Investigating Clustering in Trisomy 18 and Trisomy 13

James Phillip Cook

Submitted in Partial Fulfillment of the Requirements of the Degree of Doctor of Philosophy

Queen Mary University of London

STATEMENT OF CONTRIBUTION

I, James P Cook, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: James P Cook

Date:25/11/2013

ABSTRACT

Trisomies 18 and 13 are rare genetic conditions (occurring around 1 in 6,000 and 10,000 newborns respectively) which are caused by an extra copy of either chromosome 18 or 13, similar to trisomy 21 (Down syndrome). The only known risk factor for these syndromes is maternal age, however previous cluster analyses have linked trisomy risk to a number of alternate factors, including radiation exposure and infection.

Cases of trisomies 18 and 13 from the National Down Syndrome Cytogenetic Register (NDSCR) were scanned for temporal and spatial clusters throughout England and Wales between 2004 and 2010. No temporal clusters were detected, however there were multiple significant spatial clusters detected for both trisomies in London. These clusters were likely caused by advanced maternal age in the region, and it is also possible that regional differences in gestational age at the time of prenatal screening could have contributed to these clusters.

In order to account for maternal age and gestational age at diagnosis, a novel method was developed in R to directly weight cases based on these factors. Applying weights to cases directly allowed both factors to be simultaneously accounted for by multiplying weights together. This method was evaluated using synthetic data and compared with an alternate method in the widely used program SaTScan. Both programs returned similar results when the weighting method was mild, but when extreme weights were applied at random significant clusters were observed in SaTScan but not in R.

The NDSCR data was weighted and then rescanned for spatial clusters in both programs. No evidence of clustering was detected using the novel method, while SaTScan returned multiple highly significant clusters. These findings, combined with those obtained using the synthetic data, indicate that the novel method produces more reliable results than SaTScan when extreme adjustment is applied.

GLOSSARY OF TERMS

Amniocentesis – Medical procedure used to prenatally diagnose congenital anomalies.

Chorionic villus sampling (CVS) – Medical procedure used to prenatally diagnose congenital anomalies.

Congenital anomaly – A medical condition that a baby has at birth.

Government office region (GOR) – Administrative region of government in the United Kingdom.

National Down Syndrome Cytogenetic Register (NDSCR) – Congenital anomaly register collecting all cytogenetically diagnosed cases of trisomies 21, 18 and 13 in England and Wales.

Primary care trust (PCT) – Administrative region of the National Health Service (NHS) in the United Kingdom.

R – Freely available programming language and software environment for statistical computing and graphics.

SaTScan – Freely available software for the analysis of spatial, temporal or spatio-temporal data.

Stata – Statistical software package.

Trisomy – Chromosomal disorder characterised by the presence of a third copy of a particular chromosome.

ACKNOWLEDGEMENTS

Enormous thanks to my main supervisor Professor Joan Morris, who was a fantastic source of guidance and advice during the entire duration of the PhD, and mysecond supervisor, Dr Jane Warwick, whose comments and insight proved invaluable on a number of occasions.Thanks also to everyone else who provided academic assistance, including (but not limited to) Professor Stephen Duffy and Dr Richard Hooper, who gave me very useful feedback and fresh impetus during my 18 month examination, and to Queen Mary University of London for funding this thesis.

It was fantastic to be part of the team at the NDSCR, my thanks to Linda, Haiyan, Lynda, Beth, and everyone who made me feel welcome at the Wolfson Institute. In particular I would like to thank Mo and Becca, who kept me going through the challenging latter half of the project.

My family, as always, were full of support during my PhD. To my parents, thank you for always being there for me, both during this PhD and throughout life. Special mention must be made of my granddad, to whom my education and academic achievement meant everything.

Completing this thesis is one of the most challenging things I have ever accomplished, and I could not have done it without the friendship of other PhD students. Thanks especially to Tom, Adrian and Claire for sharing the burden in Reading, and to Michael for picking up the slack in Leicester.

And finally a very special thanks to Cat, without whom I would never have got this far.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Cluster detection in biosurveillance is a useful tool for discovering new risk factors and improving the overall understanding of medical conditions. Previous cluster analyses of trisomy 21 cases have suggested the possibility of several environmental factors which may increase a woman's risk of having an affected pregnancy. The only confirmed risk factor for all 3 trisomy syndromes is maternal age. Since mean maternal age is currently increasing in England and Wales and varies geographically, the detection of age-independent clusters of trisomies 18 and 13 will require cluster detection methods which adjust for maternal age. The presence of clusters after adjusting for maternal age could indicate the involvement of alternate risk factors.

In this thesis, trisomy 18 and trisomy 13 data from the National Down Syndrome Cytogenetic Register (NDSCR) between 2004 and 2010 were scanned for temporal and spatial clusters using commonly used cluster detection methods developed and proposed in the mid 1990s.The main focus of this thesis is the development of a novel weighting method to statistically account for maternal age, the strongest trisomy risk factor. The method is also used to account for gestational age at diagnosis, which is not a risk factor but may still contribute to the formation of spatial clusters. This novel weighting method is compared with a corresponding adjustment technique in SaTScan, a freely available piece of software for the detection of temporal and spatial disease clusters. Comparison between the novel method and SaTScan is performed in spatial cluster analyses using both synthetic data and observed data from the NDSCR.

In the remainder of chapter 1, a general introduction to Edwards and Patau syndromes is presented, along with a review of previously reported trisomy clusters. Finally, an overview of cluster detection methods is presented.

Section 1.1: Epidemiology and Aetiology of Edwards and Patau Syndromes

Edwards Syndrome (trisomy 18) and Patau Syndrome (trisomy 13) are, along with Down Syndrome (trisomy 21), the only viable autosomal trisomies in humans (Edwards *et al*1960, Patau *et al* 1960, Lejeune *et al* 1959*)*. The word "trisomy" is given to a genetic condition which is caused by the presence of either a whole or part of a third copy of a chromosome (in humans), while "autosomal" means that the affected chromosome is one of the numbered autosomes (1-22), not one of the sex chromosomes X and Y.

While it is estimated that 1-2% of human conceptions are affected with triploidy, most are aborted at the very early stages of pregnancy resulting in all triploid conditions being rare in live borns. The severity of a trisomy increases as the amount of genetic material on the affected chromosome increases, as a result of which only 3 autosomal trisomies are capable of surviving to term. These are, in order of least to most severe, trisomies 21, 18 and 13.

Recently reported fetal loss rates in England and Wales between the time of amniocentesis (18 weeks) and birth are 65% for trisomy 18 and 42% for trisomy 13 (Morris & Savva 2008). These figures compare with 25% for trisomy 21 (Savva *et al*2010), and around 2% in the general population between 14 weeks and term (Rodeck & Whittle 2008).

Section 1.1.1: Trisomy 21 (Down Syndrome)

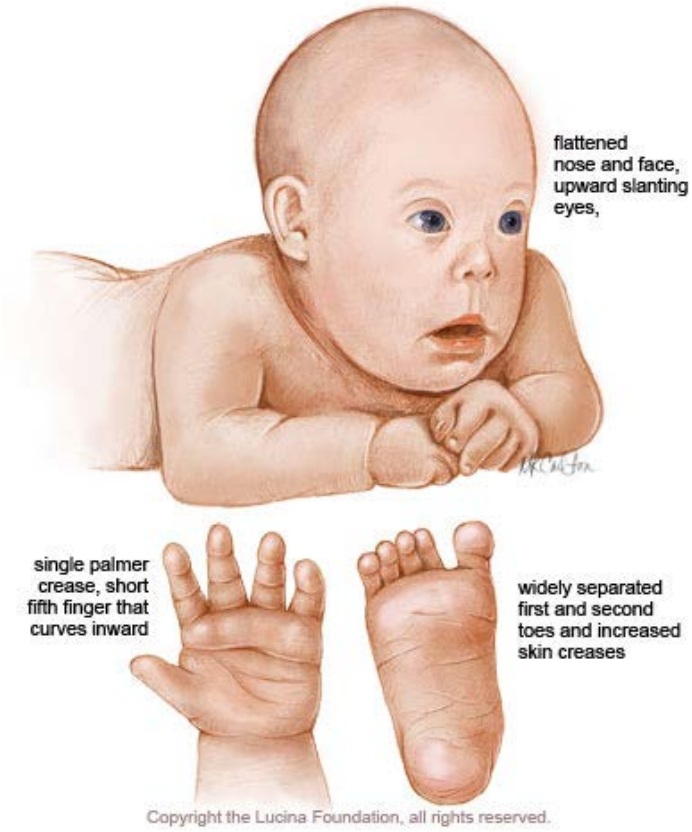

Figure 1.1.1.Common visible symptoms of trisomy 21.

Trisomy 21 occurs about 1 in every 800 live births and is the most common trisomy in humans. Individuals with trisomy 21 can often be identified by their facial features, while other common symptoms include intellectual disability, poor muscle tone, short neck and reduced growth. Cardiac defects are present in around 40% of trisomy 21 cases (Freeman *et al*, 1998). Some common visible symptoms of trisomy 21 are shown in figure 1.1.1.

Data from the NDSCR Annual Report in England and Wales in 2009 reported that 62% of trisomy 21 cases were prenatally diagnosed, consisting of 76% terminations, 5% births, 3% stillbirths and 16% are awaiting outcome. 6% of cases with unknown outcome are expected to result in a live birth. The remaining 38% of cases were diagnosed post natally, consisting of 96% live births and 4% were stillbirths.

Section 1.1.2: Trisomy 18 (Edwards Syndrome)

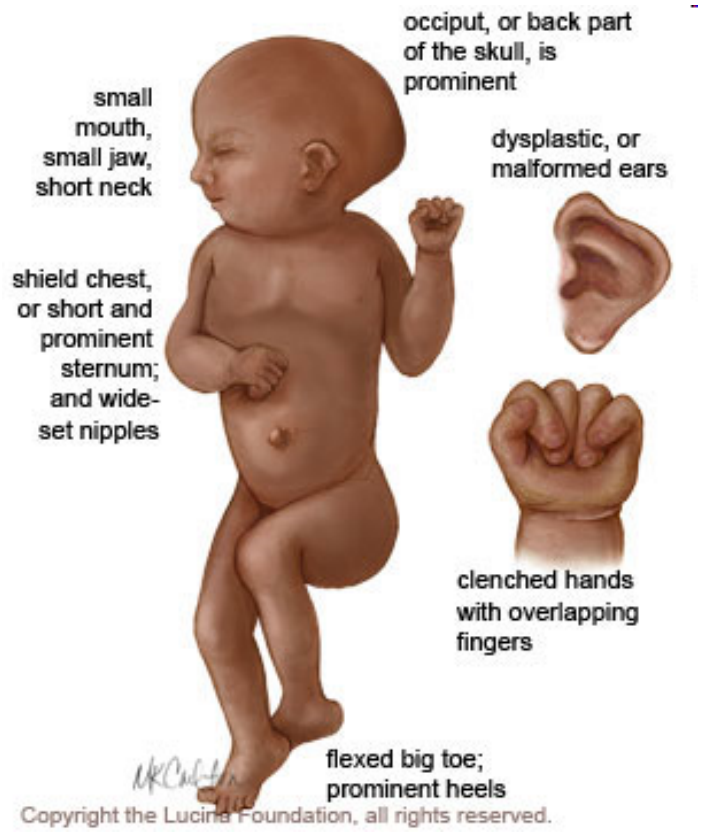

Figure 1.1.2.Common visible symptoms of trisomy 18.

Trisomy 18 occurs about 1 in every 3,000 live births and is the 2nd most common trisomy in humans. There are a wide range of symptoms, including heart defects, kidney problems, digestive tract defects, small jaw, small head, excess amniotic fluid, severe developmental delays and delayed growth. There is strong evidence that trisomy 18 is more lethal in males, with female live born survival rates up to double those of males (Niedrist et al, 2006). Some common visible symptoms of trisomy 18 are shown in figure 1.1.1.

Data from the NDSCR Annual Report in 2009 reported that 91% of trisomy 18 cases were prenatally diagnosed, of which 74% were terminations, 2% live births, 7% were stillbirths and 17% are awaiting outcome. 3% of cases with unknown outcome are expected to result in a live birth. The remaining 9% of cases were diagnosed post natally, consisting of 65% live births and 35% stillbirths.

Section 1.1.3: Trisomy 13 (Patau Syndrome)

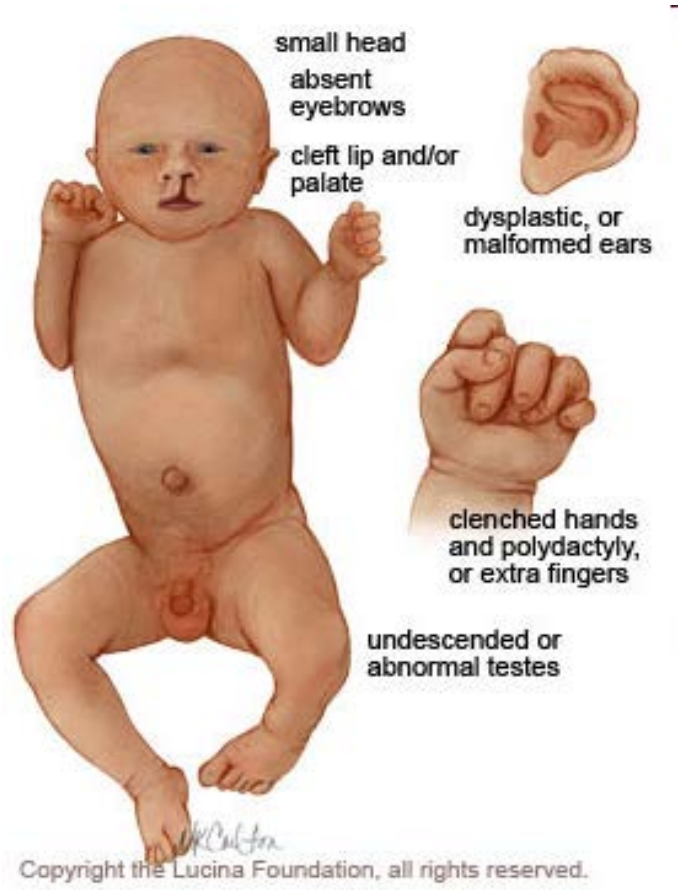

Figure 1.1.3.Common visible symptoms of trisomy 13.

Trisomy 13 occurs about 1 in every 16,000 live births and is the 3rd most common trisomy in humans. Common symptoms include heart defects, brain or spinal cord abnormalities, very small or poorly developed eyes (microphthalmia), extra fingers and/or toes, an opening in the lip (a cleft lip) with or without an opening in the roof of the mouth (a cleft palate), and weak muscle tone (hypotonia). Some common visible symptoms of trisomy 13 are shown in figure 1.1.1.

The NDSCR Annual Report in 2009 showed that there were too few trisomy 13 cases to present outcomes according to time of diagnosis, but that 66% were terminated, 12% were live born, 7% stillborn and 15% are awaiting outcome, of which 4% are expected to result in a live birth.

Section 1.1.4. Survival After Birth

Many studies examining survival of trisomies 18 and 13 have focused on survival after birth, where both trisomies also exhibit very low survival rates. Tables 1 and 2 show reported survival percentages for trisomy 18 (table 1.1.1) and trisomy 13 (table 1.1.2) from numerous studies carried out around the world. In each table the top row shows the primary author on the report, while the second row shows the year of publication.

Survival of trisomy 18 live borns was reported between 60 and 98% after a day according to the studies in table 1, with this figure falling to around 30-55% according to most studies after a week. All studies reported less than 10% survival at 1 year of age. Survival for trisomy 13 was comparable, with studies reporting between 69 and 92% survival after a day and 38 – 72% after a week. All studies reported less than 10% survival after a year with the exception of Magenis *et al* in 1967, who found that 12% of trisomy 13 live borns were still alive after a year.

For trisomy 18 live borns, the reported median survival was most commonly between 3 and 6 days with 3 studies (Irving *et al* 2011, Vendola *et al* 2010, and Nembhard *et al* 2001) not providing a median survival time and 2 studies (Weber, 19j67 and Rasmussen *et al*, 2003) reporting greater than 10 days life expectancy. The Weber paper in 1967 noted that females with trisomy 18 were more likely to survive for longer than affected males, and presented survival estimates for each gender individually. Probability of survival for male trisomy 18 cases was reported to be 0.638 at 2 months, while the study estimated that the likelihood of affected females to survive to 2 months was 0.769, to 6 months was 0.311 and to 3 years was 0.175. These survival estimates are much greater than those in more recent studies. One possible explanation for this is that some of these cases were incorrectly diagnosed and were actually cases of trisomy 21, which has a much longer life expectancy. Cases in this study were obtained from 9 previously published studies of trisomy 18 mortality and survival, and personal communication to the author. It is possible that some of these cases were incorrectly diagnosed, or that the studies used to ascertain cases were biased in favour of longer surviving cases.

The Rasmussen study in 2003 also reported a longer life expectancy (14.5 days) than any of the other recent estimates. This study collected liveborn cytogenetically confirmed cases from the Metropolitan Atlanta Congenital Defects Program (MACDP) between 1968 and 1999. Again, it is possible that either cases were incorrectly diagnosed or that there was a form of bias present in favour of longer surviving cases, as the median survival time in this study was more than double the next highest estimate (excluding the Weber study in 1967).

Recent median survival estimates for trisomy 13 live borns were more varied; falling within 2-10 days according to those studies which provided median survival estimates. This is likely due

to scarcity of data making it difficult to accurately estimate survival. The study by Magenis *et al* in 1968 estimated a much longer median survival of between 30 and 60 days. This study used data from the laboratory at the University of Oregon medical school, cases from previously published literature, and through direct communication with cytogenetic laboratories around the world. Similarly to the 1967 Weber study into trisomy 18 survival, it is possible that either some of the cases were misdiagnosed or that some of the studies used to ascertain cases were biased in favour of longer surviving cases.

In congenital anomaly studies, prevalence (the total number of cases in the population at a given time point) is studied rather than incidence (the total number of 'new' cases in the population at any given time point). This is because is it not possible to ascertain all 'new' cases as a proportion of affected pregnancies will miscarry before diagnosis is possible. There is also no population estimate of the total number of pregnancies at risk due to miscarriages and terminations. Therefore, prevalence estimates are provided for congenital anomalies per 1,000 or 10,000 total births (live births and stillbirths).

Table 1.1.1. % Survival of trisomy 18 live borns.

	Weber	Carter	Goldstein	Root	Embleton	Nembhard	Brewer	Rasmussen	Lin	Niedrist	Vendola	Irving
Year	**1967**	**1985**	**1988**	**1994**	**1996**	**2001**	**2002**	**2003**	**2006**	**2006**	**2010**	**2011**
Cases	192	43	76	64	34	68	84	114	39	161	140	67
1 day	98	60	60	86	70	84	88	86	95	68	-	-
1 week	89	35	44	45	-	56	43	63	46	40	52	36
1 month	72	11	21	34	15	40	25	39	16	22	30	27
6 months	13	5	3	9	-	-	-	11	3	9	-	-
1 year	8	4	-	5	0	-	2	8	3	6	3	6
Median (days)	**70**	**5**	**6**	**4**	**3**	**-**	**6**	**14.5**	**6**	**4**	**-**	-

Table 1.1.2. % Survival of trisomy 13 live borns.

	Magenis	Goldstein	Wyllie	Nembhard	Brewer	Rasmussen	Lin	Vendola	Irving
Year	**1968**	**1988**	**1994**	**2001**	**2002**	**2003**	**2007**	**2010**	**2011**
Cases	172	76	16	27	84	114	28	200	30
1 day	92	77	69	85	75	86	89	-	-
1 week	72	40	38	59	50	56	61	42	43
1 month	56	23	13	44	28	30	29	20	30
6 months	20	10	0	-	-	11	7	-	-
1 year	12	-	0	-	3	9	4	3	3
Median (days)	**30-60**	**2.5**	**4**	**-**	**8.5**	**7**	**9**	**-**	-

Section 1.1.5: Trisomy Screening Procedures

Screening is defined by the National Down's Syndrome Screening Programme for England as "a public health service in which members of a defined population, who do not necessarily perceive they are at risk of, or are already affected by a disease or its complications, are asked a question or offered a test to identify those individuals who are more likely to be helped than harmed by further tests or treatments to reduce the risk of disease or its complications" (National Down's Syndrome Screening Programme for England Handbook for Staff, 2004). Prenatal diagnosis of a trisomy requires a sample of foetal DNA, taken via invasive methods which present a small risk of miscarriage to the developing foetus. In order to minimise the number of such procedures performed, mothers are screened using a variety of non-invasive measures which assess the risk of the foetus being affected with a trisomy. Only those pregnancies at high risk (e.g. a risk of ≥ 1 in 250, Wald *et al* 1999) are then diagnosed using an invasive method. Although these tests are primarily designed for use in prenatally screening for trisomy 21, they are also effective in detecting trisomies 18 and 13. The UK National Screening Committee recommends that all pregnant women are offered trisomy 21 screening (UK National Screening Committee Annual Report 2010-11), and as such, all 3 trisomies are routinely screened for in England and Wales.

Screening can be performed in the first and second trimesters. One commonly used screening method is to measure the concentrations of maternal blood markers which are present in different quantities in affected compared to unaffected pregnancies. In the first trimester, levels of free (beta) human chorionic gonadotrophin (hCG) and pregnancy-associated plasma protein A (PAPP-A) can be measured, and combined with the ultrasonic measurement of nuchal translucency to give the trisomy combined test (Nicolaides *et al* 1992). Fluid at the nape of the neck of the fetus is measured (in mm) in the nuchal translucency scan, which can be performed between 11 and 14 weeks. In trisomic foetuses, the nuchal translucency is expected to be greater than in unaffected foetuses. However, the reliability of the nuchal translucency measurement has led to questions about whether first or second trimester screening is preferable (Wald *et al* 1999).

Second trimester blood markers used in screening are maternal serum alpha-seroprotein (MSAFP, Dimaio et al), unconjugated estriol (uE3), human chorionic gonadotrophin (hCG) and inhibin-A (INH-A). Measurements of MSAFP, uE3 and hCG can be combined with maternal age to make the triple test, which becomes a quadruple test when INH-A is added (Benn 2002). It is possible to combine blood marker results from the first and second trimesters to make up the trisomy integrated test (Wald *et al* 1999).

During the study period, the quality of screening in England and Wales improved. In 2004, the National Down's Syndrome Screening Programme for England Handbook for Staff recommended that by 2004/05, screening programmes should be able to detect at least 60% of cases with a false positive rate of 5% or less. Programmes which met these requirements were the quadruple test which measures the level of 4 different hormones in maternal blood, and the serum integrated test which measures the level of 5 different hormones in maternal blood across the first and second trimesters of pregnancy.

They also recommended that by 2007 screening programmes should offer a detection rate of greater than 75% with a false positive rate of less than 3%. Programmes meeting these requirements were the combined test (detailed above), and the integrated test which combines the NT measurement with the age of the mother, gestational age of the foetus, and level of 5 pregnancy hormones in both the first and second trimester. The integrated test performs better than the combined test, however the integrated test takes longer to perform due to measuring second trimester blood markers (the integrated test is ideally carried out at 15-16 weeks of pregnancy rather than 11-13 weeks for the combined test). The big jump in performance between 2004 and 2007 was made possible by the more widespread uptake of the nuchal translucency scan. The National Screening Committee 2010-11 Annual Report recommended that the preferred test from 2007-2010 where available was the combined test.

After screening, high risk pregnancies can be diagnosed using either amniocentesis or CVS. In an amniocentesis, a small amount of fluid is taken from the amniotic sac surrounding the fetus, while in a CVS a sample of cells is taken from chorionic villi on the placenta. Amniocenteces are generally performed between 15 and 20 weeks, compared to 10-12 weeks for CVS. However, early amniocenteses may be performed between 11 and 13 weeks at higher risk to the fetus. Performing invasive diagnostic tests carries a small risk of miscarriage, so it is important to only test those pregnancies which are at high risk (Wilson *et al* 1998, Rhoads *et al* 1989). A recent 11 year national registry study in Denmark (Tabor *et al* 2009) estimated that after performing an amniocentesis or CVS, 1.4% and 1.9% of unaffected foetuses respectively were lost.

However, in future it may be that diagnosis can also be performed using noninvasive methods. In 2012 Zimmermann *et al* reported the development of a noninvasive diagnostic test to detect aneuploidy of chromosomes 21, 18, 13, X and Y. Cell-free fetal DNA was analysed using 11,000 genetic markers on the 5 previously mentioned chromosomes, and the result was that 145 out of 166 individuals in the study were correctly diagnosed (the remaining 21 were excluded after failing a DNA quality test). Successful adoption of this technique would allow foetuses to be safely diagnosed without any additional risk of miscarriage as a result of the test.

Section 1.1.6: Genetic Mechanisms

Human trisomies are caused by errors in cell division which result in there being a whole or part of an extra chromosome in the cells of an affected individual. There are two types of cell division in humans; mitosis which is the production of two identical daughter cells from a single parent cell, and meiosis which is the production of four unique daughter cells from a single parent cell. Errors in either mitosis or meiosis can result in a trisomy. Mitosis is used by somatic cells (all cells in the body other than those producing sex cells) for growth and repair, while meiosis is used only by reproductive cells to produce gametes (sex cells). The human chromosome number is two, meaning that somatic cells in the human body are diploid (contain two copies of each chromosome). During the fertilisation process of sexual reproduction two gametes fuse and combine their DNA, so gametes contain half of the number of each chromosome in somatic cells and are referred to as haploid cells.

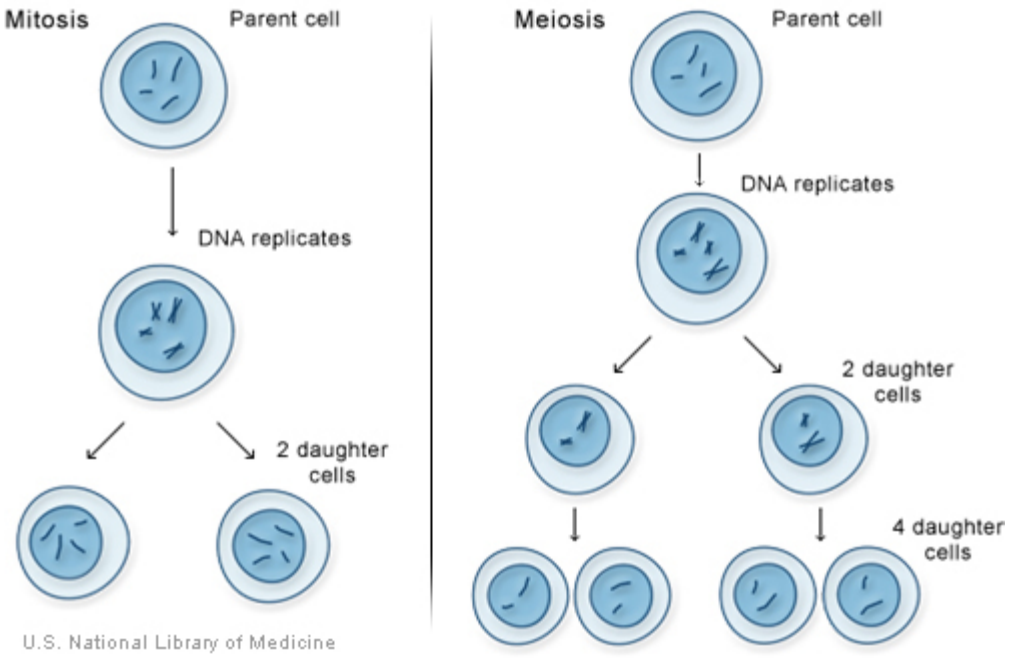

Figure 1.1.4.The key difference between mitosis and meiosis.

Figure 1.1.4 shows that meiosis initially follows the same steps as mitosis, but then has an additional cell division to halve the chromosome number. Both processes begin with a cycle of DNA replication to give each daughter cell the same number of chromosomes as the parent. The production of four daughter cells in meiosis allows for a greater degree of genetic variation in future generations, enabling adaptation and eventually evolution.

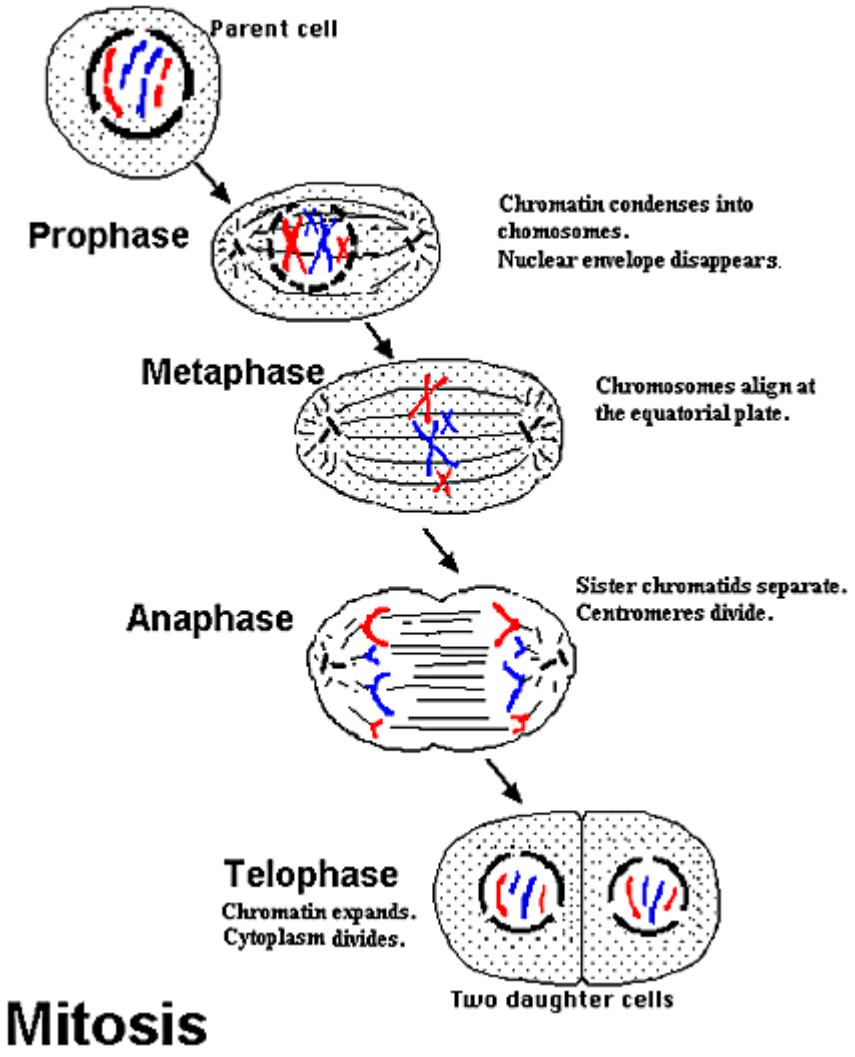

Figure 1.1.5.The 4 key stages of mitosis.

http://www.accessexcellence.org/RC/VL/GG/mitosis.php

Figure 1.1.5 shows two chromosomes going through mitosis. Suppose the larger chromosome is chromosome 1 and the smaller is chromosome 2. Similarly, let the red chromosomes be those inherited from the mother and the blue chromosomes be those inherited from the father. When DNA replicates before cell division it does so by creating sister chromatids, which are identical copies of every chromosome. Therefore at the start of mitosis there are actually 4 copies of each chromosome in the cell; 2 from the mother and 2 from the father. The sister chromatids are pulled to opposite ends of the cell, so that at the end of the cycle the daughter cells each have 1 maternal and 1 paternal copy of each chromosome, identical to the parent cell.

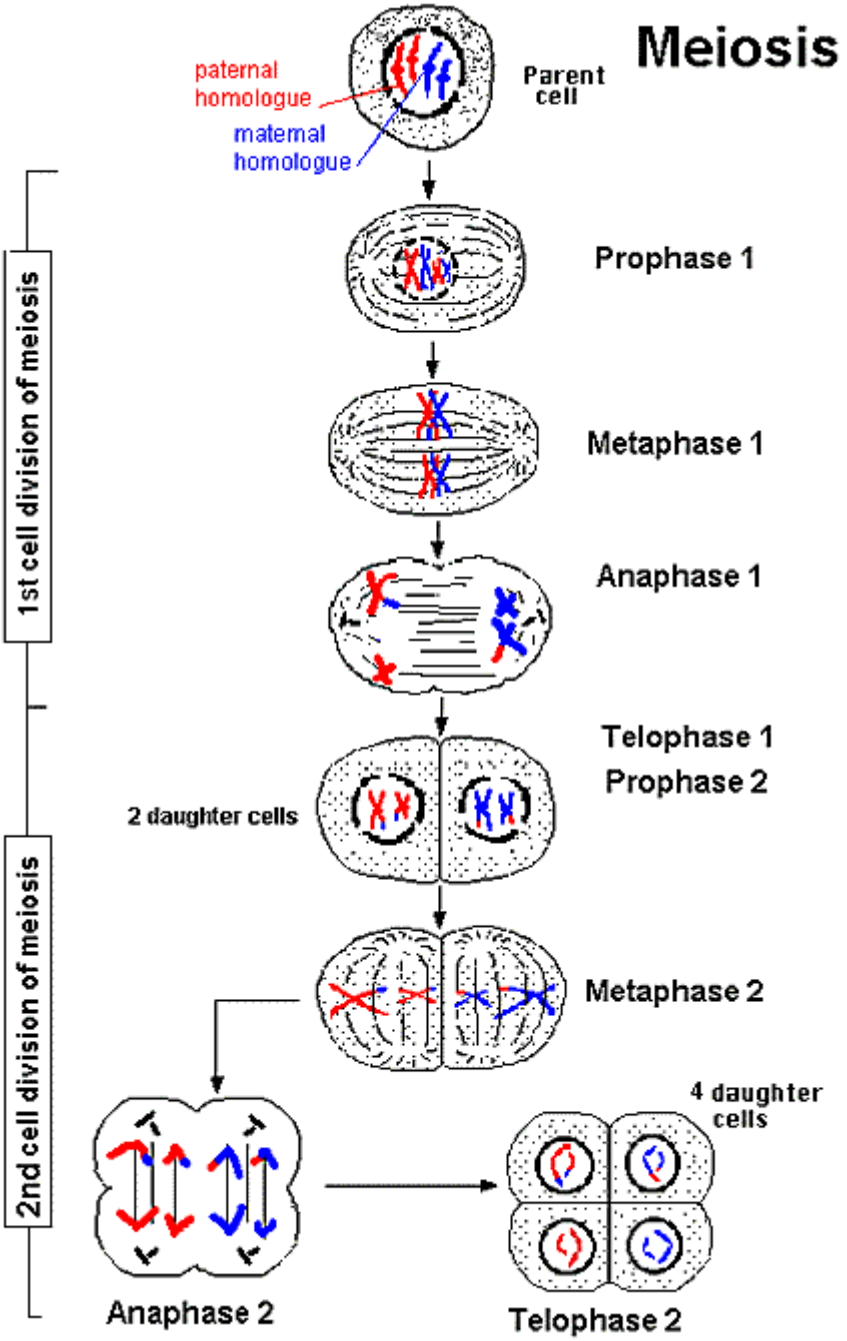

Figure 1.1.6.The key stages of meiosis.

http://www.accessexcellence.org/RC/VL/GG/meiosis.php

Figure 1.1.6 shows two chromosomes going through meiosis using the same stages as seen in mitosis. Again, suppose the larger chromosome is chromosome 1, the smaller is chromosome 2, red is maternally inherited and blue is paternally inherited. In meiosis, homologous chromosomes are separated in the first stage of cell division. This means that rather than 1 copy each of the maternal and paternal chromosomes, the daughter cells inherit both copies of either the maternal or paternal chromosome. This helps to ensure that the 4 daughter cells are all genetically unique.

The reason that the majority of trisomy cases arise due to errors in meiosis is that an error during gamete production will be present in all offspring cells if that gamete is involved in fertilisation. Errors later in development will result in only a fraction of the cells being affected.

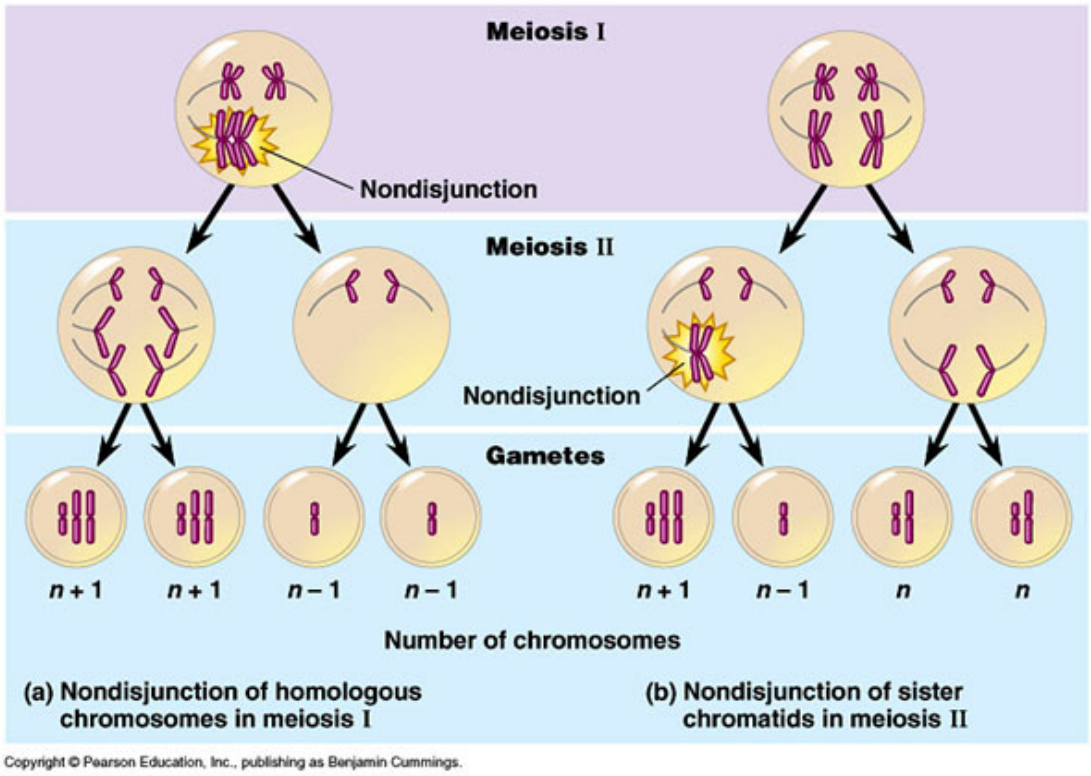

Figure 1.1.7.Two types of nondisjunction in meiosis.

The most common mechanism which can lead to a trisomy is nondisjunction, which literally means “not coming apart” and refers to pairs of chromosomes not separating properly during cell division. This problem largely occurs during meiosis and results in gametes having either 1 copy too many or 1 copy too few of one or more chromosomes. In meiosis there are two possible stages of nondisjunction, both of which are shown in figure 1.1.7. A nondisjunction in the first stage of meiosis will result in two gametes with one too many chromosomes and two gametes with one too few, while a nondisjunction in the second stage of meiosis will result in two normal gametes and one each with one extra and one missing chromosome. Nondisjunction during meiosis 1 accounts for around 96% of trisomy 21 cases, with 88% due to maternal (from the mother) nondisjunction and 8% due to paternal (from the father) nondisjunction (Nicolaidis& Petersen 1998), and it is reasonable to assume similar figures for trisomies 18 and 13. Mitotic nondisjuncture events during early embryonic development will result in some of the foetus’ cells being affected with a trisomy and some being unaffected.

Individuals carrying a mixture of affected and unaffected cells are called genetic mosaics. How early this process occurs determines the proportion of affected cells (Fishler& Koch, 1991).

The next most common mechanism is translocation, in which a partial trisomy can occur due to a chromosomal translocation where the long arm (q) of chromosome 18 (for example) is attached to another chromosome (often 14) and vice versa. An individual carrying this will appear normal as there are still only two copies of chromosomes 18 and 14, but any offspring who inherit this translocation will be affected by the trisomy. One variation of this process is a Robertson translocation, in which the two long arms of chromosomes fuse, losing the two short arms. Common Robertsonian translocations are limited to chromosomes 13, 14, 15, 21 and 22, as the short arms (p) of these chromosomes do not contain necessary genetic material (chromosomes with very small p arms are referred to as acrocentric). Around 1/3 of all Robertsonian translocations are between chromosomes 13 and 14.

Finally, in very rare events part of the chromosome can be duplicated, resulting in a partial trisomy. The severity and effects of the condition depend on how large the duplicated section was and which genes it contained.

Section 1.2: Risk Factors

Section 1.2.1: Maternal Age

Trisomies 21, 18 and 13 have one consistent risk factor, which is increasing maternal age (Hook 1981, Hassold& Jacobs 1984). This association is so strong that it was noticed in 1933 (Penrose), 25 years before it was discovered that Down Syndrome was caused by trisomy 21 (Hassold& Hunt 2001).

In 2010, Savva*et al* reported the age specific risks of having a pregnancy affected with trisomy 18 or 13 (age specific trisomy 21 risk was reported by Savva*et al,* 2006). The observed prevalence figures they obtained are shown in figures 1.2.1, 1.2.2 and 1.2.3 for trisomy 21, 18 and 13 respectively.

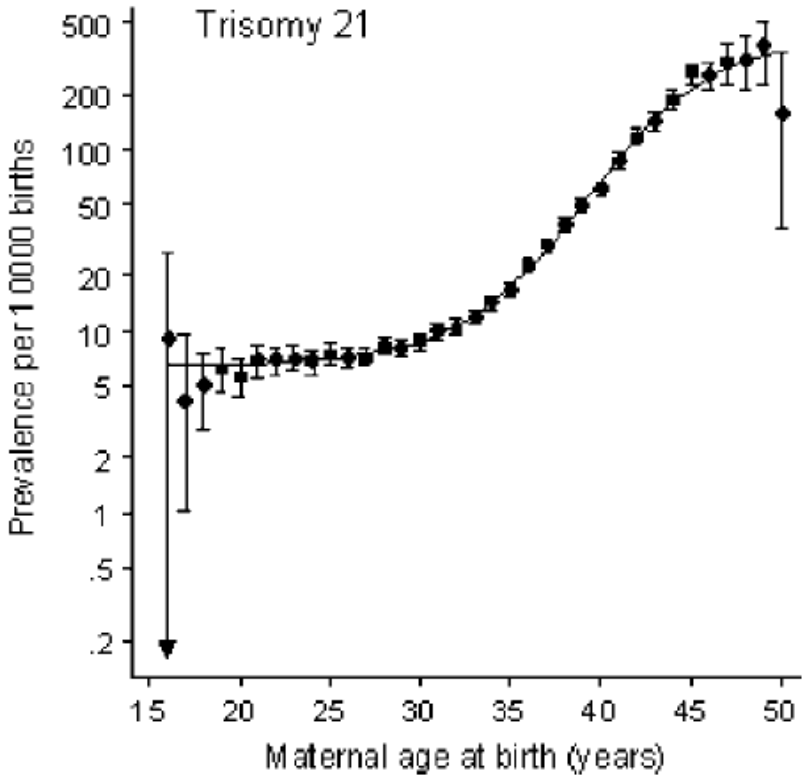

Figure 1.2.1.The age specific risk of having a pregnancy affected with trisomy 21.

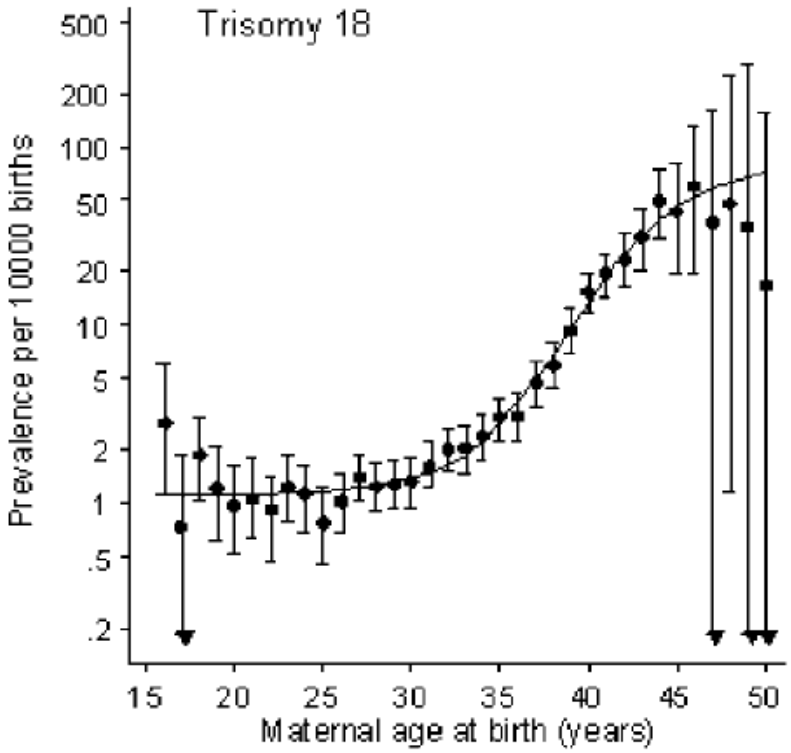

Figure 1.2.2.The age specific risk of having a pregnancy affected with trisomy 18.

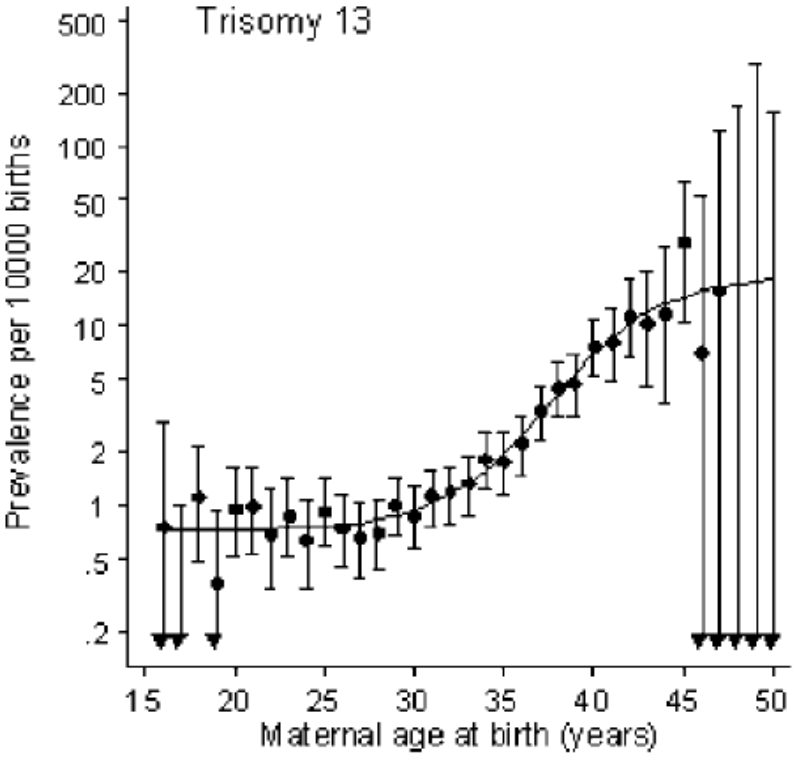

Figure 1.2.3.The age specific risk of having a pregnancy affected with trisomy 13.

Figures 1.2.1, 1.2.2 and 1.2.3 all show that the age specific risk of a trisomy stays low until about the age of 30, when it begins to increase. At the age of 35 the graphs become steeper, with the risk increasing rapidly until the age of around 45, when the graphs start to flatten out and the risk increases more slowly. All three graphs look similar, showing the scale of the maternal age effect. A woman aged 45 is around 40 times more likely to have a pregnancy affected with either trisomy 18 or trisomy 21 and around 20 times more likely to have a pregnancy affected with trisomy 13 than a woman aged 25.

There have been a variety of mechanisms proposed to explain how maternal age increases a woman's risk of having a pregnancy affected by a trisomy (Hassold& Jacobs 1984). The simplest mechanism is that, as a woman ages, cellular components which control DNA segregation weaken, increasing the likelihood of a nondisjunction event (Jones 2008). The two most likely components which degrade over time are:

- The protein complex controlling cohesion between homologous chromosomes in cell division, which is formed during initial egg production and must remain functional decades later. An error with this complex could result either in chromosomes not lining up properly in the centre of the cell or chromosomes not segregating when being pulled to opposite ends of the cell. The latter will result in a nondisjunction event like that seen in figure 1.1.7. In 2004 Revenkova*et al* demonstrated that mutation of the meiosis-specific cohesion complex component SMC1β in mice made both sexes sterile. However, in 2005 Hodges *et al* reported that the genetic material present in the eggs of the mutant female mice changed over time. As DNA takes the form of pairs of homologous chromosomes when meiosis is arrested, examining the proportion of DNA in this form in eggs of different ages reflects the effect of maternal age in humans. At 1 month old, 98.8% of chromosomes in eggs of the mutant mice were present as pairs of homologous chromosomes, while at 2 months old this number had dropped to just 35% (in healthy control mice 100% of chromosomes were present as pairs of homologous chromosomes at 6 months old, decreasing to 98.8% at 9 months old). This decease suggests that proteins involved in cohesion deteriorate over time and need to be replaced, a function which the mutant mice were unable to perform. Similar deterioration in humans could result in pairs of homologous chromosomes failing to properly segregate and the incorrect production of gametes.
- The spindle assembly checkpoint (SAC) which alerts the cell to any problems when chromosomes line up along the centre of the cell. When this system functions properly, any improper alignment of chromosomes is detected and cell division halted until the chromosomes align properly. In 2001, Steuerwald*et al* used quantitative

analysis to show that the concentration of two genes believed to be involved in this checkpoint (MAD2 and BUB1) was reduced in older women, which could lead to a failure in the meiotic checkpoint system and cells carrying the incorrect number of chromosomes being allowed to continue through meiosis.

In 1996, Lamb *et al* proposed a "Two hit" model for trisomy 21, theorising that two separate events are required to result in an age-dependant trisomy. The first was the age-independant establishment of a poorly recombined pair of homologous chromosomes during meiosis 1, and the second was the inability of the cell to detect this failure, allowing chromosomes to segregate incorrectly when dividing. As decades elapse between the two stages, the second stage is highly age dependant, and allows for the reduction in concentration of cell cycle components or the failure of cohesive units as previously described.

In 2010, Hulten*et al* proposed the Oocyte Mosaicism Selection Model (OMS). Rather than the risk of nondisjunction increasing over time, this model suggests that a proportion of oocytes (immature egg cells in women) carry a trisomy when they are first produced. The affected oocytes develop more slowly than unaffected oocytes, so that over time the proportion of affected oocytes increases as more unaffected oocytes are lost. As this proportion increases so does the probability of one of the affected oocytes being released and subsequently fertilised, leading to a trisomic pregnancy. Figure 1.2.4 shows the theorised proportions of both types of oocyte over time. This model is particularly interesting because it accounts for the trisomy risk levelling out at the age of about 45, which other models do not explain.

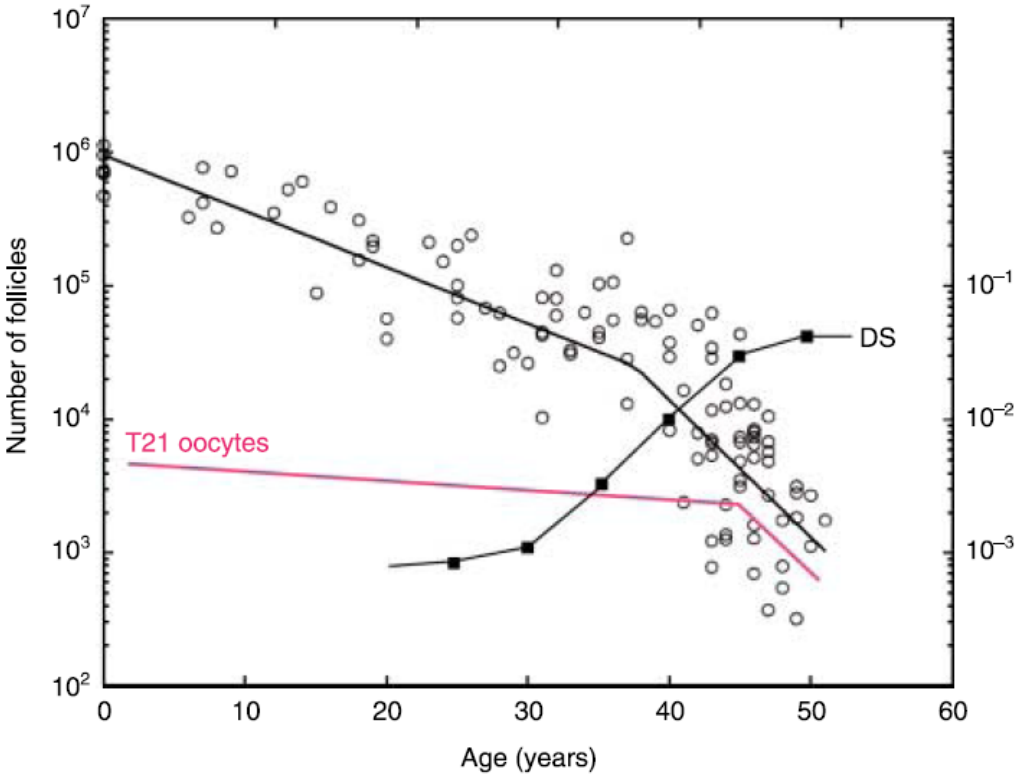

Figure 1.2.4. The Oocyte Mosaicism Selection Model (Hulten*et al*, 2010). The red line marked T21 oocytes represents the number of oocytes carrying trisomy 21, while the line marked DS shows Down Syndrome prevalence.

Section 1.2.2: Ionising Radiation

Besides maternal age, the most commonly linked risk factor to the human trisomies is ionising radiation. The opportunity to first study the relationship between radiation and DNA damage arose in 1945 during the second world war when the United States of America conducted atomic bombings of the cities of Hiroshima and Nagasaki in Japan. In Hiroshima, approximately 70,000 people in were killed directly as a result of the bombing, but a further 20,000 – 96,000 died by the end of 1945, and some estimates state that as many as 200,000 had died as a result of the immediate and long-term effects of the blast by 1950. Those who died in the aftermath of the bombings died from a variety of conditions, including burns, cancer and radiation poisoning. Those who survived exhibited similar symptoms, and it was believed that the radiation they had been exposed to had affected their DNA structure in some way. In the early 1960s there were reports linking DNA mutations to ionising radiation from X-Rays and nuclear reactors, and a few years later the same results were found amongst survivors of the atomic bombings.

In 1966 Bloom *et al* reported that among 94 exposed survivors of the bombings (51 from Hiroshima and 43 from Nagasaki) and 94 matched controls who were all under 30 years old at the time of the atomic bombings, complex chromosomal abnormalities (either breaks in chromosomes or large scale chromosomal rearrangements) were found in 0.6% of cells of the exposed and 0.01% cells of the controls. A year later Bloom *et al* presented similar findings among those over 30 years old at the time on the bombings, using 77 heavily exposed survivors of the atomic bombings and 80 controls. The second report showed complex chromosome aberrations in 61% of the exposed cases and 16% of the controls. Among these individuals, the abnormalities were found in 1.5% of cells in the cases and 0.3% of cells in the controls. In 1971 Awa *et al* demonstrated that the frequency of cells with radiation-induced chromosome damage was proportional to radiation dose. These findings were important in establishing a link between ionising radiation and chromosomal damage, and in the future this would lead to investigations into how radiation exposure affected a woman's chance of having a pregnancy affected by a trisomy. However, due to the high mortality rates in the aftermath of the bombings there were not enough babies born in affected areas of Japan in the following years to investigate radiation as a risk factor for the trisomy syndromes.

Chernobyl

On April 26th 1986, the disaster at the Chernobyl nuclear power plant sent a plume of radioactive fallout into the air. This plume drifted over large parts of Eastern Europe, contaminating most of Europe. Most severely affected were Belarus, Ukraine and parts of Russia, as shown in figure 1.2.5. As ionising radiation was a known mutagen by this time, this disaster gave investigators a chance to examine the proposed link between ionising radiation and Down syndrome. Any increase in the prevalence of Down syndrome could occur in the short term when radiation levels were highest or in the long term after prolonged radiation exposure.

The map below shows the distribution of radiation around Europe from Caesium-137 in the aftermath of the Chernobyl disaster. Particularly high levels of radioactive fallout can be seen in the countries neighbouring the Ukraine, as well as parts of the Scandinavian countries, South Germany and Austria. Caesium-137 was first created during atomic weapons testing and does not occur naturally (Okumura2003), and has been found to be lethal when injected into Beagles (Redman *et al* 1972).

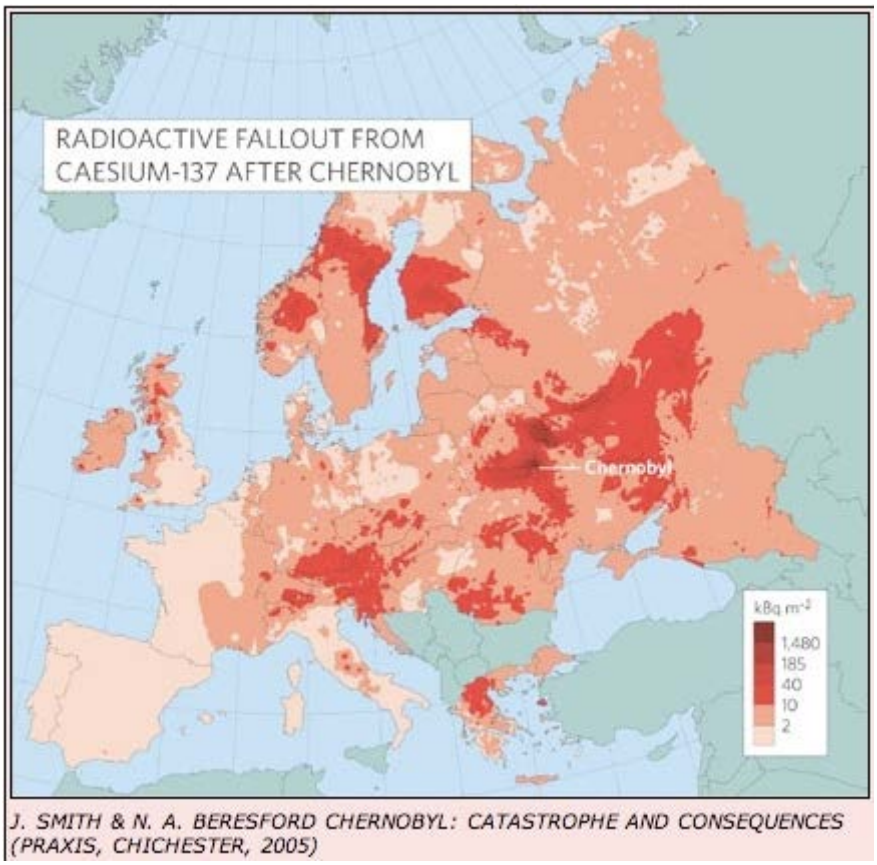

Figure 1.2.5. Map of fallout levels from Chernobyl.

Reports were published from different areas of Europe throughout the 1990s indicating clusters of Down Syndrome in early 1987, although the authors of these all noted that the increase could not be definitely linked to Chernobyl. In 1994 Ericson and Kallen reported that more exposed areas of Sweden (those with at least 5 kBq/m^2) showed a Down syndrome risk ratio of 1.47 (95% CI 1.01 – 2.14) in 1987, falling to 1.08 in 1988 and 1.00 in 1989. Analysing only the most severely affected areas (those with at least 30 kBq/m^2 of radiation), the relative risk rose to 1.61 (95% CI 0.80 – 3.24), but the reduced number of cases led to wider confidence intervals and the result was not statistically significant.

A similar finding was reported by Ramsay *et al* in 1991, showing that there were more Down Syndrome cases in the Lothian region of Scotland in 1987 than any other year between 1978 and 1989 (26 cases in 1987 compared to the next highest annual total of 15 in 1986 and 1988). This increase was significant when compared to the mean from the whole time period ($p < 0.001$). It was interesting to note that in Scotland, the most severe increase in Down syndrome risk was observed in mothers aged 35 or older, suggesting that cells in older women are the most susceptible to radiation damage. However, the authors of this study believe that the levels of ionising radiation in Scotland from Chernobyl were too low to cause any extra cases of trisomy 21, stating that a greater understanding is needed of the effects of low level radiation exposures.

There were also two important studies carried out in Germany. The first was in Berlin (Sperling*et al*, 1994) and showed that in January 1987 there was a Down Syndrome prevalence of 6.8 per 1,000 live births, a large increase from the usual prevalence of 1.56 per 1,000 births. The second was in Bavaria (Burkart *et al*, 1997), and showed that the Down Syndrome prevalence peaked in December 1986 at 2.45 per 1,000 births compared to the usual rate of 0.94 per 1,000 births. The Berlin cluster could be due to Chernobyl radiation as it took place exactly 9 months after the disaster. The same cannot be said of the Bavaria cluster because it occurred a month too early, and the monthly Down Syndrome prevalence figures for the region show similarly sized peaks earlier in 1986 and in 1985. The Burkart paper also reported that the Chernobyl radiation in Germany was at a low level and would be expected to affect fewer than 0.5% of foetal cells. This would suggest that radiation is not the cause of the observed cluster in Bavaria, and instead this could merely be the result of random variation in the monthly Down syndrome prevalence.

In 2007, Zatsepin *et al* published an article showing Down Syndrome time-clustering in Belarus in January 1987. The close proximity of Belarus to Chernobyl meant that large parts of the country were exposed to high levels of ionising radiation in the short term and low levels in the long term, so it is an ideal place to see if there is a link between radiation and Down Syndrome. They reported that in January 1987 there was a statistically significant cluster of 26 cases

observed compared to 9.84 expected (observed: expected ratio of 2.64; 95% CI 1.72-3.76), however there were no apparent long term effects and in February 1987 the prevalence had returned to usual levels. The Down syndrome prevalence during January 1987 was substantially larger than during any other month between 1981 and 1992, and the highest prevalence was observed in the most contaminated areas.

Table 1.2.1 shows the time and place of all 5 of these significant trisomy 21 clusters, with observed and expected numbers of Down syndrome births. Expected Down syndrome figures were reported in each article, and were calculated based on the previously observed Down syndrome prevalence and the number of births in the specified time period. The locations of each cluster are then shown in figure 1.2.6. These results stand to date as the strongest evidence of a trisomy risk factor besides maternal age.

Table 1.2.1. Comparison of studies of Down Syndrome clusters due to Chernobyl radiation.

				DS Infants		
Reference	Year	Cluster Time	Location	Observed	Expected	O/E
Ericsen	1994	1987	Sweden	30	15.87	1.89
Ramsay	1991	1987	Scotland	26	11.22	2.32
Sperling	1994	Jan-87	Berlin	12	2.75	4.36
Burkart	1997	Dec-86	Bavaria	10	4.4	2.27
Zatsepin	2007	Jan-87	Belarus	26	9.84	2.64

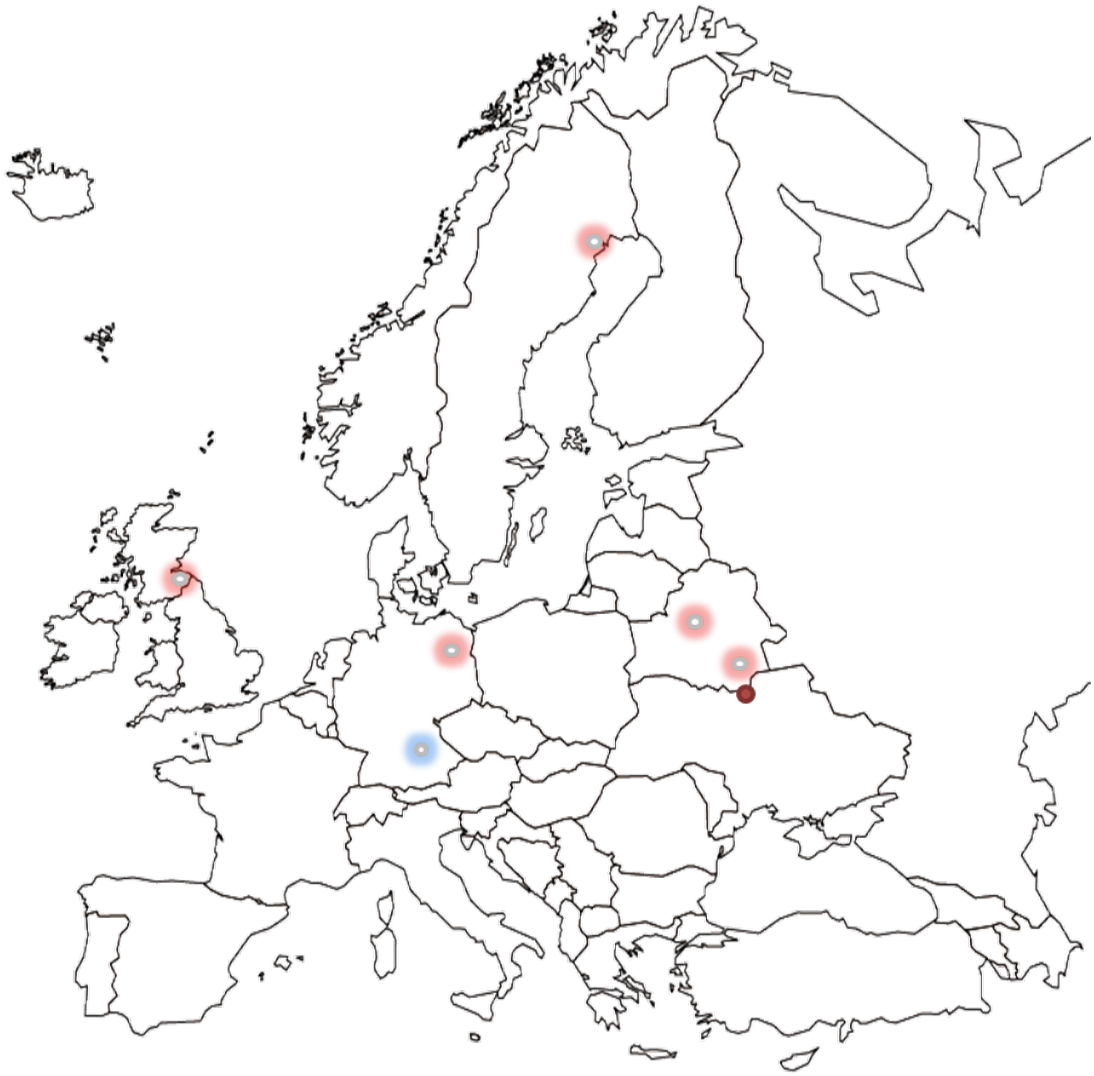

Figure 1.2.6. Map of Europe showing potential trisomy 21 clusters arising due to Chernobyl radiation. The red outlined areas denote clusters observed in January 1987, the blue shaded area denotes a cluster observed in December 1986 and the solid red circle is the location of Chernobyl.

Semipalatinsk

Between 1949 and 1989, the Soviet Union performed 456 nuclear tests at the Semipalatinsk test site in North East Kazakhstan. The location of this site is shown in figure 1.2.7. These tests were comprised of atmospheric and surface explosions from 1949-1963, and underground tests from 1963-1989. Since the testing was ceased in 1989, the population of the irradiated zone has been a unique resource in terms of examining the effects of radiation on the body.

Figure 1.2.7. Map of Kazakhstan showing the Semipalatinsk nuclear test site. The test site is the red shaded area, and is of a similar size to Wales. Map taken from http://www.spiegel.de/international/world/0,1518,741679,00.html

In 2003, Abil'dinova *et al* studied 149 individuals who were born and had permanently lived in the irradiated zone, and reported that the frequency of mutated cells in this group was 1.7-3.0 times higher than that in an unexposed control group. The authors also noted that the highest mutation rates were observed in cases born in the irradiated zone before 1960, as this was when air and surface nuclear tests were still being performed. A similar study was carried out in 2002 by Svyatova *et al*, which demonstrated that populations exposed to higher levels of radiation displayed higher mutation rates. There was a statistically significant increase in the mutation rate from a population at low radiation exposure to a population at very high exposure.

These findings were followed by an article in 2002 by Dubrova *et al*, who compared 40 three-generation families inhabiting rural areas around the nuclear test site with 28 three-generation families from a geographically similar (but non-irradiated) region of Kazakhstan. It was reported that 85% of the effective radiation dose for the case families was attributable to four surface explosions in 1949, 1951, 1953 and 1956. These shall henceforth be referred to as major events. Amongst parents in the first generation, who were living in the irradiated zones for all four of the major events, there was a statistically significant 1.8-fold rise in the cell

mutation rate. Among the second generation, some of whom were born after the four major events had occurred, the rise in mutation rate had decreased to 1.5-fold (but remained statistically significant). Among those born after 1960, when the major events had already occurred (air and surface testing stopped altogether in 1963), there was no significant rise above the mutation rate among the control group. These findings suggest that long term, low level radiation exposure does not cause significant DNA damage, which in turn could mean that the Down Syndrome prevalence would be expected to be no higher than unexposed areas of Kazakhstan.

These findings suggest that if there was ever an increase in Down Syndrome prevalence in the irradiated region, it would have occurred in the most heavily exposed zones between 1949 and 1960. However, it was only discovered in 1958 that Down Syndrome was caused by trisomy 21, and without cytogenetic technology it could have been very difficult for doctors in the region to accurately diagnose Down Syndrome among the other congenital anomalies and malformations that affected newborns. Information regarding Down Syndrome prevalence is therefore very scarce, but a technical report for the United States Defence Threat Reduction Agency (Balmukhanov *et al*, 2006) was published that referenced an article from 1993 (Kuderinov *et al*) which studied 490 children born to permanent residents of a highly contaminated area around the test site and estimated a Down Syndrome prevalence of 12%. There is no more information available as the original article cannot be located. The same technical report also reported Down Syndrome prevalence in two other highly contaminated areas in 1995 as 1 case in 408 births and 1 case in 653 births. This prevalence is roughly double that in England and Wales in 2010 (Morris, NDSCR Annual Report 2010), but such small numbers cannot be relied upon as accurate evidence. In 2006, Chebotarev *et al* reported that the Down Syndrome prevalence in Semipalatinsk between 1979 and 1996 was 0.11%, only slightly higher than that in England and Wales in 2010. There may have been an increased Down Syndrome rate in Semipalatinsk due to the nuclear tests, but information is both scarce and unreliable.

Other Radiation Studies

Chernobyl fallout is not the only proposed link between ionising radiation and Down syndrome. In 1995, Bound *et al* examined data from Lancashire in the North West of England between 1957 and 1991 for trisomy 21 clusters, and discovered a strong peak among cases conceived between October 1962 and November 1964. During this time period, the Down syndrome prevalence rose from 0.0012 to 0.0031 (p = 0.01). A secondary peak in late 1958 was found to be non-significant (p = 0.1). Neither of these peaks was due to a change in maternal age. These clusters are interesting because they coincide with high levels of atmospheric radiation in the sample area due to the testing of atomic weapons, the maximum level of which occurred between 1963 and 1964, with the next highest level in 1958. Graphs showing the coinciding peaks of Down syndrome prevalence and radiation levels are shown in figures 1.2.8 and 1.2.9.

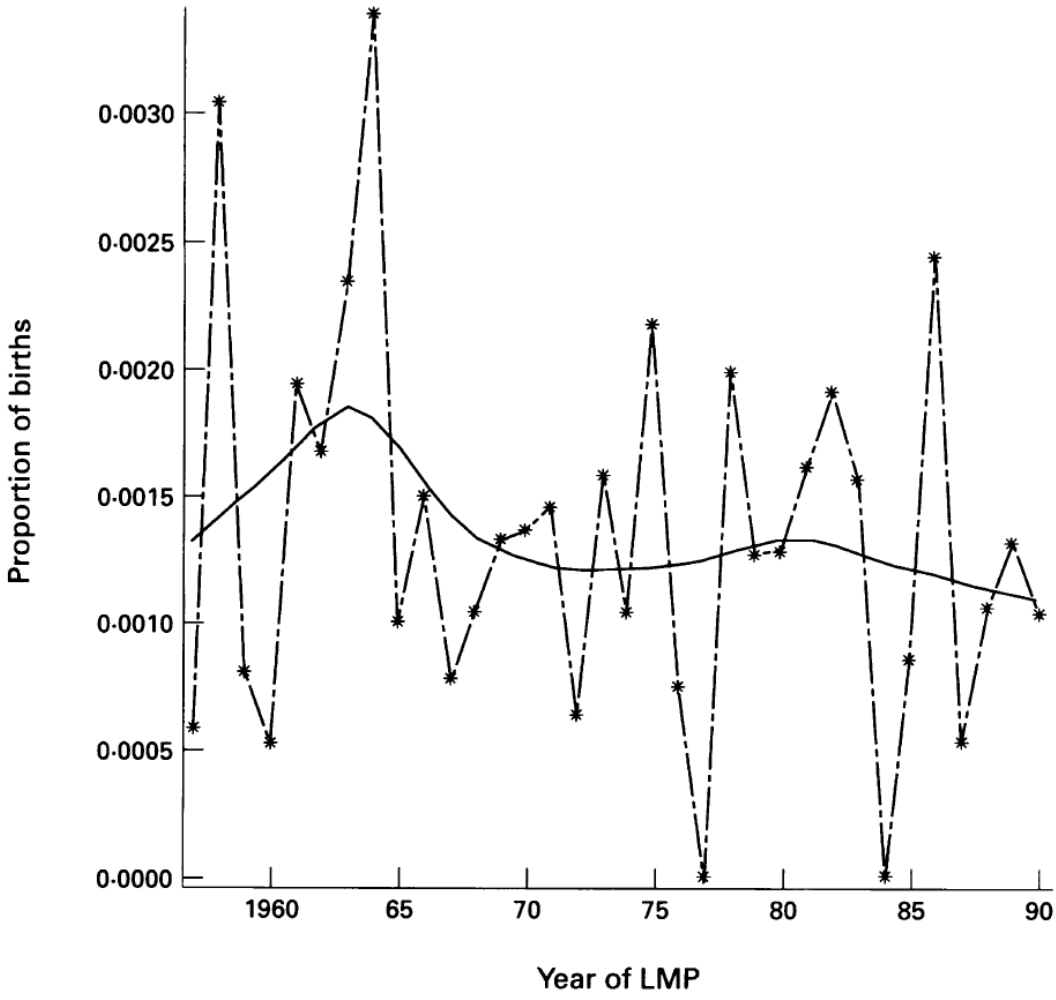

Figure 1.2.8. Prevalence of Down Syndrome cases by year. LMP stands for last menstrual period.

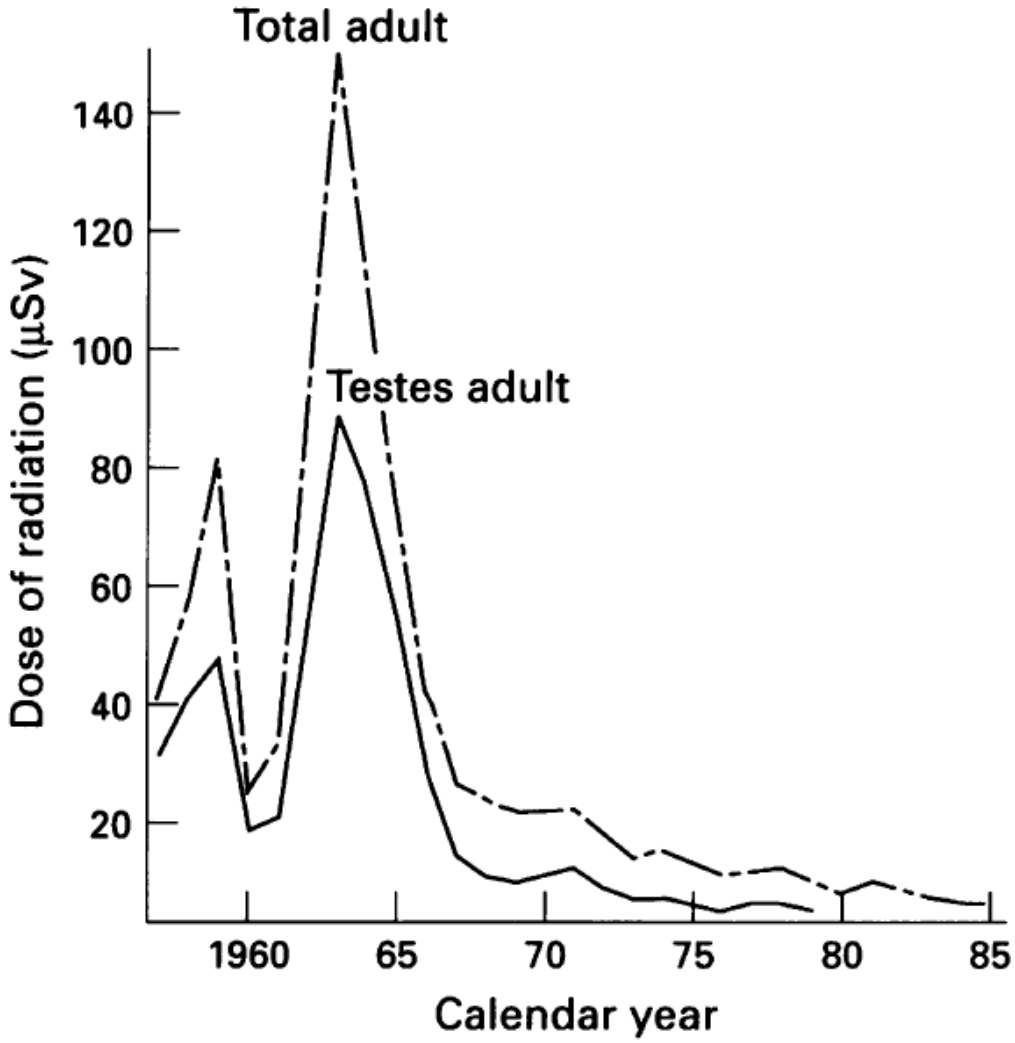

Figure 1.2.9.Estimated dose equivalents of radiation from weapons fallout.

A study carried out by Verger in 1997 examined the link between ionising radiation and Down syndrome prevalence in two zones of Southern India, one with high levels of naturally occurring radiation (14 to 28 mSv per yr) and a control area with much lower levels of radiation (0.9 mSv per yr). There was no significant difference between the two areas in terms of age and sex distribution, mean maternal age, mean family size and abortion rate. The high radiation zone had a trisomy 21 prevalence of 0.9% among 12,918 persons compared to the control zone which reported no trisomy 21 cases among 5,938 persons. For comparison, the region of Madras in Southern India has a Down syndrome prevalence of 0.26%. Verger also compared the rate of chromosomal aberrations (anomaly type and cell not specified) between 46 exposed subjects and 39 controls, and found that the mean frequency of aberrations was significantly higher in exposed individuals (1.9 ± 3.1 vs. 0.2 ± 0.6, $p < 0.01$).

These results, along with those from the Chernobyl based research, provide strong evidence that trisomy 21 (and by association, trisomies 18 and 13) is associated with high levels of atmospheric radiation. Exposure to ionising radiation from accidents, such as Chernobyl, are believed to pose only a very small additional risk for many potential mothers, but for those living in areas with high levels of natural or artificial radiation this may be sufficient to greatly increase their risk of having a pregnancy affected by a trisomy.

Section 1.2.3: Other Explanations

Ionising radiation has been investigated extensively with regard to Down syndrome, but there have been very few other risk factors proposed. In 2008, McNally *et al* examined data taken from the Northern Region of England between 1985 and 2003 for space-time clustering in trisomies 13, 18 and 21, finding statistically significant evidence of trisomy 21 clustering only. This study used both a fixed boundary approach (cluster scans take place using a set of pre-selected critical distances) and a nearest neighbour approach (uses distance between pre-selected numbers of individuals rather than fixed distances, and included 1084 cases of trisomy 21, 240 cases of trisomy 18 and 116 cases of trisomy 13. Cases of trisomy 21 were found to cluster in more densely populated areas ($p = 0.01$ by geographical distance and $p = 0.02$ by nearest neighbour), whereas there was insufficient evidence of clustering in either trisomy 18($p = 0.37$ and $p = 0.06$, respectively) or trisomy 13($p = 0.57$ and $p = 0.19$, respectively). Investigators hypothesised that transient environmental factors, such as an infectious agent, might explain these study findings. The theory of an infectious agent, which can be spread more easily in areas of high population density, is consistent with the finding that clustering of trisomy 21 was associated with cases from more densely populated areas and is a plausible (and potentially replicable) explanation of this result.

There have also been several trisomy clusters recorded with no hypothesised risk factor. These were observed clusters which did not use any specialised statistical methodology to identify them. In 2000, Dean *et al* reported a trisomy 21 cluster in offspring of former pupils of a girls' school in Ireland in 1956-57. There were 387 live births to such women, amongst which were 6 cases of trisomy 21 compared to 0.69 expected ($p < 10^{-4}$). 5 of these births were to women under the age of 30, compared with 0.15 expected ($p < 10^{-6}$). The cluster was confined to that particular year group at that school only, with no increase seen either in the surrounding area of Ireland or in other schools. It was suggested that an influenza epidemic or contamination from a nearby nuclear reactor fire could be responsible (both of which occurred in October 1957), but 3 of the women had left the school and moved away by the time the accident occured (while another of the cases arose due to an error after fertilisation). Despite the apparently high level of statistical significance, the characteristics of the cases led the investigators to believe that this cluster arose purely due to chance.

In the late 1970s there were observed clusters of trisomy 13 and trisomy 18 in hospitals in the USA. Warburton *et al* (1977) found a trisomy cluster in New York in the winter of 1976, Pai*et al*(1978) found a trisomy 13 cluster in Maryland between August 20 and October 31 1977 and Barsel-Bowers *et al*(1980) found a trisomy 18 cluster in Rhode Island and Massachusetts

between November 8 1977 and May 12 1978. They suggested no biological explanation reason for these clusters.

Section 1.2.4: Strength of Evidence for Risk Factors

While there appears to be strong evidence to link high levels of atmospheric radiation to trisomy risk, the lack of biological explanation has led investigators to advise caution regarding this link. Levels of radiation were believed to be insufficient to cause several of the trisomy 21 clusters reported in the aftermath of the Chernobyl meltdown, with even the investigators of the cluster in nearby Belarus stating that the involvement of radiation was far from certain. The study by Bound *et al* (1995) reported a Down Syndrome cluster in Lancashire, England between 1962 and 1964 which coincided with high levels of atmospheric radiation in the area due to the testing of atomic weapons. All of this evidence shows time-specific increase in trisomy 21 prevalence which coincided with high levels of atmospheric radiation, and as such it is possible that the findings are merely coincidental.

The discovery of clusters by Verger *et al* (1997) in India provides evidence of a different nature. This study compared 2 similar populations, one of which was situated in an area of high natural radiation and the other in an area of low radiation. Results showed that the rate of chromosomal aberrations was significantly higher in the population which was exposed to more radiation. Besides radiation, no biological link has been proposed to explain this difference. Overall, there is evidence that exposure to radiation can increase trisomy risk both in the short term and the long term. However, it appears that the level of radiation present in order to cause clusters must be very high. The Chernobyl meltdown sent a strong burst of concentrated radiation around Europe, while the difference in atmospheric radiation between the two regions of India in the Verger study was somewhere between 15:1 and 30:1. Data analysed in this thesis was taken from England and Wales between 2004 and 2010, during which time there has been no large scale catastrophic nuclear event such as Chernobyl. There is also no previous evidence to suggest that trisomy clusters in England and Wales have been caused by high levels of natural radiation. Therefore, while the evidence linking radiation and trisomy risk is strong, it may be difficult to repeat that observation in this thesis.

The McNally *et al* (2008) study in the North of England found space-time clustering in trisomy 21 and suggested an infectious agent as the cause, but no clusters were detected for trisomies 18 or 13. Given that all 3 syndromes are caused in a similar way, it is unlikely that clusters would exist for one but not the others. It is more likely that the data used in this investigation was too sparse to have sufficient power to detect clustering of trisomies 18 and 13 (especially if the clustering is weak), a possibility which is briefly discussed by the authors. Their 19 years of data (1985 – 2003) included 240 cases of trisomy 18 and 116 cases of trisomy 13, which both fall far short of the 1,084 cases of trisomy 21 in the same study. The greater number of

cases available in the NDSCR should make it possible to detect weaker clusters in the study presented in this thesis, although it is still possible that even the NDSCR data has insufficient numbers to detect very weak clusters caused by risk factors with a small effect size. The size of effects which may be detected using different sample sizes is evaluated further in chapters 3 and 4.

Section 1.3: Introduction to Cluster Detection Techniques

Section 1.3.1: Why Look For Clusters?

In disease surveillance, the presence of clusters can indicate the presence of a localised excess risk which leads to the presence of more affected individuals. This excess risk can be localised either in time, in which case temporal clusters would be observed, or in space, in which case spatial clusters would be observed. For example, a factor which only affected risk in winter would be expected to produce seasonal, temporal clusters while a risk factor resulting from a rare environmental exposure (such as a hazardous landfill site) would be expected to produce spatial clusters. Some risk factors would be expected to produce clusters in space and time, such as the trisomy 21 clusters resulting from the Chernobyl disaster. In that instance, all clusters were observed in the same time period (January 1987), but the strongest clusters were observed closest to the meltdown site.

Aside from the trisomy 21 clusters arising from Chernobyl, evidence was presented in section 1.2.3 (McNally *et al*, 2008) suggesting that trisomy 21 clustering in the North of England may be caused by an infectious agent. The same study did not find any evidence of clustering in trisomies 18 and 13, however this may be attributable to low sample sizes for both syndromes. The NDSCR is a national, high quality data source which collects information on all cytogenetically confirmed cases of trisomies 21, 18 and 13 in England and Wales, and has higher sample sizes available than those used in the McNally *etal* study. Scanning the NDSCR data would therefore have greater power to detect clusters, potentially arising due to exposure to infection or some other environmental hazard.

For clusters of some diseases, it is necessary to adjust for previously known risks, the effects of which could mask the effect of alternate risk factors. In the case of trisomy clusters, adjustment for maternal age is required before searching for clusters, as the strong effect of maternal age would be expected to produce age dependant clusters. While maternal age is the strongest risk factor for the trisomy syndromes, other risk factors have been linked with the conditions, most notably radiation arising from the Chernobyl disaster (detailed in section 1.2.2). Other highly radiated areas also appeared to show increased trisomy prevalence. Since radiation related clusters have only been reported in extreme circumstances (either through a catastrophic event or an abnormally high level of naturally occurring radiation), it is not necessary to adjust for radiation levels. It would also be difficult to obtain accurate data regarding radiation levels in England and Wales. However, any regional differences in prenatal

screening procedures (such as some areas screening earlier than others) could produce clusters and need to be accounted for.

Using cluster analysis, the aetiology of multiple diseases has been improved. In 1997, Kulldorff *et al* used the spatial scan statistic (*Kulldorff & Nagarwalla* 1995, further explanation in chapter 1.3.3.1) to perform scans of breast cancer cases from 1988-1992 in the Northeast USA, discovering several significant clusters in multiple cities and providing further evidence of the involvement of several previously hypothesised risk factors for which the investigators could not account in the analysis. In 2013, Massenet *et al* performed spatial and temporal cluster analysis of tuberculosis cases in the city of Saint-Louis in Senegal, identifying a densely populated fishing area as a highly likely cluster for further study.

Cluster analyses have also been used in biosurveillance to predict disease outbreaks. Mostashari *et al* (2003) proposed an early warning system for the emerging threat of West Nile virus (WNV) in the Eastern USA, which used spatial cluster analyses of dead bird numbers to identify areas of increased human risk. WNV is most commonly spread by mosquitos and birds, so an elevated number of dead birds in an area could indicate the increased presence of WNV. This method was implemented prospectively in 2001, enabling pre-emptive reduction of mosquito breeding 4 weeks before WNV activity was confirmed by laboratory analysis.

In 1999 English *et al* used a distance based clustering test (Cuzick & Edwards 1990, more details in chapter 1.3.3.3) to test for association between childhood asthma and traffic flow in a low-income population in San Diego county, California in 1993. Cases residing near high traffic flows were found to be more likely than those residing in low traffic areas to have 2 or more hospital visits for asthma during the year. Although there was no evidence of association between asthma risk and traffic counts, the study provided strong evidence that repeated exposure to air pollutants from traffic exhaust exacerbates symptoms in individuals already diagnosed with asthma. The same method was used in 2011 by Schmeidel *et al* to test for clustering of leukaemia and type 1 diabetes (T1D) in children aged 0-14 years Denmark. Their discovery of clusters of both leukaemia and TID within both the defined subsets and whole age range led the investigators to propose that the diseases shared etiological factors which varied by geographical location, including exposure to infections and agents which may compromise the immune system.

Section 1.3.2: Temporal Cluster Detection

In 1965, the first temporal scan statistic was proposed by Joseph Naus. This method scanned through one dimensional temporal data using a scanning window of fixed size to identify the maximum number of events occurring within the window.

In 1995, Nagarwalla proposed a method entitled the "Scan Statistic with a Variable Window". This technique differed from those previously available because it used a scanning window of variable size and position, meaning that any investigators would not have to judge the size of potential clusters before performing the analysis. The mechanics of the temporal scanning window are illustrated in figure 1.3.1.

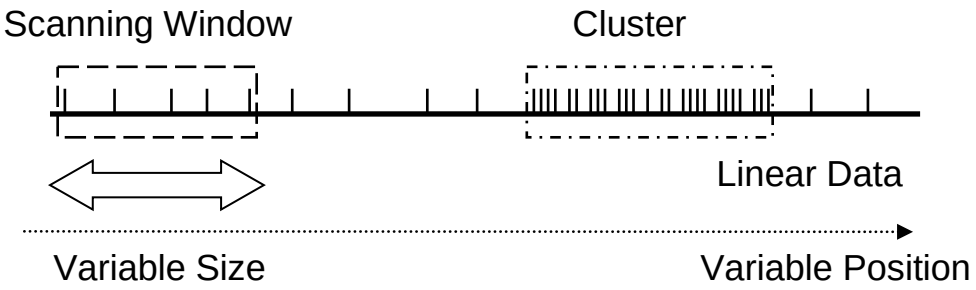

Figure 1.3.1.The Nagarwalla scanning window method of temporal cluster analysis.

The method scans through every combination of cases and finds the minimum amount of time for every possible cluster size to occur, starting at two cases and continuing until a predetermined limit is reached. This limit can be imposed by the investigator. It then generates scan statistics for those clusters using the formula below.

The statistical formula is a generalized likelihood ratio test for a null hypothesis of uniform distribution against an alternative hypothesis of non-random clustering, and is shown below.

$$\lambda[n] = \left(\frac{n}{N}\right)^{n}\left(\frac{N-n}{N}\right)^{N-n}\left(\frac{1}{d_{\min}}\right)^{n}\left(\frac{1}{1-d_{\min}}\right)^{N-n}$$

Where n = the number of cases in the cluster, N = the total number of cases, d_{min} = the minimum number of days for n cases to occur and $\lambda[n]$ = the scan statistic for that value of n.

Section 1.3.3: Spatial Cluster Detection

The detection of spatial clusters is a much-studied subject area, and there have been many different methods developed to enable such analysis. The most useful of these methods can be divided into two classes: quadrat-based and distance-based (Ripley, 1977). Quadrat-based methods take measurements of a population using overlapping spatial windows, where the most likely cluster location will be that which has the highest number of cases for the number of non-cases contained within the window. Distance-based methods (also referred to as a nearest-neighbour approach) examine the distance between each case and its nearest neighbour, takes the average of all of these nearest neighbour distances and compares this with the expected figure of a hypothetical random spatial distribution.

Quadrat-Based Methods

The quadrat-based methods are well suited to searching for disease clusters in an unevenly distributed population as the method accounts for the population distribution. The data must include cases and a background population at risk, and is usually divided into small administrative zones for ease of analysis. All cases and non-cases within each zone are assigned to a specified point within the zone (usually the administrative centre). A standard quadrat method would then involve placing a grid over a map of the study region and counting the number of cases and non-cases within each square, where the null hypothesis states that cases are evenly distributed throughout the study region and the alternative hypothesis states that the cases cluster in one or more squares. However, greater power can be achieved by generating a series of overlapping quadrats instead as this improves the coverage of the study region and makes it less likely that a cluster will be missed due to the positioning of the grid.

The first attempt to use this type of method to discover clusters of a rare disease was detailed in 1987 by Openshaw et al, who examined 853 cases of acute lymphoblastic leukaemia (ALL) in the North of England diagnosed before their 15th birthday between 1968 and 1985, using a background population of 1,544,963 children registered within the study area in the 1981 UK census. They proposed the Geographical Analysis Machine (GAM), which laid a closely spaced grid over the study region and generated circles with radii from 1km to 25km at each grid point (figure 1.3.2). This method was particularly useful because it avoided a pre-selection bias by scanning the entire study area equally. The number of leukaemia cases within each circle was recorded and statistical significance was judged using Monte Carlo testing, which involved generating 499 sets of data in which 853 individuals in the population were randomly selected to have leukaemia. Any circle containing two or more cases which was more extremely ranked (i.e. contained more leukaemia cases) than the 499 equivalent simulated circles was judged to be statistically significant at the 0.002 level. In total there were 812,993 circles generated,

1,792 of which were significant, most of which were clustered around five distinct areas. Only one of these areas had previously been identified as a leukaemia hotspot.

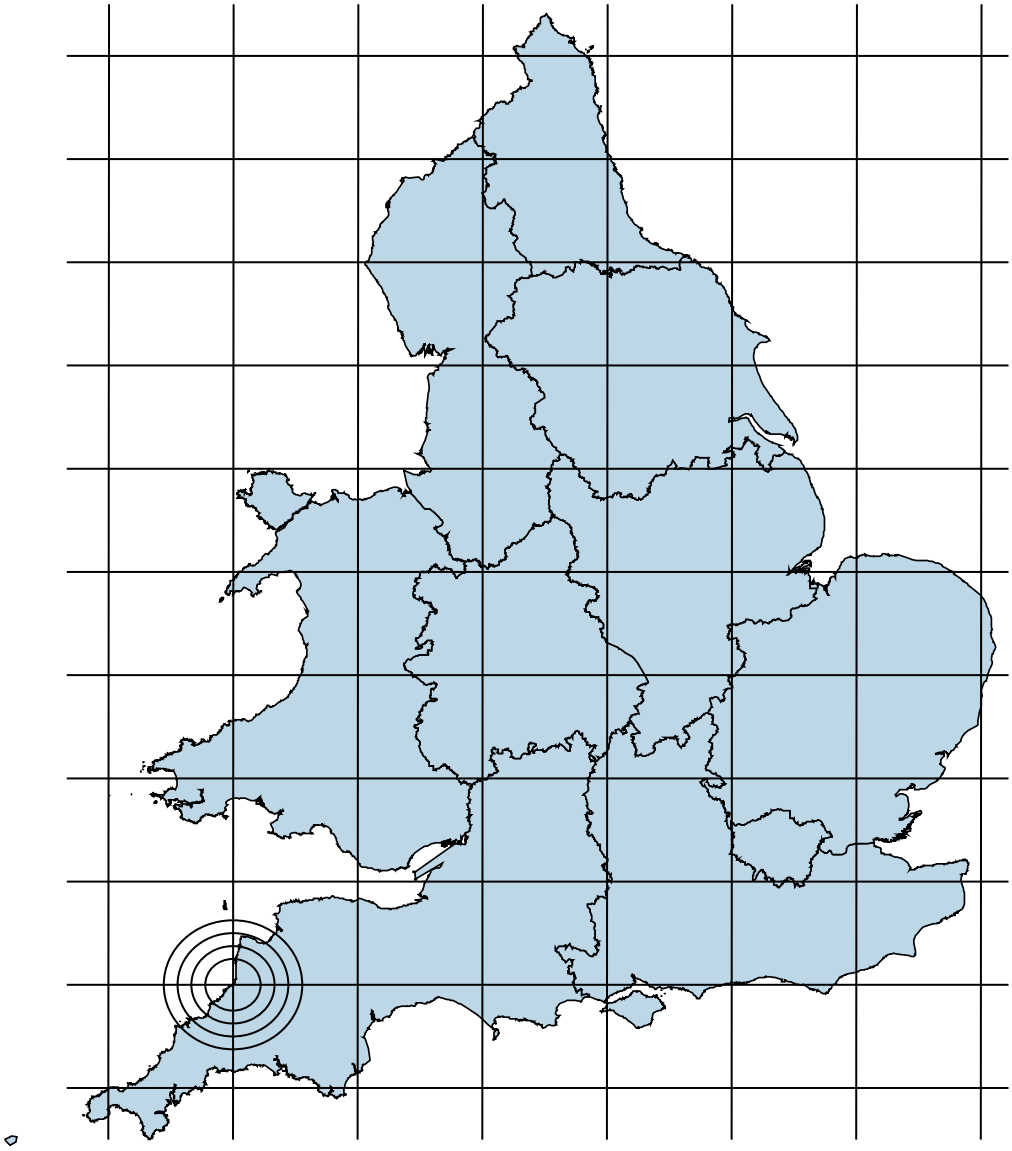

Figure 1.3.2. Graph illustrating the Openshaw *et al* method of spatial cluster detection.

While this method was a new way to search for disease clusters, it received criticism for having no control for multiple testing beyond accounting for the number of circle radii tested (Besag and Newell, 1991). Carrying out sufficient numbers of tests over a large region will result in a high likelihood of finding some clusters in the data purely by chance, but accounting for this in the analysis could then result in the method failing to detect genuine clusters. On a smaller scale, large numbers of circle radii are tested on a high resolution grid in order to determine the most likely position of any detected clusters, leading to each scan being comprised of many thousands of individual tests. However, this is also not accounted for in the calculation of significance levels.

In 1989, Turnbull et al proposed a method similar to the GAM by generating a circle of preselected radius at each administrative centre instead, while also considering techniques used in the distance-based methods (detailed later in this chapter). This method, described as the Cluster Evaluation Permutation Procedure (CEPP), was less computer intensive than the grid-based GAM, but still gave good coverage of the region as areas with more dense population tend to have more administrative zones, and therefore were covered by multiple circles. This method differed from the GAM as it required all circles to contain a fixed population R (say, for example, 5,000 individuals), which was chosen to be larger than the population of any individual administrative zone. The population of each zone (referred to as zone i) was added to a fraction of that of its nearest neighbour (referred to as zone j) in order to reach the fixed population. So if the population of cells i + j was equal to the fixed population then the entire population of j was included, but if the combined population was above the fixed population then only the necessary fraction of zone j was considered. In the event that the population of zones i + j did not contain the required population, the next nearest neighbour was considered and so on. Besag and Newell detailed a similar method in 1991 which used nearest neighbour areas to look for clusters of a fixed number of cases; this method will be detailed in the distance-based methods section. Similar to the Openshaw method, the CEPP used a Monte Carlo simulation to determine whether results were statistically significant by performing repeated analysis on sets of data in which the cases had been randomly allocated from the background population. However, the CEPP only uses the circle which contains the maximum number of cases. This eliminates the problem of multiple testing as only one significance test is being performed, and as a result the method needs no further multiple testing corrections (however, if multiple values of R are used adjustment for multiple testing is required). Therefore, if the observed number of cases in a test exceeds that of 99 replicates in a Monte Carlo simulation, the result has an estimated p value of 0.01 and is significant at the 1% level.

In 1995, Kulldorff and Nagarwalla proposed a circular scan statistic which was capable of detecting clusters of variable size and location on a map without making any prior assumptions regarding either of these variables. This method used a combination of techniques seen in the GAM and CEPP procedures, but with a maximum likelihood ratio test rather than simply the number of cases in each quadrat. Similarly to the GAM and CEPP, a number of circular zones are generated which act as quadrats. These can either be generated on a fixed grid as in the GAM, or at the centroids of census districts (or some other measure of population aggregation) as demonstrated by the CEPP. However, the circular scan statistic differs by generating circles of continuously increasing radius size at each position. This allows the method to simultaneously search for clusters of varying size and position without having to

account for multiple testing, as only the largest result is stored. This is an extension of the method seen in the CEPP.

The model being used takes one circular zone at a time, referred to as z, within which there is a probability p of being a case. Individuals outside the circle have a probability q of being a case. The statistical method is derived from the likelihood function under a binomial model where cases are labelled as ones and non-cases are labelled as zeros, which is:

$$L(z,p,q) = p^{c_z}(1-p)^{n_z-c_z}q^{C-c_z}(1-q)^{(N-n_z)-(C-c_z)}$$

Where N is the total number of individuals in the data, n_z is the number of individuals inside the circle, C is the total number of cases in the data and c_z is the number of cases in zone z.

Therefore, for a fixed zone z in which the ratio of cases to non-cases is greater than that outside the zone, the likelihood ratio can be expressed:

$$L(z) = \left(\frac{c_z}{n_z}\right)^{c_z}\left(\frac{n_z-c_z}{n_z}\right)^{n_z-c_z}\left(\frac{C-c_z}{N-n_z}\right)^{C-c_z}\left(\frac{N-n_z-(C-c_z)}{N-n_z}\right)^{(N-n_z)-(C-c_z)}$$

The most likely cluster is that for which the outcome of the above equation is maximised.

In the same article, the authors used this method to scan for leukaemia clusters in New York, and the results were compared with the GAM and CEPP methods. The GAM failed to detect any evidence of clustering using circle radii of 1, 2 and 4km, while the CEPP detected a statistically significant cluster with the population size fixed at 20,000. With fixed population sizes of 2,500, 5,000 and 10,000 the CEPP failed to detect any clusters. The circular scan statistic, using the likelihood ratio test, detected two statistically significant clusters. Subsequent power evaluations also revealed the circular scan statistic to be equally or more powerful than the CEPP under a range of simulated conditions (the GAM was not included in the comparison). The circular scan statistic has since become a widely used spatial cluster scanning technique due to the method being adaptable, powerful and relatively easy to apply. Professor Kulldorff has since published this method as part of the freely available software package SaTScan, which also includes scan statistics for temporal and space-time data analysis. All methods detailed in this thesis focus on the detection of clusters where observed > expected, i.e. where the risk is abnormally high. As a result, one-sided p values are sought exclusively.

Detecting Irregularly Shaped Clusters

The main problem with many cluster detection techniques is that they have substantially less power for searching for irregularly shaped clusters. Consider the circular scanning method that generates circles at designated points on a map with a defined population (shown in figure 1.3.2), and assume that there is a cholera outbreak along a river floodplain. The resulting cluster would appear on a map to be long and thin, and therefore could be difficult to detect using circles. The three most likely results in this instance are that the method does not identify the cluster at all, the method identifies the entire cluster in a large circle which also incorporates areas of low risk, or the method identifies the cluster as a series of adjacent small circles. As such, there have been several methods developed recently which are capable of detecting clusters of irregular shape.

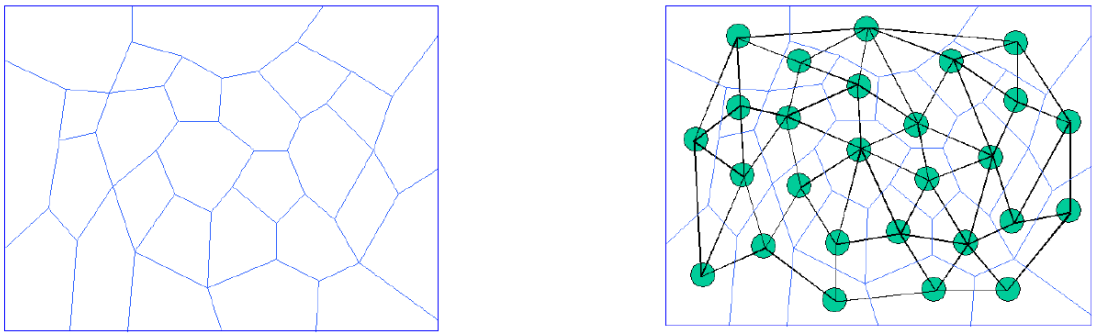

Figure 1.3.3. Cluster scanning using a new graph-based strategy.

Figure 1.3.3 is taken from a report by Duczmal and Assuncao in 2004, in which a method for detecting irregularly shaped clusters is detailed. The diagrams show how a map can be broken down into different areas, which are then connected using a central point (the administrative centre of a borough, or capital of a state for example). Any combination of connected zones can then form a cluster of arbitrary shape. The technique used to determine which connected subset of zones is the most likely cluster is referred to as a simulated annealing method. The name is derived from the annealing process in metallurgy, which is the heating and cooling of a piece of metal in order for the atoms to arrange themselves into their least energetic (and therefore ideal) state. In this context, the technique scans through every combination of regions using the Kulldorff and Nagarwalla likelihood function to generate a scan statistic. Only the largest scan statistic is tested; so when a new combination of areas is scanned, the scan statistic is recorded if it exceeds the previously stored result.

This method was compared with the circular scanning method using homicide data from Belo Horizonte, a city in Southeast Brazil with over 2 million inhabitants. Both methods located areas of high excess risk, with the simulated annealing method detecting a cluster with a

mortality rate of 490% higher than the remaining areas of the city, compared to the circular scanning method which only detected a cluster with 228% excess mortality rate.

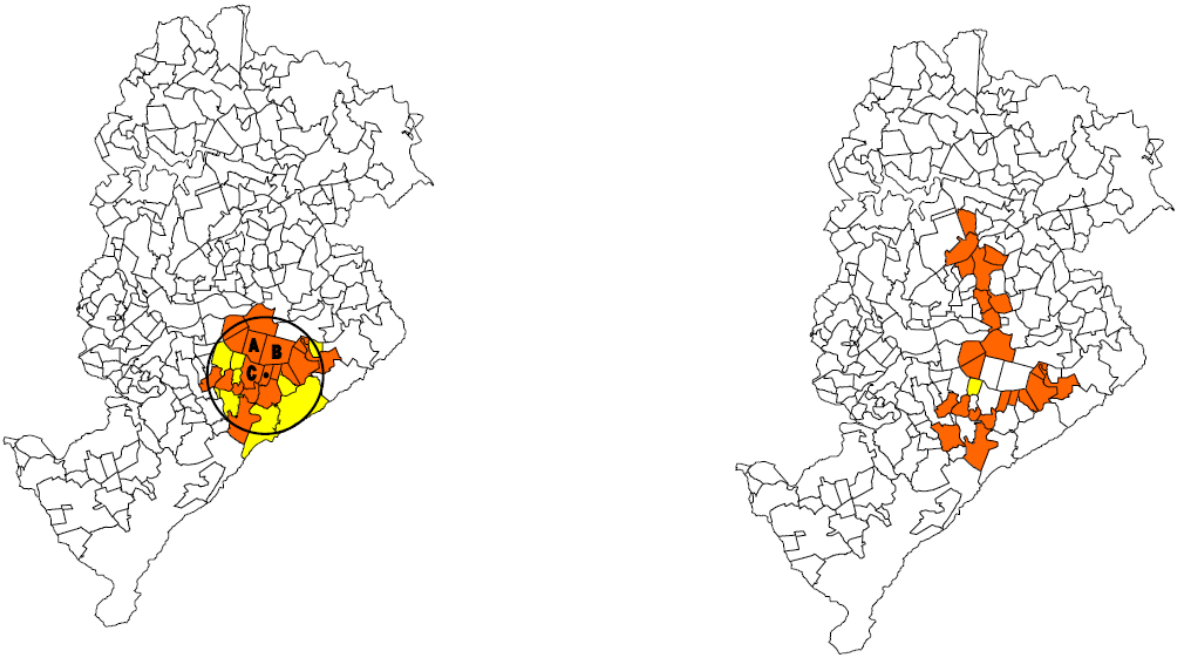

Figure 1.3.4. Differences in homicide clusters using the Kulldorff and Nagarwalla method (left) and the simulated annealing method (right).

The different clusters are illustrated in figure 1.3.4. In both maps, the cluster location is shown by the shaded regions. Yellow regions indicate those which were included in the cluster despite no homicides occurring there during the study period. Three regions included in the circular cluster (labelled A, B and C) had below-average homicide rates. This evidence appears to suggest that the simulated annealing method performs better when searching for clusters of irregular shape, although the validity of the cluster they detected has been questioned given that it included a region in which there were no homicides during the study period. The authors note that the algorithm included this region because it linked two smaller homicide clusters, but the homicide rate in other included regions was at least 10 per 100,000 citizens. The raw homicide figures in each region were not included in the report.

In 2005, Tango and Takahashi proposed a flexibly shaped spatial scan statistic for detecting small clusters of irregular shapes. This method uses a map split into small regions and generates circles at designated points, similarly to the circular scanning method. Every combination of connected regions within each circle is then scanned, allowing the maximum cluster size to be more easily controlled. This gives the advantage of being more flexible than

the circular scanning method, but restricting the size of the cluster gives more believable results than some of those produced by the simulated annealing method. For example, the cluster detected by the simulated annealing method in figure 2 is of a very unusual shape and contains a region with no homicides, and could therefore be overestimating the true size of the cluster. P-values were obtained using Monte Carlo simulation at 999 repeats, and the statistical formula (shown below) was derived from the likelihood ratio test.

$$\sup_{\boldsymbol{Z} \in Z} \left(\frac{n(\boldsymbol{Z})}{\lambda(\mathbf{Z})}\right)^{n(\boldsymbol{Z})} \left(\frac{n(\boldsymbol{Z}^c)}{\lambda(\boldsymbol{Z}^c)}\right)^{n(\boldsymbol{Z}^c)} I\left(\frac{n(\boldsymbol{Z})}{\lambda(\mathbf{Z})} > \frac{n(\boldsymbol{Z}^c)}{\lambda(\boldsymbol{Z}^c)}\right)$$

Where window **Z** is an element of region Z, $\mathbf{Z}^c$ represents the area of region Z outside window **Z**, n() denotes the observed number of cases within the specified area, λ() denotes the expected number of cases in the specified area under the null hypothesis and I() is the indicator function.

The same article also included a comparison between the flexibly shaped scan statistic, circular scanning method and simulated annealing method when scanning for disease clusters in and around Tokyo, Japan.

Figure 1.3.5. Comparison of clusters detected using the circular scanning method (a) and flexibly shaped scan statistic (b) with a maximum cluster size of 15 regions. The same results were obtained with a maximum cluster size of 20 regions.

Figure 1.3.5 shows that the flexibly shaped spatial scan statistic identified the combination of regions 14, 15, 26, 27 and 33 as the most likely cluster location, while the circular scanning method identified regions 14 and 15 as the strongest cluster.

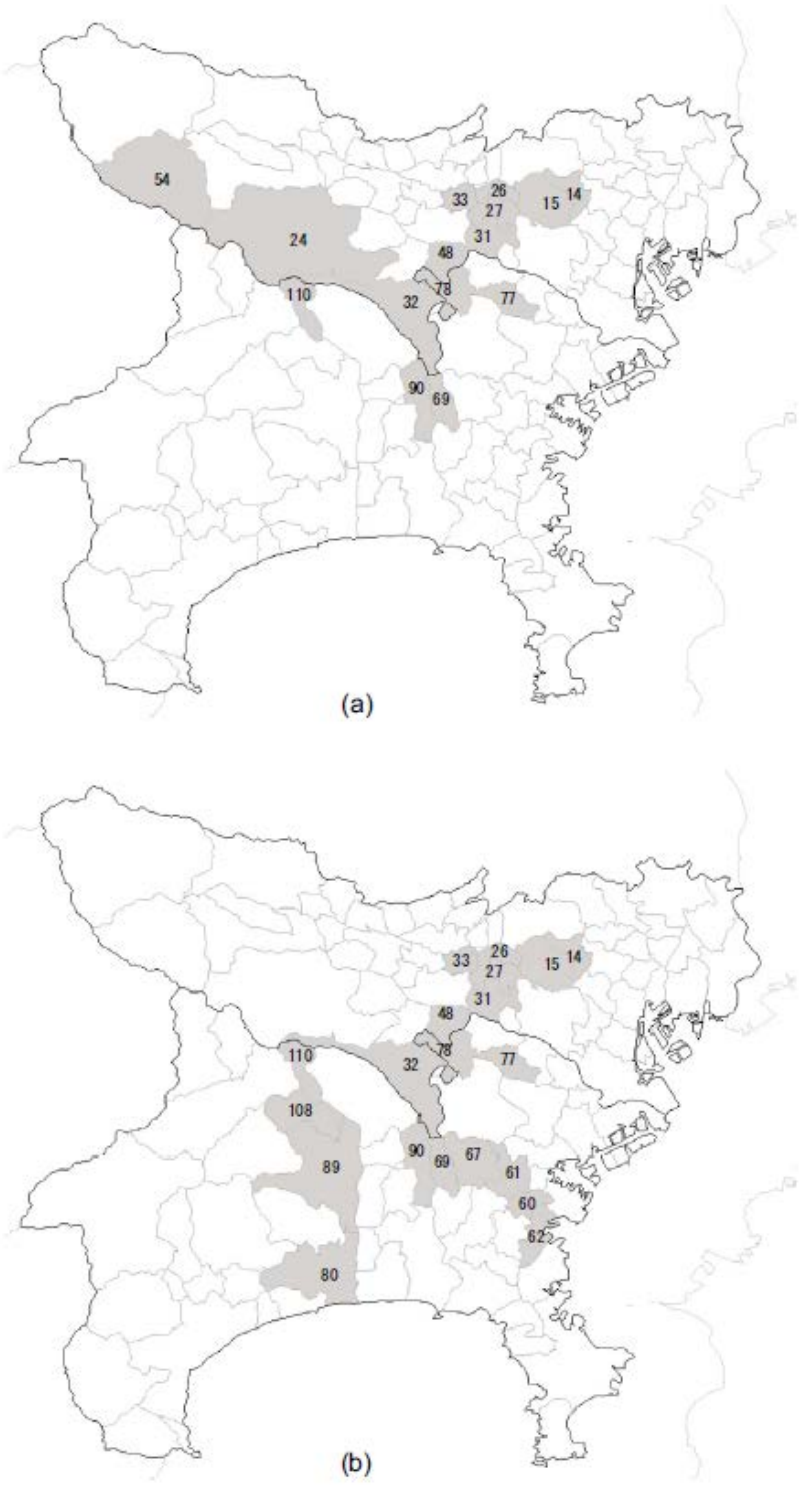

Figure 1.3.6. Comparison of clusters detected using the simulated annealing method with a maximum cluster size of (a) 15 regions and (b) 20 regions.

Figure 1.3.6 shows that the simulated annealing method has identified very large, irregularly shaped clusters when using maximum cluster sizes of 15 or 20.

Table 1.3.1. Most likely cluster locations using the three scanning methods with a maximum cluster size of 15 regions. “Circular” refers to the Nagarwalla and Kulldorff scan statistic, “Flexible” to the Tango and Takahashi flexible scan statistic and “Duczmal et al” to the simulated annealing method.

region no.	population	observed no. cases	expected no. cases	relative risk estimated (true)	Log likelihood ratio (LLR) and estimated relative risk $\hat{\theta}$ for the most likely cluster		
					Circular	Flexible	Duczmal et al.
14	319,687	14	3.794	3.69 (3.0)	*	*	*
15	529,485	21	6.283	3.34 (3.0)	*	*	*
					LLR = 20.1		
					$\hat{\theta}$ = 3.47		
26	139,077	6	1.650	3.64 (3.0)		*	*
27	165,564	6	1.964	3.05 (3.0)		*	*
33	105,899	4	1.257	3.18 (1.0)		*	*
						LLR = 29.7	
						$\hat{\theta}$ = 3.41	
24	466,347	8	5.534	1.44 (1.0)			*
31	197,677	3	2.346	1.27 (1.0)			*
32	349,050	5	4.142	1.20 (1.0)			*
48	58,635	1	0.696	1.43 (1.0)			*
54	3,808	1	0.045	22.12(1.0)			*
69	119,575	3	1.419	2.11 (1.0)			*
77	177,742	5	2.109	2.37 (1.0)			*
78	125,127	2	1.485	1.34 (1.0)			*
90	194,866	5	2.312	2.16 (1.0)			*
110	21,535	1	0.256	3.91 (1.0)			*
							LLR = 31.8
							$\hat{\theta}$ = 2.41

Table 1.3.1 shows that the cluster identified using the circular scanning method had the lowest log likelihood ratio LLR but highest relative risk of the three, due to the inclusion of only two regions each with very high relative risks. The flexible scanning method identified a larger cluster with a very small drop in relative risk but a large increase in LLR. The cluster identified using the simulated annealing method had the highest LLR of the three but the lowest relative risk, due to the inclusion of more regions with lower relative risks. Nevertheless, the authors suggested that future methods could impose a penalty based on the complexity of cluster shaped, as clusters such as either of those in figure 1.3.5 are likely to be overestimates of the true cluster size. In 2008, Takahashi et al proposed an extension to this method which could be used to scan for space-time clusters.

A method was detailed in 2004 by Patil and Taillie which used an upper level set scan statistic to detect irregularly shaped spatial clusters. The upper level set method considers only the regions with above average prevalence, ensuring that low risk areas are not incorporated in the potential cluster region. Imagine an overall region Ω, divided into many cells (denoted by *a*). Taking Y_a as the case count for each cell and A_a as the population size of each cell, candidate cells for possible clusters are those with the largest values for the rate:

$$G_a = \frac{Y_a}{A_a}$$

Using these cell rates, region Ω can be reduced to Ω_{ULS}, including only those regions with large values of G_a. This is referred to as taking an upper level set of Ω, and different upper level sets can be taken based on different threshold values for G_a. The zones in Ω_{ULS} can be considered plausible hotspots as they include only those cells with the highest prevalence.

Figure 1.3.7 shows how the upper level set results can change based on the threshold value of G_a. At the threshold value of g, there are three components highlighted; Z_1, Z_2 and Z_3. When the threshold is reduced to g`, zones Z_4, Z_5 and Z_6 emerge. The graph shows how zones Z_1 and Z_2 join to become zone Z_4, zone Z_3 grows and becomes zone Z_5 and zone Z_6 emerges at the lower threshold point. Figure 1.3.8 then shows how this method can be applied to a region which has been aggregated into population groups.

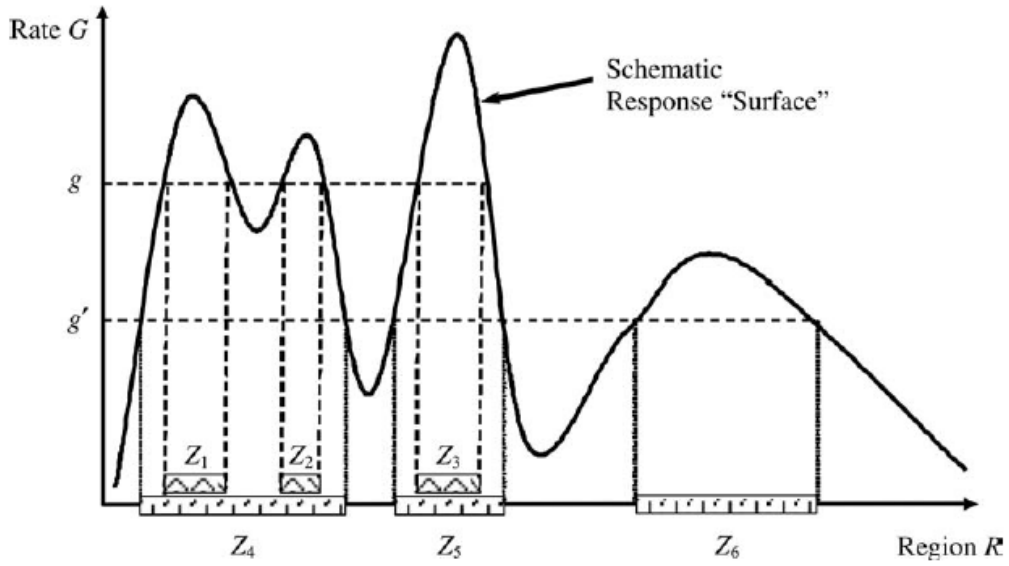

Figure 1.3.7. Graph demonstrating the upper level set method theory.

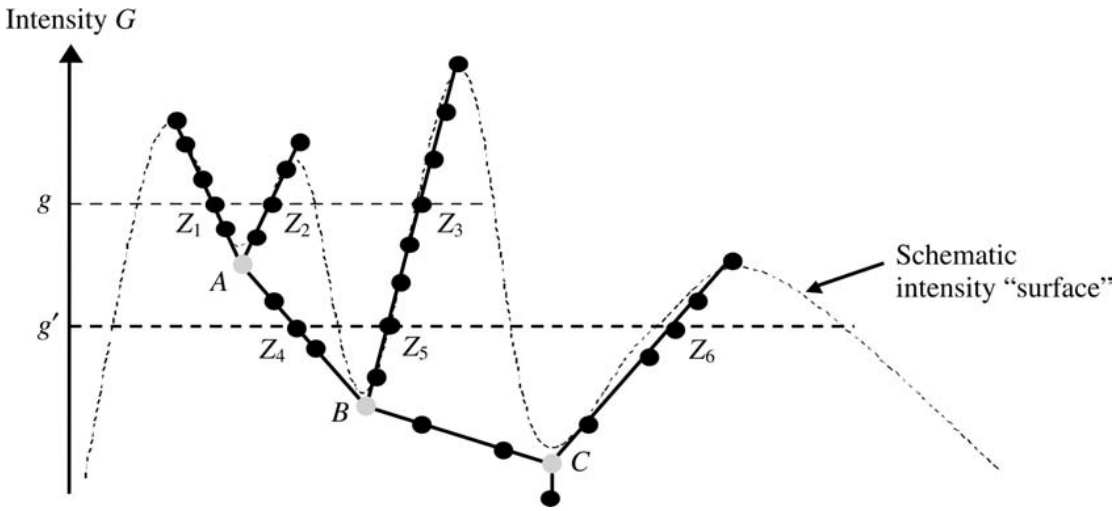

Figure 1.3.7. Graph demonstrating how the upper level set method could be used to search for clusters. Each node on the graph represents a region in space, and any combination of connected regions which exceeds the upper level can be considered a cluster.

Distance Based Methods

Rather than measuring the case :noncase ratio using overlapping quadrats, distance based methods examine the distribution of cases in the study area. This can involve comparing the position of cases either with other cases or with controls selected from the population at risk. The following two examples illustrate these two different approaches.

In 1987, Whittemore *et al* proposed a distance based method which examined the mean distance between all pairs of cases in the study region. The formula for this method is shown below, where $\binom{n}{2}$ is the total number of pairs of cases and $\Delta(i_1, i_2)$ denotes the distance between the i_1th and i_2th case.

$$\delta = \binom{n}{2}^{-1} \sum_{i_1 < i_2} \Delta(i_1, i_2)$$

This statistic is then compared with the expected value and variance of δ under the null hypothesis that all cases are independent Poisson variables and that there are no clusters in the data. The authors also proposed that, in the event of a disease having varying prevalence in different subgroups of the population, the population could be stratified and the disease be treated as an independent Poisson process in each stratum. In the case of trisomies 13 and 18, cases could be partitioned into subgroups based on maternal age.

In 1990, Cuzick and Edwards detailed a method which can detect clusters in non-uniformly distributed populations without prior knowledge of population density in terms of age and gender. Cases are compared with a group of controls selected from the risk population, and the statistics are based on whether the nearest neighbour to each case is a case or control. Information regarding the population at risk can be obtained from a variety of sources, but must be representative of the case population in terms of both age and sex. For example, if examining a condition which affects children, the population at risk could be obtained from local school records.

This method is a non-homogenous Poisson process where the expected disease density is a function of the age and sex-adjusted population density. Assume that n_0 cases are observed in some predefined region Ω, and that their positions are denoted (x_1,…,x_{n0}). From all individuals at risk in region Ω, a set of n_1 controls are selected whose positions are denoted (y_1,…,y_{n1}). This information is then combined such that the cases and controls are denoted (z_1,…,z_n, $n=n_0+n_1$), where z_i contains no information regarding whether the individual is a case or control. For i=1,…,n, δ_i is defined as 1 if z_i is a case and 0 if z_i is a control. For each i, d_i is defined as 1 if 1 if the nearest neighbour is a case and 0 if the nearest neighbour is a control.

The nearest neighbour test is defined as:

$$T = \sum_{i=1}^{n} \delta_i d_i$$

This test can be extended to count the number of cases among a preset number of nearest neighbours for each case, giving additional power when there are a few large clusters rather than many smaller ones.

Section 1.3.4: Aim Of This Thesis

Using the methods detailed in section 1.3.3.1, it is unlikely that clusters arising due to any risk factors besides maternal age could be detected due to the strength of the maternal age effect. It is possible to statistically adjust for maternal age which would enable the discovery of alternate clusters, this is further explored in chapter 5 of this thesis. Methods detailed in section 1.3.3.2 demonstrate progress which has been made in the detection of irregularly shaped (i.e. non circular) clusters, however these methods are also likely to require some form of adjustment to be made before non-age related clusters can be detected. During the study period, the standardised mean maternal age after accounting for changes in population structure increased from 29.0 in 2004 to 29.5 in 2010 (Office for National Statistics Characteristics of Mother 1, England and Wales, 2011).

There are two distinct objectives which will be addressed in this thesis. The primary methodological aim of this work is to develop a method which is capable of accounting for pre-existing risk factors (in this case, maternal age and gestational age at diagnosis) when searching for clusters. Such a method would be able to take advantage of the detailed information available on each case from the NDSCR, in order to accurately adjust for both risk factors for each case individually. The second objective is to apply this method to NDSCR data in order to attempt to discover any clusters which may arise due to alternate risk factors.

Specific aims of this thesis are:

1. To scan trisomy data in England and Wales from the NDSCR for temporal and spatial clusters.
2. To develop a statistical method capable of accounting for two risk factors simultaneously, and to demonstrate the effectiveness of this method using simulated data.
3. To scan the same data for clusters after adjusting for maternal age and gestational age at diagnosis using the novel method.
4. To assess the impact of the findings in the context of both previously undertaken and future congenital anomaly research, and NHS policy.

Chapter 2: Data Cleaning

Data cleaning is the process of improving the quality of a dataset by detecting and correcting or removing errors. Such errors may include inconsistancies within the data, typing errors during data entry, duplicate entries and incomplete records. When examining a dataset for clusters, especially when analysing conditions as rare as trisomies 18 and 13, a single case can be the difference between a cluster being classed as significant or not, so it is essential that data is thoroughly cleaned before it is scanned. All NDSCR data is routinely checked and cleaned to ensure that it is the highest possible quality and that all available information is recorded, however there are still cases in the dataset which are missing vital information for one or more of the cluster scanning techniques. For some cases vital missing information can be estimated from other information provided in order to include these cases in the relevant scan(s).

Data was scanned for temporal and spatial clusters. Temporal scans were performed within Government Office Region using sample date and then using estimated conception date. Spatial scans were initially performed using the mother's Primary Care Trust (PCT) at diagnosis and then subsequently adjusting for the mother's age and for the gestational age at diagnosis . For each of these procedures, the required information and number of cases lost due to missing data is provided as an overview in table 2.1, and then explained in more detail in the rest of the chapter. Temporal and spatial data cleaning processes are detailed separately as each requires different case information.

Table 2.1. Overview of Data Cleaning Processes.

		Trisomy 18		Trisomy 13	
Scan		**Number**	**%**	**Number**	**%**
Temporal by Sample Date	**Total Cases**	**3507**	**100**	**1473**	**100**
	Excluded - No GOR	6	0.2	5	0.3
	Excluded - No Sample Date	70	2.0	36	2.4
	Total Cases Excluded	76	2.2	41	2.8
	Total Cases Remaining	**3431**	**97.8**	**1432**	**97.2**
Temporal by Conception Date	**Total Cases**	**3507**	**100**	**1473**	**100**
	Excluded - No GOR	6	0.2	5	0.3
	Excluded - No Conception Date	175	5	103	7
	Excluded - Out of Timeframe[1]	107	3.1	46	3.1
	Total Cases Excluded	288	8.2	154	10.5
	Total Cases Remaining	**3219**	**91.8**	**1319**	**89.5**
Spatial without Adjustment	**Total Cases**	**3507**	**100**	**1473**	**100**
	Excluded - No PCT	186	5.3	91	6.2
	Total Cases Excluded	186	5.3	91	6.2
	Total Cases Remaining	**3321**	**94.7**	**1382**	**93.8**
Spatial with Adjustment	**Total Cases**	**3507**	**100**	**1473**	**100**
	Excluded - No PCT	186	5.3	91	6.2
	Excluded - No Maternal Age	91	2.6	61	4.1
	Excluded - No Gestational Age	105	3	67	4.5
	Excluded - Fetal Loss Under 20 Wks	488	13.9	213	14.5
	Total Cases Excluded	870	24.8	432	29.3
	Total Cases Remaining	**2637**	**75.2**	**1041**	**70.7**

[1]Out of timeframe refers to periods of time at the start and end of the timeframe during which the NDSCR would not have complete data. There is a full explanation later in the chapter on page 73.

Section 2.1.Preparation of Data Before Applying the Temporal Cluster Detection Method

Government Office Region

The first essential piece of information for the temporal cluster scanning method is government office region, as each region is analysed individually. Government Office Region is determined from the mother's postcode at diagnosis or the mother's home town (free text). In total, 14 cases of trisomy 18 (0.4%) and 11 cases of trisomy 13 (0.4%) were missing both fields of information and so could not be assigned a government office region.

Each case is diagnosed and subsequently reported to the NDSCR by one of 21 cytogenetic laboratories. All cases therefore have an assigned cytogenetic lab, which can be used to impute government office region if the laboratory has a sufficiently large proportion of cases originating in the same region. The threshold for this proportion was set at 90%, so if a lab takes over 90% of its cases from a one region, it would be reasonable to assume that any case from that lab with missing GOR comes from the same region. There is a small chance of error in doing this (up to 10% depending on the lab), but it is preferred to losing the case altogether. This threshold may appear overly conservative, but as the scans are performed separately by GOR and a single case can make the difference between a cluster being significant or not, high accuracy is essential when imputing GOR. Using this information 7 / 14 trisomy 18 cases and 5 / 11 trisomy 13 cases could be assigned to a government office region (specific case details are shown in the appendix), leaving 7 and 6 cases respectively still unassigned. The number and proportion of diagnoses from each government office region by cytogenetic lab are shown in the appendix tables 2.9 and 2.10 respectively. Tables 2.11 and 2.12, also in the appendix, list all cases which were assigned to a government office region on the basis of cytogenetic lab information.

Of the remaining cases, one trisomy 18 case was diagnosed at lab 30, but had the referring hospital on the form. Since 31 of the 35 cases in the database from this same hospital were from government office region H, this case was assignedto region H. The 6 remaining trisomy 18 cases (0.2% of the total) and 6 trisomy 13 cases (0.3% of the total) could not be assigned a government office region, and were excluded.

Date of Sample

Initially, temporal scans are performed using the date of sample as the time point for each case. Date of sample refers to the date when the tissue sample was taken for cytogenetic testing and diagnosis. The samples ranged in age from a gestation of 10 weeks to 40 weeks. 70 cases of trisomy 18 and 36 cases of trisomy 13 did not have complete sample date information recorded and had to be excluded from this analysis, which was a loss of 2.0% of trisomy 18 cases and 2.4% of trisomy 13 cases. All of these cases were missing both month and day of sample except for 2 cases of trisomy 18 and 1 cases of trisomy 13, which were only missing day of sample. There was no way to accurately infer this information from other variables in the data.

Date of Conception

After these scans had been performed, a further series of scans were run using the date of conception as the time point for each case. Conception date is not reported to the NDSCR, but can be estimated by subtracting either gestational age at sample from date of sample or gestational age at outcome from outcome date. Both methods are equally accurate as outcome and sample dates are recorded accurately by day. Sample information was used initially as this was more likely to be complete. If the sample information was incomplete the outcome information was used. Gestational age at sample and outcome is reported to the NDSCR either in complete weeks only or in weeks and days, however weeks and days are only available for only a small proportion of cases. For this reason, whenever gestational age information is used throughout this thesis only complete weeks are used. The percentage of cases included in each scan by date of conception is shown in table 2.5.

Table 2.2. Initial estimates of conception dates using either complete sample or outcome information.

	Trisomy 18		Trisomy 13	
Calculated Using	**Number**	**%**	**Number**	**%**
Sample Date – Gestational age at Sample	3124	89.1	1258	85.4
Outcome Date – Gestational age at Outcome	193	5.5	104	7.1
Total Calculated	3317	94.6	1362	92.5
Could not Calculate Conception Date	184	5.2	106	7.2
Previously Excluded (Missing GOR)	6	0.2	5	0.3
Total Missing	190	5.4	111	7.5

Table 2.2 shows that 3,124 cases of trisomy 18 (89.1%) and 1,258 cases of trisomy 13 (85.4%) had complete sample date and gestational age at sample information, and therefore their conception date could be estimated using these variables. Around half of the remaining cases had complete information regarding outcome date and gestational age, and so could have their conception date estimated from gestation at outcome and outcome date. Date of conception could not be estimated for 184 cases of trisomy 18 (5.3% of total) and 106 cases of trisomy 13 (7.2% of total). These are detailed in table 2.3.

Table 2.3.Breakdown of missing information showing proportions of complete data.

	Trisomy 18		Trisomy 13	
Information Present	**Number**	**%**	**Number**	**%**
Sample date, no gestational age at sample	146	4.2	84	5.7
Gestational age at sample, no sample date	18	0.5	11	0.7
Neither sample date nor gestational age	20	0.6	11	0.7
Total	184	5.2	106	7.2
Outcome date, no gestational age at outcome	85	2.4	44	3
Gestational age at outcome, no outcome date	15	0.4	12	0.8
Neither outcome date nor gestational age	84	2.4	50	3.4
Total	184	5.2	106	7.2

The largest proportion of cases in table 2.l3 are those with a recorded sample date but no gestational age at sample, and it was thought that the type of cytogenetic sample used could provide an estimate of gestation at sample as different tests are performed depending on the gestational age of the pregnancy; CVS, generally performed early during the pregnancy, and amniocentesis, usually performed later in the pregnancy. Table 2.4 shows the number and proportion of cases which were confirmed using either CVS or amniocentesis.

Table 2.4. Number and proportion of total cases missing conception date which were diagnosed using a CVS or amniocentesis sample.

	Trisomy 18		Trisomy 13	
Tissue Sample	**Number**	**%**	**Number**	**%**
CVS	37	1.1	19	1.3
Amniocentesis	27	0.8	10	0.7

Table 2.4 shows that around 1.9% of total cases for trisomy 18 and 2.0% for trisomy 13 could be included in the analysis if their gestation at sample could be accurately estimated by tissue

sample type. All NDSCR cases between 2004 and 2010 (including trisomy 21) which were diagnosed after either a CVS or amniocentesis were analysed for gestational age at sample. The same analysis was also performed using each trisomy individually to ensure that the distributions did not significantly differ, which they did not. However, after examining the distribution of gestational ages in all CVS and amniocentesis diagnoses, there was insufficient evidence to be able to accurately impute gestation at sample using tissue type. The distributions of gestational ages for cases diagnosed using either CVS or amniocentesis are included in the appendix in tables 2.13 and 2.14.

For cases with gestation at sample but no sample date, it could be possible to obtain the sample date using the unique identifier assigned to the sample by the cytogenetic laboratory. In most labs, cases from each calendar year are divided by tissue type and assigned a reference number in the order in which they are processed. For example, the first CVS sample to arrive in 2006 would be given the laboratory ID C060001, the second would be C060002 etc. Amniocentesis samples would be similarly labelled, with the first sample of 2006 labelled A060001. For each case missing date of sample, the dates of the previous and following cases can be used to estimate sample date.

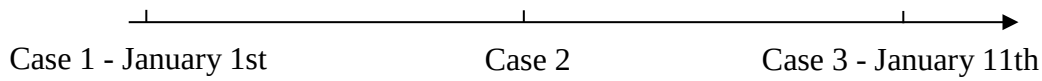

Figure 2.1. Imputing sample dates based on the case sample dates immediately before and after the case being examined.

For example, figure 1 shows a hypothetical trio of cases at the start of a calendar year. Assuming that case 1 occurs on January the 1st and case 3 occurs on January the 11th, case 2 can be estimated to have occurred in the middle of this time period on January the 6th. In order to maintain uniformity across all sample dates estimated in this way, the assigned sample date was always the middle of the timeframe. Where the timeframe spanned an even number of days and the exact middle fell across 2 days, the sample date was assigned to the first of these days. For example, if case 1 occurred on January the 1st and case 3 occurred on January the 10th, case 2 would be assigned the sample date January the 5th rather than the 6th. As sample date accuracy is so important to this analysis, it was decided that the sample date estimate could be no less precise than the process of estimating conception dates. As gestational ages are processed in whole weeks only, there is a maximum potential error of 6 days when subtracting the gestational age at sample from the sample date. Therefore, the

maximum error in estimating sample dates using adjacent cases is 6 days. This means that the difference between the previous and following case sample dates should be no more than 12 days. Using this cutoff, 3 cases of trisomy 13 and 9 cases of trisomy 18 were assigned sample dates and could be included in the analysis. Full details regarding all cases which had sample date estimated using this method are included in the appendix in tables 2.15 and 2.16.

Timeframe

It must also be considered that using estimated date of conception changes the temporal study period from 2004 – 2010 to 2003 – 2010. 126 cases of trisomy 18 and 47 cases of trisomy 13 were conceived in 2003, as were other cases which were diagnosed in 2003 and therefore not included in this study. This means that the data we have on cases conceived in 2003 is incomplete. It would be possible to exclude all cases which were conceived in 2003 from the analysis, but this would incur a further loss of 3.8% of trisomy 18 cases and 3.5% of trisomy 13 cases. Instead, we can use a cutoff time in late 2003 for which we are confident that the majority of cases would not be detected prenatally until 2004. Prenatal screening starts at around 10 weeks gestation, and it is unlikely that any trisomy 18 or trisomy 13 case would be diagnosed before 10 weeks. Cases diagnosed in the final 10 weeks of 2003 can therefore be included. This allowed 71 cases of trisomy 18 and 25 cases of trisomy 13 to be included despite being conceived in 2003. Overall, 55 trisomy 18 cases (1.6% of total) and 22 trisomy 13 cases (1.4% of total) were conceived before the final 10 weeks of 2003 and were excluded from the study.

Similarly, there were cases conceived in the latter part of 2010 which would not be expected to be detected until 2011. While this could potentially occur any time in the last 9 months of 2010, excluding all cases from such a large timeframe could have a significantly detrimental effect on the power of the analysis. Therefore, the data was analysed by gestation at sample for each trisomy, and it was found that around 85% of cases for both trisomies had been diagnosed by the gestational age of 21 weeks. Excluding only the last 21 weeks of 2010 includes as many cases as possible without compromising the overall quality of the dataset. A further 52 trisomy 18 cases and 24 trisomy 13 cases were conceived in the final 21 weeks of 2010 and subsequently excluded. In total, 8.2% of trisomy 18 cases and 10.5% of trisomy 13 cases were excluded from the temporal scans by date of conception.

Sensitivity Analysis

For the temporal scans by conception date, a large proportion of cases were excluded due to the larger amount of information required for the scans, and the desire to only include cases for which those variables were either known or could be imputed with very high accuracy. Cases whose conception date fell outside the temporal window for which the NDSCR had complete data available were also excluded.

In order to assess the effect of these exclusions, a second analysis was undertaken which relaxed the imputation accuracy requirements. In these scans, date of conception was imputed for cases which had no gestational age at sample recorded by assigning the mean gestational age of the tissue sample used in their diagnosis. For example, a case diagnosed by CVS would be assigned the gestational age at sample of 12 weeks, whereas a case diagnosed by amniocentesis would be assigned a gestational age of 19 weeks. This was undertaken for 9 different tissue types (CVS, amniotic fluid, fetal blood, fetal skin, placenta, product of conception, villi, umbilical cord and postnatal), and the resulting gestational ages at sample were subtracted from the date of sample (where available) to provide an estimate of the conception date. There were also no timeframe restrictions applied, so all cases reported to the NDSCR between 2004 and 2010 were included in the scans even if they were diagnosed early in 2003 or late in 2010.

Using this approach, the proportion of cases excluded from the scans by date of conception could be significantly reduced in every government office region (table 2.5). For trisomy 18, the percentage of cases included increased from 91.8% to 98.4%, while the trisomy 13 the percentage rose from 89.5% to 98.0%. For trisomy 18 all regions had less than 3% of cases excluded, while for trisomy 13 all had less than 3% excluded except the East of England where 3.1% were removed. The percentage of cases excluded was consistently low across all regions in the sensitivity analysis. The results from this analysis are shown and discussed in chapter 3.

Table 2.5. Percentage of cases excluded in scans by conception date.

	Conception Date Scans		Sensitivity Analysis	
Government Office Region	**T18**	**T13**	**T18**	**T13**
North East	4.3	7.5	1.2	0
North West	12.5	15	2	2.3
Yorkshire and the Humber	12.1	10.3	2.2	2.6
East Midlands	8.9	10.6	1.4	0
West Midlands	8	11.6	2.9	2.5
East of England	5.7	8.1	0.5	3.1
London	9.9	11.7	1.3	2.1
South East	5.2	7.7	0.6	0.4
South West	4.2	8.8	0	2
Wales	9.2	9.4	1.8	0

Section 2.2.Preparation of Data Before Applying the Spatial Cluster Detection Methods

Sample Year

Cases are divided by year of sample so as to avoid excess data loss at the start and end of the study period which would be incurred if date of conception was being used (as demonstrated in the temporal scanning method). Years can be analysed individually or all together to give a single scan for the entire dataset. All cases in the NDSCR database have a recorded sample year, so no data is excluded on this basis.

PCT

Part of the NDSCR data checking and cleaning procedure is to assign PCTs to cases based on either postcode or maternal home town. After this process had been performed, 301 trisomy 18 cases (8.6% of the total) and 134 trisomy 13 cases (9.1% of the total) were missing PCT information. Table 2.6 shows what location information was present for cases which were missing pct.

Table 2.6. Information present regarding the location of cases with missing PCT.

	Trisomy 18		Trisomy 13	
Information Present	**Number**	**%**	**Number**	**%**
Maternal Postcode Only	6	0.2	3	0.2
Maternal Home Town Only	83	2.4	37	2.5
Postcode and Home Town	121	3.5	48	3.3
Neither Postcode Nor Home Town	91	2.6	46	3.1
Total	301	8.6	134	9.1

As the NDSCR procedure assigns PCTs using an automated process, it may be that any incorrectly entered or incomplete PCTs could be assigned with manual checks. An online PCT finder widget was used to attempt to assign cases to PCTs. The URL for this widget is provided here and in the references.

https://www.ndtms.org.uk/emids/cgi-bin/ons_locale.cgi

This method was implemented in three stages. Firstly, the full postcode was entered into the widget. For postcodes which were incorrect, the area code (first part of postcode plus the first character of the second part) was entered. For area codes which were incorrect, the postal

district (first part of postcode) was entered. Postal districts and area codes can span multiple PCTs (postcodes cannot), and it is important to ensure that cases are not incorrectly assigned to PCTs as a single case can make a big difference in the spatial scans. In the event of an area code spanning multiple PCTs, the widget returns the percentage probability of the address being located in each PCT. PCTs were only assigned when the probability of the PCT being incorrect was less than 10%, as this ensured a low risk of including incorrect data while including as many cases as possible. Using this approach, a further 90 trisomy 18 cases and 39 trisomy 13 cases were assigned PCTs. Maternal home towns could also be checked in the same way, as some towns fall entirely within a single PCT and can therefore be used to accurately obtain PCT information. For each maternal home town, all potential postal districts were run through the same PCT finder tool using the same maximum 10% error rate. A further 25 trisomy 18 cases and 4 trisomy 13 cases were assigned to PCTs using maternal home town information. Full tables showing which cases were processed using these methods are included in the appendix.

Table 2.7 shows the number and proportion of total cases which were assigned to PCTs based on either their postcode or maternal home town. A total of 5.3% of trisomy 18 cases and 6.2% of trisomy 13 cases could not be assigned to PCTs and were excluded from all spatial analyses.

Table 2.7. Number and percentage of total cases for whom PCT was imputed using either postcode or maternal home town information.

	Trisomy 18		Trisomy 13	
Information Used	**Number**	**%**	**Number**	**%**
Postcode	22	0.6	7	0.5
Area Code	32	0.9	14	1
Postal District	36	1	18	1.2
Maternal Home Town	25	0.7	4	0.3
Total	115	3.3	43	2.9
Total Still Missing After Cleaning	186	5.3	91	6.2

Two cases in the NDSCR database were from the PCT “5QV” (Hertfordshire), which was formed in 2008 by merging two existing PCTs (East and North Hertfordshire, and West Hertfordshire). The birth figures from the ONS were obtained before this merge happened, so these cases were manually reassigned to the correct PCT.

Maternal Age at Sample

Spatial scans were run adjusting for maternal age at diagnosis and therefore for some analyses 91 trisomy 18 cases and 61 trisomy 13 cases with missing maternal age at sample were excluded from the adjusted analysis. There was no way to impute this information from other information on the form. Figures 2.2 and 2.3 shows the distribution of maternal age for both trisomies during the study period.

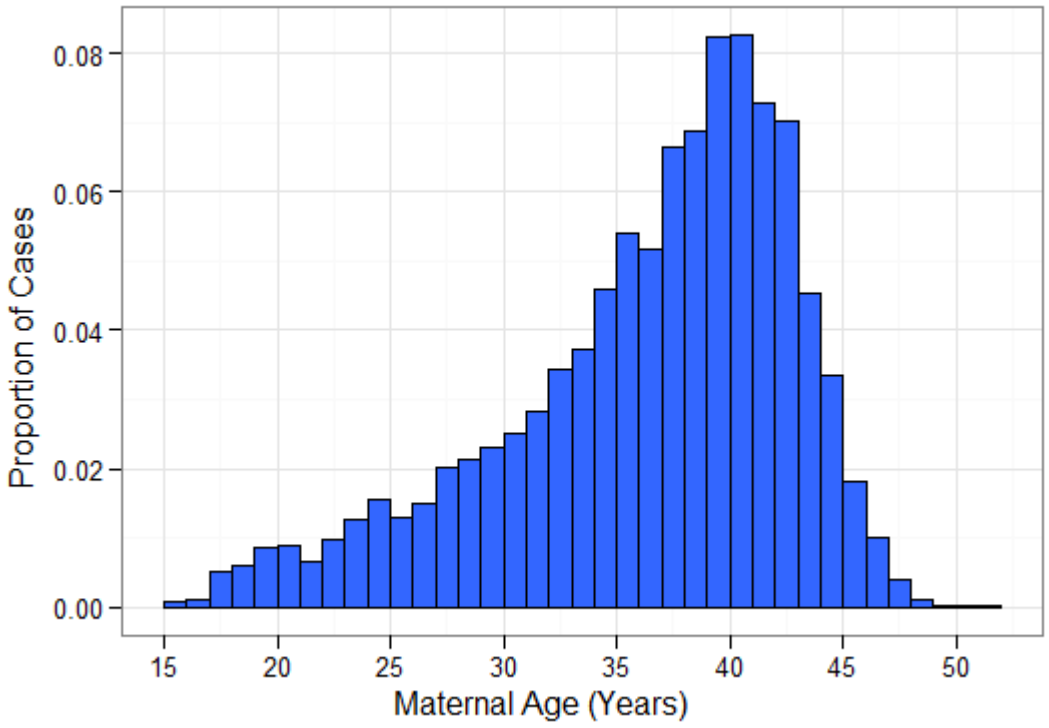

Figure 2.2. Graph of maternal age for trisomy 18 cases reported to the NDSCR during the study period.

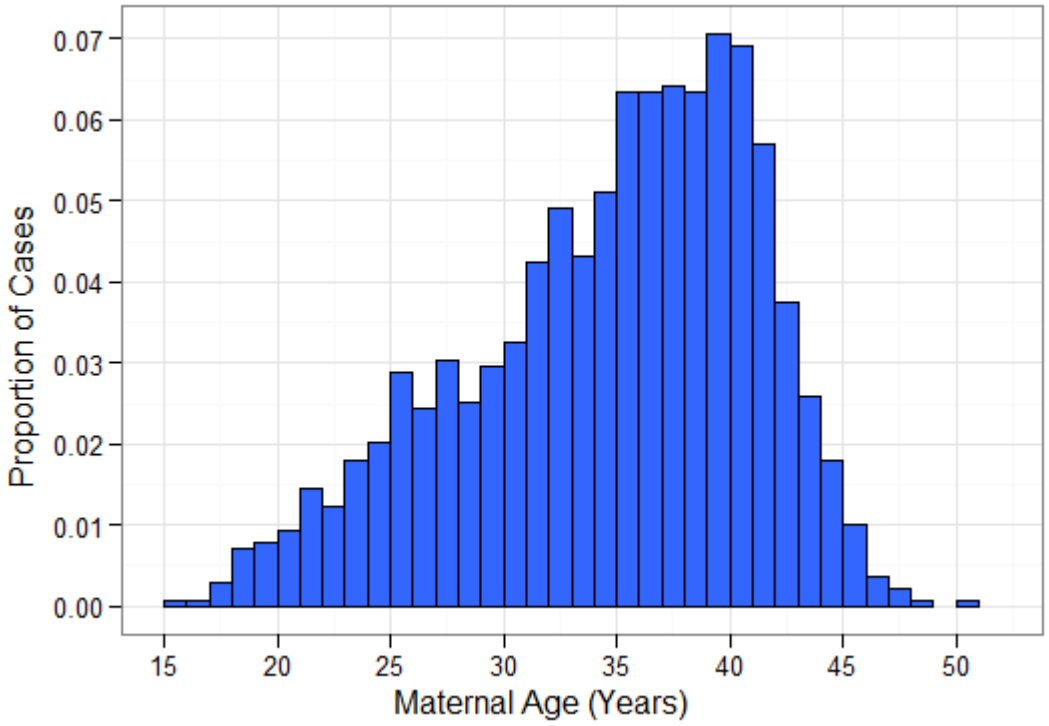

Figure 2.3. Graph of maternal age for trisomy 13 cases reported to the NDSCR during the study period.

Gestational Age at Sample

Spatial scans were run adjusting for gestational age at diagnosis in cases that were subsequently terminated (see chapter 5 for full explanation of weighting methods). Fetal loss rates are very high in trisomy 18 and trisomy 13 pregnancies. Areas in which routine prenatal screening occurs from 10 weeks gestation will diagnose more cases than those which screen later. It is assumed that all pregnancies surviving to 20 weeks are diagnosed. Therefore gestational age weights are only applied to terminated cases under 20 weeks of gestation at sample. All cases with a gestational age greater than 20 weeks at diagnosis are given a weight of 1. All fetal losses under 20 weeks were excluded from the analysis, resulting in a loss of 488 cases of trisomy 18 and 213 cases of trisomy 13. This means that all cases diagnosed at or after 20 weeks will receive a default weighting of 1. The weighting method applies variable weights to cases diagnosed between 14 and 20 weeks, with all cases diagnosed before 14 weeks of age being assigned the same weight. Therefore if the majority of cases (a minimum of 80%) diagnosed by a certain tissue type are diagnosed either after 20 or before 14 weeks of age, the equivalent weight can be assigned to cases which have known tissue type but are missing gestational age at sample. The 80% limit was chosen as it is important to minimise the amount of cases which are excluded from the scans while still maintaining a high level of accuracy in

the imputed data. The distribution of gestational age at diagnosis among cases of trisomies 18 and 13 during the study period is showed in figures 2.4 and 2.5.

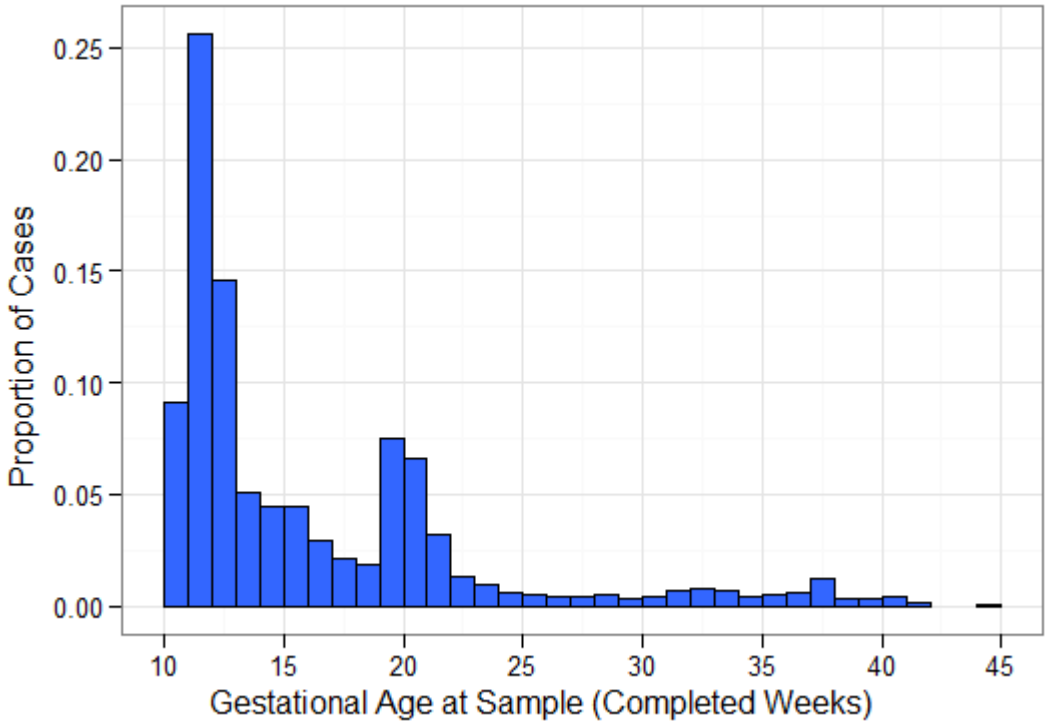

Figure 2.4. Graph of gestational age at diagnosis for trisomy 18 cases reported to the NDSCR during the study period.

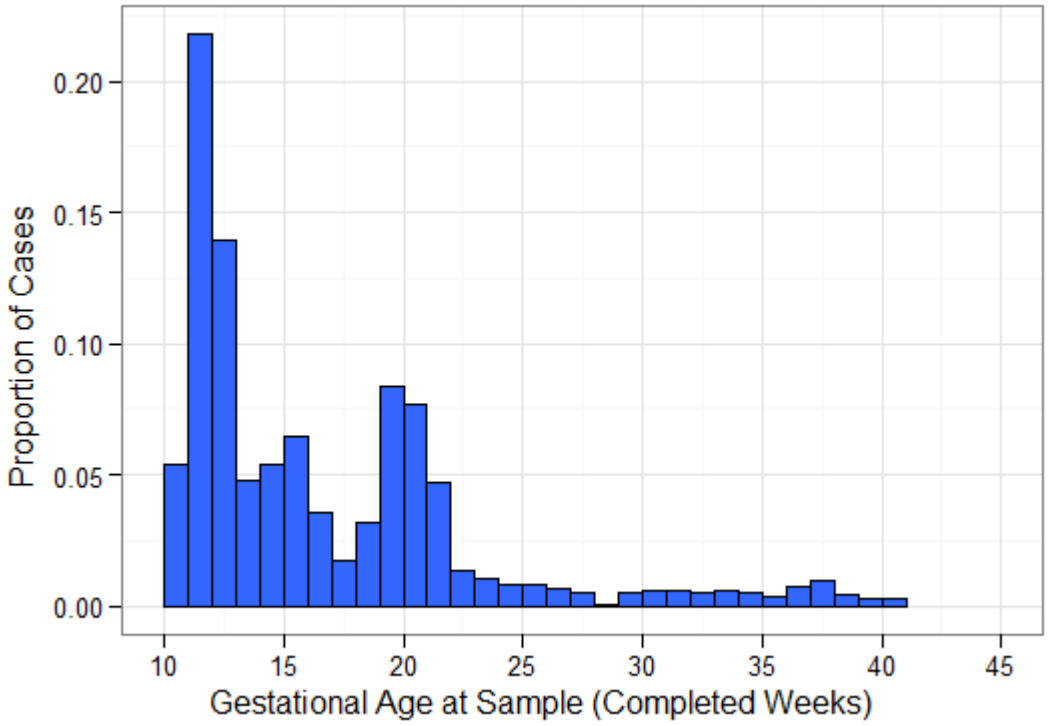

Figure 2.5. Graph of gestational age at diagnosis for trisomy 13 cases reported to the NDSCR during the study period.

There are 9 named types of tissue in NDSCR data which were sampled for cytogenetic diagnosis. The timing of five tissue types (amniotic fluid, fetal skin, placenta, villi and umbilical cord) was distributed between 14 and 20 weeks and could not be used to reliably estimate gestational age at diagnosis, however four tissue types (CVS, product of conception, fetal blood and postnatal) could be accurately used to estimate gestational age with a minimum of 80% accuracy.. Table 2.8 shows the timing and accuracy of each of these four tissue types.

Table 2.8. Tissue sample types which can be used to estimate gestational age at diagnosis.

Tissue	Timing of Test	Proportion of Cases
CVS	Before 14 weeks	90%
Product of Conception	Before 14 weeks	85%
Fetal Blood	After 19 weeks	80%
Postnatal	After 19 weeks	80%

Table 2.89 shows that a total of 136 cases of trisomy 18 and 61 cases of trisomy 13 were assigned gestational age at diagnosis based on their tissue sample type. 105 cases of trisomy 18 and 67 cases of trisomy 13 could not have gestational age estimated using their tissue type and were excluded from all adjusted spatial scans.

Table 2.9. Number of cases for which gestational age at diagnosis was estimated using tissue sample type.

	Trisomy 18		Trisomy 13	
	Number	**%**	**Number**	**%**
Missing Gestational Age	241	6.9	128	8.7
CVS	34	1	15	1
Product of Conception	2	0.1	0	0
Blood	33	0.9	14	1
Postnatal	67	1.9	32	2.2
Total Estimated	136	3.9	61	4.2
Total Still Missing	105	3	67	4.5

Chapter 3: Temporal Cluster Detection

Section 3.1. Introduction

Temporal disease clusters occur when increased disease prevalence is reported over a period of time before returning to the previously observed prevalence. Such clusters arise due to the presence of risk factors which are localised in time, such as the radiation related trisomy 21 clusters in Europe in January 1987. The discovery of temporal clusters can be useful in identifying novel risk factors and improving the general understanding of medical conditions. In the case of the trisomy syndromes, temporal clusters which could be linked to alternate risk factors could be used to provide prospective mothers with more accurate information relating to their risk of having an affected pregnancy.

Recent evidence suggests that the prevalence of trisomies 18 and 13 has been consistent in England and Wales in recent years (NDSCR Annual Reports 2004 – 2010). However, there is some evidence to suggest that increased trisomy 21 risk can be caused by radiation and infection (see literature review chapter 1.4). It is expected that the same factors which caused trisomy 21 clusters could also cause clusters of trisomies 18 and 13, however both syndromes are so rare that it is often difficult to obtain sufficient data to perform cluster analyses. The NDSCR is a large, complete register of all cytogenetically confirmed trisomy 21, 18 and 13 cases in England and Wales and therefore has the power to detect clustering effects which smaller databases may have been unable to discover. However, the only risk factor which been identified in England and Wales which influence the observed prevalence of trisomies 18 and 13 is maternal age at conception, which would be more likely to lead to a long term change in trisomy prevalence rather than the presence of distinct temporal clusters.

Section 3.2. Methods

Temporal cluster analyses used NDSCR data from the 1st of January 2004 to the 31st of December 2010, which consisted of 1,218 cases of trisomy 13 and 2,941 cases of trisomy 18. All data was analysed within government office region, which have been the primary classification for regional statistics since 1996. This allows the location of any clusters to be more accurately determined, and allows regional comparison of results. For example, clusters found at the same time in adjacent regions could indicate a risk factor affecting both regions simultaneously. It is possible that clusters occurring on the border of 2 or more regions could be missed using this type of analysis, however it is likely that such clusters would be detected in all contributing regions, albeit not as strongly. The 10 different regions are described in table3.1.1 and shown in figure 3.1.1.

Table 3.1.1. Characteristics of government office regions in England and Wales. Population estimates in millions and annual births shown are from 2007 (middle of study period).

Code	Region Name	Population (m)	Annual Births
1	North East	2.6	29,582
2	North West	6.9	85,947
3	Yorkshire and the Humber	5.2	64,191
4	East Midlands	4.4	52,482
5	West Midlands	5.4	70,098
6	East of England	5.7	69,311
7	London	7.6	125,505
8	South East	8.3	101,238
9	South West	5.2	57,003
10	Wales	3.0	34,414

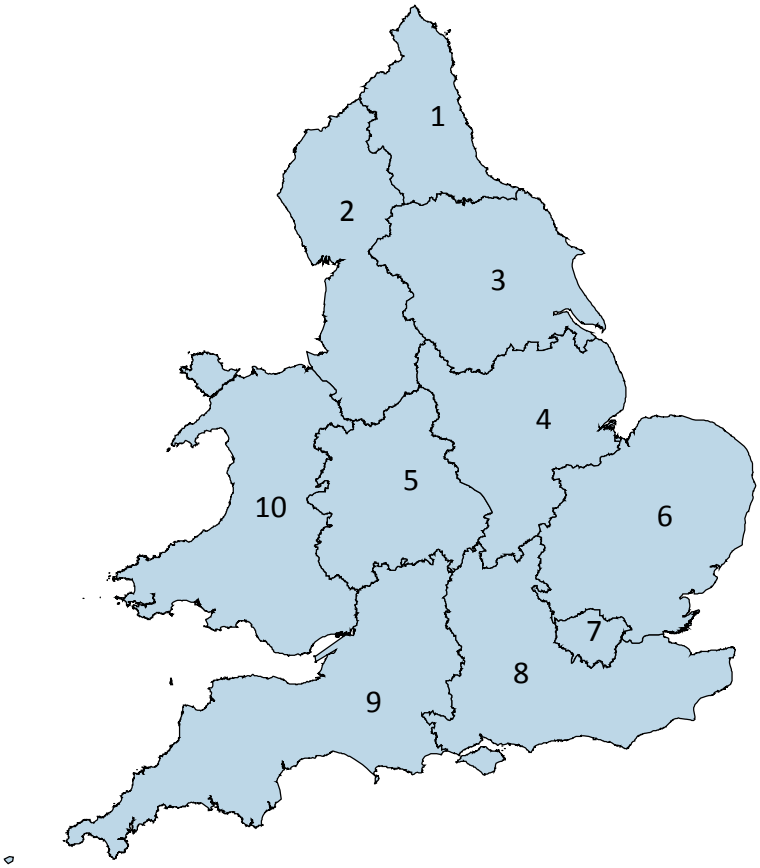

Figure 3.1.1. Map of government office regions in England and Wales. This map was generated in STATA using the spmap function.

Analysis was performed using the method detailed by Nagarwalla in 1996 (full details in chapter 1), which is a generalised likelihood ratio test for the null hypothesis that events happening within a set time period are uniformly distributed against the alternative hypothesis that the events are not evenly distributed.

This method finds the shortest time for a set number of cases to occur and uses Monte Carlo simulation to determine whether this time is significantly less than would be expected by chance. This technique differs from those previously available because it uses a scanning window of variable size and position. Using a fixed window means that the investigator must judge the size of potential clusters in advance without examining the data, so the variable window avoids any potential bias.

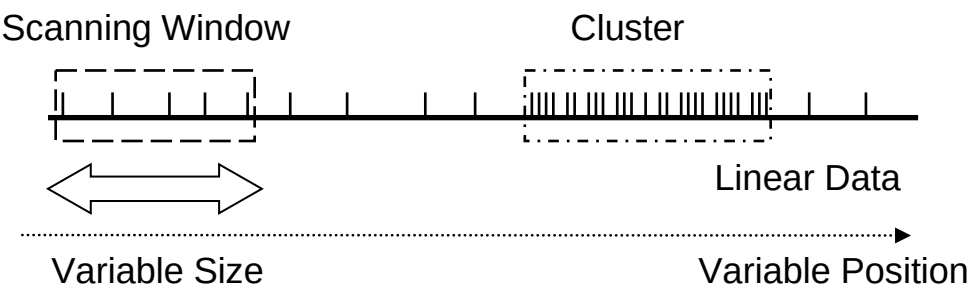

Figure 3.1.2.The scanning window method of temporal cluster analysis.

The scanning window searches through every combination of consecutive cases and finds the minimum time for every possible cluster size to occur, ranging from 2 to 2 less than the total number of cases in the data. For example, if there are 20 cases in the data the method will search for clusters of size 2-18 in sequence and for each cluster size store the most likely cluster; i.e. that which occurs over the least number of days. Clusters must have a minimum of 2 cases and occur over a minimum of 2 days. A scan statistic is then calculated for each of these potential clusters using the formula below.

Find: $\lambda[n] = \left(\frac{n}{N}\right)^{n}\left(\frac{N-n}{N}\right)^{N-n}\left(\frac{1}{d_{\min}}\right)^{n}\left(\frac{1}{1-d_{\min}}\right)^{N-n}$

Where n = the number of cases in the cluster, N = the total number of cases, d_{min} = the minimum number of days for n cases to occur as a proportion of the total number of days and $\lambda[n]$ = the scan statistic for that value of n.

Section 3.3. Programming and Implementation

The temporal scanning method was programmed using R, which is a freely available language and environment for statistical computing and graphics.

The computer program first reads in a file containing every case and the relevant date (either date of conception or diagnosis), and then ranks cases by date. It then uses the previously detailed variable scanning window to search through the data and find the cluster with the largest scan statistic for every possible cluster size and returns the 10 largest scan statistics. When data is analysed by individual region, the top 10 results per region are returned. Some regions have small numbers of cases and less than 10 results in total; in these instances all results are returned regardless of size. The idea of returning the top 10 results per region is to have the ability to potentially detect multiple non-overlapping clusters, in such an event p-values would be calculated for all non-overlapping clusters. However, in this thesis this scenario was not encountered as the top 10 results were either overlapping or obviously not significant.

The original publication which detailed this method included test data of 35 cases of the congenital anomalies oesophageal atresia and tracheo-oesophageal fistula from Birmingham, U.K. between 1950 and 1955. After writing the program in R, I tested it on the test data to ensure that the published results matched the results that my program returned. Both results were identical and demonstrate that the method is running correctly in the R program.

The probability of obtaining a result at least as extreme as that observed, under the null hypothesis, (p value) is calculated using Monte Carlo simulation, which uses repeated random sampling to calculate a p-value.

To perform the Monte Carlo simulation, cases were reallocated in time by generating random numbers based on the number of days included in the scan. For example, there are 2,557 days from 01/01/2004 to 31/12/2010 so each case would be assigned a random value from 1 to 2,557. The data was then rescanned and the largest scan statistic was stored. This process was repeated a total of 999 times. The position of the observed scan statistic amongst the list of simulated scan statistics then gives an estimate of the p-value, for example is the observed scan statistic is greater than all the simulated scan statistics $p < 1/999$, i.e. $p = 0.001$.

As both the cluster scans and Monte Carlo simulations return their strongest clusters, there is no need to account for multiple testing within a single region. However, it is necessary to account for each of the 10 regions being individually scanned. A Bonferroni correction was applied to all observed results, which divides the p-value used to determine significance by the

number of tests being performed, so for each trisomy the significance threshold was 0.05 / 10 = 0.005.

Four sets of temporal scans were run for each trisomy. Separate scans were performed using cases arranged by date of sample and by estimated date of conception. Cluster scanning using sample dates is a useful tool for quickly analysing data for clusters as soon as case data becomes available and when there may be minimal information for each case. However, trisomies 18 and 13 are genetic conditions where the initial error must occur at the time of conception for an individual to carry a full trisomy, so more informative results can be obtained using estimated date of conception.

Both of these tests were performed with and without a limit imposed on maximum cluster size. It is possible that the program could detect very large clusters spanning several years, but these would be more likely to be the result of a long term change in maternal age leading to higher trisomy rates over time than a distinct temporal cluster of cases. Imposing a limit removes the possibility of detecting such clusters. When applied, the size limit was chosen to be 20% of the total number of cases in the scan. This limit was chosen as a suitable limit to enable the detection of reasonably sized clusters spanning up to around 18 months in length without also picking up excessively large clusters. Scans were performed using a range of size limits and 20% was consistently found to be a suitable threshold. To avoid excessive repetition in this thesis, only the results obtained with a 20% limit are shown.

Section 3.4. Results

a. Characteristics of Sample

All cases of trisomies 18 and 13 which were reported to the NDSCR between January 1st 2004 and December 31st 2010 and did not have any other trisomy (i.e. were reported as having both trisomies 21 and 18) were eligible for inclusion in the analyses undertaken in this thesis. In total, there were 4 pairs of twins with trisomy 18 and 1 pair of trisomy 13 in the data. Each twin pair was taken to be 1 trisomy event, resulting in the exclusion of 1 case from each pair prior to finalising the dataset. Cases which had either an unaffected twin or one affected by a trisomy 21 were included in the analyses. There were no twin pairs reported to the NDSCR between 2004 and 2010 in which one twin had trisomy 18 and the other had trisomy 13.
In total, 3,507 cases of trisomy 18 and 1,473 cases of trisomy 13 were included in the dataset analysed in this thesis. An overview of the characteristics of these cases is shown in tables 3.4.1 to 3.4.4.

Table 3.4.1 shows that during the study period, the prevalence of trisomies 18 and 13 increased by around 25% for trisomy 18 and 29% for trisomy 13. This is likely due to a combination of increasing maternal age in England and Wales during this time period and advances in prenatal screening techniques resulting in more cases being detected prior to fetal loss and reported to the NDSCR.

Table 3.4.1. Birth totals and prevalence per 10,000 births in England and Wales of trisomies 18 and 13 by year.

	England and Wales	Trisomy 18		Trisomy 13	
	Births	Births	Prevalence	Births	Prevalence
2004	639,721	393	6.14	171	2.67
2005	645,835	466	7.22	186	2.88
2006	669,601	503	7.51	229	3.42
2007	690,013	518	7.51	224	3.25
2008	708,711	531	7.49	217	3.06
2009	706,248	541	7.66	199	2.82
2010	723,165	555	7.67	247	3.42
Total	4,783,294	3,507	7.33	1,473	3.08

Table 3.4.2. Sex distribution of cases of trisomies 18 and 13 during the study period.

	Trisomy 18 (%)	Trisomy 13 (%)
Male	1487 (42.4)	737 (50.0)
Female	1836 (52.4)	646 (43.9)
Unknown	184 (5.2)	90 (6.1)

Table 3.4.3. Maternal age mean and range for trisomies 18 and 13 during the study period.

Age (years)	Trisomy 18	Trisomy 13
Mean	36	34
Minimum	15.4	15.1
Maximum	51.8	50.8

b. Date of Sample with No Cluster Size Restrictions

Table 3.4.4 shows no evidence of clustering for trisomy 18 when cases were arranged by sample date.

Table 3.4.4. Date of Sample with No Cluster Size Restrictions – Trisomy 18 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	161	08-May-04	09-May-04	3	0.409
North West	284	19-Dec-06	20-Dec-06	3	0.865
Yorkshire and the Humber	229	10-May-09	20-May-09	6	0.852
East Midlands	210	19-Oct-10	30-Dec-10	15	0.078
West Midlands	305	03-Dec-06	05-Dec-06	4	0.650
East of England	378	31-Oct-06	10-Jun-10	240	0.068
London	847	13-Oct-10	15-Oct-10	7	0.250
South East	529	11-Apr-10	28-Apr-10	14	0.341
South West	333	06-Aug-08	07-Aug-08	4	0.325
Wales	163	17-Oct-06	19-Oct-06	3	0.834

Table 3.4.5 shows no evidence of clustering for trisomy 13 when cases were arranged by sample date.

Table 3.4.5. Date of Sample with No Cluster Size Restrictions – Trisomy 13 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	93	11-Dec-08	12-Dec-08	3	0.075
North West	129	14-Nov-07	18-Nov-07	4	0.384
Yorkshire and the Humber	116	02-Jun-08	03-Jun-08	3	0.131
East Midlands	84	17-Oct-10	20-Oct-10	3	0.465
West Midlands	117	30-May-07	01-Jun-07	4	0.031
East of England	153	13-Jun-08	16-Jun-08	3	0.939
London	319	09-Jan-06	10-Jan-06	5	0.012
South East	229	24-Jul-08	25-Jul-08	3	0.561
South West	145	06-Apr-09	08-Apr-09	3	0.744
Wales	52	13-Jul-05	24-Jul-05	3	0.895

c. Date of Sample with 20% Upper Limit

Table 3.4.6 shows that after the cluster size limit was imposed, there was no evidence of clustering for trisomy 18. There was 1 change in the results due to the size limit, which was in the East of England. The most likely cluster contained 240 out of 378 cases without the size limit and 35 cases with the limit.

Table 3.4.6. Date of Sample with 20% Upper Limit – Trisomy 18 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	161	08-May-04	09-May-04	3	0.402
North West	284	19-Dec-06	20-Dec-06	3	0.860
Yorkshire and the Humber	229	10-May-09	20-May-09	6	0.851
East Midlands	210	19-Oct-10	30-Dec-10	15	0.073
West Midlands	305	03-Dec-06	05-Dec-06	4	0.644
East of England	378	31-Oct-06	14-Feb-07	35	0.156
London	847	13-Oct-10	15-Oct-10	7	0.250
South East	529	11-Apr-10	28-Apr-10	14	0.330
South West	333	06-Aug-08	07-Aug-08	4	0.321
Wales	163	17-Oct-06	19-Oct-06	3	0.829

Table 3.4.7 shows that after the cluster size limit was imposed, there was no evidence of clustering for trisomy 13. No cluster sizes or locations changed in any region as a result of the upper limit.

Table 3.4.7. Date of Sample with 20% Upper Limit – Trisomy 13 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	93	11-Dec-08	12-Dec-08	3	0.075
North West	129	14-Nov-07	18-Nov-07	4	0.373
Yorkshire and the Humber	116	02-Jun-08	03-Jun-08	3	0.129
East Midlands	84	17-Oct-10	20-Oct-10	3	0.455
West Midlands	117	30-May-07	01-Jun-07	4	0.031
East of England	153	13-Jun-08	16-Jun-08	3	0.937
London	319	09-Jan-06	10-Jan-06	5	0.012
South East	229	24-Jul-08	25-Jul-08	3	0.552
South West	145	06-Apr-09	08-Apr-09	3	0.738
Wales	52	13-Jul-05	24-Jul-05	3	0.886

d. Date of Conception with No Cluster Size Restrictions

Table 3.4.8shows no evidence of clustering for trisomy 18 when cases were arranged by date of conception.

Table 3.4.8. Date of Conception with No Cluster Size Restrictions – Trisomy 18 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	156	03-Feb-09	05-Feb-09	3	0.701
North West	259	23-Feb-10	24-Feb-10	3	0.851
Yorkshire and the Humber	203	25-Sep-07	26-Sep-07	3	0.567
East Midlands	195	14-Jun-10	16-Jun-10	4	0.244
West Midlands	288	13-May-07	16-May-07	6	0.043
East of England	361	30-Jan-06	27-Sep-09	238	0.061
London	775	03-Jul-08	04-Jul-08	5	0.683
South East	513	02-Mar-09	03-Mar-09	5	0.095
South West	321	23-Dec-04	31-Dec-04	7	0.556
Wales	148	05-Nov-06	06-Nov-06	3	0.272

Table 3.4.9 shows no evidence of clustering for trisomy 13 when cases were arranged by date of conception.

Table 3.4.9. Date of Conception with No Cluster Size Restrictions – Trisomy 13 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	86	16-May-10	18-May-10	3	0.298
North West	113	04-Jan-10	06-Jan-10	3	0.563
Yorkshire and the Humber	104	20-Jul-04	28-Jul-04	4	0.734
East Midlands	76	14-Sep-05	21-Sep-05	3	0.993
West Midlands	107	09-Feb-09	11-Feb-09	3	0.491
East of England	147	27-May-06	29-May-06	3	0.729
London	288	24-Apr-07	25-Apr-07	4	0.215
South East	215	15-Jun-09	16-Jun-09	3	0.598
South West	135	10-Nov-05	05-Apr-10	105	0.941
Wales	48	19-Aug-08	03-Sep-08	4	0.618

e. Date of Conception with 20% Upper Limit

Table 3.4.10. Date of Conception with 20% Upper Limit – Trisomy 18 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	156	03-Feb-09	05-Feb-09	3	0.691
North West	259	23-Feb-10	24-Feb-10	3	0.849
Yorkshire and the Humber	203	25-Sep-07	26-Sep-07	3	0.557
East Midlands	195	14-Jun-10	16-Jun-10	4	0.240
West Midlands	288	13-May-07	16-May-07	6	0.042
East of England	361	31-Aug-06	07-Sep-06	8	0.130
London	775	03-Jul-08	04-Jul-08	5	0.682
South East	513	02-Mar-09	03-Mar-09	5	0.093
South West	321	23-Dec-04	31-Dec-04	7	0.438
Wales	148	05-Nov-06	06-Nov-06	3	0.267

Table 3.4.10 shows that after the cluster size limit was imposed, there is no evidence of clustering for trisomy 18. The only change in results due to the size limit was a much smaller cluster identified in the East of England.

Table 3.4.11. Date of Conception with 20% Upper Limit – Trisomy 13 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	86	16-May-10	18-May-10	3	0.294
North West	113	04-Jan-10	06-Jan-10	3	0.552
Yorkshire and the Humber	104	20-Jul-04	28-Jul-04	4	0.726
East Midlands	76	14-Sep-05	21-Sep-05	3	0.881
West Midlands	107	09-Feb-09	11-Feb-09	3	0.484
East of England	147	27-May-06	29-May-06	3	0.724
London	288	24-Apr-07	25-Apr-07	4	0.212
South East	215	15-Jun-09	16-Jun-09	3	0.594
South West	135	14-Nov-07	18-Nov-07	3	0.971
Wales	48	19-Aug-08	03-Sep-08	4	0.603

Table 3.4.11 shows that after the cluster size limit was imposed, there is no evidence of clustering for trisomy 13. The only change in results due to the size limit was a much smaller cluster identified in the South West.

f. Date of Conception: Sensitivity Analysis

In order to assess the effect of excluding around 10% of cases for both trisomies for the temporal scans by date of conception, further analyses were performed after imputing date of conception for more individuals and removing the timeframe restriction (more details in chapter 2).

As the temporal scans are based on the distribution of cases across the entire timeframe being studies, incomplete data at the beginning and end of the temporal period could cause the method to detect very large clusters including the majority of cases in the scan. For this reason, the previously used maximum cluster size limit of 20% of cases in the scan was applied.

Table 3.4.12. Date of Conception with 20% Upper Limit – Trisomy 18 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	161	03-Feb-09	05-Feb-09	3	0.665
North West	290	03-May-06	24-Mar-07	57	0.550
Yorkshire and the Humber	226	23-Jul-04	30-Jul-04	6	0.300
East Midlands	211	07-Apr-10	09-Apr-10	4	0.220
West Midlands	304	13-May-07	16-May-07	6	0.027
East of England	381	31-Aug-06	07-Sep-06	8	0.108
London	849	28-Jan-09	31-Jan-09	7	0.313
South East	538	02-Mar-09	03-Mar-09	5	0.046
South West	335	23-Dec-04	31-Dec-04	7	0.431
Wales	160	05-Nov-06	06-Nov-06	3	0.175

Table 3.4.13. Date of Conception with 20% Upper Limit – Trisomy 13 Results

Government Office Region	Total Cases	Start Date	End Date	Cases in Cluster	P Value
North East	93	16-May-10	18-May-10	3	0.278
North West	130	04-Jan-10	06-Jan-10	3	0.463
Yorkshire and the Humber	113	28-Oct-09	10-Nov-09	5	0.569
East Midlands	85	11-Dec-04	17-Dec-04	3	0.859
West Midlands	118	07-Aug-03	16-Aug-03	5	0.268
East of England	155	27-May-06	29-May-06	3	0.644
London	319	10-Jul-08	11-Jul-08	4	0.216
South East	232	08-Apr-09	09-Apr-09	4	0.073
South West	145	07-Sep-07	11-Sep-07	3	0.971
Wales	53	19-Aug-08	03-Sep-08	4	0.615

The results in tables 3.4.12 and 3.4.13 show that imputing additional information and increasing the proportion of cases which can be included in the analyses has changed the size and time of clusters in some regions, however there are still no statistically significant results. The inclusion of the additional cases has therefore not made much difference to the results overall.

Section 3.5. Interpretation

The results presented in this chapter show no evidence of temporal clustering of trisomies 18 and 13 in England and Wales between 2004 and 2010 by estimated date of conception.

Trisomies 18 and 13 are rare conditions with only a few hundred cases with each trisomy reported to the NDSCR every year. When this data is analysed by individual Government Office Region there are some regions which have very few cases, even when analysing 7 years of data. The fewest cases in one scan was for trisomy 13 in Wales when analysing by date of conception, in which there were 48 cases. The most cases in one scan was for trisomy 18 in London when analysing by date of sample, in which there were 847 cases. In section 2.3 (Programming and Implementation), it was stated that an example was provided in the initial publication of the temporal scanning method. This example included test data of 35 cases of the congenital anomalies oesophageal atresia and tracheo-oesophageal fistula from Birmingham, U.K. between 1950 and 1955. It therefore appears unlikely either that the sparseness of the NDSCR data was problematic or that the method cannot adequately identify clusters in sparse data, as the method had previously been tested in a scenario with fewer cases than any of the individual scans performed in this chapter.

Scanning each Government Office Region individually allows the location of any identified clusters to be more accurately determined and can be more informative, however there is also the possibility that clusters spanning 2 or more regions could go undetected using such an approach. The majority of the clusters identified were very small; including fewer than 10 cases and taking place over less than a week, even without applying the cluster size limit. These clusters also occurred at different points in the timeframe, with no temporal correlation. It therefore seems unlikely that significant clusters would be revealed by analysing regions together.

The cluster detection method used in this chapter is primarily intended to search for clusters rather than long term increases in prevalence which never return (or at least, not within the timeframe of the study) to the original level, and may not be able to detect increases of this nature. While long term increase in prevalence is a big concern, it is already known that the prevalence of both trisomies 18 and 13 increased during the study period (shown in table 3.4.1), mainly as a result of increasing maternal age in England and Wales (as shown in figure 1.3.8). The purpose of this thesis is to search for clusters which appear independently of this increase, which is why methods designed to detect long term changes in prevalence were not applied.

The most likely explanation for the lack of clusters present is that there was no temporal clustering in England and Wales between 2004 and 2010.

Chapter 4: Spatial Cluster Detection

Section 4.1. Introduction

Spatial cluster detection is the search for regions of space in which the occurrence of cases of interest are higher or lower than the surrounding areas. Spatial disease clusters occur wherever the prevalence of disease cases is significantly higher than expected. The discovery of spatial clusters can lead to the discovery of new risk factors and improvement in the overall understanding of the condition, which can be used to provide more accurate medical advice. One recent example of reported spatial disease clustering is the trisomy 21 clustering in the North of England reported by Mcnally *et al* in 2008.

It would be expected that trisomies 18 and 13 cluster in the area of England and Wales with the highest maternal age as this is such a strong risk factor for both conditions. Trisomy risk starts to rapidly increase from the age of 35 (Savva *et al* 2010), so an increase in the number of mothers aged 35 and above would be expected to lead to an increased prevalence of trisomies 18 and 13.

Section 4.2.Methods

Scanning Window

The spatial method uses a variable circular window to scan through the data and identify the most likely clusters. This window moves in two dimensional space (latitude and longitude coordinates are commonly used to identify spatial data points) to analyse the entire study region. Circles of increasing size are generated at specified points in the study region, and the total number of cases and controls within the circle is compared with those outside the circle (but within the study region). Different circle sizes are used to search for clusters of varying size.

Grid Method

There are two ways to determine the positions at which the circles are to be generated. The first is to lay a grid over a map of the study region and generate a circle at each grid intersection. This approach examines the entire study region, and can therefore be very informative when data fills the study region. However, when data is unevenly distributed throughout the study region or when the study region is a non-uniform shape there will be grid intersections which are not situated close to any data points and are therefore non-informative.

Aggregated Method

Alternately, when data has been aggregated into population groups it is possible to centre the circles at each aggregation point. Aggregating data is useful when cases / controls do not have sufficient information to be individually located, and instead groups individuals into population blocks. Aggregation can be by any population subgroup; commonly used groups include health regions and census districts. The main advantage of generating circles around aggregation points is that all information is included, but the program does not waste time scanning areas where there are no cases.

Formula

The spatial cluster scanning method uses a formula derived from the likelihood function under a binomial model, and is shown below. The formula is applied for each circle (z), where c_z and n_z are the number of cases and controls inside circle z, and C and N are the total number of cases and controls in the study region.

$$L(z) = (\frac{c_z}{n_z})^{c_z}(\frac{n_z - c_z}{n_z})^{n_z - c_z}(\frac{C - c_z}{N - n_z})^{C - c_z}(\frac{N - n_z - (C - c_z)}{N - n_z})^{(N - n_z) - (C - c_z)}$$

Monte Carlo

Once the test statistic has been calculated, statistical significance is determined using Monte Carlo simulation (see Chapter 3), where each case is redistributed at random in the study region, the region is rescanned and the maximum test statistic is stored. This is repeated 999 times to give a p value between 1 and 0.001.

Section 4.3.Implementation

Case and population birth data between 2004 and 2010 were obtained from the NDSCR and Office for National Statistics respectively. Case data was available for most cases by maternal postcode, while the smallest population unit for which birth data was available was maternal primary care trust (PCT). There were 174 PCTs in England (152) and Wales (22) during the study period, with an average population of around 330,000 in 2006 and a mean of around 3,847 births per PCT in the same year. Figure 4.1 shows the coordinates for the administrative centre of each PCT, which was taken to be the PCT centroid location. The postcode for each PCT administrative centre was obtained from the NHS Organisation Codes Service Access Database and converted to latitude and longitude coordinates. The NDSCR Annual Report stated that in 2006, 461 cases of trisomy 18 and 204 cases of trisomy 13 were reported to the register. This equals an expected prevalence of between 1 and 3 cases per PCT. The use of PCTs is an effective compromise between scan resolution and ensuring that cases are present at as many aggregation sites as possible. Use of aggregated districts smaller than PCTs would result in many districts having no cases present when running scans by individual year.

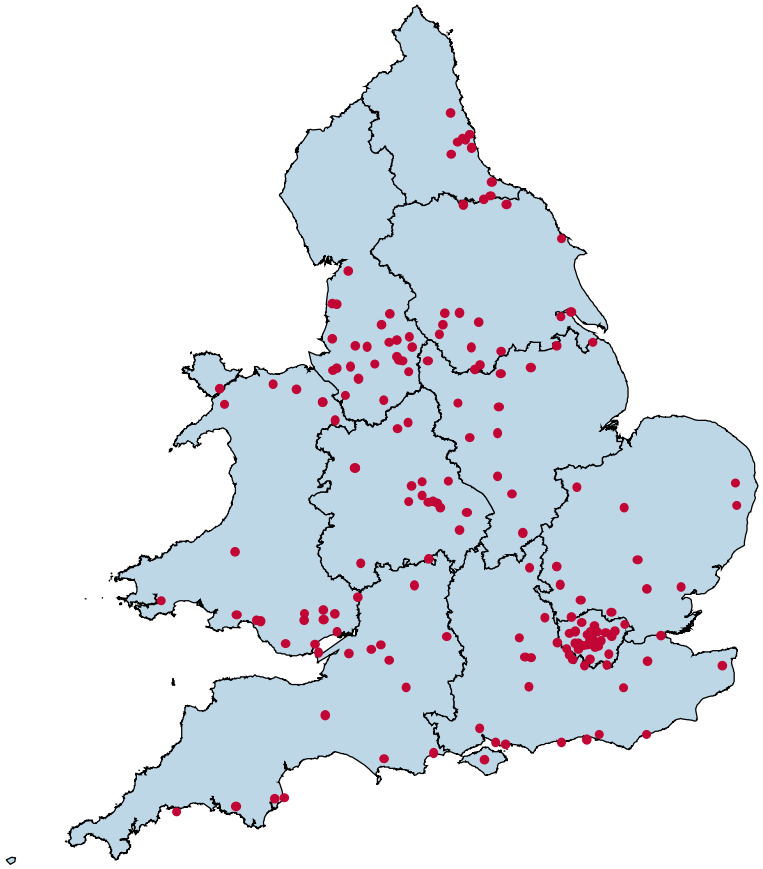

Figure 4.1. Map of PCT administrative centres in England and Wales. Map generated in STATA using the spmap function.

PCT administrative centres are clustered around centres of population. This is illustrated in figure 4.1, where clusters of PCTs can be observed in England around major cities including London, Birmingham and Manchester. In a single year, the total number of births per PCT can vary between less than 1,000 to over 10,000 births. Table 4.1 shows the number of births per PCT in 2008 arranged into 8 groups on the basis of total births.

Table 4.1.Number of births per PCT in 2008.

Total Births in 2008	Number of PCTs
Less than 1000	5
Between 1000 and 2000	28
Between 2000 and 3000	31
Between 3000 and 4000	39
Between 4000 and 5000	18
Between 5000 and 6000	22
Between 6000 and 10000	29
Greater than 10000	2

PCT locations and boundaries underwent major restructuring in 2006, when the total number of PCTs in England was reduced from 303 to 152. Case location data was converted to PCTs according to the 2006 boundaries. Population data was provided using the same boundaries for 2006, 2007 and 2008 but was unavailable for 2009 and 2010. Case data from 2004 and 2005 was compared to birth data from 2006, while cases in 2009 and 2010 were compared to birth data from 2008. Over time, the number of births in England and Wales is increasing. 669,376 births were reported in 2006 compared with 689,771 in 2007 and 708,549 in 2008, which is a 3.05% increase in the number of births from 2006 to 2007 and a 2.71% increase from 2007 to 2008. It therefore seems likely that of the available data, 2006 most closely resembles the expected birth figures from 2004 and 2005 while 2008 most closely resembles 2009 and 2010.

For scans using all 7 years data combined, population birth data was generated by combining 3 sets of 2006 and 2008 births with one set of 2007 births. Scans were performed using all data combined and also individually by calendar year. A Bonferroni correction for multiple testing was applied for scans using all 7 years data individually. This reduced the significance threshold to 0.05 / 7 = 0.007.

When using aggregated data, it is appropriate to generate circles at each aggregation point rather than using intersection points on a grid. Using this method does not adversely affect

power as areas with high concentrations of cases are still scanned, however the program runs more efficiently as no circles are generated in parts of the region where there are no cases.

Section 4.4.Programs

The Kulldorff and Nagarwalla spatial cluster detection method was programmed into 2 programs using the R statistical package. This method is henceforth referred to as the R Program. The first program read in a data file containing PCT location and the total number of cases and noncases within each PCT. The scan statistics of the 10 strongest clusters were then returned. P-values were then calculated using Monte Carlo simulation in the second program. A copy of both programs is included in the appendix.

Results generated in the R Program were compared with a second set of scans using Kulldorff's widely used SaTScan program. Results were expected to be comparable but not identical due to differences between the way in which the programs generate circles. Circles in R were generated using latitude and longitude values as radii, while SaTScan generates circles using kilometres as the radius size variable. This is significant as latitude and longitude values are not always equivalent, for example in England and Wales a latitude difference of 1 is equivalent to around 111km compared to a longitude difference of 1 which is around 70km. Consequently, circles generated in R were elliptical in shape while those in SaTScan were circular. Neither approach is superior or inferior to the other, however the difference would be expected to lead to a slight variation in cluster locations and p values. As another result of this difference, the methods generate circles of different sizes. Again this is only expected to lead to small changes in results.

In order to prevent both methods from returning excessively large clusters, the maximum circle radius was restricted to 90km in SaTScan and 1 unit of latitude and longitude in R. A 90km limit was chosen for SaTScan as the resulting circle would be of a similar area to the largest circle generated by the R Program (25,444km^2 in SaTScan vs 24,408km^2 in R). A comparison of results is provided between R, SaTScan with a 90km radius limit and SaTScan with no radius limit.

In addition to differences in circle size and shape, there are 2 other main differences between R and SaTScan which would not be expected to lead to any difference in the results.

Gumbel Approximation in SaTScan

SaTScan uses a Gumbel Approximation method which allows the program to calculate p-values lower than 0.001 even when using 999 sets of simulated results. The maximum likelihood ratio from each iteration of the Monte Carlo simulation is used to fit a Gumbel distribution to the data, and the resulting estimate of the p-value is the probability that this distribution generates a value greater than the observed maximum likelihood ratio for the real dataset. This method has been demonstrated to fit the data very well using the SaTScan spatial scan.

Sequential Monte Carlo Simulation in SaTScan

SaTScan uses a sequential Monte Carlo simulation method (Besag& Clifford, 1991) which allows the program to terminate the scan when the p-value is insignificant. Scans terminate when a fixed number of replicates (the default value is 50 for a 999 replicate Monte Carlo scan) have a test statistic greater than that of the observed data. This method does not negatively affect power at the 0.05 level, and allows the program to run more quickly when there is no evidence of clustering.

Section 4.5.Synthetic Data Testing

Section 4.5.1.Method

Detection performance for both methods was evaluated using synthetic data. The aim of this testing was to evaluate the power of the novel method in R and SaTScan and ensure that both programs returned expected results. A data set was created with a fixed number of cases (*x*) and controls (*y*) in each PCT, with 1 distinct region selected as a simple model cluster. Within the cluster, the initial number of cases in each PCT was set at *x + 1* and increased by 1 until a highly significant p value had been obtained in both programs.

The test cluster was located in 12 PCTs in the North East of England. This region was chosen because it could be identified using both R and SaTScan without including any other PCTs. As the programs use different circle sizes and shapes, the test cluster had to be an isolated group of PCTs.

Figure 4.2.Location of the synthetic test cluster.

The test data was chosen to approximate the real data as closely as possible by assigning each test PCT the mean number of births per PCT in 2008. In 2008, there were a total of 708,459 births in 174 PCTs, equal to a mean of 4071 births per PCT. For convenience when calculating

case: control ratios, this figure was rounded down to 4,000 births per PCT, which was a total of 696,000 births in the test data.

The programs were compared using a range of disease prevalence values. In PCTs outside the cluster, values of *x* were evaluated from 1 to 6. This range represents all prevalence figures which will be observed in the NDSCR case data (observed prevalence ranges from under 1 case per PCT for trisomy 13 to over 3 cases per PCT for trisomy 18) while also testing higher prevalence scenarios.

Section 4.5.2.Results

Results in tables 4.5.1 – 4.5.6 demonstrate that under simple simulated conditions, the p values returned by both programs are similar. The test cluster is always identified as the strongest cluster by both programs unless the total number of cases per PCT is the same inside and outside the test cluster region. The R program returns lower p values than SaTScan does, but these differences are small and can be accounted for by slight differences between the programs such as using different circle shapes and sizes.

Table 4.5.1. 1 case per PCT outside cluster

Cases per PCT			P Value	
Inside Cluster	Outside Cluster	Relative Risk	SaTScan	R
1	1	1.00	1	1
2	1	2.00	0.52	0.5
3	1	3.00	< 0.001	0.001

Table 4.5.2. 2 cases per PCT outside cluster

Cases per PCT			P Value	
Inside Cluster	Outside Cluster	Relative Risk	SaTScan	R
2	2	1.00	1	1
3	2	1.50	1	0.99
4	2	2.00	0.016	0.006
5	2	2.50	< 0.001	0.001

Table 4.5.3. 3 cases per PCT outside cluster

Cases per PCT			P Value	
Inside Cluster	Outside Cluster	Relative Risk	SaTScan	R
3	3	1.00	1	1
4	3	1.33	1	1
5	3	1.67	0.14	0.096
6	3	2.00	< 0.001	0.001

Table 4.5.4. 4 cases per PCT outside cluster

Cases per PCT			P Value	
Inside Cluster	Outside Cluster	Relative Risk	SaTScan	R
4	4	1.00	1	1
5	4	1.25	1	1
6	4	1.50	0.59	0.21
7	4	1.75	0.005	0.001
8	4	2.00	< 0.001	0.001

Table 4.5.5. 5 cases per PCT outside cluster

Cases per PCT			P Value	
Inside Cluster	Outside Cluster	Relative Risk	SaTScan	R
5	5	1.00	1	1
6	5	1.20	1	1
7	5	1.40	0.77	0.654
8	5	1.60	0.016	0.011
9	5	1.80	< 0.001	0.001

Table 4.5.6. 6 cases per PCT outside cluster

Cases per PCT			P Value	
Inside Cluster	Outside Cluster	Relative Risk	SaTScan	R
6	6	1.00	1	1
7	6	1.17	1	1
8	6	1.33	0.89	0.726
9	6	1.50	0.064	0.03
10	6	1.67	< 0.001	0.001

Section 4.6.Analysis

Section 4.6.1.Distribution of Cases

Prevalence of trisomies 18 and 13 between 2004 and 2010 in each PCT were calculated and presented in figures 4.3 and 4.4. Groups of high prevalence PCTs may indicate the location of possible clusters.

Figure 4.3 shows one strong cluster location with several high prevalence PCTs in central and west London. This location is the only grouping of PCTs with prevalence of greater than 10 in 10,000 births, and would be expected to be detected as an extremely strong cluster by both R and SaTScan.

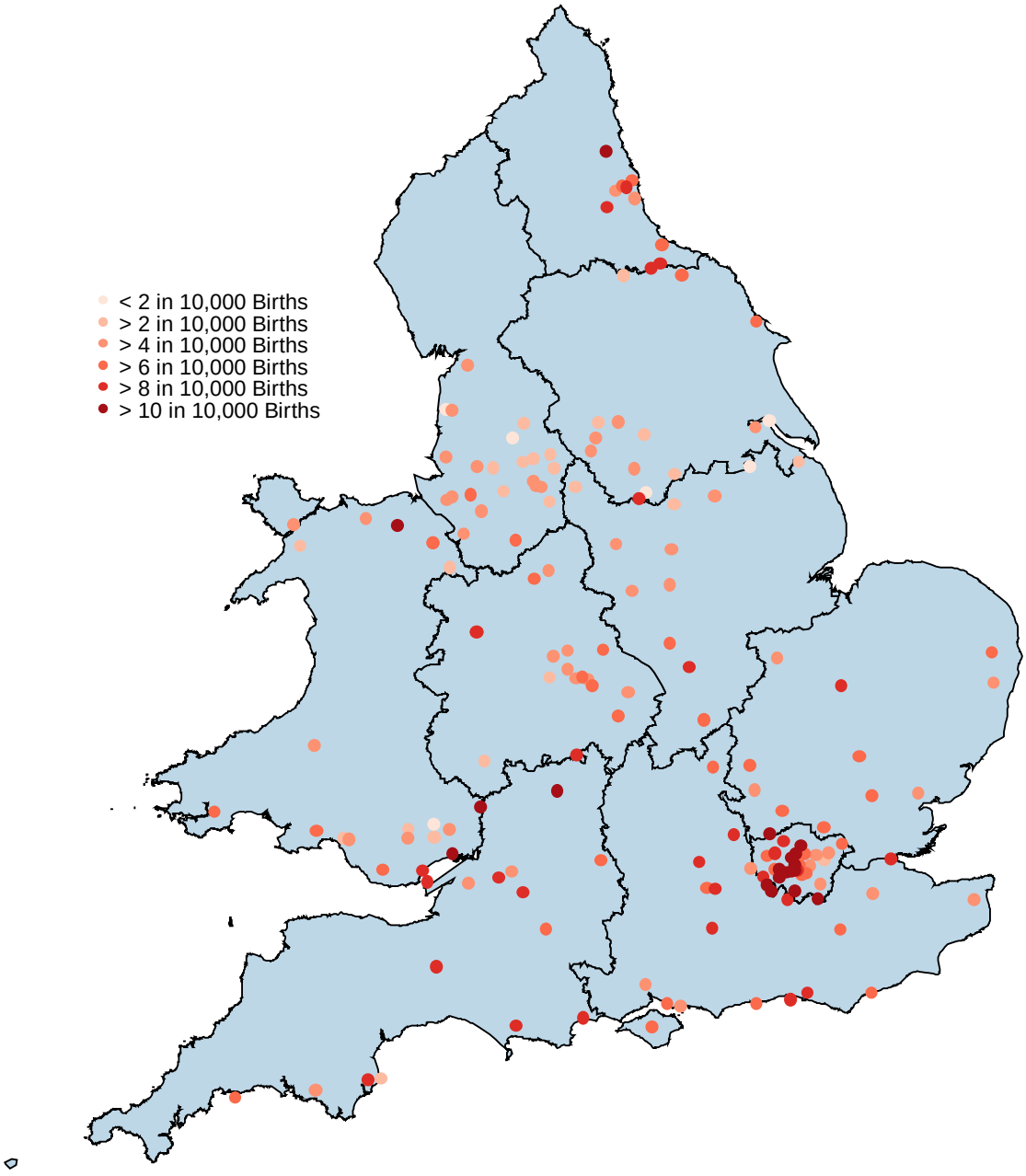

Figure 4.3.Prevalence of trisomy 18 according to PCT in England and Wales from 2004 to 2010.

Figure 4.4 appears to show two potential cluster locations for trisomy 13; one in central and west London and another in the North East. The London cluster is larger and contains more PCTs, however the north east cluster shows a localised group of 5 PCTs all with prevalence of greater than 5 in 10,000 births. It is probable that the larger cluster will be more easily detected, however the north east could also be identified as a secondary cluster hotspot.

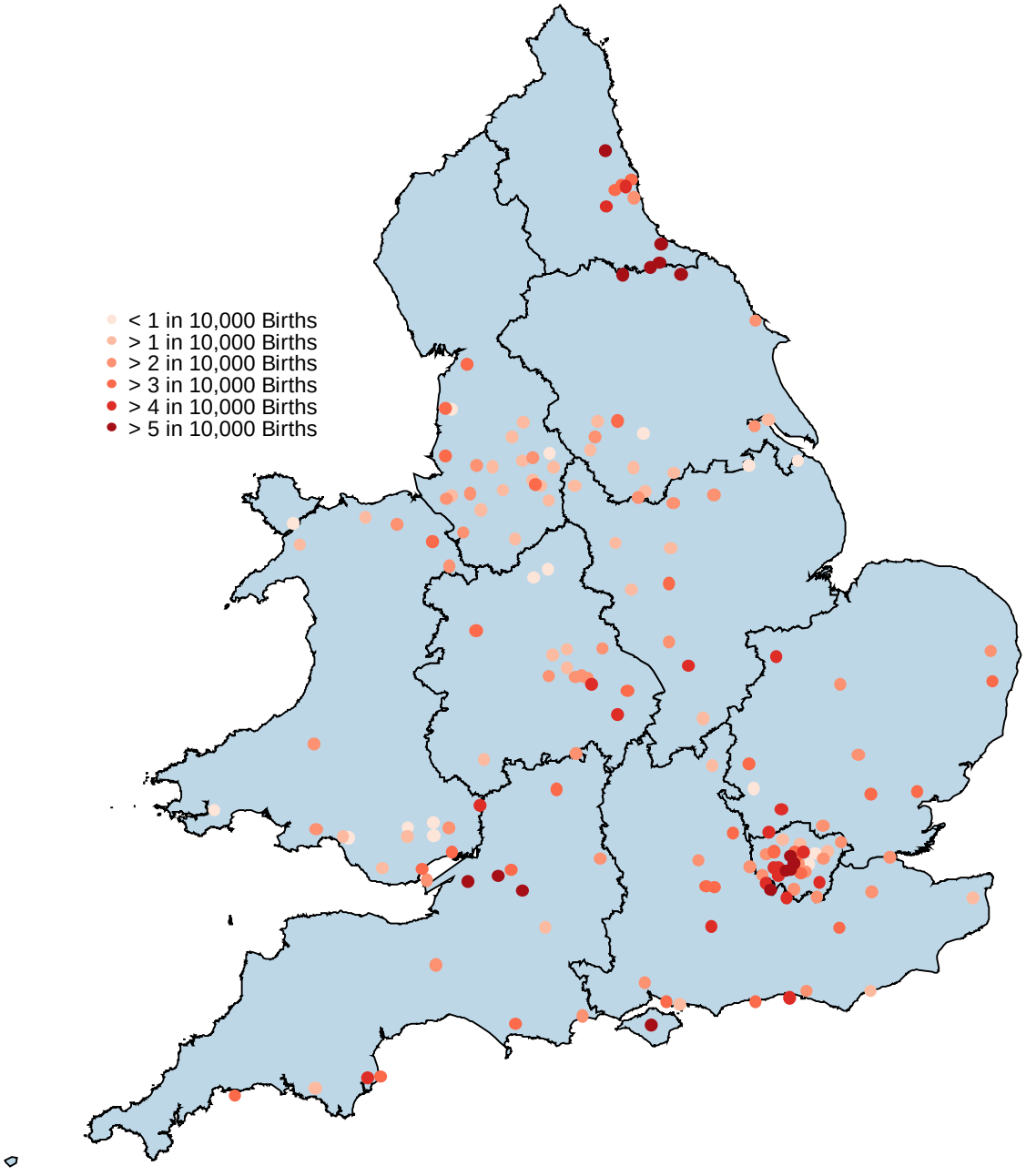

Figure 4.4.Prevalence of trisomy 13 according to PCT in England and Wales from 2004 to 2010.

Section 4.6.2.Trisomy 18 Scan Results

By Individual Calendar Year

Table 4.6.1 shows that for trisomy 18 data analysed by individual calendar year there was significant evidence of clustering detected by both R and SaTScan. The programs detected overlapping clusters in each of the 7 years scanned, and after applying a Bonferroni correction for multiple testing the same 5 clusters (2004, 2006, 2008, 2009 and 2010) were judged significant by both programs.

Table 4.6.1. Trisomy 18 scan results by calendar year.

		Centroid Coordinates		Circle Radius			
Year	**Cases in Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	*174*	*51.908 N*	*0.443 W*	*111*	*70*	*0.001*	*R*
	186	*51.908 N*	*0.443 W*	*81.73*	*81.73*	*<0.0001*	*S*
2005	*148*	*50.841 N*	*0.144 W*	*99.9*	*63*	*0.072*	*R*
	76	*51.509 N*	*0.477 W*	*25.86*	*25.86*	*0.026*	*S*
2006	*117*	*51.908 N*	*0.443 W*	*66.6*	*42*	*0.001*	*R*
	148	*51.908 N*	*0.443 W*	*55.27*	*55.27*	*<0.0001*	*S*
2007	*194*	*51.574 N*	*0.071 E*	*99.9*	*63*	*0.018*	*R*
	148	*51.879 N*	*0.544 E*	*70.19*	*70.19*	*0.023*	*S*
2008	*235*	*51.391 N*	*0.302 W*	*77.7*	*49*	*0.001*	*R*
	245	*51.391 N*	*0.302 W*	*72.73*	*72.73*	*<0.0001*	*S*
2009	*176*	*51.206 N*	*0.797 W*	*88.8*	*56*	*0.001*	*R*
	170	*51.206 N*	*0.797 W*	*62.99*	*62.99*	*<0.0001*	*S*
2010	*113*	*51.460 N*	*0.239 W*	*22.2*	*14*	*0.001*	*R*
	77	*51.351 N*	*0.165 W*	*17.87*	*17.87*	*<0.0001*	*S*

Indicates overlapping clusters detected in R and SaTScan.

SaTScan Radius Limit Comparison

Table 4.6.2 shows that imposing a 90km limit on the maximum cluster size detected by SaTScan had no major effect on the reported clusters. There were small changes in cluster centroids for 6 out of 7 years, however the identified clusters showed significant overlap and would have included many of the same PCTs. While it would be expected that clusters with fewer cases were detected using the 90km limit, results in 2004 show that the most significant clusters with and without the 90km radius limit included the same number of cases. This is due to a change in the cluster centroid location which meant that the cluster detected using the 90km limit was in a more densely populated region and therefore included more cases relative to the cluster radius.

Table 4.6.2.Trisomy 18 comparison of SaTScan results using a 90km radius limit or no geographical limit.

		Coordinates					
Year	**Cases In Cluster**	**Latitude**	**Longitude**	**Radius**	**P-Value**	**Limit**	**Overlap?**
2004	186	50.749 N	1.880 W	155.1	<0.0001	None	Yes
	186	51.908 N	0.443 W	81.73	<0.0001	90km	
2005	197	50.701 N	1.306 W	134.54	0.017	None	Yes
	76	51.509 N	0.477 W	25.86	0.026	90km	
2006	248	52.022 N	0.791 W	92.81	<0.0001	None	Yes
	148	51.908 N	0.443 W	55.27	<0.0001	90km	
2007	263	51.404 N	0.774 W	112.03	0.017	None	Yes
	148	51.879 N	0.544 E	70.19	0.023	90km	
2008	327	51.908 N	0.443 W	125.68	<0.0001	None	Yes
	245	51.391 N	0.302 W	72.73	<0.0001	90km	
2009	307	50.704 N	1.306 W	170.66	<0.0001	None	Yes
	170	51.206 N	0.797 W	62.99	<0.0001	90km	
2010	77	51.351 N	0.165 W	17.87	<0.0001	None	Yes
	77	51.351 N	0.165 W	17.87	<0.0001	90km	

All Data from 2004 - 2010

R

Cases		Centroid Coordinates		Radius (km)		
Total	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**
3321	1068	51.509 N	0.477 W	55.5	35	0.001

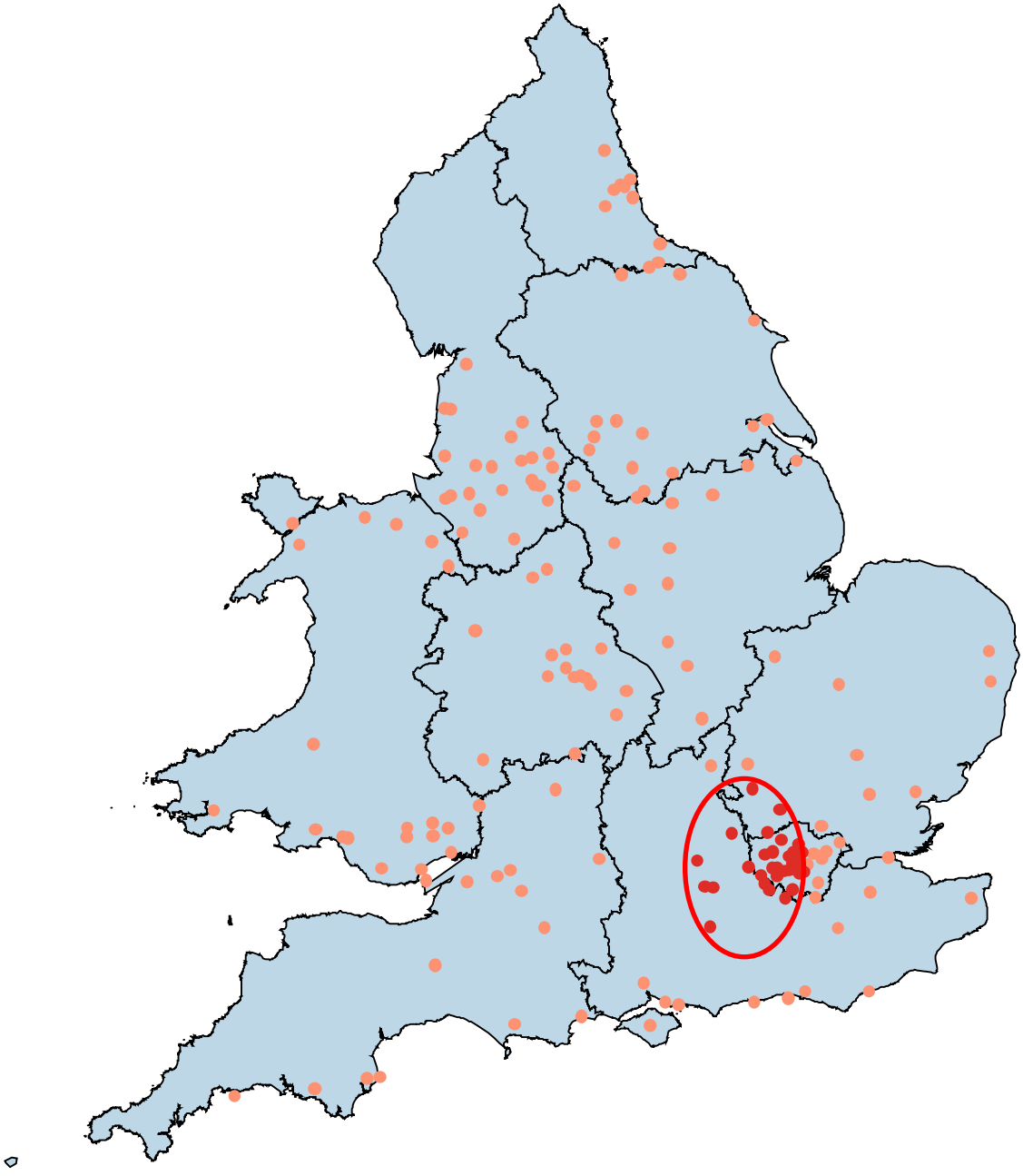

Figure 4.5. Map of trisomy 18 cluster by PCT using all data from 2004-2010 in R.

SaTScan

Cases		Centroid Coordinates			
Total	**In Cluster**	**Latitude**	**Longitude**	**Radius (km)**	**P Value**
3321	1109	51.404 N	0.774 W	60.21	1E-17

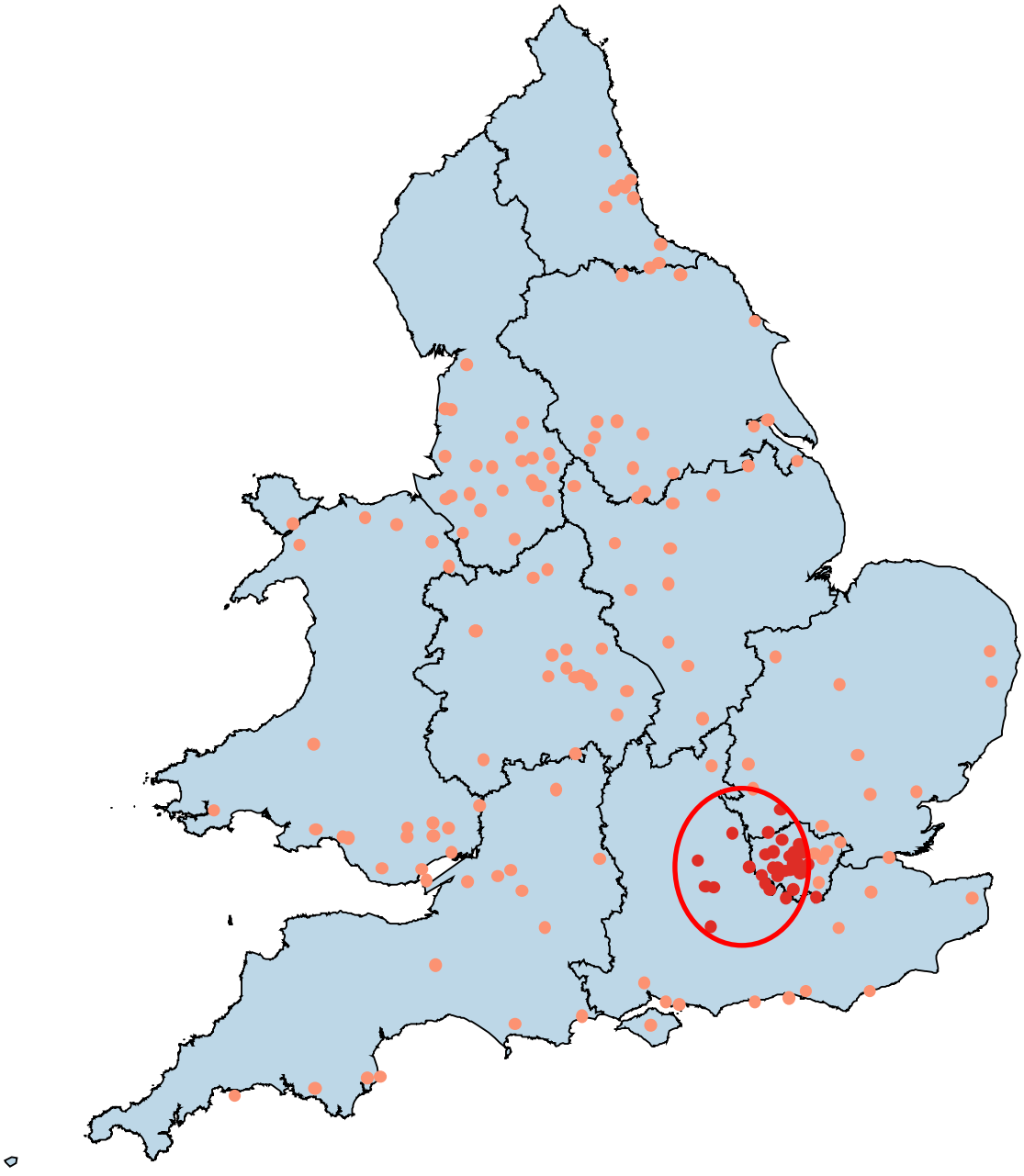

Figure 4.6. Map of trisomy 18 cluster by PCT using all data from 2004-2010 in SaTScan.

Figures 4.5 and 4.6 show that when analysing all trisomy 18 data together, R and SaTScan both detect very strong clusters in London which are almost identical in size and location.

Section 4.6.3 Trisomy 13 Scan Results

By Individual Calendar Year

Table 4.6.3 shows that for trisomy 13 data analysed by individual calendar year there was significant evidence of clustering detected by both R and SaTScan. After applying the Bonferroni correction, significant clusters were detected by both programs in 2006, 2009 and 2010. Clusters in 2006 and 2010 were located in London while the centroid of the 2009 cluster was situated in Bristol.

Table 4.6.3. Trisomy 13 scan results by calendar year.

		Centroid Coordinates		Circle Radius			
Year	**Cases in Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	*18*	*50.806 N*	*1.062 W*	*55.5*	*35*	*0.23*	*R*
	27	*50.820 N*	*0.424 W*	*63.96*	*63.96*	*0.059*	*S*
2005	4	55.161 N	1.691 W	11.1	7	0.69	R
	14	51.009 N	3.115 W	62.89	62.89	0.58	S
2006	*27*	*51.391 N*	*0.302 W*	*33.3*	*21*	*0.001*	*R*
	50	*51.509 N*	*0.477 W*	*25.86*	*25.86*	*0.0002*	*S*
2007	*42*	*51.904 N*	*2.097 W*	*77.7*	*49*	*0.25*	*R*
	24	*51.904 N*	*2.097 W*	*59.82*	*59.82*	*0.096*	*S*
2008	*76*	*51.206 N*	*0.797 W*	*88.8*	*56*	*0.011*	*R*
	72	*51.206 N*	*0.797 W*	*62.99*	*62.99*	*0.039*	*S*
2009	*14*	*51.461 N*	*2.592 W*	*33.3*	*21*	*0.001*	*R*
	16	*51.461 N*	*2.592 W*	*39.15*	*39.15*	*0.0062*	*S*
2010	*90*	*51.492 N*	*0.137 W*	*11.1*	*7*	*0.005*	*R*
	27	*51.492 N*	*0.137 W*	*7.83*	*7.83*	*0.006*	*S*

Indicates overlapping clusters detected in R and SaTScan.

SaTScan Radius Limit Comparison

Table 4.6.4 shows that when imposing a 90km limit on the maximum cluster size detected by SaTScan, the geographical location of the primary cluster either remained the same or had a very strong overlap with the previously identified cluster when there was evidence of clustering present in the data. Limiting the size of the cluster did affect the p value on 2 occasions (2007 and 2008), changing the result for that year from apparent clustering to no apparent clustering. In 2009 the introduction of the 90km limit did not change the cluster location or size, but did marginally change the p value.

Table 4.6.4.Trisomy 13 comparison of SaTScan results using a 90km radius limit or no geographical limit.

		Centroid Coordinates					
Year	Cases In Cluster	Latitude	Longitude	Radius	P-Value	Limit	Overlap?
2004	27	50.820 N	0.424 W	64.0	0.061	None	Yes
	27	50.820 N	0.424 W	64.0	0.059	90km	
2005	76	50.749 N	1.880 W	153.3	0.397	None	No
	14	51.009 N	3.115 W	62.9	0.58	90km	
2006	50	51.509 N	0.477 W	25.9	<0.0001	None	Yes
	50	51.509 N	0.477 W	25.9	<0.0001	90km	
2007	119	51.549 N	1.731 W	113.7	0.0021	None	Yes
	24	51.904 N	2.097 W	59.8	0.096	90km	
2008	94	50.704 N	1.306 W	129.2	0.007	None	Yes
	72	51.206 N	0.797 W	63.0	0.039	90km	
2009	16	51.461 N	2.592 W	39.2	0.0084	None	Yes
	16	51.461 N	2.592 W	39.2	0.0062	90km	
2010	132	51.908 N	0.443 W	114.2	0.0033	None	Yes
	27	51.492 N	0.137 W	7.8	0.006	90km	

All Data from 2004 - 2010

R

Cases		Centroid Coordinates		Radius (km)		
Total	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**
1382	219	51.587 N	0.270 W	22.2	14	0.001

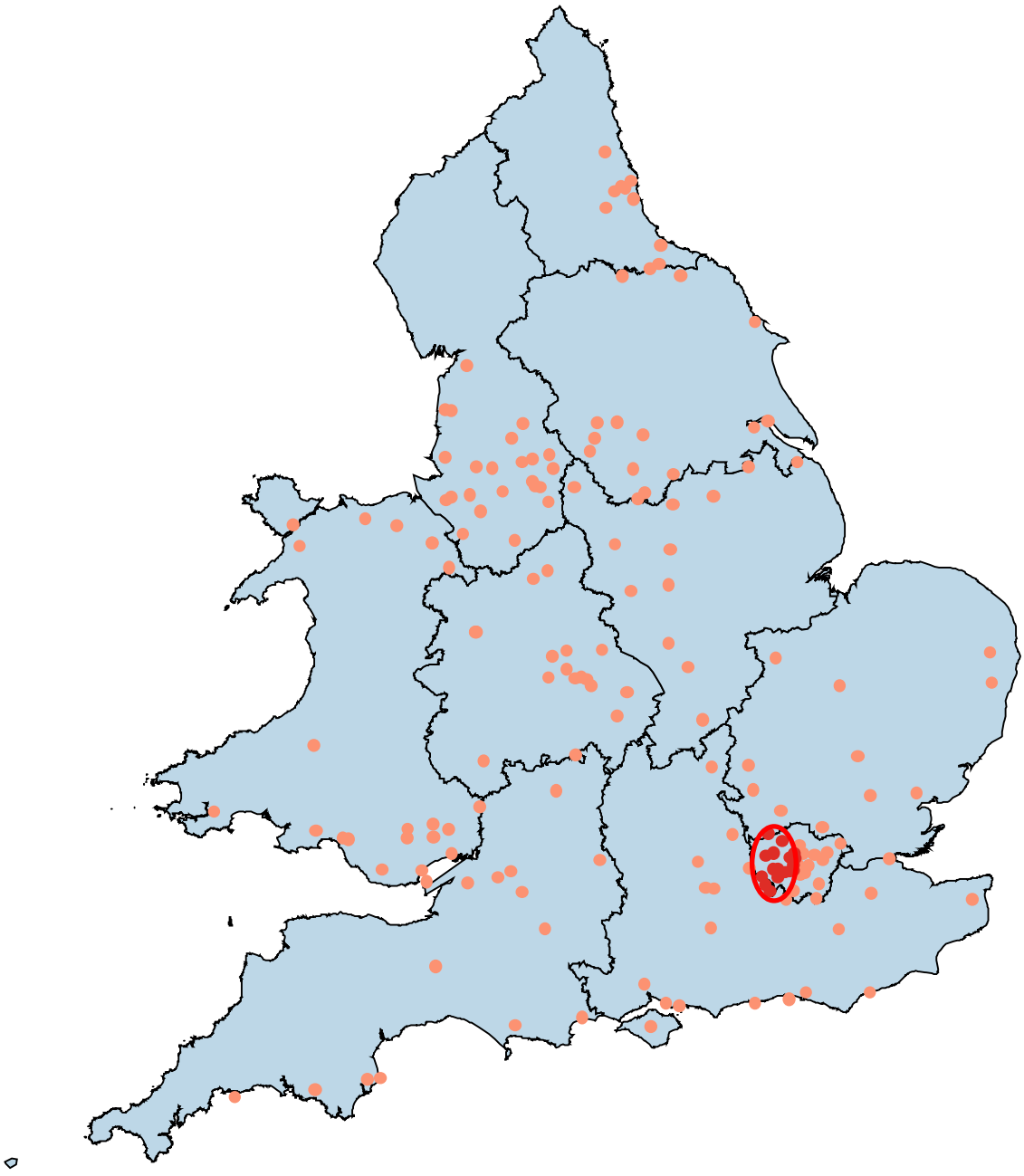

Figure 4.7. Map of trisomy 13 cluster by PCT using all data from 2004-2010 in R.

SaTScan

Cases		Centroid Coordinates			
Total	**In Cluster**	**Latitude**	**Longitude**	**Radius (km)**	**P Value**
1382	334	51.408 N	0.843 W	52.87	5.5E-11

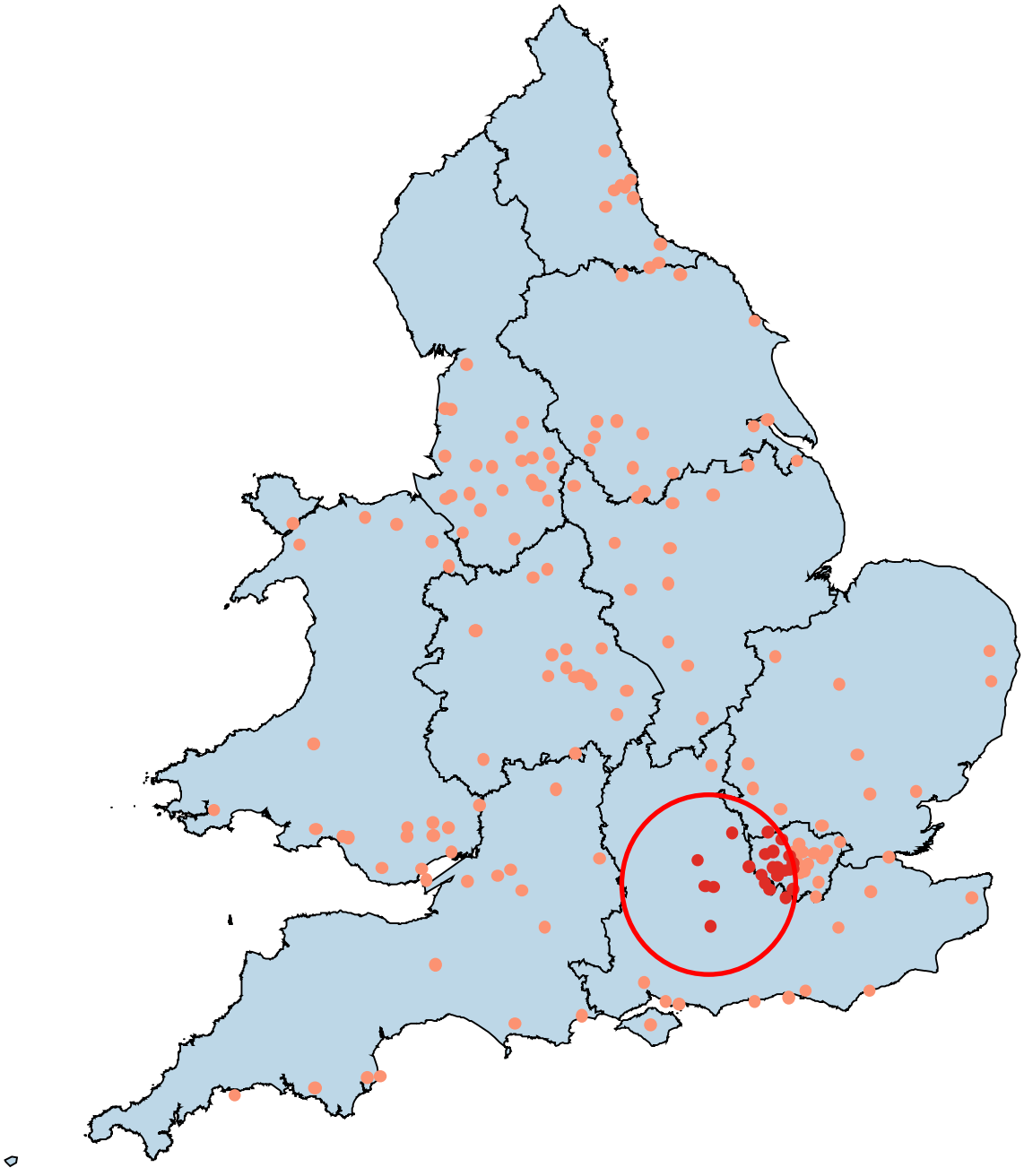

Figure 4.8. Map of trisomy 13 cluster by PCT using all data from 2004-2010 in SaTScan.

Figures 4.7 and 4.8 show that R and SaTScan both detect very strong evidence of clustering in London when analysing trisomy 13 data from 2004 to 2010. The cluster reported by R is

smaller than that of SaTScan and the centroid is at a different, although close, PCT site. The SaTScan cluster includes the entire R cluster and some additional PCTs to the west of London.

Section 4.7.Summary

R and SaTScan detect the same clusters when significant clusters are present. This was assessed using synthetic data, which showed that both methods could detect the same clusters and returned comparable p values in a simple scenario with one cluster present in a study region of otherwise uniform prevalence. There were small differences in p-values between the programs, which are most likely caused by differences in circle size and shape having a small effect during Monte Carlo simulation.

Use of synthetic data demonstrated that both methods can detect clusters with a relative risk of between 1.8 and 3 compared to the remainder of the study region. The required relative risk changes depending on the prevalence of the condition; as the disease prevalence increases, the required relative risk decreases. Given the observed prevalence of trisomies 18 and 13, a relative risk of between 2 and 3 would be required for a cluster to be deemed significant. Required relative risk may also depend on the number and size of PCTs within the cluster. Larger clusters will be easier to detect and will require a smaller relative risk to be significant, while large PCTs (i.e. PCTs with a large number of births compared to other PCTs) are also likely to be easier for the methods to detect.

An isolated cluster was chosen for the synthetic data testing in order to make comparisons between results simpler. If a cluster had been chosen in a more densely populated region then the strongest cluster detected by each method may include other PCTs. The different shapes (cluster or ellipse) may favour one method over the other, depending on the position of other PCTs. Using an isolated cluster favours neither shape. However the results are likely to be generalizable to a cluster in another area.

The primary purpose of chapter 4 was to establish whether there were spatial clusters of trisomies 18 and 13 in England and Wales from 2004 to 2010. Significant evidence of clustering was detected for both trisomies using both R and SaTScan, the majority of which was in London.

In this chapter, the method detailed by Kulldorff and Nagarwalla was written into a custom program in R and compared with SaTScan, a widely used free piece of software for the detection of spatial clusters. The main weakness of both of these programs is that each is optimised to search for clusters of a predetermined shape. SaTScan is best suited to detecting circular clusters, while R is best suited to detecting elliptical clusters. Both programs lose power when searching for clusters which are not of their optimal shape (Duczmal & Assuncao, 2004). It is possible that clusters of alternate shape (such as long, thin clusters), or clusters of irregular shape, could be present but cannot be detected using either program.

To conclude, it appears likely that the observed clusters in London were caused by advanced mean maternal age in London compared to other parts of England and Wales.

Chapter 5: Adjusted Spatial Cluster Detection

Section 5.1.Introduction

In chapter 4, evidence for clustering in both trisomies 18 and 13 was presented. The majority of clusters were located in London. Table 5.1 shows that in 2008, the highest proportion of births at the high risk maternal age groups (35-39 and 40-44) was observed in London compared to other government office regions in England and Wales (Office for National Statistics Birth Statistics 2008 Report). The next highest figure for both age groups was observed in the South East region. It would therefore be expected that if any age related clusters were present in the data, they would be in London.

Table 5.1.Proportion of births by government office region and maternal age group in 2008.

Government Office Region	15-19	20-24	25-29	30-34	35-39	40-44
North East	9.1	20.8	29.2	26.9	11.9	2.2
North West	7.4	20.2	29.4	27.5	12.9	2.5
Yorkshire and the Humber	8.0	20.2	28.4	28.3	12.6	2.4
East Midlands	6.9	19.5	29.8	28.3	12.8	2.7
West Midlands	7.3	21.2	29.0	27.2	12.6	2.6
East of England	5.2	18.3	29.1	29.8	14.7	2.9
London	5.9	17.5	22.8	28.2	20.0	5.5
South East	5.1	16.6	27.0	31.3	16.4	3.5
South West	5.7	17.2	28.9	30.1	15.0	3.1
Wales	8.0	20.5	29.3	27.2	12.7	2.4

It is also possible that the appearance of clusters could be caused by regional differences in gestational age at diagnosis. Fetal loss rates are very high in trisomy 18 and trisomy 13 pregnancies, and areas which routinely screen from 10 weeks gestation will diagnose more cases than those which screen later. The observed trisomy prevalence will therefore be higher in areas which screen early compared to those which screen late. This difference may be sufficient to produce (or at least contribute to) clusters. Prenatal screening from 10 weeks gestation is currently offered in all areas of England and Wales, but at the time of data collection many areas did not routinely screen at such early gestations,

In order to detect age-independent clusters, a novel method was developed to weight cases individually based on their maternal age and gestational age at diagnosis. Each case was assigned a weight which was then applied to the total number of cases in each PCT; so if a PCT contained one case with weight 0.5 and one case with weight 1, there would be a total of 1.5 cases in that PCT. Weights were retained for each case in Monte Carlo simulation.

Methods to account for maternal age and gestational age at diagnosis were required as both could be expected to lead to the formation of clusters. One significant advantage of weighting cases individually is that for each case, more than 1 weight can be applied at once by assigning and multiplying weights, to give an overall weight for each case. This theoretically allows a direct case weighting approach to account for as many factors as required.

Section 5.2.Methods

Section 5.2.1.Adjusting in R

Gestational Age

All terminations under 20 weeks of gestation at diagnosis were weighted based on their likelihood of surviving to 20 weeks based on gestational age-specific fetal loss rates for trisomies 18 and 13 in England and Wales (Savva and Morris, 2008) (See table 5.2.1).

Table 5.2.1.Gestational age-specific weights.

	Weight	
Gestational Age at Diagnosis (Weeks)	**Trisomy 18**	**Trisomy 13**
< 14	0.754	0.844
14	0.792	0.869
15	0.831	0.894
16	0.864	0.922
17	0.921	0.941
18	0.941	0.962
19	0.971	0.983
> 19	1	1

Maternal Age

Maternal age weights were calculated using previously published maternal age-specific trisomy 18 and 13 prevalence figures (Savva*et al*, 2010). The median population maternal age of England and Wales is 29.5, and cases born to women aged 29 were assigned a weight of 1. All other maternal ages were weighted according to the formula below.

For maternal age *x:* Weight = $\frac{Prevalence\ at\ age\ 29}{Prevalence\ at\ age\ x}$

Table 5.2.2.Maternal age-specific weights.

Maternal Age	T13 Prevalence	T13 Weight	T18 Prevalence	T18 Weight
< 24	0.7	1.2857	1.1	1.1818
24	0.8	1.1250	1.1	1.1818
25	0.8	1.1250	1.2	1.0833
26	0.8	1.1250	1.2	1.0833
27	0.8	1.1250	1.2	1.0833
28	0.8	1.1250	1.2	1.0833
29	0.9	1.0000	1.3	1.0000
30	0.9	1.0000	1.4	0.9286
31	1.0	0.9000	1.5	0.8667
32	1.1	0.8182	1.7	0.7647
33	1.3	0.6923	1.9	0.6842
34	1.6	0.5625	2.3	0.5652
35	1.9	0.4737	2.8	0.4643
36	2.5	0.3600	3.7	0.3514
37	3.2	0.2813	4.9	0.2653
38	4.2	0.2143	6.0	0.2167
39	5.5	0.1636	9.7	0.1340
40	7.0	0.1286	13.6	0.0956
41	8.6	0.1047	18.8	0.0691
42	10.3	0.0874	25.1	0.0518
43	11.9	0.0756	32.2	0.0404
44	13.4	0.0672	39.6	0.0328
45	14.6	0.0616	46.8	0.0278
46	15.6	0.0577	53.3	0.0244
47	16.4	0.0549	58.9	0.0221
48	17.0	0.0529	63.5	0.0205
49	17.4	0.0517	67.2	0.0193

Section 5.2.2.Adjusting in SaTScan

It is not possible to enter non-integer totals of cases in SaTScan, so another method of weighting cases had to be used. While it was possible to apply the weights and round the total for each PCT to give an integer value, this would not have been very inaccurate due to the small case numbers present in many PCTs and likely errors when reassigning cases for Monte Carlo simulation.

There is a method in SaTScan to adjust for known relative risks by using an adjustment file. From the SaTScan user guide:

"Sometimes it is known a priori that a particular location and/or time has a higher or lower risk of known magnitude, and we want to detect clusters above and beyond this, or in other words, we want to adjust for this known excess/lower risk. One way to do this is to simply change the population at risk numbers in the population file. A simpler way is to use the adjustments file. In this file, a relative risk is specified for any location and time period combination. The expected counts are then multiplied by this relative risk for that location and time. For example, if it is known from historical data that a particular location typically have 50 percent more cases during the summer months June to August, then for each year one would specify a relative risk of 1.5 for this location and these months. A summer cluster will then only appear in this location if the excess risk is more than 50 percent."

Using the mean of the case weights, specific risks can be calculated for each PCT and applied in the adjustments file. For example, a PCT has 3 cases, all of which have older mothers and have subsequently been assigned low weights. The resulting adjustment in SaTScan should be greater than 1 as more cases are expected in that PCT due to high maternal age. The case weights are 0.25, 0.5 and 0.75, with a mean case weight of 0.5. In SaTScan the adjustment is applied to the expected number of cases rather than the observed number, so the inverse of the mean case weight must be taken. In this example the SaTScan adjustment is 2.

There are three major differences between this weighting method and the direct case weighting method used in R which may affect the results. Firstly, the adjustment is being applied to the expected number of cases in a PCT rather than the observed number. Secondly, weights are static and cannot be reassigned when calculating p-values. All cases are reassigned without weighting, but a PCT will have the same relative risk for the initial scan and then for each iteration of the Monte Carlo simulation. Finally, any PCT which has no cases present in the scan cannot be adjusted and must instead be assigned a default risk of 1. This is not a problem when scanning all cases from 2004-2010 as there are very few PCTs with no cases, but for single year scans there are a high proportion of PCTs with no cases in the scan. For example, in 2004 there were 91 PCTs (over 50%) without any trisomy 13 cases. This means that

when calculating p-values, some weighted PCTs may be assigned no cases and therefore the weighting for that PCT will be lost. Applying case weights directly always retains full weighting information in Monte Carlo simulation.

Section 5.3.Test Data

Section 5.3.1. Accounting for Simulated Clusters in Synthetic Data

Methods

In chapter 4, it was demonstrated that R and SaTScan obtained similar results when calculating cluster significance by using an idealised test data set with uniform disease prevalence apart from a cluster of 12 PCTs in the North East.

A similar method can be used to test whether the programs return expected results when accounting for a previously detected cluster. Starting with a highly significant cluster, cases within the cluster were weighted down gradually until the total number of weighted cases in each PCT within the cluster was equal to the total number of cases in all other PCTs. For example, starting with 1 case per PCT outside the cluster and 4 cases per PCT within the cluster, cases within the cluster were weighted to give a total of 3 cases per PCT, followed by 2 cases per PCT and finally 1 case per PCT. In SaTScan the corresponding adjustment was calculated and applied. As the number of cases per PCT outside the cluster increased, so did the number inside the cluster. When there were 2 cases per PCT outside the cluster, the initial number within the cluster was set as 5. Similarly, when there were 4 and 6 cases per PCT outside the cluster the initial number of cases within was 8 and 11 respectively. These numbers were chosen to ensure that the unadjusted scan always returned a very strong cluster in order to assess the effect of applying the weights.

Results

Tables 5.3.1 to 5.3.4 show that R and SaTScan both successfully account for a single test cluster. Differences between p-values are mostly small although can exceed an order of magnitude, for example the relative risk of 2 in table 5.3.1. However, these differences can be attributed to the weights being applied in different ways and differences in Monte Carlo simulation which were discussed in chapter 4, which can have a large effect in conditions of low prevalence such as in the aforementioned tests. For comparison, the tests in table 5.3.4 produce much closer p-values when the prevalence is increased to 6 cases per PCT outside the cluster.

Table 5.3.1. Case adjustment from 4 to 1 cases per PCT within test cluster

SaTScan Analysis		R Analysis	
Adjustment	**P Value**	**P Value**	**Weighted Cases**
1.00	6.60E-10	0.001	4
1.33	0.000004	0.001	3
2.00	0.026	0.304	2
4.00	1.000	1.000	1

Table 5.3.2. Case adjustment from 5 to 2 cases per PCT within test cluster

SaTScan Analysis		R Analysis	
Adjustment	**P Value**	**P Value**	**Weighted Cases**
1.00	0.000088	0.001	5
1.25	0.00078	0.006	4
1.67	0.800	0.942	3
2.50	1	1	2

Table 5.3.3. Case adjustment from 8 to 4 cases per PCT within test cluster

SaTScan Analysis		R Analysis	
Adjustment	**P Value**	**P Value**	**Weighted Cases**
1	0.0000054	0.001	8
1.14	0.00083	0.001	7
1.33	0.115	0.214	6
1.6	1	1	5
2	1	1	4

Table 5.3.4. Case adjustment from 11 to 6 cases per PCT within test cluster

SaTScan Analysis		R Analysis	
Adjustment	**P Value**	**P Value**	**Weighted Cases**
1.00	0.0000014	0.001	11
1.10	0.0061	0.001	10
1.22	0.011	0.030	9
1.38	0.640	0.726	8
1.57	1.000	1.000	7
1.83	1.000	1.000	6

Section 5.3.2.Adjustment using Randomly Allocated Weights

Methods

While there were only small differences between p-values when accounting for a single cluster, it is possible that the different ways in which weights are applied could create differences between the results in R and SaTScan. In order to investigate whether both methods give expected results, a set of data was generated which contained a randomly allocated set of weights. As all weights were random, no clusters would be expected in this data.

Weights were allocated to PCTs at random in a uniform population of 4 cases and 4,000 births per PCT. For the purpose of this test, all cases in a given PCT were assigned the same weight. Weights were calculated using a random number generator in R between 1 and a set maximum value of 10, 20 or 40, to give a minimum case weight of 0.1, 0.05 or 0.025 respectively. These figures were selected in order to determine how well each program could handle more extreme weighting methods, such as that used to account for maternal age. In order to reduce the risk of obtaining anomalous results, 5 sets of random weights were calculated for each weighting range.

Results

Tables 5.3.5 and 5.3.6 show the results obtained in R and SaTScan when weights between 0.1 and 1 were applied to PCTs at random. Table 5.4.1 shows that in R, no clusters were detected and for all tests the most likely cluster had a p-value of over 0.5. Table 5.4.2 shows that in SaTScan, much stronger clusters were detected with 4 of the 5 p-values falling below 0.1. Three p-values fell below 0.05 and would be considered significant at the 5% level. This shows a clear difference between R and SaTScan, and given that the randomly assigned weights would not be expected to lead to clustering, implies that the weighting method in SaTScan is not working as intended.

Table 5.3.5. R, Relative Risk Range 1-10

Cases		Centroid		Radius (km)		
Total	In Cluster	Latitude	Longitude	Horizontal	Vertical	P Value
380.4	47.6	51.574 N	0.071 E	14	22.2	> 0.5
410.0	42.8	51.681 N	0.614 W	28	44.4	> 0.5
388.0	11.6	52.521 N	2.012 W	7	11.1	> 0.5
371.2	40.8	50.820 N	0.424 W	49	77.7	> 0.5
353.2	16.8	51.700 N	4.121 W	42	66.6	> 0.5

Table 5.3.6.SaTScan, Relative Risk Range 1-10

Cases		Centroid		Radius (km)		
Total	In Cluster	Latitude	Longitude	Horizontal	Vertical	P Value
696	12	53.544 N	2.128 W	12.37	12.37	0.021
696	8	52.265 N	0.870 W	27.51	27.51	0.083
696	16	53.158 N	1.605 W	30.93	30.93	0.032
696	24	52.580 N	0.249 W	62.55	62.55	0.037
696	20	51.700 N	4.121 W	48.51	48.51	0.266

Tables 5.3.7 and 5.3.8 show the results obtained in R and SaTScan when weights between 0.05 and 1 were applied to PCTs at random. No clusters were detected in R, while SaTScan detected significant clusters ($p < 0.05$) in 4 of the 5 tests. In 2 of these tests, the p-value was below 0.001.

Table 5.3.7. R, Relative Risk Range 1-20

Cases		Centroid		Radius (km)		
Total	In Cluster	Latitude	Longitude	Horizontal	Vertical	P Value
370.6	76.8	52.265 N	0.870 W	70	111.0	> 0.5
368.6	26.6	53.424 N	1.351 W	35	55.5	> 0.5
354.4	16.2	51.587 N	0.270 W	7	11.1	> 0.5
389.4	14.6	51.348 N	1.400 E	63	99.9	> 0.5
345.6	30.6	54.983 N	1.552 W	42	66.6	> 0.5

Table 5.3.8.SaTScan, Relative Risk Range 1-20

Cases		Centroid		Radius (km)		
Total	In Cluster	Latitude	Longitude	Horizontal	Vertical	P Value
696	8	51.196 N	0.278 E	22.01	22.01	0.016
696	4	53.578 N	2.389 W	0	0	0.151
696	8	53.601 N	3.034 W	18.11	18.11	0.00089
696	12	51.681 N	0.614 W	22.73	22.73	0.0051
696	12	51.495 N	3.228 W	20.18	20.18	0.00028

Tables 5.3.9 and 5.3.10 show the results obtained in R and SaTScan when weights between 0.025 and 1 were applied to PCTs at random. Again, no clusters were detected in R while SaTScan detected significant clusters ($p < 0.05$) in all 5 tests. In 4 of these tests, the p-value was less than or equal to 0.001.

Table 5.3.9. R, Relative Risk Range 1-40

Cases		Centroid		Radius (km)		
Total	In Cluster	Latitude	Longitude	Horizontal	Vertical	P Value
361.6	20.5	53.255 N	3.437 W	35	55.5	> 0.5
362.5	36.0	53.556 N	0.484 W	63	99.9	> 0.5
365.2	11.1	52.056 N	2.709 W	35	55.5	> 0.5
355.9	28.1	54.919 N	1.447 W	28	44.4	> 0.5
358.1	20.3	53.781 N	1.589 W	21	33.3	> 0.5

Table 5.3.10.SaTScan, Relative Risk Range 1-40

Cases		Centroid		Radius (km)		
Total	In Cluster	Latitude	Longitude	Horizontal	Vertical	P Value
696	8	52.485 N	1.884 W	3.02	3.02	0.00011
696	12	53.754 N	0.435 W	22.27	22.27	0.0014
696	8	53.514 N	1.115 W	16.88	16.88	0.0009
696	8	51.584 N	2.974 W	13.64	13.64	0.001
696	20	53.701 N	1.777 W	26.76	26.76	0.000032

Tables 5.3.6, 5.3.8 and 5.3.10 all show that SaTScan detects significant clusters in a scenario in whichthere would not be expected to be any clusters present. More significant clusters are detected as the weights become more extreme. When weights between 0.1 and 1 were applied the mean and median p-values were 0.088 and 0.037 respectively. These figures compare to a mean of 0.035 and median of 0.0051 when weights between 0.05 and 1 were applied, and a mean and median of 0.0007 and 0.0009 when using weights between 0.025 and 1.

Assuming that there genuinely would not be expected to be clusters present in randomly generated data of this type, there are 2 possible explanations for these results. Either the weights do not work in SaTScan, or this application of the weighting does not.The results shown in section 5.3.1 indicate that this is an effective weighting method in a highly simplified scenario. However it is likely that complications could arise when scanning more complex data,such as that generated in this simulation.

Discrepancies in the results generated in R and SaTScan were greatest when the weighting structure being simulated was the most extreme. When simulating a more mild weighting method (relative risks between 1 and 10), the difference in p values between R and SaTScan was smaller than when the weighting method generated relative risks between 1 and 40. The most apparent explanation of these results is that weighting in SaTScan is unreliable when using an extreme weighting method. Were this the case, it would create a problem when applying the maternal age weighting in SaTScan for both trisomy 18 (relative weight range 1 – 61.2) and trisomy 13 (relative weight range 1 – 24.9).

Section 5.3.3.Adjustment using Different Case Distributions

Methods

One advantage of direct case weighting in R compared to PCT adjustment in SaTScan is that R takes into account the exact case weights during Monte Carlo simulation.

Weights remain static at each PCT in SaTScan while in R weighted cases are randomly reassigned during each iteration of the Monte Carlo simulation. This means that in SaTScan only the mean weight per PCT is used, whereas in R the weighting structure could also affect results even if the mean weight per PCT does not change.

When weights are used, those cases with the largest weights contribute the most to potential clusters. So, for example, a case with weight 2 is equal to 2 cases with weight 1. The more extreme the weighting method, the greater the difference becomes between the greatest and least weighted cases, and the most highly weighted cases have an even greater contribution to clusters. It may be that having very highly weighted cases in the data increases the chance of observing clusters after data has been reassigned during Monte Carlo simulation, as having only small number of highly weighted cases in a region could be sufficient to create a cluster. This effect would only be visible in R.

A synthetic set of data was generated to test this effect. For the sake of convenience, this data had a uniform prevalence of 4 cases and 4,000 births per PCT, with the exception of one cluster (the standard test cluster region consisting of 12 PCTs) in which there were 8 cases and 4,000 births. In the test cluster, each PCT was assigned a mean weight of 0.25 to represent the region being populated by older mothers. The remaining PCTs were assigned a mean weight of 0.625 to represent these regions being of average maternal age. However, one region of 51 PCTs in and around West London (region chosen for convenience) was simulated to be populated by younger mothers, and all PCTs within this region were assigned a weight of 1. The expectation in this simulation is that the 12 PCTs with 8 cases per PCT would appear as a cluster in the absence of weighting, however once the weights were applied this cluster would be accounted for. Meanwhile the region of younger mothers, despite having average disease prevalence, would appear as a cluster due to the relatively high weight assigned to PCTs within this region. A cluster of this type can be termed an emerging cluster. This scenario could potentially appear in the NDSCR data, but has been simplified for the purpose of this simulation.

In R, a series of tests were performed by changing the case structure in each region using 2 different case weights in each region, so for example in PCT *x*, rather than weighting all cases 1, half would be weighted 1.5 and the other half 0.5. This would have no effect on the mean

weight so the SaTScan p-value would remain the same, but in R this change could lead to a change in the calculated p-value. This is another simplification of an event which will appear in the NDSCR data, as PCTs with multiple cases are likely to contain multiple weights.

Results

Table 5.3.11 shows that changing the case weights in R could change the p-value by up to approximately 3 orders of magnitude compared to SaTScan. However, the weights have to changed dramatically in all 3 regions to observe such a change. Tests 2, 3 and 4 all simulated having a half and half mixture of younger and older women in region 1.

Table 5.3.11. R vs SaTScan, Cluster Emerging as the Result of Weighting

	Weighting Structure			P Value	
Test	**Region 1**	**Region 2**	**Region 3**	**R**	**SaTScan**
1	1	0.625	0.25	0.001	0.000022
2	1.5, 0.5	1, 0.25	0.25	0.001	0.000022
3	1.75, 0.25	1, 0.25	0.25	0.001	0.000022
4	1.98, 0.02	0.625	0.25	0.001	0.000022
5	1.98, 0.02	1.23, 0.02	0.48, 0.02	0.017	0.000022

Results in table 5.3.11 show that it is possible to obtain a p-value in R which is around 3 orders of magnitude higher than that in SaTScan, but this requires that a very extreme weighting method is used. In test 5, half of cases in each were assigned very low weights which the remaining half were assigned a much higher weight which varied depending on which region the PCT was in. The weighting ratio for each region in test 5 was 99:1 in region 1, 123:2 in region 2 and 24:1 in region 3.

Section 5.3.4.Summary of Test Data Results

Previously in section 5.3, 3 tests were performed using synthetic data. Firstly in section 5.3.1, a single cluster was successfully accounted for in R and SaTScan by assigning low weights to each case / PCT within the cluster. In section 5.3.2, weights were applied at random to assess the effect of using weighting scales of varying severity. Finally in section 5.3.3, the effect of using different weights in R was tested by altering the weights within each PCT while keeping the mean weight in each PCT the same.

Results varied between R and SaTScan to varying degrees in all of these tests. When accounting for a single cluster, small differences were observed which can be attributed to the different ways in which weights are applied in R and SaTScan. However, in the other 2 tests there were greater differences between the methods. In the random weights tests, significant clusters were observed in SaTScan while the p value in R never dropped below 0.5. Clusters in SaTScan became more significant as the weights used became more severe. In the third test, changing the weighting structure in R did have an effect on the resulting p value, but only when very extreme changes had been made. When smaller changes were made, no difference was observed.

Having observed some significant differences between R and SaTScan in these simple tests, it seems likely that when scanning NDSCR data similar differences may be present. As the NDSCR data is much more complex, with the potential for a different trisomy prevalence in each PCT and tens of possible weights for each trisomy, it is also possible that these differences may be much greater when scanning NDSCR data.

Section 5.4. Analysis

Section 5.4.1. Cases excluded from the analysis

Calculating weights for cases requires additional information compared to unadjusted scans. In this chapter maternal age and gestational age at diagnosis are being accounted for individually and simultaneously. To allow unbiased comparison between results and to compare the effect of the weighting methods, cases which were missing either maternal age or gestational age at diagnosis were excluded from all weighted scans. The full data cleaning process was detailed in chapter 2.

Table 5.3.1 shows the number and percentage of cases which were excluded from the weighted spatial scans due to missing information. 5.3% of trisomy 18 cases and 6.2% of trisomy 13 cases were missing PCT and could not be included in all spatial scans, including the unadjusted scans detailed in chapter 4. 2.6% of trisomy 18 cases and 4.1% of trisomy 13 cases were excluded from weighted scans due to missing maternal age, in addition to 3.0% of trisomy 18 cases and 4.5% of trisomy 13 cases which were excluded due to missing gestational age. In total, 10.9% of trisomy 18 cases and 14.8% of trisomy 13 cases were excluded from weighted scans due to missing information, compared to 5.3% and 6.2% for trisomies 18 and 13 respectively from the unadjusted scans.

A further 13.9% of trisomy 18 cases and 14.5% of trisomy 13 cases were excluded as they were classed as fetal losses under 20 weeks of age at diagnosis. The intention of weighting for gestational age is to eliminate the effect of regional differences in prenatal screening methods which may create or contribute to clusters. It is assumed that all cases in England and Wales which survive to 20 weeks are detected and diagnosed, as any fetal losses from this age onwards are rare and would be genetically karyotyped. Fetal losses under the age of 20 weeks may be detected in some regions but not others. For the purpose of this weighting method, all fetal losses under 20 weeks had to be excluded.

Table 5.4.1. Cases Excluded from Weighted Spatial Scans

	Trisomy 18		Trisomy 13	
	Number	**%**	**Number**	**%**
Total Cases	**3507**	**100**	**1473**	**100**
Excluded - No PCT	186	5.3	91	6.2
Excluded - No Maternal Age	91	2.6	61	4.1
Excluded - No Gestational Age	105	3.0	67	4.5
Excluded - Fetal Loss Under 20 Wks	488	13.9	213	14.5
Total Cases Excluded	870	24.8	432	29.3
Total Cases Remaining	**2637**	**75.2**	**1041**	**70.7**

In total, 24.8% of trisomy 18 cases and 29.3% of trisomy 13 cases were excluded from all weighted scans. These are very high percentages of cases which cannot be included, and demonstrate what could be considered to be a major weakness of utilising weighting methods in cluster scanning. In the unadjusted scans, a total of 94.7% and 93.8% of cases of trisomies 18 and 13 could be included. When using both weighting methods, only 75.2% and 70.7% can be included respectively.

Section 5.4.2. Distribution of Weights

In chapter 4.6.1, the distribution of cases for trisomies 18 and 13 was displayed on PCT maps of England and Wales. It was thought that groups of high prevalence PCTs could indicate the location of potential clusters, and for both trisomies it was apparent that the prevalence in London was greater than elsewhere in England and Wales. The same analysis can be used to attempt to visualise whether there are regions of particularly high or low case weights.

The distribution of weights was mapped for each trisomy by calculating the mean maternal age, gestational age and both weights combined in each PCT. It would be expected that maternal age weights would be low in London as that is where maternal age is highest. Gestational age weights may also be low in London if gestational age contributed to the clusters reported in chapter 4. The combined weight should reflect the maternal age weight distribution as maternal age weight is the stronger of the two weights.

Maternal Age

Figure 5.1 shows the trisomy 18 maternal age distribution. The figure shows a cluster of PCTs in London with low mean weight, which may be sufficient to account for the very strong clusters reported in chapter 4. There are no other regions in England and Wales with such a uniformly low weight.

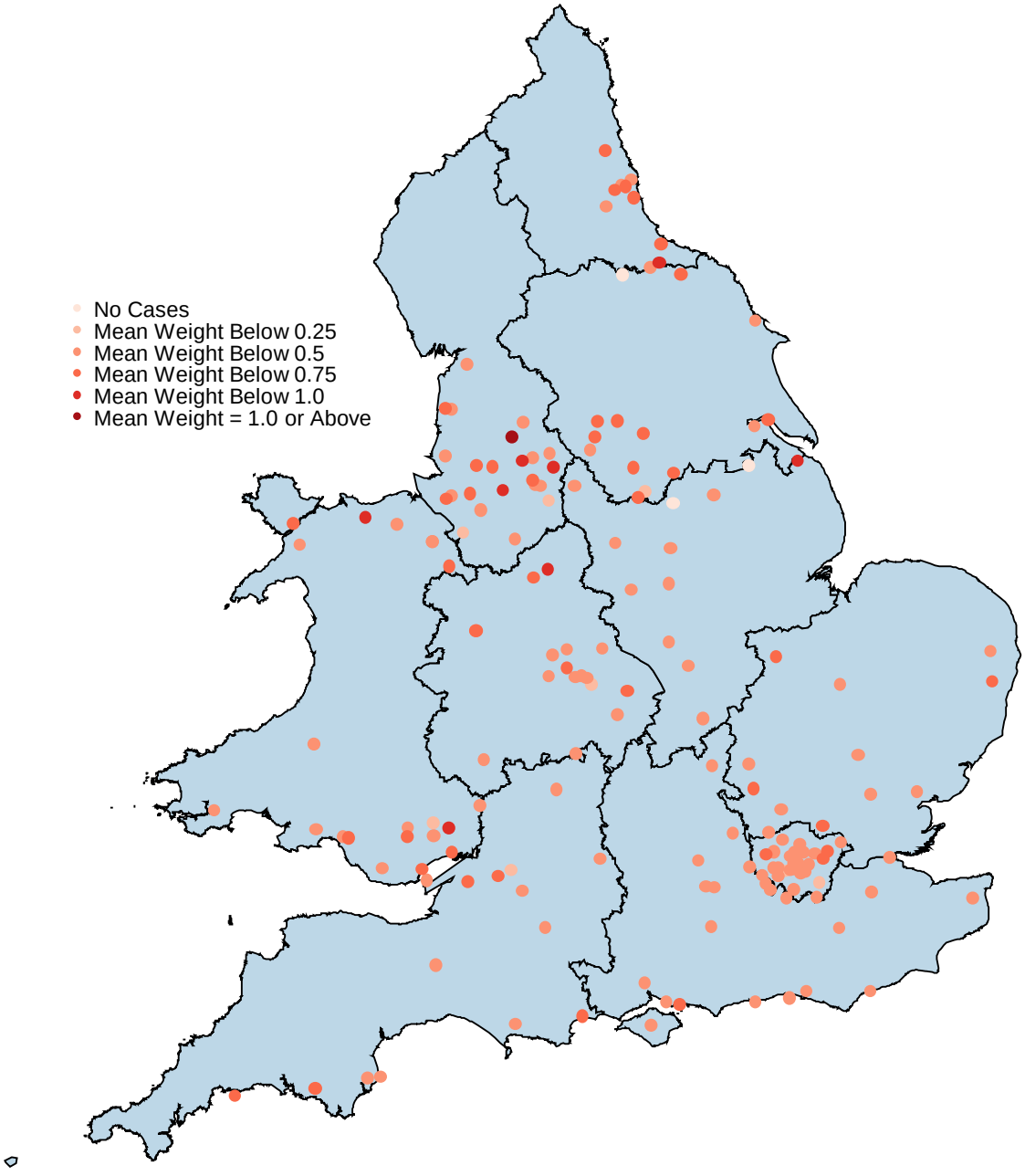

Figure 5.1. Trisomy 18 Mean Maternal Age Weight by PCT.

Figure 5.2 shows the trisomy 13 maternal age distribution. This figure shows that many PCTs in London have been assigned low mean weights, however PCTs are not all heavily weighted down as they were for trisomy 18. In London there are several coloured PCTs which have a

mean weight of above 0.75. This may be sufficient to account for the clusters reported in chapter 4 as trisomy 13 clusters were not as strong as those of trisomy 18.

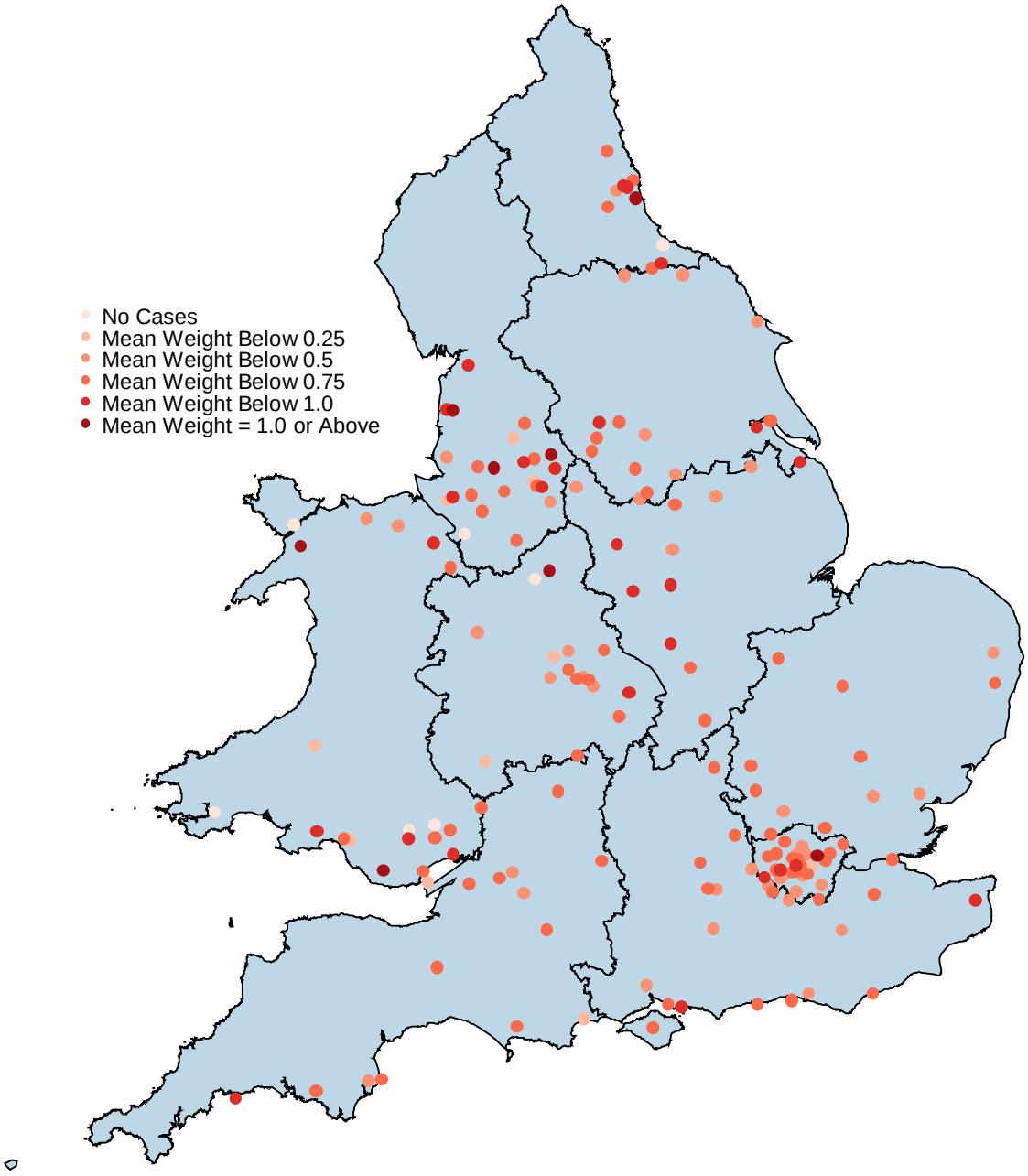

Figure 5.2. Trisomy 13 Mean Maternal Age Weight by PCT.

Gestational Age

Figure 5.3 shows the trisomy 18 gestational age distribution. The map appears to show that PCTs in London are mostly weighted down compared to the rest of England and Wales. London is the only large group of low-weighted PCTs. Although the gestational age weights are mild compared to those for maternal age, this may still account for some of the clustering observed in chapter 4.

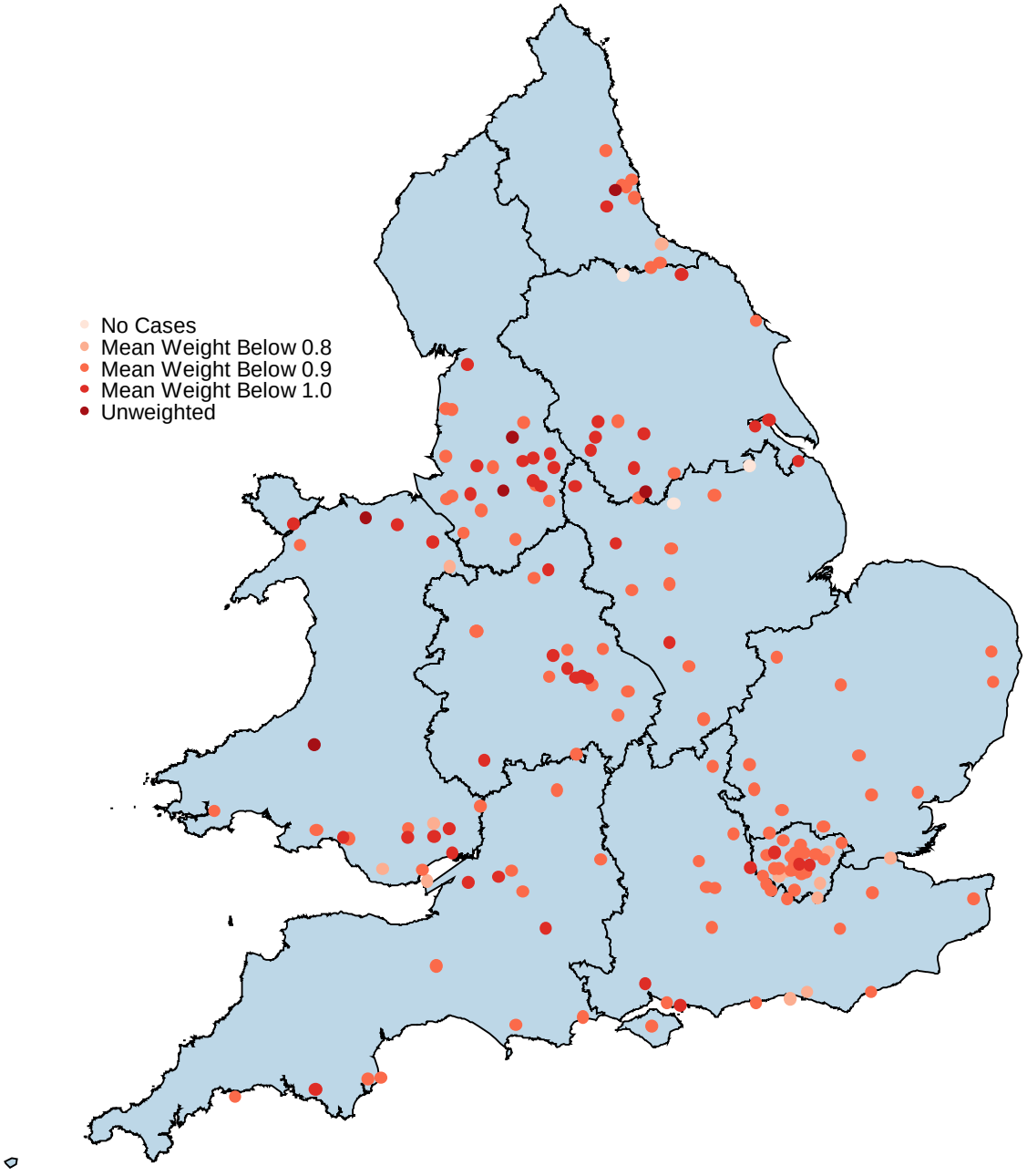

Figure 5.3. Trisomy 18 Mean Gestational Age Weight by PCT.

Figure 5.4 shows the trisomy 13 gestational age distribution. The majority of PCTs in England and Wales have been assigned weights between 0.9 and 1. There is not much of a pattern to the gestational age weights, although there is a higher concentration of PCTs with mean

weight below 0.9 in London than anywhere else in England and Wales. This may be sufficient to account for a small part of the clusters reported in chapter 4, however the gestational age weights for trisomy 13 are mild and unlikely to have a large effect on the results.

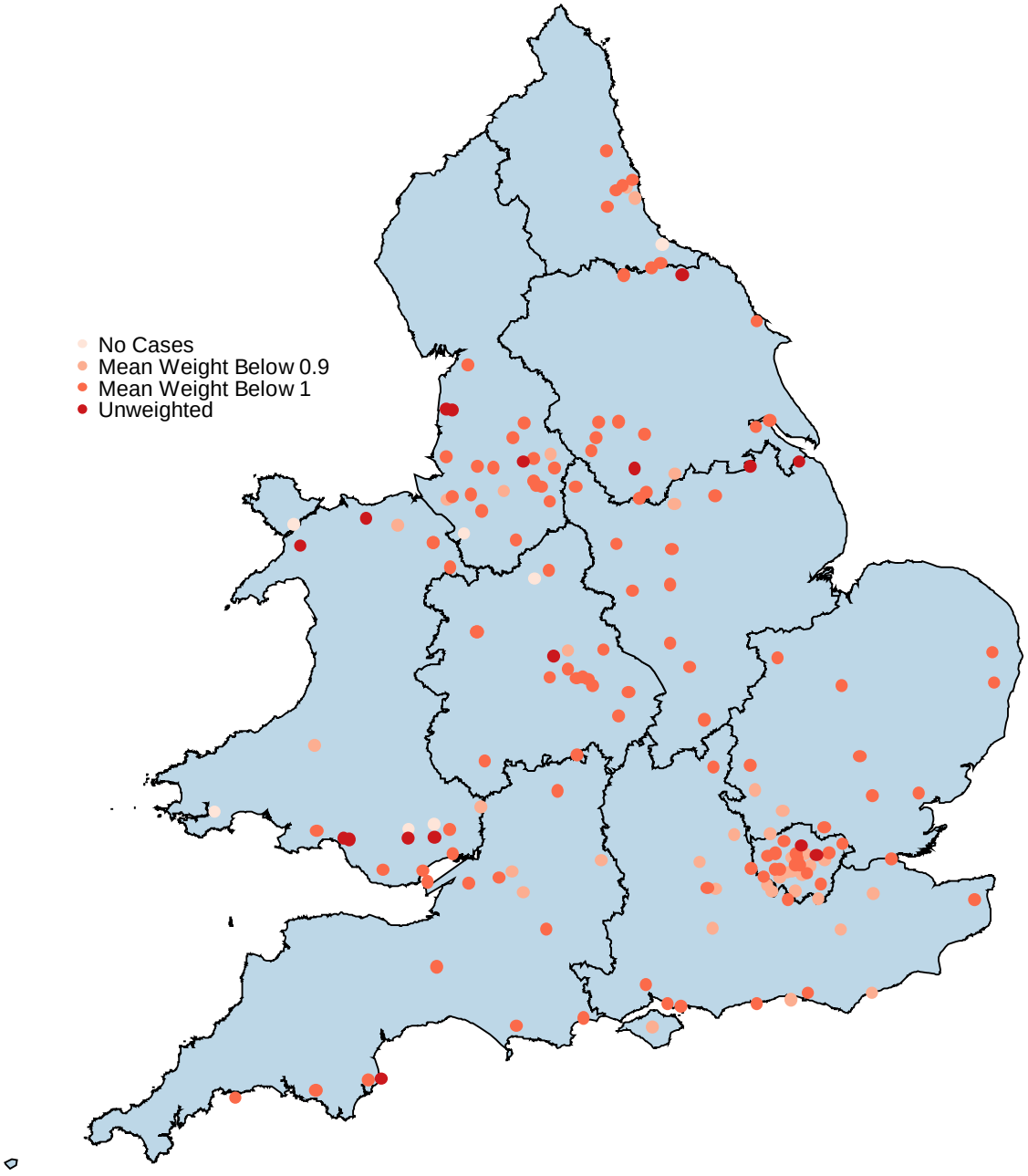

Figure 5.4. Trisomy 13 Mean Gestational Age Weight by PCT.

Both Weights Combined

Figure 5.5 shows the trisomy 18 combined weight distribution. When both weights are combined there is clearly a low weighted region in London compared to the rest of England and Wales. The weighting effect is stronger combined than for each weighted variable combined, and is therefore more likely to be sufficient to account for clusters reported in chapter 4. There are no other regions which show such consistently low weights, although many PCTs outside London were still heavily weighted down.

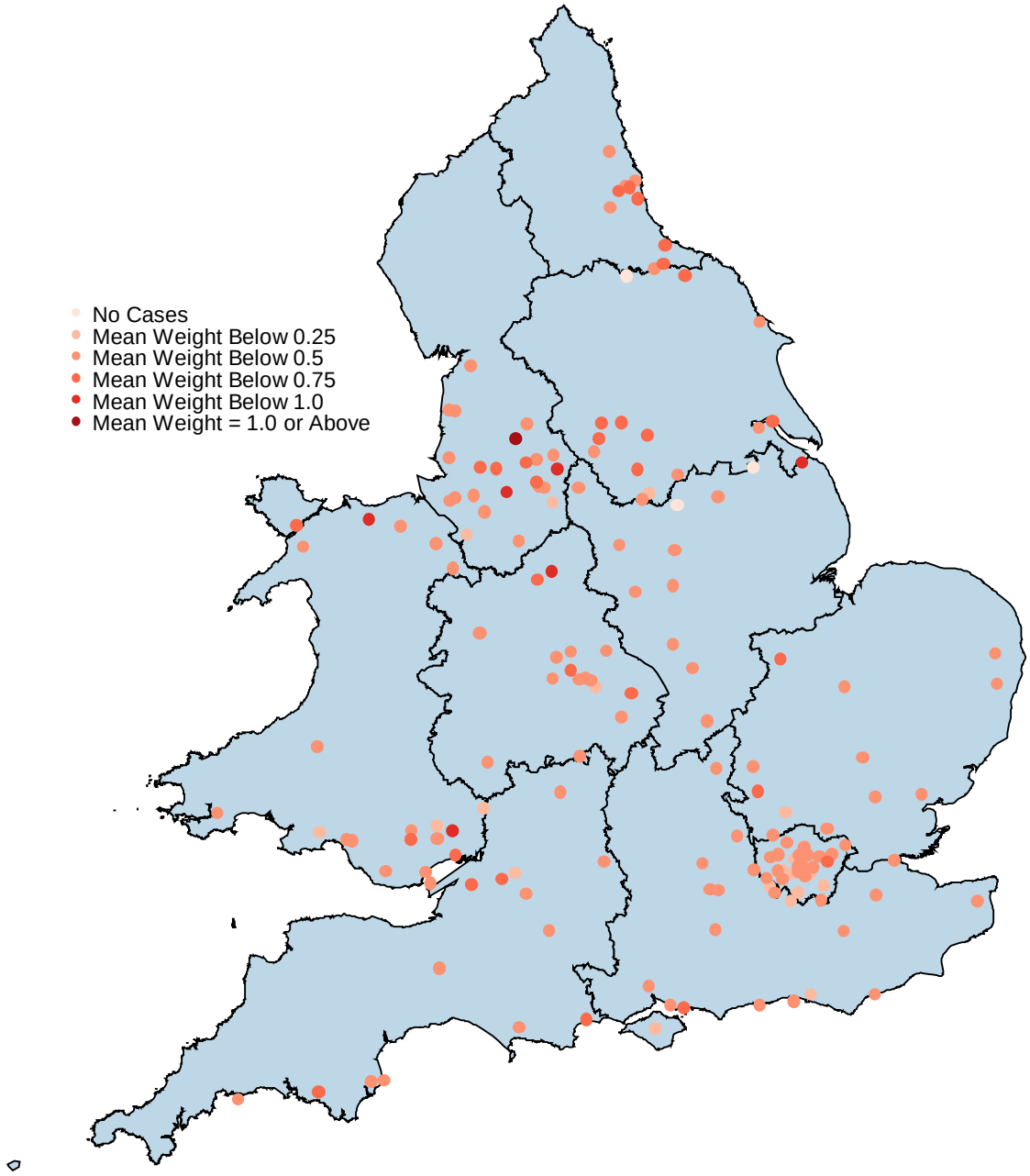

Figure 5.5. Trisomy 18 Combined Weight by PCT

Figure 5.6 shows the trisomy 13 combined weight distribution. There is a higher concentration of low-weighted PCTs in London than anywhere else in England and Wales, however this is less

apparent than it is for trisomy 18. Weights may still be sufficient to account for clusters as trisomy 13 clusters were weaker than those for trisomy 18, and therefore less adjustment is required to account for those clusters.

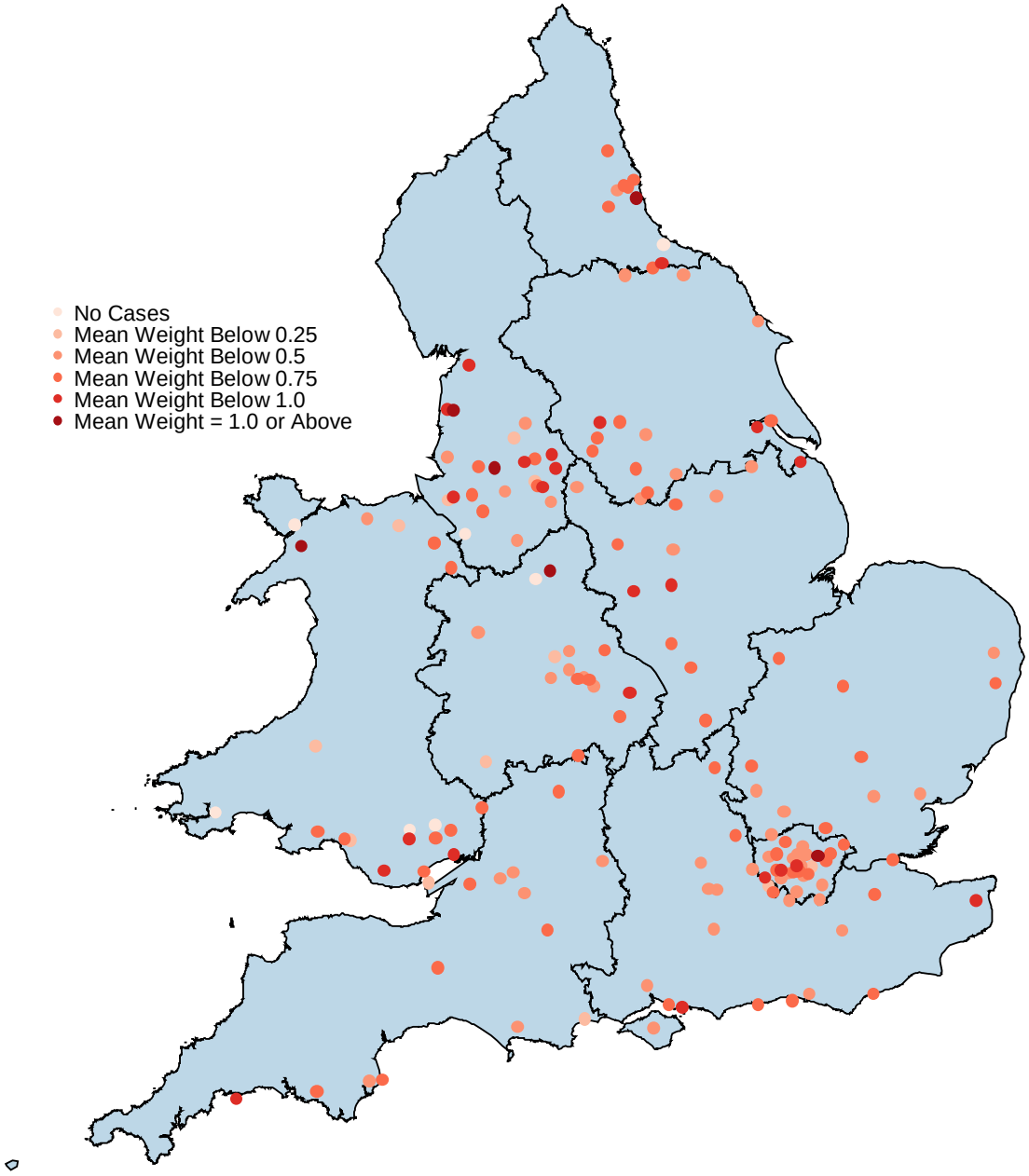

Figure 5.6. Trisomy 13 Combined Weight by PCT

Section 5.4.3. Scan Results

Trisomy 18

Table 5.4.2shows that without weights applied, both R and SaTScan identify 4 clusters, although the programs are not entirely consistent. Both methods return significant trisomy 18 clusters in 2004, 2008 and 2010, however R also shows a significant cluster in 2007 while SaTScan shows a cluster in 2006. Without weighting applied there were 5 significant clusters identified by each program with p = 0.001 in R and p < 0.0001 in SaTScan. Therefore, excluding cases which were ineligible for the weighted scans has had a considerable effect on the results.

Table 5.4.2.Unweighted Trisomy 18 Scan Results

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	290	146	51.908 N	0.443 W	111	70	0.001	R
	290	146	51.908 N	0.443 W	81.73	81.73	< 0.0001	S
2005	357	37	51.497 N	0.107 W	11.1	7	0.101	R
	357	59	51.486 N	0.169 W	14	14	0.103	S
2006	362	110	52.035 N	0.484 W	77.7	49	0.01	R
	362	115	51.908 N	0.443 W	55.27	55.27	0.0023	S
2007	406	187	51.624 N	0.054 W	99.9	63	0.007	R
	406	123	51.879 N	0.544 E	70.19	70.19	0.036	S
2008	410	214	51.646 N	0.202 W	111	70	0.001	R
	410	193	51.391 N	0.302 W	72.73	72.73	< 0.0001	S
2009	410	135	51.206 N	0.797 W	88.8	56	0.017	R
	410	132	51.206 N	0.797 W	62.99	62.99	0.01	S
2010	402	154	51.408 N	0.843 W	88.8	56	0.001	R
	402	61	51.351 N	0.165 W	17.87	17.87	< 0.0001	S

Tables 5.4.3shows that there is a very large difference between results generated in R and SaTScan. Weights in R appear to have worked as expected, and accounted for all previously observed clustering. In SaTScan, highly significant clusters have been identified in every year. 6 out of 7 of these clusters include PCTs in West London, a region which was shown in chapter 5.3.b to be weighted down on average.

Table 5.4.3. Trisomy 18 Scan Results, Cases Weighted for Maternal Age

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	117.1	56.9	51.908 N	0.443 W	70	111	0.052	R
	290	79	51.509 N	0.477 W	28.82	28.82	< 0.0001	S
2005	152.8	9.5	51.433 N	2.844 W	21	33.3	0.012	R
	357	63	51.549 N	1.731 W	86.83	86.83	0.0007	S
2006	151.6	4.6	51.009 N	3.115 W	7	11.1	0.655	R
	362	32	51.549 N	1.731 W	61.22	61.22	< 0.0001	S
2007	169.8	10.9	52.475 N	1.940 W	7	11.1	0.853	R
	406	103	53.181 N	2.445 W	88.63	88.63	< 0.0001	S
2008	166.2	45	52.022 N	0.791 W	56	88.8	0.053	R
	410	140	52.022 N	0.791 W	74.49	74.49	< 0.0001	S
2009	187.3	4.9	51.555 N	0.700 E	7	11.1	0.313	R
	410	94	51.404 N	0.774 W	47.58	47.58	< 0.0001	S
2010	163.1	6.7	50.340 N	4.799 W	49	77.7	0.397	R
	402	116	51.206 N	0.797 W	59.38	59.38	< 0.0001	S

Table 5.4.4 shows that the gestational age weighting method has been effective in both R and SaTScan, and has reduced the significance of all results in both programs. The reduction is small as the weighting method is mild, but the 2007 cluster in R has been accounted for and is no longer significant. All other clusters identified in tables 5.3.9 and 5.3.10 are still present although not as strongly.

Table 5.4.4. Trisomy 18 Scan Results, Cases Weighted for Gestational Age

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	250.8	122.7	51.908 N	0.443 W	70	111	0.001	R
	290	146	51.908 N	0.443 W	81.73	81.73	0.0003	S
2005	312.2	31.5	51.497 N	0.107 W	7	11.1	0.196	R
	357	37	51.527 N	0.104 W	6.98	6.98	0.234	S
2006	313.5	93.7	52.033 N	0.484 W	49	77.7	0.018	R
	362	115	51.908 N	0.443 W	55.27	55.27	0.0044	S
2007	351.3	160	51.624 N	0.054 W	63	99.9	0.075	R
	406	123	51.879 N	0.544 E	70.19	70.19	0.105	S
2008	353	180.5	51.646 N	0.202 W	70	111	0.001	R
	410	193	51.391 N	0.302 W	72.73	72.73	< 0.0001	S
2009	350.9	112.8	51.206 N	0.797 W	56	88.8	0.153	R
	410	132	51.206 N	0.797 W	62.99	62.99	0.030	S
2010	341.1	72.4	51.467 N	0.370 W	21	33.3	0.001	R
	402	61	51.351 N	0.165 W	17.87	17.87	< 0.0001	S

Table 5.4.5shows that when trisomy 18 cases were weighted for both maternal age and gestational age at diagnosis, all clusters are accounted for in R whereas strong clusters were detected in every year in SaTScan. 4 of these clusters are situated in and around London, an area which would be expected to be accounted for by the maternal and gestational age weighting.

Table 5.4.5. Trisomy 18 Scan Results, Both Weights Combined

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	103.8	48.3	51.908 N	0.443 W	70	111	0.314	R
	290	148	51.908 N	0.443 W	88.65	88.65	< 0.0001	S
2005	138.1	9.3	51.433 N	2.845 W	21	33.3	0.008	R
	357	12	51.433 N	2.845 W	19.05	19.05	< 0.0001	S
2006	134.5	4.3	51.009 N	3.115 W	7	11.1	0.598	R
	362	32	51.549 N	1.731 W	61.22	61.22	< 0.0001	S
2007	150.5	9.9	52.475 N	1.940 W	7	11.1	0.852	R
	406	50	53.544 N	2.128 W	50.18	50.18	< 0.0001	S
2008	146.2	39.7	52.022 N	0.791 W	56	88.8	0.065	R
	410	195	52.022 N	0.791 W	87.37	87.37	< 0.0001	S
2009	163.8	3.8	51.555 N	0.700 W	7	11.1	0.670	R
	410	28	51.348 N	1.400 E	82.76	82.76	0.0004	S
2010	140.6	6.1	50.340 N	4.799 W	49	77.7	0.330	R
	402	173	52.022 N	0.791 W	84.4	84.4	< 0.0001	S

Tables 5.4.3 and 5.4.5 show that weighting cases for maternal age and for maternal; and gestational ages combined on the basis of risk has been successful in trisomy 18 cluster scans in R, but not in SaTScan. Adjusting for gestational age alone has been successful using both programs. Results indicate that adjusting for maternal age has a much larger effect on results than adjusting for gestational age as the weights are much more extreme.

Trisomy 13

Table 5.4.6shows scan results without any weighting applied. In chapter 5.3.a, the number of cases lost in weighted scans due to missing information was found to be very high with 24.8% of total trisomy 18 cases and 29.3% of trisomy 13 cases having to be excluded. It is possible that excluding these cases could affect the findings which were reported in chapter 4.

After applying the Bonferroni correction there is no evidence of clustering in R and 1 significant cluster in SaTScan, observed in London in 2006. The equivalent scan results before cases were excluded were presented in chapter 4 and featured 3 significant clusters in both R and SaTScan.

Table 5.4.6. Unweighted Trisomy 13 Scan Results

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	128	17	50.806 N	1.062 W	55.5	35	0.057	R
	128	25	50.820 N	0.424 W	63.96	63.96	0.017	S
2005	130	3	55.161 N	1.691 W	11.1	7	0.900	R
	130	14	52.079 N	0.436 E	54.12	54.12	0.910	S
2006	162	43	51.391 N	0.302 W	33.3	21	0.010	R
	162	33	51.571 N	0.336 W	16.76	16.76	0.0023	S
2007	177	9	51.461 N	2.592 W	22.2	14	0.100	R
	177	20	51.904 N	2.097 W	59.82	59.82	0.143	S
2008	156	67	50.820 N	0.424 W	99.9	63	0.010	R
	156	49	50.820 N	0.424 W	78.36	78.36	0.058	S
2009	134	9	51.433 N	2.845 W	33.3	21	0.044	R
	134	20	51.393 N	0.105 W	12.24	12.24	0.028	S
2010	154	55	52.265 N	0.870 W	111	70	0.010	R
	154	21	51.206 N	0.797 W	40.12	40.12	0.054	S

Table 5.4.7 shows that after weighting for maternal age only, p-values in R are much higher and those in SaTScan are much lower than when scans were performed without weighting. Similarly to the results presented in chapter 5.2.c, this is likely due to the inability of SaTScan to accurately scan when an extreme weighting method has been applied. In SaTScan a total of 4 significant clusters were detected, all of which included the area of West London which was identified in chapter 4 to be the primary location for maternal age related clusters. In chapter 5.3.b it was demonstrated that using the proposed maternal age weighting method, cases in this area would on average be weighted down compared to other regions of England and Wales.

Table 5.4.7. Trisomy 13 Scan Results, Cases Weighted for Maternal Age

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	67.1	4.4	54.562 N	1.304 W	7	11.1	0.018	R
	128	5	54.562 N	1.304 W	5.09	5.09	0.0077	S
2005	71.9	3.3	54.071 N	2.855 W	7	11.1	0.565	R
	130	9	52.606 N	1.548 E	89.67	89.67	0.06	S
2006	93.1	19.8	51.587 N	0.270 W	14	22.2	0.203	R
	162	31	51.587 N	0.270 W	15.03	15.03	< 0.0001	S
2007	102.6	25.9	51.549 N	1.731 W	70	111	0.943	R
	177	35	51.549 N	1.731 W	83.31	83.31	0.0029	S
2008	90.3	6.0	51.426 N	0.339 W	7	11.1	0.812	R
	156	28	50.806 N	1.062 W	83.86	83.86	0.033	S
2009	74.1	5.5	51.433 N	2.845 W	42	66.6	0.186	R
	134	50	51.562 N	0.135 W	57.78	57.78	< 0.0001	S
2010	94.2	34.7	52.265 N	0.870 W	70	111	0.043	R
	154	35	51.408 N	0.843 W	49.76	49.76	0.007	S

Table 5.4.8shows that the gestational age weighting has had a small effect on p values in both R and SaTScan. In both programs, p values are mostly higher than those calculated without weighting applied, although this has had very little effect on cluster significance. The 2006 cluster is SaTScan is still identified as the only significant cluster in either program, although the same cluster is identified in R and is only marginally non-significant after applying the Bonferroni correction.

Table 5.4.8. Trisomy 13 Scan Results, Cases Weighted for Gestational Age

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	117.7	15.4	50.806 N	1.062 W	35	55.5	0.174	R
	128	25	50.820 N	0.424 W	63.96	63.96	0.021	S
2005	120.8	3	55.161 N	1.691 W	7	11.1	0.957	R
	130	14	52.079 N	0.436 E	54.12	54.12	0.940	S
2006	151.2	39.9	51.391 N	0.302 W	21	33.3	0.008	R
	162	33	51.571 N	0.336 W	16.76	16.76	0.003	S
2007	161.9	28.7	51.904 N	2.097 W	49	77.7	0.205	R
	177	20	51.904 N	2.097 W	59.82	59.82	0.111	S
2008	143.8	52.1	51.206 N	0.797 W	56	88.8	0.232	R
	156	49	50.820 N	0.424 W	78.36	78.36	0.084	S
2009	123.3	8.3	51.433 N	2.845 W	21	33.3	0.098	R
	134	20	51.393 N	0.105 W	12.24	12.24	0.084	S
2010	140.2	50.3	52.265 N	0.870 W	70	111	0.014	R
	154	21	51.206 N	0.797 W	40.12	40.12	0.116	S

Tables 5.4.9shows that when both weights are combined there are no clusters present in R and 4 in SaTScan.

Table 5.4.9. Trisomy 13 Scan Results, Both Weights Combined

	Cases		Centroid		Radius (km)			
Year	**Total**	**In Cluster**	**Latitude**	**Longitude**	**Vertical**	**Horizontal**	**P Value**	**Scan**
2004	62.9	4.2	54.562 N	1.304 W	11.1	7	0.024	R
	128	5	54.561 N	1.304 W	5.09	5.09	0.0082	S
2005	67.9	3.3	54.071 N	2.855 W	11.1	7	0.524	R
	130	9	52.606 N	1.548 E	89.67	89.67	0.037	S
2006	88.1	18.5	51.587 N	0.270 W	22.2	14	0.32	R
	162	31	51.587 N	0.270 W	15.03	15.03	< 0.0001	S
2007	94.5	23.8	51.549 N	1.731 W	111	70	0.926	R
	177	35	51.549 N	1.731 W	83.31	83.31	0.002	S
2008	84.3	5.6	51.426 N	0.339 W	11.1	7	0.824	R
	156	28	50.806 N	1.062 W	83.86	83.86	0.045	S
2009	69.2	6.4	51.009 N	3.115 W	66.6	42	0.238	R
	134	50	51.562 N	0.135 W	57.78	57.78	< 0.0001	S
2010	85.8	31.9	52.265 N	0.870 W	111	70	0.056	R
	154	35	51.408 N	0.843 W	49.76	49.76	0.013	S

Tables 5.4.6 to 5.4.9 show similar results for trisomy 13 as were observed for trisomy 18 in tables 5.4.2 to 5.4.5. Weighting for gestational age alone was successful in both methods, however adjusting for maternal age and both weights combined was successful in R but not SaTScan. Again, adjusting for maternal age had a much larger effect on the results.

Section 5.5.Summary

In chapter 5, a novel weighting method was detailed which assigns weights directly to cases based on observed and quantifiable risk factors. This was tested in a custom program in R by comparing results with those using an alternate weighting method in SaTScan. Weights could not be assigned directly to cases in SaTScan, so instead the inverse of the mean case weight for each PCT was applied as a PCT adjustment value.

Synthetic data was used to compare R and SaTScan in 3 simple scenarios. Firstly, a single test cluster in a uniform population was accounted for by assigning progressively lower weights to cases within the cluster. Both methods successfully accounted for the test cluster. Secondly, weights were applied at random to PCTs throughout the study region in order to test how each method handled extreme weights in a scenario in which no clusters would be expected. In this test, there was no evidence of clustering in R while SaTScan reported significant clusters which became more significant as the weights used became more extreme. Finally, a dataset was created to simulate how different case weights affected the p-value in R. The mean case weight for each PCT remained constant so that SaTScan would always return the same p-value. As the weighting method became more extreme, it was possible to obtain a p-value in R which was around 3 orders of magnitude greater than that in SaTScan. However, to observe a difference of this magnitude weights a very extreme weighting had to be applied which would not realistically be used in an actual study. When more moderate weights were applied, there was no observed difference between R and SaTScan.

It is therefore reasonable to assume that direct case weights in R are more accurate than PCT adjustments in SaTScan as results were consistent between the two programs except for the randomly assigned weights test, where SaTScan detected significant clusters in data which in all likelihood contained no clusters.

Cases of trisomies 18 and 13 were then weighting according to their maternal age and gestational age at diagnosis in order to attempt to account for the observed clustering in chapter 4. For both trisomies, it was observed that direct case weighting in R had worked as expected when weighting for both variables individually and combined. Results in R suggest no evidence of clustering beyond maternal age and gestational age at diagnosis. Results in SaTScan were not consistent with those in R when extreme weights were used, and showed very strong clusters when cases were weighted by maternal age or both weights combined. When cases were weighted by gestational age at diagnosis alone, results were consistent with R and showed a slight decrease in cluster significance.

Assuming that results in R are accurate, there is no evidence of trisomy 18 or trisomy 13 clustering in England and Wales between 2004 and 2010 after adjusting for maternal age and

gestational age at diagnosis. Results indicate that the clusters reported in chapter 4 were mostly caused by advanced maternal age in London, with gestational age at diagnosis contributing a small amount to these clusters. Table 5.1 on page 123 shows that the proportion of births in high risk age groups was higher in London than any other region in England and Wales.

As part of the NDSCR data processing and quality control, data are matched with the Regional Congenital Anomaly Registers to improve data quality. A previous study showed that the NDSCR data was over 94% complete (Savva & Morris 2009). Most labs provide data within 6 months of diagnosis, and regular requests are also sent for additional information on cases with incomplete data (NDSCR Annual Report 2010). Any obvious gaps in the data (for example, time periods with noticeably few cases reported) are also identified and raised with the appropriate lab(s). The analysis for the spatial cluster scans was undertaken in 2012 on cases diagnosed between 2004 and 2010, meaning that there was no substantial outstanding data for any of the labs, and followup had been undertaken to ensure that the data was as complete as possible. While it is reasonable to believe that there may be slight differences in reporting to the NDSCR by the individual labs, it is therefore unlikely that there is substantial variation in terms of the proportion of cases reported in each region.
Differences in ascertainment are also caused by regional differences in prenatal diagnosis methods, as regions which use the earliest or most effective screening procedures detect more cases than those which do not. This was accounted for both by gestational age weighting and excluding foetal losses under 20 weeks (full explanation in chapter 5.1).

Another issue which must be considered is whether multiplying the weights was an appropriate way to account for both risk factors simultaneously. It is possible that more affluent mothers may choose to wait before starting a family and be older when having children. The fact that trisomy risk increases with maternal age is well publicised, and these women may have been more active in selecting the earliest screening methods possible, either through the NHS or private care. The maternal age weight is applied as trisomy risk is higher in older mothers, while the gestational age weight is applied as the risk of fetal loss varies by gestational age. It may be that older mothers are more likely to miscarry than younger mothers, however this has not been demonstrated in trisomy 18 or 13 pregnancies. There has also been no reported evidence that more affluent mothers are less likely to have either a miscarriage or an affected pregnancy than less affluent mothers. I therefore believe that multiplication of the weights is a valid way to account for both risk factors.

Chapter 6: Conclusions

In section 1.3.4, the aims of this thesis were stated as:

1. To scan trisomy data in England and Wales from the NDSCR for temporal and spatial clusters.
2. To develop a statistical method capable of accounting for two risk factors simultaneously, and to demonstrate the effectiveness of this method using simulated data.
3. To scan the same data for clusters after adjusting for maternal age and gestational age at diagnosis using the novel method.
4. To assess the impact of the findings in the context of both previously undertaken and future congenital anomaly research, and NHS policy.

The first three aims on this list have already been completed. A summary and analysis of the findings in this thesis is presented in section 6.1. In section 6.2, a more detailed discussion of the wider impact of the findings of this thesis is presented. Finally, section 6.3 looks to future work which could be carried out building on what has been developed and performed during this thesis.

Section 6.1. Summary and Analysis of Findings

Temporal Cluster Detection

Analysis of cases from the NDSCR found no evidence of temporal clustering for either trisomy in England and Wales between 2004 and 2010, using the Nagarwalla temporal cluster scanning method published in 1996. Maternal age at conception would not be expected to produce distinct, visible temporal clusters, however as prenatal diagnosis techniques improve and mean maternal age continues to rise, both of these factors may contribute to a long term increase of observed prevalence of both trisomy 18 and trisomy 13.

The only published evidence of temporal trisomy clustering remains that linking Down syndrome risk to ionising radiation exposure caused by the Chernobyl disaster, in which a strong, quick burst of radiation was detected in multiple European countries. While extreme events may have the potential to increase trisomy risk sufficiently to produce clusters, in the absence of such events it may be that no individual factor is strong enough to produce clusters which can be detected using this method.

The Nagarwalla temporal cluster scanning method has been proven to be effective both in the original publication and numerous studies since then. The method is quick to run, can easily be adapted to the needs of the investigator and requires only the bare minimum of information meaning that very few cases are excluded from analysis on the basis of missing data. Also of note is the ability to search for clusters of any size in a single scan without having to adjust for multiple testing.

Data was scanned for clusters by individual government office region to give the analysis some resolution in space and to allow comparison of regional results. Analysing data in this way is advantageous as it allows regions to be scanned as soon as complete data is available, but it is possible that clusters occurring on the border or 2 or more regions could be missed. In the event of significant or nearly significant results being obtained in 2 or more connecting regions, it would be possible to combine these regions and perform another scan which would be able to detect clusters occurring at the boundary. This analysis was not performed in this thesis due to the lack of evidence of temporal clusters for either trisomy in any region.

Separate scans were performed using cases arranged by date of sample, and by estimated date of conception. Date of sample scans are useful in biosurveillance where cases are routinely scanned for clusters, but their relevance is limited in trisomy scanning as the genetic errors occur at the time of conception. Samples can be taken from a range of gestational ages from approximately 10 to 40 weeks, so a cluster of cases by date of sample does not necessarily mean that a genuine cluster has occurred. Running scans by date of conception is necessary to search for trisomy clusters, however there are 2 main disadvantages to such an

approach. Date of conception is estimated by subtracting gestational age at sample from the sample date, or gestational age at outcome from the outcome date. In both cases, gestational age is only recorded in completed weeks so the estimated conception date is only accurate to the week. Also, estimating date of conception requires more information than scanning by date of sample so fewer cases can be included in the scan. 2.2% of trisomy 18 cases and 2.8% of trisomy 13 cases were excluded from the scans by sample date due to missing information, compared to 8.2% of trisomy 18 cases and 10.5% of trisomy 13 cases when using estimated date of conception. A sensitivity analysis was performed by imputing date of conception using less accurate assumptions and imputing gestational age at sample as the mean gestational age for the sample tissue taken, which resulted in excluding only 1.6% of trisomy 18 and 2.0% of trisomy 13 cases. Including these cases in the analysis did not have any effect on the results and no significant clusters were detected. Weighted temporal scans were not performed in this thesis due to the expectation that the factors being adjusted for would not result in temporal clusters, and to focus on the weighted spatial analyses in which regional variation in maternal age and gestational age at diagnosis could be masking other possible clusters. However, such scans could be performed in future analyses.

Unadjusted Spatial Cluster Detection

Strong evidence of spatial clustering in London was observed for both trisomies when cases were unadjusted. When analysing data by individual calendar year, 5 trisomy 18 and 3 trisomy 13 clusters were identified. Very strong clusters were observed for both trisomies when all years were analysed together. The high maternal age in this region compared to the remainder of England and Wales was very likely to be the primary cause of all observed clusters for both trisomies, and it was also possible that regional differences in prenatal screening methods contributed to the clustering effect.

Results in R were compared to those using the freely available program SaTScan, which is a commonly used tool for temporal, spatial and space-time cluster scanning. Both programs used the same statistical methodology but differed slightly in the way circles were generated at each PCT site. These differences were responsible for small differences in p-values, but overall the same findings were obtained using both programs.

Given that maternal age has a very strong influence over trisomy risk, it is surprising that clusters were not observed in London in every scan when data was analysed by calendar year. Of the 8 clusters (5 for trisomy 18 and 3 for trisomy 13), 7 were in London with one small cluster in Bristol (trisomy 13 in 2009). Therefore, what could be considered as the expected result was only observed in half of the scans. It is worth noting that all 5 trisomy 18 clusters had p-values of 0.001 in R and below 0.0001 in SaTScan. In the remaining 2 years, 1 cluster was

observed with $p < 0.05$ in R and both years had clusters with $p < 0.05$ in SaTScan, however these results were non-significant after applying a Bonferroni correction for multiple testing. Evidence of trisomy 13 clustering was not as strong, with only 2 clusters discovered in London. However, this may be due to trisomy 13 being less prevalent than trisomy 18. There was an additional year in which both methods identified a cluster with $p < 0.05$ which was not significant after applying a Bonferroni correction.

The power of both of these methods was evaluated using a set of synthetic data. A set of 12 PCTs was selected as an artificial cluster site in the North East of England which could be uniquely identified using both programs. It was found in both programs that as the prevalence outside the cluster increased, the excess risk required for the cluster to be judged to be significant decreased. This is important as trisomies 18 and 13 are rare conditions, and therefore any potential clustering effect would have to be strong in order to be detected. Trisomy 18 is more common than trisomy 13, which may explain the relative strength of the observed trisomy 18 clusters compared to those of trisomy 13.

Cases were analysed by maternal PCT, which was the smallest available population unit for which population birth data was available. During the study period there were 174 PCTs in England and Wales, and on average in an individual calendar year there would be expected to be between 2 and 3 cases of trisomy 18 and between 1 and 2 cases of trisomy 13. As such, PCTs are an acceptable population unit for this analysis as there are not so many cases per PCT that smaller population blocks would be obviously more accurate, and there are not commonly too many PCTs with 0 cases present. There were a number of years in which around half of all PCTs had 0 trisomy 13 cases, however zero count PCTs do not adversely affect the statistical method. It is therefore unlikely that use of smaller population blocks would have been more accurate, or that the results would have been different using smaller population units. This could again suggest that the scarcity of trisomy 13 cases compared to trisomy 18 explains why trisomy 18 clusters were more prevalent and stronger than those of trisomy 13.

Adjusted Spatial Cluster Detection

A novel method was proposed to account for risk factors in spatial cluster analyses. By weighting each case directly, the program is able to effectively account for previously observed clusters on the basis of risk as the case structure is taken into account during Monte Carlo simulation to calculate p-values. Cases could be directly weighted in R but not in SaTScan. In SaTScan, the inverse of the mean weight for each PCT was calculated and applied as a PCT specific adjustment.

This method was tested thoroughly using synthetic data. A simple dataset was created and weights were applied in both programs to adjust for advanced maternal age in a single test

cluster. The cluster was successfully accounted for using both methods, and both returned similar p-values. Two further scenarios were then created. The first involved weights being applied at random throughout the study region. This was not expected to lead to any clusters being present, but examined how each method responded to the application of extreme weights. In this test, no clusters were present at all in R but in SaTScan significant clusters were detected which became more significant as the weights became more extreme. The final test using synthetic data tested how the effect of different case weights affected the p-value in R. Different case weights were applied to cases in R which did not change the mean weight in each PCT, so the weighting (and consequently, p-value) in SaTScan remained the same. By increasing the extremity of the weighting method, it was possible to obtain a p-value in R which was around 3 orders of magnitude greater than that in SaTScan (0.017 in R compared to 0.000022 in SaTScan). However, the weighting scale used to obtain this p-value was very extreme and ranged from 0.02 to 1.98 (relative scale 1-99).

These results show that there is a clear difference between R and SaTScan when weights are applied. The direct case weights in R appear to offer increased accuracy, reliability and flexibility compared to SaTScan. Results in R depend on the exact weights being used rather than just the average, and the significant clusters in SaTScan when weights were applied at random indicate that SaTScan cannot accurately use extreme weighting methods.

Maternal age was adjusted for by weighting cases based on maternal age at diagnosis. Maternal age weights ranged from 0.019 to 1.182 for trisomy 18 (relative scale 1 – 62), and from 0.052 to 1.286 for trisomy 13 (relative scale 1 – 25). Both of these weighting ranges are extreme as the largest weight is worth several times that of the smallest. These weighting ranges were comparable to those in the randomly assigned weight test using the synthetic data set. The weighting range for trisomy 18 was greater than any range tested using synthetic data, while the range of trisomy 13 cases fell between the 1 – 20 and 1 – 40 ranges which were tested.. No clusters were detected for either trisomy after weights were applied in R, while SaTScan detected strong trisomy 18 clusters in every year scanned and significant trisomy 13 clusters in 4 of 7 years.

Gestational age was adjusted for by weighting cases based on their likelihood of surviving to 20 weeks gestation. Gestational age weights ranged from 0.754 – 1 for trisomy 18 and from 0.844 – 1 for trisomy 13. Compared to the maternal age weights this is a minor adjustment, as there is only a small difference between the largest and smallest weight for either trisomy. After weights were applied there was a small decrease in significance in all scans for both trisomies in both methods.

A final set of scans were then performed using both weights combined. For each case, weights were multiplied to give a combined weight which was applied directly in R and as a PCT

adjustment in SaTScan. As the maternal age weight was much more extreme, it would be expected that results obtained using both weights combined would be similar to those obtained using the maternal age weight alone. No clusters were detected for either trisomy after weights were applied in R, while SaTScan detected the same clusters which were found after applying the maternal age weight only.

The scan results using trisomy 18 and trisomy 13 case data provide further evidence that R outperforms SaTScan when the weighting method used is extreme. After weighting for maternal age, it would be expected that the clusters discovered in chapter 4 would be accounted for as they were most likely age-dependant. This result was observed in R, but SaTScan appeared to amplify existing clusters rather than account for them. However, both methods performed as expected when the relatively minor adjustment for gestational age at diagnosis was applied. This suggests that SaTScan is suitable to use when the weights being applied are small.

The novel weighting method offers a number of advantages over methods available in SaTScan. Weights are accurate and highly adaptable, and can be calculated based on whatever criteria is required by the investigator. It is theoretically possible to account for as many different factors as is desired, although there could be adverse effects caused by using too many weights simultaneously.

Calculating case weights requires the availability of precise information regarding every weighting variable for every case. In total, 24.8% of trisomy 18 cases and 29.3% of trisomy 13 cases were excluded from weighted analyses due to either missing information or ineligibility (all fetal losses under 20 weeks were excluded). Losing this many cases resulted in a change of results even without applying weights. In chapter 4, 7 clusters of trisomy 18 and 3 clusters of trisomy 13 were identified using both R and SaTScan. In chapter 5, only 4 trisomy 18 clusters were detected with no evidence of trisomy 13 clustering. It may be that the cases excluded were those responsible for the clusters reported in chapter 4, but it is more likely that simply excluding such a high proportion of cases negatively affected the power of both methods to detect genuine clusters. One potential conclusion already outlined in this chapter stated that both programs may have struggled to identify trisomy 13 clusters compared to trisomy 18 due to lower prevalence. Reducing sample size results in a loss of power, so excluding 20-30% of all cases from the analysis would also have a similar effect. Adjusting for maternal age resulted in a loss of 2.6% of trisomy 18 cases and 4.1% of trisomy 13 cases, compared to a loss of 16.9% and 19% of total cases respectively when adjusting for gestational age. In addition to the unadjusted results from chapter 5, it would be worth performing a sensitivity analysis to assess the effect of excluding such a high proportion of cases. It may also be worth adjusting for

maternal age alone in order to avoid having to exclude so many cases from the weighted scans.

The effect of losing cases is compounded when multiple factors are being accounted for simultaneously. Each case in the combined weighted analysis must have complete information regarding each weighting variable in order to maintain accuracy. Therefore, as more factors are being adjusted for, the more likely it is that each individual case will be excluded due to missing information. It is impossible to impose an arbitrary limit on the number of cases which can be adjusted for simultaneously, but there will come a point where the benefit of adjustment is outweighed by the number of cases being lost.

However, a large percentage of the cases excluded from the weighted scans in chapter 5 were excluded due to being fetal losses under 20 weeks rather than for missing information. In total, 56% of trisomy 18 cases and 49% of trisomy 13 cases which were excluded were fetal losses under 20 weeks, while only 23% of trisomy 18 exclusions and 29% of trisomy 13 exclusions were due to missing maternal age or gestational age at diagnosis. It seems that the maximum number of factors which can be accounted for simultaneously depends on the proportion of cases which will be lost when the weights are calculated.

Strong clusters were detected in chapter 4 when cases were not adjusted, and adjustment is therefore necessary in order to account for known risk factors. However, in chapter 5 there were a large number of cases excluded from analysis due to missing gestational age information. Adjusting for gestational age had a relatively small effect on the results compared to maternal age, and it may be that the benefit of adjusting for gestational age did not merit the large number of cases which were lost and subsequent loss of power. The best analysis for trisomy clustering may therefore be to adjust for maternal age only.

These analyses have led to the initial conclusion that in scenarios where there is a significant variation in the size of case weights being applied, the novel method in R uniformly outperforms SaTScan and is therefore recommended for use. However, when cases are not weighted this new method offers no noticeable advantage over SaTScan.

Section 6.2. Comparison with Other Studies

In 2008, McNally *et al* found evidence of clustering of trisomy 21 in the North of England between 1985 and 2003 using a nearest neighbour method after accounting for maternal age. The authors theorised that an infectious agent may be responsible for this, which is consistent with their finding that cases clustered most strongly within densely populated where infection can spread more easily. In the same study, the investigators looked for clustering in trisomies 18 and 13, finding no evidence of clustering for either trisomy.

In section 1.3.1 of this thesis it was speculated that the lack of evidence for clustering in this study may have been due to their low available sample size of trisomies 18 and 13, and that the greater sample size available in the NDSCR data may enable the detection of weaker clustering effects. There are three main differences between the cluster scans carried out in this thesis compared to those in the McNally *et al* study, which are the cluster detection method, the timeframe of data included, and the scale of the area scanned. However, this is still the best available study to compare with the work carried out in this thesis as the majority of trisomy cluster analyses are performed on cases of the much more common trisomy 21. The results reported in this thesis are consistent with those in the McNally *et al* report, and the scans performed were unable to detect any evidence of clustering.

Section 6.3. Wider Impact of Findings

Implications for Policy

The only clusters detected in this thesis were demonstrated to be the result of regional variation in maternal age, which is already a known risk factor for trisomic pregnancies and an important part of all screening procedures. Owing to the lack of novel findings, there are no recommended changes to be made to policy as a result of the findings made in the cluster scans performed in this thesis.

Recommendations for Research

The NDSCR remains a very valuable resource for future trisomy research, in particular being able to monitor trisomy prevalence and trends as mean maternal age continues to change and more accurate screening methods are developed and adopted. As a result of the lack of temporal and spatial clusters detected (after adjustment) of trisomies 18 and 13 detected in England and Wales in 2004-2010, I would not recommend that such scans were added to the NDSCR routine analysis. However, the findings in this thesis do not mean that there is no potential for trisomy clusters to appear in future. The majority of trisomy clusters reported to date occurred apparently as a result of the nuclear disaster in Chernobyl, and while it is still disputed whether all of these clusters could have been caused by the expected levels of radiation dispersed over Europe, there is very strong evidence that exposure to ionising radiation and trisomy prevalence are linked. The recent nuclear disaster arising as a result of the Japan earthquake in 2011 provides an opportunity to examine this link further, however it is possible that low ascertainment of trisomy cases in Japan may hinder this analysis. In 1999, Hoshi *et al* reported a study of trends in Down syndrome prevalence in Japan between 1980 and 1997, using data obtained from the Japan Association of Obstetricians and Gynecologists (JAOG). JOAG operates a hospital based register of congenital anomalies, however this is only maintained in selected hospitals in Japan and ascertainment in the Hoshi *et al* study was estimated at 60-70%.

The study of other diseases could also be improved using this method. For example, gastroschisis is a congenital anomaly in which the abdominal contents can protrude through a small hole in the anterior abdominal wall, and is more common amongst younger mothers (Kirby *et al* 2013). Multiple studies have indicated that maternal aspirin use during pregnancy can increase gastroschisis risk (Werler *et al* 2002), with one large study by the California Birth Defects Monitoring Program (Torfs *et al* 1996) finding that aspirin could quadruple gastroschisis risk. It is possible that, using cluster analysis weighting for maternal age, alternate risk factors may be identified.

Use of this method is not restricted to the study of congenital anomalies. Age related conditions such as Alzheimer's disease are not well understood and although prevalence increases with age, there is a wide range of proposed factors which could increase the risk of developing the disease. One particularly strong link is that between Alzheimer's and exposure to aluminium (Crapper *et al* 1976, Martyn *et al* 1989). Scanning Alzheimer's cases for clusters weighting for age at diagnosis could help to identify environmental effects which contribute to the disease risk. This could also be extended to look for regions with unusually low prevalence, which could help to identify factors with a protective effect.

Section 6.4: Future Work

The weighting method developed in this thesis can be extended to have a wider range of uses in biosurveillance. The method has been demonstrated to be effective in quadrat based spatial cluster scanning, and the most apparent extension of the method is to include quadrat based temporal and spatio-temporal cluster scanning. Both of these techniques also use the ratio of cases inside and outside the quadrat to determine cluster location and significance, so implementation would be very similar to that in the spatial scanning method in chapter 5. Besides maternal age, risk factors previously theorised as being associated with trisomy prevalence (such as ionising radiation and exposure to an infectious agent) have presented as space time effects arising from a single time point in a single location. While the temporal and spatial scans in this thesis also have some resolution in space and time respectively, dedicated spatio-temporal methods would have greater power to detect space-time clusters.

Another potential extension of the method is to examine temporal trends in addition to clusters, which would be able to detect risk factors which do not result in the creation of clusters. For example, the changes in maternal age in England and Wales during the time period studied in this thesis are not detected through temporal cluster scanning as the result is an increase in prevalence of both trisomies rather than the occurrence of clusters. Testing the temporal data for trend would reveal this steady increase. Weighting the data by maternal age would also enable the detection of alternate risk factors simultaneously causing long term increases or decreases in prevalence which are too small to be detected in addition to the very strong maternal age effect. For example, when studying trisomy data weighted by maternal age the underlying trend may reflect the improvement in prenatal screening over time resulting in more cases being detected prior to fetal loss.

It is also important to consider potential extension of this weighting method to distance based spatial clustering methods. These methods examine the distribution of cases to test for an overall clustering effect, rather than scanning through the study area to find cluster locations. In chapter 1, 2 such approaches were considered. The first, detailed by Whittemore *et al*, was a test which measured the distance between each case and its nearest neighbour case and compared the findings to an expected distance assuming all cases were independent and there was no clustering present. The second was detailed by Cuzick and Edwards, and examined the distance between each case and its nearest neighbour control, where controls were randomly selected from the population at risk. These methods are completely different to the quadrat based methods which the weighting technique was developed for, and it would be more difficult to adapt the novel method to such analyses. It would be easier and more effective to

stratify cases on the basis of one observed variable and look for clustering within each individual stratum. Currently, it is not envisaged that the weighting method can be easily adapted for use in distance-based spatial cluster detection methods.

The weighting method has been demonstrated to be effective when accounting for 2 clustering factors simultaneously, and is technically capable of accounting for as many factors as required by multiplying each weight to give an overall weight for each case. However, introducing too many weighting variables could diminish the effect of the weighting method. Chapter 5 examined the effect of weighting for maternal age at conception, gestational age at diagnosis, and both factors combined. The weighting method was effective in all 3 scenarios. These factors were very different in scale, as the maternal age weighting is an extreme weight with a factor of around 60 dividing the largest and smallest weights, while the gestational age weighting was comparatively mild and only weighted cases between 0.75 and 1. In this sense, maternal age can be classified as a major weight and gestational age as a minor weight. It may be that attempting to use 2 major weights in the same analysis diminishes the effect of both weights as they cancel each other out. Further work is necessary to understand the possible outcomes when multiple extreme weighting methods are applied simultaneously.

There are also applications for this type of weighting outside biosurveillance. For example, clusters of road traffic accidents could be searched for by weighting the driver based on age and sex, as there is a much higher risk of accidents being caused by young men than by any other demographic.

In conclusion, this thesis has developed a simple method to adjust for risk factors when searching for spatial clusters. The method could be adapted for use in searching for temporal, spatial or spatio-temporal clusters for other congenital anomalies or other diseases.

BIBLIOGRAPHY

1. Abil'dinova GZ, Kuleshov NP, Svyatova GS. 2003. Chromosomal Instability Parameters in the Population Affected by Nuclear Explosions at the Semipalatinsk Nuclear Test Site. Russian Journal of Genetics 39;8:944-947
2. Awa AA, Honda T, Sofuni T, Neriishi S, Yoshida MC, Matsui T. 1971. Chromosome-aberration frequency in cultured blood-cells in relation to radiation dose of A-Bomb survivors. Lancet 2;903-905.
3. Balmukhanov SB. 2006. Defense Threat Reduction Agency Technical Report: Medical Effects and Dosimetric Data from Nuclear Tests at the Semipalatinsk Test Site. United States of America Department of Defense.
4. Barsel-Bowers G, Pueschel SM, Abuelo DN, Charley J. 1980. A cluster of infants born with Trisomy 18. Lancet 1;120.
5. Benn PA. 2002. Advances in prenatal screening for Down syndrome: 1. General principles and second trimester testing. ClinicaChimicaActa 323;1-2:1-16
6. Besag and Newell. 1991. The Detection of Clusters in Rare Diseases. Journal of the Royal Statistical Society 154:1;143-155
7. Bloom AD, Awa AA, Archer PG, Neriishi S, Honda T. 1967. Chromosome aberrations in leucocytes of older survivors of the atomic bombings of Hiroshima and Nagasaki. Lancet 2;802-805
8. Bloom AD, Neriishi S, Kamada N, Iseki T, Keehn RJ. 1966. Cytogenetic investigation of survivors of atomic bombings of Hiroshima and Nagasaki. Lancet 2;672.
9. Bound JP, Francis BJ, Harvey PW. 1995. Down's syndrome: prevalence and ionising radiation in an area of north west England 1957-91. Journal of Epidemiology and Community Health 49:164-170
10. Brewer CM, Holloway SH, Stone DH, Carothers AD, FitzPatrick DR. 2002. Survival in trisomy 13 and trisomy 18 cases ascertained from population based registers. American Journal of Medical Genetics 39:e54
11. BurkartW, Grosche B, Schoetzau A. 1997. Down Syndrome Clusters in Germany after the Chernobyl Accident. Radiation Research 147;321-328.
12. Carter P, Pearn J, Bell J, Martin N, Anderson N. 1985. Survival in trisomy 18. Clinical Genetics 27:59-61
13. Chebotarev AN, Bochkov NP, Vinogradova EI. 2006. Periodic constituent of Down syndrome frequency dynamics. Doklady Biological Sciences 409:305-307
14. Crapper DR, Krishnan SS, Quittkat S. 1976. Aluminium, neurofibrillary degeneration and Alzheimer's disease. A Journal of Neurology 99:67-80

15. Cuzick J, Edwards R. 1990.Spatial Clustering for Inhomogeneous Populations. Journal of the Royal Statistical Society. Series B (Methodological) 52:1;73-104
16. Dean G, Nevin NC, Mikkelsen M, Karadima G, Petersen MB, Kelly M, O'Sullivan J. 2000. Investigation of a cluster of children with Down's syndrome born to mothers who had attended a school in Dundalk, Ireland. Occupational and Environmental Medicine 57;793-804
17. Dimaio MS, Baumgarten A, Greenstein RM, Saal HM, Mahoney MJ. 1987. Screening for fetal Downs-syndrome in pregnancy by measuring maternal serum alpha-fetoprotein levels. New England Journalof Medicine 317;6:342-346.
18. Dubrova YE, Bersimbaev RI, Djansugurova LB, Tankimanova MK, Mamyrbaeva ZZ, Mustonen R, Lindholm C, Hulten M, Salomaa S. 2002. Nuclear Weapons Tests and Human Germline Mutation Rate. Science 295:1037.
19. Duczmal L, Assuncao R. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics and Data Analysis 45;2:269-286
20. Edwards JH, Harnden DG, Cameron AH, Crosse VM, Wolff OH. 1960. A New Trisomic Syndrome. Lancet 1;787-90.
21. Embleton ND, Wyllie JP, Wright MJ, Bum J, Hunter S. 1996. Natural history of trisomy 18. Archives of Disease in Childhood 75:F38-F41
22. English P, Neutra R, Scalf R, Sullivan M, Waller L, Zhu L. Examining associations between childhood asthma and traffic flow using a geographic information system. Environmental Health Perspectives 107;9:761-767
23. Ericson A, Kallen B. 1994. Pregnancy Outcome in Sweden after the Chernobyl Accident. Environmental Research 67;149-159.
24. Fishler K, Koch R. 1991. Mental Development in Down syndrome mosaicism. American Journal on Mental Retardation 96:3;345-351
25. Freeman SB, Taft LF, Dooley KJ, Allran K, Sherman SL, Hassold TJ, Khoury MJ, Saker DM. 1998. Population-Based Study of Congenital Heart Defects in Down Syndrome. American Journal of Medical Genetics 80:213-217.
26. Goldstein H, Nielsen KJ, 1988. Rates and survival of individuals with trisomy 13 and 18. Data from a 10-year period in Denmark. Clinical Genetics 34:366-372
27. Hassold T, Hunt P. 2001. To Err (Meiotically) is Human: The Genesis of Human Aneuploidy. Nature Reviews Genetics 2;280-291
28. Hassold TJ, Jacobs PA. 1984. Trisomy in Man. Annual Review of Genetics 18;69-97

29. Hodges CA, Revenkova E, Jessberger R, Hassold TJ, Hunt PA. 2005. SMC1β-deficient female mice provide evidence that cohesins are a missing link in age-related nondisjunction. Nature Genetics 37;1351-1355
30. Hook EB. 1981. Rates of chromosome abnormalities at different maternal ages. Obstetrics andGynecology 58;282-5
31. Hoshi N, Hattori R, Hanatani K, Okuyama K, Yamada H, Kishida T, Yamada T, Sagawa T, Sumiyoshi Y, Fujimoto S. 1999. Recend Trends in the Prevalence of Down Syndrome in Japan, 1980-1997. American Journal of Medical Genetics 84:340-345
32. Hulten MA, Patel S, Jonasson J, Iwarsson E. 2010. On the origin of the maternal age effect in trisomy 21 Down syndrome: the Oocyte Mosaicism Selection Model. Reproduction 139;1-9
33. Irving C, Richmond S, Wren C, Longster C, Embleton ND. 2011. Changes in fetal prevalence and outcome for trisomies 13 and 18: a population-based study over 23 years. Journal of Maternal-Fetal and Neonatal Medicine 24:137–141.
34. Jones KT. 2008. Meiosis in oocytes: predisposition to aneuploidy and its increased incidence with age. Human Reproduction Update 14;143-158
35. Kirby RS, Marshall J, Tanner JP, Salemi JL, Feldkamp ML, Marengo L, Meyer RE, Druschel CM, Rickard R, Kucik JE, for the National Birth Defects Prevention Network. 2013. Prevalence and correlates of gastroschisis in 15 U.S. states: 1995-2005. Obstetrics & Gynecology 122;2:275-281
36. Kuderinov TK, Sarsembaev A, Shulakov TE. 1993. Psychiatric Status of Children Born in Karkarala District in 1950-1960. Specificity of Therapeutical Pathology in Semipalatinsk Region. Proceedings of Semipalatinsk Medical Institute, vol 1 (Unclassified).
37. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. 1997. Breast cancer clusters in the Northeast United States: A geographic analysis. American Journal of Epidemiology 146;2:161-170.
38. Kulldorff M, Nagarwalla N. 1995. Spatial Disease Clusters: Detection and Inference. Statistics in Medicine 14;799-810.
39. Lamb NE, Freeman SB, SavageAustin A, Pettay D, Taft L, Hersey J, Gu YC, Shen J, Saker D, May KM, Avramopoulos D, Petersen MB, Hallberg A, MikkelsenM, Hassold TJ, Sherman SL. 1996. Susceptible chiasmate configurations of chromosome 21 predispose to non-disjunction in both maternal meiosis I and meiosis II. Nature Genetics 14;400-405
40. Lejeune J, Turpin R, Gautier M. 1959. Chromosomal diagnosis of mongolism. Archives françaises de Pédiatrie16;962-3

41. Lin HY, Lin SP, Chen YJ, Hsu CH, Kao HA, Chen MR, Hung HY, Ho CS, Chang JH, Huang FY, Tsai TC, Lin DS, Chan WT. 2007. Clinical characteristics and survival of trisomy 13 in a medical center in Taiwan, 1985-2004. Pediatrics International 49:380-386
42. Lin HY, Lin SP, Chen YJ, Hung HY, Kao HA, Hsu CH, Chen MR, Chang JH, Ho CS, Huang FY, Shyur SD, Lin DS, Lee HC. 2006. Clinical Characteristics and Survival of Trisomy 18 in a Medical Center in Taipei, 1988-2004. American Journal of Medical Genetics Part A 140A:945-951
43. Magenis RE, Hecht F, Milham S. 1968. Trisomy 13 (D1) syndrome: studies on parental age, sex ratio and survival. The Journal ofPediatrics 73:222-228
44. Martyn CN, Osmond C, Edwardson JA, Barker DJP, Harris EC, Lacey RF. 1989. Geographical relation between Alzheimer's disease and aluminium in drinking water. The Lancet 333:61-62
45. Massenet D, Fall D, Diop M, Tall SA, Huttinger E, Riveau G. 2013. Spatiotemporal distribution of tuberculosis cases in the city of Saint-Louis Senegal from 2008-2011. Revue D Epidemiologie et de Sante Publique 61;5:421-428
46. McNally RJQ, Rankin J, Shirley MDF, Rushton SP, Pless-Mulloli T. 2008. Space-time analysis of Down syndrome: results consistent with transient pre-disposing contagious agent. International Journal of Epidemiology 37;5:1169-1179.
47. Morris JK, Savva GM. 2008. The risk of fetal loss following a prenatal diagnosis of trisomy 13 or trisomy 18. American Journal of Medical Genetics Part A 146A:7;827-832
48. Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V. 2003. Dead bird clusters as an early warning system for West Nile virus activity. Emerging Infectious Diseases 9;6:641-646.
49. Nagarwalla N. 1996. A Scan Statistic with a Variable Window. Statistics In Medicine 15;845-850
50. National Down's Syndrome Screening Programme for England. A Handbook for Staff. 2004
51. Naus J. 1965. The Distribution of the Size of the Maximum Number of Points on a Line. Journal of the American Statistical Association 60:310;532-538
52. Nembhard WN, Waller DK, Sever LE, Canfield MA. 2001. Patterns of first-year survival among infants with selected congenital anomalies in Texas,1995-1997. Teratology 64:267–275.
53. Nicolaides KH, Azar G, Byrne D, Mansur C, Marks K. 1992. Fetal nuchal translucency: ultrasound screening for chromosomal defects in first trimester of pregnancy. British Medical Journal 304;867-869.

54. Nicolaidis P, Petersen MB. 1998. Origin and mechanisms of non-disjunction in human autosomal trisomies. Human Reproduction 13;313-319
55. Niedrist D, Riegel M, Achermann JM, Schinzel A. 2006. Survival with trisomy 18 – Data from Switzerland. American Journal of Medical Genetics Part A 140A;9:952-959
56. Office for National Statistics. 2013. Characteristics of Mother 1, England and Wales, 2011.
57. Okumura 2003. The material flow of radioactive cesium-137 in the U.S. 2000. United States Environmental Protection Agency online, www.epa.gov
58. Openshaw S, Craft AW, Charlton M, Birch JM. 1988. Investigation of Leukaemia Clusters by use of a Geographical Analysis Machine. Lancet 331;272-273
59. Pai GS, Valle D, Thomas G. 1978. Cluster of Trisomy 13 live births. Lancet 1:613.
60. Palomaki GE, Haddow JE, Knight GJ, Wald NJ, Kennard A, Canick JA, Saller DN, Blitzer MG, Dickerman LH, Fisher R, Hansmann D, Hansmann M, Luthy DA, Summers AM, Wyatt P. 1995. Risk-based prenatal screening for trisomy-18 using alpha-fetoprotain, unconjugated estriol and human chorionic gonadotrophin. Prenatal Diagnosis 15;8:713-723
61. Patau K, Smith DW, Therman E, Inhorn SL, Wagner HP. 1960. Multiple Congenital Anomaly Caused by an Extra Autosome. Lancet 1;790-3.
62. Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environmental and Ecological Statistics 2004, 11;183-197
63. Penrose L. 1933. The relative effects of paternal and maternal age in mongolism. Journal of Genetics 27;219-224
64. Public Health England. www.screening.nhs.uk. 2013
65. Ramsay CN, Ellis PM, Zealley H. 1991. Down's Syndrome in the Lothian region of Scotland. Biomedicine& Pharmacotherapy 45;267-272.
66. Rasmussen SA, Wong LY, Yang Q, May KM, Friedman JM. 2003. Population-based analyses of mortality in trisomy 13 and trisomy 18. Pediatrics 111:777-784
67. Redman HC, McClellan RO, Jones RK, Boecker BB, Chiffelle TL, Pickrell JA, Rypka EW. 1972. Toxicity of 137 CsCl in the beagle, Early biological effects. Radiation Research 50;3:629-648.
68. Revenkova E, Eijpe M, Heyting C, Hodges CA, Hunt PA, Liebe B, Scherthan H, Jessberger R. 2004. Cohesin SMC1β is required for meiotic chromosome dynamics, sister chromatid cohesion and DNA recombination. Nature Cell Biology 6;555-562
69. Rhoads GG, Jackson LG, Schlesselman SE, Delacruz FF, Desnick RJ, Golbus MS, Ledbetter DH, Lubs HA, Mahoney MJ, Pergament E, Simpson JL, Carpenter RJ, Elias S, Ginsberg NA, Goldberg JD, Hobbins JC, Lynch L, Shiono PH, Wapner RJ, Zachary JM. 1989. The safety

and efficacy of chorionic villus sampling for early prenatal-diagnosis of cytogenetic abnormalities. New England Journal of Medicine 320;10:609-617.

70. Ripley BD. Modelling Spatial Patterns. Journal of the Royal Statistical Society 1977, 39:2;172-212
71. Rodeck CH, Whittle MJ (2008). Fetal medicine, basic science and clinical practice. Churchill Livingstone, in print, p837.
72. Root S, Carey JC. 1994. Survival in trisomy 18. American Journal of Medical Genetics 49:170-174
73. SaTScan. http://www.SaTScan.org/
74. Savva GM, Morris JK. 2009. Ascertainment and accuracy of Down syndrome cases reported in congenital anomaly registers in England and Wales. Archives of Disease in Childhood, Fetal, and Neonatal Edition 94:23-27
75. Savva GM, Morris JK, Mutton DE, Alberman E. 2006. Maternal age-specific fetal loss rates in Down syndrome pregnancies. Prenatal Diagnosis 26;499-504
76. Savva GM, Walker K, Morris JK. 2010. The maternal age-specific prevalence of trisomies 13 and 18 compared to trisomy 21 (Down syndrome). Prenatal Diagnosis 30;57-64
77. Schmeidel S, Jacquez GM, Blettner M, Schuz J. 2011. Spatial clustering of leukaemia and type 1 diabetes in children in Denmark. Cancer Causes & Control 22;6:849-857
78. Sperling K, Pelz J, Wegner RD. 1994. Significant increase in trisomy-21 in Berlin 9 months after the Chernobyl reactor accident – temporal correlation or causal relation. British Medical Journal 309;158-162
79. Steuerwald N, Cohen J, Herrera RJ, Sandnlinas M, Brenner CA. 2001. Association between spindle assembly checkpoint expression and maternal age in human oocytes. Molecular Human Reproduction 7;49-55
80. Svyatova GS, Abil'dinova GZ, Berezina GM. 2002. Results of a Cytogenetic Study of Populations with Different Radiation Risks in the Semipalatinsk Region. Russian Journal of Genetics 38;3:294-299.
81. Tabor A, Vestergaard CHF, Lindegaard O. 2009. Fetal loss rate after chorionic villus sampling and amniocentesis: an 11-year national registry study. Ultrasound in Obstetrics & Gynecology 34;1:19-24
82. Takahashi K, Kulldorff M, Tango T, Yih K. A flexibly-shaped space-time scan statistic for disease outbreak detection and monitoring. International Journal of Health Geographics 2008, 7:14
83. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. International Journal of Health Geographics 2005, 4:11;4-11
84. The National Down Syndrome Cytogenetic Register 2004 Annual Report

85. The National Down Syndrome Cytogenetic Register 2005 Annual Report
86. The National Down Syndrome Cytogenetic Register 2006 Annual Report
87. The National Down Syndrome Cytogenetic Register 2007 Annual Report
88. The National Down Syndrome Cytogenetic Register 2008 Annual Report
89. The National Down Syndrome Cytogenetic Register 2009 Annual Report
90. The National Down Syndrome Cytogenetic Register 2010 Annual Report
91. Torfs CP, Katz EA, Bateson TF, Lam PK, Curry CJR. 1996. Maternal medications and environmental exposures as risk factors for gastroschisis. Teratology 54:84-92
92. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for Clusters of Disease; Application to Leukemia Incidence in Upstate New York. American Journal of Epidemiology 1990, 132 (Supp1);136-143
93. UK National Screening Committee. 2011. Screening in England 2010-11 Annual Report.
94. Vendola C, Canfield M, Daiger SP, Gambello M, Hashmi SS, King T, Noblin SJ, Waller DK, Hecht JT. 2010. Survival of Texas Infants Born With Trisomies 21, 18, and 13. American Journal of Medical Genetics Part A 152A;2:360-366
95. Verger P. 1997. Down Syndrome and Ionizing Radiation. Health Physics 73(6):882-893.
96. Wald NJ, Kennard A, Hackshaw A, McGuire A. 1997. Antenatal screening for Down's syndrome. J Med Screen 4:181-246.
97. Wald NJ, Watt HC, Hackshaw AK. 1999. Integrated Screening for Down's Syndrome Based on Tests Performed during the First and Second Trimesters. New England Journal of Medicine 341:461-467
98. Warburton D, Kline J, Stein Z, Susser M (1977). Trisomy cluster in New York. Lancet 2:201.
99. Weber W. 1967. Survival and the sex ratio in trisomy 17-18. American Journal of Human Genetics 19:369-377
100. Werler MM, Sheehan JE, Mitchell AA. 2002. Maternal medication use and risks of gastroschesis and small intestinal atresia. American Journal of Epidemiology 155;1:26-31
101. Whittemore AS, Friend N, Brown BW, Holly EA. A Test to Detect Clusters of Disease. Biometrika 1987, 74:3;631-635
102. Wilson RD, Johnson JM, Dansereau J, Singer J, Drinnan SL, Winsor EJT, Soanes S, Kalousek D, Hillier J, Ho MF, Brown J, Wyatt P, Farrell S, O'Farrell J, Chadwick D, Natale R, Mitchell C, Fan YS, Mohide P, Beecroft ML, Tomkins DJ, Carter R, Benzie R, Maclean B, Wang HS, Dallaire L, DesRochers M, Tihy F, Eydoux P, Provencher S, Cartier L, Armson BA, Hogg H, Welch JP, Harman C, Babaian K, Dawson A, Okun N, Park M, Lin JCC, Robinson G, Summers A. 1998. Randomised trial to assess safety amid fetal outcome of early and midtrimester amniocentesis. Lancet 351:242-247.

103. www.lucinafoundation.org
104. Wyllie JP, Wright MJ, Burn J, Hunter S. 1994. Natural history of trisomy 13. Archives of Disease in Childhood 71:353-345
105. Yang Q, Rasmussen SA, Friedman JM. 2002. Mortality associated with Down's syndrome in the USA from 1983 to 1997: a population-based study. Lancet 359;1019-1025
106. Zatsepin I, Verger P, Robert-Gnansia E, Gagniere B, Tirmarche M, Khmel R, Babicheva I, Lazjuk G. 2007. Down syndrome time-clustering in January 1987 in Belarus: Link with the Chernobyl accident? Reproductive Toxicology 24;289-295
107. Zimmermann B, Hill M, Gemelos G, Demko Z, Banjevic M, Baner J, Ryan A, Sigurjonsson S, Chopra N, Dodd M, Levy B, Rabinowitz M. 2012. Noninvasive prenatal aneuploidy testing of chromosomes 13, 18, 21, X and Y, using targeted sequencing of polymorphic loci. Prenatal Diagnosis 32;13:1233-1241.

Appendices

Appendix 1. Data Cleaning Tables

Table 2.9. Number of diagnoses for trisomies 21, 18, and 13 combined from each cytogenetic laboratory by Government Office Region.

	Government Office Region									
Lab	**A**	**B**	**D**	**E**	**F**	**G**	**H**	**J**	**K**	**W**
1	0	1	0	3	7	1	1	1	1286	15
2	0	0	5	383	1	2	0	0	0	0
3	0	2	1	0	0	289	1085	9	1	0
4	1	1068	7	19	9	1	3	0	1	1
5	776	89	6	1	0	0	0	1	0	0
6	0	0	0	0	1	11	357	453	1	1
7	0	1	0	2	1	0	2	654	268	3
8	1	0	397	163	1	0	1	0	0	0
9	5	31	986	4	0	0	2	1	1	0
13	0	0	0	259	5	1	1	0	0	1
14	0	2	1	0	0	1	501	351	1	0
15	0	601	0	0	4	1	1	0	0	54
16	0	2	0	2	3	406	944	10	1	1
19	0	0	0	0	1	0	48	22	0	0
22	8	4	2	44	1597	3	1	1	2	20
23	0	0	0	205	4	20	3	722	117	0
24	0	0	0	29	0	550	0	1	1	0
26	0	0	0	1	0	169	1	0	0	0
27	0	1	0	0	6	1	0	0	3	654
28	0	1	0	0	0	0	0	0	0	0
30	2	8	4	16	15	406	1109	571	24	36

Table 2.10. Proportion of diagnoses for trisomies 18 and 13 combined from each Government Office Region in each Cytogenetic Laboratory.

	Government Office Region									
Lab	**A**	**B**	**D**	**E**	**F**	**G**	**H**	**J**	**K**	**W**
1	0	0.1	0	0.2	0.5	0.1	0.1	0.1	97.8	1.1
2	0	0	1.3	98	0.3	0.5	0	0	0	0
3	0	0.1	0.1	0	0	20.8	78.2	0.6	0.1	0
4	0.1	96.2	0.6	1.7	0.8	0.1	0.3	0	0.1	0.1
5	88.9	10.2	0.7	0.1	0	0	0	0.1	0	0
6	0	0	0	0	0.1	1.3	43.3	55	0.1	0.1
7	0	0.1	0	0.2	0.1	0	0.2	70.2	28.8	0.3
8	0.2	0	70.5	29	0.2	0	0.2	0	0	0
9	0.5	3	95.7	0.4	0	0	0.2	0.1	0.1	0
13	0	0	0	97	1.9	0.4	0.4	0	0	0.4
14	0	0.2	0.1	0	0	0.1	58.5	41	0.1	0
15	0	90.9	0	0	0.6	0.2	0.2	0	0	8.2
16	0	0.1	0	0.1	0.2	29.7	69	0.7	0.1	0.1
19	0	0	0	0	1.4	0	67.6	31	0	0
22	0.5	0.2	0.1	2.6	94.9	0.2	0.1	0.1	0.1	1.2
23	0	0	0	19.1	0.4	1.9	0.3	67.4	10.9	0
24	0	0	0	5	0	94.7	0	0.2	0.2	0
26	0	0	0	0.6	0	98.8	0.6	0	0	0
27	0	0.2	0	0	0.9	0.2	0	0	0.5	98.3
28	0	100	0	0	0	0	0	0	0	0
30	0.1	0.4	0.2	0.7	0.7	18.5	50.6	26.1	1.1	1.6

Highlighted numbers show proportions which were sufficiently high to impute government office region from cytogenetic lab.

Table 2.11. Trisomy 18 cases assigned to a government office region using cytogenetic lab information.

Case	Cytogenetic Lab	Government Office Region
31488	Lab 1	K
31512	Lab 4	B
36159	Lab 27	W
39405	Lab 4	B
40656	Lab 27	W
41195	Lab 1	K
49350	Lab 9	C

Table 2.12. Trisomy 13 Cases Assigned to a Government Office Region using cytogenetic lab information.

Case	Cytogenetic Lab	Government Office Region
31586	4	B
38176	27	W
42583	15	B
50645	4	B
50652	9	D

Table 2.13. Gestational ages for all CVS diagnoses between 2004 and 2010.

Week of Gestation	% Diagnosed
< 11	0.48
11	10.20
12	42.76
13	32.80
14	9.31
> 14	4.45
Total	100

Table 2.14. Gestational ages for all amniocentesis diagnoses between 2004 and 2010.

Week of Gestation	% Diagnosed
< 15	4.38
15	13.46
16	19.27
17	15.16
18	8.94
19	5.21
20	11.34
> 20	22.24
Total	100

Table 2.15. Trisomy 18 cases assigned sample dates using the immediately adjacent cases.

				Previous Sample		Following Sample			
Case	Lab ID	Lab	Sample Year	Lab ID	Date	Lab ID	Date	Accuracy (days)	Sample Date Estimated
30954	C040222	30	2004	C040084	08/01/2004	C040316	28/01/2004	21	-
31564	A043835	30	2004	A0403567	29/06/2004	A0403849	13/07/2004	15	-
33542	C0503530	22	2005	C0502813	07/02/2005	C0503656	16/02/2005	10	11/02/2005
33546	C0507590	22	2005	C0507156	31/03/2005	C0507840	08/04/2005	9	04/04/2005
33565	C050002	30	2005	-	-	C050058	10/01/2005	10	05/01/2005
34733	C050522	3	2005	C050482	19/09/2005	C050527	19/10/2005	31	-
34772	F0518810	22	2005	F0518427	08/08/2005	F0519167	15/08/2005	8	11/08/2005
35445	P051118	30	2005	P051104	29/12/2005	-	-	3	30/12/2005
36116	F006206	9	2006	F005406	22/02/2006	F006506	01/03/2006	8	25/02/2006
36157	F060093	24	2006	F060068	02/02/2006	F060110	23/02/2006	20	-
36172	P060021	30	2006	-	-	P060036	09/01/2006	9	05/01/2006
37203	P060457	30	2006	P060375	20/04/2006	P060623	06/07/2006	78	-
37204	P060505	30	2006	P060375	20/04/2006	P060623	06/07/2006	78	-
37807	T0616334	22	2006	T0612441	14/05/2006	T0620785	14/08/2005	93	-
37808	T0622299	22	2006	T0620785	14/08/2005	T0622735	06/09/2005	28	-
38058	F031406	9	2006	F030506	13/12/2006	-	-	19	-
40571	AF071433	4	2007	AF071410	15/10/2007	AF071440	19/10/2007	5	17/10/2007
46885	09CV0174	23	2009	09CV0172	26/08/2009	09CV0178	03/09/2009	9	30/08/2009

Table 2.16. Trisomy 13 cases assigned sample dates using the immediately adjacent cases.

				Previous Sample		Following Sample			
Case	Lab ID	Lab	Sample Year	Lab ID	Date	Lab ID	Date	Accuracy (days)	Sample Date Estimate
30369	T0425969	22	2004	T0419608	21/09/2004	-	-	102	-
34294	P050470	30	2005	P050403	19/05/2005	P050482	10/06/2005	23	-
34825	C050566	3	2005	C050565	08/11/2005	C050597	28/11/2005	21	-
34827	AF051490	4	2005	AF051489	30/09/2005	AF051565	17/10/2005	18	-
34845	T0525427	22	2005	T0525340	28/10/2005	T0526668	14/11/2005	18	-
34858	A0503797	30	2005	A0503591	04/08/2005	A0503814	23/08/2005	20	-
35482	A0505159	30	2005	A0505010	18/11/2005	A0505203	05/12/2005	19	-
36221	P060015	30	2006	-	-	P060036	09/01/2006	9	05/01/2006
37254	C060959	30	2006	C060931	13/04/2006	C061000	21/04/2006	9	17/04/2006
38188	P060702	30	2006	P060623	06/07/2006	P060728	03/08/2006	29	-
38192	P061057	30	2006	P061055	15/11/2006	P061059	16/11/2006	2	15/11/2006

Appendix 2. R Code for Temporal Scanning Program

```
begTime<- Sys.time()

data=read.csv("T18K.csv", header = TRUE, sep = ",", quote="\"", dec=".", fill=TRUE)

sample = sum(data[,2])

days=2557 # 7 years from Jan 01 2004 to Dec 31 2010, including 2 leap years.
sdays=nrow(data)
min=2 # Sets the minimum cluster size.
div=5 # Sets the maximum cluster size as a division of the total number of cases.

N=sample
remain=0
counter=1

case<- array(0, dim=sample)
cases<- array(0, dim=sample)
results<- matrix(0, nrow=sample, ncol=3)
likely<- matrix(0, nrow=10, ncol=5)

dminhold = 0

dmin<- matrix(1, nrow=sample, ncol=2)

for(i in 1:sdays){

repeat{
case[counter]=data[i,1]

data[i,2]=data[i,2]-1

counter=counter+1

if(data[i,2]==0) break

}

}

for(i in 1:sample){

cases[i]=case[i]/days

}

for(j in 2:sample){

for(i in 1:sample){

if(i+j<=(sample+1)) dminhold=(cases[i+j-1]-cases[i])
```

```
if(dminhold<dmin[j,1]) dmin[j,2]=i

if(dminhold<dmin[j,1]) dmin[j,1]=dminhold

}

}

for(j in min:(sample/div)){

results[j,1]=((j/N)^j)*(((N-j)/N)^(N-j))*((1/dmin[j,1])^j)*((1/(1-dmin[j,1]))^(N-j))

results[j,2]=dmin[j,2]

results[j,3]=j

if(results[j,1]>(5e+100)) results[j,1]=0

if(results[j,1]>likely[10,1]) likely[10,2]=results[j,2]

if(results[j,1]>likely[10,1]) likely[10,3]=results[j,3]

if(results[j,1]>likely[10,1]) likely[10,4]=case[likely[10,2]]

if(results[j,1]>likely[10,1]) likely[10,5]=case[likely[10,2]+likely[10,3]-1]

if(results[j,1]>likely[10,1]) likely[10,1]=results[j,1]

likely<- likely[order(likely[,1]),,drop=FALSE]

likely<- likely[rev(order(likely[,1])),]

}

likely # Returns a table of the 10 strongest clusters.

##### Calculates P-Value using Monte Carlo Simulation

caseno = sample # number of cases.
repeats = 999 # Number of repeats.

cas<- array(0, dim=caseno)
result<- matrix(0, nrow=caseno, ncol=3)
output<- matrix(0, nrow=repeats, ncol=3)
output2 <- array(0, dim=repeats)

N=caseno

dminvalue=0

set.seed(37587) # Seed values are entered "at random" by the user.
```

```
for(k in 1:10){
output[k,1]=likely[k,1]
output[k,2]=k
}
for(k in 11:repeats){
cas<- array(0, dim=caseno)
for(i in 1:caseno){
cas[i]=as.integer(runif(1,1,days))
cas[i]=cas[i]/days
}
cas=cas[order(cas)]
holding=0
dmin<- matrix(1, nrow=caseno, ncol=2)
for(j in 2:caseno){
for(i in 1:caseno){
if(i+j<=(caseno+1)) dminhold=(cas[i+j-1]-cas[i])
if(dminhold<dmin[j,1]) dmin[j,2]=i
if(dminhold<dmin[j,1]) dmin[j,1]=dminhold
}
}
for(j in min:(caseno/div)){
result[j,1]=((j/N)^j)*(((N-j)/N)^(N-j))*((1/dmin[j,1])^j)*((1/(1-dmin[j,1]))^(N-j))
result[j,2]=dmin[j,2]
result[j,3]=j
if(result[j,1]>(5e+100)) result[j,1]=0
if(result[j,1]>holding) holding=result[j,1]
}
```

```
output[k,1]=holding

}

output<- output[order(output[,1]),,drop=FALSE]

for(i in 1:repeats){

output2[i]=output[i,1]
output[i,3]=i

}

output[order(output[,2]),,drop=FALSE]

likely

sample

runTime<- Sys.time()-begTime

runtime
```

Appendix 3. R Code for Spatial Scanning Program

Appendix 3.1. Identifying Cluster Locations

```
begTime<- Sys.time()

data=read.csv("04T13mod.csv", header = TRUE, sep = ",", quote="\"", dec=".", fill=TRUE)
pcts = nrow(data)

diameter = 10 # Maximum quadrat size of the circle (this value is actually the radius)
population= sum(data[,4]) # Total population.
cases= sum(data[,3]) # Number of total population who are also cases.
noncases=population-cases
set.seed(1) # Sets seed for the random number generator.
results<- matrix(0, nrow=(pcts*diameter), ncol=6)
counter=1
cluster<- array(0, dim=7)
holding=-1000000
pctclust<- array(0 ,dim=50)

likely<- matrix(-1000000, nrow=10, ncol=7)

xhold=0
yhold=0
dist=0

casin=0
nonin=0
totin=0

outcome=0

lc=log(cases)
lp=log(population)
lzero=log(cases)*cases + log(population-cases)*(population-cases) -
log(population)*population
```

```
for(g in 1:pcts){ # Scans through different PCT locations.

i=data[g,5]
j=data[g,6]

for(k in 1:diameter){ # Scans through different diameter sizes for sample circles.

for(l in 1:pcts){ # Uses Pythagoras' theorem to determine how far each individual is from the centre of the sample circle.

xhold=data[l,5]-i
yhold=data[l,6]-j
dist=yhold^2+xhold^2

if(dist<k^2) totin=totin+data[l,4] # Adds up how many of the population fall within the boundaries of the sample circle.
if(dist<k^2) casin=casin+data[l,3] # Adds up how many cases fall within the boundaries of the sample circle.

}

nonin=totin-casin

results[counter,1]=i
results[counter,2]=j
results[counter,3]=k
results[counter,4]=casin
results[counter,5]=nonin

lcas=log(casin)
ltot=log(totin)

if(nonin==0) outcome=lzero # Sets the outcome statistic to Lzero if there are no non-cases within the sample circle.
```

else if((casin/nonin)<((cases-casin)/(population-totin))) outcome=lzero # Sets the outcome statistic to Lzero if the ratio of cases to non-cases is greater outside the sample circle than within it.

else outcome= (lcas-ltot)*casin + (log(totin-casin) - ltot) * (totin-casin) + (log(cases-casin) - log(population-totin)) * (cases-casin) + (log(population-totin-(cases-casin)) - log (population-totin)) * (population-totin-(cases-casin))
Works out Lz for all zones where the case:non-case ratio is higher inside the sample circle than it is outside it.

results[counter,6]=outcome

if(outcome>likely[10,6]) likely[10,1]=results[counter,1]
if(outcome>likely[10,6]) likely[10,2]=results[counter,2]
if(outcome>likely[10,6]) likely[10,3]=results[counter,3]
if(outcome>likely[10,6]) likely[10,4]=results[counter,4]
if(outcome>likely[10,6]) likely[10,5]=results[counter,5]
if(outcome>likely[10,6]) likely[10,6]=results[counter,6] # If the current outcome statistic is greater than the current greatest outcome this saves all information about that cluster to a separate array.

likely<- likely[order(likely[,6]),,drop=FALSE]

likely<- likely[rev(order(likely[,6])),]

if(outcome>holding) holding=outcome

counter=counter+1

casin=0
nonin=0
totin=0
outcome=0 # Resets the counting variables.

}

```
}

for(i in 1:10){

lambda=likely[i,6]-lzero # Works out lambda by dividing the largest value of Lz by Lzero.

likely[i,7]=exp(lambda)

likely[i,1] = (likely[i,1] / 10) + 49

likely[i,2] = (likely[i,2] / 10) - 6

}

likely # Y Coordinate, X Coordinate, Radius, Number of Cases in Circle, Number of Noncases in Circle, Log of the Scan Statistic, Scan Statistic

cases

runTime<- Sys.time()-begTime

runtime
```

Appendix 3.2.Calculating P-Values

```
begTime<- Sys.time()

data=read.csv("testrc3c.csv", header = TRUE, sep = ",", quote="\"", dec=".", fill=TRUE) # Read in data file containing cases and population figures aggregated by PCT.
pcts = nrow(data)

test=1324.97 # Scan Statistic to calculate p-value for.

diameter = 10 # Maximum quadrat size of the circle (this value is actually the radius)
population= sum(data[,4]) # Total population.
cases= sum(data[,3]) # Number of total population who are also cases.
noncases=population-cases
set.seed(1) # Sets seed for the random number generator.

repeats=999 # Number of repeats to perform.
monte<- matrix(0, nrow=7, ncol=repeats)
carlo<- array(0, dim=repeats+1)
casecheck<- array(0, dim=repeats)


for(a in 1:repeats){


pctclust<- array(0 ,dim=50)
counter=1
results<- matrix(0, nrow=(pcts*diameter), ncol=6)
cluster<- array(0, dim=7)


datagen<- matrix(0, nrow=pcts, ncol=2)

holding=0

for(h in 1:pcts){
```

```
datagen[h,1]=holding+data[h,4]

holding=datagen[h,1]

}

for(h in 1:cases){
case<- sample(1:population, 1)
if(case<=datagen[1,1]) datagen[1,2]=datagen[1,2]+1
for(i in 2:pcts){
if(datagen[i-1,1]<case && case<=datagen[i,1]) datagen[i,2]=datagen[i,2]+1
}
}
for(i in 1:pcts){
data[i,3]=datagen[i,2]
}
holding=-1000000
xhold=0
yhold=0
dist=0
casin=0
nonin=0
totin=0
outcome=0
lc=log(cases)
lp=log(population)
lzero=log(cases)*cases + log(population-cases)*(population-cases) -
log(population)*population
for(g in 1:pcts){ # Scans through different PCT locations.
i=data[g,5]
j=data[g,6]

for(k in 1:diameter){ # Scans through different diameter sizes for sample circles.
```

for(l in 1:pcts){ # Uses Pythagoras' theorem to determine how far each individual is from the centre of the sample circle.

xhold=data[l,5]-i
yhold=data[l,6]-j
dist=yhold^2+xhold^2

if(dist<k^2) totin=totin+data[l,4] # Adds up how many of the population fall within the boundaries of the sample circle.
if(dist<k^2) casin=casin+data[l,3] # Adds up how many cases fall within the boundaries of the sample circle.

}

nonin=totin-casin

results[counter,1]=i
results[counter,2]=j
results[counter,3]=k
results[counter,4]=casin
results[counter,5]=nonin

lcas=log(casin)
ltot=log(totin)

if(nonin==0) outcome=lzero # Sets the outcome statistic to Lzero if there are no non-cases within the sample circle.

else if((casin/nonin)<((cases-casin)/(population-totin))) outcome=lzero # Sets the outcome statistic to Lzero if the ratio of cases to non-cases is greater outside the sample circle than within it.

else outcome= (lcas-ltot)*casin + (log(totin-casin) - ltot) * (totin-casin) + (log(cases-casin) - log(population-totin)) * (cases-casin) + (log(population-totin-(cases-casin)) - log (population-totin)) * (population-totin-(cases-casin))

```
# Works out Lz for all zones where the case:non-case ratio is higher inside the sample circle
than it is outside it.

results[counter,6]=outcome

if(outcome>holding) cluster[1]=results[counter,1]
if(outcome>holding) cluster[2]=results[counter,2]
if(outcome>holding) cluster[3]=results[counter,3]
if(outcome>holding) cluster[4]=results[counter,4]
if(outcome>holding) cluster[5]=results[counter,5]
if(outcome>holding) cluster[6]=results[counter,6] # If the current outcome statistic is greater
than the current greatest outcome this saves all information about that cluster to a separate
array.

if(outcome>holding) holding=outcome
counter=counter+1

casin=0
nonin=0
totin=0
outcome=0 # Resets the counting variables.

}
}

lambda=cluster[6]-lzero # Works out lambda by dividing the largest value of Lz by Lzero.
cluster[7]=exp(lambda)
for(i in 1:7){
monte[i,a] = cluster[i]

}

carlo[a]=cluster[7]

casecheck[a]=cases
}
```

```
carlo[repeats+1]=test
carlo<- carlo[order(carlo),drop=FALSE]
rank=1
for(i in 1:(repeats)){
if(test>carlo[i]) rank=rank+1
}

rank
runTime<- Sys.time()-begTime
runTime
```