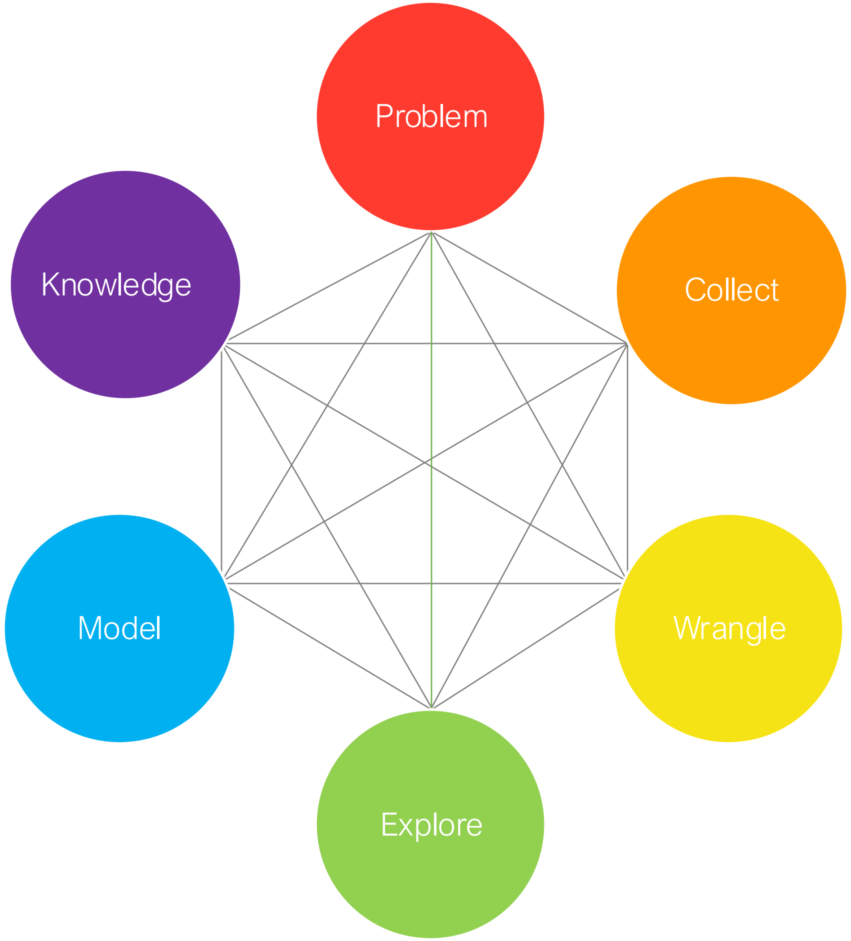
GEOG0114: PRINCIPLES OF SPATIAL ANALYSIS

WEEK 9: SPATIAL MODELS (PART 3)

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

Contents

- Introduction to Spatial Risk Models for Areal Data
- 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 & spatiotemporal data)
- Model formulation from a Bayesian Framework
- Examples and interpretation



Recap

Spatial Models

Week 7

Spatial Lag and Error Models

Week 8

Geographically Weighted Regression

Week 9

Spatial Risk Models

What are spatial risk models

Definition:

A **Spatial Risk Model** estimates the probability or intensity of an event (e.g., disease, infestation, flooding, crime) at different geographic locations. These are probabilistic models which are operationalised to quantify how risk varies across space, and time.

The models we have been using previous allows us to quantify relationships globally, or locally, and it was conducted within **Frequentist Approach** – where we derive an estimate for single values.

Spatial Risk Models are different – they are performed within a **Bayesian framework**, which relies heavily on **probabilistic approach,** accounting spatial dependence and heterogeneity, with **the main goal to map and understand the geographic variation in risk of an outcome and identify the drivers of that variation in risk!**

Spatial risk models fall under the family of mainly Poisson & Logistic regression models.

Packages: RStan, INLA, JAGS, WinBUGS

Here is a board overview:

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc.,	**Linear regression**
Binary measures (1 = "present" or 0 = "absent"): e.g., Person's voting for a candidate, Lung cancer risk, house infested with rodents etc.,	**Logistic Regression**
Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with rodents, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc.,	**Logistic Regression**
Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc.,	**Poisson Regression**
Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc.,	**Survival Analysis with Cox regression**

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 the 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 geographic areas where data is available. We can use spatial risk 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

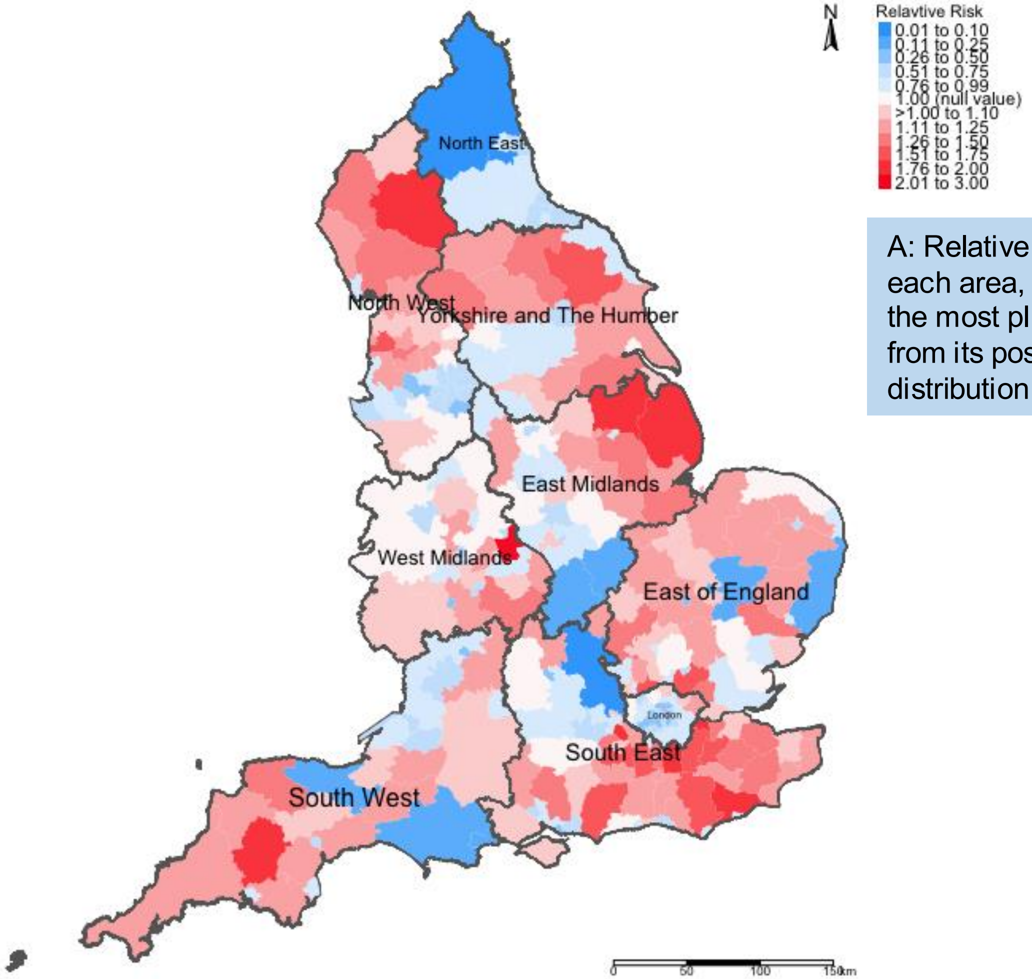
Exceedance Probabilities (or Marginals) is a statistical measure describing the probability that an estimated risk value for an areal-unit exceeds a given threshold.

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 a lot to detect clusters of areas with exceedingly high risk of a disease events.

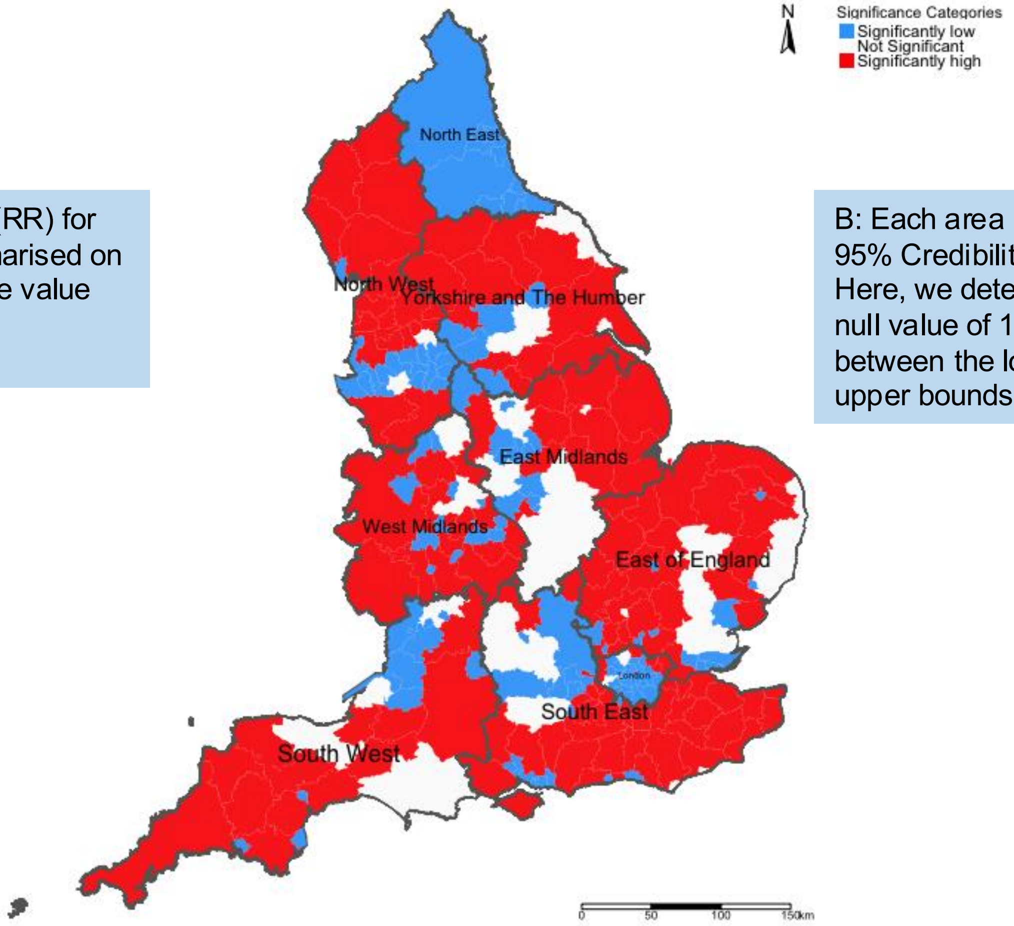
Example: Risks of Road-related casualties in England 2015-2020 [1]

Outcome: Aggregate counts of incident road accidents at Local Authorities



A: Relative Risk (RR) for each area, summarised on the most plausible value from its posterior distribution

Relative Risks (RR)



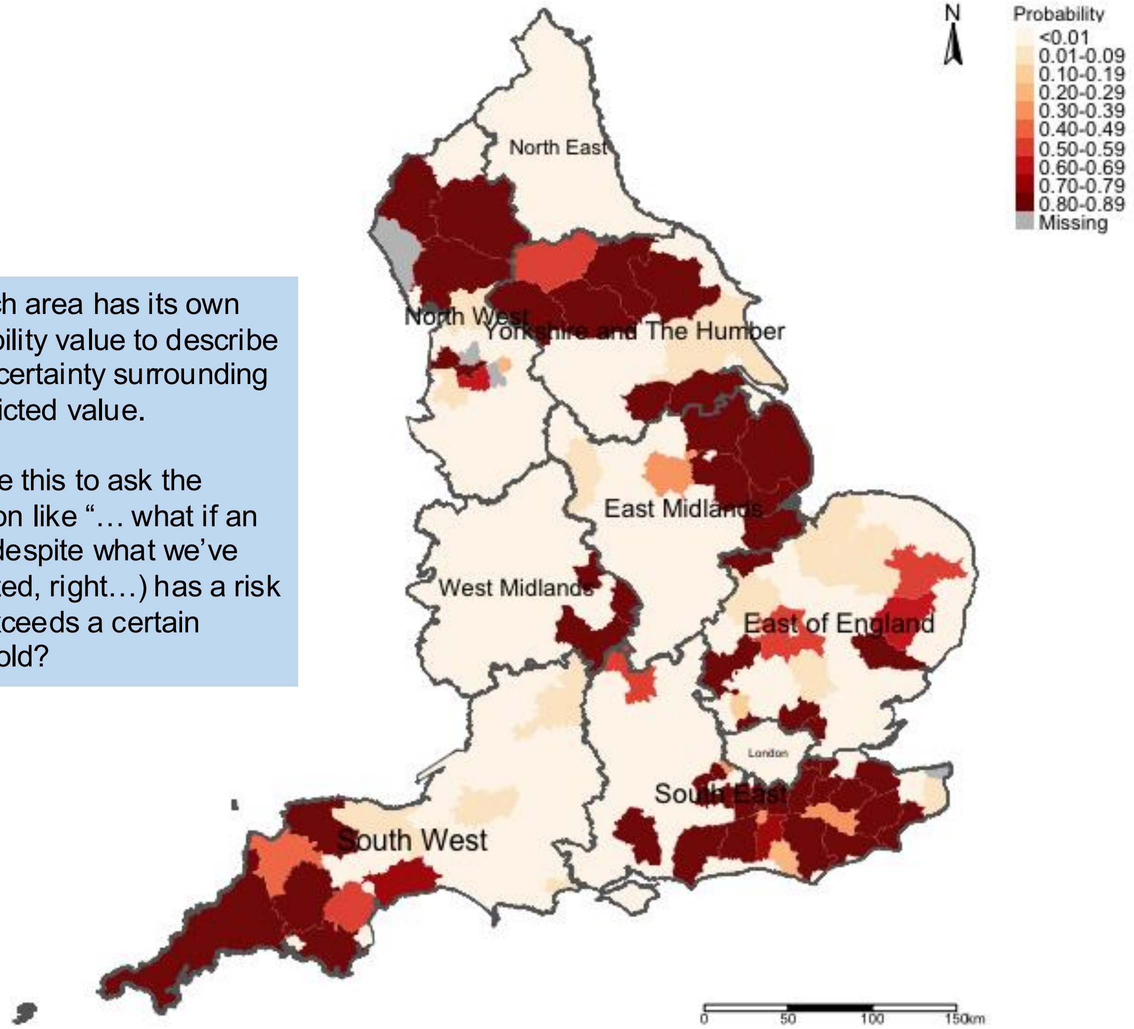
B: Each area has its own 95% Credibility Intervals. Here, we determine if the null value of 1 does not lie between the lower and upper bounds.

Overall significance (95% Credibility Intervals)

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

C: Each area has its own probability value to describe the uncertainty surrounding a predicted value.

We use this to ask the question like "… what if an area (despite what we've predicted, right…) has a risk that exceeds a certain threshold?



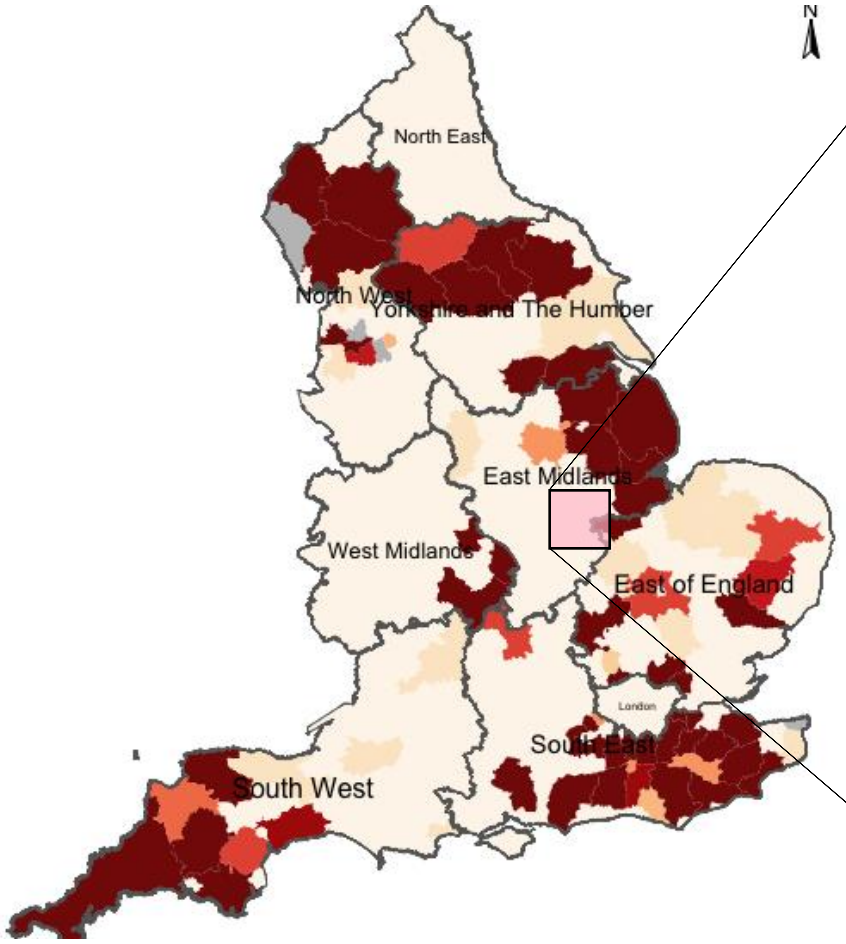
Don't you think a 40% increased risk of road accidents in an area is too high?

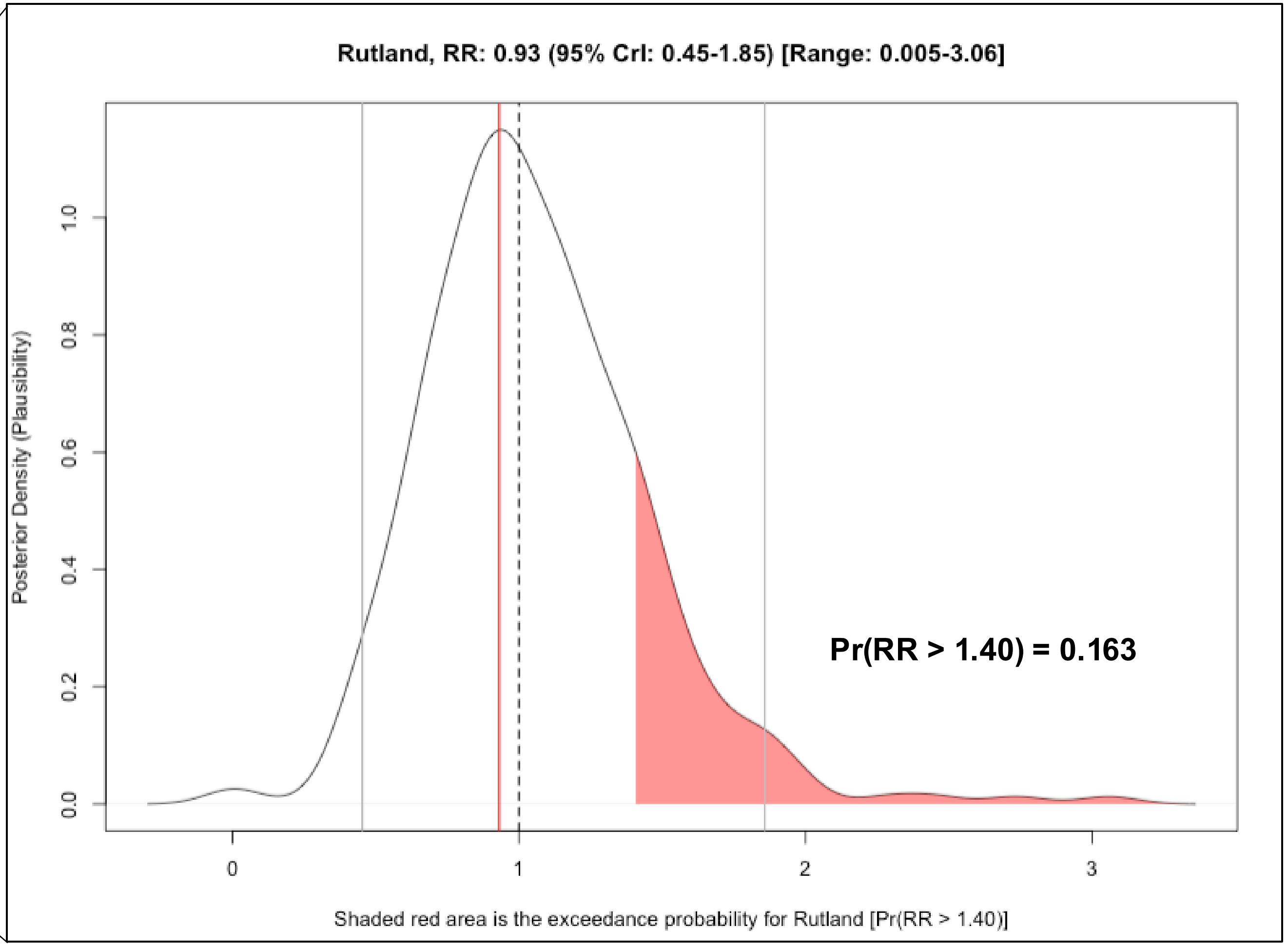
The areas in darker reds are perhaps priority areas for some road safety policy should be implemented.

Here, we have operationalised the concept of exceedance probabilities to compute the probability of an area having a relative risk that is 1.40 or greater than 1.40.

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

Example: Risks of Road-related casualties in England 2015-2020 [3]





Geographical distribution the exceedance probability for the RR being above 1.40

This graph illustrates the predicted posterior distribution for the relative risk of road accidents in Rutland. The most plausible risk value is 0.93 (95% CrI: 0.45-1.85) [Range: 0.005-3.06]. With exceedance probabilities, we want to know what is the probability that a set of estimates lie above a certain threshold. The exceedance probability for Rutland to have a relative risk of road accidents at/above 1.40 is 16.3%

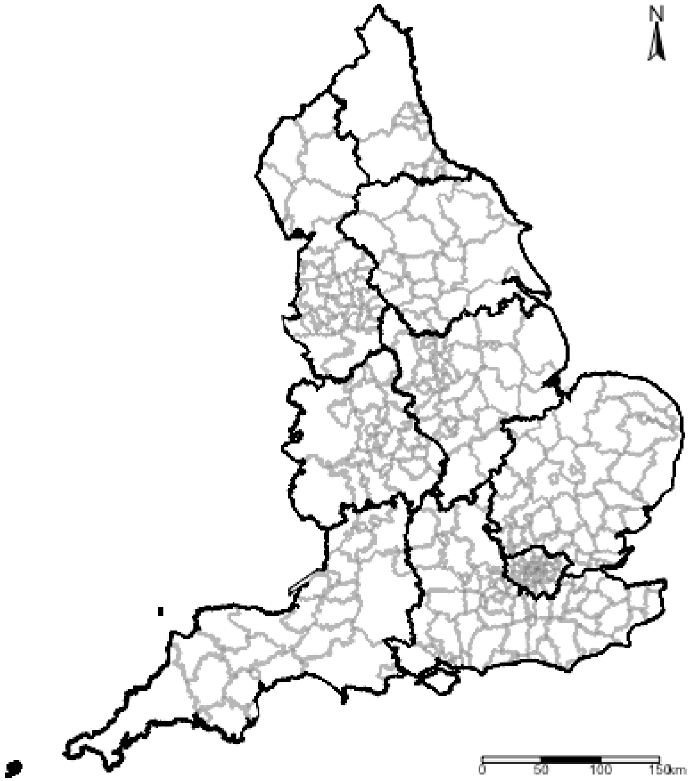
Spatial Conditional Auto-regressive models (CARs/iCARs)

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

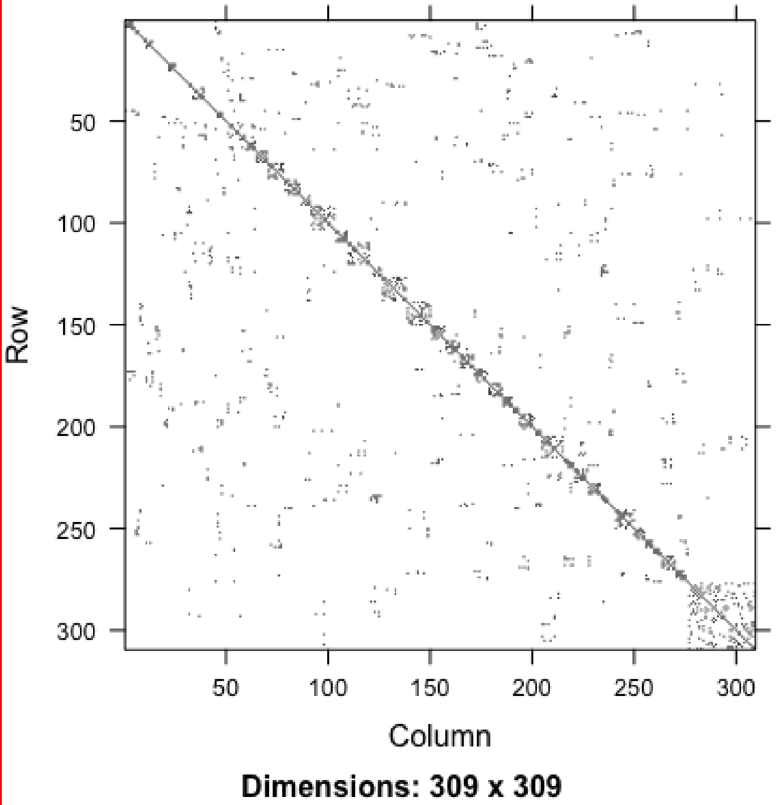
This is a popular spatial model which takes into account that the data may potentially be spatially correlated and the observations in the neighbouring areas may be more similar than observations in areas that are distant from each other.

- This is a type of spatial risk 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 ONLY 'fully specified' spatial effect – Conditional Auto-Regressive model (CAR)
  - ❖ BYM model that has BOTH a spatial effect term which is treated a **structured random effect**, and the non-spatial effect term which is treated as an **unstructured random effects** – Intrinsic Conditional Auto-Regressive Model (ICAR). This is the best option.
- 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

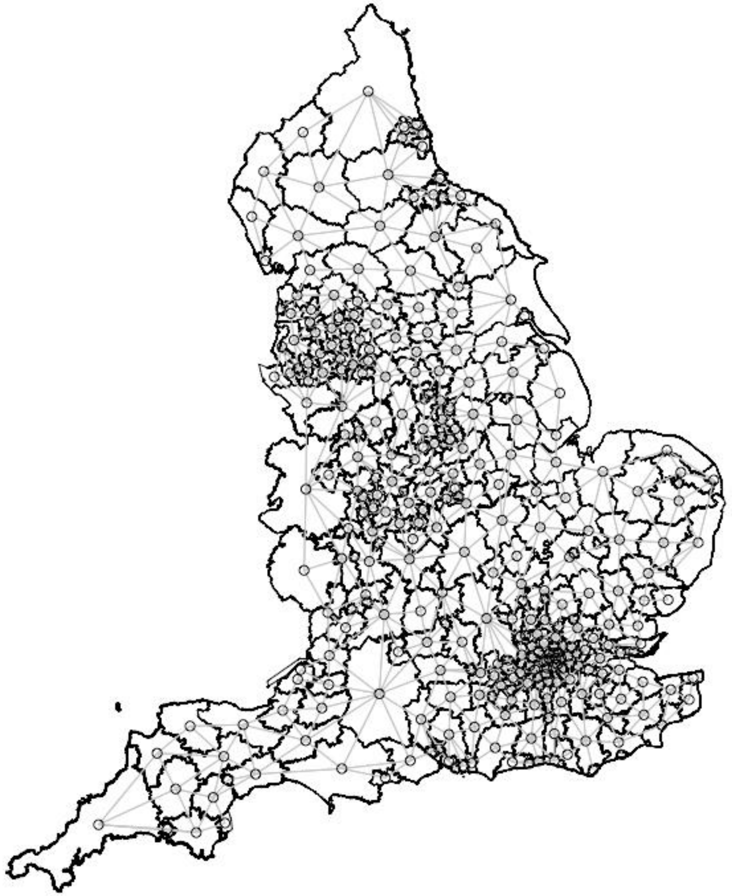


Adjacency matrix translated to graph format



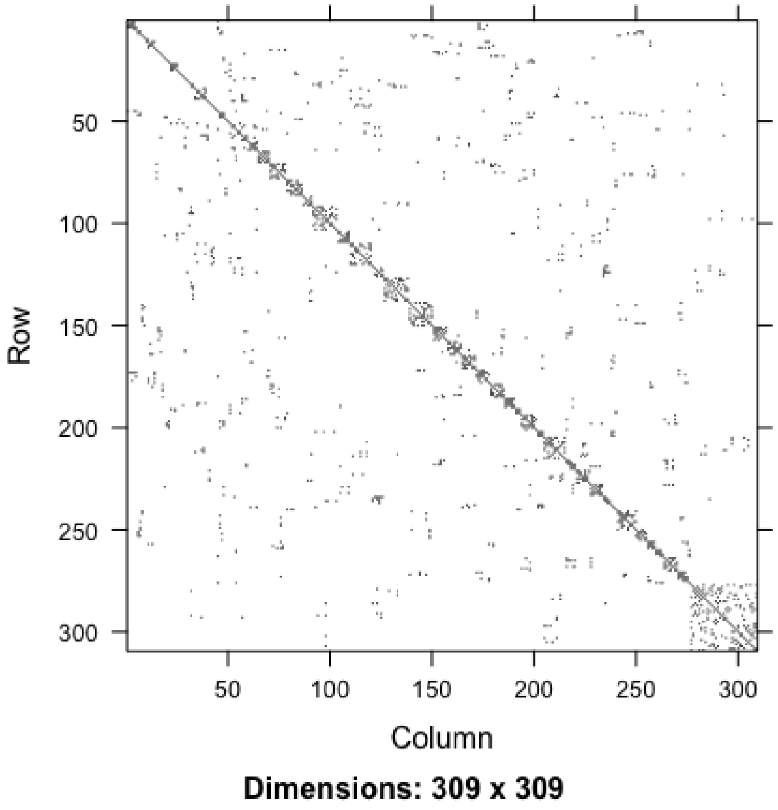
INLA uses adjacency matrix (Queen contiguity-based).

Adjacency matrix translated to nodes and edge format



Stan only uses the nodes and edges format to reconstruct the adjacency matrix (Queen contiguity-based).

Spatial structured and unstructured random effects [1]



Adjacency matrix translated in graph 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 $\phi + \theta$

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 $\sigma(\sqrt{(1-\rho)} * \theta + \sqrt{(\rho)} * \phi)$

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 often counts – thus it will be Poisson (with log as the link function).

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

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

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

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

- Build Bayesian model

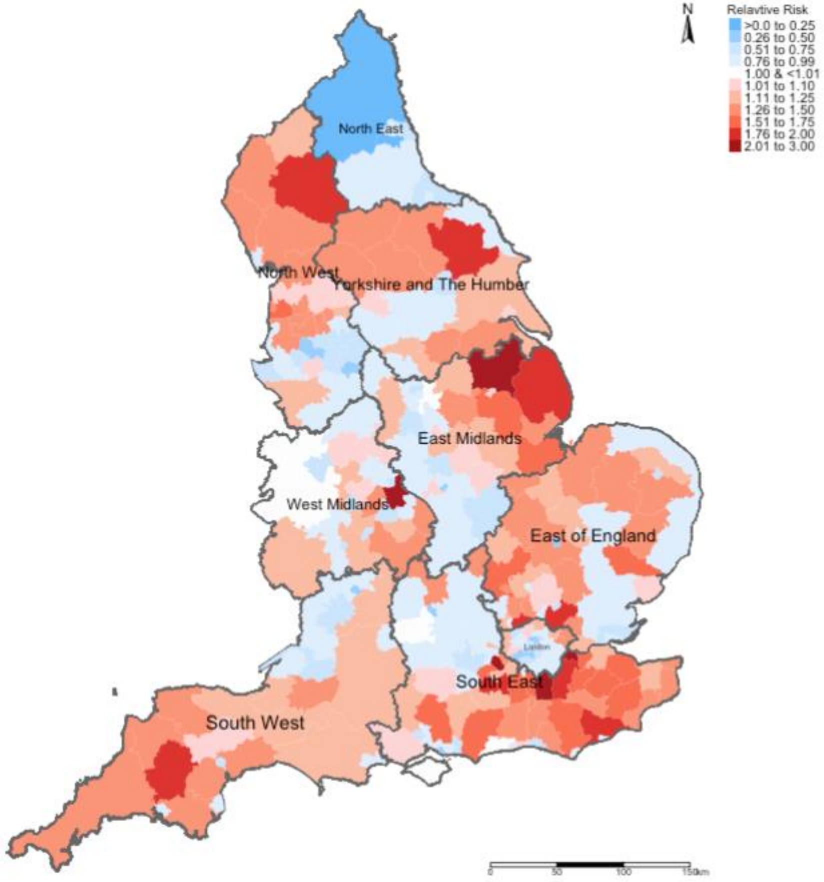
Recall the Bayes' Rule: $P(\theta|Y) \propto P(Y|\theta)P(\theta)$

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

Do not worry about this slide – its to show what's underneath the bonnet of vehicle!!!

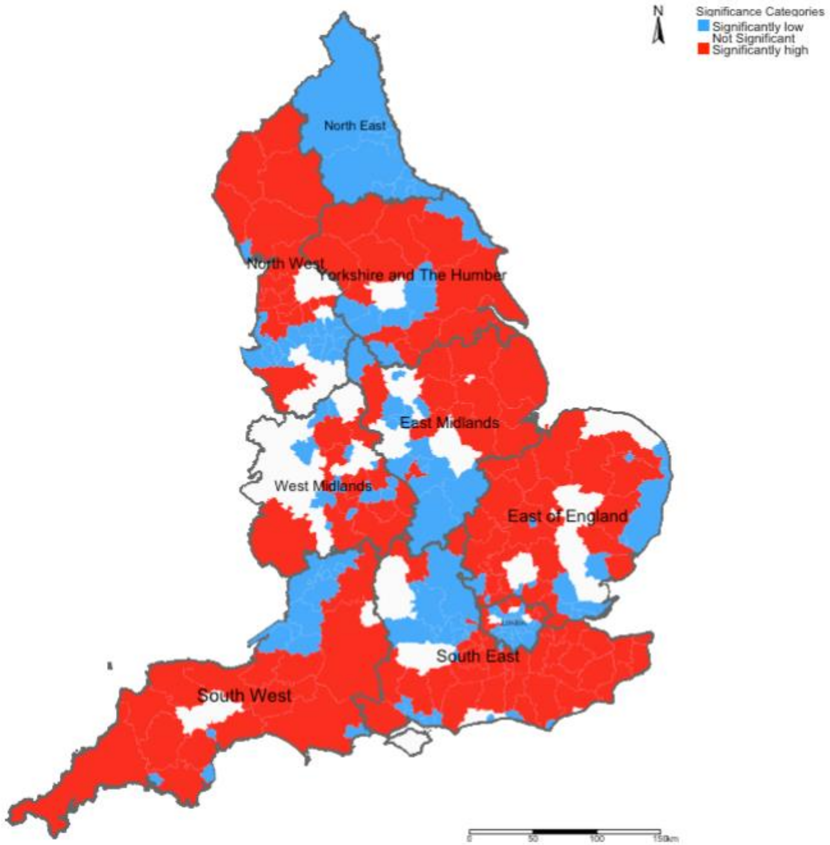
Expected output [1]

Relative risk ratios (RR)



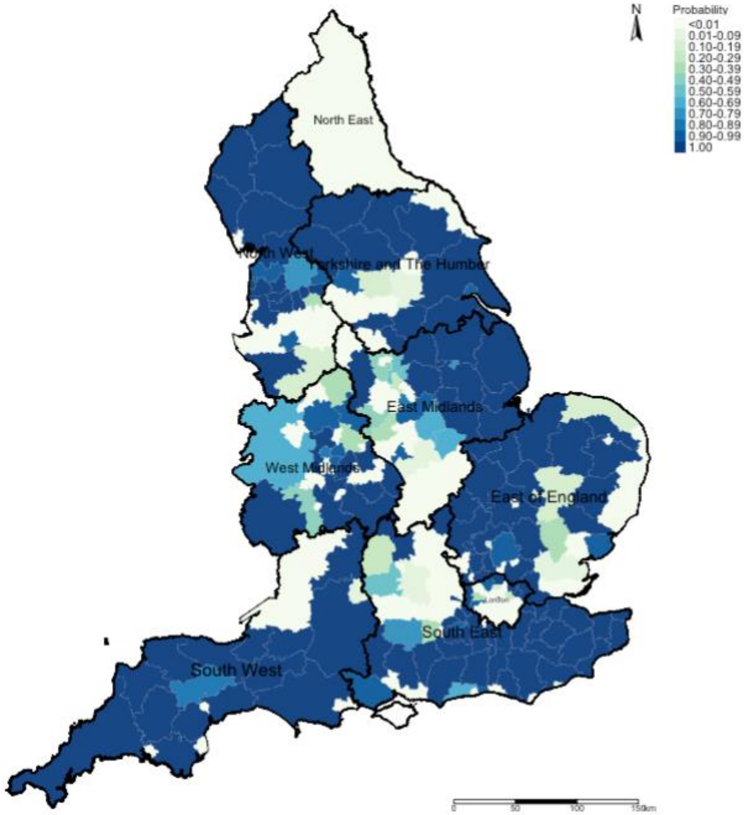
Here, we use this output to describe the burden of an outcome

Statistical Significance



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

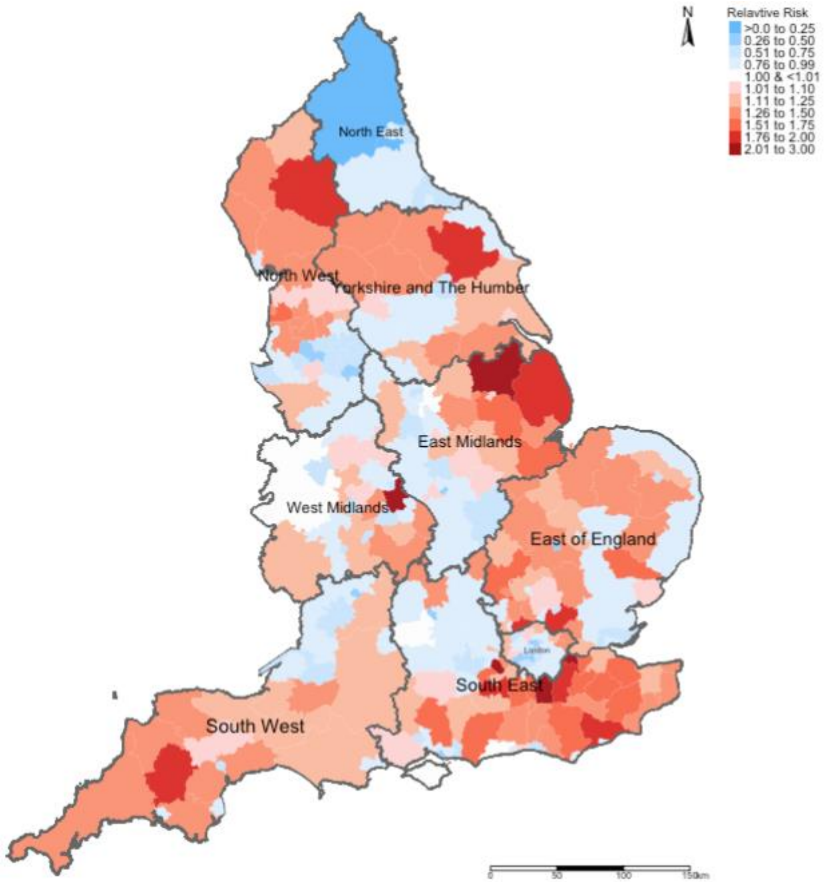
Exceedance Probabilities



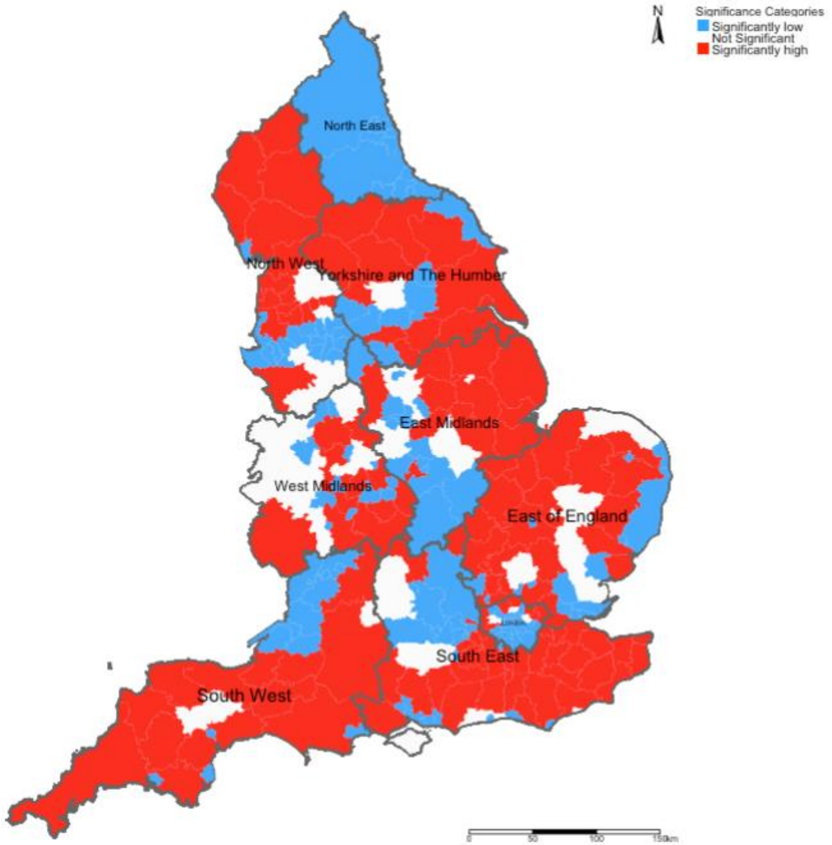
This output is used to describe the uncertainty that surrounds the risks we found in the first map when we explore P(RR>1.00)

Expected output [2]

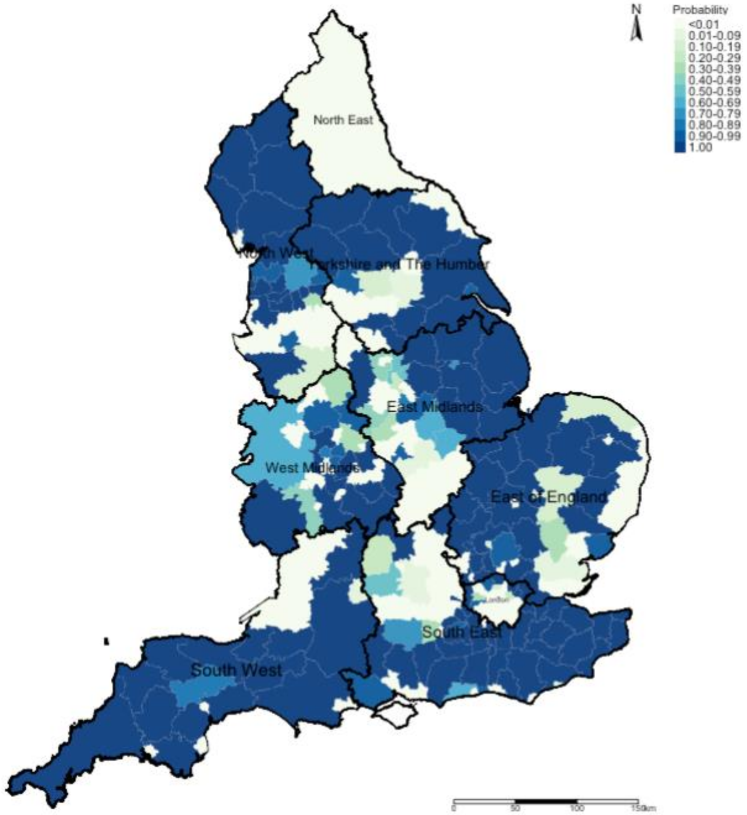
Relative risk ratios (RR)



Statistical Significance

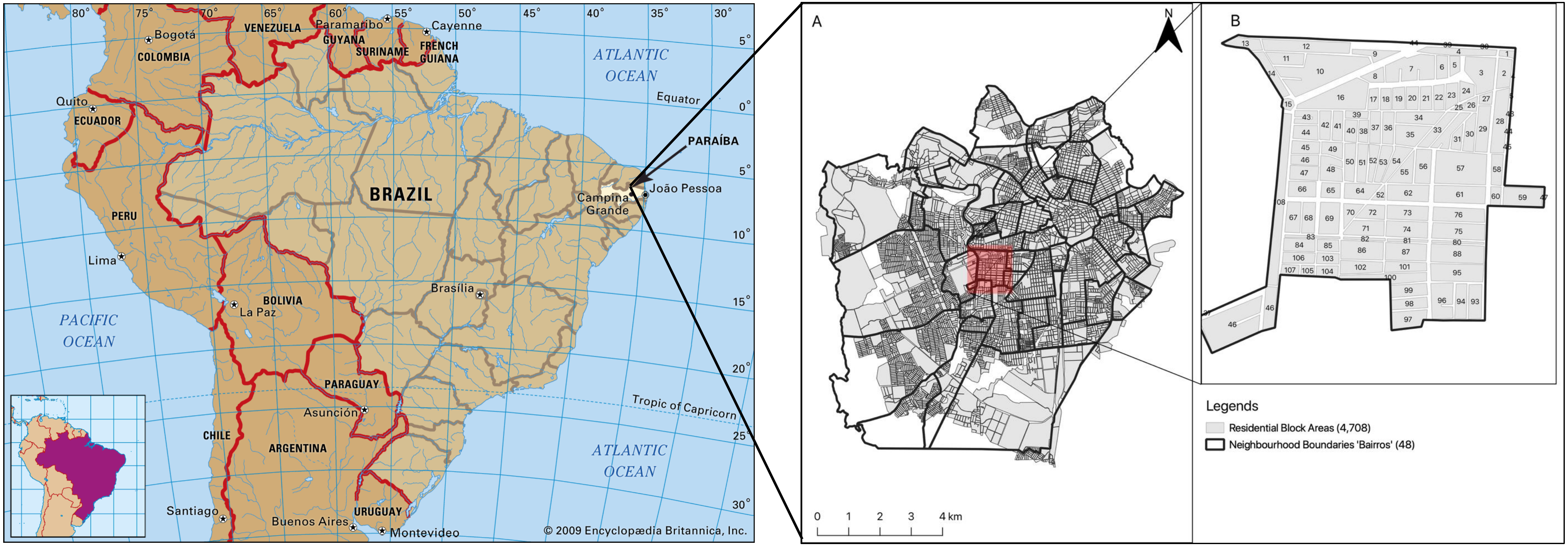


Exceedance Probabilities



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 Southwest 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 Southwest area.

Application 1: Risk assessment and mapping of infestation in Campina Grande (cross-sectional)



- 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.

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			

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.

Cross-sectional (single time-point scenario)

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \alpha + X_{k,i}\beta_k + \log(E_i) + \sigma\, C_i$$

$$C_i = \left(\sqrt{1-\rho_s}\right)\theta_i + \left(\sqrt{\rho_s}\right)\phi_i$$

combined = non-spatial + spatial

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.

Y_i is observed counts of houses with mosquito infestation in neighbourhood i
E_i is expected number of houses with mosquito infestation in neighbourhood i
α is the overall baseline risk of mosquito infestation in the whole of the study area
$X_{k,i}$ list of independent variables, where k = 4 (temperature, precipitation, NDVI and urbanisation)
β_k list of coefficients for our independent variables, where k = 4 (temperature, precipitation, NDVI and urbanisation)
σ global variation or standard deviation levels of infestation for the whole of the study area
C_i represents the combined effects from non-spatial (local) and spatial effects in neighbourhood i
θ_i estimated unstructured non-spatial (or local) effect for each neighbourhood
ϕ_i estimated structured spatial effect for each neighbourhood
ρ_s proportion (0 to 1) for how much we want to partition our model i.e., non-spatial versus spatial structure

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

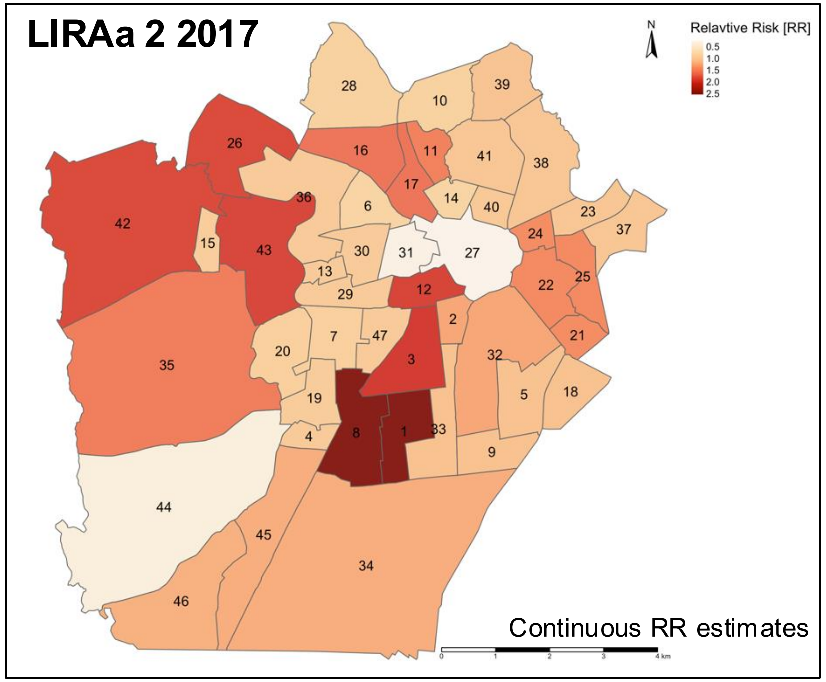
RR: Relative risks; Pr(RR > 1): Exceedance probabilities (the probability that RR being greater than 1)

Interpretation (examples):

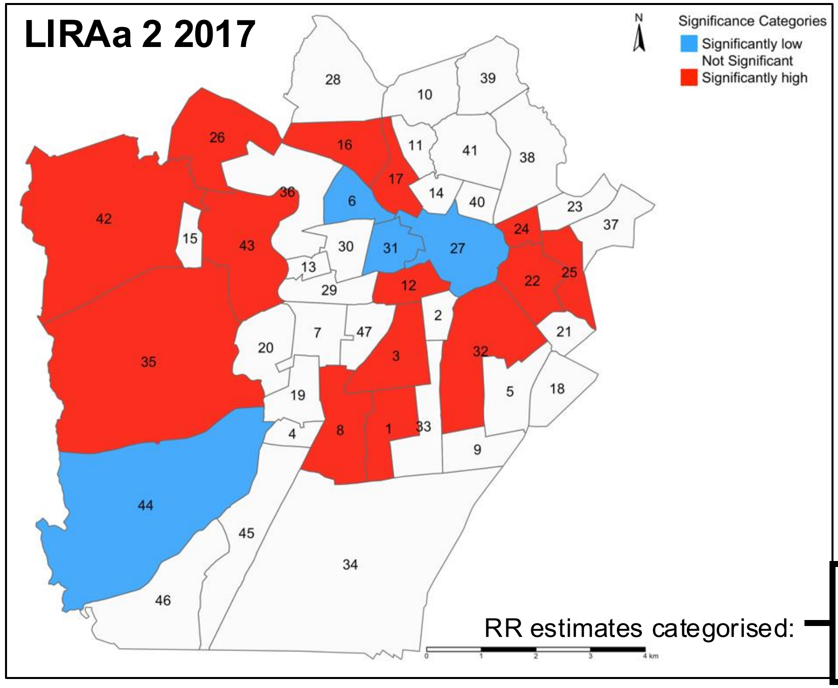
- **Intercept**: The overall baseline risk of mosquito-borne infestation is 1.64 times (or 64%) **greater** in the population of Campina Grande. The overall probability that there's excess risk of infestation (i.e., RR > 1.00) is 51%.
- **Temperature**: In relation to temperature, the risk of mosquito-borne 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



PART 3: 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)

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!

Application 2: Risk assessment fire-related casualties in England 2010-19 (Spatiotemporal)

Longitudinal (spatiotemporal scenario)

$$Y_{i,t} \sim \text{Poisson}(\lambda_{i,t})$$

$$\log(\lambda_{i,t}) = \alpha + X_{k,i,t}\beta_{k,t} + \log(E_{i,t}) + \sigma\, C_{i,t}$$

$$C_{i,t} = \left(\sqrt{1-\rho_s}\right)\theta_{i,t} + \left(\sqrt{\rho_s}\right)\phi_{i,t} + \left(\sqrt{\rho_t}\right)\tau_t$$

combined = non-spatial + spatial + temporal

Notes:
- $\exp(\alpha)$ is the overall risk ratio for study area
- $\exp(\beta)$ is the overall risk ratio for coefficient
- $\exp(\alpha + \sum \beta_{k,t} X_{i,k,t} + C_{i,t}\sigma)$ by adding $+C_{i,t}\sigma$ to the α allows the risks to vary for each area at a given time. By adding $+\sum \beta_{k,t} X_{i,k,t}$ you are also adjusting for the effects of each variables

$Y_{i,t}$ is observed counts of fire-related casualties in an FSA (Fire Service Area) i in England
$E_{i,t}$ is expected number of fire-related fire-related casualties in an FSA i in England
α is the overall baseline risk of fire-related fire-related casualties throughout England
$X_{k,i,t}$ list of independent variables, where k = 4 (Living Environment, Crime, Housing and Barriers and Education)
$\beta_{k,t}$ list of coefficients for our independent variables, where k = 4 (Living Environment, Crime, Housing and Barriers and Education)
σ global variation or standard deviation levels of fire-related casualties in England
$C_{i,t}$ represents the combined effects from non-spatial (local) and spatial effects in FSA i in England at a given time point t
$\theta_{i,t}$ estimated unstructured non-spatial (or local) effect for each FSA at a given time point t
$\phi_{i,t}$ estimated structured spatial effect for each FSA at a given time point t
ρ_s proportion (0 to 1) for how much we want to partition our model i.e., non-spatial versus spatial structure
ρ_t proportion (0 to 1) of temporal variation (scale value)

Table results illustrates the overall association between socioeconomic deprivation factors and risk of fire-related casualties in England (2010-19).

Multivariate spatiotemporal Bayesian regression models that explores the overall association with deprivation indexes with dwelling fire-related SCRs and random effects.

Domains of Deprivation in England	Adjusted Relative Risk (95% Credibility Intervals)		
	Estimates	Percentage	
Living Environment	1.241 (1.164–1.329)	+24.1%	(16.4%–32.9%)
Education, Skills & Training Deprivation	1.181 (1.124–1.245)	+18.1%	(12.4%–24.5%)
Housing & Barriers to Public Services	1.137 (1.094–1.184)	+13.7%	(9.4%–18.4%)
Crime	1.010 (1.007–1.013)	+1.0%	(0.7%–1.3%)
Random effects	**Median (95% Credibility Intervals)**		
τ^2: estimate of temporally varying spatial variation	0.035 (0.028–0.044)		
ρ_S: estimate of spatial autocorrelation	0.014 (0.001 to o.065)		
ρ_T: estimate of temporal autocorrelation	0.936 (0.885–0.986)		

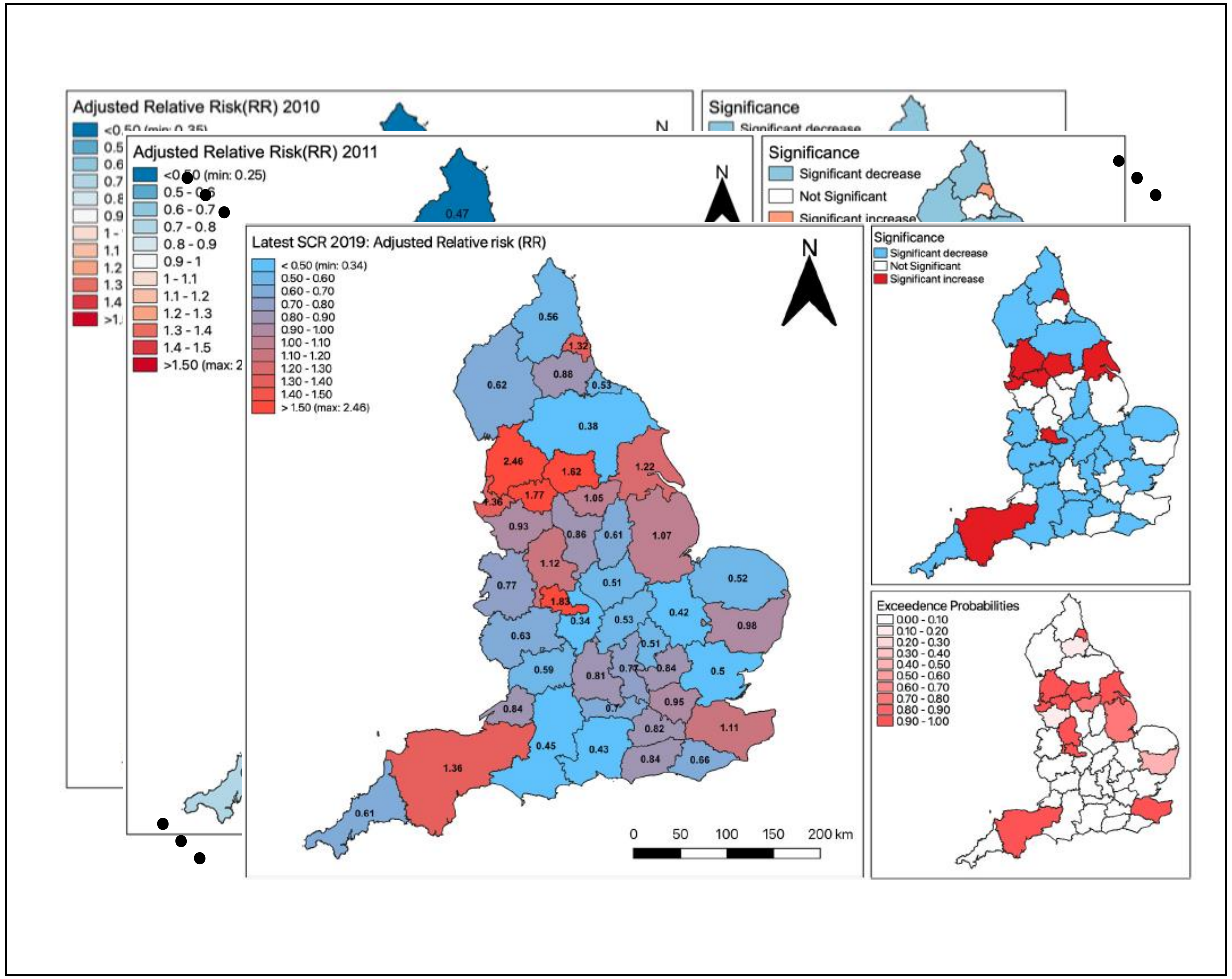
Source: Li et al, (2022), Ecological study exploring the geospatial associations between socioeconomic deprivation and fire-related dwelling casualties in England (2010-2019), DOI: https://doi.org/10.1016/j.apgeog.2022.102718

Interpretation:

- **Living environment**: In the context of this socioeconomic deprivation domain – **increased levels significantly increases the risk of fire-related casualties by 24.1% (1.241 times higher).**
- **Barriers to housing and public service**: In the context of this socioeconomic deprivation domain – **increased levels significantly increases the risk of fire-related casualties by 13.7% (1.137 times higher).**
- **Education, Skills & Training**: In the context of this socioeconomic deprivation domain – **increased levels significantly increases the risk of fire-related casualties by 18.1% (1.181 times higher).**
- **Crime**: In the context of this socioeconomic deprivation domain – **increased levels marginally increases the risk of fire-related casualties by 1.0% (1.01 times higher).** While this is statistically significant, the significance can be seen as something that is marginal.

NOTE: The random effects are reported in this instance to show the variation in risk. Most of it is captured in the time component (0.936 = 93.6%) meaning that there is a very strong temporal autocorrelation in the data.

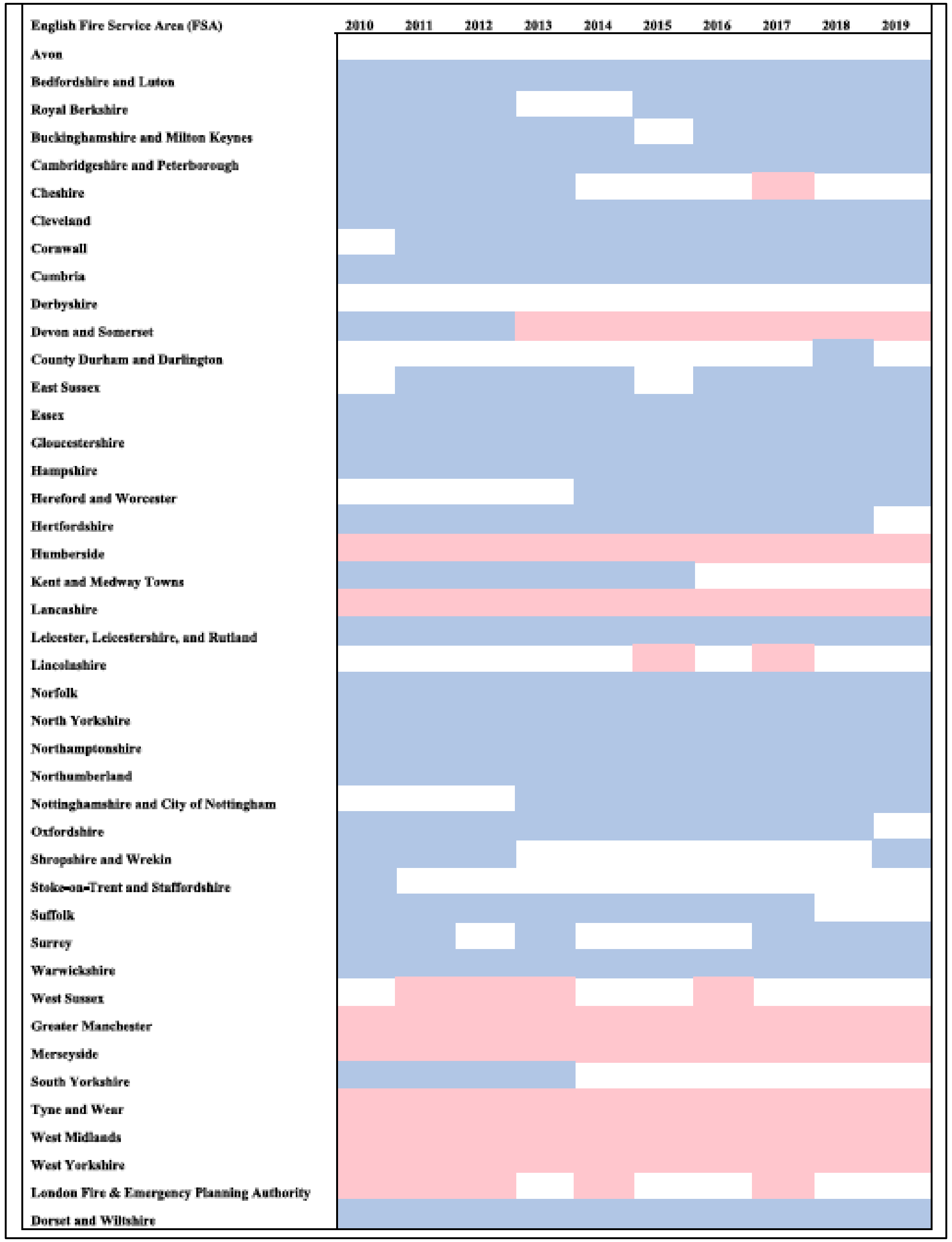
Spatial and temporal variation in risk of fire-related casualties



Interpretation: On a year-on-year basis, the increased risks are significantly sustained throughout the 10-year period for Humberside, Lancaster, Greater Manchester, Merseyside, Tyne and Wear, West Midlands, West Yorkshire, and Devon & Somerset.

>>>

Translate maps to show significant risk trajectories



Source: Li et al, (2022), Ecological study exploring the geospatial associations between socioeconomic deprivation and fire-related dwelling casualties in England (2010-2019). DOI: https://doi.org/10.1016/j.apgeog.2022.102718

Any questions?