What is spatial data?

Geostatistics

Regression modelling

Spatial correlation

Spatial models as GAMs

Spatial modelling with GAMs in R

Summary

References

Modelling Areal Data

ISAIR

Ruoran Li

2025-01-28

What is spatial data?

Observed variables with spatial coordinates associated with them whose values may vary according to some unobserved spatial process.
Three main classes of spatial data:

- **Areal data** - data within a polygon
- **Point pattern data** - events observed at a given point
- **Geostatistical data** - discrete measurements of a phenomenon that occurs over continuous space

There are specific statistical analyses for each of these types of spatial data.

Geostatistical data

Data that vary continuously over space, but are measured only at discrete locations.

Consist of pairs of data $(Y_i, \boldsymbol{s}_i)$, where

- Y_i is the value observed/measured at a fixed location $\boldsymbol{s}_i$.
- $\boldsymbol{s}_i$ a vector of coordinates in a coordinate system, e.g.
  - longitude and latitude
  - UTM coordinate reference system (metres east and north of a reference point)

Examples could be rainfall, temperature, soil characteristics.

Example

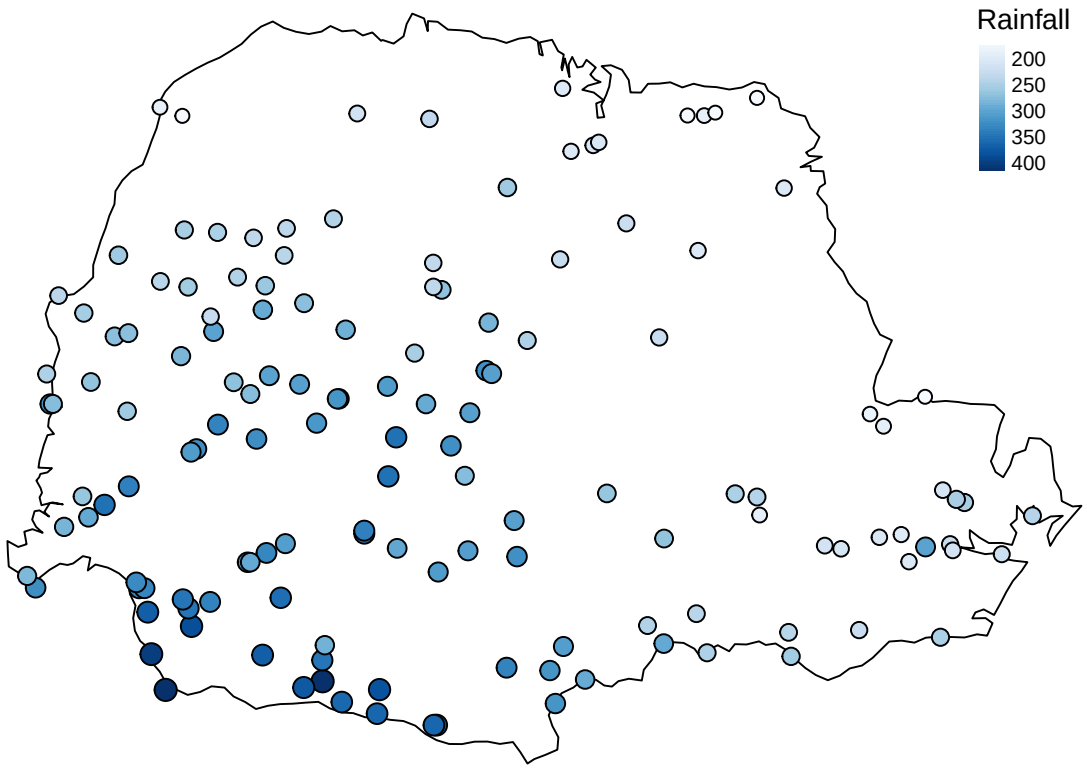


Figure 1: Average rainfall over different years for the period May-June collected at 143 recording stations throughout Parana state, Brasil.

Point pattern data

Consist of a set of locations s_i of objects/events occurring in a study region.

- More concerned with the presence/absence of an event rather than value of a measurement at a point

- e.g trees in a forest, animal nets, crimes, domiciles of new cases of a certain disease

A point pattern may also be:

- marked (e.g. which species of tree), or

- unmarked (just that a tree exists there)

Example

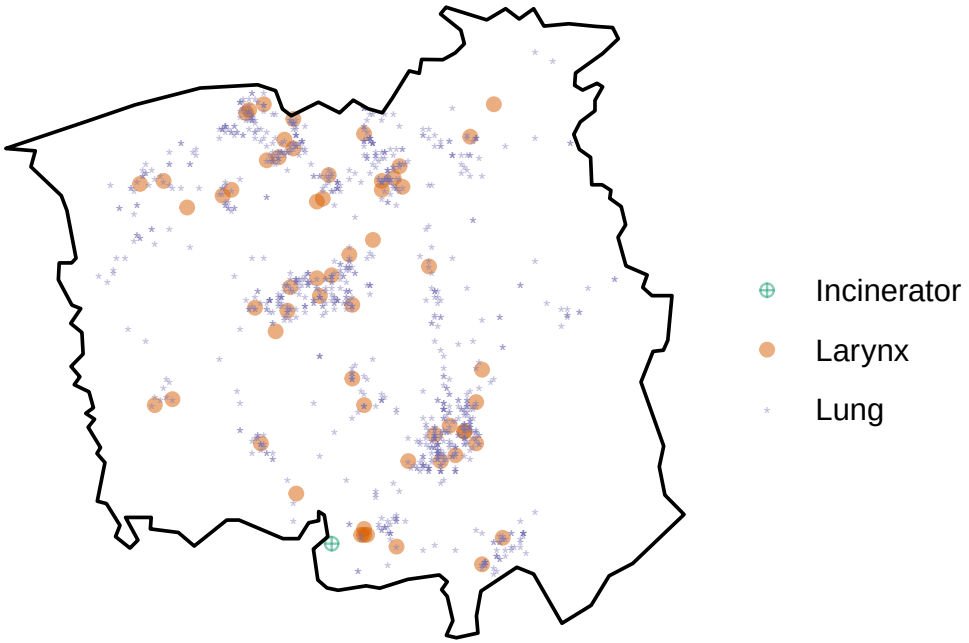


Figure 2: Chorley-Ribble data. Spatial locations of cases of cancer of the larynx and the lung and a disused industrial incinerator.

Areal data

Similar to geostatistical data, but are aggregated to an area that can have a **regular** (e.g. a square grid) or **irregular** shape (e.g. census districts).
Consist of pairs of data (Y_i, A_i), where

- Y_i is the value registered for the area A_i.
- A_i is a specific area (polygon) in our study region.

Examples include vaccine coverage per NHS Trust catchment, number of disease cases per county, national birth rates in a global study.

Modelling Areal Data

Ruoran Li

What is spatial data?

Geostatistics

Regression modelling

Spatial correlation

Spatial models as GAMs

Spatial modelling with GAMs in R

Summary

References

Example

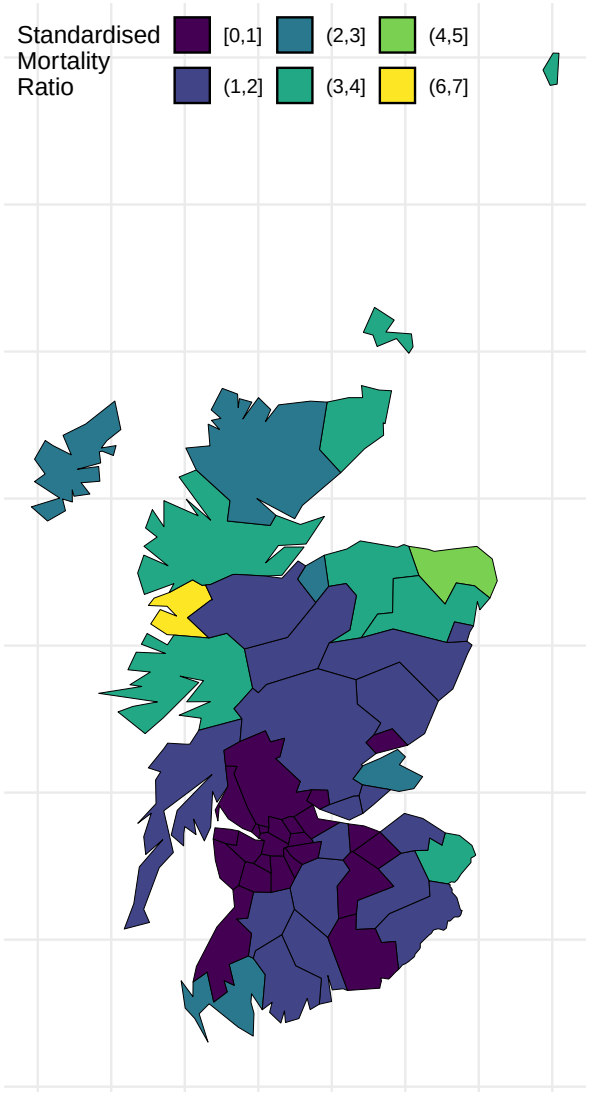


Figure 3: Standardised mortality ratio of lip cancer cases for the period 1975 to 1986, for 56 districts in Scotland.

Past vs. Present

Traditionally, a self-contained methodology for spatial prediction, developed at Ecole des Mines, Fontainebleau, France (Matheron 1963).

Nowadays, that part of spatial statistics which is concerned with data obtained by (spatial) sampling of continuous spatial phenomena (Banerjee, Carlin, and Gelfand 2014).

Model-based Geostatistics

The application of general principles of statistical modelling and inference to geostatistical problems (P. J. Diggle, Tawn, and Moyeed 1998)

- formulate a model for the data;
- use likelihood-based methods of inference;
- answer the scientific question.

Independence

- One of the main assumptions of Generalised Linear Models is that the observations y_i are mutually independent.

- The value of one observation does not give me any information about the value of another observation.

$$\text{Covariance}(y_i, y_j) \equiv 0$$

Do you think this assumption is still valid for spatial data?

Independence

First law of geography: close things are more related than distant things (Tobler 1970)

Two main consequences for violating this assumption:

1. Narrower confidence intervals for the regression parameters that leads to an increase in Type I error.

2. We don't exploit the spatial correlation when doing predictions.

    - The **effective sample size** for correlated data will be smaller than the actual sample size

Statistical workflow

1. Exploratory analysis.
2. Model formulation.
3. Model fitting.
4. Model validation. Evidence against the assumptions?

    - Yes: go back to point 2.
    - No: you can generate predictions and visualise uncertainty.

See also chapter 7 of Peter J. Diggle and Chetwynd (2011)

What is spatial data?

Geostatistics

Regression modelling

Spatial correlation

Spatial models as GAMs

Spatial modelling with GAMs in R

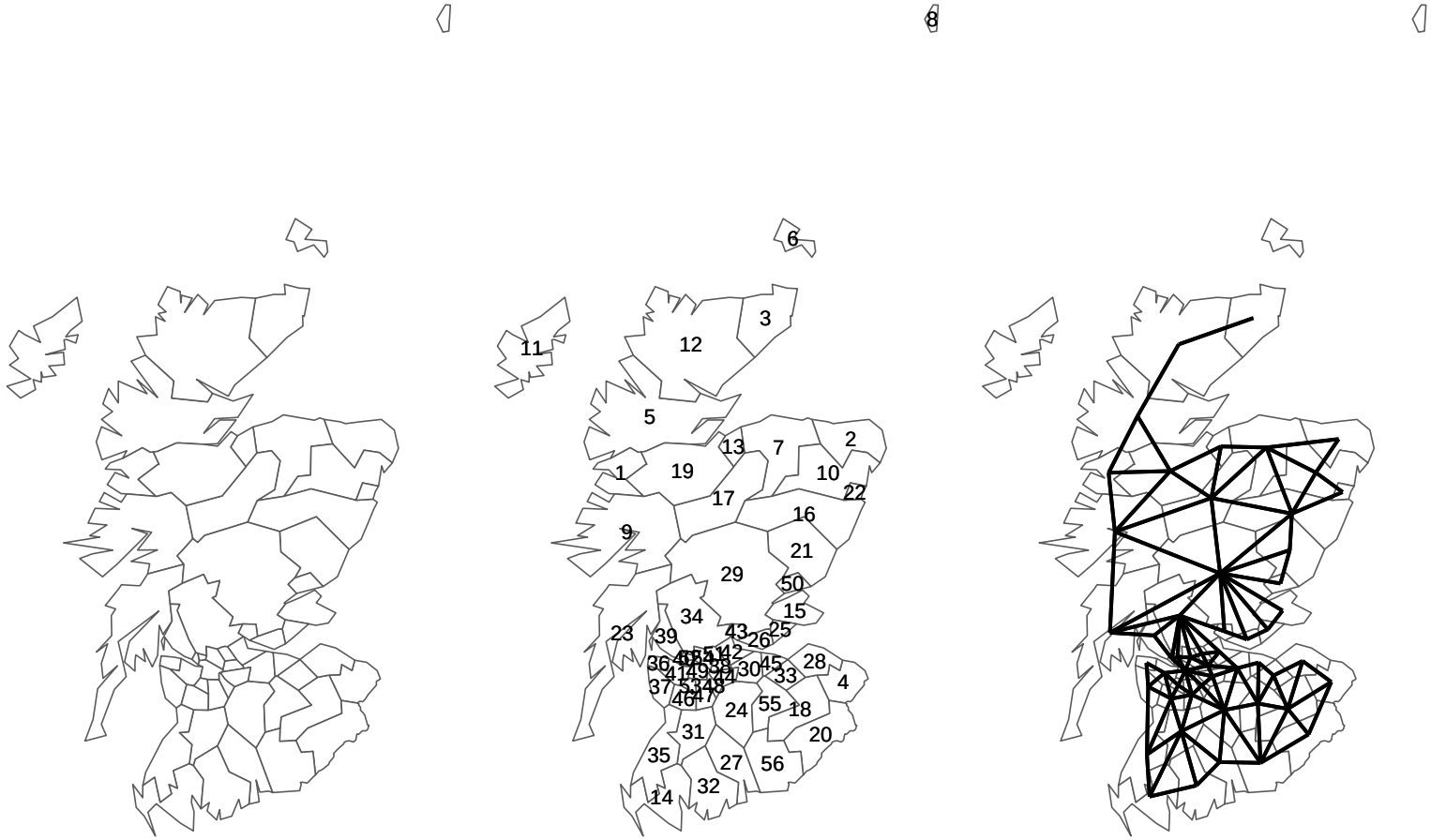
Summary

References

Spatial correlation

- Our GLM assumes that $\text{Covariance}(y_i, y_j) \equiv 0, \forall i \neq j$
- Moran's I generalises Pearson correlation, allowing us to look at spatial data
    - consider the values
    - consider their spatial location
    - how close geographically are they?
- Need to define a labelling scheme and calculate distance matrix and spatial weights

Spatial correlation



Spatial correlation

- For areal data, weights are based on whether two areas are neighbours
- If area i and area j have boundaries that share an edge, $w_{ij} = 1$

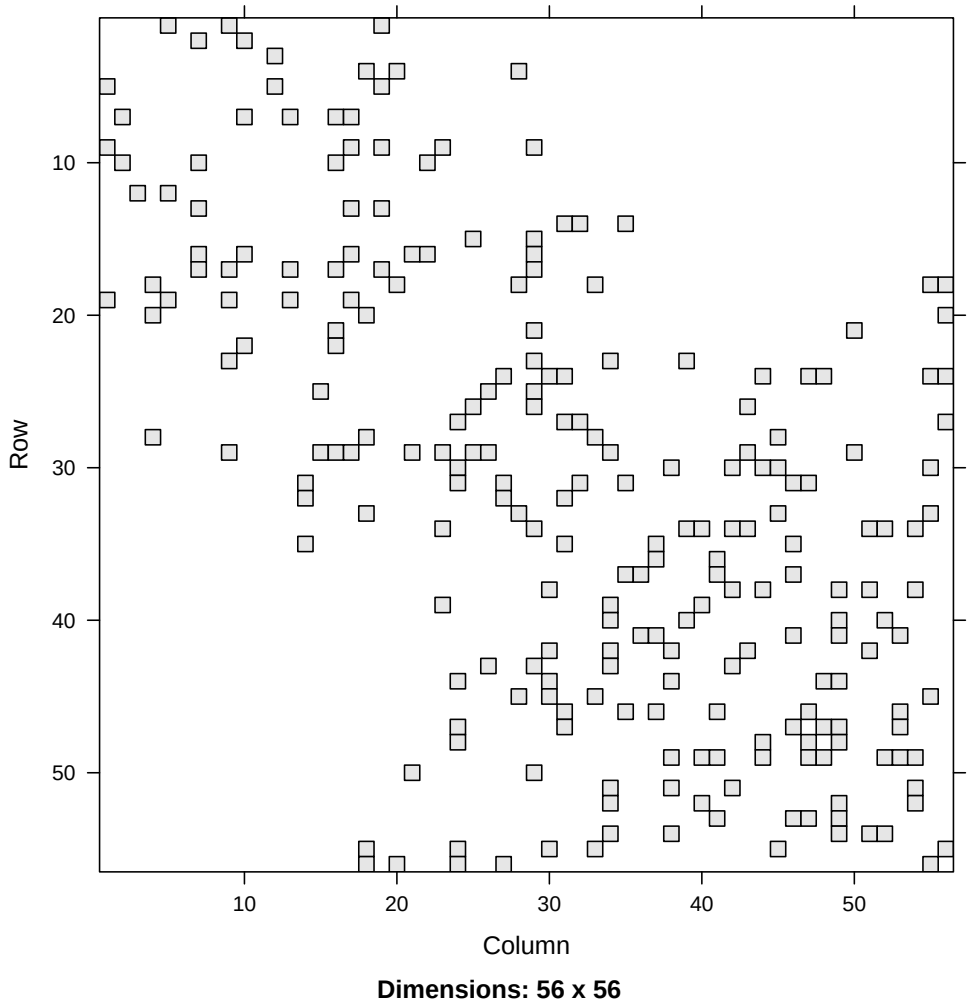
```
W <- st_intersects(scotland_shp)
W
```

```
## Sparse geometry binary predicate list of length 56, where the predicate
## was 'intersects'
## first 10 elements:
##  1: 1, 5, 9, 19
##  2: 2, 7, 10
##  3: 3, 12
##  4: 4, 18, 20, 28
##  5: 1, 5, 12, 19
##  6: 6
##  7: 2, 7, 10, 13, 16, 17
##  8: 8
##  9: 1, 9, 17, 19, 23, 29
##  10: 2, 7, 10, 16, 22
```

- NB: Area 1 and 5 are neighbours, so $w_{1,5} = 1 \leftrightarrow w_{5,1} = 1$
- NB: Area 6 has no neighbours, making it hard to smooth spatially

Spatial correlation - as a matrix

- Consider our Scottish Lip Cancer data
- Need to define a neighbourhood structure for the regions
- Requires a labelling scheme



Spatial Correlation - Moran's I

- Moran's I is a weighted version of the Pearson correlation (Moran 1950)

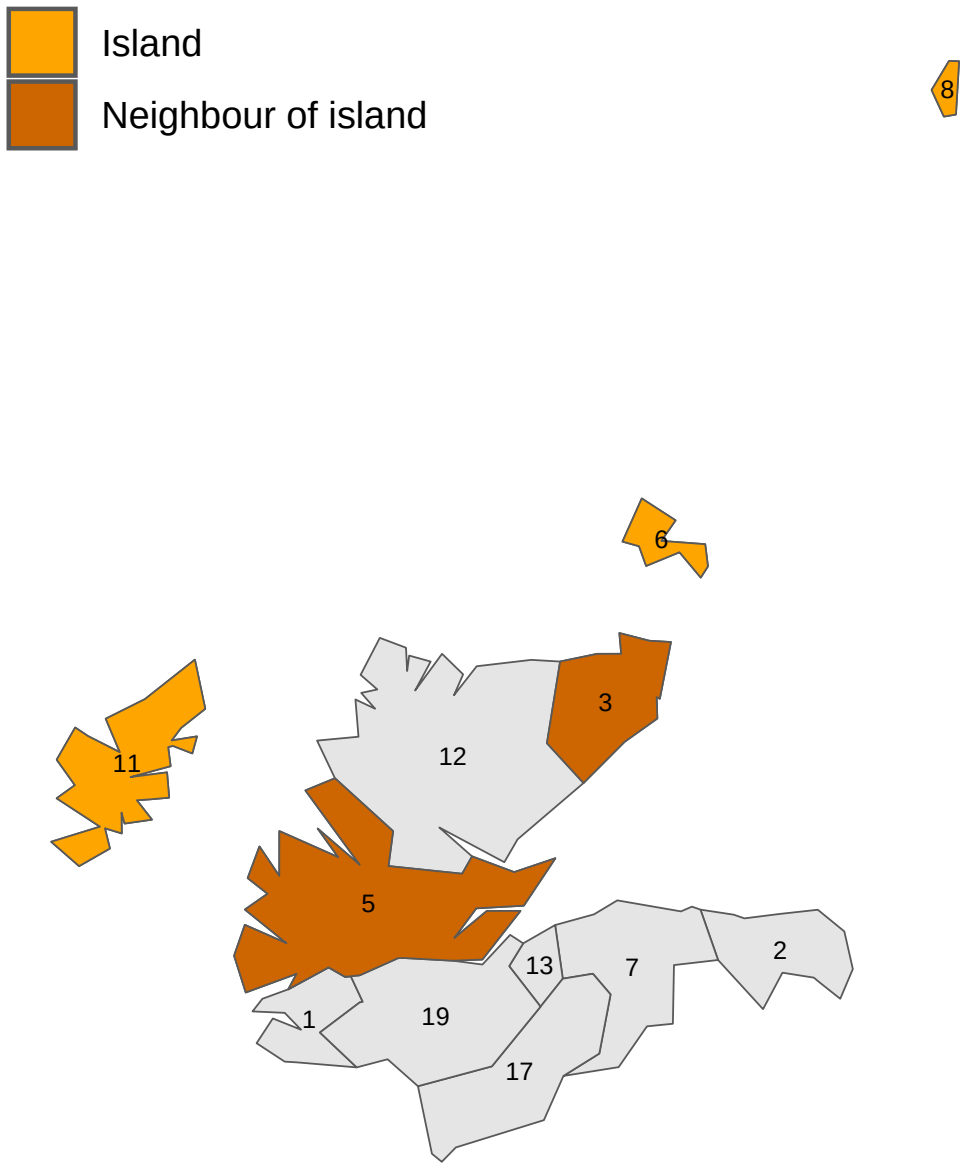
$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - \overline{y}\right)\left(y_j - \overline{y}\right)}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

- I takes values in range $(-1, 1)$, just like Pearson's ρ
- The second term considers the covariance as being weighted by adjacency
- The first term weights the covariance by the degree of connectivity in the spatial network
- Remember that $w_{ii} = 0$, a location can't be its own neighbour

Spatial Correlation - Lip cancer

- To ensure we can calculate Moran's I with our Scottish Islands, we need to fill in the gaps
- We can manually edit the matrix of weights to enforce adjacency
- For our islands, we want to make them neighbours with the nearest region
- Here, areas 6, 8, 11 are closest to areas 3, 3, 5, respectively so we set their $w_{ij} = 1$

Spatial Correlation - Lip cancer



Spatial Correlation - Lip cancer

Consider the following three models

- GLM with no covariates, offset only
- GLM as above but with AFF as covariate
- GAM replacing linear AFF term with spline

GAM can penalise the spline to be linear (model 2) or zero (model 1)

Spatial Correlation - Lip cancer

Model	Moran's I	AIC
GLM with offset only	0.558	591.175
GLM with AFF	0.332	450.820
GAM with AFF	0.352	419.949

- I on residuals on log scale
- Including AFF accounts for some of the spatial variation
- Given our plot of the residuals, it makes sense to explicitly model the correlation in the residuals

IID model

- We can assume independent and identically distributed spatial random effect values
- This is the simplest spatial model and considers each region as independent from the others

$$\begin{aligned} \boldsymbol{y} &\sim f(\boldsymbol{y}, \boldsymbol{\theta}) \\ g\left(\mathbb{E}(\boldsymbol{y})\right) &= X\boldsymbol{\beta} + \boldsymbol{v} \\ \boldsymbol{v} &\sim N\left(\mathbf{0}, \tau_v I\right) \end{aligned}$$

- This is equivalent to fitting our random effects mean model for the rats data previously
- We can refer to $\boldsymbol{v}$ as an **unstructured** spatial random effect
- $\boldsymbol{v} \equiv Z\boldsymbol{u}$ for convenience
  - $\boldsymbol{u}$ has length equal to number of regions
  - $\boldsymbol{v}$ has length equal to number of observations

CAR model

- The conditional autoregressive (CAR) model is one of the most famous spatial models (Besag 1974, 1975, 1986)
- Original work was Bayesian but has been adapted to non-Bayesian methods

$$\begin{aligned} \boldsymbol{y} &\sim f(\boldsymbol{y}, \boldsymbol{\theta}) \\ g\left(\mathbb{E}(\boldsymbol{y})\right) &= X\boldsymbol{\beta} + \boldsymbol{u} \\ \boldsymbol{u} &\sim N\left(\mathbf{0}, \Sigma\right) \end{aligned}$$

where Σ is the *variance-covariance matrix* and gives information about the spatial structure of the data. Typically easier to discuss $Q = \Sigma^{-1}$, the *precision matrix*

CAR model

The intrinsic CAR model's precision matrix is built as follows:

- if areas i and j are neighbours, $R_{ij} = -w_{ij} = -1$
- diagonals are $R_{ii} = \sum_{i \neq j} w_{ij}$
- $Q = \tau R$ where τ controls the amount of smoothing and usually has some prior, $\tau \sim p(\tau)$

Further detail on prior precision penalty matrices can be found in Havard Rue and Held (2005)

BYM model

- Another spatial model that extends the CAR model is the BYM model (Besag, York, and Mollié 1991)
    - combines CAR model for u with iid spatial random effect

$$g\left(\mathbb{E}(\boldsymbol{y})\right) = X\boldsymbol{\beta} + \boldsymbol{u} + \boldsymbol{v}$$

- This allows for spatial smoothing but also region-specific differences that can't be explained by a smooth spatial term
- These models are implemented in the `diseasemapping` package (Brown and Zhou 2018) which makes use of the R-INLA package (Håvard Rue, Martino, and Chopin 2009; Lindgren, Rue, and Lindström 2011; Martins et al. 2013)

MRF model

- Instead of the Bayesian CAR and BYM models, the Markov Random Field approach doesn't rely on Bayesian statistics
- Specify a neighbourhood matrix, e.g.
    - as for Moran's I
    - as for CAR with $w_{ii} = \sum_{j \neq i} -w_{ij}$
    - as for IID with $w_{ij} = \delta(i = j)$
- We have freedom to define whatever neighbourhood structure we want
- NB: the CAR specification is analagous to the wiggliness penalty on penalised splines (Havard Rue and Held 2005)

Spatial modelling

Consider a set of models with different spatial random effects

- None
- IID
- Smoothing with automatically calculated neighbourhood structure from list of polygons
- MRF with specified penalty matrix (as above but adjusted for islands)

Fit these models to Scottish lip cancer data with and without AFF

Spatial modelling

Figure 4: Predicted cancer cases and AIC for each model

GAMs

- The above are all examples of Generalised Additive Models
- These models replace the linear terms of a GLM with *scatterplot smoothers*
- Incredibly useful when you don't know the functional form of a relationship
- Implemented in `mgcv` package (Wood 2017; Wood et al. 2016)

What is spatial data?

Geostatistics

Regression modelling

Spatial correlation

Spatial models as GAMs

Spatial modelling with GAMs in R

Summary

References

GAMs

To introduce spatial terms, we specify that we want a smooth random effect, with `s()`

```
CANCER ~ offset(log(CEXP)) + s(RECORD_ID, bs="re")
```

Here we are telling `gam()` that we want a **s**mooth function of `RECORD_ID`, the spatial location, to be fit as a **r**andom **e**ffect
This is different to the syntax yesterday, and integrates better with the rest of mgcv

GAMs

If we want to specify a structured spatial effect, we need to pass info about the spatial relationships, e.g. the MRF with the neighbourhood structure used for Moran's I

```
CANCER ~ offset(log(CEXP)) +
  s(factor(RECORD_ID), bs = "mrf",
    xt = list(polys = scotland_nb))
```

Here we pass a list of polygons and R will check which polygons are neighbours
MRFs set up a fixed effect basis

GAMs

If we want to specify a structured random effect, we need to pass info about the spatial relationships, e.g. the MRF with the CAR-style smoothing

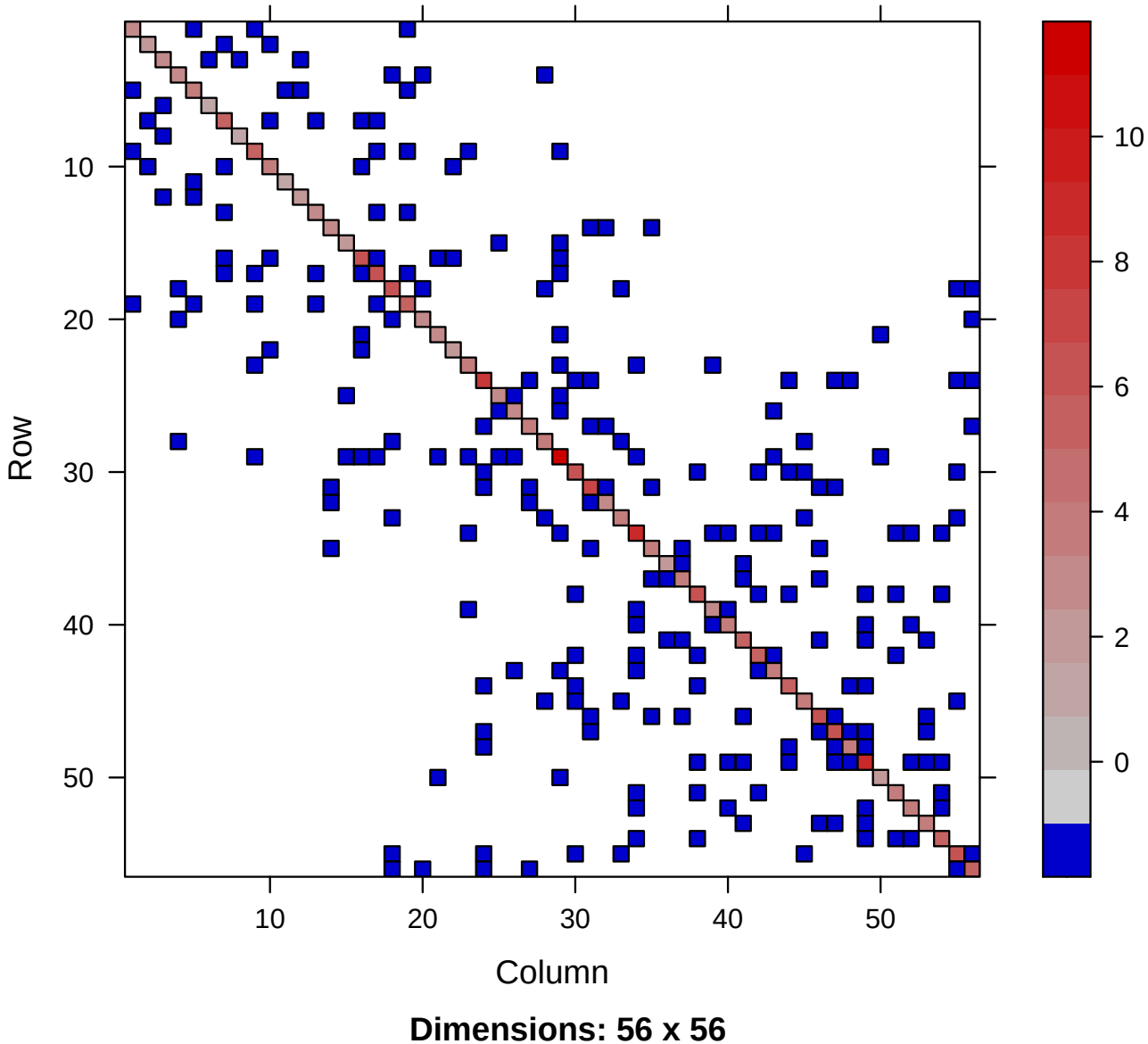
```
scotland_penalty <- -W_mat + diag(rowSums(W_mat))

CANCER ~ offset(log(CEXP)) +
  s(factor(RECORD_ID), bs = "mrf",
    xt = list(penalty = scotland_penalty))
```

Here we're telling `gam()` that we want a Markov Random Field model with a penalty matrix given by the Q

GAMs

The penalty matrix looks like:



Dimensions: 56 x 56

GAMs

- The MRF above has one spatial fixed effect term for each spatial location
- We can't add explanatory variables to this models because we only have one observation per location
- If we want to add `AFF` to predict lip cancer, we have to reduce the spatial complexity of the model
- This is done by fitting a low-rank MRF

```
CANCER ~ offset(log(CEXP)) +
  s(factor(RECORD_ID), bs = "mrf",
    xt = list(penalty = scotland_penalty), k = 28)
# instead of 56, use a basis of size 28
# fewer degrees of freedom consumed
```

GAMs - low rank smoothers

- In summary, low rank means we use a small number of basis functions to approximate the behaviour of the full rank system
- e.g. we might be able to approximate a complex polynomial with a quadratic
- We do lose some spatial detail by doing this
- More detail on low rank smoothers can be found in Wood (2003) and Wood (2017) section 5.8.1

Spatial correlation again

- Which of our models's residuals exhibited the least spatial correlation, and why?

Table 2: Moran's I for residuals from fitted models

Model	Moran's I	AIC
Offset only	0.558	591.175
IID random effect	0.088	222.928
Neighbours	-0.062	297.005
MRF smooth	0.242	310.906
AFF	0.352	419.949
IID + AFF	0.083	224.891
Neighb. + AFF	-0.151	293.902
MRF + AFF	0.208	326.828

Some questions

- Are our spatial terms fixed or random effects?

What is spatial data?

Geostatistics

Regression modelling

Spatial correlation

Spatial models as GAMs

Spatial modelling with GAMs in R

Summary

References

Some questions

- Are our spatial terms fixed or random effects?
- Why can't we use `AFF` as an explanatory variable with a full rank spatial smoother?

What is spatial data?

Geostatistics

Regression modelling

Spatial correlation

Spatial models as GAMs

Spatial modelling with GAMs in R

Summary

References

Some questions

- Are our spatial terms fixed or random effects?
- Why can't we use `AFF` as an explanatory variable with a full rank spatial smoother?
- Under what conditions could we use a full rank smoother *and* explanatory variables?

Some questions

- Are our spatial terms fixed or random effects?
- Why can't we use `AFF` as an explanatory variable with a full rank spatial smoother?
- Under what conditions could we use a full rank smoother *and* explanatory variables?
- How do we assess spatial autocorrelation?

Some questions

- Are our spatial terms fixed or random effects?
- Why can't we use `AFF` as an explanatory variable with a full rank spatial smoother?
- Under what conditions could we use a full rank smoother *and* explanatory variables?
- How do we assess spatial autocorrelation?
- How do we encode the spatial relationships in our model?

What is spatial data?

Geostatistics

Regression modelling

Spatial correlation

Spatial models as GAMs

Spatial modelling with GAMs in R

Summary

References

Summary

- Random effects models allow us to model unexplained variability inherent in group structure
- Spatial modelling requires idea of adjacency of observational units
- Want to make sure adding more parameters to model (e.g. spatial random effect) is worth it in terms of goodness of fit

References I

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed. Chapman; Hall/CRC.

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (2): 192–236. http://www.jstor.org/stable/2984812.

———. 1975. "Statistical Analysis of Non-Lattice Data." *Journal of the Royal Statistical Society. Series D (The Statistician)* 24 (3): 179–95. http://www.jstor.org/stable/2987782.

———. 1986. "On the Statistical Analysis of Dirty Pictures." *Journal of the Royal Statistical Society. Series B (Methodological)* 48 (3): 259–302. http://www.jstor.org/stable/2345426.

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian Image Restoration, with Two Applications in Spatial Statistics." *Annals of the Institute of Statistical Mathematics* 43 (1): 1–20. https://doi.org/10.1007/bf00116466.

Brown, Patrick E, and L Zhou. 2018. *Diseasemapping: Modelling Spatial Variation in Disease Risk for Areal Data*. https://CRAN.R-project.org/package=diseasemapping.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. "Model-Based Geostatistics." *Journal of the Royal Statistical Society. Series C: Applied Statistics* 47 (3): 299–325.

Diggle, Peter J., and Amanda G. Chetwynd. 2011. *Statistics and Scientific Method*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199543182.001.0001.

Lindgren, Finn, Håvard Rue, and Johan Lindström. 2011. "An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (4): 423–98. https://doi.org/10.1111/j.1467-9868.2011.00777.x.

Martins, Thiago G., Daniel Simpson, Finn Lindgren, and Håvard Rue. 2013. "Bayesian Computing with INLA: New Features." *Computational Statistics & Data Analysis* 67 (November): 68–83. https://doi.org/10.1016/j.csda.2013.04.014.

References II

Matheron, Georges. 1963. “Principles of Geostatistics.” *Economic Geology* 58 (8): 1246–66. https://doi.org/10.2113/gsecongeo.58.8.1246.

Moran, P. A. P. 1950. “Notes on Continuous Stochastic Phenomena.” *Biometrika* 37 (1/2): 17. https://doi.org/10.2307/2332142.

Rue, Havard, and Leonhard Held. 2005. *Gaussian Markov Random Fields: Theory and Applications*. CRC press.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. “Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations.” *Journal of the Royal Statistical Society: Series b (Statistical Methodology)* 71 (2): 319–92.

Tobler, W. R. 1970. “A Computer Movie Simulating Urban Growth in the Detroit Region.” *Economic Geography* 46 (sup1): 234–40. https://doi.org/10.2307/143141.

Wood, S. N. 2003. “Thin-Plate Regression Splines.” *Journal of the Royal Statistical Society (B)* 65 (1): 95–114.

———. 2017. *Generalized Additive Models: An Introduction with r*. 2nd ed. Chapman; Hall/CRC.

Wood, S. N., N., Pya, and B. Säfken. 2016. “Smoothing Parameter and Model Selection for General Smooth Models (with Discussion).” *Journal of the American Statistical Association* 111: 1548–75.