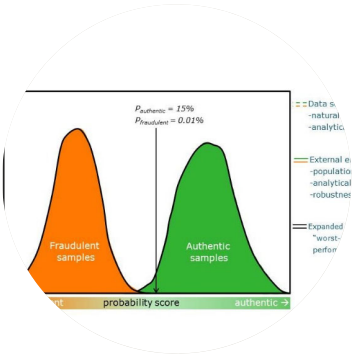
Quality markers and performance characteristics for non-targeted testing methods

Eurachem 2022 – Quo Vadis pt. 1

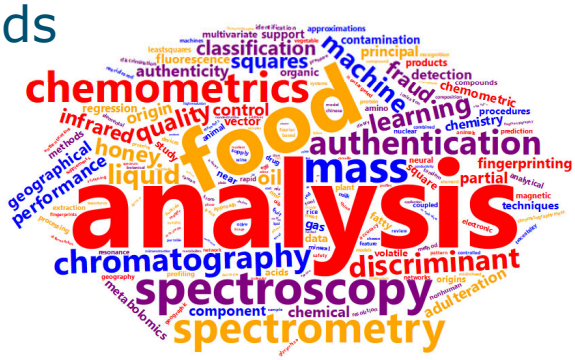
Martin Alewijn - WFSR (Wageningen University and Research), the Netherlands



WAGENINGEN UNIVERSITY & RESEARCH

Non-targeted testing methods

- .. As just defined *(Stephan Walch)*
- Yields vast amounts of data
- Data workflow may vary with application
- *This presentation:* focus on (food) authentication methods



Wordcloud from titles+keywords Scopus search:
(TITLE-ABS-KEY (non-targeted OR "non targeted" OR untargeted OR fingerprinting OR profiling OR chemometrics OR "machine learning") AND TITLE-ABS-KEY (food AND (fraud OR authenticity OR authentication)))

Analytical data + **Known samples** + **Model** = **Decision**

WAGENINGEN UNIVERSITY & RESEARCH

Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL EUROPE SECTION

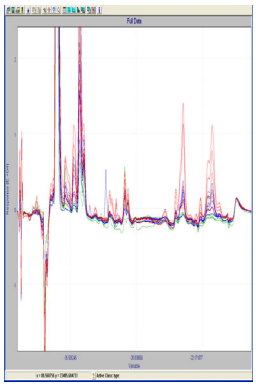
Egg-sample of a non-targeted testing method

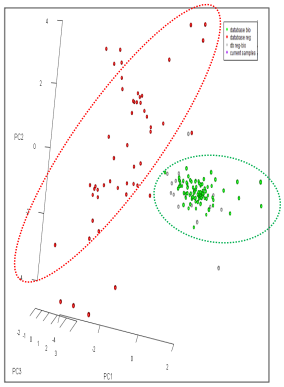
450 eggs
Balanced over NL

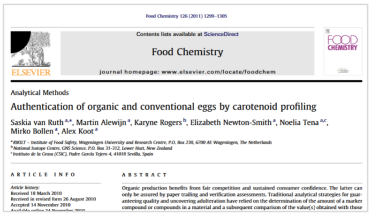
HPLC-DAD
carotenoid analysis

Classification
modelling

Publication







Accreditation!
(Sept 2022)

DUTCH ACCREDITATION COUNCIL

Accreditation Nr. L014
https://www.rva.nl/en

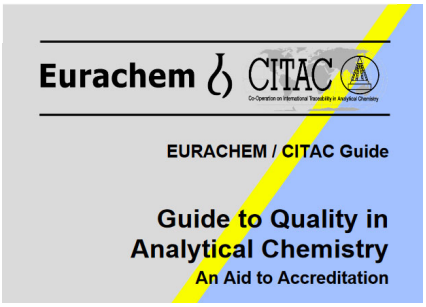
WAGENINGEN
UNIVERSITY & RESEARCH

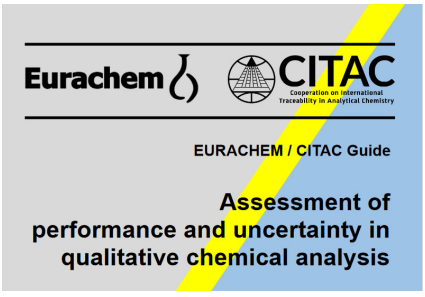
Method performance

Well-established for targeted analytical methods

But hard to translate for non-targeted testing methods...

- Many “analytes”
- Typically, no direct relation between analytes and the property measured
- Binary result
- Result based on reference sample set





WAGENINGEN
UNIVERSITY & RESEARCH

Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL
EUROPE SECTION

AM0

(some) Performance characteristics

Conventional/targeted:

Accuracy
Precision
Linearity
Selectivity
Specificity
Application range
LOD
LOQ
Recovery
Robustness
Repeatability
Reproducibility
Measurement Uncertainty
CCα
CCβ

Classification:

Accuracy
Precision/positive predictive value
Recall/Sensitivity/True positive rate
Selectivity/Specificity/True neg. rate
False positive rate/Type I error
False negative rate/Type II error
F-1 Score (F-β score)
Likelihood ratio
Youden's index (J)
Kappa (Cohen/Fleish)
MCC (Matthews correlation coeff)
Kolmogorov-Smirnov statistic
Log-loss
Brier score
AUROC (AUC score)

Lack of established performance metrics

- Makes it hard to appraise methods or compare them
- Hinders (official) use
- Some are difficult to interpret, sensitive to dataset balanced-ness & non-normal behavior, discrete nature
- (No performance metrics for reference set quality?)
- Aim to develop a harmonized validation protocol for non-targeted testing methods in food authenticity testing in CEN TC/460 "Food authenticity"

CEN/TC 460/WG 5

WAGENINGEN UNIVERSITY & RESEARCH

Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL EUROPE SECTION

Slide 5

AM0 https://neptune.ai/blog/evaluation-metrics-binary-classification
Alewijn, Martin; 2022-11-08T20:18:17.632

Quality levels in non-targeted testing methods

3) Routine use quality (extrapolation)

Power to predict correct results:
- using new (routine) samples – extrapolation of reference set
- time after method development – analytical & population drifts

2) Developed model quality (optimisation)

Power to predict correct results:
- based on reference samples
- using a mathematical model

1) Analytical quality

- Performance of the analytical method

WAGENINGEN UNIVERSITY & RESEARCH

1) Analytical quality metrics

- Conventional QC on $\sqrt{n}$ (or selected) variables
- Apply conventional QC to raw model score:
  - r, R, long-term monitoring reference samples
  - Quantify replicate suitability
  - (may not always follow a normal distribution)
- Alternatively, use *n* PCA scores or mahalanobis distances
  - Use appropriate scaling
  - Excludes model's variable weight

WAGENINGEN UNIVERSITY & RESEARCH

2) Developed model quality

- Accuracy? $\left(\frac{TP+TN}{TP+TN+FP+FN}\right)$
- True negative rate at a predefined acceptable false negative rate

For example:

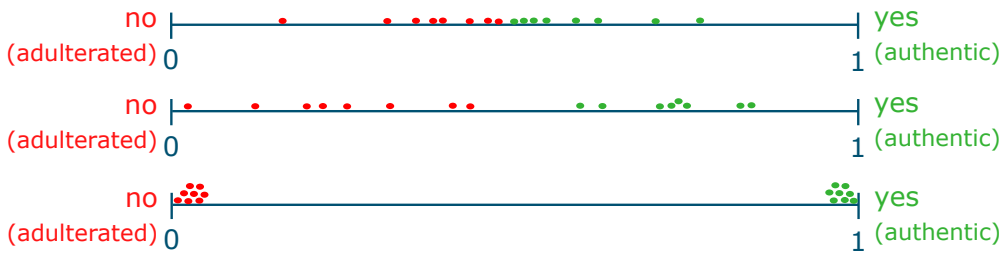
➢ Accepting ≤0.1% false classification of authentic samples, the method correctly detects ≥80% of non-authentic samples

- *How to obtain this information?*

WAGENINGEN UNIVERSITY & RESEARCH

2) Developed model quality

- During method development: rather a metric based on scores



- This mitigates the resolution problem:
- And allows parameter estimation...

Table 4. The minimum number of analyses to find one or more false (positive or negative) result(s).

	Confidence level	
False result rate	95 %	99 %
0.5 %	598	919
1 %	299	459
5 %	59	90

From: "R Bettencourt da Silva and S L R Ellison (eds.) *Eurachem/CITAC Guide: Assessment of performance and uncertainty in qualitative chemical analysis.* First Edition, Eurachem (2021)."

WAGENINGEN UNIVERSITY & RESEARCH

Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL EUROPE SECTION

2) Developed model quality

- To check model quality, cross-validation is usually used.
  - Not really "validation"!
- Split dataset, use part of the data to build a model, predict left-outs

➢ Don't be nice: no leave-one-out. Use random splits or splits where groups of similar samples are left out together.

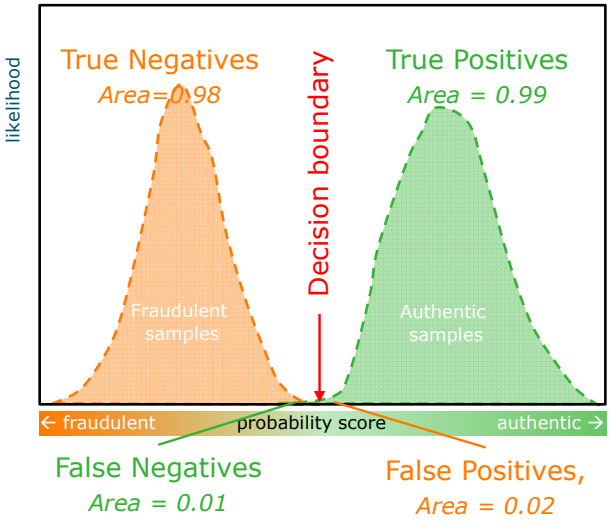
  ➢ For model tuning, inner-loop CV could be used

- Check the performance based on all left-outs...

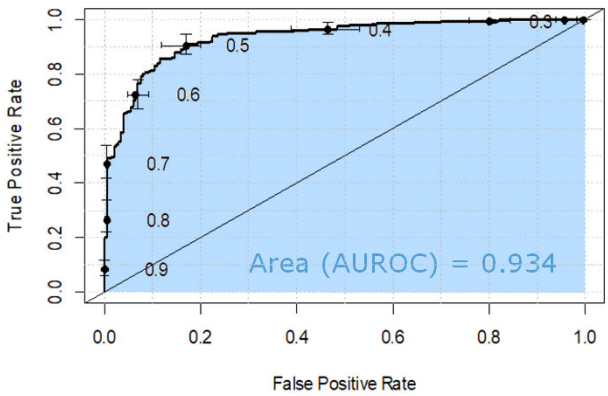
WAGENINGEN UNIVERSITY & RESEARCH

2) Model quality: AUROC and density plot

Model score density plot*



ROC curve



WAGENINGEN UNIVERSITY & RESEARCH

*Empirical probability estimates

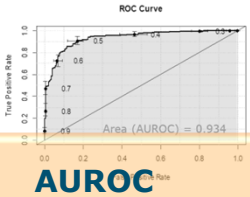
Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL EUROPE SECTION

2) Model quality: AUROC and density plot

Density plots

- Graphical representation of discriminating ability
- Shape diagnostic for subgrouping
- Allow extraction of performance characteristics
- Allows finding a suitable decision boundary to calculate accuracy



AUROC

- One number reflecting discriminating ability
- Based on raw/probability scores
- Insensitive to imbalanced-ness
- For model optimising/tuning
- Performance measure for sets (CV, validation, ...)

WAGENINGEN UNIVERSITY & RESEARCH

3) Routine use quality

- After design, method is fixed (data set *and* analytical workflow *and* data treatment procedures)
- New samples are always an extrapolation, use care!
- Finalise development with an independent validation set:
  - New samples (within scope)
  - New harvest/year/suppliers/technicians/devices (within scope)
- *If possible:* multi-lab trial

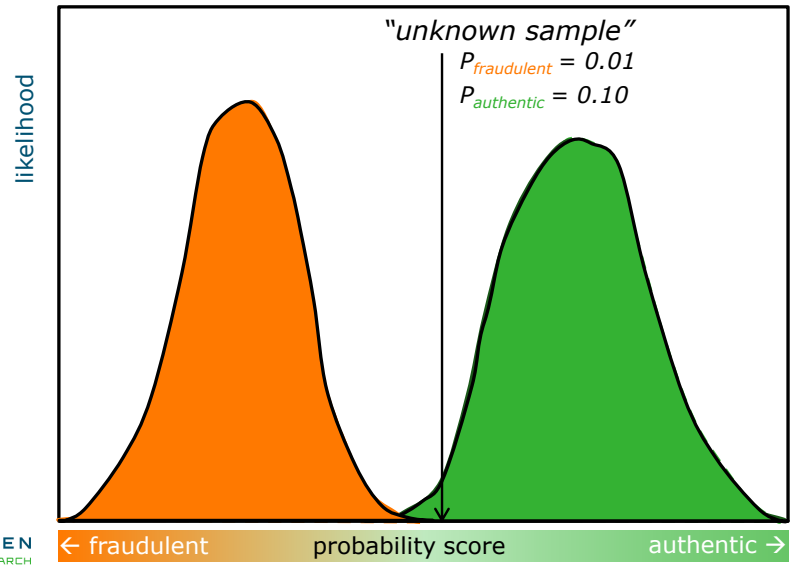
➢Validation samples should meet requirements (AUROC, Spec, FNR, Acc)

WAGENINGEN UNIVERSITY & RESEARCH

Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL EUROPE SECTION

Method sources of error (uncertainty)



Training set error:
-natural variation
-analytical variation
Levels 1) + 2)

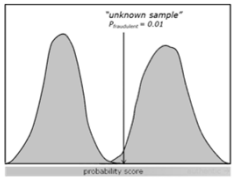
External errors (validation set):
-population change
-analytical drift
-robustness
Level 3)

Expanded (widened) **distribution:**
-"worst-case" model performance
Σ 1) + 2) + 3)

WAGENINGEN UNIVERSITY & RESEARCH

15

3) Routine use quality



Two levels of performance statements can be extracted from "worst-case distribution":

- *Global:* on average, method gives ≥ x% TNR @ ≤ y% FNR
- *Per sample*: classification *and* a confidence statement:
  - Decision (yes/no, authentic/non-authentic)
  - $p_{authentic}$ and $p_{non\text{-}authentic}$
- Maintenance: analytical - model score stability reference sample(s)
- Maintenance: dataset – adding new (QC) samples, update datasets

WAGENINGEN UNIVERSITY & RESEARCH

16

Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL EUROPE SECTION

Discussion

- A few relatively easy to interpret performances metrics suggested
- Easy to obtain, but not from (all) commercial software

- Far from a real-life proven & accepted validation protocol
- More research, and discussion needed
- (A.o. in CEN TC460/WG5)

WAGENINGEN UNIVERSITY & RESEARCH

Thank you for your attention

martin.alewijn@wur.nl

WAGENINGEN UNIVERSITY & RESEARCH

Eurachem
A focus for analytical chemistry in Europe

AOAC INTERNATIONAL EUROPE SECTION