Supplementary material to "Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing"

Jan van de Kassteele, Paul H.C. Eilers, Jacco Wallinga

eMethods

Constrained P-spline smoothing

In order to obtain stable estimates of both the smooth surface and the day-of-the-week effect, we include prior information on the reporting process as additional constraints: the surface is unimodal in the reporting delay dimension, is (nearly) zero at the predefined maximum delay, and has a presumed shape at the beginning of the outbreak. Furthermore, the regression coefficients $\boldsymbol{\beta}$ are regularized to avoid extreme estimates in a sparse data setting.

We show how these constraints are constructed an how they are applied as penalizations on the Negative Binomial log-likelihood function $\ell(\mathbf{n} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$.

Constraint 1: controlling the roughness in two dimensions

If there are no abrupt changes in the reporting process, we expect the true reporting intensity to be a smooth surface in time of symptoms onset and in the reporting delay dimension. We use this information to extrapolate the smooth surface outside the reporting trapezoid.

The number of the univariate B-spline basis functions K_T and K_D should be sufficiently large to capture all variations in the trend surface. To prevent overfitting as the number basis functions, and therefore the number of coefficients, increases, we regularize the estimation of the unknown coefficients $\boldsymbol{\alpha}$ with a roughness penalty. These penalized B-splines are called P-splines[1] which can straightforwardly be extended in into two dimensions[2]. The P-spline method uses B-splines as the basis for the regression and modifies the log-likelihood by a difference penalty on the regression coefficients. For computational reasons, usually a quadratic penalty is taken. The penalized Negative Binomial log-likelihood function for the bivariate smoothing then becomes

$$\ell^{*} = \ell(\mathbf{n} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta) - \tfrac{1}{2}\lambda_T \boldsymbol{\alpha}' \mathbf{D}_T' \mathbf{D}_T \boldsymbol{\alpha} - \tfrac{1}{2}\lambda_D \boldsymbol{\alpha}' \mathbf{D}_D' \mathbf{D}_D \boldsymbol{\alpha}. \qquad (1)$$

Matrices $\mathbf{D}_T$ and $\mathbf{D}_D$ are difference operator matrices working on vector $\boldsymbol{\alpha}$ in the direction of the time of symptoms onset and direction of the reporting delay, respectively. They are obtained by first calculating mth order difference operator matrices $\mathbf{D}_T^{(m_T)}$ and $\mathbf{D}_D^{(m_D)}$, having dimensions $(K_T - m_T) \times K_T$ and $(K_D - m_D) \times K_D$ respectively, corresponding to the regression coefficients of the univariate B-splines bases, and then expand these matrices using Kronecker products:

$$\mathbf{D}_T = \mathbf{I}_{K_D} \otimes \mathbf{D}_T^{(m_T)} \text{ and } \mathbf{D}_D = \mathbf{D}_D^{(m_D)} \otimes \mathbf{I}_{K_T}. \tag{2}$$

Hence, $\mathbf{D}_T$ and $\mathbf{D}_D$ get dimensions $(K_T - m_T)K_D \times K_T K_D$ and $(K_D - m_D)K_T \times K_T K_D$ respectively.

An important feature of our method is that extrapolation is a natural consequence of the smoothing process. The choice of the difference order is critical here, since it determines the form of the extrapolation: by penalizing the first order differences of adjacent coefficients, a trend is extrapolated as a constant, while penalizing second order differences, a trend is extrapolated linearly[3]. For example, taking $m_T = m_D = 2$, our default choice, our method allows linear extrapolation (on a log-scale) of possible trends in both the time of symptoms onset (epidemic trends) and reporting delay (shape of the reporting delay distribution).

Finally, the roughness is controlled by the smoothing parameters λ_T and λ_D, which can be found by minimizing an information criterion. See the section on parameter estimation.

Constraint 2: unimodal distribution of reporting delay times

Because the surface can become unstable when information about the true number of reported symptomatic cases is missing, extrapolation is challenging. However, we have prior information on the reporting process. We know it has low intensities at both short delays and long delays, and shows a maximum in the reporting intensity somewhere in between. We may therefore assume that the intensity of the reporting process is unimodal in the reporting delay dimension and can expect that this unimodality will be valid outside the reporting trapezoid.

Mathematically, a unimodality of the reporting delay distribution is reached by forcing a log-concave shape of the smooth surface in the reporting delay dimension. This constraint can be enforced by introducing asymmetric penalties[4,5]. To ensure concavity on a log-scale we only allow negative second order differences of $\boldsymbol{\alpha}$ in the reporting delay dimension. The modified penalized log-likelihood function now becomes

$$\ell^{**} = \ell^{*} - \kappa_u \boldsymbol{\alpha}' \mathbf{D}_u' \mathbf{V}_u \mathbf{D}_u \boldsymbol{\alpha}. \tag{3}$$

The penalty looks very much like the roughness penalty in equation (1). Matrix $\mathbf{D}_u = \mathbf{D}_u^{(2)} \otimes \mathbf{I}_{K_T}$ is identical to the one in equation (2); a second order difference operator matrix working on $\boldsymbol{\alpha}$ in the reporting delay dimension. The introduction of weight matrix $\mathbf{V}_u = \text{diag}(\mathbf{v}_u)$ enforces the unimodality by

$$\mathbf{v}_u = \begin{cases} 1 & \text{if} \quad \mathbf{D}_u \boldsymbol{\alpha} \geq 0 \\ 0 & \text{if} \quad \mathbf{D}_u \boldsymbol{\alpha} < 0 \end{cases}. \tag{4}$$

Typically $\kappa_u = 10^6$, which results in a very high log-likelihood if the surface is not concave in the reporting delay dimension. This ensures concavity. In turn, the log-link function ensures log-concavity and therefore unimodality.

Constraint 3: boundary constraints

The prior information on the reporting process can further be utilized in the form of boundary constraints. First, we know that the reporting intensity should go to (approach) zero at the maximum delay, because almost all cases should have been reported by then. Second, given the type of infection and the health reporting system, we already have an idea of what the delay mechanism will be. For example, for Influenza we expect delays in the order of days.

A rough guess of a prior distribution for the reporting delay is sufficient to construct the boundary constraints. For example, from historical data on measles we know that the average reporting delay is approximately 10 days and almost all cases, say 99%, get reported within six weeks (42 days)[6]. Furthermore, at $t = 1$, day one of an outbreak, we can assume that at most one case can be expected. This constraint is important to obtain stable estimates of the trend surface, as only few observations are available. On the other hand, in case we are dealing with endemic data, more cases can be expected at $t = 1$ and this constrained becomes less important as more information is available. Without having to know the exact numbers and under the assumption that reporting delay follows a Negative Binomial distribution (not to be confused with the distribution for $n_{t,d}$), these numbers entirely define the value of $\mu_{1,d}$ for $d = 0, \dots, D$. At $d = D$, the maximum delay, we set $\mu_{t,D} = \mu_{1,D}$ for $t = 1, \dots, T$. The smooth surface is not allowed to be larger than these values. Mathematically, these boundary constraints can again be enforced with asymmetric penalties on $\boldsymbol{\alpha}$.

Let $\mathbf{g}$ be a vector of length $T(D + 1)$ with the predefined maximum log-reporting intensities. We then must ensure that $\mathbf{B}\boldsymbol{\alpha} < \mathbf{g}$ at the locations where this constraint should be applied. We can now write the modified penalized log-likelihood function as

$$\ell^{***} = \ell^{**} - \tfrac{1}{2}\kappa_b(\mathbf{B}\boldsymbol{\alpha} - \mathbf{g})'\mathbf{V}_b(\mathbf{B}\boldsymbol{\alpha} - \mathbf{g}), \tag{5}$$

where $\mathbf{V}_b = \mathrm{diag}(\mathbf{b}\mathbf{v}_b)$ is a matrix with asymmetric weights specified as

$$\mathbf{v}_b = \begin{cases} 1 & \text{if} \quad \mathbf{B}\boldsymbol{\alpha} \geq \mathbf{g} \\ 0 & \text{if} \quad \mathbf{B}\boldsymbol{\alpha} < \mathbf{g}' \end{cases} \tag{6}$$

and $\mathbf{b}$ is a vector of length $T(D + 1)$ with elements equal to 1 at the locations where the constraint should be applied, and 0 otherwise. Typically, $\kappa_b = 10^6$. A large penalty will be activated at all locations where $\mathbf{B}\boldsymbol{\alpha} \geq \mathbf{g} \wedge \mathbf{b} = 1$. This will ensure that the smooth surface remains below the pre-specified intensities at the given locations, here at $t = 1 \wedge d = 0, \dots, D$ and $t = 1, \dots, T \wedge d = D$.

Constraint 4: avoiding extreme parameter estimates

The final prior information is that very large values are impossible: the smooth surface and the size the day-of-the-week effects cannot be infinitely large or small. At the beginning of an outbreak, only little information on the day-of-the-week effect is provided by the data, which may result in large estimates. To obtain finite estimates of the day-of-the-week effects $\boldsymbol{\beta}$, we add a ridge penalty with a small fixed parameter $\kappa_w = 0.01$. Furthermore, to make the estimation of the smooth surface numerically stable, a ridge penalty with a very small fixed smoothing parameter $\kappa_s = 10^{-6}$ is added to the log-likelihood. The modified penalized log-likelihood function then becomes

$$\ell^{****} = \ell^{***} - \tfrac{1}{2}\kappa_w \boldsymbol{\beta}'\boldsymbol{\beta} - \tfrac{1}{2}\kappa_s \boldsymbol{\alpha}'\boldsymbol{\alpha}. \qquad (7)$$

Parameter estimation

The smooth surface and the day-of-the-week effects are estimated simultaneously. We can write our method as a penalized generalized linear model, with a Negative Binomial error distribution, a log-link function, a model matrix and, additionally, a penalty matrix. We can therefore use the penalized version of the iterative weighted least squares (PIWLS) algorithm. Given the smoothing parameters λ_T and λ_D and overdispersion parameter θ in an outer loop, the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are found in an inner loop by iteratively solving the system

$$(\mathbf{U}'\mathbf{W}\mathbf{U} + \mathbf{P})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} = \mathbf{U}'\mathbf{W}\mathbf{z} + \begin{pmatrix}\kappa_b \boldsymbol{B}'\mathbf{V}_b\mathbf{g}\\ \mathbf{0}_{k_w}\end{pmatrix}. \qquad (8)$$

Here $\mathbf{U} = [\mathbf{B}|\mathbf{X}]$ is the combined model matrix of the smooth surface and the day-of-the-week effects. $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1}(\mathbf{n} - \boldsymbol{\mu})$ is the working variable and $\mathbf{W} = \mathrm{diag}(\mathbf{rw})$ is the weight matrix. For the Negative Binomial likelihood, the weight vector is given by $\mathbf{w} = \boldsymbol{\mu}^2/(\boldsymbol{\mu} + \boldsymbol{\mu}^2/\theta)$.

We introduce additional weights $r_{t,d}$ ($\mathbf{r}$ in vector notation), which take the value 1 if $t \leq T - d$, i.e., the element lies within the reporting trapezoid, and 0 otherwise. The zero weights disable contributions to the log-likelihood function outside the reporting trapezoid, but the smoothness penalty automatically generates predictions there.

Because we have separated the smooth surface from the day-of-the-week effects, the penalty matrix $\mathbf{P}$ is a block diagonal matrix, given by

$$\mathbf{P} = \mathrm{blockdiag}(\lambda_T \mathbf{D}_T'\mathbf{D}_T + \lambda_D \mathbf{D}_D'\mathbf{D}_D + \kappa_u \mathbf{D}_u'\mathbf{V}_u\mathbf{D}_u + \ \kappa_b \boldsymbol{B}'\mathbf{V}_b\boldsymbol{B} + \kappa_s \mathbf{I}_{K_T K_D}, \kappa_w \mathbf{I}_{K_W}). \qquad (9)$$

The overdispersion parameter θ is found by maximizing the log-likelihood given the current estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Smoothing parameters λ_T and λ_D are found by minimizing the Bayesian information criterion (BIC) using a greedy grid search algorithm[7]. The BIC is given by

$$BIC = -2\ell^{****} + edf \log\left(\sum_{t=1}^{T} \sum_{d=0}^{D} r_{t,d}\right), \tag{10}$$

where effective degrees of freedom edf[8] is obtained by

$$edf = \mathrm{trace}[(\mathbf{U}'\mathbf{W}\mathbf{U} + \mathbf{P})^{-1}\mathbf{U}'\mathbf{W}\mathbf{U}]. \tag{11}$$

Alternatively, the Akaike information criterion (AIC) could be used, but there is evidence that the AIC tends to undersmooth the data[9].

References

1. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci*. 1996;11(2):89-121. doi:10.1214/ss/1038425655

2. Eilers PHC, Marx BD. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemom Intell Lab Syst*. 2003;66(2):159-174. doi:10.1016/S0169-7439(03)00029-7

3. Fahrmeir L, Kneib T, Lang S, Marx B. *Regression; Models, Methods and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013.

4. Eilers PHC. Unimodal smoothing. *J Chemom*. 2005;19(5-7):317-328. doi:10.1002/cem.935

5. Hofner B, Kneib T, Hothorn T. A unified framework of constrained regression. *Stat Comput*. 2016;26(1-2):1-14. doi:10.1007/s11222-014-9520-y

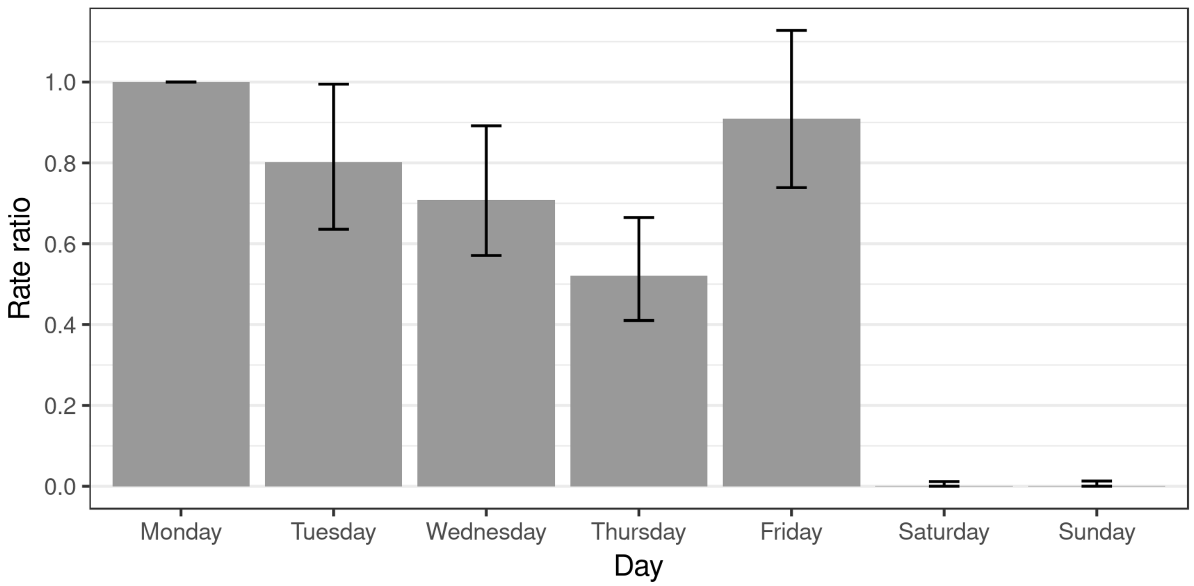
6. Marinović AB, Swaan C, van Steenbergen J, Kretzschmar M. Quantifying Reporting Timeliness to Improve Outbreak Control. *Emerg Infect Dis*. 2015;21(2):209-216. doi:10.3201/eid2102.130504

7. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to Algorithms, Third Edition*. 3rd ed. The MIT Press; 2009.

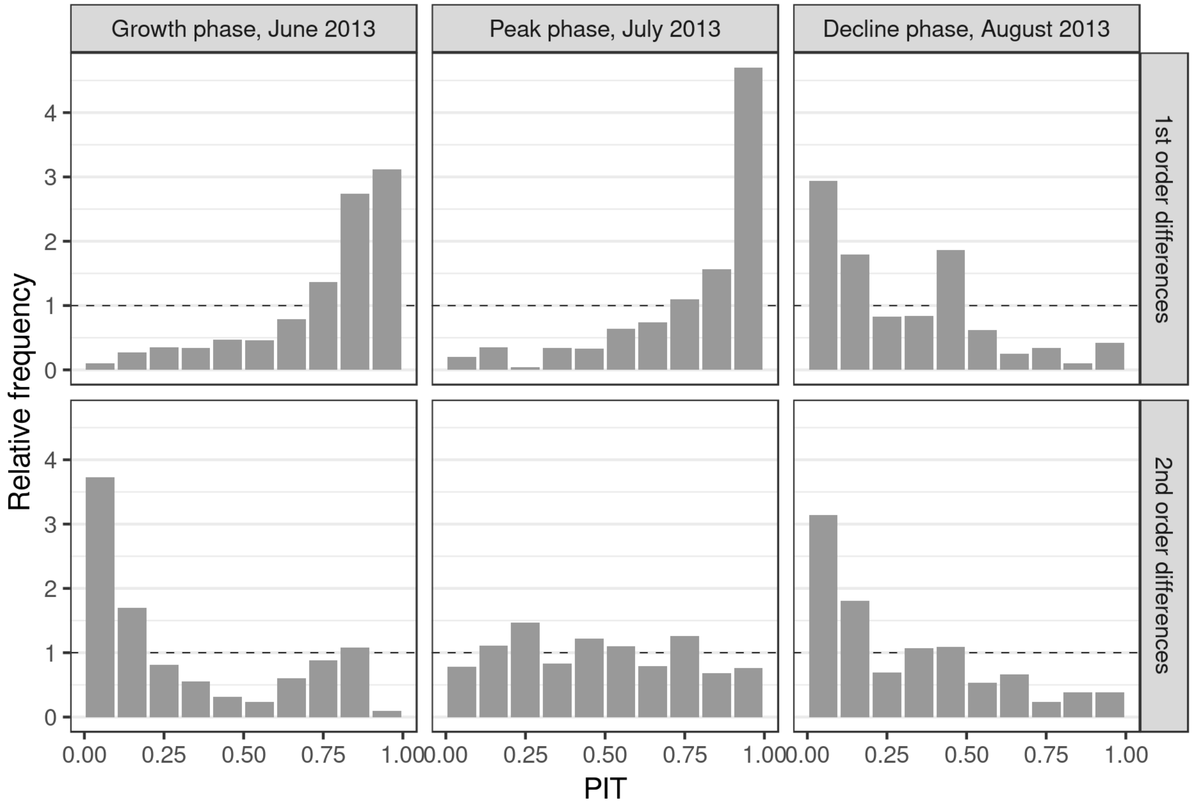
8. Eilers PHC, Marx BD, Durbán M. Twenty years of P-splines. *SORT-Stat Oper Res Trans*. 2015;39(2):149-186.

9. Currie ID, Durban M, Eilers PH. Smoothing and forecasting mortality rates. *Stat Model*. 2004;4(4):279-298. doi:10.1191/1471082X04st080oa

eFigures



eFigure 1. Day-of-the-week effects expressed as rate ratios including 95% confidence intervals. Nowcast date is August 10, 2013. Monday is taken as the reference day (RR = 1).



eFigure 2. PIT histograms showing the performance of nowcasts during the Measles outbreak. Columns: growth phase, peak phase and decline phase of the outbreak. Rows: penalization of the 1st order and 2nd order differences on the adjacent coefficients in the time of symptoms onset dimension.