Spatial epidemiology
Four tutorial style BRG seminars, Thursdays 4pm

Gavin Shaddick
Department of Mathematical Sciences, University of Bath, UK

gavin@stat.ubc.ca / shaddick@stat.ubc.ca

www.stat.ubc.ca/∼gavin

With many many thanks to Prof. J. Wakefield,
University of Washington

Spatial Textbooks:

Elliott, P., Wakefield, J., Best, N. and Briggs, D. (2000). *Spatial Epidemiology: Methods and Applications*, Oxford University Press.

Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*, Wiley, New York.

Epidemiology Textbooks:

Breslow, N.E. and Day, N.E. (1980, 1987). *Statistical Methods in Cancer Research. Volume I: The Analysis of Case-Control Studies, Volume II: The Design and Analysis of Cohort Studies*, IARC Scientific Publications Nos. 32 and 82, Lyon.

Rothman, K. and Greenland, S. (1998). *Modern Epidemiology, Second Edition*, Lipincott-Raven.

Section 1: Preliminaries

Motivation of the need for spatial epidemiology; Types of spatial study and examples; Overview of epidemiological framework; Overview of Statistical Techniques; `R`.

Section 2: Disease Mapping

Non-spatial and spatial smoothing models; Bayesian inference and computation (`WinBUGS` software); Statistical models; Examples.

Section 3: Spatial Regression

Simple approaches via logistic and Poisson regression; The ecological fallacy; Sophisticated approaches; Geostatistical regression for point data; Methods for point sources with count data; Examples.

Spatial Epidemiology

Epidemiology: The study of the distribution, causes and control of diseases in human populations.

Disease risk depends on the classic epidemiological triad of person (genetics/behavior), place and time – spatial epidemiology focuses on the second of these.

Place is a surrogate for exposures present at that location, e.g. environmental exposures in water/air/soil, or the lifestyle characteristics of those living in particular areas.

Types of Data

An important distinction is whether the data arise as:

- *Point data* in which "exact" residential locations exist for cases and non-cases, or
- *Count data* in which aggregation (typically over administrative units) has been carried out. These data are *ecological* in nature, in that they are collected across groups, in spatial studies the groups are geographical areas.

We will only consider non-infectious diseases, though many of the issues transfer to infectious diseases.

Need for Spatial Methods

All epidemiological studies are spatial!

When do we need to "worry", i.e. acknowledge the spatial component?

- Are we explicitly interested in the spatial pattern of disease incidence? e.g. disease mapping, cluster detection.
- Is the clustering a nuisance quantity that we wish to acknowledge, but are not explicitly interested in? e.g. spatial regression.

If we are interested in the spatial pattern then, if the data are not a complete enumeration, we clearly we need the data to be randomly collected in space

Growing interest in spatial epidemiology due to:

- Public interest in effects of environmental "pollution", *e.g.* Sellafield, UK.
- Development of statistical/epidemiological methods for investigating disease "clusters".
- Epidemiological interest in the existence of large/medium spread in chronic disease rates across different areas.
- Data availability: collection of health data at different geographical scales.
- Increase in computing power and methods (Geographical Informations Systems).

It is convenient to distinguish three types of study:

1. Disease mapping – provide information on a measure of disease occurrence across space. Mapping studies exploit spatial dependence in order to smooth rates and provide better predictions.
2. Spatial regression – specifically interested in the association between disease risk and exposures of interest. For count data we examine the association between risk and exposures at the area level via ecological regression; Poisson regression is the obvious framework. For point data logistic regression is the obvious approach though we may also use "geostatistical" methods. In this context spatial dependence is a hindrance to the use of standard statistical tools (and interpretation is difficult due to the potential for "confounding by location").

3. Clustering/Cluster detection – the former examines the tendency for disease risk (or better to think of residual risk, after controlling for population distribution, and important predictors of disease that vary by area such as age and race) to exhibit "clumpiness", while the latter refers to on-line surveillance or retrospective analysis, to reveal "hot spots". Here understanding the form of the spatial dependence is the aim.

Disease Mapping

Aims:

- Simple description – a visual summary of geographical risk.
- Provide estimates of risk by area to inform public health resource allocation.
- Give clues to etiology via informal examination of maps with exposure maps, components of spatial versus non-spatial residual variability may also provide clues to source of variability (e.g. environmental exposures usually have spatial structure). The formal examination is carried out via spatial regression.
- In general mapping is based on count data (which is more routinely available) – may also be carried out with point data but much less common (case-control studies are explicitly carried out to examine an exposure of interest, and cannot inform on risk without additional information).

- Provide a context within which specific studies may be placed. For example:
  - Surveillance of disease registries will be greatly helped if we have a knowledge of the variability in residual spatial risk, and the nature of that variability (spatial versus non-spatial), i.e. what is the "null" distribution (distribution in absence of a "hot spot").
  - Regression will be aided if we have a "prior" on the background variability.
- More recently there has been increased interest in statistical models for disease mapping in *time and space*.

Example: Scottish Lip Cancer Data

Incidence rates of lip cancer in males in 56 counties of Scotland, registered in 1975–1980. These data were originally reported in the mapping atlas of Kemp, Boyle, Smans and Muir (1985).
The form of the data is:

- Observed and "expected" number of cases (based on the county age populations, details shortly) – allows the calculation of the standardized morbidity ratio, the ratio of the observed to the expected cases.
- A covariate measuring the proportion of the population engaged in agriculture, fishing, or forestry (AFF).
- The projections of the longitude and latitude of the area centroid, and the "position" of each county expressed as a list of adjacent counties.

County No. i	Obs Cases Y_i	Exp Cases E_i	Prop AFF	SMR	Project N (km)	Project E (km)	Adjacent Counties
1	9	1.4	0.16	6.43	834.7	162.2	5,9,19
2	39	8.7	0.16	4.48	852.4	385.8	7,10
3	11	3.0	0.10	3.67	946.1	294.0	12
4	9	2.5	0.24	3.60	650.5	377.9	18,20,28
5	15	4.3	0.10	3.49	870.9	220.7	1,12,19
6	8	2.4	0.24	3.33	1015.2	340.2	Island
7	26	8.1	0.10	3.21	842.0	325.0	2,10,13,16,17
8	7	2.3	0.07	3.04	1168.9	442.2	Island
9	6	2.0	0.07	3.00	781.4	194.5	1,17,19,23,29
...							
47	2	5.6	0.01	0.36	640.8	277.0	24,31,46,48,49,53
48	3	9.3	0.01	0.32	654.7	282.0	24,44,47,49
49	28	88.7	0.00	0.32	666.7	267.8	38,41,44,47,48,52,53,54
50	6	19.6	0.01	0.31	736.5	342.2	21,29
51	1	3.4	0.01	0.29	678.9	274.9	34,38,42,54
52	1	3.6	0.00	0.28	683.7	257.8	34,40,49,54
53	1	5.7	0.01	0.18	646.6	265.6	41,46,47,49
54	1	7.0	0.01	0.14	682.3	267.9	34,38,49,51,52
55	0	4.2	0.16	0.00	640.1	321.5	18,24,30,33,45,56
56	0	1.8	0.10	0.00	589.9	322.2	18,20,24,27,55

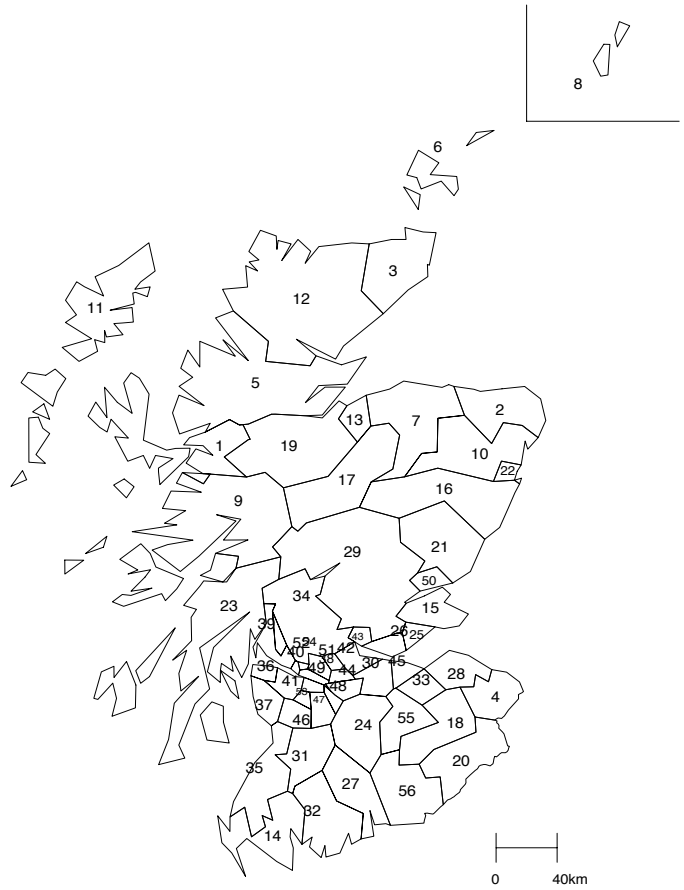

Figure: Labels for 56 counties of Scotland.

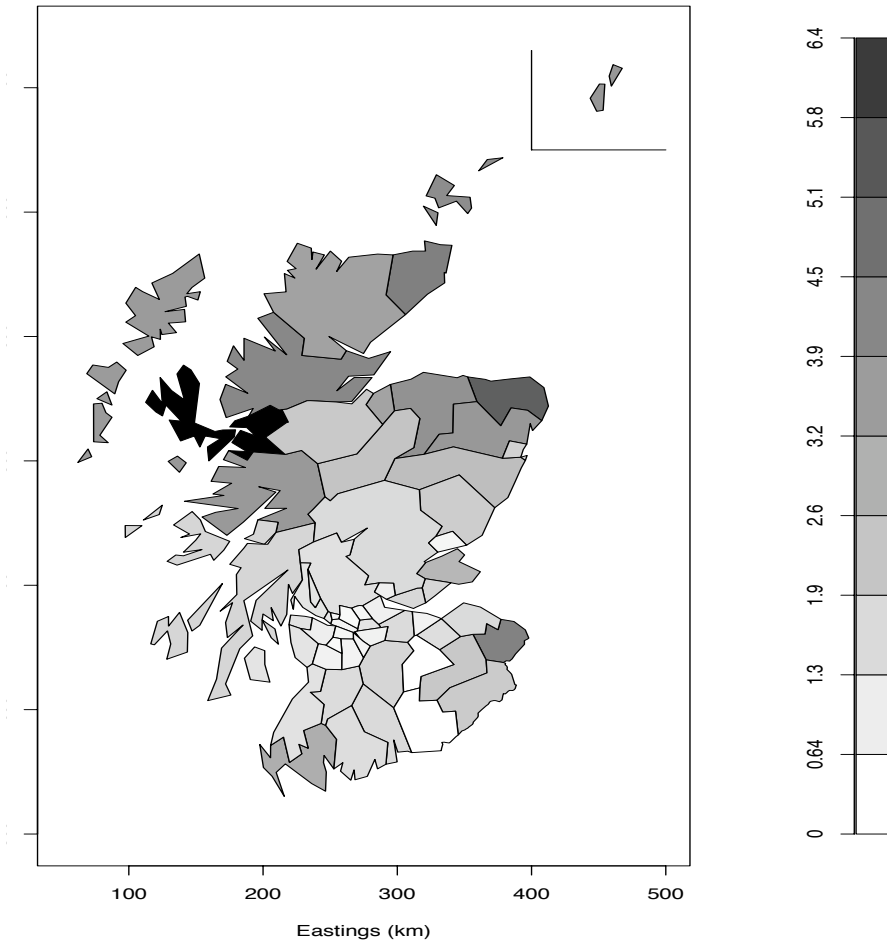

Figure: SMRs for male lip cancer in 56 counties of Scotland.

Example: Lung and Brain cancer in the North-West of England (Chapter 7, EWBB)

This study was used as an illustration of smoothing techniques using a variety of hierarchical models.

Two tumors were chosen to contrast mapping techniques for relatively non-rare (lung), and relatively rare (brain) cancers.

The absence of information on smoking means that for lung cancer in particular the analysis should be viewed as illustrative only (since a large fraction of the residual variability would disappear if smoking information were included).

This is important point – residual spatial dependence is induced by missing variables that are predictive of disease outcome (or data errors/model misspecification).

Study details:

- Study period is 1981–1991.
- Incidence data by postcode, but the analysis is carried out at the ward level of which there are 144 in the study region. For brain cancer the median number of cases per ward over the 11 year period is 6 with a range of 0 to 17. For lung the median number is 20 with range 0–60.
- "Expected counts" were based on ward-level populations from the 1991 census, by 5-year age bands and sex.

The SIRs are shown in Figures 3 and the smoothed rates for lung and brain in Figures 4 5, respectively.
Notice that for lung the smoothed area-level relative risk estimates are not dramatically different from the raw versions in Figure 3(a) – the large number of cases here mean that the raw SIRs are relatively stable. For brain we see a much greater smoothing of the estimates as compared to the raw relative risks in Figure 3(b).

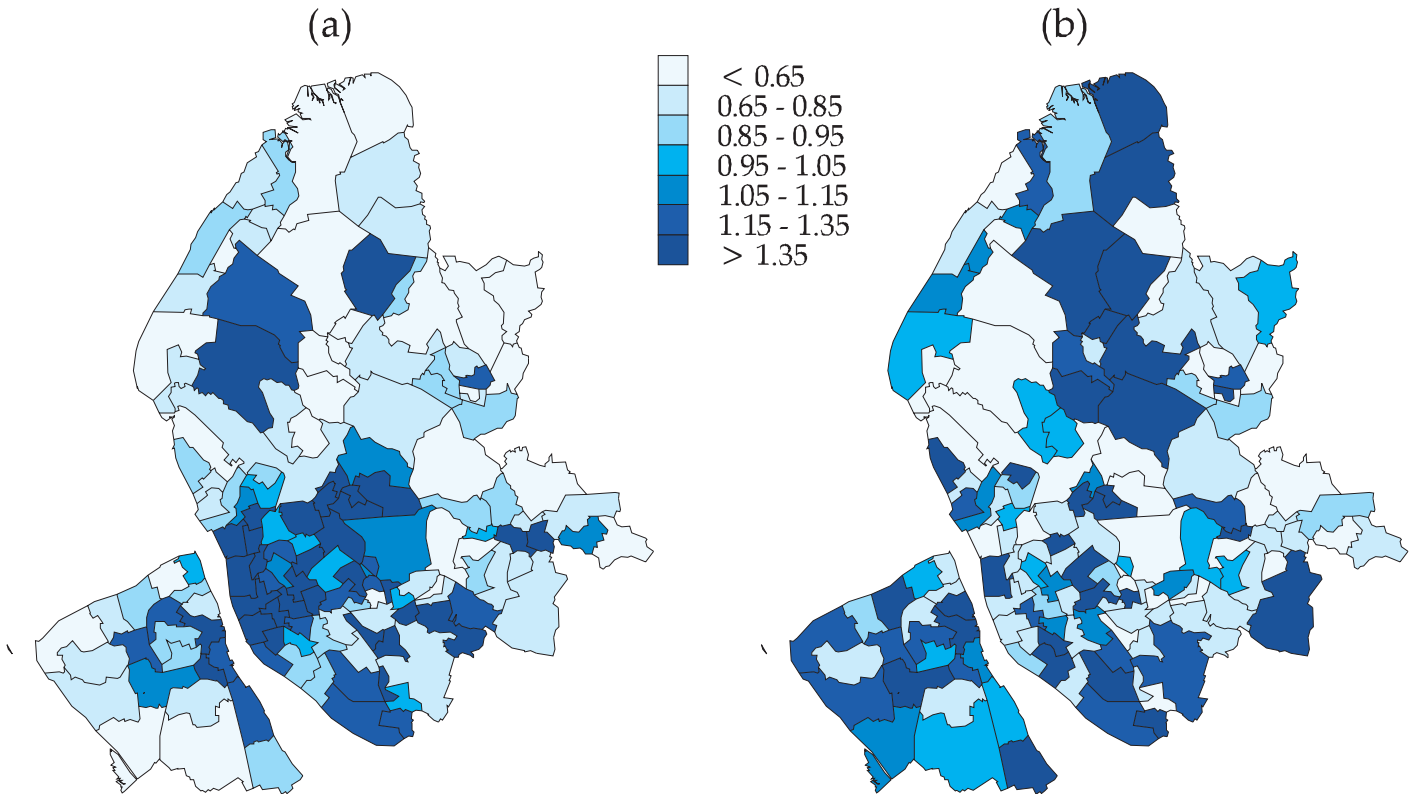

Figure: SIRs for (a) lung cancer, and (b) brain cancer.

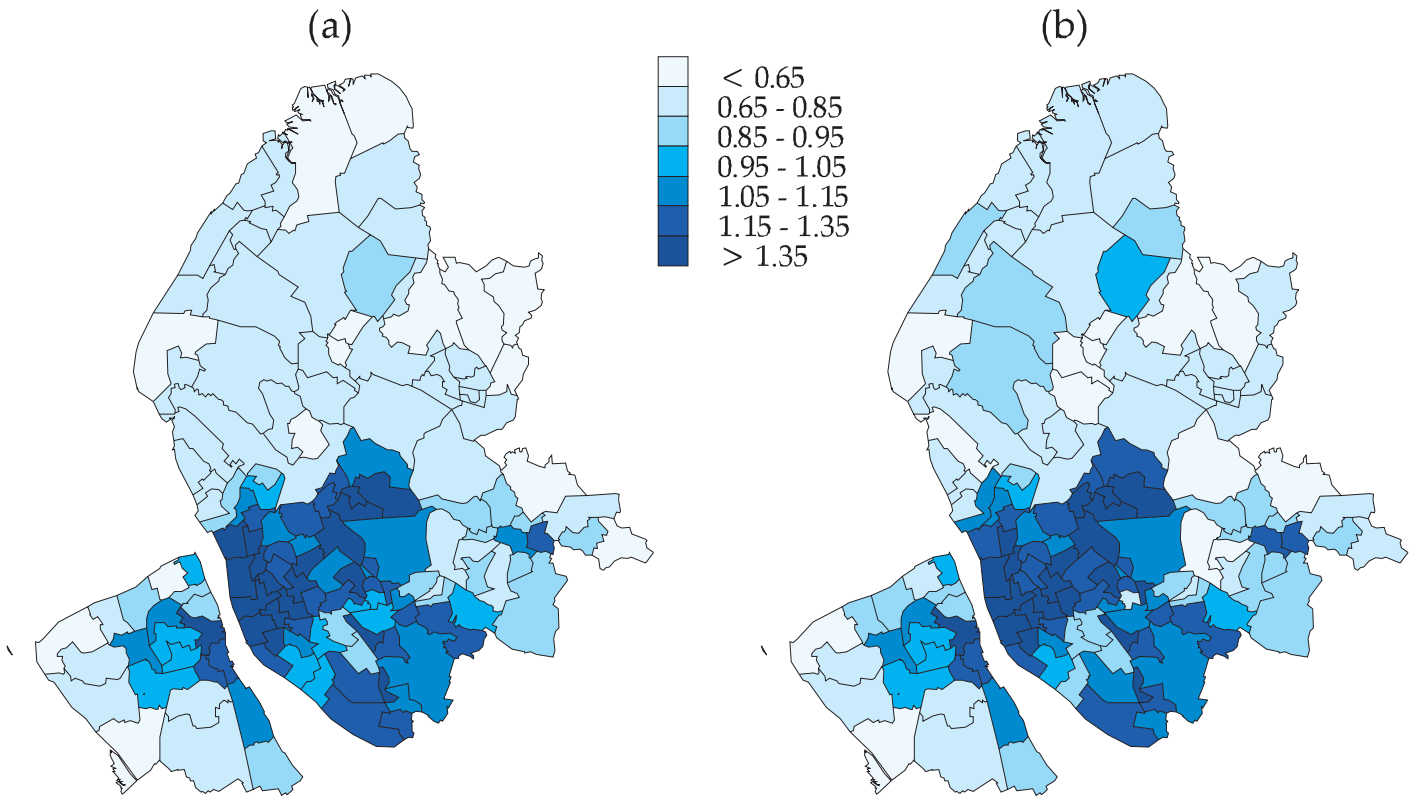

Figure: Smoothed SIRs for lung cancer under (a) a conditional spatial model, and (b) a marginal spatial model.

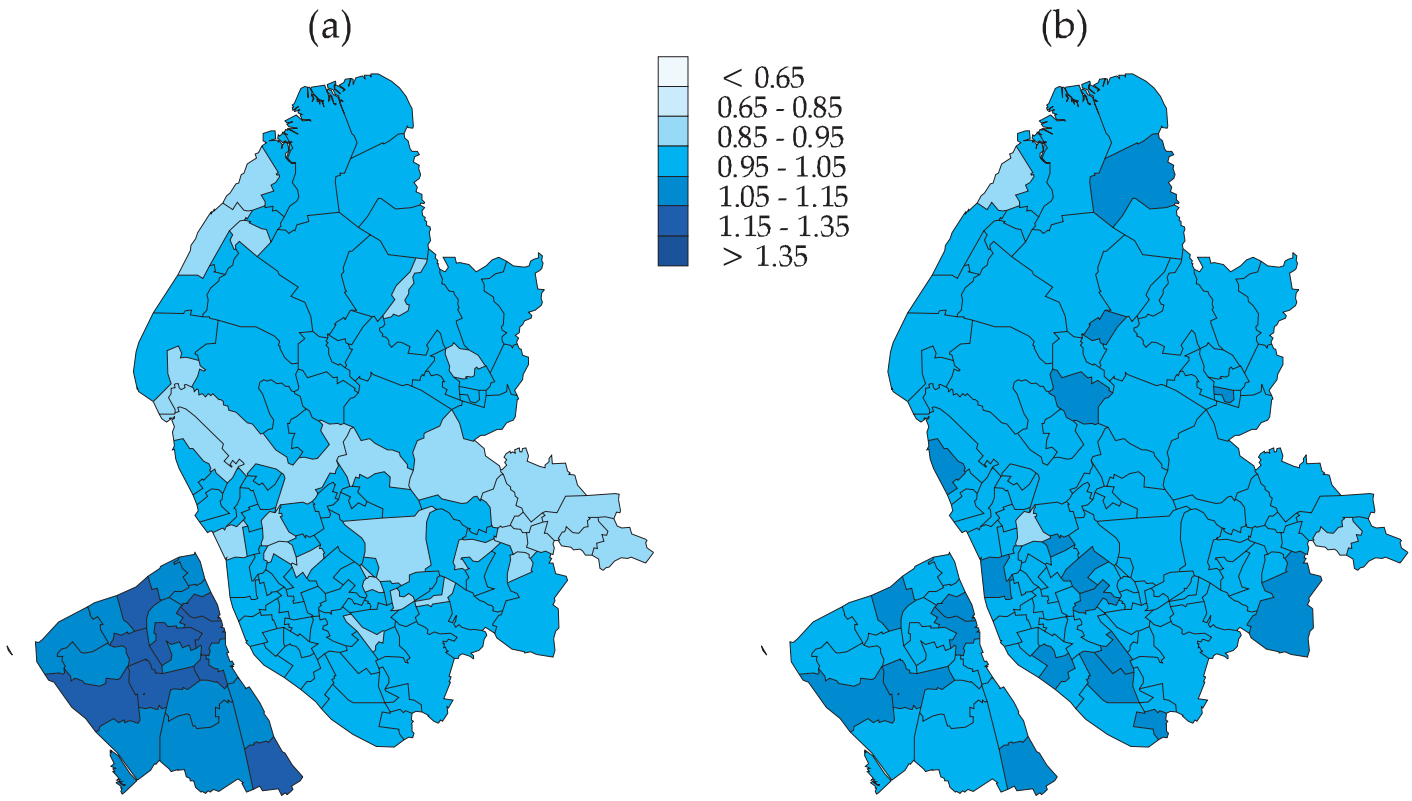

Figure: Smoothed SIRs for brain cancer under (a) a conditional spatial model, and (b) a marginal spatial model.

Example: Colorectal cancer in the West Midlands of England (Kelsall and Wakefield, 2002, JASA)

We include this example to illustrative one useful way of thinking about disease mapping, in terms of a *continuous spatial risk surface*.

Study details:

- Study period is 1989.
- Incidence data by postcode but the analysis is carried out at the ward level, of which there are 39 in the study region. There are a total of 568 cases with a range of 5–27, and a median of 14 per ward.
- Expected counts were based on ward-level populations from the 1991 census, by 5-year age bands and sex, total population is approximately 1 million.

Figure 6 shows the raw SIRs, while Figure 7 gives the smoothed surface – based on a model that assuming a particular smoothing model for the relative risks (a Gaussian process model on the log scale).

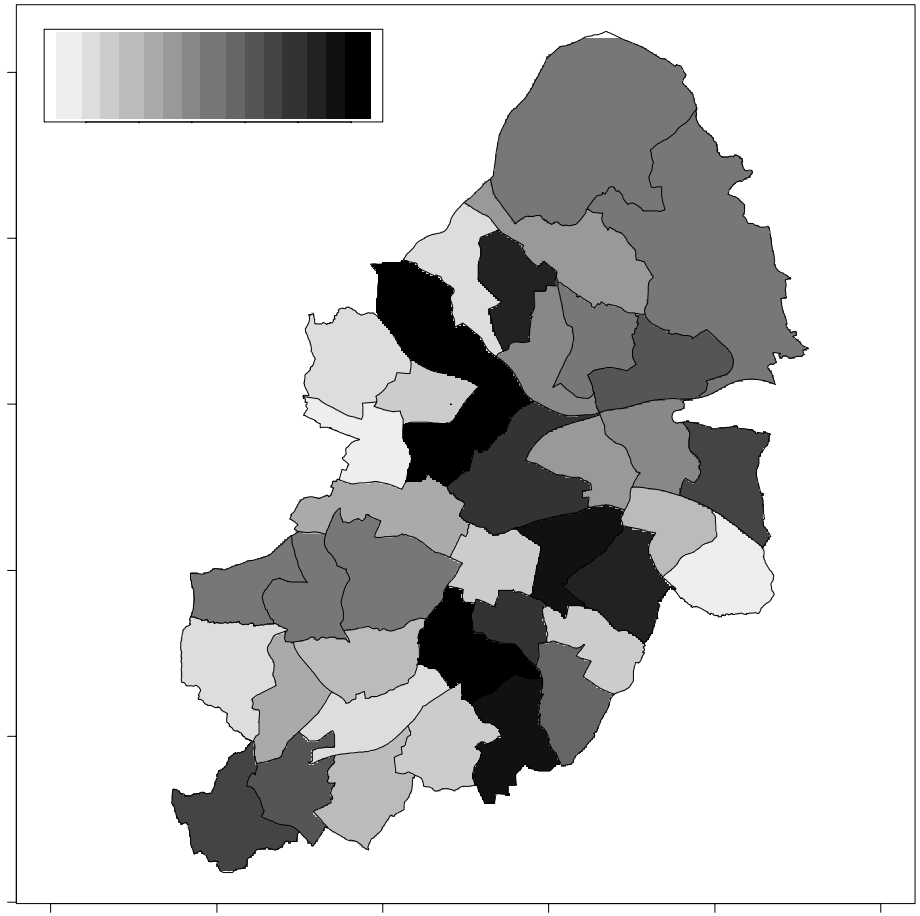

Figure: SIRs for colorectal cancer in the West Midlands.

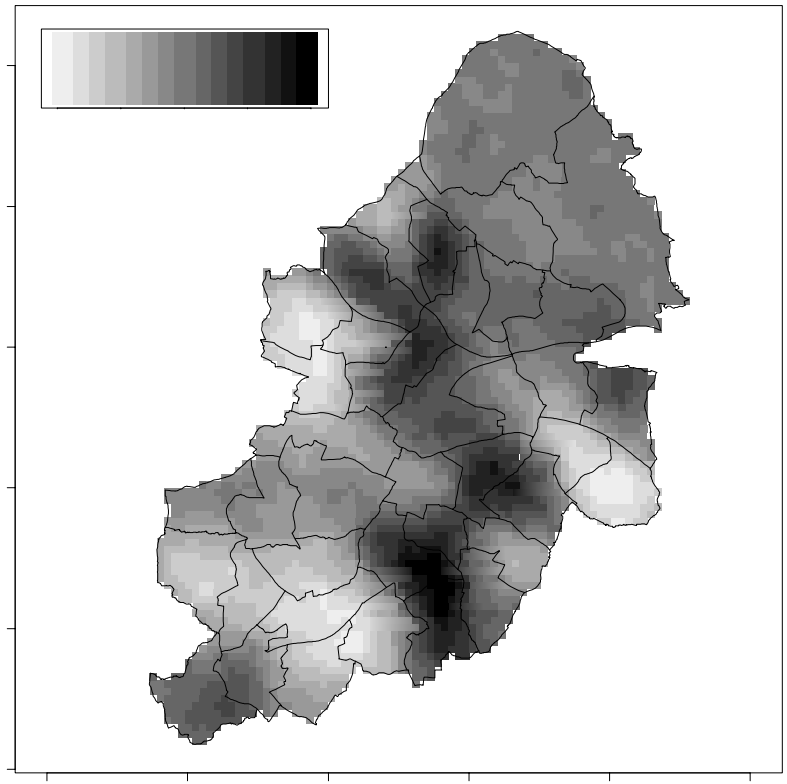

Figure: Smoothed relative risk surface for colorectal cancer in the West Midlands – note the arbitrariness of the area boundaries.

Spatial Regression

Aims:

- Examination of the association between disease outcome and explanatory variables, in a spatial setting, using regression models.
- Conventional modeling approaches such as logistic regression for point data, and loglinear models for count data may be used though if there is significant residual variation methods must acknowledge this in order to obtain appropriate standard errors.
- Also included in this enterprise is the examination of risk with respect to a specific point or line putative source of pollution.
- For count data in particular, the disease mapping models we describe may be extended to incorporate a regression component.

Example: Childhood asthma in Anchorage, Alaska

Study details:

- Data were collected on first grade children in Anchorage, with questionnaires being sent to the parents of children in 13 school districts (the return rate was 70% which has implications for interpretation).
- We analyze data on 905 children, with 885 aged 5–7 years. There were 804 children without asthma, the remainder being cases.
- The exposure of interest is exposure to pollution from traffic. Traffic counts were recorded at roads throughout the study region and a 50m buffer was created at the nearest intersection to the child's residential address and within this buffer traffic counts were aggregated (for confidentiality reasons the exact residential locations were not asked for in the survey).

Figure 8 shows the residential location of the cases and non-cases in Anchorage.

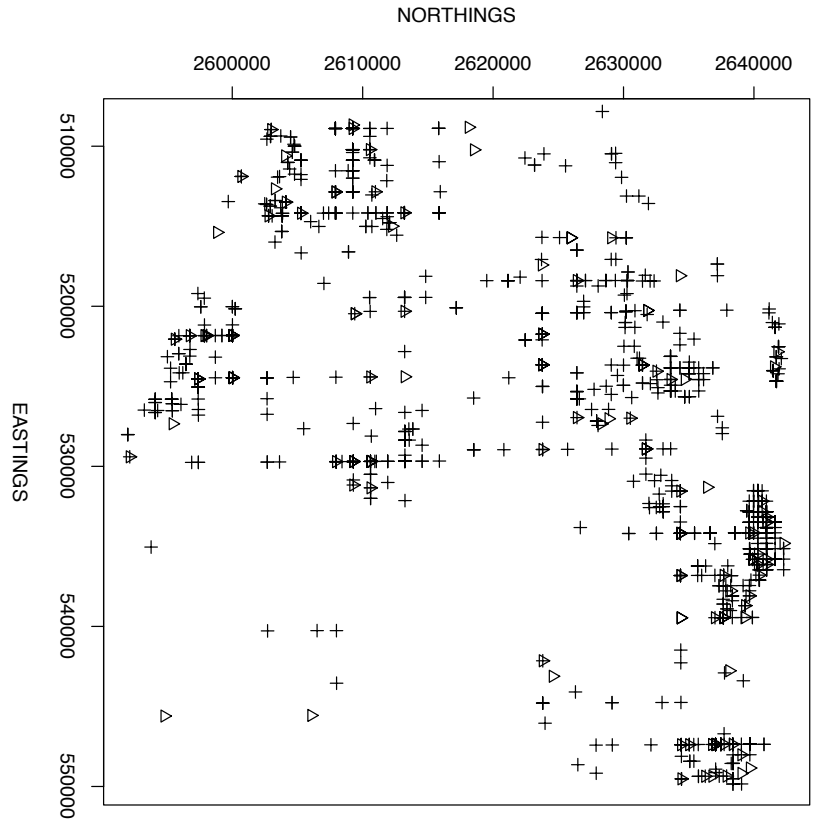

Figure: Asthma cases ($\triangle$) and non-cases ($+$) in Anchorage.

Naive non-spatial logistic modeling

- Initially we may ignore confounding and the spatial nature of the data and fit a logistic regression of asthma incidence on exposure (with the exposure variable scaled to lie between 0 and 10).
- Such an analysis gives an odds ratio of 1.09 with a 90% confidence interval of 1.00–1.18.
- This analysis assumes that, given exposure, the Bernoulli 0/1 labels are independent. Due to unmeasured variables with spatial structure this may not be true which will result in inappropriate standard errors.
- At this stage we make the important point that the sophistication of any analysis should be consistent with the quality of the data. In any observational study, the potential for bias due to confounding and data selection and recording procedures should be carefully considered. In a regression setting, accounting for spatial dependence will often be of secondary importance when compared to these other issues.

Example: Stomach cancer in the North-East of England (Wakefield and Morris, 2001, JASA)

Study objective: examination of risk close to a municipal incinerator.

- Cases are aggregated from post code to Enumeration District – there are 85 counts with 0–10 cases per Enumeration District.
- Populations are available from the decennial census and are aggregated to 44 Enumeration Districts (by age/sex) – the total population in the study region is 36,824.
- Standard caveat: Why are we carrying out the statistical investigation? Is this a random incinerator? Or was it selected because the risk appears elevated, in which case standard measures of measuring the departure from the null (no effect) are not appropriate.
- No exposure measurements are available here, so instead the spatial location of each enumeration district population-weighted centroid relative to the location of the incinerator was used.

Exploratory Analyses

- Figure 9 gives a number of exploratory plots for this example. Panel (a) gives positions of Enumeration District centroids in relation to the incinerator (represented by the origin); the large concentric circle represents the extent of the study region and the smaller circle has radius 3km. The additional circles are centered on the ED centroids and have radii proportional to the expected number of cases.
- In panel (b) we plot Standardized Incidence Ratios versus distance and see a decrease in risk with increasing distance from incinerator (assumed isotropic effect, i.e. no directional effects).
- In panel (c) a census-based index of socio-economic status is plotted versus distance – relatively poorer areas are closer to the pollution source.
- Finally, panel (d) gives SIRs plotted versus socio-economic status. The solid lines on (b)–(d) denote local smoothers. Panels (b)–(d) indicate that confounding of the distance-risk relationship by socio-economic status could be a problem here.

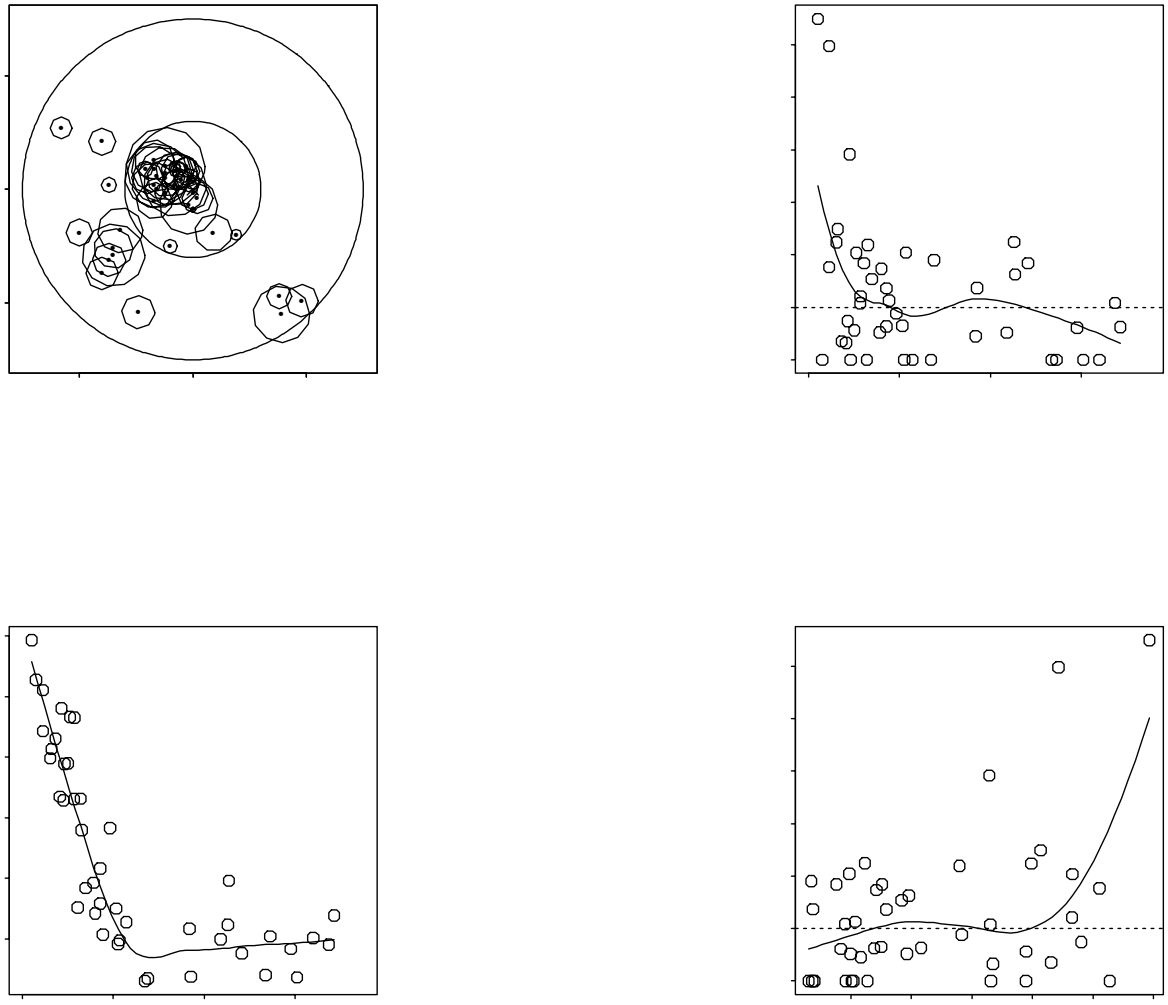

Figure: Exploratory plots for the municipal incinerator example.

Clustering/Cluster Detection

Clustering Aims:

- Examination of *residual spatial clustering* in order to gain clues to disease etiology. For example, is there an infectious agent? Or, a significant unmeasured risk factor?
- The formal comparison of geographical risk with risk factors is the subject of spatial regression. So the examination of clustering step may be seen as exploratory.
- Detection of clustering may also aid in model building in spatial regression settings.

Cluster Detection Aims:

- Cluster detection – examination of spatially-indexed data in order to detect "clusters". May be carried out retrospectively, or prospectively – in the latter case the operation is referred to as *surveillance*.
- Surveillance may offer clues to etiology, but also has a public health role, for example, to determine whether screening programs are being taken up universally (across space).
- Cluster examination in response to an inquiry is subtly different because the hypothesis of increased risk may be the result of data dredging.
- Surveillance is data dredging, but we have set the rules for dredging and so can attempt to adjust significance levels.

Example: Chorley-Ribble point data (Chapter 8, EWBB)

Study details:

- The data for this example consist of the residential locations of 467 cases of larynx cancer and 9191 cases of lung cancer that, for purposes of illustration, will be considered as a set of controls.
- These data were collected in the Chorley-Ribble area of Lancashire in the UK over the period 1974–1983.

Figure 10(a) shows the locations of the cases and controls and Figure 10(b) a perspective view of a kernel density estimate of the controls alone.

The non-uniform distribution of residences is clearly the major source of variation in the spatial distribution of cases and controls.

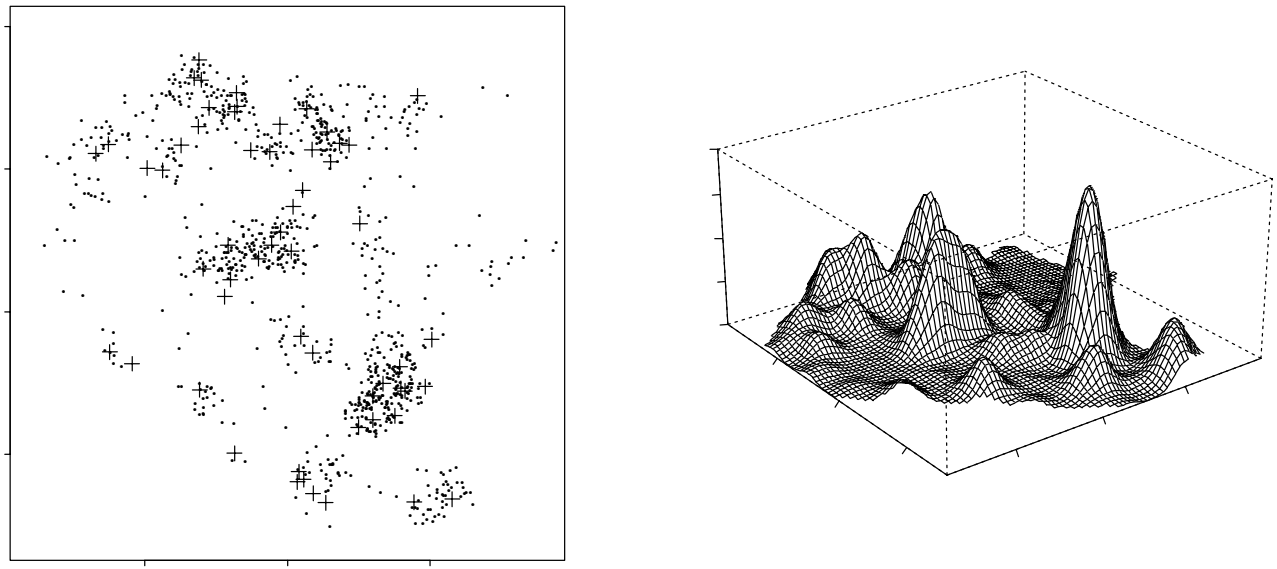

Figure: Case-control data in the Chorley-Ribble area of England: (a) Larynx cancer cases (+) and controls (·), (b) Perspective view of kernel density estimate of control data.

Example: Cluster detection for lung and brain cancer (Chapter 8, EWBB)

- Study details were previously described.
- Here we use these data to determine if there are any "clusters" that might merit further investigation.
- Figure 11 shows one particular method, that of Openshaw et al. (1987), which scans across the map highlighting collections of cases that are "significant" with respect to the underlying population. The circles indicate potential clusters.

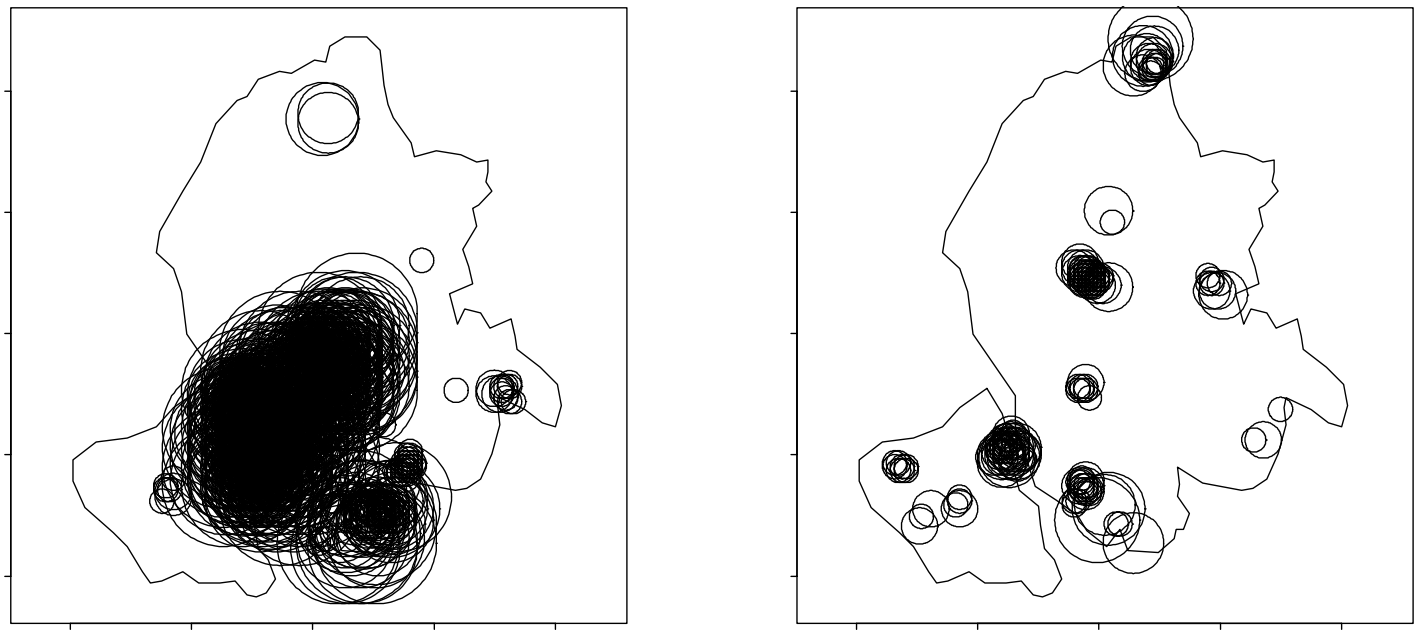

Figure: Results of Openshaw's Geographical Analysis Machine for (a) lung cancer, and (b) brain cancer data. Axes labels are in meters.

- The *incidence proportion* measures the proportion of people who develop the disease during a specified period.
- The *prevalence proportion* is the proportion of people with the disease at a certain time.
- The *risk* is the probability of developing the disease within a specified time interval – estimated by the incidence proportion (note: a probability so between 0 and 1).
- The *relative risk* is the ratio of risks under two exposure distributions.

Precise definitions of the outcomes and exposures under study are required.
The majority of epidemiological studies are observational in nature (interventions provide an example of an experimental study).

Cohort studies select a study population and obtain exposure information, and then the population is followed over time to determine incidence. Requires large numbers of individuals (since diseases are usually statistically rare), and long study duration (for most exposures).

Case-control studies begin by identifying "cases" of the disease and a set of "controls", exposure is then determined. Although subject to selection bias, can overcome the difficulties of cohort studies.

Cross-sectional studies determine the exposure and disease outcome on a sample of individuals at a particular point of time.

Ecological studies use data on groups, areas in a spatial setting. No direct linkage between individual disease and exposures/confounders.

Semi-ecological studies collect individual-level data on disease outcome and confounders, and supplement with ecological exposure information.

Rothman and Greenland (1998) give the following criteria for a confounder:

1. A confounding factor must be a risk factor for the response.
2. A confounding factor must be associated with the exposure under study in the source population.
3. A confounding factor must not be affected by the exposure or the response. In particular it cannot be an intermediate step in the causal path between the exposure and the response.

Note that if a variable is assigned its value before the exposure is assigned, and before the response occurs, then it cannot be caused by either exposure or response.

An Example of When to Adjust

Suppose Y is the rate of lung cancer, X the smoking rate, and Z represents diet and alcohol variables. In this case Z is a confounder under the above definition since it satisfies 1.–3. The causal diagram in Figure 12 illustrates one plausible mechanism for this situation, U denotes unmeasured variables; U could represent education level (or poverty) here. If we obtain data on X, Z, Y then we will see an association between X and Y, but also between Z and Y, hence we must control for Z.

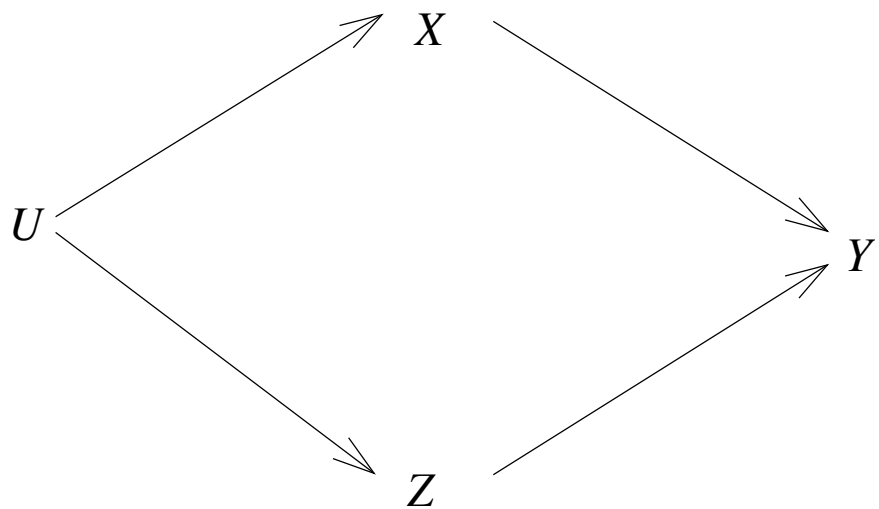

Figure: U denotes unmeasured variables.

Variables on the Causal Pathway

We give an example of a variable that satisfies 1. and 2. but not 3.
In Figure 13, U denotes unmeasured variables, X is smoking, Z is a variable representing tar deposits, and Y is lung cancer.
If we looked at the marginal associations we would find relationships between X and Y but also between Z and Y.
In this case we should not adjust for Z because this would dilute the causal effect of X on Y.
In this example Z is not a confounder because it is on the causal pathway between X and Y (thus invalidating criteria 3.), Z is known as an intermediary variable.

Figure: Z is an intermediary variable, and should not be controlled for.

Variables Affected by the Response

To further illustrate a situation in which a variable satisfies 1. and 2. but contradict 3., we consider an example given by Greenland, Pearl and Robins (1999) in which Y represents endometrial cancer, X estrogen and Z uterine bleeding. The latter could be caused by X or Y and so, under this scenario, we have the causal diagram represented by Figure 14. Again we should not adjust for Z since the estimated causal effect of X on Y would be distorted. Note that we would observe marginal associations between X and Z and Y and Z and so 1. and 2. are satisfied.

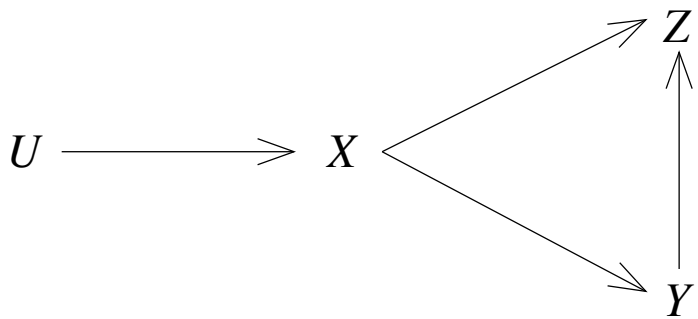

Figure: Z is caused by X and Y, and should not be controlled for.

Spatial Examples

Example 1: Exposure to Sunlight

In the absence of a direct measurement we might use Northings as a surrogate for exposure to sunlight.

However, if we have an available measure (average hours of sunshine by location) then we would not want to include Northings in the model.

Example 2: Confounding by Location

Often, spatial analyses will not contain all of the information on confounders and spatial location will be included in the model to act as a surrogate for the unmeasured variables.

If we are interested in estimating the association between a health outcome and an environmental pollutant then great care must be taken in modeling space: including a complex term for the spatial model will dilute the estimate of the effect of pollution (since this has spatial structure), and including a very simple term may not be sufficiently subtle to control for the unmeasured confounders. See Figures 15 and 16.

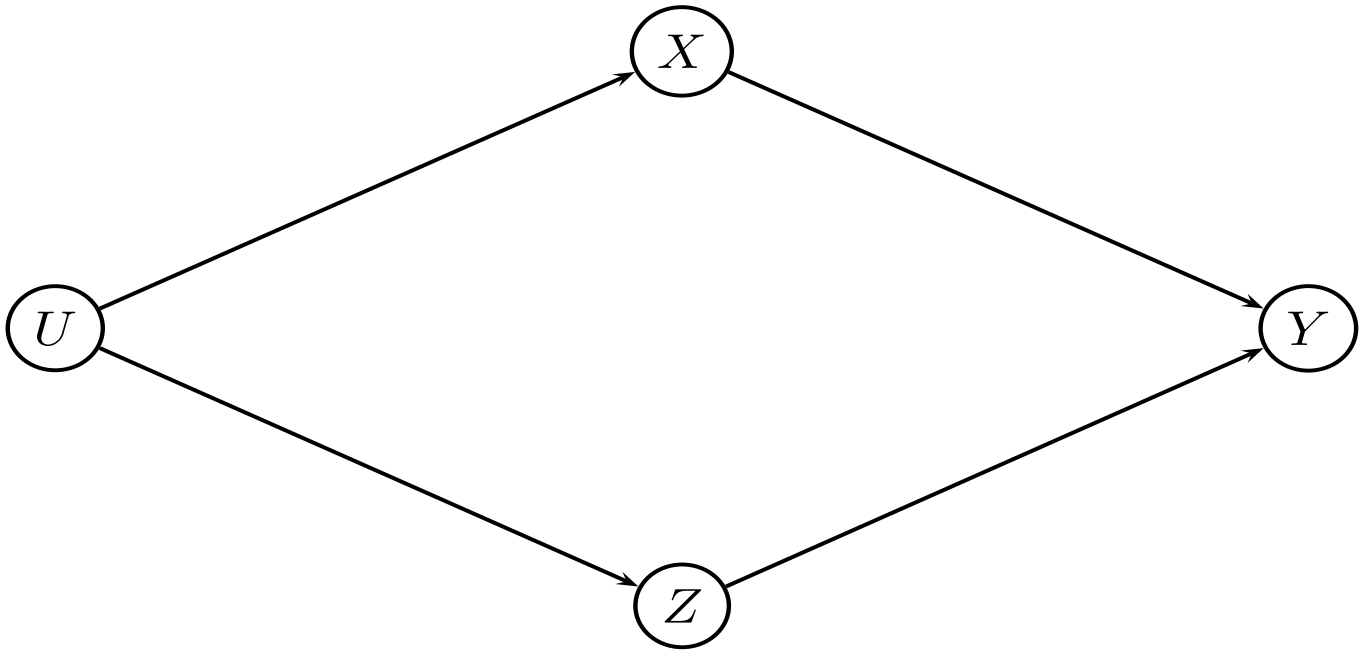

Figure: With confounders Z.

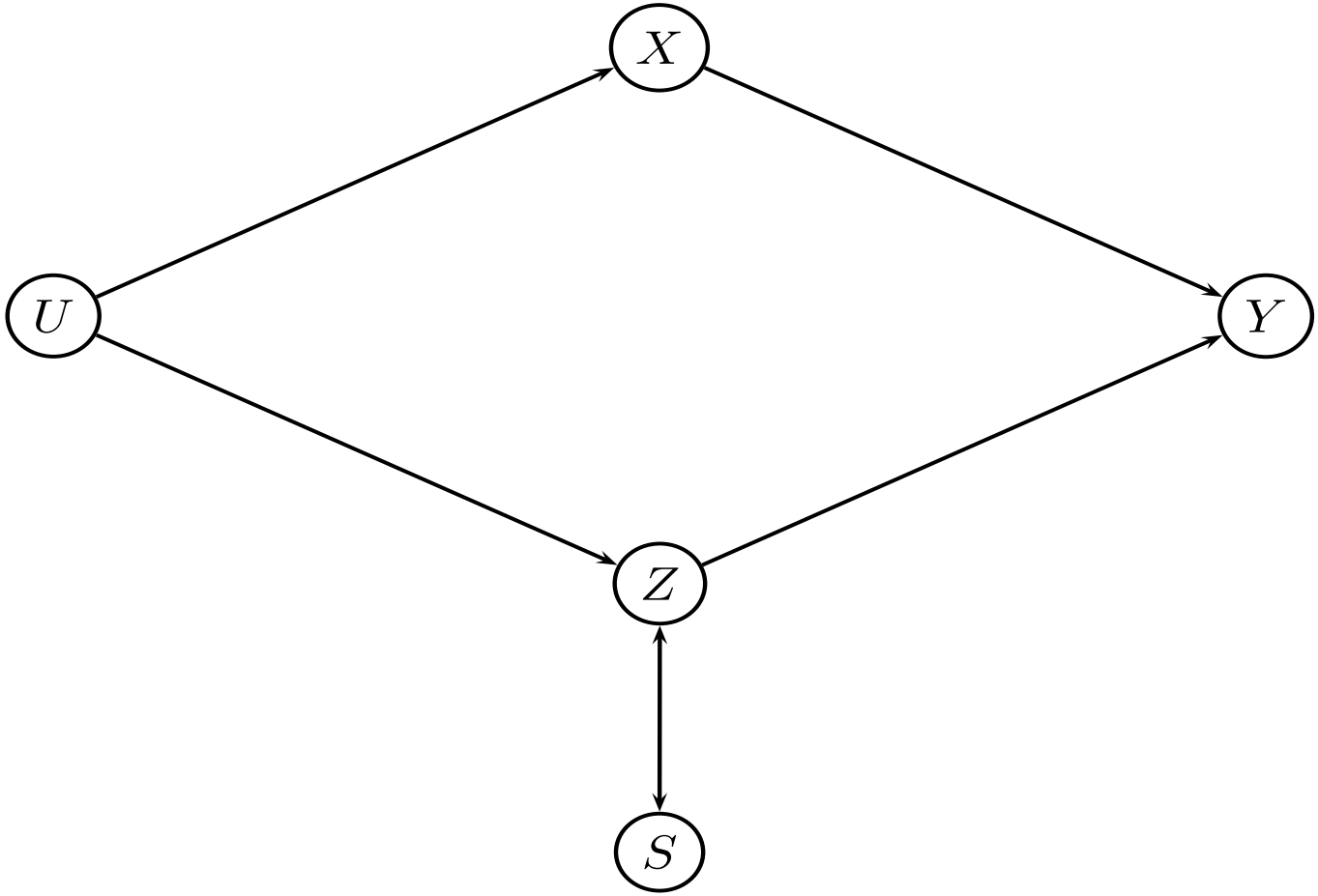

Figure: With confounders Z and space S – with Z (or components there-of) unmeasured an association is induced between S and Y. In this case S may act as surrogates for these components.

Socio-economic confounding

- In spatial epidemiological applications that use count data, population data are obtained from the census and so while one can control for known factors such as age and gender (and sometimes race), information is not available on other possible confounders.
- In such situations it has become common to control for a measure of *socio-economic status.*
- Across various scales of aggregation, measures of deprivation have been shown to be powerful predictors of a variety of health outcomes.
- Deprivation may be viewed as a surrogate for individual-level characteristics such as smoking, diet and alcohol consumption.
- Could be true area-level effects, however, for example, access to health care services.
- Relationship between health, socio-economic status and exposure to environmental pollution is complex since ill-health may cause deprivation (e.g. lose job) so that Y causes Z.

Carstairs Index

A number of area-level indices of deprivation have been created in the UK (e.g., Carstairs, Jarmen, Townsend). In the US income and education may be used as surrogates.

The *Carstairs index* has been extensively used by the Small Area Health Studies Unit (SAHSU), a number of whose studies we shall use as illustration in this course. This index measures (from the census) the proportion of individuals within each ED who: are unemployed; live in overcrowded accommodation; lack a car; have a head of the household who is in low social class.

These variables are standardized across the country and then added together to give a continuous area-based measure with high values indicating increased deprivation.

Important Point: since control is at the ecological, and not the individual, level, the control is not likely to be strong – casting doubt in situations in which *small* relative risks are observed.

Age, for example, will almost always need controlling for – different disease risks in different area may reflect differences in the age population. There are a number of ways to control for confounding, and one method is direct or indirect standardization. Let Y_{ij} denote the number of cases, within some specified period (in years) of S (which is assumed the same for each individual in the study population) in area i and confounder stratum j, and N_{ij} be the corresponding population at risk, $i = 1, ..., m$, $j = 1, ..., J$. Let Z_j and M_j denote the number of cases and population in stratum j in a "reference", or standard, population.

The risk of disease in confounder stratum j in area i, over the time period T, is $\widehat{p}_{ij} = Y_{ij}/N_{ij}$. The rate of disease is $r_{ij} = 1000 \times Y_{ij}/[N_{ij} \times T]$ per 1000 person years. Note that a rate does not need lie between 0 and 1. The crude rate in area i is given by $1000 \times Y_i / \left\{ \sum_{j=1}^{J} N_{ij} \times S \right\}$ per 1000 person years.

The directly standardized rate in area i is given by $\sum_{j=1}^{J} r_{ij} w_j$, per 1000 person years, where $w_j = M_j / \sum_j M_j$ is the proportion of the population in stratum j (these weights may be based on the world, or a uniform, population.

The directly standardized rate is a weighted average of the stratum-specific risks, and corresponds to a "counter-factual" argument in which the estimated rates within the study region are applied to the standard population.

If $q_j = 1000 \times Z_j / [M_j \times T]$ is a standard disease rate in stratum j then the comparative mortality/morbidity figure (CMF) for area i is given by:

$$\text{CMF}_i = \frac{\sum_{j=1}^{J} r_{ij} w_j}{\sum_{j=1}^{J} q_j w_j}$$

In small-area studies in particular the CMF is rarely used since it is very unstable, due to small counts by stratum in area i, Y_{ij}.

The method of indirect standardization produces the standardized mortality/morbidity ratio (SMR):

$$\frac{Y_i}{\sum_{j=1}^{J} N_{ij}\widehat{q}_j}$$

where $Y_i = \sum_j Y_{ij}$ is the total number of cases in area i, and $\widehat{q}_j = Z_j/M_j$ is a reference risk.
The indirectly standardized rate compares the total number of cases in an area to those that would result if the rates in the reference population were applied to the population of area i.
Which reference rates to use? In a regression setting dangerous to use internal standardization in which $\widehat{q}_j = \sum_i Y_{ij} / \sum_i N_{ij}$.
External standardization uses risks/rates from another area.

Expected Numbers

The expected numbers $E_i = \sum_{j=1}^{J} N_{ij} q_j$ follow from assuming the proportionality model

$$p_{ij} = \theta_i q_j$$

where θ_i is the relative risk associated with area i (this assumption removes the need to estimate J risks in each area). Since

$$\mathsf{E}[Y_{ij}] = N_{ij} \theta_i q_j$$

we obtain

$$\mathsf{E}[Y_i] = \sum_{j=1}^{J} N_{ij} \theta_i q_j = \theta_i E_i.$$

The SMR is therefore given by

$$\mathsf{SMR}_i = \frac{Y_i}{E_i}.$$

If incidence is measured then also known as the Standardized Incidence Ratio (SIR). Control for confounding may also be carried out using regression modeling.

In routinely carried out investigations the constituent data are often subject to errors.

Population data

- Population registers are the gold standard but counts from the census are those that are typically routinely-available.
- Census counts should be treated as estimates, however, since inaccuracies, in particular underenumeration, are common.
- For inter-censual years, as well as births and deaths, migration must also be considered.
- The *geography*, that is, the geographical areas of the study variables, may also change across censuses which causes complications.

Health data.

- For any health event there is always the possibility of diagnostic error or misclassification.
- For other events such as cancers, case registrations may be subject to double counting and under registration.

In both instances *local knowledge* is invaluable.

Exposure data

- Exposure misclassification is always a problem in epidemiological studies.
- Often the exposure variable is measured at distinct locations within the study region, and some value is imputed for all of the individuals/areas in the study.
- A measure of uncertainty in the exposure variable for each individual/area is invaluable as an aid to examine the sensitivity to observed relative risks.

Wakefield and Elliott (1999, Statistics in Medicine) contains more discussion of these aspects.
In terms of combining the population, health and exposure data, this is easiest if such data are *nested*, that is, the geographical units are non-overlapping.

- A GIS is a computer-based set of tools for collecting, editing, storing, integrating, displaying and analyzing spatially referenced data.
- A GIS allows linkage and querying of geographically indexed information. So for example, for a set of geographical residential locations a GIS can be used to retrieve characteristics of the neighborhood within the locations lies (e.g. census-based measures such as population characteristics and distributions of income/education), and the proximity to point (e.g. incinerator) and line (e.g. road) sources.
- Buffering – a specific type of spatial query in which an area is defined within a specific distance of a particular point, line or area.

- Time activity modeling of exposures – we may trace the pathway of an individual, or simulate the movements of a population group through a particular space-time concentration field, in order to obtain an integrated exposure.
- In this course, we will not use any GIS tools, but use capabilities within `R` and `WinBUGS/GeoBUGS` to display maps.

Examples

Figure 17 shows a map of Washington state with various features superimposed; this was created with the Maptitude GIS.
Figure 18 smoothed relative risk estimates for bladder cancer.
Figure 19 shows 16 monitor sites in London – a GIS was used to extract mortality and population data within 1km of the monitors, and the association with SO_2 was estimated.

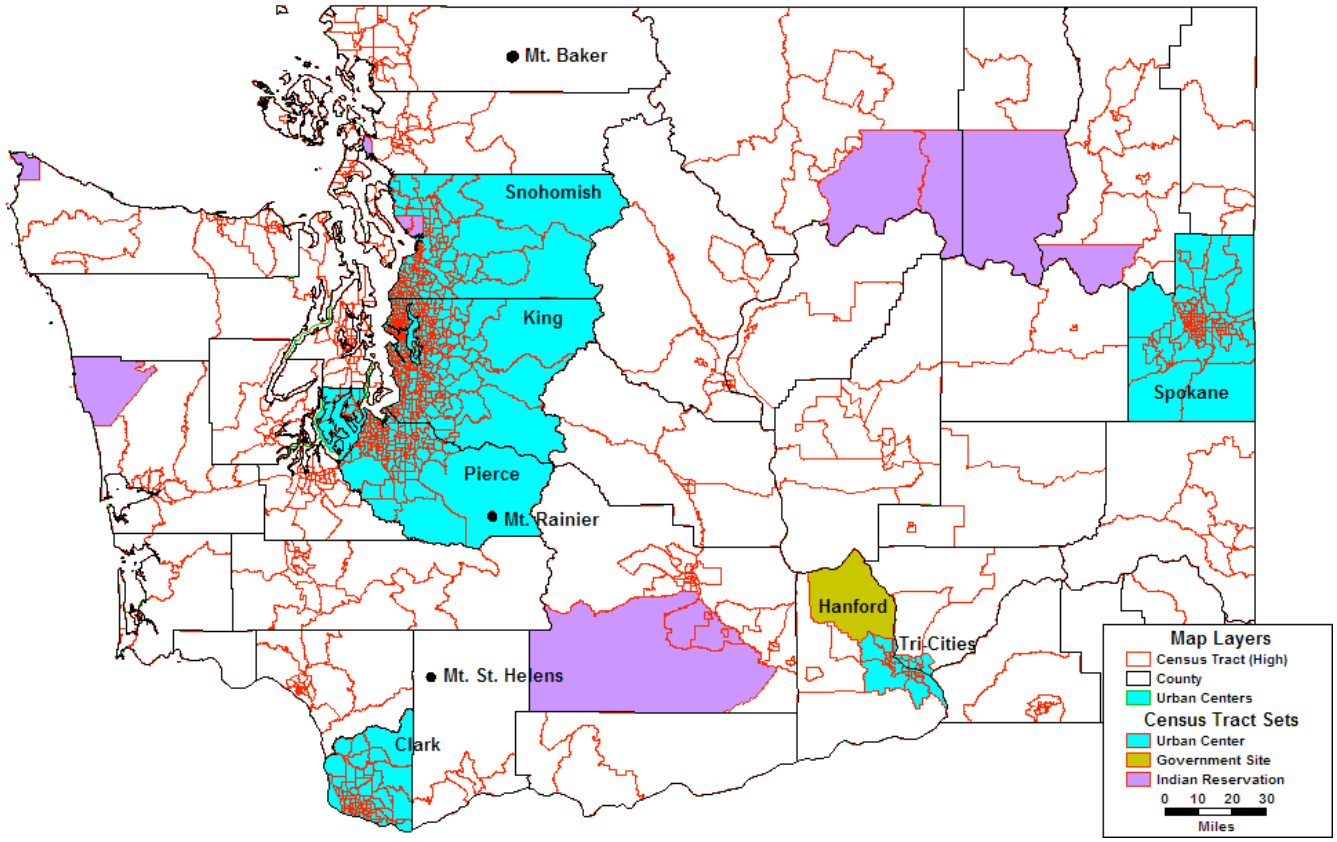

Figure: Features of Washington state, created using a GIS.

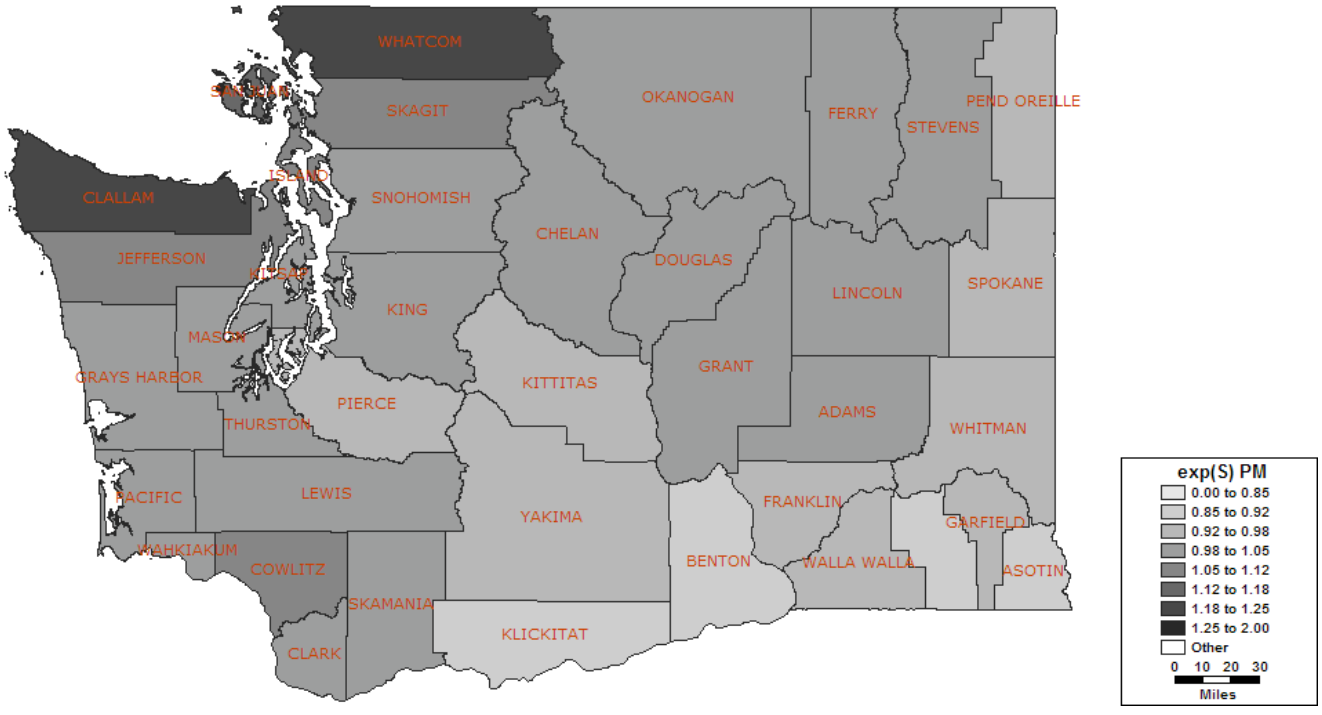

Figure: Smoothed relative risk estimates for bladder cancer in 1990–2000 for counties of Washington state.

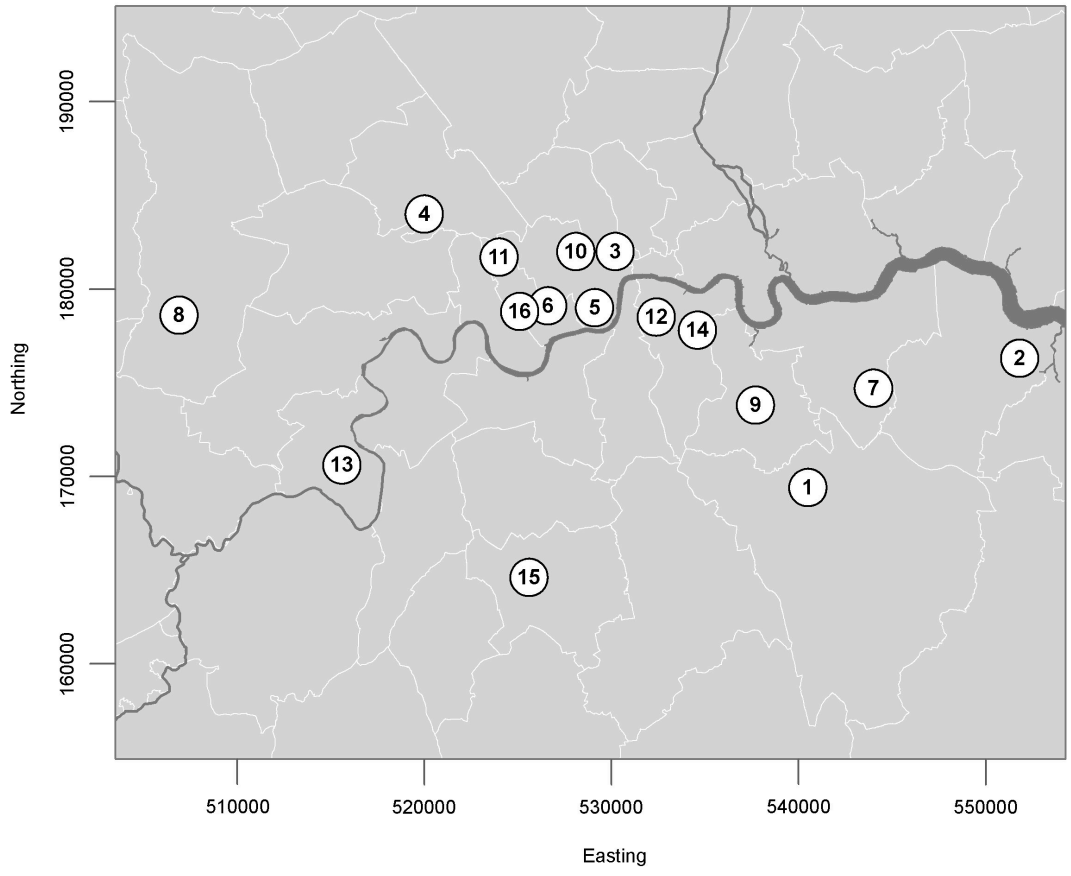

Figure: Air pollution monitor sites in London.

GLMs are a convenient family for fitting a range of data types – we will use the `glm` function in `R`. A GLM is defined by:

- The data arise as an independent sample from an exponential family probability distribution; this family includes the normal, binomial and Poisson distributions.
- A link function linking the mean function, $\mu = \mathrm{E}[Y]$ to a linear predictor $g(\mu) = \mathbf{x}\beta$; logistic regression and log-linear models form two common examples.

By assuming a linear predictor certain aspects of inference are simplified, both in terms of computation and properties of the resultant estimates.

In their original form, GLMs assume independent data, GLMMs extend this to allow dependence induced by random effects.
The link function now has

$$g(\mu_i) = \mathbf{x_i}\beta + \mathbf{b_i},$$

where b_i represents a *random effect*.
The random effects are then assigned a distribution, and in a spatial setting it is natural to assume

$$\mathbf{b} = (\mathbf{b_1}, ..., \mathbf{b_m})^{\mathbf{T}} \sim_{\mathbf{iid}} \mathbf{N_m}(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is an $m \times m$ covariance matrix with (i, j)-th element defining the covariance between random effects at locations i and j.
A simple choice model is $\mathbf{\Sigma_{ij}} = \sigma^{\mathbf{2}}\rho^{\mathbf{d_{ij}}}$, for $i, j = 1, ..., m$, with $\sigma^2 > 0, 0 < \rho < 1$ and d_{ij} the distance between the centroids of areas i and j. This model is *isotropic* since the covariance only depends on the distance between points.

Estimation of parameters may be based on specifying mean and variance of the data only, as in quasi-likelihood, or on specifying the complete probability distribution of the data, as in likelihood and Bayesian approaches.

The likelihood function is the probability distribution viewed as a function of the unknown parameter, and maximum likelihood estimation (MLE) the estimation criteria that chooses that value of the parameter that gives the highest probability to the observed data.

For most models the MLE is asymptotically normal which allows confidence intervals/tests to be constructed.

Example:

Poisson likelihood.
Suppose we have a count Y in an area with expected number E.
Assumed probability model for data, for fixed λ:

$$\Pr(Y = y|\lambda) = \frac{\mathrm{e}^{-E\lambda}(E\lambda)^y}{y!}$$

for $y = 0, 1, ...$ Here λ is the relative risk.
For fixed y we have the likelihood function:

$$l(\lambda) = \frac{\mathrm{e}^{-E\lambda}(E\lambda)^y}{y!} \propto \mathrm{e}^{-E\lambda}\lambda^y$$

for $\lambda > 0$.

Example: Seascale excess

Figure 20 gives an example for $y = 4$, $E = 0.25$ for which the MLE is $\widehat{\lambda} = 16 = \mathrm{e}^{\widehat{\alpha}} = \mathrm{e}^{2.773}$ with 95% asymptotic confidence interval

$$(\mathrm{e}^{2.773-1.96\times0.5}, \mathrm{e}^{2.773+1.96\times0.5}) = (6.0, 42.6).$$

Here is the `R` code for finding the MLE and the standard error:

```
> y <- 4; E <- 0.25
> summary(glm(y~1+offset(log(E)),family=poisson))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.773      0.500   5.545 2.94e-08 ***
```

The "offset" is the known multiplier in the log-linear mean function:

$$\log \mu = \log E + \alpha$$

and $\sim$1 denotes the intercept.
Notice that the parameter is on the linear predictor scale.

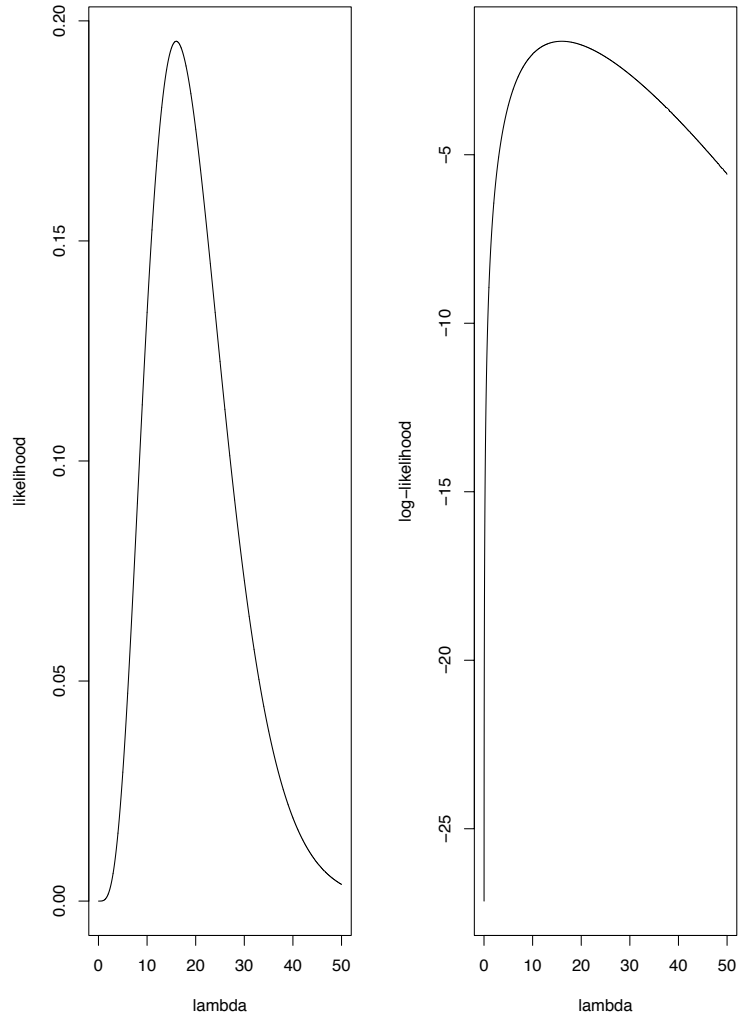

Figure: Likelihood (left) and log-likelihood (right) for Poisson data with $y = 4$, $E = 0.25$.

For the log-linear model

$$Y_i \sim_{ind} \text{Poisson}(\mu_i),$$

with

$$\log \mu_i = \log E_i + \alpha + \beta X_i$$

$i = 1, ..., n$ the MLEs for α and β are not available in closed form but reliable maximization routines are available in all statistical packages.

Example: Simple regression in the Scottish Lip Cancer Data

The file `scotdat.txt` contains the Scottish data as a `list`:

```
z <- list(N = 56, Y = c( 9, 39, 11 ... 1, 0, 0),
          E = c( 1.4, 8.7, 3.0... 7.0, 4.2, 1.8),
          X = c( 0.16, 0.16, 0.10 ... 0.01, 0.16, 0.10))
> source(``scotdat.txt'')
> SMR <- z$Y/z$E
> postscript("scot_smr.ps",horiz=F)
> par(mfrow=c(1,2)) # creates a 1 x 2 plot
> hist(SMR,xlab="SMR")
> plot(z$X,SMR,type="n")
> text(z$X,SMR)
> lines(lowess(z$X,SMR))
> dev.off()
```

This code creates a postscript file for Figure 21.

We carry out likelihood analyses using the `glm` function and the log-linear mean function

$$\log \mathsf{E}[Y_i] = \log E_i + \alpha + \beta X_i, \quad i = 1, ..., 56.$$

$\exp(\beta)$ represents the difference in area-level relative risk between areas with all the population in AFF and zero of the population in AFF.

```
> summary(glm(Y~offset(log(E))+X,data=z,family=poisson(link="log
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.54227    0.06952   -7.80 6.21e-15 ***
X            7.37322    0.59557   12.38  < 2e-16 ***

(Dispersion parameter for poisson family taken to be 1)
```

So $\widehat{\alpha} = -0.542$ (0.070) and $\widehat{\beta} = 7.37$ (0.60) – the relative risk describing the area-based association between incidence and AFF is exp(7.37) = 1588!!!

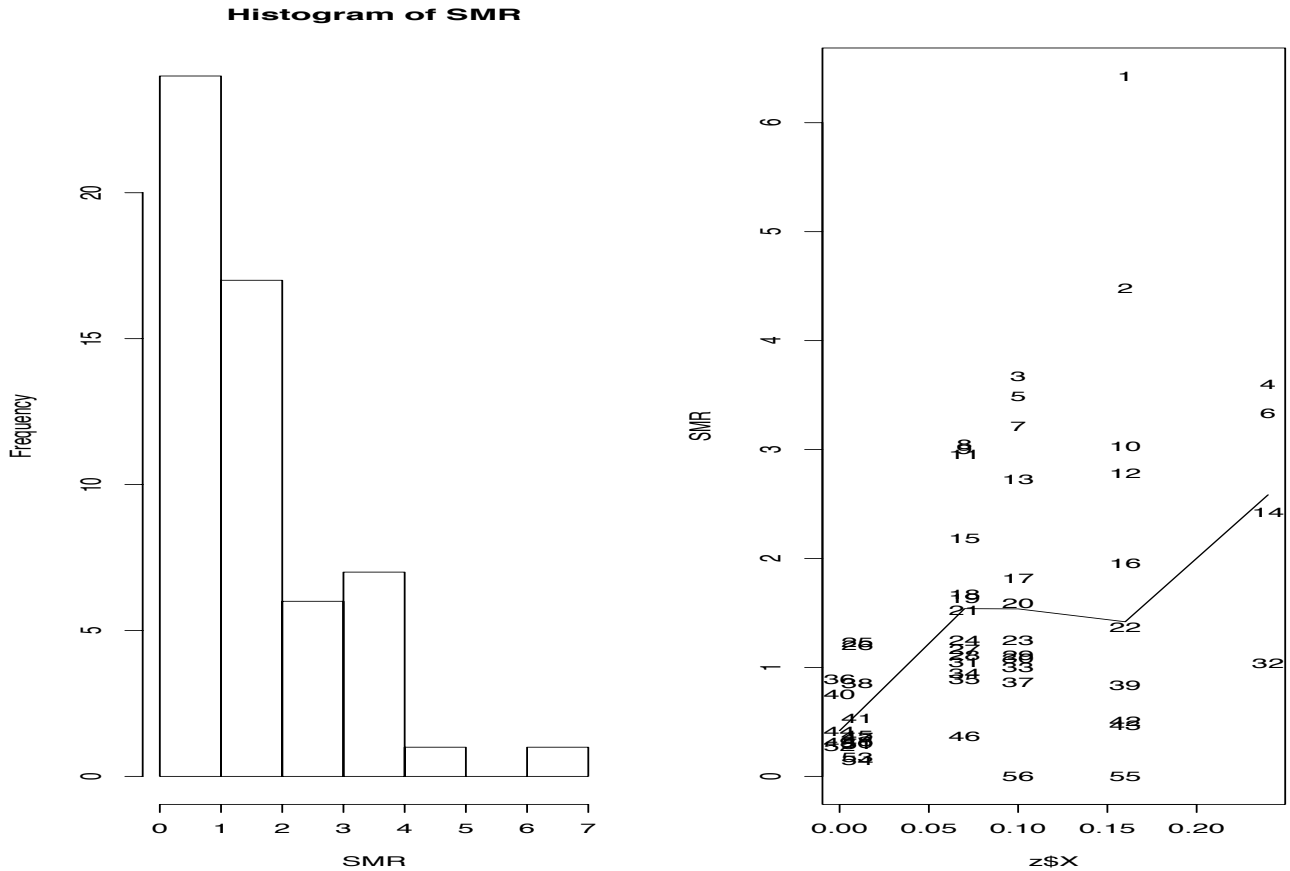

Figure: Histogram of SMRs for the Scottish data, and SMR versus proportion in AFF X.

The Poisson model is restrictive in the sense that the variance is constrained to equal the mean.
In a quasi-likelihood approach we assume

$$\text{var}(Y_i) = \kappa \text{E}[Y_i]$$

where κ allows *overdispersion* and is estimated as

$$\widehat{\kappa} = \frac{1}{n-p} \sum_{i=1} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i^2}$$

where n is the number of counts, and p the number of estimated parameters.
Point estimates are identical to those from likelihood, but standard errors are multiplied by $\widehat{\kappa}^{1/2}$.

To fit a quasi-likelihood model:

```
> summary(glm(Y~offset(log(E))+X,data=z,family=quasipoisson(link
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5423     0.1542  -3.517 0.000893 ***
X             7.3732     1.3208   5.583 7.89e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasipoisson family taken to be 4.9179
```

So $\widehat{\alpha} = -0.542$ (0.15) and $\widehat{\beta} = 7.37$ (1.32) – identical point estimates and standard errors multiplied by $\sqrt{4.92} = 2.22$; large overdispersion here, and the Poisson model is clearly inadequate.

For disease mapping and spatial regression in particular, the Bayesian approach to inference is particularly convenient.
Bayes' Theorem:

$$p(\theta|y) = \frac{p(y|\theta) \times p(\theta)}{p(y)}.$$

Three key elements:

- The *prior* distribution $p(\theta)$.
- The *likelihood* $p(y|\theta)$.
- The *posterior* distribution $p(\theta|y)$.

The crucial difference to likelihood inference is that θ is viewed as a random variable and y as fixed.
The normalizing constant $p(y) = \int l(\theta)p(\theta)\ d\theta$ is often ignored to give:

$$p(\theta|y) \propto p(y|\theta) \times p(\theta).$$

Inference is made through the posterior distribution, and derived quantities, and is based on probability.
To summarize a one-dimensional posterior distribution we might:

- Report the complete posterior distribution.
- Summarize in terms of posterior moments, for example the posterior mean or posterior standard deviation, or quantiles, for example the posterior median or a 90% credible interval.

In general it is not possible to obtain the above summaries analytically (and it's even worse for more than one parameter), but obtaining samples from the posterior is more straightforward.

Example: Poisson data

Suppose

$$Y|\theta \sim \text{Poisson}(E\theta),$$

where Y is the number of disease events, E is the expected number, and θ is the relative risk – viewing this probability distribution as a function of θ gives the *likelihood* function.
For a Bayesian analysis we need a prior for θ.
Consider the *gamma* prior, $\text{Ga}(a, b)$ which has the form:

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta)$$

which has mean a/b and variance a/b^2.
To use this as a prior we need to specify a and b to reflect what we believe about the relative risk *before* we see the data.

Densities and Samples

We emphasize the duality between densities and samples. Figure 22 shows the densities for three choices of a, b and histograms of samples from these densities.

```
> nsim <- 10000
> upper <- 4
> a1 <- b1 <- 5
> thetavals <- seq(0,upper,.1)
> plot(thetavals,dgamma(thetavals,a1,b1),type="n",ylab="Density"
> lines(thetavals,dgamma(thetavals,a1,b1))
> theta1 <- rgamma(nsim,a1,b1)
> hist(theta1,main="",xlim=c(0,upper))
```

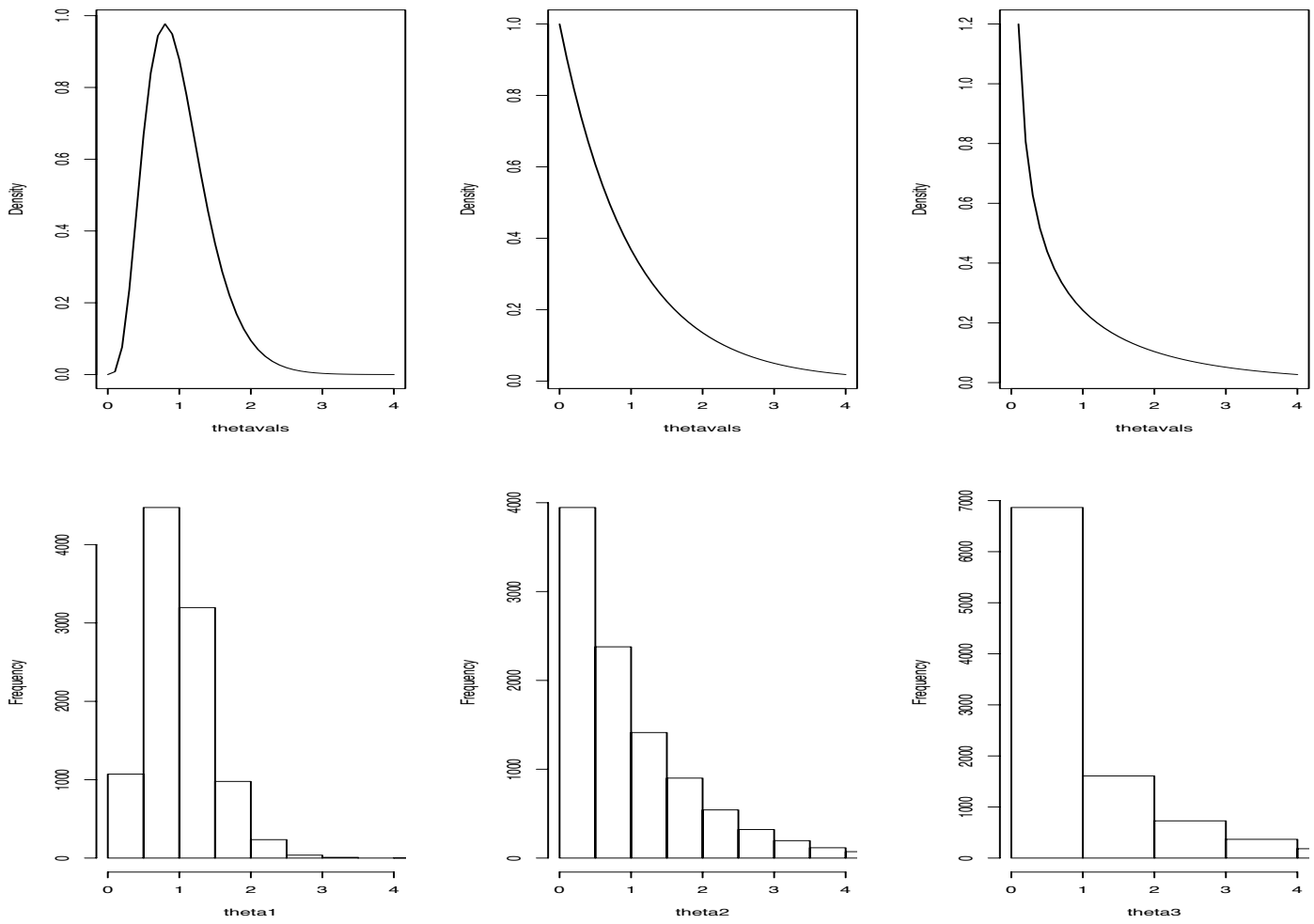

Figure: Densities and samples for the three gamma distributions: Ga$(5, 5)$ (left column), Ga$(1, 1)$ (middle column), Ga$(0.5, 0.5)$ (right column).

Similarly we can calculate theoretical and sample-based prior quantiles and moments.

```
# Theoretical quantiles of a gamma Ga(a1,b1) distribution
# with a1 = b1 = 5.
> qgamma(c(.025,.05,.25,.5,.75,.95,.975),a1,b1) #
[1] 0.3246973 0.3940299 0.6737201 0.9341818 1.2548861 1.8307038
# Sample quantiles of a sample of 10,000 from a gamma Ga(a1,b1)
> quantile(rgamma(nsim,a1,b1),probs=c(.025,.05,.25,.5,.75,.95,.9
     2.5%         5%        25%        50%        75%        95%
0.3177495 0.3875716 0.6712929 0.9290584 1.2504534 1.8519038 2.04
# Theoretical and sample mean and variance
> a1/b1
[1] 1
> a1/b1^2
[1] 0.2
> mean(rgamma(nsim,a1,b1))
[1] 1.007807
> var(rgamma(nsim,a1,b1))
[1] 0.1982659
```

Example: Sellafield

We now illustrate the duality between samples and densities, and the sensitivity of inference to the prior distribution.

- We consider the famous Sellafield nuclear site located in the north-west of England on the coast of West Cumbria. The Sellafield plant re-processes spent fuel from nuclear power plants in Britain.
- In the period 1968–1982 there were four cases of lymphoid malignancy among 0–14 year olds in the village of Seascale which lies 3km to the south of the site, compared with an expected number of 0.25 based on registration rates for the Northern region of England.
- Statistical Model: If all disease risk factors were controlled for in the expected numbers then we might expect the count to follow a Poisson distribution – we start with this model but acknowledge that due to unmeasured risk factors and data anomalies (in particular errors in the population counts), we would expect *overdispersion*.

Gamma Prior Analysis

- With a gamma prior $\mathrm{Ga}(a, b)$ on θ we obtain a posterior of

$$\begin{aligned} p(\theta|y) &\propto l(\theta) \times p(\theta) \\ &\propto \exp(-E\theta)\theta^y \times \theta^{a-1}\exp(-\theta b) \\ &= \theta^{a+y-1}\exp(-\theta[b+E]), \end{aligned}$$

  which is a gamma distribution with parameters $a + y$ and $b + E$.
- This is an example of a *conjugate analysis*, in which the prior and posterior are of the same form.
- The posterior mean is

$$\mathrm{E}[\theta|y] = \frac{a+y}{b+E} = w\frac{a}{b} + (1-w)\frac{y}{E}$$

  where the "weight" $w = b/(b + E)$.
- For a *reference* analysis we may pick $a = b = 0$ (an improper prior), and in this case the posterior mean coincides with the MLE.

- The quantiles are given in Table 1 and indicate that under this prior there is strong evidence that the relative risk associated with the Seascale area is elevated. These values were obtained using the distribution function for a gamma in R.

```
> y <- 4
> E <- 0.25
> a <- b <- 0
> qgamma(c(0.025,0.05,0.5,0.95,0.975),a+y,b+E)
[1]  4.359461  5.465274 14.688243 31.014626 35.069092
> 1-pgamma(1,a+y,b+E)
[1] 0.9998666 # posterior probability of exceedence of
```

- In fact $\Pr(\theta > 1|y = 4) = 0.99987$.

Probability	0.025	0.05	0.5	0.95	0.975
Quantile	4.4	5.5	14.7	31.0	35.1

Table: Posterior quantiles

Gamma Informative prior

The choice Ga(a=5.66,b=5.00) gives $\Pr(\theta < 0.5) = 0.059$ and $\Pr(\theta > 2) = 0.948$. The prior and posterior from the informative Gamma prior and Poisson likelihood analysis are shown in Figure 23 – it is clear that the results will be highly dependent on the prior (since the prior is very influential).

In this example we don't need the sample quantities because the theoretical versions are available – but in general priors and likelihoods do not combine in such a convenient way, but if we can produce samples from the posterior, we can reconstruct summaries of interest.

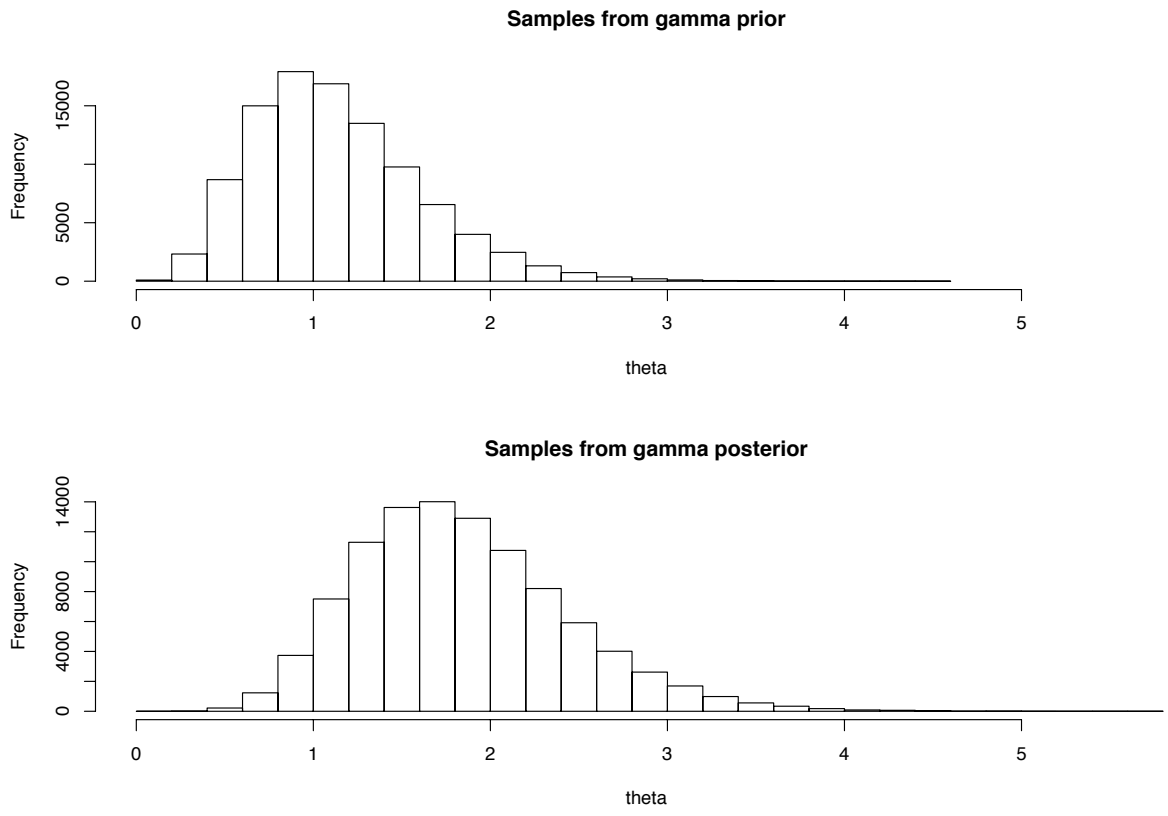

Figure: Gamma prior $\text{Ga}(5.66, 5.00)$ and gamma posterior $\text{Ga}(5.66 + 4, 5.00 + 0.25)$.

Two-Dimensional Example

Suppose we have $Y_j \mid p_j \sim \text{Binomial}(n_j, p_j)$, $j = 1, 2$, with independent beta prior $\text{Be}(a, b)$ distributions. e.g. no. of respiratory health events close ($j = 1$) and not-close ($j = 2$) to a point source that has experienced a discharge of pollutants.
The posteriors are available analytically as

$$\begin{aligned} p(p_j \mid y_j) &\propto p(y_j|p_j)p(p_j) \\ &\propto p_j^{y_j}(1-p_j)^{n_j-y_j} p_j^{a-1}(1-p_j)^{b-1} \\ &= p_j^{a+y_j-1}(1-p_j)^{b+n_j-y_j+1} \end{aligned}$$

a $\text{Be}(a + y_j, b + n_j - y_j)$ distribution.
But suppose we are interested in inference for the odds ratio

$$\phi = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}$$

and for the relative risk $\theta = \frac{p_1}{p_2}$ for which known distributions are not available.

The following is R code to simulate from $\phi \mid y_1, y_2$ when $n_1 = 35, n_2 = 45,\ y_1 = 30, y_2 = 10$:

```
> n1 <- 35; n2 <- 45; y1 <- 30; y2 <- 10
> nsamp <- 1000
> p1 <- rbeta(nsamp,y1+1,n1-y1+1); p2 <- rbeta(nsamp,y2+1,n2-y2+
> oddsrat <- (p1/(1-p1))/(p2/(1-p2));  rr <- p1/p2
> par(mfrow=c(2,2))
> hist(p1,xlim=c(0,1))
> hist(p2,xlim=c(0,1))
> hist(oddsrat)
> hist(rr)
> sum(oddsrat[oddsrat>10])/sum(oddsrat) # Posterior prob that od
                                        # is > than 10
[1] 0.945683
```