Please read: Copyright and licensing

Under the terms of this license, you may copy, redistribute and adapt the work for **non-commercial purposes**, provided the work is appropriately cited, as indicated below.

In any use of this work, there should be no suggestion that WHO endorses any specific organisation, products or services.

The use of the WHO logo and other institutions is not permitted.

If you adapt the work, then you must **license your work** under the same or equivalent Creative Commons license.

If you create a **translation** of this work, you should add the following disclaimer along with the suggested citation: “This translation was not created by the World Health Organization (WHO). WHO is not responsible for the content or accuracy of this translation. The original English edition shall be the binding and authentic edition”.

Please read: Disclaimer

This training material was developed by the World Health Organization (WHO). It is intended to be used as educational material.

All reasonable precautions have been taken by WHO to verify the information contained in this training. However, the content is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the online training lies with the reader. In no event shall WHO be liable for damages arising from its use.

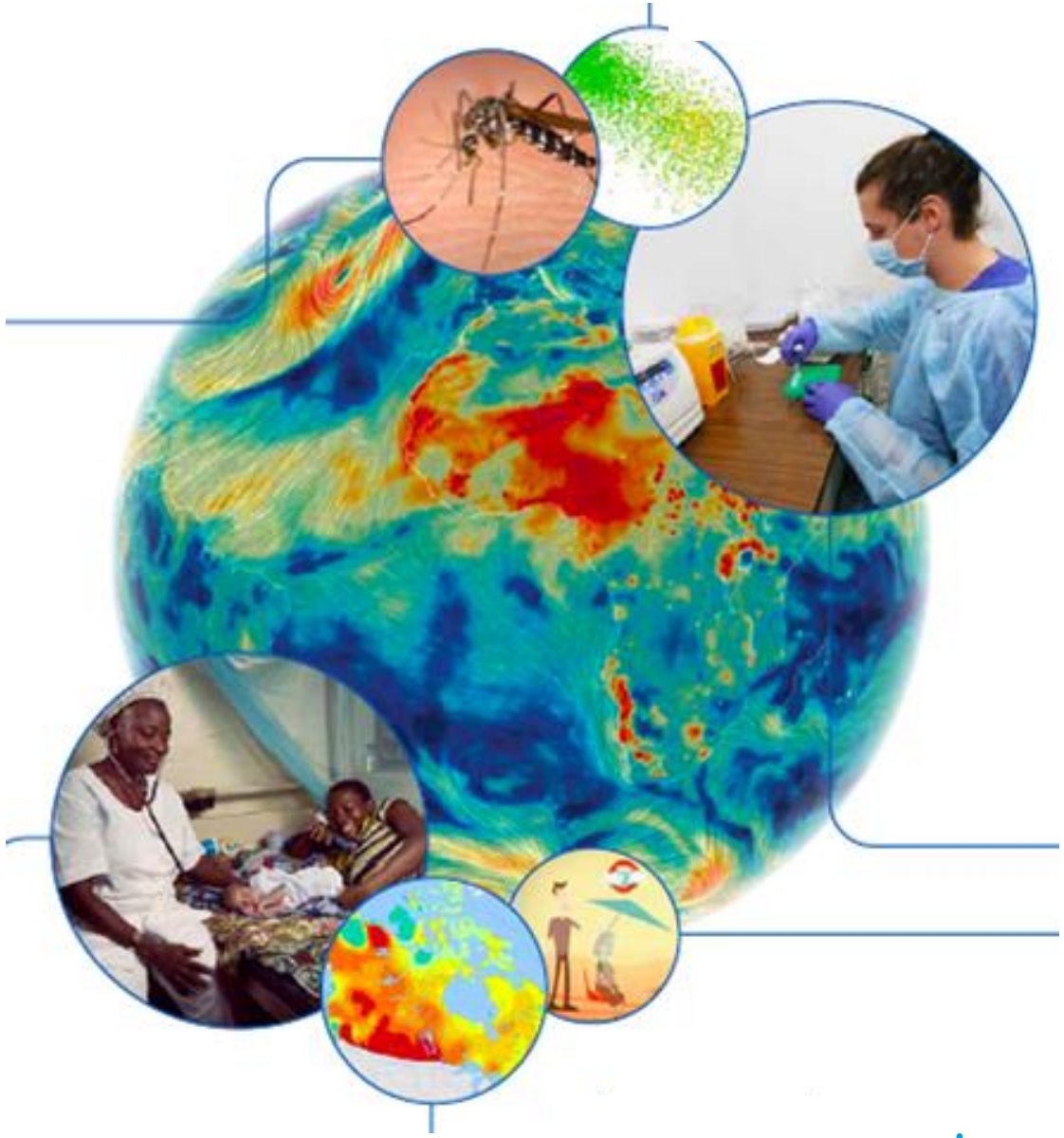
The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific projects or entities does not imply that they are endorsed or recommended by WHO in preference to others of a similar nature that are not mentioned.

Part 3

INTEGRATING CLIMATE-HEALTH DATA THROUGH MODELLING

Enhanced health decision-making using climate information



UMEÅ UNIVERSITET

UF UNIVERSITY of FLORIDA

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

東京大学 THE UNIVERSITY OF TOKYO

World Health Organization

Part 3: INTEGRATING CLIMATE-HEALTH DATA THROUGH MODELLING

Sections

3.1 Best practices for the attribution and quantification of climate and weather effects on disease risk (accounting for other environmental/SES drivers) by epidemiological studies - which analyses and modelling techniques best suit your problem

3.2 Identifying associations by exploring geographic patterns of health risks, and how to account for clustering, hotspots and ecological niche models

3.3 Process-based models (mathematical/simulation models)

3.4 Identifying associations by exploring temporal patterns of health risks: time lags and time series models

Section 3.1:

Best practices for the attribution and quantification of climate and weather effects on disease risk

- **Learning objective**: Understand the notion of hazard, exposure and vulnerability in the context of climate change and attribution of risk

Case study:

Examples of the impacts of heat on human health: A. M. Vicedo-Cabrera et al., (2021). *Nature Climate Change* (11): 492–500

Further reading:

- Ebi KL, Ogden NH, Semenza JC, Woodward A (2017). Detecting and attributing health burdens to climate change. Environ Health Perspect 125(8):085004.https://doi.org/10.1289/EHP1509
- Ebi K, et al. (2020). Using detection and attribution to quantify how climate change is affecting health. Health Affairs 39(12):2168–2174.
- Stone D, et al. (2013). The challenge is to detect and attribute the effects of climate change on human and natural systems. Climatic Change 121(2):381–395.
- Vicedo-Cabrera AM, et al. (2021). The burden of heat-related mortality attributable to recent human-induced climate change. Nature Climate Change 11:492–500.
- Bhaskaran K, et al. (2013). Time series regression studies in environmental epidemiology. Int J Epidemiol 42(4):1187–1195.
- World Weather Attribution, extreme-event attribution resources: https://www.worldweatherattribution.org/

3.1. Best practices for the attribution and quantification of climate and weather effects on disease risk

Dr Jan C. Semenza
Umeå University

© WHO /Bart Verweij

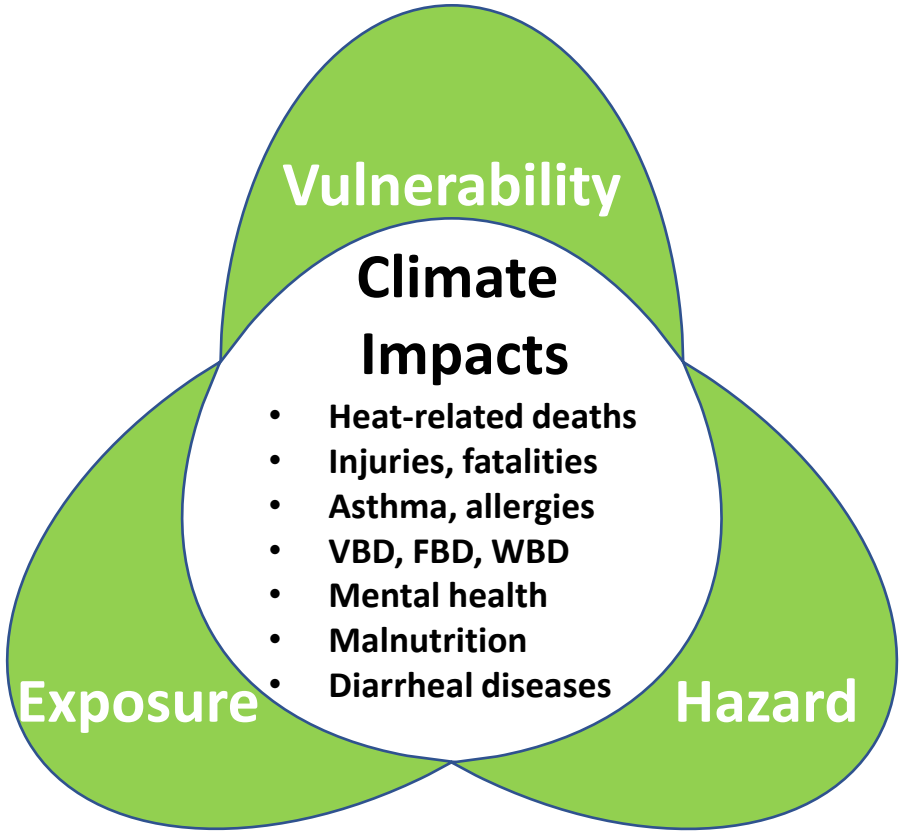
UMEÅ UNIVERSITET

World Health Organization

Learning objective

- Define hazard, exposure and vulnerability in the context of climate change
- Differentiate attribution *versus* detection
- Describe the steps of attribution
- Understand how can heat-related deaths can be attributed to climate change
- Describe the process of extreme event attribution

Attributed climate impacts due to hazard, exposure, and vulnerability



Selected observed climate change impacts due to hazard, exposure and vulnerability

Climate change impacts result from dynamic interactions between climate-related hazards and the exposure and vulnerability of the affected human or ecological system

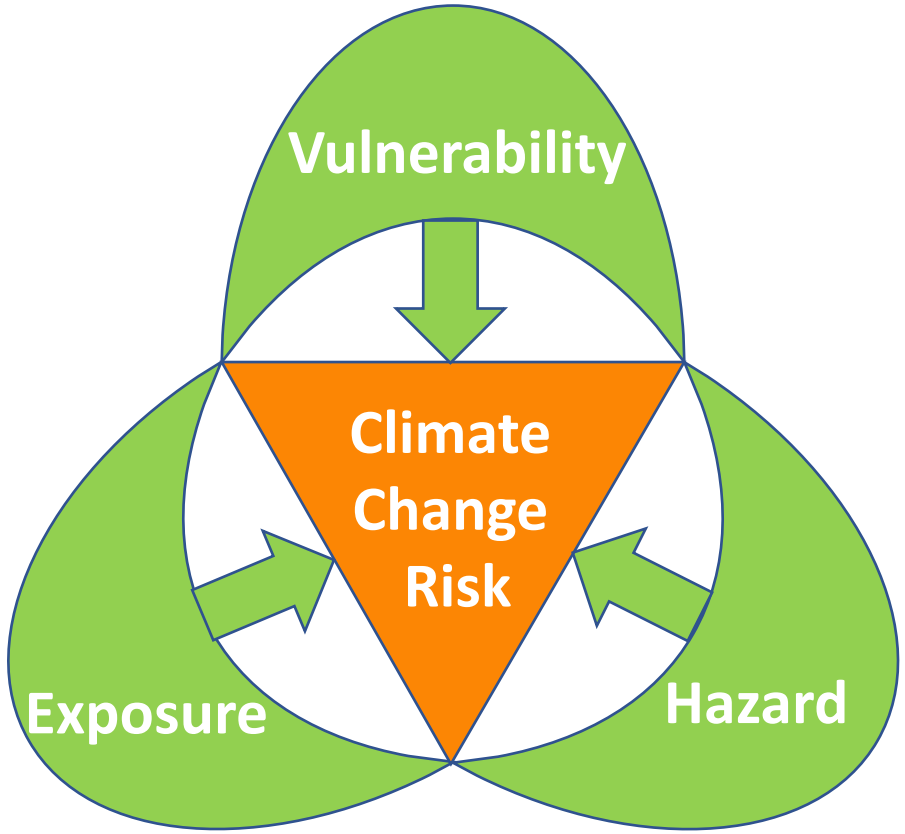
Vulnerability is defined as the propensity or predisposition to be adversely affected. Vulnerability encompasses a variety of concepts and elements, including sensitivity or susceptibility to harm, and lack of capacity to cope and adapt

Hazards are climate events that can be sudden, e.g., a heat wave or heavy rain event, or more slow onset, e.g., land loss, degradation, and erosion linked to multiple climate hazards

Exposure is the state of having no protection from something harmful, such as a climate hazard

Semenza JC, Paz S. (2021) – Climate change and infectious disease in Europe: Impact, projection and adaptation, *The Lancet Regional Health – Europe.*

Climate change risk due to hazard, exposure, and vulnerability



The arrows in the hazard, exposure and vulnerability blades highlight the role of specific actions to reduce climate change risks in the future

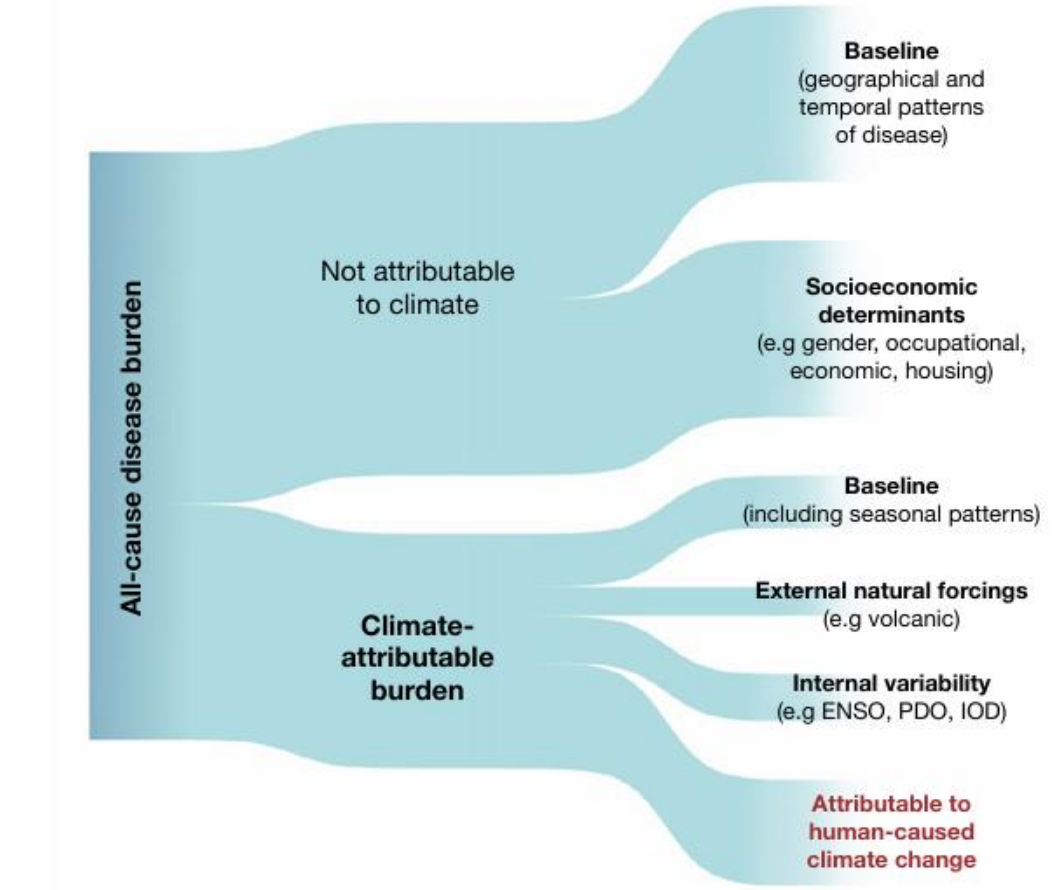
Observed health impacts can be attributed to climate change through **attribution and quantification**

Expected risks to human health can be reduced in the future by acting on hazard, exposure and vulnerability through **mitigation and adaptation**

Semenza JC, Paz S. (2021) – Climate change and infectious disease in Europe: Impact, projection and adaptation, *The Lancet Regional Health – Europe.*

Attribution

- Attribution addresses the question of the magnitude of the contribution of climate change to disease burden
- It indicates how much of the observed change (e.g. heat-related mortality) is due to climate change, with an associated confidence statement
- Therefore, attribution requires the evaluation of the contributions of all external drivers to the system change



(ENSO: El Niño-Southern Oscillation; PDO: Pacific Decadal Oscillation; IOD: Indian Ocean Dipole.)

Drivers and sources of variation in all-cause disease burden

CJ Carlson, et al. 2024- Detection and attribution of climate change impacts on human health: a data science framework.

Ebi et al., (2017) - Detecting and Attributing Health Burdens to Climate Change, *Environ Health Perspect*.

Attribution vs. detection

- Detection of a climate signal does not necessarily suggest significant attribution
- The signal needs to be significantly different from what can be expected from natural variability
- If so, it can be attributed

Ebi et al., (2017) - Detecting and Attributing Health Burdens to Climate Change, *Environ Health Perspect*.

Plausible vs possible

1. **Hypothesis**: Identification of potential impact
2. **Climate change**: documentation of a trend in climate at relative scales
3. **Climate baseline**: specification of system behavior in a stable climate
4. **Impact detection**: demonstration of consistency of observations with a climate effect and inconsistency with plausible alternatives
5. **Attribution**: determination of the magnitude of the relative contribution of climate change

Hansen et al., 2015. Linking local impacts to changes in climate: a guide to attribution, *Reg Environ Change*
DOI 10.1007/s10113-015-0760-y

Attribution vs. detection

- **Detection**:
  - Over the last decade, has a change been **observed** in the system, relative to normal conditions ("over and beyond a predefined baseline)?

- **Attribution**:
  - To what extent has this change been **caused** by observed changes in climate (e.g. warming, rainfall)

Ebi et al., (2017) - Detecting and Attributing Health Burdens to Climate Change, *Environ Health Perspect*.

Challenges to attribution and quantification

- There are two major challenges: observations and process
- Observations:
  - Lack of data:
    - High-quality and long-term **health data** are often not available or missing
    - High-quality and long-term contextual data on **determinants** of health are often not available or missing
- Process
  - Lack of understanding:
    - Climate change, in conjunction with other factors, may affect the system in question
      - nonlinear—for example, involving **threshold** effects
      - non-local in both **space and time**
      - lagged **responses**
      - **trans-regional effects** due, for example, to trade or migration

Ebi et al., (2017) - Detecting and Attributing Health Burdens to Climate Change, *Environ Health Perspect*.

Objective

- Building the evidence base on detection and attribution is the basis for better evidence-based risk management to reduce current and plan for future climate change-related health burdens, and to inform advocacy for actions to mitigate greenhouse gas emissions.

- Attribution studies typically focus on whether and to what extent a system has changed in response to observed climate change.

© Nick Youngson / Pix4free.org

Background

- Detection and attribution studies determine:

a) Whether a system is changing beyond a specified baseline that characterises behaviour in the absence of climate change; and

b) Whether climate change has contributed substantially to the observed change in a system.

Stone D, et al., (2013). The challenge to detect and attribute effects of climate change on human and natural systems, *Climatic Change* 121(2):381–395

Methods

- Multi-step attribution links separate single-step approaches that
  a) Attribute is an observed change in a variable of interest to a change in climate or other environmental variable, and
  b) attribute the change in climate to external drivers (e.g., greenhouse gas emissions).
  c) assess the relationship between climate and the variable of interest
- For example:
  - The first step links temperature increases to heat-related mortality ('impact attribution').
  - The second step links the temperature increase to anthropogenic greenhouse gas emissions ('climate attribution').
  - The assessment of the relationships between climate and the variable of interest may involve a process model or statistical association. This approach has been used for detection and attribution (D&A) in ecosystem studies.

Methods

- Specifying the causal factor to which a change is being attributed is important. Many climate detection and attribution studies focus on expected fingerprints of changes in weather patterns associated with climate change and their associated uncertainties, where fingerprints are metrics or space–time patterns of the response of climate variables to anthropogenic (e.g., greenhouse gas emissions) or natural (e.g., solar radiation changes) forcing.

- Examples of fingerprints include changes in global mean surface temperature, precipitation, and sea ice extent.

Case study

Heat-related mortality

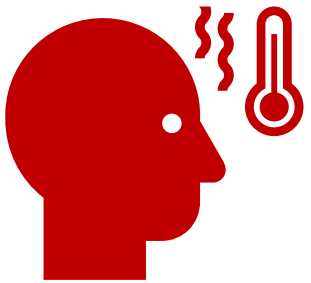
- There is robust evidence that:
- climate change is affecting the frequency, intensity, and duration of heatwaves; and
- exposure to high ambient temperatures is associated with excess morbidity and mortality

© WHO/Mahmoud Hamda

Attributing heat-related deaths to climate change

- In the first step, time-series regression techniques can be applied to observed temperature and mortality data to estimate location-specific exposure-response functions.

- These functions characterise the complex **relationship between daily mean temperature and mortality** from all causes (or non-external causes) by simultaneously accounting for the nonlinear and delayed dependencies typically found in this type of assessment.



A. M. Vicedo-Cabrera et al., (2021). *Nature Climate Change* (11): 492–500

Case study

Attributing heat-related deaths to climate change

- The usual regression method of choice for analysing count data is **Poisson regression**
- **Time-series analyses** for observed temperature and mortality data over the warmest consecutive months in each location
- **Delayed exposure** effects: lagged association

Delayed exposure and 'lags'

It would be ideal to relate temperature on one day to the mortality on that day. However, often there is a delayed association.

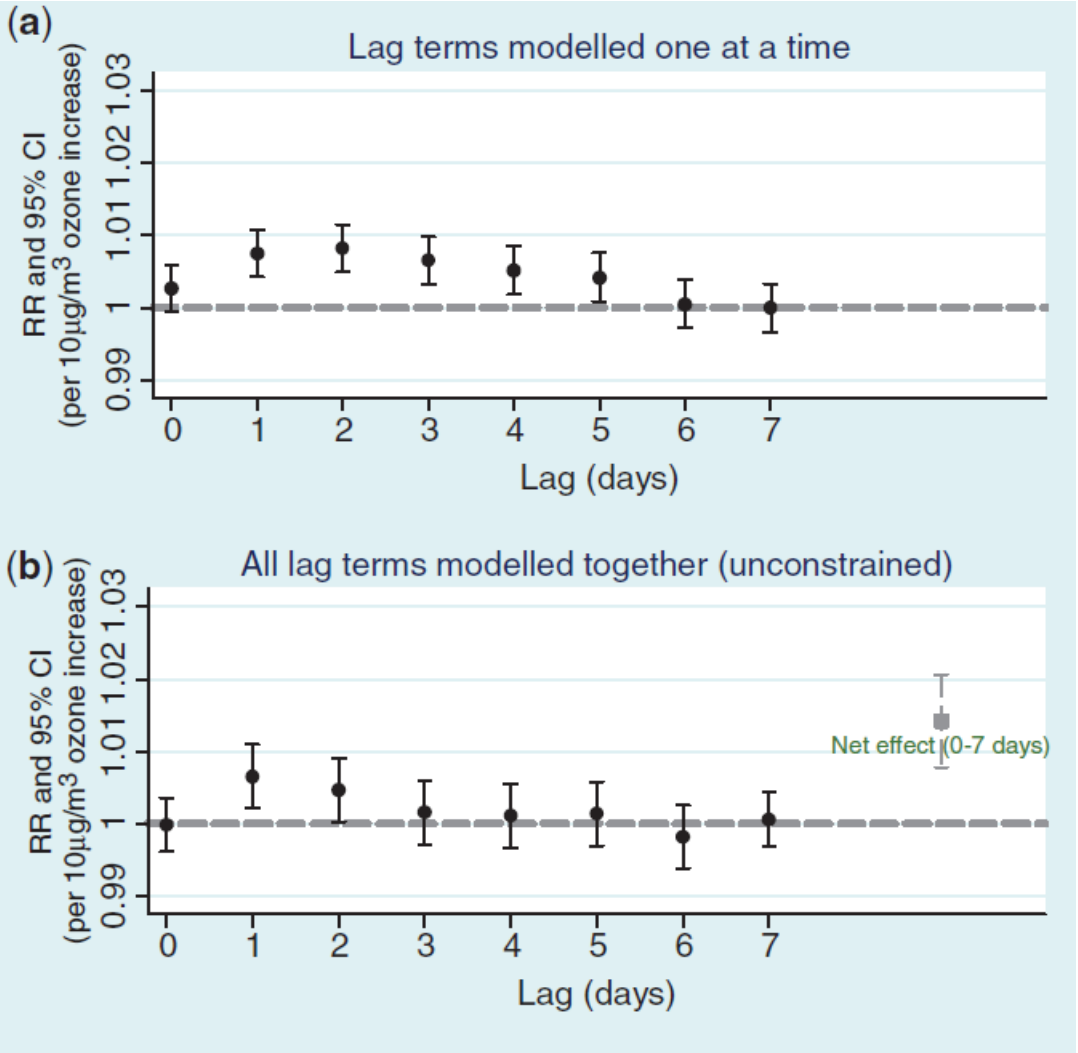
However, by shifting the temperature variable and including it in the model the association between the current day mortality and temperature the previous day can be explored.

A. M. Vicedo-Cabrera et al., (2021). *Nature Climate Change* (11): 492–500

Case study

Allowing for delayed exposure effects

- Estimating the association between previous day exposure (e.g. temperature) and the current day's mortality risk is a question of shifting the temperature series forward in time
- The lag can be increased from 0 to 7 days. There seems to be an association between exposure and mortality when the lag time is between 1 and 5 days (panel a)
- Different lag effects can be adjusted for each other by entering them simultaneously into the model (panel b)

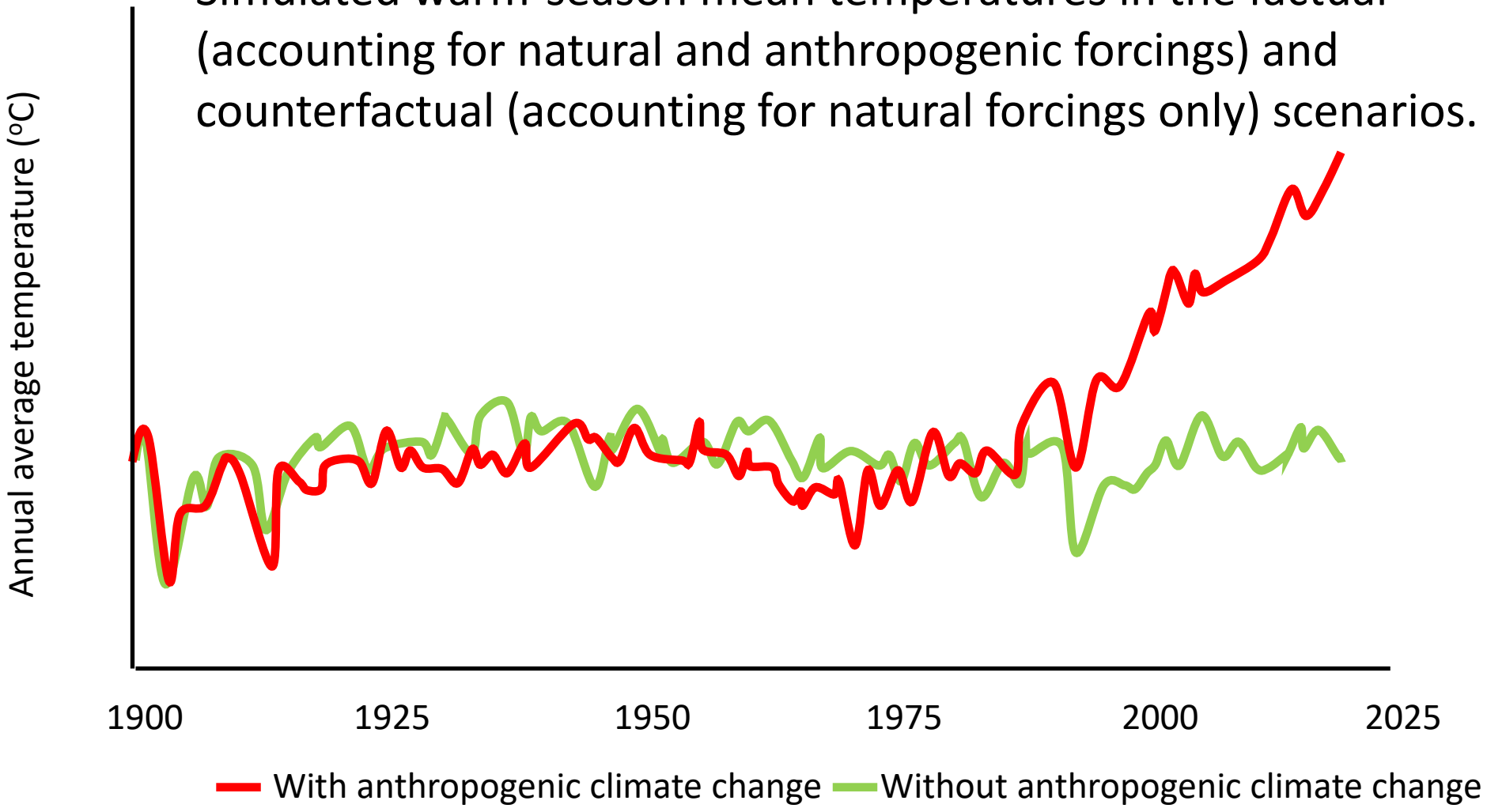


Bhaskaran et al., (2013). *International Journal of Epidemiology*, 42(4): 1187–1195

Temperature modelled with and without anthropogenic climate change

Case study

Simulated warm-season mean temperatures in the factual (accounting for natural and anthropogenic forcings) and counterfactual (accounting for natural forcings only) scenarios.

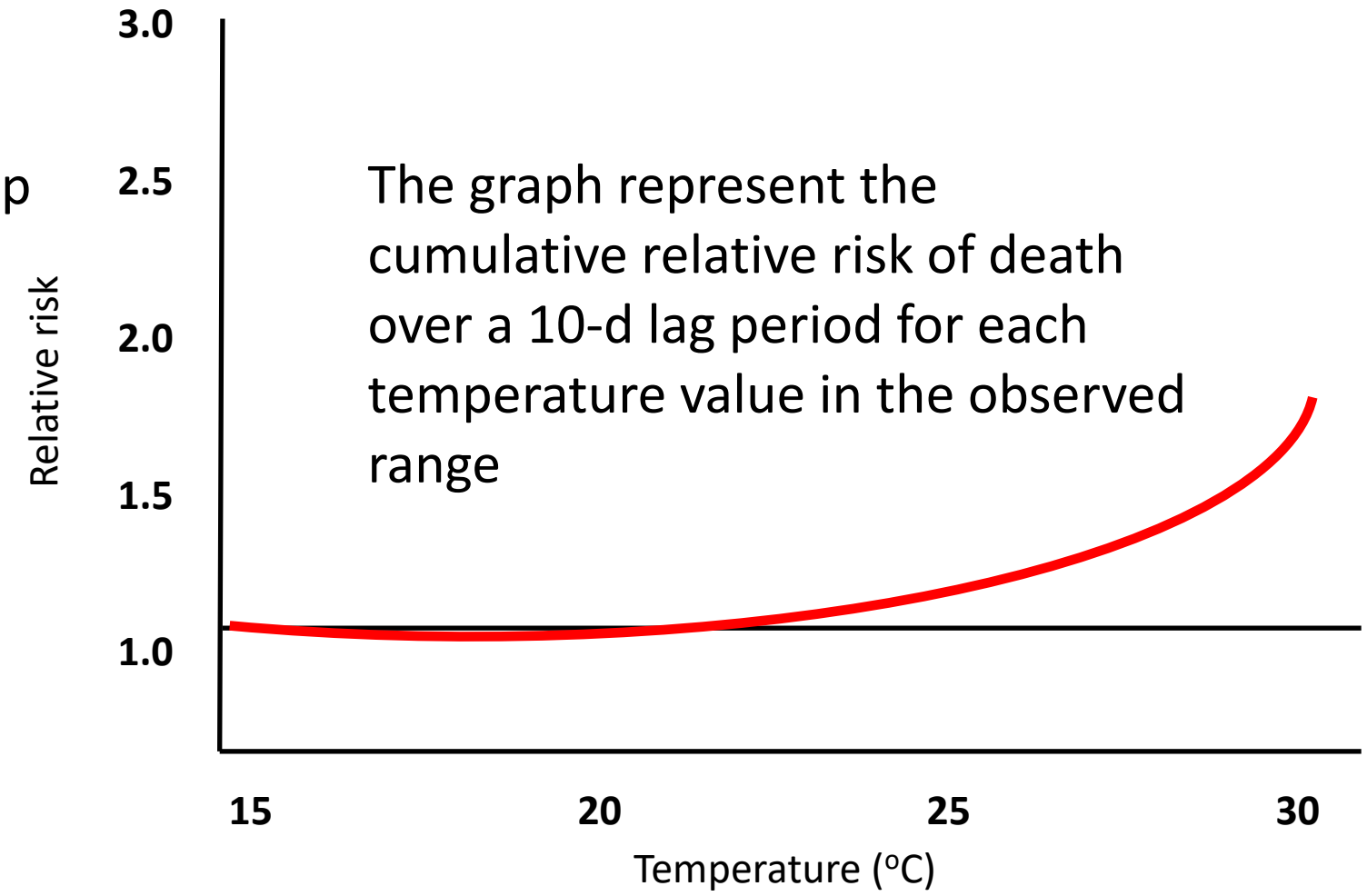


Schematic based on Vicedo-Cabrera et al. (2021). *Nature Climate Change* (11): 492–500

Case study

Location-specific temperature-mortality relationships

Step 1.
Exposure-response relationship of the relative risks for a warm-season temperature, versus the optimum temperature, corresponding to the temperature of minimum mortality



The graph represent the cumulative relative risk of death over a 10-d lag period for each temperature value in the observed range

Schematic based on Vicedo-Cabrera et al. (2021). *Nature Climate Change* (11): 492–500

Attributing heat-related deaths to climate change

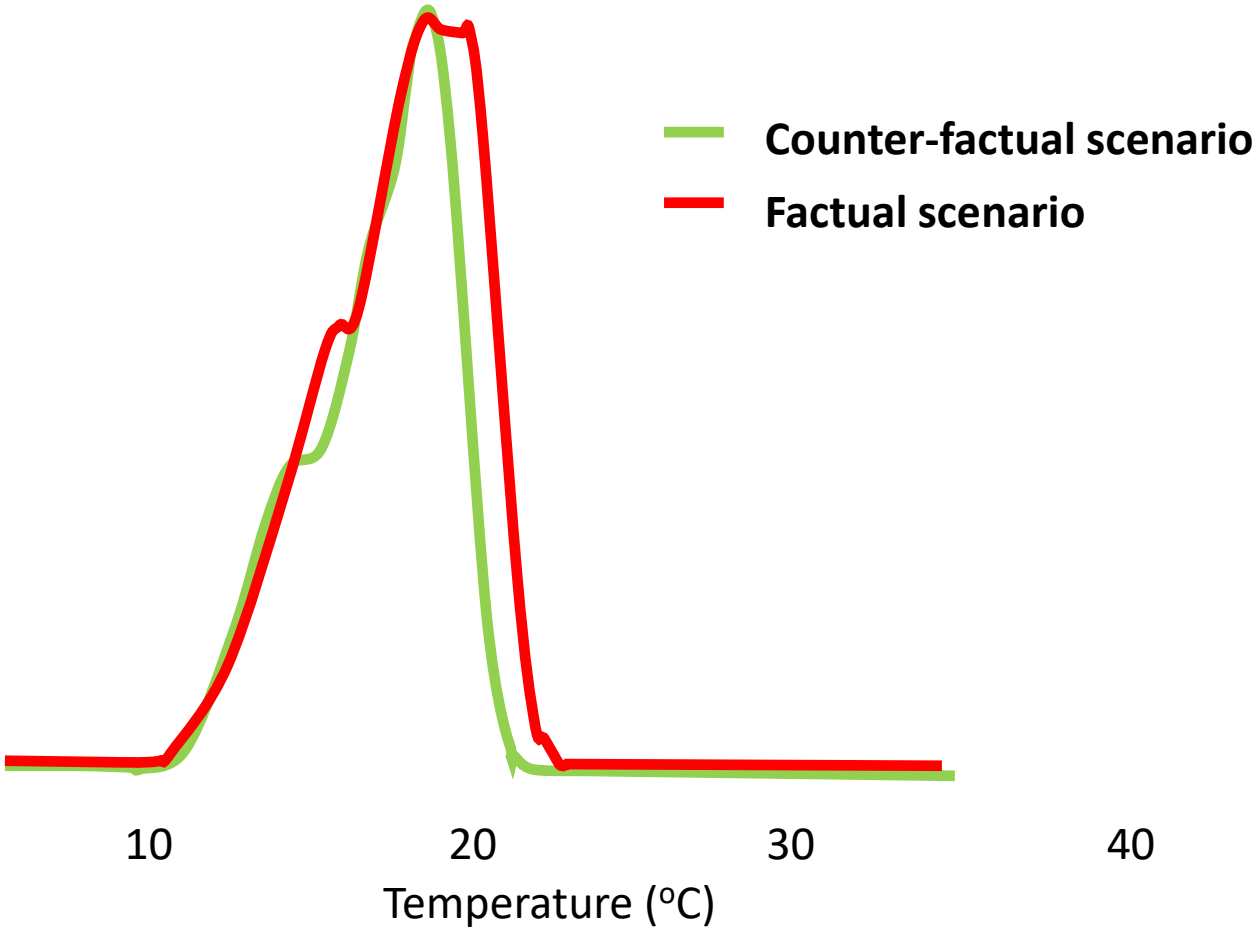
- In the second step, the estimated exposure-response function is used to compute the heat-related mortality burden for a baseline, under **two scenarios**:
  - A **factual** scenario consisting of simulations of historical climate (all climate forcings)
  - A **counterfactual** scenario where climate simulations are driven by natural forcings only, thus approximating the climate that would have occurred in a world without human-induced or anthropogenic climate change.

A. M. Vicedo-Cabrera et al., (2021). *Nature Climate Change* (11): 492–500

Attributing heat-related deaths to climate change

- Location-averaged warm-season temperature distributions in each scenario

- The burden attributable to recent human-induced climate change is defined as the difference in heat-related mortality during the warm season between the factual and counter-factual scenarios



Schematic based on Vicedo-Cabrera et al. (2021). *Nature Climate Change* (11): 492–500

Attributing heat-related deaths to climate change

- Heat-related mortality fractions are estimated as the number of deaths attributed to heat (days above the optimum) divided by the total number of deaths during the warm season in each location.
- The level of uncertainty of the impact estimates is expressed in terms of 95% CI, which accounts for both:
  - the statistical uncertainty when estimating the exposure-response function and
  - the variability in the temperature series across model-specific simulations

A. M. Vicedo-Cabrera et al., (2021). *Nature Climate Change* (11): 492–500

Attributing heat-related deaths to climate change

- The difference between the scenarios:
  - with anthropogenic climate change and
  - without anthropogenic climate change

is interpreted as the proportion of total deaths during the warm season attributable to human-induced climate change.

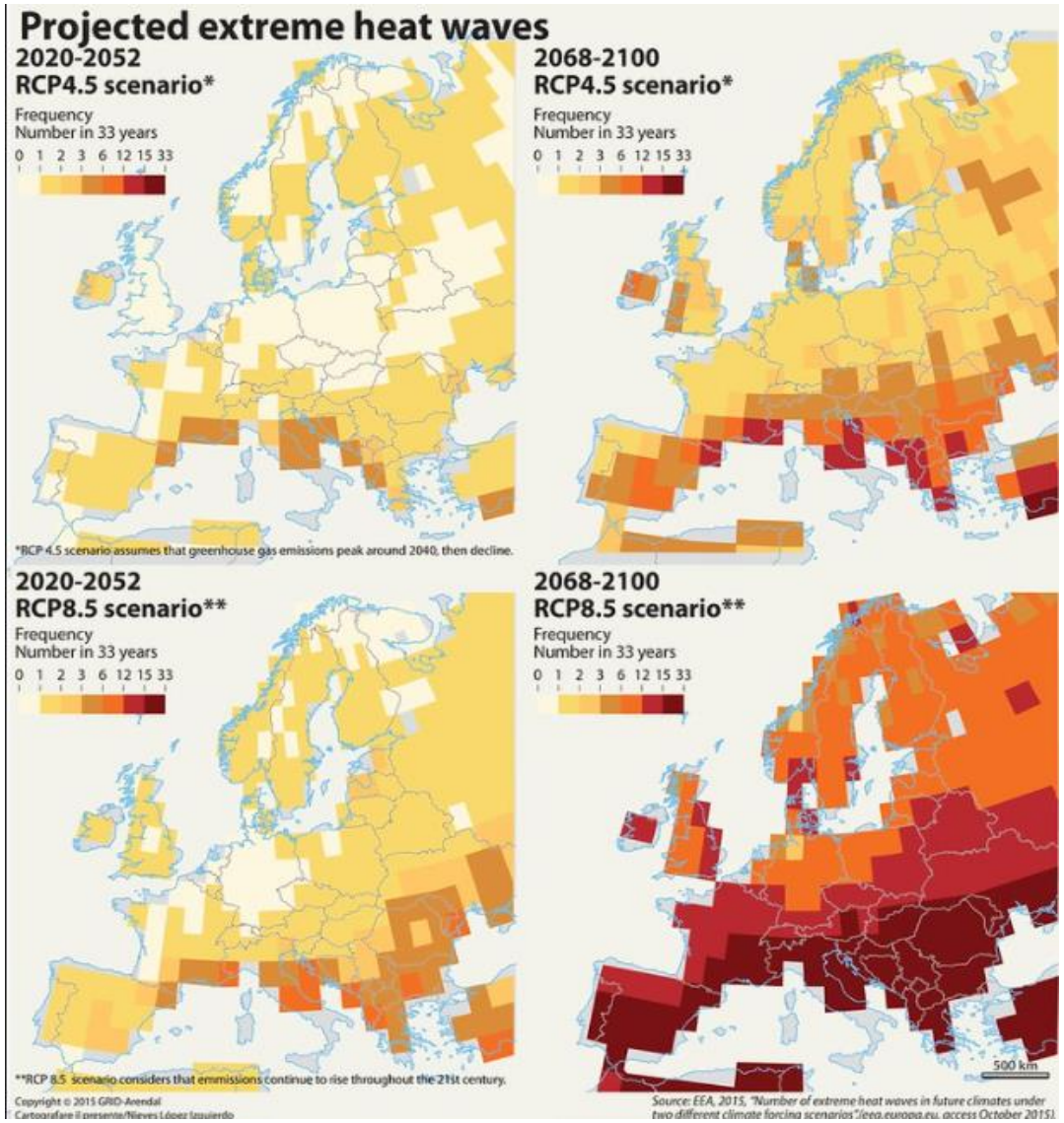
A. M. Vicedo-Cabrera et al., (2021). *Nature Climate Change* (11): 492–500

Case study

Attributing heatwave-related deaths to climate change

- Attributing a proportion of deaths during a heatwave to climate change requires multiple steps, taking into account that natural climate variability and anthropogenic climate change jointly contribute to the occurrence and intensity of a heatwave

- Model simulations can compare the probability of an event occurring with and without anthropogenic climate change



Source: GRID Arendal/Nieves Izquierdo

Extreme event attribution: process

1. The trigger: which studies to perform
2. The event definition: which aspect of the extreme event were most relevant
3. Observational trend analysis: how rare was it and how has that changed
4. Climate model evaluation: which models can represent the extreme
5. Climate model analysis: what part of the change is due to climate change
6. Hazard synthesis: combine the observational and model information
7. Analysis of trends in vulnerability and exposure
8. Communication of the results

https://www.worldweatherattribution.org/pathways-and-pitfalls-in-extreme-event-attribution/

Extreme event attribution: process

Analysis trigger: prioritize events that have a large impact on society, or that provoked a strong discussion in society

Event definition with meteorological or hydrological quantities, like temperature, wet bulb temperature, rainfall, river discharge

Observational trend analysis: how rare the event is in the current climate, and how much this has changed over the period with observations.

https://www.worldweatherattribution.org/pathways-and-pitfalls-in-extreme-event-attribution/

Extreme event attribution: process

Climate model evaluation: how often extreme events occur in the computed weather in the climate model.

Climate model analysis: how much more likely or intense the extreme event has become due to anthropogenic emissions of greenhouse gases and aerosols.

Hazard synthesis: how has the probability and intensity of the physical extreme event has changed?

https://www.worldweatherattribution.org/pathways-and-pitfalls-in-extreme-event-attribution/

Extreme event attribution: process

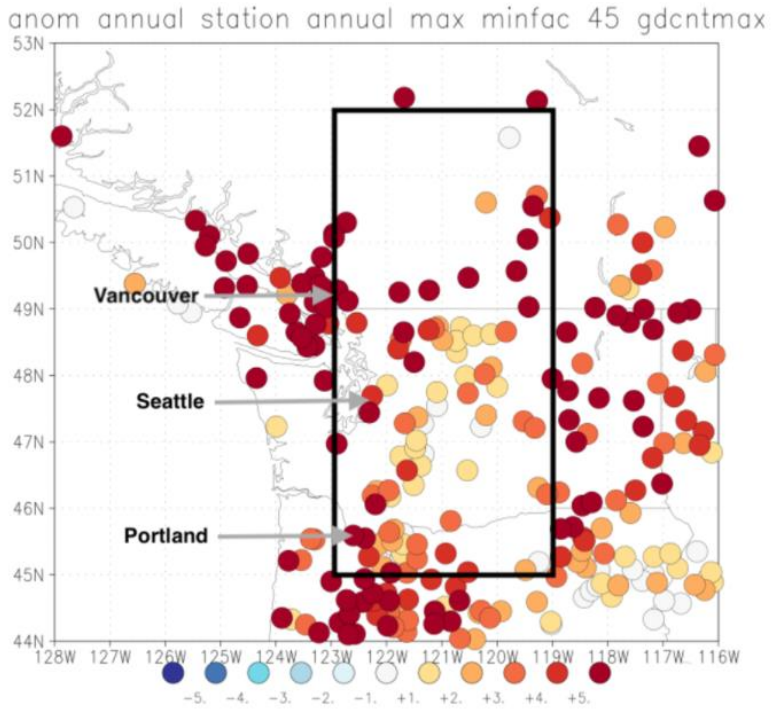
Vulnerability and exposure combine with the changes in the physical extremes that were computed in the previous step

Communication stratified according to their level of expertise: scientists, policy-makers, emergency management agencies, media outlets and the general public.

https://www.worldweatherattribution.org/pathways-and-pitfalls-in-extreme-event-attribution/

NW American extreme heat virtually impossible without human-caused climate change: June 2021

- Based on observations and modelling, the occurrence of the June 2021 heatwave with maximum daily temperatures was virtually impossible without human-caused climate change.
- The observed temperatures were so extreme that they lie far outside the range of historically observed temperatures. This makes it hard to confidently quantify how rare the event was. In the most realistic statistical analysis, the event is estimated to be about a 1-in-1000-year event in today's climate.



Fiure 1: Station data anomalies of the 2021 event relative to the mean of the highest daily maximum temperature of the year in the time series. Note that some stations do not have data up to the peak of the heatwave yet and hence underestimate the event. Negative values certainly do not include the heatwave and have therefore been deleted. The black box shows the study region. Source: GHCN-D downloaded 4 July 2021.

https://www.worldweatherattribution.org/western-north-american-extreme-heat-virtually-impossible-without-human-caused-climate-change/

Resources

Abatzoglou, J. T. and A. P. Williams, 2016: Impact of anthropogenic climate change on wildfire across western US forests. *PNAS*, 113(42), 11770-11775

Arbuthnott, K., Hajat, S., Heaviside, C., & Vardoulakis, S. (2020). Years of life lost and mortality due to heat and cold in the three largest English cities. *Environment International*, 144, 105966.

Astrom, D. O., Ebi, K. L., Vicedo-Cabrera, A. M., & Gasparrini, A. (2018). Investigating changes in mortality attributable to heat and cold in Stockholm, Sweden. *International Journal of Biometeorology*, 62(9), 1777-1780.

Bhaskaran K, et al. Time series regression studies in environmental epidemiology. *Int J Epidemiol*. 2013. PMID: 23760528

Diaz, J., Carmona, R., Miron, I. J., Luna, M. Y., & Linares, C. (2019). Time trends in the impact attributable to cold days in Spain: Incidence of local factors. *Science of the Total Environment*, 655, 305-312.

Diffenbaugh, N. S. and M. Burke, 2019: Global warming has increased global economic inequality. *PNAS*, 116(20), 9808-9813.

Diffenbaugh, N. S. et al., 2017: Quantifying the influence of global warming on unprecedented extreme climate events. *PNAS*, 114(19), 4881-4886.

Ebi KL, Ogden NH, Semenza JC, Woodward A. Detecting and Attributing Health Burdens to Climate Change. *Environ Health Perspect*. 2017 Aug 7;125(8):085004. doi: 10.1289/EHP1509.

Ebi, K., Astrom, C., Boyer, C., Harrington, L., Hess, J., Honda, Y., . . . Otto, F. (2020). Using Detection And Attribution To Quantify How Climate Change Is Affecting Health. *Health Affairs*, 39(12), 2168-2174.

Hajat, S. (2017). Health effects of milder winters: A review of evidence from the United Kingdom. *Environmental Health*, 16, 109.

Hanigan, I. C., Dear, K. B. G., & Woodward, A. (2021). Increased ratio of summer to winter deaths due to climate warming in Australia, 1968-2018. *Australian and New Zealand Journal of Public Health*. doi: 10.1111/1753-6405.13107.

Lee, S., Lee, H., Myung, W., Kim, J., & Kim, H. (2018). Mental disease-related emergency admissions attributable to hot temperatures. *Science of the Total Environment* (616), 688-694.

Mengel, M., S. Treu, S. Lange and K. Frieler, 2020: ATTRICI 1.0–counterfactual climate for impact attribution. *Geoscientific Model Development Discussions*, 1-26.

Mitchell, D., Heaviside, C., Vardoulakis, S., Huntingford, C., Masato, G., Guillod, B., . . . Allen, M. (2016). Attributing human mortality during extreme heat waves to anthropogenic climate change. *Environmental Research Letters*, 11.

Stone D, Auffhammer M, Carey M, Hansen G, Huggel C, Cramer W, et al.2013. *Climatic Change* 121(2):381–395

Vicedo-Cabrera A. M. et al., (2021). *Nature Climate Change* (11): 492–500

Section 3.2:

Geographic patterns of health risks: overview of spatial distributions and methods to describe clustering and hotspots

Learning objective: Understand the methods available to associate climate/weather with disease risk based on geographic variations in climate/weather and disease hazard occurrence. Understand best practices for their use and the need to account for possible sources of bias in data, such as spatial clustering.

Case studies:

- Predicting *Aedes aegypti* presence in Ecuador;
- A temperature-driven trait-based model for *Aedes* spp. disease transmission,
- Temperature-driven model for current and future potential global disease risk expansion.

Further reading:

- R package "spatstat" for spatial statistics website: http://spatstat.org/
- Using R as a GIS – a straightforward website with simple pieces of code: http://pakillo.github.io/R-GIS-tutorial/#plot
- Spatial Data Science with R – open access, open source tutorial set by Robert Hijmans https://rspatial.org/
- SatScan free software for conducting space-time scan statistics (as seen in the Domnica case study) https://www.satscan.org/
- R package 'sdm' contains multiple types of ENMs to run on datasets – the website biogeoinformatics.org has a nice tutorial (with data) for walking through these and comparing metrics of goodness/fit
- To run distribution models such as BIOCLIM, and visualise them, Diva-GIS, an open source software, has a nice platform and free spatial datasets to work through (Diva-gis.org).
- The model output gridded data for our global *Aedes* distribution models, and the case study input data for ENMs (https://dataverse.harvard.edu/dataverse/Aedesmaps)

3.2 Geographic patterns of health risks: overview of spatial distributions and methods to describe clustering and hotspots

Dr Catherine A. Lippi and Dr Sadie J. Ryan

University of Florida

Learning objectives

- Define clustering, randomness, and uniformity; gain familiarity with several commonly used techniques to statistically describe clustering in the context of climate-health; be aware of spatial and temporal scale issues and data size limitations.
- Understand what landscape regressions entail, what ecological niche models represent, in the context of diseases

Specific objectives

General Objective:

- Understand basic geographic principles and techniques as they apply to public health problems

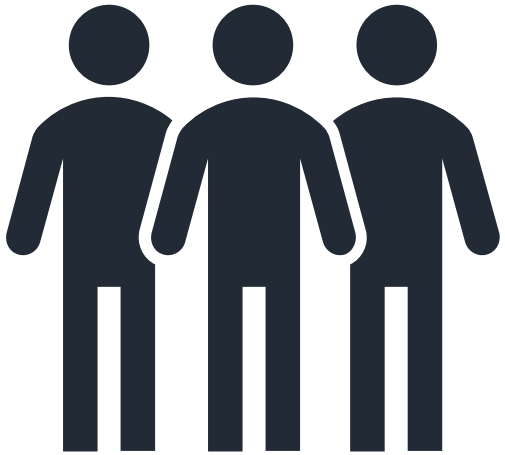
Specific Objectives:

- Be aware of spatial and temporal scale issues, and data size limitations.
- Define clustering, dispersion, and spatial randomness
- Gain familiarity with several commonly used techniques to statistically describe geographic patterns and clustering in health data
- Understand what landscape regressions entail, and what ecological niche models represent, in the context of diseases

WHAT IS MEDICAL GEOGRAPHY?

Medicine: investigate causes of diseases in individuals

Epidemiology: investigate health-related issues and risk factors in human populations

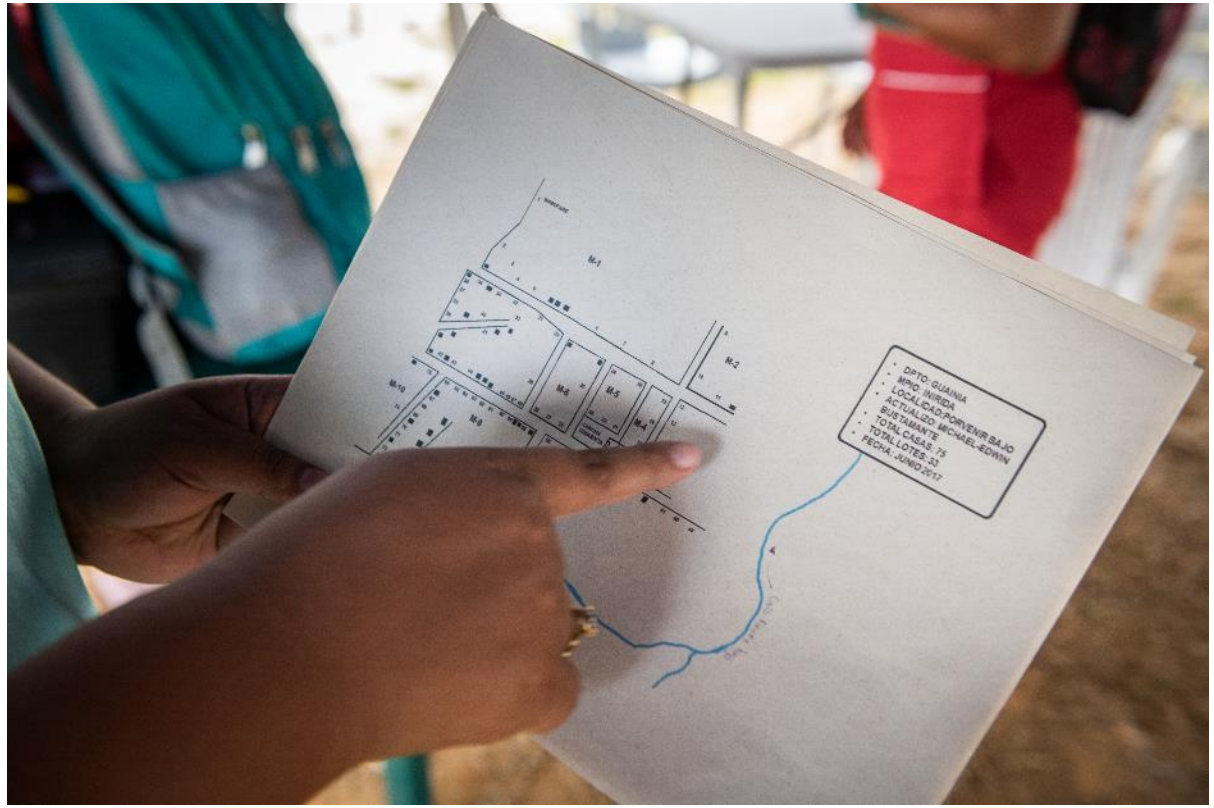


Medical Geography: investigate health-related issues and risk factors in populations, in a spatial context, using concepts and methods taken from Geography

LOCATIONS AND HEALTH OUTCOMES

- **Describe** unique properties of an area, such as land cover, climate, hydrology, socioeconomic factors, culture, and health metrics

- **Compare:**
  - Compare locations and explore underlying factors associated with health outcomes
  - Compare changes over time, or movements across locations through time



© WHO/Nadège Mazars

SCALE

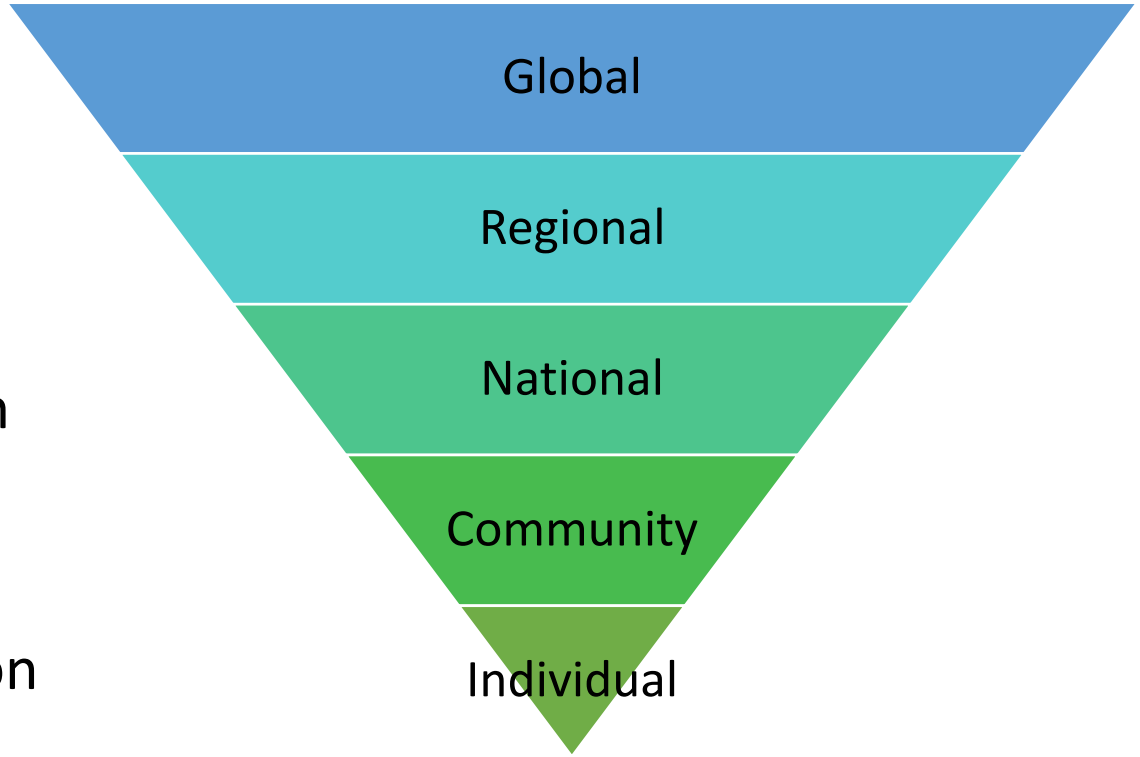
Scale is important when defining disease outbreaks

(e.g. endemic, epidemic, pandemic)

Temporal scale (i.e. time period) is also important

Spatial and temporal scales must be defined when conducting analyses

- Multiple data sources must match in resolution
- Changes in spatial resolution/units of aggregation change findings and the appropriate application



SCALE AND DATA

The scale of data preparation and analysis must be appropriate for the study question and defined at the start of an investigation

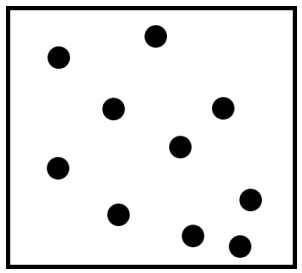
Mixing scales of data input in analyses can lead to incorrect results

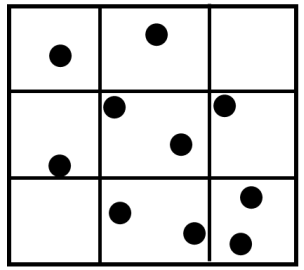
Ecological fallacy: applying conclusions or observations from one scale to another

e.g. assuming that national health trends apply to individuals

COMMON GEOREFERENCED DATA FORMATS

- Point data: discrete locations where an object is located, or an event occurred
  - E.g, individual cases plotted with latitude/longitude coordinates



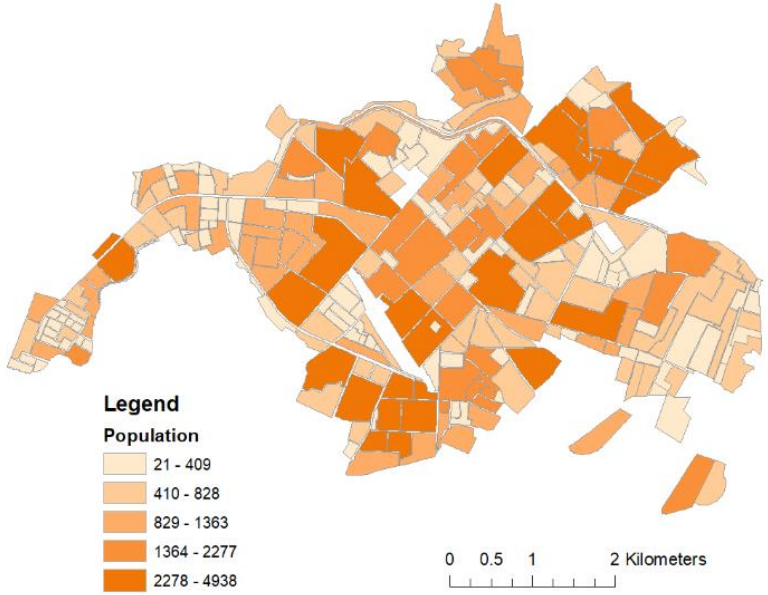


1	1	0
1	2	1
0	2	2

- Raster data: continuous surface of cells or pixels
  - Common when working with environmental datasets; used to record data such as land cover type, elevation, temperature, precipitation, etc.

- User-specified units of aggregation: overlaying a grid on the study area and aggregating points to the grid

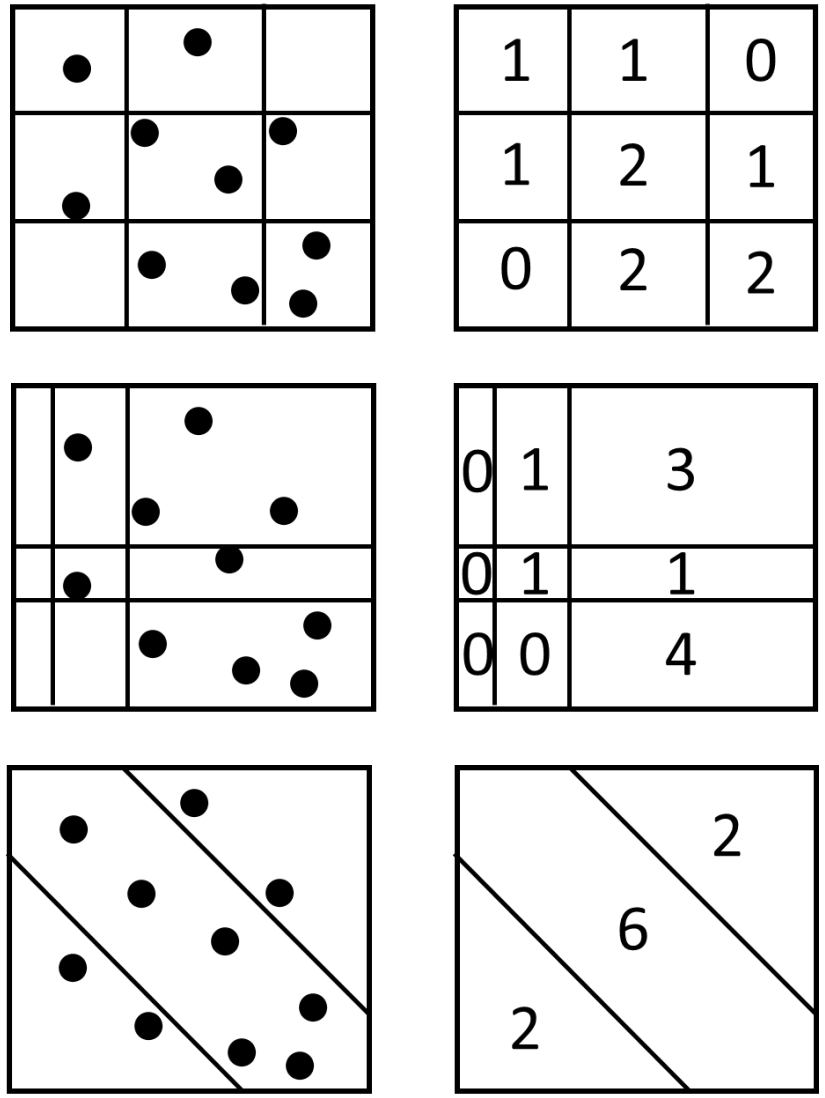
- Administrative Units/Census Blocks/Zip Codes (e.g. socioeconomic data, disease surveillance data)



Lippi et al. 2020, Int J Health Geogr

MAUP

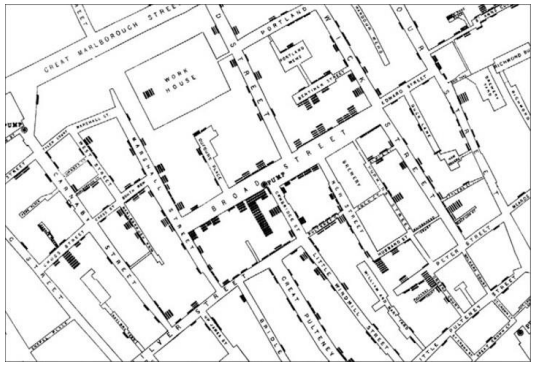
- MAUP = Modifiable Areal Unit Problem
  - Important to consider when designing and conducting spatial analyses
- MAUP arises because changes in how data are aggregated can alter observed patterns
- Size and shape of aggregating units influence how data are presented
- When aggregated units are used as input for statistical analyses, this can influence results
- Good practice: run spatial statistics across multiple scales or aggregating units when possible to assess the effects of aggregation on observable patterns
- At minimum, have a rationale for use of units, such as aggregating case counts to operational health districts or census units where policies and interventions are enacted

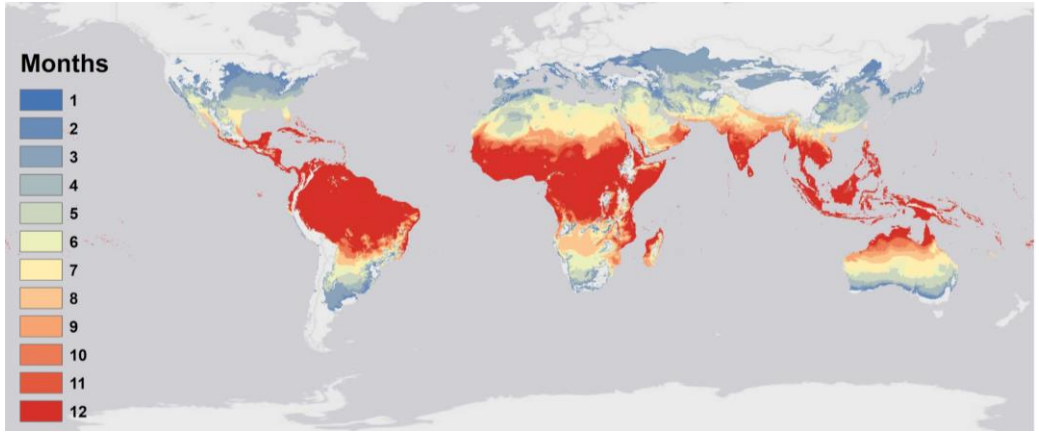


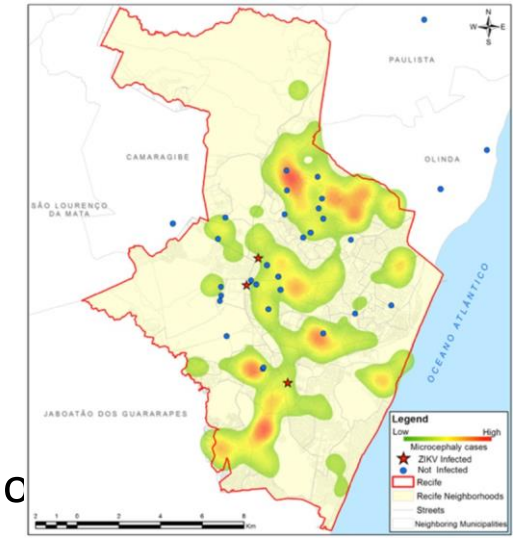
This example shows that the same underlying patterns of points can appear to have very different spatial spreads, depending on how we aggregate them.

MAPPING AND DESCRIPTIVE STATISTICS

- Dot maps of cases
  - Plots locations of events
- Choropleth mapping
  - Uses colour and shading to indicate quantities
- Mean Centre and Spatial Standard Deviation
  - Describes spatial density and overall direction of dispersion
- Kernel Density Estimation (KDE)
  - Uses the spatial decay function to plot concentratic of events continuously across the landscape







ABOVE LEFT: Dot map - John Snow's display of cholera cases during the 1854 outbreak in London's Broad St. area

ABOVE RIGHT: Choropleth map - Modelled number of suitable months for *Aedes aegypti* presence (red areas indicate longer periods of suitability), taken from Ryan et al. 2019

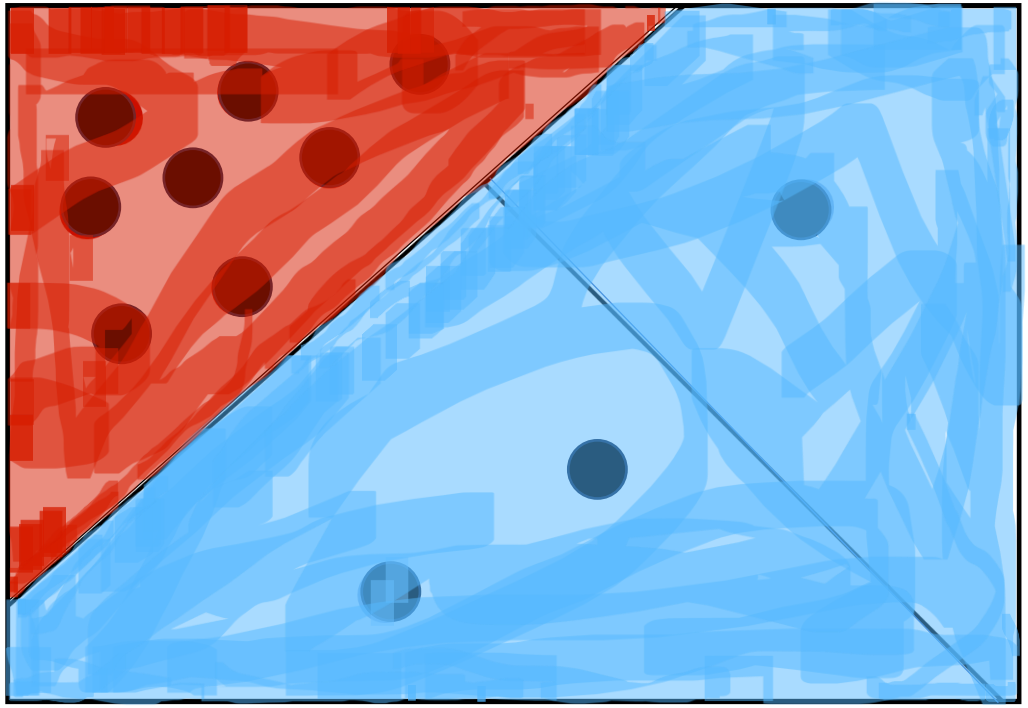
LEFT: KDE of Zika-associated microcephaly cases, taken from Guedes et al. 2017

SPATIAL CLUSTERING AND DISPERSION

The terms “clustered” and “dispersed” are used to describe the spatial distribution of occurrences (e.g. disease cases) in the study area

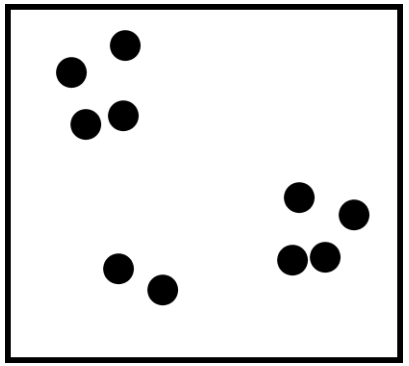
In the context of public health, clusters of high disease activity may be referred to as “hotspots” (red)

In contrast, areas of dispersion, or low disease activity, may be called “cold spots” (blue)

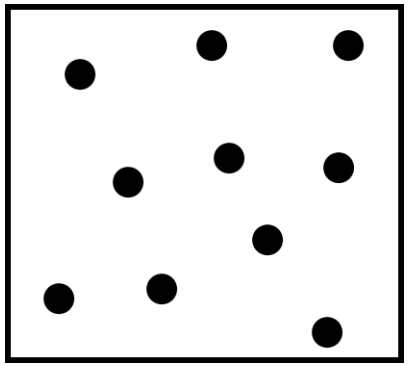


SPATIAL CLUSTERING AND DISPERSION

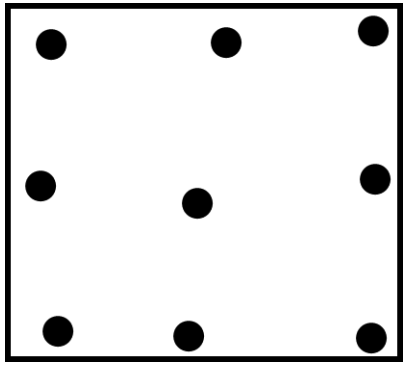
Clustered: closer together than expected



Random: no spatial pattern



Dispersed: further away than expected



SPATIAL AUTOCORRELATION

- Detecting Spatial Autocorrelation in data allows us to measure how similar or dissimilar nearby objects or events are

- Tobler's First Law of Geography – "Everything is related to everything else, but near things are more related than distant things"

- If occurrences or processes in neighbouring locations are similar, spatial autocorrelation is positive

- If neighbouring locations are not similar, then spatial autocorrelation is negative

***THESE PATTERNS ARE DEPENDENT UPON SCALE AND DATA AGGREGATION**

TWO TYPES OF CLUSTERING

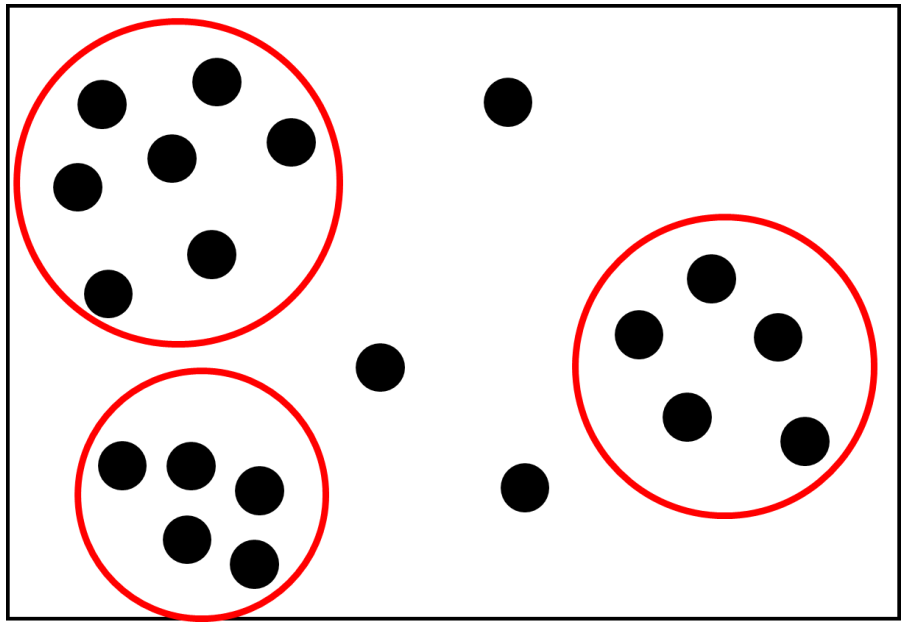
Global Clustering:

Is there clustering in the study area?

Note that global tests for spatial autocorrelation detect presence of clustering or dispersion in the study area, but do not tell us WHERE those clusters occur

Local Clustering:

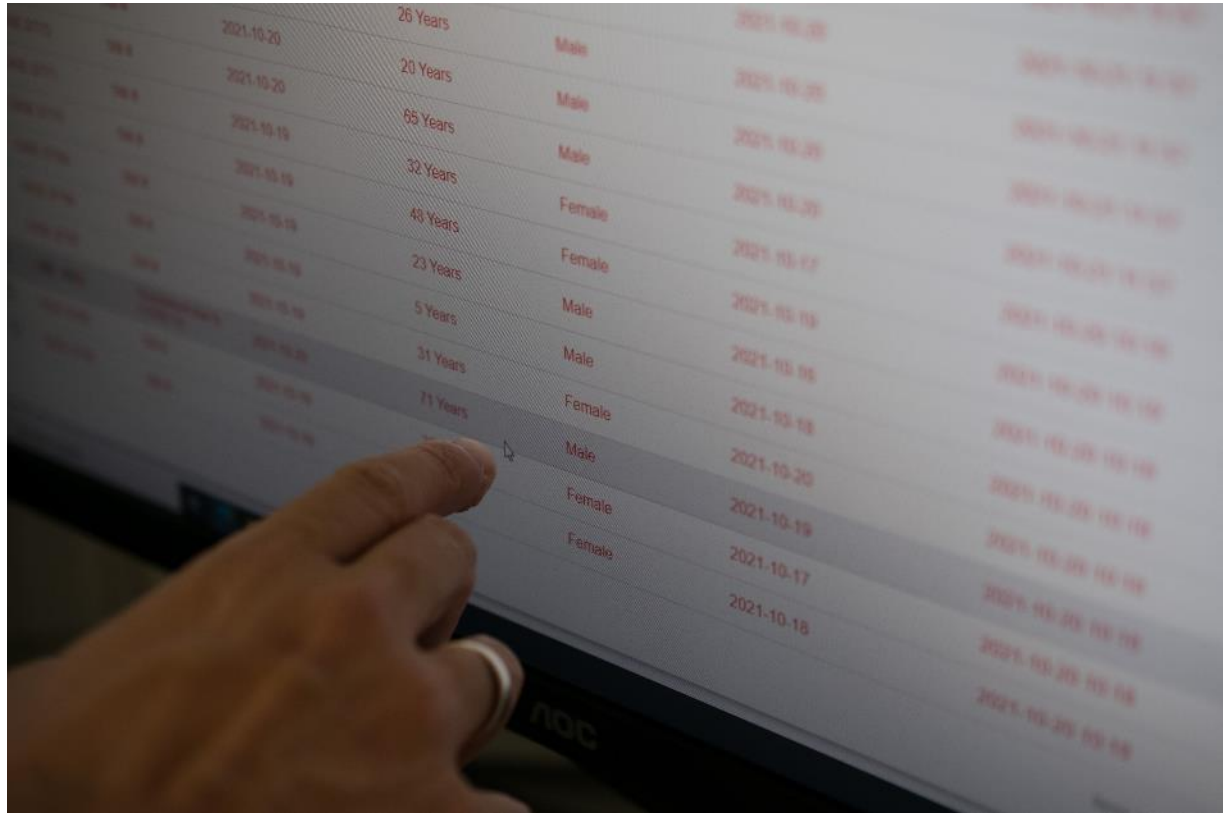
WHERE are clusters located within the defined study area?



STATISTICAL METHODS

- Unlike descriptive statistics, these tests allow the user to measure the degree of clustering (or dispersion), and assess if it is statistically significant (i.e. differences are greater than what we would expect if occurrences were truly random)

- Provides a framework for experimental design and hypothesis testing



© WHO/Joanna Demarco

AVERAGE NEAREST NEIGHBOR INDEX (ANNI)

- Ratio comparing the distance between observed locations of points (e.g. cases) and expected distances if points were spatially random
- ANNI allows for testing of global clustering

- Values less than 1 indicate clustering
- Values near or equal to 1 indicate a random distribution of points
- Values greater than 1 indicate dispersion

$$ANN = \frac{\overline{D}o}{\overline{D}_E}$$

$$\overline{D}o = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$\overline{D}_E = \frac{0.5}{\sqrt{n/A}}$$

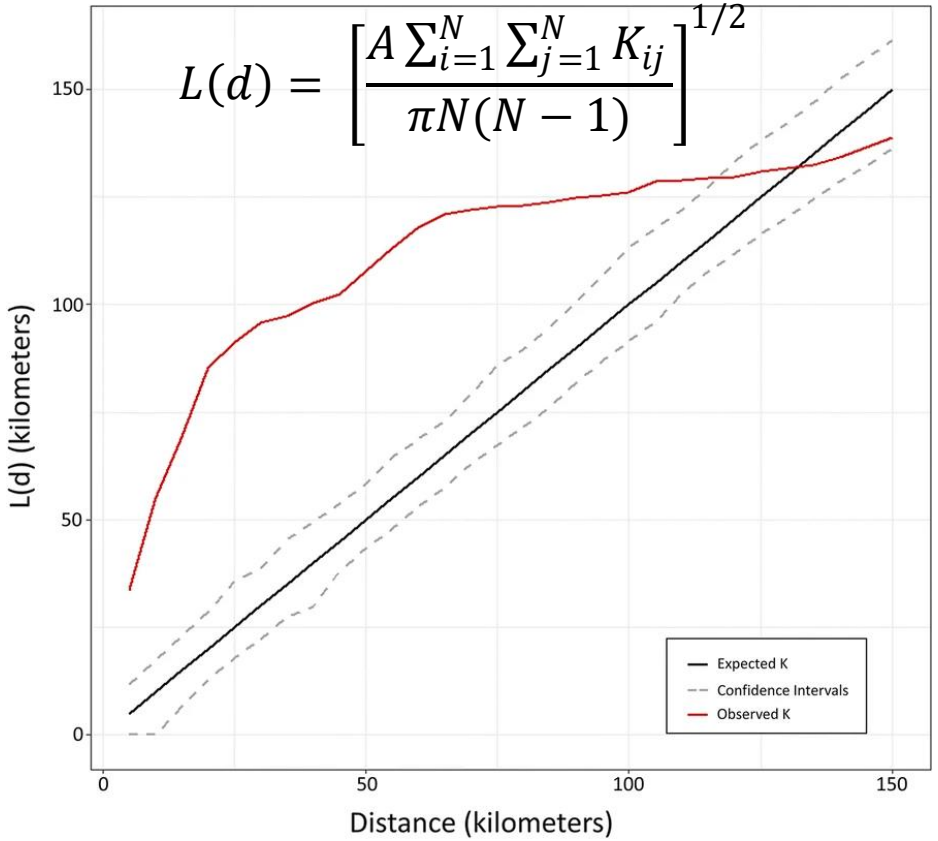
$\overline{D}o$ = **Observed mean distance between each feature and its nearest neighbor**

$\overline{D}_E$ = **Expected mean distance for features under the assumption of complete spatial randomness**

RIPLEY'S K

- Allows us to detect the PRESENCE and spatial SCALE of statistically significant clustering
- Allows us to assess the observed and expected number of cases at each location, compared to the number of neighbouring locations within a specified radius
- Bandwidth defines the neighbourhood size around each point; smaller bandwidths may capture local clustering, larger bandwidths may capture more widespread patterns
- Statistics must be run for many distances (i.e. bandwidths) to detect the scale of clustering
- Values that exceed confidence intervals for expected values indicate statistically significant clustering or dispersion

$$L(d) = \left[\frac{A\sum_{i=1}^{N}\sum_{j=1}^{N}K_{ij}}{\pi N(N-1)}\right]^{1/2}$$



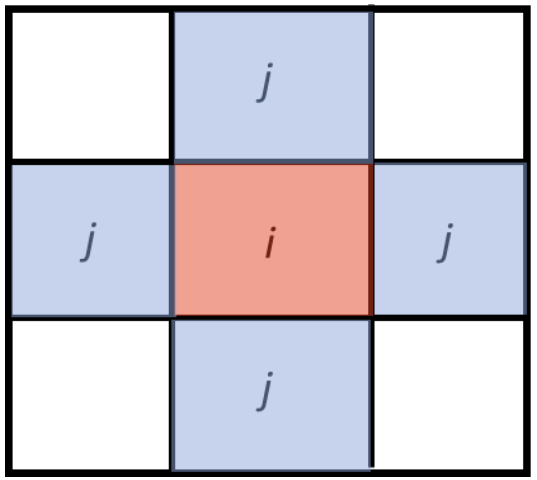
Results of a Ripley's K analysis to detect clustering of insecticide resistance in mosquitoes, taken from Mundis et al. 2020; the black line shows our expectation of spatial randomness and the red line shows the value of our test statistic, indicating significant clustering along different spatial scales. Significant dispersion would be indicated at distances where the statistic is lower than the black line of expected values.

MORAN'S I

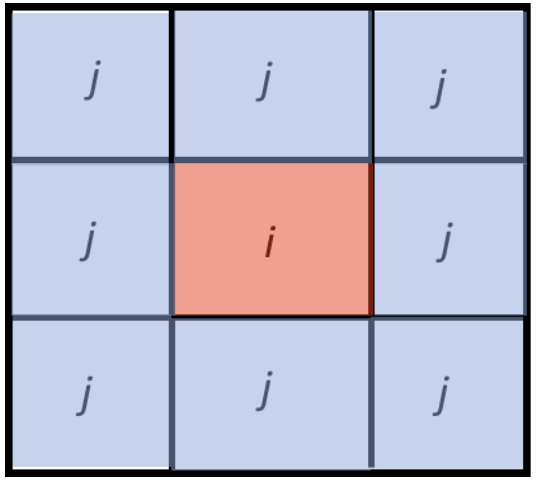
- Another method of assessing the study area for global clustering

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \; x \; \frac{\sum_i \sum_j w_{ij} \; (xi - \bar{x})(xj - \bar{x})}{\sum_i (xi - \bar{x})^2}$$

- Values (e.g. case counts) at each location (i) and neighbouring locations (j) are compared to values for the rest of the study area
- Neighbours (j) are defined by a spatial weights matrix w_{ij})
- Weights matrix determines which locations are considered neighbours and can be specified in a number of ways (e.g. distance, inverse distance, contiguity)
- Values greater than 0 indicate clustering
- Values lower than 0 indicate dispersion



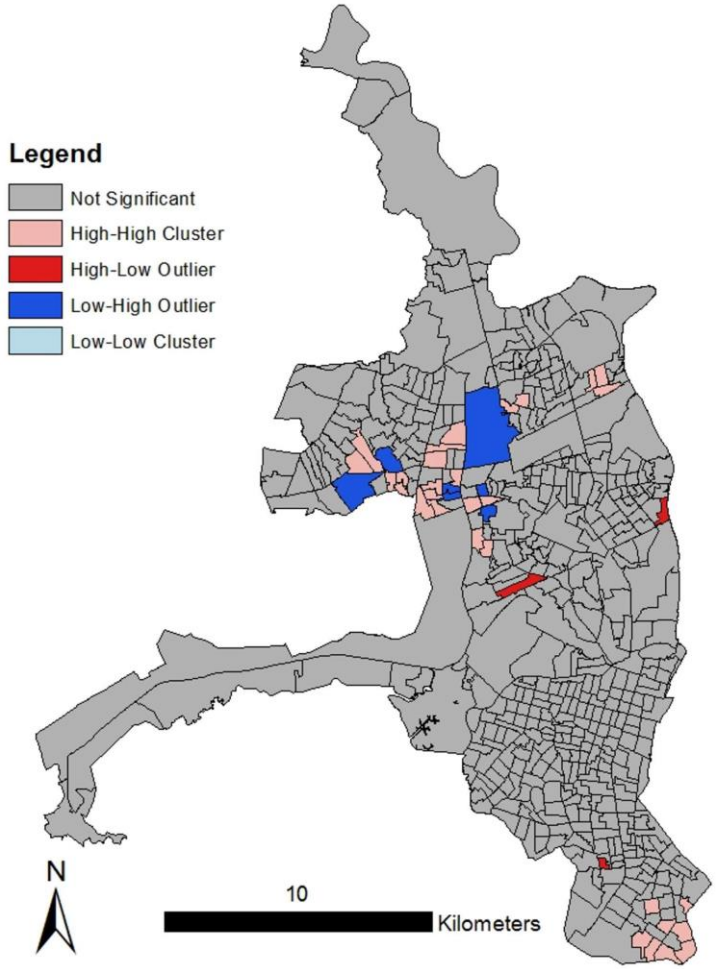
Rook contiguity (more conservative)



Queen contiguity (less conservative)

LOCAL INDICATORS OF SPATIAL ASSOCIATION

- LISA – tells you WHERE clustering occurs in the study area
- Also useful for identifying locations of spatial outliers (e.g., areas of high disease activity adjacent to low activity areas)
- But susceptible to high variance due to small underlying populations, as we often see in disease rates
- Anselin Local Moran's I is a commonly used test for local spatial autocorrelation



Anselin's Local Moran's I analysis for the 2012 Guayaquil outbreak. Cases of dengue were significantly clustered in the North Central and Southern areas of the city.
Lippi et al. 2018

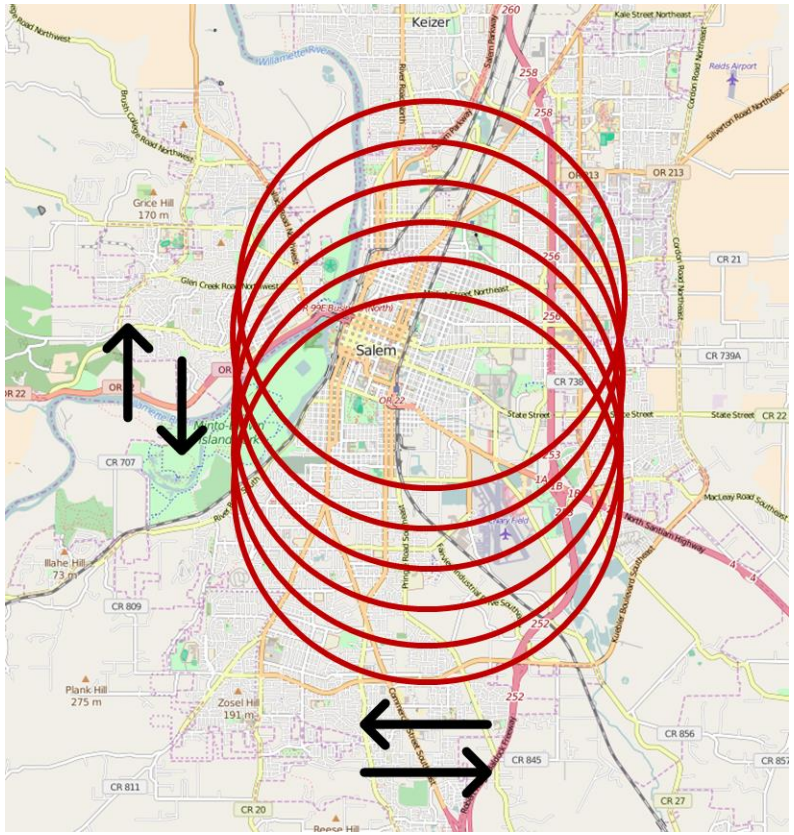
CONSIDERATIONS AND LIMITATIONS

- Be aware of the MAUP
- Defining your neighbours is important
  - User-specified definitions of neighbouring areas can influence results and observed “hotspot” patterns
- Edge effects can introduce bias near the boundaries of the study area
- Small numbers problem

***Statistical methods need a rationale, and tests are typically performed many times to assess the effects of scale and data preparation**

SPATIAL SCAN STATISTICS

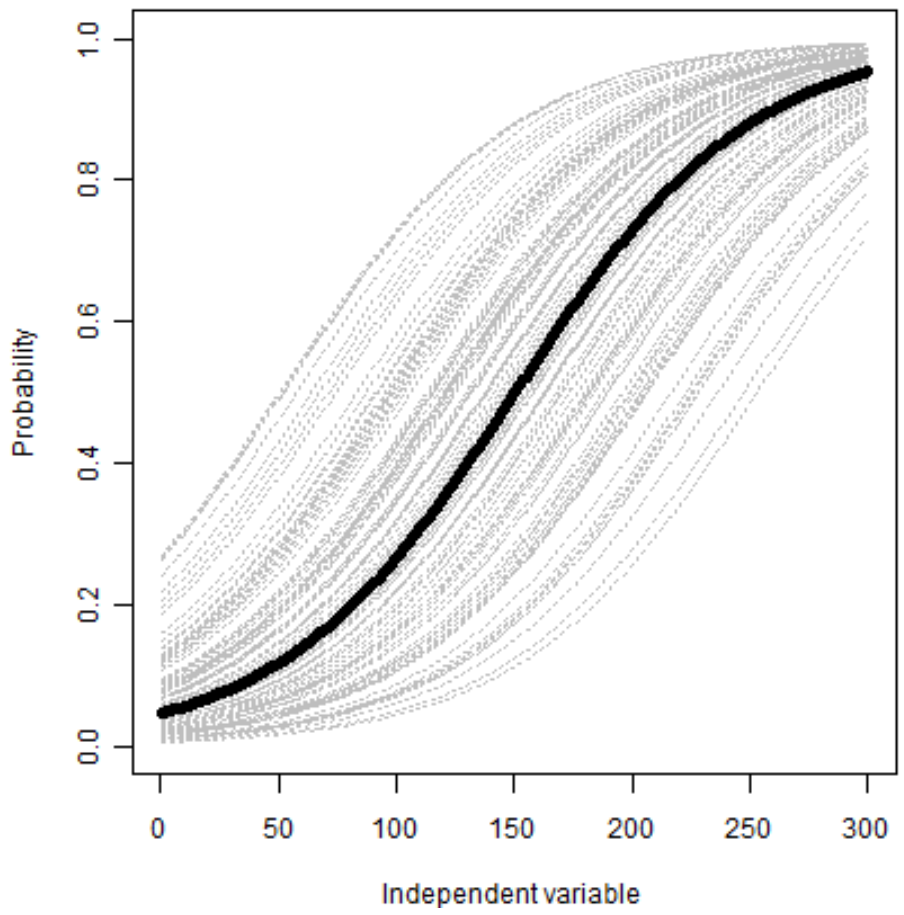
- Spatial scan statistics for cluster detection
- Circular windows of variable bandwidth move across the study area, comparing disease incidence inside the window to the area outside of the window
- Options for spatiotemporal analyses allow for cylindrical windows
- Tell us WHERE and WHEN clustering or dispersion of disease events occur



- SaTScan, developed by Martin Kulldorff, is an open-source software to perform spatial scan statistics

LOGISTIC REGRESSION

- Common statistical model used to model binary outcomes (e.g. cases and non-cases)
- Frequently used in public health to predict outcomes and estimate coefficients for factors of interest
- Typically conducted without spatial information
- Spatial autocorrelation violates the statistical assumptions of linear regression
- Failing to account for spatial effects in the model can lead to skewed or incorrect results



SPATIAL EXTENSION OF LR

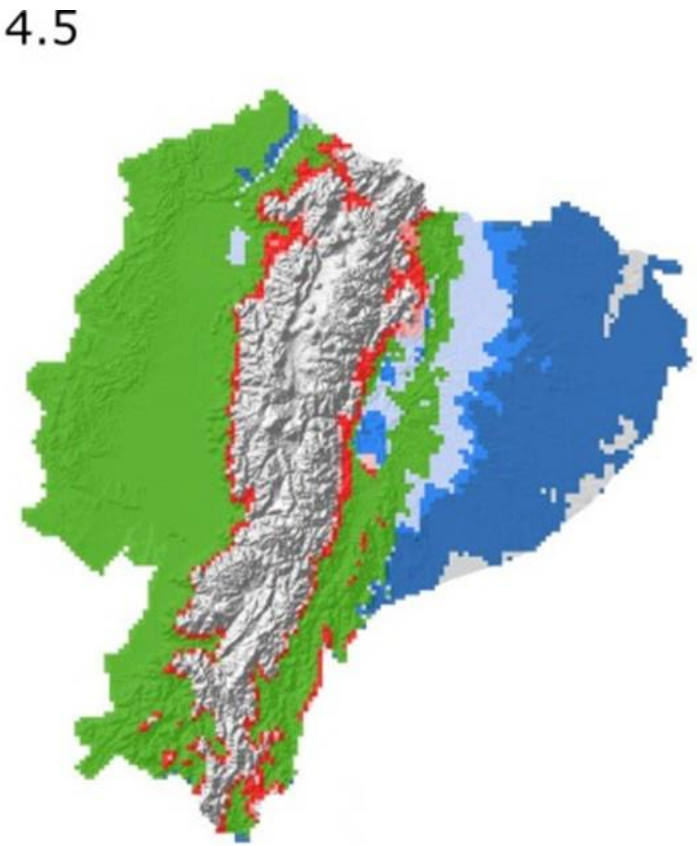
- If spatial autocorrelation is detected in LR residuals, there are methods to control for these effects
  - Spatial error term and spatial lag terms -> Model terms that can be incorporated to account for model error arising from spatial dependence
  - Geographically Weighted Regression (GWR) -> explicitly models spatial relationships

- If continuous spatial variables are used as input, model output can be projected to the landscape

ECOLOGICAL NICHE MODELS

Species Distribution Model (SDM) is often used interchangeably

- Use known locations of occurrence and underlying environmental conditions to predict geographic distribution in areas that have not been sampled
- In the context of public health, it is used to estimate geographic ranges and potential risk of exposure to pathogens, disease vectors, etc.
- Climate products for future conditions allow us to project models to different time points, estimating potential shifts in exposure and risk
- Many ENM algorithms used, including MaxEnt, GARP, BRT, RF, and more



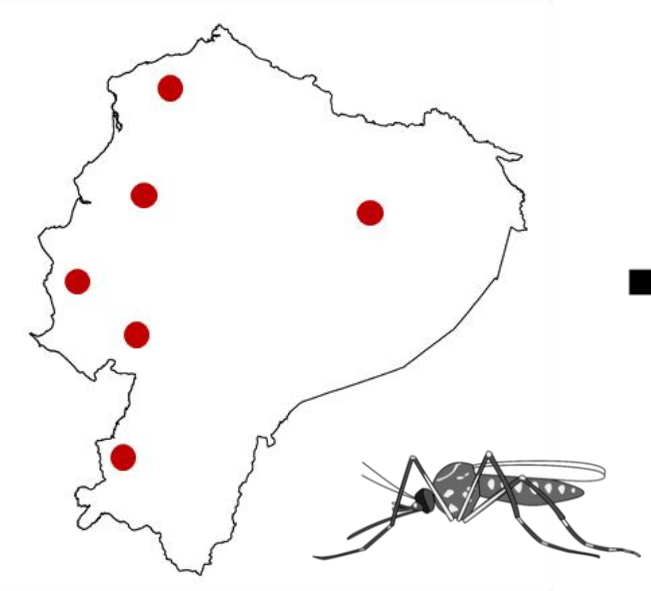
ENMs were used to estimate potential range expansion of *Aedes aegypti* in Ecuador, taken from Lippi et al. 2019. Under this scenario of climate change, projected models show potential areas of range expansion on the periphery of the Andes mountain range, indicating areas where habitat may become suitable to mosquitoes in the future.

ECOLOGICAL NICHE MODELING

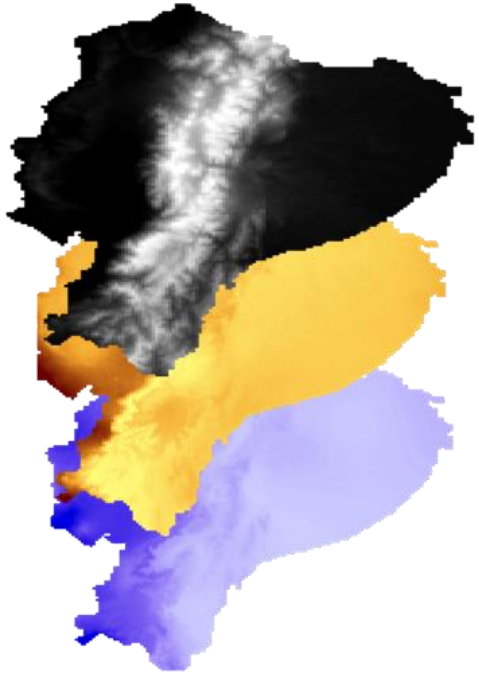
Georeferenced Species Records
Museum Collections
Public Databases
Scientific Literature
Public Health Surveillance
Field Sampling

Layers of Environmental Data
Elevation
Climate
Vegetation
Soil Conditions

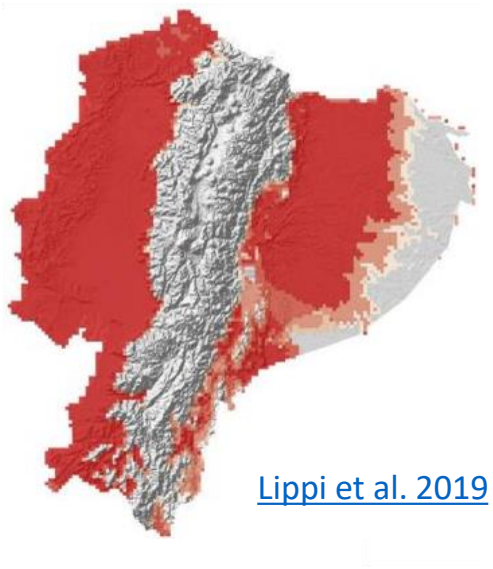
Objective
Estimate the geographic distribution for the species of interest



+



ENM Algorithm →



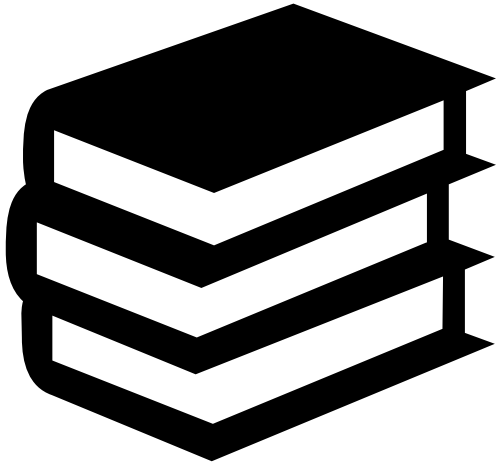
Lippi et al. 2019

BIBLIOGRAPHIC REFERENCES

Case studies

- Lippi et al., 2018. The Social and Spatial Ecology of Dengue Presence and Burden during an Outbreak in Guayaquil, Ecuador, 2012. Int. J. Env. Res. & Pub. Health. 15:4 https://www.mdpi.com/1660-4601/15/4/827/pdf (Open Access)

- Nsoesie EO, et al. 2015 Spatial and Temporal Clustering of Chikungunya Virus Transmission in Dominica. PLoS Negl Trop Dis 9(8): e0003977. https://doi.org/10.1371/journal.pntd.0003977



ADDITIONAL RESOURCES

- R package “spatstat” for spatial statistics website: http://spatstat.org/
- Using R as a GIS – a straightforward website with simple pieces of code: http://pakillo.github.io/R-GIS-tutorial/#plot
- Spatial Data Science with R – open access, open source tutorial set by Robert Hijmans https://rspatial.org/
- SatScan free software for conducting space-time scan statistics (as seen in the Domnica case study) https://www.satscan.org/

Section 3.3:

Process based models

- **Learning objective**: To gain a basic understanding of process based models

Further reading:

- Keeling MJ & Rohani P (2008). Modelling infectious diseases in humans and animals. Princeton University Press.
- Keeling MJ & Ross JV (2008). On methods for studying stochastic disease dynamics. J R Soc Interface 5:171–181. https://doi.org/10.1098/rsif.2007.1106
- Randolph HE & Barreiro LB (2020). Herd immunity: understanding COVID-19. Immunity 52:737–741.
- DiSera L, et al. (2020). The mosquito, the virus, the climate: an unforeseen Réunion in 2018. GeoHealth 4:e2020GH000253.
- CDC, mosquito life cycle and biology: https://www.cdc.gov/mosquitoes/

MODULE 3.3

Process-based models

Dr Joacim Rocklöv
Umeå University

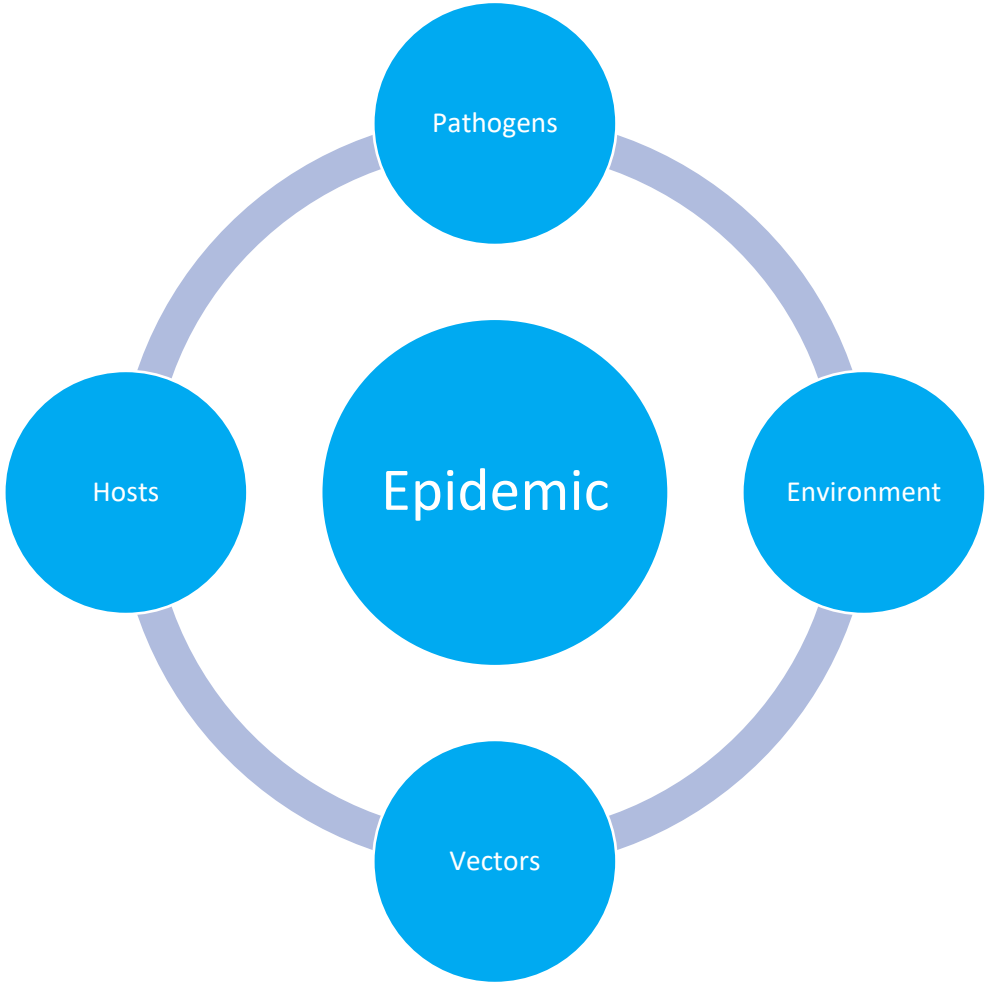
© WHO/Thilo Hub

UMEÅ UNIVERSITY

World Health Organization

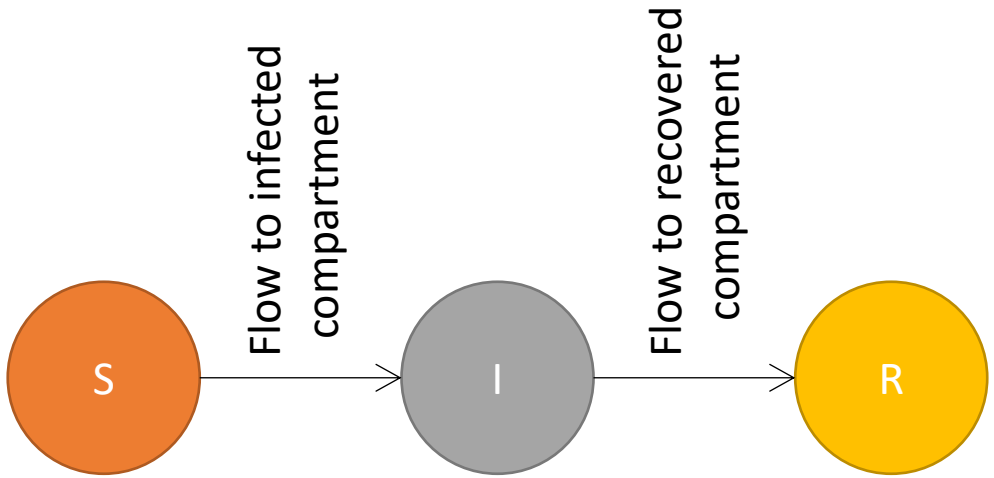
A SYSTEMS PERSPECTIVE ON EPIDEMICS

- The epidemiological progression
  In epidemics, it is a complex dynamical system.
- A complex set of features determines the epidemiological outcome over time.



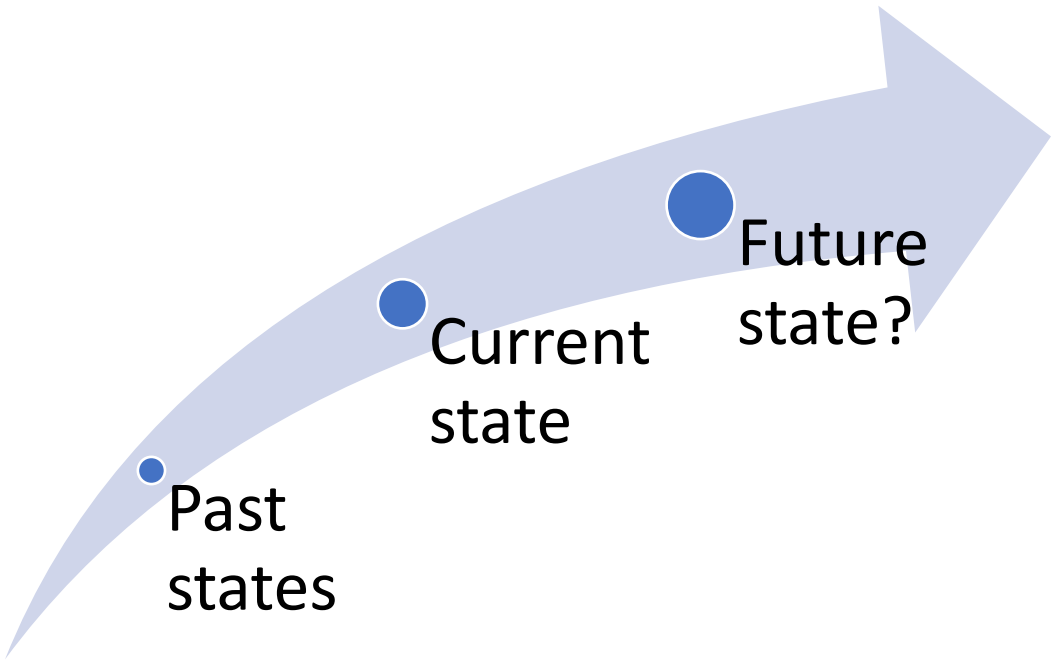
THE CONVENTIONAL FORMALISM – COMPARTMENTAL MODELS IN EPIDEMIOLOGY

- The standard SIR model can exemplify the concept.
- **(S):** The average number of **S**usceptible individuals that can get infected.
- **(I):** The average number of **I**nfected and infectious individuals.
- **(R):** The average number of **R**ecovered individuals. Or, more generally, individuals who cannot be infected because they are removed from the transmission dynamics after infection, e.g., because they were immunised (recovered), moved to isolation, or died.



STATE SPACE AND CONTROL SPACE

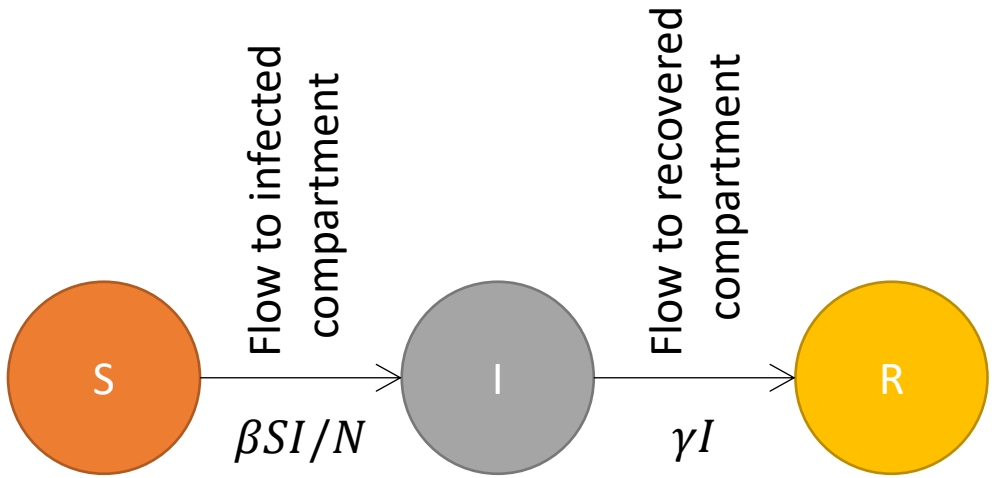
- The state-space is the set of values that variables can take, e.g., the number of infected persons *S*.
- The control space is the set of values that parameters can take, e.g., the rate of transmission.
- We typically know the past states, the current state and the rate at which it is changing, but not the long-term outcome.
- Process-based models provide an outcome depending on the current state and the associated process rates of change.



PROCESS RATES DESCRIBE THE SPEED OF CHANGE IN A DYNAMICAL SYSTEM – FROM ONE STATE TO ANOTHER

Box 1. Mathematical representation of the change in S, I and R per unit time.

$$\frac{dS}{dt} = -\beta SI/N$$
$$\frac{dI}{dt} = \beta SI/N - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$
$$N = S + I + R$$



- The average number of susceptible becoming infected per unit time is $\beta SI/N$.
- The average number of infected recovering per unit time is γI.
- The infection rate per susceptible individual is equal to $\beta I/N$; the "force of infection" (λ).
- The recovery rate per infected individual is equal to γ.
- The parameter β is often called the infection rate or transmission rate.

THE FORCE OF INFECTION λ

- The force of infection equals the per capita rate at which susceptible individuals contract the infection.
- For a standard SIR-model, the force of infection is $\beta I/N$.
- The force of infection can take any plausible form, but there are two typical and conceptually important forms:
- (1) Frequency-dependent force of infection.
- (2) Density-dependent force of infection.

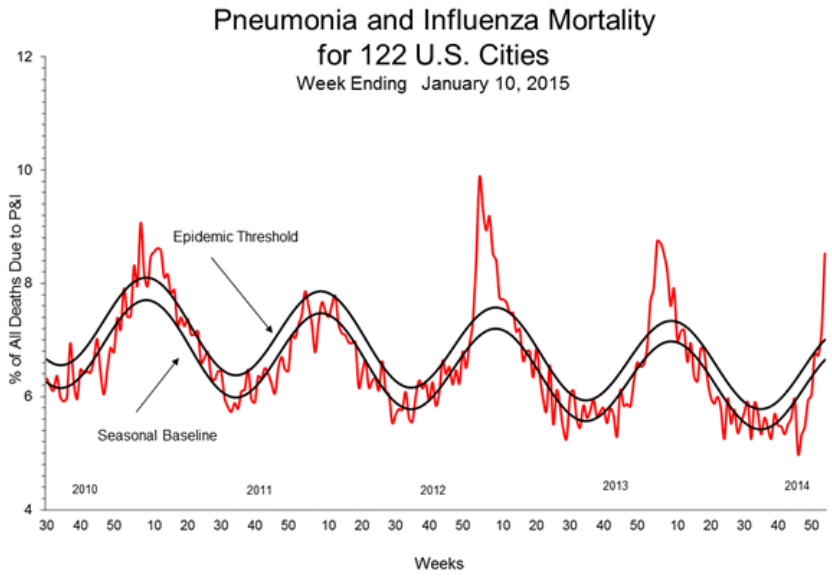
Box 2. Two important forms of force of infection.

$$\lambda = \frac{\beta I}{N} \qquad (1)$$

$$\lambda = \beta I \qquad (2)$$

THE TRANSMISSION RATE β

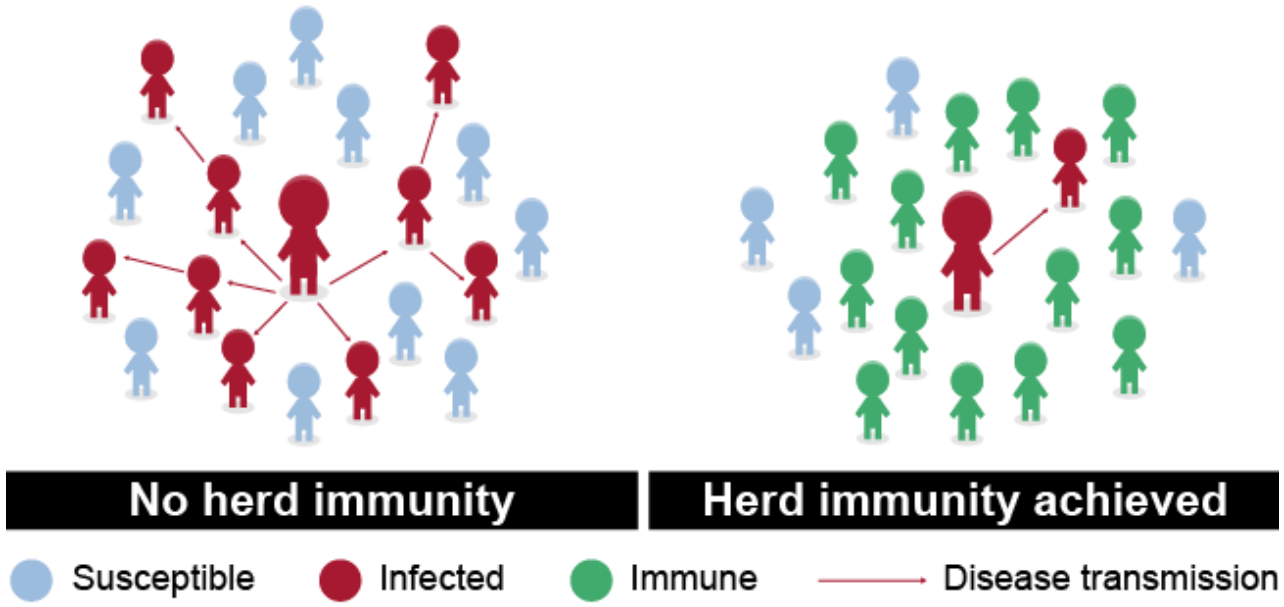
- In simple models, β is often assumed to be constant.
- It is thought of as the product between the average contact rate c between people and the probability τ that an infection event takes place during a contact event.
- The transmission rate is often, however, dependent on the environment over time, e.g., seasonality of influenza or social/physical distancing implementations within populations, etcetera.
- Accordingly, it is often useful to implement β as a time-dependent function, leading to a "temporally forced model".



https://www.cdc.gov

THE BASIC REPRODUCTION NUMBER $\boldsymbol{R_0}$ AND HERD IMMUNITY

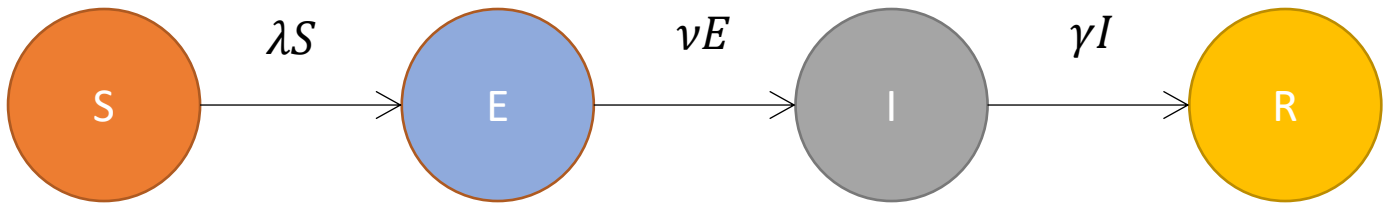
- The basic reproduction number is defined as:
  The average number of secondary cases arising from an average primary case in an entirely susceptible population.
- In a standard SIR-model $R_0 = \beta/\gamma$.
- Herd immunity is achieved when the immune fraction of the population is greater than $1 - 1/R_0$.



Source: GAO adaptation of NIH graphic. | GAO-20-646SP

ADDING A LATENT PERIOD: THE SEIR-MODEL

- There is often a latent period between the time of transmission and the onset of symptoms or infectiousness.
- A latent period is well captured by an SEIR model, which extends the SIR model by an "exposed" compartment start equation, with



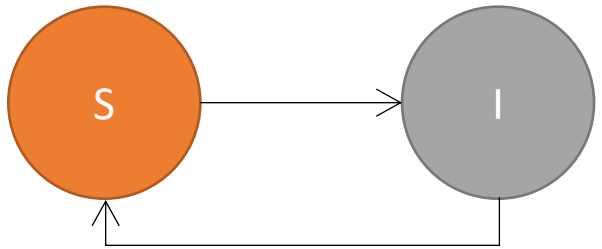
Box 3. Mathematical representation of the change in S, E, I and R per unit time.

$$\frac{dS}{dt} = -\lambda S$$
$$\frac{dE}{dt} = \lambda S - \nu E$$
$$\frac{dI}{dt} = \nu E - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

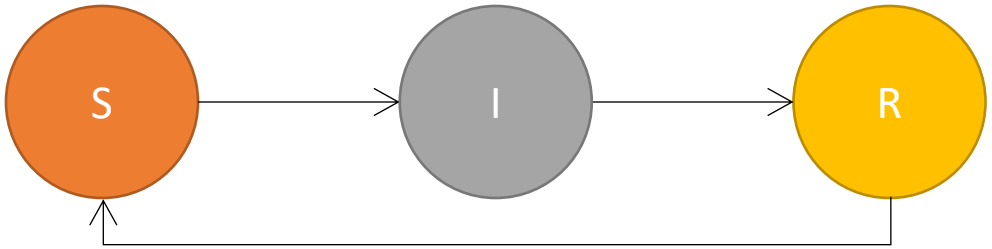
OTHER COMMON MODEL STRUCTURES

- The SIS model for when immunity does not develop.
- The SIRS for when immunity is temporary.
- In this type of manner, the structure of compartmental models in epidemiology can be altered and adapted to apply to specific pathogens or circumstances.
- The expression for R_0 is dependent on the model structure.

SIS-model



SIRS-model



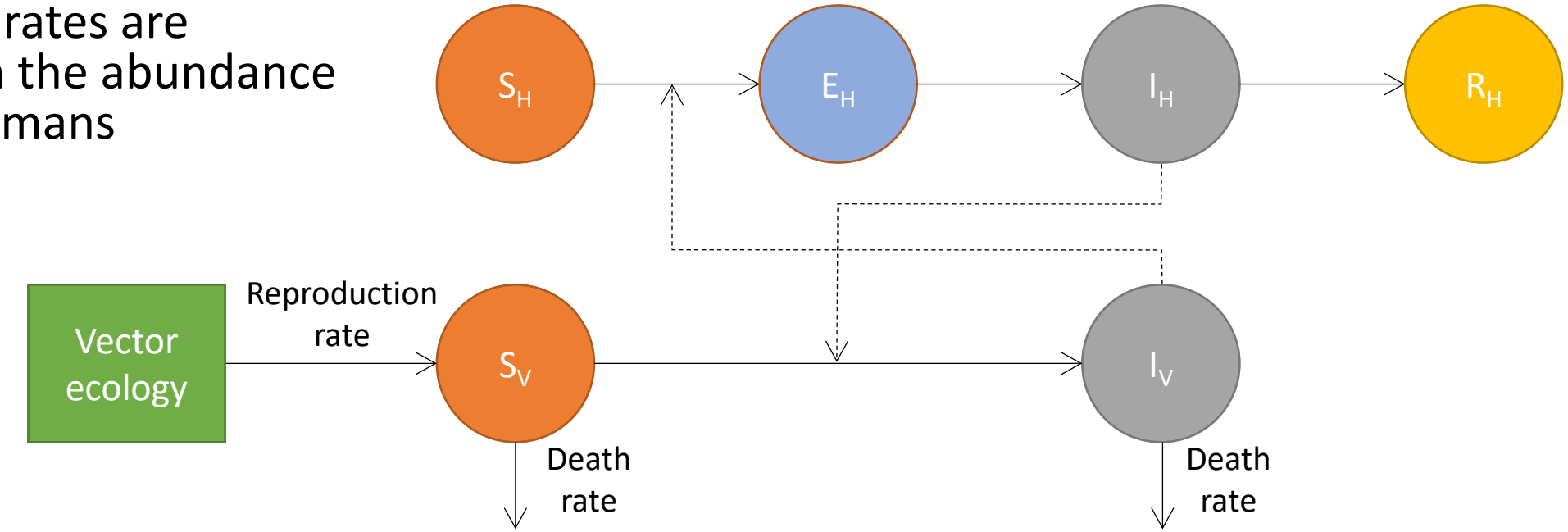
VECTOR-BORNE DISEASES

- For vector-borne diseases, like Malaria or dengue fever, transmission occurs across humans and a vector species.
- To model this type of epidemiology, the human and the vector-species population must be coupled.
- The ecology of the vector species must be taken into account.

© WHO/Hedinn Halldorsson

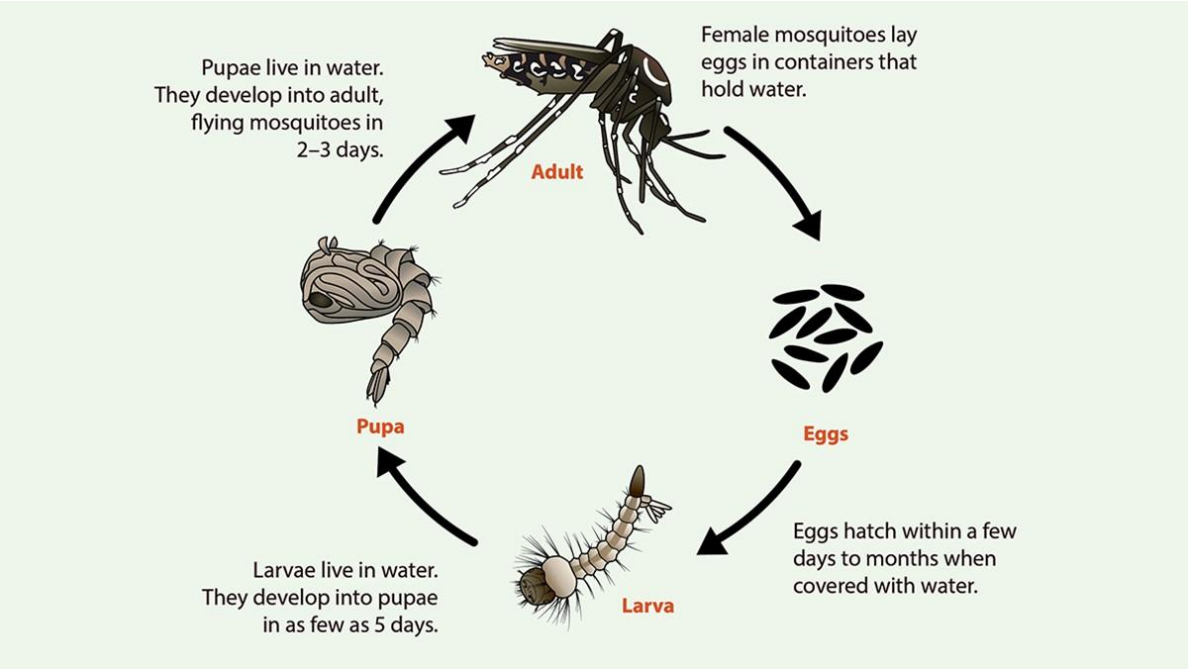
VECTOR-BORNE DISEASES

- This flow diagram is an example of a model of some vector-borne disease.
- A human SEIR model is coupled with a vector SI model with vector ecology (vector demographics).
- The dashed arrows indicate that infection rates are dependent on the abundance of infected humans and infected vectors.



WEATHER AND CLIMATE DEPENDENCE

- Weather and climate become central for vector-borne diseases.
- Primarily due to the fact that it directly affects the ecology of vector species.
- For example, the juvenile stages of mosquito vectors are directly dependent on water supply, and the vector is generally sensitive to temperature.



Source: CDC. https://www.cdc.gov/mosquitoes/images/Aedes-life-cycle.jpg

VECTORIAL CAPACITY $\boldsymbol{V}$

- Vectorial capacity V is generally a measure the rate of transmission of a pathogen from a vector to humans.
- It has been expressed in several mathematical forms.
- It relates closely to being a counterpart to contact rate c or transmission rate β in, for instance, the SIR model.
- Vectorial capacity can be expressed as a function of R_0.
- Vectorial capacity is often weather and climate dependent.

$$V = R_0 \frac{\gamma}{\tau}$$

Vectorial capacity, where γ is the recovery rate of humans and where τ is the probability that a human becomes infected once bitten by an infectious vector individual.

R_0 FOR VECTOR BORNE DISEASES

- The basic reproduction number R_0 for vector borne diseases differs conceptually from that of human-to-human diseases.

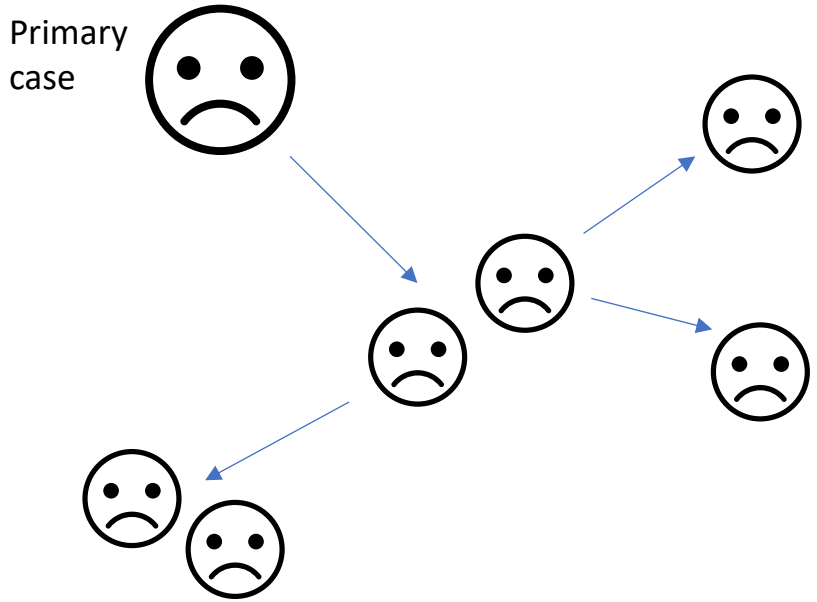
- R_0 can be given by the product between the expected number N_{H} of humans in a susceptible population that becomes infected by an infectious vector individual and the expected number of vector individuals N_{V} that in turn becomes infected from an infectious human; because of this two-way transmission process for vector borne diseases.

$$R_0 = N_H \, N_V$$

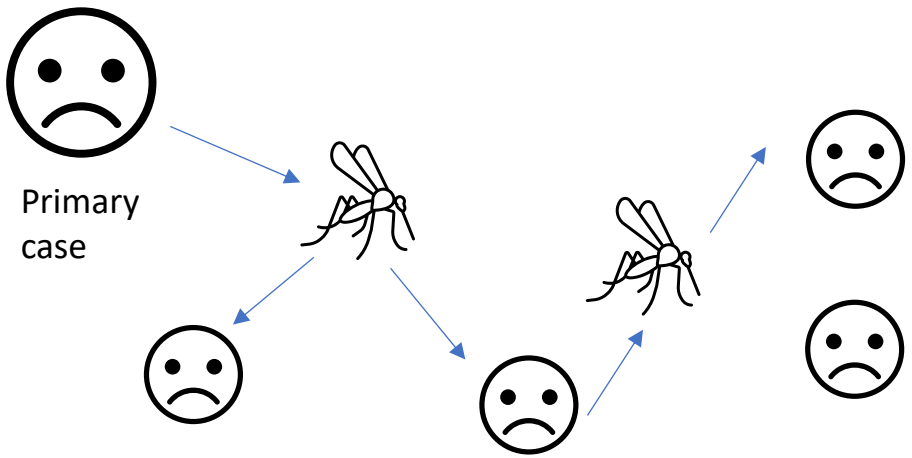
The basic reproduction number, where N_{H} is expected number of humans in a susceptible population that becomes infected by an infectious vector individual, and N_{V} is the expected number of vector individuals that in turn becomes infected from an infectious human.

Comparison of $\boldsymbol{R_0}$ and Vectorial Capacity $\boldsymbol{V}$

R_0: Average number of secondary cases produced by a single infection into a completely susceptible population



V: Average daily number of secondary cases generated by one primary case introduced into a completely susceptible population



$V = R_0 / T_h$ where T_h is the duration of viremia in humans if probability per infectious bite is 1.

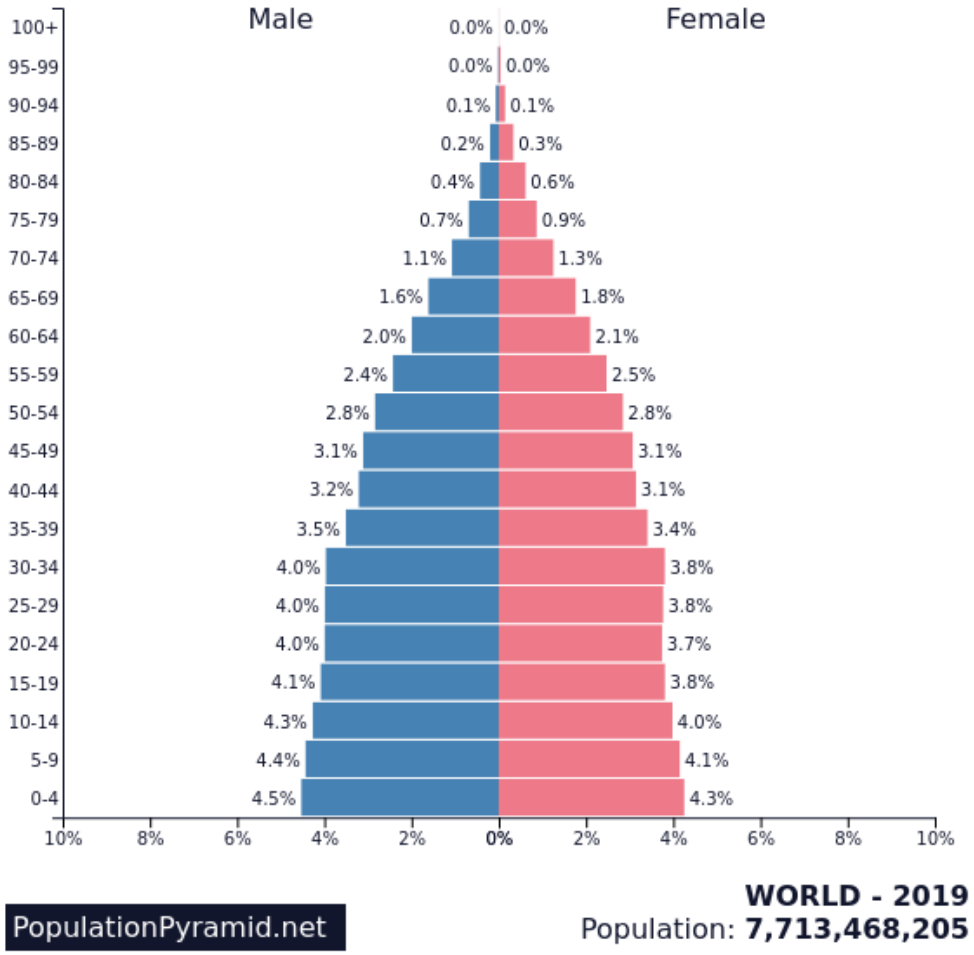
PARAMETERIZATION AND STRUCTURE OF PROCESS-BASED MODELS

- While theoretical process-based models are good at providing conceptual understanding of the qualitative dynamics of the epidemiology of infectious diseases, data-guidance is crucial for understanding and predicting the progression of practical epidemics.

- For a given epidemiological problem, data can help to determine, for instance:
  - the relevant state-space; i.e., the relevant compartments (state-variables) and their potential values.
  - the relevant control space, i.e., the parameters (such as infection rates, recovery rates, contact rates, etc.) and their potential values.

- Values on some parameters may be difficult to know from data. These values can be estimated by fitting the model to previous situations, e.g., data on previous or historical incidence, number of infected, infection fatality ratio, among other things.

STRUCTURED MODELS

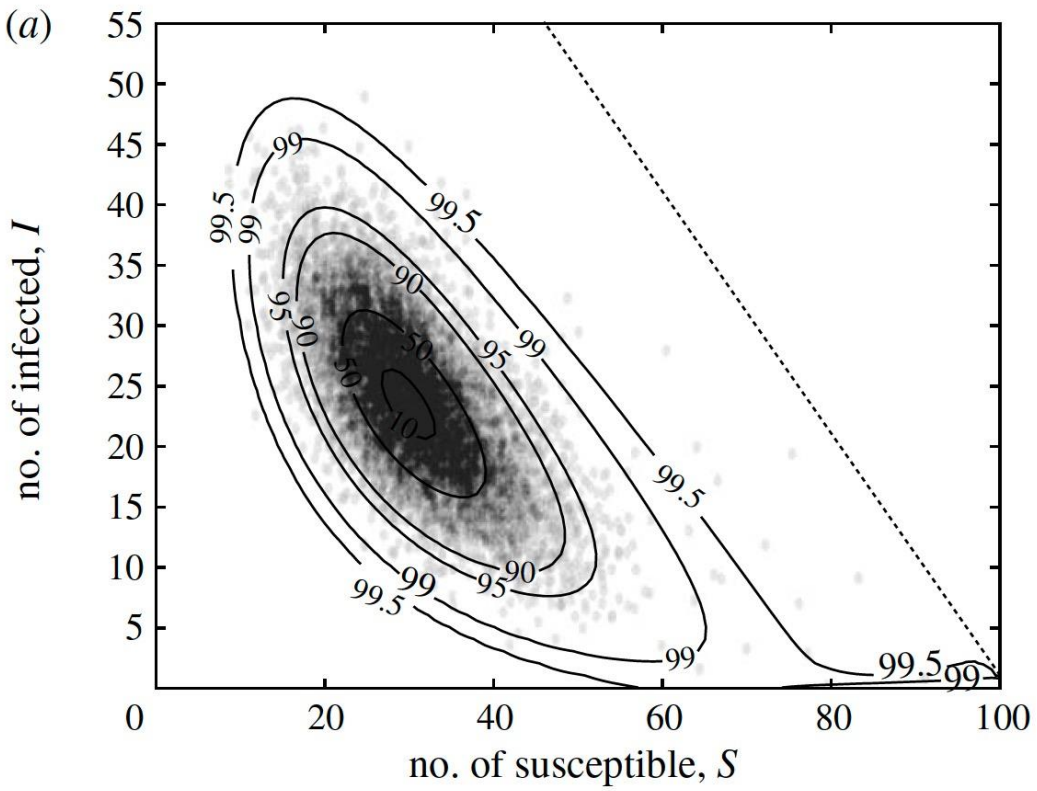


https://www.populationpyramid.net/world/2019/

- Simplified process-based models are very important for conceptual understanding.
- To capture real-life processes, it is often useful to account for relevant structures, for example, contact structures, spatial distribution, travel processes, or age structures in the population.

STOCHASTIC MODELS

- Deterministic models (e.g., the standard SIR model) account for the expected number of individuals, or the fraction of individuals in any compartment.
- This is a good approximation in large populations and is often applicable.
- Stochastic models account for discrete individuals and for the probability of a certain number of individuals in any compartment.
- Stochastic models become important when considering processes that act on a smaller population or on subsets of a population.



Keeling and Ross (2007), doi: 10.1098/rsif.2007.1106

BIBLIOGRAPHIC REFERENCES AND FURTHER READING

- Randolph, H. E. & Barreiro, L. B. Herd Immunity: Understanding COVID-19. *Immunity* **52**, 737–741 (2020).
- DiSera, L. *et al.* The Mosquito, the Virus, the Climate: An Unforeseen Réunion in 2018. *GeoHealth* **4**, e2020GH000253 (2020).
- Keeling, M. J. & Ross, J. V. On methods for studying stochastic disease dynamics. *J R Soc Interface* **5**, 171–181 (2008).
- Keeling, M. J. & Rohani, P. *Modeling Infectious Diseases in Humans and Animals*. (Princeton University Press, 2008).

World Health Organization

Section 3.4 :

Identifying associations by exploring temporal patterns of health risks: time lags and time series models

- **Learning objective**: To understand the basic principle of the time-series study design and best practices for their use. Gain a basic understanding of how to interpret the effect estimates from time-series regression modelling

Further reading:

- Gasparrini A, et al. (2015). Mortality risk attributable to high and low ambient temperature: a multicountry observational study. The Lancet 386:369–375.
- Bhaskaran K, et al. (2013). Time series regression studies in environmental epidemiology. Int J Epidemiol 42(4):1187–1195.
- Peng RD & Dominici F (2008). Statistical methods for environmental epidemiology with R: a case study in air pollution and health. Springer.
- Peng RD, et al. (2006). Model choice in time series studies of air pollution and mortality. J R Stat Soc Ser A 169(2):179–203.
- Schwartz J (2000). The distributed lag between air pollution and daily deaths. Epidemiology 11(3):320–326.
- Hajat S & Kosatky T (2010). Heat-related mortality: a review and exploration of heterogeneity. J Epidemiol Community Health 64:753–760.
- DLNM time-series workshop materials (Kim et al., ISEE-AC 2016): http://hosting03.snu.ac.kr/~hokim/iseeac2016/

3.4 Identifying associations by exploring temporal patterns of health risks: time lags and time series models

Dr Yoonhee Kim[1] and Dr Ho Kim[2]

[1]Graduate School of Medicine, The University of Tokyo, Japan
[2]Graduate School of Public Health, Seoul National University, Republic of Korea

東京大学
THE UNIVERSITY OF TOKYO

Learning objectives

- Understand the basic principles of the time-series study design.
- Learn best practices for using time-series studies and how to interpret effect estimates from the time-series regression modelling.

Time-series studies

- An ecological study design
- Substantial developments since the 1980s

- Multiple names
  - Time-series regression model
  - Time-series log-linear model
  - Time-series Poisson model

- The measurements of each variable constitute a time series (e.g., days or weeks).

Table 1 Example rows of time series data from the London dataset showing daily levels of environmental variables and daily number of deaths

Date	Ozone ($\mu g/m^3$)	Temperature (°C)	Relative humidity (%)	*n* deaths
1 Jan 2002	4.59	−0.2	75.7	199
2 Jan 2002	4.88	0.1	77.5	231
3 Jan 2002	4.71	0.9	81.3	210
4 Jan 2002	4.14	0.5	85.4	203
5 Jan 2002	2.01	4.3	93.5	224
6 Jan 2002	2.4	7.1	96.4	198
7 Jan 2002	4.08	5.2	93.5	180
8 Jan 2002	3.13	3.5	81.5	188
9 Jan 2002	2.05	3.2	88.3	168
10 Jan 2002	5.19	5.3	85.4	194
11 Jan 2002	3.59	3.0	92.6	223
12 Jan 2002	12.87	4.8	94.2	201

Source: Bhaskaran et al. 2013, International Journal of Epidemiology

Acute (short-term) effects of time-varying exposure

Research question:

Is there an association between **day-to-day variation** in exposure and **day-to-day variation** in the health outcome? (or any short-term discrete time interval)

Here, we assume that the population size in a city/location today **does not change much** from yesterday.

The London smog, 1952

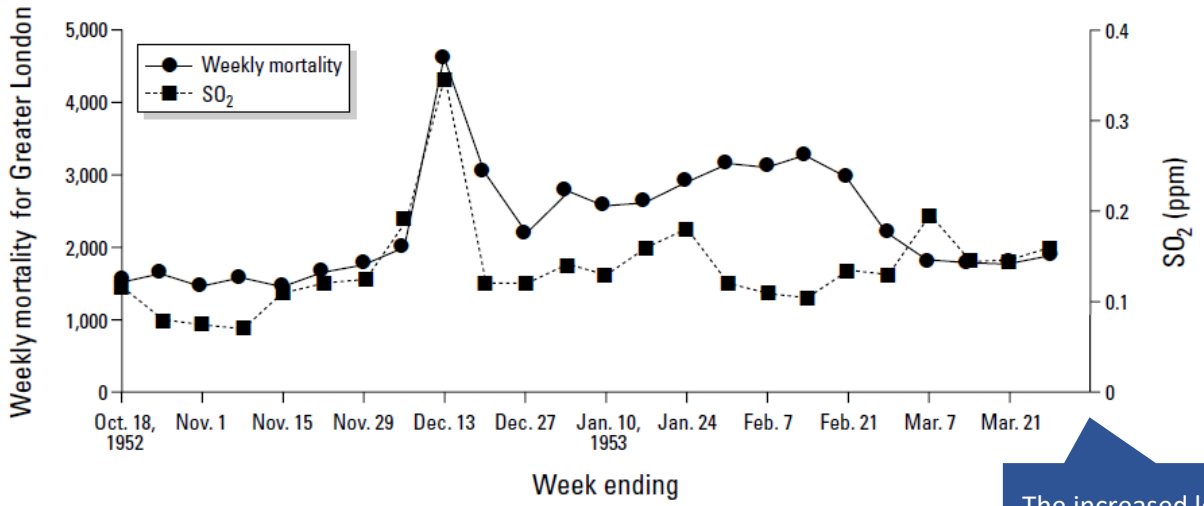
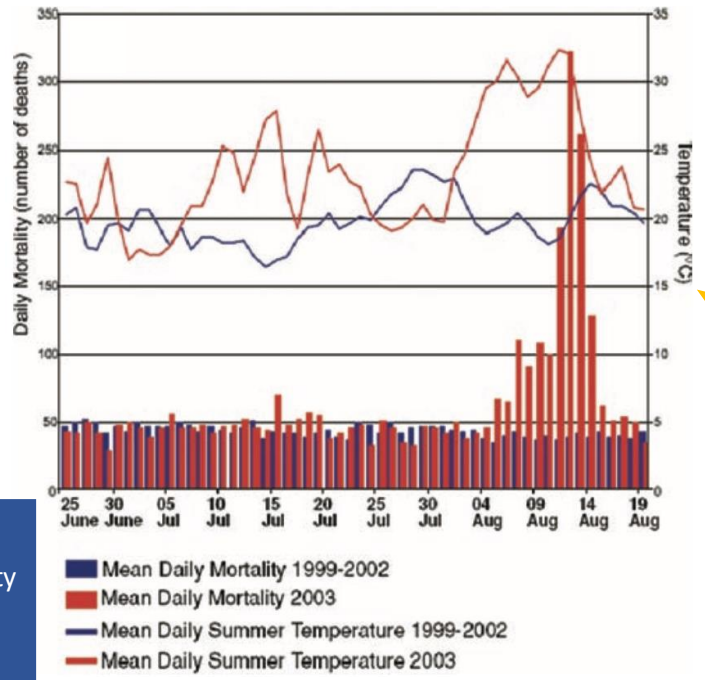


Figure 1. Approximate weekly mortality and SO_2 concentrations for Greater London, 1952–1953.

Bell and Davis (2001) Environmental Health Perspectives

The increased level of sulfur dioxide and mortality counts in the week of Dec 13 suggests a short-term association between air pollution and mortality

Heatwave in Europe, 2003



IPCC (2007) Chapter 8 of the 4th Assessment Report by Working Group 2

Suggests a short-term association between high ambient temperature and mortality observed during the 2003-heatwave in France.

Time-series regression modelling

Aim

- To examine whether the **short-term** variation in the outcome is explained by the **short-term** variation in the exposure of interest

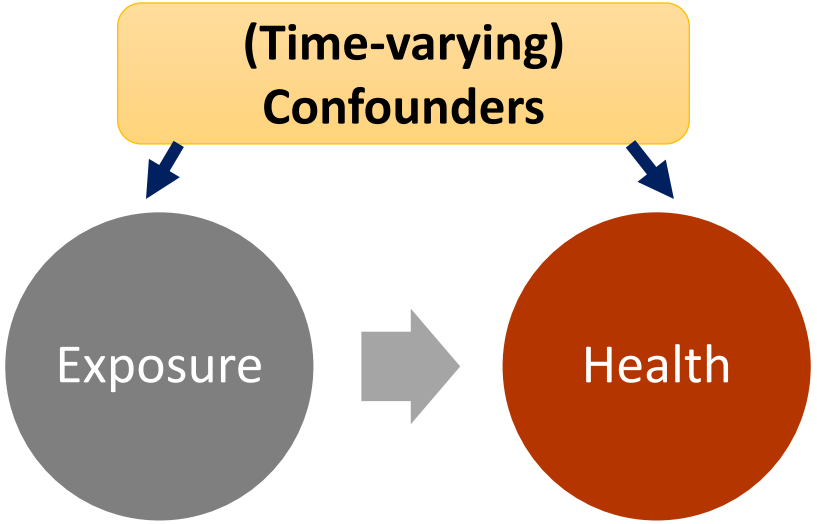
Health outcomes

- Count variable (e.g., daily death counts) → Poisson distribution assumed

Exposure of interest: time-varying environmental stressors (e.g., air pollutants, ambient temperature, etc.)

Time-varying confounders

- **Unmeasured confounders**
  - i.e., seasonal and long-term time trends
- **Measured confounders**
  - Other environmental variables
  - Day-of-week and/or holidays

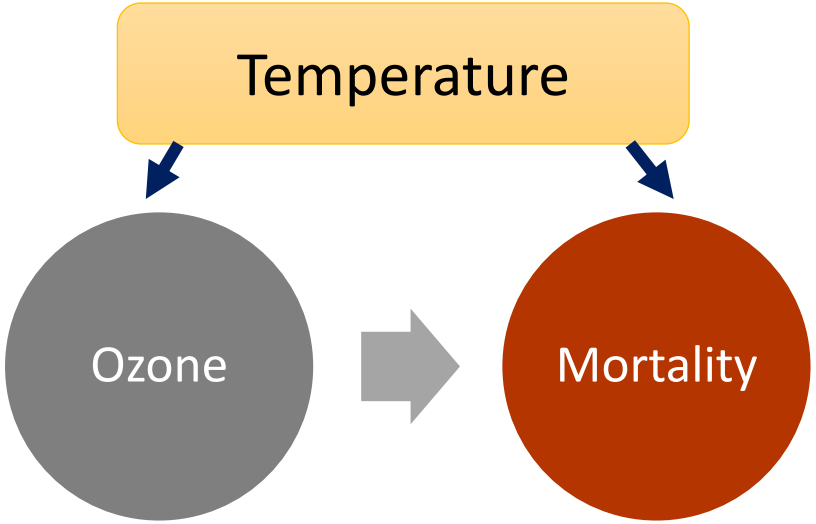


Time-varying confounders (measured) – *Example 1*

Factors with short-term temporal variability associated with the exposure of interest and the health outcome

Important time-varying confounders

- Meteorological variables in air pollution studies

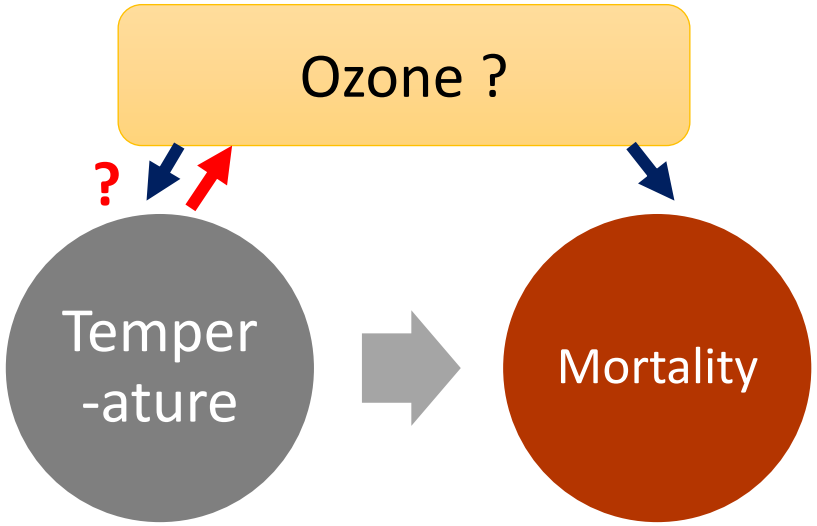


Time-varying confounders (measured) – *Example 2*

Factors with short-term temporal variability associated with the exposure and the outcome

Important time-varying confounders

- Air pollution variables in temperature-health studies



"We assert that the **role of ozone** in studies of temperature and mortality is **a causal intermediate** that is affected by temperature and that can also affect mortality, rather than a confounder."
(Reid et al. Environmental Health Perspectives 2012)

Time-varying confounders (measured) – *Example 3*

Factors with short-term temporal variability associated with the exposure and the outcome

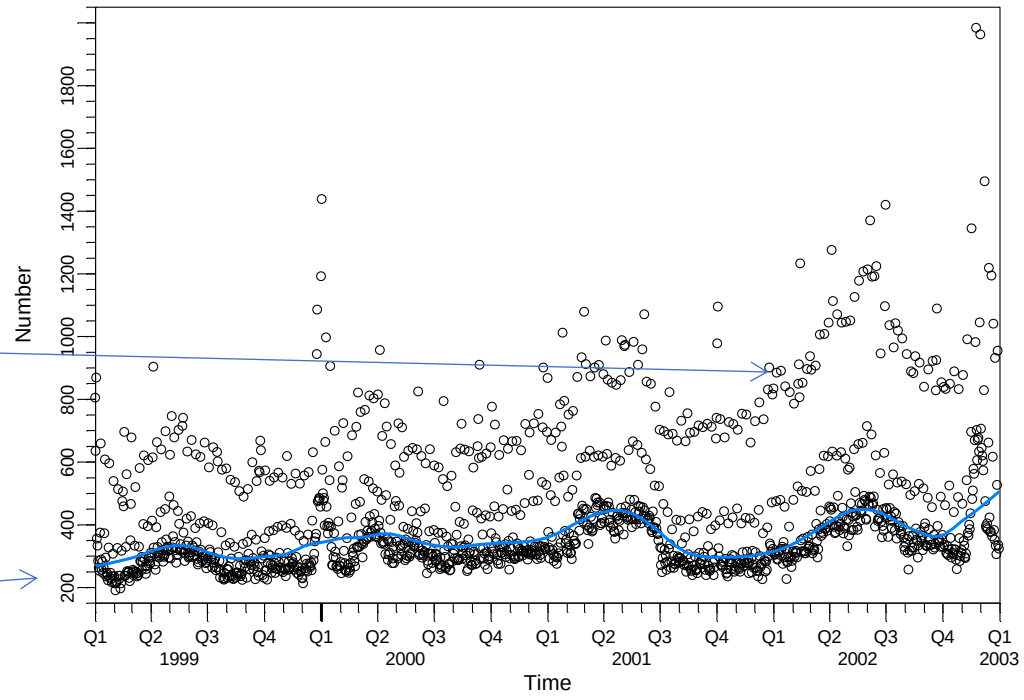
Important time-varying confounders

- Adjustment of the **day-of-week** is essential.

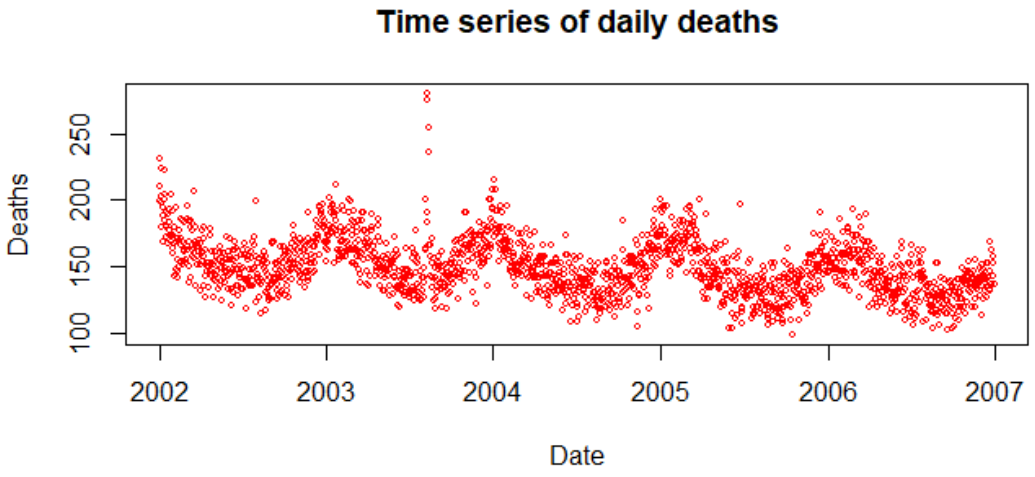
Daily hospitalizations in Seoul

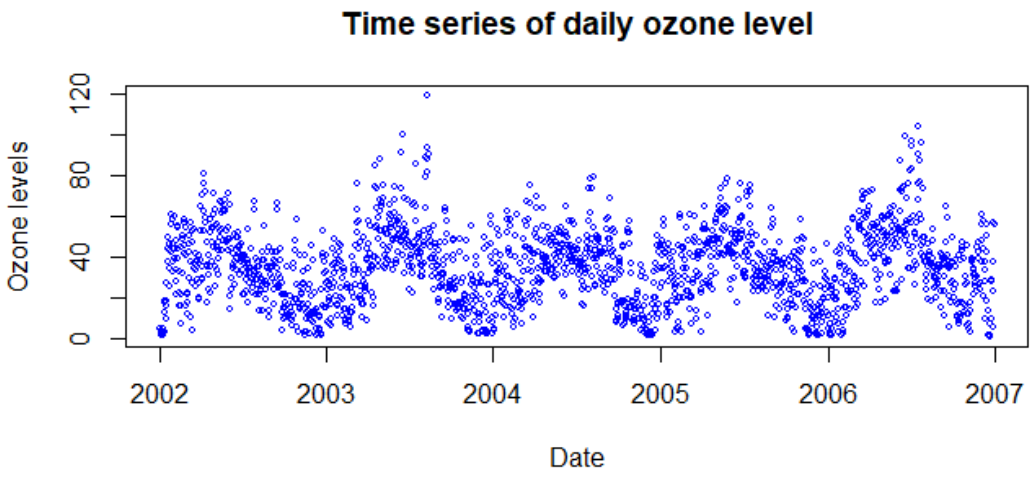
Mondays

Sundays and holidays



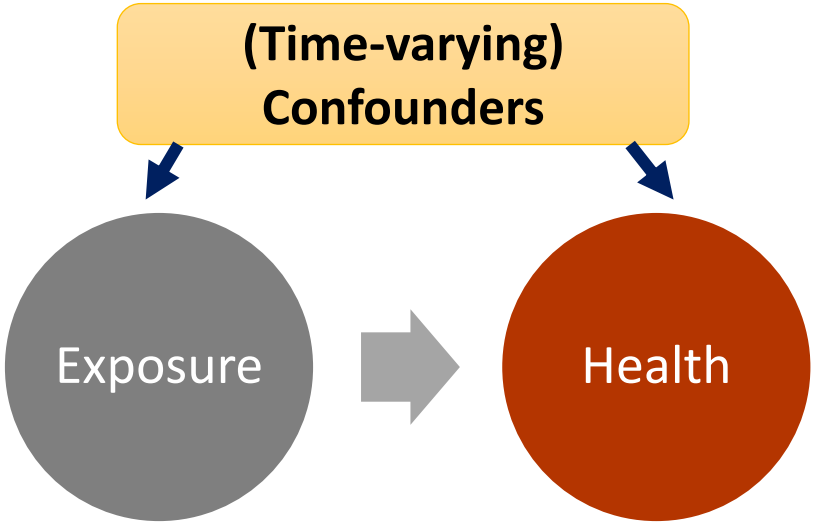
Time-varying confounders (unmeasured)



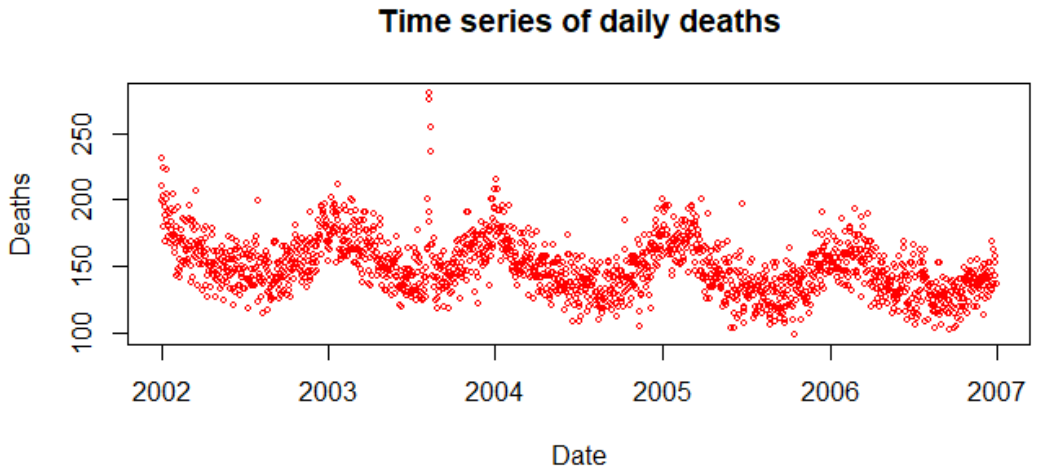


Seasonal patterns and **long-term trends** of both exposure and outcome can dominate their crude association

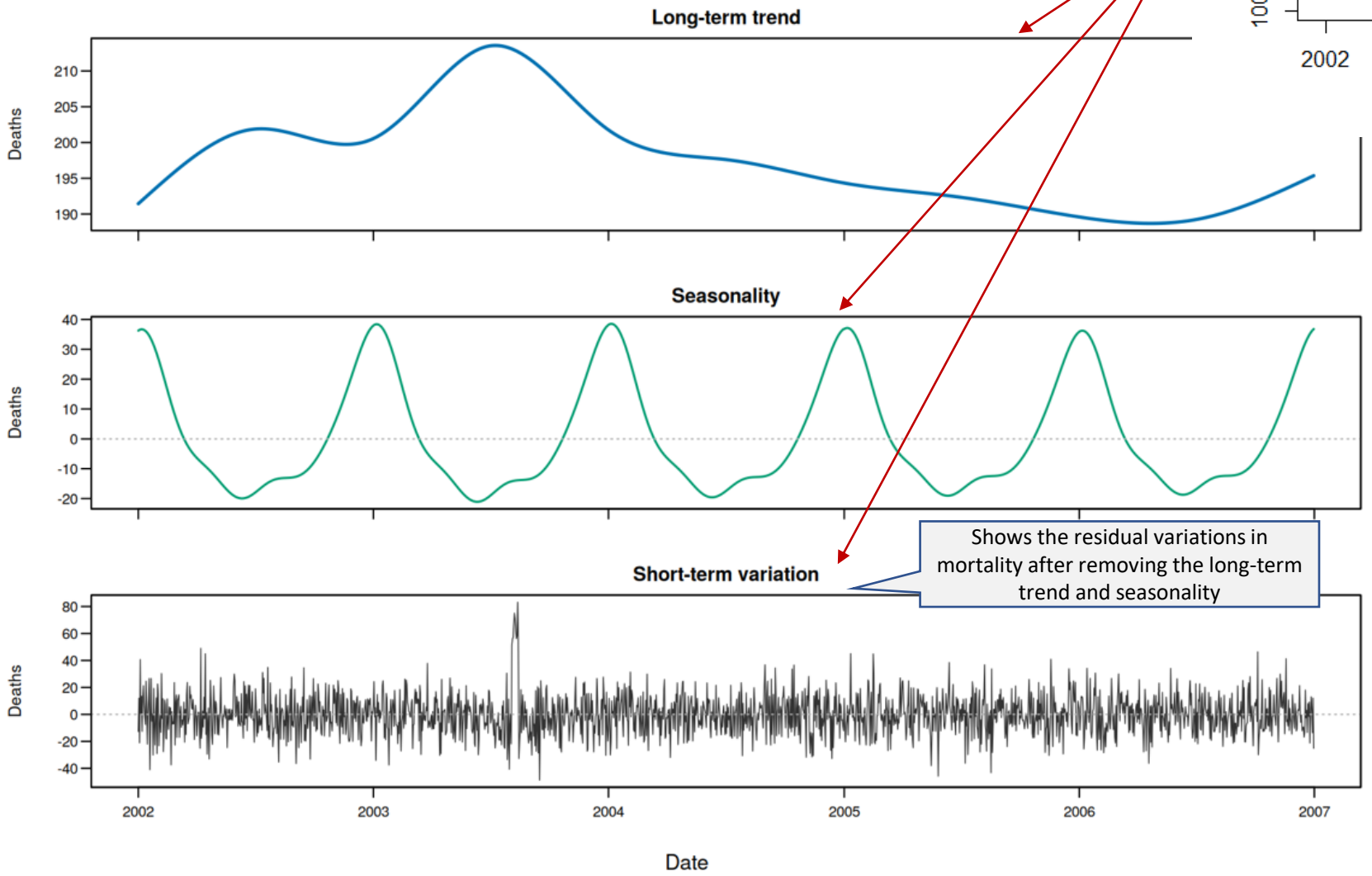
→ Necessary to **separate the trends** from the short-term association



Timescale decomposition



Decomposed into three parts



Shows the residual variations in mortality after removing the long-term trend and seasonality

Possible drivers for long-term time trend

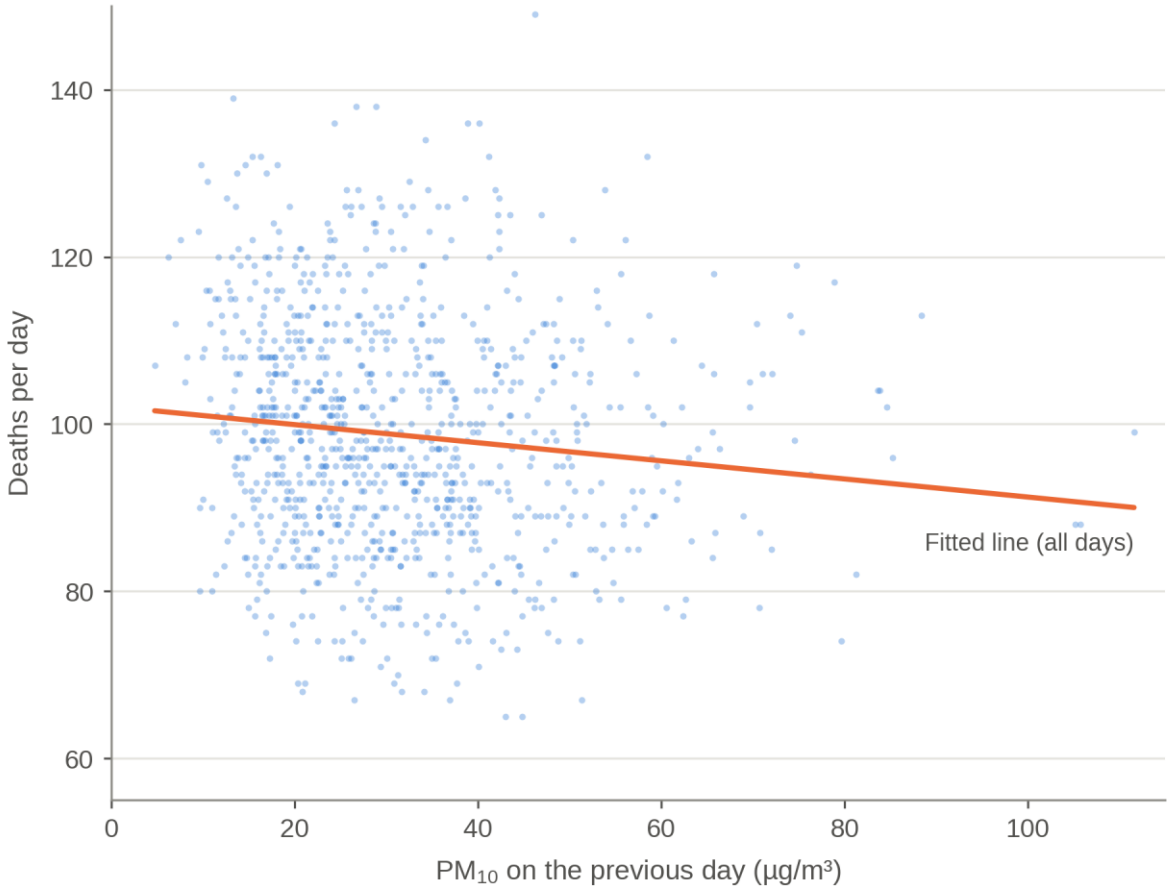
- Population size
- Population ageing
- Smoking rate
- Technology developments

for seasonal trends

- Infectious diseases
- Diet
- Human behaviors

Naïve approach (w/o adjustment of the unmeasured confounders)

for the association between PM10 and mortality

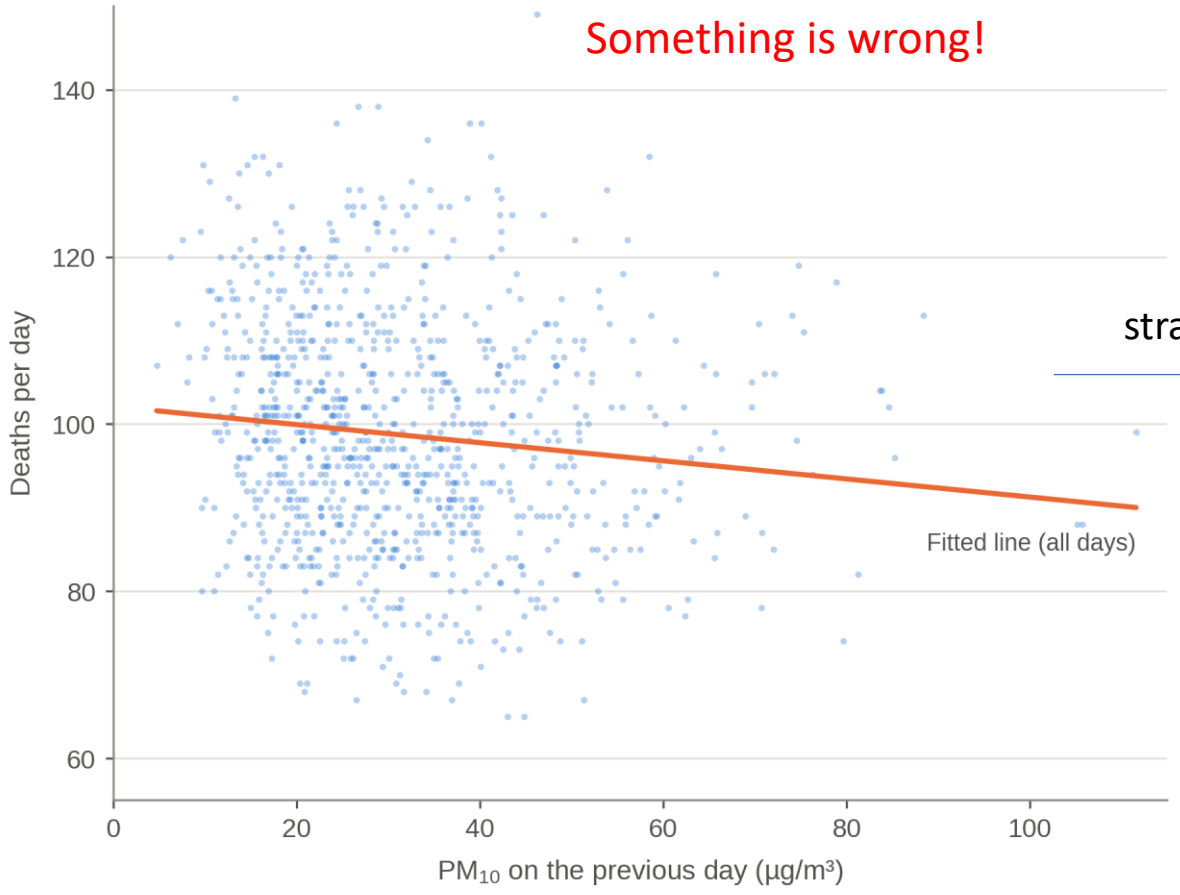


This crude association does not make sense!

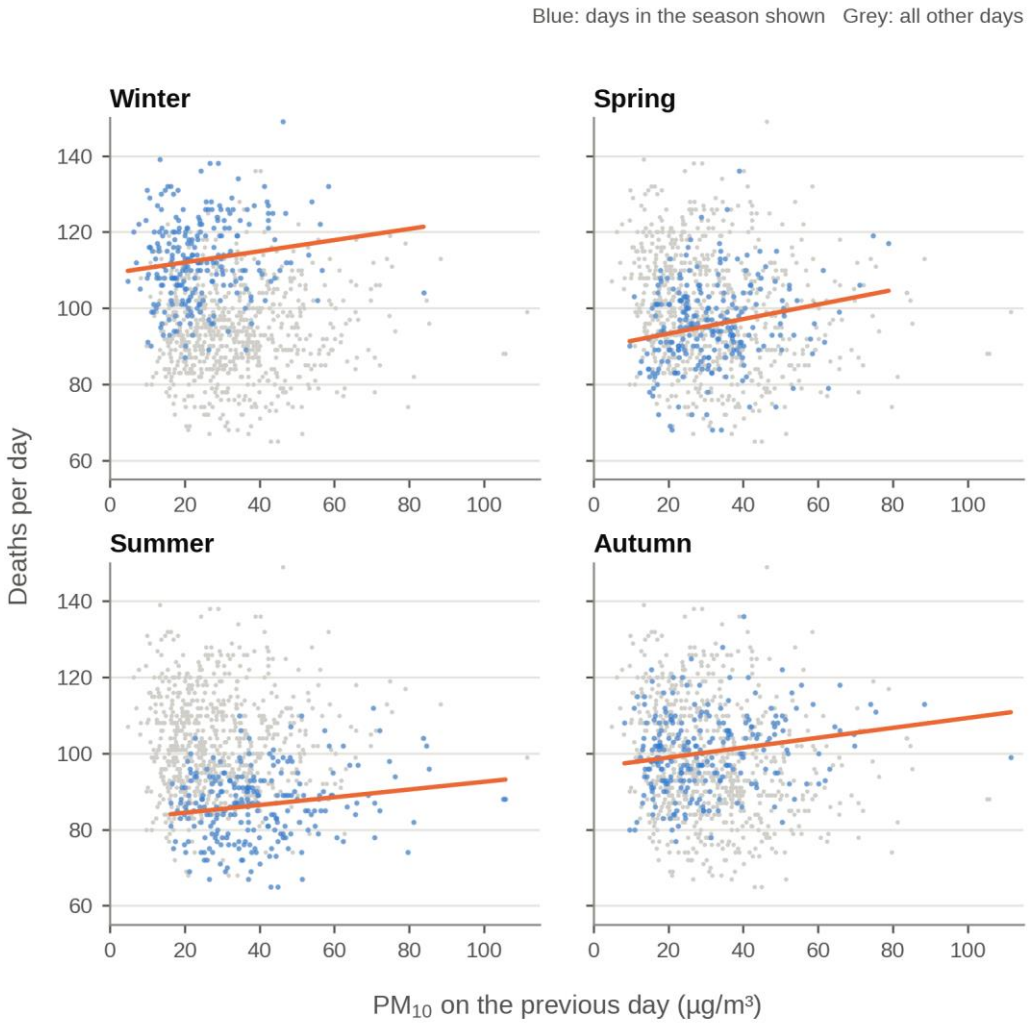
Something is wrong!

Schematic based on Peng and Dominici (2008) Statistical Methods for Environmental Epidemiology in R

Simple adjustment (stratified by seasons) - a little better



stratified



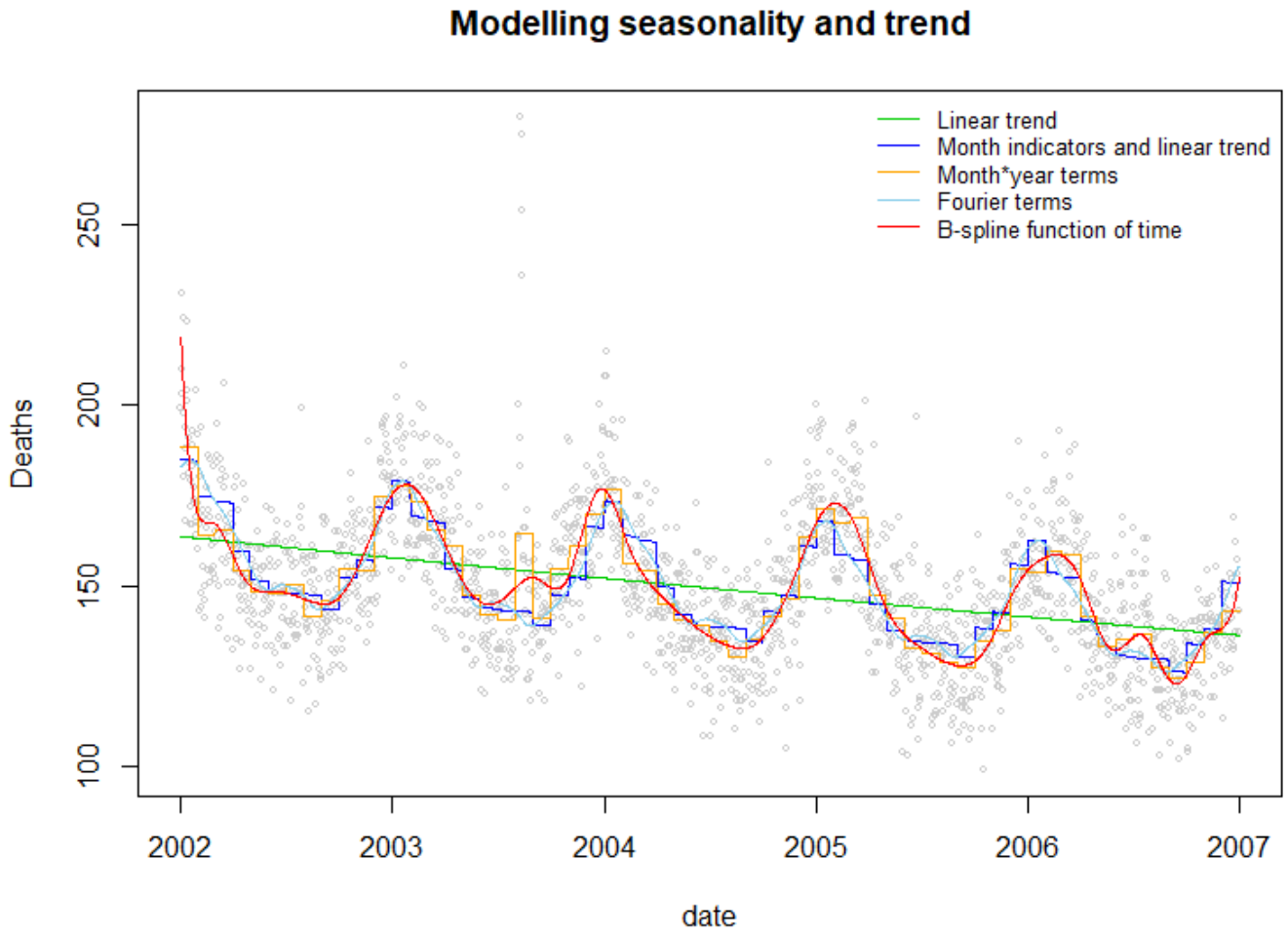
Schematic based on Peng and Dominici (2008) Statistical Methods for Environmental Epidemiology in R

How to adjust for season and long-term trends

How to effectively separate the trends from the short-term association?

Options

1. **Time-stratified model**
   (e.g., indicators of month)
2. **Periodic functions**
   (e.g., Fourier terms)
3. **Spline functions**
   (e.g., natural cubic splines, B-splines, etc.)



Bhaskaran et al. (2013) International Journal of Epidemiology
Adapted from the pre-conference workshop materials (Gasparrini A. et al.; ISEE 2018)

1. Time-stratified approaches

Split the whole period into intervals

Pros:

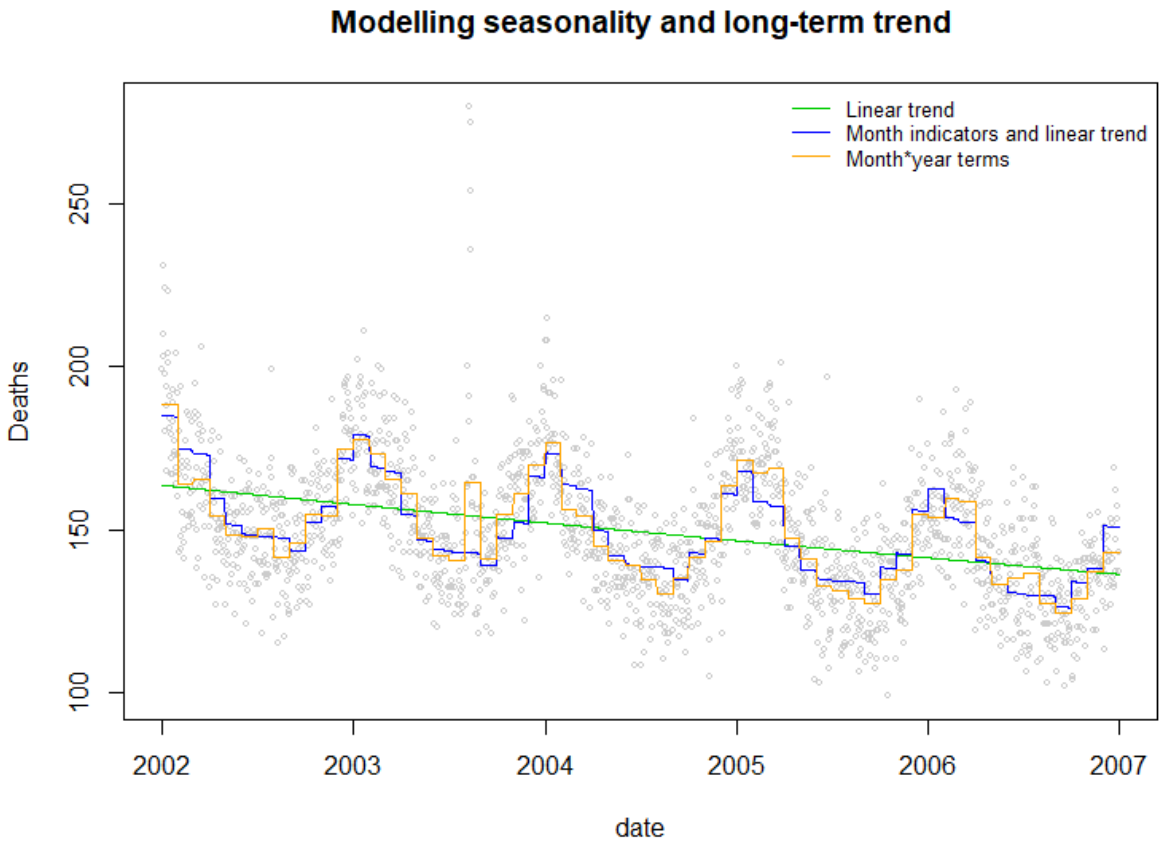
- Easy to understand
- Often captures the main long-term patterns well

Cons:

- Large number of parameters
- Assumes biologically implausible jumps in risk between adjacent time intervals

Interaction b/w month and year (in yellow)

→ Assumes different seasonal patterns between years



Bhaskaran et al. (2013) International Journal of Epidemiology

2. Periodic functions

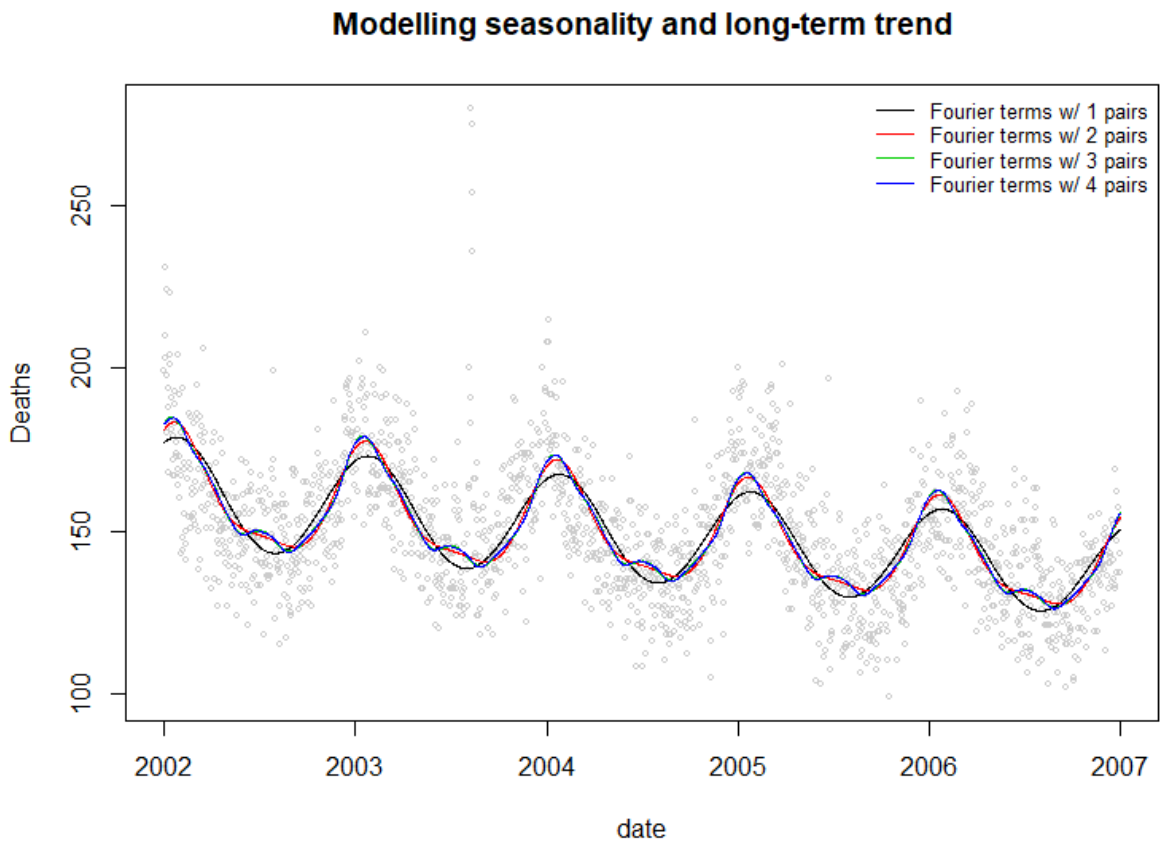
Model regular seasonal patterns using pairs of sine and cosine functions of time

Pros:

- Models smoother patterns
- Requires relatively few parameters

Cons:

- More mathematically complex than the time-stratified model
- Regular seasonal patterns may not always reflect the data
- Fourier terms alone cannot capture a long-term trend, but this can be addressed by adding a linear term for calendar time



Bhaskaran et al. (2013) International Journal of Epidemiology

3. Spline functions

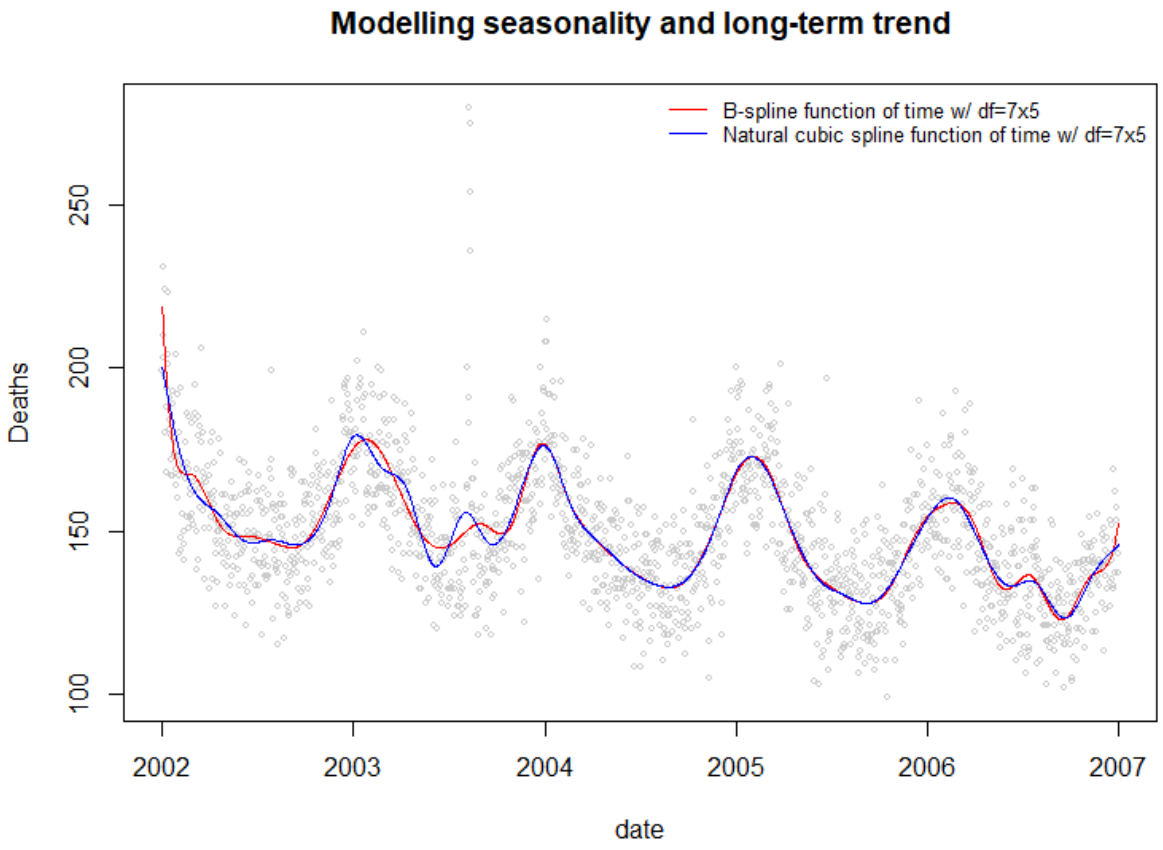
Polynomial curves join end to end at each knot to cover the full period, using a set of smoothlybasis variables of time, depending on the types of splines

Pros:

- Captures seasonal patterns and long-term trends smoothly, for both regular or irregular patterns depending on functional forms

Cons:

- More mathematically complex than other functions



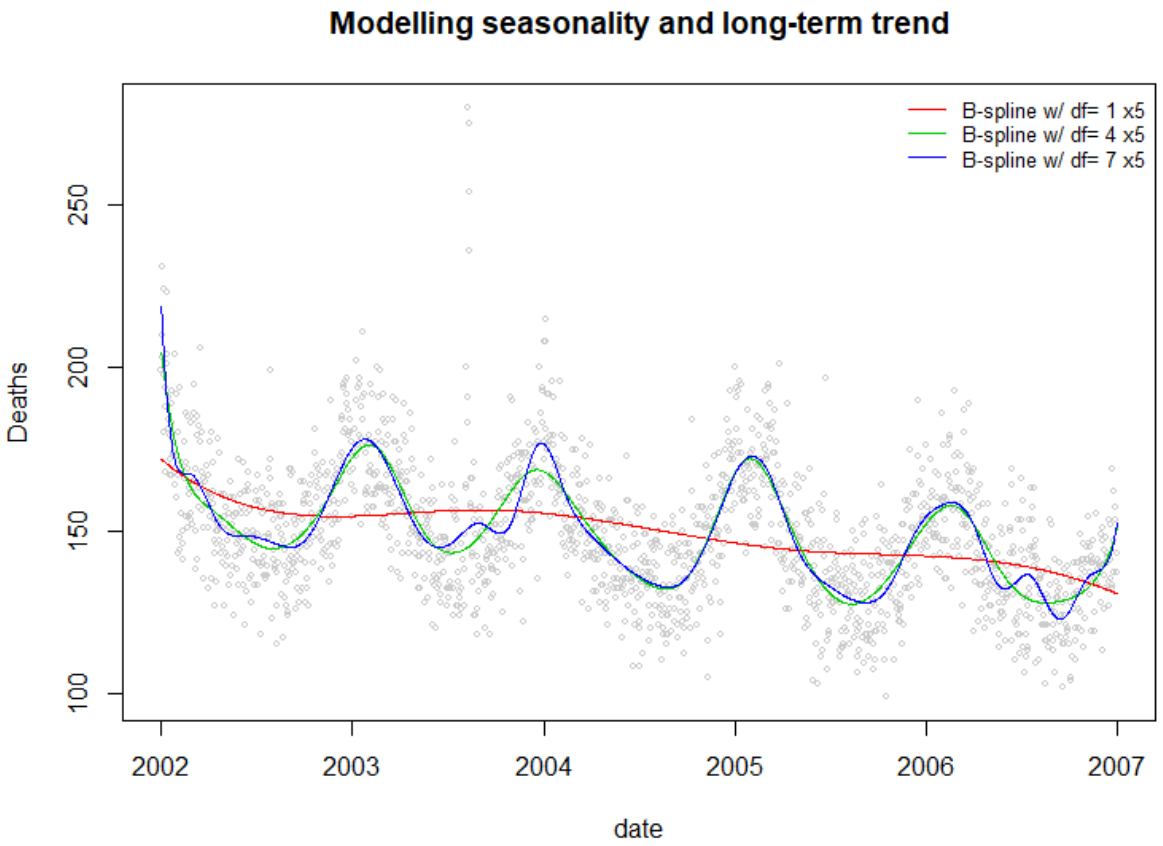
Bhaskaran et al. (2013) International Journal of Epidemiology

3. Spline functions

Choice for flexibility is determined by degrees of freedom (df)

In other words, it's determined by the number of knots (join-points) or end-to-end curves.

- Too few df will fail to capture the patterns.
- Too many df will result in a wiggly function, which may compete with the exposure of interest and lead to a wider confidence interval.



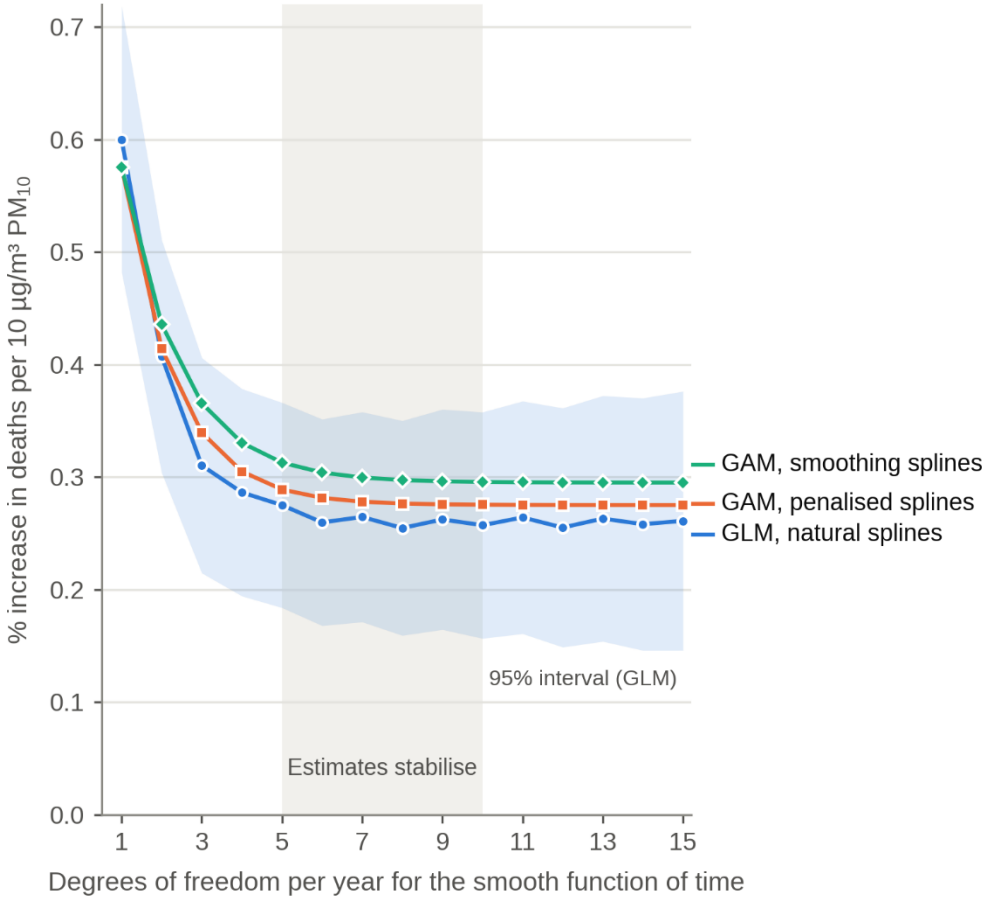
Bhaskaran et al. (2013) International Journal of Epidemiology

Choosing the degree of freedom for a smooth function of time

A model selection criterion

- e.g., Akaike information criteria (AIC) and/or Bayesian information criteria (BIC)
- Choose the optimal df from the model by minimizing the criterion

Generally, the graph shows that the effect estimates were relatively unstable with a lower degree of freedom but became more stable over the df range of 5 to 10, indicating that models with higher df provide more robust estimates. However, the best model should balance robustness of the results with appropriately capturing the exposure of interest, as mentioned in the previous slide.



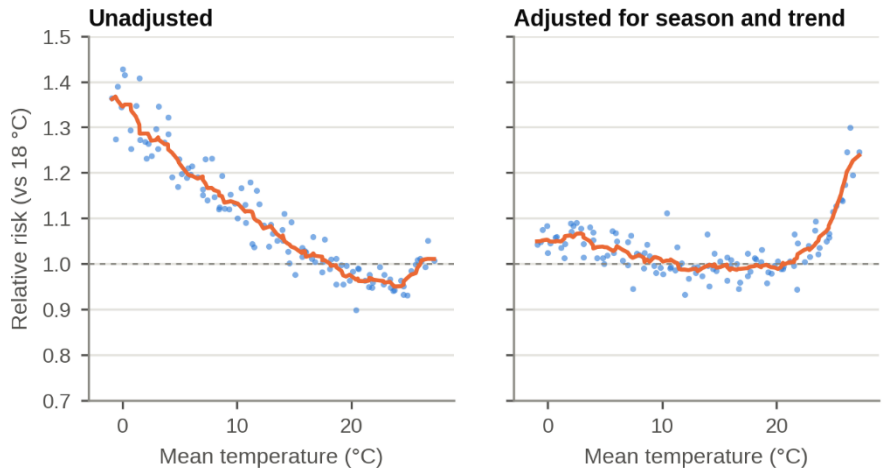
Sensitivity analysis

- Explore the sensitivity of the effect estimates by changing the df
- Choose the optimal df from the model providing the robust estimates

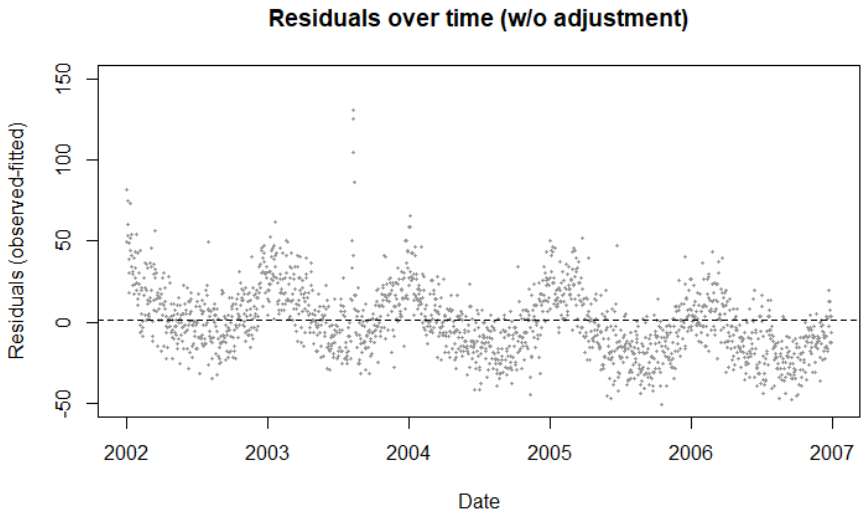
Schematic based on Peng and Dominici (2008) Statistical Methods for Environmental Epidemiology with R

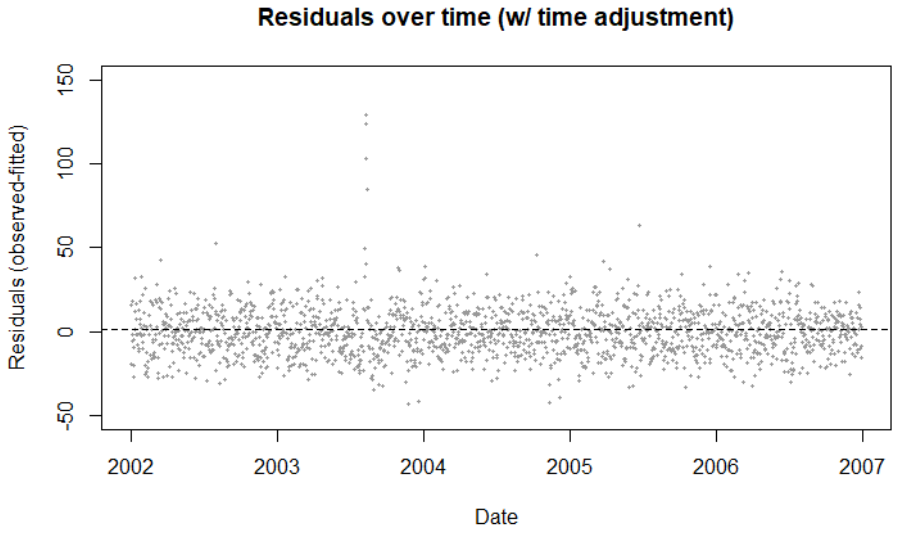
After the time adjustment

- Check the residual variations in daily deaths **before and after** ‘removing’ season and long-term time trends
- Investigate the remaining short-term variations by adding the exposure variable into the model; additional adjustment of other (measured) time-varying confounders may be needed.
- Check for **autocorrelation** to make sure that the assumption of independent residuals is not violated.



Daily cardiovascular disease mortality and temperature; Schematic based on Armstrong (2006) Epidemiology



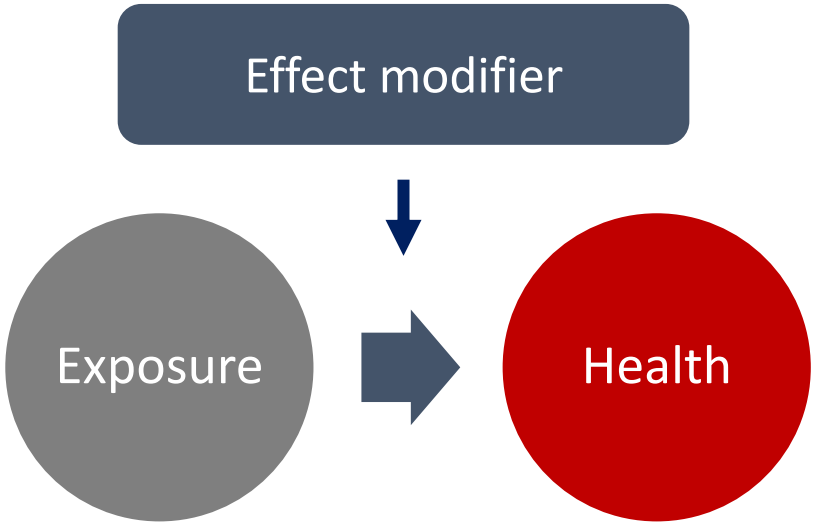


Bhaskaran et al. (2013) International Journal of Epidemiology

Time-invariant (within a short interval) factors

Personal behaviors/characteristics ??

- Assumed they reflect long-term effects
- Not considered as a confounder
- Not expected to be associated with the short-term temporal change in environmental exposure
- Possible to consider them as an effect modifier through stratified analysis



Pros and cons

Advantages

- Cheap and easy to apply
- Free from time invariant confounding within a short time period

Limitations

- Possibility of misclassifying exposure due to a few monitoring sites
  - ✓ It is assumed that people living in the same location were exposed to the same level of exposure aggregated within the location.
- Individual variability in sensitivity cannot be studied

How to quantify the short-term exposure-response association, after adjusting for the time-varying confounders?

1. Linear assumption for the association
2. Non-linear exposure-response associations

Time-series regression modelling

Let's look at a **linear** association, first.

$$Y_t \sim Poisson\,(\mu_t)$$
$$\log \mu_t = \alpha + \beta x_t + \boldsymbol{\eta}' \boldsymbol{z_t} + f(t;\lambda) + \varepsilon_t$$

Assume a Poisson distribution because the outcome is a count (log-linear Poisson model)

$\boldsymbol{Y_t}$ observed count of the relevant health outcome (e.g., mortality) on day t

$\boldsymbol{\mu_t}$ expected count of the relevant health outcome

$\boldsymbol{\beta}$ the log-relative risk associated with the **exposure of interest** x_t (e.g., air pollutant)

$\boldsymbol{z_t}$ a vector of **measured covariates** that we want to adjust for directly in the model, i.e., time-varying confounders (e.g., ambient temperature, humidity, etc.)

$\boldsymbol{f(t;\lambda)}$ the smooth function of time to capture a baseline trend of **unmeasured time-varying confounders**, changing slowly over time (e.g., seasonal and long-term time trends)

Key model assumptions and practical solutions

Common violations in the raw time-series count data

Autocorrelation

- Nearby observations more similar than distant ones
- Violation of independence assumption
- Substantially reduced after adjusting for seasonality and long-term trends
- Verify residual autocorrelation via checking diagnostic plots, e.g.:

✓ Scatter plot of deviance residuals vs. time

✓ Partial autocorrelation function (PACF) plot of deviance residuals

Overdispersion

- **Variance** of outcome counts **higher** than expected under Poisson
- Requires a simple adjustment

$$var(Y_t) = \boldsymbol{\varphi \mu_t}$$

- μ_t expected count of the health outcome
- $\boldsymbol{\varphi}$ over-dispersion parameter

→ Commonly referred to as a **quasi-Poisson** time-series regression model

Bhaskaran et al. (2013) International Journal of Epidemiology

Risk estimates, assuming a linear association

$$Y_t \sim Poisson\,(\mu_t)$$
$$\log \mu_t = \alpha + \beta x_t + \boldsymbol{\eta}' \boldsymbol{z_t} + f(t;\lambda) + \varepsilon_t$$

With the log-linear Poisson distribution,

β the effect estimate for the exposure of interest (**log-relative risk**)

- In other words, it is the expected increment of log(death) per a unit increase of the exposure of interest on a single day.

Mortality risk ratio (RR) = exp(β)

- RR represents mortality risk changes per **a unit increase** of the exposure on a single day

Let's look at how the RR works.

- RR = ratio of two mortality risks
- logRR = β; RR = exp(β)

- Suppose that we examine an association between ozone (denoted as 'Z') and mortality.

$$\log(\text{death; when } Z+1) - \log(\text{death; when } Z) = \beta \times (Z+1) - \beta \times Z$$

- Recall that subtracting logB from logA results in a ratio, i.e., log(A/B).

$$\log(\ (\text{death; when } Z+1) \ / \ (\text{death; when } Z)\ ) = (\beta \times Z) + \beta - (\beta \times Z) = \beta$$

- After exponentiating both sides,

$$RR = (\text{death; when } Z+1) \ / \ (\text{death; when } Z) = \exp(\beta)$$

Risk estimates, assuming a linear association

Mortality risk ratio **(RR) = exp(β)**

A 1-unit increase in exposure may represent only minor changes, so many studies present the relative risk (RR) per 10-unit increase or per an interquartile range (IQR) increase in exposure levels. The larger unit can be directly multiplied by the effect estimates, or alternatively, a scaled exposure variable can be incorporated into the time-series regression model.

- If interpreting RR per **10 unit** increase of the exposure, simply exp($\beta \times 10$)
- As for **95% CI**,

  exp($\beta \pm 1.96 \times se$) per a unit; exp($10 \times (\beta \pm 1.96 \times se)$) per 10 unit
- Possible to scale the variable (x'=x/10) and interpret β per 10-unit increase

In addition to the relative risk (RR), reporting the percent change (%) is also a useful way to quantify the association.

- If interpreting it as the **percent change** (PC),

  $\mathrm{PC}\ (\%) = 100 \times (\exp(\beta) - 1)$, representing the percent change in mortality per unit increase in the exposure
- If interesting the PC (%) **per an IQR increase** in the exposure,
  - $\mathrm{PC}\ (\%) = 100 \times (\exp(\beta \times IQR) - 1)$
  - $95\%\ \mathrm{CI} = 100 \times (\ \exp(\ (\beta \pm 1.96 \times se) \times IQR\ ) - 1\ )$

An example

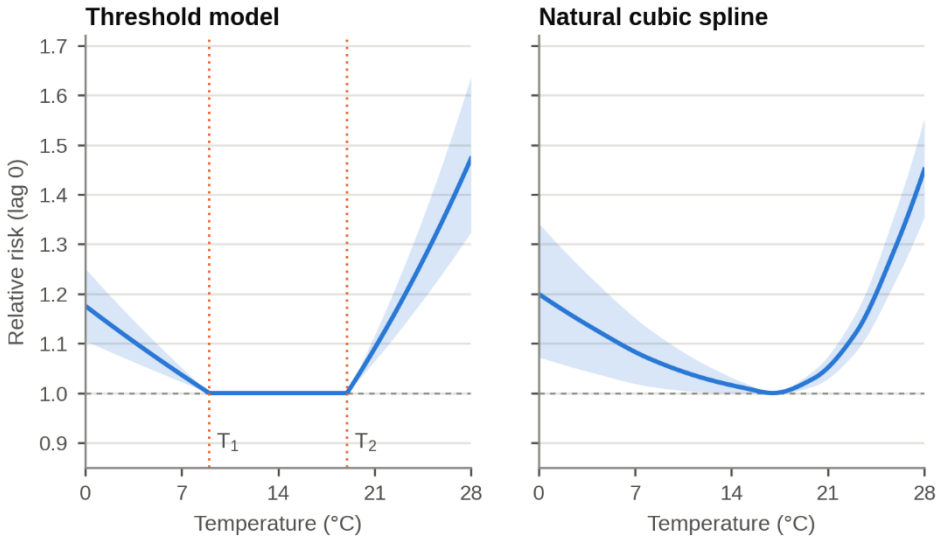
Suppose that we quantify the RR of mortality per every 10-unit increase in ozone.

1. Ozone variable scaled to (Z′=Z/10)
2. Let's say, the effect estimates for the scaled ozone are:
   - β=0.006608 and se=0.001599
3. exp(β)=RR per 10 $\mu g/m^3$ = 1.00663, approximately 1.007
4. exp($\beta \pm 1.96 \times se$)=95% CI per 10 $\mu g/m^3$ = 1.003481 to 1.00979
5. Percent change (%) = 100 x (RR-1) = 0.7% (95% CI, 0.3 to 1.0) risk increase per 10 $\mu g/m^3$ of ozone

- Small effect size but large impact if entire populations are exposed

How to quantify **nonlinear** risks (heat & cold)

0. **Linear** model
1. **Dummy** variable approach
   - How to define the dummy variables?
2. **Nonparametric smoothing**
   - Pros: intuitive graphics
   - Cons: How to define a risk??
3. **Piecewise linear model**
   - Correlation between threshold & slope (effect)
   - Can be extended to piecewise polynomial model
4. **Spline model**
   - Comparing two risks between two specific temperatures
   - Possible to use quantiles of temperature

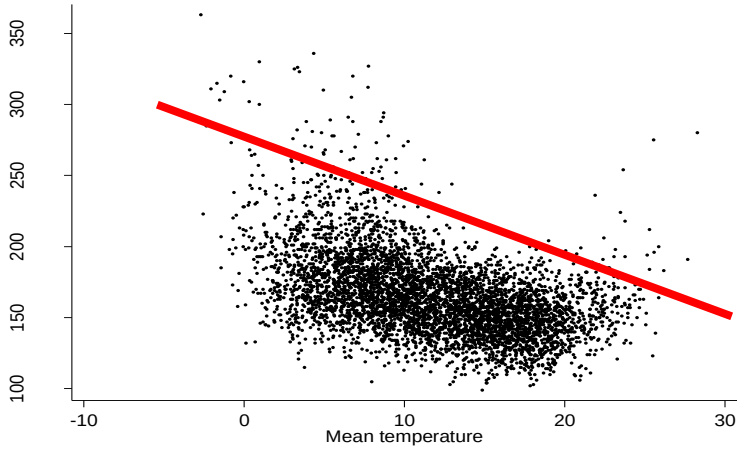


Thresholds and natural cubic spline models for lag 0 temperature (Cardiovascular disease deaths over a 10 year period)

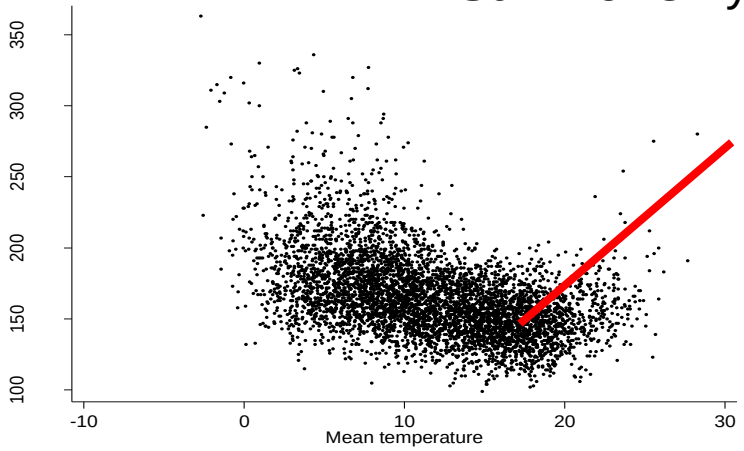
Schematic based on Armstrong (2006) Epidemiology

Use a linear model for the temperature-mortality association?

- Not right !
- Stratified by season (or month)
  - often used in the case-crossover settings



Summer Only



Piecewise linear model (a.k.a. linear thresholds model)

- Identifying a reasonable **threshold** and estimate **a heat slope** above the threshold

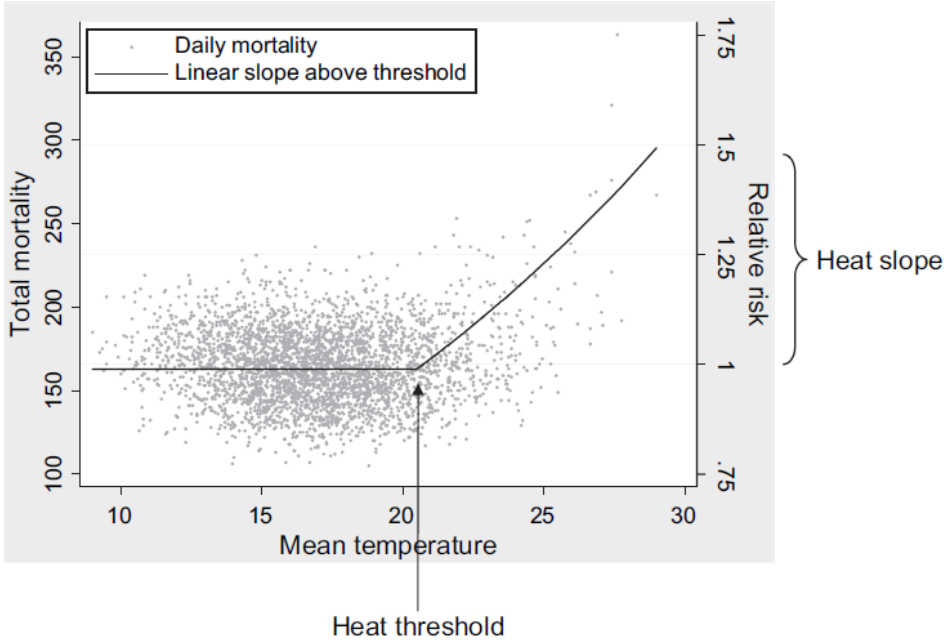
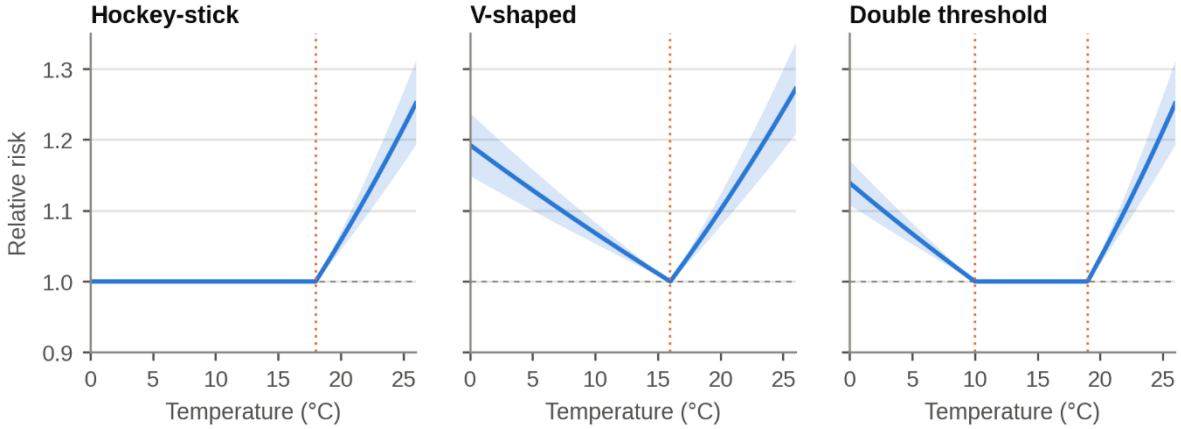


Figure 1 Relationship between daily mortality and summertime daily mean temperature in London, 1976–2003.

Hajat and Kosatky (2010) J Epidemiol Community Health

Armstrong (2006) Epidemiology

1. **Hockey-stick model** -- a special case of the linear thresholds model, when a heat slope or a cold slope is equal to zero
2. **V model** -- a special case of the linear thresholds model, where the two thresholds are the same
3. **Double thresholds model**



Schematic based on Armstrong (2006) Epidemiology

Spline model

$$y_t \sim Poisson\ (\lambda_t)$$

$$\text{Log}\ \lambda_t = \alpha_0 + f(\boldsymbol{x}_t; \boldsymbol{\beta}) + \sum_{p=1}^{P} h_p\left(z_{pt}; \boldsymbol{\gamma}_p\right) + s(t; \boldsymbol{\theta})$$

Purpose: To examine a nonlinear short-term association between temperature and mortality

Outcome: Daily death counts → quasi-Poisson distribution (allowing for overdispersion)

Exposure of interest: ambient temperature

β the effect estimate for the exposure of interest

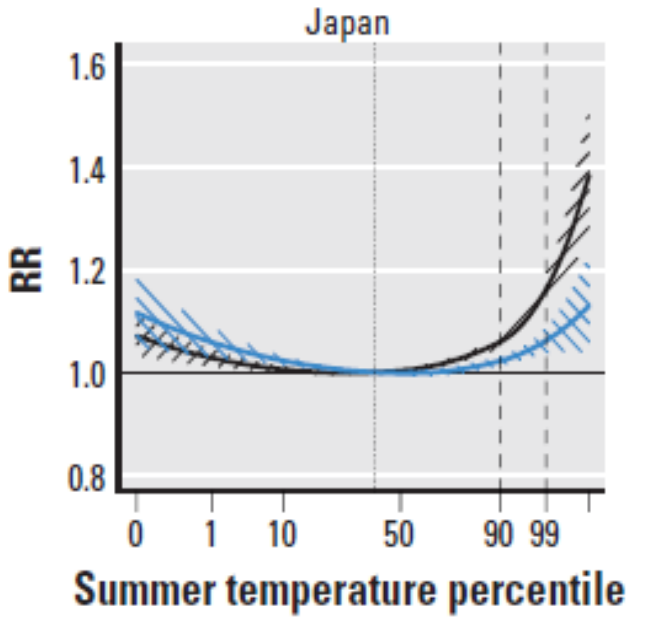
z_{pt} measured time-varying confounders

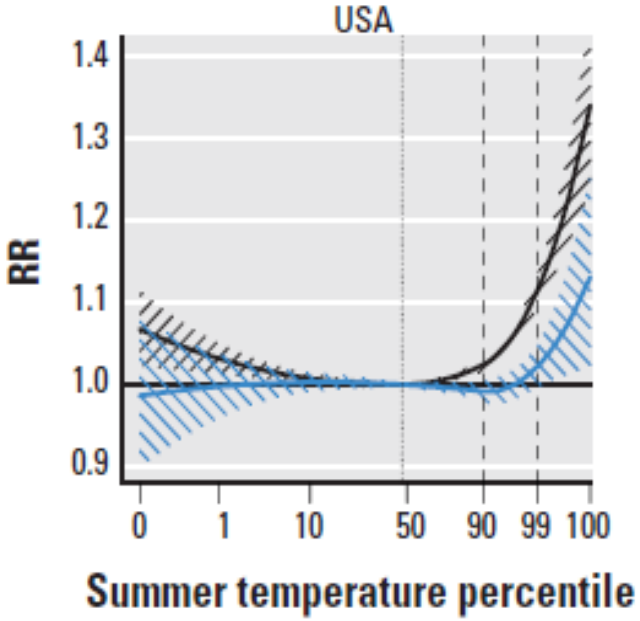
h_p smooth functions for each *p*

$s(t; \boldsymbol{\theta})$ baseline trend that captures unmeasured time-varying confounders

Analytical procedure

1. Generate a set of **basis variables** for the exposure of interest using a spline function
2. Incorporate the basis variables into the log-linear Poisson **model**
3. **Predict** the expected outcome values (risks) for each of exposure levels
   - Consider what level of exposure you want to use as a reference for relative risks (RRs)
4. Calculate the **RRs and 95% CIs** based on standard errors (SE)
   - **By comparing two risks** of mortality between two specific temperatures





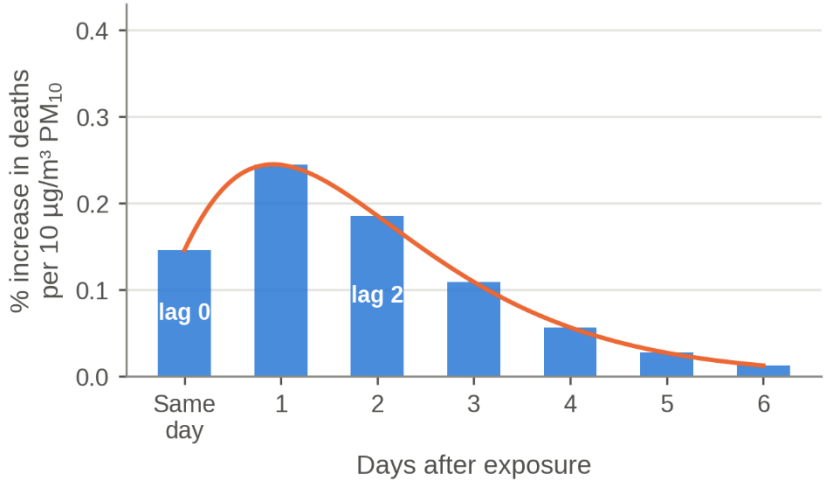
Gasparrini et al. (2015) Environmental Health Perspectives
Black line = 1993; Blue line = 2006

Delayed effects

Lag: a time period between changes in the exposure and the consequent changes in the outcome

- "lag 0 effect" –the same day effect
- "lag 2 effect"
  - the effect of the exposure two days before on the current day health outcome, or
  - the effect of the current day exposure on the health outcome two days later

An example of the hypothesized curve, representing the risks of death after exposure to PM_{10}



Schematic based on Schwartz (2000) Epidemiology

Summary

The **basic principles** of the time-series study design

- Time-varying **confounders** (measured and unmeasured)
- The techniques of adjusting for seasonality and long-term **time trend**
- **Other essentials to be considered** when using the time-series regression models
  - Overdispersion
  - Autocorrelation of the residuals and the trend adjustments
  - Sensitivity analysis (model selection and choice of *df* for time trend and/or covariates)
  - Outliers, influential points, and missing values

Quantifying and interpreting the short-term exposure-response association

- Risk estimates assuming the linear or nonlinear association

References

- Bell ML and Davis DL (2001). Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution. Environ Health Perspect 109:389-394.
- Bhaskaran et al. (2013) International Journal of Epidemiology 42:1187–1195. Link
- Gasparrini A, et al. (2015) Mortality risk attributable to high and low ambient temperature: A multicountry observational study. The Lancet 386:369-375.
- Gasparrini A. et al. (2018) International Society of Environmental Epidemiology (ISEE) 2018 Pre-conference workshop. Advanced modeling techniques for time series analysis using R.
- Hajat S and Kosatky T (2010) Heat-related mortality: A review and exploration of heterogeneity. Journal of epidemiology and community health 64:753-760.
- Kim H et al. (2010) International Society of Exposure Science and International Society of Environmental Epidemiology Joint meeting (ISES-ISEE) 2010 Pre-conference workshop. Statistical methods for evaluating air pollution and temperature effects on human health. http://hosting03.snu.ac.kr/~hokim/isee2010/
- Kim H, Honda Y, Hashizume M, et al. (2016) International Society of Exposure Science and International Society of Environmental Epidemiology Joint meeting Asian Chapter (ISES-ISEE AC) 2016 Pre-conference workshop. Time-series regression analysis in environmental epidemiology: concepts of DLNM (distributed lag non-linear model) and its application. http://hosting03.snu.ac.kr/~hokim/iseeac2016/
- Peng RD and Dominici F. (2008) Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health. Springer.
- Peng RD et al. (2006). Model choice in time series studies of air pollution and mortality. Journal of the Royal Statistical Society: Series A 169(2):179-203.
- Reid CE et al. (2012) The role of ambient ozone in epidemiologic studies of heat-related mortality. Environ Health Perspect 120(12):1627-1630.
- Schwartz J. (2000). The distributed lag between air pollution and daily deaths. Epidemiology 11(3):320-326.