Practical I: epidemiological surveillance and modelling (non-spatial)

26/27th January

In this first practical the student is expected to learn about epidemiological case count data and standard techniques to visualise and analyse them. Further, the student is expected to reflect on the issues these data present and provide a perspective on how to improve disease surveillance in order to better estimate epidemiological parameters in the future.

The practical will start with a short 15-20 minute presentation by the course convenor and demonstrators. The next 2h will be spent answering predefined questions before a short wrap up and reflection at the end.

There will be a 15 minute break after ca. 1h ½.

Practical I Questions

Q1: How is public health surveillance for infectious diseases defined? (ca. 5 minutes)

Model answer:

The overall aim of public health surveillance is to ensure that the right information is available at the right time to make informed public health decisions.

Public health surveillance needs to be adaptable so that demographic and environmental driving changes in disease incidence and prevalence can be understood.

Public health surveillance has a long history and often falls under the remit of public health agencies in different countries. In the UK for example surveillance is carried out by the United Kingdom Health Security Agency (UKHSA) and performed by clinics, GPs, and hospitals among others.

Q2: What are common disease surveillance data and what are their pros and cons? Name 4 at most. (ca. 15 minutes)

Model answer:

Table 1: Table continues below

Data.Type	Pros	Cons
Syndromic surveillance	• Early detection • Relatively cheap	• Higher chances for misdiagnosis • Little information on pathogen evolution
Genomic surveillance	• Detailed pathogen information • Allows for downstream phylogenetic analysis and tracking of evolution (e.g. Drug resistance and Mutations) • Identify new/emerging pathogens	• Relatively expensive • Requires special expertise • Tends to be slower at detecting new outbreaks
Serological surveillance	• Overall burden of disease in population • Past exposure to pathogen(s)	• May not reflect current infections • Susceptible to cross-reactivity • Expensive (REACT, ONS Surveys)
Wastewater surveillance	• Early detection • Can measure overall disease burden • Cost-effective + non-invasive	• Difficult to identify individuals infected

Q3: What are some of the common challenges with public health surveillance at the beginning of disease outbreaks? (ca. 5 minutes)

Model answer:

- Speed of collection, cleaning and sharing of data
- Non-representativeness
- Changing case definitions
- Testing inaccuracies (false negatives / positives)

Q4: Describe the key data sources necessary to estimate the time-varying reproduction number Rt: https://academic.oup.com/aje/article/178/9/1505/89262? Focus on main manuscript and high level insights (ca. 20 minutes)

Model answer:

- Date of onset of cases (if daily data are reported)
- Number of cases
- Serial interval distribution (the time between the onset of symptoms in a primary case and the onset of symptoms of secondary cases)

Q5: Please install the R-package 'EpiEstim' and load it (https://cran.r-project.org/web/packages/EpiEstim/index.html), and familiarize yourself with the documentation. Remove any objects from your workspace. Copy the code to install the package below. (ca. 15 min)

Model answer: rm(list = ls()) install.packages('EpiEstim') library(EpiEstim)

Q6: Install the most used visualisation package in R that is recommended by 'EpiEstim'

Model answer: install.packages('ggplot2') library(ggplot2)

Q7: Load and visualise flu incidence data contained in the package. Plot the figure and add a caption to the plot.

```
data(Flu2009)
library(incidence)
plot(as.incidence(Flu2009$incidence$I, dates = Flu2009$incidence$dates))
```

Q8: Estimating time varying reproduction number, Rt.

We can run estimate_R on the incidence data to estimate the time varying reproduction number R. For this, we need to specify i) the time window(s) over which to estimate Rt and ii) information on the distribution of the serial interval (SI).

For i), the default behavior is to estimate Rt over weekly sliding windows (for example: window 1 = 01Jan2020 to 07Jan2020, window 2 = 02Jan2020 to 08Jan2020, window 3 = 03Jan2020 to 09Jan2020, etc). This can be changed through the config$t_start and config$t_end arguments (see below, "Changing the time windows for estimation"). For ii), there are several options, specified in the method argument.

The simplest is the parametric_si method, where you only specify the mean and standard deviation of the SI.

In this example, we only specify the mean and standard deviation of the serial interval. In the following example, we use the mean (2.6 days) and standard deviation (1.5) of the serial interval for flu from Ferguson et al., Nature, 2005: https://www.nature.com/articles/nature04017

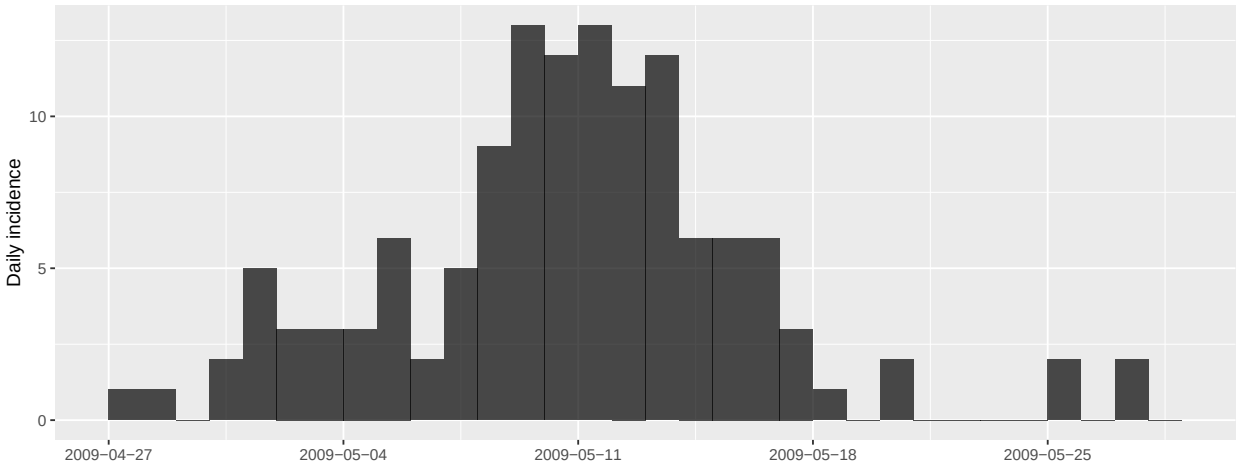
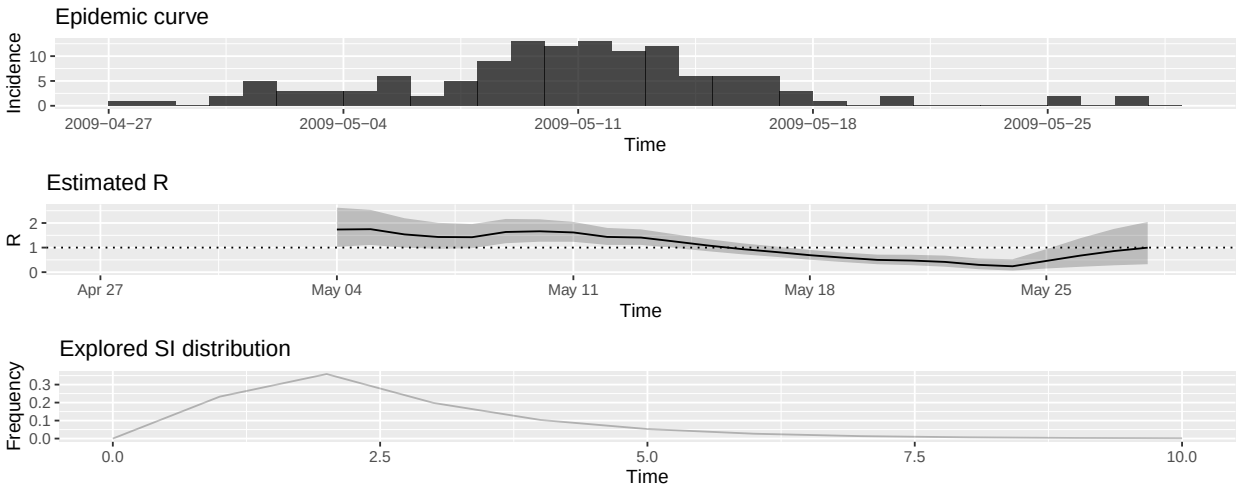


Figure 1: Daily incidence of Flu between April and May 2009.

Plot the code below and explain each panel. (ca. 10 minutes)

```
res_parametric_si <- estimate_R(Flu2009$incidence,
method="parametric_si",
config = make_config(list(
mean_si = 2.6, std_si = 1.5)))
plot(res_parametric_si, legend = FALSE)
```



The first plot shows the daily number of cases. The second plot shows the time varying reproduction number, Rt and the third plot the serial interval distribution used to estimate the time varying reproduction number.

Q9: Based on the above plot, describe the epidemiological dynamics of the outbreak by referencing the change in the time-varying reproduction number, Rt. (ca. 10 minutes)

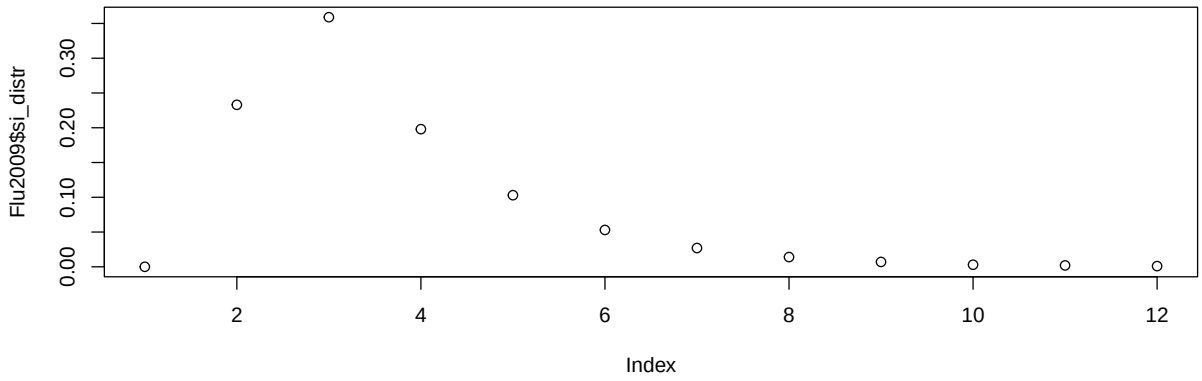
Model answer: Sporadic cases detected in the first week and half of the outbreak when cases started rising quickly from the 7-11th of May. Cases plateaued and declined. Rt dropped below 1 around 14/15 of May.

Q10: Estimating Rt with a non parametric serial interval distribution

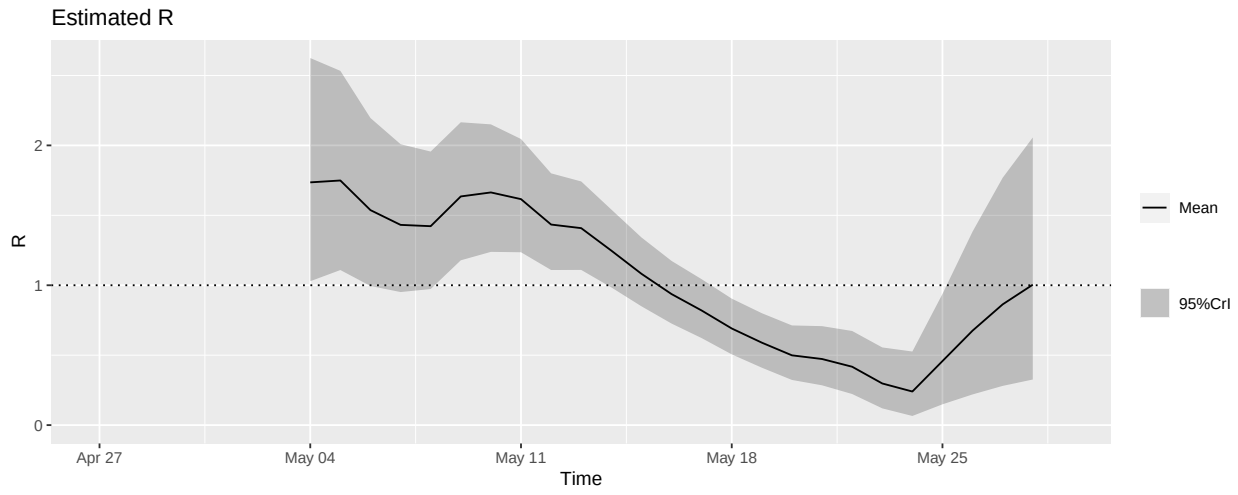
If one already has a full distribution of the serial interval, and not only a mean and standard deviation, this can be fed into estimate_r as follows.

Plot serial interval distribution and outputs from this model. Provide a perspective on what the SI distribution means for estimating transmission. (ca. 15 minutes)

```
plot(Flu2009$si_distr)
```




```
res_non_parametric_si <- estimate_R(Flu2009$incidence,                    method="no
config = make_config(list(si_distr = Flu2009$si_distr)))
plot(res_non_parametric_si, "R")
```



The estimation method used in this work is developed for the ideal situation in which times of infection are known and the infectivity profile may be approximated by the distribution of the generation time (i.e., time from the infection of a primary case to infection of the cases he/she generates). However, times of infection are rarely observed, and the generation time distribution is therefore difficult to measure. On the other hand, the timing of onset of symptoms is usually known, and such data collected in closed settings where transmission can reliably be ascertained (e.g., households) can be used to estimate the distribution of the serial interval (time between onset of symptoms of a case and onset of symptoms of his/her secondary cases). Therefore, in practice, we apply our method to data consisting of daily counts of onset of symptoms where the infectivity profile is approximated by the distribution of the serial interval.Non-parametric distribution implies that the distribution of SI is "distribution-free", i.e. we do not have to make any assumptions about the shape or form of the distribution. In practice when we have empirical data on serial interval (or generation time) distribution, we can use that instead of specifying a mean and standard deviation.

Q11: Estimating R accounting for uncertainty on the serial interval distribution

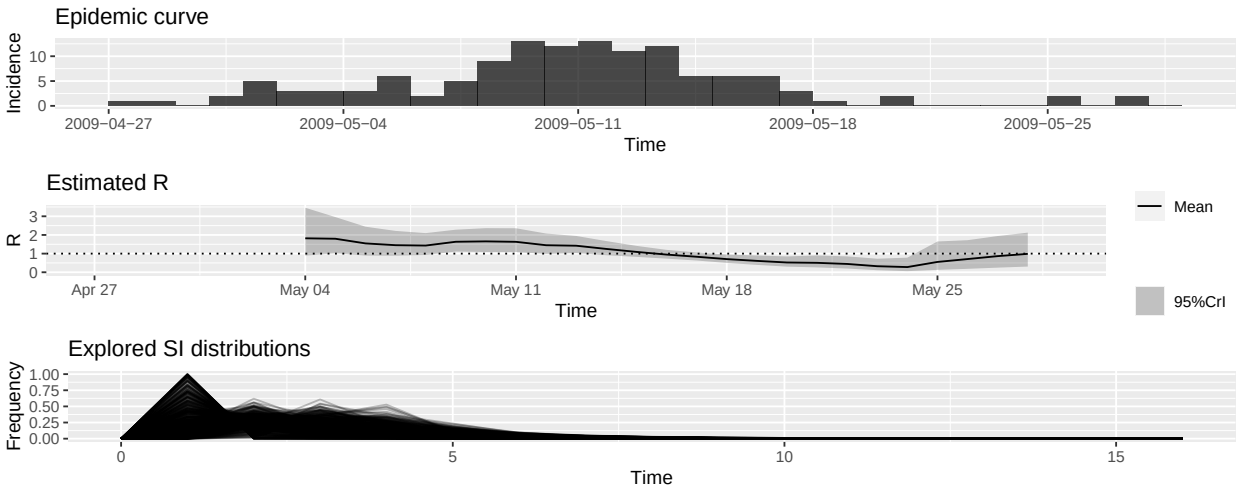
Sometimes, especially early in outbreaks, the serial interval distribution can be poorly specified. Therefore estimate_R also allows integrating results over various distributions of the serial interval. To do so, the mean and sd of the serial interval are each drawn from truncated normal distributions, with parameters specified by the user, as in the example below:

We choose to draw:

- The mean of the SI in a Normal(2.6, 1), truncated at 1 and 4.2. This truncated normal distribution is centred around 2.6, which was the mean of the SI distribution in Q8. Here we add uncertainty in the mean parameter, but restrict it to lie between 1 and 4.2.
- The sd of the SI in a Normal(1.5, 0.5), truncated at 0.5 and 2.5. Similarly, here we add uncertainty to the standard deviation of 1.5 described in Q8, and limit this to lie between the reasonable range of 0.5 and 2.5.

Provide the outputs from the model and describe the bottom plot in light of the uncertainty about the SI distribution.

```
config <- make_config(list(mean_si = 2.6, std_mean_si = 1,
min_mean_si = 1, max_mean_si = 4.2,
std_si = 1.5, std_std_si = 0.5,
min_std_si = 0.5, max_std_si = 2.5))
res_uncertain_si <- estimate_R(Flu2009$incidence,
method = "uncertain_si", config = config)
plot(res_uncertain_si)
```



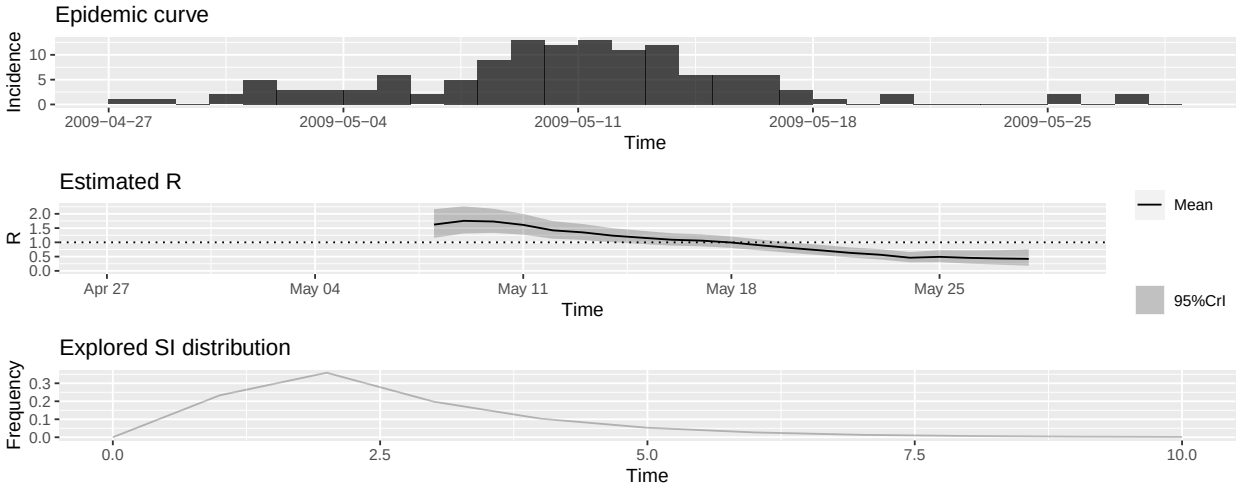
The bottom plot shows the SI distributions explored in this model. In contrast to Q8 where we only explored one SI distribution.

Q12: Changing the time windows for estimation

The time window can be specified through arguments config$t_start and config$t_end. For instance, the default weekly sliding windows can also be obtained by specifying:

Describe how the estimate of Rt compares to previous estimates (ca. 15 minutes):

```
T <- nrow(Flu2009$incidence)
t_start <- seq(2, T-10)
t_end <- t_start + 10 # adding 10 to get 11-day windows as bounds included in window
res_weekly <- estimate_R(Flu2009$incidence,
method="parametric_si",
config = make_config(list(t_start = t_start,
t_end = t_end,mean_si = 2.6, std_si = 1.5)))
plot(res_weekly)
```



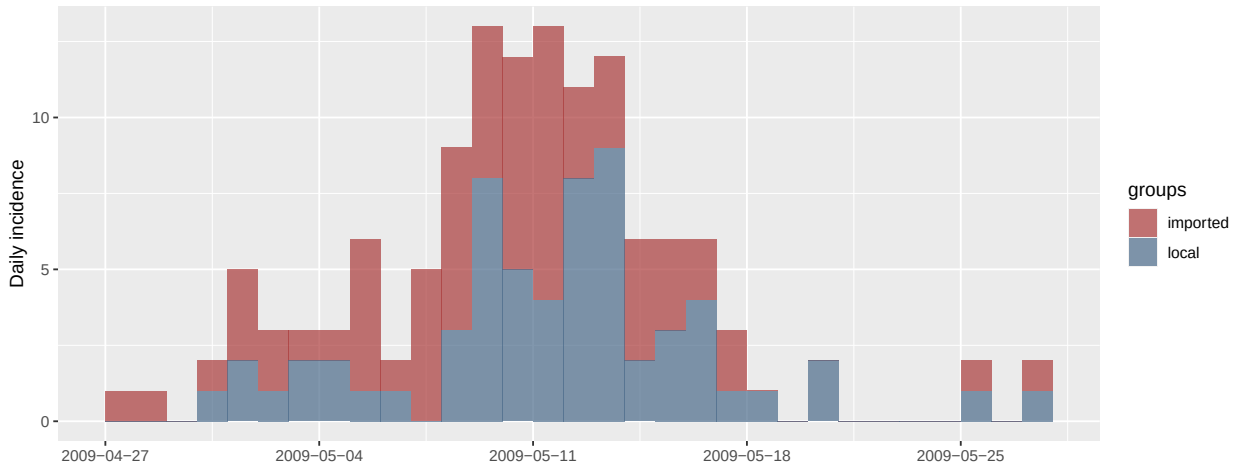
Estimates are more smooth due to aggregation of cases through time. Small variations in case numbers do not impact Rt estimation. However we would not want to use sliding windows of size 30 days because we estimate one R per time frame and assuming a constant reproduction number for 30 days would not be realistic in the case of a fast spreading pathogen such as Influenza.

Q13: Specifying imported cases

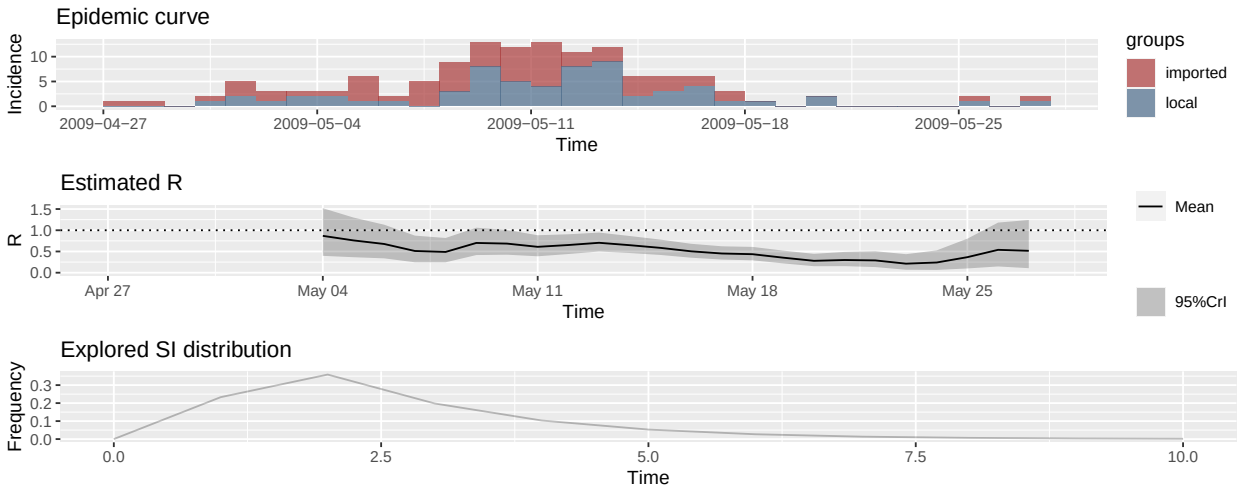
All of the above assumes that all cases are linked by local transmission (eg- within country). Sometimes you may have information (from field epidemiological investigations for instance) indicating that some cases are in fact imported (from another location (eg internationally), or from an animal reservoir). See some more details on the estimation in this publication: https://www.sciencedirect.com/science/article/pii/S1755436519300350?via%3Dihub We allow to include such information, when available, as illustrated in the example below: generating fake information on our cases:

Compare estimates of Rt for this model compared to earlier models that do not consider imported cases. What is the impact on local Rt? (ca. 15 minutes)

```
dates_onset <- Flu2009$incidence$dates[unlist(lapply(1:nrow(Flu2009$incidence),
  function(i)
  rep(i, Flu2009$incidence$I[i])))]

location <- sample(c("local","imported"), length(dates_onset), replace=TRUE)

location[1] <- "imported" # forcing the first case to be imported

## get incidence per group (location)
incid <- incidence(dates_onset, groups = location)

plot(incid)
```



```
## Estimate R with assumptions on serial interval:
res_with_imports <- estimate_R(incid, method = "parametric_si",
config = make_config(list(mean_si = 2.6, std_si = 1.5)))

# Default config will estimate R on weekly sliding windows.
# To change this change the t_start and t_end arguments.
plot(res_with_imports, add_imported_cases=TRUE)
```



Earlier drop in Rt below 1. Note that in the above the estimated reproduction number is, as expected, much lower than estimated before when assuming that all cases were linked by local transmission. Find more details here: https://www.sciencedirect.com/science/article/pii/S1755436519300350?via%3Dihub

Q14: Name other sources of uncertainty in the Rt estimation?

Model answer:

- Under-reporting
- Spatial scale of reporting
- Health care seeking