

# A Model-based View on Outbreak Detection

Leonhard Held, Michaela Paul and Daniel Sabanés Bové

Biostatistics Unit  
Institute of Social and Preventive Medicine  
University of Zurich

Conference on

Statistical Methods for Outbreak Detection

Open University, Milton Keynes, 19 May 2010

# Outline

- 1 Introduction
- 2 Model-based outbreak prediction
- 3 Model validation
- 4 Discussion

# 1. Introduction

- This talk is about outbreak detection for routinely collected surveillance data seen as (multiple) time series of counts.
- Our view is that outbreak detection should be viewed as a **probabilistic prediction** problem:

outbreak detection → **outbreak prediction**

- Such a **model-based** view avoids problems in the definition of outbreaks necessary to evaluate the performance of standard outbreak detection algorithms currently in use.
- Instead, statistical methodology to evaluate the performance of **one-step-ahead forecasts** can be used to validate models for **outbreak prediction**.

# Evaluating the performance of outbreak detection algorithms

- Evaluation (Sensitivity, specificity, probability of false alarm, etc.) of an outbreak detection algorithm requires **complete information** on observed outbreaks.
- However, there is no universal agreement on the **definition** of an outbreak and only far from complete information on possible outbreaks does typically exist.
- **Simulation studies** are often performed where artificial outbreaks are added to a reference model.
- However, results from such studies will depend on the assumed form of the reference model and the artificial outbreak and will typically not reflect the vagaries of surveillance data.

## 2. Model-based outbreak prediction

Goal: Development of a **realistic** stochastic model for time series of infectious disease counts.

Features that should be taken into account:

- Count data, possibly overdispersion
- Epidemic nature, residual autocorrelation
- No information about number of susceptibles
- Long-term temporal trends, seasonality
- Model should provide “reasonable” probabilistic one-step-ahead forecasts.

## Our model approach

- A compromise is needed between **mechanistic** and **empirical** modelling.
- Our model is based on a generalized **branching process** with immigration.
- Note: Branching process is a useful approximation of SIR-models in the absence of information on susceptibles.
- Explicit decomposition of the incidence in **endemic** and **epidemic** component (Held et al., 2005).
- Past counts act **additively** on disease incidence  
→ model is not a GLM

## Basic model

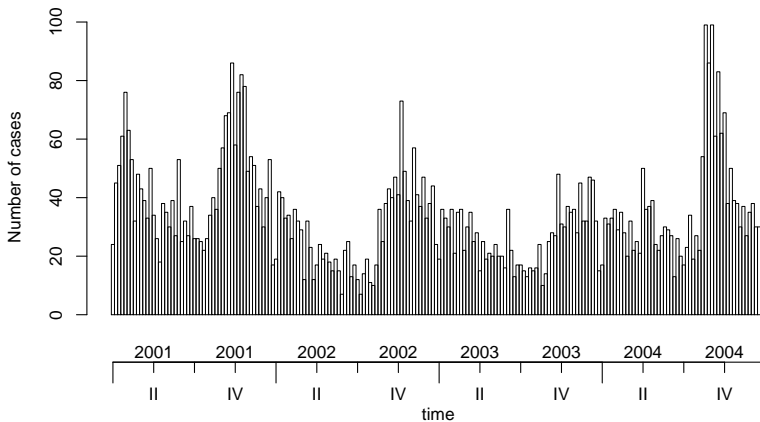
$$\begin{aligned}
 X_t &\sim \text{Po}(\nu_t) \\
 Y_t &\sim \text{Po}(\lambda Z_{t-1}) \\
 \rightarrow Z_t = X_t + Y_t &\sim \text{Po}(\underbrace{\nu_t + \lambda Z_{t-1}}_{\mu_t})
 \end{aligned}$$

- Autoregressive coefficient  $\lambda \geq 0$  determines stationarity of  $Z_t$ , can be interpreted as **epidemic proportion** if  $\lambda < 1$ .
- $\log \nu_t$  is modelled parametrically with long-term temporal and seasonal trends ( $\omega_s = 2\pi s/52$  for weekly counts):

$$\log(\nu_t) = \alpha + \beta \cdot t + \sum_{s=1}^S (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t))$$

- Adjustments for **overdispersion** straightforward: Replace  $\text{Po}(\mu_t)$  by  $\text{NegBin}(\mu_t, \psi)$ -Likelihood.

# Example: Hepatitis A in Germany 2001-2004



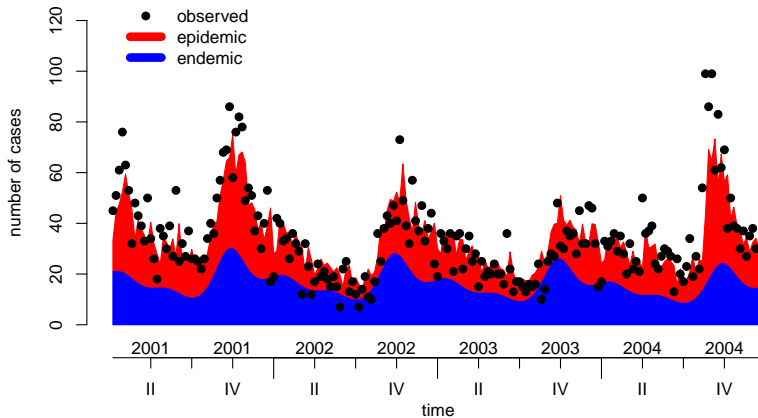
## Parameter estimates

Results from model fit using the R-package `surveillance`:

$S$	$\hat{\lambda}$ (se)	$\hat{\psi}$ (se)	$\log L$	$p$	AIC
1	0.62 (0.03)	-	-890.2	5	1790.5
1	0.60 (0.06)	0.07 (0.01)	-769.9	6	1551.9
2	0.55 (0.06)	0.07 (0.01)	-766.2	8	1548.5
3	0.52 (0.06)	0.06 (0.01)	-763.2	10	1546.4
4	0.51 (0.06)	0.06 (0.01)	-762.7	12	1549.4

Note:  $Z \sim \text{NegBin}(\mu, \psi)$  with  $E(Z) = \mu$ ,  $\text{Var}(Z) = \mu(1 + \psi\mu)$

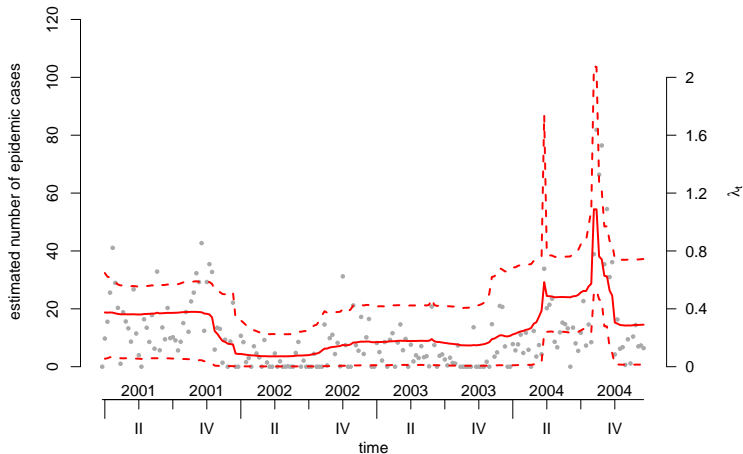
## Fitted values



## Piecewise constant epidemic parameter

- Model can be extended to allow for **piecewise constant** autoregressive parameter  $\lambda_t$  (Held et al., 2006) using a **multiple changepoint model**.
- Number and location of changepoints unknown
- Bayesian inference using MCMC possible
- However, computation of **one-step-ahead forecast distribution** is tedious, as it involves a complete re-fit of the model.

# Results with piecewise constant autoregressive parameter



## A sequential algorithm for outbreak prediction

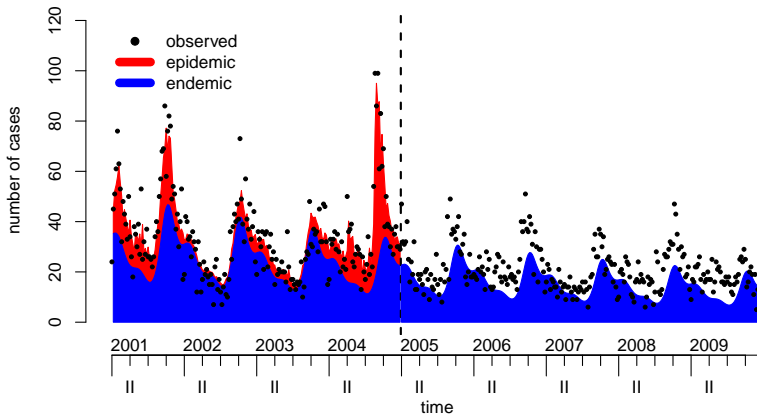
- 1 Fit Held et al. (2006) model to **training data**.
- 2 Extrapolate endemic component  $X_t$  to **test data**.
- 3 Compute epidemic component

$$Y_t = \max(Z_t - X_t, 0)$$

and adjust for overdispersion (estimated from **training data**) to ensure  $E(Y_t) \approx \text{Var}(Y_t)$ .

- 4 Choose prior for  $\lambda_t$  based on estimates from **training data**.
- 5 Fit **sequential** Monte Carlo change-point algorithm (Fearnhead, 2006; Fearnhead and Liu, 2007) to epidemic component  $Y_t$  and generate iid samples from one-step-ahead forecast distributions.

# Visualization



# Outbreak prediction

Suitable quantities:

- Compute one-step-ahead probability

$$\mathbb{P}(Y_{t+1} \geq c | Y_{1:t})$$

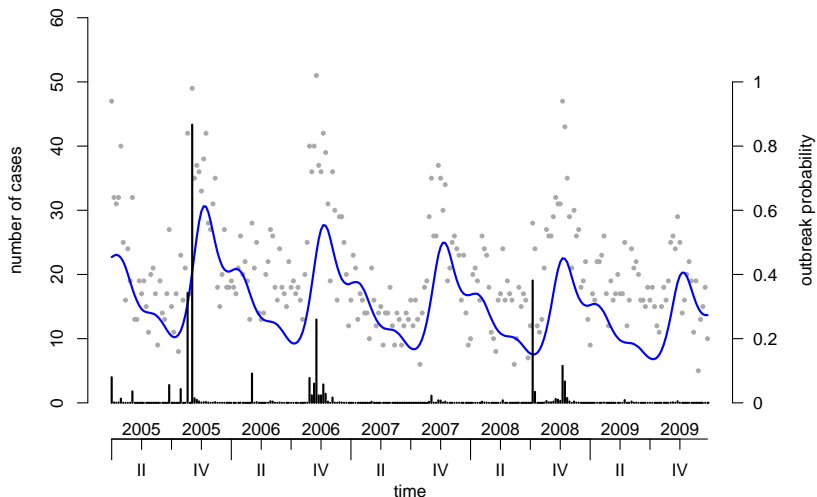
for suitably chosen constant  $c$ .

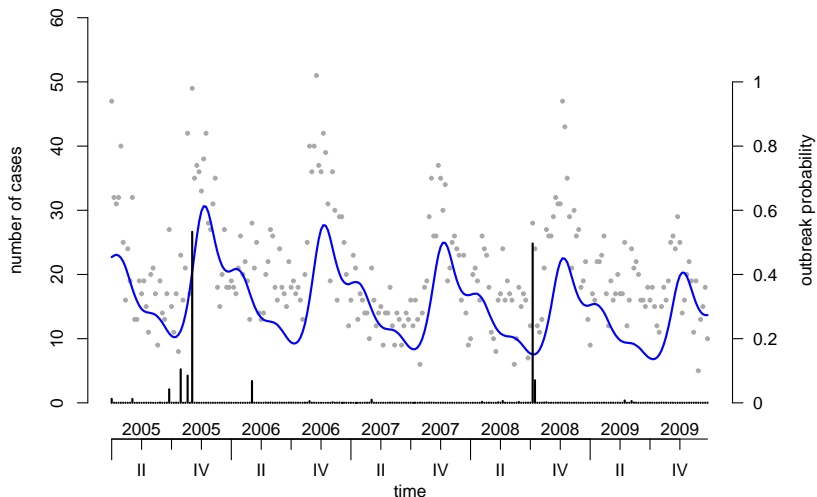
- Compute one-step-ahead probability

$$\mathbb{P}(\lambda_{t+1} \geq 1 | Y_{1:t})$$

of **non-stationarity**, i.e. eventual **explosion** of the process.

# One-step-ahead probability $\text{Prob}(Y_{t+1} \geq 30 | \mathbf{y}_{1:t})$



One-step-ahead probability  $\text{Prob}(\lambda_{t+1} \geq 1 | \mathbf{y}_{1:t})$ 

### 3. Model validation

- We validate the models based on **probabilistic one-step-ahead predictions**.
- In the absence of a suitable loss function, **proper scoring rules** (Gneiting and Raftery, 2007), which assess both **calibration** and **sharpness**, are recommended.
- **Calibration** alone is assessed using **PIT histograms** for count data (Smith, 1985; Czado et al., 2009):

$$\text{PIT}_{t+1} = \text{Prob}(Y_{t+1} \leq y_{t+1} | \mathbf{y}_{1:t}),$$

here  $y_{t+1}$  denotes the actually observed value.

- $\text{PIT}_{t+1}$  is uniformly distributed if forecast is well calibrated.
- The **mean squared error** (MSE) score does not incorporate prediction uncertainty.

## Proper scoring rules

- The **log score** is strictly proper and defined as

$$\text{LogS}(Y_{t+1}, y_{t+1}) = -\log f_{Y_{t+1}}(y_{t+1}),$$

the log predictive density ordinate at the observed value  $y_{t+1}$ .

- Note: Minus the sum of the log-scores of the one-step-ahead forecasts equals the log marginal likelihood  $\log f(\mathbf{y})!$
- A popular strictly proper score which is less sensitive to outliers but sensitive to distance is the so-called **ranked probability score**

$$\text{RPS}(Y_{t+1}, y_{t+1}) = \sum_{s=0}^{\infty} (\text{Prob}(Y_{t+1} \leq s) - \mathbf{1}(y_{t+1} \leq s))^2,$$

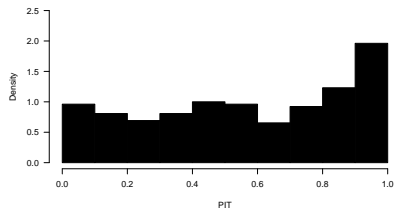
the sum of the **Brier scores** for binary predictions at all possible thresholds  $s$ .

# Comparison of models with and without changepoints

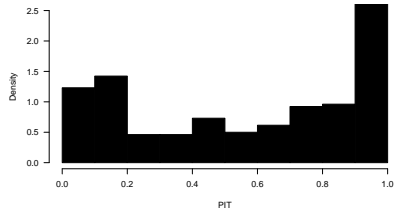
	with changepoints	without changepoints
$\log f(\mathbf{y})$	-604.29	-670.88
$\overline{\text{LogS}}$	2.32	2.58
$\overline{\text{RPS}}$	1.61	1.63
MSE	10.41	8.30

# PIT histograms

with changepoints



without changepoints



## Summary/Discussion

- A **model-based** approach for **outbreak prediction**
- Extrapolation of endemic component and prior elicitation based on **training data**
- Application of efficient algorithms to compute **one-step-ahead forecast distributions**
- Model validation using proper scoring rules and PIT histograms based on **test data**
- Refinement of the model may be warranted.
- An extension to **spatio-temporal** model-based outbreak detection is desirable, see Paul et al. (2008) for suitable model extensions.

## Literature

- Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics* 65, 1254–1261.
- Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing* 16(2), 203–213.
- Fearnhead, P. and Z. Liu (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society. Series B (Methodological)* 69(4), 589–605.
- Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102, 359–378.
- Held, L., M. Hofmann, M. Höhle, and V. Schmid (2006). A two-component model for counts of infectious diseases. *Biostatistics* 7, 422–437.
- Held, L., M. Höhle, and M. Hofmann (2005). A statistical framework for the analysis of multivariate infectious disease surveillance data. *Statistical Modelling* 5, 187–199.
- Paul, M., L. Held, and A. M. Toschke (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* 27, 6250–6267.
- Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting* 4, 283–291.