Journal of the Royal Statistical Society Series C: Applied Statistics, 2024, **00**, 1–22
https://doi.org/10.1093/jrsssc/qlae016

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series C Applied Statistics

Original Article

Generalized functional additive mixed models with (functional) compositional covariates for areal Covid-19 incidence curves

Matthias Eckardt[1], **Jorge Mateu**[2] **and Sonja Greven**[1]

[1]Chair of Statistics, Humboldt-Universität zu Berlin, Berlin, Germany
[2]Department of Mathematics, University Jaume I, Castellón, Spain

Address for correspondence: Matthias Eckardt, Chair of Statistics, Humboldt-Universität zu Berlin, Unter den Linden 6 (UL6), 10099 Berlin, Germany. Email: m.eckardt@hu-berlin.de

Abstract

We extend the generalized functional additive mixed model to include compositional and functional compositional (density) covariates carrying relative information of a whole. Relying on the isometric isomorphism of the Bayes Hilbert space of probability densities with a sub-space of the L^2, we include functional compositions as transformed functional covariates with constrained yet interpretable effect function. The extended model allows for the estimation of linear, non-linear, and time-varying effects of scalar and functional covariates, as well as (correlated) functional random effects, in addition to the compositional effects. We use the model to estimate the effect of the age, sex, and smoking (functional) composition of the population on regional Covid-19 incidence data for Spain, while accounting for climatological and socio-demographic covariate effects and spatial correlation.

Keywords: compositional data analysis, Covid-19, functional compositions, functional data analysis, functional regression, function-on-function regression

1 Introduction

Understanding the infectious disease dynamics of the Covid-19 (Coronavirus disease 2019) pandemic and its potential associations with different exogenous environmental, socio-economic, and health-related variables has become an important challenge of current interdisciplinary research. Although massive data are collected, the interplay of the local numbers of daily Covid-19 cases with sets of different (potentially time-varying) risk factors still remains an open and challenging topic. In particular, this includes the effects of the composition of the local population (age, sex, smoking, etc.) on the spread of the disease, which has not been investigated to the best of our knowledge. This paper aims to fill this gap by extending the generalized functional additive mixed model (GFAMM) of Scheipl et al. (2016), which allows to model the particular spatio-temporal correlation structure of such data, to the case where parts of the covariate set are finite or infinite compositions, i.e. multivariate or functional covariates carrying relative information of a whole. In particular, in addition to the effects of climatological and socio-economic covariates, we aim to estimate the effects of the local male-to-female and smoking habits compositions as well as age distributions on the spread of the disease. The proposed formulation allows to model the local Covid-19 dynamics conditional on various types of local exogenous variables, including among others the population density (a scalar), the average temperature (a function of time), the smoking status (a finite composition of smokers, non-, ex-, and occasional smokers), the age composition (a functional composition also known as infinite composition or density) and the regional structure

Received: January 21, 2022. **Revised:** February 26, 2024. **Accepted:** February 26, 2024

(a grouping factor with spatial correlation). To this end, using areal Covid-19 incidence data collected in Spain daily until the vaccination onset, the local incidence counts over time for 52 Spanish provinces are modelled as generalized functional responses, as a function of various potential scalar, functional, compositional, and functional compositional risk factors, within one unified framework taking a functional data analysis (Ramsay & Silverman, 1997) perspective. Our contributions are thus twofold and include on the applied side a careful investigation of risk factors on the spread of Covid-19 in Spain, and on the methodological side an extension of the flexible GFAMM model for generalized functional responses to include new compositional and functional compositional covariate effects. Our proposed framework not only allows to estimate the effects of different types of covariates but also takes into account the expected spatial and temporal dependence structure of the disease curves. In particular, the view as spatially correlated functional data allows to naturally model the non-stationarity of the temporal correlation that is to be expected in Covid-19 data collected over several waves. This is in contrast to potential alternative spatio-temporal models for such data such as auto-regressive Poisson models (Congdon, 2022), for which additionally no compositional, functional, or functional compositional effect terms are available to the best of our knowledge.

In functional data analysis, the response curves are considered as realizations of some stochastic process with continuous support, such as time. While the process itself could theoretically be observed for any point at arbitrary resolutions, the curves are only measured on a discrete grid. A suitable class of functional data analysis techniques for the present purpose are functional regression models with functional responses and scalar as well as functional covariates, see Morris (2015) and Greven and Scheipl (2017a) for an overview. Regressions for non-Gaussian functional responses (e.g. counts) include the generalized function-on-scalar model (Goldsmith et al., 2015) and also the GFAMM (Scheipl et al., 2016), which provides a flexible regression framework for possibly non-Gaussian functional responses on potentially irregular or sparse grids using basis function representations of the fixed and/or random effects of scalar and/or functional covariates. We note that a complementary approach for (non-generalized) functional regression models observed on equidistant grids is presented in Morris and Carroll (2006) and subsequent work in a Bayesian framework, see Morris (2017) and Greven and Scheipl (2017b) for a comparison. Besides the above regression context, generalized functional data have also been considered in other contexts such as generalized functional principal component analysis (Gertheiss et al., 2017).

Although different (generalized) functional regression specifications exist, only a very small body of the literature discusses extensions of functional regression models to the case where the responses or parts of the covariate set are finite or infinite compositions. Predominantly, the literature focused on extensions to density-valued (functional compositional) responses while extensions to density-valued covariates appeared only rarely (see Petersen et al., 2021 for a recent review of different statistical approaches to density-valued quantities). Sierra et al. (2015) treat density-valued explanatory variables and Park and Qian (2012) treat both density-valued explanatory variables and responses as Hilbertian random variables in the standard L^2 space, which does not account for the constrained nature of these variables. The additive regression model of Han et al. (2020) first maps density-valued responses to the L^2 space in a pre-processing step, while Happ et al. (2019) point out instabilities of these pre-transformations. The implicit model formulation of Petersen and Müller (2019) in a non-linear space does not allow for straightforward interpretation of regression coefficients. Different from the above approaches, Arata (2017), Talská et al. (2018), and Maier et al. (2021) applied a Bayes Hilbert space formulation (van den Boogaart et al., 2014) to include density-valued responses into a functional regression framework where the estimation uses the centred log-ratio (clr) transformation to map the response to a sub-space of the L^2 with integration-to-zero constraint. Although this approach provides a promising framework for density-valued response regression, extensions which (additionally) allow for density-valued explanatory variables remain extremely limited. A first linear Bayes Hilbert space regression model for scalar responses and density-valued covariates was proposed by Talská et al. (2021) using a constrained spline representation (Machalová et al., 2021) and extended further by Scimone et al. (2021), allowing for both density-valued responses and covariates. Both models allow for linear effects of the functional composition on the scalar/density response. No existing work covers the setting of generalized functional response with flexible additive models including linear, non-linear and time-varying effects and where parts of the predictor are finite or infinite compositions.

We fill this gap within the flexible GFAMM framework by extending the functional additive mixed model predictor by (functional) compositional covariate effects. Additive mixed models have previously been extended for scalar responses to include finite compositional covariates in Verbelen et al. (2018). In comparison, our model allows for (a) (generalized) functional responses and (b) functional (infinite) compositions as covariates.

In particular, in our framework both finite and infinite compositions are included into a general structured predictor through suitable basis function representations that account for the constrained covariate nature and allow for interpretable additive effect estimates for the generalized functional responses. All data and R code to reproduce the proposed model are made publicly available in a github repository https://github.com/MatkcE/CoDaGFAMM.

We note that the phrase *functional composition* was also used by Sun et al. (2020) to refer to finite compositional covariates which could additionally vary over time. Different from this paper, we follow Hron et al. (2016) to indicate a constrained function that is a composition of infinitely many parts (i.e. a density).

The remainder of this paper is structured as follows. Section 2 briefly reviews the history of the pandemic in Spain and recent findings on potential risk factors for the spread of the disease that inform our selection of covariate effects. An introduction to the GFAMM model is presented in Section 3. In particular, extensions of the covariate effects to finite and infinite compositions are discussed in Section 3.2. More information on the different data sources we used to compile the Spanish Covid-19 incidence data for our analysis, and an application of the proposed model extension to these data are given in Section 4. The paper concludes with a discussion in Section 5.

2 Covid-19 data for Spain

2.1 A small history of the Covid-19 pandemic in Spain

Within 3 months of the official notification of a small regional outbreak in Wuhan, China, in late December 2019, Spain was facing one of the highest infection and, in particular, mortality rates among the European countries (Soriano & Barreiro, 2020). Within 4 weeks of the first tourist-based case on the Island of La Gomera in January and the first official domestic hospitalization on 15 February 2020, Spain witnessed a large but spatially strongly heterogeneous increase in numbers of Covid-19 infections with a clear concentration in large metropolitan conurbations (Henríquez et al., 2020). These marked regional differences and early local peaks in Madrid may in part be explained by the regional mobility from and to the Spanish capital (cf. Mazzoli et al., 2020). Using data for Catalonia, Coma Redon et al. (2020) provided some evidence that early local Covid-19 cases may have been masked by excess of influenza notifications between 4 February 2020 and 20 March 2020 as polymerase chain reaction (PCR) tests were restricted to hospital-admitted patients only and general practitioners were asked to diagnose Covid-19 infections without PCR confirmation. Due to this uncertainty, early cases and deaths in private and nursing homes may have been excluded from the official reports for this period. However, previously undetected symptom-based cases and deaths were subsequently added to the official notification system such that we can plausibly assume that any remaining bias can be neglected in the present study (while a sensitivity analysis will later investigate this point).

In response to the rapid increase in mortality, particularly among the elderly (potentially multimorbid) parts of the population, and the transmission dynamics of the disease, the Spanish National Government imposed a series of global regulatory interventions to suppress the spread of the virus. A first strict lockdown excluding only essential services (e.g. food, health) and some economic subsectors was imposed on 14 March. This measure was tightened in subsequent actions by placing strict entry refusal at the Spanish borders on 17 March and prohibiting any non-essential activities within the period from 30 March to 12 April 2020. In consequence, these restrictions yielded a dramatic reduction in overall regional mobility. Ended on 21 June, this lockdown remains the only global action of the Spanish Government against the Covid-19 pandemic. Although facing severe increases in numbers of Covid-19 cases in the second half of 2020 and early 2021, the public health responsibilities were delegated back to the local governments of the autonomous communities by 21 June and only local but no further global restrictions were reimposed. These (mostly soft) local measures, however, show a strong heterogeneity, and

information on the exact timing, nature and extend of the different imposed restrictions was not available from official sources for our study.

2.2 Recent findings on risk factors for the spread of Covid-19

Apart from a higher risk of Covid-19 infections caused by the increase in immunosenescence for the older ages (Crimmins, 2020), a clear association of higher age with the development of severe symptomatic Covid-19 infections, hospitalization and fatality rates is stressed in the literature (see, e.g. Tiruneh et al., 2021). Besides certain health conditions and comorbidities (Du et al., 2021), Wolff et al. (2021) identified smoking as an important co-factor. In particular, active smokers or non-active smokers with a clear smoking history face an increased risk for the development of symptomatic Covid-19 (Gülsen et al., 2020; Hopkinson et al., 2021). Apart from these findings, e.g. Moosa and Khatatbeh (2021) reported a close relation between densely populated regions and contact rates between different (potentially infected) individuals, which positively impact the disease transmission and, in turn, the reproduction rate of the disease. This idea is also supported by Paez et al. (2021) who reported a clear positive effect of mass transport systems on the incidence. These authors also reported a positive association of the disease with wealthier regions, i.e. regions with a higher GDP per capita, with a potential explanation via a connection between wealth and the regional level of globalization, i.e. international trade and travel.

The effect of climatological and environmental covariates on the spread of the disease is less consistent. In line with results on similar pathogens, which suggest that the virus is more stable and transferable in conditions of low temperature and low humidity, Paez et al. (2021) found a negative effect of higher values in temperature and humidity on the incidence of the disease, contrasted with a positive impact of sunshine. In a systematic review, Mecenas et al. (2020) found a positive effect of cold and dry weather conditions on the seasonal viability and transmissibility of Covid-19. Takagi et al. (2020) reported an inverse association of temperature, air pressure, and ultraviolet light with the prevalence of the disease, while Hossain et al. (2021) draw mixed conclusions based on both positive and negative effects of the weather characteristics. Shahzad et al. (2020) highlighted a clear positive effect of bad air quality on the transmission of Covid-19, whereas temperature serves only as a contributory factor, with higher temperatures reducing the spread of the disease. Summarizing the findings of 23 articles in a systematic review, McClymont and Hu (2021) found a clear association of temperature and Covid-19 and also a significant association of humidity and Covid-19 which, however, was derived from mixed results. Using a non-linear effect specification, Wu et al. (2020) found a negative association of high temperature and also high humidity with the daily number of Covid-19 cases and associated deaths.

3 Model specification

We will model the Spanish case counts over time as generalized functional data, including as covariates those variables that arise as potentially important from the literature discussed in Section 2. To account for linear, non-linear and time-varying effects as well as spatial correlation, we will use the GFAMM (Scheipl et al., 2016) framework, which we thus summarize in Section 3.1. As this model does not yet allow for compositional and density covariates, such as available for the sex, age, and smoking compositions in the Spanish regions, we extend the functional additive predictor to include corresponding interpretable additive effects in Section 3.2.

3.1 The GFAMM

We adopt a general structured additive regression model for generalized functional responses $Y_i(t) \sim \mathcal{F}(\mu_i(t))$, where $Y_i(t)$ in our setting is the number of Covid-19 cases in province $i = 1, \ldots, 52$ at time $t \in \mathcal{T}$. In our notation, we follow the usual convention of using capital letters for random and small for observed quantities. In general, we assume that $Y_i(t)$ point-wise follows a (here count) distribution $\mathcal{F}$ with conditional expectation $\mathbb{E}[Y_i(t) \mid x_{it}, t] = \mu_i(t)$ and is recorded over a domain $\mathcal{T}$, here covering the 381 days of observations. Building on Wood et al. (2016), Wood (2017), and Scheipl et al. (2016), $\mathcal{F}$ can be an exponential family distribution, a Tweedie, Negative Binomial, Beta, zero-inflated Poisson, or scaled and shifted t-distribution. An overdispersed Poisson model can be estimated using a quasi-likelihood (Wood, 2017).

Ordered categorical responses are also possible, in which case $\mu_i(t)$ is not the conditional mean of $Y_i(t)$ but of a latent variable determining the response category (cf. Wood et al., 2016).

To achieve high flexibility of the model, the mean $\mu_i(t)$ is related to a structured additive predictor $\eta_i(t)$ through a known link function g,

$$g(\mu_i(t)) = \eta_i(t) = \sum_{r=1}^{R} f_r(x_{rit}, t).$$

Here, r indexes the R structured additive model terms, with each such $f_r(x_{rit}, t)$ being a smooth function of the argument t of the outcome—also implying smoothness of the response mean $\mu_i(t)$—and of a subset x_{rit} of the complete covariate set x_{it}.

The above formulation allows for linear, non-linear, and time-varying effects of grouping factors, functional and potentially time-varying scalar covariates, as well as functional random effects, where the form of $f_r(x_{rit}, t)$ is determined by the covariates in x_{rit} and the chosen effect type. For example, for a functional intercept that varies over t, as the baseline rate of Covid-19 cases in our application, x_{rit} is the empty set and $f_r(x_{rit}, t)$ simplifies to $\beta_0(t)$. For a smooth effect of a scalar x_{ir} that is constant over t, i.e. $x_{rit} \equiv x_{ir}$, $f_r(x_{rit}, t)$ becomes $f_r(x_{ir})$ (and $f_r(x_{ir}, t)$ in the time-varying case), whereas linear effects of x_{ir} that vary over t (as for the effect of coastal provinces) correspond to $x_{ir}\beta(t)$. Linear time-varying effects of a functional covariate $x_{rit} \equiv x_{ir}$ with values $x_{ir}(s)$, $s \in \mathcal{S}$, are included as $f_r(x_{ir}, t) = \int_{\mathcal{S}} x_{ir}(s)\beta(s, t)\,\mathrm{d}s$, while a concurrent effect for $\mathcal{S} = \mathcal{T}$ can be included as $f_r(x_{ir}(t), t)$ or $f_r(x_{ir}(t))$ as for the climatological variables in our data. For a grouping variable $x_{rit} = c$ with M levels, scalar and functional random effects γ_c and $\gamma_c(t)$ are included as zero mean Gaussian variables with a potentially general correlation structure, and as Gaussian processes $\mathcal{T} \times \{1, \ldots M\} \to \mathbb{R}$ with general covariance function $\mathbb{C}\mathrm{ov}(\gamma_c(t), \gamma_{c'}(t'))$ that is smooth in t, t'. This specification also allows to control for the spatial correlation of the different levels of $c = i$ (formalized through the precision matrix of a Markov random field (MRF) with known correlation based on the planar neighbourhood structure) in the construction of (potentially spatially correlated) smooth residual curves for the individual locations. In the case of multiple random effects, a mutual independence assumption is placed between the individual random terms. In our application, we can include spatially correlated smooth functions per province and additional uncorrelated smooth curves per community to capture heterogeneity due to local Covid-19 measures. See Scheipl et al. (2016, 2015) for a full detailed list of potential covariate specifications. We note that in case $\eta_i(t)$ includes (functional) random effects $f_r(x_{ir}, t) = \gamma_c(t)$ or $f_r(x_{ir}, t) = \gamma_c$, the modelled conditional expectation also conditions on the vector of random effects (functions) $\boldsymbol{\gamma}$ and is taken to be $\mu_i(t) = \mathbb{E}[Y_i(t) \mid x_{it}, t, \boldsymbol{\gamma}]$.

Having n generalized functional observations $y_i(t)$ on a grid of T_i points $\mathbf{t}_i = (t_{i1}, \ldots, t_{iT_i})^\top$ available, the model can be fitted through a penalized (quasi-)likelihood approach based on the $(\sum_{i=1}^{n} T_i)$-vectors $\mathbf{y} = (\mathbf{y}_1{}^\top, \ldots, \mathbf{y}_n{}^\top)^\top$ and $\mathbf{t} = (\mathbf{t}_1{}^\top, \ldots, \mathbf{t}_n{}^\top)^\top$ of the concatenated response curves and their arguments, respectively, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})$ with $y_{il} = y_i(t_{il})$, $l = 1, \ldots, T_i$. While $T_i \equiv 381 =: T$ and equal $\mathbf{t}_i$ for all i in the present application, we note that the proposed framework also allows for differently spaced arguments t_{ij}. On the grid, we can simplify notation and write the predictor as $\eta_{il} = \sum_{r=1}^{R} f_r(\mathbf{x}_{rit_l}, t_{il})$. Each of the R terms $f_r(\mathbf{x}_r, \mathbf{t})$, containing evaluations $f_r(\mathbf{x}_{rit_l}, t_{il})$, $l = 1, \ldots, T_i$, $i = 1, \ldots, n$ in a vector, can then be represented through a tensor product basis function expansion

$$f_r(\mathbf{x}_r, \mathbf{t}) \approx (\boldsymbol{\Phi}_{xr} \odot \boldsymbol{\Phi}_{tr})\boldsymbol{\vartheta}_r = \boldsymbol{\Phi}_r\boldsymbol{\vartheta}_r,$$

where $\mathbf{A} \odot \mathbf{B} = (\mathbf{A} \otimes \mathbf{1}_b{}^\top) \cdot (\mathbf{1}_a{}^\top \otimes \mathbf{B})$ denotes the row tensor product of the matrices $\mathbf{A}$ $(h \times a)$ and $\mathbf{B}$ $(h \times b)$, with $\mathbf{1}_d$ the d-vector of ones and $\cdot$ the element-wise multiplication, and $\boldsymbol{\Phi}_{xr}$ and $\boldsymbol{\Phi}_{tr}$ contain the evaluations of the (K_{xr}, respectively, K_{tr}) marginal basis functions for the covariate effects and over t, respectively, discussed in Subsection 3.1.1 below. The effect shape is determined by the unknown vector of coefficients $\boldsymbol{\vartheta}_r$. To provide sufficient flexibility of the model, the approximation uses a large set of basis functions, which is regularized by an anisotropic quadratic penalty term for the coefficients in the (quasi-)log-likelihood

$$\mathrm{pen}(\boldsymbol{\vartheta}_r \mid \lambda_{tr}, \lambda_{xr}) = \boldsymbol{\vartheta}_r{}^\top(\lambda_{xr}\mathbf{P}_{xr} \otimes \mathbf{I}_{K_{tr}} + \lambda_{tr}\mathbf{I}_{K_{xr}} \otimes \mathbf{P}_{tr})\boldsymbol{\vartheta}_r.$$

Here, λ_{xr} and λ_{tr} are positive smoothing parameters, $\mathbf{P}_{xr}$ and $\mathbf{P}_{tr}$ are known and fixed positive (semi-)definite marginal penality matrices corresponding to the basis matrices $\mathbf{\Phi}_{xr}$ and $\mathbf{\Phi}_{tr}$, and $\mathbf{I}_{K_{xr}}$ and $\mathbf{I}_{K_{tr}}$ are identity matrices of dimensions K_{xr} and K_{tr}, respectively (see Scheipl et al., 2016, 2015). For given smoothing parameters, the unknown coefficients can then be estimated through a penalized (quasi-)maximum likelihood approach, maximizing

$$\ell_p(\boldsymbol{\vartheta}, \boldsymbol{\lambda}, \boldsymbol{\nu} \mid \mathbf{y}) = \ell(\boldsymbol{\vartheta}, \boldsymbol{\nu} \mid \mathbf{y}) - \frac{1}{2}\sum_{r=1}^{R} \text{pen}(\boldsymbol{\vartheta}_r \mid \lambda_{tr}, \lambda_{xr})$$

where $\boldsymbol{\lambda} = (\lambda_{t1}, \lambda_{x1} \ldots, \lambda_{tR}, \lambda_{xR})$, $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1{}^\top, \ldots, \boldsymbol{\vartheta}_R{}^\top)^\top$, and $\ell(\boldsymbol{\vartheta}, \boldsymbol{\nu} \mid \mathbf{y})$ in most cases is the log-likelihood based on $\mathbf{y}$ implied by the respective chosen response distribution $\mathcal{F}$ for $Y_i(t)$ or else the quasi-log-likelihood as for the overdispersed Poisson (see Scheipl et al., 2016), optionally depending on additional nuisance (e.g. dispersion) parameters $\boldsymbol{\nu}$. Smoothing parameters are estimated using a (Laplace approximated) marginal (extended quasi-)likelihood (Nelder & Pregibon, 1987), which extends the usual marginal likelihood approach that corresponds to restricted maximum likelihood in the case of a Gaussian likelihood, and which is known to work well for the choice of smoothing parameters (Wood et al., 2016).

3.1.1 Basis function representations for different covariate effects

The marginal basis matrices and corresponding penalty matrices are suitably chosen depending on the specified covariate effects. A full description is given in Scheipl et al. (2016, 2015); we here list some common choices also used in the model for the Covid-19 data for illustration and completeness, restricting to the case of equal grids for ease of presentation. For covariate effects that are constant over t, $\mathbf{\Phi}_{tr} = \mathbf{1}_{nT}$ is a vector of length nT containing ones and $\mathbf{P}_{tr} = 0$, while smooth time-varying effects are achieved when choosing $\mathbf{\Phi}_{tr}$ as a matrix of spline evaluations with $\mathbf{P}_{tr}$ a corresponding penalty matrix (e.g. based on finite differences of B-spline coefficients). (Functional) intercepts β_0, $\beta_0(t)$ are obtained through $\mathbf{\Phi}_{xr} = \mathbf{1}_{nT}$ and $\mathbf{P}_{xr} = 0$. For effects $x\beta$ and $x\beta(t)$ that are linear in x, $\mathbf{\Phi}_{xr}$ changes to $\mathbf{\Phi}_{xr} = \mathbf{x} \otimes \mathbf{1}_T$ where $\mathbf{x} = (x_1, \ldots, x_n)^\top$ and $\mathbf{P}_{xr} = \mathbf{0}$. In case of a non-linear effect specification for x, i.e. $f(x)$ and $f(x, t)$, $\mathbf{\Phi}_{xr}$ corresponds to a suitable marginal spline basis matrix over x and $\mathbf{P}_{xr}$ is specified accordingly.

For linear effects of a functional covariate $x(s)$, $s \in \mathcal{S}$, a tensor product spline representation for $\beta(s, t)$ is used with marginal spline basis functions Φ_{k_s}, $k_s = 1, \ldots, K_{xr}$ over $\mathcal{S}$ and Φ_{k_t}, $k_t = 1, \ldots, K_{tr}$ over $\mathcal{T}$. This yields

$$\int_{\mathcal{S}} x_i(s)\beta(s, t_l)\, \mathrm{d}s \approx \int_{\mathcal{S}} x_i(s) \sum_{k_s=1}^{K_{xr}} \sum_{k_t=1}^{K_{tr}} \Phi_{k_s}(s)\Phi_{k_t}(t_l)\vartheta_{r,k_s,k_t}\, \mathrm{d}s.$$

Then $\mathbf{\Phi}_{tr} = [\Phi_{k_t}(t_l)]_{\substack{l=1,\ldots,T \\ k_t=1,\ldots,K_{tr}}} \otimes \mathbf{1}_n$ and $\mathbf{\Phi}_{xr} = [\int_{\mathcal{S}} x_i(s)\Phi_{k_s}(s)\, \mathrm{d}s]_{\substack{i=1,\ldots,n \\ k_s=1,\ldots,K_{xr}}} \otimes \mathbf{1}_T$ with marginal penalty matrices corresponding to the chosen marginal spline bases. In practice, the integral is approximated using numerical integration. A concurrent effect $f(x(t))$ or $f(x(t), t)$ of a functional covariate $x(s)$, $s \in \mathcal{T}$, is constructed analogous to $f(x, t)$ above.

Finally, functional random intercepts for groups $c = 1, \ldots, M$, $c(i)$ being the group level of observation i (e.g. the community or province), are associated with a marginal basis $\mathbf{\Phi}_{xr} = [\delta_{c(i)m}]_{\substack{i=1,\ldots,n \\ m=1,\ldots,M}} \otimes \mathbf{1}_T$, with δ_{cm} the indicator for $c = m$. The matrix $\mathbf{P}_{xr}$ then is an $M \times M$ precision matrix defining the dependence structure between levels of c.

3.2 Compositional predictor

While the GFAMM provides a rich methodological toolbox, compositional and functional compositional covariates have not yet been included into this framework, as in our context the age-, sex- and smoker-compositions of the provinces. We thus extend the predictor $\eta_i(t)$ by $\eta_i^{\text{comp}}(t)$ to include effects of both the compositional and functional compositional covariates into the GFAMM framework. We note that this extension is similar in spirit to Verbelen et al. (2018), who incorporated the effect of a (*non-functional*) compositional covariate into the additive

predictor of generalized additive models for scalar responses. We first discuss existing methods for the case of finite compositions as covariates and scalar responses in Section 3.2.1, before introducing the proposed extensions to functional compositional covariates and/or (generalized) functional responses in Section 3.2.2.

3.2.1 Finite compositional covariates and scalar responses

Following principles of compositional data analysis (Aitchison, 1986), we formalize vector-valued covariates describing D parts of a whole summing to a constant as compositions of D parts living on the simplex

$$\mathbb{S}^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D)^\top : x_d > 0, d = 1, \ldots, D; \sum_{d=1}^{D} x_d = \kappa \right\}.$$

Instances of such variables in our data at hand are the regional sex and smoking status compositions with $D = 2$ and $D = 4$, respectively. The simplex is provided with a finite $(D-1)$-dimensional Euclidean vector space structure isometric to the $\mathbb{R}^{D-1}$ (cf. e.g. Pawlowsky-Glahn & Egozcue, 2001), when equipped with the perturbation $\mathbf{x} \oplus \mathbf{u} = \mathrm{cls}(x_1 u_1, \ldots x_D u_D)$ and the powering $\alpha \odot \mathbf{x} = \mathrm{cls}(x_1^\alpha, \ldots, x_D^\alpha)$ operations, where $\mathbf{x}, \mathbf{u} \in \mathbb{S}^D$, $\alpha \in \mathbb{R}$ and $\mathrm{cls}(\mathbf{x}) = (\kappa x_1 / \sum_{j=1}^D x_j, \ldots, \kappa x_D / \sum_{j=1}^D x_j)^\top$ is the closure operator, as well as the inner product $\langle \mathbf{x}, \mathbf{u} \rangle_A = (2D)^{-1} \sum_d \sum_j \log(x_d/x_j) \log(u_d/u_j)$. Noting this correspondence, a central idea in compositional data analysis is to map compositions isometrically to $\mathbb{R}^{D-1}$, perform well-established statistical analysis methods there, and then potentially back-transform the result onto $\mathbb{S}^D$ using inverse operations.

Common transformations include first the centred log-ratio transformation

$$\mathrm{clr}(\mathbf{x}) = \left[\log \frac{x_1}{m(\mathbf{x})}, \ldots, \log \frac{x_D}{m(\mathbf{x})} \right],$$

where $m(\mathbf{x})$ is the geometric mean of $\mathbf{x}$. The clr projects the composition onto the clr-plane $\mathcal{H}^D$, a $(D-1)$-dimensional sub-space of $\mathbb{R}^D$ whose components add to zero. By contrast, the isometric log-ratio (ilr) transformation (Egozcue et al., 2003) returns $(D-1)$ coordinates with respect to an orthonormal system on the clr-plane $\mathcal{H}^D$, which is equivalent to the logit-function used in logistic regression for $D = 2$. For $D > 2$, infinitely many orthonormal basis systems exist. As we use the ilr only internally for estimation, the choice does not affect the interpretation and we use *pivot coordinates* (Fišerová & Hron, 2011) yielding $D = 1$ and $D = 3$ ilr coordinates for the sex and smoking composition in our application, respectively.

Making use of these isometric isomorphisms, linear effects of compositional covariates can be modelled as $\langle \mathbf{x}, \mathbf{b} \rangle_A = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{b}) \rangle = \langle \mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{b}) \rangle$ (e.g. Verbelen et al., 2018). With $\mathbf{b} = \mathrm{ilr}^{-1}(\boldsymbol{\beta})$ the inverse of the coefficients in $\mathbb{R}^{D-1}$, the compositional effect is

$$\langle \mathbf{b}, \mathbf{x} \rangle_A = \sum_{j=1}^{D-1} \beta_j \mathrm{ilr}_j(\mathbf{x}). \tag{1}$$

Recalling Barceló-Vidal et al. (2011), $\mathbf{b}$ represents the simplicial gradient of the predictor with respect to the composition $\mathbf{x}$, and can be interpreted as the direction of perturbation of compositions in $\mathbb{S}^D$ which yields the largest effect on the outcome (cf. Verbelen et al., 2018). That is, for $\mathbf{x} \oplus \mathbf{b}^*$ with $\mathbf{b}^* = \mathbf{b}/\|\mathbf{b}\|_A$, the increase is

$$\langle \mathbf{b}, \mathbf{x} \oplus \mathbf{b}^* \rangle_A = \langle \mathbf{b}, \mathbf{x} \rangle_A + \|\mathbf{b}\|_A. \tag{2}$$

Further, to quantify the effect of a change in the composition on the predictor, say an increase of the non-smoking share in the local smoking compositions, Verbelen et al. (2018) suggested to

perturb the composition into the direction of each part. For example, a change in the relative ratio of the first compositional component of $\mathbf{x}$ by some $\alpha \neq 1$, while keeping the relative ratios for all other components constant, leads to a perturbation $\mathbf{x} \otimes \text{cls}'$ of $\mathbf{x}$ by $\text{cls}' = \text{cls}(\alpha, 1, \ldots, 1)^\top$ and a resulting change on the predictor by $\langle \mathbf{b}, \text{cls}' \rangle_A = \log(\alpha)\text{clr}_1(\mathbf{b})$, where clr_1 is the first component of the clr transformation. In particular, for a log-link relation of the expected outcome and the predictor, as used in our application, the effect of a relative ratio change in the first component on the response scale simplifies to a change by the factor $\alpha^{\text{clr}_1(\mathbf{b})}$.

3.2.2 *Extensions to functional compositional covariates and functional responses*

The above formulation of equation (1) allows us to extend the GFAMM to include compositional covariates such as the sex and smoking composition in our application by including the $(D-1)$ ilr-transformed coordinates as scalar covariates with linear effects, such that $\mathbf{\Phi}_{xr} = \text{ilr}(\mathbf{X}) \otimes \mathbf{1}_T$ and $\mathbf{P}_{xr} = \mathbf{0}$ for $\mathbf{X} = (\mathbf{x}_{id})_{i,d}$ and a row-wise application of the ilr-transform. Combination with suitable $\mathbf{\Phi}_{tr}$ and $\mathbf{P}_{tr}$ as discussed above newly allows such effects also for (generalized) functional responses, with time-constant $\langle \mathbf{b}, \mathbf{x} \rangle_A$ or time-varying effect $\langle \mathbf{b}(t), \mathbf{x} \rangle_A$, respectively.

Treating density functions as infinite (functional) compositions, we extend the previous results to such covariates. A typical example of a functional composition is the age density in our application, which due to integration-to-one and strict positivity constraints cannot be treated within the classic functional data setting. As there is no extension of the ilr transformation used by Verbelen et al. (2018) to functional compositions, we take a different approach in the functional case. The idea is to use an isometric isomorphism between the space of functional compositions and a sub-space of the L^2 space of functions via a functional clr transformation, and to then treat the transformed functional composition as a functional covariate within the GFAMM framework using a suitably adapted basis function specification.

A suitable space in this context is the Bayes Hilbert space of densities

$$B^2(\mathcal{T}) = \left\{ f : \mathcal{T} \to (0, +\infty),\ \int_\mathcal{T} f(t)\,\mathrm{d}t = 1,\ \int_\mathcal{T} \left[\log(f(t))\right]^2 \mathrm{d}t < \infty \right\}$$

(Egozcue et al., 2006; van den Boogaart et al., 2014). It generalizes the Aitchison geometry from compositional data and provides a suitable geometric framework for the analysis of density functions. We here focus on some basic properties that are relevant in our setting and refer to van den Boogaart et al. (2014) for a more formal definition and further mathematical details. Analogous to $\mathbb{S}^D$, $B^2(\mathcal{T})$ has a vector space structure with perturbation and powering operations. For $f, h \in B^2(\mathcal{T})$, $t \in \mathcal{T}$ and $\alpha \in \mathbb{R}$, the perturbation ($\oplus$) and powering ($\odot$) operations are defined by $(f \oplus h)(t) = f(t)h(t) / \int_\mathcal{T} f(t)h(t)\,\mathrm{d}t$ and $(\alpha \odot f)(t) = f(t)^\alpha / \int_\mathcal{T} f(t)^\alpha\,\mathrm{d}t$, respectively. Additionally, the inner product $\langle \cdot, \cdot \rangle_{B^2}$ on $B^2(\mathcal{T})$ generalizes the Aitchison inner product,

$$\langle f, h \rangle_{B^2} = \frac{1}{2|\mathcal{T}|} \int_\mathcal{T} \int_\mathcal{T} \log\frac{f(t)}{f(s)} \log\frac{h(t)}{h(s)}\,\mathrm{d}s\,\mathrm{d}t,$$

where $f, h \in B^2(\mathcal{T})$ and $|\cdot|$ is the Lebesgue measure of the argument. In particular, noting that $\langle f, h \rangle_{B^2} = \langle \text{clr}(f), \text{clr}(h) \rangle_{L^2}$, where $\text{clr}(f)(t) = \log(f(t)) - |\mathcal{T}|^{-1} \int_\mathcal{T} \log(f(s))\,\mathrm{d}s$, the $B^2(\mathcal{T})$ can be shown to be a separable Hilbert space and to be isometrically isomorph to the sub-space $L_0^2(\mathcal{T})$ of functions in $L^2(\mathcal{T})$ integrating to zero with the usual L^2 metric (van den Boogaart et al., 2014). While this allows a transformation of densities to the $L^2(\mathcal{T})$, the additional integration-to-zero constraint of $L_0^2(\mathcal{T})$ needs to be accounted for and, in general, prohibits a direct application of standard functional data analysis techniques to the transformed densities.

For the GFAMM, functional compositions $x_i(s), s \in \mathcal{S}$, such as age in our case are included into the regression with a linear effect in terms of the scalar product in B^2, using the equivalence

$$\langle x_i, b(., t) \rangle_{B^2} = \langle \text{clr}(x_i), \text{clr}(b(., t)) \rangle_{L^2} = \int_\mathcal{S} u_i(s)\beta(s, t)\,\mathrm{d}s,$$

with $u_i = \text{clr}(x_i)$ and $\beta(., t) = \text{clr}(b(., t))$ for each t. Note that β is a surface with $\beta(., t) \in L_0^2(\mathcal{T})$ fulfilling an integration-to-zero constraint for each t. Thus, we can estimate the effect similarly to a linear function-on-function regression term, with the modification of this additional constraint. We achieve this through the specification of a tensor product basis, which places an integration-to-zero constraint on the marginal basis for β over s (but not on the marginal basis over t) to get terms with integration-to-zero-for-each-t constraints (see Wood, 2017, Chapter 5.6).

This model formulation results in interpretable linear effects of the functional composition. In a post estimation step, the functional composition surface $b(s, t)$ with $b(., t) \in B^2(\mathcal{T})$ for all t can be computed through the inverse clr transformation, $b(., t) = \text{clr}^{-1}(\beta(., t))$ for each t. Similar to finite compositions, $b(., t)$ can then be interpreted for each t as the preferential direction in which to perturb the functional composition to yield the largest increase in the outcome, from $\langle x_i, b(., t)\rangle_{B^2}$ to $\langle x_i, b(., t)\rangle_{B^2} + \|b(., t)\|_{B^2}$. Alternatively, to suitably extend the second interpretation of compositional covariate effects to our functional setting, we derive in Section 2.3 of the online supplementary material an interpretation based on the effect of a change in the relative ratio of the functional composition on some subinterval $A \subseteq \mathcal{S}$ relative to $A^C = \mathcal{S}\backslash A$. This corresponds to perturbing x_i to $x_i \oplus \text{cls}'$ with $\text{cls}' = (\alpha \mathbb{1}_A + \mathbb{1}_{A^C})/(\alpha|A| + |A^C|)$ and $\mathbb{1}_A$ the indicator function on A, which we show changes the additive predictor at time t by $(+\log(\alpha)\beta_A(t))$ with $\beta_A(t) = \int_A \beta(s, t)\, ds$, and changes the mean response at time t under a log-link by a factor $\alpha^{\beta_A(t)}$.

4 Application to the Spanish Covid-19 data

4.1 Data

4.1.1 Data sources and variables

The data were compiled from different sources, most commonly information provided by the regional governments. It originates from a collaborative data project by the geovoluntarios community (https://www.geovoluntarios.org), Centro de Datos Covid-19 and ESRI Spain and provides information on the daily numbers of Covid-19 cases for 52 Spanish provinces, each of which subsumes numerous local administrative units. It covers the period from 5 January 2020 to 19 January 2021 until just before vaccinations became more widespread. Covid-19 cases are defined as probable infections without test information or confirmed infections based on positive test results derived from (a) PCR, antibody, and antigen detection or Elisa techniques and (b) reported by other laboratories—showing a clear majority of results derived from positive PCR tests. In contrast, notifications based on antibody tests (with less precise timing information on the time of infection) contributed only at rather small and also spatially varying rate. Restricted to the first Spanish Covid-19 wave only, the highest regional proportion of antibody based test results relative to all cases reported appeared for the provinces of Cuenca (5.06%) and Albacete (2.34%). We thus use the complete incidence counts based on all tests. To account for a potential delay between the date of the test and the notification date caused by the individual testing procedures, the dates were shifted back by the provider using a 3 days lag. Different from data used in this study, official periodical data releases on the Covid-19 pandemic through the Spanish National Government cover only information at a coarser level of spatial aggregation, i.e. 18 so-called *autonomous communities*, to which the 52 Spanish provinces belong. Note that 2.05 % (resp. 0.02 %) of the Spanish population received the first (resp. second) vaccination before 19th January 2021. Due to this relatively small proportion of partly vaccinated inhabitants, any potential confounding effects of the vaccination action on the incidence data is assumed to be negligible.

To investigate (potentially time-varying) effects on the spread of the pandemic over space and time, we linked these data to climatological, socio-economic and demographic information, recorded for the provinces and time period under study (see Table 1 in the online supplementary material for a detailed description of the variables). Daily climatological information including the average daily temperature (in °C), humidity (in %), maximal wind speed (in km/hr), sun hours (in hr) and precipitation (in mm) were collected for each province from the State Meteorology Agency and the Ministry of Agriculture, Fisheries, and Food. The original weather data exhibited some missing values (days) across all 52 provinces and days under study, in particular 0.5% for average temperature, 0.77% for maximal wind speed, 2.07% for sun hours, 5.34% for precipitation, and 0.5% for humidity—most likely caused by transmission or technical problems. Any of

these missing values were imputed by linear interpolation using the `imputeTS` package (Moritz & Bartz-Beielstein, 2017) in R. As information on the daily solar exposure was not provided at all for Malaga, missing values for sun hours for Malaga were replaced by average values computed from the neighbouring provinces, i.e. Cadiz and Granada. For the precipitation variable, the reported values show a large number of zeros in the daily amount of precipitation at the province level (ranging from 36 days at minimum to 215 days at maximum) as well as skewness with extreme peaks of 150 mm. For this reason, we summarized the original information into a binary variable indicating the absence or presence of rain per province on a daily basis. In addition to this rain indicator, we computed the log transformation of the non-zero precipitations. Next, all weather information was shifted using a 5-day lag to account for the time lag between infection and symptom onset, as we want to investigate weather effects on the infection probability and assume symptom onset to be strongly correlated with the timing of the performed Covid-19 test. While the 5-day lag is supported by the findings of Linton et al. (2020) who reported an average incubation period (defined as the time from the infection to symptom onset) by around 5 days, we tested for the effect of different lag specifications in a sensitivity analysis.

Regional socio-demographic information on the number of inhabitants, the gross domestic product (GDP) per capita, the proportion of males, and age pyramids (0–100 years) were collected from public data provided by the national statistics institute. This source was also used to compute the smoking composition of the population (categorized in daily, occasional, ex-, and non-smokers) at the provided coarser level of the *autonomous communities*, which we then assigned to all provinces within this community.

4.1.2 Generated variables

In addition to the above data, we generated different variables to control for the regional and geographical characteristics of the individual spatial units. First, to account for a potential impact of public mass transportation systems on the transmission and spread of the virus, we generated a binary variable indicating whether or not a province offers access to a metro or subway system. In addition, we generated a second binary variable indicating whether or not a province offers direct access to the Mediterranean Sea or the Atlantic Ocean. Besides a higher population density in the coastal regions compared to the inland provinces (except for the metropolitan regions), the coastline is commonly strongly affected by high numbers of incoming tourists during the summer and public vacation periods, which might serve as an acceleration factor for the risk of infection. Finally, to account for (a) potential temporal variation in the regional notification systems and (b) the effect of the global lockdown measures, we generated 6 weekday dummy variables treating Sundays as reference and three lockdown dummies. Each of these indicates one of the three successive global lockdown periods with different measures imposed during the first wave in 2020, i.e. (a) 11–24 May, (b) 25 May–7 June, and (c) 8–21 June.

4.2 Data description

The response curves show strong regional heterogeneity, with the highest numbers of cases in the provinces of Madrid and Barcelona. The highest peak ($y = 6,750$) appeared for Madrid on 18 September 2020, contrasted with only $y = 77$ on that date recorded for the province of Lleida. To better understand the similarities among the spatial disease patterns, we calculated the regional incidence rates per 100,000 inhabitants, and its average version, computed as the mean rate per region over all 381 days (see Figure 1). The incidence curves (left panel) reflect a clear positive deviation for Madrid (green) and also Lleida (blue) from the mean curve (red) for the period from June to October 2020—with a temporal delay in the second wave for Madrid compared to Lleida. The corresponding regional averages (see right panel) indicate a clear spatial pattern, with Palencia and Cuenca showing the highest average incidence rates with 20.5 and 19.66 cases per 100,000 inhabitants, respectively. These high average values are contrasted with relatively small reported rates for Lugo (6.77). The clustering of small and high average rates suggests a positive spatial autocorrelation structure in the data, which is supported by highly significant results for Moran's I (Moran, 1950) index (restricted to continental provinces). See Figures 1 and 2 in the online supplementary material for the spatial distribution over all Spanish provinces and communities including the Spanish islands and African enclaves.

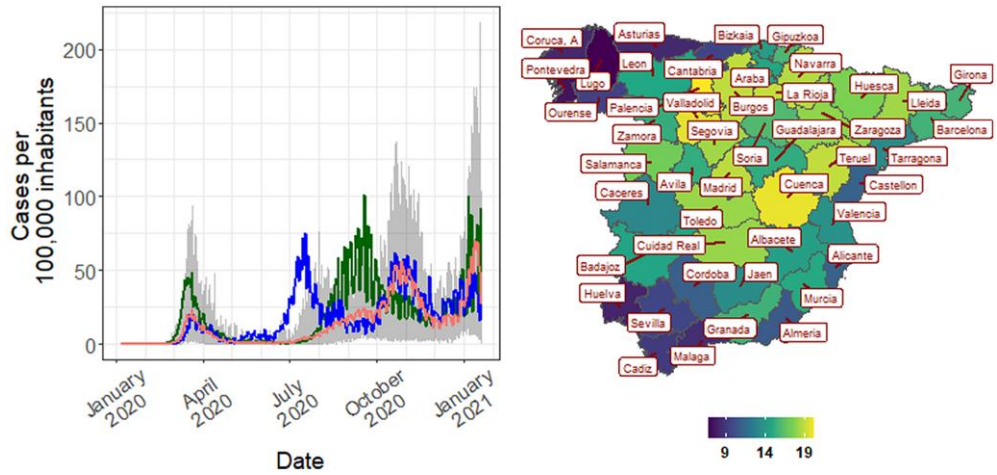

Figure 1. Distribution of daily Covid-19 cases per 100,000 inhabitants over time and space: (left) mean (red) and regional incidence for all provinces (grey), as well as for Madrid (green) and Lleida (blue) over 381 days, and (right) average incidence over 381 days at province level, restricted to continental Spain.

Figure 2 illustrates the patterns of the scalar, functional and (functional) compositional covariates. Both the averaged spatial patterns over time of mean temperature and solar exposure (left column of this plot) show a general increase from the northern to the southern parts of Spain. The temporal means for humidity and precipitation reflect some spatial heterogeneity with higher values computed for the northern areas contrasted with lower values for both variables at the Mediterranean coastline. Different from the other four weather variables, the maximum wind speed exhibits little spatial correlation, with the highest values reported for Gipuzkoa in the north. Over time (central column), both average temperature and sun hours show high values during the summer contrasted with low values in the winter. At the same time, humidity, precipitation and maximum wind speed reflect less clear temporal patterns. Although some variation and higher peaks are shown for humidity and wind speed in winter, spring and autumn compared to the summer period, the mean precipitation levels remain constantly at low values over time.

For the compositional covariates depicted in the right column of this plot we found a clear dominance of non-smokers over daily and ex-smokers in all 52 provinces, with occasional smokers (occ) constituting the smallest part of the regional populations. For the sex composition, a small dominance of females over males exists in all provinces. The normalized age curves computed from the age pyramids show a clear mode, with the largest population mass around 50 years. A second smaller mode and large variation can be seen for younger ages of around 10 years, while the densities decrease roughly monotonically and consistently across provinces for ages older than 55 years. Finally, the spatial patterns for the socio-demographic time-constant variables, shown in the bottom two right panels of Figure 2 reflect a clear spatial variation of the individual GDP (in 10,000 Euro), with lower values in the south contrasted with higher values in the northern provinces of Spain and especially Madrid. The population density shows a strong heterogeneity with the highest values for the metropolitan provinces of Madrid and Barcelona, but also the provinces of Bizkaia and Gipuzkoa.

4.3 Model specification for Spanish Covid-19 incidence curves

We include in our model for the Spanish Covid-19 incidence all variables with a possible effect (and an interaction for temperature and humidity, as low temperature and low humidity may interact to increase transferability Paez et al., 2021) according to the literature discussed in Section 2.2. We generally aimed for model parsimony given the limited amount of available data, but to include non-linear or time-varying model terms where there was some prior indication that time-constant linear terms might not be sufficient. As variables are selected based on expert opinion, we do not undertake any automated model selection, but do investigate several modelling

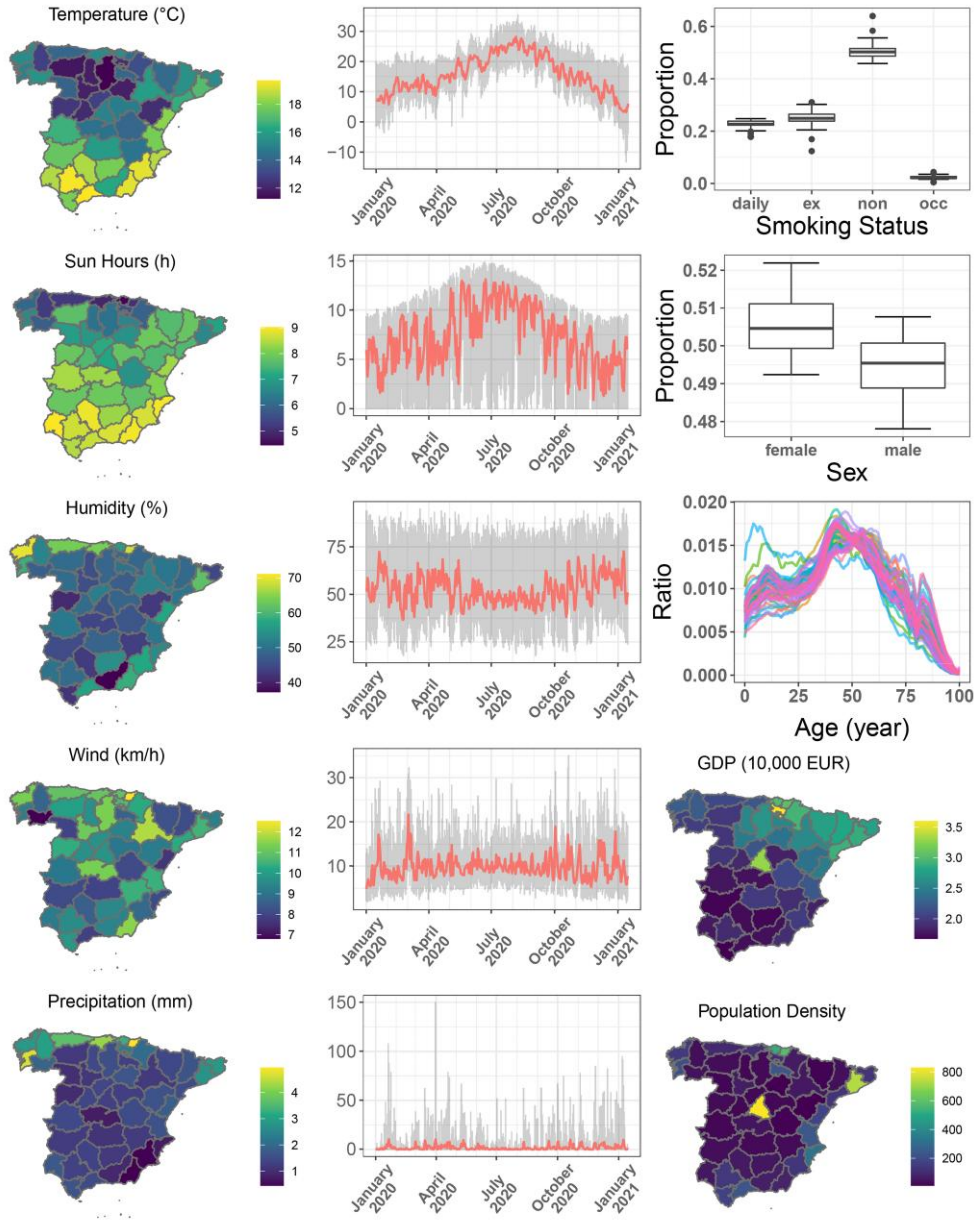

Figure 2. Distributional characteristics of the climatological, compositional, and socio-demographic covariates: temporally averaged spatial variation (left column), daily variation over time (central column, mean in red) of regional weather characteristics at province level, and regional variation of compositional and scalar covariates (right column).

choices in sensitivity analyses discussed below. Based on expert opinion, the aim of model parsimony and the inspection of the roughly time-constant effect patterns, we considered the three lockdown indicators $x_{ld_l,i}$, $l = 1, 2, 3$, the rain indicator $x_{\text{rain},i}$, the weekday indicators $x_{\text{day}_d,i}$, $d = 1, \ldots, 6$, the GDP $x_{\text{gdp},i}$ and the transport system indicator $x_{\text{tra},i}$ to have time-constant effects. In contrast, recalling the hypothesized impact of overcrowded areas on the disease transmission, in particular during the initial stages of the pandemic, and the strong variation in the size of the population over the different seasons for the coastal regions, we modelled the effect of the population density $x_{\text{dens},i}$ and the coastline indicator $x_{\text{sea},i}$ through linear time-varying

effects. For the 5-day lagged weather variables temperature ($x_{\text{temp},i}$), sun hours ($x_{sun,i}$), humidity ($x_{\text{hum},i}$) and wind speed ($x_{wind,i}$) and the log-transformed non-zero precipitation ($x_{\text{lprec},i}$), we also considered smooth, time-varying effect specifications, but as estimated effect surfaces of all covariates were roughly constant over time and monotone in the specific covariates, all weather covariates were specified to have time-constant smooth effects to reduce the complexity of the model. For precipitation, we used the effect form $x_{\text{rain},i}(t-5)(\beta_{\text{rain}} + f_5(x_{\text{lprec},i})(t-5))$, where $x_{\text{rain},i}$ is an indicator for positive precipitation amounts, $(t-5)$ indicates a 5-day lag and $x_{\text{rain},i}(t-5)f_5(x_{\text{lprec},i})(t-5)$ is defined to be zero if $x_{\text{rain},i}(t-5) = 0$, i.e. if there is no rain, with $f_5(x_{\text{lprec},i})(t-5)$ centred around the constant β_{rain}. This specification allows for a smooth continuous effect of the (log-)precipitation amount if positive, but a discontinuous difference between no rain at all and a small positive amount of precipitation. In addition to these terms, we also considered a smooth interaction of $x_{\text{temp},i}$ and $x_{\text{hum},i}$ to control for the interaction between these two terms reported in the literature. To account for the observed spatial correlation among the provinces, a spatially correlated functional random effect $\gamma_i(t)$ was included, using an MRF specification for the marginal basis $\boldsymbol{\Phi}_{xr}$. The structure for this MRF was derived from a Gabriel graph (Matula & Sokal, 1980) to control for the strong economic and social interrelations of continental Spain and the Spanish islands and African enclaves. In addition, we included independent smooth functional random intercepts $\gamma_{0,\text{com}_i}(t)$ for the 18 community spatial units to control for potential spatially nested effects and unobserved heterogeneity of the local Covid-19 measures on community level. For the compositional covariates, we included the effect of the smoker status composition $\mathbf{x}_{\text{smoke},i} = (x_{\text{daily},i}, x_{\text{occ},i}, x_{\text{ex},i}, x_{\text{non},i})^\top$ as a time-constant linear function-on-composition term (internally using the ilr transformation). The sex composition $\mathbf{x}_{\text{sex},i} = (x_{\text{male},i}, x_{\text{fem},i})^\top$ effect was modelled with a time-varying linear function-on-composition term to account for the strong heterogeneity in proportions of males and females within the public health and the nursing sectors—with a clear majority of female workers—which yielded high numbers of infected females already at the beginning of the pandemic. Finally, for the age densities $x_{\text{age},i}$ we considered a linear function-on-functional composition term (internally specified through a tensor product interaction smooth of the clr transformed age curves and time).

Combining these terms and writing $\boldsymbol{\kappa}_i = \{\boldsymbol{x}_{it}, t, \gamma_i(t), \gamma_{0,\text{com}_i}(t); t \in \mathcal{T}\}$, where the vector $\boldsymbol{x}_{it}$ contains all covariates at time t, and $W = \{\text{temp, sun, hum, wind}\}$, the expected number of Covid-19 cases $\mathbb{E}[Y_i(t) \mid \boldsymbol{\kappa}_i]$ for province i is specified through the following regression equation

$$
\begin{aligned}
\log\{\mathbb{E}(Y_i(t) \mid \boldsymbol{\kappa}_i)\} &= \log(N_i) + \beta_0(t) + x_{\text{rain},i}(t-5)\beta_{\text{rain}} + x_{\text{gdp},i}\beta_{\text{gdp}} + x_{\text{tra},i}\beta_{\text{tra}} \\
&\quad + x_{\text{sea},i}\beta_{\text{sea}}(t) + x_{\text{dens},i}\beta_{\text{dens}}(t) + \sum_{d=1}^{6} x_{\text{day}_d,i}\beta_{\text{day}_d} + \sum_{l=1}^{3} x_{ld_l,i}\beta_{ld_l} \\
&\quad + \sum_{k \in W} f_k(x_{k,i}(t-5)) + x_{\text{rain},i}(t-5)f_5(x_{\text{lprec},i}(t-5)) \\
&\quad + f_6(x_{\text{hum},i}(t-5), x_{\text{temp},i}(t-5)) + \gamma_i(t) + \gamma_{0,\text{com}_i}(t) \\
&\quad + \langle \mathbf{x}_{\text{smoke},i}, \boldsymbol{b}_{\text{smoke}} \rangle_A + \langle \mathbf{x}_{\text{sex},i}, \boldsymbol{b}_{\text{sex}}(t) \rangle_A + \langle x_{\text{age},i}, b_{\text{age}}(., t) \rangle_{B^2},
\end{aligned}
$$

using a log-link, where $\log(N_i)$ is an offset for the population size N_i in province i. To account for potential overdispersion of the response, we here assume a quasi-Poisson model for the Covid-19 incidences such that the variance is related to the mean through the overdispersion parameter ξ, i.e. $\mathbb{V}\text{ar}[Y_i(t) \mid x_{it}, t] = \mu_i(t)\xi$. We note that alternative suitable distributions include the negative binomial where the variance is a quadratic instead of a linear function of the mean as under the present quasi-Poisson specification, and the Conway–Maxwell–Poisson (COM), a flexible two parameter extension of the Poisson distribution which also allows for under-dispersion (Sellers & Premeaux, 2021; Shmueli et al., 2005). Although not considered here, we note that the proposed model also allows for COM distributions by extending the approach of Chatla and Shmueli (2018) to the present context.

We also fit a negative binomial model as a sensitivity analysis, with results given in the online supplementary material. Comparing the fit of both models (see Figures 7 and 8 in the online supplementary material), both models performed similarly. Where there are differences between fits, the quasi-Poisson tends to fit better for larger sites such as Barcelona or Granada, while the fit can be better for the negative binomial model for smaller regions such as Ceuta or Teruel. From a theoretical perspective, the negative binomial imposes a concave relation of the estimation weights to the means such that larger weights are assigned to smaller means while the weights are proportional to the means under the quasi-Poisson. In this respect, the quasi-Poisson appears to be more natural to estimate the global disease dynamics over all sites which we suspect to be dominated by the larger sites (Ver Hoef & Boveng, 2007).

All computations were performed in R version 4.1.1 (R Core Team, 2021), using in particular the `compositions` (van den Boogaart et al., 2021) and `refund` (Goldsmith et al., 2020) packages. The model was implemented using the `pffr()` function from the `refund` package. Basis sizes before application of possible constraints are as follows. The marginal basis functions of the smooth effects in the t direction were specified through penalized B-splines (Eilers & Marx, 1996) using $K_{tr} = 30$ knots (yielding around 1 knot per 13 days), and $K_{tr} = 28$ knots for the global functional intercept. Covariate effect marginal bases were chosen smaller due to the smoothness of effects: $K_{xr} = 10$ for the smooth effects of temperature and humidity, $K_{xr} = 9$ for sun hours and $K_{xr} = 7$ for wind speed. Tensor product interactions were specified with 5×5 knots, which applied to the smooth interaction of temperature and humidity as well as to the function-on-function effect for clr(age). Using an Intel(R) Core(TM) i-7 processor with 1.5 GHz and 16 GB RAM, the computation of the proposed model requires about 30 minutes with an effective 568.711 model degrees of freedom and a sample size of $n \times T = 52 \times 381 = 19{,}812$.

We now discuss the results of the proposed model for the Spanish Covid-19 data. We note that our analysis is based on the complete time frame provided by the geovoluntarios project, spanning from 5 January 2020 to 19 January 2021. While Covid-19 cases initially appeared in official surveillance reports at the end of February, the provided data were continuously updated post-hoc by previously undetected cases and also falsely assigned respiratory infections by the geovoluntarios project and is thus the best data available. To account for the remaining inherent uncertainty in the true disease onset and early disease numbers in Spain, we additionally run as a sensitivity analysis two separate models starting on 1 February 2020 and 1 March 2020. A detailed description of the corresponding results for both reduced time frames is provided in the online supplementary material of this paper. Almost all effects remain similar to the ones on the full data, with the notable exception of the lockdown effects, which can change sign and significance when using data from different time frames. This indicates that results for the lockdown effects should be interpreted with care, as lockdown effects are difficult to disentangle from overall smooth time trends, especially when little data before the lockdown is available and when effects are likely more complex than is possible to capture with a constant effect during lockdown time periods. Under the above specification, the model explains 96% of the deviance. The estimated dispersion parameter under the quasi-Poisson specification is 15.1. Table 1 shows the estimated covariate effects which are treated as time constant, where the reported p-values are based on a Wald test for the null hypothesis that each parametric term is zero (Wood, 2017). All effects except for the indicators for the second and third global lockdown periods and the second ilr component are significant at a significance level of $\alpha = 0.05$. We found a negative effect of the first global lockdown period (Lockdown 1), indicating a reduction by around 12% of the daily numbers of Covid-19 cases by the imposed measures (compared to the trend under no lockdown), which however has to be interpreted with care as discussed above.

The weekday effects show a clear positive impact on the numbers of Covid-19 notifications, which is similar for Monday through Friday and smaller for Saturday, compared to Sunday. This heterogeneity over the weekdays may be due to daily variation in the availability at local authority levels and of tests. In line with the findings of Paez et al. (2021), the coefficient for *transport* suggests an increase in Covid-19 cases by around 56% if higher order transit systems are available. The expected incidence decreases with increasing GDP by around 31% per 10,000 EUR, ceteris paribus. Lastly, the rain indicator (representing days with non-zero levels of precipitation using a 5-day lag) shows a positive effect, leading to around 5% more Covid-19 cases after rainy days.

Table 1. Estimated time-constant linear effects in the functional generalized additive model for the Spanish Covid-19 incidence and corresponding Wald test results

	β	exp(β)	Standard error	Test statistic	p-Value
Intercept	3.44	31.12	1.30	2.65	0.008
Lockdown 1	−0.13	0.88	0.04	−2.96	0.003
Lockdown 2	−0.02	0.98	0.14	−0.16	0.875
Lockdown 3	0.04	1.05	0.13	0.35	0.724
ilr(smoke 1)	−5.13	0.01	1.73	−2.98	0.002
ilr(smoke 2)	−1.09	0.34	1.23	−0.89	0.374
ilr(smoke 3)	−5.79	0.00	1.96	−2.95	0.003
Monday	0.34	1.41	0.01	33.81	0.000
Tuesday	0.41	1.51	0.01	41.00	0.000
Wednesday	0.37	1.45	0.01	36.11	0.000
Thursday	0.33	1.39	0.01	32.00	0.000
Friday	0.38	1.47	0.01	38.09	0.000
Saturday	0.14	1.15	0.01	13.09	0.000
Transport	0.45	1.56	0.01	29.53	0.000
GDP	−0.37	0.69	0.02	−15.29	0.000
Rain	0.05	1.05	0.01	4.47	0.000

Figure 3 shows the functional intercept and the linear functional (time-varying) effects of the scalar covariates. The functional intercept (left) has its highest peak during the second Covid-19 wave, with maximum numbers of Covid-19 infections in mid-September.

The effect for the population density at province level (centre) reflects a clear positive impact on the expected number of infection notifications, with the strongest impact during the early stages of the first wave up to mid-March 2020. This finding is consistent with the association of densely crowded areas with the spread of the disease stated in the literature. The effect of *coast* (right) suggests smaller incidences for coastal compared to non-coastal provinces, in particular starting from mid-April 2020 onwards and reaching a minimum at around mid-July. We hypothesize that this negative effect might be explained by an increased public risk awareness and protective travelling behaviour caused by the aftermaths of the recent Covid-19 and lockdown experiences. Indeed, facing the massive impact of the disease on the Spanish population and health system during the first wave, overcrowded regions including the coast and metropolitan conurbations suffered a larger exodus of the population, and rural areas and the countryside became a favourite travelling destination. In addition, imposed national travelling restrictions and strict quarantine regulations for incoming and/or homecoming travellers yielded a strong reduction in numbers of international tourists and travellers. In a recent paper, Sun et al. (2021) reported a reduction of global scheduled flights for Spain by over 90% for April to June 2020 compared to those month in 2019, which decreased to 65.7% for July 2020. The observed negative effect of coast on the number of Covid-19 cases slowly vanishes towards the fall and winter of 2020, which could potentially be due to less protective individual travelling behaviour.

4.4 Concurrent functional effects of weather on Covid-19 cases

The estimated non-linear time-constant concurrent effects of the lagged weather covariates and the interaction surface for the lagged mean temperature and lagged humidity are depicted in Figure 4.

For the lagged mean temperature (upper left), we found a negative effect of higher temperatures on the expected number of cases, which is consistent with Wu et al. (2020) and Paez et al. (2021). The non-linear effect for the lagged sun hours (upper central) only shows small positive and negative departures from zero, with confidence bands indicating high uncertainty especially for large

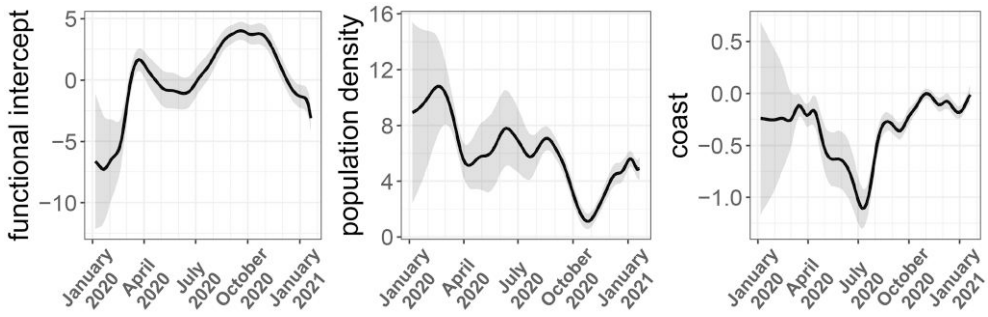

Figure 3. Functional intercept and covariate effects: Smooth effects on expected number of daily Covid-19 cases. Functional intercept (left) and linear functional (time-varying) effects of the scalar covariates population density (central), and coastline (right). Effects are given on the predictor level (on the log-mean) and point-wise 95% confidence bands are shaded in grey.

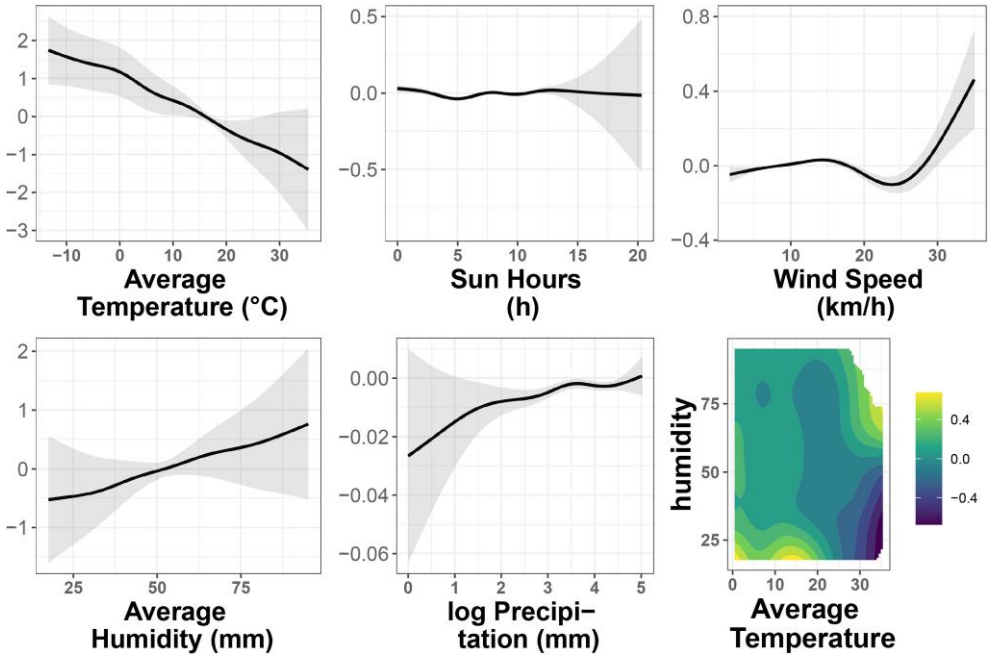

Figure 4. Non-linear time-constant concurrent effects of different weather characteristics on the log-mean number of daily Covid-19 cases considering a 5-day lag. Upper panels: effects for lagged mean temperature, average sun hours, and maximum wind speed on the log-mean number of daily cases. Lower panels: effects for the average lagged humidity, log-transformed non-zero levels of the precipitation variable, and interaction effect surface for lagged mean temperature and lagged humidity (including main effect functions of temperature and humidity).

values above 12 hr. The non-linear effect for maximum wind speed (upper right) shows a small monotone increase in excepted daily notifications until around 15 km/h, with a decrease and increase for higher wind speeds becoming increasingly uncertain due to small numbers. The average humidity and the log-transformed non-zero precipitation values (lower left and central panels) show a roughly linearly increasing effect on the expected incidence. Finally, the interaction surface of the lagged average temperature and the lagged humidity (lower right panel) suggests a positive effect for low temperatures and low levels of humidity on the spread of the disease dynamics, contrasted with a negative impact of high temperatures and low humidity levels. The interaction indicates that smooth main effects of temperature and humidity should be interpreted with care, as

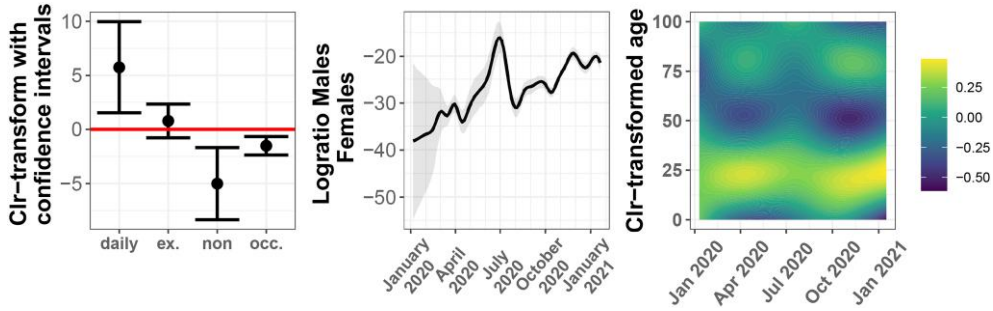

Figure 5. Effects of the compositional covariates smoking status (left, clr-transformed), sex (middle, ilr-transformed), and age (right, clr-transformed) on the log-mean number of Covid-19 cases.

they average over parts of the interaction surface that depend on the data range. This might potentially explain the mixed results on the effects of climate variables on the spread of the disease observed in different climatological regions, and is also seen in our sensitivity analysis below.

4.5 Compositional effect of smoking behaviour, sex and age on Covid-19 cases

The effects of the compositional covariates on the expected number of daily notifications are shown in Figure 5. To interpret the time-constant effect of the individual smoking habits, we obtained the simplicial gradient via inverse ilr transformation of the corresponding coefficients on ilr-level, see Table 1, indicating that the largest increase in the expected incidence is obtained by increasing the proportion of smokers [see online supplementary material, simplicial gradient roughly (1, 0, 0, 0)].

Applying the clr transformation to the simplicial gradient (see Section 3.2.1) allows to evaluate the effect of a multiplicative change in the relative ratio of one component while holding all other ratios constant. Depicted as sum-to-zero constrained effect estimates for the clr-transformed composition with corresponding 95% confidence intervals, we found a clear positive effect of a larger fraction of daily smokers on the disease incidence, contrasted with a negative effect of a larger fraction of occasional and in particular non-smokers (see left panel) which is in line with the results of Hopkinson et al. (2021) and Gülsen et al. (2020). A relative ratio increase by 10% for daily smokers yields a multiplicative increase in the expected daily Covid-19 incidence by the factor $1.1^{5.747} = 1.729$, i.e. by 73%. An analogous 10% relative ratio increase for non-smokers ($1.1^{-5.009}$) yields a 38% decrease in the expected incidence.

The estimated effect of the sex log-ratio (central panel) suggests a negative effect of an increase in the male-to-female ratio on the mean Covid-19 counts, particularly early in the pandemic. A possible explanation could be the described heterogeneity among the sexes in terms of employment in high-contact jobs such as in the retail and medical fields.

The right panel shows the effects of the clr-transformed age compositions on the disease dynamics, with estimated effect surface constrained to fulfil an integration-to-zero constraint, $\beta(., t) \in L_0^2(\mathcal{S})$, for each time point t. The effect surface shows clear variation over time and over the different ages. The strongest positive effects on the number of Covid-19 cases appear for the younger and also for the very old parts of the population, with a clear mode for around 25 year olds. To interpret the effect of the age distribution on the incidence, we applied the inverse clr transformation to the estimated surface $\beta(., t) \in L_0^2(\mathcal{S})$ for each t, to obtain for each time point the direction $b(., t) \in B^2(\mathcal{S})$ of change in the age composition leading to the largest increase in the mean incidence analogously to equation (2). Inspecting the time trend of $b(., t)$ depicted in Figure 5 of the online supplementary material, all age curves show a clear mode for the younger ages and a second, but smaller, mode for around 80 year olds, with small variations in the exact density shape over time. This suggests that provinces with high proportions of young people (and to a lesser extend old people) are more strongly affected by Covid-19 cases (see Sections 2.2 and 2.3 of the online supplementary material for a more detailed discussion of the results including computations of changes in the mean response for different changes in the age distribution).

4.6 Spatio-temporal effects

A discussion of the results for both the spatially correlated functional random intercepts per province and the spatially uncorrelated community-specific functional random intercepts is given in online supplementary material. Both effects exhibit some variation in sign and effect size over the 52 provinces and 19 communities, respectively (see Figure 4 in the online supplementary material).

4.7 Model diagnostics, sensitivity analyses and evaluation in simulations

Comparing the fitted and the observed incidence curves (Figure 7 in the online supplementary material) shows only small deviations of the fitted from the observed daily values over the study period. The scaled Pearson residuals and the autocorrelation (Figure 9 in the online supplementary material) suggest overall good model fit, with some amount of heterogeneity in variation and autocorrelation of the residuals, in particular some structure corresponding to the three Covid-19 waves, remaining.

We also performed a range of sensitivity analyses to assess the effects of considering different lags (4-, 6-, 7-, and 8-day) for the weather covariates, a different spatial neighbourhood specification, a negative binomial response distribution, or a different time window leaving out early more uncertain incidences. Overall, results were largely similar and general conclusions did not change with the exception of the lockdown effects discussed above (see online supplementary material for a detailed discussion). Note that while the interaction surface for temperature and humidity was stable under different model specifications and datasets, the main effects did change when considering continental Spain only or data beginning 1 March 2020 only, due to differences in the variable distribution over which the main effects average the interaction surface.

Finally, we conducted a simulation study to evaluate estimation performance (see Section 4 in the online supplementary material). We simulated 500 datasets mimicking our Covid-19 data, based on the estimated negative binomial model, to focus on the new model terms compared to (extensive simulations in) Scheipl et al. (2016) and on how well model terms can be estimated for our given model complexity and data size. All effects were recovered well, with no or little bias. Variability was usually small, while somewhat larger for complex interaction terms, and well captured by the confidence intervals/bands estimated on the Covid-19 data. This confirms that our proposed model extension can identify the effect of compositional and functional compositional covariates in addition to different spatial, functional and scalar model terms.

5 Conclusion

This paper has extended the GFAMM to the case when some of the covariates in the predictor are finite or infinite (functional) compositions summing or integrating to a whole. We use an equivalence between the scalar product in the Bayes Hilbert space and a constrained linear (functional) term for the clr-transformed compositional covariate to embed the new model terms into the existing model framework. For the transformed functional composition, the linear effect was modelled with a tensor product basis with a bivariate spline for the effect function, placing centring constraints on the marginal basis for the covariate effect to account for the integration-to-zero constraint. We also discuss interpretation of the next effect. Although not considered here due to the increased model complexity given the sample size, the proposed model in principle also allows to include non-linear effects of the transformed (functional) covariates.

The proposed model was applied to spatially correlated daily Covid-19 count data for Spain to quantify potential impacts of population compositional, socio-economic, weather, and regional covariates on the disease dynamics. The information at hand was retrieved from various source, including the state meteorology agency, the ministry of agriculture, fisheries, and food, the National Statistics Institute and the a collaborative data collection project by different geodata providers. The sampled data were collected from 5 January 2020 to 19 January 2021, just before a large-scale nationwide immunization programme was imposed in February 2021, which minimizes unknown effects on our results of the regionally varying and heterogeneous vaccination regimes. We note that, although the analysis was restricted to a pre-vaccination setting, daily information on the proportions of non-/partly/fully vaccinated people per region if available could be included into the proposed regression framework in the form of an additional compositional

covariate. The available data have some limitations. First, the reported numbers on a given day likely represent a mixture of counts over neighbouring (unobserved) true dates of symptom onset, given that the reconstruction of symptom onset dates by a 3-day lag from the positive test results is only an approximation. There is even more uncertainty regarding the true infection date, even if the average incubation period is 5 days. We may thus underestimate the weather effects, if the lagged weather variables only approximately measure weather on the date of infection. However, we did not detect large variations of the estimated results in our sensitivity analysis considering alternative lag specifications. Second, the data do not provide separate infection counts for subgroups of the population according to sex, age and smoking habits. While we incorporate the effects of these variables on overall infections via compositional covariates measuring the composition of the population, we have to acknowledge the typical risks of ecological inference here. For instance, for the increasing effect of a larger share of smokers in the population on the Covid-19 incidence, we cannot distinguish whether this is due to a higher infection risk for smokers or due to a higher risk of Covid-19 positive smokers to infect others. Third, while the compiled data for 52 provinces and 381 days have better temporal and spatial resolution than other publicly available datasets, the data size still limits the complexity of the model in terms of the number of non-linear and/or time-varying effects. Taking these limitations into account, our model highlights a clear effect of the population composition according to sex, age and smoking habits, of weather variables (rain, temperature, wind speed and humidity), of GDP, population density, coast and public transit on the number of Covid-19 notifications, as well as spatial and temporal heterogeneity.

Acknowledgments

The authors gratefully thank Dr. Fabian Scheipl for his helpful comments on the *refund* R package. Further, the authors acknowledge support by public health experts from the Spanish Statistical Institute (INE, https: // www. ine. es/ covid/ covid_ inicio. htm) and the National Institute of Health Carlos III (ISCIII) (see https: // www. isciii. es/ Paginas/ Inicio. aspx).

Conflict of interest: None declared.

Funding

The authors gratefully acknowledge financial support by the German Research Association and the Spanish Ministry of Science and Innovation. J.M. was funded by grant PID2022-141555OB-I00 from the Spanish Ministry of Science and Innovation. S.G. was funded by grant GR 3793/8-1 from the German Research Foundation.

Data availability

The R code and data used in the real data applications are made publicly available in a github repository https://github.com/MatkcE/CoDaGFAMM.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

Aitchison J. (1986). *The statistical analysis of compositional data*. Chapman & Hall.

Arata Y. (2017). A functional linear regression model in the space of probability density functions (*Discussion Papers 17015*). Research Institute of Economy, Trade and Industry (RIETI).

Barceló-Vidal C., Martín-Fernández J. A., & Mateu-Figueras G. (2011). *Compositional differential calculus on the simplex* (Chap. 13, pp. 176–190). John Wiley & Sons, Ltd.

Chatla S. B., & Shmueli G. (2018). Efficient estimation of COM-Poisson regression and a generalized additive model. *Computational Statistics & Data Analysis*, *121*, 71–88. https://doi.org/10.1016/j.csda.2017.11.011

Coma Redon E., Mora N., Prats-Uribe A., Fina Avilés F., Prieto-Alhambra D., & Medina M. (2020). Excess cases of influenza and the coronavirus epidemic in catalonia: A time-series analysis of primary-care electronic

medical records covering over 6 million people. *BMJ Open*, *10*(7), e039369. https://doi.org/10.1136/bmjopen-2020-039369
Congdon P. (2022). A spatio-temporal autoregressive model for monitoring and predicting COVID infection rates. *Journal of Geographical Systems*, *24*(4), 583–610. https://doi.org/10.1007/s10109-021-00366-2
Crimmins E. M. (2020). Age-related vulnerability to coronavirus disease 2019 (Covid-19): Biological, contextual, and policy-related factors. *Public Policy & Aging Report*, *30*, 142–146. https://doi.org/10.1093/ppar/praa023
Du P., Li D., Wang A., Shen S., Ma Z., & Li X. (2021). A systematic review and meta-analysis of risk factors associated with severity and death in Covid-19 patients. *Canadian Journal of Infectious Diseases and Medical Microbiology*, *2021*, 6660930. https://doi.org/10.1155/2021/6660930
Egozcue J. J., Díaz–Barrero J. L., & Pawlowsky-Glahn V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica, English Series*, *22*(4), 1175–1182. https://doi.org/10.1007/s10114-005-0678-2
Egozcue J. J., Pawlowsky-Glahn V., Mateu-Figueras G., & Barceló-Vidal C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*(3), 279–300. https://doi.org/10.1023/A:1023818214614
Eilers P. H. C., & Marx B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–121. https://doi.org/10.1214/ss/1038425655
Fišerová E., & Hron K. (2011). On the interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, *43*(4), 455–468. https://doi.org/10.1007/s11004-011-9333-x
Gertheiss J., Goldsmith J., & Staicu A.-M. (2017). A note on modeling sparse exponential-family functional response curves. *Computational Statistics & Data Analysis*, *105*, 46–52. https://doi.org/10.1016/j.csda.2016.07.010
Goldsmith J., Scheipl F., Huang L., Wrobel J., Di C., Gellar J., Harezlak J., McLean M. W., Swihart B., Xiao L., Crainiceanu C., & Reiss P. T. (2020). *refund: Regression with functional data*. R package version 0.1-22. https://CRAN.R-project.org/package=refund.
Goldsmith J., Zipunnikov V., & Schrack J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, *71*(2), 344–353. https://doi.org/10.1111/biom.12278
Greven S., & Scheipl F. (2017a). A general framework for functional regression modelling. *Statistical Modelling*, *17*(1–2), 1–35. https://doi.org/10.1177/1471082X16681317
Greven S., & Scheipl F. (2017b). Rejoinder (for a general framework for functional regression modelling). *Statistical Modelling*, *17*(1-2), 100–115. https://doi.org/10.1177/1471082X16689188
Gülsen A., Yigitbas B. A., Uslu B., Drömann D., & Kilinc O. (2020). The effect of smoking on Covid-19 symptom severity: Systematic review and meta-analysis. *Pulmonary Medicine*, *2020*, 7590207. https://doi.org/10.1155/2020/7590207
Han K., Müller H.-G., & Park B. U. (2020). Additive functional regression for densities as responses. *Journal of the American Statistical Association*, *115*(530), 997–1010. https://doi.org/10.1080/01621459.2019.1604365
Happ C., Scheipl F., Gabriel A.-A., & Greven S. (2019). A general framework for multivariate functional principal component analysis of amplitude and phase variation. *Stat*, *8*(1), e220. https://doi.org/10.1002/sta4.220
Henríquez J., Gonzalo-Almorox E., García-Goñi M., & Paolucci F. (2020). The first months of the Covid-19 pandemic in Spain. *Health Policy and Technology*, *9*(4), 560–574. https://doi.org/10.1016/j.hlpt.2020.08.013
Hopkinson N. S., Rossi N., El-Sayed Moustafa J., Laverty A. A., Quint J. K., Freidin M., Visconti A., Murray B., Modat M., Ourselin S., Small K., Davies R., Wolf J., Spector T. D., Steves C. J., & Falchi M. (2021). Current smoking and Covid-19 risk: Results from a population symptom app in over 2.4 million people. *Thorax*, *76*(7), 714–722. https://doi.org/10.1136/thoraxjnl-2020-216422
Hossain M. S., Ahmed S., & Uddin M. J. (2021). Impact of weather on Covid-19 transmission in south Asian countries: An application of the ARIMAX model. *Science of the Total Environment*, *761*, 143315. https://doi.org/10.1016/j.scitotenv.2020.143315
Hron K., Menafoglio A., Templ M., Hřzová K., & Filzmoser P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, *94*, 330–350. https://doi.org/10.1016/j.csda.2015.07.007
Linton N. M., Kobayashi T., Yang Y., Hayashi K., Akhmetzhanov A. R., Jung S.-m., Yuan B., Kinoshita R., & Nishiura H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine*, *9*(2),538. https://doi.org/10.3390/jcm9020538
Machalová J., Talská R., Hron K., & Gába A. (2021). Compositional splines for representation of density functions. *Computational Statistics*, *36*(2), 1031–1064. https://doi.org/10.1007/s00180-020-01042-7
Maier E., Stöcker A., Fitzenberger B., & Greven S. (2021). 'Additive density-on-scalar regression in Bayes Hilbert spaces with an application to gender economics', arXiv, arXiv:2109.xxx, preprint: not peer reviewed.
Matula D. W., & Sokal R. R. (1980). Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane. *Geographical Analysis*, *12*(3), 205–222. https://doi.org/10.1111/gean.1980.12.issue-3

Mazzoli M., Mateo D., Hernando A., Meloni S., & Ramasco J. J. (2020). Effects of mobility and multi-seeding on the propagation of the covid-19 in spain. *medRxiv*. https://www.medrxiv.org/content/early/2020/05/18/2020.05.09.20096339

McClymont H., & Hu W. (2021). Weather variability and Covid-19 transmission: A review of recent research. *International Journal of Environmental Research and Public Health*, *18*(2), 396. https://doi.org/10.3390/ijerph18020396

Mecenas P., Bastos R. T. D. R. M., Vallinoto A. C. R., & Normando D. (2020). Effects of temperature and humidity on the spread of Covid-19: A systematic review. *PLoS One*, *15*(9), e0238339–21. https://doi.org/10.1371/journal.pone.0238339

Moosa I., & Khatatbeh I. (2021). The density paradox: Are densely-populated regions more vulnerable to Covid-19? *The International Journal of Health Planning and Management*, *36*(5), 1575–1588. https://doi.org/10.1002/hpm.v36.5

Moran P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1–2), 17–23. https://doi.org/10.1093/biomet/37.1-2.17

Moritz S., & Bartz-Beielstein T. (2017). imputeTS: Time series missing value imputation in R. *The R Journal*, *9*(1), 207–218. https://doi.org/10.32614/RJ-2017-009

Morris J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, *2*(1), 321–359. https://doi.org/10.1146/statistics.2015.2.issue-1

Morris J. S. (2017). Comparison and contrast of two general functional regression modelling frameworks. *Statistical Modelling*, *17*(1–2), 59–85. https://doi.org/10.1177/1471082X16681875

Morris J. S., & Carroll R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(2), 179–199. https://doi.org/10.1111/j.1467-9868.2006.00539.x

Nelder J. A., & Pregibon D. (1987). An extended quasi-likelihood function. *Biometrika*, *74*(2), 221–232. https://doi.org/10.1093/biomet/74.2.221

Paez A., Lopez F. A., Menezes T., Cavalcanti R., & Pitta M. G. D. R. (2021). A spatio-temporal analysis of the environmental correlates of Covid-19 incidence in Spain. *Geographical Analysis*, *53*(3), 397–421. https://doi.org/10.1111/gean.v53.3

Park J. Y., & Qian J. (2012). Functional regression of continuous state distributions. *Journal of Econometrics*, *167*(2), 397–412. https://doi.org/10.1016/j.jeconom.2011.09.024

Pawlowsky-Glahn V., & Egozcue J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, *15*(5), 384–398. https://doi.org/10.1007/s004770100077

Petersen A., & Müller H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, *47*(2), 691–719. https://doi.org/10.1214/17-AOS1624

Petersen A., Zhang C., & Kokoszka P. (2021). Modeling probability density functions as data objects. *Econometrics and Statistics*, *21*, 159–178. https://doi.org/10.1016/j.ecosta.2021.04.004

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Ramsay J., & Silverman B. (1997). *Functional data analysis*. Springer.

Scheipl F., Gertheiss J., & Greven S. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics*, *10*, 1455–1492. https://doi.org/10.1214/16-EJS1145

Scheipl F., Staicu A., & Greven S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, *24*(2), 477–501. https://doi.org/10.1080/10618600.2014.901914

Scimone R., Menafoglio A., Sangalli L., & Secchi P. (2021). *A look at the spatio-temporal mortality patterns in Italy during the Covid-19 pandemic through the lens of mortality densities* (Technical Report MOX-Report No. 27/2021). Dipartimento di Matematica, Politecnico di Milano, Miland.

Sellers K. F., & Premeaux B. (2021). Conway–Maxwell–Poisson regression models for dispersed count data. *WIREs Computational Statistics*, *13*(6), e1533. https://doi.org/10.1002/wics.1533

Shahzad K., Shahzad U., Iqbal N., Shahzad F., & Fareed Z. (2020). Effects of climatological parameters on the outbreak spread of Covid-19 in highly affected regions of Spain. *Environmental Science and Pollution Research*, *27*(31), 39657–39666. https://doi.org/10.1007/s11356-020-10551-3

Shmueli G., Minka T. P., Kadane J. B., Borle S., & Boatwright P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *54*(1), 127–142. https://doi.org/10.1111/j.1467-9876.2005.00474.x

Sierra C., Ordóñez C., Saavedra A., & Gallego J. (2015). Element enrichment factor calculation using grain-size distribution and functional data regression. *Chemosphere*, *119*, 1192–1199. https://doi.org/10.1016/j.chemosphere.2014.10.024

Soriano V., & Barreiro P. (2020). Why such excess of mortality for Covid-19 in Spain? *Therapeutic Advances in Infectious Disease*, *7*, 2049936120932755. https://doi.org/10.1177/2049936120932755

Sun X., Wandelt S., Zheng C., & Zhang A. (2021). Covid-19 pandemic and air transportation: Successfully navigating the paper hurricane. *Journal of Air Transport Management*, *94*, 102062. https://doi.org/10.1016/j.jairtraman.2021.102062

Sun Z., Xu W., Cong X., Li G., & Chen K. (2020). Log-contrast regression with functional compositional predictors: Linking preterm infants' gut microbiome trajectories to neurobehavioral outcome. *The Annals of Applied Statistics*, *14*(3), 1535–1556. https://doi.org/10.1214/20-AOAS1357

Takagi H., Kuno T., Yokoyama Y., Ueyama H., Matsushiro T., Hari Y., & Ando T. (2020). Higher temperature, pressure, and ultraviolet are associated with less Covid-19 prevalence: Meta-regression of Japanese prefectural data. *Asia Pacific Journal of Public Health*, *32*(8), 520–522. https://doi.org/10.1177/1010539520947875

Talská R., Hron K., & Grygar T. M. (2021). Compositional scalar-on-function regression with application to sediment particle size distributions. *Mathematical Geosciences*. https://doi.org/10.1007/s11004-021-09941-1

Talská R., Menafoglio A., Machalová J., Hron K., & Fišerová E. (2018). Compositional regression with functional response. *Computational Statistics & Data Analysis*, *123*, 66–85. https://doi.org/10.1016/j.csda.2018.01.018

Tiruneh S. A., Tesema Z. T., Azanaw M. M., & Angaw D. A. (2021). The effect of age on the incidence of Covid-19 complications: A systematic review and meta-analysis. *Systematic Reviews*, *10*(1), 80. https://doi.org/10.1186/s13643-021-01636-2

van den Boogaart K. G., Egozcue J. J., & Pawlowsky-Glahn V. (2014). Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, *56*(2), 171–194. https://doi.org/10.1111/anzs.2014.56.issue-2

van den Boogaart K. G., Tolosana-Delgado R., & Bren M. (2021). *compositions: Compositional data analysis*. R package version 2.0-2. https://CRAN.R-project.org/package=compositions.

Ver Hoef J. M., & Boveng P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, *88*(11), 2766–2772. https://doi.org/10.1890/07-0043.1

Verbelen R., Antonio K., & Claeskens G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *67*(5),1275–1304. https://doi.org/10.1111/rssc.12283

Wolff D., Nee S., Hickey N. S., & Marschollek M. (2021). Risk factors for Covid-19 severity and fatality: A structured literature review. *Infection*, *49*(1), 15–28. https://doi.org/10.1007/s15010-020-01509-1

Wood S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Boca Raton.

Wood S. N., Pya N., & Säfken B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Wu Y., Jing W., Liu J., Ma Q., Yuan J., Wang Y., Du M., & Liu M. (2020). Effects of temperature and humidity on the daily new cases and new deaths of Covid-19 in 166 countries. *Science of the Total Environment*, *729*, 139051. https://doi.org/10.1016/j.scitotenv.2020.139051