2016 SISMID Module 16
Lecture 1: Introduction and Overview

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Course Logistics

Epidemiology Terminology

Standardization

Data Quality

Geographical Information Systems

Coordinate Reference Systems, Projections and Maps

Texts

Primary Book:

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

Supplementary Book:

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

Epidemiology Books:

- Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research. Volume I: The Analysis of Case-Control Studies*, IARC Scientific Publications Nos. 32, Lyon.
- Breslow, N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research. Volume II: The Design and Analysis of Cohort Studies*, IARC Scientific Publications Nos. 82, Lyon.
- Rothman, K. and Greenland, S. (1998). *Modern Epidemiology, Second Edition*, Lipincott-Raven.

Supplementary Texts

R Computing Environment:

- Bivand, R.S., Pebesma, E.J. and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R, Second Edition*, Springer. Available on-line at UW libraries.
- Hills, M., Plummer, M. and Carstensen, B. (2006). *Statistical Practice in Epidemiology with R*. Available on class web site.
- Krause, A. and Olson, M. (2005). *The Basics of S-Plus, Fourth Edition*, Springer-Verlag. Available online from UW libraries.

Supplementary Texts

Additional Spatial Books:

- Baddeley, A., Rubak, E. and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*, CRC Press.
- Banerjee, S., Gelfand, A.E. and Carlin, B.P. (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*, CRC Press.
- Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-Temporal Bayesian models with R-INLA*, John Wiley and Sons.
- Diggle, P.J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.
- Diggle, P.J. and P.J. Ribeiro (2007). *Model-Based Geostatistics*, Springer.
- Gelfand, A.E., Diggle, P.J., Fuentes, M. and Guttorp, P. (2010). *Handbook of Spatial Statistics*, CRC Press.
- Lawson, A.B. (2006). *Statistical Methods in Spatial Epidemiology, 2nd Edition*, John Wiley and Sons.
- Lawson, A.B., Browne, W.J. and Rodeiro, C.L.V. (2003). *Disease Mapping with WinBUGS and MLwiN,* John Wiley and Sons.
- Schabenberger, O. and Gotway, C.A. (2004). *Statistical Methods for Spatial Data Analysis*, CRC Press.
- Shaddick, G. and Zidek, J. (2015). *Spatio-Temporal Methods in Environmental Epidemiology*, CRC Press.
- Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*, Springer.

Logistics

Demonstrations of methods via R implementations will be carried out in class. Students are encouraged to follow along.

Code and other materials (course notes, papers) are available at the course website:

http://faculty.washington.edu/jonno/SISMIDspatial.html

R command files containing R code are on the website.

Course Outline

DAY 1:

- Mon 8.30–10.00: Lecture 1 (Wakefield) Introduction: Overview of course, Motivation, Likelihood and Bayes inference, GIS in R.
- Mon 10.30–12.00: Lecture 2 (Waller) Initial examinations of spatial data, Questions that can be asked. More on GIS.
- Mon 1.30–3.00: Lecture 3 (Waller) Point processes, K functions.
- Mon 3.30–5.00 Lecture 4 (Wakefield) Space and space-time disease mapping, INLA for implementation.

DAY 2:

- Tues 8.30–10.00: Lecture 5 (Waller) Spatial regression including geostatistics.
- Tues 10.30–12.00: Lecture 6 (Wakefield) Clustering and cluster detection with aggregate data, Scan statistics.
- Tues 1.30–3.00: Lecture 7 (Waller) Slippery Slopes: Spatially Varying Coefficients
- Tues 3.30–5.00: Lecture 8 (Waller) Disease Ecology.
- Tues 5.00–6.00: R session: Exercises.

DAY 3

- Wed 8.30–10.00: Lecture 9 (Wakefield) Disease dynamics/infectious diseases, illustrated with measles and flu examples.
- Wed 10.30–12.00: Lecture 10 (Wakefield) Small-area estimation.

Motivation: 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.

Time, which may be measured on different scales (age/period/cohort), is also a surrogate for aging processes and exposures/experiences accrued.

In a perfect world we would have data on residence history, so that we could examine space-time interactions in detail.

Types of Data

Key Point: People are not uniformly distributed in space, therefore we need information on the background spatial distribution of the population at risk in order to infer whether the spatial distribution of cases differs.

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.

Overview

- Introduction
  - Motivation of the need for spatial epidemiology
  - Types of spatial study and examples
  - Background:
    - Epidemiological concepts
    - Overview of approaches to statistical inference
    - Overview of `R`.
- 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.
  - Non-spatial and spatial smoothing models
  - Bayesian inference and computation (`INLA` software)
  - Models for aggregate data
  - Space-time models:
    - Random walk smoothing models in time
    - Age-Period-Cohort models and the Lexis diagram
    - Prediction
  - Prevalence mapping
  - Exposure mapping
  - Examples.

Overview

- Spatial Regression
    - Simple approaches via logistic and Poisson regression
    - Ecological bias and the ecological fallacy.
    - Sophisticated approaches
    - Geostatistical regression for point data for prevalence mapping
    - Methods for pollution point sources
    - 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 a (statistically) rare outcome.
    - For point data logistic regression is the obvious approach though we may also use "geostatistical" methods which model the spatial risk surface.
    - 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").
    - Examples.

Overview

- Clustering and 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”.
  - The latter refers to on-line surveillance or retrospective analysis, to reveal “hot spots”.
  - Understanding the form of the spatial dependence is often an aim.
  - Distance/adjacency methods
  - Moving window methods
  - Risk surface estimation
  - Examples.

Overview

- Infectious Disease Modeling
  - Gain clues of space-time dynamics of a disease.
  - Fit a model which allows the effect of interventions (e.g. vaccination) to be assessed.
  - Epidemic/endemic models in time and space
  - Chain binomial models
  - Examples.
- Small-Area Estimation
  - Estimate the total number of events (or the proportion) of interest in a geographical area, based on a sample which may be “small”.
  - May be used for planning, for example, interventions.
  - Introduction to survey data
  - Weighted estimators
  - Spatial smoothing
  - Examples.

Need for Spatial Methods

All epidemiological studies are spatial!

But often the study area is small and/or there is abundant individual-level information and so spatial location is not acting as a surrogate for risk factors.

When do we consider the spatial component?

- When we are explicitly interested in the spatial pattern of disease incidence? e.g. disease mapping, cluster detection.
- When we want to leverage spatial dependence in rates to improve estimation, e.g. small area estimation.
- The clustering may be a nuisance quantity that we wish to acknowledge, but are not explicitly interested in? For example, in spatial regression we want to get appropriate standard errors.

Need for Spatial Methods

If we are interested in the spatial pattern then, if the data are not a complete enumeration, we clearly we would prefer the data to be "randomly sampled in space", i.e., not subject to selection bias with the extent of bias depending on the spatial location of the individual.

For example, in a matched case-control study, we may match controls on the geographical region of the cases, which will clear distort the geographical distribution of controls (so that they will not be representative of the population at risk).

In small-area estimation and prevalence mapping this is a serious consideration because the available data are often gathered via complex survey designs in which, for example, cluster sampling, stratified sampling and over-sampling of certain populations is carried out for reasons of logistics, or because of power considerations.

Need for Spatial Methods

In complex survey design settings, design weights are typically reported, these reflect the design, and can often be interpretated as the number of individuals represented by that surveyed individual.

In this case, careful thought is required to determine whether the weights should be incorporated in the analysis.

Example: Suppose we are interested in the prevalence of diabetes across areas with samples being taken and diabetes status determined. Suppose we oversamples African Americans, if one ignores the design we will obtain biased estimates of prevalence because diabetes is associated with race.

Motivation

Growing interest in spatial epidemiology due to:

- Public interest in effects of environmental "pollution", *e.g.* Sellafield, UK (Gardner, 1992), Three-Mile Island, US.
- Acknowledgment that many environmental/man-made risk factors may be detrimental to human health.
- 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, population and exposure data at different geographical scales.
- Evidence-based decision making, regarding interventions, for example, requires point and interval estimates for relevant quantities. Prevalence mapping and small area estimation are both endeavors that provide estimates with associated measures of uncertainty.
- Increase in computing power and tools such as Geographical Informations Systems (GIS).

Measures of disease occurrence

- The *incidence (proportion)* is the proportion of people in a population who develop the disease during a specified period.
- The *prevalence (proportion)* is the proportion of people in a population with the disease at a certain time.
- The *(incidence) risk* is the probability of developing the disease within a specified time interval – can be estimated by the incidence proportion.
- In general, summaries can be reported as rates or risks, the former are positive while the latter are between 0 and 1.
- The *relative risk* is the ratio of risks under two exposure distributions (e.g., exposed and not exposed).

Precise definitions of the outcomes and exposures under study are required.

The majority of epidemiological studies are observational in nature.

In contrast, an intervention provides an example of an experimental study.

Observational study types

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

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

Matched case-control studies are case-control studies in which cases are matched with controls on the basis of confounders.

- Efficiency considerations lead to the conclusion that 3–5 controls to each case is usually sufficient.
- Frequency matched studies ensure that the ratio of controls to cases is roughly constant within broad confounder bands (e.g. 10-year age bands). Individually matched studies carry out precise matching.

Observational Study Types

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

The nested case-control study starts with a cohort and identifies cases that have already occurred, or as they occur. For each case, a specified number of controls is selected from within the cohort among those in the cohort who have not developed the disease by the time of disease occurrence in the case[1]. Some form of time-matching is carried out.

[1] Theoretically every case-control study takes place within a cohort, but identifying the cohort is often difficult

Observational Study Types

A case-cohort study is a variant on the nested case-control study without matching. Hence, the same sub-cohort (a random sample of the complete cohort) can be used for multiple disease outcomes.

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.

Thomas (2014) provides a good summary on study designs.

Confounding

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.

Risks and rates

Suppose we observe a cohort of people over P years and we are interested in the disease incidence over this period.

For an individual in the cohort let T be the survival time.

Let $P = h \times m$ where h is some interval of time, e.g. $P = 5$ years and $h = 0.5$ (6 months) so that $m = 10$.

Now split the interval $[0, P)$ into sub-intervals $[t_i, t_{i+1})$ with

$$t_i = (i - 1) \times P/m,$$

$i = 1, \ldots, m$.

Example: $P = 5$ years and $h = 0.5$ with $m = 10$ gives $t_1 = 0$, $t_2 = 0.5$, $t_3 = 1, \ldots, t_m = 4.5$ and intervals:

$$[0, 0.5), \quad [0.5, 1), \quad [1, 1.5), \ldots, [4.5, 5).$$

The $[,)$ notation here means that a time exactly at the break point is included at the start of an interval, so that 0.5 is in the second interval.

Risks and rates

The probability of failure in an interval,

$$[t_i, t_i + h) = \left[(i-1) \times \frac{P}{m}, i \times \frac{P}{m}\right),$$

given survival to the start, is

$$\begin{aligned} \pi(t_i) &= \Pr(t_i \leq T < t_i + h | T \geq t_i) \\ &= \Pr(\text{ failure in } [t_i, t_i + h) \mid \text{survival until } t_i \text{ }) \\ &\approx \lambda(t_i) \times h \end{aligned}$$

where $\lambda(t_i)$ is the hazard rate, i.e. the instantaneous probability of failure, $i = 1, \ldots, m$.

Note

$$\lambda(t) \approx \frac{\pi(t)}{h}$$

is a rate.

Risks and rates

The probability of failure before P is[2]

$$\begin{aligned} {}_P p_0 &= \Pr(\text{ failure in } [0, P) \text{ }) = 1 - \Pr(\text{ survival over } [0, P)) \\ &= 1 - [1 - \pi(t_1)] \times [1 - \pi(t_2)] \times \cdots \times [1 - \pi(t_m)] \\ &= 1 - \prod_{i=1}^{m} [1 - \pi(t_i)] \end{aligned}$$

For example, suppose $P = 3$ years, $m = 36$ and conditional probabilities of failure of $\pi(t_i) = \pi = 0.0005$ (i.e. constant), $i = 1, \ldots, m$, the probability of failure in any 1-month interval, given survival until this point.

Then the probability of failure in 3 years is

$${}_3 p_0 = 1 - [1 - \pi]^{36} = 0.0178.$$

[2] using demography notation ${}_n p_x = \Pr(\text{event before } x + n \mid \text{no event by } x)$

Risks and rates

Suppose we have a constant π, which:

- for continuous survival times correspond to exponential survival times, and
- for discrete survival times correspond to geometric survival times.

This gives a constant hazard, and the (cumulative) survival probability is

$$(1-\pi)^m = (1-\lambda h)^m$$

and taking logs

$$m \log[1-\lambda h] \approx -m\lambda h = -P\lambda$$

for small h.

Hence,

$$\log(\text{ Cumulative survival probability }) = -\text{ cumulative failure rate}$$

or

$$\begin{aligned}\text{Cumulative survival probability} &= \Pr(\text{ survival over } [0, P)) \\ &= \exp(-\lambda P).\end{aligned}$$

Risks and rates

Under another approximation

$$\begin{aligned} \text{Probability of disease in } [0, P) &= 1 - \exp(-\lambda P) \\ &\approx 1 - [1 - \lambda P] \\ &= \lambda P \end{aligned}$$

so that the failure probability (risk) is approximated by the cumulative rate.

The above may be extended to a time varying hazard rate $\lambda(t)$ for $0 \leq t < P$.

Survival analysis link

The survivor function is $S(t) = \Pr(T \geq t)$ and the probability density function is

$$f(t) = -\frac{dS(t)}{dt}.$$

The hazard function is

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

The cumulative hazard function is

$$\Lambda(t) = \int_o^t \lambda(s)ds.$$

Since

$$\frac{d}{dt}\Lambda(t) = \lambda(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t),$$

we can express the survivor function as

$$S(t) = \exp\left\{-\Lambda(t)\right\}.$$

Risks and rates

The cumulative survival probability is

$$\Pr(\text{ survival over } [0, P) \text{ }) = \exp\left\{ -\int_0^P \lambda(s) ds \right\}.$$

In some models $\lambda(t)$ is allowed to vary, for example, with different rates in different years; these are sometimes known as period effects.

The other obvious extension is to allow different hazard rates at different ages.

Finally, there may be different rates for different cohorts, i.e. individuals born at different times.

Combining the three different time scales give age-period-cohort models (Clayton and Schifflers, 1987a,b; Carstensen, 2007).

Risks and rates

We now examine how the above can be extended to the situation in which we have N individuals in an area.

For each of the N individuals we have, under exponential survival times,

$$\Pr(T \geq P) = \exp(-\lambda P).$$

Let Y be the number of disease cases (failures) in $[0, P)$, then, assuming independence of failures[3]

$$Y|\lambda \sim \text{Binomial}(N, 1 - \exp[-\lambda P])$$

or (approximately)

$$Y|\lambda \sim \text{Binomial}(N, \lambda P).$$

[3] which would not be reasonable in an infectious disease context

Risks and rates

Under a rare disease assumption

$$Y|\lambda \sim \text{Poisson}(N\lambda P), \tag{1}$$

where the mean of the distribution is the (approximate) person years NP, times the failure rate λ.

Take Away Message: Suppose we have counts and populations Y_i and N_i in areas $i = 1, \ldots, n$, over some study of length P; a starting model is

$$Y_i|r_i \sim \text{Poisson}(N_i r_i).$$

By comparison with (1) we see that r_i is approximating the risk ($= \lambda_i P$, allowing different risks in different areas) in area i, and the rate ($= \lambda_i$) is r_i/P.

Risks and rates

We can estimate the probability of failure (the risk) in $[0, P)$ and area i by

$$\widehat{r}_i = \frac{Y_i}{N_i}.$$

The rate is estimated by

$$\widehat{\lambda}_i = \frac{\widehat{r}_i}{P} = \frac{Y_i}{N_i P},$$

i.e., the number of events divided by the person years.

A rate does not need to lie between 0 and 1, but cannot be negative.

The rate is sometimes expressed as a function of some number of years, for example, the rate per 1000-person years is

$$1000 \times \widehat{\lambda}_i = 1000 \times \frac{\widehat{r}_i}{P} = 1000 \times \frac{Y_i}{N_i P}.$$

Spatial context of risks and rates

We extend the concepts to a spatial context in which we have confounding (say by age).

We will almost always have to account for age in the analysis, since different disease risks in different area may reflect differences in the age population.

There are a number of ways to control for confounding, and two common methods are direct or indirect standardization.

Let

- Y_{ij} denote the number of cases, within some specified period 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 denote the number of cases in a "reference", or "standard", population.
- Let M_j be the population in stratum j in this "reference", or "standard", population.

Spatial Context of Risks and Rates

The risk of disease in confounder stratum j in area i, over the time period $[0, P)$ (in years, say), is

$$\widehat{r}_{ij} = \frac{Y_{ij}}{N_{ij}}.$$

The rate of disease per 1000 person years is

$$1000 \times \widehat{\lambda}_{ij} = \frac{1000 \times Y_{ij}}{P \times N_{ij}}.$$

The crude rate in area i per 1000 person years is

$$\frac{1000 \times Y_i}{P \times N_i}$$

where $Y_i = \sum_j Y_{ij}$ is the total number of cases and $N_i = \sum_{j=1}^{J} N_{ij}$ is the total population in area i.

Standardization

The directly standardized rate, per 1000 person years, in area *i* is

$$1000 \times \sum_{j=1}^{J} \widehat{\lambda}_{ij} w_j,$$

where

$$w_j = \frac{M_j}{\sum_j M_j}$$

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 rates, and corresponds to a counter-factual argument in which the estimated rates within the study region are applied to the standard population.

Standardization

If

$$\widehat{\psi}_j = \frac{Z_j}{P \times M_j}$$

is a standard disease rate in stratum j then the comparative mortality/morbidity figure (CMF) for area i is:

$$\text{CMF}_i = \frac{\sum_{j=1}^{J} \widehat{\lambda}_{ij} w_j}{\sum_{j=1}^{J} \widehat{\psi}_j w_j}.$$

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

Hence, for data aggregated over areas we will concentrate on the Standardized Mortality/Morbidity Ratio (SMR).

Indirect Standardization

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

$$\text{SMR}_i = \frac{Y_i}{\sum_{j=1}^{J} N_{ij} q_j}$$

where q_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 risks in the reference population were applied to the population of area i.

Which reference rates should be used?

In a regression setting dangerous to use internal standardization in which $\widehat{q}_j = Y_j / N_j$ since we might distort the effect of the exposure of interest.

External standardization uses risks/rates from another region or another time period.

Expected Numbers

We have $\mathrm{E}[Y_{ij}] = N_{ij} r_{ij}$ where r_{ij} is the risk in area i and stratum j.

In a general model we would try to estimate all parameters r_{ij} which, for small areas in particular, is not possible.

As an alternative we can assume the proportionality model

$$r_{ij} = \theta_i \times q_j$$

so that θ_i is the relative risk associated with area i,since

$$\theta_i = \frac{r_{ij}}{q_j},$$

for all j, i.e., a ratio of risks.

Expected Numbers

We then have

$$\mathrm{E}[Y_{ij}] = N_{ij}\theta_i \times q_j.$$

Summing over stratum:

$$\mathrm{E}[Y_i] = \mathrm{E}\left[\sum_{j=1}^{J} Y_{ij}\right] = \theta_i \sum_{j=1}^{J} N_{ij} q_j = \theta_i E_i$$

where the expected numbers are defined as

$$E_i = \sum_{j=1}^{J} N_{ij} q_j.$$

This assumption removes the need to estimate J risks in each area and we have the model

$$Y_i | \theta_i \sim \mathrm{Poisson}(E_i \theta_i)$$

and the aim is often to model θ_i as a function of space and spatially-referenced covariates.

Expected Numbers

In some cases the outcome of interest will not be available by stratum for each area (i.e., we don't have Y_{ij} only Y_i).

But we may have access to rates by stratum (nationally or regionally, for example).

Hence, if we have population counts by stratum (which are often available) then we can still fit the model

$$Y_i|\theta_i \sim \text{Poisson}(E_i\theta_i).$$

This allows better control for confounding than including some summary of the strata in a regression model (e.g., proportion in each group, or proportion female).

Checking the Proportionality Assumption

One can assess the proportionality assumption by examining SMRs calculated over collections of strata.

Consider groups defined by collections of strata (for example, females and males); let k index these groups (e.g. $k = 1, 2$ for females, males).

Then we can assume proportionality for each group k:

$$r_{ij}^{(k)} = \theta_i^{(k)} \times q_j$$

where we again assume q_j is known.

We can then form expected numbers for each group k:

$$\mathrm{E}[Y_i^{(k)}] = \sum_{j=1}^{J} N_{ij}^{(k)} \theta_i^{(k)} \times q_j = \theta_i^{(k)} E_i^{(k)}$$

with expected numbers $E_i^{(k)} = \sum_{j=1}^{J} N_{ij}^{(k)} \times q_j$.

Checking the Proportionality Assumption

We can then form stratum-specific relative risk estimates

$$\widehat{\theta}_i^{(k)} = Y_i^{(k)}/E_i^{(k)}$$

and if these look roughly similar (across all areas) for each k, then we can make the assumption that we have proportionality.

A simple way of informally comparing the estimates is to plot the relative risk against each other.

For example, if the groups are gender we plot $\widehat{\theta}_i^{(1)}$ against $\widehat{\theta}_i^{(2)}$.

More formally, one may carry out likelihood ratio tests by including different interaction terms in (Poisson) loglinear models (e.g. interactions between area and gender and age and gender).

SMRs

The SMR in area i is

$$\text{SMR}_i = \frac{Y_i}{E_i}$$

and is an estimate of θ_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.

If there is a concern with confounding then we can not collapse the data, and fit models to Y_{ij} (and so estimate the q_j along with other parameters in the model).

Summary on risks, rates and standardization

It is sometimes useful to think about the underlying survival model which produces the rates or risks that we model within a binomial or Poisson model.

In these latter two models, a count s being modeled, but these aggregates are summaries of individual-level continuous survival times.

Often, standardization for age is carried out within the expected numbers, and time of the event of interest (i.e., the period) is summarized via a count; hence, we are using an age-period model.

Sometimes it is beneficial to consider cohort effects also.

A proportionality assumption underlies indirect standardization: if it doesn't hold, then be careful on interpretation.

Data Quality Issues

In routinely carried out investigations the constituent data are often subject to errors; *local knowledge* is invaluable for understanding/correcting these errors.

Wakefield and Elliott (1999) contains more discussion of these aspects.

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.

Data Quality Issues

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.

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.

Combining the population, health and exposure data is easiest if such data are *nested*, that is, the geographical units are non-overlapping.

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.

It is important to attempt to control for confounding when one is wishing to estimate the association between disease risk and an exposure.

In such situations it has become common to control for a measure of *socio-economic status (SES).*

Across various scales of aggregation, measures of SES have been shown to be powerful predictors of a variety of health outcomes.

SES may be viewed as a surrogate for individual-level characteristics such as smoking, diet and alcohol consumption.

True area-level effects could be present, however, for example, access to health care services.

Area-Based Indices

Relationship between health, SES and exposure to environmental pollution is complex since ill-health may cause loss of job (for example) so that Y causes Z.

A number of area-level indices of SES have been created in the UK (e.g., Carstairs, Jarmen, Townsend). See the discussion in Carstairs (2000).

In the US income and education are often used.

If one is interested in disease mapping (which is more a problem in prediction), then an area-based measure of SES may be useful for providing better, i.e., small mean squared error(MSE)[4], estimates.

In this case one is not interested in interpretation of the response-SES relationship, or worrying about whether confounding is "being accounted for".

[4] The MSE is defined as the sum of the squared bias and the variance of an estimator

Area-Based Indices

The *Carstairs index* has been extensively used by the Small Area Health Studies Unit (SAHSU), where JW worked for 3 years.

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 (i.e., area) level, and not the individual level, the control is not likely to be strong, which casts doubt on the validity of the findings in situations in which *small* relative risks are observed.

Geographical Information Systems (GIS)

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 of pollution.

Buffering – a specific type of spatial query in which an area is defined within a specific distance of a particular point, line or area.

Geographical Information Systems (GIS)

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 exclusive GIS products, but use capabilities within `R` to combine datasets and display maps.

Examples

- Figure 1 shows a map of Washington state with various features superimposed; this was created with the Maptitude GIS.
- Figure 2 smoothed relative risk estimates for bladder cancer.
- Figure 3 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.

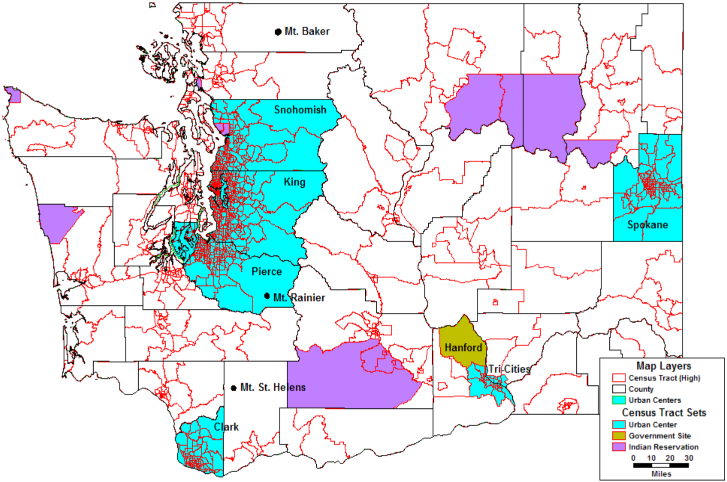

Figure 1 : Features of Washington state, created using a GIS.

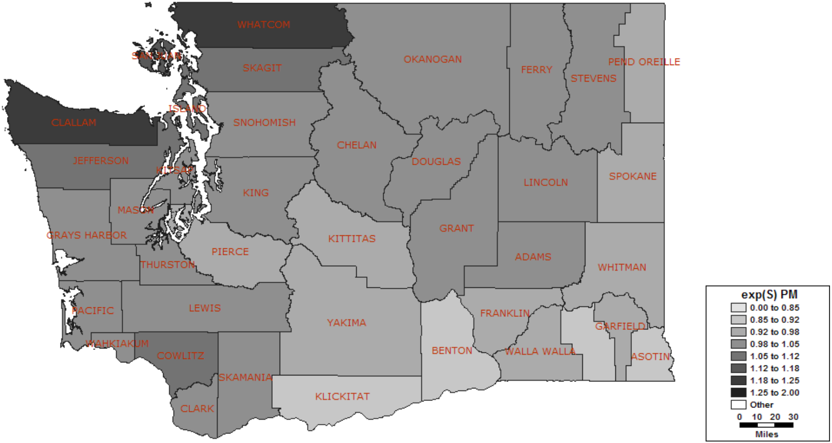

Figure 2 : Smoothed relative risk estimates for bladder cancer in 1990–2000 for counties of Washington state.

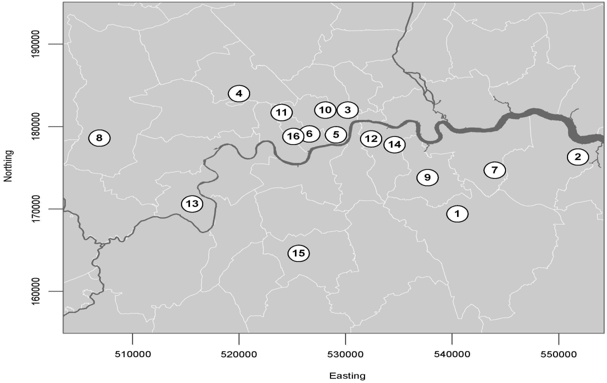

Figure 3 : Air pollution monitor sites in London.

Coordinate Reference Systems

Coordinate reference systems (CRS) are key to GIS.

The CRS allows, amongst other things, datasets to be combined and a reference for calculated distances.

The objective is to represent attributes within some region on the face of the earth on the plane.

A geographic CRS can be expressed in degrees and gives a representation of a model for the shape of the earth — a prime meridian and a datum, which anchors a CRS to an origin in 3D, including a height from the center of the earth or above a standard measure of sea level.

Most countries have multiple CRS.

Projections

If we have a set of points on the surface of the Earth they may be represented by their associated latitude and longitude (one possible coordinate system).

Lines of longitude pass through the north and south poles. The origin is the line set to 0° and is the line of longitude passing through the Greenwich Observatory in England.

Longitude can be measured in degrees (0° to 180°) east or west from the 0° meridian.

Projections

Due to the curvature of the Earth the distance between two meridians (line of longitude) depends on where we are.

Latitude references North-South position and lines of latitude (called parallels) are perpendicular to lines of longitude.

The equator is defined as 0° latitude.

Once a coordinate system is decided upon once must decide on which projection is to be used for display, i.e., the map projection.

Projections

Different map projections distort areas, shapes, distances and directions in different ways.

When move from 3-dimensions to 2-dimensions, it is intuitively obvious that we will lose some information.

Over the years, many weird and wonderful projections have been invented.

Conformal (e.g. Mercator) projections preserve local shape.

Equal-area (e.g. Albers) projections preserve local area.

Once projection has taken place a grid system must be established.

One common system is the Universal Transverse Mercator (UTM) coordinate system, which is a conformal mapping system.

Conclusions

Key questions to ask:

- Why are we interested in space?
- What is the aim of the study: description, exploration, specific risk factors of interest, estimation of prevalence/incidence/totals?
- What is the interpretation of the parameters of the model?
- How does the sampling scheme impact modeling/interpretation?
- What are the important confounders and how can we control for these variables?

References

Carstairs, V. (2000). Socio-economic factors at areal level and their relationship with health. In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial Epidemiology: Methods and Application*, pages 51–67. Oxford University Press, Oxford.

Carstensen, B. (2007). Age–period–cohort models for the lexis diagram. *Statistics in medicine*, **26**, 3018–3045.

Clayton, D. and Schifflers, E. (1987a). Models for temporal variation in cancer rates. I: age–period and age–cohort models. *Statistics in medicine*, **6**, 449–467.

Clayton, D. and Schifflers, E. (1987b). Models for temporal variation in cancer rates. II: age–period–cohort models. *Statistics in medicine*, **6**, 469–481.

Gardner, M. (1992). Childhood leukaemia around the sellafield nuclear plant. In P. Elliott, J. Cuzick, D. English, and R. Stern, editors, *Geographical and Environmental Epidemiology: Methods for Small-area Studies*, pages 291–309, Oxford. Oxford University Press.

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

Thomas, D. C. (2014). *Statistical Methods in Environmental Epidemiology*. Oxford University Press, Oxford.

Wakefield, J. and Elliott, P. (1999). Issues in the statistical analysis of small area health data. *Statistics in Medicine*, **18**, 2377–2399.