Austrian Journal of Statistics
2026, Volume 55, 79–104.
http://www.ajs.or.at/
doi:10.17713/ajs.v55i2.1983

Modelling Wastewater Data from Austria Using Generalized Additive Models for Location, Scale and Shape (GAMLSS)

Roman Pfeiler
Institute of Applied Statistics
Linz, Austria

Karin Weyermair
AGES
Graz, Austria

Hans Peter Stüger
AGES
Graz, Austria

Sabrina Kuchling
AGES
Graz, Austria

Patrick Hyden
AGES
Graz, Austria

Helga Wagner
Institute of Applied Statistics
Linz, Austria

Abstract

Generalized Additive Models for Location, Scale and Shape (GAMLSS) are a flexible alternative to standard regression models, where only the mean of the response variable is modelled in terms of covariates. However, GAMLSS currently are seldom used in statistical applications. In this paper, we analyze data from wastewater measurements in Austria with respect to COVID-19. The goal is to model the viral load in the wastewater in terms of covariates. The results show that both the vaccination rate and the dominant virus variant are important covariates in the analysis of COVID-19 related viral load in the wastewater. Moreover, complex GAMLSS based on the four-parametric Box-Cox t distribution clearly outperformed simpler Generalized Additive Models (GAMs) based on the Gamma distribution. Moreover, GAMLSS show also better predictive performance than GAMs.

Keywords: GAMLSS, Gamma distribution, Box-Cox t distribution, wastewater analysis, COVID-19, statistical applications.

1. Introduction

The outbreak of the SARS-CoV-2 virus (or Corona) and the lockdowns influenced the Austrian population in various forms, see, e.g., Aigner, Bacher, Hasengruber, Pfeiler, and Nnebedum (2022); Mayerl, Stolz, and Freidl (2021); Zartler, Dafert, and Dirnberger (2022). In Austria, up until June 2023 the number of positive COVID-19 tests and the incidence rate of COVID-19 were used to assess the number of infected people. The incidence rate is defined as the number of reported cases per 100,000 people in the population over a certain period, usually 7 or 14 days, see, e.g., Statista (2023). An alternative to the incidence based strategy is the analysis of the amount of COVID-19 virus copies in the wastewater, which has the advantage of being independent of the testing intensity in the population. In this paper, longitudinal data of wastewater measurements taken at different Austrian wastewater treatment plants are analyzed with regression models to better understand the development of the viral load over time.

Due to the complexity of the data - the panel is highly unbalanced as wastewater measurements were irregularly obtained for the different treatment plants - standard time series regression models for positive continuous outcomes (Prass, Pumi, Taufemback, and Carlos 2025) cannot be used. Instead, we consider Generalized Additive Models for Location, Scale and Shape (GAMLSS), which were proposed by Rigby and Stasinopoulos (2005) and can be regarded as a natural extension of both the Generalized Linear Model (GLM), see Nelder and Wedderburn (1972), and the Generalized Additive Model (GAM), see Hastie and Tibshirani (1986). GAMLSS allow to model various parameters of the response distribution rather than the conditional mean only, which is useful to model skewness and kurtosis, see, e.g., van Ogtrop, Vervoort, Heller, Stasinopoulos, and Rigby (2011) who modelled streamflow intensity data of semi-arid catchments in South Western Queensland using the right-skewed Box-Cox t distribution. As the unbalanced wastewater trajectories are also non-stationary, we model time non-linearly with splines in a similar vein as van Ogtrop *et al.* (2011) and Villarini, Smith, and Napolitano (2010).

Even though GAMLSS have already been used in applied sciences such as in psychology (Timmerman, Voncken, and Albers 2021; Correa, Kneib, Ospina, Tejada, and Marmolejo-Ramos 2023) and educational data mining (Marmolejo-Ramos, Tejo, Brabec, Kuzilek, Joksimovic, Kovanovic, González, Kneib, Bühlmann, Kook, Briseño-Sánchez, and Ospina 2022), many applied researchers still rely on standard regression models, which only allow for the modelling of the mean (Kneib, Silbersdorff, and Säfken 2023) and, hence, might miss effects on other aspects of the response distribution.

This paper is structured as follows: Section 2 highlights the goal of this study while Section 3 gives an overview of the data and variables used. This includes results from the descriptive analysis as well as information on data quality and management. Section 4 briefly describes the theory and practical aspects of GAMLSS. Section 5 presents the main findings of this study and Section 6 concludes.

2. Study goal

The main goal of this study is to analyze the COVID-19 related viral load found in the waste-water in terms of covariates. In particular, we are interested in (1) the influence of the vaccination rate and (2) the effect of the currently dominant virus variant. Regarding the former, we assume that a higher vaccination coverage in the population leads to a reduction in severe infections and consequently a lower viral load in the wastewater. Concerning the effect of the dominant virus variant, we assume that some variants are more aggressive than others and, thus, are related to a higher viral load.

Aside from inference, we will also investigate the predictive performance of GAMLSS and compare it to that of GAMs in the context of our study. For this, we will evaluate predictions of various models with respect to their out-of sample prediction error.

3. Data and descriptive analysis

We analyze data from $m = 32$ wastewater treatment plants in Austria for which a total of $n = 5180$ wastewater samples were collected between 28.09.2020 and 10.10.2022.
This section gives an overview of the most important variables. The wastewater signal, which is the response variable in this study, is described in Subsection 3.1. Results from the descriptive analyses of the vaccination rates and the dominant virus variant are provided in Subsections 3.2 and 3.3, respectively. Subsection 3.4 contains a brief discussion on additional covariates, which are included in the regression models to control for potential confounding.

3.1. Wastewater signal

The *wastewater signal* variable contains information on the amount of COVID-19 related viral load found in the wastewater. More precisely, it is a normalized value, where normalization is based on the assumption that 11 g of nitrogen (N), 8 g of ammonium (NH_4) and 120 g of chemical oxygen demand (CSB) correspond to the excretion of 1 resident (BML 2022). The set {N,NH_4,CSB} is therefore referred to as the set of *normalization types*. Based on the selected type, the wastewater signal is computed as

$$\begin{aligned}
\text{wastewater.signal}_{\text{N}} &= \frac{11 \times \text{ virus copies } [10^6\#/\text{mL}]}{\text{N } [\text{mg/L}]}, \\
\text{wastewater.signal}_{\text{NH}_4} &= \frac{8 \times \text{ virus copies } [10^6\#/\text{mL}]}{\text{NH}_4 \text{ } [\text{mg/L}]}, \\
\text{wastewater.signal}_{\text{CSB}} &= \frac{120 \times \text{ virus copies } [10^6\#/\text{mL}]}{\text{CSB } [\text{mg/L}]}.
\end{aligned}$$

It is assumed that the normalization types vary in their accuracy (BML 2022). The organic chemical oxygen demand parameter might overestimate the population size. This is especially the case for wastewater treatment plants with a high degree of pollution caused by industry and agriculture. Contrary to the organic CSB parameter, nitrogen-based normalization is primarily based on human and animal excretion and is therefore considered to be a more accurate marker. Compared to ammonium, nitrogen is considered as more robust, since it includes both hydrolyzed and non-hydrolyzed components (BML 2022). Therefore, if multiple normalization types were available for a specific wastewater measurement, nitrogen was preferred over ammonium and ammonium was preferred over chemical oxygen demand. When none of the three markers was available, normalization was based on the population size with respect to the catchment area of the wastewater treatment plant.

The resulting wastewater signal variable is then interpreted as the *viral load in the wastewater per person present in the catchment area of the treatment plant.* Hence, the variable has values on $\mathbb{R}^+$, where higher values correspond to a higher virus concentration in the wastewater.

Figure 1 shows the wastewater signal over time for four exemplary treatment plants. For "Plant 01" measurements start relatively late in time, whereas for "Plant 16" no measurements are available over a long period. Overall, the univariate time series indicate that measurements are obtained irregularly at the different wastewater treatment plants. Time series for all $m = 32$ treatment plants are given in Appendix A.

3.2. Vaccination score

To obtain information on the vaccination rates, data were aggregated at the level of treatment plants. Hence, the vaccination rates of one treatment plant are based on the total number of daily administered doses in all municipalities that are associated with the corresponding plant. Time series of the first three vaccination rates for each treatment plant are shown in Figure 2.

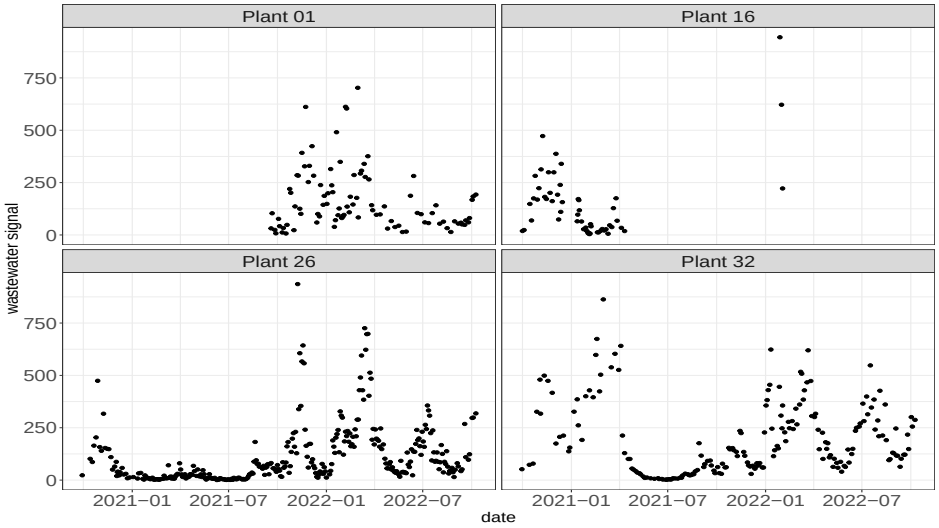

Figure 1: Wastewater signal for selected plants (points are measurements)

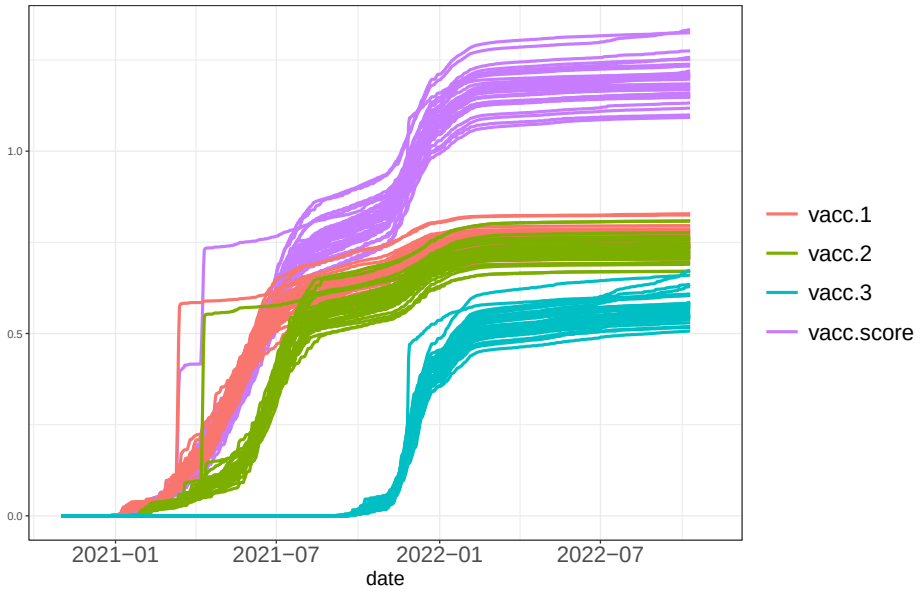

Figure 2: Time series of vaccination rates and the vaccination score for 32 wastewater plants

Since data on the degree of immunization in the population are not available, we apply principal component analysis based on the three single vaccination rates to create a composite *vaccination score*, which is considered as a proxy for immunization. This composite vaccination score is based on the loadings of the first principal component and was computed as

$$\texttt{vacc.score} = 0.6490 \cdot \texttt{vacc.1} + 0.6369 \cdot \texttt{vacc.2} + 0.4162 \cdot \texttt{vacc.3}.$$

The component weights are very similar for the first two vaccination rates and slightly lower for the third. Moreover, the first principal component already accounts for 90.4 % of the total variation among the three vaccination rate variables. The vaccination score is shown over time together with the vaccination rates in Figure 2. The purple lines represent the plant-specific time series of the composite vaccination score. The large increase of the first and second rate for one specific treatment plant is due to a vaccination campaign of the municipality Schwaz. The rates as well as the score are almost constant in 2022 and, hence, do not capture loss of immunization over time.

3.3. Dominant virus variant

Information on the *virus variant* (*Alpha-Beta*, *Delta*, *Omicron-1-2* and *Omicron-4-5*) was obtained through a biological sequence analysis of the wastewater sample. Hence, it was possible to partition the viral load of a single wastewater sample into parts of different variant types (e.g., 30 % of the entire viral load can be attributed to the *Delta* variant etc.). The dominant virus variant is the one, which is associated with the highest percentage from the sequence analysis. There also exist some unclassified variants, as the percentages of the biological sequence analysis do not necessarily add up to 100 %. These unclassified variants were not excluded and instead classified to the residual variant type (*rest/unknown*).

Information on the biological sequence analysis was, however, not available for all wastewater measurements. In order to avoid any information loss, some of these missing data were imputed. Since it is plausible that the dominant virus variant does not change rapidly, missing values were imputed by carrying the value of the last observed value forward, if the last valid observation was at most 14 days before the date of the missing value. However, after imputation there were still 20 % missings in the data set due to a violation of the imputation condition. These missing virus variants were also classified as *rest/unknown* to avoid any information loss.

3.4. Additional covariates

To control for confounding, additional covariates were included in the predictors of the regression models. Those variables are the *federal state*, where the wastewater treatment plant is located, the *laboratory*, which conducted the measurement, the *calendar week*, the *weekday* of the measurement and a *holiday* indicator variable. Due to confidentiality, results regarding the laboratory variable are not reported in this paper.

Table 1 reports the number of plants, laboratories and measurements of each federal state in Austria. There is at least one wastewater treatment plant in each federal state and the number of plants varies from 1 plant in Burgenland and Vienna to 6 in Vorarlberg.

Table 1: Number of plants, laboratories and wastewater measurements by federal state

Federal State	# of Plants	# of Laboratories	# of Measurements
Burgenland	1	0	95
Carinthia	4	1	713
Lower Austria	5	1	686
Salzburg	4	1	591
Styria	3	1	320
Tyrol	5	2	444
Upper Austria	3	1	281
Vienna	1	1	196
Vorarlberg	6	0	1854
$\sum$	32	8	5180

4. Modelling the wastewater signal with GAMLSS

4.1. General aspects of GAMLSS

In a GAMLSS, the random response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is assumed to follow a parametric distribution $\mathcal{D}$, which can be characterized by up to four distributional parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)^\top$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^\top$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$. Each distributional parameter is then modelled by additive predictors, where the single predictors may contain different covariates and smooth terms. Conditional on regressor matrices $\mathbf{X}_1, \ldots, \mathbf{X}_4$ a GAMLSS is then given as

$$\begin{aligned}
\mathbf{Y}|\boldsymbol{\gamma} &\overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}), \\
g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 &\equiv \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{1j}\boldsymbol{\gamma}_{1j}, \\
g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 &\equiv \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{2j}\boldsymbol{\gamma}_{2j}, \\
g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 &\equiv \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{3j}\boldsymbol{\gamma}_{3j}, \\
g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 &\equiv \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{4j}\boldsymbol{\gamma}_{4j}.
\end{aligned} \tag{1}$$

In Equation (1), $g_1(\cdot), \ldots, g_4(\cdot)$ are link functions, and $\mathbf{Z}_{11}, \ldots, \mathbf{Z}_{1J_1}, \ldots, \mathbf{Z}_{4J_4}$ are design matrices and $\boldsymbol{\gamma}_{11}, \ldots, \boldsymbol{\gamma}_{1J_1}, \ldots, \boldsymbol{\gamma}_{4J_4}$ are random effects, e.g., for nonparametrically modelled smooth functions.

When smooth terms are present, estimation of the parameters in Equation (1) is based on maximizing the penalized log-likelihood

$$\ell_{\text{pen}} = \sum_{i=1}^{n} \log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i) - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{kj}^\top \mathbf{T}_{kj}(\boldsymbol{\lambda}_{kj})\boldsymbol{\gamma}_{kj}, \tag{2}$$

where $\boldsymbol{\gamma}_{kj}^\top \mathbf{T}_{kj}\boldsymbol{\gamma}_{kj}$ is a quadratic penalty which is used to prevent overfitting. The penalty matrix $\mathbf{T}_{kj}$ may depend on a set of hyperparameters $\boldsymbol{\lambda}_{kj}$, which are treated as fixed when maximizing Equation (2), see Stasinopoulos, Rigby, Heller, Voudouris, and De Bastiani (2017) for details on estimation algorithms that maximize the penalized log-likelihood.

For model assessment *normalized quantile residuals* are used. These residuals were introduced by Dunn and Smyth (1996) and are defined as

$$r = \Phi^{-1}(u). \tag{3}$$

In Equation (3), $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard Normal distribution and u is a realisation of the random variable $U = F_Y(y)$, where $F_Y(y)$ is the cumulative distribution function of Y. When the GAMLSS is correctly specified, then

$$\hat{r} = \Phi^{-1}(\hat{u}) = \Phi^{-1}[F(y|\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})]$$

approximately follows a standard Normal distribution. The *wormplot* (van Buuren and Fredriks 2001) can be used to assess the adequacy of the model graphically. A wormplot is a detrended QQ-plot of the normalized quantile residuals. The fit is considered as adequate if the 95 % confidence interval contains roughly 95 % of the residuals.

Finally, an important task for fitting a GAMLSS is the selection of covariates in the predictors of the distributional parameters. In contrast to standard regression models, the selection of

covariates in GAMLSS is complicated by the fact that not only one, but up to four predictors might be present. In order to perform variable selection for GAMLSS, we follow Voudouris, Gilchrist, Rigby, Sedgwick, and Stasinopoulos (2012) as well as Stasinopoulos *et al.* (2017) and use their "Method A", where stepwise selection is performed sequentially for the predictors. For a model with four distributional parameters, "Method A" comprises the following seven steps (Stasinopoulos *et al.* 2017):

1. Perform forward selection to select an appropriate model for μ with σ, ν and τ fitted against a constant.

2. Given the selected model for μ use forward selection to fit an appropriate model for σ with ν and τ fitted against a constant.

3. Given the selected models for μ and σ use forward selection to fit an appropriate model for ν with τ fitted against a constant.

4. Given the selected models for μ, σ and ν use forward selection to fit an appropriate model for τ.

5. Given the models for μ, σ and τ use backward selection on the model for ν to fit an appropriate model for ν.

6. Given the models for μ, ν and τ use backward selection on the model for σ to fit an appropriate model for σ.

7. Given the models for σ, ν and τ use backward selection on the model for μ to fit an appropriate model for μ and then **stop**.

For a large number of potential covariates, boosting (Mayr, Fenske, Hofner, Kneib, and Schmid 2012) can be a useful alternative approach.

4.2. Distributions for wastewater modelling

Rigby, Stasinopoulos, Heller, and De Bastiani (2019) discuss over 100 parametric distributions for modelling data using the GAMLSS methodology, which makes the selection of a suitable probability distribution a challenging but important task for applied scientists.

First of all, the support of the distribution should match the support of the response. Hence, when modelling count data obvious candidates are the Poisson distribution and the Negative Binomial distribution, whereas a Gamma distribution is a suitable candidate for response variables with positive support. This support-based strategy already provides a first filter for selecting a distribution. One can then select among the class of suitable distributions a one- or two-parametric distribution and investigate the wormplot of the residuals obtained from this model to assess the adequacy of the distribution. When the fit is not optimal, the model might be too simplistic and, thus, a three- or four-parametric distribution should be considered instead. Stasinopoulos *et al.* (2017) mention that the shape of the worm indicates which distributional parameter (location, scale or shape) requires a more flexible modelling strategy. Table 2 gives an overview on the sources of model misspecification. Hence, when, e.g., a two-parametric Gamma distribution yields a U-shaped wormplot, the skewness of the response variable is not properly modelled.

Since the wastewater signal has restricted support on $\mathbb{R}^+$, positive continuous distributions are selected as suitable candidates for $\mathcal{D}$. A natural choice is the Gamma distribution with density

$$f(y|\mu,\sigma) = \frac{y^{(1/\sigma^2)-1}\exp\left(-\frac{y}{\sigma^2\mu}\right)}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)}, \quad y>0,\ \mu>0,\ \sigma>0, \tag{4}$$

Table 2: Wormplot diagnostics table (based on Stasinopoulos *et al.* (2017))

Feature	Shape of Worm / Curve	Fitted Distribution
level	above origin below origin	location too low location too high
slope	positive negative	scale too low scale too high
U-shape	U-shape inverted U-shape	skewness too low skewness too high
S-shape	left bent down left bent up	tails too light tails too heavy

where $\Gamma(\cdot)$ is the Gamma function. When the Gamma distribution is parameterized as in Equation (4), μ is the mean, σ is the coefficient of variation and $\mu^2\sigma^2$ is the variance of the distribution (Stasinopoulos *et al.* 2017).

As an alternative to the two-parametric Gamma distribution, we consider the more flexible Box-Cox family of distributions. This family is defined by the distribution of the variable Z resulting from a Box-Cox transformation of Y as

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right], & \nu \neq 0, \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right), & \nu = 0, \end{cases}$$

where $\mu > 0, \sigma > 0$ and $\nu \in (-\infty, \infty)$.
For $\nu \neq 0$, $Y > 0$ results when the support of Z is restricted to

$$\begin{aligned} -\frac{1}{\sigma\nu} < Z < \infty, & \qquad \text{if } \nu > 0, \\ -\infty < Z < -\frac{1}{\sigma\nu}, & \qquad \text{if } \nu < 0. \end{aligned}$$

If Z follows a truncated standard Normal random variable, then Y has a Box-Cox-Cole-Green distribution (Cole and Green 1992), i.e. $Y \sim \text{BCCG}(\mu, \sigma, \nu)$. If Z follows a truncated Student t distribution with $\tau > 0$ degrees of freedom, then Y has a Box-Cox t distribution (Rigby and Stasinopoulos 2006), i.e. $Y \sim \text{BCT}(\mu, \sigma, \nu, \tau)$.

Compared to other three- or four-parametric distributions, the parameters of the BCCG and the BCT distribution are easy to interpret: μ is the median, σ is an approximate centile-based coefficient of variation and ν is a skewness parameter. Furthermore, the parameter τ of the BCT distribution is a kurtosis parameter. More details on the various GAMLSS distributions are given in Rigby *et al.* (2019). For the application of GAMLSS and their implementation in `R` see Stasinopoulos *et al.* (2017) and Stasinopoulos and Rigby (2007).

5. Results

In this section we discuss results of the fitted regression models. In Subsection 5.1 we give a brief overview on the fitted models and in Subsection 5.2 we show the results with an emphasis on model assessment and interpretation. Finally, Subsection 5.3 compares the models with respect to their predictive accuracy. Data analysis was conducted using the open-source statistical software `R` (R Core Team 2024).

5.1. Fitted models

The variables time (or calendar week) and vaccination score are modelled using P-splines (Eilers and Marx 1996). Categorical predictors are dummy coded with baseline values *Delta* (dominant virus variant), *Vienna* (federal state), *Sunday* (weekday), *no holiday* (holiday indicator) and *nitrogen* (normalization types).

Model 1 is a simple Gamma model, where the mean depends on covariates. Model 2 is a double Gamma model, where both parameters of the Gamma distribution are modelled in terms of covariates. Model 3 is based on the four-parametric Box-Cox t (BCT) distribution. We also tried the three-parametric Box-Cox-Cole-Green (BCCG) distribution, but this resulted in convergence problems, which were not present with the BCT model. This is most likely due to the robustness of the t-distribution. Models 1 to 3 contain a random intercept only in the predictor of μ. To account for plant-specific heterogeneity in the scale parameter, Model 4 is a BCT model, which contains a random intercept not only in the predictor of μ, but also in the predictor of σ. As the trajectories of wastewater signal differ considerably between plants, this would suggest to include an interaction effect of time and plant in the predictor of μ. However, as there are not enough measurements available for all wastewater treatment plants to estimate such an effect, we modelled P-splines that vary with federal state instead of treatment plant. Moreover, as noted by a referee, a random intercept might not be adequate as the implied correlation between measurements of one treatment plant would be the same irrespective of the time lag. We therefore considered further models where the interaction of federal state and time in μ is modelled by a penalized varying coefficient: Thus, Model 5 extends Model 4 and Model 6 additionally includes a random slope in the predictors of μ and σ. Models 7 and 8 are the same as Models 5 and 6 without random effects for σ. Table 3 gives an overview on the eight regression models that are discussed in this paper.

Table 3: Overview on regression models for the wastewater signal

	Model Type	Random Intercept	Random Slope	Time*State Interaction
Model 1	simple Gamma	X		
Model 2	double Gamma	X		
Model 3	four parametric BCT	X		
Model 4	four parametric BCT	X		
Model 5	four parametric BCT	X		X
Model 6	four parametric BCT	X	X	X
Model 7	four parametric BCT	X		X
Model 8	four parametric BCT	X	X	X

We do not consider an explicit spatial term for modelling the wastewater signal as the inclusion of the federal state variable together with random intercepts in the predictors should be sufficient to capture spatial dependence. Variable selection was based on "Method A".

5.2. Results of the fitted models

Table 4 gives an overview of the covariate effects and smooth terms selected by variable selection and reports Cox-Snell's R^2 as well as BIC for each model. The table provides information on whether a covariate or a random effect was included in a predictor after the selection process. (-) means that the covariate (or random effect) was not included in any of the predictors after selection, whereas 'NA' means that the covariate (or random effect) was not considered for this model. The simple Gamma model (Model 1), the double Gamma model (Model 2) and the BCT-1 model (Model 3) contain a random intercept only in the predictor of μ, whereas the BCT-2 model (Model 4) contains a random intercept in the predictors of both μ and σ. In the BCT-3 model (Model 5), an interaction between calendar week and federal state was added in the predictor of μ, whereas in the BCT-4 model (Model

6) a random slope was included in the predictors of μ as well as σ. Models BCT-5 and BCT-6 (Models 7 and 8) only include random effects for μ and hence are extensions of Model 3.

Table 4: Details on models after variable selection

covariate	Model 1 SG	Model 2 DG	Model 3 BCT-1	Model 4 BCT-2	Model 5 BCT-3	Model 6 BCT-4	Model 7 BCT-5	Model 8 BCT-6
vacc.score	μ	μ	μ	μ	μ	μ	μ	μ, ν
variant	μ	μ, σ	μ, σ, ν	μ, σ	μ, σ	μ	σ, ν	μ, σ, ν
time	μ	μ, σ	μ, σ, τ	μ, σ, ν	μ, σ, ν	μ, σ	μ, σ	μ, σ
state	(-)	σ	σ	(-)	(-)	(-)	σ	σ
*time*state*	NA	NA	NA	NA	μ	μ	μ	μ
laboratory	μ	μ, σ	μ, σ, ν	μ, σ, ν	μ, σ, ν	μ, σ, ν	μ, σ	μ, σ, ν
normaliz.	μ	μ	μ	μ	μ	μ, ν	μ, σ	μ, σ
weekday	μ	μ	μ, τ	μ, τ	μ, τ	μ, σ	μ, σ	μ, σ
holiday	μ	μ	μ	μ	μ	μ	μ	μ
rand.interc.	μ	μ	μ	μ, σ	μ, σ	μ, σ	μ	μ
rand.slope	NA	NA	NA	NA	NA	μ, σ	NA	μ
BIC	56219.6	55450.0	55396.8	55408.0	54968.2	55082.1	55021.3	55017.2
R^2	0.71	0.77	0.78	0.78	0.83	0.83	0.82	0.83

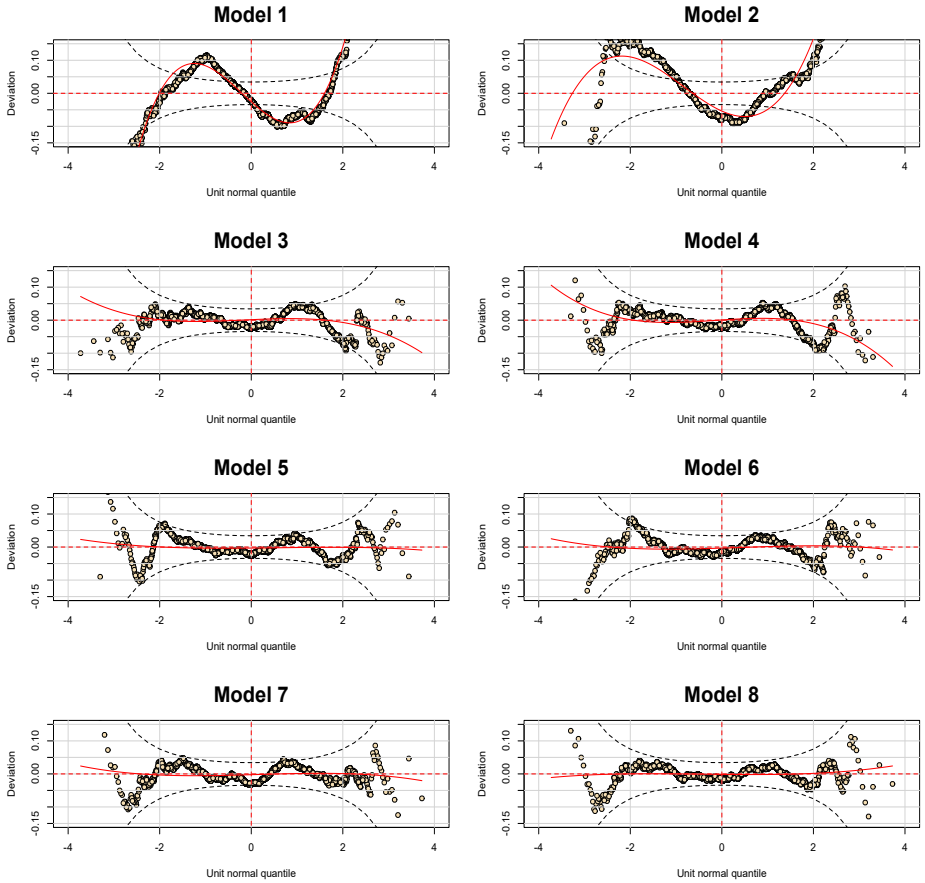

Figure 3: Wormplots of the eight regression models

For both the simple Gamma model with constant scale parameter (Model 1) as well as the double Gamma model (Model 2), wormplots are S-shaped whereas the wormplots for the six BCT models basically suggest an appropriate fit (see Figure 3). The best model in terms of BIC is Model 5, which contains an interaction in the predictor of μ and random intercepts

in the predictors of μ and σ. The main results of this model ae presented in the following. Figures 4 and 5 show covariate effect plots for the location parameter μ, i.e. the median and the scale parameter σ, i.e. the approximate centile-based coefficient of variation. The effect plots for the skewness parameter ν and kurtosis parameter τ of Model 5 are given in Appendix B.5.

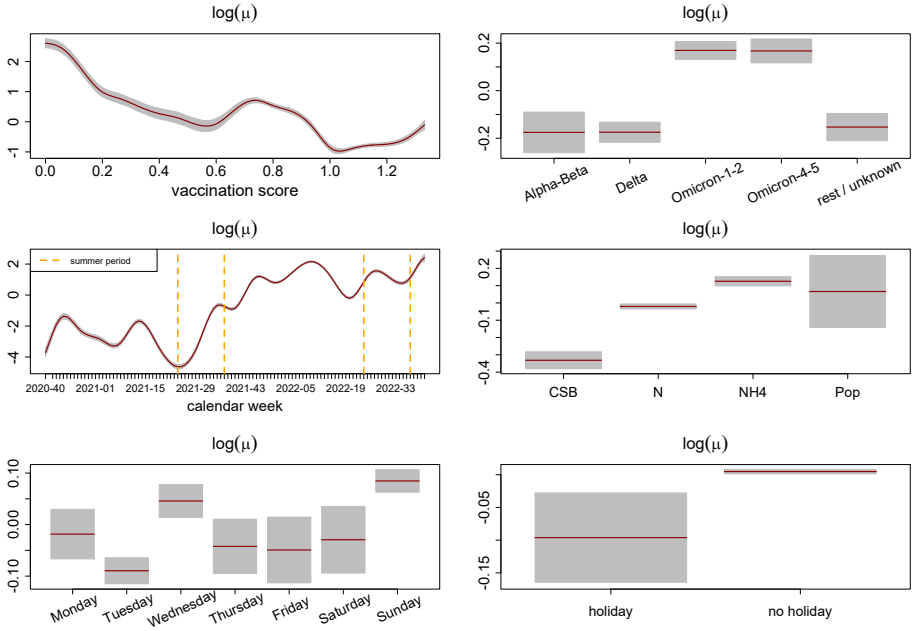

Figure 4: Covariate effect plots of Model 5 for $\log(\mu)$

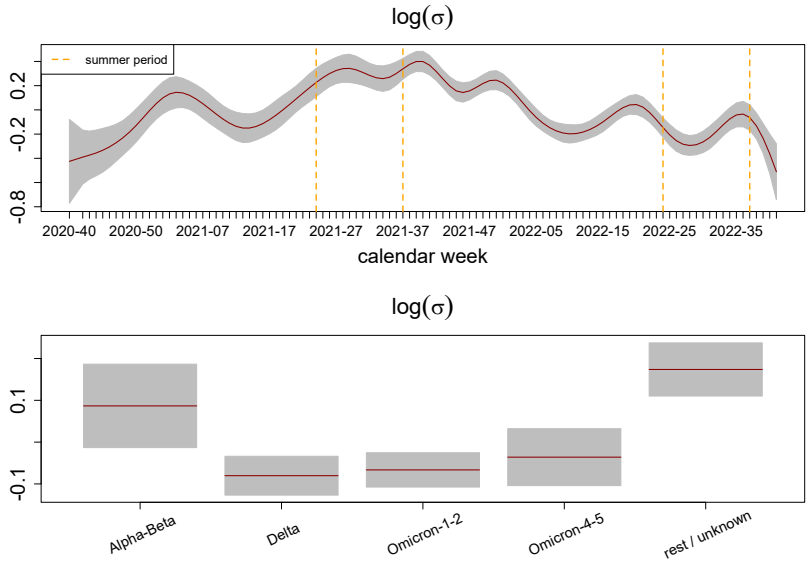

Figure 5: Covariate effect plots of Model 5 for $\log(\sigma)$

The effect of the composite vaccination score on μ decreases nonlinearly over time suggesting that a higher vaccination coverage corresponds to a lower COVID-19 signal in the wastewater. This result is consistent among all models except for Model 6, where the effect is highly nonlinear (see Appendix B.6), and for Model 7, where a positive linear relationship was estimated (see Appendix B.7). The somewhat surprising linear effect in Model 7 is due to the fact that the dominant virus variant was excluded from the predictor of μ during variable selection. If the dominant virus variant is included in the predictor of μ, then the effect of the

vaccination score is nonlinear decreasing and similar to the corresponding effect of the other models.

The *Omicron* subvariants are related to a higher excretion in the wastewater than both the *Alpha-Beta* and *Delta* types. There is, however, no difference among *Omicron* subvariants. The estimated effects of the dominant virus variant differ considerably between the models: In the double Gamma model (Model 2) and the best main effects BCT model (Model 3), the *Delta* type is related to a higher wastewater signal compared to the *Alpha-Beta* type and the *Omicron* subvariants are less similar, whereas in the simple Gamma model (Model 1), there is no substantial difference in the estimated effects of *Delta*, *Omicron-1-2* and *Omicron-4-5*. Finally, the effects of the virus variant in Models 6 and 8 are relatively similar to the ones estimated by Model 5, i.e. the best model (see the corresponding figures in the Appendix).

CSB and nitrogen are related to a lower wastewater signal than ammonium and population based normalization. Only 0.48 % of the measurements are related to population based normalization and, thus, the respective confidence interval is rather wide.

The wastewater signal is considerably lower on holidays and tends to be higher on Wednesday and Sunday, but the weekday effect is small.

The effect of time is nonlinear and increases in the second half of the observation period.

Concerning the effects on the scale parameter σ in Model 5, the effect of time is clearly nonlinear with higher values in 2021 than in 2022. The *Alpha-Beta* and *rest/unknown* types are related to higher variability in wastewater measurements than *Delta* and *Omicron* types. The relatively high variation for the *rest/unknown* type is reasonable as many different variant types are included in this group.

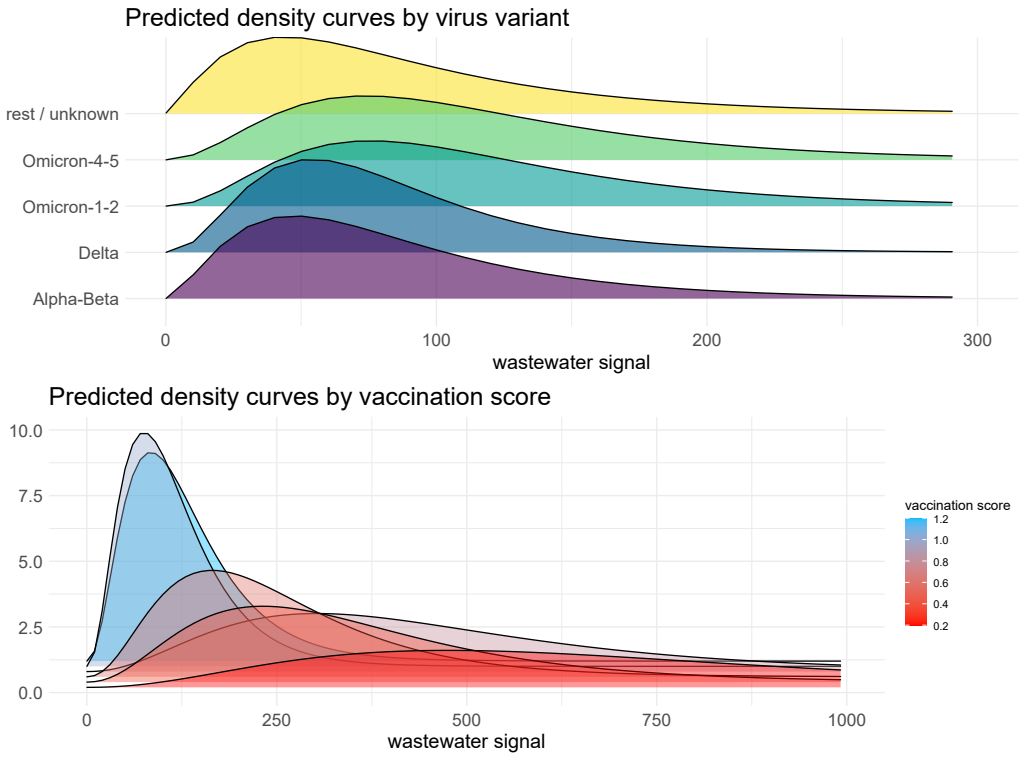

Figure 6: Predicted densities of Model 5. *The figures show partial effects. All the other covariates are fixed at either their median (continuous covariates) or their most frequent value (categorical covariates), i.e. Vorarlberg (federal state), CSB (normalization type), Omicron-1-2 (dominant virus variant), Sunday (weekday) and no holiday (holiday indicator).*

As some covariates enter multiple predictors in the BCT model, the effects of covariates should also be assessed jointly by plotting the predicted density curves for different covariate values. Figure 6 shows the predicted density of the wastewater signal for categories of the vaccination

score as well as the dominant virus variant while holding the other covariates constant. As the vaccination score is a quantitative variable, the predicted density of the wastewater signal is shown for a set of equidistant values, i.e., $\{0.20, 0.40, 0.60, 0.80, 1.00, 1.20\}$. The results support the conclusions based on the single covariate effect plots. The distribution of the wastewater signal is more concentrated around smaller values when more people have received vaccination. The mode is shifted towards higher signal values when the *Omicron* subvariants are the dominant virus variant.

Figure 7 shows the effect of calendar week on the location parameter μ with federal state used as effect modifier. Note that the plots do not contain the main effects but show only the contributions of the interaction term. There are no measurements available for the plants in Burgenland, Styria and Tyrol in the first half of the observation period. In Burgenland, Salzburg, Styria and Tyrol, the main effect of time already accounts for most variation in the wastewater measurements. The interaction structure is relativley similar in the remaining states, i.e. a higher degree of nonlinearity at the first half of the observation period compared to the second half.

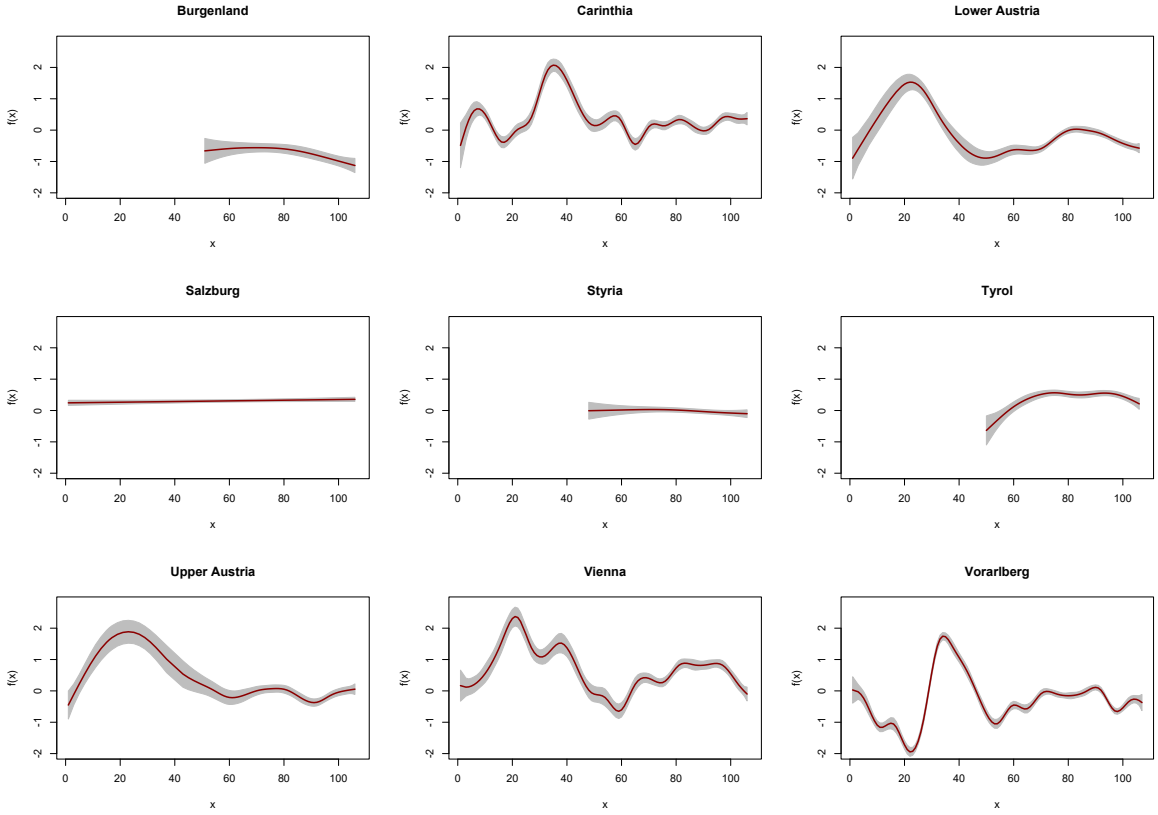

Figure 7: Penalized smoothing of calendar week modified by federal state (Model 5)

5.3. Predictive modelling with GAMLSS

To assess the predictive performance of the models, we split the original data into a training and a test set: the training set consists of all measurements from September 28, 2020 until August 31, 2022 and data reported from September to October, 2022 are used as test set. Hence, the training data consists of $n_{\text{train}} = 4877$ observations, and the test set comprises $n_{\text{test}} = 303$ observations.

For our prediction analysis, we consider the covariate values as known and, thus, predict the wastewater signal based on their values. As point predictions we use the (conditional) mean and the (conditional) median. We assess the performance of the model with respect to the

criterion for which these values are optimal, i.e. the root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(y_i^{\text{true}} - y_i^{\text{predicted}} \right)^2}$$

for which the mean is optimal, as well as the mean absolute error (MAE)

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| y_i^{\text{true}} - y_i^{\text{predicted}} \right|$$

for which the median is optimal. As the mean of the BCT distribution is not of closed form, we use a Monte Carlo approach and estimate it via simulation.

The results are given in Table 5. Overall, the BCT models are performing better than the two models based on the Gamma distribution. Furthermore, the BCT model without interactions and only a random intercept in the predictor of μ (Model 3) gives the lowest RMSE, whereas the BCT model with interactions as well as random intercepts in the predictor of μ (Model 7) has lowest MAE.

Table 5: Evaluation of prediction models

	Model Type	RMSE	MAE
Model 1	Simple Gamma	136.61	97.60
Model 2	Double Gamma	130.77	94.63
Model 3	BCT-1	112.94	87.73
Model 4	BCT-2	115.84	87.77
Model 5	BCT-3	127.22	85.79
Model 6	BCT-4	127.69	85.37
Model 7	BCT-5	120.25	82.80
Model 8	BCT-6	125.81	84.89

6. Concluding remarks

This study used GAMLSS to analyze the COVID-19 related viral load in the wastewater that was measured at 32 treatment plants across Austria. The best model in terms of BIC was a GAMLSS based on the four-parametric Box-Cox t distribution that included an interaction term of time and federal state (as an aggregated proxy for treatment plant) in the predictor of the median wastewater signal μ as well as random intercepts in the predictors of μ and σ. According to the estimated effects, a higher vaccination score was related to a lower wastewater signal and the *Omicron* subvariants to a higher wastewater signal than the *Delta* type.

Finally, the predictions obtained from GAMLSS were better than the ones obtained from the GAM based on the Gamma distribution.

A shortcoming of the analysis is the irregular spacing of the wastewater measurements over time with large gaps between measurements for some wastewater treatment plants. Also for some treatment plants, wastewater measurements were recorded very late in time, which makes general conclusions for these plants over the entire study period difficult. Another limitation is that the vaccination score is only a crude proxy-variable for the degree of immunization and that actual data on the immunization status in the society was not available for our study. Especially since the score is constant in 2022, it might not capture the development of the immunization status appropriately. Finally, we used federal state as a proxy for treatment plant when modelling the plant-time interaction. This was due to the small number of measurements for some plants.

Despite those limitations, this study still demonstrated the potential of GAMLSS in the context of wastewater analysis. Aside from generally higher predictive accuracy, GAMLSS were useful for modelling the wastewater signal as the Box-Cox t distribution is flexible enough to describe the highly skewed response variable. In general, GAMLSS are more likely to outperform standard regression models when the dispersion, skewness or kurtosis are functions of covariates. Therefore, GAMLSS is a useful toolkit for practicioners, but in order to make GAMLSS more attractive, hands-on applications in all scientific fields, where this methodology could be of relevance, are needed.

References

Aigner P, Bacher J, Hasengruber K, Pfeiler R, Nnebedum C (2022). "Familien mit Migrationshintergrund und die Herausforderungen des Coronabedingten Homeschoolings: Eine Typologie Unterschiedlicher Bewältigungsstrategien zwischen Innovativem Krisenmanagement und Dysfunktionalität." *SWS-Rundschau*, **62**(4), 449–470.

BML (2022). "Coron-A - Nachweis und Überwachung von SARS-CoV-2 Infektionen in Österreichs Bevölkerung mittels Abwasseranalysen." URL https://info.bml.gv.at/service/publikationen/wasser/coron-a.html.

Cole TJ, Green PJ (1992). "Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood." *Statistics in Medicine*, **11**(10), 1305–1319. doi:10.1002/sim.4780111005.

Correa JC, Kneib T, Ospina R, Tejada J, Marmolejo-Ramos F (2023). "Assessing Potential Heteroscedasticity in Psychological Data: A GAMLSS Approach." *The Quantitative Methods for Psychology*, **19**(4), 333–346. doi:10.20982/tqmp.19.4.p333.

Dowle M, Srinivasan A (2023). *data.table: Extension of 'data.frame'*. R package version 1.14.8, URL https://CRAN.R-project.org/package=data.table.

Dunn PK, Smyth GK (1996). "Randomized Quantile Residuals." *Journal of Computational and Graphical Statistics*, **5**(3), 236–244. doi:10.2307/1390802.

Eilers PHC, Marx BD (1996). "Flexible Smoothing with B-splines and Penalties." *Statistical Science*, **11**(2), 89–121. doi:10.1214/ss/1038425655.

Hastie T, Tibshirani R (1986). "Generalized Additive Models." *Statistical Science*, **1**(3), 297–310. doi:10.21236/ada147454.

Kneib T, Silbersdorff A, Säfken B (2023). "Rage Against the Mean - A Review of Distributional Regression Approaches." *Econometrics and Statistics*, **26**, 99–123. doi:10.1016/j.ecosta.2021.07.006.

Marmolejo-Ramos F, Tejo M, Brabec M, Kuzilek J, Joksimovic S, Kovanovic V, González J, Kneib T, Bühlmann P, Kook L, Briseño-Sánchez G, Ospina R (2022). "Distributional Regression Modeling via Generalized Additive Models for Location, Scale and Shape: An Overview through a Data Set from Learning Analytics." *WIREs Data Mining and Knowledge Discovery*, **13**(1). doi:10.1002/widm.1479.

Mayerl H, Stolz E, Freidl W (2021). "Longitudinal Effects of COVID-19-Related Loneliness on Symptoms of Mental Distress among Older Adults in Austria." *Public Health*, **200**, 56–58. doi:10.1016/j.puhe.2021.09.009.

Mayr A, Fenske N, Hofner B, Kneib T, Schmid M (2012). "Generalized Additive Models for Location, Scale and Shape for High Dimensional Data - A Flexible Approach Based on Boosting." *Journal of the Royal Statistical Society Series C*, **61**(3), 403–427. doi:10.1111/j.1467-9876.2011.01033.x.

Nelder JA, Wedderburn RWM (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A*, **135**(3), 370–384. doi:10.2307/2344614.

Prass TS, Pumi G, Taufemback CG, Carlos JH (2025). "Positive Time Series Regression Models: Theoretical and Computational Aspects." *Computational Statistics*, **40**, 1–31. doi:10.1007/s00180-024-01531-z.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rigby RA, Stasinopoulos MD (2005). "Generalized Additive Models for Location, Scale and Shape." *Journal of the Royal Statistical Society, Series C*, **54**(3), 507–554. doi:10.1111/j.1467-9876.2005.00510.x.

Rigby RA, Stasinopoulos MD (2006). "Using the Box-Cox t Distribution in GAMLSS to Model Skewness and Kurtosis." *Statistical Modelling*, **6**(3), 209–229. doi:10.1191/1471082x06st122oa.

Rigby RA, Stasinopoulos MD, Heller GZ, De Bastiani F (2019). *Distributions for Modeling Location, Scale and Shape - Using GAMLSS in R*. CRC Press, Boca Raton, London, New York. doi:10.1201/9780429298547.

Sarkar D (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. URL http://lmdvr.r-forge.r-project.org.

Stasinopoulos MD, Rigby RA (2007). "Generalized Additive Models for Location, Scale and Shape (GAMLSS) in R." *Journal of Statistical Software*, **23**(7), 1–46. doi:10.18637/jss.v023.i07.

Stasinopoulos MD, Rigby RA (2023). *gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape*. R package version 6.1-1, URL https://CRAN.R-project.org/package=gamlss.dist.

Stasinopoulos MD, Rigby RA, De Bastiani F (2024). *gamlss.ggplots: Plotting Functions for Generalized Additive Model for Location Scale and Shape*. R package version 2.1-12, URL https://CRAN.R-project.org/package=gamlss.ggplots.

Stasinopoulos MD, Rigby RA, Heller GZ, Voudouris V, De Bastiani F (2017). *Flexible Regression and Smoothing - Using GAMLSS in R*. CRC Press, Boca Raton, London, New York. doi:10.1201/b21973.

Statista (2023). "Incidence of Coronavirus (COVID-19) Cases in the Past Seven Days in Europe as of March 13, 2023, by Country." URL https://www.statista.com/statistics/1139048/coronavirus-case-rates-in-the-past-7-days-in-europe-by-country/.

Timmerman ME, Voncken L, Albers CJ (2021). "A Tutorial on Regression-Based Norming of Psychological Tests with GAMLSS." *Psychological Methods*, **26**, 357–373. doi:10.1037/met0000348.

van Buuren S, Fredriks M (2001). "Worm Plot: A Simple Diagnostic Device for Modelling Growth Reference Curves." *Statistics in Medicine*, **20**(8), 1259–1277. doi:10.1002/sim.746.

van Ogtrop FF, Vervoort RW, Heller GZ, Stasinopoulos MD, Rigby RA (2011). "Long-Range Forecasting of Intermittent Streamflow." *Hydrology and Earth System Sciences*, **15**(11), 3343–3354. doi:10.5194/hess-15-3343-2011.

Villarini G, Smith JA, Napolitano F (2010). "Nonstationary Modeling of a Long Record of Rainfall and Temperature over Rome." *Advances in Water Resources*, **33**(10), 1256–1267. doi:10.1016/j.advwatres.2010.03.013.

Voudouris V, Gilchrist R, Rigby RA, Sedgwick J, Stasinopoulos MD (2012). "Modelling Skewness and Kurtosis with the BCPE Density in GAMLSS." *Journal of Applied Statistics*, **39**(6), 1279–1293. doi:10.1080/02664763.2011.644530.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.3, URL https://CRAN.R-project.org/package=dplyr.

Wickham W (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Zartler U, Dafert V, Dirnberger P (2022). "What Will the Coronavirus Do to Our Kids? Parents in Austria Dealing with the Effects of the COVID-19 Pandemic on Their Children." *Journal of Family Research*, **34**(1), 367–393. doi:10.20377/jfr-713.

A. Development of wastewater signal by treatment plant

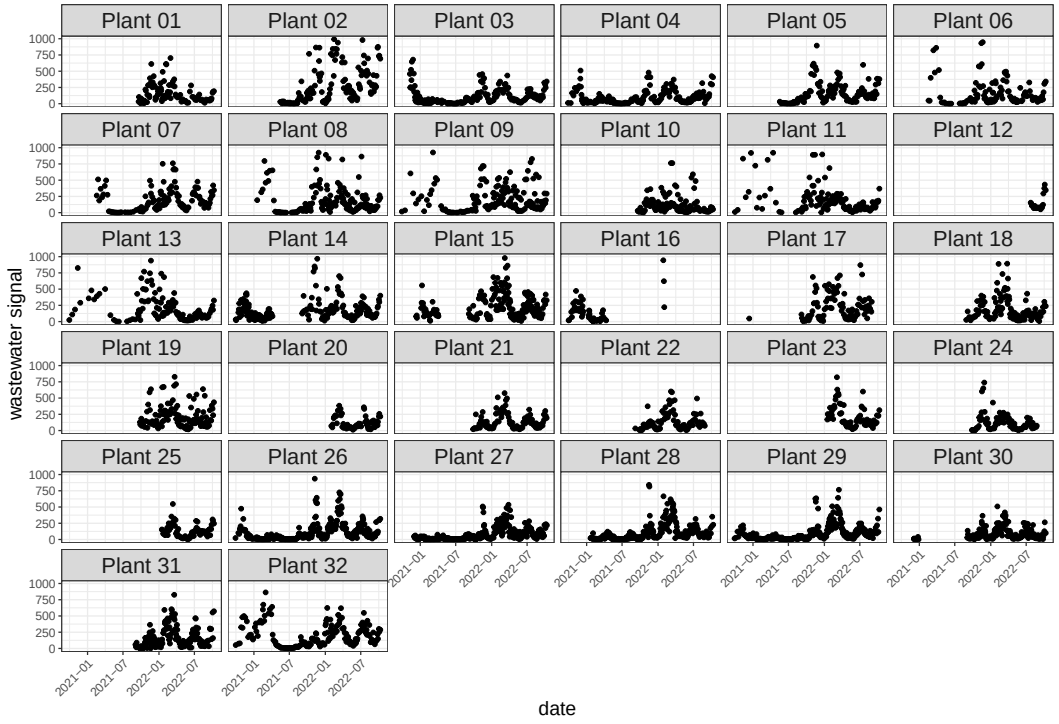

Figure 8: Time series of wastewater signal for all $m = 32$ treatment plants

B. Additional results

B.1. Effect plots of Model 1 (SG)

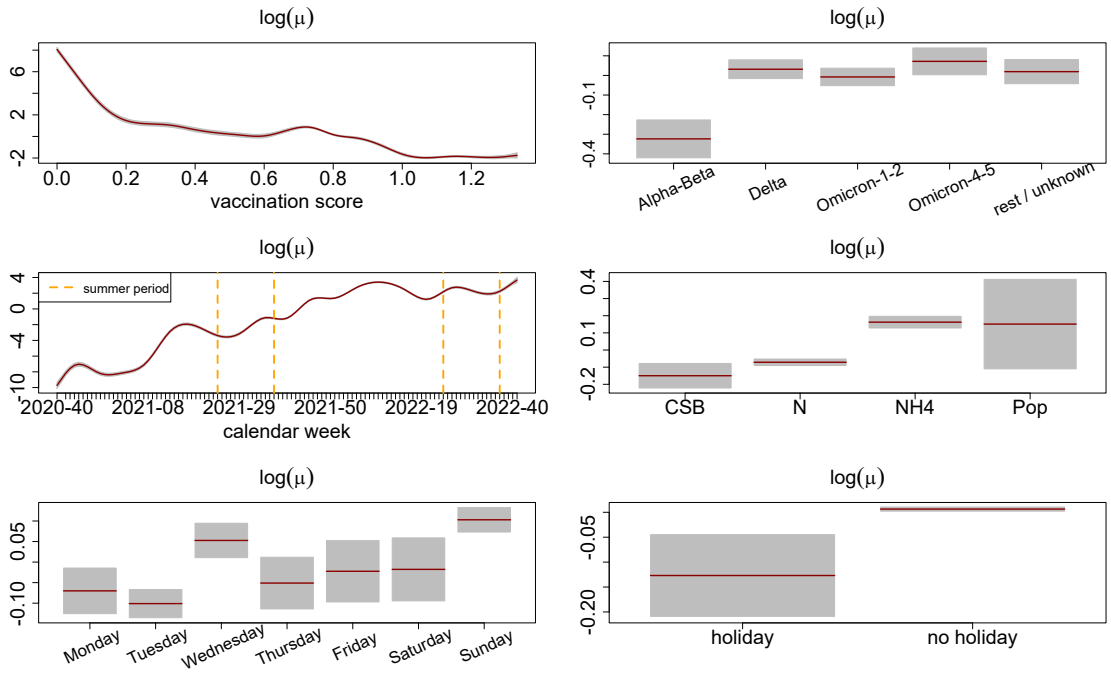

Figure 9: Covariate effect plots of Model 1 for $\log(\mu)$

B.2. Effect plots of Model 2 (DG)

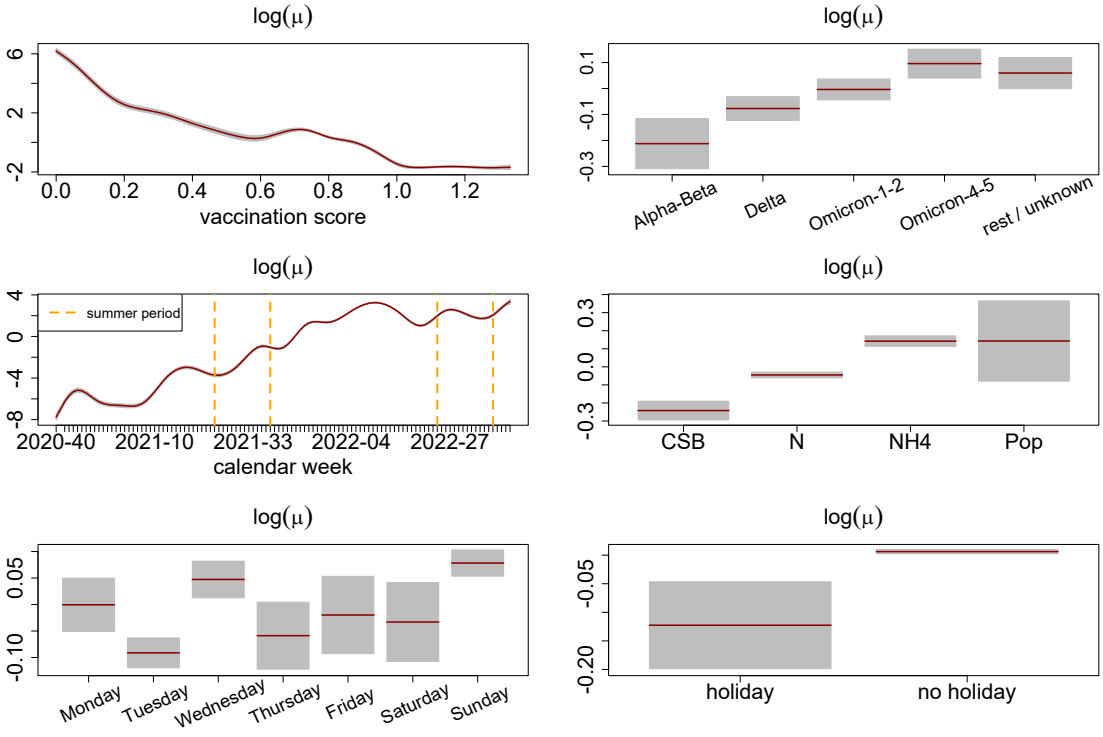

Figure 10: Covariate effect plots of Model 2 for $\log(\mu)$

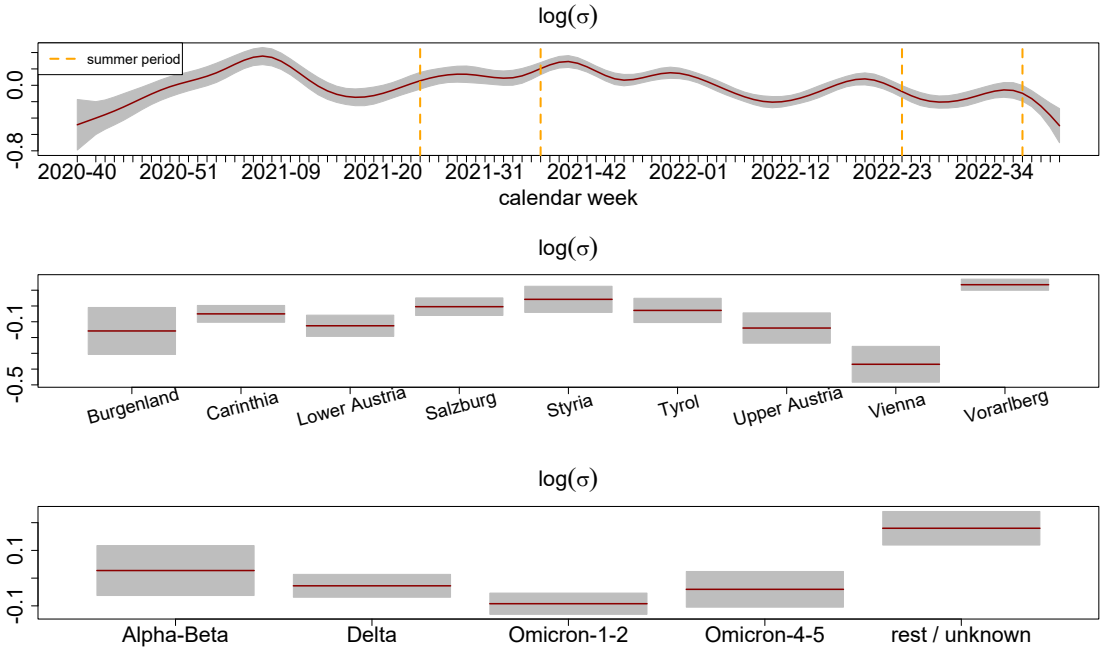

Figure 11: Covariate effect plots of Model 2 for $\log(\sigma)$

B.3. Effect plots of Model 3 (BCT-1)

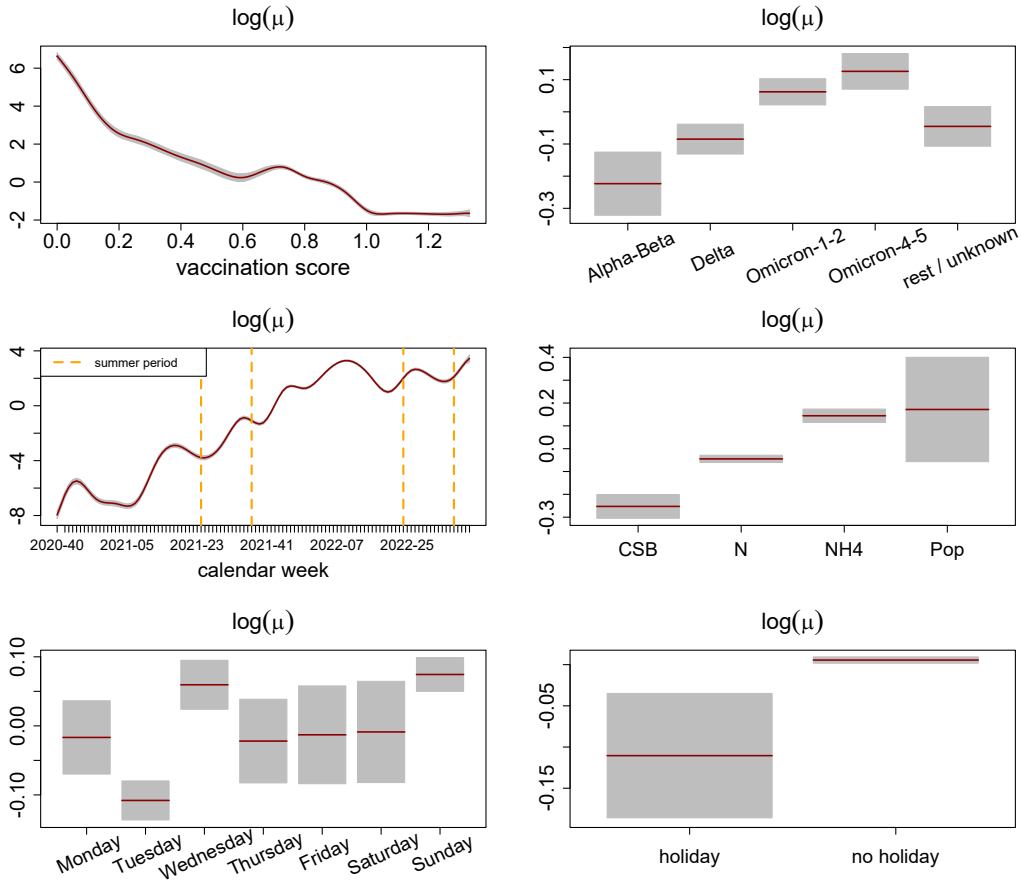

Figure 12: Covariate effect plots of Model 3 for $\log(\mu)$

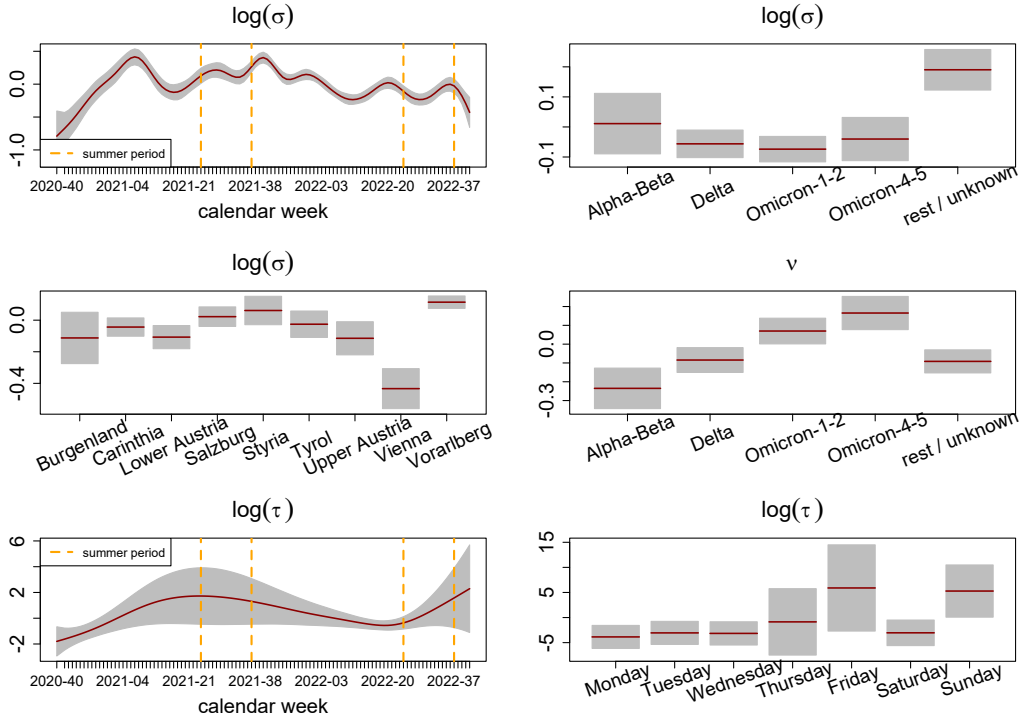

Figure 13: Covariate effect plots of Model 3 for $\log(\sigma), \nu$ and $\log(\tau)$

B.4. Effect plots of Model 4 (BCT-2)

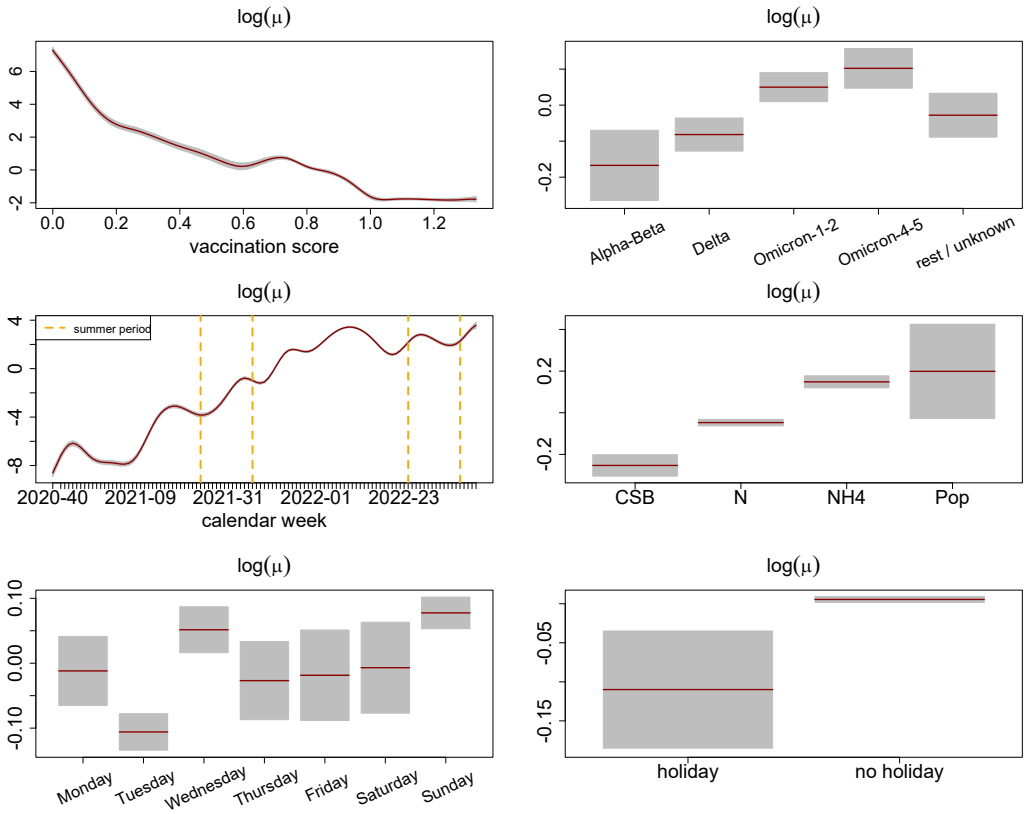

Figure 14: Covariate effect plots of Model 4 for $\log(\mu)$

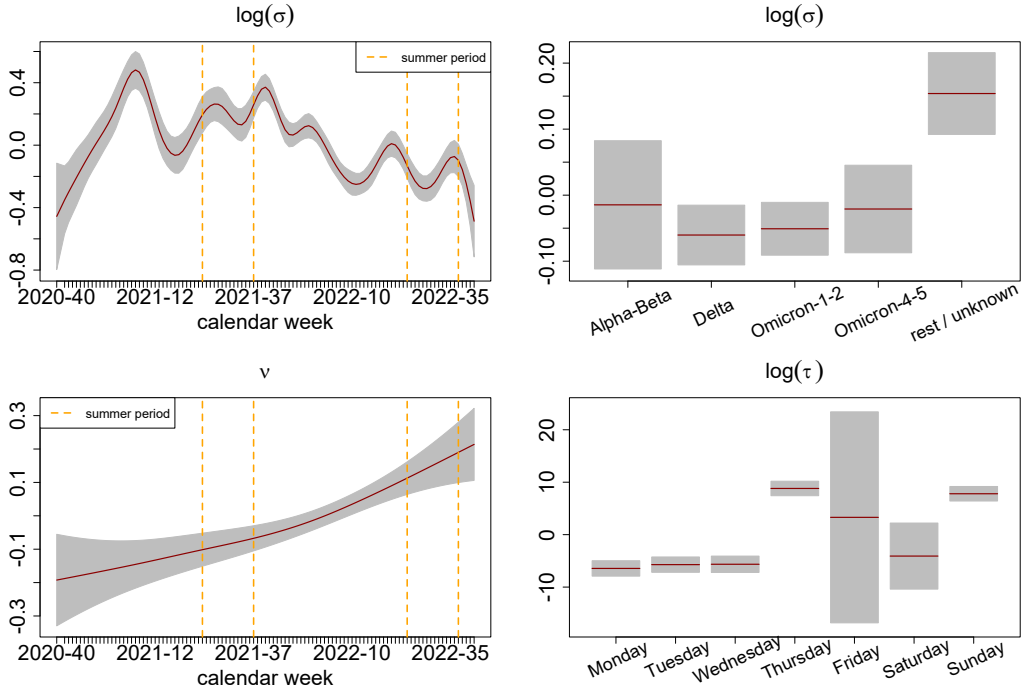

Figure 15: Covariate effect plots of Model 4 for $\log(\sigma), \nu$ and $\log(\tau)$

B.5. Additional effect plots of Model 5 (BCT-3)

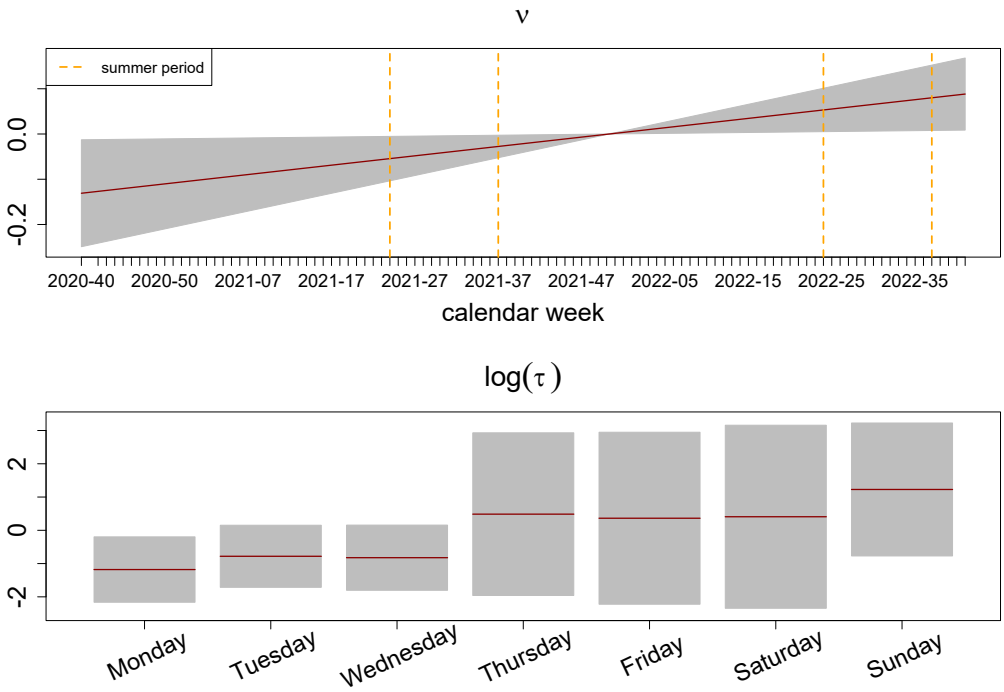

Figure 16: Covariate effect plots of Model 5 for ν and $\log(\tau)$

B.6. Effect plots of Model 6 (BCT-4)

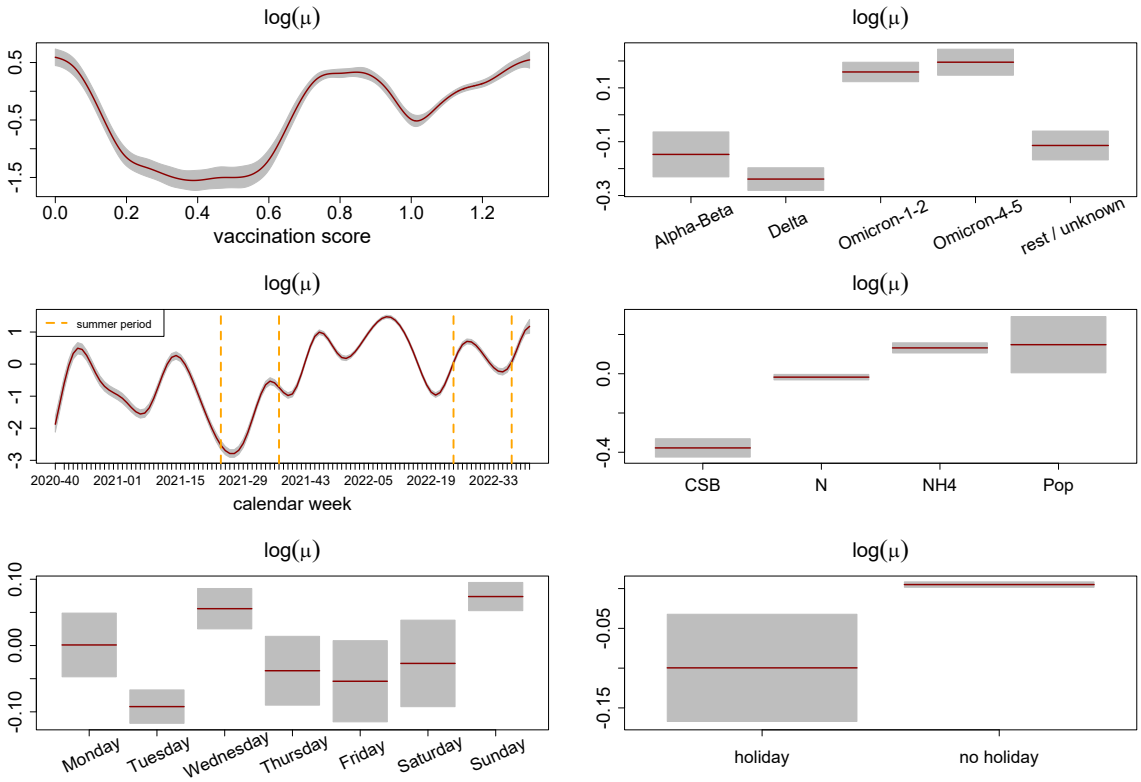

Figure 17: Covariate effect plot of Model 6 for $\log(\mu)$

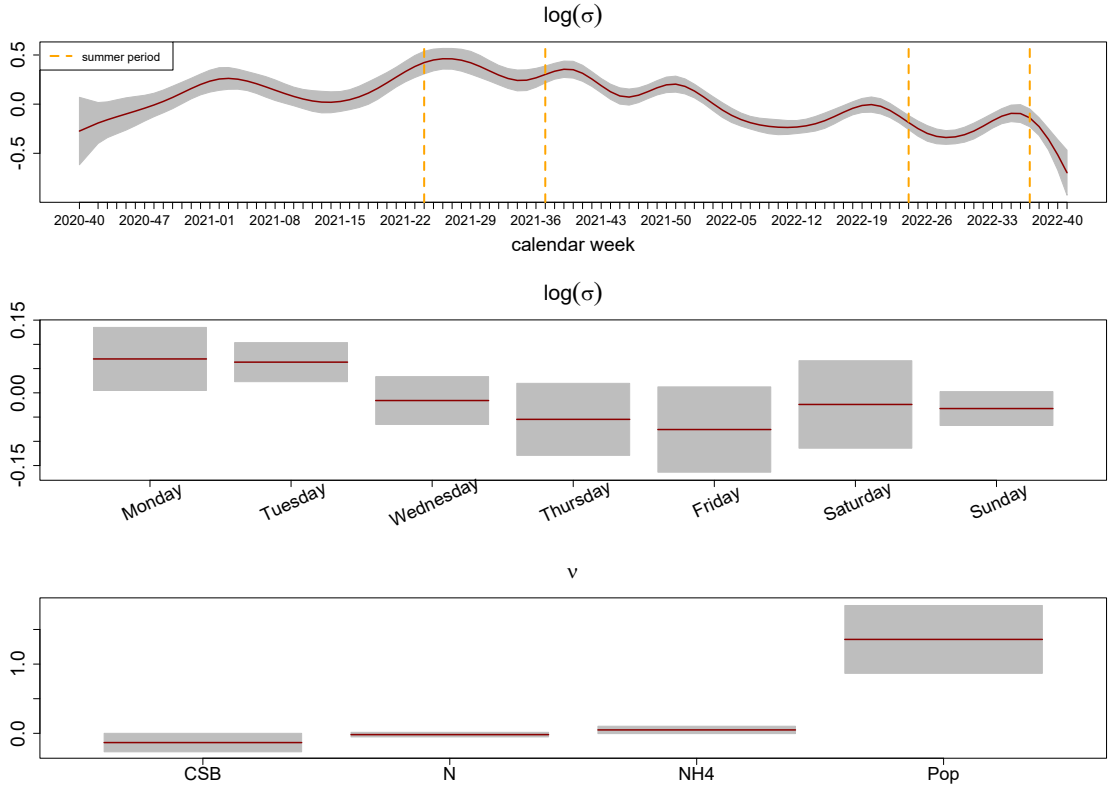

Figure 18: Covariate effect plot of Model 6 for $\log(\sigma)$ and ν

B.7. Effect plots of Model 7 (BCT-5)

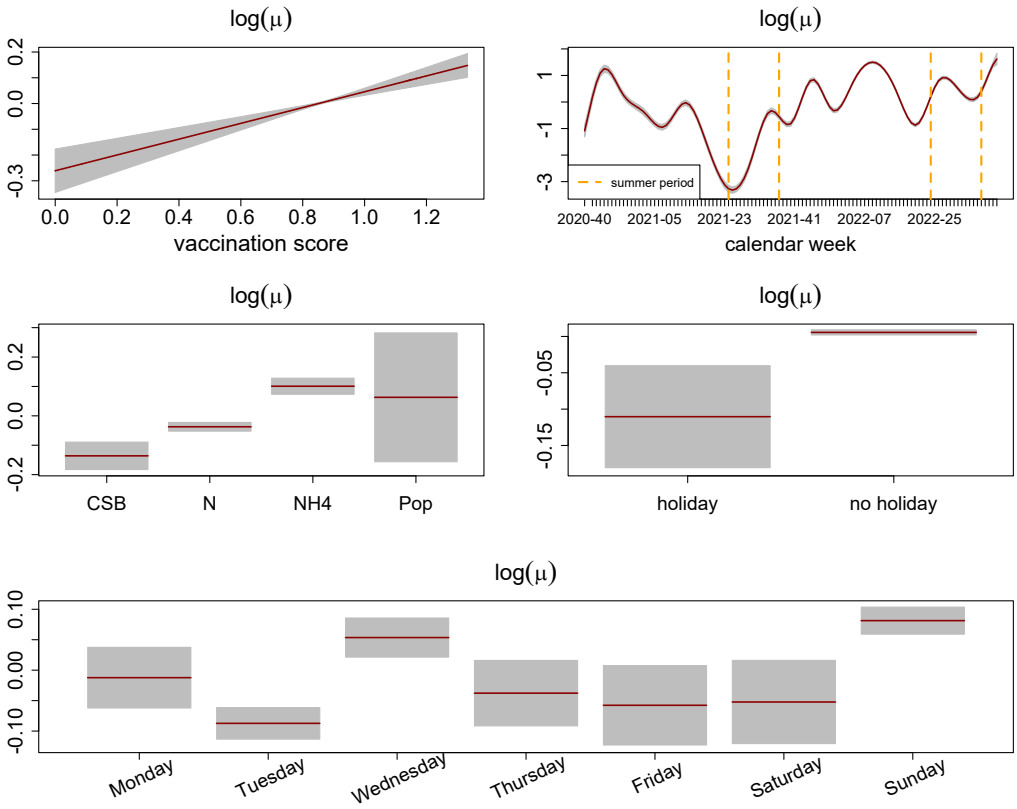

Figure 19: Covariate effect plot of Model 7 for $\log(\mu)$

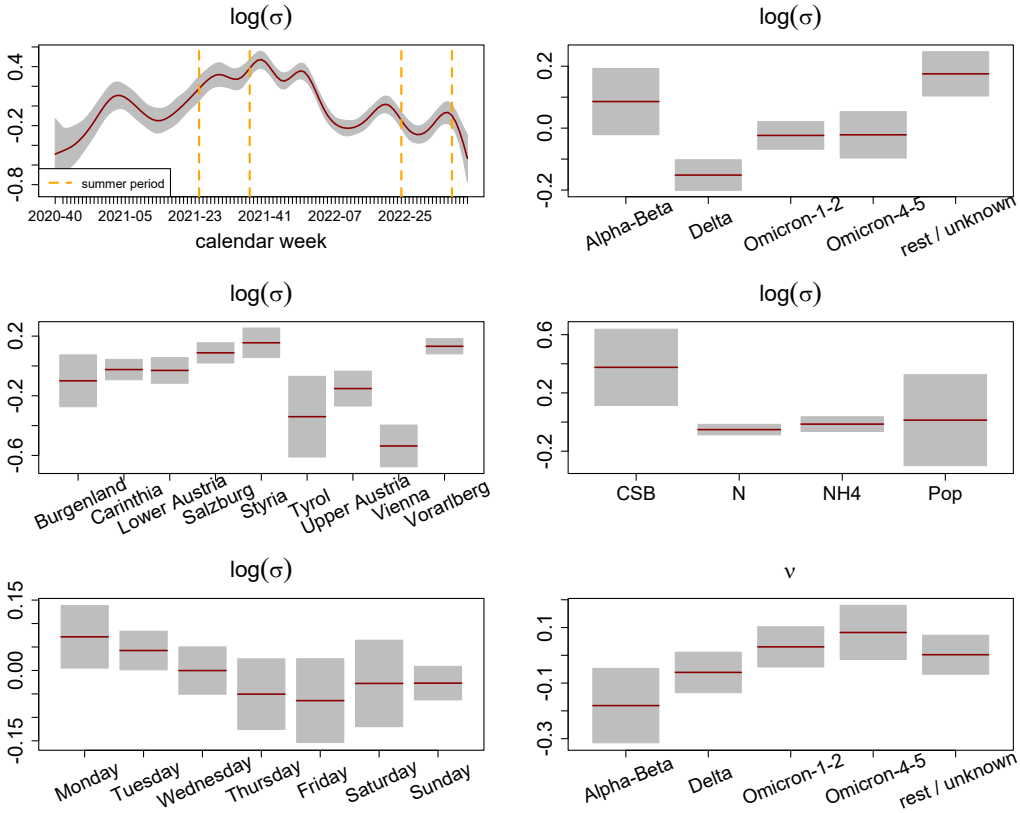

Figure 20: Covariate effect plot of Model 7 for $\log(\sigma)$ and ν

B.8. Effect plots of Model 8 (BCT-6)

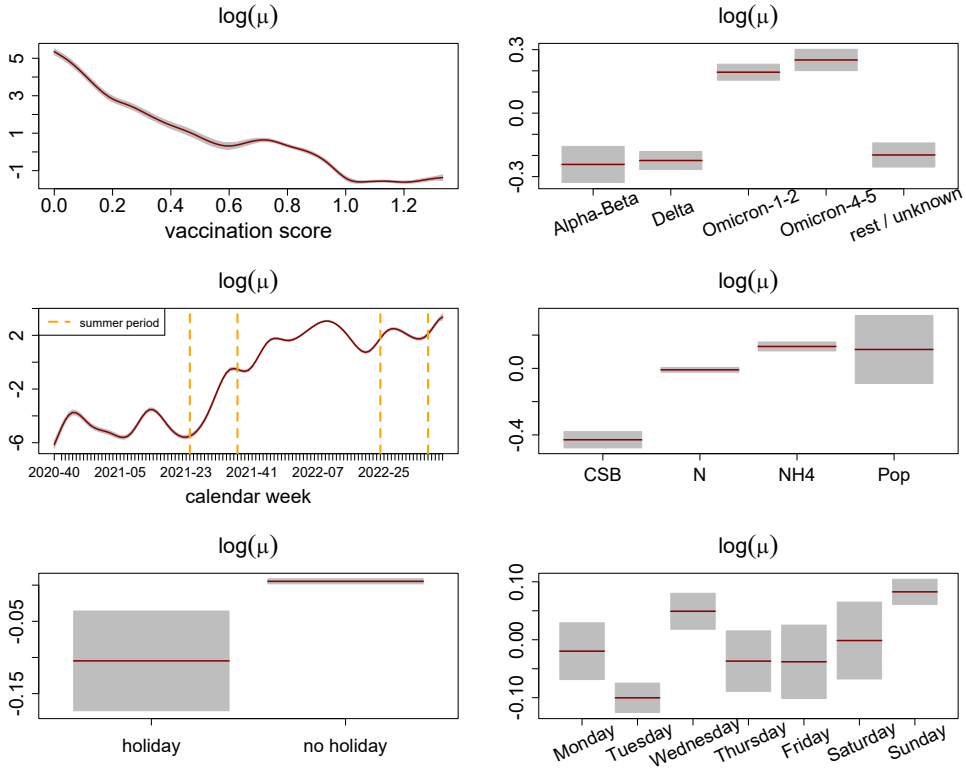

Figure 21: Covariate effect plot of Model 8 for $\log(\mu)$

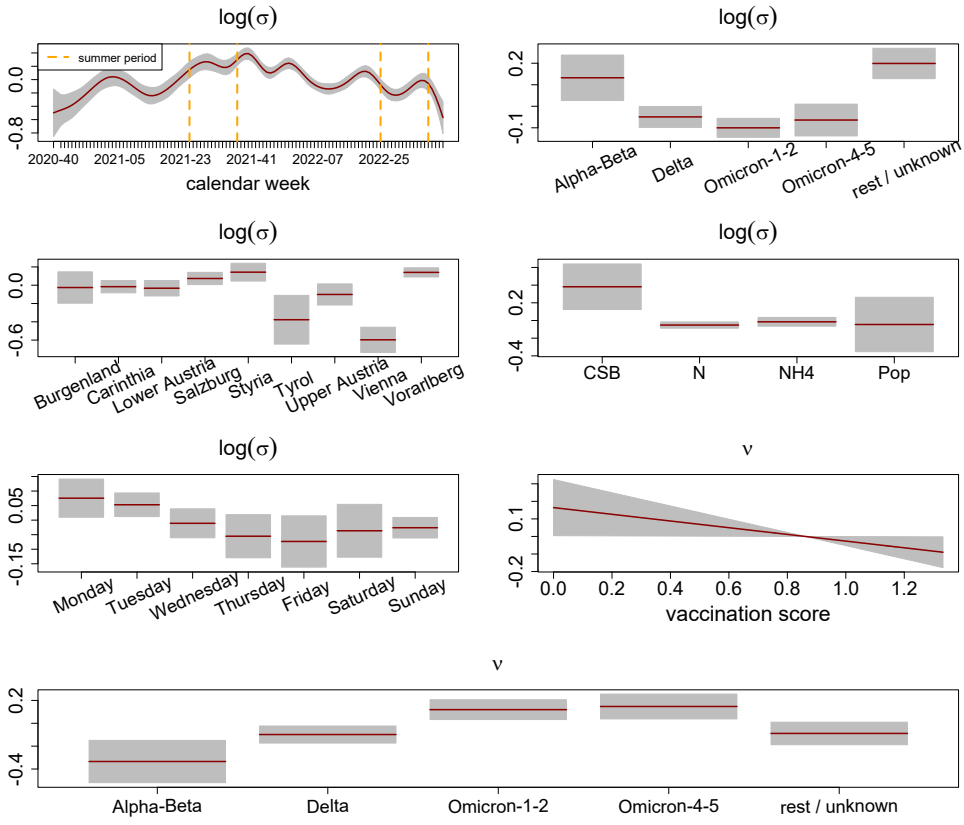

Figure 22: Covariate effect plot of Model 8 for $\log(\sigma)$ and ν

Affiliation:

Roman Pfeiler
Institute of Applied Statistics
Johannes Kepler University
Linz, Austria
E-mail: roman.pfeiler@jku.at
URL: https://www.jku.at/institut-fuer-angewandte-statistik/

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 55
2026

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2024-06-21
Accepted: 2026-02-06