Received: 6 August 2020 | Accepted: 31 August 2021

DOI: 10.1111/rssa.12753

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

On the interplay of regional mobility, social connectedness and the spread of COVID-19 in Germany

Cornelius Fritz | **Göran Kauermann**

Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany

Correspondence
Cornelius Fritz, Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539, Munich, Germany.
Email: cornelius.fritz@stat.uni-muenchen.de

Funding information
European Cooperation in Science and Technology [COST Action CA15109 (COSTNET)]; Munich Center for Machine Learning (MCML); German Federal Ministry of Education and Research (BMBF), Grant/Award Number: 01IS18036A

Abstract

Since the primary mode of respiratory virus transmission is person-to-person interaction, we are required to reconsider physical interaction patterns to mitigate the number of people infected with COVID-19. While research has shown that non-pharmaceutical interventions (NPI) had an evident impact on national mobility patterns, we investigate the relative regional mobility behaviour to assess the effect of human movement on the spread of COVID-19. In particular, we explore the impact of human mobility and social connectivity derived from Facebook activities on the weekly rate of new infections in Germany between 3 March and 22 June 2020. Our results confirm that reduced social activity lowers the infection rate, accounting for regional and temporal patterns. The extent of social distancing, quantified by the percentage of people staying put within a federal administrative district, has an overall negative effect on the incidence of infections. Additionally, our results show spatial infection patterns based on geographical as well as social distances.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. Journal of the Royal Statistical Society: Series A (Statistics in Society) published by John Wiley & Sons Ltd on behalf of Royal Statistical Society

KEYWORDS

COVID-19, infectious disease modelling, semiparametric regression, social connectedness, social networks, spatial network data

1 | INTRODUCTION

The COVID-19 virus outbreak originating in mainland China leapt over to Europe and quickly evolved to a global pandemic in March 2020. Only through strict non-pharmaceutical interventions (NPI) could most national health systems rapidly react to this new threat. In numerous scientific efforts, physical distancing measures were discovered to be the most effective interventions (Prem et al., 2020) and found to be necessary maybe until 2022 (Kissler et al., 2020). The measures' effectiveness emanates from researchers confirming that the main form of virus transmission is person-to-person interaction (Chan et al., 2020). The virus can be spread by inhaling microscopic aerosol particles that contain COVID-19 and remain viable in the air with a half-life of about 1 h (Asadi et al., 2020) or direct contact through the exchange of virus-containing droplets with infected individuals (Guan et al., 2020). Since also a high proportion of cases is asymptomatic (Lavezzo et al., 2020) and gets infected by cases in the presymptomatic stage (Li et al., 2020b), human mobility can explain the spread of COVID-19 to a considerable extent (Kraemer et al., 2020).

Stemming from the consequential need to account for contact patterns when investigating the spread of COVID-19, Oliver et al. (2020) list multiple possibilities of how one may utilise mobile phone data to do so. To enable this type of research, Facebook extended the *Data for Good* program to a broader audience of researchers and provided so-called *Disease Prevention Maps* for multiple countries (Maas et al., 2019). This database includes measurements on quantities like co-location, user counts and movement ranges on a regional level derived from information of more than 26 million Facebook users. Additionally, a measure for the social connectedness between geographical regions is supplied (Bailey et al., 2018). In various studies, this data source was employed to demonstrate how the impact of lockdown measures in Italy was more severe for municipalities with higher fiscal capacities (Bonaccorsi et al., 2020), quantify social and geographical spillover effects from relaxations of shelter-in-place orders (Holtz et al., 2020) and predict the number of infections on a granular spatiotemporal resolution using contact tracing data (Lorch et al., 2020).

This article uses the same data source to analyse how regional differences in mobility patterns and friendship proximity affect the spread of COVID-19 in Germany. While NPIs, for example, the nationwide shutdown in Germany that started 22 March, had an evident impact on national human mobility and ceased the exponential spread of the virus (Flaxman et al., 2020), the effect of the relative movement between regional districts was not yet fully assessed. So far, studies concerning human movements during the current pandemic are focused mainly on how the lockdown affected national human mobility (Galeazzi et al., 2020) or specific regions regarding their economic status (Bonaccorsi et al., 2020). To fill this gap, we derive covariates from the mobility data to quantify the overall dispersion of meeting patterns and compliance with social distancing. Through weekly standardisation of the covariates, we control for the dynamics therein, which are, in turn, driven by NPIs. As a result, our research enables a quantitative assessment of different mobility strategies relative to the

national average. Also, we infer positions of the federal administrative regions in a social space from the information on the relative friendship links among them using multidimensional scaling (Cox & Cox, 2000). Subsequently, we relate the processed data to Germany's weekly rate of local COVID-19 infections between 3 March and 22 June 2020. This time frame permits the analysis of the dynamic spread starting with the WHO declaring COVID-19 a pandemic (WHO, 2020).

We employ a spatiotemporal regression model for the ratio of local COVID-19 infections that takes autoregressive structures, age- and gender-specific effects, contagion by nearby districts in the geographical and social space, as well as latent heterogeneities between the districts into account. Our method is closely related to the surveillance model introduced by Held et al. (2005). They extend generalised linear models to analyse surveillance data from epidemic outbreaks. This approach was expanded to handle multivariate surveillance data (Paul et al., 2008), control for seasonality and spatial heterogeneity (Held & Paul, 2012) and include neighbourhood information from social contact data (Meyer & Held, 2017). In contrast to this type of model, our model's objective is to investigate the connection between mobility patterns, social connectivity and the spread of COVID-19 in an interpretable manner. While forecasting infections is undoubtedly a central objective in epidemic surveillance, this is not the main focus of our work (see also Held et al., 2017).

The rest of the article will be structured as follows: We discuss the data sources, its measures on social interaction as well as mobility in Section 2. In Section 3, we detail our proposed modelling approach. We propose an imputation model for missing onset dates and use a semiparametric spatiotemporal model to analyse the ratio of local COVID-19 infections with a specific disease onset date. The results of the analysis are presented in Section 4. Section 5 concludes the article.

2 | DATA DESCRIPTION

2.1 | Data on infections

Our application's outcome of interest is the ratio of COVID-19 infections in a federal administrative district (NUTS-3 level), which we define as the quotient of the number of COVID-19 infections over the corresponding population size. In Germany, there are $n = 401$ federal administrative districts (a complete list is given by the German Federal Statistical Office). At a higher hierarchical level, each federal district also belongs to a federal state (NUTS-1 level). In most figures, for example Figure 2, we colour-code the district-specific time series according to this allocation. If we refer to a specific district in the text, we generally specify the corresponding federal state in brackets.

Infection count: The Robert-Koch-Institute provides timely data on the daily number of COVID-19 infections in Germany for each federal district. We limit the present analysis to individuals between 15 and 59,years old due to the age structure in the Facebook population. Besides, the given surveillance counts are stratified by age group (15–35 and 36–59) and gender. For each entry, dates of symptom onset and reporting are given, although the onset date is partially missing. Our principal analysis is based on the disease onset date since it ensures more valid information on the infection incidence (Günther et al., 2020). Imputation of the missing values is required (we present our method in Section 3.1). By $y_{i,g,t}$ we denote the observed (and partially imputed) counts of new onsets within district i, age/gender-group g and week t. For completeness, we define with $x_{i,g}$ the corresponding indicator for the age/gender group.

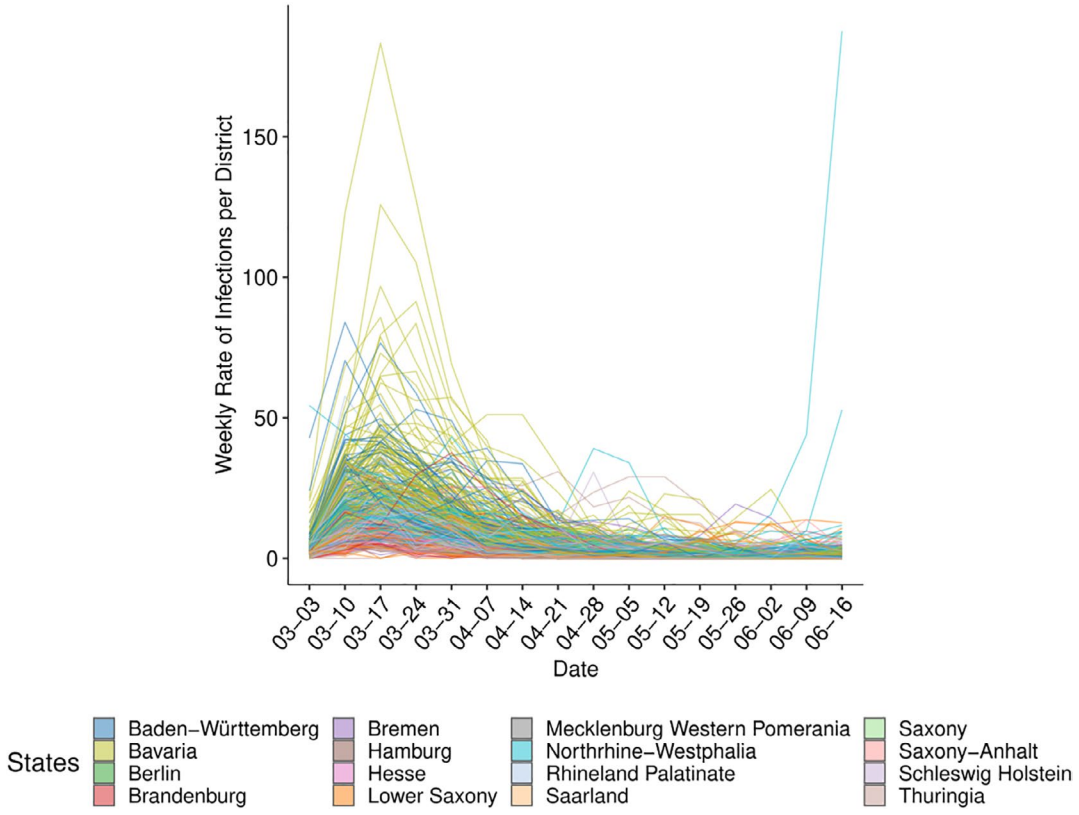

FIGURE 1 Observed Rate of Weekly Infections for each federal district. The colour of the lines indicate the federal state in which each district is located and the dates (mm:dd) are the first day of the corresponding week

Population: We obtained district-, age- and gender-specific population data from the German Federal Statistical Office. To guarantee a consistent definition of age-groups, we categorised the data according to the two primary age groups according to which the infection data are reported, namely people between 15–35 and 36–59 years old. The corresponding time-constant covariate is denoted for age/gender-group g in district i by $x_{i,g,pop}$.

The observed rates per 10.000 inhabitants $\tilde{y}_{i,g,t} = \frac{10.000 y_{i,g,t}}{x_{i,g}}$ are visualised in Figure 1 colour-coded according to the different states. For each week, we plot the rate of disease onsets that we partially impute in case of missingness, as described in detail in the next section. Once the first peak of infections could be overcome, the cases in the aftermath are increasingly attributed to local outbreaks. Two districts, namely Guetersloh and Warendorf (North Rhine-Westphalia), experience a local outbreak in a meat factory during the last weeks of the observational period (Kottasová, 2020). This local outbreak encompasses 48% of all infections with disease onset in the week starting on 16 June.

2.2 | Data on social activity during COVID-19

All data related to social activities during the COVID-19 pandemic are generated from approximately 10 million Facebook users in Germany, who enabled geolocation features in the Facebook

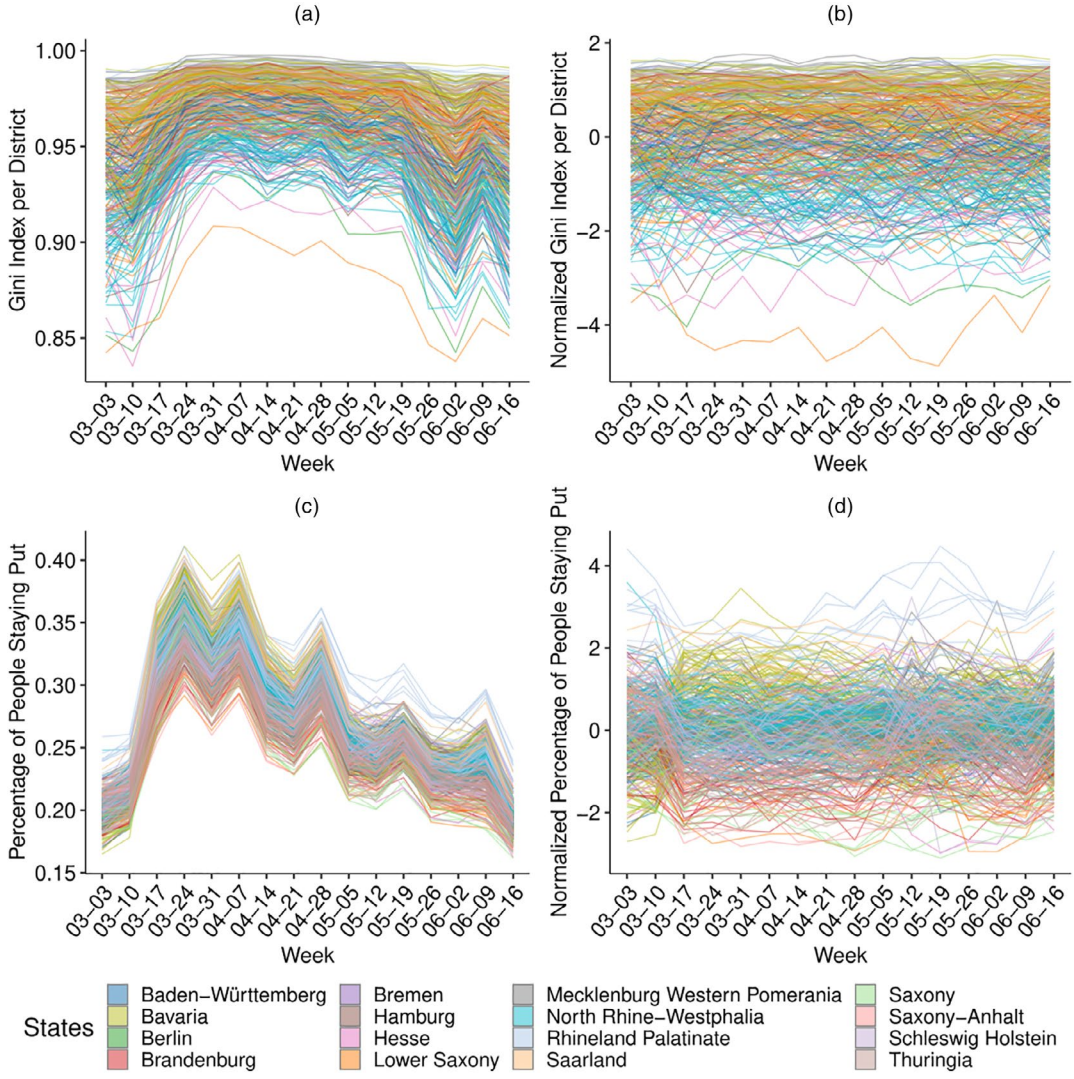

FIGURE 2 (a) Gini indices for each district over time. (b) Standardised Gini indices for each district over time. (c) Percentages of people staying put for each district over time. (d) Standardised percentages of the people staying put. The colour of the lines indicate the state in which each district is located and the dates (mm:dd) are the first day of the corresponding week

app on their mobile devices. To abide by the privacy policies, the observations are anonymised through aggregation onto tile bing polygons, censoring if we observed not enough users in the spatial region, as well as randomisation using additional noise and spatial smoothing (Maas et al., 2019). We aggregate the polygons to the same spatial units for our application on which we have the infection data. We propose the following measures describing social interaction and mobility. All measurements are taken weekly, where we use simple averaging for quantities available at a more granular temporal resolution.

Co-location: Co-location in week t is measured by the probability $p_{ij,t}$ of a random person from district i to be located in the same 0.6 km × 0.6 km square as another random person from district j (Iyer et al., 2020). These probabilities are then used to construct a district-wise quantity for the concentration of meeting patterns using the Gini index, which is given by:

$$x_{i,t,gini} = \frac{\sum_{m,l \neq i} |p_{im,t} - p_{il,t}|}{2(n-1)\sum_{j \neq i} p_{ij,t}}.$$

If we were to observe the maximal value of 1 in $x_{i,t,gini}$, all people within federal district i would only meet people (i.e. Facebook users) from only one further district. This behaviour is exemplary of extremely restricted mobility. Conversely, a lower value heuristically indicates dispersed meeting patterns. Due to this intuitive interpretation, we opt for the Gini index as a measure of concentration. The Gini indices' temporal paths for the 401 districts in Germany are depicted in Figure 2a. Overall, the meeting patterns become more concentrated on a few other districts as the crisis evolves. This behaviour contrasts rather dispersed practices before the pandemic. An upward trend is visible until the nationwide lockdown on 22nd of March, 2020.1 Thereupon, meeting patterns continue to be overall condensed, although the indices slowly decline. To enable a meaningful comparison between the respective estimates in the regression setting of Section 3, we standardise the Gini indices per week. The standardised covariate $\tilde{x}_{i,t,gini}$ is shown in Figure 2b and given by

$$\tilde{x}_{i,t,gini} = \frac{x_{i,t,gini} - \hat{\mu}_{t,gini}}{\hat{\sigma}_{t,gini}}, \tag{1}$$

where $\hat{\mu}_{t,gini} = \frac{1}{n}\sum_{j=1}^{n} x_{j,t,gini}$ and $\hat{\sigma}_{t,gini} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{j,t,gini} - \hat{\mu}_{t,gini})^2}$.

Percentage staying put: Besides the relative attribution of co-location probabilities to other districts, we investigate a measure that expresses how people (Facebook users) comply with social distancing. We quantify this concept by the covariate $x_{i,t,sp}$, which is defined as the average percentage of people in district i staying put during week t. Respective data were collected using geolocation traces of mobile devices and users are defined to be staying put, if they are only observed in one 0.6 km × 0.6 km square throughout a day (Facebook, 2020). In Figure 2c, clear break-points are visible, giving evidence of the temporary lockdown that started between 17 and 24 March. During the following weeks, the observed values gradually level off around pre-lockdown values. We also observe some peaks in the weeks starting on 7 and 28 April, which could be traced back to the different mobility behaviour during national holidays, namely *Good Friday* on 10 April and *Labour Day* on 1 May 2020.

Similarly to the treatment of the Gini index, we standardise the percentages in the regression setting. While the visual impression from Figure 2c insinuates that the dynamics of people staying put are similar between districts, the standardised paths, given in Figure 2d, reveal local differences between them. For instance, the early look-down in Bavaria resulted in a substantial relative increase of the respective districts between the 10th and 17th of March, see the yellow-green lines.

Friendship distance: Spatial distance is found to be strongly associated with the spread between regions (Kang et al., 2020). Beyond the geographical proximity, Cho et al. (2011) argued that friendship ties explain specifically long-distance mobility, which is fundamental for understanding the early spread of the pandemic (Chinazzi et al., 2020). To accommodate this possible line of infection, we include a measure for the strength of friendship ties between the districts of Germany. More precisely, we employ the social connectedness index proposed by Bailey et al. (2018), which is based

[1]In Bavaria, the lockdown started already on 16 March 2020.

on an anonymised snapshot of all active Facebook users and their friendship networks from April 2020. For the administrative district i and j, the time-invariant measure $x_{ij,soc}$ is given by:

$$x_{ij,soc} = \frac{\#\{\text{Friendship Ties between users in district } i \text{ and } j\}}{\#\{\text{Users in district } i\}\#\{\text{Users in district } j\}} \quad \forall \; i,j \in \{1,\ldots,n\}. \tag{2}$$

In a note, Kuchler et al. (2021) uncover high correlations between the social connectedness indices and the spread of COVID-19. This index is further processed to provide a spatial allocation based on social instead of Euclidean distances. To do so, we first transform social connectedness to social distance d_{soc} by taking the reciprocal of connectedness, that is, $d_{ij,soc} = \frac{1}{x_{ij,soc}}$. Consecutively, we process these distances to coordinates using *multidimensional scaling* (Cox & Cox, 2000). In our application, this procedure's result is a two-dimensional representation of each district's location in the network defined through Equation (2) that is only identifiable up to the scale and rotation. Using *Procrustes analysis*, we map the rotation of the inferred coordinates in the friendship space to be most similar to the geographical coordinates (Cox & Cox, 2008). Technical details on both procedures are given in Annex A. The outcome of the algorithm for each district i is denoted by $x_{i,soc}$ and gives the geo-coordinates in the friendship space as shown in Figure 3. Robust connectivity within federal states and neighbouring districts are visible in the friendship coordinates. We also observe that the capital, Berlin, is situated in the very centre, reflecting its unique and highly connected position. One can also detect a persisting corridor between districts located in former East- and West Germany. Next

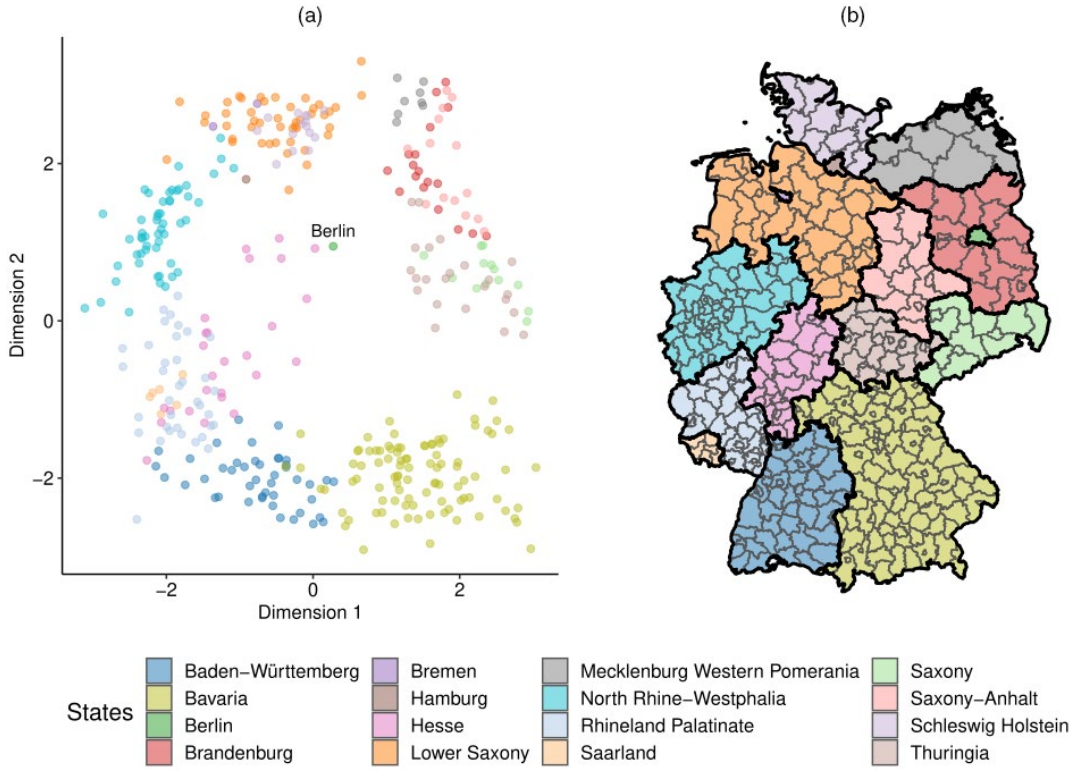

FIGURE 3 (a) Coordinates representing the friendship distances. The colour of the points indicates the state in which each district is located. (b) Map representing the colour legend. The thick black lines represent borders between federal states, while the thinner grey borders separate federal districts

to the social coordinates, we incorporated each district's geographical coordinates $x_{i,coord}$, that is, the longitude and latitude of each districts centroid, in our application.

3 | MODELLING

We start by proposing a model to impute missing dates of the disease onset. Subsequently, these partially imputed infection data are modelled with a negative binomial regression.

3.1 | Imputation model

We can see in Figure 4 that approximately 30% of the onset dates are missing. To still make use of all available information, we propose to impute missing disease onset dates under the assumption of missingness at random. This allows for unbiased findings which are not guaranteed when using a complete case analysis (Little & Rubin, 2002). In particular, we leverage the fact that the chronologically later reporting date is available for all cases. Thereby, the problem of imputing the date of disease onset for a single infection is reduced to imputing the time between onset of disease and its reporting through a positive test, which we call test delay. Following Günther et al. (2020), we use the subset of all data without any missing disease onset dates to fit a distributional regression model for this test delay. In the next step, we predict all distributional parameters under this model for all cases with a missing disease onset date and sample the missing onset date.

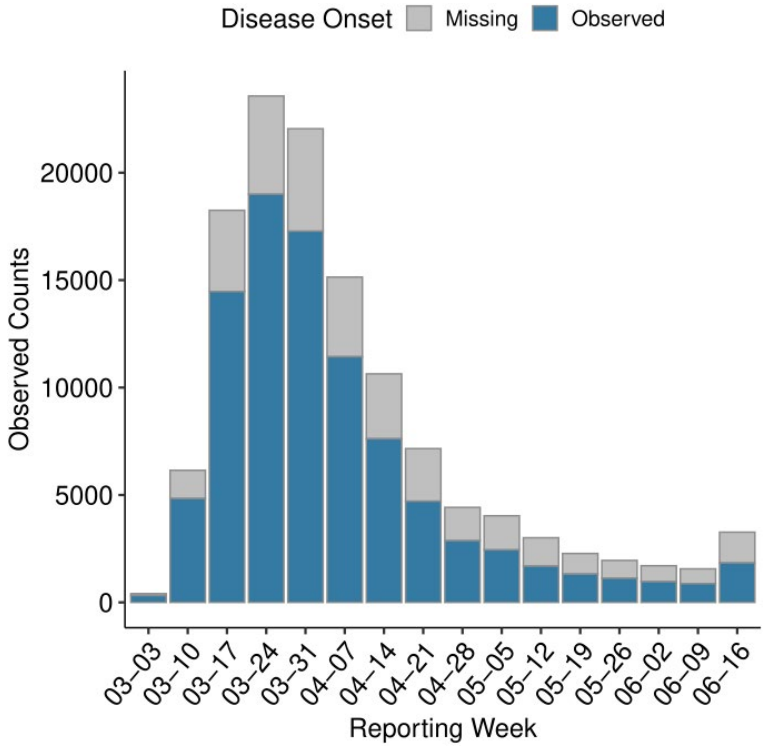

FIGURE 4 Count of missing and observed disease onset dates per reporting week

To fit the imputation model, we first disaggregate the given surveillance counts to the patient level. For each complete case l, the data include the age/gender group indicator ($x_{l,g}$), an indicator whether the reporting date was during a weekend ($x_{l,weekend}$), and the state ($x_{l,state}$) and district ($x_{l,district}$) where it was observed. Regarding the temporal information of each infection, we are given the date of disease onset ($t_{l,o}$) and its reporting ($t_{l,r}$). For complete-case data, the test delay is then given by $d_l = t_{l,r} - t_{l,o}$. As regressors in the imputation model, we include dummy covariates $x_l = (x_{l,g}, x_{l,weekend}, x_{l,state}, x_{l,district})$ and the metric covariate $t_{l,r}$ itself to account for changing testing strategies, for example, during the early spread the test capacities were limited and patients needed to wait longer for a test to be conducted. We assume that d_l is a realisation of random variable D_l, which follows a negative binomial model:

$$D_l \,|\, x_l, t_{l,r} \sim NB\left(\mu_l = \exp\left\{\theta_\mu^\top x_l + f_\mu(t_{l,r})\right\}, \sigma_l = \exp\left\{\theta_\sigma^\top x_l + f_\sigma(t_{l,r})\right\}\right), \tag{3}$$

where $\mathbb{E}(D_l \,|\, x_l,\, t_{l,r}) = \mu_l$ and $\mathrm{Var}(D_l \,|\, x_l,\, t_{l,r}) = \mu_l + \sigma_l \mu_l^2$ holds. A discrete-valued distribution appears most suitable since the patient-level data are available daily, making the test delay inherently discrete. As indicated in Equation (3), we model the location and scale parameters of the distribution by separate linear predictors. Note that the linear predictors are defined by $\eta_\mu = \theta_\mu^\top x_l + f_\mu(t_{l,r})$ and $\eta_\sigma = \theta_\sigma^\top x_l + f_\sigma(t_{l,r})$ for the corresponding distributional parameters and that the linearity only refers to linearity in the coefficients not in the covariates. Therefore, the model lies within the family of generalised additive models for location, scale and shape (Rigby & Stasinopoulos, 2005). While all components of x_l have a log-linear effect, we parameterise the trend effect of the reporting date $t_{l,r}$ by nonlinear penalised splines (see Eilers & Marx, 1996 for details). The district-specific effects are assumed to be Gaussian. After having obtained the estimates, we calculate $\widehat{\mu}_{\tilde{l}} = \exp\{\widehat{\theta}_\mu^\top x_{\tilde{l}} + \widehat{f}_\mu(t_{\tilde{l},r})\}$ and $\widehat{\sigma}_{\tilde{l}} = \exp\{\widehat{\theta}_\sigma^\top x_{\tilde{l}} + \widehat{f}_\sigma(t_{\tilde{l},r})\}$ for all observations $\tilde{l}$ with missing disease onset. We can now simulate $d_{\tilde{l}}$ from Equation (3) to acquire a full data set by setting $t_{\tilde{l},o} = t_{\tilde{l},r} - d_{\tilde{l}}$. Through aggregation from the daily patient-level data to the infection counts per district i and age/gender group g with disease onset in week t, denoted by $y_{i,g,t}$, we build a single partially imputed data set. This procedure is repeated K times to represent the uncertainty associated with the missing information of all disease onsets.

3.2 | Infection model

To model the rate of infections with partially imputed data, we apply a negative binomial '*observation-driven*' model for count data including the population as an offset term (Cox, 1981). By doing so, we assume

$$Y_{i,g,t} \,|\, x_{i,g,t-1}, a_i, b_i \sim NB(\mu_{i,g,t}, \sigma), \;\; \forall \; i \in \{1, \ldots, 401\}, \;\; g \in \mathcal{G}, \;\; \text{and} \;\; t = 2, \ldots, T, \tag{4}$$

where $x_{i,g,u} = (u, x_{i,g}, x_{i,g,pop}, \tilde{x}_{i,u,gini}, \tilde{x}_{i,u,sp}, x_{i,coord}, x_{i,soc}, \tilde{y}_{i,g,t-1})$ are the covariates at arbitrary week u specified in Section 2 and $\mathcal{G}$ denotes the set of age/gender groups used from the data. Furthermore, let T be the final week of data we use in the analysis. We assume in Equation (4) that the random variable $Y_{i,g,t}$ follows a negative binomial distribution conditional on $x_{i,g,t-1}$, a_i and b_i to compensate overdispersion in the observed counts.

Aligned with models for the spread of infectious diseases (Held et al., 2005), we decompose $\mu_{i,g,t}$ into an endemic and epidemic component:

$$\mu_{i,g,t} = \exp\left\{\nu_{i,g,t}^{END} + \nu_{i,g,t}^{EPI}\right\}, \tag{5}$$

where each part is parameterised as follows:

$$\nu_{i,g,t}^{EPI} = \theta_{AR(1)} \log(\tilde{y}_{i,g,t-1} + c) \tag{6}$$

$$\begin{aligned}\nu_{i,g,t}^{END} = {} & \theta_t + \theta_{gen} \mathbb{1}(x_{i,gen} = \text{“Male”}) + \theta_{age} \mathbb{1}(x_{i,age} = \text{“36}-\text{59”}) \\ & + \theta_{age:gen} \mathbb{1}(x_{i,gen} = \text{“Male”}) \cdot \mathbb{1}(x_{i,age} = \text{“36}-\text{59”}) + \theta_{t,gini} x_{i,t-1,gini} \\ & + \theta_{t,sp} x_{i,t-1,sp} + f_{coord}(x_{i,coord}) + f_{soc}(x_{i,soc}) + a_i + b_i \mathbb{1}(t = T) + \log(x_{i,g,pop}).\end{aligned} \tag{7}$$

We include a first-order autoregressive term of this rate, since path dependencies and self-exciting behaviour are common with infectious diseases and should therefore be accounted for Held et al. (2005). In addition, we transform the respective term by $h(x) = \log(x + c)$ to bypass problems with absorbing states of the implied counting process when $\tilde{y}_{i,g,t-1} = 0$. The value $c \in (0, 1]$ is estimated from the data. More general types of these autoregressive models are proposed by Zeger and Qaqish (1988).

As is evident from Equation (5), we constitute that both the epidemic and endemic components have a multiplicative effect on the observed infection rates. As an alternative, Held et al. (2005) replace the log link by an identity link, although Fokianos et al. (2020) argue for the logarithmic link implied in Equation (5) if additional covariates are available. They further derive theoretical properties, such as ergodicity, in the case of Poisson-distributed target variables under the condition $\theta_{AR(1)} < 1$.

Time-varying effects: For the endemic part (7), the temporal trend is reflected by piecewise constant fixed effects separately for each week, θ_t. By means of group-specific covariates we control for gender- and age-related effects and their interaction, θ_{gen}, θ_{age} and $\theta_{age:gen}$ (Walter & Mcgregor, 2020). The principal covariates, Gini Index and Percentage Staying Put, are modelled by piecewise constants in each week for maximal flexibility. To account for the stylised fact, that the incubation period, that is, the time between being infected and symptom onset, for COVID-19 is around 5 days (Li et al., 2020a), we lag the information on Gini Index and Percentage Staying Put by one week as indicated in Equation (7).

Isotropic smooth effects: The bivariate functions $f_{coord}(x_{i,coord})$ and $f_{soc}(x_{i,soc})$ display the effects of geographical coordinates and social coordinates on the incidence rate. To properly incorporate $x_{i,coord}$ and $x_{i,soc}$ in our regression framework, we propose the usage of isotropic splines. These kind of flexible functions were proposed by Duchon (1977) to model multiple covariates by a multivariate term as an alternative to anisotropic tensor products. Isotropic smoothers have the property of giving the identical predictions of the response under arbitrary rotation and reflection of the respective covariates (Wood, 2017). This characteristic is commonly reasonable when working with geographical coordinates $x_{i,coord}$ and in accordance with the uniqueness of the multidimensional scaling results, thus also for $x_{i,soc}$. With respect to the form of the smooth terms, we follow Wood (2003) and use a low-rank approximation of the thin-plate splines introduced in Duchon (1977). To obtain a smooth fit, we impose a penalty that is controlled by τ_{soc} and τ_{coord} for the respective isotropic splines.

Random effects: Because super spreader events such as carnival sessions (Streeck et al., 2020) or local outbreaks in major slaughterhouses (Dyal et al., 2020) lead to unobserved heterogeneities, our model comprises two district-specific Gaussian random effects. The random effect a_i governs long-term heterogeneities, while short-term dependencies, that is, sudden locally confined outbreaks as visible in the last week of Figure 1, are captured by b_i. We assume $a = (a_1, \ldots, a_n)^\top \sim N(0, I_n \tau_a^2)$ and $b = (b_1, \ldots, b_n)^\top \sim N(0, I_n \tau_b^2)$. Relying on the duality

between semiparametric regression and random effects (Ruppert et al., 2003), we can equivalently write the random effects as semiparametric terms. Hence we may replace a_i and b_i by $f_a(i) = (a_1, \ldots, a_n)^\top X_a$ and $f_b(i) = (b_1, \ldots, b_n)^\top X_b$, respectively, and introduce a ridge penalty for each coefficient vector. In this context, the design matrices X_a and X_b each consist of n dummy variables indicating to which district a specific observation refers. As a result of this reformulation, we can estimate the additional parameters τ_a and τ_b as tuning parameters in semiparametric regression (see Annex B for further information).

Modelling rates via count regression: Effectively, we model the rate of infections by including the term $\log(x_{i,g,pop})$ as an offset in Equation (7) since the infections rates $\tilde{Y}_{i,g,t}$ relate to the counts through $Y_{i,g,t} = \tilde{Y}_{i,g,t} x_{i,g,pop}$ via (note the slight abuse of notation as we here do not regard the infection rate among 10.000 inhabitants but the percentage of people infected with a disease onset in a specific week). As a byproduct, we implicitly assume that the entire population is susceptible, which is reasonable when considering the low prevalence of COVID-19 in Germany during the first wave. However, the model is still applicable in the later stages of the pandemic by replacing this offset with the number of susceptible inhabitants in each region.

3.3 | Estimation

At first, we propose an estimation procedure for the imputation model from Section 3.1. Given a partially imputed data set, we specify how to get estimates for the infection model from Section 3.2. Finally, the multiple imputation scheme combining both approaches is presented. Generally, we carry out all computations conditional on the observations in $t = 1$, that is, the week between the 3rd and 9th of March.

Imputation model: We get estimates for the imputation model through maximising the likelihood function resulting from Equation (3). As mentioned in Section 3.2, we can rewrite all random effects as smooth terms and penalise this likelihood to obtain smooth functions. By repeatedly updating the estimators through a backfitting algorithm, we optimise this objective (see Rigby & Stasinopoulos, 2005 for details). This procedure is readily implemented in the software package `gamlss` (Stasinopoulos et al., 2020).

Infection model: The infection model is characterised by the parameters c and θ, relating to the log-transformation of the autoregressive component and all other parameters. Given a partially imputed data set, we first consider θ to be a nuisance parameter and find c via a profile likelihood approach. Here the profile likelihood is given by

$$\mathcal{L}_{Profile}(c) = \max_{\theta} \mathcal{L}(c, \theta) = \mathcal{L}(c, \widehat{\theta}(c)),$$

where $\mathcal{L}(c, \theta)$ is the joint likelihood resulting from Equation (4) and $\widehat{\theta}(c)$ is the maximum likelihood estimator of θ for a fixed value of c. For any c, we can find $\widehat{\theta}(c)$ by carrying out the estimation as explained in Annex B, hence it is straightforward to evaluate $\mathcal{L}_{Profile}(c)$. Building on this result, we use standard optimisation software, that is, the `optimise` routine within the software environment `R` (R Core Team, 2020), to obtain $\widehat{c} = \arg\max_c \mathcal{L}_{Profile}(c)$. In the consecutive step, we fix c at $\widehat{c}$ to get $\widehat{\theta}$ again by following Annex B.

Multiple imputation: Since information on the onset of symptoms is missing for approximately 30% of the cases, we proposed an imputation model in Section 3.1 to generate K partially imputed data sets. To correct the uncertainty quantification of the infection model for this multiple imputation procedure, we use the Rubin's rule. At first, we sample K imputed

data sets according to Section 3.1. Let $\widehat{\vartheta}_{(k)} = (\widehat{\theta}_{(k)}, \widehat{c})$ be the resulting estimator from the two-stage maximum profile likelihood procedure explained in the previous paragraph given the partially imputed data set from the kth imputation step. By $\widehat{V}_{(k)}$ we denote the corresponding variance estimate that results from Bayesian large sample properties (Wood, 2013). We then average the coefficients over all K iterations to obtain $\widehat{\vartheta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\vartheta}_{(k)}$ and estimate its variance through:

$$\widehat{\mathrm{Var}}(\widehat{\vartheta}_{MI}) = \overline{V} + (1 + K^{-1})\overline{B},$$

where its components are given by

$$\overline{V} = \frac{1}{K}\sum_{k=1}^{K}\widehat{V}_{(k)}$$
$$\overline{B} = \frac{1}{K-1}\sum_{k=1}^{K}(\widehat{\vartheta}_{(k)} - \widehat{\vartheta}_{MI})(\widehat{\vartheta}_{(k)} - \widehat{\vartheta}_{MI})^{\top}.$$

In our application, setting $K = 20$ proved to be sufficient since the estimates of different imputed data sets varied only marginally.

4 | RESULTS

We only report the findings of the infection model detailed in Section 3.2. A detailed analysis of the imputation model as well as a robustness check for the infection model can be found in the Supplementary Material.

4.1 | Temporal effect

To start, the estimate of θ_t is shown in Figure 5. The progression of the weekly estimates confirms generally decreasing infection rates over time. Due to the standardisation employed for the principal covariates in the analysis, the temporal trend can be interpreted as the log-transformed expected infection rate of female individuals aged between 15 and 35 in a district where the standardised Gini Index and Percentage Staying Put are zero. Since observing a zero in the standardised covariates translates to the mean observed values where we observed most information, the standard errors are also extremely narrow.

4.2 | Sociodemographic and epidemic effects

The linear time-constant estimates are given in Table 1 and exhibit in general a negative effect on male patients compared to female patients, 3% in the younger and 9.6% in the older age cohort.[2]

[2]One can derive these percentages by computing the expected multiplicative change that results from alternating the prediction from one to another demographic group. For instance, $\exp\{0.03\} \approx 0.97$, which is equivalent to a 3% decrease, is the multiplicative change ceteris paribus between females and males both aged between 15 and 35.

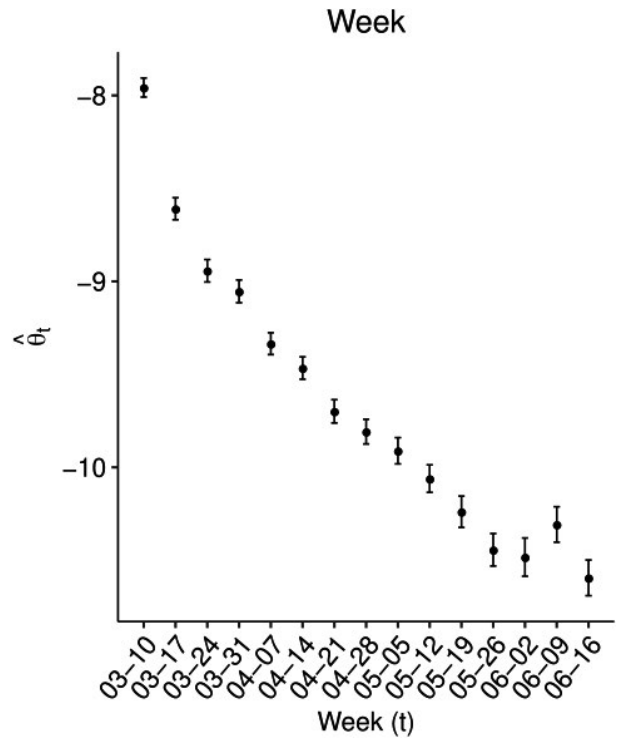

FIGURE 5 Estimate of temporal effect θ_t. The 95% confidence interval accompanies the estimates, and the shown dates (mm:dd) on the *x*-axis are the first days of the corresponding weeks

TABLE 1 Estimates of linear time-constant effects

Covariable	Estimate (standard error)	exp{Estimate} (standard error)
Male	−0.03	0.97
	(0.015)	(0.014)
A35–A59	−0.031	0.969
	(0.014)	(0.013)
Male: A35–A59	−0.071	0.931
	(0.02)	(0.017)
$\log(\tilde{y}_{i,g,t-1} + c)$	0.623	1.865
	(0.009)	(0.031)

Notes The reference group are female individuals aged between 15 and 35. By use of the delta rule, we approximated the standard errors of the transformed coefficients in the third row. The value *c* is estimated at 0.499 with a standard error of 0.027.

According to its partial effect, we also predict that the older age group has a lower infection rate than the younger group encompassing individuals aged between 15 and 35, for men 9.7% and women 3.1%. The autocorrelation coefficient $\theta_{AR(1)}$ expresses that one more infection among 10.000 inhabitants in a district during the past week almost doubles the predicted infections for the present week. This dominant finding confirms strong path dependencies in the data. In this context, we need to remark that the coefficients are partial effects that condition on all other

covariates used in the model. Therefore, a positive coefficient of a dummy variable does not necessarily translate to the same finding in the raw numbers.

4.3 | Mobility effects

The time-varying estimates regarding the relative mobility pattern are displayed in Figure 6. Overall, the estimated effects of the measures proposed in Section 2.2 on the rate of local COVID-19 infections are negative. In regards to relative importance, both variables rank similarly during the lockdown period that persists until early May. Subsequently, the Gini Index in a region gains weight, while the effect of People Staying Put becomes more volatile. The temporal changes of the respective estimates illustrate nonlinearities, which would not have been sufficiently captured by linear effects.

Gini index of co-location: Given all other covariates, Figure 6a suggests that inhabitants with meeting patterns that are centred around a few other districts entail reduced infection rates for a specific district. This tendency is only suspended in the week starting on March 17th during the early lockdown in Bavaria. The corresponding estimate is positive and significant. Right after the national lockdown on 22 March 2020, is ordered, the effect is not significantly different from zero for one week (03–24). The estimated effects remain low but negative until the German government introduces compulsory masks in public areas on 22 April (Mitze et al., 2020). Thereupon, the effect has a clear downwards tendency. Once policymakers slowly lift the lockdown measures, the estimate declines further until its maximum in the penultimate week of our observational period. This development may be viewed as evidence that a more focused attribution of co-location probabilities in a district becomes more crucial over time.

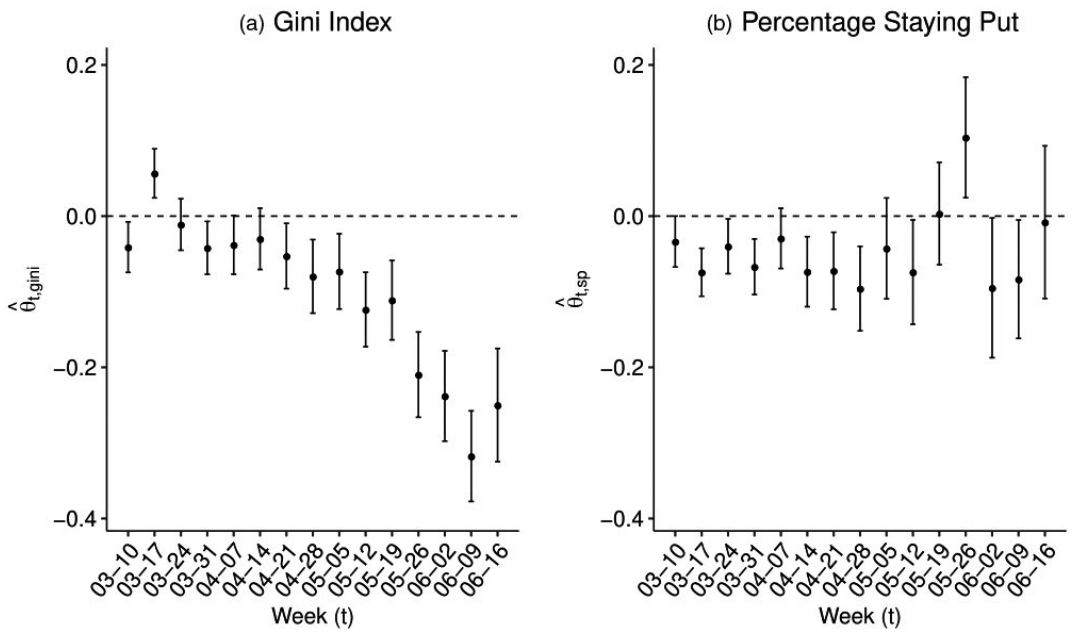

FIGURE 6 (a) Time-varying effects of the Gini index $\widehat{\theta}_{t,gini}$. (b) Time-varying effects of the Percentage of People Staying Put $\widehat{\theta}_{t,sp}$. The 95% confidence interval accompanies the estimates, and the shown dates (mm:dd) on the *x*-axis are the first days of the corresponding weeks

Percentage staying put: Suppose the percentage of inhabitants in a district staying put is large relative to the national tendency. In that case, we expect the incidence of infections throughout the lockdown period to be lower. We deduce this result from the largely negative estimates in Figure 6b for the weeks between 10 March and 12 May. Once the orders are relaxed, on the other hand, the standard errors of the respective covariate become relatively large, and the effect vanishes in the final week of the study. A possible explanation for this phenomenon is that when daily infections decline, most diseases are related to local outbreaks (as already mentioned in Section 2). These breakouts, in turn, cannot be associated with the percentage of people staying put. One exception to this finding is the estimate in the week starting on 26 May, where we encounter a significant positive effect.

4.4 | Spatial and social connectedness effects

In our model specification, we incorporate the friendship coordinates and geographical coordinates as two spatial effects. In combination with the two unstructured latent variables, we can disentangle separate influences on the local infection rates of spatial and friendship proximity as well as short- and long-term district-specific deviations from it.

Spatial effects: Let us start with the smooth spatial effect in Figure 7. Overall, the geographical effects within federal states, indicated by the black borders in Figure 7, are mostly heterogeneous. To give some examples, an almost uniformly augmented risk of infections is estimated in Baden-Württemberg and Thuringia. At the same time, we remark a negative spatial effect in Germany's northern districts, that is, Schleswig Holstein and Mecklenburg Western Pomerania. On the other hand, the fit for districts in North Rhine-Westphalia varies between positive, negative and no effect.

We visualise the result of the friendship coordinates in two manners. One may plot the smooth bivariate function in the friendship space, Figure 8a, or map the smooth fit on the geographical space, Figure 8b. The re-mapping allows for sharp edges in the geographical coordinates. Broadly, the fit differentiates between districts allocated in former East Germany (corresponding in Figure 8a to MDS coordinates located in the first quadrant) and former West Germany. We observe that the predicted infections are ceteris paribus lower if a district is situated in former East Germany. Districts allocated in the second and fourth quadrant of Figure 8a (mainly including districts from the states Bavaria, North Rhine-Westphalia and parts of Lower Saxony) are negatively affected by social proximity. Figure 8b demonstrates how the partial effects sometimes change abruptly between large cities and neighbouring districts. For instance, Berlin's central position is unrelated to the infection rates compared to the negative effect evaluated in Brandenburg. We observe a similar phenomenon for Hamburg when contrasting its partial effect with surrounding districts in Schleswig Holstein and Lower Saxony.

Unobserved heterogeneity effect: In Figure 9, the posterior modes of both random effects evince strong heterogeneities between districts and underpin local differences in the spread of COVID-19. Noticeable estimates of the long-term random effects, Figure 9a, reflect early outbreaks in the districts Greiz (Thuringia) and Coesfeld (North Rhine-Westphalia). Some estimates may also be related to heterogeneous testing practices between the districts.

We can trace back most high estimates of the short-term random effect to locally confined outbreaks, for instance, Guethersloh and Warendorf (North Rhine-Westphalia). As already stated in Section 2 the proportion of infections attributed to these local events rises once the general level of new cases declines. This result is supported by the different scales of the two types of random

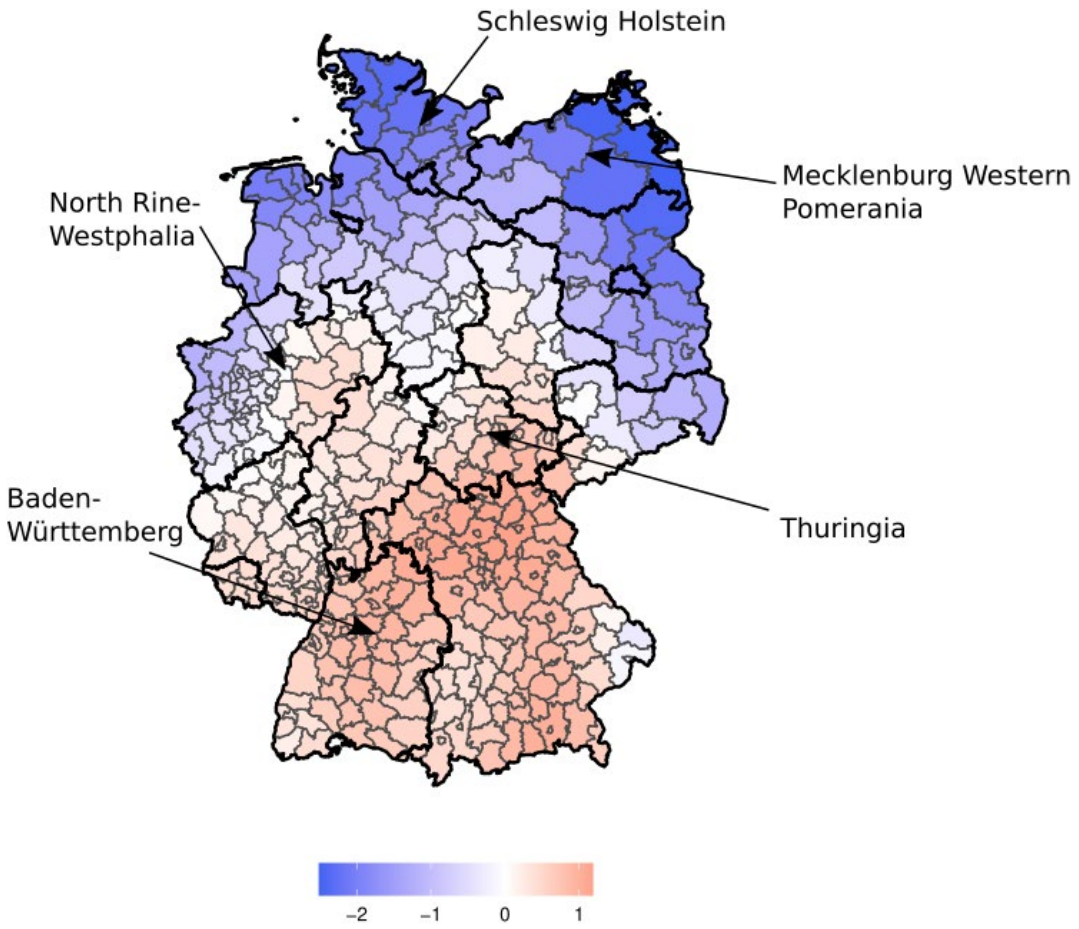

FIGURE 7 Estimated smooth spatial effect f_{coord}. The thick black lines represent borders between federal states, while the thinner grey borders separate federal districts. Through arrows, we highlight selected states mentioned in the text

effects and apparent in the estimates $\widehat{\tau}_a = 0.2 < \widehat{\tau}_b = 0.585$. Therefore, the posterior modes of the short-term effects exhibit higher variances and are larger in absolute terms than the long-term effects.

4.5 | Model assessment

We compare various alternative model specifications to check the robustness of our conclusions. In particular, we estimate separate models, adding dummy covariates for each state and leaving out one of the spatial terms, the Gini index, the Percentage of People Staying Put, all Facebook-related covariates and random effects. For this endeavour, we utilise the corrected Akaike information criterion (cAIC) introduced by Wood et al. (2016) since the effective degrees of freedom need to adjusted for the additionally estimated variance components if random effects are included (we average the respective values over the results of all imputed data sets). The results in Table 2 support the appropriateness of our final model since the corresponding cAIV value is the lowest. Besides, the change in the cAIC value to the model (4), denoted by ΔcAIC, permits

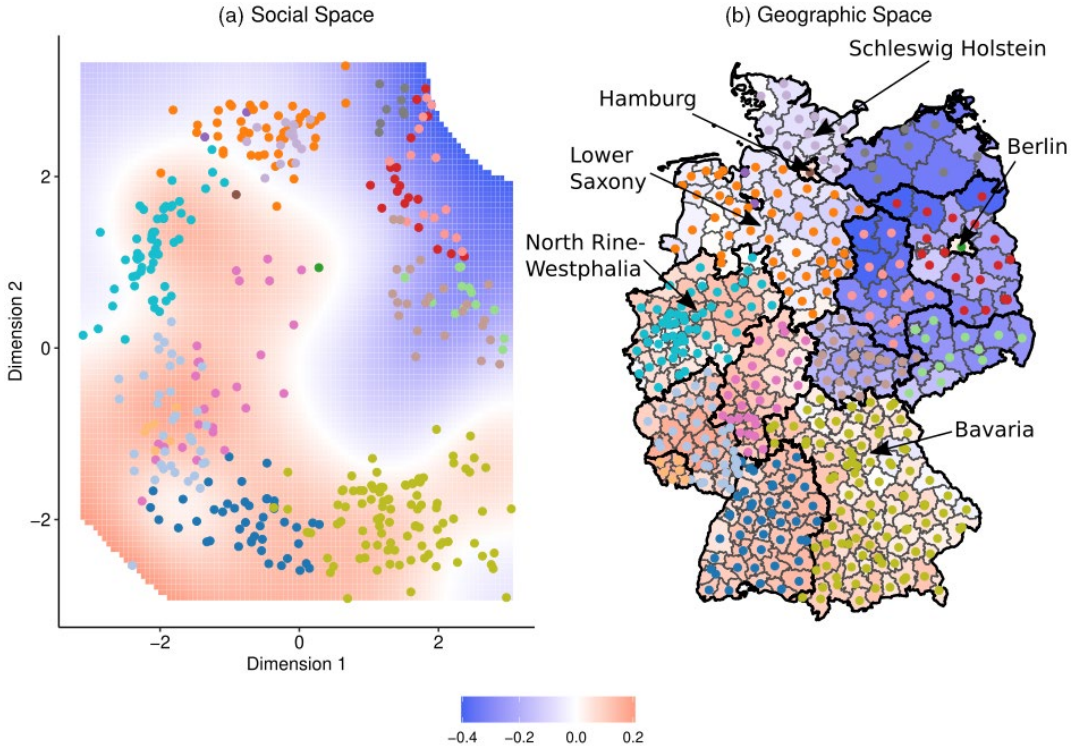

FIGURE 8 (a) Coordinates of the districts in the friendship space with the smooth partial effect of f_{soc} in the background. We only show the predictions in the range of observed values. (b) Coordinates of the districts in the geometric space with the smooth partial effect of f_{soc} again shown in the background for each district. The thick black lines represent borders between federal states, while the thinner grey borders separate federal districts. Through arrows, we highlight selected states mentioned in the text

an evaluation of the variable importance of each eliminated covariate. We can conclude from Table 2 that the exclusion of the Gini Index induces the highest loss in cAIC value. Concerning the different types of distances, the friendship distance is more important than the geographical distance.

For further validation, we plot one draw of the randomised quantile residuals in Figure 10a. Dunn and Smyth (1996) proposed this type of residual based on the result that evaluating the cumulative distribution function at all observed values of $y_{i,g,t}$ under the estimated parameters should yield uniformly distributed random variables. Transforming these uniform values by the quantile function of the standard normal gives the quantile residuals. To obtain continuous residuals, the values are randomised since the negative binomial distribution in Equation (4) has discrete support. On average, the empirical quantiles are close to the theoretical expectations and do not indicate problems regarding the statistical fit. At the right tail of the distribution, 38 (out of 24.060) observations exhibit higher deviations from the normal quantiles, which we coloured in red. The underlying counts are mainly credited to local outbreaks that could not be completely captured by the random effects, namely Coesfeld (Thuringia), Cuxhaven (Lower Saxony), Aichach-Friedberg (Bavaria), Guetersloh and Warendorf (North Rhine-Westphalia). Additionally, we assess the predictions of the final model through plotting the predicted infections against the observed infections, Figure 10b, and a rootogram proposed by Kleiber and

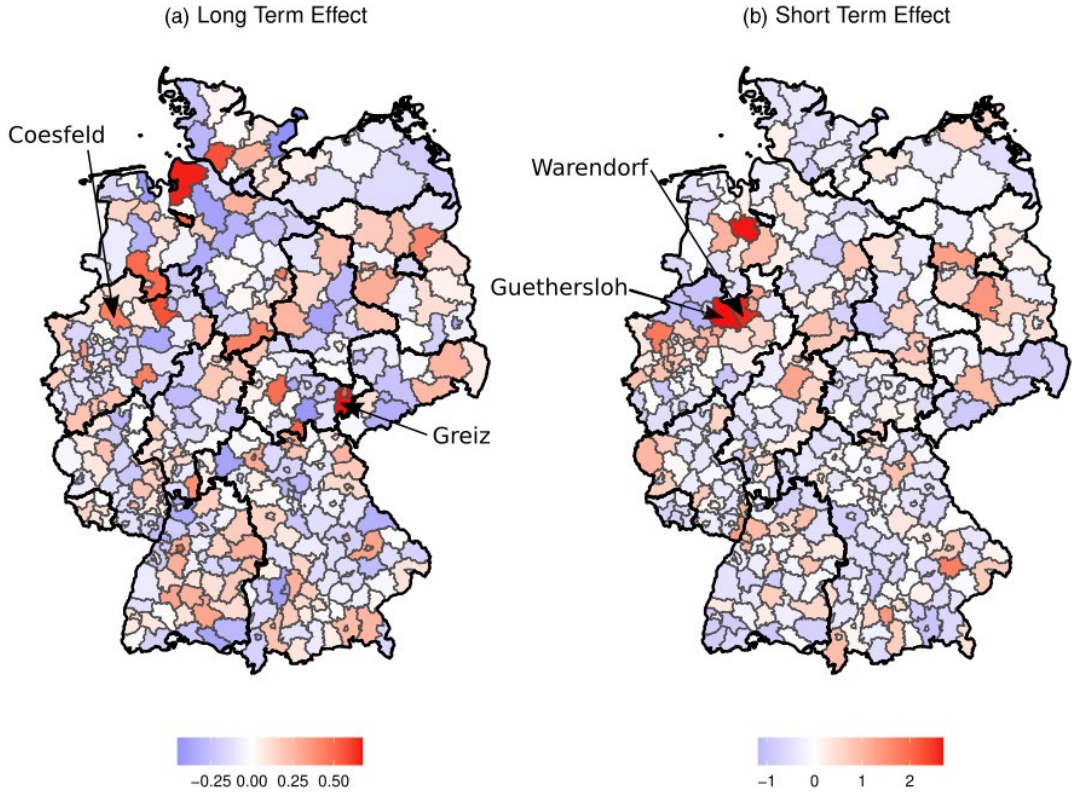

FIGURE 9 (a) Maximum posterior modes of the long-term random effects a_i. (b) Maximum posterior modes of the short-term random effects b_i. The thick black lines represent borders between federal states, while the thinner grey borders separate federal districts. Through arrows, we highlight selected districts mentioned in the text

TABLE 2 Alternative model specifications with resulting corrected Akaike information criterion (cAIC) value and change in corrected AIC value when compared to our model from Section 3

Model description	cAIC (Model)	ΔcAIC (Model)
Our model	86694	–
With state effect	86694.79	0.790
Without geographical distance	86699.42	5.422
Without friendship distance	86701.26	7.262
Without Age:Gender interaction	86707.34	13.336
Without percentage staying put	86732.46	38.461
Without Gini index	86974.32	280.319
Without Facebook covariates	87033.87	339.867
Without long-term effect	87452.62	758.620
Without short-term effect	87900.03	1206.034
Without long- and short-term effect	88624.38	1930.382

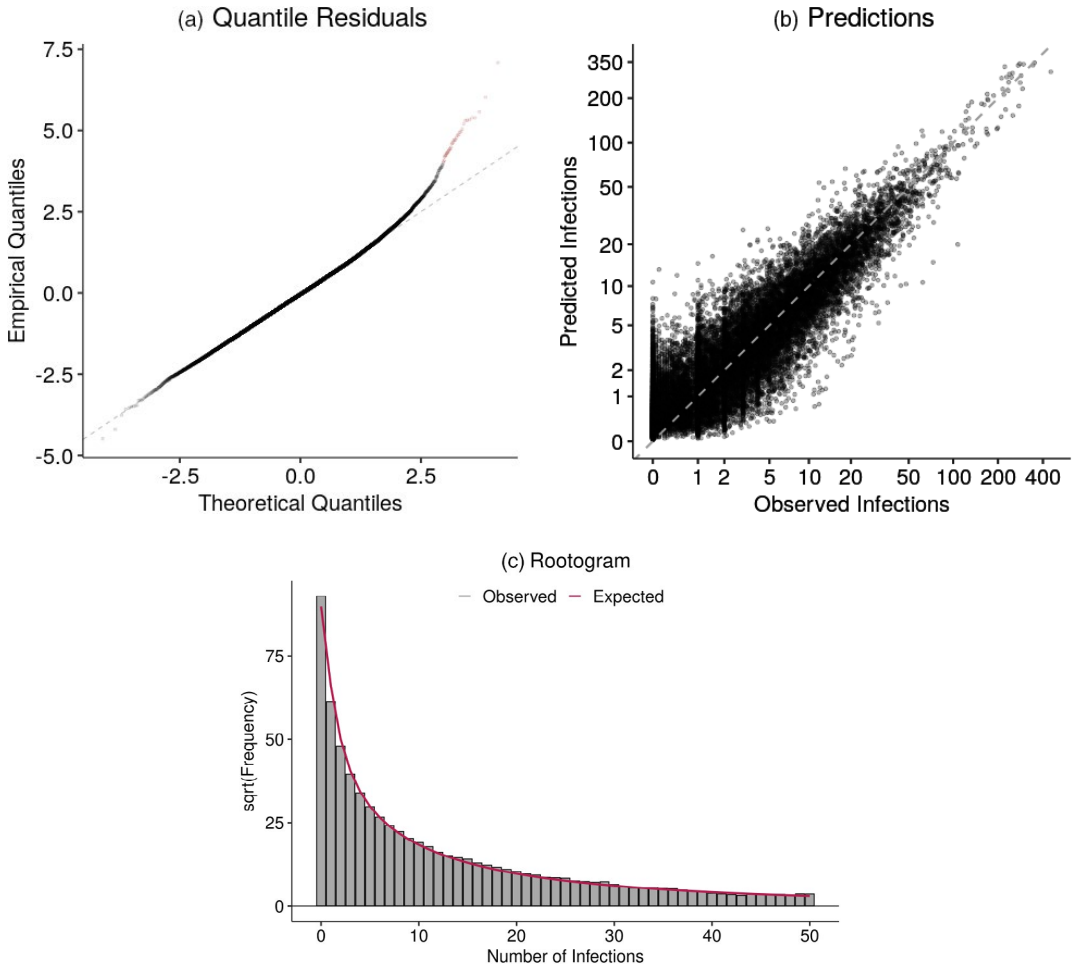

FIGURE 10 (a) QQ Plot of randomised quantile residuals, observations with a distance larger than 1 to the theoretically expected values are drawn in red. (b) Scatter plot of the observed and predicted infection count, for the x and y-axis, we used a $\log(\cdot + 1)$ scale. The dotted grey line is the best-case scenario of the prediction and has intercept 0 and slope 1. (c) Rootogram comparing the observed and expected counts. The grey barplot specifies the observed counts, while the red line gives the expected values under Equation (4)

Zeileis (2016), Figure 10c. Both visualisations confirm a strong fit of the presented model and proof that the model can sufficiently capture the observed counts of infected individuals. Due to the multiple imputation scheme specified in Sections 3.1 and 3.3, we carry the model assessment out for each imputation separately and report the averaged results.

5 | CONCLUSION

In this writing, our contributions are twofold. First, we used state-of-the-art regression models to quantify the importance of human mobility for understanding the spread of COVID-19 on a local level accounting for their temporal dynamic, latent effects and other covariates. Concerning the relative importance, the Gini index of meeting probability attribution proved to be a primary

driver of the infection rates. Second, we used methods from multivariate statistics to derive friendship coordinates for the federal districts in Germany. Consecutively, we coupled the result with a regression model via isotropic splines and, thereby, revealed a perpetual clustering of communities in former East- and West Germany that remains existent for COVID-19 infections because the social geographical system proves to be an essential regressor in our application. Moreover, our findings enable an evaluation of the district-wise policies undertaken between March and June 2020. The results corroborate the usefulness of interventions limiting trans-district movements and concentrating meeting patterns. Especially during the last weeks of this study, local lockdowns could mitigate further national outbreaks.

Still, we need to address some limitations of our work, which require additional investigation. The data sources for the infection data include all individuals in Germany that tested positive on COVID-19. During the peak phase in March, these tests were mainly carried out with patients who showed symptoms or had contact with an infected individual. Due to an unknown dark figure of infected persons missing in the public records (Lavezzo et al., 2020), the observed data are a proxy for the current epidemiological situation. To control for this possible bias, further research on the prevalence of COVID-19 in Germany and the representability of the official statistics of the real infection occurrence akin to the REACT Study in England (Riley et al., 2021) would be necessary.

Even with these caveats, the combination of infection, mobility and connectivity data can serve for a fruitful application of other methods as well. Contrasting our approach, one may tackle the regression task in Section 3 by incorporating the spatial dependencies directly in the correlation structure, as is done in the literature on spatial econometric models (LeSage & Pace, 2009). We could also employ novel clustering algorithms that naturally exploit different proximity dimensions, such as the geographical and social space, to identify similar districts while taking into account spatial dependencies (D'Urso & Vitale, 2020; D'Urso et al., 2019). Furthermore, the research questions posed in this article would greatly benefit from an examination through the lens of analytical sociology (Hedström & Bearman, 2011). Nevertheless, this type of analysis usually necessitates individual-level data, which are not readily available. Therefore, we can only verify some of the theoretical results of Block et al. (2020) on the macro scale, which does not necessarily translate to the micro scale (Stadtfeld, 2018). Therefore, additional empirical work on the implications of individual behaviour on the spread of COVID-19 is still needed. Nevertheless, our work can give valuable pointers in that regard contingent on the assumption that the corresponding district average adequately represents the mobility patterns of an individual.

6 | DATA AND CODE AVAILABILITY

Facebook collected the anonymised mobility and connectivity data. We cannot share the raw data due to a data agreement. Still, we are allowed to provide all data aggregated onto the level of federal districts. To guarantee the replicability of our results, we make the complete code to obtain the results from this article available online. We also supply a visualisation of the entire pipeline of our analysis in the Supplementary Material for transparency.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their careful reading and constructive comments. Further, we thank Alex Pompe from Facebook for his useful comments and explanations of the data from Facebook. The project was supported by the European Cooperation in Science and

Technology [COST Action CA15109 (COSTNET)]. This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibility for its content.

ORCID


Cornelius Fritz http://orcid.org/0000-0002-7781-223X

REFERENCES

Asadi, S., Bouvier, N., Wexler, A.S. & Ristenpart, W.D. (2020) The coronavirus pandemic and aerosols: does COVID-19 transmit via expiratory particles? *Aerosol Science and Technology*, 54(6), 635–638.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J. & Wong, A. (2018) Social connectedness: measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3), 259–280.

Block, P., Hoffman, M., Raabe, I.J., Beam Dowd, J., Rahal, C., Kashyap, R. et al. (2020) Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nature Human Behavior*, 4, 588–596.

Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F. et al. (2020). Economic and social consequences of human mobility restrictions under COVID-19. *Proceedings of the National Academy of Sciences*, 117(27), 15530–15535.

Borg, I., Groenen, P.J.F. & Mair, P. (2013) *Applied multidimensional scaling*. New York: Springer.

Cailliez, F. (1983) The analytical solution of the additive constant problem. *Psychometrika*, 48(2), 305–308.

Chan, J.F.W., Yuan, S., Kok, K.H., To, K.K.W., Chu, H., Yang, J. et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 395(10223), 514–523.

Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S. et al. (2020) The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, 368(6489), 395–400.

Cho, E., Myers, S.A., & Leskovec, J. (2011) Friendship and mobility: user movement in location-based social networks. In: *KDD '11: proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*.

Cox, D.R. (1981) Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, 8(2), 93–115.

Cox, T. & Cox, M. (2000) *Multidimensional scaling*. Boca Raton: CRC Press.

Cox, M. & Cox, T. (2008) Multidimensional scaling. In: Chen, C., Härdle, W.K.. & Unwin, A. (Eds.) *Handbook of data visualization*. Berlin: Springer, pp. 315–347.

D'Urso, P. & Vitale, V. (2020) A robust hierarchical clustering for georeferenced data. *Spatial Statistics*, 35, 1–33.

D'Urso, P., De Giovanni, L., Disegna, M. & Massari, R. (2019) Fuzzy clustering with spatial–temporal information. *Spatial Statistics*, 30, 71–102.

Duchon, J. (1977) Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: Schempp, W. & Zeller, K. (Eds.) *Constructive theory of functions of several variables*. Berlin: Springer, pp. 85–100.

Dunn, P.K. & Smyth, G.K. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3), 236–244.

Dyal, J.W., Grant, M.P., Broadwater, K., Bjork, A., Waltenburg, M.A., Gibbins, J.D. et al. (2020) COVID-19 among workers in meat and poultry processing facilities—19 states, April 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69(18), 557–561.

Eilers, P.H. & Marx, B.D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–102.

Facebook (2020) GeoInsights Help. Available from https://www.facebook.com/help/geoinsights.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584, 257–261.

Fokianos, K., Støve, B., Tjøstheim, D. & Doukhan, P. (2020) Multivariate count autoregression. *Bernoulli*, 26(1), 471–499.

Galeazzi, A., Cinelli, M., Bonaccorsi, G., Pierri, F., Schmidt, A.L., Scala, A.. et al. (2020). Human mobility in response to COVID-19 in France, Italy and UK. *arXiv preprint*.

Guan, W.-J., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J.-X. et al. (2020). Clinical characteristics of Coronavirus disease 2019 in China. *New England Journal of Medicine*, 382(18), 1708–1720.

Günther, F., Bender, A., Katz, K., Küchenhoff, H. & Höhle, M. (2020) Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal* (OnlineFirst).

Hedström, P. & Bearman, P. (2011) What is analytical sociology all about? An introductory essay. In: Bearman, P. & Hedström, P. (Eds.) *The Oxford handbook of analytical sociology*. Oxford: Oxford University Press.

Held, L. & Paul, M. (2012) Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal*, 54(6), 824–843.

Held, L., Höhle, M. & Hofmann, M. (2005) A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3), 187–199.

Held, L., Meyer, S. & Bracher, J. (2017) Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Statistics in Medicine*, 36(22), 3443–3460.

Holtz, D., Zhao, M., Benzell, S.G., Cao, C.Y., Rahimian, M.A., Yang, J. et al. (2020) Interdependence and the cost of uncoordinated responses to COVID-19. *Proceedings of the National Academy of Sciences of the United States of America*, 117(33), 19837–19843.

Iyer, S., Karrer, B., Citron, D., Kooti, F., Maas, P., Wang, Z. et al. (2020) Large-scale measurement of aggregate human colocation patterns for epidemiological modeling. *medRxiv preprint*.

Kang, D., Choi, H., Kim, J.H. & Choi, J. (2020) Spatial epidemic dynamics of the COVID-19 outbreak in China. *International Journal of Infectious Diseases*, 94, 96–102.

Kimeldorf, G.S. & Wahba, G. (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2), 495–502.

Kissler, S.M., Tedijanto, C., Goldstein, E., Grad, Y.H. & Lipsitch, M. (2020) Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*, 368(6493), 860–868.

Kleiber, C. & Zeileis, A. (2016) Visualizing count data regressions using rootograms. *The American Statistician*, 70(3), 296–303.

Kottasová, I. (2020) Germany reports 650 new coronavirus cases in a meat processing plant - CNN. Available from https://edition.cnn.com/2020/06/18/europe/germany-meat-processing-plant-coronavirus-cases-intl.

Kraemer, M.U., Yang, C.H., Gutierrez, B., Wu, C.H., Klein, B., Pigott, D.M. et al. (2020) The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 368(6490), 493–497.

Kuchler, T., Russel, D. & Stroebel, J. (2021) The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook. *Journal of Urban Economics* (OnlineFirst).

Lavezzo, E., Franchin, E., Ciavarella, C., Cuomo-Dannenburg, G., Barzon, L., Del Vecchio, C. et al. (2020) Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. *Nature*, 584, 425–429.

LeSage, J. & Pace, R.K.P. (2009) *Introduction to spatial econometrics*. Boca Raton: CRC Press.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y. et al. (2020a) Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13), 1199–1207.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W. et al. (2020b) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, 368(6490), 489–493.

Little, R.J.A. & Rubin, D.B. (2002) *Statistical analysis with missing data*. New York: Wiley.

Lorch, L., Trouleau, W., Tsirtsis, S., Szanto, A., Schölkopf, B. & Gomez-Rodriguez, M. (2020) A spatiotemporal epidemic model to quantify the effects of contact tracing, testing and containment. *arXiv preprint*.

Maas, P., Shankar, I., Gros, A., Wonhee, P., Mcgorman, L., Nayak, C. et al. (2019) Facebook disaster maps: aggregate insights for crisis response & recoveryIn: *Proceedings of the 16th ISCRAM conference*, pp. 1–12.

Mardia, K.V. (1978) Some properties of classical multi-dimensional scaling. *Communications in Statistics-Theory and Methods*, 7(13), 1233–1241.

Meyer, S. & Held, L. (2017) Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*, 18(2), 338–351.

Mitze, T., Kosfeld, R., Rode, J. & Walde, K. (2020) Face masks considerably reduce COVID-19 cases in Germany. *Proceedings of the National Academy of Sciences of the United States of America*, 117(51), 32293–32301.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D. et al. (2020) vegan: community ecology package. R package version 2.5-7.

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Delataille, S., De Nadai, M. et al. (2020) Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances*, 6(23), 1–6.

Paul, M., Held, L. & Toschke, A.M. (2008) Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27(29), 6250–6267.

Prem, K., Liu, Y., Russell, T.W., Kucharski, A.J., Eggo, R.M., Davies, N. et al. (2020) The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health*, 5(5), e261–e270.

R Core Team. (2020) R: a language and environment for statistical computing.

Rigby, R.A. & Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3), 507–554.

Riley, S., Ainslie, K.E.C., Eales, O., Walters, C.E., Wang, H., Atchison, C. et al. (2021) Resurgence of SARS-CoV-2: detection by community viral surveillance. *Science*, 372(6545), 990–995.

Ruppert, D., Wand, M. & Carroll, R.J. (2003) *Semiparametric regression*. Cambridge: Cambridge University Press.

Stadtfeld, C. (2018) The micro–macro link in social networks. *Emerging Trends in the Social and Behavioral Sciences* (OnlineFirst).

Stasinopoulos, M., Rigby, B., Voudouris, V. & Kiose, D. (2020) gamlss.add: extra additive terms for generalized additive models for location scale and shape. R package version 5.1-6.

Streeck, H., Schulte, B., Kuemmerer, B., Richter, E., Hoeller, T., Fuhrmann, C. et al. (2020) Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *Nature Communications*, 11(5829), 1–12.

Walter, L.A. & Mcgregor, A.J. (2020) Sex-and gender-specific observations and implications for COVID-19. *Western Journal of Emergency Medicine*, 21(3), 507–509.

WHO (2020) Coronavirus disease (COVID-2019) situation reports. Available from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

Wood, S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95–114.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), 3–36.

Wood, S.N. (2013) On *p*-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S.N. (2017) *Generalized additive models: an introduction with R*. Boca Raton: CRC Press.

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.

Young, G. & Householder, A.S. (1938) Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1), 19–22.

Zeger, S.L. & Qaqish, B. (1988) Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, 44(4), 1019–1031.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Fritz, C., Kauermann, G. (2021) On the interplay of regional mobility, social connectedness and the spread of COVID-19 in Germany. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 00, 1–25. https://doi.org/10.1111/rssa.12753

APPENDIX A

A. MULTIDIMENSIONAL SCALING AND PROCRUSTES ANALYSIS

In order to determine the information given in the pairwise social connectedness indices $x_{soc} = (x_{ij,soc})_{i,j=1,\dots,n}$ for explaining the spread of COVID-19 in Germany, we use techniques from multivariate statistics (Cox & Cox, 2000). Thereby, we can derive a low-dimensional representation of the network on the actor level and guarantee interpretable as well as transparent results. More specifically, we apply *metric multidimensional scaling* (MDS) to represent dissimilarity matrices in a lower-dimensional geometric space that preserves the dissimilarities through Euclidean distances (Borg et al., 2013). To illustrate the application of this algorithm, one can think of MDS as a technique to reverse-engineer geographical coordinates that are unique up to scale and rotation from distances between cities (Young & Householder, 1938).

At first, we transform the similarities expressed by the counts of friendship ties between the districts x_{soc} to dissimilarities. In our application, the measure of dissimilarity is given by $d_{soc} = (\frac{1}{x_{ij,soc}})_{i\neq j=1,\dots,n}$ and $d_{ii,soc} = 0$. While this dissimilarity matrix is symmetric and nonnegative, there is no general guarantee that the entries of d_{soc} are Euclidean. Therefore, we add the constant c to the off-diagonal elements to ensure that the distances between the found coordinates are Euclidean (Cailliez, 1983; Mardia, 1978). In order to estimate these p-dimensional coordinates $x_{i,soc} = (x_{i,1}, \dots, x_{i,p})\ \forall\ i = 1, \dots, n$ from the dissimilarity matrix d_{soc}, the objective is to minimise the squared error between the pairwise entries of d_{soc} and the Euclidean distances calculated with the respective coordinates:

$$x_{soc} = \left(x_{1,soc}^{\top}, \dots, x_{n,soc}^{\top}\right) = \underset{\tilde{x}\in\mathbb{R}^{p\times n}}{\operatorname{argmin}} \left(\sum_{i\neq j} (d_{ij,soc} + c - \|\tilde{x}_i - \tilde{x}_j\|^2) \right)^{1/2}, \tag{A1}$$

in our case we set $p = 2$. See Cox and Cox (2000) and Borg et al. (2013) for methods to find x such that Equation (A1) holds, which are implemented in the R-package `stats` (R Core Team, 2020).

Since arbitrary transformations, rotations and reflections of any coordinates that optimise (A1), represented by $x_{soc} = (x_{1,soc}, \dots, x_{n,soc})$, are equally valid, we further process the solution to guarantee uniqueness and an intuitive understanding of the result. To achieve this goal, we use *Procrustes Analysis* (Cox & Cox, 2008) and find an optimal solution x_{soc} to Equation (A1) that is also most similar to the geographical coordinates $x_{coord} = (x_{1,coord}, \dots, x_{n,coord})$ given in Figure 8. As a measure of similarity between the matrices x_{soc} and x_{coord}, commonly $R^2 = \sum_{i=1}^{n} (x_{i,soc} - x_{i,coord})^{\top}(x_{i,soc} - x_{i,coord})$ is used. Furthermore, we can parameterise the desired class of functions that transform an according to Equation (A1) optimal solution $x_{soc,i}$ to $\tilde{x}_{soc,i}$ by:

$$\tilde{x}_{soc,i} = \rho \mathcal{A}^{\top} x_{soc,i} + b, \tag{A2}$$

where ρ is scalar determining the dilation, $\mathcal{A}$ an orthogonal matrix defining the rotation and reflection, and b a two-dimensional vector for a possible translation. From an optimisation point of view, we now have to find ρ, $\mathcal{A}$, and b such that the resulting R^2 is minimised, which we can do in closed form (see Cox & Cox, 2000). This type of transformation is implemented in the R-package `vegan` (Oksanen et al., 2020) and does not change the estimates or inference because we apply isotropic smooth terms.

B. ESTIMATION OF θ GIVEN c AND COMPLETE DATA

From Equation (4), we construct a likelihood for each district and age/gender group tuple. Combining these separate contributions under independence leads to a joint logarithmic likelihood given by:

$$\ell(\theta, c) \propto \sum_{i=1}^{n} \sum_{g \in \mathcal{G}} \sum_{t=1}^{T} \log\left(\frac{\Gamma(\phi + y_{i,g,t})}{y_{i,g,t}!\Gamma(y_{i,g,t})}\right) + \phi \log\left(\frac{\phi}{\phi + \mu_{i,g,t}}\right) + y_{i,g,t} \log\left(\frac{\mu_{i,g,t}}{\phi + \mu_{i,g,t}}\right). \quad \text{(B1)}$$

note that ϕ is the dispersion parameter of the negative binomial distribution and that the likelihood of the imputation model from Equation (3) in Section 3.1 has the same form with $\phi^{-1} = \sigma_l$. Suppose we plug $\mu_{i,g,t}$ as defined in Equation (5) into (B1) and fix the value of c. In that case, we observe that the result is a function of θ and resembles the likelihood of a generalised additive model with negative binomial distributed target variables and denote the likelihood by $\ell(\theta|c)$ (Ruppert et al., 2003). To obtain a smooth fit of θ, we extend this function by an additive penalisation component:

$$\ell_p(\theta|c) = \ell(\theta|c) - \tau^\top S, \quad \text{(B2)}$$

where $\tau = \left(\tau_a, \tau_b, \tau_{coord}, \tau_{soc}\right)^\top$ are smoothing parameters weighting the term-specific penalties $S = \left(S_a, S_b, S_{coord}, S_{soc}\right)^\top$. The choice of these penalties differs between the random effects and bivariate spacial effects. For the random effects, we follow Ruppert et al. (2003) and define S_a and S_b through ridge penalties, hence, for instance, $S_a = \sum_{i=1}^{2} a_i^2$. In the case of the isotropic semiparametric terms, we chose the penalty terms in accordance with Duchon (1977). Here, S_{coord} penalises the roughness of the bivariate function $f_{coord}(x_{i,coord}) = f_{coord}(x_{i,coord,1}, x_{i,coord,2})$, where $x_{i,coord,p}$ denotes the pth dimension of $x_{i,coord}$ $\forall$ $p \in \{1, 2\}$, in our application the longitude and latitude of district i. Given this notation, we can state the functional form of the penalty term:

$$S_{coord} = \int \frac{\partial^2}{\partial^2 x_{coord,1}} f_{coord}(x_{coord})^2 + 2\frac{\partial^2}{\partial x_{coord,1} \partial x_{coord,2}} f_{coord}(x_{coord})^2 + \frac{\partial^2}{\partial^2 x_{coord,2}} f_{coord}(x_{coord})^2 \, dx_{coord,1} \, dx_{coord,2}.$$

Besides we ensure identifiability of all smooth effects by incorporating a sum-to-zero constraint per term, which translates to $\sum_{i=1}^{n} f_{coord}(x_{i,coord}) = 0$ for $f_{coord}(\cdot)$ (Wood, 2017).

To maximise (B2) in terms of θ and τ, we follow the nested optimisation approach of Wood (2011). Hence, we find $\widehat{\tau}$ in an outer iteration and $\widehat{\theta}$ consecutively in an inner iteration. Generally, the validity of this procedure rests on the finding that $\widehat{\theta}$ is the posterior mode of $\theta|y$ under the assumption that θ follows a zero-mean normal prior with improper variance (Kimeldorf & Wahba, 1970). Viewing θ as random coefficients enables us to estimate all smoothing parameters τ via restricted maximum likelihood estimation. More specifically, we set up $f(y, \theta|c)$ given $\ell(\theta|c)$ and $f(\theta)$. Through integrating θ out of $f(y, \theta|c)$ by deploying a Laplace approximation we obtain an approximate REML criterion, which is a function of τ and ϕ, the dispersion parameter from Equation (B1). Maximising the derived function in terms of these parameters gives $\widehat{\tau}$ and $\widehat{\phi}$ (see Wood, 2011 for additional details). Given the tuning parameters, we consecutively find $\widehat{\theta}$ through standard penalised iterative re-weighted least squares estimates (PIRLS, Wood, 2017) in the inner iteration. We repeat this iterative scheme until convergence to obtain $\widehat{\theta}$ and $\widehat{\tau}$ given a fixed value of c. A scalable implementation of this routine that we used is available in the software package `mgcv` (Wood, 2017).