An evaluation of small-area statistical methods for detecting excess risk: with applications in breast and colon cancer mortality in Scotland 1986-1995

Grant Mark Andrew Wyper

A Dissertation Submitted to the

University of Glasgow

for the degree of

Master of Science

Department of Statistics

June 2009

Acknowledgements

There are a number of people who have helped me throughout my period of research and postgraduate degree. Firstly I would like to extend my appreciation to ISD Scotland as an organisation for providing the funding to allow my period of research to take place. I hope my research can help to add value to the department of Epidemiology and Statistics or at the very least present information for thought on alternative ways of analysing their data. Within this department I would like to thank Roger Black who has overseen the research from ISD Scotland. Roger's expertise in the area of Epidemiology has both been inspirational and invaluable.

During my final year of my postgraduate degree, I have been employed by ISD Scotland to work in the Primary Medical Services Information team. I would like to thank my team manager and co-workers for being understanding and for providing me with the opportunity to continue my research.

The staff and postgraduate students at the University of Glasgow have helped my period of study be both useful in developing my education and an enjoyable experience. The level of support and educational insight that my supervisor Professor E. Marian Scott has put into this thesis has helped guide me through my research. I am indebted for her support and I appreciate the time and effort that she has spent during the production of this thesis.

Finally I would like to thank both my family and my fiancé for standing by me throughout this period. There have been some difficult times experienced over my period of study and without support coming from home and work, the production of this thesis would not have been possible.

Abstract

The need to report data at small-area level is constantly increasingly. In a society which is both health-conscious and environmentally aware, statistics at small-area level have a high degree of political significance. This type of data is required to plan and implement regional policies and apportion health care in accordance to the differing needs of the population. Recent advances in computer power has brought many advances to this area of study. For all the advances in technology and methodology, the problem of small numbers consistently appears. Is there an excess risk or is it down to chance? This is a question which is paramount in small-area statistics and will be addressed in this thesis.

An overview of the thesis is provided below:

Chapter 1 introduces the concept of small-area statistics and some of the social and political issues connected with this topic. There is a discussion of the analysis of small-area health data and the principal ideas that need to be considered in a statistical, political and social sense in this area of work. The aims of ISD Scotland are introduced and how they can be linked to this field of study.

Chapter 2 describes an overview of the methods used in small-area statistics. The chapter begins by firstly considering the Standardised Incidence Ratio (SIR) which is the technique mainly used in the basic analysis done by ISD Scotland. Other techniques are then considered, however not all of these techniques are directly comparable to each

other. The strengths and weaknesses of these techniques in previous research are discussed to give an idea of how the techniques perform in different scenarios.

Chapter 3 is a simulation study of three of the techniques discussed in Chapter 2, these being the SIR, Circular Spatial Scan and Flexibly-Shaped Spatial Scan. The reason for this simulation study is to evaluate these techniques on simulated data arising from real scenarios. The strengths and weaknesses of these techniques are then highlighted which will prove helpful when analysing the data in Chapter 4.

Chapter 4 provides an analysis of the mortality of breast and colon cancer in Scotland in the ten-year time period from 1986 to 1995. Using data provided by ISD Scotland, the analysis is carried out to identity any potential mortality clusters in both diseases.

Chapter 5 provides a conclusion to this research by providing a summary of findings of the thesis and gives recommendations based upon these findings. A discussion is also given for potential further study in this field that could provide some value to ISD Scotland as they look to other ways of analysing their small-area data.

Glossary of Abbreviations

ISD	Information Services Division
CSA	Common Services Agency
ESG	Epidemiology and Statistics Group
GIS	Geographic Information System
HRT	Hormone Replacement Therapy
ICD	International Statistical Classification of Diseases and Related Health Problems
MLE	Maximum Likelihood Estimator
NHS	National Health Service
RIF	Rapid Inquiry Facility
RRF	Relative Risk Function
ScotPHO	Scottish Public Health Observatory
SAHSU	Small Area Health Statistics Unit
SEHD	Scottish Executive Health Department
SIMD	Scottish Index of Multiple Deprivation
SIR	Standardised Incidence Ratio
SMR	Standardised Mortality Ratio

Contents

List of Tables

List of Figures

Chapter 1

Introduction to Small-Area Spatial Epidemiology

1.1 Introduction to Epidemiology

Epidemiology is an investigation into disease, and its occurrence, in different groups of people. This information is used to prepare and assess different strategies to limit or prevent illness. It is also used as a guide to the management of patients where the disease has already developed [1].

A key concept of epidemiology that differentiates it from clinical medicine is that epidemiologists are concerned with both people who get a disease and those that do not and how these two groups of people differ. Epidemiologists are concerned with whole groups and communities whereas clinicians direct their questions at particular patients.

To further study disease, Epidemiologists study the distribution of disease amongst groups. Epidemiologists ask the questions 'Who?', 'When?' and 'Where?' [2]. The first question refers to groups of people under investigation, for example males and females or people of different ages. The second refers to the time period over which the study takes place. The last question refers to where it is that the study is taking place *i.e.* referring to the geographical region, for example it may be a city or a country.

After answering these questions the next step in the process is to find out why some groups of people are at a higher risk. This question can be answered by considering associations between certain risk factors and increased risk of disease. These risk factors can vary from environmental exposures to lifestyle factors such as diet. The relationship between risk factors and disease is the fundamental concept of Epidemiology.

1.2 Introduction to Small-Area Statistics

Nowadays in our increasingly health-conscious and environmentally aware society there is a need to report on data at a scale fine enough to meet the demands of the local population. Advances in computer technology and statistical methods have made possible the collecting and reporting of statistics at a small-area level [3].

Small-area statistics have a high degree of political significance. Small communities can often unite based on ethnicity, gender, location or age and have strong feelings surrounding their identity and how they are portrayed [4]. It is not uncommon to see tabloid headlines portraying a community in a negative way. These headlines can tarnish the reputation of communities since many people believe them to be true in every sense. Small-area statistics are useful since they can objectively and independently report on specific populations on issues such as crime rates and disease rates on a fine scale.

The need for data reporting at this level is constantly increasing. The data are required to carry out regional policies and apportion health care in accordance to the need and

demand for it. Problems can only be dealt with efficiently when data on disease is explained well and can be followed back to a defined area.

1.3 Spatial Epidemiology

Spatial data has been studied for millennia through the construction of maps. The need to reduce these data to numbers is a relatively new idea, made possible by the surge in computer power over the past few decades [5]. Spatial Epidemiology is the account and analysis of geographically indexed health data in respect to demographic, environmental, behavioural, socio-economic, genetic and infectious risk factors [6].

Epidemiologists are primarily concerned with the distribution and determinants of health related conditions or events in a specified population. Epidemiology is the study and application of these concerns in an attempt to exert a level of control over health problems. Studying the distribution of a defined disease in a specified population, rather than individuals, allows us to seek which factors are causing the condition. The identification of these factors not only alerts us to which types of people are at risk, it aids us in making changes or introducing measures in an effort to control the risk.

Spatial Epidemiology deals with an array of issues. Given below are some practical examples of questions that may be addressed.

- How does cancer vary between local government regions in Scotland?

- Do cases cluster together in a particular area?
- Is there a raised incidence of leukemia around a nuclear installation?

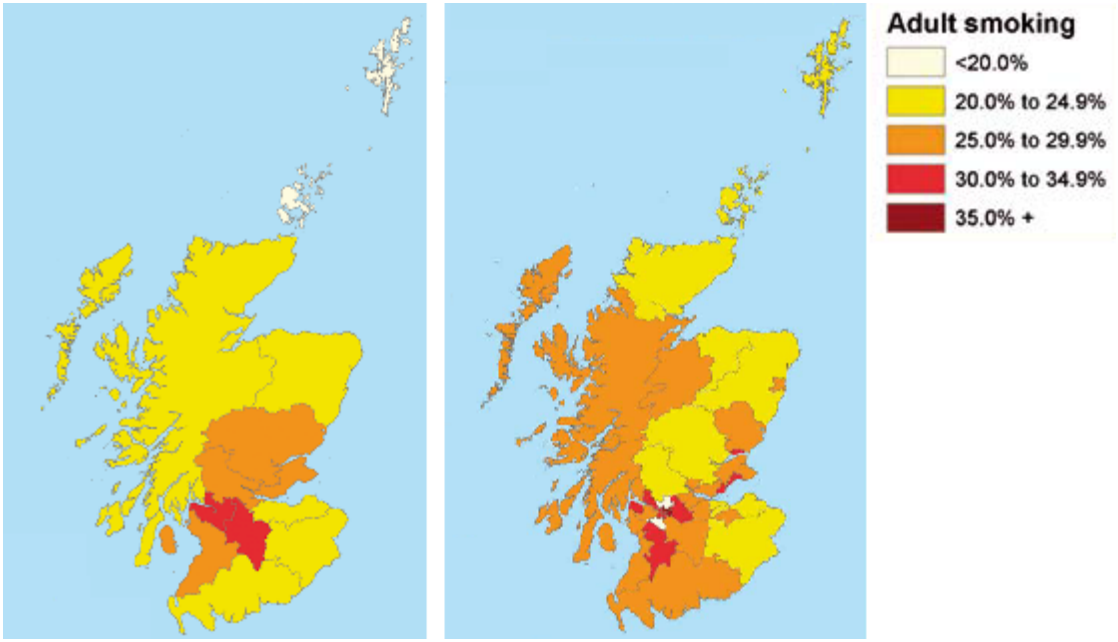
These practical issues are amongst the basic types of questions that arise in this area. To deal with these sorts of questions effectively, Spatial Epidemiology must be split into a number of areas and depending on the type of issue, it may fall within one of these sub-sections. These areas include disease mapping, geographic correlation studies and disease clusters and surveillance. Although these areas of study are considered to be different they often overlap.

1.3.1 Disease Mapping

Disease mapping is a visual summary of the geographic indexed health data on the disease under inspection. This helps to create a picture of how risk varies across the entire study area. Typically this involves the mapping of Standardised Incidence Ratios (SIR) for each area nested within the overall study area. The SIR is a method which can be used to assess the risk of disease and its methodology will be discussed further in Chapter 2.

By constructing a visual summary of the risk of disease this can help highlight patterns or any causes for concern which may have been overlooked in any previous analysis [6]. This mapping tool has been used from as early as 1936 where it was used to explain the variation in cancer mortality in England and Wales [7].

This function can aid surveillance, as any high-risk areas are visually noticeable. Although this technique can be very helpful it also has several drawbacks so care must be taken when these maps are constructed. In creating these maps there are many choices for consideration and one that can drastically change any inferences is the number and geographic boundary of sub-areas which are used. If maps of the same area are produced, using different numbers of sub-areas and different geographical boundaries for these sub-areas, they can suggest very different things [7, 8].



(i) *(ii)*

Figure 1.1: Estimates of smoking prevalence in the adult population in Scotland

The example above is health data taken from the website of the Scottish Public Health Observatory (ScotPHO). Figure 1.1 above shows estimates of smoking prevalence in the adult population (ages 16 or over) in Scotland by *(i)* NHS Board and by *(ii)* Community

Health Partnership (CHP) [9]. Both of these graphs are plotted using the same scale and the same data but produce very different pictures. In *(i)* some of the higher prevalence rates are lost due to them being grouped together with lower rates. The difference between the two also displays the need for health care to be apportioned according to the need for it, rather than uniformly over larger areas, like *(i)* would suggest.

When these maps are constructed, the size of units and the method to aggregate units must be selected to highlight the features we are interested in. Groups that are aggregated must display similar characteristics for any conclusions to be meaningful. Varying the scales and aggregation approaches can lead to different characteristics of the data being displayed, though in practise we seek to choose geographic units which are as small as possible although the option is not usually available as the accessibility of the data usually dictates the selection. Since the data are often sparse in small area statistics there is a trade-off between homogeneity within small geographic units and precision of risk estimates.

1.3.2 Geographic Correlation Studies

In geographic correlation studies our primary concern is how geographical variations in disease data relate to geographical variations in risk factors. The problems arising from disease mapping are also true for geographic correlation studies. Another drawback is the potential for correlation between the confounding variables. For example it is not

unusual for people living in a deprived area to be living close to a hazardous environmental source.

A crucial consideration is the ecological fallacy. This is when relationships that are observed at group level are incorrectly taken to imply association at an individual level. A study carried out by Davies (1997) is an example of the ecological fallacy in application. In this study the deprivation score was classified using a number of indicators [10]. The fallacy occurs in that many people view individuals on low incomes as deprived.

1.3.3 Disease Clusters

The connotation of the word clustering suggests a gathering of events within a small area. A disease cluster exists when the frequency of the number of cases of a particular disease in a defined population is greater than it is expected to be over a given time period.

The proposition of disease clusters can surface when people observe a raised incidence of a given disease within a defined population. In many instances members of the public start to observe a similar disease pattern within family members, neighbours or co-workers. These perceptions can trigger a public outcry if the media pick up on these insights. The reaction to these reports can cause problems of post-hoc inference. These problems occur when information that has no relationship is manipulated until it appears to have meaning. This is known as the Texas Sharpshooter fallacy.

From an ethical viewpoint it is essential that Epidemiologists take these views into account rather than disregard them [11]. In considering these views Epidemiologists can devise a framework to make the required decision on whether or not to address the publics' uncertainty. Trumbo (2000) documents the framework which was used in the United States in 1997. In this study the importance of cluster investigations relative to other tasks was scored 3.5 where 1 is unimportant and 7 is important so it shows that whilst these claims are not the top priority, they are still taken seriously.

For concerns to be valid the perceived raised incidence must be between diseases which are the same, if not very similar. If the observed disease is different in each sample under observation, the proposed disease cluster is not as likely to be a true cause for concern since a combination of different factors cause different diseases. A more frequent occurrence of a rare disease should instantly raise the level of alert compared to common disease since there may be something to explain this raised incidence in the rare disease.

The differences between a surplus of cases in a small-area or around a hazard source and the general tendency for clustering must be taken into account. This is a key idea within the Spatial Epidemiology framework, where numbers could simply be down to chance alone rather than some underlying mechanism. Only clusters that have a disease rate that is statistically significantly greater than the rate of the general population are investigated.

1.4 Confounding

Confounding occurs when one or more factors distort the association being studied between two variables due to a strong correlation between the factors and the two variables being studied. An example of this is that alcohol appears to cause lung cancer. In this example the confounder is smoking, since increased alcohol consumptions tens to be associated with increased smoking, which is in turn associated with lung cancer.

Let us suppose that the risk of disease amongst our population is increased for those individuals living near a nuclear installation where the harmful substances given off are perceived to be having an adverse affect on a given disease. The population around and nearby this site does not represent a random sample of the general population; instead they have to live in an area that is subject to many social weaknesses [12].

On the whole when diseases display a difference between social classes, the more deprived individuals experience a higher risk of disease [13]. These socio-economic factors act as a confounder to the link between disease and exposure. If these socio-economic factors are not taken into account then the relationship between disease and location may suffer from bias.

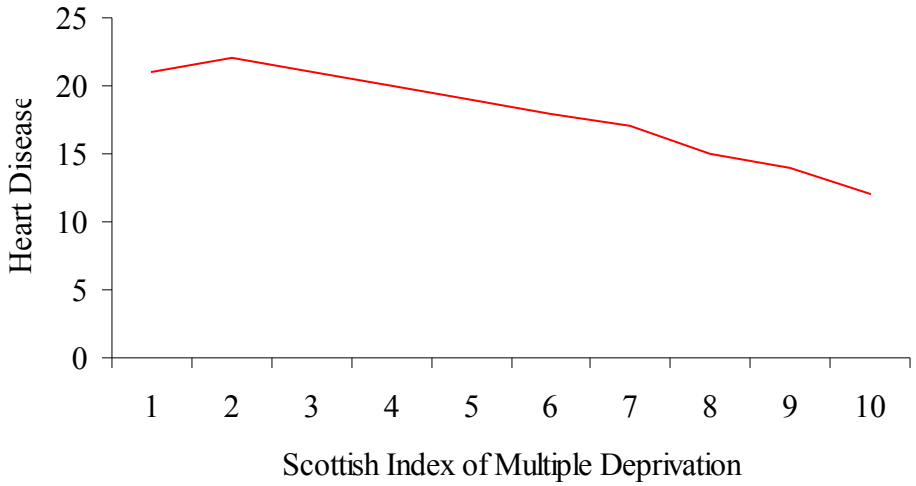


Figure 1.2: Rate of heart disease per 1,000 population plotted against SIMD decile

Figure 1.2 is an example which was published by ISD Scotland on the website of the Scottish Government [14]. The study aimed to link ill health or mortality with deprivation, in this example heart disease. The Scottish Index of Multiple Deprivation 2006 (SIMD 2006) is used to allocate individuals into a deprivation category based on their postcode of residence.

SIMD 2006 is based across seven domains which are: current income, employment, health, education, housing, geographic access and crime. The overall index score is a weighted sum of each of these scores in the seven domains. The weighting given to each domain is based upon the relative importance of the domain in measuring multiple deprivation. The weights which were used in SIMD 2006, expressed as a percentage of the overall SIMD score, were as follows: current income (28%), employment (28%),

health (14%), education (14%), geographic access (9%), crime (5%) and housing (2%) [15].

In Figure 1.2, decile 1 is the most deprived where as decile 10 is the least deprived. In this example the rate of heart disease per 1,000 people, for hospital inpatients and day cases are most frequent in the most deprived deciles. By looking at Figure 1.2 it is clear that the rate of heart disease per 1,000 people decreases as it shifts to the least deprived deciles.

1.5 Types of Data

Our goal in epidemiological research is to be able to quantify the occurrence of disease in a defined population. To be able to do so we must be able to clearly define what is meant by a case, the population from which the case originates and the time period over which the data were collected. The epidemiological definition of a case is not always the same as the clinical definition and epidemiologists usually have to rely on tests which are less reliable and cheaper than a clinicians test.

Data can be collected through registries, notification systems, death certificates, abstracts or clinical records or surveys of the general population. In the instance of measuring the occurrence of disease in a town, we must include all the cases from that town, even those which have been diagnosed elsewhere. Cases where the samples normally reside elsewhere, such as a term-time student, must be omitted from our calculations. When

analysing small-areas, there are many other factors which should be considered. People commuting to and from a defined area will be subject to a different risk of disease than people that rarely or never leave the specified area [16].

To be able to infer conclusions about a certain disease or exposure, it is essential that the size of the population which the cases occur in is known. For instance it cannot be concluded that a disease is more frequent in city A than city B if it simply has more cases, instead comparing the frequency of disease with the population size allows more meaningful conclusions to be made. If there are any others who are not subject to the risk of living in the defined population then they should be excluded. In addition to this the time period must be clearly defined. Most health related events are not constant through time so we cannot interpret the measure of occurrence without clearly defining the time period.

Firstly we must measure our exposure variable by obtaining information surrounding it. This information can be collected in many forms such as personal interviews, questionnaires, diaries, records, biological measurements and measurements in the environment. In the event that the subject under observation cannot disclose the required information then the data can be obtained from a proxy respondent. There are many difficulties faced when dealing with exposure due the difficulties faced when measuring consistently over time. Data on the history of residential locations is not readily available in the United Kingdom [17]. Knowing these data would benefit analysis of longer

latency health outcomes as the history of exposure could be more readily reconstructed rather than the location at time of diagnosis or death.

The duty of care does not only lie with the subject under examination. If we are to receive the information which we desire, then it is essential that our questions reflect this. Generalised questions must not be asked if we require detailed information *e.g.* we should enquire about specific forms of a disease rather than just generally the disease.

Defining exposure is rarely a simple question because exposure is seldom simply present or absent. Some exposures are quantitative variables and relate to different extremes of exposure, for instance, radiation could be measured by extent where the people living closest to it would be subject to greater doses than people living further away would. In studying the link between radiation and cancer, we can further analyse our conclusions if there is a trend of increasing numbers of cases of cancer with increasing exposure to radiation.

We can obtain our data on a defined outcome of interest through various sources. Similarly to collecting data on our exposure variable, we can use questionnaires or interviews amongst many other methods to collate information on the samples' health status. An individual's status can be monitored by following them up using techniques such as periodic interviews or check-ups to update the information on their health state. With cancer cases it is more likely that our information would be obtained from hospital records, cancer registrations or death certificates. By using records our data can be

limited because outcomes are often recorded routinely and may omit other data that we may be interested in.

There are two types of data which can be used in the analysis of small area statistics, the first of which being point data. Point data is where the exact location of each case is known. On the other hand there is count data, where the number of cases over a defined area has been aggregated. In order to allow for different age and social differences amongst populations, we can allow for the division of groups where differences in the population are clear. For example age specific expected rates can be calculated.

When using health data, it is essential that consideration is paid to the protection of individuals. At ISD Scotland, the Disclosure Control protocol sets out procedures to guard, not only, patient level data but also potentially disclosive data such as small numbers. This consideration is especially relevant in the field of small area statistics due to the scale of some of the analysis.

1.6 ISD Scotland

This M.Sc. project is carried out in conjunction with ISD Scotland. The role of ISD needs to be reactive to the requirements of the NHS due to the developing delivery of healthcare. Its dynamic approach helps to resolve and advise how to best use information effectively to guarantee the highest standard of patient care [18]. The project is run by the Epidemiology and Statistics Group (ESG) which maintains the area of work on

disease surveillance, public health, the evaluation of health and social care interventions and quality improvement in health services.

ISD Scotland is the Information Services Division of the National Health Service (NHS) National Services Scotland, which is formerly the Common Services Agency (CSA). ISD Scotland has been in existence for over 40 years. The division provides a support service to NHS Scotland and the Scottish Executive Health Department (SEHD) in response to the needs of NHS Scotland as the delivery of health and social care changes over time. The ISD Scotland website [18] documents ISDs vision of success in the future. The vision is to be an essential partner in providing better health and better care for people in Scotland through:

- Information that leads to action
- Proactive and innovative approaches
- Working in partnership with our customers

ISD Scotland brings together a range of data about the individual and stores it in the national database. It works with a widespread range of organisations from hospitals, General Practitioners and voluntary organizations, amongst many more, to build the national database. The information which is collected can be broadly defined as: patient and activity data workforce and earnings data and NHS Scotland complaints data.

From January 2001 the majority of the statistics published by ISD have been covered by National Statistics. During June 2000, National Statistics was introduced to the UK, changing the way which official statistics were governed and how accountability was assigned. The main objective of National Statistics is to improve the quality, timeliness and relevance of its service to customers [19]. Every publication that is published under National Statistics must have a planned release date and all publications must be pre-announced on the ISD website.

In addition to publications, information can be requested through ISDs Information Request Service. This method is another opportunity to better understand and explore health and social care. This service is subject to resource constraints and a prioritisation process. Priority is given to NHS Scotland, Scottish Executive, Local Authorities, Audit Scotland and independent contractors such as Dentists or General Practitioners.

1.7 Aims of Thesis

In this project the aim is to review some of the current methods used in the analysis of small-area statistics data and to compare those to the current methodology used in ISD Scotland. Firstly this will begin by reviewing the theory of some of the commonly used techniques in small-area analysis that will lead to a discussion on the apparent advantages and drawbacks which these methods exhibit in different scenarios.

Whilst the theory and previous analysis provides a deep insight to how these techniques perform, a simulation study will be carried out in Chapter 3 to asses their performance. Once the methods have been documented and analysed, an investigation of breast and colon cancer in Scotland will take place. This is an ISD Scotland dataset which will be analysed for the first time.

The final aim of the project is to make recommendations based upon the research which can hopefully benefit as they look towards other ways of improving their analysis of small-area data.

Chapter 2

Comparison of Small-Area Statistical Methods

2.1 Introduction

The following chapter documents the theory of some of the most commonly used techniques in small-area statistics. The advantages and disadvantages of each technique as discussed in academic literature will also be covered. This is to give an indication of the level of importance placed on each of the techniques in small-area data analysis.

The methods which will be covered are:

- Standardised Incidence Ratio (SIR)
- Besag and Newell Cluster Test
- Circular Spatial Scan
- Flexible Spatial Scan
- Bithell's Linear Risk Score

The methodology of these techniques greatly differ however all these techniques look at count data which has been aggregated to small-area level. In ISD Scotland, the Standardised Incidence Ratio (SIR) is the technique which is most commonly used. This technique takes a set region and calculates the risk for that region. The other methods

implement a range of techniques to scan the entire study area. These types of tests can further be segregated. Besag and Newell and the Spatial Scan techniques are based upon drawing a centroid or another flexible shape which scans for clusters. Bithell's Linear Risk Score differs in that it is concerned with the relationship of risk over a distance; usually how risk varies in relation to a point source.

The first technique which will be documented is the SIR. The technique is highlighted in much more detail since it is the technique which is primarily used in ISD and will, in some forms, re-appear within the other techniques. The theory of the other techniques is documented to give a basic understanding, however the use of these techniques to support the findings of this thesis are used mainly in R software and WinBUGS.

2.2 Standardised Incidence Ratio (SIR)

2.2.1 Overview and assumptions

To get an estimate of the relative risk of disease the SIR can be used. The notation of the method is Standardised Mortality Ratio (SMR) if the data are death rates rather than incidence of disease. This measure of risk is a single summary of incidence, which can be used to compare the risk of similar regions within a study area. It is for this reason and its ease of application that the SIR method is widely used when analysing small-area health data.

2.2.2 Methodology

When using the SIR method the data being used must be count data. The test statistic is concerned with the number of cases within a geographically defined area, rather than point data where the exact location of each case is known. The null hypothesis of the test is set such that the $\theta_i = 1$. It is assumed that the observed number of cases O_i are drawn from a Poisson distribution with mean $\theta_i E_i$ and relative risk θ_i, that is $O_i \sim Pois(\theta_i E_i)$, where i=1,2,…, is the number of regions in the study area.

Firstly the notation required for the method is defined as:

O_i = the number of observed cases of a defined disease in a region i

N_i = the population-at-risk in a region i

r_i = the incidence of a defined disease in a region I for a standard population

E_i = the number of expected cases of a defined disease in a region i

θ_i = the relative risk of a defined disease in a region i

A step-by-step guide to the estimation of the SIR is provided below:

1. Calculate E_i where $E_i = N_i \times r_i$. To allow for different age or social structures within the population this is calculated as $E_{ij} = \sum_j N_{ij} \times r_{ij}$ where N_{ij} is the population-at-risk in a sub-group j in a region i and r_{ij} is the incidence of a defined disease in a sub-group j in a region i.

2. Recalling that $O_i \sim Pois(\theta_i E_i)$, the Maximum Likelihood Estimator (MLE) of θ_i

$$SIR_i = \frac{O_i}{E_i} = \hat{\theta}_i.$$

An SIR of 1 would occur if the observed number of cases for a given population equals the number we would expect. If the SIR < 1 then this suggests that the incidence rate for the defined population is lower compared to a standard population. Finally the case where SIR > 1 is what we are primarily concerned with. This situation arises when there is a surplus of observed cases suggesting that the incidence rate is greater for the defined population when compared to a standard population. This can also be expressed as a percentage by multiplying the MLE by 100 [20]. For example an SIR of 125% would mean that there were 25% more cases observed in the population under study compared to that if the incidence was that of the standard population.

Now we wish to test whether there is evidence that there is an excess risk for a defined area. A surplus of observed cases may simply be down to chance so we cannot conclude there is an excess risk if an SIR > 1. The case where SIR < 1 is not considered due to the fact this represents the case where fewer cases are observed than we expected and does not raise any concerns.

3. Now to test under the null hypothesis that the mean $O_i \sim Pois(E_i)$ the probability of observing at least O_i cases by chance is derived from a Poisson distribution with mean E_i,

$$\Pr(X \geq O_i \mid E_i) = 1 - \Pr(X < O_i \mid E_i)$$

$$\Pr(X \geq O_i \mid E_i) = 1 - \sum_{X=0}^{O_i - 1} \frac{E_i^X e^{-E_i}}{X!} = p. \tag{2.1}$$

In Equation 2.1, the value p is the probability that represents the one sided p-value. If $p \leq \alpha$ where α is the chosen significance level then we would reject the null hypothesis and conclude that the risk of disease is significantly greater at the α% level. If this is the case then there must be some further investigation into the cause of this excess risk [21].

2.2.3 Summary

The SIR is one of the most widely used methods in the analysis of spatial data. There are many reasons for the use of the technique, but there are also many disadvantages or dangers to using the technique.

Disease mapping is an area where the use of SIRs are important. Mapping the SIR of each of the regions in a study area can give a good picture of how the risk is varying amongst the entire study region. However there is a difficulty in deciding what size of geographic area should be used. When areas have larger populations then the rates of disease are more stable. Although there is the danger that small areas with high or low rates will be smoothed out with the choice of a larger area [22]. There are many possible

solutions or ways to get around this problem. Extending the data collection over a larger time period could help as areas with a higher risk may just be down to chance [23].

Unfortunately this does not always overcome the problem. A study of the risk of leukaemia among children living near the Solway coast of Dumries and Galloway Health board area in Scotland was carried out for the period 1975-2002 [24]. This study investigated two similarly sized time periods 1975-1989 and 1990-2002 to analyse the incidence of childhood leukaemia. During this study confidence intervals were used for the SIRs. Confidence intervals can be useful when SIRs are being used. If the confidence interval for the SIR does not contain 1, then this can be used as a test for a statistically significant excess risk of disease. However during this study the width of the confidence intervals were large due to the numbers of cases being small.

Another difficulty that needs to be considered when using SIRs is that the data can show extra Poisson variability where $\mathrm{Var}(O_i) > E(O_i)$ [25]. The reason for this may be due to the data being dependent on an unknown or missing confounding factor.

A major advancement in the area of SIRs is the use of smoothing for risk estimation [26]. Using smoothing techniques allows the user to make use of the data on disease rates in nearby regions which is useful if there a spatial dependence between rates in nearby or neighbouring regions. This can also be a useful technique to overcome small numbers and unstable rates. The way smoothing works is that the smoothing estimate borrows

precision from data in nearby areas that depends on the precision of the raw estimate for each area.

There are some disadvantages to using smoothing methods. In using these techniques the raw data is being adjusted. Many people are apprehensive when numbers are adjusted, especially if money or power is to be allocated based upon the numbers [27]. Although borrowing strength from other nearby regions can be an advantage, it can also be problematic as it introduces autocorrelation [28]. Another disadvantage to using smoothing techniques is that it can smooth out any high rates which may represent a true elevated risk of disease.

2.3 Besag-Newell Cluster Test

2.3.1 Overview and assumptions

The method proposed by Besag and Newell [29] was formulated to detect clusters of disease. The method was first used to search for clusters of childhood leukaemia in northern England. The test statistic searches for clusters of a set observed number of cases. At each region where a case is observed, the number of nearby regions needed to reach the set number of cases is computed. If there are too many observed cases in a small number of regions then it the result is that is a potential cluster.

When using this technique some questions that need to be considered are:

What time period of data to use?

There are many factors which could influence data arising from different time periods. The Besag and Newell method is based upon identifying historical clusters so it cannot be used to make any forecasts on future clusters or clustering.

What is the risk of disease?

The risk of disease can be calculated in many different ways. Firstly the risk of disease may be the same for each individual. The Besag and Newell method assumes that the risk of disease is the same for everyone, however this may not be the case in some regions.

What number of cases defines a cluster?

This is one of the main selection criteria of the Besag and Newell test statistic. Users of the test must pick a value, corresponding to the number of cases, which is the minimum number of cases that can appear as a cluster. This value is vital as different numbers of cases can have an affect on which clusters appear to be statistically significant. When choosing the cut-off value for a cluster to be statistically significant, the value must not be too strict as some clusters on the borderline may not be detected.

2.3.2 Methodology

Firstly the region of observation must be defined. For each region, the centroid is calculated by taking the population-weighted centre of the region. In order to calculate this, it is essentiall that the weighted centre is based upon the population-at-risk of disease and not the general population. If it is based upon the general population then if a specialised group of the population is being studied then the centroid will not be accurate.

Once the centroid of each area is set then the closest centroid in distance is added to the current region up until the number of cases which is specified is attained. The null hypothesis is set to state that the observed number of cases is distributed at random among the population-at-risk. This test is for data that is aggregated to areal units.

Firstly the following terminology is defined:

k = the minimum number of cases defined for the test statistic for each region

O_j = the number of observed cases in a region j

p_j = the population of a region j

t = the incidence of disease across the entire region

A_0 = the region where the cluster occurs

A_i = {1,2,3,4,…} which is determined by the increasing distances of the centroids from A_0.

The method is outlined below:

1. Firstly calculate $D_i = \left(\sum_{j=o}^{i} O_j \right) - 1$ where $D_o, ..., D_i$ are the summed number of cases in regions $A_o, ..., A_i$.
2. The numbers of population-at-risk $u_i = \left(\sum_{j=0}^{i} p_j \right) - 1$ where $u_0 \le u_1 \le ...$ in regions i=1,2,…
3. Now let $M = \min\{i : D_i \ge k\}$ so that regions $A_o, ..., A_M$ but not $A_o, ..., A_{M-1}$ contain at least k other cases.
4. The probability under the null hypothesis is the hypergeometric probability that s individuals amongst u_m are cases. This can be approximated by the Poisson term $\frac{e^{-\lambda}\lambda^s}{s!}$ where $u_m t = \lambda$ and $t = \frac{O}{p}$.
5. $\Pr(M \le m) = 1 - \Pr(M > m)$ where M related to the minimum number of regions required to sum to at least k cases, so

$$\Pr(M \le m) = 1 - \sum_{s=0}^{k-1} \frac{e^{-\lambda}\lambda^s}{s!}. \qquad (2.2)$$

The resulting p-value in Equation 2.2 is the Monte Carlo p-value which is determined from comparing the observed test statistic to the reference distribution created by randomising the cases across the study area. If the value of the test statistic is large, that

is when there are much more observed cases in just a few regions with low expected cases, then the null hypothesis of no clustering is likely to be rejected.

2.3.3 Summary

The Besag and Newell cluster test is designed to be used as a screening test to detect clusters. Further analysis may be required if the test identifies a cluster of several regions. One of the main criticisms of the Besag and Newell test is that it does not control for multiple testing through the definition of consistent clusters [30]. This can lead to many false positive clusters arising. It is recommended that this test is run multiple times to try to better detect the presence of clustering [31]. Song and Kulldorff (2003) evaluated the power of numerous methods of detecting disease clusters [32]. It concluded that the Besag and Newell technique is a good choice if the size and scale of clustering is roughly known. The reason for this is because the technique is greatly dependent on the choice of parameter [33]. Knowing the size and scale of a cluster can help because there is less difficulty in choosing a minimum number of cases to define a cluster.

2.4 Circular Spatial Scan Statistic

2.4.1 Overview and assumptions

Rather than investigating the risk of a disease in a single region, spatial scan statistics scan the full study region for clustering. The circular spatial scan, allows users to investigate data on disease across an entire number of sub-areas and to detect if there is a circular cluster of disease amongst a circular window imposed around an area.

When running the spatial scan, the scanning window is in the form of either a circle or an ellipse. The window with the Maximum Likelihood Estimator is the most likely cluster which is the cluster which is least likely to be due to chance.

The spatial scan statistic used in this thesis is based on the spatial scan used in the SaTScan [34] software. The only difference is that for this thesis the number of cases in each region is assumed to be Poisson distributed. When using the Poisson model, the requirements are as follows:

- Case and population counts are defined for each region
- The geographical location of each region is defined
- The time period over which the analysis is to take place is defined

2.4.2 Methodology

The theory for this method is based upon the method as outlined by Kulldorff (1997) [35].

Firstly let N stand for a spatial point process where the number of random points is defined as $N(A)$ in the set $A \subset G$ where G is a geographical space of which A is a subset of. As the scan statistic window scans the study area it defines a set $\mathcal{Z}$ of zones where $Z \subset G$. There is just one zone $Z \subset G$ where individuals have probability p of being a point, probability q of being outside the zone and μ is the underlying intensity governing the distribution of points under the null hypothesis such that $N(A) \sim Pois(p\mu(A \cap Z) + q\mu(A \cap Z^C)) \forall A.$. The null hypothesis can be defined as $p = q$ whilst on the other hand, the alternative hypothesis is such that $p > q, Z \in Z.$ $N(A) \sim Pois(p\mu(A)) \forall A$ under the null hypothesis. The null hypothesis is that the risk of disease is the same in the entire study area, where as the alternative is that there is an excess risk in the circular scan window.

Now the likelihood of the Poisson model is defined. The probability of n_G points in the study region is defined as

$$\frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}\left[p\mu(Z)+q(\mu(G)-\mu(Z))\right]^{n_G}}{n_G!} \tag{2.3}$$

The density function of a specific point being observed at location x, $f(x)$ is

$$\begin{cases} \dfrac{p\mu(x)}{p\mu(Z)+q(\mu(G)-\mu(Z))} & \text{if } x \in Z \\ \dfrac{q\mu(x)}{p\mu(Z)+q(\mu(G)-\mu(Z))} & \text{if } x \notin Z \end{cases} \tag{2.4}$$

The likelihood function can now be written in the following form:

$$L(Z,p,q) = \frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}\left[p\mu(Z)+q(\mu(G)-\mu(Z))\right]^{n_G}}{n_G!} \times \prod_{x_i \in Z} \frac{p\mu(x_i)}{p\mu(Z)+q(\mu(G)-\mu(Z))} \prod_{x_i \notin Z} \frac{q\mu(x_i)}{p\mu(Z)+q(\mu(G)-\mu(Z))}$$

$$L(Z,p,q) = \frac{e^{-p\mu(Z)-q(\mu(G)-\mu(Z))}}{n_G!} p^{n_Z} q^{(n_G-n_z)} \prod_{x_i} \mu(x_i). \tag{2.5}$$

The likelihood ratio λ can be defined as $\lambda = \dfrac{\sup_{Z \in Z, p>q} L(Z,p,q)}{\sup_{p=q} L(Z,p,q)} = \dfrac{L(\hat{Z})}{L_0}$. (2.6)

We have $L_o = \sup \dfrac{e^{-p\mu(G)} p^{n_G}}{n_G!} \prod_{x_i} \mu(x_i) = \dfrac{e^{-n_G}}{n_G!} \left(\dfrac{n_G}{\mu(G)} \right) \prod_{x_i} \mu(x_i)$. (2.7)

Now we take the supremum over all p and q for a fixed Z. It takes a maximum when

$$p = \frac{n_Z}{\mu(Z)} \text{ and } q = \frac{(n_G - n_Z)}{(\mu(G)-\mu(Z))}. \tag{2.8}$$

We now have

$$L(Z)=\begin{cases}\dfrac{e^{-n_G}}{n_G!}\left(\dfrac{n_Z}{\mu(Z)}\right)\left(\dfrac{n_G-n_Z}{\mu(G)-\mu(Z)}\right)^{n_G-n_z}\prod_{x_i}\mu(x_i) & \text{if } \dfrac{n_Z}{\mu(Z)}>\dfrac{\left(n_G-n_Z\right)}{\left(\mu(G)-\mu(Z)\right)} \\ \dfrac{e^{-n_G}}{n_G!}\left(\dfrac{n_G}{\mu(G)}\right)\prod_{x_i}\mu(x_i) & \text{otherwise.}\end{cases} \tag{2.9}$$

The likelihood ratio can now be written as

$$\lambda=\frac{\sup_{z\in Z}L(Z)}{\dfrac{e^{-n_G}}{n_G!}\left(\dfrac{n_G}{\mu(G)}\right)\prod_{x_i}\mu(x_i)}$$

$$\lambda=\sup_{z\in Z}\frac{\left(\dfrac{n_Z}{\mu(Z)}\right)^{n_Z}\left(\dfrac{n_G-n_Z}{\mu(G)-\mu(Z)}\right)^{n_G-n_Z}}{\left(\dfrac{n_G}{\mu(G)}\right)^{n_G}}I\left(\frac{n_Z}{\mu(Z)}>\frac{n_G-n_Z}{\mu(G)-\mu(Z)}\right) \tag{2.10}$$

when there is at least one Z where $\dfrac{n_Z}{\mu(Z)}>\dfrac{n_G-n_Z}{\mu(G)-\mu(Z)}$ and $\lambda=1$. If this is not the case then $I(\ )$ is the indicator function.

To find the distribution of the test statistic under the null hypothesis, Monte Carlo hypothesis testing is used. The *p*-value of the test is based upon the null distribution of likelihood ratio test statistic with a recommended large number of Monte Carlo replications of the specified data set generated under the null hypothesis.

2.4.3 Summary

Where the spatial scan differs from other methods is that it is concerned with comparing what is inside the spatial scanning window with what is outside, for example the number of cases. This is seen to be an advantage as there may be a raised risk in the whole study area.

Unlike the Besag and Newell method, the circular spatial scan is useful if the size and scale of clustering is not known [36]. The spatial scan may be a better choice to gain inferences about possible clusters and then it may be useful to test using the Besag and Newell method once information on clustering is generally known.

Another advantage of this method is that it does not have multiple testing problems [37]. The test is said to be good at pinpointing hot spot clusters, but a limitation of the circular spatial scan is that the shape of the clusters is restricted, that being the shape of a circle of an ellipse [38].

2.5 Flexible-Shaped Spatial Scan Statistic

2.5.1 Overview and assumptions

Tango and Takahashi (2005) proposed a new method of scanning the study area to detect clusters [39]. The underlying reason behind this proposal was that the previous scan

statistic proposed by Kulldorff (1997) [35] was set up to use a circular window to define possible clusters. This is seen is a disadvantage, as there are many situations where clusters may take a different shape. The flexible spatial scan, allows users to explore disease data across an entire number of sub-regions and to detect if there is an irregularly shaped cluster of disease amongst a circular window imposed around an area. An example of this may be on a coastline. Tango and Takashi (2005) discusses the use of a scan statistic for detecting non-circular clusters [39].

2.5.2 Methodology

The methodology discussed is based upon the findings of Tango and Takahashi (2005). The use of this method in this thesis is utilised using FleXScan software, which is software for the flexible spatial scan statistic [40]. This software is similar to SaTScan [34].

The first step in the method is to consider the entire region being split into m sub-regions. In each sub-region i, the number of cases is given by the random variable N_i where the observed value is given by n_i where i=1, 2, …, m. The null hypothesis H_0 is that there is no clustering in the region. We define N_i as independent Poisson variables under H_0 such that

$$H_0 : E(N_i) = \xi_i,\ N_i \sim \text{Pois}(\xi_i),\ i=1, 2, \ldots, m.$$

In the above equation Pois(e) is the Poisson distribution with mean e and the ξ_i are the expected number of cases in the sub-region i under the null hypothesis. For the purposes of this thesis the geographical position of each sub-region i will be the population-weighted centroid. The null hypothesis is such that the underlying risk of disease is the same in the entire study area, where as the alternative is that there is an excess risk in at least one scan window.

The circular scan statistic uses a circular window $\mathbf{Z}$ on each population-weighted centroid of each sub-region i. The radius of the circle d varies from 0 to a fixed maximum distance, or a set maximum number of regions K which are to be included in the cluster. If the population-weighted centroid of a sub-region i is incorporated within the scan window then the entire region is included in the window. A number of overlapping circles that are possible clusters are created, each of different scale and location. If $\mathbf{Z}_{ik}$ is the scan window, where k=1,2, …, K, created by the (k – 1) nearby neighbouring regions to i then the windows to be scanned by the circular spatial scan statistic are included in the set.

$$Z_1 = \{\mathbf{Z}_{ik} \mid 1 \le i \le m,\ 1 \le k \le K\}\,. \tag{2.11}$$

On the other hand, the flexible scan statistic creates a set of irregularly shaped windows which has k sub-regions which include the sub-region i. The connected sub-regions are forced to be subsets of the set of sub-regions i and (K – 1) nearest neighbours to the sub-region i where K is the maximum length of the cluster which is pre-specified.

If $\mathbf{Z}_{ik(j)}$ is the jth window where j=1,2, …, j_{ik} and is a set of k sub-regions which are joined, starting at the sub-region i where j_{ik} is the number of j which satisfies $\mathbf{Z}_{ik(j)} \subseteq \mathbf{Z}_{ik}$ when k = 1, 2, …, K. All the windows scanned are included within the set

$$Z_2 = \{\mathbf{Z}_{ik(j)} \mid 1 \le i \le m, 1 \le k \le K, 1 \le j \le j_{ik}\}. \qquad (2.12)$$

Where this differs from the circular spatial scan statistic is that the flexible spatial scan statistic considers K concentric circles and all the possible permutations of connected regions and including the chosen sub-region i where the centroids are within the Kth largest concentric circle. Therefore the size of Z_2 is much greater than the size of Z_1 which is a maximum of mK.

Now we consider the alternative hypothesis where there is at least one scan window $\mathbf{Z}$ where the risk is higher in the scan window than it is on the outside

$$H_0 : E(N(\mathbf{Z})) = \xi(\mathbf{Z}), \text{ for all } Z;$$

$$H_1 : E(N(\mathbf{Z})) > \xi(\mathbf{Z}), \text{ for some } Z.$$

$N(\ )$ is the random number of cases and $\xi(\ )$ is the number of cases under the null hypothesis in the scan window. In each window the likelihood is computed in order to examine the observed number of cases in and outside the specified window. The test statistic is

$$\sup_{\mathbf{z}\in Z}\left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{n(\mathbf{Z})}\left(\frac{n(\mathbf{Z}^C)}{\xi(\mathbf{Z}^C)}\right)^{n(\mathbf{Z}^C)} I\left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{n(\mathbf{Z}^C)}{\xi(\mathbf{Z}^C)}\right) \tag{2.13}$$

where $\mathbf{Z}^C$ refers to every region which lies outwith the window $\mathbf{Z}$, $n(\ )$ is the number of cases observed within the window and $I(\ )$ is the indicator function. The window $\mathbf{Z}^*$ which has the maximum likelihood is defined as the most likely cluster.

In order to find the distribution of the test statistic under the null hypothesis, we need to use Monte Carlo hypothesis testing. To compute the *p*-value of the test, we base it upon the null distribution of likelihood ratio test statistic, where the number of Monte Carlo replications of the dataset generated under the null hypothesis should be as large as possible.

2.5.3 Summary

In proposing the method, Tango and Takahashi (2005) compared the performance of the circular spatial scan to that of the flexible spatial scan [39]. The findings of the analysis confirmed that the circular spatial scan statistic is highly accurate in detecting circular regions and correctly identifying hot spot regions as part of the most likely cluster. The flexible spatial scan is shown to have good power in detecting circular clusters. The main strength of the technique is the ability to pinpoint non-circular hot spot clusters more accurately than the circular spatial scan method.

One drawback of the analysis was that the flexible spatial scan statistic only works well for small or reasonable cluster sizes of up to around 30. The method was found not to be practical for larger clusters sizes.

2.6 Bithell's Linear Risk Score

2.6.1 Overview and assumptions

An alternative method of analysing small-area data is to consider the relationship between risk and the distance from a point source. One of the most common techniques to use when dealing with this sort of data is Bithell's Linear Risk Score [41, 42]. This is a focused cluster test which examines how the risk of a disease varies as the distance from a defined hazard source changes.

This method assumes that there is hazard source, or point of reference from which the distance can be measured to centroids as defined earlier in subsection 2.3.2. It also assumes that the alternative hypothesis is defined by a Relative Risk Function (RRF) to outline how the relative risk varies with distance.

2.6.2 Methodology

The method, known as Bithell's *T* statistic is formulated by assigning a risk score to each region.

The notation required for the method is defined as follows:

O_i = the number of observed cases of a defined disease in a region *i*

θ_{1i} = the relative risk for a defined region *I* based on the alternative hypothesis

The steps required in the analysis are provided below:

1. Firstly define the coordinates of the hazard source or point of reference.
2. Define the alternative hypothesis in the form of a RRF.

The null hypothesis is that the observed counts are Poisson distributed with relative risk 1. The alternative hypotheses are drawn up to show how the relative risk varies with distance. An example of some of these relative risk functions are given below in Figure 2.1

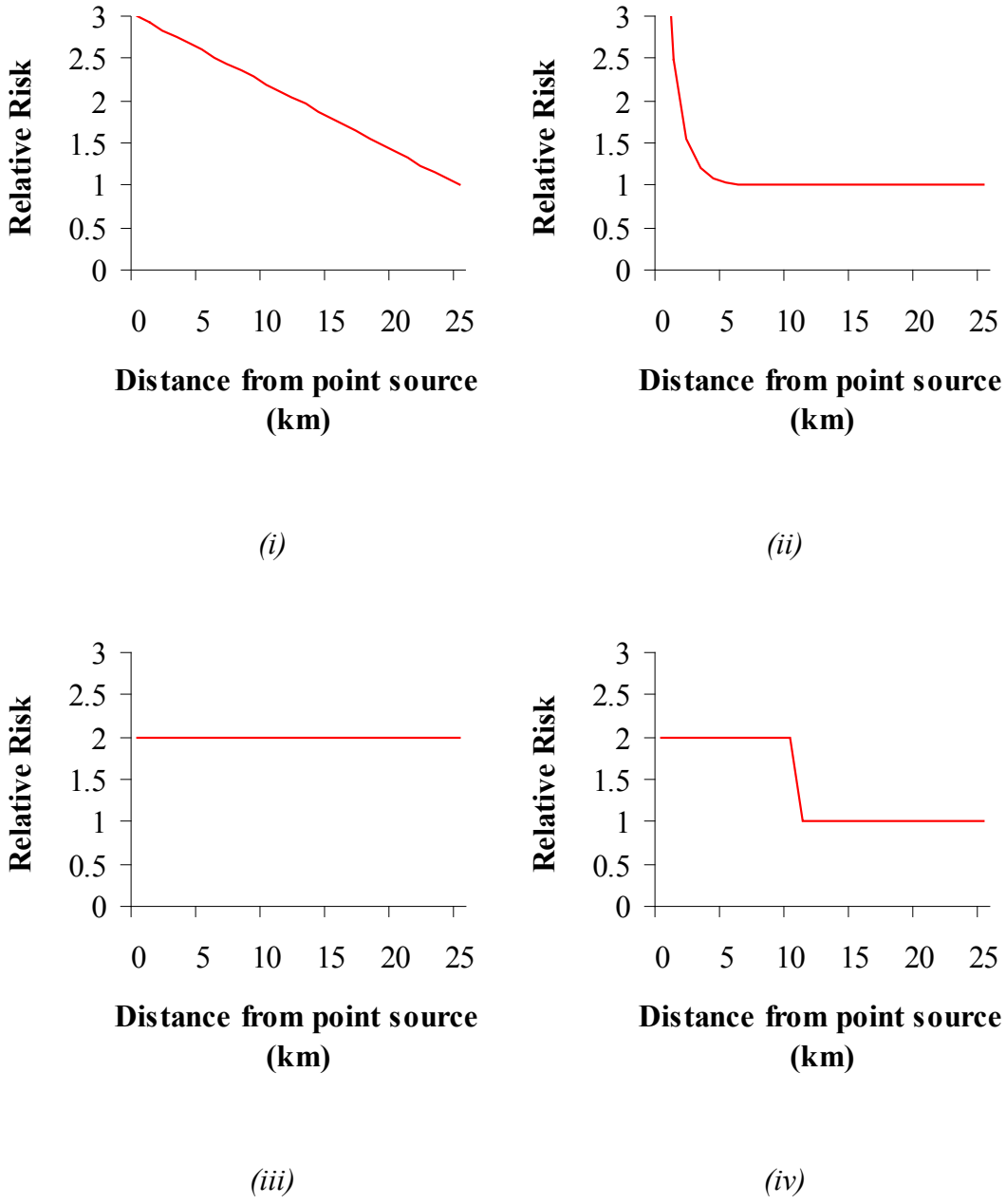


Figure 2.1: Examples of relative risk functions which can be used to define the alternative hypothesis

In Figure 2.1 we have the following:

- In *(i)* the relative risk is 3 at the hazard source which declines linearly to 1 at 25km from the source

- In *(ii)* the relative risk is 3 at the hazard source which declines exponentially to 1 at 5km

- In *(iii)* the relative risk in the entire 25km region is raised at 2

- In *(iv)* the relative risk is 2 at the hazard source up to 10km from the point source then the relative risk instantly drops to 1

The reason behind using these functions is that they can incorporate how risk can vary to exposure in distance and direction. Figure 2.1 *(iii)* and *(iv)* however show that risk can vary in a different way and does not need to follow a set drop. In the case of *(iii)* the risk has a constant rise, and in the case of *(iv)* the risk takes a sudden drop. For these examples, the risk function would still be identical in all directions.

To simulate these scenarios, the expected number of cases would be generated using mathematical functions set-up to generate cases in a way that would reflect the relative risks relationship with distance from the point source.

3. The next step is to calculate $T = \sum_i O_i \log(\theta_{1i})$

The *p*-value of the test is based upon Monte Carlo simulations, for which a large number is recommended.

2.6.3 Summary

Recently the study of risk in relation to distance from a point source has been a central area of investigation. The political, social and health aspects of areas surrounding many nuclear installations in the UK have prompted a great deal of research within this area. The Black Advisory Group was set up in 1983 in response to a documentary connecting raised risk of cancer within the surrounding area of a nuclear installation and the radioactivity which this installation released [43]. The report released by the group [44] concluded that there was a higher incidence of childhood leukaemia, however the estimated radiation source could not account for this increased incidence on the basis of the knowledge available at that time.

Since the publication of this report, there have been countless studies and re-analysis of data and data of a similar nature, such as the study into the incidence of childhood leukaemia and non-Hodgkin's lymphoma in the vicinity of nuclear sites in Scotland from 1968-93 [45]. This study was an investigation into seven nuclear sites in Scotland, namely: Dounreay, Chapelcross, Hunterston, Torness, Faslane, Holy Loch and Rosyth. This study found that there was no statistically significant excess risk of disease within the vicinity, however the numbers of cases were low. This is a problem with using

methods such as the linear risk score, as the areas can be small, so expected numbers of rare diseases can be very small.

One of the pioneers of these methods is the Small Area Health Statistics Unit (SAHSU) [46]. SAHSU have been involved in the area of risk and exposure assessment since its foundation in 1987. SAHSU have developed a software tool for this type of analysis, called the Rapid Inquiry Facility (RIF) [47]. The RIF is a tool that is provided as an extension to Geographic Information Software (GIS) technology. The software can quickly perform risk analysis around a hazard source by generating standardised rates and relative risks for a defined disease. The RIF was intended for internal use at SAHSU but has since been available to other organisations. A recent analysis [48] found the RIF to be useful in enhancing the interpretation of spatial scan results.

Chapter 3

Simulation study

3.1 Size and power of a study

When we are assessing error there are two sorts of statistical error. In a study there is a null hypothesis which relates to the status quo. When testing there must be a defined state which relates to the opposite situation than that of the null hypothesis, namely the alternative hypothesis. When a hypothesis is tested, the main aim is to precisely conclude if the null hypothesis can be discarded or not. There are two results which can occur when a test is carried out *e.g.* a positive result or a negative result. However mistakes can occur, so if the result of the test is not the same as the true condition then an error in testing has occurred. On the other hand if the conclusion is the same as the true condition then the conclusion is correct.

The two types of error which can be made are known as a type I and type II error. When making reference to these, a type I error (α) is when the null hypothesis is rejected when it is in fact true. Therefore it is the error when we reject the null hypothesis in favour of the alternative hypothesis. This error relates to claiming that something is positive when it is not *e.g.* a population has higher than expected risk of a disease when it does not. On the other hand there is a type II error (β) which is the error when the null hypothesis is not rejected when the alternative hypothesis is the true state of nature. This error means that the test has failed to recognise a difference when there actually is one *e.g.* it is

concluded that a population has the same risk as expected when it has a higher risk than is expected. The probability of not making a type II error is known as the power of a test which is equal to 1- β.

When assessing the power of a test we are assessing the probability that the test will reject a false null hypothesis. This is the same as the probability of not making a type II error. A preset significance level α which is usually set to 0.05 (5%), which is known as a type I error. The premise behind a type I error is that we should be minimising the chance of it occurring [49]. For example if we look to set α as 0.05 then there is roughly a 5 in 100 chance that the observed result is down to chance. When dealing with power, any increase results in the chance of a type II error decreasing and vice versa.

3.2 Introduction to SIR simulation

The first method considered in this simulation is the SIR. In order to reflect a real life scenario the leukaemia incidence in Scotland for 2003 will be used. The reason for using leukaemia is that the distribution of risk in leukaemia is varied, enabling tests of different statistical size to take place.

Assuming that there is no underlying process affecting the risk of leukaemia and therefore the cases are occurring at random, leukaemia incidence will be monitored to assess the performance of techniques used in small area statistics in relation to disease.

Table 3.1: Rates of leukaemia per 100,000 person-years at risk in Scotland in 2003

Age	≤ 9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	≥ 80
Male	18.9	8.5	5.2	5.4	11.9	26.8	60.7	159.6	223.0
Female	6.7	6.4	3.9	4.0	10.8	17.5	40.5	79.3	114.3

In order to model leukaemia, the historical trends in incidence for Scotland are used. The most up-to-date rates are the 2003 rates found on the ISD Scotland Cancer webpages. These rates are chosen as the expected rates. Table 3.1 shows these data which will be used for the analysis. The data includes eighteen age groups that are separated by five-year age gaps, but for this analysis, the age groups have been aggregated to give nine age groups.

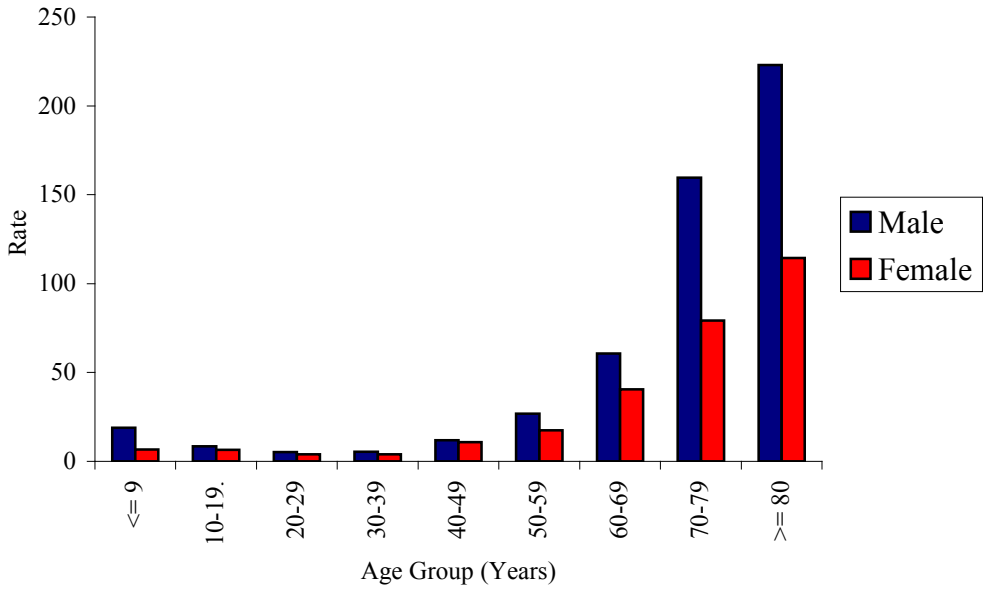


Figure 3.1: Plotted rates of leukaemia per 100,000 person-years at risk in Scotland in 2003

From Figure 3.1 it is clear that overall the rate of leukaemia is greater in males than females. The rate of leukaemia is also much greater in those over 50 years old as we observe a sharp increase in rate for both genders, especially males. On the whole the rate seems fairly steady between the ages of 10 and 40.

3.3 SIR – Simulating under the null hypothesis

When simulating under the null hypothesis we aim to assess the empirical size of the study. In assessing this size, we are looking to measure the amount of times the null hypothesis is rejected when it is actually true. In order to properly define the null hypothesis it is important to think about what information we are receiving from the test statistic regarding the population. When testing the SIR our null hypothesis takes the following form:

$$H_0: \theta = 1$$

which indicates that the relative risk is 1, meaning that there is no excess risk of disease in the set region.

In order to duplicate the conditions under which cases occur, the cases must be simulated under a similar premise. Suppose that the number of cases which are observed in a given population are each located randomly in a fixed region. The assumption of complete

spatial randomness is that the locations of these cases do not influence each other and that risk is uniform throughout the region [50]. To model this type of behaviour the Poisson distribution can be used to simulate the number of cases, with the formula outlined in Equation 3.1, given below

$$O = \text{Pois}(E). \tag{3.1}$$

The Poisson distribution is regularly used to model the number of events that occur in a fixed spatial region when they are taking place at random, such as cases of a rare non-infectious disease. This expresses the probability that a number of events will occur in a fixed period of time and a fixed geographical region given that the expected number of these events is known from average rates. For non-infectious diseases it is assumed that the observed number of cases are independent of one another, which the Poisson distribution fulfils.

The foundations upon which we will be testing have been set up so we look to define our population and region which will be subject to the formal analysis. To do this, information is required on the age, sex and location of the cases which we observe. Although leukaemia is one of the more common diseases, the rates are still fairly low in some age groups so grouping them further will give higher rates within each of the age groups, as displayed in Table 3.1. From Equation 3.1, E is calculated from the figures given in Table 3.1 which contains the leukaemia rates per 100,000 people in Scotland in 2003. When population is allowed to vary during the simulation study, E will be adjusted according to the rates given in Table 3.1.

As previously highlighted, a problem with the SIR is its performance when there are low expected counts. The method is evaluated in a set region, allowing the performance to be assessed over a widespread population range to allow the tests to be conducted in both sparse and dense populations. The problems of low expected counts also arise when the disease under scrutiny is rare. By changing the population size, the expected numbers of disease cases within the population will vary, which will allow the SIR to be evaluated in both low and high extremes.

3.3.1 Empirical size

Looking back, the null hypothesis states that θ=1 meaning that there is no excess risk of disease in the defined region. When we simulate from the Poisson distribution the random variation suggests that the null hypothesis may not always be satisfied. We now test to see how many times we observe a type I error over 1,000 simulations *i.e.* when we reject the null hypothesis when it is the true state of nature.

For this we will test at the 10%, 5% and 1% significance levels to see how the choice of rejection criteria affects the results. At these significance levels, as a rule of thumb, the null hypothesis should be rejected roughly 100, 50 and 10 times respectively from 1,000 simulations. The results of the entire analysis can be viewed in Appendix A. Due to the amount of output in results the discussion will be based on males aged 20-29 years and females aged 80 years and over. The males aged 20-29 years group has the smallest expected number of leukaemia cases in males and the females aged 80 years and over

group has the highest expected number of leukaemia cases, so the results will reflect the low and high expected numbers scenarios.

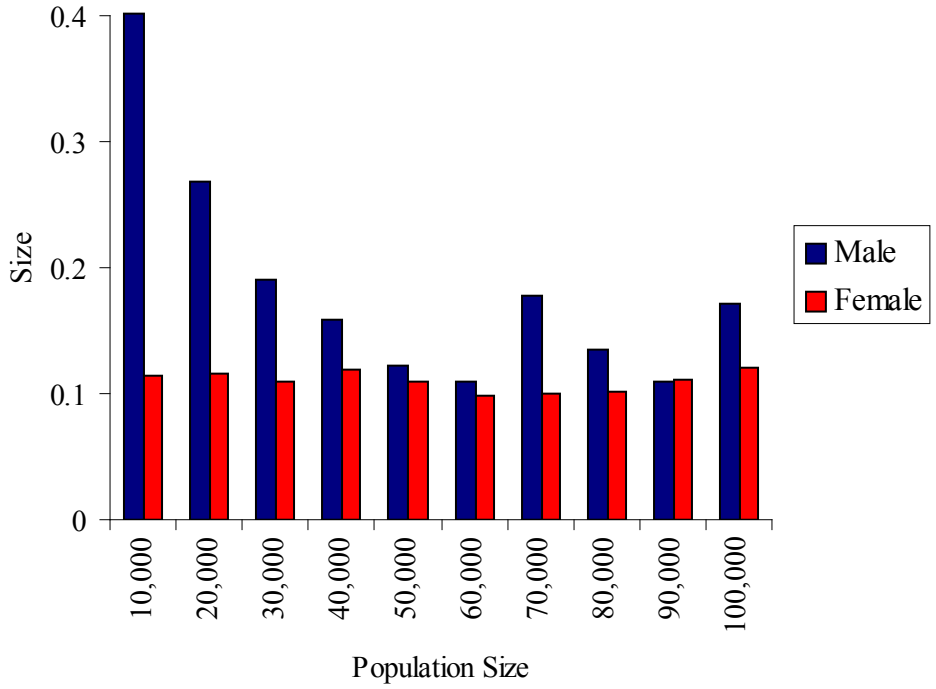


Figure 3.2: Comparison of SIR size calculations when α=0.1

The results when testing at the 10% significance level are displayed in Figure 3.2. The results in the male aged 20-29 years group are very unsteady. The data displays that as the population size increase *i.e.* as the expected number increases, there are less type I errors being made. The results of the female aged 80 years and over group are steady at the 10% level. As a rule of thumb the size calculations should be approximately 0.1 for each of the populations. This criterion is met by the female group but not by the male group where the results are very volatile. Recalling that the male rates used in Figure 3.2

were chosen because it is the smallest rate in males, then this result underscores the problem with small numbers.

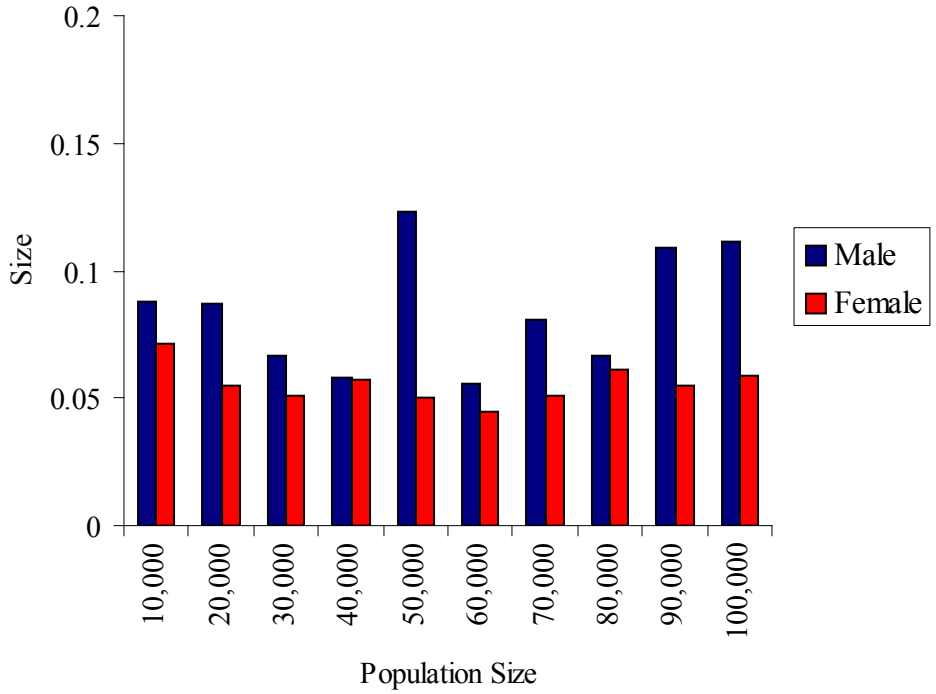


Figure 3.3: Comparison of SIR size calculations when α=0.05

Figure 3.3 compares the results of the SIR size calculations at the 5% significance level. The results of these should approximately lie around 0.05. The male group again displays a lot of variation, which can be attributed to the small number of cases. The amount by which the results fluctuate across the differing population sizes is not as large as in Figure 3. This implies that although the testing conditions are stricter, the SIR is still fairly unstable, although not as much as in the previous case. The female data is

again fairly steady and is approximately around the 0.05 level which we would expect it to be.

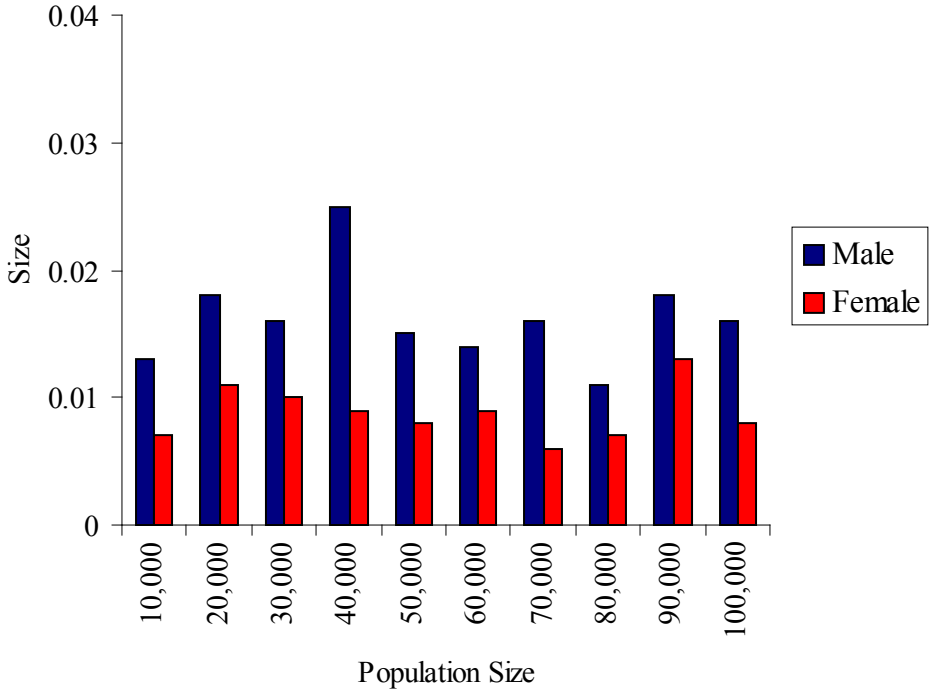


Figure 3.4: Comparison of SIR size calculations when α=0.01

Finally Figure 3.4 displays the results when testing at the 1% significance level. This significance level relates to when the conditions for rejection are very tight. The results for the male age group are a lot closer to what they should be. The results should be close to 0.01. The female age group results are very close to 0.01. Overall both age groups are fairly good when the grounds for rejection are very tight. From Figure 3.4 you can see that the results for the male age group are slightly higher than the female age group. This highlights that although the results are better at this significance level, the fact the expected numbers are low in the male age group affects the quality of the results.

From applying the SIR method to the data, there are clearly some major strengths and weaknesses. The size of the SIR is very unstable when the population is sparse from around the ages 10 to 49 years, especially when the significance level is set as 10% and 5%. The reason for this is that setting the significance level as small as possible helps to protect the null hypothesis and can help prevent false claims being made. Around the ages of 50 and over the size is fairly stable. Looking to the plot of the leukaemia rates in Figure 3.1 there is a clear trend. Between the ages of 10 and 49 this is when the rates are at the lowest. It is at this stage where the null hypothesis is being rejected the most meaning this is where type I errors are most likely to be made. Around the age of 50 the rates start to rise and this rise in numbers is reflected in the size since the size appears to be very stable and closer to the preset significance level around this point.

3.4 SIR – Simulating under the alternative hypothesis

It is not unreasonable for a member of the public to perceive that there may be an underlying mechanism causing the risk of disease in an area to rise. It is for this reason the power of a statistical hypothesis test is measured, since this computes the tests' ability to reject the null hypothesis when it is actually false. The alternative hypothesis is used to assess the power of the study. The alternative hypothesis takes the following form

$$H_1: \theta > 1$$

which means the relative risk is greater than 1 for the region meaning that there is an excess risk of disease in the defined region.

The alternative hypothesis will be simulated from a Poisson process just like the null hypothesis was. The difference this time is that the expected numbers will be manipulated to create a scenario where the relative risk is raised for each age group

$$O = \text{Pois}(\theta E) \tag{3.2}$$

where θ in Equation 3.1 is the relative risk. In the case when simulating the null hypothesis the relative risk was set to 1. When simulating the alternative hypothesis the relative risk will be set to 2 meaning that the number of cases being simulated will be double what we actually expect.

3.4.1 Power

Now we look to assess how many times a type II error occurs, that is when the null hypothesis is not rejected when the alternative hypothesis is the true state. As before, in the empirical size calculations, the males aged 20-29 years and females aged 80 years and over age groups will be used in the analysis. The results for the other age groups can be found in Appendix A.

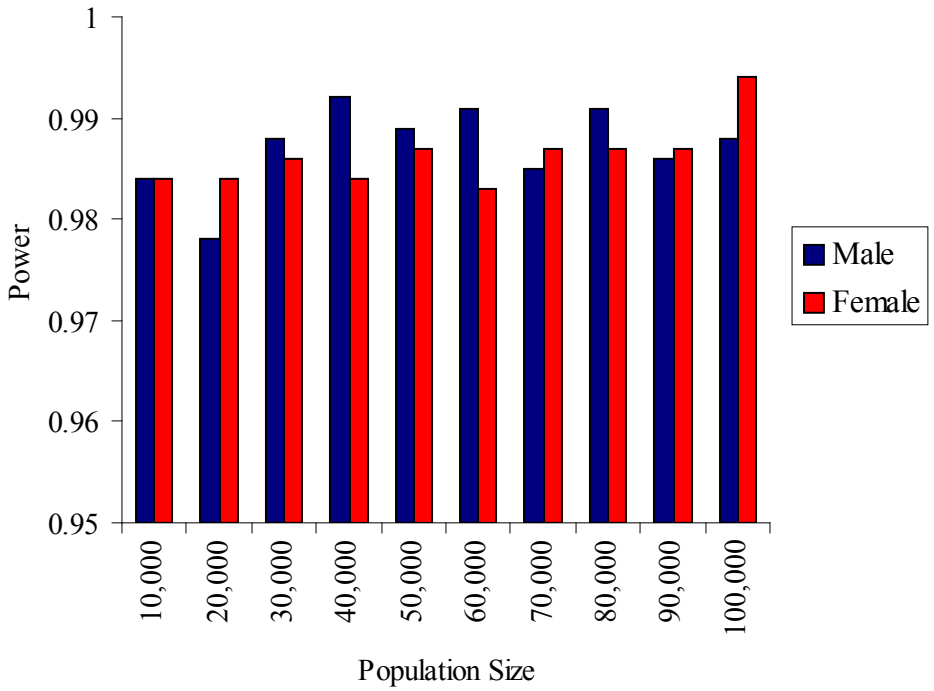


Figure 3.5: Comparison of power of SIR when α=0.1

The power at the 10% significance level is good. The results of the analysis are displayed in Figure 3.5. The power is strong in both the male and female age groups. The lowest power is around 97% that is observed in the male group.

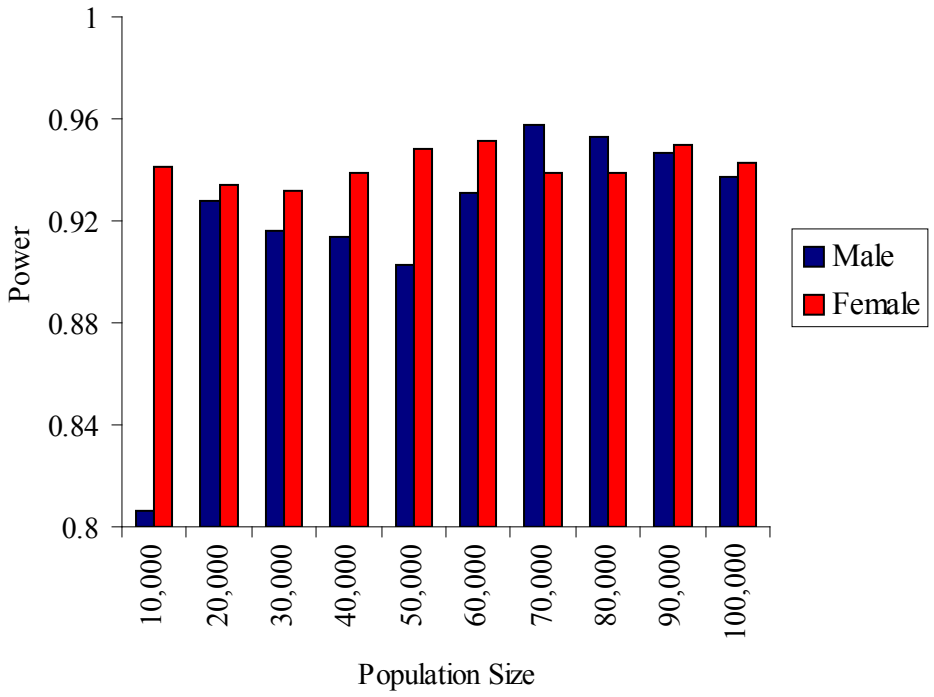


Figure 3.6: Comparison of power of SIR when α=0.05

The results in Figure 3.6 are not as powerful than the previous example, due to the tightening of the rejection region. However, the results display that the female group has good power, roughly around 95%. The male group also has good power, however the low population size of 10,000 is exhibiting weaker power at around 80% than the rest of the population groups.

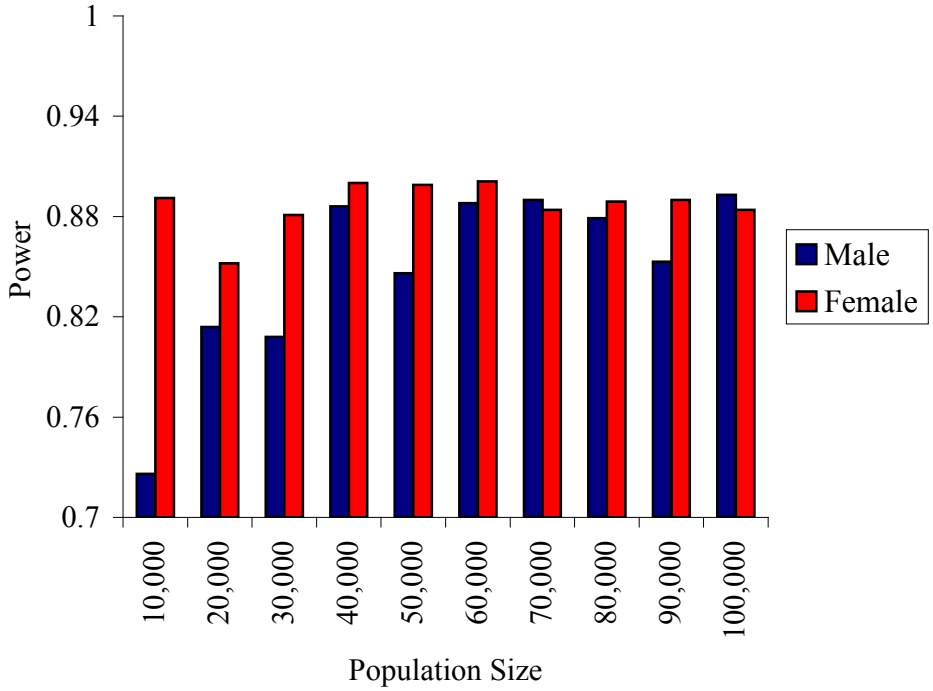


Figure 3. 7: Comparison of power of SIR when α=0.01

Figure 3.7 displays the power of the SIR method when the significance level is set to 1%. The female group displays good power at this significance level. The power of the male group is fairly unsteady in comparison to the female group - at the lower population sizes the power is significantly lower than at the higher population size.

The results of the powers calculations, mirror that of the empirical size calculations. This underscores the important of size and power. The analysis reinforces the previous knowledge on the SIR. The SIR performs poorly when sparse data is involved, especially when expected numbers of cases are below or around 1. This analysis underscores the fact that there are often cases when a region may observe far more cases than expected which can produce a result which we cannot find a reason for.

3.5 Introduction to Spatial Scan simulation

The direction of the study now takes a different approach by considering an alternative scenario. Leukaemia incidence in children is a controversial and well-researched area so the section of the simulation will be structured to represent practical examples which would be likely to prompt a research interest. Likewise in the previous section the historical trends in incidence for Scotland are used. This time the population of interest is restricted to those boys and girls under 15, to represent childhood leukaemia cases. The rates are averaged over the ten-year period so that the chosen rates are more stable.

The breakdown of childhood leukaemia rates by gender is provided in Table 3.2.

Table 3.2: Rates of childhood leukaemia per 100,000 person-years at risk in Scotland in 1994-2003

Age	< 5	5-9	10-14
Male	84.2	36.4	19.7
Female	67.2	23.5	25.6

3.6 Spatial scan – Simulating under the null hypothesis

The form of the null hypothesis is such that it defines the state of no clustering. Essentially this means that the risk inside the scan window is not significantly different from the risk outside of the window.

The null hypothesis is the simplest model to define in the spatial scan scenario. To do this regions are set-up using Scotland's local government district [51]. There are 56 local government districts in Scotland. The first step is to calculate the number of childhood leukaemia cases that occur in each of these 56 areas. The spatial scan methods used are purely based upon observed and expected numbers so the locations of these cases do not need to be determined. Instead we are only interested in the number of cases that occur within each of these local government districts, the point of reference is the population-centroid that is calculated by finding the weighted centre point of the population-at-risk.

Similarly to the first scenario of the study, the Poisson distribution is used to calculate the observed number of cases within each local government region. To mimic the conditions of leukaemia incidence, we simulate the cases based upon their expected numbers. In order to calculate these expected numbers, we use data on the age, sex and grid reference of the cases which we observe.

3.6.1 Empirical size

By simulating the null hypothesis using the expected number, we look to see how many times we observe a type I error. Due to limitations in computing time of the method, this type I error is observed over 100 simulations and the population at risk is combined by age and sex, giving us a single expected number for each region. These restrictions may also limit the conclusions which can be drawn from this study since due to the small number of simulations used and may not be enough to draw meaningful conclusions

about the data. In this analysis, small differences between the methods may be down to chance, where as large variations between methods are more likely to be real differences.

Taking both the observed and expected number of cases for each local government district, we calculate the empirical size for both the circular spatial scan statistic and the flexible spatial scan statistic.

Table 3.3: Empirical size of the spatial scan statistic when using the circular spatial scan

Significance Level	10%	5%	1%
Empirical Size	0.11	0.06	0.01

Table 3.4: Empirical size of the spatial scan statistic when using the flexible spatial scan

Significance Level	10%	5%	1%
Empirical Size	0.08	0.05	0.01

Table 3.3 shows the empirical size results of the circular spatial scan method when applied to the childhood leukaemia data. Looking at the figures, the empirical size calculations look to be around what would be expected. At the 10% significance level there is only one more case that we would expect, which similarly is the case at the 5% significance level. At the 1% level, we get exactly the same, as we would expect to observe as a rule of thumb. However if we observed one more case then we would be observing double the number of cases that we would expect to observe. This highlights

the main problem with using 100 simulations. From these results, we can conclude that the size of the circular spatial scan method is seemingly stable.

The empirical size results of the flexible spatial scan method are displayed in Table 3.4. On investigation, the empirical size looks steady. At the 10% significance level we observe two less type I errors than we would expect where as the 5% and 1% significance levels give us results which we would expect.

3.7 Spatial scan - Simulating under the alternative hypothesis

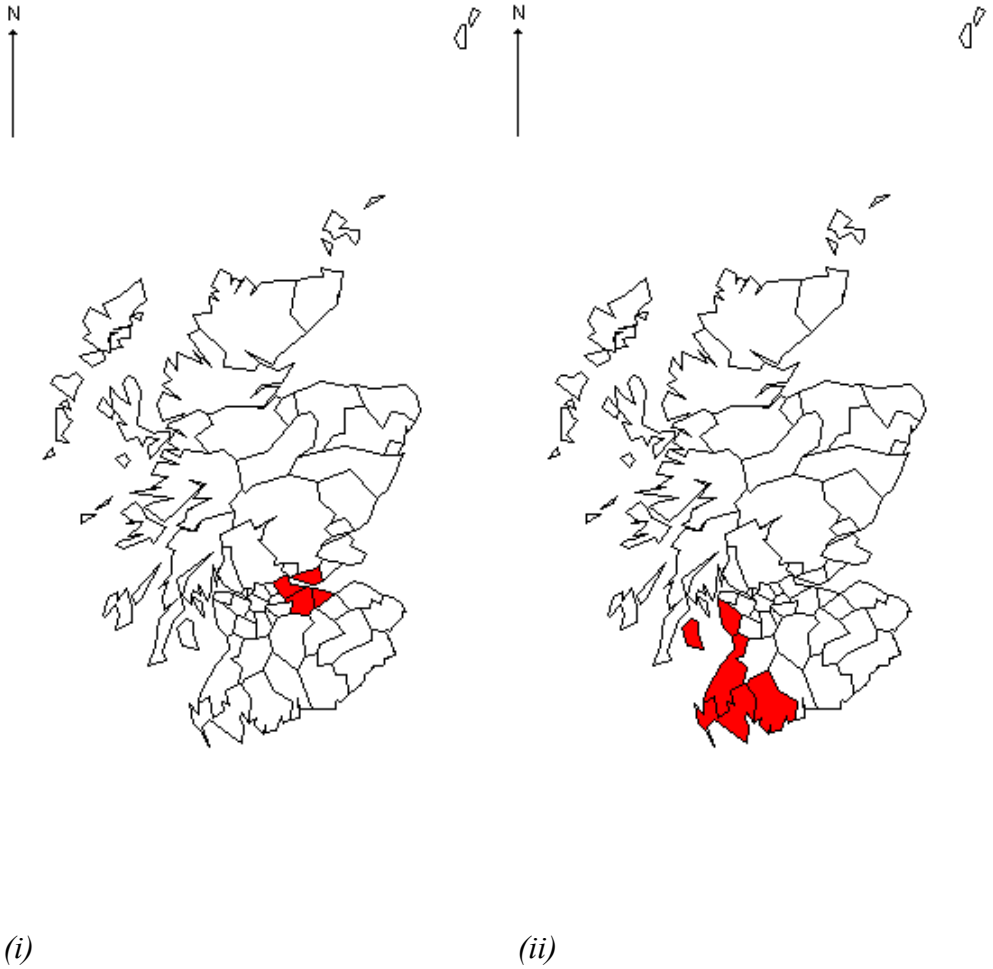
The alternative hypothesis of the spatial scan statistic will take two forms. The reason behind this is because of the two scan statistic methods which are being used, namely the circular and flexible spatial scans. It is important that specialised scenarios are identified to assess their performance. The form each alternative hypothesis will take is that four local government districts will take the form of a cluster. The relative risk in each of these four areas is set to be 2. Each of the local government districts in the cluster must be connected to at least one another local government district within the cluster.

The alternative hypotheses are set-up as in Figure 3.8. The local government regions included in the clusters are as follows:

Circular Cluster *(i)*: Falkirk, Dunfermline, Edinburgh City and West Lothian

Flexible Cluster *(ii)*: Stewartry, Wigtown, Cunninghame and Kyle & Carrick

Looking at Figure 3.8 gives a feel for how the clusters are set-up. Cluster *(i)* is closely representative of a circular cluster since the local government districts are set-up in a manner that represents a circular shape. Cluster *(ii)* on the other hand does not exhibit the characteristics of a circular cluster; instead it has been set-up to cover the south-west coast line of Scotland. A flexible cluster such as this one could emerge from the presence of a risk-factor on the coastline, which could possibly explain increased incidence along or around this area. Another applicable example of a cluster of this form is if the water from a river running through the country was contaminated and this was a risk-factor responsible for raising the incidence of disease in areas adjacent to the river.



(i) *(ii)*

Figure 3.8: Alternative hypothesis defined for a circular and flexible cluster

3.7.1 Power

In this section, the results of the circular scan statistic are compared against the results of the flexible spatial scan to highlight the advantages and disadvantages of each method. We now look to see the probability that each of the methods have of rejecting the null hypothesis of no clustering when there is a cluster present, either *(i)* or *(ii)*. Due to the limitations in computing power, 100 simulations are used. When investigating the results

small differences are likely to be due to chance and large differences are more likely to be true variations.

Firstly we investigate the circular cluster. Table 3.5 contains the results of the power calculations when both the circular and flexible spatial scan statistic methods are used.

Table 3.5: Results of power calculations for the circular cluster

Circular Spatial Scan Method

Significance Level	10%	5%	1%
Power	0.99	0.98	0.96

Flexible Spatial Scan Method

Significance Level	10%	5%	1%
Power	0.98	0.98	0.94

The results show that the power is very high for both the circular and flexible spatial scan methods when carried out upon the data with the circular cluster as defined in *(i)*. A difference in the results occurs at the 10% level, where the circular spatial scan method has a power of 0.99 with the flexible spatial scan method yielding a power of 0.98. The only other difference is at the 1% significance level where the circular spatial scan method yields a slightly higher power of 0.96 compared to the flexible spatial scans power of 0.94. The difference in power between the circular and flexible scan are very small, which indicates that they are likely to have occurred through chance.

The power is a good indication of how well the method is performing, however it does not give us any indications of the type or size of the significant cluster in each simulation. Table 3.6 details the proportion of times the areas from the 'true' cluster were actually part of the most likely cluster according to both the circular and flexible spatial scan methods.

Table 3.6: Proportion of times the MLC contains a true district from the circular cluster

Area	Circular Method	Flexible Method
Falkirk	0.93	0.83
Dunfermline	0.95	0.65
Edinburgh City	0.95	0.93
West Lothian	0.97	0.94

Table 3.6 gives us more of an idea of how the actual method is performing. Although the flexible spatial scan method was yielding very similar power to the circular spatial scan, it is clearer that the areas with the raised relative risk *i.e.* the districts from the true cluster, were not part of the significant cluster as much for the flexible method than they were the circular method. This helps provide evidence that the circular spatial scan statistic is a better at picking up areas which are closely joined together and represent a circular shape.

Now we examine the results of the power calculations for the alternative hypothesis as describes in *(ii)*. Table 3.7 displays the results of the power comparisons.

Table 3.7: Results of power calculations for the flexible cluster

Circular Spatial Scan Method

Significance Level	10%	5%	1%
Power	0.81	0.68	0.50

Flexible Spatial Scan Method

Significance Level	10%	5%	1%
Power	0.82	0.71	0.57

For the alternative hypothesis *(ii)* the power is lower than the power achieved under the alternative hypothesis *(i)*. However the flexible spatial scan method yields higher power than the circular spatial scan at all three significance levels, 10%, 5% and 1%, albeit, just slightly higher power. Although the power is not as high, it is fairly good at the 10% significance level. Again the differences between both methods are fairy negligible and are likely to be down to chance rather than being a true difference.

Table 3.8 outlines how often we observe a district from the defined cluster in each of our simulations.

Table 3.8: Proportion of times the MLC contains a true district from the flexible cluster

Area	Circular Method	Flexible Method
Stewarty	0.36	0.63
Wigtown	0.36	0.67
Cunninghame	0.73	0.92
Kyle & Carrick	0.80	0.95

The results in Table 3.8 help us investigate the findings of the previous power calculations. The circular spatial scan statistic does not have a high proportion of the

districts from the true cluster *(ii)* appearing in the most likely cluster. For the districts of Stewartry and Wigtown, these areas appear almost two times more often when the flexible spatial scan method is used compared with the circular spatial scan. The flexible spatial scan has a higher percentage of true hot spot districts in it, and has slightly greater power when the cluster in *(ii)* is used, providing more support to the findings that the flexible spatial scan is a better choice of method to pick up non-circular clusters.

3.8 Conclusions

In analysing the performance of the SIR in different scenarios, the strengths and weaknesses that were discussed in Chapter 2 can be further highlighted. In the instance where the population was low, therefore the expected rates were low, the SIR proved to be very unstable. This underscores the fact that when dealing with sparse data, the performance of the SIR can be very volatile. These weaknesses lead us to consider different techniques. However it must be noted that when the data is dense, therefore expected numbers were high, the SIR mostly yielded the correct results, indicating that in this scenario the technique performs very well.

Both spatial scan techniques yielded high power, indicating that the technique performs well. The circular and flexible scan statistics both performed well regardless of the shape of the cluster. However, the choice of scan used is very important, since the circular scan was poor in correctly pinpointing the correct cluster hotspots, when the cluster took a non-circular shape.

Chapter 4

Mortality of breast and colon cancer in Scotland

4.1 Epidemiology of breast cancer

The most common cancer for women in the United Kingdom is breast cancer. Around 44,000 cases of breast cancer occur each year. The lifetime risk of breast cancer amongst women is one in nine and it accounts for one-third of all cases of cancer in women [53].

Breast cancer can develop in the milk-producing glands within the breasts or from the passages from which milk is delivered to the nipples. It can spread to the surrounding tissues as well as other body parts. However due to earlier detection and improved treatments for breast cancer, the death rates in the United Kingdom have fallen by one-fifth in the last decade. Although breast cancer is predominantly a female cancer, it is still found in males but cases are very rare. Each year in the United Kingdom there are roughly 300 cases diagnosed.

With breast cancer there are a number of risk factors which can increase the chance of a patient being diagnosed with the disease. The following factors affect the risk of breast cancer:

- **Age** – This is the main risk factor for breast cancer which is not gender-specific is age. Breast cancer is more commonly found in older women and is considered rare in woman under 30

- **Children** – If a woman has more children then this lowers her risk of breast cancer, the younger she has these children will also decrease the risk.

- **Menstruation and menopause** – Women that start their periods early or have a late menopause are subject to an increased risk.

- **Contraceptive pill** – Women who take the pill have a slightly raised risk of breast cancer however this risk returns to normal once a woman stops taking it.

- **Hormone replacement therapy (HRT)** – Women undergoing HRT are at an increased risk. This risk increases if they are taking combined oestrogen/progestagen HRT and the longer they take it. This risk goes back to normal around 5 years after the treatment stops.

- **Weight** – If a woman is overweight once she has gone through the menopause then she is at an increased risk due to body fat affecting the level of oestrogen in her body. A balanced diet and regular exercise can help a woman keep a health body weight.

- **Alcohol** – A woman who drinks alcohol every day will increase her risk of breast cancer. If a woman drinks more every day then the risk will get greater.

- **Family** – If one or more of a woman's close family have had breast cancer then the woman's risk of being diagnosed is greater.

- **Breastfeeding** – The only recognised risk factor in reducing the risk of being diagnosed with breast cancer is breastfeeding. A woman who breastfeeds her children for longer will lower her risk of being diagnosed.

With breast cancer being a common disease it is not unusual for other members of a family to be diagnosed with breast cancer through chance. However there are cases of breast cancer being hereditary due to a faulty gene, although only a small amount of breast cancer is thought to be attributable to this.

4.2 Analysis of breast cancer mortality in Scotland 1986-1995

4.2.1 Introduction

Breast cancer mortality data in Scotland for the years 1986-1995 were obtained from the General Register Office for Scotland for analysis. The data included information on deaths due to breast cancer in Scotland for both men and women between the ages of 20-39 years. In this time period there were 382 deaths from breast cancer in this age group

in Scotland. Two of the cases were male deaths. Due to the main risk factor associated with breast cancer being gender specific, the male cases were removed from the data. The number of cases is so small it cannot merit an analysis on its own right. The individual cases had data on the date of death, the age of death, the cause of death given by ICD 9 coding and the output area where this death occurred where the output areas correspond to the postcode sectors of the 1991 census geography, which were linked to the output areas corresponding local government distict. There are 5 cases of breast cancer where the census output area is not defined. For the purposes of summary statistics these cases were included, however they were excluded from all other analysis.

Table 4.1: Cases of breast cancer mortality for females aged 20-39 years in Scotland 1986-1995

		Age group (years)				
		20-24	25-29	30-34	35-39	Total
Year	1986	1	0	9	26	36
	1987	0	1	11	22	34
	1988	0	5	7	22	34
	1989	0	2	8	24	34
	1990	0	6	11	25	42
	1991	0	2	17	21	40
	1992	1	5	15	19	40
	1993	0	3	13	24	40
	1994	0	1	15	25	41
	1995	0	3	12	24	39
	Total	2	28	118	232	380

The data published on the numbers of breast cancer deaths by the NHS are broken up into age groups of five-year age groups. The breast cancer mortality in Scotland in women aged between 20-39 years is displayed in Figure 4.1.

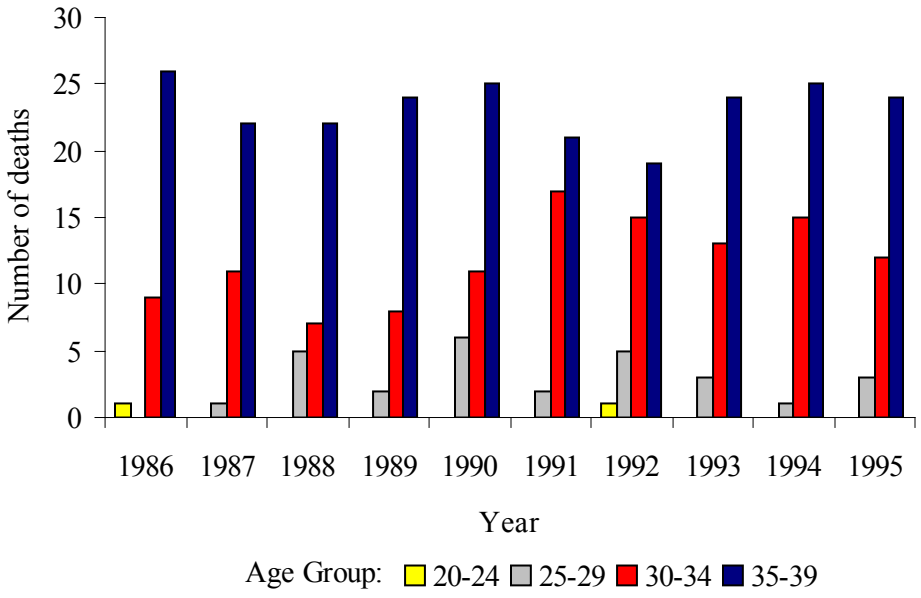


Figure 4.1: Trend of breast cancer mortality in females aged 20-39 years in Scotland 1986-1995

The breast cancer mortality in females between 20-24 years has 2 cases over a 10-year period. The number of deaths due to breast cancer in females aged 25-29 years starts to rise as the time period advances. In general the number of cases fluctuate however the changes seem to be due to random variation. Mortality in females aged 30-34 years is greater than the previous age group. The number of deaths due to breast cancer varies over the years however there is a very slight increase in the number of deaths as a general trend. For females between the ages 35-39 years there are a lot more cases. These fluctuate like the other groups and no general trend is apparent. It can be seen that age is a risk factor in Figure 4.1 since the cases are more prevalent in the older age groups.

4.2.2 Methodology

The rationale behind investigating the breast cancer mortality between females aged between 20-39 years is that it has been suggested that younger breast cancer patients have a poorer survival rate. Although the incidence is far higher in older women, breast cancer in younger women is said to be more aggressive and is more deadly in younger women [54]. By looking at a disease map for the whole of Scotland it will pinpoint the fact that different areas are exposed to different risk factors. By considering that there is a hereditary nature in breast cancer mortality in younger women we would expect any excess in cases to occur closed to each other due to families' members mostly living in similar districts.

In order to analyse the data the expected numbers are needed. To calculate these expected numbers, the age-standardised mortality rates for Scotland are used. These rates are adjusted in proportion to the population-at-risk within each local government district. For each of the 56 local government districts we now have the expected numbers of deaths due to breast cancer. The first step in our analysis is to calculate the SMR for the 56 local government districts and the 12 local government regions. Once this has been carried out then a spatial scan analysis of the data will be carried out, using both a circular and flexible shaped scan window. Finally the Besag and Newell test statistic will be carried out to add to any exploratory findings from the data.

4.2.3 Results of analysis

To gain further information into the data the disease map is plotted in Figure 4.2. Plotting the SMR for each district provides a useful insight on how breast cancer mortality is varying across the local government regions and more importantly Scotland as a whole. Without considering statistical significance and just focusing purely on the SMR, many inferences can be made about the way the mortality of breast cancer varies across the country. The north-west area of the Highlands and Orkney appear to have an excess of mortality due to breast cancer. The rest of the Highland districts do not exhibit this trend. In the Highland districts the population is sparse so there may be some difficulties in interpreting the SMR due to small numbers.

The north-east area of Scotland, notably the Grampian region has a level of mortality greater than is expected. This trend continues down into the Lothian region and is displayed most noticeably across central Scotland. There are some districts in the south that also display this behaviour but most of the districts surrounding this area are implying there is less deaths due to breast cancer than we expect so there is not much more information that we can gather from this.

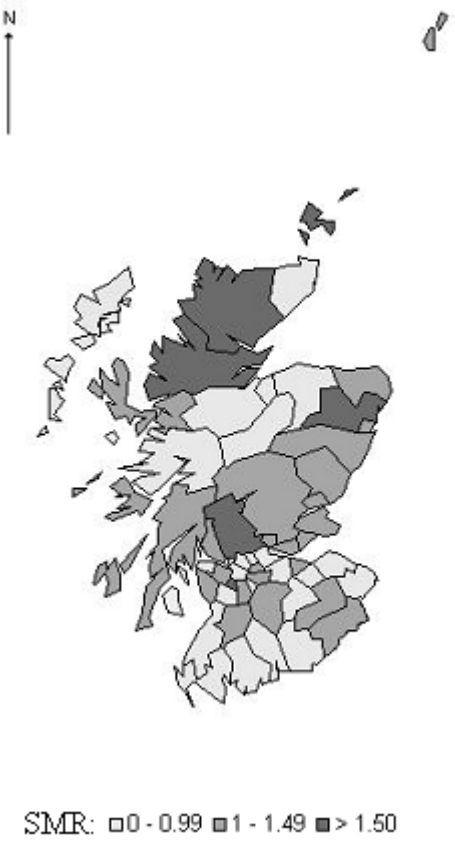


Figure 4.2: Disease map of breast cancer mortality in females in Scotland 1986-1995

The ideas gathered from analysing the disease map can now be formally tested by considering the statistical significance of the SMR for each district. SMR results for each of the 56 local government districts, as well as for each of the 11 local government regions, are given in Table 4.2.

Table 4.2: Results of SMR analysis for breast cancer mortality in females in Scotland 1986-1995

Local Government District	Observed (O_i)	Expected (E_i)	SMR	p-value
Borders	7	7.10	0.99	0.42
Berwickshire	1	1.26	0.79	0.36
Ettrick & Lauderdale	3	2.40	1.25	0.22
Roxburgh	3	2.39	1.25	0.22
Tweeddale	0	1.05	0	0.65
Central	22	19.59	1.12	0.25
Clackmannan	5	3.50	1.43	0.14
Falkirk	8	10.44	0.77	0.71
Stirling*	9	5.64	1.59	0.06
Dumfries & Galloway	4	10.31	0.39	0.98
Annandale & Eskdale	0	2.61	0	0.93
Nithsdale	2	4.17	0.48	0.79
Stewartry	1	1.51	0.66	0.44
Wigtown	1	2.02	0.50	0.60
Fife	29	24.95	1.16	0.18
Dunfermline	12	9.54	1.26	0.17
Kirkcaldy	11	10.77	1.02	0.39
North East Fife	6	4.63	1.30	0.19
*Grampian***	52	38.87	1.34	0.02
Aberdeen City	20	15.56	1.29	0.11
Banff & Buchan*	9	6.05	1.49	0.09
Gordon**	12	6.91	1.74	0.02
Kincardine & Deeside	6	4.42	1.36	0.16
Moray	5	5.92	0.84	0.54
Highland	12	14.93	0.80	0.73
Badenoch & Strathspey	0	0.75	0	0.53
Caithness	0	1.83	0	0.84
Inverness	2	5.03	0.40	0.88
Lochaber	0	1.43	0	0.76
Nairn	0	0.72	0	0.51
Ross & Cromarty**	7	3.54	1.98	0.03
Skye & Lochalsh	1	0.82	1.22	0.20
Sutherland**	2	0.81	2.46	0.05
Lothian	55	56.87	0.97	0.56
East Lothian	5	6.38	0.78	0.61
Edinburgh City*	40	32.49	1.23	0.08
Midlothian	4	6.19	0.65	0.74
West Lothian	6	11.81	0.51	0.95
Strathclyde	160	167.57	0.95	0.70
Argyll & Bute	5	4.39	1.14	0.28
Bearsden & Milngavie	1	3.02	0.33	0.80
Clydebank	1	3.17	0.32	0.83
Cumbernauld & Kilsyth	4	5.03	0.79	0.57
Cumnock & Doon Valley	4	3.01	1.33	0.19
Cunninghame	13	10.10	1.29	0.14
Dumbarton	6	6.03	1.00	0.40
East Kilbride	9	6.56	1.37	0.13
Eastwood*	8	5.03	1.59	0.07
Glasgow City	37	48.32	0.77	0.94
Hamilton	8	8.20	0.98	0.44
Inverclyde	3	6.31	0.48	0.87
Kilmarnock & Loudoun	3	6.12	0.49	0.86
Kyle & Carrick	6	8.20	0.73	0.71
Clydesdale	5	4.51	1.11	0.30
Monklands	10	7.59	1.32	0.15
Motherwell*	15	10.33	1.45	0.06
Renfrew	18	14.81	1.22	0.17
Strathkelvin	4	6.83	0.59	0.81
Tayside	31	27.52	1.13	0.22
Angus	7	6.82	1.03	0.37
Dundee City	12	11.82	1.02	0.40
Perth & Kinross	12	8.88	1.35	0.12
Orkney	2	1.32	1.52	0.15
Shetland	2	1.67	1.19	0.24
Western Isles	1	1.75	0.57	0.52

4.2.4 Discussion

Using the SMR method there are a significant excess risk of mortality at both the local government district level and local government region at the 5% and 10% significance levels. Those districts which produced significant results at the 10% significance level but not at the 5% significance level were Stirling, Banff & Buchan, Edinburgh City, Eastwood and Motherwell.

At the 5% level the most statistically significant results are at the districts Gordon, Ross & Cromarty and Sutherland. The local government region of Grampian had a significant excess risk at the 5% level. These results are very much like was expected from examining the disease map. Although the disease map did not provide a definitive response to any concerns, it is useful to get a visual representation of the geographic variation.

Firstly we look to the Highland region where there are two regions which the analysis suggests have an excess risk of breast cancer mortality at the 5% significance level. Looking to the plot we see Ross & Cromarty and Sutherland areas are neighbouring districts. For the Sutherland district there are only 2 deaths due to breast cancer in the ten-year period which is over two times what we expect. This excess risk could just be down to chance due to the small amount of deaths involved. However since it is a neighbouring district of Ross & Cromarty, which has almost double the amount of deaths

than it is expected to have, there could possible be an excess risk of death due to breast cancer in females aged 20-39 years.

Now we look to the Grampian region. Banff & Buchan and Gordon display a significant excess risk of mortality at the 10% and 5% significance levels respectively. Although the Grampian region contains three other districts which do not display an excess risk of mortality, there are two of them which have an SMR greater than 1. Overall the results from the analysis suggest that there is an excess risk of breast cancer mortality in females between the ages of 20-29.

The other 3 districts which are significant at the 10% level are from different local government regions. Stirling, Edinburgh City, Eastwood and Motherwell all have a significant excess risk of mortality however they are located close to or in the central belt of Scotland. Although they are from different regions, geographically they are located a great distance from one another.

In order to fully analyse the data for a possible elevated risk of death due to breast cancer it is necessary that we look to further investigate the possibility of one or more mortality clusters. To do so, we carry out the circular and flexible spatial scan. Summary results of this analysis are provided below in Table 4.3. More detailed results of this analysis are provided in Appendix B.

Table 4.3: Spatial Scan analysis of Breast Cancer Mortality

	Circular Scan	Flexible Scan
Overall Relative Risk	1.20	1.41
p-value	0.37	0.63

The results of the analysis did not find any significant most likely clusters. The most likely cluster which was found using the circular spatial scan had a *p*-value of 0.374. Using the flexible spatial scan the most likely cluster had a *p*-value of 0.627.

Recalling that the Besag and Newell test statistic is useful if the basic size and shape of a cluster is known, we can now call upon that. The advantage in setting the size of the cluster prior to the analysis is that we can set it in relation to the number of deaths observed within a region, or a combination of districts. However the application of the Besag and Newell test statistic after the SMR analysis raises post-hoc testing issues. Due to this, the results of the analysis are considered to add to investigative findings, rather than definitive conclusions.

First of all we consider the Highland region. In this region the SMR analysis suggests that there may be an excess risk of mortality in both Sutherland and Ross & Cromarty. Since these are neighbouring districts we look to see if there could be a possible clustering of deaths. Testing using the Besag and Newell method for a cluster of size 9 yields a *p*-value of 0.079 which is statistically significant at the 10% level. Accumulating the cases for both regions gives 9 observed deaths where only 4.35 were expected. The evidence of a significant cluster at the 10% level, made up of Sutherland and Ross &

Cromarty, gives us an indication that there may be an excess risk of mortality within this sub-region of the Highlands.

Now we consider the Grampian local government region. The analysis appeared to suggest that living in this region during the study time period put females aged 20-39 years at a raised risk of mortality of breast cancer. We look for a cluster of 52 cases, which is the amount of deaths observed in the region. Whilst 52 deaths are observed there are only 38.86 deaths expected in the Grampian region. Testing using the Besag and Newell method results in a p-value of 0.042 for the Grampian region which is significant at the 5% level. This may suggest that there is evidence of clustering of breast cancer mortality in the Grampian region. The geographical location of the clusters are represented by Figure 4.3.

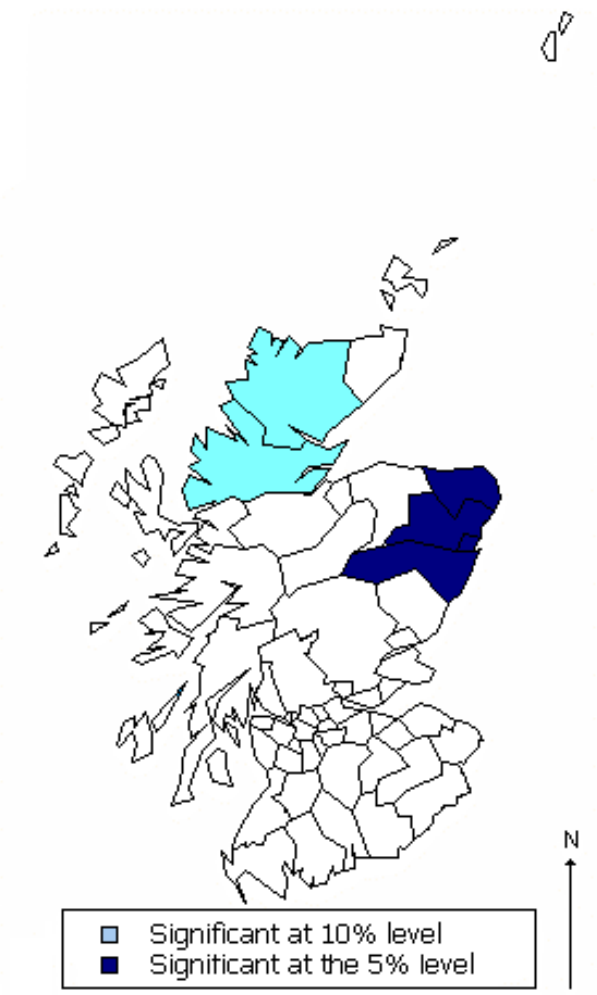


Figure 4.3: Clusters of breast cancer mortality in females aged 20-29 in Scotland 1986-1995

Overall the analysis has suggested that there may be a possibility of a cluster of breast cancer mortality amongst females aged 20-29 from 1986-1995 in a small sub-section of the Highlands (Sutherland and Ross & Cromarty). There may also be evidence of a cluster of breast cancer mortality in the Grampian region, where the most significance cluster is made up of the Aberdeen City, Banff & Buchan, Gordon, Moray and Kircardine districts. Although the tests have provided some evidence to suggest this, neither of the

scan statistics showed displayed any evidence of clustering. Another issue to consider is multiple testing, whereby some false positive results may occur by chance.

4.3 Epidemiology of colon cancer

Cancer of the colon is more commonly known as bowel cancer. It is a very common cancer, being the third most common in men and the second in females in the United Kingdom. Colon cancer accounts for around 13% of all cancers with around 21,617 and 13.389 new cases in men and women respectively each year in the UK [54].

This form of cancer occurs in the colon and takes around five to eight years to develop. The first stage is often a small growth, called a polyp or adenoma, which grows on the wall of the bowel. Colon cancer can spread to other body parts, most frequently the liver.

The factors which can result in a raised risk of colon cancer are outlined below:

- **Age** – Bowel cancer is most frequent in older people and as you become older the risk of developing this cancer increases. Around eight out of ten cases of are people that are over 60.

- **Weight** – Overweight or obese people are at a higher risk.

- **Physical ability** – Those who do not exercise and are inactive are at a higher risk than those who don't. Moderate exercising can aid to lower the risk of the disease.

- **Diet** – Diets which are high in red or processed meat and fat and low in fruit, fibre, folate and vegetables can be at a higher risk

- **Smoking and alcohol** – Those who smoke may be at an increased risk, especially those who are also heavy drinkers. The consumption of alcohol can possibly lead to a raised risk, especially in people that have low levels of folate in their diet.

- **Previous diagnosis** – Those that have previously been diagnosed with bowel cancer or people that have had a polyp in their bowel are at a raised risk. If the polyp was an adenomatous polyp then it is possible that there is another increase in the risk.

- **Crohn's disease** – Anyone that has had Crohn's disease may have a small increased risk.

- **Family** - Colon cancer can run in families. Some people inherit a faulty gene from one of their parents which puts them at a higher risk than normal. Families that contain this gene usually have a history of bowel cancer within their family.

4.4 Analysis of colon cancer mortality in Scotland 1986-1995

4.4.1 Introduction

Data on colon cancer mortality were gathered from the GRO Scotland. The time period of the study is from 1986-1995. The information obtained was deaths due to colon cancer for males and females aged between 15-39 years. There were 104 deaths due to colon cancer in Scotland between 1986 and 1995. Of these 104 deaths, 53 of these were males and 51 of them females. For each case, data were provided on the date of death, the age of death, the cause of death and the output area which the death occurred in, which was then matched to the output areas corresponding local government district.

The data published regarding numbers of colon cancer deaths were gathered from ISD Scotland, where the mortality is broken up into age groups of five year gaps. A breakdown of the difference in mortality trends between both genders in the study time period is displayed in Tables 4.4 and 4.5.

Table 4.4: Cases of colon cancer mortality for males aged 15-39 years in Scotland 1986-1995

		Age group					
		15-19	20-24	25-29	30-34	35-39	Total
Year	1986	0	0	1	1	2	4
	1987	0	2	0	1	5	8
	1988	0	0	0	1	4	5
	1989	0	0	0	1	0	1
	1990	1	0	2	2	3	8
	1991	0	0	0	0	4	4
	1992	0	2	0	1	4	7
	1993	0	0	0	1	3	4
	1994	0	0	0	3	2	5
	1995	0	0	1	2	4	7
	Total	1	4	4	13	31	53

Table 4.5: Cases of colon cancer mortality for females aged 15-39 years in Scotland 1986-1995

		Age group					
		15-19	20-24	25-29	30-34	35-39	Total
Year	1986	0	0	0	1	1	2
	1987	0	0	1	0	6	7
	1988	0	0	1	1	1	3
	1989	0	1	1	2	3	7
	1990	1	0	1	0	1	3
	1991	1	0	0	0	2	3
	1992	0	1	1	1	4	7
	1993	1	2	0	0	4	7
	1994	0	0	0	1	7	8
	1995	0	0	0	3	1	4
	Total	3	4	5	9	30	51

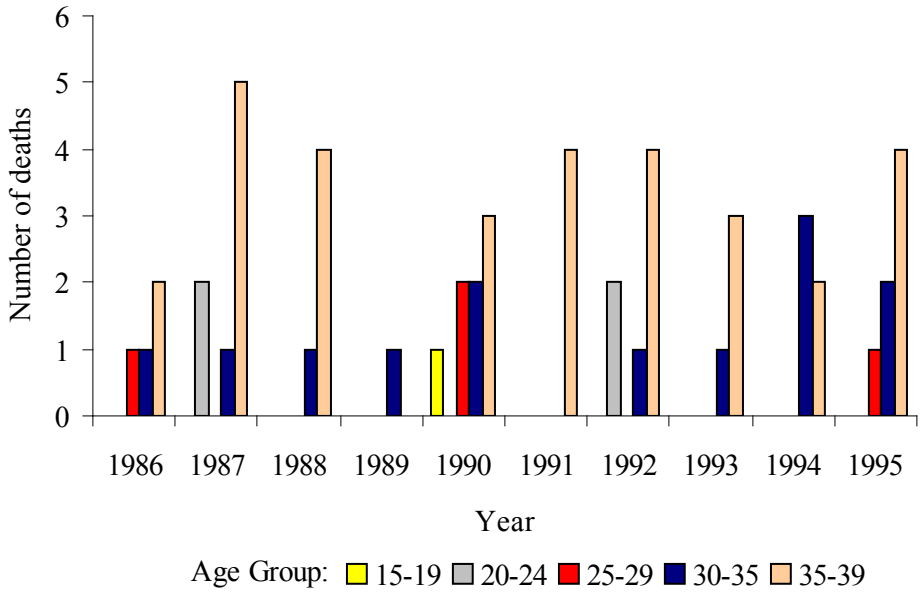


Figure 4.4: Trend of colon cancer mortality in males aged 15-39 years in Scotland 1986-1995

There are 53 male deaths due to colon cancer in males aged 15-39 years of age. Over the ten year period there was only 1 death due to colon cancer in males aged 15-19 years. This number rises to 4 deaths in both the 20-24 and 25-29 years age groups. There is a rising number of deaths as the age groups get older confirming that colon cancer is more frequent in older males, which can be seen in Figure 4.4.

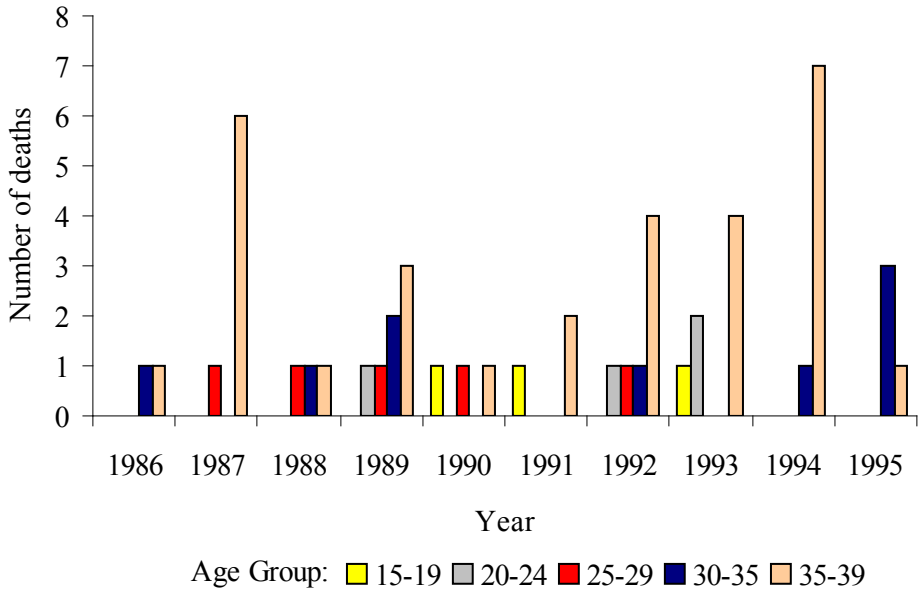


Figure 4.5: Trend of colon cancer mortality in females aged 15-39 years in Scotland 1986-1995

Over the 10-year study period there are 51 colon cancer deaths in females aged 15-39 years. Figure 4.5 shows that the data follows the same trend as the male deaths in that colon cancer mortality is more frequent in older age groups than it is in younger age groups.

4.4.2 Methodology

The analysis is an investigation into colon cancer mortality in both males and females aged 15-39 years in Scotland. Since the aforementioned risk factors were not gender specific the sex of the individual is not an issue so it can be omitted from the analysis. Just like the previous breast cancer mortality analysis, we would expect any excess in

mortality of colon cancer to occur close together representing that family members live close by to one another.

Before analysing the data, expected numbers were needed. These were calculated using the age-standardised mortality rates for Scotland. The rates were then adjusted to represent the proportion of the population-at-risk living in each local government region. For each local government district, data are now available for the numbers of deaths observed and the number of deaths that were expected due to colon cancer in the population-at-risk. Firstly we calculate the SMR for each of the local government districts and the local government regions. A spatial scan analysis of the data will then be carried out, using both a circular and flexible shaped scan window. The final step is to apply the Besag and Newell test statistic that will allow us to add to any exploratory findings.

4.4.3 Results of analysis

Before looking to formally test the results of the SMR analysis, the disease map of colon cancer mortality is plotted in Figure 4.6. This gives us an idea on how colon cancer mortality is varying across Scotland as a whole and not just in each local government region. The most notable point is that there is an excess of colon cancer mortality which seems to cluster around the central to the south of Scotland. This excess spreads mostly across the Tayside, Central and Strathclyde regions of Scotland. There seems to also be an excess risk in the Western Isles, Caithness and Banff & Buchan.

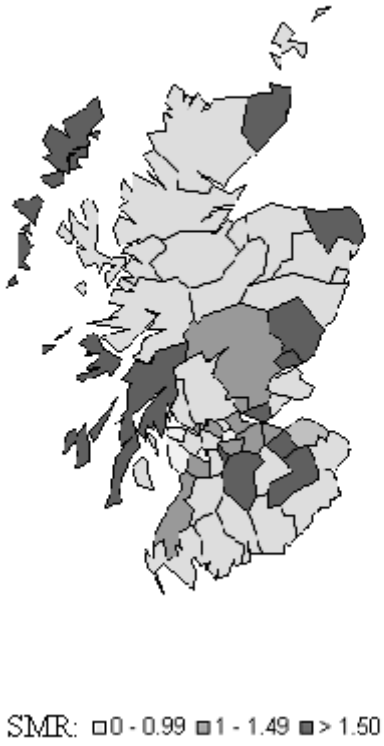


Figure 4.6: Disease map of colon cancer mortality in Scotland 1986-1995

These subjective impressions from investigating the disease map are now formally tested with the results displayed in Table 4.6 where the statistical significance of the SMR for each local government district and region is considered.

Table 4.6: Results of SMR analysis for colom cancer mortality in Scotland 1986-1995

Local government district	Observed (O_i)	Expected (E_i)	SMR	p-value
Borders	1	1.90	0.53	0.57
Berwickshire	0	0.33	0.00	0.28
Ettrick & Lauderdale	1	0.66	1.52	0.14
Roxburgh	0	0.63	0.00	0.47
Tweeddale	0	0.28	0.00	0.24
Central	6	5.11	1.17	0.25
Clackmannan	1	0.93	1.08	0.24
Falkirk	4	2.71	1.48	0.14
Stirling	1	1.47	0.68	0.43
Dumfries & Galloway	1	2.71	0.37	0.75
Annandale & Eskdale	0	0.69	0.00	0.50
Nithsdale	1	1.09	0.92	0.30
Stewartry	0	0.39	0.00	0.32
Wigtown	0	0.54	0.00	0.42
Fife	6	6.56	0.92	0.48
Dunfermline	4	2.51	1.59	0.11
Kirkcaldy	1	2.81	0.36	0.77
North East Fife	1	1.24	0.80	0.35
Grampian	8	10.36	0.77	0.71
Aberdeen City	0	4.23	0.00	0.99
Banff & Buchan ***	6	1.62	3.70	0.00
Gordon	1	1.78	0.56	0.53
Kincardine & Deeside	0	1.15	0.00	0.68
Moray	1	1.58	0.63	0.47
Highland	2	3.95	0.51	0.75
Badenoch & Strathspey	0	0.20	0.00	0.18
Caithness *	1	0.48	2.08	0.08
Inverness	1	1.30	0.77	0.37
Lochaber	0	0.38	0.00	0.31
Nairn	0	0.20	0.00	0.18
Ross & Cromarty	0	0.94	0.00	0.61
Skye & Lochalsh	0	0.22	0.00	0.20
Sutherland	0	0.23	0.00	0.20
Lothian *	20	14.87	1.35	0.08
East Lothian	1	1.66	0.60	0.49
Edinburgh City *	12	8.59	1.40	0.10
Midlothian *	3	1.58	1.90	0.08
West Lothian	4	3.05	1.31	0.19
Strathclyde	45	43.23	1.04	0.36
Argyll & Bute ***	5	1.15	4.33	0.00
Bearsden & Milngavie	0	0.77	0.00	0.54
Clydebank	1	0.80	1.25	0.19
Cumbernauld & Kilsyth	1	1.30	0.77	0.37
Cumnock &Doon Valley**	2	0.78	2.57	0.04
Cunninghame	0	2.59	0.00	0.92
Dumbarton	1	1.57	0.64	0.47
East Kilbride	1	1.73	0.58	0.52
Eastwood	1	1.26	0.79	0.36
Glasgow City	12	12.60	0.95	0.49
Hamilton *	4	2.07	1.93	0.06
Inverclyde	0	1.65	0.00	0.81
Kilmarnock & Loudoun	2	1.52	1.32	0.20
Kyle & Carrick	3	2.05	1.46	0.15
Clydesdale **	3	1.18	2.55	0.03
Monklands	1	1.97	0.51	0.58
Motherwell	4	2.68	1.49	0.13
Renfrew	2	3.83	0.52	0.74
Strathkelvin	2	1.73	1.16	0.25
Tayside ***	13	7.07	1.84	0.01
Angus **	4	1.79	2.23	0.04
Dundee City **	6	3.02	1.99	0.03
Perth & Kinross	3	2.26	1.33	0.19
Orkney	0	0.36	0.00	0.30
Shetland	0	0.47	0.00	0.38
Western Isles ***	2	0.50	3.98	0.01

4.4.4 Discussion

There is a significant excess risk of mortality at both the local government district and region level at the 1%, 5% and 10% significance level. The districts which produced significant results at the 1% significance level were Banff & Buchan and Argyll & Bute. The local government regions of Tayside and the Western Isles had an excess of mortality which was significant at the 1% level. At the 5% level the districts of Clydesdale, Angus, Dundee City and Cumnock & Doon Valley produced significant results. At the 10% significance level the excess of mortality was significant in the districts of Hamilton, Midlothian, Edinburgh City, Caithness and in the region of Lothian.

These results pair up with the impressions gathered from examining the disease map. Most of the districts and regions which had a significant excess risk of mortality were located in the central area of Scotland going south to the region of Strathclyde. The others which fell outwith these bounds were located towards the north of Scotland and concerns were already raised regarding these districts.

To further investigate the possibility of one or more mortality clusters, the circular and flexible spatial scan statistics are used on the data. Summary results of the analysis can be found in Table 4.7. More detailed results are provided in Appendix B.

Table 4.7: Spatial Scan analysis of Colon Cancer Mortality

	Circular Scan	Flexible Scan
Overall Relative Risk	4.06	1.45
p-value	0.46	0.73

The results of the spatial scan analysis did not find any significant most likely clusters. The most likely cluster which was found using the circular spatial scan had a *p*-value of 0.46. Using the flexible spatial scan the most likely cluster had a *p*-value of 0.731.

To go one step further it is necessary to investigate the possibility of a mortality cluster in the south-central area of Scotland. To do so the Besag and Newell test for clusters is carried out. Due to the uncertainty regarding a perceived cluster due to the number of regions exhibiting an potential excess being a lot larger, we will search for clusters of size k where $k = 1, \ldots, N$ where N is the total number of cases of colon cancer. Again it must be noted that the application of the Besag and Newell test statistic after the SMR analysis raises post-hoc testing issues, therefore any results are considered as exploratory rather than definitive.

Table 4.8: Significant clusters of colon cancer mortality in Scotland 1986-1995

Cluster centre	Observed	Expected	*p*-value
Kirkcaldy	55	41.69	0.095
	57	43.42	0.096
	61	45.82	0.072
North East Fife	44	32.15	0.088
East Lothian	59	44.72	0.086
	61	46.45	0.086
West Lothian	36	25.57	0.091
	37	26.5	0.094
	41	28.57	0.058
	43	30.58	0.067
	44	32.05	0.085
Angus	10	4.81	0.078

Table 4.8 contains a summary of the findings of analysing the data with the Besag and Newall test statistic. Testing around south-central Scotland has found many possible clusters. All the possible clusters are significant at the 10% level, indicating that there may be an indication of a colon cancer cluster in south-central Scotland. Looking to Figure 4.6, the disease map indicated that there was an excess risk of mortality around the south-central area of Scotland. The number of results in Table 4.8 indicates that there are many possible clusters in this area. This suggests that there may be possible evidence that there was a clustering of colon cancer mortality in south-central Scotland. Once again the numbers are very small during this study. Problems with small numbers and multiple testing prohibit these results from being underlined. This is because the use of these methods on small counts can be misleading due to unstable results. The caveats associated with the small numbers only allow us to view these findings as investigative rather than definitive.

Chapter 5

Summary and Further Research

5.1 Summary of Thesis

The purpose of this thesis was to introduce the issues that arise in the analysis of small-area data and to examine some commonly used methods in the analysis of small-area spatial health data. Firstly this was achieved by reviewing the methods, taking into account their strengths and weaknesses and noting the scenarios in which they performed both well and poorly in. In Chapter 1, the political, social and health issues that arise in small-area statistics were discussed, noting the fundamental concepts of spatial epidemiology.

In Chapter 2, the relative merits of five techniques were discussed, namely the Standardised Incidence Ratio, Besag and Newell Cluster Test, Circular Spatial Scan, Flexibly-Shaped Spatial Scan and Bithell's Linear Risk Score. The main emphasis was placed upon exploring the benefits of using the SIR as a method of detecting the risk of disease in small-areas, with the performance of the others being taken into account to provide a possible alternative to analysing small-area health data.

Chapter 3 extended on the theory and academic findings of each of the five methods through a detailed simulation study, helping underscore historical findings by using

empirical results as evidence. The simulation study, detailing the performance of the circular and flexibly-shaped spatial scan, confirmed the prior findings, however the limitations in computing power placed some uncertainties over the outcomes.

Chapter 4 rounded off the thesis with an analysis of the mortality of breast and colon cancer in Scotland for the ten-year time period 1986-1995. The analysis was carried out using the SIR method, the circular spatial scan and flexibly-shaped spatial scan, with the Besag and Newell method also being considered after the original analysis, due the method performing better when the size and scale of clustering is known. The findings of this analysis were that there might be evidence of a cluster of breast cancer mortality in the Grampian region. The post-hoc problems associated with using the Besag and Newell test statistic prohibits the results from being anything other than exploratory. The analysis also suggested evidence of clustering of colon cancer mortality in south-central Scotland.

5.2 Conclusions

From the thesis, many conclusions can be drawn. The SIR method is very useful as it provides a quick and basic summary of risk. Its performance in areas with small population or small numbers of cases is the main drawback of using the SIR. Using smoothing techniques, or constructing confidence intervals can help combat these problems, however there is always a difficulty faced when numbers are very small.

Methods such as the spatial cluster scan statistics and the Besag and Newell cluster tests are useful to get a general feeling around the data. Observations from the Besag and Newell are difficult to underscore due to the post-hoc application of this method.

Many more advances are being made to address the relationship between risk and distance from a point source, due to the political and social significance attached to this association. SAHSUs development of the RIF has proved a big advance in this field, due to the fact that it draws concentric circles and calculates the risk of the area inside the circle. The drawback of the older methods is that the cases are assumed to fall upon the population-weighted centroid. The reliance on the distance between centroids is a main drawback since any cases falling within the distance, but outwith a centroid would not be included.

At ISD Scotland, the SIR is the primary technique used in the basic analysis of small-area data. For a basic analysis of the data, the SIR method is very useful for illustrating how risk is varying amongst the entire study area. With any method there will always be some limitations, however at a basic level the caveats associated with the SIR method are not enough to justify scrapping the use of the method. There will be times where conclusions cannot be drawn from a purely SIR analysis and this is where ISD Scotland could rely on other methods.

ISD Scotland should consider the use of spatial scan statistics and cluster tests to drill down another level in the data, to try to draw more specific conclusions. Methods such

as the circular and flexible spatial scan statistics would be useful to ISD Scotland to help gain a deeper insight into risk within a study region. These techniques are powerful in detecting hot spot clusters, with the flexible spatial scan being able to identify clusters of any shape. If an SIR analysis starts to indicate that there is an excess risk of disease among similarly located areas, then the use of cluster tests can also help to validate or reject such suggestions. This would be very useful to ISD Scotland since there would be more concrete evidence to present to its customers to support or reject any claims, thus reinforcing any conclusions made.

Appendix A: Appendix for Chapter 2

A.1 Empirical Size calculations of SIR method

A.1.1 Size calculations for Male age-groups when α=0.1

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.127	0.195	0.136	0.152	0.160	0.131	0.118	0.126	0.129	0.111
	10-19	0.217	0.230	0.130	0.142	0.134	0.147	0.158	0.144	0.161	0.162
	20-29	0.401	0.268	0.190	0.159	0.123	0.110	0.178	0.135	0.109	0.172
	30-39	0.112	0.307	0.240	0.171	0.142	0.120	0.174	0.156	0.113	0.203
	40-49	0.138	0.254	0.134	0.129	0.130	0.131	0.145	0.103	0.127	0.135
	50-59	0.111	0.161	0.119	0.137	0.141	0.092	0.141	0.143	0.123	0.110
	60-69	0.150	0.108	0.107	0.116	0.130	0.127	0.109	0.096	0.124	0.087
	70-79	0.140	0.120	0.110	0.125	0.118	0.103	0.131	0.127	0.104	0.094
	80 +	0.142	0.123	0.128	0.099	0.110	0.103	0.114	0.095	0.106	0.110

A.1.2 Size calculations for Female age-groups when α=0.1

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.139	0.149	0.161	0.134	0.131	0.116	0.105	0.165	0.185	0.137
	10-19	0.133	0.130	0.145	0.111	0.106	0.155	0.172	0.147	0.130	0.126
	20-29	0.323	0.190	0.121	0.210	0.132	0.107	0.162	0.189	0.171	0.113
	30-39	0.338	0.194	0.123	0.216	0.151	0.111	0.137	0.114	0.157	0.098
	40-49	0.295	0.198	0.105	0.147	0.172	0.132	0.132	0.165	0.105	0.144
	50-59	0.110	0.146	0.144	0.167	0.100	0.102	0.128	0.123	0.131	0.133
	60-69	0.107	0.125	0.121	0.161	0.133	0.112	0.108	0.098	0.125	0.118
	70-79	0.089	0.133	0.135	0.116	0.101	0.118	0.125	0.107	0.102	0.122
	80 +	0.114	0.116	0.109	0.119	0.109	0.099	0.100	0.101	0.111	0.120

A.1.3 Size calculations for Male age-groups when α=0.05

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.127	0.095	0.074	0.087	0.057	0.066	0.078	0.052	0.052	0.071
	10-19	0.060	0.088	0.130	0.061	0.079	0.055	0.092	0.084	0.076	0.055
	20-29	0.088	0.087	0.067	0.058	0.123	0.056	0.081	0.067	0.109	0.111
	30-39	0.112	0.091	0.089	0.060	0.058	0.067	0.086	0.070	0.054	0.119
	40-49	0.138	0.108	0.064	0.067	0.063	0.063	0.088	0.057	0.074	0.061
	50-59	0.043	0.094	0.064	0.085	0.060	0.053	0.073	0.076	0.091	0.056
	60-69	0.070	0.077	0.073	0.053	0.073	0.066	0.060	0.058	0.072	0.043
	70-79	0.059	0.068	0.050	0.068	0.059	0.055	0.068	0.061	0.043	0.047
	80 +	0.078	0.063	0.055	0.037	0.052	0.050	0.061	0.032	0.060	0.055

A.1.4 Size calculations for Female age-groups when α=0.05

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.139	0.149	0.050	0.055	0.050	0.057	0.105	0.093	0.097	0.072
	10-19	0.133	0.130	0.145	0.111	0.106	0.054	0.094	0.065	0.073	0.067
	20-29	0.062	0.190	0.121	0.067	0.132	0.057	0.064	0.100	0.079	0.113
	30-39	0.062	0.194	0.123	0.076	0.057	0.058	0.066	0.114	0.071	0.042
	40-49	0.104	0.078	0.105	0.068	0.093	0.058	0.078	0.054	0.068	0.090
	50-59	0.110	0.076	0.074	0.051	0.053	0.064	0.071	0.071	0.058	0.059
	60-69	0.051	0.080	0.085	0.074	0.070	0.062	0.062	0.056	0.058	0.065
	70-79	0.047	0.062	0.062	0.062	0.058	0.061	0.075	0.055	0.048	0.065
	80 +	0.071	0.055	0.051	0.057	0.050	0.045	0.051	0.061	0.055	0.059

A.1.5 Size calculations for Male age-groups when α=0.01

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.007	0.015	0.021	0.014	0.020	0.015	0.022	0.012	0.013	0.012
	10-19	0.010	0.025	0.019	0.028	0.013	0.019	0.022	0.011	0.010	0.015
	20-29	0.013	0.018	0.016	0.025	0.015	0.014	0.016	0.011	0.018	0.016
	30-39	0.024	0.022	0.023	0.019	0.021	0.022	0.017	0.012	0.009	0.028
	40-49	0.035	0.016	0.010	0.037	0.016	0.007	0.012	0.014	0.012	0.010
	50-59	0.011	0.020	0.015	0.017	0.012	0.014	0.014	0.012	0.024	0.013
	60-69	0.015	0.014	0.016	0.014	0.016	0.014	0.015	0.011	0.010	0.006
	70-79	0.012	0.011	0.006	0.014	0.020	0.013	0.011	0.011	0.010	0.010
	80 +	0.013	0.011	0.008	0.005	0.005	0.006	0.014	0.007	0.014	0.008

A.1.6 Size calculations for Female age-groups when α=0.01

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.027	0.018	0.015	0.025	0.022	0.025	0.027	0.020	0.022	0.016
	10-19	0.029	0.006	0.012	0.012	0.022	0.012	0.018	0.009	0.016	0.017
	20-29	0.062	0.049	0.032	0.021	0.012	0.021	0.016	0.022	0.014	0.024
	30-39	0.062	0.048	0.034	0.018	0.022	0.018	0.028	0.022	0.010	0.019
	40-49	0.024	0.022	0.020	0.013	0.007	0.013	0.011	0.013	0.009	0.013
	50-59	0.030	0.024	0.017	0.013	0.019	0.013	0.017	0.013	0.011	0.022
	60-69	0.019	0.024	0.022	0.014	0.016	0.014	0.009	0.013	0.014	0.008
	70-79	0.012	0.012	0.010	0.007	0.015	0.007	0.013	0.008	0.007	0.010
	80 +	0.007	0.011	0.010	0.009	0.008	0.009	0.006	0.007	0.013	0.008

A.2 Power calculations of SIR method

A.2.1 Power calculations for Male age-groups when α=0.1

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.986	0.992	0.986	0.979	0.992	0.984	0.991	0.992	0.980	0.991
	10-19	0.968	0.983	0.986	0.987	0.984	0.981	0.988	0.995	0.979	0.993
	20-29	0.984	0.978	0.988	0.992	0.989	0.991	0.985	0.991	0.986	0.988
	30-39	0.972	0.982	0.980	0.984	0.981	0.978	0.989	0.982	0.989	0.986
	40-49	0.993	0.980	0.991	0.986	0.993	0.987	0.988	0.984	0.993	0.995
	50-59	0.980	0.985	0.976	0.992	0.991	0.993	0.985	0.987	0.990	0.982
	60-69	0.981	0.986	0.981	0.985	0.990	0.987	0.992	0.984	0.988	0.988
	70-79	0.988	0.989	0.992	0.994	0.989	0.983	0.985	0.995	0.992	0.991
	80 +	0.982	0.988	0.982	0.989	0.990	0.994	0.988	0.979	0.992	0.990

A.2.2 Power calculations for Female age-groups when α=0.1

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.986	0.980	0.977	0.987	0.982	0.984	0.987	0.990	0.994	0.989
	10-19	0.989	0.981	0.973	0.991	0.989	0.978	0.985	0.975	0.988	0.987
	20-29	0.956	0.974	0.985	0.988	0.978	0.979	0.994	0.992	0.987	0.986
	30-39	0.954	0.978	0.988	0.985	0.986	0.979	0.992	0.977	0.985	0.979
	40-49	0.985	0.985	0.979	0.988	0.987	0.986	0.985	0.985	0.988	0.988
	50-59	0.969	0.991	0.993	0.990	0.985	0.984	0.986	0.993	0.979	0.986
	60-69	0.987	0.988	0.981	0.987	0.992	0.982	0.988	0.986	0.989	0.993
	70-79	0.986	0.994	0.989	0.994	0.986	0.976	0.982	0.988	0.987	0.988
	80 +	0.984	0.984	0.986	0.984	0.987	0.983	0.987	0.987	0.987	0.994

A.2.3 Power calculations for Male age-groups when α=0.05

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.856	0.944	0.947	0.911	0.923	0.931	0.952	0.925	0.948	0.953
	10-19	0.850	0.887	0.900	0.928	0.925	0.940	0.931	0.942	0.925	0.943
	20-29	0.806	0.928	0.916	0.914	0.903	0.931	0.958	0.953	0.947	0.937
	30-39	0.813	0.917	0.901	0.888	0.958	0.932	0.948	0.921	0.940	0.929
	40-49	0.940	0.924	0.923	0.938	0.921	0.955	0.936	0.917	0.938	0.954
	50-59	0.929	0.948	0.926	0.929	0.942	0.928	0.943	0.964	0.935	0.937
	60-69	0.960	0.942	0.957	0.929	0.944	0.926	0.942	0.942	0.947	0.945
	70-79	0.946	0.943	0.945	0.947	0.940	0.942	0.952	0.947	0.951	0.955
	80 +	0.941	0.934	0.932	0.939	0.948	0.951	0.939	0.939	0.950	0.943

A.2.4 Power calculations for Female age-groups when α=0.05

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.856	0.944	0.947	0.911	0.923	0.931	0.952	0.925	0.948	0.953
	10-19	0.850	0.887	0.900	0.928	0.925	0.940	0.931	0.942	0.925	0.943
	20-29	0.806	0.928	0.916	0.914	0.903	0.931	0.958	0.953	0.947	0.937
	30-39	0.813	0.917	0.901	0.888	0.958	0.932	0.948	0.921	0.940	0.929
	40-49	0.940	0.924	0.923	0.938	0.921	0.955	0.936	0.917	0.938	0.954
	50-59	0.929	0.948	0.926	0.929	0.942	0.928	0.943	0.964	0.935	0.937
	60-69	0.960	0.942	0.957	0.929	0.944	0.926	0.942	0.942	0.947	0.945
	70-79	0.946	0.943	0.945	0.947	0.940	0.942	0.952	0.947	0.951	0.955
	80 +	0.941	0.934	0.932	0.939	0.948	0.951	0.939	0.939	0.950	0.943

A.2.5 Power calculations for Male age-groups when α=0.01

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.839	0.841	0.898	0.846	0.905	0.892	0.886	0.893	0.887	0.911
	10-19	0.750	0.887	0.871	0.859	0.845	0.850	0.849	0.866	0.857	0.889
	20-29	0.726	0.814	0.808	0.886	0.846	0.888	0.890	0.879	0.853	0.893
	30-39	0.705	0.849	0.880	0.851	0.824	0.872	0.872	0.832	0.888	0.879
	40-49	0.804	0.889	0.890	0.899	0.857	0.851	0.873	0.898	0.881	0.895
	50-59	0.827	0.883	0.858	0.871	0.905	0.906	0.890	0.870	0.868	0.879
	60-69	0.879	0.892	0.859	0.893	0.882	0.885	0.898	0.872	0.901	0.908
	70-79	0.884	0.894	0.871	0.877	0.878	0.873	0.886	0.907	0.915	0.900
	80 +	0.878	0.894	0.895	0.893	0.909	0.909	0.886	0.884	0.904	0.889

A.2.6 Power calculations for Female age-groups when α=0.01

		Population Size									
		10,000	20,000	30,000	40,000	50,000	60,000	70,000	80,000	90,000	100,000
Age Group (Years)	< 10	0.856	0.856	0.879	0.826	0.869	0.885	0.837	0.872	0.899	0.874
	10-19	0.850	0.887	0.801	0.857	0.868	0.891	0.879	0.843	0.873	0.847
	20-29	0.806	0.797	0.783	0.793	0.903	0.885	0.911	0.903	0.830	0.819
	30-39	0.813	0.783	0.763	0.888	0.901	0.882	0.903	0.860	0.886	0.892
	40-49	0.837	0.837	0.871	0.830	0.871	0.903	0.852	0.882	0.865	0.907
	50-59	0.859	0.824	0.885	0.888	0.881	0.880	0.889	0.913	0.889	0.904
	60-69	0.900	0.868	0.904	0.894	0.917	0.878	0.900	0.878	0.889	0.898
	70-79	0.881	0.889	0.898	0.904	0.901	0.882	0.908	0.881	0.890	0.899
	80 +	0.891	0.852	0.881	0.900	0.899	0.901	0.884	0.889	0.890	0.884

Appendix B: Appendix for Chapter 4

B.1 Spatial Scan Analysis of Breast Cancer Mortality

B.1.1 Results of Circular Spatial Scan

SUMMARY OF DATA:

Limit length of cluster: 15
Number of census areas.: 56
Total cases .....: 377

MOST LIKELY CLUSTER:

Census areas included .: 5, 12, 13, 14, 15, 17, 18, 28, 29, 51, 52, 53
Maximum distance.....: 15917.8 (areas: 17 to 29)
Number of cases .....: 148 (123.158 expected)
Overall relative risk .: 1.20171
Log likelihood ratio ..: 3.60947
Monte Carlo rank .....: 374/1000
P-value .....: 0.374

B.1.2 Results of Flexible Spatial Scan

SUMMARY OF DATA:

Limit length of cluster: 15
Number of census areas.: 56
Total cases .....: 377

MOST LIKELY CLUSTER:

Census areas included .: 15, 16, 17, 18, 23, 25, 26, 27, 53
Maximum distance.....: 23724.2 (areas: 16 to 26)
Number of cases .....: 69 (49.0216 expected)
Overall relative risk .: 1.40754
Log likelihood ratio ..: 4.23017
Monte Carlo rank .....: 627/1000
P-value .....: 0.627

B.2 Spatial Scan Analysis of Colon Cancer Mortality

B.2.1 Results of Circular Spatial Scan

SUMMARY OF DATA:

Limit length of cluster: 15
Number of census areas.: 56
Total cases .....: 104

MOST LIKELY CLUSTER:

Census areas included .: 32
Maximum distance.....: 0 (areas: 32 to 32)
Number of cases .....: 5 (1.23172 expected)
Overall relative risk .: 4.05936
Log likelihood ratio ..: 3.3068
Monte Carlo rank .....: 460/1000
P-value .....: 0.46

B.2.2 Results of Flexible Spatial Scan

SUMMARY OF DATA:

Limit length of cluster: 15
Number of census areas.: 56
Total cases .....: 104

MOST LIKELY CLUSTER:

Census areas included .: 5, 6, 12, 29, 30, 31, 51, 52, 53
Maximum distance.....: 9740.16 (areas: 31 to 51)
Number of cases .....: 41 (28.3188 expected)
Overall relative risk .: 1.4478
Log likelihood ratio ..: 3.61798
Monte Carlo rank .....: 731/1000
P-value .....: 0.731

References

[1] Coggan, D., Rose, G. and Barker, D. (1997). *Epidemiology for the uninitiated.* BMJ Publications.

[2] Dos Santos Silva, I. (1999). *Cancer Epidemiology: Principles and Methods*. International Agency for Research on Cancer, Lyon.

[3] Elliot, P., Cuzick, J., English, D. and Stern, R. (1992). *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford University Press, New York.

[4] United Nations Economics Commission for Europe (26 June 2001). "*Small area statistics – an indispensable tool for decision makers at all levels"*. Press release. www.unece.org/press/pr2001/01stat07e.htm - Accessed on 03/12/2007.

[5] Ripley, B. (1981). *Spatial Statistics*. Wiley, New York.

[6] Elliot, P. and Wartenberg, D. (2004). *Spatial Epidemiology: Current Approaches and Future Challenges*. Environ. Health Perspect. 112(9):998-1006.

[7] Stocks, P. (1936). *Distribution in England and Wales of cancer of various sites*. Ann. Rep. Br. Empire Cancer Campaign 13:239-280.

[8] Monmonier, M. (1997). *How to Lie with Maps*. Chicago: The University of Chicago Press.

[9] "An Atlas of Tobacco Smoking in Scotland", *NHS Health Scotland.* www.scotpho.org.uk/tobaccoatlas - Accessed on 10/07/2008.

[10] Davies, H., Joshi, H. and Clarke, L. (1997). *Is it Cash that the Deprived are Short Of?*, J. R. Statistical Society Series A 160, Part 1, pp. 107-126.

[11] Trumbo, C (2000). *Public Requests for cancer cluster investigations: a survey of state health departments*. Am J Public Health 90:1300-1302.

[12] Jolley, D., Jarman, B. and Elliot, P. (1992). *Socio-economic confounding*. In Elliot, P., Cuzick, J., English, D. and Stern, R. (1992). *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford University Press, New York, pp 72-88.

[13] Leon, D. (1988). *Longitudinal Study: Social distribution of cancer*, OPCS Series LS, No. 3. HMSO, London.

[14] "Social Focus on Deprived Areas". *Scottish Government*. www.scotland.gov.uk/Publications/2005/09/2792129/21311 - Accessed on 16/08/2008.

[15] “Scottish Index of Multiple Deprivation”. *Scottish Government*. www.scotland.gov.uk/Topics/Statistics/SIMD - Accessed on 20/08/2008.

[16] Arnold, R. (1999). *Counts in Small Area Studies: Implications for Studies of Environment and Health*. Studies on Medical and Population. Subjects No.62. London:U.K. Office of National Statistics, 10–23.

[17] Nelson, P. (2003). *Geographical Epidemiology of Hypospadias: Small Area study of Birth Prevalence*. University of London.

[18] “Our Organisation”, *ISD Scotland*. www.isdscotland.org/isd/846.html - Accessed on 20/02/2007.

[19] “National Statistics”, *ISD Scotland*. www.isdscotland.org/isd/776.html - Accessed on 13/05/2007.

[20] Esteve, J., Benhamou, E., Raymond, L. (1994). *Statistical Methods in Cancer Research Vol. IV: Descriptive Epidemiology*. IARC Scientific Publications No. 128, International Agency for Research on Cancer, Lyon.

[21] Best, N., Elliot, P. and Richardson, S. *Spatial Epidemiology: A short course*. Imperial College, London. www.stats.ma.ic.ac.uk/n/ngb30/public_html/ - Accessed on 12/04/2007.

[22] Boyle, P., Muir, C. and Grundmann, E. (1989). *Cancer Mapping (Recent Results in Cancer Research)*. Springer-Verlag, New York.

[23] Gatrell, A. (2002). *Geographies of Health: An introduction.* Wiley-Blackwell.

[24] Stark, J., Black, R. and Brewster, D. (2007). *Risk of leukaemia among children living near the Solway coast of Dumfries and Galloway Health Board area, Scotland, 1975-2002.* Occup. Environ. Med.

[25] Elliot, P., Wakefield, J., Best, N. and Briggs, D. (2000). *Spatial Epidemiology: Mehtods and Applications.* Oxford University Press, Oxford.

[26] Lawson, A., Browne, W., Carmen, L. and Rodeiro, V. (2003). *Disease Mapping with WinBUGS and MlwiN.* Wiley and Sons.

[27] Waller, L. and Gotway, C. (2004). *Applied spatial statistics for public health data*, Wiley and Sons.

[28] Gelman, A. and Price, P. (1999). *All maps of parameter estimates are misleading.* Statistics in Medicine, 18:3221-3234.

[29] Besag, J., and Newell, J. (1991). *The detection of clusters in rare diseases.* J. R. Statistical Society Series A 154: 143-155.

[30] Aignaux, J, Cousens, S *et* al (2002). Analysis *of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998.* International Journal of Epidemiology 2002;31:490-495.

[31] Kim, A, Al-Rumaizan, C. *et al* (2004). *A brief evaluation of statistical methods for detecting disease clusters in time and/or space*. Center for Computational Mathematics, Japan.

[32] Song, C., and Kulldorff, M (2000). *Power evaluation of disease clustering tests.* Department of Statistics, University of Connecticut.

[33] Costa, M. and Assunçã, R. (2005). *A fair comparison between the spatial scan and the Besag and Newell disease clustering test.* Environmental and Ecogological Studies, 301-319.

[34] Kulldorff, M. and Information Management Services Inc. (2006). SaTScanTM v7.0: Software for the spatial and space-time scan statistics.
www.satscan.org

[35] Kuldorff, M. (1997). *A spatial scan statistic.* Communications in Statistics: Theory and Methods, 26:1481-1496.

[36] Kuldorff, M., Song, C., Gregorio, D. *et al* (2006). *Cancer Map Patterns: Are they random or not?* Am. J. Prev. Med. 30:37-49.

[37] Schabenberger, O. and Gotway, C. (2005). *Statistical Methods for Spatial Data Analysis.* CRC Press.

[38] Ozdenerol, E., Williams, B., Kang, S. and Magsumbol, M. (2005). *Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters.* Int. J. Health Geogr. 4:19.

[39] Tango, T. and Takahashi, K. (2005). *A flexibly shaped spatial scan statistic for detecting cluster.* Int. J. Health Geogr. 4:11.

[40] Takahashi K., Yokoyama T. and Tango T. (2007). *FleXScan v2.0: Software for the Flexible Scan Statistcs*, Department of Technology Assessment and Biostatistics, National Institute of Public Health, Japan.,
www.niph.go.jp/english/index.html

[41] Bithell, J., Dutton, S., Draper, G. and Neary, N. (1994). *Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. BMJ.* 1994;309(6953):501–505.

[42] Bithell, J. (2005). *The choice of test for detecting raised disease risk near a point source.* Stat. Med. 14(21–22):2309–2322

[43] Committee on Medical Aspects of Radiation in the Environment (COMARE) (2005). *Tenth Report – The incidence of childhood cancer around nuclear installations in Great Britain.*NRPB, Chilton.

[44] Black, D. (1984). *Investigation of the possible increased incidence of cancer in Wes Cumbria.* Report of the Independent Advisory Group. HMSO, London.

[45] Sharp, L., Black, R., Harkness, P. and McKinney, P. (1996). *Incidence of childhood leukaemia and non-Hodgkin's lymphoma in the vicinity of nuclear sites in Scotland 1968-93.* Occup. and Environ. Medicine 53:823-831.

[46] "About us", *SAHSU*.
www.sahsu.org/about_us.htm - Accessed on 01/02/2007.

[47] "Related Studies", *SAHSU*
www.sahsu.org/sahsu_related_studies.htm#RIF - Accessed on 10/02/2007.

[48] Ball, W., LeFevre, S., Jarup, L., and Beale, L. (2008). *Comparison of Different Methods for Spatial Analysis of Cancer Data in Utah*. Environ. Health Perspect. 116(8):1120-4.

[49] Pearson, E. and Neyman, J. (1930). *On the Problem of Two Samples*., reprinted at pp.99-115 in Neyman, J. & Pearson, E.S., *Joint Statistical Papers*, Cambridge University Press, (Cambridge), 1967 (originally published in 1930).

[50] Diggle, P. (1983). *Statistical analysis of spatial point patterns.* Arnold, London.

[51] “Publications 2004 Information Paper”, *Local Government Boundary Commission for Scotland.*

www.lgbc-scotland.gov.uk - Accessed on 27/06/2007.

[52] Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). *WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility*. Statistics and Computing, 10:325--337.

[53] “Breast cancer at a glance”, *Cancer Research UK.*

info.cancerresearchuk.org/cancerandresearch/cancers/breast/ - Accessed on 03/10/2007.

[54] “Bowel cancer at a glance”, *Cancer Research UK.*

info.cancerresearchuk.org/cancerandresearch/cancers/bowel/ - Accessed on 03/10/2007.