Guidelines for validation of chemometric models for food authentication

RIKILT Report 2011.022

G. van der Veer, S.M. van Ruth and W. Akkermans

Guidelines for validation of chemometric models for food authentication

G. van der Veer, S.M. van Ruth and W. Akkermans

Report 2011.022 **December 2011**

Project number:	120. 72.174.01
BAS-code:	WOT-02-005-006
Project title:	multivariate validation/accreditation
Project leader:	S.M. van Ruth

RIKILT – Institute of Food Safety
Wageningen UR (University & Research centre)
Akkermaalsbos 2, 6708 WB Wageningen, The Netherlands
P.O. Box 230, 6700 AE Wageningen, The Netherlands
Tel. +31 317 480 256
Internet: www.rikilt.wur.nl

This research was funded by Ministry of Economic Affairs, Agriculture and Innovation (EL&I).

Distribution list:

- New Food and Consumer Product Safety Authority (nVWA); dr. Karin Visser

This report from RIKILT – Institute of Food Safety has been produced with the utmost care. However, RIKILT does not accept liability for any claims based on the contents of this report.

Extended summary

Chemical fingerprinting methods (e.g. NMR, NIR and chromatography) in combination with chemometric techniques provide a powerful tool for verifying the authenticity of food and related commodities. In this integrated approach, a multivariate classification model is "trained" to distinguish between authentic and non-authentic food samples based on chemical fingerprinting data. After training, such a model can be used to determine whether a suspect food sample is authentic or not at a certain level of confidence. These authentication methods generally have the following characteristics:

- The test is specifically developed for a single combination of commodity/product and authenticity question.
- The test is based on a multivariate classification model.
- The classification model is empirically derived (i.e. based on a reference dataset).
- The fingerprinting data used for building the classification model is often of a non-targeted or semi-targeted nature.

For application of such food authentication methods in a commercial context, integrated validation of the analytical and statistical aspects of the method is of utmost importance. However, protocols for validating such advanced methods are currently lacking. The aim of this report therefore is to describe a set of generic guidelines for validation of methods that are based on chemometric classification techniques. In addition, the report provides guidelines on how to develop and optimize such classification models. For the purpose of this report we limit the case to binary classifiers.

Procedures for validation of regular analytical methods are laid down in various protocols such as Commission Decision 2002/657/EC and ISO 17025. Herein it is described 1) which performance characteristics need to be determined, and 2) which criteria have to be met for accepting the performance of the method. The methods described in this report are qualitative methods and according to existing protocols, performance characteristics required for validation of qualitative techniques include:

- Detection capability
- Specificity
- Ruggedness
- Stability

For several reasons, however, these performance characteristics are not, or not directly, suitable in a multivariate context. Therefore, several modifications and alternatives are provided that can be used to assess the performance of the analytical aspects (fingerprinting method) as well as the statistical aspects (classification model).

To assess the performance of the fingerprinting method, it is here proposed to use the average standard deviation of:

- The duplicate measurements of the validation set.
- The repeated measurements of several internal standards over time (stability).

The performance of the classification model can be assessed using a number of procedures to calculate the true positive rate (TPR) and true negative rate (TNR):

- Using the individual duplicate (or triplicate) results.
- Internal cross-validation using bootstrapping.
- External validation using a new set of samples.
- Using a permutation test.

The most crucial aspect of the proposed validation procedure concerns external model validation. Only by external validation the generalization ability of the model can be assessed.

The proposed criteria for acceptance of the performance characteristics for fingerprinting method are based on the average standard deviation of the duplicate samples that were used to build the model (i.e. training set), whereas the criteria for acceptance of the model performance are based on minimum values of the TPR and TNR as determined from the bootstrap distribution used for internal cross-validation.

The generic approach described in this report is meant as a general guideline for validation of methods based on chemical fingerprinting in combination with chemometric modelling. It should be kept in mind that for each of the performance characteristics and criteria mentioned here, various alternatives exist which are equally valid or may even be better suited. As such, the optimal approach to validation will depend on the specific question at hand.

Contents

1 Introduction

The basic goal of food authenticity testing is to objectively verify the acclaimed specifications of some product or commodity with respect to its composition, typicality, production method and/or geographical origin. Such problems are commonly addressed by chemical fingerprinting methods such as NMR, MIRS, NIRS and chromatography in combination with chemometric classification techniques (see e.g. Charlton et al., 2002; Møller et al., 2005; Bevin et al., 2006; Bertelli et al., 2010; Van Ruth et al., 2010).

Common questions in food authentication studies for example include: "Are these eggs organic or not", or "can we discern between wines from Bourgogne and wines from Bordeaux?" In statistics this translates to a binary classification problem. Binary classification is the task of classifying the members of a given set of objects into two groups on the basis of whether they have some property or not. More complex situations arise when one wants to discriminate between a set of objects into a number of groups (i.e. a multiclass classification problem). For the purpose of this report, however, we focus on binary classification problems only.

A common approach in food authenticity testing is to develop a classification model (classifier) on the basis of a representative set of reference samples. In case of a binary classification problem, the reference samples belong to either one of the two classes and the allocation to these classes is known. The final classification model then provides a means to predict whether an unknown food sample is authentic or not at a certain level of confidence.

Such an authenticity test has the following general characteristics:

- The test is specifically developed for a single combination of commodity/product and authenticity question.
- The test is based on a (binary) classification model (classifier).
- The classification model is empirically derived (i.e. based on a reference dataset).
- The fingerprinting data used for building the classification model is often of a non-targeted or semi-targeted nature.

Validation of such an authenticity test requires that both the analytical as well as the statistical aspects are integrally evaluated and tested for compliance with the relevant performance criteria. Because such tests are based on an empirical model, it is of utmost importance to assess their generalization ability by external model validation.

Official protocols and standards for validation of these aspects are currently lacking and the aim of this report is to describe a set of generic guidelines for in-house validation of a method for authenticity testing that are based on a combination of chemical fingerprinting techniques and chemometric classification models. In chapter 2 the different aspects relevant to the development of binary classification models are reviewed and discussed, whereas chapter 3 provides a proposed methodology for in-house validation of the aforementioned approach to authenticity testing.

2 Development of a classification model

Development of a binary classification model consists of a number steps which are summarized in Figure 1. These steps are only briefly reviewed in this chapter, and the reader is referred to Massart et al. (1998) and Otto (1999) for further details.

2.1 Raw data

The raw fingerprinting data consist of multivariate/multichannel measurements (e.g. spectra) from a set of samples that belong to a number of known classes, which are regarded as different (sub)populations. For the purpose of this report the number of classes is limited to two (binary classification) and the measurement data is expected to be of a one dimensional nature.

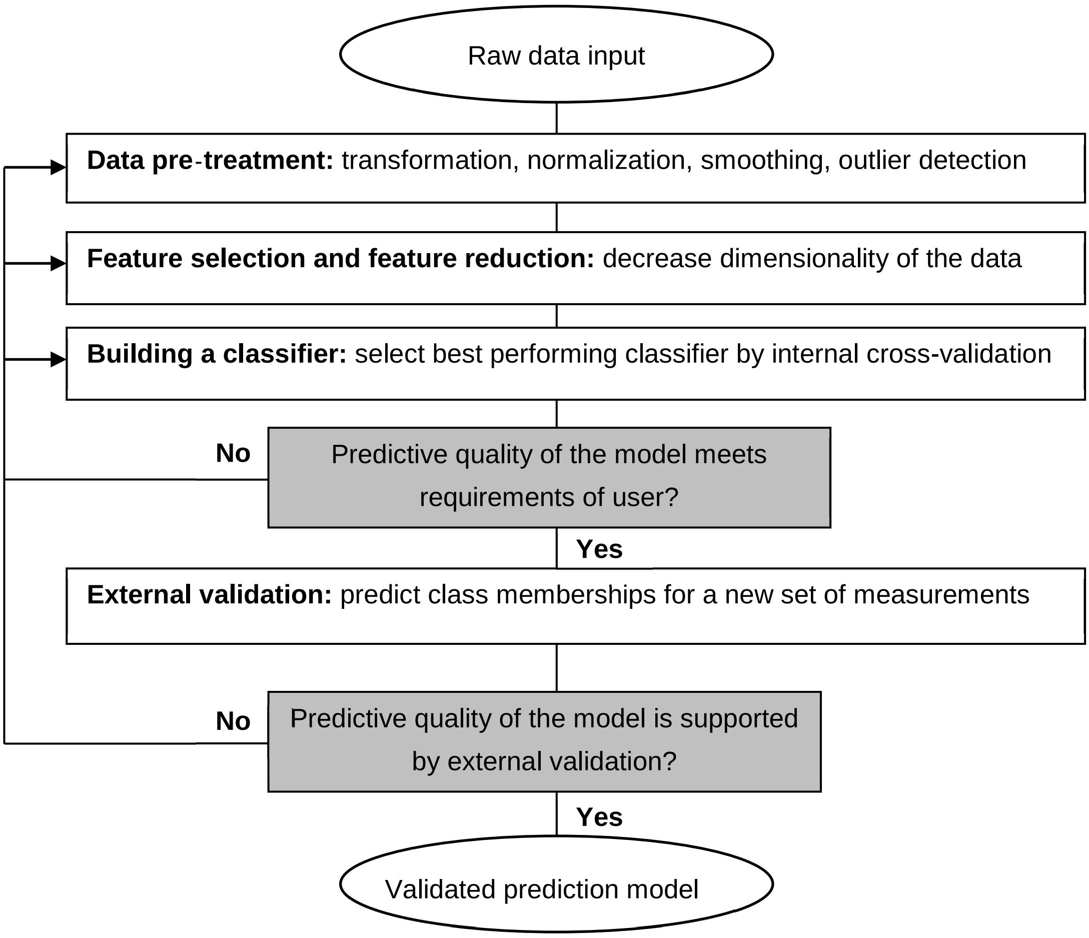
Raw data input
Data pre-treatment: transformation, normalization, smoothing, outlier detection
Feature selection and feature reduction: decrease dimensionality of the data
Building a classifier: select best performing classifier by internal cross-validation
No
Predictive quality of the model meets requirements of user?
Yes
External validation: predict class memberships for a new set of measurements
No
Predictive quality of the model is supported by external validation?
Yes
Validated prediction model

Figure 1 Flow chart showing the different steps in the model development and validation process.

Obviously, one of the most important question in model development is to decide how many samples are required to built a realistic model that has sufficient generalization ability. Unfortunately the answer to this question is not trivial and depends on many factors including the size of - and variation within - the populations of interest, the extent up to which differences between the groups are reflected by the chemical fingerprinting method as well as the stability of these differences over time.

When there is no information about the variation in the data available beforehand, the proper amount of samples required for building a model can only be guessed. In the case that preliminary data is available from both populations, power analysis could be used to determine the minimum amount of samples required for model development (see e.g. Cohen, 1988). Power analysis is based on the effect size, which is usually calculated as the difference between the means of the two population normalized by the standard deviation of one population (assuming equal variance). Next, the effect size is used to calculate the amount of samples needed from both populations using a t test with a predefined power and α error. For an effect size below ~0.6, however, the amount of samples calculated by this method strongly depends on the effect size (Figure 2). An extension of this approach for a multivariate case, which is based on the Mahalanobis distance, has been described in Morse (1999).

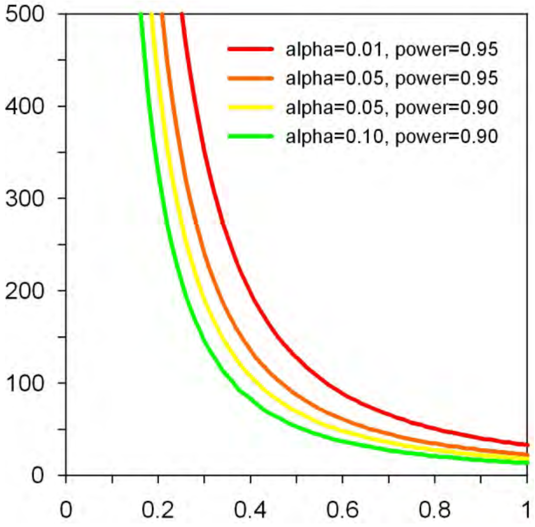

Figure 2 The amount of samples required for each class to test for a significant difference between the classes as a function of the effect size for a univariate case. Calculations assume equal variances of the classes and an equal amount of samples.

Although power analysis can give some guidelines on the amount of samples needed for model building, one should be cautious when the outcome suggests that less than say 50 to 100 samples are required for each class. Such sparse models might not only give under- or overoptimistic results, but they could furthermore hamper feature selection (e.g. Jain and Zongker, 1997). At last it should be noted that multivariate classification methods can be very sensitive to large unbalances in the data, an it is therefore advisable to have roughly an equal amount of samples in each of the two classes (see e.g. Berrueta et al., 2007).

2.2 Data pre-treatment

Data pre-treatment is a crucial step and includes one or more mathematical operations on the data including:

- Pre-processing(e.g. mean centering and scaling)
- Transformation (derivation, smoothing, baseline subtraction, Fourier transforms etc.)
- Aggregation (e.g. merging multiple copies of the same measurements)

During data pre-treatment, missing values should furthermore be dealt with, either by row-wise deletion, replacement by some average value or by estimation of their value(s) using imputation. When missing values occur the categorical information about the samples (i.e. the class to which they belong), the class membership might be guessed based on additional information. Alternatively the sample could be left out for model building.

At this point, also the variables that will be used for modelling are pre-selected whereas variables that contain no relevant information are discarded. This selection step is different from feature selection (section 2.3) in a sense that it is based on knowledge of the researcher. For example because a part of the spectrum is known to contain no useful information, or because the values of some variables are all below the detection limit. Although this seems trivial, this step should be clearly described so that the same procedures can be followed for the validation set.

Furthermore, the raw data should be checked for outliers before further feature selection and model building. This is usually done by drawing a PCA score plot which reveals samples deviating from the bulk of the samples. Alternative approaches for multivariate outlier detection can be found in e.g. Hadi (1994) and Singh (1996). The decision to regard a sample as an outlier is up to a large extent arbitrary and leaving an odd sample out might imply that the initial model performs well under internal cross-validation, but lacks generalization power (i.e. the power to correctly predict the class membership of a new sample). In any case, the criteria used to define an outlier should be clearly described, moreover because the same criteria should be applied to the validation set.

When the data set consists of duplicate or triplicate measurements, a more complex situation might arise; it can be that only one of the two or three duplicate or triplicate measurements is outlying, or it can be that the duplicate or triplicate measurements as a whole are outlying compared to the other sets. In the first case it would make sense to remove the one outlying measurement, but in the later case the choice for exclusion is not so trivial as the duplicate or triplicate results are internally consistent.

In any case therefore, detection of outliers should be done before merging of the individual measurements by e.g. averaging. Merging of the data is important because retaining the duplicate or triplicate measurements in the dataset to artificially increase sample size should be avoided at all times (e.g. Berrueta et al., 2007). For triplicate measurements, taking the median values in stead of the averages will provide more robust results.

2.3 Feature selection

In many fingerprinting datasets, the amount of variables outnumbers the amount of samples by far. This is often referred to as “the curse of dimensionality”, which easily leads to overfitting of the model during model building and optimization. Overfitting takes place if the model learns the idiosyncrasy of the data; then, the noise is modelled as well, and the model looses its generalization ability (Berrueta et al., 2007). Overfitting can be expected when the number of variables is larger than (N – c)/3, where N is the number of samples and c is the number of classes (see Defernez and Kemsley, 1997).

Feature selection, also referred to as variable selection or feature reduction, aims to reduce the dimensionality of the data whereby maintaining the information content that is present. Especially when the data is of a high dimensional nature and/or has a high degree of redundancy, feature selection is strongly recommended. Different feature selection algorithms have been developed which can broadly be classified into three categories: filter, wrapper and embedded models (Zhao et al., 2010).

The filter model relies on the general characteristics of data and evaluates features without involving any learning algorithm. Examples of such methods include standard deviation ranking or feature selection based on Fisher weights. In the first method only variables with a high standard deviation are retained for modeling, whereas in the second method variables are selected by evaluating the difference between the mean of the classes compared to their variance. This approach is similar to calculating the effect size (see section 2.1).

The wrapper model requires a predetermined learning algorithm and uses its performance as evaluation criterion to select features. Algorithms with embedded models incorporate variable selection as a part of the training process, and feature relevance is obtained analytically from the objective of the learning model. Berrueta et al. (2007) also provides an overview of different variable selection as well as reduction approaches. As for the data pre-processing, it is important to clearly describe the variables selected for model building and to use the same set of variables in the external validation phase.

2.4 Building a classifier

There are many different supervised classification algorithms available, - e.g. support vector machines (SVM), artificial neural networks (ANN), classification and regression trees (CART), partial least squares discriminant analysis (PLS-DA), soft independent modelling of class analogy (SIMCA) etc. -, and selection of the optimal classification method is not trivial. An overview of available algorithms and their characteristics can be found in Massart et al. (1998) and Berrueta et al. (2007).

A general distinction can be made between soft and hard classification methods. Soft classification techniques such as SIMCA and UNEQ build frontiers between each class and the rest of the universe. As these class boundaries are allowed to overlap, in some cases samples will be attributed to two or even more classes. When a sample falls outside all boundaries, it is regarded as an outlier. Hard classification techniques such as SVM, ANN and PLS-DA in contrast divide the hyperspace in as many regions as the number of classes. Because these regions do not overlap, such methods will attribute samples to a single class only. In most cases, hard models produce a better discrimination between the classes and are preferred in the framework of authenticity testing.

The choice of classifier should moreover be based on knowledge about the characteristics of the data in terms of the number of variables, the presence of multi-variate normality and the presence of non-linearity. Note that most classification tasks can be performed using linear methods such as LDA, CVA and PLS-DA, and non-linear methods such as SVM and ANN are rarely needed (Beruetta et al., 2007). Furthermore, LDA, UNEQ, CART, ANN and SVM appear especially sensitive to overfitting (Beruetta et al., 2007). Overfitting can also be an issue in PLS-DA and occurs when too many latent components are selected. These latent components, which are sometimes referred to as factors, are uncorrelated linear transformations of the original predictor variables, and are basically used to reduce the dimension of the data before performing DA.

Finding the optimal classifier is usually this selection is done empirically, i.e. by comparing performance of various classification methods using internal cross-validation such as k-fold-cross-validation or a bootstrapping approach. Note that leave-one-out cross-validation should be avoided because it has a strong tendency to overfitting and underestimating the true prediction error (Baumann, 2003). In general it is better to use leave-multiple-out cross-validation (Baumann, 2003), repeated-k-fold cross-validation (Cruciani et al., 1992; Baroni et al., 1992) or bootstrapping (see e.g. Efron and Gong, 1983).

Internal cross-validation normally proceeds by taking the following steps:

1. Split the samples into two groups: a training and a test set
2. Train each classifier on training set, and test the classifier on the test set
3. Repeat the previous steps many times
4. Collect performance characteristics of each classification method
5. Choose the classification method with the best performance

The choice of the optimal classifier could be based on the best overall performance in terms of true positive rate and true negative rate or a similar measure of model performance. The true positive rate of the model is defined as:

- True positive rate = True positives / (True positives + False negatives)

And the true negative rate as:

- True negative rate = True negatives / (True negatives + False positives)

The true positive rate (also referred to as sensitivity, hit rate or recall) and the true negative rate (also referred to as specificity) are the basic performance characteristics of a classification model and can be easily derived from the confusion matrix (see Table 1). Herein, the authentic samples are the “positives” and the non-authentic samples the “negatives” under the hypothesis:

H_0 : The sample belongs the authentic population
H_1 : The sample does not belong to the authentic population (i.e. is not authentic)

In authenticity testing, one is generally more worried about the false negatives (i.e. saying a truly authentic sample is not authentic, Type II error) than about the false positives (i.e. not noticing a not authentic sample, Type I error) because of financial and juridical consequences. As such, classification models with a high true positive rate but a lower true negative rate are generally preferred over models with a high true negative rate and lower true positive rate. Obviously it is up to the user to decide whether the model's predictive power is sufficient or not.

If the predictive quality of tested models does not meet the desired criteria, an alternative set of classification algorithms, feature selection approach or data pre-processing could be considered. Alternatively, another fingerprinting method may be employed that provides a better classification model. Care should be taken when the model performs well under a specific set of conditions, but not under any of the other conditions. This might be an indication that the model is overfitted or otherwise overoptimistic.

Table 1 Example of a 2x2 contingency table (also know as "confusion matrix").

		Outcome of test	
		Authentic	Not authentic
True identity	Authentic	True positive (p=1-α)	False negative (p=β)
	Not authentic	False positive (p=α)	True negative (p=1-β)

2.5 External validation of the classifier

External validation is of great importance to assess the generalization ability of the method. In this approach, the model performance is evaluated using a new set of samples – the external validation set – which is then compared against the previously established performance criteria. This procedure is explained in more detail in the next chapter.

2.6 Model conformance and interoperability

To ensure model conformance and interoperability it is of importance that all details concerning the final model are recorded in a standardized way. A common standard for complex statistical models is the Predictive Model Markup Language (PMML). PMML is an XML-based markup language developed by the Data Mining Group (DMG; see http://www.dmg.org/). The markup language provides a way for applications to define models related to predictive analytics and data mining and to share those models between PMML-compliant applications.

PMML consists of the a number of components in which information about various modelling aspects such as pre-processing steps, dealing with missing values, model details and validation are stored in a standardized way. PMML is supported by a range of software products including R, SPSS and Statistica.

3 Method validation

This report focuses on in-house method validation, for which the procedures are laid down in RIKILT protocol RSV F0052 and related protocols. The procedures described herein are directly based on Council Directive 2002/657/EC and ISO 17025. For in-house validation at RIKILT, the method should be described using a standard protocol (RSV), and the validation process should be laid down in a validation plan (see RVS F0052). In the validation plan the following topics should be addressed:

- Goal and scope of the method
- Method used for validation
- Status of the method
- Type of method (qualitative/quantitative)
- Performance characteristics
- Criteria for evaluation of performance characteristic
- Experimental design
- Description of deviations from standard protocols for method validation

These topics will be reviewed and discussed in the following sections.

3.1 Goal and scope

As mentioned in the introduction, the goal of authenticity testing is to verify the acclaimed specifications with respect to composition, typicality, production method and/or geographical origin of some commodity or commodity at a given level of confidence. The null hypothesis and alternative hypothesis for such a test would be:

H_0 : The sample is authentic
H_1 : The sample is not authentic

In principle, the hypotheses could be defined the other way around and it is important to explicitly state the null and alternative hypothesis used for testing.

With respect to the scope of the test the following points should be described:

- The commodity or product for which the test applies
- The production area or producers for which the test applies
- Optional: the production period for which the test applies

It is important to describe the commodity or product for which the test applies in detail. This includes not only the truly authentic commodity or product (i.e. positive group), but also similar not authentic commodities or products (negative group). For the authentic commodity this description should furthermore include information about which property or characteristic the authenticity is based on (e.g. geographical origin, typical composition/ingredients, type of production method or year of production).

Because the model used for testing is based on a representative set of samples from a confined region and/or number of producers, the test will – at least at first instance – only be valid for that region and/or those producers, or similar producers within the region concerned. Preferentially, the generalization ability of the model should be motivated using additional knowledge about the variation in the data.

Unless the data is representative for all relevant production periods, or unless the temporal variation is deemed of lesser importance, also the specific production period(s) for which the test is valid should be included in the description of the scope. Again, this should preferentially be motivated by additional knowledge about the variation in the data.

3.2 Method of validation

In general, the method of validation depends on the matrix-analyte combination (see RVS F0052). For certain matrix-analyte combinations legislative protocols have already been defined, and these should be followed. In all other cases, validation should proceed according to RVS A0906. This in principle also applies for the validation of methods for authenticity testing as described in this report. However, RVS A0906 was not developed for integral validation of untargeted fingerprinting methods in combination with binary classification models. This report therefore provides a set of alternative procedures for validation of such methods.

3.3 Status of the method

With respect to status, a method can be classified as “conformable to a reference method”, “similar to a reference method” or “own method” (see RVS F0052). The methods for authenticity testing will at first instance be developed as an own, or in-house, method, and this implies that the relevant requirements are to be defined by the RIKILT, or alternatively by the customer for which the method is developed.

3.4 Type of method

Two basic types of methods can be discerned; quantitative and qualitative methods. In general, an authenticity test is a qualitative test (yes/no) for which quantitative statements are made about the confidence level of the test. In Council Directive 2002/657/EC, two types of qualitative methods are recognized: screening methods and confirmatory methods. Whereas screening methods are generally used as a rapid method to detect suspicious samples at a given level of confidence, confirmatory methods allow to confirm the presence or absence of a substance beyond any reasonable doubt. As such, a method for authenticity testing as described in this context will generally classify as a screening method.

3.5 Relevant performance characteristics

The relevant performance characteristic that are required for method validation depend on the type of method as well as the status of the method. The relevant performance characteristics that are required for validation of a screening method according to RVS A0906 and Council Directive 2002/657/EC include:

- Detection capability
- Specificity
- Ruggedness

- Stability

For a confirmatory method, in addition also the decision limit (CCα) is required. For clarity, the definitions of these terms are summarized in Table 3. To be applicable in the context of a method for authenticity testing, the above performance characteristics have to be translated in some way. When this is not possible, suitable alternatives are provided.

3.5.1 Decision limit and detection capability

The decision and limit (CC_α) and detection capability (CC_β) are related to the analytical precision around a certain permitted limit of some substance. In case no permitted limit has been established, blanks or fortified blanks are used to determine the decision limit and detection capability (see Council Directive 2002/657/EC). The decision limit and detection capability are required to ensure that decisions about compliance of a sample are not faulty because of measurement uncertainty.

In the case of authenticity testing, the terms compliant and non-compliant would usually translate to authentic and non-authentic. In this case, however, the permitted limit cannot be expressed as a single value because the boundaries of the classes are determined by some multivariate function. It would moreover be impractical to determine the analytical precision of the method along the trajectories of such functions. It is therefore concluded that CC_α and CC_β can not be directly translated to a multivariate setting and that alternative performance characteristics have to be defined.

As an alternative to the CC_α and CC_β it is proposed to determine the predictive power of the model under additional measurement uncertainty. When for example the samples were measured in duplicate or triplicate, - and assuming the model has been built from their average values -, the effect of measurement uncertainty can be evaluated by predicting the class memberships based on the individual measurement series. The false positive and negative rate can then be determined for each individual series, which can then be compared to the criteria for acceptance as described in section 3.6.

When no duplicate or triplicate measurements are available, alternative test sets can be prepared by adding increasing amounts of noise to the existing dataset. This can for example be done by adding (Gaussian) noise to each of the variables by letting the standard deviation of the noise vary proportionally with the standard deviation of the variable. Using these sets to assess the model robustness against noise could be used to find a cut-off value for the acceptable amount of analytical variation. Because the artificial noise will in general have different properties than real noise, this approach might yield less realistic results.

At last, it is important to realize that none of these tests tell something about the generalization ability of the model (i.e. how well the model can predict new samples), and as such they are not part of the external model validation procedure. Such a test does however provide a cheap way of assessing the susceptibility of the model to analytical variation and gives a first indication of how well the model can accommodate additional variation in the data.

Table 2 Definition of relevant performance characteristics for validation of screening methods according Council Directive 2002/657/EC. Definition of stability taken from RIKILT protocol F0052. Note that the decision limit is not required for screening methods (only for confirmative methods).

Performance characteristic	Definition
Decision limit (CC_α)	The limit at and above which it can be concluded with an error probability of α that a sample is non-compliant.
Detection capability (CC_β)	The smallest content of the substance that may be detected, identified and/or quantified in a sample with an error probability of β. In the case of substances for which no permitted limit has been established, the detection capability is the lowest concentration at which a method is able to detect truly contaminated samples with a statistical certainty of 1 – β. In the case of substances with an established permitted limit, this means that the detection capability is the concentration at which the method is able to detect permitted limit concentrations with a statistical certainty of 1 – β.
Specificity	The ability of a method to distinguish between the analyte being measured and other substances. This characteristic is predominantly a function of the measuring technique described, but can vary according to class of compound or matrix.
Ruggedness	The susceptibility of an analytical method to changes in experimental conditions which can be expressed as a list of the sample materials, analytes, storage conditions, environmental and/or sample preparation conditions under which the method can be applied as presented or with specified minor modifications. For all experimental conditions which could in practice be subject to fluctuation (e.g. stability of reagents, composition of the sample, pH, temperature) any variations which could affect the analytical result should be indicated.
Stability	The susceptibility of an analytical method as a result of sample storage and analyses.

3.5.2 Specificity

In the context of the analytical methods described in Council Directive 2002/657/EC, the specificity refers to the ability of the method to discriminate between the analyte and any other compounds (see Table 2). This definition is slightly different from the definition of specificity – or true negative rate – as used in chemometrics and machine learning (see Table 1). In terms of a method for authenticity testing, the specificity could be translated as the ability to correctly classify samples from the not authentic class, whereas the sensitivity – or true positive rate – refers to the ability to correctly classify samples from the authentic class.

The true positive rate and true negative rate are often used to choose an optimal classification algorithm as well as to optimize the classification model by a leave-multiple-out cross-validation or a bootstrap approach (see section 2.4). After model optimization, the final TPR and TNR are then calculated by autoprediction (i.e. calculating the TPR and TNR using all the samples), and presented as the final power of the test. It is well known, however, that the true positive rate and true negative rate as derived in this way are optimistic estimators of the “true” model TPR and TNR.

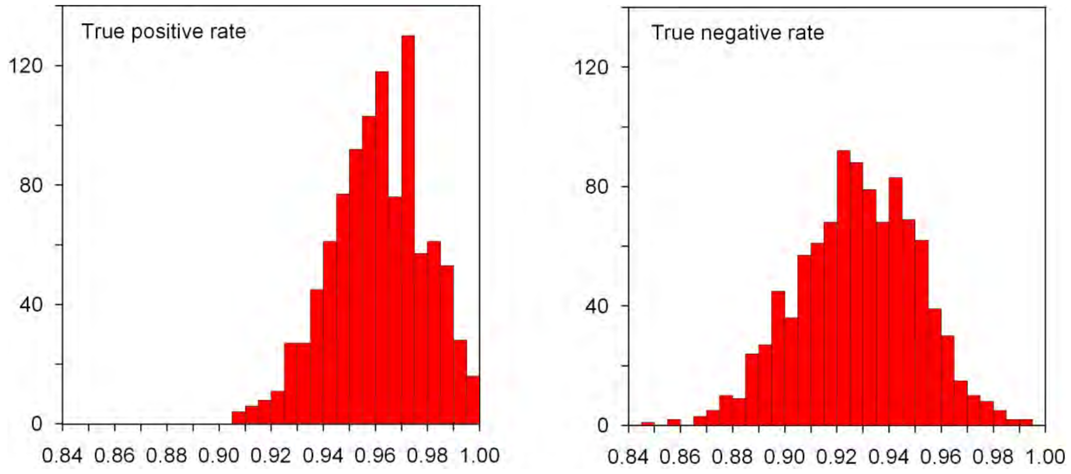

Figure 3 Bootstrap distribution of the TPR and TNR of an artificial dataset consisting of two partially overlapping clusters classified using PLS-DA with 6 latent components. Both clusters consist of 1000 samples and 20 variables of which 5 are relevant (i.e. non-noisy). Bootstrap distribution generated by 1000 realizations leaving out 200 samples. The true positive and true negative rate calculated using all samples (autoprediction) is 0.96 and 0.93 respectively.

As an alternative to the autopredicted true positive rate and true negative rate it is suggested to use the mean or median of the TPR and TNR as calculated from many non-parametric bootstrap cycles (typically >1000) of the original model. This approach does not only allow to determine the mean or median TPT and TNR, but also allows to calculate some measure of dispersion (e.g. standard variation). In non-parametric bootstrapping, a test set of N samples are repetitively selected randomly from the training set with replacement (e.g. Efron and Gong, 1983). The distribution of the true positive rate and negative rate is then established from the model outcomes at each of the cycles. Herein it is important that the test sets are sufficiently large to prevent erroneous outcomes (e.g. Brereton, 2006; Isaksson et al., 2008). An example based on data from two artificial clusters classified using PLS-DA is shown in Figure 3.

3.5.3 Ruggedness (minor variations)

The ruggedness of a method is mainly related to its susceptibility to variations in the experimental conditions. To evaluate the ruggedness of a method, samples should be measured under varying experimental conditions such as pH, time of preparation, device setting etc, which is usually done using a factorial design (see Council Directive 2002/657/EC).

A similar approach can be used to test the ruggedness of a classification model. In this case, however, one would be more interested in generalization ability of the model, - i.e. how well the it predicts the class memberships for a set of newly collected samples -, than in the effects caused by using different experimental conditions. Not so much because the effects of varying experimental conditions are deemed irrelevant, but the effects are considered of lesser importance compared to the differences within the population of authentic and non-authentic samples. Moreover, the analytical method used for fingerprinting will often be already validated, albeit maybe for a more specific purpose (e.g. determination of water content by NIR, fatty acid content by GC etc.).

Testing the model's predictive power using a new set of samples is generally referred to as external model validation. The essential questions for the external validation procedure are: I) which samples to use, and II) how many samples to include? Selection of new samples should preferably be based on knowledge about the main sources of variation within the population. Moreover, they should fall within the previously defined scope of the method (see section 3.1). In general it is advisable to collect new samples both from additional producers as well as from previously sampled producers/areas. The latter allows to evaluate the temporal effects. If deemed important, other factors could be considered as well such as the production method or storage conditions.

To determine the appropriate sample size the power analysis approach as discussed in section 2.1. could be used to provide a rough guideline. Obviously, the sample size shouldn't be too small, because otherwise the performance characteristics such as the TPR and TNR cannot be assessed with sufficient resolution. For example when only 20 samples are taken from the authentic group and 20 samples from the non-authentic group, the "resolution" of the predicted TPR and TNR is limited to 0.05 (e.g. 0.95, 0.90, 0.85 etc.).

Before the new set of measurements is used to evaluate the classification model, it is advisable to determine analytical quality of the new measurements and compare it with the results for the original ones (see section 3.6.2). Assuming that the samples were analyzed in duplicate the average standard deviation of the validation set (SD_{val}) can be calculated using:

$$SD_{val} = \frac{\sum_{j=1}^{k} \sqrt{\left(\frac{\sum_{i=1}^{N} d_{i,j}^{2}}{2N} \right)}}{k}$$

Where $d_{i,j}$ is the difference between the i-th duplicate pair for the j-th variable, N is the number of of duplicate pairs in the set of external validation samples used to validate the classification model and k is the number of variables.

3.5.4 Stability

The stability expresses the susceptibility of the method as a result of sample storage and analysis. Guidelines for evaluation of the stability can be found in Council Directive 2002/657/EC. The approach basically boils down to repeated analysis of a set of samples stored under for several periods. For testing the stability of an analyte in a matrix, sample material should be analyzed at T=0 (fresh) and after one, two, four and 20 weeks while stored at least at -20 °C or lower if required.

In the context of authenticity testing a similar approach can be followed by using a set of reference samples of both the positive and negative group. By repeated analysis of these reference samples, for example according to the scheme in Table 4, the average standard deviation $SD_r(T)$ can be calculated from the difference of each measurement at T>0 and the measurement at T=0 using:

$$SD_r(T) = \frac{\sum_{j=1}^{k} \sqrt{\left(\frac{\sum_{i=1}^{N} d_{i,j}^2(T)}{2N} \right)}}{k}$$

Where $d_{i,j}(T)$ is the difference between the i-th duplicate pair for the j-th variable, N is the number of duplicate pairs and k is the number of variables. Note that a duplicate pair in this context refers to the measurement of a reference sample at T=0 and at T>0. The time for which the $SD_r(T)$ falls within the limits provided in section 3.6.2 can then be taken as the maximum storage time.

Table 4 Example of a measurement scheme of different aliquots for testing the stability.

Group	T=0 (fresh)	T = 1 week	T = 2 weeks	T = 4 weeks	T = 20 weeks
Positive	5	5	5	5	5
Negative	5	5	5	5	5

3.5.5 Additional performance characteristic; permutation test

A very useful measure of model performance can be derived a using permutation test. This test evaluates whether the specific classification of the individuals in the two designed groups is

significantly better than any other random classification in two arbitrary groups (Golland et al. 2005; Mielke and Berry 2001). Using this approach, the distribution of the true positive rate and true negative rate of the model are determined during many cycles in which each time a different randomly ordered classification scheme is used.

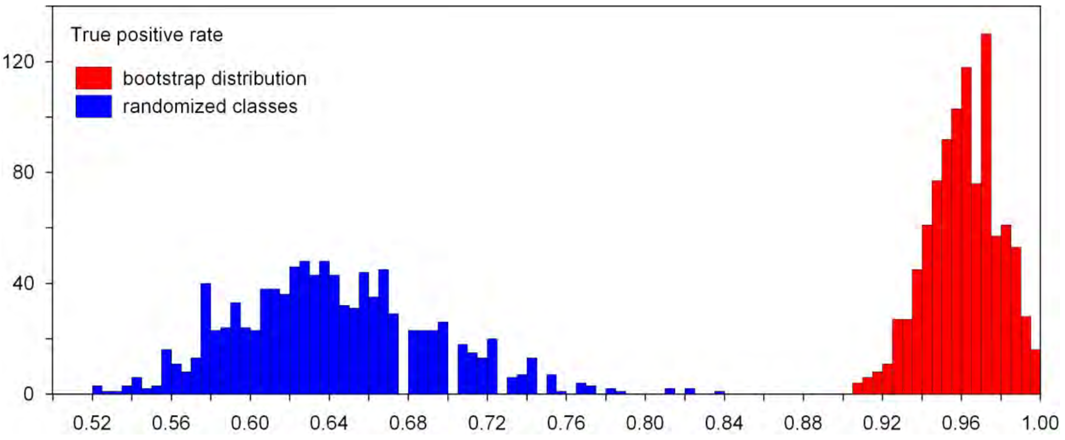

Figure 4 Distribution of the true positive rate of a PLS-DA classification of a two-cluster dataset using many random permutations of the true class labels (left), and the bootstrap distribution of the TPR using the true classes and leaving out 20% of the data during each cycle (right; same dataset as used for Figure 3). The 95th percentile of the TPR of the randomized classification is 0.75, which is considerably lower than the lower range of the TPR as determined by bootstrapping the model using the true classification.

Using the calculated mean and standard deviation, the upper limit of e.g. the 95% confidence interval can be established for both the true positive rate and true negative rate. Alternatively, the upper 95th or 99th percentile of the distribution can be used as the upper limit of the one-sided 95% or 99% confidence interval. This gives an objective measure against which the mean TPR and TNR can be evaluated against (see section 3.6.1). An example based on a PLS-DA classification of an artificial dataset is given in Figure 4.

3.6 Criteria for acceptance of performance characteristics

In section 3.5, a number of alternative performance characteristics were reviewed which are summarized in Table 5. The next step in model validation is to compare these performance characteristics with predefined criteria for acceptance. For regular screening methods, these criteria are laid down in Council Directive 2002/657/EC. These are however not suitable for evaluation of the alternative performance characteristics as reviewed in this report, and alternative criteria are discussed in section 3.6.2.

3.6.1 Criteria for acceptance according to Council Directive 2002/657/EC

The performance characteristics required for validation of screening methods according to Council Directive 2002/657/EC include the detection capability, the specificity, ruggedness and stability. For three of the characteristics, - including the detection capability, ruggedness and stability -, no specific criteria for acceptance are provided by the Council Directive. This is because these characteristics basically describe the limitations of the method in terms of measurement uncertainty, factors that influence the results as well as the maximum time a sample can be stored. It is not possible to formulate general criteria for these characteristics and it depends on the requirements of the user whether the performance of the characteristics is acceptable or not.

The only performance characteristic for which a strict criterion is formulated concerns the specificity of the method. According to Council Directive 2002/657/EC the chance of a Type II error should be < 5%, which implies that the TPR should be larger than 95%). This criterion is used as well for purpose of validation of a method for authenticity testing (TPR_{mean} > 0.95%; see Table 5).

Table 5 Overview of performance characteristics for screening methods according to Council Directive 2002/657/EC and alternative performance characteristics for validation of classification models for authenticity testing and criteria for acceptance.

Performance characteristic	Alternative for authenticity test	Criteria for acceptance
Detection capability	Model robustness against analytical variation/noise	$TPR_{var} > TPR_{min}$ $TNR_{var} > TNR_{min}$
Specificity	Mean TPR and TNR from bootstrap distribution	$TPR_{mean} > 0.95$ TNR_{mean} > user defined TPR
Ruggedness	Analytical variation of validation set and model performance under external validation	$SD_{val} < SD_{lim}$ $TPR_{val} > TPR_{min}$ $TNR_{val} > TNR_{min}$
Stability	Stability (modified approach)	$SD_r(T) < SD_{lim}$
-	Upper limit of TPR and TNR from permutation test	$TPR_{upper} < TPR_{min}$ $TNR_{upper} < TNR_{min}$

3.6.2 Alternative criteria for acceptance

In the previous sections, the true positive rate and true negative rate determined at various instances:

1. TPRvar and TNRvar: TPR and TNR calculated by predicting the class memberships of the individual duplicate or triplicate measurements using the final classification model.
2. TPRmean and TNRmean; mean (or median) TPR and TNR derived from bootstrap distribution of the TPR and TNR.
3. TPRval and TNRval: TPR and TNR calculated by predicting the class memberships for a set of new samples using the final classification model.
4. TPRupper and TNRupper: The upper limit (e.g. 95th percentile) of the distribution of the TPR and TNR derived using randomized class memberships (permutation test).

As a criterion for acceptance of the different TPRs and TNRs calculated under point 1 to 4, a value for TPR_{min} and TNR_{min} should be defined. For this purpose it is proposed to use the lower e.g. 5-th percentile of the bootstrap distribution of the TPR and TNR.

Furthermore, the criterion for acceptance of the TPR_{mean} is that it should be larger than 0.95, which is similar to the criterion for the sensitivity as defined by Council Directive 2002/657/EC. For acceptance of the TNR_{mean} no criterion is provided, and this should be determined by the user itself.

According to the set-up described in section 3.5.3 and 3.5.4, the average standard deviation is calculated for both the duplicate measurements of the external validation set as well for the analysis of the references samples after different storage times. These values should be compared to some previously established measure of the analytical variation of the method. Provided that the samples that were used to build the model were analyzed in duplicate, the average standard deviation of the duplicates of the original data can be used to formulate a criterion. A criterion $SD_{\lim}$ is here defined as:

$$SD_{\lim} = 1.96 \frac{\sum_{j=1}^{k} \sqrt{\left(\frac{\sum_{i=1}^{N} d_{i,j}^2}{2N} \right)}}{k}$$

Where $d_{i,j}$ is the difference between the i-th duplicate pair for the j-th variable, N is the number of of duplicate pairs in the set of samples used to build the classification model and k is the number of variables.

Obviously, the analytical variation in the external validation set should be smaller than the criterion provided above. If this is not the case, the external validation procedure will most likely fail to produce a good prediction. For determining the stability of the method, the storage period before $SD_r(T) > SD_{\lim}$ should be denoted as the maximum storage time for the sample material. It is however possible that $SD_r(T)$ does not show a clear trend, for example because it decreases again after some period. In this case, the data should be rigorously checked for outlying or anomalous results.

3.7 Experimental design

The experimental design for validation of a method for authenticity testing involves assessment of relevant performance characteristics (section 3.5) which are to be compared to the criteria for acceptance (section 3.6). For the order in which the different performance characteristics are described we have followed Council Directive 2002/657/EC. In practice it is however more cost-efficient to first determine the characteristics that do not require additional sampling and analysis which include:

- The predicted true positive and negative rate based on the duplicate measurements (section 3.5.1)
- The mean/median of the true positive and negative rate as determined by bootstrapping (section 3.5.2)
- The upper limit of the true positive and negative rate as determined by the permutation test (section 3.5.5).

Only when these parameters give acceptable results, other characteristics can be determined such as model performance under external validation (section 3.5.3) and stability (section 3.5.4).

External validation is probably the most important aspect of the validation of any classification model and special attention should be paid to setting up a rigorous sampling design. Apart from estimating the minimum amount of samples to be collected from both the authentic and non-authentic populations (see section 3.5.3), general guidelines for a design are difficult to provide. In each specific case, different choices will be made depending on the scope of the test as well as the sources of variation within the population.

At last it should be stressed that the chemical analysis of the external validation samples should be performed in the exact same manner as for the original samples. The same applies to the data pre-treatment, outlier detection and feature selection steps. Preferentially all data pre-treatment feature selection, model building and model validation procedures are therefore laid down in the form of a number of scripts or as a mark-up language such as PMML.

3.8 Description of deviations from standard protocols for method validation

In this section of the validation plan, any deviations from the standard protocol should be described. Because there is no standard protocol available for validation of the methods described in this report, it is advisable to describe the complete procedure used for model development and validation in the validation plan.

4 References

A0906, 2010. Het vaststelling van prestatiekenmerken op basis van initiële validatiekenmerken. RIKILT RSV (internal document): 1 - 10.

Baumann, K., 2003. Cross-validation as the objective function for variable-selection techniques Trends Anal. Chem. 22:. 395-406.

Baroni, M., Clementi, S., Cruciani, G., Costantino, G., Riganelli, D., Oberranch, E., 1992. Predictive ability of regression models. Part II: Selection of the best predictive PLS model. J. Chemom. 6: 347-356.

Berrueta, L.A., Alonso-Salces, R.M., Héberger, K., 2007. Supervised pattern recognition in food analysis. Journal of Chromatography A 1158: 196–214.

Bertelli, D., Lolli, M., Papotti, G., Bortolotti, L., Serra, G., Plessi, M., 2010. Detection of Honey Adulteration by Sugar Syrups Using One-Dimensional and Two-Dimensional High-Resolution Nuclear Magnetic Resonance. J. Agric. Food Chem. 58 (15): 8495–8501.

Bevin, C.J., Fergusson, A.J., Perry, W.B., Janik, L.J., Cozzolino, D., 2006. Development of a Rapid “Fingerprinting” System for Wine Authenticity by Mid-infrared Spectroscopy. J. Agric. Food Chem. 54 (26): 9713–9718.

Brereton, R.G., 2006. Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. Trends in Analytical Chemistry, Vol. 25, No. 11: 1103 – 1111.

Charlton, A.J., Farrington, W.H.H., Brereton, P.B., 2002 Application of 1H NMR and Multivariate Statistics for Screening Complex Mixtures: Quality Control and Authenticity of Instant Coffee. J. Agric. Food Chem. 50 (11), pp. 3098–3103.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences (2nd ed.), Lawrence Erlbaum Associates, New Jersey.

Cruciani, G., Baroni, M., Clementi, S., Costantino, G., Riganelli, D., Skagerberg, B., 1992. Predictive ability of regression models. Part I: Standard deviation of prediction errors (SDEP). J. Chemom. 6: 335-346.

Commission Decision 2002/657/EC, OJ L 221, 17.08.2002. URL: http://eur-lex.europa.eu/LexUriServ /LexUriServ.do?uri=OJ:L:2002:221:0008:0036 :EN:PDF [last accessed December 2010]

Defernez, M., Kemsley, E.K., 1997. The use and misuse of chemometrics for treating classification problems. Trends Anal. Chem. 16, 4: 216-221.

Efron, B., Gong, G., 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician, Vol. 37, No. 1: 36-48.

F0052, 2010. Uitgangpunten bij validatie van analysemethoden. RIKILT RSV (internal document): 1-10.

Golland, P., Liang, F., Mukherjee, S., Panchenko, D., 2005. Permutation tests for classification. Lecture notes in Computer Science 3559: 501–515.

Hali, A.S., 1994. A modification of a method for the detection of outliers in multivariate samples. J. Roy. Stat. Soc. B56: 393 – 396.

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. Pattern Analysis and Machine Intelligence 19, no.2:153-158.

Mielke, P.W. Jr, Berry, H., 2001. Permutation methods: A distance function approach. New York: Springer.

Morse, D.T., 1999. Minsize2: a Computer Program for Determining Effect Size and Minimum Sample Size for Statistical Significance for Univariate, Multivariate, and Nonparametric Tests. Educational and Psychological Measurement 59: 518 - 531.

Isaksson, A., Wallman, M., Göransson, H., Gustafsson, M.G., 2008. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters 29: 1960–1965.

ISO 17025, 1999. General requirement for the competence of calibration and testing laboratories.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., De Jong, S., 1998. Handbook of chemometrics and Qualimetrics. Two Volume set, Elsevier Amsterdam.

Møller, J.K.S., Catharino, R.R., Eberlin, M.N., 2005. Electrospray ionization mass spectrometry fingerprinting of whisky: immediate proof of origin and authenticity. Analyst, 130: 890–897.

Otto, M., 1999. Chemometrics. Wiley-VHC, Weinheim.

Singh, A., 1996. Outliers and robust procedures in some chemometric applications. Chemom. Intell. Lab. Syst. 29: 75 – 100.

Van Ruth, S.M., Villegas, B., Rozijn, M., Akkermans, W., van der Kamp, H., 2010. Prediction of the identity of fats and oils by their fatty acid, triacylglycerol and volatile compositions using PLSDA. Food Chemistry, 118, 948-955.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H., 2010. Advancing Feature Selection Research - ASU Feature Selection Repository. URL: http://featureselection.asu.edu /featureselection_techreport.pdf [last accessed December 2010].

RIKILT - Institute of Food Safety is part of the international knowledge organisation Wageningen UR (University & Research centre). RIKILT conducts independent research into the safety and quality of food. The institute is specialised in detecting and identifying substances in food and animal feed and determining the functionality and effect of those substances.

RIKILT advises national and international governments on establishing standards and methods of analysis. RIKILT is available 24 hours a day and seven days a week in cases of incidents and food crises.

The research institute in Wageningen is the National Reference Laboratory (NRL) for milk, genetically modified organisms, and nearly all chemical substances, and is also the European Union Reference Laboratory (EU-RL) for substances with hormonal effects.

RIKILT is a member of various national and international expertise centres and networks. Most of our work is commissioned by the Dutch Ministry of Economic Affairs, Agriculture and Innovation and the new Dutch Food and Consumer Product Safety Authority. Other parties commissioning our work include the European Union, the European Food Safety Authority (EFSA), foreign governments, social organisations, and businesses.

More information: www.rikilt.wur.nl