Introduction to environmental time-series analysis

Aurelio Tobías

Spanish Council for Scientific Research

Nagasaki University School of Tropical Medicine and Global Health

XIV Summer School UPC-FME

Barcelona, 2021/06/29

Outline

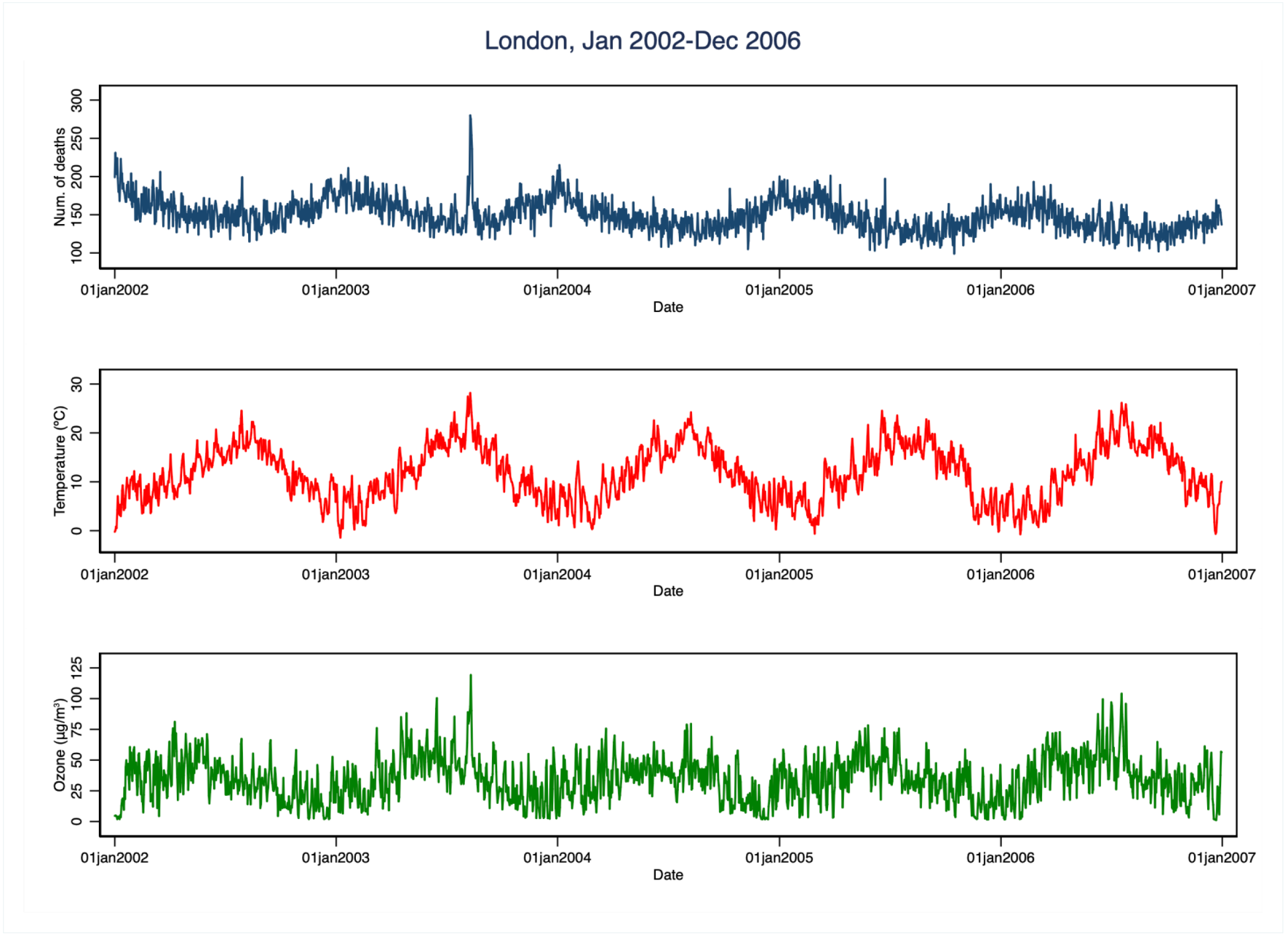
- Introduction
- Time-series design
- Exposure-response
- Lagged effects
- Summary

Introduction

- **1952** – Air pollution episode (fog) in London
- **1990's** – Time-series regression in air pollution multi-centre studies in EU and US (Schwartz et al. 1990, Katsouyanni et al. 1996, Samet et al. 2000)
- **2000's** – Generalised Additive Models (Hasite and Tibshirani 1990) in **Splus** (`gam` function)
- **2003** – Heatwave in Europe
- **2010** – Distributed Lag Non-Linear Models (Gasparrini et al. 2010) in **R** (dlnm library)

Introduction

- Health outcomes (e.g., mortality, hospital admissions) and environmental exposures (e.g., temperature, air pollution) are characterized by similar time-trends
- Measures of individual predictors are usually not available
- We need a study design that relies on between-day comparison within the same population and able to control for time-trends


London, Jan 2002-Dec 2006
Num. of deaths
100
150
200
250
300
01jan2002
01jan2003
01jan2004
01jan2005
01jan2006
01jan2007
Date
Temperature (°C)
0
10
20
30
01jan2002
01jan2003
01jan2004
01jan2005
01jan2006
01jan2007
Date
Ozone (µg/m³)
0
25
50
75
100
125
01jan2002
01jan2003
01jan2004
01jan2005
01jan2006
01jan2007
Date

Time-series data

- A time-series is a sequence of measurements equally spaced through time *(Zeger et al. 2006)*
- The unit of analysis use to be the day (t), not the individual person (i)
- But it could be annual, monthly, weekly or hourly
- The outcome is a count (e.g., number of deaths)

- Example
  - First 10 rows of time-series data (London, Jan 2002 – Dec 2006)

obs	date	deaths	temp	ozone
1.	01jan2002	199	-0.2	4.6
2.	02jan2002	231	0.1	4.9
3.	03jan2002	210	0.9	4.7
4.	04jan2002	203	0.5	4.1
5.	05jan2002	224	4.2	2.0
6.	06jan2002	198	7.1	2.4
7.	07jan2002	180	5.2	4.1
8.	08jan2002	188	3.5	3.1
9.	09jan2002	168	3.2	2.1
10.	10jan2002	194	5.3	5.2

Time-series design

- **Research question** – “Is there an association between day-to-day variation in the environmental exposure (X_t) and daily risk of health outcome (Y_t)”?
- Series should be long enough to identify day-to-day variation necessary to disentangle short-term effects from time-trends (e.g., at least 3 consecutive years for daily data)

Time-series design

- Strengths
  - Use of administratively-collected data
  - Same population is compared with itself – focus is day to day variation
  - Time-invariant or slowly-varying individual risk factors controlled by design (e.g., age, gender, smoking)
- Limitations
  - Ecological design based on aggregated, not individual data
  - Sensitive to choices for modelling time-trends
  - Not applicable to estimate log-term (chronic) effects

Time-series regression

- Similar in principle to any regression analysis but with some specific features
- Poisson regression

  $\mathbf{Y|x \sim Poisson(\mu)}$

  $\mathbf{E(Y|x) = \mu}$

  $\mathbf{\log(\mu) = \alpha + \beta x}$

  with $\mathbf{V(Y|x) = \mu}$

- Why not to use ARIMA?
  - Lack of knowledge/training
  - Epidemiological interpretability
  - Non-normal distribution for cause-specific health outcomes

Poisson regression (technicalities)

- Measure of effect
  - $\log(\mu) = \alpha + \beta x$

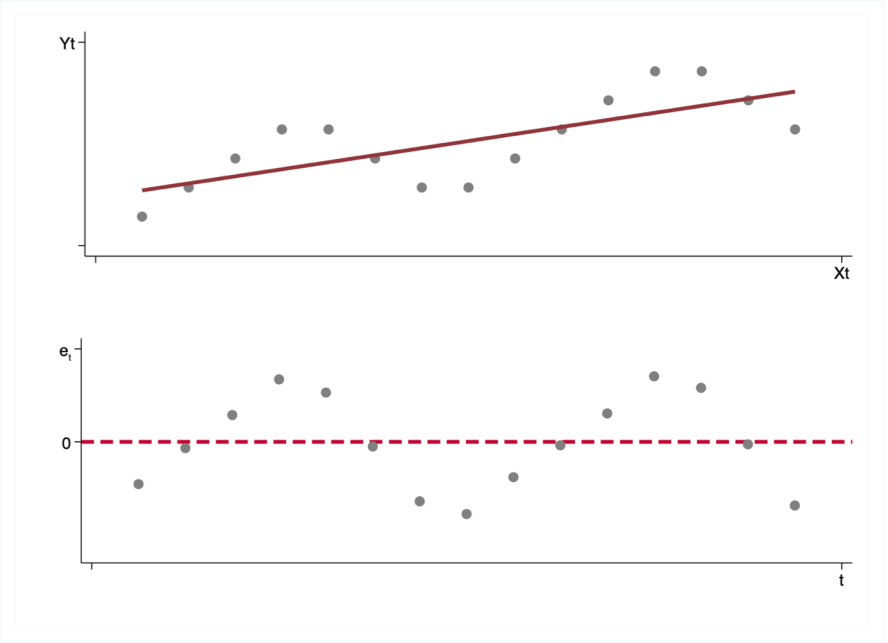
    $\exp(\alpha) = \mu_0$ when x = 0

    $\exp(\beta) = \mu/\mu_0 \Rightarrow$ Relative Risk (RR) for 1 unit increase of x
  - Percentage increase of risk as, (RR-1)×100%
  - However, we often find data that exhibit today's population size in a city does not change much from yesterday's population
  - So denominator (underlying population size) is not a big concern

- Overdispersion
  - However, we often find data that exhibit over-dispersion, with $V(E|x)$ larger than $E(Y|x)$
  - It affects standard errors and should be taken into account to make valid inferences
  - Quasi-Poisson, with $V(Y|x) = \phi\mu$

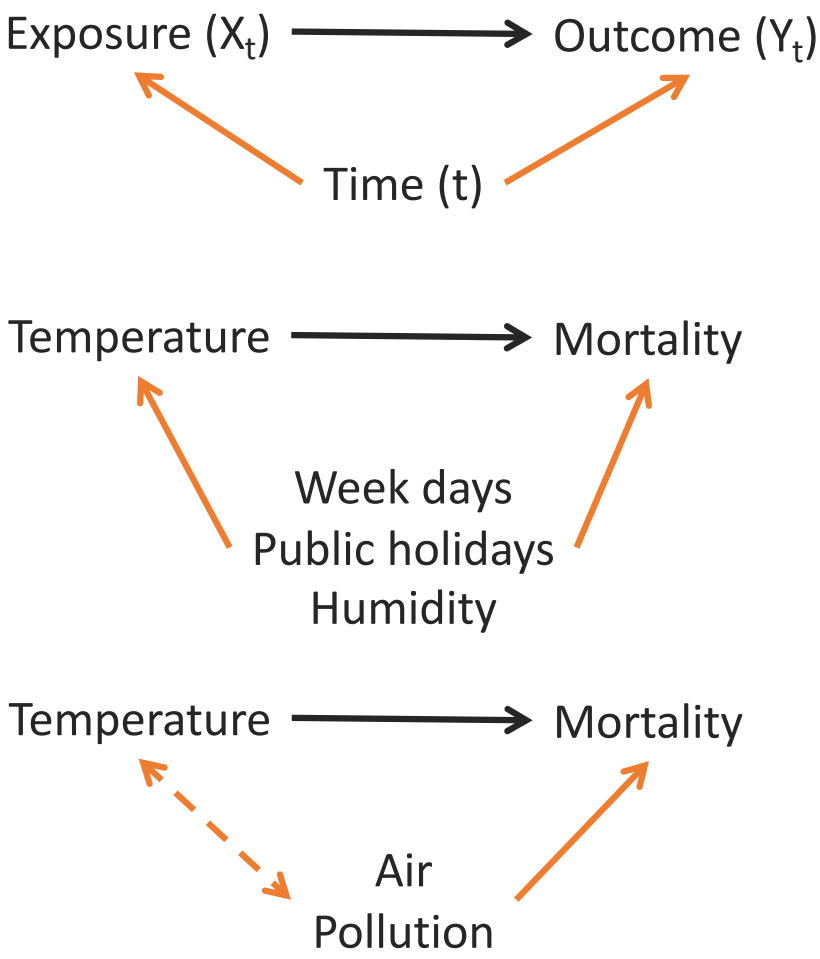
Autocorrelation

- The key assumption of any regression model is that the outcome measures are independent
- With time-series, closer events tend to be more similar than those further apart in time
  - It affects standard errors and should be taken into account to make valid inferences
  - After controlling for time-trend, residual variation without temporal pattern **(e_t ~ iid)**

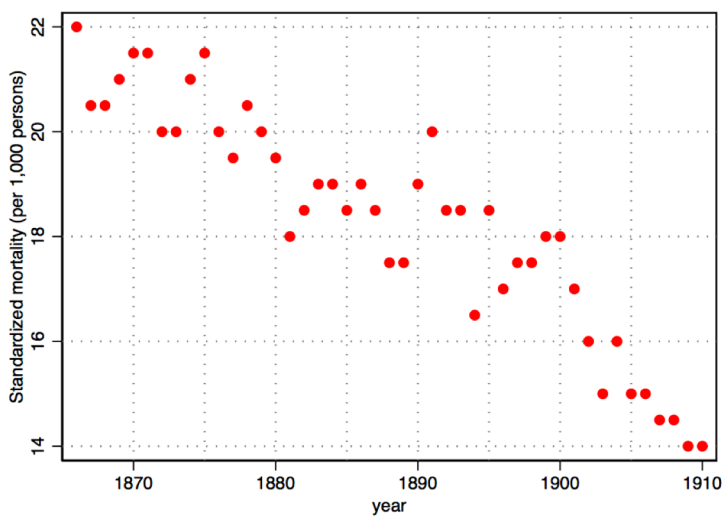


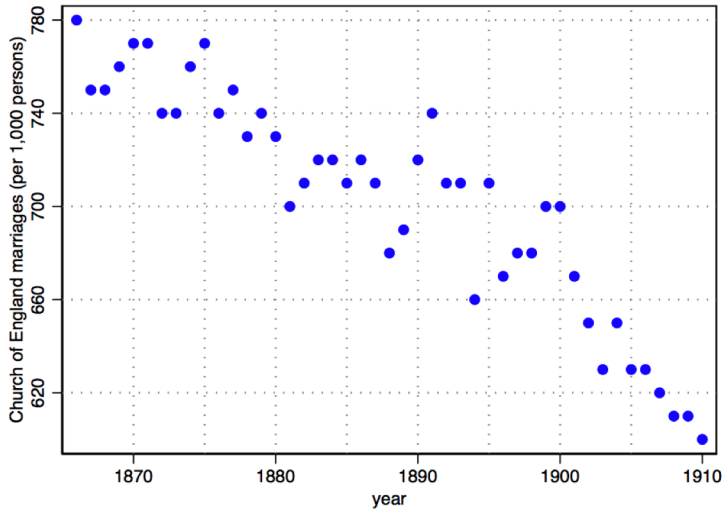
Confounding

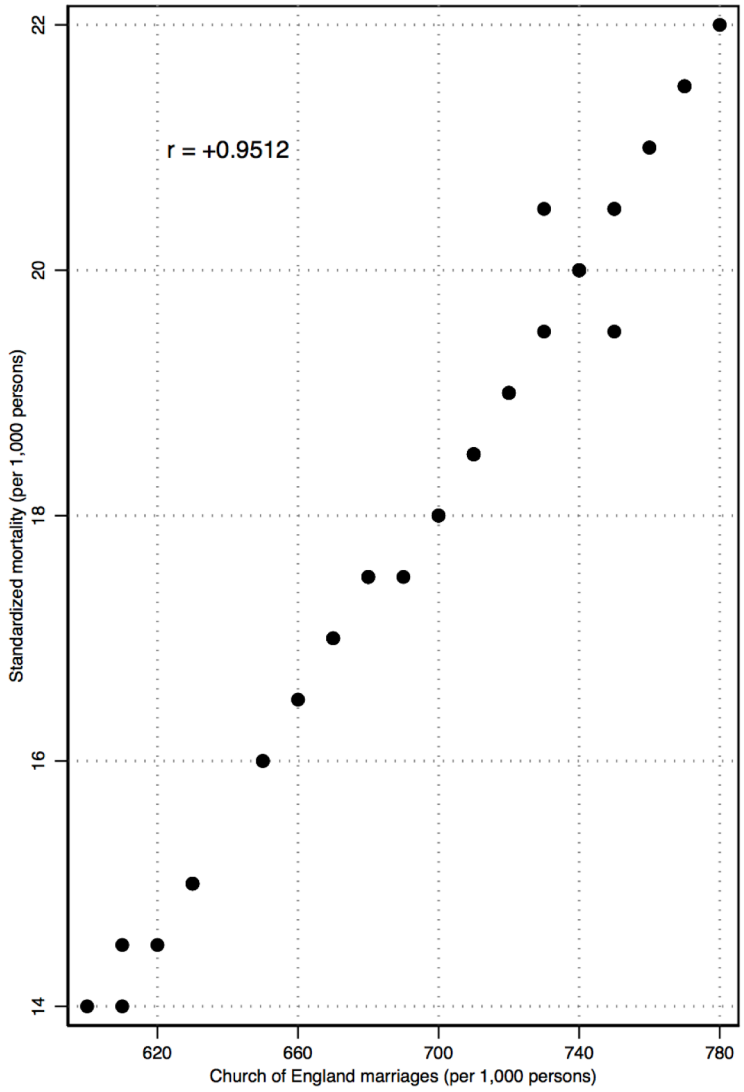
- It must be associated with the exposure (X) being investigated
- It must be independently associated with the outcome (Y) being investigated
- It must not be on the causal pathway between exposure (X) and outcome (Y)



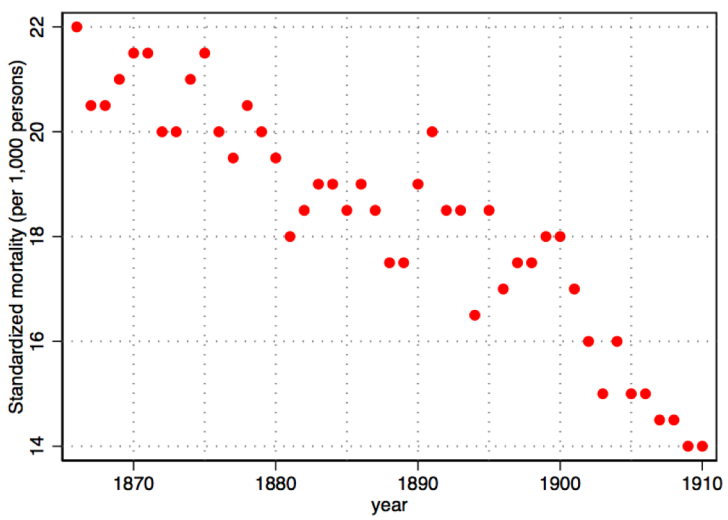
Yule GU. **Why do we sometimes get nonsense correlations between time series?** J Royal Stat Soc Sci. 1926;89:1-64.

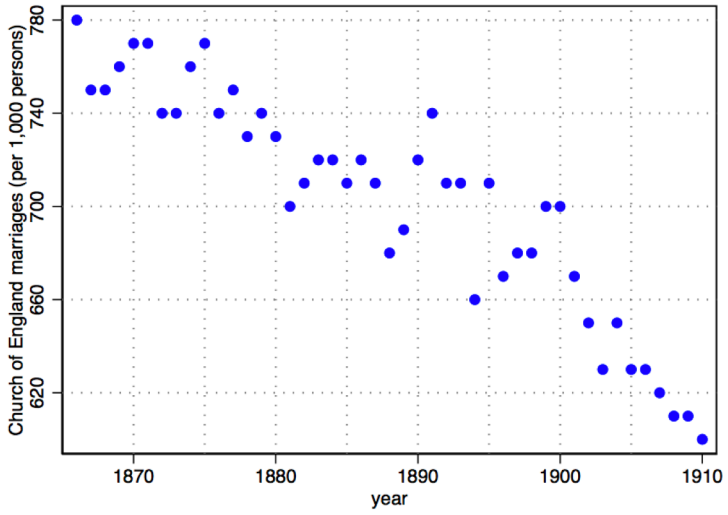


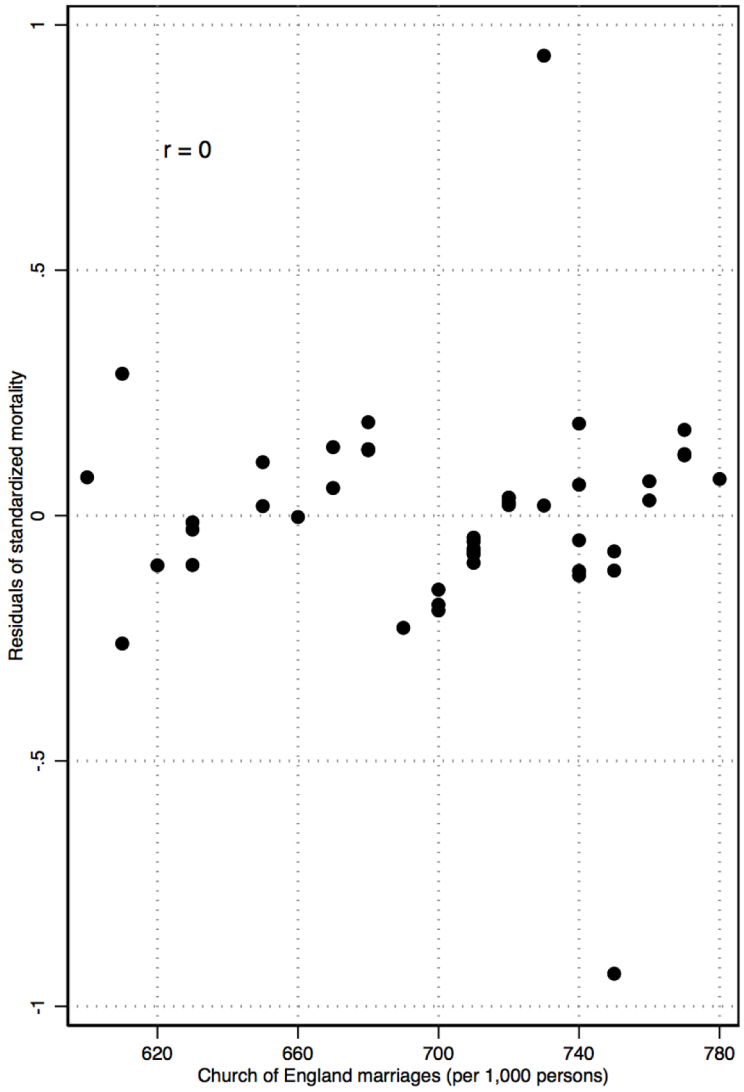




Yule GU. **Why do we sometimes get nonsense correlations between time series?** J Royal Stat Soc Sci. 1926;89:1-64.





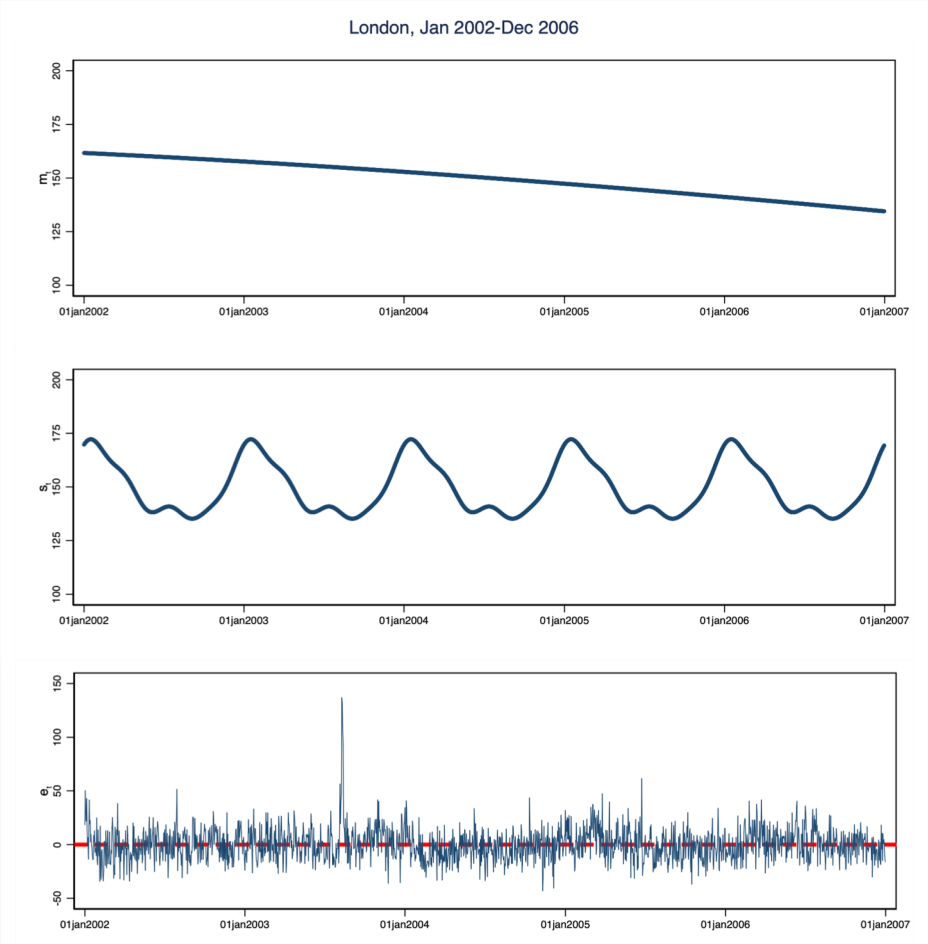


Modelling framework

- Temporal decomposition,

$$Y_t = m_t + s_t + e_t$$

  - *With m_t and s_t as time components (long trend and seasonality) and e_t as residual series*
- Underlying trends are filtered-out from the time series, allowing the inspection of associations at shorter time scale



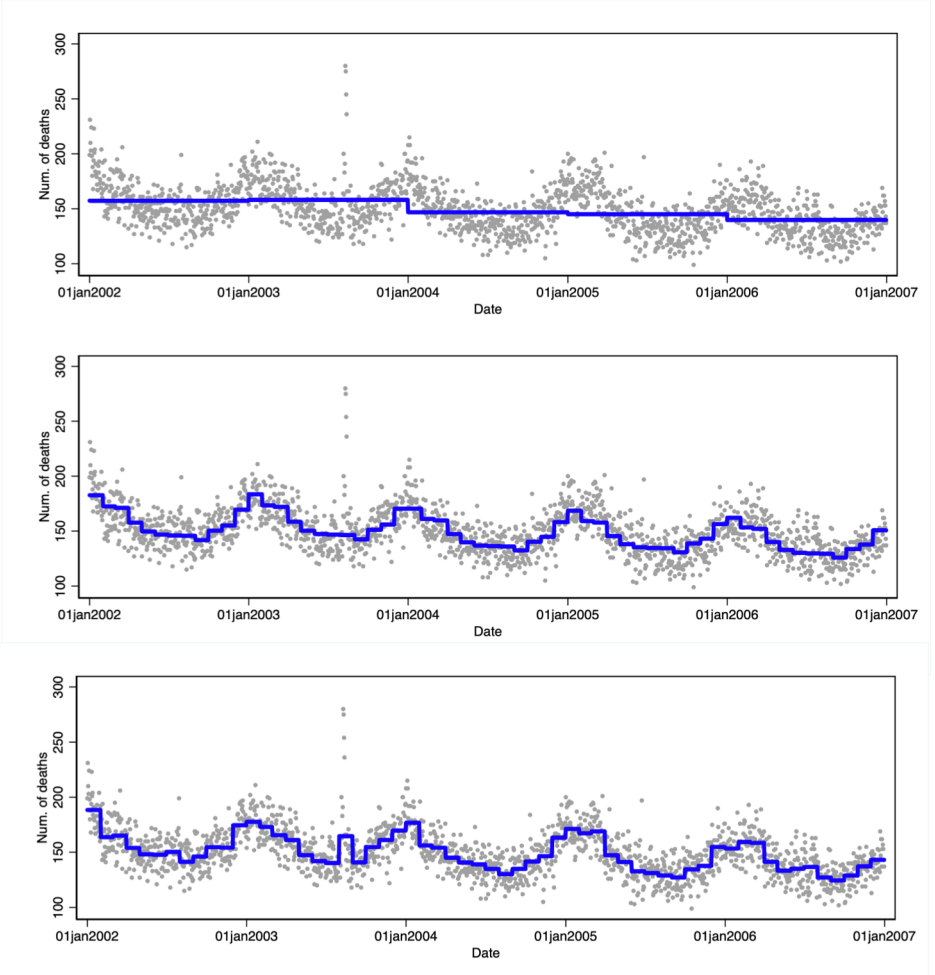
Time-stratified model

- Split the study period into time-intervals estimating a different baseline mortality risk
- Use of indicator variables for year and month

$$Y_t = \alpha + \sum \beta_i year_i + \sum \delta_j month_j$$

  - Easy to understand, and often captures main long-term patterns
  - Implicitly assumes biologically implausible jumps in risk between adjacent months

(Bhaskaran et al. 2013)

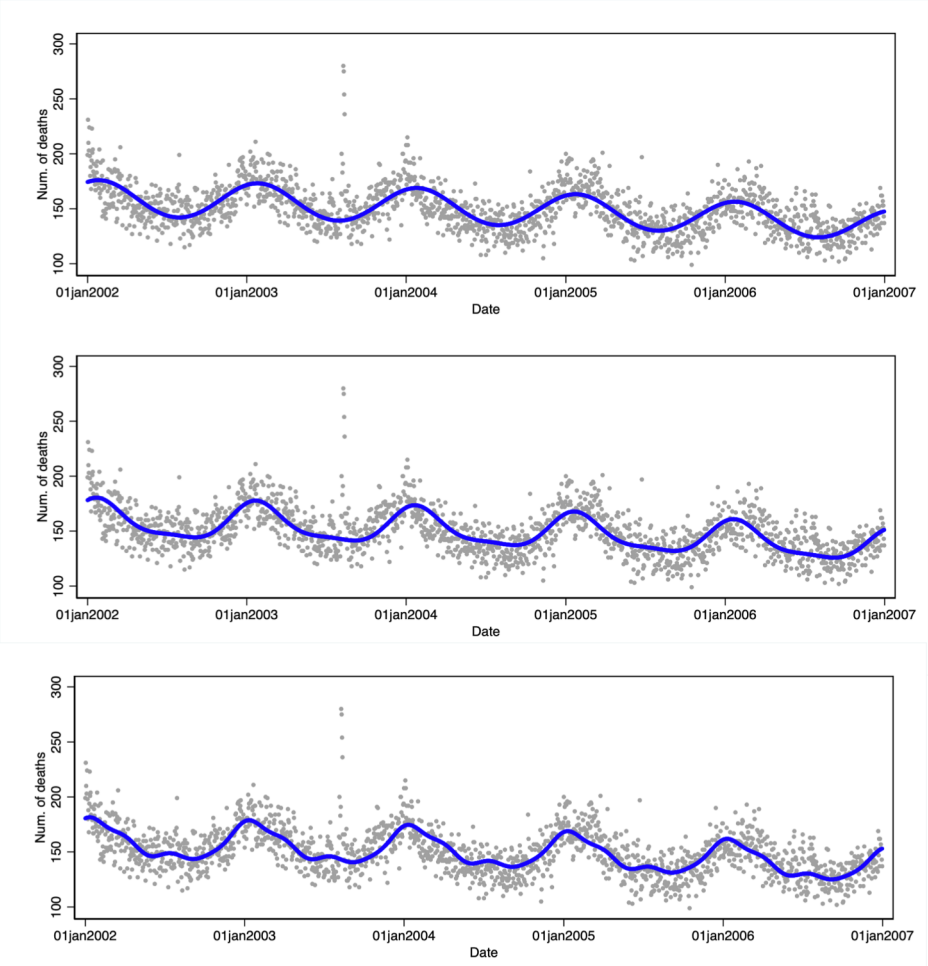


Periodic functions

- Fourier terms (pairs of sine and cosine functions of time) to model seasonal variation in the outcome as a regular wave each year

$$Y_t = \alpha + \beta t + \sum \delta_k \sin(^{k\pi t}/_{T}) + \sum \gamma_k \cos(^{k\pi t}/_{T})$$

  - Suitable to capture very regular seasonal patterns
  - The modelled seasonal pattern is forced to be the same for each year, which may not reflect the data well



(Bhaskaran et al. 2013)

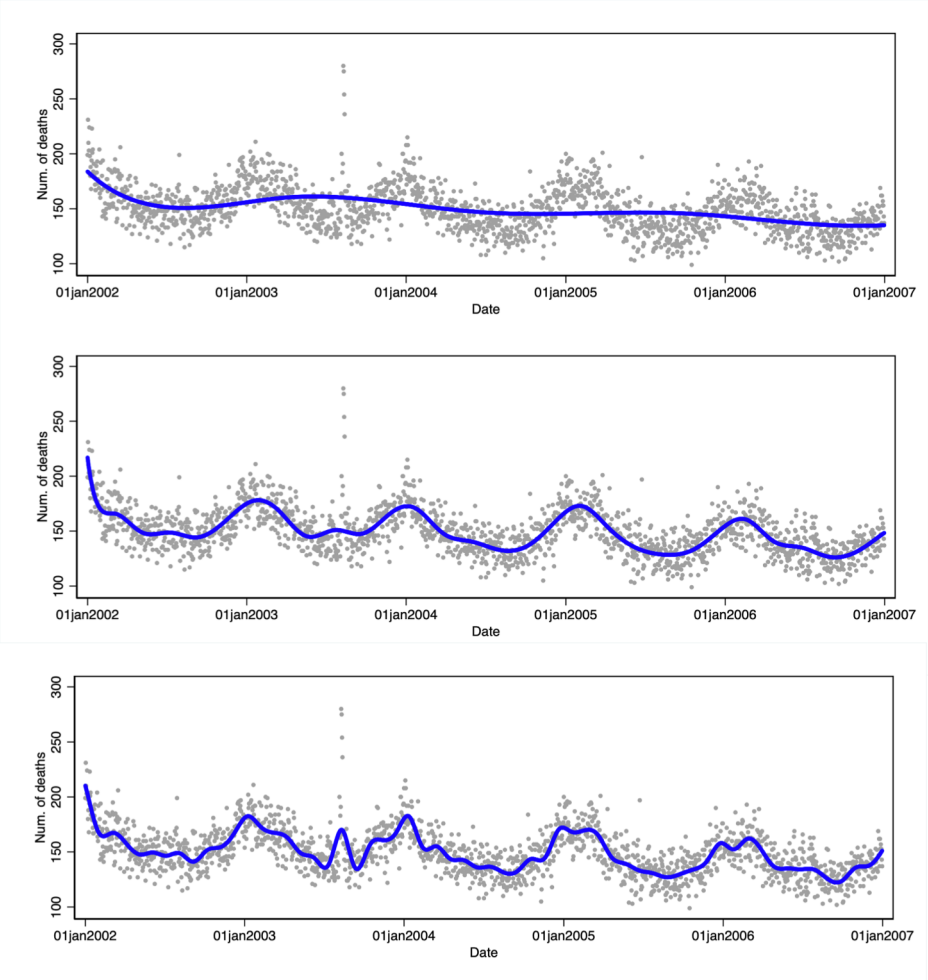
Spline function

- A set of polynomial curves (commonly cubic) that are joined smoothly end-to-end to cover the study period
- Basis variables are functions of the calendar time

$$Y_t = \alpha + f(t,\beta)$$

  - Capture seasonal patterns in a way allowed to vary from each year
  - It is necessary to decide how many knots there should be, which governs how flexible the curve will be

(Bhaskaran et al. 2013)



Autocorrelation and overdispersion

- Time-stratified model

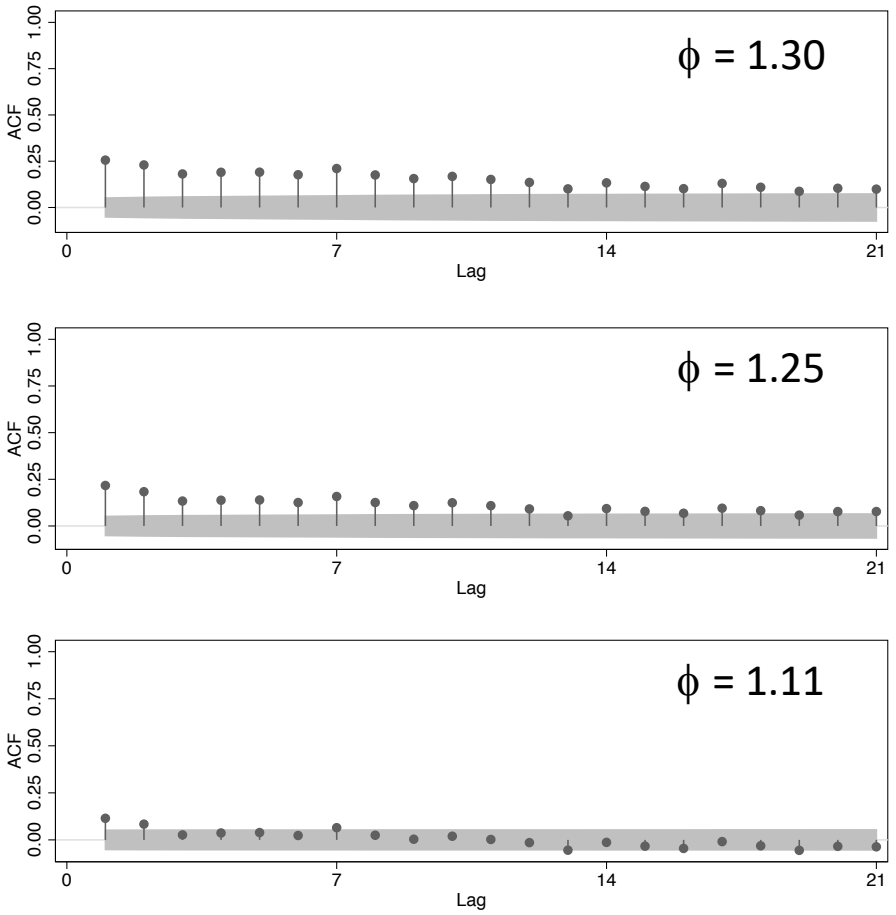
$$Y_t = \alpha + \sum\beta_i year_i + \sum\delta_j month_j$$

- Periodic functions

$$Y_t = \alpha + \beta t + \sum\delta_k sin(^{k\pi t}/_T) + \sum\gamma_k cos(^{k\pi t}/_T)$$

- Spline function

$$Y_t = \alpha + f(t,\beta)$$



Exposure-response and lagged effects

- Associations between environmental exposures and health outcomes
  - Can show different type of shapes, usually a linear association with air pollution and non-linear with temperature
  - Are often characterized by lagged effects
- We need to model potentially complex temporal patterns of risk due to time-varying exposures
- It requires a knowledge previous knowledge about the shape of the exposure-response function and the lagged effects

Exposure-response

- Linear association

$Y_t = \alpha + \beta x_t$

- Quadratic association

$Y_t = \alpha + \beta_1 x_t + \beta_2 x_t^2$

- Piecewise linear model

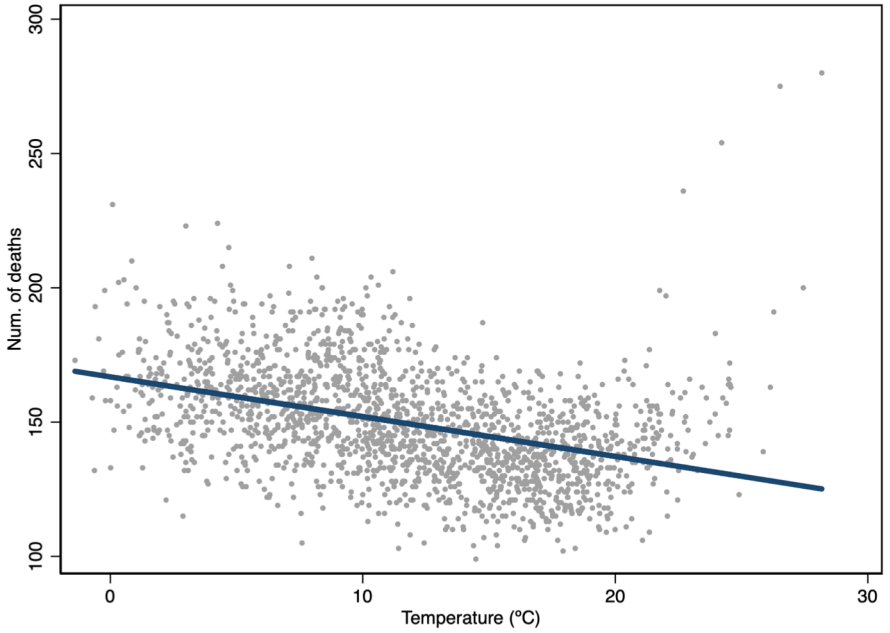
$Y_t = \alpha + \beta_C x_{Ct} + \beta_H x_{Ht}$

$x_C = max[(\tau_C - x_i), 0]$

$x_H = max[(x_i - \tau_H), 0]$

- Spline function

$Y_t = \alpha + f(x_t, \beta)$



- Unique effect as, β decreases of deaths for a 1ºC rise of temperature

Exposure-response

- Linear association

$Y_t = \alpha + \beta x_t$

- Quadratic association

$\mathbf{Y_t = \alpha + \beta_1 x_t + \beta_2 x_t^2}$

- Piecewise linear model

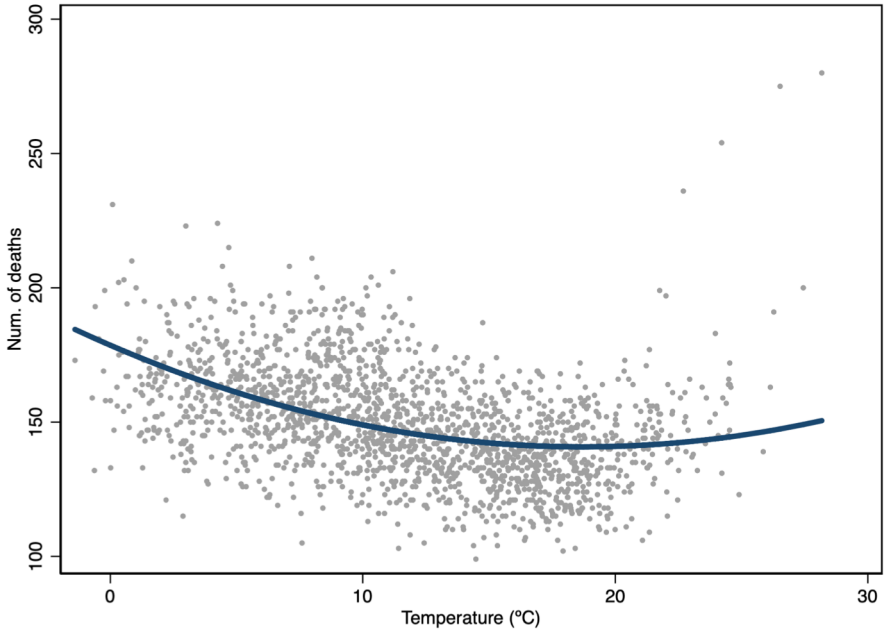
$Y_t = \alpha + \beta_C x_{Ct} + \beta_H x_{Ht}$

$x_C = max[(\tau_C - x_i), 0]$

$x_H = max[(x_i - \tau_H), 0]$

- Spline function

$Y_t = \alpha + f(x_t, \beta)$



- Multiple effects as, $\beta_1 + 2\beta_2 temperature$ increases of deaths for a 1ºC rise
- Minimum mortality temperature (MMT) at $-\beta_1/2\beta_2$

Exposure-response

- Linear association

$Y_t = \alpha + \beta x_t$

- Quadratic association

$Y_t = \alpha + \beta_1 x_t + \beta_2 x_t^2$

- Piecewise linear model

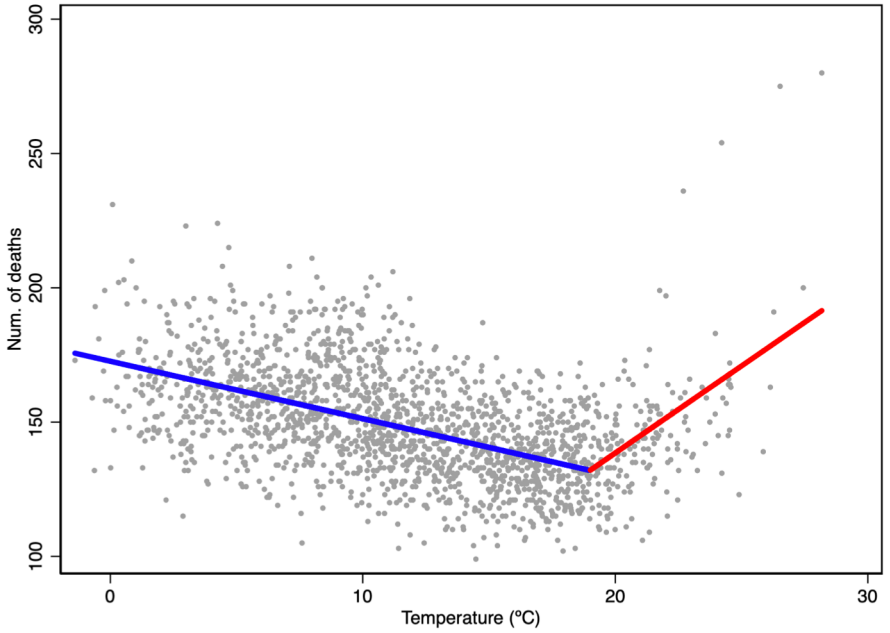
$\mathbf{Y_t = \alpha + \beta_C x_{Ct} + \beta_H x_{Ht}}$

$x_C = max[(\tau_C - x_i), 0]$

$x_H = max[(x_i - \tau_H), 0]$

- Spline function

$Y_t = \alpha + f(x_t, \beta)$



- Two effects as, β_1 decreases of deaths for a 1ºC rise below 18ºC (cold) and β_2 increases of deaths for a 1ºC rise over 18ºC (heat)

Exposure-response

- Linear association

$Y_t = \alpha + \beta x_t$

- Quadratic association

$Y_t = \alpha + \beta_1 x_t + \beta_2 x_t^2$

- Piecewise linear model

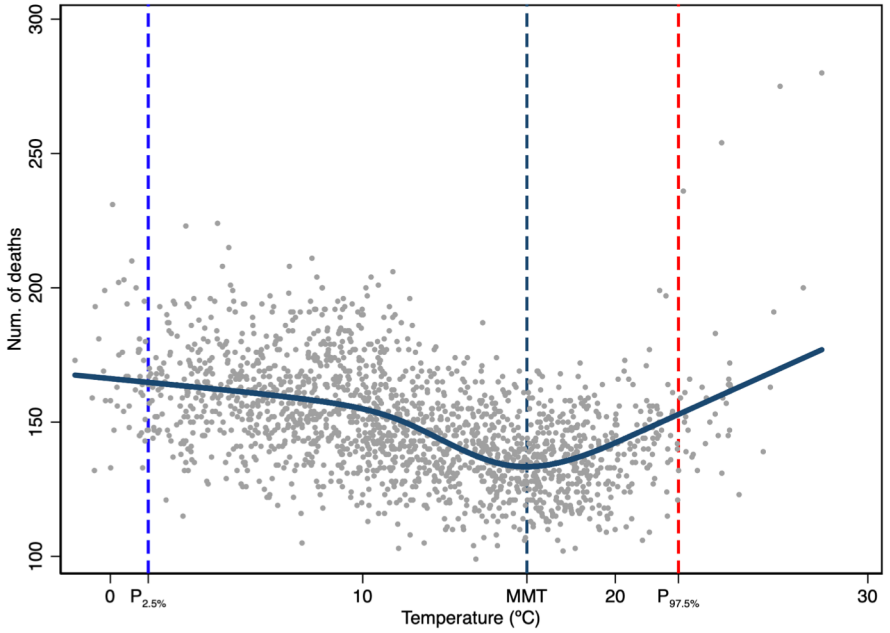
$Y_t = \alpha + \beta_C x_{Ct} + \beta_H x_{Ht}$

$x_C = max[(\tau_C - x_i), 0]$

$x_H = max[(x_i - \tau_H), 0]$

- Spline function

$Y_t = \alpha + f(x_t, \beta)$



- Multiple effects, comparing a low temperature ($P_{2.5\%}$) versus the MMT (cold) and a high temperature ($P_{97.5\%}$) versus the MMT (heat)
- MMT identified empirically

Lagged effects

- Independent

$Y_t = \alpha + \beta_j x_{t-j} \qquad \forall j=0, \ldots, k$

- Unconstrained

$Y_t = \alpha + \sum \beta_j x_{t-j}$

- Stratum

$Y_t = \alpha + \sum \beta_j x_{t-j}$

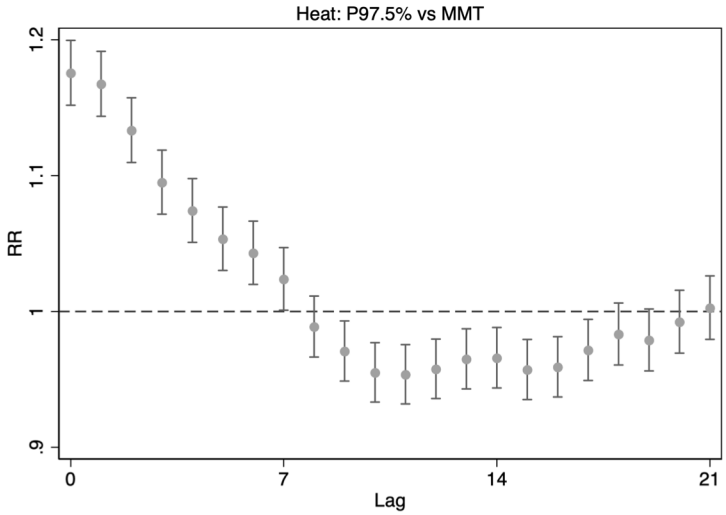
$\beta_1 = \beta_2 = \ldots = \beta_{l-k}$

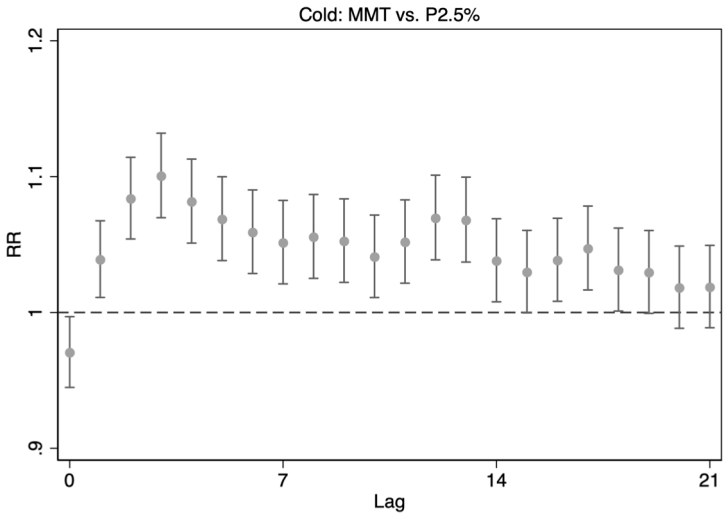
- Spline function

$Y_t = \alpha + \sum \beta_j x_{t-j}$

$\beta_j = f(\beta, j)$

(Armstrong 2006)





Lagged effects

- Independent

$Y_t = \alpha + \beta_j x_{t-j} \qquad \forall j=0, \ldots, k$

- Unconstrained

$\mathbf{Y_t = \alpha + \sum \beta_j x_{t-j}}$

- Stratum

$Y_t = \alpha + \sum \beta_j x_{t-j}$

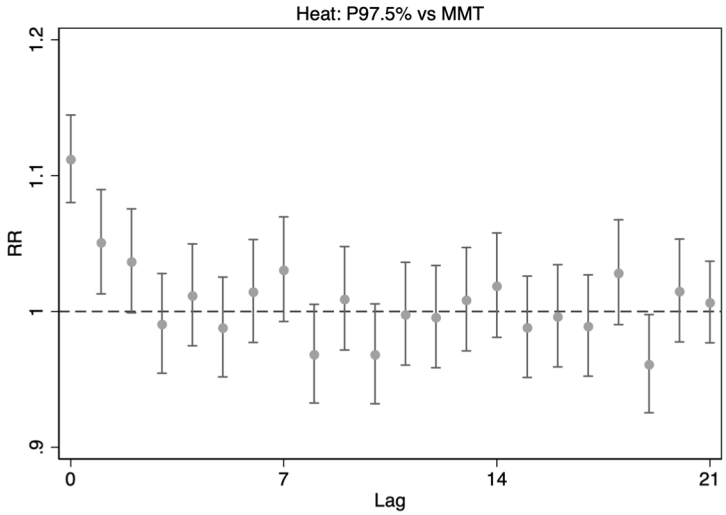
$\beta_1 = \beta_2 = \ldots = \beta_{l-k}$

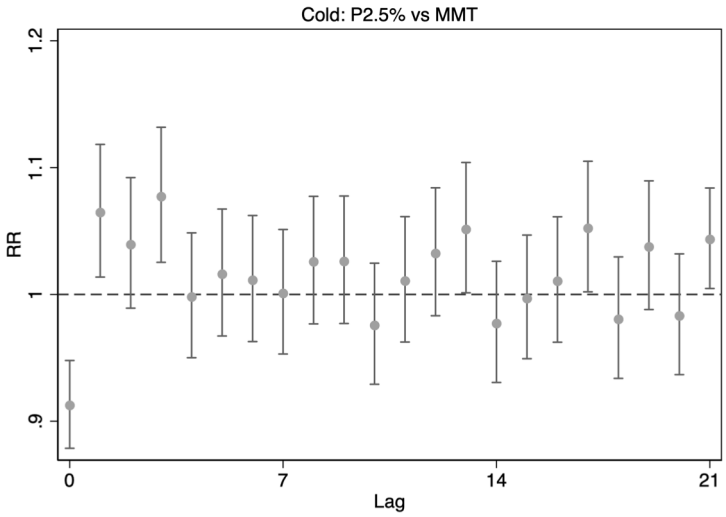
- Spline function

$Y_t = \alpha + \sum \beta_j x_{t-j}$

$\beta_j = f(\beta, j)$

(Armstrong 2006)





Lagged effects

- Independent

$Y_t = \alpha + \beta_j x_{t-j} \qquad \forall j=0, \ldots, k$

- Unconstrained

$Y_t = \alpha + \sum \beta_j x_{t-j}$

- Stratum

$\mathbf{Y_t = \alpha + \sum \beta_j x_{t-j}}$

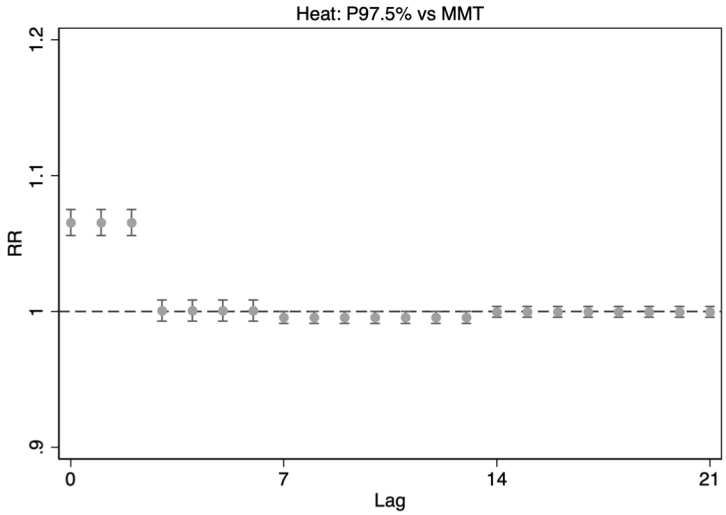
$\beta_1 = \beta_2 = \ldots = \beta_{l-k}$

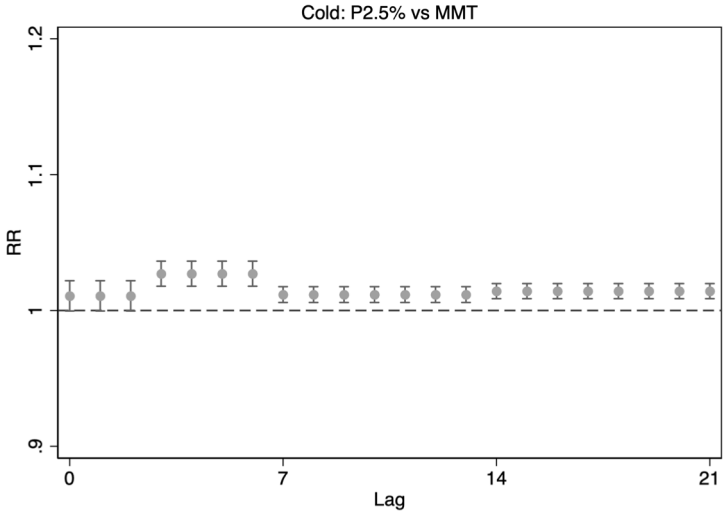
- Spline function

$Y_t = \alpha + \sum \beta_j x_{t-j}$

$\beta_j = f(\beta, j)$

(Armstrong 2006)





Lagged effects

- Independent

$Y_t = \alpha + \beta_j x_{t-j} \qquad \forall j=0, \ldots, k$

- Unconstrained

$Y_t = \alpha + \sum \beta_j x_{t-j}$

- Stratum

$Y_t = \alpha + \sum \beta_j x_{t-j}$

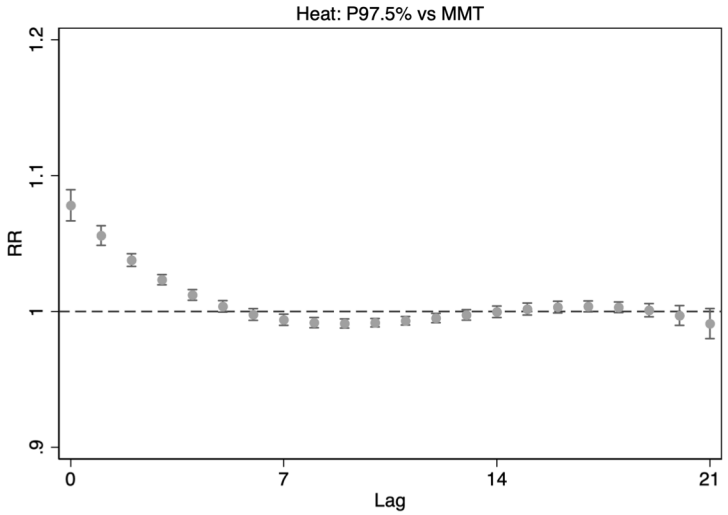
$\beta_1 = \beta_2 = \ldots = \beta_{l-k}$

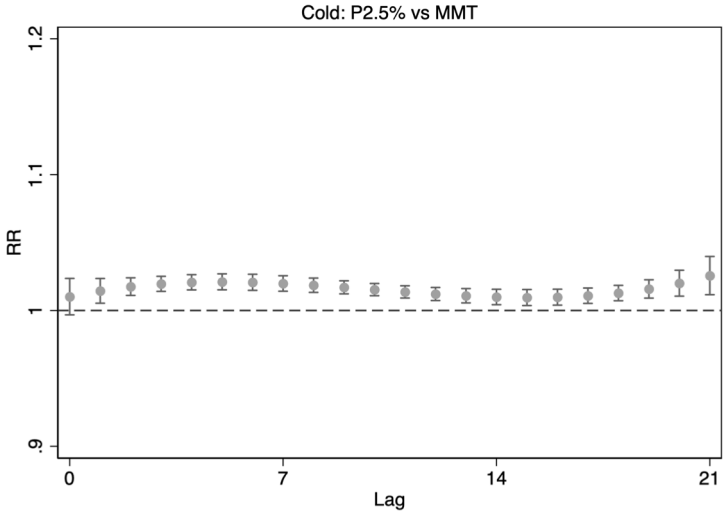
- Spline function

$\mathbf{Y_t = \alpha + \sum \beta_j x_{t-j}}$

$\beta_j = f(\beta, j)$

(Armstrong 2006)

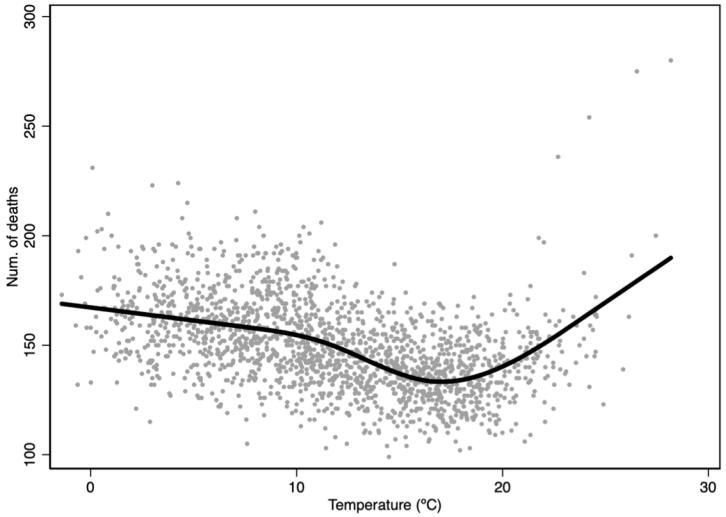




Distributed lag non-linear models

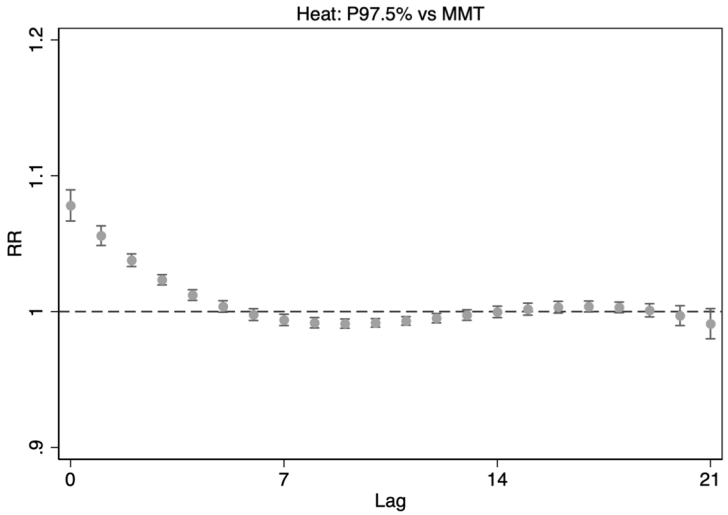
(Gasparrini et al. 2010)

- Exposure-response



×

- Lagged effects



- Function ***f(x)*** for the exposure-response

× function **W(l)** for the lagged effects

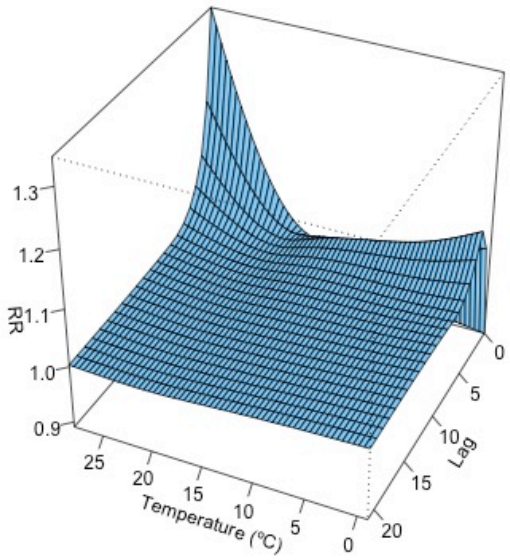
= Bi-dimensional exposure-response × lagged effects ***f·w(x,l)***

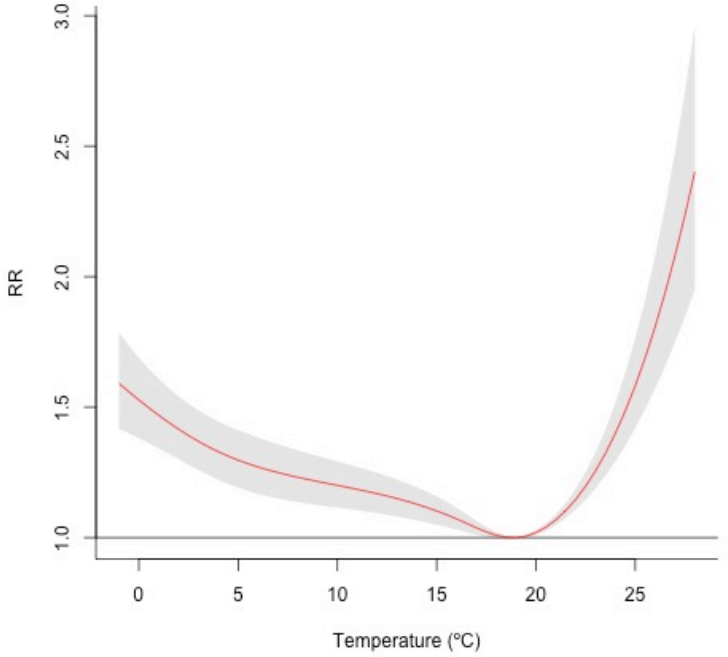
$$Y_t = \alpha + \sum f \cdot w(x_{t-l}, l)$$

Distributed lag non-linear models

(Gasparrini et al. 2010)

- Overall effect

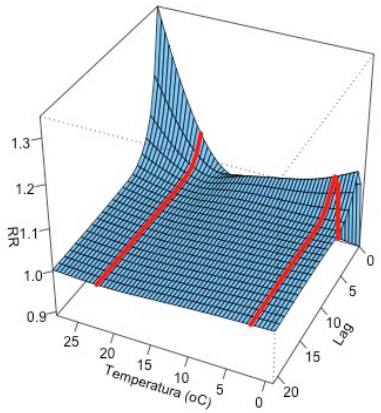


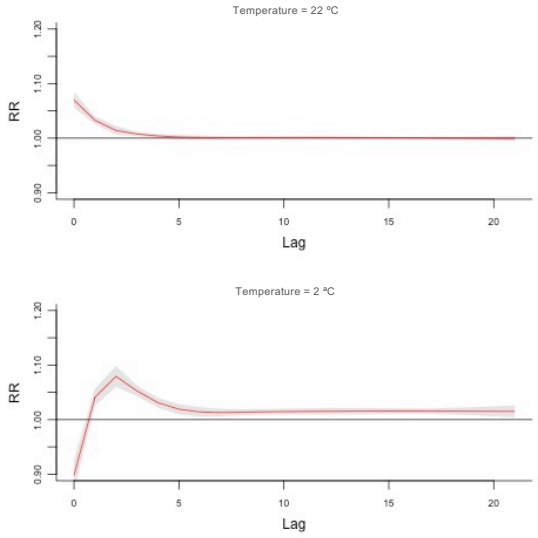


Distributed lag non-linear models

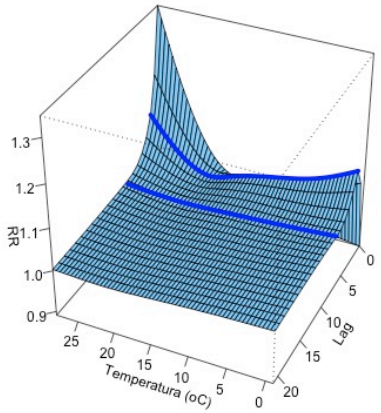
(Gasparrini et al. 2010)

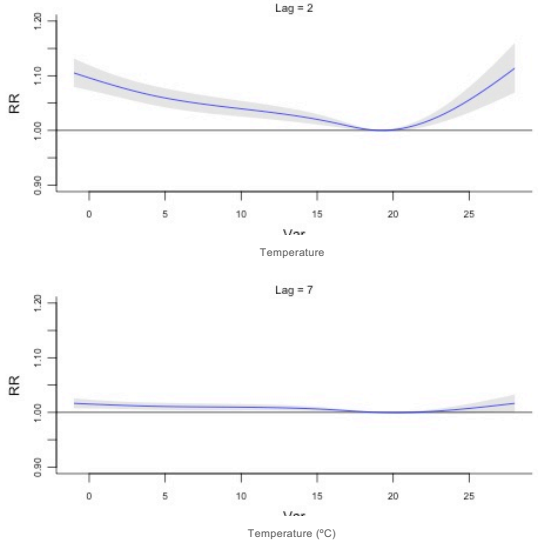
- by temperature





- by lag





Other developments within MCC study

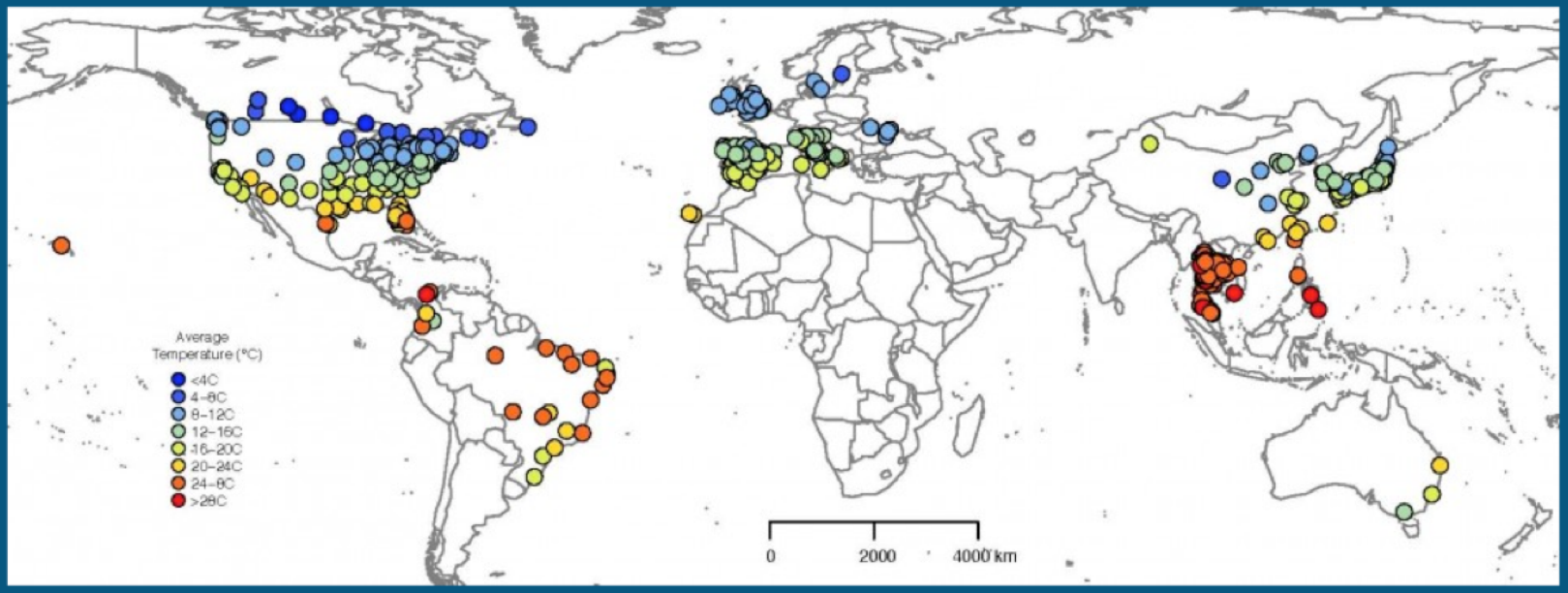
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

MCC Collaborative Research Network

An international research program on the associations between weather and health

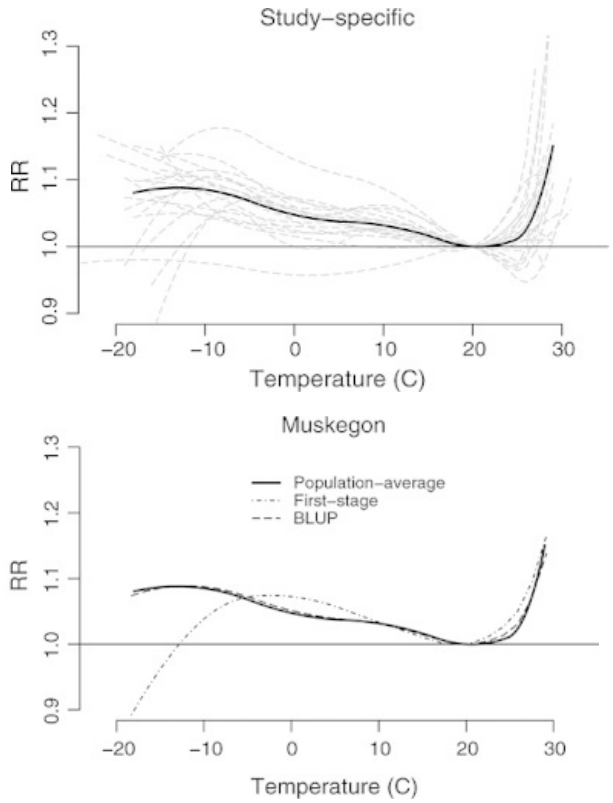
Search

Home | Participants | Research | Publications | Conferences | Media | Private | Contact Us

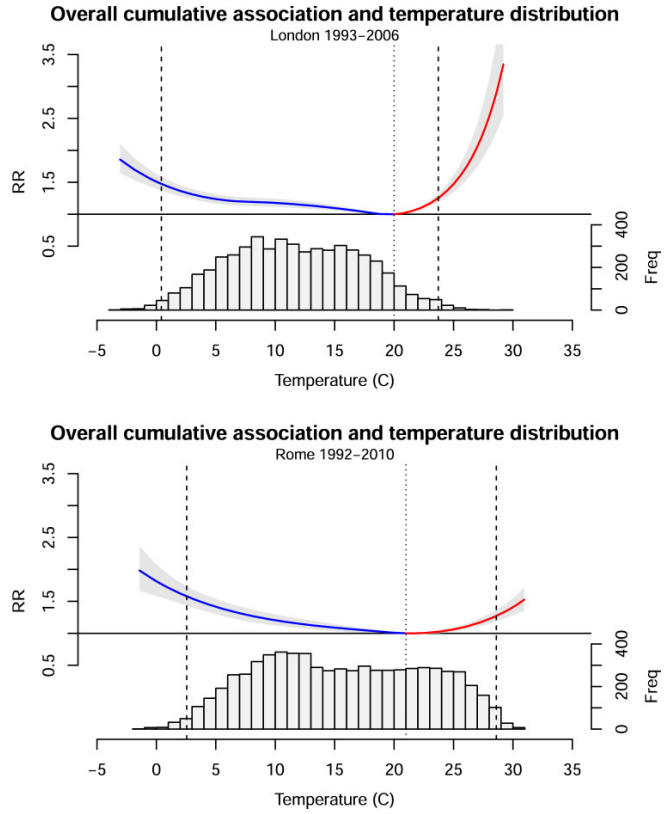


Other developments within MCC study

- Multivariate meta-analysis
  (Gasparrini et al. 2012)

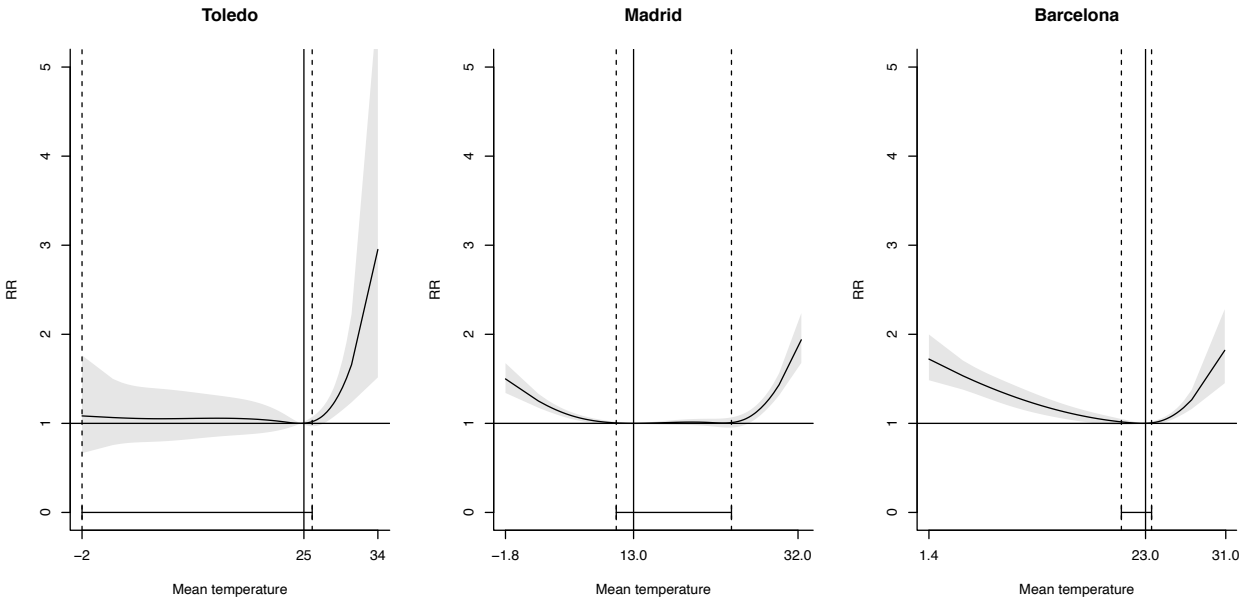


- Attributable fraction
  (Gasparrini and Leone 2014)



Other developments within MCC study

- MMT precision
  (Tobías et al. 2016)



Summary

- Time-series studies provide evidence on short-term associations between environmental exposures and health outcomes
- Time-series regression is similar in principle to any regression analysis but with some specific features
  - Residual autocorrelation
  - Controlling for time-trends and time-varying confounders
  - Be aware of non-linear associations and lagged effects