

Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011

Michael Höhle^{1,2} and Matthias an der Heiden³

¹Department of Mathematics, Stockholm University, Sweden

²Federal Institute for Quality Assurance and Transparency in Healthcare, Germany

³Department for Infectious Disease Epidemiology, Robert Koch Institute, Germany

XXVIIIth International Biometric Conference
Victoria, Canada, 12 July 2016

Outline

- 1 Motivation: STEC/HUS Outbreak in Germany 2011
- 2 Nowcasting
 - The problem and some notation
 - A Bayesian approach
- 3 Evaluating the Nowcasts
- 4 Discussion

Outline

- 1 Motivation: STEC/HUS Outbreak in Germany 2011
- 2 Nowcasting
- 3 Evaluating the Nowcasts
- 4 Discussion

STEC/HUS Outbreak in Germany 2011 (1)

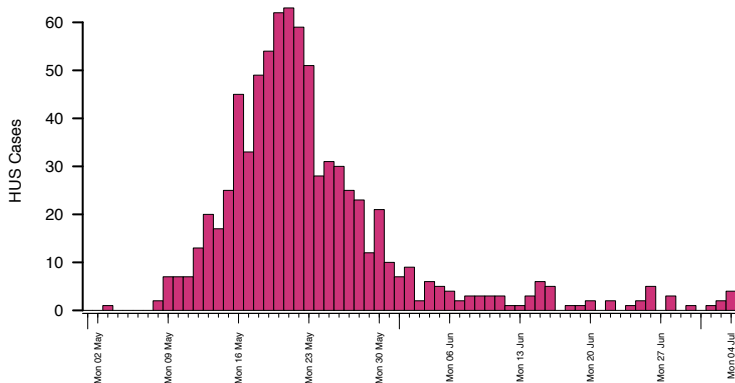
- Outbreak of *E. coli* (STEC) O104:H4 in Germany May–June 2011 associated with sprouts (Frank et al., 2011; Buchholz et al., 2011):

	STEC	HUS
N (% of total)	2987 (78)	855 (22)
Median age (years)	46	42
Female (%)	58	68
Deaths	18	35
Case-fatality-ratio (%)	0.6	4.1

- Hemolytic-uremic syndrome (HUS) is a disease characterized by hemolytic anemia, thrombocytopenia and acute kidney failure.
- HUS can be a complication of an STEC infection.
- Onset of HUS occurs a median of 5 days (IQR: 4–7) days after onset of the STEC related diarrhea.

STEC/HUS Outbreak in Germany 2011 (2)

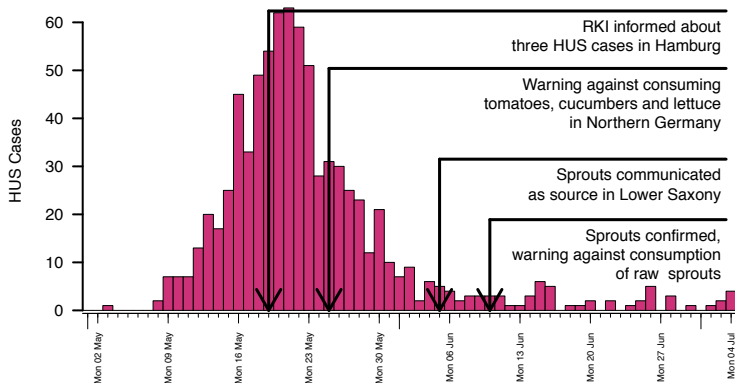
- Retrospective epidemic curve for the hospitalization date of HUS cases (where available, 658 cases) :



- However, *during* the outbreak the situation is not as clear due to reporting delays.

STEC/HUS Outbreak in Germany 2011 (2)

- Retrospective epidemic curve for the hospitalization date of HUS cases (where available, 658 cases) :



- However, *during* the outbreak the situation is not as clear due to reporting delays.

Outline

- 1 Motivation: STEC/HUS Outbreak in Germany 2011
- 2 **Nowcasting**
 - The problem and some notation
 - A Bayesian approach
- 3 Evaluating the Nowcasts
- 4 Discussion

Nowcasting – what's the situation?

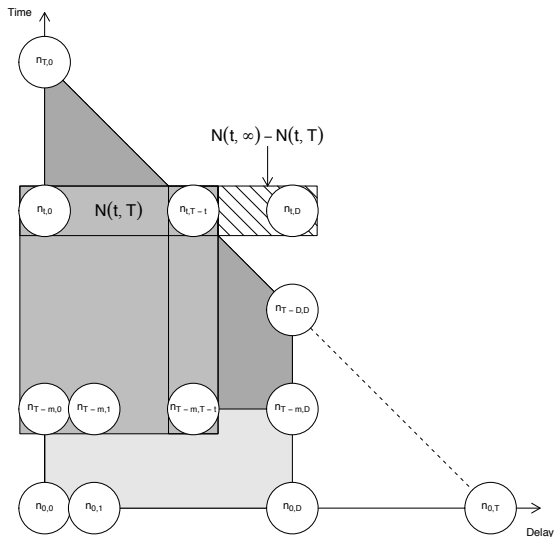
- Opposite to *forecasting*, we just want to know what the situation is “now”, i.e. in an ideal situation without reporting delay.
- The term **nowcasting** is a revival of what has been extensively studied as adjustment for *occurred-but-not-yet-reported* events during the AIDS/HIV epidemic, see e.g. Kalbfleisch and Lawless (1989).
- There is a close connection between nowcasting and *claims reserving* in actuarial sciences.
- Our aim was to assess current epidemic trends during the outbreak in order to judge if the outbreak is ongoing, assess the impact of control measures and perform capacity planning.

Nowcasting Notation (1)

- Let $n_{t,d}$ be the number of cases which occur on day t and become available with a delay of d days, where $t = 0, \dots, T$ – with T being *now* and $d = 0, \dots, D$.
- Problem: $n_{t,d}$ is unknown when $d > T - t$ – see reporting triangle
- $N(t, T) = \sum_{d=0}^{\min(T-t, D)} n_{t,d}$ is the number of cases which occurred on t and who are reported until time T
- Aim of nowcasting: predict the total number of cases, i.e.

$$N(t, \infty) = \sum_{d=0}^{\infty} n_{t,d} = \sum_{d=0}^D n_{t,d}.$$

Nowcasting Notation (2) – Reporting triangle



Frequentist Nowcasting

- What we did during the outbreak was the simple estimate

$$\hat{N}(t, \infty) = \frac{N(t, T)}{\hat{F}(T - t)},$$

where $\hat{F}(\cdot)$ was the naive ECDF estimator of the delay distribution ignoring truncation.

- Lawless (1994) contains a more rigorous treatment using $\hat{F}_{\text{trunc}}(T - t)$ together with an approximate asymptotic normal predictive distribution for $N(t, \infty)$.
- Once the outbreak was over we wanted to compare and improve approaches in order to be better prepared for the **next outbreak**.

Bayesian Nowcasting (1)

- Advantages of using a Bayesian approach for the problem:
 - ▶ natural framework for handling predictive distributions
 - ▶ truncation as a missing data problem → just more parameters
- We used a contingency table setup known from AIDS/HIV modelling to describe the problem:

$$n_{t,d} \sim \text{Po}(\mu_{t,d}), \quad (t, d) \in A_T^m, \quad (1)$$
$$\log(\mu_{t,d}) = \log(\lambda_t) + \log(f_{t,d}).$$

- For homogeneous delay we show that the *generalized Dirichlet distribution* is the conjugate prior for the reversed delay distribution under right-truncated multinomial sampling.
- If, furthermore, $\lambda_t \stackrel{iid}{\sim} \mathcal{Ga}(\alpha_\lambda, \beta_\lambda)$ then fast non-MCMC computation of the predictive posterior for $N(t, \infty)$ is possible.

Bayesian Nowcasting (2)

- Addition: 2nd order TP basis spline to smooth the epidemic curve:

$$\log(\lambda_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{i=1}^L b_i \max(t - \kappa_i, 0)^2, \quad (2)$$

where $(b_1, \dots, b_L)' \sim N(\mathbf{0}, \sigma_b^2 \mathbf{I})$.

- Addition: Time-varying delay distribution using a discrete time hazard model, i.e. one models

$$h_{t,d} = P(\text{delay} = d | \text{delay} \geq d, \mathbf{W}_{t,d}), \quad d = 0, \dots, D$$

for a case occurring at time t by

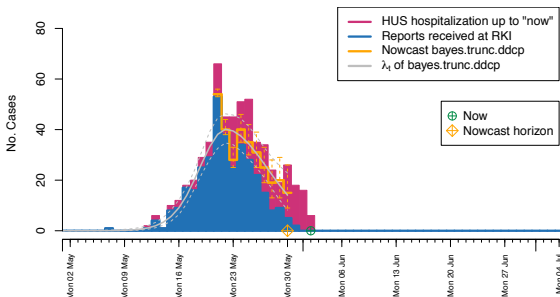
$$\begin{aligned} \text{logit}(h_{t,d}) &= \gamma_d + \mathbf{W}'_{t,d} \boldsymbol{\eta}, \quad d = 0, \dots, D-1 \quad \text{and} \\ h_{t,D} &= 1. \end{aligned} \quad (3)$$

Bayesian Nowcasting (3)

- STEC example: Intervention towards faster reporting on 23 May 2011 is represented by a single change-point, i.e.

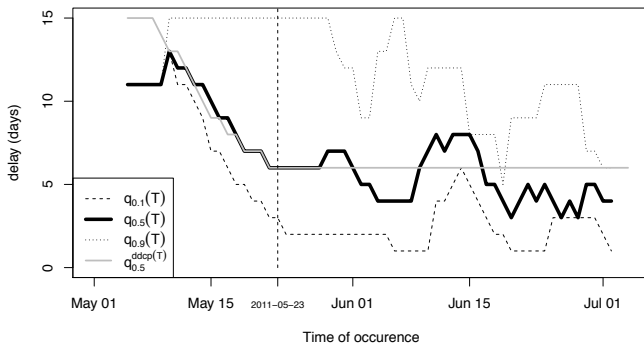
$$W_{t,d} = \mathbb{1}(t + d \geq 2011-05-23).$$

- Inference for the hierarchical Bayesian model consisting of (1), (2) and $f_{t,d}$ resulting from (3) is done using Markov Chain Monte Carlo with JAGS (Plummer, 2003) for each day T :



Results: Nowcasting during the STEC outbreak (1)

- Daily smoothed empirical delay distribution obtained from all hospitalizations occurring within a moving window of $t - 2, \dots, t + 2$.



- One observes a distinct delay reduction due to the intervention on 23 May 2011.

Outline

- 1 Motivation: STEC/HUS Outbreak in Germany 2011
- 2 Nowcasting
- 3 Evaluating the Nowcasts**
- 4 Discussion

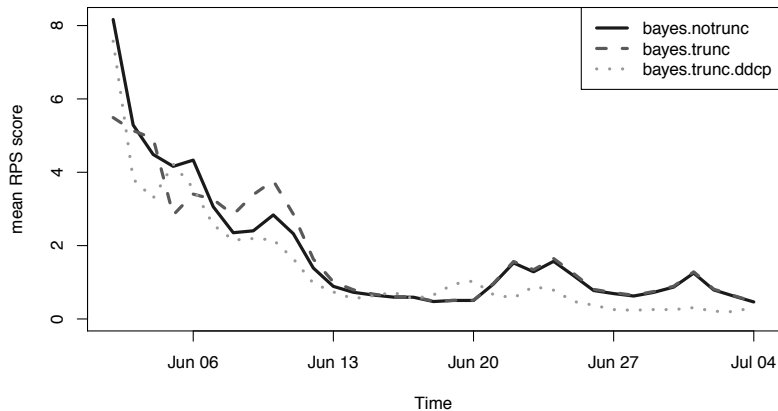
Evaluation of the nowcasts (1)

- Since, retrospectively, the true number of hospitalizations is available we were able to evaluate the predictive quality of different delay-distribution adjustment methods.
- We use *proper scoring rules* for count-data (Czado et al., 2009) evaluating both calibration and sharpness of the probabilistic forecasts.
- Nowcasts are evaluated between 2011-06-02 and 2011-07-04, for each time T we consider the 10 lags $T - 12, \dots, T - 3$.
- The mean overall score of all 330 nowcasts is

	notrunc	trunc	trunc.ddcp	unif
logS	2.28	2.02	Inf	5.69
RPS	1.79	1.77	1.39	95.10
dist.median	2.37	2.51	1.87	143.94
outside.ci	0.07	0.05	0.09	0.84

Evaluation of the nowcasts (2)

Mean RPS score as a function of the nowcasted time points:



Outline

- 1 Motivation: STEC/HUS Outbreak in Germany 2011
- 2 Nowcasting
- 3 Evaluating the Nowcasts
- 4 Discussion**


Discussion (1)

- Ignoring truncation for the specific outbreak was an ad-hoc way to compensate for the fact that, early in the outbreak, delays reduced significantly.
- However, right-truncation adjusted procedures are to be recommended as proper adjustment methods.
- Visualized nowcasts were circulated among the stakeholders at the RKI on a daily basis.
- Retrospective animations after the outbreak were an effective tool for communicating consequences of delays in the German reporting system to political stakeholders.

Discussion (2)

- In order to promote application, the proposed methods and the outbreak data are available in the R package **surveillance** (Salmon et al., 2016) as function `nowcast` and dataset `hus0104Hosp`.
- The contingency table approach has been further extended to facilitate delay adjusted outbreak detection algorithms (Salmon et al., 2015).
- Details of the talk:

Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011 (2014), Höhle M and an der Heiden M. *Biometrics*, 70(4):993–1002.

- For a less serious version of this talk check out the work on *zombie pReparedness* ( [m_hoehle](#)).

Literature I

- Buchholz, U., Bernard, H., Werber, D., Böhmer, M. M., Remschmidt, C., Wilking, H., Deleré, Y., a. d. Heiden, M., Adlhoch, C., Dreesman, J., Ehlers, J., Ethelberg, S., Faber, M., Frank, C., Fricke, G., Greiner, M., Höhle, M., Ivarsson, S., Jark, U., Kirchner, M., Koch, J., Krause, G., Lubert, P., Rosner, B., Stark, K., and Kühne, M. (2011). German outbreak of *Escherichia coli* O104:H4 associated with sprouts. *New England Journal of Medicine*, 365:1763–1770.
- Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65:1254–1261.
- Frank, C., Werber, D., Cramer, J. P., Askar, M., Faber, M., an der Heiden, M., Bernard, H., Fruth, A., Prager, R., Spode, A., Wadl, M., Zoufaly, A., Jordan, S., Kemper, M. J., Follin, P., Müller, L., King, L. A., Rosner, B., Buchholz, U., Stark, K., and Krause, G. (2011). Epidemic profile of shiga-toxin–producing *Escherichia coli* O104:H4 outbreak in Germany. *New England Journal of Medicine*, 365(19):1771–1780.
- Höhle, M. and an der Heiden, M. (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002. Animations available from <http://dx.doi.org/10.1111/biom.12194>.
- Kalbfleisch, J. D. and Lawless, J. F. (1989). Inference based on retrospective ascertainment: An analysis of the data on transfusion related AIDS. *Journal of the American Statistical Association*, 84(406):360–372.
- Lawless, J. F. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *The Canadian Journal of Statistics*, 22(1):15–31.

Literature II

- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.
- Salmon, M., Schumacher, D., and Höhle, M. (2016). Monitoring count time series in R: Aberration detection in public health surveillance. *Journal of Statistical Software*, 70(10). Also available as vignette of the R package surveillance.
- Salmon, M., Schumacher, D., Stark, K., and Höhle, M. (2015). Bayesian outbreak detection in the presence of reporting delays. *Biometrical Journal*, 57(6):1051–1067. <http://dx.doi.org/10.1002/bimj.201400159>.

Generalized Dirichlet Distribution

- The density of the $GD(\boldsymbol{\alpha}, \boldsymbol{\beta})$ distribution with parameters $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_{D-1})'$ and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{D-1})'$ is proportional to

$$f(\text{rev}(\mathbf{p})) \propto \prod_{i=0}^{D-1} p_{D-i}^{\alpha_i-1} (1 - p_D - \dots - p_{D-i})^{\gamma_i},$$

where the probabilities p_D, \dots, p_1 and $p_0 = 1 - p_D - \dots - p_1$ are all non-negative and sum to one.

- Here, $\gamma_i = \beta_i - \alpha_{i+1} - \beta_{i+1}$ for $i = 0, \dots, D - 2$ and $\gamma_{D-1} = \beta_{D-1} - 1$.
- In the above, $\text{rev}(\cdot)$ denotes the reverse delay order, i.e. where delay d is mapped to the new scale $D - d$ in analogy with the reverse time hazard function.

Delay distribution model

- The corresponding delay distribution can be computed from the $h_{t,d}$'s in (3) on slide 13 as

$$f_{t,d} = \left(1 - \sum_{i=0}^{d-1} f_{t,i} \right) h_{t,d}, \quad d = 0, \dots, D.$$

Model details: Priors

TP basis spline setup used in H. and an der Heiden (2014):

$$\begin{aligned}\beta &\sim \text{dmnorm}(\text{beta.mu}, \text{beta.prec}) \\ 1/\sigma_b^2 &\sim \text{dgamma}(0.001, 0.001) \\ \gamma_d &\sim \text{dnorm}(\text{mu.gamma}[d], \text{tau.gamma}[d]),\end{aligned}$$

with μ_γ and τ_γ s.t. a uniform distribution is obtained for the $f_{t,d}$'s

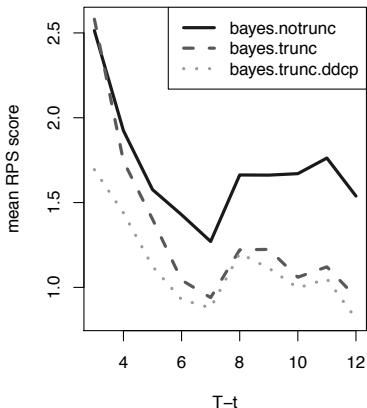
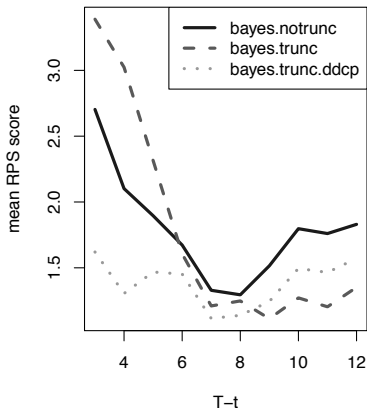
$$\eta \sim \text{dmnorm}(\text{eta.mu}, \text{eta.prec}),$$

with $\text{eta.mu} = \mathbf{0}$ and $\text{eta.prec} = \text{diag}(1)$.

Model details: Knots

- Knots are placed at $\{\kappa_1, \dots, \kappa_L\}$
- We place $\min(\lfloor T/6 \rfloor, 20)$ knots evenly between time zero and T .
- In order to avoid over-extrapolations at the end, where uncertainty is large, we remove any knots which are between time $T - D/2$ and T .

Evaluation of the nowcasts (3)



Left pane: RPS score as function of $T - t$ for the outbreak data.

Right pane: Same but now for simulated data based on the true epidemic curve and simulated homogeneous delays.

Scoring Rules

- *logarithmic score* (logS)

$$\log S(P_t^T, N(t, \infty)) = -\log(f_{P_t^T}(N(t, \infty))),$$

- *ranked probability score* (RPS)

$$\text{RPS}(P_t^T, N(t, \infty)) = \sum_{k=0}^{N_{\max}} \left(F_t^T(N(t, \infty)) - \mathbb{1}(N(t, \infty) \leq k) \right)^2$$