



**Joanna Trzęsiok**

University of Economics in Katowice, Faculty of Finance and Insurance, Department of Economic and Financial Analysis, [joanna.trzesiok@ue.katowice.pl](mailto:joanna.trzesiok@ue.katowice.pl)

## Outliers vs Robustness in Nonparametric Methods of Regression

**Abstract:** The article addresses the question of how robust methods of regression are against outliers in a given data set. In the first part, we presented the selected methods used to detect outliers. Then, we tested the robustness of three nonparametric methods of regression: PPR, POLYMARS, and RANDOM FORESTS. The analysis was conducted applying simulation procedures to the data sets where outliers were detected. Contrary to a relatively common conviction about the robustness of nonparametric regression, the study revealed that the models built on the basis of complete data sets represent a significantly lower predictive capability than models based on the sets from which outliers were discarded.

**Keywords:** outliers, robustness, nonparametric regression methods

**JEL:** C14

## 1. Introduction

The assumption of the homogeneity of a given data set is one of key assumptions in regression analysis. Its adoption means that we treat data used for analysis as a set of observations coming from the same population. In data sets, however, especially real data sets, there may be data points that are distant from other observations. They require particular attention as they may cause the model based on such a data set to be inappropriate for the analysed phenomenon. Accordingly, it is highly likely that inference, prediction and decision making based on such a model will be erroneous.

Robustness is another complex problem. In most general terms, the application of a robust regression method means that we have a model that follows a tendency manifested by the majority of observations. The robustness of regression, however, may be approached from a number of angles.

A regression method can be robust to:

- 1) the occurrence, in a training set, of *distant (outlying) points* which may disturb and significantly alter the equation of the regression function;
- 2) *random disturbances in the value of a dependent variable* (e.g.: random measurement errors with a normal distribution);
- 3) the occurrence, in a training set, of *insignificant variables* that do not have an impact on the model and the value of a dependent variable;
- 4) *sampling* of a training set that is the basis for the construction of a given model;
- 5) *the lack of values* of some variables in a training set;
- 6) the method *falling short of expectations*.

While referring to robustness of regression, we tend to equal it with the insensitivity of the model to the quality of data, so – primarily – with the presence of distant (outlying) observations in a training set. They may be a result of the disturbances in the value of both a dependent variable and explanatory variables. This is the context in which we will discuss the robustness of selected regression methods presented in the article

It attempts to identify distant observations using three criteria: Ward's cluster analysis, multidimensional scaling, and the Mahalanobis distance amended by Filzmoser, Maronna and Werner (2008). While the method applying the Mahalanobis distance to outlier detection is quite commonly used, the approach based on taxonomic analysis and multidimensional scaling is the author's original idea. However, the main goal of the article was not to identify outliers, but to verify the hypothesis about the robustness of nonparametric regression methods to the occurrence of outliers.

## 2. Outliers and their identification

The notion of an outlier does not have a single unequivocal definition in the literature. On the contrary, it is defined in many ways. This article adopts the definition proposed by Hawkins (1980), who argues that an outlier is “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism”.

In terms of the causes of their occurrence, outliers can be divided into (Rousseeuw, Leroy, 2003):

- 1) outliers originating from a number of different types of errors: measurement errors, errors involved in data collection and entry, deliberate dishonesty in reporting, unsuitable research methodology, poor sampling, or wrong assumptions;
- 2) outliers for the heavy-tailed distribution;
- 3) influential observations which have a significant impact on a given model and may lead to interesting hypotheses.

The detection of outliers and the ways of handling them are important issues related to the notion of robustness in statistics (Trzpiot, 2013). The literature provides many approaches to the identification of outliers. The most popular ones are: a one-dimensional quantile criterion (Tukey, 1977), methods based on Cook’s distance (Cook, 1977), estimates based on the Mahalanobis distance (Healy, 1968), and the method involving the local outlier factor (Breunig, Kriegel, Ng, Sander, 2000).<sup>1</sup>

A number of researchers showed interest in the topics related to outliers and non-parametric regression. Outliers detection and identification were, for example, discussed by:

- 1) Majewska (2015), who, apart from classical methods, uses non-traditional methods based on robust PCA in her work;
- 2) Batóg (2016), whose work is based on the comparison of methods that enable the identification of spatial outliers;
- 3) Ganczarek-Gamrot (2016), who used electricity market data to present methods for detecting outliers within time series;
- 4) Trzęsiok (2014), who discussed outliers in the context of data quality.

In the context of robust regression, on the other hand, applications and comparisons of various robust methods, with particular emphasis on the regression depth concept, were proposed by Kosiorowski (e.g.: 2007; 2012).

This article uses three criteria.

1. **Criterion based on the Mahalanobis distance** (Healy, 1968):

$$MD(\mathbf{x}) = \sqrt{(\mathbf{x} - \hat{\boldsymbol{\mu}}) \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}})^T} \quad (1)$$

<sup>1</sup> The criteria for outlier detection were discussed in detail, *inter alia*, in Trzęsiok (2014).

where  $\hat{\boldsymbol{\mu}}$  is a mean value, while  $\hat{\boldsymbol{\Sigma}}$  – a variance-covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x} - \hat{\boldsymbol{\mu}})^T (\mathbf{x} - \hat{\boldsymbol{\mu}}) \quad (2)$$

According to this criterion, we treat an observation as an outlier when it is matched by a high value of  $MD(\mathbf{x})$  compared to critical values in  $\chi^2$  distribution tables.

The major weakness of this method is that it draws on classic statistics, which are very sensitive to outliers and, in consequence, the values of the measure  $MD$  cannot always be deemed as reliable. For this reason, the literature proposes many modifications in the Mahalanobis distance. One of such modifications is the  $MD^*$  approach, developed by Filzmoser, Maronna, and Werner in 2008, which applies principal component analysis to outlier detection. This method is presented in detail in (Filzmoser, Maronna, Werner, 2008).

2. **Ward's method**, or hierarchical cluster analysis, is one of the agglomerative methods which is the most frequently applied and which yields the best results. It involves the successive merging of clusters into increasingly larger ones. The way that the method works (just as in the case of all hierarchical methods) can be represented with a dendrogram, which allows the reconstruction of the classification process. A dendrogram also enables the visualisation and graphic representation of the results of clustering. Hierarchical methods, including Ward's method, were discussed in (Walesiak, Gatnar, 2009).

The application of clustering methods to outlier detection has attracted criticism in the literature (Breunig, Kriegel, Ng, Sander, 2000), due to their other – primary – goal. However, in this case, we intended to apply a few criteria, in a way complementary and enabling the visualisation of multi-dimensional observations.

3. **Multidimensional scaling** is a method that allows the visualisation of relations between individual cases in a data set. It involves transforming original observations to space that has fewer dimensions (most frequently 2 or 3 dimensions), so that the distances between the objects in a new coordinate system are possibly the closest to the original distances between the relevant observations. This enables the identification of outliers in fewer dimensional space (e.g.: two-dimensional). This method also has the advantage of being able to generate the graphic representation of analysis results. Multidimensional scaling was presented in more detail in, for example, (Walesiak, Gatnar, 2009).

Outlier detection is not a simple task. Moreover, it is only the first step in the analysis. Outliers are not always a negative occurrence. They may result from a measurement error, yet they may also be influential observations, which should

not be removed from a data set, since they may carry meaningful and potentially useful information. However, the discovery of the nature of an observation is a complex and difficult task, so the right decision seems to be preserving outliers in a data set and applying robust statistical methods for further analysis. The question arises which methods are robust to the occurrence of outliers in a data set.

### 3. Regression methods used in the study

Robustness is of particular importance in the case of nonparametric regression models, which are characterised by high flexibility and the capacity for an adaptive and precise fit to data, accounting for variability caused by disturbances. The question arises how nonparametric models built on training sets disturbed by outliers behave.

In view of the above, nonparametric methods may generate models that are not robust to the occurrence of outliers in training sets, have poor predictive capabilities, and, as a result, do not hold a substantive cognitive value for researchers. On the other hand, however, many of these methods have an in-built regularisation mechanism which reduces the problem of the overfitting of a model to a training set. The mechanism involves adopting a certain compromise between the fit of a model and its complexity (Trzęsiok, 2011), which results in the increased predictive capabilities of the model. The question, however, arises to what extent the mechanism is effective and whether the methods are really robust to outliers.

The study used three selected nonparametric methods that are frequently applied in comparative analyses and possess good predictive capabilities (Meyer, Leisch, Hornik, 2003):

- 1) projection pursuit regression PPR (Friedman, Stuetzle, 1981),
- 2) multivariate adaptive regression splines POLYMARS (Kooperberg, Bose, Stone, 1997),
- 3) random forests (Breiman, 2001).

### 4. Research procedure

As mentioned above, the study aimed not only to detect outliers in data sets but also to test nonparametric methods for robustness to the occurrence of such observations. Accordingly, the analytical procedure applied in the study can be presented in the following steps:

1. Outlier detection:
  - the three outlier detection criteria presented above were used to analyse the data sets,

- then, the majorisation rule was applied to classify as outliers the observations that were detected as such according to all three criteria.
2. The construction of nonparametric regression models:
    - based on the entire original data set,
    - based on the data set from which outliers were eliminated.
  3. The comparison of the models in terms of their predictive capabilities, using the mean squared error  $MSE_{CV}$  calculated with the cross-validation method (involving the breakdown of a data set into 10 parts).

The robustness of the selected regression methods to the occurrence of outliers in a training set was tested on three data sets:

- 1) *crime*, proposed in (Agresti, Finlay, 2009); it is a set of real data on criminal activity in the US states (51 observations); it contains three outliers;
- 2) *hbk*, presented in (Rousseeuw, Leroy, 2003); it is a computer generated data set, containing 75 observations, 14 of which are outliers;
- 3) *flats* is a set of real data generated based on the information about sale transactions of flats provided by the online service [www.oferty.net](http://www.oferty.net); the data concern sale transactions completed from June 2007 to September 2009; the *flats* data set contains 747 observations described by 8 explanatory variables (5 of which are variables measured in interval or ratio scales)<sup>2</sup>.

Ward's method does not field unequivocal identifications (apart from detecting the object DC – District Columbia) of outliers in the *crime* set. Multidimensional scaling detects 3 outliers, whereas the  $MD^*$  method – Mahalanobi distance amended by Filzmoser, Maronna and Werner (2008) – identifies 4 such *observation* points. The final conclusion is that the following states are outliers: MS (Mississippi), DC (District Columbia) and LA (Louisiana).

In the case of the *hbk* set, all three criteria indicated that the first 14 observations in the set were outliers.

In the *flats* set, Ward's method showed 23 outliers (they belong to the smallest of the classes created as a result of breaking down the set into 8 groups in accordance with the silhouette index). Multidimensional scaling identified 31 such observations, while the Mahalanobis distance amended by Filzmoser, Maronna and Werner (2008) – 68 outliers.

As mentioned above, we conducted the two variations of analysis for each set. First, the model was built based on the set containing the outliers, then the outliers were removed and the new model was constructed. In each case (for each set and each regression method), we cross-validated the mean squared error  $MSE_{CV}$ . The results are presented in Table 1.

---

<sup>2</sup> As for the *flats* set, we do not know the number of outliers because it is a set of real data. There is, therefore, no objective way to detect all outliers and different methods yield different results.

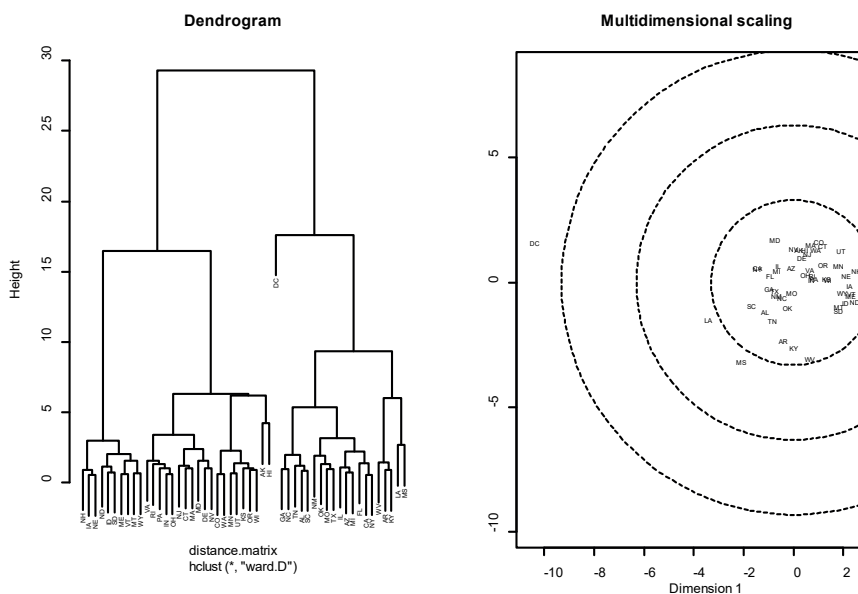


Figure 1. The dendrogram for Ward’s method and the visualisation of multidimensional scaling for the *crime* set

Source: own computation

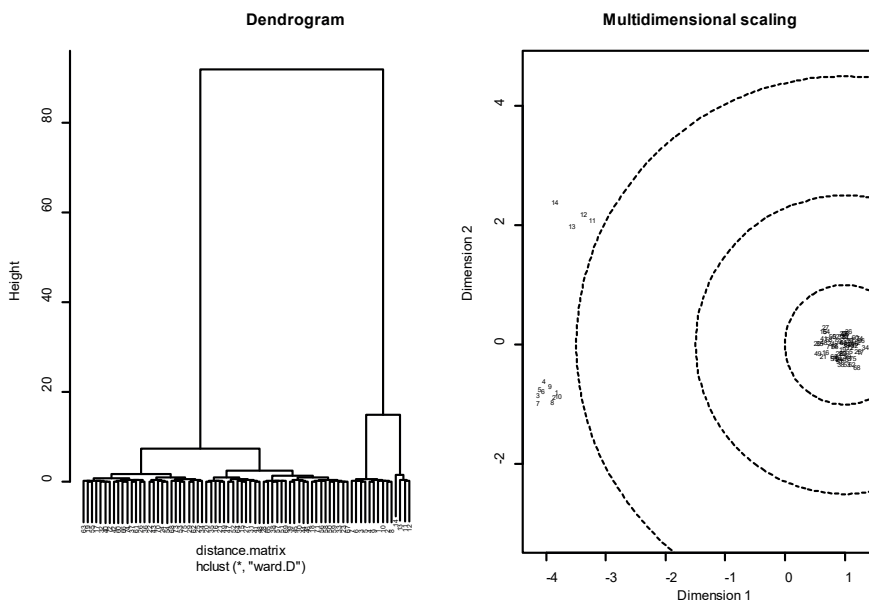


Figure 2. The dendrogram for Ward’s method and the visualisation of multidimensional scaling for the *hbk* set

Source: own computation

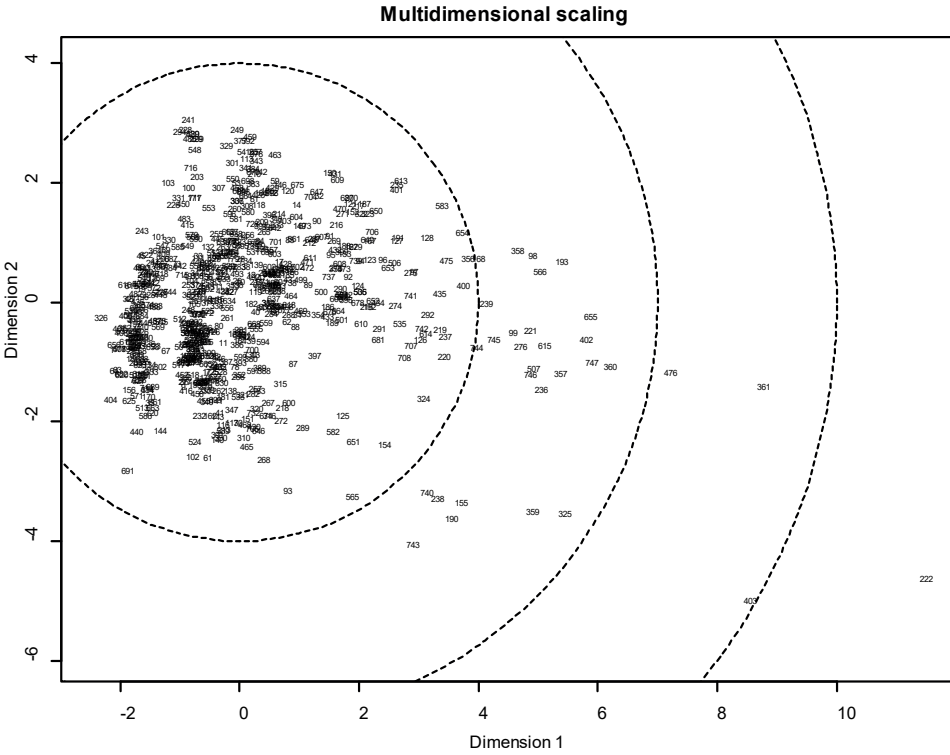


Figure 3. The visualisation of multidimensional scaling for the flats set

Source: own computation

Table 1. The values of the mean squared errors  $MSE_{CV}$ , calculated for different regression models built on the data sets with and without outliers

Methods	Data sets					
	<i>crime</i>	<i>crime without outliers</i>	<i>hbk</i>	<i>hbk without outliers</i>	<i>flats</i>	<i>flats without outliers</i>
PPR	78 236	31 311	2.72	0.29	11 321	3 566
POLYMARS	109 334	29 628	1.74	0.33	10 348	3 275
R.FORESTS	61 893	21 669	0.81	0.22	8 037	1 804

Source: own computation

While analysing the results for particular methods, presented in Table 1, we should compare the pairs of  $MSE_{CV}$  values obtained for the models constructed based on:

- 1) the set containing outliers, and
- 2) the set from which outlier were removed.



It is not important which model adopts the lowest values of  $MSE_{CV}$ , but how these values (in corresponding pairs) change as a result of removing outliers. Comparing figures in columns 2 and 3, 4 and 5, as well as 6 and 7 in Table 1, we can observe that in each case there was a relatively large decrease in the value of the mean squared error, which means that none of the methods under consideration is robust to the occurrence of outliers in a training set.

## 5. Conclusion

The article presents selected outlier detection methods which enable the preliminary analysis of a data set and, as a result, can bring certain anomalies occurring in the set to a researcher's attention. However, we cannot be certain that these methods will detect all outliers in real data sets.

It is also worth emphasising that the occurrence of outliers does not mean the immediate necessity to remove them from a data set. On the contrary, they may have a significant but positive influence on a given model. Therefore, a good solution is to apply robust methods to the analysis of such a data set. This study tested three nonparametric regression methods – PPR, POLYMARS and RANDOM FORESTS – for robustness to outliers.

The studies on the topics related to outliers mentioned in Part 2 focused primarily on the identification and detection of these observations. This article was only the initial stage of the study, as it aimed to examine the properties of selected regression methods that are commonly considered robust. The results of the examination, however, clearly show that the selected regression methods adopt significantly lower values of the mean squared errors  $MSE_{CV}$  after the removal of outliers from the data sets. Thus, the research hypothesis proposed in the introduction was verified negatively and rejected. These nonparametric regression methods cannot be considered robust to the occurrence of outlying observations in a training set.

## References


- Agresti A., Finlay B. (2009), *Statistical Methods for the Social Sciences*, 4th ed., Pearson, New Jersey.
- Batóg J. (2016), *Identyfikacja obserwacji odstających w analizie skupień*, [in:] K. Jajuga, M. Walesiak (eds.), *Taksonomia 26. Klasyfikacja i analiza danych*, “Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu”, no. 426, pp. 13–21.
- Breiman L. (2001), *Random Forests*, “Machine Learning”, no. 45, pp. 5–32.
- Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. (2000), *LOF: Identifying Density-Based Outliers*, Proceedings of the 29th ACM SIGMOD International Conference on Management of Data (SIGMOD 2000), Dallas.
- Cook R.D. (1977), *Detection of Influential Observations in Linear Regression*, “Technometrics”, no. 19(1), pp. 15–18.
- Filzmoser P., Maronna R.A., Werner M. (2008), *Outlier Identification in High Dimensions*, “Computational Statistics & Data Analysis”, no. 52, pp. 1694–1711.
- Friedman J., Stuetzle W. (1981), *Projection Pursuit Regression*, “Journal of the American Statistical Association”, no. 76, pp. 817–823.
- Ganczarek-Gamrot A. (2016), *Obserwacje odstające na rynku energii elektrycznej*, “Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach”, no. 288, pp. 7–20.
- Hawkins D. (1980), *Identification of Outliers*, Chapman and Hall, London.
- Healy M.J.R. (1968), *Multivariate Normal Plotting*, “Applied Statistics”, no. 17, pp. 157–161.
- Kooperberg C., Bose S., Stone C. (1997), *Polychotomous Regression*, “Journal of the American Statistical Association”, no. 92, pp. 117–127.
- Kosiorowski D. (2007), *O odpornej analizie regresji w ekonomii na przykładzie koncepcji głębi regresyjnej*, “Przegląd Statystyczny”, vol. 54, pp. 109–121.
- Kosiorowski D. (2012), *Statystyczne funkcje głębi w odpornej analizie ekonomicznej*, Wydawnictwo UEK w Krakowie, Kraków.
- Majewska J. (2015), *Identification of Multivariate Outliers – Problems and Challenges of Visualization Methods*, “Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach”, no. 247, pp. 69–83.
- Meyer D., Leisch F., Hornik K. (2003), *The Support Vector Machine under Test*, “Neurocomputing”, vol. 1–2, no. 55, pp. 169–186.
- Rousseeuw P., Leroy A. (2003), *Robust Regression and Outlier Detection*, John Wiley & Sons Inc., New York.
- Trzęsiok J. (2011), *Przegląd metod regularyzacji w zagadnieniach regresji nieparametrycznej*, [in:] K. Jajuga, M. Walesiak (eds.), *Taksonomia 18. Klasyfikacja i analiza danych*, “Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu”, no. 176, pp. 330–339.
- Trzęsiok M. (2014), *Wybrane metody identyfikacji obserwacji oddalonych*, [in:] K. Jajuga, M. Walesiak (eds.), *Taksonomia 22. Klasyfikacja i analiza danych – teoria i zastosowania*, “Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu”, no. 327, pp. 157–166.
- Trzpiot G. (ed.) (2013), *Wybrane elementy statystyki odpornej*, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.
- Tukey J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Boston.
- Walesiak M., Gatnar E. (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa.

## Obserwacje odstające a problem odporności

**Streszczenie:** Artykuł poświęcony jest zagadnieniu odporności metod regresji na obserwacje odstające występujące w zbiorze danych. W pierwszej części przedstawiono wybrane metody identyfikacji obserwacji nietypowych. Następnie badano odporność trzech nieparametrycznych metod regresji: PPR, POLYMARS i RANDOM FORESTS. Analiz dokonano za pomocą procedur symulacyjnych na zbiorach danych, w których wykryto obserwacje odstające. Mimo dosyć powszechnych przekonań o odporności regresji nieparametrycznej okazało się, że modele zbudowane na całych zbiorach danych mają istotnie mniejsze zdolności predykcyjne niż modele uzyskane na zbiorach, z których usunięto obserwacje nietypowe.

**Słowa kluczowe:** obserwacje odstające, odporność, nieparametryczne metody regresji

**JEL:** C14

	<p>© by the author, licensee Łódź University – Łódź University Press, Łódź, Poland.          This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license CC-BY          (<a href="http://creativecommons.org/licenses/by/3.0/">http://creativecommons.org/licenses/by/3.0/</a>)</p>
	<p>Received: 2016-12-17; verified: 2018-04-11. Accepted: 2018-06-18</p>