Biodiversity and Conservation (2018) 27:3295–3300
https://doi.org/10.1007/s10531-018-1602-2

COMMENTARY

A conceptual framework to deal with outliers in ecology

Jacinto Benhadi-Marín[1,2]

Received: 24 April 2018 / Revised: 18 July 2018 / Accepted: 30 July 2018 / Published online: 1 August 2018

Abstract
Research on ecology commonly involves the need to face datasets that contain extreme or unusual observations. The presence of outliers during data analysis has been of concern for researchers generating a lot of discussion on different methods and strategies on how to deal with them and became a recurrent issue of interest in debate forums. Systematic elimination or data transformation could lead to ignore important ecological processes and draw wrong conclusions. The importance of coping with extreme observations during data analysis in ecology becomes clear in the context of relevant environmental aspects such as impact assessment, pest control, and biodiversity conservation. In those contexts, misinterpretation of results due to an incorrect processing of outliers may difficult decision making or even lead to failing to adopt the best management program. In this work, I summarized different approaches to deal with extreme observations such as outlier labeling, accommodation, and identification, using calculation and visualization methods, and provide a conceptual workflow as a general overview for data analysis.

Keywords Extreme values · Data analysis · Environment · Conservation

A widespread method of inference in ecology relies on statistical hypothesis testing using the mean and considering the standard deviation as a measure of the normal variation of the studied processes. In this context, the likelihood of type-I and type-II error can be significantly increased by extreme observations on the left and/or right side of the data distribution altering the chances that the means to be compared have to be dissimilar, or if both low and high deviant observations make the standard deviation increase (Cousineau and Chartier 2010).

Data collected by or available for researchers in ecology usually contain unusual or deviant observations that behave as extreme values, the so-called outliers. Generally speaking, an outlier is an observation that differs so much from other observations as to arouse

Communicated by Dirk Sven Schmeller.

✉ Jacinto Benhadi-Marín
jbenma@hotmail.com

[1] Centro de Investigação de Montanha (CIMO), Escola Superior Agrária, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

[2] Department of Life Sciences, Centre for Functional Ecology, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal

suspicion that it was generated by a different mechanism (Hawkins 1980). Cook (1979) considered an observation as being influential if "important features of the analysis are altered substantially when the observation is deleted" and those observations are candidates that may cause model misspecification, biased parameter estimation, and incorrect results (Ben-Gal 2005). In addition, a false outlier is a data point which occurs naturally in the population but might even be identified by statistics as an outlier (Iglewicz and Hoaglin 1993; Rousseeuw and Hubert 2011).

In univariate regression analyses, outliers are commonly considered as points lying over three standard deviations from either side of the regression line and having large residuals (Wiggins 2000), and multivariate outliers are cases with an unusual combination of scores on two or more variables (Tabachnick and Fidell 1996). The main question towards data analysis and interpretation is whether or not such points actually represent influential observations and whether they should be dropped from the dataset or not, since although outliers are often considered as an error or noise, they may carry relevant ecological information. There is a huge amount of technical information on how to deal with outliers among literature mostly provided by books focused on data analysis. Here, an easy-to-follow conceptual workflow for extreme observation management during data exploration is provided by gathering different approaches following Cleveland (1993), Iglewicz and Hoaglin (1993), Osborne and Overbay (2004), Zuur et al. (2009, 2010), Manoj and Kannan (2013) and Kannan et al. (2015) (Fig. 1).

In order to avoid drawing wrong interpretations and conclusions, a first data exploration in this context should filter out any typing mistakes, identify possible outliers, and may also provide some ideas about how to conduct subsequent data analyses (Zuur et al. 2009).

Firstly, in order to flag extreme observations, both visual and calculation methods can be used to *label* potential outliers. Visual methods such as the classical *boxplots* (that consider the interquartile range to highlight very large or very small values), as well as *Cleveland dotplots* (each dot represents a quantitative value according to a categorical variable), allow taking a quick view on how observations deviate from the mass of dots. On the other hand, among calculation methods, the labeling can be done using:

– *The Z-score method* that uses the mean and standard deviation as:

$$Z_{score}\ i = \frac{x_i - \bar{x}}{s},$$

where $X_i \sim N(\mu,\sigma^2)$, and $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i - \bar{x}^2}$, so that Z-scores exceeding 3 in absolute value could represent potential outliers.

– *The modified Z-scores method*, using the median and the median of the absolute deviation (MAD) as:

$$MAD = median|x_i - \tilde{x}|,$$

where $\tilde{x}$ is the sample median, then $M_i = \frac{0.6745(x_i - \bar{x})}{MAD}$ where $E(MAD) = 0.675\sigma$ for large normal data, and if $|M_i| > 3.5$ the observation is labeled as outlier.

– *The median absolute deviation (MADe) method*, using the median and median absolute deviation as:

$$2MAD_e Method = Median \pm 2MAD_e,$$
$$\text{or } 3MAD_e Method = Median \pm 3MAD_e,$$
$$\text{and } MAD = median\ |x_i - median(x)|,\ i = 1, 2, \ldots, n,$$

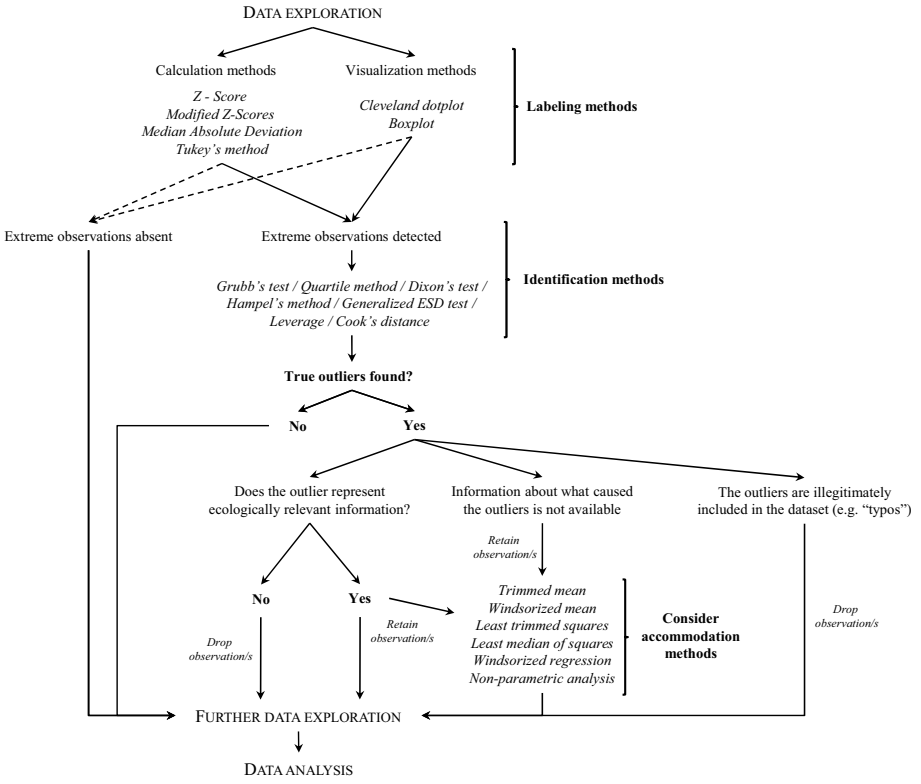

Fig. 1 A conceptual workflow to deal with outliers during data exploration

where $MAD_e = 1.483 \times MAD$ is an estimator of the spread in a dataset similar to the standard deviation for large normal data being a robust measure of central tendency.

– *The Tukey's method* uses the interquartile range to filter out very large or small numbers as follows by constructing a bulk of data that leaves out the potential outliers.

$$Low\ outliers = Q_1 - 1.5\ Q_3 - Q_1 = Q_1 - 1.5\ IQR,$$
$$\text{and } High\ outliers = Q_1 + 1.5\ Q_3 - Q_1 = Q_1 + 1.5\ IQR,$$

where Q_1 is the first quartile, Q_3 is the third quartile and IQR is the interquartile range.

Once extreme observations have been detected within a dataset and these are suspected of being outliers, those extreme values that could move the data to an inappropriate distributional model for subsequent analysis must be *identified*. This can be done by testing if the target observation behaves as an outlier in the data distribution using methods such as:

– *The Grubb's test*

It is based on the following hypothesis contrast:

H_0: The dataset does not contain outliers; H_1: At least a single outlier is present within the dataset.

The Grubb's statistic is calculated as $G = \frac{\max |Y_i - \bar{Y}|}{s}$ where Y_i is the target observation, Y is the sample mean and s is the standard deviation. According to the selected significance (α), G is compared with the critical value of the Grubb's test.

– *The quartile method*

Consist in calculating the upper quartile (Q3), the lower quartile (Q1) and the gap between them, H = Q3 – Q1, then an observation lower than Q1 – 1.5 × H and higher than Q3 + 1.5 × H can be considered a mild outlier, whereas it would be considered an extreme outlier if it is lower than Q1 – 3 × H and higher than Q3 + 3 × H.

– *The Dixon's test*

It can be applied for small sample sizes (n ≤ 40) under different criteria for samples contamined with "location" or "scalar" errors (see Dixon 1950 for detailed mathematical explanation), and consist in marking an observation X_n as outlier if the statistic $R^{(n)}$ exceeds the selected significance level (α).

– *The Hampel's method*

Consist in calculating the median (Me) of the dataset and the deviation from the median (r_i) for each observation as $r_i = (x_i - M_e)$, where x_i is an observation and Me the median, then the median of r_i is calculated as $Me_{|r_i|}$, and if $|r_i| \geq 4.5 \times Me_{|r_i|}$ then the observation can be considered an outlier.

– *The generalized ESD (extreme Studentized deviate) test*

It requires an upper bound r and is based on the hypothesis contrast: H_0: Outliers were not found in the dataset; H_1: There are up to r outliers in the dataset.

It consists in removing sequentially the observation that maximizes $|x_i - \bar{x}|$ and calculating $R_i = \frac{\max |x_i - \bar{x}|}{s}$ until r observations have been removed from the dataset (see details on the critical region in Manoj and Kannan 2013).

Also, the *leverage* (that tells how different an individual observation is compared to the other observations of the explanatory variables) and the *Cook's distance* (that in linear regression gives information on the variation of the regression parameters if the observations were sequentially omitted one by one telling how influential an observation is on the estimated parameters) can be used to assess whether an extreme observation is or not an influential observation.

Finally, regarding when to drop or not extreme observations, final decisions largely depend on the researcher and the work objective, however, removing observations must be in any case justified (Wiggins 2000). Iglewicz and Hoaglin (1993) suggested that if the cause of the outlier could not be identified, the observation should be included in the data analysis. Also, if the outlier carries important ecological information it should be retained in the dataset as well.

In these cases, different *accommodation* methods allow performing robust analyses based on the mean by censoring extreme observations at both ends of the sample such as the use of a *trimmed mean*, *Windsorized mean*, *least trimmed squares* and *least median of squares*, or applying *Windsorized regression* or *non-parametric analysis* (see Osborne and Overbay 2004).

Although more laborious, providing the results with and without outliers could help to uncover the causes and consequences of the extreme observations by comparing the interpretation of results in both situations (i.e. data analysis including and excluding outliers). This approach can help unmasking ecological processes that could be involved in the generation of the atypical observations, and it is probably worth to try it if not regarding publication at least to ensure that reliable conclusions are drawn.

Iglewicz and Hoaglin (1993) stated that one of the most common sources of outliers is recording errors (mechanical or human). Of course, data recording and management are intrinsically susceptible to error and entails the additional problem for subsequent users of having no information about how the data was collected and the extent to which it could contain errors or misinterpretations (Morgan 2004). In addition, outliers can also come from different natural situations such as changes in the system behavior or through natural deviation in populations (Hodge and Austin 2004).

Lintott and Mathews (2018) demonstrated that the misapplication of basic statistics such as the mean can have important implications for assessing environmental risk due to extreme observations, and suggested that an overinflated estimate of bat activity or the underestimation of habitat usage could lead to unnecessary or insufficient mitigation actions respectively within the Environmental Impact Assessments context. Another example in which extreme observations play an important role is the pest control in agroecosystems. Pest outbreaks are defined by the phenology of the key species and they are strongly influenced by environmental conditions (Tonnang et al. 2017). Considering the variation of the pest population abundance along time, it is clear that extreme observations correspond to pest outbreaks and encompass the most relevant ecological information in order to develop phenological models towards pest outbreak prediction. Moreover, considering research on the species' natural history, several examples can be regarded involving clear extreme observations such as dispersal and recolonization events (e.g. Öberg et al. 2008) and natural occurring blooms (e.g. Smith and Daniels 2018) involving important environmental concerns.

The systematic application of a data exploration process to deal with extreme observations such as the proposed in this work prior to data analysis could help researchers to minimize the risk of misinterpretation of results due to the presence of outliers in their datasets and thus to maximize the accuracy on their data interpretation and conclusions.

Acknowledgements Jacinto Benhadi-Marín is grateful to the Portuguese Foundation of Science and Technology for financial support through the Ph.D. grant SFRH/BD/97248/2013 and has no conflict of interest to disclose.

References

Ben-Gal I (2005) Outlier detection. In: Maimon O, Rockach L (eds) Data mining and knowledge discovery hand book: a complete guide for practitioners and researchers. Springer, Berlin

Cleveland WS (1993) Visualizing data. Hobart Press, Summit

Cook RD (1979) Influential observations in linear regression. J Am Stat Assoc 74(365):169–174

Cousineau D, Chartier S (2010) Outliers detection and treatment: a review. Int J Psychol Res 3(1):58–67

Dixon WJ (1950) Analysis of extreme values. Ann Math Statist 21(4):488–506

Hawkins D (1980) Identification of outliers. Springer, Berlin

Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22(2):85–126

Iglewicz B, Hoaglin DC (1993) How to detect and handle outliers. ASQC Quality Press, Milwaukee

Kannan KS, Manoj K, Arumugam S (2015) Labeling methods for identifying outliers. IJSS 10(2):231–238

Lintott PR, Mathews F (2018) Basic mathematical errors may make ecological assessments unreliable. Biodivers Conserv 27:265–267

Manoj K, Kannan KS (2013) Comparison of methods for detecting outliers. Int J Sci Eng Res 4(9):709–714

Morgan JH (2004) Remarks on the taking and recording of biometric measurements in bird ringing. Ring 26:71–78

Öberg S, Mayr S, Dauber J (2008) Landscape effects on recolonisation patterns of spiders in arable fields. Agric Ecosyst Environ 123:211–218

Osborne JW, Overbay A (2004) The power of outliers (and why researchers should ALWAYS check for them). Pract Assess Res Eval 9(6):1–8

Rousseeuw PJ, Hubert M (2011) Robust statistics for outlier detection. WIREs Data Mining Knowl Discov 1:73–79

Smith GJ, Daniels V (2018) Algal blooms of the 18th and 19th centuries. Toxicon 142:42–44

Tabachnick BG, Fidell LS (1996) Using multivariate statistics, 3rd edn. Harper Collins College Publishers, New York

Tonnang HEZ, Hervé BDB, Biber-Freudenberger L, Salifu D, Subramanian S, Ngowi VB, Guimapi RYA, Anani B, Kakmeni FMM, Affognon H, Niassy S, Landmann T, Ndjomatchoua FT, Pedro SA, Johansson T, Tanga CM, Nana P, Fiaboe KM, Mohamed SF, Maniania NK, Nedorezov LV, Ekesi S, Borgemeister C (2017) Advances in crop insect modelling methods—towards a whole system approach. Ecol Modell 354:88–103

Wiggins BC (2000) Detecting and dealing with outliers in univariate and multivariate contexts. Annual Meeting of the Mid-South Educational Research Association, Bowling Green

Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer, New York

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14