








# Statistical Models for Infectious Diseases Dynamics

---

Prof. Luca Scrucca

-  Università degli Studi di Perugia
-  [luca.scrucca@unipg.it](mailto:luca.scrucca@unipg.it)
-  <http://www.stat.unipg.it/luca>
-  [luca-scr](#)
-  [luca\\_scr](#)

17 September 2021

1. Parametric growth curve regression model to fit incidence data for real-time monitoring and short-term forecasting of main epidemiological indicators of COVID-19 pandemic (Alaimo Di Loro et al., 2021).

**StatGroup-19** <https://www.facebook.com/StatGroup19>

Dashboard: <https://statgroup19.shinyapps.io/Covid19App>

2. COVINDEX, a statistical model-based index for near real-time monitoring of pandemic evolution (Scrucca, 2021).

- **Dipartimento Protezione Civile** (Civil Protection Department, CPD) collected data since Feb 24th, 2020, at the national and regional levels. These are provided every day in a public GitHub repository at <https://github.com/pcm-dpc/COVID-19>.
- The [epidemiological data](#) provided by CPD can be distinguished into two basic types:
  1. [incidence](#) indicators (flows)
  2. [prevalence](#) indicators (stocks)
- COVID-19 public Italian data present several [data quality issues](#):
  - information is gathered and reported at a regional level, with each regional healthcare organization having a different transmission and data collection system;
  - measurement and data entry errors are expected;
  - delays in reporting are sometimes substantial (e.g. deaths are counted on the day of the reporting, not on the day of the outcome);
  - swabs and positive cases are not time-aligned, i.e. positives are counted on the day that test results are received, with swabs being processed from one day to several days after symptoms. No distinction between symptomatic and asymptomatic patients was made.
  - Finally, important to recall that people diagnosed with COVID-19 disease are only a small fraction of the people infected by the virus.

## Growth curve modelling for nowcasting COVID-19 incidence indicators

---

# Growth curve modelling for nowcasting COVID-19 incidence indicators

## Incidence indicators

Measure the number of individuals with a particular condition, related with the epidemic, recorded during a given period.

## Daily incidence counts

- positives
  - hospitalized (either in regular wards or in ICU)
  - isolated-at-home
- recovered
- deceased

**Cumulative incidence indicators** refer to longer time intervals and are obtained simply by cumulating indicators over time.

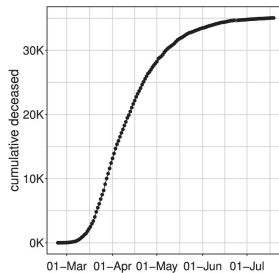
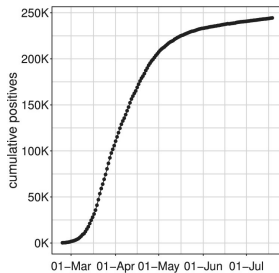
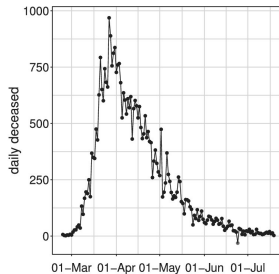
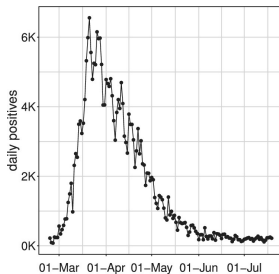
$$Y_t^C = Y_{t-1}^C + I_t$$

$Y_t^C$  = cumulative indicator at time  $t$

$Y_{t-1}^C$  = cumulative indicator at time  $t - 1$  ( $Y_0^C = 0$ )

$I_t$  = incidence indicator at time  $t$

- cumulative positives
- cumulative deceased
- cumulative recovered



Italian daily incidence (top) and cumulative (bottom) indicators  
during the 1st wave of COVID-19 pandemic

## Prevalence indicators

Measure the number of individuals with a particular condition, related with the epidemic, at a given instant in time or at a given short (e.g. a day) interval of time.

$$Y_t^P = Y_{t-1}^P + I_t - O_t$$

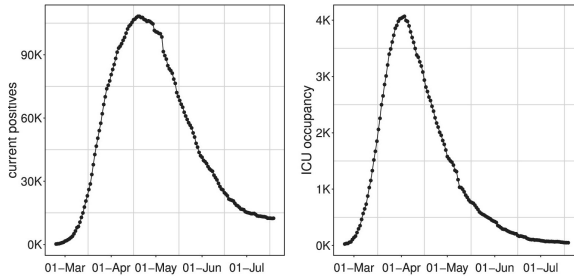
$Y_t^P$  = prevalence indicator at time  $t$

$Y_{t-1}^P$  = prevalence indicator at time  $t - 1$  ( $Y_0^P = 0$ )

$I_t$  = input at time  $t$

$O_t$  = output at time  $t$

- current positives (at time  $t$  they are given by the current positives at time  $(t - 1)$  plus the daily positives at day  $t$  and minus the sum of deceased and recovered at day  $t$ )
- current intensive care units (ICU) occupancy



Italian daily prevalence indicators during the 1st wave of COVID-19 pandemic



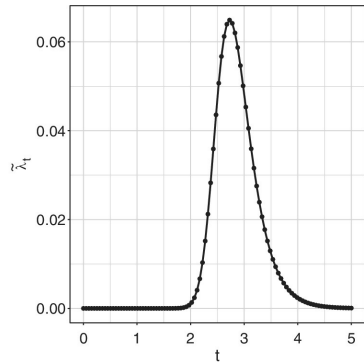
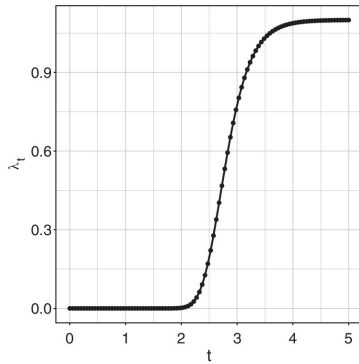
## Modelling incidence indicators

- The response function for a cumulative incidence indicator  $\{Y_t^C\}_{t=0}^T$  assumes that the expected value follow a Generalized Logistic growth curve, also known as the **Richards' growth curve**, given by the equation

$$\mathbb{E}(Y_t^C) = \lambda_{\gamma}(t) = b + \frac{r}{(1 + 10^{h(p-t)})^s}$$

where  $\gamma = [b, r, h, p, s]^T$  is the vector of parameters to be estimated.

- Parameters interpretation:
  - $b \in \mathbb{R}^+$ : lower asymptote
    - $r \in \mathbb{R}^+$ : distance between the upper and the lower asymptote, hence  $b + r$  is the final epidemic size
    - $h \in \mathbb{R}$ : infection/growth rate
    - $p \in \mathbb{R}$ : lag-phase of the trajectory (when the curve growth speed slows down) and determines the peak position
    - $s \in \mathbb{R}$ : asymmetry parameter regulating differences in the behaviour of the ascending and descending phase of the outbreak; if  $s = 1$  then the curve reduces to the logistic function, and if  $s \rightarrow \infty$  then the curve converges to the Gompertz function.
- Growth rate** is given by the first derivative of growth curve  $\tilde{\lambda}_{\gamma}(t) = \frac{\partial \lambda_{\gamma}(t)}{\partial t}$ .



Richards' curve and its derivative

## Richards GLM growth model

- The proposed model is an [Extended Generalized Linear Model](#) with response function given by the first differences of the Richards' curve

$$\tilde{\lambda}_{\boldsymbol{\gamma}}(t) = \mathbb{E}(Y_t^C) - \mathbb{E}(Y_{t-1}^C)$$

which is used to model the daily expected values

$$\tilde{\mu}_{\boldsymbol{\theta}}(t) = a + \tilde{\lambda}_{\boldsymbol{\gamma}}(t)$$

from the observed incidence counts  $y_t$  for  $t = 1, \dots, T$ .

In the unknown parameter vector  $\boldsymbol{\theta} = (a, \boldsymbol{\gamma})$ ,  $a$  is a kink effect/baseline parameter which can be interpreted as the endemic steady state incidence rate.

## Response distribution for incidence indicators

- Poisson is a standard distribution for counts:

$$Y_t | \boldsymbol{\theta} \sim \text{Pois}(\tilde{\mu}_{\boldsymbol{\theta}}(t)) \quad \text{for } t = 1, \dots, T$$

- However, since often counts are overdispersed, the [Negative Binomial distribution](#) is a better choice because includes a dispersion parameter  $v \in \mathbb{R}^+$ :

$$Y_t | \boldsymbol{\theta} \sim \text{NB}(\tilde{\mu}_{\boldsymbol{\theta}}(t), v) \quad \text{for } t = 1, \dots, T$$

## Covariates

- Covariates  $\mathbf{x}(t)$  can be included by specifying the [linear predictor](#)  $\boldsymbol{\beta}^\top \mathbf{x}(t)$ , so an additive effect of the covariates on the mean is expressed as

$$\tilde{\mu}_{\boldsymbol{\theta}}(t) = \exp\{\boldsymbol{\beta}^\top \mathbf{x}(t)\} + \tilde{\lambda}_{\gamma}(t)$$

- Note that an intercept in the linear predictor is equivalent to  $\log(a)$ .
- When modelling new positives a further term (a dummy variable) should be included to account for weekends (and holidays as well) effect.

## Model estimation

Estimates of unknown parameters  $\boldsymbol{\theta}$  can be obtained by the method of [maximum likelihood](#). However, maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{t=1}^T y_t \log \tilde{\mu}_{\boldsymbol{\theta}}(t) - \sum_{t=1}^T \tilde{\mu}_{\boldsymbol{\theta}}(t) + \text{constant}$$

is a difficult task:

- no analytical solution;
- numerical maximization required;
- given the nonsmooth shape of the objective function, a multistart strategy must be adopted based on a combination of genetic and gradient descent algorithms.

## Inference

Inference on model parameters can be obtained from the asymptotic distribution of MLEs,  $\widehat{\theta} \sim N(\theta, \widehat{V}^R(\theta))$ , where  $\widehat{V}^R(\theta)$  is a robust estimate of the covariance matrix given by the Huber's "sandwich estimator".

## Predictions

Predictions for both incidence and cumulative incidence counts can be derived using a simulation approach ([parametric double bootstrap](#)):

- resampled trajectories  $\{Y_i\}_{i=1}^B$  are obtained by simulating  $B$  sets of parameters from their asymptotic distribution and computing  $B$  mean functions trajectories  $\{\mu_{\theta_i}(t)\}_{i=1}^B$ .
- an artificial time series of counts is then simulated for each of the  $B$  trajectories and 95% confidence intervals are obtained by computing the pointwise 2.5% and 97.5% quantiles.

## Peak date prediction

- An important epidemiological information to predict is the [date of the peak](#).
- Once the model is estimated, and under the assumption that no modifications occur in epidemiological strategies, the peak can be computed analytically as

$$\widehat{t}_y = \widehat{p} + \frac{\log_{10}(\widehat{s})}{\widehat{h}}$$

# R Lab

└ 01_nowcasting.R	# code for data analysis
└└ RichardsGrowthGLM.R	# code implementing model fitting
└└ data.R	# code for download & reading data

## COVINDEX

---

## Motivation

- Detecting [changes in COVID-19 disease transmission over time](#) is a key indicator of epidemic growth.
- [Near real-time monitoring](#) of the pandemic growth is crucial for policy makers and public health officials who need to make informed decisions about whether to enforce lockdowns or allow certain activities.
- The [effective reproduction number](#)  $R_t$  is the standard index used in many countries for pandemic surveillance, but it has several drawbacks:
  - does not provide a timely snapshot of the evolution of the pandemic;
  - two-weeks delay is typically present due to the time lag between infection and case registration;
  - different estimation methods available provide different results, and no one showed to be uniformly superior to the others.
- COVINDEX is proposed as a simple near [real-time index for monitoring the evolution of the COVID-19 pandemic](#).



## Modelling test positive rate

- Consider the [test positive rate](#) (TPR) at time  $t$  defined as

$$y_t = \frac{P_t}{T_t}$$

where  $P_t$  is the number of new positive cases and  $T_t$  the number of tests (PCR – polymerase chain reaction – or molecular swabs).

- Since  $\text{TPR} \in [0,1]$ , a [beta distribution](#) can be postulated

$$y_t \sim \text{Beta}(\mu_t, \phi),$$

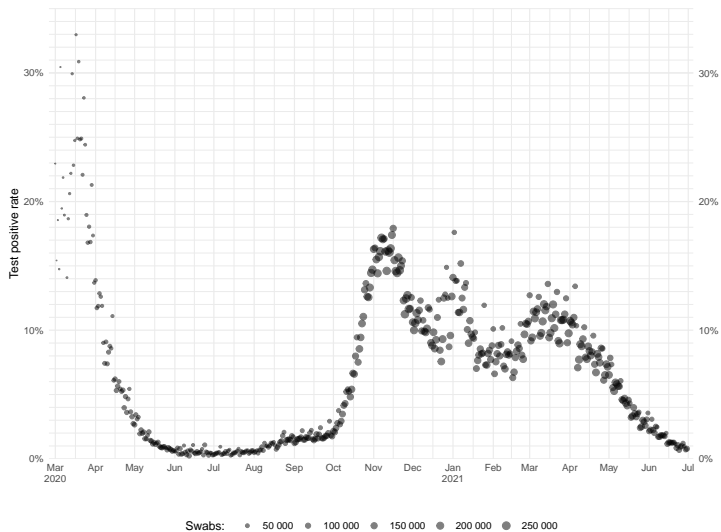
with

$$\mathbb{E}[y_t] = \mu_t \quad \text{and} \quad \mathbb{V}[y_t] = \frac{\mu_t(1 - \mu_t)}{1 + \phi}.$$

- The [beta regression model](#) expresses the mean  $\mu_t$  as a function of the linear predictor  $\eta_t = \boldsymbol{\beta}^\top \mathbf{x}_t$  using the logistic function

$$\mu_t = \text{logistic}(\eta_t) = \frac{\exp(\eta_t)}{1 + \exp(\eta_t)} = \frac{1}{1 + \exp(-\eta_t)},$$

where  $\boldsymbol{\beta}$  is a vector of unknown regression coefficients, and  $\mathbf{x}_t$  is the vector of observed values on a set of predictors.

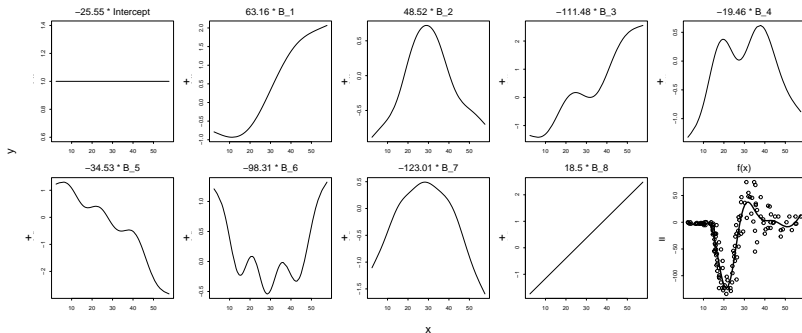


True positive rate as function of time in Italy

- [Generalized Additive Models](#) (GAMs) allows to model the dependence of a response variable in a flexible way using smooth functions of the predictors.
- In our context the time-dependence is modelled using a smooth function of time plus the inclusion of a dummy variable term (WE) to account for weekends (and holidays) effect:

$$\eta_t = \beta_0 + \sum_{k=1}^K \beta_k B_k(x_t) + \beta_{WE}(WE),$$

where  $\{B_k\}_{k=1}^K$  is the basis expansion of a thin plate regression spline (other smooth functions could be used as well).



- Estimation of the GAM model can be pursued by [REstricted Maximum Likelihood](#) (REML), which amounts to maximize the penalized log-likelihood

$$\ell_P(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\lambda\boldsymbol{\beta}^\top \mathbf{S}\boldsymbol{\beta},$$

where  $\ell(\boldsymbol{\beta}) = \sum_{t=1}^n \ell(y_t|\boldsymbol{\beta})$  is the log-likelihood for the observed values  $y_t$ .

The last term represents the smoothing penalty, with  $\lambda$  a smoothing parameter, and  $\mathbf{S}$  a known penalty matrix.

- The [number of administered swabs is not constant over time](#) because (i) on weekends and holidays the number of swabs drops drastically, (2) during periods of strong expansion of the pandemic, the monitoring system is unable to carry out effective surveillance and only symptomatic patients are likely to be tested, a [weighted penalized log-likelihood criterion](#) is adopted:

$$\ell_W(\boldsymbol{\beta}) = \sum_{t=1}^n w_t \ell(y_t|\boldsymbol{\beta}),$$

where  $w_t = T_t/\bar{T}$  are prior weights specifying the contribution of each data point to the log-likelihood.

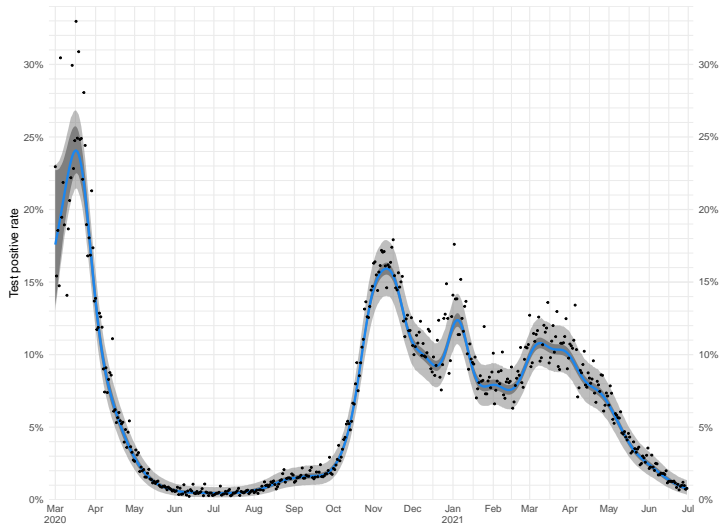
- Selection of the [smoothing parameter](#)  $\lambda$  can be obtained by minimizing the conditional Akaike's information criterion (AIC).

- The penalized likelihood approach described above has also a [Bayesian interpretation](#) by assuming an improper multivariate normal prior on  $\beta$ .
- REML estimates of  $\beta$  coefficients are asymptotically the [maximum a posteriori \(MAP\) of the Bayesian posterior distribution](#)

$$\beta | (y, \lambda) \sim N(\hat{\beta}, (\hat{I} + \lambda S)^{-1}),$$

where  $\hat{I}$  is the observed information matrix (Hessian of the negative log-likelihood) at  $\hat{\beta}$ .

- Approximate [credible intervals](#) for any function of  $\beta$ , including predictions, can be obtained by simulating from the posterior.



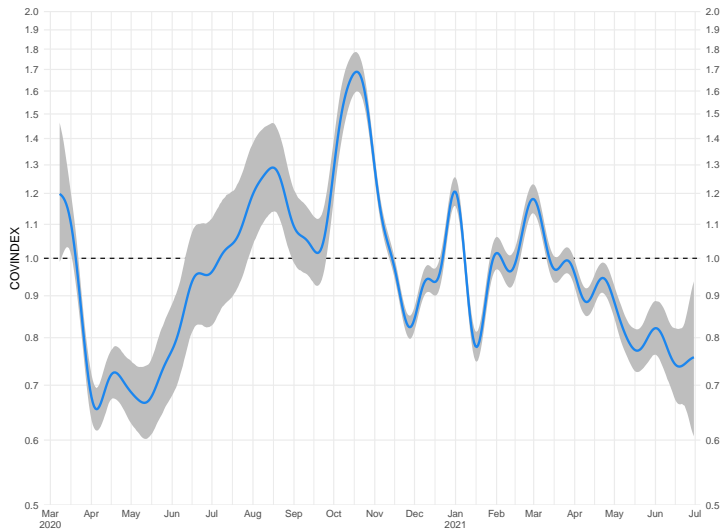
GAM beta regression fit with predictions and approximate 95% simulated credible intervals (dark grey for the mean and light grey for the single predictions) as function of time

## COVINDEX as a near real-time monitoring and decision-making tool

- COVINDEX is an attempt to compute a [synthetic index summarizing the evolution of the COVID-19 pandemic](#), which can be useful to policy makers and public health officials for monitoring local and national outbreaks.
- COVINDEX is defined as the [ratio of the predicted positive rate at time  \$t\$  to the prediction 7 days earlier](#):

$$\text{COVINDEX}_t = \frac{\hat{\mu}_t}{\hat{\mu}_{t-7}}$$

- Interpretation:
  - values larger than 1.0 means that the pandemic is growing
  - values smaller than 1.0 indicates that new infections are slowing down
- The comparison to 7-days back is chosen because it is approximately the expected incubation time for COVID-19 and because it corresponds to the observed weekly fluctuation in testing.
- Of course, uncertainty also affects the COVINDEX and the simulation approach adopted for TPR can be used here as well.



COVINDEX (on logarithmic scale) as function of time with approximate 95% simulated credible intervals



## R Lab

└─ 02_covindex.R	# code for data analysis
└─ CovindexGamBetaReg.R	# code implementing model fitting
└─ data.R	# code for download & reading data

- Alaimo Di Loro P, Divino F, Farcomeni A, Jona Lasinio G, Lovison G, Maruotti A, Mingione M. (2021) [Nowcasting COVID-19 incidence indicators during the Italian first outbreak](#). **Statistics in Medicine**, 40: 3843-3864. <https://doi.org/10.1002/sim.9004>
- Scrucia L. (2021) A COVINDEX based on a GAM beta regression model with an application to the COVID-19 pandemic in Italy. Under review. [arXiv:2104.01344](#) Pre-print available at <https://arxiv.org/abs/2104.01344>