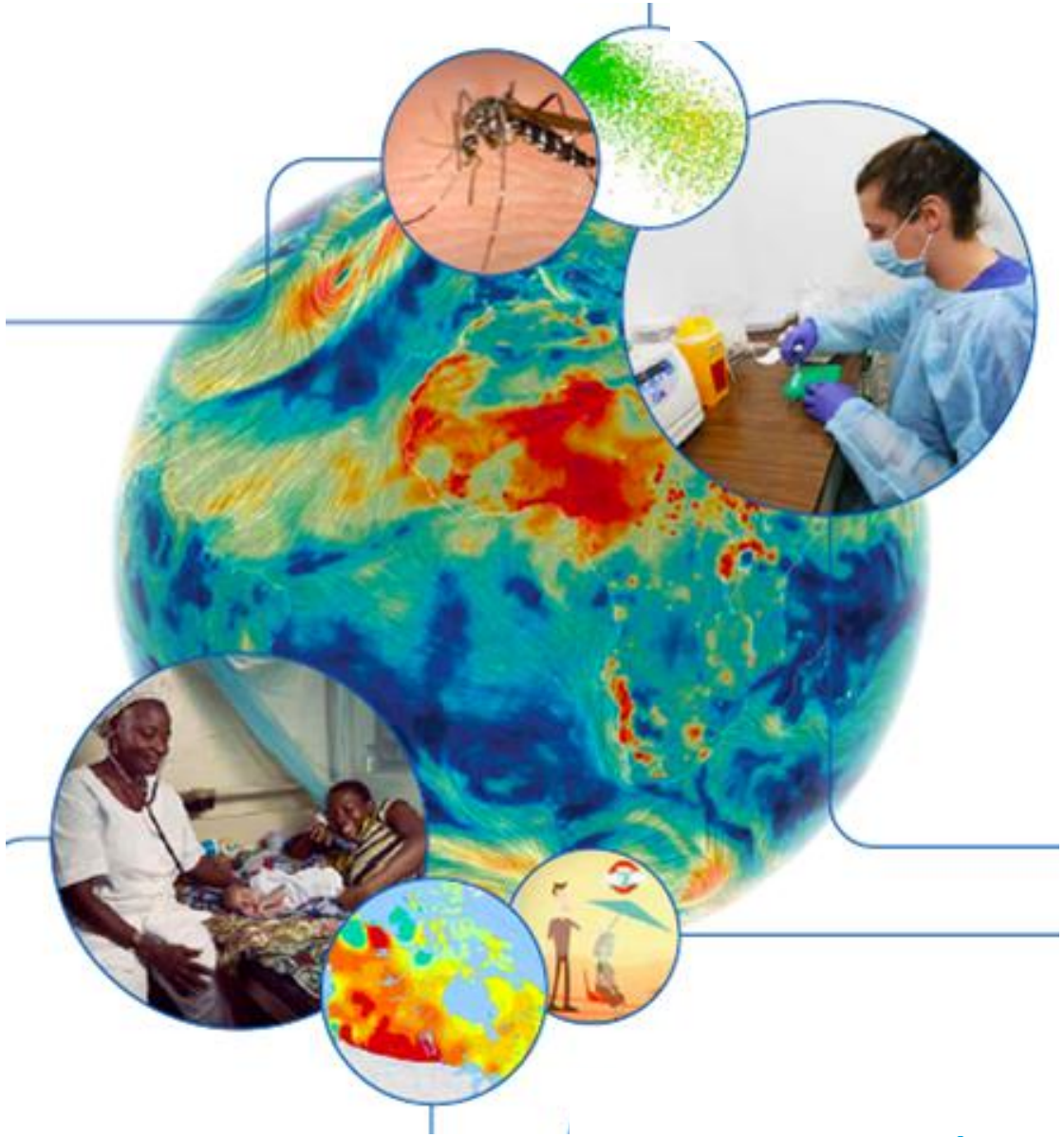
Part 2

EPIDEMIOLOGY BASICS

Improving public health decision-making in a new climate



UF UNIVERSITY of FLORIDA · PAHO · UMEÅ UNIVERSITY · Agence de la santé publique du Canada · LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE · World Health Organization

Part 2: EPIDEMIOLOGY BASICS

Section 2.1. **Overview of basic epidemiological and disease transmission terms and data use**

1. Case definition
2. Infection vs. disease
3. Clinical course of disease
4. Incidence and prevalence
5. Outbreaks, epidemics, disease curves
6. Transmission
7. Disease modelling – simple/compartmental
8. R_0 – basic reproductive number

Section 2.2. **Study design and confounding factors**

1. Climate-epidemiological study designs
2. Standard epidemiological study designs
3. Bias
4. Autocorrelations
5. Causal pathways diagrams
6. Adjusting for confounding
7. Inference

Section 2.3 **Types of data on health outcomes and disease risk – their strengths, weaknesses and best practices for their use**

1. Methodology
2. Specific and sensitivity
3. Spatio-temporal resolution
4. Representativeness
5. Ethics
6. Weather-/climate-driven decision tools for Lyme and WNV
7. Linking health data and weather and climate data

Section 2.4 **Visualising spatial and temporal variation in climate and health data**

1. Fundamental properties of spatial and spatiotemporal data
   - Spatial data visualisation
   - Types of spatial data
   - Commonly used spatial graphs
   - Colour scales
2. Visualising variation across time and space

ANNEX: Introduction to open-access data and key software

Section 2.1:

Overview of basic epidemiological and disease transmission terms and data use

Learning objective:

To be able to describe and define standard epidemiological terms, as encountered in surveillance data used in climate-health analyses for vector-borne diseases.

Case study: Descriptive epidemiology of a Zika outbreak

- Ryan et al, 2018, Zika Virus Outbreak, Barbados, 2015–2016, *Am. J. Trop. Med. Hyg,* 98(6), 2018. doi:10.4269/ajtmh.17-0978 https://www.ajtmh.org/content/journals/10.4269/ajtmh.17-0978

Further reading:

- CDC basic epidemiology definitions https://archive.cdc.gov/#/details?url=https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section1.html
- PAHO Dengue, chikungunya and Zika Regional Epidemiological Update (e.g. 10 June 2024) https://www.paho.org/en/documents/epidemiological-update-dengue-chikungunya-and-zika-10-june-2023
- Roberts, M.G., Heesterbeek, J.A.P. 1993 Bluff your way in Epidemic Models. *Trends in Microbiol.*, Vol. 1, No. 9, p.343- Open Access PDF link: https://www.sciencedirect.com/science/article/pii/0966842X93900753?via%3Dihub
- S.J. Ryan's basic excel SIR model space – Excel file available with this module.

2.1 Overview of basic epidemiological and disease transmission terms and data use

Dr Sadie J. Ryan

University of Florida

Learning objective

To be able to describe and define standard epidemiological terms, as encountered in surveillance data used in climate-health analyses for vector-borne diseases.

- Most terms will therefore be from infectious disease epidemiology, but we will remind the audience of chronic disease terminology, as this is also likely to be encountered by the learner.

© WHO/Joel Lumbala

UF UNIVERSITY of FLORIDA | Pan American Health Organization / World Health Organization Americas | PAHO

World Health Organization

Case definition

- What is a case?
- Important to know – many infectious diseases have asymptomatic cases, in an outbreak situation, or for a novel disease, this may change
  - Case Definition – may be from symptoms, involving a differential diagnosis (Dx)
    - e.g. what is dengue, the recommended case definition involves diagnostic clinical tests, 70-80% cases are asymptomatic, but serologically show infection
    - Case definition for severe dengue, and DSS
    - May take years to establish a case definition
- Why does this matter?
  - How big is the outbreak? How many people have or have had the disease? How can clinicians identify infections? How can clinicians distinguish cases (e.g. dengue or Zika)?
  - Case fatality rate (CFR) – the proportion of people who are defined as cases who then die of that disease
    - people want to know the 'risk' of dying of a disease, particularly in novel disease emergence, or novel outbreak situations, such as will occur in a changing climate
  - Defining 'case' is therefore fundamental
  - Important for comparing impacts between outbreaks or locations
  - Essential to public health intervention

PAHO Recommended case definition
Dengue fever

***Clinical description*:** An acute febrile illness of 2-7 days duration with 2 or more of the following: headache, retro-orbital pain, myalgia, arthralgia, rash, hemorrhagic manifestations, leucopenia.

***Laboratory criteria for diagnosis*:**
One or more of the following:
- Isolation of the dengue virus from serum, plasma, leukocytes, or autopsy samples,
- Demonstration of a fourfold or greater change in reciprocal IgG or IgM antibody titers to one or more dengue virus antigens in paired serum samples,
- Demonstration of dengue virus antigen in autopsy tissue by immunohistochemistry or immunofluorescence or in serum samples by EIA,
- Detection of viral genomic sequences in autopsy tissue, serum or CSF samples by polymerase chain reaction (PCR).

Case classification
Suspected: A case compatible with the clinical description.
Probable: A case compatible with the clinical description, with one or more of the following:
- supportive serology (reciprocal hemagglutination-inhibition antibody titre greater than 1280, comparable IgG EIA titre or positive IgM antibody test in late acute or convalescent-phase serum specimen),
- occurrence at the same location and time as other confirmed cases of dengue fever.

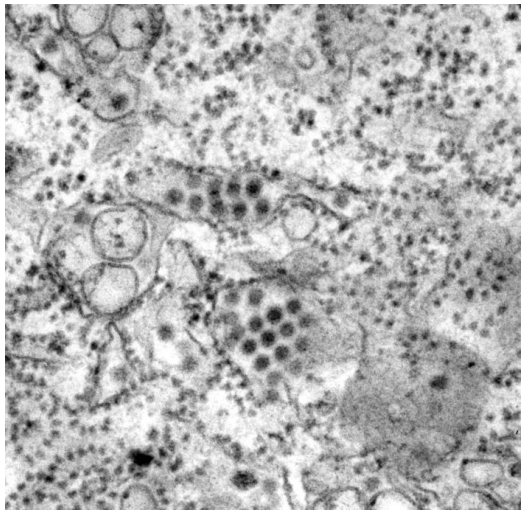
Confirmed: A case compatible with the clinical description, laboratory-confirmed.
https://www.paho.org/english/sha/be_v21n2-cases.htm#Dengue

Infection or Disease?

- **Infection**: Invasion and multiplication of infectious agents inside an organism.

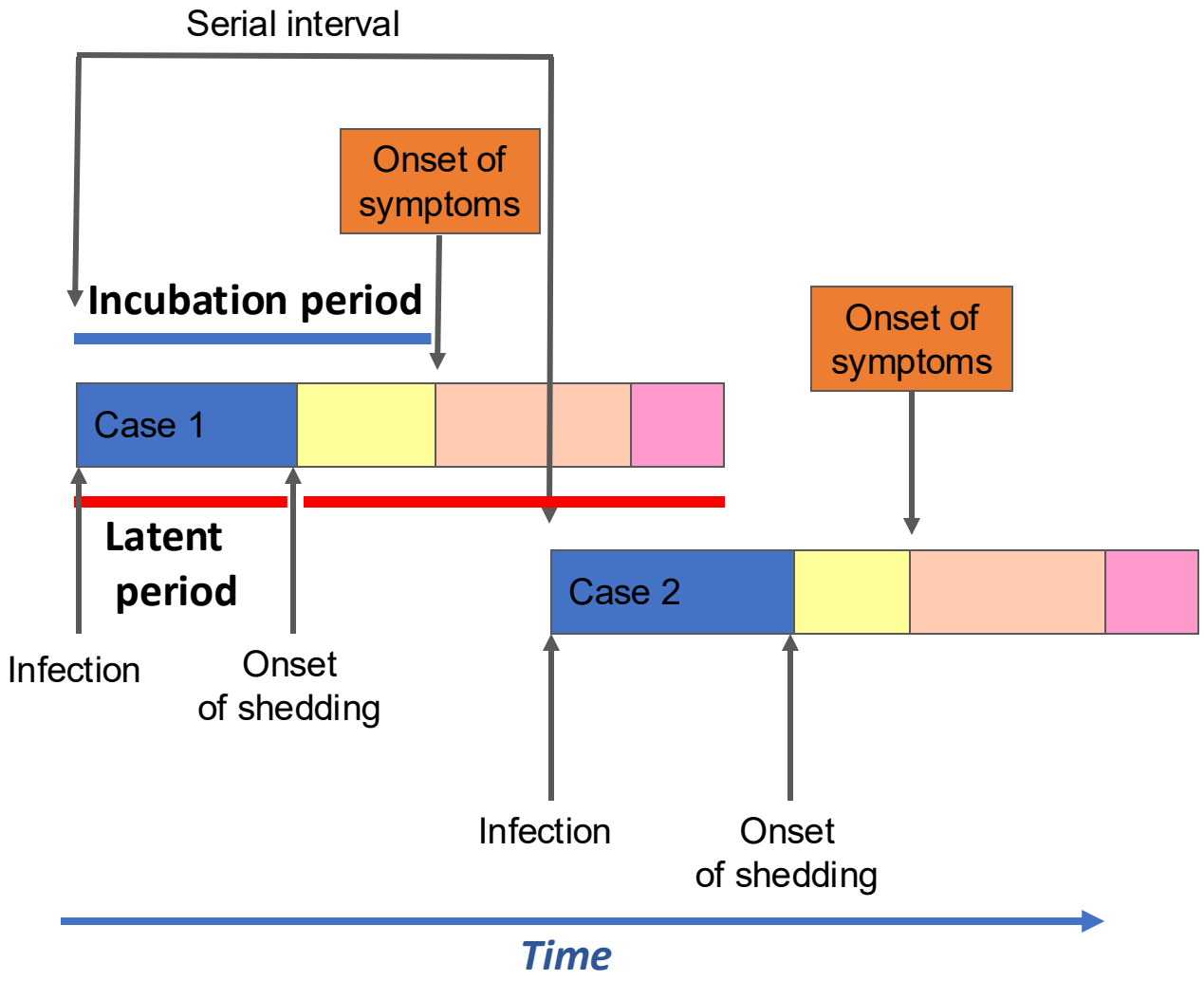
- **Disease**: A deviation from the normal physiological status of an organism that negatively affects its survival or reproduction.

- The state of the individual as infected, infectious, and having the disease, is important for understanding larger dynamics of diseases in populations, and for tracking cases through reporting and treatment or intervention.



Description: This transmission electron microscopic (TEM) image depicts a number of round, Dengue virus particles that were revealed in this tissue specimen.
Link:
https://phil.cdc.gov/Details.aspx?pid=12493
Photo Credit: CDC/ Frederick Murphy

Clinical Course of Disease



- **Incubation period**: time from infection to appearance of symptoms
- **Latent period**: time from infection to beginning of transmission
- **Infectious period**: time during which an individual can transmit the disease
- **Generation time (serial interval)**: Time from infection of one host to infection of a secondary case caused by that host (gives an idea of how fast it spreads through a population).

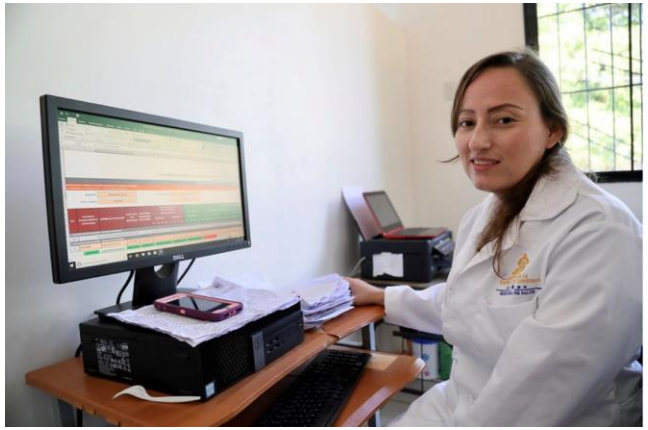
Incidence and prevalence

Two terms often confused

- **Incidence** is the number of **new cases** in a population in a specified period of time
- **Prevalence** is the proportion of **cases** in a population in a specified period of time
- These are often referred to as ‘rates’ because there is a time component; this can be confusing, as it suggests a dynamic that is not necessarily there.

Incidence and prevalence (1)

- **Incidence** is usually expressed in terms of the population, e.g., cases per 10,000, cases per 100,000, but sometimes is simply expressed as new cases, without a denominator
- **Prevalence** is usually expressed as a percentage or fraction – e.g., 0.003% prevalence, or 1 in 3000
- Note: the time component for both prevalence and incidence is often forgotten, so it is useful to remember to think about it – many times country-level rates of disease are expressed as incidence per 10,000 population, reflecting annual estimates. For instance, incidence might refer to cumulative incidence over a year, and prevalence might refer to point prevalence during one month of a year.
- Recalling a previous slide, **case definition** is also important to understanding what incidence and prevalence reflect – usually the number of reported symptomatic cases.
- As pointed out by the Global Burden of Disease report (WHO), people may get malaria or diarrhoea more than once in a reporting year, so they use '**case events**' as the reported unit, using it to generate incidence and prevalence indicators.



https://www.flickr.com/photos/pahowho/48949083442/in/album-72157711471192973/

Photo Credit: PAHO

Puerto Lempira, Honduras - 2019 Malaria Champion

Self-check

- In a population of 100,000 people, a new disease emerged in January, and a case definition was developed in the first month of emergence, wherein cases had a purple rash, high fever for 3 days, coughing, extreme fatigue, and itchy legs.

- In January, there were 300 cases,
  In February 400, and in March 200.
  - *What was the incidence in January?*
  - *What was the 3-month incidence?*
  - *What was the 3-month prevalence?*

https://www.flickr.com/photos/pahowho/49871984971/in/album-72157714224268743/

Photo Credit: PAHO
Vacinando no Brasil contra H1N1

Incidence and Prevalence (2)

CDC recommends the terms incidence proportion and incidence rate:

- **Incidence proportion -** proportion of the population that develops the disease during a specific period of time
    - Synonyms: attack rate, risk, probability of catching disease, cumulative incidence
- **Incidence rate** - person-time incidence, used in the context of long-term cohort studies.
    - For example, a set of people is studied for 10 years, and their disease state is recorded yearly over those 10 years. The incidence rate is therefore the ratio of the number of cases to the total time the population is at risk (people × time in the study).
- The distinction between incidence and prevalence in the CDC description is that **prevalence** includes all cases of the disease, both new and pre-existing
- **There are many ways in which disease rates are reported, so it is important to understand how any given report defines them**
    - **Similarly, when reporting rates, it is best to include your definition very explicitly and clearly**

https://archive.cdc.gov/#/details?url=https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section1.html

Incidence and Prevalence (3)

Photo Credit: PAHO
Puerto Lempira, Honduras - 2019 Malaria Champion

https://www.flickr.com/photos/pahowho/48948342968/in/album-72157711471192973/

- **Attack rate** – the rate of change in case numbers – identical to incidence in outbreak descriptions
- Usually used to describe small-number situations

Attack rate = number of new cases in the population at risk/

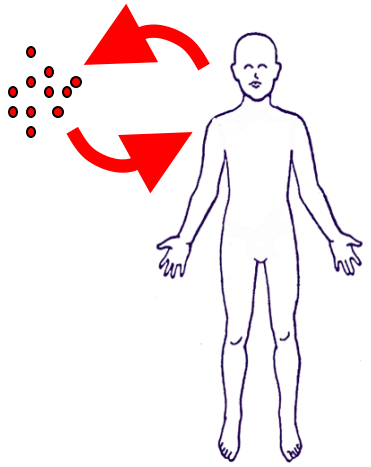
number of people in the population at risk

  - If we are in a room of 10 people and tomorrow 3 people are infected, in those 10 people, that is an attack rate of 30%.
  - Attack rates are useful in closed population epidemiology – where contact tracing is possible.
- **Burden** – this term is used to describe the number of cases OR to measure the impact of disease, often for non-lethal outcomes
  - Two common burden indicators are DALYs and QALYs
  - Disability adjusted life years (DALYs) and Quality adjusted life years (QALYs) – estimated years of life lost due to disability or morbidity impacts
  - Burden is often estimated as a financial or productivity loss to a country
- **Seroprevalence** - related to prevalence – it measures the serological indicator (e.g., antigens) that someone has been infected, and thus is a measure of the proportion testing positive of having had the infection.

Outbreaks, epidemics, disease curves

- An **outbreak** is defined as the occurrence of cases in excess of what would normally be expected in an area, community, or timeframe. However, for outbreaks of a novel infection, or infections new to the area, even 3 cases can constitute an outbreak.
- An **epidemic** is a continued or sustained outbreak, which is defined by an increasing curve of new cases, and, usually, a subsequent decrease in cases to a low or no cases time period
- When a disease continuously circulates within a population, it is **endemic**, that is, there is year-round transmission
    - However, seasonal increases in transmission can create epidemic behaviour, so this is often referred to as an epidemic, even if cases do not decrease to zero between waves
    - It is important to characterise the shape of epidemics
        - If there are start and end timing signals, this is important to climate-health data analyses
        - If there is a clear peak timing, that may also correspond to a climate signal for climate-sensitive diseases

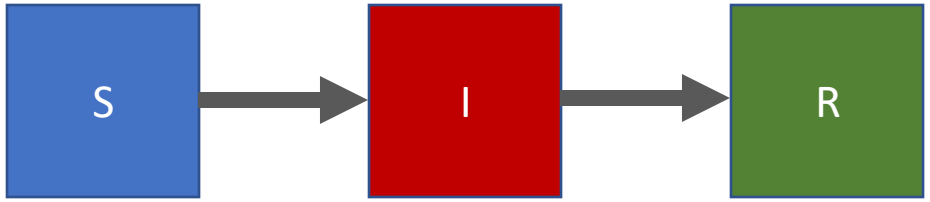
Transmission

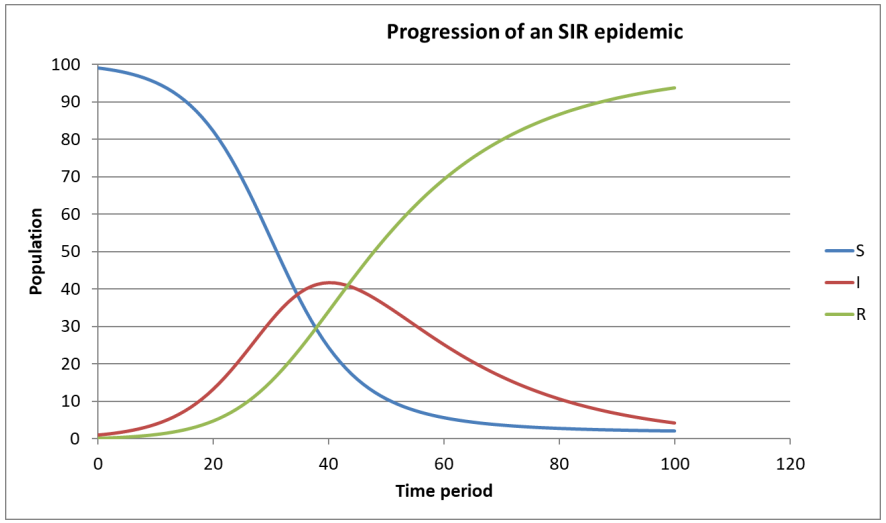


- Two main types of infectious disease transmission addressed in this course are:
  - **Direct** – for example, respiratory diseases, wherein aerosol droplets are the mode of spread – you cough on/near someone, and they inhale infectious particles
  - **Indirect** – many types of indirect spread exist
    - Extending the aerosol concept, **contaminated surface** transmission can occur
    - **Faecal-oral** transmission (infectious particles re-enter orally)
    - **Water or food** contamination
    - **Vector-borne** transmission
      - Human-arthropod-human (malaria, dengue, tick-borne diseases, typhus)
      - Human-water-snail-water-human (and/or other vertebrates) (schistosomiasis)
      - Spillover-spillback to wildlife/livestock reservoirs (influenzas, Ebolas)

Disease modeling – simple compartmental

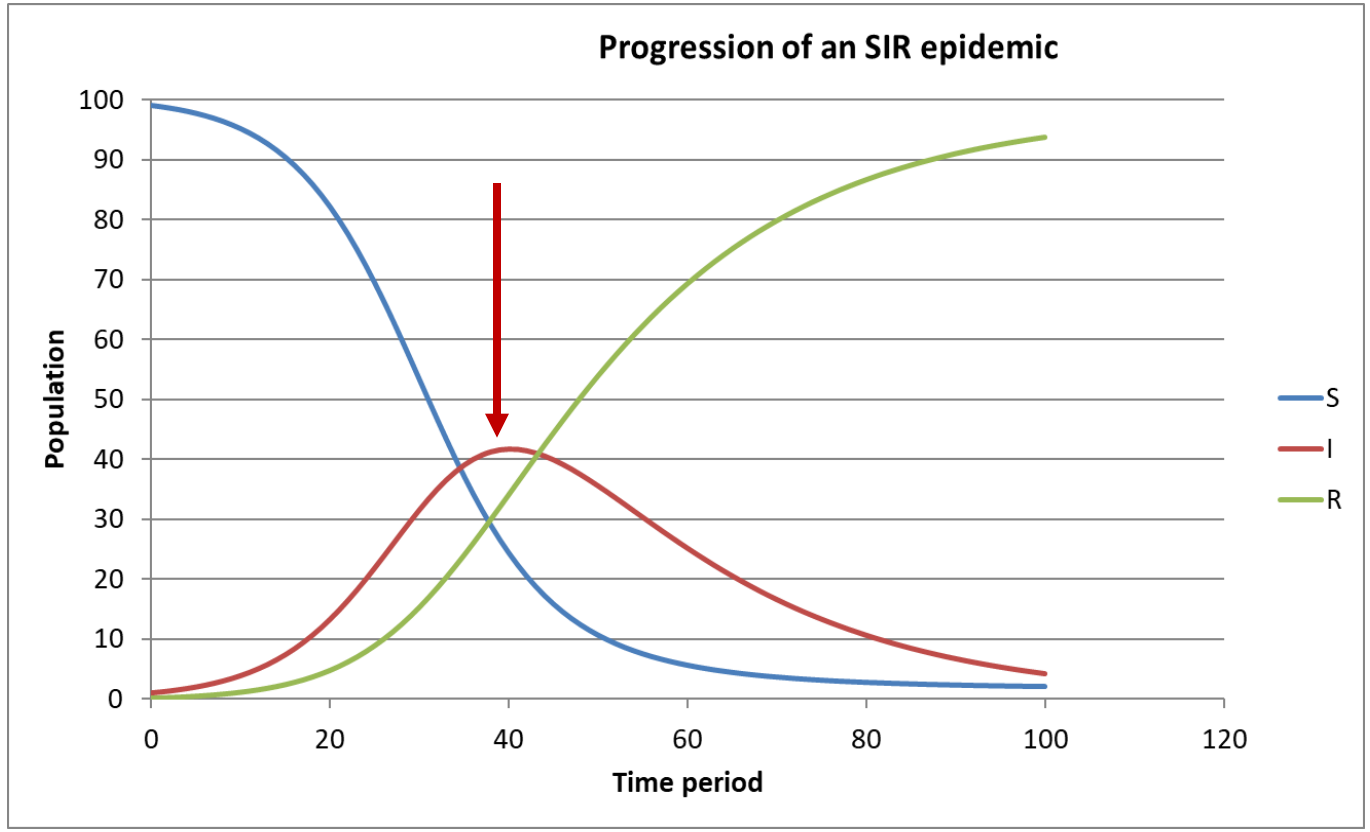
- For infectious disease spread within a population, we assume a starting point of a naive and uninfected population- Susceptible, S
- As susceptible individuals contact infected individuals and become infected, they transition to Infected, I
- If we assume that Infected individuals recover from infection, have immunity to the infection, and do not die of the infection, they transition to the recovered class, R
- We refer to this as S-I-R progression, and the SIR model is a fundamental model in disease modelling
- We assume a simple, closed population, with no birth or death dynamics, and this leads to simple progression.





Why does an epidemic peak in a simple SIR?

- In this very simple case illustration, we see that in a population of 100 individuals, the susceptible pool, S, is drained as it becomes infected, I; however, I transitions to recovered, R, also reducing I.
- This means that the epidemic rises to a peak, and declines, and when and how high is determined by the rates of transitions between the 3 states in the population.
- This graph is generated using a spreadsheet created by S.J. Ryan, which is available with this course, to adjust the rates and understand the effects this has on the shape of the curves and the speed of the epidemic progression.



UF UNIVERSITY of FLORIDA PAHO

World Health Organization

R_0: The Basic Reproductive Number

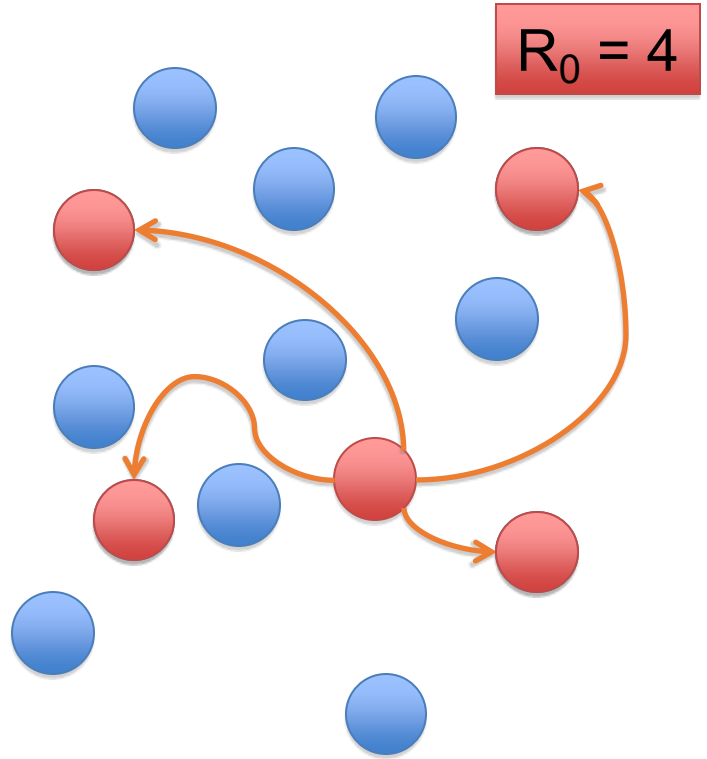
The average number of secondary infections that an infected host produces in an otherwise susceptible population.

Threshold criterion:

If $R_0 < 1$, disease dies out

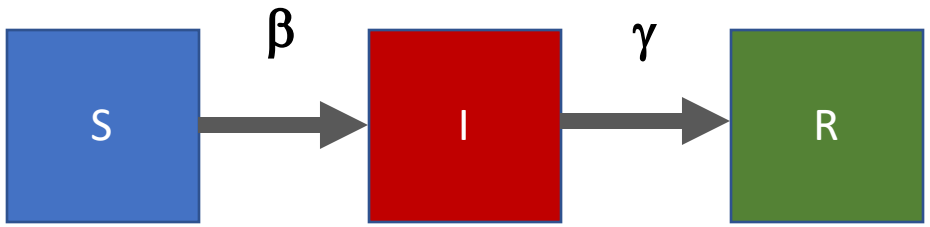
If $R_0 > 1$, disease persists

The graph on the right demonstrates the process in a small population



In this example, R_0 (pronounced R naught) = 4, thus, this infected individual produces 4 secondary infections.

SIR Compartmental model



Bringing together the concepts for the compartmental model illustration on the process of an infectious disease moving through a population, and the threshold criterion for establishment and spread, R0, **rate models** can be used as a function of time (t) to describe the movement of S to I and I to R, in terms of the transmission rate β (beta), and the recovery rate γ (gamma).

Rates

$dS/dt = -\beta SI$

$dI/dt = \beta SI - \gamma I$

$dR/dt = \gamma I$

$R_0 = \beta/\gamma$ you can infect β susceptibles and have $1/\gamma$ time to do it

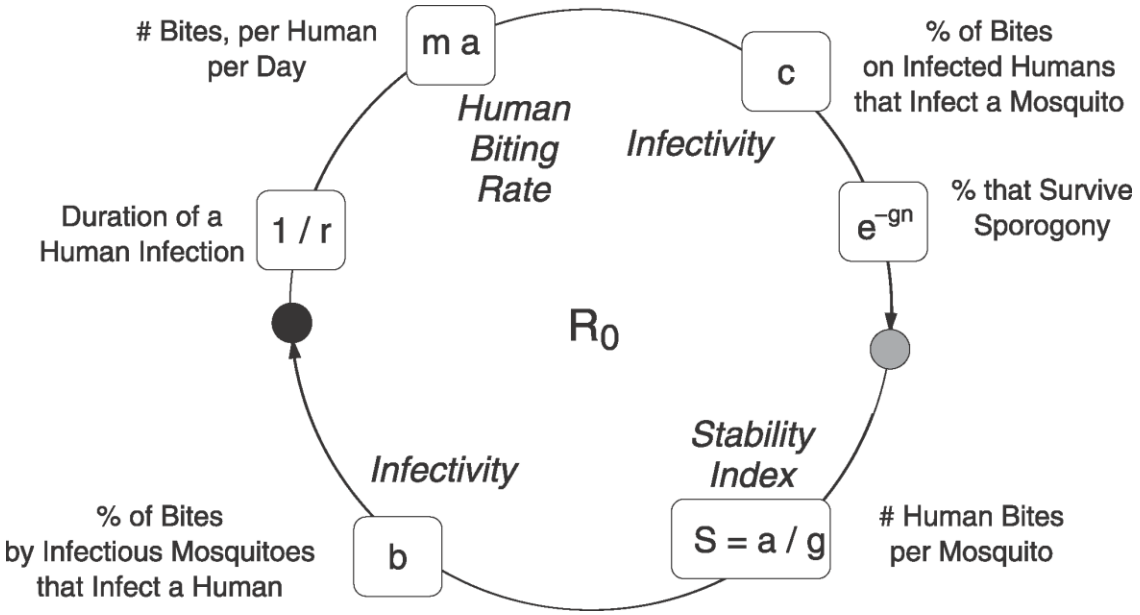
Recent estimates of R_0 for seasonal and pandemic flu typically range from 1.5 to 3. Estimates of the reproductive number (R) from England and Wales (1958-1973), for a mixture of influenza types and subtypes, ranged from 1.4 to 2.6. (1)

In contrast, SARS had an R_0 of 3 (excluding super-spreaders), and measles has an R_0 of 10 to 15, pertussis (16 - 18) or polio (8 - 12).

(1) http://www.globalsecurity.org/security/ops/hsc-scen-3_flu-transmission.htm, retrieved 06/19/2020

R_0 for vector-borne diseases

- Estimating R_0 for vector-borne diseases is complicated by the fact that this is indirect transmission, and therefore involves both a vector infection and a human infection cycle.
- For malaria, we see the description of R_0 on the right
- R_0 for malaria is estimated to range from 1-3000, in an analysis of 121 studies in Africa (D. Smith, 2007)



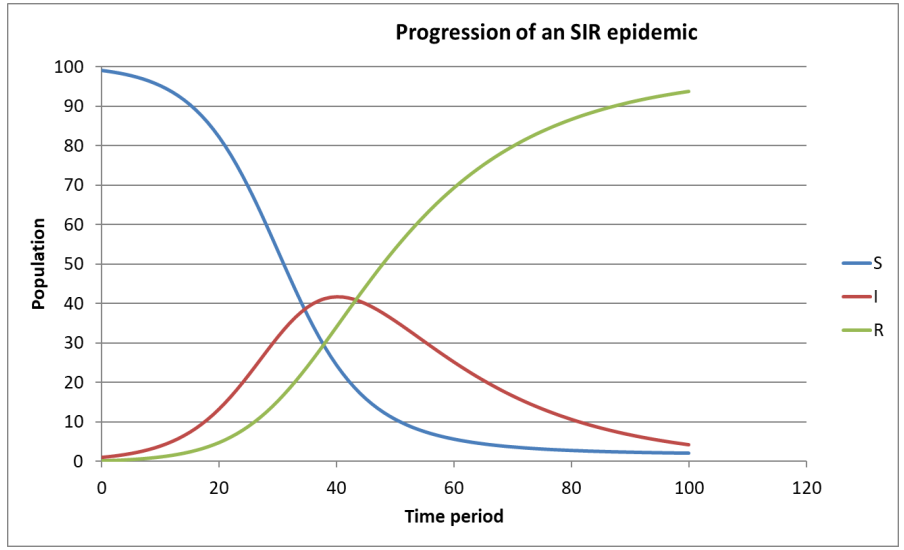
https://doi.org/10.1371/journal.pbio.0050042.g001

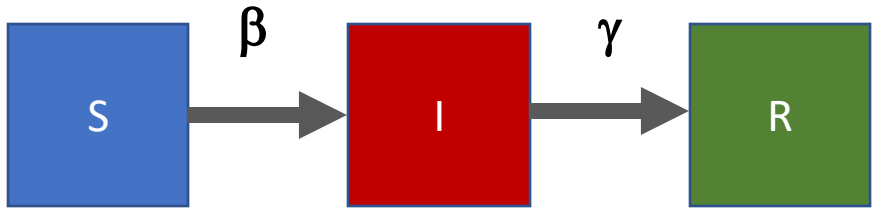
$$R_0 = ma^2bce^{-gn}/rg$$

Can we measure R_0?

Instead of constructing a model of rates, can we measure R_0 more directly during an epidemic or outbreak?

- If $R_0 = \beta/\gamma$, we can estimate these rates by following the cases closely – **contact tracing**
  - β – the rate of infection - is a combination of τ (tau), **transmission per contact**, and c - the contacts made (number of contacts while infectious)
  - To get an **estimate of** τ, we can start to estimate it from contact tracing of infected individuals, and via lab infection experiments, such as exposing mice to aerosolised pathogens at differing concentrations
  - **To measure** γ – the rate of recovery, we can use a combination of case recall information (date of onset of symptoms, date of recovery), and refine tests to differentiate between active infectious status and past infections. With these, we can narrow down the recovery timeline.



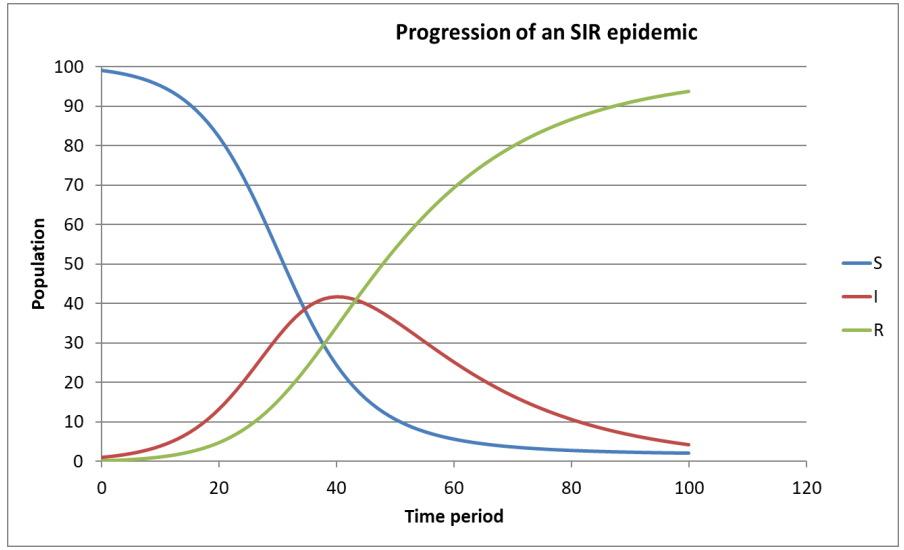


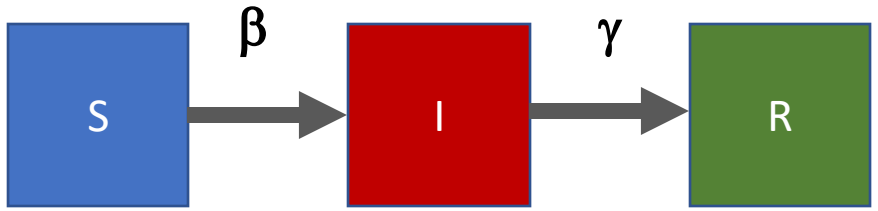
UF UNIVERSITY of FLORIDA PAHO

World Health Organization

Can we measure R_0?

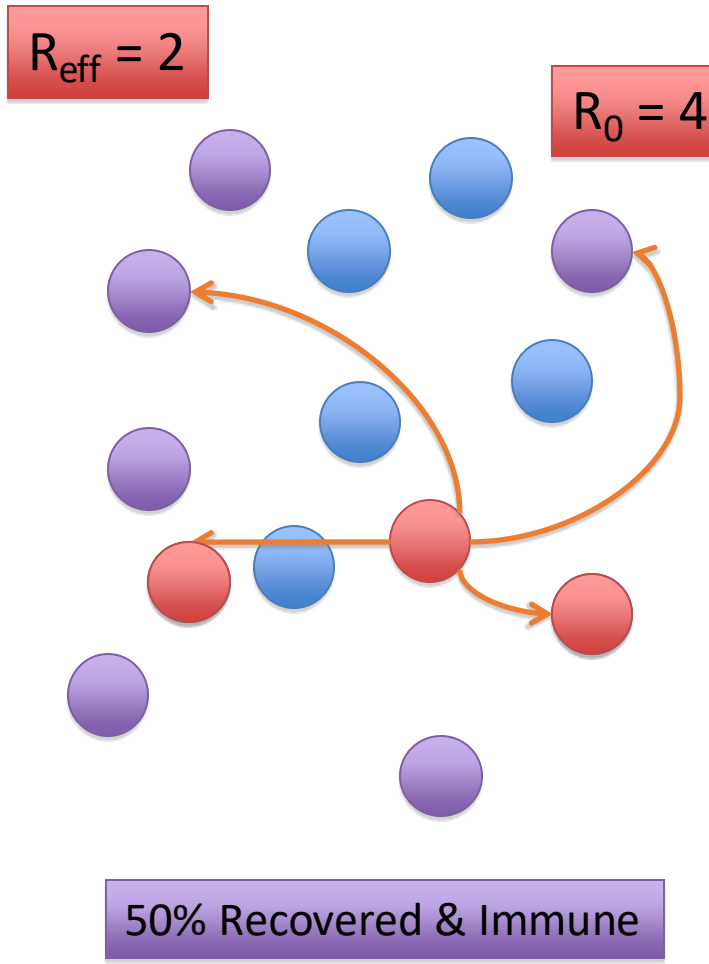
- Contact tracing is very logistically intensive, and not ideal at the beginning of an epidemic, when the information would be most useful.
- Thus, the other main approach is curve-fitting to reported case data early in an epidemic, and estimating how fast cases are increasing in given populations – the rate of change of incidence. Recall the curve for SIR.
- That is, at the onset of an epidemic, R_0 can be directly measured as the rate of change of incidence – the slope of the curve of new infections in a population, assuming this is a novel infection in a susceptible population
- This is limited by how rapidly case reporting occurs and the capacity to identify cases.
    - Recalling the epidemic curve, we would need to know where in the curve we are when we measure the rate of change.
    - At the start of the curve, we usually see $R_0>1$, and this continues until we are approaching the peak. At this point, we are reaching our threshold, and due to saturating the susceptible population, we are no longer measuring R_0, but R_{eff} ('R effective')– the effective rate of reproduction of disease in the population. This will be more apparent in the following slides.





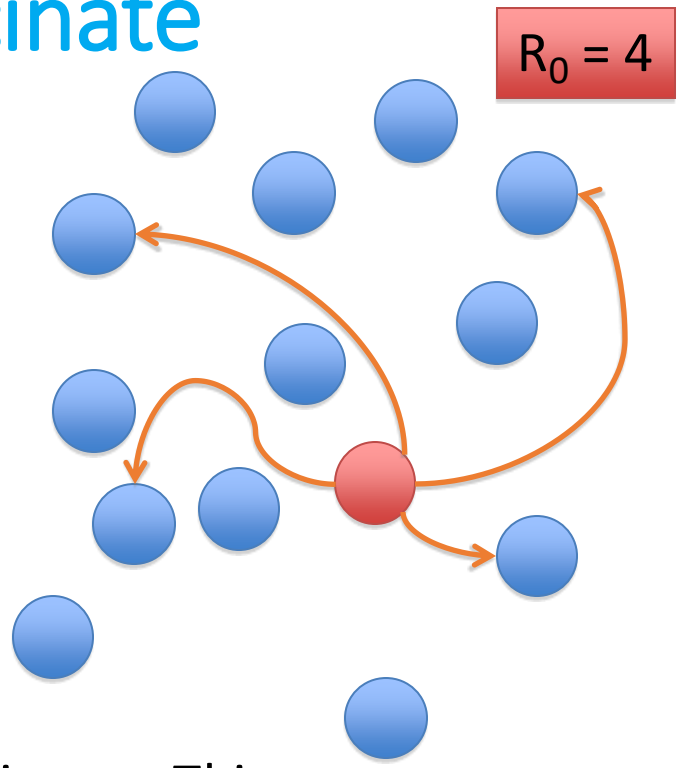
Why we want to know R_0, and R_{eff}: The Effective Reproductive Number

- R_{eff}: The average number of secondary infections that an infected host produces in a real or specified population.
- Example: 50% of the population is immune (due to previous infection or vaccination).
- $R_{eff} = R_0 * (S/N)$
  - S is susceptible, N is the total population, so S/N is the susceptible proportion
- R_{eff} allows us to ask the important public health question of what proportion of the population we need to vaccinate to drive R_{eff} to below 1.



While R_0 assumes a completely suspectible population, R_{eff} assumes that some inidviduals are immune

Using R_0 - Proportion to vaccinate



- For a disease to die out, $R_0 * (S/N) < 1$
  This is the same as saying $R_{eff} < 1$
  Rearrange this a little:
  $(S/N) < 1 / R_0$

- Meaning the proportion that must be vaccinated (e.g., not susceptible) is:
  $1 - (S/N) > 1 - 1/R_0$
  $1 - (S/N) > (R_0 - 1)/R_0$

- So you must vaccinate at least $\boldsymbol{(R_0 - 1)/R_0}$ to locally eliminate a disease. This is also referred to as the critical vaccination threshold, **Vc**, or the threshold for herd immunity

- **Hence, to eliminate a disease, it is not necessary to vaccinate everyone**

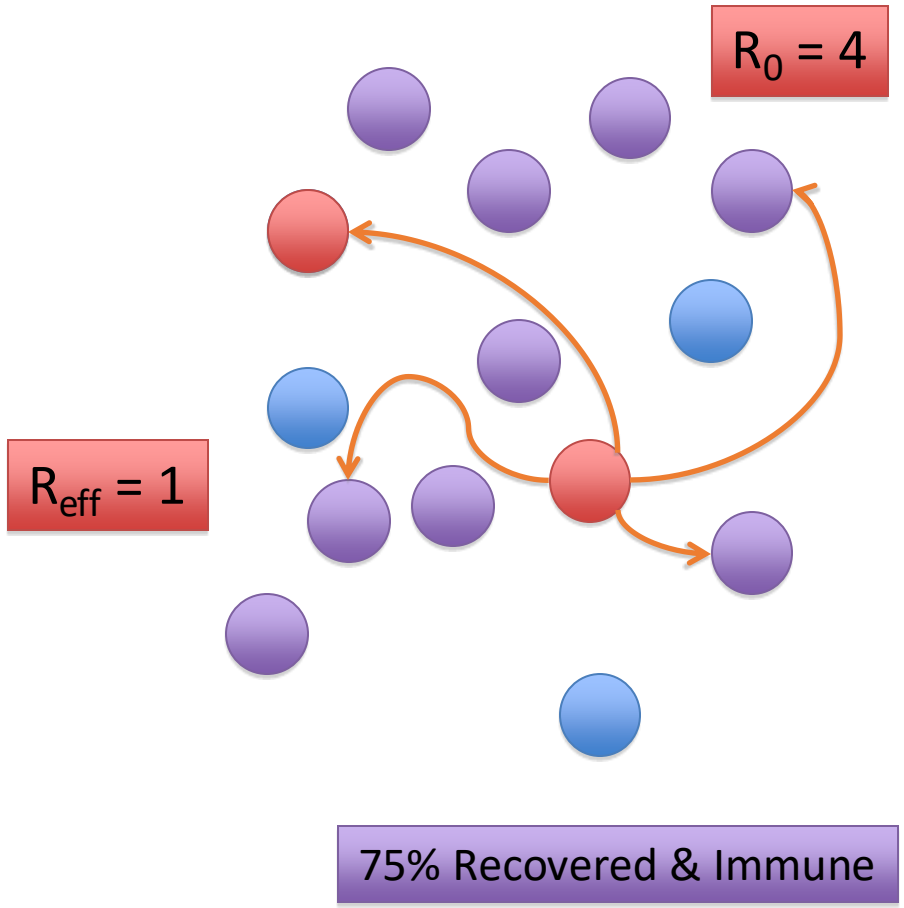
Exercise in proportion to vaccinate

Using our example again

$R_0 = 4$

$Vc = (R_0 - 1)/R_0 = 3/4 = 0.75$

- In the graph, we see that even with R_0 of 4, this disease is reduced to R_{eff} of 1, the threshold for persistence. Thus, vaccination at a rate greater than 0.75 would lead to local elimination.



Reflection exercise

1. An earlier slide described R_0 for a few diseases:

"In contrast, SARS had an R_0 of 3 (excluding super-spreaders), and measles has an R_0 of 10 to 15, pertussis (16 - 18) or polio (8 - 12)"

What is the critical vaccination level for these diseases, and what does this suggest for childhood disease (e.g. pertussis) vaccination levels?

2. Malaria R_0 was estimated in a range of 0-3,000 – in scenarios at the higher end of this extreme, how realistic is vaccination coverage?

Photo Credit: PAHO
Vacinando no Brasil contra H1N1

https://www.flickr.com/photos/pahowho/49871985081/in/album-72157714224268743/

Reflect and calculate before moving to the next slide

Reflection exercise - Results

1. SARS Vc = 2/3 or 0.66666;

Measles Vc = 0.90 - 0.93;

Pertussis Vc =0.9375-0.9444;

Polio Vc = 0.875-0.9167

=> very high coverage needed for the childhood diseases, which is why outbreaks keep coming back as soon as we let our guard down

2. Malaria – Vaccination rate is not realistic for very high biting burden scenarios; other prevention methods will reduce impact, such as biting prevention, transmission blocking, insecticides, etc.

Photo Credit: PAHO
Vacinando no Brasil contra H1N1

https://www.flickr.com/photos/pahowho/49871985081/in/album-72157714224268743/

BIBLIOGRAPHIC REFERENCES

- Smith DL, McKenzie FE, Snow RW, Hay SI (2007) Revisiting the Basic Reproductive Number for Malaria and Its Implications for Malaria Control. PLoS Biol 5(3): e42. https://doi.org/10.1371/journal.pbio.0050042

Additional materials and readings

Case study: Descriptive epidemiology of a Zika outbreak

- Ryan et al, 2018, Zika Virus Outbreak, Barbados, 2015–2016, *Am. J. Trop. Med. Hyg,* 98(6), 2018. doi:10.4269/ajtmh.17-0978 https://www.ajtmh.org/content/journals/10.4269/ajtmh.17-0978

Case study: Description of 2 years of dengue and Chikungunya epidemiology and characteristics in a cluster-based study.

- Stewart Ibarra et al., 2018. The Burden of Dengue Fever and Chikungunya in Southern Coastal Ecuador: Epidemiology, Clinical Presentation, and Phylogenetics from the First Two Years of a Prospective Study. *Am J. Trop Med & Hyg*. https://doi.org/10.4269/ajtmh.17-0762 (Open Access)

Resources:

- CDC basic epidemiology definitions https://archive.cdc.gov/#/details?url=https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section1.html
- PAHO Dengue, chikungunya and Zika Regional Epidemiological Update (e.g., June 2023) https://www.paho.org/en/documents/epidemiological-update-dengue-chikungunya-and-zika-10-june-2023
- Roberts, M.G., Heesterbeek, J.A.P. 1993 Bluff your way in Epidemic Models. *Trends in Microbiol.*, Vol. 1, No. 9, p.343- Open Access PDF link: https://dspace.library.uu.nl/bitstream/handle/1874/8075/heesterbeek_93_bluff_epidemic.pdf
- S.J. Ryan's basic excel SIR model space – Excel file available with this module – Email us

Section 2.2:

Study design and confounding factors

Learning objective: To gain a basic understanding of climate-epidemiological study design and the notion of confounding factors.

Further reading:

- Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. Lippincott Williams & Wilkins; 2008.
- Peng RD, Dominici F. Statistical methods for environmental epidemiology with R. R: a case study in air pollution and health. 2008.
- Moraga P. Geospatial health data: Modelling and visualisation with R-INLA and shiny. CRC Press; 2019 Nov 21.

2.2 Study design and confounding factors

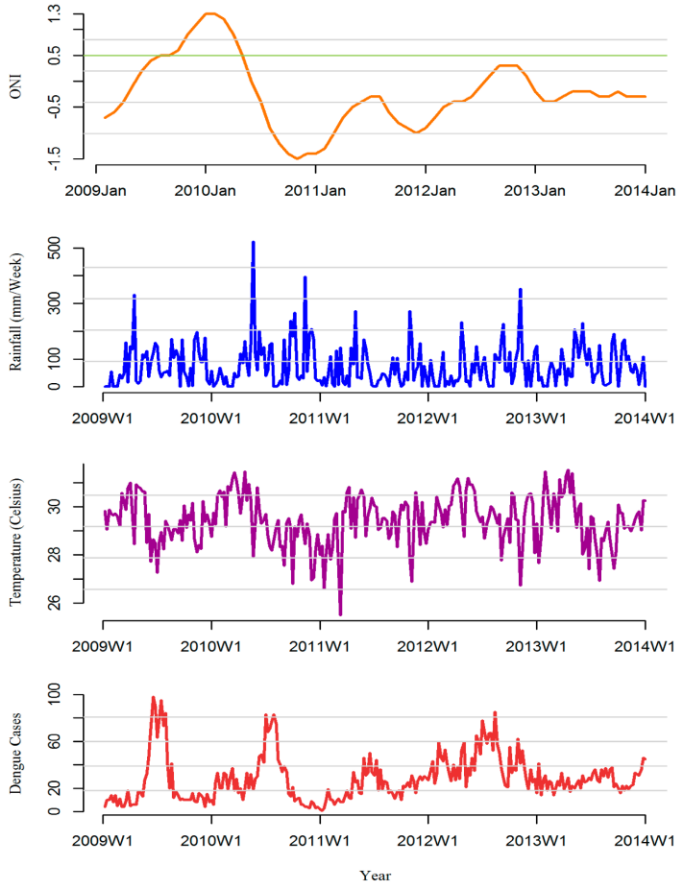
Dr Joacim Rocklöv

Umeå University

World Health Organization

Climate-epidemiological study designs

- Study designs that leverage information on the temporal or spatial variability and contrasts in hazard, exposure, vulnerability, and disease risk are often used
- These designs are often aggregated to a higher level than the individual level, i.e. a neighbourhood, a city, or a region
- Such designs include:
  - **Time series & interrupted time series designs:** Variables of study vary with time
  - **Spatial and geographical designs: Variables** of study vary between geographical areas and locations
  - **Spatio-temporal designs:** Variables of study vary over both time and space



Time series of monthly Oceanic Niño Index (ONI), weekly dengue cases, weekly cumulative rainfall and weekly mean temperature averaged across 10 MOH divisions in Kalutara district, Sri Lanka, 2009–2013.

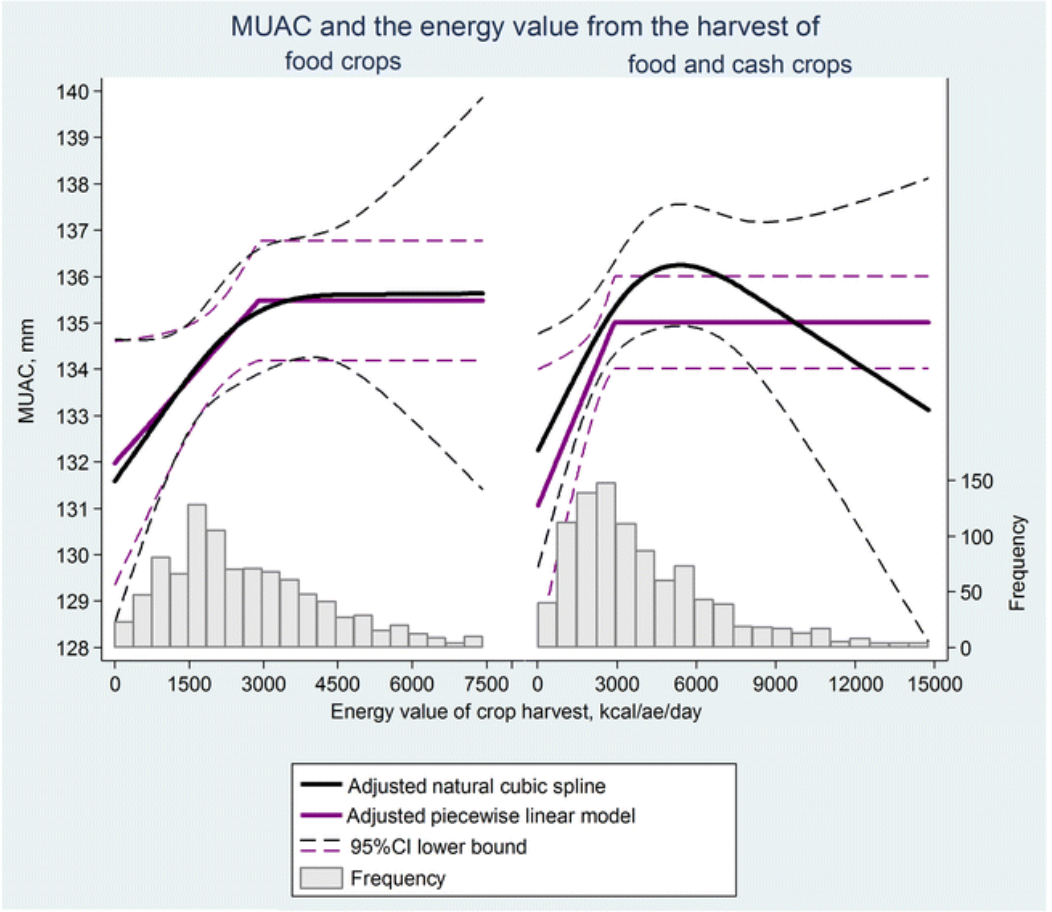
Liyange et al. 2016
https://www.mdpi.com/1660-4601/13/11/1087/htm

UMEÅ UNIVERSITY

World Health Organization

Standard epidemiological study designs

- Standard epidemiological study designs are therefore less frequently used in climate-epidemiology, but they are still applied
- For example:
  - **Survival analysis** based on cohort data is used to study long-term risks at the individual level
  - **Case-crossover designs** capturing spatio-temporal variability in exposure at the individual level
  - **Cross-sectional analysis** using regression methods based on surveys



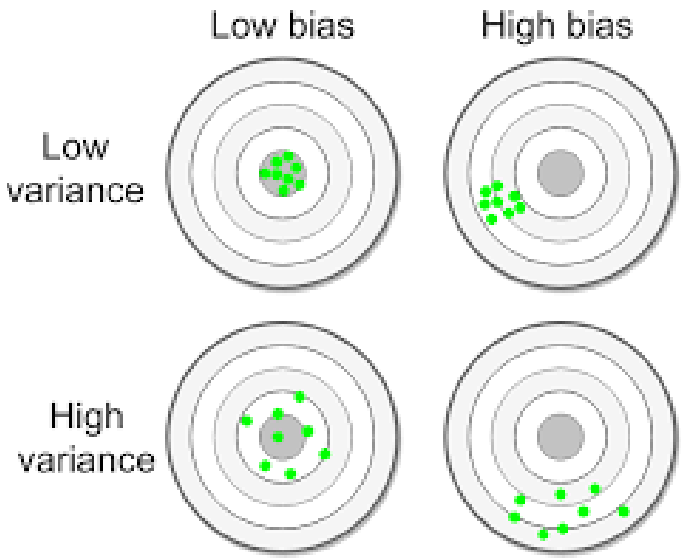
Restricted natural cubic spline and piecewise linear models of the associations of children's middle-upper arm circumference (MUAC) with food energy production. On the left: food energy estimates are based on food crop harvest alone. On the right: food energy estimates are based on food and cash crop harvest combined.

https://ehjournal.biomedcentral.com/articles/10.1186/s12940-017-0258-9

UMEÅ UNIVERSITY

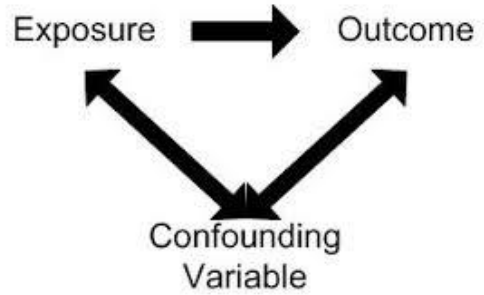
World Health Organization

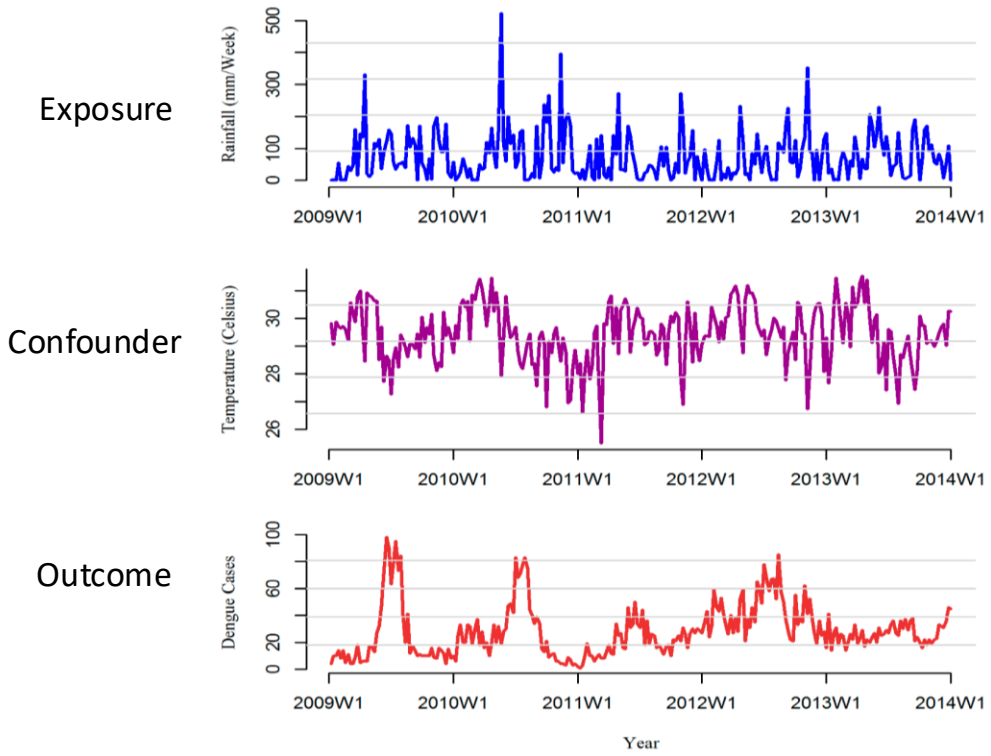
Bias



- Bias is a systematic source of error
- Study designs and analyses aim to minimise the risk of bias and disclose information about the underlying relationship between the climate and the disease
- Bias is different from variance

Confounding bias





- Confounding bias is falsely affecting the estimate of the association positively or negatively
- It is caused by a third factor – a confounder – which is associated with the climate variable of study (but does not lie in the causal pathway)
- When the confounder is introduced into a model at the same time as the climate variable, we say that we adjust for the confounding in the model. This should remove the bias.

Confounders to be adjusted for can be:

- seasonality, time trends, calendar patterns, interventions, behaviour/mobility in temporal designs
- socioeconomics, geographical vulnerability & exposure differences in spatial designs

UMEÅ UNIVERSITY

World Health Organization

Autocorrelations

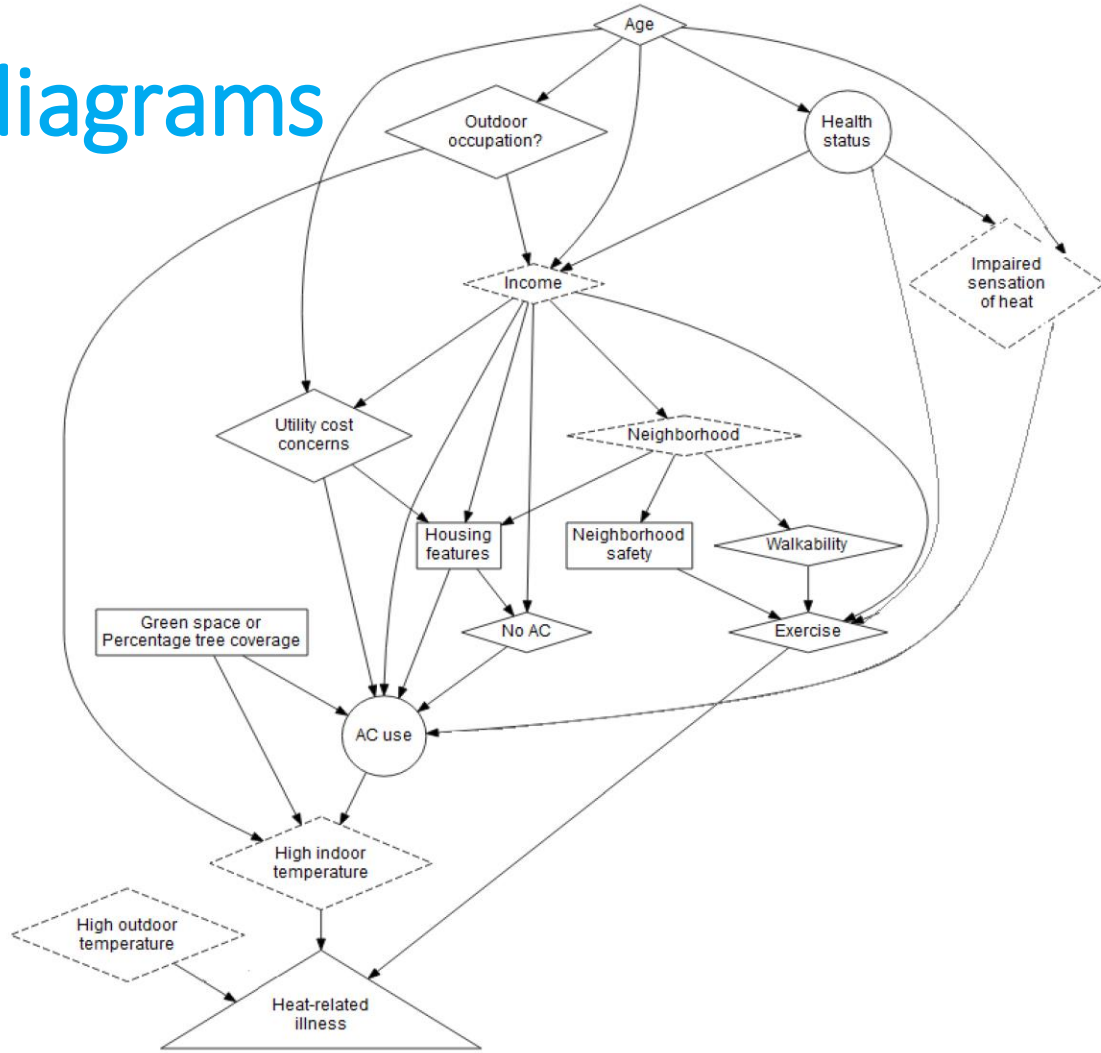
- Autocorrelation describes the relationship between a variable's current value and past value.

- Temporal or spatial autocorrelations observed in model residuals should be adjusted for.

- They could be caused by unadjusted confounding.
- They could, if significant and left unadjusted, cause bias in the estimation of standard errors and confidence intervals.

Causal pathway diagrams

- To understand and describe the role of a potential confounder, it can be helpful to create diagrams that outline its role in the pathway from climate to disease.
- These diagrams can be directed acyclic graphs (DAG´s).
- DAGs provide an understanding of the roles of variables and potential confounders within the disease system and how they interact.



Directed acyclic graph (DAG) of proposed associations between air conditioning (AC) use and heat exhaustion, and self-reported health status and heat exhaustion.

https://www.mdpi.com/1660-4601/17/16/5704/htm

UMEÅ UNIVERSITY

World Health Organization

Adjusting for confounding

- Can be made **analytically** or by **design of the study**
- When adjusting for a confounder, it is important to describe the effect of the confounder in the best possible way, including lagged effects and non-linearities
  - There is an inherent delay or **lag**, usually variable and unpredictable, between any environmental stressor and the human response.
  - **Nonlinearity** is a term used in statistics to describe a situation where there is not a straight-line or direct relationship between an independent variable and a dependent variable. In a nonlinear relationship, changes in the output do not change in direct proportion to changes in any of the inputs.
- Analytical adjustment is often done by including the confounder as an independent variable in a multiple regression model
  - **Multiple regression** is an extension of linear regression models that allows predictions of systems with multiple independent variables. Multiple regression is specifically designed to model relationships between a single dependent variable and multiple independent variables.

Outcome ~ Exposure + Confounder

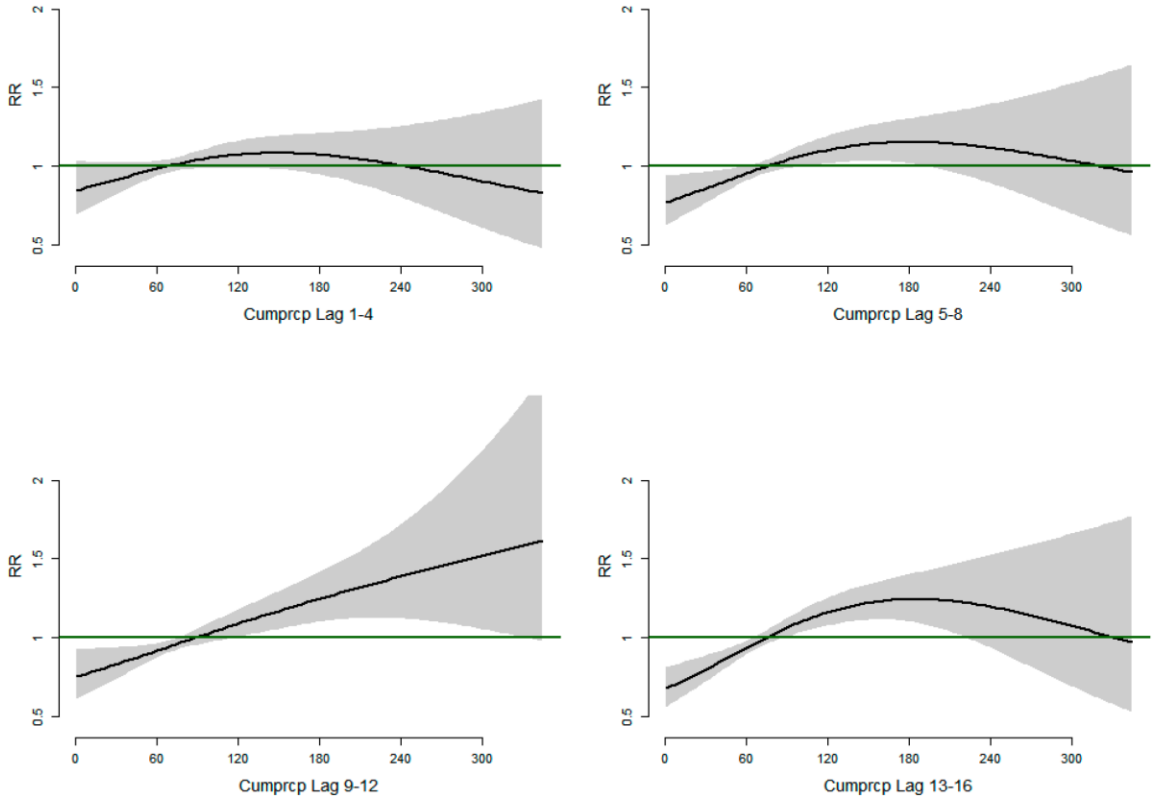
A Confounder is a variable whose presence affects the variables being studied so that the results do not reflect the actual relationship.

There are various ways to control or exclude confounding variables, including Randomisation, restriction, and Matching. But all these methods are applicable during the **study design**. When experimental designs are premature, impractical, or impossible, researchers must rely on **statistical methods** to adjust for potentially confounding effects. These Statistical models (especially regression models) are flexible to eliminate the effects of confounders.

Inference

- Association or potential cause-and-effect relationship
- Confidence intervals are important, says much more than p-values
- Estimates of association and confidence intervals are correct if we assume no bias
- Estimates of association can be obtained through
  - Relative Risk (RR),
  - Odds Ratio (OR),
  - Hazard Ration (HR),
  - coefficient (linear),
  - cumulative lag effects (net total effect over several lags),
  - etc.
- Sometimes hard to summarise as one figure, but better illustrated in graphs

Inference



https://www.mdpi.com/1660-4601/12/2/1983/htm

Relative risks of Malaria/Anemia mortality in children under five years with weekly mean temperature (Tmp) at different lag strata.

The figure shows the relative risks of malaria/anaemia mortality with different weekly lags of mean temperature.

At lags of 1–4 weeks, we observe a U-shaped relationship between temperature and mortality, with mortality declining at low temperatures but increasing above 24 °C; however, this is not statistically significant, as all confidence intervals for the Relative risks include 1.

In the last 5–8 weeks, we have seen an inverted U-shaped relationship between temperature and mortality. Mortality increases until 24 °C and then steadily declines.

At higher lags, 9–12 weeks and 13–16 weeks, the temperature mortality relationship exhibits a J-shape, with the relative risk of mortality increasing linearly with increasing temperatures from 24 °C, with greater risk observed in lag strata 13–16 weeks.

At temperatures below 24 °C, the risk of mortality increases at higher lags, starting at 9 weeks. The malaria-temperature lag relationships show a clearer pattern with longer lag times, with the strongest relationship at lag 13–16.

UMEÅ UNIVERSITY

World Health Organization

BIBLIOGRAPHIC REFERENCES

- Rothman KJ, Greenland S, Lash TL, editors. Modern epidemiology. Lippincott Williams & Wilkins; 2008.
- Peng RD, Dominici F. Statistical methods for environmental epidemiology with R. R: a case study in air pollution and health. 2008.
- Moraga P. Geospatial health data: Modelling and visualisation with R-INLA and shiny. CRC Press; 2019 Nov 21.

UMEÅ UNIVERSITY

World Health Organization

Section 2.3:

Data on health outcomes and disease risk: strengths, weaknesses and best practices for their use

Learning objective: To understand the types of data that can be used as measures or proxies for health outcomes and/or disease risk/hazard, and understand issues of spatio-temporal resolution of the data, specificity and sensitivity, representativeness and ethical considerations.

Case studies:

West Nile virus and Lyme disease surveillance in Canada: human case, sentinel animal and entomological surveillance.

Further reading:

- Surveillance systems: https://www.cdc.gov/training-publichealth101/php/training/introduction-to-public-health-surveillance.html
- Attributes of surveillance systems and data: https://www.cdc.gov/mmwr/preview/mmwrhtml/00001769.htm

World Health Organization

2.3 Data on health outcomes and disease risk: strengths, weaknesses and best practices for their use

Dr Nick Ogden

Public Health Agency of Canada

World Health Organization

Learning Objectives

Understand the types of data that can be used as measures or proxies for health outcomes and/or disease risk/hazard:

- Human case surveillance and hospital records data;
- Entomological surveillance data;
- Data from other environmental sources;
- Surveillance data from sentinel animals

and understand issues of spatio-temporal resolution of the data, specificity and sensitivity, representativeness and ethical considerations.

© WHO/Zakarya Safari

Methodology

- Example infectious diseases will be used to explore key aspects of health or health hazard data that need to be considered when trying to relate these data to weather or climate:
  - specificity and sensitivity;
  - spatio-temporal resolution of the data;
  - representativeness;
  - ethical considerations

  Will be elaborated on the following slides
- The main examples we will use are two temperate zone vector-borne zoonoses* that co-exist in North America:
  - Tick-transmitted Lyme disease; and
  - Mosquito-transmitted West Nile virus infection
- For each of these diseases, which are weather/climate sensitive, there is established surveillance for human cases, entomological surveillance, and surveillance in sentinel animals

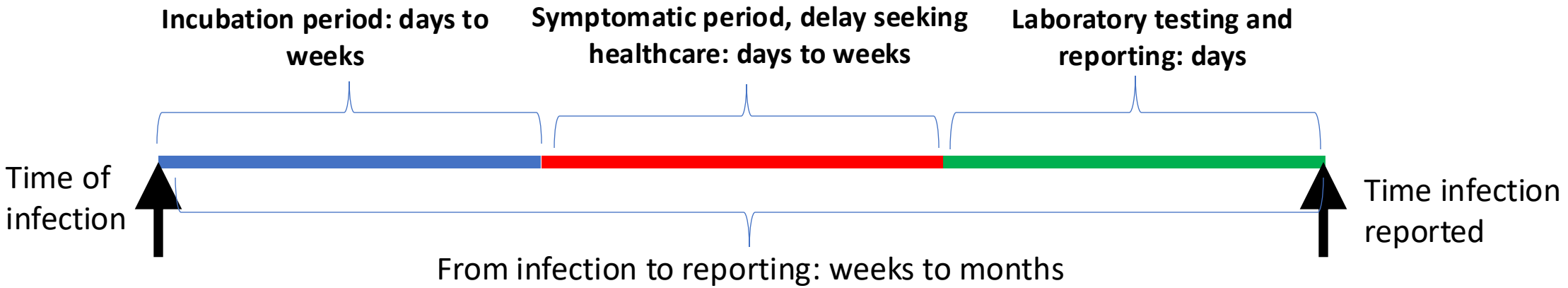
Specificity and Sensitivity

- Sensitivity is the proportion of truly infected people/animals/samples that are correctly identified
- Specificity is the proportion of uninfected people/animals/samples that are correctly identified as uninfected
- Specificity and sensitivity may refer to:
  - laboratory tests
  - clinical diagnosis on the basis of clinical manifestations (but diagnosis may be made on a combination of clinical manifestations and laboratory test results)
  - surveillance case definitions, which could be simply a positive laboratory test result, or a clinical diagnosis, but may include a combination of test results, clinical manifestations, and other epidemiological information on the likelihood of exposure
- In surveillance programs involving human and animal cases, the less specific the laboratory test, the more additional information on clinical manifestations and exposure is needed to enhance the specificity of the case definition
- Specificity and sensitivity of case definitions will determine the accuracy of any weather/climate/infection associations obtained from surveillance data

Reference: https://academic.oup.com/bjaed/article/8/6/221/406440

Spatio-temporal resolution

- To develop and validate climate/weather-driven decision tools, we need to identify associations between disease occurrence and climate/weather data (as well as other relevant environmental, demographic and socioeconomic data)
- To do this, we need to know where and when an infected person or sentinel animal acquired infection and/or an infected vector/environmental sample was found, and link time and place with data on climate/weather at this time and place (using geographic information systems and/or statistical software)
- However, several issues impact how precise we can be as to when and where detected infections were actually acquired:
  - Some wildlife and vectors may disperse long distances from where they acquired infection (examples in later slides)
  - For infected people and domesticated animals, the time lag between infection and diagnosis results in an increased possibility of recall bias – i.e. the people or domesticated animal owners cannot remember where they were when they acquired the infection:



Representativeness

- Representativeness means that the data on the occurrence of infection are actually representative of the population from which they are obtained
- Representativeness is important because, to understand the presence and absence of infection in a population, we can only realistically obtain data from studies on a subset of that population
- The study population may not be representative if (for example):
  - Sampling is biased towards some members of the population and/or against others because of weaknesses in study/surveillance system design
  - Sampling focuses on only one group of the population, which (for example, due to genetic differences) is somewhat different from other members of the population
  - Sampling occurs in only a narrow time window that does not capture all possible occurrences

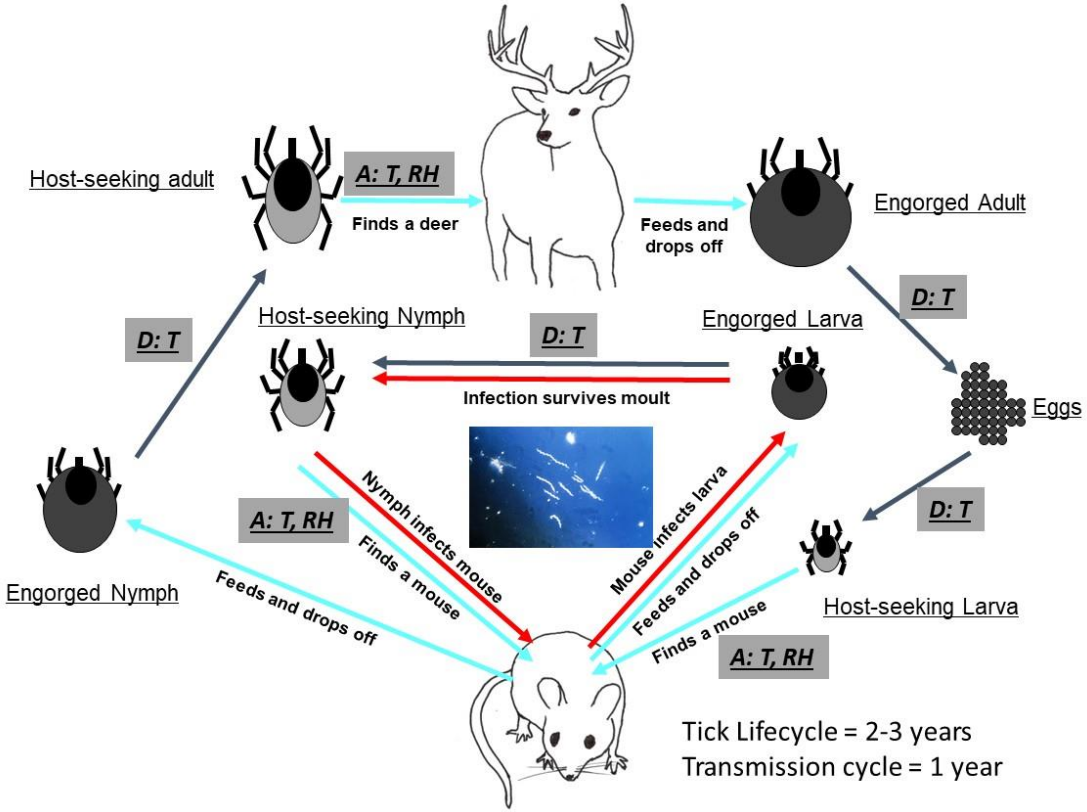
Reference: CDC MMWR, surveillance attributes - https://www.cdc.gov/mmwr/preview/mmwrhtml/00001769.htm

Ethics

- It is important to recognise and respect two key aspects of health data that determine who has access, what the data can be used for, and the extent to which the data can be shared by a user:
  - Privacy regarding the person affected by the disease (or the owner of an animal, particularly of livestock species)
  - Ownership of the data
- The extent to which privacy and data ownership issues limit the use of data, and the degree of information provided, will vary greatly amongst diseases and countries.
- In general, privacy issues centre around the degree to which information provided may be used to identify the person affected (or the farm/production enterprise in the case of infected livestock)
- This may limit the:
  - Spatio-temporal precision with which the time and location that the person (or a domesticated animal) acquired infection are reported to, and by, data users; and
  - Metadata on the person/production enterprise (e.g. age, sex, occupation, exposures)
- Issues of ownership of the data mean that appropriate permissions need to be obtained before the data are used

Reference: https://lsspjournal.biomedcentral.com/articles/10.1186/s40504-018-0078-x

Lyme disease



Ogden, N. H., & Lindsay, L. R. (2016). Effects of climate and climate change on vectors and vector-borne diseases: ticks are different. Trends in parasitology, 32(8), 646-656.

- Lyme disease has initially mild symptoms, including a skin rash, then progresses to more serious neurological, cardiac and articular disease if untreated
- The causal bacteria are spirochaetes of the *Borrelia burgdorferi* sensu lato group, whose natural reservoirs are mostly wild rodents and birds. Deer are important hosts for the ticks, but don't take part in the transmission cycle
- The bacteria are transmitted by hard-bodied ticks of the genus *Ixodes*, which live in woodlands and have long (2-3 year) life cycles involving three feeding stages that are all parasitic on animals
- The cycle of transmission involves i) infected nymphal ticks infecting the wild animal hosts they feed on; ii) larval ticks acquiring infection from infected hosts; and iii) survival of the bacterium as fed larvae moult into nymphs
- Humans don't take part in the natural transmission cycle and acquire infection when they acquire tick bites in the woodlands where the ticks occur

Points at which weather and climate (and potentially climate change) may impact the tick lifecycle and *B. burgdorferi* transmission cycle are indicated by the grey-filled boxes in which A = effects on tick activity, D = effects on tick inter-stadial development rates, T = effects of temperature and RH = effects of humidity. All images from the US Centres for Disease Control and Prevention

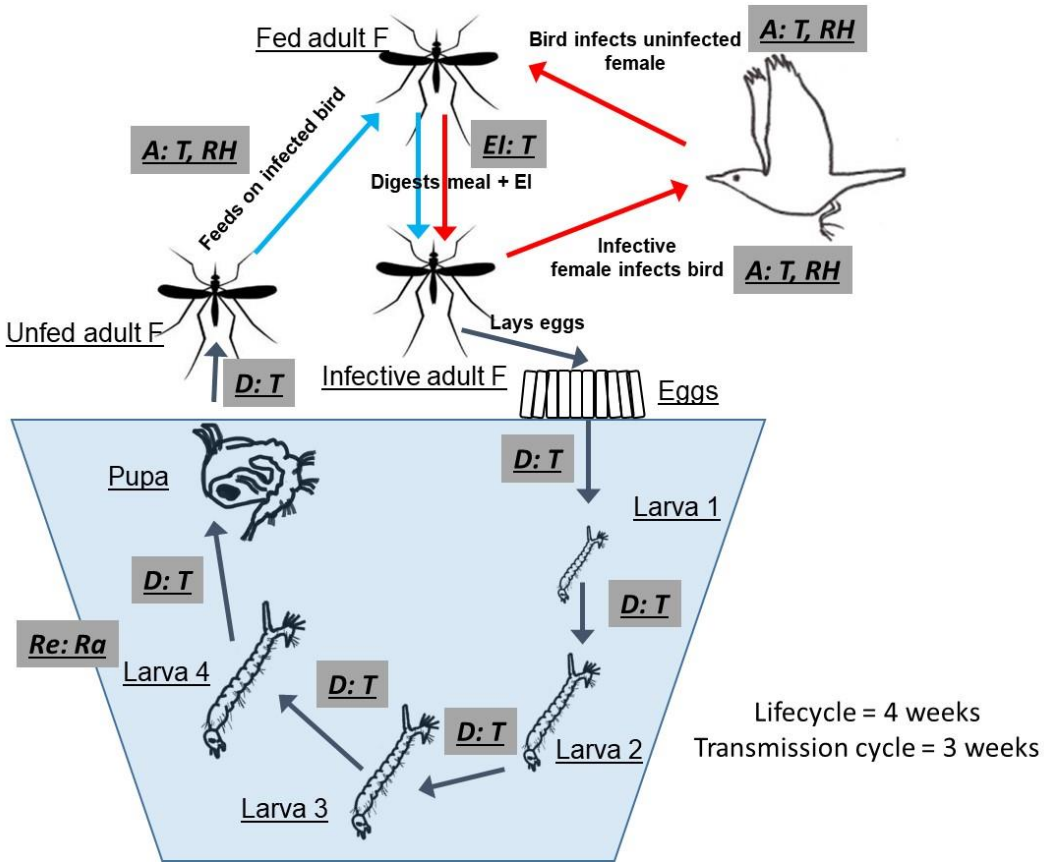
Agence de la santé publique du Canada

World Health Organization

West Nile virus infection

- West Nile virus (WNV), which is caused by a flavivirus, is mild in most cases, but in some cases, infection results in severe neurological disease that may result in death
- In North America, WNV is transmitted mostly by *Culex* species mosquitoes, which occur in a wide range of habitats, including urban and suburban areas
- The mosquitoes spend most of their life as free-living insects in fresh water. Only adult females take blood meals from their favoured hosts (birds) to provide protein for egg production
- WNV is transmitted amongst wild bird reservoir hosts by the adult female mosquitoes
- As with Lyme disease, humans can acquire infection but don't take part in transmission cycles

Ogden, N. H., & Lindsay, L. R. (2016). Effects of climate and climate change on vectors and vector-borne diseases: ticks are different. Trends in parasitology, 32(8), 646-656.



Points at which weather and climate (and potentially climate change) may impact the tick lifecycle and WNV transmission cycle are indicated by the gray-filled boxes in which A = effects on mosquito activity, D = effects on mosquito inter-stadial development rates, EI = effects on the extrinsic incubation incubation period of WNV in the mosquito, Re = effects on mosquito reproduction, T = effects of temperature, RH = effects of humidity, and Ra = effects of rainfall.

Weather-/climate-driven decision tools for Lyme and WNV

- Lyme disease and WNV are both weather- and climate-dependent vector-borne zoonoses, but there are differences between them
- *Borrelia burgdorferi* is maintained in multi-year transmission cycles by ticks with multi-year lifecycles, and expected weather and climate change effects are on:
  - The duration of the season of tick activity and risk to people
  - Tick population survival and thus the geographic range of ticks and *B. burgdorferi*
- WNV is maintained in weeks-long transmission cycles by mosquitoes that have lifecycles that can be completed in weeks, so expected weather and climate change effects are on:
  - Outbreaks of WNV driven by rapid, weather-responsive changes in mosquito lifecycles and WNV transmission
  - Mosquito and WNV population survival and thus geographic range changes
- Decision tools include:
  - Assessment of geographic range changes for both vector-borne diseases
  - Forecasting of WNV outbreaks where the disease is endemic
  - Geographic occurrence is therefore important for both, but temporal occurrence is more important for WNV

Reference: https://www.sciencedirect.com/science/article/pii/S1471492216300320

Lyme disease: Human case data

Specificity and sensitivity:

- Laboratory diagnosis of Lyme disease is mostly by serological methods, which have some limitations regarding specificity, so diagnosis requires information on both clinical manifestations and exposure[1]
- As diagnosis in humans requires information on exposure, human case surveillance alone has limited sensitivity to detect low-risk environments, including new emerging areas of risk

Spatio-temporal resolution:

- It is common that those who acquire Lyme disease do not recognise the infective tick bite as the bite is painless, and nymphal ticks, which transmit most cases, are small and difficult to find
- Lyme disease develops over a period that may extend for months after the bite of an infected tick, so diagnosis may not occur until weeks or months later
- Thus, there is almost always uncertainty as to the time and place of infection, and recall bias regarding time and place of infection is common.
- Together, these limit the spatio-temporal resolution human case data can provide on environmental risk[2]
- This spatiotemporal imprecision must be accounted for when developing predictive models by choosing appropriate scales for the Lyme disease cases data, and thus for explanatory variables (weather, climate and other environmental data)

Representativeness

- Apart from differences in genospecies of *Borrelia burgdorferi*, that result in well-recognised inter-country differences in manifestations of Lyme disease, there is little in human case surveillance for Lyme disease that challenges representativeness

Ethics

- Human case data are the most challenging regarding issues of privacy, and that applies to human Lyme disease cases

1 https://www.nature.com/articles/nrdp201690
2 https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-7219-x

Lyme disease: sentinel animal data

The slide focuses on wild animal hosts and domestic dogs as sentinels. Wild animal hosts (mostly rodents and birds) acquire infection but do not (or rarely) suffer disease. Dogs acquire infection and, in some cases, develop illness

Specificity and sensitivity:

- Detection of infection in wild rodents is frequently by molecular methods to detect *B. burgdorferi* DNA in their blood or tissues, which generally has high specificity but moderate sensitivity; however, the sensitivity of detection of *B. burgdorferi* at a site is enhanced by the usual practice of capturing multiple rodents at the same time, and by the generally persistent infection in wild rodents.
- Detection of infection in dogs is generally by the C6 serological test, which has high sensitivity and specificity

Spatio-temporal resolution:

- Detection of infection in wild rodents has high spatio-temporal resolution given their generally small home range size (<1 km). Combined with the high specificity of positive test results (when using well-performing tests), this means that testing wild rodents is the gold-standard method for detecting Lyme disease risk.
- Positive serological test results in dogs may mean the dog acquired the infection recently or a year or more ago. Therefore, as for test results in humans, there is almost always uncertainty as to the time and place of infection, which needs to be recognised in choosing appropriate scales for the serological data, and the explanatory variables (weather, climate and other environmental data)

Representativeness:

- Detection of infection in animals does not have significant issues that challenge representativeness

Ethics:

- Capture of wild animals frequently requires licensing that considers impacts on animal welfare and wildlife conservation

Lyme disease: entomological data

There are two types of entomological data: i) those collected by passive surveillance (ticks submitted by the public or ticks found on humans and animals by participating medical and veterinary clinics); and ii) those collected by active field surveillance, which involves the collection of ticks from captured wild animals and/or the collection of host-seeking ticks from the herbage by drag sampling. Ticks are then usually tested by molecular methods to detect *B. burgdorferi* DNA[1]

Specificity and sensitivity:

- Both identification of tick vectors by appropriate keys and detection of *B. burgdorferi* DNA by molecular methods have high specificity and sensitivity
- As methods to detect the presence of tick vectors and *B. burgdorferi*, passive surveillance (particularly using dogs) and capture of wild animals have high sensitivity, with drag sampling having slightly less sensitivity[2]

Spatio-temporal resolution:

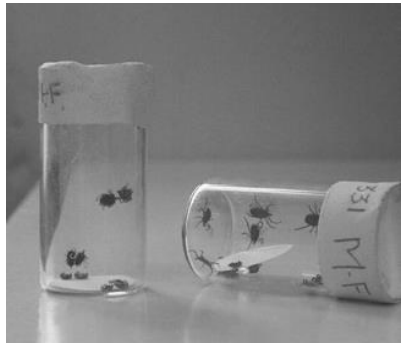
- Detection of ticks and *B. burgdorferi* by active surveillance has high spatiotemporal resolution and is the gold standard
- The detection of ticks in passive surveillance does not mean that the location they are found is the location of the reproducing tick population they came from. This is because ticks can be dispersed hundreds of miles by migratory birds. Methods have been developed to use relative numbers of ticks collected per human population to better identify ticks that are more likely to come from local reproducing populations[3]

Representativeness:

- Ticks have relatively fixed seasonal activity periods each year, although these vary geographically, by tick stage and by tick species, so surveillance has to be conducted at the right time of year to be representative
- As the seasons are predictable from one year to the next, a few well-timed site visits are needed to obtain data
- Passive surveillance for ticks is only possible where there are human populations capable of participating in the surveillance, and many ticks detected may have been dispersed from far away by migratory birds, and these issues have to be accounted for in developing associations between tick presence and the explanatory variables (weather, climate and other environmental data)

Ethics:

- Capture of wild animals frequently requires licensing that considers impacts on animal welfare and wildlife conservation



1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466818/
2. https://www.canada.ca/content/dam/phac-aspc/migration/phac-aspc/publicat/ccdr-rmtc/14vol40/dr-rm40-05/assets/pdf/ccdrv40i05a01-eng.pdf
3. https://academic.oup.com/jme/article/49/2/400/907346

WNV: Human case data

Specificity and sensitivity:

- Laboratory diagnosis of WNV is mostly by serological methods, and cross-reactivity with other flaviviruses is possible. However, where infection with more than one flavivirus is possible, confirmation using WNV-specific tests is standard practice. Most case definitions require appropriate clinical manifestations, exposure history and laboratory test results for confirmation of a case[1,2]
- This combination optimises sensitivity and specificity of test results, but the low percentage of infections that are clinical means that human case surveillance may have low sensitivity to detect WNV risk

Spatio-temporal resolution:

- The time from infection to development of symptoms, serological and clinical diagnosis, and then reporting to public health may take a month[3]
- Therefore, there is almost always uncertainty as to the time and place of infection, but this is not as much of a problem as for Lyme disease.
- Nevertheless, a degree of spatiotemporal imprecision must be considered and accounted for when developing predictive models by choosing appropriate scales for the WNV case data, and thus for explanatory variables (weather, climate and other environmental data)

Representativeness:

- Apart from the possibility of false seropositivity due to other flaviviruses, which can be managed by confirmatory serological tests, there is little in human case surveillance WNV that challenges representativeness

Ethics:

- Human case data are the most challenging regarding issues of privacy, and that applies to human WNV cases

© WHO/Yoshi Shimizu

1. National case definition: West Nile virus - Canada.ca
2. Arboviral Diseases, Neuroinvasive and Non-neuroinvasive 2015 Case Definition | CDC
3. Ogden et al. 2019-Weather-based forecasting of mosquito-borne disease outbreaks in Canada, Can Communc Dis Repo, 45(5): 127-32.

WNV: Sentinel animal data

The slide focuses on crows and other corvids, as well as horses as sentinels. Corvids, as well as unvaccinated horses, are highly sensitive to infection, suffer severe manifestations and often die rapidly.*[1,2] *Other sentinel animals, such as chickens, have been explored and used in some jurisdictions

Specificity and sensitivity:

- Detection of infection in dead corvids and horses is often post-mortem with combinations of molecular, antigen detection and serological tests that have high specificity and sensitivity.
- As corvids and unvaccinated horses are very sensitive to infection, the presence of dead corvids and severely ill horses is a potentially sensitive method of identifying times and places where levels of WNV transmission are increasing[3]

Spatio-temporal resolution:

- The spatio-temporal resolution of the sentinel animals is high, because the onset of severe manifestations and death occurs very quickly (days) after infection. So the occurrence of dead corvids or severely ill horses is usually a precise indicator of when and where WNV risk is emerging[4]

Representativeness:

- Detection of WNV transmission by sentinel animals requires the presence of unvaccinated horses and/or members of the public willing to participate in surveillance to report the presence of dead birds. This means there is no universal and equal geographic coverage for surveillance using sentinel animals, which needs to be adjusted for in developing associations between WNV occurrence and climate, weather and other environmental variables[3]

Ethics:

- There is little impact of ethical consideration regarding surveillance using dead corvids; there may be privacy issues regarding data from equines

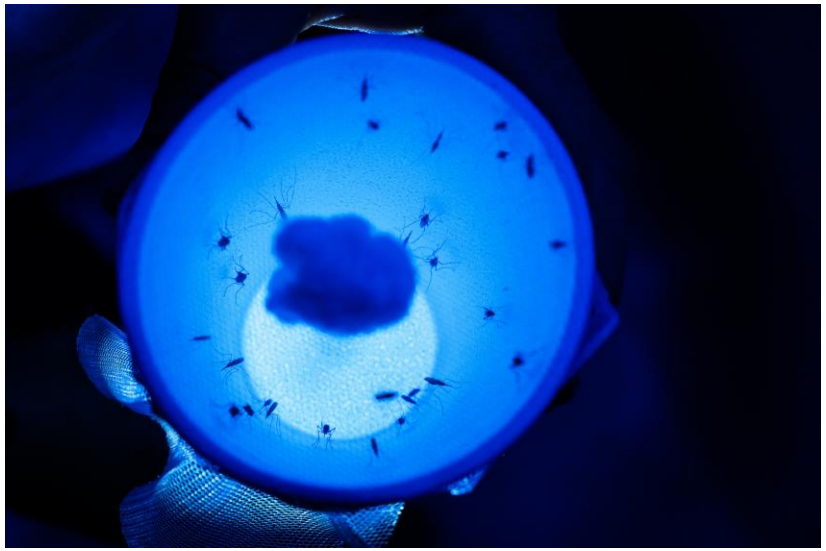
1. https://onlinelibrary.wiley.com/doi/full/10.1111/tbed.13452
2. https://nyaspubs.onlinelibrary.wiley.com/doi/full/10.1111/j.1749-6632.2001.tb02687.x
3. https://www.frontiersin.org/articles/10.3389/fvets.2019.00483/full
4. https://wwwnc.cdc.gov/eid/article/9/3/02-0628_article

WNV: Entomological data

Surveillance for emerging/re-emerging WNV risk through mosquito capture and WNV testing by molecular methods is widespread in North America and many parts of Europe.[1,2] As with other entomological surveillance programs for mosquito-borne diseases, the objective is to detect the presence of (or increasing risk from) the mosquito-borne disease by estimating both the abundance of mosquito vectors and the prevalence of infection in mosquitoes.

Specificity and sensitivity:

- Detection and identification of mosquito vectors, and detection of WNV by molecular methods, have high sensitivity and specificity, provided appropriate methods are used
- Sensitivity of detection of WNV transmission depends, however, on the design of the surveillance program: i) there has to be a sufficient number of traps per unit area to catch a large enough, and geographically representative (i.e. captures are made in most habitats where vector mosquitoes occur[3]), sample of mosquitoes; and ii) the duration of the program of capture during the year has to be long enough to capture seasonally variable changes in mosquito abundance and WNV prevalence



© WHO/Yoshi Shimizu

Spatio-temporal resolution:

- The spatio-temporal resolution of the entomological surveillance is high, because adults of most vector mosquito species have limited capacity for dispersion from the site where they emerged from the pupal stage

Representativeness:

- Provided that the design of the surveillance program meets the criteria for sensitivity described above, mosquito surveillance for WNV may be representative at the geographic scales equivalent to US states. However, at larger geographic scales, the ecology of WNV transmission may vary due to differences in mosquito vector species[4] or to more subtle ecological differences[3,5]. Therefore, there are limits to the extent to which WNV risk identified by mosquito surveillance in one region is representative of the risk in another. Precisely what the limits of representativeness are is the subject of ongoing research.

Ethics:

- There is little impact of ethical consideration regarding mosquito surveillance for WNV

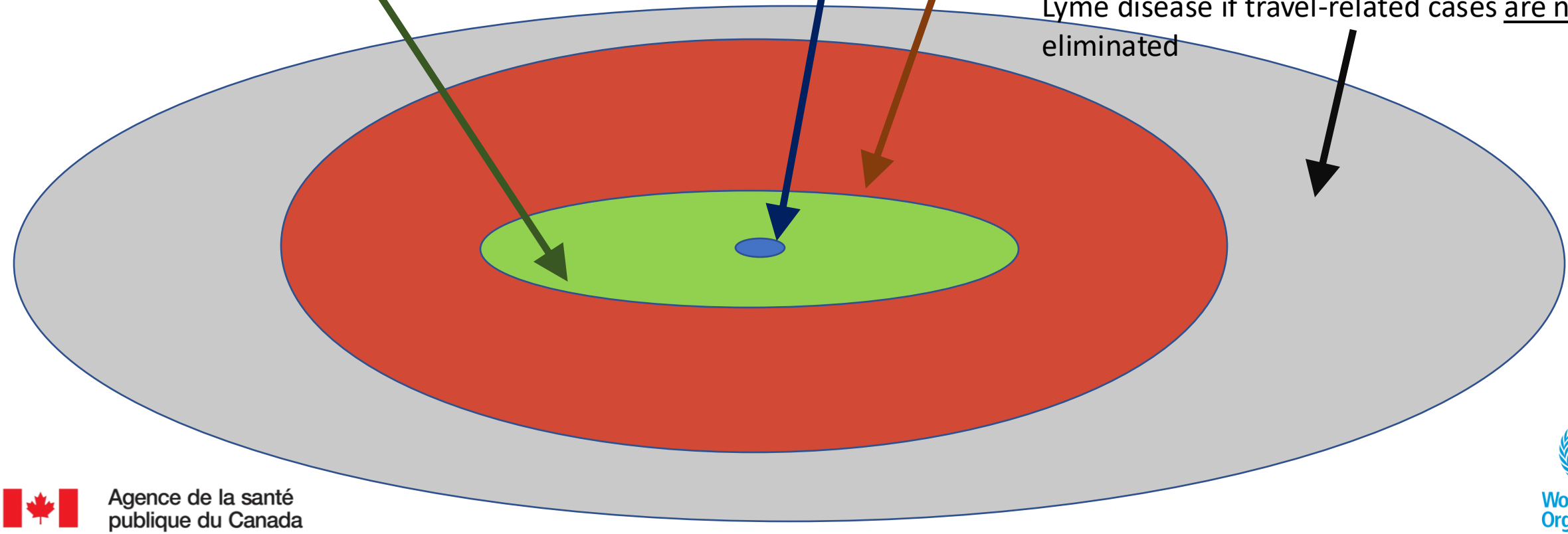
Spatial resolution of surveillance data

Circa 10km (extent of a municipality) – Human case and sentinel animal surveillance for WNV; human case surveillance for Lyme disease if travel-related cases <u>are</u> eliminated; passive tick surveillance <u>with</u> accounting for tick dispersion by migratory birds

< 1km - Mosquito surveillance for WNV
Active field surveillance to Lyme disease risk

Ca 500km – Passive tick surveillance for Lyme <u>without</u> accounting for tick dispersion by migratory birds

1000s of km – human case surveillance for Lyme disease if travel-related cases <u>are not</u> eliminated



Agence de la santé publique du Canada

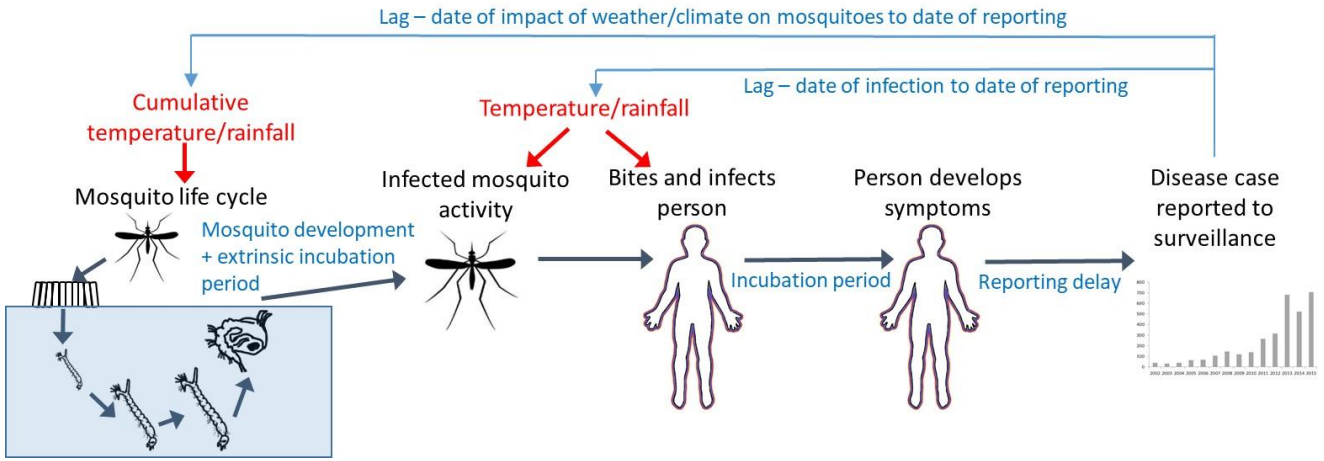
World Health Organization

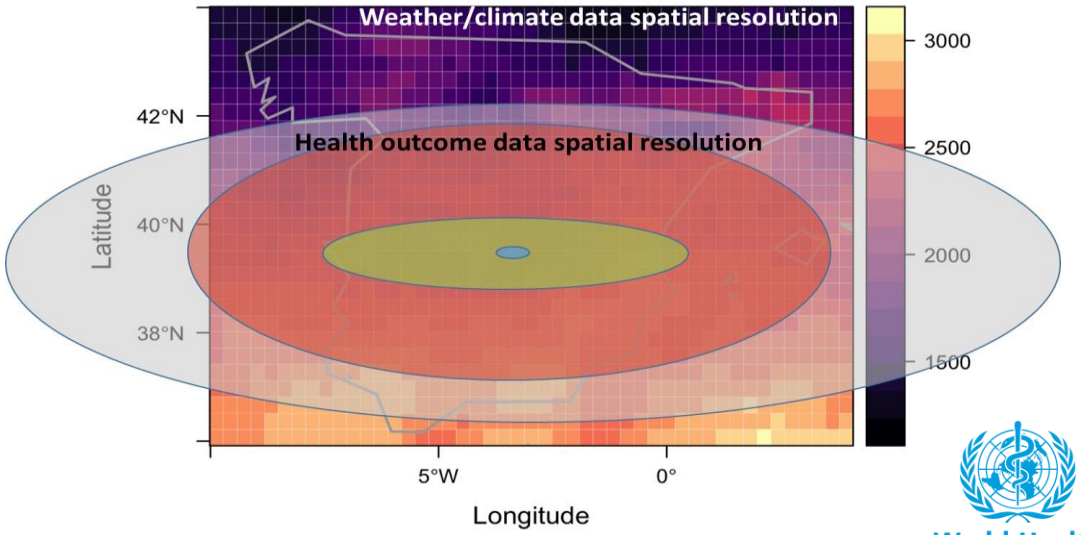
Summary of qualities of surveillance data

Surveillance data type	Specificity	Sensitivity	Spatio-temporal resolution	Representativeness	Ethical issues
Lyme – human case	Moderate	Low	Low to moderate	High	Yes
Lyme – sentinel animals	High	High	High (wildlife) Low to moderate (dogs)	High	Slight
Lyme – entomological surveillance	High	High	High (active) Low to moderate (passive)	High (active) Moderate (passive)	Slight
WNV – human case	High	Low	Moderate to high	High	Yes
WNV – sentinel animals	High	High	High	Moderate to high	Slight
WNV – entomological surveillance	High	High	High	Low	Slight

LINKING HEALTH DATA AND WEATHER/CLIMATE DATA

- First, explore health outcome and climate data for bias and, if it exists, account for it (section 2.2)
- Second, map out how, biologically, weather/climate could impact the detected health outcome, consider:
  - Temporal aspects - lags between infection and reporting of cases, lagged effects of weather/climate on hazard in the environment, whether effects of weather on hazard are instant or due to cumulative effects over time
  - Spatial aspects – what is the spatial resolution of the health outcome data, and the weather/climate data you want to link to it?
- Third, select the appropriate weather/climate data to link to the health outcome datum for analysis
- Fourth, decide what is an appropriate spatial resolution for the weather/climate data, considering the spatial resolution of the health outcome data – do you use a point value, an average over 30m^2, 30 km^2, 100km^2?





Agence de la santé publique du Canada

World Health Organization

Section 2.4:

Visualizing spatial and temporal variation in climate and health data

Learning objective: To gain a basic awareness of spatial data and the common methods used in visualising spatial and temporal variation in climate and health data; additionally, appreciate why data visualisation is important for data analysis.

Case Study:

Colón-González et al. 2013 - The Effects of Weather and Climate Change on Dengue, PloS Negl Trop Dis
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3828158/

World Health Organization

2.4 Visualising spatial and temporal variation in climate and health data

Dr Felipe J. Colón-González

World Health Organization

Aim

- To introduce some of the most common approaches and tools used for the visualisation of spatial and spatiotemporal data in climate and health contexts.

Intended learning outcomes

By the end of the module, you should be able to:

- Recognise the importance of data visualisation for science communication.
- Select appropriate methods for data visualisation depending on the type of data.
- Critique the adequacy of different visualisations.
- Generate effective visualisations for their data.

World Health Organization

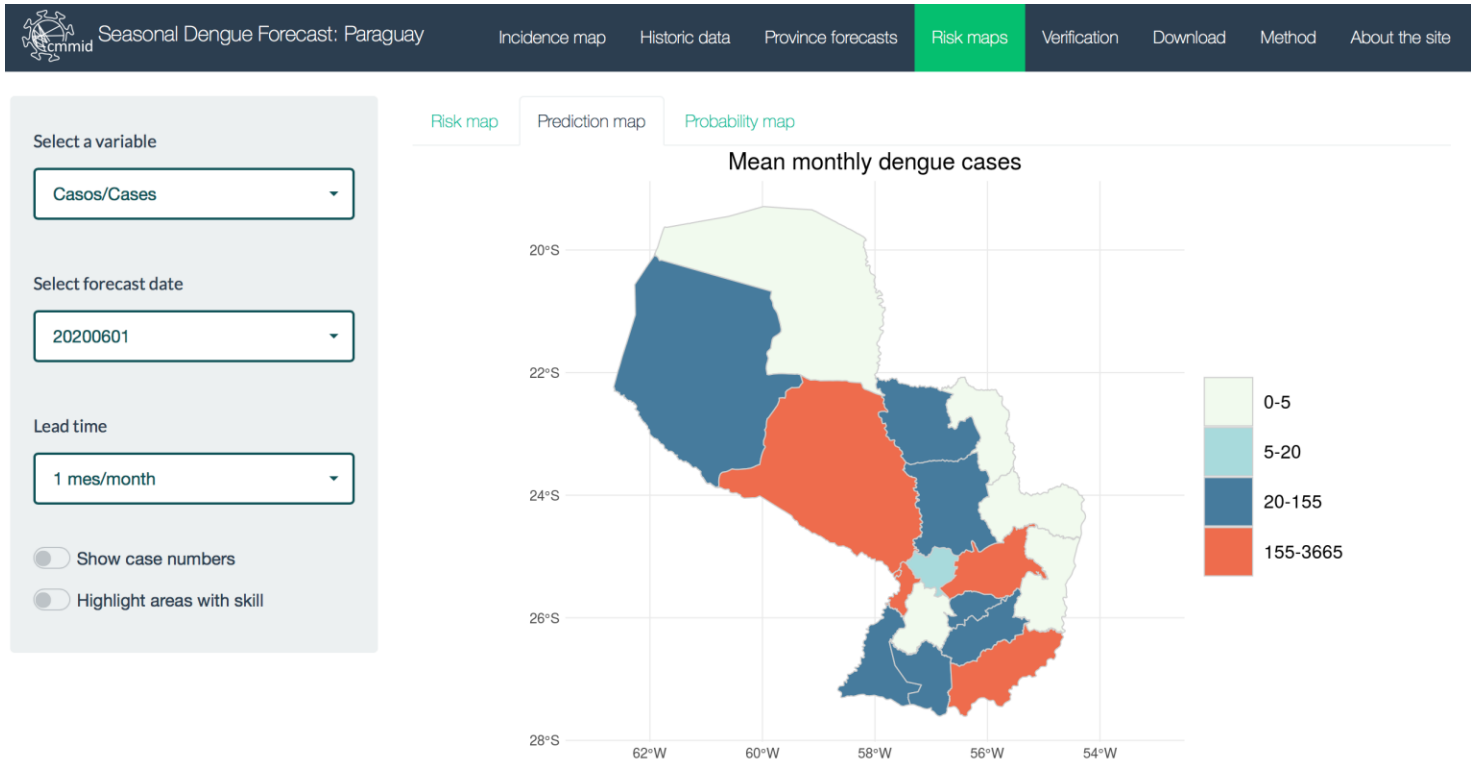
Outline

- Why data visualisation matters
- Fundamental properties of spatial and spatiotemporal data
- Spatial data visualisation
  - Types of spatial data
  - Commonly used spatial graphs
  - Colour scales
  - Example: Lip cancer in Scotland
- Visualising variation across time and space
  - Example: Dengue fever in Mexico

World Health Organization

Why data visualisation matters

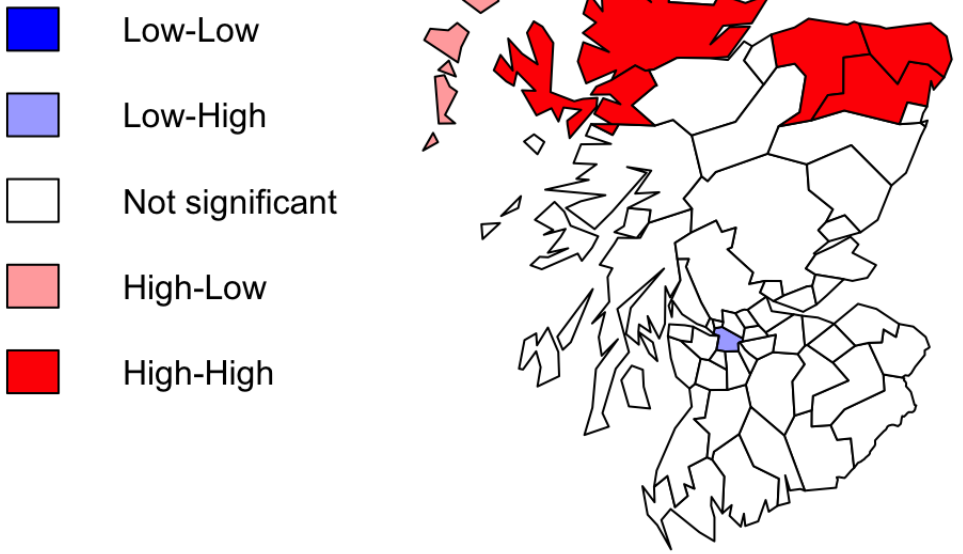
- Often, we need to convey the information contained in thousands of rows of data.
- How to make sense of such information efficiently?
- **Data visualisation** offers a great solution.



World Health Organization

Why data visualisation matters

- Data visualisation enables us to:
  - **Identify patterns** or areas that need more attention (e.g. hotspots, clusters of disease).
  - **Analyse varying risks** of infection across time and space.
  - **Characterise and make comparisons** between groups, times, or regions.
    - For example, we may be interested in exploring the differences in risk between urban and rural populations, northern and southern provinces, or winter versus summer periods.



World Health Organization

Spatial reference

- Spatial and spatiotemporal are directly or indirectly **referenced to a location** on the surface of the Earth.
- This reference may be in the form of geopolitical boundaries, or coordinate values plus a coordinate reference system (CRS).
- CRS are a mathematical representation of the Earth into a 2D map.
- Spatiotemporal data also have a **time reference**.

vibrio_data_by_state

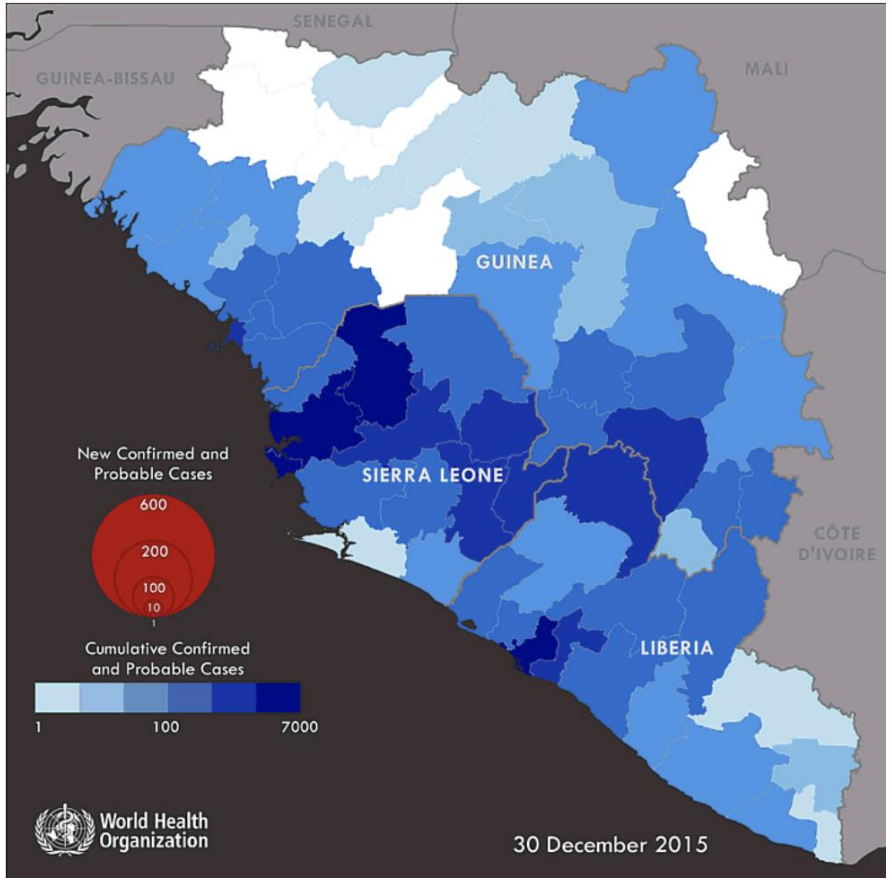
state	date	cases	epiweek	month	year
Alabama	1988-05-02	0	18	5	1988
Alabama	1988-05-09	0	19	5	1988
Alabama	1988-05-16	0	20	5	1988
Alabama	1988-05-23	0	21	5	1988
Alabama	1988-05-30	0	22	5	1988
Alabama	1988-06-06	0	23	6	1988
Alabama	1988-06-13	1	24	6	1988
Alabama	1988-06-20	0	25	6	1988
Alabama	1988-06-27	0	26	6	1988

Fig 4. An example of a spatiotemporal dataset. The data set shows the number of *Vibrio spp*. cases recorded in the State of Alabama, USA over the period May to June 1988.

World Health Organization

Dependency

- In epidemiology, many events show space/time **continuity**.
- If we know the value of an attribute at one point, we could estimate its value at a nearby point.
- This situation is known as **dependency**.
  - Dependency is one of the fundamental characteristics of spatial and spatiotemporal data.
- As we move further away from a point, the association between locations and time periods gradually decreases.
- Data visualisation allows for the visual analysis of potential dependency in the data.



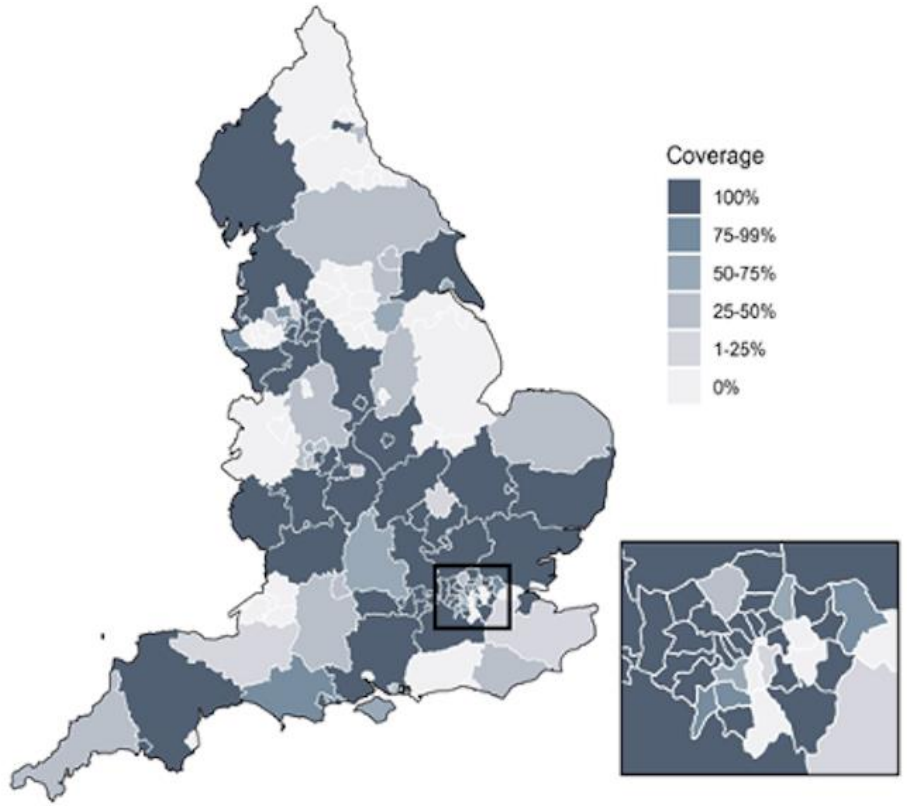
The figure shows the cumulative count of Ebola cases in three African countries over the period 2014 to 2015. We can see that, with some exceptions, adjacent areas tend to be more alike than those that are far apart. In spatiotemporal data, the value observed in a spatial unit tends to be more similar to that of its spatial and temporal neighbours. As we move further from a point in time or space, dependency gradually decreases.

Source: WHO, Emergencies preparedness & response, Ebola Maps.

World Health Organization

Heterogeneity

- The risk of infection varies between geographical areas and across time.
- This property is called **Heterogeneity**.
- Heterogeneity may arise from multiple factors, e.g.:
  - changes in diagnostic methods (e.g., improvements in diagnostic methods)
  - different areas use different diagnostic methods,
  - different surveillance system coverage,
  - urbanisation levels,
  - deprivation,
  - topography



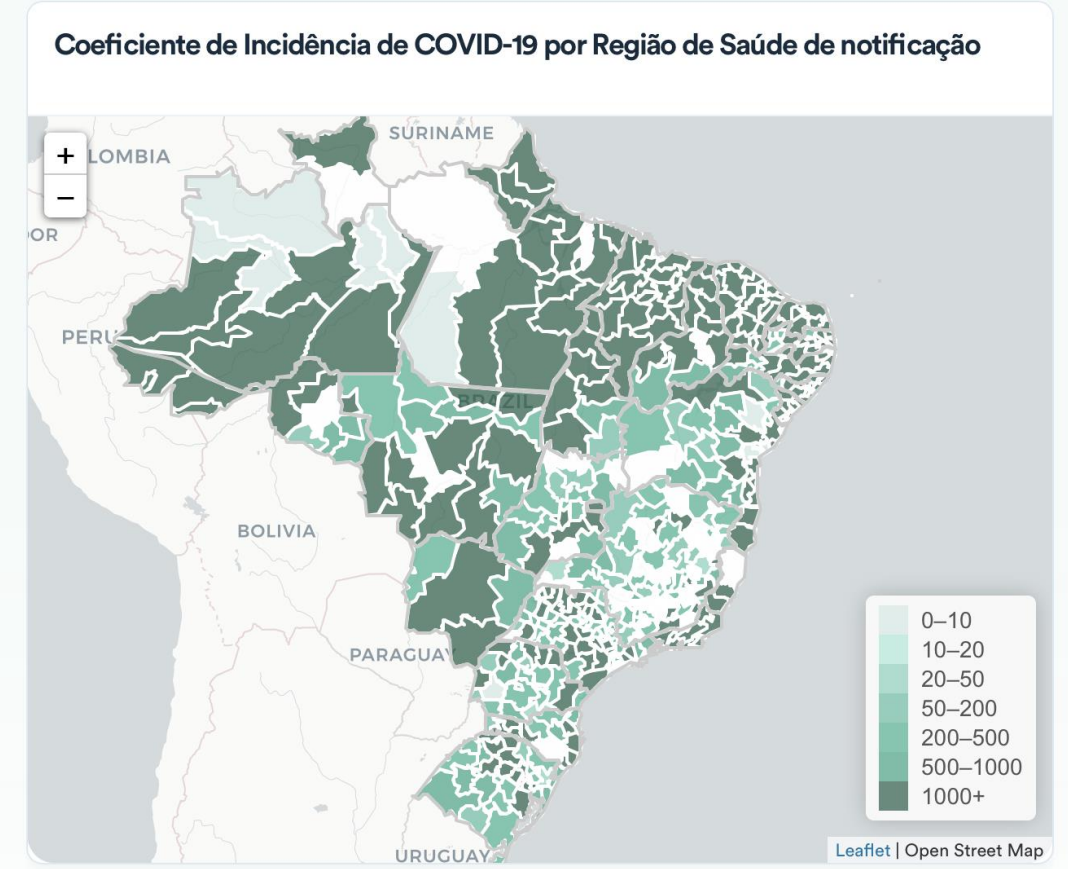
Source: Demographic and socioeconomic patterns in healthcare-seeking behaviour for respiratory symptoms in England: a comparison with non-respiratory symptoms and between three healthcare services | BMJ Open | DOI: 10.1136/bmjopen-2020-038356

This figure illustrates the property of heterogeneity. It shows that the coverage of the general practitioner out-of-hours syndromic surveillance system significantly varies between upper tier local authorities in England.

World Health Organization

Sparsity

- Each year, more data are made available at increasingly finer space/time resolutions.
- While it is tempting to conduct analyses at finer scales, high resolution may pose challenges for statistical analyses due to **insufficient data** at the chosen scale.
- Sparsity may arise from two scenarios:
  - Data arises from small populations (e.g., commune-level data)
  - Events are uncommon (e.g., cases of myeloid leukemia)



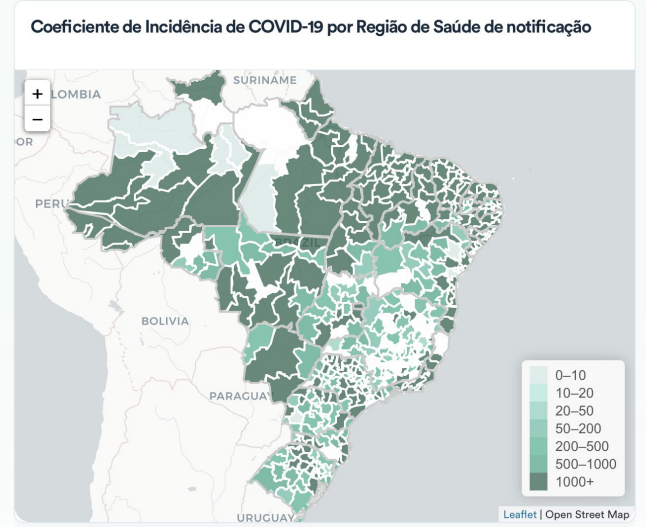
Source: https://covid.saude.gov.br

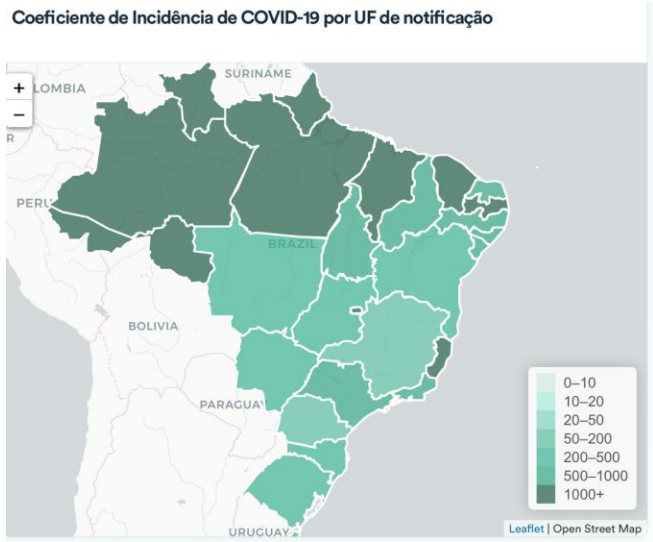
The graph shows COVID-19 incidence rates in Brazil. Notice that there are multiple municipalities with very low counts, which may lead to data sparsity and insufficient data for analysis.

World Health Organization

Aggregation

- **Aggregated data** is often used because individual-level data can be difficult to obtain due to confidentiality or technical issues.
- However, Aggregation leads to **loss of information** about within-area variability.
- Dependency and heterogeneity depend on the **spatial and temporal aggregation** of the data.
- Aggregation introduces **smoothing** (i.e., variation is “covered up” or averaged out).





The graphs show the effect of aggregation on data heterogeneity. The level of spatial heterogeneity is significantly larger in the top figure than in the bottom one.

Source: https://covid.saude.gov.br

World Health Organization

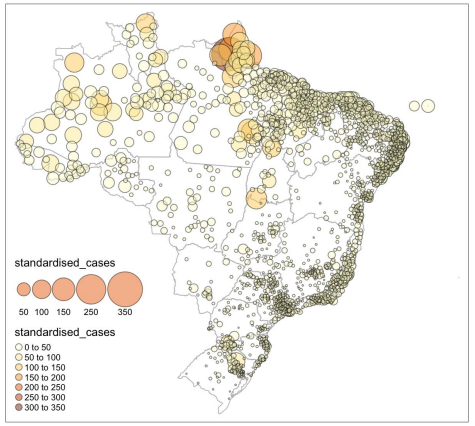
Spatial data

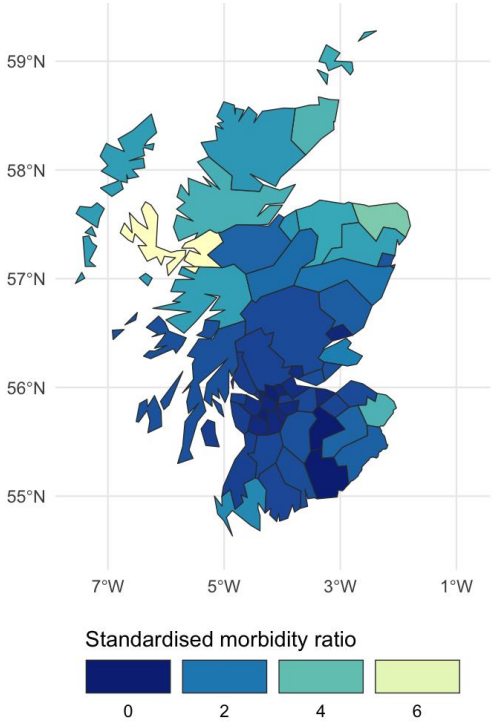
- Due to mobile broadband penetration and distributor servers, the use of spatial data has become more common.
- GPS receivers also make it possible to capture data in the field with accurate positional information.
- Remote sensing stations (such as meteorological and air quality stations) and satellites also make substantial contributions to spatial data acquisition.
- Many computational packages provide solutions for spatial data visualisation.

Common types of spatial data

Point-referenced data

- Measurements of attributes collected at particular locations, described by a single pair of coordinates
- Commonly used for serological and survey data



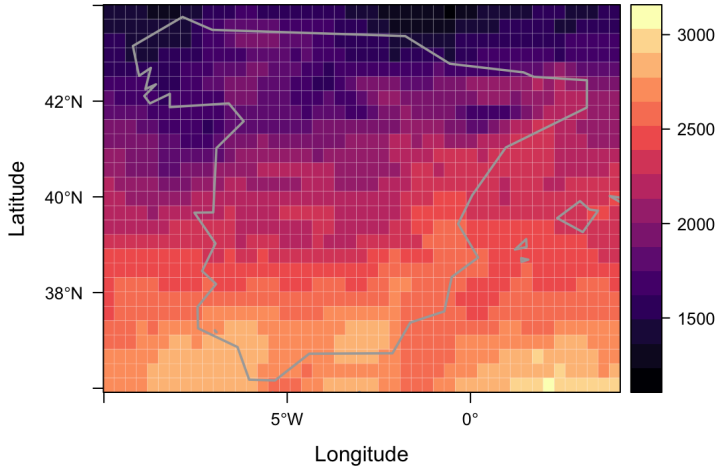


Polygon data

- Well-defined irregular boundaries typically related to geopolitical areas
- Commonly used for health and government data

Gridded data / Raster data

- Well-defined regular boundaries typically shaped as squares or rectangles
- Commonly used for climate and environmental data



World Health Organization

Reflect

- With reference to your area of study, think of examples of the most common visualisation methods used.
- What are the main challenges you face in visualising spatiotemporal data in your field?

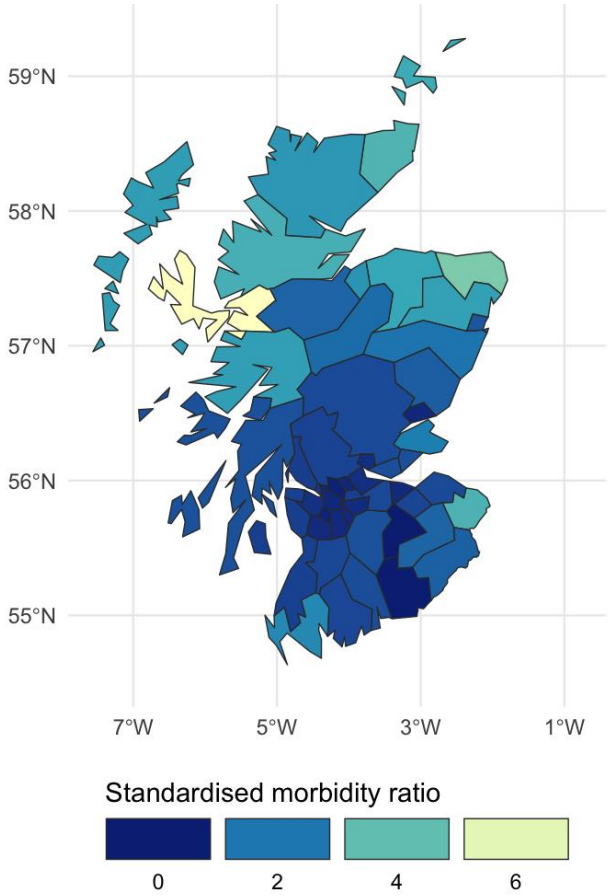
World Health Organization

Spatial mapping

- One of the most common visualisation methods for spatial data is spatial mapping. It has become a critical component of many professionals' toolkits.
- There are multiple visualisation methods and thematic types with different applications depending on the type of data and type of analysis. Some of the **most common methods** for spatial mapping are:
  - Choropleth maps
  - Heat maps
  - Proportional symbol maps
  - Dot density maps
  - Cartograms

Choropleth maps

- Choropleth maps **represent data** through shading patterns on **predefined geographic areas** or polygons.
- While they typically handle numerical data, they can also be adapted to represent categorical data.
- They are useful to represent data variability across geopolitical boundaries.
- The variable should be normalised (i.e., transforming data relative to other variables) by population, area, or another relevant measure to enable comparisons between areas of different sizes or populations.
- Add a legend.

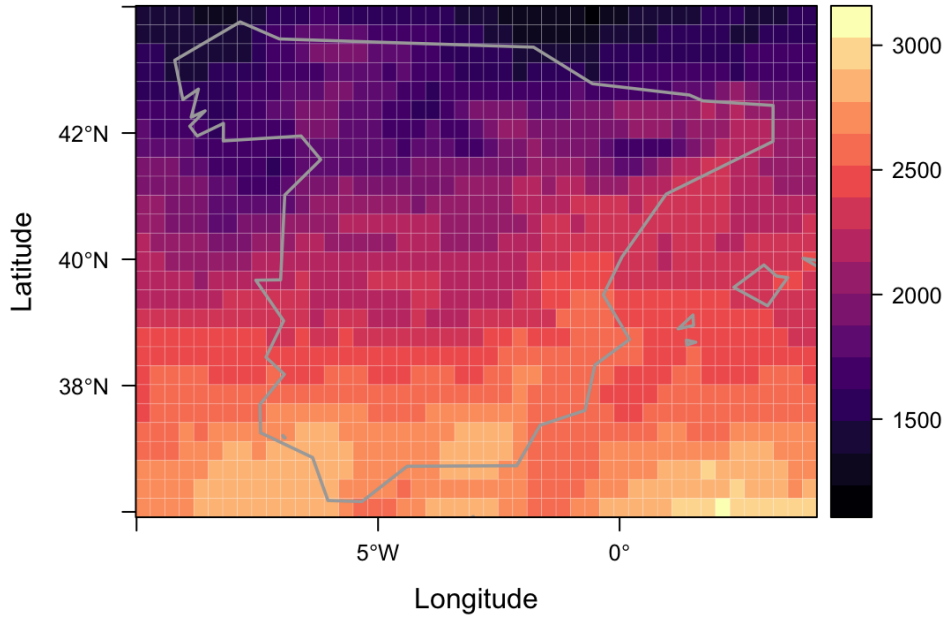


Adapted from Moraga (2019), Geospatial Health Data with R.

World Health Organization

Heat maps

- Heat maps represent the **intensity** of a variable using colour gradients.
- Unlike choropleth maps, **heatmaps do not use geopolitical boundaries** to present data.
- Heat maps are commonly used to visualise climatic and environmental data.

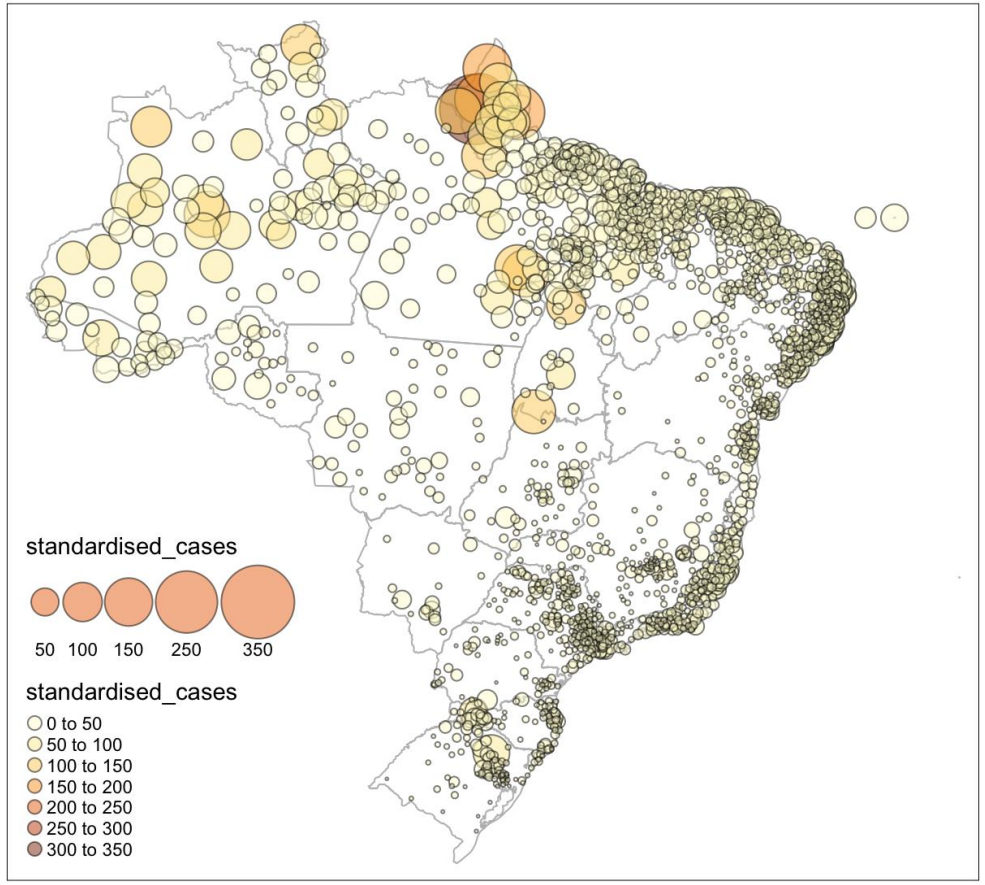


Adapted from Moraga (2019), Geospatial Health Data with R.

World Health Organization

Proportional symbol maps

- Represent data using symbols.
- The size of each symbol may be **proportional to the value** you want to visualise.
- Data may be binned into 3-5 classes to facilitate comparison and classification.
- Adding colour, shape, and transparency can enhance **visual discrimination** between classes.



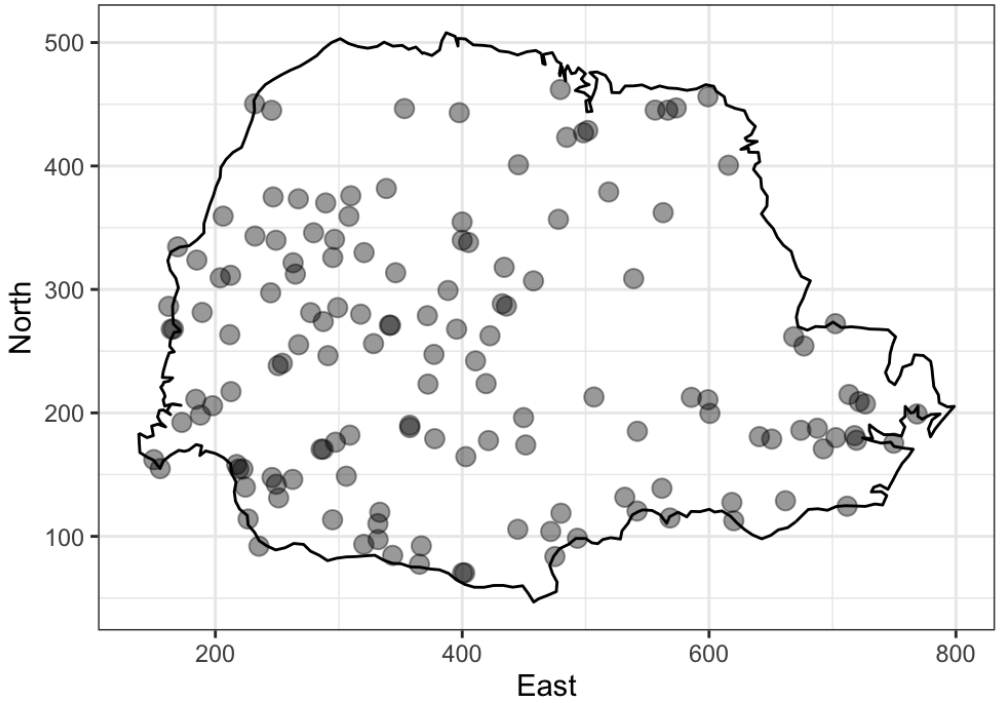
In this example, disease cases are represented by proportional symbol maps indicating the number of cases per location. Each shape and size represent a class, and color has been added to aid in the discrimination between classes.

Adapted from Moraga (2019), Geospatial Health Data with R.

World Health Organization

Dot density maps

- Dot density maps use dots to represent attributes across space.
- Dot density maps may represent attributes **one-to-one** (i.e. each dot represents a single occurrence) or **one-to-many** (i.e. each dot represents a set of aggregated data, e.g. 100 cases).
- They are useful for **identifying clusters**.

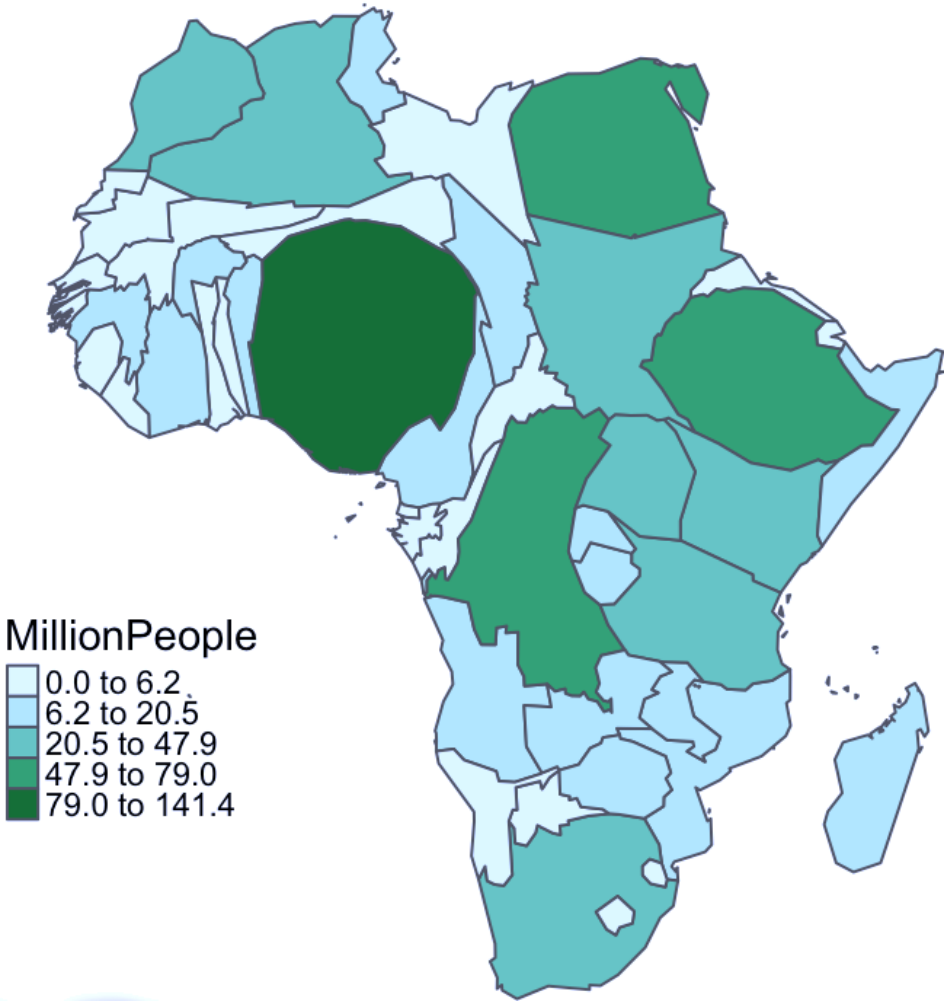


Adapted from Moraga (2019), Geospatial Health Data with R.

World Health Organization

Cartograms

- Cartograms aim to correct the bias introduced by choropleth maps.
- In choropleth maps, regions with very few data points may look as important as a region with many data points.
- **Cartograms** offer a solution to visualise geopolitical boundaries proportional to the value of an attribute in a region.
- Cartograms **distort the geography** so that each region is proportional to the value of the attribute.



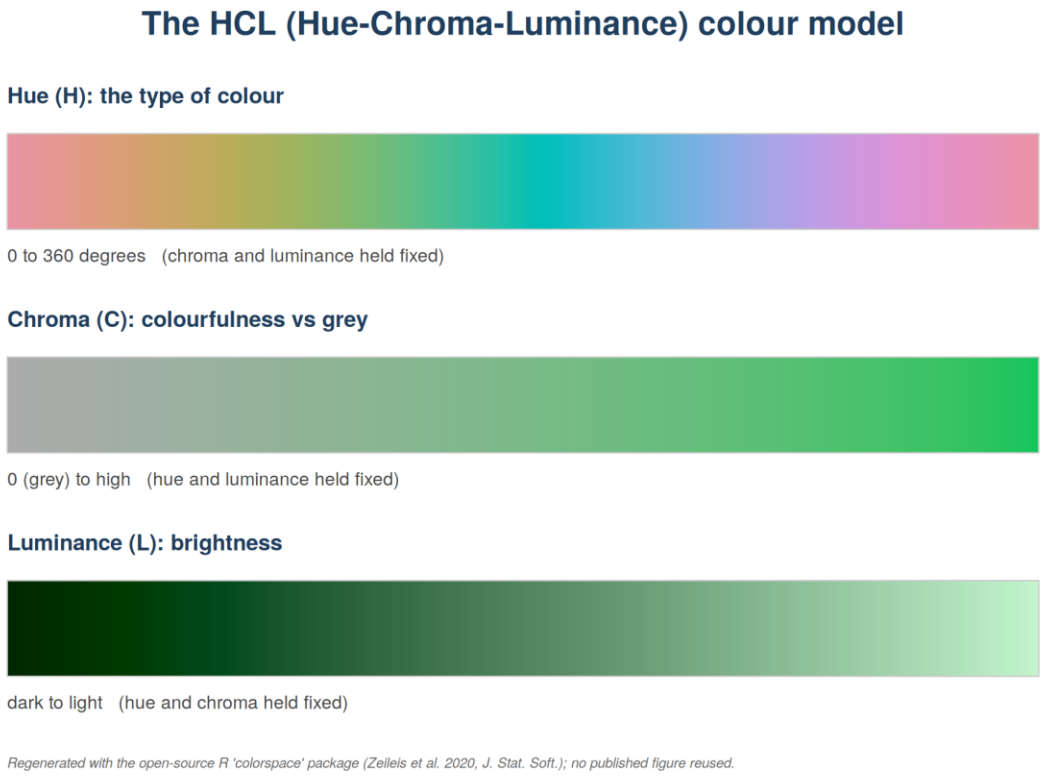
The figure on the right shows a cartogram of the population of African countries in 2005. Notice that Nigeria, Egypt, and Ethiopia appear much larger in the figure than in a classic map, as they have larger populations.

Adapted from Moraga (2019), Geospatial Health Data with R.

World Health Organization

Colour scales

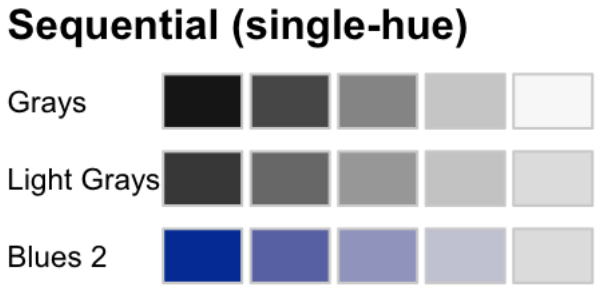
- Our perception of colour is **trichromatic.**
- We describe colour in **three perceptual dimensions**:
  - Hue – Describes colour based on its predominant wavelength; it is unordered and has no inherent lower or higher value.
  - Chroma – Describes intensity, ranging from grey tc maximum intensity.
  - Luminance – Describes the lightness of a colour, ranging from white to black.
- Chroma and luminance are ordered and thus may be suited for values that are ordered.

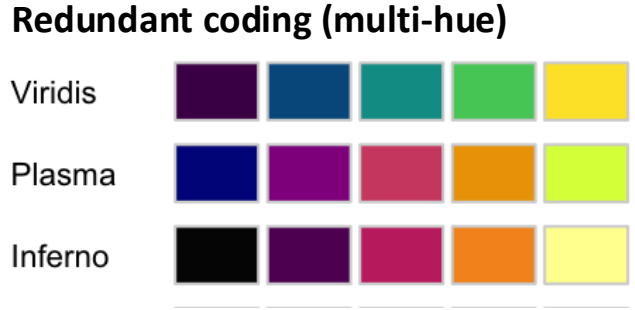


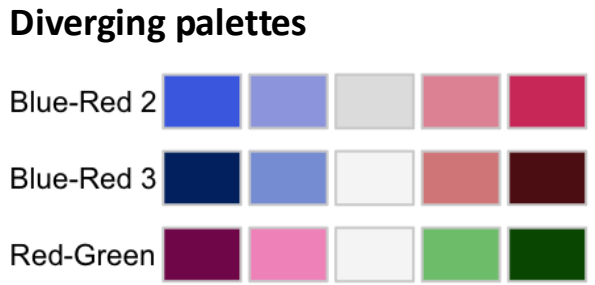
World Health Organization

Colour scales

- There are also different colour scales or palettes you can select depending on the data you want to visualise:
  - Sequential – to represent values from low to high
  - Diverging – to represent values above or below a central value
  - Qualitative – typically only varying in hue
- Online resource to pick the best colour scale given your data: http://colorbrewer2.org/

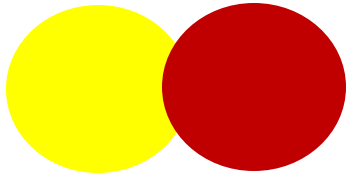




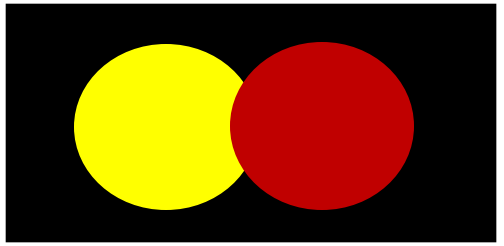


World Health Organization

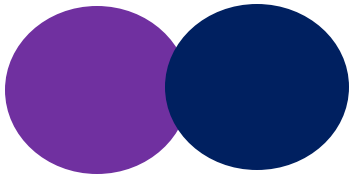
Considerations About Colour



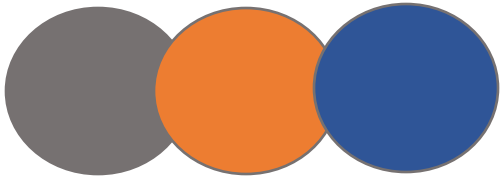
- Some colours are harder to see depending on the background
  - For example, the yellow circle at the top right is harder to see than the red circle when using a white background.



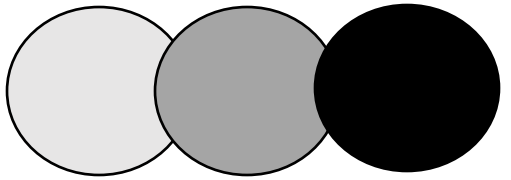
- Some colours are perceptually close and hard to discriminate
  - For example, the blue and purple dots are perceptually close and may lead to confusion when interpreting a map or plot



- Colour vision deficiency (colour blindness)
  - Colour palettes using grey, blue, and orange, or blue and red/blue and brown, are typically more colour-blind friendly.



- Consider access to colour printing
  - Changes in luminance can ensure that visualisations are grayscale-friendly



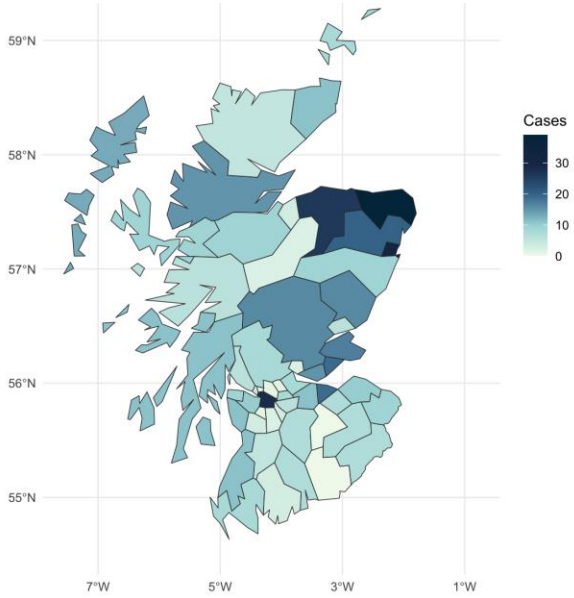
World Health Organization

Example: Lip cancer in Scotland

We will use a dataset on age-standardised relative risks of lip cancer in Scotland, as utilised by Clayton and Kaldor in their 1987 paper.

This dataset includes information on the number of lip cancer cases per county, the expected number of cases, and the latitude and longitude coordinates of each county's centroid.

- Since the data are aggregated at the county level, we will use a choropleth map for visualisation.
- We need to select a colour palette appropriate for the attribute "cases," which is an ordered, numerical, and continuous variable. Therefore, we will choose a sequential colour scale. We will opt for a diverging scale (suitable for attributes not centred on a central value) and select a palette that is both colourblind- and grayscale-friendly.



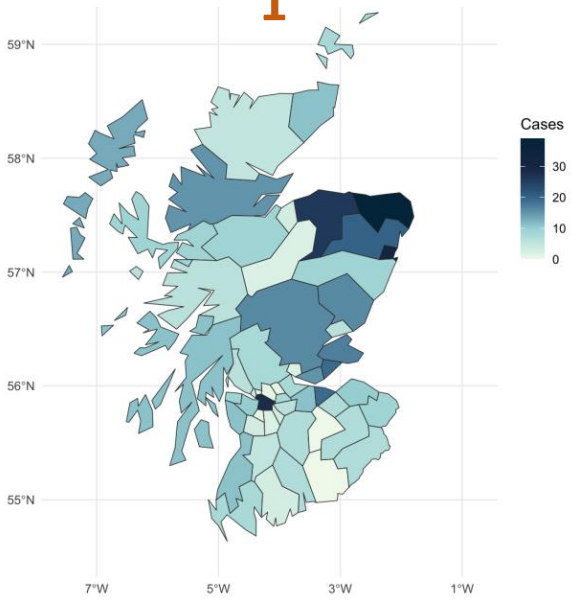
Data: lip cancer in Scotland (Clayton & Kaldor 1987, Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, Biometrics 43:671-681).

World Health Organization

Spatial data visualisation

Example: Lip cancer in Scotland

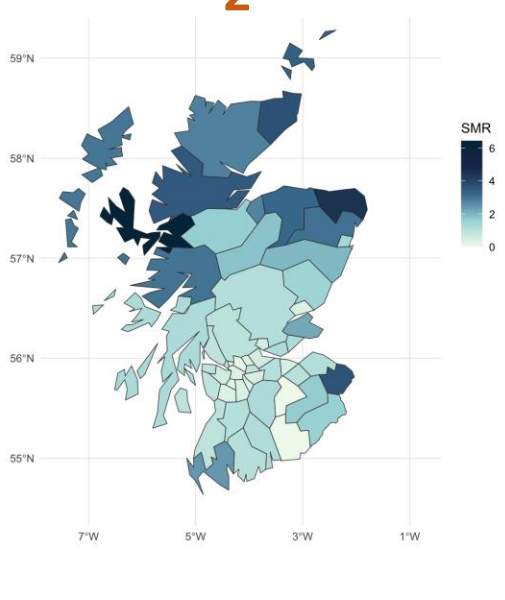
1



Graph #1 shows the number of lip cancer cases per county.

But plotting the number may be misleading, as counties differ in population.

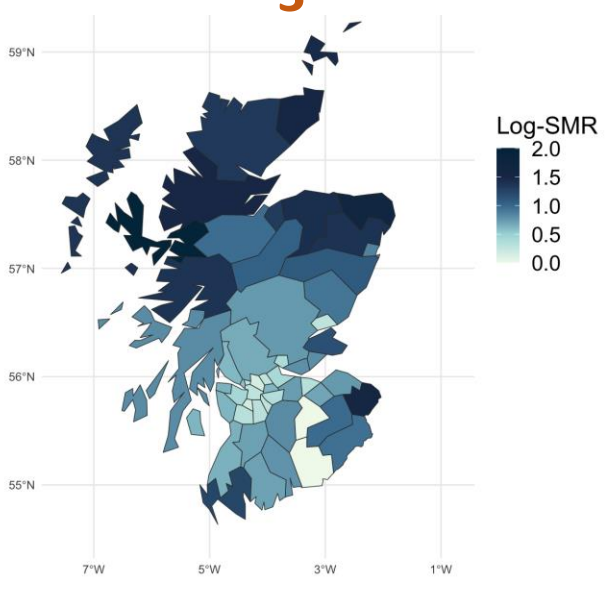
2



Graph #2 shows the standardised morbidity ratio (SMR; observed number of cases is divided by the expected number of cases), i.e., the data has been normalised.

However, it is hard to discriminate between low values in the southern counties.

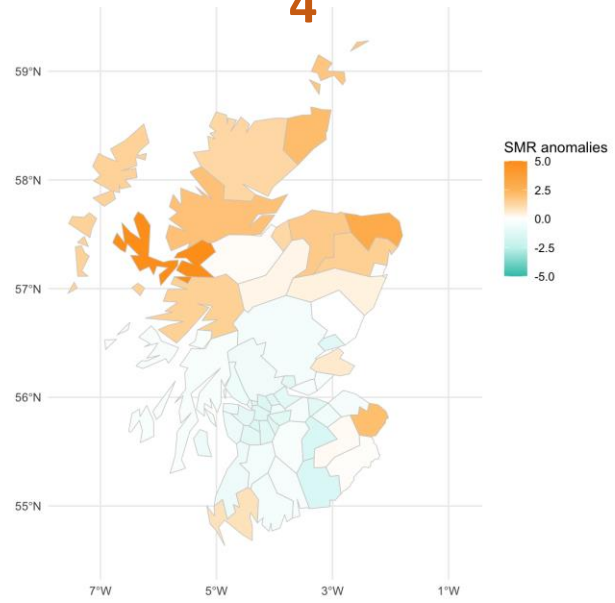
3



Graph #3 shows the log-transformed SMR, enabling better discrimination between counties across Scotland.

We observe that most counties with SMR values above the mean are concentrated in the north.

4



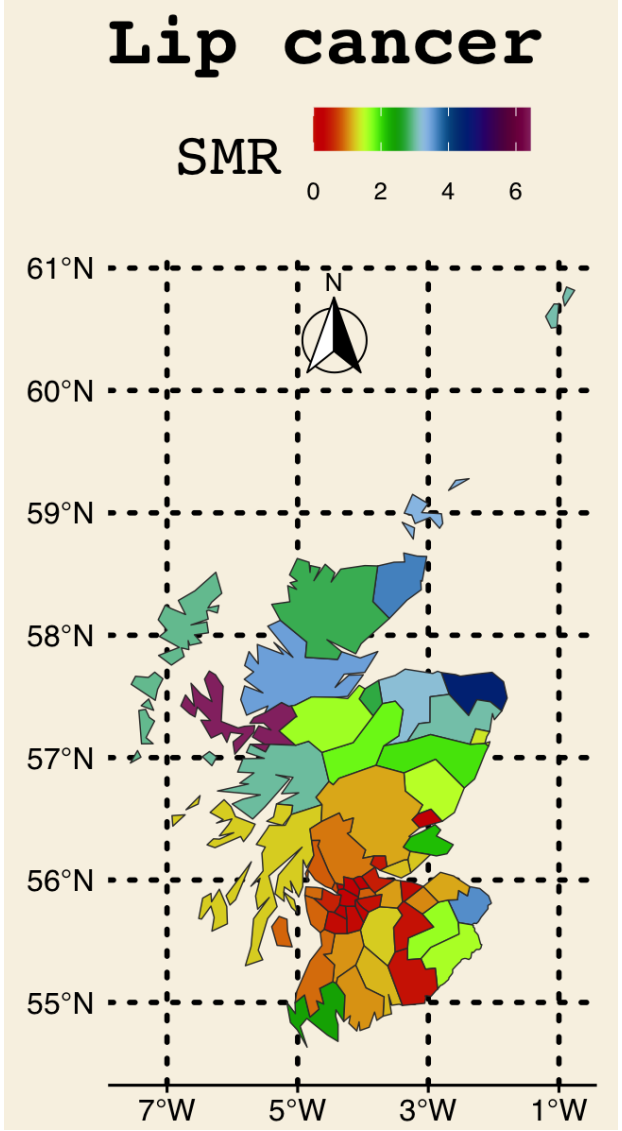
Graph #4 uses a diverging scale to highlight counties with SMR values above and below the mean. SMR anomalies are calculated by subtracting the mean SMR from the county-specific SMR to centre values at the mean SMR.

It highlights that the above-mean-value cluster is in the north, and the below-mean-value cluster is in the south.

Lip cancer in Scotland (Clayton & Kaldor 1987, Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, Biometrics 43:671-681).

Exercise

- Spend some time critically analysing the map on the right-hand side.
- What aspect of this map could be improved?
- What aspects of the visualisation would you keep and why?



Lip cancer in Scotland (Clayton & Kaldor 1987); maps after Moraga (2019).

World Health Organization

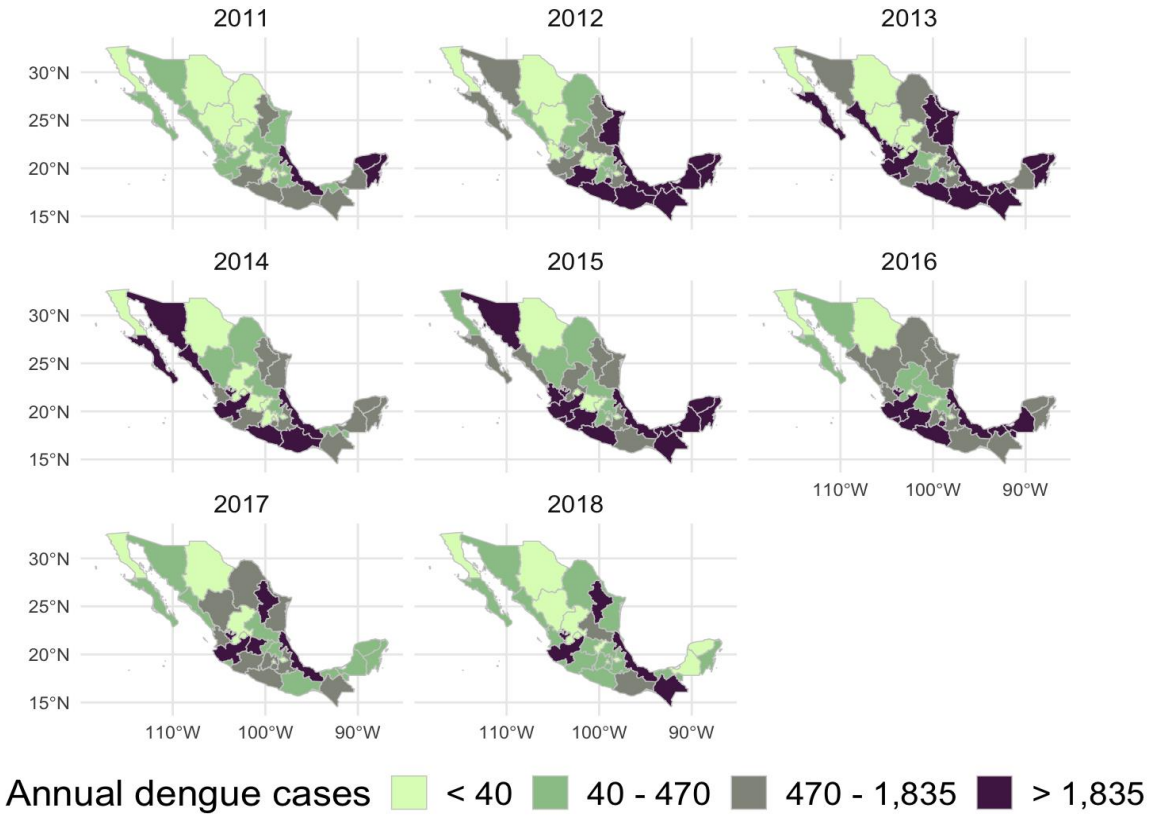
What is spatio-temporal data

- In spatiotemporal data, observations refer to attributes that are measured at **specific locations and times,** incorporating both spatial and temporal references. In tabular data, these references typically correspond to columns.

- One of the simplest yet most effective ways to visualise spatiotemporal data is to generate a sequence of maps **chronologically** in a lattice.

- It is important to use the **same categories** across all maps to ensure they are directly comparable

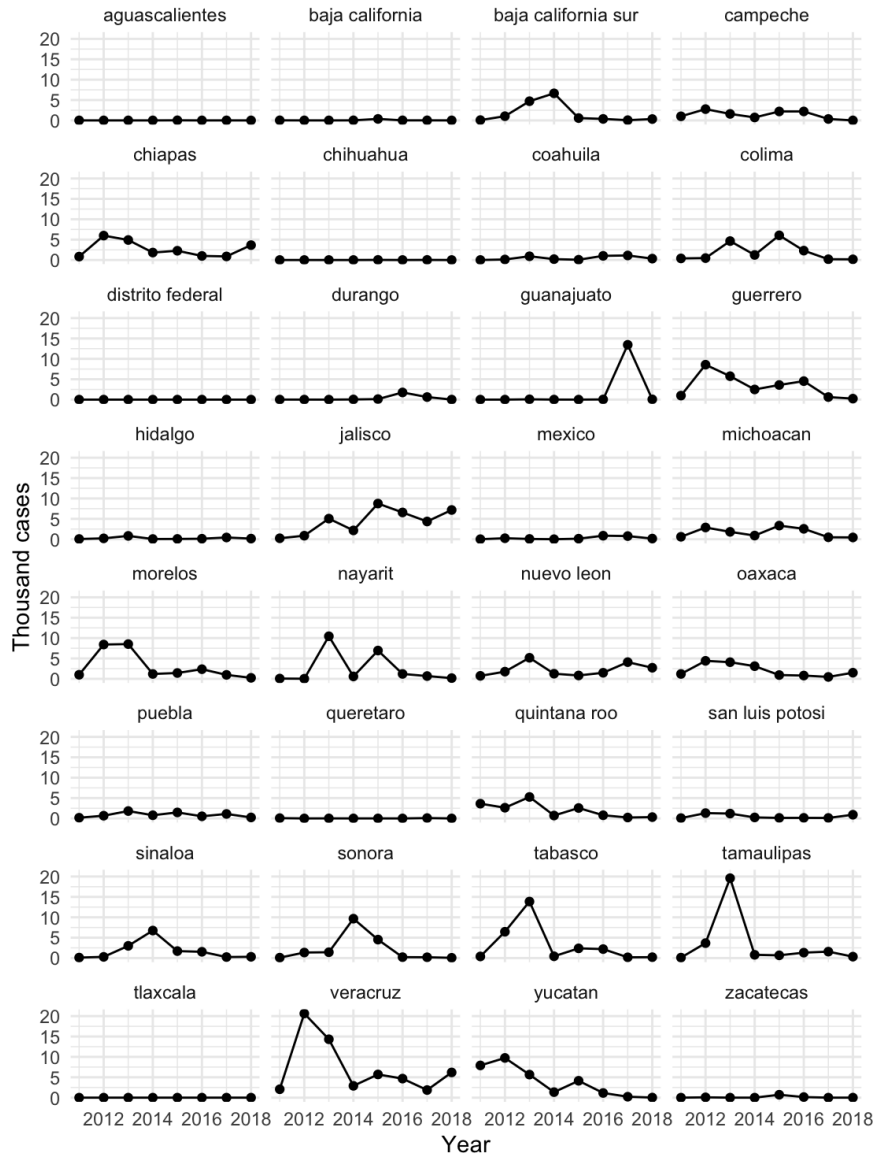
- In this example, we can see that there is substantial **inter-annual variation** in the data.



World Health Organization

Multi-panel time series plots

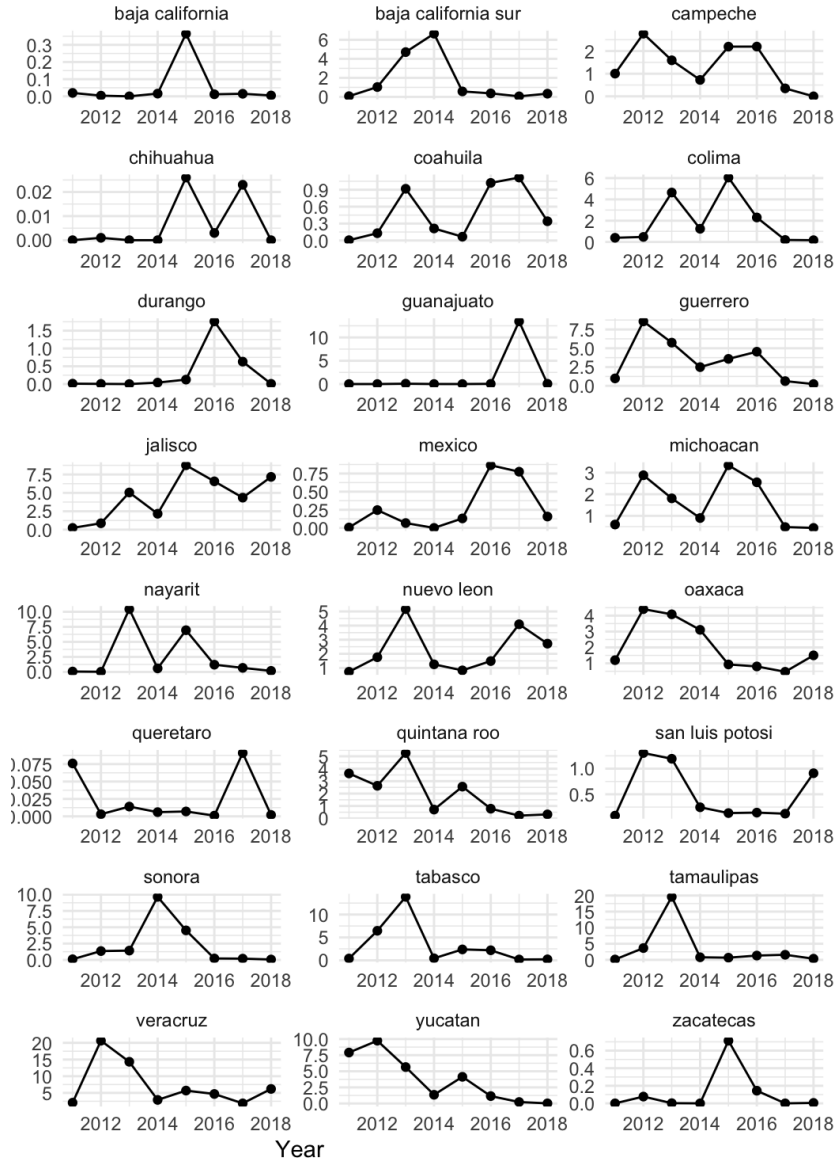
- In these plots, each of the boxes represents a Mexican State, and consecutive observations are joined by straight lines.
- These plots are extremely useful for identifying **long-term trends**.
- In this example, we see that several States in Mexico show an increase in the number of Dengue cases at the beginning of the time series. We also see large differences across states.



Spatio-temporal data

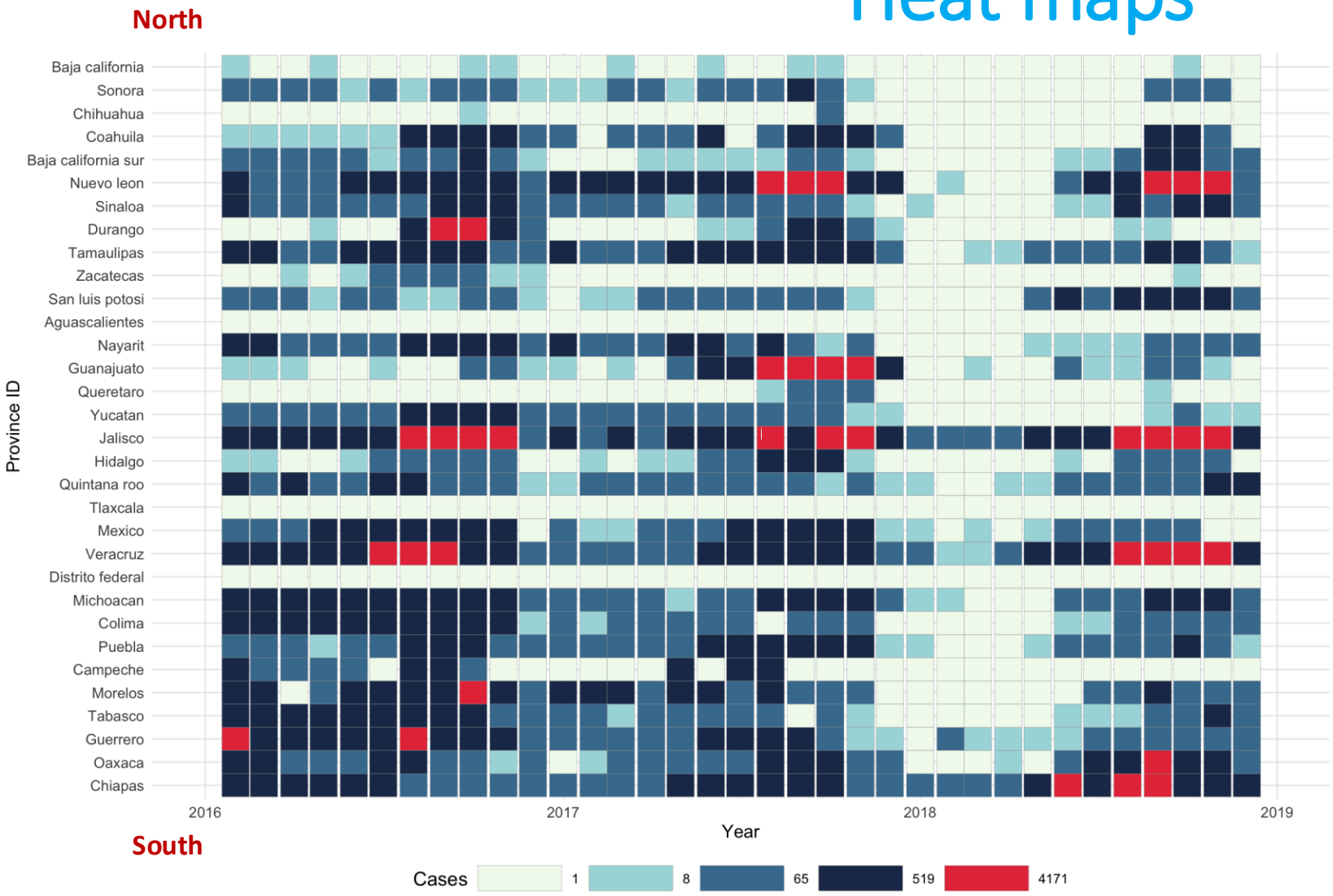
Multi-panel time series plots

- Under some circumstances, you may want to have **independent X, Y axes** for each region.
- This is particularly useful when we want to have a **deeper inspection** of the temporal trends in the data.
- In this example, we can better discern trends that were not obvious when using the same axes for all states.



Spatio-temporal data

Heat maps



- Heat maps are simple but very effective
- Arrange by latitude, longitude or altitude to detect trends
  - Rows indicate regions
  - Columns indicate time steps
- Example: Heatmap of dengue cases across all 32 Mexican states from Jan 2016 to Dec 2018. States are arranged by latitude (northern States are at the top).
- We notice: (1) dengue is present across the country with a few exceptions; (2) there are large inter-annual variations; (3) there is significant seasonality.

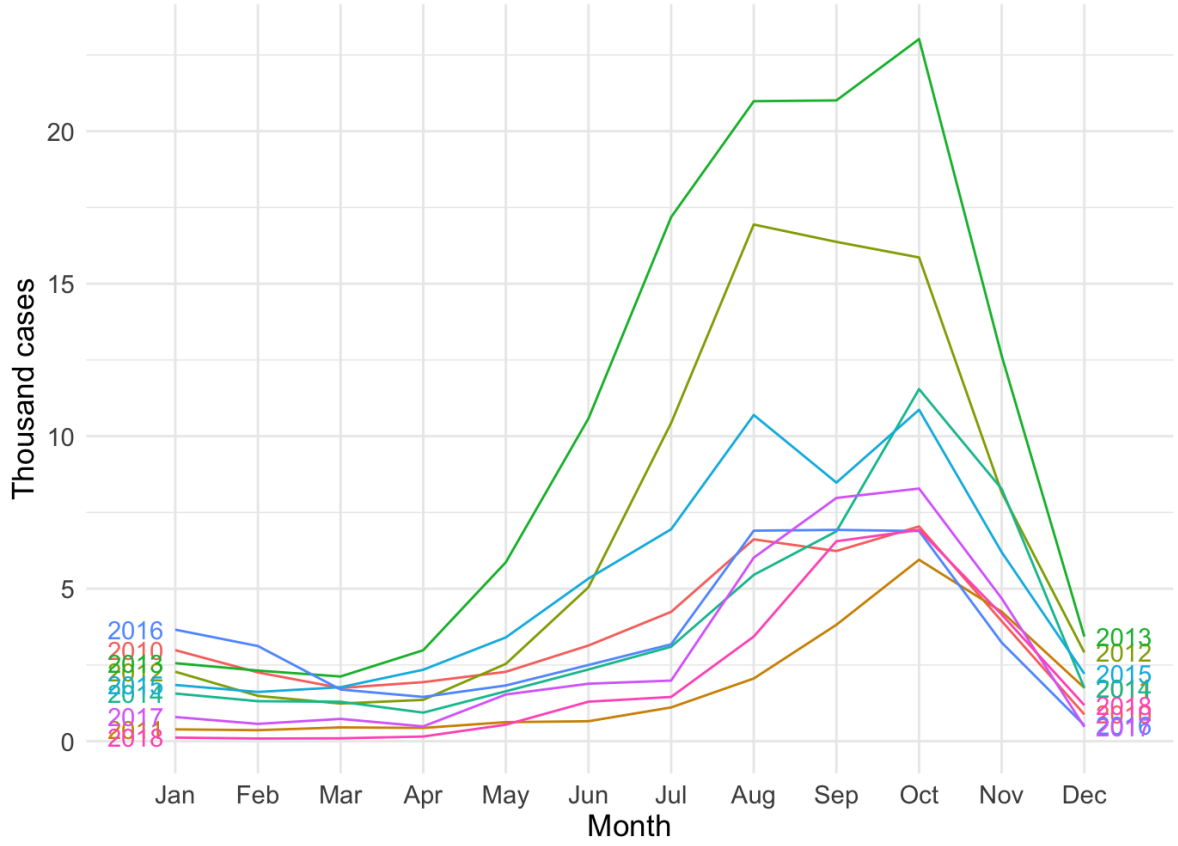
World Health Organization

Spatio-temporal data

Seasonal Spaghetti plots

- Displays data against individual seasons or months.
- Each line represents a year in the series.
- One drawback is that they can easily become messy.

In the graph, we observe that inter-annual variation in dengue transmission in Mexico is greater during the second semester. We also see that the peak of the transmission season is typically between August and October.

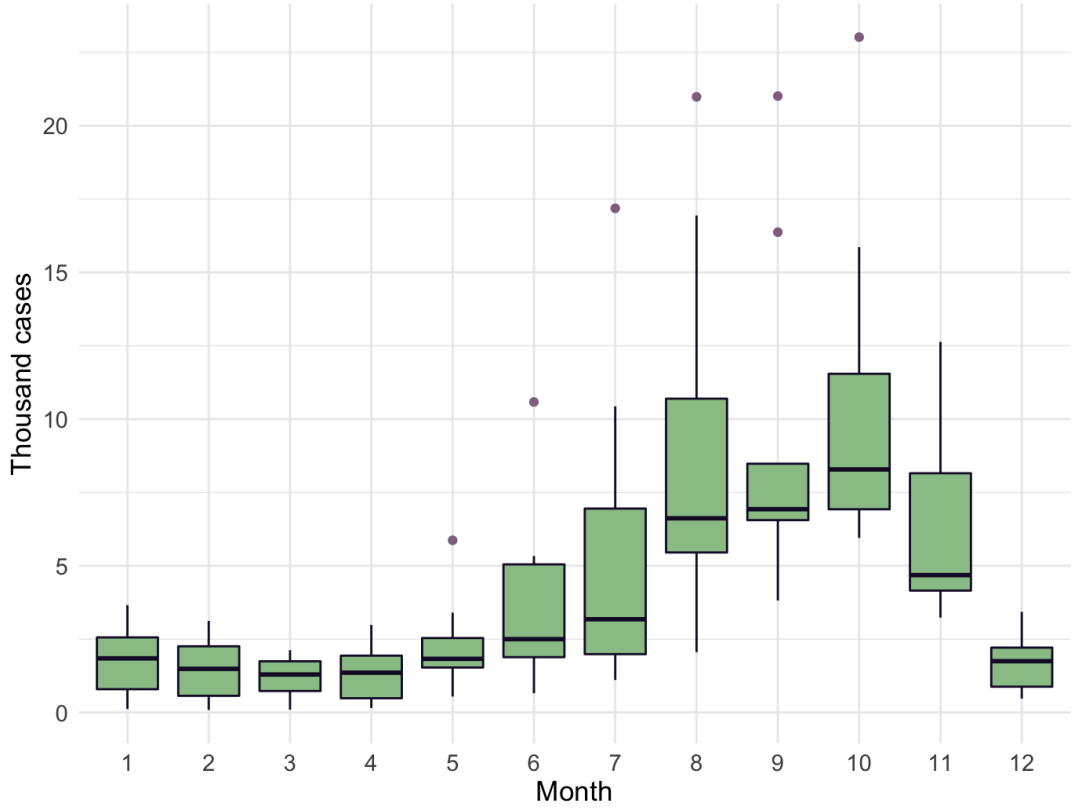


World Health Organization

Box plots

- **Box plots** are a more effective way to visualise seasonal trends.
- The relative length of the boxes and whiskers provides information about the **inter-annual variation.**

In this example, we are presenting the seasonal trends in dengue cases across the whole of Mexico over the period Jan 2010 to Dec 2018, aggregated at the national level. The boxes indicate the interquartile range (i.e., the 25th to 75th percentiles) in the distribution of dengue cases. The line inside the boxes indicates the median number of dengue cases per month. The vertical lines or ”whiskers” indicate the first and fourth quartiles of the distribution of dengue cases. The dots at the top of the figure indicate outliers.



Here, we see that the transmission season begins in May and peaks in October. There are multiple months with large interannual variation (e.g., July, August, October, and November), but there are also other months with very little year-to-year variation (e.g., December to May).

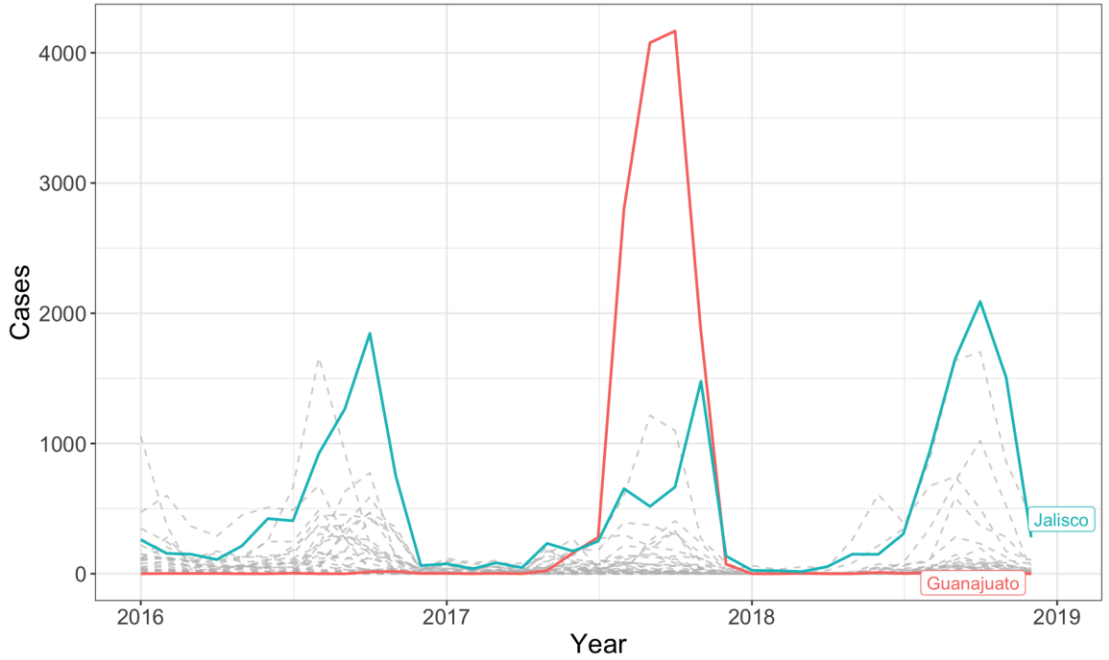
World Health Organization

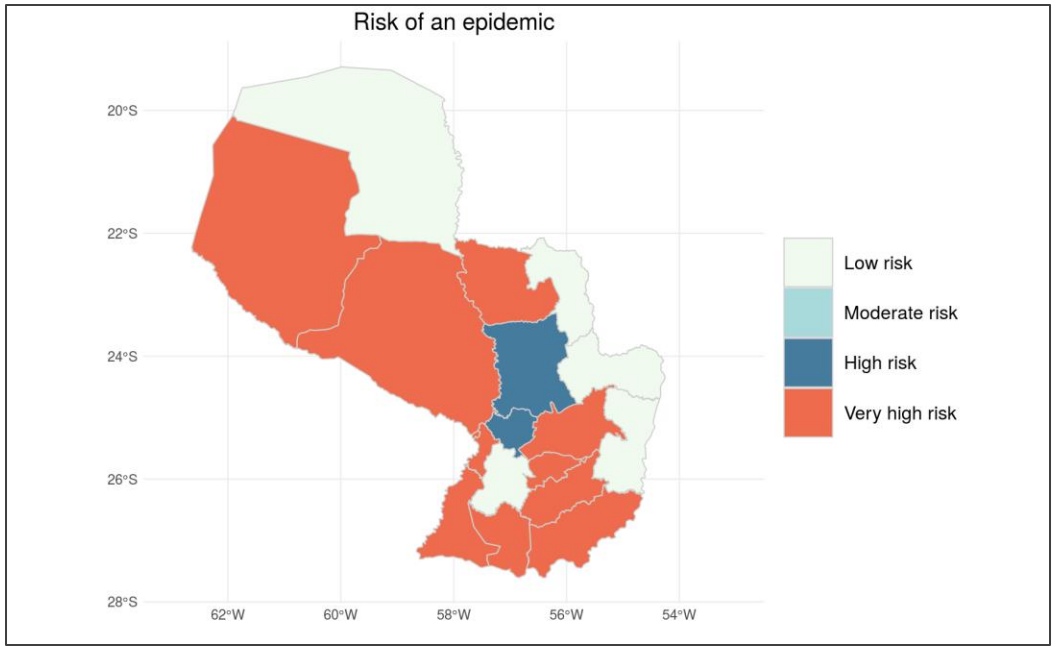
Highlighting

- Highlighting may be useful to make some features conspicuous to the user.
- Highlighting can be drawn using line-widths, colours, and stroke style combinations.

Top figure: In the time series plot example, we used line type and colour to highlight Mexican States where the monthly number of dengue cases exceeds 2,000 at any point in time. The plot shows that only two states are highlighted.

Bottom figure: In the second example, we used colour to highlight areas of very high risk for a dengue epidemic in Paraguay.





World Health Organization

Resources

- Moraga P (2020) Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny https://paula-moraga.github.io/book-geospatial
- Palette selection https://colorbrewer2.org
- Free mapping application https://www.esri.com/en-us/arcgis/products/arcreader
- Free Geographical Information System https://www.qgis.org/
- R Spatial programme https://www.r-spatial.org
- Tableau Public https://public.tableau.com/s/
- Geographic Resources Analysis Support System https://grass.osgeo.org
- GeoDa https://geodacenter.github.io
- gVSIG http://www.gvsig.com/en

World Health Organization

ANNEX: Introduction to open-access data and key software

Dr Simon Hales

World Health Organization

Outline

- Review of desktop software options for data management and analysis
- Online tools/sources for environmental and GIS data

World Health Organization

Software for data input and storage

- MS Excel: easy to use, but easy to make mistakes
  - OK only for simple analyses of small datasets
- Many software systems for data management and statistical analysis: SPSS, Stata, R, Python
- Web-based tools for subsetting and downloading environmental data
- Understanding how to use these systems requires specific training

World Health Organization

Statistical analysis software

- Open access (free)
  - R: powerful, recommended for advanced users, but has a steep learning curve
  - Use with R Studio, see:

    https://rstudio-education.github.io/hopr/starting.html

- Commercial ($)
  - Stata: a relatively easy-to-use alternative to R, particularly good for data management prior to formal analysis

    https://www.stata.com/

Reference: R Project - https://www.r-project.org/about.html

World Health Organization

Geographic Information Systems (GIS)

- Open access (free):
  - Google Earth: good for visualising maps, less powerful for spatial analysis, relatively easy to use.
  - Earth Engine: powerful but requires programming:
    https://developers.google.com/earth-engine/datasets/catalog/
  - R: very powerful, excellent for analysis of time series, spatial, and space-time analyses
    https://www.r-project.org/
  - Q-GIS: similar to the commercial ArcGIS.
    https://qgis.org/en/site/
- Commercial ($$):
  - ArcGIS: powerful, complex software. Easy to create high-quality maps once the data has been entered into the system
    - Educational users can get a 12-month license at a relatively low cost. There are several free licensing options:
      https://www.esri.com/en-us/industries/health/segments/public-health/modernization-eligibility

World Health Organization

Data file formats

- Most statistical analyses use two-dimensional data (like a spreadsheet)
  - Time series: variables in columns, time in rows
  - Spatial analysis: variables in columns, map locations in rows
- GIS data is more complex
  - Vector (e.g. boundaries, for precise locations) vs. Raster formats (matrix of cells/pixels, often used for continuous data like satellite images)
  - often requires several linked files working together
- Geo-scientific data (e.g. climate):
  - Weather station data; 2-D
  - Model output; 3-D (eg. Grib, HDF, NetCDF-- see example)

World Health Organization

Data types

- Human-readable (e.g. ASCII text) vs. binary
- Storage types for numerical data
  - int, byte, double, long, float, string
  - technical but important to understand to avoid pitfalls

World Health Organization

Data and analysis terminology

- Command line vs “drop-down menu” driven
  - For simple analysis, you can use drop-down menus in SPSS, Stata or R-studio
- Documentation
  - Unlike Excel, statistical software allows you to document the analysis as you go.
- Scripting
  - Scripts are human-readable files that automate repeated operations in an analysis task
- Packages
  - Packages are collections of functions and data sets developed by the community. They extend basic functions for specific purposes.
- Programming languages (e.g. C, Java, Python, HTML, SQL)
  - Provide “vocabulary” and syntax for instructing a computer to perform specific tasks.

Reference: Python - https://www.python.org/about/gettingstarted/

World Health Organization

Data management

- Need to be able to combine data from different sources (often in different formats)
- This is becoming easier, but requires skill and experience to avoid pitfalls
- Time series analyses can use data from a single weather station, but that may not be ideal:
  - Alternatively, estimate exposures at the population level (averaged over administrative regions)
- Depending on the data needs, there are many online tools that can help.

World Health Organization

Online tools

- Manage multi-dimensional data remotely, for basic as well as advanced users
- Obtain subsets of the data (for specific regions or time periods)
- Collapse (eg, average) space or time dimensions
  - Average over space to produce a time series of climate for a specific location (eg, city, province, small country)
  - Average data over time to produce a map for spatial analysis
- In some cases, perform statistical analyses remotely

World Health Organization

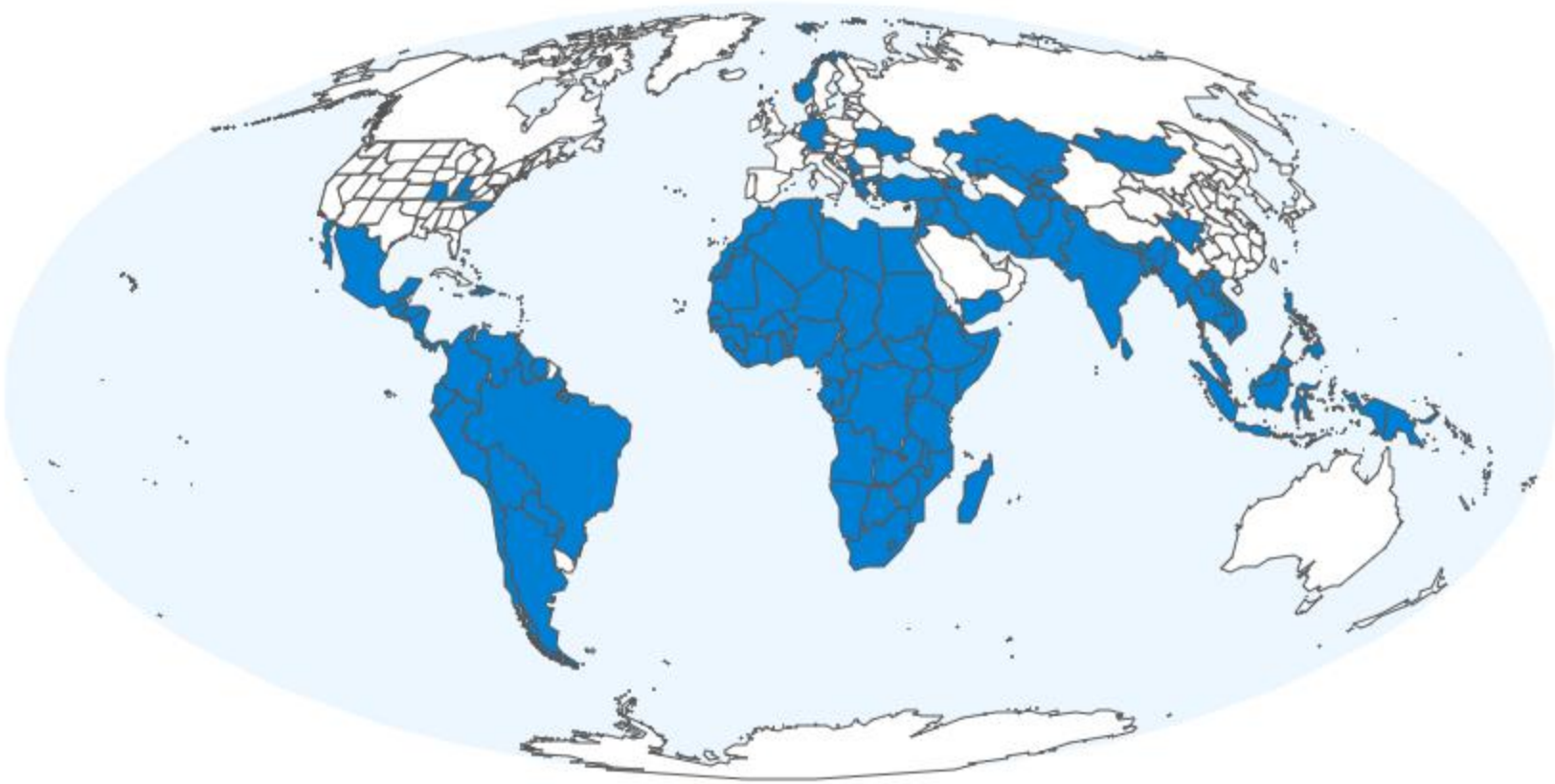
Statistical analyses

- What method should I use?
- This is why we have statisticians. Ask – preferably in advance!
- See section 4 of this course

World Health Organization

DHIS2: open source software for health systems

https://www.dhis2.org/inaction



DHIS2 is an open source, web-based software platform for data collection, management, and analysis. Today, DHIS2 is the world's largest Health Information Management System (HMIS) platform, used by over 130 countries.

World Health Organization

Climate re-analyses

- A climate reanalysis gives a numerical description of the recent climate, produced by combining models with observations.
- Which to use?
  - Best to ask advice. One of the most detailed (hourly resolution, 0.1 degree grid): https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview

World Health Organization

Climate data

All are free of charge, but may require user registration

- Copernicus climate data store

https://cds.climate.copernicus.eu

- IRI data library

http://iridl.ldeo.columbia.edu/

- Climate explorer

http://climexp.knmi.nl/start.cgi?id=someone@somewhere

- Earth Explorer

https://earthexplorer.usgs.gov/

- Giovanni

http://disc.sci.gsfc.nasa.gov/giovanni

- Servir

https://www.servirglobal.net/

- Specific systems for early warning
  - ENACTS malaria (IRI) https://iri.columbia.edu/resources/enacts/
  - EWARN dengue (TDR/WHO): Early Warning and Response System for climate-sensitive diseases (EWARS-csd)

Resources: Copernicus CDS training - https://climate.copernicus.eu/user-learning-services ; RCMRD training - http://rcmrd.org

World Health Organization

Readily available global spatial (GIS) data

- Administrative boundaries
- Population
- Topography, elevation, land use
- Risk indices (extreme events, vector-borne disease)

Reference: Global exposure and vulnerability to multi-sector development and climate change hotspots. Environ. Res. Lett. 13(5),2018. DOI: 10.1088/1748-9326/aabf45

World Health Organization