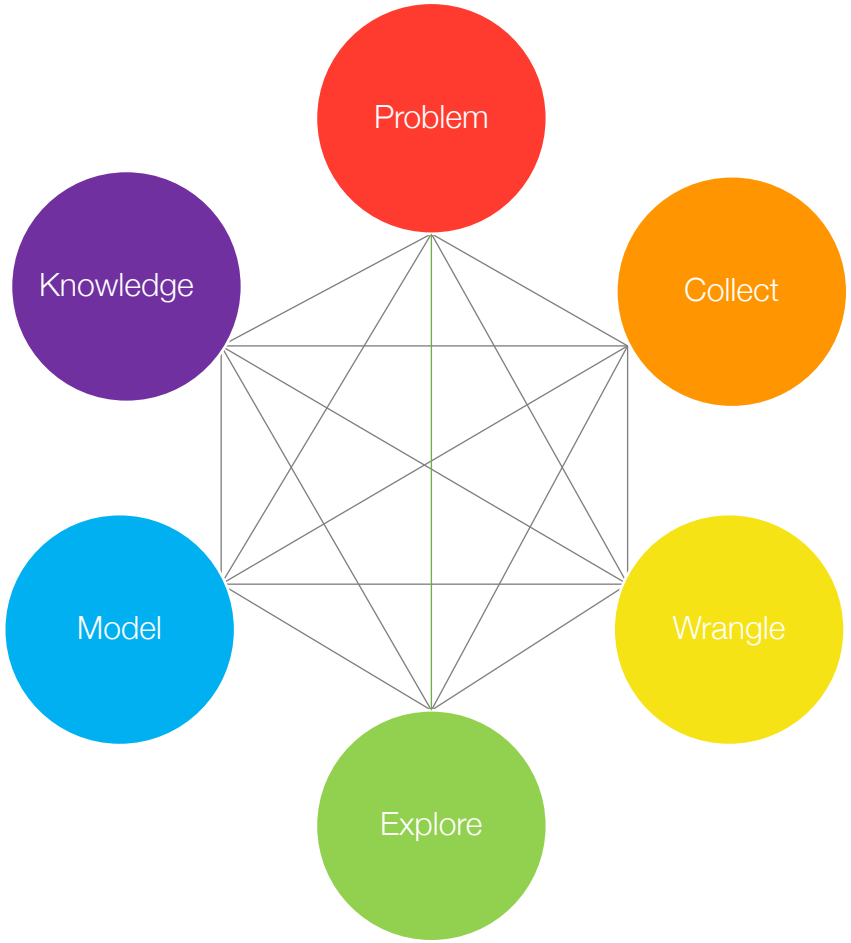
GEOG0125
ADVANCED TOPICS IN SOCIAL AND GEOGRAPHIC DATA SCIENCE

BAYESIAN SPATIAL RISK MODELLING IN STAN

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)
- Alternatives (R-INLA)



Quick recap on hierarchical regression models

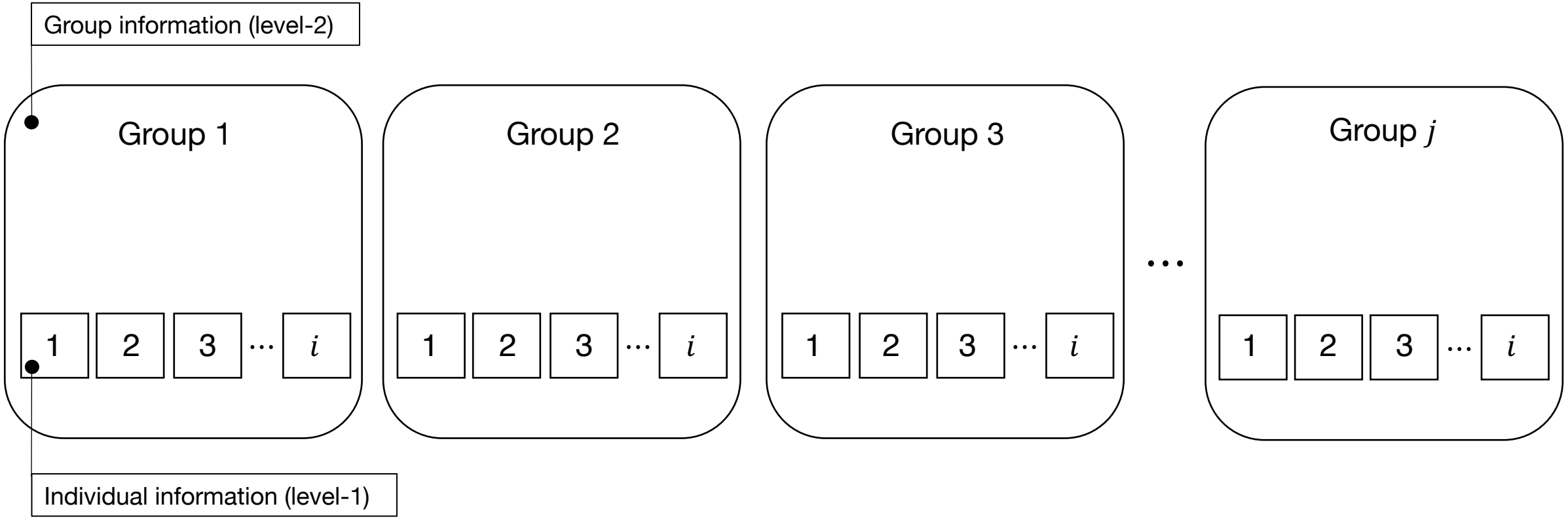
Definition:

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:

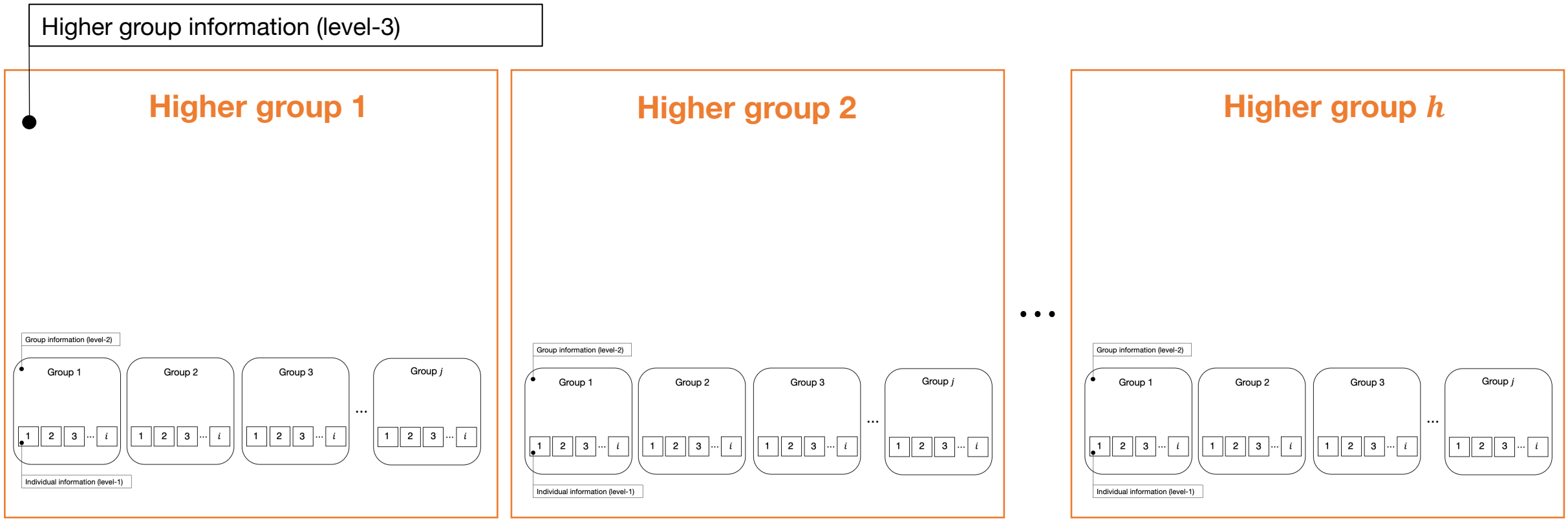
- It is an elegant way to model datasets that have varying scales in their measurements (- this artefact is caused by the multilevel or hierarchical structure in the dataset)
- It is an robust approach for accounting for **variations across individual units**, and at the same time, the "**within-group 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]



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]



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.

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	$y_{3,3}$	$y_{1,3,3}$	$x_{3,3,3}$	$x_{3,3,3}$	$\cdots$	$x_{k,3,3}$
⋮	⋮	⋮	⋮	⋮	⋮	$\cdots$	⋮
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

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

- 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

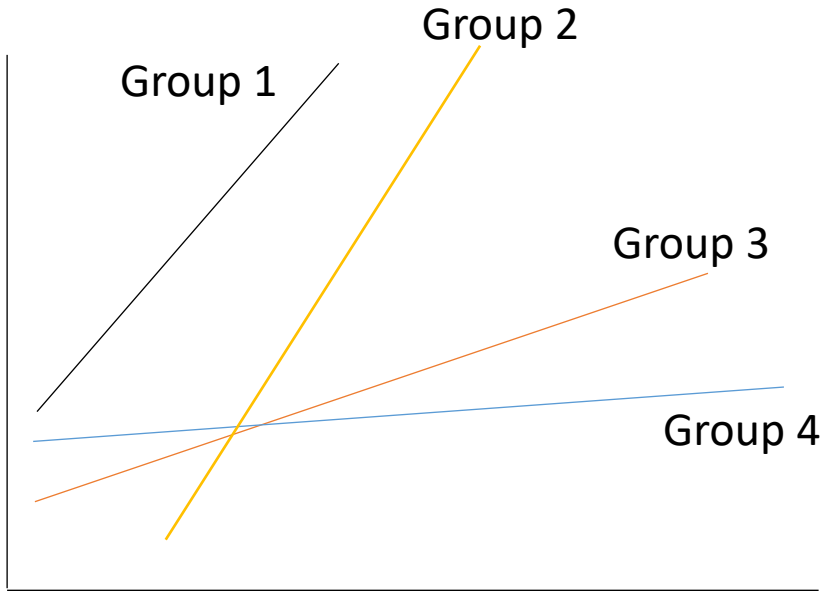
$$\beta_{0,j} = \gamma_{00} + u_{0,j}$$
$$\beta_{1,j} = \gamma_{10} + u_{1,j}$$
$$\beta_{2,j} = \gamma_{20} + u_{2,j}$$
$$\vdots \quad \vdots \quad \vdots$$
$$\beta_{k,j} = \gamma_{k0} + u_{k,j}$$

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

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

Model's true form



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

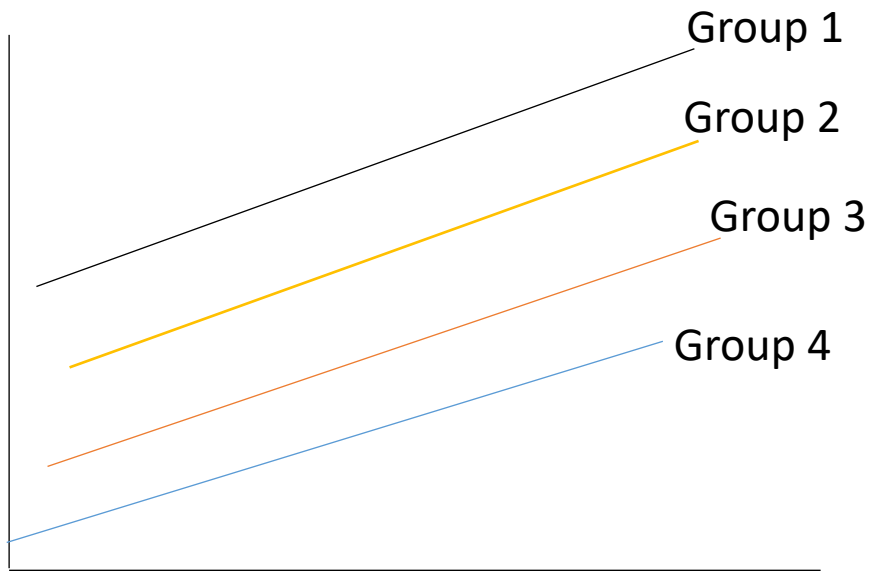
$$y_{i,j} = \beta_{0,j} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + \cdots + \beta_{k,j}x_{k,i,j} + \varepsilon_{i,j}$$ Level 1 Equation

$$\beta_{0,j} = \gamma_{00} + u_{0,j}$$ Level 2 Equation

Here, the model is much simpler:

$$y_{i,j} = \underbrace{\gamma_{00} + \beta_{1,j}x_{1,i,j} + \beta_{2,j}x_{2,i,j} + \cdots + \beta_{k,j}x_{k,i,j}}_{\text{Fixed part}} + \underbrace{u_{0,j} + \varepsilon_{i,j}}_{\text{Random part}}$$ Model's true form

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



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

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

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:

$$\begin{aligned}
\Rightarrow y_{i,j} = {} & \gamma_{00} + \gamma_{01}Z_1 + \\
& \gamma_{10}x_{1,i,j} + \gamma_{20}x_{2,i,j} + \cdots + \gamma_{k0}x_{k,i,j} + \\
& \gamma_{11}Z_1x_{1,i,j} + \gamma_{20}Z_1x_{2,i,j} + \cdots + \gamma_{k0}Z_1x_{k,i,j} + \\
& u_{0,j} + u_{1,j}x_{1,i,j} + u_{2,j}x_{2,i,j} + \cdots + u_{k,j}x_{k,i,j} + \varepsilon_{i,j}
\end{aligned}$$

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

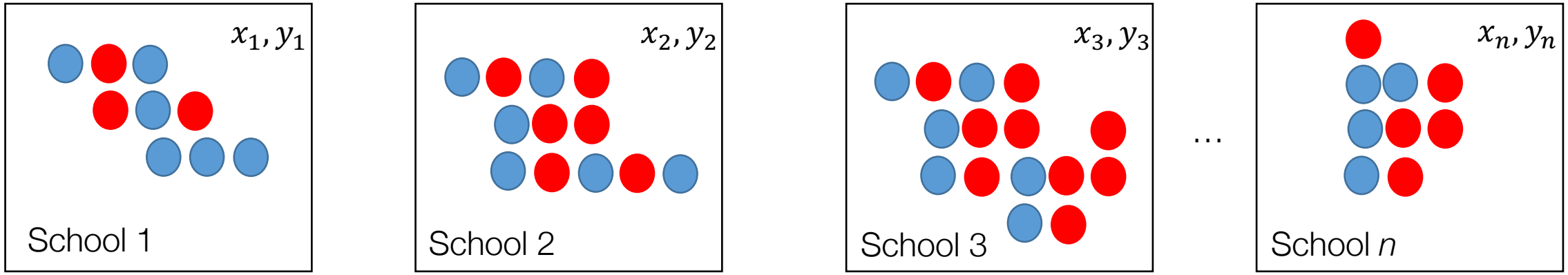
These are random coefficients for the interacting variables from the level 1 & 2 equation

These are the random effects

Situating hierarchical models within a spatial context:

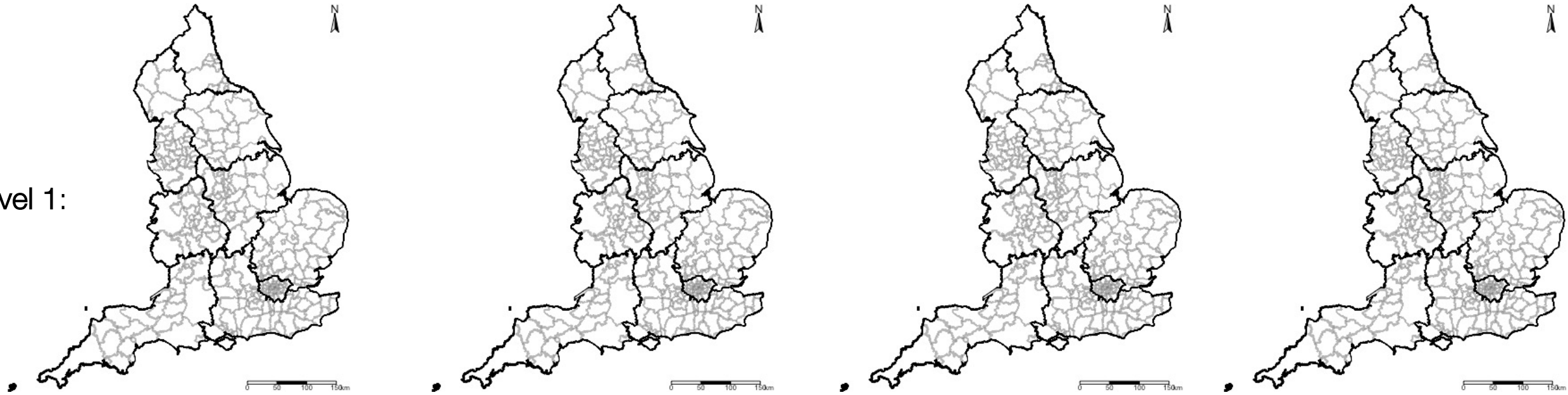
Bayesian inference, are often used in hierarchical modelling, which are models commonly used in the quantification of spatial and spatiotemporal areal data.

- 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



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

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)

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

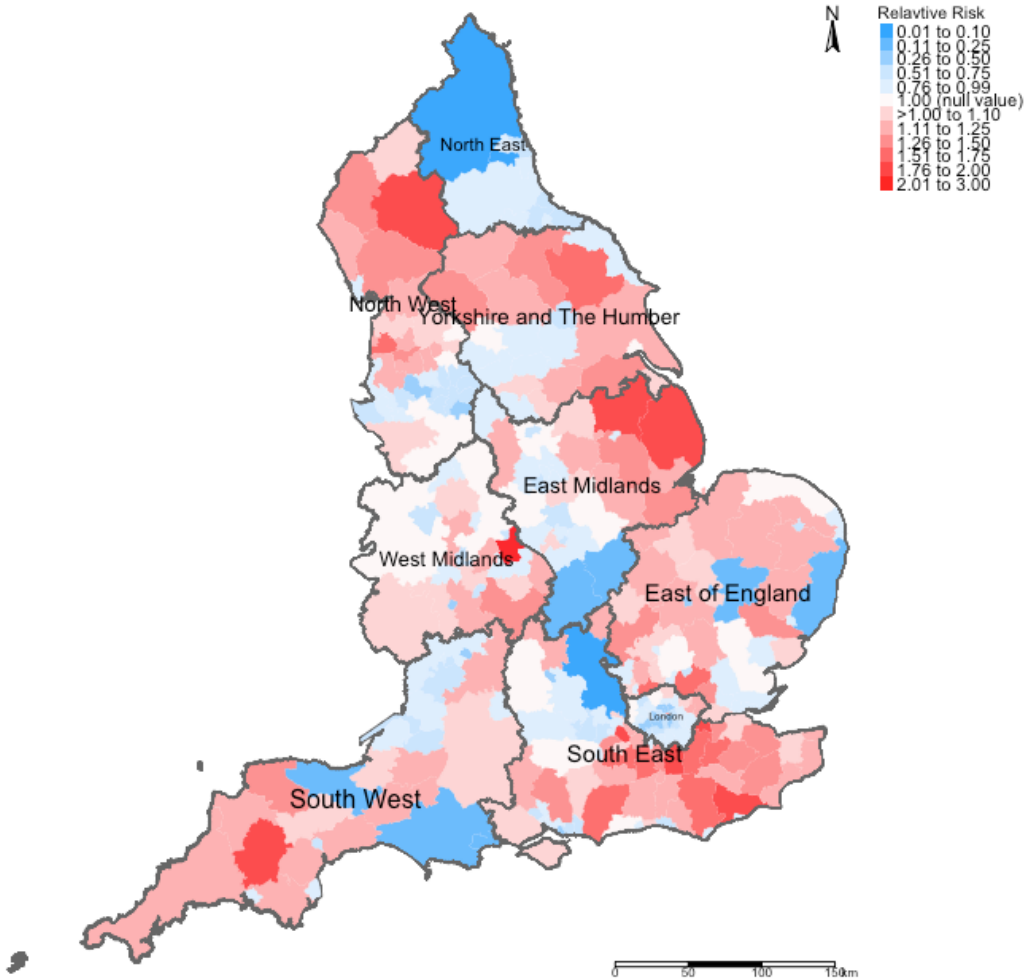
Exceedance Probability

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

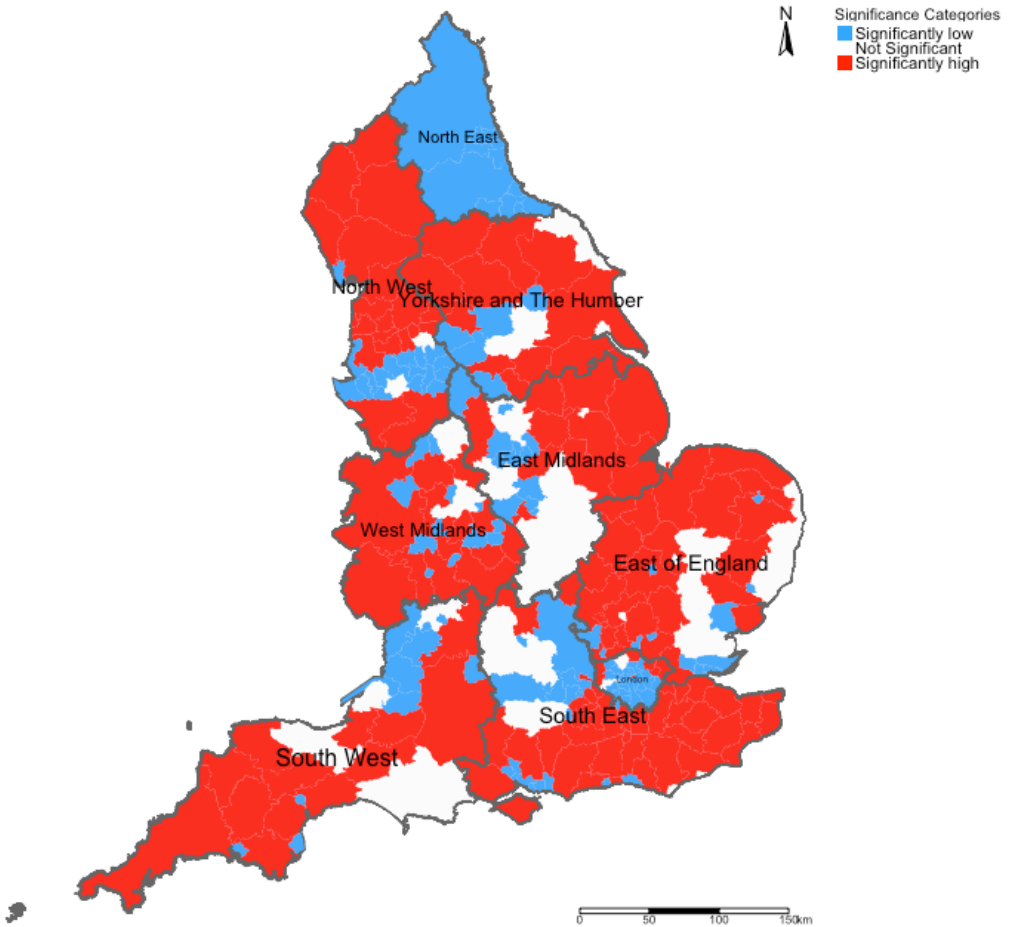
A common example used in every day application are disease risk models, we are usually concerned about areas that have excess risk of a disease type i.e., P(RR > 1)

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]

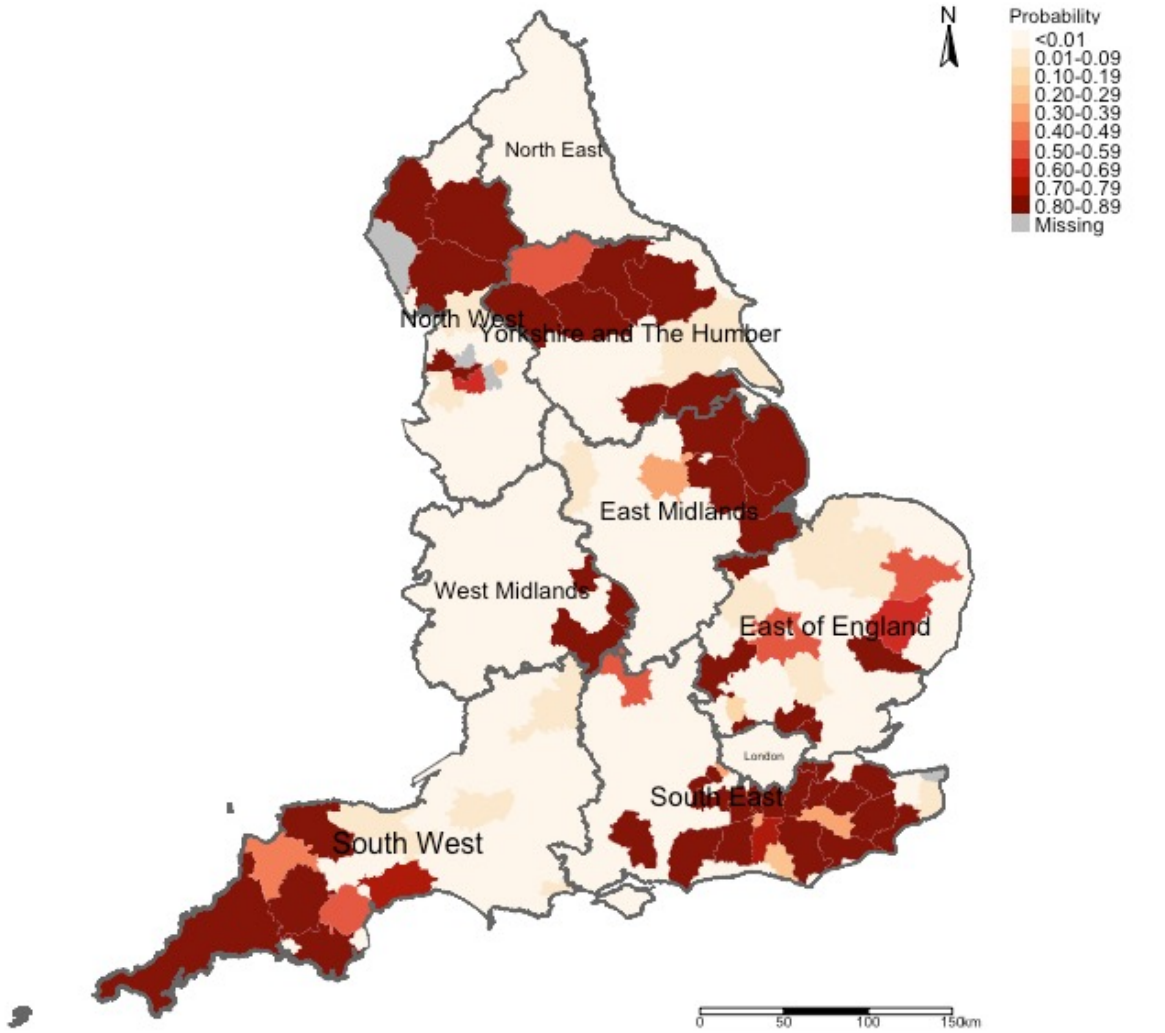


Relative Risks (RR)



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)

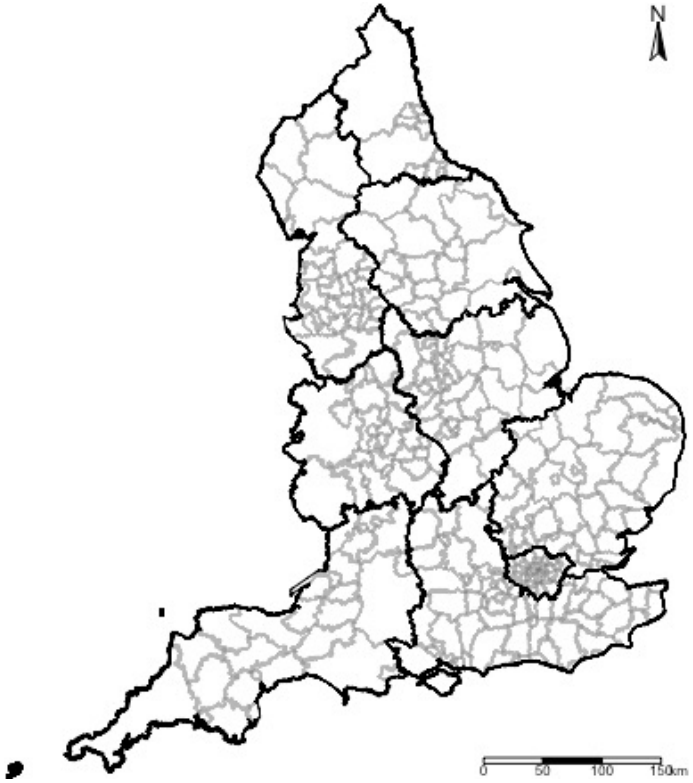
Spatial Intrinsic Conditional Autoregressive models (iCARs)

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

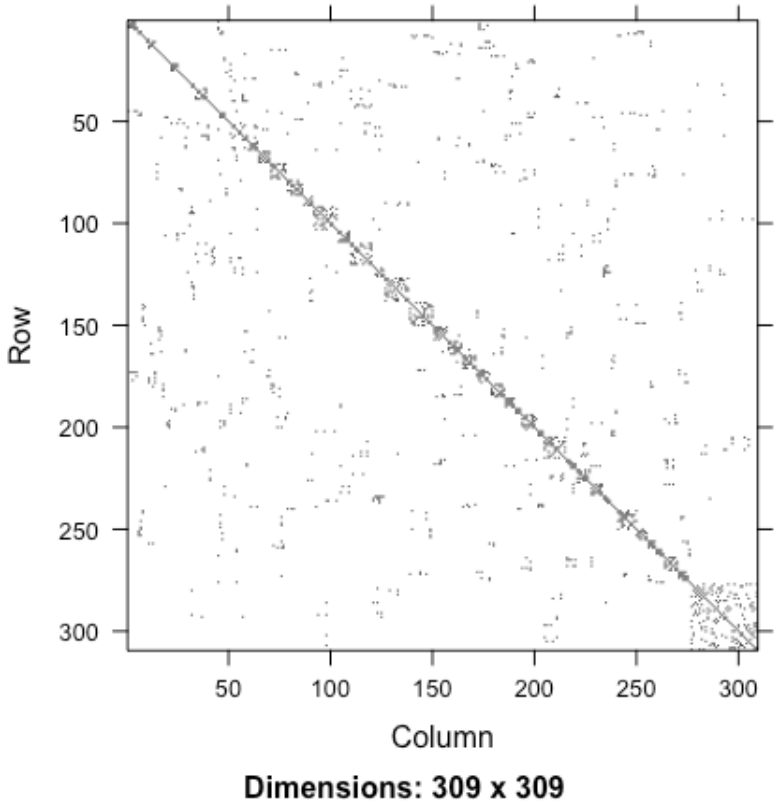
This is a popular spatial model which takes into account that the data may potentially be spatially correlated and the observations in the neighbouring areas may be more similar than observations in areas that are distant from each other.

- This is a type of hierarchical model which includes a spatial random effect,
- It is heavy dependent on 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).

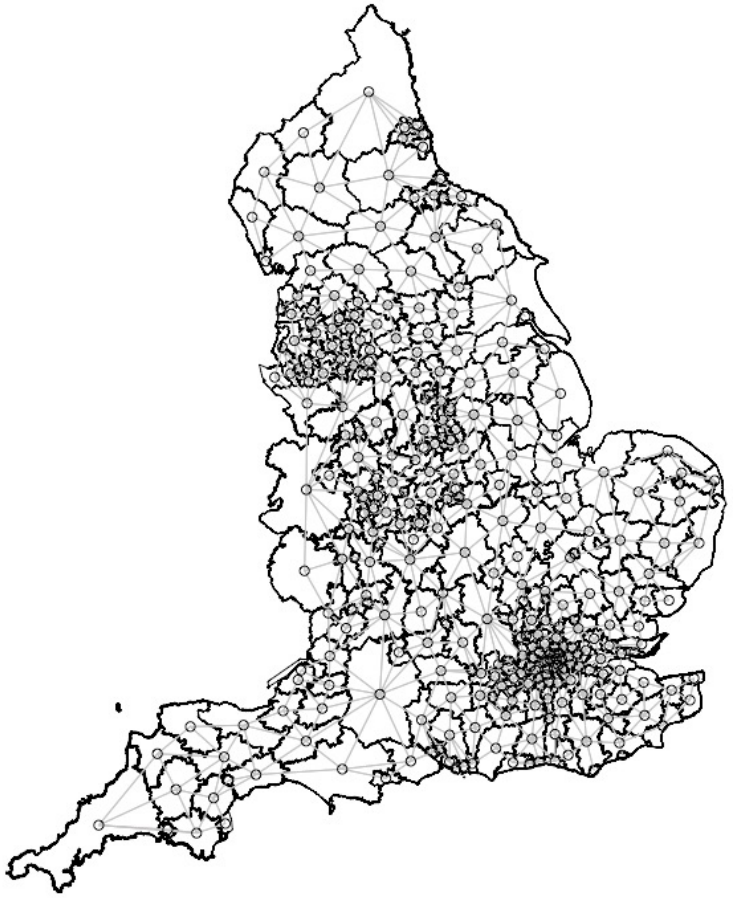
Geographically accurate neighbourhood structure



Adjacency matrix translated to graph format

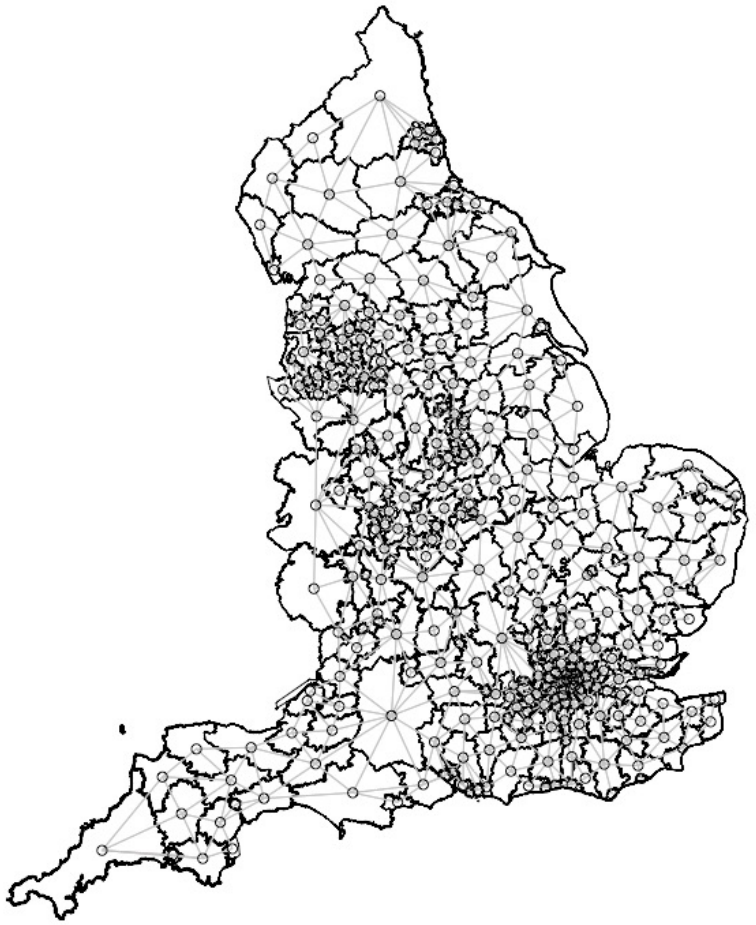


Adjacency matrix translated to nodes and edge format



Stan only uses the nodes and edges format to reconstruct the adjacency matrix

Adding the spatial effect component to the model



Adjacency matrix translated to nodes and edge format

- It is through the reconstruction of our adjacency matrix through nodes and edges is what we use to set our priors on the spatial effect φ
- In a typical regression model we derive and overall intercept which shows us the average of the outcome (when all other independent variables a held fixed) on a population
- By adding the spatial effect component to the population average, it allows the quantities to vary across areal units.

In Stan:

```
functions {
      real icar_normal_lpdf(vector phi, int N, array[] int node1, array[] int node2) {
                                          return -0.5 * dot_self(phi[node1] - phi[node2]);
}
```

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.

Prior: $\varphi \sim \text{icar normal}(\text{N}, \text{nodes1}, \text{nodes2})$ (Specified in the model block)

Model formulation from a Bayesian framework

Model components

Y_i are observed counts of cases (outcome)
E_i are expected numbers of cases in an area (offset)
φ_i are the area-specific spatial effects
ρ_i is some area-specific rates
α is the overall risk of the study area
σ an overall error term multiplied to the spatial effects

Note:
$\exp(\alpha)$ is the overall risk ratio for study area
$\exp(\beta)$ is the overall risk ratio for coefficient
$\exp(\alpha + \sum \beta_k X_{i,k} + \varphi_i \sigma)$ is risk ratio for each area

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

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

$$[1]\ \log(\lambda_i) = \alpha + \varphi_i \sigma + \log(E_i) \text{ (no variables)}$$
$$[2]\ \log(\lambda_i) = \alpha + \sum \beta_k X_{i,k} + \varphi_i \sigma + \log(E_i)$$

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

$$\alpha \sim \text{Norm}(0, 5)$$
$$\beta \sim \text{Norm}(0, 5)$$
$$\sigma \sim \text{Norm}(0, 5) \text{ (alternatives are gamma(0.001, 0.001))}$$
$$\varphi \sim \text{icar normal}(\text{N}, \text{nodes1}, \text{nodes2})$$

- Build Bayesian model

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

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

Model formulation from a Bayesian framework

Stan code

```
functions {
real icar_normal_lpdf(vector phi, int N, array[] int node1, array[] int node2) {
                        return -0.5 * dot_self(phi[node1] - phi[node2]);
                        }
}
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=0>[N] E;
}
transformed data {
vector[N] log_offset = log(E);
}
parameters {
real alpha;
real<lower=0> sigma;
vector[N] phi;
}
model {
phi ~ icar_normal(N, node1, node2);
Y ~ poisson_log(log_offset + alpha + phi*sigma);
alpha ~ normal(0.0, 1.0);
sigma ~ normal(0.0, 1.0);
sum(phi) ~ normal(0, 0.001*N);
}
generated quantities {
vector[N] eta = alpha + phi*sigma;
vector[N] mu = exp(eta);
}
```

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

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

$$[1]\ \log(\lambda_i) = \alpha + \varphi_i \sigma + \log(E_i) \text{ (no variables)}$$
$$[2]\ \log(\lambda_i) = \alpha + \sum \beta_k X_{i,k} + \varphi_i \sigma + \log(E_i)$$

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

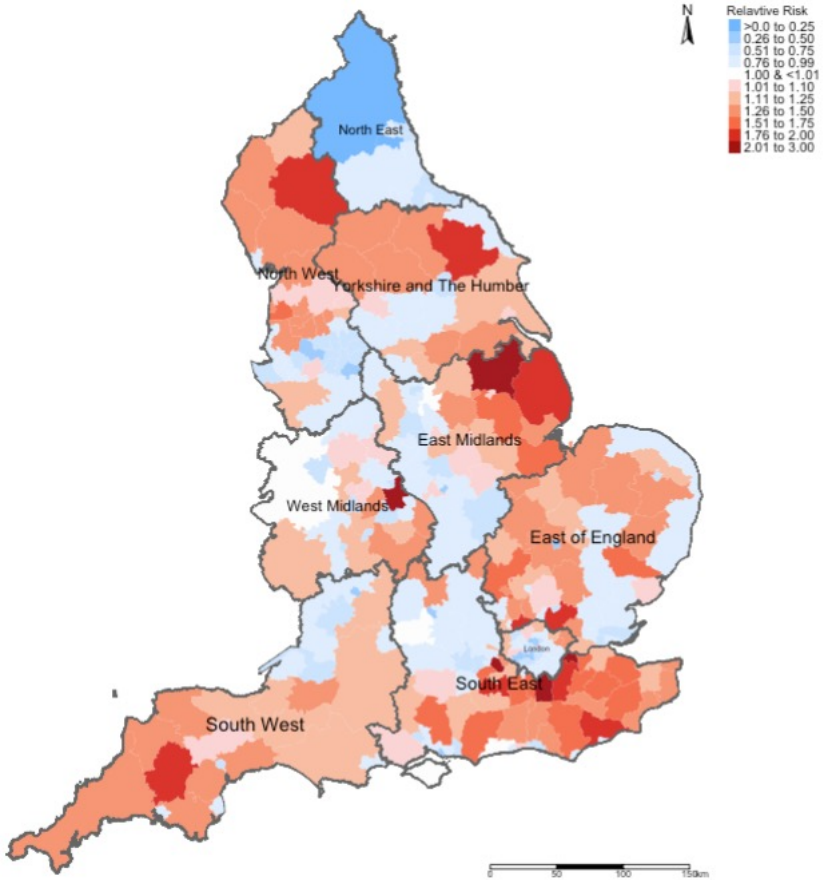
$$\alpha \sim \text{Norm}(0, 5)$$
$$\beta \sim \text{Norm}(0, 5)$$
$$\sigma \sim \text{Norm}(0, 5) \text{ (alternatives are gamma(0.001, 0.001))}$$
$$\varphi \sim \text{icar normal}(\text{N}, \text{nodes1}, \text{nodes2})$$

- Build Bayesian model

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

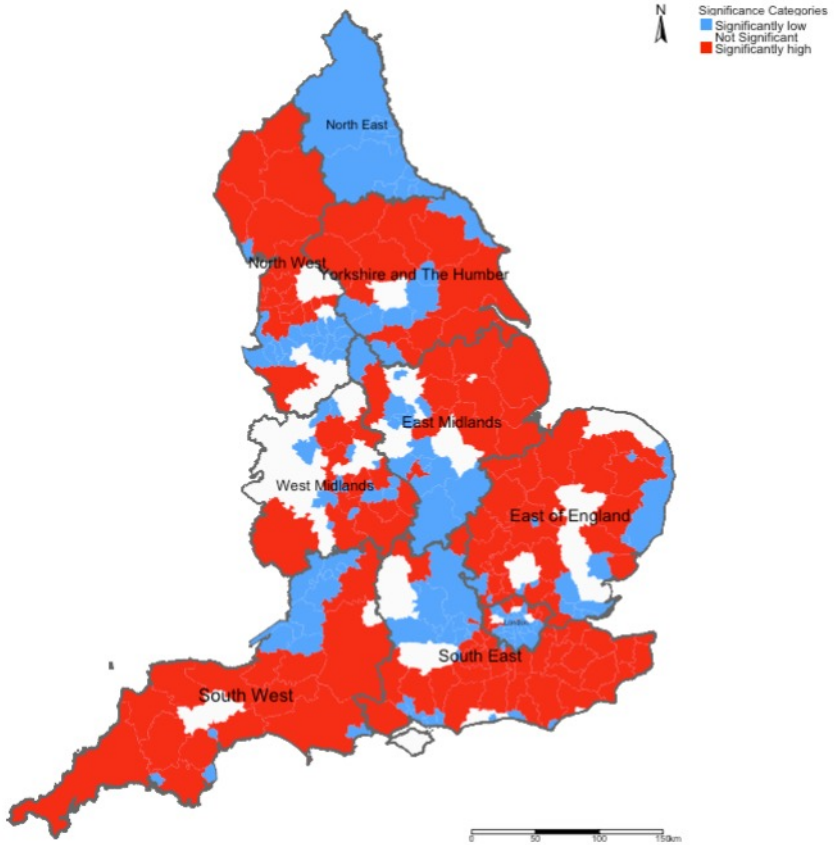
$$P(\alpha, \beta_k, \sigma, \varphi_i | \lambda_i) \propto P(\lambda_i | \alpha, \beta_k, \sigma, \varphi_i)\, P(\alpha) P(\beta_k) P(\sigma) P(\varphi_i)$$

Relative risk ratios (RR)



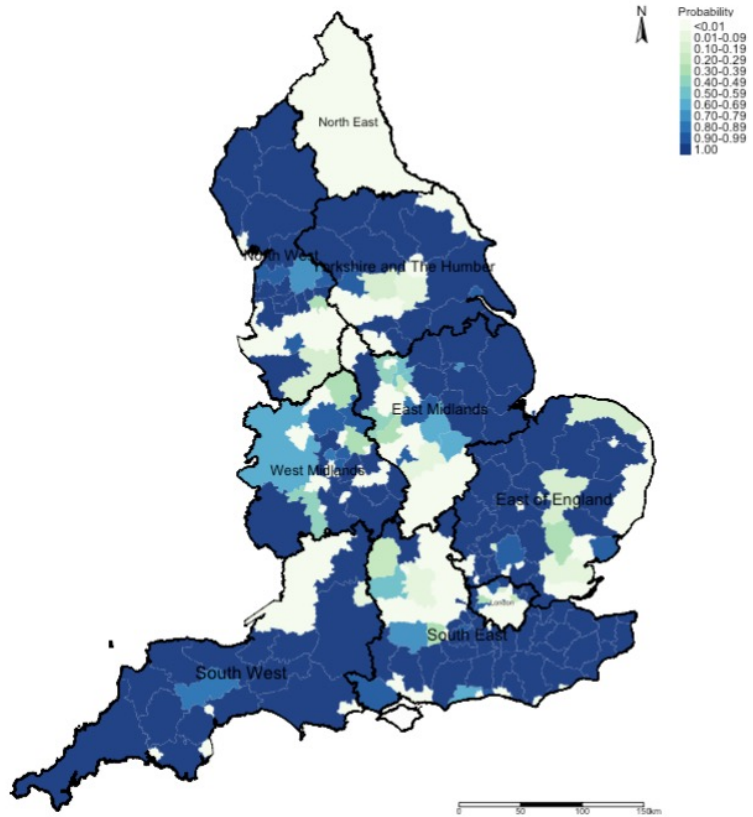
Here, we use this output to describe the burden of an outcome

Statistical Significance



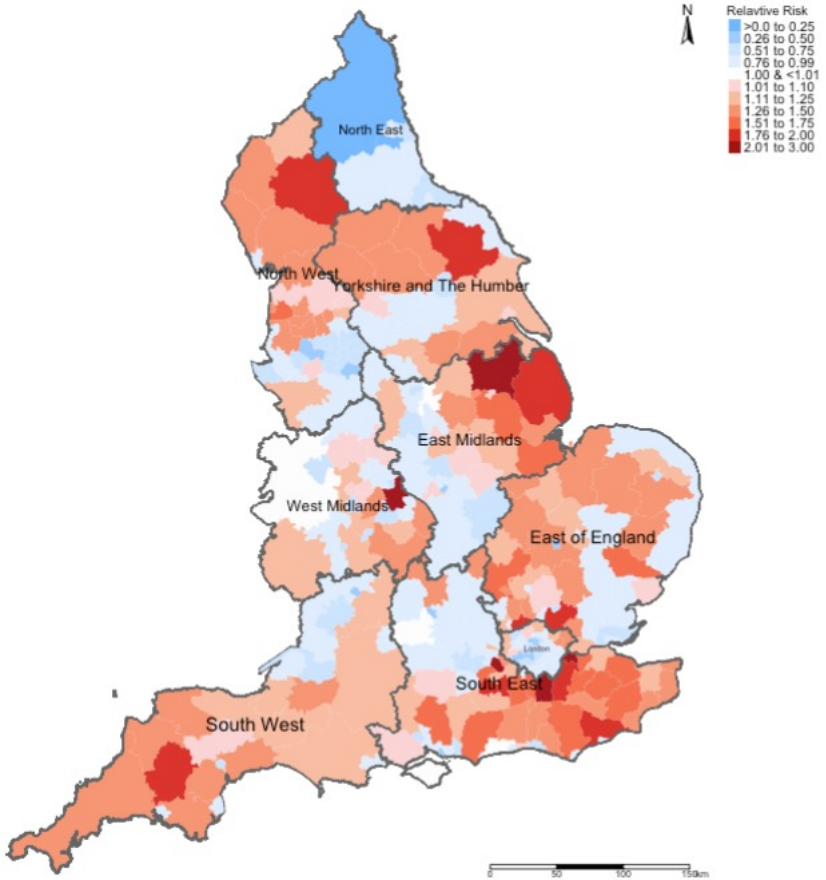
This output is to valid whatever hypothesis we had about the described outcome's burden in the first map

Exceedance Probabilities

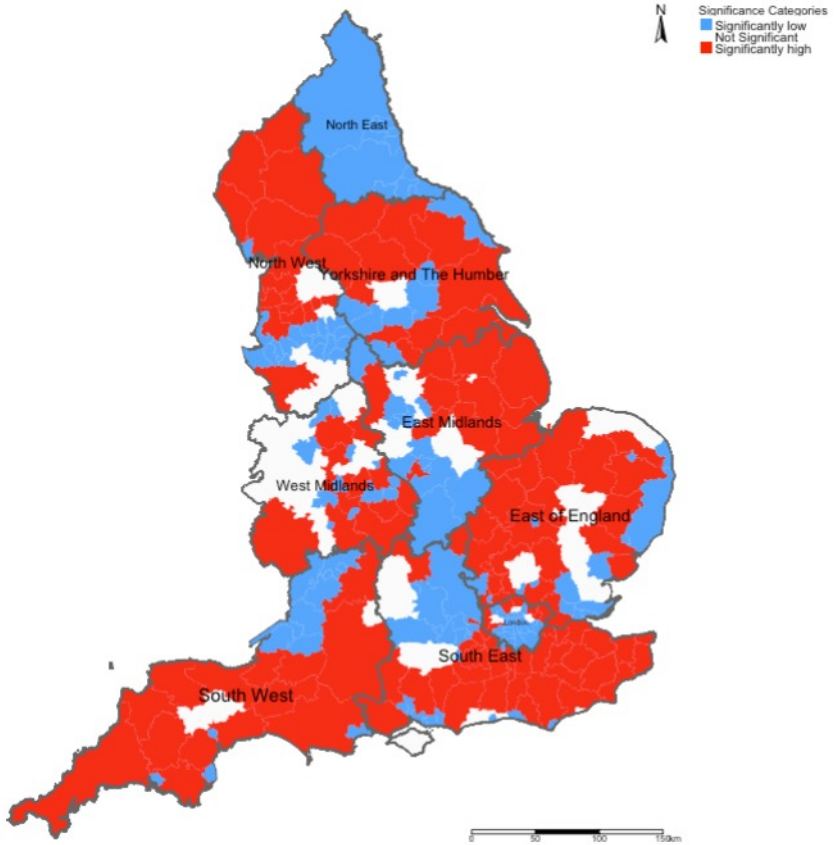


This output is used to describe the uncertainty that surrounds the risks we found in the first map

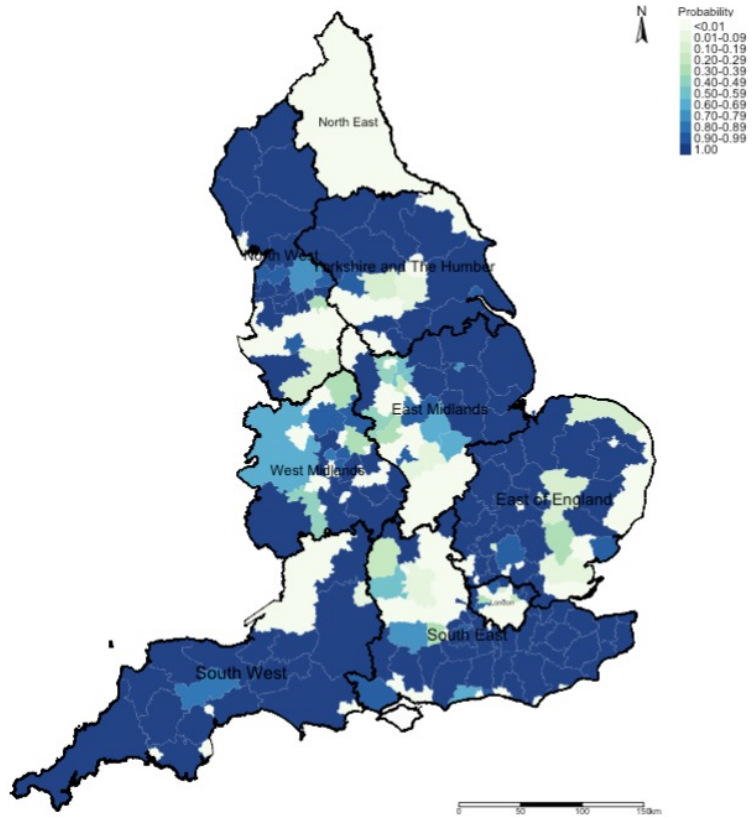
Relative risk ratios (RR)



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.

Any questions?