ISEE Young 2024 – Pre-conference workshop
Rennes, France – 5 June 2024

Methods for risk extrapolation from multi-location studies in environmental epidemiology

Pierre Masselot

Environment & Health Modelling (EHM) Lab, LSHTM, UK

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Multi-location studies

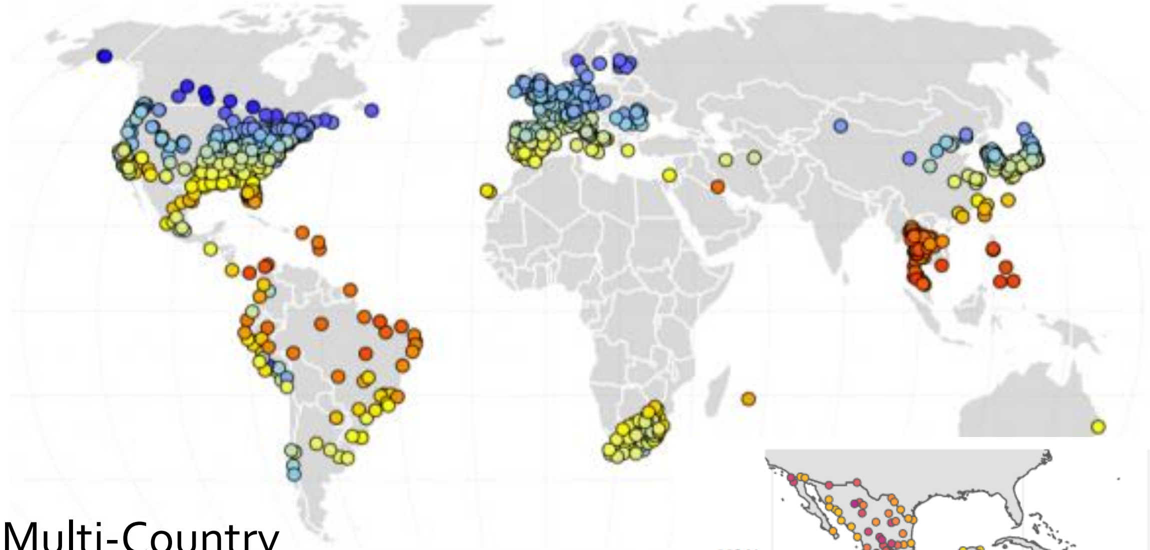
Environmental stressors can necessitate large populations to uncover risks

Recent studies of environment-health associations include multiple locations, groups or populations

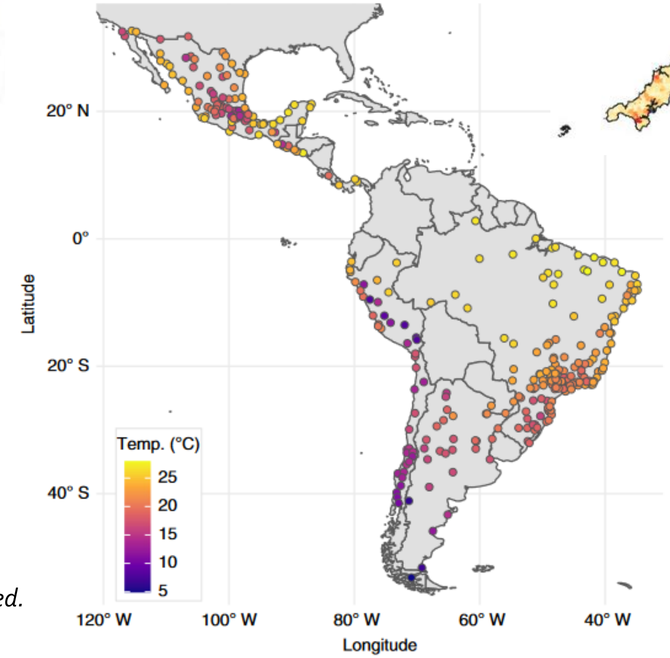
- Larger population: higher statistical power to estimate risks
- Possibility to characterize heterogeneity between populations

Such multi-location studies have become the standard

Examples

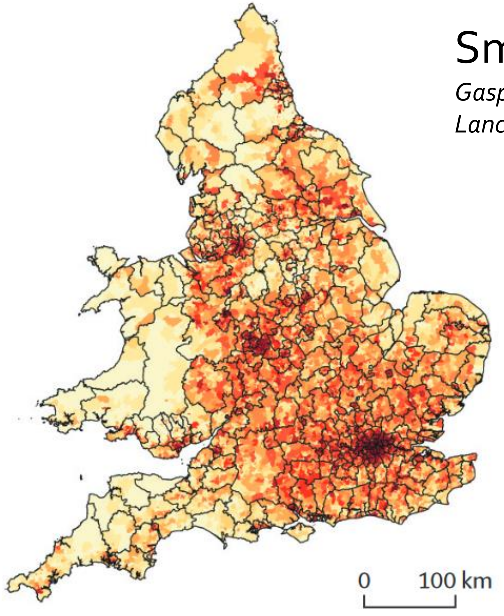


Multi-Country
Multi-City (MCC)



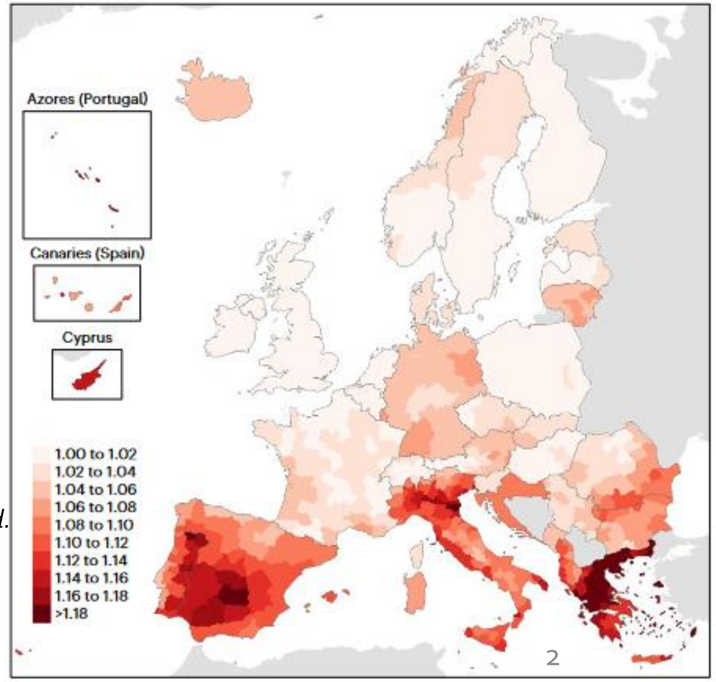
SALURBAL

Kephart et al. (2022) Nat. Med.



Small areas

Gasparrini et al. (2022) The Lancet Pl. Health.



EARLY-ADAPT

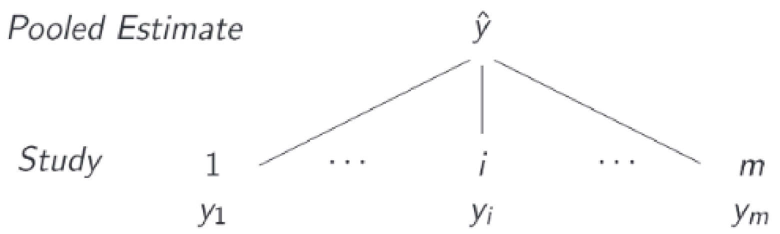
Ballester et al. (2023) Nat. Med.

Two-stage framework

Multi-location studies are often performed using a two-stage framework

1. Estimate a location-specific environment-health association

2. Estimates are pooled in a meta-analytical model



Sera et al. (2019) Stats. Med.

Advantages of the two-stage framework

Computationally efficient

- Each location represents a manageable model
- Only a subset of estimated parameters are pooled in the second-stage

Flexible approach

- Maintains acceptable statistical power
- Makes the integration of spatial information easier

First-stage regression model

Estimation of the association at location i

Association of interest

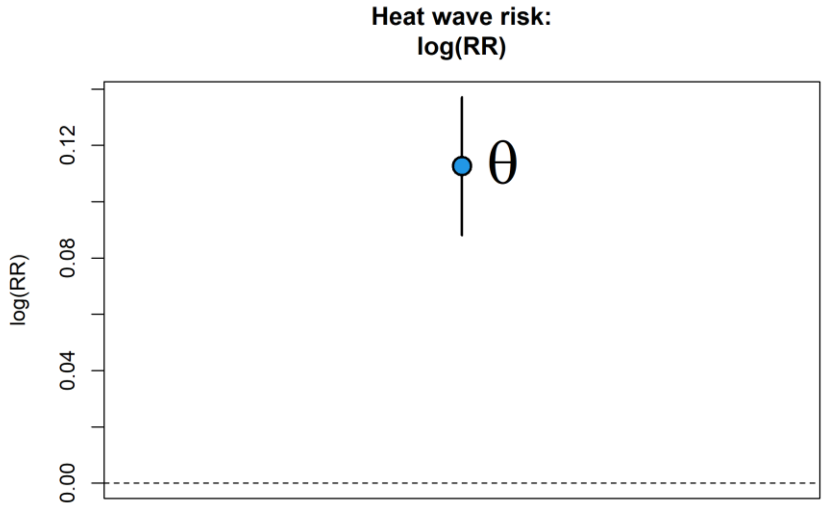
Health outcome

$$g[E(y_{it})] = \alpha + f(x_{it}; \boldsymbol{\theta}_i) + \sum_p h_p(z_{ipt})$$

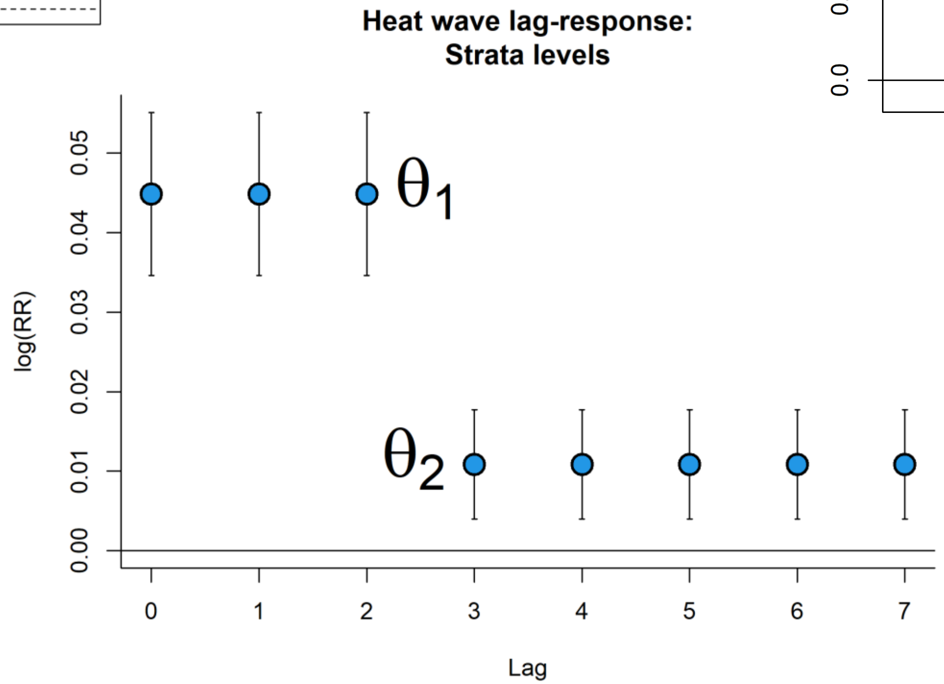
Confounders

- Time series of counts or case-crossover for instance
- Most common model $f(x_{it}; \boldsymbol{\theta}_i)$: distributed lag nonlinear model (DLNM)
    - Indexed by the parameter vector $\boldsymbol{\theta}_i$

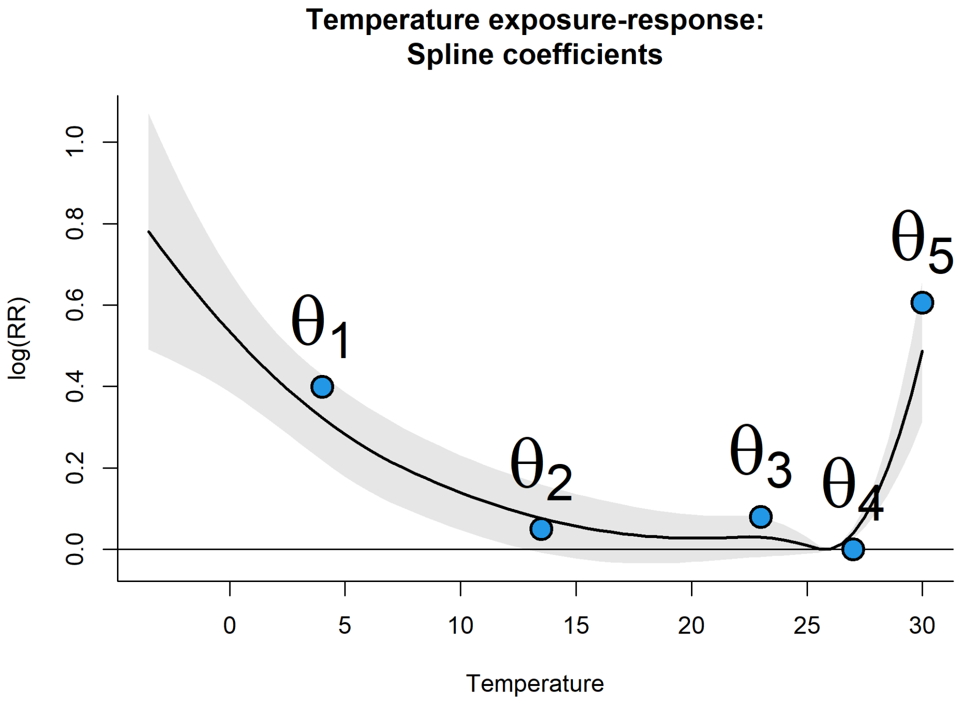
Examples



Heat wave risk:
log(RR)
log(RR)
0.00
0.04
0.08
0.12
θ



Heat wave lag-response:
Strata levels
log(RR)
0.00
0.01
0.02
0.03
0.04
0.05
θ1
θ2
0
1
2
3
4
5
6
7
Lag



Temperature exposure-response:
Spline coefficients
log(RR)
0.0
0.2
0.4
0.6
0.8
1.0
θ1
θ2
θ3
θ4
θ5
0
5
10
15
20
25
30
Temperature

Second-stage meta-analytical model

We then pool and model the estimated location specific parameters $\boldsymbol{\theta}_i$ using a meta-regression model

Fixed effect meta-predictors

Captures first-stage uncertainty

$$\boldsymbol{\theta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

First-stage association parameter(s)

Random effects

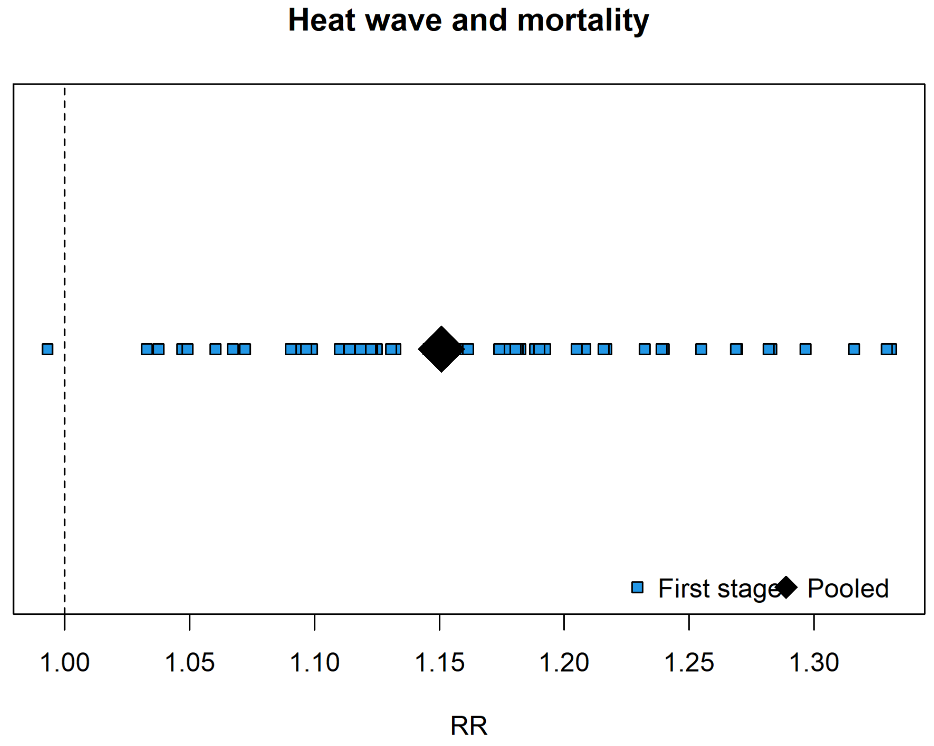
Distributional assumptions

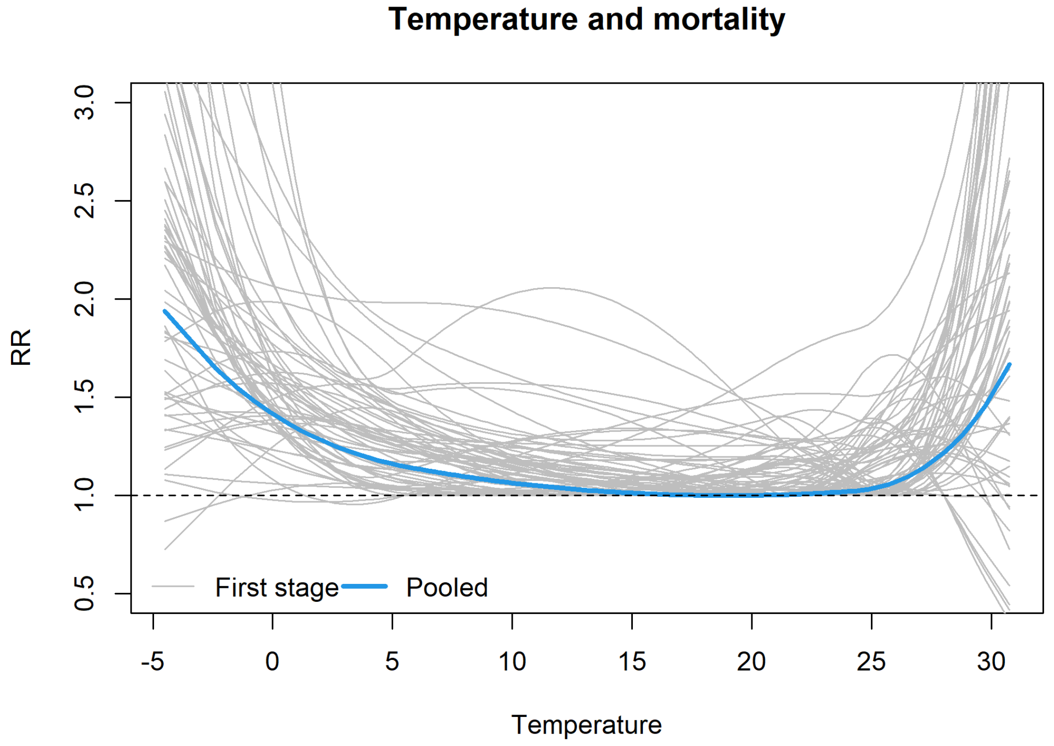
- $\boldsymbol{b}_i \sim N(0, \boldsymbol{\Psi}_\mathrm{i})$
- $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{S}_i) \rightarrow \boldsymbol{S}_i$ estimated in first-stage

Examples (1)

Simple meta-analysis

$$\theta_i = \beta + b_i + \epsilon_i$$

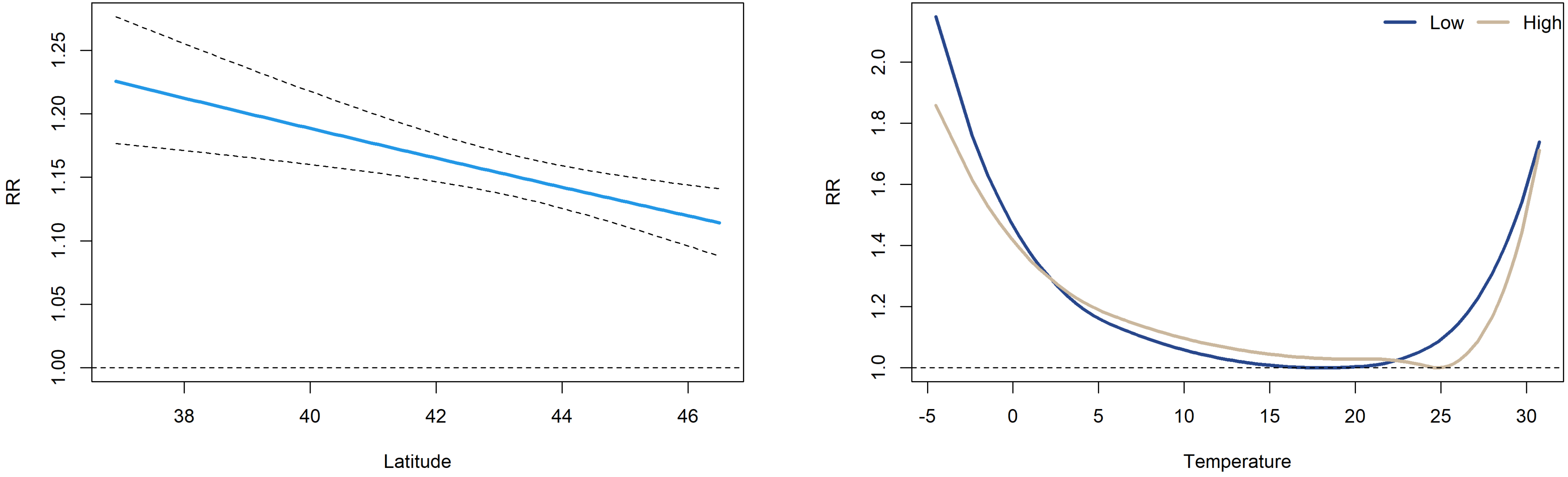




Examples (2)

Meta-regression

$$\theta_i = lat * \beta + b_i + \epsilon_i$$



Best linear unbiased predictions (BLUP)

The two-stage design allows to improve location-specific estimates through the BLUP

The BLUP is defined as

Fixed effect prediction

Proportion of variance attributed to random effects

$$\hat{\boldsymbol{\theta}}_i^* = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\hat{\boldsymbol{\Psi}}_i\boldsymbol{Z}_i\left(\hat{\boldsymbol{\Psi}}_i + \boldsymbol{S}_i\right)^{-1}(\boldsymbol{\theta}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})$$

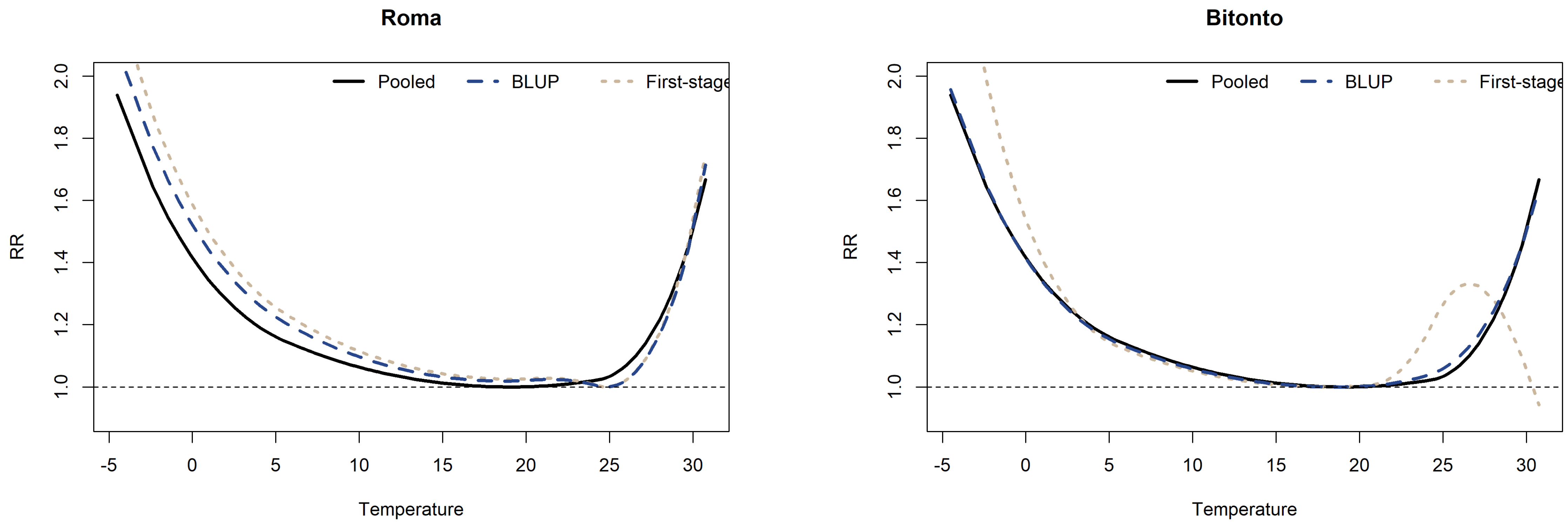
Random effect prediction

Residual

Trade-off between overall mean and location-specific

- Pulls (shrinks) the estimates toward the overall mean
- Shrinking is stronger for inaccurate first stage estimates

BLUP example

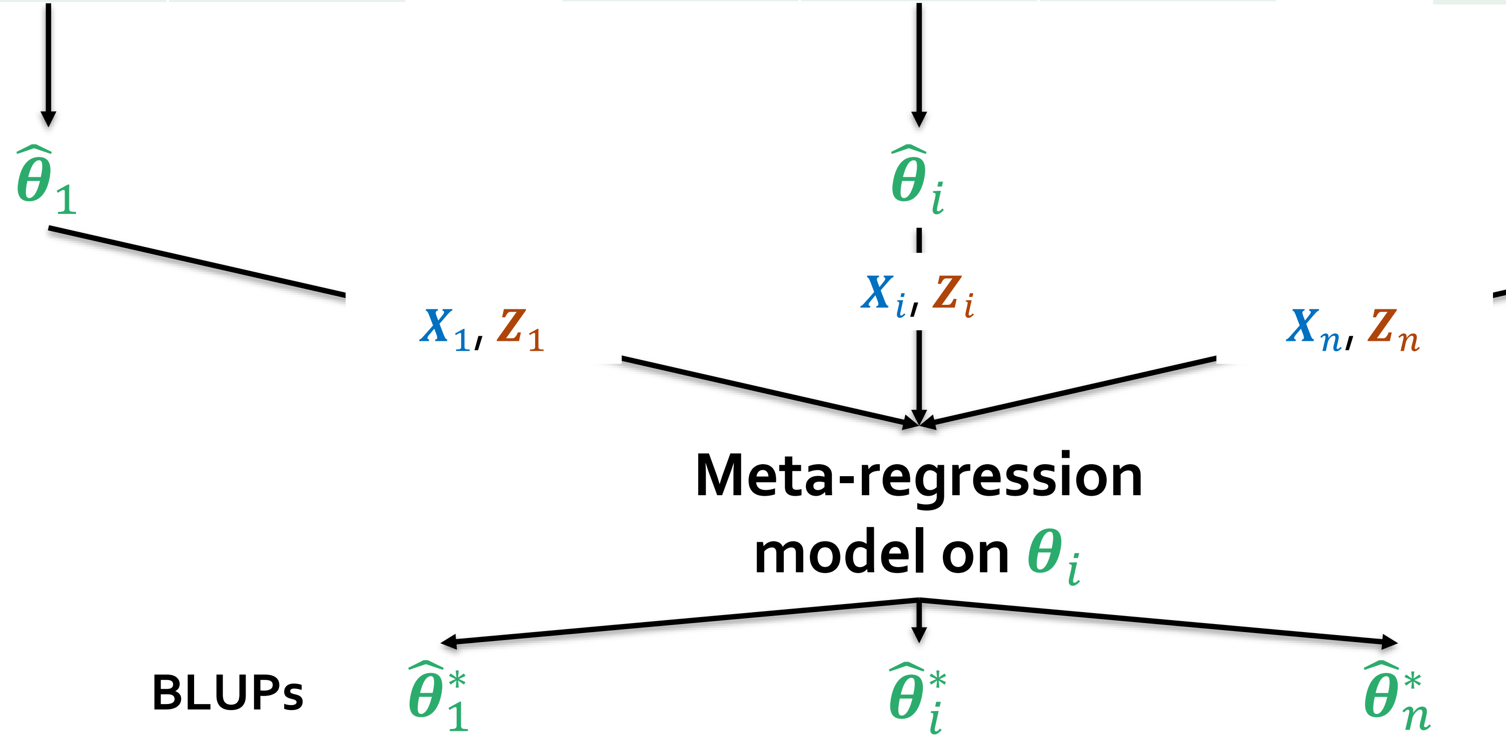


Recap of the two-stage framework

Date	Mortality	Temperature
t_{11}	y_{11}	x_{11}
t_{12}	y_{12}	x_{12}
t_{13}	y_{13}	x_{13}
⋮	⋮	⋮

Date	Mortality	Temperature
t_{i1}	y_{i1}	x_{i1}
t_{i2}	y_{i2}	x_{i2}
t_{i3}	y_{i3}	x_{i3}
⋮	⋮	⋮

Date	Mortality	Temperature
t_{n1}	y_{n1}	x_{n1}
t_{n2}	y_{n2}	x_{n2}
t_{n3}	y_{n3}	x_{n3}
⋮	⋮	⋮



Extrapolation

LONDON SCHOOL *of* HYGIENE &TROPICAL MEDICINE

Limitations of the two-stage framework

Sub-populations need to be "observed"

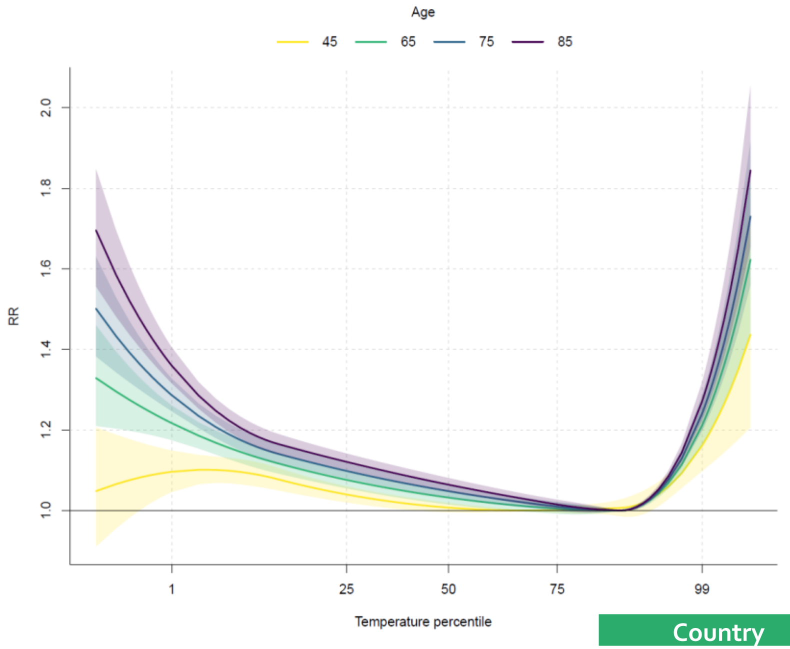
- Have first-stage estimates to be pooled in the meta-regression model
- Necessary for the random effect derivation → computation of BLUPs

Prevents extrapolation to different populations

- Inconsistent sub-population grouping (age or disease group)
- Locations with unavailable time series data

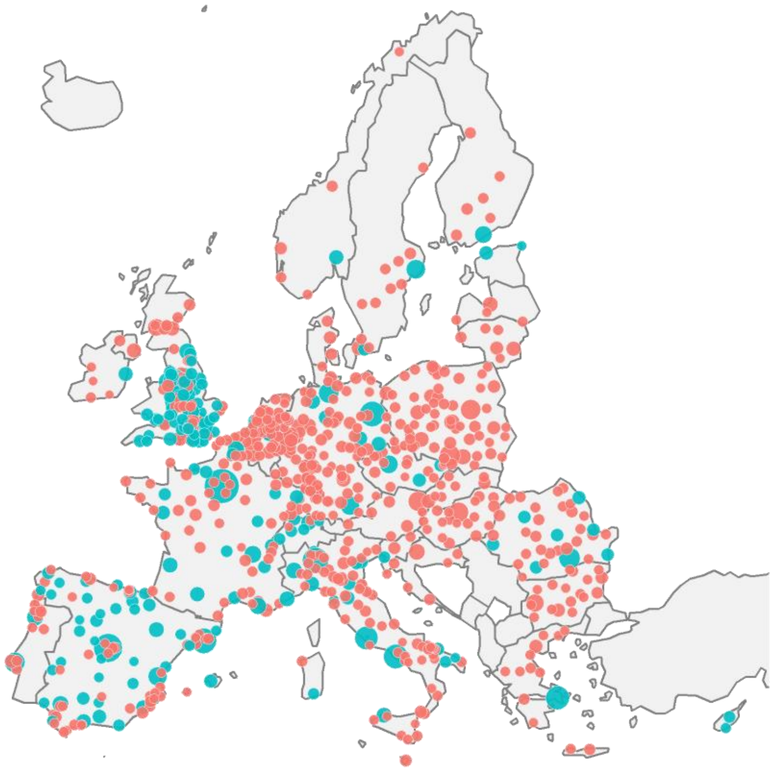
Example: European cities study (Masselot et al. 2023 The Lancet Planet. Health)

Inconsistent age groups



Country	Available age groups
Finland	00-64; 65-99
Norway	00-74; 75-99
Sweden	00-14; 15-64; 65-74; 75-84; 85-99
United Kingdom	00-01; 02-14; 15-44; 45-64; 65-74; 75-84; 85-99
France	00-64; 65-99
Switzerland	00-64; 65-74; 75-84; 85-99
Czechia	00-29; 30-59; 60-74; 75-99
Cyprus	00-44; 45-64; 65-74; 75-84; 85-99
Greece	00; 01-14; 15-64; 65-74; 75-84; 85-99
Portugal	00-64; 65-99
Spain	00-04; 05-14; 15-44; 45-64; 65-74; 75-99

Unavailable locations



Meta-prediction

Throwback to the meta-regression model

$$\boldsymbol{\theta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

💡 One can get the information $\boldsymbol{X}_{i'}$ and $\boldsymbol{Z}_{i'}$ for a new population i'

$\rightarrow$ Predictions $\widehat{\boldsymbol{\theta}}_{i'}$ for new location

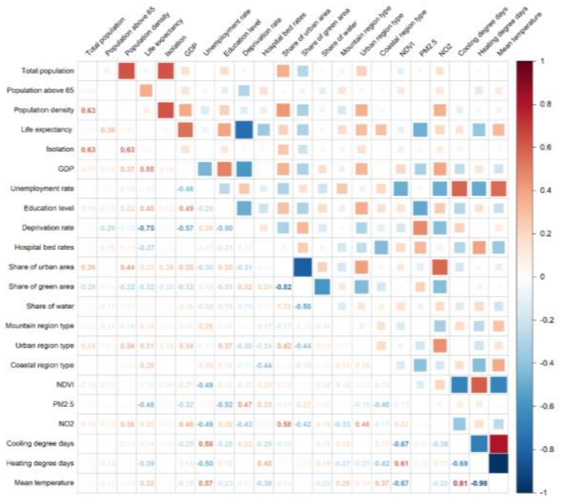
Meta-prediction: shortcomings

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

1. Which information to use for inconsistent subgroup?
   - Cannot be easily summarised by a factor

2. The determinant of risk/vulnerability $\boldsymbol{\theta}_i$ can be numerous and correlated
   - Very large matrix $\boldsymbol{X}_i$
   - Difficult model to fit numerically

3. We cannot estimate $\boldsymbol{b}_{i'}$ when no first-stage estimate is unavailable
   - No shrinkage towards the pooled effect
   - No population-specific effect



Masselot et al. (2023)
The Lancet Planet. Health

Shortcoming 1: Inconsistent subgroups

Country	Available age groups
Finland	00-64; 65-99
Norway	00-74; 75-99
Sweden	00-14; 15-64; 65-74; 75-84; 85-99
United Kingdom	00-01; 02-14; 15-44; 45-64; 65-74; 75-84; 85-99
France	00-64; 65-99
Switzerland	00-64; 65-74; 75-84; 85-99
Czechia	00-29; 30-59; 60-74; 75-99
Cyprus	00-44; 45-64; 65-74; 75-84; 85-99
Greece	00; 01-14; 15-64; 65-74; 75-84; 85-99
Portugal	00-64; 65-99
Spain	00-04; 05-14; 15-44; 45-64; 65-74; 75-99

One possibility is to merge groups

- Needs common breaks
- Lose potentially important resolution

Other (preferred) possibility: attribute continuous value

$$A_{ia} = \left(\sum_{k=l}^{u} d_{ik}\right)^{-1} \sum_{i=l}^{u} o_{ik}\, d_{ik}$$

- Average age of death weighted by death rates
- Age-specific life expectancy can be used

Meta-prediction (2)

Let's update our meta-regression model

$$\hat{\boldsymbol{\theta}}_{ia} = A_{ia}\boldsymbol{\alpha} + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_{ia}$$
$$= \boldsymbol{X}_{ia}\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_{ia}$$

The subgroup variable A_{ia} can be added as a fixed effect

Shortcoming 2: Dimension reduction

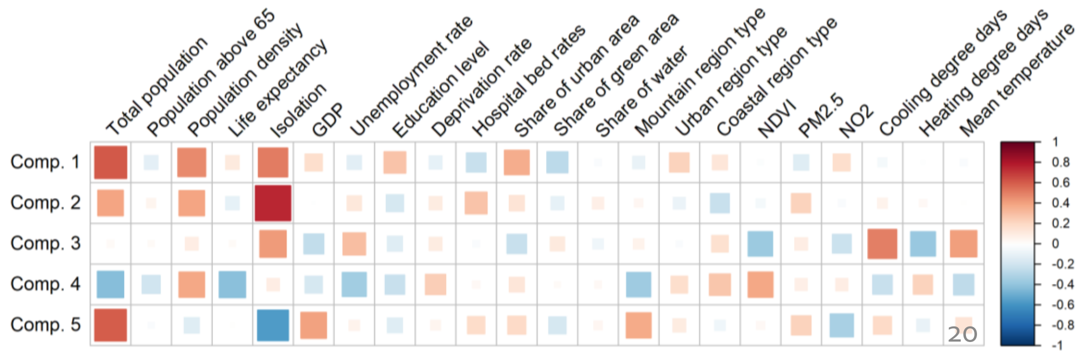
Objective: manage large matrices $\boldsymbol{X}_i$ (or $\boldsymbol{X}_{ia}$)

We can reduce the information contained in $\boldsymbol{X}_i$ into a smaller number of variables

- Principal component analysis (PCA)
- Partial least-squares (PLS)

We can obtain a reduced number of new variables $\boldsymbol{W}_i$

- Composite indices of vulnerability



Meta-prediction (3)

Let's update our meta-regression model

$$\hat{\boldsymbol{\theta}}_{ia} = \boldsymbol{W}_{ia}\boldsymbol{\gamma} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_{ia}$$

We replace the full fixed effect matrix $\boldsymbol{X}_{ia}$ by the reduced matrix $\boldsymbol{W}_{ia}$

Shortcoming 3: random effects

For a full prediction we need to estimate random effect part $\boldsymbol{Z}_{i'}\boldsymbol{b}_{i'}$

Recall the BLUP formula:

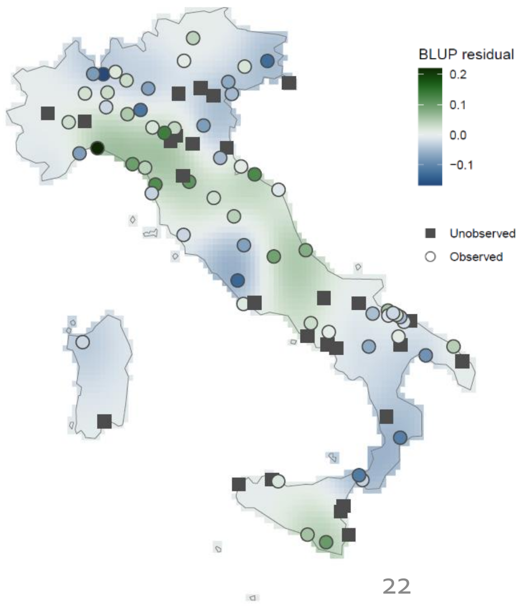
$$\hat{\boldsymbol{\theta}}_i^* = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\hat{\boldsymbol{\Psi}}_i\boldsymbol{Z}_i\left(\hat{\boldsymbol{\Psi}}_i + \boldsymbol{S}_i\right)^{-1}\left(\boldsymbol{\theta}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}\right)$$

Estimation for i

This is not available for i'

Extrapolation of random effect (BLUP residuals)

1. Estimate ξ_i the random part for observed populations i
2. Extrapolate this random part at locations i'
   - IDW
   - Kriging



Meta-prediction (final)

1. Fit the meta-regression model

$$\hat{\boldsymbol{\theta}}_{ia} = \boldsymbol{W}_{ia}\boldsymbol{\gamma} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_{ia}$$

2. Extract the BLUP residuals for observed populations

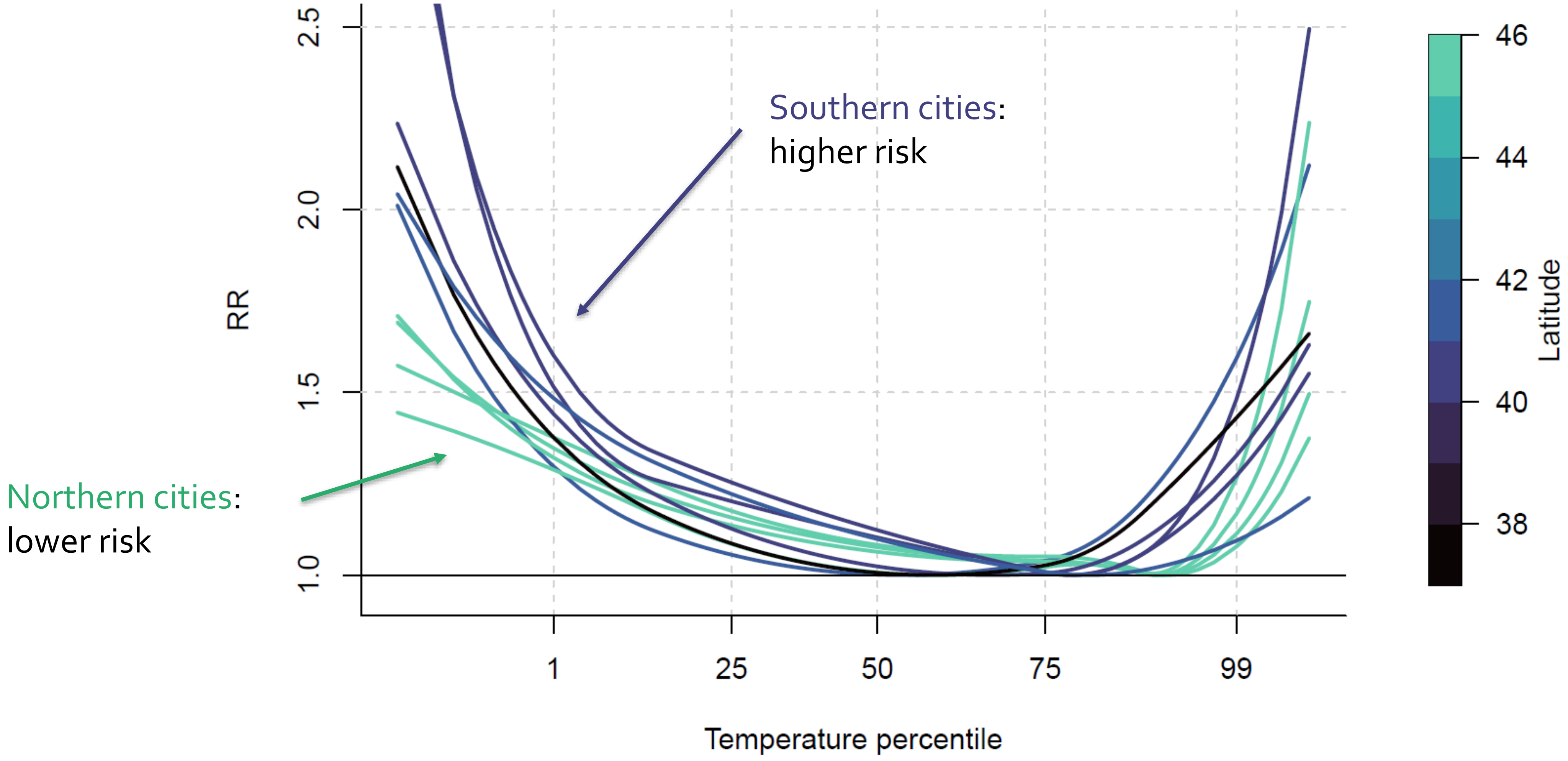
$$\hat{\boldsymbol{\xi}}_i = \boldsymbol{Z}_i\hat{\boldsymbol{\Psi}}_i\boldsymbol{Z}_i\left(\hat{\boldsymbol{\Psi}}_i + \boldsymbol{S}_{ia}\right)^{-1}(\hat{\boldsymbol{\theta}}_{ia} - \boldsymbol{W}_{ia}\hat{\boldsymbol{\gamma}})$$

3. Extrapolate the BLUP residuals to obtain $\hat{\boldsymbol{\xi}}'_i$ by (e.g.) Kriging

4. Predict the risk in new populations

$$\hat{\boldsymbol{\theta}}^*_{i'a} = \boldsymbol{W}_{i'a}\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\xi}}'_i$$

Example: predicted temperature-related mortality



Impact and uncertainty

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

Mortality attribution

For a given exposition x, we can estimate the associated relative risk

$$RR_x = e^{f(x;\boldsymbol{\theta}_i)}$$

Can be BLUP or first-stage

This RR_x can then be transformed into impact measures

Attributable Fraction: $$AF_x = \frac{RR_x - 1}{RR_x}$$

Attributable Number: $$AN_x = m.AF_x$$

Excess rate: $$E_x = \frac{AN_x}{p} = \frac{m}{p} AF_x$$

- p is the population and m the total number of cases

Example: temperature

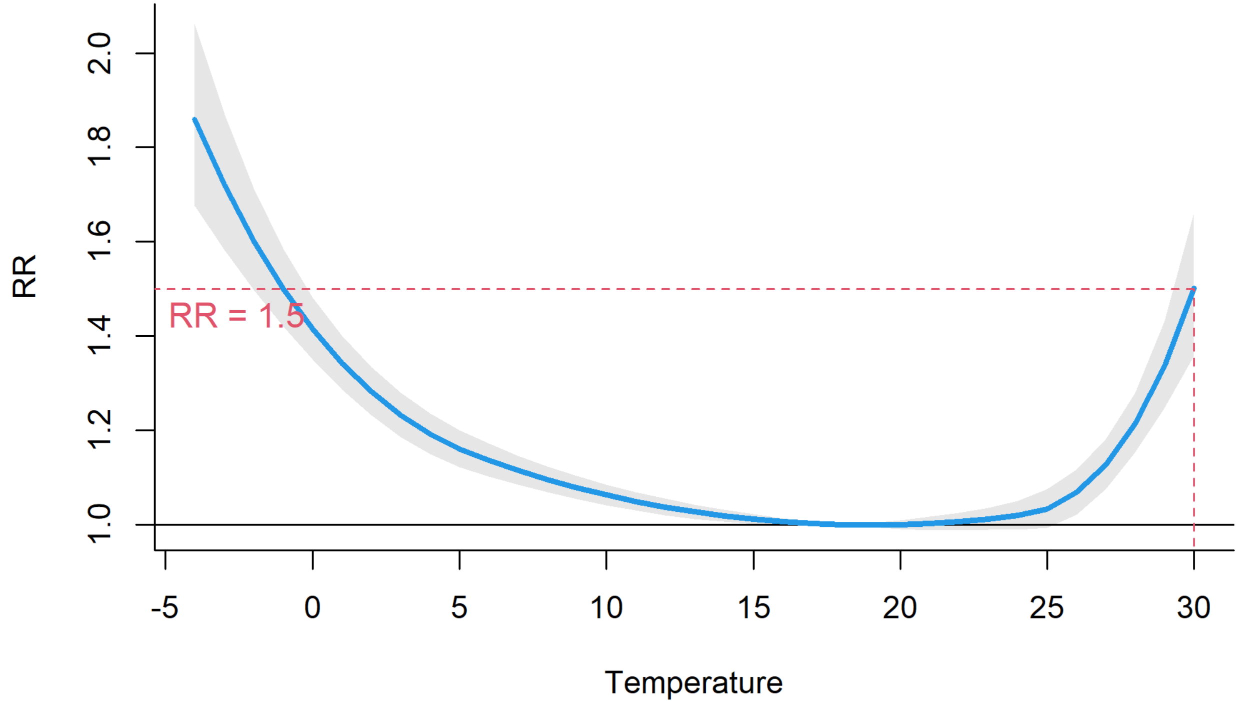
In a population of 1M, there was a total of 150 deaths a day with $x = 30°C$

We estimate that $RR_{x=30} = 1.5$

$$AF_{x=30} = \frac{0.5}{1.5} = \frac{1}{3} = 33\%$$

$$AN_{x=30} = \frac{150}{3} = 50$$

$$E_{x=30} = \frac{50}{1M} = 5.10^{-5}$$



Comparison of populations

E is sensitive to the baseline mortality

- Mortality without the exposure
- For temperature, we use the minimum mortality temperature as baseline (MMT)

Baseline mortality depends on many factors

- In particular the age distribution of population
- Other examples are the male/female ratio or prevalence of specific diseases

With identical vulnerability to an exposure, two populations can have different AN / E

→ To compare populations, E can be standardised

Standardised death rates

We can estimate E_{ia} in location i for age-group a

- i and a can represent any population and sub-group of this population

Standardised excess rates are then computed as a weighted average

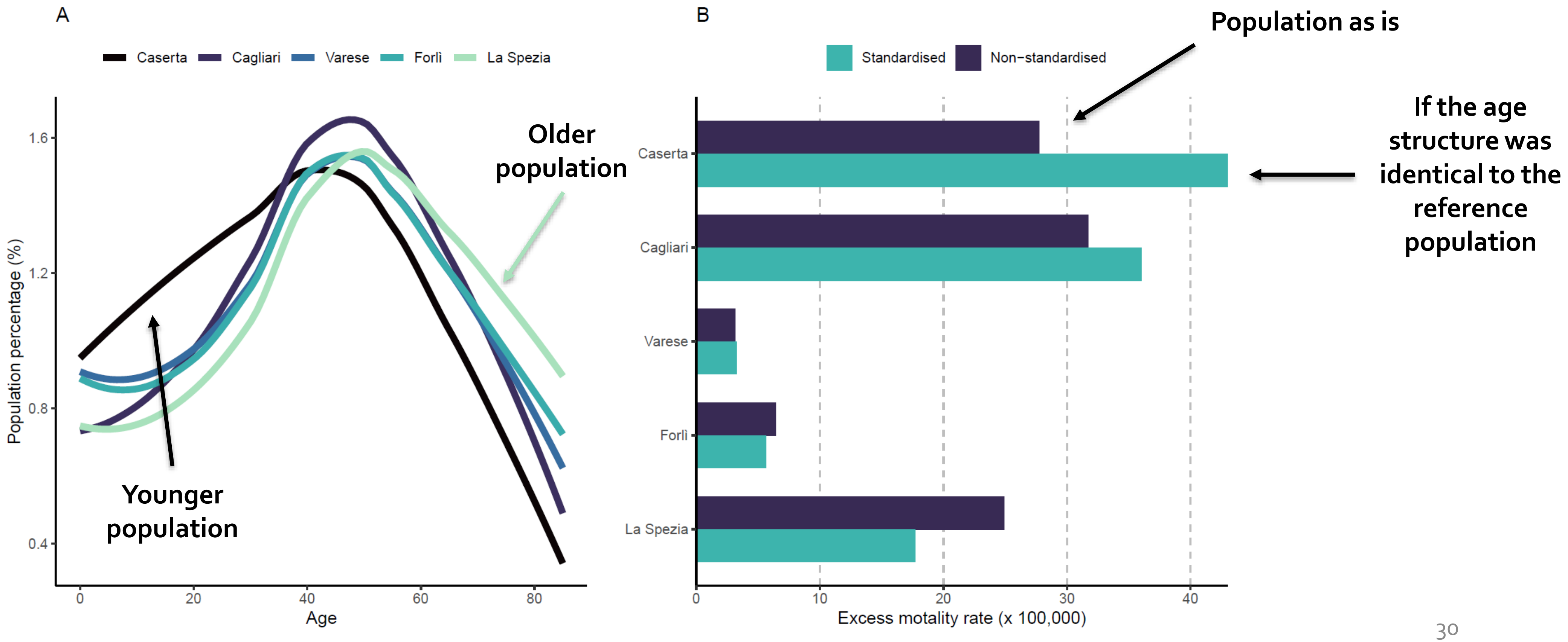
$$E_i = \frac{\sum_a w_a E_{ia}}{\sum_a w_a}$$

Where w_a represents the number of cases (deaths) in a reference population

Example: age-standardisation

Temperature-related excess death rates in five Italian cities

- Reference population: 2013 standard European population



Uncertainty assessment

The impact measures don't have an obvious distribution to obtain confidence intervals

We assess uncertainty and obtain confidence intervals by Monte Carlo simulations

1. We obtained predictions $\hat{\boldsymbol{\theta}}_{ia}^{*}$ and their uncertainty $\mathrm{V}(\hat{\boldsymbol{\theta}}_{ia}^{*})$

2. We can simulate a large number of coefficients

$$\boldsymbol{\theta}_{ia}^{*b} \sim N\left(\hat{\boldsymbol{\theta}}_{ia}^{*}, \mathrm{V}(\hat{\boldsymbol{\theta}}_{ia}^{*})\right)$$

3. From these simulated coefficients we compute measures of impact

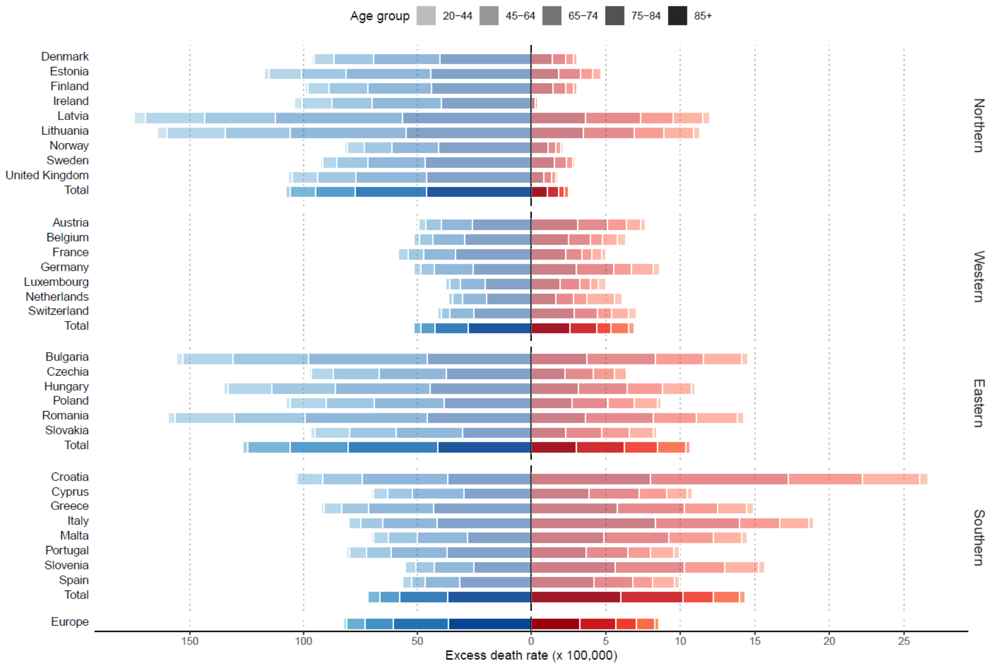
$$AF_i^b, AN_i^b, E_i^b$$

4. We compute the empirical quantiles of simulated measures as confidence intervals

Spatial correlation and aggregation

It is common practice to aggregate impact measures at a higher level

- For instance, by country or continent



Masselot et al. (2023)
The Lancet Planet. Health

The usual Monte-Carlo method considers that locations are independent

When aggregating, it underestimates uncertainty

- Random deviations cancel each other

The $\widehat{\boldsymbol{\theta}}^{*}_{ia}$ inherit from the same meta-regression model

$\rightarrow$ They are correlated

Update to uncertainty assessment

1. We simulate new parameters from the meta-regression model

$$\boldsymbol{\gamma}^b \sim \boldsymbol{N}\left(\widehat{\boldsymbol{\gamma}}, \boldsymbol{V}(\widehat{\boldsymbol{\gamma}})\right), \qquad \xi_i'^b \sim N(\hat{\xi}_i', V(\hat{\xi}_i'))$$

Kriging allows obtaining $V(\hat{\xi}_i')$

2. From the simulated coefficients we obtain predictions $\boldsymbol{\theta}_{ia}^{*b}$ and their uncertainty

3. From these simulated coefficients we compute measures of impact

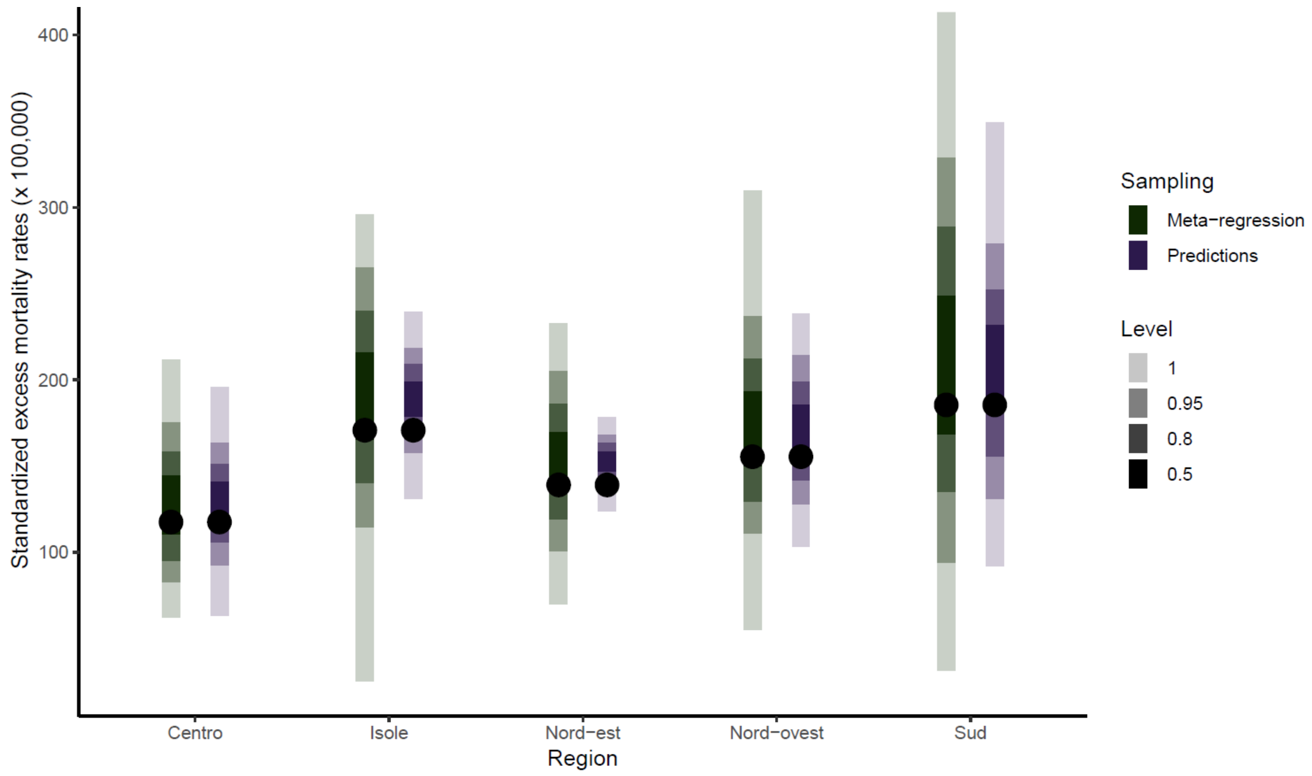
$$AF_i^b, AN_i^b, E_i^b$$

4. We compute the empirical quantiles of simulated measures as confidence intervals

Example: Italian cities

We estimate temperature-related standardised excess deaths for 87 Italian cities and aggregate into five regions

The usual approach underestimates uncertainty compared to the meta-regression approach



Summary

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Extrapolation of risk through the two-stage method

The two-stage analysis is an efficient framework for multi-location studies

- Has been extensively used for the past decade
- Still the subject of methodological development

We use this framework to extrapolate the risk to unobserved populations

- New locations
- Inconsistent subgroups between locations
- Computation of standardised measures
- Uncertainty assessment

Summary

1. Estimate the location and age-specific exposure-response function $\hat{\boldsymbol{\theta}}_{ia}$
2. Compute an average age of the population A_{ia}
3. Create composite indices of vulnerability $\boldsymbol{W}_{ia}$ from characteristics X_{ia}
4. Fit the meta-regression model $\hat{\boldsymbol{\theta}}_{ia} = \boldsymbol{W}_{ia}\boldsymbol{\gamma} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_{ia}$
5. Extract the BLUP residuals $\hat{\boldsymbol{\xi}}_i$ from the meta-regression model and extrapolate by Kriging
6. Predict the exposure-response function in all locations of interest $\hat{\boldsymbol{\theta}}^*_{i'a} = \boldsymbol{W}_{i'a}\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\xi}}'_i$
7. Compute standardised excess rates E_i
8. Assess uncertainty by Monte Carlo simulations from 6

Some additional references

Ballester, J., Quijal-Zamorano, M., Méndez Turrubiates, R.F., Pegenaute, F., Herrmann, F.R., Robine, J.M., Basagaña, X., Tonne, C., Antó, J.M., Achebak, H., 2023. Heat-related mortality in Europe during the summer of 2022. Nat Med 1–10. https://doi.org/10.1038/s41591-023-02419-z

Dominici, F., Samet, J.M., Zeger, S.L., 2000. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. Journal of the Royal Statistical Society: Series A (Statistics in Society) 163, 263–302. https://doi.org/10.1111/1467-985X.00170

Gasparrini, A., Armstrong, B., Kenward, M.G., 2010. Distributed lag non-linear models. Statistics in Medicine 29, 2224–2234. https://doi.org/10.1002/sim.3940

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklöv, J., Forsberg, B., Leone, M., De Sario, M., Bell, M.L., Guo, Y.-L.L., Wu, C., Kan, H., Yi, S.-M., de Sousa Zanotti Stagliorio Coelho, M., Saldiva, P.H.N., Honda, Y., Kim, H., Armstrong, B., 2015. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. The Lancet 386, 369–375. https://doi.org/10.1016/S0140-6736(14)62114-0

Gasparrini, A., Leone, M., 2014. Attributable risk from distributed lag models. BMC Medical Research Methodology 14, 55. https://doi.org/10.1186/1471-2288-14-55

Gasparrini, A., Masselot, P., Scortichini, M., Schneider, R., Mistry, M.N., Sera, F., Macintyre, H.L., Phalkey, R., Vicedo-Cabrera, A.M., 2022. Small-area assessment of temperature-related mortality risks in England and Wales: a case time series analysis. The Lancet Planetary Health 6, e557–e564. https://doi.org/10.1016/S2542-5196(22)00138-3

Kephart, J.L., Sánchez, B.N., Moore, J., Schinasi, L.H., Bakhtsiyarava, M., Ju, Y., Gouveia, N., Caiaffa, W.T., Dronova, I., Arunachalam, S., Diez Roux, A.V., Rodríguez, D.A., 2022. City-level impact of extreme temperatures and mortality in Latin America. Nat Med 1–6. https://doi.org/10.1038/s41591-022-01872-6

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30, 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Sera, F., Armstrong, B., Blangiardo, M., Gasparrini, A., 2019. An extended mixed-effects framework for meta-analysis. Statistics in Medicine 38, 5429–5444. https://doi.org/10.1002/sim.8362

Sera, F., Gasparrini, A., 2022. Extended two-stage designs for environmental research. Environmental Health 21, 41. https://doi.org/10.1186/s12940-022-00853-z

Thank you for your attention

pierre.masselot@lshtm.ac.uk

@MasselotPierre

This method has been implemented in:

Masselot, P., Mistry, M., Vanoli, J., Schneider, R., Iungman, T., Garcia-Leon, D., Ciscar, J.-C., Feyen, L., Orru, H., Urban, A., Breitner, S., Huber, V., Schneider, A., Samoli, E., Stafoggia, M., de'Donato, F., Rao, S., Armstrong, B., Nieuwenhuijsen, M., Vicedo-Cabrera, A.M., Gasparrini, A., 2023. Excess mortality attributed to heat and cold: a health impact assessment study in 854 cities in Europe. The Lancet Planetary Health 7, e271–e281. https://doi.org/10.1016/S2542-5196(23)00023-2

A methodological paper is currently under review