From data to modelling: why statistics is fundamental to manage the epidemic

Antonello Maruotti

Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne.
Libera Università Maria Ss Assunta, Via Pompeo Magno 22 - 00192 Roma
a.maruotti@lumsa.it

September 2021

We are talking about...

Lessons from the pandemic: using and communicating data

Statistical modelling
- Short-term forecasting
- Medium-term forecasting

Concluding remarks

The number of new cases

TV news - 09/02/2021
“The number of new cases is increasing”

The number of new cases

TV news - 09/02/2021
“The number of new cases is increasing”

- New cases as of 09/02 ⇒ 10630
- New cases as of 08/02 ⇒ 7970
- Growth ⇒ +2660 → +33%

The number of new cases

TV news - 09/02/2021
"The number of new cases is increasing"

- New cases as of 09/02 ⇒ 10630
- New cases as of 08/02 ⇒ 7970
- Growth ⇒ +2660 → +33%

Though:

The number of new cases

TV news - 09/02/2021
"The number of new cases is increasing"

- New cases as of 09/02 ⇒ 10630
- New cases as of 08/02 ⇒ 7970
- Growth ⇒ +2660 → +33%

Though:

- Swabs as of 09/02 ⇒ 274263
- Swabs as of 08/02 ⇒ 144270
- Growth ⇒ +12993 → +90%

The number of new cases

TV news - 09/02/2021
"The number of new cases is increasing"

- New cases as of 09/02 ⇒ 10630
- New cases as of 08/02 ⇒ 7970
- Growth ⇒ +2660 → +33%

Though:

- Swabs as of 09/02 ⇒ 274263
- Swabs as of 08/02 ⇒ 144270
- Growth ⇒ +12993 → +90%

"...the positivity rate decreases..."

- Positivity rate as of 09/02 ⇒ 3.9%

The number of new cases

TV news - 09/02/2021
“The number of new cases is increasing”

- New cases as of 09/02 ⇒ 10630
- New cases as of 08/02 ⇒ 7970
- Growth ⇒ +2660 → +33%

Though:

- Swabs as of 09/02 ⇒ 274263
- Swabs as of 08/02 ⇒ 144270
- Growth ⇒ +12993 → +90%

“...the positivity rate decreases...”

- Positivity rate as of 09/02 ⇒ 3.9%
- Positivity rate as of 08/02 ⇒ 5.5%

Variants of interest

- The British variant: VOC202012/01 (lineage B.1.1.7)
  - 93.3%

Variants of interest

- The British variant: VOC202012/01 (lineage B.1.1.7)
  - 93.3% ⇒ Molise
  - 75.0%

Variants of interest

- The British variant: VOC202012/01 (lineage B.1.1.7)
  - 93.3% ⇒ Molise
  - 75.0% ⇒ Sardegna
  - 0.0%

Variants of interest

- The British variant: VOC202012/01 (lineage B.1.1.7)
  - 93.3% ⇒ Molise
  - 75.0% ⇒ Sardegna
  - 0.0% ⇒ Valle d’Aosta

Variants of interest

- The British variant: VOC202012/01 (lineage B.1.1.7)
    - 93.3% ⇒ Molise
    - 75.0% ⇒ Sardegna
    - 0.0% ⇒ Valle d'Aosta
- How many samples with a valid sequencing?
    - 15 ⇒ Molise
    - 12 ⇒ Sardegna
    - 1 ⇒ Valle d'Aosta

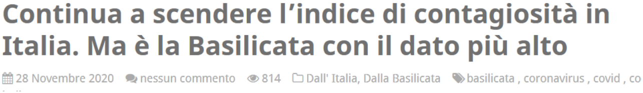
Variants of interest

- The British variant: VOC202012/01 (lineage B.1.1.7)
  - 93.3% ⇒ Molise
  - 75.0% ⇒ Sardegna
  - 0.0% ⇒ Valle d'Aosta
- How many samples with a valid sequencing?
  - 15 ⇒ Molise
  - 12 ⇒ Sardegna
  - 1 ⇒ Valle d'Aosta
- The Brazilian variant "There is a clear geographical characterization"
  - 36.2% ⇒ Umbria
  - 23.8% ⇒ Toscana
  - 13.2% ⇒ Lazio
  - 0.0% ⇒ Abruzzo
- How many samples with a valid sequencing?
  - 47 ⇒ Umbria
  - 80 ⇒ Toscana
  - 144 ⇒ Lazio; 61 ⇒ Abruzzo

Limitations of the survey

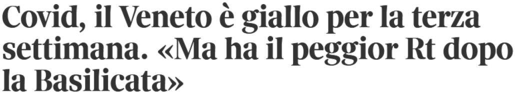
- The sample has been randomly chosen by Regional authorities, guaranteeing **some** geographical representation and **if possibile** a stratification by age.
- The sampling method may vary across Regions.
- For some Regions with low population sizes, the number of valid sequences is too low to detect some variants of interest.
- There are no further info on age stratification, geo-localization and cluster membership of recorded sequences.

The reproduction number R_t on the news



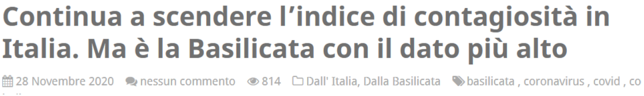
Continua a scendere l’indice di contagiosità in Italia. Ma è la Basilicata con il dato più alto

28 Novembre 2020 nessun commento 814 Dall' Italia, Dalla Basilicata basilicata , coronavirus , covid , co



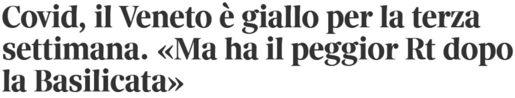
Covid, il Veneto è giallo per la terza settimana. «Ma ha il peggior Rt dopo la Basilicata»

The reproduction number R_t on the news

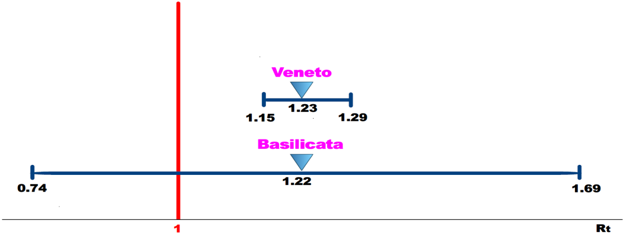


Continua a scendere l’indice di contagiosità in Italia. Ma è la Basilicata con il dato più alto

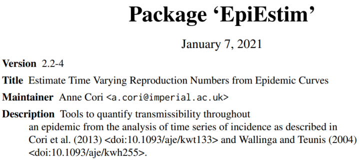
28 Novembre 2020 nessun commento 814 Dall' Italia, Dalla Basilicata basilicata , coronavirus , covid , co



Covid, il Veneto è giallo per la terza settimana. «Ma ha il peggior Rt dopo la Basilicata»



Issues about the R_t estimation



Package 'EpiEstim'

January 7, 2021

Version 2.2-4

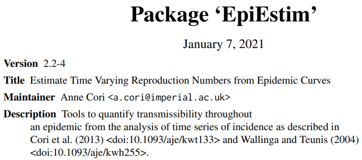
Title Estimate Time Varying Reproduction Numbers from Epidemic Curves

Maintainer Anne Cori <a.cori@imperial.ac.uk>

Description Tools to quantify transmissibility throughout an epidemic from the analysis of time series of incidence as described in Cori et al. (2013) <doi:10.1093/aje/kwt133> and Wallinga and Teunis (2004) <doi:10.1093/aje/kwh255>.

- the time window defined to estimate R_t

Issues about the R_t estimation



Package 'EpiEstim'

January 7, 2021

Version 2.2-4

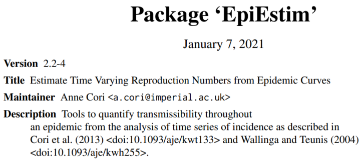
Title Estimate Time Varying Reproduction Numbers from Epidemic Curves

Maintainer Anne Cori <a.cori@imperial.ac.uk>

Description Tools to quantify transmissibility throughout an epidemic from the analysis of time series of incidence as described in Cori et al. (2013) <doi:10.1093/aje/kwt133> and Wallinga and Teunis (2004) <doi:10.1093/aje/kwh255>.

- the time window defined to estimate R_t
    - Estimates of R_t can vary considerably over short time periods, producing substantial negative autocorrelation
    - Small values lead to more rapid detection of changes in transmission but also more statistical noise; large values lead to more smoothing, and reductions in statistical noise

Issues about the R_t estimation



Package 'EpiEstim'

January 7, 2021

Version 2.2-4

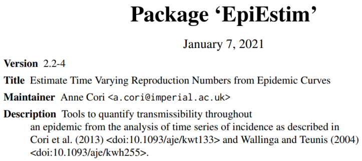
Title Estimate Time Varying Reproduction Numbers from Epidemic Curves

Maintainer Anne Cori `<a.cori@imperial.ac.uk>`

Description Tools to quantify transmissibility throughout an epidemic from the analysis of time series of incidence as described in Cori et al. (2013) <doi:10.1093/aje/kwt133> and Wallinga and Teunis (2004) <doi:10.1093/aje/kwh255>.

- the time window defined to estimate R_t
    - Estimates of R_t can vary considerably over short time periods, producing substantial negative autocorrelation
    - Small values lead to more rapid detection of changes in transmission but also more statistical noise; large values lead to more smoothing, and reductions in statistical noise
- the distributions assumed to model the number of new cases

Issues about the R_t estimation



Package 'EpiEstim'

January 7, 2021

Version 2.2-4

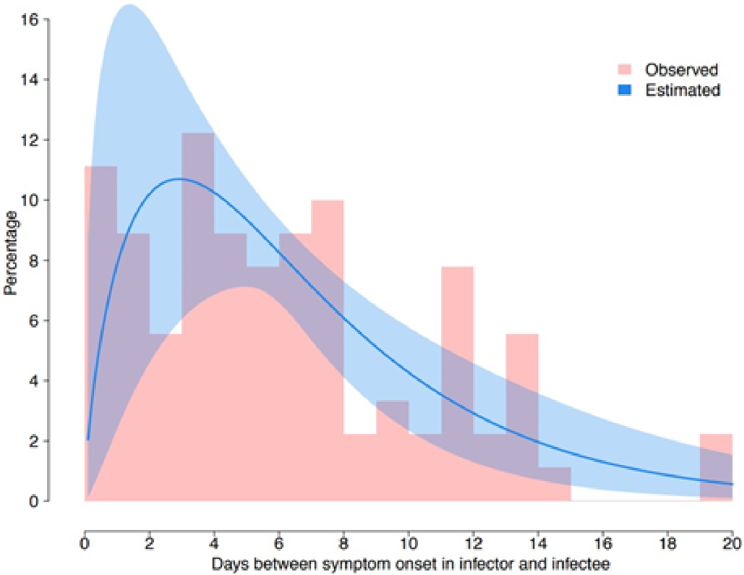
Title Estimate Time Varying Reproduction Numbers from Epidemic Curves

Maintainer Anne Cori `<a.cori@imperial.ac.uk>`

Description Tools to quantify transmissibility throughout an epidemic from the analysis of time series of incidence as described in Cori et al. (2013) <doi:10.1093/aje/kwt133> and Wallinga and Teunis (2004) <doi:10.1093/aje/kwh255>.

- the time window defined to estimate R_t
    - Estimates of R_t can vary considerably over short time periods, producing substantial negative autocorrelation
    - Small values lead to more rapid detection of changes in transmission but also more statistical noise; large values lead to more smoothing, and reductions in statistical noise
- the distributions assumed to model the number of new cases
    - the distribution of infectiousness through time after infection is independent of calendar time and follows a Poisson process
    - It is well-known that Poisson-based estimates are biased if overdispersion arises in the data.
- the generation time

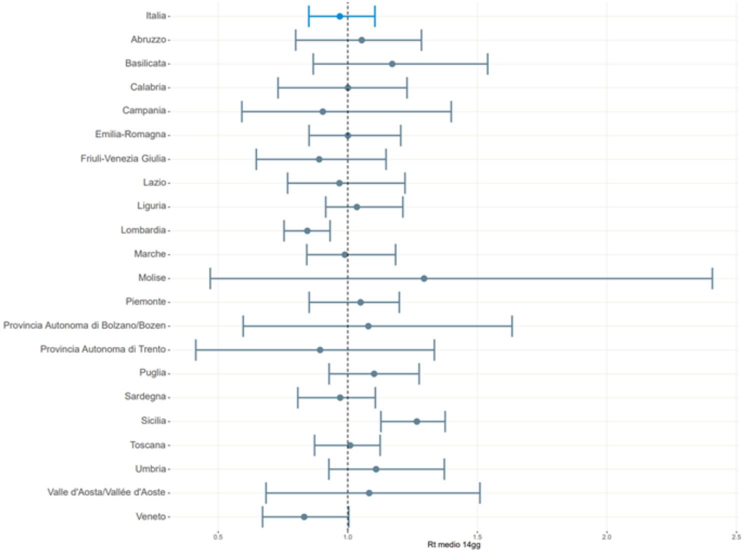
The generation time



The generation time

- It is based on 90 pairs of cases in Lombardy in February, where the authors found an infector–infectee relationship and have the dates of symptom onset of both cases
- This estimate is taken for granted for all the other Italian regions, that is, the same serial interval is assumed for all the regions and never updated.
- A crucial assumption for the adopted model is poorly estimated, wrongly applied to very heterogeneous contexts, and not checked again after the early phase of the first outbreak.
- We are puzzled about it, as the model by Cori et al.2 accepts any parametric or empirical discrete distribution with support on positive values to approximate the serial interval and the generation time, and not only estimated values from a Gamma distribution.
- Gostic et al. illustrate the consequences of misspecifying the form and the variance on the serial interval distribution.

Uncertainty



ICU occupation: Motivation

- Careful and reliable planning of resources can also aid substantially in controlling the consequences of the epidemics, and likely increase the likelihood of early diagnoses and better care.
- To respond to the looming threat of shortage of ICU beds, hospitals urgently need to establish and implement policies that more fairly allocate these scarce resources.
- If hospitals can plan in advance how many ICU beds shall be made available for the nearly following days, capacity can be increased (or decreased) to match the demand. This would avoid the ethical dilemma of severe triaging patients and not admitting those whose lives are *not worth saving*.

Our proposal

- Our approach is based on optimally combining two forecasting methods.
- The first is based on Poisson mixed effects regression
- The second one is a region-specific time series model for counts, taking into account time-dependence over time.
- The count outcome is appropriately modeled as a Poisson conditionally on observed time trends and unobserved heterogeneity including dependence, as implied by random effects or by the auto-regressive structure of the time-series models
- The averaged predictions give an optimal balance between pooling information over different areas (which targets a low variance prediction) and adaptation at the specific area (which targets a low bias prediction).

Random effects modelling of longitudinal count data

We start assuming that the observed daily ICU admissions for region i at day t, y_{it}, are realizations of independent Poisson random variables Y_{it} with parameter μ_{it}, $\forall\, i = 1, \dots, I,\ t = 1, \dots, T$.

$$\log(\mu_{it}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \times t + (\beta_2 + b_{2i}) \times t^2 + \log(residents_i) \quad (1)$$

where a canonical link has been adopted, the offset term $\log(residents_i)$ accounts for different population exposures, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ represents the vector of shared fixed-effects regression parameters, and $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})$ represents the random coefficients, i.e. the region-specific intercept and slopes, with

$$\mathbf{b}_i \sim N(\mathbf{0}, \Sigma_B)$$

Random effects modelling of longitudinal count data

- Predictions are based on the posterior estimates of the random effects and the maximum-likelihood estimate (MLE) of the fixed-effect parameters.
- Predictions intervals are found through non-parametric block bootstrap using 500 replicates. Block bootstrap involves resampling regions, and once a region is included its entire time-series is used for model estimation of the resampled data.
- The best covariance structure, which has been then used for all estimates and predictions, has turned out to be:

$$B = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & 0 \\ \sigma_{01} & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix} \tag{2}$$

Region-specific integer-valued autoregression modelling

At the second step, we fit and obtain predictions for regional time series separately. In other words, 20 different models are fitted, as time-series models for counts.

We model Y_{it} as a conditional Poisson distribution where the expectation μ_{it} at time t depends on both past counts and past covariates:

$$\mu_{it} = \alpha_0 \mu_{it-1} + \alpha_1 y_{it-1} + \boldsymbol{\gamma}^T \mathbf{x}_{it-1}, \quad t > 1. \tag{3}$$

where the coefficients α_0 and α_1 represent the effects of the expectation μ_{it-1} of the previous day and the number of ICU admissions of the previous day y_{it-1}, respectively.

Region-specific integer-valued autoregression modelling

- For each region, we compare stationary, linear, quadratic and cubic trends.
- We select the best model specification for each one separately, according to the Bayesian information criterion.
- Parameters in equation are estimated via conditional maximum quasi-likelihood estimation, using the function `tsglm` in the **tscount** `R` package.
- Prediction intervals are approximated numerically through a parametric bootstrap procedure: parameter estimates are plugged in, and several random draws are made from Poisson distributions with the resulting parameter. The approximated prediction intervals are obtained from the empirical 2.5% and 97.5% quantiles of the boostrap-based predictions.

Model averaging

- The final prediction is

$$\hat{y}_{iT+1} = w_{iT+1}\hat{y}_{iT+1}^{(GLMM)} + (1 - w_{iT+1})\hat{y}_{iT+1}^{(INAR)}, \tag{4}$$

  for some $w_{iT+1} \in (0, 1)$.
- One could simply fix $w_{iT+1} = 0.5$, but this would not lead to any optimality properties of the resulting final prediction $\hat{y}_{iT+1}$.
- We thus first repeat model estimation excluding y_{iT} for $i = 1, \ldots, I$; obtaining leave-last-out predictions $\hat{y}_{iT}^{(GLMM)}$ and $\hat{y}_{iT}^{(INAR)}$; and then we solve the optimization problem

$$w_{iT+1} = \arg \inf_{x \in (0,1)} \left| x\hat{y}_{iT}^{(GLMM)} + (1 - x)\hat{y}_{iT}^{(INAR)} - y_{iT} \right|.$$

- The rationale is that of selecting the weight that minimizes, for each region, the absolute difference between the final prediction at time T (when temporarily ignoring y_{iT}), and the actually observed count at time T.

Results

The reliability and goodness of our approach can be assessed by checking the next-day performance as:

- median absolute error over the twenty Italian regions,
- mean relative error over the twenty Italian regions,
- proportion of prediction intervals that do not contain the actually observed occupancy,
- proportion of observed occupacies above the upper limit of the prediction interval.

To summarize:

- The daily absolute error has a median of 4 beds, with first quartile 1 and third quartile 8.
- The daily relative error over the twenty regions has first quartile 2%, median 5%, third quartile 12%. Its mean is 9.2%.
- For prediction intervals we used a nominal level of 99%. Out of the 840 intervals produced, 99.4% indeed contained the observed ICU occupation.

Focus on...

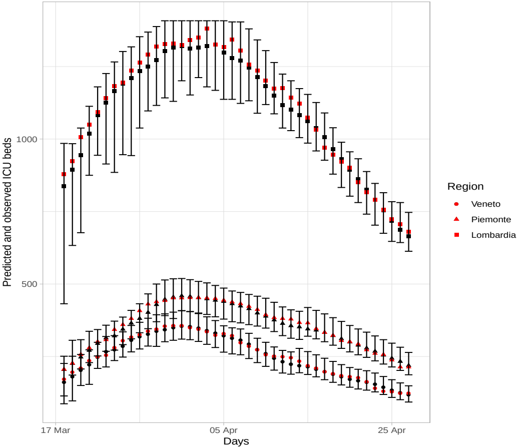
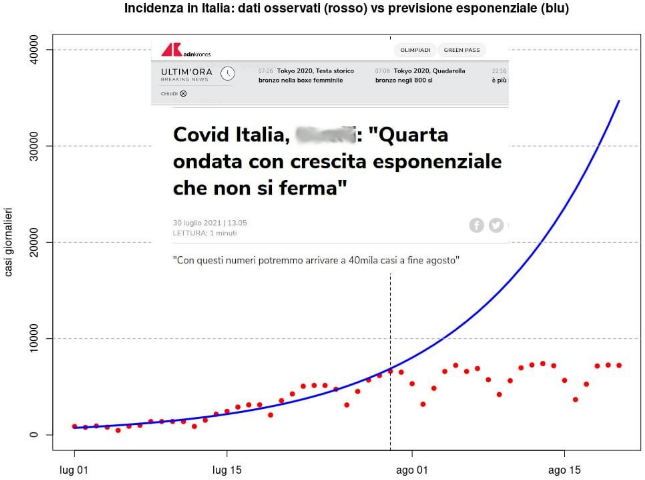


Figure: Observed (black) and predicted (gray) values with 99% Prediction Intervals for the three northern regions Lombardia, Piemonte and Veneto.

Unreliable predictions about COVID-19 infections and hospitalizations make people worry.

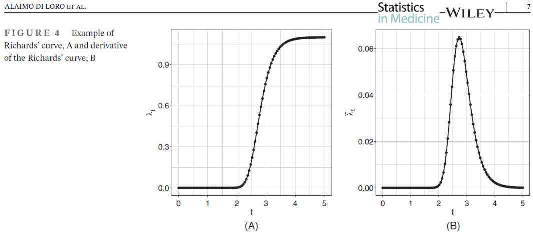


Covid Italia, [redacted]: "Quarta ondata con crescita esponenziale che non si ferma"

30 luglio 2021 | 13:05
LETTURA: 1 minuti

"Con questi numeri potremmo arrivare a 40mila casi a fine agosto"

Richards' curve



ALAIMO DI LORO ET AL.

Statistics in Medicine – WILEY 7

FIGURE 4 Example of Richards' curve, A and derivative of the Richards' curve, B

$$\mathbb{E}\left[Y_t^c\right] = \lambda_\gamma(t) = b + \frac{r}{\left[1+10^{h(p-t)}\right]^s}, \qquad \gamma^T = [b, r, h, p, s]$$

- $b \in \mathbb{R}^+$ → lower asymptote
- $r > 0$ → distance between the upper and the lower asymptote
- $h \in \mathbb{R}$ → represents the infection/growth rate (*hill*)
- $p \in \mathbb{R}$ → tells when the curve growth speed slows down
- $s \in \mathbb{R}$ → asymmetry parameter regulating differences in the behavior of the ascending and descending phase of the outbreak

Modeling key-points

- Mantain the **Richard's growth** behavior of the **cumulative counts**

$$\mathbb{E}\left[\, Y_t \right] = \mathbb{E}[\, Y_t^c] - \mathbb{E}[\, Y_{t-1}^c] = \lambda_\gamma(t) - \lambda_\gamma(t-1) = r \cdot \widetilde{\lambda}_\gamma(t),$$

where:

$$\widetilde{\lambda}_\gamma(t) = \left(\left[1 + 10^{h(p-t)} \right]^{-s} - \left[1 + 10^{h(p-(t-1))} \right]^{-s} \right)$$

Add a **baseline** (*endemic rate*)

$$\mathbb{E}\left[\, Y_t \right] = \mu_\gamma(t) = \alpha + r \cdot \widetilde{\lambda}_\gamma(t), \quad \alpha > 0$$

Modeling key-points

- Consider the effect of **covariates** through a link function

$$\eta_{\boldsymbol{\beta}}(\mathbf{X}) = \boldsymbol{\beta}\mathbf{X} \quad \Rightarrow \quad \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{X}) = \exp\left\{\eta_{\boldsymbol{\beta}}(\mathbf{X})\right\}$$

  - **Additive**:

$$\mu_{\boldsymbol{\theta}}(t,\mathbf{X}) = \alpha_{\boldsymbol{\beta}}(\mathbf{X}) + r \cdot \widetilde{\lambda}_{\gamma}(t), \quad \alpha_{\boldsymbol{\beta}}(\mathbf{X}) = g_{\boldsymbol{\beta}}(\mathbf{X})$$

  - **Multiplicative**:

$$\mu_{\boldsymbol{\theta}}(t,\mathbf{X}) = \alpha + r_{\boldsymbol{\beta}}(\mathbf{X}) \cdot \widetilde{\lambda}_{\gamma}(t), \quad r_{\boldsymbol{\beta}}(\mathbf{X}) = g_{\boldsymbol{\beta}}(\mathbf{X})$$

Modeling key-points

- Behold to the **discrete** nature of counts

Poisson

$$Y_t \sim Pois(\mu_{\boldsymbol{\theta}}(t)) \;\rightarrow\; f(Y_t|\boldsymbol{\theta}) = \frac{e^{\mu_{\boldsymbol{\theta}}(t)}}{y_t!}\mu_{\boldsymbol{\theta}}(t)^{y_t}$$

Negative Binomial

$$Y_t \sim NegBin(\mu_{\boldsymbol{\theta}}(t), \nu) \;\rightarrow\; f(Y_t|\boldsymbol{\theta}) =$$

$$= \frac{\Gamma(\nu + y_t)}{\Gamma(\nu)} \cdot \left(\frac{\nu}{\nu + \mu_{\boldsymbol{\theta}}(t)}\right)^{\nu} \left(\frac{\mu_{\boldsymbol{\theta}}(t)}{\nu + \mu_{\boldsymbol{\theta}}(t)}\right)^{y_t}$$

Input data

- Comparison of three different specifications of $\mathbf{W}$:
  - $\mathbf{W}_{\mathbf{Ind}} = \mathbf{0}$ → spatial independence
  - $\mathbf{W}_{\mathbf{Flow}}$ based on proximity flows (direct HV trains, flights, ferries)
  - $\mathbf{W}_{\mathbf{Geo}}$ based on regions' mutual geographical position

Input data

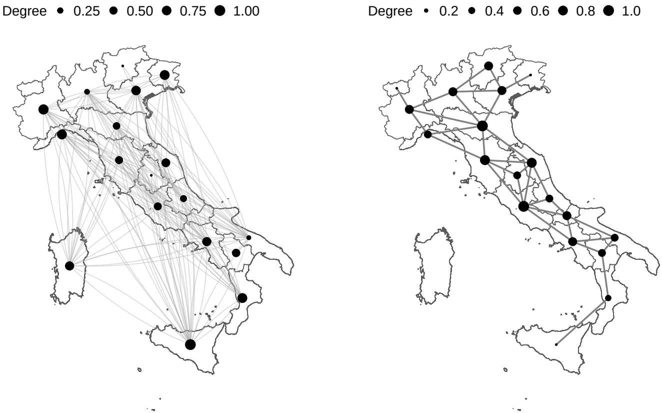

Degree ● 0.25 ● 0.50 ● 0.75 ● 1.00
Degree ● 0.2 ● 0.4 ● 0.6 ● 0.8 ● 1.0

Figure: Network structure: $\mathbf{W}_{\textbf{Flow}}$ (left) and $\mathbf{W}_{\textbf{Geo}}$ (right).

Model selection and validation

Wave	**Metric**	M_{Ind}	M_{Flow}	M_{Geo}
I	Coverage	0.98	0.98	0.98
	PIW	1535	1178	1144
	RMSE	423	**184**	**272**
	WAIC	2869	**2650**	**2774**
	LOO	3087	**2849**	**2982**
II	Coverage	0.96	0.97	0.92
	PIW	33393	4497	4046
	RMSE	12841	**910**	**995**
	WAIC	4112	**3820**	**3971**
	LOO	4393	**4080**	**4252**

Table: Out-of-sample predictive performances for the first and the second wave.

Model results - Fitting

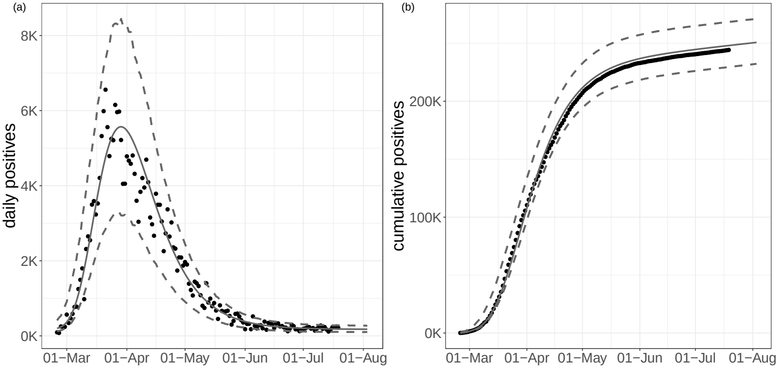


Figure: Model fitting - Daily positives - Negative Binomial.

Model results - Fitting

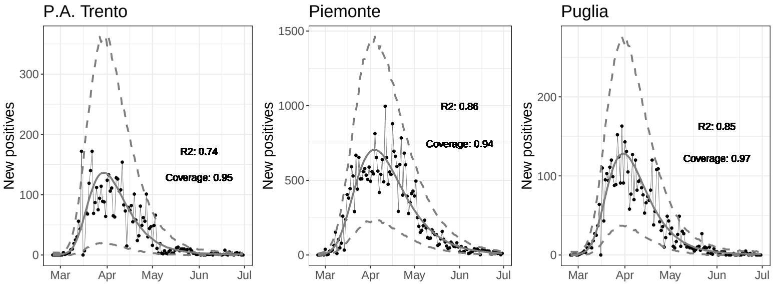


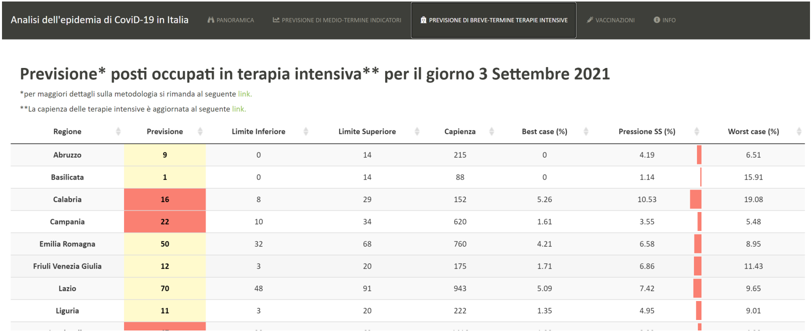
Figure: Model fitting - Daily positives - Negative Binomial.

Shiny APP

Just to corroborate the results, we provide an example of the forecasts shown daily at

$$\mathtt{https://statgroup19.shinyapps.io/StatGroup19-Eng.}$$

$$\mathtt{https://statgroup19.shinyapps.io/Covid19App/}$$


Analisi dell'epidemia di CoviD-19 in Italia
PANORAMICA
PREVISIONE DI MEDIO-TERMINE INDICATORI
PREVISIONE DI BREVE-TERMINE TERAPIE INTENSIVE
VACCINAZIONI
INFO

Previsione* posti occupati in terapia intensiva** per il giorno 3 Settembre 2021

*per maggiori dettagli sulla metodologia si rimanda al seguente link.

**La capienza delle terapie intensive è aggiornata al seguente link.

Regione	Previsione	Limite Inferiore	Limite Superiore	Capienza	Best case (%)	Pressione SS (%)	Worst case (%)
Abruzzo	9	0	14	215	0	4.19	6.51
Basilicata	1	0	14	88	0	1.14	15.91
Calabria	16	8	29	152	5.26	10.53	19.08
Campania	22	10	34	620	1.61	3.55	5.48
Emilia Romagna	50	32	68	760	4.21	6.58	8.95
Friuli Venezia Giulia	12	3	20	175	1.71	6.86	11.43
Lazio	70	48	91	943	5.09	7.42	9.65
Liguria	11	3	20	222	1.35	4.95	9.01

Other examples

The role of vitamin D in the prevention of coronavirus disease 2019 infection and mortality by Ilie et al. (2020)
with more than 200 citations!!!

“The crude association observed in the present study may be explained by the role of vitamin D in the prevention of COVID-19 infection or more probably by a potential protection of vitamin D from the more negative consequences of the infection.”

- Nominal p-values are greater than 0.05, though those reported are exactly 0.05.
- Any parametric test is based on some assumptions. In this case, they are not met.
- The linear regression model predicts -16 deaths per million people in Slovakia (resurrection!)

Remarks

- Wrong assumptions
  - one model fits all
  - the model is correct; the data are wrong
  - the fact that you apply statistics does not mean that you are a statistician or a data analyst
- The data are telling us a story. Are we good enough to read it?

References

- P. Alaimo Di Loro, F. Divino, A. Farcomeni, G. Jona-Lasinio, G. Lovison, A. Maruotti, M. Mingione (2021). Nowcasting COVID-19 incidence indicators during the Italian first outbreak. *Statistics in Medicine*, 40: 3843-3864.
- D. Böhning, I. Rocchetti, A. Maruotti, H. Holling (2020). Estimating the undetected infections in the Covid-19 outbreak by harnessing capture-recapture methods. *International Journal of Infectious Disease*, 97: 197–201.
- F. Divino, M. Ciccozzi, A. Farcomeni, G. Jona-Lasinio, G. Lovison, A. Maruotti (2021). Unreliable predictions about COVID-19 infections and hospitalizations make people worry: the case of Italy. *Journal of Medical Virology*, to appear.
- A. Farcomeni, A. Maruotti, F. Divino, G. Jona-Lasinio and G. Lovison (2021). An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biometrical Journal*, 63: 503–513.
- A. Maruotti, M. Ciccozzi, F. Divino (2021). On the misuse of the reproduction number in the COVID-19 surveillance system in Italy. *Journal of Medical Virology*, 93: 2569–2570.
- M. Mingione, P. Alaimo Di Loro, A. Farcomeni, F. Divino, G. Lovison, A. Maruotti, G. Jona-Lasinio (2021). Spatio-temporal modelling of COVID-19 incident cases using Richards' curve: an application to the Italian regions. *Spatial Statistics*, to appear.