International Conference on Mathematical Analysis and Applications in Science and Engineering
ICMA^{2}SC'24
ISEP Porto-Portugal, June 20 - 22, 2024

EXPLORING FORMS OF DISAGGREGATING COVID-19 DATA: AN EXAMPLE

Conceição Leal[1,2⋆] **Teresa A. Oliveira**[1,2] **Amitava Mukherjee**[3]
Amílcar Oliveira[1,2]

[1] Universidade Aberta, Portugal
[2] CEAUL, FCUL, Portugal
[3] XLRI-Xavier School of Management, India

[1⋆]conceicao.leal2010@gmail.com

Abstract: COVID-19 data provided by Portuguese public health authorities lack consistency in periodicity and metrics. To facilitate time series analysis, we transformed those data to achieve homogeneous periodicity and metrics. We present one method we used and assess the potential introduced bias and its impact on spatial distribution models of COVID-19 in Portugal, using spatial and non-spatial models. Comparing models fitted with transformed data to those with observed data for two specific days, we found no clear evidence of a worse fit for the disaggregated data.

keywords: Times Series Disaggregation; Modelling COVID-19 data; Model fit.

MSC2020: 49-XX; 34-XX; 92-XX.

1 Introduction

The SARS-CoV-2 virus, which causes COVID-19, was a significant public health threat. Spatial epidemiology can elucidate the disease's distribution and spread due to its varying incidence rates across locations, leading to uneven impacts. The dynamics of the disease feature a lag between exposure and detection, with asymptomatic transmission affecting public health responses. Given human-to-human transmission of the virus and the influence of atmospheric conditions on human behavior, this study incorporates meteorological variables (temperature, humidity, and rainfall) in modelling daily COVID-19 cases.

Portuguese public health authorities' COVID-19 data are inconsistent in periodicity and metrics. From March 24, 2020, to March 24, 2021, we transformed the data, potentially introducing bias. We evaluate if such bias affects geographic risk distribution maps of COVID-19. We modelled daily COVID-19 cases in Mainland Portugal, considering spatial data and meteorological factors, to capture pandemic evolution. Weather data, from IPMA, included maximum and minimum temperatures, humidity, and rainfall with a 7-day lag. Using Latent Gaussian Models via the R-INLA package, we estimated daily cases in 278 counties, analysing data from June 1, 2020 and March 24, 2021, to assess the transformation's impact on model accuracy.

The best model for 1st June 2020, using observed daily data, included only unstructured spatial effects. For 24th March 2021, using transformed data, the best-fitting model also included only unstructured spatial effects. Minimum temperature was significant on 24th March 2021, and minimum humidity on 1st June 2020.

2 Results and discussion

2.1 A Latent Gaussian Model (LGM)

Suppose y_{it} the number of cumulative daily cases of COVID-19 in county i of mainland Portugal (i=1,2,...,278) on day t (t=1,2,...,142), modelled as:

$$y_{it} \sim Poisson(\mu_{it} E_i) \tag{1}$$

with link function

$$\eta_{it} = ln(\mu_{it}) \tag{2}$$

defining Latent Gaussian Field as $x = (\eta, \alpha, f, \beta)$. Since η, α, f and β are random variables, they require parameters. Suppose *f(i)* is a parameter for y_i and Ψ is a parameter for x, then we can build a hierarchical structure for this.

In the first stage, we used a latent Gaussian model as follows.

$$\eta_{it} = ln(\mu_{it}) = \alpha + \sum_{k=1}^{m} \beta_i x_i^k + \sum_{j=1}^{c} P\left(x_{it-lag}^j\right) + s_i + \nu_i + \gamma_t + \varepsilon_t, i = 1, ...n \quad \text{and} \quad t = 1, .., T \tag{3}$$

where μ_{it} represents the mean or relative risk; α represents the intercept term (overall mean or risk); $s_i = f_1(i)$ represents spatially structured residuals in county i; $\nu_i = f_2(i)$ represents spatially unstructured residuals in county i; $\gamma_i = f_3(t)$ represents temporally structured residuals in period t; $\varepsilon_t = f_4(t)$ represents temporally unstructured residuals in period t; $\beta_i x_i^i$ represents linear effects of jth covariate in county i; $P\left(x_{it-lag}^j\right)$ represents a polynomial that allows establishing a non-linear relationship between temporally-lagged covariates, X_{it-lag} and μ_{it}.

In the second stage, we considered spatial structure with two distributions as follows:

1. s_i spatially-structured effect on county i with an intrinsic conditional autoregressive (iCAR) structure (BYM model [1])

$$s_i | s_{(j \not= i)} \sim Normal\left(\frac{1}{N_i}\sum_{j=1}^{n} w_{ij} s_j; \frac{\sigma_s^2}{N_i}\right) \tag{4}$$

   where N_i is the number of neighbours that county i has and w_{ij} is the element (i, j) of the row-standardised matrix W of dimension $n \times n$ that represents the neighbourhood matrix for the counties: $w_{ij} = \frac{1}{N_i}$ if counties i and j are neighbours, otherwise $w_{ij} = 0$ and σ_s^2 represents the variance of the spatially-structured effect.

2. ν_i spatially-unstructured effect over the counties, an independent and identically distributed Gaussian prior is considered

$$\nu_i \sim Normal\left(0; \sigma_\nu^2\right) \tag{5}$$

   where σ_ν^2 represents the variance of the spatially-unstructured effect of the model.

3. γ_t temporally-structured effect was modelled through a second-order random walk

$$\gamma_t | \gamma_{t-1}, \gamma_{t-2} \sim Normal\left(2\gamma_{t-1} + \gamma_{t-2}; \sigma_\gamma^2\right) \tag{6}$$

   where σ_γ^2 represents the variance component.

4. ε_t temporally-unstructured effect, an independent and identically distributed Gaussian prior is considered

$$\varepsilon_t \sim Normal\left(0; \varepsilon^2\right) \tag{7}$$

$E_i = \sum_j f_j P_{ij}$ where f_i is the incidence of cases in the age group j for the whole population of the country and P_{ij} the population in county corresponding to age group j ([2]) represents the expected number of cases and $ln(E_i)$ is used as the offset term of the linear predictor.

Furthermore, in the third stage, we defined the hyperparameter as follows:

$$\Psi = \left(\sigma_s^2; \sigma_\nu^2\right) \tag{8}$$

Informative prior for the hyperparameter are specified by INLA's (R-INLA package) default:

$$\sigma_s^2 \sim Gamma^{-1}(1; 0.0005) \text{ and } \sigma_\nu^2 \sim Gamma^{-1}(1; 0.0005) \tag{9}$$

The neighbourhood matrix W for the counties is defined by

$$w_{ij} = \begin{cases} \frac{1}{N_i}, & \text{counties } i \text{ and } j \text{ are neighbours (a common geographical border)} \\ 0, & otherwise \end{cases}$$

2.2 COVID-19 data and IPMA data

The analysis focused on the counties of mainland Portugal (i.e., excluding Azores and Madeira islands), a contiguous study area. COVID-19 data were sourced from the online repository "Data Science for Social Good Portugal" (https://github.com/dssg-pt/covid19pt-data).

The study period spans from 24 March 2020 to 24 March 2021. The health authorities' data were inconsistent in metrics throughout this period. From 24 March to 4 July 2020, daily cumulative cases were reported; from 14 July to 26 October 2020, cumulative cases were reported weekly. From 11 November 2020, data were published weekly as cumulative incidence per fortnight (14 days) per 100k inhabitants, calculated using the population estimates from 31 December 2019 by the National Statistics Institute (as noted in the DGS Situation Reports from 16 November 2020). Since fortnightly reports are weekly, the first week of each fortnight overlaps with the last week of the previous fortnight, preventing direct calculation of new daily or weekly cases.

We explored various methods of disaggregating data to address this inconsistency. For the data from 11th November 2020 to 24th March 2021, in one of the methods, we assumed a uniform distribution of new cases across each 14-day period, U(0,14). To estimate the daily number of new cases in each overlapping interval, we calculated the average of the estimated daily new cases from the uniform distribution of one fortnight with those from the next fortnight.

Given x_{di} the number of new cases on day d in each week i, with $d = 1, 2, \ldots, 7$ and $i = 1, 2, \ldots, 21$, starting from November 4, 2020; a_{dj} the number of new daily cases in day d of the first week of each fortnight j, estimated by the uniform distribution $U(0, 14)$ and $a_{(d+7)(i-1)}$ the number of new daily cases in day d of the second week of each fortnight $j-1$, estimated by the uniform distribution U(0,14)(Fig. 1), then:

$$x_{di} = \frac{a_{(d+7)(j-1)} + a_{dj}}{2} \tag{10}$$

for $i = 1, \ldots, 20$ and $d = 1, 2, ..., 7$.

y_i are new cases obtained by adding the number x_{di} of cases between two reporting days.

Covariates were max. and min. temperature, max. and min. humidity and maximum rainfall, (IPMA data). Per the literature on incubation time of COVID-19, the parameter Lag was varied from Lag=2 to Lag=14 days. We used 7-days lag.

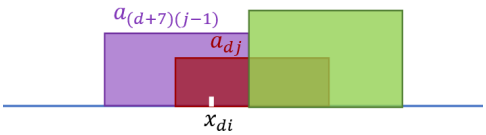
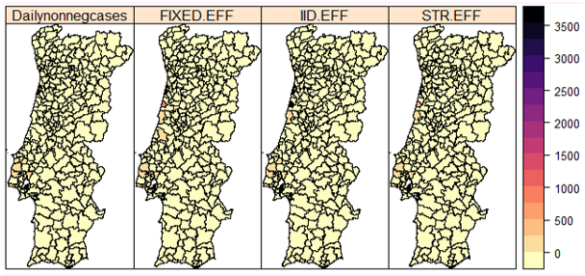


Figure 1: Estimation of the number of COVID-19 cases from the 14-day incidence

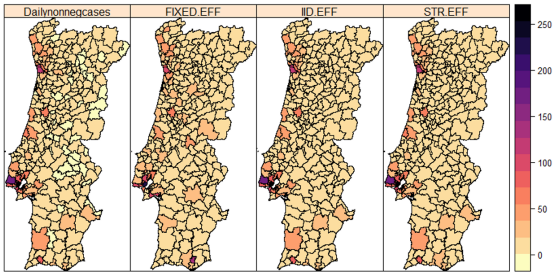
Using Integrated Nested Laplace Approximation (INLA)[3], we fitted: fixed effects model; unstructured spatial effects model; a BYM [3] (structured and unstructured) spatial effects model (iCAR) with weather covariates (7-days lag), to data for June 1, 2020 (non-transformed) and March 24, 2021 (transformed). We selected best-fitting models using DIC and WAIC criteria.

MODEL	DIC		WAIC	
	2020/6/1	2021/3/24	2020/6/1	2020/3/24
Fixed effects	403.78	946.55	3.86e+18	972.66
Fixed effects + spatially-structured(+)unstructured effects	260.78	756.72	5.68e+22	756.72
Fixed effects + spatially-unstructured effects (iCar)	258.76	434.41	6.38e+16	423.98

Table 1: Comparative table of models



(a) June 1, 2020: observed daily number of cases; estimated daily number of cases by the three models



(b) March 24, 2021: observed daily number of cases; estimated daily number of cases by the three models

Figure 2: Comparison of observed and estimated daily number of cases

In both cases the unstructured spatial effects model (iCAR) better captures the distribution of the epidemic across the counties of Portugal. DIC criterium supports this conclusion.

The maps of Figure 2 shows the daily cumulative number of cases observed for June 1st, 2020 and for March 24th, 2021, and the daily cumulative number of cases predicted by the models that considers all covariates and 7-days lag. Observing the maps, we can conclude that models with spatial effects fitted with observed data and with transformed data well reflect the spatial distribution of the number of observed cases, without clear evidence of a worse fit to the transformed data.

References

[1] Besag, J., York, J., Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. Annals of the institute of statistical mathematics, 43, 1-20.

[2] Briz-Redón, Á., Serrano-Aroca, Á. (2020). A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. Science of the total environment, 728, 138811.

[3] Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology, 71(2), 319-392.