2021 SISMID Module 5
Lecture 1: Introduction and Overview

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

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.

An obvious but important point

Key Point: Units are not uniformly distributed in space, therefore we need information on the background spatial distribution of the units in order to infer whether the spatial distribution of units of interest (e.g., cases) differs.

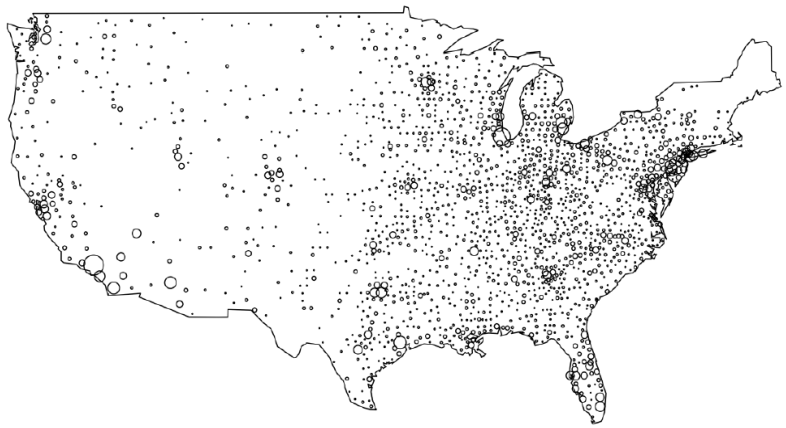


Figure 1: A SRS of 10,000 voter locations from the USA.

Three Types of Analyses

- *Geostatistical data* in which "exact" residential locations exist for the points, and spatial regression and/or prediction is of interest.
- *Area 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.
- *Point data* in which we are interested in the actual configuration of a set of points.

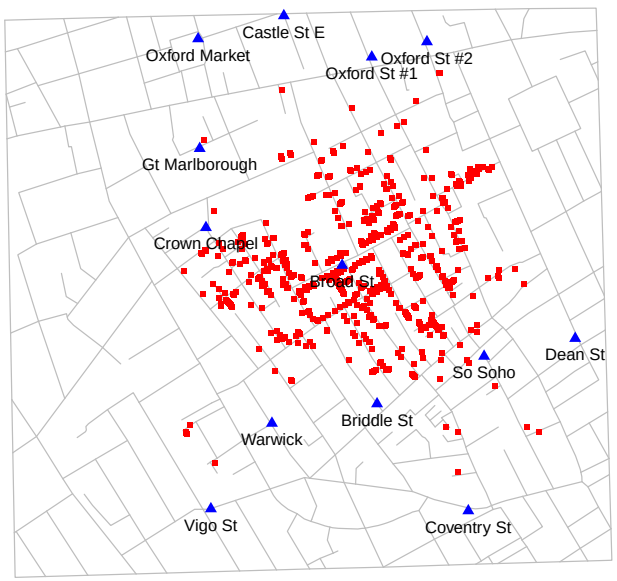


Figure 2: John Snow: Deaths locations in red, pumps in blue.

Need for spatial methods

All investigations 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 explicitly consider the spatial component?

- When we are explicitly interested in the spatial pattern of disease incidence? e.g., disease mapping, small area estimation/prevalence mapping.
- For these endeavors we want to leverage spatial dependence in rates to improve estimation.
- Cluster detection is usually concerned with localized increases in risk.
- Clustering of events may be of direct interest, or may be a nuisance quantity that we wish to acknowledge, but are not explicitly interested in.
- In spatial regression we want to get appropriate standard errors – acknowledging spatial dependence will often change the slope estimate, which is known as confoudning by location.

The selection mechanism

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 clearly distort the geographical distribution of controls (so that they will not be representative of the population at risk).

Motivation

Growing interest in spatial epidemiology due to:

- Public interest in effects of environmental "pollution". Growing interest since: Sellafield, in the UK (Gardner, 1992) and Three-Mile Island in the 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.

Motivation

Growing interest in spatial methods due to:

- 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).
- Cataloging the geographical inequity of disease risk.

Motivating examples: Area disease counts

Data: Male Scottish lip cancer incidence in 1975–1980, with an ecological covariate, proportion who work in agriculture, fishing and farming.

Objectives/Issues: Disease mapping, Spatial regression, ecological bias.

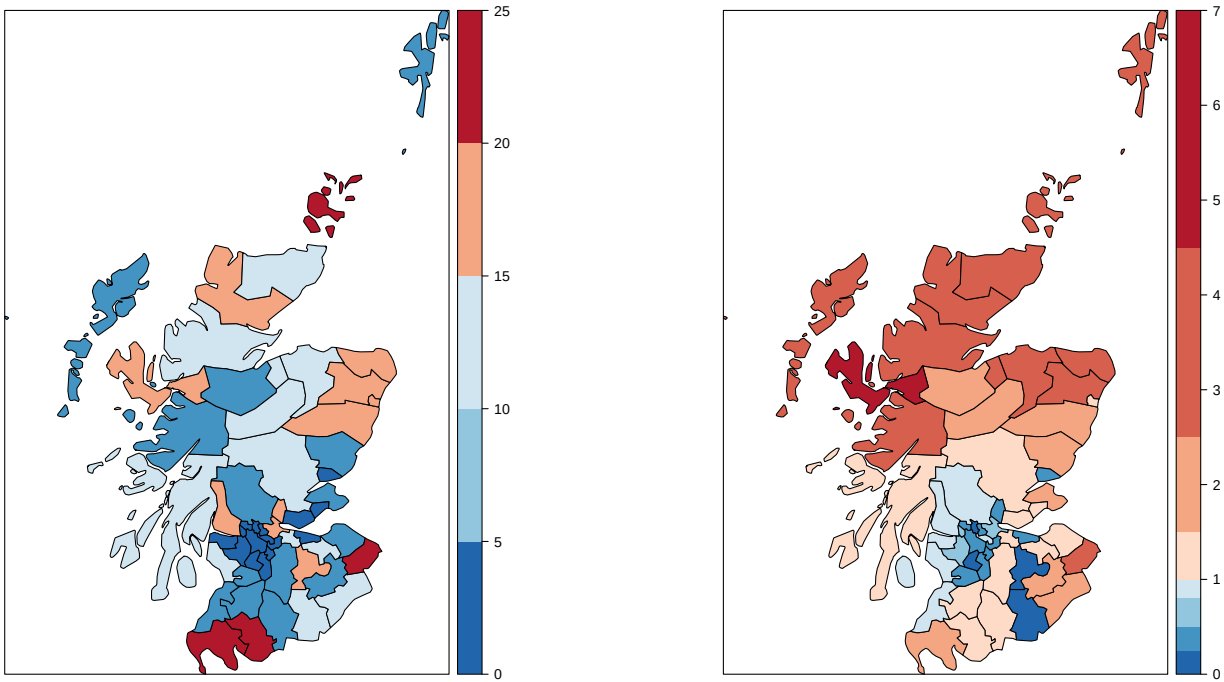


Figure 3: Left: proportion in agriculture, fishing and farming. Right: Standardized incidence ratios (SIRs).

Motivating examples: Point-Level Data

Data: HIV prevalence among women 15–29 on Malawi based on the 2015–16 Demographic Health Survey.

Objectives/Issues: Prevalence mapping.

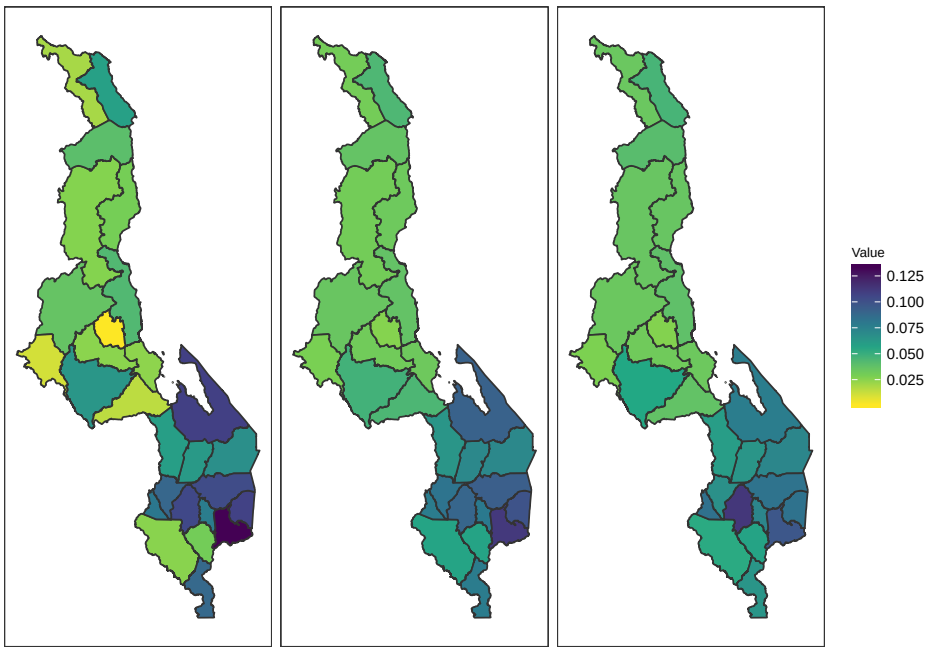


Figure 4: Top: HIV prevalence estimates from weighted, smoothed weighted and point level approaches.

Motivating Examples: Assessment of Clustering, Point Data

Data: Residential locations of larynx cases and lung cancer cases (which are treated as.a control).

Objective: Clustering of larynx cancer cases.

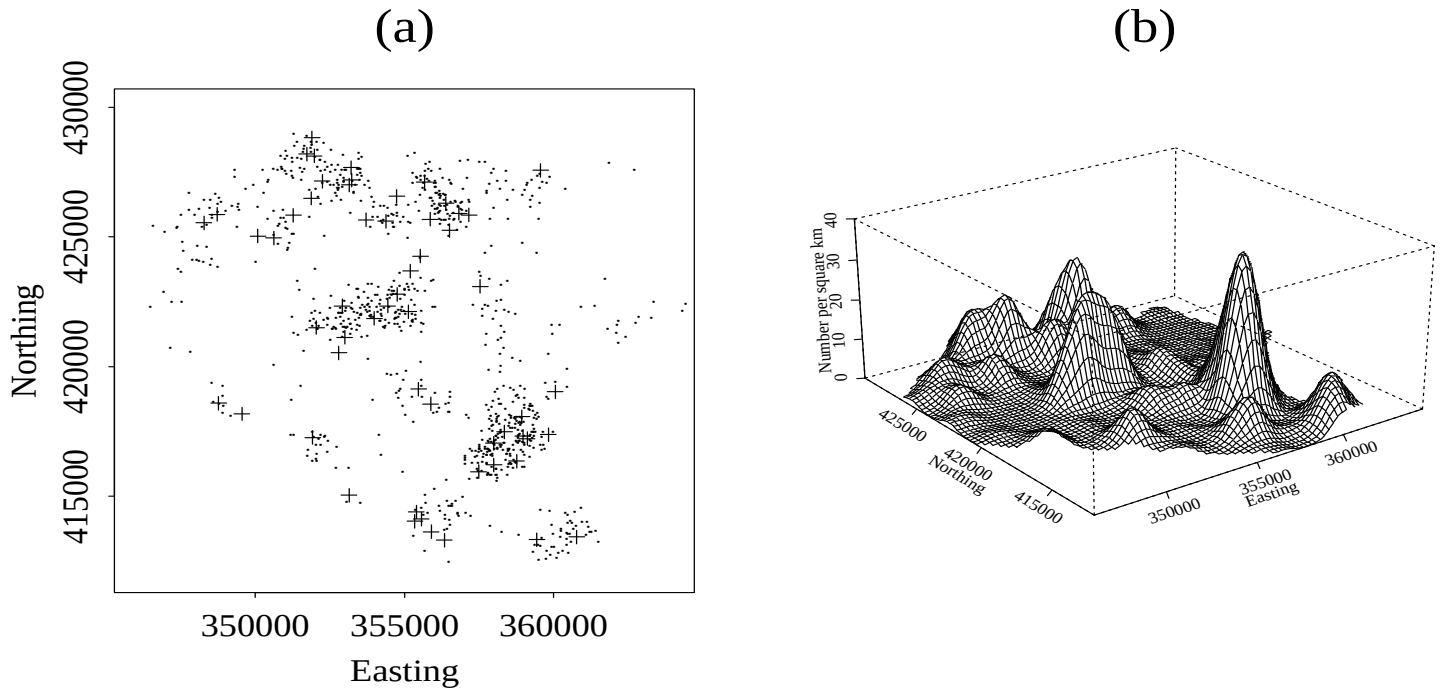


Figure 5: Case-control data in the Chorley-Ribble area of England: (a) Larynx cancer cases (+) and controls (·), (b) Perspective view of kernel density estimate of control data.

Data Quality

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.

Data Quality Issues

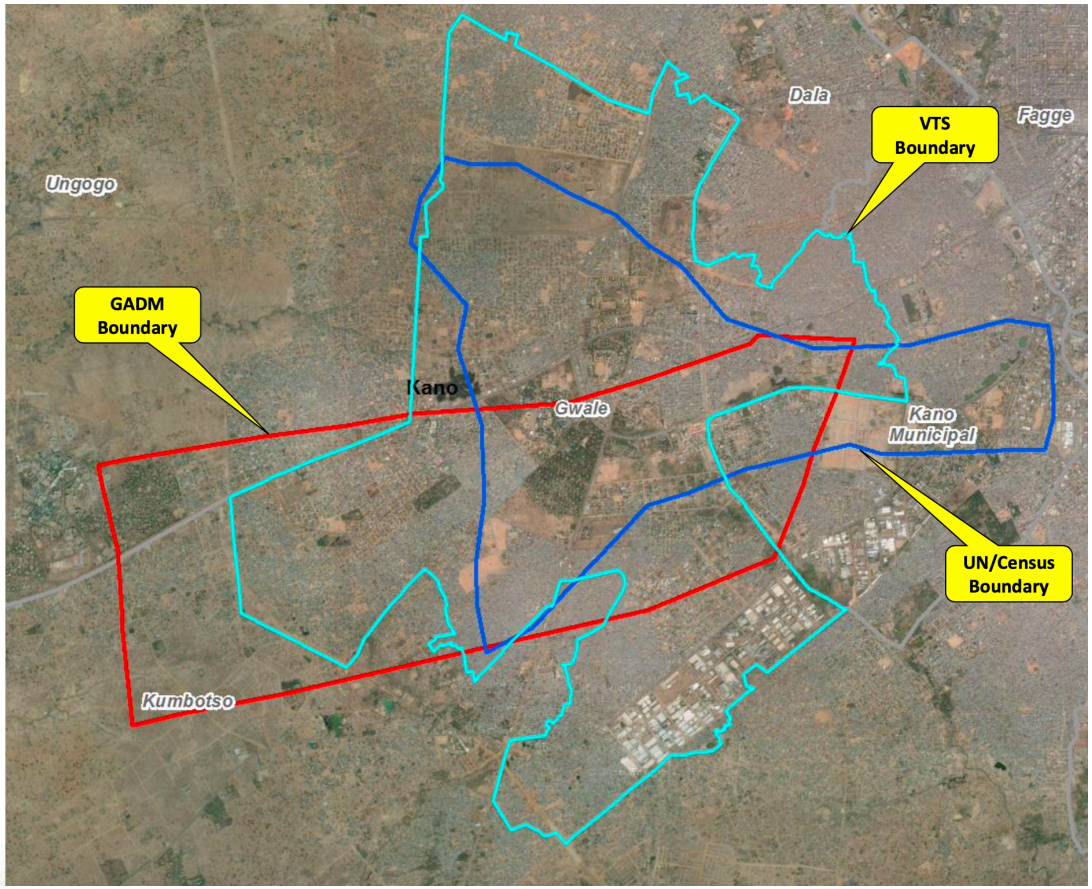
GIS data

- Digitized boundaries from different sources can be wildly different.
- These may have an impact on population estimates within areas.
- Point locations may not map correctly to geographical areas.
- Residential address locations may be very inaccurate, particularly in rural areas.

Uncertainty in Boundaries

Nigeria Sub-National Boundaries from VTS[1], GADM[2] and UN-WHO (Census) all Differ

Gwale LGA, Kano State
Jan 2015



[1]*VTS = Vaccination Tracking System and polio Nigeria geodatabase: http://www.geopode.world*

[2]*GADM = internationally-recognized global boundary resource developed by Robert Hijmans & colleagues at the University of California, Berkeley and the University of California, Davis (Alex Mandel): http://www.gadm.org/*

Figure 6: From Vince Seaman's talk which can be found here: `http://ggim.un.org/meetings/GGIM-committee/7th-Session/side_events/`

Uncertainty in Boundaries

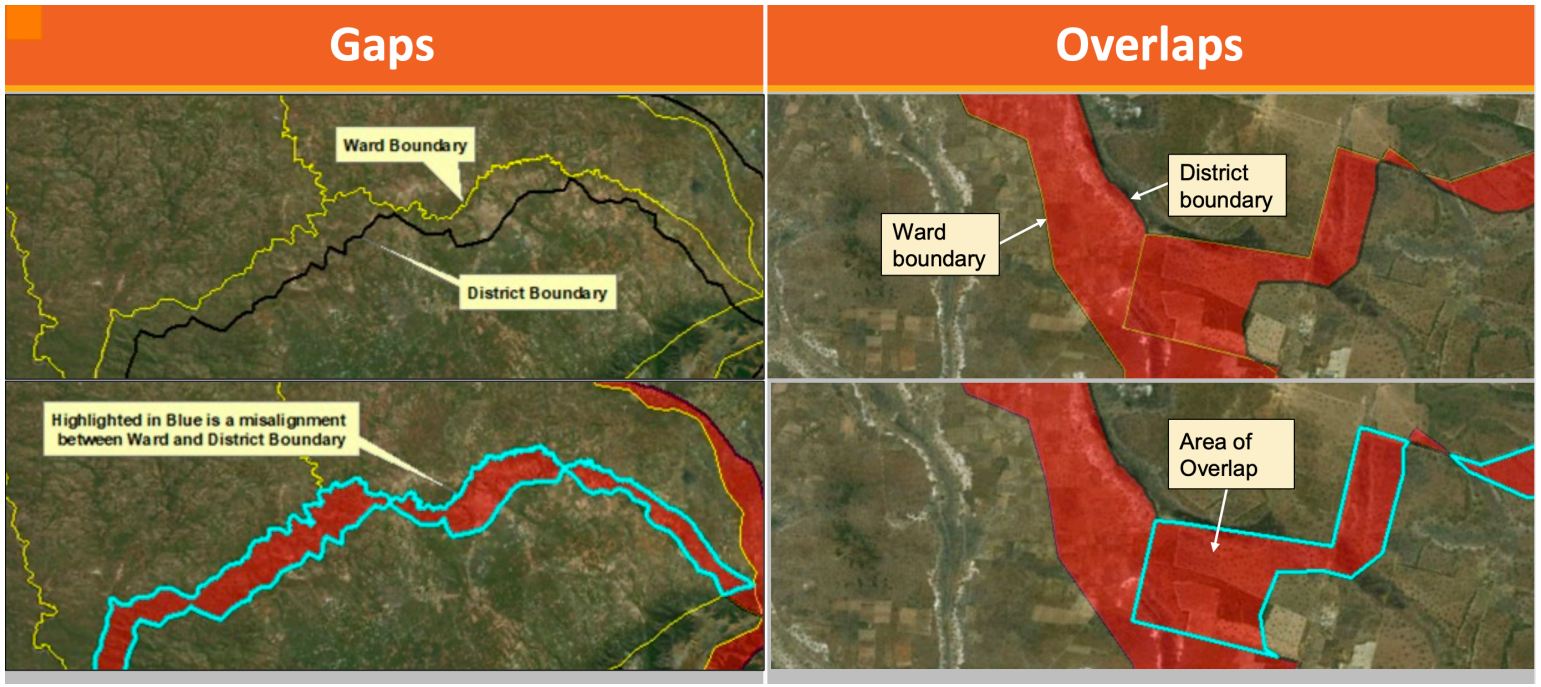


Figure 7: From Lina Pistolesi's talk which can be found here: `https://sedac.ciesin.columbia.edu/binaries/web/global/news/2019/efgs_2019_pistolesi_final.pdf`

Need for Smoothing

Motivating Example: Binomial Count Data

As a motivating example, consider the 48 health reporting areas (HRAs) of King County.

Suppose samples of size n_i in HRA i are taken using simple random sampling (SRS) from the total population N_i, $i = 1, \ldots, 48$.

For each sampled individual, let $d = 0/1$ represent their diabetes status.

The objective is to estimate, in each HRA i:

- The true fractions with diabetes $q_i = D_i/N_i$.

This is a simple examples of prevalence mapping.

Motivating Examples: Normal and Binomial Data

To motivate smoothing, we simulate data in HRAs using simple random sampling in each area.

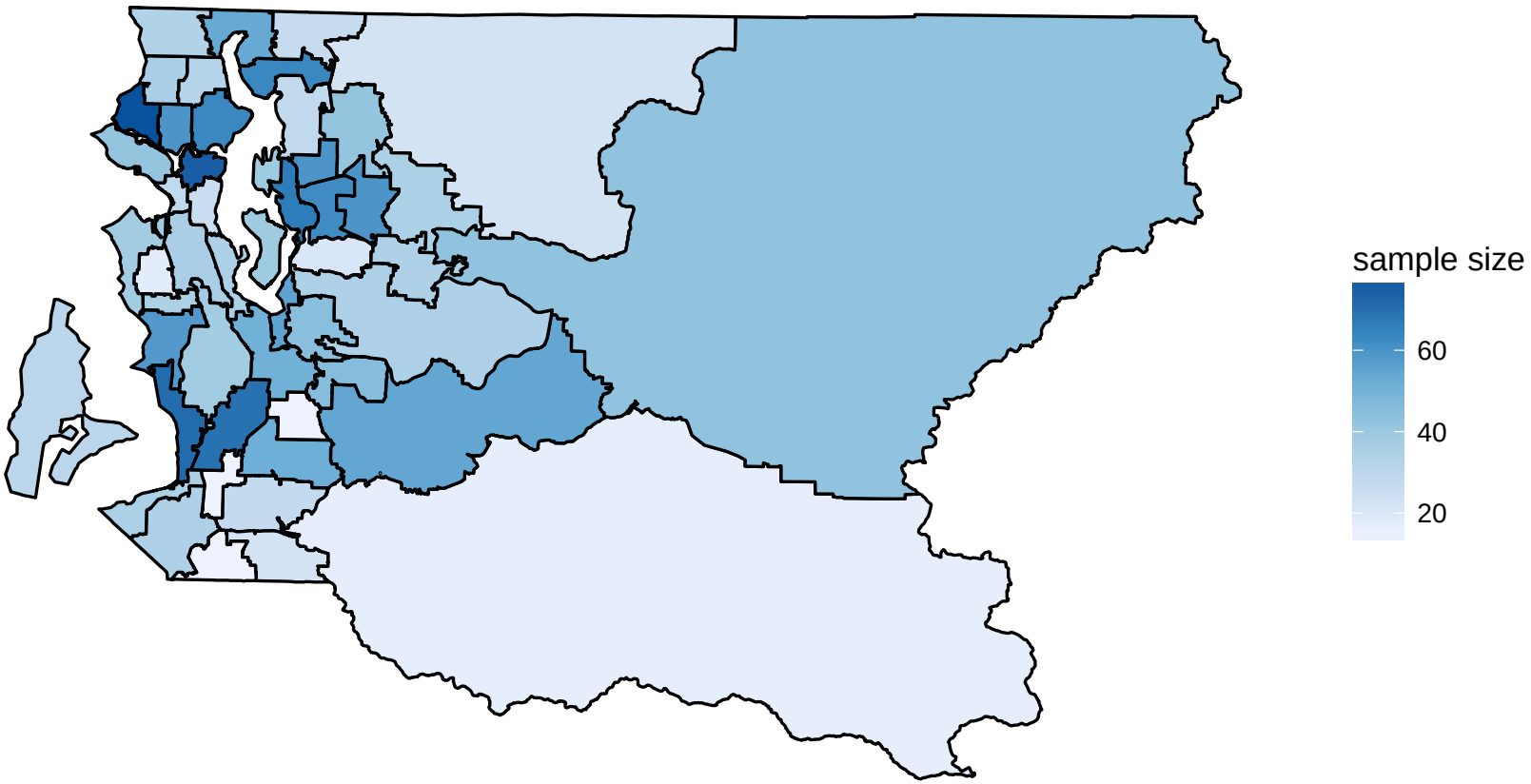


Figure 8: Sample sizes of simulated survey.

Motivating Example: Binomial Data

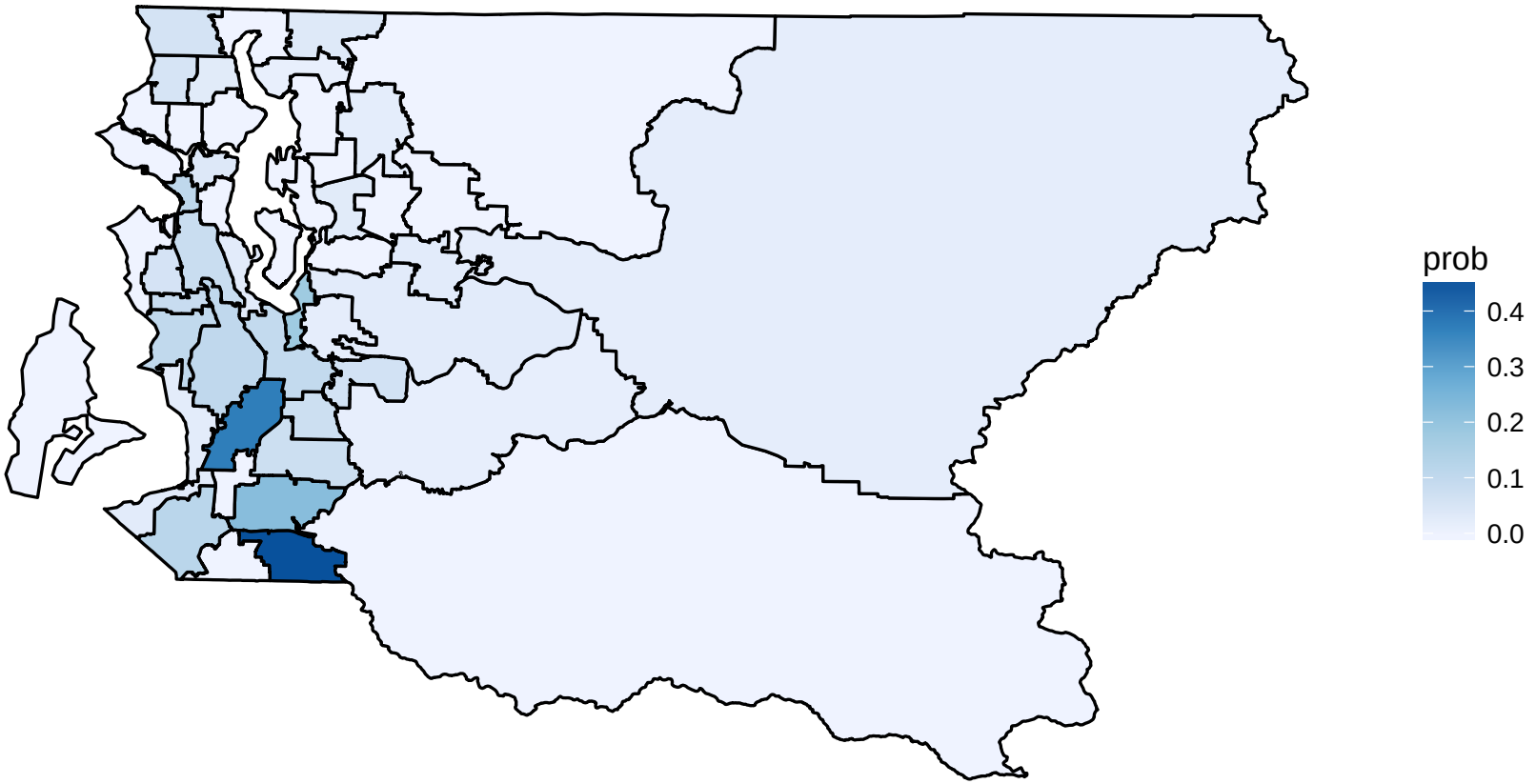


Figure 9: Sample fractions with diabetes.

Motivating Example: Binomial Data

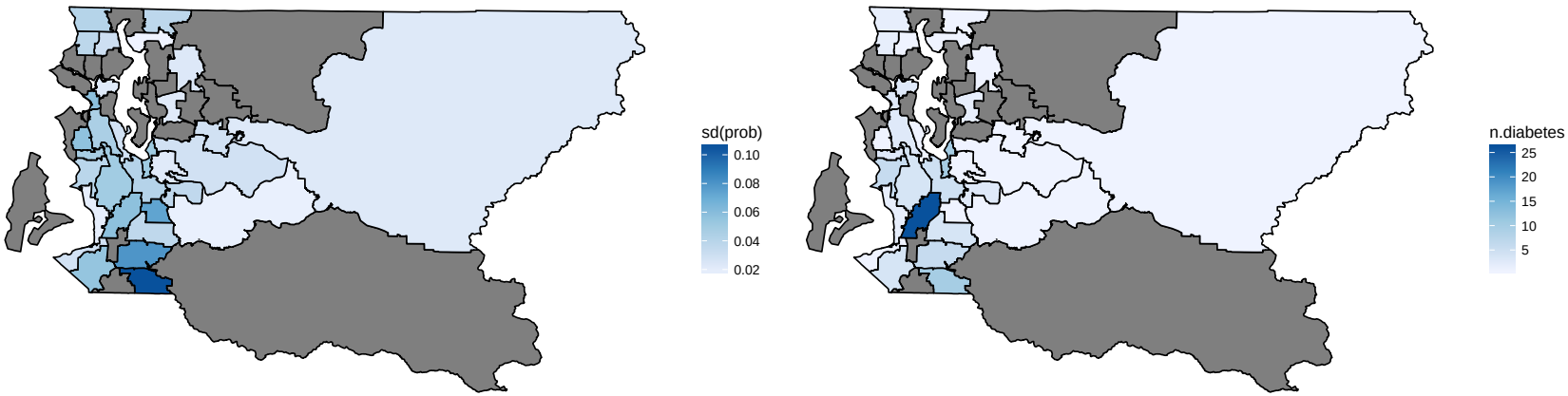


Figure 10: Left: Standard errors of fractions with diabetes. Gray areas correspond to areas with zero counts, and hence an estimated standard error of zero. Right: Number of individuals in the sample with diabetes; zero counts are again indicated in gray. We need to use all the data to assist in the areas with no/little data.

Map Projections and Coordinate Reference Systems

Overview

See Waller and Gotway (2004, Chapter 3) and Bivand et al. (2013, Chapter 4).

It is clearly fundamentally important to have a mechanism to numerically represent spatial location.

Two key ingredients are:

- A coordinate reference system (CRS).
- A map projection.

The CRS allows, amongst other things, datasets to be combined (we may need to transform between projections, which may be achieved using the `spTransform` function in `R`).

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

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; this an angular 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.

Latitude and longitude can be approximated based on a model for the shape of the Earth – known as a datum:

- A simple datum is a spheroid (a sphere that is flattened at the poles and bulges at the equator.
- The most commonly used datum is called WGS84 (World Geodesic System 1984).

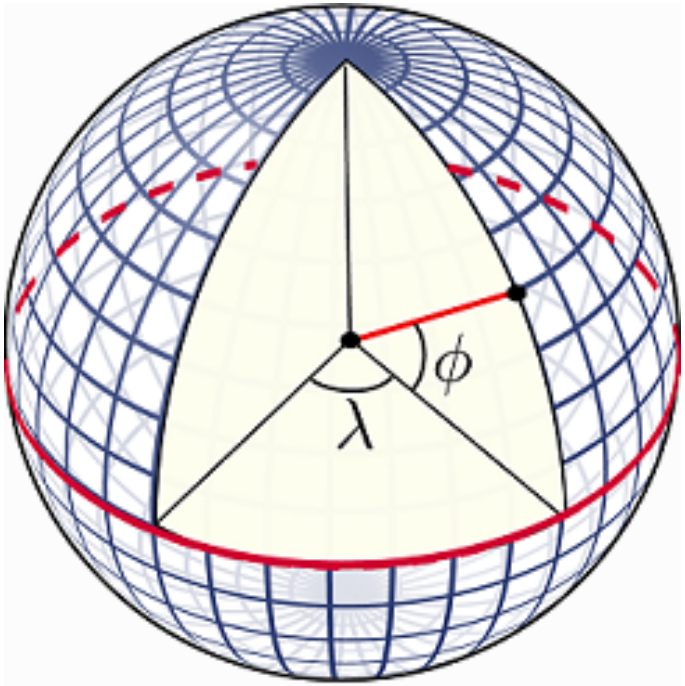


Figure 11: The latitude (ϕ) of a point is the angle between the equatorial plane and the line that passes through a point and the center of the Earth. Longitude (λ) is the angle from a reference meridian (lines of constant longitude) to a meridian that passes through the point.

From `http://www.rspatial.org/spatial/rst/6-crs.html`

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

The process of creating map projections can be understood by considering the positioning of a light source inside a transparent globe on which opaque Earth features are placed.

Then project the feature outlines onto a two-dimensional flat piece of paper.

Different ways of projecting can be produced by surrounding the globe in

- a cylindrical fashion,
- as a cone, or
- as a flat surface.

Each of these methods produces what is called a map projection family.

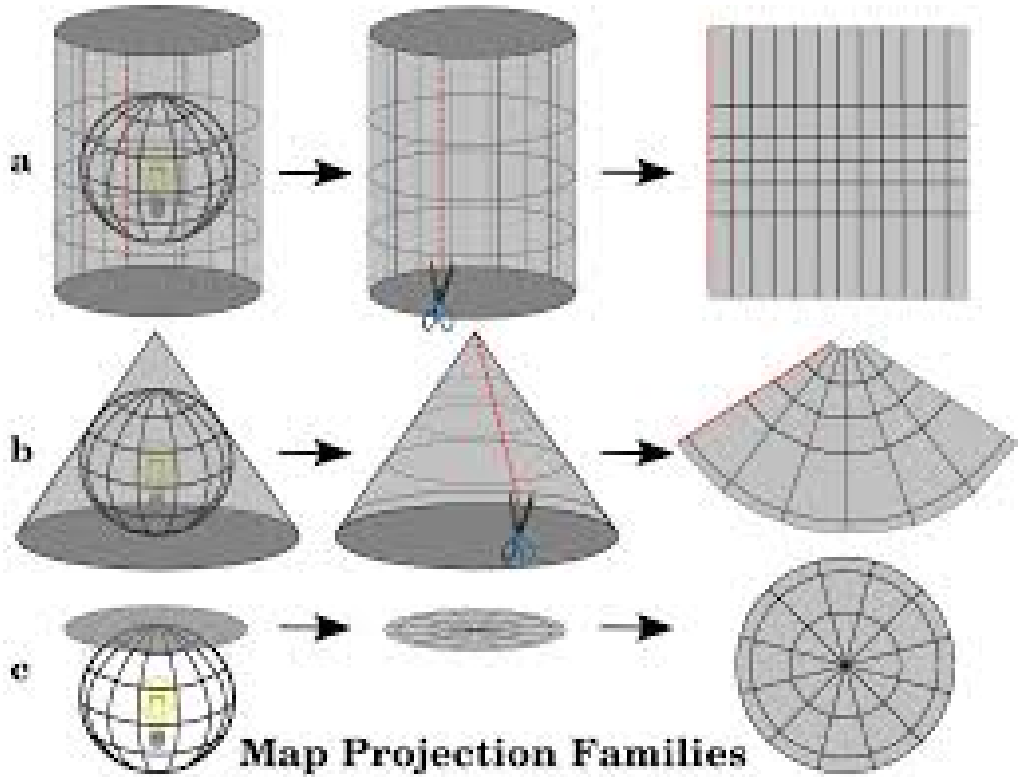


Figure 12: The three projections: a) cylindrical projections, b) conical projections, c) planar (azimuthal) projections.

From: `https://docs.qgis.org/testing/en/docs/gentle_gis_introduction/coordinate_reference_systems.html`

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.

See `https://en.wikipedia.org/wiki/List_of_map_projections` for many examples.

Figure 13: The Mercator projection is a cylindrical projection that is conformal.

Source: By Strebe - Own work, CC BY-SA 3.0,
`https://commons.wikimedia.org/w/index.php?curid=17700069`

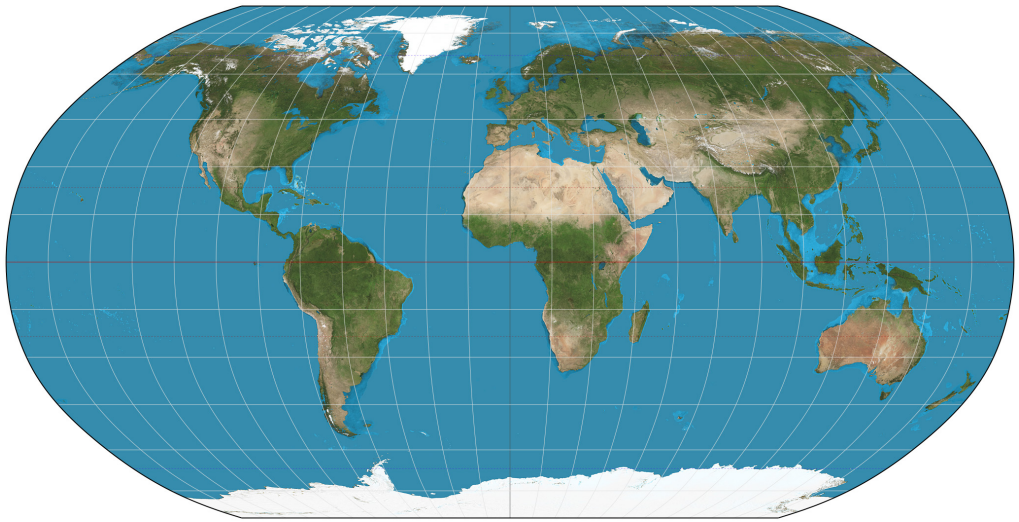


Figure 14: Robinson pseudocylindrical projection is neither equal-area nor conformal, but goes for a compromise.

`https://en.wikipedia.org/wiki/List_of_map_projections`

Figure 15: The UN logo uses the azimuthal Equidistant projection.

Takeaways

- Understand the context!
- Decide on what's the question, and whether the spatial aspect is a blessing or a curse.
- Data quality/abundance is key to dictating the complexity of the spatial analysis that can be performed (if any...).

References

Bivand, R., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R, 2nd Edition*. Springer, New York.

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.

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

Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons.