International Journal of
Environmental Research and Public Health

MDPI

Addendum

Addendum: Wang et al. A Bayesian Approach to Real-Time Monitoring and Forecasting of Chinese Foodborne Diseases. *Int. J. Environ. Res. Public Health*, 2018, 15(8):1740; doi:10.3390/ijerph15081740

Xueli Wang [1], Moqin Zhou [1], Jinzhu Jia [2], Zhi Geng [3] and Gexin Xiao [4,*]

[1] School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, China; wangxl@bupt.edu.cn (X.W.); interpreter_q@hotmail.com (M.Z.)
[2] School of Public Health, Center of Statistical Science, Peking University, Beijing 100871, China; jzjia@math.pku.edu.cn
[3] School of Mathematical Sciences, Center of Statistical Science, Peking University, Beijing 100871, China; zhigeng@pku.edu.cn
[4] China National Center for Food Safety Risk Assessment, Beijing 100022, China
* Correspondence: xiaogexin@cfsa.net.cn

Received: 16 April 2019; Accepted: 16 April 2019; Published: 23 April 2019

The authors wish to update the Abstract and Section 3 in their paper published in the *International Journal of Environmental Research and Public Health (IJERPH)* [1].

They would like to rewrite the abstract as follows:

Abstract: Foodborne diseases have a big impact on public health and are often underreported. This is because a lot of patients delay treatment when they suffer from foodborne diseases. In Hunan Province (China), a total of 21,226 confirmed foodborne disease cases were reported from 1 March 2015 to 28 February 2016 by the Foodborne Surveillance Database (FSD) of the China National Centre for Food Safety Risk Assessment (CFSA). The purpose of this study was to make use of the daily number of visiting patients to forecast the daily true number of patients. Our main contribution is that we take the reporting delays into consideration and apply a Bayesian hierarchical model for this forecast problem. The data shows that there were 21,226 confirmed cases reported among 21,866 visiting patients, a proportion as high as 97%. Given this observation, the Bayesian hierarchical model was established to predict the daily true number of patients using the number of visiting patients. We use several scoring rules to assess the performance of different nowcasting procedures. We conclude that Bayesian nowcasting with consideration of right truncation of the reporting delays has a good performance for short-term forecasting and could effectively predict the epidemic trends of foodborne diseases. Meanwhile, this approach could provide a methodological basis for future foodborne disease monitoring and control strategies, which are crucial for public health.

In the end of the first paragraph in Section 3, the authors would like to update the last two sentences of the paragraph and add some citations. The revised sentences are as follows:

In this paper, we apply a Bayesian nowcasting model proposed by Höhle and an der Heiden [2] to forecast the daily total number of cases. Thanks to Salmon et al. [3], who provided a convenient R package "surveillance", the inference for the model could be easily performed. The R package surveillance also contains a few other nowcasting methods that we also tried and did comparisons with using the scoring rules implemented in the package. The results are shown in Section 4. Below we review the model.

The authors revised Section 3.2 to describe the inference approach proposed by Höhle and an der Heiden (in Section 3.2 of [2]) in some greater detail. This part is now as follows:

For the convenience of the reader, we describe the inference approach proposed by Höhle and an der Heiden (in Section 3.2 of [2]) in some greater detail.

Define p_d as the (time-homogeneous) probability that a case will have a reporting delay of d days. The p_d's satisfy the following equation: $\sum_{d=0}^{D} p_d = 1$. Following Kalbfleisch and Lawless's previous work [4] and Zeger et al.'s previous work [5], we assume that the occurrence time of cases follows an underlying inhomogeneous Poisson process. A reasonable data generating process for the daily number of cases is thus as follows:

$$\begin{aligned} N_t|\lambda_t &\sim \mathrm{Po}(\lambda_t) \\ (n_{t,D}, n_{t,D-1}, \cdots, n_{t,0})'|N_t, \mathbf{p} &\sim \mathrm{MN}(N_t, (p_D, p_{D-1}, \cdots, p_0)') \end{aligned}$$

where $\mathrm{Po}(\lambda)$ denotes the Poisson distribution with expectation $\lambda > 0$ and $\mathrm{MN}(N, \mathbf{p})$ denotes the multinomial distribution with size parameter N and probability vector $\mathbf{p}$. Nowcasting for a given time T can thus be divided into steps of determining the λ_t's, estimating the unknown delay distribution (i.e., the p_d's), and finally predicting the unobserved $n_{t,d}$'s in order to compute the total N_t. As T increases, and if the assumption about a time-homogeneous delay distribution is acceptable, the available data make it possible to estimate the delay distribution better and better, and hence the quality of the predictions near T improves with time.

Consider a fixed time T and define $\mathbf{p}_T = (p_{T,D}, p_{T,D-1}, \cdots, p_{T,1})'$ as the probability vector denoting that a case is reported with a delay of d days given the observed incomplete information at time T, i.e., the set of $n_{t,d}$, where $t + d \leq T$. We choose as prior distribution the generalised Dirichlet distribution $GD(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with fixed constants $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_D)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_D)'$. Now we use Property 3 in the Web appendix of [2] that shows that the posterior of $\mathbf{p}$ under right-truncated multinomial sampling is again a GD distribution with parameters $\boldsymbol{\alpha}^*_T, \boldsymbol{\beta}^*_T$ given by

$$\begin{aligned} \alpha^*_{T,i} &= \alpha_i + \sum_{\tau=0}^{T-D+i} n_{\tau,D-i} \\ \beta^*_{T,i} &= \beta_i + \sum_{\tau=0}^{T-D+i} (N_{\tau,\tau+D-i} - n_{\tau,D-i}), \qquad i \in \{0, \cdots, D-1\}. \end{aligned}$$

hence, for a given T we can assume the following model hierarchy for the time points $t \in \{T - D, \cdots, T\}$:

$$\begin{aligned} \mathbf{p}_T &\sim GD\left(\boldsymbol{\alpha}^*_T, \boldsymbol{\beta}^*_T\right) \\ \lambda_t &\sim Ga(a_\lambda, b_\lambda) \\ N_t|\lambda_t &\sim Po(\lambda_t) \\ N_{t,T}|N_t, \mathbf{p}_T &\sim \mathrm{Bin}(N_t, q_{T,T-t}), \end{aligned}$$

where $q_{T,d} = \sum_{\delta=0}^{d} p_{T,\delta}$ is the proportion reported within a delay of d days. We denote by $Ga(a_\lambda, b_\lambda)$ the gamma distribution with parameters $a_\lambda > 0, b_\lambda > 0$. For this hierarchical model, the marginal distribution of N_t is a negative binomial distribution with the following mean and variance:

$$\begin{aligned} E(N_t) &= \mu_\lambda = a_\lambda b_\lambda \\ \mathrm{Var}(N_t) &= \mu_\lambda + \frac{\mu^2_\lambda}{a_\lambda}. \end{aligned}$$

To estimate N_t given the observed counts $n_{t,d}$ at time T, we have to perform two steps: (1) update the delay distribution $\mathbf{q}_T$ and (2) update the prediction for N_t:

(1) For the given T we compute α^*_T, β^*_T as stated above. We then draw for $k = 1, \ldots, K$ random vectors $\mathbf{p}_T^{(k)} \sim GD\left(\boldsymbol{\alpha}^*{}_T, \boldsymbol{\beta}^*_T\right)$ by the algorithm of Wong [6] and calculate

$$q_{T,d}^{(k)} = \sum_{\delta=0}^{d} p_{T,\delta}^{(k)}.$$

(2) Given the updated delay distribution $\mathbf{q}_T^{(k)}$ and the observed counts $n_{t,d}$, we can now update the prediction of $N_t, t = T - D, \cdots, T$. For $n \in \{0, 1, 2, \cdots\}$ we approximate by Monte Carlo sampling

$$f(N_t = n | N_{t,T}) \approx \frac{1}{K} \sum_{k=1}^{K} f_{n,t}^{(k)}, \qquad t \in \{T - D, \cdots, T\}$$

An application of Bayes theorem provides $f_{n,t}^{(k)} = \tilde{f}_{n,t}^{(k)} / c_t^{(k)}$, where $c_t^{(k)} = \sum_{n=0}^{\infty} \tilde{f}_{n,t}^{(k)}$ is the normalization constant and

$$\tilde{f}_{n,t}^{(k)} = f(N_{t,T} | N_t = n, q_{T,T-t}^{(k)}) f(N_t = n | \lambda_t) f(\lambda_t)$$

for all $t \in \{T - D, \cdots, T\}$. The factors of the last equation can be evaluated using the distributional assumptions of the model hierarchy. For numerical convenience we do not sum over the entire support $\{0, 1, 2, \cdots\}$ to get the normalization, but instead approximate

$$c_t^{(k)} \approx \sum_{n=0}^{N_{\max}} \tilde{f}_{n,t}^{(k)},$$

where $N_{\max}$ is chosen sufficiently large.

Finally, the authors would like to update "proposed in this paper" to "described in this paper".

The changes do not affect the results. The manuscript will be updated and the original will remain online on the article webpage, with a reference to this addendum.

References

1. Wang, X.; Zhou, M.; Jia, J.; Geng, Z.; Xiao, G. A Bayesian Approach to Real-Time Monitoring and Forecasting of Chinese Foodborne Diseases. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1740. [CrossRef] [PubMed]
2. Höhle, M.; An der Heiden, M. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics* **2014**, *70*, 993–1002. [CrossRef] [PubMed]
3. Salmon, M.; Schumacher, D.; Höhle, M. Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance. *J. Stat. Softw.* **2016**, *70*, 1–35. [CrossRef]
4. Kalbfleisch, J.D.; Lawless, J.F. Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS. *J. Am. Stat. Assoc.* **1989**, *84*, 360–372. [CrossRef]
5. Zeger, S.L.; See, L.C.; Diggle, P.J. Statistical methods for monitoring the AIDS epidemic. *Stat. Med.* **1989**, *8*, 3–21. [CrossRef] [PubMed]
6. Wong, T.T. Generalized Dirichlet distribution in Bayesian analysis. *Appl. Math. Comput.* **1998**, *97*, 165–181. [CrossRef]