ESTIMATING THE BASELINE INCIDENCE OF A SEASONAL DISEASE INDEPENDENTLY OF EPIDEMIC OUTBREAKS

Bohumír Procházka[1, 2], **Jan Kynčl**[3, 4]

[1]Unit for Biostatistics, National Institute of Public Health, Prague, Czech Republic
[2]Department of Child and Youth Health, 3rd Faculty of Medicine, Charles University in Prague, Prague, Czech Republic
[3]Unit for Infectious Diseases Epidemiology, National Institute of Public Health, Prague, Czech Republic
[4]Department of Epidemiology, 3rd Faculty of Medicine, Charles University in Prague, Prague, Czech Republic

SUMMARY

In epidemiology, it is very important to estimate the baseline incidence of infectious diseases, but the available data are often subject to outliers due to epidemic outbreaks. Consequently, the estimate of the baseline incidence is biased and so is the predicted epidemic threshold which is a crucial reference indicator used to suspect and detect an epidemic outbreak. Another problem is that the “usual” incidence varies in a season dependent manner, i.e. it may not be constant throughout the year, is often periodic, and may also show a trend between years. To take account of these factors, more complicated models adjusted for outliers are used. If not adjusted for outliers, the baseline incidence estimate is biased. As a result, the epidemic threshold can be overestimated and thus can make the detection of an epidemic outbreak more difficult. Classical Serfling’s model is based on the sine function with a phase shift and amplitude. Multiple approaches are applied to model the long-term and seasonal trends. Nevertheless, none of them controls for the effect of epidemic outbreaks. The present article deals with the adjustment of the data biased by epidemic outbreaks. Some models adjusted for outliers, i.e. for the effect of epidemic outbreaks, are presented. A possible option is to remove the epidemic weeks from the analysis, but consequently, in some calendar weeks, data will only be available for a small number of years. Furthermore, the detection of an epidemic outbreak by experts (epidemiologists and microbiologists) will be compared with that in various models.

Key words: censored data, baseline of incidence, threshold, season dependent model, long-term trend

Address for correspondence: B. Procházka, Unit for Biostatistics, National Institute of Public Health, Šrobárova 48, 100 42 Prague, Czech Republic. E-mail: bohumir.prochazka@szu.cz

http://dx.doi.org/10.21101/cejph.a4800

INTRODUCTION

To illustrate the models proposed, the surveillance systems which are in place in the Czech Republic (CR) will be used. These are the EpiDat system for surveillance of other infectious diseases and the system of surveillance of acute respiratory infections (ARI) including influenza-like illness (ILI). The data pool available for this purpose includes ARI data since 1982, ILI data since 2005, as well as the data on other infectious diseases reportable to the EpiDat system such as varicella (B01) from 1993 to 2014. The diseases listed as well as other diseases have been monitored in the Czech Republic much longer, but for data comparability, only the data from the period where both the above-mentioned surveillance systems were in place will be used.

The aim of this paper is to describe methods for estimating the baseline and threshold incidence of a disease outside epidemics. Different diagnoses were used to illustrate this approach. The statistical and epidemiological methods and terminology used in this paper correspond to those indicated by Armitage (1) and Gail (2). The calculations were made using the R software (3).

RESULTS AND DISCUSSION

The aim is to estimate the baseline incidence depending on season (calendar week of disease onset) and long-term trend. This estimate is the basis for further epidemiological considerations, in particular for the prediction and determination of the epidemic threshold – the cut-off alerting to the epidemic occurrence of the disease monitored.

Let us assume that the incidence has a nearly log-normal distribution. An obstacle to estimating the incidence of a disease is the emergence of epidemics, i.e. of unexpected numbers of cases. This paper focuses on how to reduce the effect of these outliers, how to estimate the baseline (common) incidence which would cover the two types of trends. Different approaches to estimating the long-term trend and seasonal (annual) periodicity will be presented. A model attempting to reduce the effect of epidemics on the general incidence of a disease which can have both the seasonal and long-term trends will be considered. This modification is based on the censored data methods suggested by Kaplan and Meier (4) which are used for survival analysis or analysis of the data below the detection limit. To solve this problem, nonparametric methods can be used, e.g. running median estimation or l-1 estimations of the regression median or regression quantiles (5). This approach does not require a prior assumption of the distribution shape, which may be an advantage, but if the type of distribution is known, this helpful information is sacrificed. Another approach consists in using models for longitudinal data analysis, ARMA or ARIMA models derived from Box and Jenkins (6). This approach provides estimates and predictions. It correlates the values found with those observed previously and therefore, the interpretation of this model is less illustrative. The suggested model has the advantage of separating the outliers (epidemic incidence) from other data. Using regression quantiles eliminates the effect of outliers, but without the identification of epidemic weeks. The ARMA or ARIMA models provide estimates of the incidence and cover the epidemic outbreaks, if any, but the aim of this paper is to estimate the expected incidence while controlling for such outbreaks.

To construct a model, it is crucial to know when precisely the epidemic started. In principle, there are two possibilities:

1) The epidemiologists are able to derive from other information (related to place or time of outbreak or laboratory results) when the epidemic begins.
2) The epidemic can be also estimated based on excess cases – a large deviation from the model suggested. It is to be noted that what is classified as high incidence for ILI in summer can be as considered unusually low incidence in winter.

If the weeks where epidemic outbreaks occurred are known, the possibility of excluding these weeks from the analysis can be considered. Nevertheless, problems may arise from this step, e.g. the incidence will be difficult to assess in certain weeks as the relevant incidence data will only be available for a small number of years, the accuracy of the estimate will become worse and information on the incidence in these weeks, even if subject to inaccuracy, will be lost. At least the value obtained designates the upper limit for the common incidence.

As was already mentioned above, a log-normal distribution is assumed and should be reflected in the model. A similar assumption was made by the authors (7), but without taking account of epidemics, and to construct the log incidence estimate, a log linear model, the lm() function of the R software, was used.

For a simpler situation with outliers where it is known that an epidemic outbreak occurred, the survreg() function from the survival library of the R software will be used. This solution is similar as in the work of Kyncl et al. (8).

First of all, the serial week is defined in accordance with the work of Procházka and Kynčl (7), i.e. week 1 of the year is always from 1 January to 7 January, week 2 from 8 January to 14 January, and so on. Furthermore, inc_i is used to designate the incidence in week yw_i (where yw_i is the sum of week w_i with the 52-fold multiple of year y_i, i.e. $yw_i = 52 \,.\, y_i + w_i$), vector c contains the censoring information, i.e. $c_i = 1$ if inc_i is the usual incidence (with no epidemic increase) and $c_i = 0$ if the incidence in week w_i is considered as excess incidence, it means that the cases are in excess of normal expectancy.

We can create the simplest model where the incidence is assumed to be constant throughout the year and epidemics are identified by experts (who determine whether $c_i = 0$ or 1). The model can be expressed as follows:

```
survreg(Surv(inc,c,type="left")~1,dist="loggaussian")
```

We obtain two parameters of log Gaussian distribution (mean log incidences μ and its scale σ).

From this model, Figure 1 is derived. The estimate of the epidemic threshold is constructed as a prediction interval (95%) for the incidence with a five percent error margin, i.e. with not more than five percent of all values falling above the model (and the model is constructed in such a way that it is not influenced by epidemics). The calculation is done as follows: for non-antilogged baseline estimate exp(μ), the upper 5% prediction interval is calculated from α quantile of normal distribution u_α and estimated standard deviation σ of log incidence, and the obtained limits are transformed as well as the average logs

$$exp\ (\mu + u_\alpha \sigma)$$

where μ is the estimated log incidence and σ is its standard deviation. In reality, more weeks can fall above this threshold due to possible epidemics, and namely this weeks with high incidence are suspected to be epidemic. From Figure 1, it can be clearly seen that such a threshold, the boundary of the grey zone, results in errors on both sides. Errors also arise from the fact that the incidence is a random quantity. The constructed threshold relies upon the historical incidence data and the expert's opinion.

To identify an epidemic, it may be unsuitable to use a constant threshold, without taking account of the long-term trend and seasonality. The solid line in Figure 1 represents the baseline estimated from the incidences after removing the influence of weeks designated by experts as epidemic. Points (daggers) represent the incidences (not considered as epidemic) while circles indicate the incidences in the weeks identified as epidemic by experts. The solid line is the estimated model (of the incidence outside epidemics) and the upper limit of the grey zone is the threshold derived from the baseline (the non-epidemic incidence should not fall above the threshold in more than in 5% of weeks).

The numerical values of the baseline incidence and values of the epidemic threshold for the constant model are indicated below the figure title. Other, more complicated models will be presented below.

To model the incidence of ILI, it was usually assumed that annually, 16% of the weeks are epidemic as inferred by epidemiologists based on long-term experience. But this percentage varies with the diagnosis.

Let us consider the simplest model, assuming that the incidence does not vary in a time interval (year or season – the constant model). Let us suppose that 16% of the weeks are epidemic, try to find the weeks suspected to be epidemic, and display the weeks designated by experts in Fig. 2.

If the epidemic weeks are not known, they can be sought iteratively:

Let us assume that over many years, the percentage of epidemic weeks of ILI is $\prod$. First of all, the above mentioned model will be calculated, and no epidemic outbreak will be assumed (therefore, $c_i = 1$ for any i). For this model, an i is found for which the positive residual is the highest. For this i, let us switch to $c_i = 0$. It means that if there were no epidemic outbreak in this week, the incidence would be lower than or equal to the value obtained. These steps are repeated until 100 $\prod$% of weeks with $c_i = 0$ are achieved.

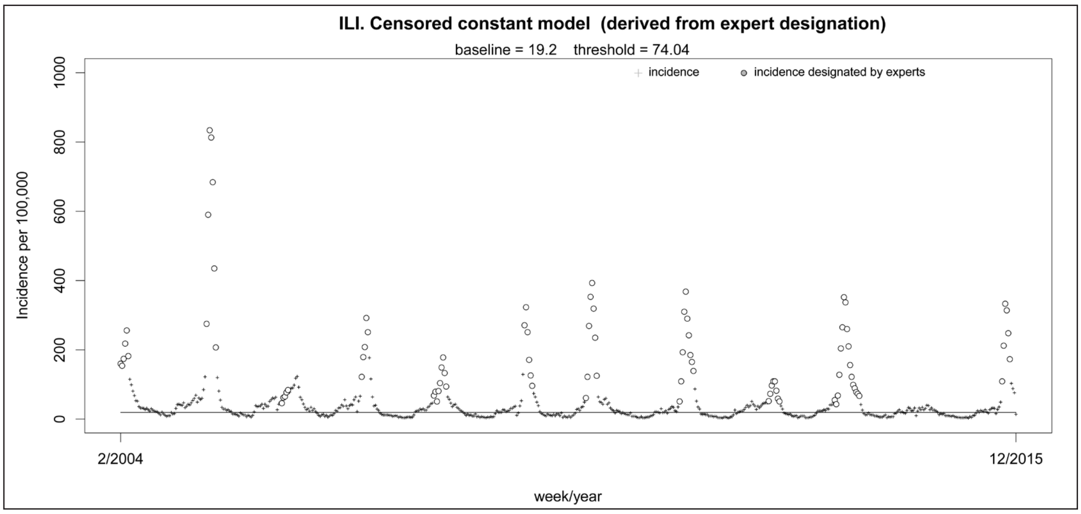

Fig. 1. Model of the incidence of ILI. A constant model for the incidence as identified by experts.

The cut-off percentage of epidemic weeks varies with diseases and should be derived from long-term experience. No unambiguous recommendation can be made and moreover, this cut-off value depends on the season. Despite these considerations, a cut-off value needs to be determined for the purposes of this paper. A 16% censoring will be used for ARI and ILI and a 10% censoring will be assumed for other diseases such as mumps or hepatitis.

These steps can be considered for models of various complexity. Some of them are presented below:

Let us compare the identification of epidemics derived from data (see the procedure above) with that assigned by experts. In Figure 2, dots are incidences, circles are incidences in weeks assumed to be epidemic by experts or the model, and crosses are the remaining weeks. There are high-incidence weeks which were not considered as epidemic by experts (white circles), and conversely, there are low-incidence weeks e.g. with alarming laboratory characteristics (grey circles), or there are weeks identically labelled as epidemic by both experts and the model (black circles). The solid line represents the estimated baseline where the effect of the16% of the highest values (identified as the highest, but not considered as epidemic by epidemiologists; therefore they correspond with the black and grey dots). Grey zones of different widths represent the periods labelled as epidemic by experts.

Figure 3 compares the two approaches from Figures 1 and 2, i.e. the epidemics identified by experts, by the model used or by both approaches (experts and model). One of the models does not take account of the information that an epidemic occurred (baseline – solid line and the epidemic threshold – dotted. In the other model, the solid line and dark grey zone were obtained by identifying the high-incidence weeks as indicated above and then by controlling for the effect (of the epidemics identified this way) on the estimate. Circles represent high incidence weeks and crosses other weeks. The baseline and threshold values represented in Figure 2 correspond to the estimates which are constructed to control for the effect of epidemics on the calculation of the estimate. The disagreement in the baseline and epidemic threshold values between Figures 1 and 2 results from the fact that the expert assumption may not fully correspond with the incidence data as can be seen from the circles in Figure 2 (disagreement is represented by grey and white circles). On the other hand, the disagreement between two estimates (censored and uncensored) in Figure 3 results from whether the effect of epidemics on the estimated baseline and epidemic threshold is considered or not (expert assumption is not considered). If this model is used to detect epidemics, the censored estimate is a clearly superior option.

The fact that the disagreement between the model values e.g. in Figure 3 seems to be substantially smaller in comparison with

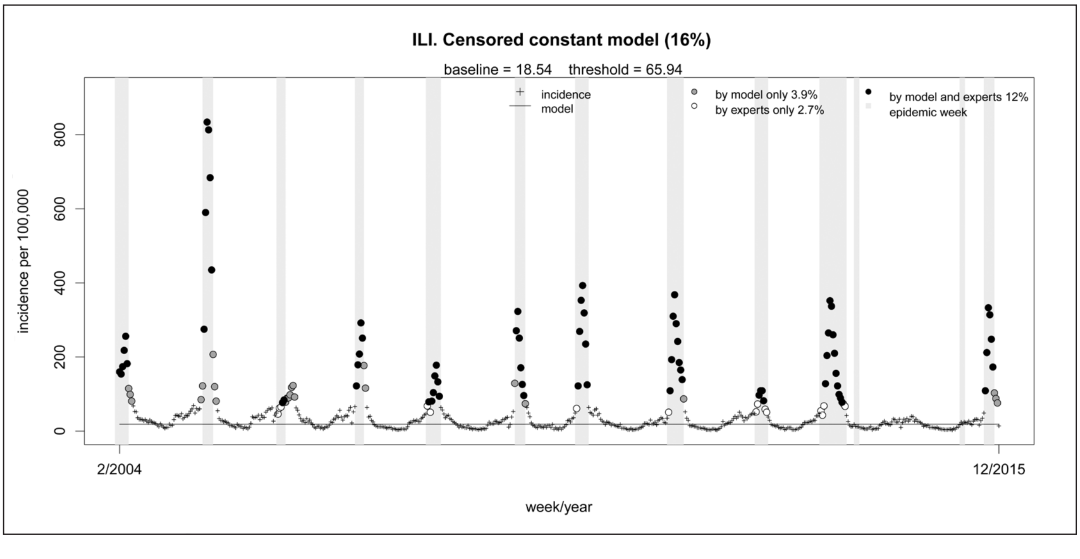

Fig. 2. A constant model for the incidence of ILI. Comparison of epidemics identified by experts and by the model.

that between the thresholds is mainly due to the use of the log-normal distribution which is more suitable for modelling the incidence. This difference is caused by using different models. The model using censored data is less sensitive to outliers (epidemic incidence), which is common to the log-normal distribution of incidences.

To calculate the model that assumes the incidence of ILI is constant throughout the year, the following function was used:

```
survreg(Surv(inc,c,type="left")~1,dist="loggaussian")
```

where *inc* is the incidence, *c* indicates whether an epidemic occurred ($c=0$) or not ($c=1$) in the respective week (*yw*).

In Figure 3, decline in the incidence can be seen, but the long-term trend is not taken into consideration in the models used. This illusion is likely to result from the epidemics occurring primarily in 2005.

The vertical light grey lines in the figures are the beginnings of years.

Now, more complex models will be presented below. The constant model can be generalized by substituting the constant 1 with a function, e.g.

```
survreg(Surv(inc,c,type="left")~ . . . ,dist="loggaussian")
```

where, similarly to the work of Procházka and Kynčl (7), any model can be imagined instead of . . . from the above mentioned simplest model (constant incidence) to periodicity models, Serfling's model (9), or ANOVA model with differently estimated long-term trend.

To calculate Serfling's first-order model (Fig. 4), the following function will be used:

```
survreg(Surv(inc,c,type="left")~yw+sin(2*pi*yw/52)
+cos(2*pi*yw/52),dist="loggaussian")
```

When this model is used, the week and year need to be identified (*yw*). The model is shown in Figure 4. The annual cyclicity of ILI is fully taken into account, but only a long-term linear trend is considered.

In Figures 3 and 4, it is clearly visible that, unlike the censored model, the uncensored model is influenced by epidemics (outliers); therefore, it is more difficult to detect the excess incidence weeks as the threshold is higher. Moreover, Figure 4 accounts for seasonal fluctuations and a possible linear trend.

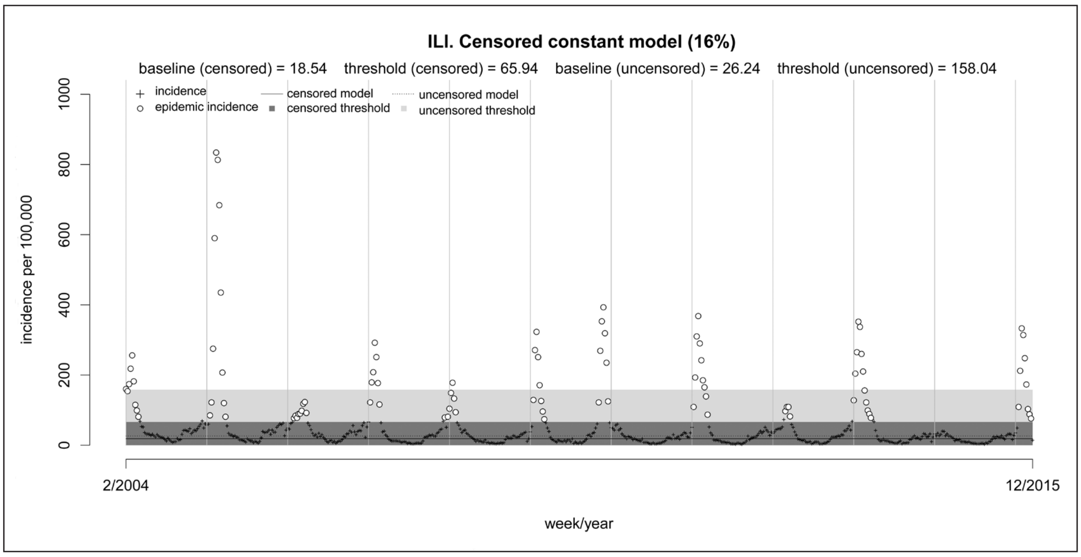

Fig. 3. *The constant model for the incidence of ILI. Epidemic weeks identified by the model, censored and uncensored.*

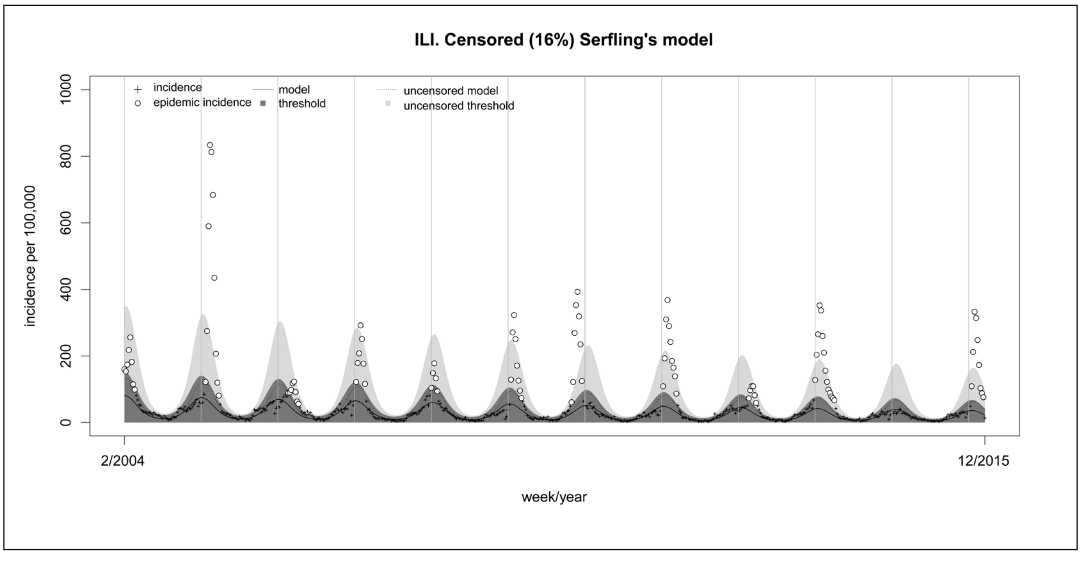

Fig. 4. *Serfling's model of the incidence of ILI, censored (16%) and uncensored.*

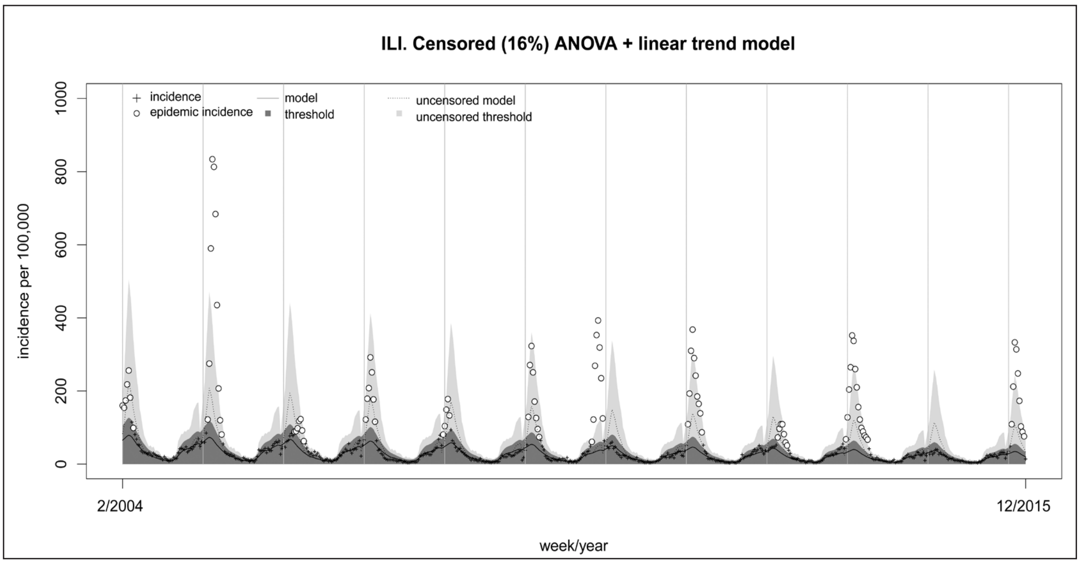

Fig. 5. Censored (16%) and uncensored ANOVA seasonality model with a linear trend.

Nevertheless, it may not be realistic to assume linearity and that is why an attempt will be made to propose a model with a more complex trend. Another problem is that it may not be always appropriate to use the sine function which has a very smooth and symmetrical shape. To get closer to the data (to find the real shape of the model), Serfling's higher order model or ANOVA model will be used:

A censored Serfling's order 5 model is calculated using the following function:

```
survreg(Surv(inc,c,type="left")~yw
+sin(2*pi*yw/52)+cos(2*pi*yw/52)
+I(sin(2*2*pi*yw/52))+I(cos(2*2*pi*yw/52))
+I(sin(3*2*pi*yw/52))+I(cos(3*2*pi*yw/52))
+I(sin(4*2*pi*yw/52))+I(cos(4*2*pi*yw/52))
+I(sin(5*2*pi*yw/52))+I(cos(5*2*pi*yw/52)),dist="loggaussian")
```

An ANOVA model with a linear trend is calculated as follows:

```
survreg(Surv(inc,c,type="left")~yw
+as.factor(w),dist="loggaussian")
```

The results obtained in these models are very similar; that is why only the ANOVA model with a linear trend for the incidence of ILI is shown (Fig. 5).

These models are appropriate for use for the diseases where a linear or constant long-term trend can be expected, but with other diseases, a more complicated trend needs to be considered, e.g. as shown in the work of Procházka and Kynčl (7). In Figure 5, the difference between uncensored and censored models is shown. The uncensored threshold from a simpler model – the light grey zone only detects the excess incidence weeks with difficulty. The model for a linear trend for these data also allows higher incidence in the beginning than in the end. The censored model reduces, but does not fully rule out, the effect of high-incidence weeks on the constructed estimate.

In Figure 5, it appears that the linear trend (in particular in the uncensored model) is not able to reduce the effect of high-incidence weeks in 2009–2011. To solve this problem, more complex models need to be used.

As shown in the work of Procházka and Kynčl (7), the data series can be decomposed into two components, one cyclic and the other, long-term trend. This step can be made by subtracting the

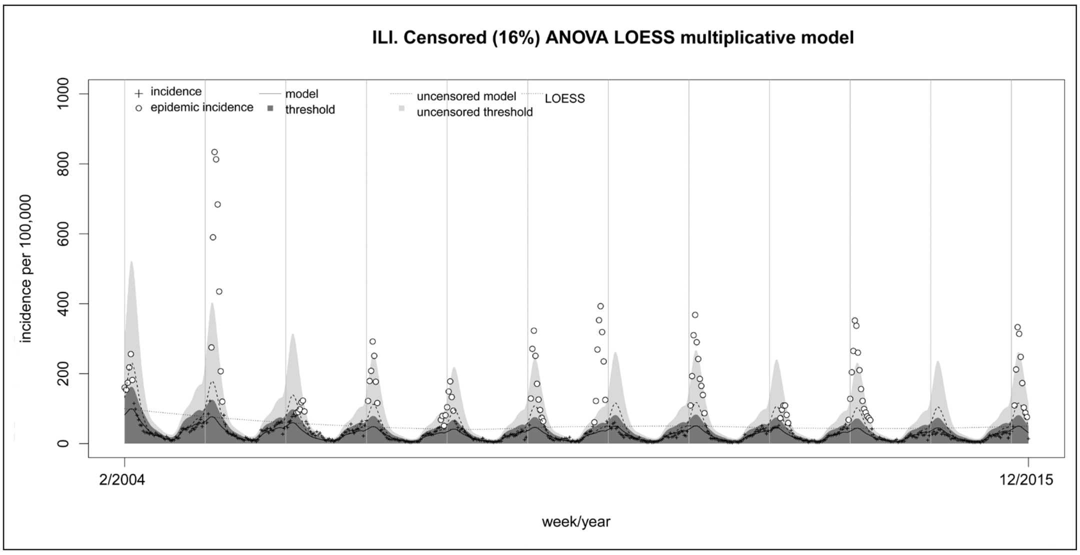

Fig. 6. Censored (16%) and uncensored ANOVA model with LOESS for ILI.

long-term trend from the data series or by dividing the data series by the long-term trend. Given that the lognormal distribution of the incidence is assumed, the multiplicative approach is more suitable (7). The multiplicative ANOVA model with the LOESS trend using the data from Figure 5 is shown in Figure 6. In comparison with the ANOVA model with the linear trend, Figure 6 shows higher estimated incidence in 2009–2011. The most marked difference appears in the end of the time series (in 2015).

The difference between the linear and ANOVA models for the incidence of ILI seems to be negligible. Other infectious diseases are addressed below.

The model of the incidence of mumps (B26) is shown in Figure 7. From the figure, it is evident that significant epidemics occurred in 2005, 2010, and 2011 and are visible even in the uncensored model, but the epidemics in 1997, 1998, 2002, 2006, 2011, and 2012 are only suspected by the censored model. The model also takes account of the long-term trend. Greater differences in high incidence weeks between the two models are due to the log scale.

In Figure 8 showing acute viral hepatitis A (B15), it can be seen that the long-term trend is estimated and hints that there is long-term cyclicity (ca 12 years), but the available data series is too short to prove it. Annual fluctuations are not found by the model, but they are insignificant and subject to accident as suggested by different shapes of the two models used (either with or without outliers).

Let us go back to Figure 2 that compares the simplest model with the constant trend, the identification of epidemics by experts, with the last model. One of the more complex models will be used that attempts to get as close as possible to the incidences outside the epidemic periods – ANOVA model with the LOESS trend shown in Figure 9. Figure 9 is constructed the same way as Figure 2, but uses a more complex model which allows a more realistic estimate of the epidemic threshold. From the calculated congruence percentages, it can be seen a slight improvement in congruence percentage between the identification of epidemics by experts and the model.

CONCLUSION

The most important point is the comparison of the models presented. Although based on different approaches, spanning from Serfling's models based on sine and cosine functions and models smoothing the weekly average incidence, all models tend to yield very similar charts. The main contribution of this article is to introduce the methodology which reduces the impact of the outlying incidences (epidemics). In addition it demonstrates the difference between the models that either do or do not eliminate the effects of the epidemic (dark grey and light grey streaks in

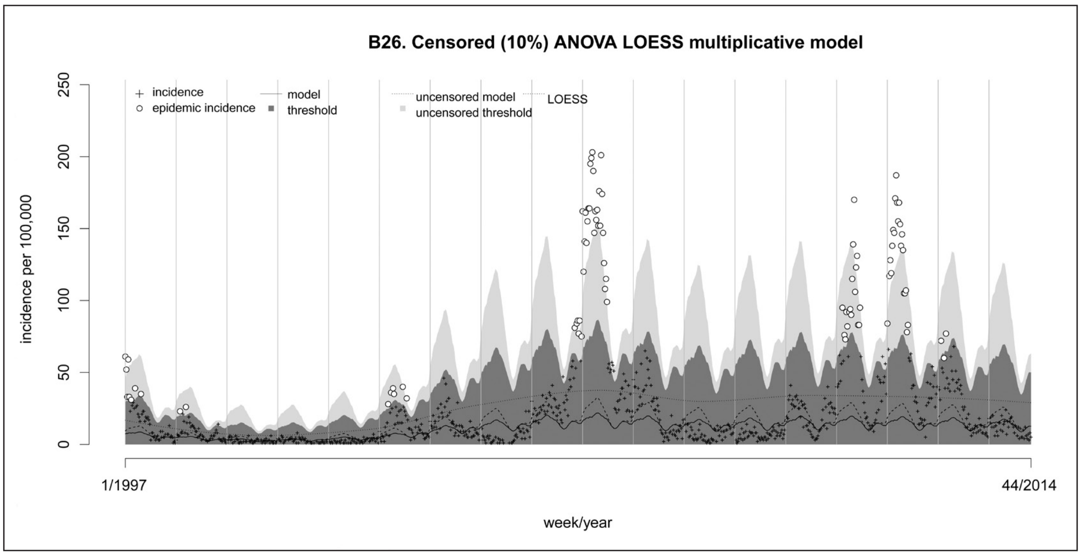

Fig. 7. *Censored (10%) and uncensored ANOVA model with LOESS for mumps.*

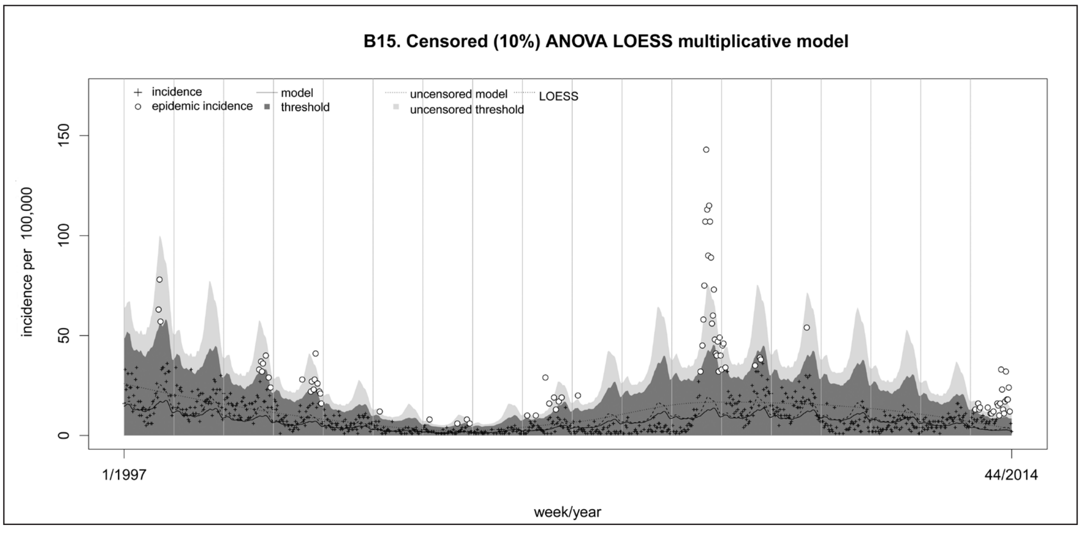

Fig. 8. *Censored (10%) and uncensored ANOVA model with LOESS for hepatitis A.*

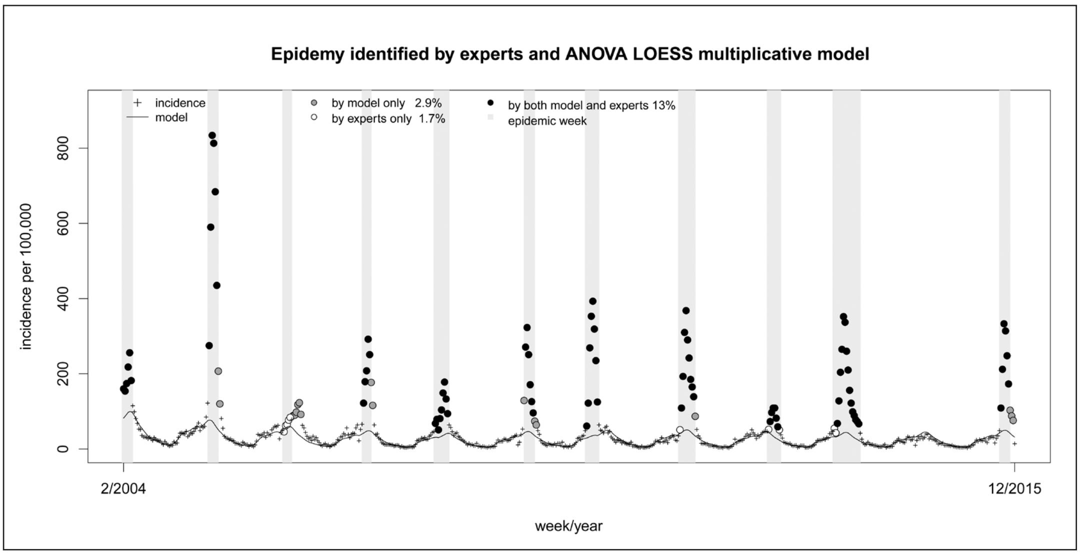

Fig. 9. *Comparison of the ANOVA model with the LOESS trend and identification of the epidemic by experts for the incidence of ILI.*

Figures 3–8). In contrast, Figures 2 and 9 illustrate the agreement between the models (ANOVA and constant model) and the identification of epidemics by experts.

Acknowledgement

Supported by MH CZ – DRO (National Institute of Public Health – NIPH, IN 75010330)

REFERENCES

1. Armitage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: John Wiley & Sons; 1998.
2. Gail HM, Benichou J. Encyclopedia of epidemiologic methods. Chichester: John Wiley & Sons; 2000.
3. R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2008.
4. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. J Am Stat Assoc.1958;53(282):457-81.
5. Koenker R, Bassett G Jr. Regression quantiles. Econometrica. 1978 Jan;46(1):33-50.
6. Box GEP, Jenkins GM. Time series analysis: forecasting and control. Rev. ed. San Francisco: Holden-Day; 1976.
7. Procházka B, Kynčl J. Estimating the baseline and treshold for the incidence of diseases with seasonal and long-term trends. Cent Eur J Public Health. 2015;23(4):352-9.
8. Kyncl J, Prochazka B, Goddard NL, Havlickova M, Castkova J, Otavova M. A study of excess mortality during influenza epidemics in the Czech Republic, 1982-2000. Eur J Epidemiol. 2005;20(4):365-71.
9. Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Rep. 1963 Jun;78(6):494-506.

Received November 13, 2015
Accepted in revised form September 23, 2016