Modeling time series of disease data by the R-package tscount

Roland Fried[1] Konstantinos Fokianos[2] Tobias Liboschik[1]

[1]Department of Statistics, TU Dortmund University, Germany

[2]Department of Mathematics and Statistics, University of Cyprus, Cyprus

Jahrestagung 2018, September 27, 2018

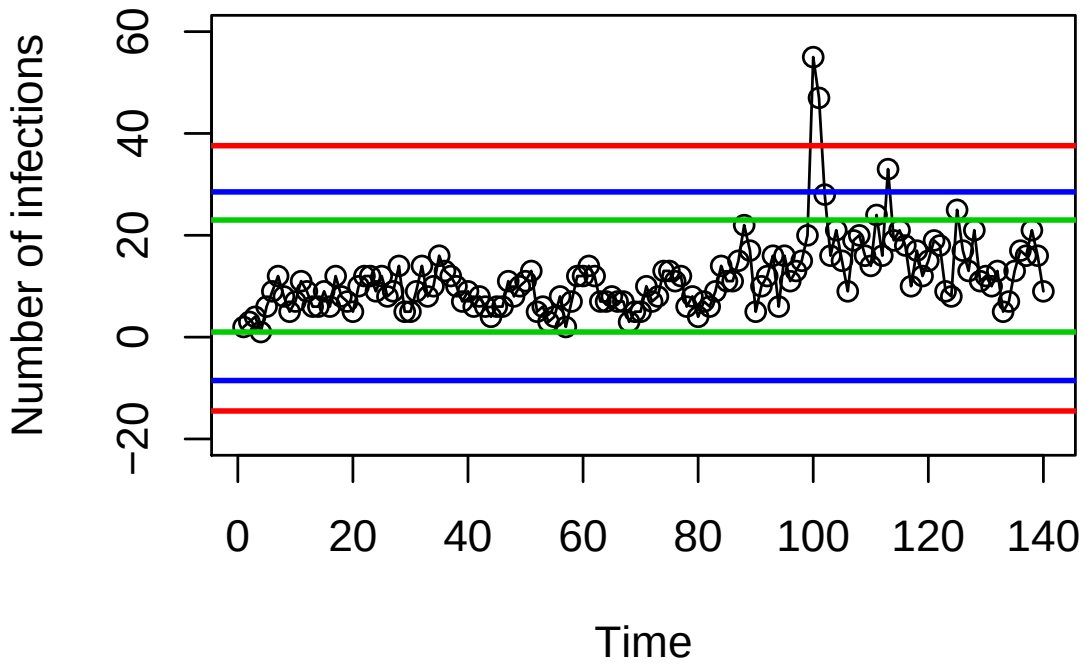
technische universität dortmund

SFB 823

Introduction

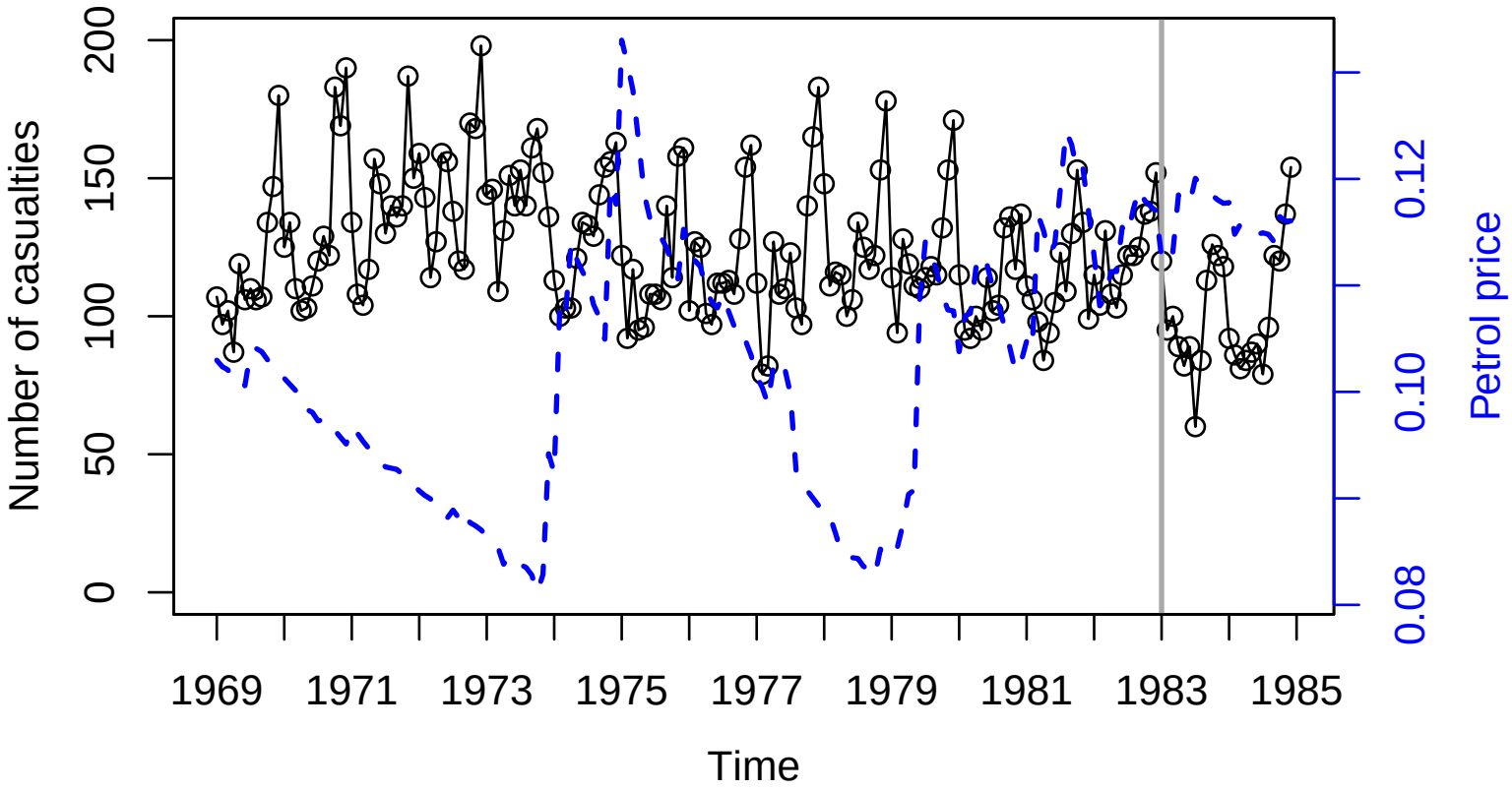
Number of campylobacterosis infections in Quebec

Ferland, Latour and Oraichi (2006)



with Gaussian and Poisson detection limits (99.982%-quantiles)

Road casualties in Great Britain



seatbelts compulsory since 1983; petrol price as covariate

Harvey and Durbin (1986)

package `tscount`

- project's website:
  `http://tscount.r-forge.r-project.org`
- available from CRAN and R-Forge
- article "tscount: An R package for the Analysis of Count Time Series Following Generalized Linear Models",
  Journal of Statistical Software,
  `vignette("tsglm", package="tscount")`)
- comprehensive help pages
- central function: `tsglm()`

Spatio-temporal Model

Meyer, Held and Höhle (2017)

$$\mathrm{E}\left(Y_{i,t}|\mathcal{F}_{t-1}\right) = \lambda_{i,t} \quad \text{with}$$

$$\lambda_{i,t} = \beta_{i,t} Y_{i,t-1} + \phi_{i,t} \sum_{j \neq i} w_{j,i} Y_{j,t-1} \; + \; \boldsymbol{\eta}_{i,t}^{\top} X_{i,t}$$

- $\{Y_{i,t} : t \in \mathbb{N}\}$ time series at site $i = 1, \ldots, I$
- $\{\lambda_{i,t} : t \in \mathbb{N}\}$ latent conditional mean process
- $\mathcal{F}_{t-1}$ history up to time t
- $\{X_{i,t} : t \in \mathbb{N}\}$ offset of expected counts at site $i = 1, \ldots, I$.

Time Series Model

Count time series following generalized linear models

$$\mathrm{E}(Y_t|\mathcal{F}_{t-1}) = \lambda_t \quad \text{with}$$

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \tilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell g(\lambda_{t-j_\ell}) \; + \; \boldsymbol{\eta}^\top \boldsymbol{X}_t$$

- $\{Y_t : t \in \mathbb{N}\}$ time series
- $\{\lambda_t : t \in \mathbb{N}\}$ latent mean process
- $\mathcal{F}_{t-1}$ history up to time t
- $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \boldsymbol{\eta}^\top)^\top \in \Theta \subseteq \mathbb{R}^{1+p+q+r}$ regression parameters
- $g : \mathbb{R}^+ \to \mathbb{R}$ link function, $\tilde{g} : \mathbb{R}^+ \to \mathbb{R}$ transformation function
- $\{\boldsymbol{X}_t : t \in \mathbb{N}\}$ with $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})^\top$ covariates

A convenient and flexible model class

- serial correlation
- deterministic and stochastic seasonality
- deterministic trends
- covariates
- popular special cases:
  - linear model of order p, q (INGARCH):

    $$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k \, Y_{t-k} + \sum_{\ell=1}^{q} \alpha_\ell \lambda_{t-\ell}$$

    Heinen (2003), Ferland et al. (2006), Fokianos et al. (2009)
  - log-linear model of order p, q:

    $$\log(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \log(Y_{t-k} + 1) + \sum_{\ell=1}^{q} \alpha_\ell \log(\lambda_{t-\ell})$$

    Fokianos and Tjøstheim (2011)

Distributional assumptions

	Poisson	Negative Binomial
conditional distribution:		
$Y_t \mid \mathcal{F}_{t-1} \sim$	$\text{Pois}(\lambda_t)$	$\text{NegBin}(\lambda_t, \psi)$
conditional variance:		
$\text{Var}(Y_t \mid \mathcal{F}_{t-1}) =$	λ_t	$\lambda_t + \lambda_t^2 \psi$
overdispersion coefficient:		
$\sigma^2 =$	0	ψ

- Negative Binomial distribution allows for more overdispersion
- Poisson is limiting case $\phi \to \infty$ of Negative Binomial

Christou and Fokianos (2014)

Estimation and inference

- quasi maximum likelihood estimation (QMLE)
- implementation:
  - constrained optimisation with quasi-Newton algorithm
  - recursive computation of log-likelihood and score vector
    $\longrightarrow$ sensitive to initialisation under strong serial dependence
  - starting value for $\boldsymbol{\theta}$ obtained by GLM or ARMA fit
- dispersion parameter ν of Negative Binomial distribution estimated afterwards using Pearson's χ^2 statistic
- inference based on asymptotic normality (for well-behaved covariate process) or parametric bootstrap
  $\longrightarrow$ `se()`, `summary()`

Model summary (Campylobacter infections)

fitting function `tsglm()` creates object of class `'tsglm'`

```
Call:
tsglm(ts = campy, model = list(past_obs = 1, past_mean = 13),
    xreg = interventions, link = "identity", distr = "poisson")

Coefficients:
             Estimate  Std. Error
(Intercept)     3.317      0.6384
beta_1          0.369      0.0564
alpha_13        0.220      0.0740
interv_1        3.086      0.7217
interv_2       41.863      7.3974
Standard errors obtained by normal approximation.

Link function: identity
Distribution family: poisson
Number of coefficients: 5
Log-likelihood: -384.9805
AIC: 779.961
BIC: 794.6692
```

Prediction

1-step-ahead prediction:

- predictor $\widehat{Y}_{n+1} = \lambda_{n+1}$ for Y_{n+1} given $\mathcal{F}_n$
- prediction intervals using cond. Poisson/Negative Binomial distribution of Y_{n+1} given $\mathcal{F}_n$
- replacing λ_{n+1} by $\widehat{\lambda}_{n+1} = \lambda_{n+1}(\widehat{\boldsymbol{\theta}})$ increases uncertainty

***h*-step-ahead prediction:**

- $\widehat{Y}_{n+h}$ by recursive 1-step ahead predictions replacing $Y_{n+1}, \ldots, Y_{n+h-1}$ by their 1-step-ahead prediction
- prediction intervals by parametric bootstrap procedure

$\longrightarrow$ `predict()`

Model selection and assessment

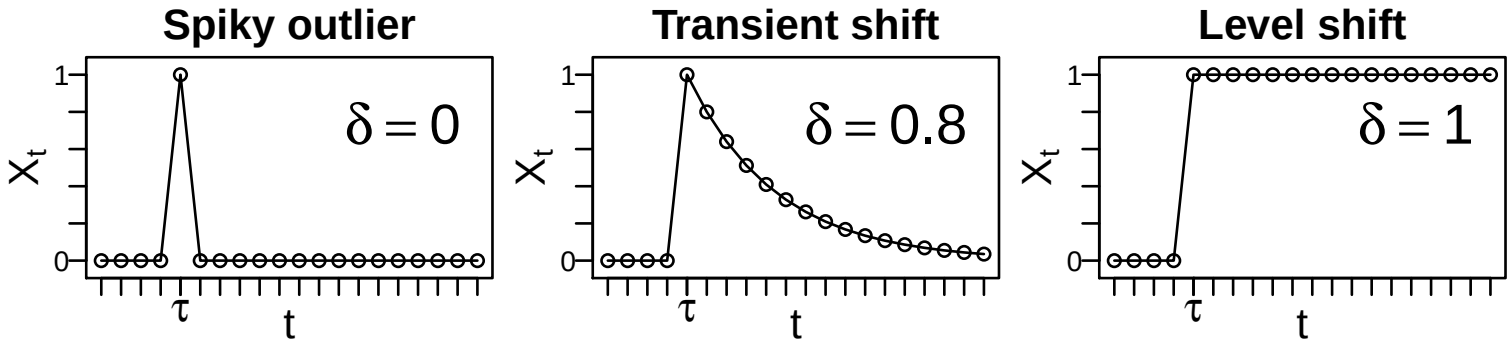
- link function
- temporal dependence structure
- seasonality
- covariates
- conditional distribution

Tools

- residual analysis (response, Pearson, Anscombe) → `residuals()`
- predictive model assessment
  - probability integral transform (PIT) histogram → `pit()`
  - marginal calibration plot → `marcal()`
  - proper scoring rules → `scoring()`
- model selection criteria → `AIC()`, `BIC()`

Retrospective intervention analysis

- model intervention at time τ by covariate $X(t) = \delta^{t-\tau} I_{[\tau,\infty)}(t)$



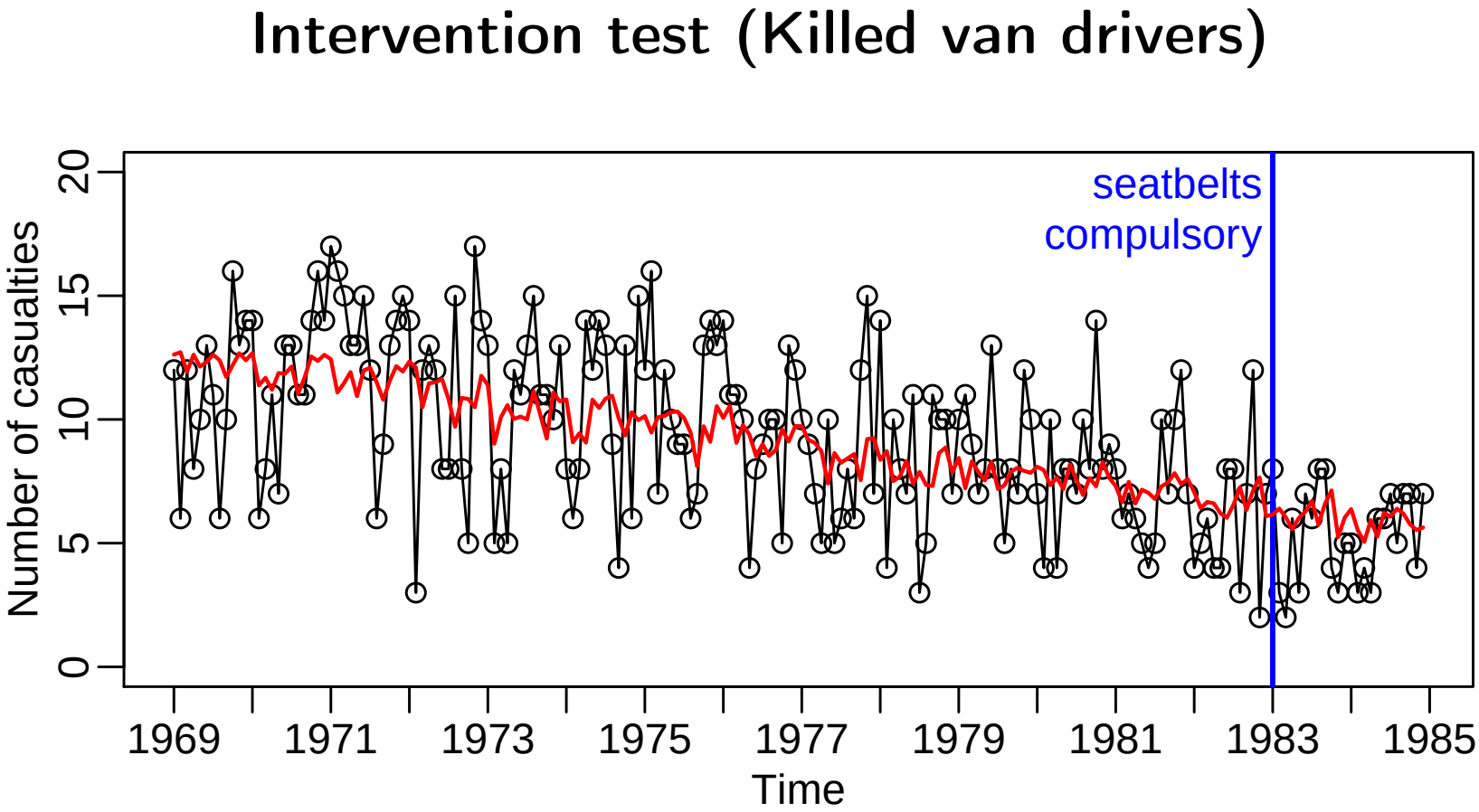
- procedures based on score test statistic:
  - estimate and eliminate intervention effects
  - test for specific type(s) of intervention(s) at known time → `interv_test()`
  - test for specific type of intervention at unknown time → `interv_detect()`
  - iterative detection of multiple interventions of unknown type at unknown times → `interv_multiple()`

Fokianos and Fried (2010, 2012), Liboschik et al. (2016)

Road casualties: known intervention

```
R> interv_test(fit_log_nbin_alldata, tau=170, delta=1,
    est_interv=TRUE)
        Score test on intervention(s) of given type at given
            time

Chisq-Statistic: 3026.885 on 1 degree(s) of freedom, p-value: 0

Fitted model with the specified intervention:

Call:
tsglm(ts = fit$ts, model = model_extended, link = fit$link,
    distr = fit$distr)

Coefficients:
(Intercept)       beta_1      beta_12  PetrolPrice      interv_1
    0.94555      0.43687      0.37327     -0.39917     -0.04514

Dispersion parameter 'size' of the negative binomial
    distribution was estimated to be 83.25749.
```

Retrospective intervention analysis



- cannot reject null hypothesis that seatbelts law has no effect
- 19.5% less van drivers killed after introduction

Campylobacterosis infections: unknown interventions

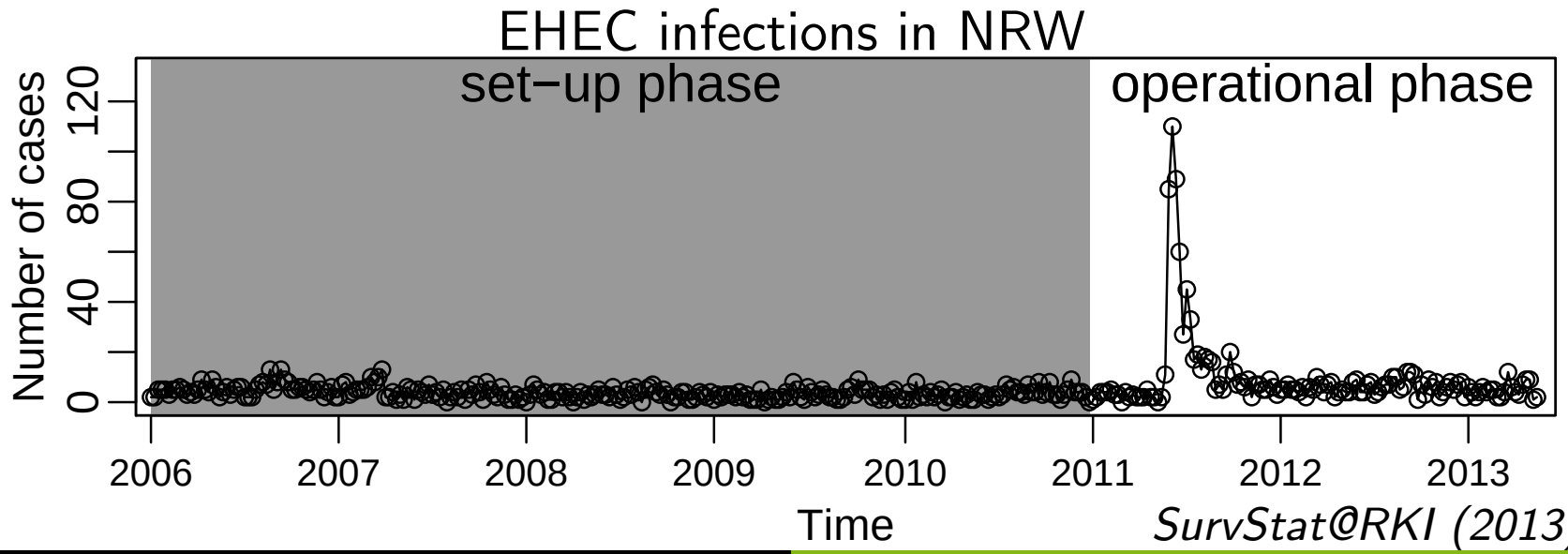
```
R> campy_fit <- tsglm(ts=campy, model=list(past_obs=1,
    past_mean=c(13)))

Coefficients:
(Intercept)       beta_1     alpha_13
     2.4306       0.5942       0.1884

R> campy_mult <- interv_multiple(fit=campy_fit, taus=80:120,
    deltas=c(0,0.8,1), B=500, signif_level=0.05)

Detect multiple intervention of unknown types at unknown times:

  tau delta      size test_statistic p_value
1  84     1  4.057509       41.19453       0
2 100     0 19.703229       42.56665       0

Fitted model with all detected interventions:

Coefficients:
(Intercept)       beta_1     alpha_13     interv_1     interv_2
     3.3169       0.3689       0.2201       3.0864      41.8628
```

Online monitoring

Application: Infectious disease surveillance

Aims:
- timely recognition of disease outbreaks
- few false alarms

Data:
- number of newly infected patients per day/week/month
- typical features: seasonality, trends, temporal dependence



Monitoring procedures

- classical procedures assume independence
- temporal dependence only considered for few specific models

Prediction-based monitoring for count time series following GLMs

1. Select & fit model to set-up phase data $y_1, \ldots, y_n$: $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\phi}_n$
2. For each $t_0 = n+1, \ldots$ in operational phase:
   a. Compute 1-step-ahead level $1-\alpha$ prediction interval for Y_{t_0}
   b. Give alarm if observation y_{t_0} lies outside interval

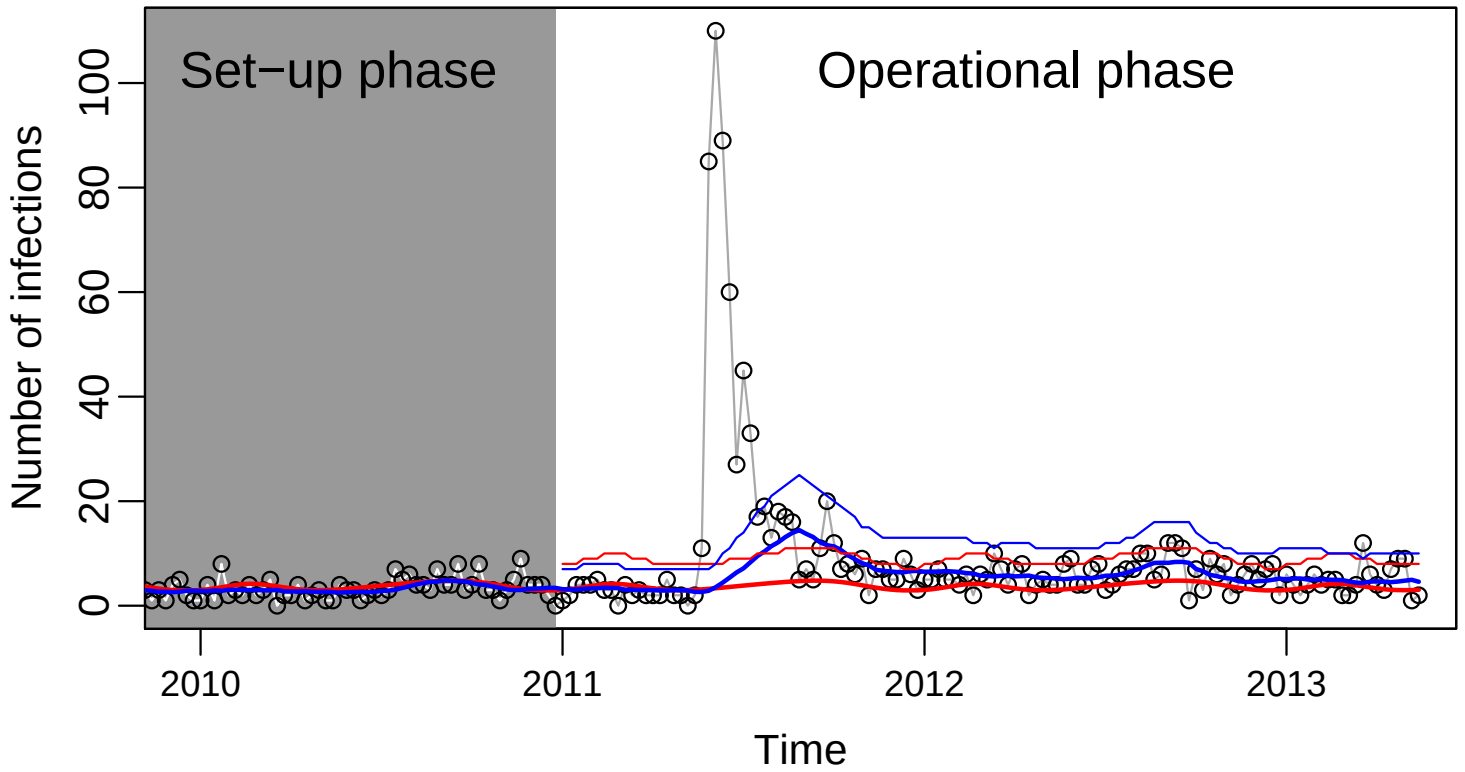
$\Rightarrow$ controls false alarm rate α

1-step-ahead prediction:

$$Y_{t_0}(1) \sim \text{NegBin}\left(\widehat{\lambda}_{t_0+1}, \widehat{\phi}_n\right) \text{ with } \widehat{\lambda}_{t_0+1} = \lambda_{t_0+1}\left(\widehat{\boldsymbol{\theta}}_n\right)$$

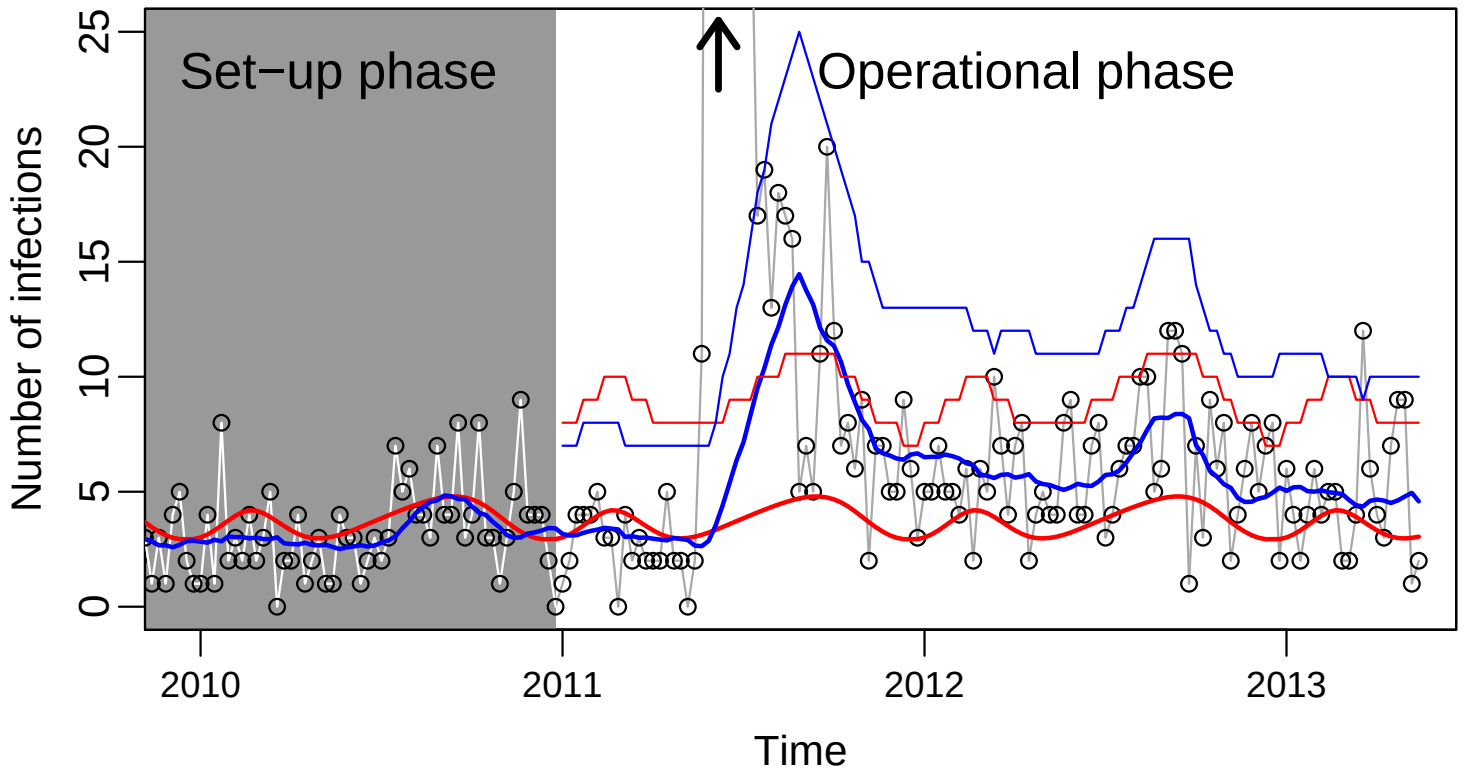
$\longrightarrow$ `predict()`

Example: EHEC infections in NRW



- 1-step-ahead pred. (bold) and one-sided 97.5% PIs (thin)
- independence model does not react to previous observations
- alarm at the same week

Example: EHEC infections in NRW

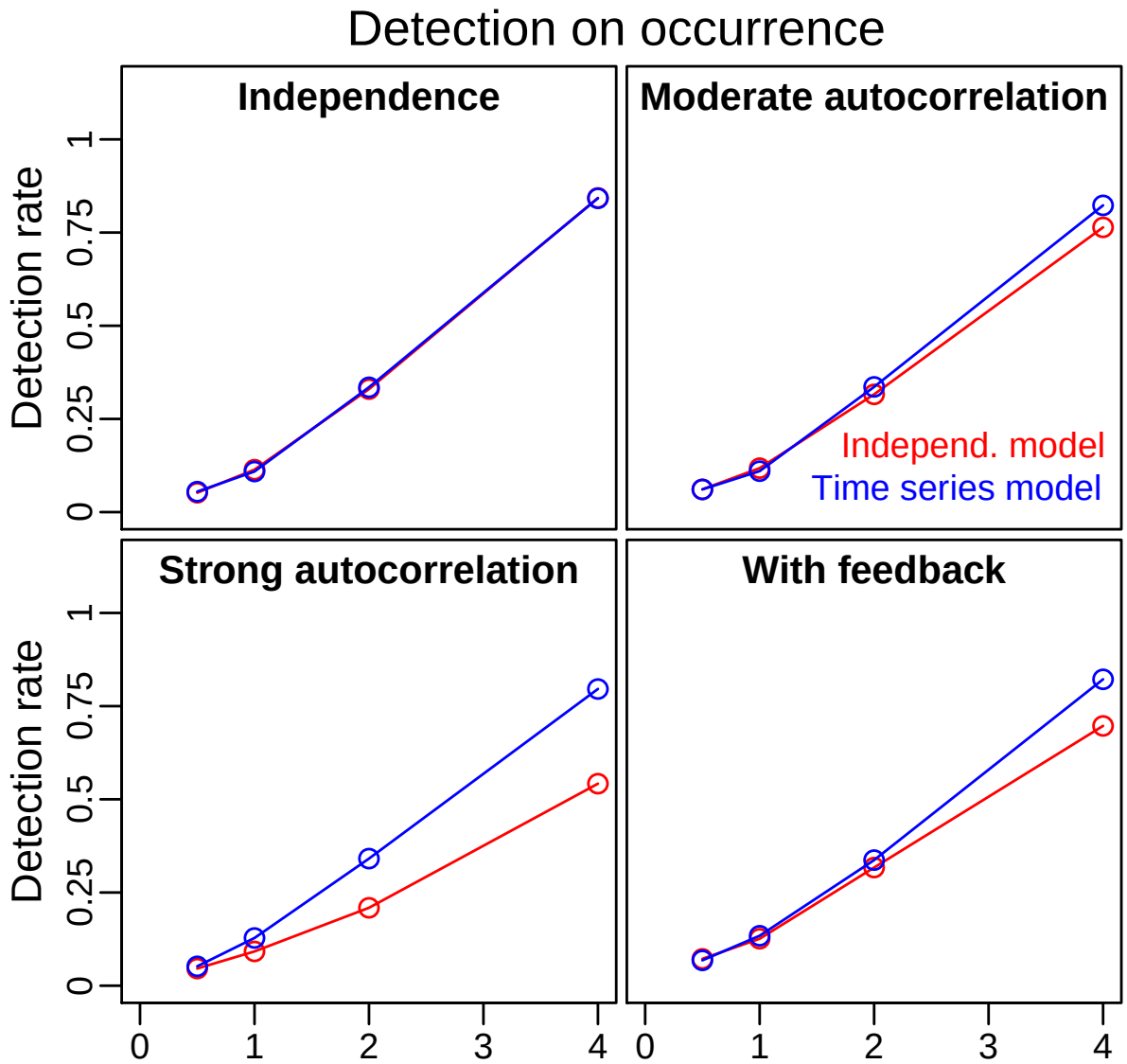


- 1-step-ahead pred. (bold) and one-sided 97.5% PIs (thin)
- independence model does not react to previous observations
- alarm at the same week

Simulation: Study design

- **Model:**
  - logarithmic link, 1st order autocorrelation (with feedback), seasonality, negative binomial distribution
  - marginal mean ≈ 5 and variance ≈ 10
  - different degrees of temporal dependence
- 9 years (468 obs.) set-up phase, 1 year operational phase
- **Outbreak scenario:**
  - one outbreak at random position lasting six weeks
  - additive effect from Poisson distribution with mean proportional to conditional standard deviation
- false alarm rate $\alpha = 2.5\%$ ($\hat{=}$ 73.2% yearly error rate)
- 1000 repetitions
- R packages **`batchJobs`** and **`batchExperiments`** (?) to conduct simulations on computing cluster

Simulation: Outbreak detection rate



- procedures hold false alarm rate
- no performance loss for independent data
- improved immediate detection for dependent data

Outlook

R package **tscount** available on R-Forge:
`http://tscount.r-forge.r-project.org`

Desirable extensions of the package:

- robust estimation *Elsaied & Fried (2014), Kitromilidou & Fokianos (2013)*
- nonlinear specifications *Fokianos & Tjostheim (2012)*
- other conditional distributions like quasi-Poisson *Ver Hoef & Boveng (2007)*
- thinning-based models *Weiß (2008)*
- multivariate and spatio-temporal models

References I

Christou, V. and Fokianos, K. (2014). Quasi-Likelihood Inference for Negative Binomial Time Series Models. *Journal of Time Series Analysis*, 35(1):55–78.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive Model Assessment for Count Data. *Biometrics*, 65(4):1254–1261.

Douc, R., Doukhan, P., and Moulines, E. (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and their Applications*, 123(7):2620–2647.

Dunsmuir, W. T. M. and Scott, D. J. (2015). The glarma Package for Observation-Driven Time Series Regression of Counts. *Journal of Statistical Software*, 67(7).

Dürre, A., Fried, R., and Liboschik, T. (2015). Robust estimation of (partial) autocorrelation. *WIREs Computational Statistics*, 7(3):205–222.

Elsaied, H. and Fried, R. (2014). Robust Fitting of INARCH Models. *Journal of Time Series Analysis*, 35(6):517–535.

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-Valued GARCH Process. *Journal of Time Series Analysis*, 27(6):923–942.

Fokianos, K. (2011). Some recent progress in count time series. *Statistics*, 45(1):49–58.

References II

Fokianos, K. and Fried, R. (2010). Interventions in INGARCH processes. *Journal of Time Series Analysis*, 31(3):210–225.

Fokianos, K. and Fried, R. (2012). Interventions in log-linear Poisson autoregression. *Statistical Modelling*, 12(4):299–322.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson Autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439.

Fokianos, K. and Tjøstheim, D. (2011). Log-linear Poisson autoregression. *Journal of Multivariate Analysis*, 102(3):563–578.

Fokianos, K. and Tjøstheim, D. (2012). Nonlinear Poisson autoregression. *Annals of the Institute of Statistical Mathematics*, 64(6):1205–1225.

Fried, R., Liboschik, T., Elsaied, H., Kitromilidou, S., and Fokianos, K. (2014). On Outliers and Interventions in Count Time Series following GLMs. *Austrian Journal of Statistics*, 43(3):181–193.

Harvey, A. C. and Durbin, J. (1986). The Effects of Seat Belt Legislation on British Road Casualties: A Case Study in Structural Time Series Modelling. *Journal of the Royal Statistical Society. Series A (General)*, 149(3):187–227.

Heinen, A. (2003). Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model. *CORE discussion paper*, 62.

References III

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

Helske, J. (2016). KFAS: Kalman Filter and Smoother for Exponential Family State Space Models. R package version 1.1.2.

Kitromilidou, S. and Fokianos, K. (2015). Robust Estimation Methods for a Class of Count Time Series Log-Linear Models. under revision.

Kourentzes, N. and Petropoulos, F. (2015). tsintermittent: Intermittent Time Series Forecasting. R package version 1.8.

Liboschik, T., Fokianos, K., and Fried, R. (2015). tscount: An R package for analysis of count time series following generalized linear models. *TU Dortmund, SFB 823 Discussion Paper*, 06/15.

Liboschik, T., Kerschke, P., Fokianos, K., and Fried, R. (2016). Modelling interventions in INGARCH processes. *International Journal of Computer Mathematics*, 93(4):640–657.

Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63(19).

Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6:1517–1549.

References IV

Meyer, S., Held, L. and Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software*, 77:11.

R Core Team (2015). R – A Language and Environment for Statistical Computing. http://www.r-project.org.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Siakoulis, V. (2015). acp: Autoregressive Conditional Poisson. R package version 2.0.

Stasinopoulos, D. M., Rigby, R. A., and Eilers, P. (2015). gamlss.util: GAMLSS Utilities. R package version 4.3-2.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Statistics and computing. Springer, New York, 4 edition.

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Wang, S. (2013). pbs: Periodic B Splines. R package version 1.1.

Yang, M., Zamba, G. K., and Cavanaugh, J. E. (2014). ZIM: Zero-Inflated Models for Count Time Series with Excess Zeros. R package version 1.0.2.

References V

Yee, T. W. (2015). VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-0.

Zhu, F. (2012). Zero-inflated Poisson and negative binomial integer-valued GARCH models. *Journal of Statistical Planning and Inference*, 142(4):826–839.