Istat | Istituto Nazionale di Statistica

Statistics Canada, 6 December 2022

Use of R in Official Statistics – uRos 2022
10th International Virtual Conference

Recent and traditional approaches to outlier detection in panel survey data

Marcello D'Orazio
Italian National Institute of Statistics - Istat | Department for Statistical Production

Outliers in Official Statistics

❑ Outlier

- ➢ UNECE Glossary (2000): A data value that lies in the tail of the statistical distribution of a set of data values

  Underlying assumption: "outliers in the distribution of uncorrected (raw) data are more likely to be incorrect"

- ➢ de Waal *et al* (2011): A value, or a record, that is not fitted well by a model that is posited for the observed data
  - a single value ⇨ **univariate** outlier
  - an entire record (or a subset of values), when the values are considered simultaneously ⇨ **multivariate** outlier

○ Univariate outliers in probabilistic sample surveys:

- Erroneous value due to a measurement error
- Non-erroneous value but extreme value (representative or non-representative)

Istat

R packages for Outlier Detection in Official Statistics (1/2)

❑ Detection of univariate Outliers:

- **univOutl** (D'Orazio, 2022)
  - Nonparametric: Boxplot based (also with moderate skewness)
  - Data following Gaussian distribution: various options (robust estimation of location and scale)
  - Hidiroglou-Berthelot (1986) approach for ratios ($r_i = y_{t,i}/y_{t-1,i}$) related to longitudinal data (panel surveys)

- **extremevalues** (van del Loo, 2020) uses statistical test after (robust) estimate of the distribution (Exponential, Weibull, LogNormal, Pareto) of the bulk of data

- **SeleMix** (Guarnera and Buglielli, 2020) fits mixture of Gaussian models (allows error free predictors)

R packages for Outlier Detection in Official Statistics (2/2)

❑ Detection of multivariate Outliers:

- **mvoutlier** (Filzmoser and Gschwandtner, 2021) distance from the distribution (Gaussian) of the bulk of data (also features for compositional data)

- **rrcov** (Todorov, 2022), **rrcovNA** (Todorov, 2020), **rrcovHD** (Todorov, 2021)
  - distance from the distribution (Gaussian) of the bulk of data (Robust estimation of mean vector and Var-Cov matrix with MVE, MCD, OGK, SD-estimator, ...)
  - robust sparse PCA, robust PLS, robust sparse classification

- **SeleMix** (Guarnera and Buglielli, 2020) fits mixture of Gaussian models (allows error free predictors)

Istat

«Recent» approaches and R packages for Outlier Detection (1/2)

- **Density (distance)-based** nonparametric approaches: many variants of ***k*-NN** (Ramaswamy et al. 2000; Angiulli and Pizzuti, 2002; …)
  - **pros**: fully nonparametric; applicable to univariate/multivariate case; assign a unique score for each obs that can be used for ranking potential outliers
  - **cons**: choice of distance function and *k*; approx. methods to deal with too many obs. (too large distance matrices)

  DDoutlier (Madsen, 2018) many density-based methods

- **Clustering-based** approaches; e.g. ***DBSCAN clustering***: outliers are the "noisy" observations not "reachable" by any other observation
  - **pros**: fully nonparametric; applicable to univariate/multivariate case; direct identification of potential outliers
  - **cons**: choice of distance function and the distance threshold that determines the "reachability"; choice of *k* that identifies the set of core observations; no score for ranking observations

  dbscan (Hahsler et al., 2019; 2022) the DBSCAN clustering algorithm and facilities to efficiently calculate the *k*-NN distance

Istat

«Recent» approaches and R packages for Outlier Detection (2/2)

- **Decision-tree** algorithms: the **isolation tree**
  - the more observations show similar *X* values, the longer (more splits) it will take to separate them in small groups (or alone) compared to less occurring *X* values: the **isolation depth** (number of splits needed to isolate a unit) is used for detecting outliers
  - **Isolation forest**: fits an ensemble of isolation trees (Liu et al., 2008 and 2012)
    - **pros**: fully nonparametric; applicable to univariate/multivariate case; assigns a unique score (from 0 to 1) for each obs that can be used for ranking potential outliers (**score>0.5 rule of thumb**); tuning parameters are not crucial as in other nonparametric approaches
    - **cons**: use in the multivariate setting requires variants of the "base" method

solitude (Srikanth, 2021): the "base" isolation forest algorithm

isotree (Cortes, 2022): the "base" algorithm and some of its variants for the multivariate setting

An application to panel survey data (1/2)

Input data are the (centered) scores (E_i) calculated in the Hidiroglou-Berthelot approach starting from the ratios $r_i = y_{t_2 i}/y_{t_1 i}$ $(i = 1,2, \dots, m)$

$$E_i = s_i \times \left[\max\left(y_{t_1 i}, y_{t_2 i}\right)\right]^U$$

$$s_i = \begin{cases} s_i = 1 - \dfrac{r_M}{r_i}, & 0 < r_i < r_M \\ s_i = \dfrac{r_i}{r_M} - 1, & r_i \geq r_M \end{cases}$$

r_M is the median of the ratios (excluding 0 and Inf)

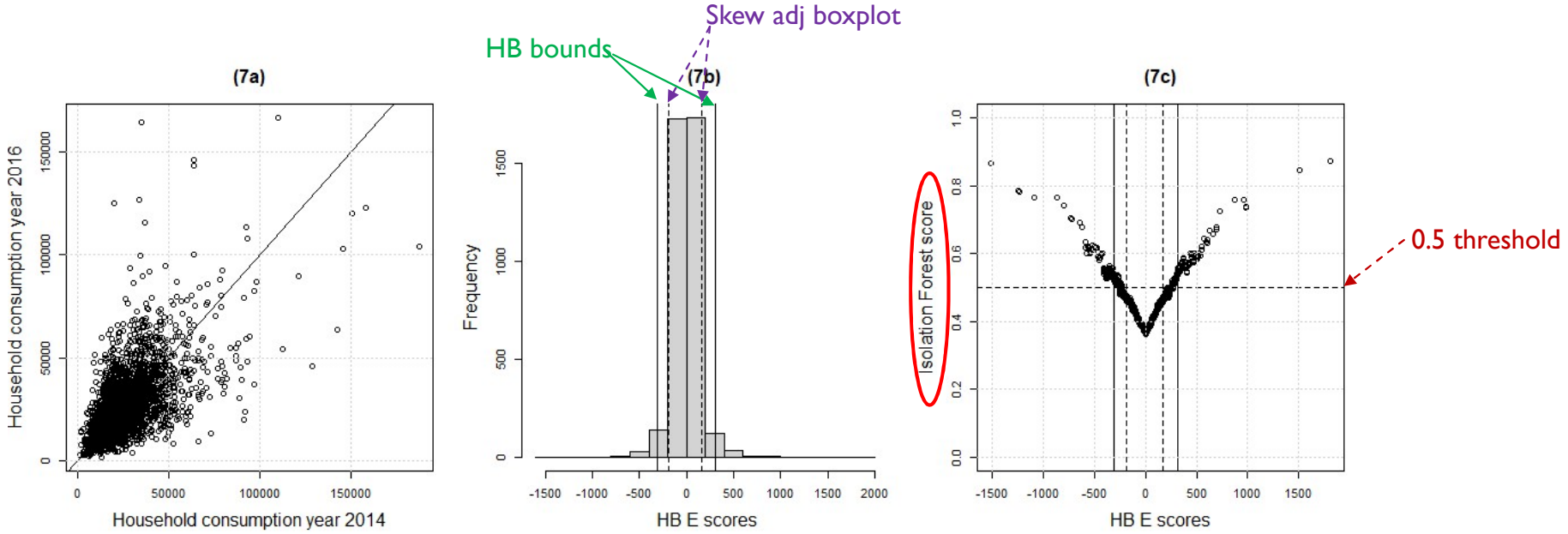
U $(0 \leq U \leq 1)$ controls the role of the magnitude in determining the importance associated to the centered ratios (often $U = 0.5$)

Istat

Example with Data: Univariate Case (1/2)

Bank of Italy, Survey on Household Income and Wealth. Public use anonymized microdata distributed for research purposes; expenditures of 3 804 households in years 2014 and 2016.

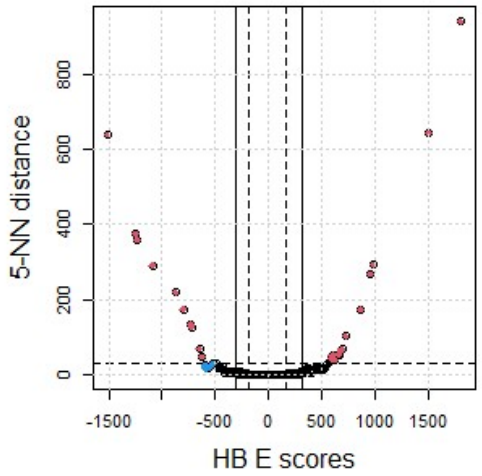
Hidiroglou-Berthelot (with $C = 7$ and $A = 0.5$); Skewness adjusted boxplot ($M = -0.024$); isolation forest

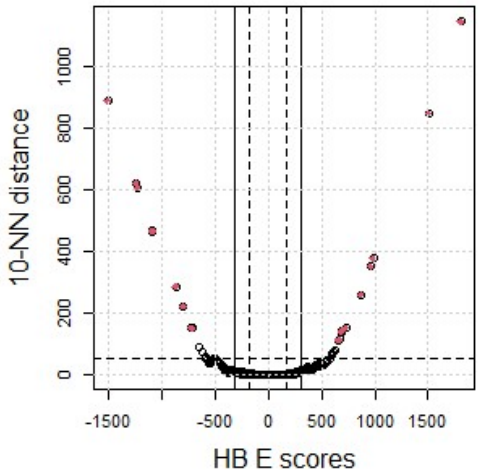


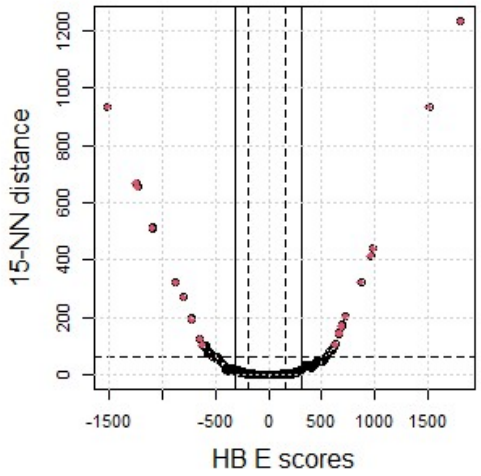
Example with Data: Univariate Case (2/2)

Bank of Italy, Survey on Household Income and Wealth. Public use anonymized microdata distributed for research purposes; expenditures of 3 804 households in years 2014 and 2016.

k-NN (*k*=5,10, 15) and DBSCAN (minPts=6 & $\varepsilon = 30$; minPts=11 & $\varepsilon = 55$; minPts=16 & $\varepsilon = 65$)







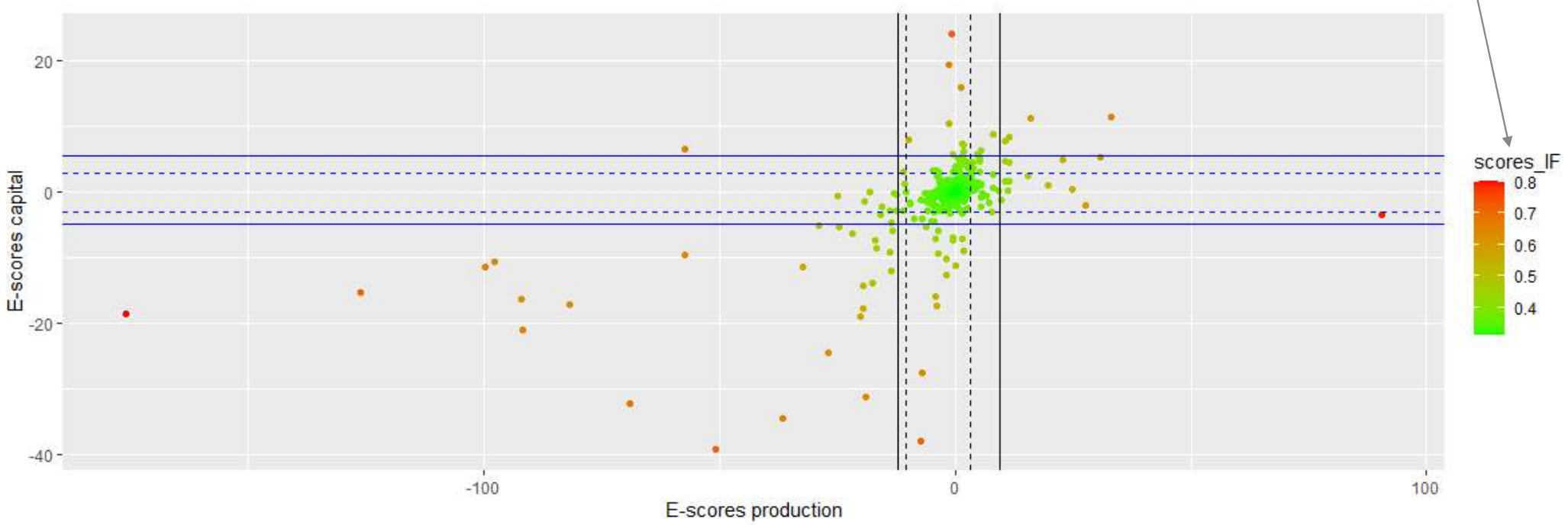
Red-color points are "noisy" points (outliers) identified by DBSCAN

Istat

Example with Data: Bivariate Case (1/3)

R&D performing US manufacturing; production and capital of 509 firms in 1982 and 1983. (https://www.nuffield.ox.ac.uk/users/bond/index.html, See also the R package **pder**)

Hidiroglou-Berthelot (with $C = 7$ and $A = 0.5$); Skewness adjusted boxplot (-0.21; -0.027); isolation forest

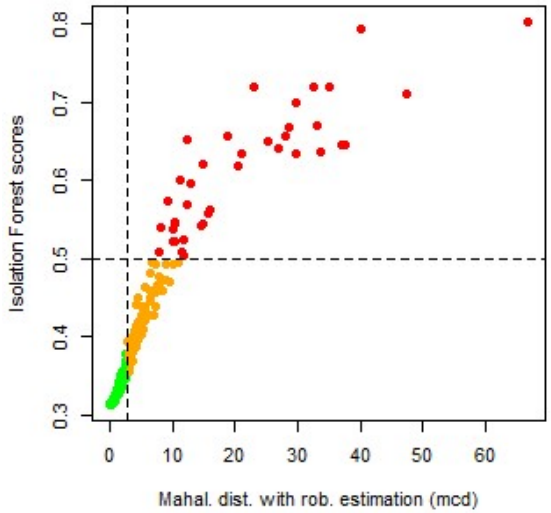


Istat

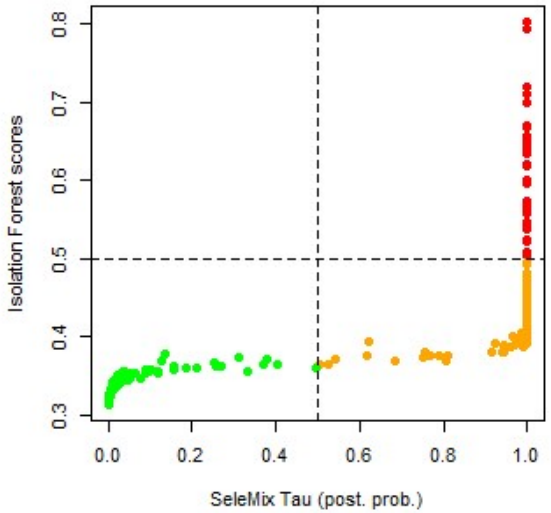
Example with Data: Bivariate Case (2/3)

R&D performing US manufacturing; production and capital of 509 firms from 1982 to 1983. (https://www.nuffield.ox.ac.uk/users/bond/index.html, See also the R package **pder**)

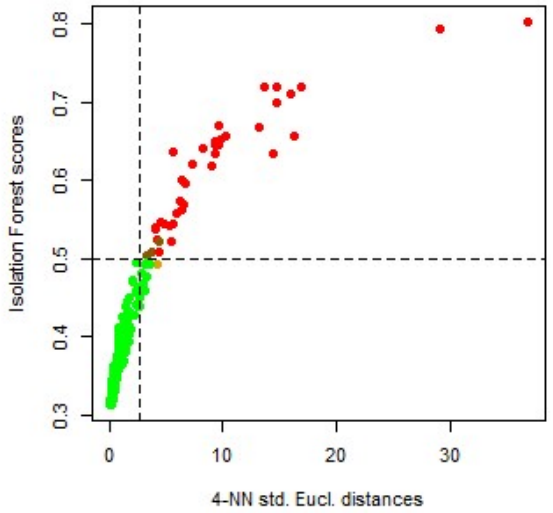
Mahalanobis dist. with rob. estimation (MCD)



SeleMix



DBSCAN



red-color: outliers identified by both the compared methods
green-color: non-outliers according to both the compared methods

Istat

Example with Data: Bivariate Case (3/3)

R&D performing US manufacturing; production and capital of 509 firms from 1982 to 1983. (https://www.nuffield.ox.ac.uk/users/bond/index.html, See also the R package **pder**)

		Isolation forest scores				
MD rob (MCD)	SeleMix	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 1]
not outlier	not outlier	380				
	outlier					
outlier	not outlier	9				
	outlier	30	50	17	17	6

		DBSCAN	
MD rob (MCD)	SeleMix	not outlier	outlier ("noisy")
not outlier	not outlier	380	0
	outlier	0	0
outlier	not outlier	9	0
	outlier	82	38

	Isolation forest scores				
DBSCAN	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 1]
not outlier	419	49	3		
outlier ("noisy")		1	14	17	6

Istat

Conclusions

Nonparametric density(distance)-based (*k*-NN and its variants):

- Sensitive to the decisions on: scaling of variables, distance function, *k*
- Produces a score, obs. with highest score are potential outliers (no thresholds, difficult to set)
- Many variants to efficiently compute (approx) distance with very large data sets but limited set of distance functions

DBSCAN clustering

- Sensitive to the decisions on: scaling of variables, distance function, distance threshold & def. of reachability
- No score to rank units
- Tend to identifies relatively few observations that have a high chance of being outliers

Isolation forest

- No need to transform variables
- Relatively few tuning parameters (just number of trees to build when there are relatively few obs.)
- score to rank units lies in the (0,1] interval
- **score>0.5** rule-of-thumb often does NOT work
- Need to use extended method in the multivariate setting (the "extended isolation forest" works quite well)

Istat

Thank You

Marcello D’Orazio | marcello.dorazio@istat.it

Istat | Istituto Nazionale di Statistica