

# Incorporating social contact data in spatio-temporal models for infectious disease spread

SEBASTIAN MEYER\*, LEONHARD HELD

*Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschgraben 84,  
CH-8001 Zürich, Switzerland*  
sebastian.meyer@uzh.ch

## SUMMARY

Routine public health surveillance of notifiable infectious diseases gives rise to weekly counts of reported cases – possibly stratified by region and/or age group. A well-established approach to the statistical analysis of such surveillance data are endemic-epidemic time-series models. The temporal dependence inherent to communicable diseases is thereby taken into account by an observation-driven formulation conditioning on past counts. Additional spatial dynamics in areal-level counts are largely driven by human travel and can be captured by power-law weights based on the order of adjacency. However, social contacts are highly assortative also with respect to age. For example, characteristic pathways of directly transmitted pathogens are linked to childcare facilities, schools and nursing homes. We therefore investigate how a spatio-temporal endemic-epidemic model can be extended to take social contact data into account. The approach is illustrated in a case study on norovirus gastroenteritis in Berlin, 2011–2014, by age group, city district and week, using additional contact data from the POLYMOD survey. The proposed age-structured model outperforms alternative scenarios with homogeneous or no mixing between age groups. An extended contact model suggests a power transformation of the survey-based contact matrix towards more within-group transmission.

*Key words:* Age-structured contact matrix; Areal count time series; Endemic-epidemic modelling; Infectious disease epidemiology; Norovirus gastroenteritis; Norwalk virus; Spatio-temporal surveillance data.

\*To whom correspondence should be addressed.

## 1. INTRODUCTION

The social phenomenon of “like seeks like” results in striking contact patterns between certain subgroups of a population. The quantification of such social mixing behaviour to model the spread of infectious diseases in human populations has been put forward by the work of [Wallinga and others \(1999\)](#). A recent review of studies measuring epidemiologically relevant social mixing behaviour is provided by [Read and others \(2012\)](#). The largest survey was conducted as part of the EU-funded POLYMOD project and recorded conversational contacts of 7 290 individuals in eight European countries ([Mossong and others, 2008](#)). Contact patterns were found to be remarkably similar across the different countries and highly assortative with respect to age, especially for school children and young adults.

The basic idea behind the combination of such social contact data with epidemic models has been termed the “social contact hypothesis” ([Wallinga and others, 2006](#)): The age distribution of contacts is proportional to the age distribution of individuals exposed to infection. For instance, for pathogens transmitted via the respiratory route, face-to-face conversation and/or physical contact is a reasonable proxy measure for exposure. [Goeyvaerts and others \(2010\)](#) use data on such contacts from the POLYMOD study to estimate the so-called “who acquires infection from whom” (WAIFW) matrix from seroprevalence data. Many other studies have now made use of the POLYMOD contact data as well ([Rohani and others, 2010](#); [Birrell and others, 2011](#); [Baguelin and others, 2013](#); [Goeyvaerts and others, 2015](#)), but none of them takes the spatial characteristics of disease spread into account. The distance of social contacts from the home location of each participant has only recently been investigated by [Read and others \(2014\)](#). Their finding that “most were within a kilometre of the participant’s home, while some occurred further than 500 km away” reflects the power-law distance decay of social interaction as determined by human travel behaviour ([Brockmann and others, 2006](#)). [Meyer and Held \(2014\)](#) found such a power law to translate to the spatial spread of infectious diseases.

The purpose of this paper is to combine both the social and spatial determinants of infectious disease spread in a time-series model for public health surveillance data. Such data are routinely available as weekly counts of reported cases of a notifiable disease aggregated across administrative districts and possibly further stratified by age group or gender. Social contact matrices reflect the amount of mixing between these strata. Since a useful statistical model should take this structure into account, we extend our previously established spatio-temporal model for areal-level count time series ([Meyer and Held, 2014](#)) to an additional stratification factor. In this paper, focus is on age-structured models, but the presented methodology equivalently applies to other strata or combinations thereof, such as age-gender groups. We investigate if incorporating a (possibly adjusted) contact matrix captures disease spread better than more simple assumptions of homogeneous or no mixing between the different groups of the population. The approach also allows us to estimate how much disease incidence in each group can be linked to previous cases in the own and in other groups – while adjusting for the spatial pattern of disease spread.

The structure of this paper is as follows. Section 2 describes our case study on weekly counts of norovirus gastroenteritis in Berlin, 2011–2014, stratified by city district and age group, as well as the age-structured contact matrix from the POLYMOD study that we make use of. Section 3 first outlines the unstratified spatio-temporal modelling framework and then extends it to an additional stratification variable accompanied with a social contact matrix. Section 4 shows the results of applying the proposed time-series model to the Berlin norovirus data. We conclude the paper with a discussion in Section 5. The supplementary material contains the data and software to reproduce the presented analysis.

## 2. CASE STUDY: NOROVIRUS GASTROENTERITIS IN BERLIN, 2011–2014

Most of the aforementioned studies associate contact patterns with the spread of influenza, whereas we here investigate the occurrence of norovirus-associated acute gastroenteritis. Both diseases are highly infectious, have a similar temporal pattern, and similar mortality in elderly persons (van Asten *and others*, 2012). However, in contrast to influenza, vaccines against noroviruses have yet to be developed (Pringle *and others*, 2015). From a statistical modelling perspective, absence of vaccination simplifies the analysis of infectious disease occurrence since vaccination coverage – potentially varying across age groups and regions – needs not to be taken into account. In the following we first give an overview of the general epidemiology of norovirus gastroenteritis, and then describe the surveillance data from Berlin and the social contact data that we use.

### 2.1 Epidemiology of norovirus gastroenteritis

The recent review of Pringle *and others* (2015) summarizes the knowledge of the epidemiology of and immunity to noroviruses. Norovirus-associated acute gastroenteritis is characterized by “sudden onset of vomiting, diarrhea and abdominal cramps lasting 2–3 days”, but the symptom profile varies by age group. O’Dea *and others* (2014) estimate an average symptomatic period of 3.35 days from several outbreaks in hospitals and long-term care facilities, where vulnerable individuals live closely together and norovirus outbreaks most commonly occur. Another group of the population frequently affected by norovirus gastroenteritis are young children in daycare centres. Norovirus incidence exhibits strong seasonality with a peak during winter, where outbreaks in childcare facilities were observed to precede those in private households, hospitals and nursing homes (Bernard *and others*, 2014).

Noroviruses are highly contagious since only few viral particles are needed for an infection, and they are thermally stable and particularly persistent in the environment (Marshall and Bruggink, 2011). Thus, noroviruses are not only transmitted directly from person to person, but also indirectly via contaminated surfaces or food. This may require adjustment of the contact data when incorporated into an epidemic model (Section 3). Furthermore, the duration of fecal norovirus shedding varies by age and is in the range of several weeks. These characteristics suggest a rather large variation in the generation time, i.e. the time between onset of symptoms in a primary and a secondary case, which ranges from within a day to more than one week with a mean of 3.6 days estimated for an outbreak across several daycare facilities in Stockholm (Götz *and others*, 2001; Heijne *and others*, 2009).

### 2.2 Incidence data

In Germany, the national public health institute (the Robert Koch Institute, RKI) provides access to incidence data of notifiable diseases through the *SurvStat@RKI 2.0* online service (<https://survstat.rki.de>) allowing for customized queries to the German notification system database. Since the last revision of the case definition for norovirus gastroenteritis in 2011, only laboratory-confirmed cases are reported to the RKI. Previously, data from epidemiologically linked cases of acute gastroenteritis were collected as well, but this procedure has been cancelled because of its unreliability and high burden on local public health departments (Robert Koch-Institut, 2010). For data generated under a consistent reporting regime, we thus restrict our attention to the time since 2011.

As to the geographic region of interest, we chose the largest city of Germany: Berlin. With respect to RKI’s reporting system, Berlin is unique in that it is divided further into 12 admin-

istrative city districts for each of which disease notification data is separately available. This enables the analysis of disease spread on smaller spatial scales. Furthermore, we expect statistical models for weekly counts by region and further stratified by age group to be more efficient with a large underlying population.

We have downloaded weekly numbers of reported cases of norovirus gastroenteritis according to the current reference definition from *SurvStat@RKI* as of 28 November 2014. The extracted time series cover three norovirus seasons from 2011-W27 to 2014-W26 and are stratified by region (the 12 city districts) and age group (0-4, 5-14, 15-49, 50-64 and 65+ years of age). The age groups were condensed from 5-year intervals to reflect heterogeneous epidemiological mixing of pre-school versus school children in a similar way as in other studies (van Kerckhove *and others*, 2013; Worby *and others*, 2015). Similarly stratified population numbers were obtained from the Statistical Information System Berlin-Brandenburg *StatIS-BBB* (<https://www.statistik-berlin-brandenburg.de/statis>) at the reference date 31 December 2011, where Berlin had 3 501 872 inhabitants in total. To determine and display spatial dependencies in the areal count time series, the geographical shapes of Berlin’s 97 local centres (subsequently aggregated to the 12 city districts) were obtained from the Statistical Office of Berlin-Brandenburg (<http://daten.berlin.de/datensaetze>) as of July 2012.

The top-left panel in Figure 1 shows the weekly numbers of reported norovirus gastroenteritis cases aggregated over all age groups and city districts. There is strong seasonality with weekly counts ranging from 7 to 214 cases, and a mean yearly incidence of 1.0 cases per 1 000 inhabitants. The seasonal pattern is very similar across the three years with a typical (reporting) bump during the Christmas break. The slightly wider epidemic curve in the second season 2012/2013 is related to a large food-borne outbreak from 20 September to 5 October 2012 (weeks 38 to 40), which affected Berlin and several other states in East Germany. The outbreak has been traced back to contaminated frozen strawberries, which were delivered almost exclusively by one large catering company serving schools and childcare facilities (Bernard *and others*, 2014). The peak incidence during this outbreak is thus most pronounced in the time series of the 5 to 14 year old children (top-right panel of Figure 1).

A comparison between the age groups confirms the findings of Bernard *and others* (2014) from analysing norovirus gastroenteritis occurrence over all of Germany from 2001 to 2009: The age groups of pre-school children and of the retired population show the highest reported incidence, and in all seasonal waves, cases in younger individuals are notified earlier than adult cases. The age-structured spatio-temporal modelling approach presented in Section 3 helps to address the question raised by Bernard *and others* (2014), “whether this reflects a pattern of disease transmission from young to old in the community” – taking the spatial aspect of disease spread into account.

How disease incidence varies across the 12 city districts of Berlin is shown in Figure 2. The general incidence level is less heterogeneous between districts than between age groups. The south-western district Steglitz-Zehlendorf tends to be affected more and the central districts tend to be affected less than the remaining districts. This pattern is roughly consistent across age groups. An exception are the two younger age groups, which exhibit a rather high incidence in the two north-western districts Lichtenberg and Marzahn-Hellersdorf, respectively. District-specific seasonal shifts are not apparent (supplementary Figure S1).

Only animated, age-stratified maps of the weekly norovirus disease counts encompass the full information from all three data dimensions. Such an animation can be found in the supplementary material and may provide additional insight into the dynamics of disease spread. However, endemic-epidemic modelling of the data as in Section 3 provides a more structured view and implicitly takes population heterogeneity into account.

### 2.3 Contact data

To inform our model about the degree of mixing and therefore transmission potential between the different age groups, we use the age-structured contact matrix from the German subset of the POLYMOD study (Mossong *and others*, 2008, Supporting Information, Table S5, 8.2). It describes the average number of contact persons (physical and non-physical) recorded per day per survey participant and is depicted in Figure 3 (left). The strong diagonal pattern reflects that people tend to mix with people of the same age, which is especially true for school children. They also have a relatively large number of contact persons. The other prominent pattern is produced by the contacts between parents and children.

To make use of the contact data in our approach, we have to aggregate the matrix to the same age groups as in the surveillance data. This is done by summing over the contact groups (columns) to be joined and calculating the weighted average across the corresponding participant groups (rows), where the weights are the numbers of study participants in the different age groups. The result of this aggregation is shown in the right-hand plot of Figure 3. It is now rather asymmetric because of the different sizes of the involved age groups. For instance, the contacts of pre-school children mainly belong to the age group 15 to 49 years, but the contact persons of that group mainly belong to itself. For the modelling in the next section, only the row-wise distributions will be relevant, i.e. the contact pattern of an infectious participant across the different age groups.

## 3. AN AGE-STRUCTURED SPATIO-TEMPORAL MODEL FOR INFECTIOUS DISEASE COUNTS

We start this section by reviewing a previously proposed endemic-epidemic modelling framework (Meyer and Held, 2014, Section 3) for aggregated spatio-temporal surveillance data. This model is subsequently extended to an additional data dimension to account for the highly assortative mixing of people known from social contact studies.

### 3.1 Spatio-temporal formulation

An endemic-epidemic multivariate time-series model for areal counts  $Y_{rt}$  from regions  $r = 1, \dots, R$  during periods  $t = 1, \dots, T$  was originally proposed by Held *and others* (2005). Conditionally on past observations,  $Y_{rt}$  is assumed to follow a negative binomial distribution with mean  $\mu_{rt}$  and region-specific overdispersion parameters  $\psi_r$  such that the conditional variance of  $Y_{rt}$  is  $\mu_{rt}(1 + \psi_r \mu_{rt})$ . The lower bound  $\psi_r = 0$  yields the Poisson distribution as a special case, and a common assumption is that  $\psi_r = \psi$  is shared across regions. In its most general formulation, the mean of this spatio-temporal model is additively decomposed into *endemic* and *epidemic* components as

$$\mu_{rt} = e_{rt} \nu_{rt} + \lambda_{rt} Y_{r,t-1} + \phi_{rt} \sum_{r' \neq r} [w_{r'r}] Y_{r',t-1} \quad (3.1)$$

with log-linear predictors

$$\begin{aligned} \log(\nu_{rt}) &= \alpha_r^{(\nu)} + \beta^{(\nu)\top} \mathbf{z}_{rt}^{(\nu)}, \\ \log(\lambda_{rt}) &= \alpha_r^{(\lambda)} + \beta^{(\lambda)\top} \mathbf{z}_{rt}^{(\lambda)}, \\ \log(\phi_{rt}) &= \alpha_r^{(\phi)} + \beta^{(\phi)\top} \mathbf{z}_{rt}^{(\phi)}, \end{aligned} \quad (3.2)$$

and normalization  $[w_{r'r}] := w_{r'r} / \sum_j w_{r'j}$ . The intercepts of the three log-linear predictors can be assumed identical across regions, region-specific, or random (Paul and Held, 2011). The regression terms often involve sine-cosine effects of time to reflect seasonally varying incidence (Held and Paul, 2012), but may also capture other explanatory variables, such as vaccination coverage (Herzog and others, 2011). The endemic mean is modelled proportional to an offset of expected counts  $e_{rt}$ , for which we typically use (the fraction of) the population living in region  $r$ . The endemic component generally accounts for cases not directly linked to previously observed cases, e.g. due to heterogeneity in reporting rates, edge effects or indirect transmission via contaminated water or food.

Disregarding the temporal dimension, the model for the mean endemic incidence is very much related to classical disease mapping approaches for non-contagious spatial processes (Wakefield, 2007). However, since surveillance data of infectious diseases exhibit autoregressive behaviour and occasional outbreaks, a so-called epidemic component driven by the observed counts  $(Y_{1,t-1}, \dots, Y_{R,t-1})$  in the previous period is superposed on the endemic component. The epidemic component in (3.1) splits up into autoregressive effects, i.e. reproduction of the disease within region  $r$ , and neighbourhood effects, i.e. transmission from other regions  $r'$ . It has proven informative to account for varying population size  $e_{rt}$  also in  $\log(\phi_{rt}) = \alpha^{(\phi)} + \tau \log(e_{rt})$  (Meyer and Held, 2014). The parameter  $\tau$  thereby determines how “attraction” to a region scales with population size (Xia and others, 2004). Furthermore, the transmission weights  $w_{r'r}$  reflect the flow of infections from region  $r'$  to region  $r$ . These weights may be based on additional movement network data (Paul and others, 2008; Schrödle and others, 2012; Geilhufe and others, 2014), but may also be estimated from the data at hand. A suitable parametric model to reflect epidemiological coupling between regions is a power-law distance decay  $w_{r'r} = o_{r'r}^{-\rho}$  defined in terms of the adjacency order  $o_{r'r}$  in the neighbourhood graph of the regions (Meyer and Held, 2014). With such a power law and the population dependence of  $\phi_{rt}$  mentioned above, the neighbour-driven component of (3.1) becomes similar to the gravity model studied by Xia and others (2004). The relation between the two models is described in detail in Höhle (2016). Note that in our model, the transmission weights are normalized such that  $\sum_r [w_{r'r}] = 1$ , i.e. the  $Y_{r',t-1}$  cases are distributed among the regions proportionally to the corresponding row vector of the weight matrix  $(w_{r'r})$ .

Estimating separate dynamics for the reproduction of the disease within a region on the one hand, and transmission from other regions on the other hand, goes back to the original model formulation of Held and others (2005), where only first-order neighbours have been incorporated:  $w_{r'r} = \mathbb{I}(r' \sim r)$ . The parametric distance weights offer an appealing alternative to reflect predominant local autoregression in a simpler model with a single epidemic component:

$$\mu_{rt} = e_{rt} \nu_{rt} + \phi_{rt} \sum_{r'} [w_{r'r}] Y_{r',t-1}, \quad (3.3)$$

where the choice  $w_{r'r} = (o_{r'r} + 1)^{-\rho}$  gives unit weight to local transmission ( $r' = r$ ) and then decays as a power law in terms of adjacency order. Apart from using less parameters, this two-component formulation has the advantage of being extensible more naturally to an additional stratification factor, here age group.

### 3.2 Extension for stratified areal count time series

The above model formulation takes into account that people do not have equally distributed contacts across all regions. The social contact structure similarly implies non-homogeneous mixing among different subgroups of the population. Extending the above spatio-temporal model to fit areal time series of counts  $Y_{g,r,t}$  stratified by group  $g = 1, \dots, G$  in addition to region, enables

us to relax the simple assumption of homogeneous mixing within each region. Our focus in this paper is on age-structured models such that  $g$  indexes different age groups. More complex strata such as the interaction of age group and gender are equally possible and can be subsumed in the single group index  $g$ . Note, however, that further stratification leads to lower counts per group and region, which might cause identifiability problems in fitting the corresponding models.

We assume that a contact matrix  $\mathbf{C} = (c_{g'g})$  is given, where each entry  $c_{g'g} \geq 0$  quantifies the average number of contacts of an individual of group  $g'$  with individuals of group  $g$ . The spatio-temporal model (3.3) then extends to a three-dimensional version with group-specific contact structure as follows:

$$\mu_{grt} = e_{grt} \nu_{grt} + \phi_{grt} \sum_{g',r'} [c_{g'g} w_{r'r}] Y_{g',r',t-1}. \quad (3.4)$$

The potentially time-varying population offset  $e_{grt}$  is now by group  $\times$  region and the endemic and epidemic predictors may gain group-specific effects. How the counts from the previous period affect the current mean in group  $g$  and region  $r$  is now determined by a product of contact and spatial weights. The product ensures that cases from group  $g'$  in region  $r'$  are ignored both if there are no contacts to group  $g$  or the geographical relation to region  $r$  suggests absence of transmission. The weights are row-normalized as in the unstratified model, now over all combinations of group and region:  $\sum_{g,r} c_{g'g} w_{r'r} = 1$ . Note that this normalization removes any differences in group-specific overall contact rates (the row sums of  $\mathbf{C}$ ). Our model therefore does not distinguish between proportionate mixing, where the rows of the contact matrix only differ by a proportionality factor, and a matrix with identical rows. Instead of group-specific *infectiousness*, the endemic-epidemic model estimates group-specific *susceptibility*: The regression term  $\phi_{grt}$  calibrates the weighted sum of past cases transmitted to group  $g$  in region  $r$ . For example, if  $\phi_{grt} = \phi_g^{(G)} \phi_{rt}^{(RT)}$ , the group-specific effects  $\phi_g^{(G)}$  will adjust the *columns* of the contact matrix.

There are two special cases of the contact structure involved in the epidemic component. First, a contact matrix with identical rows implies that the mixing pattern of the  $Y_{g',r',t-1}$  infectious cases does not depend on the group  $g'$  they belong to. An example of such homogeneous mixing is a contact matrix where each row contains the  $G$  population fractions ( $c_{g'g} = e_g$ ). If  $\phi_{grt}$  contains group-specific effects as mentioned above, a simple matrix of ones ( $c_{g'g} = 1$ ) will induce the same contact structure. The other special case is a diagonal contact matrix  $\mathbf{C} = \mathbf{I}$ , which reflects complete absence of mixing. This is equivalent to modelling the  $G$  areal time series separately using the spatio-temporal regression framework (3.3). However, also in this case of no between-group mixing, the joint model formulation has the advantage of allowing for parsimonious decompositions of  $\nu_{grt}$  and  $\phi_{grt}$  into group and region effects. Borrowing strength across groups this way is especially useful in applications with low counts. In what follows we propose a way to incorporate the given contact matrix  $\mathbf{C}$  in an adaptive way to infer how much interaction between the groups the surveillance data actually support.

### 3.3 Parameterising the contact matrix

Informing epidemic models by external contact data from sociological studies is a sensible approach. Especially for models of low-prevalence diseases, there might be no feasible alternative to taking the contact matrix as granted. However, the contact patterns from such surveys might not fully match the characteristics of disease spread. For example, social networks are known to change during illness (van Kerckhove and others, 2013). We therefore suggest a parsimonious single-parameter approach to adaptively estimate the transmission weights as a function of the given contact matrix  $\mathbf{C}$ .



Our proposal for a parametrisation of the contact matrix is borrowed from [Küchenhoff and others \(2006\)](#), who deal with misclassification in regression models. A misclassification matrix is progressively transformed to establish an association between the amount of misclassification in a covariate and the corresponding parameter estimate. The involved transformation uses the eigendecomposition of the matrix  $\mathbf{C}$  to raise it to the power of  $\kappa \geq 0$ ,

$$\mathbf{C}^\kappa := \mathbf{E}\mathbf{\Lambda}^\kappa\mathbf{E}^{-1}, \quad (3.5)$$

where  $\mathbf{\Lambda}$  is the diagonal matrix of eigenvalues and  $\mathbf{E}$  is the corresponding matrix of eigenvectors. Translated to our setting, the parameter  $\kappa$  measures the amount of transmission between the different groups of the population. Specifically,  $\kappa = 0$  corresponds to complete absence of between-group transmission ( $\mathbf{C} = \mathbf{I}$ ), whereas  $\kappa = 1$  leaves the social contact matrix unchanged. If  $\mathbf{C}$  is row-normalized, the behaviour of  $\mathbf{C}^\kappa$  for  $\kappa > 1$  can be derived from regarding  $\mathbf{C}$  as the transition matrix of a homogeneous Markov chain ([Grimmett and Stirzaker, 2001](#), Chapter 6). In particular, for  $\kappa \rightarrow \infty$ , all rows of  $\mathbf{C}^\kappa$  converge to the stationary distribution of the process (if it exists). Thus, with increasing  $\kappa$  the transmission pattern becomes independent of the group the infected individual belongs to. Because of this interpretation, numerical advantages, and row-normalization in our model anyway, we assume a row-normalized contact matrix in the remainder of this paper, but keep the simple notation  $\mathbf{C}$ .

Conditions for the existence of  $\mathbf{C}^\kappa$  are briefly discussed in [Küchenhoff and others \(2006, Appendix A\)](#). The basic requirement that  $\mathbf{C}$  can be factorized by an eigendecomposition will hold in most practical cases. However, we also need to make sure that  $\mathbf{C}^\kappa$  has non-negative entries for  $\kappa < 1$ . In our application, there is only one negative entry in  $\mathbf{C}^\kappa$  for values of  $\kappa$  close to 0. We therefore follow a pragmatic approach of truncating entries at 0 which would otherwise become negative. See [Figure 4b](#) for an illustration of how diagonal and off-diagonal entries, respectively, are affected by raising  $\mathbf{C}$  to the power of  $\kappa$ .

Overall, the parametrisation (3.5) enables an assessment of how much a social contact matrix translates to the transmission pattern between the groups of the population. [Figure 4a](#) exemplifies this for the aggregated contact matrix from [Figure 3](#).

### 3.4 Inference

Likelihood inference for the multivariate count time-series model (3.1) has been established by [Paul and Held \(2011\)](#) with extensions for parametric neighbourhood weights by [Meyer and Held \(2014\)](#). The log-likelihood is maximized numerically using the quasi-Newton algorithm available through the R function `nLminb()` ([R Core Team, 2015](#)). Supplied with analytical formulae for the score function and Fisher information, convergence is fast also for a large number of parameters. The associated modelling framework is implemented in the R package `surveillance` ([Meyer and others, 2015](#), Section 5).

The extended age-structured model (3.4) is built on top of the existing inference framework, tuning the implementation to allow for a group-specific power law and group-specific overdispersion. The power parameter  $\kappa$  of (3.5) is conveniently estimated via a profile likelihood approach (see e.g. [Held and Sabanés Bové, 2014](#), Section 5.3), where we maximize the log-likelihood of a model with fixed contact matrix  $\mathbf{C}^\kappa$  as a function of  $\kappa$ . Thus, the estimate and 95% profile confidence interval of  $\kappa$  incorporate the uncertainty with respect to all other model parameters.



## 4. RESULTS

Here we present results of applying an age-structured spatio-temporal model of the form (3.4) to the norovirus surveillance data described in Section 2. To account for the heterogeneous numbers of cases per age group, we allow for group-specific overdispersion parameters  $\psi_g$ . For the mean, the following endemic-epidemic structure is assumed:

$$\begin{aligned} \mu_{grt} = e_{gr} \exp \left\{ \alpha_g^{(G)} + \alpha_r^{(R)} + \beta x_t + \gamma_g \sin(\omega t) + \delta_g \cos(\omega t) \right\} \\ + \phi_g^{(G)} \phi_r^{(R)} e_{gr}^\tau \sum_{g', r'} [(\mathbf{C}^\kappa)_{g'g} (o_{r'r} + 1)^{-\rho}] Y_{g', r', t-1}, \end{aligned} \quad (4.6)$$

for  $g = 1, \dots, 5$  age groups (0-4, 5-14, 15-49, 50-64 and 65+ years of age),  $r = 1, \dots, 12$  city districts of Berlin, and  $t = 1, \dots, 156$  weeks with  $t = 1$  corresponding to 2011-W27. The fraction of the population in age group  $g$  and district  $r$ ,  $e_{gr}$ , enters both the endemic component as a multiplicative offset and the epidemic component as a scaling factor. The endemic predictor allows for age- and region-specific incidence levels, reduced reporting during the Christmas break ( $x_t = 1$  in calendar weeks 52 and 1, otherwise  $x_t = 0$ ), as well as age-specific yearly seasonality ( $\omega = 2\pi/52$ ). Transmission between age groups is modelled using the row-normalized version of the age-structured contact matrix from the POLYMOD study, raised to the power of  $\kappa$  via the eigendecomposition (3.5). Transmission between districts is quantified by a power law with respect to adjacency order. The proportion of the mean  $\mu_{grt}$  linked to the weighted sum of past counts from all  $5 \times 12$  strata depends on both age group (via  $\phi_g^{(G)} > 0$ ) and district (via  $\phi_r^{(R)} > 0$ ). This mainly reflects that certain age groups or districts may be more susceptible to infection than others. Furthermore, these scaling parameters are influenced by the degree of under-reporting, since missing case reports lead to missing links between cases and thus lower the epidemic component. For instance, it is known that under-reporting of norovirus illness is most pronounced in the 20 to 29 year-old persons and substantially lower in persons aged  $< 10$  years and 70 years and over (Bernard and others, 2014).

Table 1 summarizes competing models with respect to the choice of contact structure between the age groups. It turns out that a superposed epidemic model component greatly improves upon a purely endemic model, and that incorporating the contact matrix from the POLYMOD study outperforms naive models with homogeneous or no mixing between age groups. Akaike's Information Criterion (AIC) is minimal for the model with a power-adjusted contact matrix  $\mathbf{C}^\kappa$  (last row), where the exponent is estimated to be  $\hat{\kappa} = 0.64$  (95% CI: 0.45 to 0.92). This means that the epidemic part subsumes more information from cases in the own age group than suggested by the original contact matrix from the POLYMOD study (cf. Figure 4a). The difference in AIC associated with this adjustment, however, is minor compared to the improvement achieved by replacing the naive assumptions on  $\mathbf{C}$  by the POLYMOD contact matrix in the first place.

The spatial spread of the disease across city districts is estimated to have a strong distance decay with  $\hat{\rho} = 2.22$  (95% CI: 1.89 to 2.62), i.e. the weights attributed to the adjacency orders 0 to 4 are 1.00, 0.21, 0.09, 0.05 and 0.03. Since human movement patterns are known to depend on age (Read and others, 2014), we have also tried to estimate age-dependent decay parameters replacing  $\rho$  by  $\rho_{g'}$  in (4.6) (with  $\kappa = 1$  for simplicity). However, the parameter for the low-incidence group of 5 to 14 year old children was not identifiable. If we thus assumed a shared decay parameter for both the 0 to 4 and 5 to 14 year old children, the model becomes identifiable but the resulting estimates are very similar for all age groups (supplementary Figure S2). Furthermore, we have assessed the basic power-law assumption by unconstrained estimates of the order-specific weights as in Meyer and Held (2014), which turn out to be in line with the power law (Figure S2).

In accordance with the idea of a gravity model (Xia *and others*, 2004), we find that the amount of weekly disease counts linked to cases from the previous week scales with the population size of the “importing” district and age group. Similar to a previous application on influenza (Meyer and Held, 2014), the corresponding estimate  $\hat{\tau} = 0.93$  (95% CI: 0.56 to 1.30) is slightly below unity and provides strong evidence for such an association. A model with identical population offsets in both the endemic and epidemic component, i.e.  $\tau = 1$ , as assumed by Xia *and others* (2004) would provide a similar fit.

To understand how much the weekly disease incidence in the individual age groups can be linked to cases from the previous week, Figure 5 shows the endemic-epidemic decomposition of the estimated mean aggregated across districts (the supplementary Figures S3 and S4 show the district-level fit). When aggregated over all age groups, the epidemic component explains about half of the weekly disease incidence. Accordingly, when reformulating the model as a multivariate branching process with immigration (Held and Paul, 2012), the largest eigenvalue of the so-called next-generation matrix holding the estimated coefficients of  $Y_{g',r',t-1}$  is 0.72. However, this overall epidemic proportion mostly reflects the situation for the well-reported 65+ group where the within-group spread is dominating. In contrast, for the group of 5 to 14 year old school children, almost no dependence on past counts of the same or the other age groups can be identified. Interestingly, the groups of 15 to 49 and 50 to 64 year old persons seem to inherit a relevant proportion of cases from other age groups with smaller reproduction in their own group. The youngest age group, though, depends on its own cases rather than on cases in other age groups, which is probably related to this age group having the earliest onset of the yearly wave of norovirus infections. The estimated age-dependent sine-cosine effects on the endemic component, which capture these shifts, are shown in Figure 6 and confirm the findings of Bernard *and others* (2014). The group-specific modal endemic incidence is in calendar weeks 48 (0-4), 43 (5-14), 51 (15-49), 51 (50-64) and 2 (65+), respectively. The largest amplitude is estimated for the youngest and oldest age groups, while only little seasonal variation was estimated for the group of 5 to 14 year old children.

The estimated group-specific overdispersion parameters are 0.22 (0-4), 2.32 (5-14), 0.03 (15-49), 0.17 (50-64) and 0.42 (65+) in the AIC-optimal model with the adjusted contact matrix. The level of overdispersion varies remarkably between the age groups and is largest in the group of 5 to 14 year old children. This is partly due to the relatively strong peak this group experiences during the food-borne outbreak in 2012, for which the model does not explicitly account in the mean structure. The estimated overdispersion parameters are similar between the different epidemic models of Table 1, but slightly larger in the endemic-only model.

## 5. DISCUSSION

In this paper, we have shown how a social contact matrix can be incorporated in an adaptive way in a regression-oriented endemic-epidemic time-series model for stratified, areal-level infectious disease counts. This three-dimensional approach provides more detailed insight into the dynamics of disease spread than unstratified areal models, which inherently assume homogeneous mixing within each spatial unit. Furthermore, a joint model can borrow strength across districts and groups to identify parameters which could not be estimated in separate group-specific areal models, say.

Our application to age-stratified norovirus incidence in Berlin revealed superiority of the contact model compared to naive assumptions of either homogeneous or no mixing between age groups. The model further improved when adjusting the contact matrix from the POLYMOD

survey towards more transmission within the own age group. The estimated strong power-law decay in the spatial dimension means that new infections are predominantly explained by past cases from the same district, with a considerably smaller dependence on districts further apart.

The spatial power law was originally motivated by human travel behaviour on larger scales. An alternative formulation for our city-level data would be a single-step kernel which only discriminates within-district spread from homogeneous transmission to other districts, i.e.  $w_{r'r} = 1 \cdot \mathbb{I}(o_{r'r} = 0) + \rho \cdot \mathbb{I}(o_{r'r} > 0)$ . Such a kernel provides a comparable fit in our application, but we have preferred the power-law model as a more sensible formulation, in particular for spatial extrapolation. Another option to model spatial dependence comes to mind when interpreting the proposed model as a multidimensional “patch” or metapopulation model with stratification by region as well as age group. In this sense, one could try to replace the parametric power law by a social contact matrix, which is stratified by spatial distance in addition to age group. This would provide an alternative to using separate movement network data for  $w_{r'r}$  to quantify the strength of epidemiological coupling between regions.

Social contacts are assortative not only with respect to age and distance. As recently shown by [Barclay and others \(2014\)](#), school children even tend to mix according to their vaccination status. A model adaption to more complex contact structures is straightforward provided correspondingly stratified surveillance data is available. However, a potentially more severe simplification of our model is the assumption of a time-constant contact matrix. Although, the obvious weekday vs. weekend differences in contact behaviour are not relevant for weekly time-series models, there are other seasonal effects on larger time scales. Most importantly, the contact structure of school children changes considerably between regular and school holiday periods as was shown in the POLYMOD data ([Hens and others, 2009](#)). The model could thus be further tuned by estimating seasonality also in the epidemic component ([Held and Paul, 2012](#)).

## 6. SOFTWARE

All analyses were performed using the statistical software environment R 3.2.2 ([R Core Team, 2015](#)) with the package `surveillance` ([Höhle and others, 2015](#)) providing the basic infrastructure for visualization, modelling and inference. See [Meyer and others \(2015\)](#) for a recent guide to the three endemic-epidemic modelling frameworks offered by the `surveillance` package. The data and code to reproduce the analyses in this paper are available in the dedicated R package `hhh4contacts`, which is part of the supplementary material.

## SUPPLEMENTARY MATERIAL

The reader is referred to the Supplementary Materials for animated maps of norovirus infections by age group in Berlin, 2011–2014, and additional figures as referenced in Sections 2 and 4. We also provide the R source package `hhh4contacts` containing the data and code to reproduce the results of Section 4. Run `demo("hhh4contacts")` after installing and loading the package.

## FUNDING

This work was supported by the Swiss National Science Foundation [[project #137919](#)].

## REFERENCES

- BAGUELIN, M., FLASCHE, S., CAMACHO, A., DEMIRIS, N., MILLER, E. AND EDMUNDS, W. J. (2013). Assessing optimal target populations for influenza vaccination programmes: An evidence synthesis and modelling study. *PLOS Medicine* **10**(10), e1001527.
- BARCLAY, V. C., SMIESZEK, T., HE, J., CAO, G., RAINEY, J. J., GAO, H., UZICANIN, A. AND SALATHÉ, M. (2014). Positive network assortativity of influenza vaccination at a high school: Implications for outbreak risk and herd immunity. *PLoS ONE* **9**(2), e87042.
- BERNARD, H., FABER, M., WILKING, H., HALLER, S., HÖHLE, M., SCHIELKE, A., DUCOMBLE, T., SIFFCZYK, C., MERBECKS, S. S., FRICKE, G., HAMOUDA, O., STARK, K., WERBER, D. *and others.* (2014a). Large multistate outbreak of norovirus gastroenteritis associated with frozen strawberries, Germany, 2012. *Eurosurveillance* **19**(8).
- BERNARD, H., HÖHNE, M., NIENDORF, S., ALTMANN, D. AND STARK, K. (2014b). Epidemiology of norovirus gastroenteritis in Germany 2001–2009: Eight seasons of routine surveillance. *Epidemiology & Infection* **142**(1), 63–74.
- BERNARD, H., WERBER, D. AND HÖHLE, M. (2014c). Estimating the under-reporting of norovirus illness in Germany utilizing enhanced awareness of diarrhoea during a large outbreak of Shiga toxin-producing *E. coli* O104:H4 in 2011 – a time series analysis. *BMC Infectious Diseases* **14**(1), 116.
- BIRRELL, P. J., KETSETZIS, G., GAY, N. J., COOPER, B. S., PRESANIS, A. M., HARRIS, R. J., CHARLETT, A., ZHANG, X. S., WHITE, P. J., PEBODY, R. G. *and others.* (2011). Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proceedings of the National Academy of Sciences of the United States of America* **108**(45), 18238–18243.
- BROCKMANN, D., HUFNAGEL, L. AND GEISEL, T. (2006). The scaling laws of human travel. *Nature* **439**(7075), 462–465.
- GEILHUF, M., HELD, L., SKRØVSETH, S. O., SIMONSEN, G. S. AND GODTLIEBSEN, F. (2014). Power law approximations of movement network data for modeling infectious disease spread. *Biometrical Journal* **56**(3), 363–382.
- GOEYVAERTS, N., HENS, N., OGUNJIMI, B., AERTS, M., SHKEDY, Z., VAN DAMME, P. AND BEUTELS, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **59**(2), 255–277.
- GOEYVAERTS, N., WILLEM, L., VAN KERCKHOVE, K., VANDENDIJK, Y., HANQUET, G., BEUTELS, P. AND HENS, N. (2015). Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence. *Epidemics* **13**, 1–9.
- GRIMMETT, G. R. AND STIRZAKER, D. R. (2001). *Probability and Random Processes*, 3rd edition. Oxford: Oxford University Press.
- GÖTZ, H., EKDAHL, K., LINDBÄCK, J., DE JONG, B., HEDLUND, K. O. AND GIESECKE, J. (2001). Clinical spectrum and transmission characteristics of infection with Norwalk-like virus: Findings from a large community outbreak in Sweden. *Clinical Infectious Diseases* **33**(5), 622–628.

- HEIJNE, J. C. M., TEUNIS, P., MORROY, G., WIJKMANS, C., OOSTVEEN, S., DUIZER, E., KRETZSCHMAR, M. AND WALLINGA, J. (2009). Enhanced hygiene measures and norovirus transmission during an outbreak. *Emerging Infectious Diseases* **15**(1), 24–30.
- HELD, L., HÖHLE, M. AND HOFMANN, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* **5**(3), 187–199.
- HELD, L. AND PAUL, M. (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal* **54**(6), 824–843.
- HELD, L. AND SABANÉS BOVÉ, D. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Berlin: Springer.
- HENS, N., AYELE, G., GOEYVAERTS, N., AERTS, M., MOSSONG, J., EDMUNDS, J. AND BEUTELS, P. (2009). Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. *BMC Infectious Diseases* **9**(1), 187.
- HERZOG, S. A., PAUL, M. AND HELD, L. (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. *Epidemiology & Infection* **139**(04), 505–515.
- HÖHLE, M. (2016). Infectious Disease Modelling. In: Lawson, A. B., Banerjee, S., Haining, R. P. and Ugarte, M. D. (editors), *Handbook of Spatial Epidemiology*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC. Forthcoming.
- HÖHLE, M., MEYER, S. AND PAUL, M. (2015). *surveillance: Temporal and spatio-temporal modeling and monitoring of epidemic phenomena*. R package version 1.10-0.
- KÜCHENHOFF, H., MWALILI, S. M. AND LESAFFRE, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* **62**(1), 85–96.
- MARSHALL, J. A. AND BRUGGINK, L. D. (2011). The dynamics of norovirus outbreak epidemics: Recent insights. *International Journal of Environmental Research and Public Health* **8**(4), 1141–1149.
- MEYER, S. AND HELD, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics* **8**(3), 1612–1639.
- MEYER, S., HELD, L. AND HÖHLE, M. (2015). Spatio-temporal analysis of epidemic phenomena using the R package `surveillance`. Revised for the *Journal of Statistical Software*. Preprint available from <http://arxiv.org/abs/1411.0416>.
- MOSSONG, J., HENS, N., JIT, M., BEUTELS, P., AURANEN, K., MIKOLAJCZYK, R., MASSARI, M., SALMASO, S., TOMBA, G. S., WALLINGA, J., HEIJNE, J., SADKOWSKA-TODYS, M., ROSINSKA, M. *and others*. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine* **5**(3), e74.
- O’DEA, E. B., PEPIN, K. M., LOPMAN, B. A. AND WILKE, C. O. (2014). Fitting outbreak models to data from many small norovirus outbreaks. *Epidemics* **6**, 18–29.
- PAUL, M. AND HELD, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine* **30**(10), 1118–1136.

- PAUL, M., HELD, L. AND TOSCHKE, A. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* **27**(29), 6250–6267.
- PRINGLE, K., LOPMAN, B., VEGA, E., VINJE, J., PARASHAR, U. D. AND HALL, A. J. (2015). Noroviruses: Epidemiology, immunity and prospects for prevention. *Future Microbiology* **10**(1), 53–67.
- R CORE TEAM. (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- READ, J. M., EDMUNDS, W. J., RILEY, S., LESSLER, J. AND CUMMINGS, D. A. T. (2012). Close encounters of the infectious kind: Methods to measure social mixing behaviour. *Epidemiology & Infection* **140**(12), 2117–2130.
- READ, J. M., LESSLER, J., RILEY, S., WANG, S., TAN, L. J., KWOK, K. O., GUAN, Y., JIANG, C. Q. AND CUMMINGS, D. A. T. (2014). Social mixing patterns in rural and urban areas of southern China. *Proceedings of the Royal Society of London B: Biological Sciences* **281**(1785).
- ROBERT KOCH-INSTITUT. (2010). Überarbeitete Falldefinitionen zur Übermittlung von Nachweisen von Denguevirus und Norovirus und Erkrankungs- oder Todesfällen an Denguefieber und an Norovirus-Gastroenteritis. *Epidemiologisches Bulletin* **2010**(49), 494–495. In German.
- ROHANI, P., ZHONG, X. AND KING, A. A. (2010). Contact network structure explains the changing epidemiology of pertussis. *Science* **330**(6006), 982–985.
- SCHRÖDLE, B., HELD, L. AND RUE, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics* **68**(3), 736–744.
- VAN ASTEN, L., VAN DEN WIJNGAARD, C., VAN PELT, W., VAN DE KASSTEELE, J., MEIJER, A., VAN DER HOEK, W., KRETZSCHMAR, M. AND KOOPMANS, M. (2012). Mortality attributable to 9 common infections: Significant effect of influenza A, respiratory syncytial virus, influenza B, norovirus, and parainfluenza in elderly persons. *Journal of Infectious Diseases* **206**(5), 628–639.
- VAN KERCKHOVE, K., HENS, N., EDMUNDS, W. J. AND EAMES, K. T. D. (2013). The impact of illness on social networks: Implications for transmission and control of influenza. *American Journal of Epidemiology* **178**(11), 1655–1662.
- WAKEFIELD, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* **8**(2), 158–183.
- WALLINGA, J., EDMUNDS, W. J. AND KRETZSCHMAR, M. (1999). Perspective: human contact patterns and the spread of airborne infectious diseases. *Trends in Microbiology* **7**(9), 372–377.
- WALLINGA, J., TEUNIS, P. AND KRETZSCHMAR, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology* **164**(10), 936–944.
- WORBY, C. J., CHAVES, S. S., WALLINGA, J., LIPSITCH, M., FINELLI, L. AND GOLDSTEIN, E. (2015). On the relative role of different age groups in influenza epidemics. *Epidemics* **13**, 10–16.
- XIA, Y., BJØRNSTAD, O. N. AND GRENFELL, B. T. (2004). Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics. *The American Naturalist* **164**(2), 267–281.

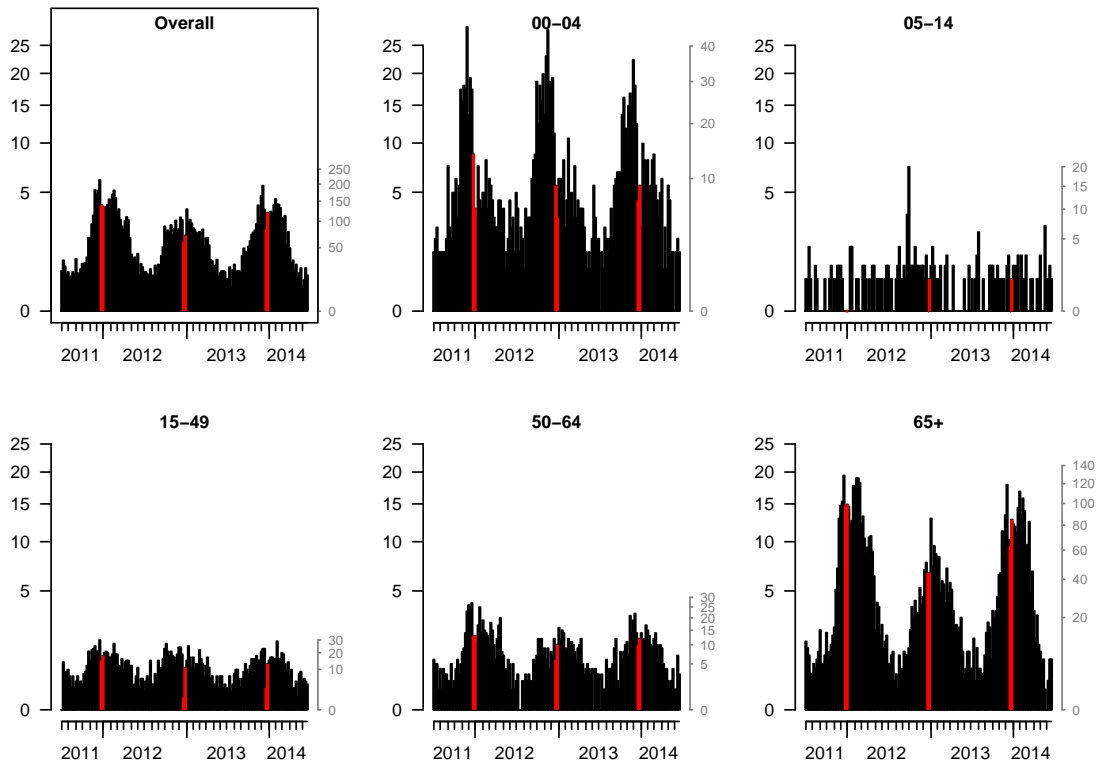


Figure 1: Weekly overall and age-stratified norovirus gastroenteritis incidence (per 100 000 inhabitants) in Berlin. The incidence on the left axis obeys the same  $\sqrt{\cdot}$ -scale in all panels, while the corresponding counts can be read off from the right axis. The population sizes of the age groups are 160 885, 271 574, 1 736 903, 669 243 and 663 267, respectively. The yearly Christmas break in calendar weeks 52 and 1 is highlighted.



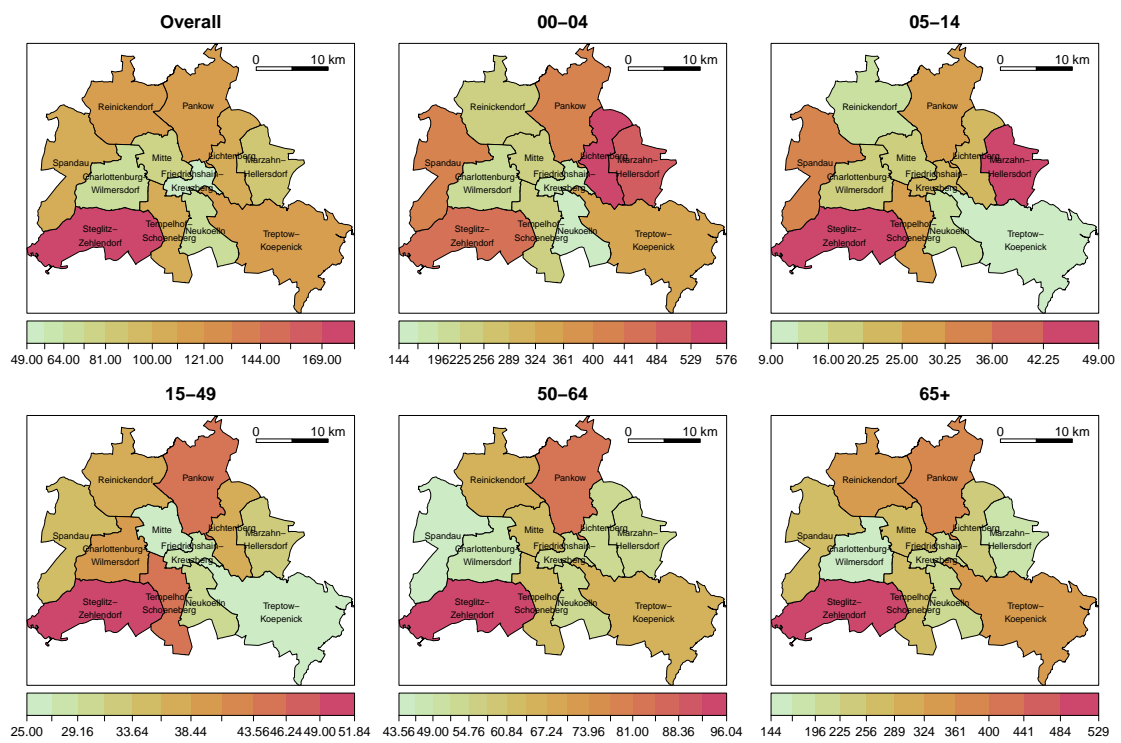


Figure 2: Overall and age-stratified maps of mean yearly norovirus gastroenteritis incidence (per 100 000 inhabitants) in Berlin's city districts, from 2011-W27 to 2014-W26.

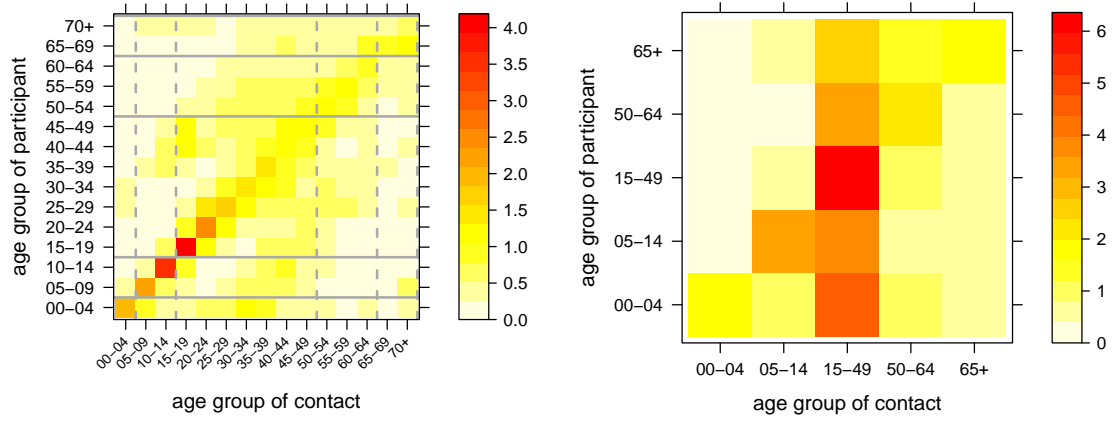
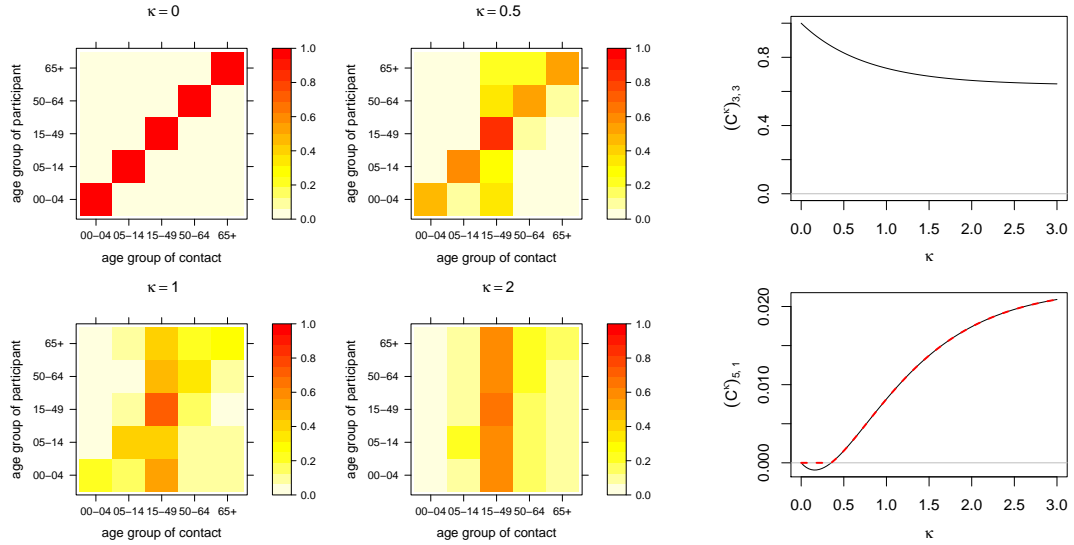


Figure 3: Age-structured contact matrix in 5-year intervals as obtained from *Mossong and others (2008)* (left) and its aggregated version with age groups matching the surveillance data (right). The entries refer to the average daily number of contact persons per participant.



(a)  $C^\kappa$  for different values of  $\kappa$ . Moving from the original matrix at  $\kappa = 1$  to the limiting case  $\kappa = 0$ , off-diagonal values are pushed row-wise towards the diagonal. For increasing  $\kappa$  the row-wise distributions approach each other.

(b) The diagonal entry  $[3, 3]$  and the off-diagonal entry  $[5, 1]$  of  $C^\kappa$  as a function of  $\kappa$ . The dashed line shows the latter truncated at zero.

Figure 4: The effect of raising the row-normalized version of the aggregated contact matrix  $C$  from Figure 3 to the power  $\kappa$  as defined in (3.5).

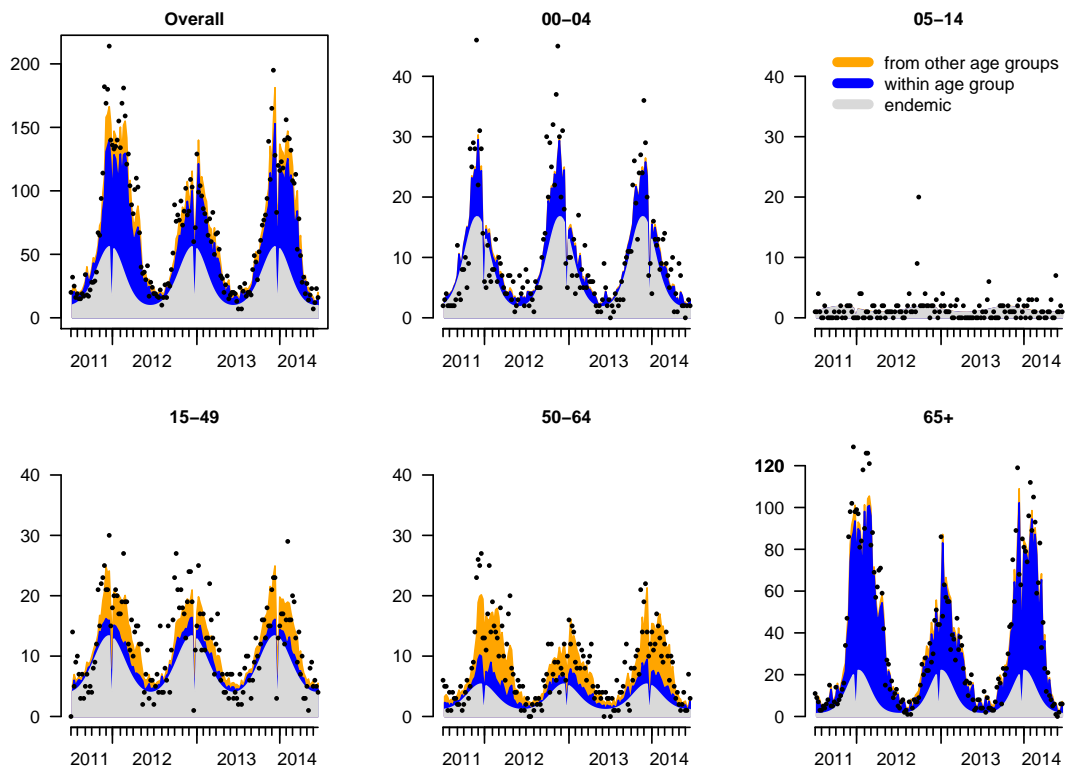


Figure 5: Fitted mean components from the AIC-optimal model, aggregated over all districts. The counts in the top-left plot are aggregated over all age groups. The dots correspond to the reported numbers of cases.

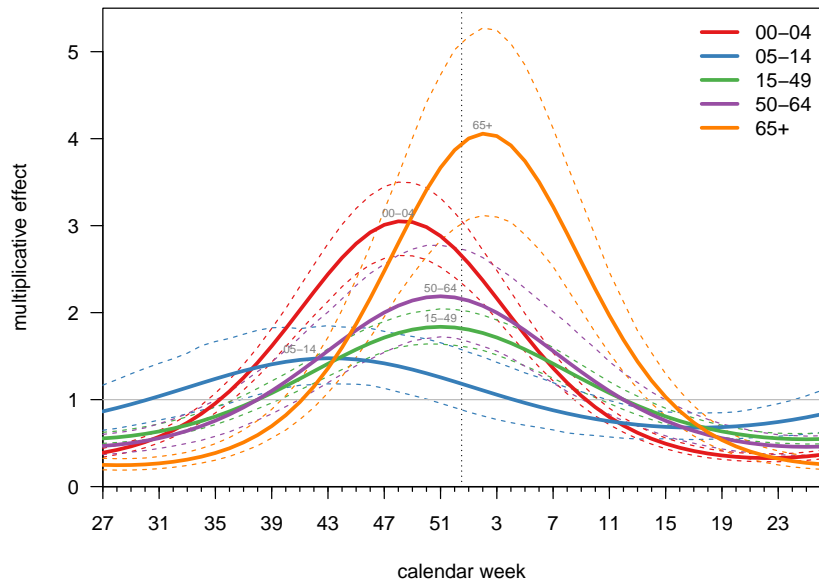


Figure 6: Estimated age-dependent seasonality of the endemic model component with 95% point-wise confidence intervals.

Table 1: Summaries of different models for the age-stratified, areal surveillance data of norovirus gastroenteritis in Berlin. For reference, the first row lists the fit of the purely endemic model, which assumes all counts to be independent. The remaining rows correspond to endemic-epidemic models with a spatial power law but varying assumptions on the contact matrix  $\mathbf{C}$  between age groups. The columns refer to the following model characteristics: the number of parameters, Akaike's Information Criterion, the power  $\tau$  of the population scaling factor, the decay parameter  $\rho$  of the spatial power law, and the power adjustment  $\kappa$  of the contact matrix. The parameter columns contain the estimate and a 95% Wald confidence interval.

	dim	AIC	$\tau$	$\rho$	$\kappa$
purely endemic model	32	20947.8	–	–	–
homogeneous mixing ( $\mathbf{C} = \mathbf{1}$ )	50	20609.7	1.10 (0.74 to 1.46)	2.44 (2.03 to 2.94)	–
no mixing ( $\mathbf{C} = \mathbf{I}$ )	50	20542.9	0.65 (0.22 to 1.08)	2.02 (1.68 to 2.43)	–
original contact matrix $\mathbf{C}$	50	20477.0	0.97 (0.61 to 1.33)	2.26 (1.92 to 2.67)	–
adjusted contact matrix $\mathbf{C}^\kappa$	51	20471.0	0.93 (0.56 to 1.30)	2.22 (1.89 to 2.62)	0.64 (0.45 to 0.92)