5th International Conference "New Challenges for Statistical Software – The use of R in Official Statistics (uRos 2017)"
Bucharest, 6-7 November 2017

Workshop:

Outlier Detection in R: Some Remarks

Marcello D'Orazio*
marcello.dorazio(at)fao.org

**Senior Researcher in Statistical Methodology*
Office of Chief Statistician, Food and Agriculture Organization of the UN
(Italian National Institute of Statistics – Istat)

Outline

1. Univariate outliers
2. Detection of univariate outliers
    a. specific focus on official statistics
    b. in R
3. Examples of detection of bivariate outliers

Outlier: ‘lies outside’

Outlier (continuous variables):

“An observation which is not fitted well by a model"

“An observation which is not close to the center of the data”

(Istat CBS SFSO Eurostat, 2007)

Univariate outlier, when dealing with only one variable

Multivariate outlier, referred to a set of variables

- ✓ Due to measurement error (e.g. unit measure error): the observed value is NOT the true value and the true value is NOT an outlier (e.g. observed 1,000,000 instead of 1,000)

- ✓ Extreme value NOT affected by error (the observed value is an outlying true value). May deserve 'special' treatment in analysis.

  In sample surveys:

  - Representative outliers
    i.e. a value observed on one sample unit, but in the population there are other non-sampled units with similar values
  - Non-representative outliers
    i.e. a value observed on one sample unit, but in the population there aren't non-sampled units with similar values ('unique' value)

Source of outliers determines the corresponding treatment:

Outlier due to measurement errors:
delete the value and substitute (impute) it

Not an error, representative:
sometimes value is deleted and substituted with a new value (Winsorization) to reduce its influence on final survey estimates

Not an error, non-representative:

- the survey weight associated to the unit is reduced (sometimes set close or equal to 1)
- the value is deleted and substituted with a fixed value (Winsorization)

Winsorization: all the values larger than a threshold k are substituted with k

- Location & Scale-based intervals (mainly referred to Gaussian Distribution)

- Boxplot Methods

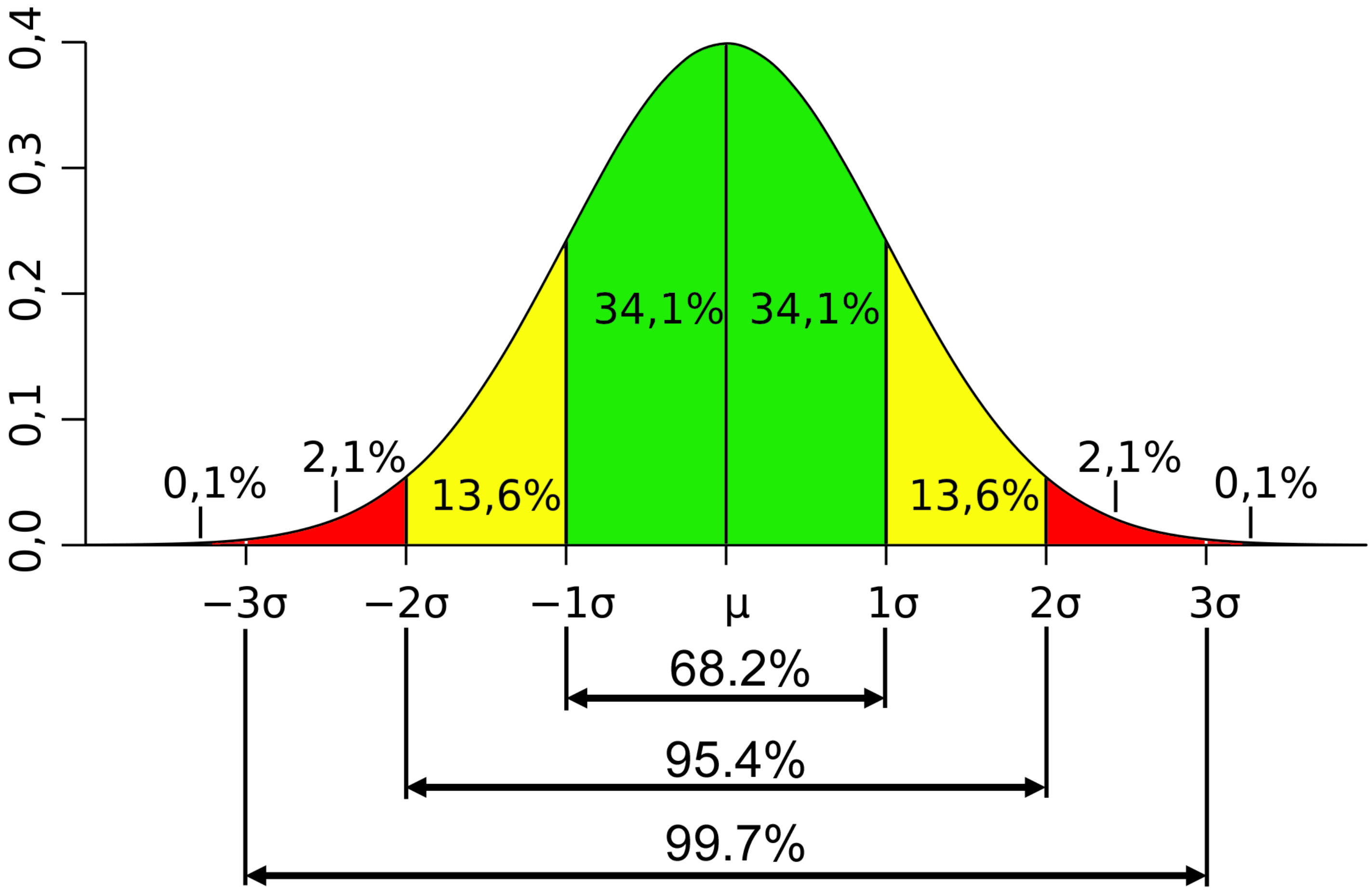
- Test-statistics, after (robust) estimation of the distribution of the bulk of data

- Methods based on fitting mixture models

- …

Detection methods usually do not depend on the source of outliers.

Gaussian distribution: $\mu \pm k \cdot \sigma$, $k = 1,2,3$



Source: http://www.muelaner.com/wp-content/uploads/2013/07/Standard_deviation_diagram.png

Outlier: observations lying outside interval $[\tilde{\mu} - k \cdot \tilde{\sigma},\ \tilde{\mu} + k \cdot \tilde{\sigma}]$

$\tilde{\mu}$ and $\tilde{\sigma}$ robust estimates of μ and σ, respectively. $k \in \{2, 2.5, 3\}$

$\tilde{\mu} = median = Q_{0.50}$ (max breakpoint of 50%)

<u>Max breakpoint</u>: fraction of obs. that can be arbitrarily changed while maintaining the estimate bounded

Various alternatives to achieve robust estimation of σ:

a) $\tilde{\sigma} = IQR/a = (Q_{0.75} - Q_{0.25})/a$ (max breakp. of 25%)

b) $\tilde{\sigma} = MAD = b \times med\left|x_i - med(x_i)\right|$ (max breakp. of 50%)

c) $\tilde{\sigma} = S_n = c \times med_i \left\{ med_j \left|x_i - x_j\right| \right\}$ (max breakp. of 50%)

d) $\tilde{\sigma} = Q_n = d \times \left\{ \left|x_i - x_j\right|; i < j \right\}_{(k)}$ (max breakp. of 50%)

e) Bi-weight estimate of σ (and μ) (max breakp. of 50%)

Gaussian distr.: $a = 1.349,\ b = 1.4826,\ c = 1.1926,\ d = 2.21914$

IQR and MAD in R package **stats** (R Core Team, 2017)

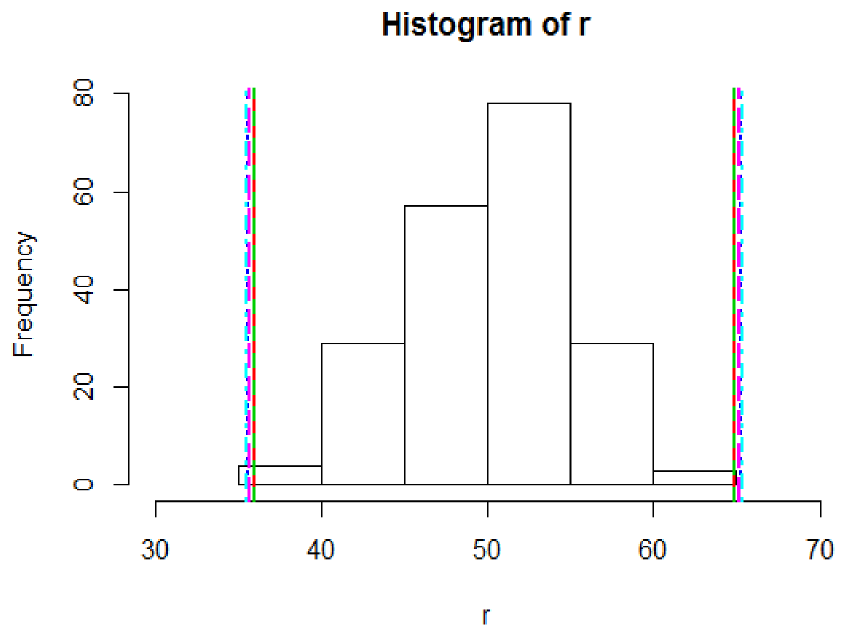
S_n, Q_n, bi-weight estimate of σ, in package **robustbase** (Maechler et al. 2016)

Wrapper in package **univOutl** (D'Orazio, 2017), by means of the function **`LocScaleB()`**

- includes all the estimators of σ
- accepts survey weights (not always, but breakp. 0%)

```
> library(sn)
> library(univOutl)

> # generate data from normal distr.
> set.seed(123456)
> r <- rsn(n=200, xi=50,
           omega=5, alpha=0)
> hist(r)
> mc(r) # medCouple skewness measure
[1] 0.006576713
```




```
# estimated parameters
            median    scale
IQR       50.41952 4.820553
MAD       50.41952 4.818536
Sn        50.41952 4.935173
Qn        50.41952 4.971679
scaleTau2 50.41952 4.916652
```

```
# Estimated bounds
          lower.low upper.up
IQR        35.95786 64.88118
MAD        35.96391 64.87512
Sn         35.61400 65.22504
Qn         35.50448 65.33455
scaleTau2  35.66956 65.16947
```

Rousseeuw and Croix (1993): S_n and Q_n can work with asymmetric distributions (changing c and d constants)

With asymmetric distribution:

i. use S_n and Q_n but change constants (c, d), or

ii. transform the data (Log, Box-Cox, etc.), or

iii. use methods accounting for (slight) skewness:

a) asymmetric intervals $\left[\tilde{\mu} - k \cdot \tilde{\sigma}_L ,\ \tilde{\mu} + k \cdot \tilde{\sigma}_R\right]$
simple example:

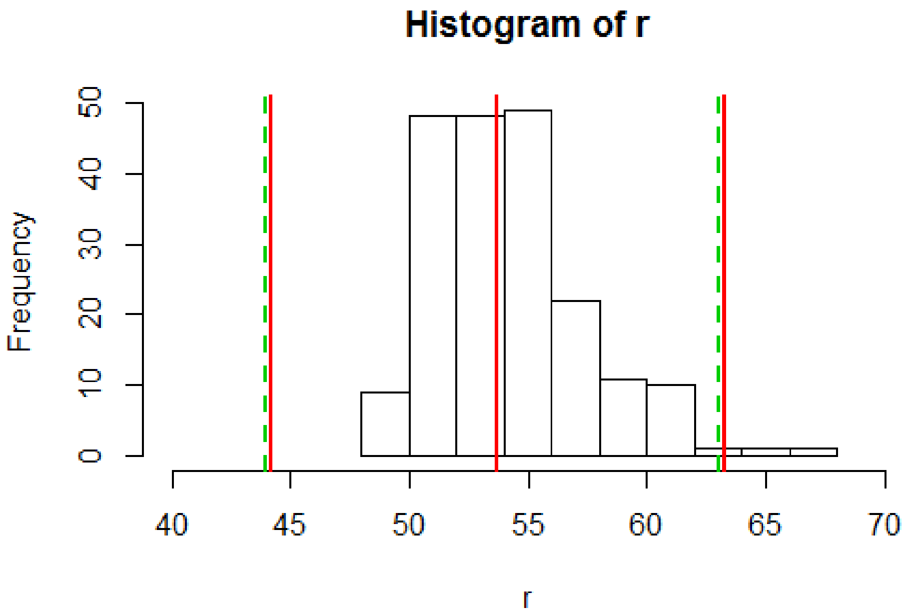
$$\tilde{\sigma}_L = \frac{Q_2 - Q_1}{0.6745} \qquad \tilde{\sigma}_R = \frac{Q_3 - Q_2}{0.6745}$$

(but max breakp. 25%)

b) more in general, “reflection” across median (Lanzante, 1996)

```
> # gen. data positive skewed normal
> set.seed(432123)
> r <- rsn(n=200, xi=50,
           omega=5, alpha=4)
> hist(r)
> mc(r) # medCouple skewness measure
[1] 0.05529924

> a1 <- LocScaleB(x=r,
                  method = "IQR")
> a2 <- LocScaleB(x=r,
                  method = "dq")
```




```
> a1$pars
   median     scale
53.709897  3.181763
> a2$pars
   median   sc.left  sc.right
53.709897  3.254857  3.108669
```

```
> a1$bounds
lower.low  upper.up
 44.16461  63.25519
> a2$bounds
lower.low  upper.up
 43.94533  63.03590
```

Outlier: observations lying outside interval $\left[f_{low},\ f_{up}\right]$

f said **fence**

Traditional:

$$f_{low} = Q_1 - k \times IQR \qquad f_{up} = Q_3 + k \times IQR \qquad k \in \{1.5, 2, 3\}$$

Asymmetric fences (slight skewness):

$$f_{low} = Q_1 - 2k \times (Q_2 - Q_1) \qquad f_{up} = Q_3 + 2k \times (Q_3 - Q_2) \qquad k \in \{1.5, 2, 3\}$$

Skewness-adjusted (moderate skewness, $-0.6 \le M \le 0.6$):

$$f_{low} = Q_1 - 1.5 \times e^{aM} \times IQR \qquad f_{up} = Q_3 + 1.5 \times e^{bM} \times IQR$$

M is the **medcouple** measure of skewness, when $M > 0$ then $a = -4$ and $b = 3$ ($a = -3$ and $b = 4$ with $M < 0$) (Vanderviere and Huber, 2008)

Boxplot -> various functions (e.g. **`boxplot.stats()`** in **grDevices**; R Core Team, 2017)

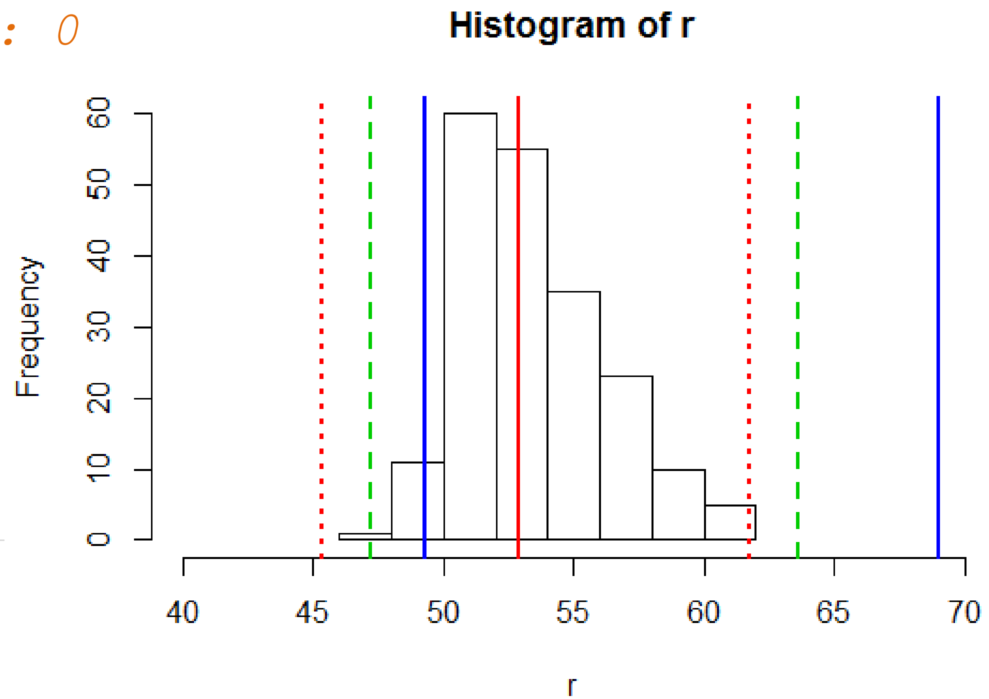
Skewness-adjusted -> function **`adjboxStats()`** in **robustbase** (Maechler et al. 2016)

Wrapper in package **univOutl** (D'Orazio, 2017), by means of the function **`boxB()`**:

- implements also asymmetric fences;
- accepts survey weights.

```
> set.seed(11122)
> r <- rsn(n=200, xi=50, omega=5, alpha=5)
> hist(r)
> mc(r) # medCouple skewness measure
[1] 0.2597695
> a1 <- boxB(x=r, k=1.5, method='resistant')
No outliers found
> a2 <- boxB(x=r, k=1.5, method='asymmetric')
No outliers found
> a3 <- boxB(x=r, k=1.5, method='adjbox')
The MedCouple skewness measure is: 0.2598
No. of outliers in left tail: 4
No. of outliers in right tail: 0
```

```
# fences
        lower    upper
std  45.32037 61.73863
asym 47.19009 63.60835
adjb 49.28308 69.05740
```



In panel surveys, same units observed in different time occasions:

$y_{1,t-1}$ $y_{1,t}$

$y_{2,t-1}$ $y_{2,t}$ → detection of outliers on **ratios** $r_i = y_{i,t} / y_{i,t-1}$

…

Hidiroglou-Berthelot (1986) method (loc-scale intervals on scores derived from the ratios):

1) $s_i = \begin{cases} 1 - r_{med}/r_i\,, & \text{if } \ 0 < r_i < r_{med} \\ r_i/r_{med} - 1, & \text{if } \ r_i \geq r_{med} \end{cases}$ r_{med} is the median of ratios

2) $E_i = s_i \times \left[max\left(y_{i,t}, y_{i,t-1} \right) \right]^U$ $0 \leq U \leq 1$ (usually $U = 0.5$)

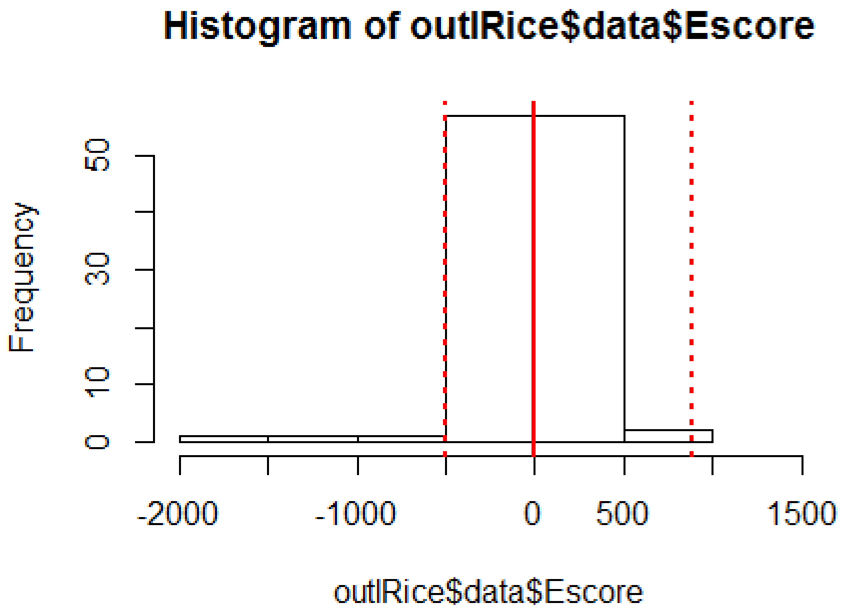
3) Outlying ratios those outside interval $\left[E_{med} - C \times d_1, E_{med} + C \times d_3 \right]$

$d_1 = max\left\{ \left(E_{med} - E_{Q_1} \right), \left| A \times E_{med} \right| \right\}$ $d_3 = max\left\{ \left(E_{Q3} - E_{med} \right), \left| A \times E_{med} \right| \right\}$

Usually $A = 0.05;$ $C \geq 4$

Implemented in function **`HBmethod()`** in package **univOutl** (D'Orazio, 2017)

```
> outlRice <- HBmethod(yt1 = rice$Prod2014,
+                      yt2 = rice$Prod2015,
+                      return.dataframe = TRUE,
+                      C=15)
MedCouple skewness measure of E scores: 0.1253
Outliers found in the left tail: 3
Outliers found in the right tail: 0
>
> outlRice$quartiles.E
      25%        50%        75%
-33.21419   0.00000   58.87622
> outlRice$bounds.E
      low         up
-498.2128  883.1434
```



```
> head(outlRice$excluded, 3)
  id yt1      yt2
1 92   0 1.345158
> head(outlRice$data, 3)
  id    yt1         yt2     ratio      Escore outliers
1  1 537000 410000.0000 0.7635009 -229.584643        0
2  2    320    273.8746 0.8558582   -3.069251        0
3  3  42288  45322.0000 1.0717461   14.658901        0
```

Hidiroglou-Berthelot method can deal with slightly skewed ratio distributions.

Recent paper by Young and Mathew (2015) based on trimming.

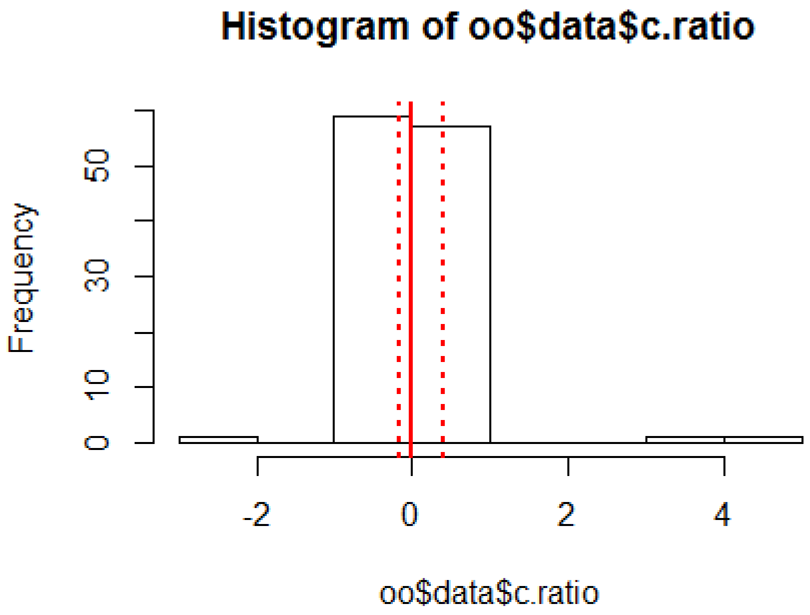
Alternative approach: use skewness-adjusted boxplot.

In **`ratioSize()`** function in **univOutl** (D'Orazio, 2017):

1) Derive skewness-adjusted fences for scores s_i
2) Inspect only 'important' outliers, i.e. those with a $z_i > h$ $(h > 0)$; size measure (z_i) can be an arbitrarily chosen variable, for instance, as in HB $z_i = \left[max\left(y_{i,t}, y_{i,t-1}\right)\right]^U$

Detection of Outliers with Ratios with Skewed Distribution

```
> outl.HB <- outlRice$data$id[outlRice$data$outliers==1]
> oo <- ratioSize(numerator = rice$Prod2015,
+                 denominator = rice$Prod2014,
+                 return.dataframe = T)
MedCouple skewness measure of centerad ratios: 0.1506
> oo$median.r
[1] 1.002703
> oo$bounds
[1] -0.1733648  0.4133880
```




```
> head(oo$data, 3)
     id numerator denominator     ratio      c.ratio outliers      size
119 120 208230000   206507392 1.0083416  0.005623246        0 208230000
49   49 156540000   157200000 0.9958015 -0.006930754        0 157200000
50   50  75397841    70846465 1.0642428  0.061373723        0  75397841
```

Detection of Outliers with Ratios with Skewed Distribution

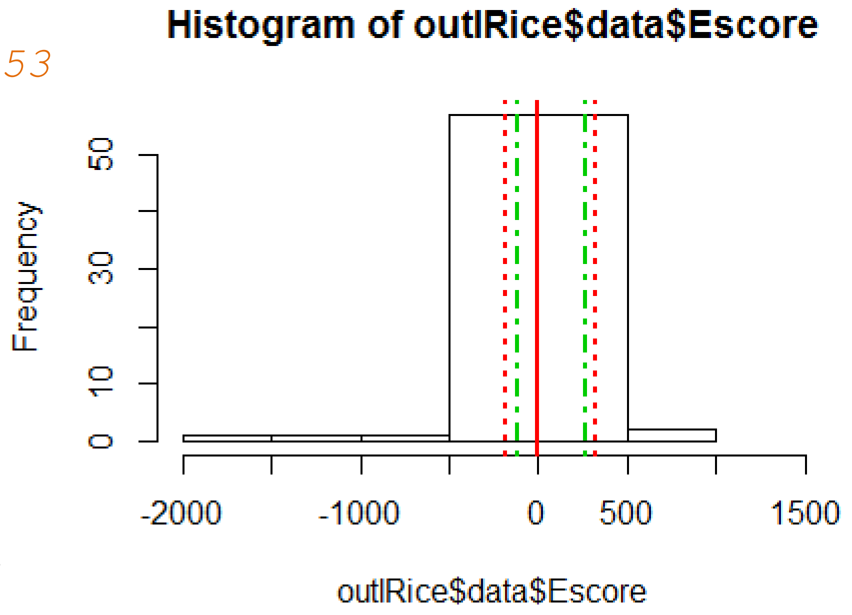
```
> oo <- ratioSize(numerator = rice$Prod2015,
+                 denominator = rice$Prod2014,
+                 return.dataframe = T, size.th = 1000)
MedCouple skewness measure of centered ratios: 0.1506
> head(oo$data)
     id numerator denominator     ratio   c.ratio outliers     size
103 104  27702191    32620160 0.8492353 -0.1807130        1 32620160
93   94    871693     1204020 0.7239855 -0.3849769        1  1204020
117 118    802015     1158056 0.6925529 -0.4478363        1  1158056
92   93    906348      559021 1.6213130  0.6169421        1   906348
6     6    690390      819276 0.8426831 -0.1898936        1   819276
29   29    418037      584800 0.7148376 -0.4027008        1   584800
```

Another approach:

Skewness adjusted boxplot on Hidiroglou-Berthelot *E*-scores

```
> outlRice <- HBmethod(yt1 = rice$Prod2014,
+                      yt2 = rice$Prod2015,
+                      return.dataframe = TRUE,
+                      C=5.4)
MedCouple skewness measure of E scores: 0.1253
Outliers found in the left tail: 12
Outliers found in the right tail: 3
>
> oo <- boxB(x=outlRice$data$Escore,
+           method = 'adjbox')
The MedCouple skewness measure is: 0.1253
No. of outliers in left tail: 18
No. of outliers in right tail: 4

> outlRice$bounds.E
      low        up
-179.3566  317.9316
> oo$fences
    lower     upper
-116.8821  260.0705
```



Detecting outliers with test after (robust) estimating the distribution of the bulk of data:

Exponential, Weibull, LogNormal, Pareto → van der Loo (2010) R package **extremevalues**

Methods based on fitting mixture models:

Mixture of Gaussian distr. → Guarnera and Buglielli (2016) package **SeleMix**

Identification of outliers in time series:

Method by Chen and Liu (1993) implemented in package **tsoutliers** (López-de-Lacalle, 2017)

Smoothing method in function `tsoutliers()` in package **forecast** (Hyndman, 2017)

Multivariate outlier: an observation with 'characteristics' different from the multivariate distribution of the majority of observations

Detection of multivariate outliers: distance from the distribution of the bulk of data; typically:

i. Multivariate Gaussian distribution is considered
ii. Robust estimation of mean vector and Var-Cov matrix (MVE, MCD, OGK, SD-estimator, ...)
iii. Mahalanobis distance of each obs. $d_{M,i} = (x_i - \tilde{x})^T \tilde{S}^{-1} (x_i - \tilde{x})$
iv. Obs. with $d^2_{M,i} > \chi^2_{p,1-\alpha}$ detected as outliers (p is the no. of vars., $\alpha = 0.025$ or 0.01 etc.).
Sometimes threshold is $\chi^2_{p,1-\alpha}$ modified according to the robust estimation method of mean and Var-Cov matrix.

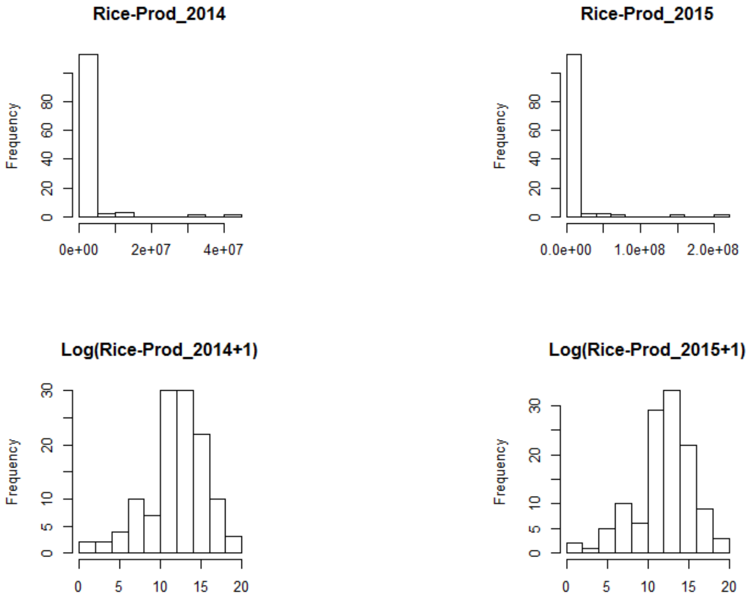
Various R packages available:

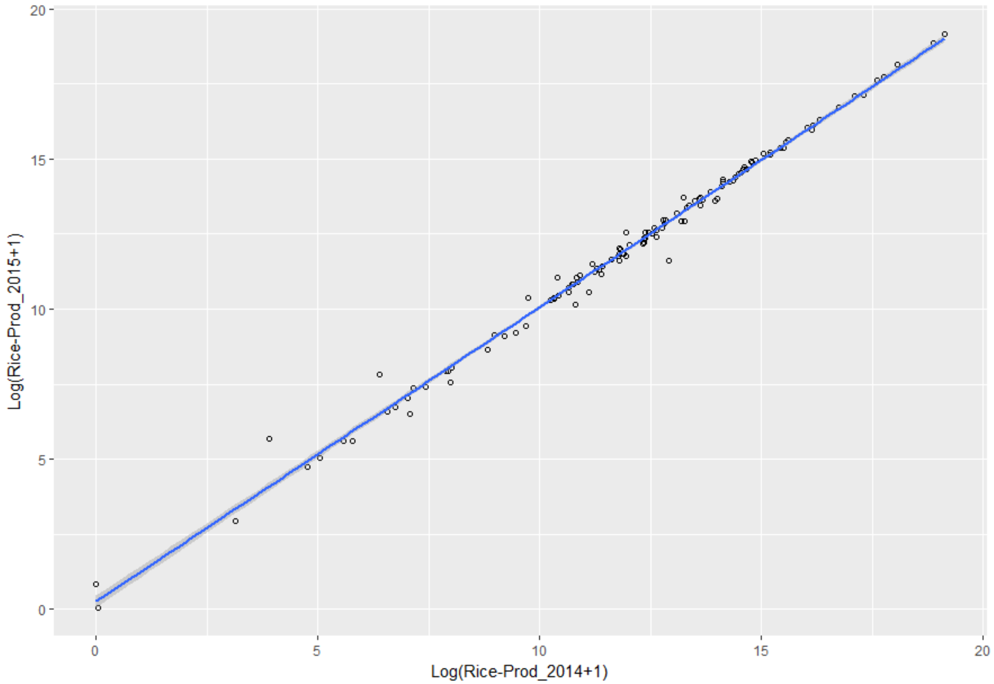
mvoutlier (Filzmoser and Gschwandtner, 2017)

rrcov (Todorov and Filzmoser, 2009), **rrcovNA** (Todorov, 2016)

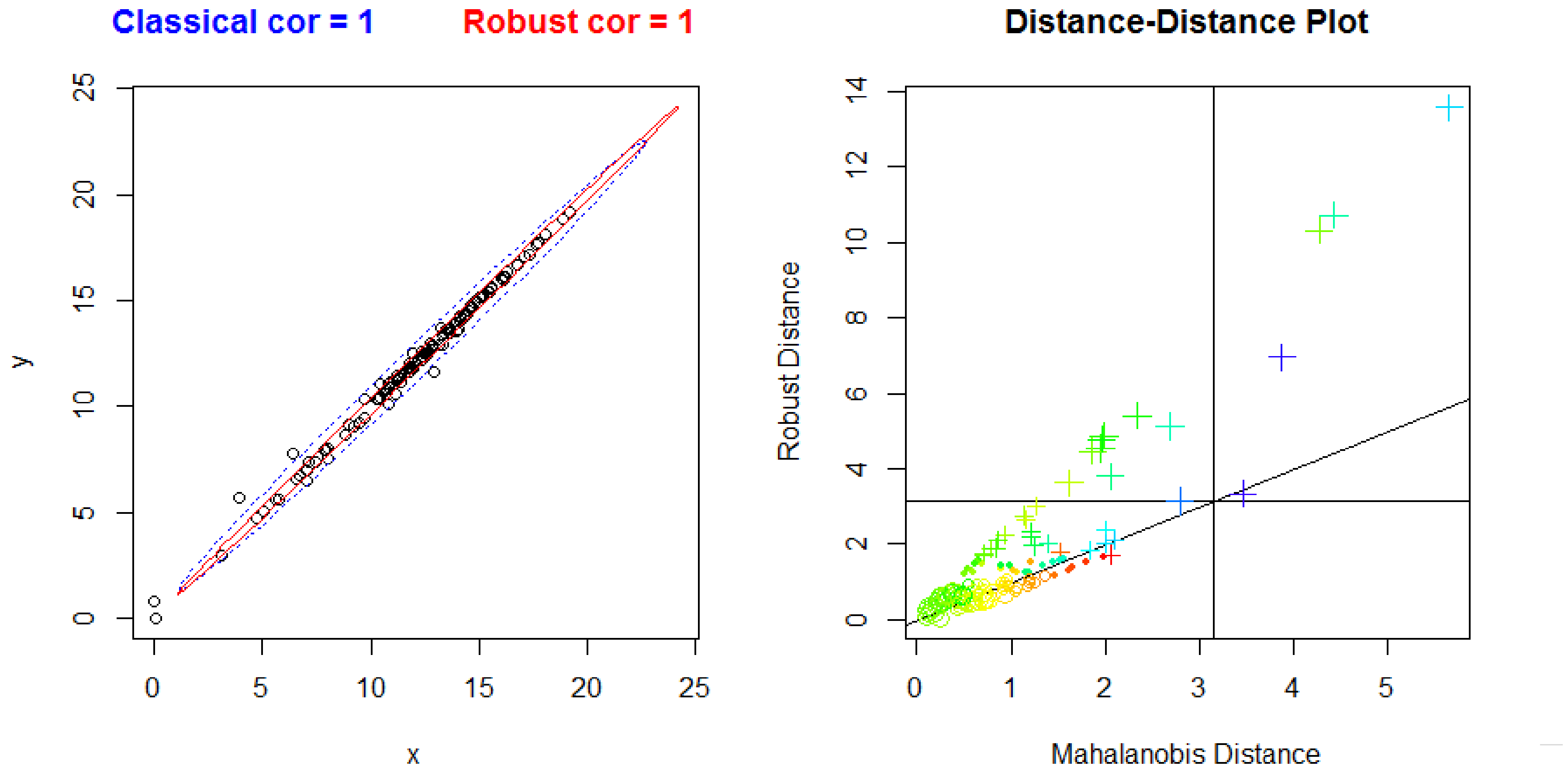
...

Let's consider Rice production data

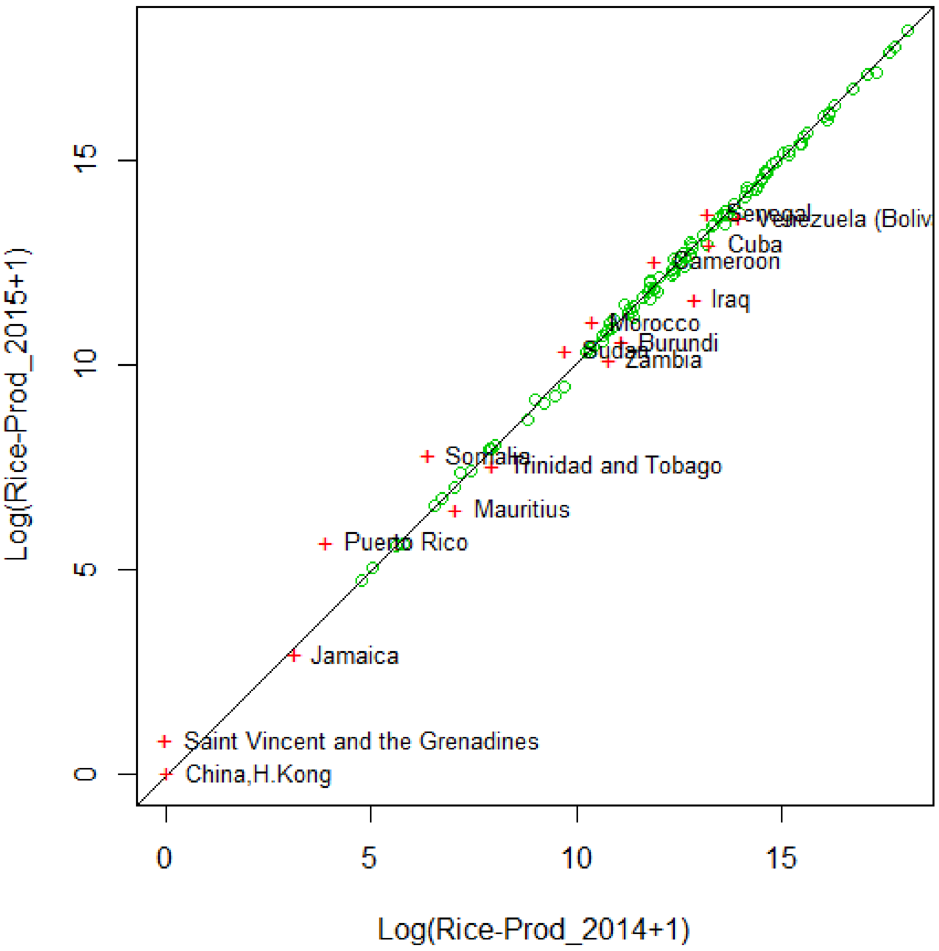




```
> library("mvoutlier")
> par(mfrow=c(1,2))
> corr.plot(rice$logProd2014, rice$logProd2015, alpha=0.01)
$cor.cla
[1] 0.9961189
$cor.rob
[1] 0.9994238
> dd <- dd.plot(rice[,c("logProd2014", "logProd2015")], alpha=0.01)
```



```
> sum(dd$outliers)
[1] 16
> outl <- dd$outliers
> head(rice[outl,], 3)
   geographicAreaM49 Geographic.Area Prod2014 Prod2015 Area2014 Area2015 logProd2015 logProd2014
17               108         Burundi    67377    38674    23730    34246    10.56295    11.11807
19               120        Cameroon   153246   278281   126901   226779    12.53639    11.93981
30               192            Cuba   584800   418037   171572   112166    12.94333    13.27903
```



The distribution for the observed data is a mixture of two Gaussian distributions. In package **SeleMix**, the two distributions:

a) share the same mean vector

b) but have different Var-Cov matrix, contaminated data (outliers due to measurement errors) have a larger Var-Cov matrix, proportional to the one of non-contaminated data.

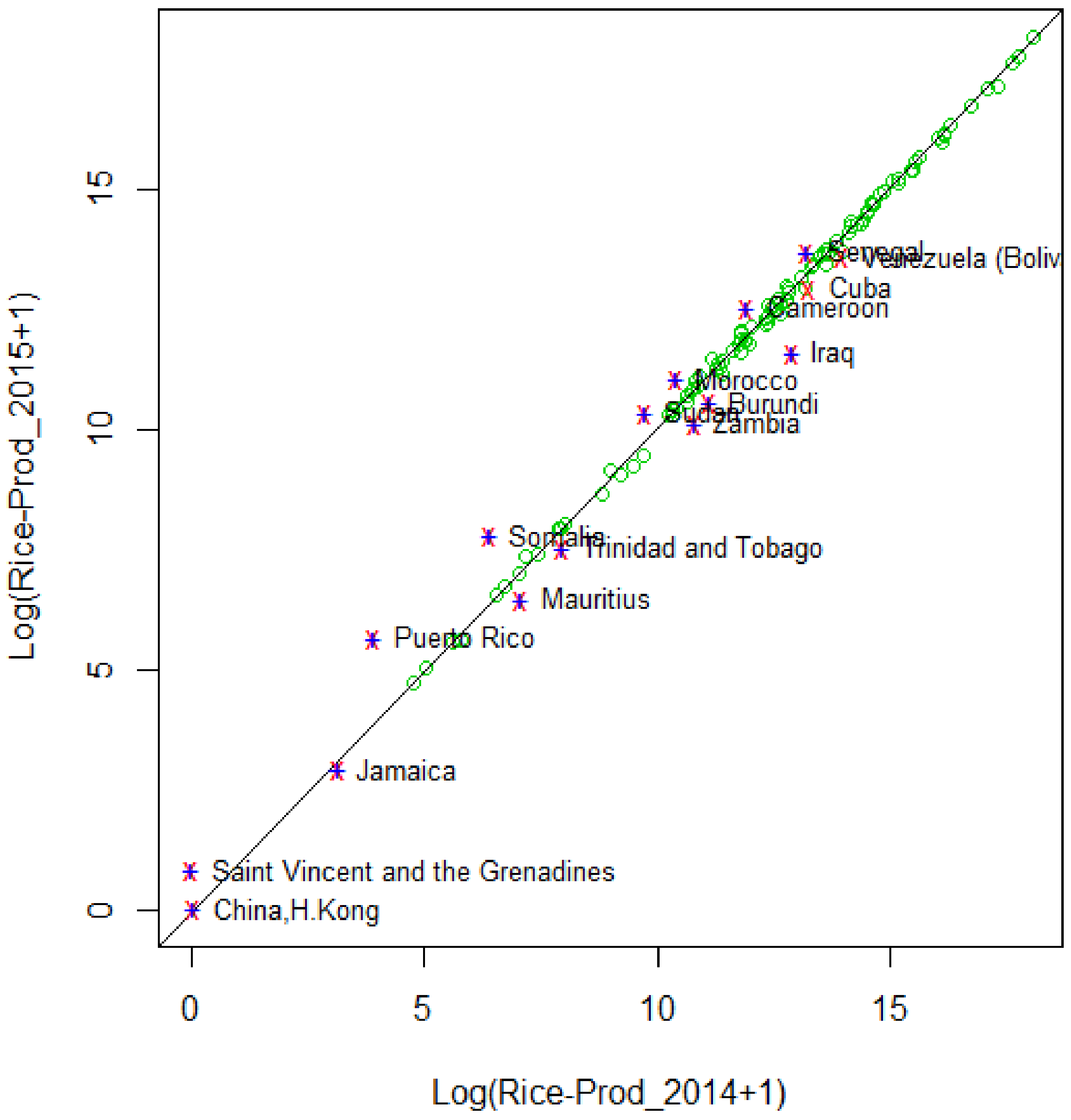
I.e. contaminated data have additive measurement errors with zero mean but variance proportional to the one of non-contaminated.

Detection based on **posterior probability of being erroneous:**

prob > t.outl (=0.5 usually)

```
> library("SeleMix")
> out.sel <- ml.est(y = rice[,c("logProd2014", "logProd2015")],
+                   model="N", w=0.005, w.fix=F, t.outl=0.5)
> out.sel$w # estimated proportion of contaminated data
[1] 0.1482978
> out.sel$lambda # estimated variance inflation factor
[1] 17.41743
> sum(out.sel$outlier) # estimated number of contaminated obs
[1] 14

> toCheck <- data.frame(Geographic.Area=rice$Geographic.Area,
+                       postProb=out.sel$tau,
+                       rice[,c("logProd2014", "logProd2015")],
+                       out.sel$ypred)
>
> toCheck <- subset(toCheck, postProb>0.5)
> toCheck <- toCheck[order(toCheck$postProb, decreasing = T), ]
> head(toCheck)
                     Geographic.Area  postProb logProd2014 logProd2015 logProd2014.p logProd2015.p
53                              Iraq 1.0000000   12.906764  11.6010279      12.54574      12.47941
90                       Puerto Rico 1.0000000    3.909018   5.6827291      12.05719      12.15807
100                          Somalia 1.0000000    6.398595   7.8042514      12.19237      12.27326
95  Saint Vincent and the Grenadines 1.0000000    0.000000   0.8523528      11.84495      11.89580
127                           Zambia 0.9995330   10.812572  10.1470218      12.43127      12.39942
70                         Mauritius 0.9992235    7.079184   6.4892049      12.22532      12.19743
```


Senegal
Venezuela (Boliv
Cuba
Cameroon
Iraq
Morocco
Burundi
Sudan
Zambia
Somalia
Trinidad and Tobago
Mauritius
Puerto Rico
Jamaica
Saint Vincent and the Grenadines
China,H.Kong
0
5
10
15
Log(Rice-Prod_2015+1)
Log(Rice-Prod_2014+1)

Thank you for your attention

Questions?

References

Chen, C. and Liu, Lon-Mu (1993). 'Joint Estimation of Model Parameters and Outlier Effects in Time Series'. *JASA*, 88, pp. 284-297.

D'Orazio M. (2017). univOutl: Detection of Univariate Outliers. R package version 0.1-3. https://CRAN.R-project.org/package=univOutl

Filzmoser P. and Gschwandtner M. (2017). mvoutlier: Multivariate Outlier Detection Based on Robust Methods. R package version 2.0.8. https://CRAN.R-project.org/package=mvoutlier

Guarnera U. and Buglielli T. (2016). SeleMix: Selective Editing via Mixture Models. R package version 1.0.1. https://CRAN.R-project.org/package=SeleMix

Hyndman RJ (2017). forecast: Forecasting functions for time series and linear models. R package version 8.2, http://pkg.robjhyndman.com/forecast

Hidiroglou, M.A. and Berthelot, J.-M. (1986) 'Statistical editing and Imputation for Periodic Business Surveys'. *Survey Methodology*, 12, pp. 73-83.

Istat, CBS, SFSO and Eurostat (2007) *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Manual prepared by the EDIMBUS Project. http://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf

Lanzante J.R (1996) "Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples including applications to historical radiosonde station data". *International Journal of Climatology*, 16, pp. 1197-1126.

López-de-Lacalle J. (2017). tsoutliers: Detection of Outliers in Time Series. R package version 0.6-6. https://CRAN.R-project.org/package=tsoutliers

Maechler M., Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., Salibian-Barrera M., Verbeke T., Koller M., Conceicao E. L. T. and di Palma M. A.(2016). robustbase: Basic Robust Statistics R package version 0.92-7. http://CRAN.R-project.org/package=robustbase

References

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rousseeuw P.J. and Croux C. (1993) "Alternatives to Median Absolute Deviation". *JASA*, 88, pp. 1273-1283.

Todorov V. (2016). rrcovNA: Scalable Robust Estimators with High Breakdown Point for Incomplete Data. R package version 0.4-9. https://CRAN.R-project.org/package=rrcovNA

Todorov V. and Filzmoser P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. Journal of Statistical Software, 32(3), 1-47. http://www.jstatsoft.org/v32/i03/

van der Loo M.P.J. (2010), extremevalues, an R package for outlier detection in univariate data, R package version 2.3 http://www.github.com/markvanderloo/extremevalues

van der Loo M.P.J (2010) "Distribution Based Outlier detection in Univariate Data". *Statistics Netherlands Discussion Paper*, 10003.

Vanderviere, E. and Huber, M. (2008) 'An Adjusted Boxplot for Skewed Distributions', *Computational Statistics & Data Analysis*, 52, pp. 5186-5201

de Waal T., Pannekoek J., Scholtus S. (2011) *Handbook of statistical data editing and imputation*. Wiley & Sons, Inc., Hoboken, New Jersey.

Young D.S. and Mathew T. (2015) "Ratio edits based on Statistical Tolerance Intervals", *JOS*, 31, pp. 77-100