Multilevel models in epidemiology

An application to chikungunya and Zika epidemics

Advanced statistical methods for physicists

Julien Riou
Institute of Social and Preventive Medicine
University of Bern

Bern, 31 May 2019

Themes

Provide a real-world example of model development in relation to a scientific question:

- Mechanistic model (data-generating processes)
- Multilevel structure in relation to data structure
- Partial pooling of information

The global invasion of *Aedes* mosquitoes

Aedes mosquitoes (i)

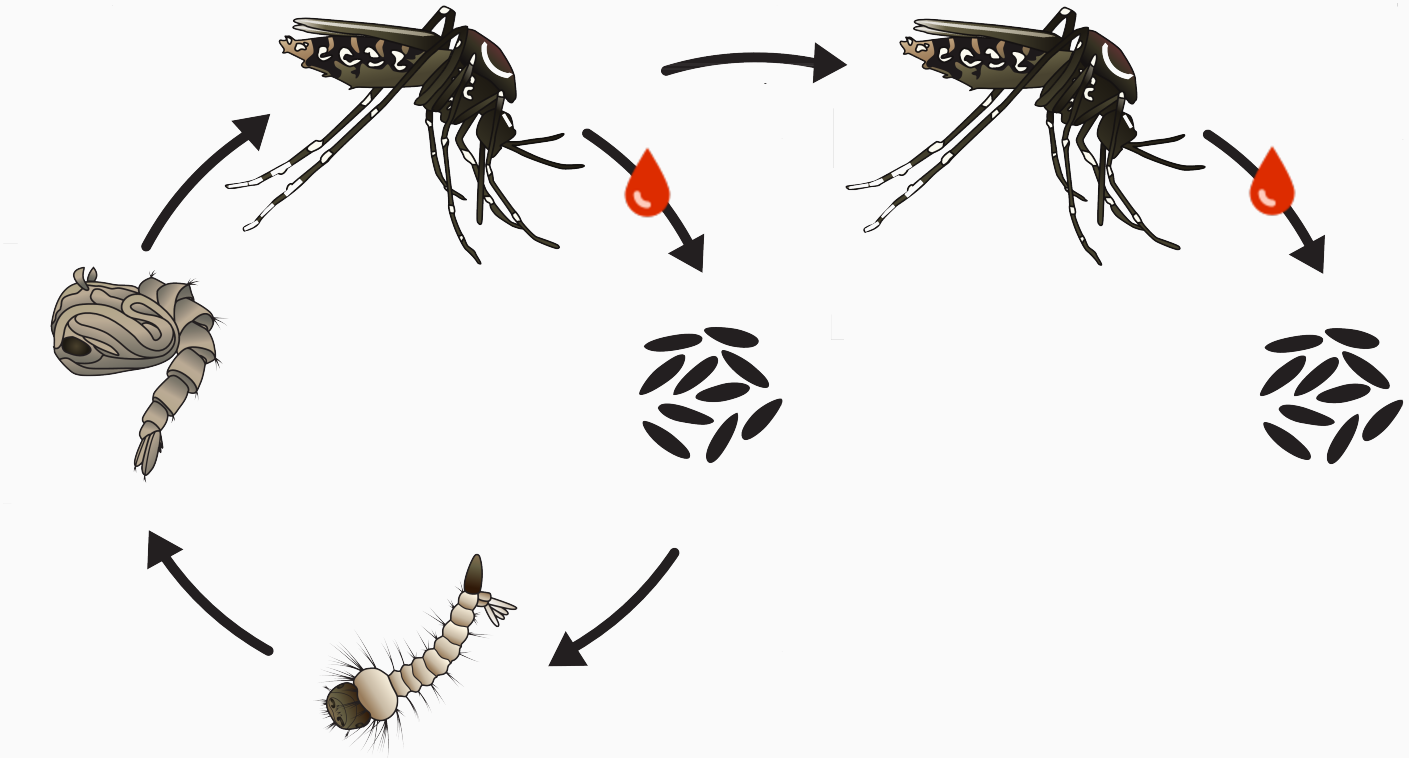


Figure 1: Life cycle of *Aedes* mosquitoes.

- Blood meal necessary to egg maturation
- Biting behaviour: gonotrophic cycle
- Transmission through saliva: specific vectorial competence

Aedes mosquitoes (ii)

Two species are important to human health:

- *Aedes aegypti* (tropical and subtropical areas)
- *Aedes albopictus* (subtropical et temperate areas)

Figure 2: Female adult specimens of *Aedes aegypti* (left) and *Aedes albopictus* (right).

The domestication of *Aedes aegypti*

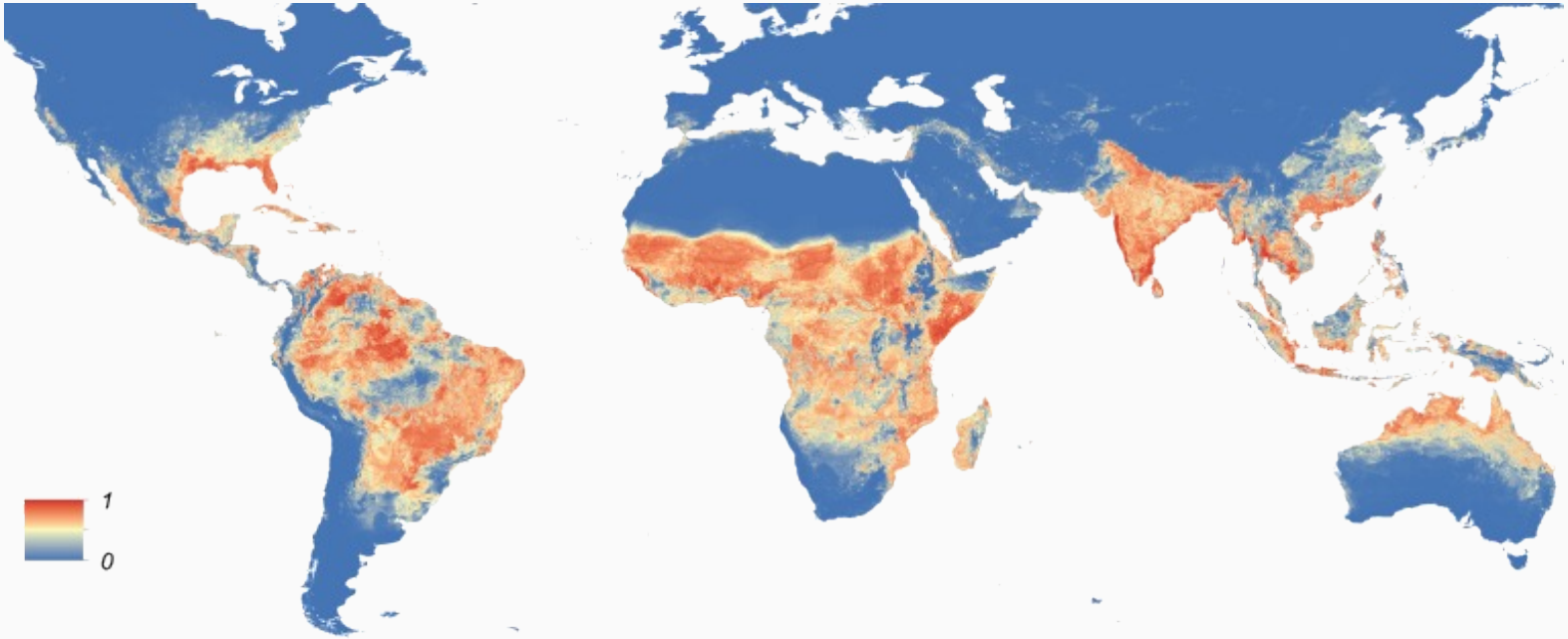


Figure 3: World distribution of *Aedes aegypti* [1].

- Originally from Africa, extension from the 15th century
- Urban and domestic species: adapted to human settlements[2]

[1] Kraemer et al., *eLife* (2015) ; [2] Powell et Tabachnick, *Memorias do Instituto Oswaldo Cruz* (2013)

The invasion of *Aedes albopictus*

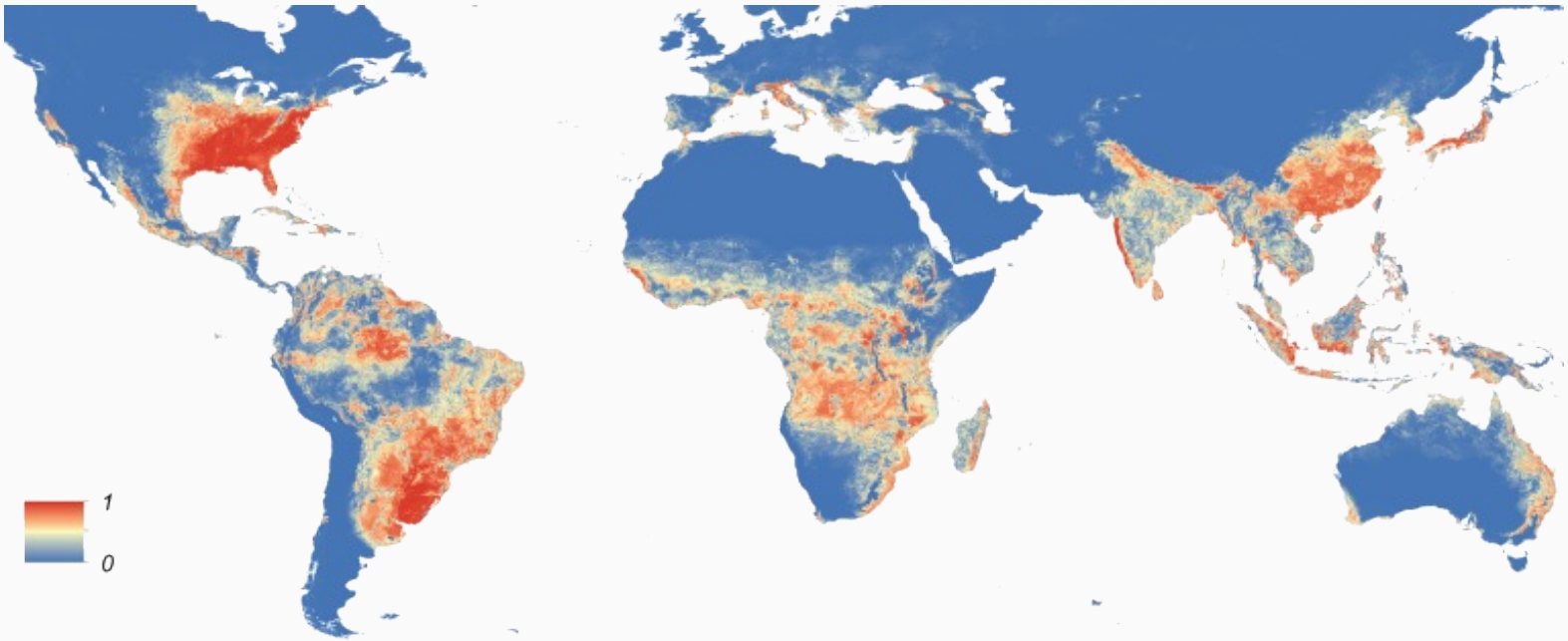
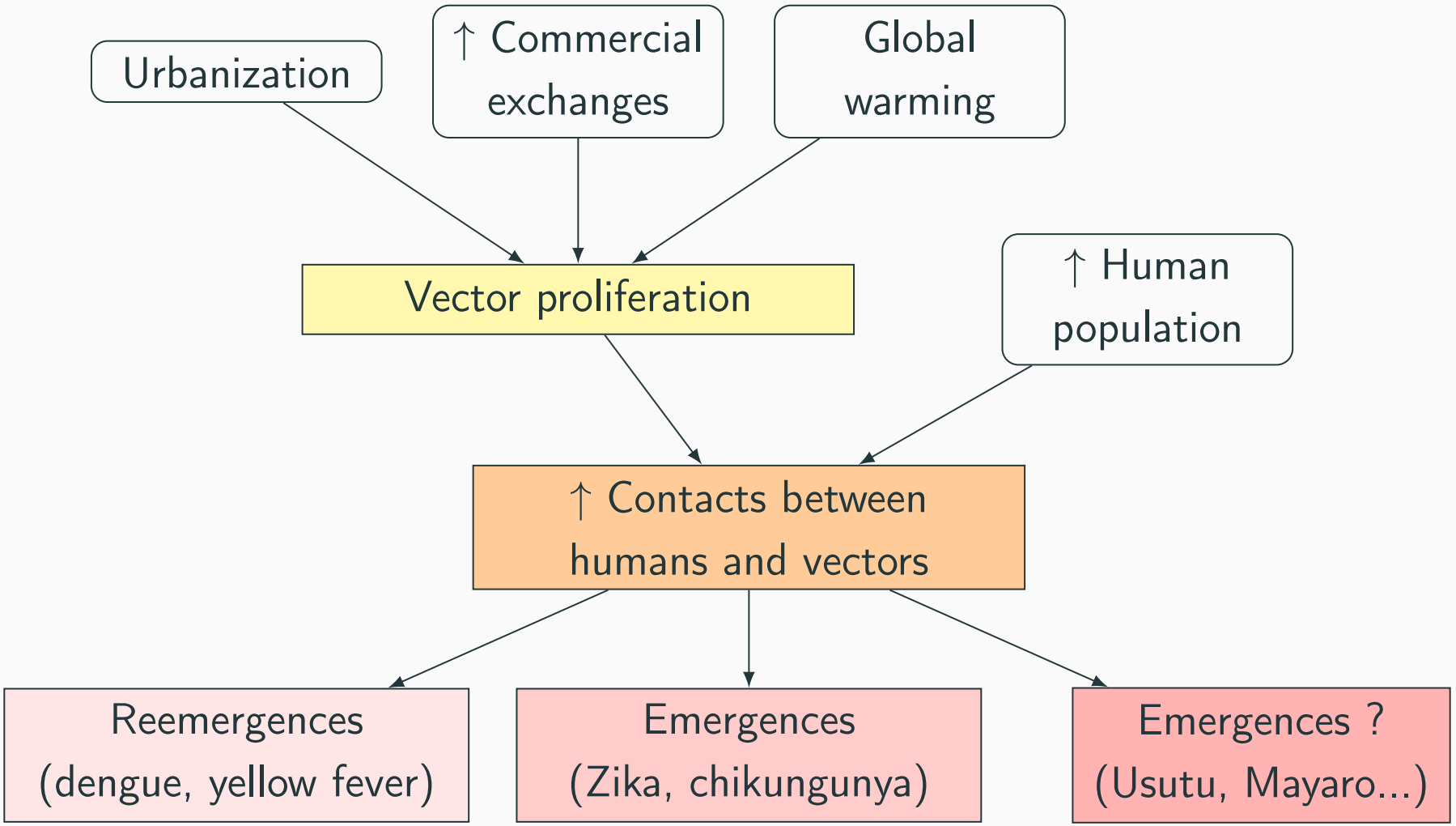


Figure 4: World distribution of *Aedes albopictus* [1].

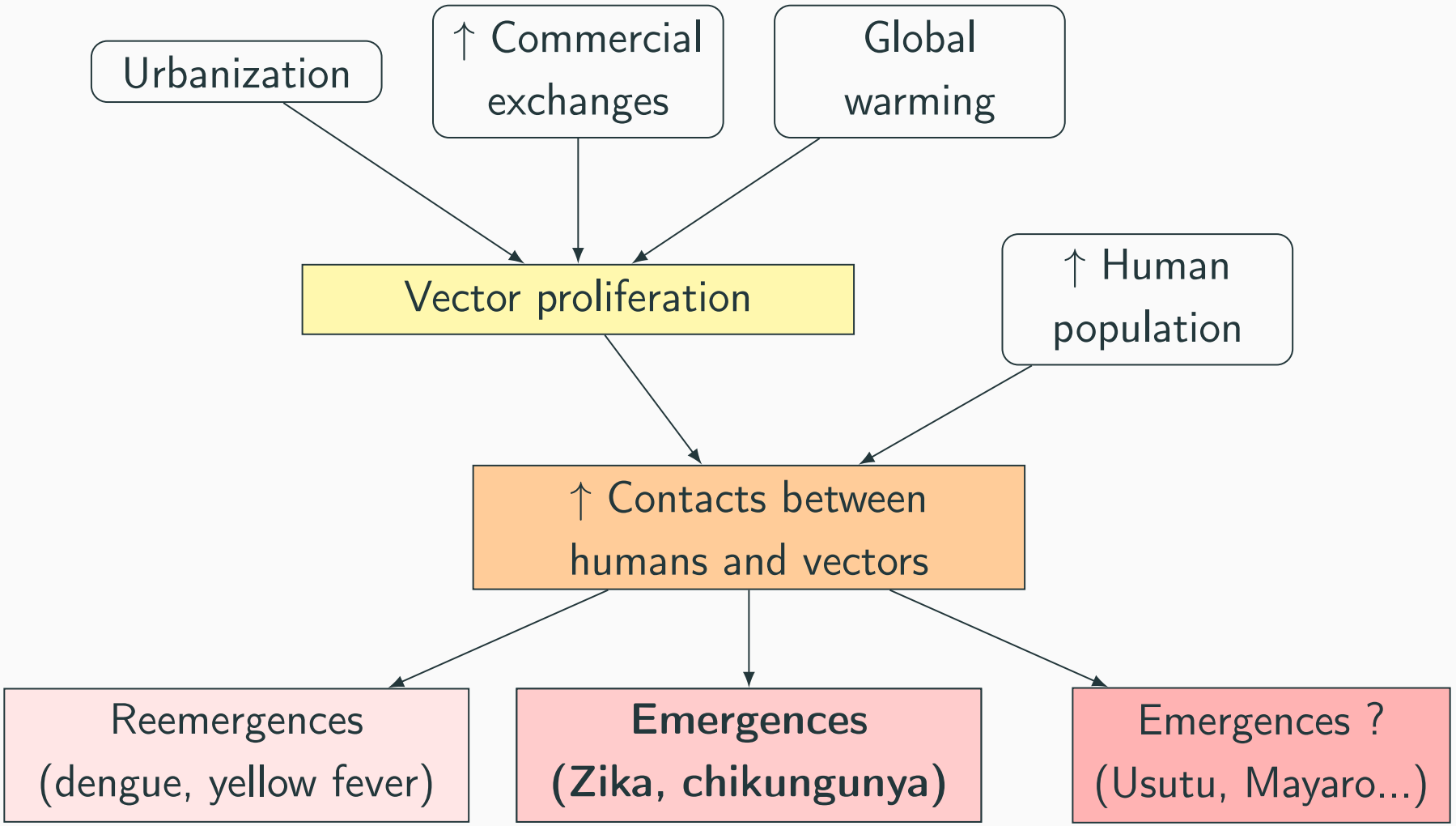
- Originally from Asia, extension from the 20th century[2]
- Invasive species: plasticity, competitive advantages[3]

[1] Kraemer et al., *eLife* (2015) ; [2] Reiter, *Journal of the American Mosquito Control Assoc.* (1998) ; [3] Paupy et al., *Microbes and Infection* (2009)

Disease emergences



Disease emergences



World propagation of chikungunya (CHIKV)

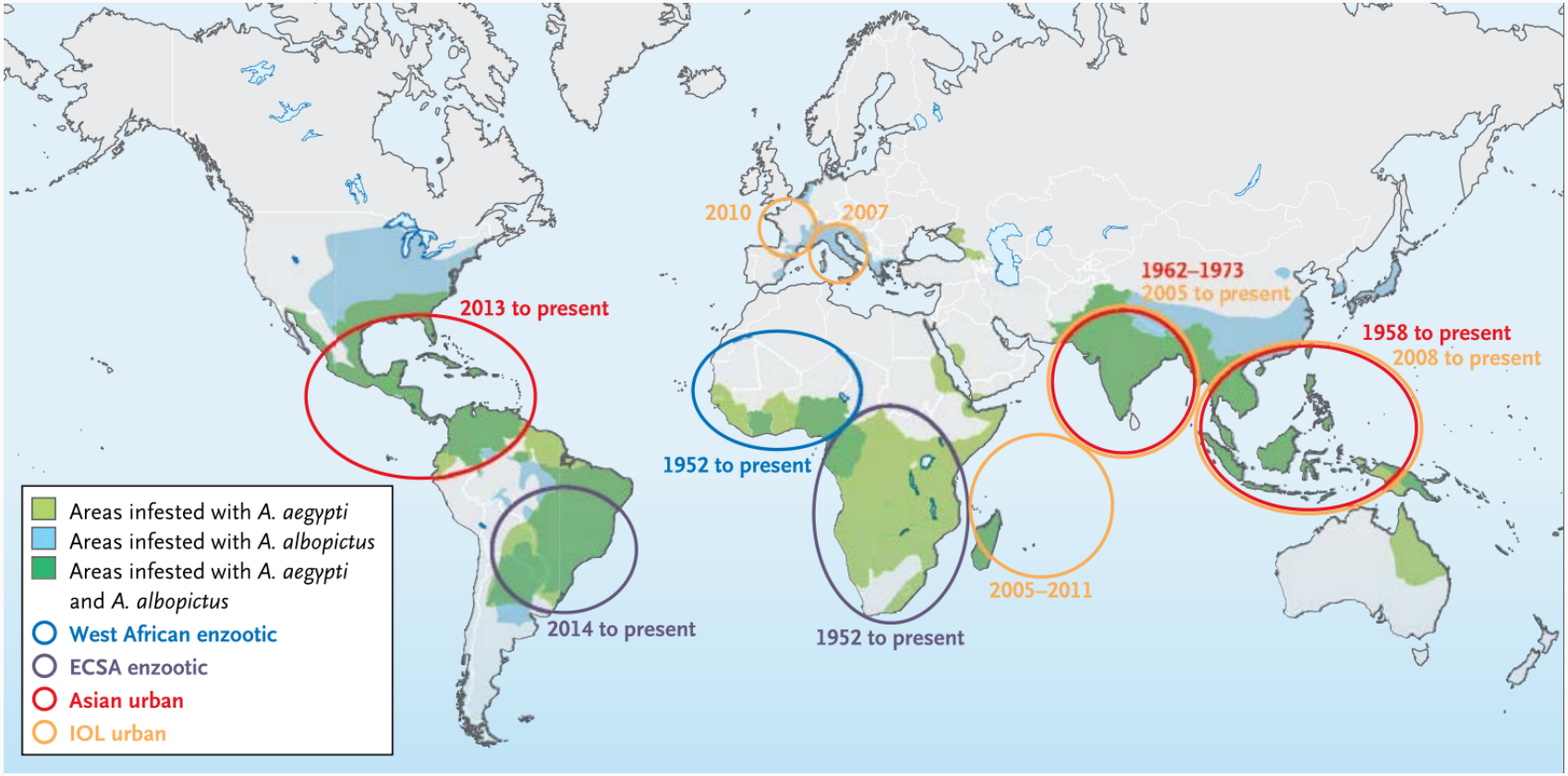


Figure 5: Origin and extension of the chikungunya virus and his vectors[1].

[1] Weaver et al., *New England Journal of Medicine* (2015)

World propagation of Zika (ZIKV)

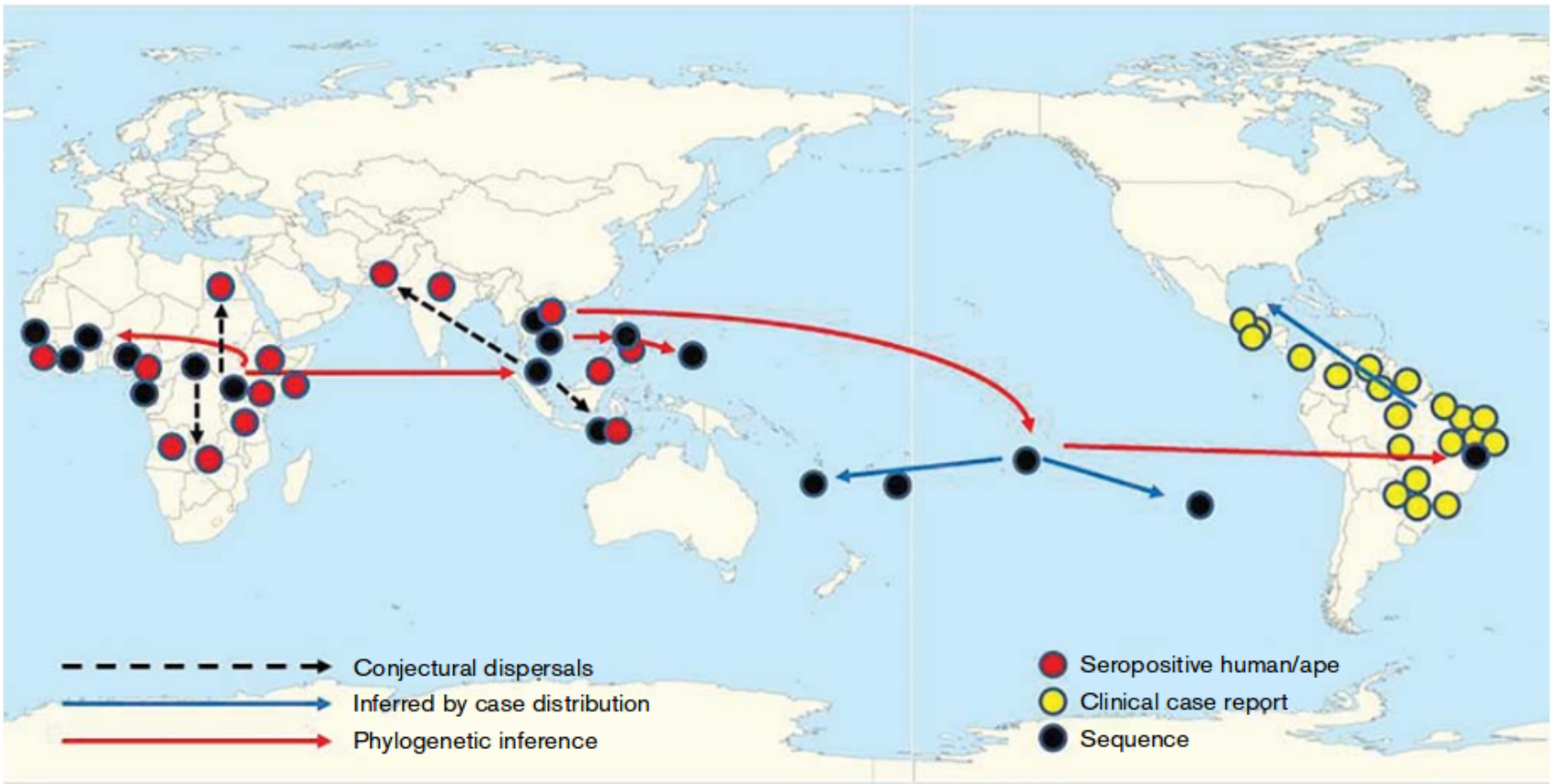


Figure 6: Origin and extension of the Zika virus[1].

[1] Gatherer et Kohl, *Journal of General Virology* (2016)

Comparing Zika and chikungunya epidemics

Strategy

Successive waves of chikungunya and Zika epidemics:

- each circulating for the first time
- in the same areas
- within a short timespan

$\Rightarrow$ Comparison of epidemics of different viruses:

- in the same populations (immunologically naive)
- in the same environments (vectors)
- observed by the same surveillance systems

Data

Time-series of incidence data for 18 outbreaks of ZIKV and CHIKV:

- weekly number of reported cases CHIKV or ZIKV
- between 2013 and 2016
- in 9 islands with similar surveillance systems

```
> load("zikachik.Rdata")
> tibble(zikachik)
# A tibble: 537 x 1
   zikachik$record… $REGION $REGION_ID $ISLAND $ISLAND_ID $ISLAND_ABB $VIRUS $VIRUS_ID $DATE      $WEEK $YEAR $N_DAY $N_WEEK $NCASES  $POP $STATION
              <int> <fct>        <dbl> <chr>        <int> <fct>       <chr>      <dbl> <date>     <int> <int>  <dbl>   <dbl>   <int> <dbl> <chr>
 1              185 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-11-09    45  2014      0       0       1  7000 91954
 2              186 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-11-16    46  2014      7       1       7  7000 91954
 3              187 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-11-23    47  2014     14       2      17  7000 91954
 4              188 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-11-30    48  2014     21       3      54  7000 91954
 5              189 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-12-07    49  2014     28       4      81  7000 91954
 6              190 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-12-14    50  2014     35       5      93  7000 91954
 7              191 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-12-21    51  2014     42       6      70  7000 91954
 8              192 FP               0 AUSTRA…          6 AUS         CHIKV          0 2014-12-28    52  2014     49       7      65  7000 91954
 9              193 FP               0 AUSTRA…          6 AUS         CHIKV          0 2015-01-04     1  2015     56       8      65  7000 91954
10              194 FP               0 AUSTRA…          6 AUS         CHIKV          0 2015-01-11     2  2015     63       9      74  7000 91954
# … with 527 more rows, and 32 more variables: $NCASES_M1 <dbl>, $NCASES_M2 <dbl>, $NCASES_M3 <dbl>, $NCASES_M4 <dbl>, $NCASES_M5 <dbl>,
#   $CUM_NCASES <dbl>, $INC <dbl>, $CUMINC <dbl>, $TOT_NCASES <int>, $AR <dbl>, $TempAvgC_W0 <dbl>, $PrecCm_W0 <dbl>, $TempAvgC_W1 <dbl>,
#   $PrecCm_W1 <dbl>, $TempAvgC_W2 <dbl>, $PrecCm_W2 <dbl>, $TempAvgC_W3 <dbl>, $PrecCm_W3 <dbl>, $TempAvgC_W4 <dbl>, $PrecCm_W4 <dbl>,
#   $TempAvgC_W5 <dbl>, $PrecCm_W5 <dbl>, $TempAvgC_W6 <dbl>, $PrecCm_W6 <dbl>, $TempAvgC_W7 <dbl>, $PrecCm_W7 <dbl>, $TempAvgC_W8 <dbl>,
#   $PrecCm_W8 <dbl>, $temp <dbl>, $GT.mean <dbl>, $GT.sd <dbl>, $Ostar <dbl>
```

https://github.com/jriou/multilevel_models_epi

Data

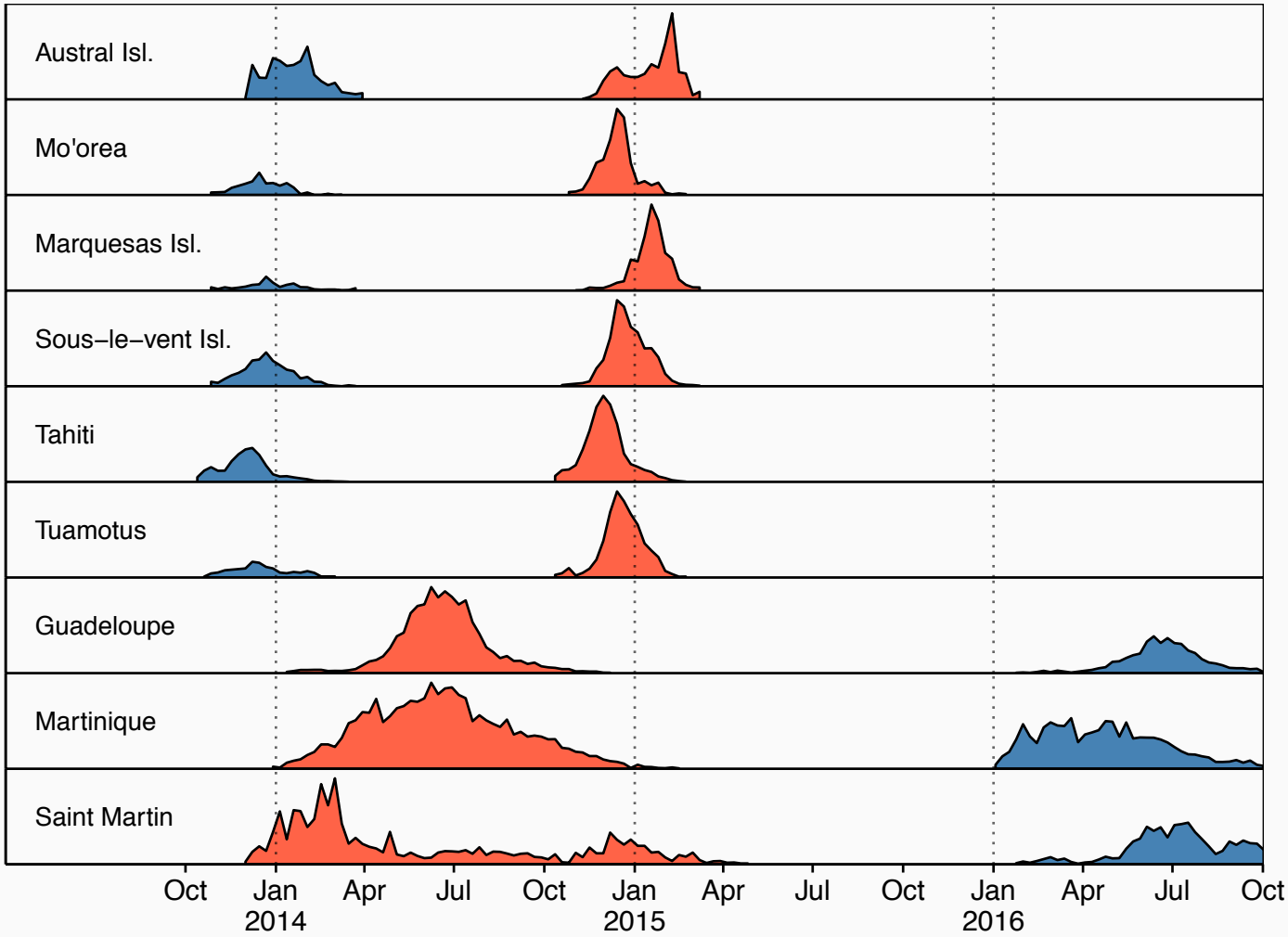


Figure 7: Profiles of CHIKV (red) and ZIKV (blue) incidence in nine territories during 2013–2016.

Data

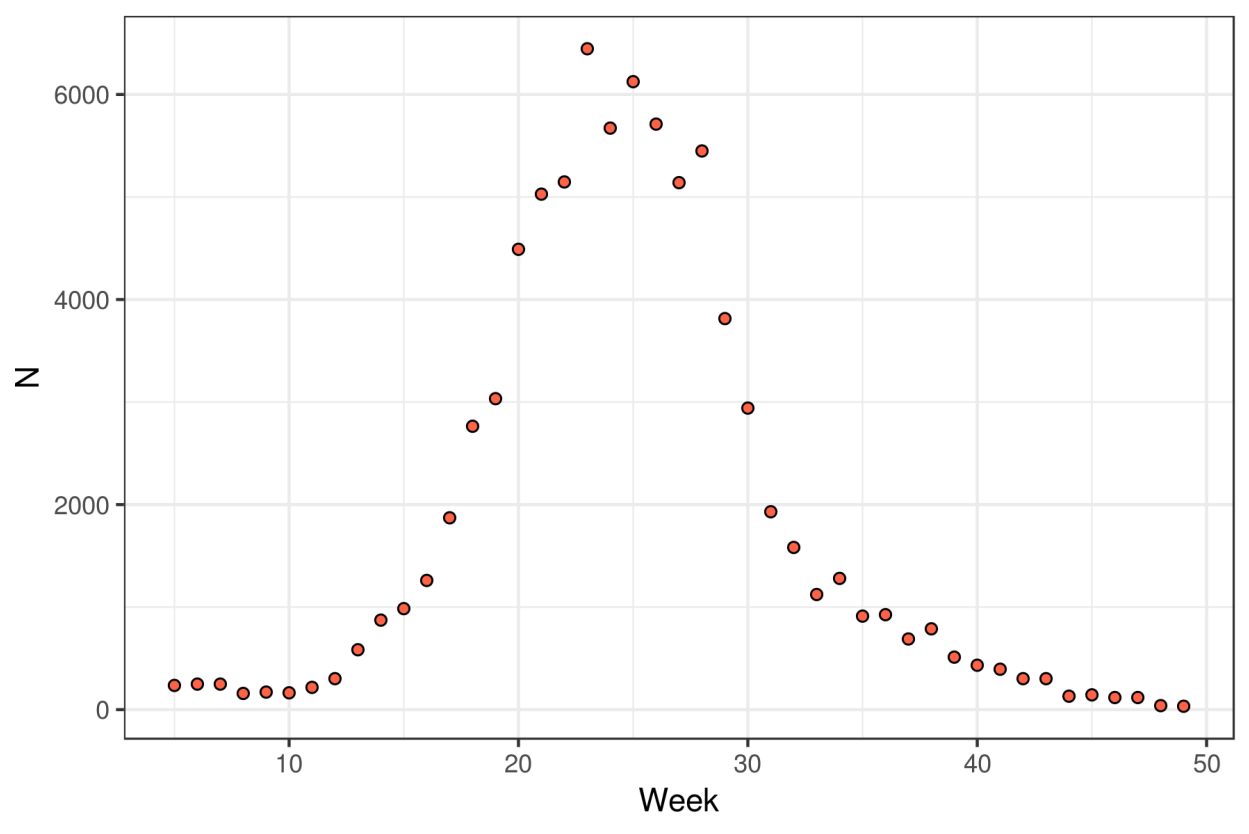


Figure 8: Weekly number of reported cases of chikungunya in Guadeloupe (Feb. – Dec. 2014).

Data-generating processes: observation

How is the number of reported cases on week t (O_t) generated?

Data-generating processes: observation

How is the number of reported cases on week t (O_t) generated?

- true number of infections by CHIKV or ZIKV (I_t)
- with symptoms (70% for CHIKV, 20-30% for ZIKV)
- sufficient to lead to a consultation with a physician
- recognized by the physician
- reported by the physician to the surveillance authorities

Data-generating processes: observation

How is the number of reported cases on week t (O_t) generated?

- true number of infections by CHIKV or ZIKV (I_t)
- with symptoms (70% for CHIKV, 20-30% for ZIKV)
- sufficient to lead to a consultation with a physician
- recognized by the physician
- reported by the physician to the surveillance authorities

At the observation level:

$$O_t \sim \text{Binom}(I_t, \rho)$$

$\Rightarrow$ Parameter ρ: probability of reporting

Data-generating processes: transmission

How it the true number of infections (I_t) generated?

Data-generating processes: transmission

How it the true number of infections (I_t) generated?

- (if we ignore importations)
- all new cases can be linked back to cases in the last few weeks
- with a delay corresponding to the actions of the vector

Data-generating processes: transmission

How it the true number of infections (I_t) generated?

- (if we ignore importations)
- all new cases can be linked back to cases in the last few weeks
- with a delay corresponding to the actions of the vector

At the transmission level:

$$I_t \sim \text{Binom}\left(S_t, \beta \boxed{\frac{1}{N}\sum_{n=1}^{5} w_{t,n} I_{t-n}}\right)$$

$\Rightarrow$ Parameter β: number of secondary cases by primary case ($\mathcal{R}_0$)

$\Rightarrow$ Exposure: depends on the serial interval w_t

The serial interval

Definition

Serial interval: the time between the disease onset of a primary case and one of its secondary cases[1]

Reconstruction using the full transmission cycle

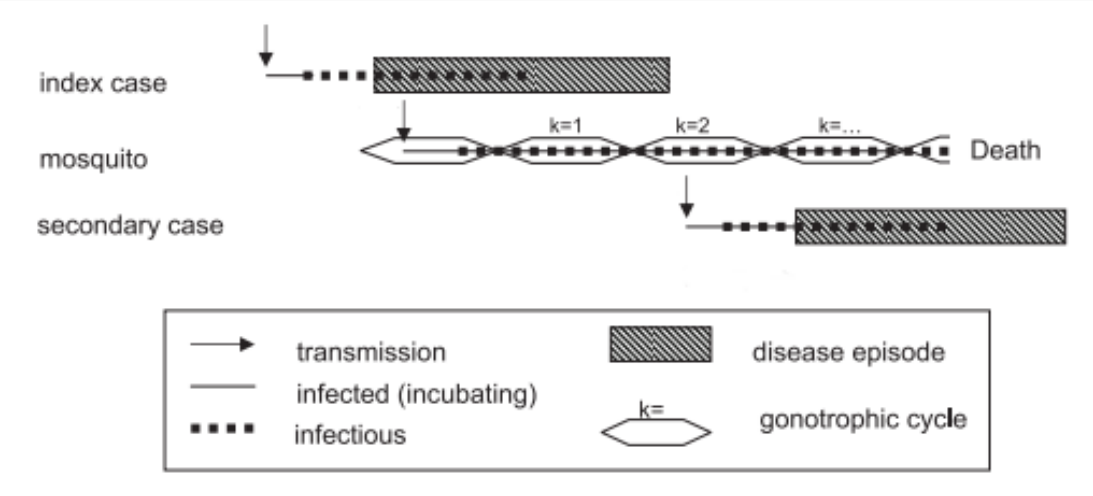


Figure 9: Framework for calculating the distribution of the serial interval[2].

[1] Svensson et al, *Math. biosciences* (2007); [2] Boëlle et al, *Vector-borne and zoonotic diseases* (2007)

The serial interval

Definition

Serial interval: the time between the disease onset of a primary case and one of its secondary cases[1]

Reconstruction using the full transmission cycle

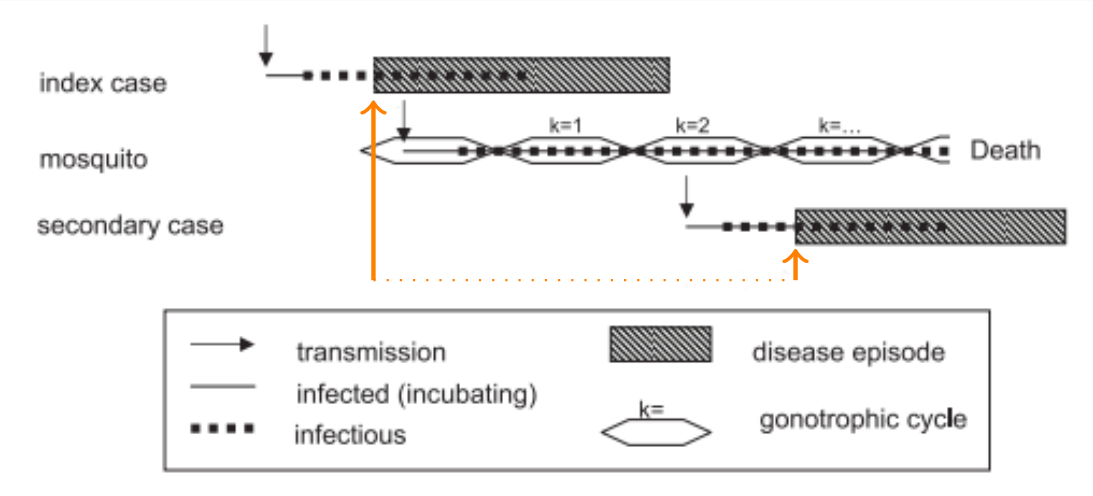


Figure 9: Framework for calculating the distribution of the serial interval[2].

[1] Svensson et al, *Math. biosciences* (2007); [2] Boëlle et al, *Vector-borne and zoonotic diseases* (2007)

The serial interval

Definition

Serial interval: the time between the disease onset of a primary case and one of its secondary cases[1]

Reconstruction using the full transmission cycle

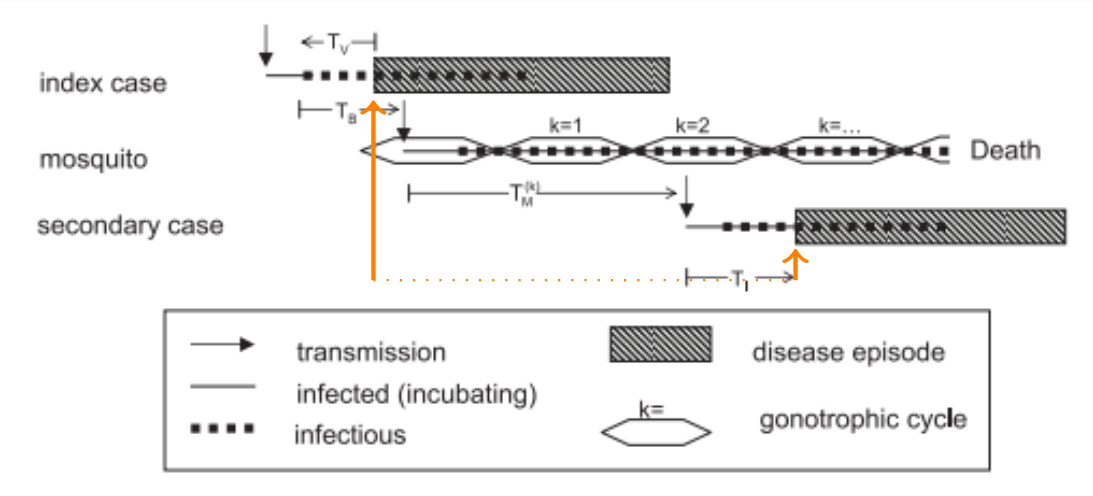


Figure 9: Framework for calculating the distribution of the serial interval[2].

[1] Svensson et al, *Math. biosciences* (2007); [2] Boëlle et al, *Vector-borne and zoonotic diseases* (2007)

The serial interval

Using published data on each stage (including dependence to local temperature T) we obtain:

$$T_{SI} = -T_V + T_B + T_M(T) + T_I$$

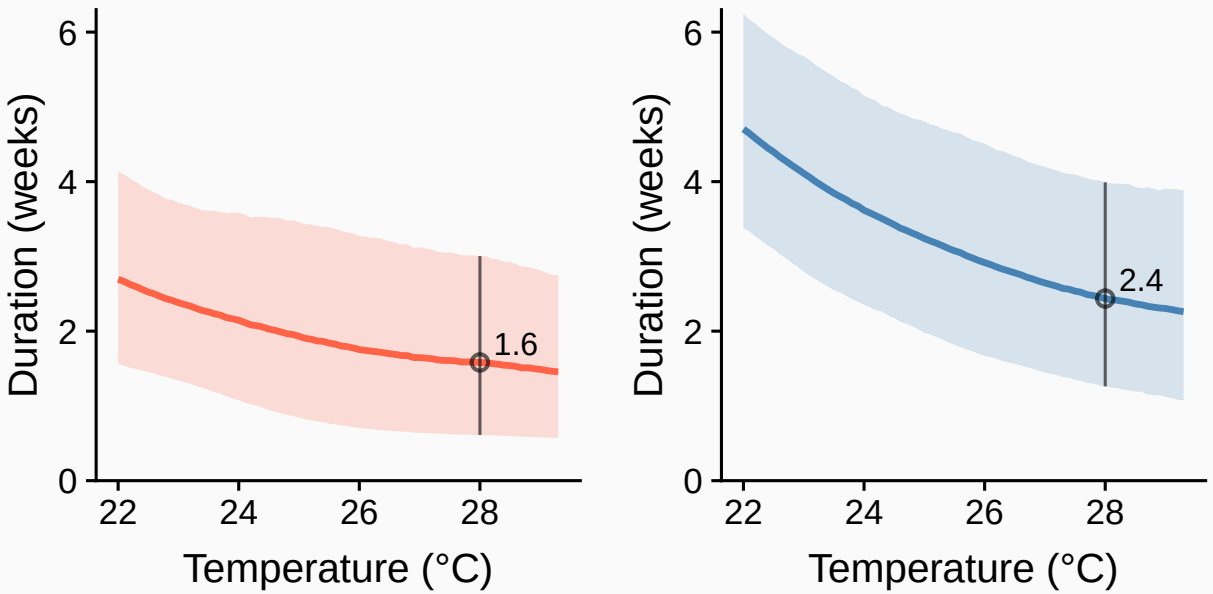


Figure 10: Distribution of the serial interval for CHIKV (red) and ZIKV (blue) according to temperature.

One disease, one island: model

Let O_t be the observed incidence in one epidemic on week t:

- at the observation level:

$$O_t \sim \text{Binom}(I_t, \rho)$$

Reporting rate

One disease, one island: model

Let O_t be the observed incidence in one epidemic on week t:

- at the observation level:

Reporting rate

$$O_t \sim \text{Binom}(I_t, \rho)$$

- at the transmission level:

Transmission

Exposure

Susceptibles

$$I_t \sim \text{Binom}\left(S_t, \beta \frac{1}{N} \sum_{n=1}^{5} w_{t,n} I_{t-n}\right)$$

One disease, one island: model

Let O_t be the observed incidence in one epidemic on week t:

- at the observation level:

Reporting rate

$$O_t \sim \text{Binom}(I_t, \rho)$$

- at the transmission level:

Transmission

Exposure

Susceptibles

$$I_t \sim \text{Binom}\left(S_t, \beta \frac{1}{N} \sum_{n=1}^{5} w_{t,n} I_{t-n}\right)$$

- both levels simplify into:

Overdispersion

$$O_t \sim \text{Neg-Binom}\left(S_t \frac{\beta}{N} \sum_{n=1}^{5} w_{t,n} \frac{O_{t-n}}{\rho}, \phi\right)$$

One disease, one island: inference

We want to infer the values of 3 parameters $\{\rho, \beta, \phi\}$ from data using a likelihood-based method:

One disease, one island: inference

We want to infer the values of 3 parameters $\{\rho, \beta, \phi\}$ from data using a likelihood-based method:

$$\Pr(O_t|\rho, \beta, \phi) = \text{Neg-Binom}\left(S_t \frac{\beta}{N} \sum_{n=1}^{5} w_{t,n} \frac{O_{t-n}}{\rho}, \phi\right)$$

One disease, one island: inference

We want to infer the values of 3 parameters $\{\rho, \beta, \phi\}$ from data using a likelihood-based method:

$$\Pr(O_t|\rho, \beta, \phi) = \text{Neg-Binom}\left(S_t \frac{\beta}{N} \sum_{n=1}^{5} w_{t,n} \frac{O_{t-n}}{\rho}, \phi\right)$$

- maximum likelihood $\rightarrow$ point-estimate + interval

One disease, one island: inference

We want to infer the values of 3 parameters $\{\rho, \beta, \phi\}$ from data using a likelihood-based method:

$$\Pr(O_t|\rho, \beta, \phi) = \text{Neg-Binom}\left(S_t \frac{\beta}{N} \sum_{n=1}^{5} w_{t,n} \frac{O_{t-n}}{\rho}, \phi\right)$$

- maximum likelihood $\rightarrow$ point-estimate + interval
- Bayesian approach $\rightarrow$ posterior distributions:

$$\Pr(\rho, \beta, \phi|O_t) \propto \Pr(O_t|\rho, \beta, \phi) \Pr(\rho, \beta, \phi)$$

One disease, one island: priors

We need to choose prior distributions for $\{\rho, \beta, \phi\}$:

- informative priors: reflect our prior knowledge about the values

[2] Gelman et al, The Annals of Applied Statistics (2008)

One disease, one island: priors

We need to choose prior distributions for $\{\rho, \beta, \phi\}$:

- informative priors: reflect our prior knowledge about the values
- non-informative priors: flat priors (problematic)

[2] Gelman et al, The Annals of Applied Statistics (2008)

One disease, one island: priors

We need to choose prior distributions for $\{\rho, \beta, \phi\}$:

- informative priors: reflect our prior knowledge about the values
- non-informative priors: flat priors (problematic)
- weakly-informative priors[2]: reflect our knowledge about the magnitude of the values

$$\rho \sim \text{Beta}(1, 1)$$

$$\beta \sim \text{Exponential}(0.1)$$

$$\phi \sim \text{Half-Cauchy}(2.5)$$

[2] Gelman et al, The Annals of Applied Statistics (2008)

One disease, one island: priors

We check the adequacy of these choices by conducting a prior predictive check[3]:

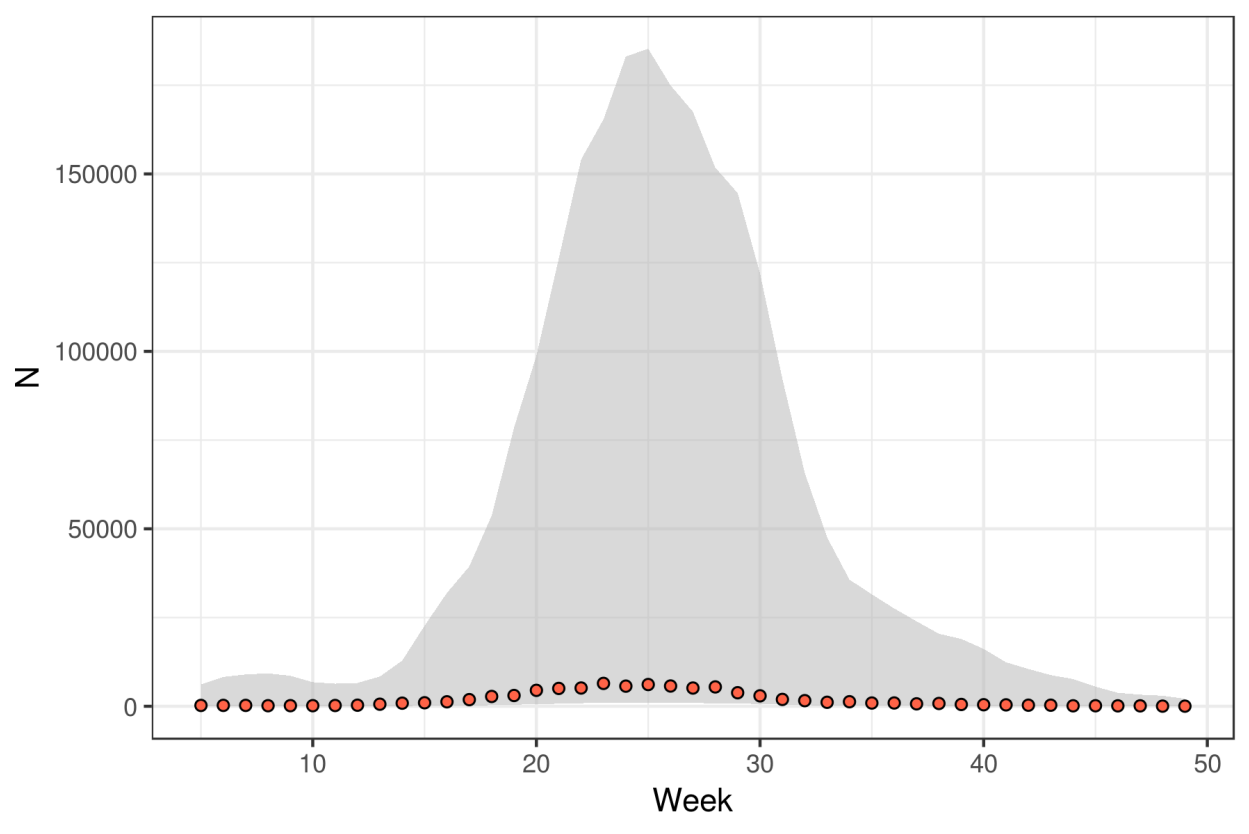


Figure 11: Prior predictive check for the epidemic of Zika virus in Guadeloupe.

[3] Gabry et al, Journal of the Royal Statistical Society (2019)

One disease, one island: implementation in Stan

In a separate `.stan` file:

- Data block:

```
data {
  int<lower=1> W; // number of records
  int<lower=0> O_t[W]; // number of reported cases at time t
  real<lower=0> Ostar_t[W]; // exposure at time t
  int<lower=0> sumO_t[W]; // cumulative number of reported cases at time t
  int<lower=0> pop; // island population
}
```

One disease, one island: implementation in Stan

In a separate `.stan` file:

- Data block:

```
data {
  int<lower=1> W; // number of records
  int<lower=0> O_t[W]; // number of reported cases at time t
  real<lower=0> Ostar_t[W]; // exposure at time t
  int<lower=0> sumO_t[W]; // cumulative number of reported cases at time t
  int<lower=0> pop; // island population
}
```

- Parameters block:

```
parameters {
  real<lower=0> beta;
  real<lower=0,upper=1> rho;
  real<lower=0> phi;
}
```

One disease, one island: implementation in Stan

- Transformed parameters block:

```
transformed parameters {
  real<lower=0> lp[W];
  real<lower=0> sampledisp[W];
  for(i in 1:W) {
    lp[i] = ( 1 - sumO_t[i] / (rho * pop)) * beta * Ostar_t[i] ;
    sampledisp[i] = lp[i]/phi;
  }
}
```

NB: $S_t \frac{\beta}{N} \sum_{n=\mathbf{1}}^{\mathbf{5}} w_{t,n} \frac{O_{t-n}}{\rho} = \left(1 - \frac{\sum_t O_t}{\rho N}\right) \beta O_t^*$

One disease, one island: implementation in Stan

- Transformed parameters block:

```
transformed parameters {
  real<lower=0> lp[W];
  real<lower=0> sampledisp[W];
  for(i in 1:W) {
    lp[i] = ( 1 - sumO_t[i] / (rho * pop)) * beta * Ostar_t[i] ;
    sampledisp[i] = lp[i]/phi;
  }
}
```

NB: $S_t \frac{\beta}{N} \sum_{n=1}^{5} w_{t,n} \frac{O_{t-n}}{\rho} = \left(1 - \frac{\sum_t O_t}{\rho N}\right) \beta O_t^*$

- Model block:

```
model {
  beta ~ exponential(0.1);
  rho ~ beta(1,1);
  phi ~ cauchy(0,2.5);
  // likelihood
  target += neg_binomial_2_lpmf(O_t|lp,sampledisp);
}
```

One disease, one island: implementation in Stan

Control `Stan` from `R` with `library(rstan)`:

```
## Select CHIKV in Guadeloupe
GUAD = filter(zikachik,ISLAND=="GUADELOUPE",VIRUS=="CHIKV",WEEK>=5)
## Format data for rstan
GUAD_L = list(W=dim(GUAD)[[1]],
              O_t=GUAD$NCASES,
              Ostar_t=GUAD$Ostar,
              sumO_t=GUAD$CUM_NCASES,
              pop=GUAD$POP[[1]])
## Sample
S_GUAD = stan("TSIR_one_island.stan",data=GUAD_L)
```

One disease, one island: implementation in Stan

Control `Stan` from `R` with `library(rstan)`:

```
## Select CHIKV in Guadeloupe
GUAD = filter(zikachik,ISLAND=="GUADELOUPE",VIRUS=="CHIKV",WEEK>=5)
## Format data for rstan
GUAD_L = list(W=dim(GUAD)[[1]],
              O_t=GUAD$NCASES,
              Ostar_t=GUAD$Ostar,
              sumO_t=GUAD$CUM_NCASES,
              pop=GUAD$POP[[1]])
## Sample
S_GUAD = stan("TSIR_one_island.stan",data=GUAD_L)
```

Results: posterior distributions of β, ρ and ϕ

```
> print(S_GUAD,pars=c("beta","rho","phi"))
Inference for Stan model: TSIR_one_island.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

      mean se_mean    sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
beta  1.54    0.00  0.06  1.42  1.50  1.53  1.58  1.66  1632    1
rho   0.33    0.00  0.02  0.30  0.32  0.33  0.34  0.38  1539    1
phi  51.04    0.24 11.47 33.31 42.74 49.42 57.74 77.75  2258    1

Samples were drawn using NUTS(diag_e) at Tue May 28 13:04:45 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

One disease, one island: implementation in Stan

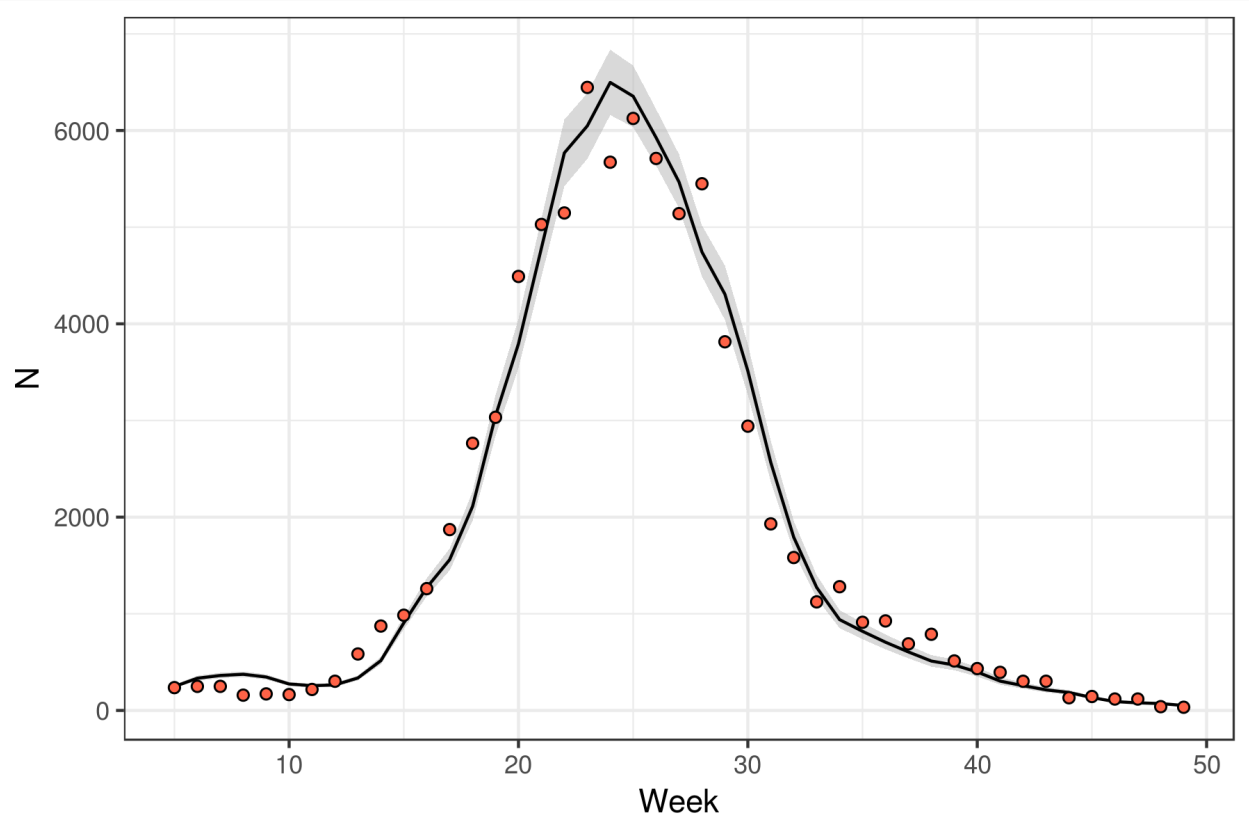


Figure 12: Model fit for the epidemic of Zika virus in Guadeloupe.

Now for 9 islands

Degrees of pooling:

- independent β_i and ρ_i for each island: no pooling

Now for 9 islands

Degrees of pooling:

- independent β_i and ρ_i for each island: no pooling
- the same β and ρ for all islands: complete pooling

Now for 9 islands

Degrees of pooling:

- independent β_i and ρ_i for each island: no pooling
- the same β and ρ for all islands: complete pooling
- *correlated* β_i and ρ_i for each island: partial pooling
  $=$ multilevel or hierarchical

Now for 9 islands

For the epidemic in island i, we have:

- a transmission parameter β_i which depends on hyperparameters μ_β and σ_β:

$$\ln \beta_{i,j=0} \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$$

Now for 9 islands

For the epidemic in island i, we have:

- a transmission parameter β_i which depends on hyperparameters μ_β and σ_β:

$$\ln \beta_{i,j=0} \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$$

- a reporting parameter ρ_i which depends on hyperparameters μ_ρ and σ_ρ:

$$\ln \frac{\rho_{i,j=0}}{1-\rho_{i,j=0}} \sim \mathcal{N}(\mu_\rho, \sigma_\rho^2)$$

Now for 9 islands

For the epidemic in island i, we have:

- a transmission parameter β_i which depends on hyperparameters μ_β and σ_β:

$$\ln \beta_{i,j=0} \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$$

- a reporting parameter ρ_i which depends on hyperparameters μ_ρ and σ_ρ:

$$\ln \frac{\rho_{i,j=0}}{1-\rho_{i,j=0}} \sim \mathcal{N}(\mu_\rho, \sigma_\rho^2)$$

$\Rightarrow$ We now also estimate μ_β, μ_ρ, σ_β and σ_ρ

Now for 9 islands and 2 diseases

We now model together the epidemics of ZIKV ($j = 1$) and CHIKV ($j = 0$) assuming proportionality:

- on the transmission parameters of ZIKV and CHIKV:

$$\beta_{i,j=1} = \eta \times \beta_{i,j=0}$$

- on the reporting parameters of ZIKV and CHIKV (on the logit scale):

$$\frac{\rho_{i,j=1}}{1 - \rho_{i,j=1}} = \omega \times \frac{\rho_{i,j=0}}{1 - \rho_{i,j=0}}$$

Now for 9 islands and 2 diseases

We now model together the epidemics of ZIKV ($j = 1$) and CHIKV ($j = 0$) assuming proportionality:

- on the transmission parameters of ZIKV and CHIKV:

$$\beta_{i,j=1} = \eta \times \beta_{i,j=0}$$

- on the reporting parameters of ZIKV and CHIKV (on the logit scale):

$$\frac{\rho_{i,j=1}}{1 - \rho_{i,j=1}} = \omega \times \frac{\rho_{i,j=0}}{1 - \rho_{i,j=0}}$$

$\Rightarrow$ We now also estimate η and ω

Now for 9 islands and 2 diseases

The model fit is acceptable for CHIKV:

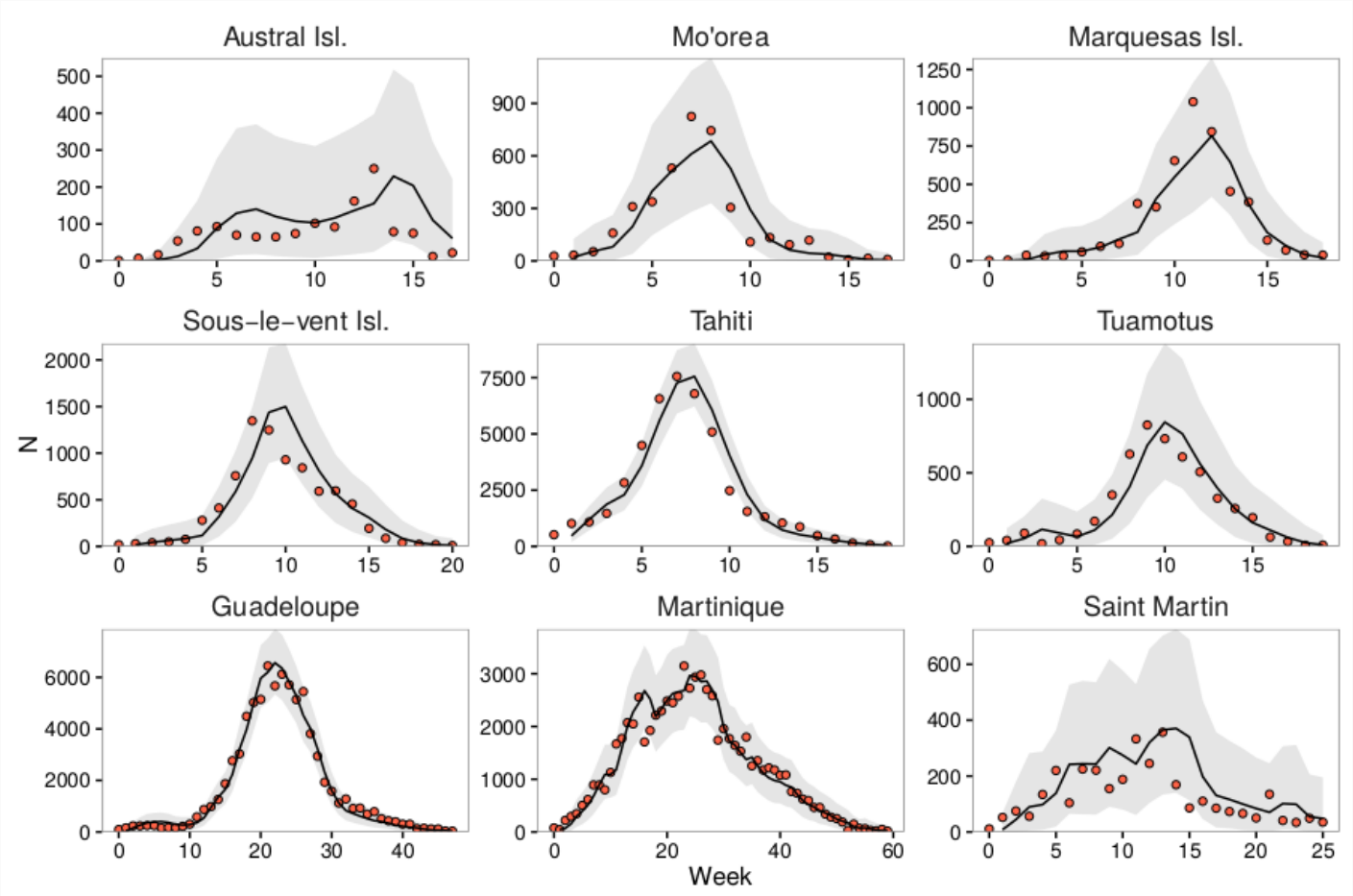


Figure 13: Model fit for the CHIKV epidemics.

Now for 9 islands and 2 diseases

And for ZIKV:

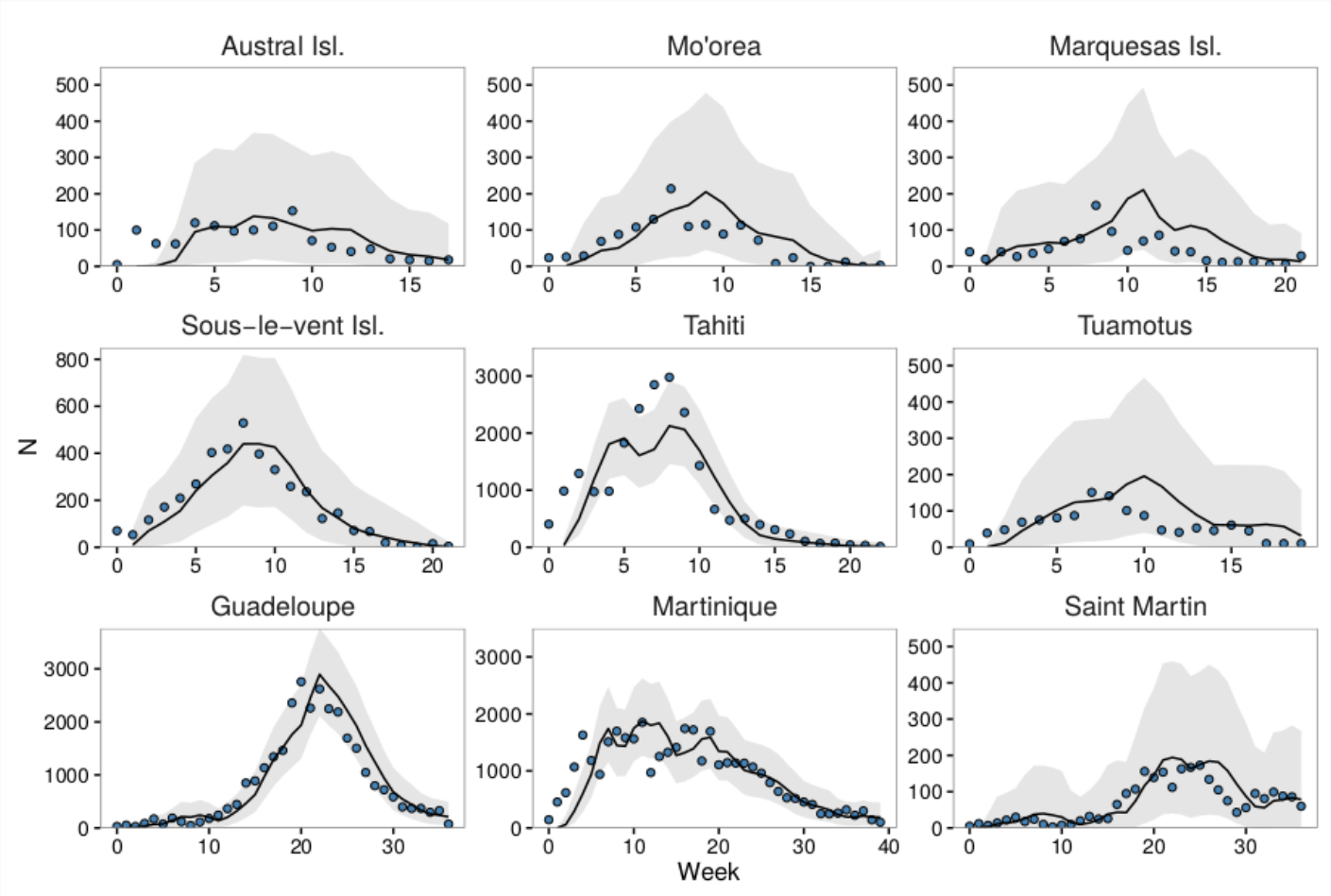


Figure 14: Model fit for the ZIKV epidemics.

Now for 9 islands and 2 diseases

The main results are:

- a similar transmissibility of CHIKV and ZIKV within an area
  - $\eta|\text{data} = 1.04\ [0.97 - 1.13]$
- a lower reporting rate for ZIKV
  - $\omega|\text{data} = 0.37\ [0.34 - 0.40]$

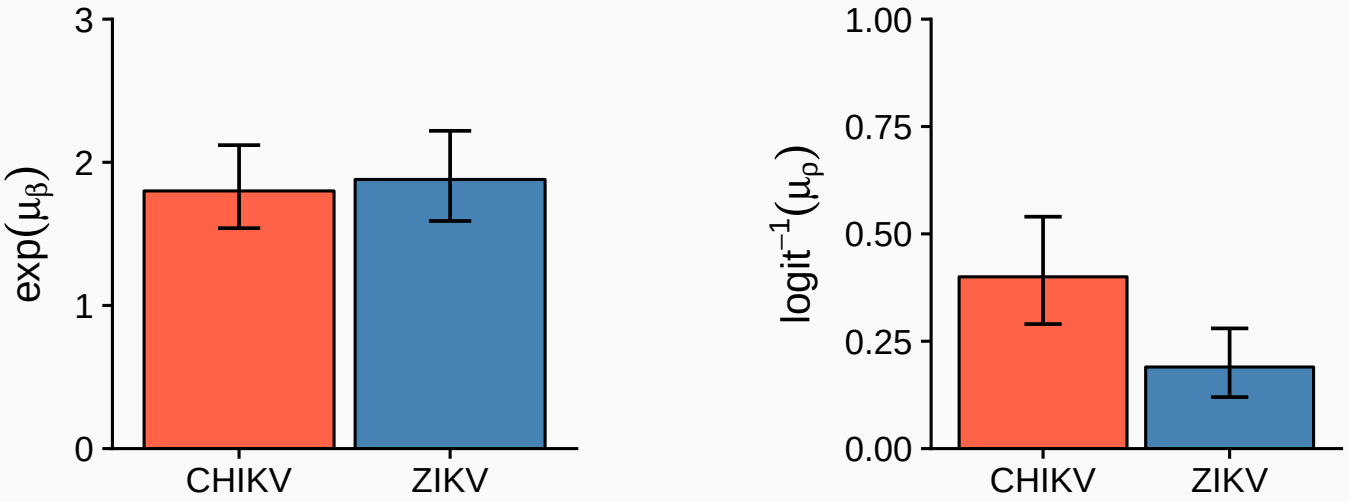


Figure 15: Posterior estimates of μ_β and μ_ρ for CHIKV and ZIKV.

Now for 9 islands and 2 diseases

We also find heterogeneity between areas:

- $\sigma^2_\beta|\text{data} > 0$, lower β in the French West Indies
- $\sigma^2_\rho|\text{data} > 0$, higher ρ in small islands and in Martinique

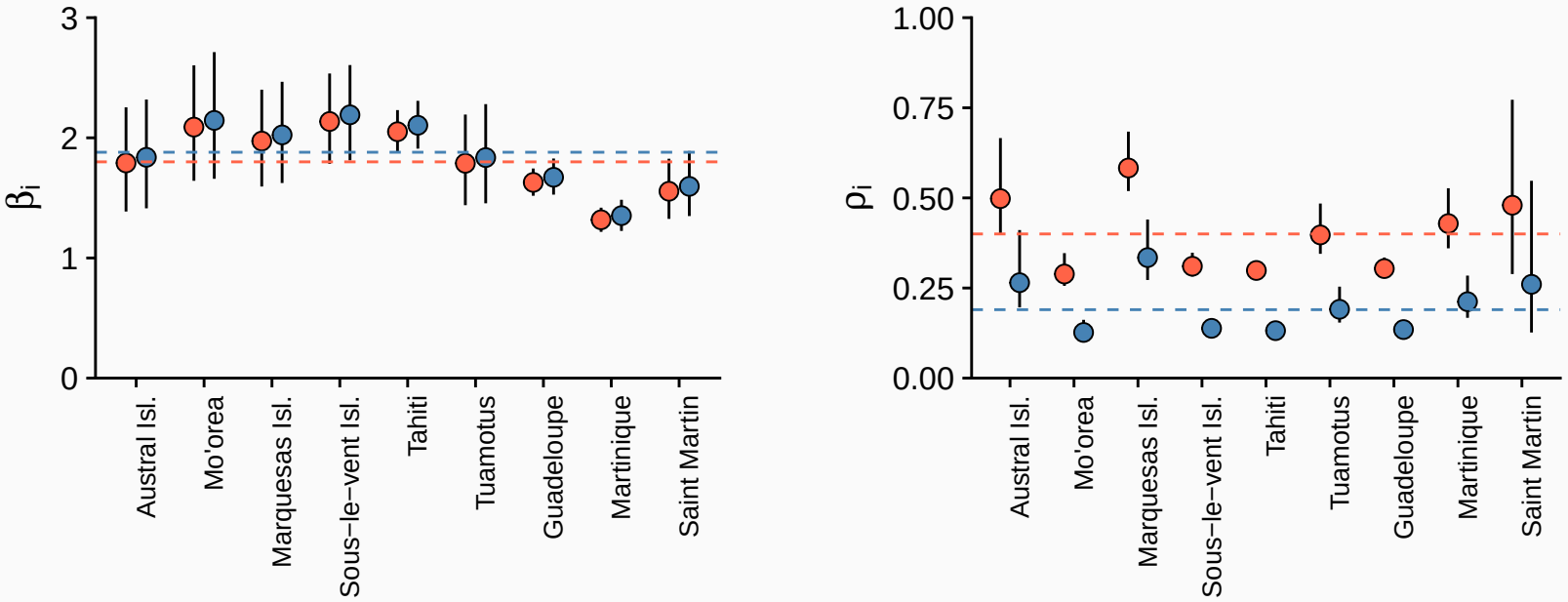


Figure 16: Island-specific posterior estimates of β and ρ for CHIKV (red) and ZIKV (blue).

Conclusion

Remember about:

- Data-generating mechanisms
- Prior predictive checks
- Plot model predictions (fit)
- Multi-level structure following data structure (partial pooling)