Multivariate Count Time Series Modeling of Surveillance Data

Leonhard Held[1] Michael Höhle[2]

[1]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland

[2]Department of Mathematics, Stockholm University, Sweden

ISDS Webinar
19 September 2016

University of Zurich[UZH]

Stockholm University

Outline

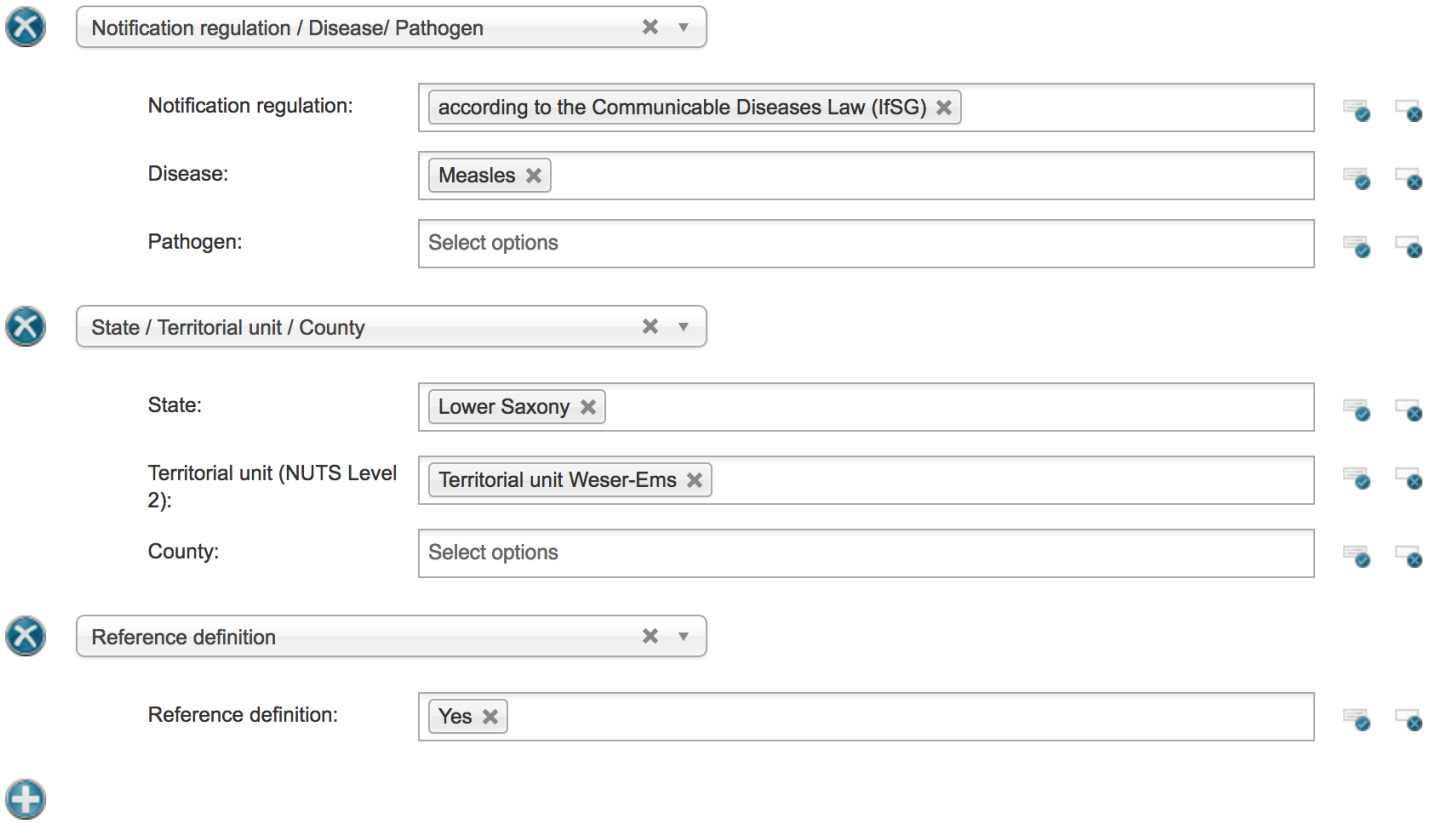
Outline

Infectious disease registry data

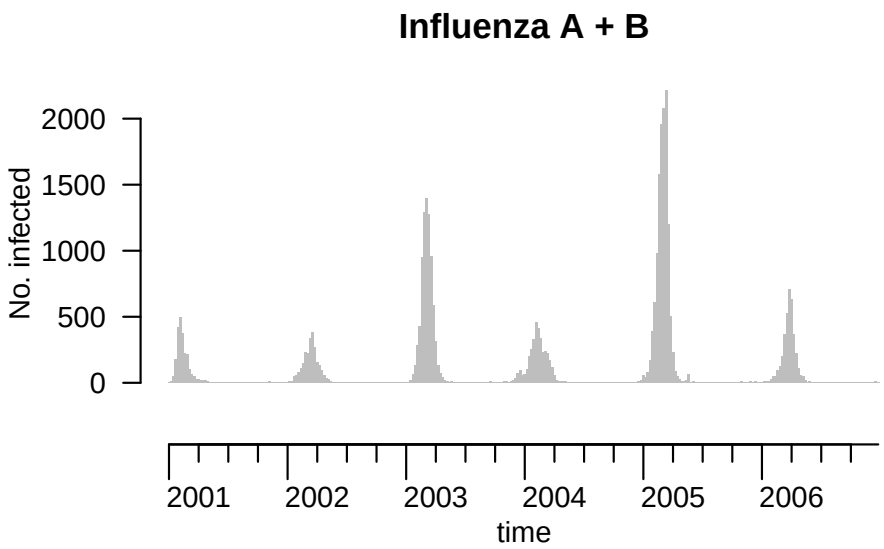
- Many countries have established surveillance systems for the routine collection of infectious disease data.
- Such surveillance data consist of individual-level, time-stamped and geo-referenced case reports of notifiable diseases.
- Publically available registry data are usually aggregated into time series of counts of new infections of a specific disease, observed in different areas or age groups.

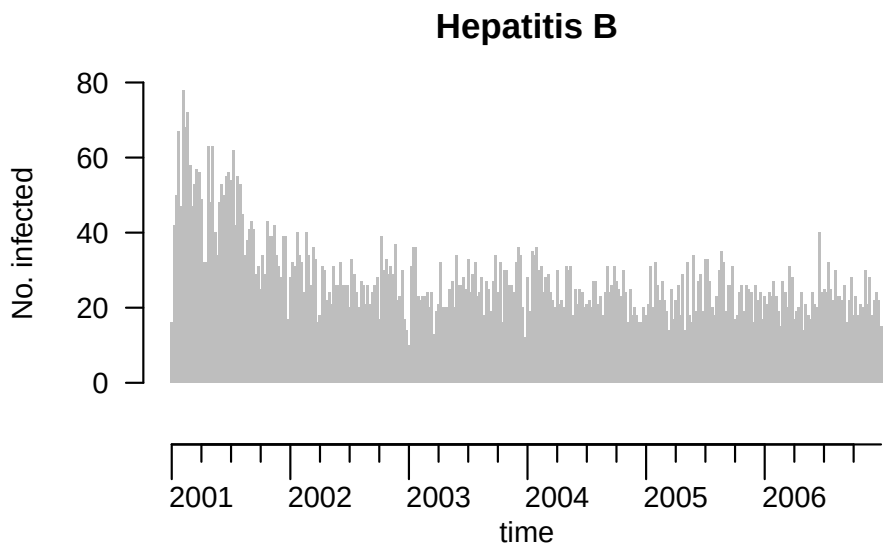
Example: SurvStat@RKI 2.0 in Germany

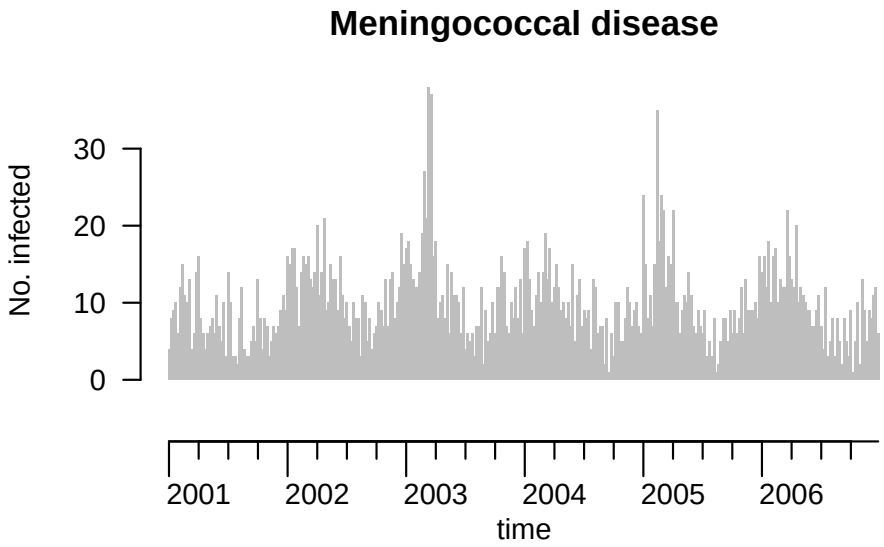
Custom online queries of aggregated data under the 'Protection against Infection Act'

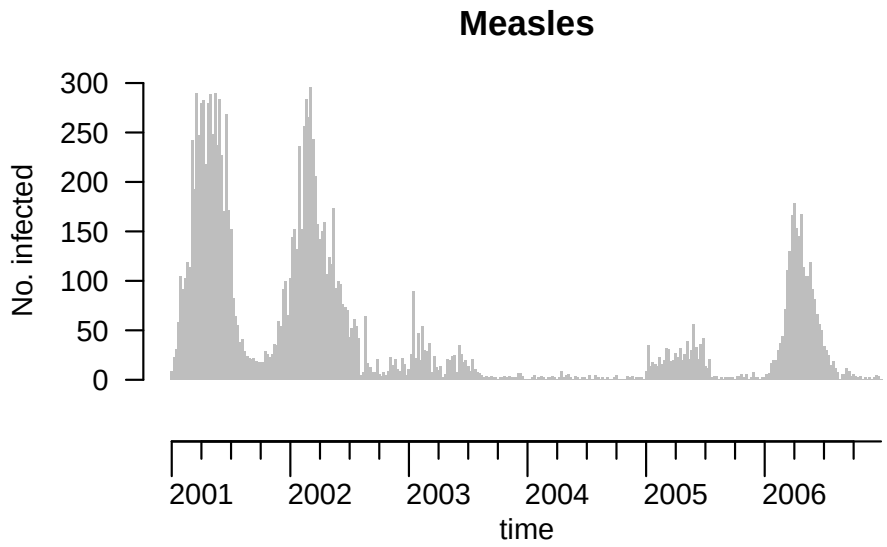


Weekly counts of different infectious diseases





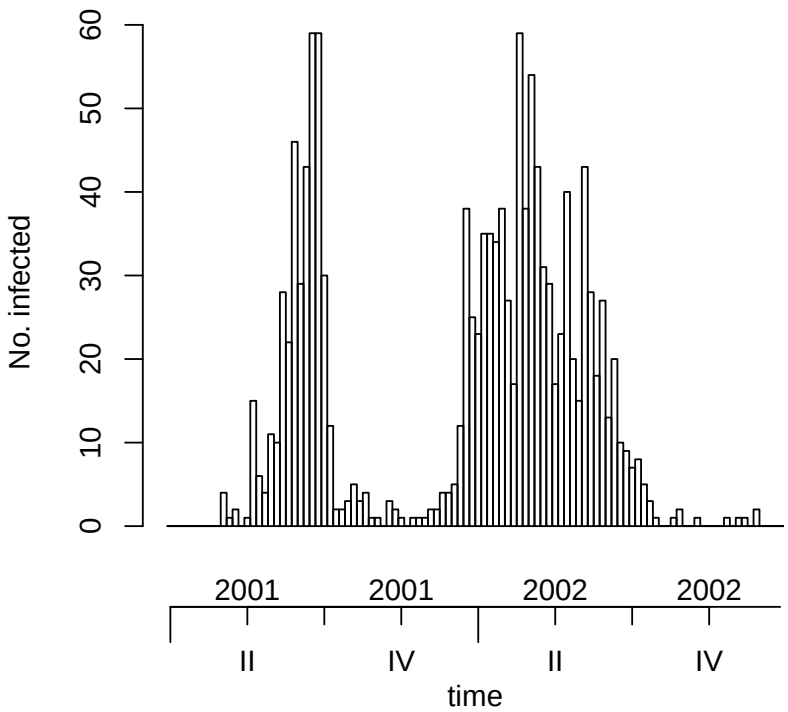




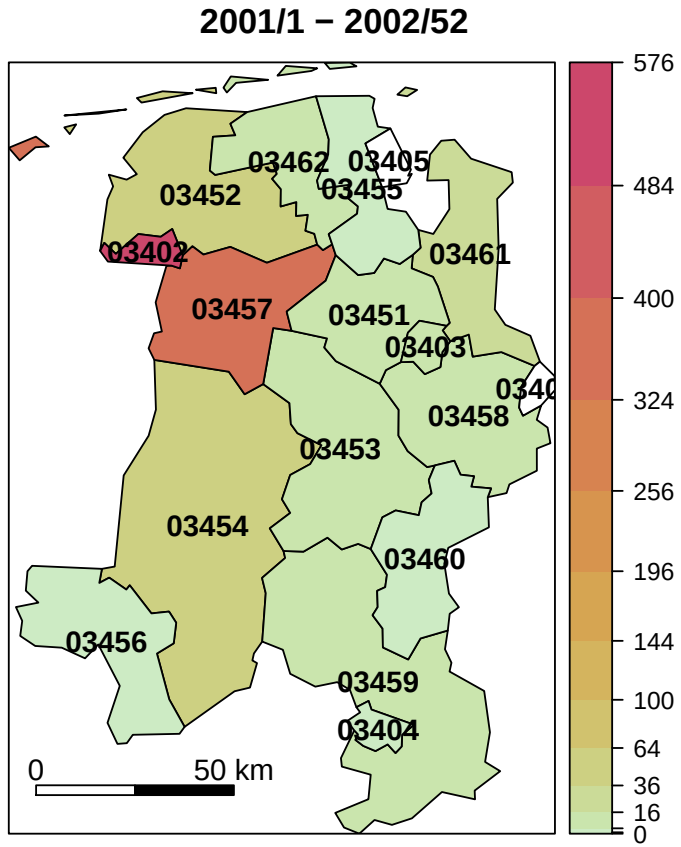
Weekly counts of measles infections

Weser-Ems region of Lower Saxony, Germany, 2001–2002

Time series of weekly counts

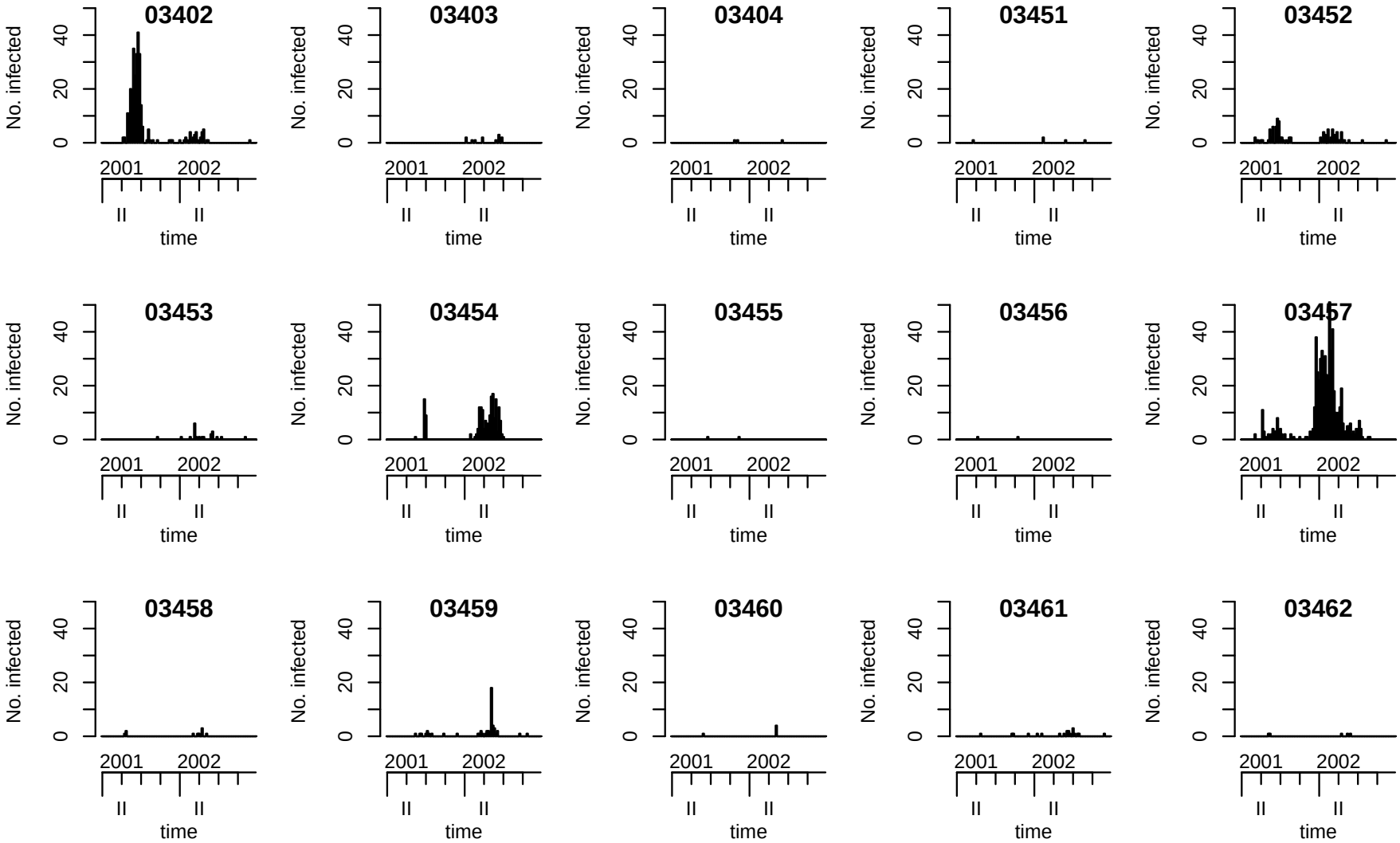


Disease incidence (per 100 000 inhabitants)



Weekly counts of measles infections

Count time series of the 15 affected districts



Characteristics of surveillance data

- Low number of cases
- Seasonality
- Occassional outbreaks
- Dependence between areas, age groups, etc.
- Underreporting, reporting delays
- No information about the number of susceptibles

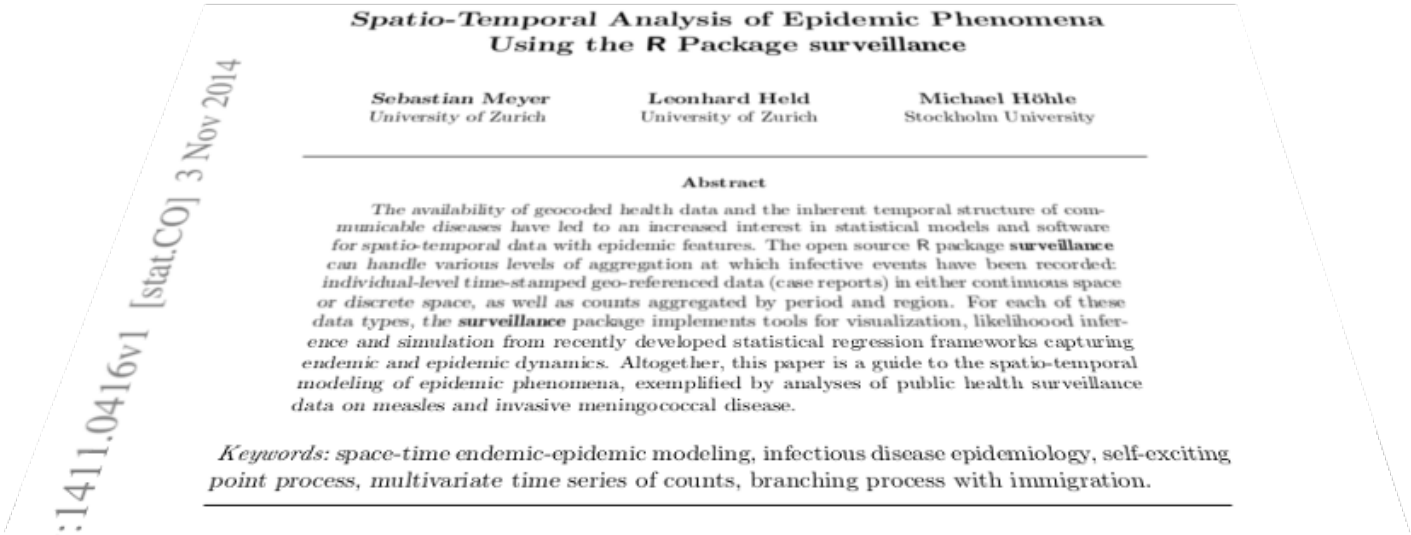
A model-based approach

We use statistical models for multivariate count time series

1. to answer relevant research questions
2. to predict future disease incidence

Aims of the Talk

- We follow the presentation in Meyer et al. (2016, Section 5) and analyse the Weser-Ems measles dataset with models of successively increasing complexity.



Spatio-Temporal Analysis of Epidemic Phenomena
Using the R Package surveillance

Sebastian Meyer *University of Zurich* | **Leonhard Held** University of Zurich | **Michael Höhle** Stockholm University

Abstract

The availability of geocoded health data and the inherent temporal structure of communicable diseases have led to an increased interest in statistical models and software for spatio-temporal data with epidemic features. The open source R package **surveillance** *can handle various levels of aggregation at which infective events have been recorded: individual-level time-stamped geo-referenced data (case reports) in either continuous space or discrete space, as well as counts aggregated by period and region. For each of these data types, the* **surveillance** *package implements tools for visualization, likelihoood inference and simulation from recently developed statistical regression frameworks capturing endemic and epidemic dynamics. Altogether, this paper is a guide to the spatio-temporal modeling of epidemic phenomena, exemplified by analyses of public health surveillance data on measles and invasive meningococcal disease.*

Keywords: space-time endemic-epidemic modeling, infectious disease epidemiology, self-exciting point process, multivariate time series of counts, branching process with immigration.

- The corresponding R code is available as part of the vignette
  "`hhh4`: Endemic-epidemic modeling of areal count time series"
  in the R package **surveillance** available from CRAN.
- Alternatively, the code can be viewed directly as file `hhh4_spacetime.Rnw` from the package subversion repository.

Outline

Univariate time series models

- A statistical framework for surveillance counts Y_t (Held et al., 2005):

$$Y_t \mid Y_{t-1} \sim \text{Po}(\mu_t) \quad \text{with} \quad \mu_t = \nu_t + \lambda Y_{t-1}$$

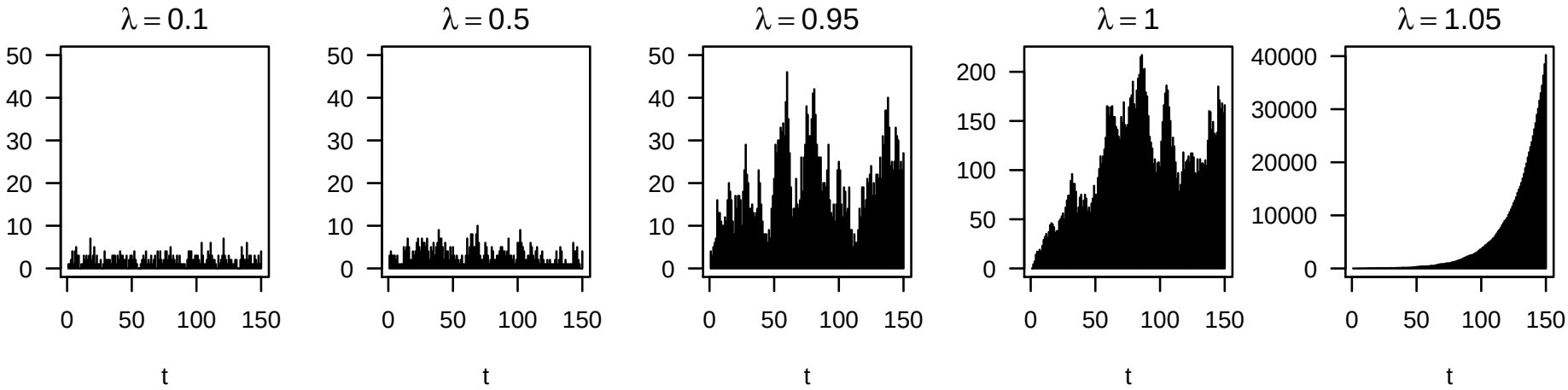
  where Y_t is the number of cases at time $t = 1, 2, \ldots$

The disease incidence is additively decomposed into

- endemic component ν_t
  which may parametrically model regular trends and seasonality
  similar to log-linear Poisson regression
- epidemic (or autoregressive) component λY_{t-1}
  which captures disease spread (if time resolution $\approx$ serial time).
- The autoregressive parameter λ may also depend on time in order to capture seasonally varying severity of pathogens (Held and Paul, 2012).

Simulations with different values of λ

$$Y_t \mid Y_{t-1} \sim \text{Po}(\mu_t) \quad \text{with} \quad \mu_t = 2 + \lambda Y_{t-1} \quad \text{and} \quad Y_0 = 1$$



- Autoregressive coefficient λ can be interpreted as epidemic proportion.
- $\lambda < 1$ ensures stationarity with mean incidence $\nu/(1-\lambda)$.
- In applications, the Poisson response distribution is often replaced by the negative binomial to adjust for overdispersion.

Weekly counts of measles infections

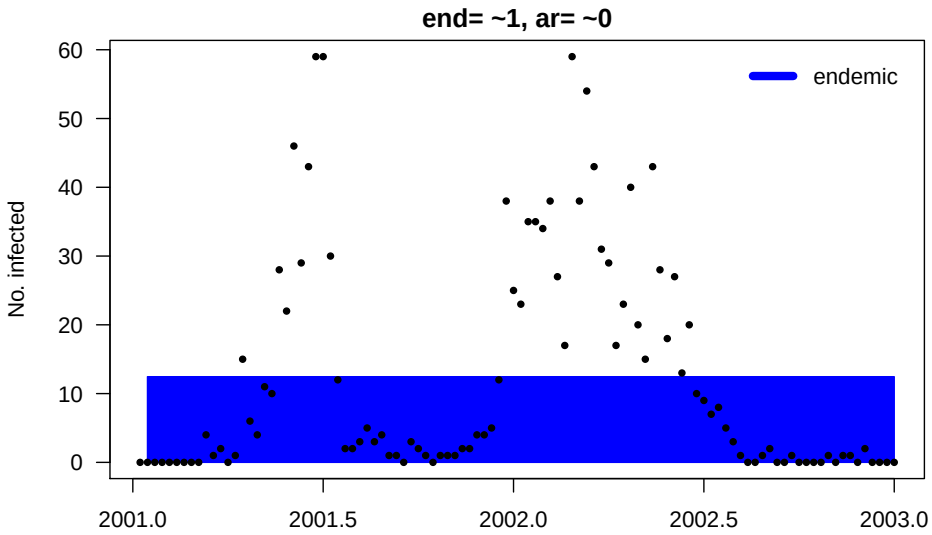
Fitting a univariate model

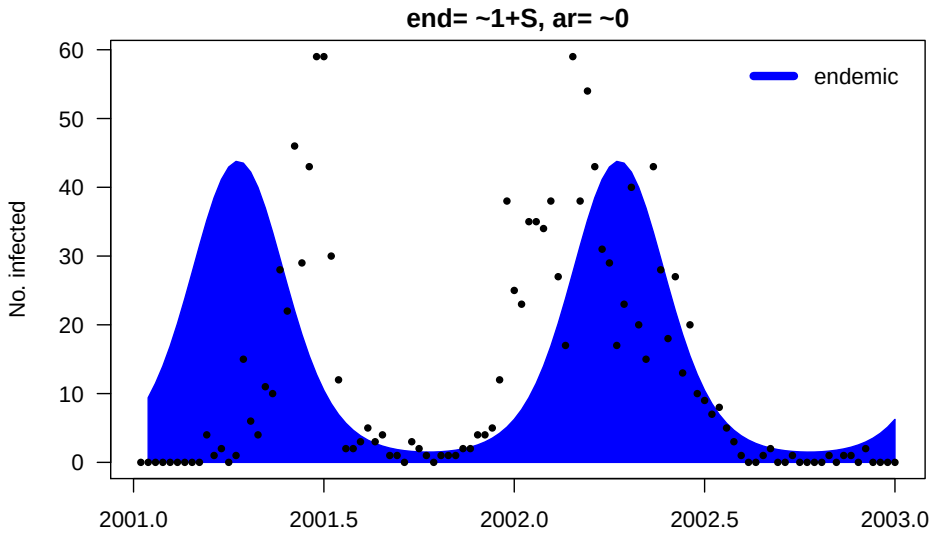
Serial time of measles $\approx$ 10 days $\rightarrow$ use weekly (or biweekly) counts.

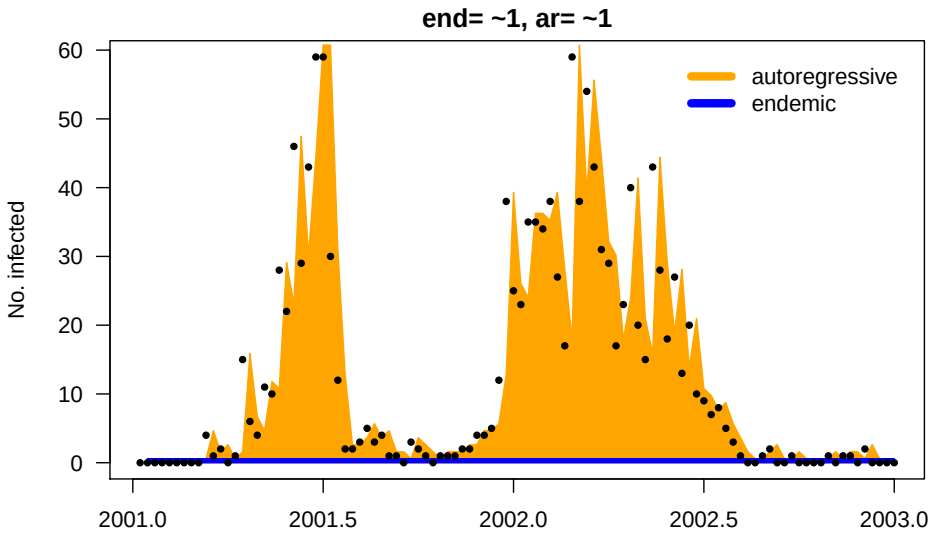
```
R> measlesModel_uni_0 <- list(end = list(f = ~1), ar = list(f = ~1), family = "NegBin1")
R> measlesFit_uni_0 <- hhh4(stsObj = counts.total, control = measlesModel_uni_0)
R> summary(measlesFit_uni_0, idx2Exp = TRUE, maxEV = TRUE)

Call:
hhh4(stsObj = counts.total, control = measlesModel_uni_0)

Coefficients:
            Estimate  Std. Error
exp(ar.1)   1.01946   0.09498
exp(end.1)  0.58555   0.15533
overdisp    0.39588   0.09097

Epidemic dominant eigenvalue:  1.02

Log-likelihood:   -276.35
AIC:              558.69
BIC:              566.6

Number of units:        1
Number of time points:  103
```

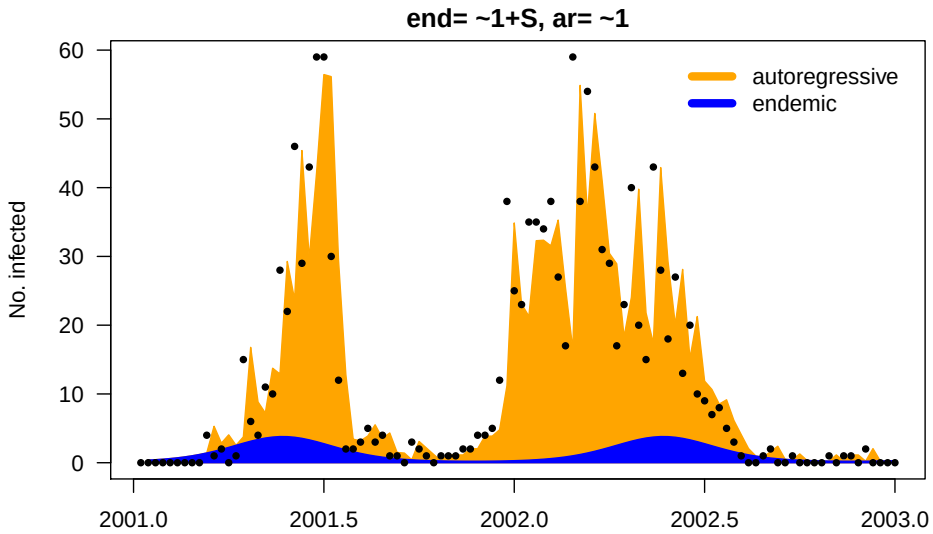
Weekly counts of measles infections – Comparison of Fit

Seasonality in endemic vs. epidemic component



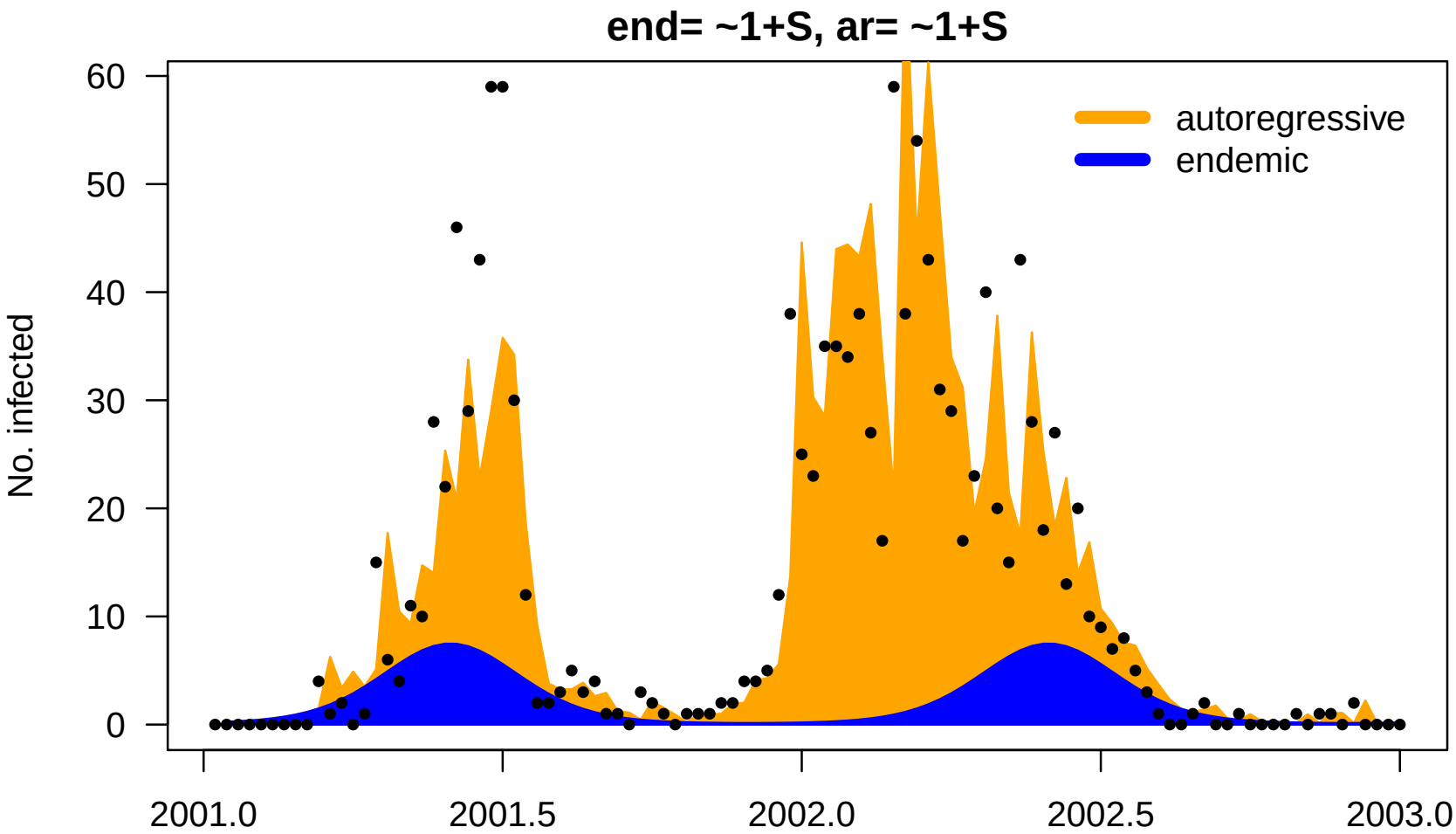






Weekly counts of measles infections – Comparison of Fit

Seasonality in endemic vs. epidemic component



Outline

Multivariate modelling

Suppose now multiple time series are available (here by district):

μ_{it}: mean number of cases in district i at time t

$$\mu_{it} = \nu_{it} + \lambda_i Y_{i,t-1}$$

Multivariate modelling

Suppose now multiple time series are available (here by district):

μ_{it}: mean number of cases in district i at time t

$$\mu_{it} = \nu_{it} + \lambda_i Y_{i,t-1}$$

- $\log(\nu_{it}) = \alpha_i +$ population offset $+$ seasonal trend $+$ covariates

Multivariate modelling

Suppose now multiple time series are available (here by district):

μ_{it}: mean number of cases in district i at time t

$$\mu_{it} = \nu_{it} + \lambda_i Y_{i,t-1}$$

- $\log(\nu_{it}) = \alpha_i +$ population offset $+$ seasonal trend $+$ covariates
- $\log(\lambda_i) = \beta_i +$ covariates

Multivariate modelling

Suppose now multiple time series are available (here by district):

μ_{it}: mean number of cases in district i at time t

$$\mu_{it} = \nu_{it} + \lambda_i Y_{i,t-1} + \phi_i \sum_{j \neq i} w_{ji} Y_{j,t-1}$$

- $\log(\nu_{it}) = \alpha_i +$ population offset $+$ seasonal trend $+$ covariates
- $\log(\lambda_i) = \beta_i +$ covariates
- neighbour-driven / spatiotemporal component:
  - $\log(\phi_i) = \gamma_i +$ covariates
  - w_{ji}: transmission weights
  - special case: first-order weights $w_{ji} = 1$ for neighboring regions

Multivariate modelling

Suppose now multiple time series are available (here by district):

μ_{it}: mean number of cases in district i at time t

$$\mu_{it} = \nu_{it} + \lambda_i Y_{i,t-1} + \phi_i \sum_{j \neq i} w_{ji} Y_{j,t-1}$$

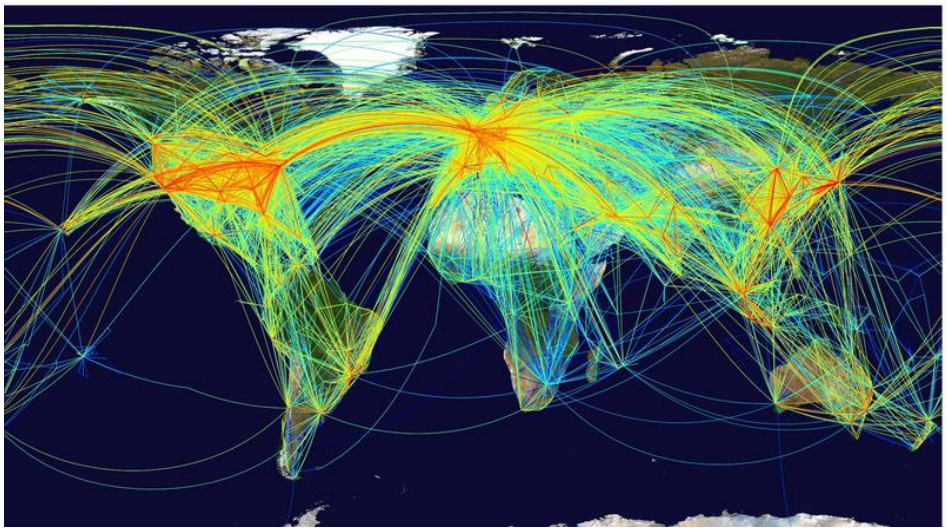
- $\log(\nu_{it}) = \alpha_i +$ population offset $+$ seasonal trend $+$ covariates
- $\log(\lambda_i) = \beta_i +$ covariates
- neighbour-driven / spatiotemporal component:
  - $\log(\phi_i) = \gamma_i +$ covariates
  - w_{ji}: transmission weights
  - special case: first-order weights $w_{ji} = 1$ for neighboring regions
- epidemic proportion now depends on all parameters and weights

  $\rightarrow$ "epidemic dominant eigenvalue"

Choice of transmission weights

Use data on connectivity between regions

- For livestock diseases: exchange of animals between farms (Schrödle et al., 2012)
- For human diseases: Use information on travel intensities of individuals (Geilhufe et al., 2014)



Source: Max Planck Institute for Dynamics and Self-Organization (`http://www.mpg.de/4406928/`)

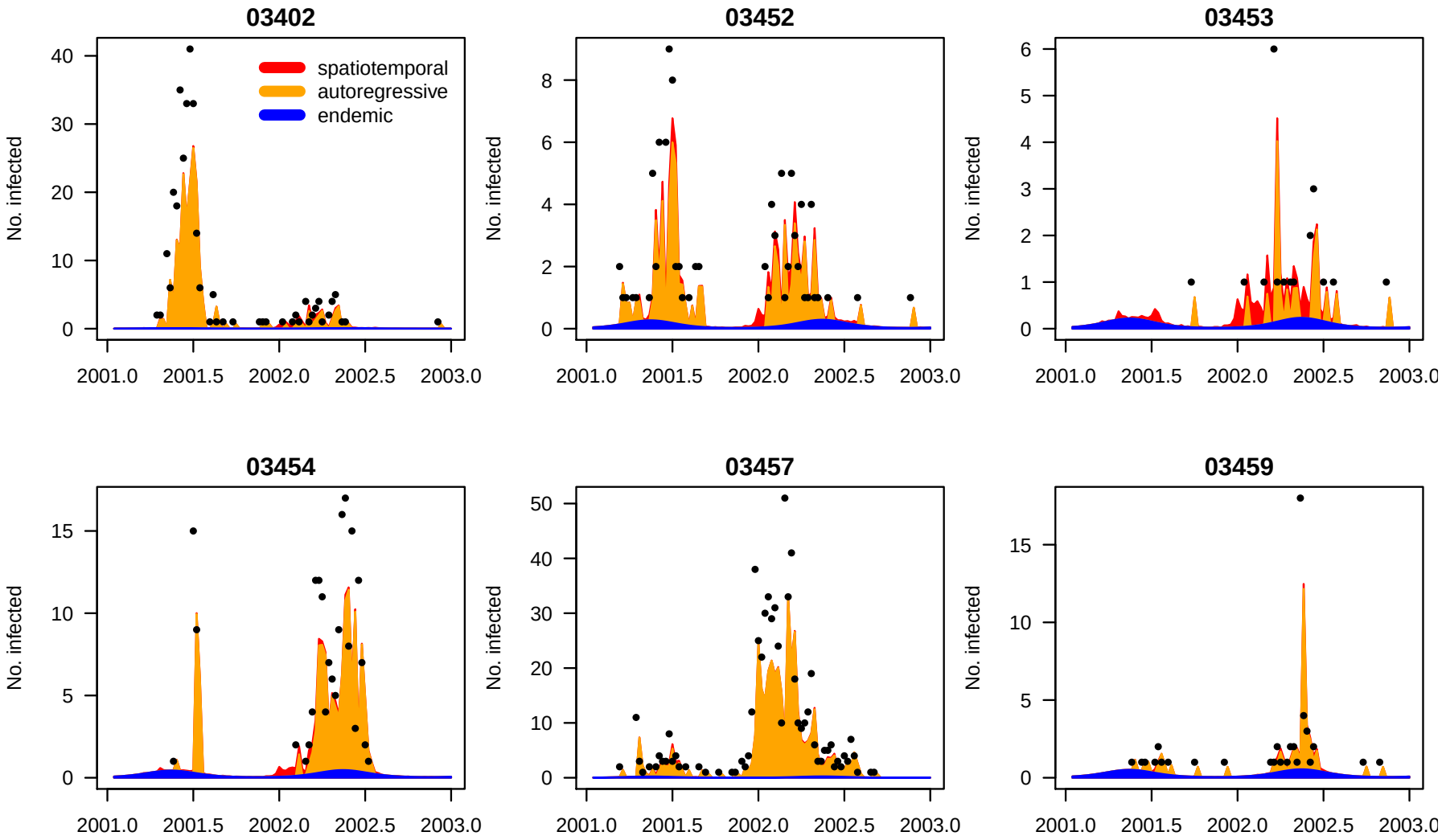
Weekly counts of measles infections

Fitting a multivariate model with first-order weights

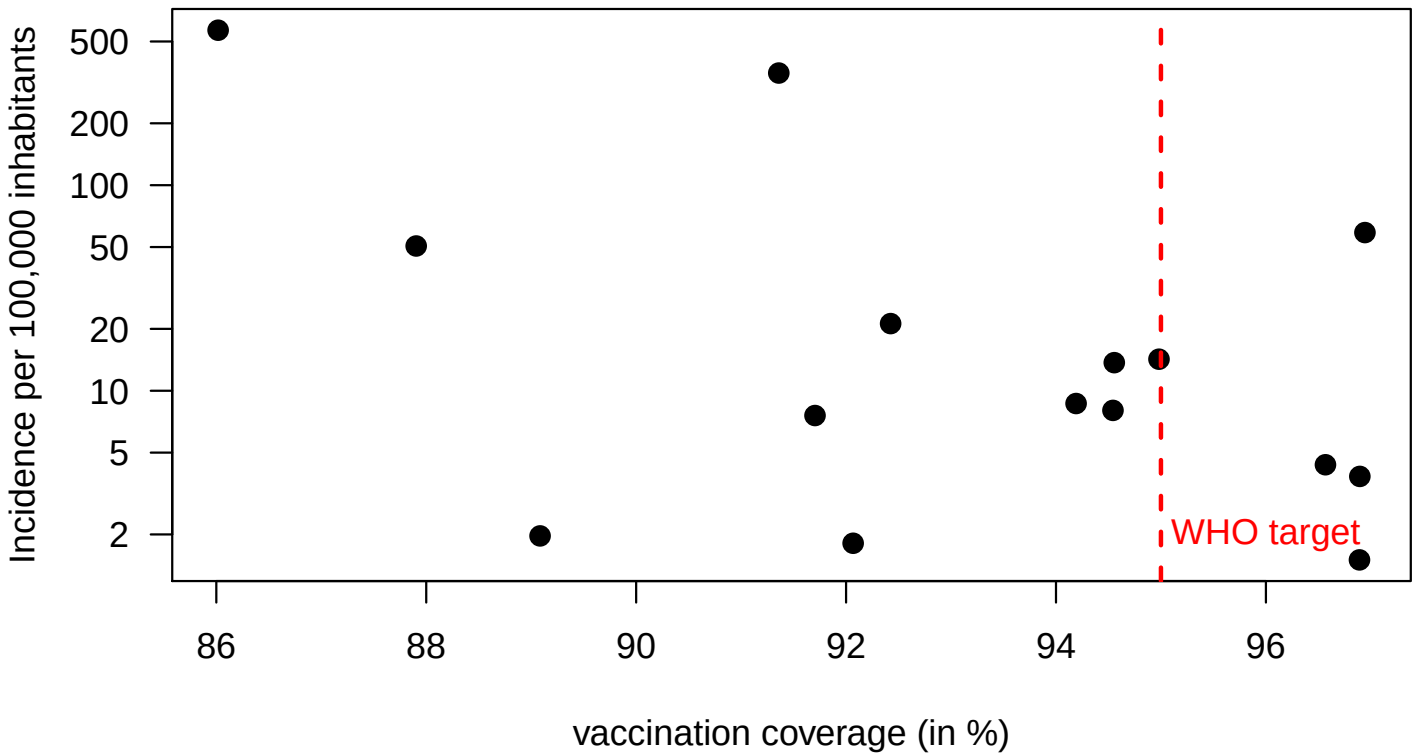
```
R> measlesModel_basic <- list(end = list(f = addSeason2formula(~1 + t,
+     period = measlesWeserEms@freq), offset = population(measlesWeserEms)),
+     ar = list(f = ~1), ne = list(f = ~1, weights = neighbourhood(measlesWeserEms) ==
+         1), family = "NegBin1")

R> measlesFit_basic <- hhh4(stsObj = measlesWeserEms, control = measlesModel_basic)
R> summary(measlesFit_basic, idx2Exp = TRUE, amplitudeShift = TRUE, maxEV = TRUE)

Call:
hhh4(stsObj = measlesWeserEms, control = measlesModel_basic)

Coefficients:
                     Estimate   Std. Error
exp(ar.1)             0.645403   0.079270
exp(ne.1)             0.015805   0.004200
exp(end.1)            1.080248   0.278839
exp(end.t)            1.001185   0.004264
end.A(2 * pi * t/52)  1.164231   0.192124
end.s(2 * pi * t/52) -0.634360   0.133500
overdisp              2.013839   0.285441

Epidemic dominant eigenvalue:  0.72

Log-likelihood:   -971.72
AIC:              1957.44
BIC:              1995.72

Number of units:        17
Number of time points:  103
```

Weekly counts of measles infections

Fitted components in the initial model for selected districts



Vaccination coverage among children in Weser-Ems region



Weekly counts of measles infections

Effect of vaccination coverage

- v_i: vaccination coverage in district i
  - $\rightarrow$ $1 - v_i$: proxy for the fraction of susceptibles
- Adding the term $\alpha_{vacc} \log(1 - v_i)$ to the endemic predictor:

```
Estimate Std. Error
  1.7181     0.2877
```

  - $\rightarrow$ Strong evidence for an association
  - $\rightarrow$ If the fraction of susceptibles doubles, the endemic measles incidence increases by a factor of $2^{1.72} = 3.3$ (95% CI: 2.2 to 4.9).
  - $\rightarrow$ AIC decreases from 1957 to 1917

A gravity model to reflect commuter-driven spread

- Scale susceptibility of districts according to population size e_i
- We add the term $\gamma_{pop} \log(e_i)$ to the spatiotemporal component:

```
> measlesFit_nepop <- update(measlesFit_vacc,
+     ne = list(f = ~log(pop)),
+     data = list(pop = population(measlesWeserEms)))
```

  $\rightarrow$ Strong evidence for such an agglomeration effect: the estimated coefficient is 2.85 (95% CI: 1.83 to 3.87) and AIC decreases from 1917 to 1887.

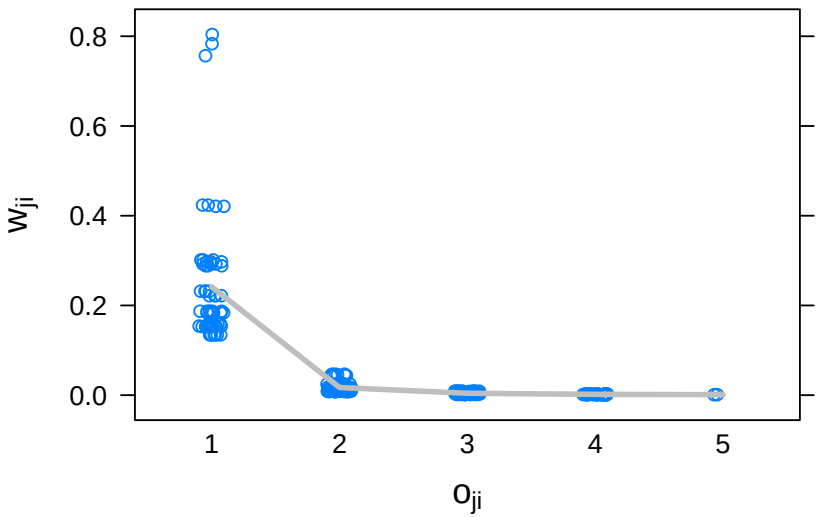
- Models where attraction to a region scales with population size are called gravity models (Xia et al., 2004; Meyer and Held, 2014).

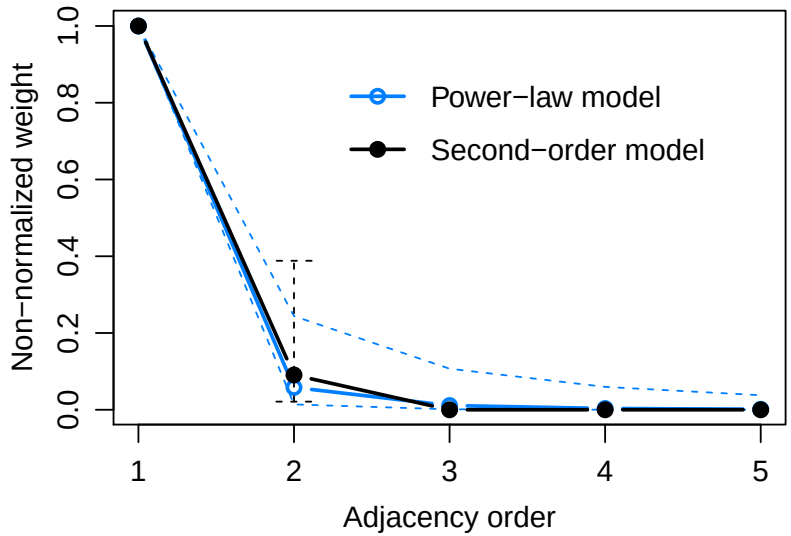
Better models for spatial dispersal

- Meyer and Held (2014) proposed to estimate the weights w_{ji} as a function of the adjacency order o_{ji} between the districts.
- A power-law model assumes $w_{ji} = o_{ji}^{-d}$ with decay parameter d.
- Normalization of the weights is recommended and applied by default.
- The resulting parameter estimate is $d = 4.1$ (95% CI: 2.0 to 6.2), which represents a strong decay of spatial interaction for higher-order neighbours.
- As an alternative to the power law, weights can be also be estimated seperately for each adjacency order.

Weekly counts of measles infections

Estimated weights and power law





AIC decreases further to 1882 for the power law.

Incorporating district-specific heterogeneity

There are several options for the district-specific parameters $\alpha_i, \beta_i, \gamma_i$:

- constant across districts, e.g., $\alpha_i = \alpha$

Incorporating district-specific heterogeneity

There are several options for the district-specific parameters $\alpha_i, \beta_i, \gamma_i$:

- constant across districts, e.g., $\alpha_i = \alpha$
- different fixed effects α_i

Incorporating district-specific heterogeneity

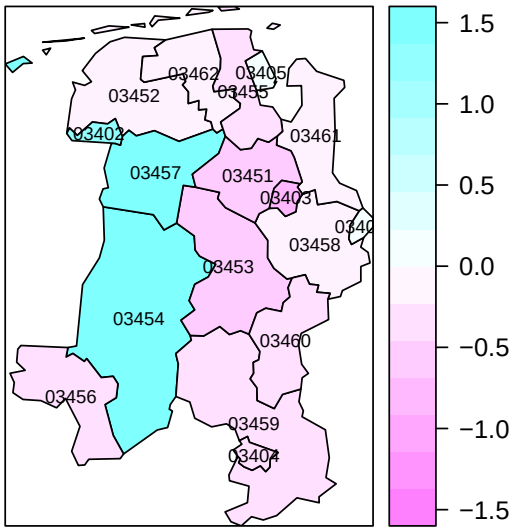
There are several options for the district-specific parameters $\alpha_i, \beta_i, \gamma_i$:

- constant across districts, e.g., $\alpha_i = \alpha$
- different fixed effects α_i
- different normally distributed random effects:
  - Independent random effects
  - Spatially correlated (CAR) priors
  - Details in Paul and Held (2011)

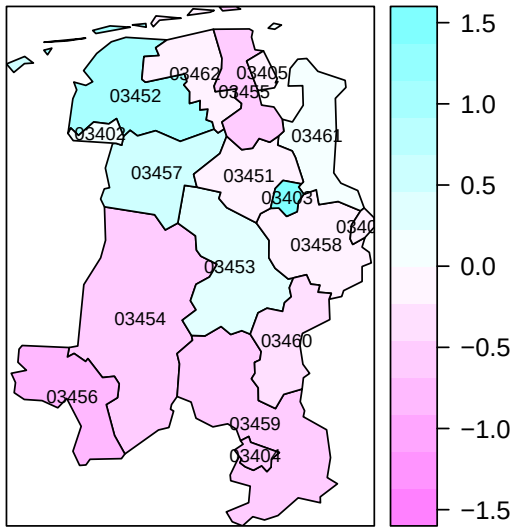
Weekly counts of measles infections

Estimated random effects in each component

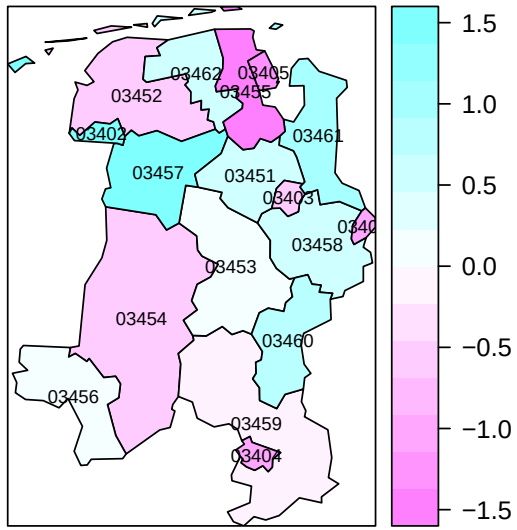
Autoregressive



Spatio-temporal

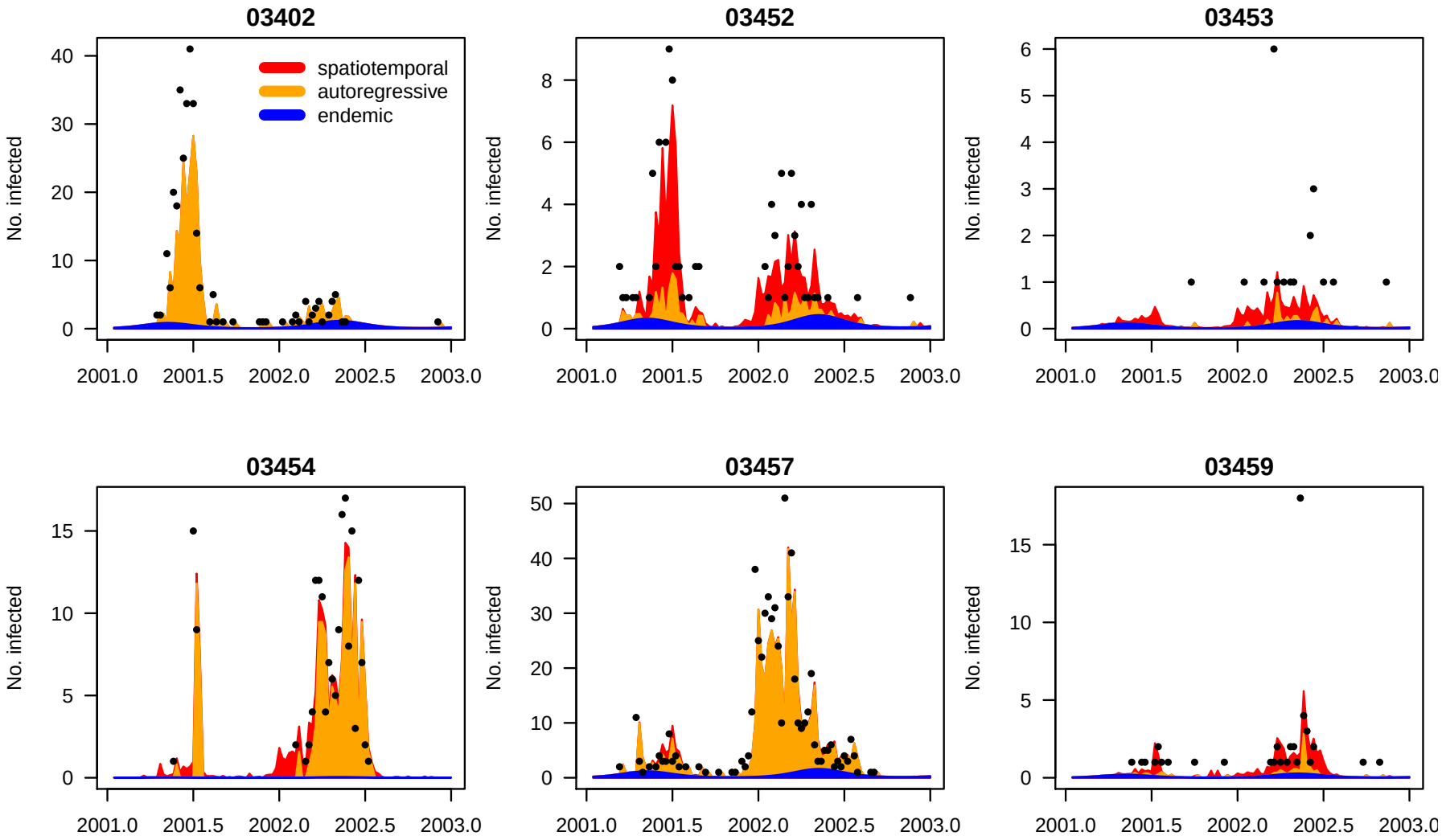


Endemic



Weekly counts of measles infections

Fitted components in the random effects model for selected districts



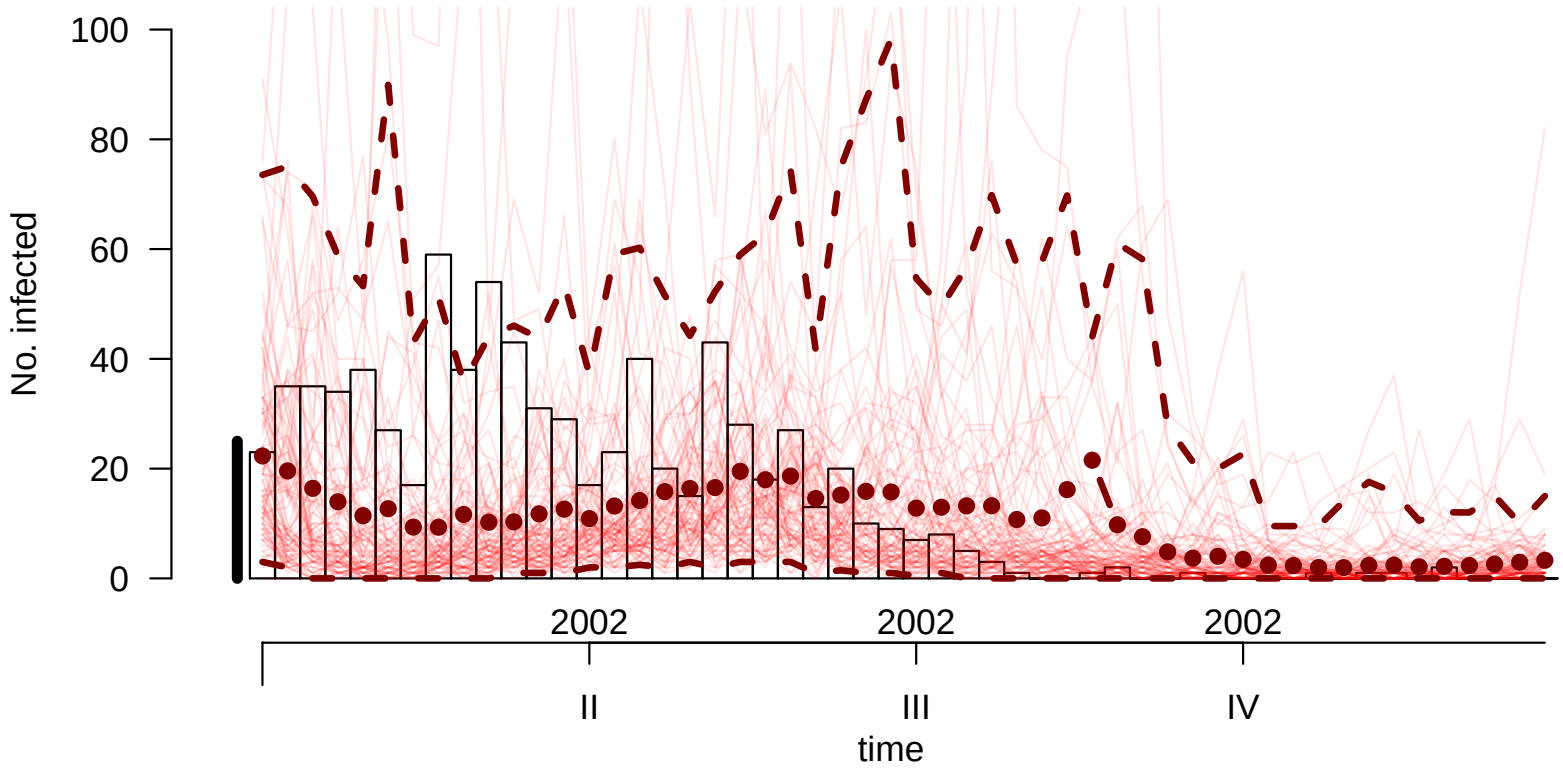
Outline

Probabilistic forecasting

- Probabilistic one-step-ahead forecasts are directly available.
- Long-term forecasts can be generated through Monte Carlo simulation.
- Suitable summary statistics can be considered, for example the final size.

Weekly counts of measles infections

Simulation-based long-term forecast for 2002



The weekly mean of the simulations is represented by dots and the dashed lines correspond to the 2.5th and 97.5th percentile. The observed counts are shown in the background.

Median final size is 424, observed is 779 (one-sided p=0.17)

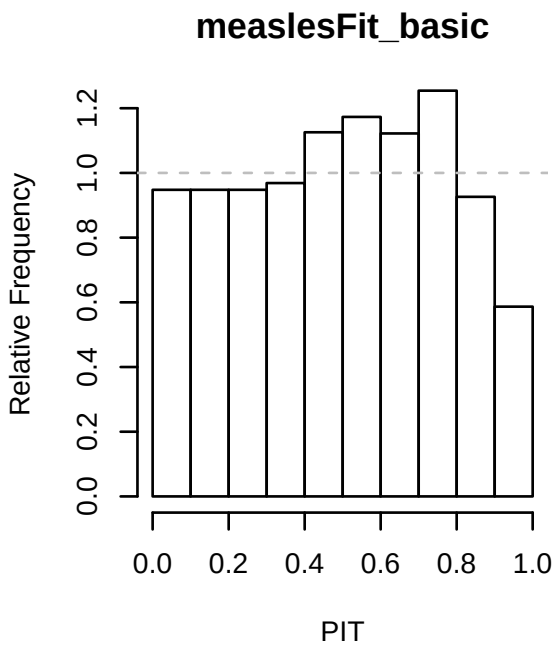
Predictive model assessment

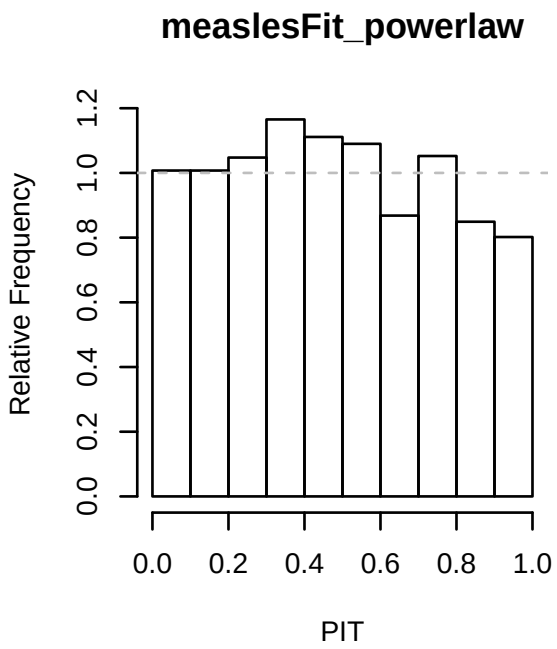
Calibration

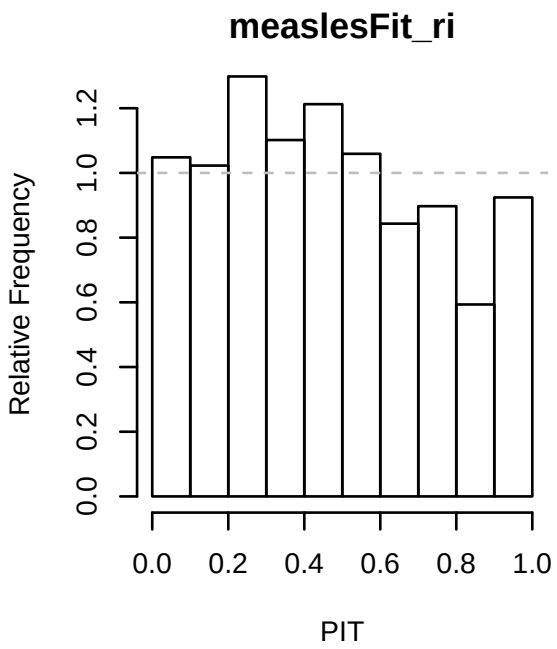
- Calibration is the statistical consistency of forecasts and observations.
- Probability integral transform (PIT) histograms can be used to assess calibration of one-step-ahead forecasts (Czado et al., 2009).
  - PIT histograms are uniformly distributed for well calibrated forecasts.
- Whether forecasts of a particular model are well calibrated can be formally investigated by calibration tests for count data based on proper scoring rules (Wei and Held, 2014).
- Examples:
  - the logarithmic score (used in CDC Epidemic Prediction Initiative)
  - the ranked probability score (RPS)

Weekly counts of measles infections

PIT histograms of the one-week-ahead predictions during the second quarter of 2002







$p = 0.21$ $p = 0.60$ $p = 0.42$

(p-values from RPS calibration test, $n = 195$)

Outline

Discussion

- We presented a statistical framework for multiple count time series.
- Meyer et al. (2016) also contains descriptions of similar endemic-epidemic models for individual-level surveillance data. Höhle (2016) links the different model classes.
- Recent work combines social contact data between age groups with the multivariate time series model (Meyer and Held, 2016).

Take home message:

1. Models are useful for prediction and understanding of infectious disease spread.
2. The `surveillance` package offers an open-source and easy-to-use implementation of the methods described.

Acknowledgments

Sebastian Meyer Michaela Paul Wei Wei



Funding:

- German Science Foundation (2003–2006)
- Munich Center of Health Sciences (2007–2010)
- Swiss National Science Foundation (2007–2010, 2012–2015)
- University of Zurich (2015–present)
- Robert Koch Institute (2012–2015)
- Swedish Research Council (2016–2019)

Literature I

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

Geilhufe, 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.

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

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

Höhle, M. (2016). Infectious disease modelling. In Lawson, A., Banerjee, S., Haining, R., and Ugarte, L., editors, *Handbook on Spatial Epidemiology*, chapter 26. CRC Press. Preprint available at `http://www.math.su.se/~hoehle/pubs/Hoehle_SpaMethInfEpiModelling2015.pdf`.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3):1612–1639.

Meyer, S. and Held, L. (2016). Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*. In revision. Preprint available at `https://arxiv.org/pdf/1512.01065.pdf`.

Literature II

Meyer, S., Held, L., and Höhle, M. (2016). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software*. Available as `http://arxiv.org/pdf/1411.0416`.

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.

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.

Wei, W. and Held, L. (2014). Calibration tests for count data. *TEST*, 23(4):787–805.

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.