# ESTIMATING THE BASELINE AND THRESHOLD FOR THE INCIDENCE OF DISEASES WITH SEASONAL AND LONG-TERM TRENDS

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

[1]Department of Biostatistics, National Institute of Public Health, Prague, Czech Republic

[2]Department of Infectious Diseases Epidemiology, Centre for Epidemiology and Microbiology, National Institute of Public Health, Prague, Czech Republic

**SUMMARY**

In epidemiology, it is very important to estimate the baseline incidence of infectious diseases. From this baseline, the epidemic threshold can be derived as a clue to recognize an excess incidence, i.e. to detect an epidemic by mathematical methods. Nevertheless, a problem is posed by the fact that the incidence may vary during the year, as a rule, in a season dependent manner. To model the incidence of a disease, some authors use seasonal trend models. For instance, Serfling applies the sine function with a phase shift and amplitude. A similar model based on the analysis of variance with kernel smoothing and Serfling's higher order models, i.e. models composed of multiple sine-cosine function pairs with a variably long period, will be presented below. Serfling's model uses a long-term linear trend, but the linearity may not be always acceptable. Therefore, a more complex, long-term trend estimation will also be addressed, using different smoothing methods. In addition, the issue of the time unit (mostly a week) used in describing the incidence is discussed.

*Key words:* seasonal trend, long-time trend, incidence, epidemic threshold, LOESS, Serfling model, ANOVA model

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

## INTRODUCTION

The surveillance systems operated in the Czech Republic, i.e. the system of surveillance of acute respiratory infections (ARI) including influenza-like illnesses (ILI) and the EpiDat system for surveillance of other infectious diseases, will be used to illustrate the methods proposed. The data pool available for this purpose includes ARI data since 1982, ILI data since 2004, as well as the data on other infectious diseases reportable to the EpiDat system such as varicella (B01) since 1993. Although the diseases listed as well as other diseases have been monitored in the Czech Republic much longer, for the reason of comparability of data from different years, only the data from the period where both the surveillance systems, i.e. EpiDat and ARI, were in place will be used (1, 2).

The situation is simpler in ARI where the data are collected as the numbers of patients per calendar week. For the infections monitored in the EpiDat system, the situation is complicated by the fact that two types of dates are entered in it: the disease report date (indicated by the physician) or the date of the onset of symptoms.

In general, the data need to be aggregated into longer time intervals than the reporting units which are often very short, e.g the EpiDat reporting unit is one day (considering days implies the problems of small numbers and weekly periodicity). By aggregating data into years, such problems are avoided, but the model only depicts a long-term trend, whose shape can vary over years, including multiple-year periodicity, but is not able to reflect the fluctuations in the incidence during the year. Annual aggregation does not pose any problem in terms of the seasonal variation, as the years are practically of equal length and the year spans all seasons which may influence the incidence of a disease. Nevertheless, this type of aggregation does not allow for the prediction for shorter time periods such as one week which may be a crucial unit for many diseases.

Using the calendar units has some drawbacks, since the months vary in length and the length of the year is not a multiple of the length of one week (when the year is divided into seven-day periods, 1.25 day is in excess, i.e. one day every year plus one day every leap year). When modelling the incidence of a disease requiring daily data, this problem can be solved as follows: a suitable time unit will not be a calendar week but a serial week, with seven-day weeks counted from the very beginning of the year (1 January), i.e. week 1 for the first seven days, week 2 for the second seven days, and so on. Instead of seven days, the last serial week, week 52, will thus contain eight or even nine days in the leap years. Anyway, from the point of view of disease reporting, the last week of the year is problematic in nature as it is exceptional. Although far from being ideal, this approach has the advantage of grouping all exceptional days into one defined week; of course, with the exception of Easter which needs separate attention.

Another factor to be considered is the type of the date reported. E.g. two different dates are entered in the EpiDat system, i.e. the disease report date and the date of the onset of symptoms. The situation may be interpreted differently depending on the type of the date reported, but the mathematical models used do not differ. From the medical perspective, the date of the onset of symptoms is more relevant.

When working with the ARI or ILI data, calendar weeks should be used, since the system does not gather data on individual cases

but age aggregate data on the weekly numbers of cases. Concerning these weekly data, the issue of weeks needs to be approached differently. Calendar week 53 is incomplete, similarly to week 1 of the following year, and therefore, it will be reasonable to merge them into one week, i.e. week 1 of the following year. Consequently, a similar effect will be produced as described above for the serial weeks. It will also result in an inaccuracy, although of a different nature. The week at the turn of the year includes a variable number of days of the Christmas holiday period that influences human behaviour, including the willingness to see a doctor. Similarly, other holidays may not overlap with the calendar weeks. The model generated will serve for estimating the expected incidence of a disease and for determining the cut-off for excess incidence. It will require smoothing and the inaccuracy resulting from using the above specified time units is considered as insignificant.

The statistical and epidemiological methods and terminology used in this paper are specified in works by Armitage and Colton, and Fleming et al. (1, 3).

## MATERIALS AND METHODS

Another aim is to determine the baseline, expected incidence (number of cases divided by the total of the population monitored) depending on the season (week of the onset of symptoms), cyclic variations and long-term trend. Let us assume the incidence has a log-normal distribution, with the average that varies depending on both the season and long-term trend. Several variants of the estimation of the long-term trend and annual periodicity pattern will be presented. Different methods are used to adjust for the long-term and seasonal trends (4–7). They are very similar to those for estimating excess deaths (8–12). All computations were made in the R software (12)*.

One of the aims of modelling the incidence is to find the expected, mean, time-dependent baseline and mainly to determine the epidemic threshold which identifies outliers. This threshold can be defined as the prediction interval in the respective time – i.e. a cut-off which is exceeded in the selected percentage of the weeks only.

To describe the methods, the following symbols are used: *inc*, vector of the incidence rates observed and *yw*, vector of the respective weeks

$$yw = 52 \cdot y + w$$

where $y$ is the year and $w$ is the week in which the respective incidence was observed.

First of all, the simplest model without cyclicity will be presented, i.e. the model that assumes a constant incidence (the logarithm of incidence) throughout the year, i.e.:

**lm(log(inc)~1)**

From the perspective of the seasonal trend of the disease, this model is very simple. It assumes the incidence of ILI is constant throughout the year (Fig. 1). The grey zone in the picture represents the 95% prediction limits for a log-normal distribution of the incidence, it means that the borderline separates 5% of the highest values that are the most distanced outliers relative to the baseline. It is the threshold sought. The selection of the percentage is a matter of personal preference. The model in this picture represents the commonly used threshold but does not take into account any variation with time. As the model considered here is constant, the model or epidemic threshold can be characterized by a single value, as shown in Figure 1.

The epidemic threshold can be calculated by computing the prediction interval for the non-anti-logged estimated baseline μ from α quantile normal distribution $u_\alpha$ and estimated standard deviation σ:

$$e^{\mu} \pm u_\alpha \cdot \sigma$$

More precisely, the incidence is 95% likely to be below the constructed borderline (the assumption is erroneous in five percent of cases).

Evidently, the model in Figure 1 does not reflect the seasonal trend or any other long-term trend in the disease (nevertheless, the data in the figure seem to be declining). The simplest way to model a drop is to use a linear model. For the linear model of the incidence of ILI (Fig. 2), a linear term, *yw*, is added:

**lm(log(inc)~yw)**

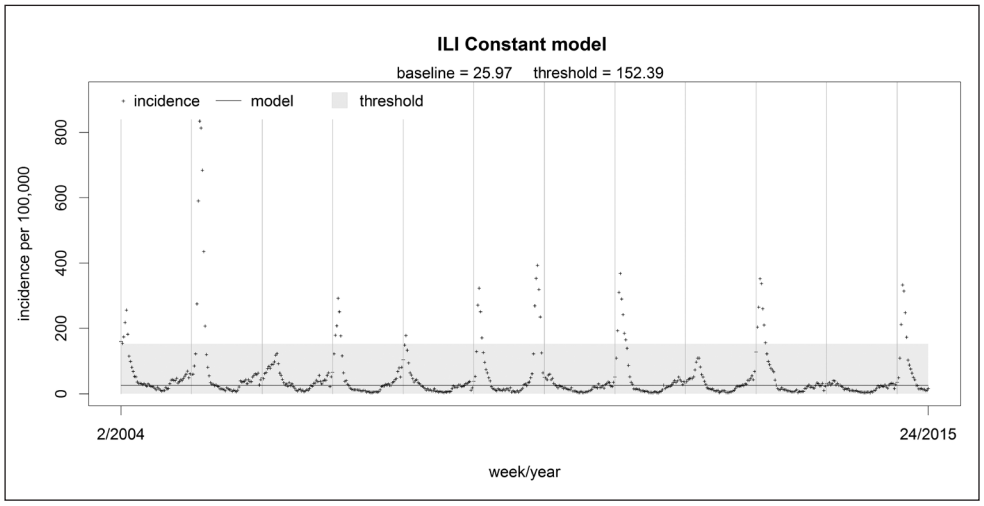


***Fig. 1.*** *Model of the incidence of ILI constant in time regardless of seasonality.*

*The R Project for Statistical Computing, http://www.R-project.org

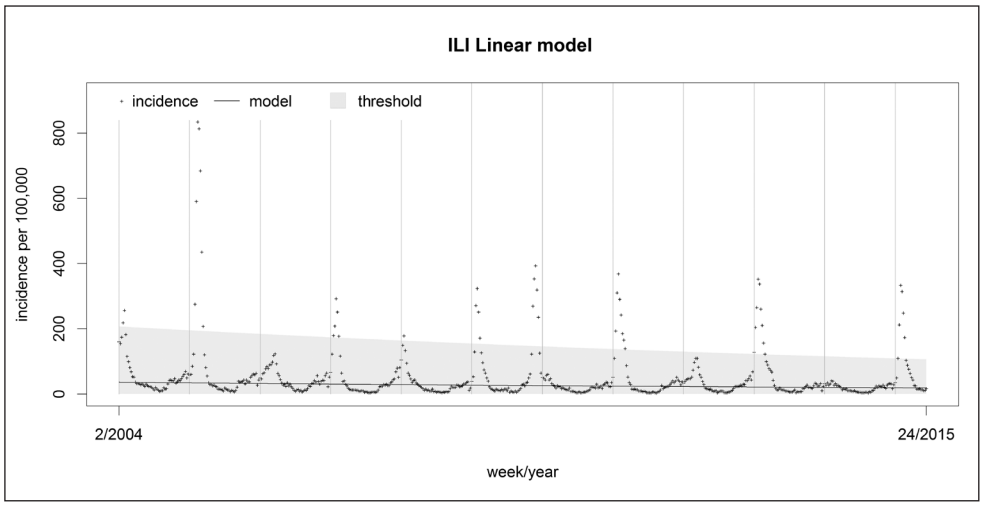


***Fig. 2.*** *Model of the incidence of ILI linearly depen-dent on time regardless of seasonality.*

Here, it is not possible to use a single number to describe the baseline incidence or epidemic threshold, but a similar calculation is applied, except that the estimation in time *yw* is used instead of $\mu$. A more complicated model with an order 2 polynomial is constructed as follows:

**lm(log(inc)~yw+I(yw^2))**

Attention will be paid to the seasonality which is a key characteristic of ILI.

Many diseases with a cyclic trend are often represented as a line chart spanning several years (Fig. 3). The last year is compared to the other years as a whole, e.g. to a bundle of years 2011 to 2015.

Figure 3 clearly shows the annual periodicity, but any long-term trend is difficult to consider. Therefore, various models are generated encompassing a long-term trend. One of these, often applied in practice, is the model using trigonometric functions, called Serfling's model (12). It uses, apart from the long-term linear trend, a cyclic component (trigonometric function) to model annual periodicity. To this end, the serial week number is multiplied by 2π so that one year represents one period. The following model will be used:

**lm(log(inc)~yw+sin(2*pi*yw/52)+cos(2*pi*yw/52))**

Thus, the timeline of incidences is fit with the weighted sum of the sine and cosine functions with the equal period, which is equivalent to fitting the sine function with two parameters – the amplitude (wave height) and the phase shift. This record using the sine and cosine functions is more suitable for calculating the model parameters, the amplitude and shift, which are helpful in the interpretation. The model is represented in Figure 4.

The parameters with the sine and cosine functions are difficult to interpret (let us designate them $\gamma_{cos}$ and $\gamma_{sin}$) but can be easily converted to the amplitude (*Ampl*) and phase shift ($\varphi$) which are much helpful. When considering the simplest Serfling's order 1 model, which uses the sine function while estimating two para-meters determining the amplitude and shift and the linear trend, it can be expressed as follows:

$$\gamma_{sin} \cdot sin(t) + \gamma_{cos} \cdot cos(t) + \gamma_0 + \gamma_1 \cdot t + \gamma_2 \cdot t^2$$

where the coefficients are estimated. The coefficients sin and cos are the coefficients of the respective trigonometric functions and the coefficients 0, 1, and 2 are the constant, linear, and quadratic terms. However, the interpretation of these coefficients poses a minor problem. As indicated above, it is more suitable to switch to the following model:

$$Ampl \cdot sin(t + \varphi)$$

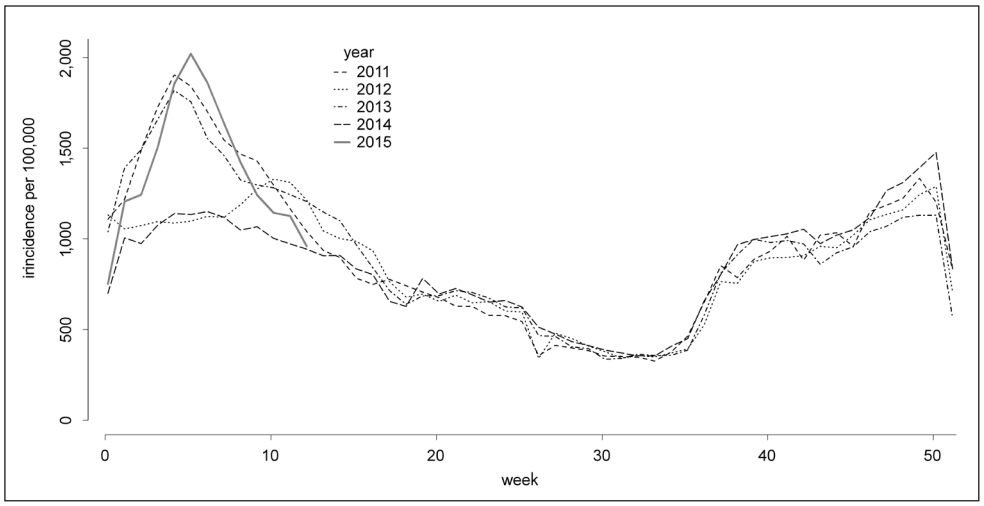


***Fig. 3.*** *Weekly incidence of ARI over several years.*

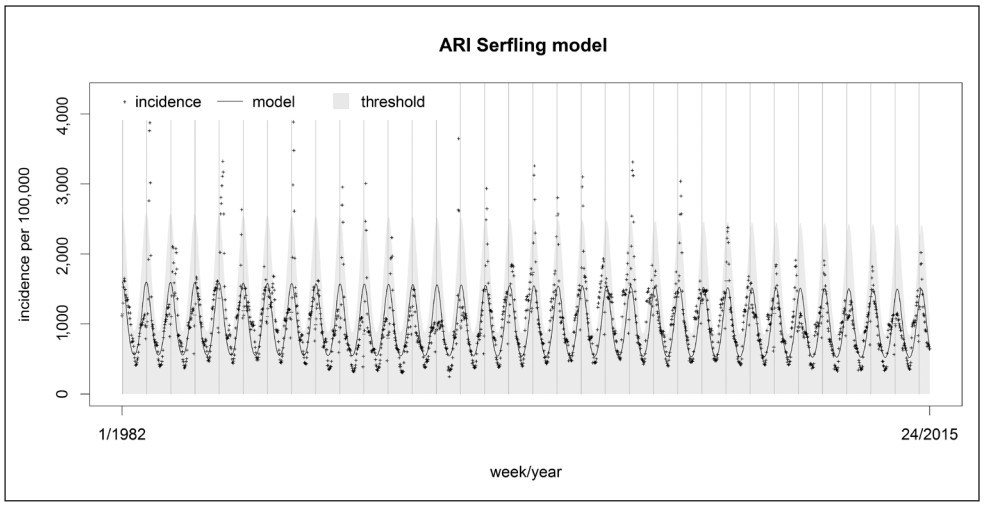


***Fig. 4.*** *Serfling's order 1 model for the incidence of ARI.*

where *Ampl* is the amplitude of the model and $\varphi$ is the time shift. These coefficients are much easier to interpret.

They are calculated from the coefficients obtained as follows:

$$\varphi = \arctan\left(\frac{\gamma_{\cos}}{\gamma_{\sin}}\right)$$

and

$$Ampl = \frac{\gamma_{cos}}{sin(\varphi)}$$

The initial parametrization is very helpful in the calculation of an estimate. However, the estimated parameters should be converted using the formulas above. The following function can be used in R for this purpose:

```
> Am_pos<-function(a)
+ {
+ fi<-atan(a[2]/a[1])
+ Am<-a[1]/(cos(fi))
+ p<-c(Am,fi)
+ names(p)<-c("Amplitude","Shift")
+ p
+ }
```

Thus, it is possible to get both the sine and cosine coefficients equal to 1 and the following is obtained:

```
> Am_pos(c(1,1))
Amplitude Shift
1.4142136 0.7853982
>
```

This model has the disadvantage of being too smooth (looks artificial) which is not consistent with the modelled timeline. One of the ways to improve this model is fitting the sum of the sine functions with a variable period length (half, third or quarter length). It can be called Serfling's order r model. Such a model (e.g. order r=5) is calculated as follows:

```
lm(log(inc)~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)))
```

The phase shift and amplitude can also be calculated for the seasonality parts thus obtained. This model is represented in Fig. 5.

Another possible way to model the incidence of a disease is derived from the analysis of variance. Such a model is proposed in the present study and is called the ANOVA model. The serial week number is considered a categorical quantity.

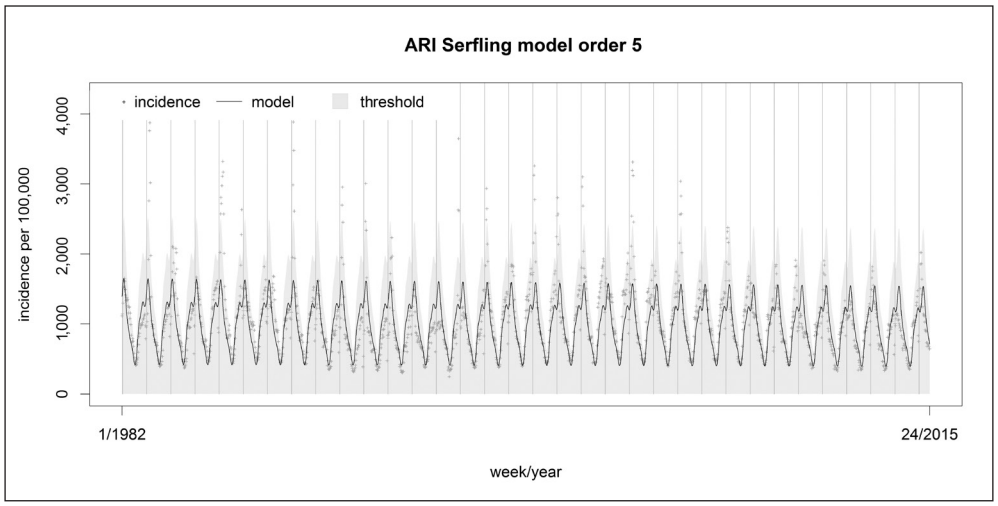


***Fig. 5.*** *Serfling's order 5 model for the incidence of ARI.*

For a specific week in different years, the mean (geometric mean because of the log normal distribution) is counted and drawn. To calculate this model, the following formula is used:

```
lm(log(inc)~0+as.factor(w))
```

But the model in Fig. 6 does not contain any long-term trend. Not using INTERCEPT does not matter, it can be even helpful in drawing the model (there is no need to sum up the parameters for individual weeks with the INTERCEPT, but it is present in the estimated parameters). This model has the advantage of not assuming the seasonality to be symmetrical in shape (unlike trigonometric functions) and is similar to Serfling's high order model. It has the disadvantage of using a rather rugged model of the period and not taking into account a linear trend as is contained in Serfling's model. The estimate's ruggedness can be overcome by using a kernel estimate with a triangular kernel with the highest weight put in the centre and decreasing towards the extremities of the window. The result of smoothing for an annual cycle is illustrated in Fig. 7. While smoothing, it should be kept in mind that the model is cyclically repetitive from year to year and therefore, the kernel at the end of one year needs to be joined to the beginning of the next year (R language function). Such smoothing should better be done prior to the addition of a long-term trend, if any, since consequently, smoothing at the beginning and at the end of the series is simplified. The extent of smoothing can be selected based on the window's width and weight values.

To smooth this model, it is possible to use the kernel estimate computed with the following algorithm:

```
kl3tyd<-function(oh){
    o<-NA
    o[1]<-(oh[52]+2*oh[1]+oh[2])
    for(i in 2:51) o[i]<-(oh[i-1]+2*oh[i]+oh[i+1])/4
    o[52]<-(oh[51]+2*oh[52]+oh[1])/4
    o
}
```

which considers a three-week window and puts double weight on the central value in comparison with both extremities. For the first or the last week of the year, the model benefits from the fact that the last week of the previous year is followed by the first week of the next year and the values are expected to be unchanged, with the exception of the long-term trend. The long-term trend will be taken into account in the resultant estimate.

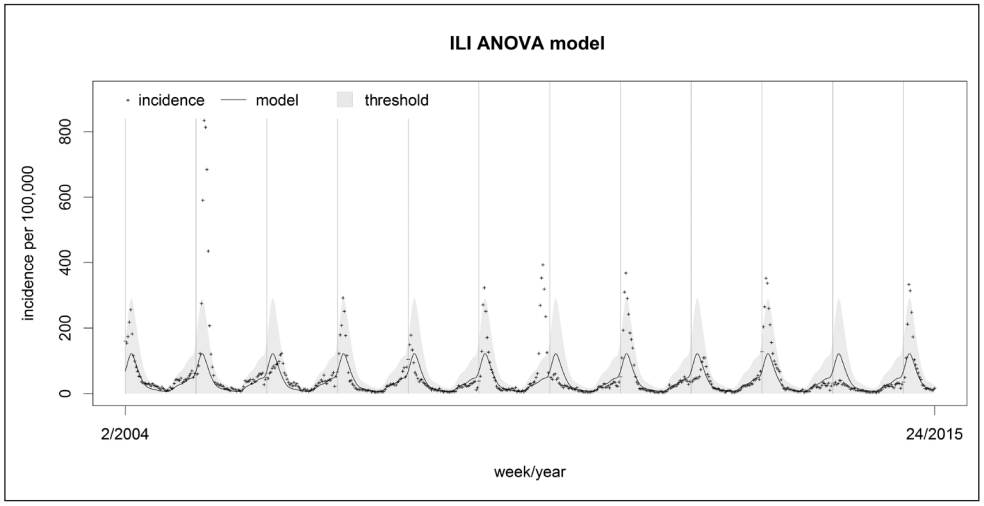


***Fig. 6.*** *ANOVA model for the incidence of ILI without trend.*

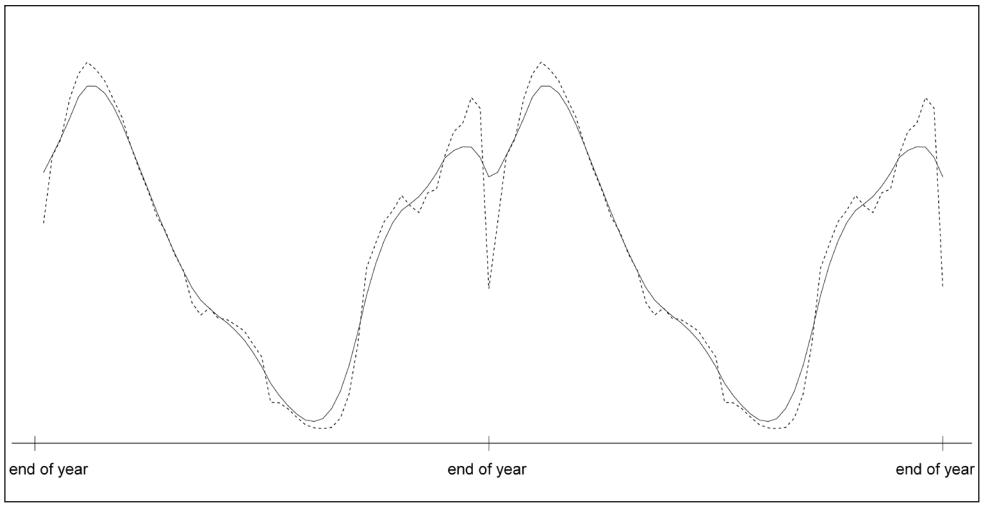


***Fig. 7.*** *ANOVA model for seasonality in the incidence of ARI prior to (dotted line) and after (full line) smoothing.*

To achieve a higher smoothing effect, the function below can be used which considers a five-week window:

```
kl5tyd<-function(oh){
      o<-NA
      o[1]<-(.5*oh[51]+oh[52]+2*oh[1]+oh[2]+.5*
      oh[3])/5
      o[2]<-(.5*oh[52]+oh[1]+2*oh[2]+oh[3]+.5*oh[4])/5
      for(i in 3:50){
      o[i]<-(.5*oh[i-2]+oh[i-1]+2*oh[i]+oh[i+1]+.5*oh[i+2])/5
      }
      o[51]<-(.5*oh[49]+oh[50]+2*oh[51]+oh[52]+.5*
      oh[1])/5
      o[52]<-(.5*oh[50]+oh[51]+2*oh[52]+oh[1]+.5*
      oh[2])/5
      o
}
```

In the picture, an artefact can be seen, i.e. the decline in the incidence in the last week of the year. This drop is evident not only for the disease reporting dates but also for the disease onset dates. A possible explanation is that in the last week of the year, many people attempt self-treatment at home rather than going to see a doctor.

To illustrate variation in the flexibility of different models, a diagnosis was selected which shows considerable changes in the long-term trend, i.e. campylobacteriosis (A04.5) or salmonellosis (A02).

A linear trend will be integrated into the ANOVA model if the following equation is considered:

```
lm(log(inc)~yw+as.factor(w))
```

This model is shown in Figure 8. It is close to Serfling's higher order model. However, the model obtained tends to overestimation in its extremities and to underestimation in its central part.

Similarly to the model without cyclicity, both Serfling's and ANOVA models can include a polynomial trend instead of the linear one, e.g. Serfling's order 1 model will include an order 3 polynomial trend as follows:

```
lm(log(inc)~yw+I(yw^2)+I(yw^3)+sin(yw)+cos(yw))
```

or for an ANOVA model:

```
lm(log(inc)~yw+I(yw^2)+I(yw^3)+as.factor(t))
```

Using a polynomial may result in shaping the data in a non-realistic way, as each polynomial tends to stretch to infinity at its extremities.

The same effect can be achieved by dividing the trend into two components – a long-term trend and a cyclic trend:

- The long-term trend (linear, polynomial, or other) will be estimated;
- The estimated trend will be subtracted from the incidence. There are two possible ways for doing so:
  *Additive:* the estimated trend will be subtracted from the incidence and in further steps, this difference will be modelled;
  *Multiplicative:* the incidence will be divided by the estimated trend and in further steps, this quotient will be modelled;
- The seasonal component will be estimated (using Serfling's model or the ANOVA model);
- The seasonal and long-term components will be put together to obtain the required estimate of the expected (baseline) incidence.

Given the assumed log normal nature of the incidence, the multiplicative approach appears to be more appropriate while in the additive model, the normal nature of the incidence is suggested. The assumed log-normal distribution and multiplicative model have the advantage of not allowing the prediction of negative values.

The ANOVA model extended with the multiplicative LOESS (local regression models) trend estimate is shown in Figure 9.

When a linear model is used and a normal incidence distribution is assumed, Figure 11 is obtained. A serious problem is evident: the prediction can yield negative values.

To calculate the running mean, the length of the window needs to be established. To free the long-term trend from the annual periodicity, the length of the window must be a multiple of one year, i.e. 52 weeks, 104 weeks, and so on. The fact that the number of weeks in a year was adjusted to 52 appears to be relevant in this context. To provide a more general insight, a model with the estimation of

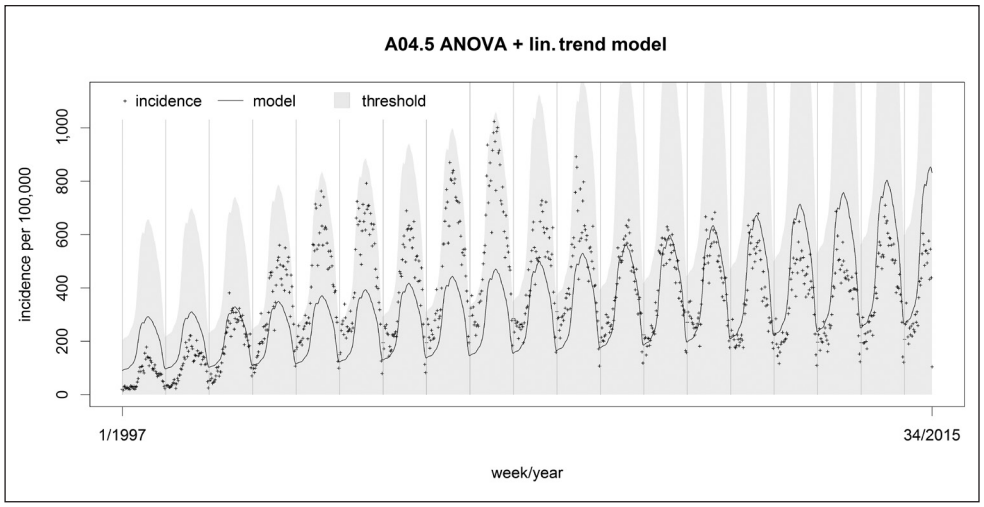


***Fig. 8.*** *ANOVA model with a linear trend for the incidence of campylobacteriosis.*

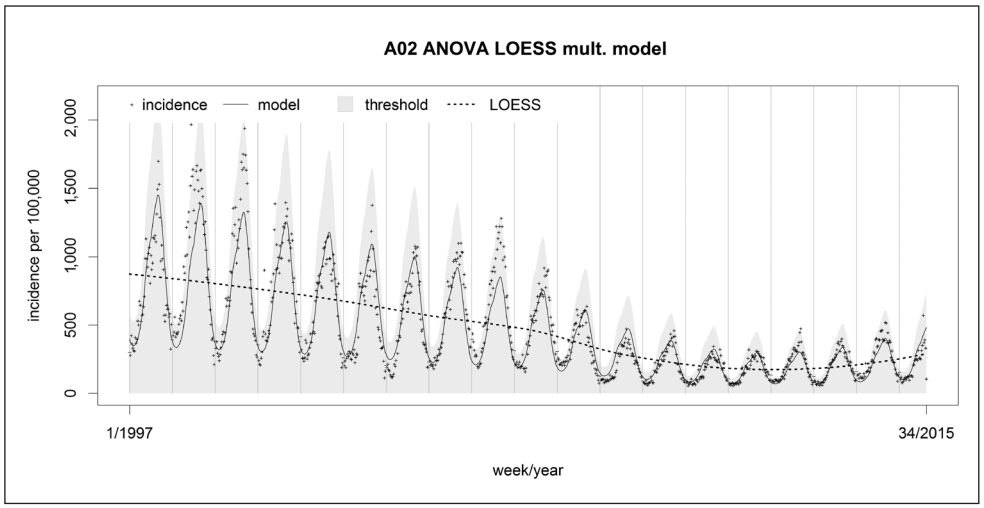


***Fig. 9.*** *ANOVA model with a multiplicative LOESS trend for the incidence of salmonellosis.*

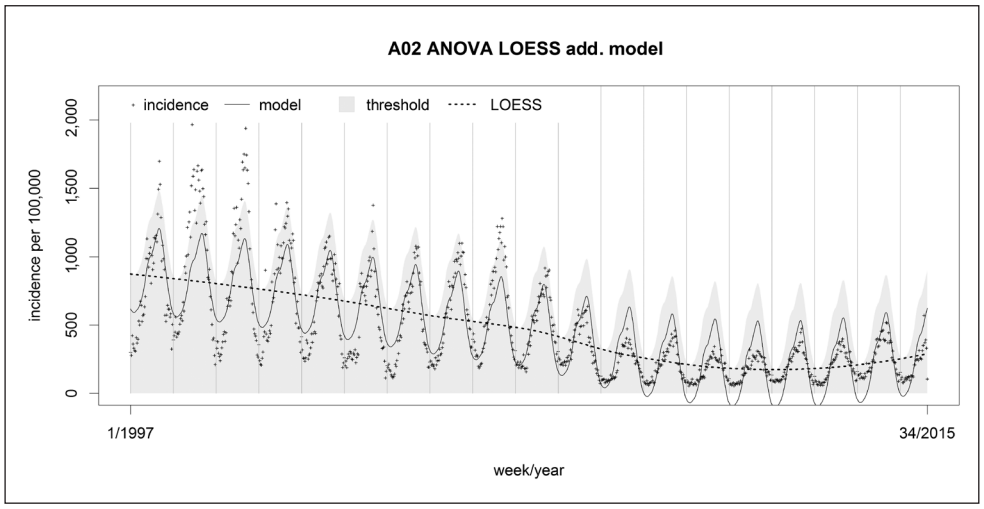


***Fig. 10.*** *ANOVA model with an additive LOESS trend for the incidence of salmonellosis.*

a long-term trend using a running mean is presented. A two-year window is used, but the LOESS estimate appears to be smoother and the model is closer to the data but at the expense of copying accidental influences, i.e. at the expense of lesser smoothness. The incidence was assumed again to have a log-normal distribution with a multiplicative long-term trend (Fig. 11).

Light smoothing or Serfling's high order model provide a description close to the real situation but has a lower predictive power and is less suitable for the determination of the epidemic threshold (is more dependent on accidental fluctuation) and thus also for issuing an alert for an excessive incidence.

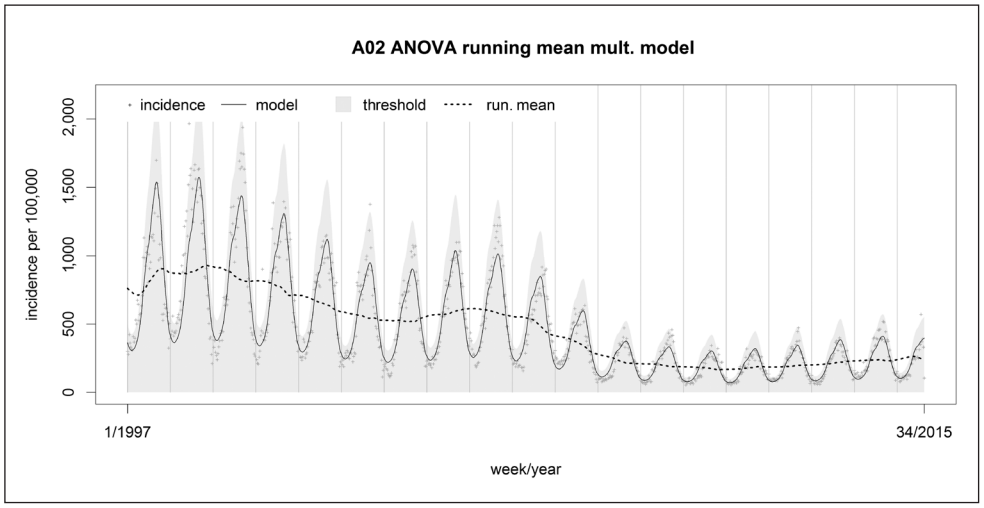


***Fig. 11.*** *ANOVA model with a multiplicative running mean (two-year window) trend for the incidence of salmonellosis.*

## CONCLUSION

This article features different methods for modelling the incidence of diseases but does not consider the effect of possible outliers (epidemics): this issue is beyond its scope and deserves separate attention.

In practice, the simplest model is often applied where one constant is considered as the threshold on a long-term basis (Fig. 1). This approach may not fit all diseases, as is the case with ILI whose incidence may differ between the summer and winter or with salmonellosis (Fig. 7) with typical seasonal fluctuations coupled with a long-term downward trend. A long-term trend may have even a more complicated shape. It also turns out that the integration of the polynomial is a worse approach than the running means of LOESS estimation since the polynomial tends to stretch to infinity at its extremities. Another consequence is that using the log-normal distribution is more suitable for modelling the incidence (the point is how many times it changes and not the size of the change) along with the multiplicative model. The ANOVA model seems to fit well and unlike the sine, to allow even for an asymmetrical period. This can be helpful e.g. in the context of ARI, characterized by a relatively high incidence in the winter, with a typical decline during the Christmas holiday due to unwillingness to go to see a doctor and to self-treatment attempts.

The main purpose of such modelling is to find a tool to predict the incidence and in particular, to identify the threshold, the exceedance of which indicates the excessive incidence and thus helps the epidemiologists, along with laboratory data, detect an epidemic.

**Acknowledgement**

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

## REFERENCES

1 Gail HM, Benichou J. Encyclopedia of epidemiologic methods. Chichester: John Wiley & Sons; 2000.
2 Kyncl J, Kriz B. Surveillance of acute respiratory infections in the Czech Republic and in Europe - example of an early warning system. In: Kocik J, Janiak M, Negut M, editors. Preparedness against bioterrorism and re-emerging infectious diseases. Amsterdam: IOS Press; 2004. p. 40-4.
3 Armitage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: John Wiley & Sons; 1999.
4 Kyncl J, Prochazka B, Havlickova M, Otavova M, Castkova J, Kriz B. Excess death attributable to influenza in the Czech Republic in 1982-2000. J Epidemiol Community Health. 2004 Aug;58 Suppl 1:A31-2.
5 Procházka B, Beneš Č. Evaluation of time trends in the weekly count of diseases. Epidemiol Mikrobiol Imunol. 1999 Apr;48(2):52-9. (In Czech.)
6 Procházka B. Biostatistics for physicians: Principles of basic methods and interpretation of results using the R statistical system. Praha: Karolinum; 2015. (In Czech.)
7 Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Rep. 1963 Jun; 78(6): 494–506.
8 Assaad F, Cockburn WC, Sundaresan TK. Use of excess mortality from respiratory diseases in the study of influenza. Bull World Health Organ. 1973; 49(3): 219-33.
9 Fleming DM, Cross KW, Watson JM, Verlander NQ. Excess winter mortality. Method of calculating mortality attributed to influenza is disputed. BMJ. 2002 Jun; 324(7349):1337.
10 Kyselý J, Kříž B. High summer temperatures and mortality in the Czech Republic in 1982-2000. Epidemiol Mikrobiol Imunol. 2003 Aug;52(3):105-16. (In Czech.)
11 Nicholson KG. Impact of influenza and respiratory syncytial virus on mortality in England and Wales from January 1975 to December 1990. Epidemiol Infect. 1996 Feb;116(1):51-63.
12 Owen J, Maillardet R, Andrew R. Introduction to scientific programming and simulation using R. New York: CRC Press; 2009.

*Received April 13, 2015*
*Accepted in revised form October 21, 2015*