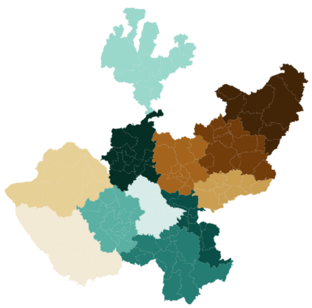


Intrinsic Auto-Regressive Models

Spatial data analysis in Stan

Sue Marquez
Data Scientist at The Rockefeller Foundation

Hi!

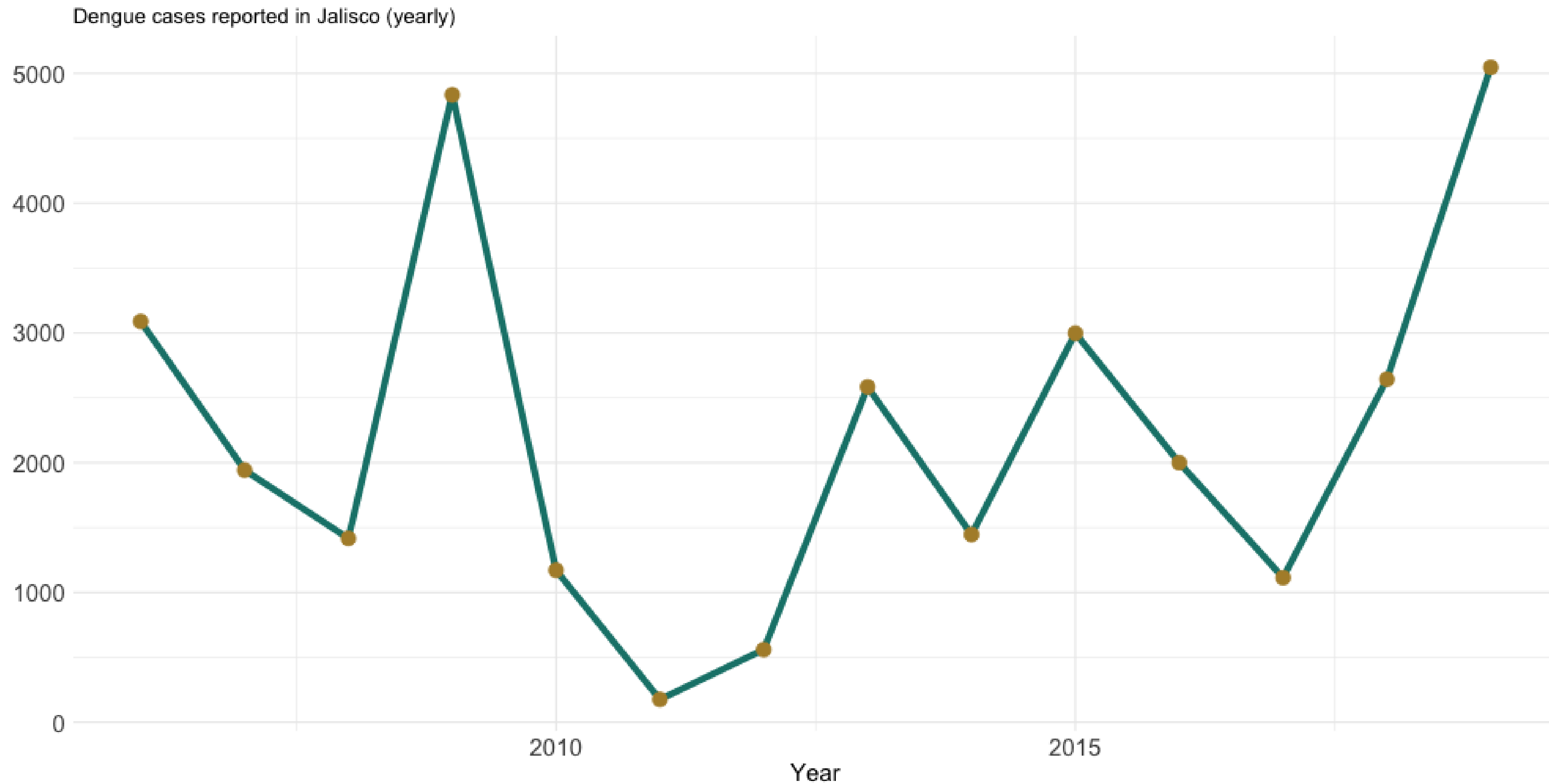
- Sue Marquez, Manager and Data Scientist at The Rockefeller Foundation
- Previously, I was a Data Scientist at BuzzFeed and a Statistical geneticist at The Feinstein Institute for Medical Research
- Hold a Graduate Diploma in Statistics and Stochastic Processes from the University of Melbourne, Australia.

Structure of this talk

1. Motivating question: Spike in cases of dengue in Jalisco.
2. Why we can't have nice things!
3. What are Conditional Auto-Regressive models (CAR)?
4. What is the intrisic part of the Intrinsic Auto-Regressive models (ICAR)?
5. Implementation of ICAR in Stan

Motivating question: Spike in cases of dengue in Jalisco.

Dengue cases in Jalisco



Jalisco polygon

Mapa general de Jalisco 2012 modificado por decreto 26837, límite estatal

Limite Estatal del Mapa General del Estado de Jalisco 2012, es un archivo vectorial con geometría de polígono que define el límite territorial del Estado de Jalisco, actualizado en su mayoría a escala 1:50,000. El polígono corresponde al Mapa General de Jalisco 2012, publicado en el Periódico Oficial El Estado de Jalisco, el 27 de marzo de 2012 y modificado por Decreto 26837/LXI/18 Mezquitic publicado en el Periódico Oficial El Estado de Jalisco, el 3 de junio de 2018.
La información aquí vertida tiene un carácter referencial, sujeta a discusión y en su caso a corrección para que llegado el momento se puedan lograr los acuerdos correspondientes y su aprobación definitiva.

Datos y recursos

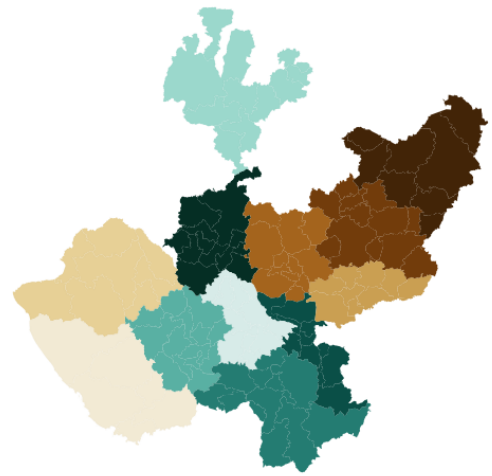
Limite estatal 2012 modificado por Decreto 26837, en shp
Polígono del límite estatal del MGJ 2012, modificado mediante Decreto 26837...

DESCARGAR

Limite estatal 2012 modificado por Decreto 26837, en kml
Polígono del límite estatal 2012, modificado mediante Decreto 26837

DESCARGAR

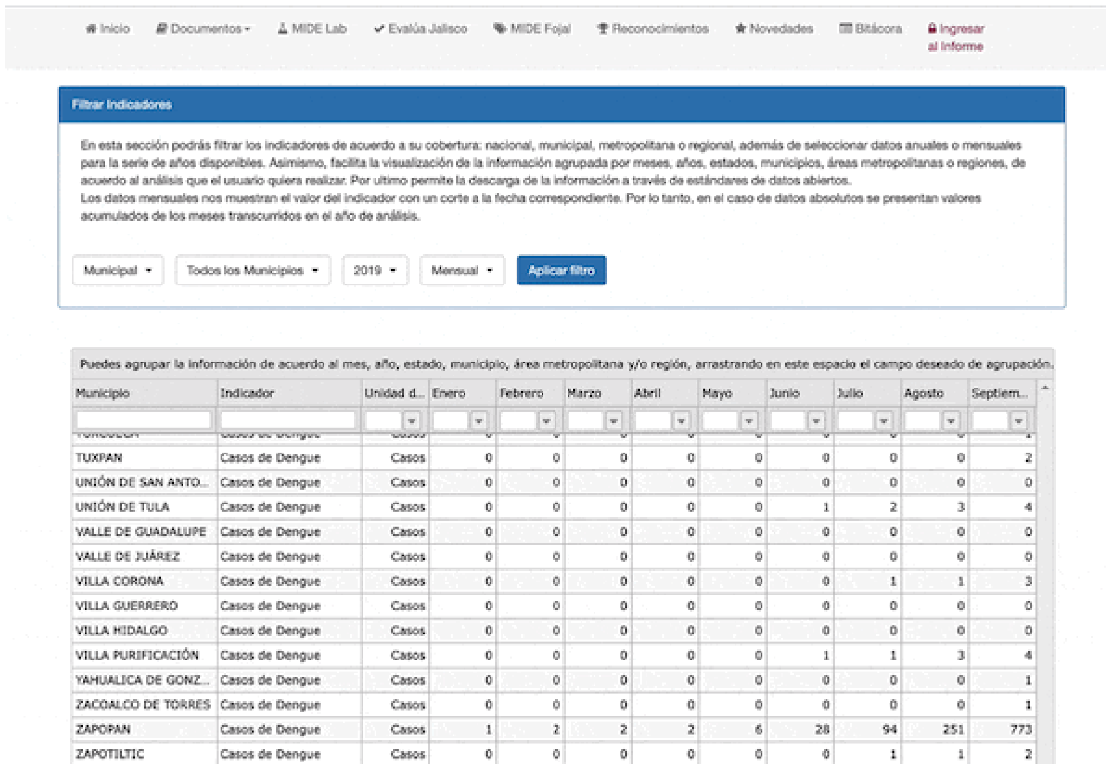
DESCARGAR TODOS



Jalisco Polygon (https://datos.jalisco.gob.mx/dataset/mapa-general-de-jalisco-limite-estatal)

Dengue data

MIDE Jalisco (https://seplan.app.jalisco.gob.mx/mide/panelCiudadano/tablaDatos?nivelTablaDatos=3&periodicidadTablaDatos=anual&indicadorTablaDatos=772&accionReg


Inicio Documentos MIDE Lab Evalúa Jalisco MIDE Fojal Reconocimientos Novedades Bitácora Ingresar al Informe

Filtrar Indicadores

En esta sección podrás filtrar los indicadores de acuerdo a su cobertura: nacional, municipal, metropolitana o regional, además de seleccionar datos anuales o mensuales para la serie de años disponibles. Asimismo, facilita la visualización de la información agrupada por meses, años, estados, municipios, áreas metropolitanas o regiones, de acuerdo al análisis que el usuario quiera realizar. Por ultimo permite la descarga de la información a través de estándares de datos abiertos.
Los datos mensuales nos muestran el valor del indicador con un corte a la fecha correspondiente. Por lo tanto, en el caso de datos absolutos se presentan valores acumulados de los meses transcurridos en el año de análisis.

Municipal Todos los Municipios 2019 Mensual Aplicar filtro

Puedes agrupar la información de acuerdo al mes, año, estado, municipio, área metropolitana y/o región, arrastrando en este espacio el campo deseado de agrupación.

Municipio	Indicador	Unidad d...	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiem...
TUXPAN	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	2
UNIÓN DE SAN ANTO...	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	0
UNIÓN DE TULA	Casos de Dengue	Casos	0	0	0	0	0	1	2	3	4
VALLE DE GUADALUPE	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	0
VALLE DE JUÁREZ	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	0
VILLA CORONA	Casos de Dengue	Casos	0	0	0	0	0	0	1	1	3
VILLA GUERRERO	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	0
VILLA HIDALGO	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	0
VILLA PURIFICACIÓN	Casos de Dengue	Casos	0	0	0	0	0	1	1	3	4
YAHUALICA DE GONZ...	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	1
ZACOALCO DE TORRES	Casos de Dengue	Casos	0	0	0	0	0	0	0	0	1
ZAPOPAN	Casos de Dengue	Casos	1	2	2	2	6	28	94	251	773
ZAPOTILTIC	Casos de Dengue	Casos	0	0	0	0	0	0	1	1	2

/

Why we can't have nice things!

Why we can't have nice things!

Spatial autocorrelation

Spatial autocorrelation

Autocorrelation is a measurement of similarity between close observations of the same phenomenon.

> **Example with temporal autocorrelation:** If you measure your weight, two observations close in time are very similar than distant ones.

Spatial autocorrelation is more nuanced because, unlike time, spatial variables are at least two-dimensional.

> **Spatial autocorrelation**: Describe the extent to which two observations from neighboring regions exhibit higher correlation than distant ones.

Autocorrelation in spatial data

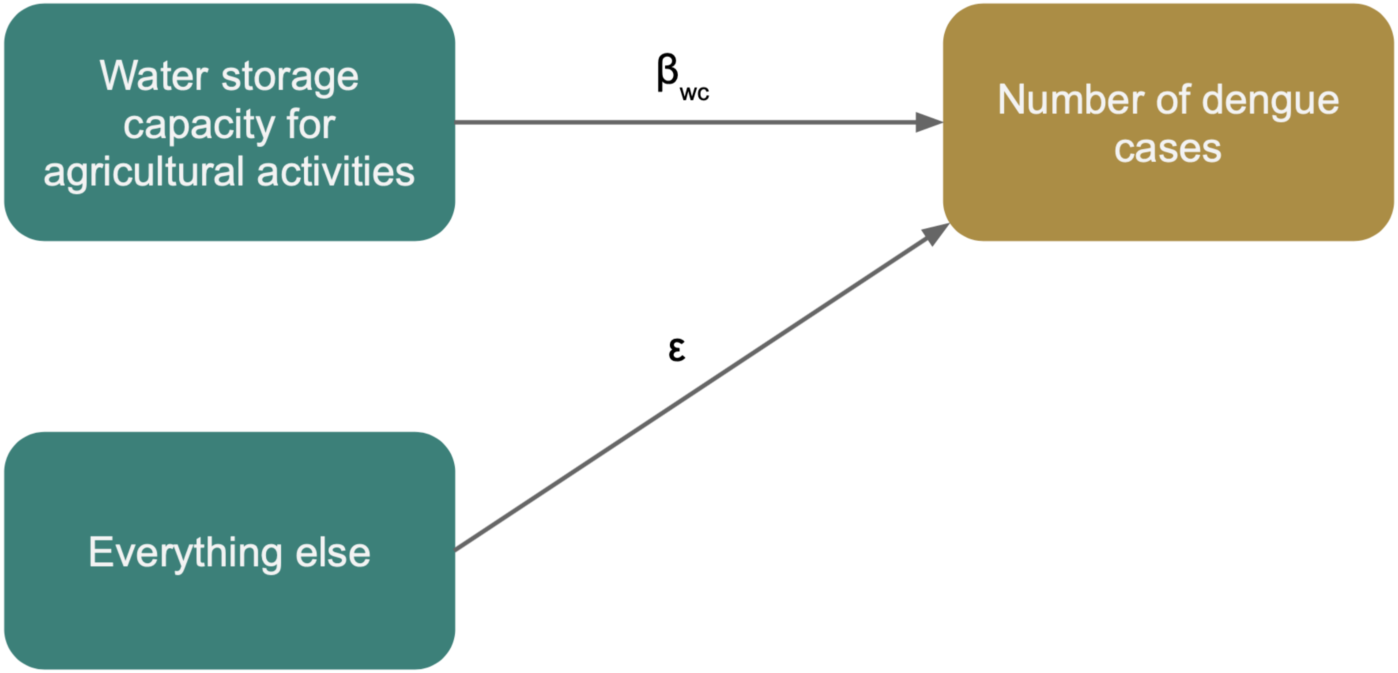
- In regression analysis, one of the standard assumptions is that errors are uncorrelated.
- Correlated errors suggest we have additional information in the data that has not been accounted for in the model as it is.
- In the case of spatial data, adjacent residuals tend to be similar and therefore *autocorrelated*.

> **Main problem:** if autocorrelation is not exploited in your model, your explanatoy variables coefficients will display an unusual explanatory power, which might be the consequence of of just fitting spatial noise.

Initial question about dengue

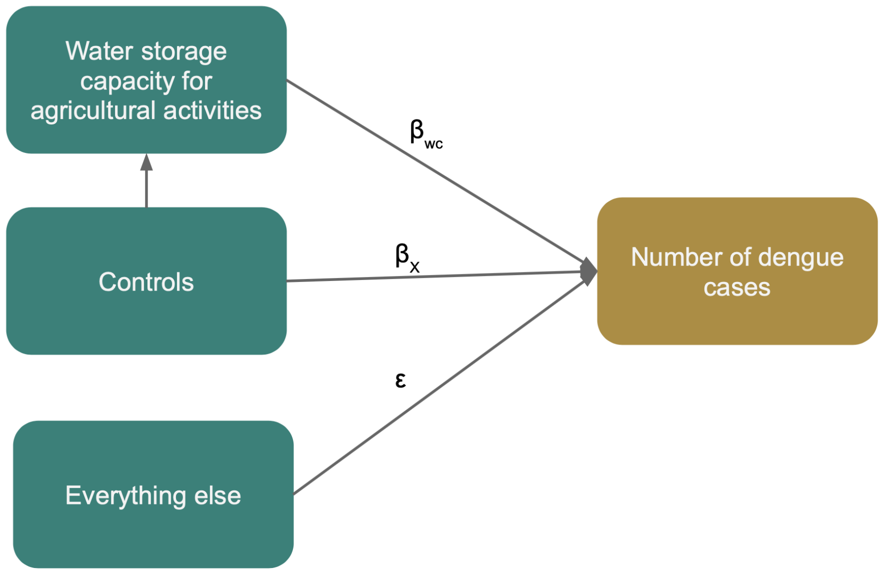
Simple model

$$y = \beta_0 + (\mathrm{WC})\beta_{wc} + \epsilon$$



Let's add covariates

$$y = \beta_0 + (\mathrm{WC})\beta_{wc} + X\beta_X + \epsilon$$



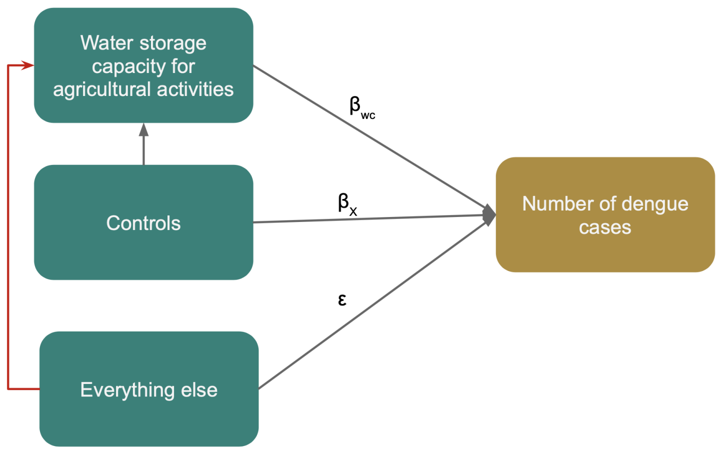
Assuming that *everything else* does not affect *water capacity* this model should be decent.

When *everything else* contains spatial correlation

We are fitting this

$$y = \beta_0 + (\mathrm{WC})\beta_{wc} + X\beta_X + \epsilon, \quad \text{which assumes } E[\epsilon|X] = 0$$

But in reality, we have this: $E[\epsilon|X] \neq 0$



> *Our coefficient estimates will be wrong!*

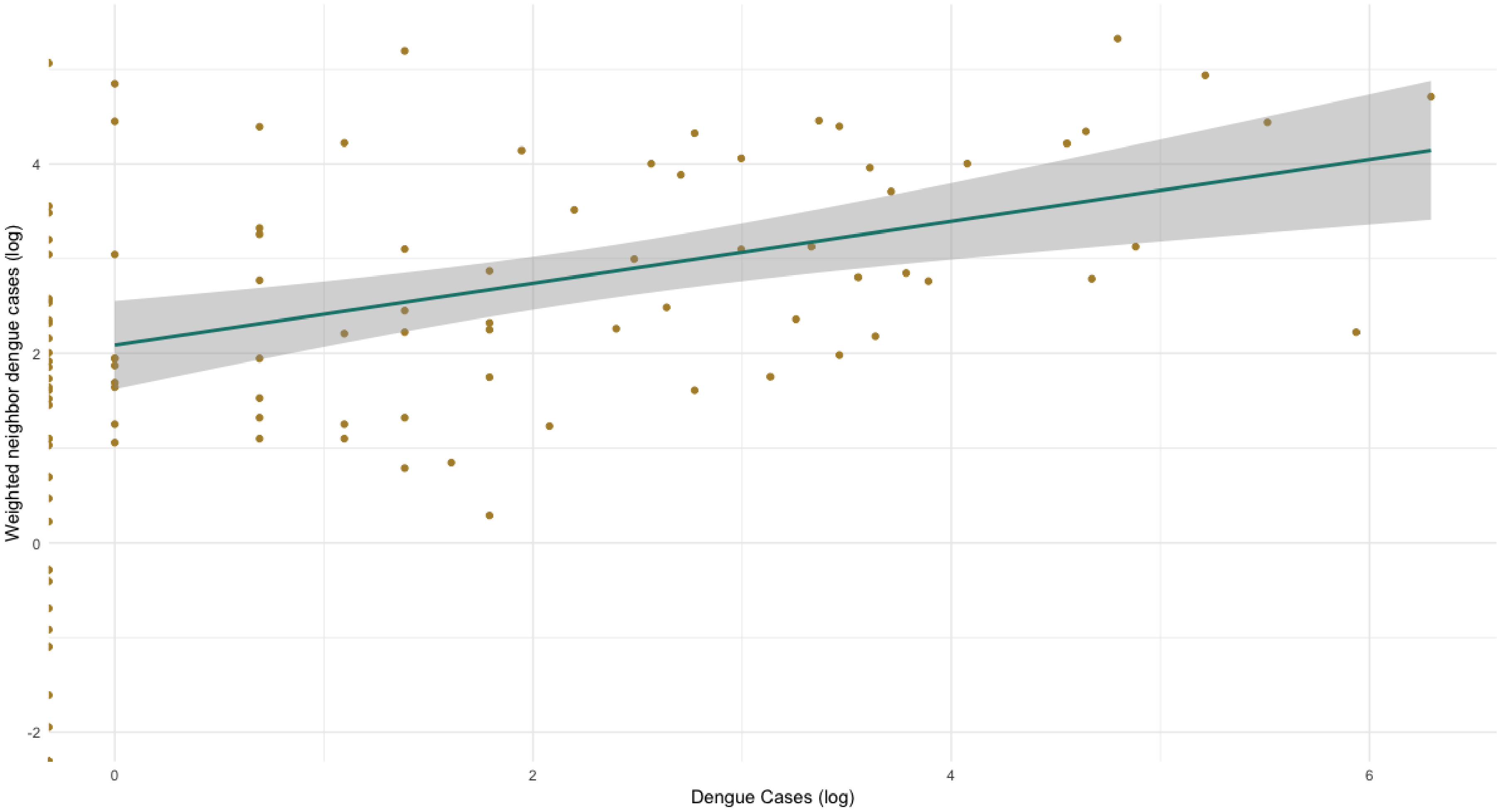
Moran's I (autocorrelation statistic)

- Analogous to the the standard correlation concept.

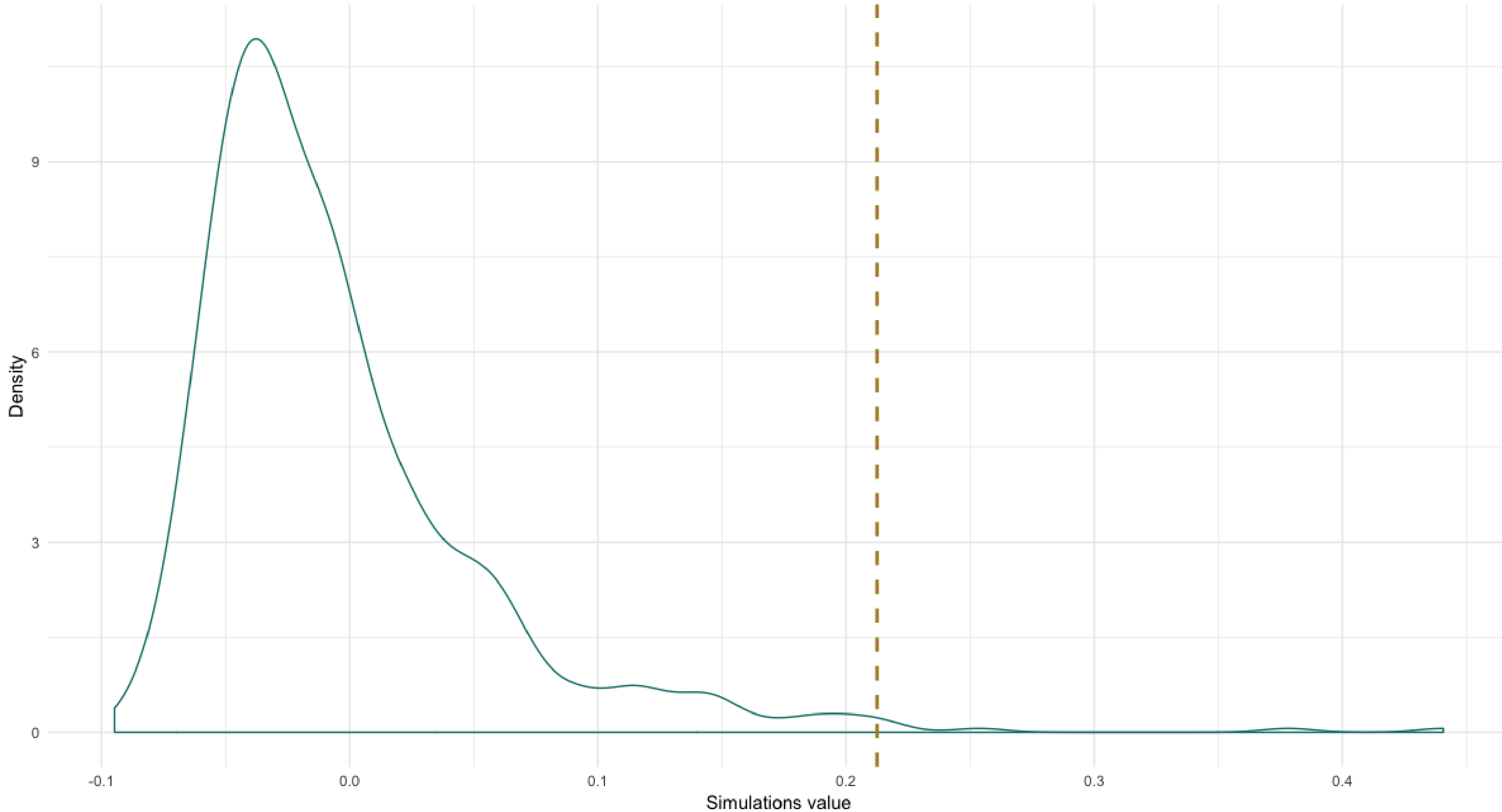
$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- Numerator measuring deviatiom from the mean for adjacent units.
- Denominator standardizes the quantity to reflect the variability of the quantity of interest.

Moran's I (Jalisco data)



Moran's I test (Jalisco data)



Moran's I test (Jalisco data)

```
Monte-Carlo simulation of Moran I

data:  moran_df$dengue_cases
weights: l_weight
number of simulations + 1: 601

statistic = 0.21246, observed rank = 597, p-value = 0.006656
alternative hypothesis: greater
```

But Sue, is this really a problem in other research areas?

> Kelly, Morgan, *The Standard Errors of Persistence* (June 3, 2019) (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3398303)

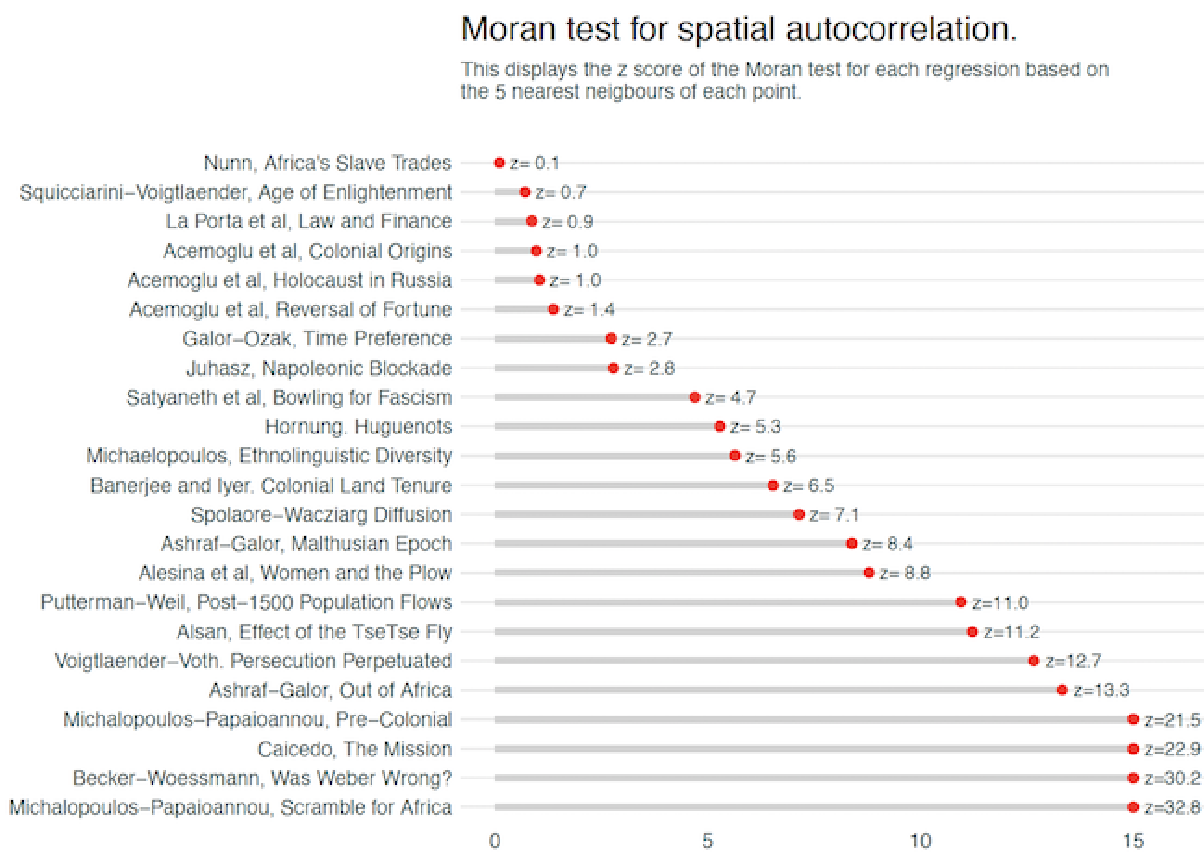


Figure 6: Z scores of Moran tests for spatial autocorrelation in regression residuals.

What are Conditional Auto-Regressive models (CAR)?

Conditional Auto-Regressive models (CAR)

- CAR models are a class of spatial models used to estimate spatial autocorrelation.
- These models are widely used in Ecology, Economics and Epidemiology.
- CAR was first developed by Julian Besag in his now classic 1974 paper *Spatial Interaction and the Statistical Analysis of Lattice Systems*.

CAR specifications

- Single aggregated measure per spatial unit, it can be continuous, binary or discrete count.

> **Example:** Number of car accidents at the county level.

- Finite set of non-overlapping spatial units.
- For spatial units, the relationship is defined in terms of adjacency.

CAR model

Let N be the total number of spatial units from a region.

A neighbor relationship is defined as $i \sim j$ where $i \neq j$.

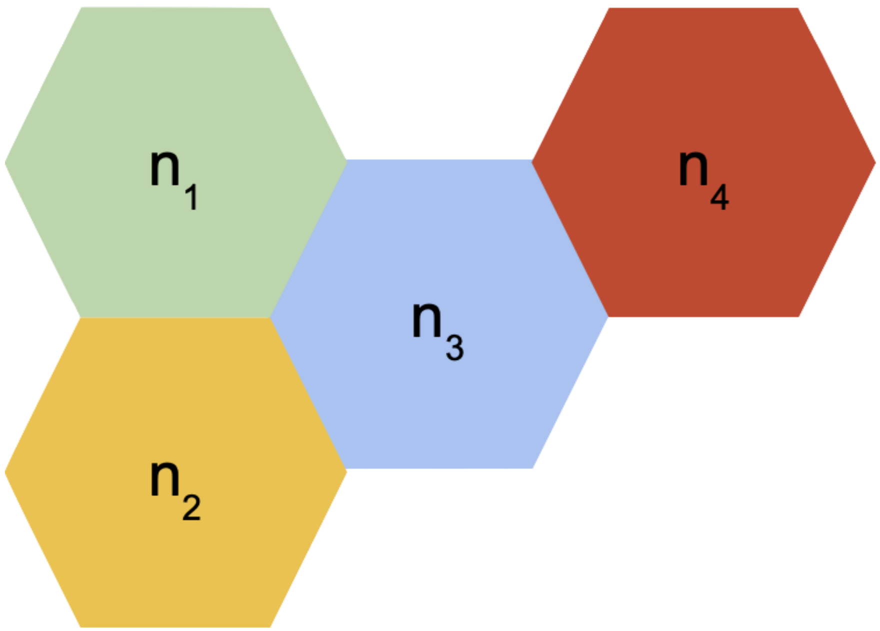
$$\begin{array}{ll} n_i \text{ and } n_j \text{ are adjacent} & 1 \\ \text{otherwise} & 0 \end{array}$$

This relationship is symmetric (i.e if $i \sim j \Rightarrow j \sim i$) but not reflexive (i.e. a region cannot be neighbor of itself).

Adjacency!

There are two matrices describing different measures of adjacency in this model.

1) Adjacency matrix W, defining neighbor relationship.

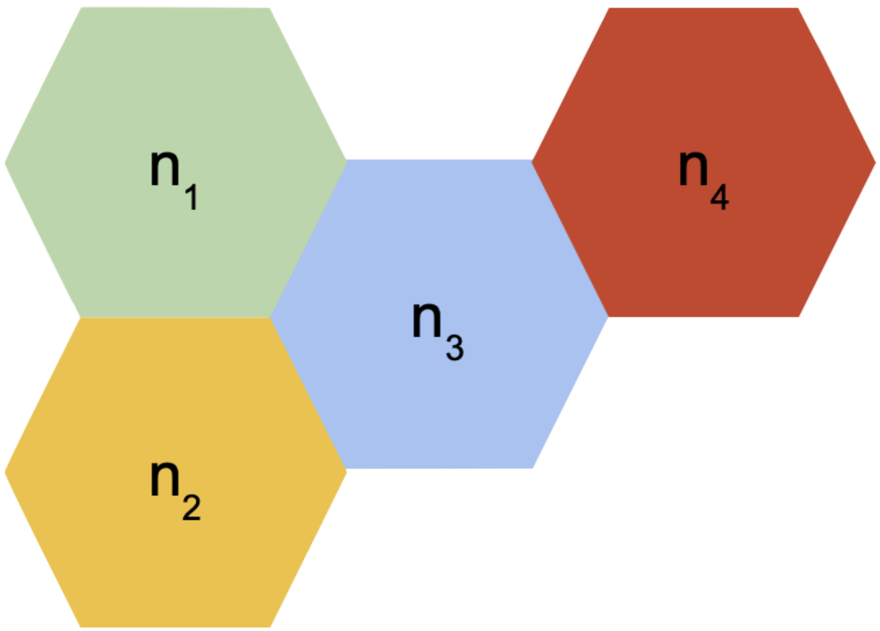


	n_1	n_2	n_3	n_4
n_1	0	1	1	0
n_2	1	0	1	0
n_3	1	1	0	1
n_4	0	0	1	0

Adjacency!

There are two matrices describing different measures of adjacency in this model.

1) Diagonal matrix D, defining number of adjacent units.



	n_1	n_2	n_3	n_4
n_1	2	0	0	0
n_2	0	2	0	0
n_3	0	0	3	0
n_4	0	0	0	1

CAR model

Spatial interaction between areal units is modelled *conditionally* as a normally distributed random variable, represented by the N-length vector $\boldsymbol{\phi}$ (i.e. $\phi = (\phi_1, \ldots, \phi_n)^T$).

Therefore, the conditional distribution of *EACH* ϕ_i is defined as follows,

$$p(\phi_i | \phi_j, i \neq j) \sim N\left(\alpha \sum_{j=1}^{n} w_{ij}\phi_j,\ \sigma^2\right) \quad i, j = 1, \ldots, n.$$

where $w_{ij}\phi_j$ is the weighted values of the neighbors.

From Banerjee, Carlin, and Gelfand, 2004, sec. 3.2, it follows that the joint distribution $\phi \sim N(\mathbf{0}, [D(I - \alpha W)]^{-1})$

CAR model

$$\phi \sim N(\mathbf{0}, [D(I - \alpha W)]^{-1})$$

Recap!

- α: between 0 and 1, it represents the strength of the spatial association, with 0 meaning spatial independence.
- D is our diagonal matrix.
- W is the adjacency matrix.

What is the intrisic part of the Intrinsic Auto-Regressive models (ICAR)?

The intrinsic conditional autoregressive (ICAR)

The difference between CAR and ICAR is that the parameter α is set to 1.

- α = 1
- *D* is our diagonal matrix.
- *W* is the adjacency matrix.

$$
\begin{aligned}
\phi & \sim N(\mathbf{0}, [D(I - \alpha D^{-1} W)]^{-1}),\ \alpha = 1 \\
& \sim N(\mathbf{0}, [D(I - D^{-1} W)]^{-1}) \\
& \sim N(\mathbf{0}, [D - W]^{-1})
\end{aligned}
$$

However, setting $\alpha = 1$ creates a challenge because $[D - W]$ becomes a singular matrix (i.e. non-invertible).

Thankfully, including the constraint $\sum_i \phi_i = 0$ solves this challenge.

Pairwise derivation

ICAR component is then defined as follows,

$$\phi \sim N(\mathbf{0}, [D - W]^{-1})$$

and after some algebra, the log probability density becomes:

$$\log p(\phi) \propto \frac{1}{2} \left(\sum_{i \sim j} (\phi_i - \phi_j)^2 \right)$$

Stan

Stan is an open-source probabilistic programming language.

It's written in C++ and and genrally speaking, it is used to specify Bayesian statistical models.

Stan estimate parameters by calculating the **log probability density**. (Try multiplying a large number of observations with tiny numbers, you will quickly run into numerical errors.)

Stan model structure

```
// The input data is a vector 'y' of length 'N'.
data
{
  int<lower=0> N;
  vector[N] y;
}


// The parameters accepted by the model.
parameters {
  real mu;
  real<lower=0> sigma;
}

// The model where 'y' is normally distributed with mean 'mu'
// and standard deviation 'sigma'.
model {
  y ~ normal(mu, sigma);
}
```

Stan model structure

```
data
{
  int<lower=0> N; // number of obs
  int<lower=0> K; // number of cols in design matrix
  int<lower=0> N_edges;
  int<lower=1, upper=N> node1[N_edges];  // node1[i] adjacent to node2[i]
  int<lower=1, upper=N> node2[N_edges];  // and node1[i] < node2[i]

  int<lower=0> y[N];    // count outcomes
  matrix[N,K] X;        //the model matrix
  vector<lower=0>[N] E; // exposure

}
```

Stan model structure

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

parameters {
  real beta0;               // intercept
  vector[K] beta;           //the regression parameters
  real<lower=0> sigma;      // overall standard deviation
  vector[N] phi;           // spatial effects
}
```

Stan model structure

```
model {
  y ~ poisson_log(log_E + beta0 + X * beta + phi * sigma);
  beta0 ~ cauchy(0,2);

  for(i in 1:K)
    beta[i] ~ normal(0.0, 1.0);

  sigma ~ normal(0.0, 1.0);
  phi ~ icar_normal_lpdf(N, node1, node2);
  // soft sum-to-zero constraint on phi
  sum(phi) ~ normal(0, 0.001 * N);
}
```

Fitting different models

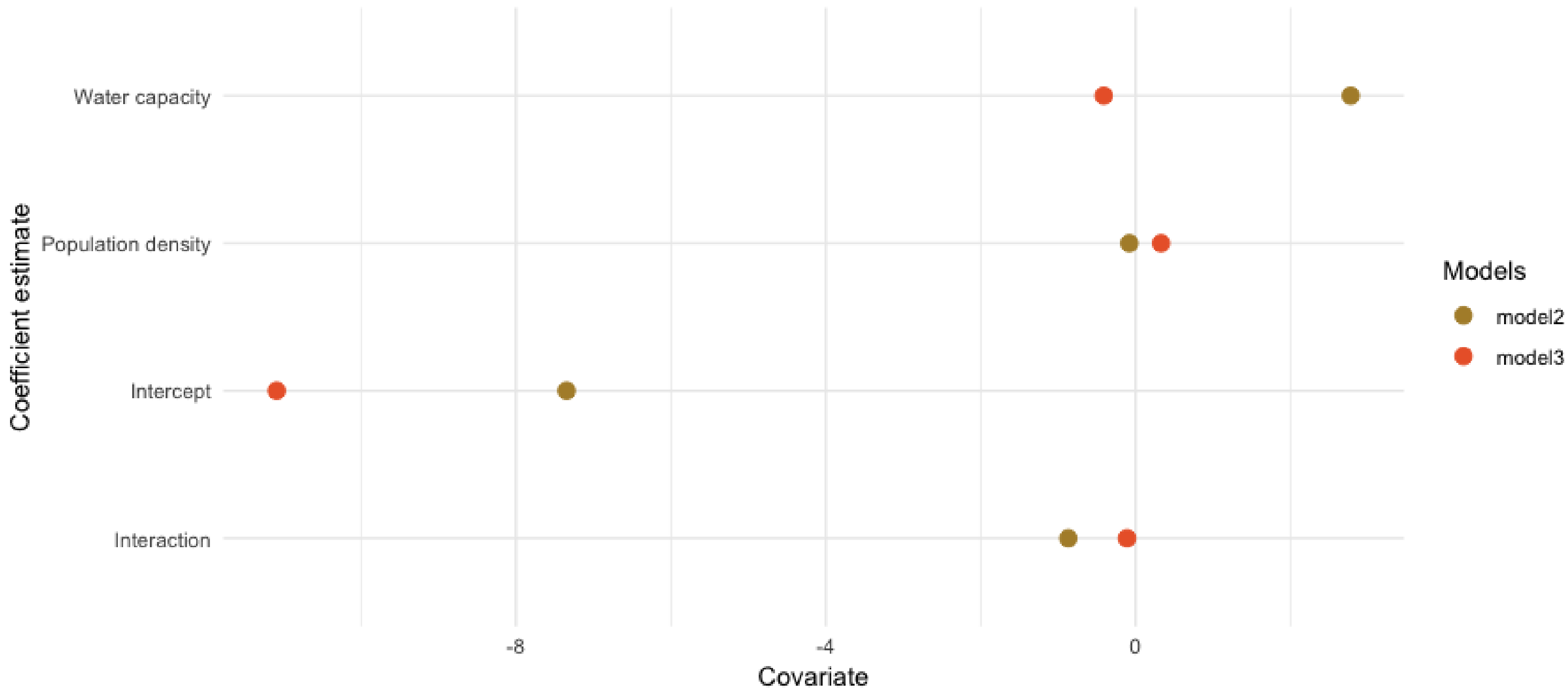
$$y \sim Pois(\lambda), \quad \text{where } \log(E[\boldsymbol{Y}|\boldsymbol{X}]) = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

a) Model with only water capacity as a covariate.

b) Model with water capacity and population density as covariates

d) Model with water capacity, pop density and an ICAR component (Stan)

Fitting different models



/

References

- Besag, J. (1974), *Spatial Interaction and the Statistical Analysis of Lattice Systems*, Journal of the Royal Statistical Society, Vol. 36, No. 2. (https://www.cise.ufl.edu/~anand/fa11/Besag_Spatial_interaction.pdf (https://www.cise.ufl.edu/%7Eanand/fa11/Besag_Spatial_interaction.pdf))

- Wheeler-Martin, Katherine; DiMaggio, Charles; Morris, Mitzi; Gelman, Andrew; Mooney, Stephen; Simpson, Daniel (2019), *"Bayesian Hierarchical Spatial Models: Implementing the Besag York Mollié Model in Stan"*, Spatial and Spatio-temporal Epidemiology, Vol 31, (http://dx.doi.org/10.17632/b5r4yztghx.2 (http://dx.doi.org/10.17632/b5r4yztghx.2))

- Morris, Mitzi, *Spatial Models in Stan: Intrinsic Auto-Regressive Models for Areal Data*, (https://mc-stan.org/users/documentation/case-studies/icar_stan.html (https://mc-stan.org/users/documentation/case-studies/icar_stan.html))

- Besag, Julian, and Charles Kooperberg.(1995) *On conditional and intrinsic autoregression.*, Biometrika.

Resources

- Spatial Data Science with R (https://rspatial.org/raster/index.html)
- CARBayes (https://cran.r-project.org/web/packages/CARBayes/vignettes/CARBayes.pdf)
- `spdep` (https://cran.r-project.org/web/packages/spdep/index.html)
- `sf` (https://r-spatial.github.io/sf/articles/sf1.html)

Thank you

- Mitzi Morris - Stan
- Evan Tachovsky - The Rockefeller Foundation
- Jim Savage - Schmidt Futures