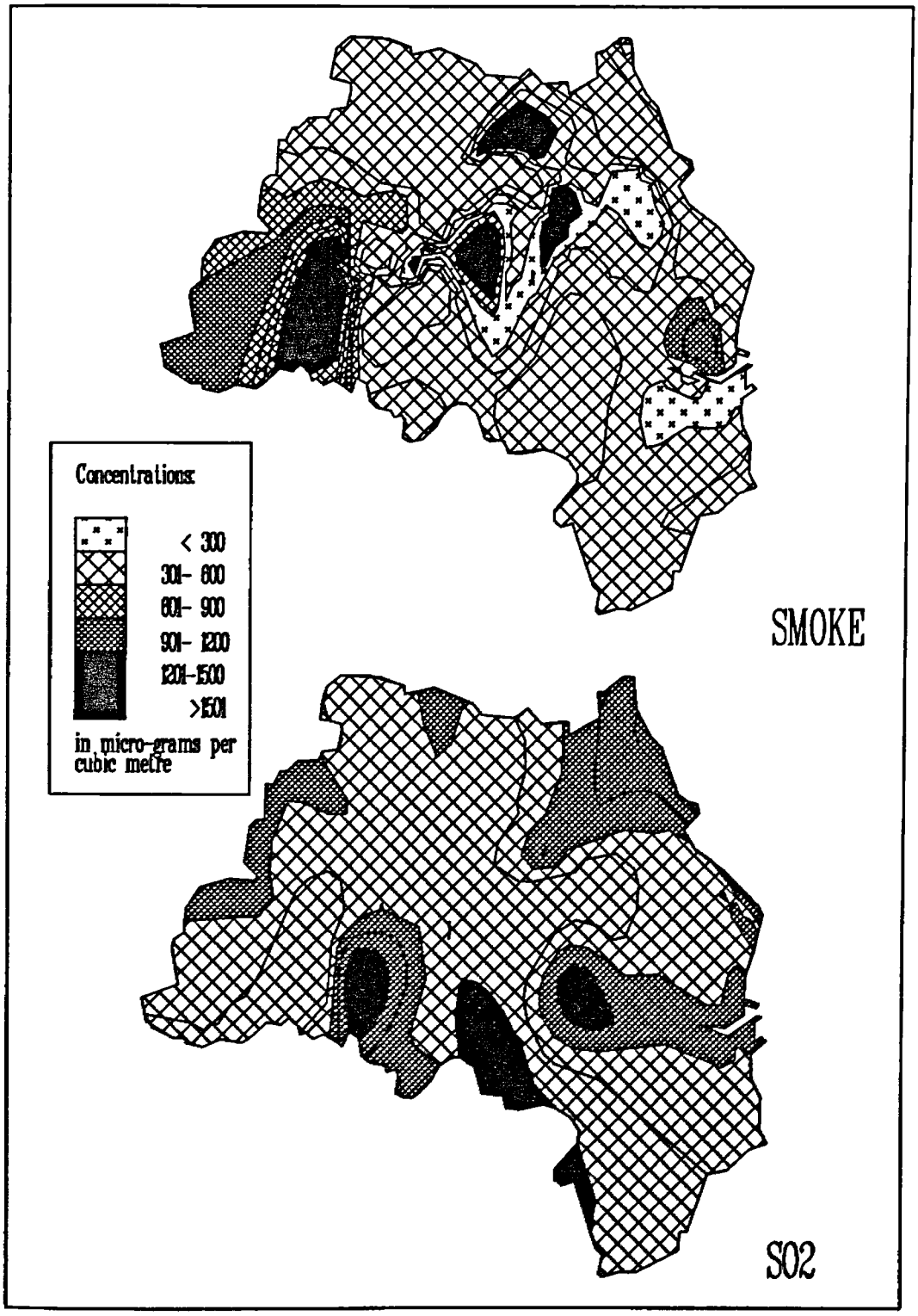
### 4.3.3 Natural environmental sources

Despite this section being headed 'natural' environmental sources, the aspects discussed, such as landuse and water related features could easily have been incorporated into the section on man-made environmental sources by virtue of the fact that these features are often adversely effected by man's intervention. In comparison to the latter databases though, the importance of the natural environment in terms of ALL causation has been the subject of far less research and scrutiny. This section therefore provides justification for the inclusion of certain databases, as well as highlighting GISs capabilities to test even vague hypotheses if a spatial component is available.

#### *4.3.3.1 Vegetation*

An inverse relationship between neuroblastoma and the proportion of land devoted to farming was a theory put forward by Rogers and Pendergrass (1987). Although this did not involve a specific investigation into ALL a link between childhood cancers

Figure 4.11: Air pollution concentrations in Tyne and Wear



was suggested and for the purpose of this research this is sufficient evidence to warrant the inclusion of this database.

Therefore a generalised landuse database was established using information obtained from the Annual Agricultural Census for England and Wales. The basic unit of measurement supplied to non-governmental bodies is the 10km National Grid square, which is subsequently divided into one hundred 1km squares. The database represents ten landuse categories in all, summarised in Table 4.4 covering the whole of the study region, Figure 4.12.

**Table 4.4** Agricultural Landuse Codes

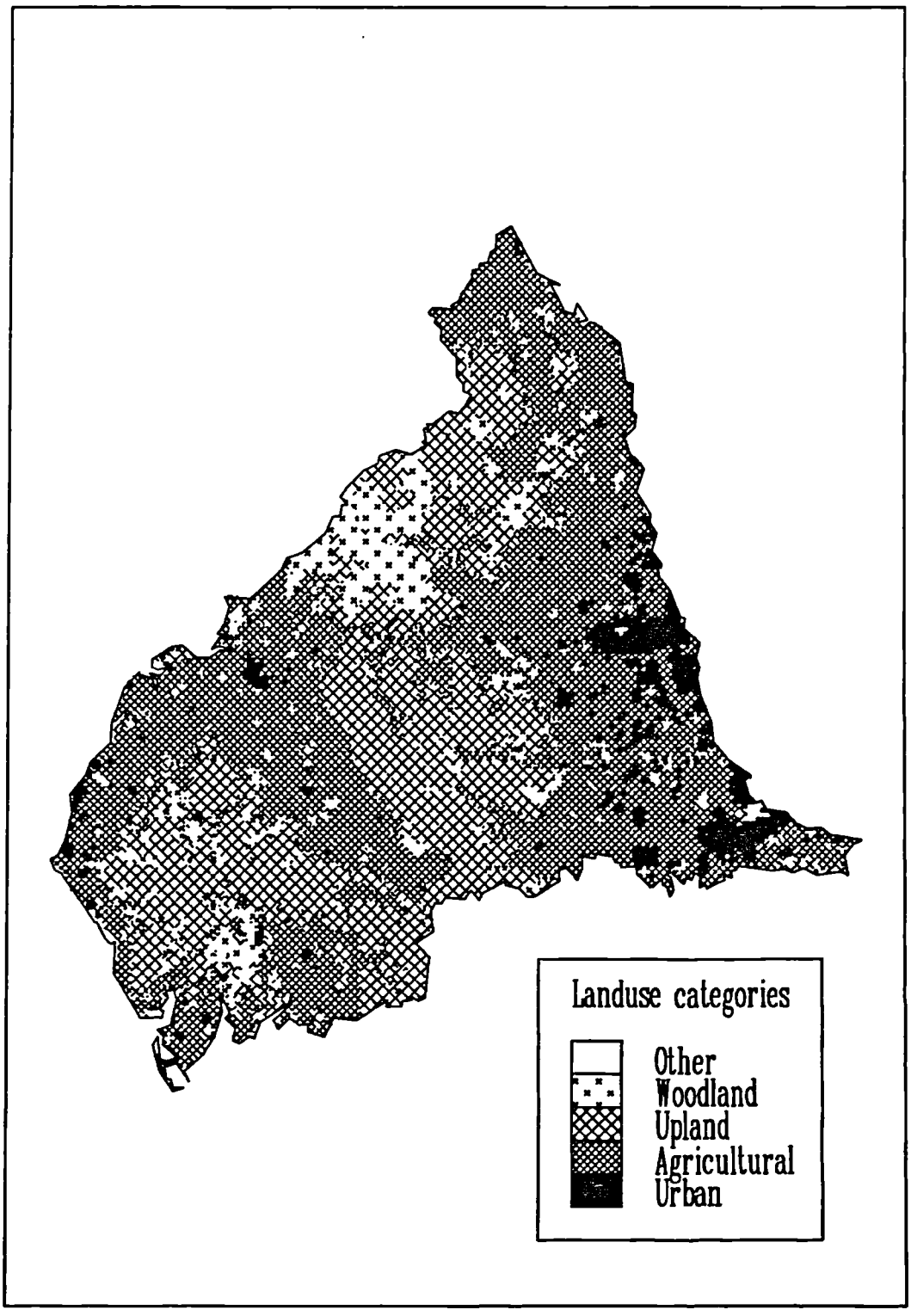
| Code | England and Wales |
|---|---|
| 1 | Urban |
| 2 | Agricultural (Grade 1&2) |
| 3 | Agricultural (Grade 3) |
| 4 | Agricultural (Grade 4) |
| 5 | Upland |
| 6 | Upland |
| 7 | Woodland |
| 8 | Non-agricultural (artificial) |
| 9 | Inland water |

Source Hotson (1988)

The coarseness of the 1km land use framework however does not lend itself to the accurate mapping of categories. For instance, the definition of woodland and inland waters can present some difficulty because these frequently cover less than half a kilometre square, and therefore may be under or over represented depending upon the subjectivity of the person coding up the data. An overall picture of urban versus rural 'landuse though should be sufficient for this health and environment GIS application.

In addition, a smaller more focused dataset was created to represent the distribution of bracken. This has been highlighted as a significant carcinogen. Evans (1976) refers to a study by Rosenberger and Heeschen (1960) who found that five cattle fed on

Figure 4.12: A general view of the landuse in Northern England



sub-lethal amounts of bracken for long periods developed haematuria and changes in the urinary bladder mucosa. Later Evans and Widdop (1966) also found that rats fed on a diet of one third bracken developed tumours, with an increased vulnerability in the younger ones. A more worrying factor is that there is evidence to suggest that bracken may constitute a direct hazard to human health, either through the consumption of the young bracken croziers, or indirectly through the contamination of milk and dairy products (Evans, 1976). This indirect effect has been attributed mainly to human stomach cancers but it cannot be ignored that this and/or other vegetation which is digested by cattle could, via the food chain, lead to increased human susceptibility to other cancers. This database was compiled by digitising a 1:200000 scale paper-based map which focused on Northumberland (Lunn, 1976). The localised nature of this database was a function of data availability, since detailed vegetation maps are usually only compiled for special studies rather than as a matter of course like geological maps.

#### *4.3.3.2 Water related sources*

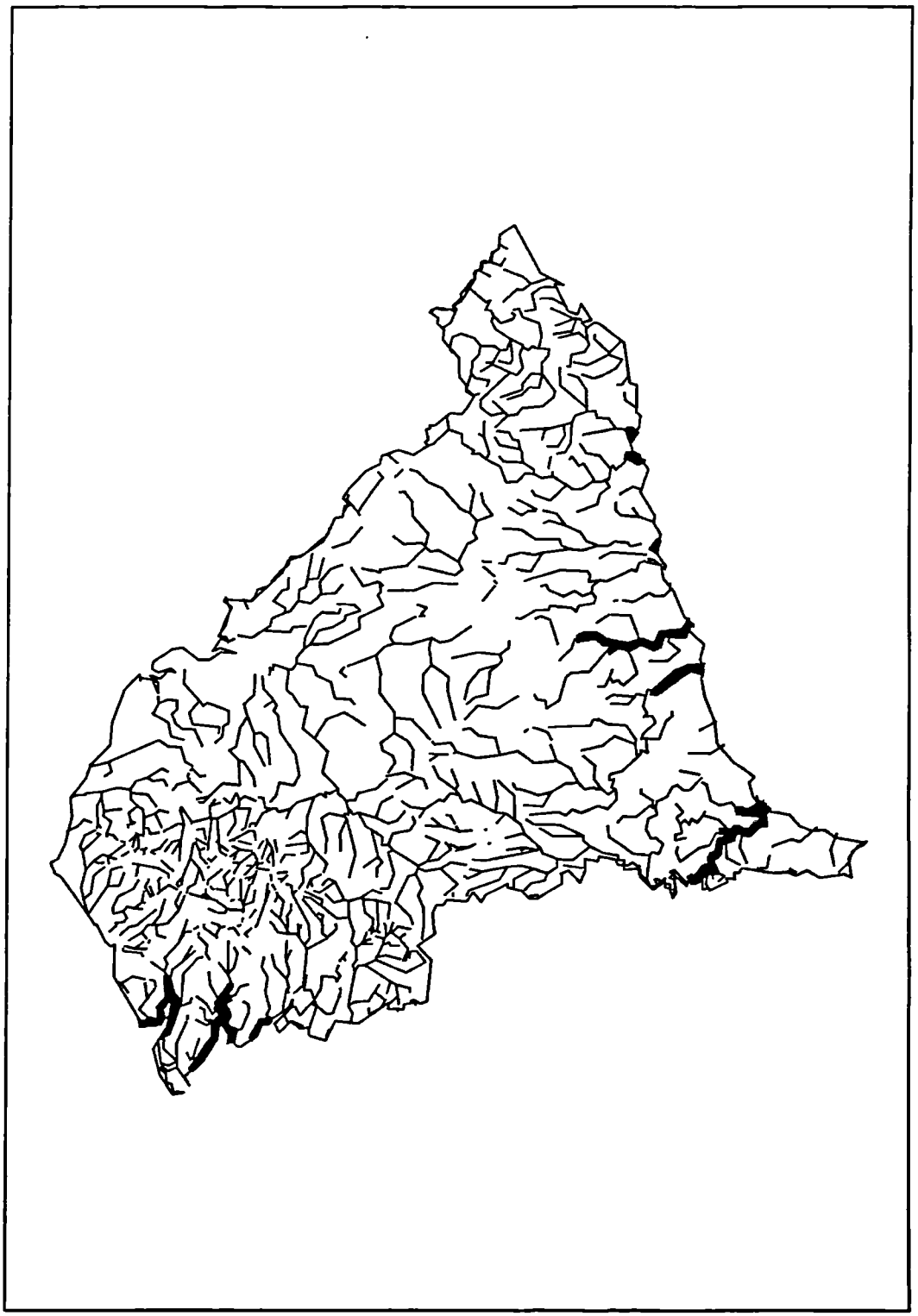
Included under this heading are estuaries, drinking water and rainfall totals.

*a) Estuaries*

This involves the investigation of leukaemia in areas adjacent to estuaries. The results to date from a new study (Alexander et al, 1990) postulate an increased risk in estuarine areas compared with coastal zones. The theory is based on the exposure of individuals to ionising radiation with increased levels being found in estuarine silts and other fine sediments, except sand. It is suggested that these higher levels are dispersed along the entire west coast of Britain. The ecological pathway is hypothesised as being a movement inwards from the oceans and stretching upwards into the tidal rivers, or alternatively movement outwards from the land mass involving heavy metals. The links to man are through the pollution of drinking water and the landward migration of radionuclides.

The definition of estuarine areas for the two coastlines involved in this study are based upon subjective delineation of the tidal regions as depicted on the OS 1:50000 map series, shown in Figure 4.13.

Figure 4.13: The drainage network and estuarine areas



*b) Drinking water*

When discussing the impact of the proximity to estuaries the importance of drinking water supplies was suggested as an important pathway for elements harmful to human health. This medium is under increasing scrutiny, especially when it is considered that apart from air, water is the most important input to the human body. Water therefore constitutes an ideal physiological opportunity for chronic toxins to be introduced into the human system.

Drinking water varies from area to area according to hardness, taste, colour, and trace elements. For instance, drinking water derived from surface sources is more likely to contain higher levels of micropollutants than those from ground water. Attention has recently been heightened by speculation over the aluminium content of drinking water supplies and the link with the brain wasting disease, Alzheimer's (Edwardson, 1988). In the Northern Region specific concern has been expressed over the 'High radon levels found in schools water' (The Journal, January 1991), in Alnwick, Northumberland, where radon levels were found to be over twenty times the national average. The National Radiological Protection Board's response was that whilst these levels were extreme they did not pose a threat to public health! Other sites in the area were also tested showing no significant levels of radon. This demonstrates the localised effect of micro-environments and the obvious impact upon selected groups of individuals.

The identification of specific catchment areas for drinking water is not simple, in addition the available Bartholomews drainage dataset (1:253440 scale) was both general and incomplete, see Figure 4.13. At this stage of building a HEGIS therefore it was decided that this information was inadequate for further analysis. Instead it is hoped that the databases which already exist in the GIS framework may be sufficient proxies for representing this environmental factor. This serves to highlight that in order to implement a GIS it is not always necessary to have all the data about every aspect of the environment before carrying out important analysis, although it is hoped that aspects such as the drainage network and associated characteristics will develop with time.

*c) Rainfall*

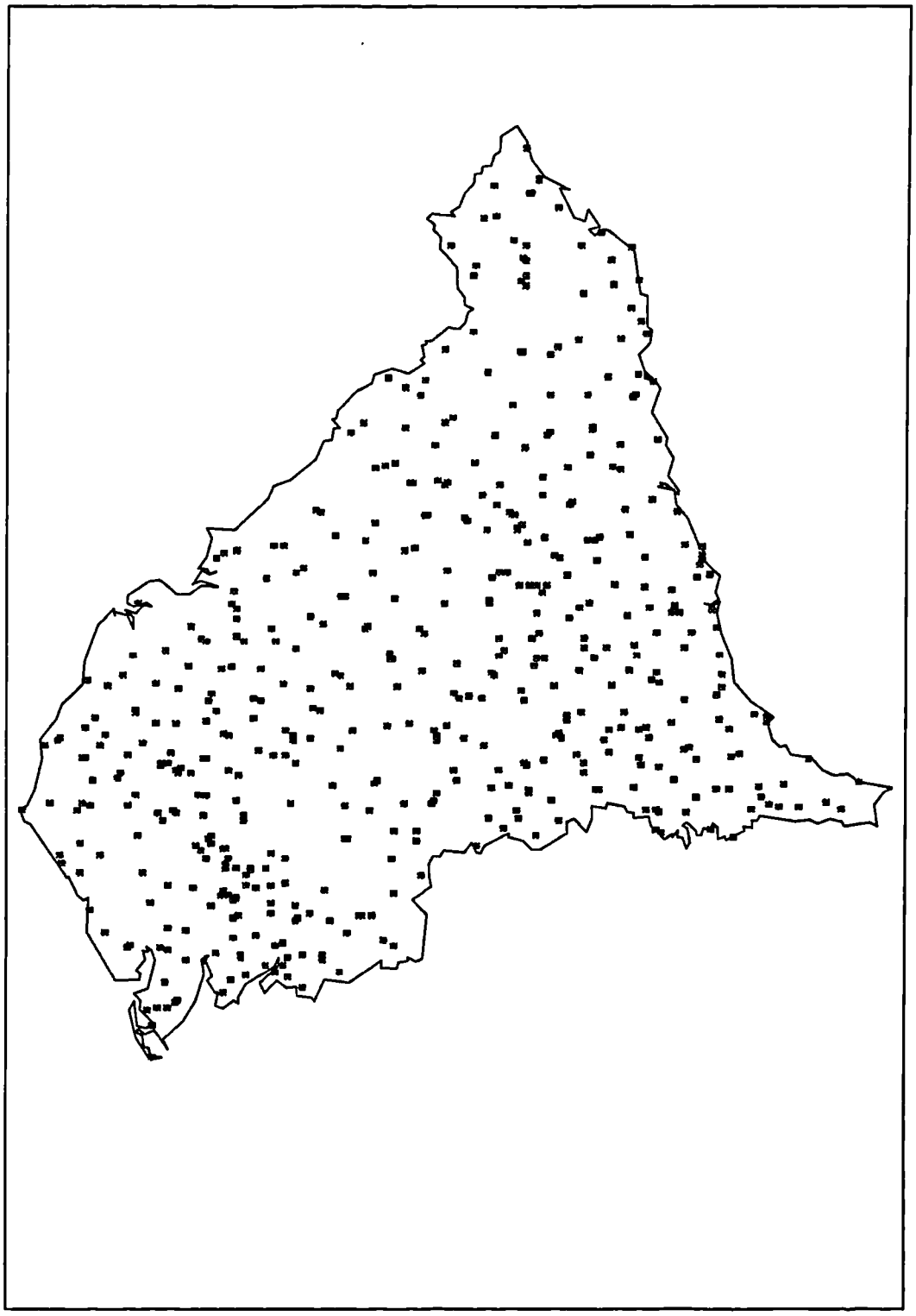
A final water related environmental factor is that of rainfall, which like rivers is assumed to be an important medium for the transportation of hazardous or toxic material from the atmosphere to the ground. In terms of the environment rainfall has often been seen to be detrimental, particularly through the destruction of vegetation and water courses due to the effects of 'acid rain'. The question this raises is that if it can have such a visible impact upon vegetation is it not possible that traces could be incorporated into agricultural crops ultimately affecting the food chain? A key example is the deposition of radioactive material over Cumbria during and after the Chernobyl disaster of 1986. The combination of atmospheric pollution, prevailing weather conditions and the onset of rain, culminated in the deposition of harmful material which lead to a ban on the sale of lamb from the area due to the possible health risk.

For this dataset daily rainfall totals covering the whole of the region from 1976 to 1986 were available form the Meteorological Office. Being a temporal based point dataset, it was subject to similar data quality and database manipulation problems that existed with the smoke and sulphur dioxide measurements. These include, discontinuous monitoring, missing data, varying density networks of recording stations and the averaging of all the data to simplify the data for analysis. Manipulation therefore involved the surfacing of the data from the resultant station points using the average rainfall total value. This created a continuous coverage for the distribution of rainfall in Northern Region, the density of recording stations is shown in Figure 4.14.

### 4.3.4 Socioeconomic sources

These include information pertaining to population counts, demographic characteristics and the social class of individuals. They may have very little to do with the causation of ALL, but factors such as the population base are essential for any system that is involved in establishing a HEGIS. It forms an integral part of medical resource management, analysis of public health and even simple calculations for disease rates cannot be achieved without fundamental background population data such as this.

Figure 4.14: Distribution of rainfall stations in Northern England



### *4.3.4.1 Population counts*

In the UK, a Census is held every 10 years and this represents the largest source of population information. Besides demographic characteristics it provides a vital link to the socioeconomic and behavioural details concerning large groups of people. Population counts are aggregated to enumeration districts (EDs), which are the smallest units that can be derived for population counts in any area. These also include counts by sex and different age groups which conveniently coincide with the descriptive epidemiological characteristics of ALL mentioned in Chapter 3, ie it gives a population count for the age groups of 0-4, 5-9 and 10-15 year olds. However these aggregations and the static nature of the data provided by the census is subject to criticism, discussed further in the overall evaluation of GIS in Chapter 10. Being the only reliable and detailed means of calculating the population at risk though the census data must be used.

In order to attach population information to the cancer database it involves matching the postcoded ALL cases to their nearest and 'supposedly' most representative ED. In this research a method of mathematical calculation was used to deduce the shortest distance between the two grid references (Reading et al 1990). Although GIS has the capability to replicate this procedure by using the NEAR command which works on the same principal by assigning the cancer in one coverage to the nearest ED stored in another, discussed in the data capture stage, Chapter 5.

Other administrative units which are employed in this thesis include ward boundaries derived from the ESRC Data archive and the digitised county boundaries from OS 1:625000 hard copy maps. These were used for preliminary analysis in Chapter 7 and as a means of simply emphasising the Northern Region, they have no particular relevance in the actual search for environmental causes of ALL.

### *4.3.4.2 Social class*

In recent years the relationship between social class and the incidence of ALL has become a favoured theory. McWhirter (1987) took a sample of childhood cancer cases from an area in Queensland Australia for 1973 to 1979. From this study it was suggested that a higher incidence of childhood leukaemia prevailed in upper social

class families. This was supported by two other independent reports, one from England and Wales (Sanders et al 1981) and one based in the USA (Browning and Gross, 1968). The explanation for the relative increase in the incidence of ALL was not clear, but several hypotheses have been put forward. The main underlying reason was that the increased number of infant deaths in lower social class families was attributable to other causes, such as infection and/or poorer health care, occurring before a malignancy could be diagnosed and thus serving to mask the true number of leukaemagenic cases.

One useful synthesis of census data therefore is 'Super Profiles' (Credit and Data Marketing Services Ltd, 1985) This contains 150 cluster classifications, from which 10 descriptive 'lifestyle' groups can be derived, see Table 3.4 in Chapter 3 for a summary of these. These socioeconomic groups were originally produced for target marketing in the private sector, but the classifications can also provide a powerful means of distinguishing national variation in socioeconomic conditions (Charlton et al, 1985). Figure 3.5 in Chapter 3 illustrated the variation of ALL according to these socioeconomic characteristics. If some inference had to be made from this graph then it may be suggested that there was a slight under representation of ALL cases in group H, 'Fading Industrial', whilst there was an even slighter increase in numbers for group A, 'Affluent Minority'. This may be interpreted as a greater prevalence in higher social classes based upon McWhirter's theory. However when dealing with rare diseases, such as ALL, the number of cases involved are small and thus significant conclusions cannot be made from such slight variations, unless of course they persist. It may be that McWhirter's study suffered from similar small number problems.

In addition, the information provided by Super Profiles may be used as a valuable proxy for other 'lifestyle' factors and the general social conditions of the areas in which children live. Thus, returning to the argument for increased incidences in the upper social class environments, it may be assumed that this is a surrogate for improved standards of nutrition or general physical health which in turn may be a contributing factor towards the development of leukaemia. This theory was supported by Tannenbaum (1940) who found a weak relationship between nutrition and cancers and identified three possible causes; a) Food additives, b) nutrient deficiencies leading to biochemical alterations and c) changes in the in-take of selected macronutrients which may produce metabolic and biochemical abnormalities. However, Reddy et al (1980) raised the important issue that the correlation between diet and certain cancers

does not prove causation. It was suggested that many factors cause cancer but the modification of just one of these contributing factors, such as diet, may be sufficient to retard the chain of causative events. Thus causes of ALL should not be viewed in isolation.

## 4.4 A Recap

This chapter has reviewed the environmental data considered relevant to study ALL. This particular focus is narrower than that which would be required under the WHO's HEGIS programme, but nevertheless it offers a useful pilot study. Even though the environmental, medical and socioeconomic datasets captured are not comprehensive they are still sufficient to tackle the problems originally established, ie. that of implementing and evaluating an environmental and health GIS.

Chapters 1 to 4 have thus constituted the feasibility stage, and in a sense form a theoretical 'Stage 0' in the development of a GIS. They are an important part of the process and should not be taken lightly, determining the requirements for database capture as well as ensuring that time and resources are efficiently focused in the next stage of 'The GIS Process', ie. data capture. It also provides vital documentation about data sources, resolution and other implications, this knowledge will ultimately lead to more accurate interpretations of the data in later stages. Both the feasibility stage outlined thus far and the data capture stage are very important to the success of GIS applications but in turn it should be noted that they are very time consuming. Chapters 5 through 7 will now document the stages of what is termed here 'The GIS Process', concentrating on both the attractive and the less desirable features that accompany the task of building a GIS for spatial epidemiology.

## **Table 4.5:** Environmental Databases: GIS Technicalities

This Table compliments that of Table 4.1 in this chapter. It highlights key aspects concerning data sources which will determine data capture and database design characteristics in Stage I, described further in Chapter 5.

| COVERAGE KEY WORD | TYPE OF SOURCE | SPATIAL REFERENCE | RESOLUTION | REGIONAL EXTENT | FEATURE TYPE |
|---|---|---|---|---|---|
| 1) ALL | Registry | Postcode | 100m | Total | Point |
| 2) Power Stations | Directory | Postcode | 100m | Total | Point |
| 3) Special Sites | Directory | Postcode | 100m | Total | Point |
| 4) Geology | Paper maps | Digitised from 1:253440 | 1m | Total | Areal |
| 5) Background Rad | Table | Grid Ref | 10km | Total | Point |
| 6) Overhead Power | Paper maps | Digitised from 1:10000 | 1m | T&W | Linear |
| 7) Substations | Paper maps | Digitised from 1:10000 | 1m | T&W | Point |
| 8) Road Network | ASCII data | Digitised from 1:253440 | 1m | Total | Linear |
| 9) Railways | ASCII data | Digitised from 1:253440 | 1m | Total | Linear |
| 10) Incinerators | Directory | Postcode | 100m | Total | Point |
| 11) Waste Disposal | Directory | Postcode | 100m | Total | Point |
| 12) Mines | Directory | Grid Ref | 100m | Total | Point |
| 13) Smoke | ASCII data | Postcode | 100m | T&W | Point |
| 14) Sulphur | ASCII data | Postcode | 100m | T&W | Point |
| 15) Landuse | ASCII data | Grid Ref | 1km | Total | Grids |
| 16) Bracken | Paper maps | Digitised from 1:200000 | 1m | N'land | Areal |
| 17) Estuaries | Paper maps | Digitised from 1:50000 | 1m | Total | Areal |
| 18) Streams | ASCII data | Digitised from 1:253440 | 1m | Total | Linear |
| 19) Rainfall | ASCII data | Postcode | 100m | Total | Point |
| 20) Population | ASCII data | EDs | 100m | Total | Point |
| 21) Wards | ASCII data | Digitised scale unknown | 1m | Total | Areal |
| 22) Counties | Paper maps | Digitised from 1:625000 | 1m | Total | Areal |
| 23) Social Class | ASCII data | Assigned to EDs | NA | Total | Point |

**NB** Despite the resolution of the original data captured, all datasets were converted to 100m grid references, ensuring compatibility with the main cancer dataset and each other.

**Note:** ASCII - American Standard Code for Information Interchange
T&W - Tyne and Wear, N'land- Northumberland

# SECTION TWO:
# THE GIS PROCESS

STAGE I: DATA CAPTURE

STAGE II: DATA STORAGE

STAGE III: DATA MANIPULATION AND ANALYSIS

STAGE IV: FURTHER ANALYSIS AND DATA PRESENTATION

# CHAPTER 5

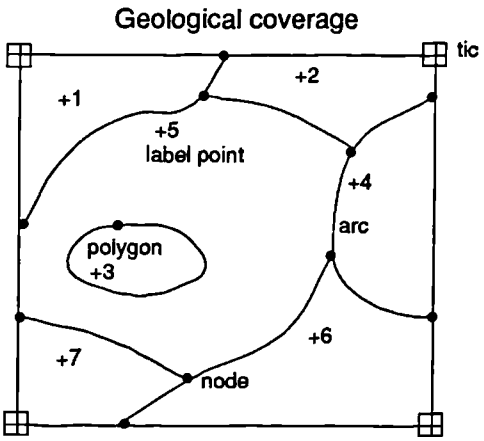
# THE GIS PROCESS: STAGE I AND II
# DATA CAPTURE AND STORAGE

This chapter is a guide to Stage I and II of 'The GIS Process' and describes the methods by which all health and environmental databases are captured and stored using a GIS. It will assist both the unskilled and the potential GIS user, by providing an explanation of the processes involved in these two very important stages. In addition it highlights some of the key decisions that need to made during any database design procedure and evaluates some of the problems which are encountered on the way. Both must be documented and tackled at this point, because if left unsolved they may produce serious repercussions at later stages. Thus the material covered in this chapter is applicable to both the person responsible for building and implementing the GIS, as well as the end-user who will access, manipulate and interpret the databases created.

As a matter of course therefore certain aspects should be documented throughout the building of a GIS application tool. For example, the database builder will make a number of key operational decisions about the method, design and eventual storage of individual datasets. These decisions will include the degree of accuracy to be employed when digitising mapped data, the resolution of data to be stored, as well as a number of subjective opinions on data standards and the coding up of groups of data. In addition, and perhaps more importantly, the user should be well informed about the actions taken by the data builder so that they too can be responsive to any inconsistencies or problems encountered in this stage, and the methods which may be adopted to overcome these. This knowledge therefore is not only background information to an application, but immensely relevant to Stages III (manipulation) and IV (presentation of data) of 'The Process', by ensuring more realistic interpretations of the data Ideally data should be captured to specified standards, which should be made readily available to all potential GIS users. This would serve to minimise the number of operational decisions made and in turn release valuable time for concentration on the actual analysis of data rather than the capturing of it, an issue which will be discussed in further detail in Chapter 10.

The end product of Stage I therefore, is the conversion of all the databases outlined in Chapters 3 and 4 into digital data files. These will contain spatial references and descriptive variables commonly referred to as 'attributes'. In order to form an integrated GIS tool for the study of spatial epidemiology though they must undergo further manipulation to ensure complete compatibility. Thus Stage II involves the cleaning up of and the removal of any errors. This is followed by the building of topology, which will establish important linkages between features found in the same database serving to create a true representation of reality. Each of the datasets referred to in Table 4.5 in chapter 4 will therefore form a single digital data layer in the system, for example geology, waste disposal sites, rainfall etc. Reality will be depicted by a set of primary features such as arcs, nodes, polygons and label points, and secondary features such as tics, geographical extent, links and annotation. Figure 5.1 demonstrates the distinction between these features and what they represent in terms of mapped data.

**Figure 5.1:** Features contained in a GIS coverage



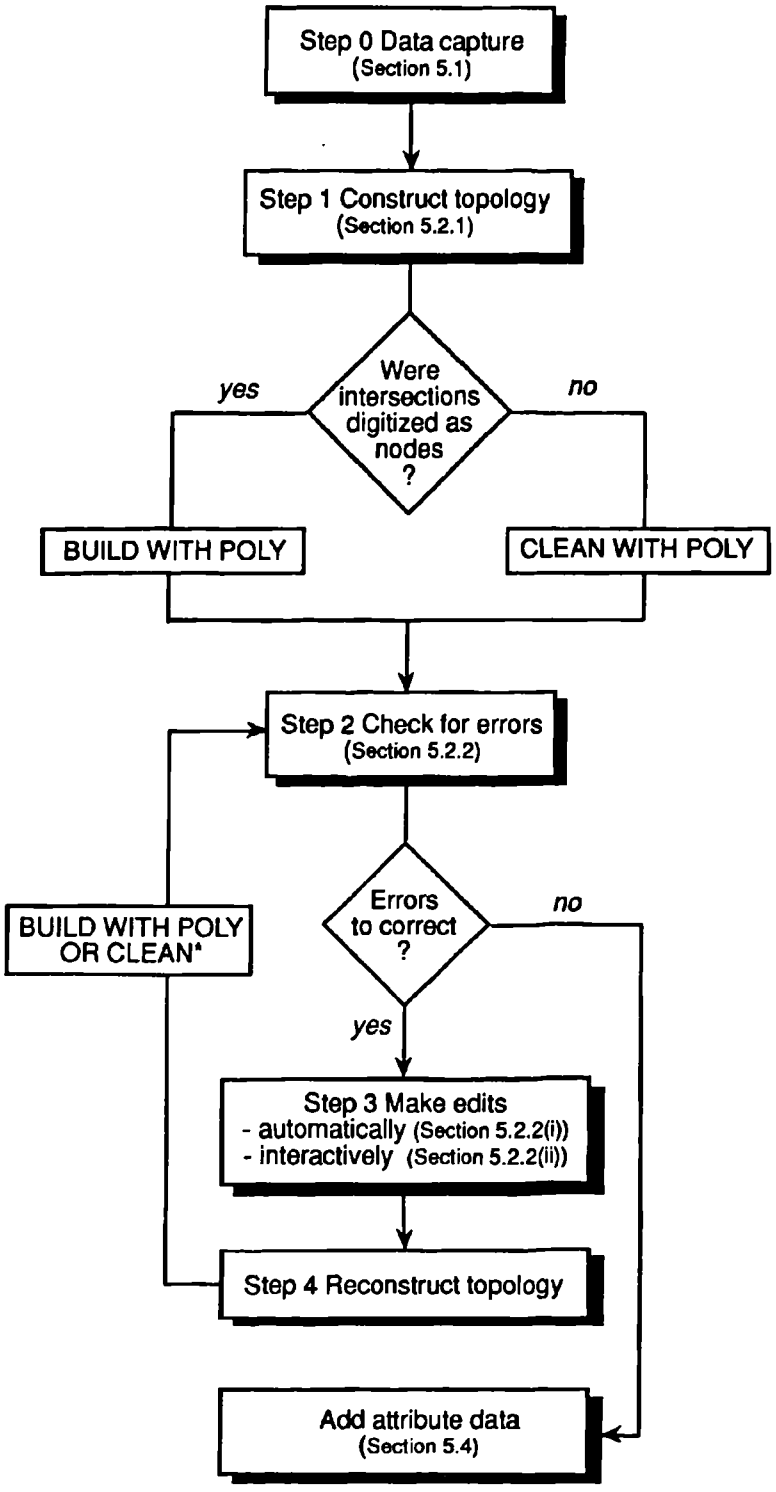

Geological attributes

| ID | GEOLOGY | CODE | SYMBOL |
|---|---|---|---|
| 1 | Limestone | 14 | 4 |
| 2 | Basalt | 2 | 1 |
| 3 | Limestone | 14 | 4 |
| 4 | Sand | 20 | 8 |
| 5 | Andesite | 3 | 13 |
| 6 | Terraces | 21 | 7 |
| 7 | Sand | 20 | 8 |

Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO Method

The end product of Stage II is a series of 'coverages' which have been produced from the transformation of raw data into a set of thematically related databases ready for subsequent manipulation and analysis. Figure 5.2 summarises the steps that are

Figure 5.2 The steps to correcting spatial data for a polygon coverage




** CLEANs on the same coverage is not recommended. In these cases, you can perform edits on the original coverage which can be re-cleaned by specifying a new output coverage for CLEAN each time it is edited.*

Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO Method

involved in the successful implementation of Stages I and II of 'The GIS Process'. These also reflect the structure of the remaining sections in this chapter.

# STAGE I

## 5.1 Data Capture

There is a danger that some GIS users will start capturing data immediately, or worst still before the acquisition of GIS technology. This problem usually stems from the management who want to see benefits from their expenditure as soon as possible. Invariably these are the same people responsible for making the decision to implement a GIS guided by the wonderful claims of software manufacturers, or impressed by some training course which provided ready made databases designed to work! As a result they tend to lack some of the basic knowledge generated during the feasibility stage, which if carried out effectively will ensure time and cost effectiveness in stages I and II, as well as flexibility and successful implementation in stages III and IV. Thus the warning here is that to overlook these initial stages will more than likely result in the production of a series of ill-conceived and irrelevant data files, as opposed to the desired integrated and complete set of GIS databases designed to meet both short and long term needs of the end-user. Chapters 3 and 4 were in effect the feasibility stage for this research. They served to highlight the health and environmental databases relevant to the spatial epidemiological problem and determine data accessibility.

The simplest form of data capture is to acquire copies of existing digital data, assuming they are complete and acceptable for use, and these can then be directly incorporated into a GIS. This would allow an almost immediate progress into Stage II or even III. Unfortunately, the task is never that simple. The epidemiologist can easily access cancer registries and immediately transfer them from their existing system to a GIS. However additional data such as ward boundaries, population counts and environmental factors which are equally as important for the development of a HEGIS will not automatically be available. These data can usually be obtained from various data collecting organisations, but this method of data acquisition can prove a very expensive option. The alternative is to build up databases from scratch. The undesirable effect of this approach is the duplication of data and resources, although on the positive side it is not only cheaper to create databases in-house, but it also

ensures that all operational decisions made at this stage, and any data problems encountered will be known from the onset of the GIS application. In this research data were acquired by both options.

The remaining subsections will outline in more detail the main methods used in data capture, with specific reference to the criteria adopted in database design and the possible impact that these will have up on the future use of these databases.

### 5.1.1 Digitising

Two options are available for capturing map based data; (a) capturing one feature at a time using a digitiser, or (b) capturing the whole map at once using an electronic scanner. The latter automatically produces a raster based image, and since ARC/INFO is a vector based system, the former of the two methods was considered preferable. This operational decision was also made in response to the content of the maps. Since existing electronic scanners are not sensitive enough to discriminate between similar shades and would have resulted in a loss of detail, especially as geological maps are invariably coded using a series of subtle variations in colour to represent different geological types, and thus certain geologies would have tended to merge into one. The databases of bracken, geology and estuaries were therefore captured by digitising hard copy base maps.

Once such technical decisions are made and the relevant maps obtained, a number of additional criteria concerning the preparation of these maps for data capture must be determined. These include; (a) the creation of the necessary workspace into which the coverage will eventually be stored and, (b) the choice of registration points for the map. The latter is the most important element of any digitising process as these registration points not only serve to locate the map on the digitiser table, but provide a basis for future transformations of data, if necessary. At least four, but preferably six registration points are located on the base map, known as TIC points. These are assigned unique ID's, of say 1 to 6, and for each of these their respective real world coordinates must be noted. Ideally they must represent features for which a specific geo-reference can be recorded, this may include landmarks, major intersections, or as was the case for the geology map the National Grid coordinates.

Another element of map preparation is the decision on how the features should be recorded. For instance, the geological map is very complex with a number of small polygons situated within larger ones. It was decided therefore to systematically work through the arcs from left to right and top to bottom respectively. This is usually referred to as 'spaghetti' digitising, because intersections are not explicitly defined during the data capture process and this results in a series of overlapping arcs which resemble a 'pile of spaghetti'.

For this exercise an 'orthogonal' Calcomp 9100 digitiser was used. This consists of a table with a fine wire grid of electronic conductors embedded below the surface. Certain guide-lines must be followed during this digitising process. (a) The map should be affixed onto the active area of the digitiser table. This is because the electrical sensitivity of the digitiser does not extend to the edges of the table thus any data captured in these areas would be lost or at best distorted. (b) The maps should also be completely flat and secure to keep the number of errors to a minimum.

The operator then traces the map features using the cross hairs in the centre of the digitiser cursor. This sets up an alternating current between the cursor and the table which logs the position of the digitiser as it moves across the map. The x- and y-coordinate positions are then registered and stored in the computer when the relevant button is pressed on the cursor pad. The ARC Digitising System (ADS) in ARC/INFO ensures that this stage is fairly user-friendly providing a menu driven interface that takes the operator systematically through the steps for registering and capturing map details. The result of this process is a reproduction of the paper-based map as a digital file, stored in map units and ready for conversion into real world geo-reference coordinates, at a later stage (section 5.5.1).

This data capture process can, and does, generate a number of errors. Some of the errors are a product of the hardware and software limitations, but most are due to the inability of the operator to replicate what they see. Even an experienced operator cannot guarantee a completely accurate replication of a map first time. Any problems must therefore be rectified in order to render the coverages clean and usable, this is discussed further in section 5.2. However the creation and introduction of data error at this stage is a complex and as yet unsolved feature of GIS technology. At best it must be hoped that the databases acquired, or the ones that are built in-house, are done so

to the best standards possible and are accompanied by documentation of any problems and operational decisions made.

### 5.1.2 Manual/ASCII file data input

Data may already be available in a digital format, such as those derived from commercial firms such as Bartholomews, the Meteorological Office and Warren Springs Laboratory. These include the road network, rainfall totals and smoke/sulphur dioxide concentrations. When these were compatible data could be input directly into the system.

ARC/INFO provides a number of standard transfer format programs which can convert certain data formats to a compatible form. Despite this though not all data formats can be catered for, thus as a generic alternative ARC/INFO also accepts digital data from simple ASCII text files containing x- and y- coordinate pairs. These coordinates can then be converted from ASCII data files to usable coverages using the ARC command GENERATE. This technique was particularly useful for the preparation of the sulphur dioxide and smoke datasets.

An ideal x- and y- system of spatial referencing is not always available. Some raw data may only possess an address as a reference point, or it may have a postcode if you are lucky! For example, data acquired from the cancer registry only gave the postcode as a spatial reference point. Consequently, such datasets required further pre-processing before the GIS data capture stage could commence. This involved assigning the relevant x- and y- coordinates to the postcodes available. Fortunately, this was made possible by a program written at Newcastle University which matches up the cancer individuals' postcoded addresses with information stored in a Postcode Address File (produced by the Post Office). This allocates a 100 metre grid-reference to the postcode. An important factor which should be emphasised about this allocation procedure, is the criteria employed to assign these coordinates. They are added from OS maps, whereby the 'first' address encountered with a new postcode will be used to attach the relevant geo-reference. This x- and y- coordinate will then be assigned to all other addresses that possess the same postcode as the first address. This raises some important questions over the accuracy of data employed, and constitutes a major issue of Chapter 10 which takes a more detailed look at the problems involved in assigning postcodes and their inherent 100 metre resolution.

The next stage in the data automation process is to ensure that the data coverage created is free from any errors and is topologically correct.

## 5.2 Making spatial data usable

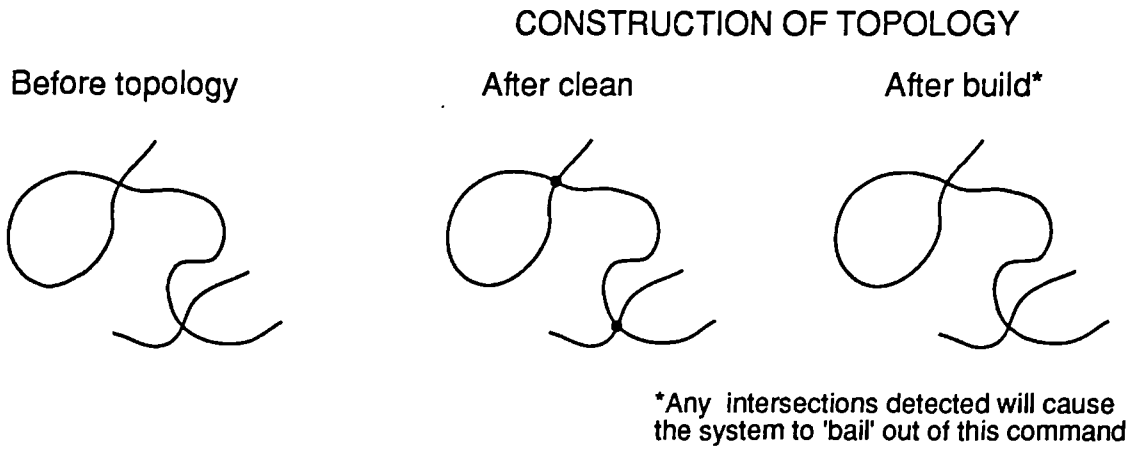
This is the intermediate stage between data capture and data manipulation. It includes the checking of coverages for any missing or unwanted data. The checking of cancer data may involve spotting rogue postcodes which for some reason fall outside the study area. Alternatively for areal coverages such as the geological dataset it may involve searching for polygons which are incomplete ie. areas where the boundaries are not enclosed.

### 5.2.1 Creating topology

Before any coverage can be edited, manipulated or have feature attributes added to it, it needs to be topologically correct. This involves the conversion of lines, points and areas, captured in their raw form in Stage I to form a true representation of the map. Creating topology therefore establishes spatial relationships and determines connections between features and the adjacency of features (contiguity). Thus before topology is constructed no polygons actually exist and arcs which may cross have no defined intersection (or node).

ARC/INFO provides two commands; BUILD and CLEAN, which automatically create topology and templates for feature attribute tables. These commands are not exactly the same since BUILD will process all types of features whereas CLEAN will only process lines and polygons. The most important difference in their usage depends on the original digitising practice adopted. BUILD assumes that the data has been entered correctly and therefore does not recognise undefined intersections (known as discrete digitising). CLEAN, on the other hand, will create new intersections wherever arcs cross. Therefore in the case of spaghetti digitising CLEAN is the most appropriate option for creating topology. A summary of these differences are shown schematically in Figure 5.3

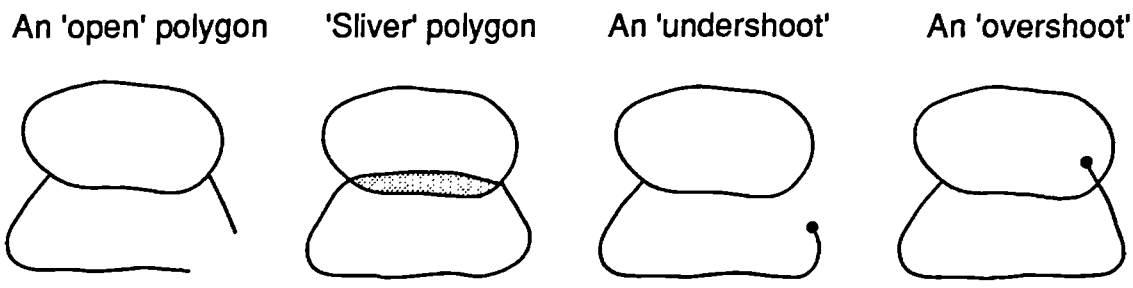
**Figure 5.3:** Creating Topology




Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO Method

Topology also serves to highlight certain types of error in the database. Those common to digitised data include dangle nodes/arcs and sliver polygons, illustrated in Figure 5.4. Although it must be noted that in some cases a dangle node or arc may be acceptable, for example the representation of streams and dead end roads. However the key to successful GIS implementation is based on clean and well designed databases. The elimination of any source of error at this stage is crucial, since even ensuring topological consistency does not necessarily mean the creation of accurate coverages!

**Figure 5.4:** Types of error which can occur with digitised data




Adapted form ESRI (1990), PC Understanding GIS: The ARC/INFO Method

### 5.2.2 Correcting for Error in captured data

The correction of error in a database results in the adding of any missing data and/or removing of any bad data. This is done manually or automatically.

*i) Automatic editing*

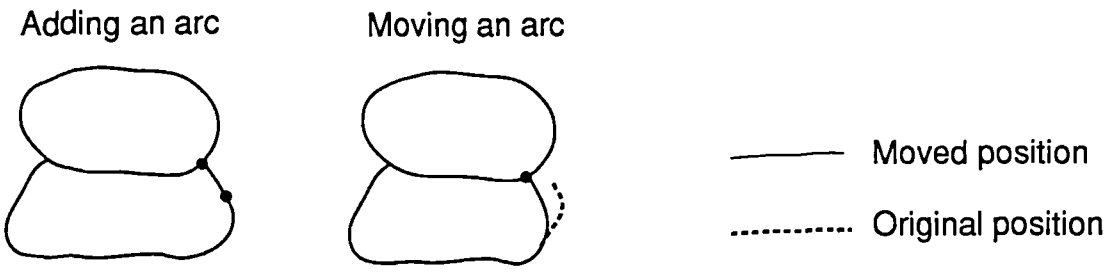
This can be achieved at the same time as the topology is created by using the ARC/INFO command CLEAN. The key options in the command are {dangle_length} and {fuzzy_tolerance}. The function of these is to allow the minimum distance to be specified between arc vertices such that any dangling arcs found to be less than the specified distance will be snapped together. If no edit tolerance is required for the cleaning up of the coverage, a {dangle_length} and {fuzzy_tolerance} of 0.0 must be explicitly stated, because ARC/INFO sets up default distances which will be used in the absence of any specifications.

It is very important to specify sensible tolerance levels, because there is a danger of losing very small but legitimate polygons when the tolerance level is set too high. Also larger edit distances lead to a greater degree of generalisation of other arcs in the database. The rule is therefore to fully understand the nature of the data to be encoded and act accordingly.

*ii) Interactive editing*

The alternative is an interactive approach to correcting data, using the ARC/INFO subsystem ARCEDIT. This allows errors to be diagnosed, feature attributes to be edited, and data to be matched across boundaries. This can be done by comparing the digitised version of a map with the original data source. In ARCEDIT features such as arcs, nodes and points can be selected and edited accordingly. Editing includes adding, deleting, moving, splitting, or annotating the features displayed. Consequently, if a coverage includes 'undershoots' they can be corrected in one of two ways; (a) by adding an arc between the two dangling nodes to connect them, or (b) by moving one of the dangling arcs to meet the other. However, the latter solution will also effect the position of the whole arc which may not be desirable, illustrated in Figure 5.5.

**Figure 5.5:** Correcting for errors in digitised data




Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO Method

This interactive approach to the editing process therefore has the potential to encourage the users to 'create' new data errors in their eagerness to make the data meet reality. Interactive editing can, and does, lead to a kind of data fabrication process which the data gatherer should be fully aware of. If simply ignored these minor errors will be carried through into every stage of 'The GIS Process', and could prove to be the difference between the incidence of ALL occurring on geology type A rather than geology type B.

After any interactive edit process the topology will have been altered and the next step will be to reconstruct this again. The fixing of these errors and the building of topology are the most fundamental steps in database creation. If ignored, they may render future calculations and analyses of the data inaccurate and possibly meaningless. For instance, if the geological database was left with open polygons the codes from one area would spill over into adjacent areas with the result that the codes assigned would be incorrect. Errors such as these can be flagged by using the ARC/INFO command LABELERRORS which will highlight all the polygons which do not contain a label point or have more than one. A glossary of the most commonly used commands in this research and their usage is given in Appendix D.

The final outcome of the editing process is a topologically correct and clean coverage, containing a unique internal number to register each feature. The coverage is now ready to be stored and manipulated within a GIS framework.

## STAGE II

## 5.3 Data Storage

Each coverage has a set of reference files which include the TICS registered at the start of digitising, and the BND file which essentially logs the total extent of the coverage. On building the topology a feature attribute table for each coverage was also generated. This can have one of three forms depending on the feature type represented in the coverage;

PAT polygon attribute table, which is produced for an areal coverage such as the geology database

AAT arc attribute table, which applies to a linear network like the road coverage

PAT point attribute table, which includes the point cancer data from the Registry

Typically, all these tables consist of columns which represent items, some of which are automatically generated in the topological process, for example, area, perimeter and user-ids. Others include attribute items which can be added as extra descriptive information to be attached to the spatial data. In terms of the medical data this would allow items such as place of diagnosis, sex, age etc. to be attached to the relevant point locations. The rows within the table are the individual records for a particular point or area. Figure 5.6 shows an example of part of the polygon attribute table for the geology coverage.

The success of GIS implementation depends on carefully designed datafiles and an overall database structure which ensures sufficient flexibility in subsequent GIS analysis. Thus the creation and management of the general work space is equally as important in the development of a HEGIS. In ARC/INFO this is likened to the concept of a filing system where an individual directory is created to contain a collection of files and these are linked to other directories in a hierarchical structure.

**Figure 5.6** A typical Polygon Attribute Table, as stored in INFO

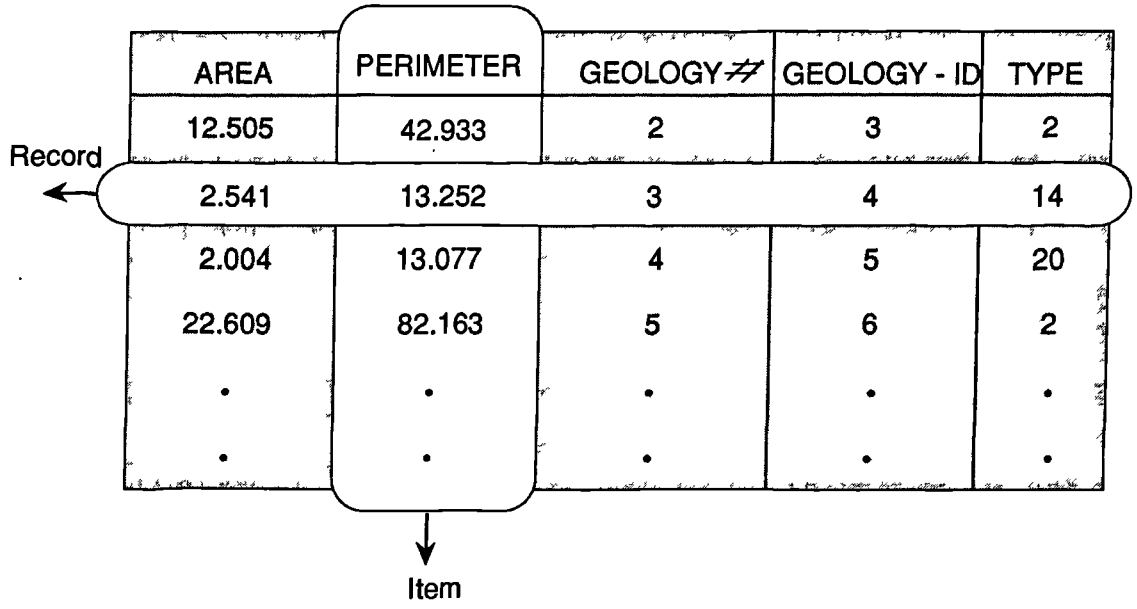
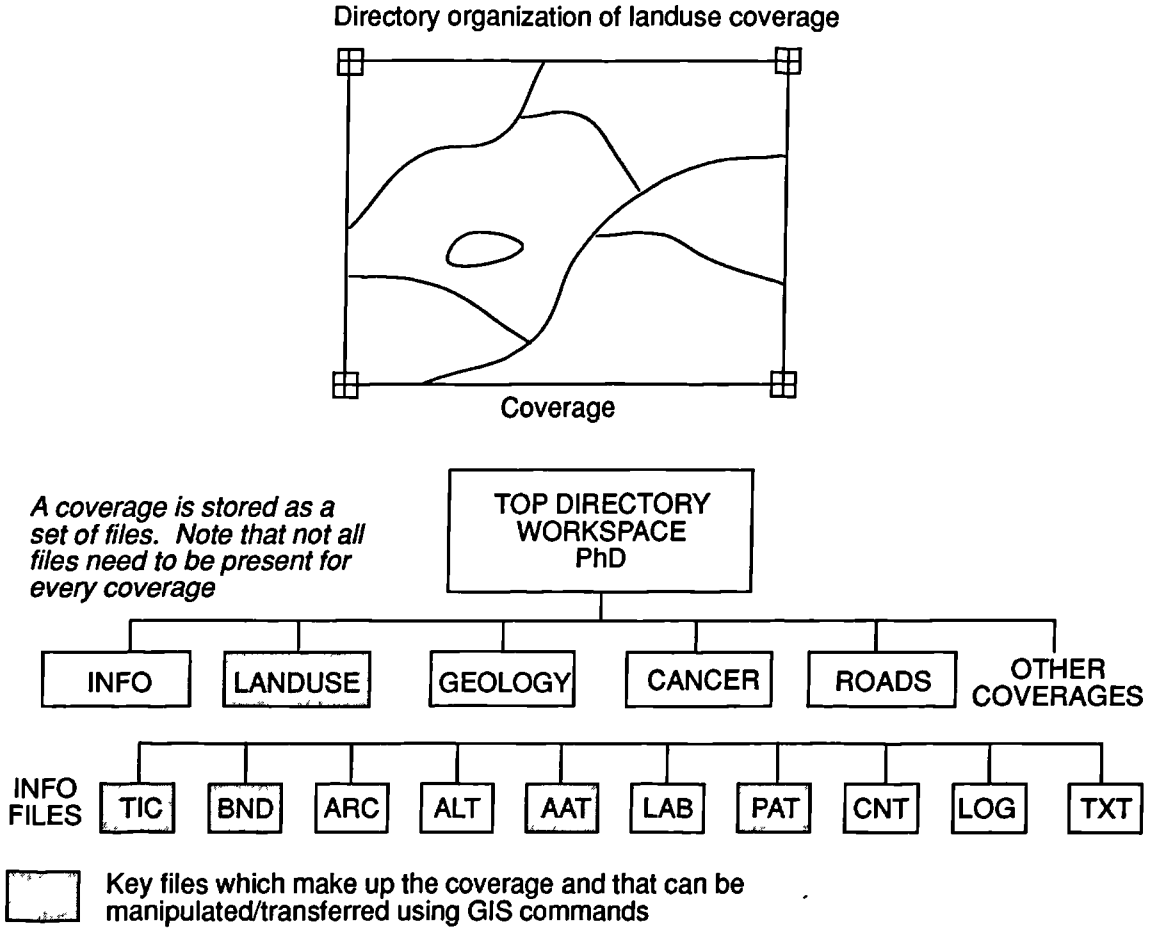



**Figure 5.7** Organising databases in a GIS




Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO Method

Directories offer a convenient means of organising work by grouping files into logical and manageable sections. Thus a particular directory may store all the data, coverages and related look-up tables for a specific environmental factor such as landuse which would include coverages for the distribution of bracken in Northumberland, and/or urban and agricultural areas. This serves to give compactness and order to the databases created, as shown in Figure 5.7 on the previous page.

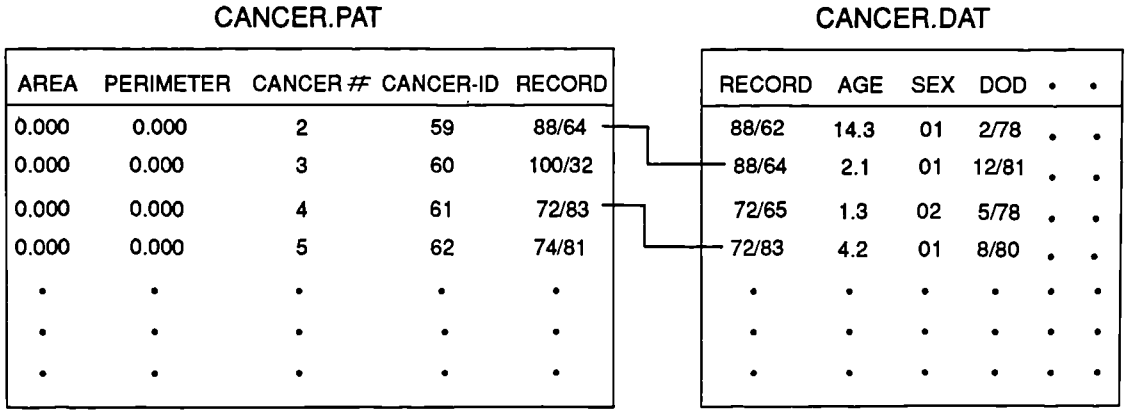
## 5.4 Assigning attribute information to spatial data

The completion of any dataset involves the addition of attribute data to the captured spatial data. This procedure is achieved by firstly assigning label points to the cleaned coverages using the command CREATELABELS. Attributes can be added from a separate file or manually using the mouse in ARCEDIT. The most common means of matching spatial data to its associated attribute information is to combine two files. This is best illustrated by outlining the development of the cancer database. The process was as follows;

A file was created in INFO to hold all the additional attributes, such as age, sex, date of diagnosis etc, but more importantly an item which would link this dataset with that containing the spatial information. This usually involves an item known as the coverage-id, because this is automatically produced for the spatial data when the coverage is created. Once a template has been defined the data can be added either manually a record at a time, or where 100's even 1000's of cases exist, they can be added all at once from a compatible ASCII file. This new datafile can now be joined to the feature attribute table generated for the spatial data, described in sections 5.1 and 5.2. With a one-to-one match between CANCER.DAT (descriptive attributes) and the CANCER.PAT (point attribute file) the files can be joined using the ARC/INFO command JOINITEM. When the merge involves one attribute being matched to more than one point in space then a RELATE option is more effective. The effect of the JOINITEM command though is summarised in Figure 5.8

However, the datasets which have been produced are not yet ready for use within the application. This is because they have all been developed independently and possess different levels of resolution and scale, for example the mapped data is still stored in unusable digitised coordinates. These datasets therefore need to be manipulated further to ensure compatibility for future use in Stages III and IV.

**Figure 5.8:** Combining data files in INFO




Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO Method

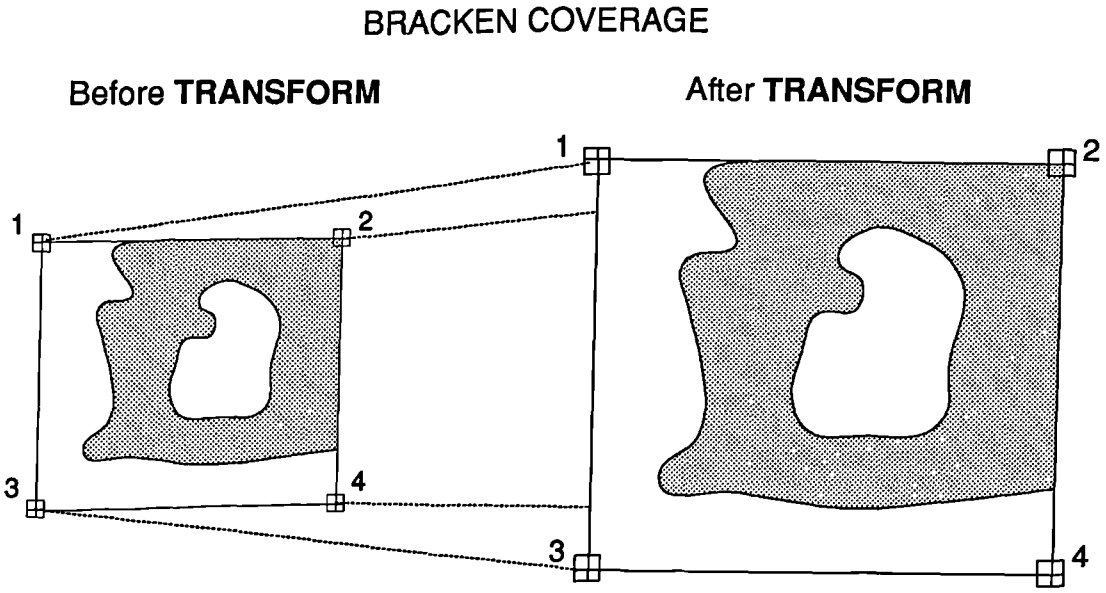
## 5.5 Database Manipulation

This section could cover a vast range of procedures ranging from map projections to simple map limit definitions. However, this research made use of only four main types of manipulation techniques; transformations, clips, surface generation and edgematching. These in turn are probably indicative of the main needs of the general end-user.

### 5.5.1 Transformations

It may be necessary to alter coverages in two ways. The first of these is the conversion of the digitised coverages from map units to the necessary real world geo-referenced coordinates. This process requires the creation of a new TIC file containing registration points which will complement the original coverage, for example, Bracken2.TIC from Bracken1.TIC. The real world coordinates noted at the onset of the digitisation process are then entered into the new TIC file. The ARC command TRANSFORM can then be used to map the digitised version of the coverage onto the new real world coordinates, as shown in Figure 5.9.

**Figure 5.9** Transforming digitised data to real world coordinates




Coverage in digitizer inches

| IDTIC | XTIC | YTIC |
|---|---|---|
| 1 | 2.74 | 2.90 |
| 2 | 2.73 | 5.33 |
| 3 | 0.39 | 2.89 |
| 4 | 0.42 | 5.31 |

Coverage in real-world coordinates

| IDTIC | XTIC | YTIC |
|---|---|---|
| 1 | 4,703.3 | 4,596.7 |
| 2 | 4,703.3 | 6,603.7 |
| 3 | 2,895.1 | 4,596.5 |
| 4 | 2,895.1 | 6,603.7 |

Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO Method

This process was needed for all coverages derived from paper-based maps, including estuaries, geology and bracken. Outside organisations would also have used this procedure, for example Bartholomews who provided a digitised version of the road and railway network.

This command is also used to transform coverages which are already in real world coordinates but are at different resolutions. For example, when building the cancer database the 100m grid references were known and entered directly into the system, however, other datasets namely the meteorological recording stations were assigned 12 digit grid references, making them accurate at the 1 metre level. In terms of a GIS application these datasets are presently incompatible for analysis, thus they need to be converted to the same resolution. In this research the scale was dictated by the detail of the health data and hence was set at 100 metres. On this occasion therefore TRANSFORM would be used to convert a 1 metre resolution file already in real world coordinates to a generalised version of the coordinates, ie. taking grid references of 485000 and 540000 and mapping them onto a 100m resolution file containing the equivalent coordinates specified at 4850 and 5400 respectively.

It should be noted therefore that constraints based on the resolution of the main dataset can lead to undesirable generalisations of the other databases in the research and reduce the accuracy of any results produced. In most cases this is unavoidable due to the spatial referencing standards that are adopted by the data gatherers concerned. Potential GIS end-users must always attempt to view GIS in the long term if they are to make the best use of their data. Therefore, wherever possible, the best policy is to capture data at the finest resolution and then aggregate when necessary. This provides a greater flexibility of datasets for future projects. This and the impact of other human decisions in 'The GIS Process' will be discussed in further detail in Chapter 10.

### 5.5.2 Map extent

Most of the databases created in this research covered the whole of the Northern Region, however there is an option to focus at a more detailed and smaller spatial unit. Thus given additional datasets for the county of Tyne and Wear, including overhead power lines and substations, it is possible to isolate this area for independent analysis. This can be achieved by reselecting on an area code, if it exists in the INFO

file, or alternatively using the ARC/INFO command CLIP to essentially cut out the information from the main database using the map extent for the county of Tyne and Wear, and this can then be saved in a separate coverage, see Figure 5.10.

### 5.5.3 Creating surfaces

As mentioned in Chapter 4 the use of point sources for atmospheric pollution and/or rainfall measurements are not always conducive to successful analysis in GIS, since the environmental impact of these sources has a much wider range than the immediate area surrounding the recording station. Simply buffering the outlets therefore is not really the most sensible solution. An alternative is the creation of a surface of concentration values which gives a regional representation of the environmental factors.

Surfaces can be created by employing TIN (an acronym for Triangular Irregular Network), which is fully integrated into ARC/INFO and can generate surfaces from any points with x, y and z concentration values. This process involves the interpolation of adjacent and non-overlapping triangles derived from the points between neighbouring stations. Obviously these derived values are not the real values but are the most likely mathematically. The result is a surface which can be used in relation with the distribution of ALL and other childhood cancer cases.

It is accepted that the surface created will be far more reliable where the number of recording stations is greater. Since this tends to be in the urban areas where the incidences of childhood cancer are most dominant too, the latter limitations should not be that problematic. A key benefit from using this technique lies with the type of displays that can be obtained from the new coverage. These can be both very impressive and visually informative, including the preview of profiles and three dimensional displays viewed at different angles, orientations and levels of resolution. In addition other coverage features can be draped over these new surfaces using the VIEW program. Figure 5.11 demonstrates the alternative ways in which the sample data for background radiation levels can be presented.

Figure 5.10: Isolating an area of interest by using the CLIP command

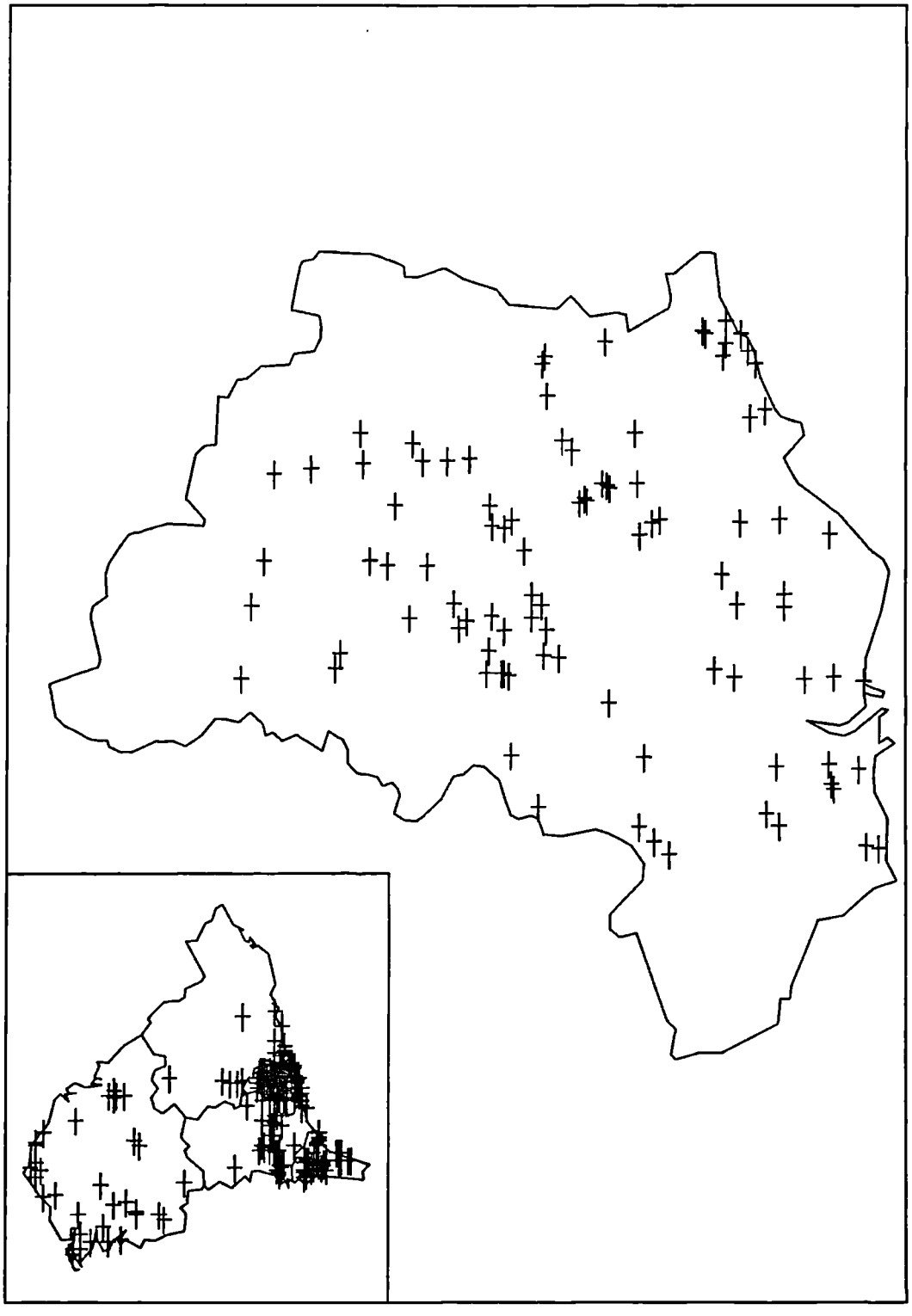
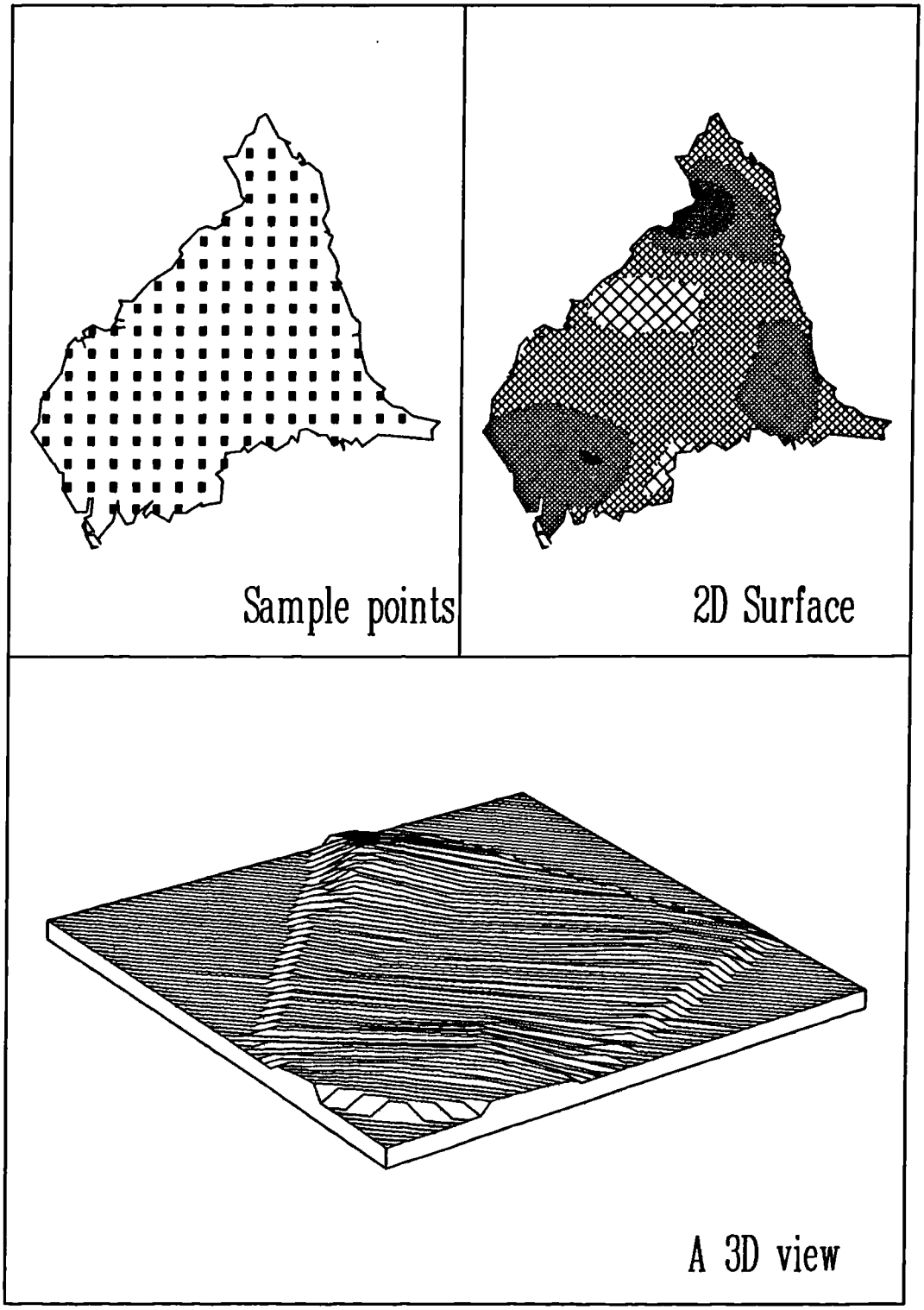


Figure 5.11: Different ways of viewing the information provided for background radiation levels



### 5.5.4 Edgematching

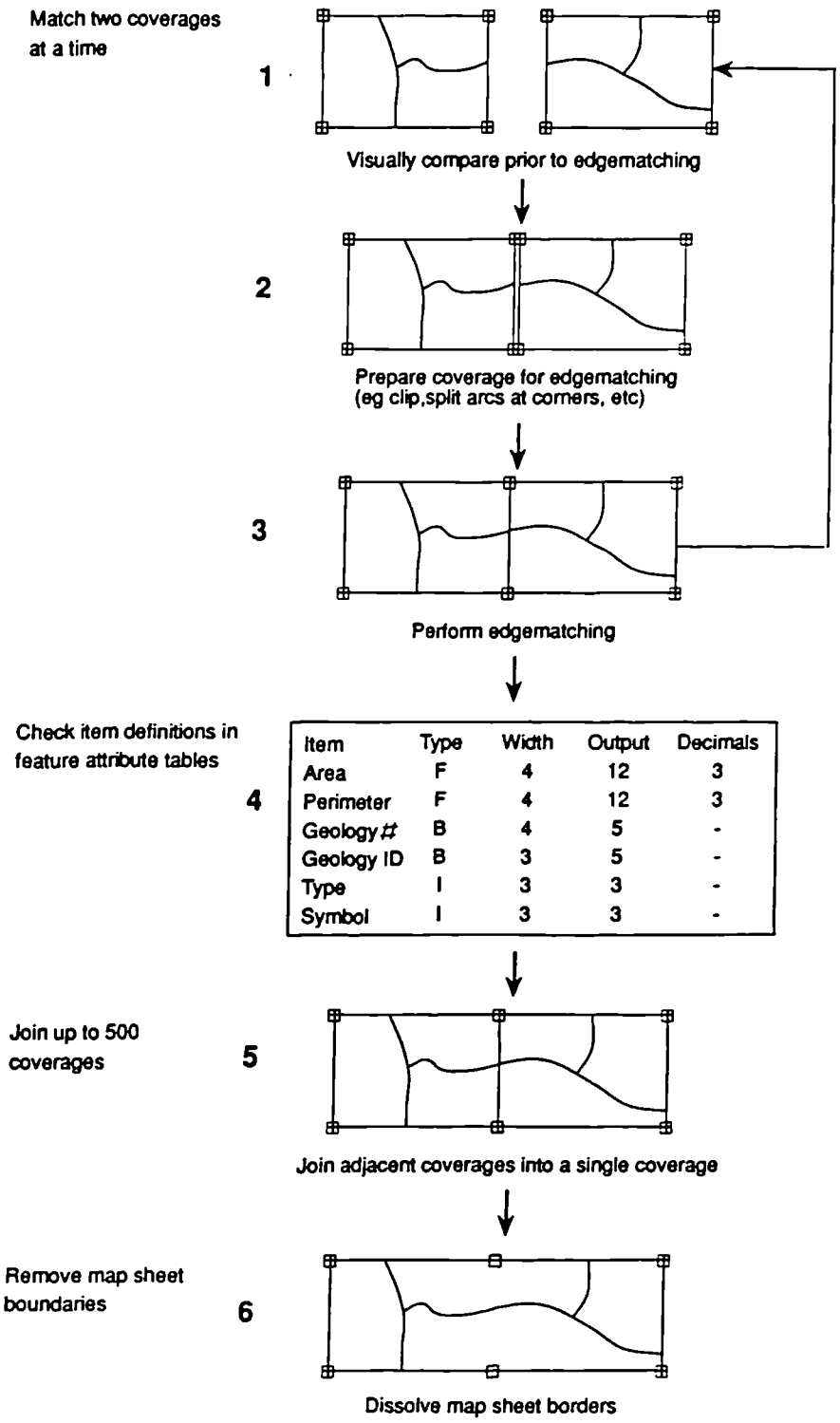
This type of database manipulation was particularly relevant to the building up of the geological dataset because the original data source extended over four different paper based maps. In order to establish a single, coherent database for the geology of the region these separately digitised maps needed to be merged together. The EDGEMATCH command in ARC/INFO allows this type of manipulation. It, takes one map as a reference to which all other maps are matched and results in the creation of a new comprehensive coverage. A problem can be experienced though when the coordinates do not exactly match along the adjacent boundaries, thus requiring coordinates to be shifted in order to make the coverages compatible. This inconsistency usually occurs due to distortions of the original paper-based maps through temperature changes, excessive handling, the printing process, and any operator errors which may have been incurred in the digitising process itself. Figure 5.12 outlines the stages involved in this type of map manipulation.

Up to 500 adjacent coverages can then be physically joined together using the MAPJOIN command. The borders between them will be retained but can be eliminated using the DISSOLVE command. It must be remembered however that the removal of the boundaries will serve to alter the topology of the resultant coverage. It is usually a good idea therefore to leave any labelling of polygons until after this type of database manipulation is complete, thereby avoiding the possibility of assigned codes being corrupted during the rebuilding of topology leading to the creation of attribute errors.

## 5.6 Summary

The importance of stages I and II in the development of a GIS application cannot really be emphasised strongly enough. Those involved must be prepared for set-backs in data acquisition, as well as being equipped with logical thinking, especially when making decisions upon alternative datasets to represent those features which are either missing or unobtainable. Estimates suggest that these stages can take Up to 60 or 80 percent of the time and development costs required to build up a GIS application. This is particularly so when taking into account both the manual data gathering,

**Figure 5.12** The process used to join adjacent geological coverages




Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO method

awaiting upon other contacts to supply data, and the considerable amount of manipulation of data that may be necessary. The implications of these stages therefore can be far reaching, since they determine key aspects of database flexibility and accuracy. The latter being the subject of greater discussion in Chapter 10 which looks at the errors which can be involved in Stages I and II and their importance in the success of GIS applications.

The satisfaction of completing this part of 'The Process' though lies in the benefits which can be accrued from subsequent analysis and the successful use of GIS as an aetiological tool. Also, a good design will lend the system to further development over time providing a system which can evolve to allow new datasets to be included whenever it is necessary or appropriate. It will also be flexible enough to allow new techniques to be used on existing data.

There is a lot of work involved in these stages which unfortunately tends to be taken for granted, especially when the end-user simply employs GIS through customised menu interfaces. Ideally there should be a continuous dialogue between the database builder and the end-user. This will not only inform the user (who in this instance is the epidemiologist) about database problems and the need for certain surrogate features, but also the data capturer who needs to fully understand the aims of GIS as a spatial epidemiological tool in order to ensure that the system built and the databases created successfully meet the epidemiologist's objectives.

In one sense Stages III and IV are therefore dependent upon these initial steps. However they can also provide positive feedback because if a data gatherer can see their data put to use and creating some interesting results then this in turn may stimulate more enthusiasm, or at least a desire to capture data as accurately as possible. Chapters 6 and 7 make extensive use of the coverages which have been completed as a result of Stages I and II, and provide a means of truly evaluating GIS as a spatial epidemiological tool. We now know it can store the necessary data but can it solve the questions that the epidemiologist wishes to ask?

# CHAPTER 6

# THE GIS PROCESS: STAGE III
# MANIPULATION AND ANALYSIS

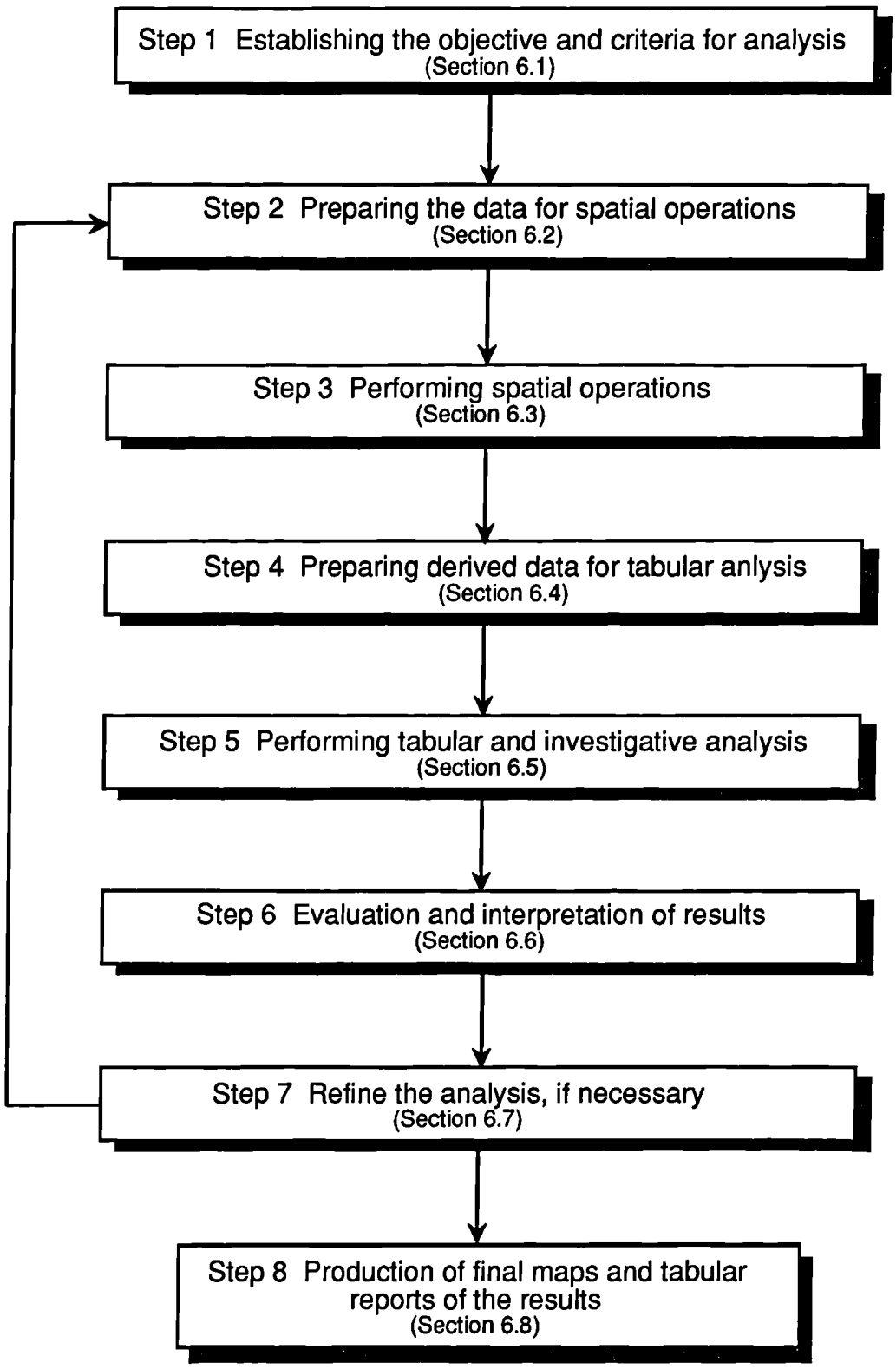
Any modern GIS will contain over 1000 different commands and several specialised modules. A sample of these have already been referred to in the development stages of this GIS application (Chapters 3 to 5), however a majority are designed to organise, manipulate and analyse established databases. This chapter will highlight a number of these commands with emphasis being put upon the purpose and result of using them rather than their actual syntax. Commands can be used either on their own, or combined to form integrated programs designed to carry out entire analysis procedures.

Stage III of 'The GIS Process' begins to fully explore available databases, particularly the potential of GIS as a spatial epidemiological tool in the search for environmental correlates of ALL. As in the manner of the entire research presentation, this stage can be rationalised into a series of steps. For example the ARC/INFO User Guide, which accompanies the software, defines the 'traditional' method to GIS analysis as eight, summarised in Figure 6.1. These are expanded upon in more detail in the sections to follow. It is important to note that the structured outline provided by ARC/INFO tends to take a rather 'simplistic' view of 'The GIS Process'. In reality the route to successful GIS analysis is often complicated, with each of the steps uncovering its own set of problems and technicalities which have to be overcome. This usually involves adopting an alternative or more obscure approach to the analytical problem in order to reach the desired goal. These too will be highlighted in this chapter.

## 6.1 Establishing the objectives and criteria for analysis

This section and that of 6.2 should have already been completed by Stage III of the process. They involve the formulation of a set of questions which typify the needs of the application. It would have been presumed at the beginning that GIS technology would be able to answer these, or at least provide a framework to help answer the spatial problems they represented. However these questions may have been changed

**Figure 6.1:** The steps to 'GIS Analysis'




Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO method

or added to during the feasibility stage (Chapters 3 and 4). At this point though they should be finalised to represent the objectives for Stage IV (Chapter 7).

In general questions vary according to the scale and subject matter of an application. Thus every GIS application will access and explore different commands and modules in order to satisfy their analysis needs. As an example, the European Health and Environment GIS initiatives will have a far wider spectrum of questions to ask. Objectives will range from simple resource management to more complicated modelling procedures necessary for the implementation of health and environmental policies, see Table 6.1.

**Table 6.1** Some of the questions that a European 'global' approach to HEGIS may ask

a) What are the safe dosage limits for radiation and other environmental pollutants?

b) Monitoring and modelling of harmful factors, especially in terms of their impact upon man

c) What are the variations in rates of mortality and associated characteristics over Europe?

d) What environmental and/or social factors vary to cause these differences?

e) Modelling industrial and water pollution, to establish safe levels for policy issues

f) Resource allocation for new medical provision, where is it needed?

g) What would be the optimum location of hospitals for accessibility and requirements? and

h) What action should be taken in response to major disasters and what types of evaluation and forecasting procedures will be adopted to accommodate for the associated environmental risks?

Whilst a smaller pilot study approach to HEGIS will have more focused priorities, as outlined in Table 6.2. This table also serves to summarise all the issues which have been referred to thus far in this research and the questions that Stage IV in 'The GIS Process' will attempt to answer.

**Table 6.2** Some of the key questions to be addressed in the search for causes of ALL

a) Where can unusual disease incidences be found?

b) Do the cases tend to occur in certain places rather than others?

c) If patterns can be observed, to what extent are they real, or do they occur by chance?

**If so..**

d) Why are there an increase in cases here and not there?

e) What may be causing these patterns?

f) Is it environmental, generic, viral etc?

**Assuming an environmental cause..**

g) What factor of the environment is sufficiently different to cause this variation over space?

h) Is there a single or multiple set of environmental factors responsible?

**If so**

i) What aspects of the environment should we be looking into further?

**Ultimately..**

j) How does this cause lead to the development of malignancies in children?

## 6.2 Preparing the data for spatial operations

This step therefore was referred to in Chapter 5 involving the creation of complete, clean and compatible databases for GIS analysis.

# STAGE III

## 6.3 Performing spatial operations

Step 6.2 provided the basic GIS database framework and coverages. However these require further manipulation in addition to transformations and the creation of surfaces, especially if they are to be used in any practical GIS analysis. This is the function of Stage III. Table 6.3 provides a summary of the main commands employed to achieve the latter, and a more comprehensive list can be found in Appendix D. Obviously the command names and structures will vary slightly from one software package to another but the usages they depict will be generic and readily available in most GISs.

**Table 6.3** Commands used heavily in the manipulation of data in this research

| Commands: | Subsystems: | Modules: |
|---|---|---|
| BUFFER | ARCEDIT | TIN |
| DISSOLVE | ARCPLOT | VIEW |
| IDENTITY | INFO | |
| INTERSECT | | |
| NEAR | | |
| POINTDISTANCE | | |
| RESELECT | | |
| STATISTICS | | |
| TABLES | | |
| UNION | | |

Spatial operations of one kind or another are performed on every database involved in this application, however in many cases the techniques used and the eventual outcome is very similar. For the purpose of this chapter a sample of databases which conveniently demonstrate many of the characteristics and requirements of analysis will be selected.

In the first instance therefore spatial operations are needed to convert basic databases into something more meaningful in terms of analysis, these include establishing, covariates, areas of impact and the tagging of data with environmental information.

### 6.3.1 Attaching population covariates

As mentioned in Chapter 4 the incidence of childhood cancer is a fairly useless statistic on its own. Some supplementary knowledge of the overall population at risk is also required. This involves combining two separate datasets; (a) the incidences of ALL from the cancer coverage, and (b) the population counts from the census database stored as enumeration districts (EDs). This can be achieved by using the NEAR command in ARC/INFO which serves to assign an individual cancer case in one coverage to the nearest ED in another. However, the task is not that simple! As with many GIS operations certain commands sound very exciting and just what are needed but the output they produce are not the ideal results you had hoped for. For example, when a NEAR is used to combine cancers with their respective EDs, the attribute information which accompanies the census database are in fact dropped. The desired link between cancer incidences and the population coverage has not yet been fully achieved. What this resultant database does contain however, is an all important internal reference number for the matched coverage, ie ED# which can then be used to join the rest of the attributes from the original datafile using the command JOINITEM. Then the coverage is complete.

### 6.3.2 Areas of Impact

The creation of areas of impact are mainly needed to establish the range of influence that both linear and point sources of air pollution can have upon the surrounding population. This is particularly relevant to the manipulation of the road network, which is employed as a surrogate for the chemicals lead and benzene, associated with exhaust fumes. Again the NEAR command could have been used, which would establish the nearest road to a point. However in terms of analysis this would at best provide a series of graphs and tables summarising the number of cases of ALL according to their proximity to certain roads. In addition, the use of this command could lead to an undesired loss of information, because only the nearest road is stored in the resultant coverage. Thus in a situation where the incident is found very close to

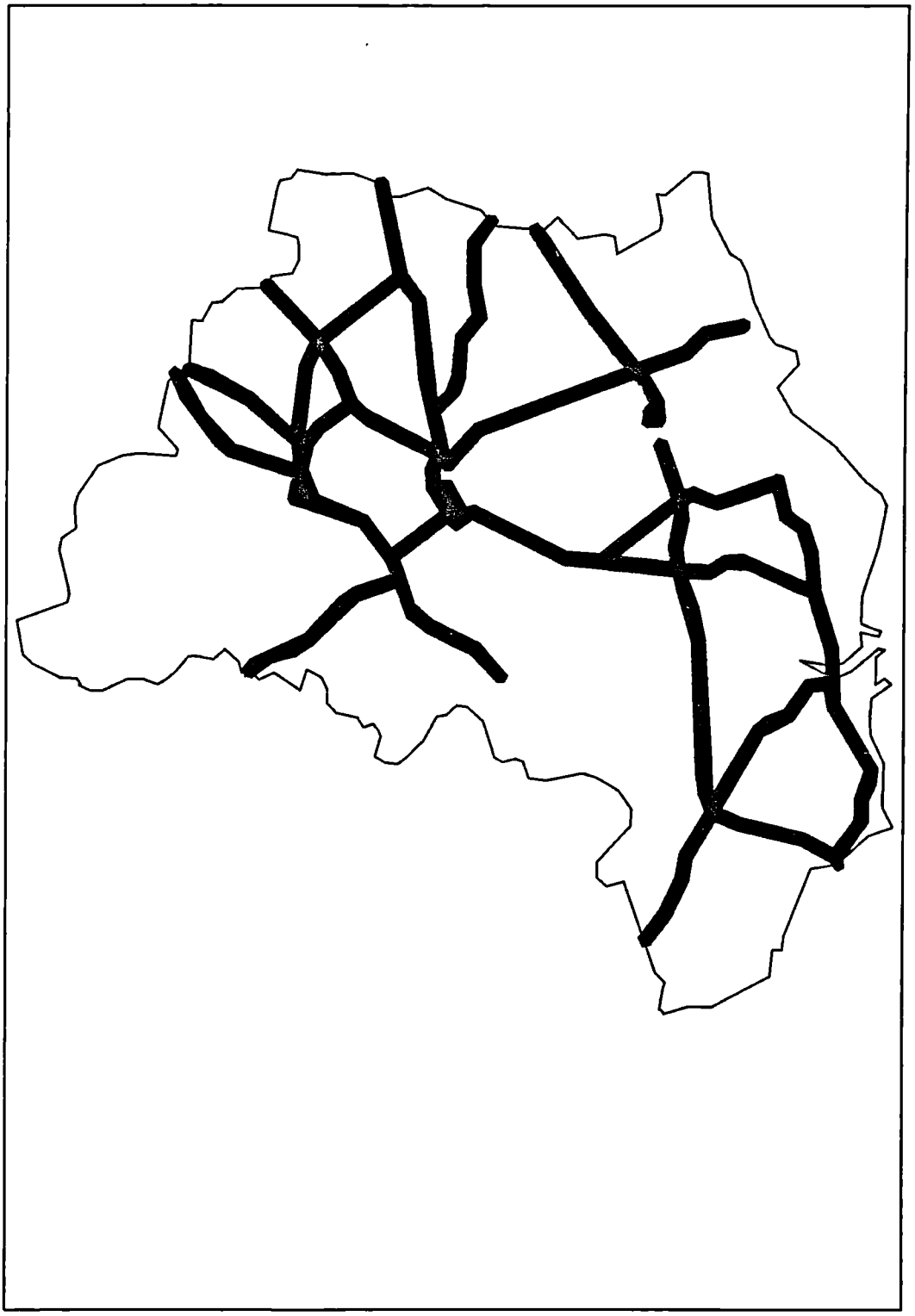
a minor road but is also within a reasonable area of impact associated with a nearby primary road the effect of lead pollution up on that individual may be under-estimated.

For the purpose of this research, the area of impact was created using the BUFFER command. This essentially creates a corridor around the linear feature. The distance can be determined subjectively or, wherever possible, based on detailed studies. Field studies relating to these types of areas of impact were in fact available from the previous Tyne and Wear research (Raybould, 1989). This suggested a maximum area of impact of 250 metres and this was constant irrespective of traffic density. However variations around this optimum distance were adopted as a means of making allowance for any errors which may have been created in the digitising of the road network and the 100 metre resolution of the cancer data. Thus initially the incidences of ALL were analysed according to their presence within 150, 250 and 350 metre corridors around the network and this was carried out irrespective of the traffic density.

One of the less desirable features of this type of spatial operation is that where buffer corridors overlap they become merged into one large area of impact, see Figure 6.2 which demonstrates this by mapping the 250 metre buffer for primary roads in Tyne and Wear. Simply buffering the road network as a whole therefore does not produce the most realistic representation of the influence of lead and/or benzene pollution. In addition, using this type of coverage can mean that any significant cancer distributions which may occur along a particular stretch of the road can not be observed. Conversely, increased incidences of ALL may go undetected because the creation of a large polygon would simply dilute the probability statistic used to locate areas of interest.

Consequently further refinements of the road network is required to render a more realistic analysis of the association of ALL and lead/benzene pollution. This is the subject of step 6.7 in this 'traditional' analysis approach and involves distinguishing between links and junctions. The isolation of roads by status to form separate coverages is also highlighted, indicating that these situations involve different pollution characteristics.

Figure 6.2: A 250m buffer for Primary roads in Tyne and Wear



### 6.3.3 Environmental tagging

In order to explore the environmental correlates of ALL in statistical terms the cancer incidences need to be physically linked with the environmental databases. Thus taking the two databases discussed in this chapter as an example; the cancer information plus population, would need to be tagged with information concerning road corridors and associated characteristics. This is achieved using the command IDENTITY. Figure 6.3 provides a map of the two coverages superimposed which enables a visual interpretation of their relationship. Alternatively, IDENTITY superimposes the coverages but also creates a separate coverage containing both sets of information which can subsequently be interrogated using the methods outlined in step 6.4. Table 6.4 summarises all the new coverages which were produced in Stage III and these in turn will form the basis for the analysis carried out in Stage IV.

## 6.4 Preparing derived data for tabular analysis

Two commands are particularly relevant in this step, including STATISTICS and FREQUENCY. These both provide an option to create a limited number of summary statistics in INFO. In addition INFO provides other commands such as CALCULATE which can also be used to achieve fairly basic statistics, for example, the rate of cancer incidences can be calculated by dividing the number of cancers by the population at risk. These are very useful if the epidemiologists' approach is one of descriptive analysis of ALL, as was the case in Chapter 3.

Unfortunately though these commands are limited to providing only the sum, the mean and the minimum/maximum values of a set of items. Problems also arise due to the structure of ARC/INFO whereby only 500 different item classes can be dealt with at any one time. This is not particularly useful when the datasets employed are fairly large. For instance, summarising cancer counts by wards in Northern England is problematic because there are 683 in total, and this would be magnified ten times over if the study region was to be extended to a national or even European scale. Thus to overcome this limitation there would be the added inconvenience of splitting up the original database into smaller and more manageable files, executing the desired

Figure 6.3: Visually superimposing Primary roads with incidences of ALL

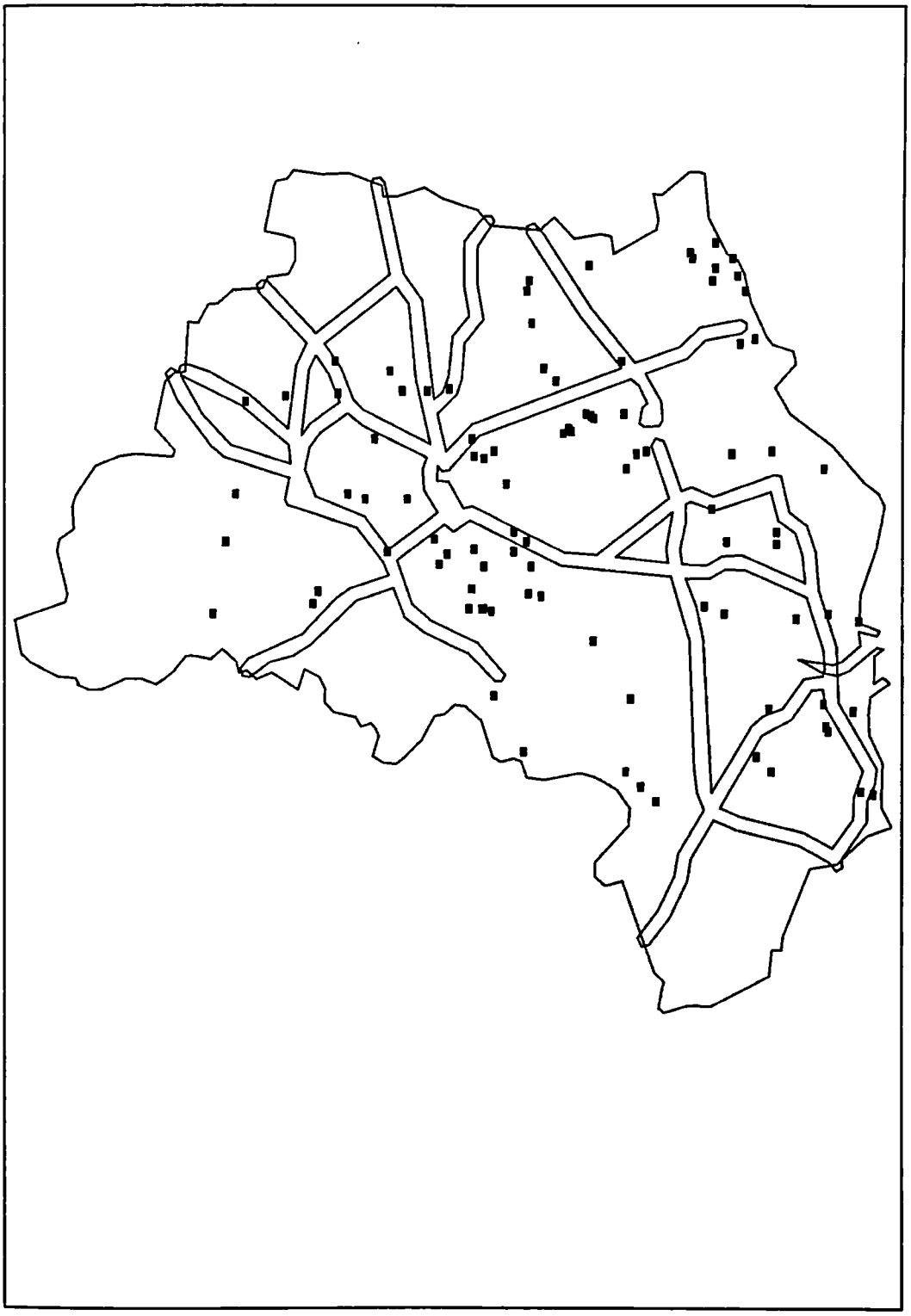


**Table 6.4:** New coverages produced in Stage III for use in GIS analysis

| FEATURE | COVERAGE | REFINEMENT | AREAS OF IMPACT/NEW COVERAGE |
|---|---|---|---|
| Linear | Roads | Buffer | 150m, 250m, 350m |
| | Roads | Reselect | Other,Main,Primary, Motorway |
| | Roads | Nodepoint | Junctions |
| | Railways | Buffer | 150m, 250m, 350m |
| | Estuaries | Buffer | 400m, 600m, 800m |
| | TW Overheads | Buffer | 100m, 200m, 500m |
| Points | Mines | Buffer | 1km, 2km, 5km |
| | Special site | Buffer | 1km, 2km, 5km |
| | Power station | Buffer | 1km, 2km, 5km |
| | Waste sites | Buffer | 1km, 2km, 5km |
| | Incinerators | Buffer | 1km, 2km, 5km |
| | TW Substation | Buffer | 100m, 200m, 500m |
| Areal | Geology | | ] |
| | Rainfall | | ] These already have |
| | Back radiation | | ] attribute codes and |
| | Landuse | | ] are therefore |
| | Nland Bracken | | ] suitable for analysis |
| | TW Smoke | | ] without further |
| | TW Sulphur | | ] manipulation |

Note: TW - Tyne and Wear
Nland - Northumberland

The buffer distances that have been used to manipulate the original databases are, wherever possible, based on field studies from previous work in this area, namely Raybould (1989). In other cases they are determined through subjective opinions as to the possible area of impact that may be realistically assumed for these types of pollution outlets, bearing in mind that the aim of the research was to explore the potential of GIS through a general search for environmental correlates of ALL rather than to determine a specific cause of childhood cancers themselves.

function and then combining the resultant statistics at a later stage. At present this is the only response to such software foibles.

Admittedly every new system will have a number of problems and these are usually written into the documentation, but not until after the aims and benefits of the command have been emphasised first. There is a strong argument therefore for an 'alternative' guide to standard GIS manuals which would act as a quick reference to software limitations, but in addition it would provide possible solutions to these common inadequacies. This could have the advantage of reducing the 'learning curve' for anyone else intending to embark on the GIS route, by preventing them from wasting valuable research time in 're-inventing the wheel' as they too discover these minor technicalities, this is an aspect which is discussed further in Chapter 10.

## 6.5 Performing investigative analysis

The importance of this step in the GIS analysis procedure is that at this stage a realistic attempt at answering some of the questions in Table 6.2, of section 6.1, can begin. Five broad areas can be identified which essentially typify the extent of GIS as an analytical and investigative tool. These concern issues of what is at? where is it? what has changed since? what spatial patterns exist? and what if? questions, all of which are illustrated in the following.

### 6.5.1 LOCATION: What is at....?

This involves querying the system to find out what features are present within a particular area of interest. Many of the figures illustrating the databases in chapters 3 and 4 were examples of 'What is at?' questions, since these provided a view of the environmental factors and their distribution throughout Northern England.

### 6.5.2 CONDITION: Where is it?

In spatial epidemiology transferring the child cancer registry into disease distribution maps may be viewed as a 'where is it?' type of GIS analysis. Figure 3.1 showed the distribution of ALL in Northern England and to the medically trained eye this in itself may be sufficient to render an additional insight and hypothesis as to the causation, or

at least something unusual in a particular location. The aim of the spatial element therefore is to provide some animation to the tables and records, as well as automating the 'sticking pins in a map' type of approach which was the epidemiologist's previous means of spatial analysis. It also increases the flexibility that the epidemiologist has to carry out map manipulation and data interrogation procedures. Thus the answer to the 'where is it?' question may provide a platform from which other more pertinent questions may evolve, some of which may be easily followed up using a GIS if the relevant databases are available.

### 6.5.3 TRENDS: What has changed since..?

This exploits the temporal element of databases studying the change in patterns of incidences over time, Figure 3.7 in chapter 3 showed maps illustrating the distribution of childhood cancers over four three yearly intervals from 1976 to 1987. This example did not provide any obvious distinctions between time periods. However, if an increased number of cases had occurred in a particular period this may have been interpreted as an indication of some event or change in the surrounding environment. Hence time can be an invaluable surrogate for other elements of change which subsequently could focus investigation in a particular location.

### 6.5.4 PATTERNS: What spatial patterns exist?

This type of query begins to take the question raised in section 6.5.2 of 'where is it? one step further by attempting to deduce whether patterns are unusual or are the norm under the circumstances. GIS is possibly ill-equipped to fully answer this, despite the fact that it is probably the most useful tool for an epidemiologist who is searching for causes of ALL.

Basic patterns can be observed, such as simple point patterns including the distribution of ALL in relation to the road network, as shown in Figure 6.3. However the information presented in this diagram is meaningless without some indication of the population at risk. Alternatively patterns can be observed that depict variations of incidence rates based on different areas and their population base, a method which is often used to present mortality and disease data, Gardner et al (1983) and Kemp et

al(1985). The impact and benefits of this form of approach to disease mapping will be discussed in Chapter 7.

### 6.5.5 MODELLING: What if?

This question is not directly relevant to this research since the causation of ALL is unknown. Therefore the modelling for a 'what if?' situation cannot be applied. However, it is perhaps one of the more useful methods of analysis that GIS offers, and is particularly relevant at the European Health and Environment GIS scale. Policy issues can be formulated on results from such models. An example of a 'what if?' model therefore is the siting of nuclear installations where selections upon key databases can be made to isolate desirable/undesirable areas based upon accessibility, population, geological criteria etc. From the combination of these databases the site which exhibits both maximum safety and minimum environmental effects can be deduced. The use of GIS in this type of approach has already been carried out successfully in the UK (Carver 1991).

Other important 'what if?' facilities in GIS include; ALLOCATE and NETWORK. ALLOCATE can model the distribution of resources from one or more centres, and thus is particularly useful in applications for ambulance and vehicle routing, and monitoring response times. Whilst NETWORK enables the simulation of all elements, characteristics and functions of networks as they appear in the real world. For example, replicating the movement of electricity from the power station to the customer across a network of power lines. Consequently ALLOCATE and NETWORK could be employed by a European HEGIS at a more strategic level.

The culmination of this step of the 'traditional' GIS analysis procedure would be a series of maps and summary tables. Thus the next course of action would be to provide an;

## 6.6 Evaluation and interpretation of results

This is the aim of Chapter 7 which summarises Stage IV of 'The GIS Process' and provides a means of evaluating the potential of this new technology as a spatial epidemiological tool.

At this point in the analysis procedure the end-user may be satisfied with the results and they may have gained sufficient insight to the problems they had set out as a part of section 6.1. Alternatively, the end-user may begin to realise that GIS does not really answer some of their more fundamental questions. They may even be left with more questions than they originally started out with, and with no available methods to satisfy their curiosity. One solution to this is provided in the next section, whilst Chapters 8 and 9 explore this area further.

## 6.7 Refine the analysis procedure

Steps 6.1 through to 6.6 highlights one of two conflicting aspects of the GIS analysis procedure, (a) some of the limitations involved in these initial stages but also (b) its ability to stimulate ideas and suggest other research avenues. In both cases the end-user may wish to respond by reformulating their objectives and/or manipulating databases so that they more readily meet their aims. For example, step 6.3.2 referred to the inadequacies of simply buffering the road network as a whole to represent lead and benzene pollution. It was suggested that benefits could be accrued from further database refinement involving more emphasis on road characteristics. Assumptions made included the fact that the higher the road status the more likely it is to have an increased traffic density and an associated increase in pollution from exhaust fumes. It would be advantageous therefore to break the roads up into different types ie motorway, primary, main and other roads and then repeat the procedures outlined in steps 6.4 and 6.6. The rates could then be interpreted based upon this new set of criteria.

Another line of thought involved the isolation of key areas of impact along the road network, such as junctions as opposed to the open road. As it may be assumed that exhaust fumes tend to be more concentrated as a result of slow moving vehicles and low gear changes. This type of modification is made possible in GIS by using the NODEPOINT command which can extract points at the intersections along the linear road network creating a separate coverage to represent junctions alone. These new points can then be buffered and the rates/probabilities of cancer can be compared between junctions and the open road.

It is noted at this point in the analysis procedure that GIS technology should be used sensibly for any database refinements. It is fine to manipulate the data by using GIS techniques if it means that more realistic results can be obtained. However the analytical questions themselves should not be altered to fit GIS capabilities, since this would defeat the object of using the new technology in the first place. Chapter 7 will highlight similar areas for concern and deficiencies in the general GIS toolbox and suggest key areas which should be reviewed with respect to both generic applications and that of the specific needs of spatial epidemiology.

## 6.8 Production of final maps and tabular reports of the results

Again this is the essence of chapter 7, and from these results the extent to which GIS can search for environmental correlates of ALL will be tested. Any lessons to be learnt, the ways forward and the problems associated with this type of GIS approach and resultant presentations will also be discussed.

# CHAPTER 7

# THE GIS PROCESS: STAGE IV
# DATA PRESENTATION, MAPS AND MORE MAPS!

This chapter outlines the final stage in the GIS Process. This is the first real test as to whether GIS can succeed in terms of a spatial epidemiological tool. The maps produced in this chapter will provide evidence of GIS's ability to execute all the functions of traditional cartographic techniques, but also demonstrate the potential for exploring spatial relationships between mapped data.

## 7.1 Map based analysis

'Map based analysis' is not a new concept. In fact the first cancer atlas was published in 1928 by Stocks, who presented Haviland's (1855) findings on the variations in mortality between counties in England and Wales. Particular attention was focussed upon the increased risk of cancer in the North of England. The use of administrative boundaries for reporting differences in cancer incidences continues to be a popular method of disease mapping. Numerous atlases have been produced based upon spatial epidemiology, the most recent of which include Gardner (1983), Kemp (1985) and Alexander(1990).

Figures 7.1(a) and 7.1(b) demonstrate the impact of mapping cancers in relation to administrative boundaries, and the variations in distributions which can be deduced from using relatively simple statistical techniques. Figure 7.1(a) emphasises the locational aspect of individual incidences by simply mapping the raw point data for ALL onto the ward boundary coverage for the Northern Region (1981). Figure 7.1(b) uses the same data, but takes into account the underlying population at risk by calculating incidence rates per 1000 population. This is simply achieved by taking the total number of cancers and dividing them by the total population at risk, ie. 0 to 15 year olds. The figure derived from this calculation is then factored up by 1000 to obtain a sensible number for interpretation and mapping purposes. In the process some detail is lost due to the aggregation of populations and cancers to ward level. However the benefits to be accrued from this technique are based on increased information regarding the identification of spatial patterns such as clusters of

Figure 7.1a:Distribution of ALL cases according to wards in Northern England

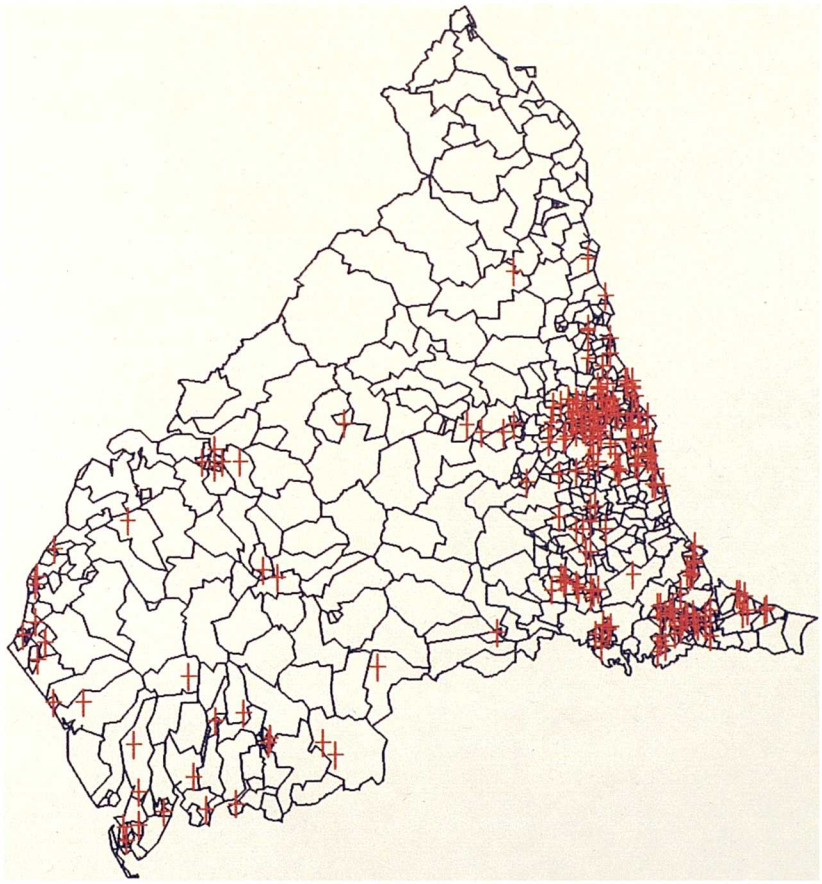
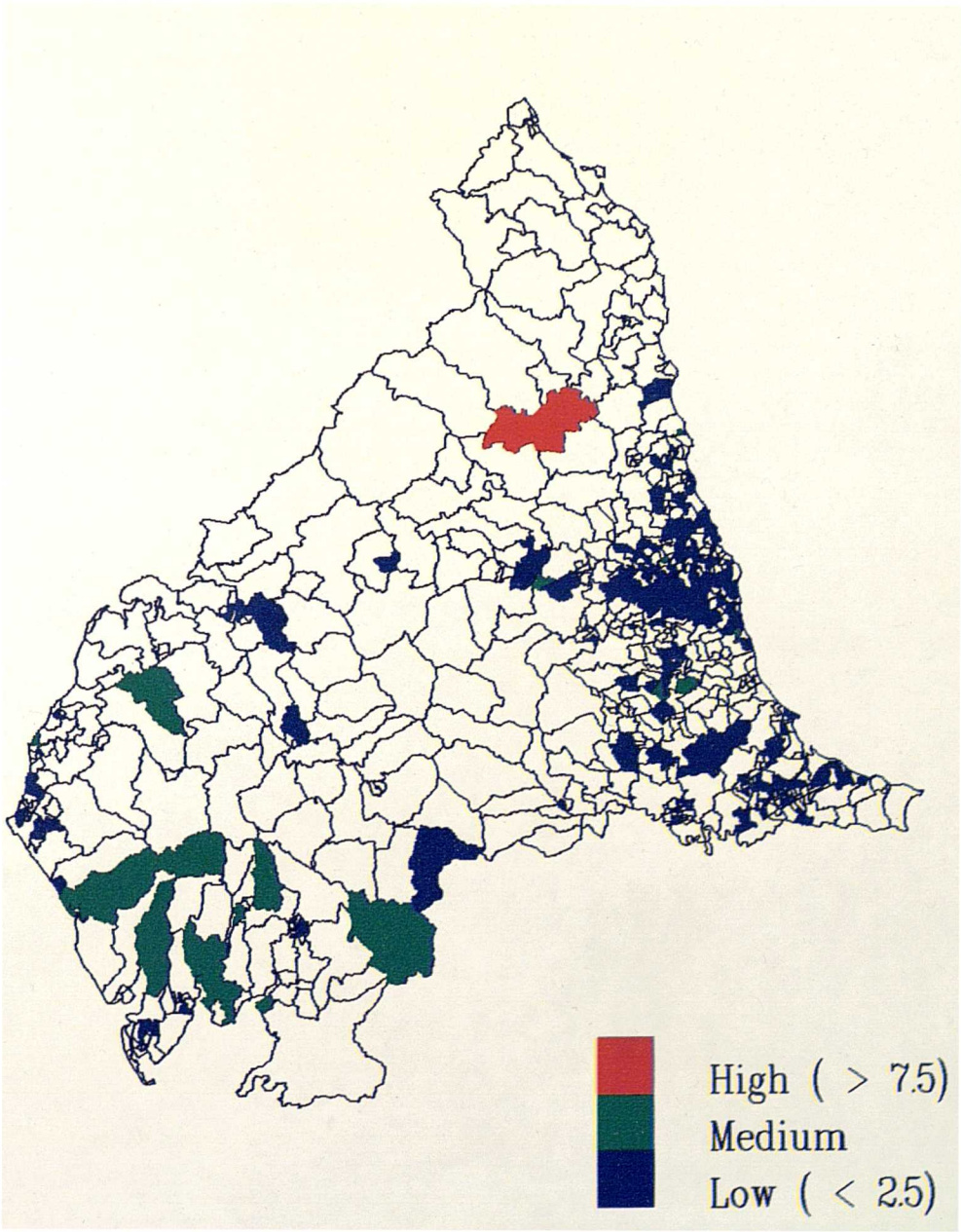


Figure 7.1b: ALL rates per 1000, by ward



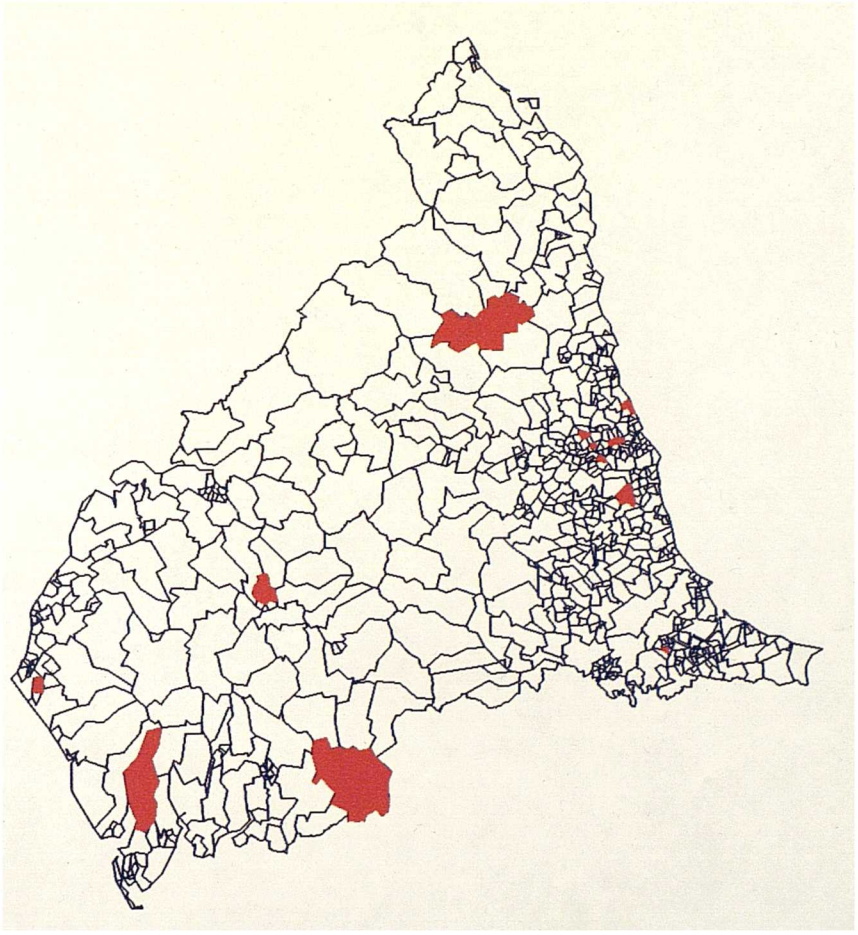
neighbouring areas which appear to have similar rates. It can also flag areas which appear to have fairly high or low rates compared to the region as a whole. These results may then be sufficient to stimulate further investigation into the possible underlying effect of features such as the environment and socioeconomic factors which characterise these particular wards.

Figures 7.1(a) and 7.1(b) represent 'rough and ready' methods for achieving a quick perspective of the distributions of cancer cases in the region. These can easily be produced within any GIS framework once the cancer and population counts are known for each ward. This would simply involve adding an extra ITEM to the existing database and using the CALCULATE command in INFO to establish the value for the ward rates. However this method is by no means sufficient in terms of providing a level of confidence in the relationships that can be observed between the ward rates mapped. In statistical terms the calculation of rates fail to take into account the effect of ward size. Yet large wards frequently have smaller populations and are typically characterised by large standard errors. Conversely, smaller more densely populated districts tend to have smaller standard errors which suggests that the actual rates observed are closer to the true rate.

The map in Figure 7.1(c) therefore uses the same ward boundary delineations and incidences of ALL, but a form of significance has now been assigned to the distributions using the Poisson probability statistic. Again this takes into account the underlying population but also helps to compensate for the problems of standard error noted with the last example on ward rates. This statistic is described in greater detail in section 7.2.

These figures illustrate the various ways in which the distribution of ALL can be presented. For instance, the impact of Figure 7.1(a) by simply overlaying point incidences with other coverage basemaps in order to interpret possible relationships is not always reliable. In this case the eye is distracted by the concentration of cases in urban areas, yet these are to be expected given the size of the associated population at risk. Figure 7.1(b) and (c) highlight the fact that it is wards with only a relatively few number of incidences which are of particular interest. Out of the three illustrations therefore Figure 7.1(c) provides the most apt view of the wards which have unusual incidences of ALL compared to the region as a whole. However even these may prove to be mis-leading. Since the significance level which was set for this exercise

Figure 7.1c: ALL distribution, significant wards at the 5% level



was that of 0.05, in other words there is a possibility that one times out of twenty the wards which are depicted as significant may have occurred by chance. Therefore given that for the Northern Region there are 683 wards in total, it would be expected that about 34 wards would appear by chance based up on this five percent significance level. Thus with only 12 significant wards actually highlighted in Figure 7.1(c) it could be suggested that there is nothing unusual happening at the ward level. The use of the multiple testing procedure could overcome this problem but the effect of this will be discussed in section 7.2.

There are a number of advantages to this type of map-based analysis approach and these can be summarised as an ability to (i) yield a synoptic picture, (ii) replicate spatial structure, (iii) provide models of the real world and (iv) identifying blatantly obvious spatial patterns. All of these can facilitate understanding and provide an extremely good communication medium for expressing and exchanging ideas about the real world. These however should be viewed in the light of at least three areas of deficiency which also accompany this approach.

Firstly, it must be remembered that the view portrayed by these maps is essentially static, for instance Figure 3.1 in Chapter 3 actually covered a decade of diagnosed cases of ALL, with no indication of how these patterns had actually evolved over time. Figure 3.7, attempted to overcome this problem by using the flexibility of GIS to reselect on temporal features stored in the cancer database, such as the date of diagnosis. At best this resulted in a series maps for each given time interval which again only served to produce a snap shot of the patterns that existed. The development of computer movies (Moellering 1980), is a step forward towards dynamic mapping, although the concept of dynamic and multimedia GIS is very much an issue for the future. The development of dynamism however would also serve to alleviate the second problem which is inherent in map-based analysis, whereby maps attempt to depict spatial structure. This is usually the product of spatial processes which will have evolved over time and thus again the static nature of maps would cause certain features to be masked or worst still misinterpreted altogether.

The third deficiency of maps concerns that of the map makers themselves. Maps can be produced using misleading or inadequate data and can then be displayed in equally misleading and inadequate ways. To some extent the maps in this section are a good example of this, because they attempt to reflect variations in the incidences of ALL

based upon totally arbitrary boundaries. These boundaries have been defined for administrative reasons as opposed to representing patterns of health and the processes which may effect these (Openshaw, 1982). Some of the significant wards only contain one cancer case and this begs the question, can this be considered a 'spatial pattern'? Thus messages which are portrayed by a map, or combination of maps, can depend upon both the users interpretation and the compilers' preconceptions, both of which serve to complicate the information which is portrayed.

This section therefore provides evidence to contradict the often expressed opinion that maps are somehow 'simple'. Maps attempt to represent models of reality which are in themselves extremely complex. At this stage of the 'map-based analysis' process many of the researchers involved in the mapping of diseases cease to offer any further interpretations on the patterns produced at this administrative scale. They pass pertinent questions of possible environmental causes on to those considered more qualified to answer them, in this case the epidemiologist.

The maps presented thus far in this chapter are no different to those which have been produced for many years using simple graphics packages and traditional cartographic techniques. So what other benefits can be achieved from storing this information within a GIS framework? In effect the development of health and environmental databases in a flexible spatial data handling tool, such as GIS, will carry the map makers contribution to new levels by providing alternative boundaries for investigation based on natural or pseudo natural delineations, such as the difference between one geological type and another. These provide an exciting medium for alternative research as well as to emphasise the limitations which behold naturally biased administrative units. Section 7.3 demonstrates the potential of GIS to employ these databases and investigates the environmentally based causative factors which may be attributable to the development of childhood cancers such as ALL. It will be interesting to see if the same locations that were flagged by this ward based analysis remain significant when the research basis is reformulated to take into account natural and/or man-made impacts on an individual child's health.

Firstly though reference is made to the main statistical technique which is employed in this research, since it forms the basis for the map-based analysis procedures in this chapter and subsequent techniques discussed in Chapters 8 and 9.

## 7.2 The Choice of Statistic

The Poisson probability statistic was adopted in this research as a means of measuring the relationship between ALL and all other coverages found in this GIS framework. The Poisson probability has a discrete distribution which assumes that each incident of ALL is independent of every other case registered, and that they occur at random in both time and space. This forms what is commonly known as the 'Null Hypothesis'. When the statistic is significant this suggests that data observed do not conform to the Null Hypothesis of randomness. It is not known what distribution ALL does conform to, thus in this case it is a very minimalistic test. Nevertheless the result will provide a measure of the deviation from the expected occurrence and in turn indicate possible clustering in a specific location. This is essentially what is demanded in many spatial epidemiological applications such as this, which involves the determination of possible patterns in rare diseases.

In general, the calculation involves establishing an overall mean of events, ie the mean incidence rate of ALL for the whole of the Northern Region. Then all the subcategories within the database under investigation, ie wards in Figure 7.1(c), are compared to this overall mean. This is achieved by comparing the expected number of cancer incidences (the mean rate for the region multiplied by the population in each individual subcategory, such as the ward) with that of the actual number of incidences ie. the number of ALL cases which have been diagnosed in the ward. These two values are used to calculate the Poisson probability employing the equation;

$$P(k) = \sum_{k=1}^{n-1} \frac{\lambda^k e^{-\lambda}}{k!}$$

Where, n is the number of the observed cases of Acute Lymphoblastic Leukaemia;

$\lambda$ is the expected number of cases;

e is the exponential constant;

P(k) is the probability that an area will contain k points.

The Poisson Probability is:

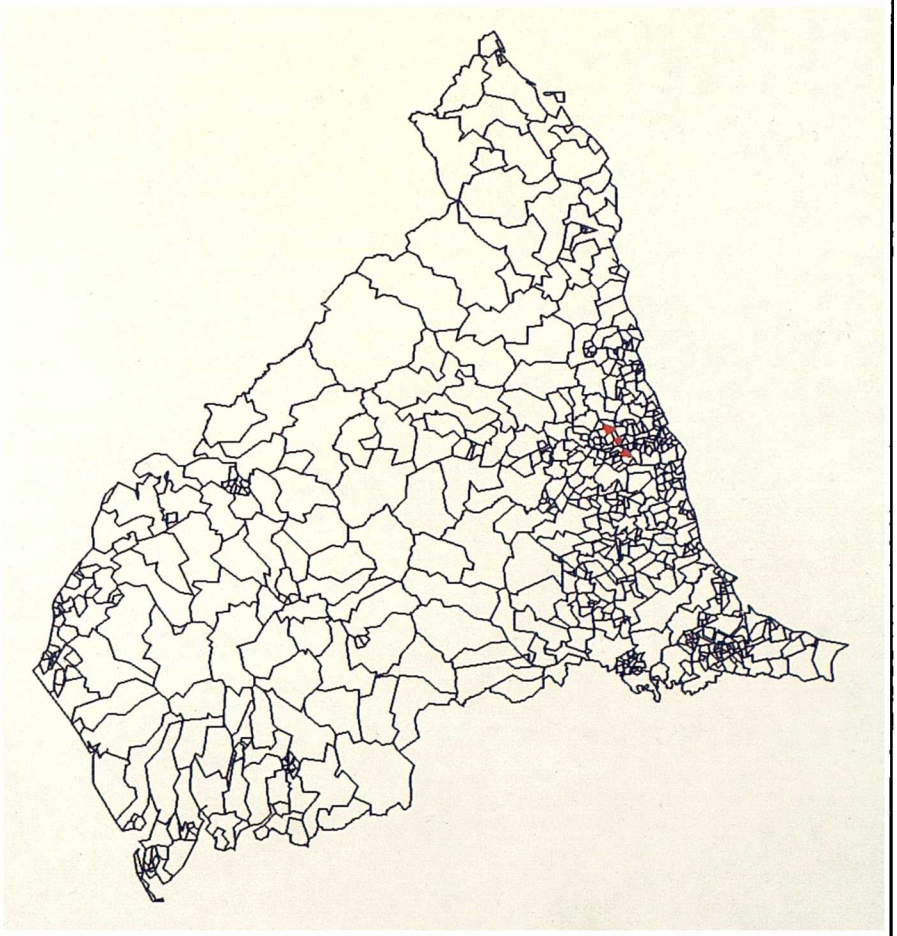
$$1 - P(k)$$

The value returned by this calculation defines the significance level for the distribution, which is usually set at 0.05. The Null Hypothesis can be rejected if the value is less than 0.05, because this suggests that the events occurring in that particular ward are more extreme than would be expected for the region as a whole, and thus refuting the idea that the distribution of ALL is random. A 5 per cent confidence level means that only once out of twenty occasions will the results be incorrect. This was discussed in section 7.1 with respect to the number of significant wards which were established under the Poisson probability. These wards should not necessarily be disregarded completely though just because of this law of the Poisson statistic, but rather that any further interpretations should proceed with caution.

Two types of error can occur when employing the Poisson probability statistic. One is a 'false positive' (Type I error), whereby the test picks out a significant distribution when in reality one does not exist. Conversely, there is an error which produces a 'false negative' (Type II error) whereby the test fails to identify a cluster which does exist. These faults however should be viewed in the light of the fact that datasets employed are not always perfect anyway. For instance, more cases may be observed in certain wards as a result of better diagnosis procedures rather than there being any physical tendency for a greater number of malignancies to develop. In this case, the test may suggest a cluster when in fact there is no real difference, rather it is an artefact of data due to the human element effecting diagnosis procedures.

However, ensuring that data are input into the system as accurately as possible should minimise these problems, and as Chapter 3 noted, the Child Cancer Registry does undergo rigourous checks to maintain an accurate database. An important factor here is that not only must the cancer data be correct but also the population at risk data, but it has already been noted that this is only taken every 10 years in the form of a census. Thus another way in which some Type I errors can be reduced is to use multiple testing techniques which lowers the significance level at which the Null Hypothesis would be rejected. Alternatively, for a less computationally demanding approach the test could be carried out at a much lower significance level, for example, the Null Hypothesis would only be rejected when the Poisson probability value returned was less than 0.01. This would provide the researcher with a 99 per cent confidence level in the clusters which may have been observed. Figure 7.1(d) represents the result of lowering the significance level to 0.01 for the ward example.

Figure 7.1d:ALL distribution, significant wards at the stricter 1% level



The effect of the latter more rigourous approach to testing could be problematic as it limits the results to only very strong clusters. However, this may not be desirable in an exploratory context especially when dealing with rare diseases such as ALL, because it may exclude clusters which are still interesting epidemiologically. Thus in this type of application it may be interpreted as only serving to reduce the power of the overall Poisson test (Besag and Newell, 1991). For Figure 7.1(c) in section 7.1 and all future use of the Poisson probability test therefore, the significance level will be set at 0.05 and any multiple testing will be ignored. This is the preferred course of action because it will ensure that small clusters are captured, leaving the final decision as to what clusters are, or are not, real to the epidemiologist. On a more practical level, multiple testing is computationally demanding and thus to some extent its use would defeat the object of having quick and effective maps at your fingertips. Also to include this type of methodology into a GIS framework would be extremely cumbersome.

Despite the Poisson probability analysis being a very basic statistical test it was not available within the GIS software package itself. Instead, a complementary FORTRAN program was written to access data exported from the INFO datafiles, this program was then run to calculate the Poisson probability and the results were returned in a suitable format to be imported back into the existing database. This was integrated into an overall ARC Macro Language program which exists in ARC/INFO. Macro languages are extremely advantageous for developing programs which repeatedly carry out the same set of procedures. For instance, calculating significance levels of ALL for all the environmental databases and adding the resultant values to the existing datafiles involved the same set of commands to be used a minimum of 20 times. Consequently, the use of an AML saved a lot of time and repetitive keystrokes. (Appendix E lists the basic AML used and shows how it accesses the FORTRAN program working on a Unix workstation).

Finally it should be emphasised that the Poisson probability is used in this research 'as an aid to intelligent thought and pattern identification rather than a substitute for it!'. Thus the maps produced and any areas highlighted as significant under this test must be interpreted in the light of the factors raised in both this section and those concerning the pitfalls inherent in map-based analysis.

## 7.3 The GIS Response: Is there an environmental cause for ALL?

This involved the analysis of the new coverages produced in Stage III of 'The GIS Process' based on the areas of impact established for certain sources of pollution. A purely visual approach to establishing relationships between the distribution of ALL and individual environmental factors was discounted from the onset. Whilst this may be of some superficial use for a quick 'where is it?' query, the complicated nature of certain maps, namely the geological database with some 30 different categories and over 500 polygons suggested that some method of summarising the visual overload was necessary.

Consequently the AML for running the Poisson probability test was used to establish whether any of the environmental coverages showed a one to one relationship with ALL. Initially this was executed at a very general level ie. the difference in probabilities were calculated according to the presence of cases inside or outside an incinerator buffer, main road, power stations etc., irrespective of its specific location. In the case of areal features such as geology, background radiation and landuse the statistic was calculated for each subcategory eg. individual rock types and landuse codes. Table 7.1(a) and (b) summarise the statistics which were derived from these preliminary analyses.

From these tables it can be seen that for the areal features the only significant categories to be picked out at the five percent level are those of rock type 21, 35 and 16 which are Andesite Lavas, Middle Lias and Silurian rocks of the Tarronian series respectively, illustrated in Figure 7.2. For landuse, agricultural land (grade 1 and 2) and woodland were highlighted as having a possible association with ALL cases, comparison to other landuse types in the rest of the region, see Figure 7.3. The other, and probably most interesting result is that of the 250m corridor around the railway network which had a Poisson probability statistic of 0.03449, shown in Figure 7.4(a). In fact out of all the point and linear coverages representing various aspects of the environment this is the only one at this stage of the analysis procedure to demonstrate any possible relationship with ALL. This is rather ironic considering that this was the sole database which was incorporated into the system simply because it was made available along with the road network, rather than it being specifically linked to ill health.

**Table 7.1a:** Poisson Probability Results: Preliminary analysis for linear and point features (at the 5% level)

| FEATURE | COVERAGE | DISTANCE | POISSON |
|---|---|---|---|
| Linear | Other roads | 150 | 0.908414 |
| | | 250 | 0.761498 |
| | | 350 | 0.347743 |
| | Primary | 150 | 0.432269 |
| | | 250 | 0.839117 |
| | | 350 | 0.811613 |
| | Main | 150 | 0.328491 |
| | | 250 | 0.908288 |
| | | 350 | 0.881161 |
| | Railway | 150 | 0.065924 |
| | | 250 | **0.034490** |
| | | 350 | 0.347743 |
| | Estuaries | 400 | 0.796143 |
| | | 600 | 0.399198 |
| | | 800 | 0.767337 |
| | Overheads | 100 | 0.727399 |
| | | 200 | 0.664217 |
| | | 500 | 0.401094 |
| Points | Mines | 1km | 0.642845 |
| | | 2km | 0.955601 |
| | | 5km | 0.346070 |
| | Special site | 1km | 0.225475 |
| | | 2km | 0.057480 |
| | | 5km | 0.126472 |
| | Powerstation | 1km | - |
| | | 5km | 0.240569 |
| | Incinerators | 1km | 0.819970 |
| | | 2km | 0.669352 |
| | | 5km | 0.280391 |
| | Waste sites | 1km | 0.359893 |
| | | 2km | 0.322039 |
| | | 5km | 0.398930 |
| | Incinerators | 1km | 0.076293 |
| | | 2km | 0.068452 |
| | | 5km | 0.060328 |
| | Substations | 100 | 0.504926 |
| | | 200 | 0.462913 |
| | | 500 | 0.405080 |

**Table 7.1b:** Poisson Probability Results: Preliminary analysis of areal features

| | | | CANCERS | | | |
|---|---|---|---|---|---|---|
| FEATURE | COVERAGE | CATEGORY | EXPECTED | OBSERVED | POPN | POISSON |
| Areal | Geology | 21 | 0.42 | 4 | 1406 | 0.001719 |
| | | 35 | 0.38 | 3 | 1268 | 0.010789 |
| | | 16 | 2.00 | 6 | 6633 | 0.032553 |
| | Landuse | 2 | 1.16 | 4 | 3483 | 0.049646 |
| | | 7 | 0.60 | 3 | 2000 | 0.035068 |
| | TWSmoke | | 1.52 | 7 | 5035 | 0.005691 |

Note: Those coverages which did not have any ALL cases associated with them are omitted from these summary tables.

Figure 7.2: Location of significant rock types under the Poisson Probability (5% level)

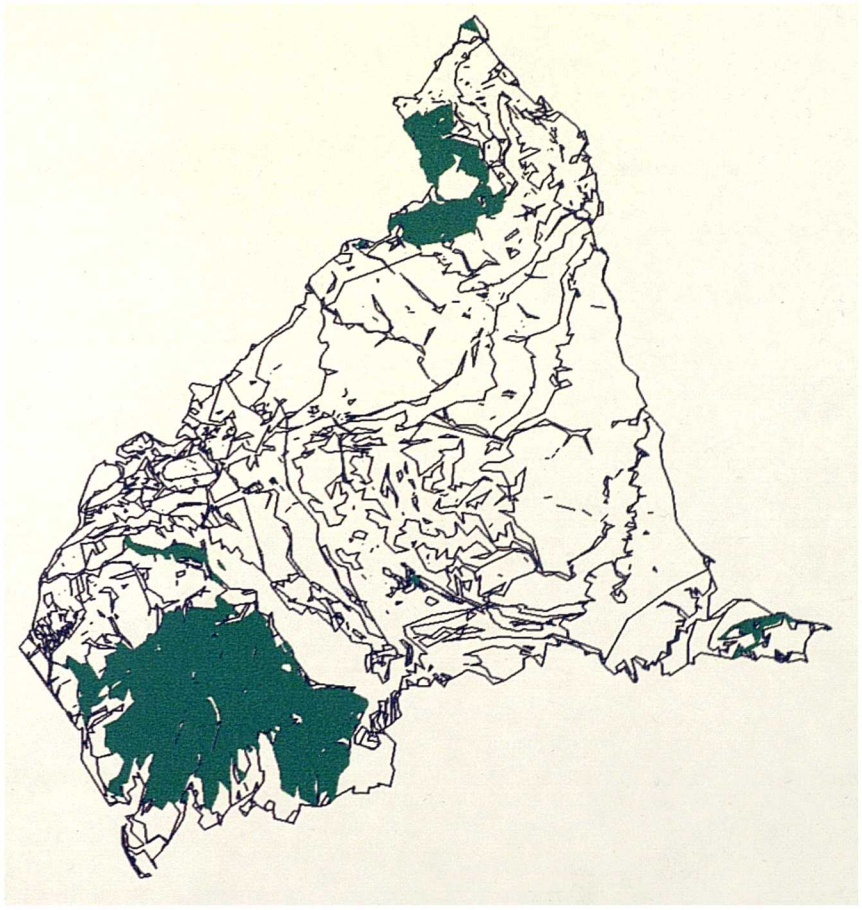


Figure 7.3: Location of significant landuse categories plus the distribution of ALL cases

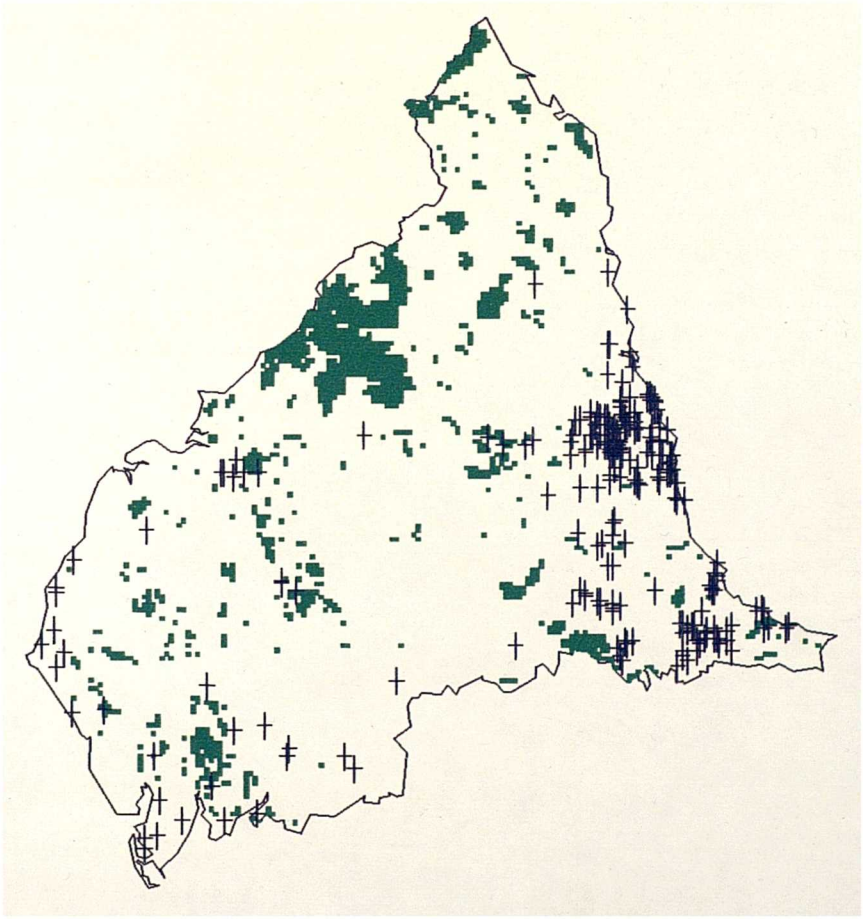
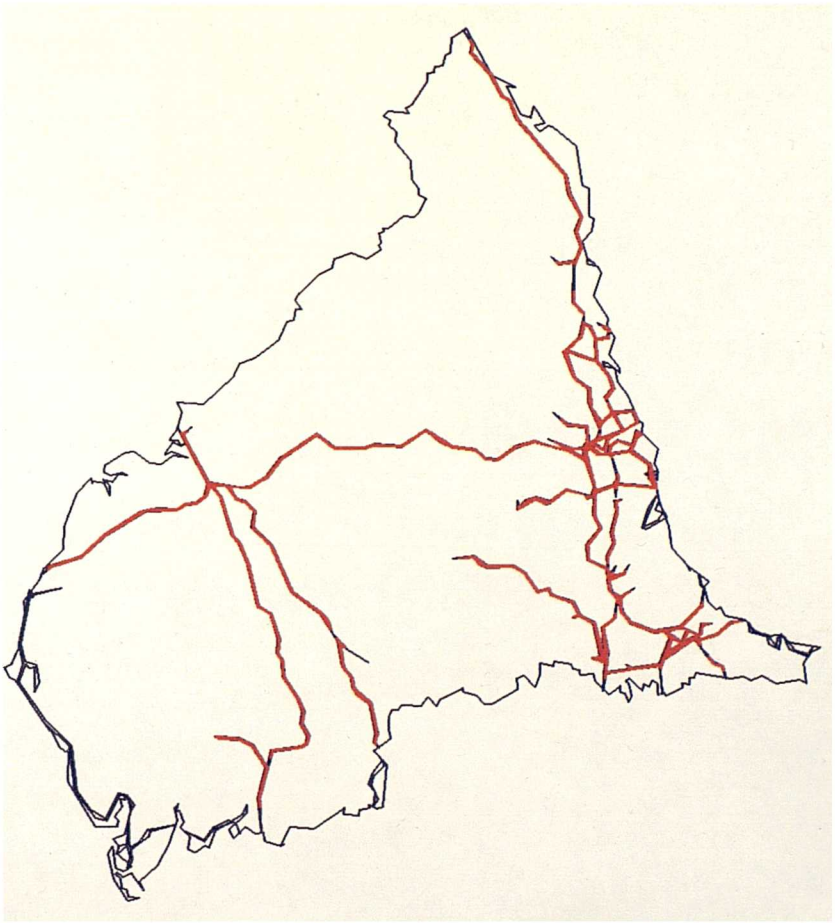


Figure 7.4a:The significant polygons for the whole of the railway network buffered at 250 metres



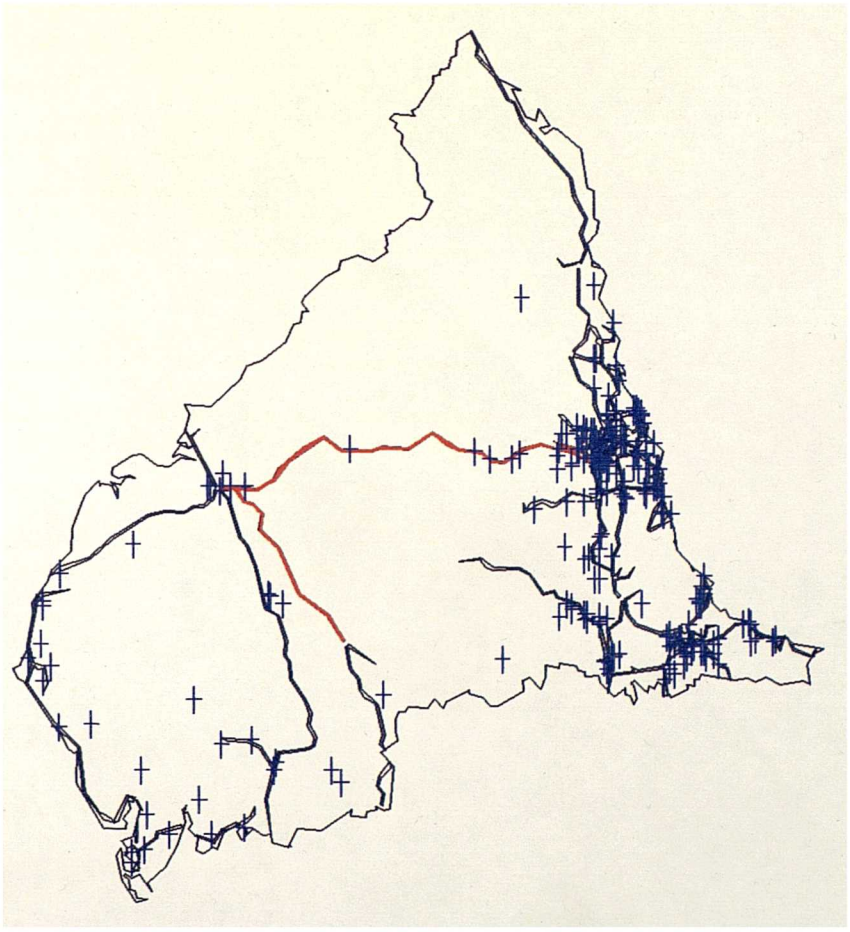
These tentative steps into the use of GIS to explore possible relationships with ALL exhibit some obvious short comings, not least of which is the loss of locational specificity due to the categorical approach of the Poisson probability statistic. For instance, two landuse types are picked out by the statistic as possibly significant, however when they are superimposed with the raw cancer data, as shown in Figure 7.3, it is apparent that this is not a universal relationship. Some of the significant areas are actually devoid of cancers and thus it may be that small areas with only one or two cancers are in fact driving the statistical analysis process. From this though it may be inferred that areas with agricultural land are a contributing factor in the distribution of ALL, but are not the only factor exerting an influence.

The key advantage of GIS at this point is the ability to re-introduce the locational aspect into the analysis, especially when certain categories have been flagged for further investigation. This therefore involves the employment of GIS for maximum benefit by exploiting its flexibility to manipulate data in a variety of ways, be it locationally, by attributes, or by overlaying and combining different datasets. This was referred to in step 6.7 of chapter 6 and involved the refining of databases to obtain a greater insight into the possible spatial processes which may be associated with these features.

For this purpose the railway network is a good example to adopt. This coverage suffers from the same problems as the road network in that the creation of a corridor around railway lines leads to almost one whole polygon covering the entire region, as shown in Figure 7.4(a). Ideally therefore it should be split up into smaller units, achieved by using the NODEPOINT command to deduce intersections and then from these the railway links and junctions can be extracted. The Poisson probability can now be calculated again but this time for individual polygons representing different parts of the railway track, using the internal number of each polygon to ensure that the locational aspect is retained.

The result of the latter was to isolate the significant area of impact to one section of the railway network, shown in Figure 7.4(b). The Poisson probability is 0.010673 and the expected number of cancers for the 74851 children at risk within this 250m buffer was only 22, but 39 actually occur. Thus it would have been these incidences which were sufficiently interesting to cause the whole network to have a significant value

Figure 7.4b: A more focused look at the impact of railways buffered at 250 metres



under the previous analysis. If the epidemiologist has started to get excited by these results, he or she may wish to refine the analysis further. This may involve a search into whether the relationship is expressed for all children with ALL in this area or whether it effects a particular age group, ie the Poisson probability could now be calculated for the main age group that is affected by ALL, that of 0-4 year olds. The latter would simply require a reselection on the cancer database to isolate the subset of cases which the epidemiologist wishes to investigate. Again a significant number of cases occur in this section of the railway line with 11 cases as opposed to the 3.9 expected. Although this only constitutes one polygon out of a maximum of 24 and thus under the Poisson probability test it could be discarded as a chance occurrence, it may be interesting to explore this avenue further given that this area persistently appears in various refinements of the analysis procedure.

This example therefore serves to demonstrate how both the health and environmental databases can be manipulated to obtain different perspectives on the possible ALL relationships. Thus the advantage of GIS in this instance was the power of inference ie. the ability to flag a particular environmental factor which may or may not be significant but from which the epidemiologist may begin to formulate hypotheses. The following provides a geographer's view of why the railway network has been highlighted. Firstly, it is noted that the relationship observed only existed at the 250m buffer level. It may be presumed therefore that for the other buffer corridors such as 150m there would be relatively few houses situated that close to a main railway in order to have an effect. Alternatively the relationship had disappeared at the 350m corridor and this can be interpreted in one of two ways; (i) The distribution exhibits an interesting distance decay effect in that whatever the environmental impact it is most prevalent at about 250m, or (ii) that the choice of 250m provided an area which suitably encapsulated the right number of cancers and population so that they appeared significant, thus the result is one of a boundary effect rather than any environmental relationship.

In the interest of epidemiology though the latter may be sufficient to stimulate a minor follow up investigation into the segment of railway that has been highlighted and its associated area. Various lines of thought can be generated such as, is there a certain type of freight which is transported along that line as opposed to the main routes through cities ie. does it carry radioactive material, harmful chemicals in the form of weed killers etc. More particularly is there a section along this line where

freight may be left to stand for any period of time during transit, which in the case of a possible leakage is more likely to effect the immediate micro-environment than say a train that is simply passing through. Alternatively, this railway line may be a proxy for some other factor of the area which exhibits a greater impact upon the causes of ALL. This may include the type of housing found near by which would lead to a possible socioeconomic influence to be hypothesised, or there maybe industrial factories in the area which deal with volatile chemicals and which are not included in the hazardous outlets set up for this application? In addition children have a tendency to play in these sort of areas and there may be some waste ground which has material illegally dumped that no one has discovered, but it is having a minor effect on the micro-environment sufficient to trigger of the development of child malignancies.

The latter all attempt to explain the possible localised effect of this particular stretch of the railway line, however a more universal relationship cannot be discounted. As with many of the databases employed in this research a significant relationship between ALL and the environment may not have been found because of the spatial reference that is available for the actual incidences. Children are mobile and thus just knowing their home address at the date of diagnosis may not be sufficient to tackle this complex problem of causation. For instance, in the last example it was suggested that children may play near railways, thus cases with homes some distance away from railways may have equal contact with this environmental factor but through their social micro-environment. It may be necessary therefore to also review more general practices that effect railway lines, in particular maintenance. For instance sewage is deposited on tracks whilst trains are in transit, and more seriously certain chemicals banned for general public use because they are known to be carcinogenic, are still employed to kill weeds on railway lines and embankments. British Rail are exempt from this ban because it is presumed that the general public do not come into any serious contact with railway tracks and thus are in no danger, but this may be a convenient assumption and not necessarily reality. This paragraph therefore has provided a few ideas based up on the possible link between ALL and the location of railways as afforded by this stage of the GIS analysis process, and the speculation could continue.

This provides an excellent example therefore of GISs ability to infer new lines of thought from subjective and exploratory investigation of spatial databases even if it is at a rather superficial level. However it may be asked, if this is a significant

relationship then why have other forms of analysis failed to pick it out? The answer may lie in the fact that other research has tended to focus on features of the environment which were considered favoured causes of ALL, railways would be well down that list, if on it at all! In addition analysis which has concentrated upon the distribution of ALL alone may not have picked out any increased incidences along the railway, since they tend to be based around a circular search rather than linear which this interesting result is depicting, this will be emphasised in Chapter 8 where cluster analysis techniques are discussed. Thus since other research projects by their very design would not have encountered this possible relationship it cannot be overwhelmingly discounted at this stage.

The next course of action would be to carry out similar analysis on the railway network in another area, or for a different time period. Preferably not in the Northern Region to avoid criticism of post-hoc hypothesis. This should be a relatively easy task for a national approach to HEGIS given the national distribution of the railway network, a national Cancer Registry and a GIS.

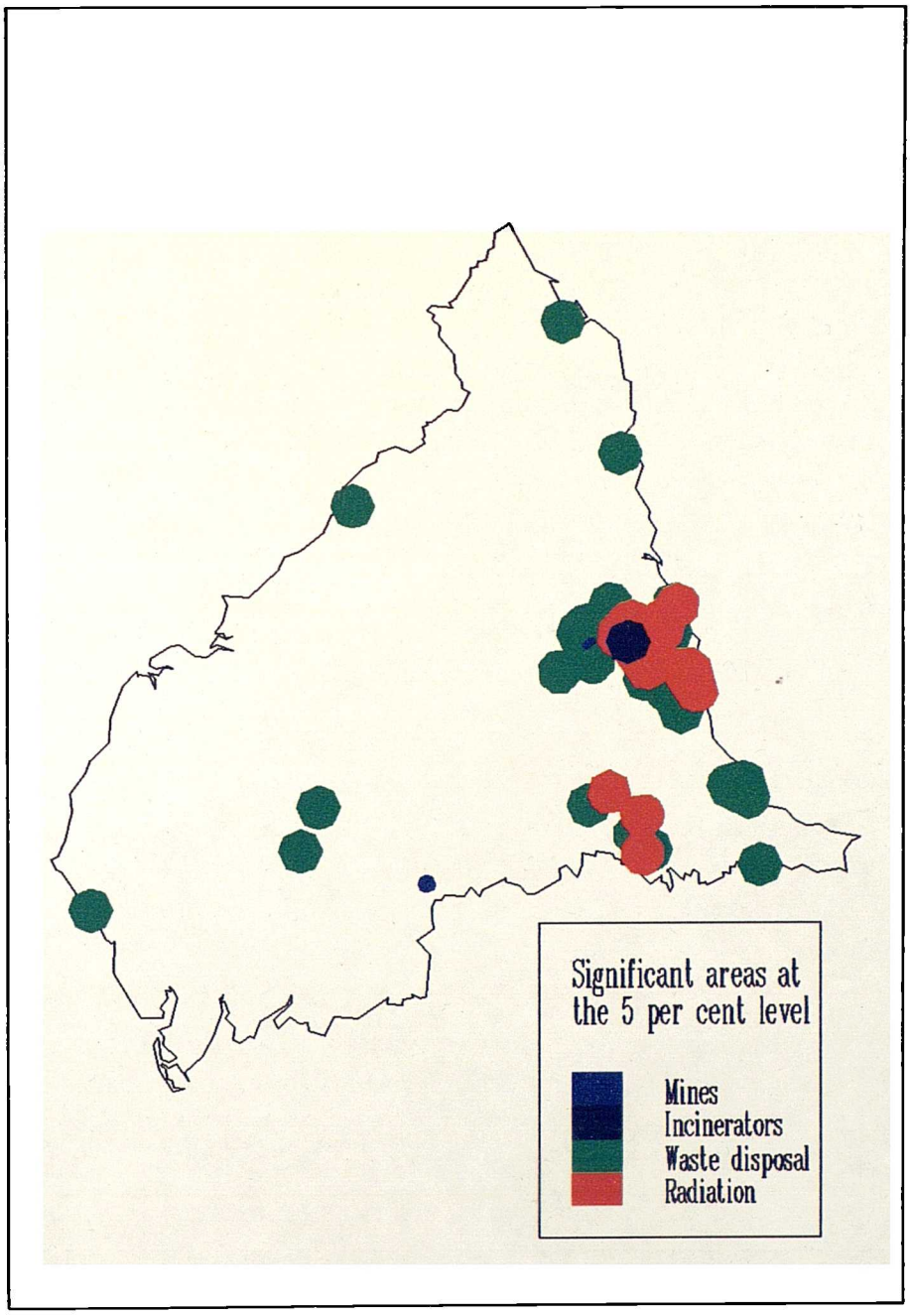
In the refinement of the railway network some benefit was accrued form isolating a certain area of impact by retaining the spatial aspect. Thus it may be interesting to now carry out this approach for the other areas of impact which did not appear significant in the preliminary analysis stage. Table 7.2 therefore summarises the specific coverage zones and the number of polygons which now appear significant under the Poisson probability. GIS has again facilitated the reformulation of analysis allowing the spatial component to be re-introduced and new areas to be flagged. Figure 7.5 for example illustrates the localised areas of impact which have appeared in this more focused analysis of point sources of pollution. Tyne and Wear appears to dominate but other highlighted areas are indiscriminately located around the region with no particular pattern. This may again be attributed to chance given that in all cases the number of significant polygons registered is less than that which would be expected under the Poisson probability, at the 5 percent level, see Table 7.2. Alternatively they may be exerting an influence on the cancers in that area but the main relationship with ALL is attributable to some other variable which is presently missing in this GIS framework. For instance, all these outlets may be associated with a similar type of housing.

**Table 7.2:** Poisson Probability: Locational analysis of the main linear and point features; Is there any additional areas of interest?

| FEATURES | COVERAGE | DISTANCE | NUMBER OF SIGNIFICANT POLYS | |
|---|---|---|---|---|
| | | | OBSERVED | EXPECTED |
| Linear | Other roads | 150m | 3.0 | 8.5 |
| | Other roads | 250m | 3.0 | 7.5 |
| | Other roads | 350m | 5.0 | 6.7 |
| | Main roads | 150m | 1.0 | 2.9 |
| | Primary roads | 150m | 1.0 | 1.5 |
| | Railways | 150m | 1.0 | 1.4 |
| | Railways | 250m | 1.0 | 1.4 |
| Point | Incinerators | 5km | 1.0 | 0.4 |
| | Mining sites | 2km | 1.0 | 2.6 |
| | Special sites (radiation) | 1km | 2.0 | 1.7 |
| | | 2km | 2.0 | 1.4 |
| | | 5km | 2.0 | 0.8 |
| | Waste disposal | 1km | 1.0 | 2.6 |
| | | 5km | 1.0 | 0.7 |

This table summarises the number of polygons which were found to be significant when the Poisson probability analysis was refined to take location into account, some of these significant areas (at the 5% level) are also illustrated in Figure 7.5. The table highlights that in many cases the number of interesting polygons are less than that which would be expected from chance occurrences. At the same time it emphasises that figures alone can mask any important spatial element which may be occurring, thus Figure 7.5 serves to demonstrate how maps can put these figures it to a more realistic, if not more comprehensible context.

Figure 7.5: Localised areas of environmental impact under the Poisson Probability



In the analysis carried out in this stage therefore a few interesting aspects have been noted but nothing glaringly obvious has been flagged as a definite cause of ALL. Given the nature of geographical data though the latter may be extent of any evidence which a GIS can support under the circumstances. It may be argued that if the relationship between ALL and the environment was simple it would have been found by now. On several occasions though it has been suggested that there may be a combination effect leading to the increased incidences of ALL in certain areas. GIS can therefore be employed to attempt to investigate this idea. The next example takes up a hypothesis which was referred to in Chapter 4 that suggested the possible accumulated effect of radiation doses within localised micro-environments.

In order to achieve this line of investigation further data manipulation is required to combine the relevant datasets. These include the sources of radiation such as sites with special licences, power stations, and the background radiation levels for the region. This is achieved by using one command in ARC/INFO, that of UNION, however the coverage that results is not sufficient for realistic interpretation as Figure 7.6(a) shows. The number of polygons have increased considerably and at present very little can be deciphered from simply superimposing the incidences of ALL on this basemap. Thus some end-user intuition is required to establish the difference in dosage which may be received by each individual ie the difference between the exposure of someone living within 2km of a special radiation site and in an area with background gamma ray levels of 6 bequerels, compared to someone who lives in an area where the only effect is background radiation of 3.5 bequerels. In other words the polygons which represent large doses of radiation should be distinguished from those with minor effects. This involved assigning a value to each individual effect giving greater weight to those areas with higher associated levels of radiation, these were then aggregated for each individual polygon. This is demonstrated in Figure 7.6(b) which provides a 3D surface of the resultant coverage, using the calculated cumulative dosage code as a z value on to which the incidences of ALL can be superimposed. This may look impressive, but on running the Poisson probability on this new coverage no significant areas were highlighted. However, the main factor is that GIS allowed this weak hypothesis driven by end-user curiosity to be investigated, be it at a very basic level.

Figure 7.6a: Combining radiation sources to look for a relationship with ALL

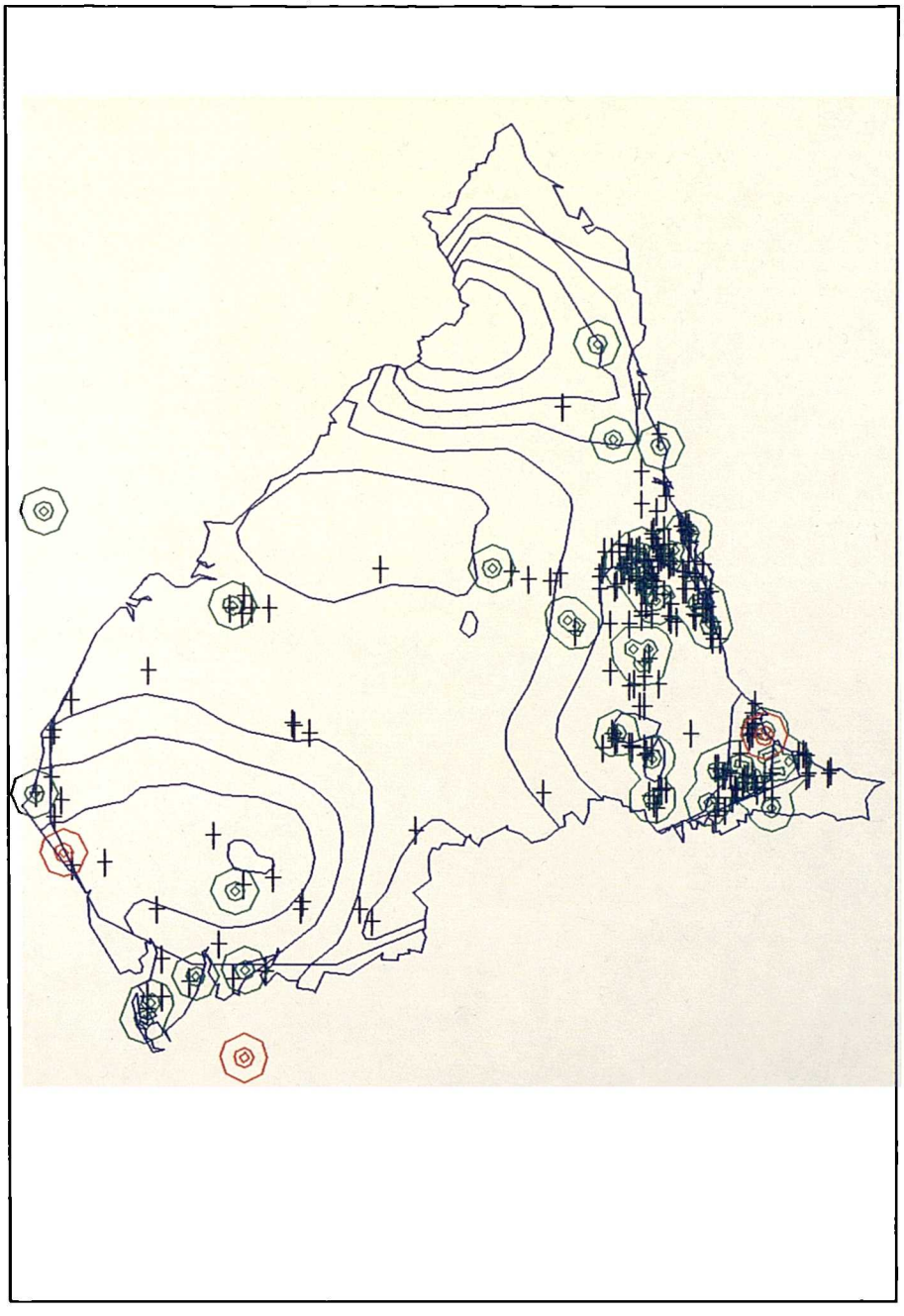
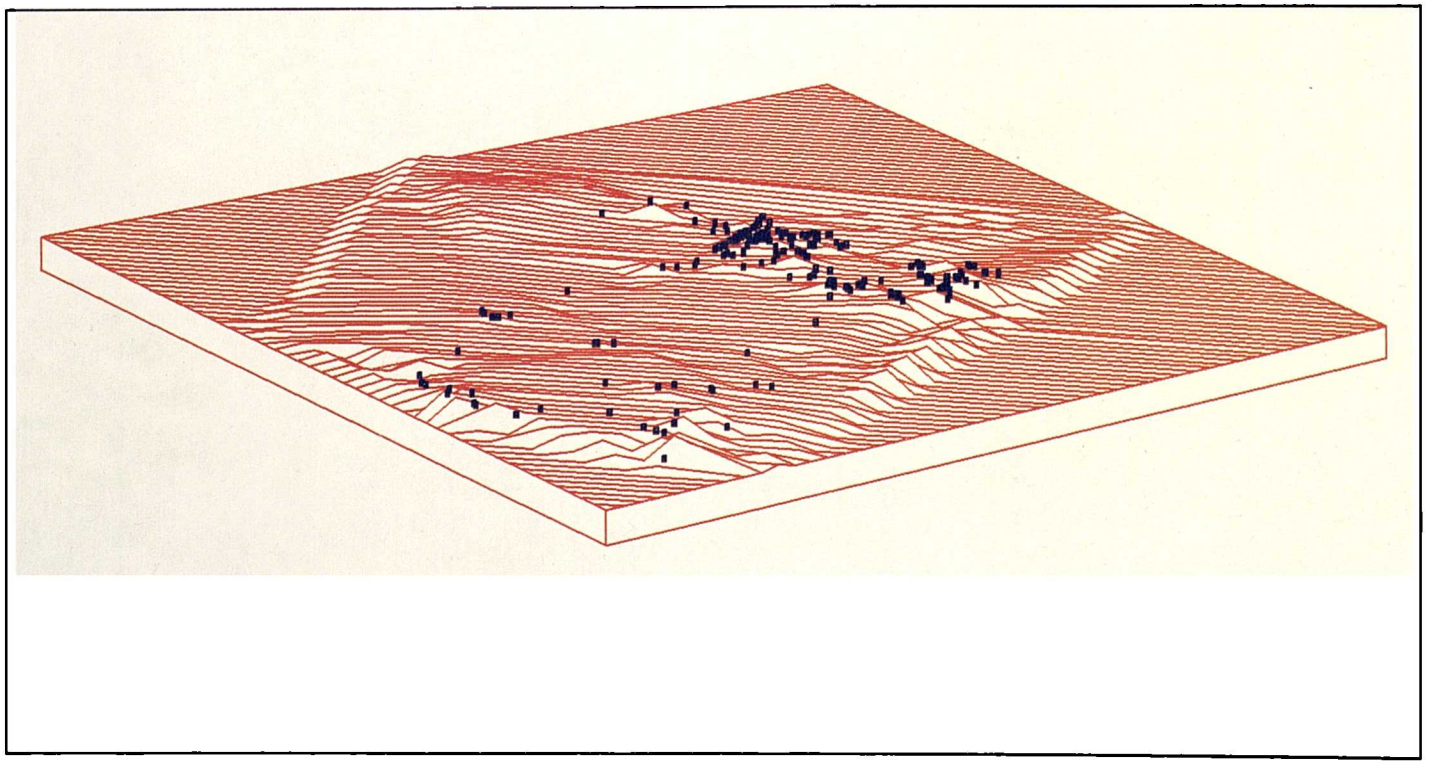


Figure 7.6b: A 3D view of the combination of radiation dosage and ALL cases

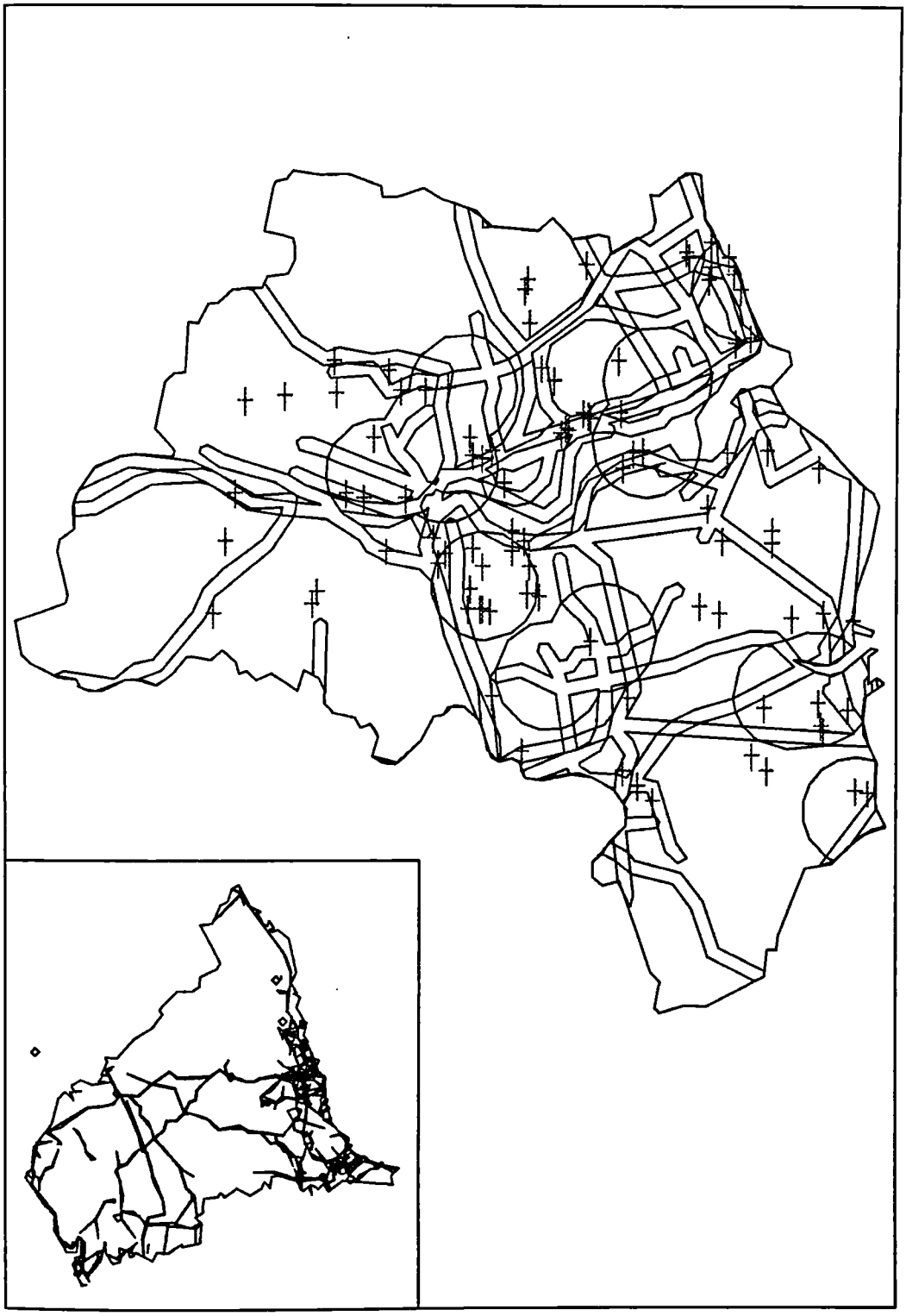


This form of analysis into the possible interaction of environmental causes is limited, despite GISs capability to overlay and combine datasets ad infinitum. This limitation is set both by the end-user and to some extent by the minor technicalities which effect GIS operations. In terms of the operator, the limit to overlays that can be assimilated in this attempt to deduce relationships between databases is at best restricted to three coverages, four at the most. This is because a subjective approach can no longer distinguish intelligently between the datasets presented, or realistically understand the spatial processes that resulted in the interaction of environments combined. Figure 7.7 shows the combination of main roads buffered at 250m the railways at 250m and special radiation sites at 2km. What does this mean?.(Remembering that two of these coverages are in fact acting as surrogates for other unmeasured aspects of the environment anyway, which only serves to add another dimension of complexity to analysis procedure). In addition by this stage the number of polygons which make up this new coverage will tend to the summary statistics limit of 500, and thus in order to perform simple Poisson probability requiring population and cancer counts for each item or polygon of interest there would be the added practical inconvenience of attaining summarised statistics. In terms of answering the epidemiological problem of ALL causation therefore GIS may have reached its limit, but as Chapter 9 attempts to demonstrate this may be overcome by exploring alternative 'relationship seekers' based on the databases already established within the GIS framework.

## 7.4 The benefits to the epidemiologist

What does all this mean in terms of exploiting GIS as a spatial epidemiological tool? Section 7.3 represented the final stage in 'The GIS Process' with a series of summary statistics and certain areas of possible interest. However even with the ability to refine the analysis procedure in GIS, there is nothing glaringly obvious to suggest an environmental causal effect of ALL, not in the eyes of the geographer at least. As stated from the onset though it is not the purpose of this research to provide 'proof', just an alternative methodology to an unsolved problem. What was deduced from the last section though may be sufficient for the epidemiologist who is just beginning to get to grips with this new technology and the potential of so much spatially referenced data for them to play around with. Already GIS provides them with capabilities they did not possess before, ie. to view the Cancer Registry spatially, to have a flexible framework for testing out their own vague ideas, and a query system to answer simple

Figure 7.7: Main roads 250m, railways 250m and special radiation sites 2km. Is there a relationship?



questions of where are they? Of course GIS may not have solved all the epidemiologists problems but in terms of a comprehensive HEGIS it must be remembered that it has a number of additional tools which are extremely useful, for instance the ability to delineate hospital catchment areas, model environmental pollution and so on. Thus the result of building and implementing GIS may have served to stimulate some interest into data spatial handling and may even get the epidemiologist slightly excited!

As a means of evaluating GIS though, the latter stage did raise a number of factors, particular concerning the potential of GIS as a spatial epidemiological tool. Essentially this research has served to anticipate some of the future needs of the epidemiologist as their attentions switch to complex spatial analysis problems rather than the production of pretty maps.

## 7.5 Visualisation: Is it the key to GIS success?

On the positive side, it may be argued that map presentations are far more advantageous than the standard means of presenting data such as graphs, tables, and pie charts. Simply compare the figures from Chapter 3 and the table in section 7.3. Which are more informative in terms of assimilating knowledge about the possible environmental causes of ALL?

Although it should be noted that epidemiologists or geographers do not need to be efficient in spatial analysis or map compilation to realise the inadequacies and problems that behold choropleth mapping, which in the long run is what GIS is good at. They suffer from areal problems which have been highlighted more than once in this chapter, and can look highly impressive yet spurious just because of the class intervals and shading criteria adopted in their production. Any hypotheses therefore which may emanate from the mapping of these simple relationships must be followed up and interpreted very cautiously.

However it may be argued that the choropleth nature of the maps may serve to partially mask specific locational aspects but at least they do not hide it completely by a series of figures arranged in columns to form an essentially aspatial table. In addition the maps produced have a very useful summarising effect, which reduces information overload and ensures that data can be processed quickly and more

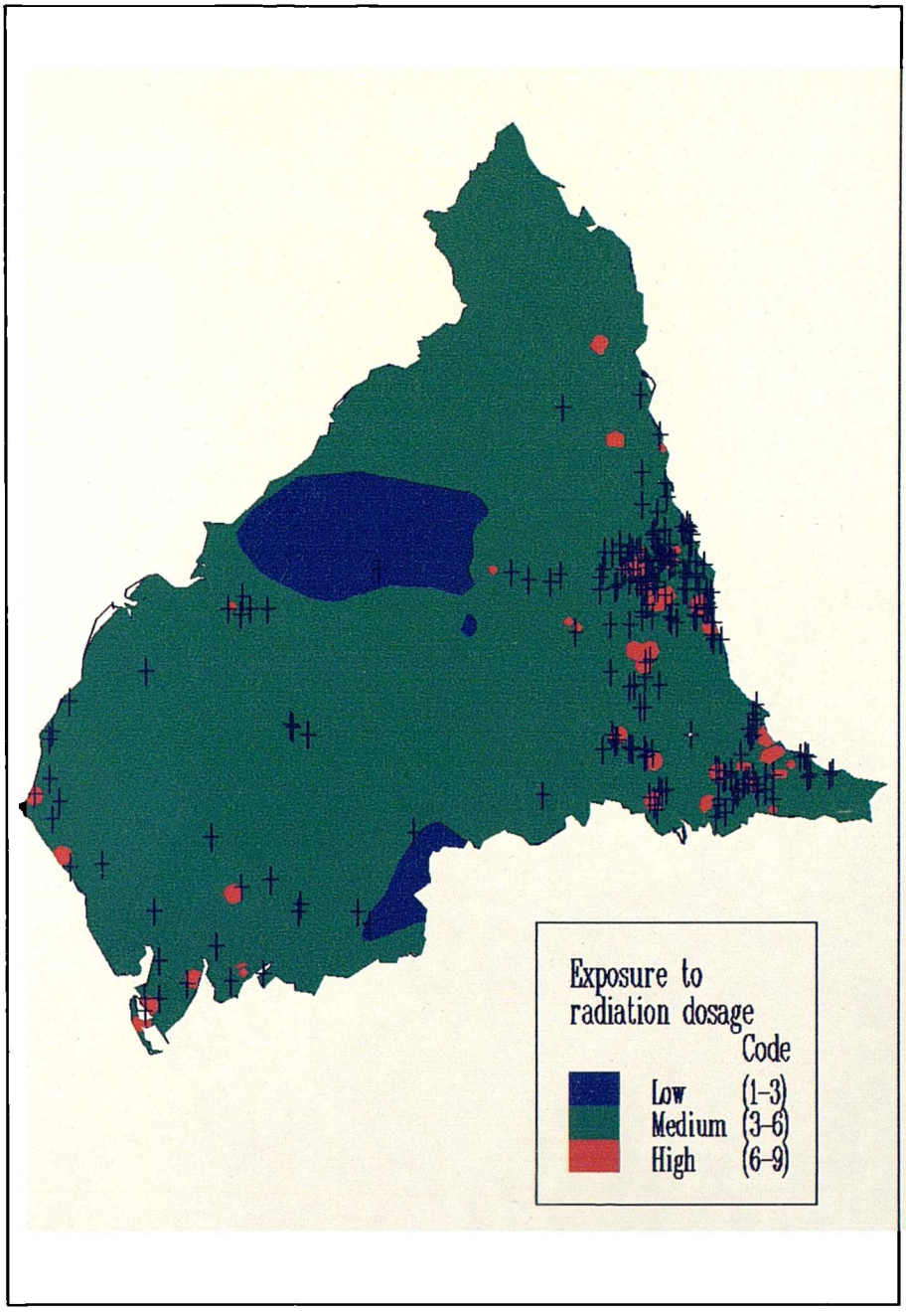
effectively by the end-user. They are therefore extremely powerful disseminators of information. Most people can read thematic maps, and this vital information can be communicated to a wider audience of expertise.

As mentioned previously, maps can also act as a positive feedback mechanism by providing the researcher with a means of following and checking the progress of research. One example is the verification of spatial encoding of the data where incidences which have obviously been assigned with the wrong postcode for the area concerned can be spotted. Secondly it is very satisfying to be able to visualise the exploits of the data capture and manipulation stages, seeing data put to good use may instill a desire for greater accuracy in the initial stages of data acquisition. This may be more relevant at the organisational level where the end-user of GIS is completely detached from the data gatherer and developer of GIS. The result being that both fail to fully appreciate the importance of each others contribution to the system and the benefits which can be accrued from the sharing in experiences and knowledge derived at every stage of the process.

A well designed map therefore offers the epidemiologist a number of benefits, both in terms of descriptive analysis and as a decision support tool. However, GIS and the maps produced should carry some sort of 'health warning'. GIS can allow poorly designed maps to be created just as easily as good ones, with the result that the spatial information conveyed is false. Figure 7.8 represents the cumulative background radiation dosages which were subjectively assigned by this researcher, these have been summarised into a 2D choropleth map in order to provide an overview of the possible differences in areas. However what does low, medium and high mean, are they actually realistic despite the message that is be conveyed by the map. This illustrates that results can be totally spurious and only really serve to demonstrate what can be achieved with a lot of imagination and very little understanding in what is required of this map-based analysis. The warning is that the power of visualisation in GIS can be easily abused. A good analogy here may be to take the saying 'garbage in garbage out' which rephrased to reflect the abuse of GIS may read 'garbage in pretty map out!'

Finally it should be noted that maps by their very nature, are not honest and objective tools for important decision making. There is a considerable amount of subjectivity involved in every stage of the GIS process. These limitations are not a criticism that

Figure 7.8: Another view of the possible cumulative effect of radiation and the distribution of ALL cases



the technology is too simple to be effective as spatial epidemiological tool, far from it. Rather that the questions being asked of the technology are going beyond the realms of description. What is needed now is some form of objective statistical testing to either prove or refute the ideas that these maps have generated. This issue is not one to be taken lightly, and in fact during the course of this research others in the GIS field have begun to realise the need for spatial analysis techniques in GIS.

## 7.6 The Missing Link!

It would seem therefore that there is one particular link in the GIS framework which is missing, and this has been identified as the need for basic spatial analysis functionality. Although this is not necessarily the view of the software manufacturer's where a review of promotional GIS software literature tends to contradict this finding. For instance;

> 'Analysis and map manipulation - capabilities of ARC/INFO include map overlay, buffer generation, file structure conversion and tabular analysis..'(Doric Computer Systems 1990)

Taking an example of another competing software manufacturer, the situation remains the same,

> '..sophisticated functions and processes that enhance the analysis of geographical data..'

different words and phrasing, but the qualification of this statement reads almost identical

> '..Included in these options are the abilities to generate buffer zones around features, remove common boundaries between area features, and pattern features'

The argument put forward in this thesis is that the techniques referred to in these definitions are extremely useful by virtue of the sophisticated mathematical functions and equations that they employ. However, they do not necessarily achieve the goals that the end-user has in mind. The problem of 'Spatial Analysis' therefore seems to lie

in its terminology! To the end-user, sections 7.3 offered nothing more than a means of altering, combining and producing new maps for visualisation and/or decision support. To them this is spatial mathematics to achieve important map manipulation requirements.

The next two chapters are designed therefore to offer some ideas and methods of alleviating the problems highlighted by this section. Drawing attention to alternative and essentially innovative techniques will hopefully serve to strengthen the GIS framework and provide it with a more optimistic and exciting future. Research like this, and academia in general, is well placed to tackle this type of initiative because it possesses the data, technology, expertise and working environment which will not passively accept the inadequacies of systems. Instead it acts proactively to overcome problems and improve upon functionality, wherever possible. The effect of this is to hopefully maximise the benefits that can be accrued from the technology available, as well as prevent less fortunate end users from failing in the GIS application world due to GIS's shortcomings or their disillusionment in the technology when their objectives are not met.

# SECTION THREE:
# EXTENDING GIS FUNCTIONALITY AND LOOKING TO THE FUTURE

# CHAPTER 8

# COMPLEMENTARY SPATIAL ANALYSIS I
# PATTERN SPOTTERS

The completion of Stage IV of 'The GIS Process' served to demonstrate that this new technology has successfully advanced the skills of the epidemiologist, particularly in terms of spatial data handling. They now possess a number of additional tools for quick query, flexible data manipulation and a highly impressive communication medium in the form of a map. Thus GIS has provided an exciting and acceptable alternative to data handling and theory testing. However, as Chapter 7 suggested once the databases are established and the epidemiologist has exhausted the data query and map making features their attentions will begin to focus on more pertinent questions. They will realise that their interpretations of causative factors and/or distributions of ALL which can be obtained from GIS results only provide a subjective approach to spatial analysis.

In the light of Chapter 7 therefore it would seem that the objectives which were set at the start of this research were perhaps overly adventurous in terms of 'state of the art' GIS technology. In turn it demonstrated how GIS techniques can be pushed to their limits. Chapters 8 and 9 are designed to offer alternative methodologies to help tackle the problem of searching for environmental causes of ALL. Thus these chapters attempt to anticipate some of the future needs of GIS as a spatial epidemiological tool and identify tools which will extend the potential of GIS databases to create a complete 'end-user relevant' package. This research application has highlighted a situation whereby 'the tool has got ahead of the knowledge' (Berry, 1987) and Chapter 7 suggested that a lack of spatial analysis functionality in the general GIS toolbox was a key factor.

GIS therefore requires considerable developments in the area of 'spatial analysis' if the latter problem is to be alleviated, and especially if it is to meet the future expectations of the epidemiologist and the elaborate definition advocated by Chorley (1987),

'GIS...is as significant to spatial analysis as the inventions of the microscope and telescope were to science, the computer to economics, and the printing press to information dissemination.' (pp 8, DoE)

It is noted that certain software packages, namely SPatial ANalysis Systems (SPANS, Intera TYDAC Technologies) do provide a wider range of spatial analysis techniques than the ARC/INFO system employed in this research. Whilst these make in-roads into the problem they are by no means the answer, and in fact even the spatial analysis modules in SPANS cannot satisfy the needs of the spatial epidemiological problems outlined here.

Chapters 8 and 9 will discuss these issues under the broad heading of 'complimentary spatial analysis'. They involve the use and adaption of existing techniques in order to overcome two types of limitations, that of identifying statistically significant distributions of ALL (this chapter), as well as offering new approaches for the search into more complex relationships between the incidences of ALL and the environment (Chapter 9). The next section though will illustrate the extent of the knowledge and techniques that are available to carry out sophisticated spatial analysis, both inside and outside the GIS framework.

## 8.1 Complementary GIS analysis

At present extensive literature on spatial analysis techniques exists both in geography and statistics, for example Upton and Fingleton (1985), Wilson (1974), and Unwin (1981). However, many of the methods and styles of analysis that are covered date back to the pre-GIS era, which was typically characterised by little data, slow computers, and an abundance of theories and concepts. Up until recently no globally useful text had been designed to specifically inform the GIS end-user about the complete range of spatial analysis methods which could be exploited.

The first real comprehensive text on GIS was Burrough in 1986. Since then a number of publications have formed a niche in the GIS field such as the International Journal of Geographical Information Systems. These are to be directed at the academic market and thus very important information that they present does not tend to reach the general application user. Although in terms of developing GIS applications, academics are probably in the best position to tackle problems of spatial analysis.

However, a wider audience of end-users including the police, health, and local government must also be exposed to these less desirable repercussion of GIS, if only to rationalise the idealistic pictures portrayed by the software manufacturers. This statement is not to say that non-academic GIS users are ignorant to such technical issues, but realistically it can be assumed that the enormity and long term effect of the limitations of the GIS toolbox are not understood. Alternatively, other end-users may simply be waiting for the academics to solve the problems.

As previously mentioned a realistic view of the situation would suggest that the problem of spatial analysis is in fact an artifact of our own desire to advance technology. Increasing computerisation has simply provided us with the capabilities to generate vast numbers of point datasets for a multitude of application purposes. So where do you start in order to compensate and overcome these new limitations? Firstly, a clear view of the type of spatial analytical functions which would be beneficial to a generic GIS end-user and their application must be outlined.

### 8.2 A Spatial Analysis Toolkit

This will be discussed briefly as a matter of context however a multitude of papers are now dedicated to this subject (Scholten and Openshaw 1990). The problem with much of the literature is that they simply concentrate on highlighting problems and suggesting possible solutions. There is another branch of researchers who spend their time criticising those making initial steps to tackle the problem, rather than offering constructive and improved techniques to help develop their ideas. After this theoretical phase has been exhausted maybe more will be done to devise actual 'workable' solutions. So whilst it is fine to adopt the attitude that 'A problem shared is a problem halved', or so the proverb goes, a little less remonstrating and little more concentration on the development of concrete application examples would be far more lucrative and could lead to a 'problem solved!'

Funding is now readily available from various research bodies including the ESRC, aimed at establishing programmes for graduates and post-graduates to tackle and develop expertise within all areas of the GIS field. This includes the development of spatial analysis tools. In addition a number of researchers have been developing new methods for handling spatial data problems and associated analysis, and these in turn can be used as a basis for creating an equivalent GIS spatial analysis platform, as

discussed in section 8.3 Table 8.1 summarises a theoretical list of the basic requirements for a spatial analysis toolkit for GIS.

**Table 8.1** Suggested spatial analysis tools

i) Pattern spotters and testers

ii) Relationship seekers and provers

iii) Data simplifiers

iv) Edge detectors

v) Auto spatial response modellers

vi) Fuzzy pattern analysis

vii) Visualisation enhancers and

viii) Spatial video analysis

Source: Openshaw et al (1990)

Out of these tools (i) and (ii) will be discussed in further detail in this chapter and that of Chapter 9 respectively. As these are particularly relevant to the needs of the spatial epidemiologist and thus may help to tackle some of the questions that remained to be answered after the completion of Stage IV of the GIS analysis procedure including incidences of ALL localised and does the environment have an impact up on this?

## 8.3 Pattern Spotting

The aim of this approach is termed 'cluster' analysis, whereby areas with a possible localised increase in the number of ALL cases are highlighted. This type of analysis has received considerable attention recently from researchers in both geography and statistics, triggered off by the renewed interest in the possible clustering of chronic diseases advocated by investigative reports of Black (1984) and COMARE (1989).

Hill and Alexander (1989) assessed many of the statistical methodologies which are now available. For the purpose of this exercise however only the two most prominent techniques will be highlighted; the Geographical Analysis Machine (Openshaw et al, 1987) and Besag and Newell's(1990) Nearest Neighbour technique. This combination will serve to demonstrate both a geographer's and statistician's approach to the same problem.

In both cases the techniques have predominantly focused upon small area analyses and have developed separately from GIS technology. The aim of this chapter though is to emphasise the potential implementation of these techniques into the system itself and their relevance in a HEGIS framework.

### 8.3.1 Geographical Analysis Machine (GAM)

The general hypothesis of the GAM is to determine whether there is an excess of observed points within x km of a specific point location. The concept is fairly simple, and the original prototype GAM (Openshaw et al, 1986) has subsequently been developed to form a family of GAMs which detect clusters based on a variety of criteria. These are based on the main components of; (i) A spatial hypothesis generator and (ii) a procedure for assessing the significance of groups of incidences, which can then be linked to a GIS to retrieve and handle the necessary data and then provide a means for geographical display and map processing

The GAM-K procedure described in this section differs from GAM/1 discussed in Chapter 1, serving to demonstrate how techniques can develop from crude, and often error prone beginnings onto far more sophisticated and acceptable techniques. The methodology is similar in that it involves the creation of a set of overlapping circles ranging from a radii of 2km to 10km, with 1km increments. The circles themselves are centred on a grid which is sufficiently fine to allow an 80 percent overlap and completely covers the Northern region. A major design criteria is that it will only detect large rather than small clusters of incidences, thus reducing the number of clusters which may be easily generated by data irregularities.

It is at this point that the GAM-K method deviates significantly from that of GAM/1, because instead of mapping a series of circles, a kernel estimation process is applied to smooth the data produced (Silvermann, 1986). A kernel is estimated for each of the

1km centroids which have been assigned a significant excess value from the Poisson probabilities which were calculated for each of the overlapping circles generated in the first part of this procedure. It is this information which can be incorporated into the GIS environment for presentation. This would use the mid points of any 1km grid square of Britain in order to determine the x- and y- coordinate for all areas which possess a cumulative estimated value greater than zero.

Table 8.2 summarises the format of the results which are produced in this analysis procedure, whilst Figure 8.1 puts these into their spatial context using GIS technology. The increasing intensity of colour from blue though to green and then red demonstrates a rasterised 2D version of the localised distribution of ALL cases in the Region. Figure 8.2 provides a more detailed view for the Tyne and Wear area using the ARC/INFO facility of TIN to reproduce the results in a 3D representation of the cluster affect. The benefits of this alternative means of viewing go beyond simple visual enhancement, as discussed in 8.5.

**Table 8.2:** A sample of the results reported back from GAM-K, which can be used for mapping in GIS

| 1km Grid Squares | | Cumulative Kernel Estimate |
|---|---|---|
| X | Y | |
| 323 | 481 | 0.000236 |
| 358 | 488 | 0.008763 |
| 362 | 492 | 0.009785 |
| 365 | 519 | 0.076544 |
| 410 | 527 | 1.253779 |
| 420 | 556 | 1.785325 |
| 435 | 556 | 2.245614 |
| 450 | 514 | 1.729135 |
| 417 | 550 | 0.257231 |

It is interesting to note from these maps that the Gateshead cluster picked out in the original GAM/1 still persists, but that there is no strong evidence for a significant distribution of ALL cases around Sellafield, between 1976 to 1986. This is rather ironic considering that it was cases in this area which originally sparked off the environmental debate of the 1960's. The latter is due to the fact that the patterns which

Figure 8.1: A 2D view of the GAM/K results

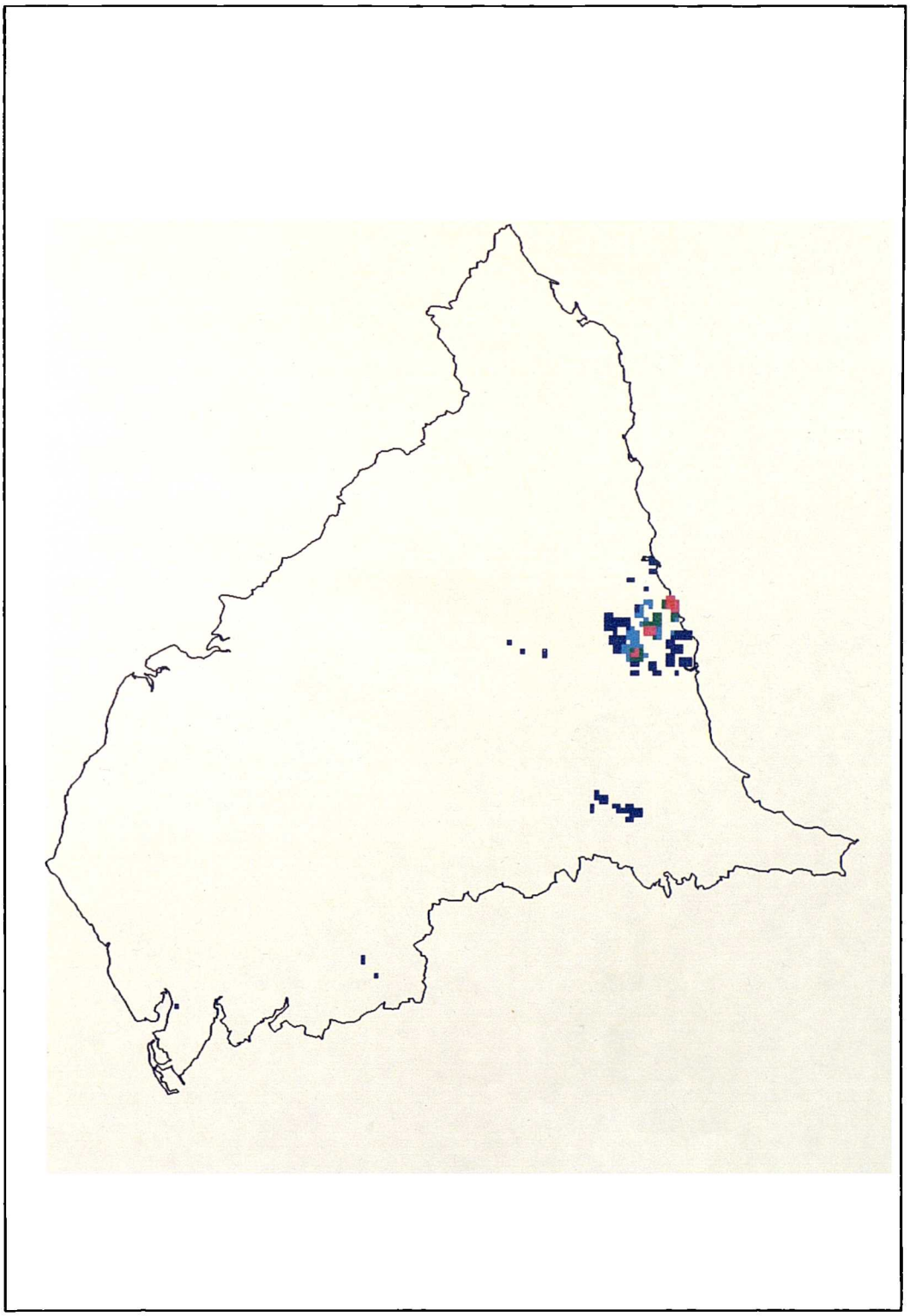
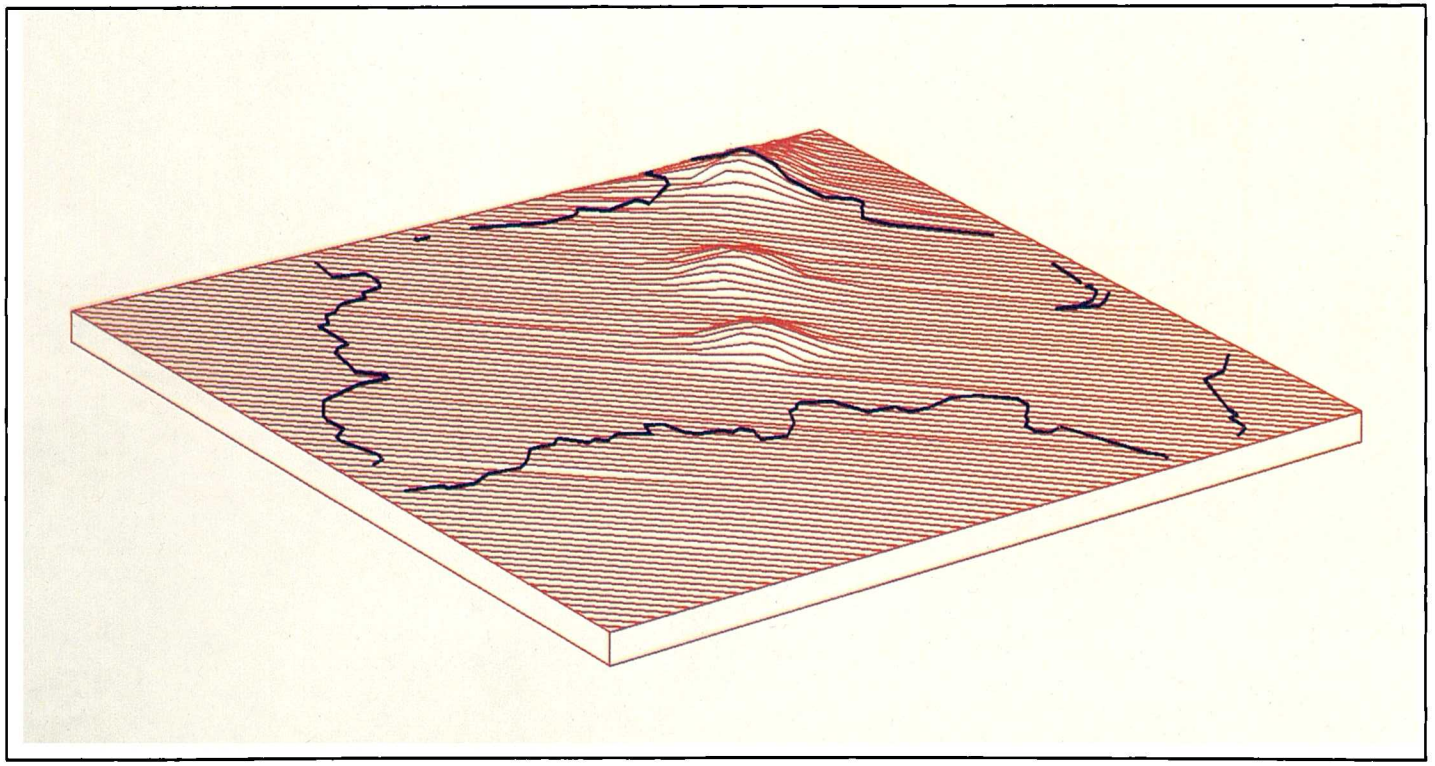


Figure 8.2: A 3D View of the GAM results for Tyne and Wear



were initially established in Sellafield concerned cases which happened over a long period of time. Thus the adoption of a more focused ten year interval means that fewer incidences can be observed on which to base this type of cluster analysis. The Gateshead cluster therefore appears to offer a strong argument for a persistent cause in a relatively small area. In turn the use of this pattern spotting technique can be used to effectively reformulate the epidemiologist's search area and their use of GIS by isolating key regions for specific observation in terms of environmental characteristics.

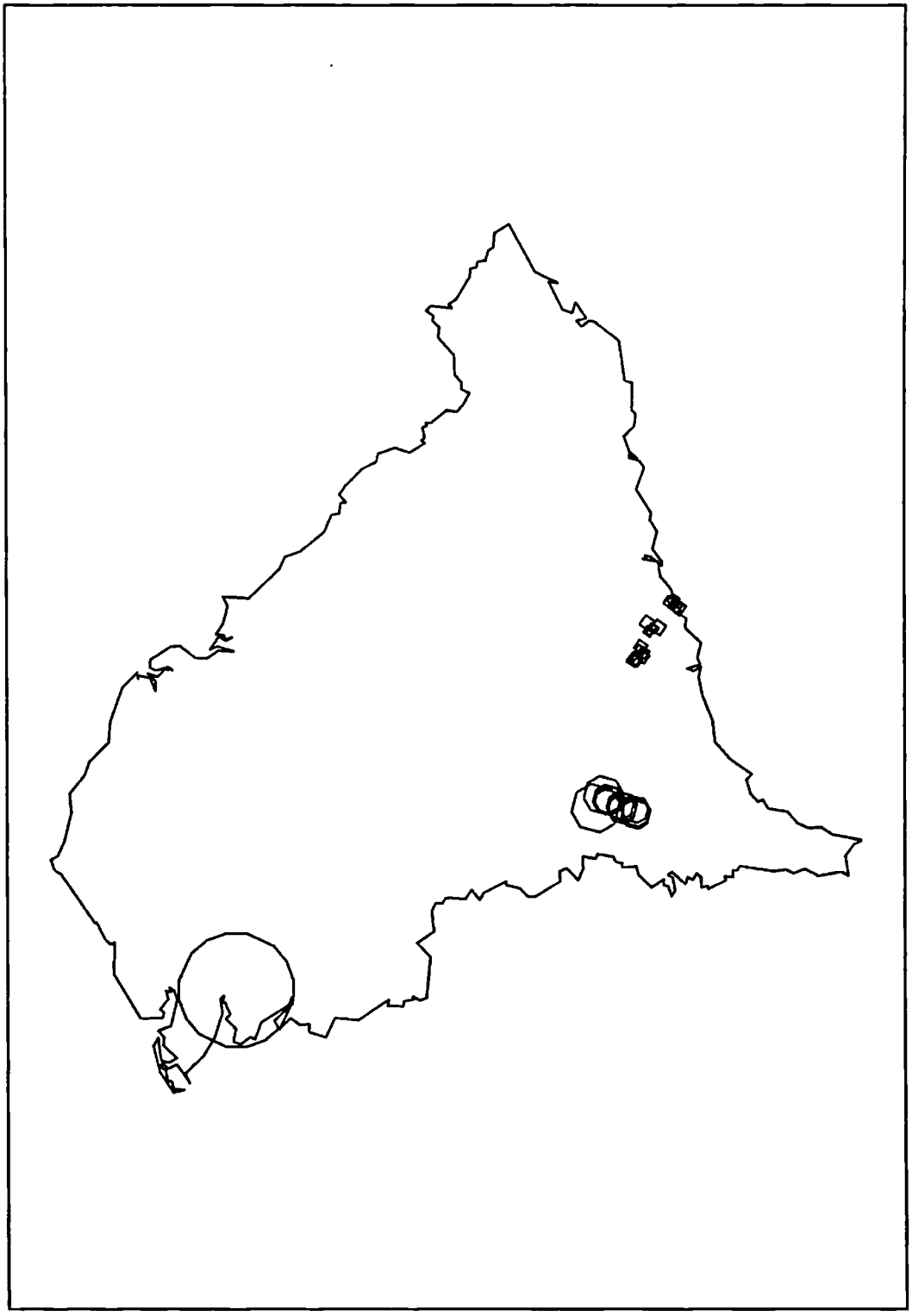
The following method provides a statisticians point of view on cluster analysis but is considered to be an improvement on the basic concept of GAM.

### 8.3.2 Besag and Newell's Nearest Neighbour Method

From the onset the authors state that their method is ad hoc but that it is considered to be more statistically acceptable and less computationally demanding than GAM. Although the methodology that is employed is in fact included within the wider GAM family.

The Nearest Neighbour Method requires two pieces of information, in this case the spatial distribution of ALL and the population at risk which are then merged to form one comprehensive dataset. Each point in this new dataset is then searched individually with all other incidences ordered by distance from a chosen central case. In the first instance a desired cancer count threshold is determined, this is flexible, but for the purpose of this exercise it was set at 5 cases. Thus, when looking for a cluster of five incidences of ALL the distance of the fifth nearest neighbour from the specified central case is determined. This distance is then registered as the radius for the circle in which the number of cancer cases and the associated population at risk is accumulated. The next step is then to calculate the Poisson probability of obtaining 5 cases for a given population at risk in the resultant circle. As with the analysis in Chapter 7 the Null Hypothesis of randomness will only be rejected if the circle probability is found to be less than the five percent threshold. If a significant value is deduced then the central point and radius of the circle will be saved and mapped. The latter procedure is repeated for each and every incidence of ALL within the database. The results of performing Besag and Newell's method for pattern spotting can be seen in Figure 8.3.

Figure 8.3: Cluster Analysis, using the Besag and Newell Nearest Neighbour Method



This methodology differs from GAM-K in that the cancer cases themselves determine the search radius rather than a predefined spatially systematic approach, and this is where the main saving on computer time is made. The advantages of this method over GAM-K though are not confined to computational issues. The zones it produces also take into account variations in population densities in a far more reliable fashion. Thus it is not prone to the problems of traditional methods where constant populations are adopted leading to the amalgamation of areas. Consequently, small rural clusters also stand a reasonable chance of being detected because the radii of the circles are continually adaptive. On the other hand it may be argued that the nearest neighbour technique has a draw back in that the cancer threshold limits the type and size of cluster that can be identified and, unlike GAM-K, no attempt is made to take into account for errors in the spatial data retrieval process.

The significant areas highlighted by this technique are very similar to those produced by GAM-K, despite their different approaches. This may be an indication that this form of pattern spotting technique is sufficiently mature for adoption into a wider GIS framework.

Both examples however ignored the use of multiple testing procedures based on the rationale that the actual benefits which would be accrued from this approach would be small in comparison to the increased magnitude in computing times and memory which would be required. In addition it is considered that when dealing with rare diseases, such as ALL, it would only serve to exclude clusters which may in fact be epidemiologically important. However there is a possibility that these clusters are simply the product of underlying population changes. For instance any spatial analysis technique which is dependent upon population data will suffer from the drawbacks of data availability, ie. that such data are only a snapshot of the population on one day in 1981. Therefore clusters may appear due to the effect of rapid population change over a short period of time, as is the case with the development of New towns or conversely when an area experiences rural decline. Both situations would not be represented in the population base used for these calculations and in turn this would lead to clusters being flagged when in fact they are not unusual but a reflection of changing circumstances in the area.

Alternatively some clusters may go undetected because the circle approach masks their true distribution. For instance, they may be located along a sector area which is effected by some distinct and hazardous atmospheric pollution and whose pathway is determined by both topological and meteorological factors unique to the region concerned. Or as Chapter 7 suggested they may be located along linear features such as the railway network.

Thus point pattern spotting techniques should be interpreted upon their merits and pitfalls. They are not the ultimate answer in spatial cluster analysis but they are a considerable step in the right direction. In the case of both techniques the authors admit that the methods are only intended as a screening device, which can be as lax or as strict as the number of cases specified for the cluster threshold. As a result these techniques still constitute descriptive spatial analysis but are far more useful than the basic toolbox that GIS can presently offer in this field.

## 8.4 Links to GIS

Both techniques were easily employed in a complimentary role to this HEGIS application. This is possible by ARC/INFO's ability to read and produce ASCII files, which in turn are compatible with the FORTRAN programs written to carry out the analysis. This serves to illustrate that there are means of overcoming GIS spatial analysis inadequacies even if it involves a temporary export of the data out of databases. The steps involved in this link are provided in Figure 8.4.

**Figure 8.4**: Linking Pattern Spotting techniques to GIS

**STEP 1:** Select a point datafile in GIS on which you wish to run cluster analysis, for instance the incidences of ALL

**STEP 2:** If the original datafiles with the raw grid references are not available then UNGENERATE the existing point coverage

**STEP 3:** LIST all the IDs and easting and northings for each point

**STEP 4:** Access the GAM or Besag and Newell FORTRAN programs or EXPORT the datafile to another system for analysis.

**STEP 5:** RUN the program using the GAM-K or the Besag and Newell method. For GAM-K the results produced should retain the mid point of the 1km grid and the associated cumulative estimate derived from the kernel estimate. For the Besag and Newell approach the central case should be retained including the grid references and the radii of the significant circles to be mapped.

**STEP 6:** Input results into GIS in the form of new grid referenced data

**STEP 7:** Create the necessary coverages. For GAM-K this involves the pinpointing of points onto 1km grids covering the study region and creating a rasterised choropleth map using the kernel estimates as a z value to generate 2D or 3D surfaces, using the TIN module in ARC/INFO. For the Besag and Newell example, the points and their radii can be automatically GENERATED as a series of circles.

**STEP 8:** Enter ARCPLOT to view the data and overlay the raw cancer point data or other environmental datasets in order to further enhance the visual impact, Figures 8.1 through to 8.3.

### 8.4.1 Integrating pattern spotting techniques into GIS

Some end-users may want this type of analysis incorporated directly as a black box in GIS, where data goes in and results come out. Whilst the Macro Languages of GIS serve to do this in terms of standard customisation of GIS applications, there are no similar bolt on techniques for accessing point pattern techniques such as GAM-K and Besag and Newell, at the moment at least. ESRI, the developers of ARC/INFO, are actively seeking researchers to develop relevant spatial analytical techniques such as these for adoption into GIS to form compatible complimentary modules.

In a bid to satisfy end-user requirements as well as a curiosity to see how amenable ARC/INFO can be in replicating other programming languages, the Besag and Newell Nearest Neighbour technique was taken and manipulated. This was chosen over

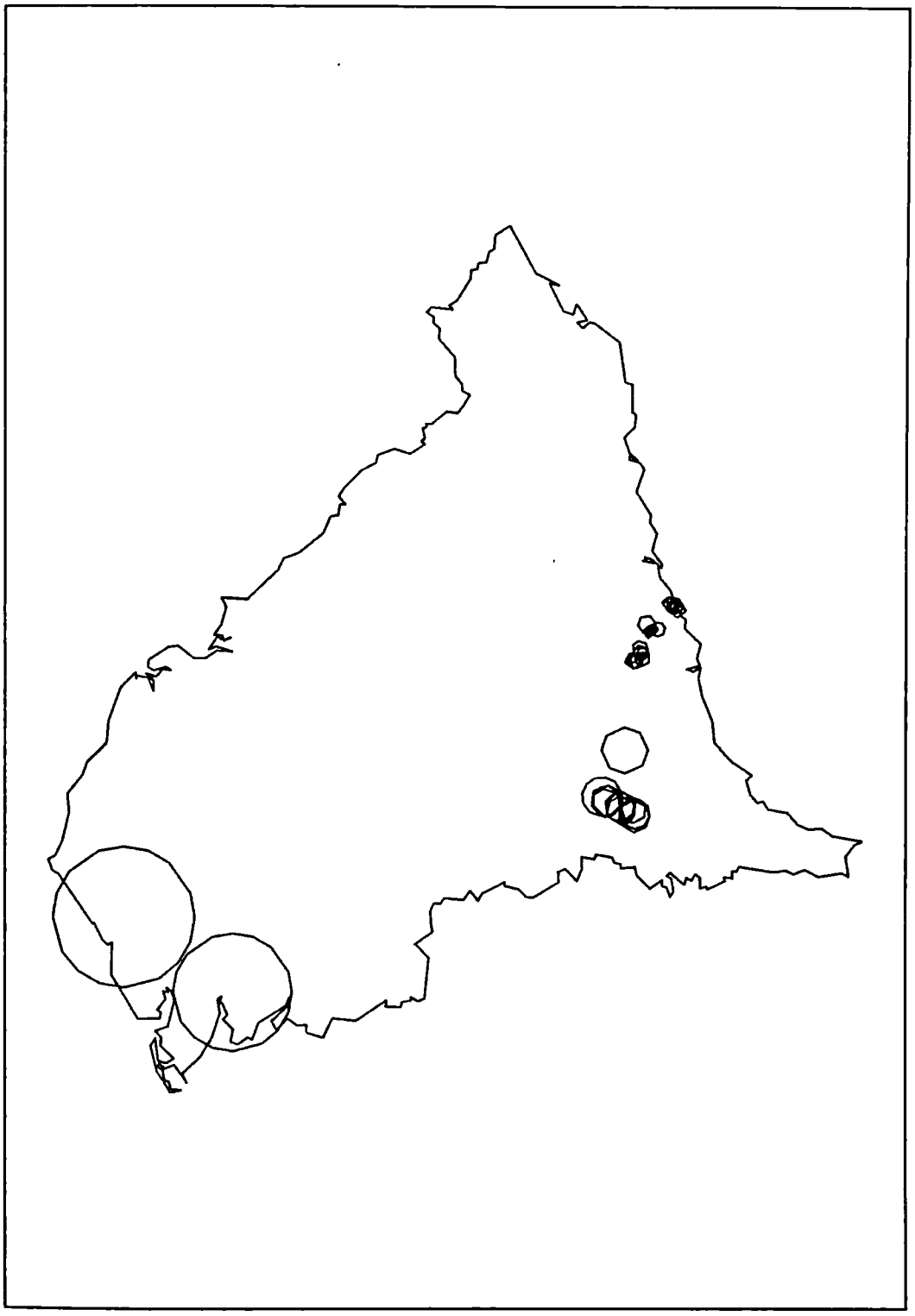
GAM-K because the method lends itself more readily to an ARC/INFO data model by virtue of its simple and rather neat philosophy. In addition, many of the steps involved in the process show considerable similarities to the functions of certain ARC/INFO commands. Thus by using the ARC Macro Language (AML), to combine ARC and INFO commands plus a little embedded FORTRAN, this technique could be replicated, see Appendix F for examples of AML functionality. The AML and associated programs are also given in Appendix G, with an explanation of how certain factors such as the critical population and the mean were deduced. This acts as a reference for others who may wish to adopt, or even refine, these initial attempts at integrating cluster analysis techniques into a GIS framework.

It is noted that ideally the calculations for the critical population and the mean could also have been written into an AML but this would only have served to slow the process down further. The main reason for writing an AML was to emphasise the similarities between the Besag and Newell procedures and the functions which are available in GIS. These included the use of the command POINTDISTANCE to establish the fifth nearest neighbour and the radius for the resultant circle. The BUFFER command could then be used to create the circle and CLIP could isolate the population/cancer cases for the area concerned. The STATISTICS command summarises the data for future calculations be it for the Poisson probability, or in this case the population which would be compared with the critical threshold for a significant number of cancers to result, ie if the population for the circle is less than the critical population the circle is considered significant and saved for mapping.

As the map in Figure 8.5 demonstrates the outcome of the GIS version of Besag and Newell is almost identical to that achieved from the original program. The absence of one or two circles may be due to the loss or inclusion of a single ED which may lead to the final probability statistic differing by 0.049 compared to 0.05, one of which would be considered significant the other not. However the main cluster in Gateshead remains.

The main difference with the GIS version of pattern spotting is the computational requirements both in terms of time and memory. The purely mathematical approach of Besag and Newell takes a matter of minutes of CPU time to complete whilst the ARC/INFO version takes considerably longer because of all the calculations which are involved to establish buffer zones, isolated areas and summary statistics for each

Figure 8.5: Cluster Analysis, ARC/INFO style!



of the 225 cases. Thus the result of this exercise is to show that spatial analysis is not completely beyond the realms of GIS, but for speed and perhaps finesse of programming, such tools may still be far more beneficial if used in a complimentary role.

## 8.5 GIS and Spatial Analysis Tools in harmony

These techniques were developed independently of GIS and the results they produce could easily have been mapped on any graphics package just as successfully. However as this chapter demonstrates these methods have an important role to play in terms of enhancing the information that can be stored in a point dataset. The question raised at this point is; Is there any way in which GIS can now be used to improve upon these cluster analysis results?

The answer is yes by increasing their descriptive power. The epidemiologist now has the ability to view data as either raw points or as clustered entities which provides a more focused approach to their search for environmental causes. This can be achieved in the GAM-K example by taking the 3D views of clustered cases and draping certain environmental coverages over this new base to observe possible relationships. Figures 8.6(a) and (b) demonstrate the effect of combining the cluster analysis results for Tyne and Wear with overlays for incinerator buffers and geology respectively. Alternatively the Besag and Newell circles could be mapped and using ARCPLOT other underlying coverages could be queried in and around the areas depicted by the significant circles.

The epidemiologist is now in a position where interesting patterns have been detected and some degree of significance has been assigned to these clusters. They can begin to ask questions as to why these distributions of ALL cases are interesting, especially in these localised areas? Chapter 9 therefore takes the next logical step which is to search for the possible combinational influence of different environmental factors upon the overall micro-environment of a child. This was attempted in Chapter 7 with limited success due to some of the practicalities of database manipulation. Thus the next chapter hopes to overcome the problems experienced in Stage IV of 'The GIS Process, by providing a far more useful and computationally workable method to multiple overlays and subsequent analysis of ALL.

Figure 8.6a: Overlaying GAM-K results with environmental coverages: incinerators

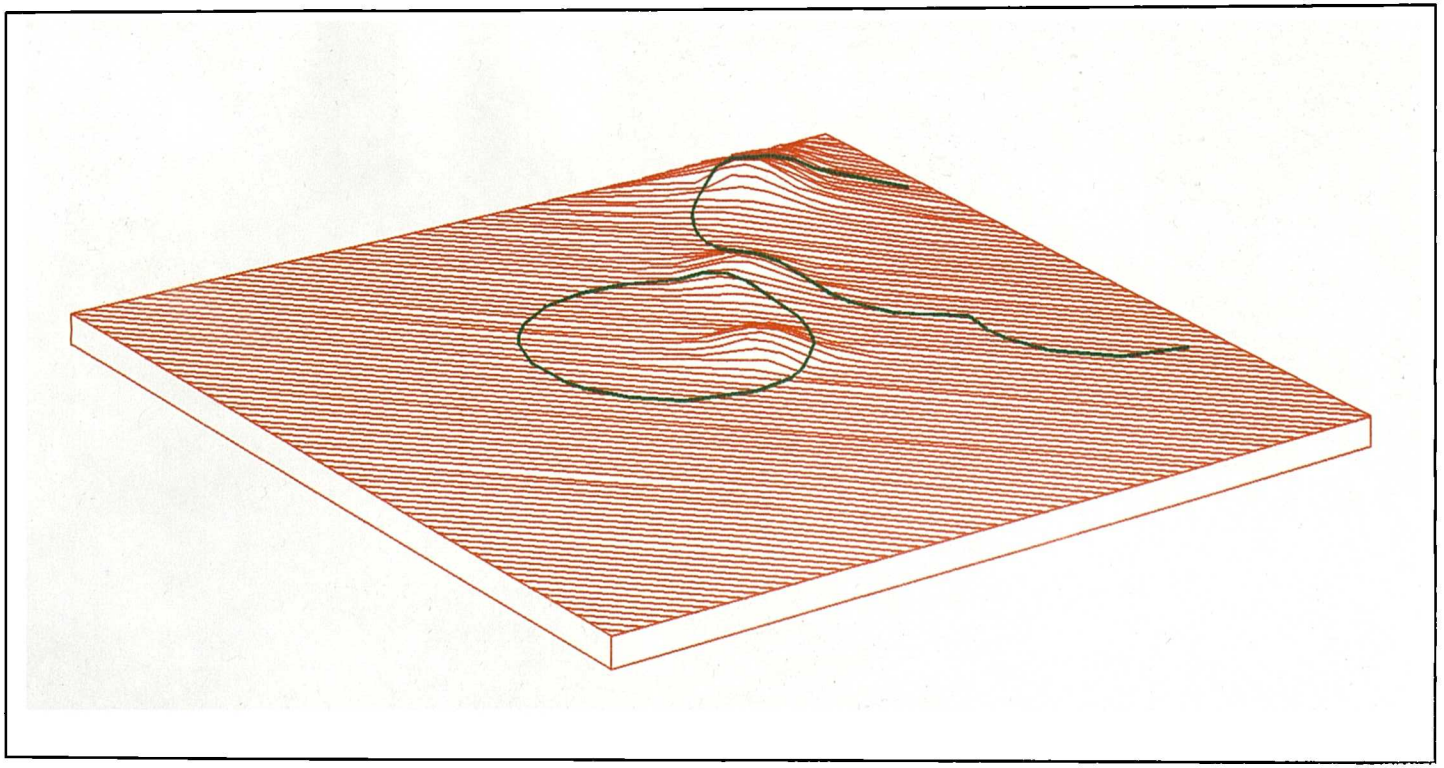
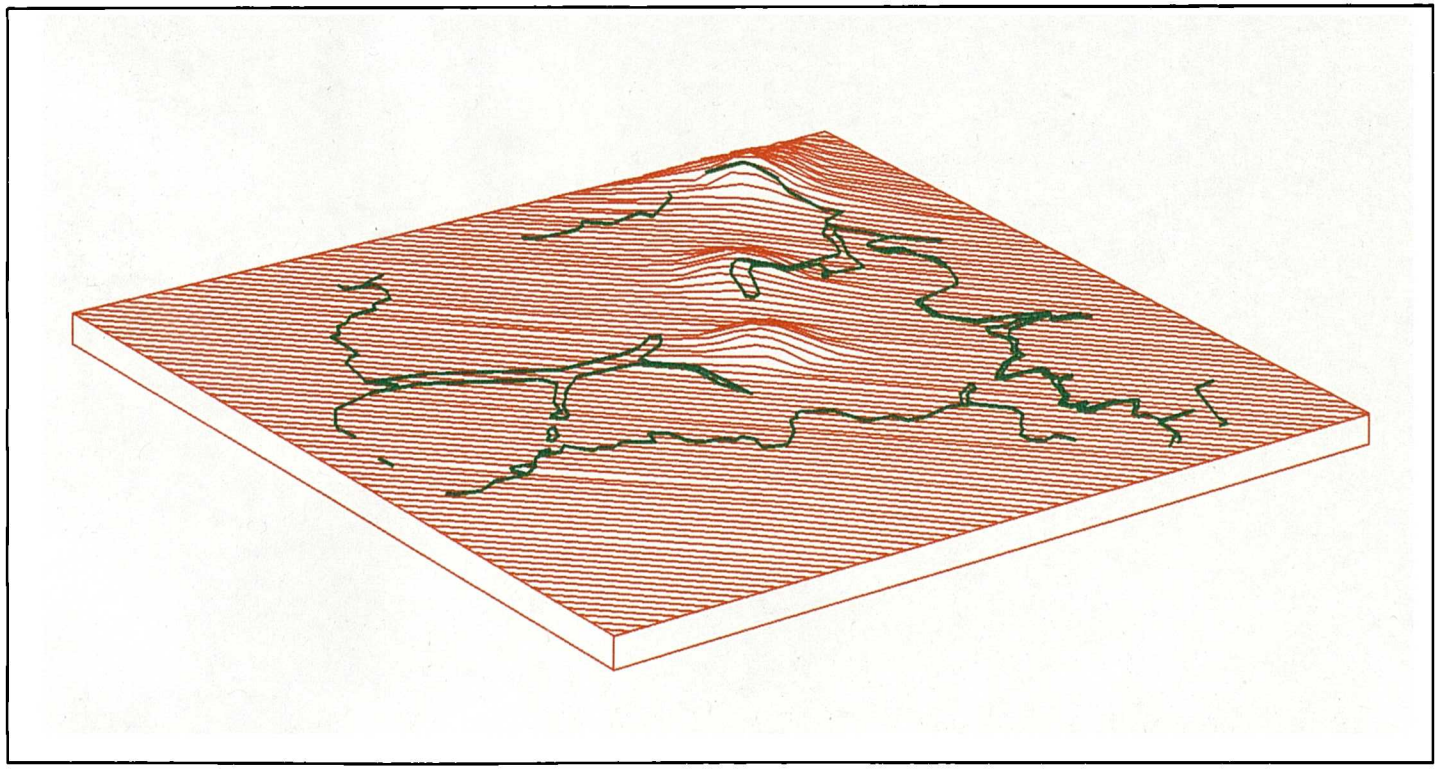


Figure 8.6b: Overlaying GAM-K results with environmental coverages: geology



# CHAPTER 9

# COMPLEMENTARY SPATIAL ANALYSIS II
# RELATIONSHIP SEEKERS

This chapter tackles a second spatial analysis technique which is considered to be highly relevant to HEGIS and in particular the needs of spatial epidemiology. It was referred to in Chapter 8 ( ii, in Table 8.1) as a 'relationship seeker'. This idea focuses on the automatic exploration of databases in terms of searching for evidences of spatial relationships in geographical data, as opposed to spatial patterns in point data which was the objective of the last chapter.

## 9.1 Using GIS to discover Spatial Relationships

Chapter 7 offered a GIS solution to searching for a relationship between ALL and environmental factors, however a number of issues were highlighted as problematic in terms of a GIS approach. The first involved the choice of predictor coverages whereby the very nature of the environmental causation of ALL debate is characterised by the fact that there are no strong a priori theories to direct subsequent analysis. Secondly the ability to overlay two or more maps may be a fairly simple procedure in terms of GIS commands, but the coverages which are produced tend to be very complex both in terms of the resultant database structure and for necessary interpretation.

Table 9.1 summarises the number of problems which can further complicate the task of selecting possible environmental predictor coverages. These are in addition to problems which immediately confront the epidemiologist such as; a fairly large number of observations (in this case 225), several possible predictors (some 16 environmental coverages at the Northern Region), the presence of spatial autocorrelation, and the end-users own ignorance over the processes being modelled.

**Table 9.1**: List of potential problems which typify this research

i) Relationships need to be established based on data availability as opposed to existing prior hypotheses

ii) It must be understood that many geographical relationships may be spatially localised rather than ubiquitous.

iii) Missing or unavailable data for certain aspects of the environment may need to be acquired in the form of surrogates, such as the road network and incinerators in order to infer zones of pollution.

iv) It has to be assumed that, until there is evidence to the contrary, most spatial relationships will tend to be complex If such relationships were simple the mapping procedures in Chapter 7 would probably have sufficed to identify the processes responsible for the causes of ALL.

Besides the complexity of maps that can be produced by the GIS process the technology also fails in terms of determining the possible interaction that may occur between one or more overlays. The resultant coverages provide no guide to which coverage may be the most important in a combination of coverages, or conversely which has such a minor effect up on the final outcome that it could have been ignored. This inability to suggest interaction between factors may carry a very high risk of significant effects being overlooked. An alternative therefore to directly using GIS to search for relationships between databases would be to link into existing statistical packages which can enhance the quantitative side of this spatial problem. The following sections offer two alternatives to GIS in the search for relationships between ALL and environmental factors.

## 9.2 Linking to statistical packages

Statistical packages such as SAS, SPSSX and GLIM offer categorical regression technique which can be used to calibrate models that link the probability of a child having ALL to various environmental factors. These produce a table of values that highlight possible interactions which may be occurring between the predictor

coverages chosen and which are considered to be statistically acceptable in terms of 'seeking relationships'.

GIS can manipulate the relevant datasets in order to produce the necessary summary of environmental characteristics to be used by these statistical packages. Each row in this table will contain a record for individual EDs with six age/sex groups for the population, six age/sex counts for the cancers, and all the codes which represent the surrounding environment for that particular ED. Table 9.2 summarises the coverages used in this analysis, the variable codes which were attached and the sub-categories that the environmental coverages were split up into in order to be compatible with the model.

In simple terms, the models which will be fit to the data involve;

P (incidences of ALL) = f (within the influence of an incinerator buffer, on a particular geology type, landuse etc.)

This equation will vary according to the type of relationship that the epidemiologist wanted to investigate. The left hand side of the equation refers to the 'dependent' variable, ie. the incidences of ALL. The right hand side consists of all the categories which form the possible set of predictors. The problem set in this research cannot use conventional ordinary least squares regression because, (i) the dependent variable is either categorical or a Poisson probability, in other words the number of cancers in a given ED, and (ii) the independent variables are also categorised ie, rock types, landuse codes, inside or outside a buffer and so on. The analysis in this section therefore will use a 'log-linear' modelling approach for searching databases for relationships, via the statistical package GLIM (Generalised Linear Interactive Modelling, Healy, 1988).

### 9.2.1 Interfacing ARC/INFO with GLIM

Other researchers in this field have already attempted to integrate statistical packages directly into GIS software, in particular ARC/INFO with GLIM (Kehris, 1990). This was in response to the primitive statistical facilities offered by ARC/INFO, such as STATISTICS. In this example the interfacing was achieved with the use of FORTRAN subroutines which was made easier by the fact that ARC/INFO is written

**Table 9.2:** A summary of the coverages and associated categories to be used in GLIM and GCEM

| VARIABLE | COVERAGE | CATEGORY | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| V1 | Coal Power station | 1km | 2km | 5km | NN | | | | | | | |
| V2 | Estuaries | 400m | 600m | 800m | NN | | | | | | | |
| V3 | Geology | Recent | ------- | J u r a s s i c | | | | | -------- | Triassic | | Permian |
| V4 | Incinerator sites | 1km | 2km | 5km | NN | | | | | | | |
| V5 | Main roads | 150m | 250m | 350m | NN | | | | | | | |
| V6 | Mines | 1km | 2km | 5km | NN | | | | | | | |
| V7 | Motorways | 150m | 250m | 350m | NN | | | | | | | |
| V8 | Nuclear Powers | 1km | 2km | 5km | NN | | | | | | | |
| V9 | Other roads | 150m | 250m | 350m | NN | | | | | | | |
| V10 | Radiation site | 1km | 2km | 5km | NN | | | | | | | |
| V11 | Railway | 150m | 250m | 350m | NN | | | | | | | |
| V12 | Waste disposal sites | 1km | 2km | 5km | NN | | | | | | | |
| V13 | Background radiation | 2.5- | 3.0- | 3.5- | 4.0- | 4.5- | 5.0- | 5.5 | | | | |
| V14 | Landuse | Urban | Agricultural | | Upland | | Wood | | Other | | | |
| V15 | Rainfall | 1000- | 2000- | 3000- | 4000- | 5000- | 9000- | 13000- | 17000 + | | | |
| V16 | Primary roads | 50m | 250m | 350m | NN | | | | | | | |

**Note:** NN denotes that the ALL case was found outside the area of impact for that particular source of pollution

in FORTRAN. However, the resultant interface was not compatible between systems and the analysis procedures took longer to execute, which is similar to the experiences that this researcher when attempting to build in cluster analysis techniques (Chapter 8). A review of GLIM in the light of this spatial epidemiological problem also identified additional short comings in this approach to enhancing spatial analysis in GIS.

### 9.2.2 Performing Log-linear modelling

Data are input to the system and accessed by GLIM via a program which reads in the data and fits the chosen models to be tested, see Appendix H. The first procedure to be executed involves the command $FIT$ which determines the 'grand mean' for the model. This assumes that the development of childhood cancer is independent from everything else and establishes an average for all the expected cells in a given model, calculated as proportions of the total with a binomial error term and a logit function (O'Brien, 1989).

The problems with this type of approach were already becoming obvious given one 'dependent' variable and 16 possible models which could be fitted, outlined in Table 9.2. Also the analysis procedure was immediately restricted by the computational inadequacies of the GLIM package which stores all the information and calculations in memory before the results are reported back. The outcome of this is that with only a limited amount of memory the larger the dataset used the quicker the system will fail when it runs out of memory. In order to even begin this exercise therefore the number of cases which could be analysed had to be reduced to a manageable size. This involved taking a twenty percent sample from the original file which contained 7065 EDs, producing 1418 cases for analysis. The implications of this will be discussed later.

The reduced dataset was then fitted to the grand mean and the environmental effect of each coverage upon the incidences of ALL was measured. Table 9.3 summarises some of the models which were tested. The first row represents the grand mean, with a scaled deviance of 285.9 and 1417 degrees of freedom ( ie. the number of cases minus one), all subsequent analysis values will be compared to this overall figure. For example, the second row in the table demonstrates the log-linear model which is produced when the presence of a coal power station is taken into account. To deduce

whether this is in fact influential in terms of ALL causation the reduction in the scaled deviance from the grand mean is calculated. This is determined by calculating the difference in degrees of freedom, in this case 3, and establishing the reduction of scaled deviance which should have occurred under the chi squared distribution, at a significance level of 0.05. In this example a 3 point drop in degrees of freedom should have lead to at least a 7.815 drop in the scaled deviance, however the actual difference is only 2.9 which suggests that the presence of a coal power station as a singular cause of ALL can be ignored. Comparing columns four and five in Table 9.3 indicates that in fact log-linear modelling does not highlight any relationships between ALL and each of the environmental factors, not even the railway network which was picked out in Chapter 7 by the Poisson Probability test.

**Table 9.3:** The Log-linear modelling results from GLIM

| Model $FIT$ | Scaled deviance | Degrees of Freedom | Under Chi-square Expected | Observed |
|---|---|---|---|---|
| | **285.9** | **1417** | | |
| V1 | 283.0 | 1414 | 7.815 | 2.9 |
| V2 | 283.5 | 1414 | 7.815 | 2.4 |
| V3 | 274.4 | 1408 | 16.919 | 11.5 |
| V4 | 281.3 | 1414 | 7.815 | 4.6 |
| V5 | 282.5 | 1414 | 7.815 | 3.4 |
| V6 | 284.2 | 1414 | 7.815 | 1.7 |
| V7 | 285.3 | 1414 | 7.815 | 0.6 |
| V8 | 285.5 | 1414 | 7.815 | 0.4 |
| V9 | 284.7 | 1414 | 7.815 | 1.2 |
| V10 | 284.0 | 1414 | 7.815 | 1.9 |
| V11 | 283.8 | 1414 | 7.815 | 2.1 |
| V12 | 284.8 | 1414 | 7.815 | 1.1 |
| V13 | 282.4 | 1412 | 11.070 | 3.5 |
| V14 | 277.4 | 1409 | 15.507 | 8.5 |
| V15 | 282.6 | 1409 | 15.507 | 3.3 |
| V16 | 283.5 | 1414 | 7.815 | 2.4 |
| V10*V3 | 262.7 | 1388 | 38.885 | 23.2 |
| **V4*V10** | **260.9** | **1403** | **23.685** | **25.0** |
| V4+V10 | 280.2 | 1411 | 12.592 | 5.7 |

However the main purpose for employing this log-linear approach was to explore the possible interaction effects between different environmental factors. In terms of syntax this is a relatively simple affair in GLIM, involving a * to establish the interaction between two or more environmental variables or + to look at them

together but not necessarily dependent upon each other. Even with this reduced sample though GLIM could only cope with calculations for a three way interaction, but as mentioned in Chapter 7 this is probably the maximum that can be assimilated by an end-user at any one time. In this initial analysis one combination of coverages did appear to be significant, that of the presence of an incinerator and a site with a licence to deal with radioactive material, record 19 in Table 9.3 representing the model V4*V10.

This finding though must be viewed in the light that the model is only being fitted to a sample of the actual data. Thus in order to test the robustness of this interaction the model was fitted to additional random samples but in each case the interaction did not appear again. The first result may be interesting but more than likely it was produced by bias in the data sample. Given the nature of rare diseases such as ALL and the impact of reducing numbers can have on any patterns that may have existed then very little credence should be given to the results of the models described in this section. So why do the analysis? Basically statistical packages like GLIM are the only means of carrying out this type of log-linear modelling. Thus this section has served to evaluate the feasibility of actually interfacing GIS directly into these available packages.

### 9.2.3 Is it worth it?

The findings from performing log-linear modelling in respect of a complex spatial epidemiological problem did not afford any strong evidence for an environmental cause of ALL. This may be good or bad depending upon the end-users aims and perceptions of the data to be analysed, however there are a number of shortcomings with using GLIM in this application. The first of these is the restraint upon computational power. The methodology of log-linear modelling works in theory but was originally designed to be used in controlled experiments where the hypotheses and predictors are well defined, consequently fewer permutations are involved. In order to accommodate for this limitation of GLIM the number of cases were reduced, and this application probably suffered more than most from this course of action. As the latter may have lead to any patterns in ALL to be removed and thus GLIM would fail to pick out any possible associations anyway.

Another problem with GLIM, particularly in relation to this application, is that the models depend upon the end-user knowing which variables they want to test. Yet the whole idea of looking for an environmental relationship with ALL is because these predictors are unknown. The result of this restriction would possibly cause the epidemiologist to concentrate on favoured hypotheses, hence failing to take advantage of the various environmental databases that can be stored in a GIS. Also log-linear modelling is designed to search for relationships between ALL and the thematic variables of each coverage, in other words it takes a global look at the possible interactions which may be occurring in the whole of the Northern Region and forfeits any localised effects that may be present.

This approach therefore fails to make the best use of the data and the flexibility that is offered by GIS technology. For example, the latter methodology serves to eliminate any locational information which is a major asset of GIS. Location can also be very useful in flagging key areas of localised incidences by acting as a surrogate for some unknown or unrepresented aspect of the environment. In addition, log-linear modelling fails to take into account the effect of spatial autocorrelation (Healy, 1988). In this spatial problem therefore where the causes are unknown, GIS has the ability to support a data led exploratory technique, but GLIM is a hypothesis tester for which the epidemiologist does not have any specific theories to test. In practice therefore GLIM does not appear to be amenable to this application, or realistically to GIS as a whole. Especially when it is considered that GIS is geared to develop, manipulate and store large datasets which these statistical packages simply cannot cope with. This does raise the question that it may be a waste of time attempting to interface into existing packages if they are not even compatible with the ethos of GIS. Modifications of these ideas though may be the way forward.

## 9.3 Developing an automated version of the GIS overlay process

The objective here is to use the data collected in this research to create new insights into the spatial phenomena of ALL. At the same time the aim was to overcome the technical and computational demands of multiple coverage overlays in GIS, as well as replicating the procedure described in the log-linear modelling section but with a spatial analytical component. The result was the development of a prototype Geographical Correlates Exploration Machine (Openshaw, Cross and Charlton, 1990)

described in detail in section 9.4 whilst section 9.5 presents some of the maps produced by this alternative approach.

## 9.4 Building a Geographical Correlates Exploration Machine (GCEM)

The task of building a map overlay exploration machine is regarded as being analogous to a spatial response modelling problem with a large and non-deterministic set of possible predictors represented by map coverages. The manual equivalent of this search would involve the overlay of different environmental coverage permutations until the analyst was either exhausted or some interesting results were obtained. The objective here is to automate this process and to remove the activity from an explicit cartographic domain to a spatial analytic equivalent. This would allow a full coverage permutation search to be completed and areas which exhibited strong evidence of a relationship would be identified. Also taking the problem out of the GIS framework will allow the search space to be extended to include other covariates which are not explicitly map based, ie socioeconomic characteristics.

This search process is functionally analogous to the regression modelling problem discussed in section 9.2 whereby the task is to identify the best subset of K predictors from M available variables. However, unlike the logistic regression runs, GCEM searches for the best overlay combinations and looks for relationships that are not only global but also restricted to certain small parts of the study region. In some ways it is similar to the GAM-K and Besag and Newell methods, described in Chapter 8, but it uses the polygon overlay process to define the areas of search rather than geometric pattern detectors. A key feature of the prototype GCEM therefore is the searching through of all 2**M-1 overlay coverage permutations, see Openshaw et al (1990).

### 9.4.1 Basic algorithm

This coverage permutation generator could perform a few thousand sets of up to K overlay operations In ARC/INFO terms (or any other GIS) this would lead to extremely prohibitive computer times. For instance, if the analysis was to be carried out for all the coverages available at the Northern Region scale this would involve 16

different factors. These in turn have additional information that must be considered, including geological types, background radiation levels, buffers for areas of impact and so on. Hence an exhaustive overlay process of these in GIS would involve 2**16-1 permutations, over 30,000 sets of possible coverage overlays. It is not an exaggeration therefore to assume that this would take an unreasonable length of time and computing resources. Fortunately, GCEM offers a simple solution to this problem, and one which works well when the dependent variable is based on point coverages. This involves the replication of the overlay process without actually having to perform any polygon overlays. This approach is outlined in the following;

**Step 1:** Define a set of M polygon coverages of interest and the main point dataset to be investigated, ie. select all the environmental coverages of interest for the region and the point incidences of ALL. In order to simplify subsequent calculations a single combined health dataset was established, containing the ED centroid and all the relevant age/sex characteristics of the population at risk and the incidences of ALL, as described in the Besag and Newell method.

**Step 2:** Using ARC/INFO a point-in-polygon operation was performed on all the coverages selected. This was achieved via the ARC/INFO command IDENTITY, which attaches the respective environmental data to each ED centroid. Then by using INFO all the necessary attributes for further analysis was selected and output for use within GCEM. These include; a unique reference number for each ED, a unique reference number for the polygons within each environmental coverage, and a value which represents the attributes for that coverage, such as a land use code, geological type and buffer zone distances.

**Step 3:** Transfer this new data file to another system, in order to run the analysis, in this case a CRAY XMP/416 supercomputer was used.

**RUN GCEM**

The basic GCEM search process involves;

**[a]** Define an a priori set of M coverages, as outlined in Table 9.2.

[b] Generate a permutation of the M coverages or else go to [g].

[c] Set an acceptable limit for K, ie the maximum number of coverages which can be sensibly overlayed at any one time, beyond which it is considered that the polygons generated become too small for realistic analysis and the spatial processes that they represent become too complicated for the end-user to interpret. In this example K was set at 4, and thus if there are more than K coverages then process returns to Step [b].

[d] Simulate the overlay process for a permutation of up to K coverages and aggregate the point data of interest for each overlay polygon.

This virtual overlay simulator identifies the points with the same record key (ie. they occur in the same polygon, based on its id and have the same attribute codes) for each coverage. From this it can be assumed that these points also lie in the same final polygon of a coverage which would have resulted if the overlay process had been carried out in a GIS environment. At this point the total population at risk and the total number of cancers are known for the polygons and thus a probability value can be calculated.

[e] Some measure of the presence of pattern is used as an indication of the interaction between coverages and therefore the possible relationship which may exist with respect to ALL. The most interesting polygon overlay combinations are saved.

[f] Return to [b] to continue the search of overlay permutations.

[g] GCEM scans the results file to pick out specific record keys based upon the polygon label numbers and the attribute codes attached.

**Step 4:** From this information, and the associated probabilities, the results can be transfered back into a GIS compatible file and used to carry out further mapping and manipulation. The crucial detail required for this convertion of results from GCEM into GIS maps are; the names of the coverages involved in the significant permutation, the characteristics of interest ie did the polygon concerned represent

urban, woodland or agricultural land? was it in or outside a road buffer? and so on. Finally the actual unique polygon id must be returned since this is the all important locational aspect. As well as the significant EDs involved in the resultant permutation because this serves to determine the difference between polygons which are not significant but still satisfy the overlay criteria. Section 9.5 provides the necessary results and gives an illustration of this point.

The use in GCEM of the so-called 'virtual polygon overlay simulator' therefore greatly reduces the computational burden of map searching in GIS. However, it is not entirely accurate because certain complex polygons may appear interesting but are in fact the product of topologically disjoint regions. In other words the very nature of the original coverage's topology can result in polygons having exactly the same internal numbers based upon the virtual overlay process but represent completely different polygons in terms of an actual overlay which would be produced by GIS. This is a relatively rare event and may be avoided to some extent by splitting up the large and complex polygons which cause this problem. Even so these errors are not catastrophic, the principal effect is to reduce the sensitivity of the method rather than to produce spurious results. It could also be argued that in relation to GCEM these are not errors but possibly a useful feature, indeed the ability to violate topology can be used advantageously to generalise the point-in-polygon process to extend virtual mapping of the data to include non-map information. As previously mentioned this could include a residential area classification which can be handled in a similar way to real map coverages. This allows the search process to examine various mixtures of spatial, implicitly spatial and non spatial covariates without having to assume that they all have to exist in a GIS digital map form.

The remaining problems discussed in this section concern that of the choice of pattern statistic and the need to handle the multiple testing problems involved in the search process.

### 9.4.2 Measuring the presence of spatial pattern

The measure of the presence or absence of any spatial patterns in the polygons created by GCEM was derived from the use of the Poisson Probability test, based on the same procedures and hypotheses as previously described in Chapter 7, ie. the distribution of

ALL was assumed to be random unless the statistic proved significant at the 5 per cent level suggesting a possible clustering of events.

### 9.4.3 Handling multiple comparison problems

If a simple Poisson probability test statistic, or any other statistic, is to be used then measures need to be taken to correct the results for multiple testing, perhaps more so in this methodology than those adopted in Chapter 7. For instance, in an overlay permutation with 200 final polygons, the occurrence of a very small Poisson probabilities may merely reflect the fact that there are 200 results to choose from and it could well be a chance occurrence that is not nearly as rare as the small size of the probability might suggest.

This multiple comparison correction procedure is only valid for a single permutation of overlays. Yet the GCEM process involves upto 2**M-1 sets of results. There are two possible strategies to overcome this problem. The first is to simply ignore the multiple comparison problem based on the fact that GCEM is at best considered a descriptive and creative tool rather than a strict test of hypotheses. The second strategy is to use Monte Carlo simulation to handle both sets of multiple comparison problems simultaneously.

Openshaw et al (1990) reported some Monte Carlo results. The problem is though that the multiple testing problem cannot be solved. GCEM creates and tests many millions of hypotheses and even under the Null Hypothesis of randomness many 'significant' results might be expected. Certainly very small Poisson probabilities can occur by chance under these conditions. Therefore Monte Carlo simulation may help but it is likely that this approach would lead to all the results reported back by GCEM being rejected. In some respects the original study by Openshaw et al (1990) was lucky that any of the results survived the Monte Carlo significance testing process. As noted in Chapter 8, when employing the GAM and Besag and Newell methods, GCEM is using inference as a filter and not as a strict test of the hypothesis in the conventional sense of the word. Thus GCEM can suggest possible hypotheses but not, by itself, prove anything. GCEM is therefore, essentially a hypothesis generator. This limited use is probably all that can be expected though given the problems that behold geographical data.

## 9.5 Searching for geographical correlates of Leukaemia: Preliminary results

Table 9.4 summarises the most significant coverage permutations which were derived from exhaustive use of the virtual map overlay process of GCEM. At a quick glance it can be seen that every coverage is listed in at least one of the permutations. However the most interesting aspect is that certain coverages tend to occur a number of times in different combinations. In terms of interaction this seems to suggest that maybe the other coverages are present but the main effect upon the increased incidences of ALL observed is the interaction between these recurrent environmental factors, ie. 16, 10, 11 and 12, which special radiation sites, railways, waste disposal sites and primary roads respectively, thus the variable codes relate to those documented in Table 9.2 in section 9.2.

Figure 9.1 provides a simplified and annotated version of the detailed results which are returned by the GCEM program.

**Figure 9.1:** An annotated example of the significant overlay entries returned by GCEM

```
ENTRY 613   613

      Number of coverages        3
      Coverage permutations    10      11      12

Overlay Coverage.....10  Polygon No(#) = 14 Category = 3
Overlay Coverage       11  Polygon No(#) =  4 Category = 1
Overlay Coverage       12  Polygon No(#) = 36 Category = 3

Population at risk  1326.0     Number of cancers = 5
Poisson probability = 0.00012836

EDs which pinpoint the significant polygons

5061,5080,5120,5370,5371,5619,5262,6042,6813,7012
```

**Table 9.4:** A summary of the significant overlay permutations returned by GCEM

| ENTRY | POISSON | POPN | CANCERS | COVERAGES PLUS SUBCATEGORIES | | | |
|---|---|---|---|---|---|---|---|
| 214 | 0.00003411 | 998 | 5 | +6.2 | -9.1 | | |
| 2390 | 0.00004160 | 1041 | 5 | -2.4 | +9.3 | -16.4 | |
| 2278 | 0.00004851 | 1766 | 6 | -6.4 | -11.4 | +12.3 | +16.1 |
| 2387 | 0.00005206 | 1092 | 5 | +9.3 | -16.4 | | |
| 2397 | 0.00011394 | 678 | 4 | +4.3 | +9.3 | -16.4 | |
| 613 | 0.00012836 | 1326 | 5 | +10.3 | +11.1 | +12.3 | |
| 2282 | 0.00013905 | 2145 | 6 | -2.4 | +11.1 | +12.3 | -16.4 |
| 2302 | 0.00021923 | 1490 | 5 | -4.4 | +6.3 | +11.1 | -16.4 |
| 2482 | 0.00041301 | 4852 | 8 | +3.5 | -4.4 | +6.3 | -16.4 |
| 2283 | 0.00064275 | 2870 | 6 | -1.4 | +11.1 | +12.3 | -16.4 |
| 2284 | 0.00086889 | 3044 | 6 | -11.1 | +12.3 | -16.4 | |
| 2290 | 0.00087247 | 4196 | 7 | -1.4 | +3.5 | +11.2 | -16.4 |
| 2297 | 0.00096904 | 2075 | 5 | +3.5 | -5.4 | +11.2 | -16.4 |
| 2356 | 0.00132317 | 2228 | 5 | -5.4 | +10.1 | -16.4 | |
| 2506 | 0.00142631 | 15263 | 14 | -2.4 | +3.5 | -4.4 | -16.4 |
| 2301 | 0.00172997 | 7517 | 9 | -5.4 | +6.3 | -11.4 | -16.4 |
| 2291 | 0.00176050 | 4751 | 7 | +3.5 | +11.2 | -16.4 | |
| 2267 | 0.00185741 | 4797 | 7 | +6.3 | +10.3 | +12.3 | +16.4 |
| 2320 | 0.00190470 | 9144 | 10 | -1.4 | -8.4 | +11.2 | -16.4 |
| 2312 | 0.00212745 | 9287 | 10 | -1.4 | -7.4 | +11.2 | -16.4 |
| 2511 | 0.00245931 | 39394 | 26 | -1.4 | -2.4 | +3.5 | -16.4 |
| 2306 | 0.00270154 | 11231 | 11 | -6.4 | -11.4 | +16.1 | |
| 2341 | 0.00278889 | 2650 | 5 | +10.1 | -16.4 | | |
| 2285 | 0.00319822 | 9842 | 10 | +11.2 | -16.4 | | |
| 2319 | 0.00321829 | 6742 | 8 | -2.4 | -8.4 | +11.2 | -16.4 |
| 2302 | 0.00347149 | 837 | 3 | +4.3 | +6.3 | -11.4 | -16.4 |
| 2501 | 0.00355796 | 8392 | 9 | +4.3 | -5.4 | -16.4 | |
| 2458 | 0.00371773 | 53431 | 32 | -1.4 | +3.5 | -7.4 | -16.4 |
| 2510 | 0.00374361 | 53457 | 32 | -1.4 | +3.5 | -16.4 | |
| 2327 | 0.00502069 | 4348 | 6 | +3.5 | -9.4 | +11.2 | -16.4 |
| 2483 | 0.00561382 | 17857 | 14 | -2.4 | -4.4 | +6.3 | -16.4 |
| 2336 | 0.00589387 | 1014 | 3 | -4.4 | +10.3 | +11.1 | -16.4 |
| 2508 | 0.00619683 | 19971 | 15 | +3.5 | -4.4 | -16.4 | |
| 2485 | 0.00698984 | 20253 | 15 | +4.3 | +6.3 | -16.4 | |
| 2486 | 0.00717755 | 1090 | 3 | +4.3 | -5.4 | -6.4 | +16.1 |
| 2337 | 0.00743058 | 1104 | 3 | +3.4 | +10.3 | -11.4 | +16.1 |
| 2507 | 0.00773503 | 18567 | 14 | -1.4 | +3.5 | -4.4 | -16.4 |

These concern all the overlay permutations which GCEM reported back as being possibly interesting in terms of a relationship between the distribution of ALL and environmental factors in the Northern Region. The positive or negative symbol found before the coverage number highlights which environmental factors are associated with the resultant overlay, ie. inside or outside a buffer zone. The decimal place in each case provides further detail on the sub-categories of interest and are linked to the codes summarised in Figure 9.2, ie +4.3 refers to coverage 4 (incinerators) and in particular category 3 (which denotes the 5km buffer corridor)

Firstly there is a reference key which links this detailed section to the summary of significant permutations. It informs the user of the coverages which were involved in that particular overlay permutation and details of population counts, cancers and the resultant poisson probability. The key feature in this section though are the coverage numbers since this isolates the coverages to be searched and the polygons which constitute the significant areas of interest, as well as the EDs which contain the significant ALL cases. These can all be used to put the results of this analysis back into a spatial context and allows the overlays that can now be created using GIS commands to be refined. Figure 9.2(a) and (b) illustrate how the information in entry 613, described in Figure 9.1, can now be represented in a GIS. Figure 9.2(a) shows the result of combining these three coverages and provides an overlay of cancer cases to demonstrate the difficulty that an epidemiologist may have if confronted with this as a means of deducing a possible relationship. Figure 9.2(b) demonstrates the areas of interest that GCEM has selected, the first of these takes into account the radiation sites buffered at 5km (i), the railway buffer at 150m (ii) and then waste disposal sites at 5km (iii), whilst (i) is the resultant area which is produced from combining these and deducing the key intersecting polygons. There is one more step which must be executed before the results from GCEM have been completely replicated and this is illustrated in Figure 9.3 which focuses on the polygons isolated in Figure 9.2 and highlights four out of the ten polygons which resulted. These are the ones that actually contain significant cases of ALL based and are distinguished by the important EDs which are returned by the GCEM program. Figure 9.4 and 9.5 provide another illustration of GCEM results this time for entry 2291.

This process can be repeated for every entry in the summary table. In some cases the resultant coverage may only retain one significant polygon, whilst others will be like the last example where several small polygons are involved. It should be noted that unlike the Poisson probability tests in Chapter 7 the GCEM results are less likely to be driven by small number problems because the program makes provision for only accepting polygons which have at least 500 people at risk.

An overall interpretation of the GCEM results demonstrated that in many cases the coverages involved in the significant permutations appeared to be operating in a negative fashion. In other words the cases were found within areas of impact for one or two environmental sources but away from some other pollution site, for instance away from the primary road corridors (variable -16) was a recurrent factor in the

Figure 92a: Entry 613, a significant overlay permutation picked out by GCEM

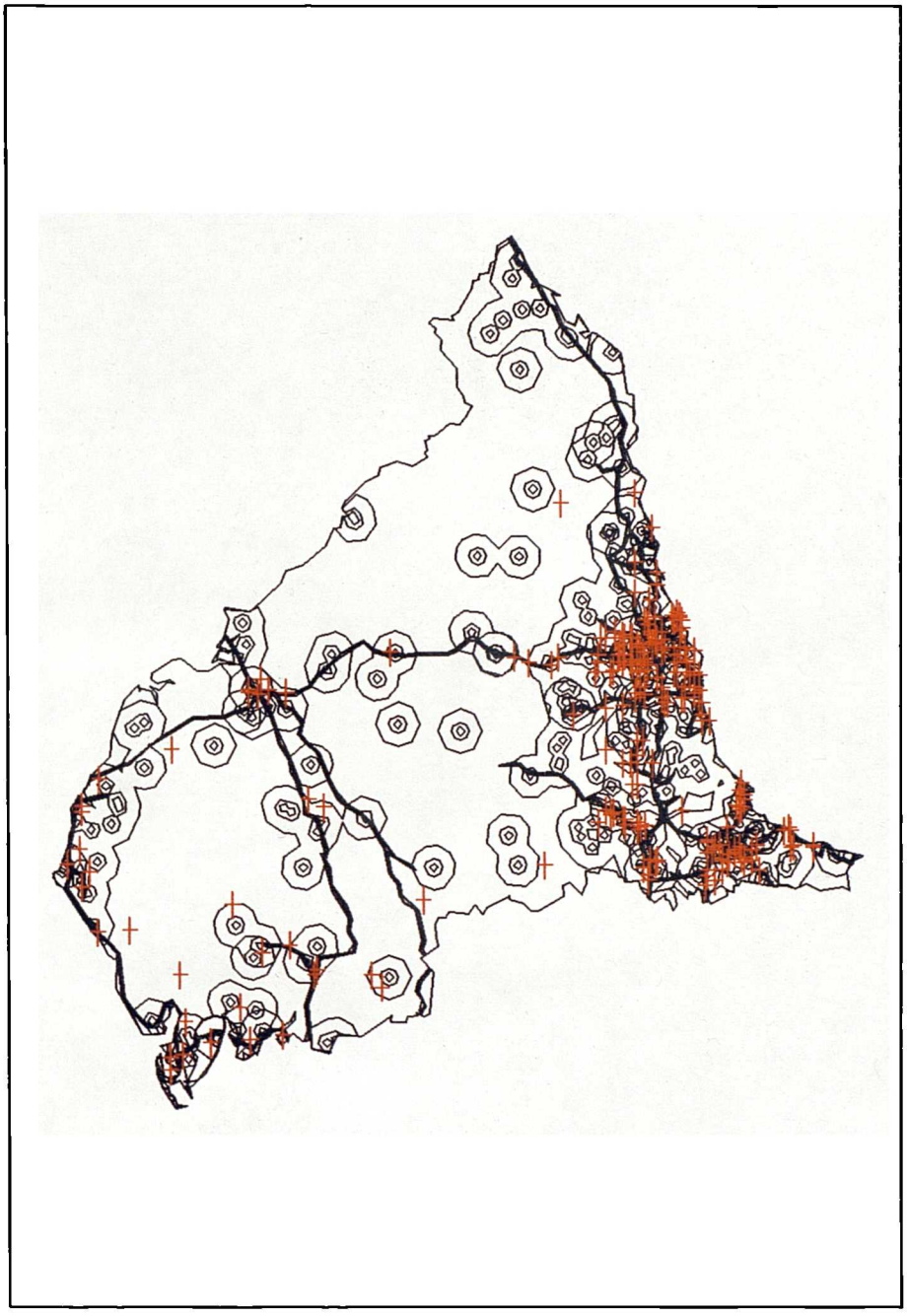


Figure 9.2b: A breakdown of the key polygons represented in GCEM Entry 613

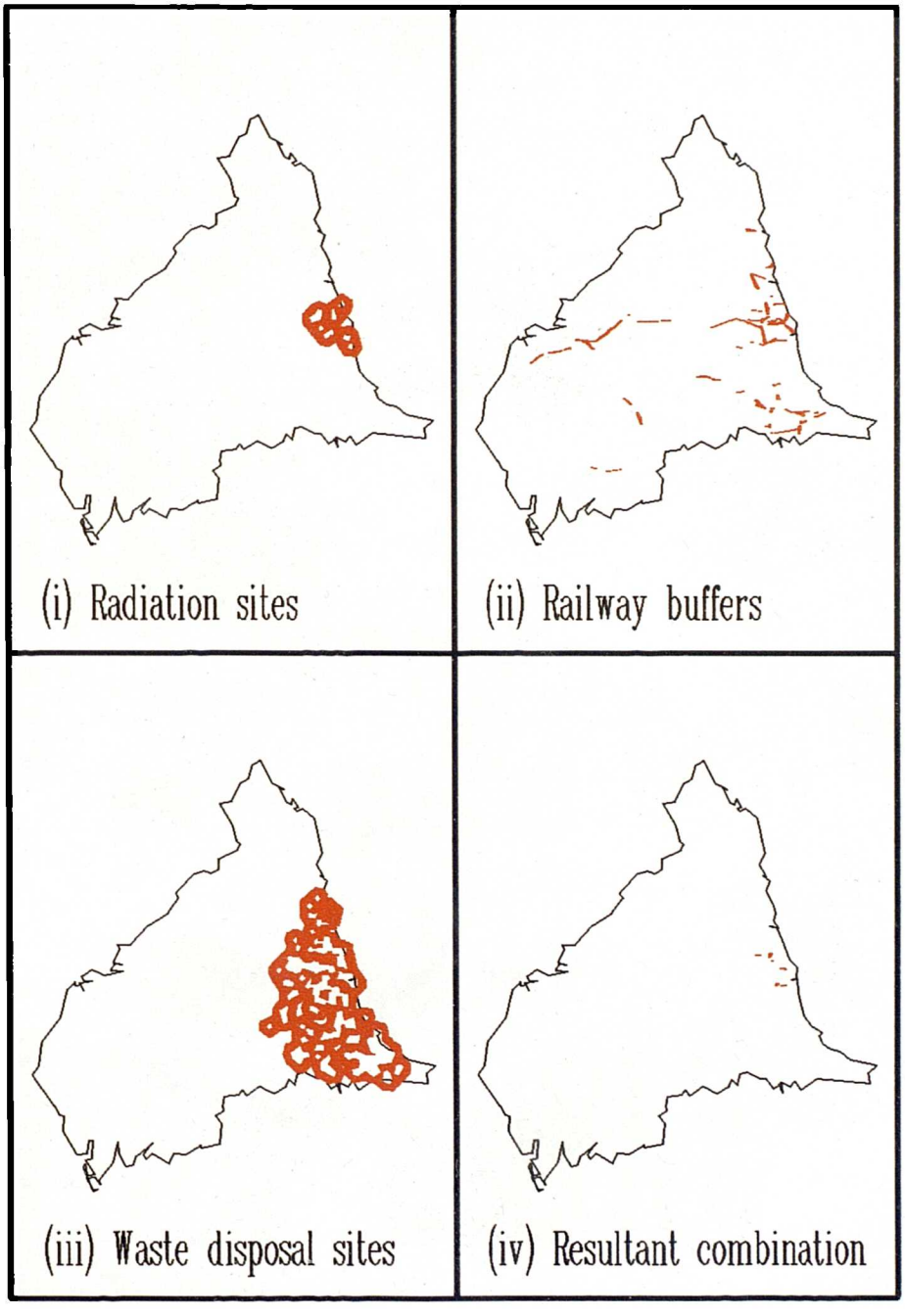


Figure 9.3: Entry 613, the polygons satisfying the GCEM overlay, showing the most significant areas

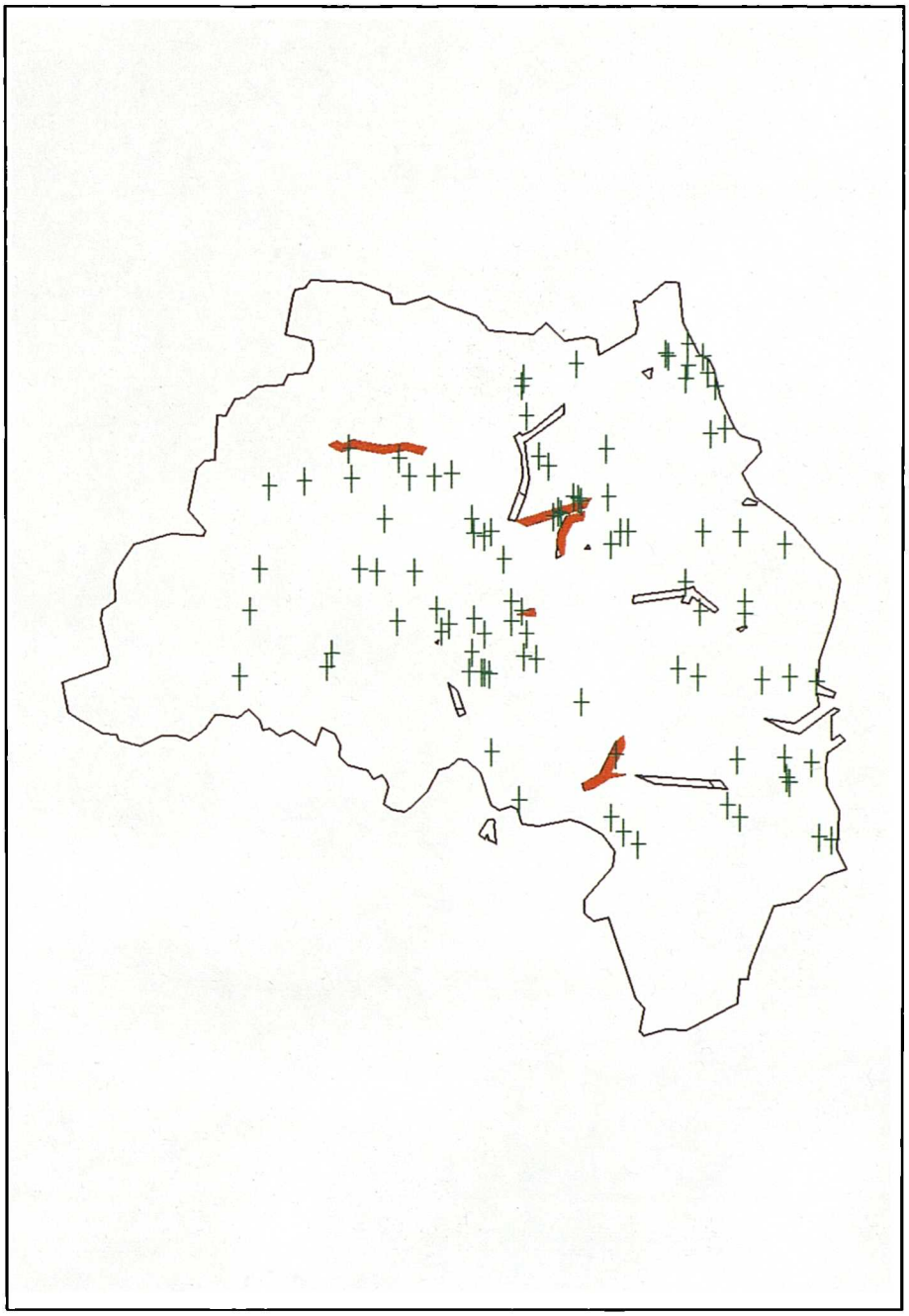


Figure 9.4: Another example of GCEM results, Entry 2291

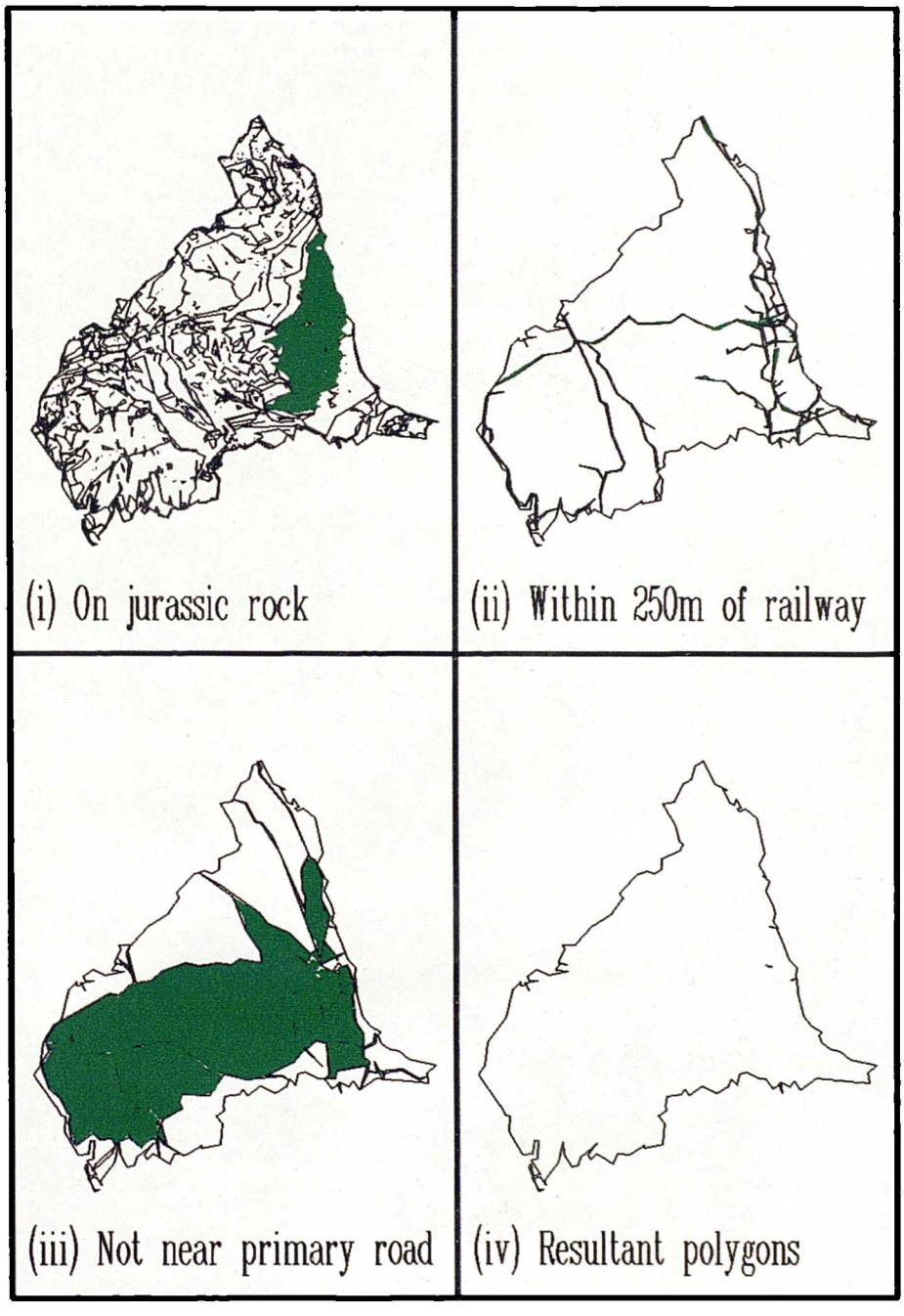
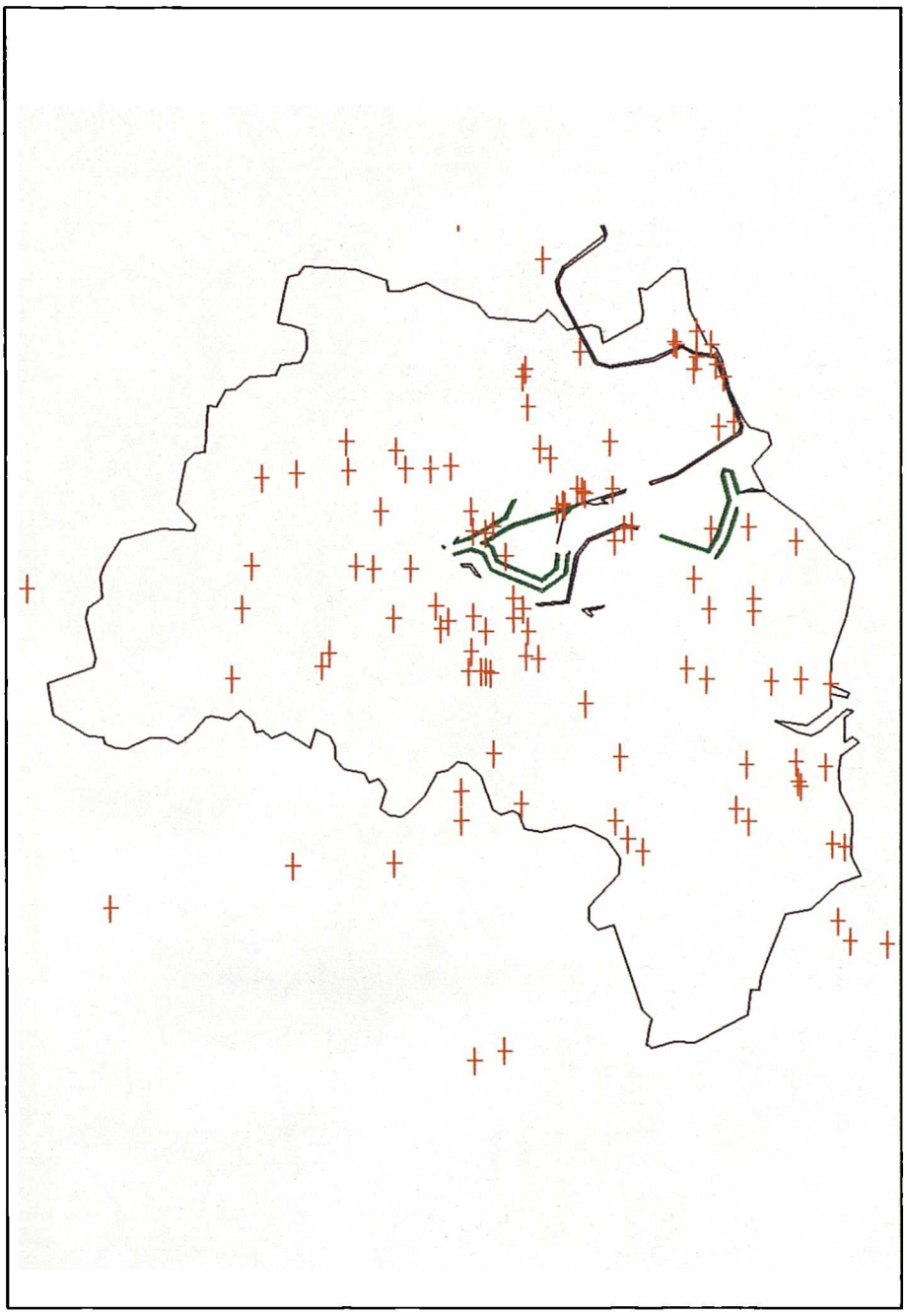


Figure 9.5: A focused view of the key areas in Entry 2291



GCEM results. What this means in terms of a possible relationship with ALL is not immediately obvious. However it may represent another example of a coverage acting as a proxy for some other unknown aspect of the environment. Thus it is not the fact that ALL is inversely related to the lead pollution associated with roads but that the area at some distance from these roads is characterised by some other interesting feature, maybe a particular type of housing.

In summary therefore GCEM only really defines areas for further research rather than specific evidence of environmental causes of ALL. Thus by comparing Figures 9.2 and 9.4 it would appear that certain areas in the region are occurring more often than others. It must also be considered therefore that location is an important factor for flagging key polygons. Further research involving different data and/or areas must be performed to clarify the associations that are being identified by GCEM. This should serve to determine whether it is the location that is driving the significant polygons (ie. different coverages every time but in the same areas), or whether it is the recurrent coverages themselves that are important (ie. the predominance of variables 10, 11, 12, and 16 noted in this study) when establishing possible relationships.

## 9.6. What has GCEM achieved?

It follows from this discussion that even if some strong spatial associations can be found they will not necessarily be directly causal but will identify spatial surrogates and proxies for other unmeasured and probably unknown factors. What these missing causative factors are cannot be determined by spatial analysis alone and requires more detailed investigation using different techniques and data, perhaps at a different scale of study. The best that spatial analysis and GIS technology can do at present is to indicate where to look and provide some guidance as to what to look for.

This limitation is absolute but it cannot be denied that any insights into complex spatial patterns are worthwhile and can provide a framework for subsequent study. However one inherent problem of research using geographical data is that some databases are at best only surrogates for other features and therefore demand that any interpretations proceed cautiously. It is realised that there comes a point when other, more detailed and less geographically based technology should take over. This is not to underestimate the immense utility of geographical information and analysis but merely to recognise that there are limits to what can ultimately be achieved, especially

in the search for complex relationships. However, it should be stated that current geographical analysis technology has by no means reached these limits yet.

It is argued that GCEM represents a potentially useful spatial analytical tool for use within GIS to search for relationships, or at least to generate hypotheses for subsequent testing. The empirical application in Northern Region identified some interesting results but it also emphasised that any hypotheses generated from the results need to be tested elsewhere, preferably outside the Region. A more general perspective would regard GCEM as a search platform within which other relationship detectors could be included.

The statistical component of GCEM should not be refined at the expense of the power of the test though because this may well render ineffective the advantages of any exploratory and descriptive style of geographical analysis. At best therefore it is hoped that the locational sensitivity retained by a GCEM approach will stimulate the fertile imagination of the user and actively assist in creating new hypotheses to be tested elsewhere. Used properly GCEM could be employed in many different areas as a source of map-based intuition which is of descriptive and suggestive value. The concept therefore seems to be worth developing further but the development of relationship seekers for use within a GIS environment also needs to be broadened to include comparisons with other methods and approaches. For instance, the ability to train artificial neural nets to identify spatial patterns and relationships could be used in subsequent versions of GCEM moving it out of a statistical analysis domain (Aleksander, 1990).

## 9.7 A summary on Spatial analysis in GIS

Chapter 7 through 9 have suggested that spatial analysis, or lack of it, within GIS is only the tip of a very big iceberg. Many end-users including academics, are beginning to realise this problem, but as yet they have not established a clear view of the types of generic spatial analysis tools required to overcome these. This evaluation has suggested that this issue of GIS must be addressed urgently, particularly if GIS is to be used in a HEGIS as a useful spatial epidemiological tool.

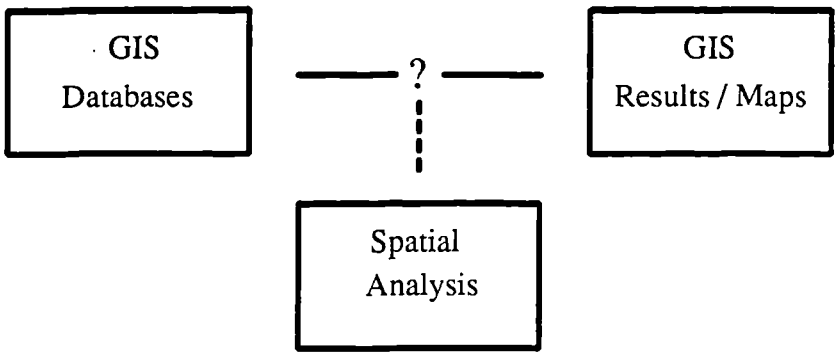
The outcome of these 'complimentary spatial analysis' chapters has been to suggest that the ability to completely embed all types of complex spatial analysis functions

within any given GIS is probably unrealistic. Chapter 8 offered an example whereby the Besag and Newell approach was formalised into the system by using the ARC Macro Language. However it ran much slower than the equivalent method running outside the system and demonstrated that the relationship between GIS databases and existing analysis packages/techniques is by no means simple and can be computationally unfeasible. GISs could be designed to interface directly into FORTRAN or C programs which access data and then perform the necessary analysis, returning results in a GIS database compatible format. A generic version of these though may not be possible given different hardware platforms which access data and run programs differently, thus any black box approach would have to be recreated for certain hardware platforms ie. Unix versus VAX versus PCs. End users may not be able to keep up with the developments in hardware in order to simply accommodate the developments in software add-ons which may involve spatial analysis in GIS. The most practical scenario may therefore be to maintain spatial analysis tools in a complimentary role to GIS, leaving GIS to do what it does best managing, manipulating and presenting data. This argument is summarised in Figure 9.6.

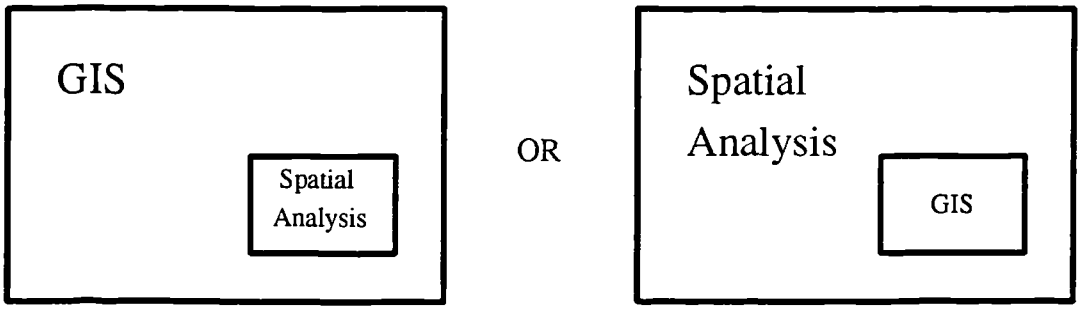
These chapters have served a dual purpose, on the one hand they have attempted to further the spatial analysis aspects of GIS and answer leading questions in the epidemiological field. On the other they have been instrumental in evaluating the limitations that presently face GIS technology. It is noted that any solutions that are developed to overcome these problems must be designed to suit both generic and application specific requirements, especially if GISs future as a spatial analytical tool, or even an aid to spatial analysis, is to be assured. Other aspects concerning the possible limitations of GIS are the subject of Chapter 10 which highlights other areas of interest that could significantly effect successful GIS implementation in application environments.

**Figure 9.6:** Summarising the possible relationship between 'Spatial Analysis' and GIS.

(i) Existing GIS environment in terms of spatial epidemiology.



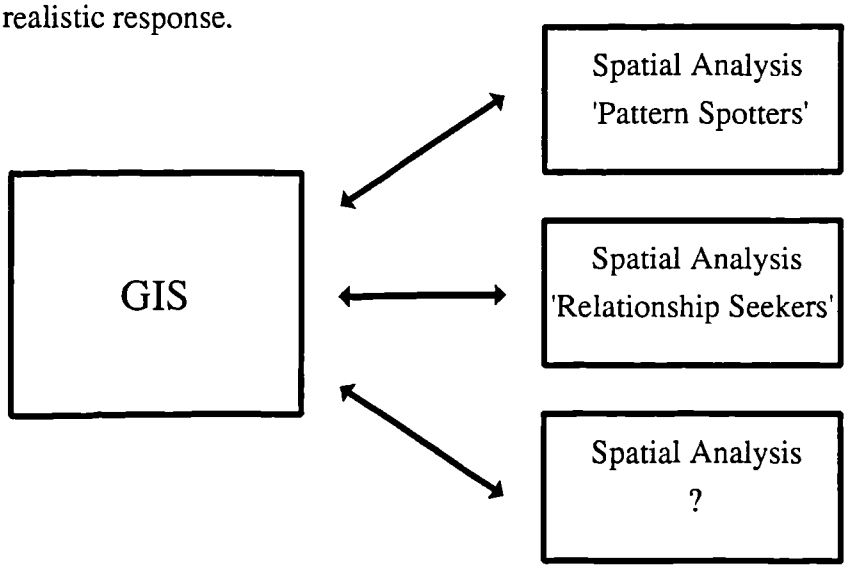

(ii) Possible solutions.




Embedding existing techniques eg. GLIM.

Embedding GIS in a standalone system.

(iii) A realistic response.




GIS and Spatial Analysis techniques: in a complementary role.

# CHAPTER 10

# BROADER ISSUES WHICH CONCERN THE BUILDING OF A HEGIS

One of the objectives of this research was to build, implement and evaluate GIS. Chapters 3 to 7 discussed the development of a GIS application based upon the available 'state of the art' technology. They also highlighted some of the factors of the new technology which are not quite so desirable and which the potential end-user is unlikely to come across by reading standard GIS promotional brochures. Thus whilst previous chapters have briefly discussed some of the problems in implementing a GIS. This chapter will go into more detail about these areas and essentially serves to set the scene for the way forward in GIS.

## 10.1 Evaluating GIS: The aftermath!

This chapter recognises three areas, in addition to spatial analysis discussed in Chapters 8 and 9, which may prove problematic in future developments of a HEGIS. These include the implications of both health and environmental spatial data, as well as the inaccuracies which can accompany every stage of 'The GIS Process'. Finally reference will also be made to the importance of the human element and the impact that individuals can have upon the development of any new technology. All of these factors are discussed in order to highlight the multitude of problems which can plague any GIS application and the areas which must be researched further if the credibility of GIS as a major scientific tool is to be sustained.

## 10.2 Implications of available data

### 10.2.1 Environmental databases

At the European scale programmes such as CORINE have made significant contributions to the availability of environmental databases. The principal challenge was to maintain consistency and quality of data. This pilot study has demonstrated that this is not a particularly simple task even when the study area is restricted to the

Northern Region. Inconsistencies in this thesis have been attributed to a variety of sources such as differences in timeliness, measurement methods, spatial coverage and density or subjectivity of sampling techniques adopted. In addition, these discrepancies are invariably hidden behind misused or misinterpreted terminology, imprecise definitions and poor documentation, not to mention those databases which each as surrogates for missing or otherwise unmeasured aspects of the environment.

In many cases the decisions which are made in the development of environmental databases are largely ad hoc and at best designed to meet the needs of the specific application being investigated. Unfortunately, until there is a widely disseminated definition of European Standards for data collection the responsibility for data will always lie with the specific application end-user, who will tend to have little regard for his/her data's potential outside their field. Only with rigourous standards will rich sources of data, produced at the national and regional scale, be able to contribute to invaluable international libraries of digital data. The function of a European HEGIS could be to provide a secure foundation for the future of GIS, by allowing a multitude of users to readily access data for their application and with the added advantage of providing a certain degree of confidence in the accuracy of such information. For instance, in this research at least one to two years could have been saved if environmental data were to be made more readily available, releasing valuable time for more fruitful analysis and research into the actual causes of ALL.

Environmental databases should not be confined to spatially referenced information. Non-spatial data sources are of equal importance and can provide key data on aspects of toxicology and chemicals. These were not available for this research but it is recognised that they would be extremely useful in terms of modelling the possible exposures of individuals to harmful chemicals in the environment, and reducing the trial and error aspect of creating areas of impact around certain sources of pollution.

### 10.2.2 Health related data

An integral part of the development of a HEGIS is the formulation of inventories and registries for all disease and mortality into a GIS compatible format. A Registry for cancer is currently recorded at a regional and national basis. However this should be extended to incorporate all types of health conditions. This may be achieved by ad hoc studies geared to assess disease incidences and the inclusion of routine data

records, such as hospital discharge and disease notification registries. Data such as these are available now but they are often not adequately geo-referenced for immediate inclusion into a GIS framework or they possess poor spatial resolution codes.

To be successful any epidemiologically related database and GIS application will need a number of guidelines on data formats and standards which would take into account major issues such as;

### *10.2.2.1 Comparability*

In general, efforts need to be made to improve consistency and comparability of datasets, especially if they are to be employed in both large scale environmental health GISs, such as the European HEGIS, as well as be effective at smaller scale studies. Standardised methods of data collection should be advocated wherever practical. This applies to both environmental and socio-demographic variables. In socioeconomic terms this may require the European HEGIS committee to provide a list of classifications for housing, occupation, employment status, and car ownership, which all Member States would then be able to follow. In addition, some reference should be made to the minimum amount of data that accompanies individual records of diagnosis, such as postcodes for place of diagnosis, place of birth, age, and sex. As a general rule, such data should be gathered at the most disaggregate level possible. Since GIS software is adequately equipped to aggregate data, as and when is necessary, but inferences from national databases to a regional scale may not be so easily attained. or for that matter acceptable.

### *10.2.2.2 Currency/Accuracy*

As mentioned on a number of occasions it is important that there is a full documentation of data characteristics and only subsequent data manipulation which will ensure a realistic interpretation of the results. Currency of data may be particularly significant if the data are to be used in the forecasting and policy making of sensitive health issues. At this point, the availability of historical information alone may not be adequate. The cancer registry employed in this research may be queried on this aspect. Since despite claims of its accuracy based on a number of rigourous cross checks, there is a time lag between the diagnosis of incidences and availability

of data for analysis. This brings into question the feasibility of GIS as a real time application tool and whether accuracy of databases may have to be forfeited in order to tackle responsive modelling of evolving disease patterns.

If GIS is to be of any use in responsive modelling, health data and complimentary databases must possess a temporal element. This is most pertinent when using population at risk data. In the UK this information is derived from the census where an estimate for the population is based upon one day in 1981 and this is used to represent the population base for cancer data covering a decade of diagnoses (1976-1986). In addition this dataset is subject to various sampling frameworks, particularly effecting the locational aspects of certain groups of the population such as students, military, and other mobile people. These problems will be carried through into the analysis stage and may serve to bias final results. Control on the quality of data to be employed within GIS is therefore a must!

#### *10.2.2.3 Duplication*

Any intervention by a national or European body must not stop at guidelines for data collection. Special coordination must also be given to prevent the duplication of time and effort in the development of any GIS application. The aim being to prevent the collection of the same data, be it at different scales or resolution, by opposing organisations. This is likely to result as a lack of communication and publication of data available and due to the cost of data capture which leads to selfishness and the sharing of data only at considerable monetary cost. Thus there is a considerable danger of vast amounts of unproductive activity resulting if this problem is not solved, particularly for large scale applications. The key must be towards mutual benefit through the sharing of centrally linked databases. This has been recognised by the European Initiatives and is reflected in the following statement, (Target 19, 1990):

> '.. information at any level is likely to be most cost-effective where the system provides for multiple applications and meets the varied needs of different user groups'(pp 6)

### *10.2.2.4 Confidentiality*

One very pertinent issue which particularly effects the building of a health GIS has to be that of confidentiality, especially when dealing with very sensitive records on individuals. The strictness of the medical ethic demands privacy at all times but raises a number of problems in the collection, transmission, storage and use of any such data relating to identifiable persons. A major HEGIS project would have to consider where the line is drawn between privacy for the individual and the benefit to the society as a whole from any analysis which maybe carried out at this more detailed level. A geographic data resolution of 100 metres, as used in this research, may well be considered sufficient to protect individual identities. In turn identities may be further masked by the general presentation of results. A useful means of overcoming the problem of confidentiality, while still allowing both individuals and groups of people to be identified, is to use of linkages.

### *10.2.2.5 Linkages*

Linkages can have a number of benefits, essentially it would involve attaching a unique personal code to each member of the population. The advantages being the ability to establish links between families or persons who have resided in the same location. In turn this may provide some distinction between the possible generic versus environmental risk of a diseases such as ALL.

The idea of linkages was not directly used in this research, but some of its advantages were expressed in Chapter 3. In particular, this refers to the inadequacies of presently available medical data where spatial referencing is confined to the individual's home address. However, working or social environments may prove more influential in terms of the well-being of the individual. Hence aspects of an individual's occupational exposure should be taken into account in the building of a comprehensive health database. Linkages would therefore enable such information to be traced, as well as establishing possible temporal movements, that would allow flows of people in space to be mapped ie. a change of address or job, since this in turn may coincide with an influential change in an individuals environment. A more continuous approach to monitoring and linkage through time therefore is logical and necessary, given that nothing in life or the environment for that matter is constant. It also provides an exciting future for a dynamic approach to GIS.

#### *10.2.2.6 Dissemination*

This subject was tackled in Chapter 7 where it was demonstrated that GIS is very good at displaying results in a variety of different ways, however at best the maps produced were only really a means of descriptive epidemiology. It was assumed however that the information contained in each map may be sufficient to promote further thought and alternative areas of research, especially if interrogated by those with the medical expertise to make qualified judgements. This should not be seen as the end stage of GIS, but rather as another step in an essentially interactive process where data can be employed for various analysis and presentations, but in turn can fuel other ideas and areas of research involving further data acquisition, development and presentation.

The culmination of these guidelines is to demand that stricter coordination and standards be outlined. These will not solve all the problems surrounding health and environmental data, but they may go some way to improving the existing situation. Some documentation administered at both the supply and demand level may be sufficient to develop an acceptable code of practice between data gatherers. This may appear straight forward, but it is a major task. At present few agencies carry out documentation for their own reference never mind public dissemination. Even academia suffers from this problem, where there is a fear of reassessing other academics' information gathering. This has resulted in vast quantities of social survey data lying untouched in the ESRC Data Archive safe from critical reanalysis (Blakemore, 1990).

This section discussed the implications of raw spatial data gathering. Once this hurdle has been crossed and a GIS implemented, a whole new set of issues become important. These specifically relate to the ways in which datasets are manipulated and transformed within GIS. This can, and invariably does, lead to inaccuracies in data, commonly referred to as 'Error in GIS'.

## 10.3 Error in GIS

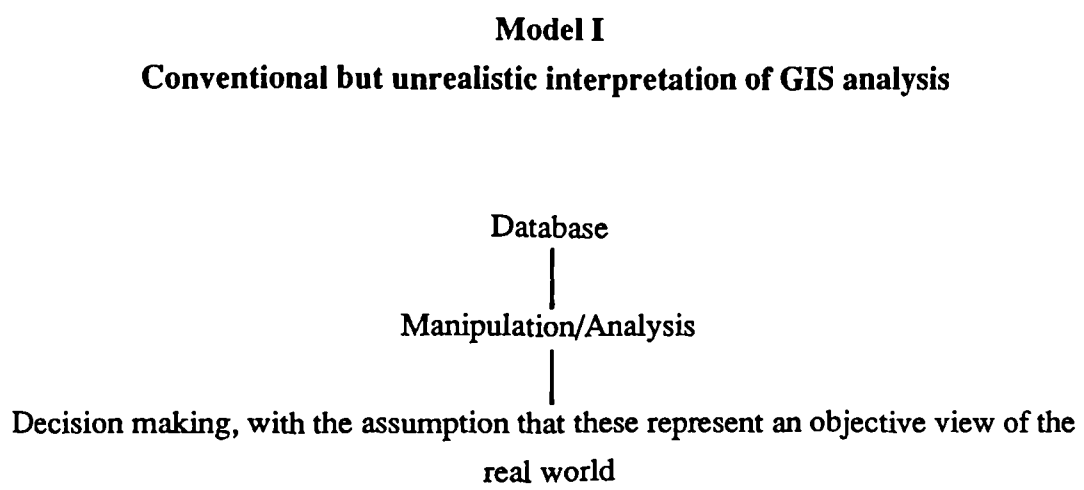
For the purpose of this research the problems of error were largely ignored. This was due to the fact that the main objective of the research was to evaluate GIS as an

epidemiological tool. Thus the ability to search for environmental causes of ALL was considered to be far more important than the need to actually find a cause or make crucial decisions upon the results produced.

Two perspectives can be adopted when considering the importance of error within any given GIS application. The first perspective is that it can be ignored, as was the case in this PhD, because the technology was used as a descriptive tool. The results are interpreted irrespective of the problems which are inherent in data acquisition and manipulation. The epidemiologist may also argue that the need to find any possible causative factor cannot wait until error free technology and/or error handlers come along to tackle these kinds of sensitive issues. The second perspective is that the impact of error may be assumed to be vital, because the results obtained would be used as a basis for establishing important policy decisions. Thus false inferences and bad decisions made in the light of GIS analysis would be costly both in time and consequences, especially if the subject to be tackled concerns that of child health.

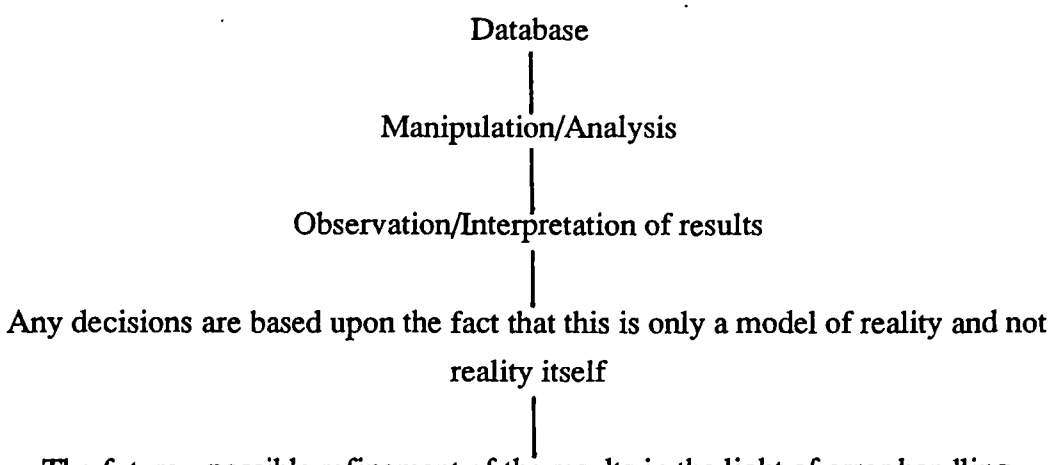
Figure 10.1 summarises these two lines of thought. The second of these is considered to be most applicable in this research where the philosophy is that whilst there are no other alternatives to 'Error in GIS' applications must make do with these inadequacies, although this should not be the long term response to these problems.

**Figure 10.1:** The possible models for rationalising 'Error in GIS'



**Model II**

**Parallel interpretation of GIS analysis**




Obviously before such techniques for data handling can be developed sources of error and how they can effect subsequent GIS analysis and map presentations need to be fully understood. This in itself is not the easiest of tasks since the inaccuracies that occur in each stage of 'The GIS Process' are not always obvious, and even if they were, there is a strong possibility that they will be ignored, or worst still, never contemplated as being problematic. The following sections will emphasise the sources of error which were evident in this application, both in terms of the sources of original data and as a product of using GIS technology. These will offer some basic methods for evaluating the possible influence that these errors can have in the future. These sections do not compensate for the errors involved, but knowledge about the nature of datasets employed may be sufficient to at least give an insight into the accuracy of the data.

### 10.3.1 Obvious and understood sources of error

Theoretically, in a well documented GIS application sources of error resulting from the data capture and manipulation stages of the process should be apparent to the end-user. These are summarised in Table 10.1 and define the error in datasets which render them only partial models of the real world.

**Table 10.1:** Factors which can contribute to error in databases

| Contributing factors to Error | A Database Example |
|---|---|
| Age of data | Geology maps (latest compilation 1952) |
| Surfacing | Air pollution datasets |
| Map scales convert to 100m | Geology, Vegetation, etc |
| Density of sampling sites | Rainfall stations |
| Relevancy - data surrogates | Road network |
| Accessibility -purchasing | Road network from Bartholomews |
| Human keyboard/digitiser | Cancer Registry |

The human impact referred to in this Table involves the ability of operators to input data and can have major implications with respect to the accuracy of spatial data encoding, as well as affecting accompanying descriptive variables which are of equal importance. In most cases these inaccuracies cannot physically be corrected but knowledge of their presence hopefully improves interpretation of subsequent analysis. Obviously, points which fall outside the area of interest will pinpoint certain spatial encoding problems, and cross checks on attribute information may allow other keyboard errors to be detected and rectified. These are the easy errors to detect. There are a number of other problems which affect the forms of spatial referencing of data received.

### 10.3.2 Inherent spatial error, partially understood

Problems of positional accuracy are perhaps to be expected in a technology designed to specifically deal with spatial data. Most researchers will be conscious of these sources of error, but at the same time are willing to live with them. In some cases they may have no choice because the problems usually arise long before the end-user encounters them. These include errors resulting from poor fieldwork, distortion of base maps and limitations of the hardware and software made available. These are best illustrated by taking two examples which serve to demonstrate the problems that affect point and areal datasets.

#### *10.3.2.1 Point datasets*

*a) Attaching Postcodes*

Many GIS applications in social and economic fields will involve the handling of address-based data. In this research the geo-referencing of ALL was derived from the patients home addresses, where the postcode was converted to a 100 metre x- and y-coordinate, as described in Chapter 5. From this the necessary locational aspect for incidences was stored as a series of points. The question is; How representative are postcodes as a form of spatial encoding?

The postcode is used to represent the centroid of approximately 8 to 10 houses, although this varies between rural and urban areas. This is a rather misleading definition because what is defined as a centroid may not even locate a single dwelling assigned to that particular postcode. For example, the postcode centroid for properties arranged along a markedly curved street will not actually lie along that street but some metres from the road concerned. Another example covers a less probable situation where the density of properties making up a particular postcode unit may vary quite considerably, with a cluster of properties at one end of a street and an isolated property elsewhere which will still be attached with the same postcode (Gatrell, 1989). In terms of simply visualising the general spatial pattern of incidences this may be acceptable. However this research used address based datasets to look for possible relationships with other defined datasets. At this point the error involved in postcoding may prove far more problematic, since 100m could be the difference between a patient being located and tagged with one type of environmental factor compared to another represented by a neighbouring polygon. The next subsection will illustrate this further when boundary effects of areal features are discussed.

*b) Attaching EDs*

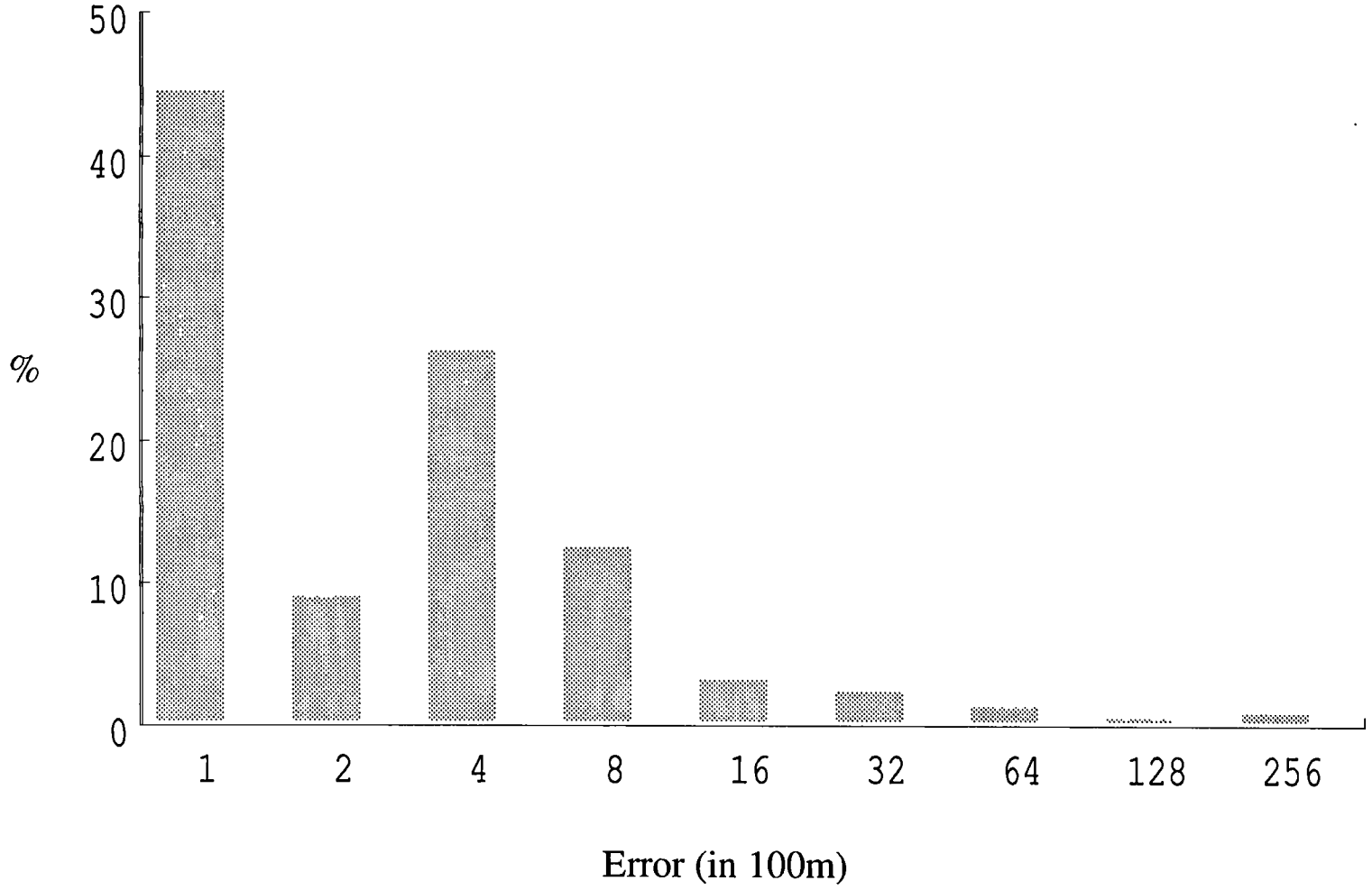
Any application which is dependent upon individuals will require a population denominator if they are to carry out any reasonable analysis concerning variations in spatial patterns. For this research it involved assigning ALL cases to the nearest enumeration district with the assumption that these were representative of the

demographic characteristics of the area in which the patient lived. Obviously this would require finding the ED that is the nearest in distance to the x- and y-coordinates of the incidences of ALL. This task is not as simple as it sounds though, which the following empirical test illustrates. A subset of cases were selected and the EDs were assigned by an epidemiologist, who used a 1:10000 Ordnance Survey census ED map to pick the ED in which the address was thought to lie. The other technique involved employing a computer program which worked out the shortest distance between two points mathematically. Figure 10.2 shows the difference between the two methods of assigning EDs. Almost 50 per cent were either the same or within 100 metres of each other. This is acceptable given the 100 metre resolution of both point datasets concerned. Also using the assumption that neighbouring EDs tend to exhibit similar characteristics a difference of 100 metres may be a feasible margin for error.

A further 40 percent of the sample were found within 800 metres of each other. This degree of error may be considered too extreme for cases found in urban districts, since basic knowledge of urban morphology suggest that housing types and associated socioeconomic characteristics can vary considerably over such distances. However, such discrepancies may well have occurred because of the rural nature of the incidences involved where the problem quite possibly could have been that the epidemiologist assigned an ED in one direction, whilst the computer program assigned it in the opposite direction, but in terms of the actual distance of the ED form the cancer case they are in fact the same. On cross checking the original dataset it became apparent that some of the problems experienced were in fact due to basic transcription errors, incorrect readings of the Ordnance Survey map, and the subjective element involved whenever decisions had to be made about incidences which could be assigned to two EDs at equal distance away. Unfortunately GIS and associated data are often riddled with such data noise. On an individual basis they may be accommodated but are compounded by error propagation and totally consumed by the technology resulting in problems being completely hidden from the end-user.

The problems with complimentary population data though should be reviewed in the light of the problems that affect the main database, ie that of the cancer database. If major spatial errors are included in the place of diagnosis then the problems associated with determining the relevant EDs may pale into insignificance. Of course

**Figure 10.2:** The error distance between EDs assigned to cases of ALL



it depends on which source of error has most impact or leverage in any particularanalysis. It may also be argued that if a rare disease exhibits a spatial pattern it will be recognised even if the population base differs by a few 100 people. Thus in a sense the assigning of wrong EDs may only be serving to compound errors which already exist in the raw cancer data, and with rare diseases these sources of uncertainty could well prove crippling.

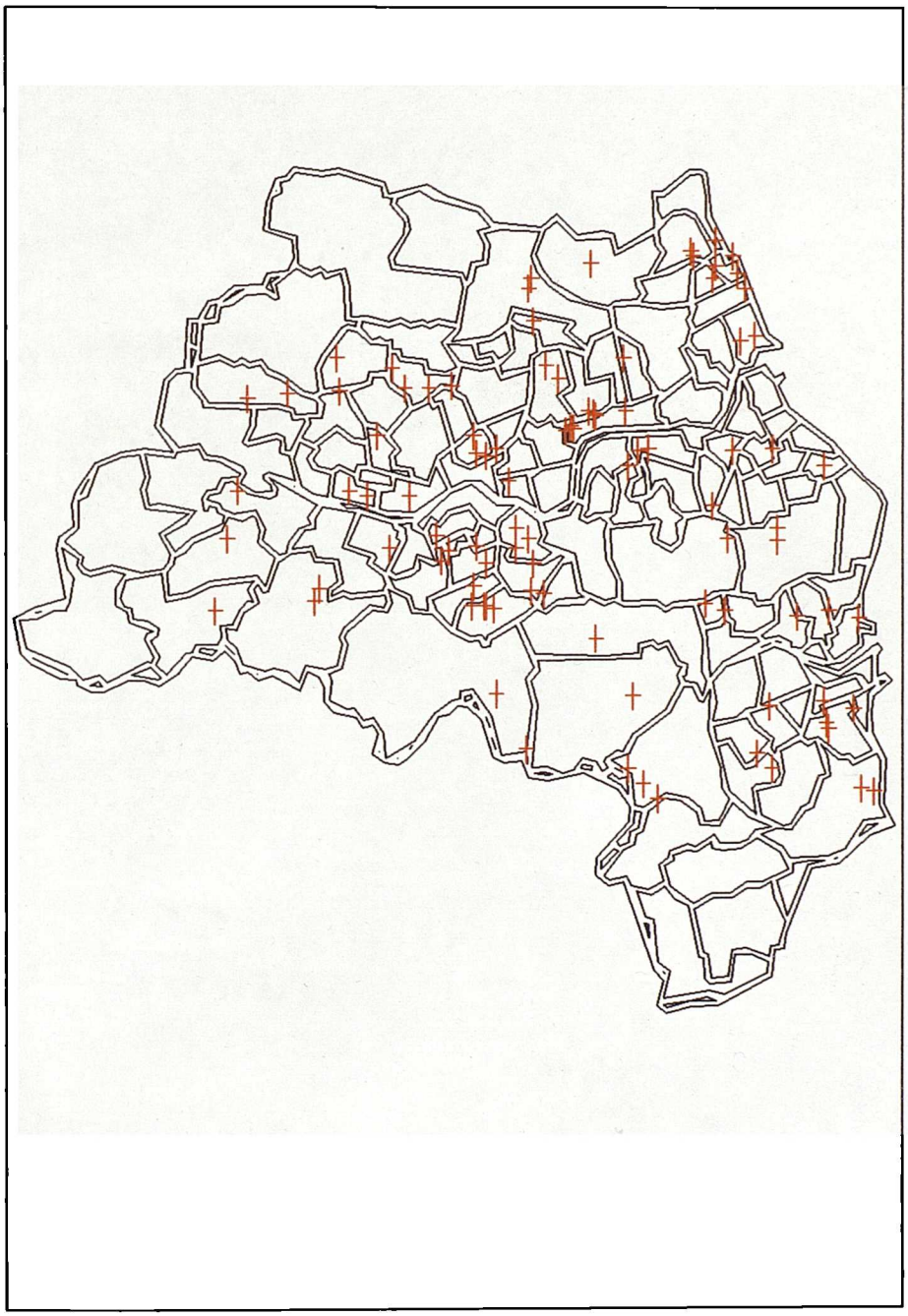
### *10.3.2.2 Areal datasets*

A number of factors will also influence how accurately areal features can be stored within a GIS framework. The data capture stage in this thesis outlined some of these key areas, including the quality of the source data, map scales, drafting skills, and the width of lines used to represent features. For example, capturing features from a map at a scale of 1:625000 (where one centimetre is equal to 250000 centimetres) with a fine drafting pen that draws lines at 1/40th of a centimetre wide, this would constitute an error margin of about 62.5 metres on the ground. In a majority of cases this may be much greater, especially as the digitiser operator attempts to recapture this map-based data into a digital form.

The effect of boundary errors will only really be a problem in cases where the incidences of ALL are found to be very close to the edge of an area. In terms of the environmental coverages used in this research, simple visualisation showed that this was not a particular problem. However the following provides an example of how 100m error in map lines can be quite influential. This example uses the ward boundaries with the exercise being to establish in which wards the incidences of ALL fall. Figure 10.3 shows these two coverages superimposed for the county of Tyne and Wear.

This possible influence of 100m error can be explored by employing a simplified version of an error application put forward by Blakemore (1984, adapted from Perkal (1966)). This estimated the uncertainty of point-in-polygon procedures caused by cartographic line errors. It involved the definition of a distance ('epsilon') about a cartographic line, where the corridor defined indicated an error band. In the original application therefore a point-in-polygon problem was used and five classes were derived according to the position of a point relative to a digitised boundary and the epsilon distance, these were;

Figure 10.3: Wards buffered at 100m to show the possible boundary effect caused by digitisation



1. within the boundary and outside epsilon distance = 'definitely in'
2. within the boundary and within epsilon distance = 'possibly in'
3. outside the boundary and within epsilon distance = 'possibly out'
4. outside the boundary and outside epsilon distance = 'definitely out'
5. exactly on the boundary = 'ambiguous'

Blakemore's empirical experiments using data geocoded at 1km resolution, overlaid on polygon boundaries of North West England employment office areas, revealed that only 60% of the points in the industrial database could be positively assigned to an employment office area. This statistic was then employed as a means of ascribing a descriptive level of certainty to the resultant map.

This methodology has been adapted for this research to provide analysis on the accuracy of points assigned to wards. In this case the ward boundaries for Tyne and Wear were buffered at 100 metres. This distance was chosen subjectively, although it is believed that it is sufficient to represent the combination of the map scale error of 62.5m and that which would have occurred from the digitisation and GIS cleaning of the captured ward boundaries. Since no cases are actually found near the county boundary or the coast, the original categories from Blakemore's example were reduced to two ie. cases definitely found inside a ward boundary and those which were found in the 'epsilon' band. Those cases found within the buffer corridor therefore may be attributable to one ward or that of its neighbour. The point-in-polygon procedure was carried out using the command IDENTITY, and this assigned ALL cases to one of the two categories. The number of cases which fell within the error corridor could then be calculated using STATISTICS.

The result of this analysis was to establish that out of 94 cases in Tyne and Wear, there was a possibility that 18 (some 19%) could easily have been wrongly assigned due to the inaccuracies involved in the definition of ward boundaries. This could be extremely influential, especially when concerning rare diseases, because a small number of cases wrongly assigned could mean the difference of one ward being considered significant in terms of spatial patterns compared to that of its neighbour.

The latter example only took into account the error which may have occurred due to the inaccurate capture or generalisation of the ward boundary. This problem may be

further complicated when the 100 metre resolution of postcoded cancer cases is also taken into account. Thus the question is, will the margin for error still remain at 19 per cent? Again the end-user without sophisticated error handling techniques can attempt to test the robustness of these distributions. This can be achieved by taking the raw grid references for the ALL cases of Tyne and Wear and 'wobbling' them randomly between plus and minus 50 metres, repeating the point-in-polygon process using the new dataset. This can be achieved using a a FORTRAN program, again embedded within an AML. The overall functions carried out being as follows;

**STEP 1:** Access the raw spatial data

**STEP 2:** Run the FORTRAN program which wobbles the data within 50 metres of the original position

**STEP 3:** Use GENERATE to create a new point data set

**STEP 4:** Carry out the IDENTITY process which tags the point data according its presence as 'definitely in' and 'maybe' inside the ward

**STEP 5:** Summarise the 94 cases according to these two categories

**STEP 6:** Repeat steps 1 to 5 one hundred times and analyse the spread of results

The outcome of this process was to show that in extreme cases 28 percent of ALL incidences could be wrongly assigned to a particular ward, as they fall within the 100m buffered region of the boundary. In the best scenario only 19 percent would be mis-placed. By taking the spread of results from this approach to simple error handling it would seem that as a rule between 21 and 24 percent of cases maybe wrongly assigned to a ward 50 percent of the time, for Tyne and Wear at least.

The AML and program which executes this procedure can be found in Appendix I. Its shortness alone demonstrates how uncomplicated this procedure is and how quickly this type of resolution problem can be tested. This of course does not provide any means for accommodating for any error because it does not generate a magical number that assigns a level of confidence to the data, but at least it provides more information about the accuracy of data.

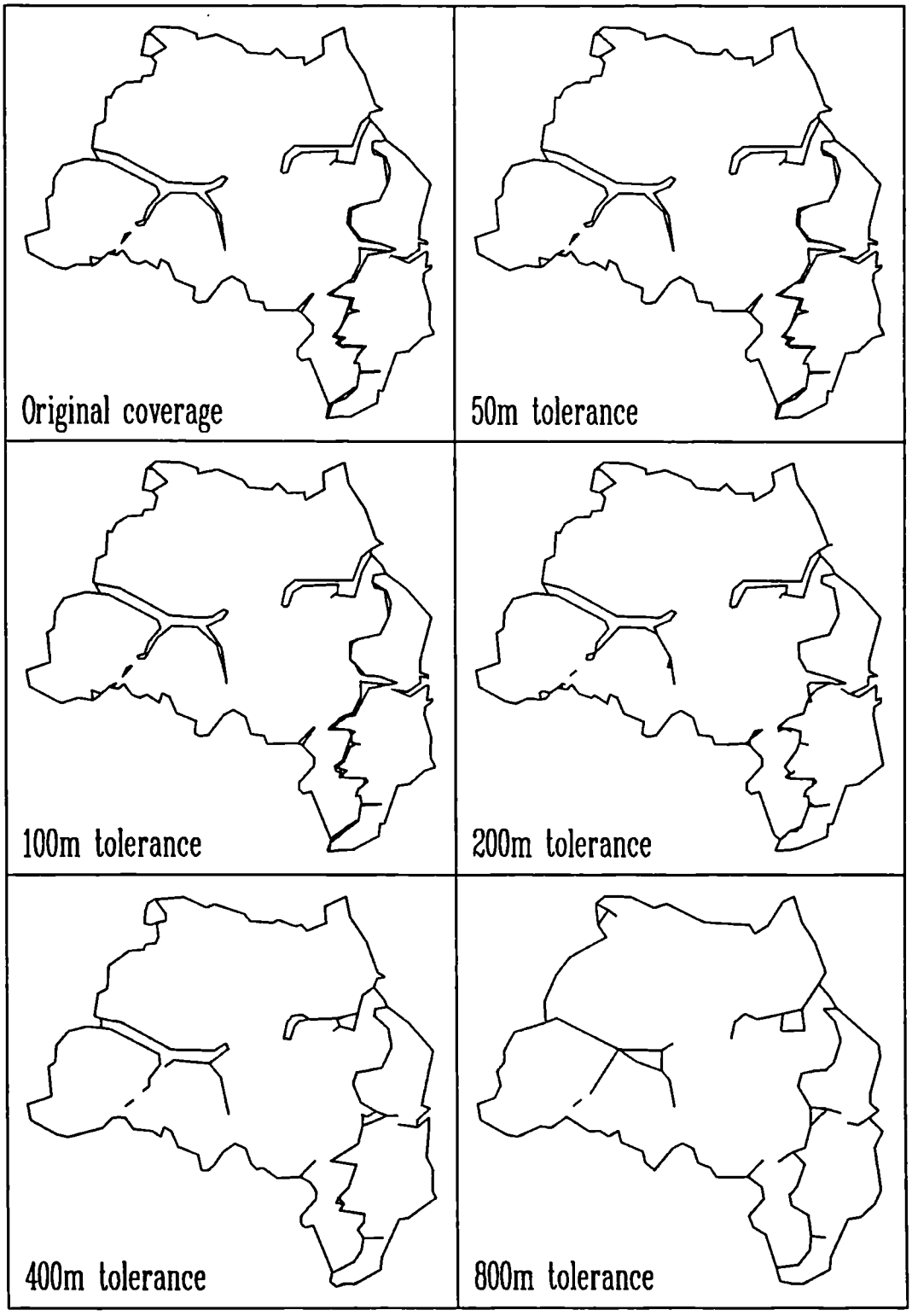
Research into error propagation in GIS is an emerging field. More sophisticated models than Blakemore's (1984) are now being developed, such as the assigning of different error models/bands to each separate segment of an arc in a particular polygon boundary. Thus rather than assuming a single distribution of errors for the whole digital map layer this suggests that it may be more realistic to assume that the error of a line with greater complexity will require a larger margin for error than that of a more simplified digitised line (see Brunsdon et al, 1990). The final set of errors which any future GIS application must also take into account are those which are created by the technology itself.

### 10.3.3 GIS error creations, invariably ignored

GIS is quite capable of manipulating, analysing and presenting data with total disregard for any error which data possesses or may have been generated by the software itself. There are a number of errors which arise simply through the processing of datasets. One is the ability of computers to process data at various levels of precision, ie decimal places, with the rounding up/down of figures which can progressively distort data through a series of calculations. For instance, INFO items can be defined to specific formats such as storing a number to four decimal places. This means however that if a value of 0.00009 is recorded it will be rounded up to 0.0 and this would be the start of an error which may get carried through into future calculations.

In addition, GIS manipulation techniques can have a considerable effect upon the topology of coverages. The flexibility of interactive editing such as ARCEDIT provided by the ARC/INFO software can encourage the database builder to fabricate errors as a means of tidying up data. GIS software packages also offer several operations for performing 'so called' CLEANs upon data. However the users inability to understand the complexity of data and the defaults which the system itself will perform can lead to the displacement of lines and undesirable generalisations because of the levels of tolerance employed. Figure 10.4 demonstrates the effect of this process upon the topology of the geological database for Tyne and Wear and the loss of detail that can result from setting up impractical tolerances for cleaning coverages.

Figure 10.4: Generalisation, the effect of cleaning tolerances



These subsections have demonstrated the errors which can effect individual databases, however these can be accumulated through every stage of 'The GIS Process', beginning with data capture through to the production of the final maps (Chrisman, 1982b). Yet all too often the data which leads to these, and the mathematical procedures involved to produce high quality cartographic products, are not supported by adequate information. Goodchild (1988) suggested that;

> 'The accuracy of a product may be defined as the magnitude of difference between the reported value and the true value'

The problem now is how to develop techniques which measure this difference. Research agendas are being devised to tackle topics of spatial data error handlers. These may not provide the complete solution or eliminate errors that presently effect GIS. However the possible development of a set of GIS compatible modules which can accommodate for certain error problems may stimulate greater confidence in the results that GIS produces. They may also have an invaluable contribution in terms of improving user awareness of the existence and importance of errors in spatial databases. The significance of informing the end-user about GIS capabilities is discussed further in the final section.

## 10.4 The Human Element

The last couple of years have seen the development of a new era in GIS, whereby it seems that researchers with four or more years of experience in GIS are now more interested in highlighting its limitations. As well as offering alternative methods to develop GIS software capabilities as was the intention of Chapters 8 through 10. The emphasis of the latter chapters however was to mainly focus upon software issues and their influence on the success of GIS in any given application. Little has been said in relation to the important contribution that the end-users themselves can have upon the potential of GIS.

This section therefore highlights another major area of concern, arguing that there is no use having very sophisticated and complicated GIS software if the end-user cannot fully understand what is involved and how to properly use it. Thus education in GIS should not be confined to the 'learning curve' that accompanies the building,

implementation and development of a GIS, but also involve changing the attitudes of end-users and organisations who will be affected by the invasion of this new technology. This in itself is a major area of further research. It is briefly discussed in this chapter as a means of concluding on future research needs in the field of GIS, as well as emphasising the impact that any end-user can have upon the completion of an application. Thus it is identifying areas where a HEGIS may fail if the human element is not considered.

> 'the concept of computer technology as a package which includes not only hardware and software but also people, personal skills, operational practices and corporate expectations.' (Campbell, 1991)

### 10.4.1 Attitudes to Information Technology (IT)

In order for any GIS to succeed the generally held perceptions and opinions on IT must change in accordance with the advances in the 'state-of-the-art' technology. This is an important issue which will need to be tackled, and in relation to a HEGIS it will involve general practitioners, Local Health Authorities, the National Health Service etc.

Initially the rapid up-take of GIS is often seen as a threat, replacing existing manual and/or computer systems, which the users understand and may wish to retain. In everyday working environments outside of academia it must be expected therefore that there will be a little inertia to up-take and a resultant backlash when the new technology is adopted. It is essential that the enthusiasm for GIS is cultivated at both the strategic and the nuts and bolts end of the hierarchical structure of any organisation. Thus the software and hardware must be accompanied by developed channels of communication, data sharing, corporate collaboration, and mechanisms for ensuring that the results are used and can be used routinely in the key areas where they are intended to contribute. Reality may be quite different however, with barriers to data access occurring due to inequalities across systems that prevent the transfer of data to systems with more limited capabilities, as well as the establishment of different vested interests for data and the results to be achieved. Again the general requirement to solve these problems appears to be the call for detailed standards and reasonable coordination. Thus it is extremely important not to under-estimate the psychological aspects of misplaced knowledge and expectations which can

accompany any advances in technology and may act as a failure mechanism in GIS development.

### 10.4.2 Realistic expectations

GIS is a highly attractive technology and in turn suffers from the problems that effect all new technologies being wrongly sold as the solution to all the spatial data handling needs of the end-user. This only serves to lead the end-user into a false sense of security heightened by the fact that the end-user invariably does not know what is wanted from a GIS, or more specifically how GIS's capabilities and limitations will effect their particular application area. Even in this research it was expected that GIS would be able to do more than produce maps. That was up until Stage III (Chapter 6) when the stark reality of GIS as a potential spatial epidemiological tool was realised. This was, and is, invaluable experience and should not be kept as the sole prerogative of the afflicted end-user. The success of technology is in many ways dependent on the availability of good advice and information which highlights both the beneficial and problematic areas of any new technology.

The presence of a well thought out strategy is also crucial for an application, as this will help to formulate the important decisions that need to be made when adopting and developing a GIS. In addition, every application will experience its fair share of set backs both technological and mental, particularly as the system fails to execute the end-user's original objectives. At this point the researcher must look to his/her goal for an incentive and a means of rationalising the problem in order to find an alternative way of tackling the problem. Documentation of GIS limitations and other end-user experiences such as this may play an important role by not only offering alternative methodologies but also reassuring other end-users that they are not alone with their GIS frustrations.

Thus the key to realistic expectations is to first of all set out application goals, which the European HEGIS has done in the form of Target 19 (1990), this will provide a solid foundation for future developments. A formula for establishing a clear view of GIS and end-user expectations of new technology may read as Figure 10.5

**Figure 10.5:** The GIS success/failure equation

What are my needs, in terms of spatial data handling?

+

What can GIS offer in terms of the problems set?

+

Is this applicable at the individual, organisational, and/or national level?

+

Who will implement, maintain and build up this system?

+

What do I expect to achieve from this?

+

Does it?????

=

The successful implementation of GIS in any application area

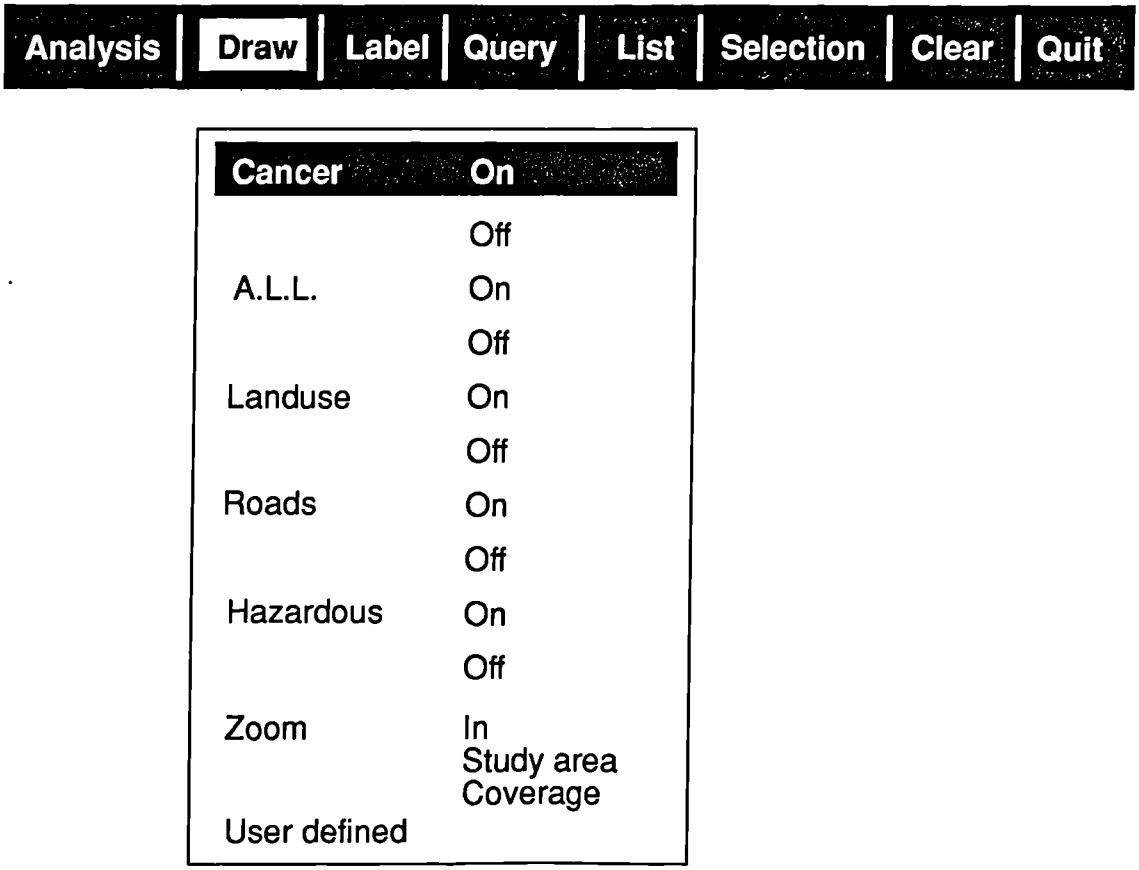
Everyone involved in this equation should also be included in the selection, implementation and the eventual use of a GIS. This would stimulate important team coordination and understanding and would constitute an essential part of the human aspect of GIS education.

### 10.4.3 Relevant customisation

In practice many of the important features of 'The GIS Process' can be obscured from the end-user in what is theoretically referred to here as 'Stage V' of 'The Process' that of customisation of GIS.

GIS vendors to some extent rely on customisation to sell their products, but is this the answer to successful GIS take-up? This process involves the conversion of the general toolbox of GIS commands and modules, into an application specific model, accessed by menu-driven interfaces. A simplified view of this is shown in Figure 10.6. All the technicalities of GIS are hidden from the end-user by high level macro languages which create simplified menu boxes requiring the occasional key word or just the press of a mouse button to obtain data and mapped information. For instance, the

**Figure 10.6:** A simple example of a menu-driven interface




Adapted from ESRI (1990), PC Understanding GIS: The ARC/INFO method

highlighted area in Figure 10.6 could include the selection of cancers which could then be viewed on the screen, it would include commands such as;

```
MAPEXTENT CANCER
MARKERSYMBOL 2
POINTS CANCER
```

This is a very basic example but it does demonstrate that the end-user can view all or part of the database without knowing anything about the syntax or structure of GIS.

As section 10.4.2 pointed out though, invariably the end-users do not know what they want, and cannot usually convey their specific needs to the vendor, this has in the past led to inadequate customisation. This can only serve to complicate the GIS rather than make it easier! Failure in the interface can lead to the whole technology being dubbed a failure because key areas of functionality cannot be accessed. Thus 'Stage V' should really be carried out in-house and should be a bonus in 'The GIS Process' rather than a substitute for the learning curve altogether. Knowledge of data requirements and original objectives/decisions will ensure that the latter customisation will be more effective. Once this stage is reached therefore the benefits of user-friendly and tolerant interfaces can be developed, to extend information to other audiences of expertise, including non-GIS trained epidemiologists. This provides a ubiquitous tool for multiple access, but the danger still remains that without important background knowledge about the implications of data and GIS, customisation simply facilitates the ability to misuse and misinterpret data.

## 10.5 Concluding remarks

The issues outlined in this section may have only dealt with the relationship between the end-user and a GIS application in a superficial manner, but they do focus on the factors which are likely to make all the difference between success and failure of GIS applications, including HEGIS. It is hoped therefore that with a little more understanding of error in data, of the development of error handlers and a greater acceptance of the importance of the human element in the GIS revolution any associated problems and/or backlashes to the new technology will not ultimately over-shadow its actual benefits. The important experiences derived from completed applications therefore should not be underestimated as a key feedback element for the

wider end-user community. It could serve to prevent others rediscovering similar, if not identical, pitfalls. In addition it can highlight key areas of deficiency within existing GIS software and indicate to the vendors that their systems are not always meeting the fundamental needs of certain end-user applications.

# CHAPTER 11

# CONCLUSION

The objective of this pilot study was to evaluate GIS in a Health and Environmental application. The specific aim was to explore GISs capabilities as a spatial epidemiological tool to search for causes of Acute Lymphoblastic Leukaemia (ALL). This chapter serves to summarise the findings and starts by answering the question originally set in Chapter 1, 'Can GIS do it?

## 11.1 Has GIS done it?

The answer to whether GIS is successful in a Health and Environment application will differ depending upon the functions that are required. This thesis did not cover all the subject matter necessary to develop a comprehensive and operational HEGIS, that is left to the European initiative (WHO, 1988). Instead it focused on one aspect of spatial epidemiology.

This research took the first tentative steps therefore in establishing GIS as a useful aetiological tool. However it found that the technology did not quite emerge as the complete answer to all spatial epidemiological problems. The findings of this study therefore partially agree with the European conclusion that GIS has;

> '...reached a sufficient level of maturity for the concept of HEGIS, available at international, national, and subnational levels, to be realizable' (pp 20, 1990)

Although it must be emphasised that whilst it provides the basis for an operational infrastructure, there are a number of areas which need to be addressed further. In other words GIS still has a bit more 'maturing' to do. Existing GIS technology may be all that is needed in these early stages. For example, the Small Area Health Statistics Unit (SAHSU), established in the UK in 1987, does not necessarily seek the answers to complex questions of causation, but a means to;

> '..examine quickly reports (formal or informal) of unusual clusters of disease in the neighbourhood of industrial installations..' (Rose, 1986)

In this case the attractiveness of data availability, flexibility and manipulation may prove to be the key to GIS success being sufficient to satisfy the majority of end-user needs. By the time their questions become more sophisticated, demanding new complex analysis procedures, they will probably be available within the GIS framework.

In addition it should be considered that this research has culminated in at least 20 new environmental digital datasets and the creation of a comprehensive cancer database for the Northern Region. This is an improvement on the situation that existed four years ago and from this foundation a HEGIS can progress further through the addition of data and techniques, as and when they become available. The 'Bible' in GIS (Maguire et al, 1991) has just been published and serves to document many of the factors discussed in this thesis. However it is interesting to note that health, and in particular spatial epidemiology, is only referred to once. This would seem to suggest therefore that the applications which have thus far tended to dominate the GIS application world have been concentrated in areas where maybe its full potential has not been recognised or contested. Thus while this new book outlines GIS development in the UK over the last decade, it would appear that HEGIS is a major research agenda for the 1990's and reflects a new era of GIS where many of the issues raised in this PhD will have to be tackled.

Thus the problems that this research encountered essentially represent the future needs of the epidemiologist, when the attraction of multi-coloured maps and overlays have been exhausted and they begin to switch their attentions to spatial and statistical analysis requirements. Chapters 8 through 10 therefore did not only constitute criticisms of GIS but anticipated GIS limitations and highlighted areas for further development, with the main problem being referred to as a 'missing link'. This 'link' would be able to efficiently interrogate the masses of data stored in a GIS, provide statistical analysis to clarify subjectively observed patterns, and basically answer epidemiologist's questions as to 'Where are they?', 'Do they cluster?' and 'Why are they there?' These may be referred to as 'higher order' questions, which in turn require a 'second generation' of analytical tools that are more appropriate within an epidemiological framework (Raybould, et al. 1991). As Chapters 8 and 9 emphasised this is not simply a 'nice' idea but one which is essential if GIS is not to become severely restricted in this field.

In the absence of any magical 'black box' module which could assist the end-user in the GIS analysis process and interpretation of distributions, Chapter 7 used the Poisson probability to attempt to flag areas of interest. This was by no means the most satisfactory solution but under the circumstances it was the best that was available. This emphasises the fact that not only does GIS software lack the necessary tools for sophisticated spatial analysis in epidemiology but that this field needs to be developed generally. Thus this situation could have an undesirable 'catch 22' effect, because whilst GIS can provide the necessary data to support epidemiology it does not have the tools to carry out the desired geographic exploration into clustering of diseases or the establishment of key relationships between phenomena. This may serve to foster an attitude that GIS is a failure and as a result the enthusiasm to keep acquiring relevant data may subside. The danger is that the ultimate response will be one of, 'what is the point?', which would lead to a lack of data and thus inhibit the further development of relevant spatial analysis tools.

GIS is presently a lucrative business and acceptable in the end-user market. The next step though should be to ensure the future of this technology. This will require a sustained input to tackle every aspect of the GIS toolbox in order to remove the types of limitations which have been discussed in this thesis. These issues include; data accuracy, coordination, analytical tools and the implications of error. This is not only important to the researcher and their specific application, but also for the purpose of the general marketing of this new technology. GIS must avoid becoming labelled an 'information manager' and drawer of 'pretty pictures'. This can be achieved by actively seeking out and developing scientific analysis in GIS which can only be beneficial in strengthening GIS appeal and applicability.

## 11.2 Pushing GIS forward

The future of GIS is, to some extent, in the hands of those who have already acquired the necessary expertise and experiences from pilot studies such as these. They have a responsibility to ensure that GIS can, is, and will be extremely useful in this area. It is not necessary to have everything done immediately, it is more important to evolve GIS and that could take another 5 to 10 years. The response to any problems that are encountered therefore should not be one of passive acceptance of the 'state-of-the-art' GIS. The moving of research goal posts to suit the technology will not push GIS

forward. Instead, completed applications should be used to flag areas for development. This was the aim of Chapters 8 to 10 in this research. This referred to both the integration of existing techniques, complimentary analysis and the suggestion for new methods such as automated 'relationship seekers'. Obviously these techniques and ideas are still in their infancy and others may express alternative views on their applicability and potential for further development, but they are an attempt to progress GIS and if nothing else the issue has been recognised. From these small beginnings GIS may then begin to flourish.

In the conclusions to Chapters 8 and 9 it was suggested that these techniques offered alternatives to spatial analysis, although initially they only really serve to provide yet another tool for descriptive epidemiology. This is still an improvement on traditional methods and, as chapter 3 emphasised, descriptive epidemiology can be extremely useful in providing an insight into the aetiology of specific diseases, especially when no other alternative is available. The preoccupation with the need for statistical techniques to provide theories with some credibility is perhaps too often over rated. Statistics can equally be manipulated and falsified through bad calculations and/or methodology, so in terms of objective inferences maps may in some cases be more reliable and comprehensible than statistical figures.

Chapter 10 suggested that GIS has probably entered a new era where people are no longer afraid to admit that GIS did not answer all their application goals. These experiences should be viewed positively because they form the basis for developing new techniques from existing methodologies or innovative ideas. One such example is GCEM which attempts to advance GIS further along its 'revolutionary' route and answers the call made by other leading GIS researchers that GIS;

> 'requires fresh thinking, and it needs to be accompanied by a new generation of spatial analytical techniques that can make the best of new opportunities. Tired, old, worn out techniques have no place in a brave new world.' (Openshaw and Scholten, EGIS 1990)

The European approach to HEGIS can make major contributions to the overall development of GIS, because the setting up of wide spread coordination and standards of data will secure the foundation of GIS. At the same time, researchers need to develop error handlers and spatial analysis techniques to ensure that the software will

gain greater credibility in the application world. At present, GIS does not offer an end result to many spatial data problems, but as Chapter 7 to 9 showed it is a very good vehicle for providing a means to an end. Thus in comparison to the definition of GIS given in Chapter 1, it is probably better typified as a;

> 'decision support system involving the integration of spatially referenced data in a problem solving environment.' (pp 1554, Cowen, 1988)

## 11.3 Technology and people

The human element of GIS was referred to in Chapter 10 and is mentioned again in this section to reinforce the opinion that end-users of GIS are as much a limiting barrier to the future of GIS as the technical problems of the software. This thesis took a spatial epidemiological problem and applied a geographical perspective to it, with a hint of epidemiological thinking. The question is, will it have more or less the same effect when the epidemiologist themselves are confronted with the same spatial data and the power of a GIS? It may be that they will require additional knowledge in order to enhance their understanding of the importance of space, and the best ways of manipulating and investigating spatial data without abusing the flexibility of the technology.

The experience of this research demonstrated how easy it is to 'expect' GIS technology to answer more questions than is actually possible based on the 'reality' of GIS capabilities. In this study the result was to strive to improve on the areas which were highlighted as deficient in a bid to solve the original spatial epidemiological problems set. Efforts must be made to reduce, or better still, remove this gap in the end-users' interpretation of what GIS can and cannot do. Pilot studies and end-user experiences therefore have a key role to play in offering alternative views on GIS and may prove invaluable to others embarking upon the GIS route.

The latter provides a strong argument for an 'alternative' guide to standard GIS manuals which would basically act as an end-user 'survival guide' to building, implementing and using GIS successfully. It would therefore be a vital document in terms of educating the GIS end-user. It could progress through each stage of 'The GIS Process', rather like any software brochure and/or GIS manual, but would concentrate on the problems that are likely to be met and suggestions for overcoming these. This

thesis provides some pointers and other researchers would probably be able to contribute a number of additional problems and possible solutions too. The response to the original Chorley report (1988) also hinted that there was a need to explore alternative areas of GIS and to educate the end-user in order to ensure that this technology was a success. It was suggested that there was a need for;

> 'More Geographic Information Systems familiarisation courses..... training opportunities in the handling of geographic information for managers and operators...' (pp 15 and 16)

## 11.4 Future prospects of GIS

There are two views on the way in which GIS may go in the future. The first of these is negative, and prophesises that GIS will fragment and disappear being nothing but a memory at the end of the century. The scenario is likened to that of the quantitative revolution in geography in the 1960s, whereby it is suggested that organisations will be looking back in ten years time and saying well 'that was GIS'. This is a possibility if the number of disillusioned end-users increase and systems fail to answer their fundamental questions. This will only occur however if the software manufacturers cannot or choose not to acknowledge the limitations of their packages and fail to develop their technology to higher levels. At some point, database management and visualisation may not prove to be strong enough glue to hold the fragments of GIS together (Goodchild, 1990).

A more positive view would be to see researchers and other end-users emerging from pilot studies with suggestions and answers for better tools which can be used to extend GIS capabilities. This should lead to increasing GIS awareness, improving data standards and data quality, training and education. The development of cooperation between the vendors, academics and the wider end-user community should lead to the emergence of a strong set of core concepts and modules. This should force the pace of GIS development and provide the basis for an exciting and long future of GIS, of which fully documented applications such as this research have already made an important start.

# APPENDICES

# APPENDIX A

## A Brief History on the Development of a European HEGIS

### 1. Background

A Consultation on Environmental Health Information Systems in Europe was organised by the WHO Regional Office for Europe at the invitation and with the support of the Federal Ministry of the Environment of the Federal Republic of Germany in Berlin (West), 21 to 25 November 1988. It was attended by 29 experts and observers from 11 countries as well as representatives of four international organisations. It was convened to determine the most appropriate strategy (or strategies) to facilitate and enhance the use of existing environmental and health data for decision-making at local, national, and regional levels in environmental health risk management.

The Consultation focused on the significance of harmful agents in the environment to public health, because of the difficulties of this subject and the extent of public concern. Research in this area was recognised as crucially important. In particular, the Consultation gave attention to developing the role of analytical epidemiology in providing direct evidence. In that respect, neither laboratory studies nor the role of viruses, bacteria, and other microbes were specifically considered. However, the influence of workplace and socioeconomic factors was not ignored completely.

In addition to establishing strategies for the use and management of environmental and health databases and relevant research, the Consultation was to identify various users of environmental health data and to review existing sources of data in the areas of exposure, toxicity, lifestyle, health care, and health outcome on local, national, and regional levels. It was also to identify differences in the priorities attached to relevant activities of national, regional, or international organisations (Commission of the European Communities (CEC), Council for Mutual Economic Assistance, Organisation for Economic Co-operation and Development) and to recommend different strategies to deal with national differences in organisation and legislation concerning health-related data collection and analysis.

This broad and wide-ranging brief encompassed the fields of chemical databases and toxicology, occupational health and exposure, database management and mapping, epidemiology, and health services policy and research.

## 2. Summary

The Environment and Health Service (EH) of the WHO Regional Office for Europe (WHO/EURO) is developing a programme of Environment and Health Information Systems (EHIS) as a health risk assessment and management tool for the European Region. The programme builds around three major project areas, each developed by a collaborating centre currently active in the specialised area and coordinated by WHO/EURO.

a) A directory of the sources and holders of information, literature and research project outcomes in the broad area of environmental risk factors, social variables and health outcomes [a metadatabase (MDB)] will be developed by an international centre with the assistance of national focal points. The MDB will be available on an electronic data system with public access wand will permit the user to obtain an overall view of the extent and nature of and methods of access to environment and health data within the Region.

b) A set of environment and health indicators appropriate for demonstrating the unevenness of environmental risk factors and health status within and between the Member States, together with a method of providing geographical mapping of these and other demographic data of interest, will be developed by a national coordinating institution with cooperation of local and national focal points [a geographic information system (GIS)], discussed in further detail in section 3.

c) A programme for the development of a network of local centres for the collection and analysis of geographically linked health data [small area health studies (SAHS) and for the development and testing of methodologies for analysis of temporal and geographic anomalies of disease incidence (cluster analysis) will be carried out under the direction of a national collaborating institution.

Each of these activities will be initiated as a pilot phase and developed in several stages. In order to take advantage of resources and avoid duplication of effort, it is

intended to use existing facilities and systems as much as possible and to begin the overall effort in autumn of 1990. Milestones will be the development of pilot models through 1990 with partial implementation by mid-1991 and full programme development by the end of 1992. Continuous effort will be made to keep Member States not participating in the pilot phase aware of progress and to extend as much as possible coverage to the Region in the period 1992 and thereafter. The overall structure of the programme will utilise WHO/EURO as coordinator but actual project activities are delegated to international collaborating centres and national focal points as indicated.

## 3. GIS indices for state of the Environment and Health

A geographical information system is one which permits information, associated with a well defined geographic identifier, to be plotted in some aggregated form, assigning average values to the areas of interest. For purposes of environment and health, the importance of a GIS lies in the ability to identify the geographical boundaries of risk factors, e.g. ground water contamination, high values for disease incidence (as in the atlas of avoidable death). The important concept is that the lower the degree of aggregation, the more useful is the data for purposes of identification of unevenness in either risk factors or of health outcomes.

A great amount of data is available in individual Member States, collected by local authorities. Usually national health statistics are made available in aggregated form only. The major task of the activity is to demonstrate the degree of unevenness in health effects and of risk factors, with two purposes in mind: 1) To use these unevenness as an indicator to governments that there is neither equality in health status, nor equality in risk factors. 2) To use such mapping of indicators to assign priorities to deal with the observed unevenness. The task of the activity will be to define a grid for data aggregated on some intermediate level, but sufficiently fine to demonstrate unevenness of risk/outcome associated with a series of important issues. The purpose of the activity is not to analyse the unevenness for origins, but to permit preliminary hypothesis generation as to potential cause based on an adequate inventory of the important risk factors associated with each separate issue. An important part of the activity will be the development of the concept of analytical indicators for the Environment and Health Service, beyond those currently available.

The identity of the participating organisations are as follows:

a) WHO/EURO: Responsible for initiating and coordinating activity at national level among member states with cooperation of international coordinating institutions. Responsible for developing indicators which will serve the environmental health service, and which are related to current issues of importance.

b) Collaborating Centre: The National Institute of Public Health and Environmental Hygiene (RIVM), Bilthoven has accepted the task of becoming the coordinating centre and is currently preparing a background paper and developing the scope and purpose document based on a planning meeting. An initial organisation meeting, 11-14 December 1990, in section 5, will develop the programme in detail and participating Member States are encouraged to develop their national grids and issue related information systems.

c) National focal points: National institutions who will develop the national grids and issue related information system. In general each member state will develop a set of indices for health and environmental issues and risk factors. An attempt will then be made to agree on a minimum set which will be common to all participants and of use to WHO/EURO. Each state will also have its own set of indicators related to health and environmental issues of national importance.

d) International organisational: CEC has under the programme CORINE together with the European Statistical Office developed a complex multifaceted multilayer GIS for demographic purposes. WHO/EURO has been invited to participate and to help develop the health related data for the system. It is necessary to make use of these existing programmes as much as possible. It is estimated that the grid system must cover a minimum of 20-30 areas for most mediums-sized countries.

## 4. Timetable

**Preliminary Outline Workplan for Geographic Information System Project**

Design phase. During the period January to June 1991, RIVM will develop its position paper and identify potential participating groups in the Region. The paper will be circulated in draft form in September and consultation convened in December 1990 to organise the initial pilot phase. At this consultation the minimum needs for participation will be developed together with the concept of indicators, suitable for health outcomes related to the holistic definition of environmental risk factors also to be developed. It is expected that agreement on communication, data review, etc. be developed at the consultation. It is anticipated that it would take an additional six months for member states to develop their pilot programmes. The design of the pilot phase should be finished by June 1991.

Implementation of pilot phase. The programme has two parallel efforts - one at national level for each participating country and one at regional level as developed by the collaborating centre (RIVM). It is expected that several countries already have functioning GIS programmes, so that these could be available on demonstration basis from January 1991 while the other countries would need from 9 to 12 months to develop their pilot models. The pilot phase therefore extends over the period January 1991 to December 1992.

Overall programme. Development of the overall programme depends on the decisions within CEC to develop CORINE and the extent to which WHO/EURO will be called upon to provide health related data. It also depends on how rapidly the utility of a GIS for decision making can be demonstrated to non-participating nations. By the end of 1991 there should be a major activity running in final form based on CEC activity and those member states which will have responded rapidly after December 1990. Many of the countries in the region which currently contribute to Health and Environment indicators could immediately provide data in semi-aggregated form since it is already existent locally. The task of convincing them of the utility of the programme must fall to WHO/EURO.

## 5. The Bilthoven Consultation

As a result, the National Institute of Public Health and Environmental Protection (RIVM) in Bilthoven was asked to take the lead in assisting the Regional Office to develop a GIS programme for environment and health in the European Region. To this end, a planning meeting was convened at RIVM on 13-14 December 1989, to help verify the need for and define the scope of such a GIS System and to point to relevant information sources. This planning meeting recommended the convening in Bilthoven of the Consultation that is the subject of the report for the Development of a Health and Environment GIS for the European Region.

The main aim of the Bilthoven Consultation, held in December 1990, was to justify the need for and define the potential uses and specifications of a spatial information infrastructure (i.e. GIS and its supporting facilities) to support the research and policy implementation of the Regional Office. Towards this goal the Bilthoven Consultation was requested to: (a) verify the usefulness of GIS, as a permanent information management system with the focus on the state of public health and environmental quality, for the Regional Office and other European and national agencies dealing with public health and environmental policies and research; and (b) evaluate the technical feasibility of creating a GIS that contains both health and environmental data in a single relational (or object-oriented) database.

This led to a detailed set of guidelines, recommendations and conclusions on the future of a European approach to Health and Environment Geographical Information System.

Taken from;

WHO (1988), Summary Report, for the Consultation on Environmental Health Information Systems in the European Region, Berlin (West), 21-25 November 1988, mimeo

WHO (1990), Development of Health and Environment Geographical Information System for the European Region, (Target 19), Report on a WHO Consultation, Bilthoven, 10-12 December 1990

# APPENDIX B

## A LARGE VERSUS A SMALL SCALE ORGANISATION OF A HEGIS

| EUROPEAN APPROACH | | PILOT STUDY APPROACH | |
|---|---|---|---|
| **Advantages** | **Disadvantages** | **Advantages** | **Disadvantages** |
| Integration of datasets | Variations in priorities between Member states | Independence | Isolation |
| Increased data sharing | Difference between Member States in the ability to exploit GIS facilities | Control over<br>-priorities<br>-form and accessibility<br>-information<br>-technical specialists<br>-equipment<br>-overall objectives | Lack of support<br>-financially<br>-resources<br>-technical<br>-training |
| Improved access to information | Differences in the level of awareness and spatial data handling skills | Clear lines of responsibility | A possibility that the system will become too application specific |
| More operationally informed at the<br>-strategic<br>-managerial<br>-decision making level | Initially there may be disagreements over information<br>-access<br>-leadership<br>-data standards<br>-equipment<br>-training | Easier benchmarking of capabilities and benefits | Raises the question can these results be replicated? |
| Rigourous standards leading increased efficiency,<br>-avoiding duplication of<br>-time<br>-staff<br>-cost<br>-datasets | Harmonisation of databases may result in the loss of data in order to ensure comparability | Detailed documentation of data<br>-resolution<br>-problems<br>-operational decisions<br>-errors<br>-nature of surrogates | |

**Adapted from** Campbell H (1991) 'The impact of Geographic Information Systems on British Local Government'

# APPENDIX C

## Cancer types and diagnosis codes used by the Children's Malignant Disease Registry

**Leukaemia**

1 Acute Lymphoblastic Leukaemia
2 Acute Myeloid Leukaemia
3 Chronic Myeloid Leukaemia
4 Other Leukaemias

**Brain Tumours**

5 Craniopharyngioma
6 Cerebellar Astrocytoma
7 Brain Stem glioma
8 Optic nerve glioma
9 Ependymoma
10 Medulloblastoma
31 Tumour of pineal region
32 Presumed glioma
33 Oligodendroglioma
34 Meningioma
35 Ganglioglioma
36 Supratentorial Astrocytoma
11 Other brain tumours

**Kidney Tumours**

12 Wilms' tumour
13 Other kidney tumour
14 Neuroblastoma
15 Ganglioneuroblastoma

**Bone Tumours**

16 Ewing's Sarcoma
17 Osteosarcoma
18 Osteoclastoma
19 Other bone tumours

**Other**

20 Non-Hodgkin's disease
21 Hodgkin's disease
22 Hisitocytosis x
23 Yolk sac tumour
24 Hepatoblastoma
25 Retinoblastoma
26 Rhabdomyosarcoma
28 Soft tissue sarcoma
29 Epithelial/carcinomas

**Misc.**

27 Miscellaneous
38 Testicular tumours
39 Malignant melanoma

# APPENDIX D

## An Alphabetical List of Commonly used ARC Commands

| | |
|---|---|
| ADDITEM | adds a new item with no data to an INFO data file. |
| ADS | starts the ADS digitizing and editing program. |
| APPEND | combines up to 500 coverages into one coverage. |
| ARCEDIT | starts the ARCEDIT digitizing and editing program. |
| ARCPLOT | starts the ARCPLOT cartographic display program. |
| BATCH | submits commands to your system's batch queue. |
| BUFFER | creates buffer polygons around features in a coverage. |
| BUILD | creates or updates feature topology and attribute tables for a specified coverage feature type (point, line, or polygon). |
| CLEAN | creates or updates polygon or line topology for a coverage and splits arcs at arc intersections. |
| CLIP | clips a coverage to the polygons of another coverage. |
| COPY | makes a copy of a coverage. |
| COPYINFO | makes a copy of an INFO data file. |
| CREATE | creates an empty coverage using an existing coverage's tics. |
| CREATELABELS | creates label points for coverage polygons that do not have them. |
| DELETETIC | deletes selected tics from coverage. |
| DESCRIBE | provides information about the contents of a coverage and its processing status. |
| DIGITIZE | starts the ADS digitizing and editing program to digitize a new coverage. |
| DISPLAY | specifies the graphic display device to be used in an ARC/INFO session. |
| DISSOLVE | removes arc boundaries between adjacent polygons that have the same attribute values for a given item. |
| DRAW | draws a plot file on the specified graphic display device. |
| DROPITEM | deletes an item from an INFO data file. |
| EDGEMATCH | is used to match arcs along the adjacent sides of two coverages. |
| EDIT | starts the ADS digitizing and editing program to edit an existing coverage. |
| ELIMINATE | merges selected, adjacent polygons by eliminating the longest shared border arc between them. |
| ERASE | overlays a coverage with a polygon coverage to produce a new coverage. The new coverage has all the features from the first coverage, except those overlapped by the polygonsinthe second coverage. |
| EXPORT | converts a coverage, INFO data file, or text file into an ARC/INFO interchange file. |
| EXTERNAL | corrects the external pathnames for the INFO files of a coverage. |
| EXTERNALALL | corrects the external pathnames for the INFO files of all the coverages in the current workspace. |
| FREQUENCY | produces a list of the unique attribute codes for selected items in an INFO data file. |

| | |
|---|---|
| GENERALIZE | weeds coordinates along coverage arcs by removing vertices within a specified tolerance. |
| GENERATE | generates features and adds them to a coverage. |
| IDEDIT | updates coverage features to comply with changes made to the User-IDs in one of its feature attribute tables. |
| IDENTITY | overlays a coverage with a polygon coverage to produce a new coverage. All the features from the first coverage, and those polygons in the second coverage that overlap them, are i intersected to create the features for the new coverage. |
| IMPORT | converts an ARC/INFO interchange file into a coverage, INFO data file, or text file. |
| INFO | starts the INFO database system. |
| INTERSECT | overlays a coverage with a polygon coverage to produce a new coverage. Those features from the two coverages that overlap each other are intersected to create the features for the new coverage. |
| JOINITEM | merges two INFO data files based on a shared item. |
| KILL | deletes a coverage. |
| LABELERRORS | lists a coverage's polygon label errors. |
| LINEEDIT | starts the LINEEDIT program for designing cartographic line symbols. |
| LIST | lists item values for all records in the specified INFO data file. |
| MAPJOIN | appends up to 500 coverages into one coverage and creates topology for this coverage. |
| MARKEREDIT | starts the MARKEREDIT program for designing cartographic marker (point) symbols. |
| MATCHCOVER | copies attributes from one coverage to another for arcs that match in both coverages. |
| MATCHNODE | snaps nodes together when they fall within the specified snap tolerance. |
| NEAR | finds the closest arc, label point, or node from each point in another coverage. |
| NODEPOINT | creates a point coverage from another coverage's nodes. |
| POINTDISTANCE | computes the distance between label points in two coverages. |
| POSTSCRIPT | converts an ESRI plot file into PostScript file. |
| QUIT | quits the ARC session. |
| RELATE | establishes, modifies, or reports on the current ARC/INFO relate environment. |
| RENAME | changes the name of a coverage. |
| RESELECT | copies selected features from an existing coverage into a new coverage. |
| ROTATEPLOT | rotates an ESRI plot file by 90 degrees. |
| SHADEEDIT | starts the SHADEEDIT program for designing cartographic shade symbols. |
| SNAPCOVER | adjusts the location of features in a coverage to match the location of features in another coverage. |
| SPLIT | splits a coverage into a number of separate coverages. |
| STATISTICS | generates summary statistics for specified items in an INFO data file. |
| TABLES | starts the TABLES database program which emulates INFO. |
| TEXTEDIT | starts the TEXTEDIT program for designing cartographic text symbols. |

| | |
|---|---|
| TOLERANCE | sets or lists the processing tolerances associated with a coverage. |
| TRANSFORM | changes coverage coordinates using an affine or a projective transformation function based on control points (tics). |
| UNGENERATE | creates text files of a coverage's feature coordinates for arcs or label points in a GENERATE format. |
| UNION | overlays two polygon coverages to produce a new coverage. All the polygons from both coverages are intersected to create the features in the new coverage. |
| UPDATE | overlays the two polygon coverages to produce a new coverage. In the new coverage, the polygons from the first coverage are replaced by those in the second coverage where they overlap. |

# APPENDIX E

## Executing Poisson Probability

The following includes an annotated copy of the program written using the ARC/INFO Macro Language. This serves to access all the environmental coverages and the cancer cases from the relevant databases. Population at risk and cancer counts are derived and then the FORTRAN program is run to calculate the Poisson probability for these coverages in order to establish whether there is a relationship between ALL and the environment.

The value for each coverage and the associated sub-categories is returned to the relevant database ready for mapping in GIS.

This AML allowed a number of repetitive procedures to be carried out a minimum of 20 times for each coverage in the GIS application and then for any subsequent refinements upon the analysis, ie. a localised view and selections according to age/sex disaggregations.

(1) AML for carrying out Poisson probability
(2) The Poisson Probability Equation
(3) The FORTRAN program accessed by the AML(1)

# APPENDIX E

## (1) AML for carrying out Poisson probability calculations

```
&echo &on
&args .name .item .namea .itema
&watch poiwatch
/*
/* For a given coverage the following will calculate the
/* population at risk and the cancer counts for each
/* polygon or category of interest
/*
identity candod7686 %.name% %.name%dod point
identity edjun91b %.name% %.name%ed point
statistics %.name%dod.pat %.name%dod.sta %.item%
sum m04
sum m59
sum m1015
sum f04
sum f59
sum f1015
end
statistics %.name%ed.pat %.name%ed.sta %.item%
sum totm04
sum totm59
sum totm1015
sum totf04
sum totf59
sum totf1015
end
/*
joinitem %.name%ed.sta %.name%dod.sta %.name%freq.sta
%.item% %.item%
/*
/*  Creating the summary statistics for calculating
/*  Poisson probabilities
/*
&data ARC INFO
ARC
CALC $COMMA-SWITCH = -1
OUTPUT /MNT/USER1/USERS/ANNA/PHD/%.namea%PP.DAT
SEL %.namea%FREQ.STA
RESEL %.itema% GT 1
DISP %.itema%,SUM-TOTM04,SUM-TOTM59,SUM-TOTM1015,SUM
TOTF04,SUM-TOTF59, ~
 SUM-TOTF1015,SUM-M04,SUM-M59,SUM-M1015,SUM-F04,SUM-
F59,SUM-F1015 PRINT
```

```
CALC $COMMA-SWITCH = -1
OUTPUT /MNT/USER1/USERS/ANNA/PHD/%.namea%PP2.DAT
SEL %.namea%FREQ.STA
DISP %.itema%,SUM-TOTM04,SUM-TOTM59,SUM-TOTM1015,SUM-
TOTF04,SUM-TOTF59, ~
 SUM-TOTF1015,SUM-M04,SUM-M59,SUM-M1015,SUM-F04,SUM-
F59,SUM-F1015 PRINT

Q
STOP
&end
/*
/* Access the FORTRAN program that will calculate
/* Poisson probabilities for the selected coverage
/*
&system cp %.name%pp.dat poisson.dat
&system cp %.name%pp2.dat poisson2.dat
&system poisson
/*
/* Create a template to store the new data
/*
copyinfo
/mnt/user1/users/anna/phd/info/!arc!%.name%freq.sta
%.name%pp.sta
dropitem %.name%pp.sta %.name%pp.sta frequency
dropitem %.name%pp.sta %.name%pp.sta sum-totm04
dropitem %.name%pp.sta %.name%pp.sta sum-totf04
dropitem %.name%pp.sta %.name%pp.sta sum-totm59
dropitem %.name%pp.sta %.name%pp.sta sum-totf59
dropitem %.name%pp.sta %.name%pp.sta sum-totm1015
dropitem %.name%pp.sta %.name%pp.sta sum-totf1015
dropitem %.name%pp.sta %.name%pp.sta sum-m04
dropitem %.name%pp.sta %.name%pp.sta sum-f04
dropitem %.name%pp.sta %.name%pp.sta sum-m59
dropitem %.name%pp.sta %.name%pp.sta sum-f59
dropitem %.name%pp.sta %.name%pp.sta sum-m1015
dropitem %.name%pp.sta %.name%pp.sta sum-f1015
/*
additem %.name%pp.sta %.name%pp.sta pprob 4 12 f 8
%.item%
/*
/*
&data ARC INFO
ARC
SEL %.namea%PP.STA
PURGE
Y
ADD FROM /MNT/USER1/USERS/ANNA/PHD/OUTPOIS.DAT
Q
STOP
```

```
&end
/*
/* Put the results from this significance test
/* back into the relevant databases
/*
joinitem %.name%freq.sta %.name%pp.sta %.name%final.sta
%.item% %.item%
&data ARC INFO
ARC
CALC $COMMA-SWITCH = -1
OUTPUT /MNT/USER1/USERS/ANNA/PHD/%.namea%FINAL.DAT
SEL %.namea%FINAL.STA
DISP %.itema%,SUM-TOTM04,SUM-TOTM59,SUM-TOTM1015,SUM-
TOTF04,SUM-TOTF59, ~
  SUM-TOTF1015,SUM-M04,SUM-M59,SUM-M1015,SUM-F04,SUM-
F59,SUM-F1015,PPROB PRINT
Q
STOP
&end
/*
/*
&label try
joinitem %.name%.pat %.name%pp.sta %.name%.pat %.item%
%.item%
/*
/* Cleaning up file space
/*
kill %.name%dod all
kill %.name%ed all
&data ARC INFO
ARC
SEL %.namea%ED.STA
PURGE
Y
ERASE %.namea%ED.STA
Y
SEL %.namea%DOD.STA
PURGE
Y
ERASE %.namea%DOD.STA
Y
SEL %.namea%FREQ.STA
PURGE
Y
ERASE %.namea%FREQ.STA
Y
SEL %.namea%FINAL.STA
PURGE
Y
ERASE %.namea%FINAL.STA
```

```
Y
SEL %.namea%PP.STA
PURGE
Y
ERASE %.namea%PP.STA
Y
Q
STOP
&end
&run sys outpois.dat
&run sys poisson.dat
&run sys poisson2.dat
&system rm rm -i
&watch &off
&return
```

# APPENDIX E

## (2) The Poisson Probability equation

$$P(k) = \sum_{k=1}^{n-1} \frac{\lambda^k e^{-\lambda}}{k!}$$

Where, n is the number of the observed cases of Acute Lymphoblastic Leukaemia;

$\lambda$ is the expected number of cases;

e is the exponential constant;

P(k) is the probability that an area will contain k points.

The Poisson Probability is:

$$1 - P(k)$$

# APPENDIX E

## (3) The FORTRAN program: Poisson probability

```
      PROGRAM POISON
*
* Creating a program that will carry out Poisson
* probabilities between childhood cancer incidences
* (ie ALL) and various environmental correlates
* which have been assigned using the GIS software
*
*
* Defining the necessary items
*
       REAL*8     MEAN, EXPECT,AMEAN,CUMP,NEXTP,PROBEX,
     X            FINALP, FACT, F
       INTEGER    TOTPOP, TOTCAN, TYPE, TM04, TM59, TM1015,
     X            TF04, TF59,TF1015, CM04,CM59,CM1015,
     X            ACTUAL,J,K,CATPOP,CATCAN
       CHARACTER*20 F1,F2,F3
*
        F1 = 'poisson.dat'
        F2 = 'outpois.dat'
        F3 = 'poisson2.dat'
        OPEN(UNIT=1, FILE=F1, STATUS='OLD')
        OPEN(UNIT=2, FILE=F2, STATUS='UNKNOWN')
        OPEN(UNIT=3, FILE=F3, STATUS='OLD')
*
*
*       Calculating the overall mean (MEAN)
*
*
        TOTPOP = 0
        TOTCAN = 0
        N = 0
*
  50    CONTINUE
        READ(3, *, END=1000) TYPE,TM04,TM59,TM1015,
     X              TF04, TF59, TF1015
     X              CM04,CM59,CM1015,CF04,CF59,CF1015
*
        TOTPOP = TOTPOP + TM04 + TF04 + TM59 + TF59 +
                 TM1015 + TF1015
        TOTCAN = TOTCAN + CM04 + CF04 + CM59 + CF59 +
                 CM1015 + CF1015
        N = N + 1
        GOTO 50
*
```

```
 1000 MEAN = (1.0*TOTCAN) / TOTPOP
      WRITE(6,*) MEAN
*
      CLOSE(UNIT=3)
*
*
* Now we have the  overall mean we can begin to calculate
* the Poisson probabilities which represent what is
* actually going on in each of the attribute categories
*
*
*      POISSON PROBABILITY -- Anna Cross Style !!!!!!!!
*
      OPEN(UNIT=1, FILE=F1, STATUS='OLD')
*
  60  CONTINUE
      CATPOP = 0
      CATCAN = 0
      ACTUAL = 0
      EXPECT = 0
      CUMP = 0
      PROBEX = 0
      NEXTP = 0
      FINALP = 0
      J = 0
      K = 0 .
      F = 0
*
      READ(1, *, END=2000) TYPE,TM04,TM59,TM1015,
     X           TF04,TF59,TF1015,
     X           CM04,CM59,CM1015,CF04,CF59,CF1015
*
      CATCAN = CM04 + CF04 + CM59 + CF59 + CM1015 +
               CF1015
      CATPOP = TM04 + TF04 + TM59 + TF59 + TM1015 +
               TF1015
*
      ACTUAL = CATCAN
      EXPECT = MEAN * CATPOP
      AMEAN  = - EXPECT
*
*
*     Avoiding certain redundant calculations ?
*
*
      IF(ACTUAL .EQ. 0) THEN
       WRITE(2, 300) TYPE
 300   FORMAT(I2, ',   0')
       GOTO 60
      END IF
```

```
       IF(CATPOP .EQ. 0) THEN
        WRITE(2, 400) TYPE
 400    FORMAT(I2, ',   0')
        GOTO 60
       ENDIF
*
       PROBEX = DEXP(AMEAN)
       CUMP = PROBEX
       K = (ACTUAL - 1)
       DO 800 J=1, K
          NEXTP = ((EXPECT ** J) * PROBEX)
          F = FACT(J)
          NEXTP = NEXTP / F
          CUMP = CUMP + NEXTP
 800   CONTINUE
*
       FINALP = 1 - CUMP
       WRITE(2, 900) TYPE, FINALP
 900   FORMAT(I2, ',', 2X, F10.8)
       GOTO 60
*
*
2000   STOP
       END
*
*        A SUB ROUTINE TO CALCULATE FACTORIALS
*
FUNCTION FACT(J)
      REAL*8 FACT,I,QFACT
      FACT = 1
      IF (J .LT. 2) THEN
       RETURN
      ELSE
       QFACT = 1
       DO 100 I = 2, J
 100    QFACT = QFACT*I
      END IF
      FACT = QFACT
      RETURN
      END
```

# APPENDIX F

# AN ALPHABETICAL LIST OF COMMONLY USED AML DIRECTIVES AND FUNCTIONS

## AML directives

| | |
|---|---|
| &ARGS | allows on AML file to capture arguments from the |
| &RUN | directive which invoked the program. |
| &DATA | submits a command to the operating system along with an input data file. |
| &DO | delimits a block of statements or directives to be executed one or more times. Variants include: &DO &UNTIL, &DO &WHILE. |
| &GOTO | causes control to be passed to the statement following the specified label directive line. |
| &IFTHEN&ELSE | allows statements to be executed conditionally. |
| &LABEL | marks the location in an AML program referenced by an &GOTO directive. |
| &POPUP | displays a scrollable text file at the terminal. |
| &RETURN | terminates an AML file or the current input source; or, if encountered in a routine block, returns control to the statement following the &CALL directive that invoked the routine. |
| &RUN | executes the specified AML file. |
| &STATION | defines the workstation environment. |
| &SYSTEM | initiates a dialog with the operating system or executes a specified operating system command. |
| &TYPE | sends the specified message to the terminal. |
| &WATCH | enables and disables a watch file. |
| [CALC] | returns the result of the calculation of an ARC expression. |
| [CLOSE] | closes the file opened on the AML file-unit. |
| [DELETE] | deletes a system file, directory or INFO file. |
| [EXISTS] | determines whether the given object (file, coverage, workspace, etc.) exists. |
| [READ] | reads a record from the file opened on the specified AML file-unit. |
| [SQRT] | returns the square root of the given variable. |
| [TRANSLATE] | translates one specified string into another in a target string. |
| [TYPE] | returns a code indicating the type specification of a string. |
| [VALUE] | returns the contents of the given variable. |
| [VARIABLE] | indicates if the given AML variable exists. |
| [WRITE] | writes a record to the file opened on the specified AML file-unit. |

# APPENDIX G

## Cluster analysis: GIS Style!

The following includes the AML for accessing the cancer database and the procedures which are designed to replicate the Besag and Newell Nearest Neighbour Method for cluster analysis of rare diseases, referred to in Chapter 8.

This could be simplified, given that the number of cancers which would form a cluster threshold was known ie. 5, and the desired significance level for the Poisson probability was also set, ie. 0.05. Thus, with only one unknown in the Poisson probability equation it was possible to determine a critical population threshold, using the bisection method. A FORTRAN program was written to carry out the latter and a critical population value was stored as a variable in the AML. All subsequent populations which were established for the circles in this analysis were then compared to this. The idea being that if the circle population is less than the critical population this suggests that the presence of 5 cancers in that area is significant and the resultant circle is retained for mapping.

This program could be generic by simply setting up a template where the end-user would be asked to specify the coverage to be analysed for clusters, the cluster threshold and the desired significance level for the analysis procedure. These would be stored as variables in the AML and would be used appropriately in the various sections of the program.

(1) AML to replicate the Besag and Newell Nearest Neighbour Method
(2) Description of the Bisection Method
(3) The FORTRAN program which carried out the necessary recursive division to determine the critical population

# APPENDIX G

## (1) AML to replicate Besag and Newell's Nearest Neighbour Method

```
&echo &on
&goto loop
/*
/* Calculating the distances between cancers and their
/*  5th nearest neighbour, this is done once at the
/*  start of the program
/*
pointdistance bescancer bescancerd pointd.dis
&s .critical = 0
&data arc info
ARC
SEL POINTD.DIS
SORT ON DISTANCE
SORT ON BESCANCER#
DEFINE POINTD.IDX
BESCANCERD#,4,6,B,0
DISTANCE,4,12,F,3

CALC $COMMA-SWITCH = -1
OUTPUT /MNT/USER1/USERS/ANNA/PHD/ALLCAN.DAT
SEL EDJUN91B.PAT
DISP TOTM04,TOTM59,TOTM1015,TOTF04,TOTM59,TOTF1015 PRINT

DEFINE POINTD.IDX
BESCANCER#,4,6,B,0
DISTANCE,4,12,f,3

Q STOP
&end
&system mean
&label loop
/*
/*  Output the 5th nearest neighbour value of interest
/*
/*  Start up a loop which takes each cancer
/*  separately and calculates the population at
/*  risk for the resultant circle this  will be compared
/*  to the critical value.
/*  At the moment, this is worked out by a FORTRAN
/*  program but it could be implemented in the form
/*  of another AML
/*
&s .incr = 0
&s .val = 173
```

```
&s .last = -167
/*
/* This determines the number of points in the file ie
/* 225 and this will act as the increment factor to
/* deduce the 5nn in the file created by POINTDISTANCE
/*
&s .no [calc %.val% * %.val%]
&do &until %.last% = 1044
&run sys next
 &s .incr = %.incr% + 1
 &s .last = %.last% + %.val%
 &data arc info
ARC
CALC $COMMA-SWITCH = -1
OUTPUT /MNT/USER1/USERS/ANNA/PHD/NEXT
SEL POINTD.DIS
RESEL $RECNO = %.last%
DISP BESCANCER#,',',DISTANCE PRINT
QUIT
STOP
&end
/*
/* Start to generate circles based on the radius
/* derived from the latter
/*
&data arc info
ARC
SEL POINTD.IDX
ADD FROM /MNT/USER1/USERS/ANNA/PHD/NEXT
QUIT
STOP
&end
/*
/* Buffer the relevant cancer case using this radius
/* Once the population at risk and number of
/* cancers are known you can either calculate the
/* Poisson probability for the circle or
/* alternatively use the critical population to
/* reject or accept this circle.
/*
joinitem bescancer.pat pointd.idx bescancer.pat
bescancer# bescancer#
buffer bescancer bescirc%.incr% distance # # # point
clip edjun91b bescirc%.incr% besinfo%.incr% point
dropitem bescancer.pat bescancer.pat distance
/*
&data arc info
ARC
SEL POINTD.IDX
PURGE
```

```
Y
Q STOP
&end
/*
/* THE CRITICAL VALUE APPROACH
/*
&label critical
additem besinfo%.incr%.pat besinfo%.incr%.pat totpop 4 8
b 0
additem besinfo%.incr%.pat besinfo%.incr%.pat link 4 5 b
0
&data ARC INFO
ARC
SEL BESINFO%.incr%.PAT
CALC TOTPOP = TOTM04 + TOTF04 + TOTM59 + TOTF59 +
TOTM1015 + TOTF1015
CALC LINK = 1
Q
STOP
&end
statistics besinfo%.incr%.pat besinfo.sta link
sum totpop
end
&data ARC INFO
ARC
CALC $COMMA-SWITCH = -1
OUTPUT /MNT/USER1/USERS/ANNA/PHD/BESINFO.DAT
SEL BESINFO.STA
DISP SUM-TOTPOP PRINT
SEL BESINFO.STA
PURGE
Y
ERASE BESINFO.STA
Y
Q
STOP
&end
/*
/* Compare the circle population with the critical
/* value returned from the FORTRAN program. If less
/* than this value save the circle and repeat the
/* process, if not discard and continue
/*
&label mistake
&if [unquote %.critical% ne 0] &then
&goto diff
&else
&s fileunit := [open critical.mu status -r]
&s mu := [read [unquote %fileunit% readstat]]
&s fileunit3 := [open mean.sta status -r]
```

```
&s meanc := [read [unquote %fileunit3% readstat]]
&s fileunit2 := [open besinfo.dat status -r]
&s popn := [read [unquote %fileunit2% readstat]]
&s closestat := [close %fileunit%]
&s closestat := [close %fileunit2%]
&s closestat := [close %fileunit3%]
&s .critical = [calc [unquote %mu% / %meanc%]]
&goto compare
/*
&label diff
&s popn = 0
&s fileunit2 := [open besinfo.dat status -r]
&s popn := [read [unquote %fileunit% readstat]]
&s closestat := [close %fileunit2%]
/*
&label compare
&if [unquote %popn%] gt [unquote %.critical%] &then
&pause
 kill besinfo%.incr% all
 kill bescirc%.incr% all
 &run sys besinfo.dat
 &s .incr = %.incr% - 1
 &end
&else
 &end
&end
&return
```

# APPENDIX G

## (2) The Bisection Method

The *bisection method* for solving the equation $f(x) = 0$ for the values of x (the *roots)* is based on the following theorem.

**Theorem:** If $f(x)$ is continuous for x between a and b and if $f(a)$ and $f(b)$ have opposite signs, then there exists at least one real root of $f(x) = 0$ between a and b.

The *procedure* is as follows: Suppose that a continuous function $f(x)$ is negative at $x = a$ and positive at $x = b$, so that there is at least one real root between a and b. (Usually a and b are found by curve sketching.) If we calculate $f[(a+b)/2]$, which is the function value at point of bisection of the interval $a < x < b$, there are three possibilities:

1) $f[(a+b)/2] = 0$ in which case $(a+b)/2$ is the root;
2) $f[(a+b)/2] < 0$ in which case the root lies between $(a+b)/2$ and b;
3) $f[(a+b)/2] > 0$ in which case the root lies between $(a+b)/2$ and a.

Presuming there is just one root, if case (1) occurs the process is terminated. If either case (2) or case (3) occurs, the process of bisection of the interval containing the root can be repeated until the root is obtained to the desired accuracy.

Source: Hosking, Joyce and Turner (1981).

# APPENDIX G

## (3) The FORTRAN Program: Calculating the critical population

```
      PROGRAM RECUR
*
*
*     This is a FORTRAN program to find the critical mu
*     for the Besag and Newell Nearest Neighbour Method,
*     using recursive division
*
*      The minimum critical value can not be less than
*      0 in this case, because negative cancers are not
*      recorded and the maximum value in this run was
*      set at 5.
*
*      Therefore a = 0 and b = 5
*
*
      REAL*8  A,B,C
      REAL*8  FUNCA,FUNCC
*
*     OPEN(UNIT=1, FILE='figure.mu',STATUS='UNKNOWN')
      OPEN(UNIT=2, FILE='critical.mu', STATUS='UNKNOWN')
*
      A =0.0D0
      B = 5.0D0
      C = 0.0D0
*
* This carries out recursive division until a stable
* number is reached, this constitutes the optimum
* number and will act as mu in the equation to
* deduce the critical population for a cancer count
* of 5, significant at the 0.05 level
*
  100 IF(DABS(B - A) .GT. 1.0E-8) GOTO 200
      WRITE(2,*) A
      GOTO 300
*
*
  200 COMP = 0.0D0
      FUNCC=0.0D0
      FUNCA=0.0D0
*
      C = (A + B) / 2
*     WRITE(6,*) C
*
```

```
        FUNCA = BTMU(A)
        FUNCC = UPMU(C)
*       WRITE (6,*) FUNCA,FUNCC
        COMP = FUNCA * FUNCC
*       WRITE(1,*) COMP
        IF(COMP .GT. 0) THEN
         A = C
        GOTO 100
        ENDIF
        IF(COMP .LT. 0) THEN
         B = C
        GOTO 100
        ENDIF
*
*
  300   STOP
        END
*
*
*           FUNCTION TO CALCULATE MU AND BETTER MU  FOR A
*
*
      FUNCTION BTMU(A)
      REAL*8 BTMU,A
*
      BTMU = 1 + A + (A**2/2) + (A**3/6) + (A**4)/24
      BTMU = (DEXP(-A)*BTMU) - 0.95
      RETURN
      END
*
*           FUNCTION TO CALCULATE MU AND BETTER MU  FOR C
*
*
      FUNCTION UPMU(C)
      REAL*8 UPMU,C
*
      UPMU = 1 + C + (C**2/2) + (C**3/6) + (C**4)/24
      UPMU = (DEXP(-C)*UPMU) - 0.95
      RETURN
      END
```

# APPENDIX H

## Using GLIM to perform Log-linear modelling

This appendix includes an example of the GLIM program written to access the datafile created in ARC/INFO. This basically demonstrates the simplicity of using statistical packages such as GLIM for developing models to observe possible relationships between environmental databases and ALL.

(1) The GLIM program

# APPENDIX H

## (1) The GLIM program for running Log-linear modelling

```
$R *GLIM  6=-RESULTS
*
* The data sample is accessed
*
$UNITS 1418$
$DATA POP CAN
      V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15
V16$
$READ
$CONTINUE WITH GLIM.DAT RETURN
$
$ERROR B POP$
$YVAR CAN$
*
* These represent the variables noted in Table 9.2
* of Chapter 9, which were created for this exercise
*
$FACTOR V1 4 V2 4 V3 11 V4 4 V5 4 V6 4 V7 4 V8 4 V9 4 V10
4
      V11 4 V12 4 V13 9 V14 10 V15 14 V16 4$
*
* A sample of the models used to look for possible
* interactions between coverages and the incidence
* of ALL are shown below
*
$ECHO$
$FIT$
$FIT V3$
$FIT V4$
$FIT V10$
$FIT V3*V10$
$FIT V4+V10$
$FIT V4*V10$
$STOP$
```

# APPENDIX I

## Steps to performing simple error modelling

This includes the AML and FORTRAN program set up to carry out a simple robustness test on the possible effect of error in a line and the effect of the 100 metre resolution of postcoded data.

The AML accesses the original grid references for the cancer database. These are then wobbled within 50 metres of their original position by generating random numbers using the FORTRAN program. The new set of grid references returned by this procedure are converted into a separate coverage and the AML performs a point-in-polygon operation with the ward coverage buffered at 100, metres this represents the possible error involved in map compilation and subsequent digitisation. The number of points that fall in this buffered region are referred to as 'maybes' and these reflect the possibility that there is some positional error in the data, and this could lead to errors in future analysis. For instance with rare diseases the positioning of one case in the wrong ward may make the difference between a ward being significant or not under the Poisson probability. This procedure was repeated 100 times to establish the spread of possible occurrences.

(1) AML to carry out the point-in-polygon procedure and the necessary summary statistics

(2) FORTRAN program for 'wobbling' the points

# APPENDIX I

## (1) AML for carrying out simple error handling

```
&echo &on
&s fileunit := [open randval.dat status -r]
&s .random := [read [unquote %fileunit% readstat]]
&s closestat := [close %fileunit%]
/*
/* Take the raw x- and y- coordinates for the cancer
/* data, these will be accessed by the FORTRAN program
/* which will create a new coverage to reflect the
/* possible effect of the 100m resolution
/* of postcoded data
/*
&s .no = 0
&s .last = 100
&do &until %.no% = %.last%
&s .no = %.no% + 1
&s .random = %.random% + 234
&s value = %.random%
&s fileunit2 := [open randnew.dat status -w]
&s writestat := [write %fileunit2% [unquote %value%]]
&s closestat := [close %fileunit2%]
&system random
/*
&run sys randnew.dat
generate cancerwob
input cancerwob.dat
points
q
/*
/* This is the new coverage, the original points have
/* been randomly 'wobbled' plus or minus 50m. This
/* will now be used to carry out another point-in-polygon
/* with the ward boundaries buffered at 100m
/*

build cancerwob point
identity cancerwob wards100m cancererr point
/*
statistics cancererr.pat cancererr.sta inside
sum cancererr#
end
/*
additem cancererr.sta cancererr.sta max 4 5 b 0
additem cancererr.sta cancererr.sta min 4 5 b 0
/*
```

```
&data ARC INFO
ARC
SEL CANCERERR.STA
RESEL INSIDE = 100
CALC MIN = FREQUENCY
CALC MAX = 94 - FREQUENCY
CALC PERCENT = MIN / MAX
CALC PERCENT = PERCENT * 100
ASEL
CALC $COMMA-SWITCH = -1
OUTPUT /MNT/USER1/USERS/ANNA/PHD/TEMP/CANCERTP.DAT
SEL CANCERERR.STA
RESEL INSIDE = 100
DISP MAX, MIN PRINT
ASEL

Q
STOP
&end
/*
/*
kill cancererr all
kill cancerwob all
&system variance
/*
&end
&return
```

# APPENDIX I

## (2) The FORTRAN program: 'Wobbles' cancer points.

```
      PROGRAM RANDOM
*
*     Converting cancer points to take into account
*     the possible effect of the 100m resolution of
*     postcoded data
*
      INTEGER ID
      REAL X,XA,Y,YB
*
      OPEN(UNIT=1, FILE='cancerraw', STATUS='OLD')
      OPEN(UNIT=2, FILE='cancerwob.dat',
                                    STATUS='UNKNOWN')
      OPEN(UNIT=3, FILE='randnew.dat', STATUS='OLD')
*
      READ(3,*) VALUE
      X = RAND(VALUE)
*
  50  CONTINUE
      READ(1,*, END=1000) ID,X,Y
      XA = X + (RAND(0) - 0.5)
      YB = Y + (RAND(0) - 0.5)
      WRITE(2,*)ID,XA,YB
      GOTO 50
*
 1000 WRITE(2,*) 'END'
      STOP
      END
```

# APPENDIX J

## Glossary of Acronyms used in the thesis

| | |
|---|---|
| AAT | Arc Attribute Table |
| ADS | Arc Digitising System |
| ALL | Acute Lymphoblastic Leukaemia |
| AML | Arc Macro Language |
| ASCII | American Standard Code for Information Interchange |
| BGS | British Geological Survey |
| CAD | Computer-Aided Design |
| CDMS | Credit and Data Marketing Services Limited |
| CEC | Commission of the European Communities |
| CEGB | Central Electricity Generating Board |
| CGIS | Canada Geographic Information System |
| CHEST | Committee of Higher Education Software Team |
| CORINE | Coordination of Environmental Information |
| CPU | Central Processing Unit |
| DOE | Department of the Environment |
| ED | Enumeration District |
| EGIS | European Geographical Information System |
| ESRC | Economic and Social Research Council |
| ESRI | Environmental Systems Research Institute |
| GAM | Geographical Analysis Machine |
| GCEM | Geographical Correlates Exploration Machine |
| GIS | Geographical Information System |
| GLIM | General Linear Interactive Modelling |
| HEGIS | Health and Environment Geographic Information System |
| HMIP | Her Majesty's Inspectorate of Pollution |
| HMSO | Her Majesty's Stationary Office |
| IT | Information Technology |
| K | Kilovolts |
| M | Metres |
| NEEB | North Eastern Electricity Board |
| NHL | Non-Hodgkins' Lymphoma |
| NRPB | National Radiological Protection Board |
| OS | Ordnance Survey |
| OPCS | Office of Population Census and Surveys |
| PAT | Point Attribute Table or Polygon Attribute Table |
| RRL | Regional Research Laboratory |
| SAHSU | Small Area Health Statistics Unit |
| SPANS | Spatial Analysis Systems |
| TIN | Triangular Irregular Network |
| UK | United Kingdom |
| WHO | World Health Organisation |

# BIBLIOGRAPHY


ABLER, R.F., (1987), 'The National Science Foundation, National Center for Geographic Information and Analysis', *Int. J.GIS* , **1**(4), pp 303-326.

ALDERSON, M. (1983), *An Introduction to Epidemiology*, Second Edition. Macmillan, London.

ALEKSANDER, I., (1990), *Neural Computing Architectures*, North Oxford Academic, London.

ALEXANDER F.E., CARTWRIGHT R.A., McKINNEY P.A. and RICKETTS T.J. (1990) 'Leukaemia incidence, social class and estuaries : an ecological analysis', *Journal of Public Health Medicine*, **12**(2), pp 109-117.

ALEXANDER F.E., RICKETTS T.J., WILLIAMS S.J. & CARTWRIGHT R.A. (1991) 'Method of Mapping and Identifying small Clusters of Rare Diseases with Applications to Geographical Epidemiology, *Geographical Analysis*, **23**(2), pp 159-173.

ASPINWALL & CO LTD, (1987), *Sitefile: A Digest of Authorised Waste Treatment & Disposal Sites in Great Britain.* Printed by Inprint.

BARTHOLOMEW, J. & Co., (1990), *GREAT BRITAIN DATA*, mimeo.

BAXTER R., (1976), *Computer and Statistical Techniques for Planners*, Methuen, London.

BERRY JK. (1987) 'Computer - Assisted Map Analysis: Potential and Pitfalls' *Photogrammetric Engineering and Remote Sensing* **53**(10), pp 1403-1410.

BESAG J. & NEWELL J. (1991) 'The Detection of Clusters in Rare Diseases' *J.R. Statist Soc. A* **154** part 1, pp 143-155.

BLACK, SIR. D. (1984) *'Investigation of the possible increased incidence of cancer in West Cumbria'*: Report of the Independent Advisory Group. HMSO, London.

BLAKEMORE, M. (1984) 'Generalisation and Error in Spatial Databases', *Cartographica* - auto carto six selected papers ed. by Douglas D.N. vol 21, pp 131-139.

BLAKEMORE, M., (1990) 'Sharing Data - Whose, why, how?' presented at the Association for Geographic Information National Conference, Brighton.

BRADLEY E.J. and GREEN B.M.R.(1984), 'Outdoor gamma - ray dose rates in Great Britain, preliminary results', *Radiological Protection Bulletin*, No 59, NRPB, Chilton.

BRANDT L., NILSSON P.G. and MITELMAN T., (1978) 'Occupational exposure to petroleum products in men with acute non-lymphocytic leukaemia' *Lancet* i,. p 553.

BRASSEL K., (1984) 'Manipulation Processes in Computer Cartography', *Basic Readings in Geographic Information Systems*, ed Marble D.F., Calkins H.W. and Peuquet D.J. SPAD sys Ltd.

BRIGGS D., (1991) 'GIS development for broad scale policy applications: The lessons from CORINE', *Association for Geographic Information Yearbook* ed. Heywood DI and Cadoux-Hudson J., pp 113-120.

BROWNING D. and GROSS S., (1968) 'Epidemiological studies of acute childhood leukaemia', *Am.J.Dis. Child.* **116** pp 576.

BRUNSDON C., CARVER S., CHARLTON M., and OPENSHAW S., (1990) 'A Review of Methods for Handling Error Propagation in GIS' Proc. for the First European Conference in GIS, Amsterdam, The Netherlands, April 1990.

BURROUGH, P.A., (1986) *Principles of Geographical Information Systems for Land Resources Assessment.* Oxford University Press, Oxford.

CALKINS, H.W.,(1984) ' A Pragmatic Approach to Geographic Information System Design'. *Basic Readings in Geographic Information Systems* ed. Marble D.F., Calkins I.W. and Peuquet D.J. SPAD Sys Ltd.

CAMERON, G.A.,(1984) 'Manual Digitizing Systems' *Basic Readings in Geographic Information Systems* ed. Marble D.F., Calkins H.W. and Peuquet D.J. SPAD Sys Ltd.

CAMPBELL H., (1991) 'The Impact of Geographic Information Systems on British Local Government' presented at the UDMS Conference.

CARVER S., (1991) 'Application of Geographic Information Systems to siting Radioactive Waste Disposal Facilities', Unpublished Thesis, Department of Geography, University of Newcastle upon Tyne.

CENTRAL ELECTRICITY GENERATING BOARD (1988) 'Power Points - Electric and magnetic fields - your questions answered'. Working Paper, CEGB, London.

CENTRAL ELECTRICITY GENERATING BOARD (1988), *Radiation Risks : In Perspective*, CEGB, London.

CHARLTON, M., OPENSHAW, S. and WYMER, C., (1985) 'Some new classifications of census enumeration districts in Britain: a poor man's ACORN' *Journal of Economic and Social Measurement* **13**, pp 69-96.

CHORLEY, R. - (1987), *Handling Geographic Information*, Report to the Secretary of State for the Environment of the Committee of Enquiry into the Handling of Geographic Information, HMSO, London.

CHRISMAN N.R., (1982) Methods of Spatial Analysis Based on Error in Categorical Maps, Unpublished Thesis University of Bristol, May 1982.

CLIFF, A.D. and HAGGETT, P., (1982) *Atlas of Disease Distribution, analytic approaches of epidemiological data*. Basil Blackwell Ltd, Oxford.

COMMISSION OF THE EUROPEAN COMMUNITIES, (1985) Council Decision of 27 June on the adoption of the Commission work programme concerning an experimental project for gathering, co-ordinating and ensuring the consistency of information on the state of the environment and natural resources in the Community (85/338/EEC). *Official Journal of the European Communities*, L176, pp 14-17.

COOK-MOZAFFARI, ASHWOOD F.L., VINCENT T., FORMAN D., ALDERSON M., (1987), Cancer incidence and mortality in the vicinity of nuclear installations, England and Wales 1959-80, for Office of Population Census and Surveys : *Studies on Medical and Population Subjects* No 51. HMSO, London.

COPPOCK T.J. and ANDERSON E.K. (1987), Editorial Review, *Int. J. GIS*, 1(1), pp 3-11.

COULTER, E., (1987), 'Mapping the Future', *Computer Weekly* Oct 15th pp 39-40.

COWEN, D.J., (1988) 'GIS versus CAD versus DBMS: what are the differences?', *Photogrammetric Engineering and Remote Sensing* 54, pp 1551-1554.

CRAFT, A.W., OPENSHAW, S. and BIRCH, J.M., (1985), 'Childhood cancer in the Northern Region, 1968-82 : incidences in small geographical areas', *Journal of Epidemiology and Community Health*, pp 53-7.

CROSS A.E.,(1990) 'Using a Geographical Information System to Explore the spatial incidence of childhood cancer in Northern England', Proc. for the First European Conference in GIS, Amsterdam The Netherlands, April 1990.

CROSS, A.E. and OPENSHAW S., (1990), 'Searching for Relationships between Map Coverages', Proc. for the Second National Conference on GIS. Ottawa, Canada, March 1990.

CROSS, A.E., (1991) 'Working for a safer society' in *Association for Geographic Information, Yearbook* ed. Heywood D. and Cadoux-Hudson J.

DANGERMOND, J., (1984), 'A Classification of Software Components Commonly Used in Geographic Information Systems', *Basic Readings in Geographic Information Systems.* ed. Marble D, Calkins, H.W. and Peuquet, D.J. SPAD Sys Ltd.

DARBY, S.C. and DOLL, R., (1987), 'Fallout, radiation doses near Dounraey, and childhood leukaemia', *Br. Med. J.*, **294**, March pp 603-607.

DENHAM, C. and RHIND, D., (1983), *'The 1981 Census and its results', A Census User's Handbook*, ed. Rhind D., Methuen, London. pp 17-89.

DEPARTMENT OF THE ENVIRONMENT (1985) *'Specific problem areas in England and Wales'*, The Hazardous Waste Inspectorate, First Report, Hazardous Management: An Overview HMSO, London.

DEPARTMENT OF THE ENVIRONMENT (1988), *Radioactive Substances Act 1960, Disposal of Radioactive Waste.* HMSO, London.

DEPARTMENT OF THE ENVIRONMENT (1988), *The Hazardous Waste Inspectorate. Third Report*, HMSO, London.

DEPARTMENT OF THE ENVIRONMENT (1988), *Handling Geographic Information. The Governments' Response to the Report of the Committee of Enquiry chaired by Lord Chorley* HMSO, London.

DEPARTMENT OF THE ENVIRONMENT, (1991) Third Annual Report 1989-90, Her Majesty's Inspectorate of Pollution, HMSO, London.

DOBSON, M., (1984), 'Effective Color Display for Map Task Performance in a Computer Environment', *Proc in the Symp. of Spatial Data Handling 2*, Zurich, Switzerland, Geographicshie Institut, pp 332-348.

DOLL, F.R. and PETO, R., (1981), *The Causes of Cancer: Quantitative Estimates of Available Risks of Cancer in the United States Today.* Oxford University Press, Oxford, pp1245-1255.

DOLL, R. (1989) 'The epidemiology of childhood leukaemia', *The Journal of the Royal Statistical Society*, series 152, pp341-351.

EDWARDSON, J.A., (1988), 'Environmental Factors and the aetiology of Alzheimer's disease', *Current opinion in Psychiatry 1*, 458-461.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE, (1981), *Arc Users Manual*, ESRI Inc, Redlands.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE, (1990), PC *Understanding GIS: The ARC/INFO Method* ESRI Inc. Redlands.

EVANS, I.A. and WIDDOP, B. (1966) 'Carcinogenic activity of bracken', *British Empire Cancer Campaign for Research, Annual Report* **377**.

EVANS, I.A., JONES, R.S. and MAINWARING-BURTON, R. (1972), 'Passage of bracken fern toxicity into milk', *Nature*, **237**, pp 107-8.

EVANS, I.A. (1976), 'Relationship between bracken and cancer', *Botanical Journal of the Linnean Society* **73**, pp 105-112.

FOSTER R. Dr, (1980) 'Tracking the micropollutant', *Water*, May, 1980.

FRANK, A.U., (1984), 'Requirements for Database Systems suitable to manage large spatial databases', *Proc of the International Symp. on Spatial Data Handling, 1*, Zurich, Switzerland. Geographische Institut pp 38-60.

FREEMAN, H. and AHN J., (1984), 'Automap - an Expert System for Automatic Map Name Placement', *Proceedings of the International Symposium on Spatial Data Handling 1*, Zurich, Switzerland. Geographische Institut.

FROLOU, Y.S. and MALING, D.H., (1969), 'The Accuracy of Area Measurement by Point Counting Techniques', *Cartographic Journal* **6**, pp 21-35.

FULTON, J.P., COBB, S., PREBLE, L., LEONE, L. and FORMAN, E. (1980) 'Electrical wiring configurations and childhood leukemia in Rhode Island', *Am. J. Epidemiol.* **111**, pp 292-296.

GARDNER, M.J., SNEE, M.P., HALL, A.J., POWELL, C.H., DOWNES, S., and TERRELL, J.D. (1990) 'Results of a case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant West Cumbria', *British Medical Journal* **300**, pp 423-429.

GARDNER M.J., WINTER P.D. and BARKER D.J.P., (1984), *Atlas of Mortality from selected Diseases, in England and Wales, 1968-1978*, Wiley, Chichester.

GARDNER, M. (1990), 'Leukaemia and nuclear installations : Occupational exposure of fathers to radiation may be the explanation', *British Medical Journal* **300**, pp 411-412.

GATRELL, A.C. and LOVETT, A.A., (1989), 'Burning Questions : Incineration of Waste and Implications for human health' presented at the Institute of British Geographers Annual Conference, Coventry Polytechnic.

GATRELL, A.C., (1989), 'On the spatial representation and accuracy of address-based data in the United Kingdom', *Int. J. GIS*, 3(4), pp335-348.

GOODCHILD, M.F.,(1987) 'A Spatial Analytical perspective in geographical information systems', *Int.J.GIS*, 1(4) pp 327-334.

GOODCHILD, M. and GOPAL, S., (1989), *Accuracy of Spatial Databases*, Taylor and Francis, London.

GOODCHILD, M., (1991), 'Towards a science of Geographic Information', *Association for Geographic Information*, Yearbook ed Heywood D. and Cadoux-Hudson J.

GREAVES, M.F. and CHAN, L.C. (1986), 'Annotation : Is spontaneous mutation the major 'cause' of childhood acute lymphoblastic leukaemia?', *British Journal of Haematology* **64**, pp 1-13.

GREEN B.M.R., LOMAS P.R. and O'RIORDAN M.C., (1987) 'Action on Radon in UK Homes', presented at the Fourth International Symposium on the Natural Radiation Environment, Lisbon, Portugal 7-11 Dec 1987.

GREEN, B.M.R., LOMAS P.R., BRADLEY E.J. and WRIXON A.D., (1987) 'Gamma-radiation levels outdoors in Great Britain' - NRPB RI91, NRPB, London.

GREEN, N. and RHIND D., (1986), 'Teach Yourself GIS', *Proceedings in Auto Carto*, London 12, pp 327-39.

GREEN, N., (1987) 'Database Design and Implementation'. *SEERL Working Report No 2*, South East Regional Research Laboratory, Birbeck College.

HARRIS, P.M., HIGHLEY, D.E. and BENTLEY, K.R. (1988), *Directory of Mines and Quarries 1988, British Geological Survey* HMSO, London.

HAVILAND, A., (1855), *The Geographical Distribution of Diseases in Great Britain*, Smith Elder, London.

HEALY, M.J.R., (1988), *Glim: An Introduction*, Clarendon Press, Oxford.

HENSHAW, D.L., EAROUGH J.P. and RICHARDSON, R.B., (1990) 'Radon as a causative factor in induction of myeloid leukaemia and other cancers', *Lancet* **335**, pp 1008-1012.

HIGGINSON, J., (1975), APHA Awards for Experience, Rosenhaus Lecture.

HILL, M.D., & COOPER J.R., (1986) 'Radiation doses to members of the population of Thurso', *NRPB report R195* HMSO, London.

HILLS, M. AND ALEXANDER, F., (1989), 'Statistical Methods used in assessing the risk of disease near a source of possible environmental pollution : a review', *J.R. Statist. Soc. A.* **152** pp 353-363.

HMSO, (1985) 'Fluoridation of Water and Cancer. A Review of the Epidemiological Evidence', *Report of the Working Party.* Chairman Professor E.G. Know.

HMSO, (1986) 'Committee on Medical Aspects of Radiation in the Environment' *COMARE First Report*, HMSO, London.

HOSKINS, R.J., JOYCE, D.C., & TURNER, J.C. (1978), *First Steps in Numerical Analysis* Hodder and Staughton, London pp. 28-30.

HOTSON, J. (1988) 'Land use and agricultural activity : an arealapproach for harnessing the Agricultural Census of Scotland'.*Working paper No 11*, RRL Scotland Regional Research Laboratory, Edinburgh.

HOUSE OF COMMONS, (1986), *First report from the Environment Committee.* Session 1985-6. Radioactive waste Vol 1, HMSO, London. Chapters 1 and 7.

HOWE, G.M., (1973) 'The Environment, Its Influences and Hazards to Health', *Environmental Medicine* ed. Howe G.M. and Loraine J.A., Ch 1 pp 1-8.

HOWE, G.M. and PHILLIPS, D.R., (1983), 'Medical Geography in the United Kingdom, 1945-1982', *Geographical Aspects of Health : Essays in honour of Andrew Learmonth*, ed. McGlashan ND and Blunden J.R. Academic Press, London.

IBM, (1983), VS *Fortran Application Programming: Language Reference*, IBM, San Jose.

INFANTE, P.F., SCHWARTZ, E. and CAHILL R., (1990) 'Benzene in petrol : a continuing hazard', *Lancet* **336**, pp 314-815.

JOHANNSEN, I.M., (1978), 'How to get Cartographic Data in computer Systems', *Computer - Assisted Cartography* International Cartographers Association, 6-11 November, Nairobi-Kenya.

KEHRIS, E., (1990), 'Interfacing ARC/INFO with GLIM' *NW Research Report No.5* for the North West Regional Research Laboratory, Lancaster University.

KELLETT, C.E., (1937) 'Acute Myeloid Leukaemia in one of identical twins', *Arch. Dis. Childh.* **12**, p 239.

KEMP, J., BOYLE, P., SMAN, M., & MUIR, C., (1985), A*tlas of cancer in Scotland, 1975-1980: incidence and epidemiological perspective*, IARC Scientific Publications No. 75.

KINLEN, L.J., CLARKE, K. and HUDSON, C., (1990) 'Evidence from population mixing, in British New towns 1946-85, of an infective basis for childhood leukaemia', *Lancet* **336**, pp 577-582.

KINLEN, L.J., (1988), 'Evidence for an infective cause of childhood leukaemia : comparison of a Scottish new town with nuclear reprocessing sites in Britain', *Lancet ii* pp. 1324-1327.

KLEINSCHMIDT, I, (1990) 'An Information System for Small Area Health Statistics', presented at the Consultation on the Development of a Health and Environmental Geographical Information System for the European Region, Bilthoven, 10-12 Dec 1990.

KNOX G., (1964) 'Epidemiology of Childhood Cancer in Northumberland and Durham', *British J.Prev. Soc. Med*, **18**, pp 17-24.

LEBRET, E, (1990) 'Indicators of Public Health and Environmental quality', presented at the Consultation on the Development of a Health and Environmental Geographical Information System for the European Region, Bilthoven 10-12 Dec 1990.

LEGON, CD, (1952) 'The Aetiological Significance of Geographical Variations in Cancer Mortality', *British Medical Journal*, **1**, pp 700-2.

LENIHAN, J. (1985) *Bonnybridge/Denny Morbidity Review*, Scottish Home and Health Department, Edinburgh.

LINET, M.S., (1985) *The Leukaemias: Epidemiological aspects.* Oxford University Press, Oxford.

LLOYD, O.L., WILLIAMS F.L.R., BERRY W.G. and FLOREY C du V.,(1987) *An Atlas of Mortality in Scotland*, Croom Helm, London.

LLOYD, O.L. (1990) 'Geographical Information in Environmental Epidemiology' presented at the Consultation on the Development of a Health and Environmental GIS for the European Region, Bilthoven 10-12 Dec.

LUCIE, N.P., (1989) 'Radon exposure and leukaemia', *Lancet ii*, pp 99-100.

MAGUIRE, D.J., GOODCHILD, M.F. and RHIND, D. (1991) *Geographical Information Systems, Volume 1: Principles and Volume 2: Applications*, Longman, London.

MARBLE, D.F., LANZON, J.P. and McGRANAGHAN, M., (1984) 'Development of a conceptual model of the manual digitising process', B*asic Readings in Geographic Information Systems* ed. Marble D.F., Calkins H.W. and Peuquet D.J. SPAD systems Ltd.

MARBLE, D.F., (1990) 'Geographic information systems : an overview', *Introductory readings* in Geographic Information Systems ed. Peuquet D. and Marble D.

McGARTH G., (1986) 'The Challenge to Educational Establishments : preparing students for a future in LIS/GIS', *Proc. Auto Carto*, London Vol 2.

McGLASHAN, N.D., (1972) 'Medical Geography : An introduction', M*edical Geography : Techniques and Field Studies.* ed. McGlashan N.D. Methuen, London, pp 3-17.

McGLASHAN, N.D., (1972) 'Medicine and Medical Geography by A.T.A. Learmonth' *Medical Geography : Techniques and Field Studies*, ed McGlashan N.D., Methuen, London, pp 17-43.

McGLASHAN, N.D. and BLUNDEN, J.R., (1983) 'Introduction', *Geographical Aspects of Health : Essays in honour of Andrew Learmonth.* ed. McGlashan N.D. and Blunden J.R. Academic Press, London.

McINNES, G., (1988), 1) 'Airborne lead concentrations in the United Kingdom 1984-1987', *Laboratory Report* LR/676 (AP). Warren Springs Laboratory.

McINNES G., (1988), 'Airborne lead Concentrates and the effects of reductions in the Lead Content of Petrol'. *LR 587 (AP)M.* Warren Springs Laboratory.

McWHIRTER, W.R., (1982) 'The relationship of incidence of childhood lymphoblastic leukaemia to social class', *Br. J. Cancer* **46**, pp 640.

MOELLERING, H., (1980) 'Strategies of real time cartography', *The Cartographic J*, **17**(1), pp12-15.

MONMONIER, M.S., (1982) *Computer Assisted Cartography Principles and Prospects*, Prentice Hall, Englewood Cliffs, N.J.

MORRISON, J.L., (1986) 'Cartography. A Milestone and its Future', *Proceedings Auto Carto* London, Vol 2 pp 1-13.

MOUNSEY H., (1990) 'From Research to Reality - The Diffusion of Innovation' presented at the Association of Geographic Information National Conference, Brighton.

MYERS, A., CLAYDON A.D., CARTWRIGHT R.A., and CARTWRIGHT S.C., (1990) 'Childhood cancer and overhead powerlines : a case-control study', *Br. J. Cancer* **62**, pp. 1008-1014.

NATIONAL CENTER FOR GEOGRAPHIC INFORMATION AND ANALYSIS, (1989) 'The research plan of the National Center for Geographic Information and Analysis', *Int. J. GIS.* 3(2), pp 117-136.

NATIONAL INSTITUTE OF PUBLIC HEALTH AND ENVIRONMENTAL PROTECTION. (1989), European Environmental Agency (RIVM, Bilthoven, The Netherlands).

NATIONAL INSTITUTE OF PUBLIC HEALTH AND ENVIRONMENTAL PROTECTION. (1989), *Concern for tomorrow, a national environmental survey 1985-2010 - highlights* (RIVM, Bilthoven, The Netherlands).

NATIONAL RADIOLOGICAL PROTECTION BOARD (1990), Comment, *Lancet* **335**, pp 220 Jan 1990.

NEWCASTLE UNIVERSITY COMPUTING LAB., (1987) *Introduction to the VAX/VMS Graphics System*, mimeo.

NORTHUMBRIAN UNIVERSITIES MULTIPLE ACCESS COMPUTER, An *Introduction to FORTRAN* (1986) (course notes) mimeo.

NYSTROM, D.A., (1980) 'Geographic Information System Developments within the US Geological Survey', *Proceedings in Autocarto*, **1**, pp 33-43.

O'BRIEN, L.G., (1989), 'The Statistical Analysis of Contingency Table Designs.' *CATMOG 51*, GeoBooks, Norwich.

O'RIORDAN, M.C., JAMES, A.C., GREEN, B.M.R. and WRIXON, A.D., (1987) 'Exposure to Radon Daughters in Dwellings', *NRPB-GS6*, NRPB.

OPENSHAW, S., (1982) 'The Modifiable Areal Unit Problem', *CATMOG 38*, GeoBooks, Norwich.

OPENSHAW, S., (1983) 'Multivariate analysis of census data : the classification of areas', *A Census Users Handbook* ed Rhind D, Methuen, London, pp 243-64.

OPENSHAW, S., CHARLTON, M. and WYMER, C., (1987) 'A Mark 1 Geographical Analysis Machine for the automated analysis of point pattern data', *Int. J. GIS* **1**(4), pp 335-358.

OPENSHAW, S., (1987) 'Learning to live with errors in spatial databases',*NorthEast Regional Research Laboratory Report*, Newcastle University.

OPENSHAW, S. & SCHOLTEN, H., (1990) 'Spatial Analysis and Geographical Information Systems : An introduction to an exciting subject!' Workshop paper in Spatial Analysis given at the First European Conference on Geographical Information Systems, Amsterdam, The Netherlands April 10.

OPENSHAW, S., CROSS, A. and CHARLTON, M., (1990) 'Building a Prototype Geographical Correlates Exploration Machine, *Int. J. GIS* **4**(3) pp 297-311.

OPENSHAW, S., (1990) 'Spatial Analysis and Geographical Information Systems : A Review of Progress and Possibilities', *Geographical Information Systems for Urban and Regional Planning* ed. H.J. Scholten and J.C.H. Stillwell, Kluwer Academic Publishers. Printed in the Netherlands. pp. 153-163.

OPENSHAW, S., CROSS, A., CHARLTON, M., BRUNSDON, C., & LILLIE, J., (1990) ' Lessons learnt from a Post Mortem of a failed GIS' presented at the Association for Geographic Information National Conference, Brighton.

OPENSHAW, S. and CRAFT, A. (1991) 'Using Geographical Analysis Machines to search for evidence of clusters in childhood leukaemia and Non-Hodgkins lymphoma in Britain' *NorthEast Regional Research Laboratory Working Paper.*

PENDERGRASS, T.W., (1989) 'Epidemiology of Acute lymphoblastic leukaemia', *Seminars in Oncology* **12** pp 80-91.

PERKAL, J., (1966), 'On the length of empirical curves', Discussion paper 10. Ann Arbor, Michigan Inter-University Community of Mathematical Geographers.

PEUCKER, T.K. (1972) 'Computer Cartography', Association for American Geographers, *Resource paper No 17.*

POST OFFICE, (1985) *The Postcode Address File Digest* (PAF), The Post Office.

PRISLEY, S.P., (1986) 'Commercial GIS's for natural resources management. What a manager needs to know', *Proceedings of Geographic Info. Sys Workshop, American Soc. for Photogrammetry and Remote Sensing*, pp 1-12.

RAYBOULD, S., (1988) 'Environmental Correlates of Childhood Cancer in Tyne and Wear'. Unpublished Thesis, University of Newcastle upon Tyne.

RAYBOULD S., NICOL, J., CROSS, A., & COOMBES, M., (1991) 'The Long term potential of GIS for Epidemiology', *The Added Value of Geographical Information Systems in Public and Environmental Health*, ed. Stern R, de Lepper M., & Scholten H. (forthcoming).

READING, R.F., OPENSHAW, S. and JARVIS, S.N., (1990) 'Measuring child health inequalities using aggregations of Enumeration Districts', *Journal of Public Health* **12**(3/4), pp 160-167.

REDDY, B.S. et al, (1980) 'Nutrition and its relationship to cancer', *Advances in Cancer Research* **32** pp 238-41.

RHIND, D., (1977) 'Computer aided Cartography', *Transactions of the Institute of British Geographers new series* **2**, pp 71-97.

RHIND, D.W., EVANS, I.S. and VISVALINGAM, M., (1980) 'Making a National Atlas of Population by Computer', *The Cartographic Journal* **17**(1) pp 3-11, The British Cartographic Society.

RHIND, D., (1983) 'Mapping census data', *A Census User's Handbook* ed, Rhind D., Methuen, London, pp 171-98.

RHIND, D. and TANNEBAUM, E., (1985) 'Linking census and other data', *A Census User's Handbook*, ed. Rhind D., Methuen, London, pp 287-300.

RHIND, D., (1984) 'Remote Sensing, Digital Mapping and Geographical Information Systems : The creation of a National Policy', *Proceeding of the Int. Symposium on Spatial Data Handling, 1* Zurich, Switzerland, Geographisches Institut, pp 6-17.

RHIND, D., (1987) 'Recent developments in geographical information systems in the UK', *Int. J. GIS.*, **1**(3), pp 229-241.

RHIND, D.W. and GREEN, W.P.A., (1988) 'Design of a Geographical Information System for a heterogeneous scientific community', *Int. J. GIS.*, **2**(2), pp 171-189.

ROBINSON, V.B. and STRAHLER, A.H., (1984) 'Issues in Designing Geographic Information Systems under conditions of inexactness'. Paper given at Hunter College - New York Medicine Processing of Remotely Sensed Data.

ROGERS, M.A. and PENDERGRASS, T.W., (1987) 'Incidences and epidemiological characteristics of neuroblastoma in the United States', *J. Epidemiol.* **126**(6) pp 1063-74.

ROGERS, R., (1982), *Lead Poison NS Report 7*, Newstatesman.

ROSE, G., (1986) *A proposal for a Small Areal Health Statistics Unit, SAHSU*, London, mimeo.

ROSENBERGER, G. and HEESCHEN, W., (1960) 'Adlerfarn (Pteris aquilina) die Ursache des sog. Stauroter der Rinder (Hae maturia vesicalis bovis chronica), Dt. tierarz. Wschr, 67, pp 201-7, referenced in Evans (1976).

RUMSEY, R.D.E., (1973) 'Radiation and Health Hazards', *Environmental Medicine* ed. Howe G.M. and Loraine J.A. pp 25-39.

RUTTER, (1983) 'Low level lead exposure : sources, effects and implications', *Lead versus Health: Sources and Effects of Low Level Lead Exposure*. ed. Rutter M. and Jones R.R. Wiley, Chichester.

SANDERS, B.M., WHITE, G.C. and DRAPER, E.J., (1981) 'Occupations of fathers of children dying from neoplasms', *J. Epid. and Comm. Health*, **35** pp 245-250.

SCHOLTEN, H.J. and de LEPPER, M.J.C., (1990) 'The Application of Geographical Information Systems in Public and Environmental Health, presented at the Consultation on the Development of a Health and Environmental Geographical Information System for the European Region. Bilthoven 10-12 December.

SHU-QUIANG, W. & UNWIN D.J., (1991), 'Modelling Landslide distribution on loess soils in China: An Investigation using GIS techniques' presented at the Seventh Colloquium on Quantitative and Theoretical Geography, Stockholm.

SILVERMAN, B.W., (1986) *Density Estimation For Statistics and Data Analysis* Chapman and Hall, London.

SMITH, T.R., MENON, S., STAR, J.W. and ESTES, J., (1987) 'Requirements and principles for the implementation and construction of large scale geographic information systems', *Int. J. GIS.* **1**(1), pp 13-31.

STAR, J.L. and CONSENTINO, M.J., (1984) 'Geographic Information Systems: Questions to ask before it is too late'. Paper given at University of California, Santa Barbara on Machine Processing of Remotely Sensed Data.

STATHER, J.W., CLARKE, R.H. and DUNCAN, K.P., (1988) 'The Risk of Childhood Leukaemia near Nuclear Establishments', *NRPB-R215*, NRPB, London.

STATUTORY INSTRUMENTS, (1980), 'The Control of Pollution (Special Waste) Regulations'. *Public Health, England and Wales* No. 1709, HMSO, London.

STOCKS, P., (1928), 'On the Evidence for a Regional Distribution of Cancer in England and Wales' *Report of the International Conference on Cancer*, London, British Empire Cancer Campaign. pp 508-519.

SWERDLOW, A.J., (1986) 'Cancer Registration in England and Wales : Some Aspects Relevant to interpretation of the Data', *J.R. Statist. Soc. A.* **149** Part 2 pp 46-60.

THE INDEPENDENT, (1989) 'An atmosphere poisoned by mistrust', The Independent Newspaper, October 3rd.

THOMAS, R.W., (1990) 'Introduction : Issues in Spatial Epidemiology', *Spatial Epidemiology* ed. R.W. Thomas. London papers in Regional Science 21.

TOBLER, W.R., (1959), 'Automation and Cartography', *The Geographical Review* **1**(49) pp. 526-36.

TOMENIUS, L., (1986) '50Hz electromagnetic environment and the incidence of childhood tumours in Stockholm County', *Bioelectro-magnetics* 7, pp 191.

TOMLINSON, R.F., (1987) 'Current and potential uses of geographic information systems: The North American experience', *Int. J. GIS.* **1**(3), pp 203-218.

UNIVERSITY OF LEEDS, (1990), *Leukaemia and Lymphoma: An atlas of distribution within areas of England and Wales, 1984-1988*, compiled by the Leukaemia Research Fund, Centre for Clinical Epidemiology, Leeds.

UNWIN, D., (1981), *Introductory Spatial Analysis* Methuen, London.

UNWIN, D. and DAWSON, J. (1985) *Computer Programming for Geographers*, Longman, London.

UPTON, G. AND FINGLETON, B., (1985), *Spatial Data Analysis by Example Volume 1: Point Pattern and Quantitative Data* Wiley, Chichester.

VAN STEENSEL-MOLL, H.A., VALKENBURG, H.A., VANDENBROUCKE, J.P. and VAN ZANEN, G.E., (1983) 'Time space distribution of childhood leukaemia in the Netherlands', *Journal of Epidemiology and Community Health*, **37** pp 145-148.

WARREN SPRINGS LABORATORY, (1986) 'United Kingdom Acid Rain Monitoring' - air pollution division, warren Springs Laboratory.

WARREN SPRINGS LABORATORY SPECIAL PUBLICATION, (1985) *The investigation of Air Pollution Directory, Part 1. Daily Observation of Smoke and Sulphur Dioxide*, part of a Cooperative Survey up to March 1985, for Dept. of Industry.

WERTHEIMER, N. and LEEPER, E., (1979) 'Electrical wiring configurations and childhood cancer', *Am. J. Epidemiol.* **109** pp 273.

WHO (1985) Draft Project Plan for Development of an Environmental Health Information System for the European Region, 24th August.

WHO (1988) Summary Report, for the Consultation on Environmental Health Information Systems in the European Region. Berlin (West) 21-25 November 1988.

WHO (1989) First European Conference in Environment and Health, Frankfurt, 7-8 December 1989.

WHO (1990) Development of Health and Environment Geographical Information System for the European Region. *(Target 19) Report on a WHO Consultation*, Bilthoven 10-12 December 1990.

WIGGINS, J.C., HORTLEY, R.P., HIGGINS, M.J. and WHITTAKER R.J., (1987) 'Computing aspects of a large GIS for the European Community', *Int. J. GIS.* **1**(1), pp 77-87.

WILSON, A., (1974), *Urban and Regional models in Geography and Planning* Wiley, Chichester.

WOOD, P.A. (1977). 'Information for Geography-cause for alarm?', *Area*, **9**(2), pp 109-113.

WRIGLEY, N. and BENNETT, R.J. (1981) *Quantitative Geography: a British view*, Routledge and Kegan Paul, London, pp 1-36.

YIAMOUYIANNIO, J., (1975) 'A definite link between fluoridation and cancer death rate'. Nat*ional Health Federation*, HMSO, London Chapters3,6 and 9.

ZIB, P., (1977) 'Urban Air Pollution Dispersion Models: A Critical Survey, Department of Geography and Environmental Sciences, University of the Witwatersand, Johannesburg.