Analysis of the seasonal incidence of acute respiratory infections including influenza (ARI) in the Czech Republic – possible contribution of the functional data boxplot in epidemiology

Ondrej Vencalek[a], Jan Kyncl[b]

Aims. The detection of an epidemic outbreak is possible only if the baseline incidence level of a given disease is well defined. The determination of the baseline is complicated by the presence of epidemic outbreaks in historical data. The aim of the paper is to provide a new way of determining the baseline.
Methods. The analyzed data containing weekly records on the incidence of acute respiratory infections including influenza (ARI) in the Czech Republic and its regions are taken from the nationwide surveillance system; data on 15 seasons from 2001/02 to 2015/16 are included. Functional boxplots of the data are constructed and five distinct methods (componentwise mean, componentwise median, median, trimmed mean, and adjusted mean) were used for the computation of the baseline level function.
Results. It was shown that the methods based on functional data analysis could successfully overcome the problems that arise when the conventional methods are used for the determination of the baseline function.
Conclusion. The functional boxplot – a new statistical tool – can bring not only a transparent visualisation of comprehensive data, but can also help epidemiologists and other public health experts to determine the baseline incidence level of a given disease as well as to detect unusual epidemic seasons.

Key words: epidemiology, acute respiratory infections, functional data analysis, boxplot, data depth

Received: February 28, 2017; Accepted with revision: September 20, 2017; Available online: October 17, 2017
https://doi.org/10.5507/bp.2017.042

[a]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacky University Olomouc, Czech Republic
[b]Department of Infectious Diseases Epidemiology, Centre for Epidemiology and Microbiology, National Institute of Public Health, Prague, Czech Republic
Corresponding author: Ondrej Vencalek, e-mail: ondrej.vencalek@upol.cz

INTRODUCTION

In recent years, a new area of statistics called functional data analysis (FDA) has been rapidly developing[1-3]. In the present paper, we show one possible application of FDA in epidemiology. The functions of interest for epidemiologists are the baseline incidence functions of various infectious diseases. Since the main goal is to detect epidemic outbreaks (unusually large numbers of cases), it is crucial to determine the baseline level of incidence (usual number of cases). This level varies in the course of the year for diseases like influenza, other acute respiratory diseases, and several others. Thus the baseline level is a function of time and can be analyzed by means of FDA.

The main complication connected to the determination of the baseline incidence is the presence of epidemic outbreaks. The reason is simple: since the baseline function should describe the usual level of incidence and the epidemic outbreak can be described as an unusually high level of incidence, the outbreaks should not be included in the estimation of the usual incidence; otherwise, the usual level will be overestimated. On the other hand, exclusion of epidemic outbreaks would lead to an underestimated baseline function. Both types of bias are undesirable.

To overcome the previously described problem, Prochazka and Kyncl have proposed taking the data from the period of the epidemic outbreak as censored and used techniques for the analysis of censored data to estimate the baseline incidence level[4]. Here we propose an alternative way to solve the problem of the baseline estimation.

The currently presented solution is based on a common statistical tool for describing what is a usual and what is an unusual value: the boxplot. The boxplot is a graphical tool for the visualisation of basic descriptive statistics of the data including the median, minimum, and maximum. The median defined as the middle value of the ordered data can be considered as a "usual" value. On the other hand, the minimal and maximal value can be called the extreme values or even outliers. The occurrence of the outliers is rare, unusual. The boxplot is commonly used for the detection of outliers. A detailed review of the boxplot is included in the Methods section.

A simple method for the determination of the baseline incidence would be based on the computation of the median separately for each week of the year. In the present paper, we call the baseline estimated by this method the componentwise median. This method is similar to the computation of the mean for each week of the year. While the mean is affected by the outliers, and afterwards the estimate is biased, using the median instead of the

mean solves this problem. However, the idea to employ the median for the baseline estimation can be enhanced by using FDA.

The awkwardness of the previously described technique lies in the fact that the median is computed separately for each week of the year. For a given week of a year, one can expect the incidence to be close to the median for most of the years. On the other hand, an incidence that is close to the median for all weeks in the year can hardly be expected. In other words, it is highly unlikely to find "usual values" during the whole year.

A possible solution to this problem is not to analyze data week by week separately, but jointly as functions. Thus, for each year, we obtain one realization of such a function. Historical data can be used to determine which function is typical in terms of level and shape. Similarly to the univariate data (numbers) which can be ordered from the smallest to the largest, the functional data (functions) can also be ordered, although not as intuitively as the univariate data since the functions usually crossing each other, which makes a straight comparison impossible. Luckily, statisticians have recently proposed a way to order the functional data by means of the so called functional data depth[5]. For each function, the depth can be computed, and the function with the highest depth value is the functional median of the data. Ordering also enables the construction of the boxplot for functional data and detection of outliers.

Respiratory virus activity is detected in Europe every winter, yet the precise timing and size of this activity are highly unpredictable[6]. Which age groups of the population are affected and how severe the illness depends on several factors, including the virus types and subtypes that circulate during a given season. The impact of influenza infection and/or acute respiratory infection in European countries is continuously monitored through a variety of surveillance systems[7]. All of these sources of information are used to assess the nature and extent of activity of influenza and other respiratory viruses and to offer guidance on the prevention and control of morbidity and mortality due to influenza at a local, national, and international level.

In this paper, we compare the baseline functions for the incidences of acute respiratory infections (ARI) in the Czech Republic and in its regions derived by five different methods: the componentwise mean, componentwise median, functional median, trimmed mean, and mean computed after the exclusion of outliers. While the first two methods are rather simple, the other three employ functional data analysis.

METHODS

Data

The analyzed data are taken from the nationwide surveillance system of acute respiratory infections including influenza (ARI) in the Czech Republic. The surveillance of ARI is based mainly on clinical surveillance (morbidity reports and mortality statistics of influenza and respiratory infections) and virological surveillance from the community and hospitals. The system now includes approximately 2230 general practitioners and 1240 pediatricians and covers approximately five million inhabitants (half of the Czech population) in all 86 districts of the Czech Republic[8]. The data contain weekly records from 15 seasons, from 2001/02 to 2015/16, and are available for the whole Czech Republic as well as for all 14 regions. One season is defined from calendar week 40 of the year to week 39 of the following year. To obtain a more realistic picture of the incidence, we did not use the data from the turn of the year (week 52 to week 1) since the reported incidence of ARI is low in these weeks.

Classical boxplot

Classical boxplot enables fast visual inspection of univariate data. An example of a boxplot can be seen in Fig. 1 (left). In this example, the univariate data consists of 15 records of ARI incidence in the Czech Republic in the first week of each of the 15 seasons (2001/02 – 2015/16).

Let us recall what is included in the boxplot and how the boxplot could be interpreted. The boxplot consists of the central box and radial whiskers with possible small circles (points) denoting outliers.

Let us start with the box. The thick line in the middle of the box denotes the median. Its value (slightly less than 1,000 per 100,000) can be found on the vertical axis. Let us notice that the boxplot might also be plotted horizontally, and there is no difference in meaning. From the median, one can say that the incidence in the considered week is usually about 1,000 per 100,000 (1%). The width of the box is determined by the lower quartile (denoted Q1) and the upper quartile (denoted Q3). Between these two values, one can find 50% of (typical) values. In the example considered, the lower quartile is about 900, and the upper quartile is slightly less than 1,150. The width of the box reflects the variability of the data. The width is known as the interquartile range (IQR).

The whiskers are determined by the minimal and maximal values of the data which are not considered as outliers. Their length also gives evidence of the variability in the data. Outliers are those data points with values more than 1.5 times IQR distant from the box, i.e. greater than Q3+1.5IQR or smaller than Q1-1.5IQR. In the example considered, there is one outlier whose value of 1,323 is higher than 1,064 + 1.5*166 = 1,312. The second highest value (1,136) is the maximal value which is not an outlier and, thus, determines the upper whisker.

Boxplot for functional data

The construction of the boxplot for functional data has been proposed quite recently[9]. The functional boxplot provides the same characteristics as the classical boxplot. An example of such a boxplot can be seen in Fig. 1 (right). The boxplot is based on 15 records where each record consists of the incidence development during one particular season.

The central "box" is coloured in grey. This is the region of 50% of the most typical functions in the data. Notice that the (vertical) width of the box is changing as

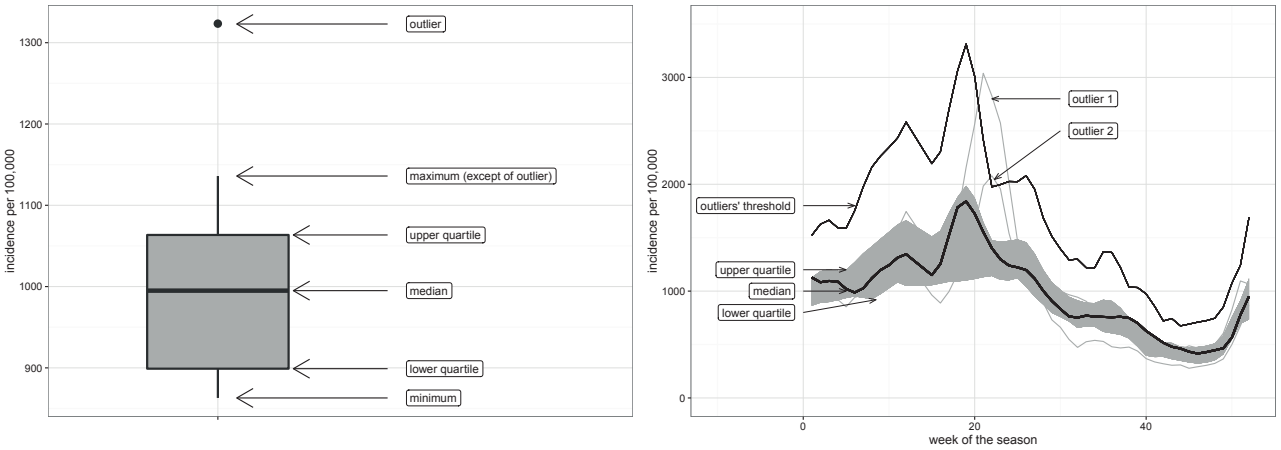

Fig. 1. Classical boxplot (left) and boxplot for functional data (right), as described in the section Boxplot for functional data.

the incidence varies in the course of the year. The highest variability is seen between weeks 10 and 25 of the season. That is just the period of occasional epidemic outbreaks. The thick black line inside the box corresponds to the most typical observation – the functional median.

The thin black line above the box outlines the border which is crossed only by outliers. Similarly to the common boxplot, it is defined by the upper bound of the box plus 1.5 times the width of the box. Therefore, its value, as well as its distance from the box, is changing during the season. When this line is crossed, the observation is considered to be an outlier. In the example presented, there were two outliers. Both are highlighted by light grey lines. In both cases, the epidemic outbreak occurred a few weeks later than usual (with peaks after week 20 of the season). In the epidemiological application suggested here, plotting the line for the detection of outliers with extremely small values can be avoided since such outliers are of no interest here.

Comparing methods for the determination of the baseline incidence

In the current paper, we compare five methods for the determination of the baseline incidence.

1) Componentwise mean – the mean value is computed separately for each week of the season. The method is believed to overestimate the usual level since the large values of epidemic outbreaks increase the mean.

2) Componentwise median – the median is computed separately for each week of the season. The value of the median is not influenced by outliers.

3) Functional median – the most typical one-year course of the incidence found in the data set is accounted for (in our case, one of the 15 years included in the data set).

4) Trimmed mean – years are ordered according to their depth (a measure of typicality), and subsequently, only the data from the most typical years (50%) are used for the computation of the componentwise mean. Thus, the extreme years are excluded and the mean is not affected by epidemic outbreaks.

5) Componentwise mean after exclusion of outliers – seasons that are outlying are detected by functional boxplot and are deleted before computing the componentwise mean.

We will also compare the "central regions" gained by the componentwise methods and functional data analysis. The quartiles included in classical boxplot determine "typical values" – after the exclusion of the smallest values (25%) and the highest values (25%), the remaining 50% of the data that can be considered as typical will be in the range determined by the quartiles. This range can be computed for each week separately. On the other hand, it is possible to order the functional data and construct the "central box" at one time. The comparison of these methods is also included in the present paper.

Technical remarks

Technical details on the functional boxplot can be found in[10]. In our study, we used functional spatial depth for ordering the data. The boxplot can be constructed for raw data without smoothing. However, the data preprocessing step consisting in smoothing may be done, but with caution (to avoid oversmoothing, particularly at the peak of an epidemic outbreak). Here we used smoothing by B-splines with penalization. The smoothing parameter was determined to optimize the generalized cross-validation criterion. The analysis was performed using the statistical software R (ref.[11]), specifically its packages fda and fda.usc.

RESULTS

Regional boxplots – description of regional specificities

The boxplot for the whole Czech Republic is shown in Fig. 1 (right) and can be compared to those for all 14 Czech administrative regions in Fig. 2. The basic shape of the curve for the incidence during the year can be roughly described as follows: it starts at a level of about 1,000 per 100,000 (1%), then it grows to reach a peak before week 20 of the season, which is followed by a long period of decrease (to a level of about only 500 per 100,000) and a short period at the end of the season when the incidence returns to a level of about 1,000 per 100,000. From the epidemiological point of view, the incidence of ARI, first

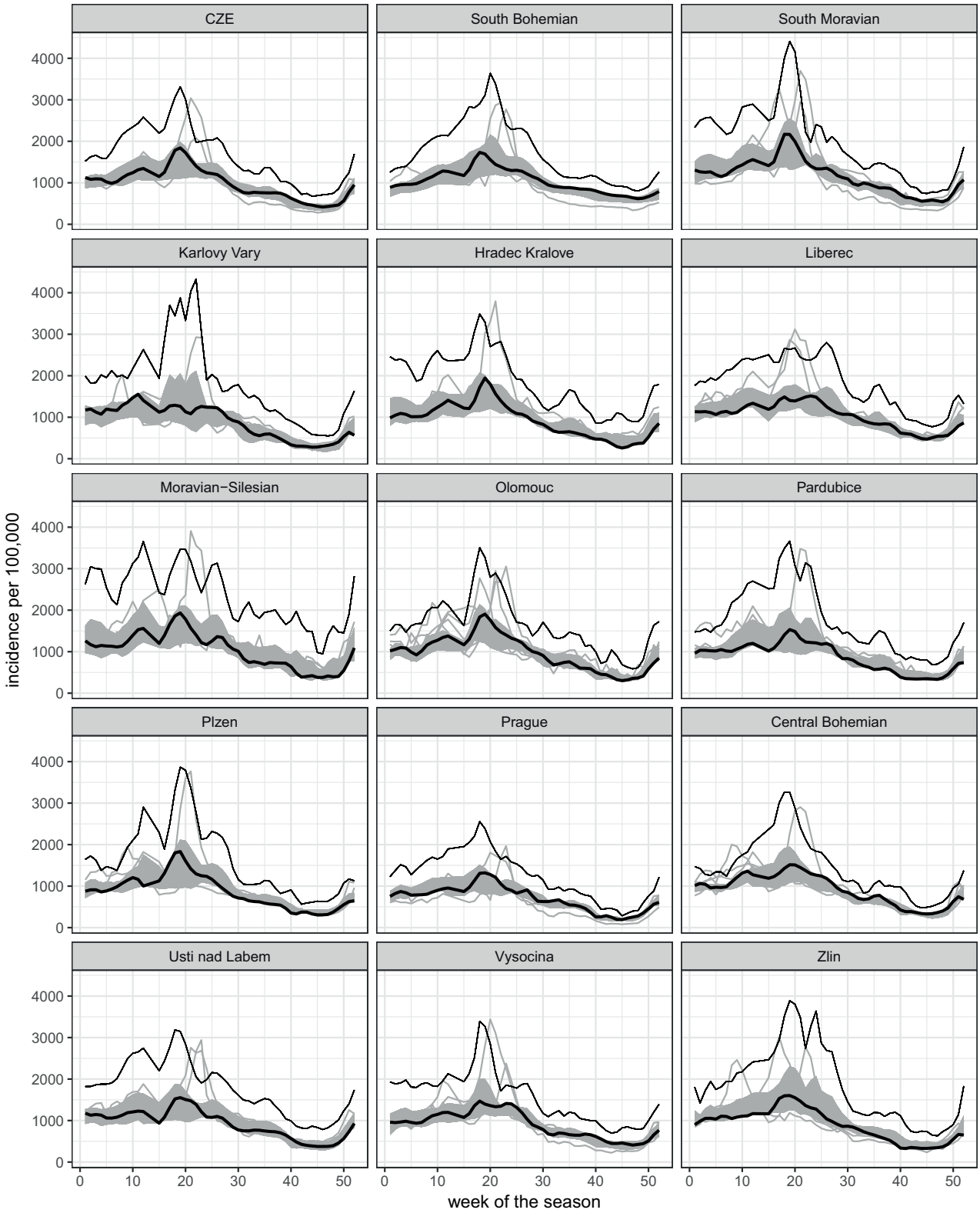

Fig. 2. Boxplots of ARI incidences in the Czech Republic and its regions, 2001/02−2015/16.

caused by non-influenza viruses, continuously increases. Then sporadic influenza cases occur and later give rise to an influenza epidemic that, in the Czech Republic, usually takes place in the second half of January and/or in February. There is a natural variability in time of onset of the outbreak. The influenza epidemic usually lasts 6-8 weeks. Its extent and severity depend on the circulating type (subtype) of the influenza virus.

Fig. 2 reveals regional differences in the incidence. The most remarkable difference is in the height of the expected peak of the incidence. For example, the maximal incidence higher than 2,600 can already be considered as an outlier in Prague, while at the national level, such values have to exceed 3,300 cases per 100,000 population or even 4,400 in the South Moravian Region. This is determined by several known and unknown factors. As

Table 1. The presence (denoted by 1) or absence (denoted by 0) of outliers in 15 seasons (2001/02-2015/16) in the Czech Republic and its 14 regions. The total numbers of outliers per season are indicated in the last row of the table.

Season	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16
CZECH REPUBLIC	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Region															
South Bohemian	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
South Moravian	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
Karlovy Vary	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
Hradec Kralove	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Liberec	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Moravian-Silesian	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Olomouc	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0
Paradubice	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Plzen	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Prague	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0
Central Bohemian	0	0	1	1	0	0	1	0	1	0	0	0	0	0	0
Usti nad Labem	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Vysocina	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
Zlin	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0
Total	**0**	**8**	**5**	**14**	**0**	**4**	**3**	**0**	**5**	**0**	**0**	**0**	**0**	**0**	0

an example, the morbidity in different regions might be also influenced by social determinants like the prevailing type of work or salary level.

Another difference is in the presence of a smaller peak (which can occur at the end of the calendar year) before the main peak. In some regions, the peak is expected (South Moravian, Karlovy Vary, Hradec Kralove, Moravian-Silesian, Olomouc, Pardubice, Plzen, and Usti nad Labem) while in others, it is rather uncommon (South Bohemian, Liberec, Prague, Central Bohemian, Vysocina, and Zlin). (This is evident from the shape of the outlier threshold line, which is a multiple of the upper quartile – the upper bound for the typical values.) There is no clear epidemiological explanation for this difference among regions.

The course of the curve in two regions – the Moravian-Silesian and Liberec Regions – differs from those in the rest of the regions by the absence of a sharp main peak. While in the first one, the variability is considerably high during the season with several expected peaks, the other region shows a small variability, and the peaks are also relatively small there.

Outliers

Two outliers were detected in the data from the whole Czech Republic: the seasons 2002/03 and 2004/05 were unusual. The courses of the incidence curves in those years are plotted in Fig. 1 (right). In both cases, the peak of the epidemic outbreak came 2 weeks later than expected.

In the season 2002/03, the incidence curves showed an outlying (unusual) course in seven out of 14 regions. In that season, the epidemic outbreak occurred later than expected in all regions. Although it was relatively strong, the same values would not be considered extreme if they occurred some two weeks earlier. From the statistical point of view, it is interesting to note that there was no excessive incidence early in that season, and it could even be considered as very low until the middle of the season. Therefore, the incidence in that season can be described as rather below average, but a relatively sharp peak occurred, later than usual. This can be explained, in part, by the co-circulation of two different influenza viruses.

A slightly different situation was observed in the season 2004/05, see Fig. 3. In that season, the incidence of ARI showed an unusual pattern in all but one region (Zlin Region). The outbreak not only occurred with a delay but also with an unexpected severity in several regions. Again, it is useful to note that the level of incidence in the first part of the season was not unusual in all regions. The same can be said for the part of the season following the outbreak. This provides evidence of difficulty in predicting an unusual epidemic outbreak. It should be noted that the length of the outbreak can also be influenced by region-specific variation in the dates of the one-week spring school holidays that results in a reduced incidence but a prolonged length of the outbreak.

In some regions, unusual situations (occurrence of outliers) were also detected in the seasons 2003/04 (five regions), 2006/07 (four regions), 2007/08 (three regions), and 2009/10 (five regions). The occurrence of outliers is summarized in Table 1. It is useful to study these phenomena.

Finally, it should be mentioned that in the season 2009/10, the time course of the incidence was really atypical in the regions denoted as outliers. The main epidemic peak occurred as early as before week 10 of the season (before the end of the calendar year). It was unusually

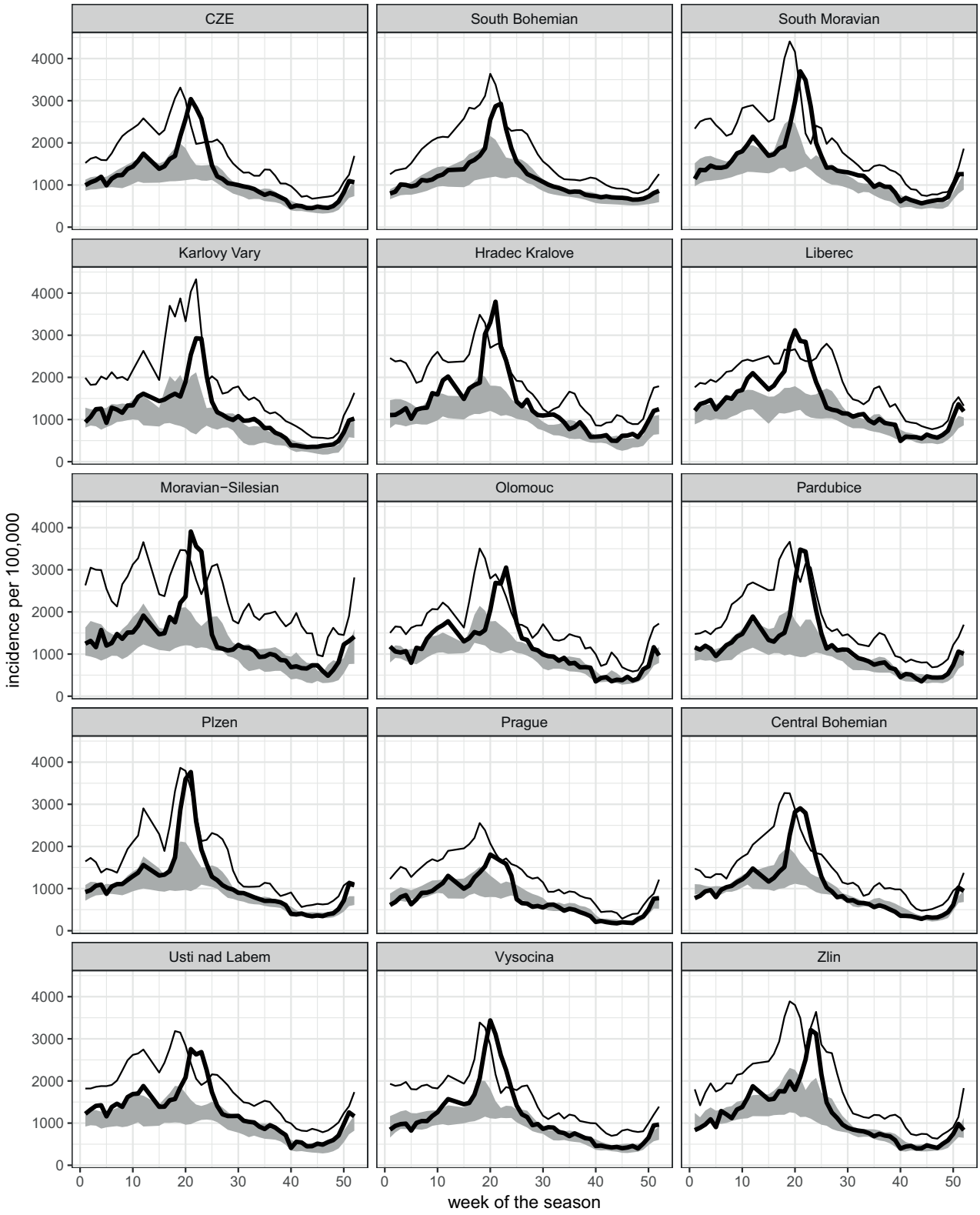

Fig. 3. Visualisation of ARI time course in the Czech Republic and its regions during the season 2004/05 (thick line) compared to the functional boxplot where the usual values fall into the grey area and the outlying values are above the thin line.

high for that period of the year but smaller than the usual yearly maximum. The usual peak observed before week 20 was absent. There is a clear epidemiological explanation for this phenomenon: the influenza pandemic that affected the world in 2009.

Comparison of the methods for determining the baseline incidence level

Fig. 4 (top) compares three methods for the determination of the baseline incidence level – the componentwise mean (cw.mean), componentwise median (cw.med), and functional median (median).

When comparing the two componentwise methods, it can be seen that the median is slightly smaller during almost the whole first half of the season. This corresponds to the skewness of the data (the presence of few high values). In the second half of the season, these two methods give practically identical results. An interesting difference can be found in the most important part of the year which contains the peak of the epidemic outbreak. While the mean indicates one peak of the outbreak, the median suggests two possible peaks.

The functional median (median) differs quite a lot from the two componentwise methods, mainly at the time of an epidemic outbreak, characterized by much larger values. This result should be interpreted with caution since the number of data for the estimation of the functional median is small (only 15 functional data points), and thus the estimation is unstable. However, this result might raise the question whether the standard componentwise methods simply underestimate the usual incidence level during the period of occasional epidemic outbreaks.

Fig. 4 (middle) compares three componentwise methods differing in the way the mean value is computed. The classical mean is compared to the trimmed mean (trim. mean) and adjusted mean (adj. mean) which is computed after the exclusion of outlying observations.

There is no substantial difference between the classical mean and adjusted mean for the major part of the year. However, there is a not negligible difference between these two methods at the time of epidemic outbreaks. In that case, the adjusted mean gives smaller values since the outliers, which make the common mean higher, are excluded.

The trimmed mean differs more from the classical mean. Its behaviour in the first part of the season (to week 15) resembles that of the componentwise median – its values are systematically lower than those of the mean. However, then the increase of the trimmed mean values is steeper than in the case of the classical mean, and the maximum is even slightly higher for the trimmed mean. Both the adjusted and trimmed means show a sharper peak than the classical mean.

There is another interesting comparison of the results gained by the componentwise methods and functional data analysis. Fig. 4 (bottom) compares the widths of the "central regions" gained by the componentwise and functional methods.

One can see that the central band based on FDA (coloured in dark grey) is wider than the band computed for each week separately (dashed line). This indicates an underestimation of the variability when using the classical componentwise methods.

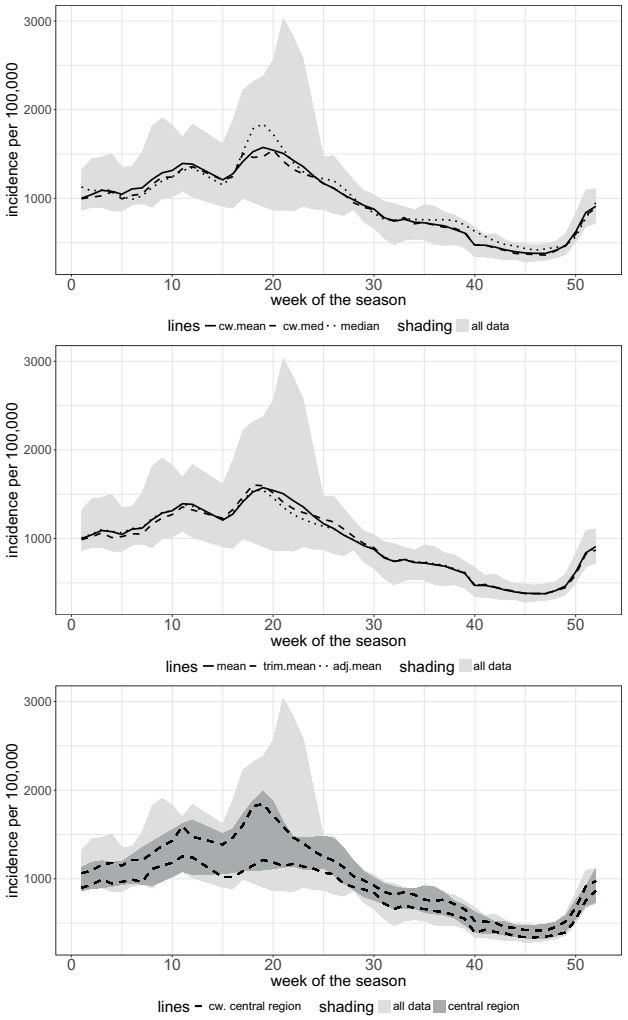

Fig. 4. Comparison of the methods for the detection of the baseline incidence: componentwise mean, componentwise median, and median (left), componentwise mean, trimmed mean, and adjusted mean (middle); comparison of the 50% central regions gained by the componentwise and functional methods (right).

CONCLUSION

Information on the incidence of infectious diseases and its precise analysis is very important for maintaining public health in Europe. In this paper, we aimed to introduce the functional data analysis (FDA) – a relatively new statistical tool – to the broader medical community.

The functional boxplot – one of the tools of FDA – was used for the detection of an unusual time course of the incidence of acute respiratory infections. The analysis of the detected outliers revealed some regional similarities. This might inspire further research to look for clusters of functional shapes and to define the possible time courses of the incidence during the year. Three new methods for the determination of the baseline incidence level were proposed – the functional median, trimmed mean, and adjusted mean. Comparison of the new methods confirmed some of the expected effects and revealed some unexpected effects like underestimation of the variability by the classical componentwise methods.

Acknowledgement: The work of Ondrej Vencalek is supported by grant No. GA15-06991S from the Czech Science Foundation. The work of Jan Kyncl is partially supported by MH CZ - DRO (National Institute of Public Health – NIPH, IN 75010330).

Author contributions: OV: statistical analysis; JK: discussion of results, epidemiological point of view.
Conflict of interest statement: None declared.

REFERENCES

1. Sorensen H, Goldsmith J, Sangalli LM. An introduction with medical applications to functional data analysis. Stat Med 2013;32(30):5222-40.
2. Wang J-L, Chiou J-M, Muller H-G. Functional data analysis. Annu Rev Statist 2016;3(1):257-95.
3. Ramsay JO, Silverman BW. Functional Data Analysis. 2nd ed. New York: Springer; 2005.
4. Prochazka B, Kyncl J. Estimating the Baseline Incidence of a Seasonal Disease Independently of Epidemic Outbreaks. Cent Eur J Public Health 2016;24(3):199-205.
5. Lopez-Pintado S, Romo J. On the concept of depth for functional data. JASA 2009;104(486):718-34.
6. Kyncl J, Havlickova M, Nagy A, Jirincova H, Piskova I. Early and unexpectedly severe start of influenza epidemic in the Czech Republic during influenza season 2012-13. Euro Surveill 2013 Feb 7 [cited 2017 Feb 28]; 18(6). Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20396.
7. European Centre for Disease Prevention and Control (ECDC)/World Health Organization Regional Office for Europe (WHO/Europe). Flu News Europe. System description. ECDC/WHO. [cited 2017 Feb 28]. Available from: http://flunewseurope.org/System.
8. Kyncl J, Paget WJ, Havlickova M, Kriz B. Harmonisation of the acute respiratory infection reporting system in the Czech Republic with the European community networks. Euro surveill. 2005 Mar 1 [cited 2017 Jun 26]; 10(3). Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=525
9. Sun, Y, Genton, MG. Functional boxplots. J Comput Graph Statist 2011;20:316-34.
10. Serfling R, Wijesuriya U. Depth-based nonparametric description of functional data, with emphasis on use of spatial depth. Comput Stat Data Anal 2017;105:24-45.
11. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.