# A framework for forecasting COVID-19 outbreak risk at subnational level in EU countries

Alexis Robert, Lloyd AC Chapman, Sebastian Funk, Adam J Kucharski
Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK, project done for the European Center of Disease Prevention and Control, based on contract ref REOP/2021/SMS/13060.

## Summary

This document presents an overview of the model we developed to forecast subnational COVID-19 outbreak risk in different European countries (so far applied to France, Czechia, and Italy). It contains a summary of the methods and data sources used to generate the case forecasts for each country.

## Methods and data sources

### Case forecasts

The case forecasts and simulations were generated using an age-stratified Endemic-Epidemic model (1–3). This framework uses a frequentist approach, and identifies factors (i.e. covariates) associated with the risks of secondary transmission and importations. Since it does not require a mechanistic implementation, the link between covariates and outcome is flexible, and adapted to changing reporting patterns and data quality. This facilitates the inclusion of different data sources and improves the robustness of the model, for instance, thanks to this flexibility, the model is able to take into account changes in the proportion of cases reported through the addition of covariates quantifying reporting capacities.

The age-stratified Endemic-Epidemic model was implemented using the R packages *surveillance* (1), *hhh4contacts* (3), and *hhh4addon* (4), which allowed us to implement distributed lags across previous time steps. The number of cases at each time step therefore did not exclusively depend on the previous time step and could follow the typical serial interval of the pathogen (centred around 5 days). The model has two sources of new cases: the epidemic component, containing within-region and cross-regional transmission, and the endemic component (i.e. importations). The risk of transmission between regions was assumed to decrease with distance according to a power law. The distance was defined as the degree of connectivity between regions, meaning that neighbours (defined as adjacent regions) had a distance of 1 (neighbours of neighbours have a distance of 2, etc…). This model proved more robust than an exponential gravity model (based on the Euclidean distance between regions) when the number of regions in the model was low.

We fitted the model to case data reported between September 2020 and the latest reported date (currently February 2023). After fitting the model, we used the parameter estimates, along with the last reported value of vaccine coverage and testing, to generate 4-week ahead forecasts. Furthermore, we simulated the impact of various targeted non-pharmaceutical interventions (NPIs), and the consequences of increase in transmission on the number of cases forecasted by the model.

Several data sources were used to implement the model:

- Since reports to centralised databases were interrupted in early 2022 in many countries, the source we used for the case data varied between countries. Similarly, vaccination and test datasets were imported from different sources (summarised in Table 1).

**Table 1: Case, death, vaccination, test and population data source used in each country.**

| Country | Metropolitan France | Czechia | Italy |
|---|---|---|---|
| Case data source | Santé Publique France (**daily age-stratified** number of cases in each **NUTS-3 region** (5)) | Ministerstvo Zdravotnictví České Republiky (**daily age-stratified** number of cases in each **NUTS-3 region** (6)**)** | Dipartimento Della Protezione Civile (**daily non-age-stratified** number of cases reported per NUTS-3 region (7)) |
| Vaccination data source | l'Assurance Maladie (weekly age-stratified number of doses in **each NUTS-3 region** (8)) | Ministerstvo Zdravotnictví České Republiky (weekly age-stratified number of doses in each **NUTS-3 region** (6)) | ECDC vaccination database (overall number of doses delivered weekly in each **NUTS-2 region** (9)) |
| Test data source | Santé Publique France (daily age-stratified number of tests (5)) | Ministerstvo Zdravotnictví České Republiky (daily age-stratified number of tests (6)) | Dipartimento Della Protezione Civile (weekly national number of tests (7)) |
| Population data source | INSEE (Number of inhabitants per NUTS 3 region and age group (10)) | Eurostat (Number of inhabitants per NUTS 3 region and age group (11)) | Google COVID-19 Open Data (Number of inhabitants per NUTS 3 region (12)) |
| Death data source | Santé Publique France (**daily age-stratified** number of deaths in | Ministerstvo Zdravotnictví České Republiky (**daily age-stratified** number of deaths in | Dipartimento Della Protezione Civile (**daily non-age-stratified** number of deaths |

| | | | |
|---|---|---|---|
| | each **NUTS-2 region** (5)) | each **NUTS-3 region** (6)**)** | reported per NUTS-2 region (7)) |

- The urban-rural status of each NUTS-3 region was taken from the Eurostat database (database labelled “urban-rural remoteness”) (13).
- The number of inhabitants per age group and region (see Table 1).

In addition, we integrated an age-specific intercept in both the neighbourhood and endemic components, and a day-of-the-week effect (considering the reporting pattern may change every day), and the impact of the Delta and Omicron variants in the neighbourhood component (14). The models also integrated two seasonality covariates, accounting for background changes in transmission during the year.
The differences in data availability led to minor differences between the equations of the model in each country.

## France

We implemented an age-stratified Endemic-Epidemic model based on Equation (1):

$$\begin{aligned} log(\phi_{ait}) = {} & \alpha^{(\phi)} + \alpha_a^{(\phi)} + \beta_{pop\_age}^{(\phi)} log(pop\_age_{ai}) + \beta_{pop\_tot}^{(\phi)} log(pop_i) \\ & + \beta_{tue} tue_t + \beta_{wed} wed + \beta_{thu} thu_t + \beta_{fri} fri_t + \beta_{sat} sat_t + \beta_{sun} sun_t \\ & + \beta_{test\_prop} log(test\_prop_t) + \beta_{test\_age} log(test\_age_{at}) \\ & + \beta_{rural}^{(\phi)} rural_i + \beta_{int\_rur}^{(\phi)} int\_rur_i + \beta_{int\_urb}^{(\phi)} int\_urb_i \\ & + \beta_{cov}^{(\phi)} log(1 - cov_{ait}) + \beta_{inc\_old}^{(\phi)} log(1 - inc\_old_{ait}) + \beta_{inc\_new}^{(\phi)} log(1 - inc\_new_{ait}) \\ & + \beta_{delta} delta_t + \beta_{omicron} omicron_t \\ & + \beta_{sin}^{(\phi)} sin(2 * \pi * t/365) + \beta_{cos}^{(\phi)} cos(2 * \pi * t/365) \end{aligned} \quad (1)$$

$$\begin{aligned} log(\nu_{ait}) = {} & \alpha^{(\nu)} + \alpha_a^{(\nu)} + \beta_{pop}^{(\nu)} log(pop_i * pop\_age_{ai}) \\ & + \beta_{rural}^{(\nu)} rural_i + \beta_{int\_rur}^{(\nu)} int\_rur_i + \beta_{int\_urb}^{(\nu)} int\_urb_i \\ & + \beta_{cases\_eur} log(cases\_eur_t) \\ & + \beta_{sin}^{(\nu)} sin(2 * \pi * t/365) + \beta_{cos}^{(\nu)} cos(2 * \pi * t/365). \end{aligned}$$

Here, $\alpha^{(\phi)}$ and $\alpha^{(\nu)}$ are fixed intercepts for the epidemic and endemic components with age adjustments $\alpha_a^{(\phi)}$ and $\alpha_a^{(\nu)}$. All covariates are defined in Table A1 in the Appendix. The cumulative incidence of cases by age and region between 1 and 12 months before day $t$, $inc\_old_{ait}$, and in the last month, $inc\_new_{ait}$, were introduced to allow the model to better capture short- and medium-term changes in immunity due to differences in the rate of spread of different variants (e.g. Delta and Omicron). The contact matrix between age groups was taken from Béraud et al (15).

## Czechia

$$log(\phi_{ait}) = \alpha^{(\phi)} + \alpha_a^{(\phi)} + \beta_{pop\_age}^{(\phi)} log(pop\_age_{ai}) + \beta_{pop\_tot}^{(\phi)} log(pop_i) \quad (2)$$

$$
\begin{aligned}
&+ \beta_{tue} tue_t + \beta_{wed} wed + \beta_{thu} thu_t + \beta_{fri} fri_t + \beta_{sat} sat_t + \beta_{sun} sun_t \\
&+ \beta_{test\_prop} log(test\_prop_t) + \beta_{test\_age} log(test\_age_{at}) \\
&+ \beta^{(\phi)}_{int\_rur} int\_rur_i + \beta^{(\phi)}_{int\_urb} int\_urb_i \\
&+ \beta^{(\phi)}_{cov} log(1 - cov_{ait}) + \beta^{(\phi)}_{inc\_old} log(1 - inc\_old_{ait}) + \beta^{(\phi)}_{inc\_new} log(1 - inc\_new_{ait}) \\
&+ \beta_{delta} delta_t + \beta_{omicron} omicron_t \\
&+ \beta^{(\phi)}_{sin} sin(2 * \pi * t/365) + \beta^{(\phi)}_{cos} cos(2 * \pi * t/365)
\end{aligned}
$$

$$
\begin{aligned}
log(\nu_{ait}) = \alpha^{(\nu)} + \alpha^{(\nu)}_{a} + \beta^{(\nu)}_{pop} log(pop_i * pop_{ai}) \\
+ \beta^{(\nu)}_{int\_rur} int\_rur_i + \beta^{(\nu)}_{int\_urb} int\_urb_i \\
+ \beta_{cases\_eur} log(cases\_eur_t) \\
+ \beta^{(\nu)}_{sin} sin(2 * \pi * t/365) + \beta^{(\nu)}_{cos} cos(2 * \pi * t/365).
\end{aligned}
$$

In comparison to France, since there was no NUTS-3 region in Czechia classified as “rural”, this level was dropped from the equation. The contact matrix was extracted from Prem et al (16).

## Italy

$$
\begin{aligned}
log(\phi_{ait}) = \alpha^{(\phi)} + \beta^{(\phi)}_{pop\_tot} log(pop_i) \\
+ \beta_{tue} tue_t + \beta_{wed} wed + \beta_{thu} thu_t + \beta_{fri} fri_t + \beta_{sat} sat_t + \beta_{sun} sun_t \\
+ \beta_{test\_prop} log(test\_prop_t) \\
+ \beta^{(\phi)}_{rural} rural_i + \beta^{(\phi)}_{int\_rur} int\_rur_i + \beta^{(\phi)}_{int\_urb} int\_urb_i \\
+ \beta^{(\phi)}_{cov} log(1 - cov_{it}) + \beta^{(\phi)}_{inc\_old} log(1 - inc\_old_{it}) + \beta^{(\phi)}_{inc\_new} log(1 - inc\_new_{it}) \\
+ \beta_{delta} delta_t + \beta_{omicron} omicron_t \\
+ \beta^{(\phi)}_{sin} sin(2 * \pi * t/365) + \beta^{(\phi)}_{cos} cos(2 * \pi * t/365)
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
log(\nu_{ait}) = \alpha^{(\nu)} + \beta^{(\nu)}_{pop} log(pop_i) \\
+ \beta^{(\nu)}_{rural} rural_i + \beta^{(\nu)}_{int\_rur} int\_rur_i + \beta^{(\nu)}_{int\_urb} int\_urb_i \\
+ \beta_{cases\_eur} log(cases\_eur_t) \\
+ \beta^{(\nu)}_{sin} sin(2 * \pi * t/365) + \beta^{(\nu)}_{cos} cos(2 * \pi * t/365).
\end{aligned}
$$

Since age-stratified data was not available in Italy, the model described in Equation (3) does not include an age-specific intercept, nor the covariates $test\_age$ and $pop\_age_{ai}$ (Appendix 1). The covariate $test\_prop$ now describes the proportion of the overall national population tested in the past 14 days.

## Death forecasts

We used simple linear regression models to generate one, two, three, and four-week ahead forecasts of the weekly number of reported deaths.
We first aggregated the number of reported cases by week, and computed the Case Fatality Ratio (CFR), considering a three week delay between weekly reported cases and reported deaths:

$$CFR_{ait} = N_{ait-3}/D_{ait}$$

(with $N$ the weekly number of cases, $D$ the weekly number of deaths, $t$ the current week, $i$ the region and $a$ the age group).
We then calculated $\Delta CFR_{ait}$, the change in CFR, and $\Delta N_{ait}$, the change in the number of cases:

$$\Delta CFR_{ait} = CFR_{ait} - CFR_{ait-x} \text{ and } \Delta N_{ait} = N_{ait} - N_{ait-x}$$

With $x$ the forecast horizon (one, two, or three weeks). We implemented a linear regression model for each forecast horizon and age group, with $\Delta CFR$ as outcome, and $\Delta N$ as explanatory variable ($\Delta CFR_{ait} = \alpha + \beta * \Delta N_{ait}$ ). We then used the estimates of $\alpha$ and $\beta$ to predict the changes in Case fatality ratio between the last week of reported case data and the forecast dates. We predict future CFR from predicted values of $\Delta CFR_{ait}$ , and use the number of cases reported in the last week of data $N_{tpred}$ to generate one, two, three, and four week-ahead forecasts of the number of deaths. Given that many factors can cause changes in CFR (vaccination, previous immunity, variant…), we only take observations from January 2022 when fitting the regression models. Since we considered a three-week shift between cases and deaths, the first three weeks of death forecasts are generated using case data and the different linear regression models. The last week of death forecasts is computed from the number of cases simulated the week after the prediction date, and parameter estimates from the three-week ahead regression models.

# References


1. Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Stat Model. 2005 Oct;5(3):187–99.
2. Meyer S, Held L, Höhle M. hhh4: Endemic-epidemic modeling of areal count time series. :23.
3. Meyer S, Held L. Incorporating social contact data in spatio-temporal models for infectious disease spread. Biostatistics. 2016 Dec 26;kxw051.
4. Bracher J, Held L. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. Int J Forecast. 2022 Jul;38(3):1221–33.
5. Santé publique France. Données de laboratoires pour le dépistage [Internet]. [cited 2022 Jul 26]. Available from: https://www.data.gouv.fr/fr/datasets/donnees-de-laboratoires-pour-le-depistage-a-compter-du-18-05-2022-si-dep/
6. Ministerstvo Zdravotnictví České republiky. COVID-19 in the Czech Republic: Open data sets and downloadable sets [Internet]. [cited 2022 Jul 26]. Available from: https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19
7. Dipartimento della Protezione Civile. Dati COVID-19 Italia [Internet]. [cited 2022 Jul 26]. Available from: https://github.com/pcm-dpc/COVID-19
8. datavaccin-covid. Données vaccination par tranche d'âge, type de vaccin et

département / région [Internet]. [cited 2022 Jul 26]. Available from: https://datavaccin-covid.ameli.fr/explore/dataset/donnees-vaccination-par-tranche-dage-type-de-vaccin-et-departement/information/
9. European Centre for Disease Prevention and Control. Data on COVID-19 vaccination in the EU/EEA [Internet]. [cited 2022 Jul 26]. Available from: https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea
10. INSEE. Estimation de la population au 1er janvier 2022 [Internet]. [cited 2022 Jul 26]. Available from: https://www.insee.fr/fr/statistiques/1893198
11. Eurostat. Population on 1 January by age group, sex and NUTS 3 region [Internet]. [cited 2023 Jan 2]. Available from: https://ec.europa.eu/eurostat/web/products-datasets/product?code=demo_r_pjangrp3
12. O. Wahltinez and others. COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2. 2020; Available from: https://goo.gle/covid-19-open-data
13. Eurostat. Methodology - Rural development [Internet]. [cited 2022 Jul 26]. Available from: https://ec.europa.eu/eurostat/web/rural-development/methodology
14. European Centre for Disease Prevention and Control. Data on SARS-CoV-2 variants in the EU/EEA [Internet]. [cited 2022 Jul 26]. Available from: https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea
15. Béraud G, Kazmercziak S, Beutels P, Levy-Bruhl D, Lenne X, Mielcarek N, et al. The French Connection: The First Large Population-Based Contact Survey in France Relevant for the Spread of Infectious Diseases. Chuang JH, editor. PLOS ONE. 2015 Jul 15;10(7):e0133203.
16. Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. Halloran B, editor. PLOS Comput Biol. 2017 Sep 12;13(9):e1005697.
17. World Health Organization. Daily cases and deaths by date reported to WHO [Internet]. 2023 [cited 2023 Jan 30]. Available from: https://covid19.who.int/WHO-COVID-19-global-data.csv

# Appendix

**Table A1: Definitions and sources of covariates included in the model**

| Variable | Definition | Source |
|---|---|---|
| $pop_i$ | Total population of region $i$ (in 100,000s) | Country dependent, see Table 1 |
| $pop\_age_{ai}$ | Proportion of population of region $i$ that is in age group $a$ | Country dependent, see Table 1 |
| $tue_t$, $wed_t$, $thu_t$, $fri_t$, $sat_t$,$sun_t$ | Binary indicator variable controlling for the day of the week effect | n/a |
| $test\_prop_t$ | Proportion of the population tested in the last 14 days nationally | Country dependent, see Table 1 |

| | | |
|---|---|---|
| $test\_age_{at}$ | Proportion of the population in age group $a$ tested in the last 14 days | Country dependent, see Table 1 |
| $rural_i$ | Binary indicator variable for whether region $i$ is rural | (13) |
| $int\_rural_i$ | Binary indicator variable for whether region $i$ is predominantly rural | (13) |
| $int\_urb_i$ | Binary indicator variable for whether region $i$ is predominantly urban | (13) |
| $cov_{ait}$ | where $cov_{ait}$ is the proportion of the population of age group $a$ in region $i$ who have received a vaccine dose in the 120 days before day $t$ | Country dependent, see Table 1 |
| $inc\_old_{ait}$ | cumulative incidence of cases in age group $a$ in region $i$ from between six months ago and 1 month ago (from day $t - 365$ up to day $t - 30$) | Country dependent, see Table 1 |
| $inc\_new_{ait}$ | cumulative incidence of cases in age group $a$ in region $i$ in the last month (from day $t - 30$ up to day $t - 1$) | Country dependent, see Table 1 |
| $delta_t$ | Binary indicator variable for whether day $t$ was in a period when the proportion of sequenced cases that were Delta was higher than 30% | (14) |
| $omicron_t$ | Binary indicator variable for whether day $t$ was in a period when the proportion of sequenced cases that were Omicron was higher than 30% | (14) |
| $cases\_eur_t$ | Number of cases in the rest of Europe in the past 30 days on day $t$ (in 100,000s) | (17) |