Joint modeling of number of recorded cases of COVID-19 and wastewater measurements daily data: A Dynamic Hierarchical approach using data augmentation technique

Pamela M. Chiroque-Solano.

September, 2021

Statistical analysis

The wastewater surveillance system has been showing to be a valuable proxy to monitor the COVID-19 pandemic. That is because the association between wastewater measurements and COVID-19 epidemiological indicators was corroborated in multiple studies (Peccia et al., 2020; Cao and Francis, 2021; Olesen et al., 2021). This correlation is maximized when case counts are lagged (Olesen et al., 2021).

Since the dynamic of the coronavirus epidemic changes, daily quantitative tools to alert bursts of disease incidence need to be implemented. A temporal multilevel model for jointly analyzing the behavior of SARS-CoV-2 wastewater concentrations and the COVID-19 cases at the ETAR-level was developed. In the first level the distribution of the numbers of contributors is conditioned on the filtered viral load measurements and follows a dynamic generalized structure (West et al., 1985; Gamerman and Migon, 1993; Migon et al., 2005). The SARS-CoV-2 wastewater concentrations were processed with error on a regular or intermittent schedule. In this level of the hierarchy three kinds of genes copies/ml of wastewater were combined to represent an effective wastewater epidemiological component. A direct consequence is the sharing of information of all types of gees. This strategy allows us to maximize the wastewater surveillance information to be filtered to the first level (Gelman and Hill, 2007). Under the Bayesian approach the data augmentation technique (Tanner and Wong, 1987) is a natural way for dealing with missing data, since the SARS-CoV-2 wastewater concentration samples were collected on a non-regular schedule. To highlight this characteristic the directed acyclic graph (DAG) is shown in Figure 1.

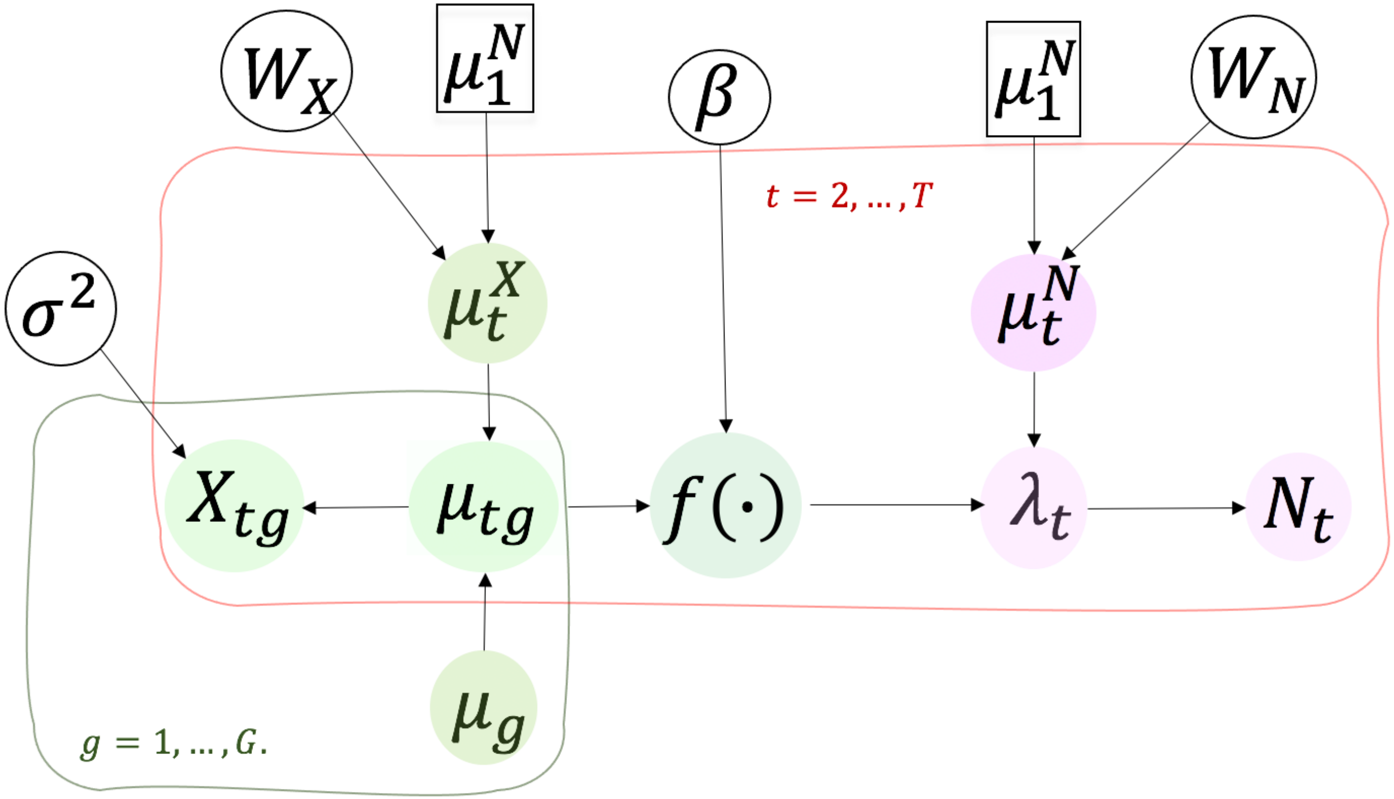

Figure 1: Directed Acyclic Graph (DAG) of the Proposal Model.

1 Model Conceptualization, Design, Implementation

The model's rationale was defined separately by location $l = 1, \ldots, L$. The index l has been suppressed for brevity. That means that the model did not share information between locations since each ETAR has its particular treatment process. The model assumes that N_t, the number of recorded cases of COVID-19, follows a Poisson distribution with rate λ_t. The $\log(\lambda_t)$ term contains the time-varying mean component μ_t^N and the predictor term $\beta f\left(\{\mu_{g,t}\}_{g=1}^{G}\right)$ relates to the wastewater information, where β is a regression parameter that describes the impact of the filtered viral load measurements on $\log(\lambda_t)$. The function $f(\bullet)$ filters the desired information from the viral loads of all genes, for example a quantile or the expected value. The component $\mu_{t,g}$ is defined below.

Let $\log X_{t,g}$ be defined as the logarithm of the total value of viral load in a given time $t = 1, \cdots, T$ for the gene $g = 1, \ldots, G$; we are assuming that $X_{t,g}$ follows a Normal distribution with mean $\mu_{t,g}$ and variance σ^2.

The component $\mu_{t,g}$ is modeled with a Normal hierarchical structure with mean equal to the dynamic component μ_t^X and variance U_g which represents the features for each gene.

The last system equations describe the two dynamic components μ_t^X and μ_t^N for the viral load and the number of recorded cases of coronavirus respectively. The w_t^N and w_t^X components are assumed to be distributed according to a Normal distribution with mean 0 and precision $1/W_N$ or $1/W_X$ respectively.

$$
\begin{aligned}
\text{Observation equation: } N_t \mid \mu_{\bullet,t}, \lambda_t &\sim \mathcal{P}(\lambda_t), \quad \lambda_t > 0, \quad \log \lambda_t = \mu_t^N + \beta f\left(\{\mu_{t,g}\}_{g=1}^{G}\right), \\
\log X_{g,t} \mid \mu_{g,t}, \sigma^2 &\sim \mathcal{N}(\mu_{g,t}, \sigma^2), \quad \sigma^2 > 0, \\
\text{Hierarchy equation: } \mu_{t,g} &= \mu_t^X + u_g, \quad u_g \sim \mathcal{N}(0, U_g), \\
\text{System/State equation: } \mu_t^X &= \mu_{t-1}^X + w_t^X, \quad w_t^X \sim \mathcal{N}(0, W_X), \quad \mu_0^N \sim \mathcal{N}(U_X, W_1) \\
\mu_t^N &= \mu_{t-1}^N + w_t^N, \quad w_t^N \sim \mathcal{N}(U_X, W_N), \quad \mu_0^X \sim \mathcal{N}(0, W_2) \qquad (1)
\end{aligned}
$$

2 Inference

The vector $\mathbf{\Theta} = \{W_{\mathcal{X}}, W_N, U_g, \sigma^2, \beta\}$ contains the unknown quantities to be estimated under the Bayesian approach. In this context, we assume prior independence for the parameters β Normal distribution with mean vector 0 and the variance is 100. The variances W_X, W_N, U_g and σ^2 each follows an inverse Gamma prior with shape and scale parameters given by 0.01, which implies infinite prior mean and variance. The posterior is not known in closed form then a Markov chain Monte Carlo (MCMC) method was used to obtain samples from the posterior distribution using the R2jags packages (v0.6-1; Yu-Sung and Masanao (2020)) in the R environment (R Development Core Team, 2020).

Four chains were run for 10,000 iterations, the first half of which were discarded to avoid transient effects. Convergence was verified via the Rhat values, meaning that all Rhat values did not exceed 1.1 as Gelman and Rubin (1992a) required. Furthermore, convergence was also visually verified using the joint trace plot of the four chain iterations.

To contrast the results for different ETARs, measurements such as (rMSD) root-mean-square deviation and (RB) Relative bias indicators was used.

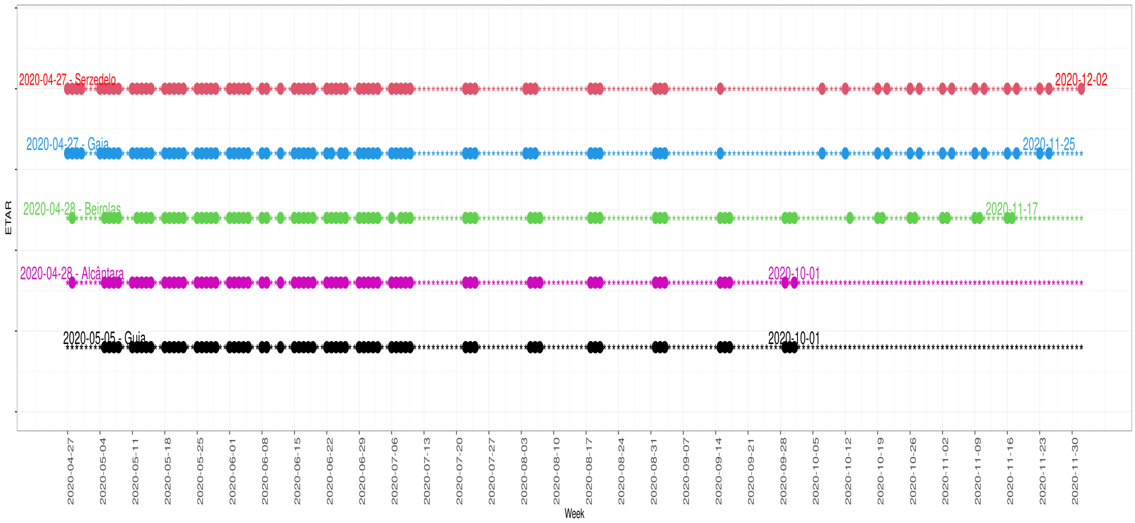

Figure 2: Sampling schedules by ETARs

3 Input Data management

The raw data relative to the number of reported COVID19 new cases were weighted using the cover population values per districts. The treatment of the viral load consists of the following steps. First to the variable "point" is filtered to the category "entrance" and "sample" to the category "composed". Then, if the viral load measure is lesser than or equal to the limit of detection (LOD) then the corrected viral load is LOD/2 times the flow value times 1000. Otherwise the viral load is updated to the current viral load times the flow value times 1000. Finally, the logarithm to base 10 is computed on the corrected viral load. The LOD varies by ETAR, that is LOD-E =2730, LOD-R = 3790 and LOD-N = 3940.

4 Application and Analysis of results

The CoviDetect runs a research programme across 5 of ETARs of Portugal in the Nort with Gaia and Serzedelo, LVT with Alcântara, Beirolas and Guia. These different epidemiological realities were covered by the sampling design throughout almost the entire year of 2020 covering. The SARS-CoV-2 wastewater concentrations were processed with error on a regular or intermittent schedule from 2020-04-27 until 2020-12-02. Figure 2 show these details.

Applying a dynamic hierarchical approach allowed an estimation of relationships between wastewater measurements time series and the number of recorded cases of COVID-19 data. The initial information at $t = 0$ is given by the equations $\mu_0^N \sim \mathcal{N}(U_N, W_1)$ and $\mu_0^X \sim \mathcal{N}(U_X, W_2)$, where U_N and U_X, are the mean value of the number of recorded cases of COVID-19 and of wastewater daily collected measurements, respectively. Let W_1 and W_2 be equal to one and the function filter was $f(\bullet) = \mu_t^X$.

By predicting the number of viral load contributors as a function of three types of viral load genes from wastewater surveillance, we found that the proposed model shows a good performance since the points are close to a regressed diagonal line and the 95% credible intervals of the marginal predictive posterior contain the observed value, see Figure 3.

ETAR	Gene			Forecasting	Fitted
	E	N	R		
Alc	0.145	0.103	0.105	2.253	1.752
Bei	0.085	-0.048	-0.006	3.340	2.897
Gai	-0.059	0.251	0.081	7.446	6.711
Gui	0.017	-0.002	0.004	0.146	1.545
Ser	0.110	-0.030	0.036	11.286	10.856

Table 1: Estimated values to evaluated the performance of the prediction and forecasting results based on the MRB criteria

Note: The mean relative bias (MRB), in percentage, is defined as $MRB = 100.\theta^{-1}\frac{\sum_{s=1}^{S}(\hat{\theta}^{(s)}-\theta)}{S}$, where $\hat{\theta}^{(s)}$ is the marginal predictor posterior mean (MPPM) respective observed valued θ. Let the forecasting number of cases of COVID-19 in the time horizon at time $T + H$, be N_{T+H}, where H is the planning horizon (H=5). Based on the decision theory fundaments if the quadratic loss is used as a utility function then the decision value (forecast value) that maximizes the utility function on the marginal predictive posterior distribution N_{T+H} is given by the posterior mean. Therefore, the forecasting number of cases of COVID-19 is obtained through the posterior mean of the N_{T+H}, which is given by:

$$p(N_{T+H} \mid D_T) = \int_{\Theta} p(N_{T+H} \mid \theta) p(\theta \mid \boldsymbol{D}) d\Theta$$

where $p(\theta \mid \boldsymbol{D})$ described the posterior distribution.

As expected the presence of the large system variability and short period of sampling (the Guia ETAR) results in 95% credible intervals very wide for days 222, 223 and 226, for instance (Figure 3(B)). This excess of uncertainty coming from the marginal predicted posterior distribution could be understood

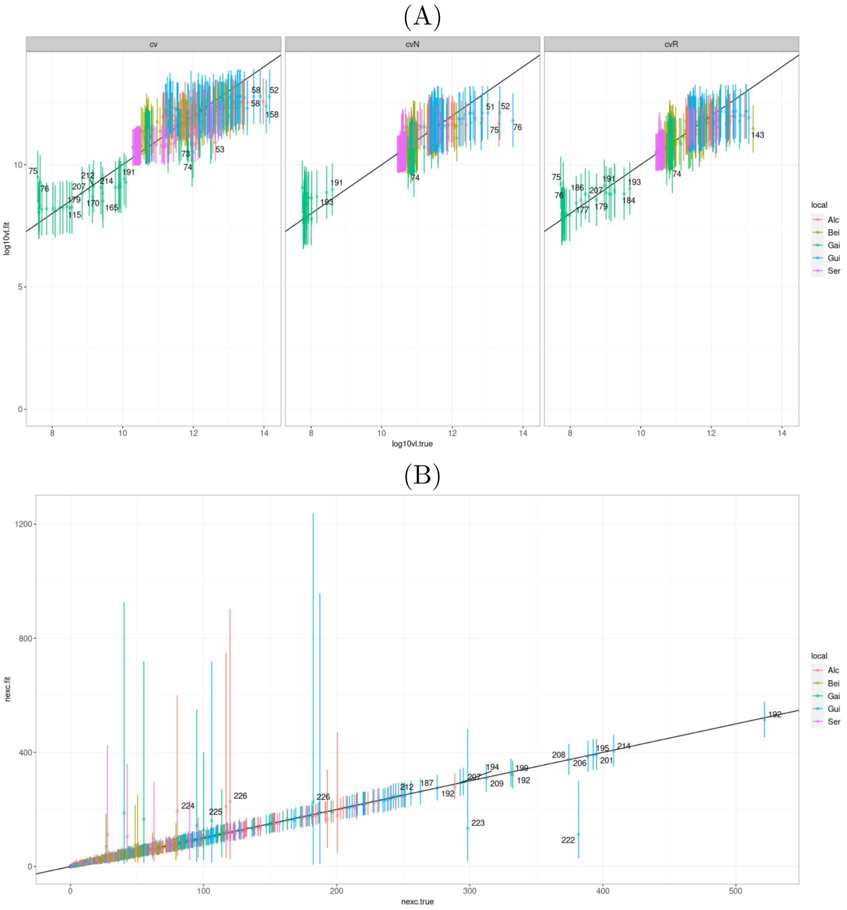

Figure 3: Scatter plots of Marginal predictive posterior mean against the observed value. The bar corresponds to the 95% credible intervals from (A) the log10 viral load distribution X_{gt} by gene E,R and N and (B) the number of recorded cases of COVID-19 N_t, respectively.

when one examines the relative bias for Serzedelo.

The results in Table 1 indicate that the fit showed better performance than the forecasting procedure. This is not unexpected. The analysis of the viral load data reveals that the Gaia, Serzedelo and Alcântara presented bigger relative bias values than the other ETARs Table 1. The viral load measures can be considered to be equivalent for the three genes analyzed. That means the credible intervals over time overlap between genes for all possible values of g (i.e. $g = 1 \dots, G$). This provides greater flexibility in the interpretation. The consistent quantification and modeling of the three genes information simultaneously helped to recover the full trend curve via the data augmentation technique. Note that, in Figure 5 this recovered trend describes particular patterns for each ETAR, Figure 4.

Figures (6,7,8,9,10) proved that for most locations (ETARs), i.e. all but Gaia, in terms of the marginal predictive distribution, the viral load measure X and the contributors count N rise and fall together over the study period. These outputs confirm that the wastewater tracking provides an effective alternative to use as an alert of COVID-19 incidence patterns.

5 Final Remarks

Our model provides realistic estimates and predicted values, allowing practitioners to set fair new strategies against the spread of the SARS-CoV-2 virus. An application was presented and a hierarchically structured model by ETARs (Gelman and Hill, 2007) was fitted. The inference procedure was done from the Bayesian approach and estimation was carried out through MCMC methods via the rjags software (Yu-Sung and Masanao, 2020). This allow us to recovered the SARS-CoV-2 wastewater concentrations were not collected in the regular schedule (2020-04-27 until 2020-12-02).

For future work, a model that incorporates heavy-tailed distributions on the viral load density and overdispersion on the distribution of the numbers of contributors could be considered. Covariates such as number of test, hospitalizations, or some epidemiological index could be included in the model to help to obtain the number of viral load contributors as a function of three types of viral load genes from wastewater surveillance.

6 Input data and code availability

The epidemiological data and code used in this study are available upon request.

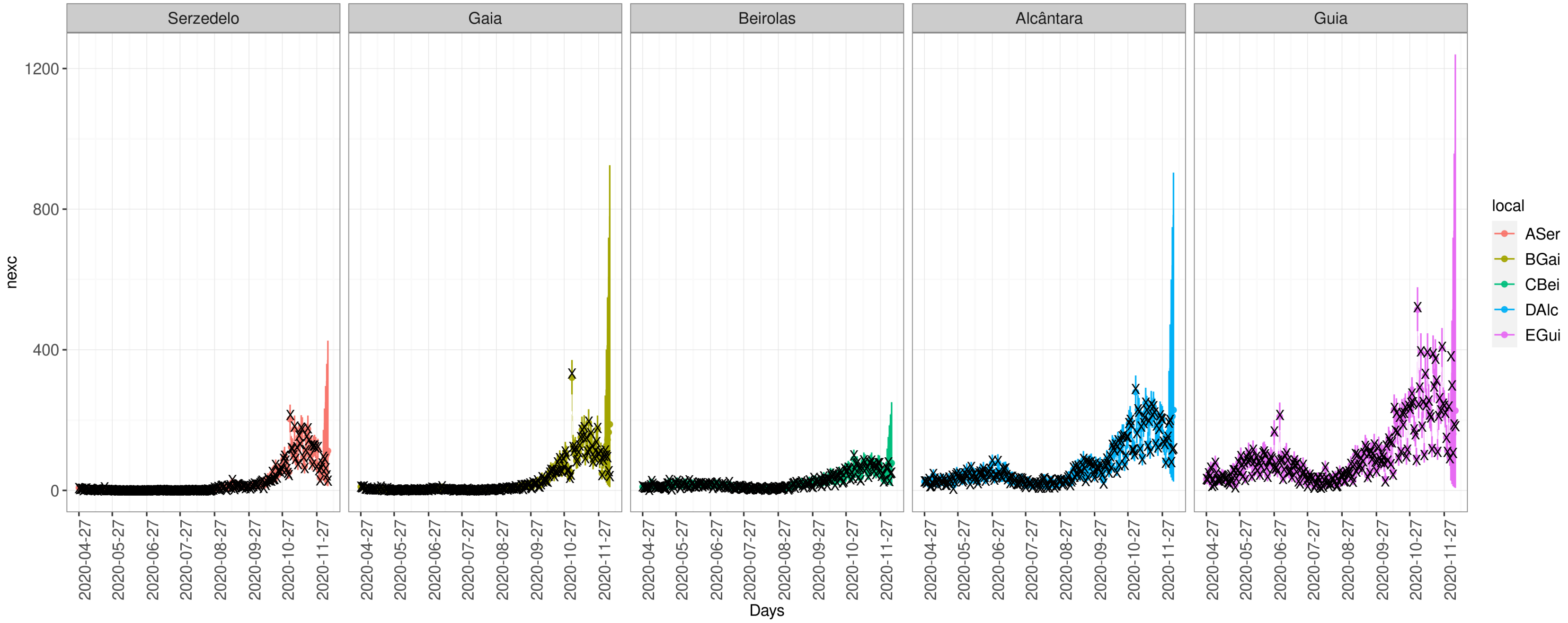
Serzedelo
Gaia
Beirolas
Alcântara
Guia
nexc
1200
800
400
0
2020-04-27
2020-05-27
2020-06-27
2020-07-27
2020-08-27
2020-09-27
2020-10-27
2020-11-27
Days
local
ASer
BGai
CBei
DAlc
EGui

Figure 4

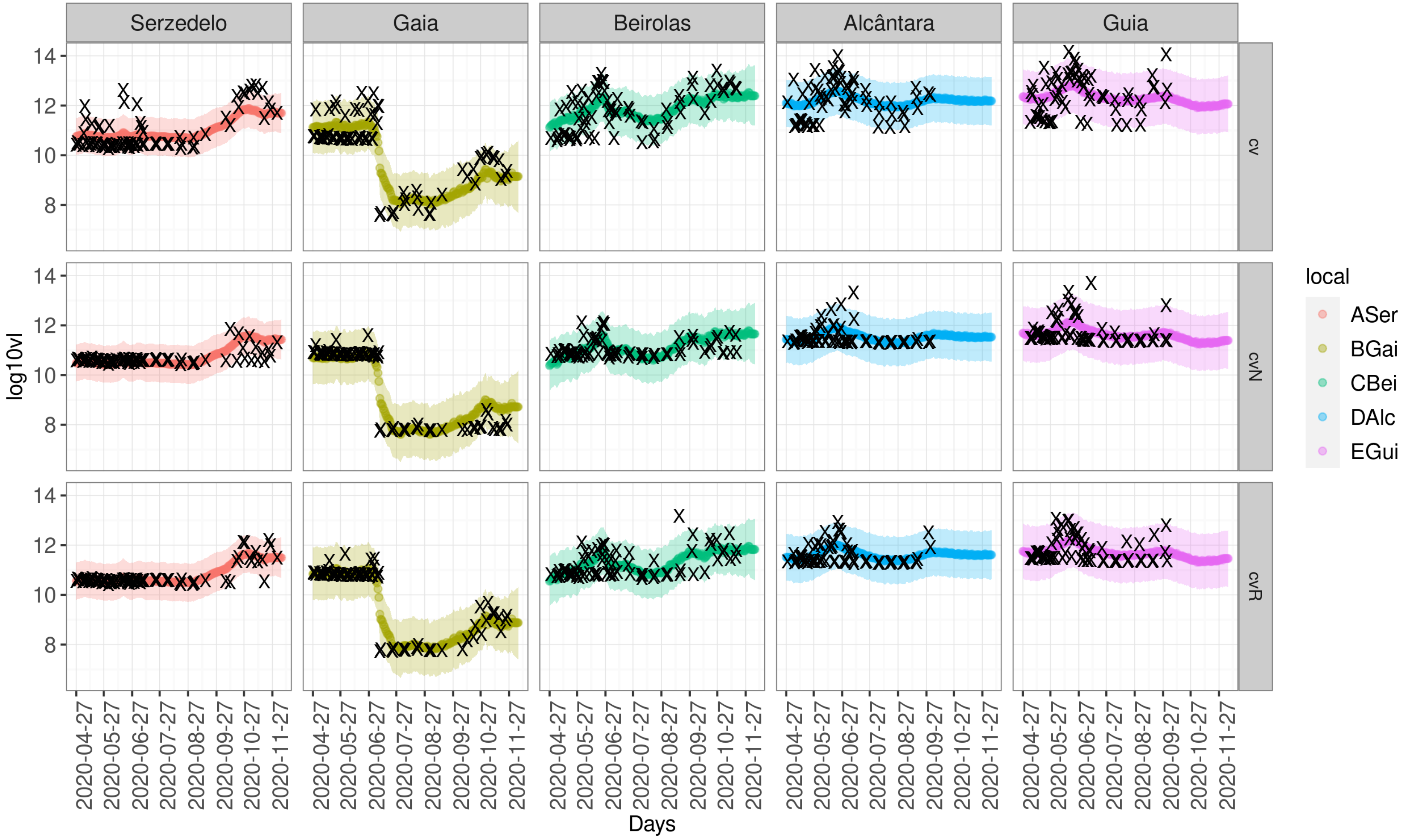
Serzedelo
Gaia
Beirolas
Alcântara
Guia
cV
cVN
cVR
log10vl
14
12
10
8
Days
2020-04-27
2020-05-27
2020-06-27
2020-07-27
2020-08-27
2020-09-27
2020-10-27
2020-11-27
local
ASer
BGai
CBei
DAlc
EGui

Figure 5

References

Cao, Y., Francis, R., 2021. On forecasting the community-level covid-19 cases from the concentration of sars-cov-2 in wastewater. The Science of the total environment 786, 147451. doi:10.1016/j.scitotenv.2021.147451.

Gamerman, D., Migon, H.S., 1993. Dynamic hierarchical models. Journal of the Royal Statistical Society B 55, 629–642.

Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York.

Gelman, A., Rubin, D., 1992a. Inference from iterative simulation using multiple sequences (with discussion). Statistical Science 7, 457–511.

Migon, H.S., Gamerman, D., Lopes, H., Ferreira, M., 2005. Dynamic models, in: Handbook of Statistics 25. Elsevier, Amsterdam, pp. 557–592.

Olesen, S., Imakaev, M., Duvallet, C., 2021. Making waves: Defining the lead time of wastewater-based epidemiology for covid-19. Water research 202, 117433. doi:110.1016/j.watres.2021.117433.

Peccia, J., Zulli, A., Brackney, D., Grubaugh, N., Kaplan, E., Casanovas-Massana, A., Ko, A., Malik, A., Wang, D., Wang, M., Warren, J., Weinberger, D., Arnold, W., Omer, S., 2020. Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. Nature Biotechnology 38, 1164 – 1167. URL: https://doi.org/10.1038/s41587-020-0684-z, doi:10.1038/s41587-020-0684-z.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540. doi:10.1080/01621459.1987.10478458.

West, M., Harrison, P.J., Migon, H.S., 1985. Dynamic generalized linear models and bayesian forecasting. Journal of the American Statistical Association 80, 73–83. doi:10.1080/01621459.1985.10477131.

Yu-Sung, S., Masanao, Y., 2020. R2jags: Using R to Run 'JAGS'. URL: https://CRAN.R-project.org/package=R2jags. r package version 0.6-1.

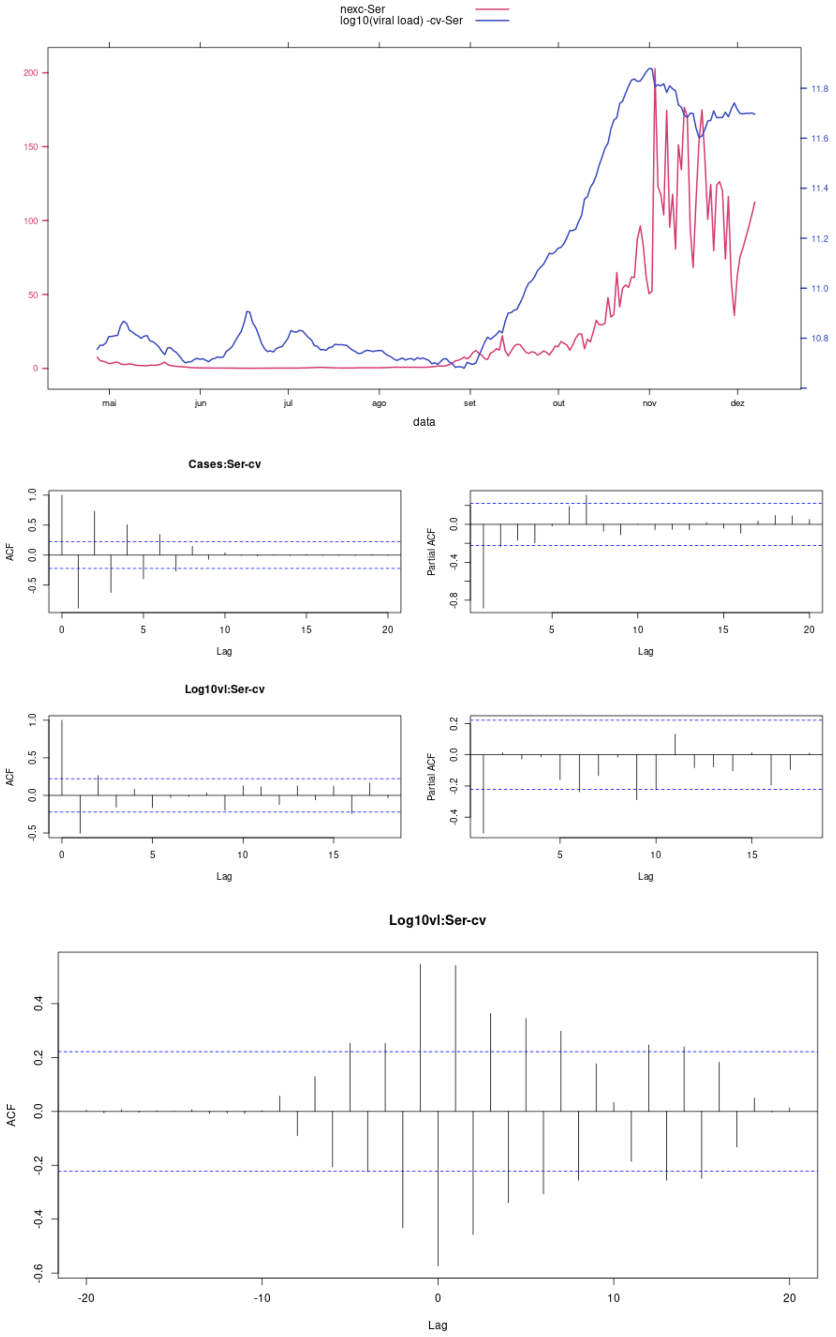
nexc-Ser
log10(viral load) -cv-Ser
200
150
100
50
0
11.8
11.6
11.4
11.2
11.0
10.8
mai
jun
jul
ago
set
out
nov
dez
data
Cases:Ser-cv
ACF
Partial ACF
Lag
Log10vl:Ser-cv
ACF
Partial ACF
Lag
Log10vl:Ser-cv
ACF
Lag
-20
-10
0
10
20

Figure 6

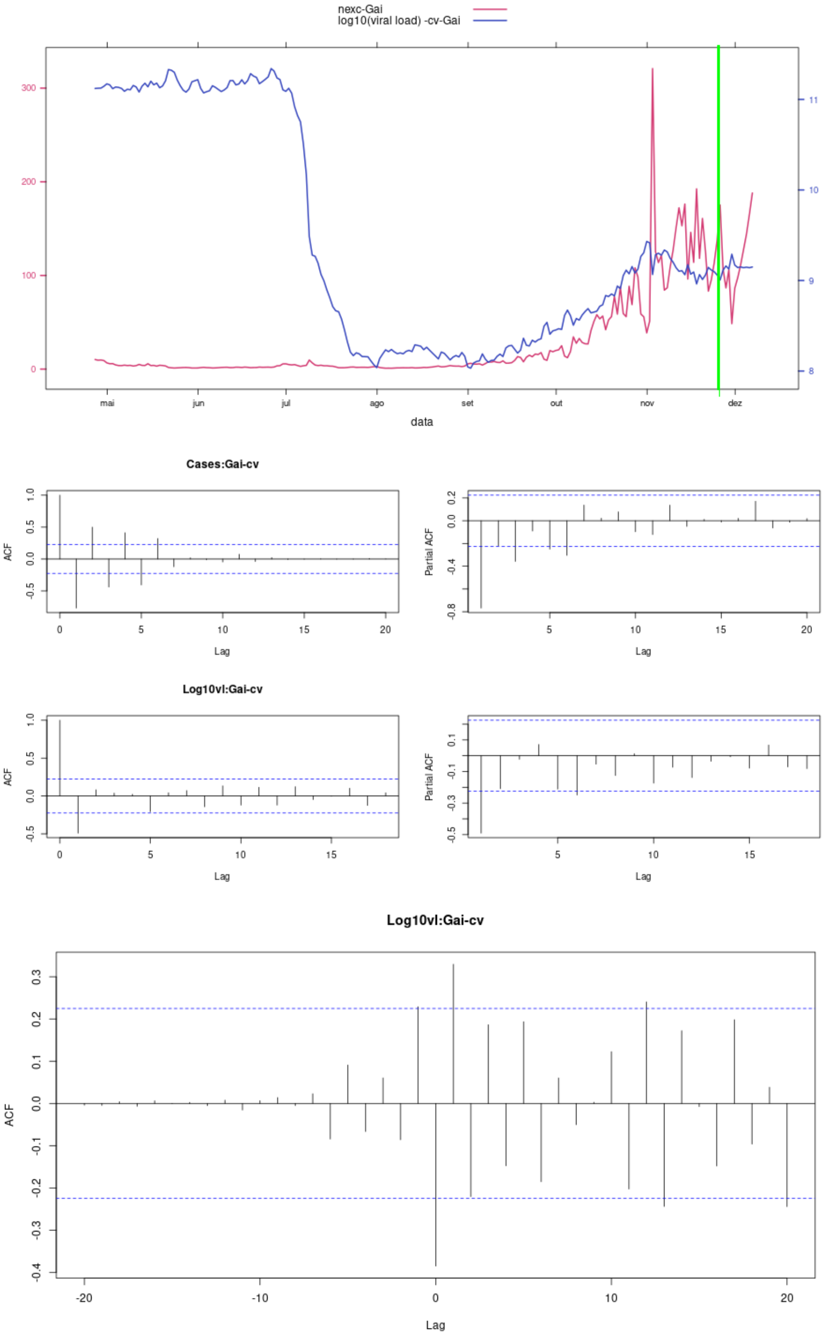
nexc-Gai
log10(viral load) -cv-Gai
300
200
100
0
11
10
9
8
mai
jun
jul
ago
set
out
nov
dez
data
Cases:Gai-cv
ACF
Partial ACF
Lag
Log10vl:Gai-cv
ACF
Partial ACF
Lag
Log10vl:Gai-cv
ACF
Lag
-20
-10
0
10
20

Figure 7

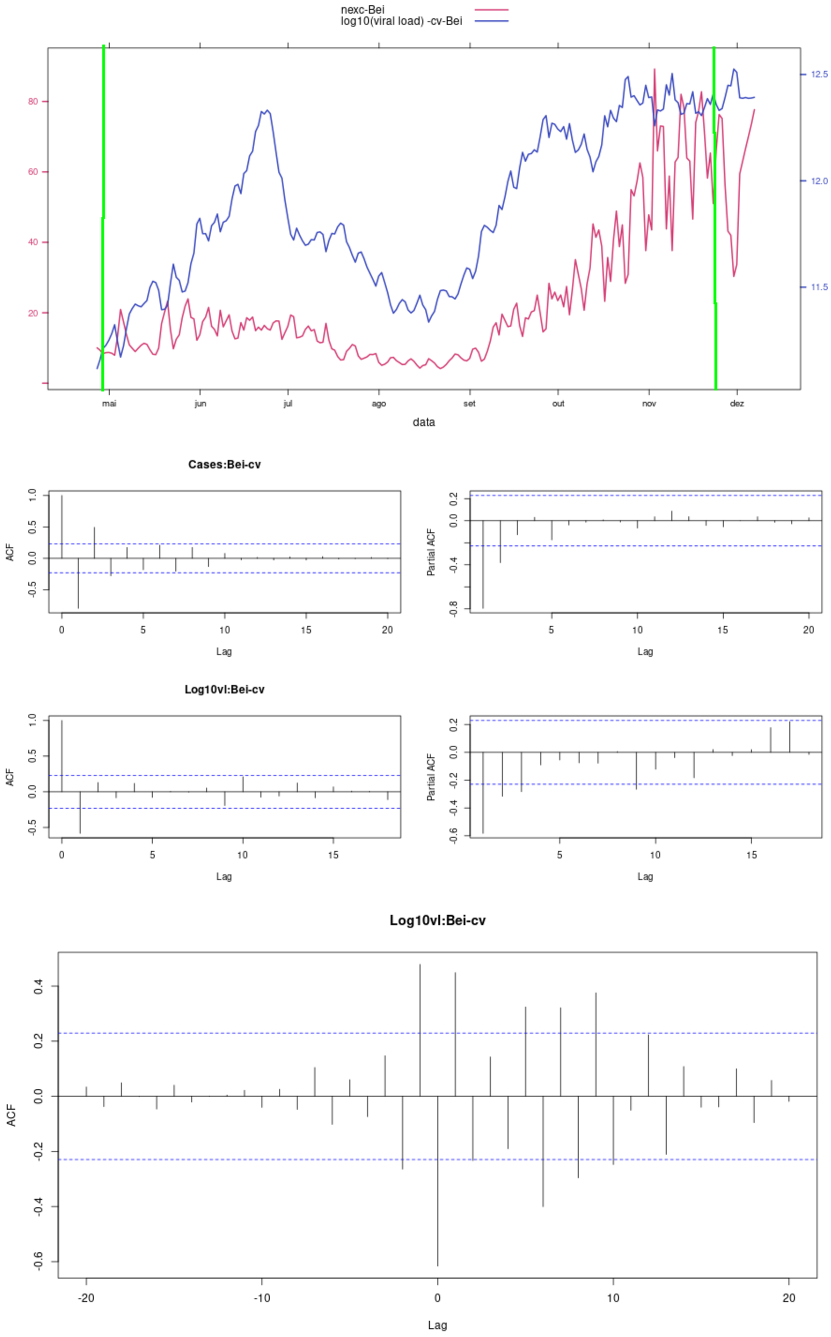
nexc-Bei
log10(viral load) -cv-Bei
80
60
40
20
12.5
12.0
11.5
mai
jun
jul
ago
set
out
nov
dez
data
Cases:Bei-cv
ACF
1.0
0.5
0.0
-0.5
0
5
10
15
20
Lag
Partial ACF
0.2
0.0
-0.2
-0.4
-0.6
-0.8
5
10
15
20
Lag
Log10vl:Bei-cv
ACF
1.0
0.5
0.0
-0.5
0
5
10
15
Lag
Partial ACF
0.2
0.0
-0.2
-0.4
-0.6
5
10
15
Lag
Log10vl:Bei-cv
ACF
0.4
0.2
0.0
-0.2
-0.4
-0.6
-20
-10
0
10
20
Lag

Figure 8

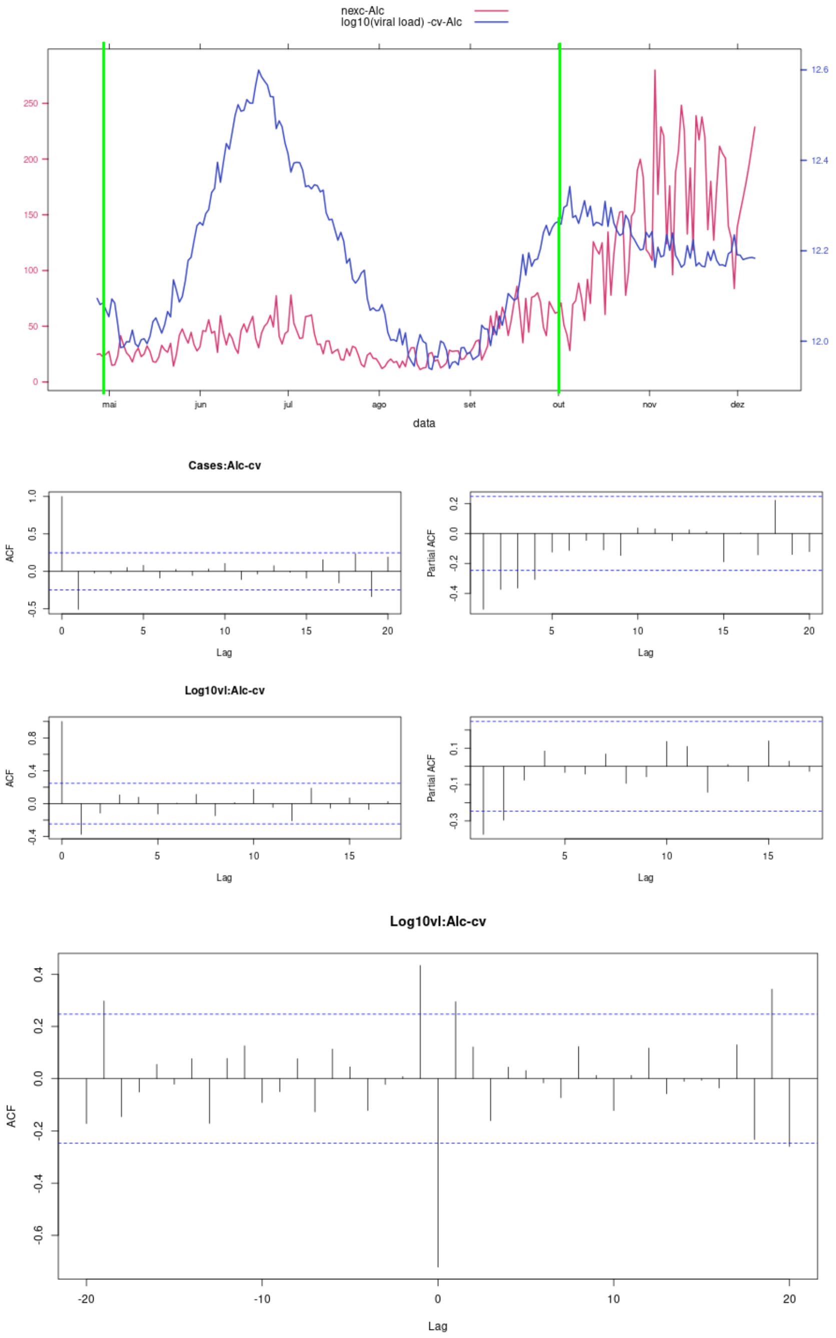
nexc-Alc
log10(viral load) -cv-Alc
0
50
100
150
200
250
12.0
12.2
12.4
12.6
mai
jun
jul
ago
set
out
nov
dez
data
Cases:Alc-cv
ACF
Partial ACF
Lag
Log10vl:Alc-cv
ACF
Partial ACF
Lag
Log10vl:Alc-cv
ACF
Lag
-20
-10
0
10
20

Figure 9

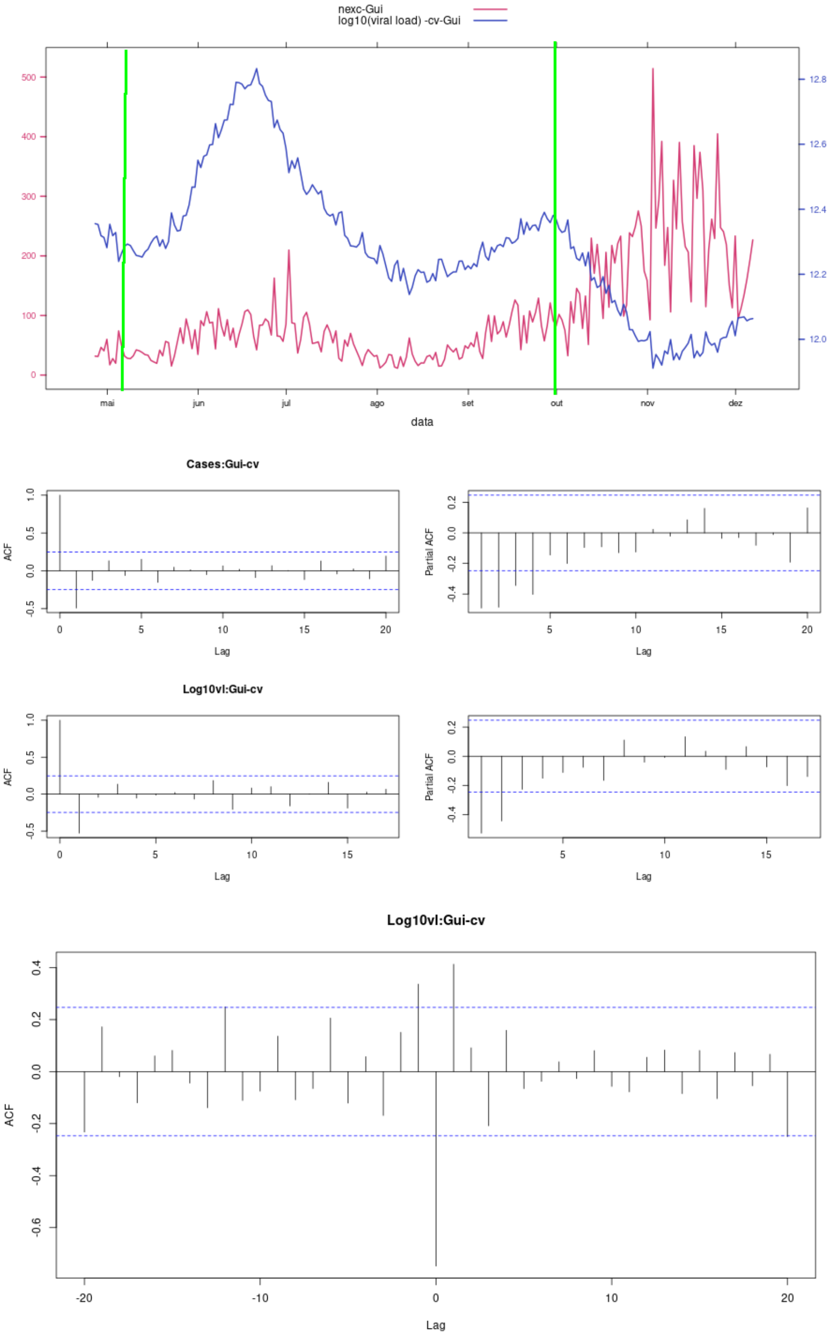
nexc-Gui
log10(viral load) -cv-Gui
500
400
300
200
100
0
12.8
12.6
12.4
12.2
12.0
mai
jun
jul
ago
set
out
nov
dez
data
Cases:Gui-cv
ACF
Partial ACF
Lag
Log10vl:Gui-cv
ACF
Partial ACF
Lag
Log10vl:Gui-cv
ACF
Lag
-20
-10
0
10
20

Figure 10