Author’s response to reviews

Title: Online detection and quantification of epidemics

Authors:

Camille Pelat (pelat@u707.jussieu.fr)
Pierre-Yves Boelle (boelle@u707.jussieu.fr)
Benjamin J Cowling (bcowling@hku.hk)
Fabrice Carrat (carrat@u707.jussieu.fr)
Antoine Flahault (flahault@u707.jussieu.fr)
Severine Ansart (severine.ansart@chu-brest.fr)
Alain-Jacques Valleron (valleron@u707.jussieu.fr)

Version: 2 Date: 6 September 2007

Author’s response to reviews: see over
Dear Dr Da-Silva,

Thank you for the very helpful comments from the reviewers, which have helped us to improve our manuscript. Please find attached our revised manuscript “Online detection and quantification of epidemics” which has been modified in accordance with the reviewers’ comments. In the following pages we have addressed each of the points raised by the reviewers.

We have, as requested, included a “Competing interests” section just before the Author’s contributions. We have also ensured that our manuscript conformed to the journal style.

Thank you again for the detailed reviews of our work and for the opportunity to re-submit.

On behalf of all authors,

Sincerely,

Camille Pelat
Reviewer: Paola Sebastiani

Reviewer’s report:

1- HMMs have been proposed as a better way to automatically distinguish between epidemics and non-epidemics phases. See for example
At least this alternative procedure to manually detect epidemic phases should be mentioned.

We agree with the reviewer that other methods have been found to outperform periodic regression models for epidemic detection in some settings. On the other hand, periodic regression models are arguably the most commonly used models for disease surveillance and the same models can also be used to evaluate the morbidity/mortality excess by comparison to a baseline. Accordingly, these lines were added at the end of the second paragraph of the Discussion:

“Alternative detection methods exist that do not rely on the hypothesis of a seasonal baseline. For instance, Hidden Markov Models assume that the observations are generated from a finite mixture of distributions governed by an underlying Markov chain.[24, 25] These methods have shown good aptitude in distinguishing epidemic and non-epidemic phases in seasonal and non-seasonal time series. Another alternative is control-chart methods, which may be calibrated on data from recent months rather from previous years.[26]”
2-a) Some more description of the format of the input file would be helpful.

We have provided a more detailed description of the format required for personal data files in the first paragraph of the “Results” section:

“Users may input their own dataset (eg incidences, mortalities, medication sales) as a plain text file (ie ASCII file) containing the time series as a single column, ie the values are separated by a carriage return. Observations must be aggregated by day, week or month. The user will be invited to specify this time step in a scrolling list. Missing values are allowed, provided they are coded by “NA”. It is assumed that the dataset will contain at least one year of data.”

b) How do you specify the time span of the training data?

In the previous manuscript, we made a distinction between the two uses of the application. To estimate the burden of the disease, we propose to use the whole dataset in the model fitting. To detect outbreaks, we propose to use only the last few years as a training period. We have clarified this last point at the end of the “Training period” subsection:

“In the latter case, the user is invited to specify the length of the training data in an input field. He can define it in number of years or in number of observations. In either case, the minimal time span accepted is one year.”
3- It is unclear why only trends up to order 3 and only up to 3 sine wave components are allowed.

Throughout this work we have tried to balance the desire for simplicity (fewer options) with the need to support adequate models (more parameters) particularly given that the model fit is automated. It would be unfortunate if the automatic fitting process tried to fit a model which was unidentifiable (too complex). In our experience, trends and seasonal terms up to the 2nd order are usually adequate to describe influenza seasonality. We have added a third seasonal term of period 3 months, as it is sometimes found in the literature and a cubic term for the trend to offer more flexibility. We have now discussed these choices in the third paragraph of the “Discussion” part:

“While first and second degree polynomial trends are frequently used in periodic regression models in the literature,[2,3] we have added the option of a third degree polynomial to offer more flexibility, only for the retrospective analysis. For the seasonal components, we included the most widely used periodicities, ie 12, 6 and 3 months. We did not propose higher degree polynomials or seasonal terms because higher order terms may be more prone to result in unidentifiable models or other problems with model fit.”

4- The model specification on page 4 is not precise. There is no error term in the equation.

Thank you, this has been corrected (in the 2nd paragraph of the “Regression equation” subsection):
“Y(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + γ_1 \cos(2πt/n) + δ_1 \sin(2πt/n) + γ_2 \cos(4πt/n) + δ_2 \sin(4πt/n) + γ_3 \cos(8πt/n) + δ_3 \sin(8πt/n) + ε(t).”

5- In the description of the model fit analysis, it is unclear how the model search works, and what are the two nested models [first paragraph, page 7].

It is true that the description of the selection algorithm was not appropriately done in the “Implementation” section. It was more easily understandable after figure 2 which was in the “Results” section. Therefore, in the new manuscript, we have moved figure 2 to the “Implementation part” (figure 2 is now figure 1). We have also detailed the algorithm description in the last part of the “Regression equation” subsection:

“In our system, automatic selection of the best fitting model is made possible by a selection algorithm (see Figure 1, which illustrates the process on an example detailed in the result section). It relies on ANOVA comparison (significance level : 0.05) to select between nested models, and on Akaike’s Criterion to select between non-nested models.[20] The algorithm starts comparing, by ANOVA, the simplest model M11 (Y(t) = a_0 + a_1 t + γ_1 \cos(2πt/n) + δ_1 \sin(2πt/n) + ε(t)) with the two models in which it is nested: M12 (Y(t) = a_0 + a_1 t + γ_1 \cos(2πt/n) + δ_1 \sin(2πt/n) + γ_2 \cos(4πt/n) + δ_2 \sin(4πt/n) + ε(t)) and M21 (Y(t) = a_0 + a_1 t + a_2 t^2 + γ_1 \cos(2πt/n) + δ_1 \sin(2πt/n) + ε(t)). If none of the alternative models (M12 and M21) is significantly better than the initial one (M11), the algorithm keeps M11 and stops. If one of the two alternative models is better than the initial one, the algorithm keeps it and goes on. If the two alternative
models are better than the initial one, the algorithm keeps the one with the lowest AIC and goes on. The process is repeated until finding the “best overall” model over the nine proposed models.”

6- The AIC criterion is known to overfit the data, why is this the criterion chosen?

We have chosen the AIC criterion because it is one of the criteria which are more widely used to select between non-nested models and should perform adequately in this scenario (Burnham, K. P. and D. Anderson (2003). Model Selection and Multi-Model Inference, Springer). We used ANOVA comparisons to provide a decision criterion for nested models and we used Akaike’s criterion for choosing between non-nested models.

Reviewer: Kenneth Mandl

Reviewer’s report:

I would encourage the authors to make the R code freely available.

We fully agree with the reviewer. The R codes have been included as a zip file in Additional file 1 and also on the application website. We have also highlighted their availability both from Additional file 1 and the website in the first lines of the manuscript “Results” section:

“The R codes are freely available in Additional file 1 and on the application website.”