Statistical modelling for infectious diseases

Part 1: Surveillance data and modelling foundation

Leo Bastos
PROCC/Fiocruz

Leonardo.bastos@fiocruz.br

@leosbastos

Summary

- Epidemiological Surveillance data
  - Descriptive analyses
- Foundation
  - Reporting uncertainty
  - Bayesian approach
- Predictive models
  - Usual model assumptions

Epidemiological Surveillance data

Time series

- Aggregated disease cases indexed by time (day, **week**, month)
- Sometimes extratified according to age groups and sex.

Spatial data

- Aggregated disease cases indexed by region (neighbourhoods, cities, states, countries)
- Spatio-temporal data is not unusal.

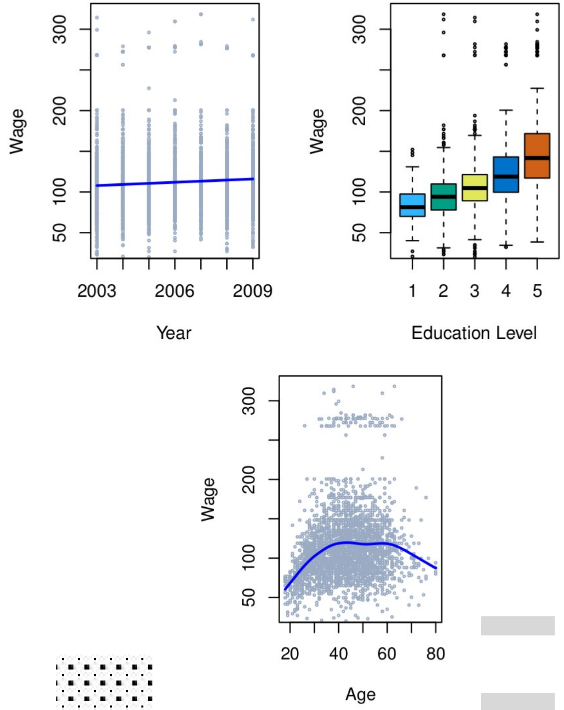
Individual level

- Information for each notified case might be available
- Usually administrative data (limited information)
- Missing information

Individual level data

Exploratory data analysis (Stats 101)

James et al. (2022, ISLR)



Variable	Overall, $N = 2501$	B.1.1.44, $N = 884$	Zeta, $N = 518$	Gamma, $N = 644$	Delta, $N = 455$
Median age in years (IQR)	21 (8-37)	19 (5-37)	25 (8-37)	22 (9-38)	17 (8-34)
Age categories	N (%)	N (%)	N (%)	N (%)	N (%)
0 to 4 years	408 (16%)	191 (22%)	79 (15%)	79 (12%)	59 (13%)
5 to 11 years	503 (20%)	167 (19%)	96 (19%)	126 (20%)	114 (25%)
12 to 17 years	250 (10.0%)	74 (8.4%)	37 (7.1%)	76 (12%)	63 (14%)
18 to 59 years	1192 (48%)	397 (45%)	268 (52%)	326 (51%)	201 (44%)
60 and older	148 (5.9%)	55 (6.2%)	38 (7.3%)	37 (5.7%)	18 (4.0%)
Female	1512 (60%)	544 (62%)	308 (59%)	379 (59%)	281 (62%)
SARS-COV-2 positive by RT-PCR	744 (12%)	139 (12%)	200 (17%)	287 (14%)	118 (7%)
SARS-CoV-2 seropositive	1479 (34%)	339 (33%)	199 (21%)	318 (21%)	623 (73%)

***Table 1*: Sociodemographic and virologic characteristics of the study participants ($N = 2501$). IQR = interquartile range.**

Carvalho et al. (2022, The Lancet Regional Health)

* Rarely available.

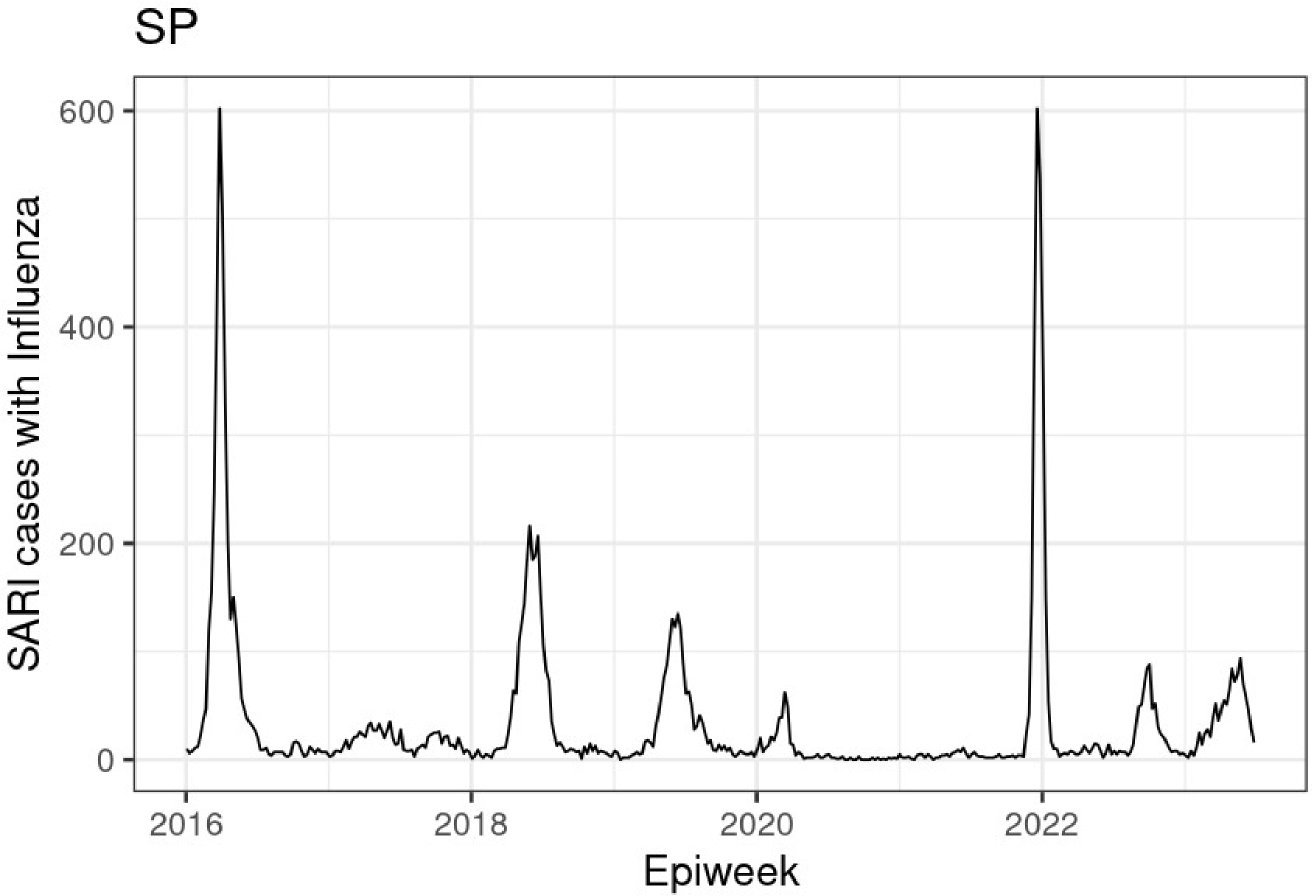
Time series

- Number of disease cases per unit of time

Y_t, X_t, N_t,,… $t = 1, 2, \ldots, T$.

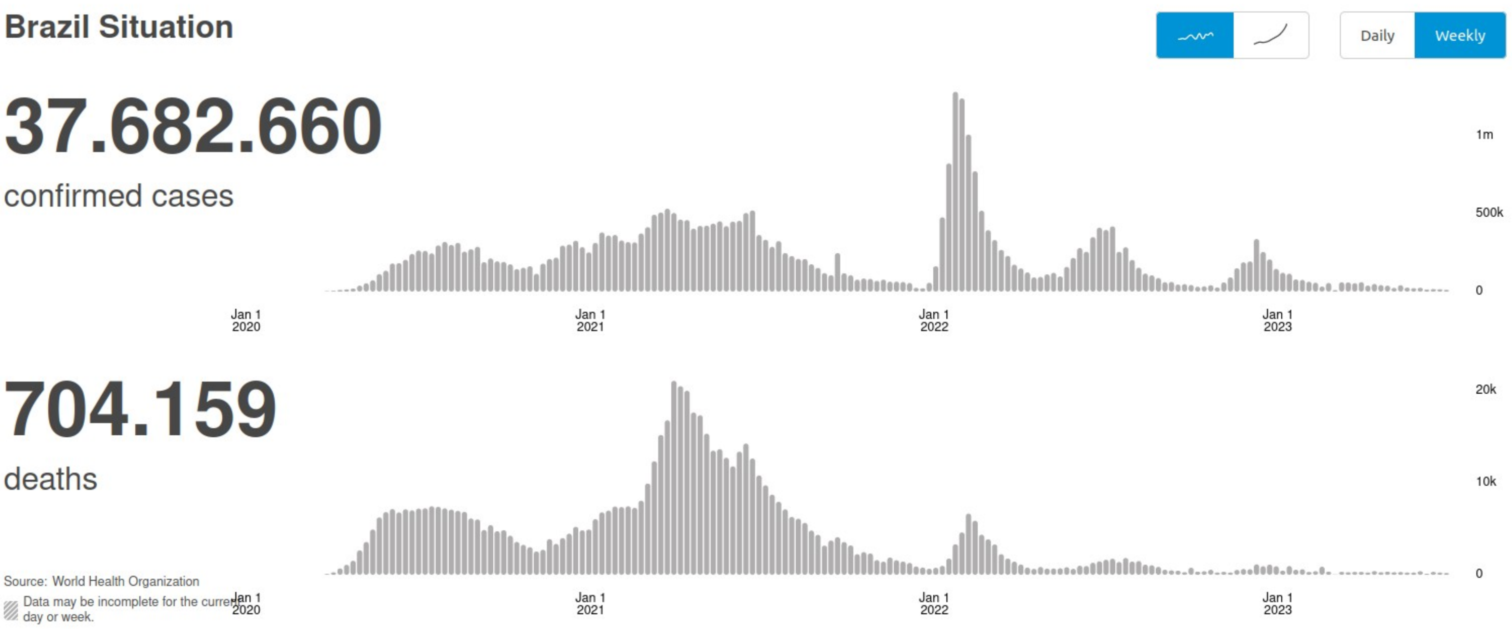
- Time could be **days**, **weeks**, months, years, ?
- There is some dependence among consecutive observations
- Most frequent (available) type of surveillance data

Time series


SP
600
400
200
0
SARI cases with Influenza
2016
2018
2020
2022
Epiweek

Time series

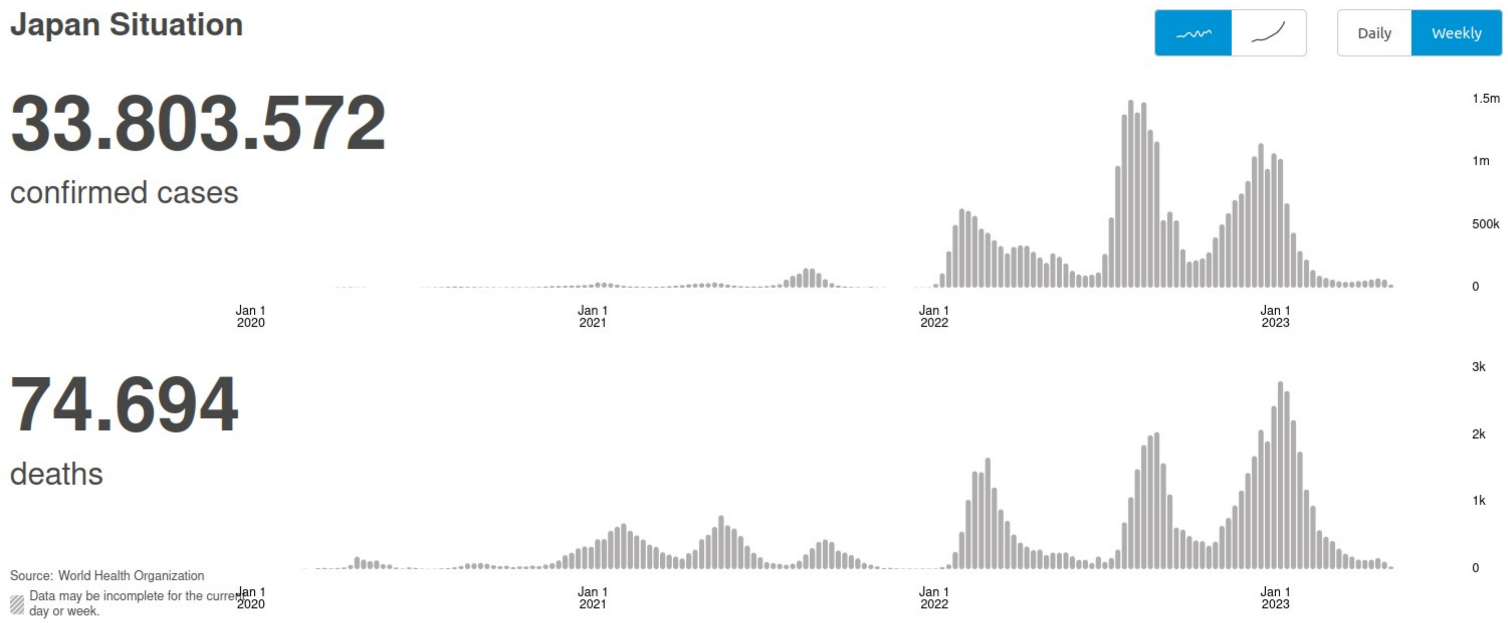
In **Brazil**, from **3 January 2020** to **12:14pm CEST, 12 July 2023**, there have been **37.682.660 confirmed cases** of COVID-19 with **704.159 deaths**, reported to WHO. As of **2 June 2023**, a total of **513.329.718 vaccine doses** have been administered.



https://covid19.who.int

Time series

In **Japan**, from **3 January 2020** to **12:14pm CEST, 12 July 2023**, there have been **33.803.572 confirmed cases** of COVID-19 with **74.694 deaths**, reported to WHO. As of **6 June 2023**, a total of **392.346.325 vaccine doses** have been administered.



https://covid19.who.int

Time series

- TS can be used to describe disease dynamics
- We may also help us to identify patterns
  - Seasonality
  - Trends
  - Disease natural history (e.g. TS by age groups)

Time series

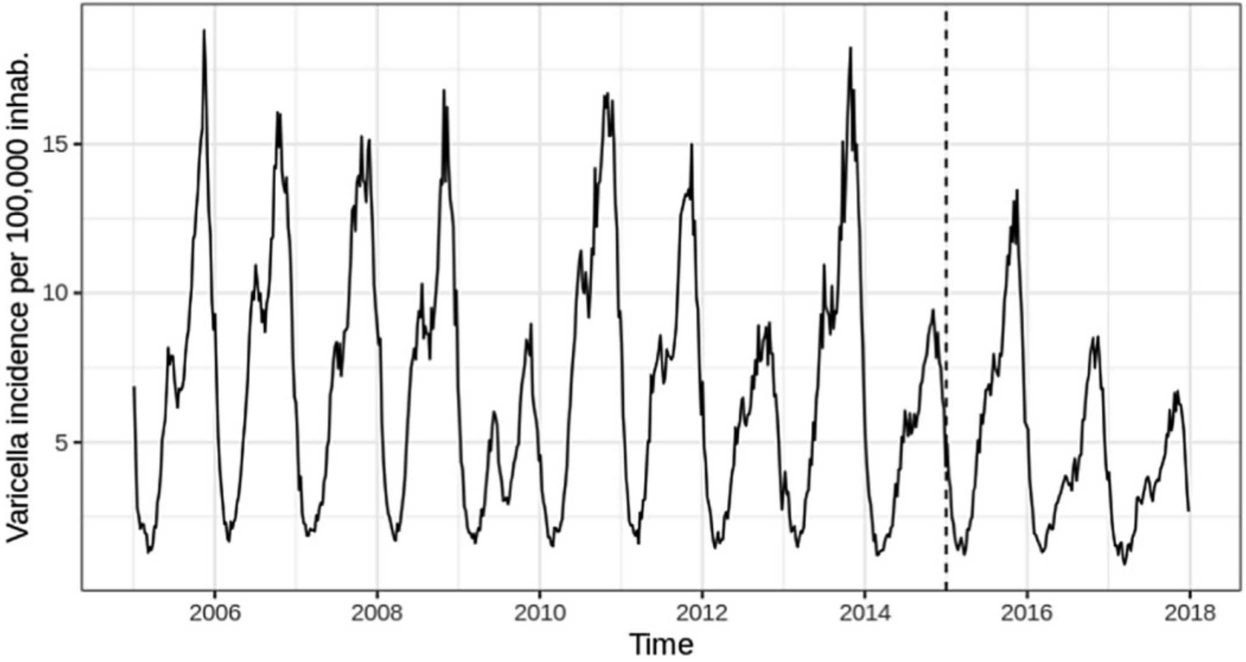
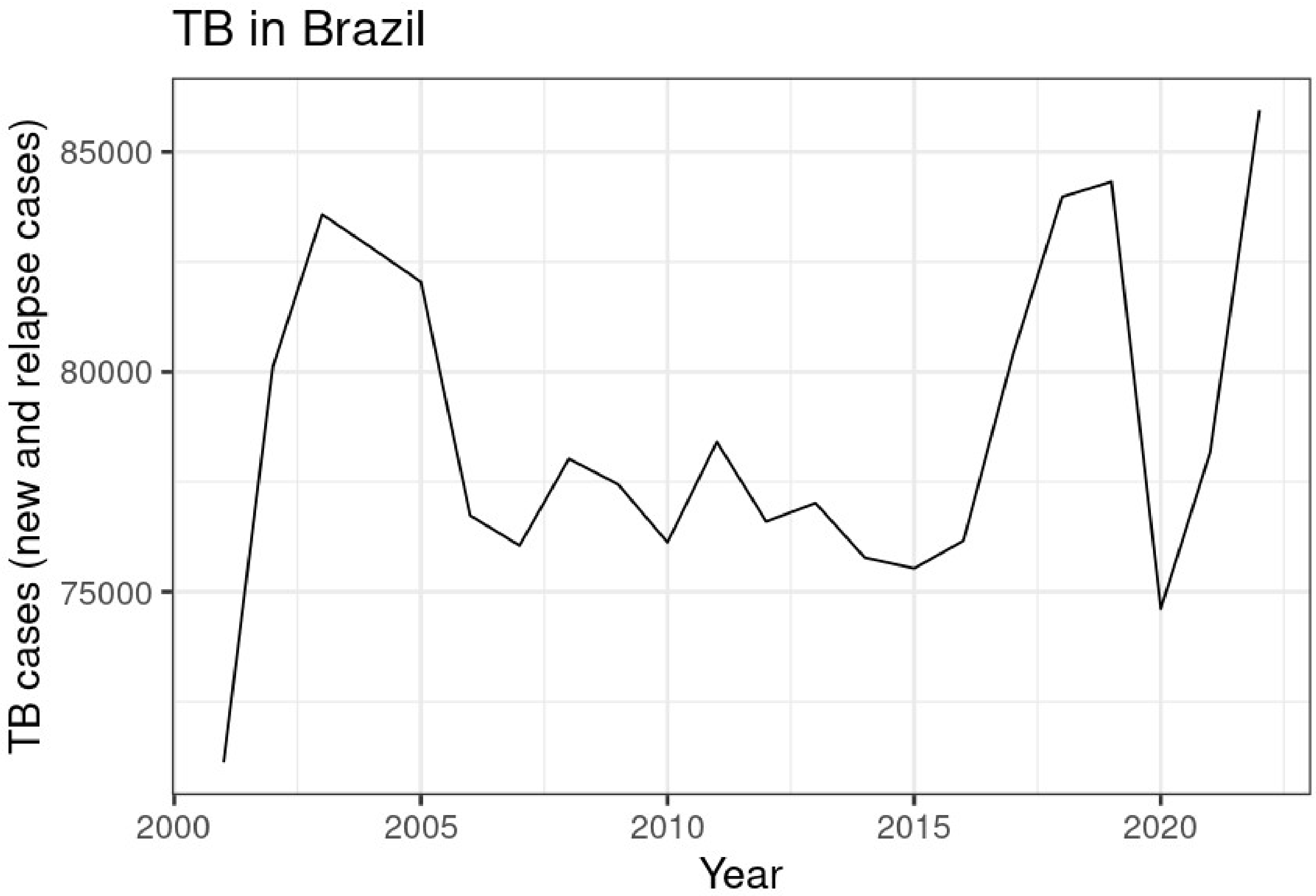
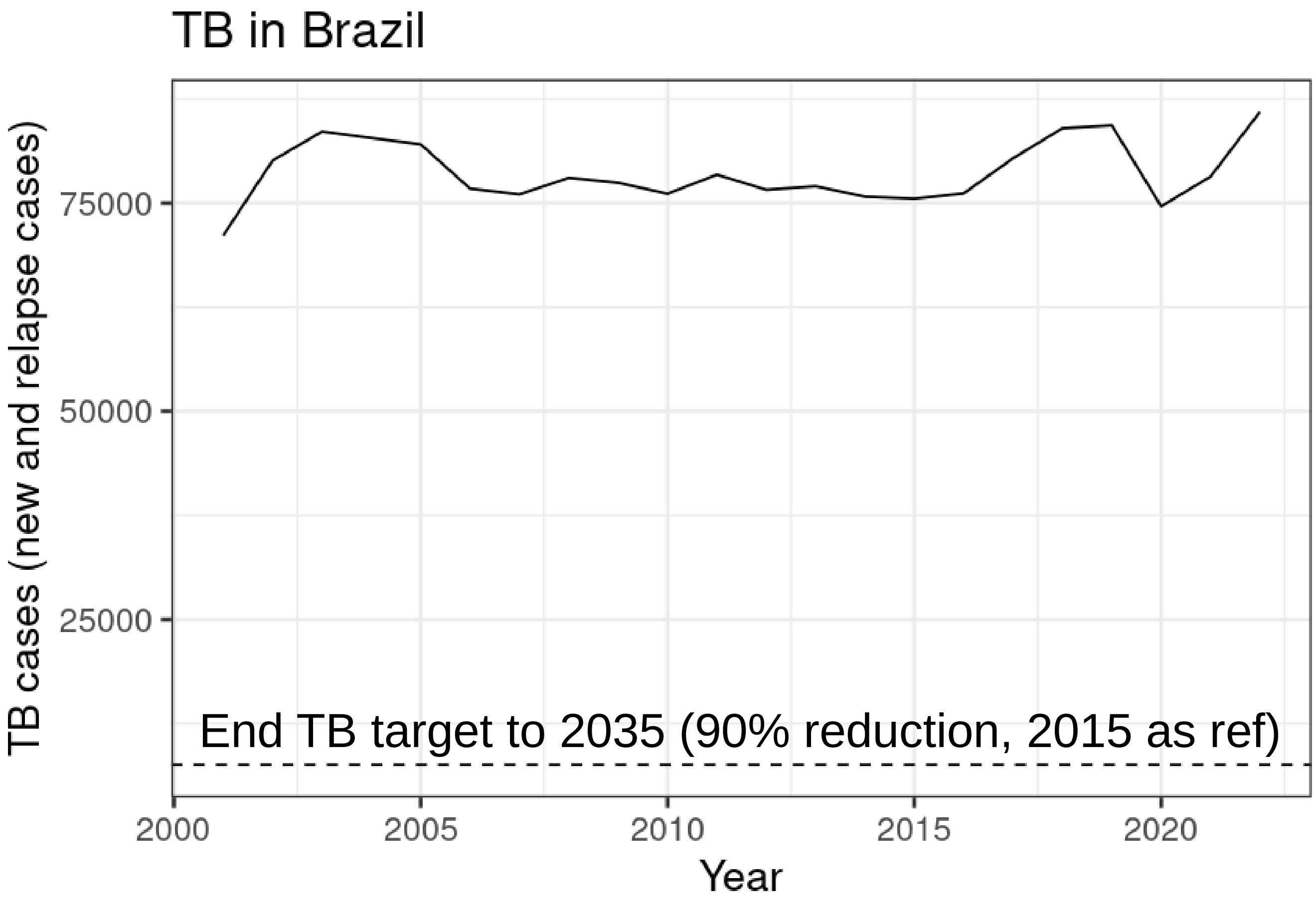


Fig. 1. Weekly varicella reported cases in Argentina from 2005 to 2017. The vertical dotted line indicates the beginning of the period when a single dose varicella vaccine become universally available to 15 month old children.

Time series


TB in Brazil
TB cases (new and relapse cases)
85000
80000
75000
2000
2005
2010
2015
2020
Year

Time series

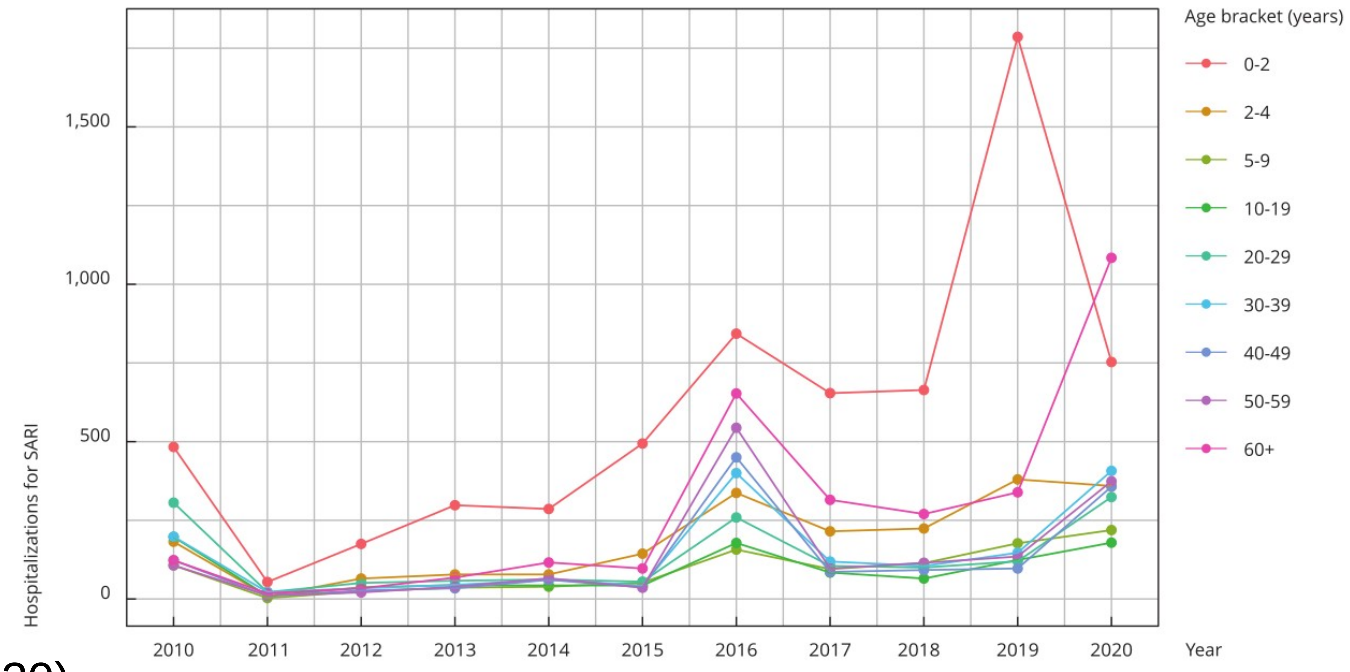


TB in Brazil
TB cases (new and relapse cases)
75000
50000
25000
End TB target to 2035 (90% reduction, 2015 as ref)
2000
2005
2010
2015
2020
Year

Time series

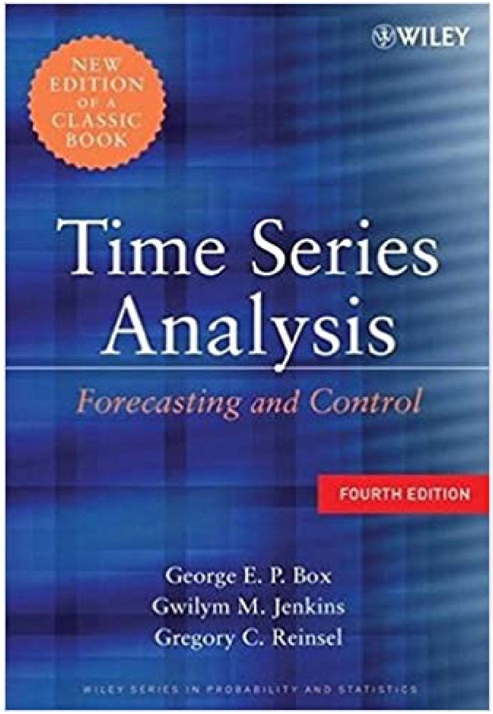
Figure 2

Absolute numbers of cases of hospitalizations for severe acute respiratory illness (SARI) in Brazil from the 9th to 12th epidemiological weeks in years 2010 through 2020, stratified by age brackets.

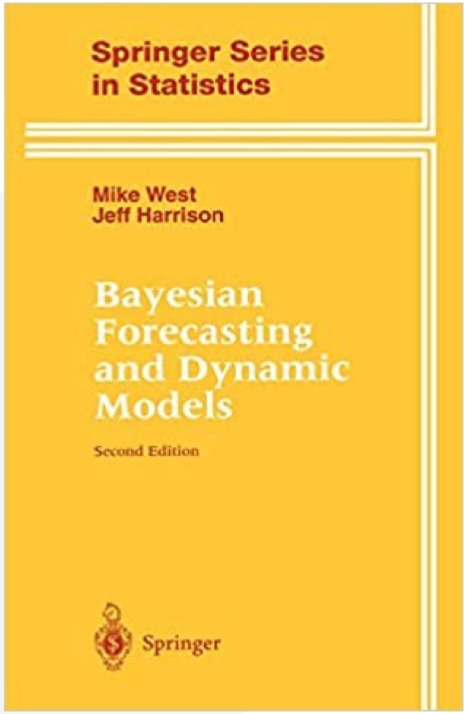


Bastos et al. (2020)

Time series books



Box, Jenkins, and Reinsel



West and Harrison

$$Y_t \sim F(\mu_t, \phi)$$

$$\mu_t = s(t, x_t, \theta)$$

$$t = 1, 2, \ldots, T$$

Spatial surveillance data

- Number of disease cases per region

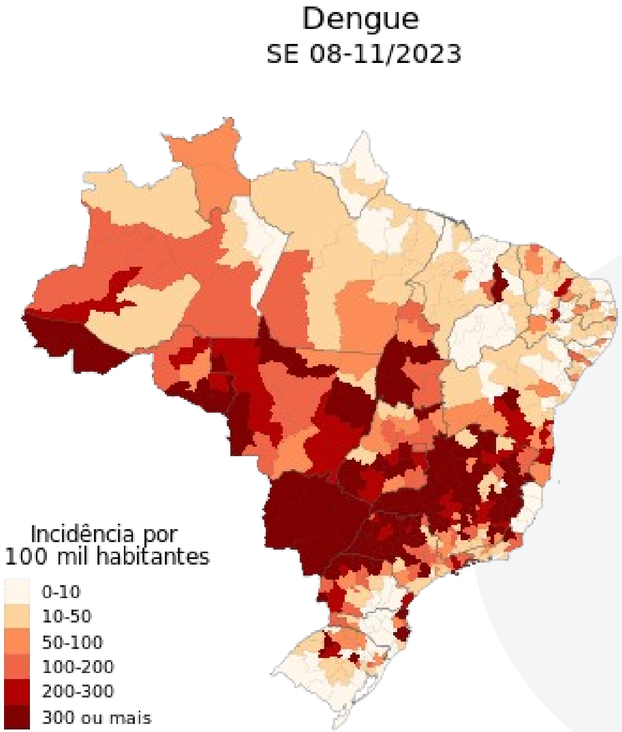
$Y_r, X_r, N_r,, \ldots \qquad r = 1, 2, \ldots, R.$

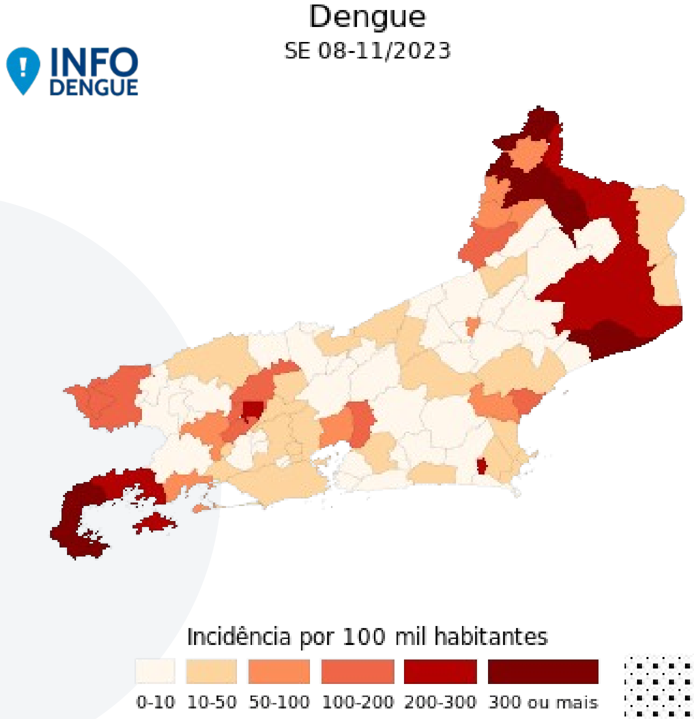
- Regions are usually neighbourhoods, **cities**, countries
- “Everything is related to everything else, but near things are more related than distant things” – Tobler

Spatial surveillance data

- There are three types of spatial data:
  - **Discrete area data** (The variable of interest occur in a well-defined region)
  - Continuous spatial data (The variable of interest can be measured anywhere over the region of interest)
  - Point process (We are interested in where the event occur)

Spatial surveillance data





Spatial surveillance data

- In area data, is common to use the neighbourhood matrix, usually a binary matrix in the form:

$$\mathrm{W}_{i,j} = \begin{cases} 1, & \text{if } i \sim j, \\ 0, & \text{otherwise.} \end{cases}$$

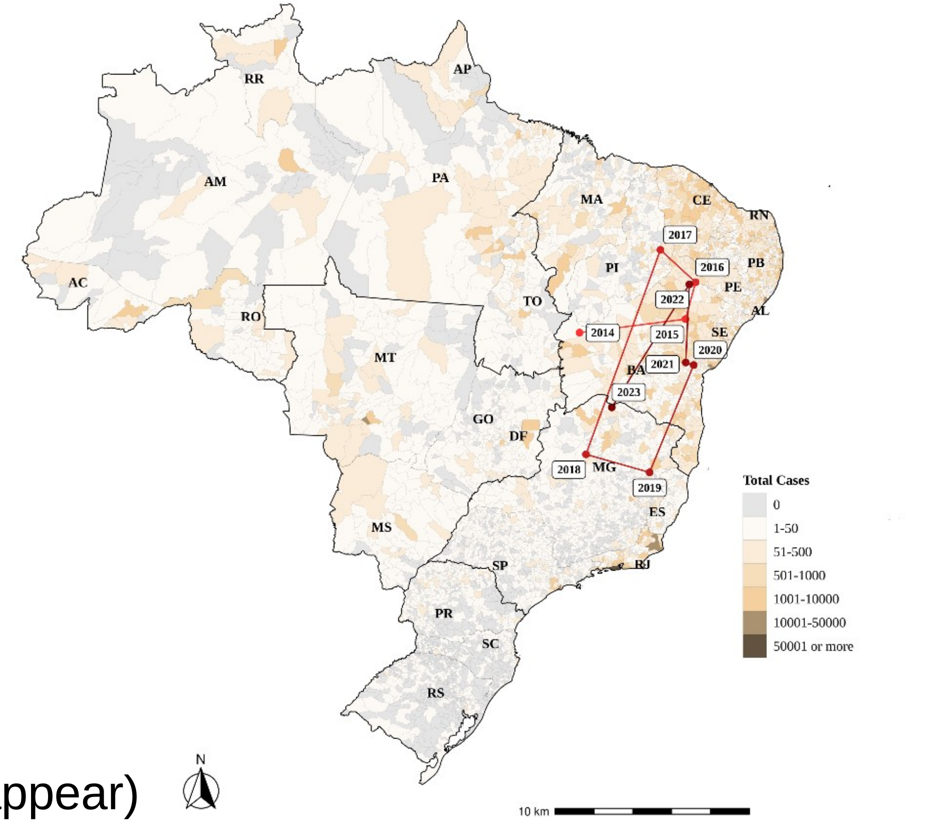
The W matrix could be used to smooth estimates or induce dependence in a model

Spatial surveillance data

- A different approach would be consider each region as a point in space, and analyse as point processes.
- E.g. calculating centroids or finding clusters

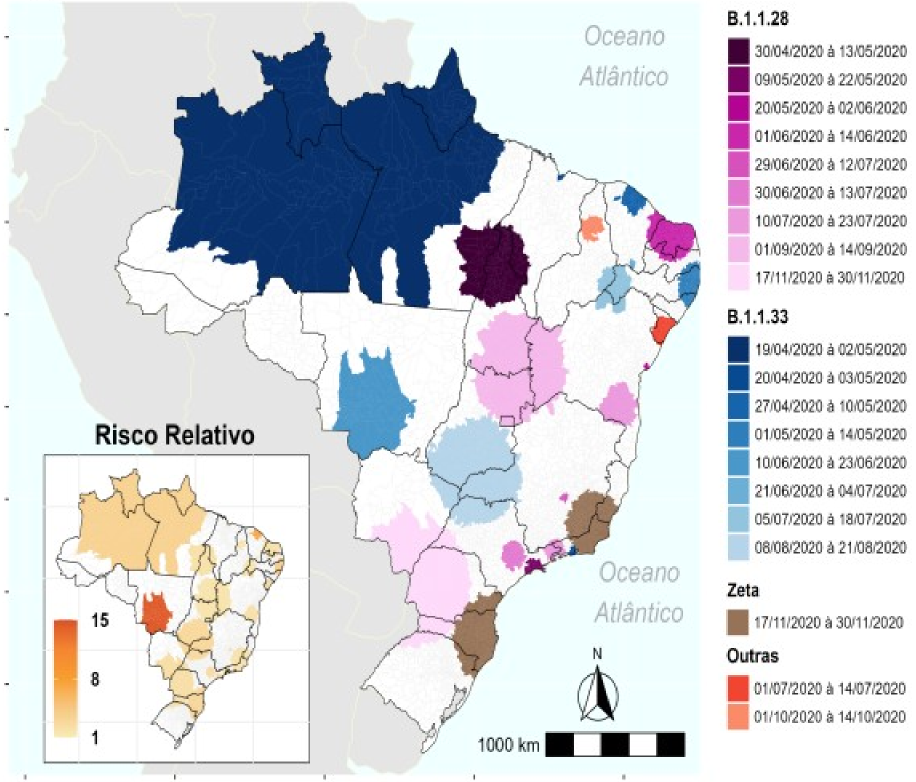
Spatial surveillance data

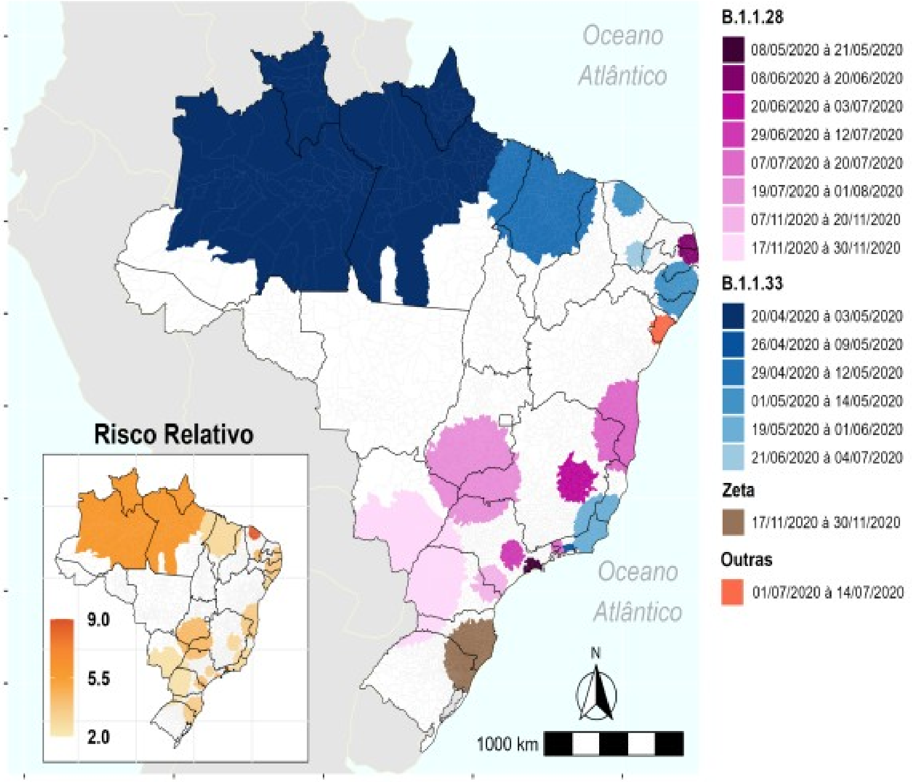
Chikungunya cases in Brazil, 2014-2023



Almeida et al. (2023, to appear)

Spatial surveillance data





Bianchi (2023)

Foundation

- Managing uncertainty
- Probabilistic approach
  - Likelihood, prior and posterior
  - Predictive distribution
- Inference methods

Managing uncertainty

- In surveillance data, there is plenty of uncertainty sources
  - What is/was/will be the number of cases of disease x at time t in region r?
  - Are we facing an epidemic? How far we are from the expected?
  - What was the impact of an intervention I? Did it reduce the number of deaths?

Managing uncertainty

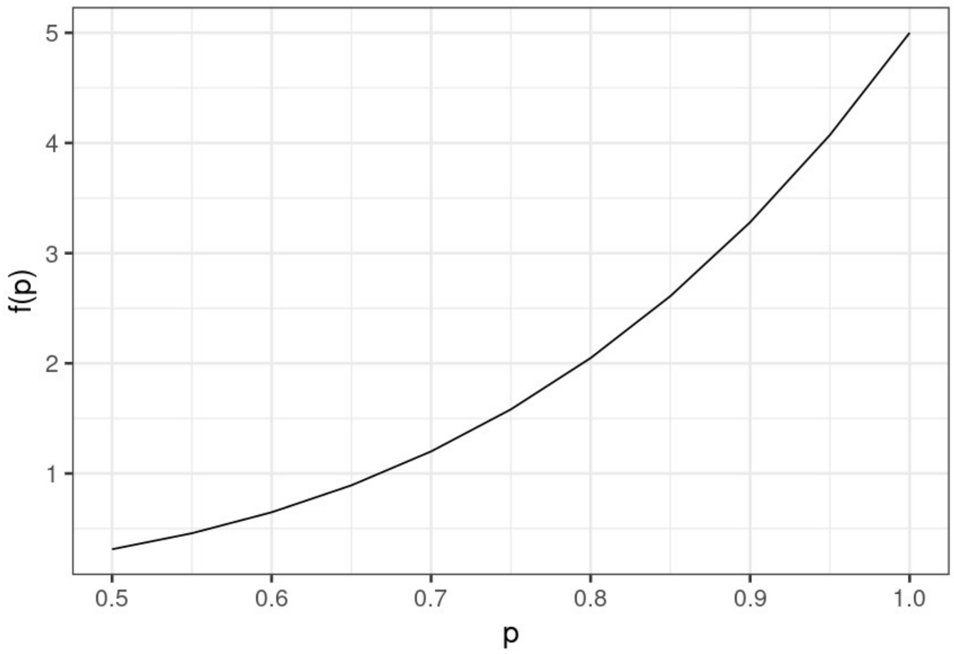
- Those question are uncertain, and we can (try to) answer them with aid of probability methods.
- In a probabilistic perspective, everything that is unknown can be represented using a probability distribution.
- This perspective is also called Bayesian perspective.

Example: COVID-19 prevalence

- I am going to elicit (built a probability distribution) of the COVID-19 prevalence of people in this room, p.
  - How likely it is? Can I define some probabilities?
    - $P(p < 0.1)$
    - $P(p < 0.5)$
    - $P(p < 0.9)$

What about mode? Mean?

Example: COVID-19 prevalence



P(p < 0.1) ~ 0
P(p < 0.5) = 0.03
P(p < 0.9) = 0.57

Mode = 1
Mean = 5/6

$$p \sim Beta(5, 1)$$

Prior

- We built a prior distribution for the COVID-19 prevalence for this audience
- We can build/elicit prior distributions for any numerical quantities that are unknown
- There are non-informative or weakly informative priors when we know little about a quantity

Likelihood / the model

- We usually try to describe our main outcome as a parametric probability distribution
- For number of cases (a counting process), we may use:

$$Y_t \sim Poisson(\theta_t)$$

$$g(\theta_t) = x_t^T \beta$$

$$Y_t \sim NegBinom(\theta_t, \phi)$$

$$g(\theta_t) = x_t^T \beta + \delta_t$$

Likelihood / the model

- The most commonly used statistical models assume independence among observations
- Then in a Poisson model

$$Y_t \sim Poisson(\theta_t)$$

$$g(\theta_t) = x_t^T \beta$$

$$L(\beta) = \prod_t p(y_t | x_t, \beta)$$

Likelihood / the model

- However, independence may be a very strong assumption (specially in the context of infectious disease)
- So we should try a different model that takes into account the dependence structure

The model

- We could use a property called conditional independence
- Given some parameter the Ys can be independent.
- So, one possible model is

$$Y_t \sim Poisson(\theta_t)$$

$$g(\theta_t) = x_t^T \beta + \delta_t$$

$$\delta_t \sim N(\delta_{t-1}, \tau_\delta^2)$$

$$L(\beta, \delta) = \prod_t p(y_t | x_t, \beta, \delta_t)$$

$$p(\delta_0, \tau_\delta^2)$$

The model

- That model is a Random effects Poisson model
- A particular case of a Bayesian generalised linear mixed model, GLMM

$$Y_t \sim Poisson(\theta_t)$$

$$g(\theta_t) = x_t^T \beta + \delta_t$$

$$\delta_t \sim N(\delta_{t-1}, \tau_\delta^2)$$

$$L(\beta, \delta) = \prod_t p(y_t | x_t, \beta, \delta_t)$$

$$p(\delta_0, \tau_\delta^2)$$

Posterior

- Combining the prior distributions and the likelihood leads to a distribution called posterior distribution
- Bayes theorem, assume two events A and B, in stats 101 we learn that

$$P(B \mid A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Posterior

- Lets supose the sample space of B could be partioned in M+1 events C_i, and B is just one of them, for simplicity lets say B= C_0.

$$P(B \mid A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{\sum_{i=0}^{M} P(A|C_i)P(C_i)}$$

Posterior

- If B is our unknown parameter, and A is our observed data. Then

$$P(\theta \mid y) = \frac{P(y|\theta)P(\theta)}{P(y)} = \frac{P(y|\theta)P(\theta)}{\sum_\theta P(y|\theta)P(\theta)}$$

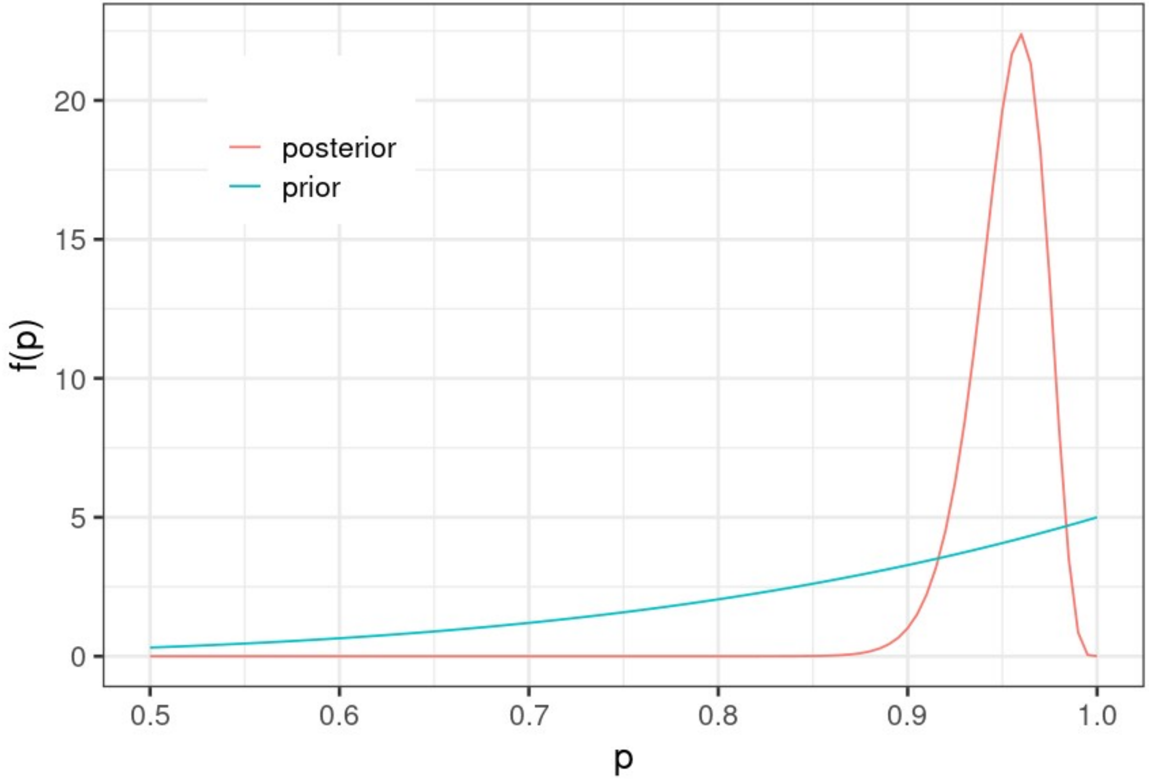
Posterior

- Our unknown parameter is usually continuous, and we have a sample of observed data

$$p(\theta \mid y) = \frac{p(\theta)\prod_i p(y_i|\theta)}{\int_\theta p(\theta)\prod_i p(y_i|\theta)d\theta}$$

$$\propto p(\theta)\prod_i p(y_i|\theta)$$

Example: COVID-19 prevalence



My guess: n = 120; y = 5

P(p < 0.1 | y) ~ 0
P(p < 0.5 | y) ~ 0
P(p < 0.9 | y) = 0.01

Mode = 0.960
Mean = 0.952

$$p \mid y \sim Beta(5 + y, n - y + 1)$$

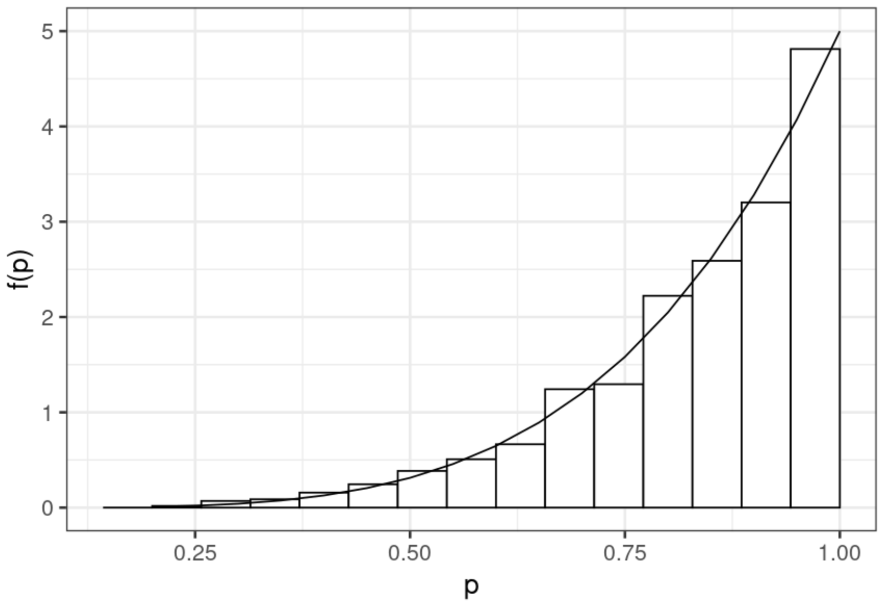
So what?

- How can we learn about the parameters?
- Can we solve that integral for complex models?
- Yes². Using Bayesian computation!

$$p(\theta \mid y) = \frac{p(\theta)\prod_i p(y_i|\theta)}{\int_\theta p(\theta)\prod_i p(y_i|\theta)d\theta}$$

Bayesian comp: Monte Carlo

- Basic idea, solve integral by sampling from the target distribution



$$\mathbb{E}[p] = \int_{p=0}^{1} p f(p) dp$$

$$\mathbb{E}[p] = \frac{5}{5+1} = 5/6 = 0.833$$

$$\mathbb{E}[p] \approx \sum_{k=1}^{M} \frac{p_k}{M} = 0.840$$

Bayesian comp: MCMC

- Basic idea: we don’t know how to sample directly so we sample from the **full conditionals** iteractively using Markov chain properties
- The samples eventually converge to samples from the full posterior.

Bayesian comp: MCMC

Initialize the chain $(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_P^{(0)})$

For k in 1:M (Monte Carlo step){

For j in 1:P (Parameter space){

Sample $\theta_p^{(k)}$ from $p(\theta_p \mid \theta_1^{(k)}, \ldots, \theta_{p-1}^{(k)}, \theta_{p+1}^{(k-1)}, \ldots, \theta_P^{(k-1)})$

}

}

Set a **BurnIn** and a **lag** to get your final MCMC sample

Bayesian comp: MCMC

- Gibbs sampling
- Metropolis-Hasting algorithm
- Slice sampler
- Halmitonian Monte Carlo

Bayesian comp: MCMC

- But, there are some good news!
- Those methods are already implemented in a set of MCMC packages/softwares

Old ones:
- WinBUGS
- OpenBUGS

NIMBLE

JAGS

https://mc-stan.org/

https://r-nimble.org/

https://mcmc-jags.sourceforge.io/

Bayesian comp: MCMC

- PROS
  - Just need data, priors and likelihood (the model)
  - Works for simple to very complex models
  - Can answer using probabilities
- CONS
  - Computational cost*

Bayesian comp: No MCMC

- There are other approximations
  - Variational Bayesian methods
  - Integrated Nested Laplace Approximation (INLA)

https://www.r-inla.org/

$$p(\theta \mid y) \approx \tilde{p}(\theta \mid y)$$

Predictions

- Can we make predictions?
- In this probabilistic approach, the values to be predicted are unknown, so they are treated as unknown parameters.
- Essentially we want

$$p\left(y^{(new)} \mid y^{(obs)}\right)$$

Predictions

- Mathematically

$$
\begin{aligned}
p(y^{(new)} \mid y^{(obs)}) &= \int_\theta p(y^{(new)}, \theta \mid y^{(obs)}) d\theta \\
&= \int_\theta p(y^{(new)} \mid \theta, y^{(obs)}) p(\theta \mid y^{(obs)}) d\theta \\
&= \int_\theta p(y^{(new)} \mid \theta) p(\theta \mid y^{(obs)}) d\theta
\end{aligned}
$$

COVID-19 prevalence example

- A new participant has just arrived, what is the probability of he/she being a prevalent COVID-19 case?

$$p(y^{(new)} = 1 \mid y^{(obs)}) = \int_0^1 p(y^{(new)} = 1 \mid 1, p)p(p \mid 120, 6)dp$$

$$= 0.952$$

Varicella in Argentina

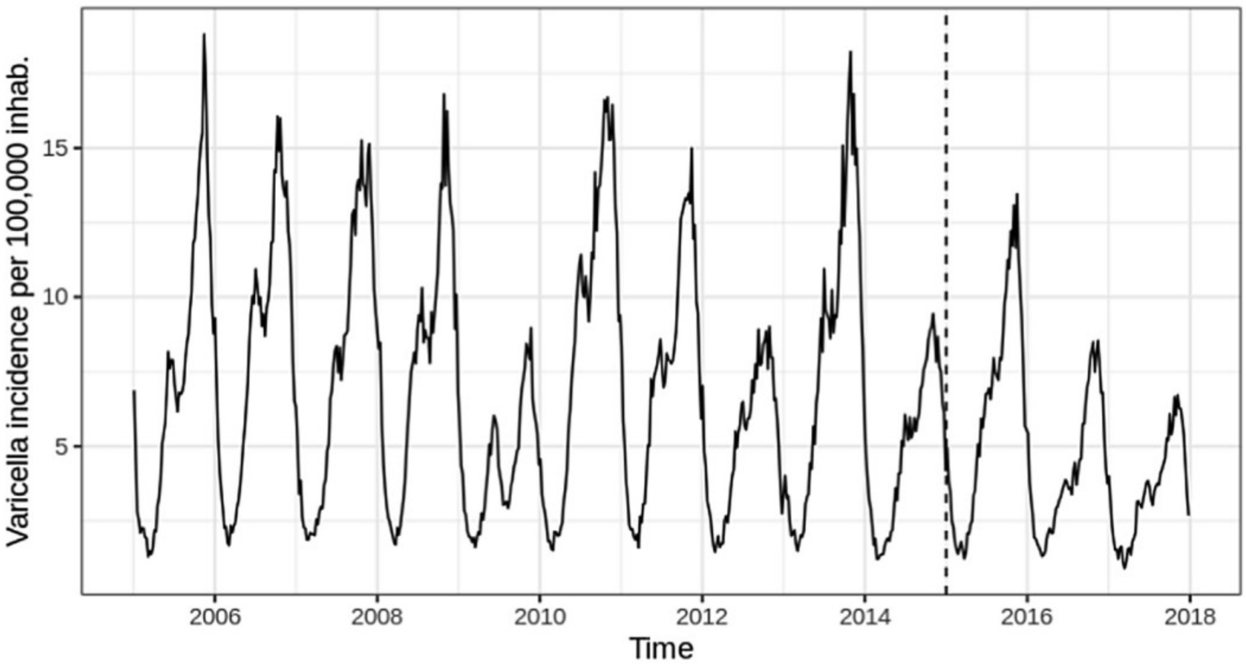


Fig. 1. Weekly varicella reported cases in Argentina from 2005 to 2017. The vertical dotted line indicates the beginning of the period when a single dose varicella vaccine become universally available to 15 month old children.

Varicella in Argentina

$$Y_t \mid \lambda_t, \phi \sim NegBin(\lambda_t, \phi)$$

$$\log(\lambda_t) = \alpha_{week[t]} + \beta_{year[t]}$$

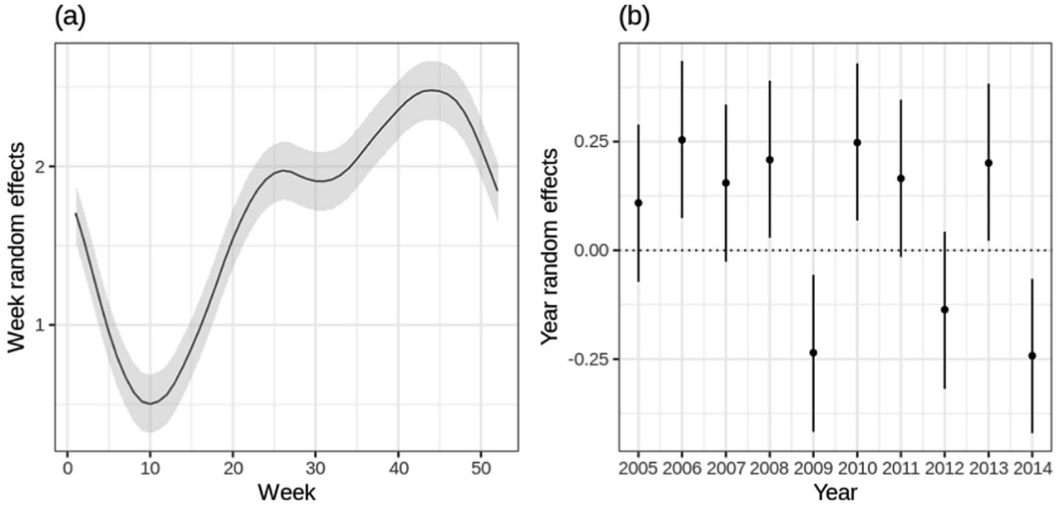


Fig. 2. Estimated random effects. (a) Week random effects; (b) Year random effects.

Varicella in Argentina

$$p(y_1^{new}, y_2^{new}, \ldots, y_{52}^{new} | data)$$