Continuing Professional Development (CPD) course
Introduction To Bayesian Inference & Modelling (2022/23)

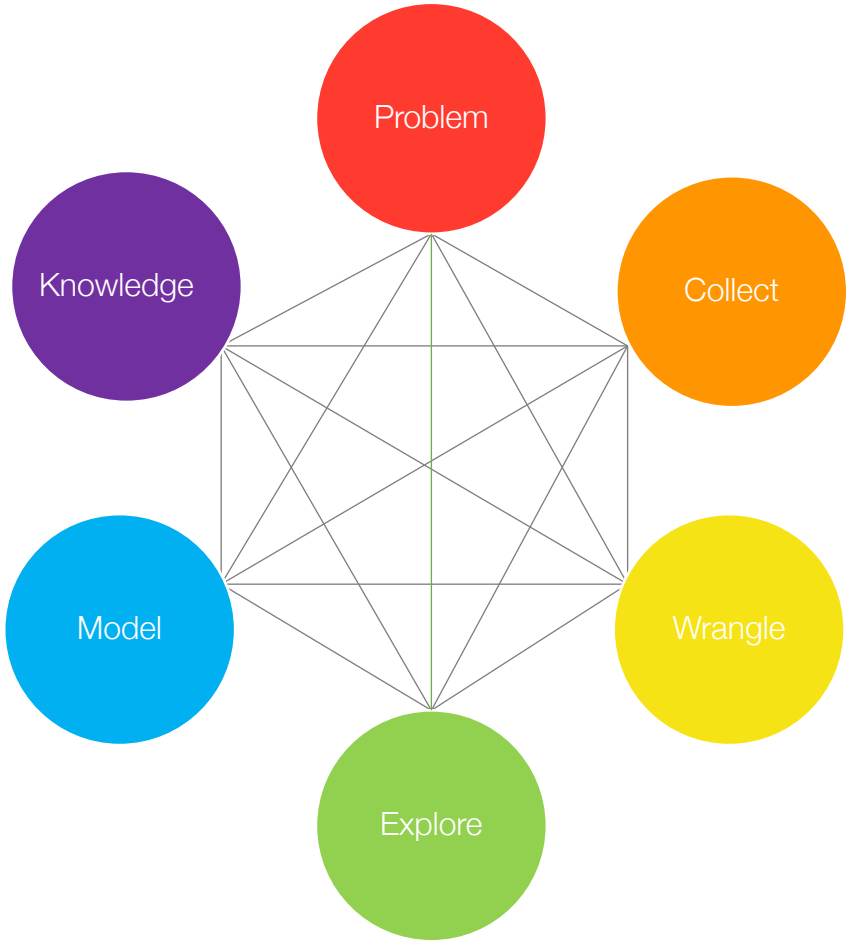
DAY 4: BAYESIAN SPATIAL RISK MODELLING IN STAN

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

Additional details:
https://www.ucl.ac.uk/social-data

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)
  - ❖ Structured and Unstructured Random Effects
- Model formulation from a Bayesian Framework
- Examples and interpretation (using Stan)



Quick recap on hierarchical regression models

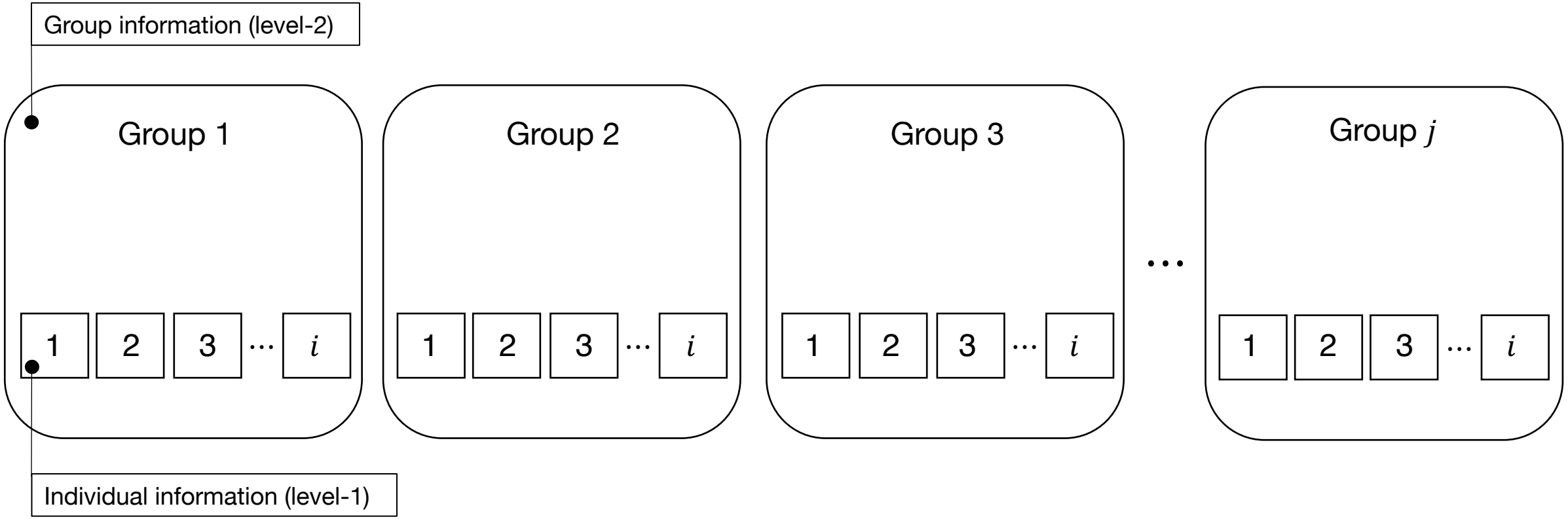
Definition:

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:

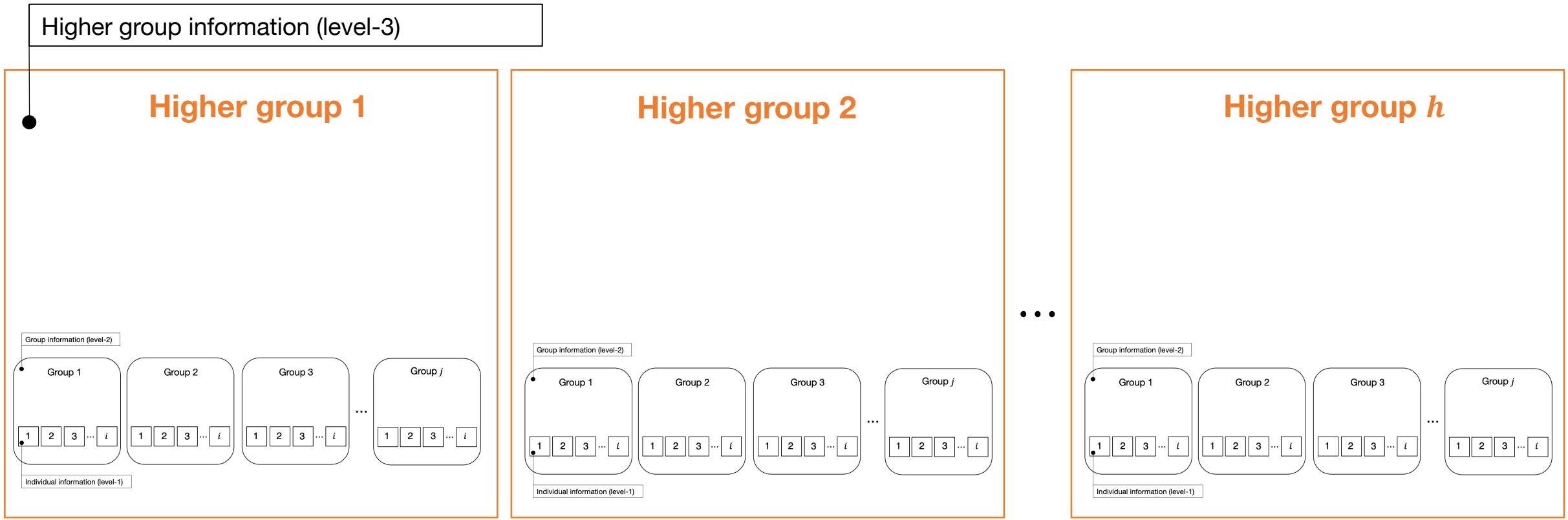
- It is an elegant way to model datasets that have varying scales in their measurements (- this artefact is caused by the multilevel or hierarchical structure in the dataset)
- It is an robust approach for accounting for **variations across individual units**, and at the same time, the "**within-group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

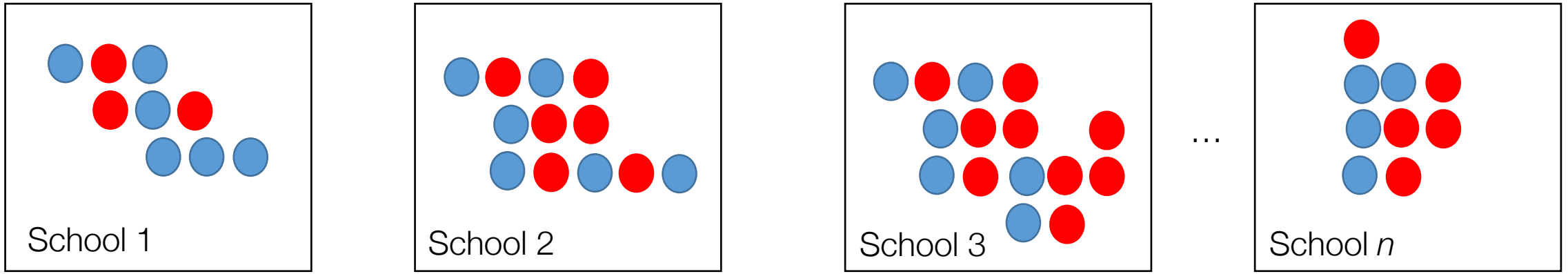
Situating hierarchical models within a spatial context:

- Hierarchical models are in fact commonly used in the quantification of certain parameters when dealing spatial and spatiotemporal areal data.

- These are statistical model written in mind to deal with multiple hierarchies formed by geographies to estimate parameters of the posterior distribution

- Example: Intestinal parasitaemia among school children in Tanzania and infection status linked with anaemia, here students (level-1) are clustered in schools, and the point location serves as the highest hierarchy (level-2).

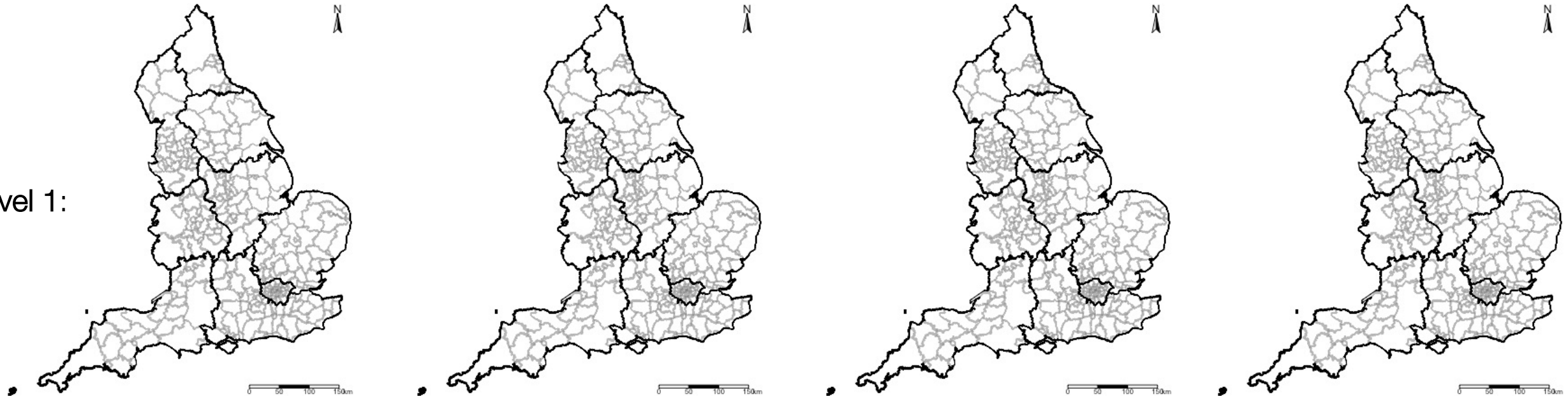
Health student = Diseased student =



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 completely 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

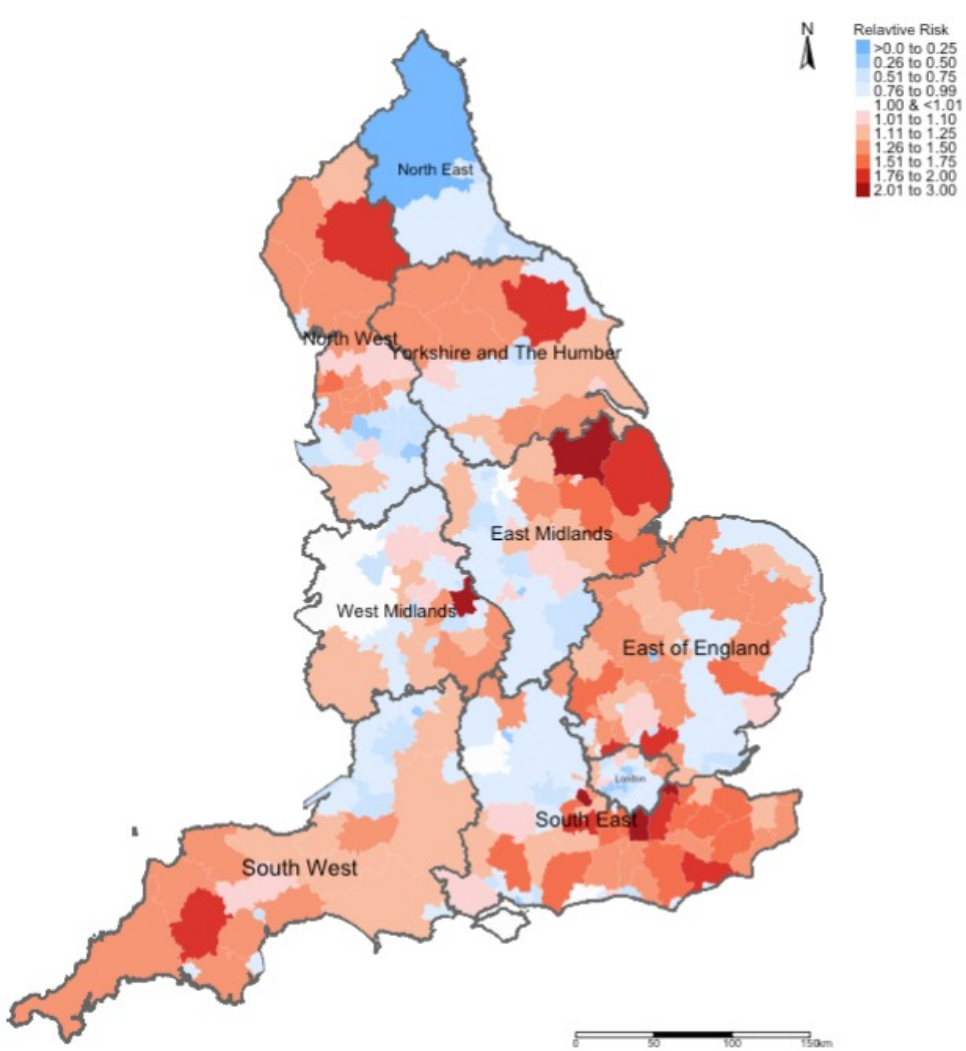
Exceedance Probability

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

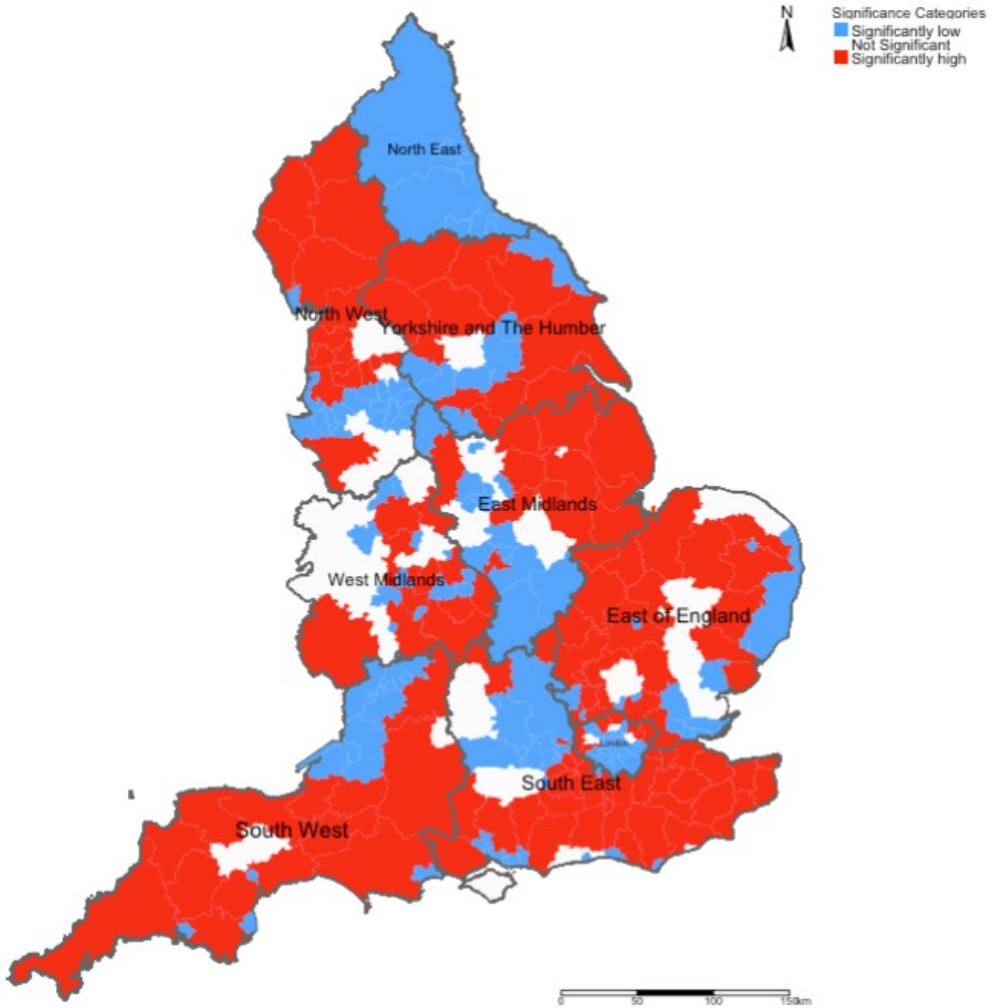
A common example used in every day application are disease risk models, we are usually concerned about areas that have excess risk of a disease type i.e., P(RR > 1.00) or P(OR > 1.00)

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]

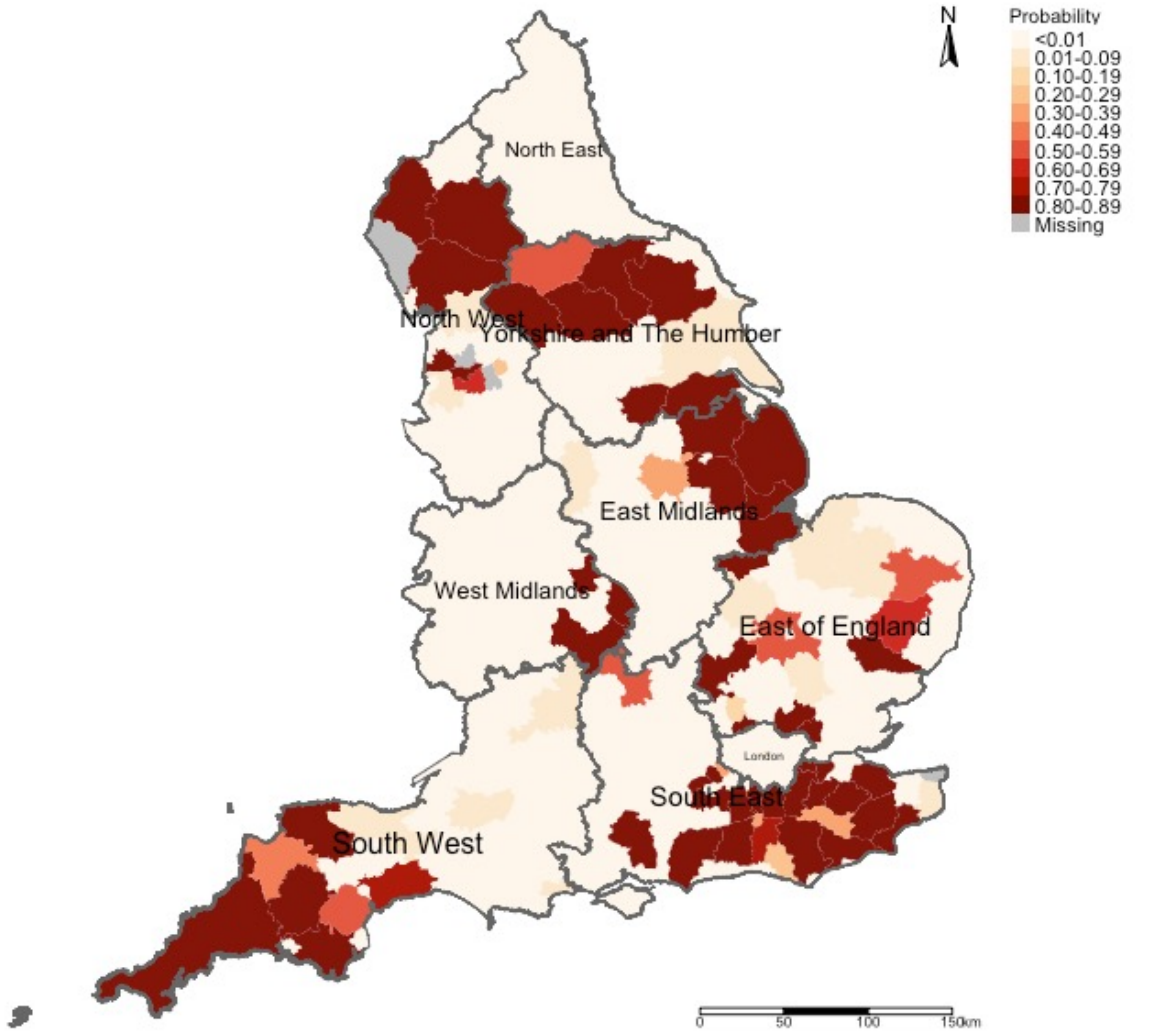


Relative Risks (RR)



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

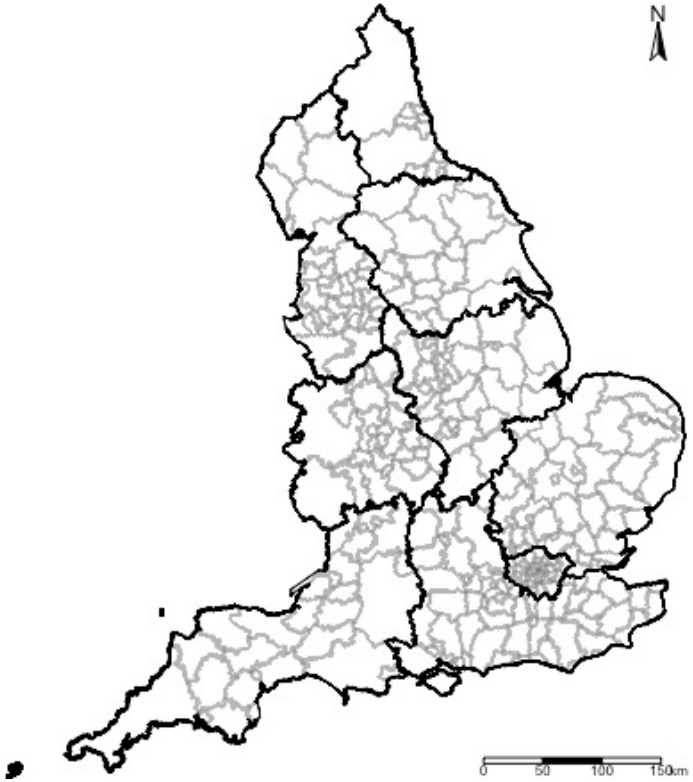
Spatial Intrinsic Conditional Autoregressive models (ICARs)

Besag-York-Mollie (BYM) model

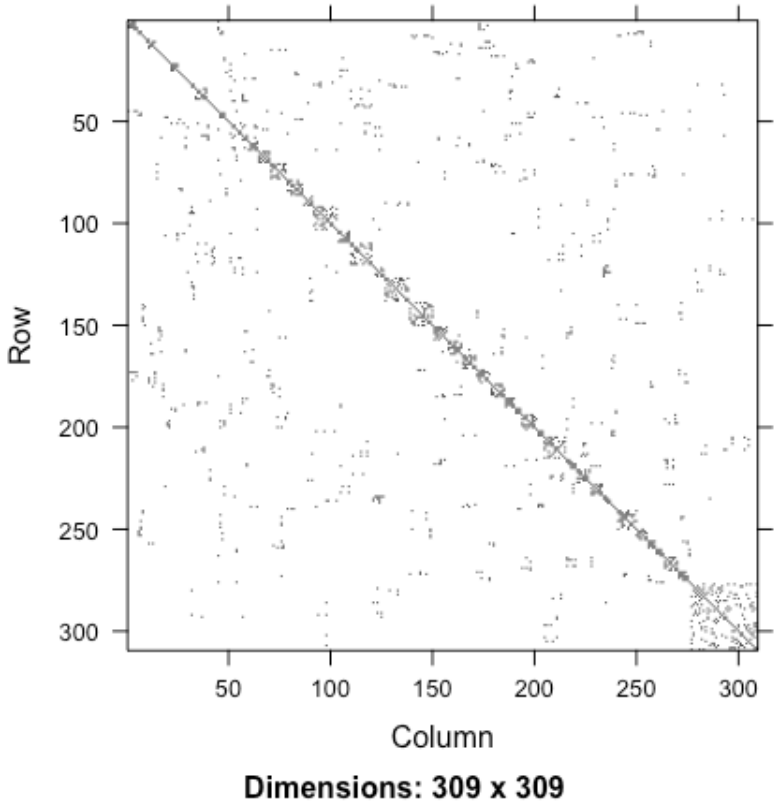
This is a popular spatial model which takes into account that the data may potentially be spatially correlated and the observations in the neighbouring areas may be more similar than observations in areas that are distant from each other.

- 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): Conditional Autoregressive Model (CAR)
  - ❖ BYM model that has both a spatial effect term which is treated a structured random effect, and the error term is an unstructured noise: Intrinisic Conditional Autoregressive Model (iICAR)
- 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). where you have counts and denominators that represent the reference population.

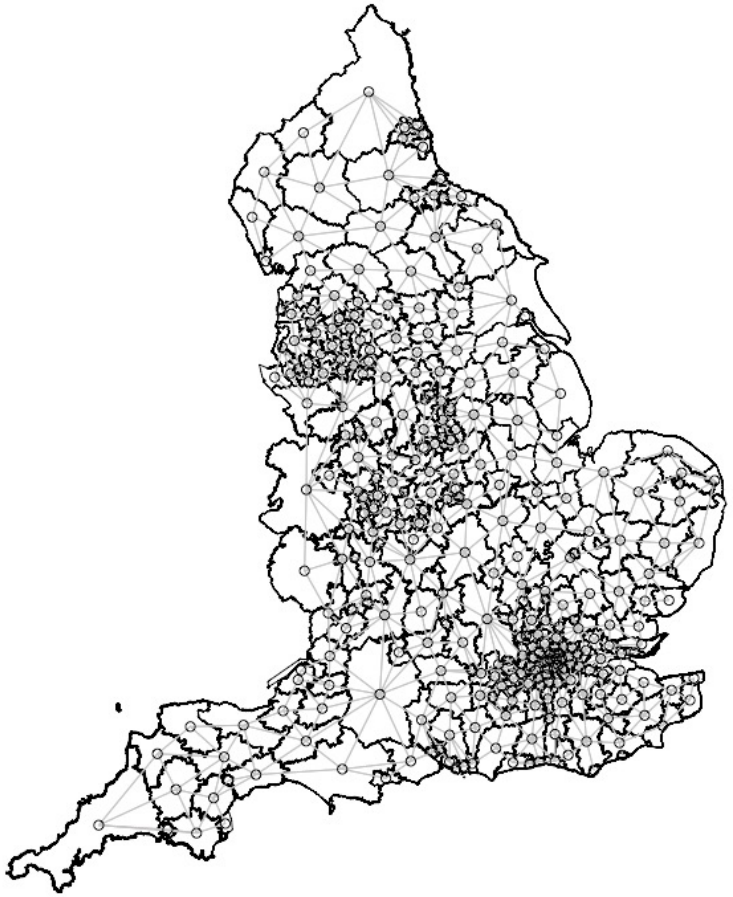
Geographically accurate neighbourhood structure



Adjacency matrix translated to graph format



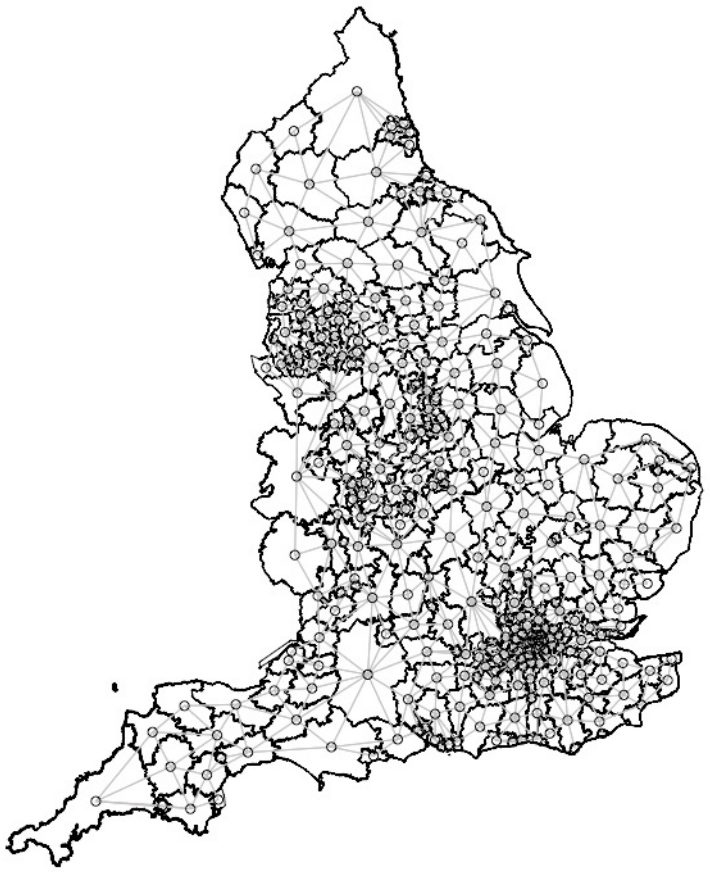
Adjacency matrix translated to nodes and edge format



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

You need to construct a neighbourhood adjacency matrix to account for the spatial configuration of the study area.

Spatial structured and unstructured random effects [1]



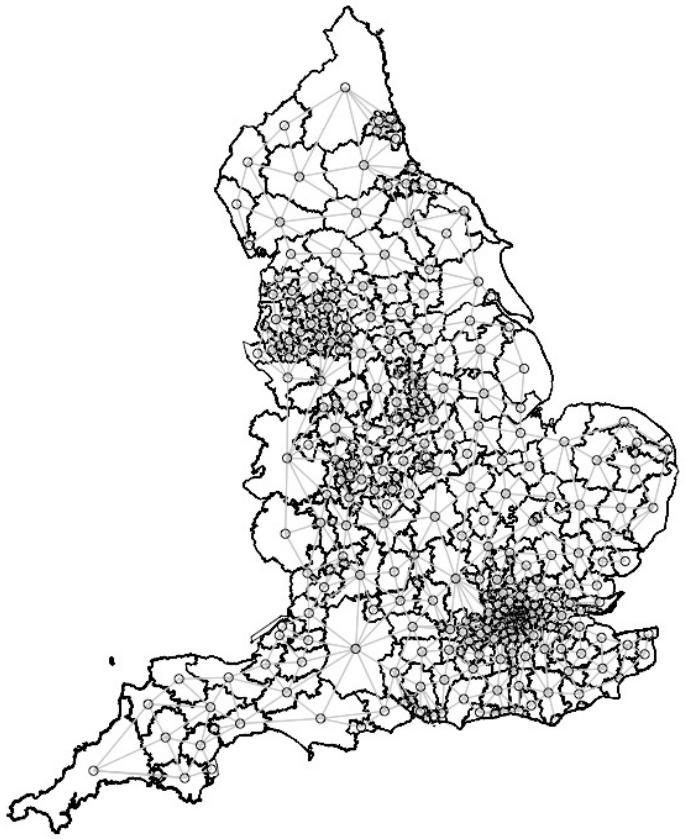
Adjacency matrix translated to nodes and edge format

- **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 $\phi + \theta$

Adding the spatial structured and unstructured effect to the model [2]



Adjacency matrix translated to nodes and edge format

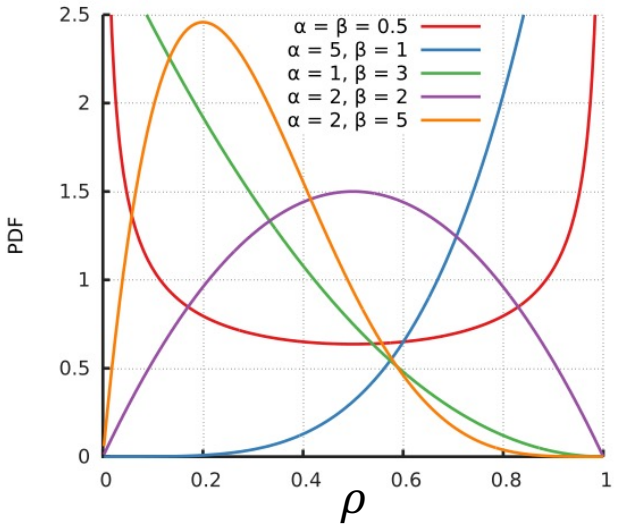
Node1 is the index area of interest;
Node2 is the neighbouring areas connected to the index area defined in Node1.
N is the total number of areas.

```
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:

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

- 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 in the a neighbourhoods (outcome)
$X_{i,k}$ independent variables
E_i are expected counts of 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

Parameters

α is the overall risk 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 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

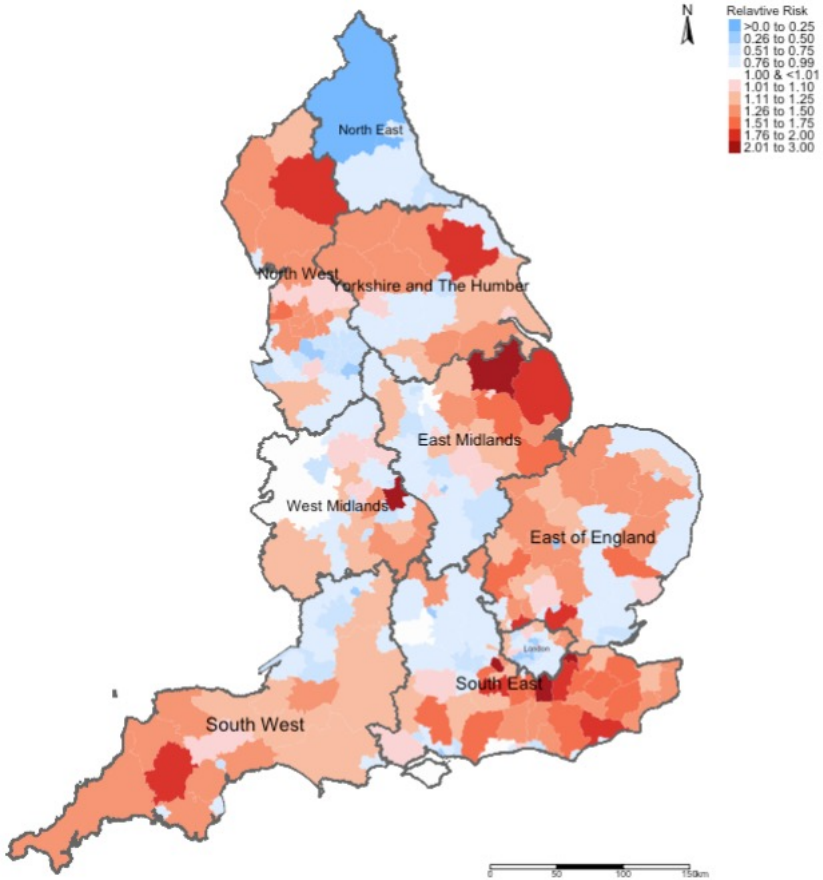
$\alpha \sim \text{norm}(0, 1)$
$\beta \sim \text{norm}(0, 1)$
$\sigma \sim \text{norm}(0, 1)$ (alternatives are $\text{gamma}(0.001, 0.001)$)
$\rho \sim \text{beta}(0.5, 0.5)$
target += -0.5*dot_self(phi[node1] – phi[node2]) (calculates weights)
sum(phi) ~ normal(0, 0.001*N)

- Build Bayesian model

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

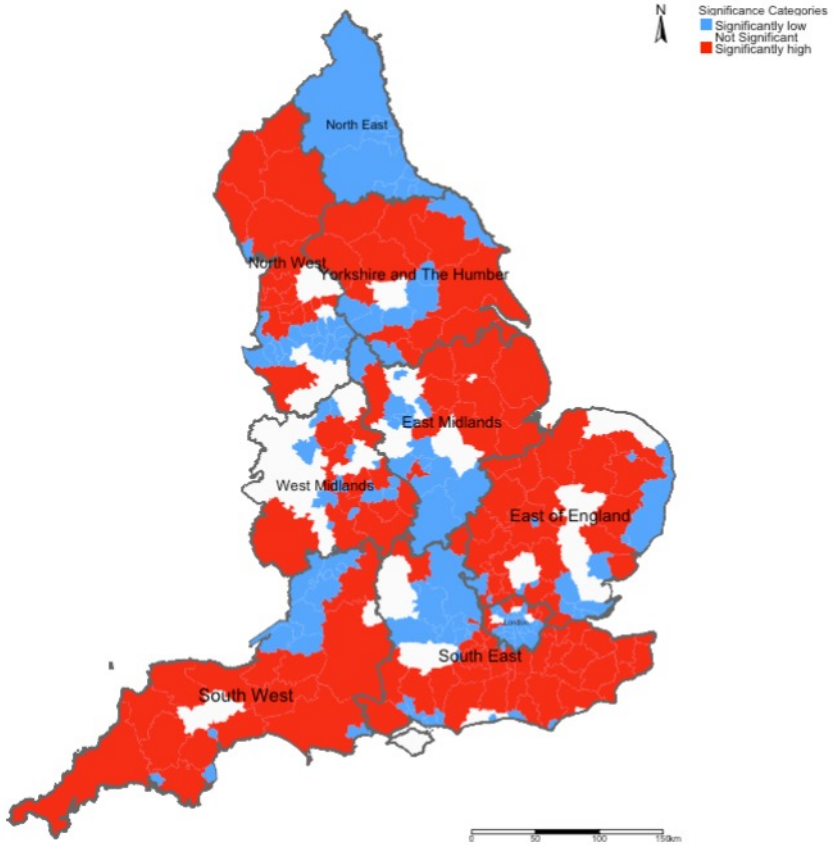
$P(\alpha, \beta_k, \sigma, \varphi_i|\, \lambda_i) \propto P(\, \lambda_i|\alpha, \beta_k, \sigma, \varphi_i)\, P(\alpha)P(\beta_k)P(\sigma)P(\phi_i)P(\rho)$

Relative risk ratios (RR)



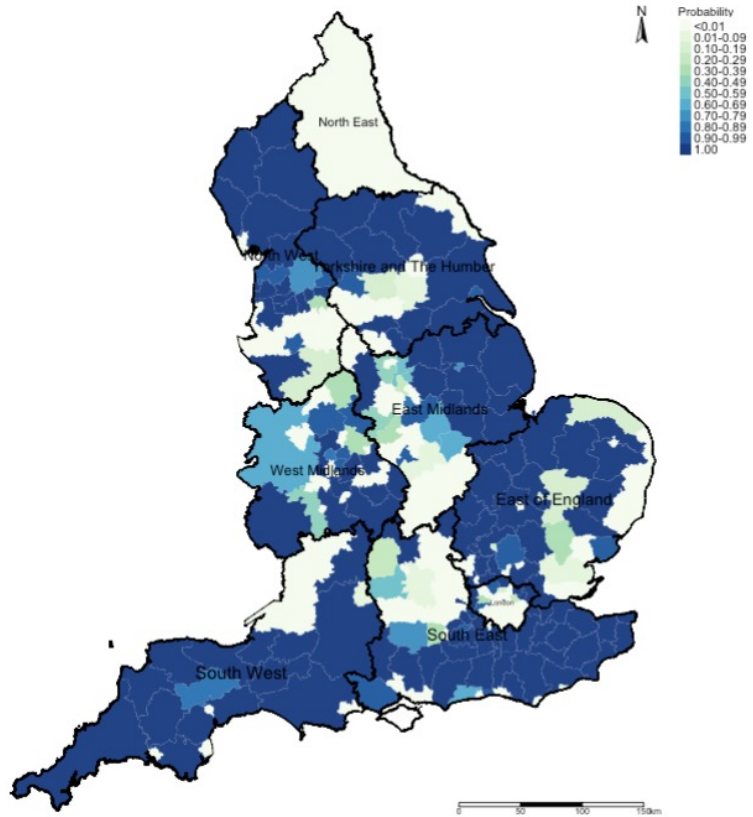
Here, we use this output to describe the burden of an outcome

Statistical Significance



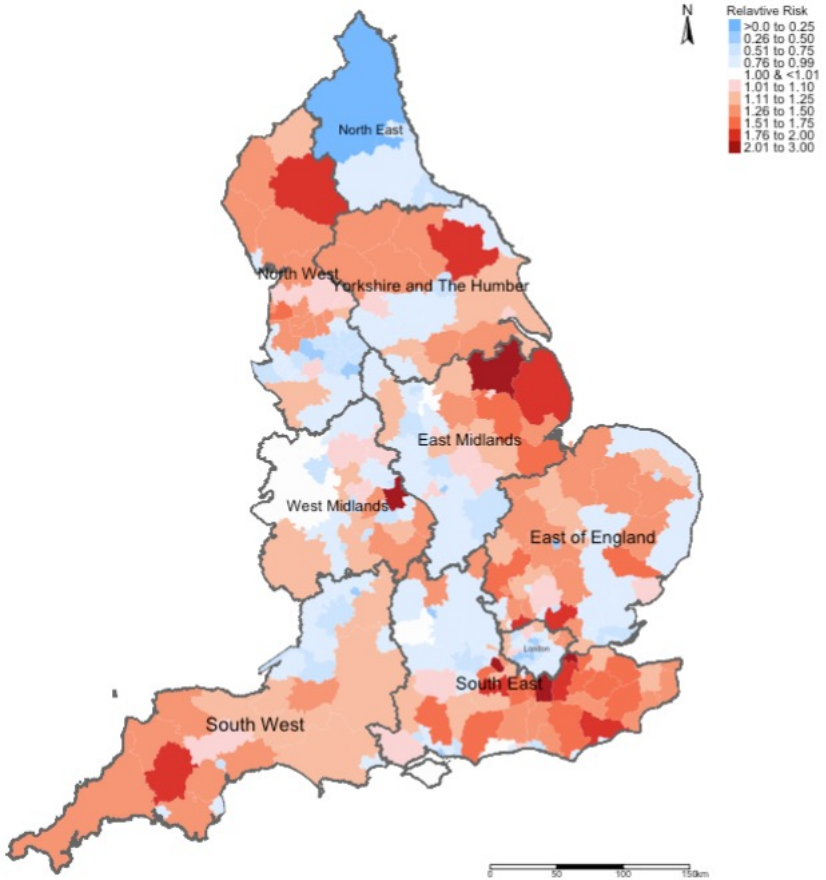
This output is to valid whatever hypothesis we had about the described outcome's burden in the first map

Exceedance Probabilities

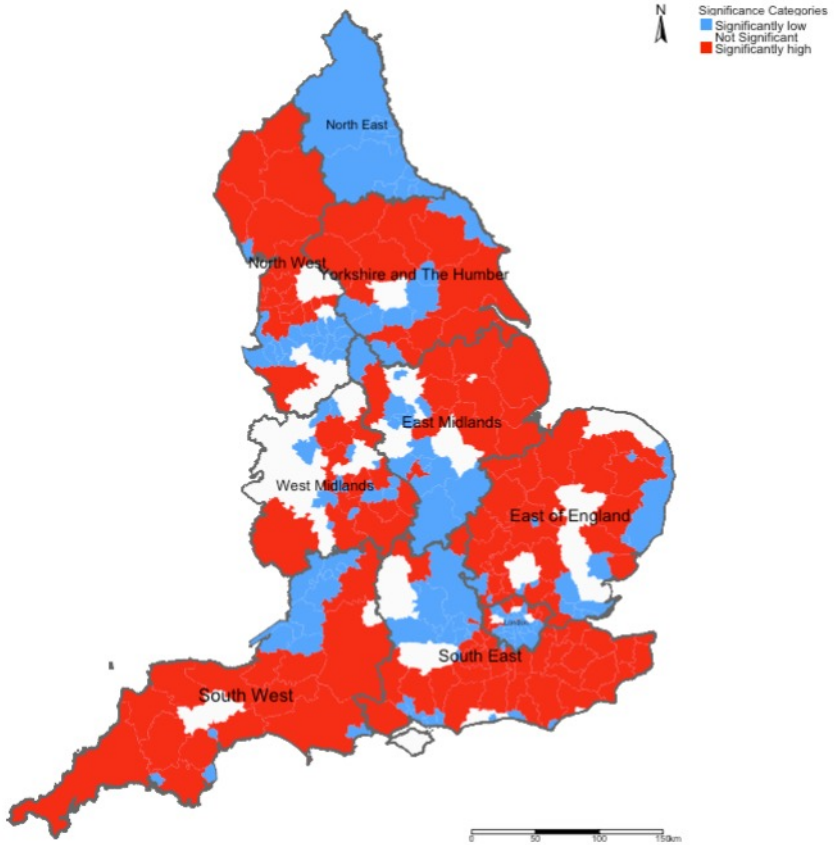


This output is used to describe the uncertainty that surrounds the risks we found in the first map when we explore P(RR>1)

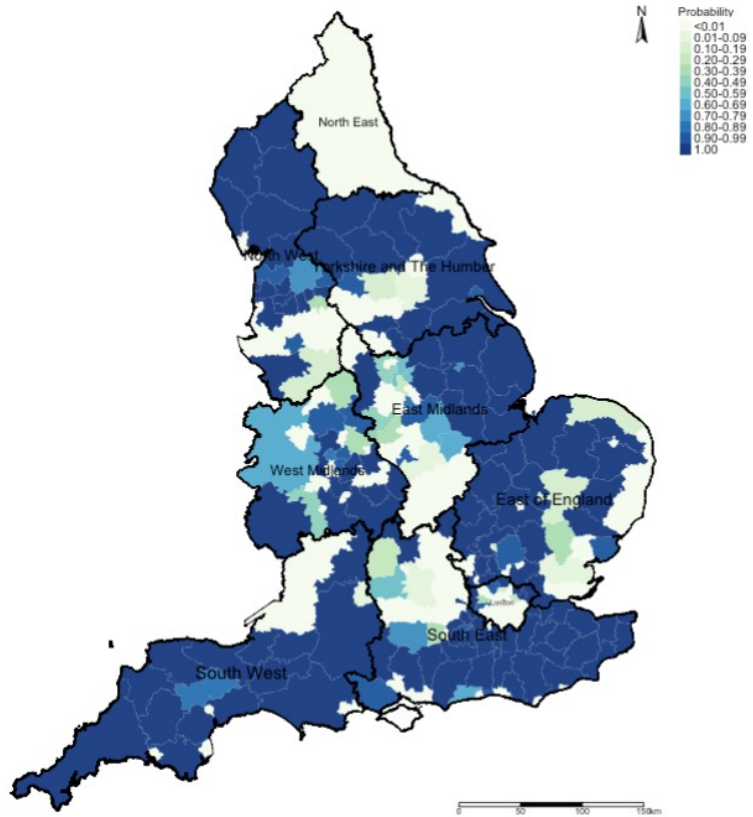
Relative risk ratios (RR)



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

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

Research Methodology & Study design

	Survey Periods				
LIRAa	**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			

Methodology:

- Population-based ecological study design within repeated 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 applied **Bayesian updating** to derive new global coefficients for covariates for each LIRA survey, as well as neighbourhood-specific relative risk estimates.

Model formulation for Spatial ICAR model

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) ~ normal(0, 0.001*N)

- Build Bayesian model

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

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

- New models with LIRAa data that follows are updated, this know as Bayesian updating.

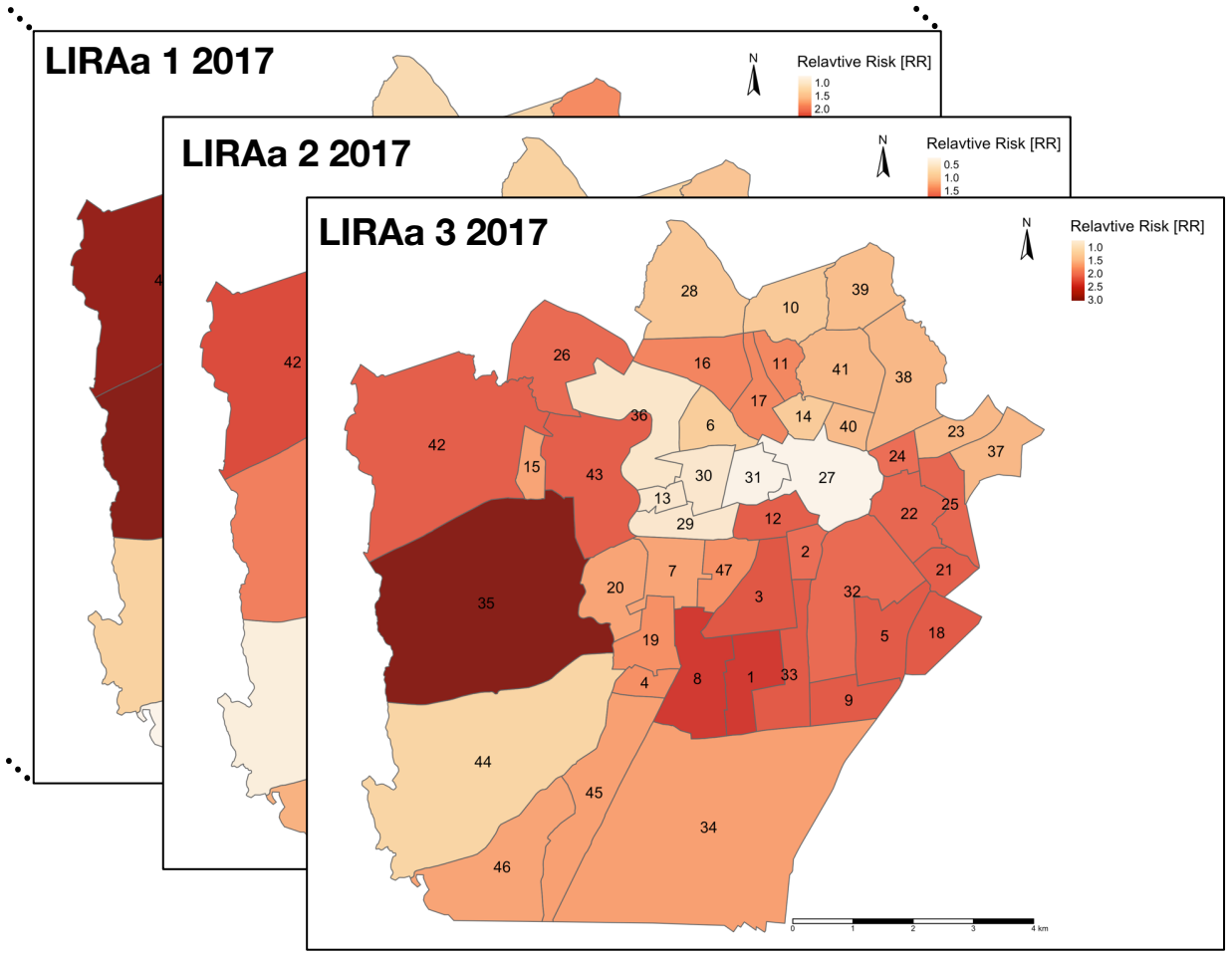
Table results illustrates the overall association between environmental, climate and anthropogenic factors and risk of infestation in Campina Grande.

2016	LIRAa 1		LIRAa 2		LIRAa 3	
	RR (95% CrI)	Pr(RR>1)	RR (95% CrI)	Pr(RR>1)	RR (95% CrI)	Pr(RR>1)
Intercept	1.53 (95% CrI: 0.13 to 6.54)	0.47	1.55 (95% CrI: 0.13 to 6.71)	0.47	1.45 (95% CrI: 0.12 to 6.26)	0.45
Temperature	0.94 (95% CrI: 0.82 to 1.07)	0.18	0.97 (95% CrI: 0.87 to 1.09)	0.33	0.95 (95% CrI: 0.86 to 1.05)	0.17
Precipitation	1.02 (95% CrI: 0.98 to 1.07)	0.87	1.02 (95% CrI: 0.92 to 1.12)	0.67	1.31 (95% CrI: 0.65 to 2.39)	0.75
NDVI	1.01 (95% CrI: 0.87 to 1.16)	0.52	1.02 (95% CrI: 0.78 to 1.31)	0.54	1.02 (95% CrI: 0.59 to 1.62)	0.48
Urbanisation	1.12 (95% CrI: 0.64 to 1.84)	0.62	1.19 (95% CrI: 0.64 to 2.02)	0.68	1.55 (95% CrI: 0.81 to 2.69)	0.91
2017	LIRAa 1		LIRAa 2		LIRAa 3	
	RR (95% CrI)	Pr(RR>1)	RR (95% CrI)	Pr(RR>1)	RR (95% CrI)	Pr(RR>1)
Intercept	1.47 (95% CrI: 0.13 to 6.27)	0.45	1.64 (95% CrI: 0.14 to 7.07)	0.51	1.81 (95% CrI: 0.15 to 7.81)	0.53
Temperature	0.92 (95% CrI: 0.82 to 1.03)	0.09	0.93 (95% CrI: 0.74 to 1.12)	0.23	1.01 (95% CrI: 0.88 to 1.15)	0.57
Precipitation	1.15 (95% CrI: 1.03 to 1.28)	0.99	1.01 (95% CrI: 0.96 to 1.07)	0.73	1.00 (95% CrI: 0.98 to 1.01)	0.61
NDVI	0.93 (95% CrI: 0.62 to 1.33)	0.32	1.09 (95% CrI: 0.71 to 1.60)	0.63	0.94 (95% CrI: 0.84 to 1.06)	0.16
Urbanisation	1.19 (95% CrI: 0.57 to 2.20)	0.64	1.18 (95% CrI: 0.37 to 2.90)	0.52	0.82 (95% CrI: 0.45 to 1.37)	0.19

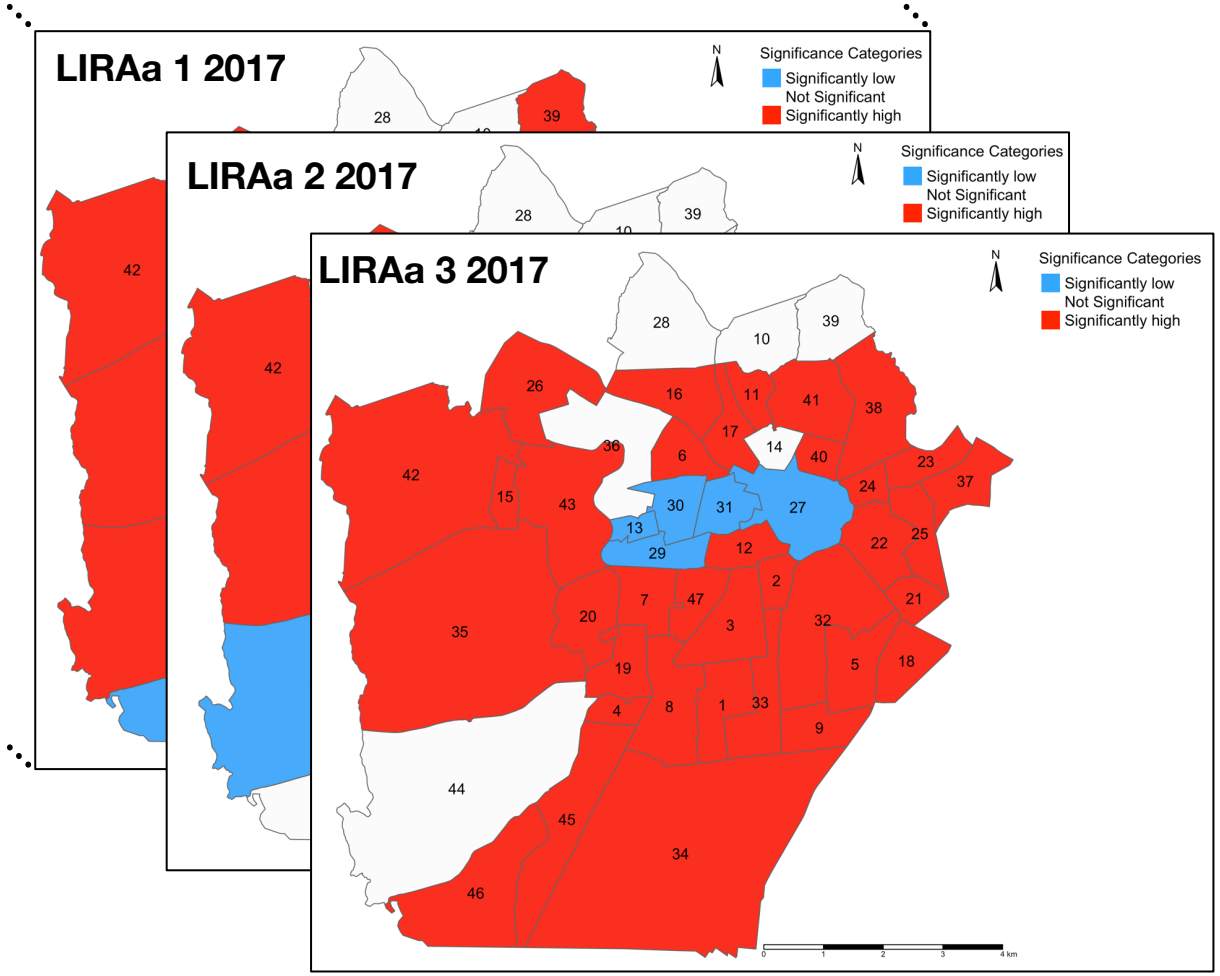
RR: Relative risks
Pr(RR > 1): Exceedance probabilities (the probability that RR being greater than 1)

Preliminary result

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



Maps on the right panel illustrates which neighbourhoods in Campina Grande have RRs that are significantly "low" or "high" risk



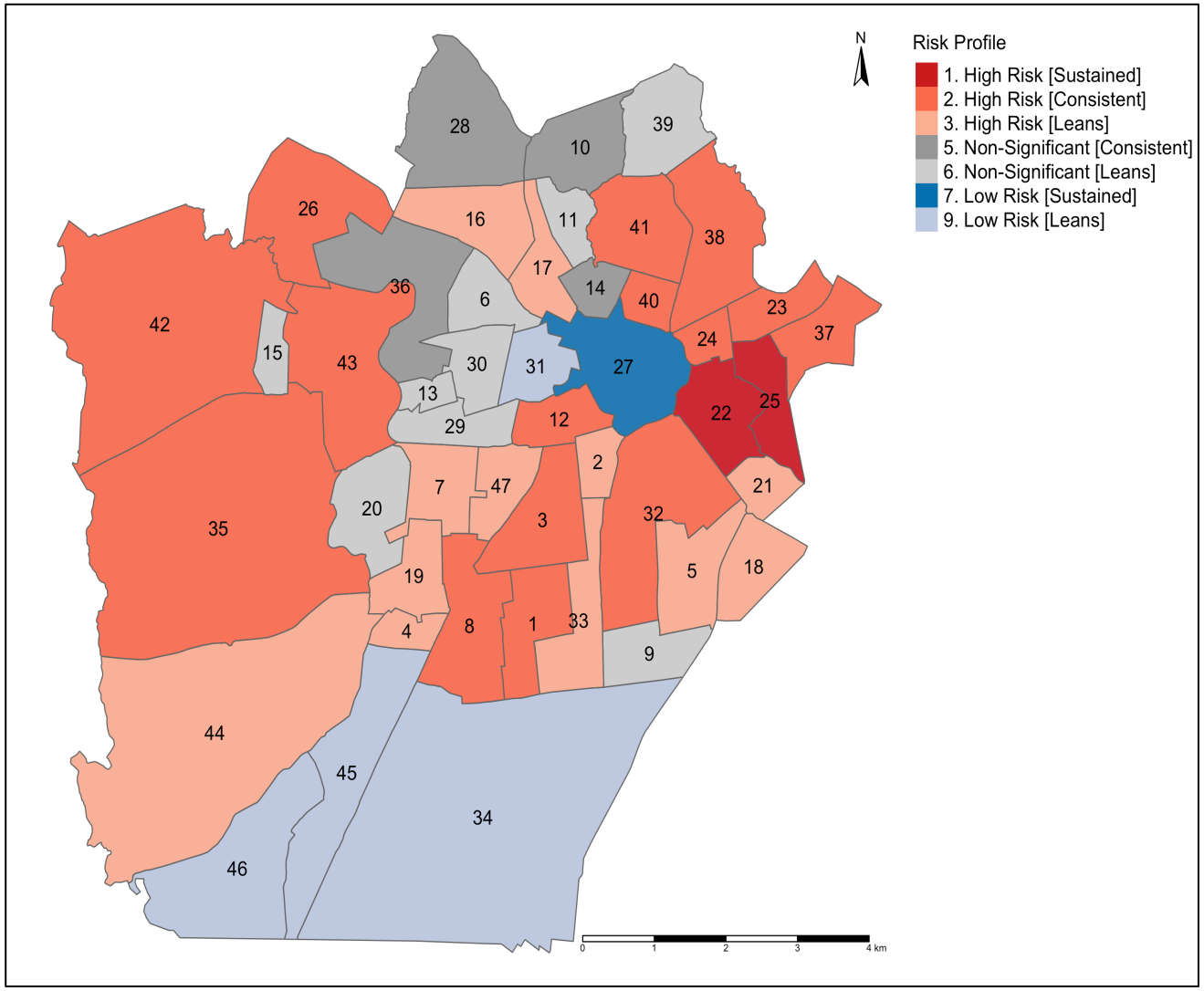
RR < 1.00 (Low risk)
RR > 1.00 (High risk)
RR = 1.00 (Non-significant risk)

Preliminary result

Modelling the risk trajectories and charting them across these LIRAa periods

Profiling the neighbourhoods accordingly to examine where these significant risks of infestation are sustained

ID	Neighbourhood	LIRAa 1 04/16	LIRAa 2 07/16	LIRAa 3 10/16	LIRAa 1 01/17	LIRAa 2 04/17	LIRAa 3 07/17	Profile
1	Jardim Paulistano							2. High Risk [Consistent]
2	Estação Velha							3. High Risk [Leans]
3	Liberdade							2. High Risk [Consistent]
4	Presidente Médici							3. High Risk [Leans]
5	Sandra Cavalcante							3. High Risk [Leans]
6	Monte Santo							6. Non-Significant [Leans]
7	Santa Rosa							3. High Risk [Leans]
8	Cruzeiro							2. High Risk [Consistent]
9	Itararé							6. Non-Significant [Leans]
10	Jardim Continental							5. Non-Significant [Consistent]
11	Louzeiro							6. Non-Significant [Leans]
12	São José							2. High Risk [Consistent]
13	Pedregal							6. Non-Significant [Leans]
14	Conceição							5. Non-Significant [Consistent]
15	Ramadinha							6. Non-Significant [Leans]
16	Jeremias							3. High Risk [Leans]
17	Palmeira							3. High Risk [Leans]
18	Vila Cabral							3. High Risk [Leans]
19	Santa Cruz							3. High Risk [Leans]
20	Dinamérica							6. Non-Significant [Leans]
21	Mirante							3. High Risk [Leans]
22	José Pinheiro							1. High Risk [Sustained]
23	Castelo Branco							2. High Risk [Consistent]
24	Santo Antônio							2. High Risk [Consistent]
25	Monte Castelo							1. High Risk [Sustained]
26	Novo Bodocongó							2. High Risk [Consistent]
27	Centro							7. Low Risk [Sustained]
28	Cuités							5. Non-Significant [Consistent]
29	Centenário							6. Non-Significant [Leans]
30	Bela Vista							6. Non-Significant [Leans]
31	Prata							9. Low Risk [Leans]
32	Catolé							2. High Risk [Consistent]
33	Tambor							3. High Risk [Leans]
34	Velame							9. Low Risk [Leans]
35	Malvinas							2. High Risk [Consistent]
36	Universitário							5. Non-Significant [Consistent]
37	Nova Brasília							2. High Risk [Consistent]
38	Jardim Tavares							2. High Risk [Consistent]
39	Nações							6. Non-Significant [Leans]
40	Lauritzen							2. High Risk [Consistent]
41	Alto Branco							2. High Risk [Consistent]
42	Serrotão							2. High Risk [Consistent]
43	Bodocongó							2. High Risk [Consistent]
44	Três Irmãs							3. High Risk [Leans]
45	Acácio Figueiredo							9. Low Risk [Leans]
46	Cidades							9. Low Risk [Leans]
47	Quarenta							3. High Risk [Leans]



Preliminary result

Any questions?