Spatio-temporal modelling of infectious disease surveillance data

Leonhard Held

Division of Biostatistics
Institute of Social and Preventive Medicine
University of Zurich

InFER2011, 31. March 2011

Joint work with Michaela Paul and Birgit Schrödle

Financial support by the Swiss National Science Foundation

Multivariate modelling of infectious disease surveillance data

Leonhard Held

Division of Biostatistics
Institute of Social and Preventive Medicine
University of Zurich

InFER2011, 31. March 2011

Joint work with Michaela Paul and Birgit Schrödle

Financial support by the Swiss National Science Foundation

Modelling infectious disease surveillance data

- Many countries have established surveillance systems for the routine collection of infectious disease data.
- Statistical analysis of such data is essential in the attempt to control and prevent disease, can be either prospective or retrospective.
- Notification data typically consist of time series of counts of new infections of a specific disease, observed in different areas, age groups, ...

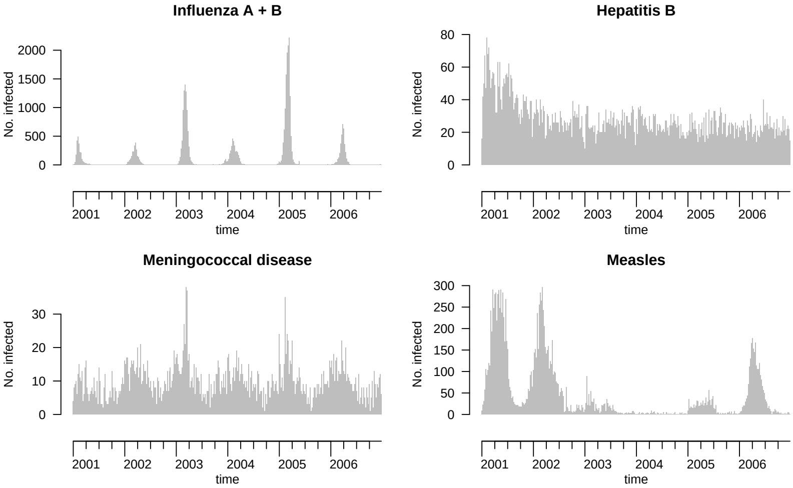
Aim

Development of a flexible model framework for the statistical analysis of surveillance data seen as multiple time series of counts.

Outline

1. Introduction
2. Modelling
3. Applications
   - Influenza and meningococcal disease
   - Measles incidence and MMR vaccination rates
   - Influenza in USA
4. Predictive validation
5. More applications
   - Influenza in Southern Germany
   - Coxiellosis in Swiss cows
6. Discussion

Surveillance data: Examples



Weekly number of cases reported to the Robert Koch Institute, Germany

Characteristics of notification data

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

How can we (statistically) analyze such data?

Accounting for temporal dependence

- A branching process with immigration is a starting point for an observation-driven model:

$$y_t | y_{t-1} \sim \text{Po}(\mu_t) \text{ with } \mu_t = \nu_t + \lambda y_{t-1}$$

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

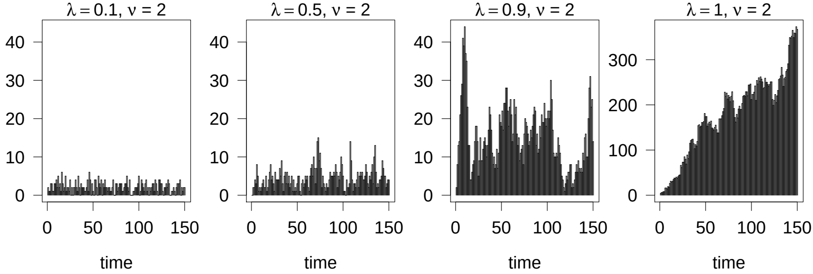
The disease incidence is additively decomposed into

- 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 will capture occasional outbreaks
- λ can be interpreted as epidemic proportion

Held et al. (2005), *Stat Model*

Accounting for temporal dependence cont.

$$y_t | y_{t-1} \sim \text{Po}(\mu_t) \text{ with } \mu_t = \nu + \lambda y_{t-1}$$



- Autoregressive coefficient $\lambda \geq 0$ determines stationarity
- In applications Poisson needs to be replaced by negative binomial distribution to adjust for overdispersion.

Multivariate formulation

Suppose now multiple time series are available:

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

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

Multivariate formulation

Suppose now multiple time series are available:

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

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

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

Multivariate formulation

Suppose now multiple time series are available:

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

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

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

Multivariate formulation

Suppose now multiple time series are available:

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

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

- $\log(\nu_{it}) = \alpha_i + \text{offset} + \text{seasonal trend} + \text{covariates}$
- $\log(\lambda_i) = \beta_i + \text{covariates}$
- $\log(\phi_i) = \gamma_i$: neighbor-driven component
- w_{ji}: known weights, e.g. adjacency-based or travel intensities

Addressing unit-specific heterogeneity

There are different options for the unit-specific parameters ν_i, λ_i, ϕ_i.

- They may be constant across units, e.g. $\phi_i = \phi$

Addressing unit-specific heterogeneity

There are different options for the unit-specific parameters ν_i, λ_i, ϕ_i.

- They may be constant across units, e.g. $\phi_i = \phi$
- They may represent different fixed effects

Addressing unit-specific heterogeneity

There are different options for the unit-specific parameters ν_i, λ_i, ϕ_i.

- They may be constant across units, e.g. $\phi_i = \phi$
- They may represent different fixed effects
- They may represent different random effects, e.g. $\alpha_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^2)$
  - Independence assumptions may be replaced with spatial (CAR) priors
  - If more than one set of parameters is taken as random, then correlation between random effects is taken into account.

Inference

- Model does not belong to the class of GL(M)Ms
- Fixed effects model:
  Maximum Likelihood estimates are obtained via a (globally convergent) Newton Raphson type algorithm

→ R package `surveillance`

- Random effects model:
  Estimation involves a multidimensional integral without closed form solution.

→ Penalized likelihood approach combined with Laplace approximation

→ R package `surveillance`

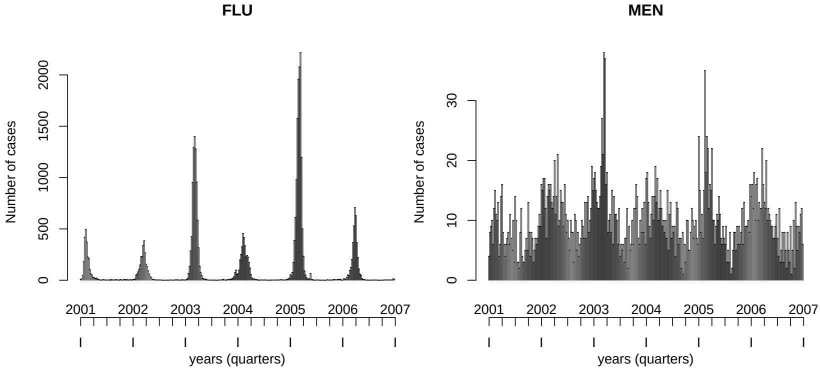
- More complex extensions require MCMC, e.g. time-varying λ
  Held et al. (2006), *Biostatistics*

Application I: Influenza and meningococcal disease

- Several studies describe an association between influenza and meningococcal disease: “Outbreaks” of meningococcal disease appear to occur at the end of influenza outbreaks
- Both influenza and meningococcal disease show seasonal variation with peak incidence rates during the winter.
- We examined whether variations in occurrence of influenza (with a delay of 1,2 weeks) were associated with changes in the incidence rate of meningococcal disease.

Paul et al. (2008), *Stat Med*

Influenza and meningococcal disease in Germany, 2001 – 2006



Modelling influenza and meningococcal disease

- Fit models with (or without) influenza cases from previous time points as explanatory variable for meningococcal disease:

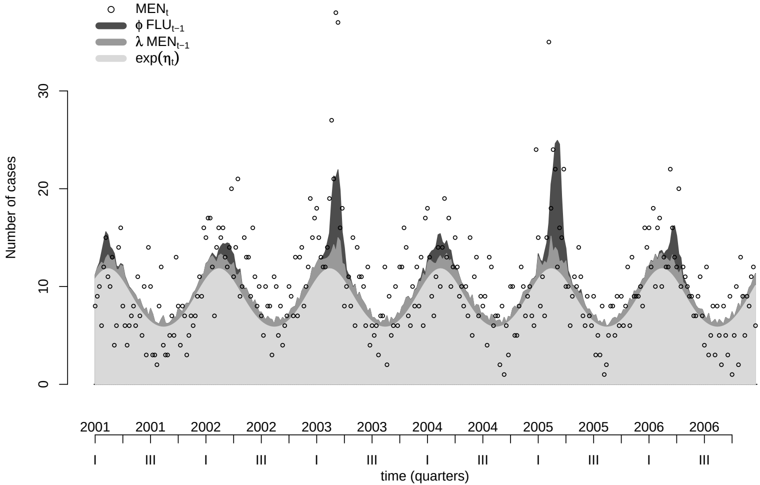
$$
\begin{aligned}
\mu_{\text{men},t} &= \nu_{\text{men},t} + \lambda_{\text{men}} y_{\text{men},t-1} + \phi y_{\text{inf},t-1} \\
\mu_{\text{inf},t} &= \nu_{\text{inf},t} + \lambda_{\text{inf}} y_{\text{inf},t-1}
\end{aligned}
$$

- Investigate also reverse direction.
- Adjust for seasonality in the endemic component for both influenza and meningococcal disease.

Results

$\hat{\lambda}$ (se)		$\hat{\phi}$ (se)			
flu	men	men → flu	flu → men	log L	p
0.74 (0.05)	0.16 (0.06)	-	-	-1889.7	14
0.74 (0.05)	0.10 (0.06)	-	0.005 (0.001)	-1881.0	15
0.74 (0.05)	0.16 (0.06)	4e-07 (1e-04)	-	-1889.7	15
0.74 (0.05)	0.10 (0.06)	4e-07 (1e-04)	0.005 (0.001)	-1881.0	16

Fitted values



Analysis with different lags

lag (weeks)	$\hat{\phi}\times 10^3$ (se $\times 10^3$)
3	2.92 (1.30)
2	4.54 (1.41)
1	5.32 (1.42)
0	5.30 (1.39)
-1	4.68 (1.31)
-2	3.73 (1.26)
-3	2.30 (1.22)

Application II: Measles in Germany

- Measles is a highly contagious disease.
- The introduction of the measles vaccine has considerably lowered the incidence level in Germany to a historical low of 2 cases per 1 000 000 inhabitants in 2004.
- However, large local outbreaks occurred in some of the federal states in recent years.
- The differences in incidence are most likely due to heterogeneous vaccination coverage rates.

Goal of analysis

Empirical investigation of the association between vaccination rates and measles epidemics using routinely collected data.

Herzog et al. (2011), *Epidemiol Infect*

Data on measles incidence and MMR vaccination rates

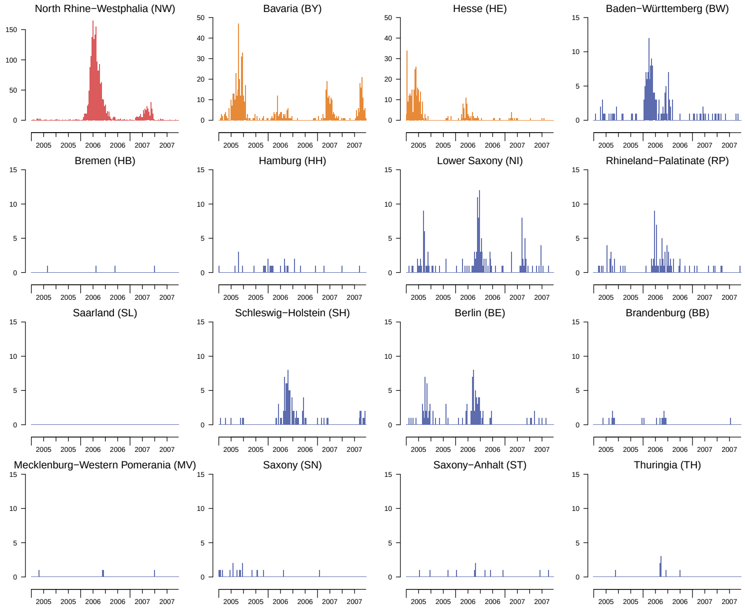
Measles incidence

- The Robert-Koch Institute (RKI) provides weekly numbers of reported cases.
- We use cases of all ages in 16 federal states for 2005 – 2007.

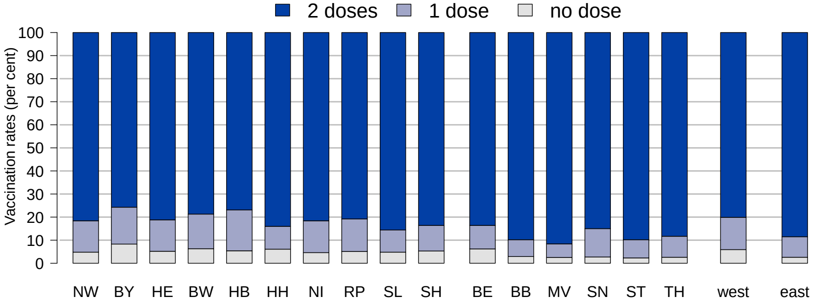
Measles-Mumps-Rubella (MMR) vaccination rates

- Coverage rates are estimated based on vaccination cards presented at yearly school entry examinations.
- They yield information about the vaccination status of children aged 4–7.
- We use state-specific rates for the 1st and 2nd dose of MMR from 2006.

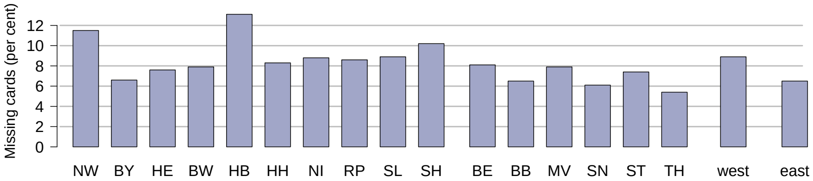
Number of reported measles cases



MMR vaccination rates in 16 federal states of Germany, estimated at school entry examinations in 2006

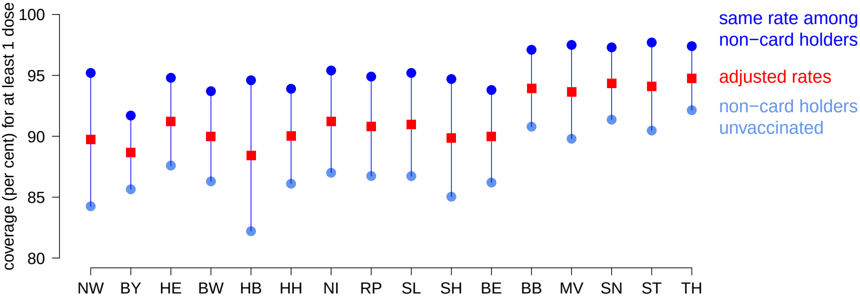


Percentage of missing vaccination cards



Adjustment of vaccination rates

- True rates are most likely overestimated by the available data as the vaccination status of card-holders is generally better.
- Nation-wide information about the degree of overestimation is not available.
- We thus assume that coverage among children without cards is half as high as among those with cards.



Model formulation for measles data

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

$$\log(\lambda_i) = \beta_0 + \beta_1 \cdot (\text{vaccination rate in state } s)$$

$$\log(\nu_{i,t}) = \text{offset} + \alpha_0 + \text{seasonal trend}$$

- Alternative model formulation: Vaccination rates enter into endemic component.

Specification of response and explanatory variables

- For measles the average time between the onset of symptoms in a primary case and a secondary case, the generation time, is about 10 days.

→ We therefore aggregate measles cases in successive biweekly periods.

- The mass action principle states:

$$\text{Rate of disease spread} \propto \begin{array}{c}\text{Susceptibles}\\ \text{(unvaccinated)}\end{array} \times \begin{array}{c}\text{Infected}\\ \text{(cases)}\end{array}$$

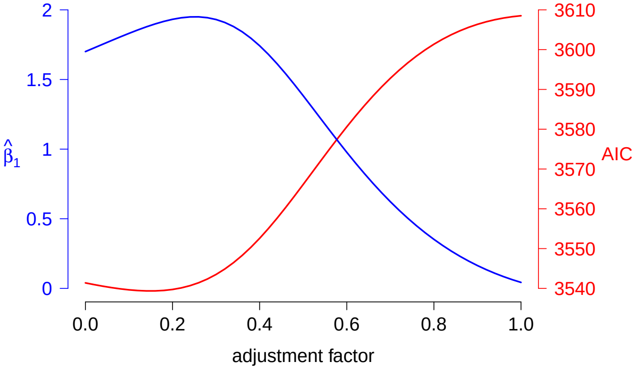
→ Taking the log proportion of unvaccinated students as covariate produces this multiplicative relation.

AIC and parameter estimates

	epidemic component		endemic component	
AIC	β_0 (se)	β_1 (se)	α_0 (se)	α_1 (se)
no covariates				
10433	-	-	3.25 (0.03)	-
3606	-0.16 (0.02)	-	1.76 (0.06)	-
log proportion of students who received at most 1 dose				
3584	1.34 (0.31)	1.02 (0.21)	1.76 (0.06)	-
3591	-0.17 (0.02)	-	3.59 (0.45)	1.17 (0.29)
log proportion of unvaccinated students				
3566	3.01 (0.52)	1.38 (0.23)	1.78 (0.06)	-
3576	-0.17 (0.02)	-	5.43 (0.69)	1.52 (0.29)

Sensitivity analysis

- We have assumed that coverage among non-card holders is 0.5 times as high as among card holders.
- We computed the AIC for several values ranging from 1 (same coverage) to 0 (all unvaccinated).

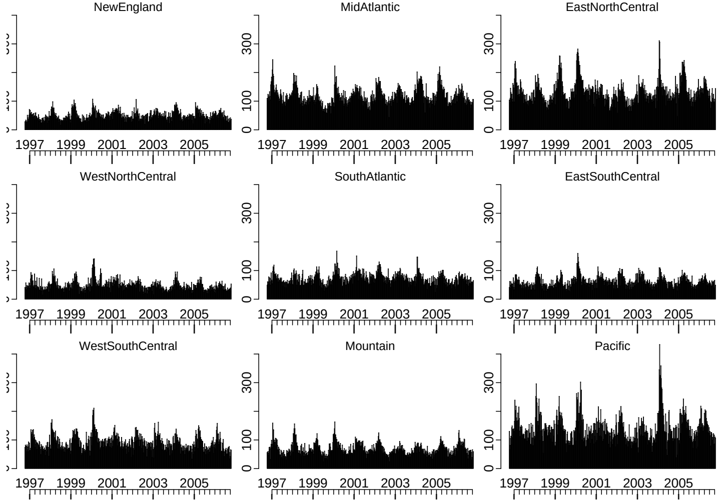


Application III: Influenza in USA, 1996 – 2006

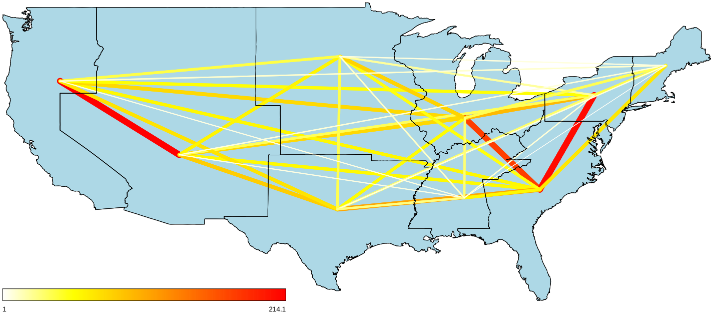
- Brownstein et al. (2006), *PLoS Med* found empirical evidence that air travel influences the annual spread of influenza in the USA
- Data on weekly mortality from pneumonia and influenza in 9 geographical regions obtained from the CDC 121 Cities Mortality Reporting system
- Data on yearly number of passengers travelling by air obtained from TranStats database, U.S. Department of Transportation

Paul et al. (2008), *Stat Med*

Data


NewEngland
MidAtlantic
EastNorthCentral
WestNorthCentral
SouthAtlantic
EastSouthCentral
WestSouthCentral
Mountain
Pacific
0
100
300
1997
1999
2001
2003
2005

Air travel data, 1997-2007



Shown is the average yearly number of passengers per 100,000

Modelling Influenza in USA

$$\mu_{it} = \exp(\nu_{it}) + \lambda y_{i,t-1} + \phi_i \sum_{j \neq i} w_{ji} y_{j,t-1}$$

Possible weights w_{ji}

- Geographical weights based on adjacencies
- Air travel information

Results - Influenza in USA

w_{ji}	$\hat{\lambda}$ (se)	$\hat{\phi}_i$ (se)	AIC	max EV
–	-	-	40300.5	-
–	0.34 (0.01)	-	39693.6	0.34
adjacencies	0.30 (0.01)	0.01 (0.01) - 0.23 (0.08)	39632.2	0.45
adjacencies (corrected)	0.30 (0.01)	0.01 (0.02) - 0.68 (0.25)	39631.6	0.44
travel	0.28 (0.01)	0.89 (3.13) - 31.58 (6.04)	39617.0	0.45
yearly travel	0.28 (0.01)	0.84 (1.09) - 28.68 (5.02)	39593.5	*

Predictive validation

- Classical model choice criteria such as AIC are problematic in the presence of random effects.
- For space-time data it is more natural to select models based on probabilistic one-step-ahead predictions.
- The often used mean squared prediction error does not incorporate prediction uncertainty.
- We use strictly proper scoring rules (Gneiting and Raftery; 2007), *JASA* which
  - compare the predictive distribution and the later observed true value y
  - simultaneously address sharpness and calibration

Proper scoring rules

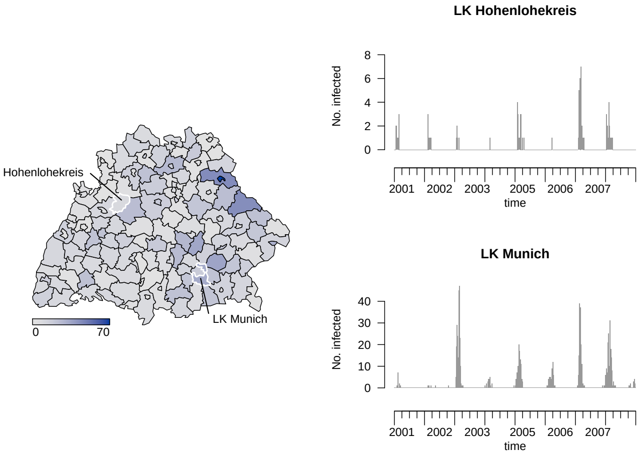
Most commonly used:

- Logarithmic score: $\text{logS} = -\log(p_y)$
- Ranked probability score: $\text{RPS} = \sum_k^{\infty}(P_k - 1(y \leq k))^2$

where p_k is the pmf and P_k is the cdf of the predictive probability distribution (Czado et al.; 2009), *Biometrics*

Note: these scoring rules are negatively oriented (the smaller the better)

Application IV: Influenza in Southern Germany



Number of laboratory confirmed influenza A and B cases in 140 administrative districts in Southern Germany, in the years 2001–2008

Paul and Held (2011), *Stat Med*

Influenza in Southern Germany

- We considered several negative binomial models, which differ depending on whether and how the autoregression is specified.
- The endemic components always includes
  - population fractions as offset
  - linear time trend and seasonal terms
  - iid random intercepts
- Model choice:
  - one-step-ahead predictions for the last two years
  - average logarithmic scores based on these predictions
  - significance of mean scores differences is investigated with a Monte Carlo permutation test

One-step-ahead predictive validation for 2007–2008

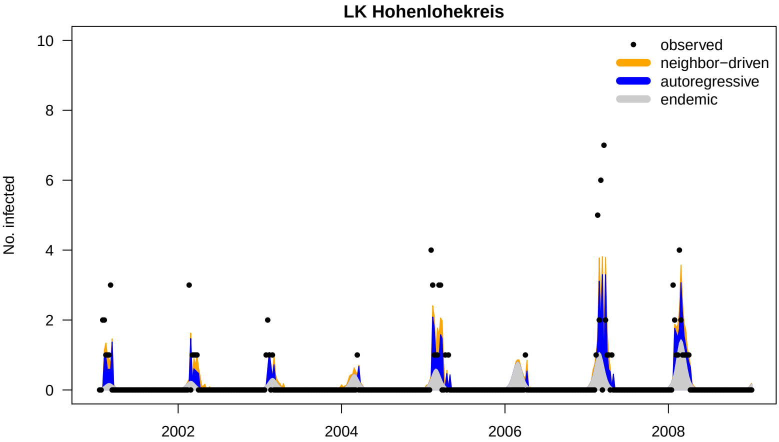
autoregressive: λ	neighbor-driven: ϕ	$\overline{\text{logS}}$
constant	random	**.563**
random	random	.564
random	constant	.565
constant	constant	.565
random	—	.569
constant	—	.569
—	random	.588
—	constant	.591
—	—	.599

One-step-ahead predictive validation for 2007–2008

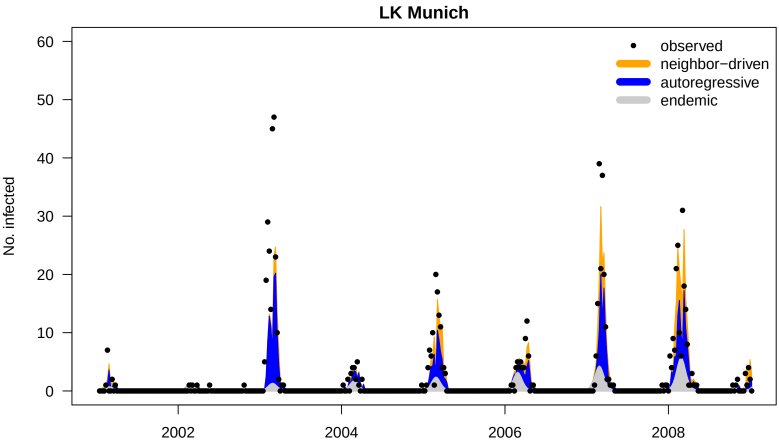
autoregressive: λ	neighbor-driven: ϕ	$\overline{\text{logS}}$	*p*-value
constant	random	**.563**	
random	random	.564	.5979
random	constant	.565	.0830
constant	constant	.565	.0353
random	—	.569	.0018
constant	—	.569	.0006
—	random	.588	.0001
—	constant	.591	.0001
—	—	.599	.0001

Monte Carlo *p*-values based on 9999 permutations

Fitted incidence



Fitted incidence



Application V: Coxiellosis in Swiss cows

- Data on Coxiellosis incidence on Swiss farms from 2004 to 2009 for 184 Swiss regions and the Principality of Liechtenstein
- A herd is denoted a case if at least one diseased animal was detected.
- Very low incidence and long reporting delays (disease is not detected until an abortion takes place) → aggregation to yearly counts
- Question: Is there spatio-temporal spread of the disease and, if yes,
    - only local (adjacency-based)?
    - or associated with cattle trade?

Schrödle et al. (2011), *submitted*

An alternative parameter-driven model

→ Latent Gaussian model

$$\mu_{it} = \lambda \mu_{i,t-1} + \phi \sum_{j \neq i} w_{ji} \mu_{j,t-1} + \epsilon_{it}$$

so $\boldsymbol{\mu}_t$ follows a vector-autoregressive (stationary) process

- Inference requires a fully Bayesian perspective using Integrated Nested Laplace Approximations (INLA)
- INLA also provides predictive distributions

→ Comparison of several parameter-driven (PM) and observation-driven (OM) models using mean predictive scores for 2009.

Results

	$\overline{\text{logS}}$		$\overline{\text{RPS}}$	
w_{ji}	*PM*	*OM*	*PM*	*OM*
–	0.583	0.624	0.239	0.257
adjacencies	0.547	0.593	0.218	0.246
cattle trade	0.549	0.590	0.214	0.236
cattle trade (relative to # herds)	0.554	0.583	0.218	0.234
$\sqrt{\text{cattle trade}}$	0.557	0.619	0.217	0.255
log cattle trade	0.575	0.624	0.230	0.257
rel. cattle trade + mean herd size	**0.549**	**0.578**	**0.212**	**0.232**

- Differences between cattle trade and adjacency-based weights are not significant.
- Difference between best *PM* and *OM* model are borderline significant ($p = 0.02$ for $\overline{\text{logS}}$, $p = 0.10$ for $\overline{\text{RPS}}$).
- The *PM* model seems better in predicting higher counts.

Discussion

- Useful statistical modelling framework for infectious disease surveillance counts.
- Ready-to-use software, easy to fit.
- Predictive validation with proper scoring rules is intuitive model choice criterion for (multiple) time series.
- Latent Gaussian hierarchical models may be a useful alternative in certain applications.

References

Brownstein, J. S., Wolfe, C. J. and Mandl, K. D. (2006). Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States, *PLoS Medicine* **3**(10): e401.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**: 359–378.

Held, L., Hofmann, M., Höhle, M. and Schmid, V. (2006). A two-component model for counts of infectious diseases, *Biostatistics* **7**: 422–437.

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

Herzog, S. A., Paul, M. and Held, L. (2011). Heterogeneity in vaccination rates explains the size and occurrence of measles epidemics in German surveillance data, *Epidemiology and Infection* **139**: 505–515.

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* . to appear.

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

Schrödle, B., Held, L. and Rue, H. (2011). Assessing the impact of network data on the spatio-temporal spread of infectious diseases, *Technical report*. submitted.