# A multivariate statistical approach to predict COVID-19 count data with epidemiological interpretation and uncertainty quantification

*Fulvia Pennoni*

*Department of Statistics and Quantitative Methods*
*University of Milano-Bicocca*
Email: fulvia.pennoni@unimib.it

joint work with *F. Bartolucci**, and *A. Mira***

**University of Perugia*, ***Università della Svizzera italiana and University of Insubria*

# Outline

- Introduction
- Proposed Dirichlet-Multinomial distribution and parameterizations
- Parameter estimation, model checking and comparison
- Application: Italian data
- Conclusions

# Introduction

- We propose Multinomial and Dirichlet-Multinomial statistical autoregressive models to jointly analyse the observed time series of COVID-19 count data

- The *official data* may present biases due to the observational nature and the delays of the collection process

- Categories include: *susceptible* not previously ill (S); *recovered* (R); positive cases in *quarantine* (Q); *hospitalized* in regular wards (H) and in *intensive care units* (ICU), together with *deceased* (D)

- The models consider that count for a every category at a certain time occasion is the sum of unobservable transitions from the same and other categories that these individuals occupied at the previous time occasion

# Introduction

- The approach is related to the SEIR (Susceptible-Exposed-Infected-Recovered) epidemiological model
- It allows us to incorporate expert prior information
- It accounts for public health non-pharmaceutical interventions (NPI) enforced to reduce the spread of the epidemic
- We estimate the persistence in each category,
- We estimate the sequence of contingency tables of the transition frequencies between two consecutive time occasions
- We include absorbing states, as that of deceased patients

# Model assumptions

- We denote by:
  - $y_{tk}$ counts over time occasions and categories $t \in \mathcal{T}$, $k \in \mathcal{K}$, where $\mathcal{T} = \{1, \ldots, T\}$ and $\mathcal{K} = \{1, \ldots, K\}$
  - $K = 6$ ordered (in terms of their severity) categories
  - $Y_{tk} = \sum_{j \in \mathcal{K}} X_{tjk}$ observed column totals at occasion $t$ of category $k$ and $Y_{t-1,j} = \sum_{k \in \mathcal{K}} X_{tjk}$ row totals
  - $X_{tjk}$ are unobserved random variables are collected in the vectors $\boldsymbol{X}_{tj} = (X_{tj1}, \ldots, X_{tjK})'$, $j \in \mathcal{K}$, $t \in \mathcal{T}'$ representing "transition frequencies"
- We assume that

$$\boldsymbol{X}_{tj} | \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \boldsymbol{\beta}_j \sim \mathrm{Mult}(y_{t-1,j}; \boldsymbol{p}_{tj}),$$

where $\boldsymbol{\beta}_j$ is the matrix of the regression parameters involved for the conditional probabilities in $\boldsymbol{p}_{tj}$

# Data structure

- $Y_{tk}$ denotes the observed count among the $K = 6$ categories for the COVID-19 application
- $X_{tjk}$ denotes the unobserved counts of number of transitions from category $j$ to category $k$ at time $t$

| | S | R | Q | H | ICU | D | Total |
|---|---|---|---|---|---|---|---|
| S | $X_{t11}$ | $X_{t12}$ | $X_{t13}$ | $X_{t14}$ | $X_{t15}$ | $X_{t16}$ | $Y_{t-1,1}$ |
| R | 0 | $X_{t22}$ | $X_{t23}$ | $X_{t24}$ | $X_{t25}$ | $X_{t26}$ | $Y_{t-1,2}$ |
| Q | 0 | $X_{t32}$ | $X_{t33}$ | $X_{t34}$ | $X_{t35}$ | $X_{t36}$ | $Y_{t-1,3}$ |
| H | 0 | $X_{t42}$ | $X_{t43}$ | $X_{t44}$ | $X_{t45}$ | $X_{t46}$ | $Y_{t-1,4}$ |
| ICU | 0 | $X_{t52}$ | $X_{t53}$ | $X_{t54}$ | $X_{t55}$ | $X_{t56}$ | $Y_{t-1,5}$ |
| D | 0 | 0 | 0 | 0 | 0 | $X_{t66}$ | $Y_{t-1,6}$ |
| Total | $Y_{t1}$ | $Y_{t2}$ | $Y_{t3}$ | $Y_{t4}$ | $Y_{t5}$ | $Y_{t6}$ | $N$ |

- $X_{t35}$ corresponds to the number of individuals that moved from category Q (number 3) at time $t-1$ into category ICU (number 5) at occasion $t$

# Model assumptions

- To account for overdispersion we alternatively assume a Dirichlet-Multinomial distribution with the vector of parameters $\boldsymbol{\alpha}_{tj}$ depending on the $\boldsymbol{\beta}$ parameters

$$\boldsymbol{X}_{tj}|\boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \quad \boldsymbol{\beta}_j \sim \mathrm{Dir-Mult}(y_{t-1,j}; \boldsymbol{\alpha}_{tj}),$$

for $t \in \mathcal{T}'$ and $j \in \mathcal{K}$

- The conditional expected value is

$$\mathrm{E}(\boldsymbol{X}_{tj}|\boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \boldsymbol{\beta}_j) \quad = \quad \frac{y_{t-1,j}}{\alpha_{tj+}}\boldsymbol{\alpha}_{tj},$$

- The variance-covariance matrix is

$$\mathrm{Var}(\boldsymbol{X}_{tj}|\boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \boldsymbol{\beta}_j) \quad = \quad \frac{y_{t-1,j}}{\alpha_{tj+}^2}[\mathrm{diag}(\boldsymbol{\alpha}_{tj}) - \boldsymbol{\alpha}_{tj}\boldsymbol{\alpha}_{tj}']\frac{n+\alpha_{tj+}}{1+\alpha_{tj+}}.$$

# Parametrization

- We denote as $p_{tjk}$ the conditional probabilities that an individual $i$ is in category $k$ at occasion $t$ given that h/she was in category $j$ at time $t-1$

- These probabilities are collected into the vector $\boldsymbol{p}_{tj}$ and are assumed to follow a multinomial logit parametrization

$$p_{tjk} = \frac{\exp(\boldsymbol{f}'_{tjk}\boldsymbol{\beta}_{jk})}{\sum_{l \in \mathcal{D}_j} \exp(\boldsymbol{f}'_{tjl}\boldsymbol{\beta}_{jl})}, \quad t \in \mathcal{T}', \, j \in \mathcal{K}, \, k \in \mathcal{D}_j,$$

  where the design column vectors $\boldsymbol{f}_{tjk}$ contain the terms of a suitable polynomial (or spline) of time $t$, (second or third order)

- The polynomials are included in the model via the regression parameter vector $\boldsymbol{\beta}_{jk}$ ($\boldsymbol{\beta}_{jj} = 0$) and $\mathcal{D}_j$ is the set of non-zero cells in the $j$-th row

# Parametrization

- We directly assume that

$$\alpha_{tjk} = \exp(\boldsymbol{f}'_{tjk}\boldsymbol{\beta}_{jk}), \quad t \in \mathcal{T}', \, j \in \mathcal{K}, \, k \in \mathcal{D}_j$$

- The parameters in $\boldsymbol{\beta}_{jk}$ are interpreted in terms of the logit of the probability of moving to category $k$ starting from category $j$

- We use common vectors across categories containing the elements of a second or third order polynomial, and we have $\boldsymbol{f}_{tjk} = (1, t, t^2, t^3)'$ for all $t$, $j$, and $k$.

- The $\boldsymbol{\beta}_{jk}$ parameters are independent with a diffuse prior distribution

$$\boldsymbol{\beta}_{jk} \sim N(0, \sigma^2 \boldsymbol{I}), \quad j \in \mathcal{K},$$

# Parametrization

- We include covariates such as dummies to study the effect of epidemic containment policies

- Some inequality constraints are introduced to better account for some epidemiological hypotheses

  - $o_{tjk}$ is the odds referred to category $k$ with respect to category $j$ at time occasion $t$, defined as $o_{tjk} = p_{tjk}/p_{tjj}$

$$a_{jk} \leq o_{tjk} \leq b_{jk}, \quad j, k = 1, \ldots, K, \; t = 2, \ldots, T^*, \; a_{jk}, b_{jk} \in R^+.$$

# Inference

- The model is estimated through a data augmented Markov chain Monte Carlo algorithm based on a Metropolis sampler repeating two steps:

  1: for all $t > 1$ update every contingency table $\boldsymbol{X}_t$ with elements $X_{tjk}$, given the current value of the parameters and the observed margins $\boldsymbol{y}_{t-1}$ and $\boldsymbol{y}_t$

  2: update the model parameters $\boldsymbol{\beta}_{jk}$ (new parameter values $\boldsymbol{\beta}^*_{jk}$ are proposed and accepted on the basis of a Metropolis-Hastings ratio)

- The algebraic algorithm of Diaconis (1998) is employed to sample tables with fixed margins

# Inference

- It consists of:
  - (*i*) randomly selecting a row and a column of the current table so that a $2 \times 2$ subtable is identified;
  - (*ii*) proposing a switch that consists in adding (or subtracting) to the two cells in the main diagonal of the subtable a random integer number, that is subtracted (or added) to the off-diagonal cells;
  - (*iii*) accepting this sub-table with probability equal to

$$\alpha = \min\left(1, \prod_{j \in \mathcal{K}} \frac{p(\boldsymbol{X}_{tj} = \boldsymbol{x}^*_{tj} | \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \boldsymbol{\beta}_j)}{p(\boldsymbol{X}_{tj} = \boldsymbol{x}_{tj} | \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}, \boldsymbol{\beta}_j)}\right),$$

    where $\boldsymbol{x}_{tj}$ is the vector of the frequencies in the $j$-th row of the current table, $\boldsymbol{x}^*_{tj}$ is that of the proposed table, and $\boldsymbol{\beta}_j$ is the matrix containing all current regression vectors $\boldsymbol{\beta}_{jk}$, $k \in \mathcal{D}_j$

- The regression parameters are updated with a Random Walk Metropolis step

# Frequency prediction

- At each iteration of the algorithm we make in-sample and out-sample predictions

- Prediction of the frequency $y_{tk}$ at step $s$ of the algorithm is given by

$$\hat{y}_{tk}^{(s)} = \sum_{j \in \mathcal{K}} y_{t-1,j} p_{tjk}^{(s)}, \quad t \in \mathcal{T}.$$

- Out-sample predictions are given by

$$\hat{y}_{tk}^{(s)} = \sum_{j \in \mathcal{K}} \hat{y}_{t-1,j}^{(s)} p_{tjk}^{(s)}$$

  for $t > T + 1$

- We consider measures of precision that take into account the variance of the posterior parameter distribution

# Estimation of a time-evolving reproduction number

- The net reproduction number $R_t$ is estimated as

$$\widehat{R}_t^{(s)} = \frac{\widehat{\Delta I}_t^{(s)}}{\sum_{r=1}^{t-1} \omega_{s,t-1} \widehat{\Delta I}_{t-r}^{(s)}},$$

  - $\omega_{r,t-1}$ is a weight obtained by normalizing the density of the Gamma distribution with parameters 1.87 and 0.28
  - $\widehat{\Delta I}_t^{(s)}$ is the number of individuals in category new positive predicted by the model for day $t$

- The overall prediction is taken as a mean across the MCMC iterations

# Model checking

- The goodness-of-fit of the model is assessed considering a discrepancy measure between observed values and in-sample predictions

$$\widehat{\mathrm{Dist}}^{(s)} = \sum_{t \in \mathcal{T}'} \sum_{k \in \mathcal{K}} \frac{(y_{tk} - \hat{y}_{tk}^{(s)})^2}{\hat{y}_{tk}^{(s)}},$$

  computed for every MCMC iteration and taking the mean of these quantities and calculating a posterior $p$-value ($\widetilde{\mathrm{Dist}}$ consider a simulated frequency from the model)

- We compare different models in terms of predictive power with the following out-sample quantity summarized by a mean

$$\widehat{\mathrm{Dist}}_t^{(s)} = \sum_{k \in \mathcal{K}} \frac{(y_{tk} - \hat{y}_{tk}^{(s)})^2}{\hat{y}_{tk}^{(s)}}, \quad t \in \mathcal{T}^{\dagger},$$

- We also provide a similar measure to establish which category (S,R,Q, H,ICU,D) presents a higher or lower prediction power

# Application: Italian data

- As an example we refer to the Italian data collected from February 24th until April 24th, 2020 (61 days)

- The goodness-of-fit of the estimated models

| Multinomial | $\widehat{\text{Dist}}$ | $\widetilde{\text{Dist}}$ | $p$-value |
|---|---|---|---|
| Model 1 (2nd order, without constraints) | 1,658.011 | 124.670 | 0.000 |
| Model 2 (2nd order, with constraints) | 2,347.274 | 68.474 | 0.000 |
| Model 3 (3rd order, without constraints) | 1,565.587 | 122.793 | 0.000 |
| Model 4 (3rd order, with constraints) | 2,203.832 | 70.512 | 0.000 |
| Dirichlet-Multinomial | $\widehat{\text{Dist}}$ | $\widetilde{\text{Dist}}$ | $p$-value |
| Model 5 (2nd order, without constraints) | 2,608.502 | 3,060.236 | 0.679 |
| Model 6 (2nd order, with constraints) | 2,992.213 | 3,629.419 | 0.750 |
| Model 7 (3rd order, without constraints) | 2,414.970 | 2,811.524 | 0.536 |
| Model 8 (3rd order, with constraints) | 2,915.772 | 3,344.208 | 0.661 |

# Application: forecast

- Discrepancy measures for the forecasted cases (Model 8, 3rd order with constraints) according to the posterior predictive distribution

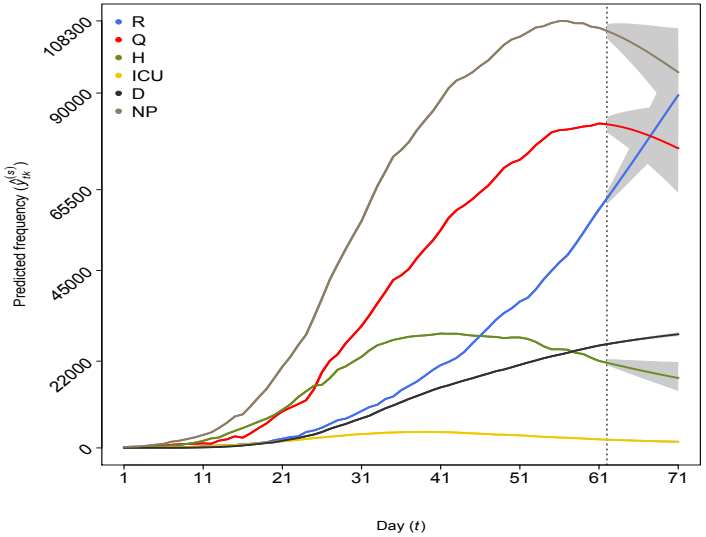
| Day | $\widehat{\mathrm{Dist}}_t$ | $\widetilde{\mathrm{Dist}}_t$ | $p$-value |
|---|---|---|---|
| 25th April | 3,231.755 | 24.523 | 0.769 |
| 26th April | 3,347.780 | 36.457 | 0.403 |
| 27th April | 2,976.716 | 19.313 | 0.198 |
| 28th April | 3,105.249 | 26.695 | 0.161 |
| 29th April | 3,216.649 | 31.738 | 0.137 |
| 30th April | 3,095.463 | 31.599 | 0.164 |
| 1st May | 2,979.734 | 37.135 | 0.118 |
| 2nd May | 3,169.230 | 47.058 | 0.103 |
| 3rd May | 3,223.772 | 58.826 | 0.095 |
| 4th May | 3,112.596 | 44.670 | 0.069 |

# Results

- The best predicted categories are D, ICU and H

| | S | R | Q | H | ICU | D | Total |
|---|---|---|---|---|---|---|---|
| $\widetilde{\mathrm{Dist}}_k^*$ | 0.000 | 1,409 | 1,397 | 372 | 31 | 12 | 3,220 |

- Daily observed and predicted counts for each category with a time horizon of 10 days and estimated 95% prediction intervals (grey)



# Results

- Estimated posterior means of the predicted transitions between categories from 25th to 26th of April, 2020 (from the 61st to the 62nd day) and 95% prediction upper and lower bounds

| | S | R | Q | H | ICU | D |
|---|---|---|---|---|---|---|
| S | 60,121,632 | 0 | 2,219 | 154 | 1 | 0 |
| R | 0 | 60,489 | 9 | 0 | 0 | 0 |
| Q | 0 | 2,665 | 79,105 | 516 | 0 | 0 |
| H | 0 | 116 | 757 | 20,925 | 73 | 197 |
| ICU | 0 | 0 | 0 | 0 | 2,023 | 149 |
| D | 0 | 0 | 0 | 0 | 0 | 25,969 |

| | S | R | Q | H | ICU | D |
|---|---|---|---|---|---|---|
| S | - | (0, 0) | (1,217, 3,188) | (0, 718) | (0, 2) | (0, 0) |
| R | - | (60,471, 60,498) | (0, 26) | (0, 0) | (0, 0) | (0, 0) |
| Q | - | (1,269, 4,357) | (77,182, 80,672) | (32, 1,479) | (0, 0) | (0, 0) |
| H | - | (0, 506) | (463, 1,129) | (20,438, 21,321) | (25, 137) | (123, 282) |
| ICU | - | (0, 0) | (0, 0) | (0, 40) | (1,963, 2,075) | (98, 210) |
| D | - | - | - | - | - | - |

# Results: obtained with data of the Lombardy region

- We also used data of the Lombardy region where the spread of the virus began in Italy and we compared the estimated values
- Estimated posterior means of the predicted transitions between categories from 25th to 26th of April, 2020

| | S | R | Q | H | ICU | D |
|---|---|---|---|---|---|---|
| S | 9,988,451 | 0 | 774 | 93 | 0 | 0 |
| R | 0 | 23,779 | 3 | 0 | 0 | 0 |
| Q | 0 | 309 | 24,123 | 389 | 0 | 0 |
| H | 0 | 170 | 379 | 8,142 | 28 | 72 |
| ICU | 0 | 0 | 0 | 0 | 703 | 52 |
| D | 0 | 0 | 0 | 0 | 0 | 13,106 |

| | S | R | Q | H | ICU | D |
|---|---|---|---|---|---|---|
| S | - | (0, 0) | (246, 1,210) | (0, 581) | (0, 1) | (0, 0) |
| R | - | (23,759, 23,782) | (0, 22) | (0, 0) | (0, 0) | (0, 0) |
| Q | - | (0, 1,337) | (22,668, 24,796) | (0, 1,565) | (0, 0) | (0, 0) |
| H | - | (0, 510) | (70, 737) | (7,683, 8,446) | (4, 67) | (30, 121) |
| ICU | - | (0, 0) | (0, 0) | (0, 5) | (668, 731) | (24, 87) |
| D | - | - | - | - | - | - |

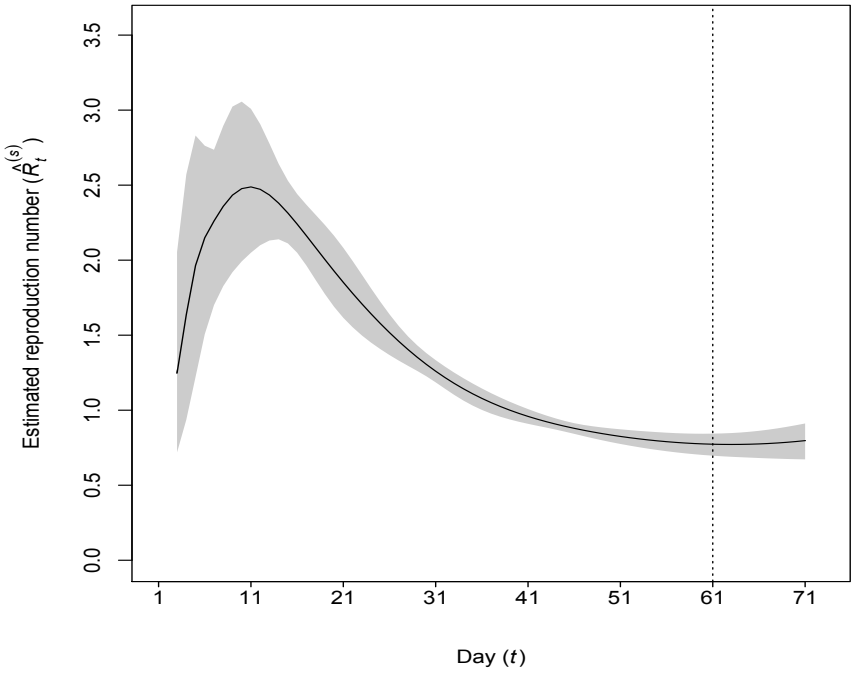
# Results

- Estimated posterior means and the 95% predicted interval for the increase in totals for H and ICU from 26th to 29th April, 2020 obtained with the Italian data

| Day | H | PI | ICU | PI |
|---|---|---|---|---|
| 25th April | -472 | (-1,047, 446) | -76 | (-140, -2) |
| 26th April | -465 | (-1,032, 462) | -73 | (-134, -2) |
| 27th April | -460 | (-1,012, 459) | -69 | (-128, -1) |
| 28th April | -450 | (-997, 465) | -67 | (-122, 0) |
| 29th April | -442 | (-981, 470) | -63 | (-118, 0) |
| 30th April | -431 | (-972, 450) | -60 | (-112, 2) |
| 1st May | -420 | (-952, 465) | -57 | (-107, 3) |
| 2nd May | -409 | (-948, 459) | -55 | (-104, 4) |
| 3rd May | -397 | (-942, 484) | -52 | (-99, 6) |
| 4th May | -384 | (-925, 454) | -50 | (-95, 7) |

# Results

- Estimated and predicted (from the horizontal line) reproduction number $R_t$ (61 observed days, prediction from 25th of April to 4th of May). Estimated 95% credibility and prediction intervals in grey



# Conclusions

- Data of the other regions can be used to make a comparison of transition rates among categories across regions
- The proposed models are formulated in a general way, and they may be adapted to a different number of categories according to data availability
- They are especially useful when the individual level information is not available, this is the case of aggregated health data, of data deriving from meta-analytic procedures and official statistics
- They could be used for the analysis of the transitions between categories of malignant tumors as in the tumor, node, metastasis classification when it is conducted on aggregated data or for the analysis of the transitions between levels of severity of other diseases

# Main References


- Bartolucci F, Farcomeni A, Pennoni F. *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman and Hall/CRC. 2013.
- Diaconis B. Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics* 1998; 26: 363–397.
- Dunson DB. Commentary: Practical Advantages of Bayesian Analysis of Epidemiologic Data. *American Journal of Epidemiology* 2001; 153: 1222–1226.
- Eleftheraki AG, Kateri M, Ntzoufras I. Bayesian analysis of two dependent $2\times2$ contingency tables. *Computational Statistics & Data Analysis* 2009; 53: 2724-2732.
- Mosimann JE. On the compound multinomial distribution, the multivariate $\beta$-distribution, and correlations among proportions. *Biometrika* 1962; 49: 65–82.
- Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; 82: 528–540.
- Zucchini W, MacDonald IL, Langrock R. *Hidden Markov Models for Time Series: An Introduction Using R*. New York: Springer-Verlag. 2017.

# Current results

- Obtained with data from 22 February to 15 May, 2021
- Estimated posterior means of the predicted transitions between categories from 15th to 16th of May 2021

| | S | R | Q | H | ICU | D |
|---|---|---|---|---|---|---|
| S | 56156904 | 6572 | 118 | 30 | 2 | 0 |
| R | 0 | 3667448 | 28033 | 998 | 2 | 0 |
| Q | 0 | 35544 | 282874 | 105 | 0 | 9 |
| H | 0 | 1267 | 125 | 10985 | 18 | 98 |
| ICU | 0 | 5 | 1 | 2 | 1732 | 66 |
| D | 0 | 0 | 0 | 0 | 0 | 124063 |

# Current results

- Predicted in Italy for the current week with the D-M model with the 2nd order polynomial: confirmed cases 43,125 and deceased 1,094
- Daily observed and predicted counts for each category with a time horizon of 10 days (from 22nd of February to 25rd of May, 2021)