FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACULTY OF MEDICINE

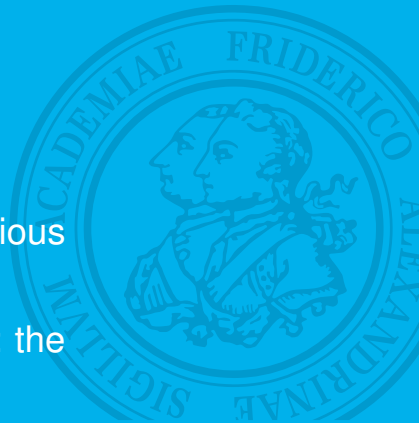
Spatio-temporal modelling and probabilistic forecasting of infectious disease counts

Sebastian Meyer
Institute of Medical Informatics, Biometry, and Epidemiology
Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
7 September 2017

Joint work with Johannes Bracher and Leonhard Held (University of Zurich):

Meyer and Held (2017): Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics* 18:338–351. DOI:10.1093/biostatistics/kxw051

Held, Meyer and Bracher (2017): Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Statistics in Medicine* 36:3443–3460. DOI:10.1002/sim.7363



World Health Organization 2014

Forecasting disease outbreaks is still in its infancy, however, unlike weather forecasting, where substantial progress has been made in recent years.

World Health Organization 2014

> *Forecasting disease outbreaks is still in its infancy, however, unlike weather forecasting, where substantial progress has been made in recent years.*

Key requirements to forecast infectious disease incidence

1. Multivariate view to predict incidence in different regions and subgroups

World Health Organization 2014

> *Forecasting disease outbreaks is still in its infancy, however, unlike weather forecasting, where substantial progress has been made in recent years.*

Key requirements to forecast infectious disease incidence

1. Multivariate view to predict incidence in different regions and subgroups
2. Stratified count time series from routine public health surveillance

World Health Organization 2014

> *Forecasting disease outbreaks is still in its infancy, however, unlike weather forecasting, where substantial progress has been made in recent years.*

Key requirements to forecast infectious disease incidence

1. Multivariate view to predict incidence in different regions and subgroups
2. Stratified count time series from routine public health surveillance
3. Useful statistical models for such dependent data

World Health Organization 2014

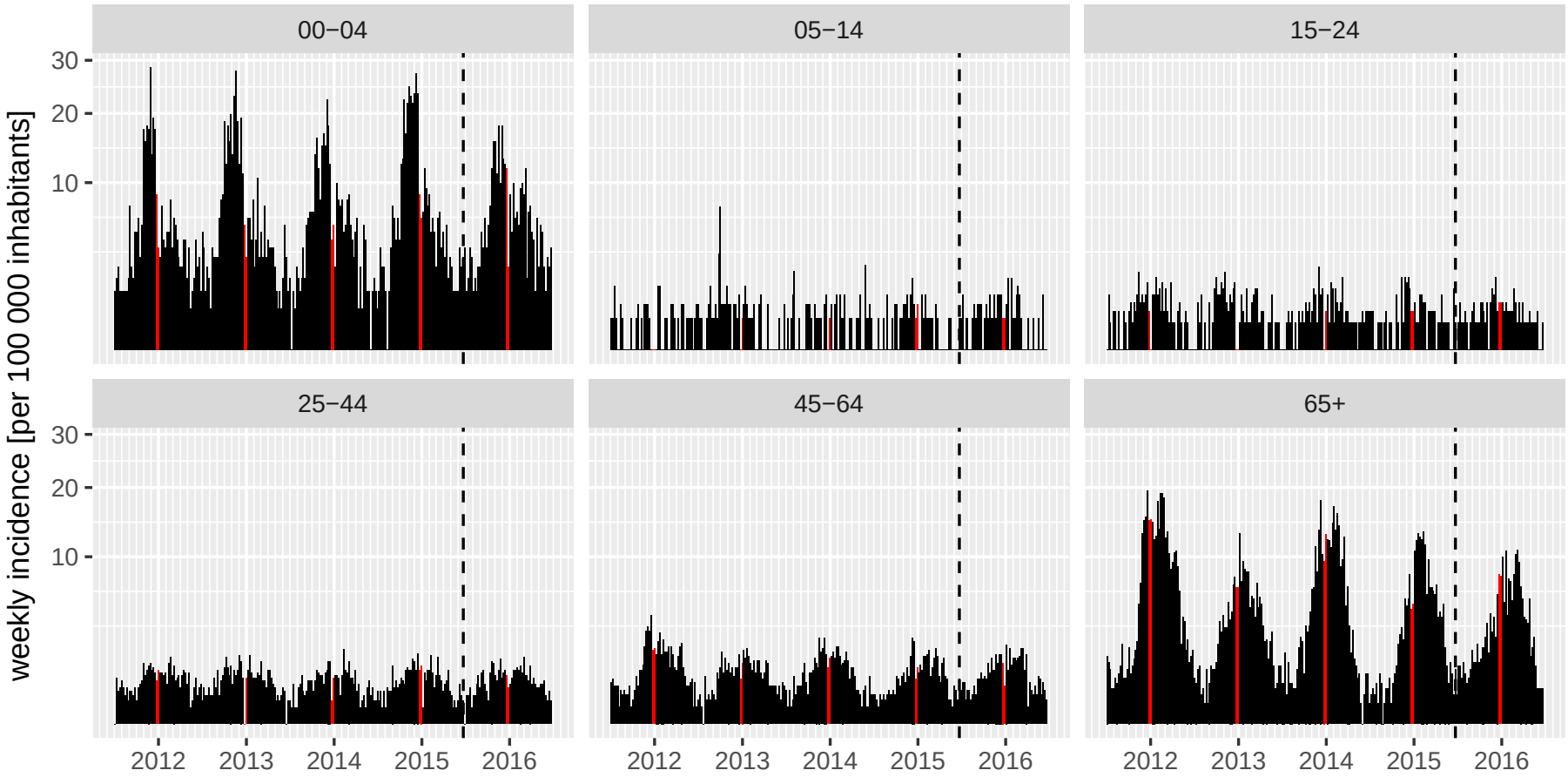
> *Forecasting disease outbreaks is still in its infancy, however, unlike weather forecasting, where substantial progress has been made in recent years.*

Key requirements to forecast infectious disease incidence

1. Multivariate view to predict incidence in different regions and subgroups
2. Stratified count time series from routine public health surveillance
3. Useful statistical models for such dependent data
4. Suitable measures for the evaluation of probabilistic forecasts

Case study: norovirus gastroenteritis in Berlin, 2011–2016

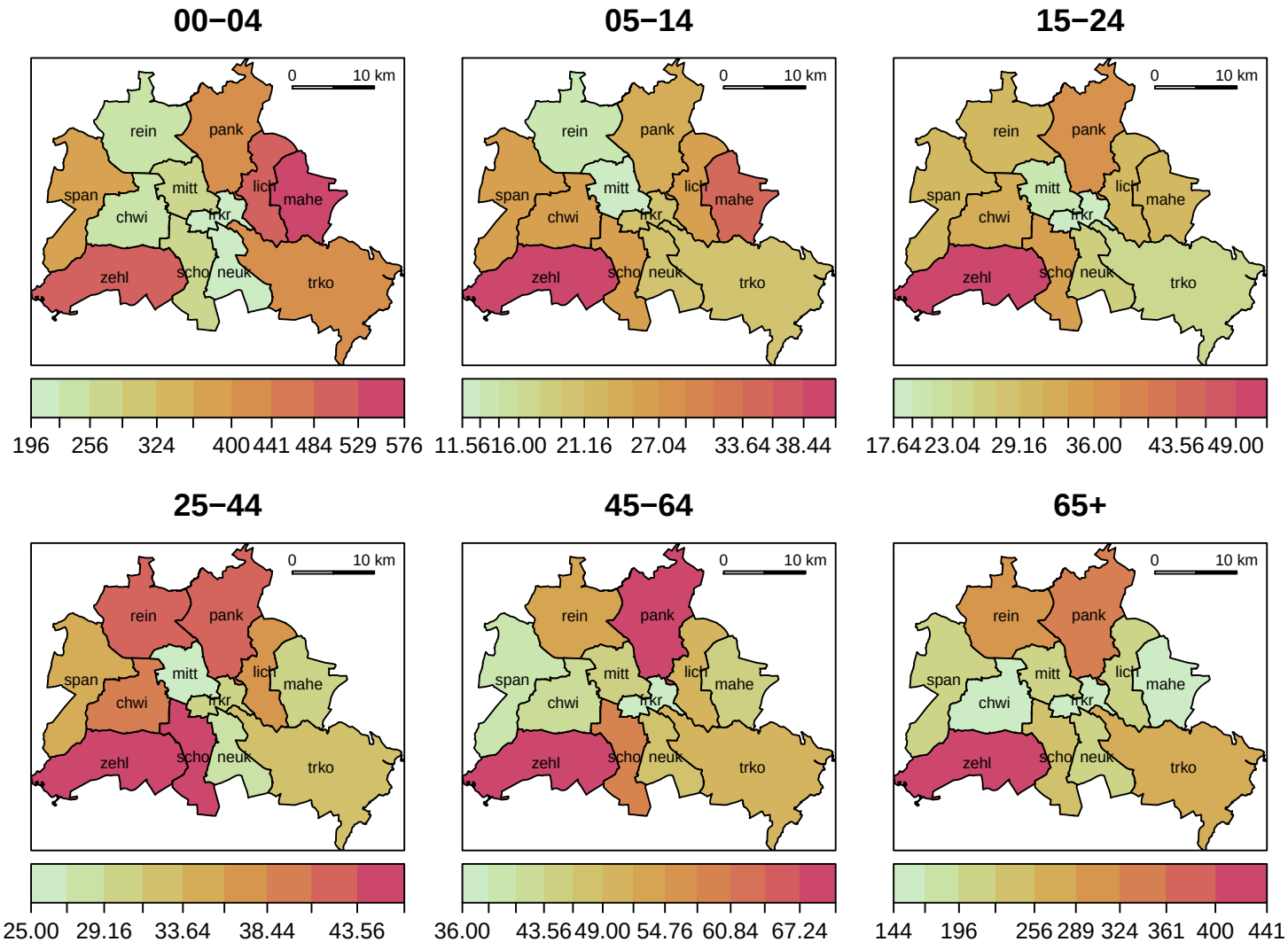
— Weekly time series by age group, aggregated over all 12 city districts



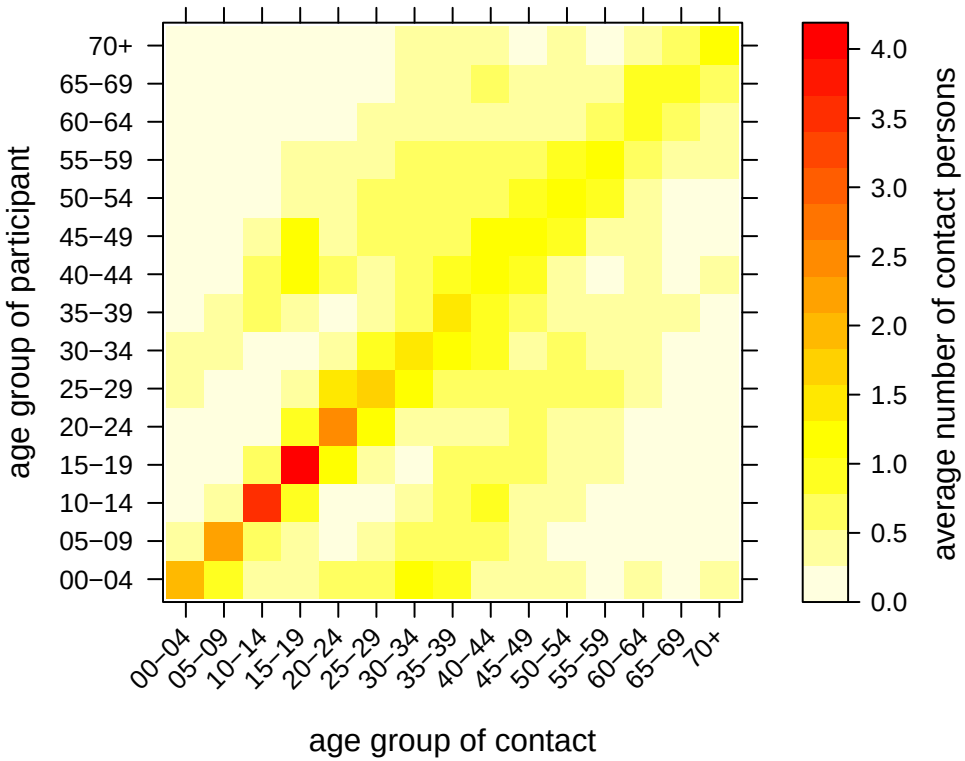
[Stratified lab-confirmed counts obtained from `survstat.rki.de`]

Case study: norovirus gastroenteritis in Berlin, 2011–2016

— Disease incidence maps by age group, aggregated over time



Infectious disease spread ~ social contacts



EU-funded POLYMOD study [Mossong et al. 2008]:

- 7 290 participants from eight European countries recorded contacts during one day
- Contact characteristics were similar across countries
- Remarkable mixing patterns with respect to age

Starting point for our statistical modelling framework

We have:

- Public health surveillance counts Y_{grt} indexed by *g*roup, *r*egion, *t*ime period
- Social contact matrix $C = (c_{g'g})$
- Maybe additional covariates (climate, socio-demographics, ...)

Starting point for our statistical modelling framework

We have:

- Public health surveillance counts Y_{grt} indexed by *g*roup, *r*egion, *t*ime period
- Social contact matrix $C = (c_{g'g})$
- Maybe additional covariates (climate, socio-demographics, ...)

We would like to model all three data dimensions (g, r, t) and account for their dependent nature:

g: social mixing patterns between age groups

r: spatial dynamics through human travel

t: temporal dependencies inherent to communicable diseases

An age-stratified, spatio-temporal, endemic-epidemic model

$$Y_{grt} = \text{NegBin}(\mu_{grt}, \psi_{gr})$$

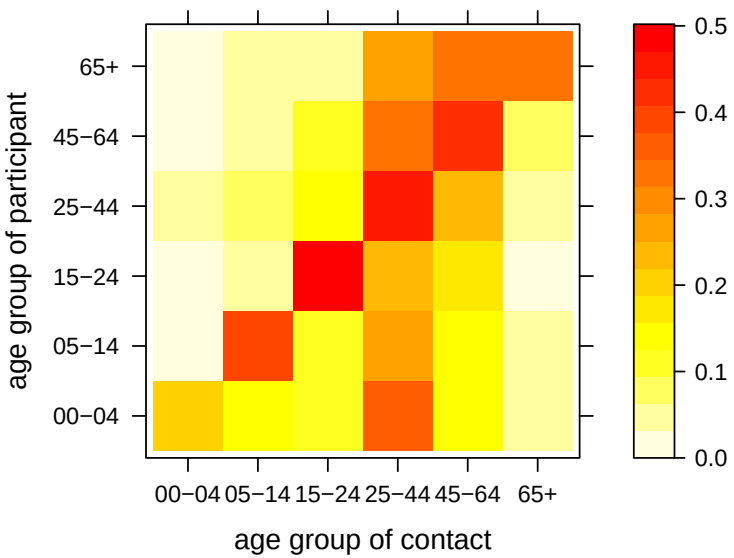
$$\mu_{grt} = \nu_{grt} + \phi_{grt} \sum_{g',r'} c_{g'g} \, w_{r'r} \, Y_{g',r',t-1}$$

An age-stratified, spatio-temporal, endemic-epidemic model

$$Y_{grt} = \text{NegBin}(\mu_{grt}, \psi_{gr})$$

$$\mu_{grt} = \nu_{grt} + \phi_{grt} \sum_{g',r'} \boxed{c_{g'g}} w_{r'r} Y_{g',r',t-1}$$

Contact matrix $(c_{g'g})$ for $g' \to g$, aggregated from POLYMOD



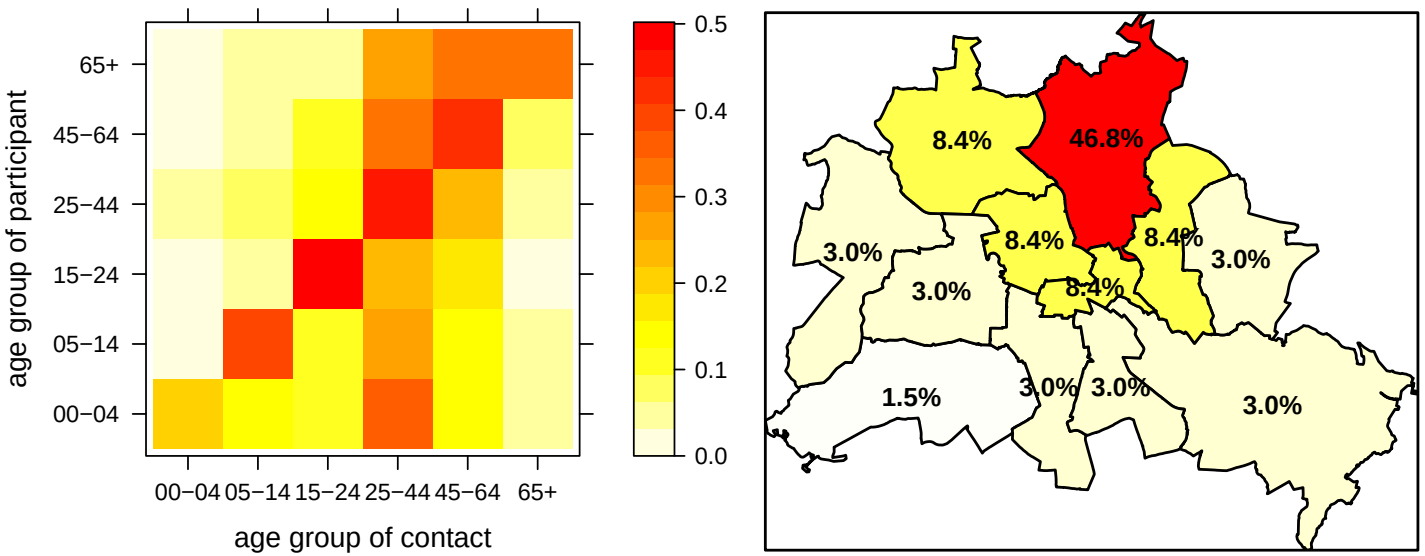
An age-stratified, spatio-temporal, endemic-epidemic model

$$Y_{grt} = \text{NegBin}(\mu_{grt}, \psi_{gr})$$

$$\mu_{grt} = \nu_{grt} + \phi_{grt} \sum_{g',r'} c_{g'g} \boxed{w_{r'r}} Y_{g',r',t-1}$$

Contact matrix $(c_{g'g})$ for $g' \to g$, aggregated from POLYMOD

Spatial weights for $r' \to r$, power-law decay $w_{r'r} = (o_{r'r} + 1)^{-\rho}$



An age-stratified, spatio-temporal, endemic-epidemic model

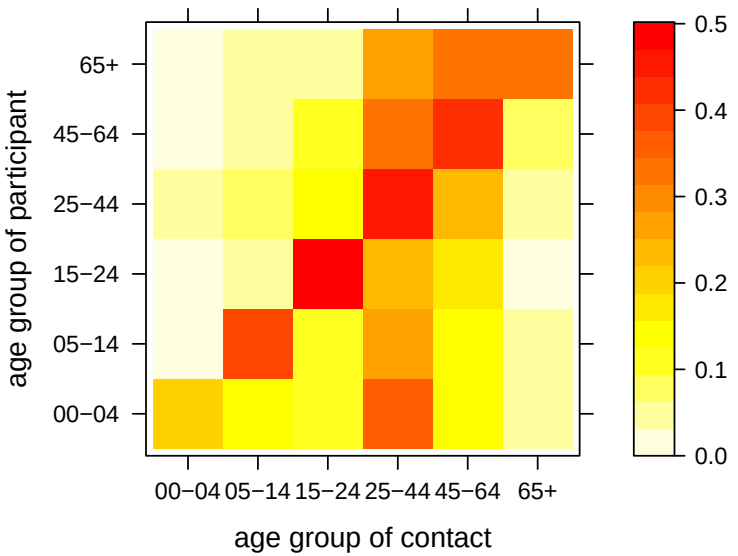
$$Y_{grt} = \text{NegBin}(\mu_{grt}, \psi_{gr})$$

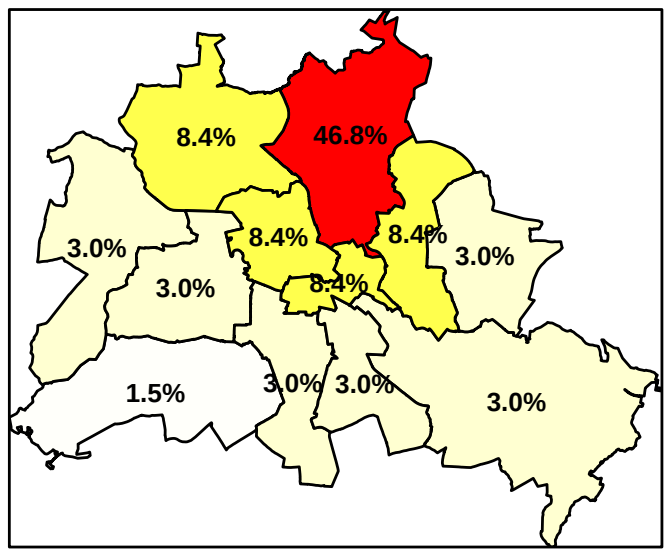
$$\mu_{grt} = \boxed{\nu_{grt}} + \boxed{\phi_{grt}} \sum_{g',r'} c_{g'g}\, w_{r'r}\, Y_{g',r',t-1}$$

Contact matrix $(c_{g'g})$ for $g' \to g$, aggregated from POLYMOD

Spatial weights for $r' \to r$, power-law decay $w_{r'r} = (o_{r'r} + 1)^{-\rho}$

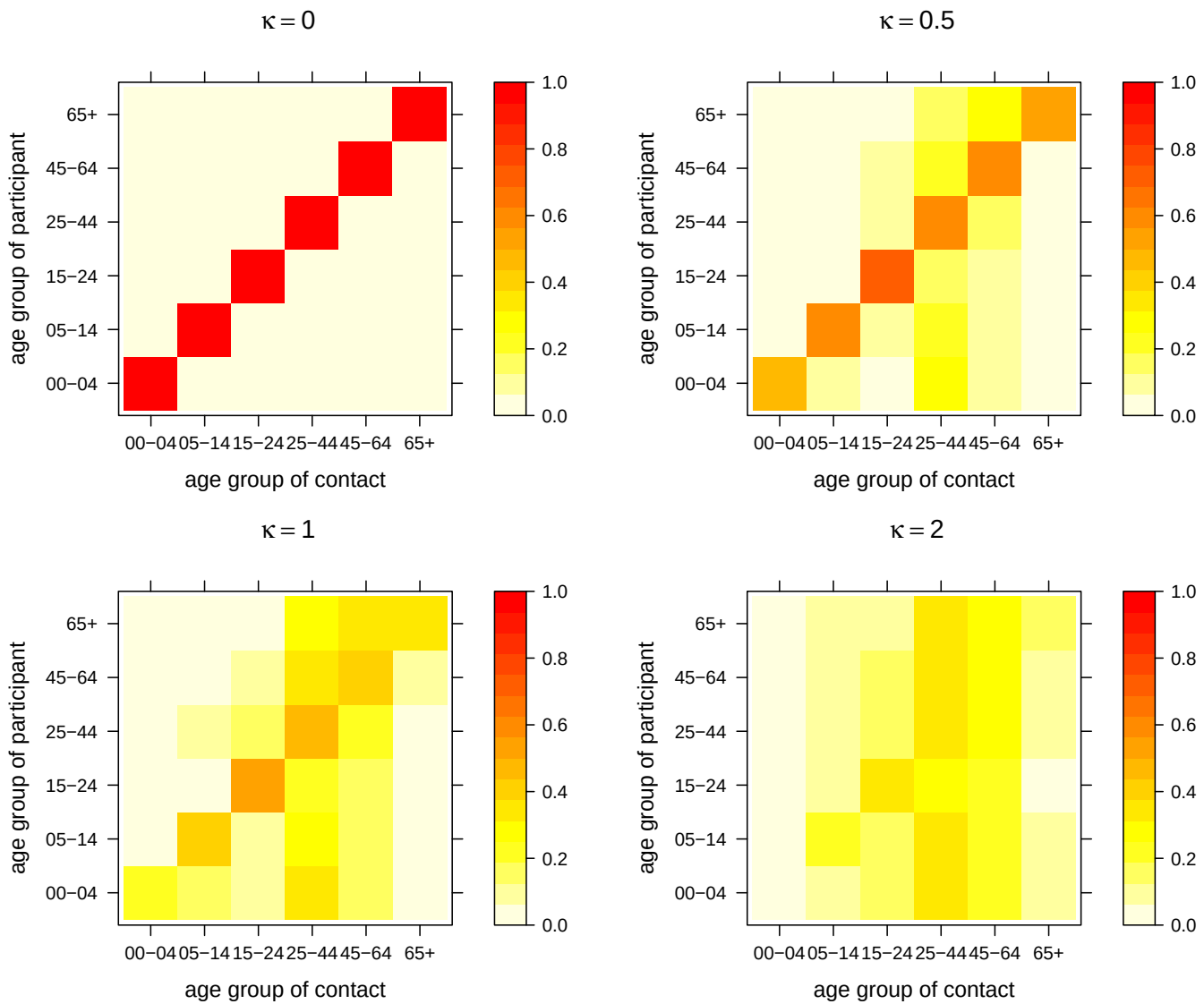
Log-linear predictors ν_{grt} and ϕ_{grt}



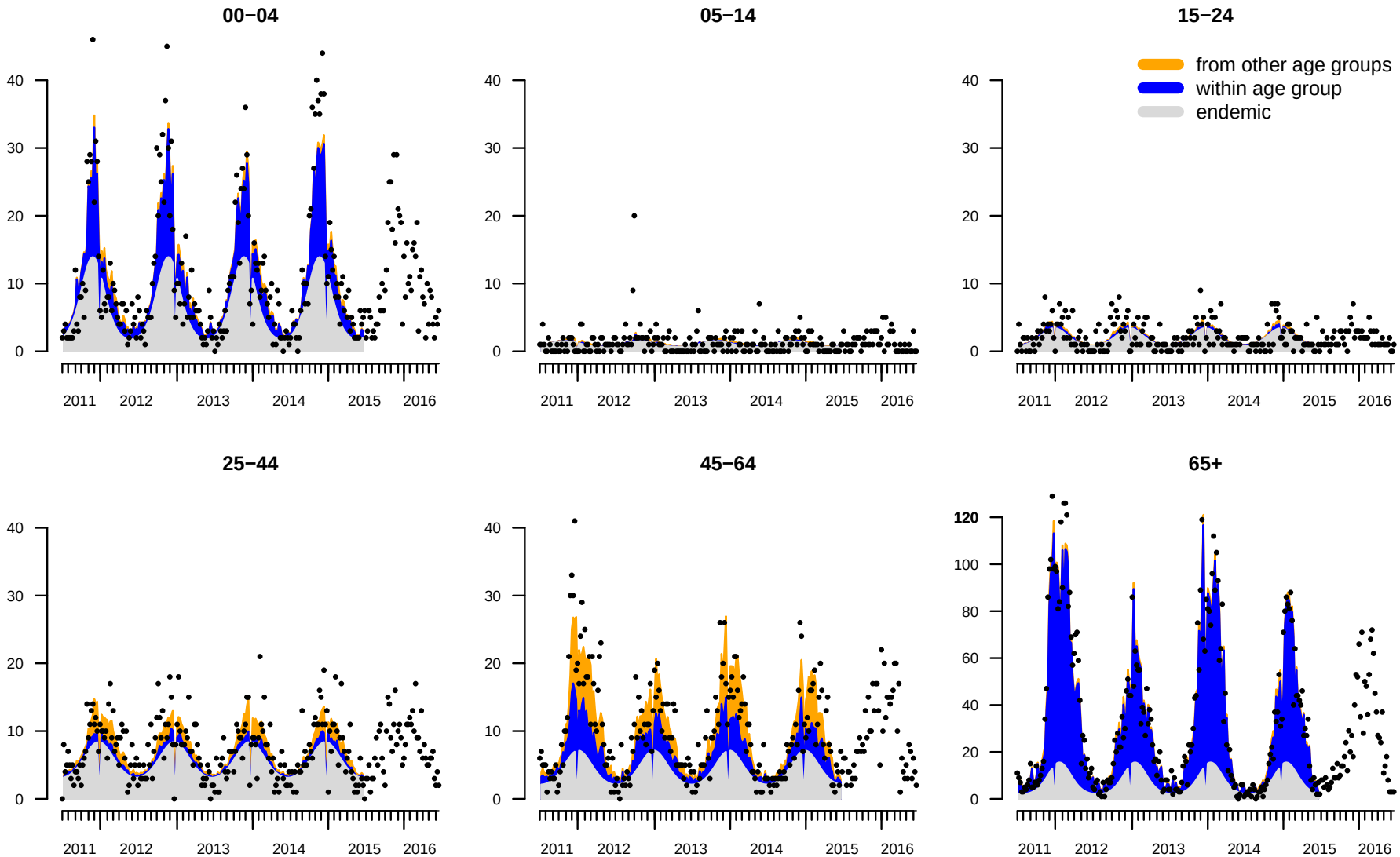


- Population
- Seasonality
- Group-specific susceptibility
- (Covariates)

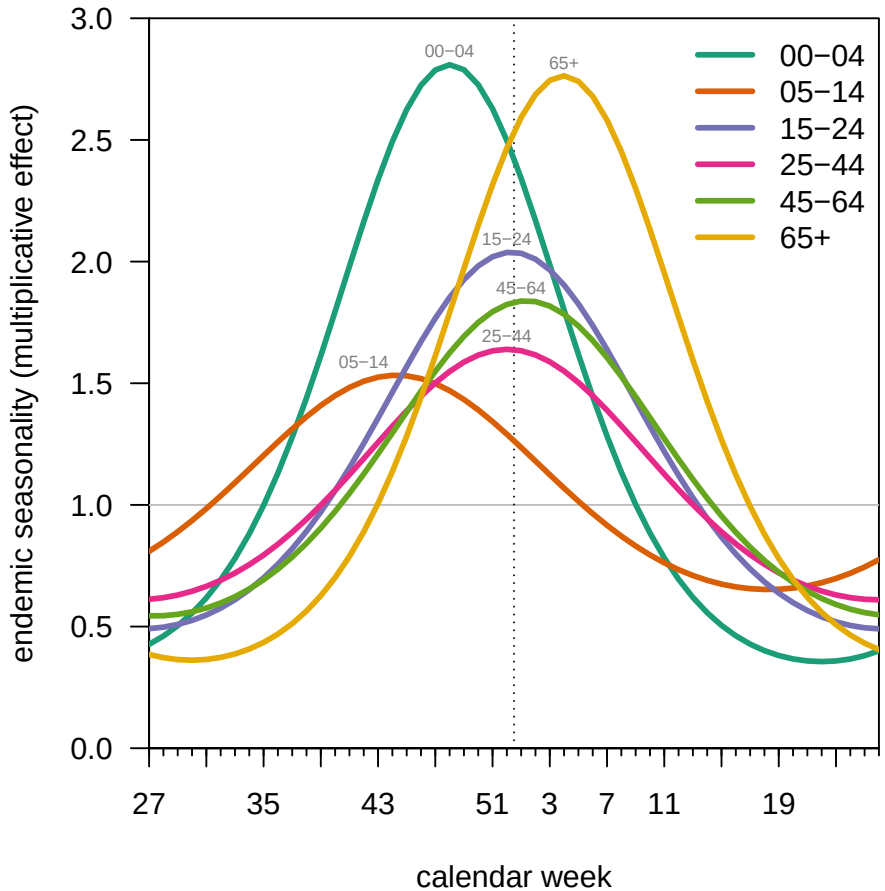
Power-adjustment of the contact matrix: $C^\kappa := E\Lambda^\kappa E^{-1}$

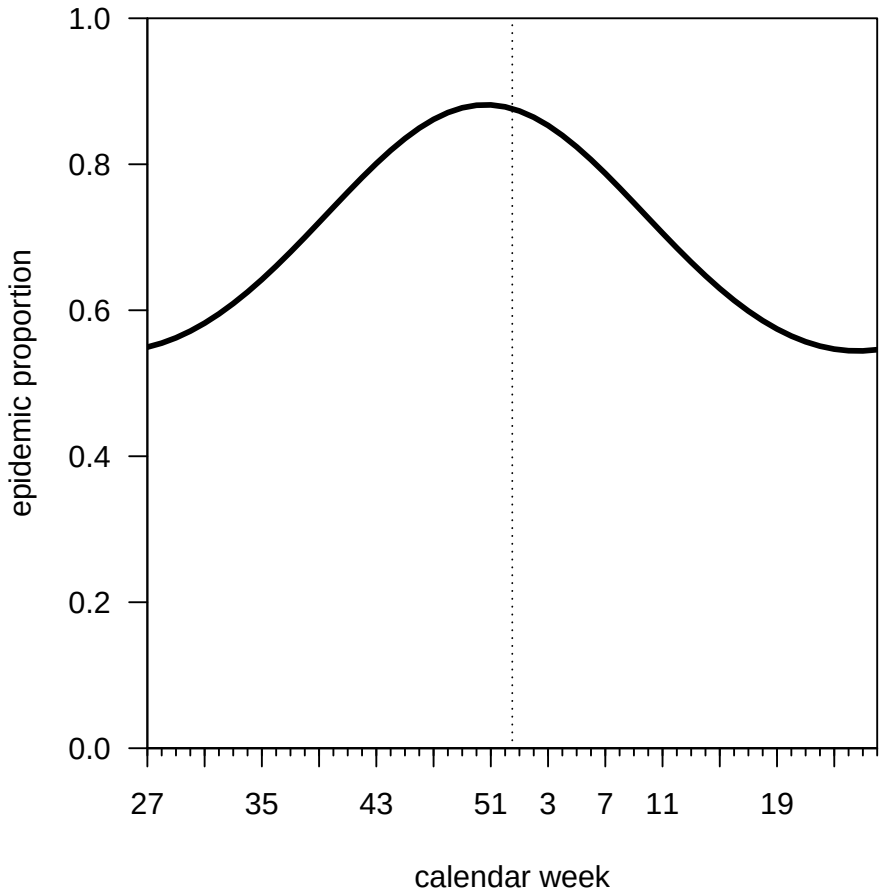


Fitted mean by age group aggregated over districts

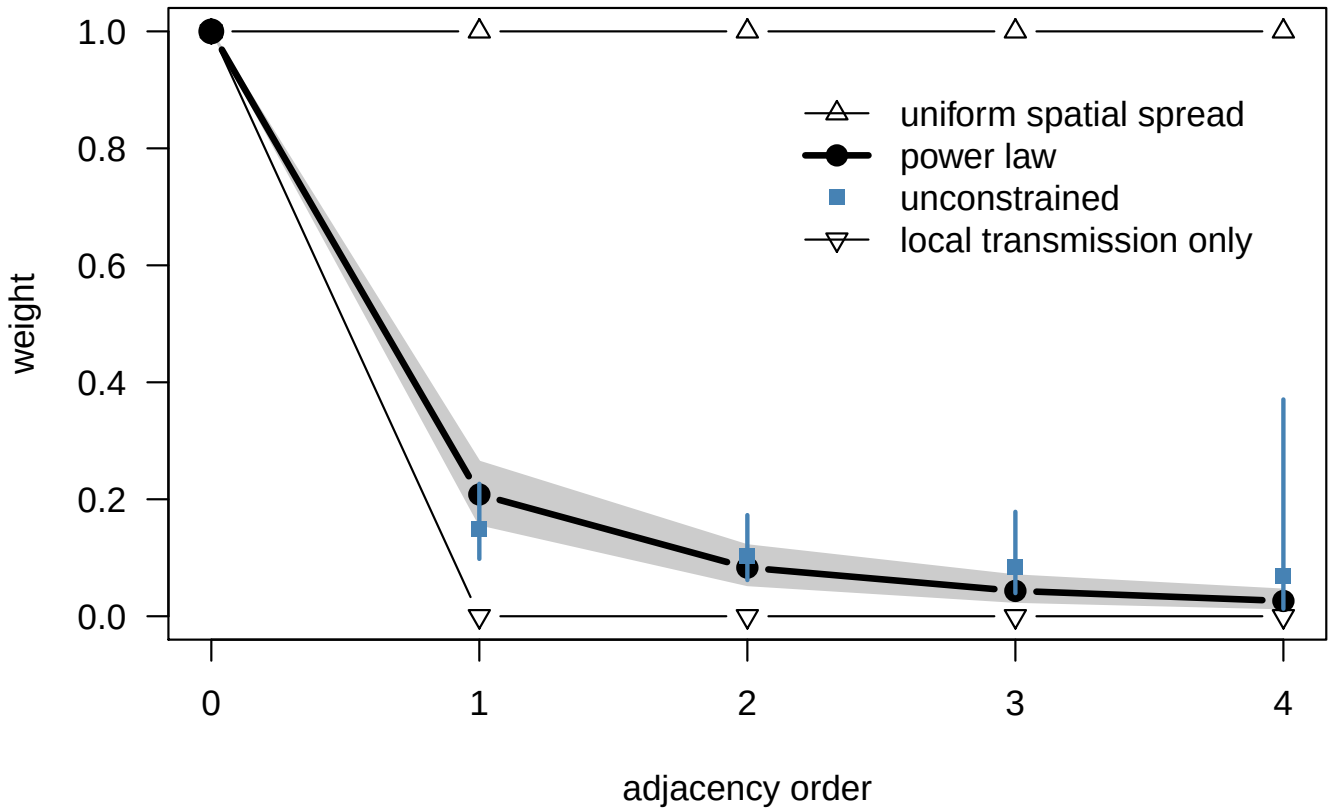


Estimated seasonality





Spatial power-law weights



Power transformation of the contact matrix

$\hat{\kappa} = 0.41$ (95% CI: 0.29 to 0.60)

Predictive model assessment

- AIC-based model comparison selects the most complex model (including the spatial power law and the adjusted contact matrix)
- Does this model also yield the best forecasts for the last season?
  - one-week-ahead: predictive distributions are negative binomial
  - long-term: via Monte Carlo simulation
- Various forecast targets exist, e.g., overall epidemic curve (weekly counts), final size (aggregated over the whole season)

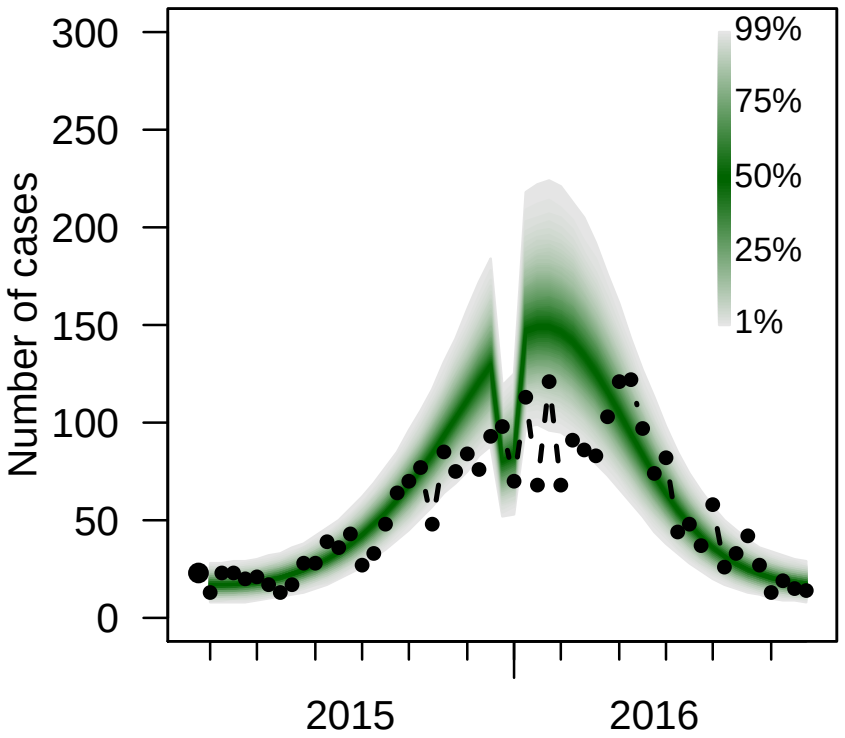
Predictive model assessment

- AIC-based model comparison selects the most complex model (including the spatial power law and the adjusted contact matrix)
- Does this model also yield the best forecasts for the last season?
  - one-week-ahead: predictive distributions are negative binomial
  - long-term: via Monte Carlo simulation
- Various forecast targets exist, e.g., overall epidemic curve (weekly counts), final size (aggregated over the whole season)
- "Best" is quantified in terms of sharpness and calibration of the forecasts
- Proper scoring rules serve as overall performance measures [Gneiting and Katzfuss 2014]
  - Assign penalty score based on the predictive distribution F and the actual observation y_{obs}
  - Example: Dawid-Sebastiani score

$$\mathrm{DSS}(F, y_{obs}) = \log|\boldsymbol{\Sigma}| + (y_{obs} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (y_{obs} - \boldsymbol{\mu})$$

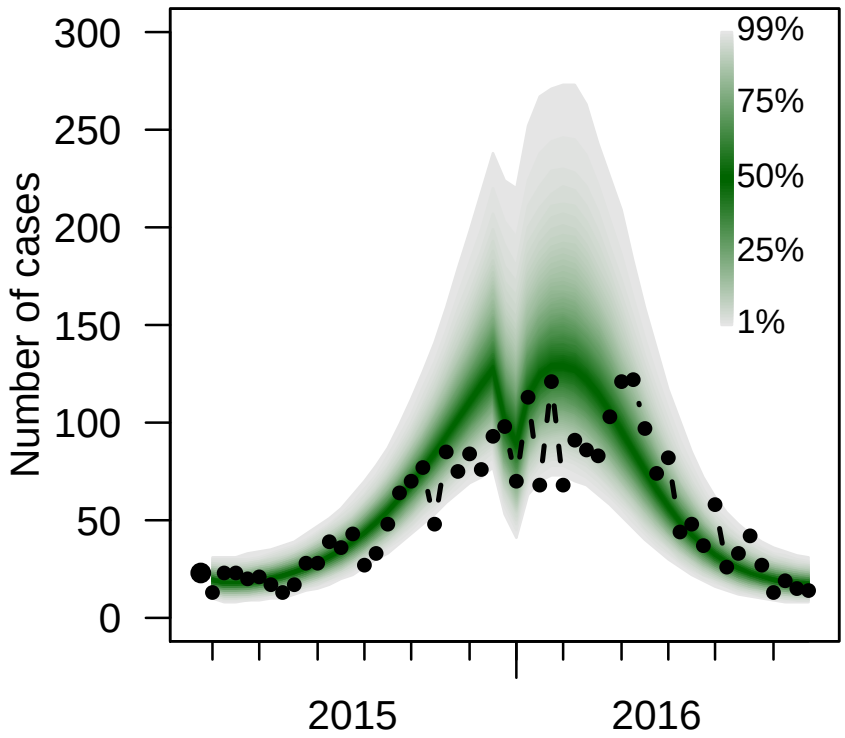
Target quantity: overall epidemic curve

purely endemic model



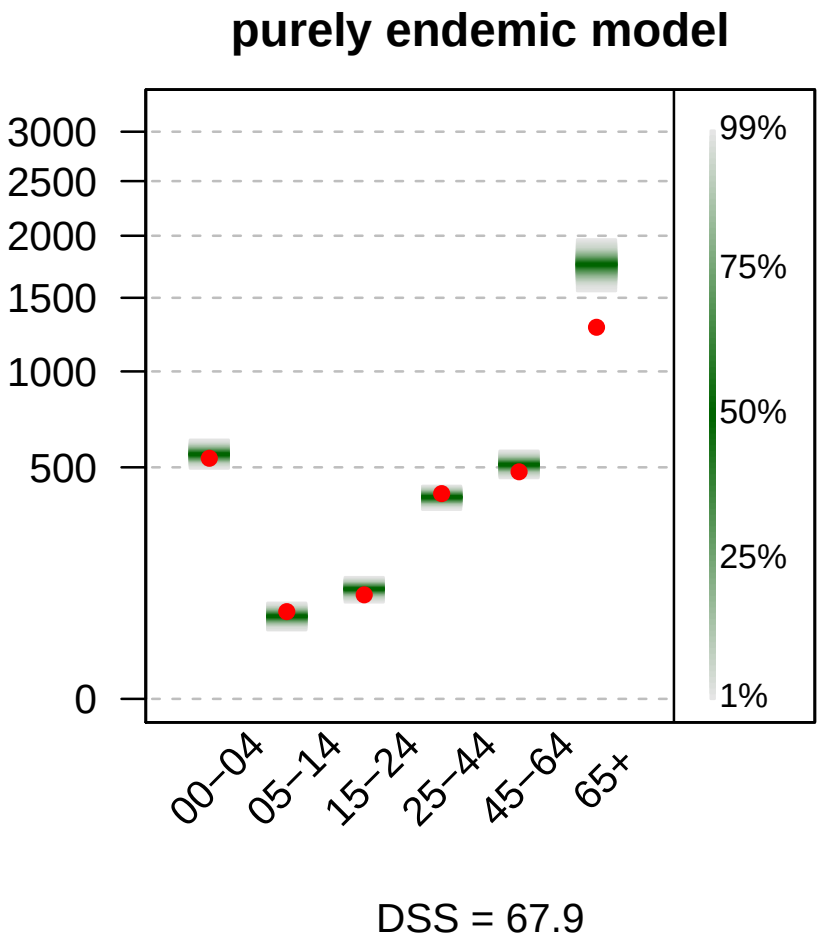
DSS = 346.5

full model

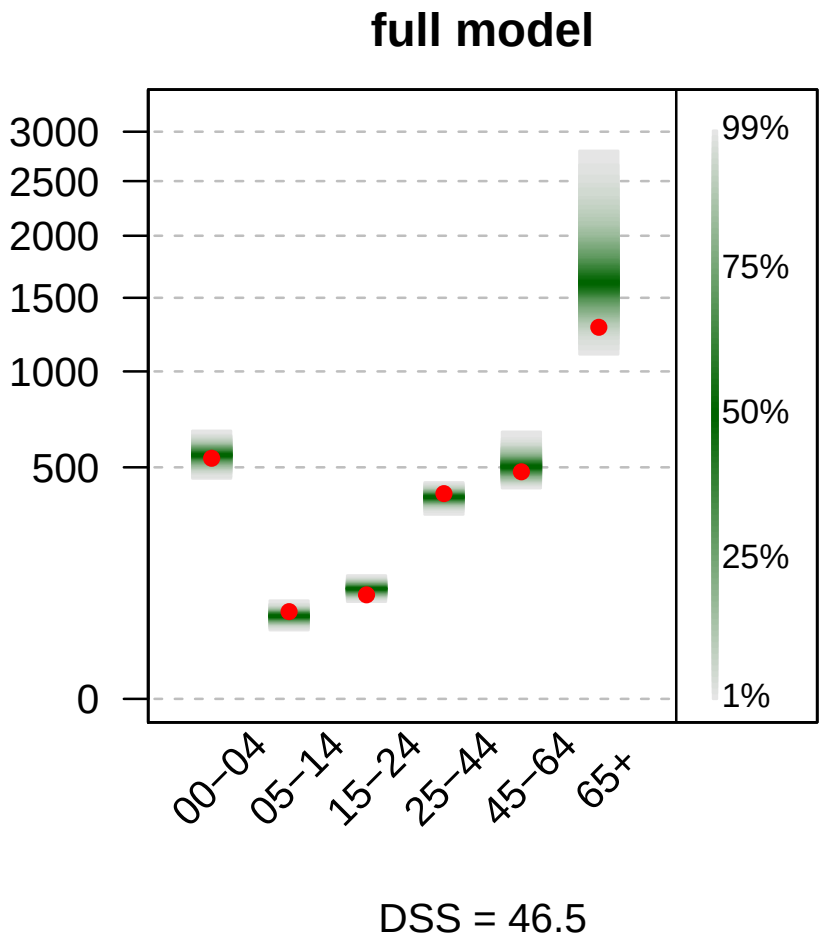


DSS = 334.0

Target quantity: final size by age group



DSS = 67.9



DSS = 46.5

Conclusion

- Models do not perfectly represent individual-level disease transmission, but are still useful for prediction of aggregate-level surveillance counts
- Endemic-epidemic modelling frameworks are implemented in `surveillance`, also for individual-level data on disease occurrence [Meyer, Held, and Höhle 2017]
- Spatial weights and social contact data improve model fit *and* predictions
- The data and code to reproduce some of the presented results is available in the supplementary package `hhh4contacts` [Meyer and Held 2017]
- If the modelling goal is forecasting, use proper scoring rules to assess the quality of probabilistic forecasts [Held, Meyer, and Bracher 2017]

References

Gneiting, Tilmann and Katzfuss, Matthias (2014). "Probabilistic forecasting". In: *Annual Review of Statistics and Its Application* 1.1, pp. 125–151. DOI: 10.1146/annurev-statistics-062713-085831.

Held, Leonhard, Meyer, Sebastian, and Bracher, Johannes (2017). "Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture". In: *Statistics in Medicine* 36.22, pp. 3443–3460. DOI: 10.1002/sim.7363.

Meyer, Sebastian and Held, Leonhard (2017). "Incorporating social contact data in spatio-temporal models for infectious disease spread". In: *Biostatistics* 18.2, pp. 338–351. DOI: 10.1093/biostatistics/kxw051.

Meyer, Sebastian, Held, Leonhard, and Höhle, Michael (2017). "Spatio-temporal analysis of epidemic phenomena using the R package **surveillance**". In: *Journal of Statistical Software* 77.11, pp. 1–55. DOI: 10.18637/jss.v077.i11.

Mossong, Joël et al. (2008). "Social contacts and mixing patterns relevant to the spread of infectious diseases". In: *PLoS Medicine* 5.3, e74. DOI: 10.1371/journal.pmed.0050074.

World Health Organization (2014). "Anticipating epidemics". In: *Weekly Epidemiological Record* 89.22, p. 244. URL: http://www.who.int/wer.

Questions? Comments? ☞ seb.meyer@fau.de

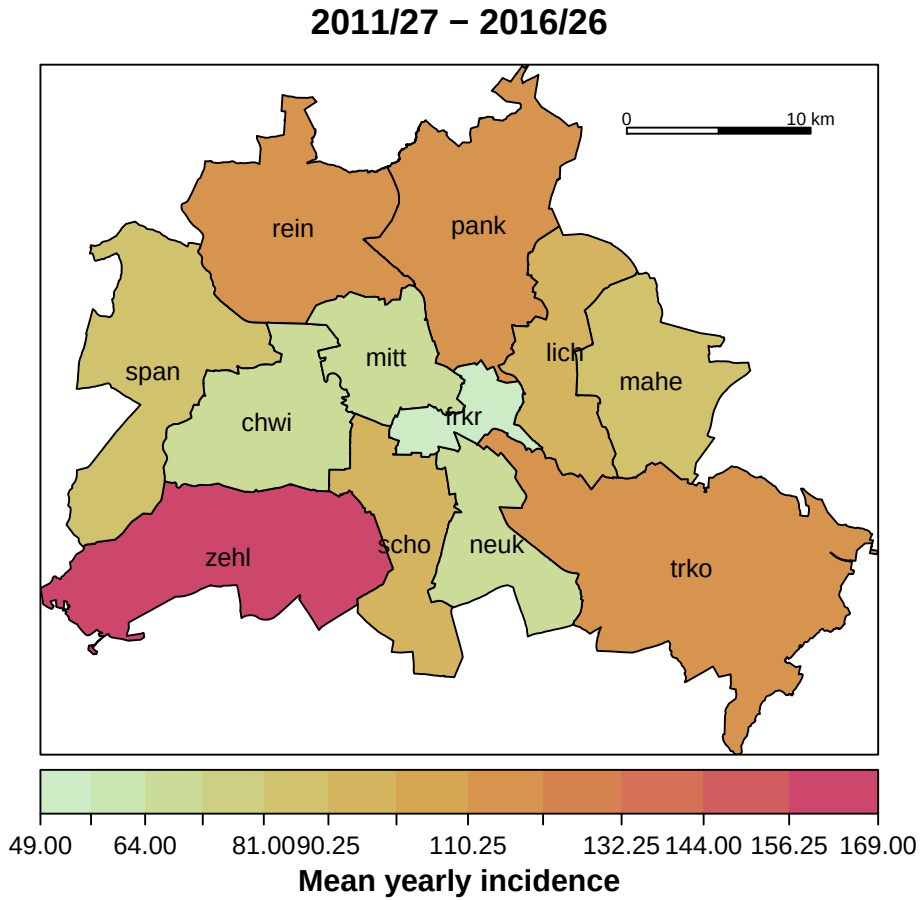
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACULTY OF MEDICINE

Appendix

Disease incidence map

```
noroBEr <- noroBE(by = "districts",
    timeRange=c("2011-w27","2016-w26"))
scalebar <- layout.scalebar(noroBEr@map,
    corner = c(0.7, 0.9), scale = 10,
    labels = c(0, "10 km"), cex = 0.6,
    height = 0.02)
plot(noroBEr, type = observed ~ unit,
     sub = "Mean yearly incidence",
     population = rowSums(pop2011) *
       (nrow(noroBEr)/52)/100000,
     labels = list(cex = 0.8),
     sp.layout = scalebar)
```



Example code for a “simple” version of the model

```
library("surveillance")  # basic "hhh4" modelling framework
library("hhh4contacts")  # norovirus and contact data

noroBEall <- noroBE(by = "all", flatten = TRUE,  # 6 x 12 = 72 columns
                    timeRange = c("2011-w27", "2016-w26"))

fit <- hhh4(stsObj = noroBEall, control = list(
    end = list(f = addSeason2formula(~1),
               offset = prop.table(population(noroBEall), 1)),
    ne = list(f = ~1 + log(pop),
              weights = W_powerlaw(maxlag = 5, log = TRUE),
              scale = expandC(contactmatrix(), 12)),
    data = list(pop = prop.table(population(noroBEall), 1)),
    family = "NegBin1", subset = 2:(4*52)))
```

Specific model formulation for the norovirus data

$$\mu_{grt} = e_{gr} \exp\left\{\alpha_g^{(\nu)} + \alpha_r^{(\nu)} + \beta x_t + \gamma_g^{(\nu)} \sin(\omega t) + \delta_g^{(\nu)} \cos(\omega t)\right\}$$
$$+ \exp\left\{\alpha_g^{(\phi)} + \alpha_r^{(\phi)} + \tau \log(e_{gr}) + \gamma^{(\phi)} \sin(\omega t) + \delta^{(\phi)} \cos(\omega t)\right\}$$
$$\sum_{g',r'} \lfloor (C^\kappa)_{g'g} (o_{r'r} + 1)^{-\rho} \rfloor \, Y_{g',r',t-1}$$

- Group- and district-specific effects $\alpha_g^{(\cdot)}$ and $\alpha_r^{(\cdot)}$
- Christmas break indicator $x_t \longrightarrow$ reduced reporting
- Group-specific endemic seasonality (sinusoidal log-rates, $\omega = 2\pi/52$)
- "Gravity model" $e_{gr}^\tau \longrightarrow$ force of infection scales with population size
- C^κ: power-adjusted contact matrix
- Power-law weights $w_{r'r} = (o_{r'r} + 1)^{-\rho}$

+ group-specific overdispersion parameters ψ_g