Istat | Istituto Nazionale di Statistica

useR! 2026

UnitMix in Production: Multivariate Gaussian Mixtures for Robust Detection of Scale Errors and Outliers

Cristina Faricelli, Renato Magistro
Istat | DCME

What are scale errors

Unit-of-measurement (scale) errors in a survey: whole blocks of records reported on the wrong scale.

€ → thousands of €	q → kg	kWh → MWh

- They are **systematic**: the same factor shifts many records at once
- Each single value is **easily mistaken for a genuine outlier**
- A **simple range check can't distinguish** between erroneous and corrected values.

Istat

Why this matters in official statistics

Accuracy

- Systematic errors don't cancel out; **they bias the estimates**, not just add noise.
- Our estimates are **weighted**, so we need the complete dataset. Dropping units breaks the **sampling design**.
- **We release the microdata**: an error correction phase is not avoidable.

Timeliness & reproducibility

- Manual scale-error correction eats a lot of time in the editing phase and is not reproducible.
- Timeliness and reproducibility are very important quality dimensions in Official Statistics.
- So we turned the method into a reusable, reproducible package instead of one-off scripts.

Istat

Detecting scale errors: no shared standard

The most extensive treatment (de Waal, Pannekoek & Scholtus, 2011) **proposes no single recognized standard and the methods used in routine practice are largely heuristic.**
The alternatives proposed include:

- **Ratio edits / bounds**

  Compare the raw value to a reference (prior period, robust median, register); flag it if the ratio falls outside [l, u].
  The bounds are a manual trade-off, wide misses errors, narrow creates false hits

- **Digit difference (Al-Hamad et al., 2008)**

  Count the order-of-magnitude gap between value and reference: a thousand-error simply shows up as a digit gap of 3.

- **Explicit modelling (Di Zio, Guarnera & Luzi, 2005)**

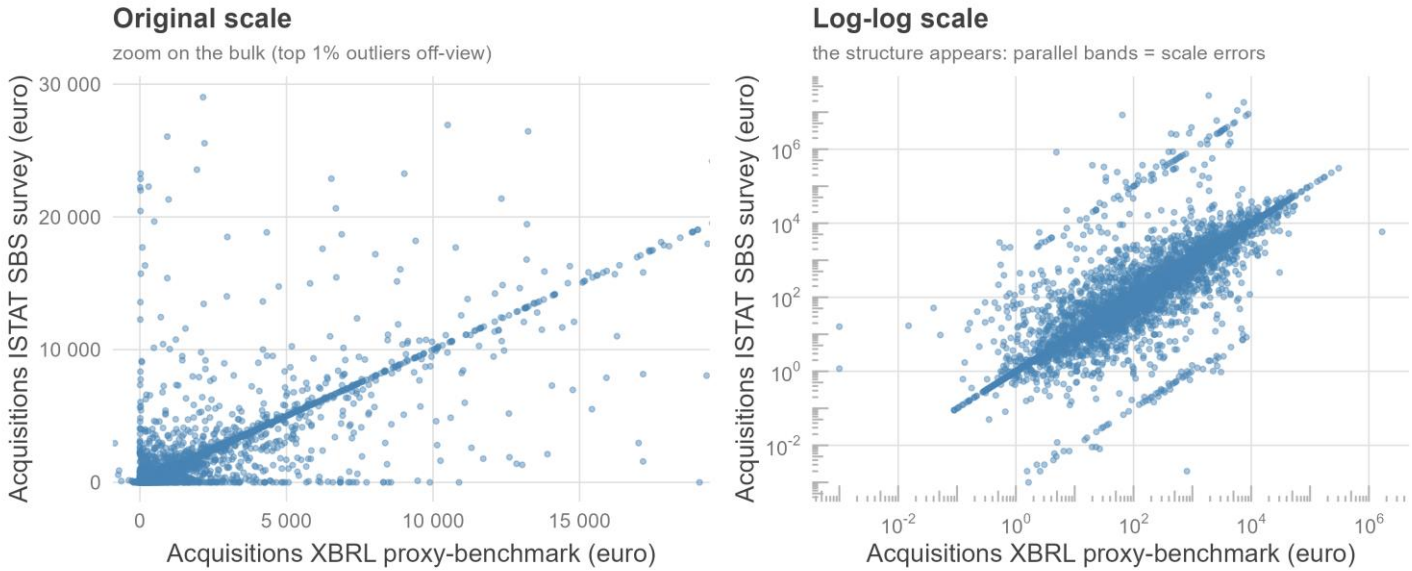
  Model the error generatively instead of by thresholds, it is a Model-based alternative. **The route UnitMix builds on.**

→ Common limits: mostly one variable vs one reference, threshold-dependent, and reliant on a clean benchmark.
UnitMix generalizes the model-based route, **multivariate, uncertainty-aware, reproducible**.

Istat

What UnitMix does

It fits a Gaussian mixture on the log-transformed variables, and assigns each record either to an **error group** or to the **correct one**, based on its posterior probability and entropy.



Assumptions & method

Assumptions

- At least two positive continuous variables with a reasonably stable ratio (they may come from different sources).
- Ideally one error-free variable, it helps, but is not strictly required.
- Log-normal data; the method still works with some skew or heavy tails.
- Typical pairs: expenditure vs quantity · same variable in two periods · value vs register.

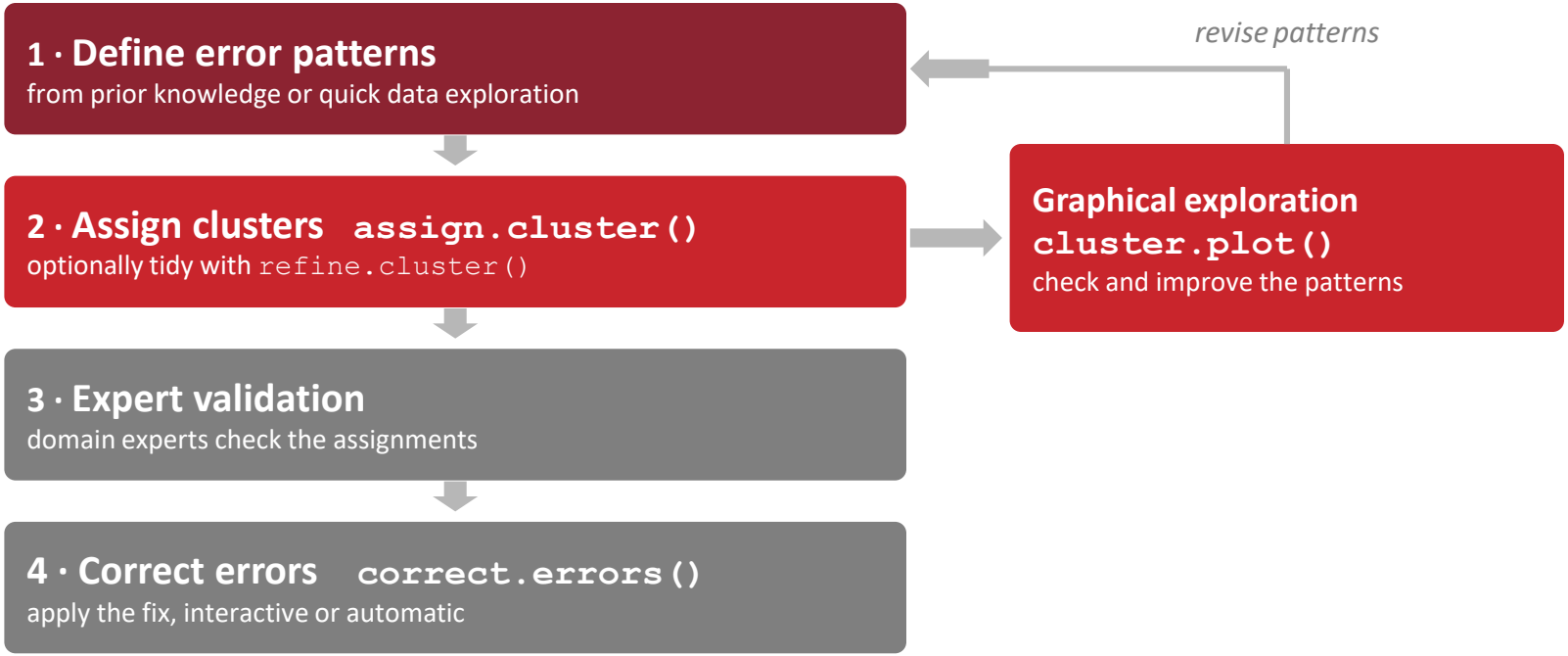
Method, in short

- A constrained EM: the cluster centres are fixed by the patterns, so only the weights and the shared covariance are estimated.
- A record is labelled by the highest posterior probability; thresholds on posterior probability and/or entropy can optionally be applied to leave uncertain records Unassigned.
- Once assigned, the correction is deterministic (after expert review).

Lightweight: depends only on `mvtnorm`.

Istat

Workflow: four steps



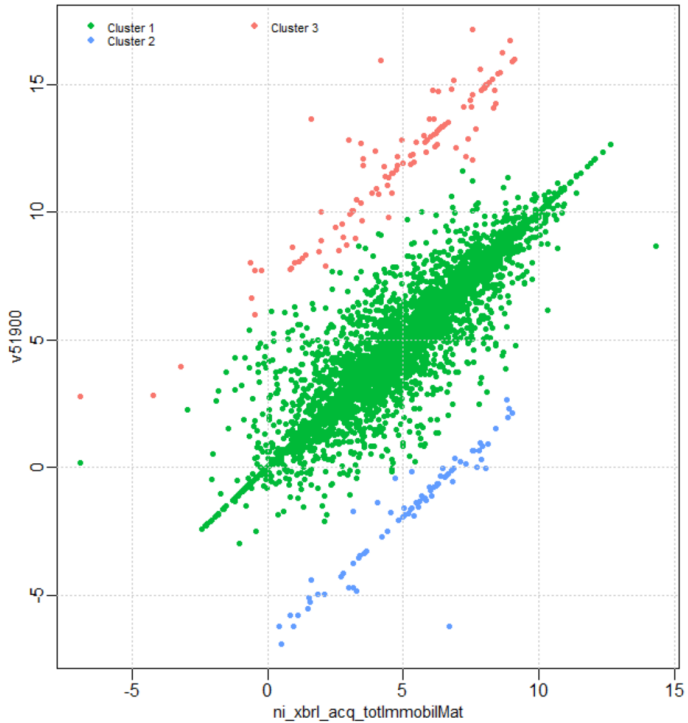
Istat

Minimal example

```
install.packages("UnitMix")    # or
remotes::install_github("cristinafaricelli/UnitMix")
library(UnitMix)
# variables to check (survey value vs XBRL benchmark)
vars <- c("v51900", "ni_xbrl_acq_totImmobilMat")
# candidate error patterns
errorPatterns <- list(
  c(0, 0),          # correct
  c(0.001, 0),      # x 1000x too small
  c(1000, 0)        # x 1000x too large
)
ac <- assign.cluster(db = data, var = vars, errorPatterns =
errorPatterns)
table(ac$data$cluster)
cluster.plot(data, vars, ac$postprob)
```


Istat

Case study 1: Enterprise acquisitions (with a benchmark)



The XBRL source (Chamber of Commerce financial statements notes) is a reliable benchmark assumed error-free. It guides the clustering of the suspect survey values, and separates consistent records from scale errors.

`c(v51900, ni_xbrl_acq_totImmobilMat)`

Pattern	Meaning	Freq.
`c(0, 0)`	correct	8769
`c(0.001, 0)`	1000× too small	113
`c(1000, 0)`	1000× too large	165

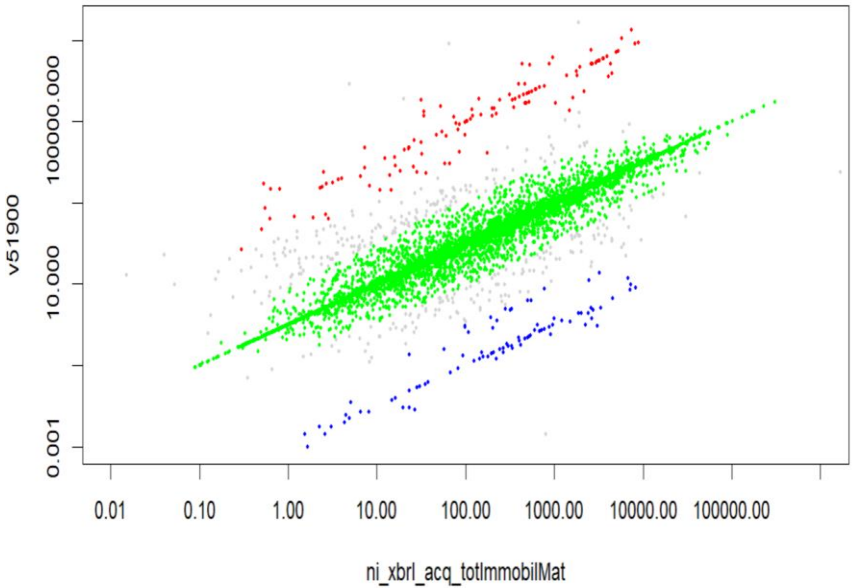
Istat

Refine & Correct: Case study 1

refine.cluster() uses Mahalanobis distance to flag borderline records as Unassigned before applying corrections. It flags 243 borderline records as Unassigned (0).

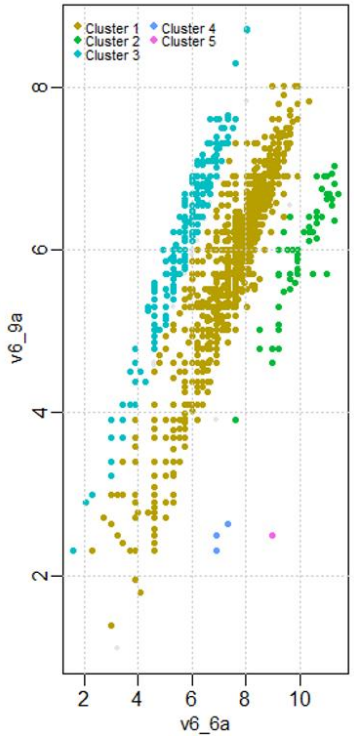
```
# Refine assignments (Mahalanobis distance)
rc <- refine.cluster(ac,
  compat_level = 0.998,
  min_good_prop = 0.001)
table(rc$data$cluster_refined)

# Apply corrections
corrected <- correct.errors(db = data, var = vars,
  corrClusters = rc$data$cluster_refined,
  errorPatterns = errorPatterns)
```

Pattern	Meaning	Freq.
	Unassigned	243
`c(0, 0)`	correct	8585
`c(0.001, 0)`	1000× too small	97
`c(1000, 0)`	1000× too large	122



Istat

Case study 2: Household energy consumption and expenditure

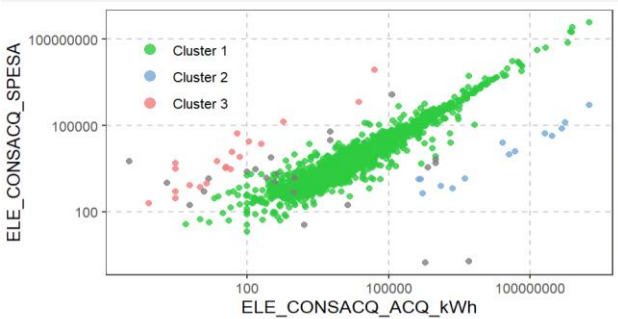


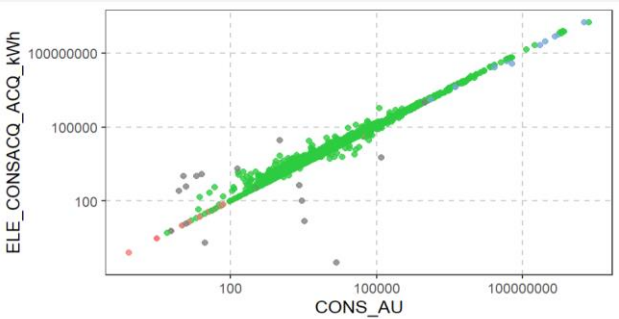
No benchmark here, both variables can be wrong. Patterns were defined with domain experts: expenditure is taken as correct, the purchased quantity as possibly wrong.

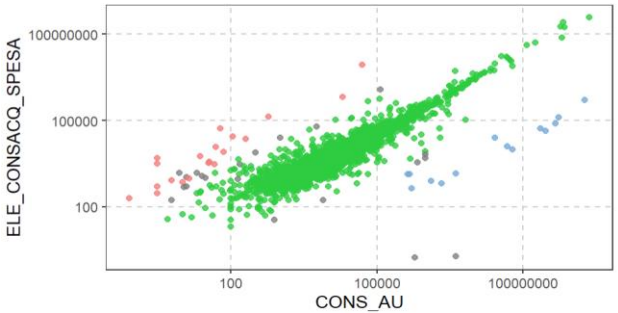
`c(v6_9a, v6_6a)` firewood consumption in kg, expenditure in €

Pattern	Meaning	Freq.
`c(0, 0)`	correct	1588
`c(100, 0)`	100× too large	51
`c(0.01, 0)`	**100× too small (q ↔ kg)**	**159**
`c(0.001, 0)`	1000× too small	3
`c(0.0001, 0)`	10000× too small	1

Case study 3: enterprise energy (three variables)







Clusters are defined jointly over three variables. The errors here sit on the expenditure variable, while the two energy quantities anchor the fit. CONS_AU represent the ELECRICITY METER consumption.

`c(ELE_SPESA_euro, ELE_ACQ_kWh, CONS_AU_kWh)`

Pattern	Meaning	Freq.
`c(0, 0, 0)`	correct	3655
`c(0.001, 0, 0)`	spend 1000× too small	14
`c(1000, 0, 0)`	spend 1000× too large	20

Future developments

- **Data-driven error patterns**

Suggest the error patterns automatically from the data, on top of the user-defined ones.

- **Relax the log-normality assumption**

Incorporate the mixture-of-mixtures model for skewed / heavy-tailed clean data (Di Zio, Guarnera & Rocci, 2007).

- **Compare with other outlier methods**

Benchmark and integrate other detectors; isolation forest already under study.

- **Shiny interface for non-R users**

Interactive validation and correction, to open the tool beyond R users.

Istat

Resources

Get UnitMix

```
install.packages("UnitMix")
```

GitHub: github.com/cristinafaricelli/UnitMix

Istat Methods & software for the statistical process:

https://www.istat.it/en/classifications-and-tools/methods-and-software-of-the-statistical-process/

Method reference

Di Zio, Guarnera & Luzi (2005) - Editing systematic unity measure errors through mixture modelling.

Related Istat tools

SeleMix → outliers and influential errors, via a contamination model.

unitOutl → a suite of methods for outlier detection.

All developed at Istat, in the same repository

Thank you!

Cristina Faricelli · cristina.faricelli@istat.it Renato Magistro · renato.magistro@istat.it

Istat