GEOG0125
ADVANCED TOPICS IN SOCIAL AND GEOGRAPHIC DATA SCIENCE

BAYESIAN SPATIAL RISK MODELLING

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

Contents

- Hierarchical models and recap

- Types of spatial risk estimation
  - ❖ Odds ratios (ORs)
  - ❖ Relative risk ratios (RRs)
  - ❖ Exceedance Probabilities

- Spatial intrinsic conditional autoregressive models (iCARs):
  - ❖ Besag-York-Mollie (within an iCAR framework)
  - ❖ Spatial model (with cross-sectional data)

- Model formulation from a Bayesian Framework

- Examples and interpretation (using Stan)

Problem

Collect

Knowledge

Wrangle

Model

Explore
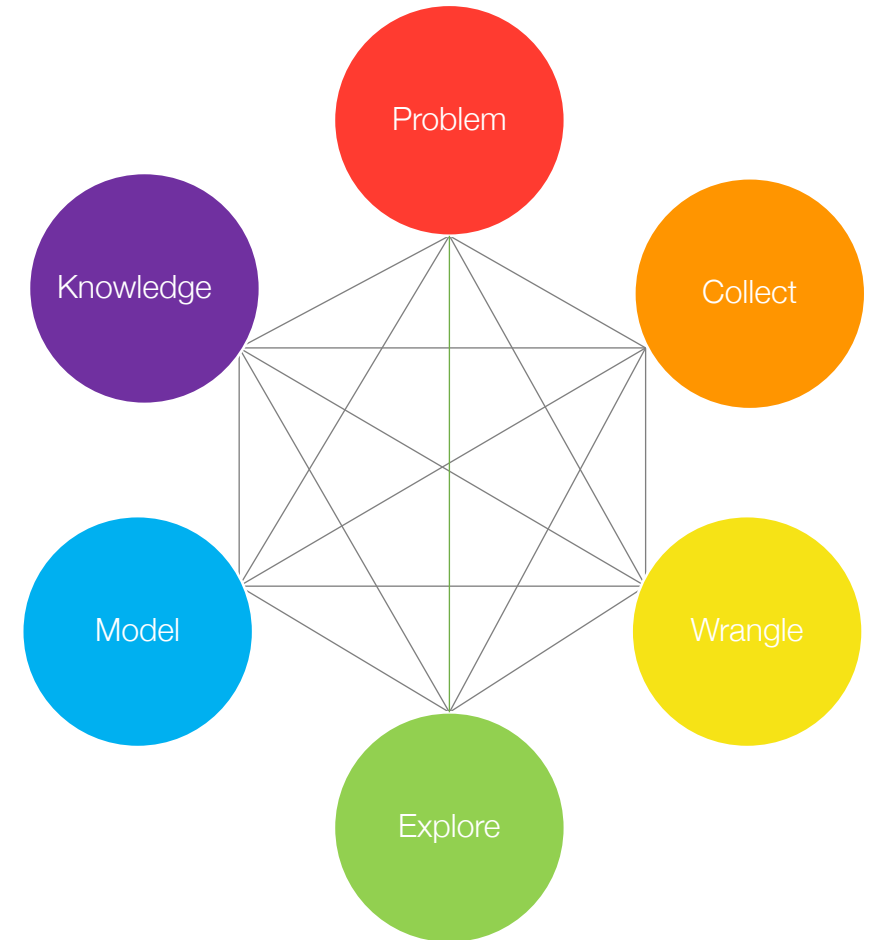
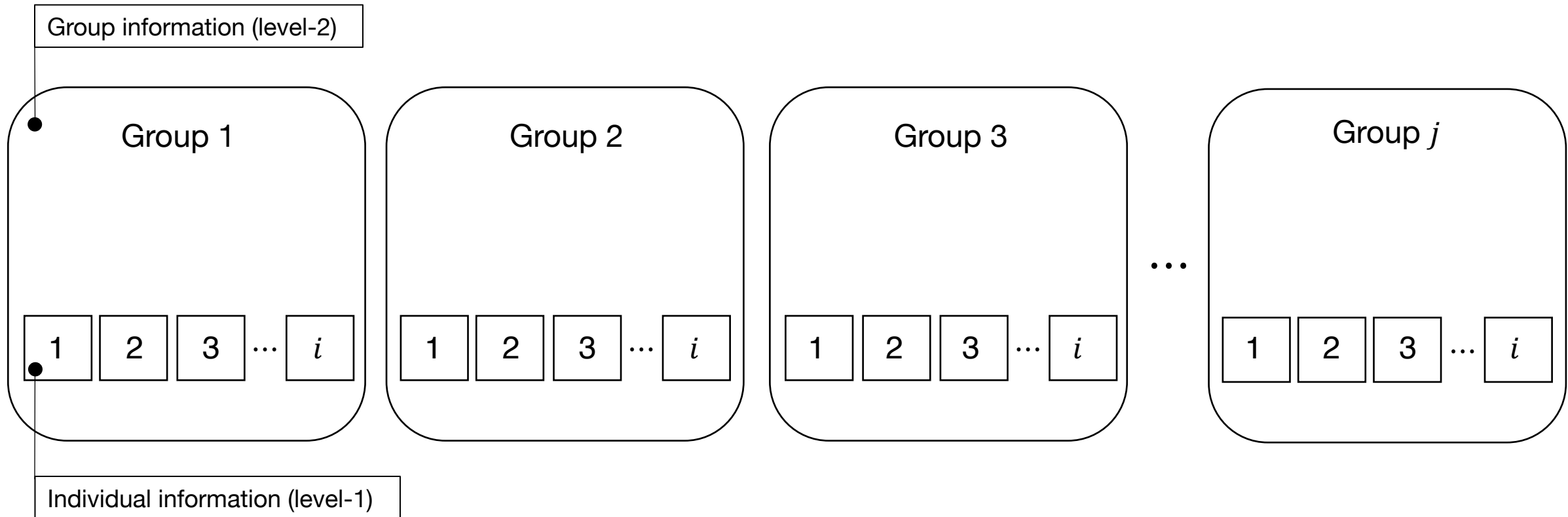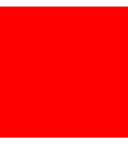# Quick recap on hierarchical regression models

Definition:

A **hierarchical regression model**, are a specialised group of regression-based models that are able to recognise the existence of hierarchies within a data structure and account for them. It is a statistical model used for exploring the relationship between a dependent variable with one or more independent variables while accounting for these hierarchical structures.

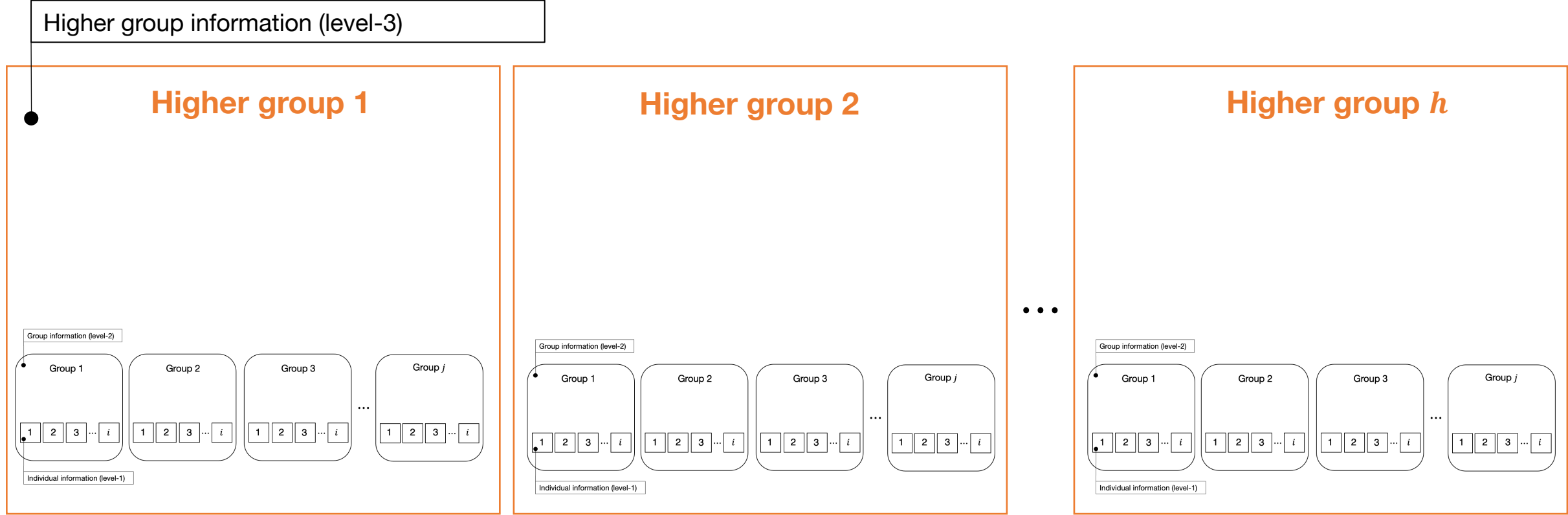## Why are hierarchical regression models important:

- It is an elegant way to model datasets that have varying scales in their measurements (- this artefact is caused by the multilevel or hierarchical structure in the dataset)

- It is an robust approach for accounting for **variations across individual units**, and at the same time, the "**within-group variations**" among groupings

- When we are modelling the direct relationship between the level-1 independent variables against the dependent variable, we can allow for direct interactions between level-1 and higher level independent variables that were measured at a group-level

- We can quantify group-specific differences as well as group-specific coefficients through the usage of "**varying-slopes**" or "**varying-coefficients**"

We are Illustrating concisely what we mean by two- or three-level model structure [1]

Group information (level-2)

Group 1	Group 2	Group 3	Group j
1 2 3 ⋯ i	1 2 3 ⋯ i	1 2 3 ⋯ i	1 2 3 ⋯ i

⋯

Individual information (level-1)

Notes: We have individual units of information that are nested or grouped within a higher measure. This is typically a **two-level structure** and a **two-level hierarchical regression** model must be used for this scenario.

We are Illustrating concisely what we mean by two- or three-level model structure [2]

Higher group information (level-3)

Higher group 1

Group information (level-2)

| Group 1 | Group 2 | Group 3 | ... | Group j |

Individual information (level-1)

Higher group 2

Group information (level-2)

| Group 1 | Group 2 | Group 3 | ... | Group j |

Individual information (level-1)

Higher group h

Group information (level-2)

| Group 1 | Group 2 | Group 3 | ... | Group j |

Individual information (level-1)

Notes: We have individual units of information that are nested or grouped within a higher measure, where by the same individuals (from the same units) are repeated (i.e., longitudinal). This is typically a **three-level structure** and so a **three-level hierarchical regression** model must be used for this scenario.

6

Recall the base model formula for a GLM

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

Mathematical reformulation of the base GLM regression model using indexes

$$y_{i,j} = \beta_{0,j} + \beta_{1,j} x_{1,i,j} + \beta_{2,j} x_{2,i,j} + \cdots + \beta_{k,j} x_{k,i,j} + \varepsilon_{i,j}$$

- When there is a hierarchical structure in the dataset, the base form of the GLM can be explicitly reformulated to show the hierarchies with indexes. For instance

    ❖ We let i represent each individual unit or observation
    ❖ We let j represent a group or cluster which an individual unit or observation i is from.
    ❖ k is the number of independent variables in the dataset

i	j	$y_{i,j}$	x_1	x_2	x_3	$\cdots$	x_k
1	1	$y_{1,1}$	$x_{1,1,1}$	$x_{2,1,1}$	$x_{3,1,1}$	$\cdots$	$x_{k,1,1}$
2	1	$y_{2,1}$	$x_{1,2,1}$	$x_{2,2,1}$	$x_{3,2,1}$	$\cdots$	$x_{k,2,1}$
3	1	$y_{3,1}$	$x_{1,3,1}$	$x_{2,3,1}$	$x_{3,3,1}$	$\cdots$	$x_{k,3,1}$
1	2	$y_{1,2}$	$x_{1,1,2}$	$x_{2,1,2}$	$x_{3,1,2}$	$\cdots$	$x_{k,1,2}$
2	2	$y_{2,2}$	$x_{1,2,2}$	$x_{2,2,2}$	$x_{3,2,2}$	$\cdots$	$x_{k,2,2}$
1	3	$y_{1,3}$	$y_{1,1,3}$	$x_{2,1,3}$	$x_{3,1,3}$	$\cdots$	$x_{k,1,3}$
2	3	$y_{2,3}$	$y_{1,2,3}$	$x_{2,2,3}$	$x_{3,2,3}$	$\cdots$	$x_{k,2,3}$
3	3	$y_{3,3}$	$y_{1,3,3}$	$x_{3,3,3}$	$x_{3,3,3}$	$\cdots$	$x_{k,3,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
i	j	$y_{i,j}$	$x_{1,i,j}$	$x_{2,i,j}$	$x_{3,i,j}$	$\cdots$	$x_{k,i,j}$

For more information about the above model, refer back to week 7's lecture notes and video.

Hierarchical regression model (true form) [1]

$$y_{i,j} = \beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + \cdots + \beta_{k,j}x_{k,i,j} + \varepsilon_{i,j}$$

Level 1 Equation

$$\beta_{0,j} = \gamma_{00} + u_{0,j}$$
$$\beta_{1,j} = \gamma_{10} + u_{1,j}$$
$$\beta_{2,j} = \gamma_{20} + u_{2,j}$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$\beta_{k,j} = \gamma_{k0} + u_{k,j}$$

- 1st equation is a random-intercept

- 2nd, 3rd and 4th and so on equations are random-slopes

- Note that these equation does not have a two-level independent variable that impacts the outcome

Level 2 Equations

- Substitute the level 2 model equations into the level 1 model equation:

$$\Rightarrow y_{i,j} = (\gamma_{00} + u_{0,j}) + (\gamma_{10} + u_{1,j})x_{1,i,j} + (\gamma_{20} + u_{2,j})x_{2,i,j} + \cdots + (\gamma_{k0} + u_{1,j})x_{k,i,j} + \varepsilon_{i,j}$$

- After substitution, we expanding the expression and rearrange as follows:

$$\Rightarrow y_{i,j} = \underbrace{\gamma_{00} + \gamma_{10}x_{1,i,j} + \gamma_{20}x_{2,i,j} + \cdots + \gamma_{k0}x_{k,i,j}}_{\text{Fixed part}} + \underbrace{u_{0,j} + u_{1,j}x_{1,i,j} + u_{2,j}x_{2,i,j} + \cdots + u_{k,j}x_{k,i,j} + \varepsilon_{i,j}}_{\text{Random part}}$$

Model's true form

Note: There are model scenarios

- γ_{00} is the global intercept from the fixed part of the model we want to report

- $\gamma_{10}, \gamma_{20}, \ldots$ and γ_{k0} are the coefficients from the fixed part of the model we want to report now

- $u_{0,j}, u_{1,j}, u_{2,j}, \ldots$ and $u_{k,j}$ as well as $\varepsilon_{i,j}$ they have variances for random part of the model we want to report

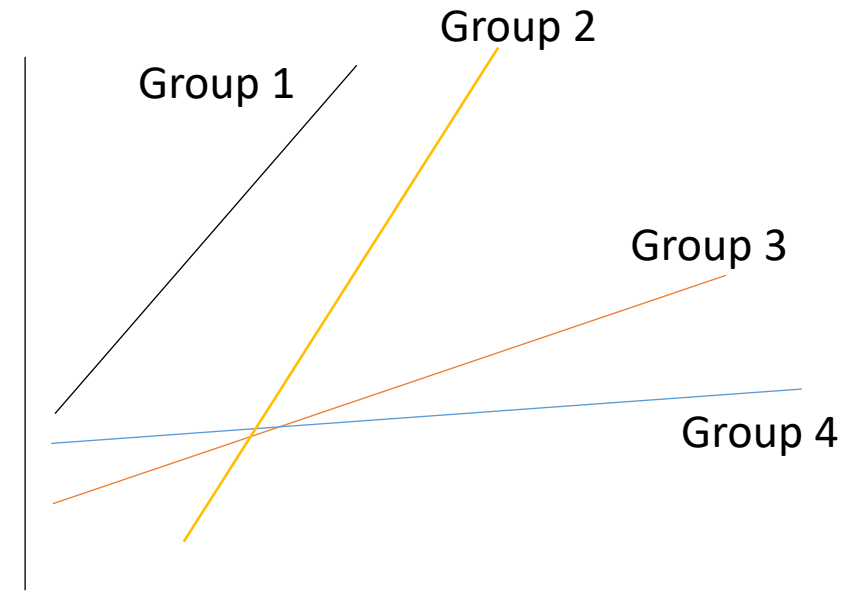# Hierarchical regression (Random-slope) model (true form) [2]

$$y_{i,j} = \beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + \cdots + \beta_{k,j}x_{k,i,j} + \varepsilon_{i,j}$$

Level 1 Equation

$$\beta_{0,j} = \gamma_{00} + u_{0,j}$$
$$\beta_{1,j} = \gamma_{10} + u_{1,j}$$
$$\beta_{2,j} = \gamma_{20} + u_{2,j}$$
$$\vdots \quad \vdots \quad \vdots$$
$$\beta_{k,j} = \gamma_{k0} + u_{k,j}$$

Level 2 Equations

This is an example of a **random-slope model** which includes both a random-intercept and random-slopes. This means there group structures causes variation in the means across groups (i.e., intercepts) and slopes

$$y_{i,j} = \gamma_{00} + \gamma_{10}x_{1,i,j} + \gamma_{20}x_{2,i,j} + \cdots + \gamma_{k0}x_{k,i,j} + u_{0,j} + u_{1,j}x_{1,i,j} + u_{2,j}x_{2,i,j} + \cdots + u_{k,j}x_{k,i,j} + \varepsilon_{i,j}$$

Fixed part

Random part

Model's true form

Group 1

Group 2

Group 3

Group 4

Hierarchical regression (random-intercept-only) model (true form) [3]

$$y_{i,j} = \beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + \cdots + \beta_{k,j}x_{k,i,j} + \varepsilon_{i,j}$$

Level 1 Equation

$$\beta_{0,j} = \gamma_{00} + u_{0,j}$$

Level 2 Equation

Here, the model is much simpler:

$$y_{i,j} = \gamma_{00} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + \cdots + \beta_{k,j}x_{k,i,j} + u_{0,j} + \varepsilon_{i,j}$$

Model's true form

Fixed part

Random part

This is an example of a **random-intercept-only model** which only includes a random-intercept and excludes the random-slopes. This means that the group structure causes variation on the means (i.e., group-specific intercepts) but not on slopes

Group 1

Group 2

Group 3

Group 4

10
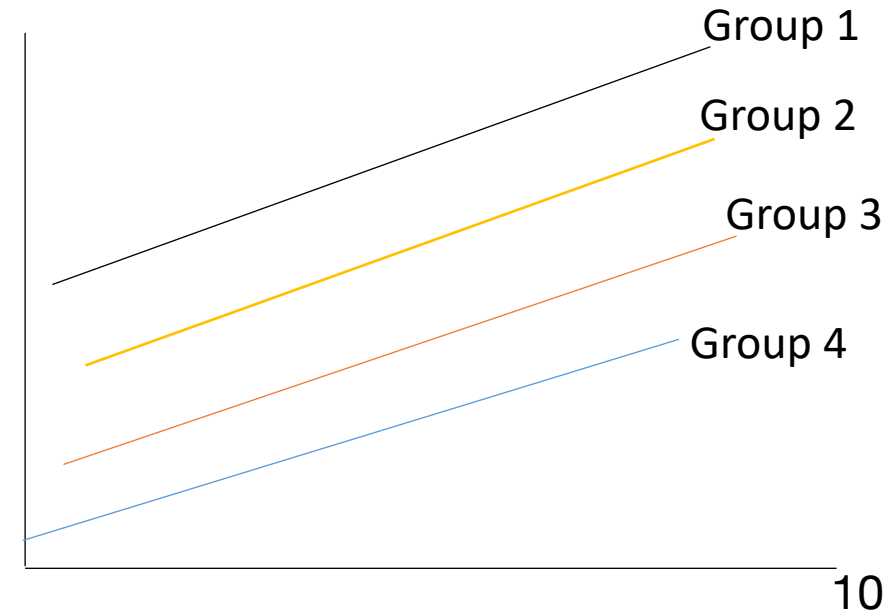
Hierarchical regression (random coefficients) model (true form) [4]

$$y_{i,j} = \beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + \cdots + \beta_{k,j}x_{k,i,j} + \varepsilon_{i,j}$$

Level 1 Equation

$$\beta_{0,j} = \gamma_{00} + \gamma_{01}Z_1 + u_{0,j}$$
$$\beta_{1,j} = \gamma_{10} + \gamma_{11}Z_1 + u_{1,j}$$
$$\beta_{2,j} = \gamma_{20} + \gamma_{21}Z_1 + u_{2,j}$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$\beta_{k,j} = \gamma_{k0} + \gamma_{k1}Z_1 + u_{k,j}$$

Level 2 Equations

Suppose we have an independent variable measure on the group-level impacting our outcome on the individual-level.

- Substitute the level 2 model equations with the variables into the level 1 model equation:

$$\Rightarrow y_{i,j} = (\gamma_{00} + \gamma_{01}Z_I + u_{0,j}) + (\gamma_{10} + \gamma_{11}Z_I + u_{1,j})x_{1,i,j} + (\gamma_{20} + \gamma_{21}Z_I + u_{2,j})x_{2,i,j} + \cdots + (\gamma_{k0} + \gamma_{k1}Z_I + u_{1,j})x_{k,i,j} + \varepsilon_{i,j}$$

- After substitution, we expanding the expression and rearrange as follows:

$$\Rightarrow y_{i,j} = \gamma_{00} + \gamma_{01}Z_1 +$$
$$\gamma_{10}x_{1,i,j} + \gamma_{20}x_{2,i,j} + \cdots + \gamma_{k0}x_{k,i,j} +$$
$$\gamma_{11}Z_1 x_{1,i,j} + \gamma_{20}Z_1 x_{2,i,j} + \cdots + \gamma_{k0}Z_1 x_{k,i,j} +$$
$$u_{0,j} + u_{1,j}x_{1,i,j} + u_{2,j}x_{2,i,j} + \cdots + u_{k,j}x_{k,i,j} + \varepsilon_{i,j}$$

Model's true form

γ_{00} is the global or population mean

γ_{01} is the random coefficient for Z_1

These are fixed effects coefficients for the variables in the level 1 equation

These are random coefficients for the interacting variables from the level 1 & 2 equation

These are the random effects

11

# Situating hierarchical models within a spatial context:

Bayesian inference, are often used in hierarchical modelling, which are models commonly used in the quantification of spatial and spatiotemporal areal data.

- Bayesian approach are incredibly good with datasets that have a hierarchical structure.

- These are statistical model written in multiple levels (i.e., hierarchical form) to estimate parameters of the posterior distribution

- Example: Intestinal parasitaemia among school children in Tanzania and infection status linked with anaemia
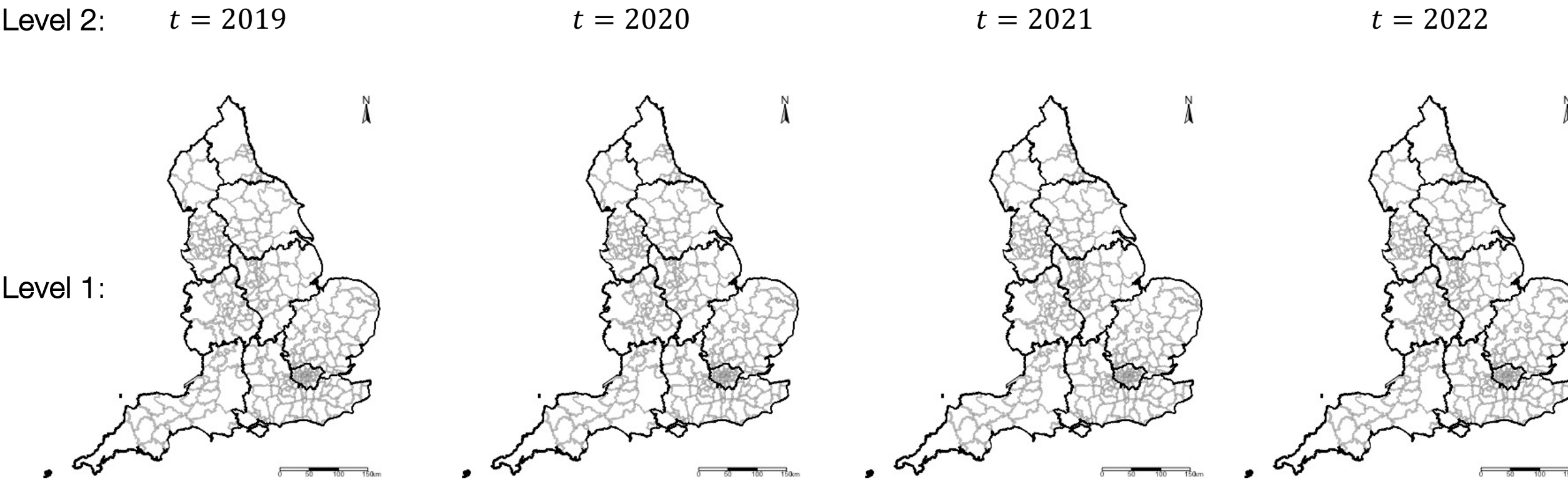
Health student = ⬤   Diseased student = 🔴

School 1   School 2   School 3   School n
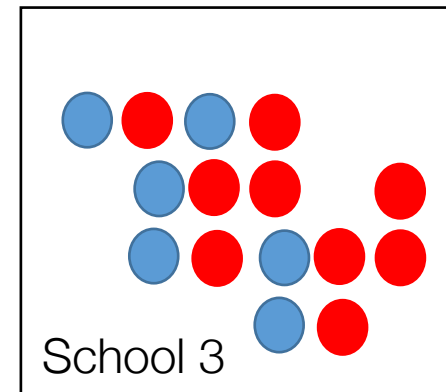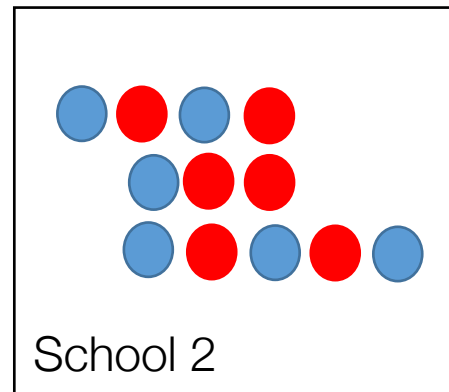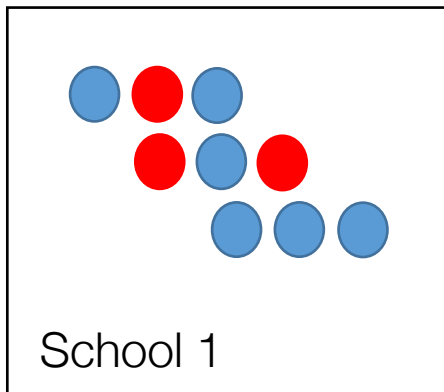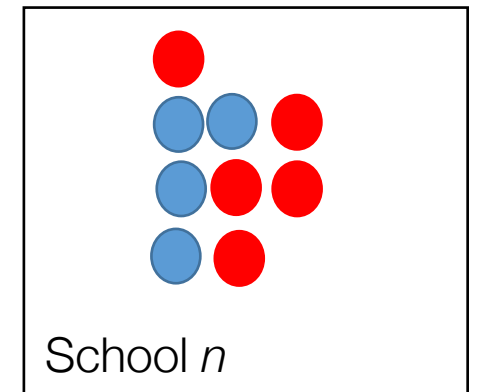
Point locations: (x_1, y_1) (x_2, y_2) (x_3, y_3) (x_n, y_n)

Level 2: $t = 2019$ $t = 2020$ $t = 2021$ $t = 2022$

Level 1:

- These models allow complete flexibility in the estimation of risks - allowing the user to account for space-time interactions

- You can make the model (in contrast to frequentist) to borrow strength across space-time, in order to improve estimation and prediction of an underlying model's feature

13

Type of spatial risk estimation

Areal data

Areal, or lattice data arise when dealing with a fixed domain that is partitioned to a finite number of sub-regions at which outcome can be aggregated too

- Examples of areal data are:
  - Number of cancer cases in counties
  - Number of road accidents in districts
  - Proportion of people living in poverty in postcode block etc.

Often, risk models aim to obtain such estimates within such areas where data is available. We can use Bayesian Hierarchical Models in this context, depending on the type of study design, to estimate the following: Odds Ratios (ORs) or Relative Risk (RRs)

15

Interpretation of Risk Ratios (RR)

RR= 1 (null value), it means that independent variable has no effect on the outcome

RR < 1, the independent variable has an impact on the outcome – in this case, its reduced effect, or reduced risk on the outcome

RR > 1, the independent variable has an impact on the outcome – and so, in this case, its increased effect, or increased risk on the outcome

From hazards models:
- Cox Proportional Hazards model
- Any Poisson model

Interpretation of Odds Ratios (OR)

OR = 1 (null value), it means that independent variable has no effect on the outcome

OR < 1, the independent variable has an impact on the outcome – in this case, its reduced effect, or reduced risk on the outcome

OR > 1, the independent variable has an impact on the outcome – and so, in this case, its increased effect, or increased risk on the outcome

From models:
-   Binary or Binomial regression model

Exceedance Probability

Exceedance Probabilities (or Marginals) is a statistical measure describing the probability that an estimated risk value for an areal-unit exceeds a given threshold.

A common example used in every day application are disease risk models, we are usually concerned about areas that have excess risk of a disease type i.e., P(RR > 1)

In epidemiology, the Exceedance Probabilities have been operationalised to detect clusters of areas with exceedingly higher risk of a disease (or adverse event).

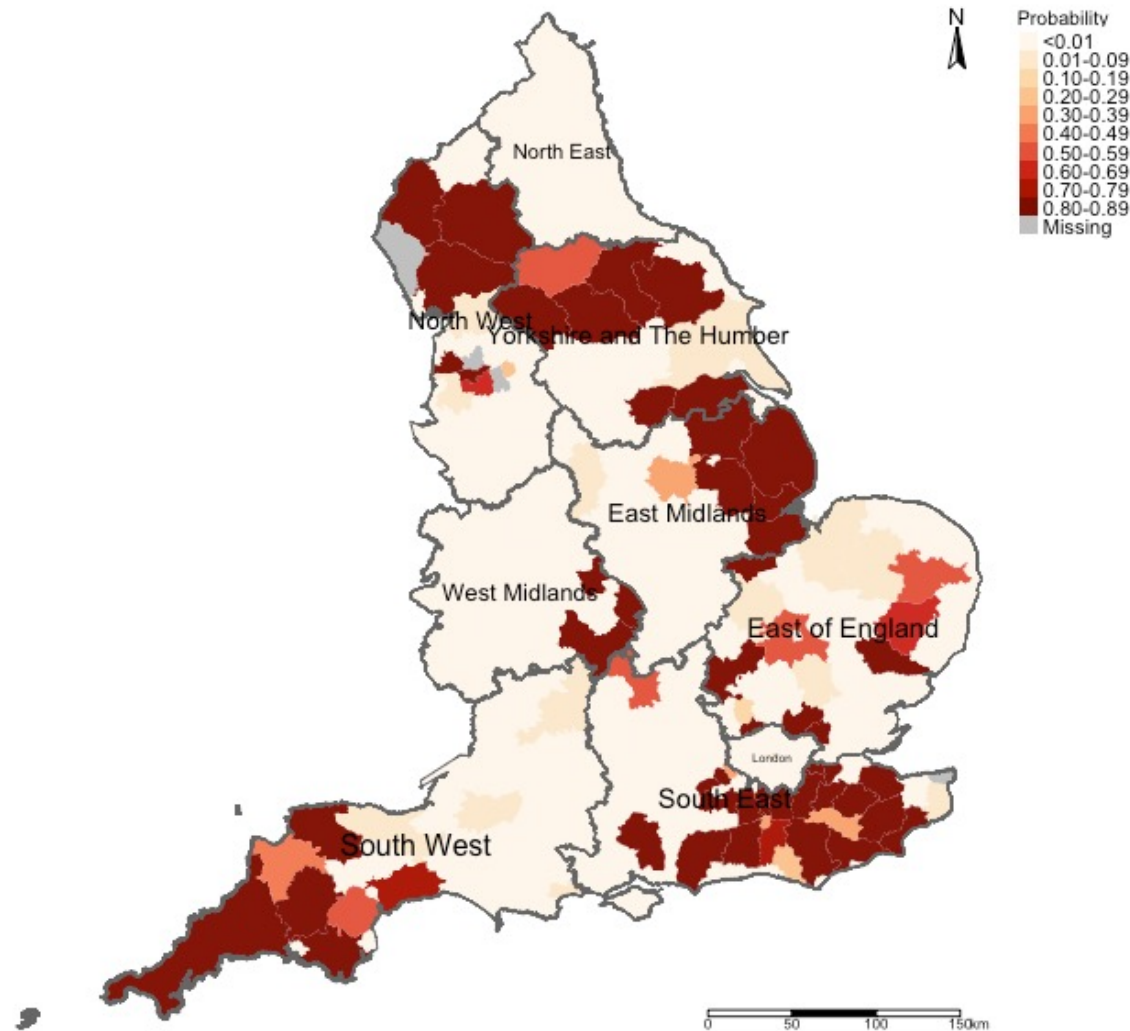# Example: Risks of Road-related casualties in England 2015-2020 [1]
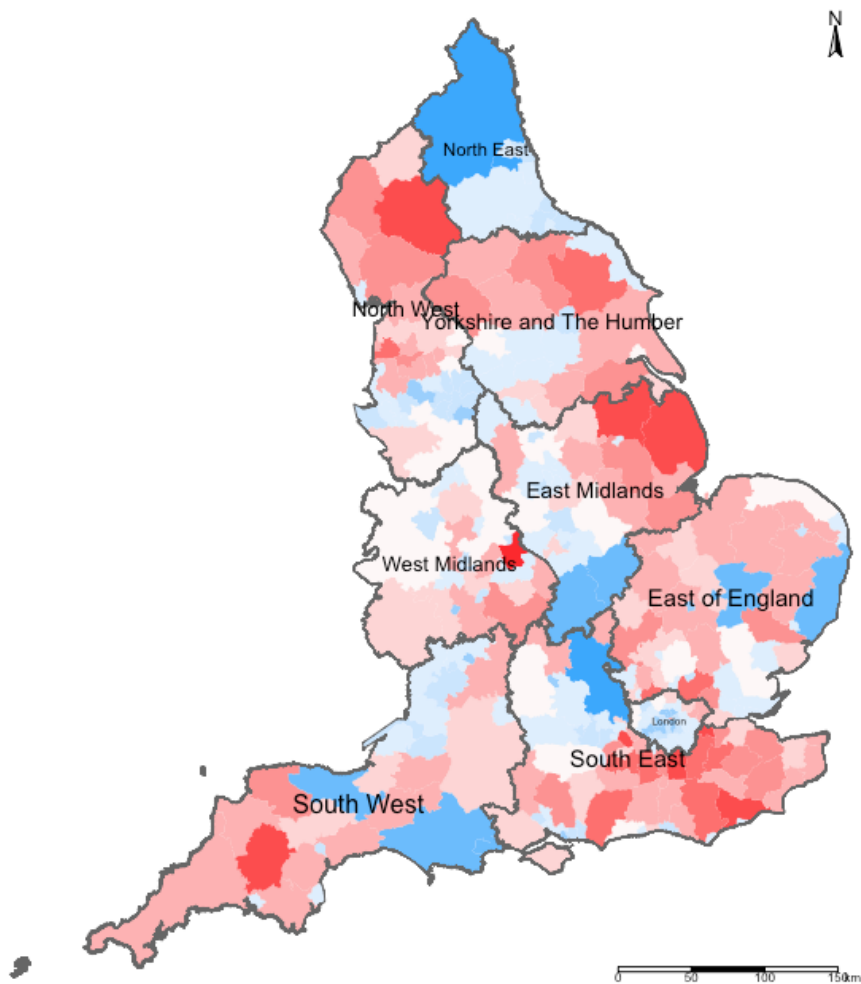
Relative Risks (RR)
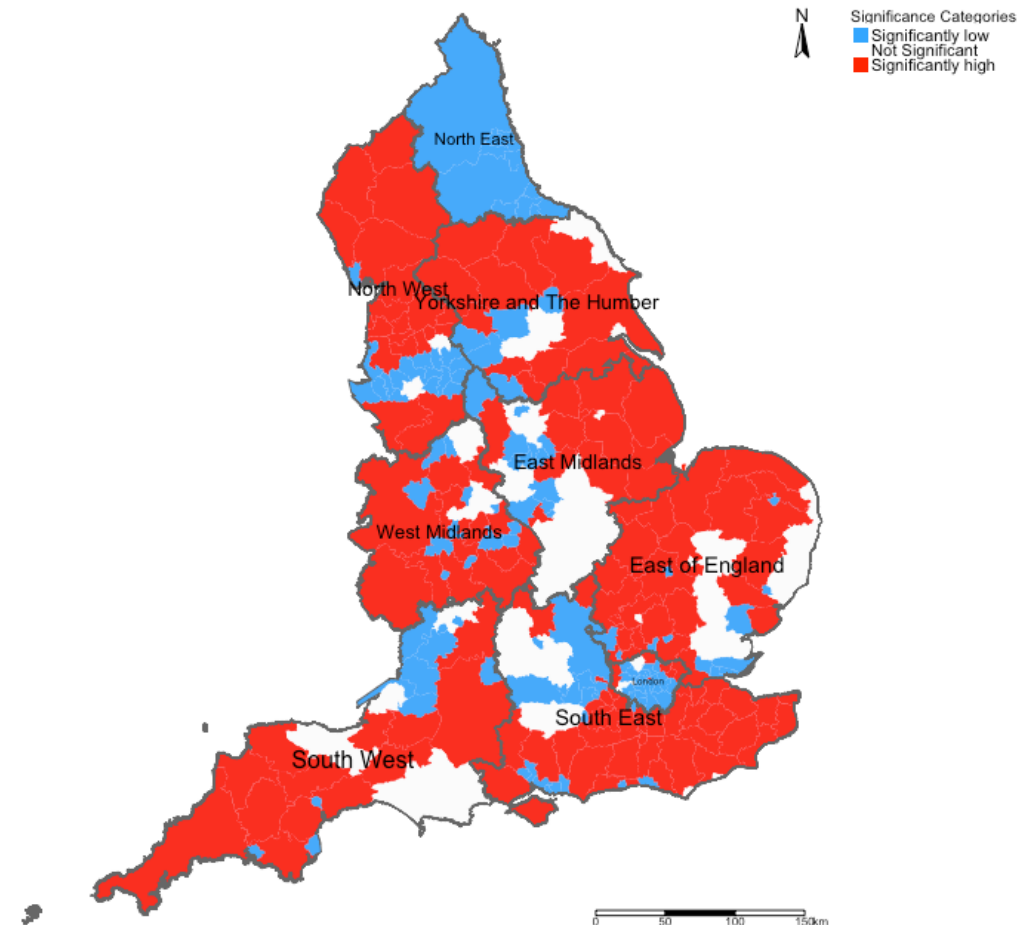
Overall significance (95% Credibility Intervals)

Example: Risks of Road-related casualties in England 2015-2020 [2]

The areas in darker reds are perhaps priority areas for some road safety policy should be implemented?

Exceedance Probability i.e., Pr(RR > 1.40) (i.e., risk are 40% higher than expected)

Spatial Intrinsic Conditional Autoregressive models (iCARs)

Besag-York-Mollie (BYM) (or CAR models)

This is a popular spatial model which takes into account that the data may potentially be spatially correlated and the observations in the neighbouring areas may be more similar than observations in areas that are distant from each other.

- This is a type of hierarchical model which includes a spatial random effect,

- It is heavily dependent on the neighbourhood adjacency matrix

- There are two versions of this model:
  - ❖ BYM model that has a spatial effect term only that's treated a smoothing term (multiplied by an error term)
  - ❖ BYM model that has both a spatial effect term which is treated a structured random effect, and the error term is an unstructured noise

- When fitting data to this type of model – the best choice of the likelihood function (i.e., statistical model) is Poisson (i.e., aggregated counts to areas).

22

Geographically accurate
neighbourhood structure
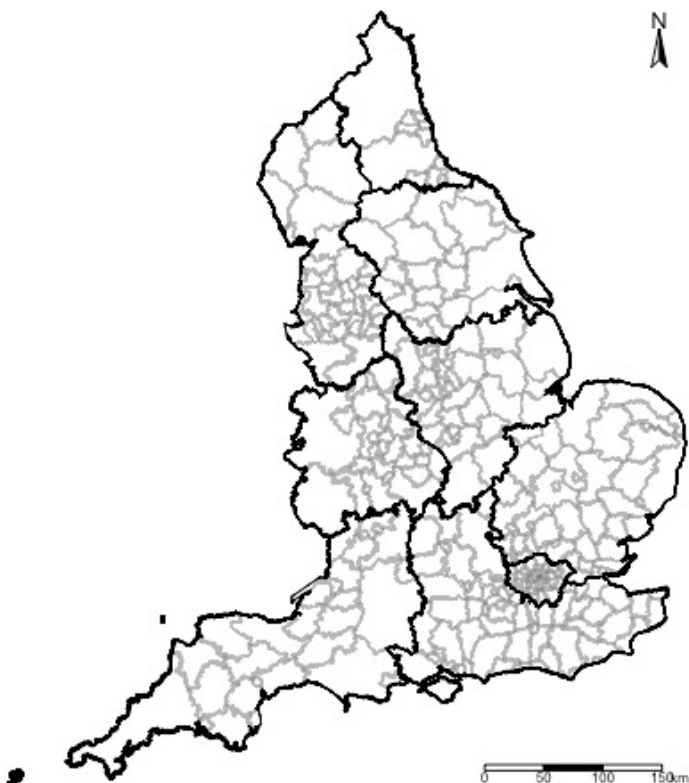
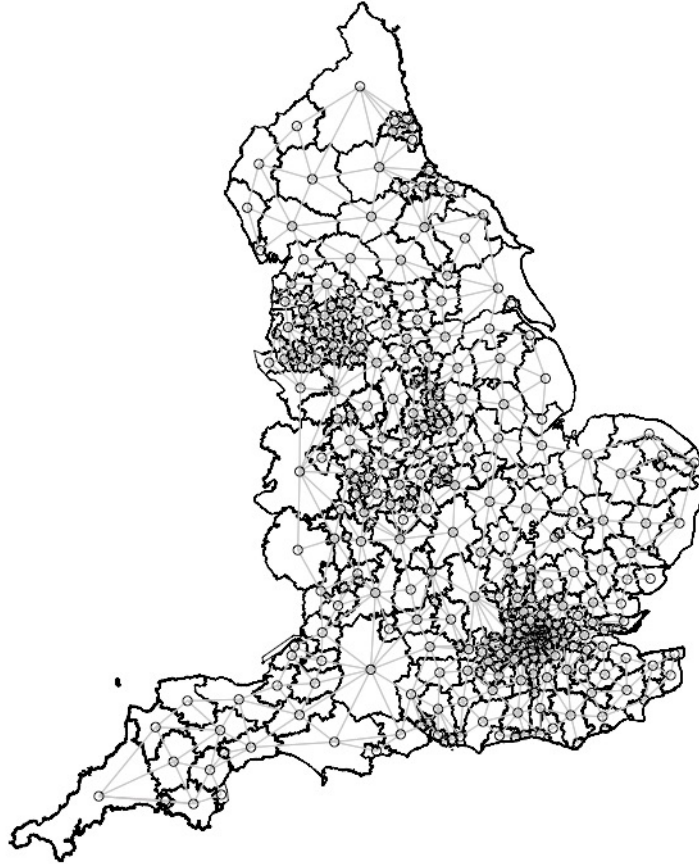Adjacency matrix translated to
graph format

Adjacency matrix translated to
nodes and edge format

Dimensions: 309 x 309
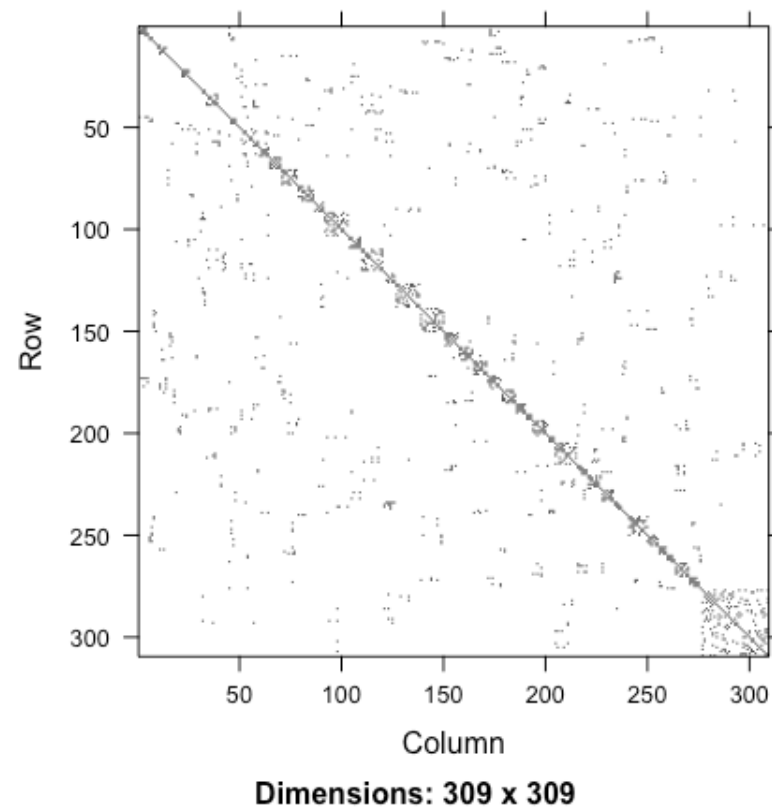
Stan only uses the nodes and edges
format to reconstruct the adjacency
matrix

# Spatial structured and unstructured random effects [1]

Adjacency matrix translated to nodes and edge format

- **Structured spatial random effects ϕ** in an ICAR model refers to the influence or impact that neighbouring locations have on each other.

- It means that the values or characteristics of one location are related to the values of its neighbouring locations

- Here, we are accounting for the spatial dependence i.e., neighbouring areas (or those closest to each other) are related than distant areas.

- Examples: Clusters of disease spread, urban development or from a climate point of view – temperature gradient or rainfall etc.,

- **Unstructured spatial random effects θ** in an ICAR model refers to the unique characteristics or behaviours of the individual locations that are not influenced by their neighbouring locations.

- It means that the values or characteristics of one location are unrelated to the values of its neighbouring locations

- Hence, there's may be no spatial dependence.

- Examples: Cultural boundaries or practices, language, unique landmarks, or a particular maybe housing style of patterns

In an ICAR model, we can account for both these types of random effects by COMBINING them as $\phi + \theta$

24

Adding the spatial structured and unstructured effect to the model [2]
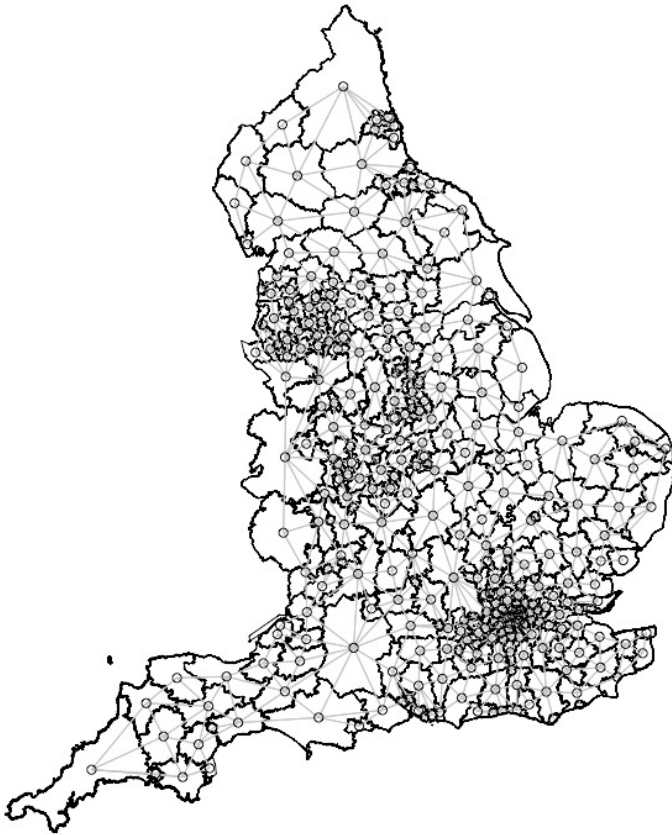
Adjacency matrix translated to nodes and edge format

Node1 is the index area of interest;
Node2 is the neighbouring areas connected to the index area defined in Node1.
N is the total number of areas.

```
target += -0.5*dot_self(phi[node1] – phi[node2])
sum(phi) ~ normal(0, 0.001*N)
```

- $C_i = \theta_i + \phi_i$ is the combined random effects which is equivalent:

$$\boldsymbol{\phi} + \boldsymbol{\theta} = \boldsymbol{\sigma}(\sqrt{(1-\rho)\boldsymbol{\theta}} + \sqrt{(\rho/s)\boldsymbol{\phi}})$$

- Where we use ρ as a proportion to represent the amount of variance that comes from the spatial random effect ϕ. While $1-\rho$ is the proportion of the variation that's unstructured θ.

- σ is the overall error

- It is proposed to use a scaling factor s on the variance for the spatial random effects. This is computed from the geometric mean of ϕ_i which are on the diagonals adjacency matrix when inverted.

- The prior distribution we use for ρ is beta(0.5, 0.5) (red-line in graph)
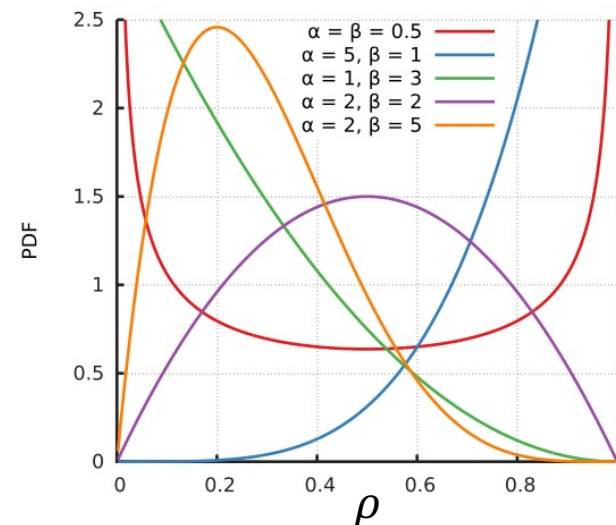
Model formulation for Spatial ICAR model

Model components

Variables
Y_i are counts of observed cases of road accidents across LAs (outcome)
$X_{i,k}$ independent variables (single variable for IMD at LA-level)
E_i are expected counts of road accident cases (derived from $Y_i R$)
R is the overall rate for the entire study location (not for each area)
r_i is some area-specific rates (this is specified in Poisson statement)

Parameters
α is the overall risk of road accidents in the entire study area (intercept)
β_k measures the overall associated risk between $X_{i,k}$ and Y_i
ϕ_i are the area-specific spatial random effects
θ_i are the area-specific unstructured random effects
σ an overall error term

Model Calibration
- ρ is the proportion that's set by the user to state the how much variance comes from either ϕ_i or θ_i
- $C_i = \theta_i + \phi_i$ is the combined random effects which is equivalent to $\sigma(\sqrt{(1-\rho)\theta} + \sqrt{((\rho/s)\phi)}$

Notes:
- $\exp(\alpha)$ is the overall risk ratio for study area
- $\exp(\beta)$ is the overall risk ratio for coefficient
- $\exp(\alpha + \sum \beta_k X_{i,k} + C_i\sigma)$ by adding $+C_i\sigma$ to the α allows the risks to vary for each area. By adding $+ \sum \beta_k X_{i,k}$ you are also adjusting the estimated risk for the variables.

Full model specification

- Specify likelihood function. The outcome is often counts – thus it will be Poisson (with log as the link function).

$$Y_i \sim \text{Poisson}(E_i r_i)$$

- $\log(\lambda_i) = \alpha + \sum \beta_k X_{i,k} + C_i\sigma + \log(E_i)$
- where $C_i = \theta_i + \phi_i = \sigma(\sqrt{(1-\rho)\theta} + \sqrt{((\rho/s)\phi)}$

- Define the priors for the intercept, coefficients and spatial and unstructured random effects as with an ICAR specification

$\alpha \sim \text{norm}(0, 1)$
$\beta \sim \text{norm}(0, 1)$
$\sigma \sim \text{norm}(0, 1)$ (alternatives are $\text{gamma}(0.001, 0.001)$)
$\rho \sim \text{beta}(0.5, 0.5)$
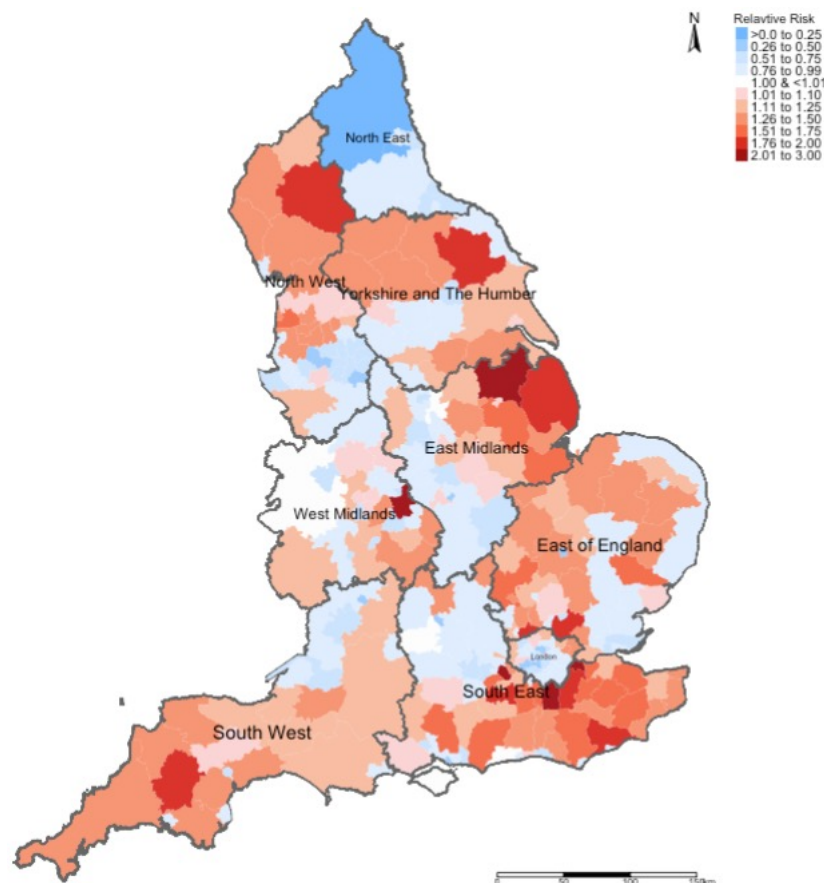$\text{target} += -0.5*\text{dot\_self}(\text{phi[node1]} - \text{phi[node2]})$ (calculates weights)
$\text{sum(phi)} \sim \text{normal}(0, 0.001*N)$

- Build Bayesian model

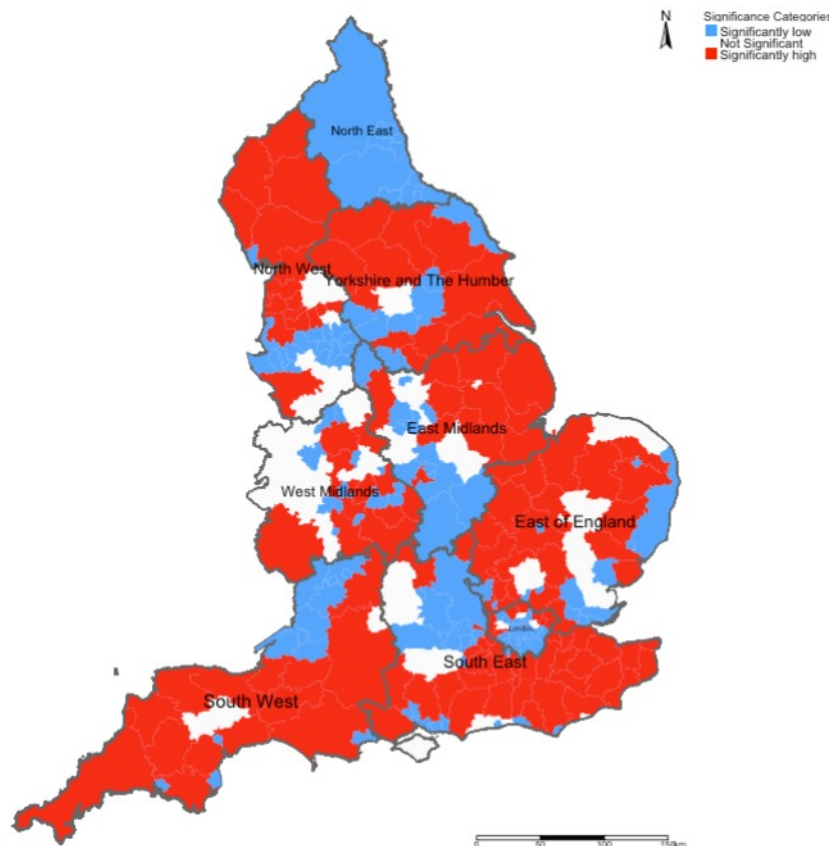Recall the Bayes' Rule: $P(\theta|Y) \propto P(Y|\theta)P(\theta)$

$$P(\alpha, \beta_k, \sigma, \varphi_i | \lambda_i) \propto P(\lambda_i | \alpha, \beta_k, \sigma, \varphi_i) P(\alpha)P(\beta_k)P(\sigma)P(\phi_i)P(\rho)$$
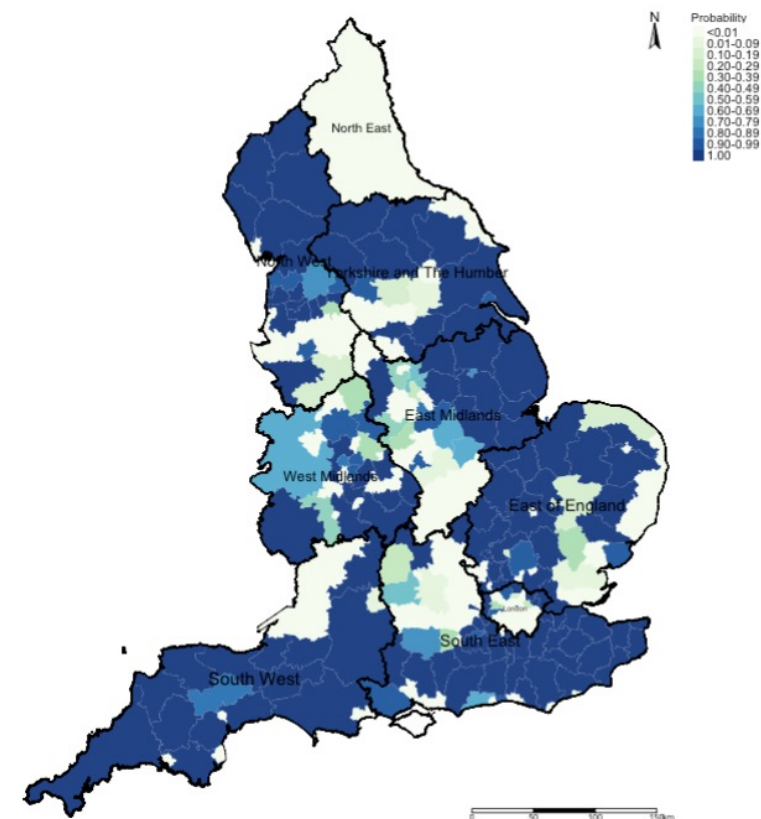
Relative risk ratios (RR)

Here, we use this output to describe the burden of an outcome

Statistical Significance

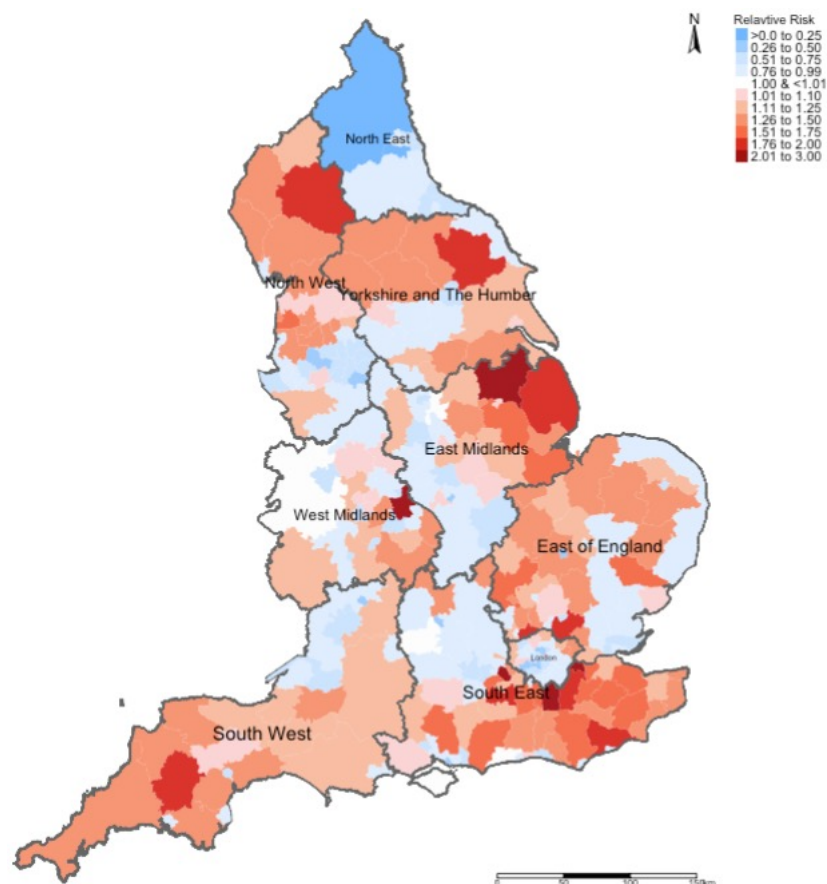This output is to valid whatever hypothesis we had about the described outcome's burden in the first map

Exceedance Probabilities

This output is used to describe the uncertainty that surrounds the risks we found in the first map when we explore P(RR>1)

27

Relative risk ratios (RR)

Statistical Significance
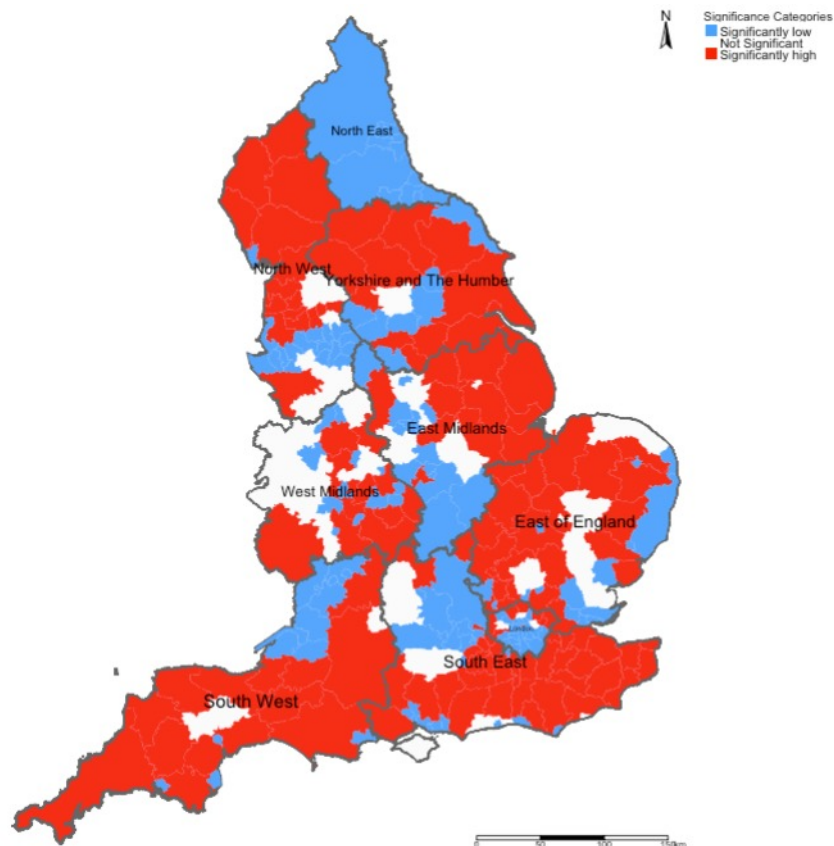
Exceedance Probabilities
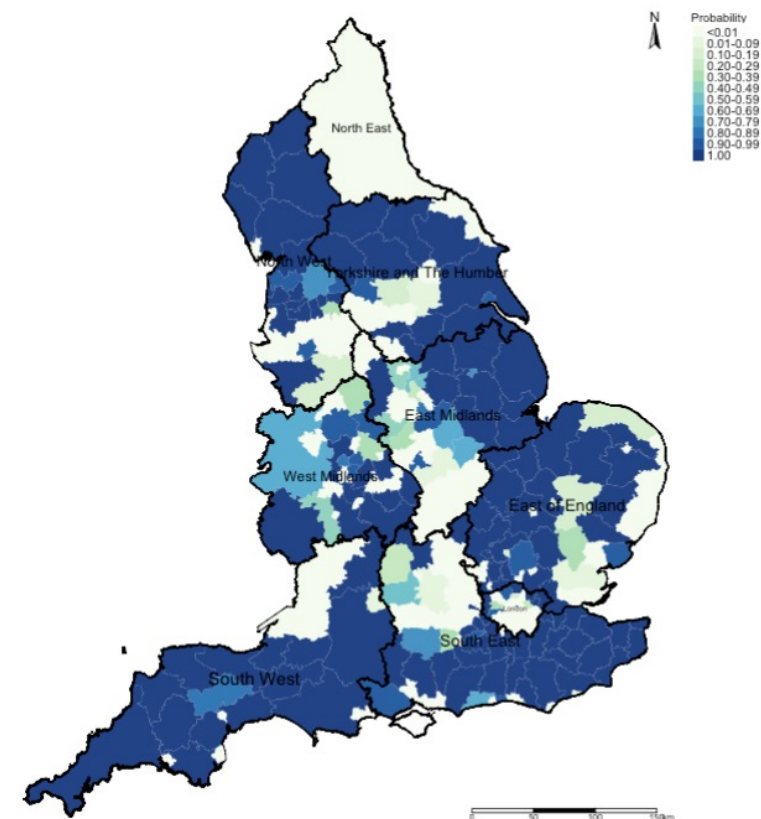
Interpretation: We can see that the risk patterns for road accidents across England are quite heterogeneous. While it is quite pronounced in all 10 regions in England, the burden is quite significant in South West region with large numbers of local authorities having an increased risk which are statistically significant. Perhaps, the Department for Transport should do an investigation on these patterns starting with the South West area.

Application: Risk assessment and mapping of infestation in Campina Grande

Legends
Residential Block Areas (4,708)
Neighbourhood Boundaries 'Bairros' (48)

- Campina Grande is into 47 neighbourhoods

- Most recent vector control data (Levantamento Rapid de Indice para Aedes a [LIRAa]): January 2013 to October 2017 (performed 3–5 times in a year)

- Baseline information – the overall number of houses in neighbourhood (as denominators); total number of households detected to be infested with larvae or adult mosquito (i.e., *Aedes aegypti*)

Aims and objectives:

- To quantify the risk trajectories of mosquito infestation on a neighbourhood-level to informs the profile of the neighbourhood (i.e., whether the risks were 'high' or 'low' in that observed LIRAa period i.e., April 2017).

- Determining the set of environmental, climate and anthropogenic risk factors that impact neighbourhood-levels of *Aedes aegypti* infestation in households.

30

Research Methodology & Study design

LIRAa	Survey Periods				
	2013	**2014**	**2015**	**2016**	**2017**
1	January	January	January	April	January
2	March	March	March	July	April
3	May	May	May	October	July
4	July	July	October		
5	October	October			

Using this snapshot time-period for April to demonstrate the application

Methodology:

• Population-based ecological study design within cross-sectional (and retrospective) framework

• For covariates, the analysis included:
1) **WorldClim (4.5km)** (**Maximum temperature** and **Precipitation**) (monthly)
2) **MOD18A1.061 Terra Vegetation Indices 16-Day Global 500m** to compute neighbourhood levels of vegetation based on the **NDVI** metrics (monthly)
3) **Worldpop.org (100m)** to extract rasters for urbanisation (which contains binary grids) to compute the fraction of surface that is urbanised for neighbourhoods (yearly).

• **Spatial risk model with Intrinsic Conditional Autoregressive (ICAR) Model**; and to derive new global coefficients for covariates for at that survey period for April 2017, as well as neighbourhood-specific relative risk estimates.

Model formulation for Spatial ICAR model [1]

Model components

Variables

Y_i are counts of infected houses in neighbourhoods (outcome)
$X_{i,k}$ independent variables ($k = 4$)
E_i are expected counts of cases infected houses an area
R is the overall rate of infestation in the study area in LIRAa period
r_i is some area-specific rates within that LIRAa period

Parameters

α is the overall risk of infestation for entire study area
β_k measures the overall associated risk between $X_{i,k}$ and Y_i
ϕ_i are the area-specific spatial random effects
θ_i are the area-specific unstructured random effects
σ an overall error term

Model Calibration

- ρ is the proportion that's set by the user to state the how much variance comes from either ϕ_i or θ_i
- $C_i = \theta_i + \phi_i$ is the combined random effects which is equivalent to $\sigma(\sqrt{(1-\rho)\theta} + \sqrt{\rho\phi})$

Notes:
- $\exp(\alpha)$ is the overall risk ratio for study area
- $\exp(\beta)$ is the overall risk ratio for coefficient
- $\exp(\alpha + \sum \beta_k X_{i,k} + C_i\sigma)$ by adding $+C_i\sigma$ to the α allows the risks to vary for each area. By adding $+ \sum \beta_k X_{i,k}$ you are also adjusting for the variables.

Full model specification

- Specify likelihood function. The outcome is often counts – thus it will be Poisson (with log as the link function).

$$Y_i \sim \text{Poisson}(E_i r_i)$$

- $\log(\lambda_i) = \alpha + \sum \beta_k X_{i,k} + C_i\sigma + \log(E_i)$
- where $C_i = \theta_i + \phi_i = \sigma(\sqrt{(1-\rho)\theta} + \sqrt{\rho\phi})$

- Define the priors for the intercept, coefficients and spatial random effects as with an ICAR specification

$$\alpha \sim \text{norm}(0, 1)$$
$$\beta \sim \text{norm}(0, 1)$$
$$\sigma \sim \text{norm}(0, 1)$$
$$\rho \sim \text{beta}(0.5, 0.5)$$

target += -0.5*dot_self(phi[node1] – phi[node2])
sum(phi) $\sim$ normal(0, 0.001*N)

- Build Bayesian model

Recall the Bayes' Rule: $P(\theta|Y) \propto P(Y|\theta)P(\theta)$

$$P(\alpha, \beta_k, \sigma, \varphi_i | \lambda_i) \propto P(\lambda_i|\alpha, \beta_k, \sigma, \varphi_i) P(\alpha)P(\beta_k)P(\sigma)P(\phi_i)P(\rho)$$

Model formulation for Spatial ICAR model [2]

Stan code

```
data {
  int<lower=0> N;
  int<lower=0> N_edges;
  array[N_edges] int<lower=1, upper=N> node1;
  array[N_edges] int<lower=1, upper=N> node2;
  array[N] int<lower=0> Y;
  vector<lower=1>[N] Offset;
  int<lower=1> K;
  matrix[N, K] X;
  real<lower=0> factor;
}

transformed data {
  vector[N] log_Offset = log(Offset);
}

parameters {
  real alpha;                        // intercept
  vector[K] beta;                     // covariates
  real<lower=0> sigma;               // overall standard deviation
  real<lower=0, upper=1> rho;        // proportion unstructured vs. spatially structured variance
  vector[N] theta;                   // unstructured random effects (heterogeneous)
  vector[N] phi;                     // spatial random effects
}
```

```
transformed parameters {
  vector[N] combined;
  combined = sqrt(1 - rho) * theta + sqrt(rho/factor) * phi;
}

model {
  Y ~ poisson_log(log_Offset + alpha + X * beta + combined * sigma);

  alpha ~ normal(0.0, 1.0);                      // prior for alpha: weakly informative
  beta ~ normal(0.0, 1.0);                       // prior for betas: weakly informative
  theta ~ normal(0.0, 1.0);                      // prior for theta: weakly informative
  sigma ~ normal(0.0, 1.0);                      // prior for sigma: weakly informative
  rho ~ beta(0.5, 0.5);                          // prior for rho: pulled for literature

  target += -0.5 * dot_self(phi[node1] - phi[node2]);
  sum(phi) ~ normal(0, 0.001 * N);
}

generated quantities {
  vector[N] eta = alpha + X * beta + combined * sigma;  // compute eta and exponentiate into mu
  vector[N] rr_mu = exp(eta);                    // output the neighbourhood-specific relative risks in mu
  vector[K] rr_beta = exp(beta);                 // output the risk ratios for each coefficient
  real rr_alpha = exp(alpha);                    // output the risk ratios for the intercept
}
```

PART 1: Table results that illustrates the GLOBAL association between environmental, climate and anthropogenic factors and risk of infestation in Campina Grande (in LIRAa 2 survey period for April 2017).

| 2017	LIRAa 2	
	RR (95% CrI)	Pr(RR>1)
Intercept	1.64 (95% CrI: 0.14 to 7.07)	0.51
Temperature	0.93 (95% CrI: 0.74 to 1.12)	0.23
Precipitation	1.01 (95% CrI: 0.96 to 1.07)	0.73
NDVI	1.09 (95% CrI: 0.71 to 1.60)	0.63
Urbanisation	1.18 (95% CrI: 0.37 to 2.90)	0.52

RR: Relative risks; Pr(RR > 1): Exceedance probabilities (the probability that RR being greater than 1)
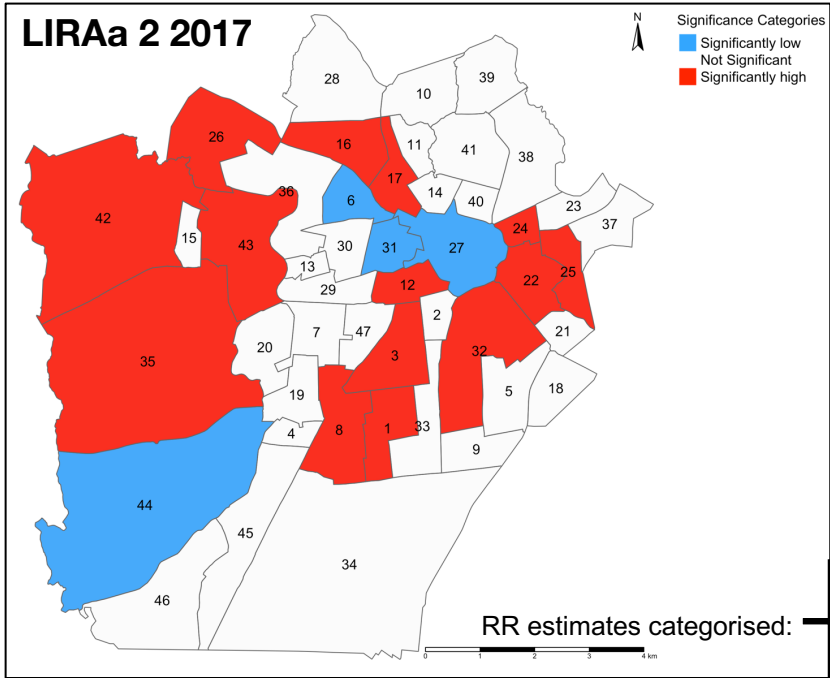
Interpretation (examples):
- **Intercept**: The overall risk of mosquito-borne infestation is 1.64 times (or 64%) **higher** in Campina Grande. The overall probability that there's excess risk of infestation (i.e., RR > 1.00) is 51%.
- **Temperature**: In relation to temperature, the risk of mosquito-borne infestation is 0.93 times (or 7%) **lower** in Campina Grande. The probability of observing an excess risk of infestation (i.e., RR > 1.00) in relation to temperature is 23%.
- **Urbanisation index**: In relation to urbanisation, the risk of mosquito-borne infestation is 1.18 times (or 18%) **higher** in Campina Grande. The probability of observing an excess risk of infestation (i.e., RR > 1.00) in relation to urbanisation is 52%.

NOTE: All relative risk estimates have the null value (1) between its lower and upper 95% credibility intervals. While the results, excluding temperature, show an increased risk of infestation – **these are all statistically not significant**.

PART 2: Maps on the left panel illustrates the relative risk (RR) of infestation across neighbourhoods in Campina Grande

LIRAa 2 2017

Continuous RR estimates
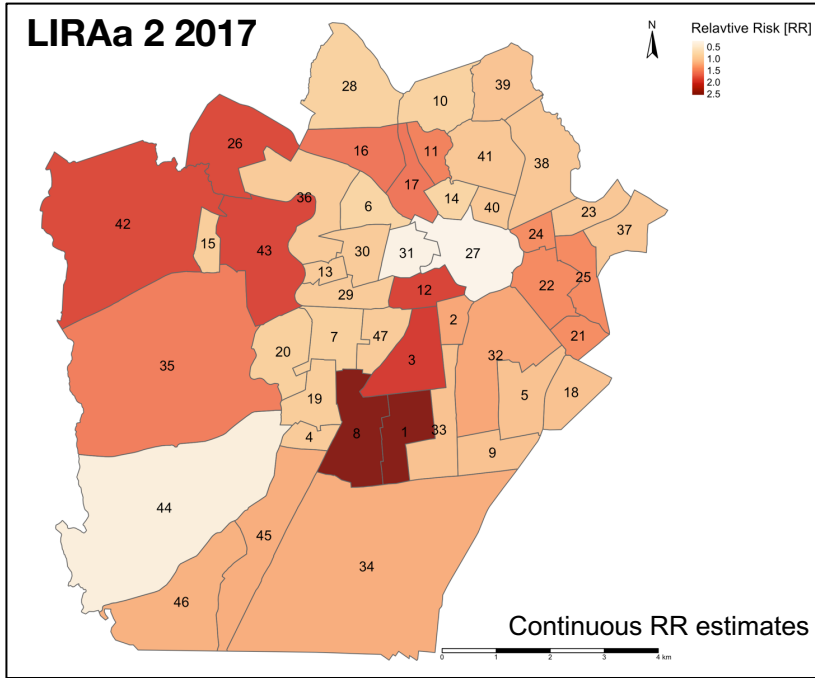
PART 3: Maps on the right panel illustrates which neighbourhoods in Campina Grande have RRs that are significantly "low" or "high" risk

LIRAa 2 2017

RR estimates categorised:

RR < 1.00 (Low risk)
RR > 1.00 (High risk)
RR = 1.00 (Non-significant risk)

Interpretation:
The following neighbourhoods in Campina Grande numbered 1, 3, 8 and 12 (for example) have RRs that are significantly above 1.00. These are examples of neighbourhoods containing households predicted to be at '**high risk**' of being infested with mosquitoes. Neighbourhoods painted in **RED** need to be monitored for mosquito breeding hotspots to prevent further infestation, which, in turn, can lead to infectious disease outbreaks e.g., Zika or Dengue viruses!

34

Any questions?