Modelling spatio-temporal data: methods, examples and challenges

Marta Blangiardo

Imperial College London
MRC-PHE Centre for Environment and Heath
m.blangiardo@imperial.ac.uk

BSU Cambridge, 11^{th} October 2017

Spatial/temporal data

Characteristics of spatio/temporal data:

- Geographically referenced and often presented ad maps.
- Temporally correlated as in time series structures.

Found in several areas:

- Epidemiology (surveillance, risk assessment);
- Environmental statistics (modelling of exposures);
- Social statistics (e.g. crime);
- Medicine (e.g. brain imaging);
- etc.

Types of spatial data

- **Point-referenced data**: the exact location of the occurrences is known
  - can be collected through specialized survey / monitoring network...
  - if location itself is *random*, e.g. measurements of where events occur
    $\Rightarrow$ point process statistical framework.

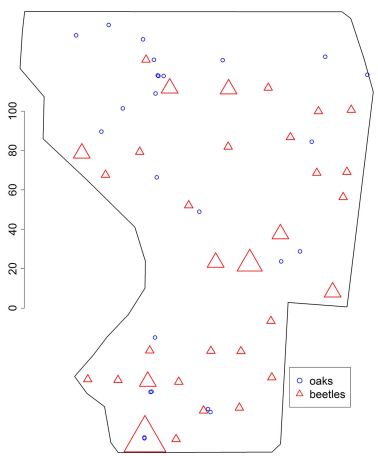


Figure: Wagner et al., iForest 2016

- Typically $D \in R^2$
- Random realisation of point pattern by $\boldsymbol{S} = \{s_1, \ldots, s_n\}$.
- $\boldsymbol{S}$ is random and characterised by
  - $N(D)$: number of points in D
  - Multivariate density D^n for $f(s_1, \ldots, s_n)$ (location density).
- Spatial point process.

Types of spatial data [cont'd]

- **Point-referenced data**: the exact location of the occurrences is known
  - can be collected through specialized survey / monitoring network...
  - if locations are *fixed* (monitoring stations, postcodes in an area) and the variable of interest is measured at each location (e.g. presence/absence of cases, pollution concentrations)
    $\Rightarrow$ geostatistical framework.
- **Area data or count data**: locations are areal units with well defined geographical boundaries, usually administrative units
  - outcome is number of cases aggregated over the area
    $\Rightarrow$ small area framework.

Small area framework

General inferential issues

Spatial (and temporal) pattern suggest that observations close to each other have more similar values than those far from each other.

- Do we want to smooth the data?
  → Disease mapping, cluster detection.
  → **In this talk**: Space-time-age disease mapping.
- Do we want to evaluate temporal trends for each area?
  → Discover unusual patterns.
  → **In this talk**: Disease surveillance.
- Is the spatial clustering and/or temporal trend a nuisance quantity that we wish to take into account but are not explicitly interested in?
  → Spatial regression
  → **In this talk**: Leukaemia and benzene in Greater London.
- Is the regression coefficient spatially varying?
  → **In this talk**: Case-crossover design for spatio-temporal epidemiological investigations.

General framework - I

- **Data** for a region of interest, a geographical level and a specific period:
  - O_i: Observed number of cases in area i
    Lung cancer deaths in males aged 45+
    Number of reported crimes
    Number of people reporting confidence in the police.

General framework - I

- **Data** for a region of interest, a geographical level and a specific period:

  - O_i: Observed number of cases in area i
    Lung cancer deaths in males aged 45+
    Number of reported crimes
    Number of people reporting confidence in the police.

  - n_i (or E_i): Population at risk in area i (or expected number of cases)
    Male population aged 45+
    Live births and stillbirths
    Resident Population or nb of dwellings
    Number of people surveyed in each borough neighborhood.

General framework - I

- **Data** for a region of interest, a geographical level and a specific period:
  - O_i: Observed number of cases in area i
    Lung cancer deaths in males aged 45+
    Number of reported crimes
    Number of people reporting confidence in the police.
  - n_i (or E_i): Population at risk in area i (or expected number of cases)
    Male population aged 45+
    Live births and stillbirths
    Resident Population or nb of dwellings
    Number of people surveyed in each borough neighborhood.
- **Parameter of interest** Relative risk λ in each area compared with the chosen reference area.

General framework - II

- Standard statistical model for rare outcomes/small areas

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

  - $SMR_i = \frac{O_i}{E_i}$ is the MLE for λ_i;
  - SMR_i very imprecise for rare events and/or areas with small populations, λ_i estimated variance is $\frac{O_i}{E_i^2}$;
    $\Rightarrow$ Highlights extreme risk estimates based on small numbers.

General framework - II

- Standard statistical model for rare outcomes/small areas

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

  - $SMR_i = \frac{O_i}{E_i}$ is the MLE for λ_i;
  - SMR_i very imprecise for rare events and/or areas with small populations, λ_i estimated variance is $\frac{O_i}{E_i^2}$;
    $\Rightarrow$ Highlights extreme risk estimates based on small numbers.

- SMR in each area is estimated independently
  $\rightarrow$ makes no use of risk estimates in other areas of the map, even though these are likely to be similar.

- Ignores possible spatial correlation between disease risk in nearby areas

General framework - II

- Standard statistical model for rare outcomes/small areas

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

  - $SMR_i = \frac{O_i}{E_i}$ is the MLE for λ_i;
  - SMR_i very imprecise for rare events and/or areas with small populations, λ_i estimated variance is $\frac{O_i}{E_i^2}$;
    $\Rightarrow$ Highlights extreme risk estimates based on small numbers.

- SMR in each area is estimated independently
  $\rightarrow$ makes no use of risk estimates in other areas of the map, even though these are likely to be similar.

- Ignores possible spatial correlation between disease risk in nearby areas

Bayesian 'smoothing' estimators in a hierarchical formulation.

Hierarchical modelling for small area data

Poisson-logNormal model

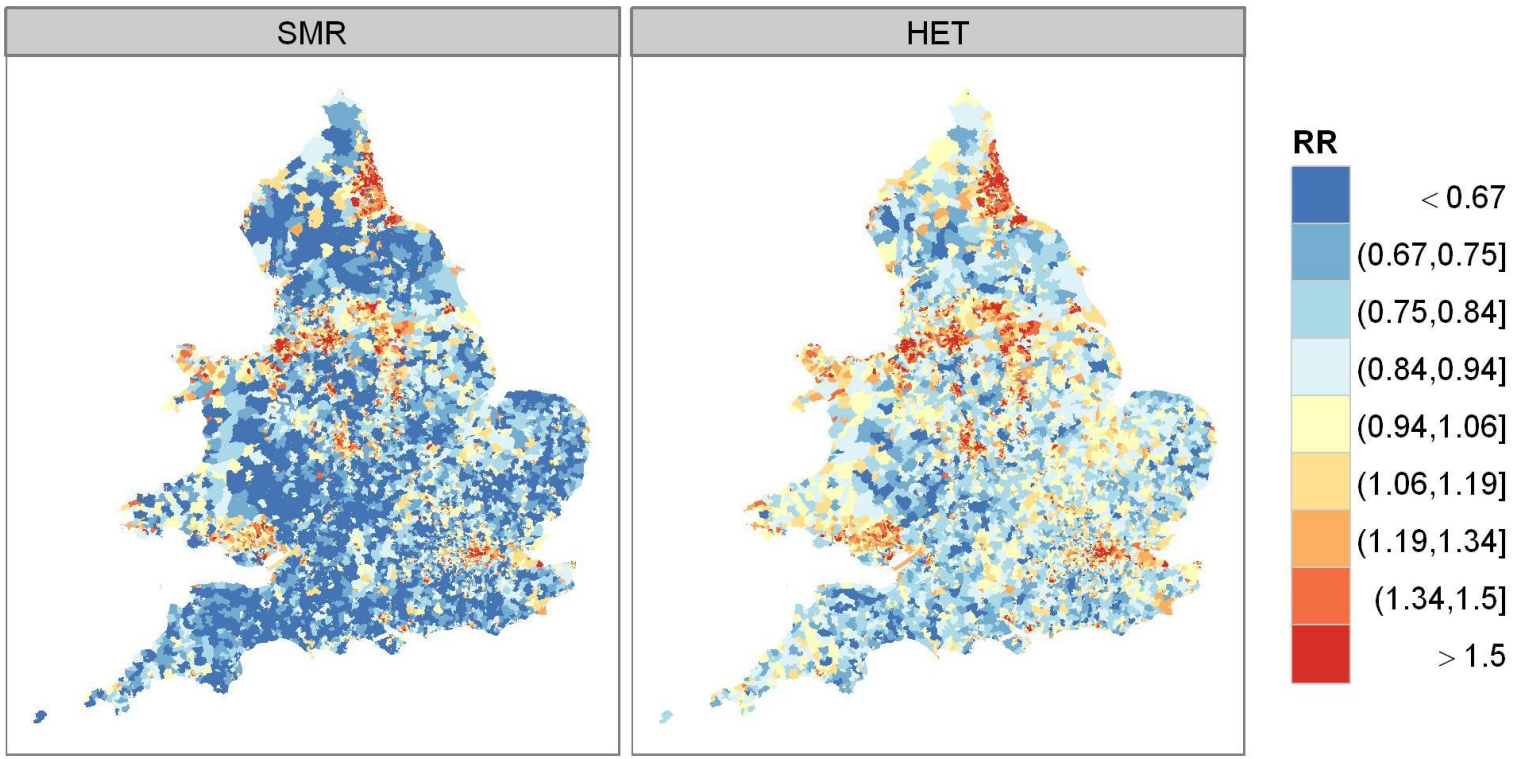
$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + v_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

where

- O_i, E_i: observed and expected nb of cases in area i
- $\lambda_i = \exp(\alpha + v_i)$: RR in area i compared with expected risk based on age and sex of population
- Parameters v_i: **area-specific random effects**
- residual RR $= \exp(v_i)$.

Lung cancer incidence in males, 1985-2009, England and Wales

RR estimates using 2 methods



SMRs and smoothed RRs

Local spatial dependency

To account for local dependency it is possible to add a spatial structure in the model:

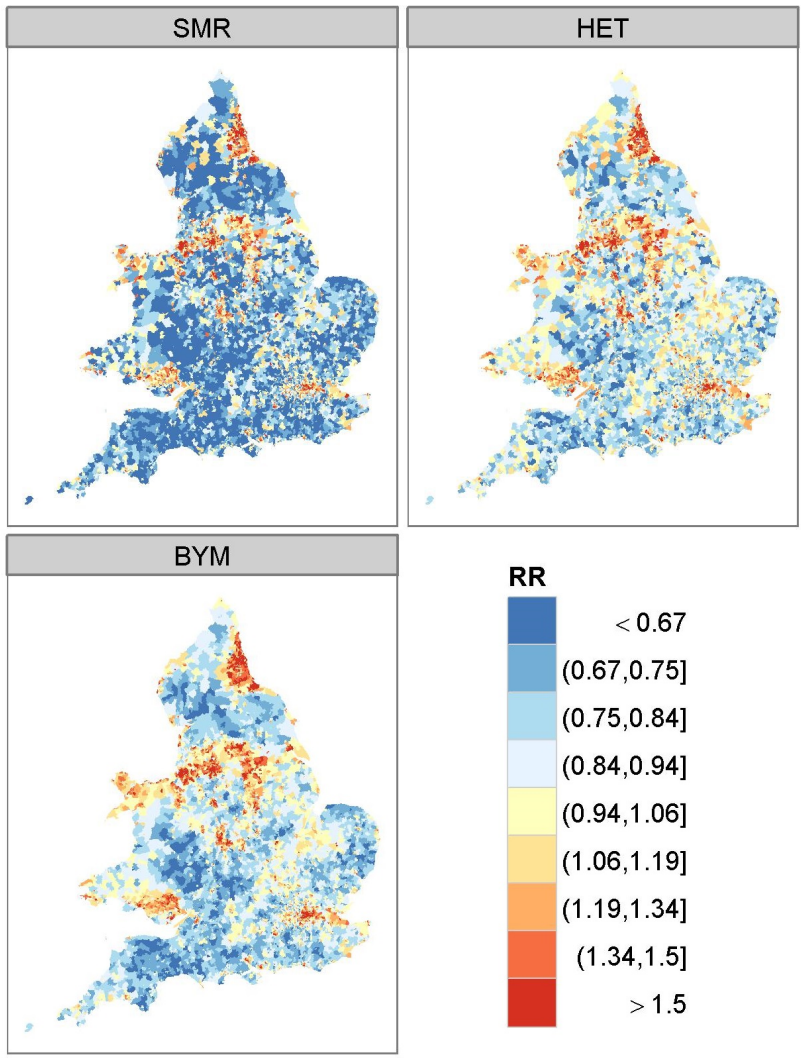
Convolution model

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\lambda_i E_i) \\
\log \lambda_i &= \alpha + v_i + u_i \\
v_i &\sim \text{Normal}(0, \sigma_v^2) \\
u_i \mid u_{-i} &\sim \text{N}\left(\frac{\sum_{k=1}^{n} w_{ik} u_k}{\sum_{k=1}^{n} w_{ik}}, \frac{\sigma_u^2}{\sum_{k=1}^{n} w_{ik}} \right)
\end{aligned}
$$

- u_i follows a conditional autoregressive specification (CAR); it assumes that only neighbouring areas contribute to the distribution of area i ($w_{ik} = 1$).
- The combination of v_i and u_i guarantees local and global smoothing (BYM model).

Besag, J., et al. (1991). "Bayesian image restoration, with two applications in spatial statistics". **Annals of the Institute of Statistical Mathematics**, 43, 1-59.

Residual RR of lung cancer incidence in males, 1985-2009, England and Wales II



SMR: non smoothed RR

HET: non spatially smoothed residual RR $\exp(v)$

BYM: spatially and non spatially smoothed residual RR $\exp(v + u)$

Extending space to space-time

- The hierarchical structure can be extended to incorporate time into a space-time model.
- The stability (or not) of the spatial pattern can aid interpretation.
- The specific space-time components of the model can potentially pinpoint unusual/ emerging risk factors.

Spatio-temporal hierarchical model

$$\begin{aligned} O_{it} &\sim \text{Poisson}(\lambda_{it} E_{it}) \\ \log \lambda_{it} &= \alpha + v_i + u_i + \phi_t + \gamma_t + \psi_{it} \end{aligned}$$

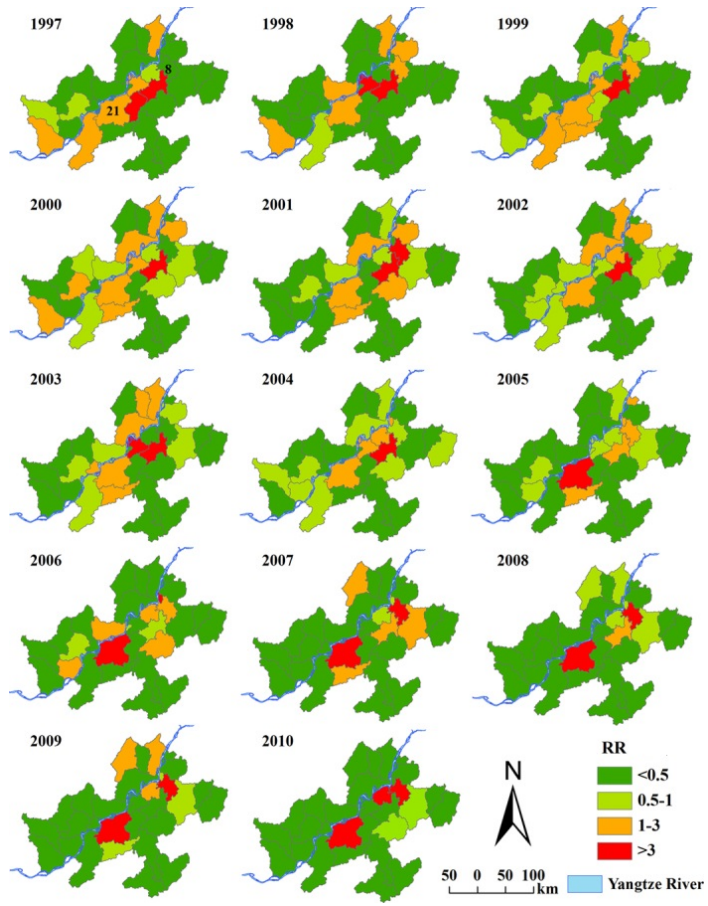
Temporal trend and space-time interaction:

- Temporal trends: $\phi_t \sim \text{Normal}(\phi_{t-1}, \sigma^2_\phi)$;
  $\gamma_t \sim \text{Normal}(0, \sigma^2_\gamma)$
- Space-time interaction: $\psi_{it} \sim \text{Normal}(0, \sigma^2_\psi)$

Knorr-Held, L., "Bayesian modelling of inseparable space-time variation in disease risk", **Statistics in Medicine**, 2000, 19(17-18): 2555:2567.

Example: schistosomiasis japonica in East China

- Spatio-temporal model of schistosomiasis japonica;
- Interested in evaluating changes in time (discover policy effects).



Hu, Y et al., "Monitoring schistosomiasis risk in East China over space and time using a Bayesian hierarchical modeling approach", **Scientific Reports**, 2012, 6: 24173.

Adding another dimension

- Detect spatio-temporal patterns according to the different age groups $j = 1, \ldots, J$

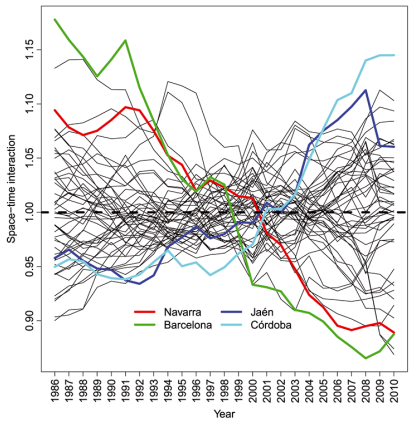
$$\log \lambda_{itj} \quad = \quad \alpha + u_i + \phi_t + \xi_j + \psi_{it}^1 + \psi_{jt}^2 + \psi_{ij}^3 + \zeta_{itj}$$

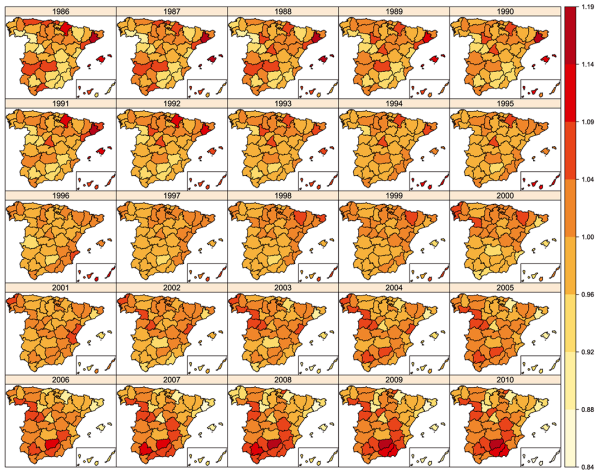
- Autoregressive structure on u_i;
- Random walk on ϕ_t and ξ_j;
- Structured interactions:
  $\rightarrow$ temporal trends are different for far apart regions but tend to be similar for adjacent regions;
  $\rightarrow$ temporal trends and spatial patterns are similar for subsequent age classes.

Goicoa T et al., "Age-space-time CAR models in Bayesian disease mapping", **Statistics in Medicine**, 2016, 35(14):2391-2405.

Example: Age-space-time CAR models in disease mapping

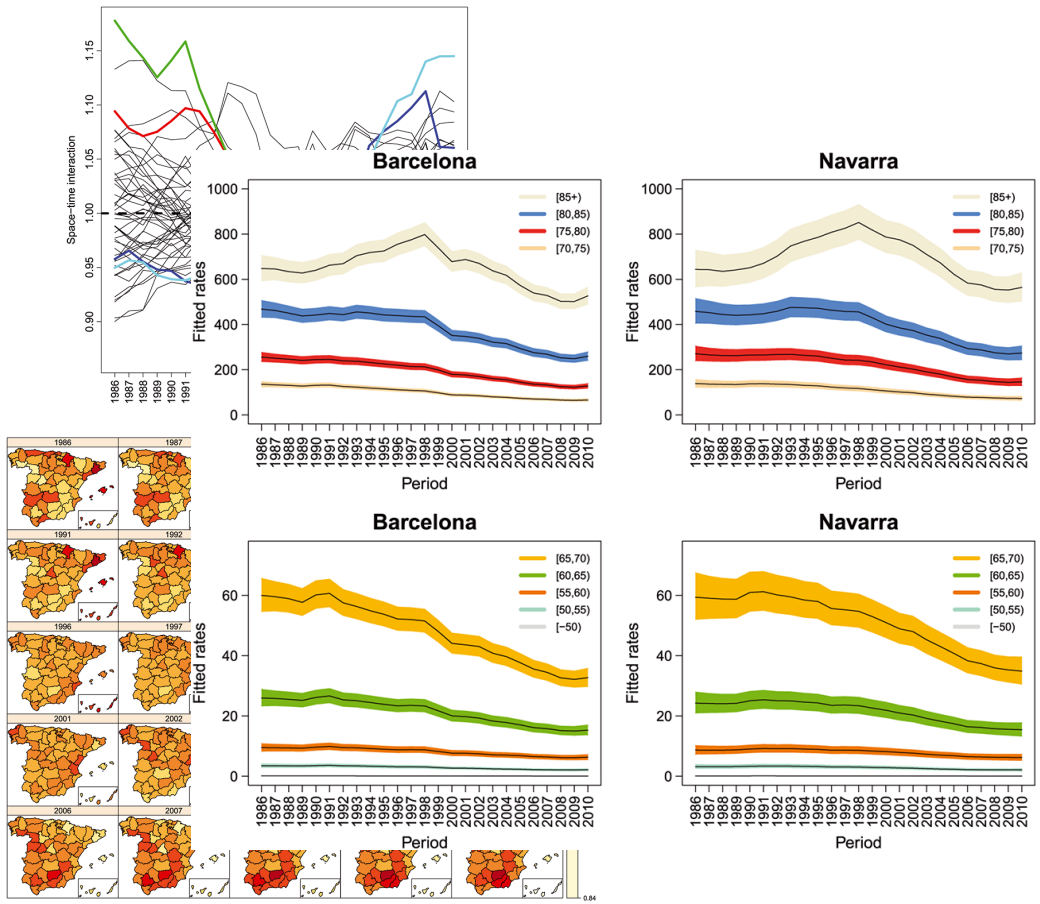
- Prostate cancer mortality in Spain.
- Data are available in 50 provinces and 9 age groups during 25 years between 1986 and 2010.





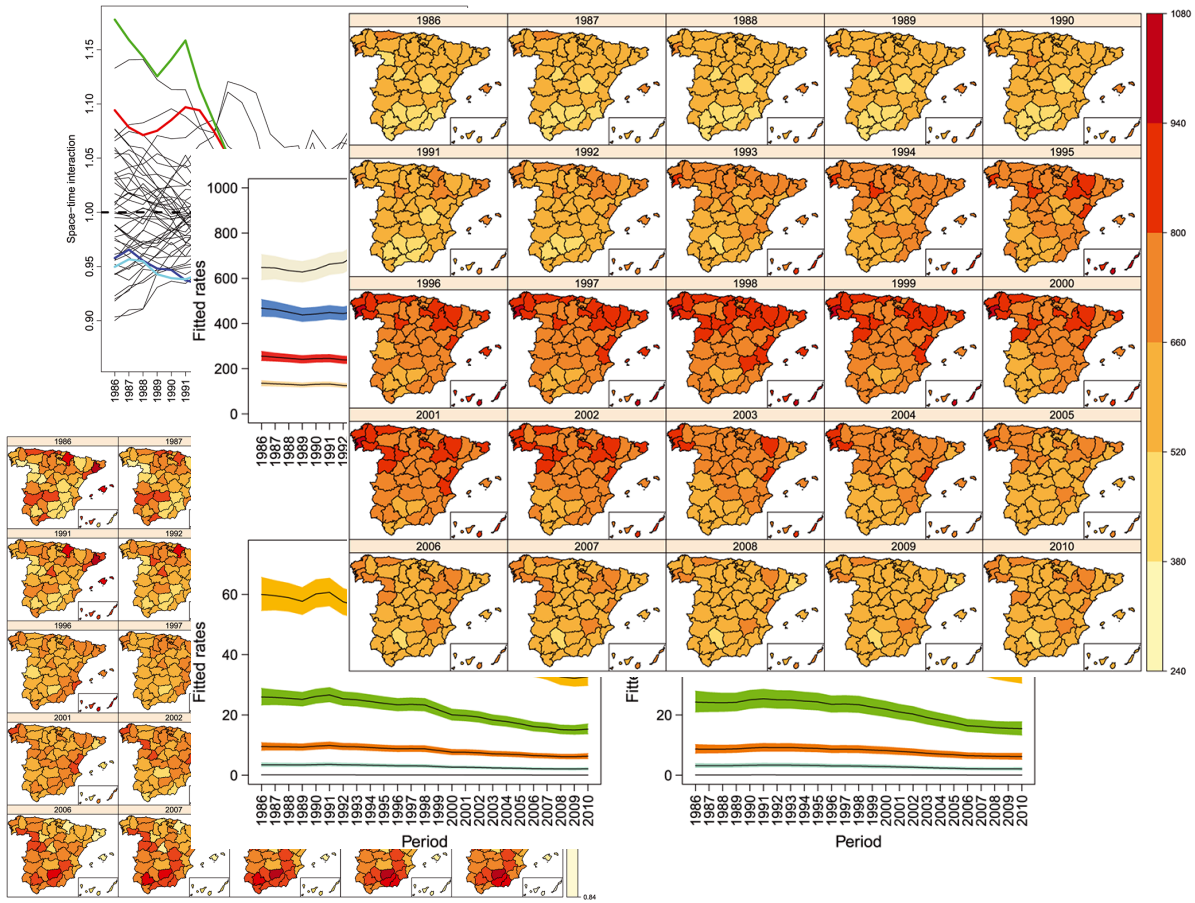
Example: Age-space-time CAR models in disease mapping

- Prostate cancer mortality in Spain.
- Data are available in 50 provinces and 9 age groups during 25 years between 1986 and 2010.



Example: Age-space-time CAR models in disease mapping

- Prostate cancer mortality in Spain.
- Data are available in 50 provinces and 9 age groups during 25 years between 1986 and 2010.



Spatio-temporal mixture model

- Extend the spatio-temporal hierarchical framework to a mixture model with two components
  - accounts for spatial and temporal correlation;
  - able to detect areas with trend different from the national one (unusual).

$$\mathrm{O}_{it} \sim \mathrm{Poisson}(\lambda_{it}\, \mathrm{E}_{it})$$

areas $i = 1, \ldots, I$; time points $t = 1, \ldots, T$.

$$\log(\lambda_{it}) = p_{it}\, \log\big(\, \lambda_{it}^{\mathrm{C}}\, \big) + (1 - p_{it})\, \log\big(\, \lambda_{it}^{\mathrm{AS}}\, \big)$$

Common Trend

- overall intercept
- spatial component
- temporal component

Area-Specific Trend

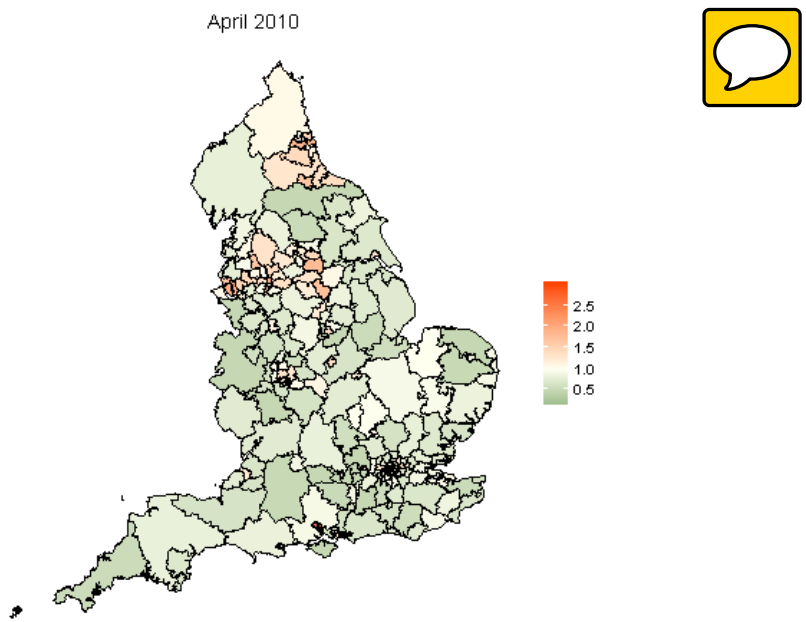
- area-specific intercept
- area-specific temporal component

Li G., et al., "BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice", **Biostatistics**, 2012, 13(4): 695-710.

Example: Unusual trend detection for COPD

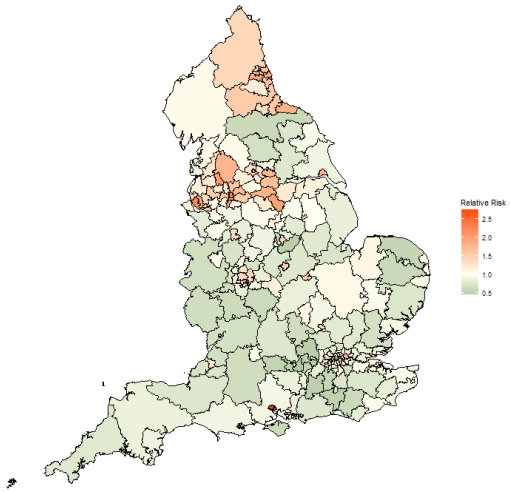
COPD hospital admissions.
Spatial resolution: 211 Clinical Commissioning Groups (CCGs)
Temporal resolution: monthly data, April 2010 - March 2011.

Figure: HES SMRs

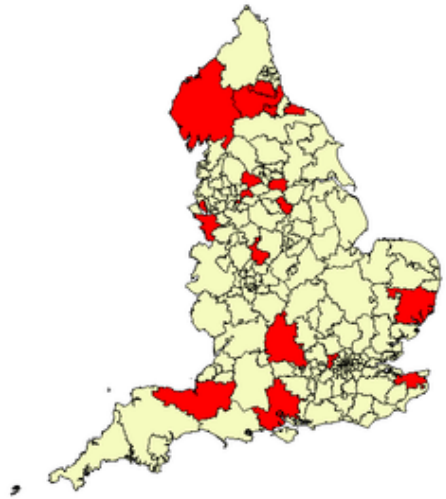


Results: Spatial pattern

Spatial Posterior Mean

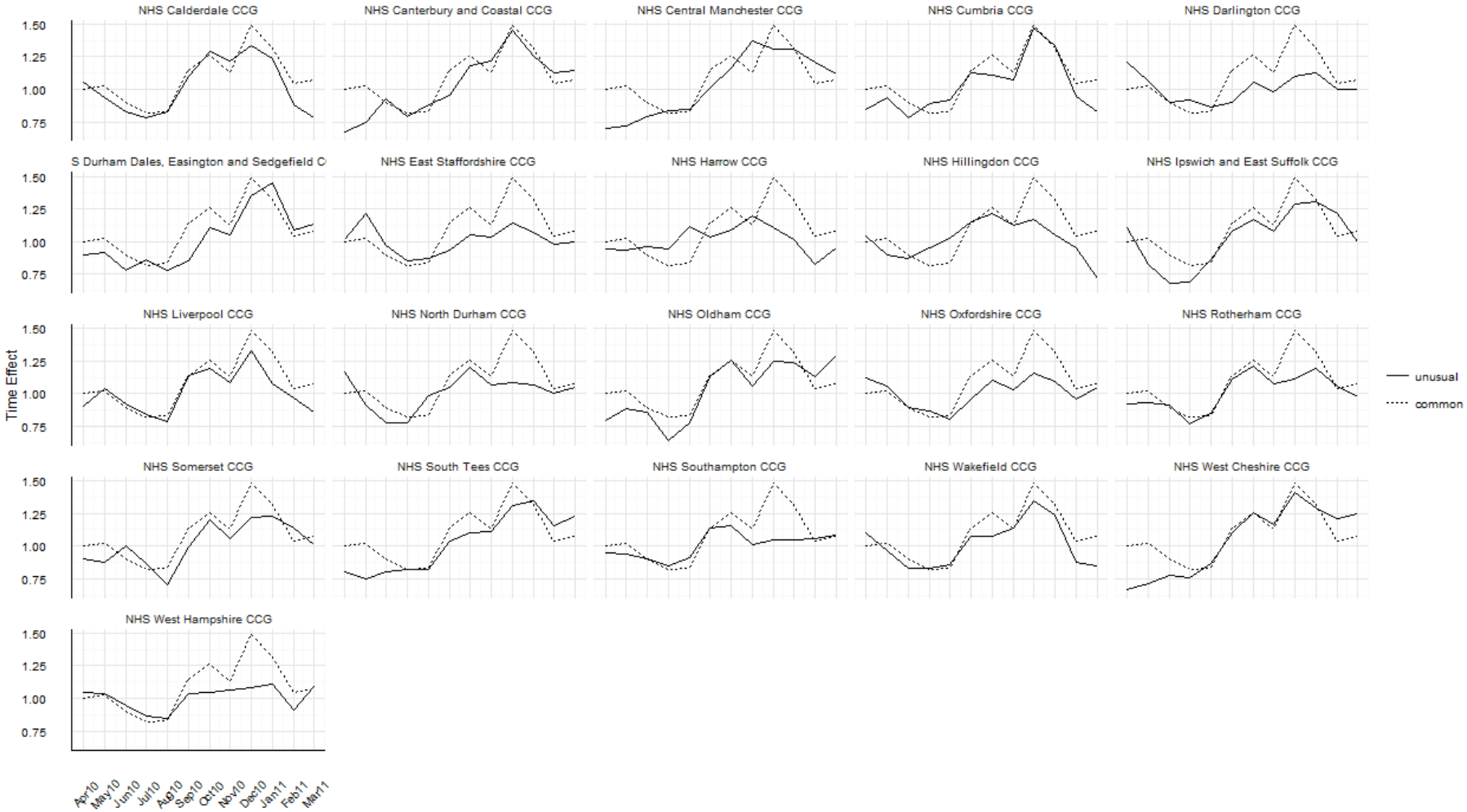


Unusual Areas



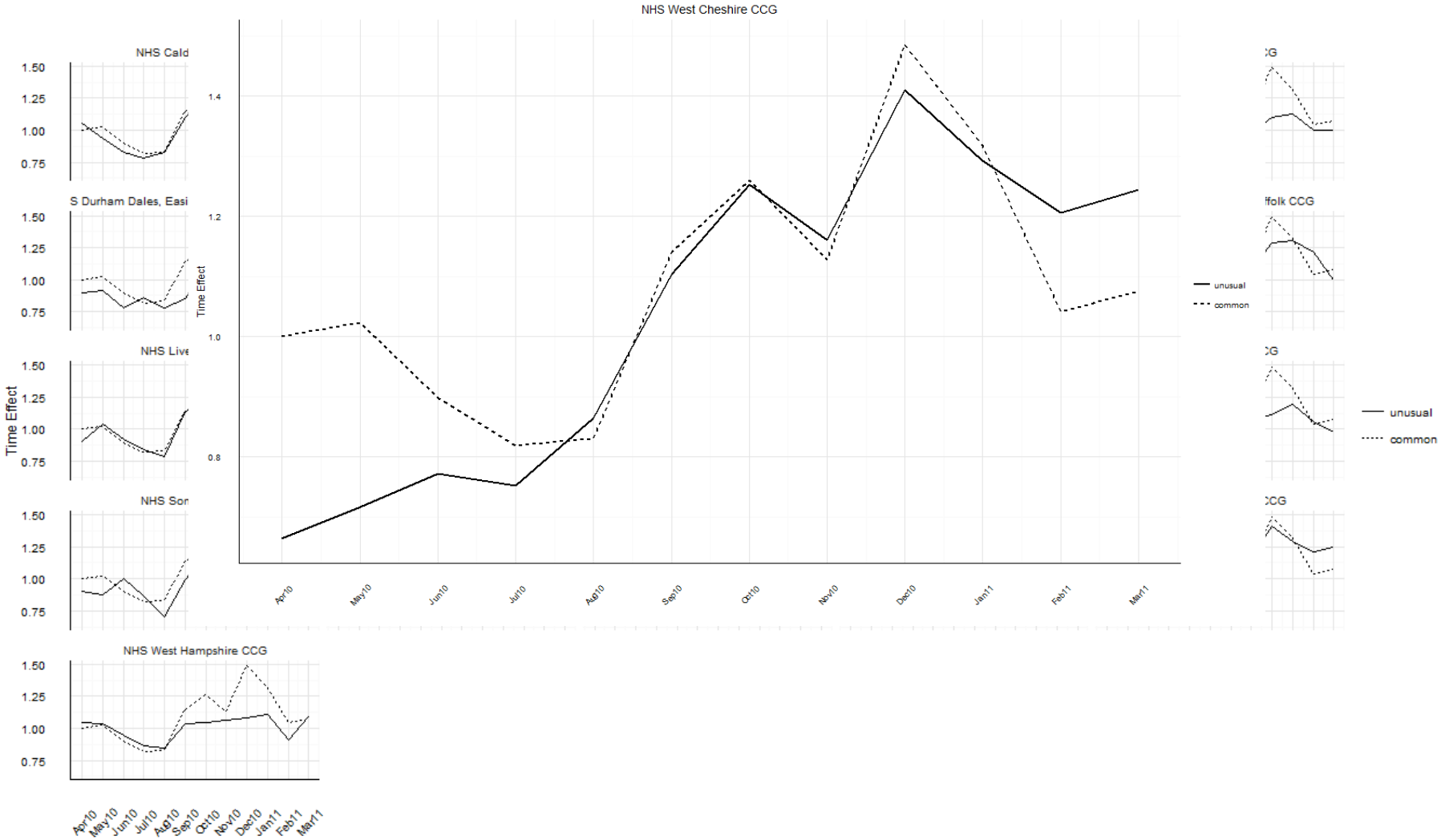
- Higher risk in the north of England across the time period.
- Unusual areas: Mostly isolated but some clustered in the North of England.

Results: Unusual trends



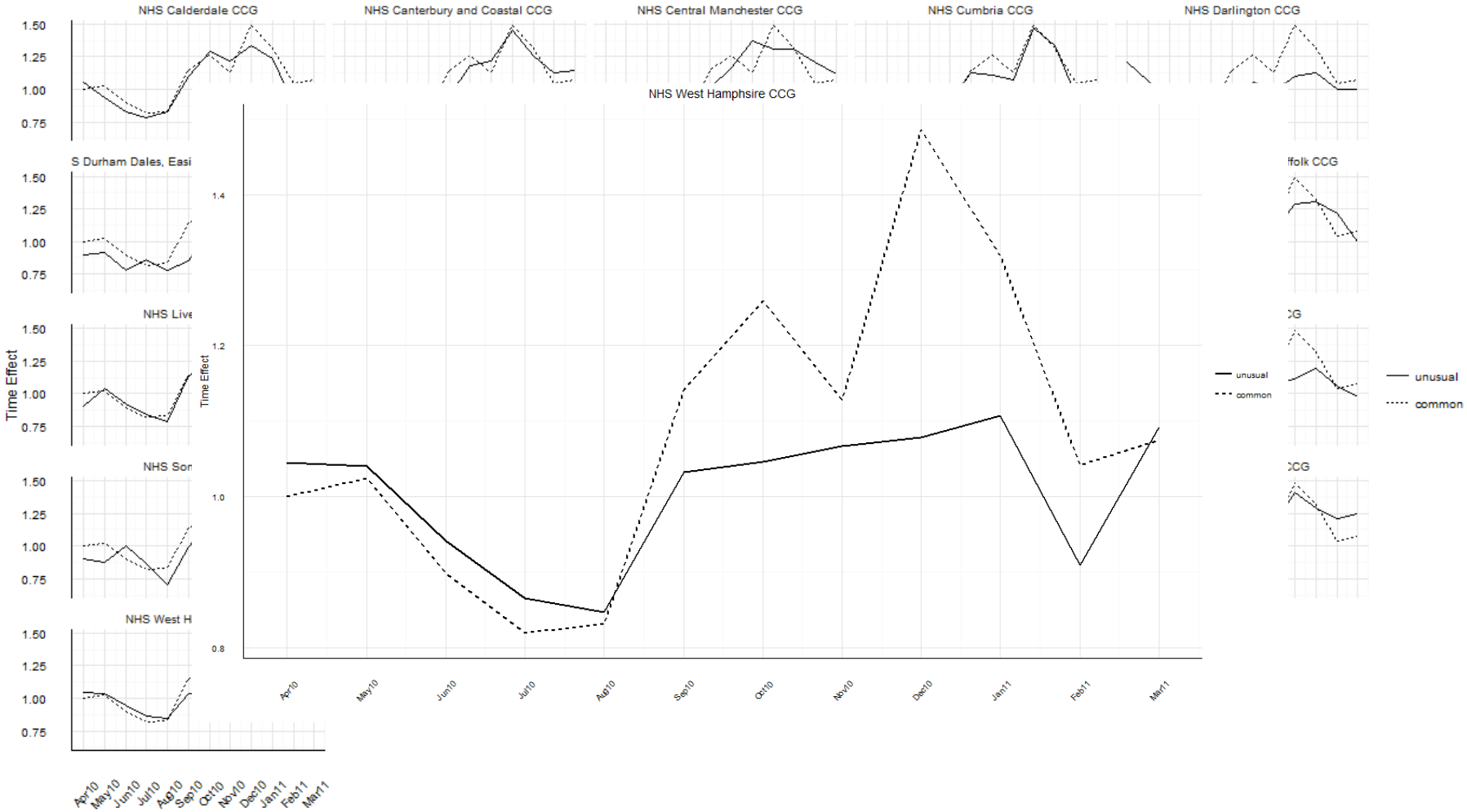
Boulieri A, et al. "Investigating trends in asthma and COPD through multiple data sources: A small area study", **Spatial and SpatioTemporal Epidemiology**, 2016, 19: 28-36.

Results: Unusual trends



Boulieri A, et al. "Investigating trends in asthma and COPD through multiple data sources: A small area study", **Spatial and SpatioTemporal Epidemiology**, 2016, 19: 28-36.

Results: Unusual trends

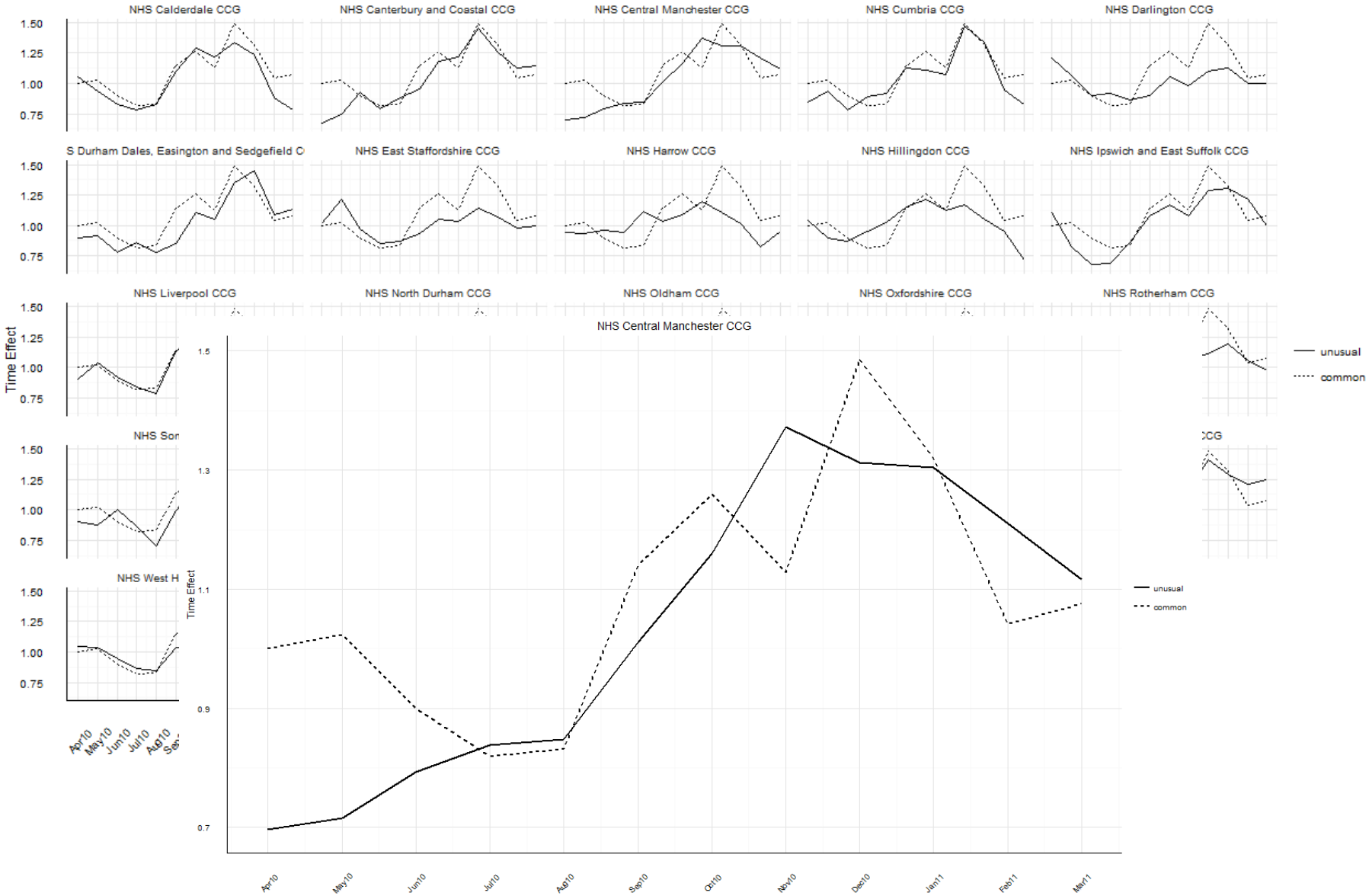


Boulieri A, et al. "Investigating trends in asthma and COPD through multiple data sources: A small area study", **Spatial and SpatioTemporal Epidemiology**, 2016, 19: 28-36.

Results: Unusual trends

Boulieri A, et al. "Investigating trends in asthma and COPD through multiple data sources: A small area study", **Spatial and SpatioTemporal Epidemiology**, 2016, 19: 28-36.

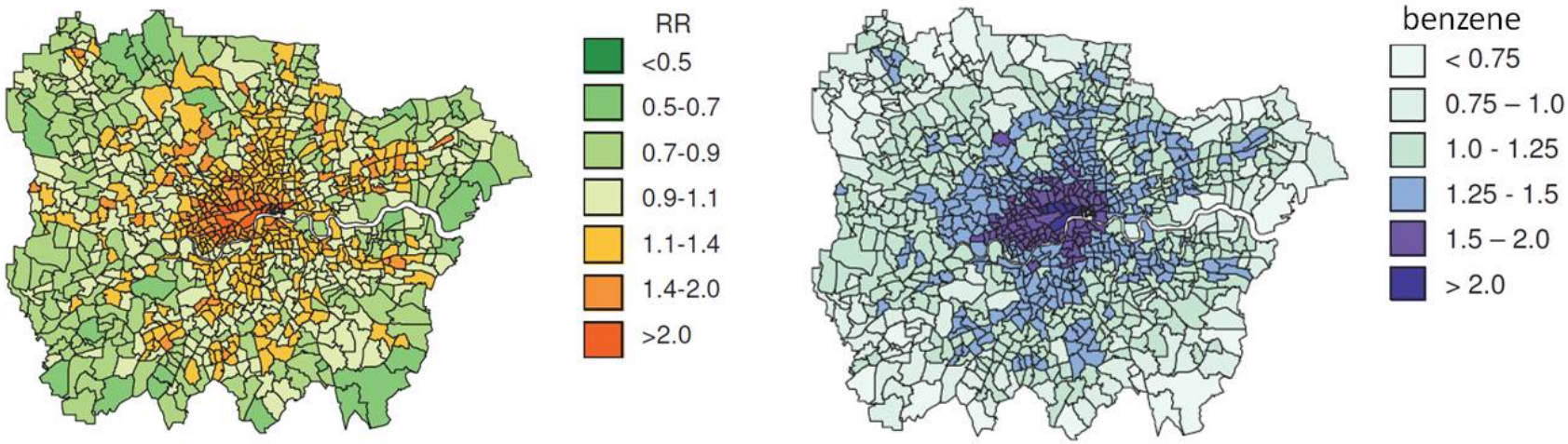
Results: Unusual trends



Boulieri A, et al. "Investigating trends in asthma and COPD through multiple data sources: A small area study", **Spatial and SpatioTemporal Epidemiology**, 2016, 19: 28-36.

Example: Childhood leukaemia and benzene

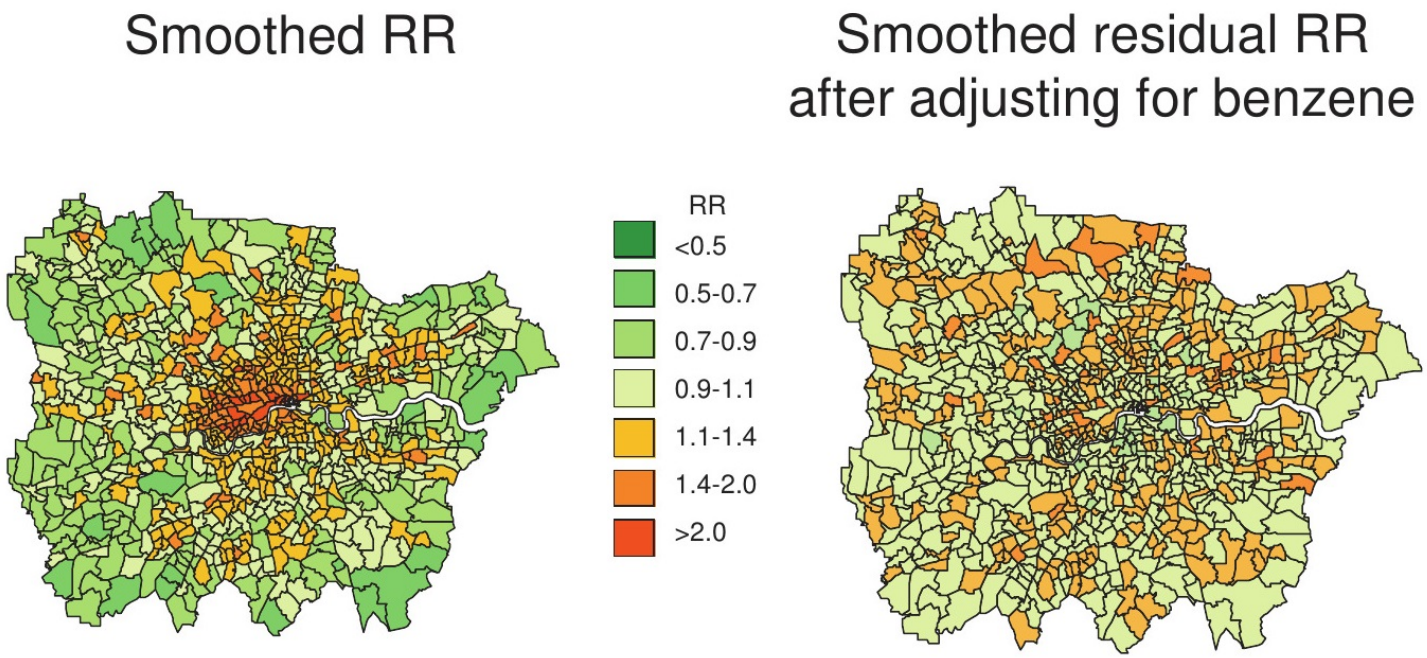
- Bayesian Disease mapping used to study leukaemia incidence in children.
- Can we explain some of the variation in risk of leukaemia by environmental exposure to benzene?
- Let X_i = average benzene emissions (tonnes per annum) in ward i.



Maps of leukaemia RR

- Disease mapping to ecological regression.

$$\log \lambda_i \quad = \quad \alpha + u_i + v_i + \beta X_i$$



Best, N. et al. (2001). "Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias". **J Roy Statist Soc, Series A**, 164, 155-74.

Example: Mortality Effects of Warm Temperature across Communities

- Investigate spatial patterns of temperature response for cardio-respiratory mortality across the 376 districts of England and Wales.

Case-crossover design - for individual i, case/control j:

$$\begin{aligned} O_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \log(\lambda_{ij}) &= \text{f}(\beta_0, \beta_{1\text{dist}_i}, \text{Temp}_{ij}) + \delta_i + \text{C}_{ij} \end{aligned}$$

- $O_{ij} = 1$ for cases ($j = 1$) and 0 for controls ($j = 2, \ldots, J_i$);
- δ_i links cases and controls from the same individual i;
- β_0 is the threshold parameter for temperature response;
- $\beta_{1\text{dist}_i}$ is the supra-threshold slope parameter.

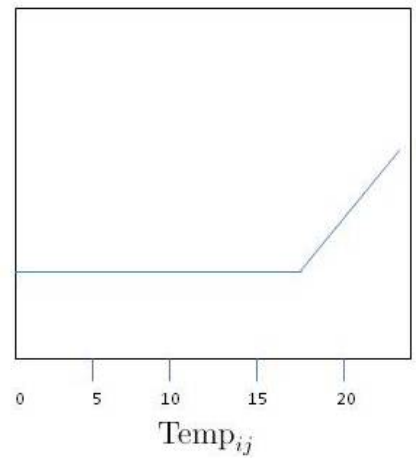
Confounders are

$$\text{C}_{ij} = \beta_{\text{PM}_{10}}\text{PM}_{10ij} + \beta_{\text{Holiday}}\text{Holiday}_{ij}$$

Bayesian Model: temperature model

The temperature threshold model $\mathrm{f}(\beta_0, \beta_{1\mathrm{dist}_i}, \mathrm{Temp}_{ij})$ is:

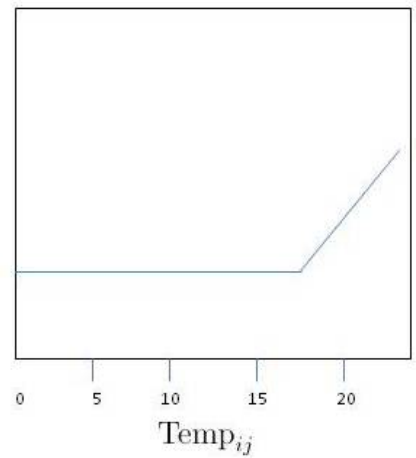
$$\begin{cases} 0 & \text{if } \mathrm{Temp}_{ij} \leq \beta_0 \\ \beta_{1\mathrm{dist}_i}(\mathrm{Temp}_{ij} - \beta_0) & \text{if } \mathrm{Temp}_{ij} > \beta_0 \end{cases}$$



Bayesian Model: temperature model

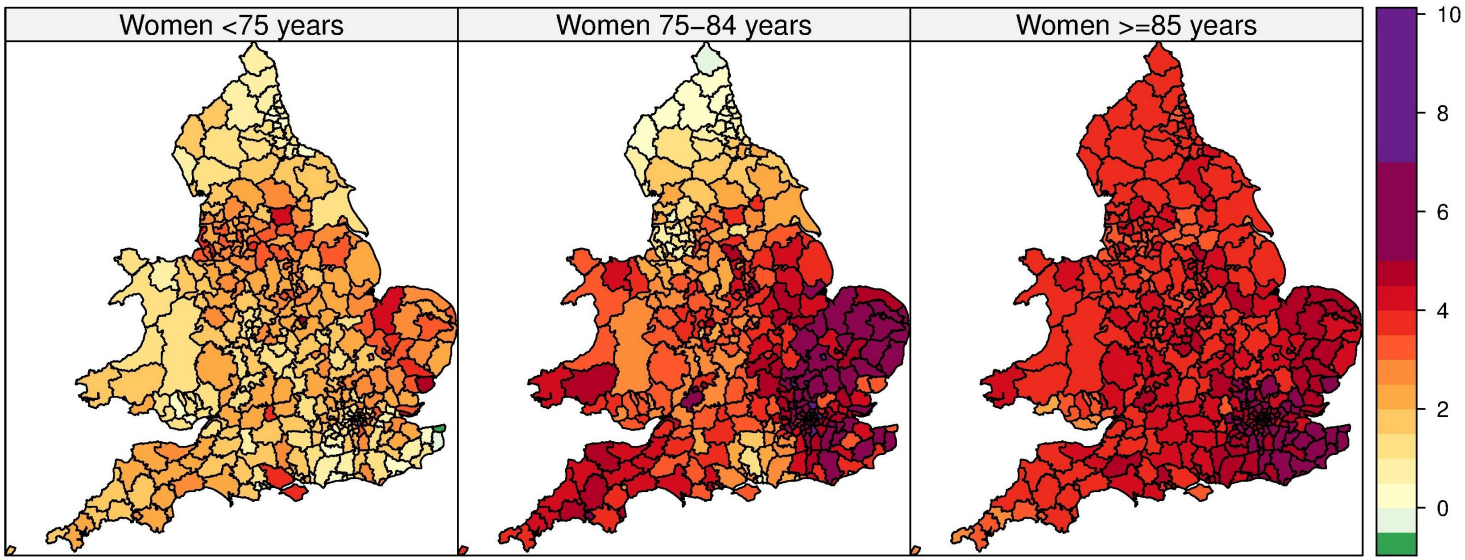
The temperature threshold model $\mathrm{f}(\beta_0, \beta_{1\mathrm{dist}_i}, \mathrm{Temp}_{ij})$ is:

$$\begin{cases} 0 & \text{if } \mathrm{Temp}_{ij} \leq \beta_0 \\ \beta_{1\mathrm{dist}_i}(\mathrm{Temp}_{ij} - \beta_0) & \text{if } \mathrm{Temp}_{ij} > \beta_0 \end{cases}$$



- Fixed nationwide threshold.
- District specific supra-threshold slope parameter modelled via a combination of structured and unstructured random effects.
- Stratified analysis for age classes and gender.

Spatial Pattern of Temperature Effects in Females

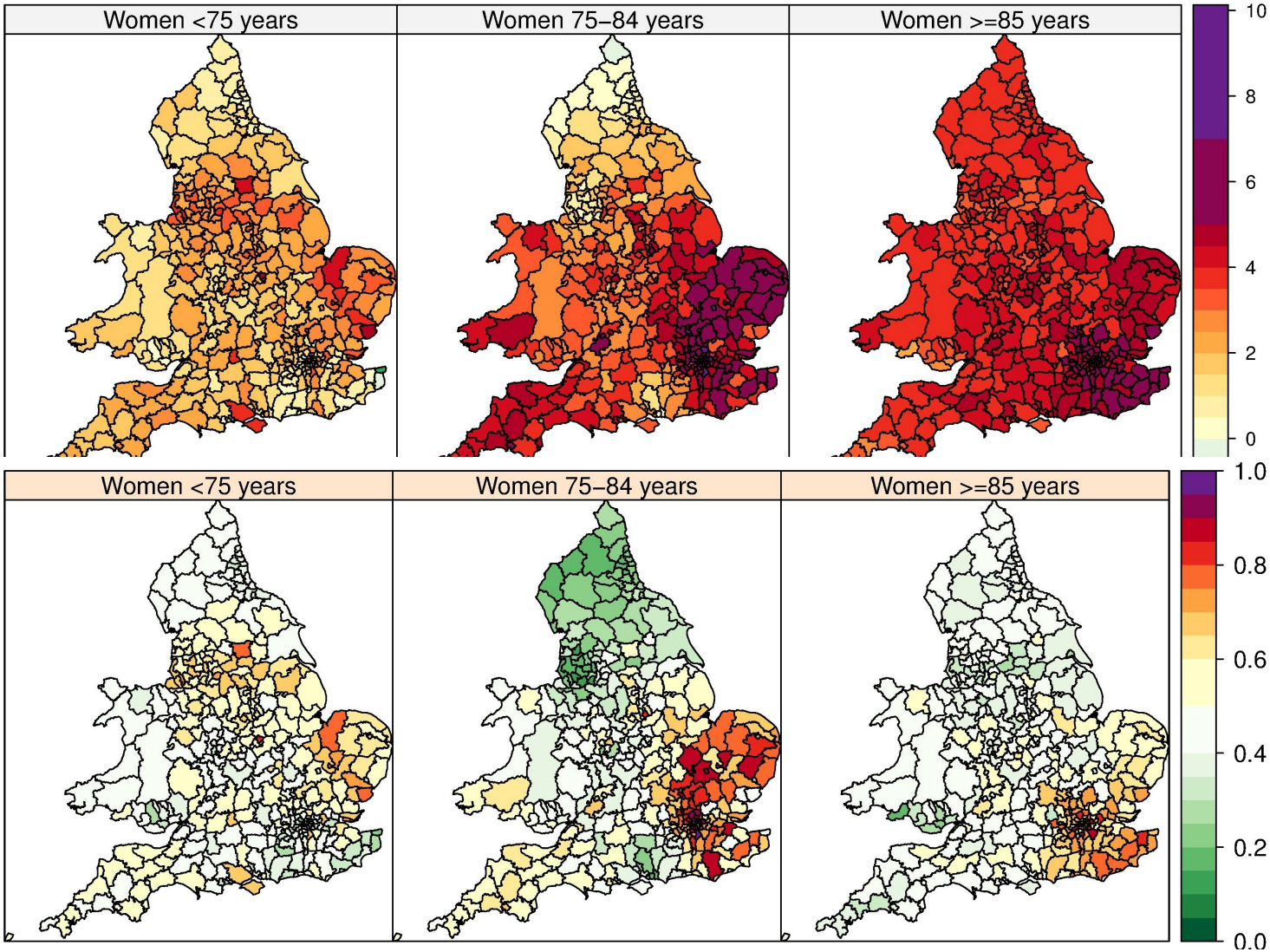


Consistent spatial pattern with age:

- Higher % increase in odds of death in the South.
- Consistent with other papers on the topic.

Bennett J, et al., "Vulnerability to the mortality effects of warm temperature in the districts of England and Wales", **Nature Climate Change** 2014, 4, 269-273.

Spatial Pattern of Temperature Effects in Females



Bennett J, et al., "Vulnerability to the mortality effects of warm temperature in the districts of England and Wales", **Nature Climate Change** 2014, 4, 269-273.

Recap: Small area framework

- Hierarchical Bayesian framework useful to deal with discrete spatial (temporal) data.
- Disease mapping: To smooth the data and look for spatial and temporal patterns.
- Mixture models: To classify areas based on their temporal trends.
- Regression models for risk assessment.

Geostatistical framework

General inferential issues

Spatial (and temporal) pattern suggest that observations close to each other have more similar values than those far from each other.
What are the aims of a geostatistical analysis?

- Reconstruct a latent spatial (temporal) surface $\boldsymbol{S}$ from a finite set of noisy observations and their spatial (temporal) location.
- Use the spatial dependence to predict values of the spatial surface (together with associated uncertainty) at locations where there are no observations.
  $\rightarrow$ **In this talk**: Spatial pattern of LF disease.

In addition:

- Do we have to deal with different spatial resolution?
  $\rightarrow$ Change of support.
  $\rightarrow$ **In this talk**: Misalignment in air pollution study.

The common framework to geostatistical models is that of Gaussian random fields.

Gaussian fields

- A spatial process $y(\boldsymbol{s})$ is a Gaussian field (GF) if for any $n \geq 1$ and for each set of locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, the vector $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))$ and spatially structured covariance matrix $\boldsymbol{\Sigma}$.
- The generic element of $\boldsymbol{\Sigma}$ is defined by a covariance function $\mathcal{C}(\cdot, \cdot)$ such that $\Sigma_{ij} = \text{Cov}\left(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)\right) = \mathcal{C}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j))$.

Gaussian fields

- A spatial process $y(\boldsymbol{s})$ is a Gaussian field (GF) if for any $n \geq 1$ and for each set of locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, the vector $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))$ and spatially structured covariance matrix $\boldsymbol{\Sigma}$.
- The generic element of $\boldsymbol{\Sigma}$ is defined by a covariance function $\mathcal{C}(\cdot, \cdot)$ such that $\Sigma_{ij} = \text{Cov}\left(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)\right) = \mathcal{C}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j))$.
- The spatial process is called second-order stationary if
  - $\boldsymbol{\mu}$ is constant (i.e. $\mu(\boldsymbol{s}_i) = \mu$ for each i)
  - the spatial covariance function depends only on the distance vector $(\boldsymbol{s}_i - \boldsymbol{s}_j) \in \mathbb{R}^2$, i.e. $\text{Cov}\left(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)\right) = \mathcal{C}(\boldsymbol{s}_i - \boldsymbol{s}_j)$.
- Moreover, a stationary process is isotropic if the covariance does not depend on the direction but just on the Euclidean distance $||\boldsymbol{s}_i - \boldsymbol{s}_j|| \in \mathbb{R}$.

  Several functions are available for the spatial covariance function (eg exponential, Matérn, spherical, etc.) parameterized by some parameters (eg spatial variance, range, etc.).

Model for geostatistical data

Data are measured (possibly with error) at n spatial locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ and are denoted by $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)) = (y_1, \ldots, y_n)$. Usually the following model is assumed

$$y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \xi(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$

where

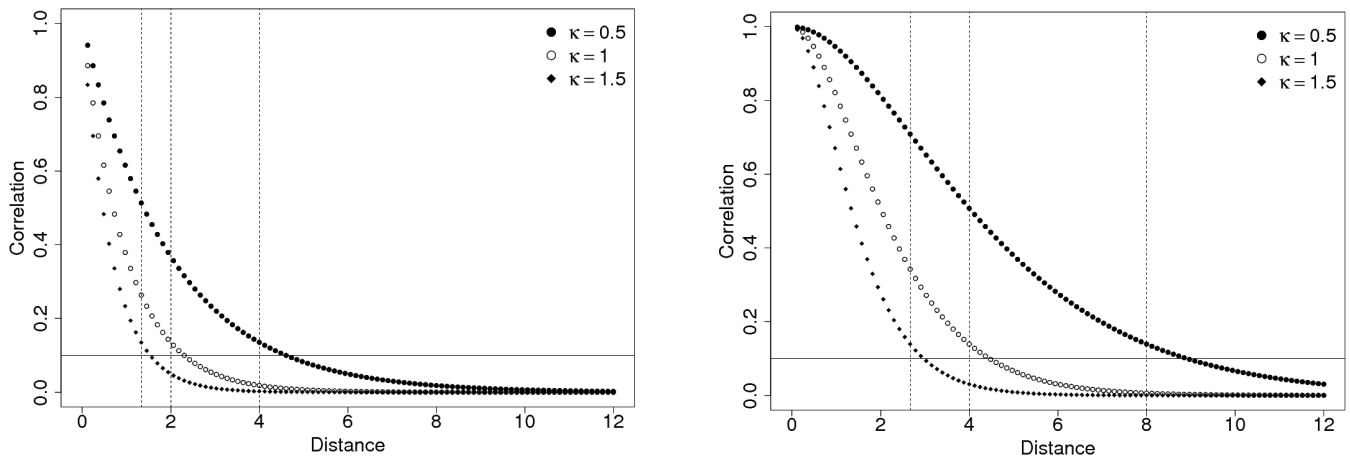
- $\mu(\boldsymbol{s})$ is the so-called large scale component, defined by some covariates.
- $\xi(\boldsymbol{s})$ is a zero mean Gaussian spatial process commonly assumed to be stationary and isotropic with covariance function $Cov(\xi(\boldsymbol{s}_i), \xi(\boldsymbol{s}_j))$ which depends only on the distance between the locations.
- $\epsilon(\boldsymbol{s})$ represents the measurement error and its variance is usually known as nugget effect.

Matèrn covariance function

$$\text{Cov}(\xi(\boldsymbol{s}_i), \xi(\boldsymbol{s}_j)) = \text{Cov}(\xi_i, \xi_j) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}} \left(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||\right)^{\lambda} K_{\lambda}\left(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||\right)$$

where

- $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ is the Euclidean distance between two generic locations $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^d$
- σ^2 is the variance
- K_{λ} denotes the modified Bessel function of second kind and order $\lambda > 0$, which measures the degree of smoothness of the process.
- $\kappa > 0$ is a scale parameter related to the range r, i.e. the distance at which the spatial correlation becomes almost null.



- The Matérn family is a very flexible class of covariance functions able to cover a wide range of spatial fields.

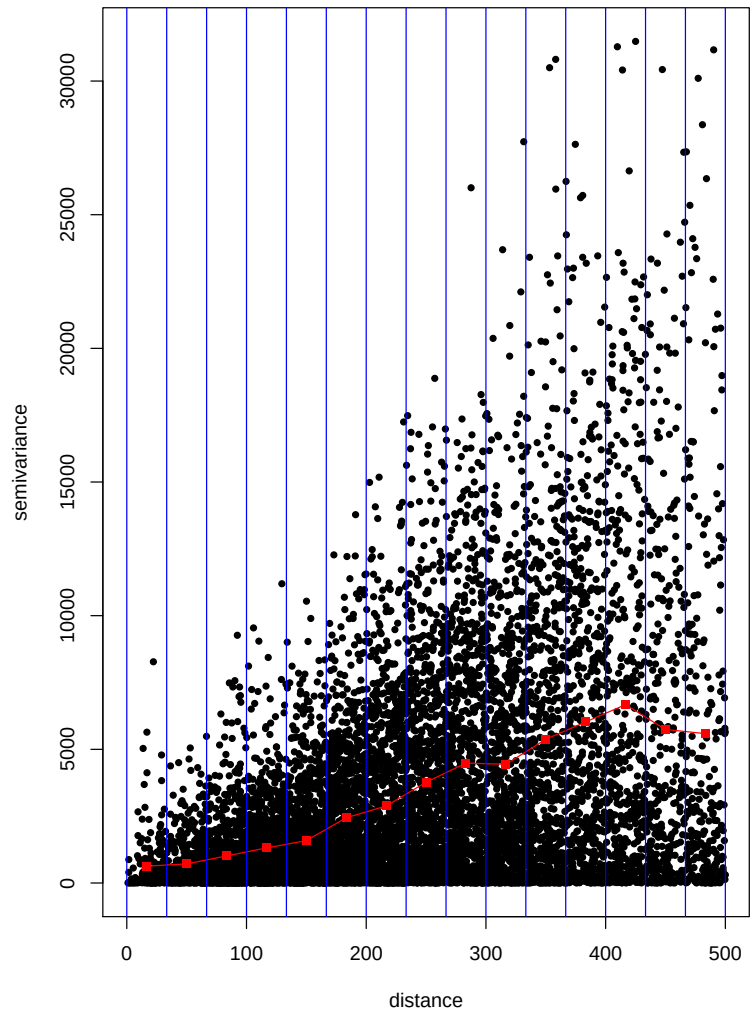
Variogram

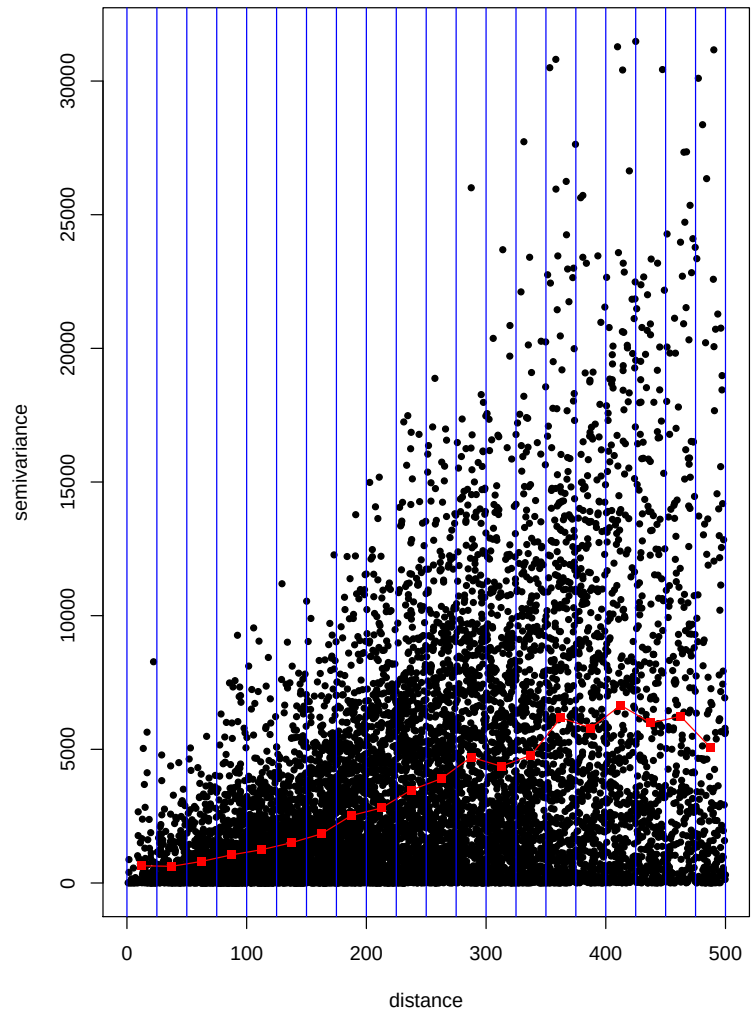
- In order to explore the decay of the spatial covariance with distance, empirical plots are useful
- The empirical variogram is a tool to visualise spatial correlation. Given a variable $y(\boldsymbol{s})$ measured at a set of locations $\boldsymbol{s}$, the empirical (semi)-variogram is defined as

$$\gamma_{ij} = \frac{1}{2}(y(\boldsymbol{s}_i) - y(\boldsymbol{s}_j))^2$$

- Values of γ_{ij} are then plotted against the distances between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ for every pair of locations to produce a variogram cloud.
- To aid interpretation the empirical variogram is often computed by averaging γ_{ij} within distance bands.

Variogram for the rain fall in Brazil example (15 (left) or 20 (right) distance classes)





Interpreting the Variogram

- The variogram measures half the squared difference between each pair of locations.

- If the data are spatially correlated, we would expect observations close together to be more similar than observations far apart (squared differences for observations close together will be smaller than for observations further apart so expect the variogram to gradually increase with increasing distance between observations).

- If data are independent, then on average we would expect the squared difference of a pair of observations that are close together to be similar to that of observations further apart (variogram should be approximately constant as distance increases).

Possible approaches for estimation

- Classical geostatistical approach: the empirical variogram is used as an exploratory tool and the mean and the covariance function parameters are estimated through least square methods usually adopting a two step procedure (first the mean is estimated and then residuals are used to make inference on the spatial parameters)
  $\rightarrow$ very simple and not computationally intensive;
  $\rightarrow$ restrictive in the models that can accommodate;
  $\rightarrow$ two-step approach to prediction;
  $\rightarrow$ difficult to account for uncertainty.

Possible approaches for estimation

- Classical geostatistical approach: the empirical variogram is used as an exploratory tool and the mean and the covariance function parameters are estimated through least square methods usually adopting a two step procedure (first the mean is estimated and then residuals are used to make inference on the spatial parameters)
  $\rightarrow$ very simple and not computationally intensive;
  $\rightarrow$ restrictive in the models that can accommodate;
  $\rightarrow$ two-step approach to prediction;
  $\rightarrow$ difficult to account for uncertainty.

- (Bayesian) Model-based approach: hierarchical model specification and posterior predictive distribution for prediction;
  $\rightarrow$ easy to allow for uncertainty;
  $\rightarrow$ flexible in accounting for any type of distribution;
  $\rightarrow$ high computationally costs, especially for factorizing the spatial covariance matrix $\mathbf{\Sigma}$ (this is known as "big n problem").

Example: Lymphatic filariasis in Africa

- Lymphatic filariasis (LF) - major vector-borne parasitic disease endemic to the tropics, including sub-Saharan Africa.
- First pan-african spatial analysis of environmental factors of LF (altitude, temperature, rain, pop. density).
- Bayesian approach - posterior predictive distribution to predict on a regular grid

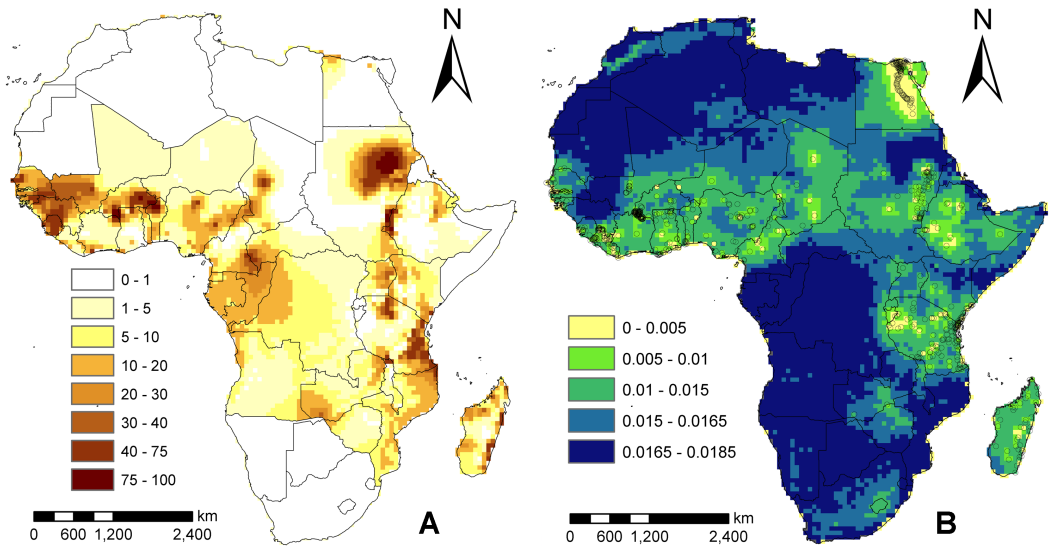


Figure: A: Posterior mean of LF prevalence; B: Uncertainty estimate

Slater H, Michael E (2013) Mapping, Bayesian Geostatistical Analysis and Spatial Prediction of Lymphatic Filariasis Prevalence in Africa. PLoS ONE 8(8):e71574.

Example:Air pollution spatio-temporal exposure model

$y(\boldsymbol{s}_j, t)$: square root of air pollution concentration at monitoring station located at site $\boldsymbol{s}_j$ $(j = 1, \ldots, J)$ and time $t = 1, \ldots, T$. We assume:

$$\begin{aligned} y(\boldsymbol{s}_j, t) &= \mu(\boldsymbol{s}_j, t) + \omega(\boldsymbol{s}_j, t) + \epsilon(\boldsymbol{s}_j, t) \\ \mu(\boldsymbol{s}_j, t) &= b_0 + b_1 \text{Temp}(\boldsymbol{s}_j, t) + b_2 \text{Baseline}(\boldsymbol{s}_j) \\ \epsilon(\boldsymbol{s}_j, t) &\sim \text{Normal}(0, \sigma_e^2) \end{aligned}$$

where

Example:Air pollution spatio-temporal exposure model

$y(\boldsymbol{s}_j, t)$: square root of air pollution concentration at monitoring station located at site $\boldsymbol{s}_j$ $(j = 1, \ldots, J)$ and time $t = 1, \ldots, T$. We assume:

$$\begin{aligned} y(\boldsymbol{s}_j, t) &= \mu(\boldsymbol{s}_j, t) + \omega(\boldsymbol{s}_j, t) + \epsilon(\boldsymbol{s}_j, t) \\ \mu(\boldsymbol{s}_j, t) &= b_0 + b_1 \text{Temp}(\boldsymbol{s}_j, t) + b_2 \text{Baseline}(\boldsymbol{s}_j) \\ \epsilon(\boldsymbol{s}_j, t) &\sim \text{Normal}(0, \sigma_e^2) \end{aligned}$$

where

- $\omega(\boldsymbol{s}_j, t)$ is a latent spatio-temporal process characterised by AR(1) temporal dynamics: which changes in time with first order autoregressive dynamics with coefficient a and spatially correlated innovations, given by

$$\omega(\boldsymbol{s}_j, t) = a\omega(\boldsymbol{s}_j, t-1) + \xi(\boldsymbol{s}_j, t)$$

with $\xi(\boldsymbol{s}_j, t)$ being a zero-mean spatial continuous process with Matérn covariance function.

Example:Air pollution spatio-temporal exposure model

$y(\boldsymbol{s}_j, t)$: square root of air pollution concentration at monitoring station located at site $\boldsymbol{s}_j$ $(j = 1, \ldots, J)$ and time $t = 1, \ldots, T$. We assume:

$$\begin{aligned} y(\boldsymbol{s}_j, t) &= \mu(\boldsymbol{s}_j, t) + \omega(\boldsymbol{s}_j, t) + \epsilon(\boldsymbol{s}_j, t) \\ \mu(\boldsymbol{s}_j, t) &= b_0 + b_1 \text{Temp}(\boldsymbol{s}_j, t) + b_2 \text{Baseline}(\boldsymbol{s}_j) \\ \epsilon(\boldsymbol{s}_j, t) &\sim \text{Normal}(0, \sigma_e^2) \end{aligned}$$

where

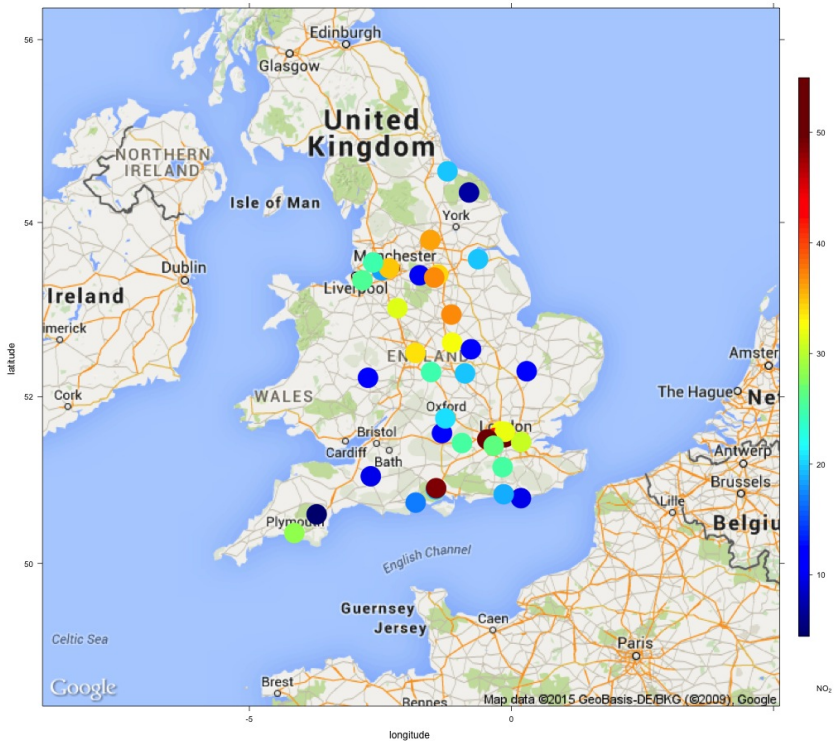
- $\omega(\boldsymbol{s}_j, t)$ is a latent spatio-temporal process characterised by AR(1) temporal dynamics: which changes in time with first order autoregressive dynamics with coefficient a and spatially correlated innovations, given by

  $$\omega(\boldsymbol{s}_j, t) = a\omega(\boldsymbol{s}_j, t-1) + \xi(\boldsymbol{s}_j, t)$$

  with $\xi(\boldsymbol{s}_j, t)$ being a zero-mean spatial continuous process with Matérn covariance function.
- $\mu(\boldsymbol{s}_j, t) + \omega(\boldsymbol{s}_j, t) = x(\boldsymbol{s}_j, t)$ is the latent field.

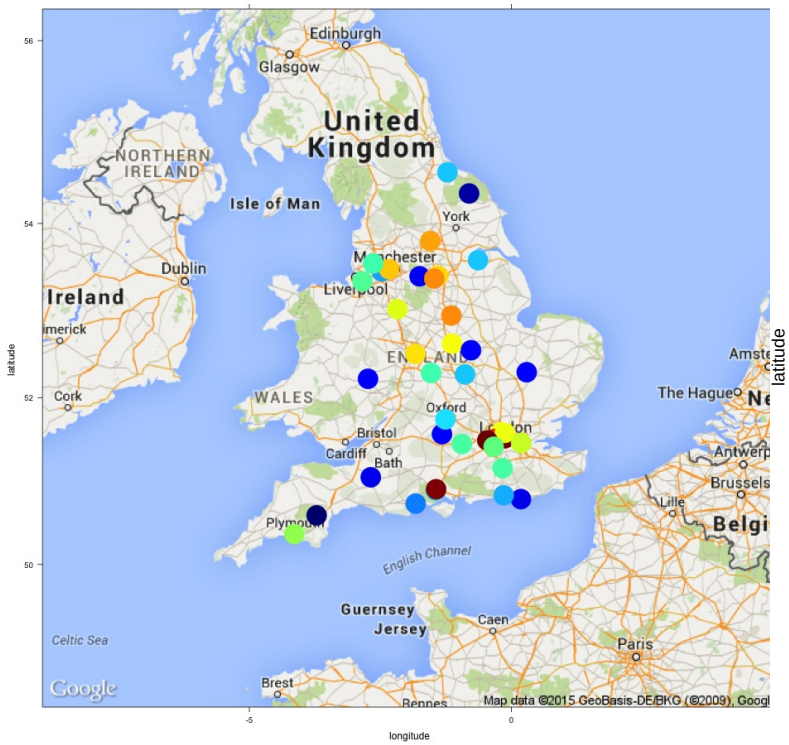
Spatial prediction: $x(\boldsymbol{s}_j, t) \rightsquigarrow x(\boldsymbol{s}_l^\star, t)$

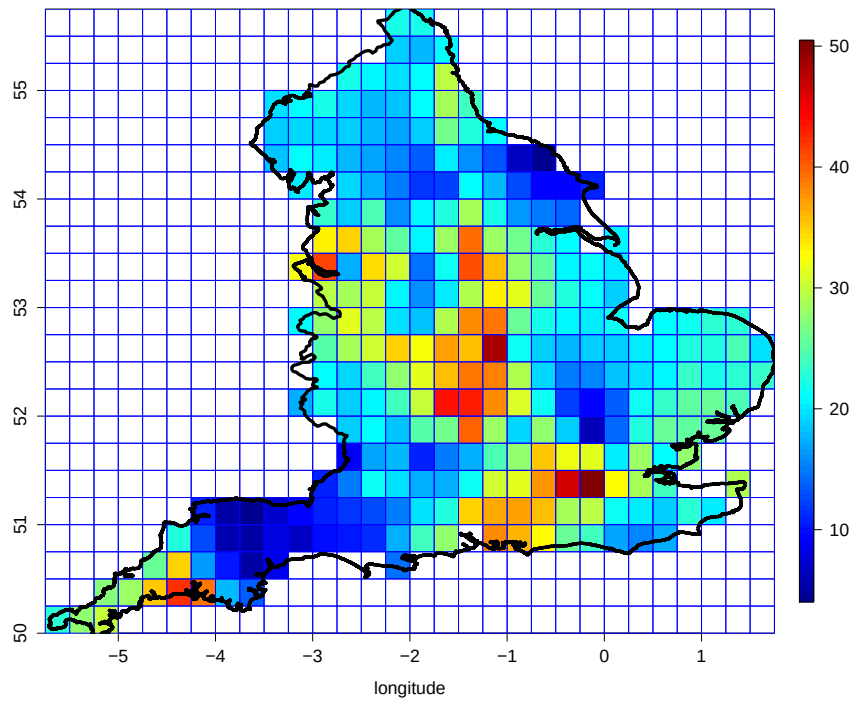
- $x(\boldsymbol{s}_j, t)$ are estimated through the model;
- To cover the spatial domain, for all the time points it is possible to sample from the posterior predictive distributions for each point $\boldsymbol{s}_l^\star$ in a regular grid.



Spatial prediction: $x(\boldsymbol{s}_j, t) \rightsquigarrow x(\boldsymbol{s}_l^\star, t)$

- $x(\boldsymbol{s}_j, t)$ are estimated through the model;
- To cover the spatial domain, for all the time points it is possible to sample from the posterior predictive distributions for each point $\boldsymbol{s}_l^\star$ in a regular grid.





Change of support

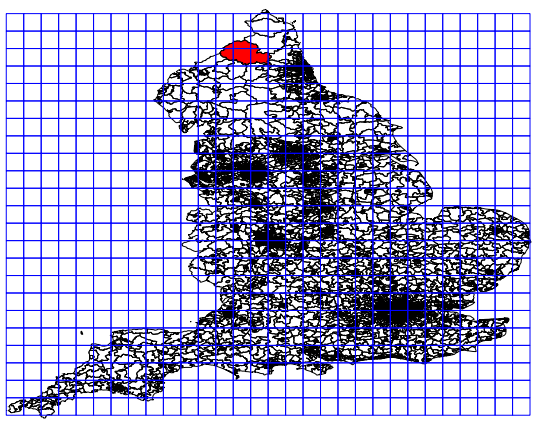
- Air pollutant concentrations available at monitoring stations in a study region, while disease data appear as counts observed in given areas.

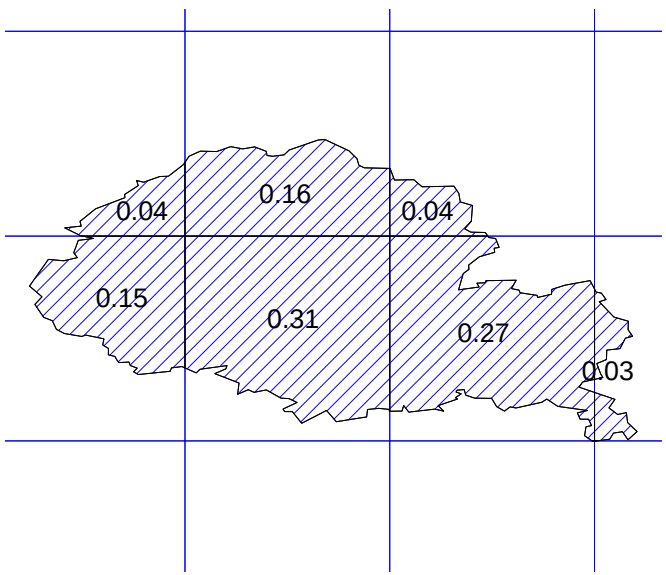
Change of support

- Air pollutant concentrations available at monitoring stations in a study region, while disease data appear as counts observed in given areas.
- *Change of support*:

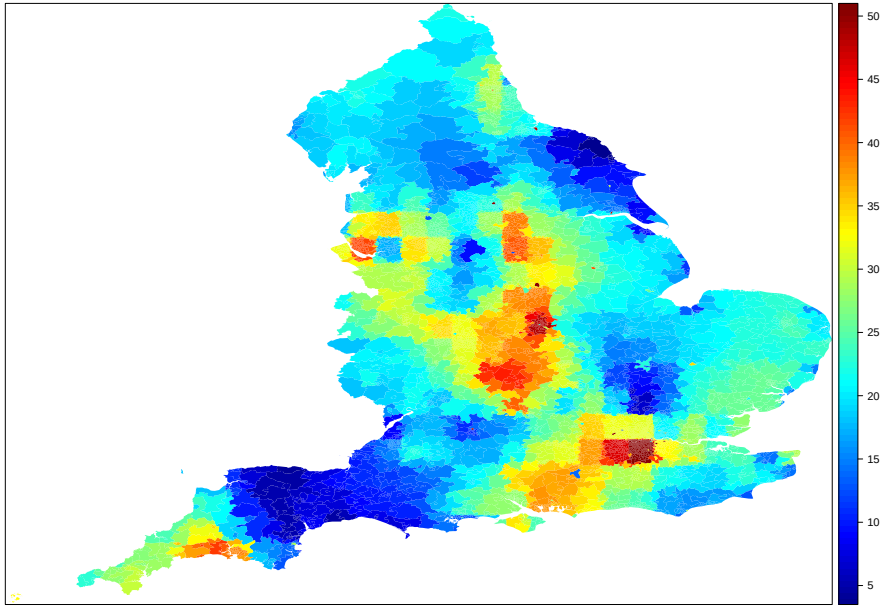
$$x_{it} = \int_{\boldsymbol{s} \in \mathcal{A}_i} x(\boldsymbol{s}, t)\, p(\boldsymbol{s})\, d\boldsymbol{s} \approx \sum_{l=1}^{N_i} x(\boldsymbol{s}^\star_{il}, t)\, p(\boldsymbol{s}^\star_{il})$$

where $x(\boldsymbol{s}^\star_{ij}, t)$ is the pollutant concentration prediction at location $s^\star_{il}$, one of N_i regular grid points inside area $\mathcal{A}_i$; moreover $\sum_{l=1}^{N_i} p(\boldsymbol{s}^\star_{il}) = 1$.

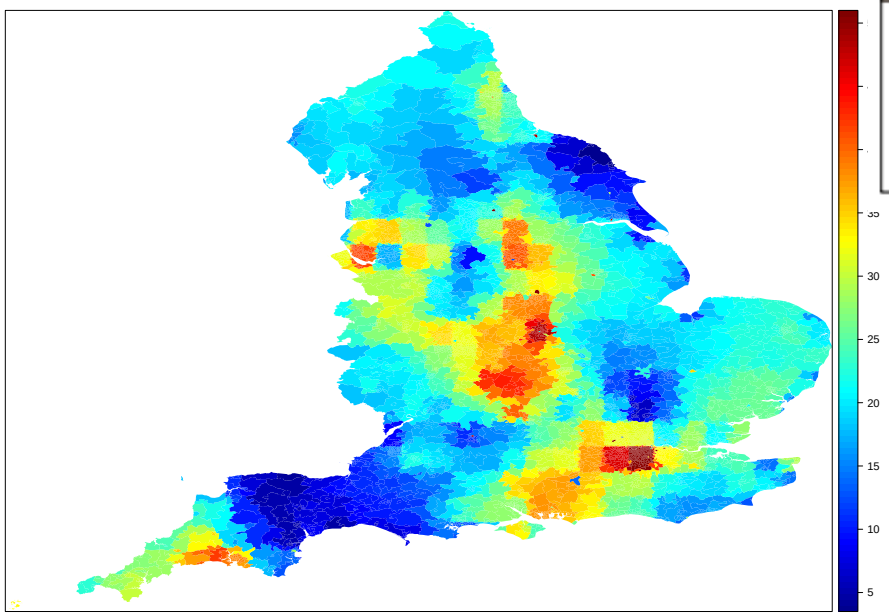


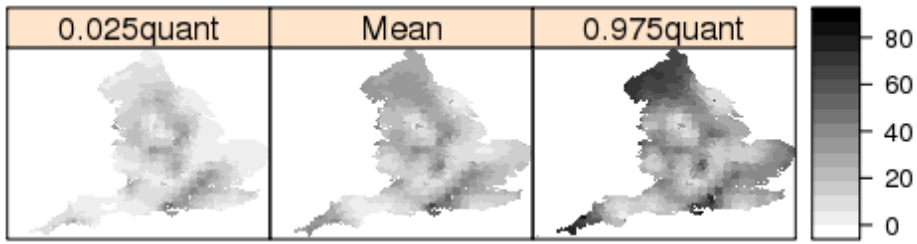


Spatial prediction: $x(\boldsymbol{s}_l^\star, t) \rightsquigarrow x_{it}$



Spatial prediction: $x(\boldsymbol{s}_l^\star, t) \rightsquigarrow x_{it}$





This estimates can be used for risk assessment in epidemiological studies.

What about uncertainty?

Recap

- Bayesian hierarchical modelling to deal with geostatistical framework
-

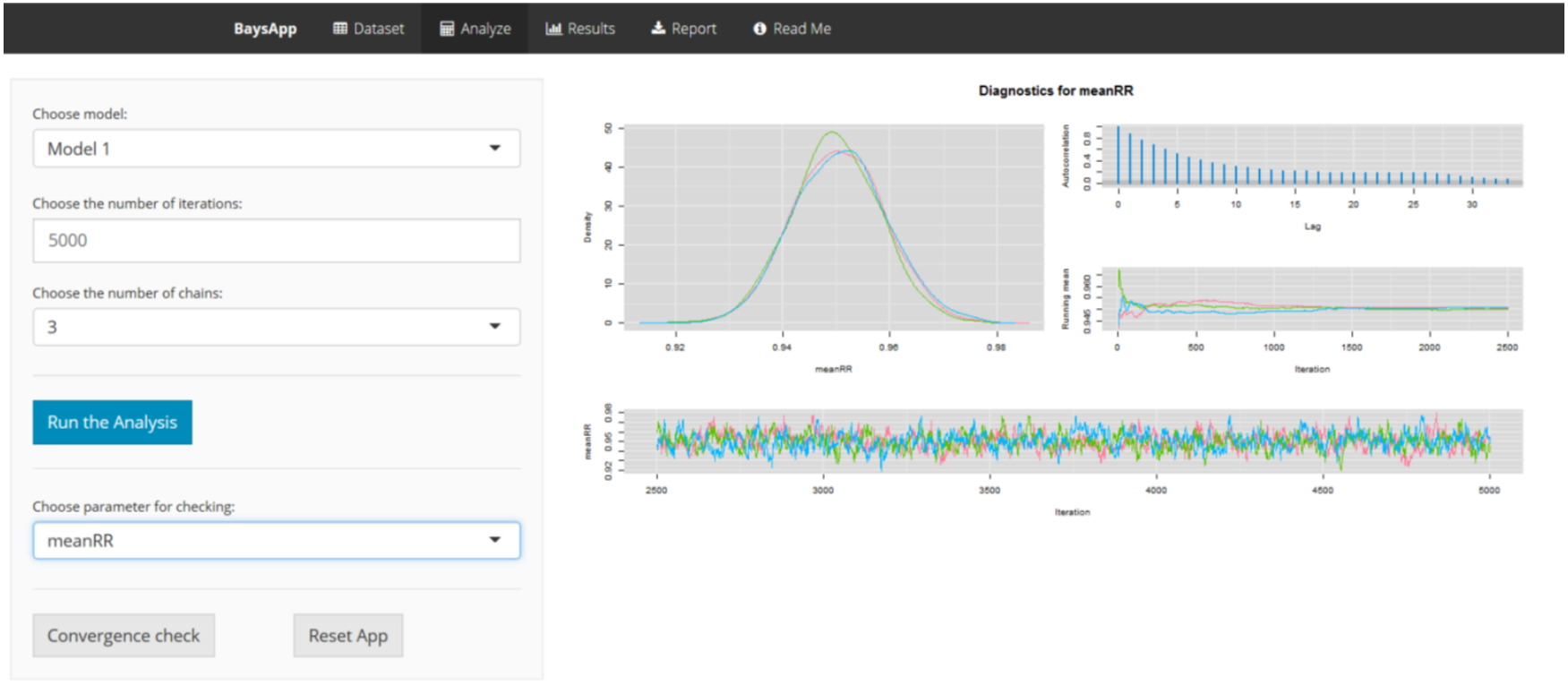
Software

1. MCMC
   - Lots of R specific packages: ex.
     `CARbayes`, spatial analysis with conditional autoregressive prior
     `geoR`, geostatistical data analysis
     `spTimer`, hierarchical models for spatio-temporal processes
   - Stand-alone general software: ex. `BUGS` (Gibbs Sampling), `Stan` (Hamiltonian Monte Carlo)
   - R packages of interface with the stand-alone software: ex. `R2OpenBUGS`, `RStan`

2. Appromixate methods
   - `R-INLA`, Laplace Approximation for Gaussian Markov Random Fields.

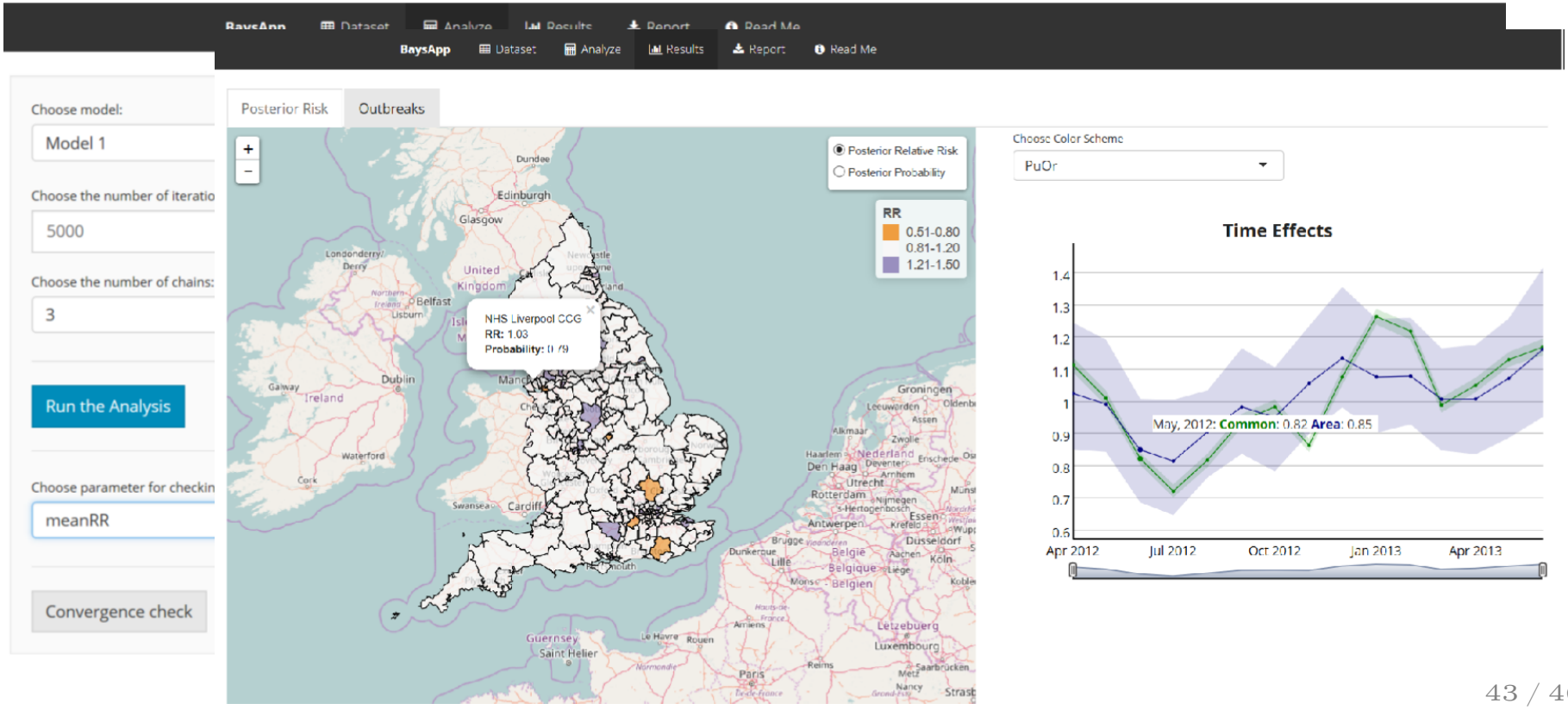
Challenge: dissemination - methods

- Key aspect to try and propose a method that can be used by other researchers (not necessarily statisticians).
- Reproducibility.
- R packages are great - sometimes difficult for public health/social scientists.
- R-shiny webapp is a step ahead.



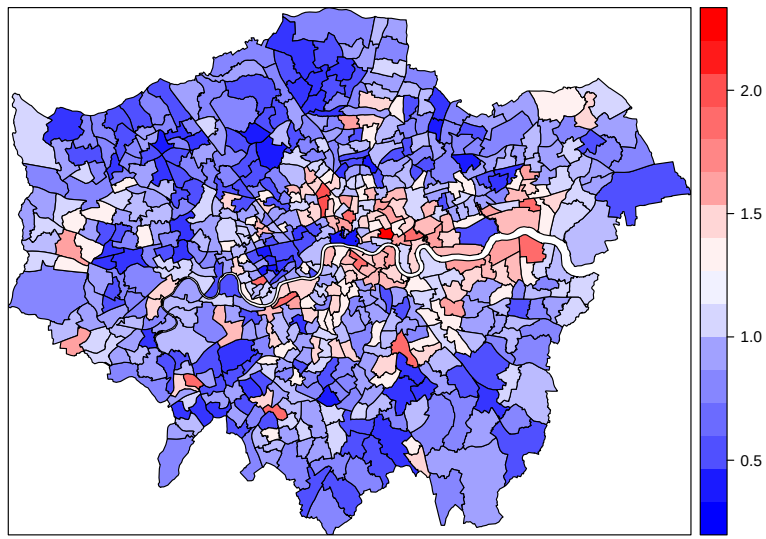
Challenge: dissemination - methods

- Key aspect to try and propose a method that can be used by other researchers (not necessarily statisticians).
- Reproducibility.
- R packages are great - sometimes difficult for public health/social scientists.
- R-shiny webapp is a step ahead.



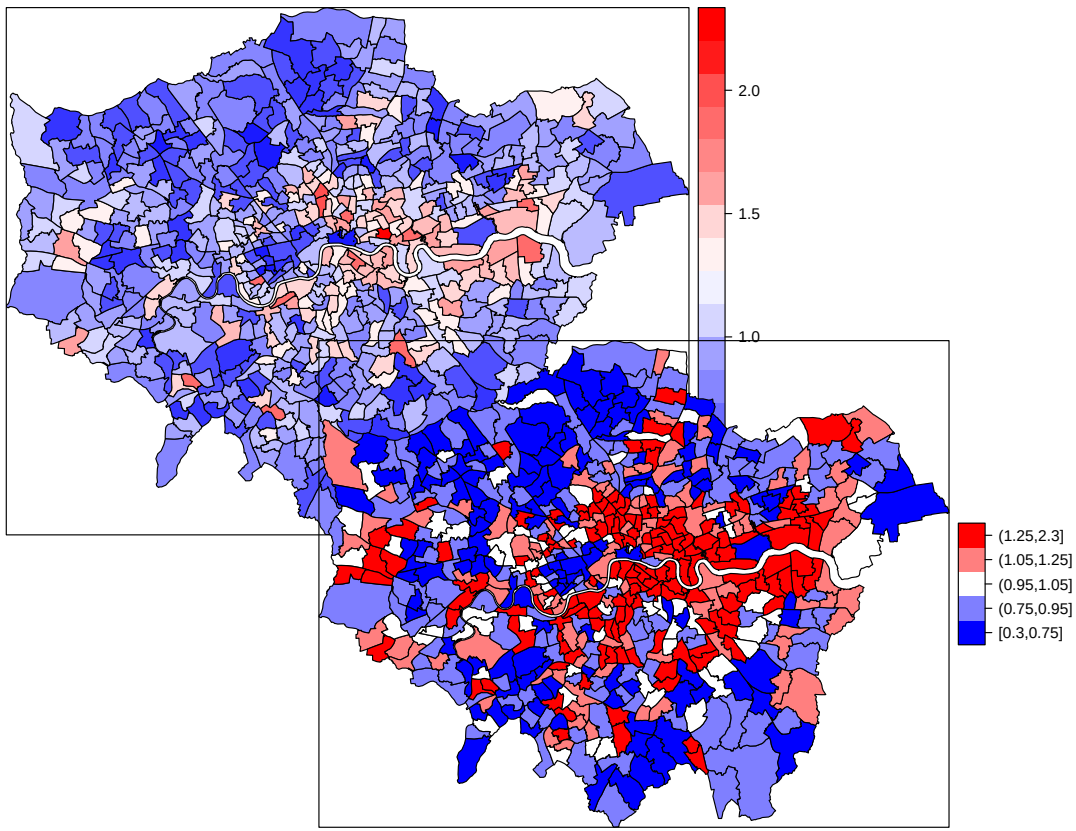
Challenge: dissemination - maps

- Cartography - scale and colors can give a different message



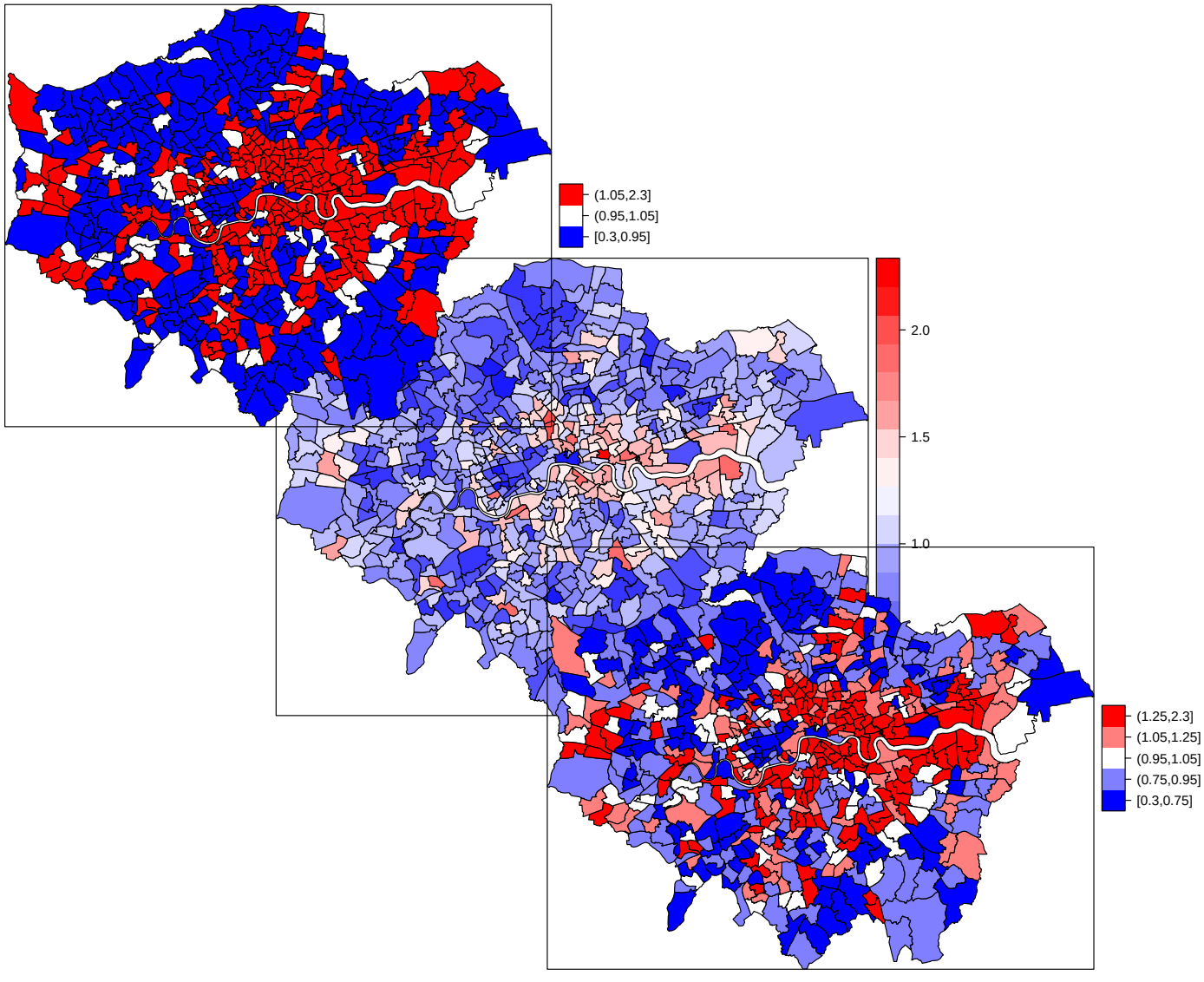
Challenge: dissemination - maps

- Cartography - scale and colors can give a different message



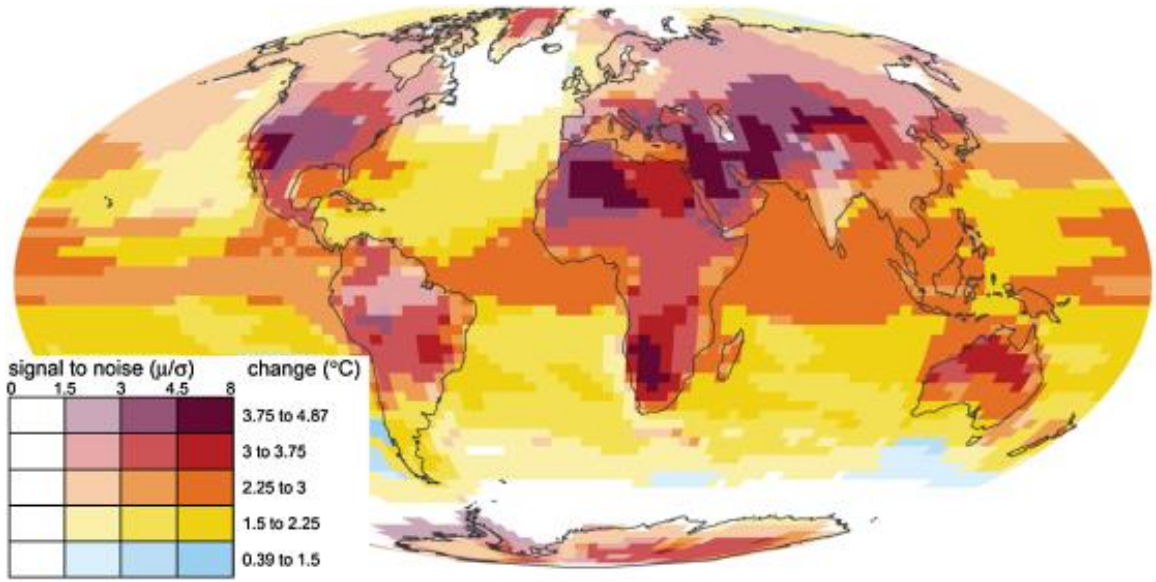
Challenge: dissemination - maps

- Cartography - scale and colors can give a different message



Challenge: dissemination - uncertainty

- How to disseminate uncertainty



International Journal of Climatology
Volume 36, Issue 3, pages 1143-1159, 14 JUL 2015 DOI: 10.1002/joc.4408
http://onlinelibrary.wiley.com/doi/10.1002/joc.4408/full#joc4408-fig-0008

- Who is the audience (stakeholders, general public, academic)?
- Focus groups can help clarify the message / to find the best way to reach the audience. ⇒Environmental & Health atlas ▸ EHA

Acknowledgements

- Areti Boulieri (IC)
- James Bennett (IC)
- Nicky Best(GSK)
- Anna Hansell(IC)
- Sylvia Richardson (MRC-BSU)
- Michela Cameletti (UNIBG)
- Virgilio Gomez-Rubio (Castillia-LaMancha)

References

1. Small area
   - Best, N., Richardson, S., and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. Statistical Methods in Medical Research, 1(14), 35-59.
   - Lee, D. (2011).A comparison of conditional autoregressive models used in Bayesian disease mapping. Spatial and Spatio-Temporal Epidemiology, 2(2), 79-89
   - Lawson A., Banjeree, S., Ugarte, L., Haining, R. (2016) Handbook of spatial epidemiology. CRC.
   - Jackson C, Best N. and Richardson S. (2008) Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. J Royal Statistical Society Series A: Statistics in Society 171(1):159-178
2. Geostatistics
   - Banerjee, S., Carlin, B., and Gelfand, A. (2004). Hierarchical Modeling and Analysis for Spatial Data. Monographs on Statistics and Applied Probability. Chapman & Hall.
   - Diggle, P. and Ribeiro, J. (2007). Model-Based Geostatistics. Springer.
   - Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors (2010). Handbook of Spatial Statistics. Chapman & Hall.
   - Wikle, C. (2003). Hierarchical models in environmental science. International Statistical Review, 71(2), 181-199.