Statistical methods for epidemiological surveillance

Leo Bastos (PROCC/Fiocruz) and Luiz Max Carvalho (FGV/EMAp)

July 5, 2022

Outline

(Infectious) Disease process

Figure 1.1 *Schematic of disease evolution.*

1. Start of etiologic process
   - Ex: Infection through a mosquito bite (dengue fever)
2. Disease begins
   - Viral replication
3. Clinical symptoms
   - Fever, headaches,...
4. Outcome of disease
   - Cure, hospitalization, death

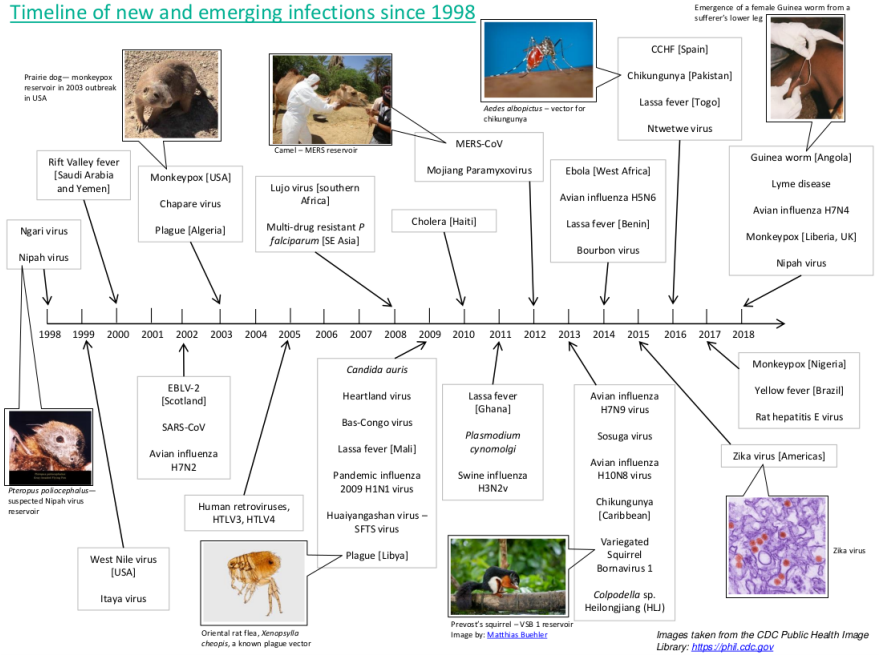
Epidemiological study of Disease process

1. Start of etiologic process
   - Can we avoid infection?
2. Disease begins
   - Is our immune system prepared?
3. Clinical symptoms
   - Can we treat or avoid evolution?
4. Outcome of disease
   - How to reduce the burden?

Infectious disease epidemiology (IDE)

- Infectious disease epidemiology (IDE) is the study of how and why infectious diseases emerge and spread among different **populations**, and what strategies can prevent or contain the spread of disease at the population level.
- Why is this important?

Timeline of new and emerging infections since 1998



Data type: In Infectious Disease Epidemiology

- Infectious disease data is mainly binary

$$Z_i = \begin{cases} 1 & \text{person } i \text{ is infected with pathogen A or has a disease D,} \\ 0 & \text{otherwise.} \end{cases}$$

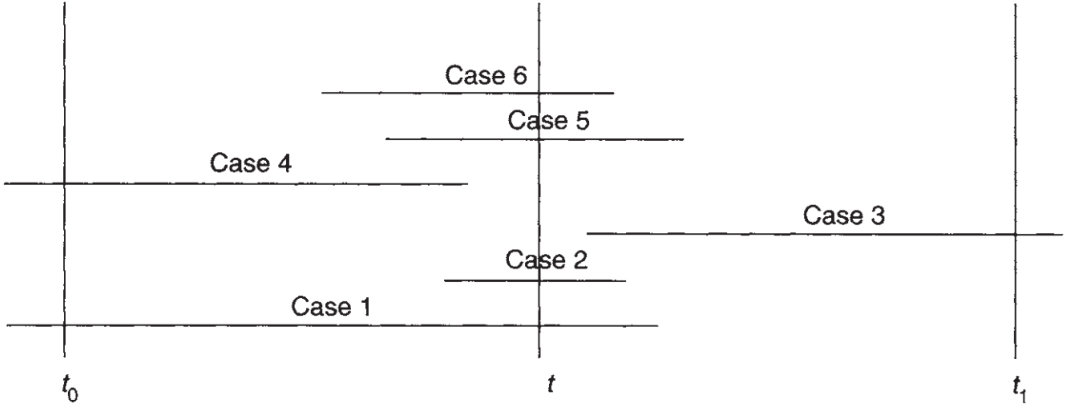
- Counts of cases through time

$$Y_t = \sum_i Z_{i,t}$$

- Induce time and spatial dependence is important, since dependence is present by definition of infectious diseases

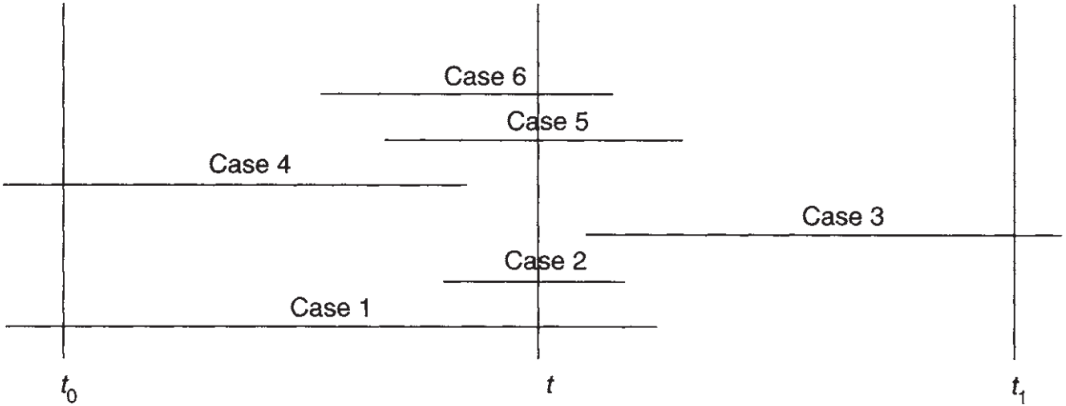
Important measures of disease occurrence

- Disease prevalence and incidence both represent proportions of a population determined to be diseased at certain times.
- Suppose there are 100 people being followed.



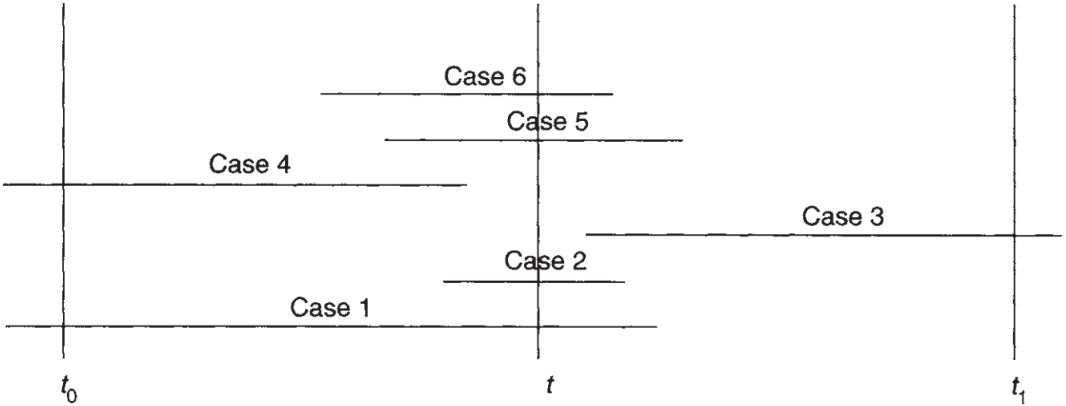
Important measures of disease occurrence

- **Disease prevalence** at time t: 4 / 100 or 4 / 99

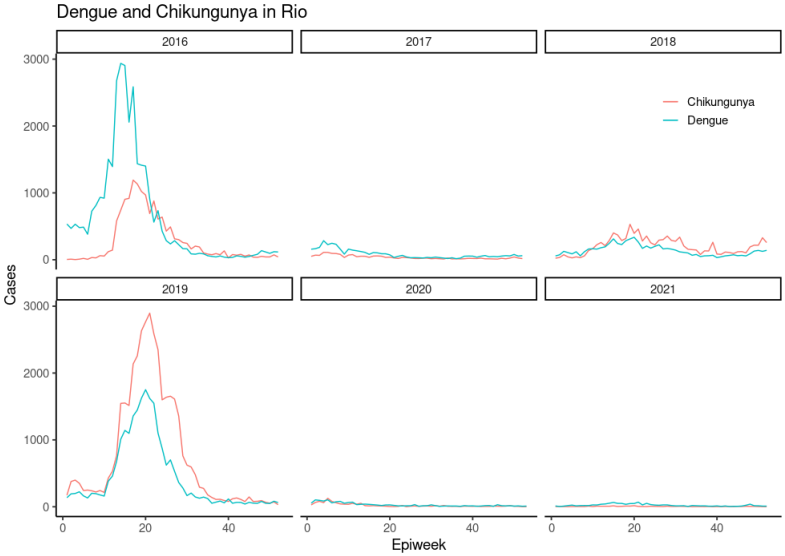


Important measures of disease occurrence

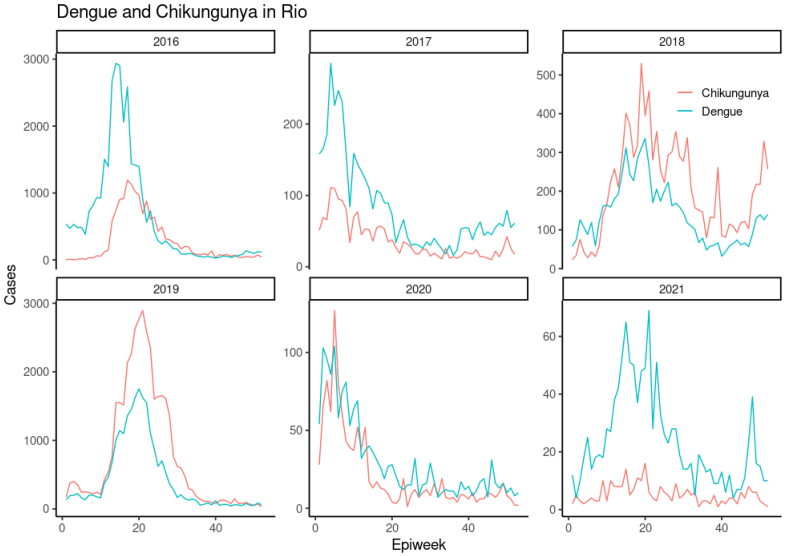
- **Disease incidence** (new cases) in time $[t_0, t_1]$: 4 / 98



Dengue and Chikungunya new cases in Rio de Janeiro



Dengue and Chikungunya new cases in Rio de Janeiro



Outbreak, epidemic, pandemic, endemic

- **Outbreak**: an unexpected increase in the number of disease cases in a small area.
  - Ex: high number of cases of ILI at Vidigal, Rio de Janeiro (Nov/21).
- **Epidemic**: an unexpected increase in the number of disease cases in a specific geographical area.
  - Ex: high number of cases of ILI in several neighbourhoods of Rio de Janeiro (Dec/21).
- **Pandemic**: Exponential growth of the number of cases; Several outbreaks/epidemics spread around the globe.
  (Influenza/H1N1, COVID-19)
- **Endemic**: When a disease is consistently present but limited to a particular region..
  (Malaria in Brazil's North region)

Modelling outbreaks

- Counting data time series models are important tools to model outbreaks and epidemics.
- the usual models have a low predictive power unless there is some knowledge about the disease dynamics, and if there is historical data (a problem for emerging diseases)
- Endemic diseases with a large data history are those possible to forecast. The long term forecasts could be seen as the expected behaviour (real time observed values greater than the expected values is a strong suggestion of an outbreak).

Defining epidemic thresholds

DOI:10.1111/j.1750-2659.2012.00422.x
www.influenzajournal.com
Original Article

Influenza surveillance in Europe: establishing epidemic thresholds by the Moving Epidemic Method

Tomás Vega,[a] Jose Eugenio Lozano,[a] Tamara Meerhoff,[b] René Snacken,[c] Joshua Mott,[d] Raul Ortiz de Lejarazu,[e] Baltazar Nunes[f]

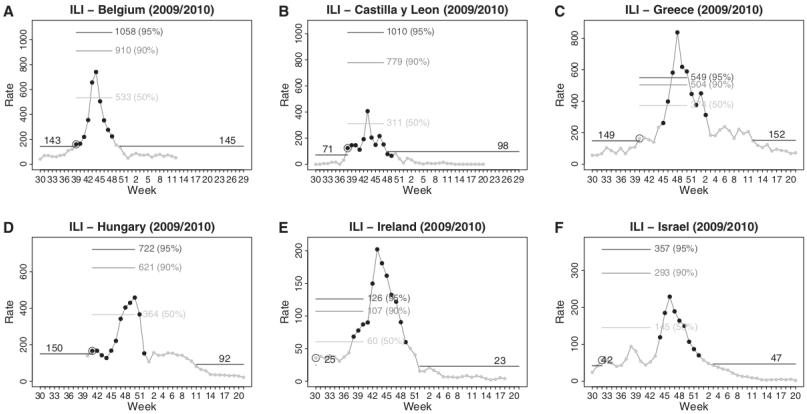
[a]Dirección General de Salud Pública, Consejería de Sanidad, Valladolid, Spain. [b]The Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [c]European Centre for Disease Prevention and Control, Stockholm, Sweden. [d]Division of Health Security, Infectious Diseases and the Environment, WHO Regional Office for Europe, Copenhagen, Denmark. [e]Centro Nacional de Gripe de Valladolid, Universidad de Valladolid, Spain. [f]Instituto Nacional de Saude Doctor Ricardo Jorge, Lisboa, Portugal.
Correspondence: Tomás Vega, Dirección General de Salud Pública, Consejería de Sanidad, Paseo de Zorrilla, 1. 47071 Valladolid, Spain. E-mail: vegaloto@jcyl.es

Accepted 27 June 2012. Published Online 16 August 2012.

MEM: Moving Epidemic Method

- Developed by Vega et al. (2012) focused on Influenza
- MEM is a three-step algorithm:
  1. Dertermine start, duration, and end of the epidemic period;
  2. Estimate epidemic thresholds and epicemic chanels defining pre- and post-epidemic levels.

Epidemic thresholds



Epidemic thresholds

R: The Moving Epidemic Method Find in Topic

The Moving Epidemic Method

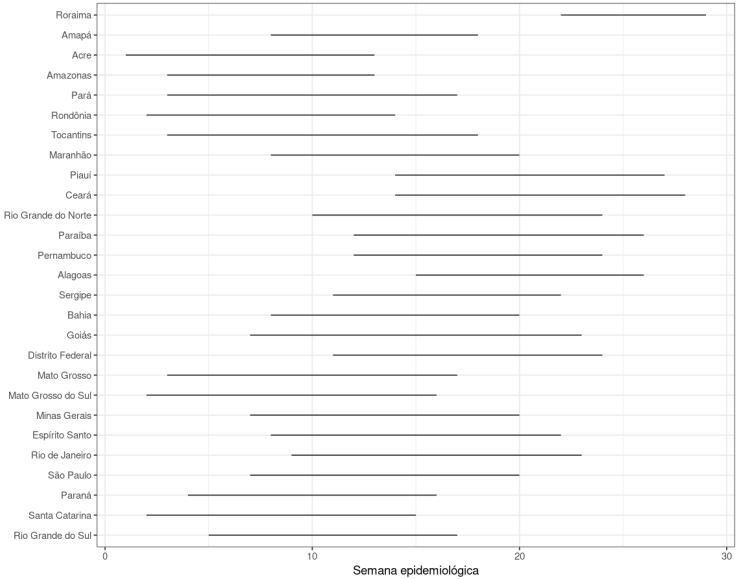
Documentation for package ‘mem’ version 2.16

- DESCRIPTION file.

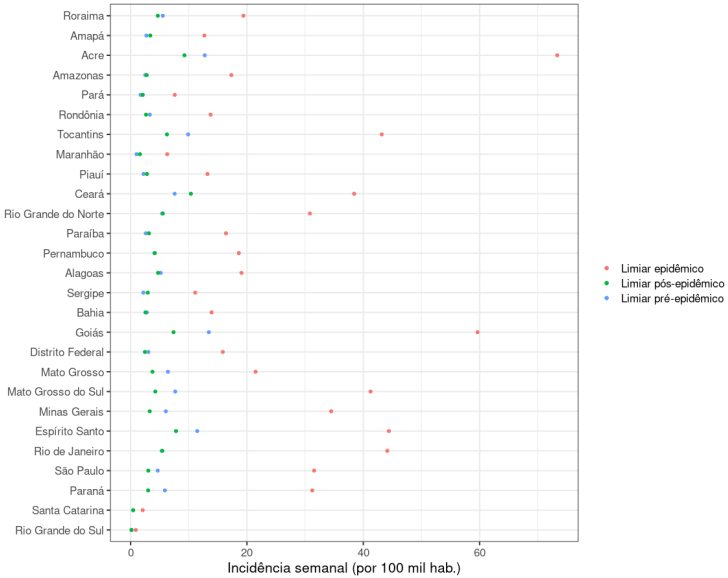
Help Pages

epimem	Deprecated function(s) in the mem package
epitiming	Deprecated function(s) in the mem package
flucyl	Castilla y Leon influenza crude rates
flucylraw	Castilla y Leon influenza standarised rates
full.series.graph	Creates the historical series graph of the datasets
memevolution	Evolution of estimators
memgoodness	Goodness of fit of the mem
memintensity	Thresholds for influenza intensity
memmodel	Methods for influenza modelization
memstability	Stability of indicators
memsurveillance	Creates the surveillance graph of the current season
memsurveillance.animated	Creates the animated graph of the surveillance of the current season
memtiming	Influenza Epidemic Timing
memtrend	Methods for influenza trend calculation
optimum.by.inspection	Inspection calcultation of the optimum
processPlots	Full process plots for mem
roc.analysis	Analysis of different indicators to find the optimum value of the window parameter
summary.epidemic,plot.epidemic,print.epidemic	Influenza Epidemic Timing
summary.flu,plot.flu,print.flu	Methods for influenza modelization
transformdata	Data transformation
transformdata.back	Data transformation
transformseries	Transformation of series of data

Start and duration of an epidemic period (dengue by states in Brazil)

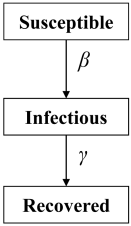


Dengue epidemic thresholds by states



A classical model

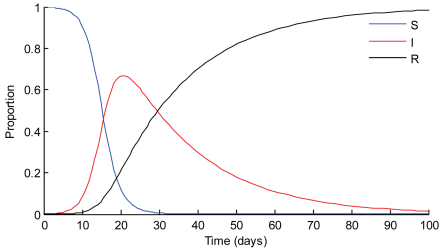
Luz, Struchiner & Galvani (2010).



$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$



Analysing a model

We have

$$S(t) + I(t) + R(t) = N \forall t$$

The basic reproductive number is

$$\mathcal{R}_0 = \frac{\beta N}{\gamma}. \tag{1}$$

Moreover,

$$\lim_{t \to \infty} I(t) = 0 \implies (S_e, 0, R_e),$$

is the only equilibrium point. Consider the Jacobian

$$J(S, I) = \begin{bmatrix} -\beta I & -\beta S \\ \beta I & \beta S - \gamma \end{bmatrix}.$$

Analysing a model (cont.)

- **Equilibria**: The characteristic polynomial is

$$\lambda^2 - (\beta S_e - \gamma)\,\lambda = 0,$$

thus $\lambda_1 = 0$ and $\lambda_2 = \beta S_e - \gamma$. This means we have neutral stability if $S_e = \gamma/\beta$, i.e., $\lambda_2 = 0$ and instability otherwise ($\lambda_2 > 0$ or $\lambda_2 < 0$).

- **Epidemic regimes**:

$$\frac{dI}{dS} = -1 + \frac{\gamma}{\beta}\frac{1}{S},$$

gives

$$I(t) = 1 - R(0) - S(t) + \frac{\gamma}{\beta}\ln\left(\frac{S(t)}{S(0)}\right).$$

From $\lim_{t\to\infty} I(t) = 0$ we know that

$$S(\infty) = 1 - R(0) + \frac{\gamma}{\beta}\ln\left(\frac{S(\infty)}{S(0)}\right).$$

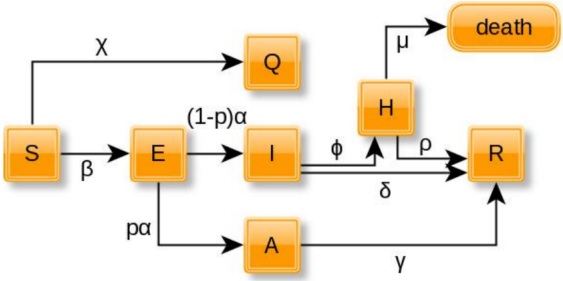
Thus **an epidemic occurs iff**

$$\frac{\beta S(0)}{\gamma} > 1.$$

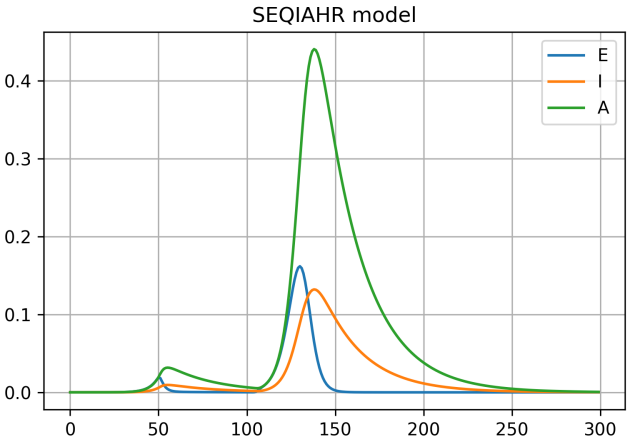
Big structured epidemic models

Coelho et al. (2020)

$$
\begin{aligned}
\frac{dS}{dt} &= -\eta[(1-\chi)S], \\
\frac{dE}{dt} &= \eta[(1-\chi)S] - \alpha E, \\
\frac{dI}{dt} &= (1-p)\alpha E - \delta I - \phi I, \\
\frac{dA}{dt} &= p\alpha E - \gamma A, \\
\frac{dH}{dt} &= \phi I - (\rho + \mu)H, \\
\frac{dR}{dt} &= \delta I + \rho H + \gamma A, \\
\eta &:= \beta(I + A).
\end{aligned}
$$

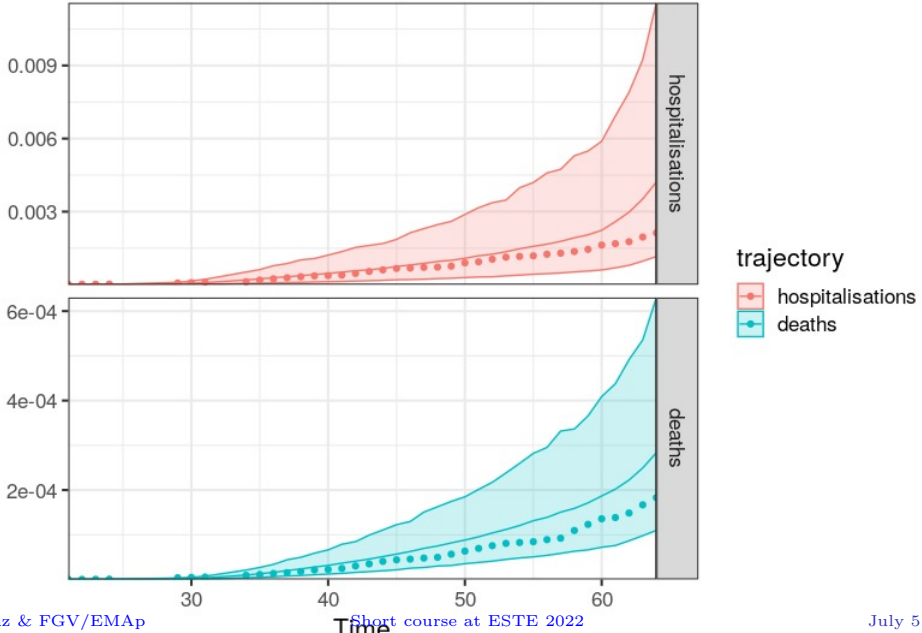


Big model (cont.)



$$\mathcal{R}_0 = \frac{\beta(1-\xi)[p(\phi+\delta)+(1-p)\gamma]}{\gamma(\delta+\phi)}$$

Why so rigid?


0.009
0.006
0.003
hospitalisations
6e-04
4e-04
2e-04
deaths
30
40
50
60
Time
trajectory
hospitalisations
deaths

Stochastic models

Description	State transition	rate
Infection	$(S, E) \to (S-1, E+1)$	$\lambda(1-\chi)S$
Exposed to I	$(E, I) \to (E-1, I+1)$	$(1-p)\alpha E$
Exposed to A	$(E, A) \to (E-1, A+1)$	$p\alpha E$
Hospitalization	$(I, H) \to (I-1, H+1)$	$\phi\ I$
Recovery of I	$(I, R) \to (I-1, R+1)$	δI
Recovery of A	$(A, R) \to (A-1, R+1)$	γA
Recovery of H	$(H, R) \to (H-1, R+1)$	ρH
Death of H	$H \to H-1$	μH

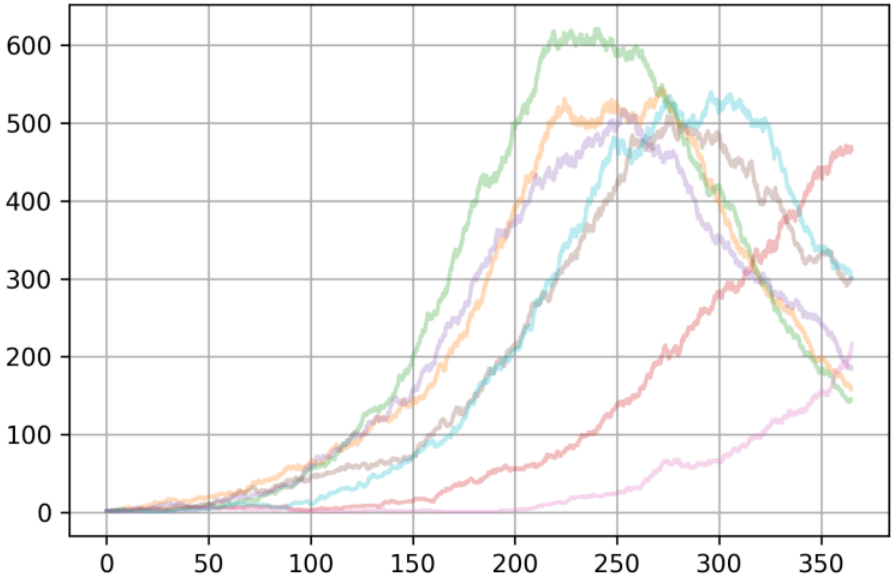
State probabilities

$$P_{s,e,i,a,h}(t) := \mathbb{P}\left(S = s, E = e, I = i, A = a, H = h\right).$$

FCK equation:

$$\begin{aligned}\frac{dP_{s,e,i,a,h}}{dt} &= P_{s+1,e-1,i,a,h}\lambda(1-\chi)(s+1) + P_{s,e+1,i-1,a,h}(1-p)\alpha(e+1) \\ &+ P_{s,e+1,i,a-1,h}p\alpha(e+1) + P_{s,e,i+1,a,h-1}\phi(i+1) + P_{s,e,i+1,a,h}\delta(i+1) \\ &+ P_{s,e,i,a+1,h}\gamma(a+1) + P_{s,e,i,a,h+1}\rho(h+1) + P_{s,e,i,a,h+1}\mu(h+1).\end{aligned}$$

Trajectories



Analysing a stochastic model

Probability-generating functions Starting with $I_i(0)$, the probability of an infected individual in state i producing offspring of type j given that $I_j(0)$ can be obtained from

$$f_i(z_1, \ldots, z_k) = \sum_{j_k=0}^{\infty} \cdots \sum_{j_1=0}^{\infty} P_i(z_1, \ldots, z_k) z_1^{j_1} \cdots z_k^{j_k}.$$

Now, define a matrix whose entries $m_{ji} = \frac{\partial f_i}{\partial u_i}|_{u=1}$ are the expected number of offspring generated in $i \to j$.

$$\mathbb{M} := \begin{bmatrix} 0 & 1-p & p \\ \frac{\beta(1-\chi)}{\beta(1-\chi)+\delta+\phi} & \frac{\beta(1-\chi)}{\beta(1-\chi)+\delta+\phi} & 0 \\ \frac{\beta(1-\chi)}{\beta(1-\chi)+\gamma} & 0 & \frac{\beta(1-\chi)}{\beta(1-\chi)+\gamma} \end{bmatrix}$$

Analysing a stochastic model (cont.)

Under some regularity conditions, we can calculate the **extinction probability**:

$$\mathbb{P}_0 = \prod_{i=1}^{3} q_i^{k_i}$$

after finding $(q_1, q_2, q_3) \in (0,1)^3$ which satisfy constraints. Here $k_1 = E(0)$, $k_2 = I(0)$ and $k_3 = A(0)$.

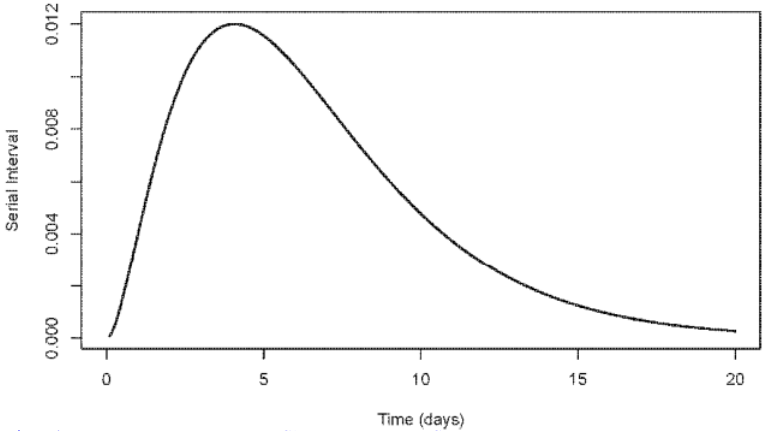
I_0	E_0	A_0	**Approx.** $\mathbb{P}_0$	**SEIAHR** $\mathbb{P}_0$	**SIR** $\mathbb{P}_0$
1	0	0	0.63	0.64	0.58
2	0	0	0.43	0.41	0.33
3	0	0	0.25	0.26	0.19
4	0	0	0.18	0.17	0.11
5	0	0	0.10	0.10	0.06

No more differential equations, please!

Start by looking at

$$R_t = \frac{I_t}{\sum_{s=1}^{t} I_{t-s} w_s}.$$

Then R_t is "the average number of secondary cases that each infected individual would infect if the conditions remained as they were at time t" (Cori et al. 2013).



Modernising the model

$$R_{t,m} = R_{0,m}\left(2\,\text{logit}^{-1}\left(-\sum_{k=1}^{4}(\alpha_k + \beta_{mk})X_{ktm}\right)\right)$$

Priors:

$$\begin{aligned}
\alpha_k &\sim \text{Normal}(0, 5);\\
\beta_{m,k} &\sim \text{Normal}(0, \gamma);\\
\gamma &\sim \text{HalfNormal}(0, 5);\\
R_{0,m} &\sim \text{Normal}(3.28, \kappa);\\
\kappa &\sim \text{HalfNormal}(0, 1/2).
\end{aligned}$$

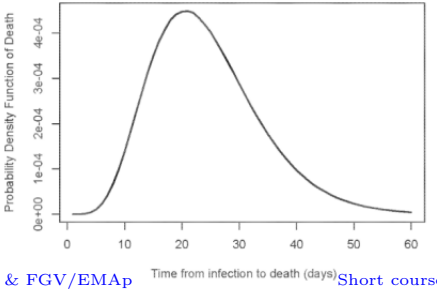
Modernising the model II: likelihood

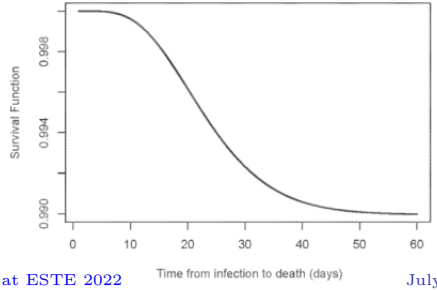
$$c_{t,m} = R_{t,m} \sum_{\tau=0}^{t-1} c_{\tau,m} g_{t-\tau},$$

$$d_{t,m} = \sum_{\tau=0}^{t-1} c_{\tau,m} \pi_{t-\tau,m},$$

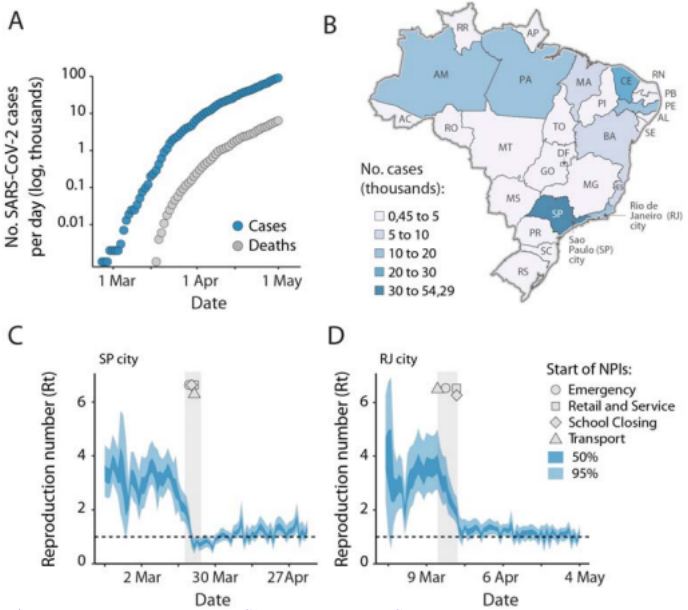
$$D_{t,m} \sim \text{NegativeBinomial}\left(d_{t,m}, \phi\right),$$

$$\phi \sim \text{HalfNormal}(0, 5).$$





The end product

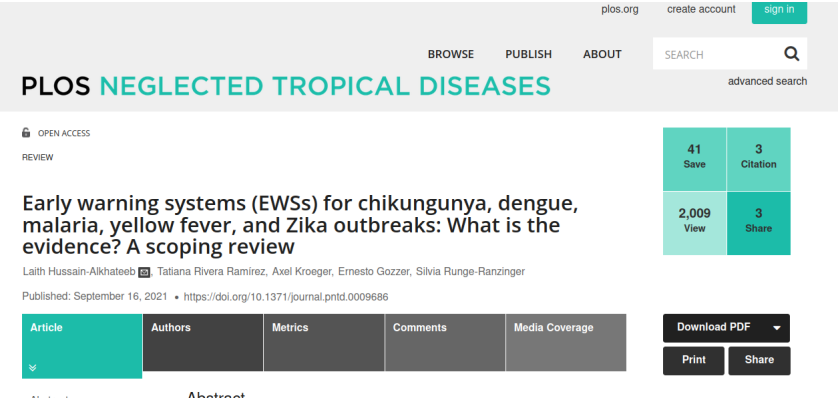


Real time analysis

- In a (nearly) real time analysis of infectious disease cases, we aim to anticipate an outbreak/epidemic
- Usually this is part of an early warning system (EWS)
- An early warning system for outbreaks depends (from an stats point of view) on
  - Organised disease data systems
  - Periodic data release
  - Data cleaning and preparing
  - Monitoring system with reports, dashboards, etc to report the warning

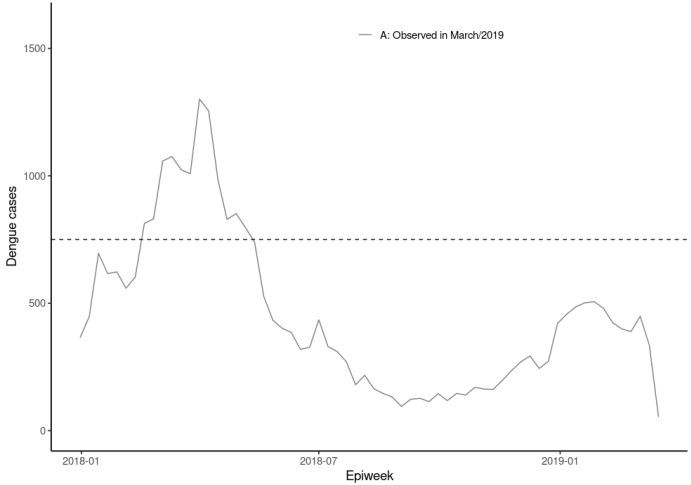
Real time analysis

- The literature on EWS for chikungunya, dengue, malaria, yellow fever, and Zika outbreaks is scarce.



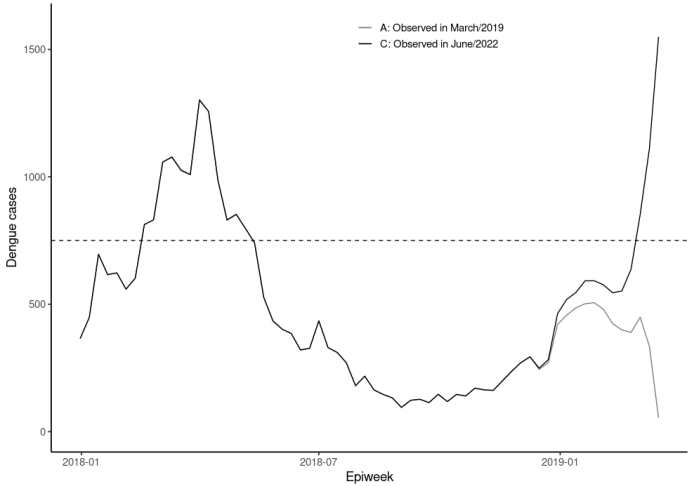
Dengue in Rio

Suppose today is 25/March/2019 and we estimate an epidemic threshold of 750 cases per week.



Dengue in Rio

Suppose today is 25/March/2019 and we estimate an epidemic threshold of 750 cases per week.



Notification delay

- Disease notification cases are in most of the cases delayed
  - Infrastructure issues
  - Epidemic level
  - Wrong diagnostics
  - etc

Data structure

Time	0	1	2	⋯	D-2	D-1	D	N	
1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$		$n_{1,D-2}$	$n_{1,D-1}$	$n_{1,D}$	N_1	Observations
2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$	⋯	$n_{2,D-2}$	$n_{2,D-1}$	$n_{2,D}$	N_2	
3	$n_{3,0}$	$n_{3,1}$	$n_{3,2}$		$n_{3,D-2}$	$n_{3,D-1}$	$n_{3,D}$	N_3	
⋮		⋮				⋮			
T-D	$n_{T-D,0}$	$n_{T-D,1}$	$n_{T-D,2}$		$n_{T-D,D-2}$	$n_{T-D,D-1}$	$n_{T-D,D}$	N_{T-D}	
T-D+1	$n_{T-D+1,0}$	$n_{T-D+1,1}$	$n_{T-D+1,2}$		$n_{T-D+1,D-2}$	$n_{T-D+1,D-1}$	$n_{T-D+1,D}$	N_{T-D+1}	Nowcasting
T-D+2	$n_{T-D+2,0}$	$n_{T-D+2,1}$	$n_{T-D+2,2}$		$n_{T-D+2,D-2}$	$n_{T-D+2,D-1}$	$n_{T-D+2,D}$	N_{T-D+2}	
T-2	$n_{T-2,0}$	$n_{T-2,1}$	$n_{T-2,2}$		$n_{T-2,D-2}$	$n_{T-2,D-1}$	$n_{T-2,D}$	N_{T-2}	
T-1	$n_{T-1,0}$	$n_{T-1,1}$	$n_{T-1,2}$		$n_{T-1,D-2}$	$n_{T-1,D-1}$	$n_{T-1,D}$	N_{T-1}	
T	$n_{T,0}$	$n_{T,1}$	$n_{T,2}$	⋯	$n_{T,D-2}$	$n_{T,D-1}$	$n_{T,D}$	N_T	
T+1	$n_{T+1,0}$	$n_{T+1,1}$	$n_{T+1,2}$		$n_{T+1,D-2}$	$n_{T+1,D-1}$	$n_{T+1,D}$	N_{T+1}	Forecasting
T+2	$n_{T+2,0}$	$n_{T+2,1}$	$n_{T+2,2}$		$n_{T+2,D-2}$	$n_{T+2,D-1}$	$n_{T+2,D}$	N_{T+2}	
⋮		⋮		⋯		⋮			
T+K	$n_{T+K,0}$	$n_{T+K,1}$	$n_{T+K,2}$		$n_{T+K,D-2}$	$n_{T+K,D-1}$	$n_{T+K,D}$	N_{T+K}	

- Total number of cases per week is given by the sum of columns

The chain-ladder model

- The likelihood

$$log(n_{t,d}) \sim N(m_{t,d}, \tau),$$

$$m_{t,d} = \mu + \alpha_t + \beta_d$$

for $t = 1, 2, \ldots, n,\ d = 0, 1, \ldots, D$.

- It is a two-way ANOVA, proposed by Renshaw (1989) with a Bayesian version proposed by Verral (1989a)
- Verral (1989b) propose a state-space representation
- Mack (1993) proposed a Poisson/glm approach with fixed effects

Correcting delays

Using historical data, reporting delays can be corrected

- Bastos et al. (2016, 2019) adapted the chain-ladder model adding spatiotemporal random effects (INLA)

Correcting delays

Using historical data, reporting delays can be corrected

- Bastos et al. (2016, 2019) adapted the chain-ladder model adding spatiotemporal random effects (INLA)
- McGough et al. (2020) also adapted the chain-ladder model with temporal random effects (JAGS + R package NoBS)

Correcting delays

Using historical data, reporting delays can be corrected

- Bastos et al. (2016, 2019) adapted the chain-ladder model adding spatiotemporal random effects (INLA)
- McGough et al. (2020) also adapted the chain-ladder model with temporal random effects (JAGS + R package NoBS)
- Rotejanaprasert et al. (2020) spatiotemporal sliding windows (INLA)

Correcting delays

Using historical data, reporting delays can be corrected

- Bastos et al. (2016, 2019) adapted the chain-ladder model adding spatiotemporal random effects (INLA)
- McGough et al. (2020) also adapted the chain-ladder model with temporal random effects (JAGS + R package NoBS)
- Rotejanaprasert et al. (2020) spatiotemporal sliding windows (INLA)
- Miller et al. (2022) use google searches to improve the correction (INLA)

Bastos et al. (2019) model

- The likelihood

$$n_{t,d} \sim NegBin(\lambda_{t,d}, \phi), \qquad \lambda_{t,d} > 0, \quad \phi > 0.$$

Bastos et al. (2019) model

- The likelihood

$$n_{t,d} \sim NegBin(\lambda_{t,d}, \phi), \qquad \lambda_{t,d} > 0, \quad \phi > 0.$$

- Adding fixed and random effects

$$\log(\lambda_{t,d}) = \mu + \alpha_t + \beta_d + \mathbf{x}'_{t,d}\boldsymbol{\gamma},$$

for $t = 1, 2, \ldots, n,\ d = 0, 1, \ldots, D$, and $\mathbf{x}_{t,d}$ a vector covariates.

Bastos et al. (2019) model

- The likelihood

$$n_{t,d} \sim NegBin(\lambda_{t,d}, \phi), \qquad \lambda_{t,d} > 0, \quad \phi > 0.$$

- Adding fixed and random effects

$$\log(\lambda_{t,d}) = \mu + \alpha_t + \beta_d + \mathbf{x}'_{t,d}\boldsymbol{\gamma},$$

  for $t = 1, 2, \ldots, n$, $d = 0, 1, \ldots, D$, and $\mathbf{x}_{t,d}$ a vector covariates.
- Time (α_t) and delay (β_d) random effects may be modelled using
  - First or second order Gaussian random walks (RW1, RW2)
  - Gaussian autoregressive processes (AR(p))
- Finally, weakly informative priors are set for $(\mu, \phi, \boldsymbol{\gamma}, \boldsymbol{\psi}_\alpha, \boldsymbol{\psi}_\beta, \cdots)$

Inference (Nowcast)

- Learn about all parameters throughout the posterior distribution
- Access the predictive distribution of unknown $n_{t,d}$s (The run-off triangle)

$$p(n_{t^*,d^*} \mid \{n_{t,d}, t+d < T\}), \quad \{(t^*, d^*) : T < t^* + d^* < T + D\}.$$

- Derive the predictive distribution of the total notifications at time t

$$p(N_t = \sum_d n_{t,d} \mid \{n_{t,d}, t+d < T\}).$$

Inference (Monte Carlo)

Sample from the predictive distribution as the following:

1. Sample $(\phi, \mu, \alpha_t, \beta_d)$ from the joint posterior;
2. Sample an unknown $n_{t,d}$ from the likelihood using the sampled parameters;
3. Calculate $N_t = \sum_d n_{t,d}$ (some values are known, others have just been sampled)

Implemented in the R package *nowcaster*

nowcaster

Nowcaster

`nowcaster` is a R package for "nowcasting" epidemiological time-series. Every single system of notification has an intrinsic delay, `nowcaster` can estimate how many counts of any epidemiological data of interest (*i.e.*, daily cases and deaths counts) by fitting a negative binomial model to the time steps of delay between onset date of the event, (*i.e.*, date of first symptoms for cases or date of occurrence of death) and the date of report (*i.e.*, date of notification of the case or death).

`nowcaster` is based on the `R-INLA` and `INLA` packages for "**I**ntegrated **N**ested **L**aplace **A**pproximation" algorithm to Bayesian inference. `INLA` is a fast alternative to others methods for Bayesian inference like **MCMC**. An introduction to `INLA` can be found here.

`nowcaster` is build for epidemiological emergency use, it was constructed for the Brazilian Severe Acute Respiratory Illness (SARI) surveillance database (SIVEP-Gripe).

Installing

Before installing the package certify you have an active installation of `INLA`, to do so you can run the follwing code:

```
install.packages("INLA",
                 repos=c(getOption("repos"),
                 INLA="https://inla.r-inla-download.org/R/stable"),
                 dep=TRUE)
```

Links

Browse source code

Report a bug

License

Full license

GPL (>= 3)

Citation

Citing nowcaster

Developers

Rafael Lopes

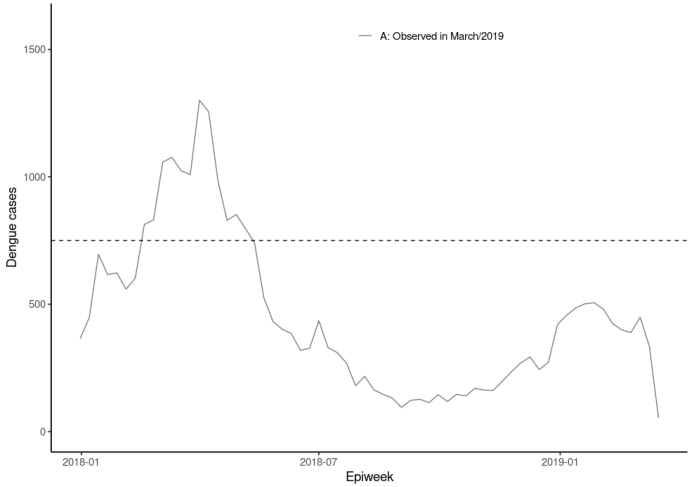
Author, maintainer

Leonardo Bastos

Author

https://covid19br.github.io/nowcaster/

Dengue in Rio



In R

- Data structure (input)

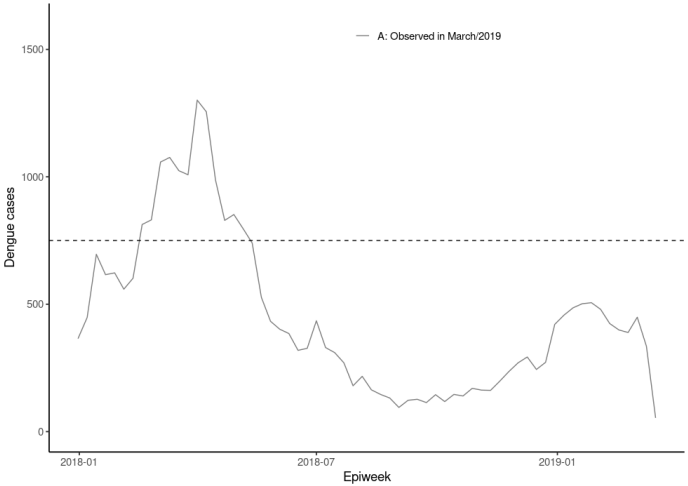
```
> head(dengue.RJ)
  SG_UF ID_MN_RESI  DT_DIGITA DT_SIN_PRI DT_NOTIFIC CS_SEXO NU_IDADE_N
1    33     330270 2019-01-10 2018-08-25 2018-09-05       F       4062
2    33     330270 2019-01-08 2018-12-12 2018-12-19       F       4053
3    33     330270 2019-01-10 2018-06-23 2018-06-29       F       4058
4    33     330270 2018-11-27 2018-06-12 2018-06-15       F       4012
5    33     330270 2018-11-14 2018-05-25 2018-06-04       M       4008
6    33     330270 2018-11-12 2018-07-01 2018-07-11       F       4058
```

- running the nowcasting model

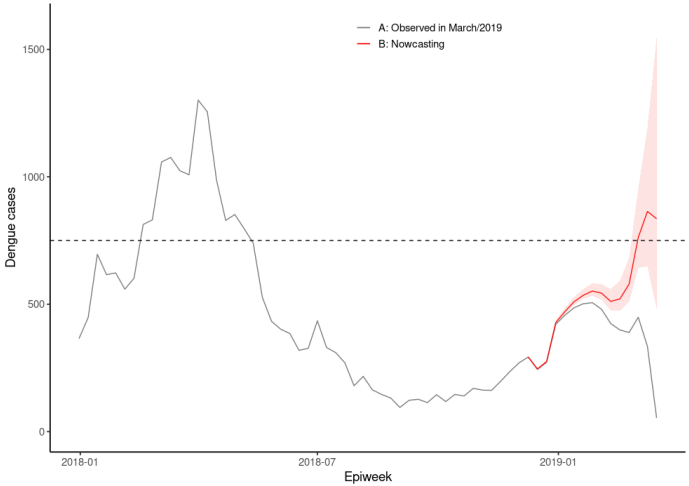
```
library(nowcaster)

dengue.RJ.now <- nowcasting_inla(dataset = dengue.RJ %>%
                                   filter(DT_DIGITA < "2019-03-25"),
                                 date_onset = DT_SIN_PRI,
                                 date_report = DT_DIGITA,
                                 data.by.week = T )
```

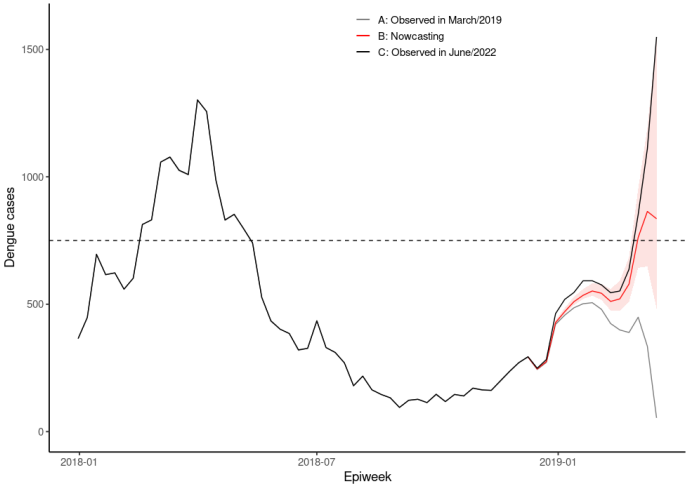
Dengue in Rio



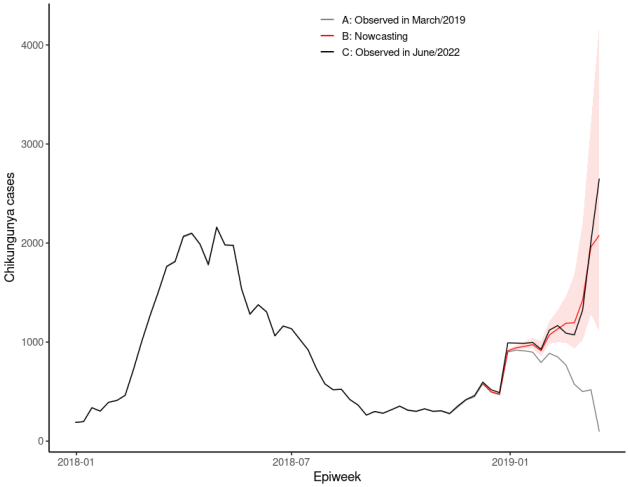
Dengue in Rio



Dengue in Rio

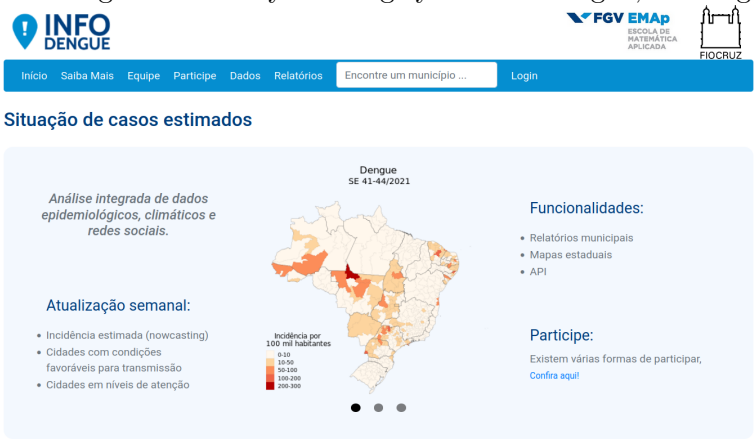


Chikungunya in Rio



Infodengue

- Infodengue is an early warning system for dengue, chikungunya and Zika



https://info.dengue.mat.br

Infodengue

- (Input) (Nearly) real time data
  - Notified cases (instead of suspected/confirmed cases)
  - Social network data (tweets)
  - Real time weather data

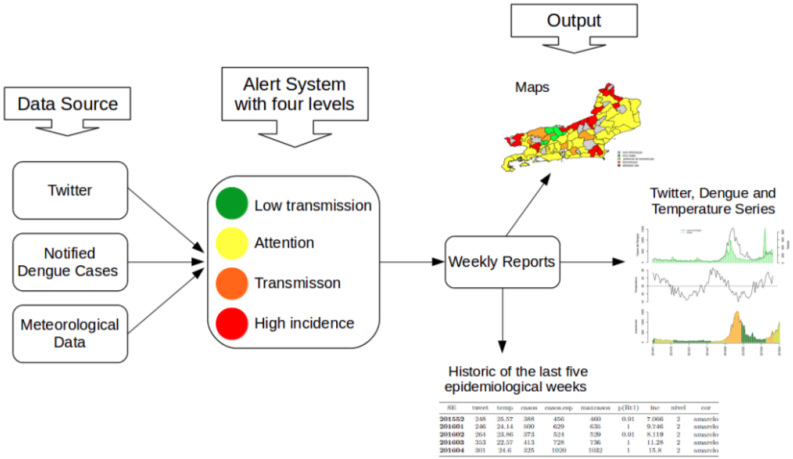
Infodengue

- (Input) (Nearly) real time data
  - Notified cases (instead of suspected/confirmed cases)
  - Social network data (tweets)
  - Real time weather data
- (Output)
  - Reports at city level
  - Corrected estimates
  - Four-color early warning mechanism

Infodengue

- (Input) (Nearly) real time data
  - Notified cases (instead of suspected/confirmed cases)
  - Social network data (tweets)
  - Real time weather data
- (Output)
  - Reports at city level
  - Corrected estimates
  - Four-color early warning mechanism
- (How?)
  - Data cleaning
  - Defining epidemic thresholds
  - Delay correction

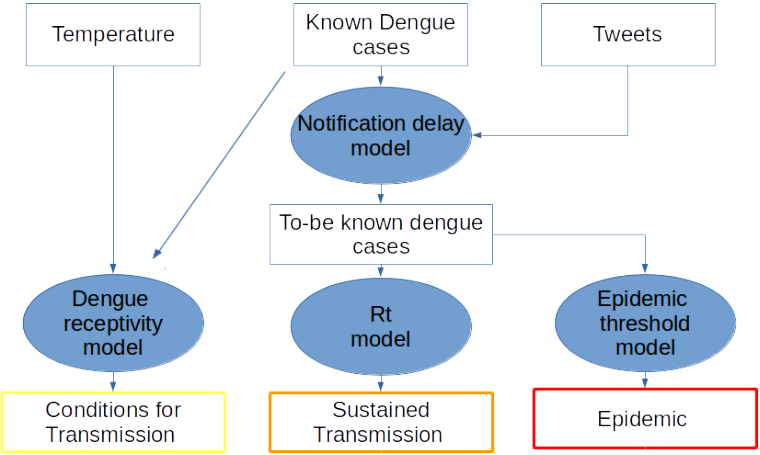
Infodengue: Data workflow



SE	tweet	temp.	casos	casos.esp.	casos.est	p(Rt>1)	Inc	nível	cor
201552	248	25.57	388	456	460	0.91	7.066	2	amarelo
201601	246	24.14	500	629	635	1	9.746	2	amarelo
201602	264	23.86	373	524	539	0.91	8.119	2	amarelo
201603	353	22.57	413	728	736	1	11.28	2	amarelo
201604	301	24.6	325	1020	1032	1	15.8	2	amarelo

https://info.dengue.mat.br

Infodengue: Model workflow



https://info.dengue.mat.br

Infodengue: Issues

- Big data: Weekly analysis for all 5500+ Brazilian municipalities

Infodengue: Issues

- Big data: Weekly analysis for all 5500+ Brazilian municipalities
- Low frequency cases versus high frequency climate data

Infodengue: Issues

- Big data: Weekly analysis for all 5500+ Brazilian municipalities
- Low frequency cases versus high frequency climate data
- Multivariate modelling (three diseases transmitted by the same mosquito)

Infodengue: Issues

- Big data: Weekly analysis for all 5500+ Brazilian municipalities
- Low frequency cases versus high frequency climate data
- Multivariate modelling (three diseases transmitted by the same mosquito)
- Misclassification (diseases with similar symptoms)

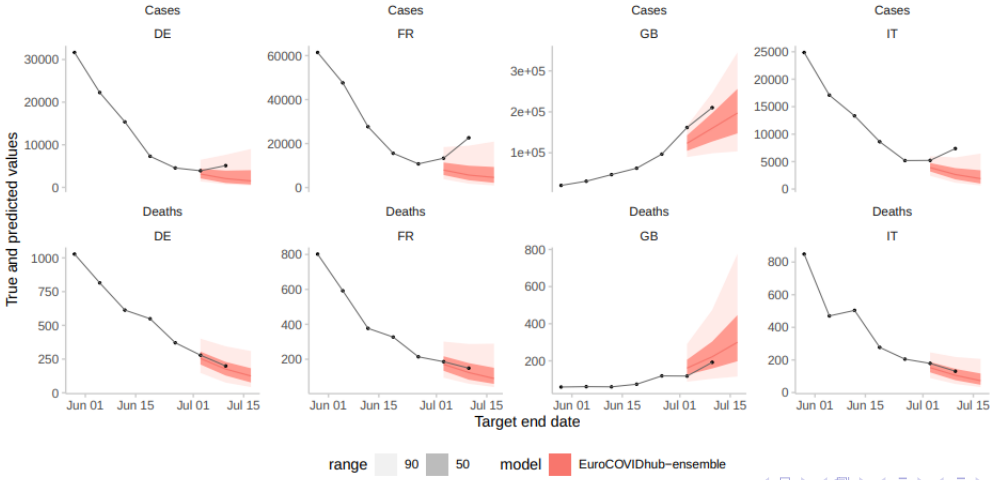
Infodengue: Issues

- Big data: Weekly analysis for all 5500+ Brazilian municipalities
- Low frequency cases versus high frequency climate data
- Multivariate modelling (three diseases transmitted by the same mosquito)
- Misclassification (diseases with similar symptoms)
- (Good) forecasting models (short 4-week ahead and long term one-year ahead)

How do you assess predictions from a model?

Desiderata:

- (Well-calibrated) Probabilistic predictions;
- Encourage careful and honest predictions;



Fixing notation

Presentation will be mostly based on Gneiting & Raftery (2007) and Bosse et al. (2022). Let $\mathcal{P}$ be a convex class of probability measures on $(\Omega, \mathcal{A})$. We call $P \in \mathcal{P}$ a probabilistic forecast.

Definition (Scoring rule)

We say $S(P, \cdot) : \Omega \to [-\infty, \infty]$ is a **scoring rule** if it is measurable and $S(P, \cdot)$ is P-quasi-integrable for all $P \in \mathcal{P}$.

The expected score under $Q \in \mathcal{P}$ if the forecast is P is

$$S(P, Q) := \int_{\Omega} S(P, \omega) \, dQ(\omega).$$

Definition (Strictly proper scoring rule)

We say S is **proper** if $S(Q, Q) \geq S(P, Q)$ for all $P, Q \in \mathcal{P}$. In addition, we say S is **strictly proper** if equality is achieved only for $P = Q$.

Categorical examples

- **Spherical score**.
  For $\alpha > 1$ we can define

$$S(\boldsymbol{p}, i) = \frac{p_i^{\alpha-1}}{\left(\sum_{j=1}^m p_j^\alpha\right)^{\frac{\alpha-1}{\alpha}}} \tag{2}$$

- **Logarithmic score**.
  When $G(\boldsymbol{p}) =$ is the Shannon entropy, we have $S(\boldsymbol{p}, i) = \log p_i$ and

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sum_{j=1}^m \log\left(\frac{q_j}{p_j}\right), \tag{3}$$

  as the Kullback-Leibler divergence.

Scoring rules for density forecasts

Let μ be a σ-finite measure on $(\Omega, \mathcal{A})$. For $\alpha > 1$ define $\mathcal{L}_\alpha$ be the space of probability measures on $(\Omega, \mathcal{A})$ such that $\nu \ll \mu$ and $p(\omega) = \frac{d\nu}{d\mu}(\omega)$ and

$$||p||_\alpha = \left(\int_\Omega p(\omega)^\alpha \, d\mu(\omega) \right)^\alpha < \infty.$$

We establish a correspondence between the forecast P and its μ-density, p. **Examples:**

- **Quadratic:**

$$\mathrm{QS}(p, \omega) = 2p(\omega) - ||p||_2^2, \tag{4}$$

  is strictly proper relative to $\mathcal{L}_2$ class of probability measures.

- **Pseudo-spherical:**

$$\mathrm{PseudoS}(p, \omega) = \frac{p(\omega)^{\alpha-1}}{||p||_\alpha^{\alpha-1}}, \tag{5}$$

- **Logarithmic score:**

$$\mathrm{LogS}(p, \omega) = \log p(\omega), \tag{6}$$

  is what happens to the pseudo-spherical score when $\alpha \to 1$.

Continuous ranked probability score

The continuous ranked probability score (CRPS):

$$\mathrm{CRPS}(F, x) = -\int_{-\infty}^{\infty} \left(F(y) - \mathbb{I}\{y \geq x\}\right)^2 \, dy, \tag{7}$$

can be seen as the integral of the Brier scores for the associated binarisation of the forecasts based on x as cutoff.

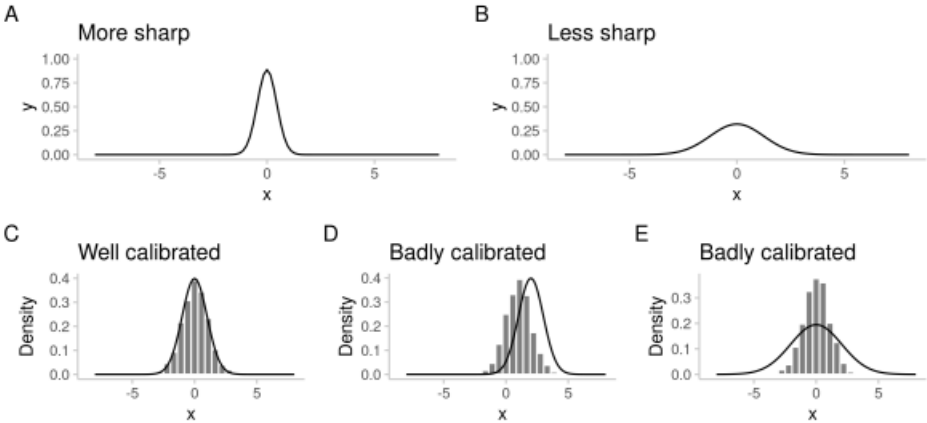
Example:

$$\mathrm{CRPS}\left(\mathrm{Normal}(\mu, \sigma^2), x\right) = \sigma \left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma}\left(2\Phi\left(\frac{x-\mu}{\sigma}\right) - 1\right)\right],$$

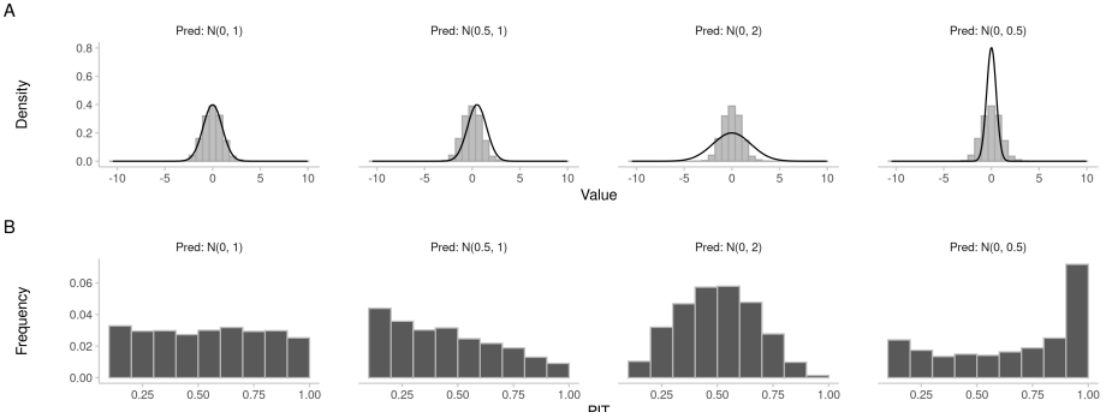
where φ and Φ are the probability density function and cumulative distribution function of a standard normal, respectively.

Being right and being sure of it

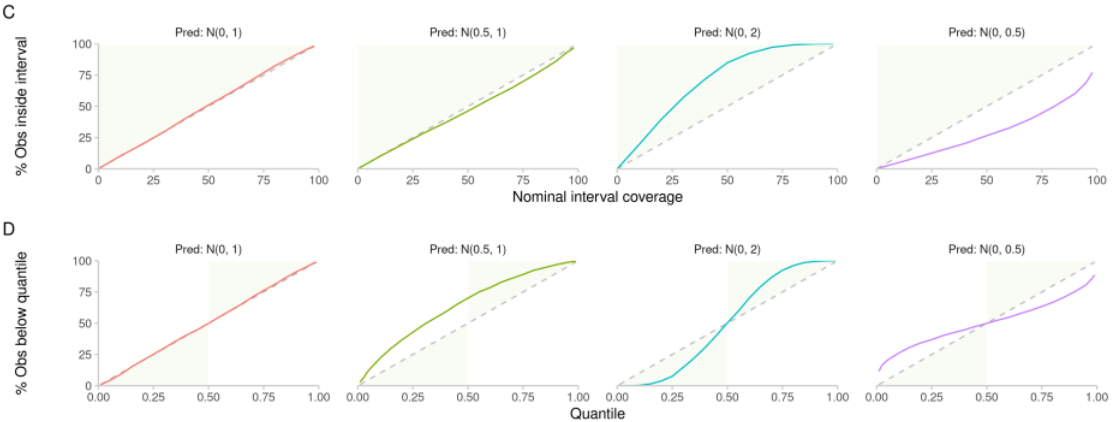
- Calibration
- Sharpness



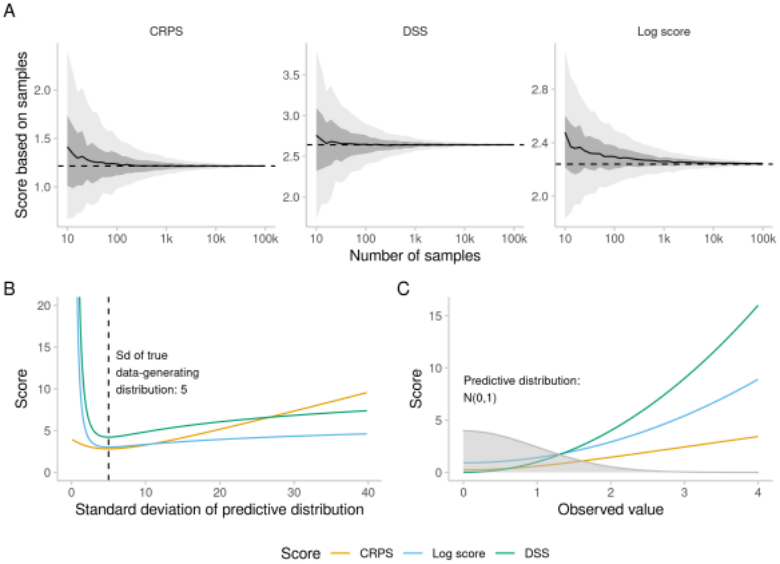
True target is Normal(0, 1)



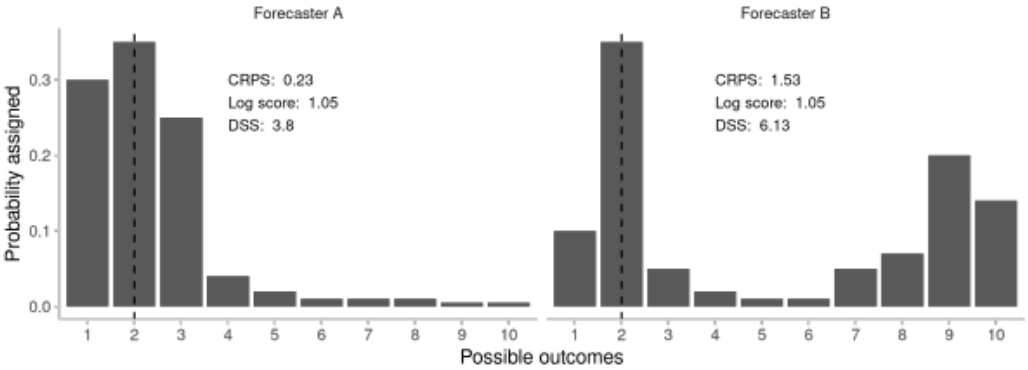
True target is Normal$(0, 1)$



Which score to pick if you're predicting hospitalisations?



Local *vs* Global



Scale dependence