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):


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

[Stratified lab-confirmed counts obtained from survstat.rki.de]
— Disease incidence maps by age group, aggregated over time
Infectious disease spread ~ social contacts

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

- 7290 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 group, region, time 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 group, region, time 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

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

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

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

Sebastian Meyer | FAU | Modelling and forecasting epidemics from surveillance counts 7 September 2017
An age-stratified, spatio-temporal, endemic-epidemic model

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

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

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

Spatial weights for \(r' \rightarrow r\), power-law decay \(w_{r'r} = (o_{r'r} + 1)^{-\rho}\)
An age-stratified, spatio-temporal, endemic-epidemic model

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

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

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 \( \nu_{grt} \) and \( \phi_{grt} \)

- 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

- **00–04**
- **05–14**
- **15–24**
- **25–44**
- **45–64**
- **65+**

- Orange: from other age groups
- Blue: within age group
- Grey: endemic
Estimated seasonality

- Endemic seasonality (multiplicative effect)
- Epidemic proportion

Calendar week:
- 27
- 35
- 43
- 51
- 3
- 7
- 11
- 19

Seasonality categories:
- 00–04
- 05–14
- 15–24
- 25–44
- 45–64
- 65+

Epidemic proportion range:
- 0.0
- 0.2
- 0.4
- 0.6
- 0.8
- 1.0

Calendar week range:
- 27
- 35
- 43
- 51
- 3
- 7
- 11
- 19
Spatial power-law weights

Power transformation of the contact matrix

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

$$DSS(F, y_{obs}) = \log|\Sigma| + (y_{obs} - \mu)\Sigma^{-1}(y_{obs} - \mu)$$
Target quantity: overall epidemic curve

purely endemic model

full model

DSS = 346.5

DSS = 334.0
Target quantity: final size by age group

purely endemic model

DSS = 67.9

full model

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


Questions? Comments? 📧 seb.meyer@fau.de
Disease incidence map

```r
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

```r
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

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

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